OPEN ACCESS

University of BRISTOL

Edwards, T. L., Annan, J., Crucifix, M., Gebbie, G., & Paul, A. (2013). Best-of-both-worlds estimates for time slices in the past. PAGES news, 21(2), 76-77.

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to publication record in Explore Bristol Research

PDF-document

University of Bristol - Explore Bristol Research

General rights

Best-of-both-worlds estimates for time slices in the past

Tamsin L. Edwards[1], J. Annan[2], M. Crucifix[3], G. Gebbie[4] and A. Paul[5]

[1]Department of Geographical Sciences, University of Bristol, UK; tamsin.edwards@bristol.ac.uk
[2]Research Institute for Global Change, JAMSTEC, Yokohama Institute for Earth Sciences, Japan; [3]Earth and Life Institute and Georges Lemaître Centre for Earth and Climate Research, Université catholique de Louvain, Belgium; [4]Department of Physical Oceanography, Woods Hole Oceanographic Institution, USA; [5]MARUM - Center for Marine Environmental Sciences and Department of Geosciences, University of Bremen, Germany

We introduce data assimilation methods for estimating past equilibrium states of the climate and environment. The approach combines paleodata with physically-based models to exploit their strengths, giving physically consistent reconstructions with robust, and in many cases, reduced uncertainty estimates.

Those seeking to understand the Earth's past usually take one of two approaches: reconstructing paleoclimate and paleo-environmental states from proxy data derived from natural archives such as ice cores and trees; or simulating them with earth system models that contain theoretical knowledge of physical processes.

Proxy-based reconstructions are based on observations of the real world, but most consider data points independently rather than accounting for correlations in space, time and between climate variables. Therefore they risk being physically inconsistent. Models incorporate aspects of physical consistency, but are imperfect and are tested during development only with present day observations.

Data assimilation produces "best-of-both-worlds" estimates that combine observational and theoretical information while not ignoring their limitations. We discuss data assimilation for estimating past equilibrium states of the earth system such as climate and vegetation. We use the term "paleodata" for measurement-based data: either the observations of proxies or the statistical reconstructions derived from them.

Aims and methods

Equilibrium state, or "time slice", data assimilation is estimation of a snapshot in time during which it is assumed the state variables are not changing. The state estimates may be the primary scientific aim or simply a "bonus" of calibrating model parameters (Annan et al., this issue).

Time slice estimation is a natural starting point in data assimilation because it is more straightforward than estimating a transient state (Brönnimann et al., this issue) and is particularly appropriate if spatial patterns are more important than temporal changes or if the model is computationally expensive. For a given computational resource time slice estimation permits more complete exploration of model uncertainties in parameters, structure, and inputs. Another advantage of a focus on time slices is that for eras studied by the Paleoclimate Model Intercomparison Project (PMIP) relatively large quantities of paleodata and simulations are available. Most data assimilation estimates of equilibrium paleo-states are therefore of the Last Glacial Maximum (LGM: 21 ka cal BP), the most recent era for which annual mean climate is substantially different to the present that also has a long history of study by PMIP.

We use model simulations in paleo-state estimation because models provide links across different locations, times (relevant to transient or multi-state estimation) and state variables. This has two advantages: it helps ensure the resulting state is physically consistent, and it also means we are not limited to assimilating the same variables we wish to estimate. We could assimilate data in one place to estimate another, or assimilate temperature data to estimate precipitation, or assimilate variables corresponding to the outputs of a model to estimate variables corresponding to the inputs. The last are termed "inversion" methods, such as estimating atmospheric variables or terrestrial carbon from paleo-vegetation records (e.g. Guiot et al. 2000; Wu et al. 2007; Wu et al. 2009; Pound et al. 2011) or estimating oceanic variables from paleo-tracer records (e.g. LeGrand and Wunsch 1995; Roche et al. 2004).

Data assimilation requires the following ingredients: paleodata with uncertainty estimates, simulations with uncertainty estimates, and a metric to quantify the dissimilarity, or "distance", between the two. Climate state estimates are obtained by searching for the simulation(s) closest to the paleodata ("optimization") or calculating a weighted combination of the two ("updating").

Distance is usually measured with the standard metric for normally distributed model-data differences, i.e. the sum of squared differences weighted by the uncertainties, though some use ad-hoc or fuzzy metrics (e.g. Guiot et al. 2000; Wu et al. 2007; Gregoire et al. 2010). For non-continuous variables, for example with a threshold, variables must be transformed or a non-Gaussian metric chosen (e.g. Stone et al. 2013).

Optimisation methods search for the simulation with the minimum distance from paleodata. One approach uses numerical differentiation of the model with respect to the parameters, essentially least-squares fitting of a line or curve to one-dimensional data (e.g. LeGrand and Wunsch 1995; Gebbie and Huybers 2006; Marchal and Curry 2008; Burke et al. 2011; Huybers et al. 2007; Paul and Losch 2012). Another approach generates an ensemble of simulations using many different parameter values and then selects the members with the smallest model-data distance ("perturbed parameter ensemble" methods; e.g. Gregoire et al. 2010).

Updating methods combine model and paleodata estimates. Typically the model estimates are generated with a perturbed parameter ensemble, which permits well-defined sampling of parameter uncertainties; the model estimates are reweighted with the model-data distance using Bayesian updating (e.g. Guiot et al. 2000; Wu et al. 2007; Wu et al. 2009; Holden et al. 2009; Schmittner et al. 2011).

Interpretation

Figure 1 illustrates some strengths of data assimilation. The model propagates information from LGM surface air temperature (SAT) reconstructions over land to other regions, and to sea surface temperatures (SST). In this example assimilating SAT reconstructions produces an SST estimate with a warming at the LGM in the northern North Atlantic, which is consistent with the SST reconstructions. Uncertainties are reduced relative to the model estimate
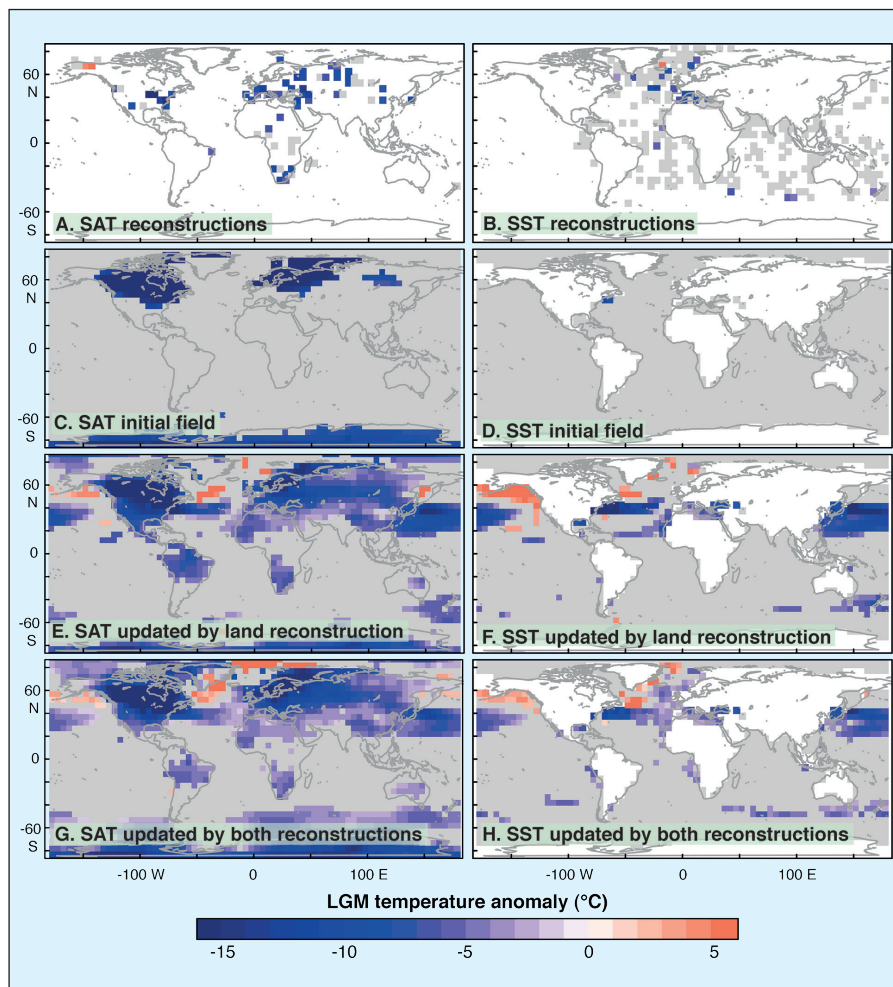
*Figure 1: LGM annual mean temperature anomalies from: **A)** surface air temperature (SAT) reconstructions based on pollen and plant macrofossils (Bartlein et al. 2010); **B)** sea surface temperature (SST) reconstructions based on multiple ocean proxies (MARGO et al. 2009); **C, D)** simulations from the HadCM3 general circulation model (mean of 17 member perturbed parameter ensemble; Edwards, unpublished data). Data assimilation estimates generated by updating with SAT reconstructions **(E, F)**, and both SAT and SST reconstructions **(G, H)**. Gray areas indicate regions with low signal-to-noise: magnitude of temperature anomaly is less than 3σ of uncertainty estimates.*

allow greater freedom in specifying the behavior of climate-proxy relationships (such as nonlinearity and multi-modal uncertainties) and enables uncertainties to cascade through the causal chain to allow full probabilistic quantification of the unknown state. Using physically-based forward models for reconstruction, i.e. data assimilation, incorporates information about the relationships between locations, times and variables and therefore minimizes the risk of physical implausibility. The long-term goal may be forward physical modeling of the whole causal chain from radiative forcings to proxy archives (e.g. Roche et al. 2004; Stone et al. 2013).

For paleo-simulations, we do not need models to be complex or state-of-the-art, but we do need to estimate their uncertainties. If they are complex it is difficult to generate their derivatives with respect to the parameters. If they are expensive it is difficult to sample, and therefore to assess, their uncertainties. Thoughtful experimental design with statisticians, and perhaps also statistical modeling of the physical model (known as "emulation"; e.g. Schmittner et al. 2011), can help in this regard. A research priority is to estimate the discrepancy between a model and reality at its best parameter values, and how this varies across different eras. New updating methods are emerging that use the PMIP multi-model ensemble to explore structural uncertainties. For example, Annan and Hargreaves (2013) use the linear combination of ensemble members that best matches the paleodata.

These challenges are worth tackling for the substantial benefits. Information from paleodata can be extrapolated to other locations, times and state variables, and uncertainties are smaller (or at worst, the same) than those of the individual model or proxy-based estimates.

in most locations (grayed out areas are reduced).

How should we interpret assimilated paleo-states? Optimization methods select a single best simulation so the state estimate is physically self-consistent according to the model. But the state estimate from updating methods is a combination of multiple model simulations and paleodata, therefore interpretation requires more care. An ensemble mean anomaly of zero might correspond to a wide spread of positive and negative results; this would be reflected in large model uncertainties. A spatially coherent signal with small uncertainty might emerge from an ensemble after assimilating a single "pinning point" from paleodata; this signal should be physically consistent because it arises from the model physics. Such considerations are common to all multi-model ensemble summaries and reanalyses.

For statistically meaningful results it is essential to use a distance metric grounded in probability theory, i.e.

corresponding to a particular distribution of model-data differences ("likelihood function" in Bayesian terms). This might preclude the use of non-standard variables such as biomes.

Data assimilation is a statistical modeling technique and should be evaluated. Testing the method with pseudo-paleodata can help avoid the (literal) pitfalls of finding local rather than global minima in high-dimensional spaces.

Future directions

Data assimilation is a formal method that not only highlights model-data discrepancies but also corrects them. It can be challenging, because it requires a process-based model and reliable estimation of uncertainties for both paleodata and simulations.

For paleodata, difficulties may arise from dating and time averaging. But improvements in estimating reconstruction uncertainties can be made by using forward modeling approaches (e.g. Tingley et al. 2012). These approaches

Selected references

Full reference list online under:
http://www.pages-igbp.org/products/newsletters/ref2013_2.pdf

Bartlein P et al. (2010) *Climate Dynamics* 37(3-4): 775-802
Guiot J et al. (2000) *Ecological Modelling* 127(2-3): 119-140
Schmittner A et al. (2011) *Science* 334: 1385-1388
MARGO et al. (2009) *Nature Geoscience* 2: 127-132
Stone EJ, Lunt DJ, Annan JD, Hargreaves JC (2013) *Climate of the Past* 9: 621-639