# Parameter estimation using paleodata assimilation

James D. Annan[1], M. Crucifix[2], T.L. Edwards[3] and A. Paul[4]

[1]RIGC/JAMSTEC, Yokohama, Japan; jdannan@jamstec.go.jp
[2]Georges Lemaitre Centre for Earth and Climate Science, Université Catholique de Louvain, Belgium; [3]School of Geographical Sciences, Bristol University, UK; [4]MARUM, University of Bremen, Germany

**In addition to improving the simulations of climate states, data assimilation concepts can also be used to estimate the internal parameters of climate models. Here we introduce some of the ideas behind this approach, and discuss some applications in the paleoclimate domain.**

Estimation of model parameter values is of particular interest in paleoclimate and climate change research, since it is the formulation of model parameterizations, rather than the initial conditions, which is the main source of uncertainty regarding the climate's long-term response to natural and anthropogenic forcings.

We should recognize at the outset that the question of a "correct" parameter value might in many cases be quite contentious and disputable. There is, for example, no single value to describe the speed at which ice crystals fall through the atmosphere, or the background rate of mixing in the ocean, to mention two parameters which are commonly varied in General Circulation Models (GCMs). Generally the best we can hope for is to find a set of parameter values, which perform well in a range of circumstances, and to make allowances for the model's inadequacies, i.e. structural errors due to inadequate equations and parameterizations. However, inadequacies will always be present no matter how carefully parameter values are chosen: this should serve as a caution against over-tuning.

It may not be immediately clear how one can use proxy-derived observational estimates of climatic state variables such as temperature or precipitation to estimate the values of a model's internal parameters. However, from a sufficiently abstract perspective, the problem of parameter estimation can be considered as equivalent to state estimation, via a standard approach in which the state space of a dynamical model is augmented by the inclusion of model parameters (Jazwinski 1970; Evensen et al. 1998). To see how this works, consider a system described by a dynamical model f, which uses a set of internal parameters θ and propagates a state vector x through time through a set of differential equations:

$$x = f_\theta(x) \qquad (1)$$

We can create an equivalent model g(x, θ) which takes as its state vector (x, θ) (in which the parameter values have simply been concatenated onto the end of the state vector), and propagates this vector through the augmented set of equations

$$\dot{x} = f_\theta(x) \qquad (2)$$

$$\dot{\theta} = 0 \qquad (3)$$

Thus, the existing methods and technology for estimating the state x can, in principle, be directly applied to the estimation of (x, θ), or in other words, the joint estimation of state and parameters.

While this approach is conceptually straightforward, there are many practical difficulties in its application. The most widespread methods for data assimilation, including both Kalman filtering and 4D-VAR, rely on (quasi-)linear and Gaussian approaches. However, the augmented model g is likely to be substantially more nonlinear in its inputs than the underlying model f , due to the presence of product terms such as $\theta_i\, x_j$ (Evensen et al. 1998).

Further challenges exist in applying this approach due to the wide disparity in relevant time scales. Often the initial state has a rapid effect on the model trajectory within the predictability time scale of the model, which is typically days to weeks for atmospheric GCMs. On the other hand, the full effect of the parameters only becomes apparent on the climatological time scale, which may be decades or centuries.

## Applications

Methods for joint parameter and state estimation in the full spatiotemporal domain continue to be investigated for numerical weather prediction, where data are relatively plentiful. But identifiability, that is the ability to uniquely determine the state and parameters given the observations, is a much larger problem for modeling past climates, where proxy data are relatively sparse in both space and time.

Therefore, data assimilation in paleoclimate research generally finds a way to reduce the dimension of the problem. One such approach is to reduce the spatial dimension, even to the limit of a global average. For example, a three-variable globally averaged conceptual model for glacial cycles has been tuned using flexible and powerful methods such as Markov Chain Monte Carlo (Hargreaves and Annan 2002) and Particle Filtering (Crucifix and Rougier 2009). Figure 1 presents the results of one parameter estimation experiment by Hargreaves and Annan (2002).

In the case of more complex and higher resolution models, the problems of identifiability and computational cost are most commonly addressed by the use of equilibrium states. Here, the full initial condition of the model is irrelevant, at least within reasonable bounds, and the dimension of the problem collapses down to the number of free parameters; typically ten at most, assuming many boundary conditions are not also to be estimated. With this approach, much of the detailed methodology of data assimilation as developed and practiced in numerical weather prediction, where the huge state dimension is a dominant factor, ceases to be so relevant.

While some attempts at using standard data assimilation methods have been performed (e.g. Annan et al. 2005), a much broader range of estimation methods can also be used. With reasonably cheap models and a sufficiently small set of parameters, direct sampling of the parameter space with a large ensemble may be feasible. A statistical emulator, which provides a very fast approximation to running the full model, may help in more computationally demanding cases (e.g. Holden et al. 2010).

One major target of parameter estimation in this field has been the estimation of the equilibrium climate sensitivity. This may either be an explicitly tunable model parameter in the case of simpler models, or else an emergent property of the underlying physical processes, which are parameterized in a more complex global climate model. The Last Glacial Maximum is a particularly popular interval for study, due to its combination of a large signal to noise ratio and good data coverage over a quasi-equilibrium interval (Annan et al. 2005; Schneider von Deimling et al. 2006; Holden et al. 2010; Schmittner et al. 2011; Paul and Losch

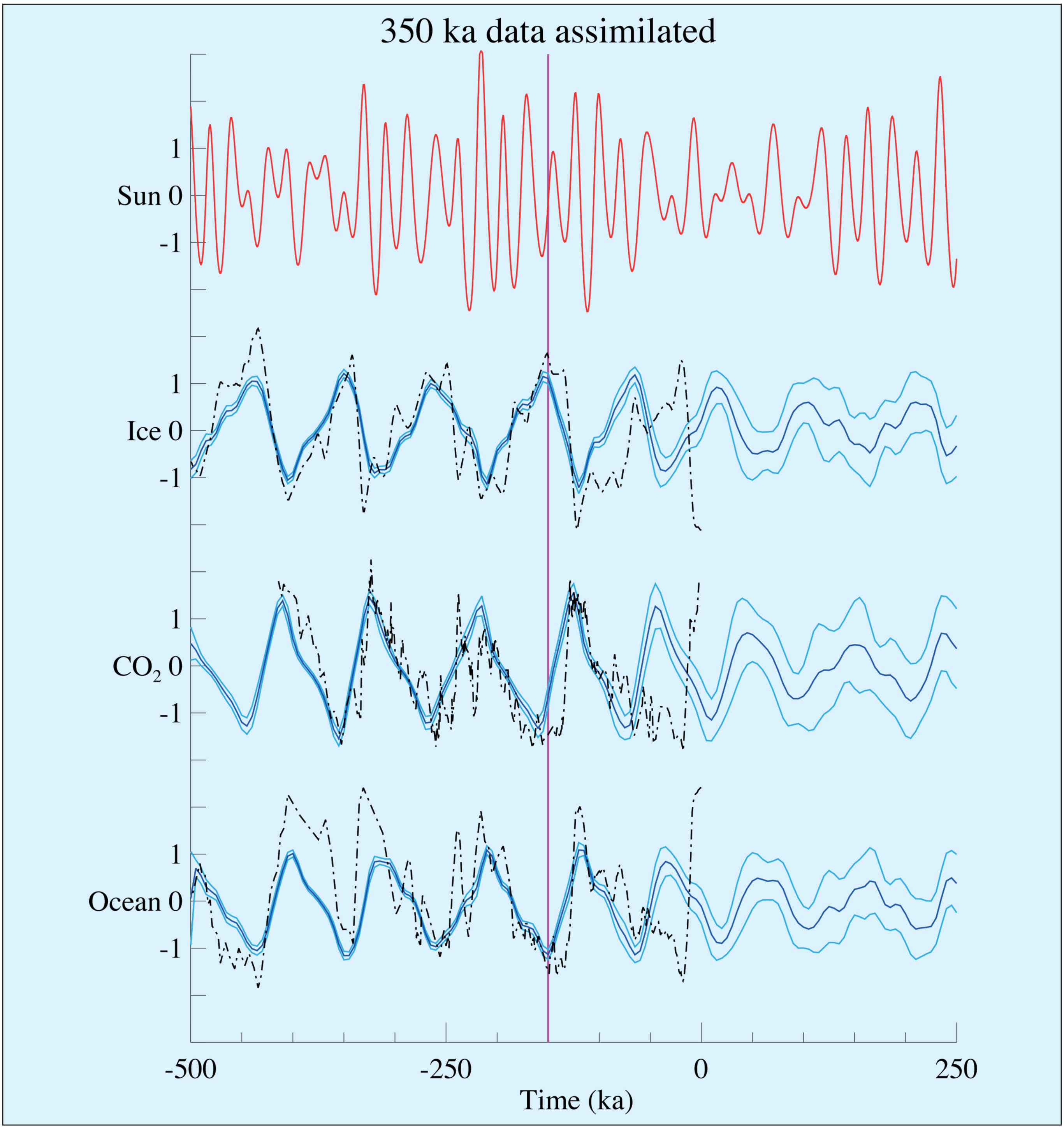


*Figure 1: Experiment with 350 ka of data assimilated. The red line at the top is the normalized summer solar insolation forcing at 65° N. The black dot-dashed lines are normalized proxy data from Vostok (ice volume and atmospheric $CO_2$ concentration) and SPECMAP (deep ocean temperature) cores. Data to the left of the vertical magenta line were used to tune parameters, with the right hand side used as validation of the model forecast, which (over a range of experiments) shows substantial skill for a duration of around 50-100 ka. The dark blue lines show the mean of the ensemble and the light blue lines show one standard deviation of the ensemble. Modified from Hargreaves and Annan (2002).*

2012). The methods used for studying the LGM in order to estimate the equilibrium climate sensitivity have covered a wide range of techniques including direct sampling of parameter spaces (with and without the use of an emulator), Markov Chain Monte Carlo methods, the variational approach using an adjoint model, and the Ensemble Kalman Filter. In general, more costly models require stronger assumptions and approximations due to computational limitations.

Approaches which aim at averaging out the highest frequencies of internal variability while still retaining a transient and time-varying forced response may make use of temporal data such as tree rings over the last few centuries (Hegerl et al. 2006). In that case, the spatial dimension can still be reduced, e.g. by averaging to a hemispheric mean. A similar approach was used by Frank et al. (2010) to estimate the carbon cycle feedback.

Paleoclimate simulations provide the only opportunity to test and critically evaluate climate models under a wide range of boundary conditions. This suggests that we need to continue to develop a broad spectrum of methods to be applied on a case-specific basis.

## Selected references

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

Annan JD et al. (2005) *SOLA* 1: 181-184

Crucifix M, Rougier J (2009) *The European Physical Journal-Special Topics* 174(1): 11-31

Hargreaves JC, Annan JD (2002) *Climate Dynamics* 19: 371-381

Holden PB et al. (2010) *Climate Dynamics* 35: 785-806

Paul A, Losch M (2012) *Perspectives of parameter and state estimation in paleoclimatology*: In, Berger A et al. (Eds) Climate Change: Inferences from Paleoclimate and Regional Aspects, Springer, 93-105