

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

4 J. D. Annan and J. C. Hargreaves

5 *RIGC/JAMSTEC, Yokohama, Japan*

---

6 **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.

7 *Key words:* parameter estimation, EMIC, SIR

---

8 **1. Introduction**

9 Climate models are one of the primary tools through which predictions  
10 of climate change can be made (Meehl et al., 2007). However, the model  
11 results can be highly dependent on the values of model parameters which  
12 are not adequately constrained either by direct process-based observations  
13 or by theoretical arguments, and therefore can only be estimated by the  
14 inverse process of comparing the model output to observations of the real  
15 world. Such calibration of models to observational data remains a significant  
16 challenge in climate science, primarily due to the vast computational

17 challenge it poses. Therefore, a range of approaches have been developed for  
18 more efficient parameter estimation in climate science in recent years (An-  
19 nan and Hargreaves, 2007). One such approach is the ensemble Kalman filter  
20 (EnKF; Kalman (1960); Evensen (2003)), which has been used for multivari-  
21 ate parameter estimation in climate models (Annan et al., 2005a). While  
22 even efficient ensemble methods such as this cannot easily be applied to the  
23 largest numerical models due to the computational costs, the development  
24 of such methods ensures that we can make effective use of Earth system  
25 models of intermediate complexity (EMICs, Claussen et al., 2002).

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

## 40 **2. An Iterative Importance Sampling method for parameter esti-** 41 **mation**

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

49 The direct Monte Carlo approach based on rejection sampling (Hammer-  
50 sley and Handscomb, 1964) is a simple and popular method which has been  
51 widely used in climate science in recent years (eg Knutti et al., 2002). In this  
52 approach, we draw samples from the prior  $p(\mathbf{x})$  and assign each one a relative

53 probability or weight defined by  $w(\mathbf{x}) \equiv p(\mathbf{o}|\mathbf{x})$ . This approach is often very  
 54 expensive. In particular, the vast majority of samples may be given negligi-  
 55 ble weight if the prior is substantially more diffuse than the posterior, and  
 56 in this case it may take a very large number of samples (each one of which  
 57 requires a model integration to evaluate the likelihood function) to populate  
 58 the posterior and achieve reasonable convergence in distribution. While this  
 59 problem is particularly severe in high dimensional problems where the en-  
 60 semble is liable to collapse to a single sample (Bengtsson et al., 2008), such  
 61 particle-based methods may still require uncomfortably large ensembles in  
 62 even problems of moderate dimension.

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

73 The method of “bridging densities” has been proposed as a means of in-  
 74 creasing the efficiency of Monte Carlo sampling in such situations (Meng and  
 75 Wong, 1996; Gelman and Meng, 1998; Del Moral et al., 2006). The basic  
 76 principle is that given an initial proposal that is some way distant from the  
 77 prior, it may be more efficient to define some intermediate “bridging” distri-  
 78 bution such that we can use the initial proposal to generate samples from the  
 79 bridging distribution, and then use the bridging distribution as a proposal  
 80 from which we generate samples from the posterior. For a suitably chosen  
 81 bridging density, this can be substantially more efficient than attempting  
 82 to directly generate the posterior by sampling from the proposal. The ap-  
 83 proach generalises directly to a larger number of bridges, or even an infinite  
 84 sequence (Neal et al., 1993; Gelman and Meng, 1998).

One natural approach is to consider the geometric family

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

85 which transforms smoothly from  $g$  to  $f$  as  $\alpha$  varies from 0 to 1. Even if it is  
 86 very inefficient to use  $g$  directly as a proposal density for  $f$ , if we select an in-  
 87 creasing sequence of closely-spaced  $\alpha_i$  we can iteratively use  $\phi_{\alpha_i}$  as a proposal

88 for  $\phi_{\alpha_{i+1}}$  and ultimately reach (or at least approach in the case of an infinite  
89 series) the target distribution  $f$ . The choice of  $g$  here may be arbitrary, but  
90 in the examples presented below we use the prior for convenience.

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

111 We have tested the approach of using bridging distributions with jitter  
112 compensated by shrinkage, but although it works well in very low dimensional  
113 problems we have found it difficult to ensure that the ensemble converges to  
114 the correct solution for more than about 3–4 parameters, with tolerable en-  
115 semble sizes. The specific difficulty we have encountered manifests itself as  
116 an over-rapid collapse of the ensemble to a narrow region of parameter space,  
117 sometimes referred to as “filter divergence”. The bridging distributions as  
118 presented above are sequentially nested and it is difficult for a distribution  
119 which is inappropriately over-narrow to recover the correct spread, since the  
120 addition of jitter (the only step whereby it can expand) is immediately coun-  
121 teracted by the shrinkage step. Therefore, we now present a minor variation  
122 of iterated Importance Sampling (IIS) which we have found to work better  
123 in our applications. Instead of using an explicit shrinkage step which is fol-  
124 lowed by importance sampling to a narrower distribution, we simply perform  
125 the importance sampling directly on the jittered ensemble, but change the

126 weighting function to account for the extra spread generated by the jitter.  
 127 As with the standard shrinkage procedure, this approach is only precisely  
 128 correct in the case of a linear Gaussian problem. However, the solutions  
 129 it generates are substantially more accurate than the EnKF approach for  
 130 the nonlinear problems we have tested, and in contrast to the conventional  
 131 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

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

132 for some  $\alpha_i$  and  $\beta_i$  (which in contrast to the established approach, are not  
 133 necessarily equal here), we firstly use this as a proposal for  $g^{1-\beta_i} f^{\alpha_i+\epsilon}$  by  
 134 reweighting the samples according to  $f^\epsilon$ , where  $\epsilon$  is a tunable parameter which  
 135 we typically set to 0.05 unless otherwise stated. Resampling with the addition  
 136 of jitter (with the jitter drawn from a Gaussian kernel fitted to the ensemble  
 137 with its variance scaled by a factor of  $\epsilon$ ) will, at least in the case where  
 138 the ensemble truly is a multivariate Gaussian, generate an ensemble which  
 139 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  $1-\beta_{i+1} = \frac{1-\beta_i}{1+\epsilon}$   
 140 respectively, this ensemble now serves as the proposal for the next iteration.  
 141 It is easily seen that over repeated applications of these steps, the sampling  
 142 distribution converges to  $g^0 f^1 = f$  as desired. Several applications below also  
 143 demonstrate the correctness of this approach. We note that the repeated use  
 144 of (a scaled version of) the likelihood function, balanced by expansion of the  
 145 ensemble around its mean, is fundamentally the same approach as previously  
 146 adopted using the ensemble Kalman filter (Annan et al., 2005b), with the  
 147 jitter here taking the place of the variance inflation step in the previous  
 148 approach, and the weighting according to the likelihood function taking the  
 149 place of the analysis step. The main difference here is that the data here enter  
 150 the process through weighting according to the likelihood function, rather  
 151 than using the Kalman equations to interpolate (or extrapolate) according  
 152 to the covariance matrix. Thus, while our new method generally requires a  
 153 somewhat larger ensemble to ensure adequate sampling, it has the benefit  
 154 of not relying so strongly on the distribution being approximately Gaussian,  
 155 and we shall demonstrate the benefit of this in some applications.

156 We mention in passing that there is an important difference between our  
 157 approach and the iterative resampling approach of West (1993), in that we  
 158 are *not* attempting to sample the true posterior  $f$  at each stage in our iterative

159 sequence. Thus, we expect our approach to be substantially less efficient in  
160 the cases where we already have a reasonable proposal distribution (including  
161 those cases where the prior is not much broader than the posterior and thus  
162 can serve as the proposal distribution). However, in many cases of interest to  
163 climate scientists, we have no reasonable proposal density and, as mentioned  
164 above, a direct attempt to construct the posterior by rejection sampling from  
165 the prior is likely to fail through an immediate collapse of the sample.

### 166 **3. Application to idealised problems**

#### 167 *3.1. Univariate problem*

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

We use one uncertain input  $x$ , a model given by

$$y = x^2$$

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

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

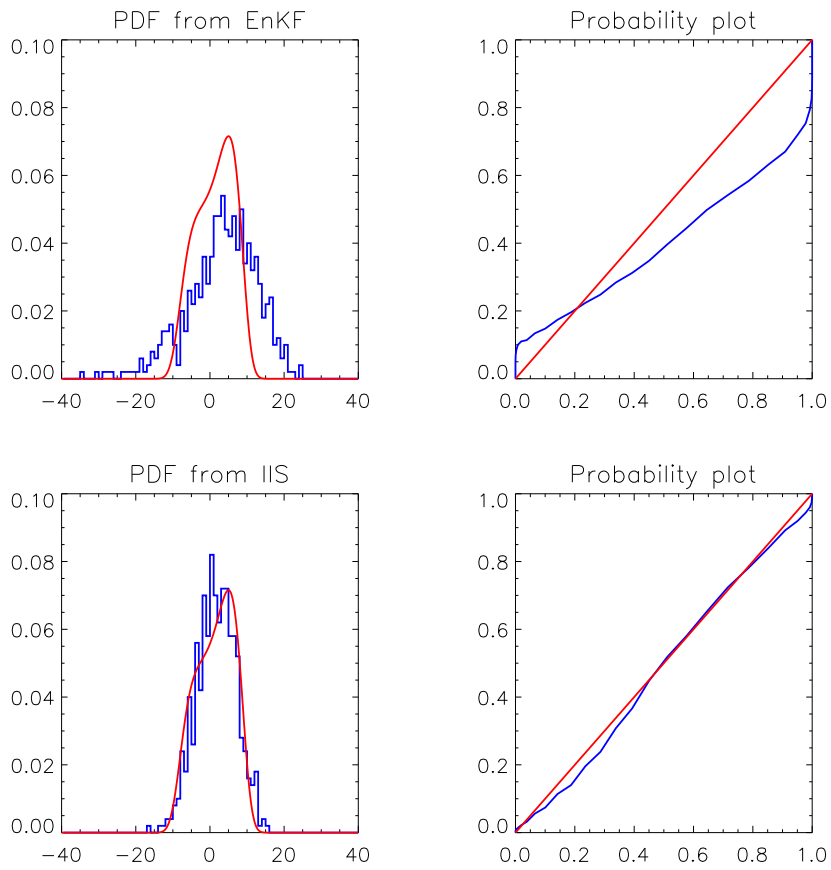


Figure 1: Comparing the performance of IIS and EnKF on a nonlinear problem. Red lines show the correct solution, blue show the experimental results. The top plots show a 500-member EnKF result, and the lower plots show the IIS result with the same size ensemble.

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

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

$N$	$p$	K-S test		Kuiper	
		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).

228 *3.2. High dimensional linear problem*

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

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

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

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

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

247 For this more computationally challenging problem, the choice of the  
 248 scaling factor  $\epsilon$  in the iterative procedure can affect the performance of the  
 249 algorithm. For a very large value, the ensemble collapses rather rapidly and

250 may converge to a incorrect solution. This is due to the curse of dimension-  
251 ality: if the prior sample is widely dispersed compared to the posterior, then  
252 the posterior weight will be concentrated on very few members and even the  
253 addition of jitter may not be enough to rescue the situation. Conversely, if  
254 the scaling factor is very large, then the weights will remain nearly uniform  
255 and the ensemble will take many iterations to converge to the true posterior.  
256 A reasonable rule of thumb arising from our experiments is to aim for a ef-  
257 fective ensemble size that is between 50% and 90% of the actual ensemble  
258 size, and so in the results presented here the value of  $\epsilon$  has been adaptively  
259 tuned to stay within these bounds.

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

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

#### 283 4. Application to a 3D EMIC

284 We now perform an identical twin experiment to demonstrate the applica-  
285 tion of the method to an earth system model of intermediate complexity, the

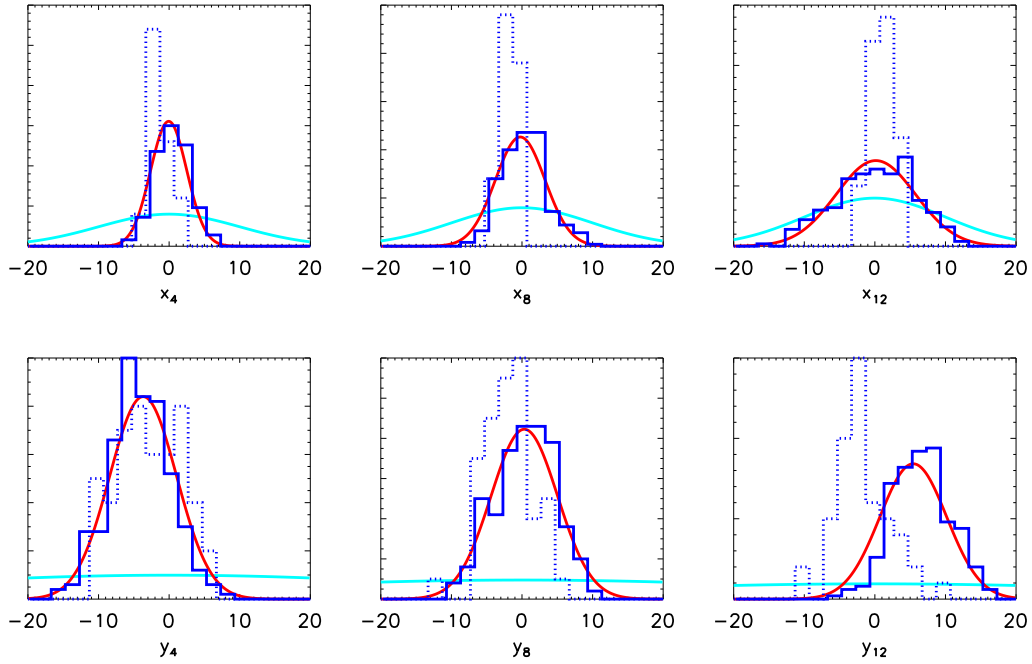


Figure 2: Testing the IIS on a 16-dimensional problem. The top row of plots show 3 of the input parameters, the lower row show 3 outputs. The cyan curves show the prior (only shown to  $\pm 20$ ), red indicates the true posterior, dark blue solid histogram shows results from a 250-member IIS calculation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

286 Grid ENabled Integrated Earth system model (GENIE: [www.genie.ac.uk](http://www.genie.ac.uk)) (Lenton  
287 et al., 2007), which is based on the fast climate model of Edwards and Marsh  
288 (2005). Previously, we have used the EnKF methodology for estimating phys-  
289 ical parameters (Hargreaves et al., 2004) or biological parameters (Ridgwell  
290 et al., 2007) separately in this model. Here we demonstrate simultaneous  
291 estimation of physical and biological parameters, using a variety of tracer  
292 data.

#### 293 *4.1. Scientific Motivation*

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

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

313 More recently, explicitly probabilistic analyses of ocean mixing have been  
314 performed by comparing ‘perturbed parameter’ ensembles of model simula-  
315 tions to observational estimates of warming over the 20th century (Knutti  
316 et al., 2002; Forest et al., 2006). Due to computational limitations, these  
317 analyses have generally been restricted to the use of greatly simplified mod-  
318 els where the ocean dynamics are limited, and mixing into the deep ocean  
319 is primarily determined by a single global vertical diffusion parameter. It  
320 is not straightforward to directly equate these parameters to those used in  
321 more complex ocean GCMs, since these latter models often include a range of  
322 mixing processes (including convection and wind stirring near the surface),

323 and the diffusion models may also incorporate spatial patterns of variable  
324 mixing. However, one striking, and perhaps worrying, aspect of the prob-  
325 abilistic analyses is that they have often assigned fairly high probability to  
326 values of global ocean mixing that are substantially lower than those com-  
327 monly obtained in GCM simulations, with strong implications for projections  
328 of climate change (Knutti and Tomassini, 2008; Sokolov et al., 2009).

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

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

#### 345 *4.2. Model*

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

358 A coupled marine biogeochemistry module based on Ridgwell et al. (2007)  
359 calculates the redistribution of tracer concentrations due to processes other

360 than transport by the circulation of the ocean, namely: air-sea gas exchange,  
361 the removal of nutrients, carbon, and alkalinity from solution as a result of  
362 biological activity in the sunlit surface ocean layer, the vertical export of  
363 particulate matter and its remineralization in the ocean interior, and the  
364 remineralization of dissolved organic matter and associated consumption of  
365 dissolved oxygen.

366 We employ a seasonal scheme for biologically-induced export out of the  
367 surface ocean based on a dual nutrient limitation of productivity by  $\text{PO}_4^{3-}$   
368 and dissolved iron ( $[\text{Fe}]$ ) derived from previously published schemes (Doney  
369 et al., 2006; Parekh et al., 2005, 2006; Ridgwell, 2001)). This differs from  
370 that described by Ridgwell et al. (2007) where it was used for an EnKF-based  
371 assimilation of marine observations, in the following ways:

- 372 1. A co-limitation of total dissolved iron on export production added, using  
373 the law of the minimum following Ridgwell (2001) and assuming a half-  
374 saturation constant for iron of  $0.1 \text{ nmol kg}^{-1}$ .
- 375 2. The effects of sub-optimal ambient light levels is implemented follow-  
376 ing Doney et al. (2006), using incident the shortwave radiation inci-  
377 dent at the ocean surface calculated by the climate model (Edwards and  
378 Marsh, 2005) and assuming a half saturation value for light of  $20 \text{ Wm}^{-2}$ .  
379 We have added a marine iron cycle based on Parekh et al. (2005, 2006),  
380 but deviating as follows:
  - 381 (a) We link the phosphate and iron cycles via an organic matter Fe:C  
382 Redfield ratio that is a function of dissolved iron availability, taking  
383 the average of the two (diatom, and non-diatom) parameterizations  
384 of Ridgwell (2001).
  - 385 (b) For iron inputs to the ocean we take the atmospheric tracer trans-  
386 port model generated dust field of Mahowald et al. (1999), and  
387 uniform iron content in dust of 3.5 wt%. However, we depart from  
388 the common assumption regarding a uniform solubility of iron in  
389 dust and instead allow solubility to vary inversely to dust loading  
390 consistent with laboratory experiments and observations (Ridgwell,  
391 2001) and with a solubility that scales inversely to the square root  
392 of dust loading (flux) (Baker and Jickells, 2006).

393 In addition to several parameters controlling aspects of the ocean carbon  
394 cycle (and hence dissolved  $\text{PO}_4$ , ALK, and  $\text{O}_2$  distributions) that we allowed  
395 to vary in previous EnKF-based assimilation work (Ridgwell et al., 2007),

396 we now include the scavenging rate of dissolved iron from the water column,  
397 and the overall (global mean) solubility of iron in dust.

### 398 *4.3. Data*

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

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

410 In our previous marine biogeochemistry data assimilation experiment (Ridg-  
411 well et al., 2007) we utilized observed 3D distributions of phosphate ( $\text{PO}_4$ ) (Conkright  
412 et al., 2002) and alkalinity (Key et al., 2004) in the ocean, to constrain model  
413 parameters controlling the marine carbon cycle. In this, observed fields of  
414  $\text{PO}_4$  help constrain the rates and distribution of  $\text{PO}_4$  uptake at the ocean  
415 surface, together with the penetration depth of particulate organic matter  
416 before remineralization and release of  $\text{PO}_4$  back to the ocean. Alkalinity  
417 (ALK) distributions place constraints on the production and dissolution of  
418 the calcium carbonate ( $\text{CaCO}_3$ ) mineral shells and (skeletons) in the ocean.  
419 The distribution of both these tracers is affected by ocean circulation. In this  
420 study we add a further 3D field of dissolved oxygen ( $\text{O}_2$ ) (Conkright et al.,  
421 2002). This is controlled not only by the remineralization of organic mat-  
422 ter and hence bacterial consumption of oxygen in the ocean interior as well  
423 as ocean circulation, but is also sensitive to ocean surface temperature and  
424 residence time as  $\text{O}_2$  is rather more soluble in colder waters and will reach  
425 equilibrium with the atmosphere only in relatively stratified conditions. We  
426 do not consider observational uncertainties in these tests.

### 427 *4.4. Parameters*

428 The physical and biological parameters we chose to vary are listed in  
429 Table 2, along with their prior 2.5–97.5% ranges. The physical parameters  
430 that we vary (shown in Table 2) are the subset of those used, and described  
431 in more detail, in previous work (Annan et al., 2005a), which were found

432 to be most influential on model behaviour. For the atmospheric physics,  
433 “Q” and “T” here refer to moisture and heat respectively. The fresh water  
434 flux adjustment (FWF) from Atlantic to Pacific, a standard procedure in  
435 EMBM-type models, is implemented here as a scaling factor on the standard  
436 0.32Sv figure of Oort (1983) rather than as an absolute value. Although  
437 presented here as an atmospheric parameter, this flux acts directly on the  
438 ocean where it strongly influences the meridional overturning circulation.  
439 The prior distributions were defined as Gaussian either in the variable or its  
440 log (for those parameters where a skewed distribution with a 50th percentile  
441 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 ( $\text{m}^2\text{s}^{-1}$ )	250	4000	615	3700	1815
2	log Diapycnal diffusion / $10^5$ ( $\text{m}^2\text{s}^{-1}$ )	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 / $10^6$ ( $\text{m}^2\text{s}^{-1}$ )	3.82	9.90	4.85	8.14	6.41
5	log Q diffusion / $10^5$ ( $\text{m}^2\text{s}^{-1}$ )	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 $\text{PO}_4$ half-saturation $\times 10^6$ ( $\mu\text{mol kg}^{-1}$ )	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 $\text{CaCO}_3$ 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’.

#### 442 4.5. Experimental Details

443 In order to validate the method and investigate the identifiability of the  
444 parameters and physical behaviour of the model, we present the results from  
445 identical twin tests here. In this case, a truth run was selected that had  
446 a reasonably realistic overall physical and biogeochemical state from a 256-  
447 member latin hypercube ensemble (McKay et al., 1979). As in previous

448 experiments, the physical observations we used consisted of climatological  
449 observations of three-dimensional ocean temperature and salinity, and the  
450 two-dimensional field of atmospheric temperature and relative humidity. For  
451 the ocean biogeochemical model, we use 3D fields of alkalinity, oxygen and  
452 phosphate.

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

476 Even though this model is computationally cheap, it would still be chal-  
477 lenging to integrate it for its full equilibration time scale of  $O(2000)$  years at  
478 each iteration. Thus we rely on the observation that adding modest amounts  
479 of jitter to the model parameters does not greatly upset the quasi-equilibrium  
480 balance of the model state, so that only a more moderate period of integra-  
481 tion (we use 200y here) is required to restore a near-equilibrium state. We  
482 checked the validity of this approximation by integrating the final ensemble  
483 on for a further 5000 years, and found that the changes were indeed very  
484 minor across the ensemble as a whole. Thus the 30 iterations of the method  
485 that we performed requires 6000y of integration time, which is only a small

486 multiple of the spin-up time of the model itself. This behaviour is comparable  
487 to what was previously found for the EnKF applied to the same model (An-  
488 nan et al., 2005a). A possible improvement for future applications would be  
489 to perturb the full state according to the covariance matrix, rather than only  
490 adjusting the parameters in isolation.

491 Gregory and Mitchell (1997) defined the ‘ocean heat uptake efficiency’  
492  $\kappa = \frac{\Delta F}{\Delta T}$  to be the heat uptake flux to the deep ocean  $\Delta F$  divided by the  
493 surface temperature anomaly  $\Delta T$ . Although this is not a fixed parameter  
494 of the climate system, it is a reliable diagnostic over a period of strongly  
495 increasing forcing such as idealised 1% pa CO<sub>2</sub> enrichment experiments or  
496 more realistic socioeconomic emissions scenarios. In order to provide a direct  
497 comparison with the  $\kappa$  values calculated by Collins et al. (2007) for their  
498 ensemble of HadCM3 results, and also by Raper et al. (2002) for the CMIP3  
499 ensemble, we also perform 1% pa CO<sub>2</sub> enrichment experiments.

#### 500 4.6. Results

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

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

530 Although the residual uncertainty in the posterior estimates of some pa-  
531 rameter values seems substantial, all parameters exhibit several significant  
532 pairwise correlations with other parameters, shown in Table 3. Thus, al-  
533 though many of the parameters cannot be individually identified with high  
534 precision, the posterior is constrained to a relatively small region of the mul-  
535 tivariate parameter space where the resulting model behaviour is reasonable.

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

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.

536 The transient warming for the prior and posterior ensembles are presented  
537 in Figure 4. The prior ensemble has a fairly broad spread in transient climate  
538 response (TCR: warming observed after 70 years of 1% pa CO<sub>2</sub> enrichment)  
539 with a 5–95% range of of 1.91–2.62C, even though the equilibrium sensitivity  
540 is essentially fixed at close to 2.9C for all samples. However, the posterior  
541 ensemble range of TCR is reduced by a factor of more than 3 compared to the  
542 prior, with the range of 2.13–2.36C clustered tightly around the true value

543 of 2.21C. The 5–95% range of effective heat uptake efficiency  $\kappa$  of the ocean  
544 is 0.47–0.85  $Wm^{-2}K^{-1}$  in the prior, narrowing substantially to 0.57–0.69 in  
545 the posterior. The true value here is 0.64  $Wm^{-2}K^{-1}$ .

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

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

#### 572 *4.7. Discussion*

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

### Prior and posterior parameter distributions

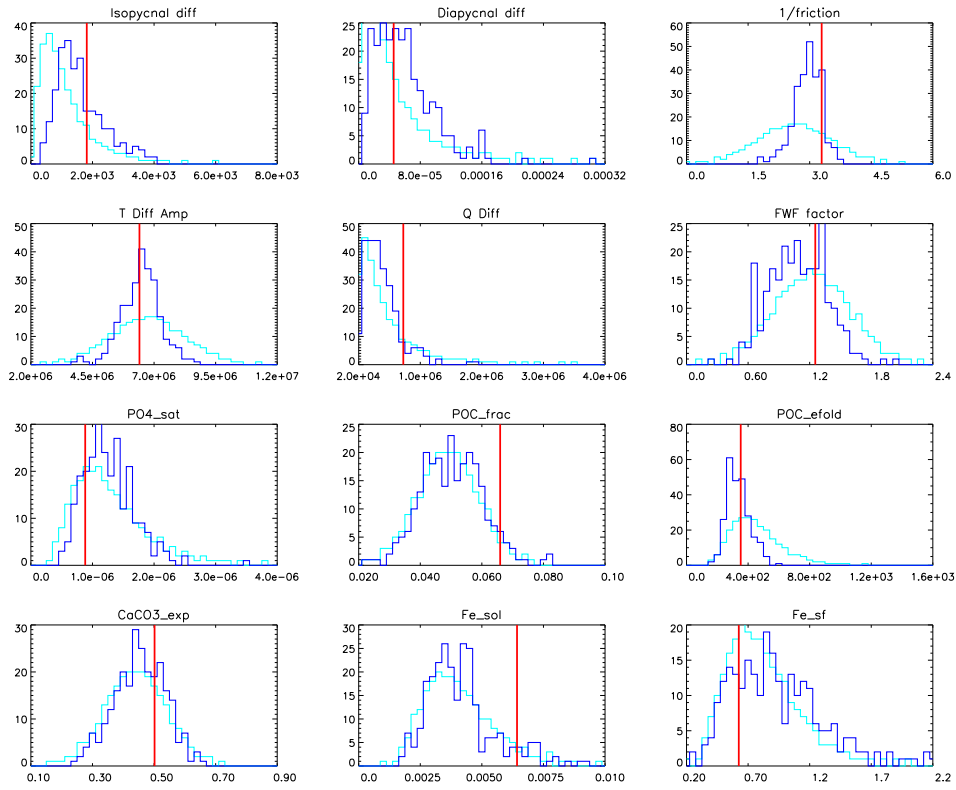


Figure 3: Prior and posterior distributions for the 12-parameter experiment described in the text. True parameter values are indicated by the vertical lines. Prior is cyan histogram and posterior is dark blue.

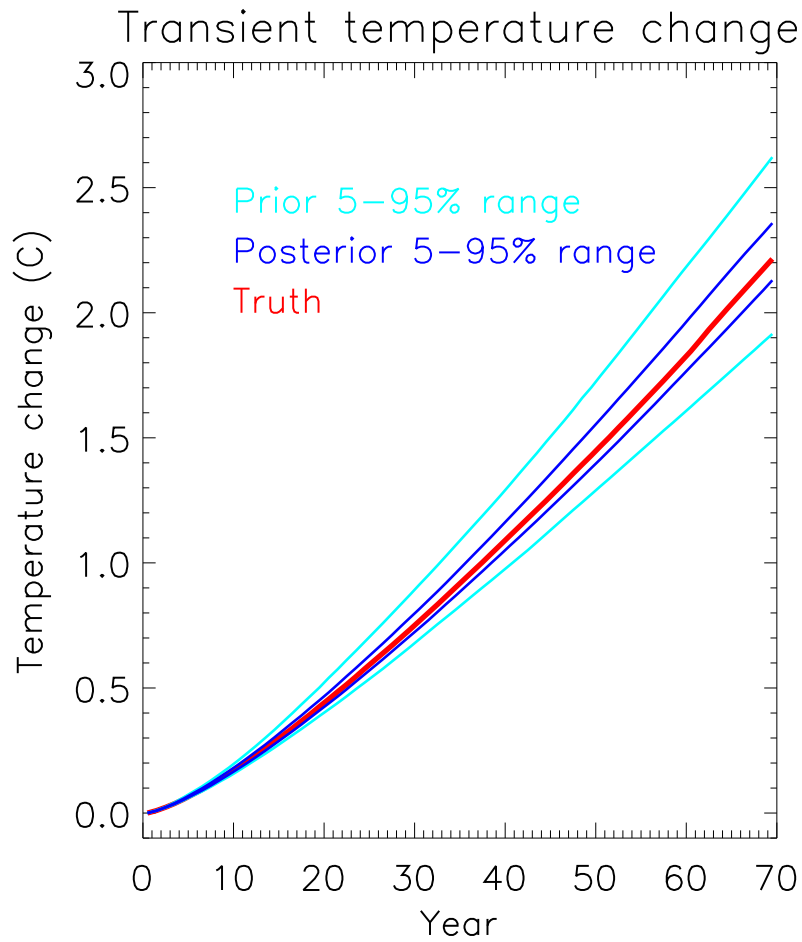


Figure 4: Results of a 70 year 1% per annum  $\text{CO}_2$  enrichment experiment, showing global mean surface temperature anomaly. Prior and posterior 5-95% ranges are indicated in cyan and dark blue respectively. Output of the truth run is shown in red.

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

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

618 estimation method is limited to equilibrium simulations.

## 619 **5. Conclusions**

620 We have presented a simple but effective method for parameter estima-  
621 tion in moderately high dimensional problems, based on an iterative impor-  
622 tance sampling approach. The method presented here shows a clear im-  
623 provement for nonlinear applications, compared to the ensemble Kalman  
624 filtering method which has been previously used. In (near-)linear problems,  
625 both methods generate good results, and the EnKF is more efficient in com-  
626 putational terms. However, in more strongly nonlinear applications, the  
627 importance sampling method is substantially more accurate. The method  
628 appears to generalise to problems of moderate dimensionality, as typically  
629 encountered in climate science, where direct sampling is computationally  
630 prohibitive. The combination of our efficient method together with a reason-  
631 ably realistic ocean model allows us to use physical and biogeochemical tracer  
632 data to constrain the dynamics of the ocean circulation for the first time.  
633 These data limit the model to a relatively small part of the multivariate pa-  
634 rameter space which strongly constrains the transient climate response. It  
635 therefore appears that observations of climatological tracer distributions in  
636 the ocean are informative about its role in the rate of global warming via  
637 heat uptake.

## 638 **6. Acknowledgments**

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

## 644 **References**

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

- 648 Annan, J. D., Hargreaves, J. C., Edwards, N. R., Marsh, R., 2005a. Param-  
649 eter estimation in an intermediate complexity Earth System Model using  
650 an ensemble Kalman filter. *Ocean Modelling* 8 (1–2), 135–154.
- 651 Annan, J. D., Lunt, D. J., Hargreaves, J. C., Valdes, P. J., 2005b. Parameter  
652 estimation in an atmospheric GCM. *Nonlinear processes in geophysics* 12,  
653 363–371.
- 654 Baker, A., Jickells, T., 2006. Mineral particle size as a control on aerosol iron  
655 solubility. *Geophys. Res. Lett* 33, 17.
- 656 Bengtsson, T., Bickel, P., Li, B., 2008. Curse-of-dimensionality revisited:  
657 Collapse of the particle filter in very large scale systems. In: *Probability  
658 and Statistics: Essays in Honor of David A. Freedman*. Vol. 2. Institute of  
659 Mathematical Statistics, pp. 316–334.
- 660 Bernardo, J., Smith, A., 1994. *Bayesian Theory*. Wiley, Chichester, UK.
- 661 Claussen, M., Mysak, L., Weaver, A., Crucifix, M., Fichefet, T., Loutre, M.,  
662 Weber, S., Alcamo, J., Alexeev, V., Berger, A., et al., 2002. Earth system  
663 models of intermediate complexity: closing the gap in the spectrum of  
664 climate system models. *Climate Dynamics* 18 (7), 579–586.
- 665 Collins, M., Brierley, C., MacVean, M., Booth, B., Harris, G., 2007. The  
666 sensitivity of the rate of transient climate change to ocean physics pertur-  
667 bations. *Journal of Climate* 20 (10), 2315–2320.
- 668 Conkright, M. E., Locarnini, R. A., Garcia, H. E., OBrien, T. D., Boyer,  
669 T. P., Stephens, C., Antonov, J. I., 2002. *World Ocean Atlas 2001: Ob-  
670 jective Analyses, Data, Statistics, and Figures*. CD-ROM Documentation,  
671 National Oceanographic Data Center, Silver Spring, MD.
- 672 Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers.  
673 *JR Statist Soc. B* 68, 411–436.
- 674 Doney, S., Lindsay, K., Fung, I., John, J., 2006. Natural variability in a stable,  
675 1000 year global coupled climate-carbon cycle simulation. *J. Climate* 19,  
676 3033–3054.
- 677 Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential Monte Carlo sam-  
678 pling methods for Bayesian filtering. *Statistics and Computing* 10 (3), 197–  
679 208.

- 680 Edwards, N. R., Marsh, R., 2005. Uncertainties due to transport-parameter  
681 sensitivity in an efficient 3-D ocean-climate model. *Climate Dynamics*  
682 24 (4), 415–433.
- 683 Evensen, G., 2003. The ensemble kalman filter: theoretical formulation and  
684 practical implementation. *Ocean Dynamics* 53, 343–367.
- 685 Forest, C., Stone, P., Sokolov, A., 2006. Estimated PDFs of climate system  
686 properties including natural and anthropogenic forcings. *Geophys. Res.*  
687 *Lett* 33.
- 688 Gelman, A., Meng, X., 1998. Simulating normalizing constants: From impor-  
689 tance sampling to bridge sampling to path sampling. *Statistical Science*,  
690 163–185.
- 691 Gregory, J., Mitchell, J., 1997. The climate response to CO<sub>2</sub> of the Hadley  
692 Centre coupled AOGCM with and without flux correction. *Geophys. Res.*  
693 *Lett* 24 (15), 1943–1946.
- 694 Hammersley, J. M., Handscomb, D. C., 1964. Monte Carlo methods. Methuen  
695 & Co Ltd, London.
- 696 Hansen, J., Russel, G., Lacis, A., Fung, I., Rind, D., 1985. Climate response  
697 times: Dependence on climate sensitivity and ocean mixing. *Science* 229,  
698 857–859.
- 699 Hargreaves, J. C., Annan, J. D., Edwards, N. R., Marsh, R., 2004. Climate  
700 forecasting using an intermediate complexity Earth System Model and the  
701 ensemble Kalman filter. *Climate Dynamics* 23 (7–8), 745–760.
- 702 Hoffert, M., Callegari, A., Hsieh, C., 1985. The role of deep sea heat storage in  
703 the secular response to climatic forcing. *Journal of Geophysical Research-*  
704 *Oceans* 85 (C11).
- 705 Huang, R., 1999. Mixing and Energetics of the Oceanic Thermohaline Cir-  
706 culation. *Journal of Physical Oceanography* 29 (4), 727–746.
- 707 Kalman, R. E., 1960. A new approach to linear filtering and prediction prob-  
708 lems. *J. Basic Engineering* 82D, 33–45.

- 709 Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin,  
710 L., Iredell, M., Saha, S., White, G., Woollen, J., et al., 1996. The  
711 NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Me-*  
712 *teorological Society* 77 (3), 437–471.
- 713 Key, R., Kozyr, A., Sabine, C., Lee, K., Wanninkhof, R., Bullister, J., Feely,  
714 R., Millero, F., Mordy, C., Peng, T., 2004. A global ocean carbon clima-  
715 tology: Results from Global Data Analysis Project (GLODAP). *Global*  
716 *Biogeochem. Cycles* 18 (4).
- 717 Knutti, R., Stocker, T. F., Joos, F., Plattner, G.-K., 2002. Constraints on  
718 radiative forcing and future climate change from observations and climate  
719 model ensembles. *Nature* 416, 719–723.
- 720 Knutti, R., Tomassini, L., 2008. Constraints on the transient climate response  
721 from observed global temperature and ocean heat uptake. *Geophys. Res.*  
722 *Lett.* 35, L09701.
- 723 Lenton, T. M., Marsh, R., Price, A. R., Lunt, D. J., Aksenov, Y., Annan,  
724 J. D., Cooper-Chadwick, T., Cox, S. J., Edwards, N. R., Goswami, S.,  
725 Hargreaves, J. C., Harris, P. P., Jiao, Z., Livina, V. N., Payne, A. J., Rutt,  
726 I. C., Shepherd, J. G., Valdes, P. J., Williams, G., Williamson, M. S., ,  
727 Yool, A., 2007. A modular, scalable, Grid ENabled Integrated Earth sys-  
728 tem modelling (GENIE) framework: Effects of atmospheric dynamics and  
729 ocean resolution on bi-stability of the thermohaline circulation. *Climate*  
730 *Dynamics* 29 (6), 591–613.
- 731 Li, Y., Peng, S., Broecker, W., Oestlund, H., 1984. The average vertical  
732 mixing coefficient for the oceanic thermocline. *Tellus. Series B, Chemical*  
733 *and physical meteorology* 36 (3), 212–217.
- 734 Lindzen, R., 2002. Do deep ocean temperature records verify models? *Geo-*  
735 *physical Research Letters* 29 (8), 95–1.
- 736 Mahowald, N., Kohfeld, K., Hannson, M., Balkanski, Y., Harrison, S., Pren-  
737 tice, I., Schulz, M., Rodhe, H., 1999. Dust sources during the last glacial  
738 maximum and current climate: a comparison of model results with paleo-  
739 odata from ice cores and marine sediments. *J. Geophys. Res* 104, 15895–  
740 15916.

- 741 McKay, M., Beckman, R., Conover, W., 1979. A comparison of three methods  
742 for selecting values of input variables in the analysis of output from a  
743 computer code. *Technometrics* 21 (2), 239–245.
- 744 Meehl, G. A., Stocker, T. F., Collins, W. D., et al., 2007. Global Climate  
745 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–  
748 845.
- 750 Meng, X., Wong, W., 1996. Simulating ratios of normalizing constants via a  
751 simple identity: a theoretical exploration. *Statistica Sinica* 6, 831–860.
- 752 Munk, W., 1966. Abyssal recipes. *Deep-Sea Res* 13, 707–730.
- 753 Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb,  
754 M. J., Collins, M., Stainforth, D. A., 2004. Quantification of modelling  
755 uncertainties in a large ensemble of climate change simulations. *Nature*  
756 430, 768–772.
- 757 Navon, I. M., 1998. Practical and theoretical aspects of adjoint parameter  
758 estimation and identifiability in meteorology and oceanography. *Dynamics of Atmospheres and Oceans* 27 (1–4), 55–79.
- 760 Neal, R., of Computer Science, D., of Toronto, U., 1993. Probabilistic infer-  
761 ence using Markov chain Monte Carlo methods. Department of Computer  
762 Science, University of Toronto.
- 763 Oliver, K., Edwards, N., 2008. Location of potential energy sources and the  
764 export of dense water from the Atlantic Ocean. *Geophysical Research Letters*  
765 35 (22), L22604.
- 766 Oort, A. H., 1983. Global atmospheric circulation statistics, 1958–1973,  
767 NOAA Professional Paper 14.
- 768 Parekh, P., Follows, M., Boyle, E., 2005. Decoupling of iron and phosphate  
769 in the global ocean. *Global Biogeochem. Cycles* 19.
- 770 Parekh, P., Follows, M., Dutkiewicz, S., Ito, T., 2006. Physical and biological  
771 regulation of the soft tissue carbon pump. *Paleoceanography* 21 (3), 1–  
772 A3001.

- 773 Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1994.  
774 Numerical recipes in Fortran: the art of scientific computing. Cambridge  
775 University Press.
- 776 Raper, S., Gregory, J., Stouffer, R., 2002. The role of climate sensitivity and  
777 ocean heat uptake on AOGCM transient temperature response. *Journal of*  
778 *Climate* 15 (1), 124–130.
- 779 Ridgwell, A., 2001. Glacial-interglacial perturbations in the global carbon  
780 cycle. Ph.D. thesis, University of East Anglia.
- 781 Ridgwell, A., Hargreaves, J. C., Edwards, N. R., Annan, J. D., Lenton,  
782 T. M., Marsh, R., Yool, A., Watson, A., 2007. Marine geochemical data  
783 assimilation in an efficient earth system model of global biogeochemical  
784 cycling. *Biogeosciences* 4 (1), 87–104.
- 785 Ridgwell, A., Singarayer, J., Hetherington, A., Valdes, P., 2009. Tackling  
786 Regional Climate Change By Leaf Albedo Bio-geoengineering. *Current Bi-*  
787 *ology* 19 (2), 146–150.
- 788 Rougier, J. C., 2007. Probabilistic inference for future climate using an en-  
789 semble of simulator evaluations. *Climatic Change* 81, 247–264.
- 790 Singarayer, J., Richards, D., Ridgwell, A., Valdes, P., Austin, W., Beck, J.,  
791 2008. An oceanic origin for the increase of atmospheric radiocarbon during  
792 the Younger Dryas. *Geophysical Research Letters* 35 (14), L14707.
- 793 Sokolov, A., Stone, P., Forest, C., Prinn, R., Sarofim, M., Webster, M.,  
794 Paltsev, S., Schlosser, C., Kicklighter, D., Dutkiewicz, S., et al., 2009.  
795 Probabilistic Forecast for 21st Century Climate Based on Uncertainties in  
796 Emissions (without Policy) and Climate Parameters. *Journal of Climate*  
797 22, 5175–5204.
- 798 West, M., 1993. Approximating Posterior Distributions by Mixture. *Journal*  
799 *of the Royal Statistical Society. Series B (Methodological)* 55 (2), 409–422.
- 800 Wilks, D. S., 1995. *Statistical methods in the atmospheric sciences*. Academic  
801 Press.