Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter.

J.D. Annan [a,*], J.C. Hargreaves [a], N. R. Edwards [b], R. Marsh [c]

[a] *Frontier Research System for Global Change*
3173-25 Showa-machi, Kanazawa-ku
Yokohama, Kanagawa 236-0001, Japan
and
Proudman Oceanographic Laboratory, Joseph Proudman Building
6 Brownlow Street, Liverpool, L3 5DA, UK

[b] *Climate and Environmental Physics, Physics Institute,*
University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

[c] *James Rennell Division, Southampton Oceanography Centre,*
Empress Dock, Southampton SO14 3ZH, UK

Abstract

We describe the development of an efficient method for parameter estimation and ensemble forecasting in climate modelling. The technique is based on the ensemble Kalman filter and is several orders of magnitude more efficient than many others which have been previously used to address this problem. As well as being theoretically (near-)optimal, the method does not suffer from the 'curse of dimensionality' and can comfortably handle multivariate parameter estimation. We demonstrate the potential of this method in identical twin testing with an intermediate complexity coupled AOGCM. The model's climatology is successfully tuned via the simultaneous estimation of 12 parameters. Several minor modifications are described by which the method was adapted to a steady state (temporally averaged) case. The method is relatively simple to implement, and with only O(50) model runs required, we believe that optimal parameter estimation is now accessible even to computationally demanding models.

Key words: data assimilation, numerical modelling, climate science

1 Introduction

All climate models, from the simplest to the most complex, are a mass of parameterisations, and climate models are widely used for predictions of future climate change (Hansen et al., 2002; Johns et al., 2003). In contrast to short-term operational weather prediction, climate forecasts depend strongly on parameterisations rather than initial conditions. Therefore, if we are to have confidence in these predictions, it is important that the parameter values should be tuned to the correct values. Moreover, given that no deterministic prediction will ever be exactly correct, it is also important to quantify the uncertainty associated with a forecast, which itself depends on the certainty with which parameter values can be determined. For these reasons, the problem of parameter estimation in climate modelling has recently attracted a great deal of attention (eg Forest et al., 2000; Andronova and Schlesinger, 2001; Knutti et al., 2002; Gregory et al., 2002). It is notable that these studies use substantially simpler models than the current state of the art models. The fundamental reason for this is that existing objective methods for parameter estimation are often computationally demanding, typically requiring thousands or even millions of model runs depending on the details of the particular implementation. At present, we have both forecasts from simple models with their (near-)optimally tuned ensemble predictions, and state of the art models each with a handful of less precisely tuned runs and limited uncertainty estimation. Ideally, one would like to see optimally tuned ensemble predictions using state of the art climate models, but the cost until now has been prohibitive.

For these reasons, an efficient ensemble-based method which can fill this gap should be of great interest to the climate modelling community. In this paper, we describe such a method. It is based on the ensemble Kalman filter (EnKF) (Evensen, 1994). It is tested on a new Earth System Model of Intermediate Complexity (EMIC) (Edwards and Marsh, 2003) which is being developed as part of the UK GENIE project (http://www.genie.ac.uk/), and the method is found to be capable of simultaneously tuning 12 independent parameters, with a very modest ensemble size of 54 members.

In section 2, we review the problem of Bayesian parameter estimation from a theoretical perspective in the context of recent practical applications in climate prediction. Section 3 contains a brief outline of the model, and the assimilation method is described in section 4. Validation of the method via identical twin testing is described in section 5, and some conclusions are presented in section 6.

* Corresponding author.
Email address: jdannan@jamstec.go.jp (J.D. Annan).

2 Estimation theory and practice

Parameter estimation using models has been studied for some time, with much of the development taking place in the field of hydrology (Chavent, 1974). Our intention here is to briefly introduce the relevant theory, and discuss it with particular reference to applications in climate prediction.

Bayes' Theorem provides the principle whereby a prior probability distribution function (pdf), for example an estimate based on theoretical arguments, can be updated with new data (observations) to form a posterior estimate (for example Wilks, 1995, Eqn 2.15):

$$P(\mathbf{\Phi}|\mathbf{X}) = \frac{P(\mathbf{\Phi})P(\mathbf{X}|\mathbf{\Phi})}{P(\mathbf{X})} \tag{1}$$

where $P(\mathbf{\Phi}|\mathbf{X})$ is the posterior pdf of the parameters $\mathbf{\Phi}$ given the observations $\mathbf{X}$. $P(\mathbf{X}|\mathbf{\Phi})$ is the pdf of the observations for a given set of parameters, also known as the likelihood when considered as a function of parameters given fixed data, $P(\mathbf{\Phi})$ is the prior pdf of the parameters and $P(\mathbf{X})$ is the pdf of the observations unconditional on parameter values (ie $P(\mathbf{X}) = \int P(\mathbf{X}|\mathbf{\Phi})P(\mathbf{\Phi})$). This equation underpins almost all modern geophysical data assimilation theory and practice. In order to calculate the pdf of the observations conditional on the parameters, we require a model that transforms a parameter set into state variables that can be compared to the data. The existence of such a model is implicitly assumed from here on, and moreover an underlying assumption of the theory is that the model is perfect in the sense that there exists a set of parameters (and initial conditions, where relevant) for which the model output matches the real world. This 'strong constraint' assumption is rarely, if ever, true in practice but nevertheless this assumption is common in estimation and data assimilation.

When we have independent observation errors, the joint pdf of the data is the product of their independent probability density functions and in the widely-assumed Gaussian case this is given by:

$$P(\mathbf{X}|\mathbf{\Phi}) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\frac{(x_i - x_i')^2}{\sigma_i^2}} = K e^{-\frac{1}{2}\sum_i \frac{(x_i - x_i')^2}{\sigma_i^2}} \tag{2}$$

where the product and sum are over the observations x_i which have observational errors σ_i, and x_i' are the corresponding model state variables which are generated by running the model with parameter set Φ. $P(\mathbf{X}|\mathbf{\Phi})$ is then the probability of observing the values x_i in a world in which the parameters take the particular values Φ. K is a normalising constant that generally need not be calculated. It is

frequently more convenient to work with the logarithm of this function, given by

$$\ln(P(\mathbf{X}|\mathbf{\Phi})) = k + -\frac{1}{2}\sum_i \frac{(x_i - x_i')^2}{\sigma_i^2}. \quad (3)$$

The prior probability for the parameters can be written similarly:

$$\ln(P(\mathbf{\Phi})) = c + -\frac{1}{2}\sum_j \frac{(p_j - p_j')^2}{\tau_j^2}. \quad (4)$$

This is the logarithm of the probability of the parameter set p_j' where the sum is over the parameters p_j', and p_j and τ_j are the mean and standard deviations of the prior estimates. The negative of the summation terms on the right hand sides of equations (3) and (4), added together, is often called the cost function.

The optimum solution for Φ is that which maximises its probability, or minimises the weighted sum of squares on the right hand side of equation (2). This is the peak of the posterior pdf, which coincides with the mean for linear models. Note the close relationship with weighted least squares and maximum likelihood estimates. An attractive feature of many of the Monte Carlo sampling methods used to solve equation (1) is that the normalising denominator $P(\mathbf{X})$, and the constant additive term in the cost function, generally need not be calculated.

One common Bayesian Monte Carlo approach, introduced by Dilks et al. (1992), is to generate a set of random samples from the joint prior pdf $P(\mathbf{\Phi})$, and assign a weight to each sample. The weights are given by the likelihood $P(\mathbf{X}|\mathbf{\Phi})$. Any statistics of the posterior can be readily evaluated from the appropriately weighted ensemble members. A minor variant of this is to make a randomised decision to accept or reject each sample, with probability of acceptance proportional to its likelihood. This is equivalent to sampling directly from the posterior pdf, so again any desired statistics of the posterior such as means and variances can be directly calculated from them. This variant has the advantage that fewer samples are retained and no weights are required, allowing potentially large economies of storage. Knutti et al. (2002) used a similar algorithm, accepting only simulations within 2 standard deviations of the observational estimates without further weighting, effectively assuming uniform pdfs. Efficiency can be improved somewhat by Latin hypercube sampling (McKay et al., 1979), in which prior probability space is divided into disjoint regions of equal probability. Edwards and Marsh (2003) used a similar procedure which is described further in section 5.1.

The principal drawback to these methods is that the number of simulations required is very large and increases dramatically with the dimension of the parameter space. Knutti et al. (2002) used 25,000 model runs to estimate the posterior pdfs of 2 parameters, but in the example of Qian et al. (2003) where 3 param-

eters are simultaneously estimated, 10,000 samples from the prior only generated only one plausible sample from the posterior. For the climateprediction.net project (http://www.climateprediction.net), the starting point is a multifactorial exploration of a 21-parameter space with each parameter taking 3 distinct values, although substantial pruning of the 10^{10} runs implied will evidently be required, even with the help of the general public via the computational grid.

In the Monte Carlo Markov Chain method a random walk is performed in parameter space, guided by the results of the analysis at each step (eg Chib and Greenberg, 1995; Harmon and Challenor, 1997; Hargreaves and Annan, 2002). This typically reduces the computational cost from an exponential to a low-order polynomial function of dimensionality, but the computational requirements are still generally large. Moreover, the algorithm is inherently serial so, in contrast to the other methods outlined, is poorly suited to massively parallel computation.

One very different and much more efficient method for the estimation of parameter values involves the use of variational techniques via an adjoint model (Derber, 1989). However this approach has several major drawbacks. Creating the adjoint is a major undertaking, with significant theoretical and practical problems associated with non-differentiable points. Moreover, it is not simple to derive the associated uncertainties for a probabilistic forecast. A further problem can arise when this method is applied to climatological tuning of chaotic models. This is because the adjoint calculates the reponse of a particular finite model run to an infinitesimal perturbation which may be out of all proportion to the true climatological sensitivity (Lea et al., 2000, 2002).

2.1 Estimation using the Kalman filter

As we have already seen, the Bayesian estimation problem is essentially the same as least squares estimation, at least under the assumption that all distributions are Gaussian. Therefore, we can also directly estimate the posterior by methods based on least squares estimation. For example, a single prior estimate f which has uncertainty σ_f (in this context, consider f to be a variable of a model which depends on our prior parameter estimates), and an observation o of the appropriate physical property with uncertainty σ_o, can be combined into an optimal estimate or analysis value a via

$$a = f + \alpha(o - f) \tag{5}$$

where

$$\alpha = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_o^2}. \tag{6}$$

The uncertainty of the analysis, σ_a, can be shown to be given by

$$\sigma_a^2 = (1 - \alpha)\,\sigma_f^2 = \frac{\sigma_o^2 \sigma_f^2}{\sigma_f^2 + \sigma_o^2}. \tag{7}$$

Thus it can be seen that the posterior estimate is less uncertain than both the prior and the data.

The multivariate version of this equation is best written in matrix form. It often appears in the guise of sequential estimation, that is to say where the prior model state is a forecast generated from an earlier time, to be updated with more recent observations. If we have a vector of model forecast values x^f with their covariance matrix P^f, observations z with error covariance matrix R and a linear transformation H which maps the field of model variables to their observational equivalents, the optimal linear estimate x^a can be written as

$$x^a = x^f + K(z - Hx^f) \tag{8}$$

where K is the Kalman gain matrix given by

$$K = P^f H^T (H P^f H^T + R)^{-1}. \tag{9}$$

We can also calculate the analysis covariance matrix P^a via

$$P^a = (I - KH)P^f. \tag{10}$$

These equations are analogous to equations (5), (6) and (7).

The analysed variables and covariances can be integrated through time via a numerical model to form the forecast at a future time, and this complete system forms the theoretically optimal iterative system for sequential multivariate estimation known as the Kalman filter equations (Kalman, 1960). Note, however, that if the model has N state variables, then the forecast covariance matrix has size $N \times N$, and if we have M observations, the calculation of the Kalman gain requires the inversion of an $M \times M$ matrix. Therefore, the immense computational demands mean that many simplifications and approximations are almost invariably used in, for example, operational oceanography and numerical weather prediction.

Our system is based on the ensemble Kalman filter (EnKF), introduced by Evensen (1994) and now widely used for model reinitialisation (eg Evensen and van Leeuven, 1996; Houtekamer and Mitchell, 1998; Keppenne, 2000; Anderson, 2001). A thorough introduction and overview of applications is given in Evensen (2003). In this method, an ensemble of model runs is used as a Monte Carlo sample of the

state pdf. The sample statistics of the ensemble provide the forecast mean and covariance, that is

$$x^f = \frac{1}{M}\sum_{i=1}^{M} x_i^f \tag{11}$$

$$P^f = \frac{1}{M-1}\sum_{i=1}^{M}(x_i^f - x^f)(x_i^f - x^f)^T \tag{12}$$

where the summation is over the M ensemble members $\{x_i^f\}$, $i = 1, \ldots, M$, and the analysis mean and covariance are defined analogously. Applying the Kalman filter equations to each ensemble member automatically generates a new ensemble with the correct statistics for x^a and P^a (Evensen (1994), but also see Burgers et al. (1998)).

Some rearrangement of the Kalman equations means that the forecast covariance matrix is never explicitly required but only appears via the smaller $P^f H^T$ and $HP^f H^T$ matrices (size $N \times M$ and $M \times M$ respectively, where generally $M \ll N$). A further gain in efficiency is due to the localised analysis which will be described in more detail later. The EnKF system generally works well in geophysical fluid dynamics applications with 40–100 members, even though theory suggests that this number may be a little on the low side (van Leeuwen, 1999). The computational cost of the algorithm is generally dominated by the integration of this ensemble. Although based on the formally optimal Kalman filter equations given above, there are a number of approximations inherent in the method, perhaps the most obvious of which is the sampling error due to the finite ensemble size. Another is that the analysis equations rely on the assumption that the error statistics are Gaussian with vanishing higher-order statistical moments (but note that a major strength of the EnKF is that the covariance matrix is propagated in time via the fully nonlinear model equations, rather than a linearised approximation as in the extended Kalman filter). Further approximations are often introduced in the detail of the computational algorithm, some of which are described below. There are various alternative ensemble Kalman filtering algorithms which have been developed for large-scale applications (Houtekamer and Mitchell, 2001; Evensen, 2003; Keppenne and Rienecker, 2001) and at this stage no individual method appears to be clearly superior. We have adopted the algorithm implemented by Keppenne (2000).

As described above, the Kalman filter equations appear to be unrelated to the problem of parameter estimation. It is merely the model state which is estimated. This is the form in which the Kalman filter has been used quasi-operationally, since short-term numerical weather prediction and ocean forecasts are primarily dependent on estimation of the initial conditions rather than model parameters.

However, over the time scale of climate forecasting, the initial conditions become less significant and the model climatology is more dependent on parameter values. It was demonstrated by Anderson (2001), using the chaotic low-order model of Lorenz (1996), that the EnKF can also be used for parameter estimation. The technique is to consider the model parameters as part of the model state, as suggested by Derber (1989), and then they can be estimated simultaneously with the rest of the model state, with the analysis state of the augmented model containing not only conventional state variables but also the newly estimated parameter values. This relatively simple extension of the EnKF has not, to our knowledge, previously been tested in applications with realistic numerical models. However, as we shall demonstrate, it appears to offer the opportunity for near-optimal multivariate parameter estimation with large-scale numerical climate models at acceptable cost.

In the remainder of this paper, we will describe our development of this technique and its subsequent application to multivariate parameter estimation in an EMIC.

3 The climate model

We use a new EMIC called C–GOLDSTEIN, which features a reduced physics (frictional geostrophic) 3–D ocean model coupled to a 2–D energy-moisture balance model of the atmosphere and a thermodynamic sea ice model. The model has been extensively described elsewhere (Edwards and Marsh, 2003), and the practicality of assimilating data into the ocean component of this model has been demonstrated by Hines and Killworth (2001). C-GOLDSTEIN is the prototype climate component of a new UK Earth System Model of Intermediate Complexity, being developed under the Grid-ENabled Integrated Earth system model (GENIE) project (http://www.genie.ac.uk). A major part of the GENIE project is the implementation of the model within a Grid-based computing environment. However the computing resource used in this research was a more traditional multiprocessor Compaq Alpha supercomputer running MPI-Fortran90. Although the ensemble members can be integrated independently, the analysis method requires all model states to be collected together at frequent intervals, and so Grid-based computation may be not be particularly appropriate for this algorithm.

The simplified, coupled global climate model comprises a 3-D ocean component, an energy and moisture balance atmosphere component and a dynamic and thermodynamic sea-ice component. Figure 1 illustrates the resolution and land geometry of the model. The resolution in this version of C-GOLDSTEIN is very coarse, 10° longitude by 3° – 15° latitude, the latter reflecting equal increments in sin(latitude) to ensure constant surface area per horizontal gridbox. There are 8 depth levels in the ocean. Despite this coarse horizontal resolution, it is significant

that C-GOLDSTEIN explicitly represents longitudinal variation, which is often implicit in models of intermediate complexity.

As in Edwards and Marsh (2003), we consider 12 of the major model parameters to be uncertain and will attempt to tune them by assimilating observations of model state variables. We do not describe the model fully here, but only outline sufficient details as are necessary to identify the rôle of these unknown parameters.

Ocean temperature and salinity are diffused along and across isopycnal surfaces with constant diffusivities, and advected by the frictional geostrophic velocity field. The frictional drag coefficient (representing nonlinear momentum terms) is spatially variable, increasing near steep topographic features and near the equator from a constant interior value. To counteract the dissipative effect of the drag we allow for a scaling factor of the wind stress. In the one-layer atmosphere, diffusivity of surface specific humidity is constant, while diffusivity of surface temperature is an asymmetric function of latitude. Here we wish to consider the effect of varying the width and asymmetry of this function and thus we define the temperature diffusivity κ_T via

$$\kappa_T = k_T \left(s_d \frac{2\theta + \pi}{\pi} + (1 - s_d) \frac{\exp\left(-(\theta/l_d)^2\right) - c}{1 - c} \right), \tag{13}$$

where θ is latitude (in radians), l_d and s_d are width and slope parameters and the constant c is given by

$$c = \exp\left(- \left(\frac{\pi}{2 l_d} \right)^2 \right). \tag{14}$$

The vertically averaged effect of atmospheric advective transport is represented by advection with a fixed surface wind field derived from NCEP analysis, scaled by a coefficient, following Weaver et al. (2001), which takes different values for moisture advection and for the zonal advection of heat. There is no meridional advection of heat in the model atmosphere. To compensate for a lack of inter-basin moisture transport produced by our simplified atmosphere we introduce a constant redistribution of moisture from the surface Atlantic to Pacific following the pattern observed by Oort (1983). This 'flux adjustment' is an important parameter of the model hydrological cycle, but is unrelated to the flux adjustments required in early coupled models to prevent climate drift, which is not a problem for this model. Sea-ice height and fractional area are advected by the surface ocean currents and diffused with constant diffusivity.

Fixing the spatial distributions of drag and atmospheric diffusivity, we thus have a set of 12 scalar model parameters which may vary. The prior distributions for all 12 parameters are given in Table 1. The ranges of the distributions are chosen

to cover, or exceed, a range of reasonable choices of appropriate values for such a model as discussed by Edwards and Marsh (2003).

The model itself is deterministic and although in some cases the initial conditions can influence its climatology (i.e. distinct steady states may exist) this only leads to a small additional uncertainty on top of that due to the wide variation in model parameters. So although the model itself has $O(10^4)$ state variables, the effective state space of this problem has only 12 significant dimensions. Parameter estimation over short time intervals, when the interactions between initial conditions, underlying model state and optimal parameter values may all be stronger, and observational density may also be an important factor, is likely to be a more difficult proposition.

4 Data assimilation implementation

4.1 Outline of analysis algorithm

Our analysis algorithm follows the method of Keppenne (2000) closely, so rather than giving a full description, we only outline the method with specific attention paid to the few minor modifications that we have made. Our ensemble contains 54 members, with each model run performed on a single processor of a Compaq Alpha SC supercomputer. This ensemble size was chosen for essentially geometrical and computational reasons, as it is the largest number that enables a regular decomposition of the grid into low aspect-ratio rectangles while not exceeding the maximum job size of 64 processors on this machine. For the analysis step, the global domain is divided into 54 subregions (arranged in a 9×6 array of 4×6 model grid points covering approximately $40^o \times 30^o$ areas, where all dimensions are quoted as longitude $\times$ latitude). Each processor collects the model data for all ensemble members over a particular subregion, along with observations made in this region, and then performs a local analysis. As well as ensuring a reasonable load balance between the processors assigned to the system, dividing up the domain reduces the number of observations to be simultaneously processed by a single processor which reduces the cost of the matrix inversion mentioned in the previous section. The model and analysis routines pass model state information between one another in the form of restart files written to disk. Although communicating in this way is rather inefficient, it is convenient to have a record of model states for post-analysis and recovery from computer malfunctions, and computational power is not a limiting factor for this model. During the local analysis step, a vector of coefficients is generated at each subdomain which must be smoothed across the model domain before the final analysis fields are created. We initially performed a distance-dependent smoothing (Keppenne, 2000, Eqn. 6) on these lo-

cal analysis coefficients, however a simple unweighted global averaging was found to work better in our application for reasons which are discussed in section 5.2. Another minor modification is that we apply a bilinear interpolation to the final regional calculation of the analysis increment, so that the analysis increment to the model fields varies smoothly between adjacent subdomains with no edge effects.

4.2 Parameter estimation in the EnKF

As described earlier, the parameters can be considered as part of the model state within the EnKF paradigm and they are thus automatically updated by the analysis along with the other model variables. Since we are only considering a steady state calculation (which in fact reduces the EnKF to a Wiener Filter, (Press et al., 1994, Chapter 13.3)) with temporally constant parameters in this work, equations for their evolution do not need to be explicitly coded, but in other applications they could be allowed to vary through time either deterministically or stochastically. Each processor collects the global parameter values for all ensemble members, together with the regional model fields. The analysis algorithm effectively treats the model as if it has regionally varying parameters, one set per subdomain. However the model itself is not designed in this way, so at the end of the analysis, the distinct regional values must be averaged globally for each ensemble member. It may be worth noting that the analysis method could trivially be generalised to estimate spatially varying parameter fields if the model allowed them — in fact this would be more natural than demanding globally uniform parameter values, as discussed further in section 5.2. However, with more free variables to be estimated from the same limited data set, the uncertainty of each estimate would necessarily grow.

We have also implemented a simple preconditioning procedure to improve the condition number (ratio of smallest to largest eigenvalue) of the covariance matrix. Each standard model variable (temperature, salinity etc.) is scaled by a rough approximation to the square root of its global variance, and the parameters are scaled by the width of their prior distributions. This treatment is especially helpful for the parameters, whose variances differ by several orders of magnitude from the largest (atmospheric temperature diffusion, $\sim 10^{12} m^4 s^{-2}$) to the smallest (ocean diapycnal diffusion, $\sim 10^{-10} m^4 s^{-2}$), and it has improved the typical condition number from 10^{-22} or worse to around 10^{-3}, which has greatly reduced the numerical noise in the analysis algorithm. Three of the diffusion parameters have $\pm 2\sigma$ ranges which span two orders of magnitude from smallest to largest values. For these, any approximations or small numerical inaccuracies in the analysis can generate negative values which will cause instability in the model. Therefore, these parameters are logarithmically transformed prior to the analysis, and then trans-

formed back to reinitialise the model for the subsequent model integration. This ensures that they remain positive at all times.

4.3 Estimation for a temporally averaged (steady state) problem

There is no seasonal variation in the forcing of this model, and moreover the energy/moisture balance model has such a strong damping effect that the coupled system converges to a steady state with next to no temporal variability when run indefinitely. Therefore, a snapshot of model variables (as in a restart file) is virtually identical to a long-term average, and temporally averaged data can be treated as equivalent to observations of these instantaneous values. For a more complex and temporally variable model, assimilation of averaged observations would require the model state (restart file) to be augmented with suitably averaged state variables along with the instantaneous values. This augmentation, which is essentially the same idea as we have implemented for the parameters themselves, is a standard technique for assimilating observations which are asynoptic and/or nonlinear in model variables (Evensen, 2003).

In principle, all we need to do to solve the steady state problem is to pick the initial conditions for the ensemble members from the assumed priors, run the models to steady state, and then perform a single analysis using the observed (climatological) data. However, in practice this does not work well. When there are many parameters, the curse of dimensionality implies that the support of the posterior pdf is likely to be very much smaller than that of the prior, so much so that with high probability none of the sampled parameter sets will generate realistic solutions. These samples which populate the extreme outer reaches of the posterior pdf cannot be expected to provide much useful information. Another, perhaps even greater, problem with this approach is that there are many approximations inherent in the method and also nonlinearities in the model, and as a result the analysed model states corresponding to x^a in equation 8 which are linear combinations of steady model solutions, are not themselves model solutions and are certainly not steady (in fact some of them are not even numerically stable). We have addressed these difficulties by using an iterative procedure which combines an ensemble inflation factor (Anderson, 2001) with repeated use of the data (including the prior estimates, which can be considered as direct observations of the parameter values), together with a compensating adjustment to the observational errors assigned to the data. As the name suggests, during ensemble inflation (also referred to as covariance inflation), the ensemble is spread by replacing each member by one whose distance from the ensemble mean has been increased by a fixed multiplicative factor along the same vector direction in parameter space. The covariance matrix is therefore increased by the square of this multiplicative factor. This procedure is commonly used to compensate for limited ensemble sizes

and model error, which can otherwise lead to 'filter divergence' in which an ensemble with too small a spread fails to be influenced by the data. In this case we are using the expansion to compensate for repeated use of the data (which would otherwise also result in an unrealistically narrow ensemble spread). After each iteration of the analysis step, the ensemble is integrated through time in order to restore balance and ensure that the correct covariances are maintained.

When the ensemble is spread and the data are repeatedly reinserted via the analysis scheme, the ensemble distribution converges to the point at which a balance is reached between the artificial expansion of the ensemble, and the narrowing caused by the repeated use of the data at each analysis step. The effect of artificially increasing the assumed observational errors is that at each iteration of the analysis, the posterior is only slightly narrower than the prior, so at each step the ensemble will contain some realistic simulations thus ensuring a meaningful distribution that will gradually converge to the final steady solution. We now show that with the appropriate adjustment to the assumed errors on the data, the covariance of the converged solution can be made to match that of the original problem.

Equations (9) and (10) can be combined into

$$P^a = R(P^f + R)^{-1} P^f \tag{15}$$

where for simplicity and clarity we are omitting the matrix H which represents the transformation from model space to observation space. Informally, one can consider augmenting the set of observations to include all state variables, by adding artificial observations with large independent errors, and consider the limit as the error on these additional observations increases without bound. Then $H = I$ and the solution is unchanged. As described in more detail in section 5, our set of observations in fact includes all the dominant dynamical variables of temperature and salinity at all grid points in the ocean, and temperature and humidity at all grid points in the atmosphere.

If we repeatedly use an inflation factor of e and assimilate both the observations and prior after making a multiplicative scalar correction factor of c to all of their errors, the converged solution will have the covariance matrix Q which from equation (15) satisfies

$$Q = c^2 P^a (e^2 Q + c^2 P^a)^{-1} e^2 Q, \tag{16}$$

since P^a is, by our previous definition, the covariance matrix obtained by combining the prior with observations. The integration of the model between analysis steps only serves to improve our estimate of the covariance matrix. In the case of a linear model and no numerical approximations it would have no effect and so could be omitted.

If we set $c^2 = e^2/(e^2 - 1)$ then $Q = P^a$ as desired and we converge to the correct solution. For our chosen expansion factor of 5%($e = 1.05$), the correction factor is given by $c \simeq 3.28$. The converged ensemble mean takes the correct value as it is the appropriately weighted average of prior and observations. We have not made any serious attempt to optimise the system with respect to different expansion factors, which no doubt could alter the computational efficiency. Since the purpose of this iterative procedure is to allow the ensemble members to remain reasonably well balanced and close to steady state, they are integrated though a time interval during which time any imbalance in model state and inaccuracy in the analysis covariances can be restored. The choice of time interval is discussed further in section 5.2.

One interesting feature of our system is that prior estimates are not necessarily required in order for it to work. Using a prior distribution based on expert opinion, and then assimilating observational data, runs the risk that the expert was already aware of the observations and had therefore been influenced by them. This re-use of data will result in an underestimate of uncertainty (Allen et al., 2002). The iterative system outlined here does not require a prior, and the starting point for the ensemble is merely an initial guess that will not (except in extreme or pathological cases) affect the converged solution. In our experiments, as will be seen in section 5.2, the data used are in fact insufficient to constrain all parameters and a prior is required in order to maintain numerical stability. Beyond that point, however, the priors have been chosen to be as broad as possible so as to minimise the effect on the final solution.

5 Experiments

5.1 Description

Edwards and Marsh (2003) performed a Monte Carlo experiment with this same version of C-GOLDSTEIN. In this experiment, 1000 runs each of 2000 year duration (to reach an approximate steady state) were performed, each with 12 parameters chosen independently from stratified prior distributions (each divided uniformly, or uniformly in log-space across their specified range). A cost function was then evaluated by comparing the results to the climatological ocean temperature and salinity of Levitus (1998) and atmospheric temperature and humidity from NCEP/NCAR averaged over 1948–2002, and the best model runs were assumed to be the solution to the estimation problem and examined further. The data used in this comparison are not an exhaustive description of all components of the Earth's climate system that are contained by the model (for example, data regarding sea ice cover, and major oceanic transports, are not used) but they form

a comprehensive and well validated constraint on the dominant dynamical variables in the model and thus can be expected to provide a robust guide to model performance. Since we are not attempting to estimate temporally-varying parameters but are only aiming for a reasonable climatology, we have not investigated the issue of observational density.

To test our method we perform an identical twin test loosely based upon this previous experiment. In identical twin testing, a preliminary run of the model itself is used to generate a synthetic 'truth' data set which can subsequently be used in assimilation experiments. Using output from the model itself means that the 'strong constraint' that the model can match the data is satisfied and, since the properties of this initial model run are known, it is straightforward to check whether the assimilation algorithm has itself worked correctly. Since the EnKF method requires an assumption of Gaussianity, we take the priors on the parameters to be described by Gaussian distributions with the 2σ width equal to the ranges chosen by Edwards and Marsh (2003). Three of the parameters vary by 2 orders of magnitude in their priors and since these are logarithmically transformed, their priors are also taken to be Gaussian in log-space. The 'truth' data is generated by running the model to equilibrium with a set of parameters selected at random from the priors. The model is well converged within 5000 years from uniform initial conditions. Some output from this run is shown in Figure 1, which also illustrates the resolution and geography of the model. Since Edwards and Marsh (2003) found that most of their 1000 randomly chosen model configurations compared rather poorly to the data, we do not expect our randomly selected truth run to look particularly realistic, but this is not the purpose of our test, which aims to check the validity and performance of the assimilation method as distinct from the model's fidelity. In order to create a data set for assimilation from this truth run (and in anticipation of performing an experiment using real observations of the same variables) we observed the same state variables as were used in the Edwards and Marsh (2003) experiment, namely three-dimensional ocean temperature and salinity, and atmospheric temperature and humidity, at all grid points. Random Gaussian errors were added to these, the magnitude of which was 1 degree error in the ocean temperature, and the errors on the other fields were scaled proportionally to the square root of the global variances of the data, resulting in values of 0.082 psu error in ocean salinity, 1.97 degree error in atmospheric temperature and 8.24e-4 error for atmospheric humidity (the unit here is the dimensionless mass ratio of water vapour to air). One reason for using these values is to achieve some comparability with Edwards and Marsh (2003), since these are the weightings used in their cost function and, as we have seen in section 2, the optimum defined by the minimum of the cost function with specific weights is equal to the the Bayesian estimate of the model state under the condition that the observational errors are equal to those weights.

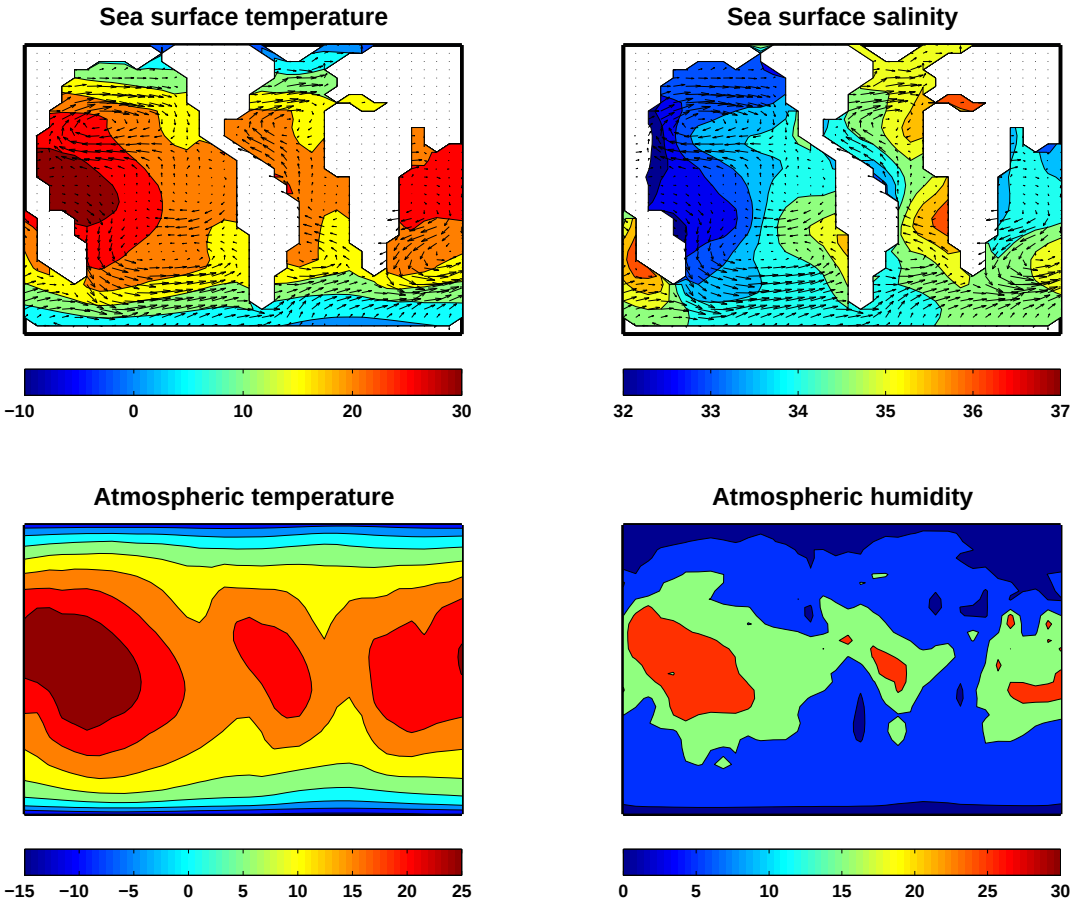

Fig. 1. The truth run

The ensemble is initialised by choosing parameters randomly from the priors, with a spatially constant ocean temperature selected randomly in the range 10–25°C (the random initial temperature has a small effect on the final steady state). Some of the sets of parameters cause numerical instability, and replacements were run as necessary. Note that this initial ensemble does not have to be sampled from the prior, it only serves as the first guess in the iterative procedure and does not affect the final solution. These models are integrated for 5000 years, and then the first assimilation is performed. The model is then integrated for a fixed time interval between analysis steps until the ensemble has converged. We have checked that the ensemble members had all reached a steady state by integrating them on with no analysis steps for another 2000 years in which time there was very little drift in individual members (and negligible drift in ensemble mean and variance).

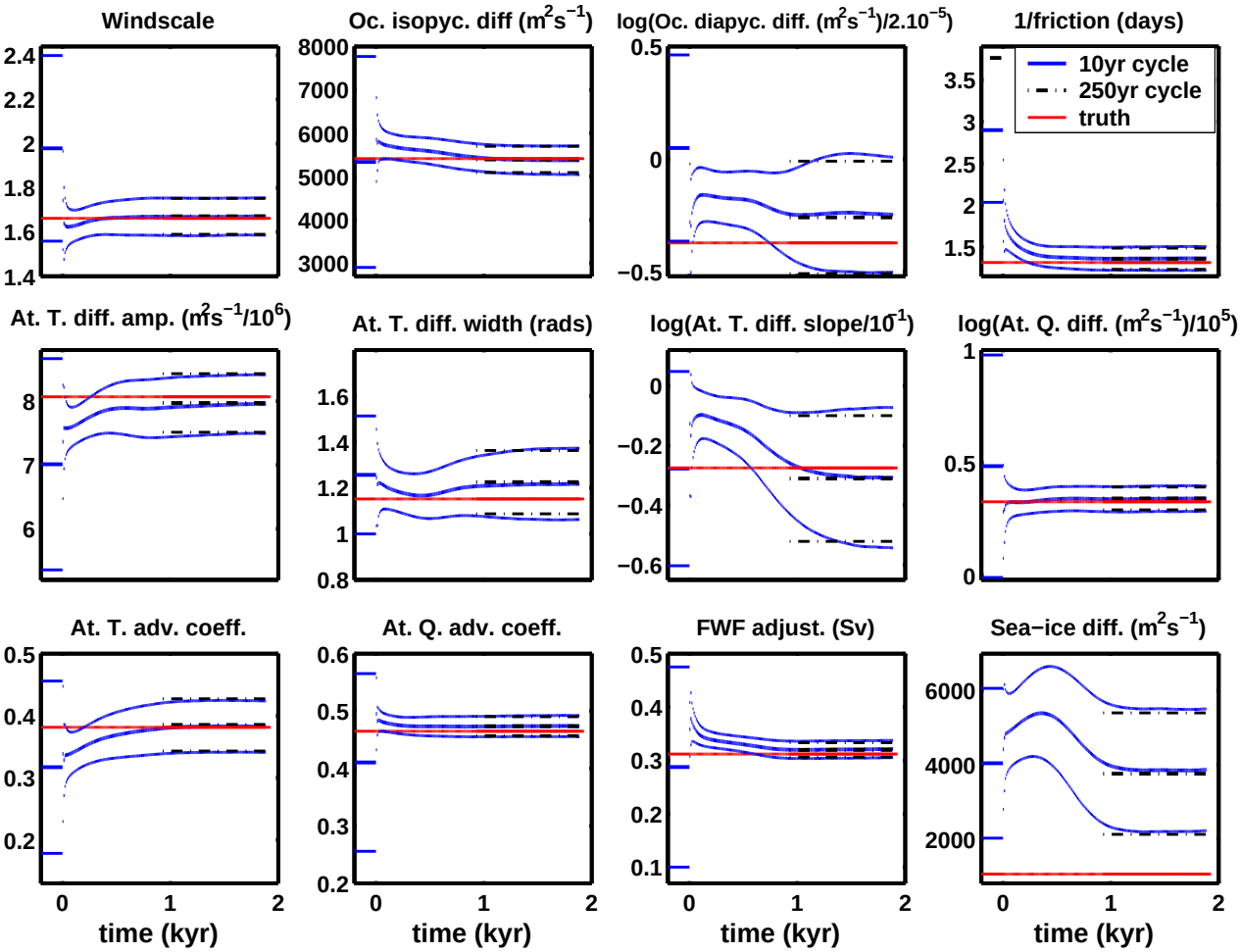

Fig. 2. Evolution of parameter distributions

5.2 Results and discussion

The evolution of the parameter distributions is shown in Figure 2. The constant parameter values over time $t < 0$ indicate the prior distribution. The parameter values used in the truth run are shown by the horizontal red lines on each plot. The thick and thin blue line indicate the mean and one standard deviation width respectively in the evolution of parameter values during a 2000 year assimilation experiment using a 10-year assimilation interval (200 analysis cycles). The black dot-dashed lines indicate the converged solution when a 250 year assimilation interval was used, for which the total run length was 15,000 years (60 cycles). It is clear that all of the 12 parameters converge to distributions consistent with the truth value and their distributions generally narrow considerably, although the sea-ice diffusivity is hardly affected by the process and several other parameters remain rather poorly constrained. The width of each posterior parameter distri-

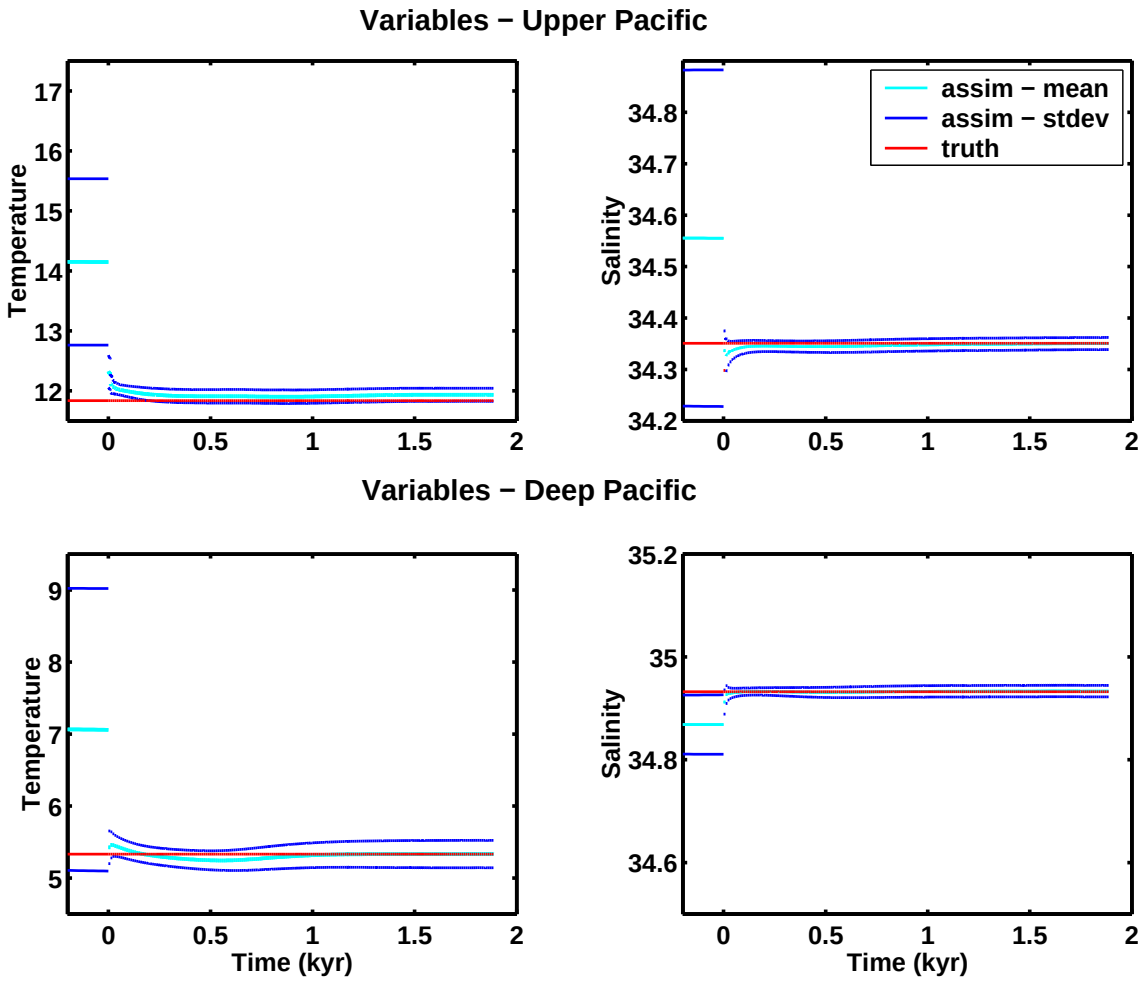

Fig. 3. Evolution of model state

bution is comparable to the difference between truth and posterior mean, which is a further confirmation that the method is working appropriately. There is clearly not enough information in the data to constrain all of the parameters precisely. Another way of putting this is to say that the model output (at least that which was assimilated) is insensitive to the values of the unconstrained parameters.

Table 1 gives the parameter values for the truth run, together with the prior and posterior distributions. The posterior parameter distributions were checked for covariances, and a few significant relationships were found. The width and amplitude parameters for atmospheric temperature diffusion had a strong negative correlation of about -0.8, which is to be expected since a locally strong diffusion that decays rapidly with increasing distance from the peak will have broadly similar net effect on poleward heat transport as a weaker diffusion that is maintained across a wider domain. The only other parameter that shows strong relationships with others is the atmospheric moisture advection coefficient which is positively

correlated with ocean isopycnic diffusion ($r^2 = 0.6$), ocean friction^{-1} (0.5) and atmospheric moisture diffusion (0.7). These parameters are all expected to have a significant effect on the meridional overturning strength in the model, and thus on the deep-ocean temperatures and salinities which have a large influence on the cost function. Moisture transport parameters affect overturning via the surface ocean freshwater forcing and are implicated in model hysteresis effects (Edwards and Marsh, 2003).

The variance of the sampling error of a small correlation, with an ensemble of N members, is about $1/N$ (Houtekamer and Mitchell, 1998) which means that a sample correlation must have absolute magnitude in excess of 0.27 before it can be said to be significantly non-zero at the 2σ level.

Although the algorithm is robust with respect to occasional model crashes (missing restart files are replaced by a random duplication of one of the existing members), a significant number of failures occur near the start of the experiment due to numerically unstable combinations of parameters. This causes the ensemble to collapse on to a lower-dimensional space through duplication of existing ensemble members. A more sophisticated replacement strategy would be to use a linear combination of stable runs to generate a new unique ensemble member, but this has not been found to be necessary. The use of the 'perturbed observations' approach (Burgers et al., 1998), in which the analysis of each ensemble member uses a different randomly perturbed set of observations, together with an ensemble expansion factor, ensures that the coincident ensemble members separate again in any case. In early testing, we found that the sea-ice parameter would occasionally become negative or large and positive and either of these will cause numerical instability in that ensemble member. Although the crashing of an individual run is not in itself a problem, occasionally the unstable model run would not get quite this far and instead produce noisy fields which contaminated the rest of the ensemble during the analysis cycle. Therefore, hard lower and upper bounds of 1000 $\mathrm{m^2 s^{-1}}$ and 9000 $\mathrm{m^2 s^{-1}}$ were placed on this parameter, which does not visibly affect the final solution but prevents this undesirable behaviour during the transient period. Logarithmic transformation of the parameter would also have been possible, and worked well with three other diffusion parameters which were so treated. Ultimately, this problem can be blamed on an inaccurately specified prior since the broad Gaussian distribution used has a small but non-negligible probability assigned to negative and large positive values even though we know these to be impossible on physical and numerical grounds respectively.

Figure 3 shows the evolution of the ocean temperature and salinity averaged in the upper and lower Pacific. The variables all converged to the appropriate values. Other regions all had similar properties. It is notable that the model state variables converge much more rapidly than the parameter values. They are of course directly adjusted towards the truth data during the assimilation step, whereas

the parameters are only indirectly altered via the model-generated covariances between themselves and the state variables, which must be recalculated through model dynamics during the iterative process.

Parameter	Truth	Prior		Posterior	
		Mean	std.dev.	Mean	std.dev.
Ocean					
Wind-scale	1.7	1.9	0.4	1.7	0.1
Isopycnal diffusion (m^2s^{-1})	5400	5100	2300	5400	300
log(diapycnal diffusion (m^2s^{-1})/2.10^{-5})	-0.37	0.07	0.44	-0.25	0.246
1/friction (days)	1.3	2.7	0.9	1.4	0.1
Atmosphere					
T diffusion amplitude (m^2s^{-1})/10^6	8.1	6.9	1.5	8.0	0.5
T diffusion width (radians)	1.2	1.3	0.3	1.2	0.2
log(T diffusion slope/10^{-1})	-0.28	-0.32	0.36	-0.31	0.24
T advection coefficient	0.38	0.35	0.17	0.38	0.04
log(Q diffusion (m^2s^{-1})/10^5)	0.33	0.57	0.43	0.34	0.06
Q advection coefficient	0.47	0.40	0.19	0.47	0.02
FWF adjustment (Sv)	0.97	1.0	0.5	1.0	0.05
Ice					
Sea-ice diffusion (m^2s^{-1})	1100	3900	1800	3900	1600

Table 1
'True' parameter values, prior and posterior distributions. Parameter descriptions are given in section 3.

Initially we used the distance-based weighting of Keppenne (2000, Eqn. 6) with a length scale of 5000 km, and so each subregion generated a set of distinct parameter values for each ensemble member. The analysis fields are by construction locally balanced with these parameter values, and therefore replacing the field of spatially varying parameter values with a single global average for each ensemble member introduced a small amount of local imbalance. This took around 100 years to die away on subsequent integration, so we initially used an analysis cycle of 250 years. The resulting system took around 15,000 years of model time to converge

(60 analysis cycles). The optimisation of a global parameter value to fit all observations in fact reveals a minor inconsistency in the underlying philosophy of the local analysis method. It is widely acknowledged that local analysis methods are useful for initial state estimation since long-range correlations between state variables are weak and unreliable (Houtekamer and Mitchell, 1998; Keppenne, 2000), even though excessive localisation can cause loss of balance (Mitchell et al., 2002). However, in our steady-state parameter estimation case, climatological observations in the Pacific should and will influence the steady state model fields in the Atlantic via their effect on global parameter values. A simple unweighted global averaging of the local coefficient vectors mentioned in section 4.1 (equivalent to smoothing with an infinite length scale) virtually eliminated the local imbalance in the analyses. In all other respects (time to convergence, and final results) the output from the assimilation was essentially identical for both of these cases when run with a 250 year analysis cycle (not shown). With the much smoother analysis fields resulting from the global smoothing, however, it is possible to substantially reduce the length of each integration interval. With a 10 year interval and the same expansion/correction factors, the solution converged much more rapidly terms of model integration time (2000 years, or 200 analysis cycles), and these are the main results plotted in Figure 2. For this highly efficient numerical model, which only has 100 model time steps per year, the total time required for this calculation is actually greater than for the 15,000 year integration on a 250 year analysis cycle, due to the larger number of analysis cycles and the consequent increase in file writing. We have not attempted shorter intervals, but see no reason why this could not be further adjusted as required if other more computationally demanding models are used. However, the convergence time will remain bounded below by model physics no matter how efficient a numerical method is used, since it is the model physics that ensures that the correct covariances are maintained between all variables.

Another factor that could in theory bound the time to convergence is that we must perform a sufficient number of iterations of the analysis cycle to ensure that the data has been fully taken account of, since we are artificially downweighting its importance in each analysis by inflating the assumed observational errors. In our example with a 5% ensemble inflation factor, at least $c^2 \simeq 11$ iterations are required before the total impact of the discounted data can equal its true value. In our examples, the model physics provides a stricter bound, but this may not be the case in all applications. For example, with a smaller ensemble inflation factor of 1%, the error correction factor would increase to 7.1 and at least O(50) iterations would be required.

6 Conclusions

We have presented an efficient method for parameter estimation and ensemble generation based on the ensemble Kalman filter. It has been successfully applied to a new computationally efficient EMIC to tune its climatology to identical twin data through simultaneous estimation of 12 parameters. The data used are not adequate to tightly constrain all of the parameters, but the posterior distributions for both parameters and state variables are consistent with the synthetic 'truth' data. Further work is underway, using real observations and applying the technique to time-varying estimation problems in tandem with the ongoing development of the GENIE model. The method presented here uses 54 ensemble members, each being a single model run which is integrated for a total time of similar order to the convergence time of the model. As such, this represents a massive increase in computational efficiency when compared to other multifactorial and Bayesian sampling methods which have been generally used for these problems.

7 Acknowledgements

NRE is currently supported by the Swiss NCCR-Climate programme. Supercomputer facilities and support were provided by JAMSTEC. This research was partly supported by the GENIE project (http://www.genie.ac.uk/), which is funded by the Natural Environment Research Council (NER/T/S/2002/00217) through the e-Science programme. We are grateful for two referees for their helpful comments.

References

Allen, M., Kettleborough, J., Stainforth, D., 2002. Model error in weather and climate forecasting. In: Proceeduings of the 2002 ECMWF predictability seminar. ECMWF, Reading, UK, pp. 275–294.

Anderson, J. L., 2001. An ensemble adjustment Kalman filter for data assimilation. Monthly Weather Review 129, 2884–2902.

Andronova, N. G., Schlesinger, M. E., 2001. Objective estimation of the probability density function for climate sensitivity. Journal of Geophysical Research 108 (D8), 22,605–22,611.

Burgers, G., van Leeuwen, P. J., Evensen, G., 1998. On the analysis scheme in the ensemble Kalman filter. Monthly Weather Review 126, 1719–1724.

Chavent, G., 1974. Identification of functional parameters in partial differential equations. In: Goodson, R. E., Polis, M. (Eds.), Identification of Parameters in Distributed Systems. ASME, New York, pp. 31–48.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49 (4), 327–335.

Derber, J., 1989. A variational continuous assimilation scheme. Monthly Weather Review 117, 2437–2446.

Dilks, D. W., Canale, R. P., Meier, P. G., 1992. Development of Bayesian Monte Carlo techniques for water quality model uncertainty. Ecological Modelling 62, 149–162.

Edwards, N. R., Marsh, R., 2003. Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model. Climate Dynamics(Submitted).

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99 (C5), 10143–10162.

Evensen, G., 2003. The ensemble kalman filter: theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367.

Evensen, G., van Leeuven, P. J., 1996. Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. Monthly Weather Review 124, 85–96.

Forest, C. E., Allen, M. R., Stone, P. H., Sokolov, A. P., 2000. Constraining uncertainties in climate models using climate change detection techniques. Geophysical Research Letters 27, 569–572.

Gregory, J. M., Stouffer, R. J., Raper, S. C. B., Stott, P. A., Rayner, N. A., 2002. An observationally based estimate of the climate sensitivity. Journal of Climate 15 (22), 3117–3121.

Hansen, J., Sato, M., Nazarenko, L., Ruedy, R., Lacis, A., Koch, D., Tegen, I., Hall, T., Shindell, D., Stone, P., Novakov, T., Thomason, L., Wang, R., Wang, Y., Jacob, D., Hollandsworth-Frith, S., Bishop, L., Logan, J., Thompson, A., Stolarski, R., Lean, J., Willson, R., Levitus, S., Antonov, J., Rayner, N., Parker, D., Christy, J., 2002. Climate forcings in Goddard Institute for Space Studies SI2000 simulations. Journal of Geophysical Research 107 (D18), 4347–.

Hargreaves, J. C., Annan, J. D., 2002. Assimilation of paleo-data in a simple Earth system model. Climate Dynamics 19 (5–6), 371–381.

Harmon, R., Challenor, P., 1997. A Markov chain Monte Carlo method for estimation and assimilation into models. Ecological Modelling 101, 41–59.

Hines, A., Killworth, P. D., 2001. An inversion-assimilation approach using hydrographic data in a coarse-resolution ocean model. Journal of Atmospheric and Oceanic Technology 18, 1503–1520.

Houtekamer, P. L., Mitchell, H. L., 1998. Data assimilation using an ensemble Kalman filter technique. Monthly Weather Review 126, 796–811.

Houtekamer, P. L., Mitchell, H. L., 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. Monthly Weather Review 129, 123–137.

Johns, T. C., Gregory, J., Ingram, W., Johnson, C., Jones, A., Lowe, J., Mitchell, J., Roberts, D., Sexton, D., Stevenson, D., Tett, S., Woodage, M., 2003. An-

thropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios. Climate Dynamics 20, 583–612.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. J. Basic Engineering 82D, 33–45.

Keppenne, C. L., 2000. Data assimilation into a primitive-equation model with a parallel ensemble Kalman filter. Monthly Weather Review, 1971–1981.

Keppenne, C. L., Rienecker, M., 2001. Development and initial testing of a parallel ensemble Kalman filter for the Poseidon isopycnal ocean general circulation model. Monthly Weather Review(Submitted).

Knutti, R., Stocker, T. F., Joos, F., Plattner, G.-K., 2002. Constraints on radiative forcing and future climate change from observations and climate model ensembles. Nature 416, 719–723.

Lea, D. J., Allen, M. R., Haine, T. W. N., 2000. Sensitivity analysis of the climate of a chaotic system. Tellus 52A, 523–532.

Lea, D. J., Haine, T. W. N., Allen, M. R., Hansen, J. A., 2002. Sensitivity analysis of the climate of a chaotic ocean circulation model. Q. J. R. Meteorological Soc. 128 (586), 2587–2605.

Levitus, S., 1998. World Ocean Atlas. Http://www.cdc.noaa.gov/.

Lorenz, E. N., 1996. Predictability: a problem partly solved. In: Proc. ECMWF seminar on predictability. ECMWF, Reading, UK, pp. 1–18.

McKay, M., Beckman, R., Conover, W., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Mitchell, H. L., Houtekamer, P. L., Pellegrin, G., 2002. Ensemble size, balance, and model-error representation in an ensemble kalman filter. Monthly Weather Review 130, 2791–2808.

Oort, A. H., 1983. Global atmospheric circulation statistics, 1958–1973, NOAA Professional Paper 14.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1994. Numerical recipes in Fortran: the art of scientific computing. Cambridge University Press.

Qian, S., Stow, C. A., Borsuk, M. E., 2003. On Monte Carlo methods for Bayesian inference. Ecological Modelling 159 (2–3), 269–277.

van Leeuwen, P. J., 1999. Comment on: "Data asimilation using an ensemble Kalman filter". Monthly Weather Review 127, 1374–1377.

Weaver, A. J., Eby, M., Wiebe, E. C., Bitz, C. M., Duffy, P. B., Ewen, T. L., Fanning, A. F., Holland, M. M., MacFadyen, A., Matthews, H. D., Meissner, K. J., Saenko, O., Schmittner, A., Wang, H. X., Yoshimori, M., 2001. The UVic Earth System Climate Model: Model description, climatology, and applications to past, present and future climates. Atmosphere–Ocean 39 (4), 361–428.

Wilks, D. S., 1995. Statistical methods in the atmospheric sciences. Academic Press.