Clim. Past, 8, 1141–1151, 2012
www.clim-past.net/8/1141/2012/
doi:10.5194/cp-8-1141-2012

Climate of the Past

Identification of climatic state with limited proxy data

J. D. Annan and J. C. Hargreaves

RIGC/JAMSTEC, 3173-25 Showamachi, Yokohama, Japan

Correspondence to: J. D. Annan (jdannan@jamstec.go.jp)

Received: 20 January 2012 – Published in Clim. Past Discuss.: 31 January 2012
Revised: 29 May 2012 – Accepted: 15 June 2012 – Published: 10 July 2012

Abstract. We investigate the identifiability of the climate by limited proxy data. We test a data assimilation approach through perfect model pseudoproxy experiments, using a simple likelihood-based weighting based on the particle filtering process. Our experimental set-up enables us to create a massive 10 000-member ensemble at modest computational cost, thus enabling us to generate statistically robust results. We find that the method works well when data are sparse and imprecise, but in this case the reconstruction has a rather low accuracy as indicated by residual RMS errors. Conversely, when data are relatively plentiful and accurate, the estimate tracks the target closely, at least when considering the hemispheric mean. However, in this case, our prior ensemble size of 10 000 appears to be inadequate to correctly represent the true posterior, and the regional performance is poor. Using correlations to assess performance gives a more encouraging picture, with significant correlations ranging from about 0.3 when data are sparse to values over 0.7 when data are plentiful, but the residual RMS errors are substantial in all cases. Our results imply that caution is required in interpreting climate reconstructions, especially when considering the regional scale, as skill on this basis is markedly lower than on the large scale of hemispheric mean temperature.

1 Introduction

Reconstructions of climate variation over recent centuries make an important contribution to our understanding of climate change, and in particular help us to place the recent anthropogenically forced changes in the context of natural variability. Therefore, it is important that we have a sound understanding of the reliability and precision of these reconstructions. Prior to the recent instrumentally observed interval (from around 1850 to the present day), direct measurements of climatic variables are not generally available, and therefore the primary sources of data are a number of proxy measurements of various types, with tree-rings being one of the best-known. Compared to the modern observational network, these proxy data are extremely limited, with there being typically tens to hundreds of observations (each representing a seasonal or annual average value) available globally during a single year. Numerous reconstructions have been presented for the mean temperature of the Northern Hemisphere, where proxies are most numerous (e.g. Jansen et al., 2007, Fig.6.10), and rather less commonly for global temperature. Most existing reconstructions are based on primarily statistical methods, in which a linear regression is used to relate the proxy data to the climate variable of interest. Tingley et al. (2012) provide a comprehensive review of the wide range of statistical methods in use, which continue to be the subject of substantial investigation and debate (e.g. Christiansen et al., 2009; Rutherford et al., 2010; Christiansen et al., 2010; Smerdon et al., 2011).

More recently, an alternative approach to climate reconstruction has been developed, in which the proxy data are assimilated into a climate model, generating what is generally referred to as a "reanalysis" of the climate state (Goosse et al., 2006; Widmann et al., 2010). In principle, such an approach could have several notable advantages over a purely statistical method. By using a dynamical model, physical relationships between climatic variables, including unobserved variables, can be directly generated from physical laws, rather than having to be inferred from limited noisy data or approximated via statistical relationships. An additional benefit arises from the temporal relationships embodied in the model: an estimate of the climate state at a given time can be enhanced by data observed both before the synoptic time (as in filtering methods) and even from data observed after this time. Such an approach is known as smoothing, but note that

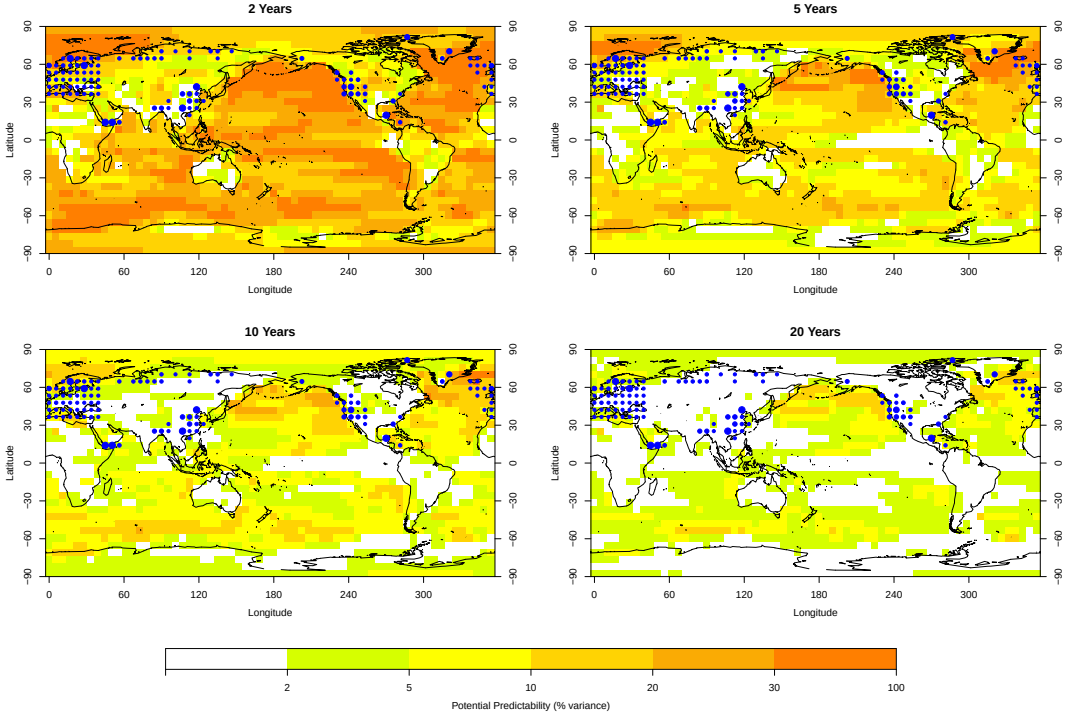

Fig. 1. Potential predictability. σ_v^2/σ^2 for different time scales, where σ_v^2 is the variance of the long time scale and σ^2 is the total variance. Also indicated are data locations (blue dots) with the size of the dot proportional to the length of time series available, which ranges between 500 yr (smallest 65 dots) and 2000 yr (largest 6 dots).

this does not refer to a simple smoothing filter such as a moving average, but rather the transfer of information through time according to the physical laws embedded in the model. In practice, however, the temporal influence of data is limited by the predictability time scale of the system. Regression-based methods do not usually account for this effect at all, with the estimate for a particular year relying purely on the proxy data associated with that time.

However, data assimilation methods also have many practical limitations. Perhaps most prominently, the computational costs may be large. Attempts to reduce the computational load typically require simplifying assumptions and approximations, which may reduce the accuracy of the results. Moreover, climate models have significant imperfections, such that the dynamical relationships that they impose on the results may not be good representations of the behaviour of the real climate system. Another potentially negative aspect of the data assimilation approach is that its output is at least in part a model product, and therefore may not be so useful as an independent test of model performance. Nevertheless, the exciting potential of such methods certainly justifies investigation into their strengths and weaknesses.

In this paper, we consider the potential of data assimilation methods to generate accurate reconstructions using limited observations. We use a particle-based approach, although the use of a formally optimal methodology and massive ensemble means that some of our conclusions must apply more generally to all Bayesian estimation methods. We have two main goals. Primarily, we investigate the precision with which it is possible to estimate the hemispheric climatic state with limited observations. Further, we also consider the viability of particle-based methods (in particular, in respect of the required ensemble size) to undertake this task. Our investigations are complementary to those of Dubinkina et al. (2011) who used a more extensive data set based on the recent observational period. We adopt an identical twin paradigm, in which pseudoproxy observations are generated from a model run (Smerdon, 2012), so as to focus specifically on the methodological aspects and theoretical performance limits.

2 Model and data

This work is based on a 101-member, 140-yr integration of the Earth system model of intermediate complexity, LOVECLIM (Renssen et al., 2005). This model primarily consists of a 3-level quasi-geostrophic atmosphere at T21 resolution, coupled to a 3°, 20 level, sea-ice-ocean general circulation model. The essential features of this model for our purposes are that, in contrast to simpler energy-balance models, it plausibly simulates the chaotic internal variability of coupled atmosphere/ocean/sea-ice system, while (unlike state-of-the-art GCMs) remaining computationally efficient enough for

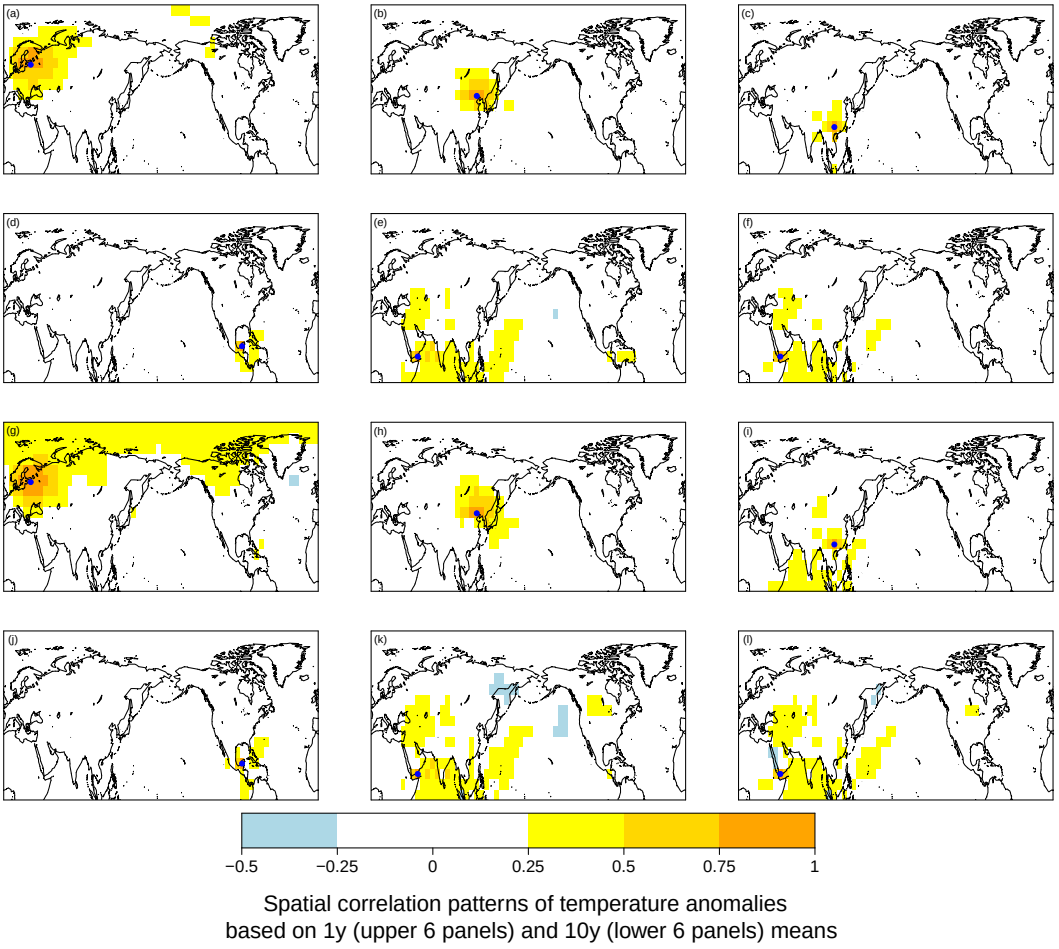

Fig. 2. Spatial correlations of annual anomalies at specified grid points (corresponding to the locations of the longest data time series).

long-term ensemble integrations to be practical. For the pseudoproxy data, we base this on the screened proxy network assembled by Mann et al. (2008), which has been previously used for data assimilation experiments with the LOVECLIM model by Goosse et al. (2010). As in that work (and most previous temperature reconstructions over the past millennium), we focus on the Northern Hemisphere where the proxy data are most plentiful. Figure 1 shows the locations of data that we use, illustrating how the density of proxies changes over time. Note that the proxy locations here are shown after binning to the model grid. In the original data set, there may have been multiple records within a single grid cell. Table 1 presents the change in data availability over time, which rises from only 6 locations where data are available for the full time span years, to 112 for the last 500 yr.

Given the sparse nature of the proxy network, one fundamental task of the data assimilation process, which underpins the generation of a hemispheric or global mean temperature anomaly, is spatial interpolation. As a first check and demonstration of the viability of the process, therefore, we consider the spatial coherence of temperature anomalies, and explore how this compares to the sparsity of the data. Figure 2 shows the spatial correlation structure of anomalies at the location of each of the 6 longest proxy data series in turn. All the coloured regions are highly statistically significant. In fact, with our large sample size, the threshold for significance at the 5 % level is a correlation of magnitude around 0.05, but such low values are of little practical value and therefore have been left blank. Around each location there is an approximate "bullseye" of high correlation, the precise shape and size of which varies across the globe but which appear roughly compatible with the O (1200 km) smoothing radius used in some observational analyses (Hansen and Lebedeff, 1987).

As a further test of the model, we check its potential predictability over the multi-annual time scale. To do this we follow the approach of Boer and Lambert (2008) and calculate the extent to which the variance of k-year means contributes to the total variance, in excess of that which would be expected if the annual temperature anomalies were serially uncorrelated white noise. Potential predictability in this sense is therefore not necessarily a direct measure of predictability of the system, but has been widely studied and

Table 1. Number of grid points containing observational data, at 100 yr intervals through the last two millennia.

50	6
150	7
250	7
350	7
450	8
550	8
650	9
750	10
850	10
950	10
1050	16
1150	22
1250	24
1350	28
1450	47
1550	112
1650	112
1750	112
1850	112
1950	112

found to correspond well with predictability in the classical sense, as discussed by Boer and Lambert (2008). Figure 1 shows the potential predictability for intervals ranging from 2 to 20 yr. This compares extremely well with Fig. 4 of Boer and Lambert (2008), in which several of the CMIP3 models were analysed. This analysis suggests that there may be some potential for the estimate of the climate state in a given year to be improved with observations from adjacent years (as could be achieved with filtering or smoothing methods from data assimilation). However, it must be noted that the potential predictability is highest over the ocean areas, whereas the data are primarily restricted to the land. This measure of potential predictability is primarily an indication of the local thermal inertia. Other aspects of model performance that may also be relevant to predictability on interannual time scales (such as modes of variability linked to ENSO and NAO) are investigated in, for example, Selten et al. (1999); Goosse et al. (2001); Timmermann et al. (2005).

3 Method

The fundamental basis of almost all data assimilation is the application of Bayes' theorem to update a prior estimate in the light of observational evidence, thereby forming a posterior estimate. The core of our approach is the simple likelihood weighting algorithm which underpins the particle filter (Arulampalam et al., 2002). In this method, in order to estimate the climate state for a given year, samples are drawn from the prior, and then each sample is weighted according to a likelihood function, which depends on the fit of the sample to the specific observations available. In the common case of Gaussian uncertainties on each observation, the weight is defined by the familiar exponential $\exp(-C)$ in which C is the quadratic cost function $C(M, O) = 0.5 \times (M - O)^T \Sigma^{-1} (\boldsymbol{M} - O)$ where $\boldsymbol{M}$ is the vector of modelled outputs corresponding to the observations O, and Σ is the covariance matrix of observational errors. In this work, we use the standard assumption that observational errors are independent, simplifying the cost function to the sum of diagonal terms $C(\boldsymbol{M}, O) = 0.5 \times \sum_i ((m_i - o_i)/\sigma_i)^2$. While some proxies may relate to seasonal parameters (such as growing season temperatures), we focus here purely on annual means for both data and model output.

After normalising the weights so that they sum to unity over the ensemble, statistics of interest such as posterior mean and variance are easily calculated from the weighted ensemble (for example, the posterior mean of a particular output x is estimated as the weighted average $\sum_i w_i x_i$, with the sum being taken over all ensemble members x_i and their associated weights w_i). In principle, this ensemble can then be used as the basis for a prior for the subsequent year, and integrated forwards in time. However, given the limited predictability of the system, an alternative approach, as used by Bhend et al. (2012), is to simply revert to the climatological prior for each individual year. We initially follow this procedure, and consider the potential of the more conventional sequential approach later. One benefit of our approach is that the entire assimilation can be performed off-line, after the ensemble integration has been completed. This enables us to investigate the effects of changing various details of the procedure (such as observational density and accuracy) at little additional computational cost. Our approach does, however, eliminate the temporal consistency of the model simulation, which would be preserved by using the posterior as the prior for the subsequent year.

Our method has some similarity to the analogue method in which a historical database of states is queried to find the best fit to observations (Barnett and Preisendorfer, 1978). Instead of choosing the single simulation with the best fit to the data, however, we calculate a weight for each member of the full ensemble, which generates a probabilistic posterior distribution. While Goosse et al. (2006, 2010) also used a degenerate particle filtering approach in which the best simulation was selected, Dubinkina et al. (2011) implemented a probabilistic approach more similar to ours, but using a far more dense network of recent observations. In contrast to the latter experiments, we make no allowance for model error, since it is absent in our identical twin experiments.

All models were initialised by making small perturbations to a long equilibrium integration, with the first 20 yr discarded as a spin-up phase. External forcing is held constant, so the ensemble samples the internal variability of the model. The final 20 yr of each integration were reserved for analysis of the predictive performance of the algorithm. We consider the case of an externally forced response in Sect. 4.3. The use of unforced simulations allows us to treat all model

years exchangeably, and thus the ensemble size for the prior is 10 000 (albeit the temporal correlation revealed in the potential predictability analysis results in a slightly smaller effective sample size), which is far greater than would be practical otherwise.

One model run at a time is used as the truth, with the remaining ensemble of 100 integrations of 100 yr used for the prior. We generate pseudoproxies from the "truth" run, by adding random noise to the model outputs at the appropriate grid points. For the standard case, we use white noise with a magnitude of 2.5 times the standard deviation of annual temperatures at that location, which results in a "signal-to-noise ratio" of 0.4, in line with the bulk of the paleoclimate literature (e.g. Mann et al., 2008; Smerdon et al., 2011). This choice implicitly defines the scaling ratio (or more complex nonlinear operator) which would in practice be required to convert the units of proxy measurements (e.g. nondimensionalised tree ring widths) into temperature. We do not consider the issue of proxy calibration further in this work. We note for the avoidance of confusion that the usage of the phrase "signal-to-noise ratio" in the paleoclimate literature, also adopted here, differs slightly from the engineering literature where it originates in that, for the latter, the ratio is generally defined in terms of power (i.e. variance), and thus amounts to the square of the paleoclimatic convention.

We emphasise that the likelihood weighting method, although simple, actually provides an exact application of Bayes' theorem in the limit of infinite ensemble size (Arulampalam et al., 2002). Its only disadvantage – albeit an overwhelming one in many cases – is its requirement for a large ensemble, as its efficiency is rather low. The problem is that, in many practical applications, the weights will be concentrated on a small proportion of the ensemble (sometimes vanishingly small), which can lead to large sampling error or even the phenomenon of "filter collapse", where the weight is focussed on a single sample and thus provides no meaningful probabilistic information (Bengtsson et al., 2008; Snyder et al., 2008). However, when the ensemble size is adequate – a point which we investigate in Sect. 4 – the method generates the correct, optimal solution to the estimation problem, which cannot be bettered by a more sophisticated algorithm, be it an ensemble-based method such as the ensemble Kalman filter or other particle filtering, variational or optimal interpolation-based methods. Simply put, the likelihood-based weighting correctly generates the full probabilistic posterior (including its uncertainty) arising from a given prior and likelihood function.

4 Experiments and results

4.1 Reconstruction of hemispheric mean temperature and spatial pattern

As mentioned above, we first consider the case of unforced internal variability. We focus on results from 4 epochs, which cover a wide range of data densities: 500–599 AD (8 grid points), 1000–1099 AD (16 grid points), 1400–1499 AD (47 grid points) and 1700–1799 AD (112 grid points). In our identical twin testing, the only difference between these epochs is that the data density and location differ between them. Each panel of Fig. 3 shows the 100-yr time series of the hemispheric mean temperature for one specific choice of truth run (blue line) and the estimated reconstruction (red line, with one standard deviation uncertainty bounds). The truth run is identical for each panel, the only differences between these experiments being the location and number of pseudoproxy data points used in the estimation, and the randomly sampled proxy errors. The correlation between posterior mean and target time series for each epoch, averaged over all experiments, is presented for each panel. All of these values are highly significant, with the correlation never being negative in any individual experiment. However, it can also be seen in the top two panels that for these small numbers of data points, the posterior is little changed from the climatological prior (which has a spread of about ± 0.2 around its mean of zero). As the data volume increases in the lower two panels, the posterior converges towards the target, but even with the maximum number of 112 data points, substantial discrepancies can easily be seen. While the correlations seem high, the RMS difference between posterior mean and target is only slightly reduced compared to the climatological prior, with the reduction in RMS error increasing from 4 % to 23 % as the data density increases. Therefore, it seems that the use of correlation as compared to residual RMS error gives a rather different impression of the level of performance. There is certainly some measurable skill even with the smallest number of data points, but there is also substantial discrepancy between the reconstruction and the target for any given year.

It is also apparent that the temporal variability of the posterior mean is substantially lower than that of the target. Loss of variance has been widely found in climate reconstruction methods (e.g. Smerdon et al., 2008), and may be unwelcome in some contexts, such as when we wish to use climate reconstructions to estimate the natural variability of the system. However, we note that it is also an inevitable consequence of the paradigm of error-minimising Bayesian estimation. At least within this paradigm, any attempt to preserve the temporal variance of the target in the reconstruction (e.g. Christiansen and Ljungqvist, 2011) cannot simultaneously generate annual error-minimising estimates of the climatic state, since such an estimate will necessarily be shrunk towards the prior mean. Hence, it is important to be clear about both the

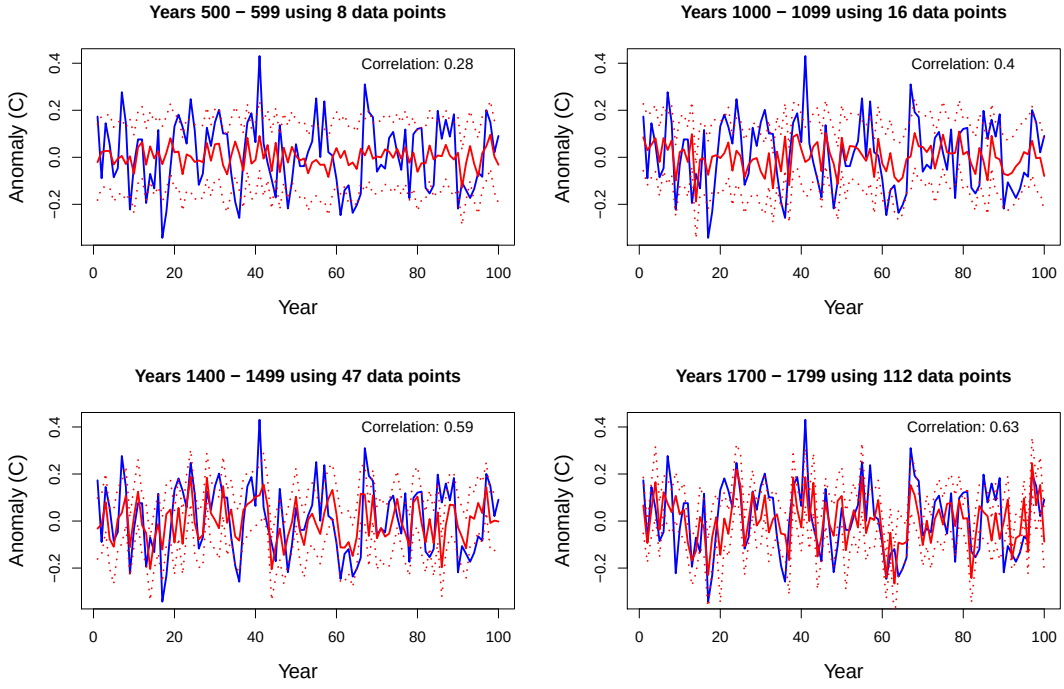

Fig. 3. Reconstruction of NH annual mean temperature anomaly in the case of internal variability, based on different number of proxy data points. Blue lines indicate target (identical in each panel). Solid red line indicates reconstruction, with dotted lines showing the ±1 standard deviation uncertainty. Correlations are averages over multiple experiments for each epoch.

goals of such an estimation, and the interpretation of the results.

The effective posterior ensemble sizes for the four panels (calculated via the standard formula $N_{\text{ef}} = 1/\sum w_i^2$ where w_i are the normalised weights on the ensemble) are around 4100 for the first panel, then 1800, 170 and 18 for the remaining panels respectively. The very large effective ensemble sizes for the first two panels, which are a large fraction of the prior sample size of 10 000, are an indication that the limited and imprecise proxy data available in these epochs do not distinguish very clearly between the prior samples, consistent with the relatively poor representation of the target by the posterior. Conversely, an ensemble as small as 18 members implies that in this case the observations are restricting the posterior to a small subset of the prior. In this case, however, sampling uncertainty may be becoming a significant factor, as 18 samples cannot be expected to accurately characterise the true posterior. However, the situation is still far less serious than seems to be the case for modern numerical weather prediction, where a prior sample size of 10 000 is argued to be completely inadequate (Bengtsson et al., 2008; Snyder et al., 2008).

As well as reconstructing the hemispheric mean temperature, we can examine how well the method reconstructs the spatial field of temperature anomalies. This result is presented in Fig. 4, where the results are presented in terms of the error on the posterior mean. These results are normalised relative to the error of the prior mean, meaning that the posterior is more accurate than the prior when the value drops below unity. There is a widespread reduction in error over large areas for all epochs. In the first two panels, however, this reduction in error appears very modest over much of the Northern Hemisphere, only dropping below the 90 % level in the immediate neighbourhood of the data points (consistent with Fig. 2). The third panel shows larger areas with substantial error reduction spreading for some distance away from the observations. Even in this case, there are large areas with marginal change, but it seems likely that the small regions where the error is actually greater than 1 are due to noise, given the finite sample from which these statistics were calculated. The final panel, however, shows a much larger region over the tropical ocean where the error of the posterior is consistently greater than that of the prior. This is a clear indication of sampling error due to the posterior ensemble being too small. It is notable that, even though Fig. 3 indicates that this experiment has skill in reproducing the hemispheric mean temperature anomaly, there are large regions where the reconstruction actually has negative skill. These results are in no way mutually inconsistent, as the local errors in the skill-free regions will tend to average out in the hemispheric mean, allowing the skilful area to impart some genuine signal into the reconstruction. They do, however, suggest that care is required in interpreting what, if anything, can be learnt about the climate on a regional basis, especially at some distance from observations. These results support those of Smerdon et al. (2011) who also showed that the skill of a large-scale

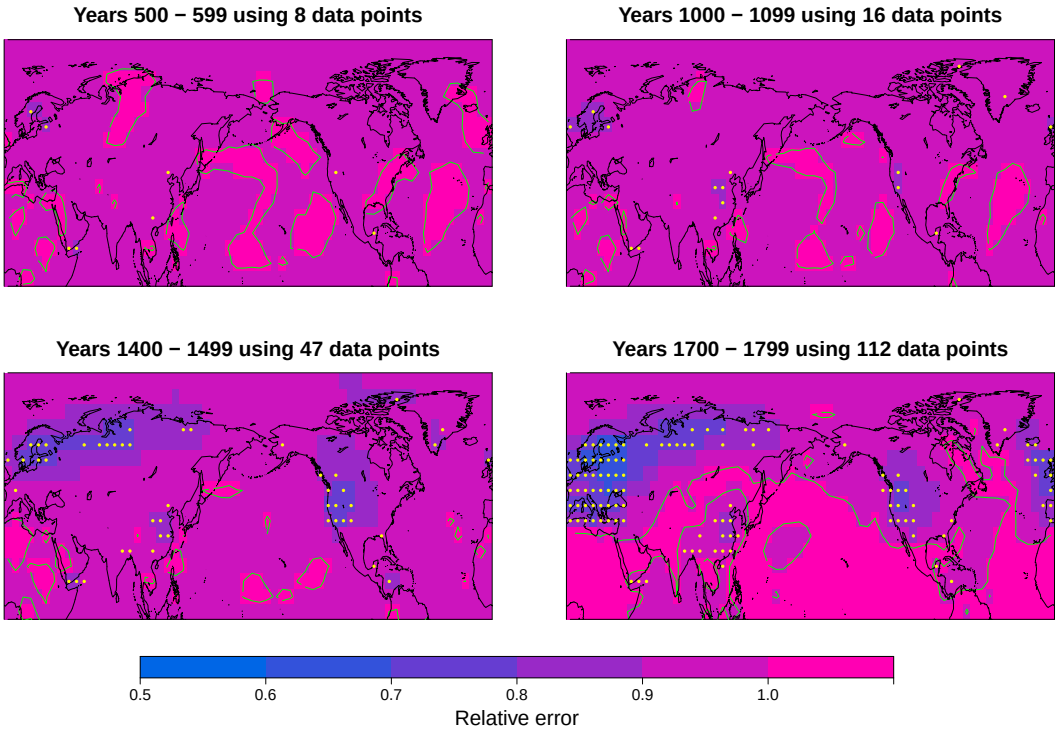

Fig. 4. Maps of normalised RMS reconstruction error for NH annual mean temperature anomalies in the case of internal variability, based on different number of proxy data points. Errors are normalised to standard deviation of climatology, with green contour indicating a normalised error of 1 (i.e. that the reconstruction is neither more nor less accurate on average than the prior climatological mean).

mean does not necessarily translate into good performance on the regional scale.

4.2 Predictive performance

Returning to the potential of data assimilation methods to improve skill though temporal smoothing, we can also consider the forecast skill arising from the use of the posterior estimate as initial conditions (prior) for the prediction of the following year. The skill (not shown here) is essentially zero for the first three cases. In these cases, the rather uncertain estimates have basically reverted to the climatological prior. In the final case, the ensemble has started to collapse, and its forecast skill is actually negative over almost the entire hemisphere due to sampling noise. While a larger ensemble could generate more skilful results, this could require a computationally infeasible ensemble (Bengtsson et al., 2008; Snyder et al., 2008). Thus, it does not seem that any sequential based or smoothing method will generate additional skill in reconstructions with this data set. This is perhaps not entirely surprising given that modern interannual prediction systems show very limited skill unless substantial volumes of ocean data are assimilated (e.g. Dunstone and Smith, 2010). However, it remains possible that proxy-based reconstructions, which use ocean-based data (such as coral records)or other data that exhibit some clear multi-year predictability, could benefit from such an approach.

4.3 Forced response

In the experiments described in Sect. 4.1, forcing was held constant and thus the procedure only accounted for the internal variability of the model. While this allowed us to take an off-line approach to the calculations, it ignores the externally forced component, which is of substantial importance in real applications. Therefore, we now extend the method to include this aspect, by superimposing an externally forced response onto the existing ensemble under the assumption that this can be considered linearly additive to the internally generated variability of the model. While this is obviously a simplification of the real system, it is a routine approximation in (for example) detection and attribution studies. We consider a single forcing, intended as a generic indication of the large-scale anomaly that may be seen in response to a radiative perturbation. As our estimate of the pattern of transient response to external forcing, we use outputs from the CMIP3 models, taking the difference between the ensemble means of 2070–2090 and 2000–2020 in the A1B simulations. This pattern, which shows the expected large-scale spatially coherent warming with land-sea contrast and polar amplification and which is very similar to those shown in Fig. 10.8 of Meehl et al. (2007), is then scaled to generate the desired forced hemispheric mean temperature anomaly, and added onto the existing unforced runs. We add a sinusoidal “forced”

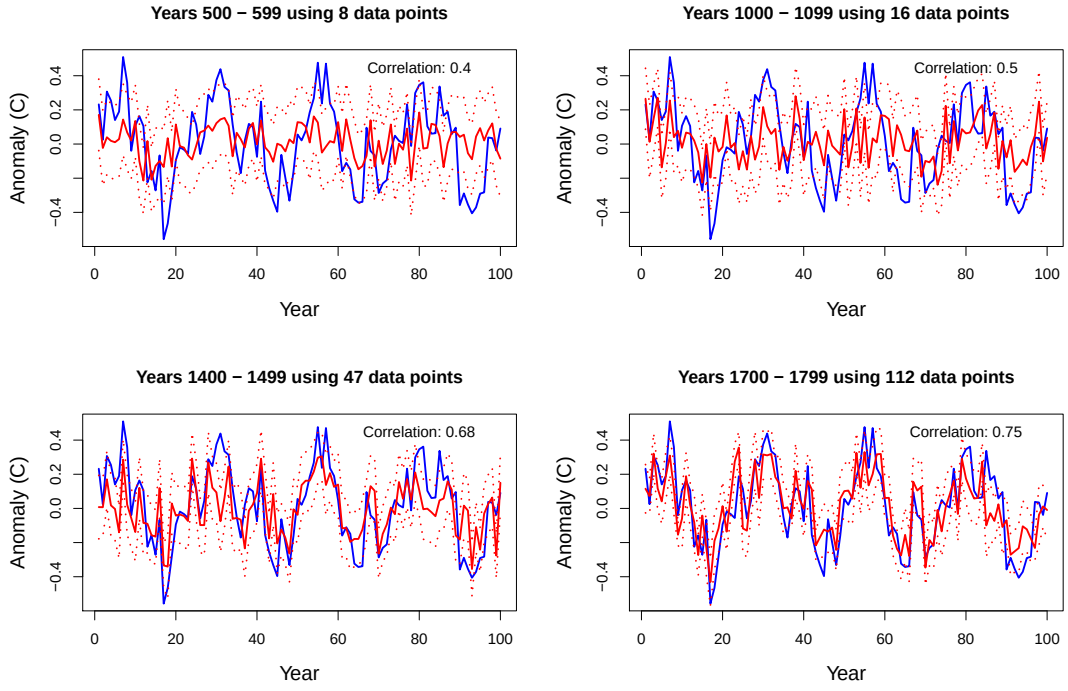

Fig. 5. Reconstruction of NH annual mean temperature anomaly for the case of externally forced changes, based on different number of proxy data points. Blue lines indicate target (identical in each panel). Solid red line indicates reconstruction, with dotted lines showing the ±1 standard deviation uncertainty. Correlations are averages over multiple experiments for each epoch.

signal onto our truth run, with a magnitude of ±0.2C (reasonably representative of the variation seen in temperature reconstructions) and a period of 25 yr. Observational errors are adjusted according to the total variance at each grid point, to maintain our chosen signal-to-noise ratio of 0.4. We also add a forced response onto each year of the ensemble, with random magnitude of the same variance as that of the signal added to the truth run.

Even though the amplitude of the forced signal was set equal for all members, its time-varying nature means that only a small proportion of the ensemble has forcing anomaly equal to that of any given year in the truth run. Interestingly, despite this (apparent) extra degree of freedom in the ensemble, the posterior ensemble size is actually fractionally larger than it was in the unforced case. This may be due to the large-scale forced response dominating some of the small-scale variability to the extent that the effective dimensionality of the problem is if anything decreased a little (Bengtsson et al., 2008). The results are shown in Figs. 5 and 6. In Fig. 5, the NH mean temperature appears to closely track the target run in the lower two panels, and even shows some hints of this in the upper two panels where much less skill was shown in the internal variability case (cf. Fig. 3). These results are reflected both in the reduction of RMS error (which ranges from 8 to 34 % across the experiments) and the correlation of reconstruction with target, which varies from 0.40 to an impressive level of 0.75. Thus, it appears that it is rather easier to identify the large-scale variation associated with external forcing than the smaller-scale variability. It is noticeable in Fig. 6 that, in all cases, the method has considerable skill in the tropical region. This may be partly an artefact of the particular climate model we are using, which has unrealistically low variability in this region. Therefore, despite the polar amplification of the forced response, it still easily dominates natural variability in the tropical region, and any skill in reproducing the forced response (which can be seen in Fig. 5) is strongly reflected in this region. Climate models with more realistic ENSO variability would probably not generate such strong results in this area. Smerdon et al. (2011) show how statistical methods, when applied to pseudoproxy experiments based on different climate models, can generate a wide range of spatially distinct patterns. Contrary to the internal variability experiments, the posterior estimates here do not show much skill in the neighbourhood of the observations themselves, where the hemispheric temperature signal represents a relatively small proportion of the total variance.

4.4 Sensitivity to observational uncertainty

We now consider the sensitivity of these results to the proxy uncertainty. While a signal-to-noise ratio of around 0.4 is typical, our implicit assumption of a single proxy record in each grid cell may be pessimistic. The Mann et al. (2008) proxy network contains over 1200 proxy records, and although fewer than half of these passed screening tests, the

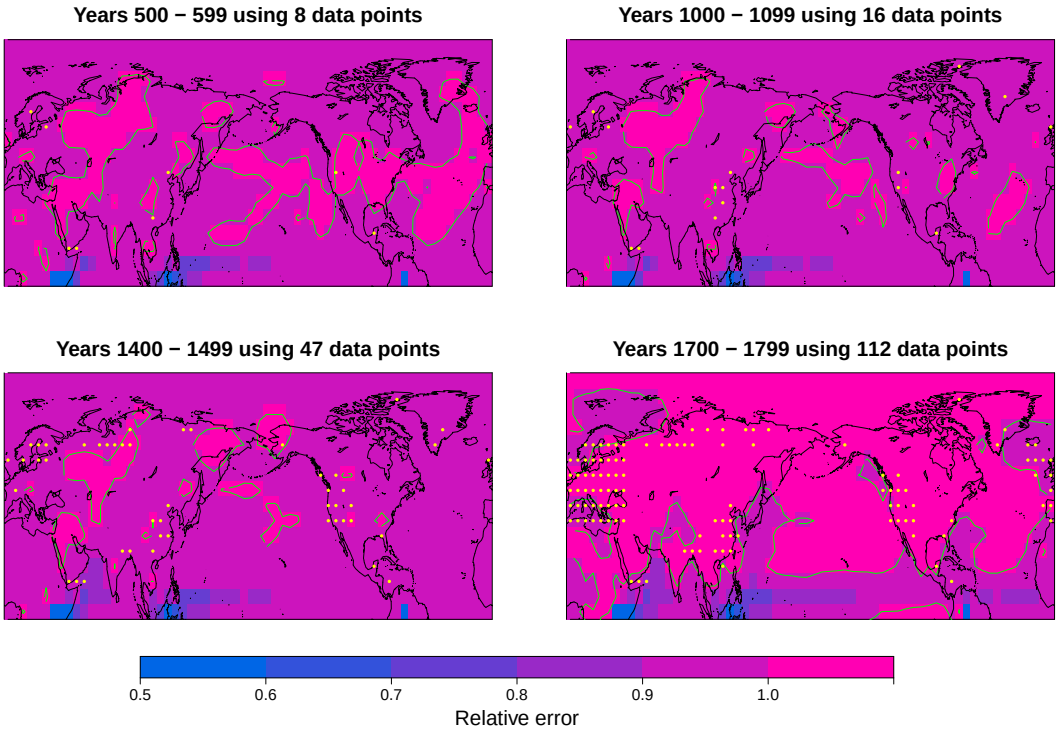

Fig. 6. Maps of normalised RMS reconstruction error for NH annual mean temperature anomalies for the case of externally forced changes, based on different number of proxy data points. Errors are normalised to standard deviation of climatology, with green contour indicating a normalised error of 1 (i.e. that the reconstruction is neither more nor less accurate on average than the prior climatological mean).

number of useful proxies is still rather larger than the number of grid cells that contain any records. We therefore consider a larger signal-to-noise ratio, to represent the situation where proxies with independent errors can be averaged to generate a more accurate signal. We use a signal-to-noise ratio of 1, which would be appropriate if each grid cell contained 4 independent proxy records, each of which had an SNR of about 0.5. We find that the reconstructions have greater skill at earlier times, but the collapse of the ensemble is also apparent at earlier times and more complete at later times, with an effective posterior ensemble size of around 2 samples in the most recent epoch. These results are qualitatively unsurprising, since greater proxy precision is largely equivalent to a larger number of proxies. The correlation of reconstruction and target is also generally higher in these experiments than for the standard case, saturating at around 0.8 at the point of ensemble collapse.

5 Conclusions

We have investigated the potential of particle-based data assimilation methods for the reconstruction of Northern Hemisphere temperatures over the past two millennia, in the context of a perfect model and well-characterised proxy uncertainty. We demonstrate that the method is successful and achieves significant skill as measured by the correlation between target and reconstruction. However, when considered in terms of residual RMS errors, the performance is less impressive. When few data points are available, the reconstruction is little changed from the prior, and RMS errors show negligible reduction. This is due simply to the sparse data providing very little information, and thus would not be changed by a different reconstruction method. For a higher data density, the reconstruction skill is higher, but the posterior ensemble has a tendency to collapse, even when a prior sample size of 10 000 is used. This is a limitation of particle-based methods in general. One interesting and encouraging attribute of our results is that this method can demonstrate skill in reconstructing the large-scale feature of hemispheric mean temperature even when most grid-point values have very little or even negative skill, as averaging out over noisy areas allows skilful regions to provide some hemispheric signal. Our results do however imply that caution must be applied when interpreting regional features of reconstructions, since smaller-scale spatial patterns and regional features may be represented with substantially less skill than an assessment of large-scale performance could suggest. We suggest that the performance of methods that are used for regional and large-scale reconstructions should be tested and demonstrated, not only in terms of correlation but also residual RMS error, in order to give a clear picture of their strengths and limitations.

The particle-based approach suffers from the requirement of a large sample size to work well, although this problem is far less severe for paleoclimate applications than has been suggested for modern numerical weather prediction, due to the relative sparsity and imprecision of proxy data. Encouragingly, the method retains substantial skill (at least by some measures) even when it has technically failed due to ensemble collapse. Alternative data assimilation methods, such as the ensemble Kalman filter (Bhend et al., 2012) or modified particle filtering approaches (van Leeuwen, 2010), could offer a way forward in respect of this problem, but it must be acknowledged that the nature of the sparse and imprecise data places fundamental limitations on our ability to reconstruct past climatic states.

Acknowledgements. We are grateful to Oliver Timm and Hugues Goosse for help and advice with LOVECLIM, Michael Mann for the data, and all three for interesting discussions.

Edited by: V. Rath

References

Arulampalam, S., Maskell, S., Gordon, N., and Clapp, T.: A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking, IEEE Trans. Signal Proc., 50, 174–188, 2002.

Barnett, T. and Preisendorfer, R.: Multifield Analog Prediction of Short-Term Climate Fluctuations Using a Climate State vector, J. Atmos. Sci., 35, 1771–1787, 1978.

Bengtsson, T., Bickel, P., and Li, B.: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems, in: Probability and Statistics: Essays in Honor of David A. Freedman, 2, 316–334, Ins. Math. Stat., 2008.

Bhend, J., Franke, J., Folini, D., Wild, M., and Brönnimann, S.: An ensemble-based approach to climate reconstructions, Clim. Past, 8, 963–976, doi:10.5194/cp-8-963-2012, 2012.

Boer, G. and Lambert, S.: Multi-model decadal potential predictability of precipitation and temperature, Geophys. Res. Lett., 35, L05706, doi:10.1029/2008GL033234, 2008.

Christiansen, B. and Ljungqvist, F.: Reconstruction of the extra-tropical NH mean temperature over the last millennium with a method that preserves low-frequency variability, J. Climate, 24, 6013–6034, doi:10.1175/2011JCLI4145.1, 2011.

Christiansen, B., Schmith, T., and Thejll, P.: A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness, J. Climate, 22, 951–976, 2009.

Christiansen, B., Schmith, T., and Thejll, P.: Reply, J. Climate, 23, 2389–2844, 2010.

Dubinkina, S., Goosse, H., Sallaz-Damaz, Y., Crespin, E., and Crucifix, M.: Testing a particle filter to reconstruct climate changes over the past centuries, Int. J. Bifurc. Chaos, 21, 3611–3618, 2011.

Dunstone, N. and Smith, D.: Impact of atmosphere and sub-surface ocean data on decadal climate prediction, Geophys. Res. Lett, 37, L02709, doi:10.1029/2009GL041609, 2010.

Goosse, H., Selten, F., Haarsma, R., and Opsteegh, J.: Decadal variability in high northern latitudes as simulated by an intermediate-complexity climate model, Ann. Glaciol., 33, 525–532, 2001.

Goosse, H., Renssen, H., Timmermann, A., Bradley, R., and Mann, M.: Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium, Clim. Dynam., 27, 165–184, 2006.

Goosse, H., Crespin, E., de Montety, A., Mann, M., Renssen, H., and Timmermann, A.: Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation, J. Geophys. Res., 115, D09108, doi:10.1029/2009JD012737, 2010.

Hansen, J. and Lebedeff, S.: Global trends of measured surface air temperature, J. Geophys. Res., 92, 13345–13372, 1987.

Jansen, E., Overpeck, J., Briffa, K., Duplessy, J.-C., Joos, F., Masson-Delmotte, V., Olago, D., Otto-Bliesner, B., Peltier, W., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R., and Zhang, D.: Palaeoclimate, in: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 6, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007.

Mann, M., Zhang, Z., Hughes, M., Bradley, R., Miller, S., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, Proc. Natl. Acad. Sci., 105, 13252, doi:10.1073/pnas.0805721105, 2008.

Meehl, G. A., Stocker, T., Collins, W., Friedlingstein, P., Gaye, A., Gregory, J., Kitoh, A., Knutti, R., Murphy, J., Noda, A., Raper, S., Watterson, I., Weaver, A., and Zhao, Z.-C.: Global Climate Projections, in: Climate Change 2007: The physical science basis, Contribution of the Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 10, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2007.

Renssen, H., Goosse, H., Fichefet, T., Brovkin, V., Driesschaert, E., and Wolk, F.: Simulating the Holocene climate evolution at northern high latitudes using a coupled atmosphere-sea ice-ocean-vegetation model, Clim. Dynam., 24, 23–43, 2005.

Rutherford, S., Mann, M., Ammann, C., and Wahl, E.: Comments on "surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness", J. Climate, 23, 2832–2838, 2010.

Selten, F., Haarsma, R., and Opsteegh, J.: On the mechanism of North Atlantic decadal variability, J. Climate, 12, 1956–1973, 1999.

Smerdon, J.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments, Wiley Interdisciplinary Reviews (WIREs), Clim. Change, 3, 63–77, doi:10.1002/wcc.149, 2012.

Smerdon, J., Kaplan, A., and Chang, D.: On the Origin of the Standardization Sensitivity in RegEM Climate Field Reconstructions, J. Climate, 21, 6710–6723, 2008.

Smerdon, J., Kaplan, A., Zorita, E., González-Rouco, J., and Evans, M.: Spatial performance of four climate field reconstruction methods targeting the Common Era, Geophys. Res. Lett., 38, L11705, doi:10.1029/2011GL047372, 2011.

Snyder, C., Bengtsson, T., Bickel, P., and Anderson, J.: Obstacles to high-dimensional particle filtering, Mon. Weather Rev., 136, 4629–4640, 2008.

Timmermann, A., An, S., Krebs, U., and Goosse, H.: ENSO Suppression due to Weakening of the North Atlantic Thermohaline Circulation, J. Climate, 18, 3122–3139, 2005.

Tingley, M., Craigmile, P., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: Statistical insights into paleoclimatic reconstructions, Quat. Sci. Rev., 35, 1–22, doi:10.1016/j.quascirev.2012.01.012, 2012.

van Leeuwen, P.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Q. J. Roy. Meteorol. Soc., 136, 1991–1999, 2010.

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