Efficient identification of ocean thermodynamics in a physical/biogeochemical ocean model with an iterative Importance Sampling method

J. D. Annan and J. C. Hargreaves

RIGC/JAMSTEC, Yokohama, Japan

Abstract

Efficient identification of parameters in numerical models remains a computationally demanding problem. Here we present an iterative Importance Sampling approach and demonstrate its application to estimating parameters that control the heat uptake efficiency of a physical/biogeochemical ocean model coupled to a simple atmosphere. The algorithm has similarities to a previously-developed ensemble Kalman filtering (EnKF) method applied to similar problems, but is more flexible and powerful in the case of nonlinear models and non-Gaussian uncertainties. The method is somewhat more computationally demanding than the EnKF but may be preferred in cases where the approximations that the EnKF relies upon are unsound. Our results suggest that the three-dimensional structure of ocean tracer fields may act as a useful constraint on ocean mixing and consequently the heat uptake of the climate system under anthropogenic forcing.

Key words: parameter estimation, EMIC, SIR

1. Introduction

Climate models are one of the primary tools through which predictions of climate change can be made (Meehl et al., 2007). However, the model results can be highly dependent on the values of model parameters which are not adequately constrained either by direct process-based observations or by theoretical arguments, and therefore can only be estimated by the inverse process of comparing the model output to observations of the real world. Such calibration of models to observational data remains a significant challenge in climate science, primarily due to the vast computational challenge it poses. Therefore, a range of approaches have been developed for more efficient parameter estimation in climate science in recent years (Annan and Hargreaves, 2007). One such approach is the ensemble Kalman filter (EnKF; Kalman (1960); Evensen (2003)), which has been used for multivariate parameter estimation in climate models (Annan et al., 2005a). While even efficient ensemble methods such as this cannot easily be applied to the largest numerical models due to the computational costs, the development of such methods ensures thats we can make effective use of Earth system models of intermediate complexity (EMICs, Claussen et al., 2002).

In this paper we have two main goals. Firstly, in Section 2, we introduce the new parameter estimation method, which is based on an iterative Importance Sampling approach. The method can be interpreted as a natural generalisation of our previous work using the ensemble Kalman filter (Annan et al., 2005a), but is more accurate and flexible in the case of nonlinear models. We test the method with some idealised examples in Section 3, which demonstrates that the new approach is substantially more accurate than the EnKF for nonlinear problems, and is capable of estimation of around 10 parameters simultaneously, at reasonable computational cost. Secondly, in Section 4, we demonstrate successful application of the method to an Earth system Model of Intermediate Complexity, using identical twin experiments to check the performance of the algorithm and investigate the identifiability of ocean heat uptake efficiency from climatological observations of tracer fields. We conclude the paper in Section 5.

2. An Iterative Importance Sampling method for parameter estimation

The generic model calibration problem is most naturally considered as a problem in Bayesian estimation (Bernardo and Smith, 1994). That is, given a prior belief $p(\mathbf{x})$ over any uncertain model parameters $\mathbf{x}$, a model M and an observational data set $\mathbf{o}$ from which we can construct a likelihood function $p(\mathbf{o}|\mathbf{x})$ which describes the relative probability of the observations for different sets of parameters, how can we efficiently estimate the posterior probability density function (pdf) $f(\mathbf{x}) \equiv p(\mathbf{x}|\mathbf{o}) = p(\mathbf{o}|\mathbf{x})p(\mathbf{x})/p(\mathbf{o})$?

The direct Monte Carlo approach based on rejection sampling (Hammersley and Handscomb, 1964) is a simple and popular method which has been widely used in climate science in recent years (eg Knutti et al., 2002). in this approach, we draw samples from the prior $p(\mathbf{x})$ and assign each one a relative probability or weight defined by $w(\mathbf{x}) \equiv p(\mathbf{o}|\mathbf{x})$. This approach is often very expensive. In particular, the vast majority of samples may be given negligible weight if the prior is substantially more diffuse than the posterior, and in this case it may take a very large number of samples (each one of which requires a model integration to evaluate the likelihood function) to populate the posterior and achieve reasonable convergence in distribution. While this problem is particularly severe in high dimensional problems where the ensemble is liable to collapse to a single sample (Bengtsson et al., 2008), such particle-based methods may still require uncomfortably large ensembles in even problems of moderate dimension.

In cases such as this, Importance Sampling may lead to large improvements (Doucet et al., 2000). In this approach, samples are drawn not from the prior, but from some "proposal distribution" $g(\mathbf{x})$ which is believed to approximate the posterior. When the weights are correctly adjusted for this biased sampling (ie by using $w(\mathbf{x}) \equiv f(\mathbf{x})/g(\mathbf{x})$), the final outcome is the same in the limit of infinite sample size but, for a well-chosen proposal distribution, convergence can be much more rapid in practice. The best possible proposal distribution would be the posterior itself (for which $w = 1$ always), but of course we do not have the ability to sample efficiently from this distribution.

The method of "bridging densities" has been proposed as a means of increasing the efficiency of Monte Carlo sampling in such situations (Meng and Wong, 1996; Gelman and Meng, 1998; Del Moral et al., 2006). The basic principle is that given an initial proposal that is some way distant from the prior, it may be more efficient to define some intermediate "bridging" distribution such that we can use the initial proposal to generate samples from the bridging distribution, and then use the bridging distribution as a proposal from which we generate samples from the posterior. For a suitably chosen bridging density, this can be substantially more efficient than attempting to directly generate the posterior by sampling from the proposal. The approach generalises directly to a larger number of bridges, or even an infinite sequence (Neal et al., 1993; Gelman and Meng, 1998).

One natural approach is to consider the geometric family

$$\phi_\alpha = g^{1-\alpha} f^\alpha,\ 0 \leq \alpha \leq 1$$

which transforms smoothly from g to f as α varies from 0 to 1. Even if it is very inefficient to use g directly as a proposal density for f, if we select an increasing sequence of closely-spaced α_i we can iteratively use ϕ_{α_i} as a proposal for $\phi_{\alpha_{i+1}}$ and ultimately reach (or at least approach in the case of an infinite series) the target distribution f. The choice of g here may be arbitrary, but in the examples presented below we use the prior for convenience.

It is well-known that in repeated applications of such particle-based methods, the weights will become increasingly concentrated on a smaller proportion of the samples, representing a reduction in effective ensemble size and therefore loss of accuracy (Doucet et al., 2000). Therefore, some procedure is required to equalise the weights, and in this paper we use the standard approach of stratified resampling. In the case of parameter estimation problems, this itself introduces a further complication. Since the model parameters are considered fixed and do not evolve in time, stratified sampling will merely result in exact duplicates of parameter sets which will do nothing to increase the effective ensemble size. To address this problem, it is common to add some jitter to the new samples. A convenient choice for the jitter kernel is a scaled version of a Gaussian approximation to the existing ensemble spread. However, the addition of jitter in this way inevitably results in an increase in the variance of the ensemble and loss of information. To address this issue, West (1993) introduced the idea of a shrinkage step in which the ensemble of jittered samples is immediately contracted towards its mean. When the magnitude of shrinkage is correctly chosen, this restores the variance of the ensemble to the original (correct) value. It should be noted that the shape of the distribution is only precisely maintained in the case of it being a multivariate Gaussian.

We have tested the approach of using bridging distributions with jitter compensated by shrinkage, but although it works well in very low dimensional problems we have found it difficult to ensure that the ensemble converges to the correct solution for more than about 3–4 parameters, with tolerable ensemble sizes. The specific difficulty we have encountered manifests itself as an over-rapid collapse of the ensemble to a narrow region of parameter space, sometimes referred to as "filter divergence". The bridging distributions as presented above are sequentially nested and it is difficult for a distribution which is inappropriately over-narrow to recover the correct spread, since the addition of jitter (the only step whereby it can expand) is immediately counteracted by the shrinkage step. Therefore, we now present a minor variation of iterated Importance Sampling (IIS) which we have found to work better in our applications. Instead of using an explicit shrinkage step which is followed by importance sampling to a narrower distribution, we simply perform the importance sampling directly on the jittered ensemble, but change the weighting function to account for the extra spread generated by the jitter. As with the standard shrinkage procedure, this approach is only precisely correct in the case of a linear Gaussian problem. However, the solutions it generates are substantially more accurate than the EnKF approach for the nonlinear problems we have tested, and in contrast to the conventional method, we have found it to work reliably for at least 10 parameters.

In detail, our modified procedure is as follows. Given an ensemble of samples drawn from the distribution

$$\phi_{\alpha_i,\beta_i} = g^{1-\beta_i} f^{\alpha_i}$$

for some α_i and β_i (which in contrast to the established approach, are not necessarily equal here), we firstly use this as a proposal for $g^{1-\beta_i} f^{\alpha_i+\epsilon}$ by reweighting the samples according to f^{ϵ}, where ϵ is a tunable parameter which we typically set to 0.05 unless otherwise stated. Resampling with the addition of jitter (with the jitter drawn from a Gaussian kernel fitted to the ensemble with its variance scaled by a factor of ϵ) will, at least in the case where the ensemble truly is a multivariate Gaussian, generate an ensemble which samples the distribution $g^{\frac{1-\beta_i}{1+\epsilon}} f^{\frac{\alpha_i+\epsilon}{1+\epsilon}}$. Defining $\alpha_{i+1} = \frac{\alpha_i+\epsilon}{1+\epsilon}$ and $1-\beta_{i+1} = \frac{1-\beta_i}{1+\epsilon}$ respectively, this ensemble now serves as the proposal for the next iteration. It is easily seen that over repeated applications of these steps, the sampling distribution converges to $g^0 f^1 = f$ as desired. Several applications below also demonstrate the correctness of this approach. We note that the repeated use of (a scaled version of) the likelihood function, balanced by expansion of the ensemble around its mean, is fundamentally the same approach as previously adopted using the ensemble Kalman filter (Annan et al., 2005b), with the jitter here taking the place of the variance inflation step in the previous approach, and the weighting according to the likelihood function taking the place of the analysis step. The main difference here is that the data here enter the process through weighting according to the likelihood function, rather than using the Kalman equations to interpolate (or extrapolate) according to the covariance matrix. Thus, while our new method generally requires a somewhat larger ensemble to ensure adequate sampling, it has the benefit of not relying so strongly on the distribution being approximately Gaussian, and we shall demonstrate the benefit of this in some applications.

We mention in passing that there is an important difference between our approach and the iterative resampling approach of West (1993), in that we are *not* attempting to sample the true posterior f at each stage in our iterative sequence. Thus, we expect our approach to be substantially less efficient in the cases where we already have a reasonable proposal distribution (including those cases where the prior is not much broader than the posterior and thus can serve as the proposal distribution). However, in many cases of interest to climate scientists, we have no reasonable proposal density and, as mentioned above, a direct attempt to construct the posterior by rejection sampling from the prior is likely to fail through an immediate collapse of the sample.

3. Application to idealised problems

3.1. Univariate problem

In order to test the validity and accuracy of this method, we start with some simple univariate applications for which an accurate solution is easily computed. Our iterative methodology has no advantage here over a more standard approach, since there is no curse of dimensionality to address. In Annan and Hargreaves (2007), a simple nonlinear toy example was used to explore the performance of the EnKF. Applying the IIS methodology to this problem generated improved results, with the error roughly halving (not shown here). However, this problem was unchallenging in that the posterior pdf was unimodal and the mapping of parameter to output was monotonic, so even the EnKF gave rather accurate results. Here we try a slightly more challenging example where the output is a quadratic function of the input parameter and has two local maxima in the observational constraint.

We use one uncertain input x, a model given by

$$y = x^2$$

and an observation of $y_o = 25 \pm 50$ (all input uncertainties are Gaussian and quoted at one standard deviation), so there are two modes in the observational likelihood at $x = \pm 5$. An off-centre prior estimate for $x_o = 5 \pm 10$ is used which prefers the positive root, but which also assigns significant prior probability to the negative one.

As can be seen from the results in Figure 1, the EnKF performs rather poorly here. This ensemble is substantially over-dispersed, with roughly 25% of the samples falling outside the central 99% probability interval of the correct solution. Encouragingly, the IIS results show a striking improvement, with the correct overall dispersion, the tails of the distribution greatly improved, and a very modest mismatch in the distributions around their modes.

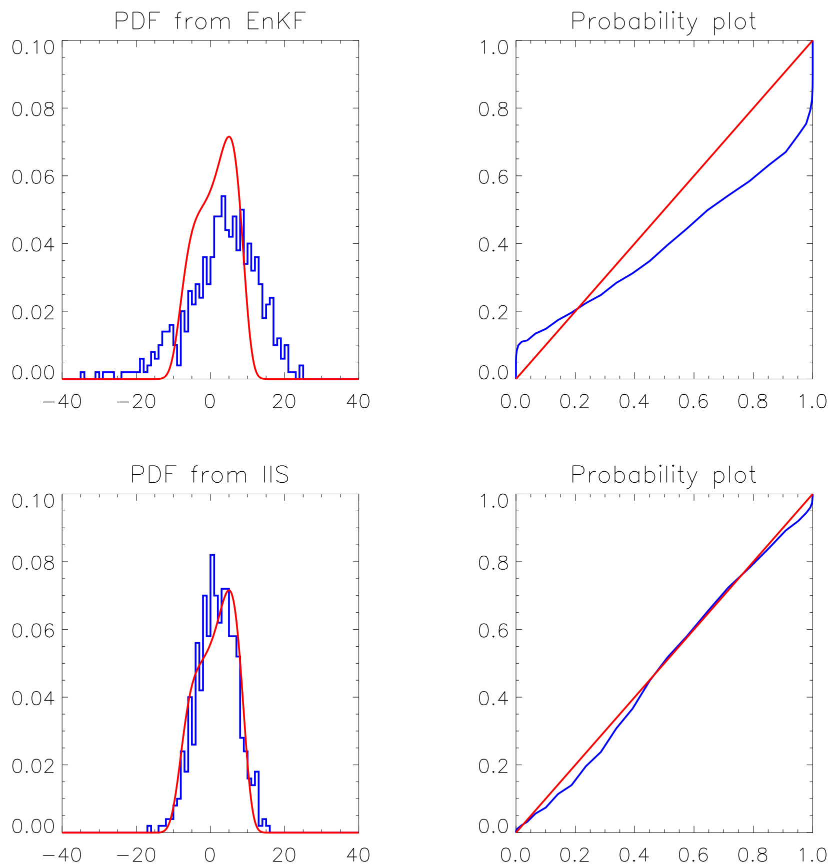

Figure 1: Comparing the performance of IIS and EnKF on a nonlinear problem. Red lines show the correct solution, blue show the experimental results. The top plots show a 500-member EnKF result, and the lower plots show the IIS result with the same size ensemble.

A common method to quantify the quality of the results is through statistical tests which aim to discriminate between different distributions, such as the Kolmogorov-Smirnov (K-S) test (Wilks, 1995, Ch. 5.2). Using this test, we investigate how confidently we can reject the null hypothesis that a finite ensemble was drawn from the true posterior distribution. Some results are presented in Table 1. With 100 replicates of an experiment using an ensemble size of 50 members, a large majority of the EnKF results are rejected as significantly different at the 1% level, whereas only a much smaller proportion of experiments using the IIS are rejected even at the 5% level. When using a 500-member ensemble, the results are even more marked, with none of the EnKF results appearing at all plausible. This is not due to the distribution changing shape with the larger ensemble (in fact it does not change detectably) but simply that with more samples, the bias in the tails of the results is more apparent and less plausibly attributable to sampling error.

Although these results do clearly indicate the greater precision of the IIS results, they also highlight a serious limitation of the K-S test in applications such as this. The K-S test statistic is based on the maximum deviation of two cumulative distributions, which will, if the samples really are drawn from the same underlying distribution, typically occur somewhere towards the median of the distributions since this is where the sample variance of a cumulative distribution is highest. However, this approach may overlook substantial differences in the tails of the distributions. A test statistic based on sampling in the tails may indicate a significant difference in the distributions even when the K-S test statistic fails to identify them as such. For example, if even as few as 5 samples from a sampled ensemble of 50 fall in the extreme tails (outside the central 99% probability interval) of a given target distribution, then this is strong evidence that the distributions are distinct, since (under the null hypothesis that the sample actually was drawn from the target) such an event can only be expected to occur with probability $\ll 1\%$. However, the absolute deviations between the cumulative distributions, of ~ 0.05 at either end, are not considered significant by the K-S test, as they would be entirely unremarkable were they to occur near the mean of the cumulative distributions. Under circumstances such as these the Kuiper test provides a stiffer hurdle to overcome (Press et al., 1994, Ch. 14.3). Using that test (also shown in Table 1), the probability of results from either method being considered significantly different from the truth increases, but the IIS method remains markedly superior.

		K-S test		Kuiper	
N	p	EnKF	IIS	EnKF	IIS
50	1%	22	97	7	91
	5%	10	87	4	84
500	1%	0	93	0	78
	5%	0	77	0	52

Table 1: Results of K-S test and Kuiper test on EnKF and IIS results with two ensemble sizes N. Values indicate number of times (out of 100 replicates) that the test does not reject at the given significance level p, that is to say the percentage probability that a single set of experimental results would be considered statistically indistinguishable from the correct solution at the p% level (according to these tests).

3.2. High dimensional linear problem

Next we test the method on a higher dimensional problem, more indicative of the input size for which the method is intended. However, in order to be able to validate the results, we revert to a linear example where the correct answer can be calculated exactly via the Kalman equations.

The example we present is very straightforward. We assume n uncertain input parameters $x_i,\ i = 1, \ldots, n$ for which we have a vague prior estimate. The linear model is a random $n \times m$ matrix M which transforms these parameters into m observed outputs y via

$$M\mathbf{x} = \mathbf{y}$$

We have a vector of observations $y_{o,j},\ j = 1, \ldots, m$, and wish to use these to generate an estimate of the inputs $\mathbf{x}$.

For the results presented here, we set $n = 16$, this being towards the high end of the number of parameters that we wish to simultaneously estimate. We also use $m = 16$, in order that the parameters are identifiable from the data (Navon, 1998). Each element in the model matrix M was an independent draw from the standard normal $N(0, 1)$. Our prior on $\mathbf{x}$ has mean 0 and standard deviation of 10 for each parameter, assumed independent. The observations of $\mathbf{y}$ are given the values $y_{o,j} = j - 8,\ j = 1, \ldots, n$ also with independent Gaussian uncertainties of magnitude 5.

For this more computationally challenging problem, the choice of the scaling factor ϵ in the iterative procedure can affect the performance of the algorithm. For a very large value, the ensemble collapses rather rapidly and may converge to a incorrect solution. This is due to the curse of dimensionality: if the prior sample is widely dispersed compared to the posterior, then the posterior weight will be concentrated on very few members and even the addition of jitter may not be enough to rescue the situation. Conversely, if the scaling factor is very large, then the weights will remain nearly uniform and the ensemble will take many iterations to converge to the true posterior. A reasonable rule of thumb arising from our experiments is to aim for a effective ensemble size that is between 50% and 90% of the actual ensemble size, and so in the results presented here the value of ϵ has been adaptively tuned to stay within these bounds.

Some typical results (using an ensemble size of 250 members) are plotted in Figure 2. It is clear that the IIS has worked correctly in this case, with the posterior suffering only from sampling error due to the finite ensemble size. It is worth emphasising the contrast in spread between the prior and posterior in this example, since this is a key motivating factor for the development of this estimation technique. The typical uncertainty of each input variable in the posterior is around 1/4 that of the prior. Therefore, a naive Monte Carlo sampling strategy would be hopelessly inefficient, as a sample from the prior has a probability of around $(1/4)^{16} \simeq 2 \times 10^{-10}$ of lying in the posterior. This problem is certainly rather more challenging than the typical application in climate science, but it gives an indication of the problem and the effectiveness of the method. The IIS method presented here has successfully populated the posterior region, using many orders of magnitude lower computational effort than direct sampling would have required.

When attempting this same problem with substantially smaller ensembles, it was not possible to reliably prevent collapse of the ensemble, and the 50-member ensemble results (also plotted in Figure 2) illustrate a typical failure. Interestingly, the EnKF approach is much more robust to such failure (not shown here), presumably through its ability to systematically interpolate and even extrapolate from the prior samples towards the posterior region, rather than relying on random jitter to perturb the locations of the samples. Therefore, in a linear application, the EnKF remains a superior choice. However, true linearity is rare in practical applications.

4. Application to a 3D EMIC

We now perform an identical twin experiment to demonstrate the application of the method to an earth system model of intermediate complexity, the

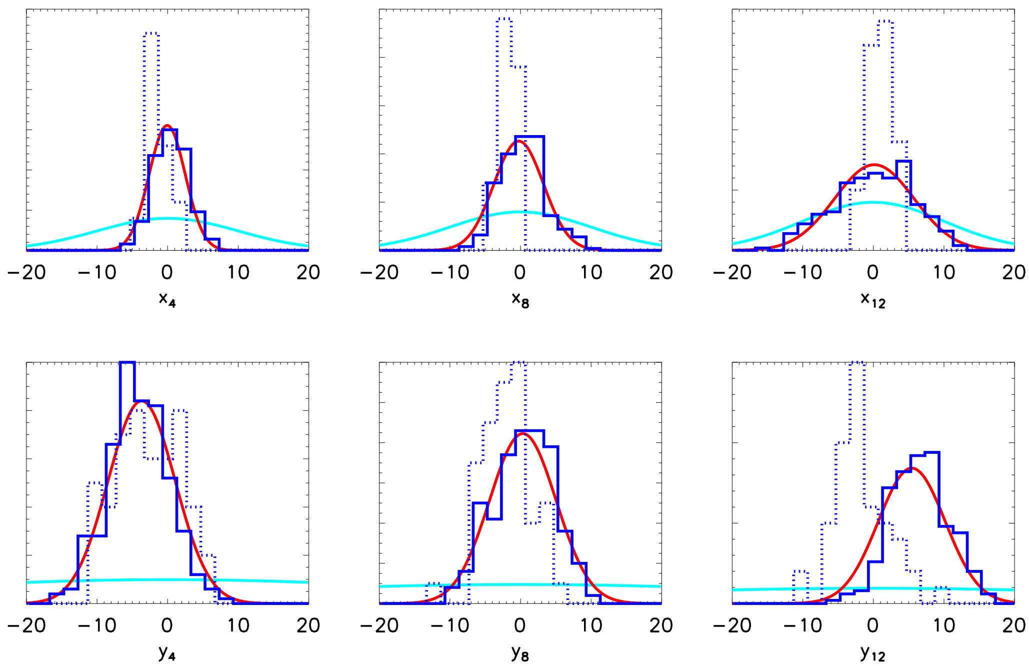

Figure 2: Testing the IIS on a 16-dimensional problem. The top row of plots show 3 of the input parameters, the lower row show 3 outputs. The cyan curves show the prior (only shown to ± 20), red indicates the true posterior, dark blue solid histogram shows results from a 250-member IIS calculation and the dotted blue histogram gives results from a 50-member ensemble which failed to converge correctly.

Grid ENabled Integrated Earth system model (GENIE: www.genie.ac.uk) (Lenton et al., 2007), which is based on the fast climate model of Edwards and Marsh (2005). Previously, we have used the EnKF methodology for estimating physical parameters (Hargreaves et al., 2004) or biological parameters (Ridgwell et al., 2007) separately in this model. Here we demonstrate simultaneous estimation of physical and biological parameters, using a variety of tracer data.

4.1. Scientific Motivation

One important model uncertainty, which is particularly relevant to ocean science, is the rate at which the surface warming (due to anthropogenic forcing) is mixed with the ocean interior. This is a first-order control on the rate of anthropogenically-forced climate change (Hansen et al., 1985). If this mixing rate is low, then the surface climate will be in near-equilibrium with the forcing, implying both relatively little committed warming at current levels of greenhouse gases, and a low rate of thermosteric sea level rise. If, however, ocean heat uptake is strong, then the thermal inertia of the ocean will allow a large radiative disequilibrium and substantial (but gradual) committed warming. Thus, this is a critical property of the climate system for understanding and addressing climate change.

Currently, there is significant uncertainty concerning estimates of mixing of the global ocean. The canonical figure of around $10^{-4}\mathrm{m}^2\mathrm{s}^{-1}$ for the overall effective diapycnal or vertical diffusion parameter (Munk, 1966) has endured fairly well (Li et al., 1984; Hoffert et al., 1985), although one more recent energy balance analysis suggests a rather lower value (Huang, 1999). These analyses all contain substantial, but poorly-quantified, uncertainties in the quantification and interpretation of the various energy sources. Thus, they do not provide adequate information for probabilistic analyses and predictions.

More recently, explicitly probabilistic analyses of ocean mixing have been performed by comparing 'perturbed parameter' ensembles of model simulations to observational estimates of warming over the 20th century (Knutti et al., 2002; Forest et al., 2006). Due to computational limitations, these analyses have generally been restricted to the use of greatly simplified models where the ocean dynamics are limited, and mixing into the deep ocean is primarily determined by a single global vertical diffusion parameter. It is not straightforward to directly equate these parameters to those used in more complex ocean GCMs, since these latter models often include a range of mixing processes (including convection and wind stirring near the surface), and the diffusion models may also incorporate spatial patterns of variable mixing. However, one striking, and perhaps worrying, aspect of the probabilistic analyses is that they have often assigned fairly high probability to values of global ocean mixing that are substantially lower than those commonly obtained in GCM simulations, with strong implications for projections of climate change (Knutti and Tomassini, 2008; Sokolov et al., 2009).

There have been very few investigations into this topic using ensembles of more complex ocean models, due primarily to the substantial computational cost this would entail. Collins et al. (2007) considered a small ensemble of ocean parameter perturbations with the fully coupled atmosphere-ocean GCM HadCM3, but could only obtain a rather small range of ocean mixing. Thus, it remains a high priority to to reconcile their results with those of Sokolov et al. (2009), and to determine which provides a more credible description of reality.

The model we use here, while computationally much cheaper than a full GCM, still has a fully three-dimensional representation of the ocean and is capable of reproducing the physical and biogeochemical properties of the global ocean reasonably well (Hargreaves et al., 2004; Ridgwell et al., 2007). The combination of computationally affordable model and efficient multivariate parameter estimation technique enables us to use various data sources for calibration of the model parameters. Thus we expect it to be a powerful tool in better constraining current estimates of ocean mixing.

4.2. Model

While the model is largely the same as used in previous work, there has been some further development which is documented here for completeness. For the physical module, we use the GENIE-1 configuration of 2D energy-moisture balance atmosphere and 3D frictional geostrophic ocean with dynamical sea ice. The ocean module is based on the 16 layer version of Singarayer et al. (2008). However, instead of modifying atmospheric temperature diffusion around Antarctica to create an appropriate cooling of high Southern latitudes in the simple energy-moisture-balance-model (EMBM) atmospheric component, we apply a zonally and annually averaged planetary albedo derived from a fully coupled GCM present-day simulation (Ridgwell et al., 2009). We also use the stratification-dependent diapycnal diffusion parameterisation of Oliver and Edwards (2008).

A coupled marine biogeochemistry module based on Ridgwell et al. (2007) calculates the redistribution of tracer concentrations due to processes other than transport by the circulation of the ocean, namely: air-sea gas exchange, the removal of nutrients, carbon, and alkalinity from solution as a result of biological activity in the sunlit surface ocean layer, the vertical export of particulate matter and its remineralization in the ocean interior, and the remineralization of dissolved organic matter and associated consumption of dissolved oxygen.

We employ a seasonal scheme for biologically-induced export out of the surface ocean based on a dual nutrient limitation of productivity by PO_4^{3-} and dissolved iron ([Fe]) derived from previously published schemes (Doney et al., 2006; Parekh et al., 2005, 2006; Ridgwell, 2001)). This differs from that described by Ridgwell et al. (2007) where it was used for an EnKF-based assimilation of marine observations, in the following ways:

1. A co-limitation of total dissolved iron on export production added, using the law of the minimum following Ridgwell (2001) and assuming a half-saturation constant for iron of 0.1 nmol kg^{-1}.
2. The effects of sub-optimal ambient light levels is implemented following Doney et al. (2006), using incident the shortwave radiation incident at the ocean surface calculated by the climate model (Edwards and Marsh, 2005) and assuming a half saturation value for light of 20Wm^{-2}. We have added a marine iron cycle based on Parekh et al. (2005, 2006), but deviating as follows:
   (a) We link the phosphate and iron cycles via an organic matter Fe:C Redfield ratio that is a function of dissolved iron availability, taking the average of the two (diatom, and non-diatom) parameterizations of Ridgwell (2001).
   (b) For iron inputs to the ocean we take the atmospheric tracer transport model generated dust field of Mahowald et al. (1999), and uniform iron content in dust of 3.5 wt%. However, we depart from the common assumption regarding a uniform solubility of iron in dust and instead allow solubility to vary inversely to dust loading consistent with laboratory experiments and observations (Ridgwell, 2001) and with a solubility that scales inversely to the square root of dust loading (flux) (Baker and Jickells, 2006).

In addition to several parameters controlling aspects of the ocean carbon cycle (and hence dissolved PO_4, ALK, and O_2 distributions) that we allowed to vary in previous EnKF-based assimilation work (Ridgwell et al., 2007), we now include the scavenging rate of dissolved iron from the water column, and the overall (global mean) solubility of iron in dust.

4.3. Data

Although the results presented here are from an identical twin experiment where synthetic data are generated from a model run, we wish in the future to apply the method to real data, and therefore the choices of data are based on those for which observational analyses are available.

The physical data we use are climatological mean fields of ocean temperature and salinity, for which global analyses such as Conkright et al. (2002) are available, and the atmospheric temperature and relative humidity which could be derived from the NCEP reanalysis (Kalnay et al., 1996). Previous work suggests that these data can constrain the ocean circulation to a reasonable state (Hargreaves et al., 2004), although a detailed quantification of the implications for heat uptake has not been performed.

In our previous marine biogeochemistry data assimilation experiment (Ridgwell et al., 2007) we utilized observed 3D distributions of phosphate (PO_4) (Conkright et al., 2002) and alkalinity (Key et al., 2004) in the ocean, to constrain model parameters controlling the marine carbon cycle. In this, observed fields of PO_4 help constrain the rates and distribution of PO_4 uptake at the ocean surface, together with the penetration depth of particulate organic matter before remineralization and release of PO_4 back to the ocean. Alkalinity (ALK) distributions place constraints on the production and dissolution of the calcium carbonate ($CaCO_3$) mineral shells and (skeletons) in the ocean. The distribution of both these tracers is affected by ocean circulation. In this study we add a further 3D field of dissolved oxygen (O_2) (Conkright et al., 2002). This is controlled not only by the remineralization of organic matter and hence bacterial consumption of oxygen in the ocean interior as well as ocean circulation, but is also sensitive to ocean surface temperature and residence time as O_2 is rather more soluble in colder waters and will reach equilibrium with the atmosphere only in relatively stratified conditions. We do not consider observational uncertainties in these tests.

4.4. Parameters

The physical and biological parameters we chose to vary are listed in Table 2, along with their prior 2.5–97.5% ranges. The physical parameters that we vary (shown in Table 2) are the subset of those used, and described in more detail, in previous work (Annan et al., 2005a), which were found to be most influential on model behaviour. For the atmospheric physics, "Q" and "T" here refer to moisture and heat respectively. The fresh water flux adjustment (FWF) from Atlantic to Pacific, a standard procedure in EMBM-type models, is implemented here as a scaling factor on the standard 0.32Sv figure of Oort (1983) rather than as an absolute value. Although presented here as an atmospheric parameter, this flux acts directly on the ocean where it strongly influences the meridional overturning circulation. The prior distributions were defined as Gaussian either in the variable or its log (for those parameters where a skewed distribution with a 50th percentile closer to the lower end was desired).

	Parameter	Prior		Posterior		Truth
		2.5%	97.5%	2.5%	97.5%	
	Oceanic physics					
1	log Isopycnal diffusion ($m^2 s^{-1}$)	250	4000	615	3700	1815
2	log Diapycnal diffusion $/10^5$ (m^2s^{-1})	0.46	26.7	1.3	16	4.54
3	1/friction (days)	0.91	4.5	2.25	3.63	3.29
	Atmospheric physics					
4	T diffusion amplitude $/10^6$ (m^2s^{-1})	3.82	9.90	4.85	8.14	6.41
5	log Q diffusion $/10^5$ (m^2s^{-1})	0.52	26	1.01	11.3	7.44
6	FWF adj ($\times 0.32$ Sv)	0.5	2.1	0.63	1.64	1.25
	Oceanic biogeochemistry					
7	log PO_4 half-saturation$\times 10^6$ (μmol kg^{-1})	0.5	3	0.69	2.22	0.88
8	Initial POC export fraction	0.03	0.07	0.033	0.07	0.066
9	log e-folding POC depth (m)	225	900	235	520	352
10	Initial $CaCO_3$ export fraction	0.25	0.65	0.30	0.61	0.50
11	log Fe solubility	0.002	0.008	0.002	0.0075	0.064
12	log Fe scavenging rate	0.4	1.6	0.4	2.2	0.62

Table 2: Prior and posterior distributions of the parameters, and the value used for the truth run. Log-normal distributions were used for the parameters prefixed by 'log'.

4.5. Experimental Details

In order to validate the method and investigate the identifiability of the parameters and physical behaviour of the model, we present the results from identical twin tests here. In this case, a truth run was selected that had a reasonably realistic overall physical and biogeochemical state from a 256-member latin hypercube ensemble (McKay et al., 1979). As in previous experiments, the physical observations we used consisted of climatological observations of three-dimensional ocean temperature and salinity, and the two-dimensional field of atmospheric temperature and relative humidity. For the ocean biogechemical model, we use 3D fields of alkalinity, oxygen and phosphate.

Although in an identical twin test it may be possible, in principle, to identify the parameters to essentially arbitrary precision, this will not be the case in any practical test with real data, since model inadequacy and observational error will always limit the precision with which the model can match the data. Thus we deliberately allow for a substantial model-data mismatch in our likelihood function, which is based on a simple sum of squares similar to that of Murphy et al. (2004) and (Edwards and Marsh, 2005) (equivalent to assuming all observational uncertainties are independent and Gaussian). In detail, we split the ocean data into 4 domains vertically (of 4 levels each), and used a cost function of the form $\sum_i \sum_j \alpha_i (x_{i,j} - o_{i,j})$ where $\alpha_i,\ i = 1, \ldots, N$ is a scaling factor over 22 disjoint subsets of the data (20 ocean, 2 atmosphere) and j indexes the spatially discrete data points in each subset. The α_i were used to normalise the contribution of each component of the cost function to the overall total, by choosing values that set each term in the sum to a value of 1 when the standard control model (not the 'truth' run in this experiment) was compared to real data. In other words, we are defining the model inadequacy to be the level of mismatch obtained by the control model, which then determines the range of uncertainty that is acceptable for the "best" set of parameters (where "best" here is used in the Bayesian sense: see (Rougier, 2007) for a more detailed description). In any realistic application the choice of cost function may have to be considered in more detail, but here we primarily wish to check that the algorithm works effectively and whether the data may be informative on the model behaviour.

Even though this model is computationally cheap, it would still be challenging to integrate it for its full equilibration time scale of O(2000) years at each iteration. Thus we rely on the observation that adding modest amounts of jitter to the model parameters does not greatly upset the quasi-equilibrium balance of the model state, so that only a more moderate period of integration (we use 200y here) is required to restore a near-equilibrium state. We checked the validity of this approximation by integrating the final ensemble on for a further 5000 years, and found that the changes were indeed very minor across the ensemble as a whole. Thus the 30 iterations of the method that we performed requires 6000y of integration time, which is only a small multiple of the spin-up time of the model itself. This behaviour is comparable to what was previously found for the EnKF applied to the same model (Annan et al., 2005a). A possible improvement for future applications would be to perturb the full state according to the covariance matrix, rather than only adjusting the parameters in isolation.

Gregory and Mitchell (1997) defined the 'ocean heat uptake efficiency' $\kappa = \frac{\Delta F}{\Delta T}$ to be the heat uptake flux to the deep ocean ΔF divided by the surface temperature anomaly ΔT. Although this is not a fixed parameter of the climate system, it is a reliable diagnostic over a period of strongly increasing forcing such as idealised 1% pa CO_2 enrichment experiments or more realistic socioeconomic emissions scenarios. In order to provide a direct comparison with the κ values calculated by Collins et al. (2007) for their ensemble of HadCM3 results, and also by Raper et al. (2002) for the CMIP3 ensemble, we also perform 1% pa CO_2 enrichment experiments.

4.6. Results

The ensemble is initialised as a 255-member latin hypercube across the prior parameter ranges listed in Table 2. As expected, the climatologies of the samples provide a very poor match to the "truth" model. However, we can see from Figure 3 that the final marginal parameter distributions all include the true values and are generally more precise (lower spread) than the initial guess. Several of the marginal distributions are constrained to values markedly closer to the true parameter value, and none are significantly worsened. We can check that the posterior ensemble includes the truth by calculating the chi-square statistic based on the Mahalanobis distance $(\mathbf{x}' - \bar{\mathbf{x}})^T C^{-1} (\mathbf{x}' - \bar{\mathbf{x}})$ where $\mathbf{x}'$ is the vector of true parameters, $\bar{\mathbf{x}}$ is the ensemble mean and C is the covariance matrix of the ensemble. Essentially, we are checking whether the truth can be considered as a member of the ensemble. This statistic remains well below the 5% significance level for the posterior ensemble, indicating that even though the ensemble has narrowed considerably in the multidimensional parameter space, it still contains the correct answer. The fit to the data for the posterior ensemble members (as indicated by the cost function) is also substantially improved, with them being generally comparable to or better than the best members of the prior sample. Therefore, although we do not have an analytical solution to compare with in distribution, the method does appear to have worked well. A number of alternate tests, with slightly different parameter sets and observational constraints, also generated similarly good results (not shown here).

However, when we tried to estimate as many as 20 uncertain parameters, the experiments failed through ensemble collapse (filter divergence), with the chi-square test strongly rejecting the hypothesis that the ensemble contained the truth. Thus, this method is still limited to problems of moderate dimensionality, and we do not claim to have eliminated the general problem described by Bengtsson et al. (2008). However, our iterative approach has helped to push the boundary of which problems can be reasonably attempted.

Although the residual uncertainty in the posterior estimates of some parameter values seems substantial, all parameters exhibit several significant pairwise correlations with other parameters, shown in Table 3. Thus, although many of the parameters cannot be individually identified with high precision, the posterior is constrained to a relatively small region of the multivariate parameter space where the resulting model behaviour is reasonable.

	2	3	4	5	6	7	8	9	10	11	12	TCR
1	0.01	**0.30**	**0.36**	**0.31**	**-0.24**	0.10	-0.12	**-0.32**	0.08	**-0.17**	0.09	-0.14
2		**-0.26**	0.04	**0.28**	-0.11	**0.26**	-0.01	-0.00	**0.21**	-0.13	0.13	-0.11
3			-0.10	**0.16**	0.05	**0.24**	**0.22**	-0.01	0.10	-0.05	-0.13	-0.08
4				-0.08	**-0.42**	0.00	**0.20**	**-0.17**	0.06	0.01	**0.24**	**-0.28**
5					**-0.16**	**0.19**	-0.08	0.06	**0.19**	**-0.19**	0.09	-0.12
6						**-0.19**	0.02	**-0.17**	**-0.17**	-0.01	-0.02	**0.29**
7							0.06	-0.10	-0.02	0.05	**-0.20**	-0.12
8								**-0.36**	0.00	-0.13	0.14	0.00
9									**0.51**	0.05	0.07	-0.08
10										**-0.28**	**0.21**	**-0.23**
11											**-0.42**	0.14
12												**-0.18**

Table 3: Pairwise correlations of parameters with each other and also with the transient climate response TCR. Parameter ordering is as for Table 2. Values that are significant at the 1% level are indicated in bold.

The transient warming for the prior and posterior ensembles are presented in Figure 4. The prior ensemble has a fairly broad spread in transient climate response (TCR: warming observed after 70 years of 1% pa CO_2 enrichment) with a 5–95% range of of 1.91–2.62C, even though the equilibrium sensitivity is essentially fixed at close to 2.9C for all samples. However, the posterior ensemble range of TCR is reduced by a factor of more than 3 compared to the prior, with the range of 2.13–2.36C clustered tightly around the true value of 2.21C. The 5–95% range of effective heat uptake efficiency κ of the ocean is 0.47–0.85 $Wm^{-2}K^{-1}$ in the prior, narrowing substantially to 0.57–0.69 in the posterior. The true value here is 0.64 $Wm^{-2}K^{-1}$.

Our ensembles reveal some interesting relationships between the ocean state and the ocean heat uptake. The dominant relationship, which we might expect on direct physical grounds, is that there is a strong correlation in the prior of around 0.85 between the stratification of the ocean (as measured here by the difference between surface and mean ocean temperature) and the TCR, and an equally strong (but negative) correlation between stratification and κ. This is perhaps not surprising since one would expect stratification to be strongly linked to mixing (at least if confounding factors such as deep water production do not vary too much). The relationship is weakened in the posterior (although still highly significant), perhaps because the range of outputs spanned by the ensemble is greatly reduced and thus the 'noise' of unrelated factors can play a larger role. There is also a negative correlation between the oxygen concentration in the ocean surface layers and the TCR, predominantly due to the direct solubility effect of the warmer (colder) ocean surface associated with weaker (stronger) mixing. The correlations between individual parameters and the TCR (also shown in Table 3) show that all of the biological parameters are correlated with various physical parameters, and two of them are directly correlated with the transient response. None of the correlations with the TCR reach a value of 0.3, so the overall narrowing in response is not directly controlled by any single parameter but instead emerges as a property of the climate system as a whole.

These encouraging results suggests that the climatological state of the ocean as determined by both biological and physical tracer distributions may be a useful constraint on transient ocean heat uptake, although more work is undoubtedly required in order to translate this idealised test into to robust practical results.

4.7. Discussion

Although our identical twin experiment precludes detailed quantitative analysis, our results exhibit interesting contrasts with previous model-based analyses of ocean heat uptake. Sokolov et al. (2009) did not explicitly present an ocean heat uptake efficiency for their results, however their posterior estimate of effective diffusivity assigns high probability to values that are very low compared to values obtained for modern GCMs. This implies that their pdf for ocean heat uptake efficiency would include values rather lower than

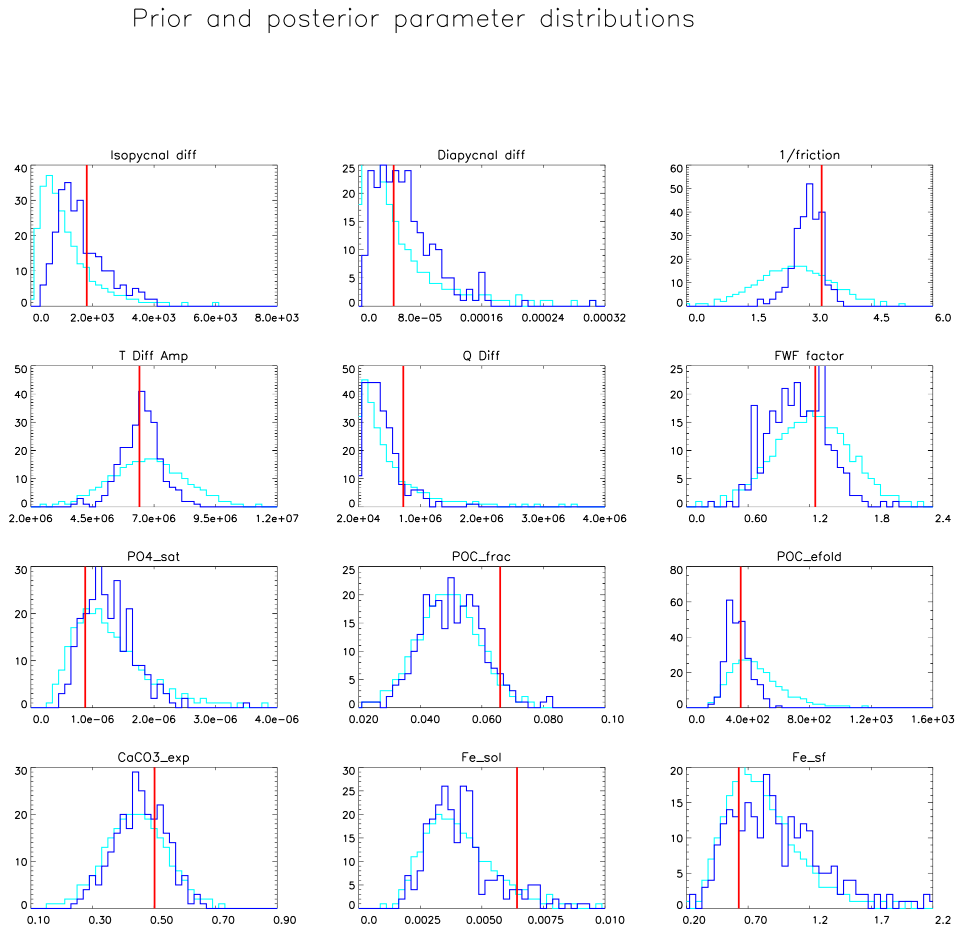

Figure 3: Prior and posterior distributions for the 12-parameter experiment described in the text. True parameter values are indicated by the vertical lines. Prior is cyan histogram and posterior is dark blue.

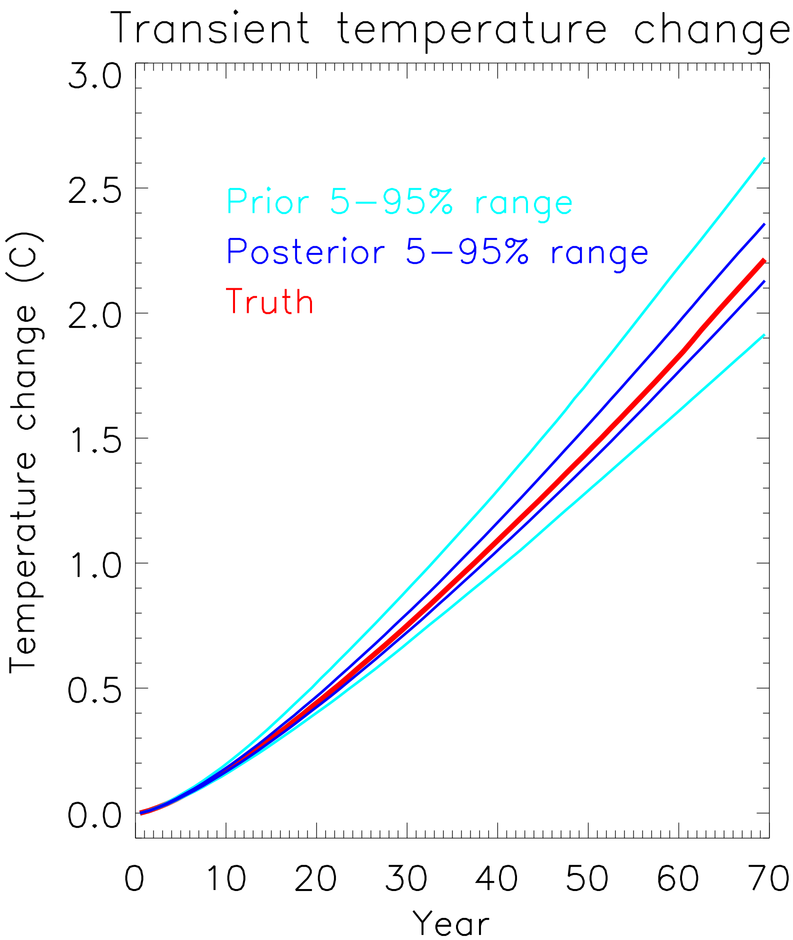

Figure 4: Results of a 70 year 1% per annum CO_2 enrichment experiment, showing global mean surface temperature anomaly. Prior and posterior 5-95% ranges are indicated in cyan and dark blue respectively. Output of the truth run is shown in red.

those provided by GCM projections. Collins et al. (2007), however, found that the parameter perturbations they made in the HadCM3 model only resulted in modest changes to the transient response. There are several possible interpretations of these results. The Sokolov et al. (2009) results may have an exaggerated range of uncertainty due to their choice of very broad prior and little data to constrain the result. In particular, they admit that both the extremely high and low values of ocean mixing parameter that they allow in their prior cannot support the observed global meridional overturning, but they did not use this information in their probabilistic analysis. The only data that they used which directly relate to the ocean is the observed ocean warming, which is known to provide only a rather weak constraint on mixing (Lindzen, 2002).

Conversely, the parameter perturbations in the HadCM3 model may have been too small to fully represent the uncertainty in their true values. Furthermore, these perturbations were applied individually, and it seems inevitable that multivariate perturbations across the same ranges would have generated a wider spread of results. It is therefore encouraging to see that our prior ensemble covers such a wide range of responses, implying that our model is fundamentally capable of simulating both very high and low overall mixing rates, with our prior 90% range of ocean heat uptake efficiency (0.47–0.85) being broader than the full range obtained from modern ocean GCMs of around 0.6–0.8 (Raper et al., 2002), let alone the even more restricted range of 0.55–0.74 obtained by Collins et al. (2007). Thus, there does not appear to be anything inherent to the model structure that artificially restricts the range of mixing rates. We emphasise that the use of a fixed atmospheric feedback (equilibrium sensitivity) in our experiments does limit the range of transient climate response, so our results cannot be directly interpreted in terms of future climate change. Nevertheless, we see that even though individual parameters are not all tightly constrained, the tracer distributions have provided a highly effective constraint on the overall ocean heat uptake. This result suggests that a practical application with real climate data could provide a significant improvement on recent predictions of climate change. We also plan in the future to consider transient simulations with realistic boundary conditions for modern anthropogenic tracers such as CFCs and radiocarbon from nuclear bomb tests. It is likely that such data will also prove to be valuable in constraining the dynamical behaviour of the ocean, as they directly relate to the penetration of a surface influence over the multidecadal time scale. However, the current implementation of the parameter estimation method is limited to equilibrium simulations.

5. Conclusions

We have presented a simple but effective method for parameter estimation in moderately high dimensional problems, based on an iterative importance sampling approach. The method presented here shows a clear improvement for nonlinear applications, compared to the ensemble Kalman filtering method which has been previously used. In (near-)linear problems, both methods generate good results, and the EnKF is more efficient in computational terms. However, in more strongly nonlinear applications, the importance sampling method is substantially more accurate. The method appears to generalise to problems of moderate dimensionality, as typically encountered in climate science, where direct sampling is computationally prohibitive. The combination of our efficient method together with a reasonably realistic ocean model allows us to use physical and biogechemical tracer data to constrain the dynamics of the ocean circulation for the first time. These data limit the model to a relatively small part of the multivariate parameter space which strongly constrains the transient climate response. It therefore appears that observations of climatological tracer distributions in the ocean are informative about its role in the rate of global warming via heat uptake.

6. Acknowledgments

We are grateful to Andy Ridgwell for help and advice concerning the bigeochemical modelling, and two reviewers for many helpful suggestions. This work was supported by Innovative Program of Climate Change Projection for the 21st Century of the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

Annan, J. D., Hargreaves, J. C., 2007. Efficient estimation and ensemble generation in climate modelling. Philosophical Transactions of the Royal Society A 365 (1857), 2077–2088.

Annan, J. D., Hargreaves, J. C., Edwards, N. R., Marsh, R., 2005a. Parameter estimation in an intermediate complexity Earth System Model using an ensemble Kalman filter. Ocean Modelling 8 (1–2), 135–154.

Annan, J. D., Lunt, D. J., Hargreaves, J. C., Valdes, P. J., 2005b. Parameter estimation in an atmospheric GCM. Nonlinear processes in geophysics 12, 363–371.

Baker, A., Jickells, T., 2006. Mineral particle size as a control on aerosol iron solubility. Geophys. Res. Lett 33, 17.

Bengtsson, T., Bickel, P., Li, B., 2008. 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. Vol. 2. Institute of Mathematical Statistics, pp. 316–334.

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, Chichester, UK.

Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M., Weber, S., Alcamo, J., Alexeev, V., Berger, A., et al., 2002. Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. Climate Dynamics 18 (7), 579–586.

Collins, M., Brierley, C., MacVean, M., Booth, B., Harris, G., 2007. The sensitivity of the rate of transient climate change to ocean physics perturbations. Journal of Climate 20 (10), 2315–2320.

Conkright, M. E., Locarnini, R. A., Garcia, H. E., OBrien, T. D., Boyer, T. P., Stephens, C., Antonov, J. I., 2002. World Ocean Atlas 2001: Objective Analyses, Data, Statistics, and Figures. CD-ROM Documentation, National Oceanographic Data Center, Silver Spring, MD.

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. JR Statist Soc. B 68, 411–436.

Doney, S., Lindsay, K., Fung, I., John, J., 2006. Natural variability in a stable, 1000 year global coupled climate-carbon cycle simulation. J. Climate 19, 3033–3054.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing 10 (3), 197–208.

Edwards, N. R., Marsh, R., 2005. Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model. Climate Dynamics 24 (4), 415–433.

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

Forest, C., Stone, P., Sokolov, A., 2006. Estimated PDFs of climate system properties including natural and anthropogenic forcings. Geophys. Res. Lett 33.

Gelman, A., Meng, X., 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science, 163–185.

Gregory, J., Mitchell, J., 1997. The climate response to CO_2 of the Hadley Centre coupled AOGCM with and without flux correction. Geophys. Res. Lett 24 (15), 1943–1946.

Hammersley, J. M., Handscomb, D. C., 1964. Monte Carlo methods. Methuen & Co Ltd, London.

Hansen, J., Russel, G., Lacis, A., Fung, I., Rind, D., 1985. Climate response times: Dependence on climate sensitivity and ocean mixing. Science 229, 857–859.

Hargreaves, J. C., Annan, J. D., Edwards, N. R., Marsh, R., 2004. Climate forecasting using an intermediate complexity Earth System Model and the ensemble Kalman filter. Climate Dynamics 23 (7–8), 745–760.

Hoffert, M., Callegari, A., Hsieh, C., 1985. The role of deep sea heat storage in the secular response to climatic forcing. Journal of Geophysical Research-Oceans 85 (C11).

Huang, R., 1999. Mixing and Energetics of the Oceanic Thermohaline Circulation. Journal of Physical Oceanography 29 (4), 727–746.

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

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al., 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77 (3), 437–471.

Key, R., Kozyr, A., Sabine, C., Lee, K., Wanninkhof, R., Bullister, J., Feely, R., Millero, F., Mordy, C., Peng, T., 2004. A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP). Global Biogeochem. Cycles 18 (4).

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.

Knutti, R., Tomassini, L., 2008. Constraints on the transient climate response from observed global temperature and ocean heat uptake. Geophys. Res. Lett. 35, L09701.

Lenton, T. M., Marsh, R., Price, A. R., Lunt, D. J., Aksenov, Y., Annan, J. D., Cooper-Chadwick, T., Cox, S. J., Edwards, N. R., Goswami, S., Hargreaves, J. C., Harris, P. P., Jiao, Z., Livina, V. N., Payne, A. J., Rutt, I. C., Shepherd, J. G., Valdes, P. J., Williams, G., Williamson, M. S., , Yool, A., 2007. A modular, scalable, Grid ENabled Integrated Earth system modelling (GENIE) framework: Effects of atmospheric dynamics and ocean resolution on bi-stability of the thermohaline circulation. Climate Dynamics 29 (6), 591–613.

Li, Y., Peng, S., Broecker, W., Oestlund, H., 1984. The average vertical mixing coefficient for the oceanic thermocline. Tellus. Series B, Chemical and physical meteorology 36 (3), 212–217.

Lindzen, R., 2002. Do deep ocean temperature records verify models? Geophysical Research Letters 29 (8), 95–1.

Mahowald, N., Kohfeld, K., Hannson, M., Balkanski, Y., Harrison, S., Prentice, I., Schulz, M., Rodhe, H., 1999. Dust sources during the last glacial maximum and current climate: a comparison of model results with paleodata from ice cores and marine sediments. J. Geophys. Res 104, 15895–15916.

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.

Meehl, G. A., Stocker, T. F., Collins, W. D., et al., 2007. 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. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., Ch. 10, pp. 747–845.

Meng, X., Wong, W., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica 6, 831–860.

Munk, W., 1966. Abyssal recipes. Deep-Sea Res 13, 707–730.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., Stainforth, D. A., 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430, 768–772.

Navon, I. M., 1998. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. Dynamics of Atmospheres and Oceans 27 (1–4), 55–79.

Neal, R., of Computer Science, D., of Toronto, U., 1993. Probabilistic inference using Markov chain Monte Carlo methods. Department of Computer Science, University of Toronto.

Oliver, K., Edwards, N., 2008. Location of potential energy sources and the export of dense water from the Atlantic Ocean. Geophysical Research Letters 35 (22), L22604.

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

Parekh, P., Follows, M., Boyle, E., 2005. Decoupling of iron and phosphate in the global ocean. Global Biogeochem. Cycles 19.

Parekh, P., Follows, M., Dutkiewicz, S., Ito, T., 2006. Physical and biological regulation of the soft tissue carbon pump. Paleoceanography 21 (3), 1–A3001.

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.

Raper, S., Gregory, J., Stouffer, R., 2002. The role of climate sensitivity and ocean heat uptake on AOGCM transient temperature response. Journal of Climate 15 (1), 124–130.

Ridgwell, A., 2001. Glacial-interglacial perturbations in the global carbon cycle. Ph.D. thesis, University of East Anglia.

Ridgwell, A., Hargreaves, J. C., Edwards, N. R., Annan, J. D., Lenton, T. M., Marsh, R., Yool, A., Watson, A., 2007. Marine geochemical data assimilation in an efficient earth system model of global biogeochemical cycling. Biogeosciences 4 (1), 87–104.

Ridgwell, A., Singarayer, J., Hetherington, A., Valdes, P., 2009. Tackling Regional Climate Change By Leaf Albedo Bio-geoengineering. Current Biology 19 (2), 146–150.

Rougier, J. C., 2007. Probabilistic inference for future climate using an ensemble of simulator evaluations. Climatic Change 81, 247–264.

Singarayer, J., Richards, D., Ridgwell, A., Valdes, P., Austin, W., Beck, J., 2008. An oceanic origin for the increase of atmospheric radiocarbon during the Younger Dryas. Geophysical Research Letters 35 (14), L14707.

Sokolov, A., Stone, P., Forest, C., Prinn, R., Sarofim, M., Webster, M., Paltsev, S., Schlosser, C., Kicklighter, D., Dutkiewicz, S., et al., 2009. Probabilistic Forecast for 21st Century Climate Based on Uncertainties in Emissions (without Policy) and Climate Parameters. Journal of Climate 22, 5175–5204.

West, M., 1993. Approximating Posterior Distributions by Mixture. Journal of the Royal Statistical Society. Series B (Methodological) 55 (2), 409–422.

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