

## An Efficient Stochastic Bayesian Approach to Optimal Parameter and Uncertainty Estimation for Climate Model Predictions

CHARLES JACKSON, MRINAL K. SEN, AND PAUL L. STOFFA

*Institute for Geophysics, The John A. and Katherine G. Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas*

(Manuscript received 31 December 2002, in final form 14 February 2004)

### ABSTRACT

One source of uncertainty for climate model predictions arises from the fact that climate models have been optimized to reproduce observational means. To quantify the uncertainty resulting from a realistic range of model configurations, it is necessary to estimate a multidimensional probability distribution that quantifies how likely different model parameter combinations are, given knowledge of the uncertainties in the observations. The computational cost of mapping a multidimensional probability distribution for a climate model using traditional means (e.g., Monte Carlo or Metropolis/Gibbs sampling) is impractical, requiring  $10^4$ – $10^6$  model evaluations for problems involving less than 10 parameters. This paper examines whether such a calculation is more feasible using a particularly efficient but approximate algorithm called Bayesian stochastic inversion, based on multiple very fast simulated annealing (VFSA). Investigated here is how the number of model parameters, natural variability, and the degree of nonlinearity affect the computational cost and accuracy of estimating parameter uncertainties within a surrogate climate model that is able to approximate the noise and response behavior of a realistic atmospheric GCM. In general, multiple VFSA is one to two orders of magnitude more efficient than the Metropolis/Gibbs sampler, depending primarily on dimensionality of the parameter space analysis. The average cost of estimating parameter uncertainties is only moderately affected by noise within the model as long as the signal-to-noise ratio is greater than 5. Also the average cost of estimating parameter uncertainties nearly doubles for problems in which parameters are nonlinearly related.

### 1. Introduction

Three-dimensional climate models are the primary tools we have to study the physics of the climate system and to make predictions of future climate. The degree of trust we place in model predictions depends on our ability to assess a model's strengths or weaknesses. So far this has largely been achieved through intermodel comparisons of simulated and observed climate (e.g., Boer et al. 1992) or through more standardized model intercomparison projects in which model forcings and experimental designs are similar (Gates et al. 1999; Joussaume and Taylor 2000). These comparisons are useful for quantifying a consensus and range of behaviors that exist for different three-dimensional atmospheric models given the task of reproducing observed climate or simulating well-preserved climate states in the geologic past. Similar intercomparison projects exist for coupled atmosphere–ocean models and models of the land surface (Covey et al. 2003; Pitman and Henderson-Sellers 1998). As useful as these calculations are

to evaluating trustworthy aspects of model predictions, there is no way to know what uncertainty limits this range of behaviors may be expressing. It is often assumed that the range of behaviors exhibited by models that participate in the intercomparison projects is representative of a realistic range of probable outcomes. This has been the assumption for various estimates of the uncertainty within climate model predictions of the climate response to projected increases in greenhouse gases (e.g., Allen et al. 2000). And although there may be some recognition of which models perform better than others, the qualitative approach to evaluating model performance does not lend itself to assigning quantitative likelihoods to model predictions. The need for more quantitative evaluations of modeling uncertainty was made explicit within the most recent 2001 Intergovernmental Panel on Climate Change report on climate change (McAvaney et al. 2001).

Because it is usual for models to be tuned towards the mean of what is observed, we may have reason to believe that the range of (unforced) model predictions given by intercomparison projects are biased towards the mean rather than being fully representative of the range. As there exists a wide range in equilibrium sensitivities to carbon dioxide forcing (Cubasch et al. 2001), we do not know what to expect for the spread

---

*Corresponding author address:* Charles Jackson, Institute for Geophysics, The John A. and Katherine G. Jackson School of Geosciences, The University of Texas at Austin, 4412 Spicewood Spring Road, Bldg. 600, Austin, TX 78759-8500.  
E-mail: charles@ig.utexas.edu

of predictions within forced experiments. The only way of determining an ensemble of models that one could be assured is representative of the combined uncertainty in the observations and model physics is to identify the parameters that are the primary sources of model uncertainty and to rank the performance of each member of that ensemble by how well it matches observations given allowances provided by the knowledge of natural variability or observational error.

One of the main limitations to evaluating climate model uncertainties in such a way is the computational cost of evaluating the uncertainty associated with any given choice of model parameter values or the uncertainty associated with any given combination of model parameter values when a climate model's response depends nonlinearly to the combined changes in model parameters (e.g., Williams et al. 2001). This exercise has been carried out within reduced dimensioned climate models (Forest et al. 2000, 2001, 2002) and is being planned on a larger scale within three-dimensional climate models (Allen 1999). A number of authors have pursued randomized Latin Hypercube sampling and kriging interpolation techniques to reduce the number of experiments that may be needed to estimate the multidimensional dependencies (Sacks et al. 1989; Welch et al. 1992; Bowman et al. 1993; Chapman et al. 1994). So far these analyses have focused on deriving functional response surfaces with additional sampling in regions that are not well represented by linear combinations of completed experiments. As will be explained in greater detail below, the stochastic representation of model parameter uncertainty is sensitive to sampling strategies. Although the Latin Hypercube sampling and kriging interpolation may prove useful for estimating parameter uncertainties, more work is needed to show how this can be done.

It is the purpose of this paper to examine the algorithmic issues related to the implementation of a particularly efficient method for estimating parameter interdependencies and uncertainties that is currently being implemented within the solid earth geophysics community (Sen and Stoffa 1996). The method is called stochastic Bayesian inversion (BSI) and is based on multiple very fast simulated annealing (VFSA). It strikes a balance between the sometimes separate objectives of estimating the optimal model parameter values with estimating uncertainties through parameter multidimensional posterior probability density functions (PPD) and covariances. Because these types of calculations have been prohibitively expensive in the past, the degree of efficiency will determine how comprehensive a given parameter uncertainty analysis can be. We use a surrogate climate model to approximate the noise and response behavior of an atmospheric-slab ocean model system to arbitrary choices in three, six, and nine model parameters. Our main goal is to consider how natural variability, the number of uncertain model parameters, and the degree of nonlinearity affect the cost

and accuracy of estimating multidimensional parameter probability density functions. In other papers (Jackson et al. 2003), we have applied this approach to optimal parameter and uncertainty estimation for 12 parameters within a land surface model (Jackson et al. 2003) and evaluated different methods of defining a metric of climate model error (Mu et al. 2004, manuscript submitted to *J. Geophys. Res.*).

## 2. Optimal parameter and uncertainty estimation

### a. Bayesian stochastic inversion formulation

Observed data are often only indirectly related to quantities or processes of interest. If an understanding exists of the processes that influenced the observed data, one may design a "forward" model of these processes and use inverse modeling tools to make inferences about forcings or properties of the system that would be difficult to obtain through more direct means. The BSI formulation of an inverse problem is based on a probabilistic approach that explicitly provides a means for accounting for uncertainties in observed data and model predictions. Sen and Stoffa (1996) provide a review of optimal parameter and uncertainty estimation of geophysical models based on Bayesian statistics. For completeness, we present here a summary of some of the important points in that formulation.

According to Bayes's rule, derived from the definition of conditional probabilities, a posterior probability density (PPD) function is defined as

$$\sigma(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \quad (1)$$

where  $\mathbf{m}$  is a model vector of random parameters,  $\mathbf{d}$  is the data vector,  $\sigma(\mathbf{m}|\mathbf{d})$  is the conditional probability for model parameters represented by vector  $\mathbf{m}$  given data in vector  $\mathbf{d}$ ,  $p(\mathbf{d}|\mathbf{m})$  is the conditional probability expressing the relative probability for data vector  $\mathbf{d}$  given model parameters  $\mathbf{m}$ ,  $p(\mathbf{m})$  is a "prior" probability for  $\mathbf{m}$  given expert judgment or other reasons to constrain the possible choices of  $\mathbf{m}$  independent of data  $\mathbf{d}$ , and  $p(\mathbf{d})$  is the probability of data vector  $\mathbf{d}$ . For our purposes, Eq. (1) expresses an outline for the solution to an inverse problem where one can substitute a set of observations  $\mathbf{d}_{\text{obs}}$  for  $\mathbf{d}$  and identify through  $\sigma(\mathbf{m}|\mathbf{d}_{\text{obs}})$  the best choice of model parameters  $\mathbf{m}$  that is consistent with  $\mathbf{d}_{\text{obs}}$ . The denominator  $p(\mathbf{d})$  is independent of  $\mathbf{m}$  and may be considered a constant within the inverse calculation (e.g., Tarantola 1987). When a particular set of observations  $\mathbf{d}_{\text{obs}}$  is used for  $\mathbf{d}$ ,  $p(\mathbf{d}|\mathbf{m})$  is called a likelihood function  $l(\mathbf{d}_{\text{obs}}|\mathbf{m})$ . When Gaussian errors are assumed within the observations and model predictions, the likelihood function takes the form

$$l(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto \exp[-SE(\mathbf{m})], \quad (2)$$

where  $E(\mathbf{m})$  is a cost function that gives a measure of mismatch between observations and model predictions

and  $S$  is a scaling factor that will be discussed more fully below. The cost function can be defined in many ways. Assuming Gaussian errors in the data, we may choose

$$E(\mathbf{m}) = \sum_{i=1}^N \frac{1}{2N} \{ [\mathbf{d}_{\text{obs}} - g(\mathbf{m})]^T \times \mathbf{C}^{-1} [\mathbf{d}_{\text{obs}} - g(\mathbf{m})] \}_i, \quad (3)$$

where there are  $N$  different sets of observations (e.g., air temperature, precipitation, winds) with each set of observations containing  $M$  data points,  $g(\mathbf{m})$  is the  $M \times 1$  vector of model predictions, and  $\mathbf{C}^{-1}$  is the  $M \times M$  inverse of the data covariance matrix that includes both the observational error and modeling uncertainty error defined separately for each model quantity.

As allowed within the Bayesian formulation, the inverse of the data covariance matrix  $\mathbf{C}^{-1}$  may be modified by a scaling factor  $S$ . This freedom is at the heart of the Bayesian formulation to provide a means to weigh different choices of  $\mathbf{m}$  according to expert opinion or other information that is not easily incorporated into the analysis. We will show that there are ways of objectively choosing  $S$  that are consistent with more classical ways of expressing uncertainty in the following section.

With these substitutions, Eq. (1) becomes

$$\sigma(\mathbf{m} | \mathbf{d}_{\text{obs}}) = \frac{\exp[-SE(\mathbf{m})]p(\mathbf{m})}{\int \exp[-SE(\mathbf{m})]p(\mathbf{m}) d\mathbf{m}}. \quad (4)$$

Although this form of the PPD assumes Gaussian errors in observations and in model predictions, because the forward model is a part of the error function, there is no expectation that the PPD would itself be Gaussian (Tarantola 1987).

Once the PPD is known, the parameter means  $\langle \mathbf{m} \rangle$  or covariances can be obtained through multidimensional integrals of the general form

$$\mathbf{I} = \int f(\mathbf{m})\sigma(\mathbf{m} | \mathbf{d}_{\text{obs}}) d\mathbf{m}, \quad (5)$$

where  $f(\mathbf{m}) = \mathbf{m}$  for the parameter means or  $f(\mathbf{m}) = (\mathbf{m} - \langle \mathbf{m} \rangle)(\mathbf{m} - \langle \mathbf{m} \rangle)^T$  for the parameter covariance matrix. Because the PPD is multidimensional, it is difficult to visualize. One approach is to display the marginal PPD, defined to be the one-dimensional projection of the multidimensional PPD [equivalent to  $f(\mathbf{m}) = 1$  in Eq. (5) and integrating over all dimensions except for the dimension of interest].

#### b. Choice of scaling factor $S$

The scaling factor  $S$  within the cost function [Eq. (3)] should be viewed as part of the inverse data covariance matrix  $\mathbf{C}^{-1}$  that performs the function of weighing the significance of model–data differences. Large values of

$S$  would imply small errors within the data and would result in sharply peaked probability distributions using Eq. (4). Fortunately, one does not need to know an appropriate choice of  $S$  to rank model parameter sets by their cost function values. After the model parameter sets have been ranked, one can select an appropriate choice of  $S$  such that the likelihood measures that we assign to parameter sets are consistent with our estimates of uncertainty: First, obtain a distribution of cost function values that represents the effect of internal variability on the cost function. This may be achieved by repeating a standard experiment a number of times with different initial conditions. From this distribution one can assess the maximum and minimum cost function values ( $\Delta E = E_{97.5} - E_{2.5}$ ) and assume that this range provides a nominal 95% credible interval. One may then apply the logic that parameter sets that are  $\Delta E$  away from the optimal parameter set will be given a likelihood measure [Eq. (2)] of  $\exp(-2)$ , which is equivalent to the 95% probability measure for a Gaussian distribution. This implies  $S = 2/\Delta E$ . In general, this method for selecting  $S$  will likely be more straightforward for climate problems where one can use a model to predict observational uncertainties that arise from natural variability. This is not typically the case for many other classes of problems, especially in geophysics where one needs to rely on expert opinion to decide which solutions appear more realistic.

### 3. Numerical sampling methods

The principal challenge in calculating integrals based on Eq. (5) is in deriving the PPD, which can be very computationally intense. We review here several numerical methods for deriving or approximating the PPD and focus on multiple VFSA, which has been highlighted by Sen and Stoffa (1996) as being particularly efficient.

#### a. Grid search

This straightforward method involves subdividing model parameter space into a number of equally spaced intervals and enumerating every possible combination of model parameters and evaluating the cost function for each of these combinations (Sen and Stoffa 1996). The disadvantages of this method are the large number of forward model calculations, many of which do not contribute substantially to the integral in Eq. (5), and the fact that the resolution is constrained by the interval spacing.

#### b. Gibbs sampler

The Gibbs sampler is a version of an ‘‘importance sampling’’ technique that improves the efficiency of the calculation by sampling model parameter sets from the Gibbs distribution which is, in effect, equivalent to the

desired PPD (Metropolis et al. 1953; Kirkpatrick et al. 1983). The concept of the Gibbs sampler originates from the numerical techniques developed in statistical mechanics to simulate the macroscopic behavior of a system with a large number of interacting particles. The numerical implementation of the Gibbs sampler is based primarily on the heat-bath algorithm in which the relative probability of different model parameter sets are evaluated in advance of any model evaluations (Geman and Geman 1984; Rothman 1986). This algorithm also requires parameter space to be subdivided into a number of equally spaced intervals. Another variant of the Gibbs sampler is based on the Metropolis algorithm (Metropolis et al. 1953; Kirkpatrick et al. 1983). In the Metropolis formulation, a starting model is selected at random and the cost function is evaluated as  $E(\mathbf{m}_i)$ . The starting model is perturbed to obtain a new model  $\mathbf{m}_{i+1}$  and new cost function evaluation  $E(\mathbf{m}_{i+1})$ . If the change in “energy”  $\Delta E = E(\mathbf{m}_{i+1}) - E(\mathbf{m}_i)$  is negative, then the new model is accepted. If the change is positive, the new model is accepted with a probability

$$P = \exp\left(\frac{-\Delta E}{T}\right), \quad (6)$$

where  $T$  is a control parameter analogous to temperature. That is, the new model is rejected if  $P$  is less than a randomly generated number between 0 and 1. Any rejected experiments are not used to calculate probability statistics. Instead any rejected iteration is assumed to repeat the last accepted experiment. After a large number of iterations at constant temperature, the PPD for  $\mathbf{m}$  [notated as  $\text{prob}(\mathbf{m})$ ] converges to the following Gibbs distribution:

$$\text{prob}(\mathbf{m}) = \frac{\exp\left[\frac{-E(\mathbf{m})}{T}\right]}{\sum \exp\left[\frac{-E(\mathbf{m})}{T}\right]}. \quad (7)$$

As noted by Sen and Stoffa (1996), the fact that the distribution for  $\text{prob}(\mathbf{m})$  with  $T = 1$  in Eq. (7) is the same as for  $\sigma(\mathbf{m} | \mathbf{d}_{\text{obs}})$  in Eq. (4) except for the presence of the prior distribution  $p(\mathbf{m})$  means that when using the Gibbs sampler at constant temperature, the distribution will converge to the distribution of the PPD without bias when assuming a constant prior. Because sampled models are not discretized in the Metropolis formulation of the Gibbs sampler, the resolution of the Metropolis/Gibbs sampler will depend on the number of iterations. The resolution in general will be much greater near the minimum in  $E(\mathbf{m})$  because those regions tend to be sampled more often. While this method can substantially improve the efficiency of evaluating the PPD, the number of forward model calculations is still prohibitively large for many purposes.

### c. Multiple very fast simulated annealing

One may use the temperature construct within the Metropolis algorithm to locate the global minimum in  $E(\mathbf{m})$  by very slowly lowering the temperature parameter within Eq. (6). This procedure is analogous to the annealing process within a physical system whereby the lowest energy state between atoms or molecules (the crystalline form) is achieved by the gradual cooling of the substance within a heat bath. Because of this physical analogy, the algorithm is known as simulated annealing. Ingber (1989) introduced within the simulated-annealing algorithm a new procedure for selecting parameter sets according to a temperature-dependent Cauchy distribution. This modified algorithm further enhances the ability of simulated annealing to converge efficiently and robustly to the global minimum in  $E(\mathbf{m})$  and is referred to as very fast simulated annealing. A comparison between the Metropolis and VFSA algorithms is shown in Fig. 1. A demonstration of the similarity in their results will not be shown until section 6a.

The selection of model parameters given an initial selection  $m_i$  within VFSA are chosen such that

$$m_i^{k+1} = m_i^k + y_i(m_i^{\text{max}} - m_i^{\text{min}}), \quad (8)$$

$$y_i \in [-1, 1], \quad \text{and} \quad (9)$$

$$m_i^{\text{min}} \leq m_i^{k+1} \leq m_i^{\text{max}}, \quad (10)$$

where  $y_i$  is generated according to a Cauchy distribution

$$y_i = \text{sgn}(\text{RND} - 0.5) T_k \left[ \left(1 + \frac{1}{T_k}\right)^{|2\text{RND}-1|} - 1 \right]. \quad (11)$$

Within Eqs. (8)–(11), subscript  $i$  is the parameter number,  $k$  is the iteration number, RND is a random number generator with a uniform distribution between 0 and 1, and  $\text{sgn}$  is the sign operator. The cooling schedule at iteration  $k$  is

$$T_k = T_0 \exp[\alpha(k - 1)^{1/\text{NM}}], \quad (12)$$

with NM equal to the number of parameters and  $\alpha$  as a tunable parameter that can be tailored for particular problems. The acceptance criterion for successive model selections is the same as for the Metropolis rule.

One of the central points in our discussion on what efficiencies can be afforded within the numerical methods that we adopt for optimal parameter and uncertainty estimation is that there needs to be a balance between the computational effort that is spent identifying optimal model parameters and the computational effort that is spent mapping the multidimensional PPD. The VFSA algorithm as presented by Ingber (1989) and used by Sen and Stoffa (1996) is an efficient method to identify optimal parameters, especially when nonlinearities are important. At the other extreme, a Monte Carlo or grid-search algorithm would provide the most accurate (non-

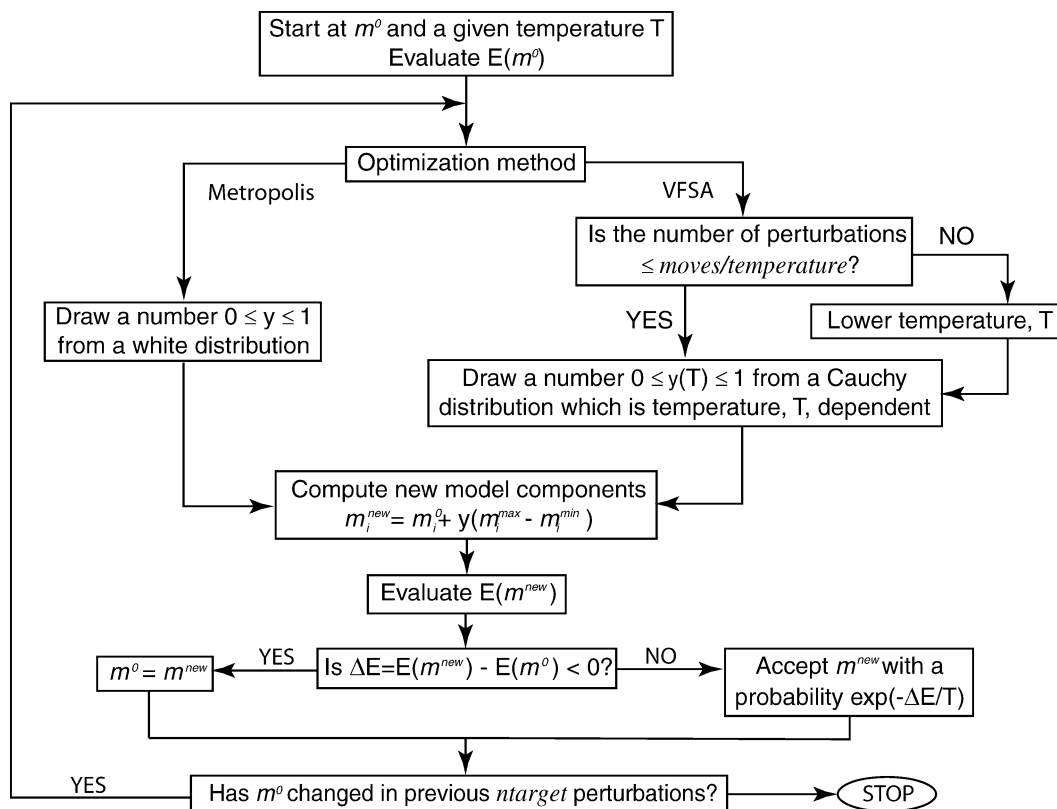


FIG. 1. Schematic diagram of the Metropolis and VFSA algorithms. Parameter *moves/temperature* gives the number of times a new model parameter set is selected and tested before lowering the temperature  $T$ . Parameter *n<sub>target</sub>* specifies the convergence criterion and is given by the maximum number of failed attempts at finding an acceptable parameter set before stopping.

biased) map of the multidimensional PPD but may require more model evaluations than one can afford to make. Sen and Stoffa (1996) argue that one can slightly adapt the VFSA cooling schedule, convergence acceptance criterion, and allow for numerous repetitions of the minimization procedure to strike a balance between these two objectives that is both efficient and effective. Depending on the application, one can save several orders of magnitude in the number of model evaluations over either the Monte Carlo or grid-search algorithms (Sen and Stoffa 1996). We refer to this use of VFSA as simply “multiple VFSA.”

Although varying NM in Eq. (12) is an effective way to tailor the algorithm to a given problem, we have found that scaling NM to the number of uncertain model parameters to be too conservative. Instead we recommend fixing NM’s value at 2 and altering the number of model evaluations at any given temperature. This change to the VFSA algorithm introduces a new parameter that we name “moves per temperature” (*moves/temperature*, Fig. 1). The effect of *moves/temperature* is slightly different from the NM parameter on the cooling schedule. For a similarly valued NM and *moves/temperature*, NM tends to produce greater cooling early and more gradual cooling as the number of parameter set perturbation at-

tempts becomes large. The  $\alpha$  parameter in Eq. (12) is not an effective means for altering the cooling schedule. We recommend fixing its value to 0.9.

Another multiple-VFSA parameter that we introduce here that is helpful for the effective use of this algorithm within Bayesian Stochastic Inversion framework is *n<sub>target</sub>* which gives the convergence criterion for a single VFSA attempt at finding a minimum of the cost function. *N<sub>target</sub>* is defined to be the maximum number of parameter set perturbations that are allowed in which no alternate set of model parameters has been identified that reduces the cost function (Fig. 1).

The PPD derived through the multiple-VFSA search algorithm is unavoidably biased towards the peaks of the PPD through its procedure to change the temperature control parameter during the selection process of model parameters. However, the VFSA algorithm may be repeated a number of times with different starting models to allow sufficient sampling of the entire model space. This reduces the biases of the PPD and improves the estimates of the model covariance matrix. While variances may be underestimated relative to what may be obtained through the Metropolis/Gibbs sampler, the normalized covariance matrix (the correlation matrix) has

been found to be nearly equivalent between the two approaches (Sen and Stoffa 1996).

#### 4. Experimental design

Experiments for testing the cost of deriving probability density functions and covariances for target parameters is broken down into two objectives: 1) What is the minimum number of model evaluations that are needed to identify the set of model parameters that minimize the cost function? And 2), what is the minimum number of model evaluations that are needed to determine statistically stable estimates of the multidimensional parameter probability density function? Each of these components of a parameter uncertainty analysis will likely depend on 1) the signal-to-noise ratio for individual model evaluations, 2) the number of model parameters being investigated, and 3) the degree of nonlinearity between model parameters.

Experiments testing the impact of each of these factors will be conducted using a surrogate climate model that can mimic the response and natural variability of a climate model to arbitrary changes in three, six, or nine linearly or nonlinearly related parameters. The main purpose of the surrogate climate model is to demonstrate the efficiency of a parameter uncertainty analysis based on multiple VFSA as compared to the Metropolis/Gibbs sampler. The model is based on surface air temperature fields generated by an atmospheric general circulation model coupled to a 50-m slab ocean in its response to continuous changes in three parameters specifying Earth's orbital geometry over the past 165 kyr (Jackson and Broccoli 2003). These three parameters are obliquity (Earth's tilt), longitude of the perihelion (celestial longitude at which Earth is at its closest approach to the sun), and eccentricity (degree of orbit circularity). The surrogate climate model "simulations" are based on the three-dimensional climate model's near-surface air temperature response to arbitrary changes in the three orbital parameters. The response can be approximated in terms of obliquity and precession components (eccentricity-scaled precession) using a least squares fitting procedure (for details see Jackson and Broccoli 2003). The residual contains a weak signal from nonlinearities and the direct response to changes in eccentricity and is dominated by internally generated noise. Therefore, the three-dimensional climate model response ( $R$ ) may be approximated by

$$R(i, j, k) \approx \text{obliquity}(i, j, k) + \text{precession}(i, j, k) + \text{noise}(i, j, k). \quad (13)$$

Here

$$\text{obliquity}(i, j, k) = \text{ob}' \frac{dT(i, j, k)}{dob}, \quad \text{and} \quad (14)$$

$$\text{precession}(i, j, k) = \text{ec} \frac{dT(i, j, k)}{dec} \cos[\text{phase}(i, j, k) - \lambda_p], \quad (15)$$

where  $i$ ,  $j$ , and  $k$  refer to the latitude, longitude, and season, respectively,  $\text{ob}'$  is the change in obliquity from its long-term mean of  $23.3513^\circ$ ,  $\text{ec}$  is the eccentricity,  $\lambda_p$  is the longitude of the perihelion,  $dT(i, j, k)/dob$  is the sensitivity of near-surface air temperature  $T(i, j, k)$  to changes in obliquity,  $dT(i, j, k)/dec$  is the corresponding sensitivity to changes in eccentricity, and  $\text{phase}(i, j, k)$  is the longitude of the perihelion relative to the autumnal equinox when near-surface air temperature is maximized. Within Eqs. (14) and (15) the only unknowns are the parameters specifying obliquity, longitude of the perihelion, and eccentricity with the other factors being derived from the Jackson and Broccoli (2003) experiment results. The component of Eq. (13) corresponding to noise is obtained from anomalies generated from a 1500-yr-long control integration. Each new experiment with the surrogate climate model includes the variability from a nonoverlapping 1-, 5-, or 10-yr-long segment (depending on the experiment) of this long integration, only repeating after sampling the entire time series. Although Eq. (13) is mathematically nonlinear, the model's response to a wide range of parameter combinations is observed to be nearly linear and serves as our "linear" surrogate climate system model with three nominal degrees of freedom [ $R_{3,L} \equiv$  Eq. (13)] stemming from potentially unknown values of obliquity, longitude of the perihelion, and eccentricity in a target dataset. A linear six- and nine-parameter climate system model was created by adding (linearly independent) spherical harmonic (SH) "signals" to Eq. (13),

$$R_{6,L} = R_{3,L} + \text{SH}_{3,L} \quad \text{and} \quad (16)$$

$$R_{9,L} = R_{3,L} + \text{SH}_{6,L}, \quad (17)$$

where

$$\text{SH}_{3,L} = A_1 Y_{mn}(3, 3) + A_2 Y_{mn}(3, 4) + A_3 Y_{mn}(3, 5) \quad (18)$$

and

$$\begin{aligned} \text{SH}_{6,L} = & \text{SH}_{3,L} + A_4 Y_{mn}(4, 3) + A_5 Y_{mn}(4, 4) \\ & + A_6 Y_{mn}(4, 5). \end{aligned} \quad (19)$$

The  $Y_{mn}$  functions in Eqs. (18)–(19) are individual spherical harmonics (associated Legendre polynomials) of wavenumbers  $m$  and  $n$ . Coefficients  $A_1$ – $A_6$  are free parameters that determine the amplitude of the respective spherical harmonic signals. For ease of notation, functional dependence on latitude, longitude, and season ( $i, j, k$ ) has been omitted.

For experiments where we investigate the effects of nonlinear dependence between model parameters, the functional dependence of the linear model was modified to permit the existence of two solutions as well as maintain an approximate signal-to-noise ratio that existed for the linear versions. For the nonlinear three-parameter surrogate climate model ( $R_{3,NL}$ )

$$R_{3,NL} = 2 \operatorname{sgn}(\text{obliquity} \times \text{precession}) \\ \times \sqrt{|\text{obliquity} \times \text{precession}|} + \text{noise}, \quad (20)$$

where  $\operatorname{sgn}()$  provides the positive or negative sign of its operand. The nonlinear six- or nine-parameter surrogate climate models ( $R_{6,NL}$  and  $R_{9,NL}$ ) are defined as

$$R_{6,NL} = R_{3,NL} + SH_{3,NL} \quad \text{and} \\ R_{9,NL} = R_{3,NL} + SH_{6,NL}, \quad (21)$$

where

$$SH_{3,NL} = 4A_1 Y_{mn}(3, 3)A_2 Y_{mn}(3, 4) + A_3 Y_{mn}(3, 5) \quad (22)$$

and

$$SH_{6,NL} = SH_{3,NL} + 4A_4 Y_{mn}(4, 3)A_5 Y_{mn}(4, 4) \\ + A_6 Y_{mn}(4, 5). \quad (23)$$

We chose a definition of the cost function of the form given in Eq. (3). Estimates of the inverse data-covariance matrix  $\mathbf{C}^{-1}$  come from the 1500-yr-long control model integration. We have chosen to only use the diagonal elements of  $\mathbf{C}^{-1}$  that give the estimates of variance for each model grid point without consideration of the covariance between grid points or autocovariance in time. This information could be very helpful in extracting signals from noisy data; however, for purposes of the present demonstration, we chose to simplify the formulation of the data-covariance matrix and explore the effect of noise more explicitly later on. Also, as the target data for model experiments is generated from the surrogate climate model, we have not added any other components to  $\mathbf{C}^{-1}$  that would account for uncertainties in measurements or theory.

## 5. Selection of optimal parameters

The purpose of this section is to identify optimal choices of moves/temperature and  $n_{\text{target}}$  for different 1) levels of noise in model integrations, 2) number of model parameters, and 3) the degree of nonlinearity between model parameters.

The target ‘‘observations’’ in all experiments is obtained from one simulation from the surrogate climate model that has been embedded in noise representative of a climatology obtained from a 10-yr mean. Because we know the set of parameters that was used for this simulation, we can measure the accuracy of the VFSA algorithm to identify those parameter values. Figures 2a–c show the quantity ‘‘fractional error’’ (defined within figure caption), which indicates the accuracy of the VFSA algorithm to identify the parameter settings of the target observations for a surrogate climate model with three, six, or nine parameters. In general, increasing  $n_{\text{target}}$  decreases the error of VFSA estimates although the rate of improvement tends to decrease with larger values of  $n_{\text{target}}$ .

Focusing first on the linear model results (circles),

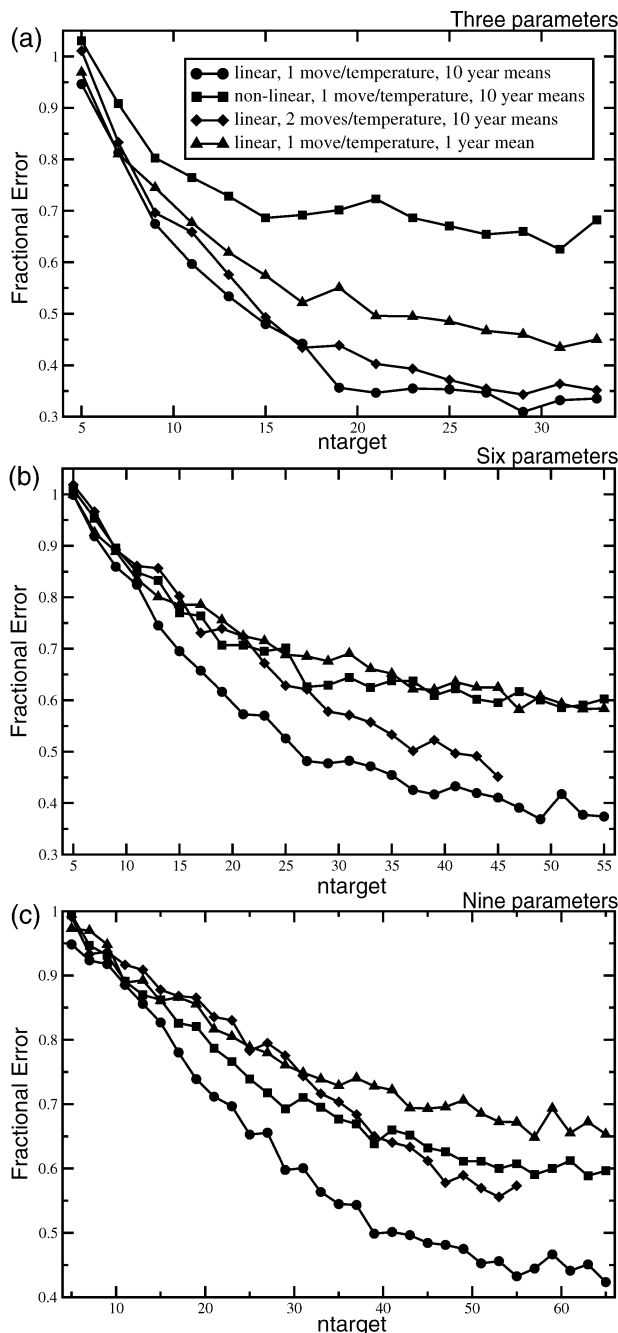


FIG. 2. Fractional error of the VFSA algorithm as a function of  $n_{\text{target}}$  for the surrogate climate model when searching over (a) three, (b) six, or (c) nine parameters. Fractional error is defined to be the average over all parameters of the ratio of errors that occur in identifying the target parameters for a given experiment relative to the errors that occur when moves/temperature = 1 and  $n_{\text{target}} = 5$ . This quantity is taken from the average of 300 VFSA convergence attempts. Results are shown for experiments that consider the linear model using one move/temperature and 10 yr means (circles), non-linear model using one move/temperature and 10-yr means (squares); and linear model using two moves/temperature [in (a)] or three moves/temperature [in (b) and (c)] and 10-yr means (diamonds); and the linear model using one move/temperature and 1-yr means (triangles).

the three-parameter problem parameters using moves/temperature = 1 and 10-yr means achieves  $\sim 90\%$  of the maximum accuracy (minimum fractional error) of all values of  $n_{\text{target}}$  tested by  $n_{\text{target}} = 19$ . In contrast, the corresponding experiments involving six or nine parameters do not achieve a similar improved accuracy until  $n_{\text{target}} = 37$  or  $n_{\text{target}} = 49$ , respectively. Experiments using the nonlinear model (squares) tend to achieve a similar degree of improved accuracy with a slightly smaller value of  $n_{\text{target}}$ . (Note that because the nonlinear model has two acceptable solutions for many of its parameters, the difference in fractional error between the linear and nonlinear model experiments does not necessarily reflect any difference in absolute errors).

The effect of noise can be observed from the comparison of experiments using 10-yr means (circles) with experiments using 1-yr means (triangles). The presence of additional noise within the surrogate climate model consistently degrades the ability of the VFSA algorithm to identify the target parameter values within a given number of model evaluations. However, the amount of noise does not strongly affect the  $n_{\text{target}}$  value at which VFSA would achieve most of its potential accuracy.

The results from experiments considering different choices of moves/temperature (diamonds) were contrary to our expectations. Increases in moves/temperature slow the rate of convergence of the VFSA algorithm and conceivably should improve its accuracy as was suggested by Ingber (1989). However, we found that increasing moves/temperature often made no difference and sometimes even degraded VFSA's accuracy for a given  $n_{\text{target}}$ . One possible explanation for this is that for this particular problem where the multidimensional cost function surface is relatively smooth with a clearly defined minimum, the VFSA can easily identify the general regions of the parameter space that minimize the cost function. However, the algorithm needs extra time to identify the best parameter combinations once it nears the cost function minimum. In our case, a small value of moves/temperature allowed it to converge quickly to a good neighborhood while increasing  $n_{\text{target}}$  gave the algorithm time to search out the best possible parameter combinations. Increasing the number of parameters being searched requires a larger  $n_{\text{target}}$  as well because it becomes increasingly difficult to find the global minimum when searching over more and more parameter combinations.

Figure 3a shows the average number of forward-model evaluations required for VFSA to converge as a function of  $n_{\text{target}}$  for the three-, six-, and nine-parameter problems using moves/temperature = 1 and 10-yr means. The steady linear increase of the number of model evaluations with increasing  $n_{\text{target}}$  indicates that the best trade-off between accuracy and cost will come from the  $n_{\text{target}}$  that provides most of the potential accuracy as determined from Figs. 2a–c. Figure 3a also shows that problems involving more parameters will require

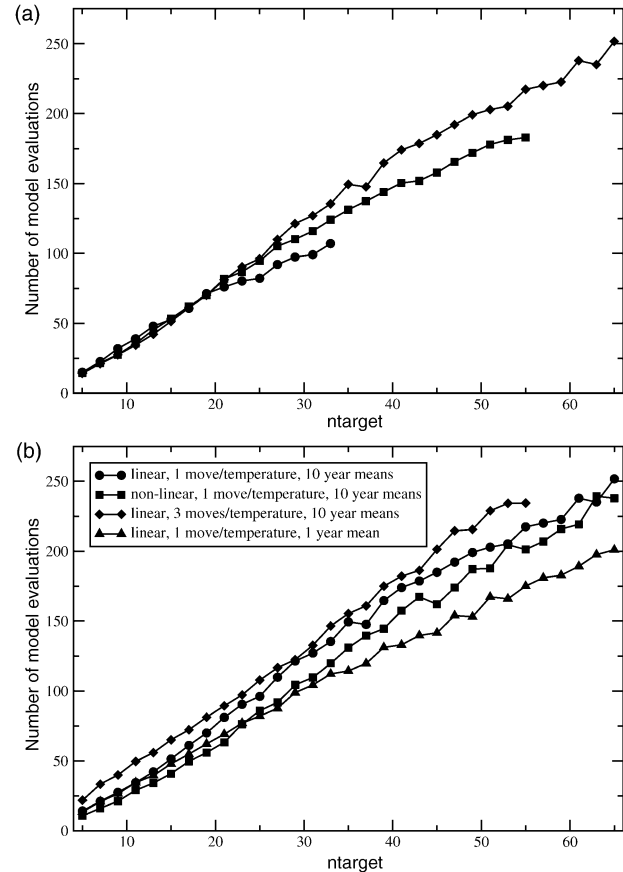


FIG. 3. Average number of model evaluations needed for convergence (among the 300 convergence attempts made for each data point) using the VFSA algorithm as a function of  $n_{\text{target}}$ . (a) Comparison of results for linear three-parameter problem (circles), six-parameter problem (squares), or nine-parameter problem (diamonds) using one move/temperature and 10-yr means. (b) Comparison of results for the nine-parameter problem with using one move/temperature and 10-yr means (circles), nonlinear model using one move/temperature and 10-yr means (squares), linear model using three moves/temperature and 10-yr means (diamonds), and the linear model using one move/temperature and 1-yr means (triangles).

more model evaluations to achieve convergence (for a given  $n_{\text{target}}$ ).

The effect of noise, nonlinearity, and changes in moves/temperature for the nine-parameter problem is shown in Fig. 3b. The results are similar for the three- and six-parameter problems. The model evaluations that include more noise (triangles) tend to require slightly fewer forward model evaluations than their less noisy counterpart (circles). This simply reflects the increased difficulty for the VFSA algorithm to identify better solutions when noise is disguising the signals. At least for this surrogate climate model, the nonlinear version (squares) requires slightly fewer model evaluations than its linear counterpart (circles). Also, increasing moves/temperature (diamonds) consistently requires more model evaluations to achieve convergence, which is as expected.

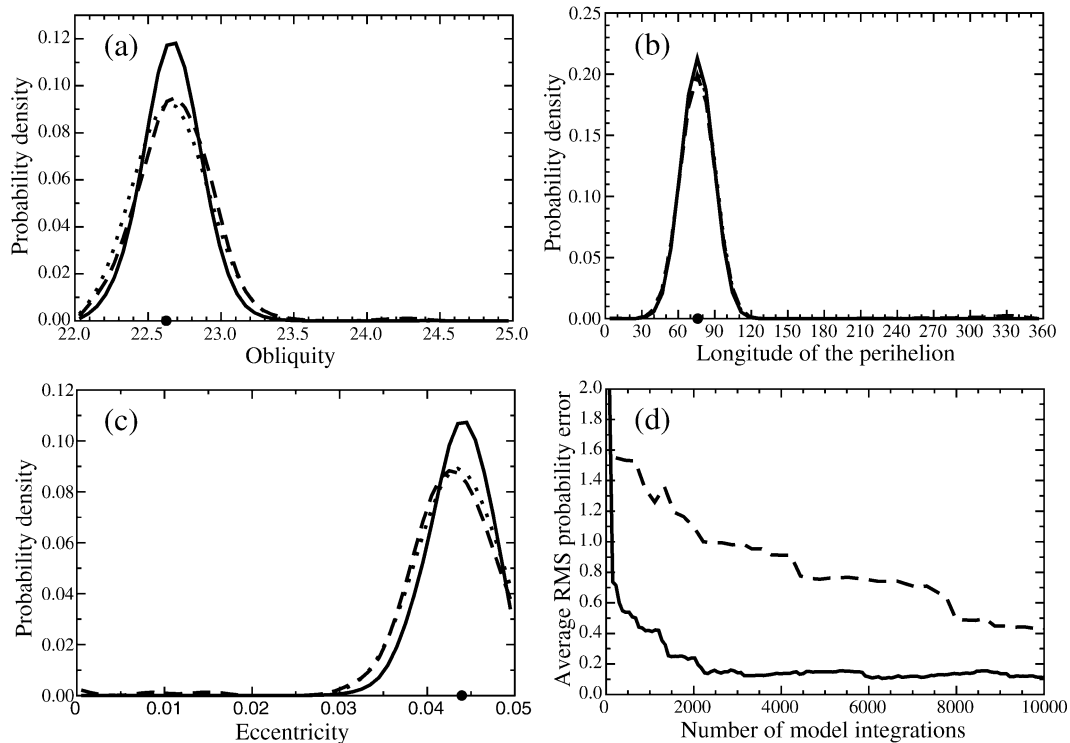


FIG. 4. Marginal probability density functions for parameters controlling (a) obliquity, (b) longitude of the perihelion, and (c) eccentricity based on VFSA (solid line), the Metropolis/Gibbs sampler (dashed line), and grid-search (dotted line) algorithms. (d) Interparameter average rms difference between the probability distributions based on an intermediate number of model evaluations and the final distribution for the VFSA (solid line) and the Gibbs sampler (dashed line). (Grid search requires 125 000 model evaluations to attain an average rms probability error of 0.) The large dot in (a)–(c) indicates the value of each parameter that was used to generate the target observations. The bounds on orbital parameter values are based on a realistic range of variation for these parameters over the past several million years.

Taking Figs. 2 and 3 together, the most accurate solutions that require the fewest forward model evaluations for the linear model with moves/temperature = 1 occur when  $n_{\text{target}} = 19$  for the three-parameter problem at an average cost of 71 forward-model evaluations,  $n_{\text{target}} = 37$  for the six-parameter problem at an average cost of 137 forward-model evaluations, and  $n_{\text{target}} = 49$  for nine-parameter problem at an average cost of 199 forward-model evaluations.

## 6. Cost of estimating parameter uncertainties

The second part of evaluating the cost of a parameter uncertainty analysis using BSI is the number of repeat convergence attempts that are required to create stable estimates of the multidimensional parameter probability density function. After comparing the cost of the deriving this function using multiple VFSA and the Metropolis/Gibbs sampler, we will evaluate how the amount of noise, number of model parameters being searched, and the degree of nonlinearity affect the number of required repeat convergence attempts and the overall cost of estimating parameter uncertainties.

### a. Multiple VFSA versus Metropolis/Gibbs sampler

To evaluate any systematic biases and the relative efficiency of the BSI parameter uncertainty estimation based on multiple VFSA relative to the Metropolis/Gibbs sampler, we will compare the rate of convergence of the multidimensional probability density functions for three, six, and nine parameters using both methods. Figure 4 shows the uncertainty in estimating the first three target parameters for a BSI analysis based on multiple VFSA (after 34 000 model evaluations) and the Metropolis/Gibbs sampler (after 100 000 model evaluations). For comparison, the distributions based on the grid search algorithm (125 000 model evaluations) are also shown and are quite similar to the results based on the Metropolis/Gibbs sampler. The probability density for the first three parameters is nearly identical for the cases when searching over six or nine parameters (not shown). There is a tendency for the distributions based on multiple VFSA to be biased relative to the distributions based on the Metropolis/Gibbs sampler. As the Metropolis/Gibbs sampling method itself does not include any inherent biases, the sharper peaks of the multiple-VFSA-based analysis show how this method can

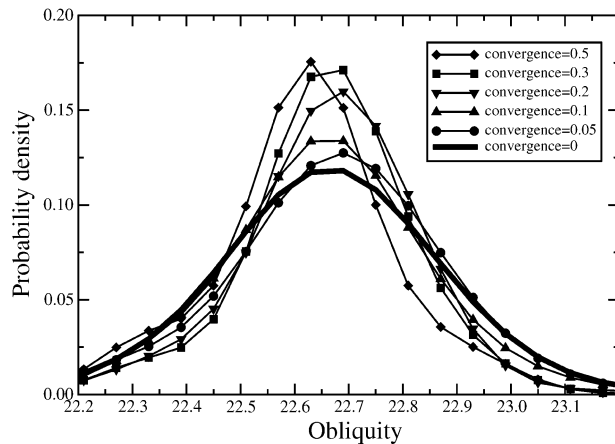


FIG. 5. Marginal probability density function for obliquity parameter derived at different levels of average rms probability error indicated in Fig. 4.

overestimate the relative likelihood of solutions near the target parameters and underestimate the width of the distributions. However, as one can surmise from Fig. 4d, the analysis based on multiple VFSA converges much more quickly than the analysis based on the Metropolis/Gibbs sampler. This measure of the relative efficiency is given by the average rms probability error in Fig. 4d and mathematically is given by the interparameter average of the root-mean-square difference between the smoothed probability distributions based on an intermediate number of model iterations and the final distribution. The convergence rate depends on the amount of smoothing or the number of bins used within the probability calculation. Each of the probability distributions presented here consists of 50 bins that have been smoothed through an averaging of the nearest neighbor bins a total of five iterations. Obviously this measure of PPD convergence would not ordinarily be available in advance of completing a large number of model evaluations. One could conceivably construct a running statistic of the change in the PPD and stop the parameter uncertainty analysis when this statistic indicates the PPD is not changing significantly.

As the total number of model evaluations needed to achieve convergence is relatively high, it would be desirable to identify a sufficient average rms probability error that represents an adequate representation of the final probability distribution. Figure 5 shows the transformation of the marginal probability density function for the obliquity parameter as we sample distributions obtained when the average rms probability error = 0.5, 0.3, 0.2, 0.1, and 0.05. The probability distributions based on fewer model evaluations (higher average rms probability error) are more sharply peaked than the final distribution. As shown in Fig. 4d, multiple VFSA is able to quickly achieve an average rms probability error of 0.2 in  $\sim 2100$  model evaluations. We therefore conclude that this distribution is the most efficient approx-

imation to the final probability distribution based on multiple VFSA. In this case, multiple VFSA is about 10 times as efficient as the Metropolis/Gibbs sampler (2100 versus 23 000 model evaluations). The relative efficiency of VFSA relative to the Metropolis/Gibbs sampler increases dramatically when considering problems involving more parameters (see section 6c). For instance, the number of model evaluations required to achieve an average rms probability error of 0.2 for a six-parameter problem is  $\sim 35$  times as great for the Metropolis/Gibbs sampler as for multiple VFSA. For a nine-parameter problem, multiple VFSA is 68 times as efficient. As will be discussed further below, the relative efficiency of multiple VFSA to the Metropolis/Gibbs sampler depends on the degree of nonlinearity.

#### b. Effect of noise

The level of noise in model simulations can be controlled directly by choosing the number of model years to integrate the model for each model experiment. We test the effect of using 1-, 3-, 5-, and 10-yr means for individual model integrations on the convergence rate of the multidimensional PDF.

Figures 6a–c shows that the final probability distributions are very similar between parameter uncertainty analyses based on experiments of different lengths. However, Fig. 6d indicates that it takes more model evaluations to reach the statistically stable distribution for experiments based on fewer model years. Even so, the total number of model years required to reach an average rms probability error of 0.2 is not as high for experiments based on 1-, 3-, or 5-yr means as it is for experiments based on 10-yr means (8.3, 12, 17.5 and 21 kyr for experiments based on 1-, 3-, 5-, and 10-yr means, respectively). Figure 6, however, does not adequately express the difficulty in identifying parameter sets that are within the 95% confidence interval of the target observations (see section 2b about the definition of this confidence interval). It is these parameter sets that are the most helpful in expressing the uncertainty and nonlinear dependencies between model parameters. It is also these parameter sets that are used to define the members of an ensemble that represent the combined uncertainty in these parameters on model predictions. By the time the analyses based on 10-yr-mean experiments achieved an average rms probability error of 0.2, there were 132 parameter sets that were within the 95% confidence interval of the target. In contrast there were only 116 parameter sets for analyses based on 5-yr-mean experiments, 38 parameter sets for analyses based on 3-yr means, and only 5 model sets for analyses based on 1-yr means. The best balance between the accuracy and cost of estimating parameter uncertainties would appear to be achieved with an analysis based on experiments with a signal-to-noise ratio of  $\sim 5$  or greater. In our case, this signal to noise ratio occurs with experiments using 5-yr means or greater. The accuracy and overall cost of

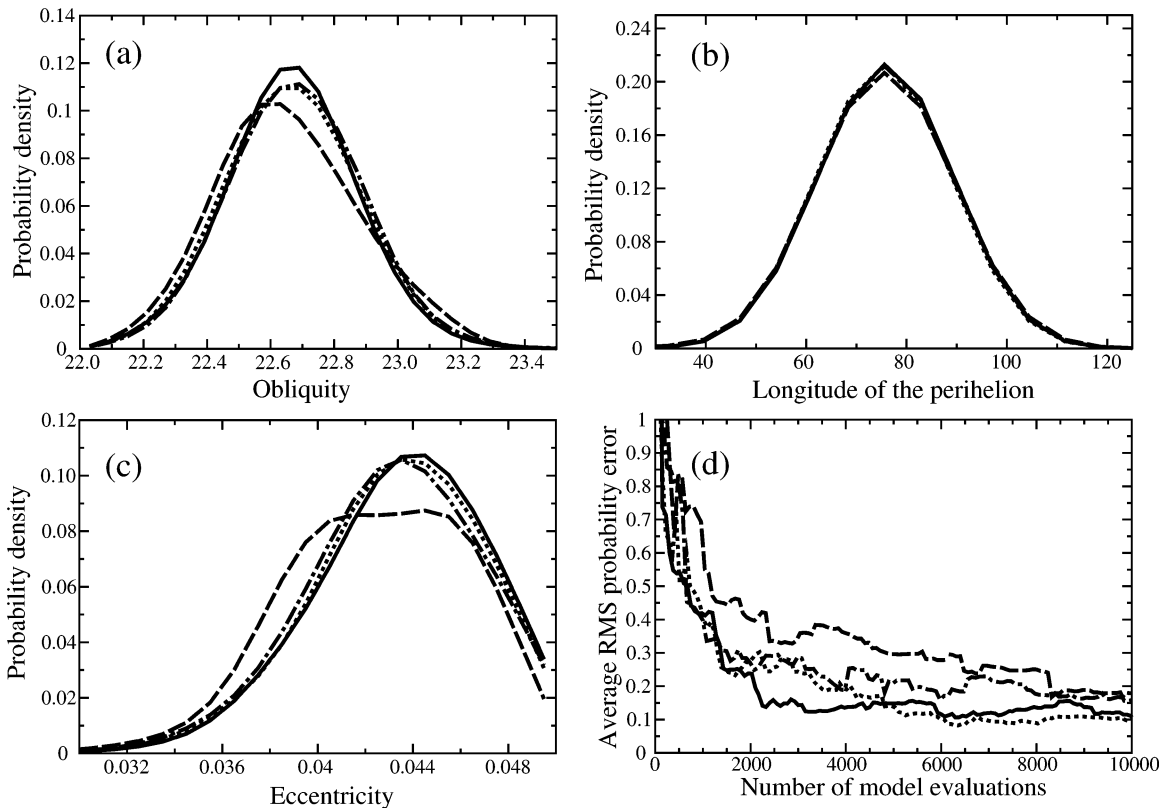


FIG. 6. (a)–(c) Marginal probability density functions based on experiments using 10- (solid), 5- (dotted), 3- (dash-dotted), and 1-yr (dashed) means. (d) As in Figure 4d, but for the above.

estimating parameter uncertainties based on 10-yr means is only about 20% greater as compared with that based on 5-yr means.

### c. Number of model parameters

The cost of estimating parameter uncertainties should scale to some degree with the number of model param-

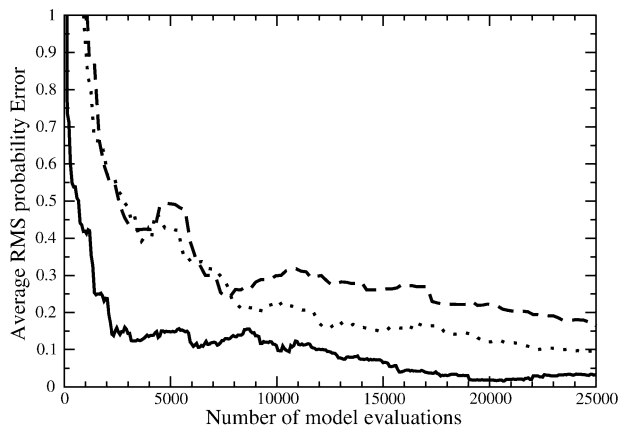


FIG. 7. Average rms probability error indicates the approach to equilibrium of the multidimensional probability distributions for problems with three (solid line), six (dotted), and nine (dashed) parameters.

eters being searched. That is, the work involved in describing the multidimensional volume of the PPD defined by Eq. (4) could grow exponentially with the number of degrees of freedom that are explored. However, all parameters are not equally important. Those parameters that affect the model solution strongly will essentially take up less volume and require fewer model evaluations to identify the most sensible solutions. Therefore, the cost of estimating parameter uncertainties based on importance sampling techniques can scale more favorably relative to purely random (Monte Carlo) or grid search algorithms that are insensitive to the effective degrees of freedom. Although the Metropolis/Gibbs sampler is an example of an importance sampling technique, it is not as sensitive as multiple VFSA to a problem's effective degrees of freedom. We consider next the impact of searching over three, six, or nine parameters within a parameter uncertainty analysis of the surrogate climate model.

Figure 7 shows the average rms probability error as a function of the number of model evaluations. The number of model evaluations required to achieve an average rms probability error of 0.2 is  $2.1 \times 10^3$ ,  $12 \times 10^3$ , and  $22 \times 10^3$  for three-, six-, and nine-parameter problems, respectively. The probability distributions for the parameters in common (obliquity, longitude of the perihelion, and eccentricity) are nearly identical, with a

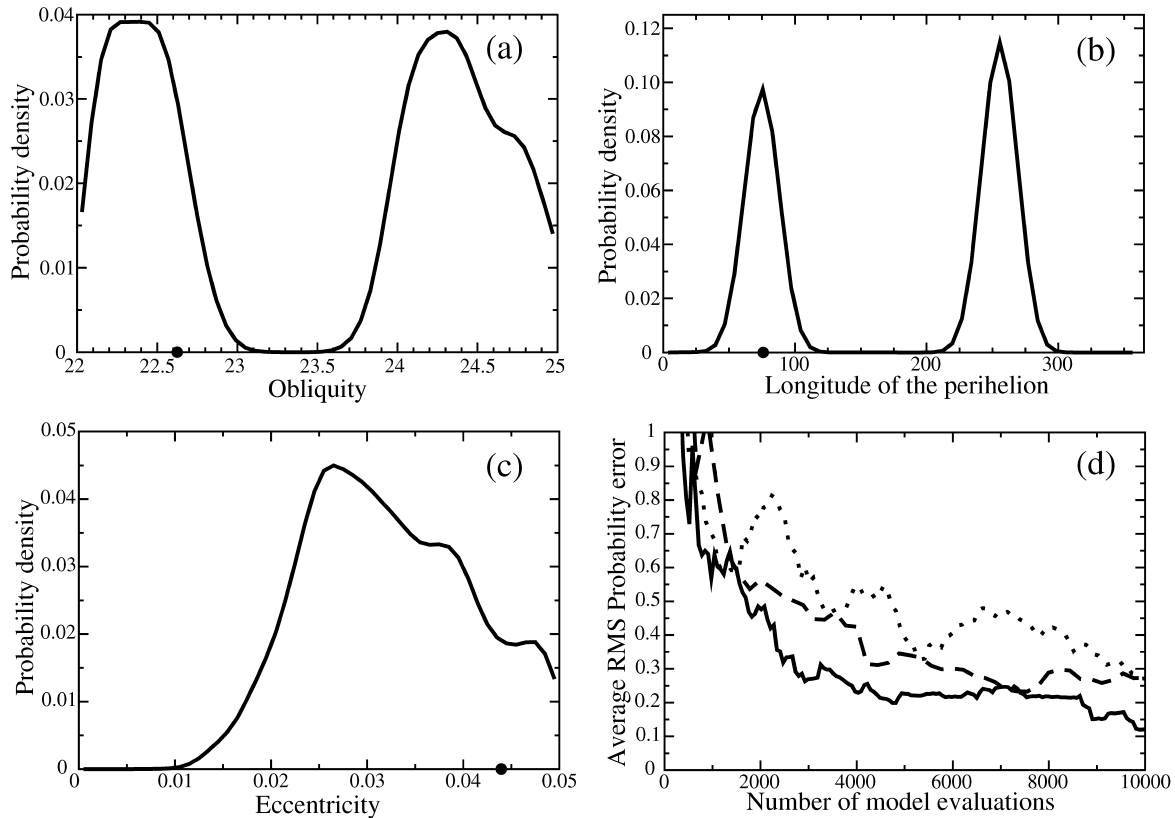


FIG. 8. Marginal probability density functions for parameters controlling (a) obliquity, (b) longitude of the perihelion, and (c) eccentricity for the three-parameter nonlinear surrogate climate model based on Eq. (20) and setting  $n_{\text{target}} = 15$ . (d) As in Fig. 4d and shows the approach to statistical stability for an analysis based on multiple VFSA with  $n_{\text{target}} = 15$  (solid line), VFSA with  $n_{\text{target}} = 19$  (dotted line), and the Metropolis/Gibbs sampler (dashed line). The probability distributions resulting from analyses with  $n_{\text{target}} = 15$  and the Metropolis/Gibbs sampler are not significantly different from what is shown in (a)–(c). The large dot in (a)–(c) indicates the value of the each parameter that was used to generate the target observations.

slight tendency for the six- and nine-parameter problems to be slightly more peaked than the three-parameter problem (not shown).

#### d. Degree of nonlinearity

According to section 5, there is not a significant difference in the number of model evaluations required to find the minimum of the cost function for the linear and nonlinear surrogate climate models. We consider now whether the degree of nonlinearity (i.e., the size of the effect of nonlinearity) affects the number of model evaluations required to map the multidimensional PPD. We also discuss the ability of the parameter uncertainty analysis to quantify relationships between parameters as can be measured through parameter correlations.

Figures 8a–c shows the marginal probability distributions for the nonlinear three-parameter surrogate climate model based on multiple VFSA using  $n_{\text{target}} = 15$ . For the parameters obliquity and longitude of the perihelion, there now are two peaks in the probability distributions indicating two separate ranges of parameter values that would allow the model to fit within the

uncertainty of the target observations. The most likely solutions indicated by the peak in the marginal PPDs for all parameters are not identical to the parameter values used to generate the target observations (indicated by a large dot). The different shaping of the probability distributions is an inherent effect of the model nonlinearities and the fact that these marginal probability distributions are one-dimensional projections of a multidimensional distribution. Very similar distributions result from an analysis based on the Metropolis/Gibbs sampler (not shown) using the same target. Also, the distributions were not significantly affected by the presence of noise within the target observations.

Figure 8d indicates the rate at which the probability distributions approach statistical stability for analysis based on VFSA using  $n_{\text{target}} = 15$  and 19 as well as an analysis based on the Metropolis/Gibbs sampler. We included the experiment with  $n_{\text{target}} = 15$  because that  $n_{\text{target}}$  value was the most cost effective at identifying the optimal solutions for the nonlinear model (Fig. 2). Indeed, this choice outperforms the analysis that uses  $n_{\text{target}} = 19$ , the value that was used for the more linear climate model. Overall, the presence of nonlinearities

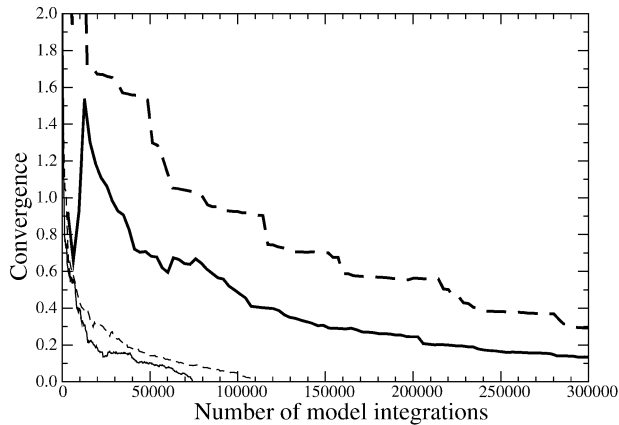


FIG. 9. The approach to statistical stability for analyses based on multiple VFSA (thin curves) and the Metropolis/Gibbs sampler (thick curves) for the nonlinear six-parameter case (continuous line) and the nonlinear nine-parameter case (dashed line). The definition of convergence is similar to previous figures.

strongly affects the cost of estimating parameter uncertainties. Depending on how one measures an equivalent degree of convergence (and not just focusing on where the lines cross the average rms probability error = 0.2), the nonlinear climate model requires 2–3 times as many model evaluations as the linear version.

The remarkable finding within Fig. 8d is that the Metropolis/Gibbs sampler for the nonlinear three-parameter problem is as efficient as the analyses based on multiple VFSA. In fact, the Metropolis/Gibbs sampler is nearly 2 times as efficient here as for the Metropolis/Gibbs sampler analysis with the linear three-parameter climate model (Fig. 4d) and about 4 times more efficient at identifying acceptable parameter sets (sets within the 95% confidence interval). We do not fully understand why the Metropolis/Gibbs sampler was so effective in the three-parameter problem. One possibility is that the step size that was chosen for the Metropolis/Gibbs sampler was better suited for the problem characteristics (more on this in the following section). For the six and nine nonlinear parameter cases, however, multiple VFSA was able to equilibrate much more quickly than the Metropolis/Gibbs sampler (Fig. 9). In these cases multiple VFSA was found to be ~12–17 times as efficient, respectively, as the Metropolis/Gibbs sampler.

## 7. Discussion

The effectiveness of any given sampling strategy depends on the appropriateness of the size of the steps used in sampling parameter space (Gelman et al. 2003). The rules governing the Metropolis/Gibbs sampler only permit fixed-size steps. In contrast, the multiple-VFSA algorithm permits a variable step size as controlled by the temperature parameter that acts to focus parameter space sampling near regions of interest (i.e., the peaks of the PPD). Thus, multiple VFSA avoids in part the

ambiguity of what the ideal step size should be for any given problem. Although there are other algorithm-specific parameters within multiple VFSA that need to be specified (such as  $n_{\text{target}}$  and moves/temperature), the ideal values for these parameters is primarily sensitive to broader characteristics of a problem such as its dimensionality and not whether nonlinearities are important. The primary advantage of the Metropolis/Gibbs sampler is that it is an example of a Monte Carlo Markov chain (MCMC). Any sampling algorithm that satisfies the properties of a Markov chain has been proven to provide asymptotically robust measures of statistical inference (like the PPD). We have shown that multiple VFSA will tend to overemphasize the peaks of the PPD and thus underestimate parameter uncertainties. Although much research has taken place over the past decade on how different sampling rules within MCMC may be adapted for different types of problems [for the most recent and basic developments see Gilks et al. (1996), Tanner (1996), Gamerman (1997), and Robert and Casella (1999)], more work is needed for problems that involve large numbers of parameters, costly model evaluations, or unusually large dataset sizes.

## 8. Conclusions

We have examined factors affecting the computational cost of identifying the optimal parameter values and their uncertainty through joint probability distributions using a realistic surrogate climate model. We discussed how the Bayesian stochastic inversion method based on multiple very fast simulated annealing is able to strike a balance between the dual objectives of identifying parameter sets that are within prescribed uncertainty limits and estimating the multidimensional parameter probability distribution. With minimal biases, analyses of three to nine parameters using an analysis based on multiple VFSA are as much as 10–68 times as efficient as the Metropolis/Gibbs sampler in estimating the multidimensional probability distribution, depending on the number of parameters and the extent of the model nonlinearities. At the same time multiple VFSA is as much as one to two orders of magnitude more efficient than the Metropolis/Gibbs sampler at identifying model parameter sets that are within the uncertainty estimates of the target observations.

Surprisingly, the ability to identify optimal parameters was not found to be sensitive to the cooling rate within the VFSA algorithm as had been previously assumed. Rather the most important parameter affecting the accuracy of VFSA is  $n_{\text{target}}$ , the parameter that specifies how long the algorithm searches near a good solution before the search is terminated. The numbers of parameters being investigated primarily dictate the best choice for  $n_{\text{target}}$ . Also, the choice of  $n_{\text{target}}$  is fairly insensitive to the amount of noise included within individual model experiments and the degree of model nonlinearity.

The amount of noise within individual model experiments and degree of model nonlinearity affects the total number of model evaluations that are required to have a stable estimate of parameter probability distributions as well as the relative efficiency in identifying parameter sets within the uncertainty of the target observations. We recommend parameter uncertainty analyses be conducted with experiments that have signal-to-noise ratios of 5 or larger. At worst, model nonlinearity only doubled the number of model experiments required to achieve statistically stable estimates of the PPD.

*Acknowledgments.* This work has been supported by the G. Unger Vetlesen Foundation and the University of Texas Institute for Geophysics. This paper has also benefited from discussions with Gabriel Huerta and two anonymous reviewers.

## REFERENCES

- Allen, M., 1999: Do-it-yourself climate prediction. *Nature*, **401**, 642.
- , P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, **407**, 617–620.
- Boer, G. J., 1992: Some results from an intercomparison of the climates simulated by 14 atmospheric general circulation models. *J. Geophys. Res.*, **97**, 12 771–12 786.
- Bowman, K. P., J. Sacks, and Y.-F. Chang, 1993: On the design and analysis of numerical experiments. *J. Atmos. Sci.*, **50**, 1267–1278.
- Chapman, W. L., W. J. Welch, K. P. Bowman, J. Sacks, and J. E. Walsh, 1994: Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *J. Geophys. Res.*, **99** (C1), 919–935.
- Covey, C., K. M. Achutarao, U. Cubasch, P. Jones, S. J. Lambert, M. E. Mann, T. J. Phillips, and K. E. Taylor, 2003: An overview of results from the coupled model intercomparison project. *Global Planet. Change*, **37**, 103–133.
- Cubasch, U., and Coauthors, 2001: Projections of future climate change. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, J. T. Houghton et al., Eds., Cambridge University Press, 525–582.
- Forest, C., M. R. Allen, P. H. Stone, and A. P. Sokolov, 2000: Constraining uncertainties in climate models using climate change detection techniques. *Geophys. Res. Lett.*, **27**, 569–572.
- , —, A. P. Sokolov, and P. H. Stone, 2001: Constraining climate model properties using optimal fingerprint detection methods. *Climate Dyn.*, **18**, 277–295.
- , P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster, 2002: Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, **295**, 113–117.
- Gamerman, D., 1997: *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, 245 pp.
- Gates, W. L., and Coauthors, 1999: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.*, **80**, 29–56.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2003: *Bayesian Data Analysis*. 2d ed. Chapman and Hall, 668 pp.
- Geman, S., and D. Geman, 1984: Stochastic relaxation, Gibbs' distribution and Bayesian restoration of images. *IEEE. Trans. Pattern An. Mech. Intell.*, **6**, 721–741.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, Eds., 1996: *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 486 pp.
- Ingber, L., 1989: Very fast simulated re-annealing. *Math. Comput. Modell.*, **12**, 967–973.
- Jackson, C., and A. J. Broccoli, 2003: Orbital forcing of Arctic climate: Mechanisms of climate response and implications for continental glaciation. *Climate Dyn.*, **21**, 539–557.
- , Y. Xia, M. Sen, and P. Stoffa, 2003: Optimal parameter and uncertainty estimation of a land surface model: A case example using data from Cabauw, Netherlands. *J. Geophys. Res.*, **108**, 4583, doi:10.1029/2002JD002991.
- Joussaume, S., and K. E. Taylor, 2000: The Paleoclimate Modeling Intercomparison Project. *Paleoclimate Modeling Intercomparison Project (PMIP): Proceedings of the Third PMIP Workshop, Canada, 4–8 October 1999*, P. Braconnot, Ed., World Meteorological Organization, WCRP-111, WMO/TD-1007, 271 pp.
- Kirkpatrick, S., C. D. Gelatt Jr., and M. P. Vecchi, 1983: Optimization by simulated annealing. *Science*, **220**, 671–680.
- McAvaney, B. J., and Coauthors, 2001: Model evaluation. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, J. T. Houghton et al., Eds., Cambridge University Press, 472–523.
- Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, 1953: Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
- Pitman, A. J., and A. Henderson-Sellers, 1998: Recent progress and results from the project for the intercomparison of land surface parameterization schemes. *J. Hydrol.*, **213**, 128–135.
- Robert, C. P., and G. Casella, 1999: *Monte Carlo Statistical Methods*. Springer Verlag, 507 pp.
- Rothman, D. H., 1986: Automatic estimation of large residual static corrections. *Geophysics*, **51**, 332–346.
- Sacks, J., S. B. Schiller, and W. J. Welch, 1989: Designs for computer experiments. *Technometrics*, **31**, 41–47.
- Sen, M., and P. L. Stoffa, 1996: Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion. *Geophys. Prospect.*, **44**, 313–350.
- Tanner, M. A., 1996: *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 2d ed. Springer Verlag, 204 pp.
- Tarantola, A., 1987: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation*. Elsevier, 613 pp.
- Welch, W. J., R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, 1992: Screening, predicting, and computer experiments. *Technometrics*, **34**, 15–25.
- Williams, K. D., C. A. Senior, and J. F. B. Mitchell, 2001: Transient climate change in the Hadley Centre models: The role of physical processes. *J. Climate*, **14**, 2659–2674.