Non-Parametric Sampling Approximation via Voronoi Tessellations

Alejandro Villagran, Gabriel Huerta, Marina Vannucci, Charles S. Jackson, and Alvaro Nosedal [1]

Abstract

In this paper we propose a novel non-parametric sampling approach to estimate posterior distributions from parameters of interest. This technique is particularly suited for models that are computationally expensive to evaluate. Starting from an initial sample over the parameter space, this method makes use of this initial information to form a geometrical structure known as Voronoi tessellation over the whole parameter space. This rough approximation to the posterior distribution provides a way to generate new points from the posterior distribution without any additional costly model evaluations. By using a traditional MCMC over the non-parametric tessellation, the initial approximate distribution is refined sequentially, allowing to sample new points at any moment. We applied this method to a couple of climate models to show that this hybrid scheme successfully approximates the posterior distribution of the model parameters without any additional forward evaluation of the model itself. The results obtained could be used not only to improve the estimation of the posterior distribution of the climate model parameters but also to search parameter space regions ideal for optimization of models in which the limitation of computational resources is a challenge.

Keywords: Non-parametric approximation, Voronoi tessellation, Parameter estimation, Climate Models.

[1]Alejandro Villagran is Senior Predictive Modeler, PEMCO Insurance. Gabriel Huerta is Professor of Mathematics and Statistics, The University of New Mexico. Marina Vannucci is Professor of Statistics, Rice University. Charles S. Jackson is Research Scientist at Institute for Geophysics, University of Texas at Austin. Alvaro Nosedal is postdoctoral associate at The University of New Mexico. Corresponding author: ghuerta@stat.unm.edu

1 INTRODUCTION

In many scientific fields, physical systems $\mathbf{T}(\cdot)$ can be approximated by complex computer models referred to as forward operators $\mathbf{g}(\cdot)$. For instance, weather forecasting uses highly sophisticated models for atmospheric pressures and humidities; the behavior of large engineering structures is typically modeled in great details as an aid to their design; astronomers and physicists have long required massive computations to model and predict movements of planets. Some cited applications in model calibration are hydrocarbon reservoir, Craig et al. (2001), photo emission computed tomography, Higdon et al. (2003), thermohaline circulation in the Atlantic, Goldstein and Rougier (2006), cosmology, Heitmann et al. (2006), and so forth. A feature of such models is that they generally require substantial amounts of computing time, even on super computers. When it is necessary to use many model runs to compute the output over a range of model configurations, the time required for each run becomes extremely relevant.

Calibration is traditionally seen as a process of fitting the computer model to the observational data, by adjusting the uncertain inputs until the model predicts those data as closely as possible. The first Bayesian approach to calibration is given by Craig et al. (1996), employing Bayes linear methods, and this is extended to a fuller treatment of calibration and model uncertainty in Craig et al. (2001). Kennedy and O'Hagan (2000) consider prediction and uncertainty analysis for complex computer codes using a Bayesian approach in which prior beliefs about the codes are represented in terms of Gaussian processes (GP). Other examples of statistical emulators based on GP are Sansó et al. (2008), and Christen and Sansó (2011). A different approach for calibration is presented in Bliznyuk et al. (2008), who construct a radial basis function interpolant of the logarithm of the posterior density based on an optimization strategy. A completely distinct approach drawn from the geophysical literature (Sambridge, 1999) is **applied in this paper and compared to other statistical computational approaches. The novelty of the method is that it is fully non-parametric and it only relies on a geometrical representation of the data called Voronoi tessellation (Voronoi, 1908). The method is conceptually simple, but is able to exhibit complex self-adaptive behavior in searching over a parameter space. Unlike previous methods, this technique makes use of some simple geometrical structures which we show can also be used to enhance existing search methods.**

The backbone of the method can be summarized in five main steps: Step (1) uses an adaptive sampling strategy to gather information from the posterior distribution of the parameters of interest; Step (2) employs the initial evaluations obtained in step (1) to build the Voronoi tessellation; Step (3) draws a sample from the approximate surface constructed in step (2) by using a Markov Chain Monte Carlo (MCMC) scheme; Step (4) adds the new points sampled in step (3) to the original data, and rebuild the tessellation; Step (5) repeats iteratively steps (2) through (4). **According to Sambridge (1999), this algorithm provides the advantage of generating new points without any additional computationally expensive evaluations of the computer model g. Potentially this technique could be applied to any computer model in which direct evaluation of the model is expensive. We test our method in a couple of climate model applications. In difference to Sambridge (1999), we evaluate the method when initial samples are produced from an Adaptive Metropolis or with a grid/factorial design instead of relying of using an stochastic optimizer such as simulated annealing.**

The organization of the article is as follows: Section 2 explains the algorithm in detail and it provides a statistical example with detailed steps to illustrate our methodology. Section 3 provides a couple of applications in climate model calibration, and presents the results obtained by using the proposed non-parametric approach. Section 4 discusses advantages, drawbacks, and alternative approaches to the method presented in this paper.

2 METHOD

In the method we are presenting, $\mathbf{g}$ is evaluated only a few times in order to explore the parameter space and to give an initial approximate posterior probability distribution (PPD) of the parameter vector $\mathbf{m}$. This first approximation is represented geometrically by a structure known as Voronoi cells (Voronoi, 1908). These are the nearest neighborhood regions defined all over the parameter space that will be used to sample new values of $\mathbf{m}$ employing a traditional MCMC. Once we have a new set of points, these will be added to the initial sample to regenerate the Voronoi cells. This procedure will be repeated several times to enhance the initial surface and therefore to improve the marginal probability distributions of $\mathbf{m}$ without any additional costly forward evaluations of $\mathbf{g}$.

2.1 ALGORITHM

The main goal of this technique is to deal with computational limitations (i.e., evaluations of the forward operator **g**) in computer models by using a completely non-parametric approach. This technique is being called hybrid-regeneration since it allows to effectively convert samples gathered during a typical computer model run into a form that can be used as a surface approximation. Once the sampling densities have been defined over the Voronoi structure, the next step is to refine the surface approximation by running a Gibbs sampler on the initial surface. The new points sampled will regenerate the surface. The method is summarized by the following steps:

1. Initial Sampling Strategy: Run **an efficient** sampler method over the parameter space, and save all the sampled model vectors and their likelihood values.

2. Voronoi Tessellation: Using the initial samples $(\mathbf{m_1}, ..., \mathbf{m_n})$ construct the Voronoi cell structure and create the approximate PPD, $\hat{\pi}(\mathbf{m}|\mathbf{d_{obs}})$, where d_{obs} is the observed data.

3. Sampling new points: Run a traditional Gibbs sampler on the approximate PPD and extract n^* new points.

4. Regeneration: Add the new points to the initial sample and regenerate the surface.

5. Repeat (2)-(4) many times.

The accuracy of the results depends mostly on the quality of the initial surface provided by step (1), which means that, the sampler must gather enough points from the entire d-dimensional space to allow the hybrid algorithm to construct an adequate d-dimensional mesh. This is why we propose to use an adaptive sampling strategy that does not require too much tuning like for example, the Metropolis-Hastings algorithm. Any serious biases in the initial sample could be inherited in the subsequent approximation of the PPD. Since new evaluations of the computer model and the posterior distribution are forbidden then steps (2) through (4) will aid by generating new points in regions where there is higher posterior density without additional cost.

2.1.1 INITIAL SAMPLING STRATEGY

In order to provide a sampling scheme that reaches a balance between efficiency and precision without the use of a burn-in period we consider the use of adaptive methods. Haario et al. (2001) suggested a method called Adaptive Metropolis (AM) that basically updates the proposal distribution with the knowledge learned that far about the target distribution. This is a non-Markovian algorithm that has the correct ergodic properties. Haario et al. (2004), and Villagran et al. (2008) applied these methods in gas profile inversion and climate calibration respectively to overcome the problems that a traditional MCMC (Metropolis (1953), Hastings (1970), Geman and Geman (1984), Sen and Stoffa (1996)) may have in selecting or tuning an effective proposal distribution.

Adaptive Metropolis (AM) algorithm

Suppose that at time $t-1$ we have sampled the states $m^{(0)}, ..., m^{(t-1)}$ where $m^{(0)}$ is a $d-$dimensional vector representing the initial state. Then a candidate point z is sampled from the proposal distribution $q_t(\cdot|m^{(0)}, ..., m^{(t-1)})$, which now may depend on the whole history. The candidate z is accepted with probability,

$$\alpha(m^{(t-1)}, z) = min\left(1, \frac{\pi(z)}{\pi(m^{(t-1)})}\right),$$

in which case we set $m^{(t)} = z$, and otherwise $m^{(t)} = m^{(t-1)}$.

The proposal distribution $q_t(\cdot|m^{(0)}, ..., m^{(t-1)})$ employed in the AM algorithm is a multivariate Gaussian distribution with mean at the current point $m^{(t-1)}$ and covariance matrix W_t. The matrix W_t is computed using the sampled covariance matrix of the parameters up to time t. The crucial aspect regarding the adaptation is how the covariance of the proposal distribution depends on the history of the chain. In the AM algorithm this is solved by setting $W_t = s_d C_{t-1} + s_d \epsilon I_d$ after an initial period, where s_d is a positive constant that depends only on the dimension d and $\epsilon > 0$ is a constant that we may choose to be very small. The role of ϵ is to ensure that C_t does not become singular. Let us define the matrix $\mathbf{M_t}$ with dimensions d by t as the matrix of sampled values up to t.

$$\mathbf{M_t} = \begin{pmatrix} m_1^{(1)} & m_1^{(2)} & \dots & m_1^{(t)} \\ \vdots & \vdots & \vdots & \vdots \\ m_d^{(1)} & m_d^{(2)} & \dots & m_d^{(t)} \end{pmatrix}$$

We select an initial time t_0 for the length of an initial period and define

$$W_t = \begin{cases} C_0, & \text{if } t \leq t_0 \\ s_d C_{t-1} + s_d \epsilon I_d, & \text{if } t > t_0 \end{cases}$$

We can compute the covariance matrix for time $t \leq t_0$ as

$$C_t = \frac{1}{t-1} \mathbf{M_t} \Big(I_t - \frac{1}{t} 1_t 1_t' \Big) \mathbf{M_t}',$$

where I_t is an identity matrix of size t, and 1_t is a row vector of ones with length t. To avoid too much computational cost we can use recursive formulas for the mean and the covariance. Then we can easily define the recursion for the vector of means as

$$\overline{m}_t = \frac{t-1}{t} \overline{m}_{t-1} + \frac{1}{t} m^{(t)}$$

and for $t > t_0$. The recursion for the covariance matrix is,

$$C_t = \frac{t-2}{t-1} C_{t-1} + \overline{m}_{t-1} \overline{m}_{t-1}' + \frac{1}{t-1} \Big[m^{(t)} m^{(t)'} - t \overline{m}_t \overline{m}_t' \Big].$$

The AM chain defined above simulates properly the target distribution π: for any bounded and measurable function $f : S \to \Re$, it holds almost surely that

$$\lim_{n \to \infty} \frac{1}{n+1} \Big(f(m_0) + f(m_1) + \cdots + f(m_n) \Big) = \int_S f(m) \pi(dm).$$

For a detailed proof of this result, see Haario et al. (2001).

2.1.2 VORONOI TESSELLATION

Voronoi cells have been known for several years (Voronoi, 1908). They have been rediscovered in a number of different fields and are primarily studied in the field of computational geometry. A survey of their many properties and algorithms to compute them can be found in Okabe et al. (2000). Recently they have been used in geophysical calibration problems (Sambridge, 1999). For any distribution of points, Voronoi cells are defined as the region about each point which is closer to that point than any other point or the nearest neighbor region of each point. In two dimensions, the cells are polygons whose edges are perpendicular bi-sectors between pairs of nodes, in three dimensions they are convex polyhedra, and in higher dimensions they are convex polytopes.

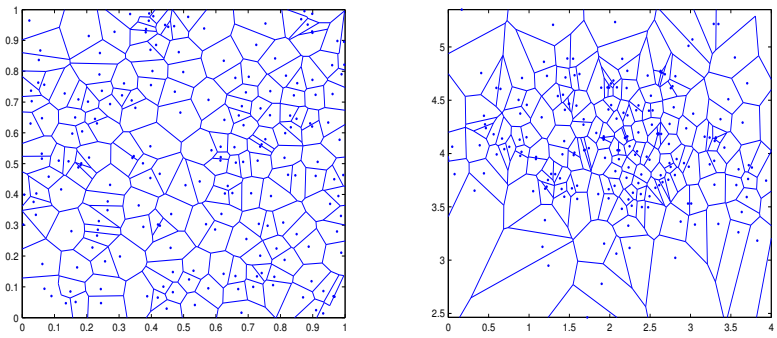

Figure 1: Voronoi tessellation from two different bivariate distributions. Top left: Uniform. Top right: Normal. Both surfaces were generated using 200 sampled points.

As formal definition of the Voronoi cells we have that, $\Omega = \{m_1, ..., m_n\}$ is a set of points in d dimensions, where $2 \leq n < \infty$, and let $m_k \neq m_j$ for $k \neq j$. The Voronoi cell around point m_k is given by

$$V(m_k) = \big\{x : ||x - m_k|| \leq ||x - m_j||, k \neq j\big\}.$$

Even though Voronoi cells are conceptually simple, they possess powerful properties. Regardless of the dimension, or how uneven or anisotropic the points distribution may be, the Voronoi cells always form local neighborhoods about each point, whose size (area, volume, etc) automatically grows, shrinks or changes shape depending on the local point density. This means that they can be used to produce an approximation to the PPD in the entire parameter space, i.e., by setting the value of the PPD in each cell to be equal to the known PPD at the point defining that cell (Okabe et al., 2000). This is called neighborhood approximation. Using this approximation we have that the density at each point is given by the reciprocal of the Voronoi cell volume. In Figure 1, we can appreciate the difference in the shape of the surface and size of the cells given a sample of 200 points coming from a uniform distribution, and a Gaussian distribution in $2D$.

2.2 REGENERATION

Since the neighborhood approximation is completely space filling, and is uniquely defined for any number of points, it may be used for sampling in the following way. After generating a set of n points (grid, uniform,

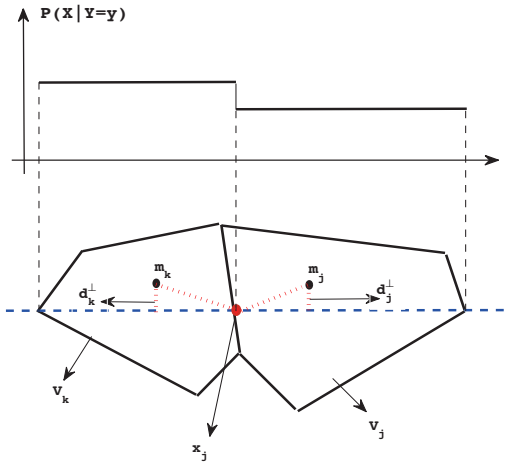

Figure 2: Two adjacent Voronoi cells, V_k and V_j, intersecting a conditional (dotted) line $Y = y_o$. On top, the conditional approximate density ($P(X|Y = y_o)$), which is constant between each cell boundary. The change in density is given by the point x_j.

MCMC), we can generate a new set of n^* samples so that they are independently distributed according to the current neighborhood approximation of the PPD. The new samples will be concentrated in regions of the parameter space which have higher density, as given by the current best estimate of the PPD based on all available information. We then update the approximation and generate n^* additional samples. As iterations proceed, the PPD approximation will be refined and each new generation will be more concentrated in regions with higher density than regions with lower density. Each new population is directly guided by all previous samples through the approximate PPD. The procedure described above can be implemented by using a traditional Gibbs sampler (Geman and Geman, 1984) over the Voronoi tessellation. Given any arbitrary value, we can compute the discrete marginal probability density inside and across all cells. This step can be done in many ways since it is a discrete piecewise distribution, however, a refined approach (Sambridge, 1999) is suggested in order to avoid simple brute force approaches. Without loss of gener-

ality, let us consider the case when the dimensionality of the parameter space is $d = 2$. In Figure 2, we show adjacent cells $V(m_k)$ and $V(m_j)$, where $m_k = (\theta_1^k, \theta_2^k)$ and $m_j = (\theta_1^j, \theta_2^j)$. Given a conditional value $Y = y_o$, we have that

$$||\mathbf{m_k} - x_j^{(y_o)}|| = ||\mathbf{m_j} - x_j^{(y_o)}||.$$

Where $x_j^{(y_o)} = (x_j, y_o)$ is the unique point in the boundary between cells k and j that defines the intersection with the axis defined by $Y = y_o$. Using Euclidian distance we have,

$$(d_k^{\perp})^2 + (\theta_1^k - x_j)^2 = (d_j^{\perp})^2 + (\theta_2^j - x_j)^2.$$

Where $(d_k^{\perp})^2 = |\theta_2^k - y_o|$ is the perpendicular distance from point m_k from the current axis $Y = y_o$. Analogously, $(d_j^{\perp})^2 = |\theta_2^j - y_o|$ is the perpendicular distance from point m_j to $Y = y_o$. Solving for the intersection point x_j we obtain,

$$x_j = \frac{1}{2}\left[\theta_1^k + \theta_1^j + \frac{(d_k^{\perp})^2 - (d_j^{\perp})^2}{\theta_1^k - \theta_1^j}\right].$$

To find out the required boundaries of the Voronoi cells, the previous equation must be evaluated for all n cells. The lower and upper boundaries for each cell are given by $max[l_d, x_j]$, and $min[u_d, x_j]$ respectively. l_d and u_d are the lower and upper bounds of the parameter space in the d dimension. Once we have all the x_j's, we can start constructing the approximate marginals $\hat{\pi}(\mathbf{m}|\mathbf{X})$ since all the points inside each cell have the same density by definition. In Figure 3, we show that we can avoid a discretization of the axis by knowing the exact intersection point between cells given any conditional line. As we can see in Figure 3, each intersection point defines the change in the density up to the boundary of the next cell. The density associated to each cell comes from the initial evaluation done by the sampler before the creation of the first Voronoi tessellation. This means that the model evaluations will be used as the only information needed to build these approximate marginal densities. Once a new point is sampled, it will have the same density as its nearest neighbor. The computational cost of this procedure for each model depends on the number of samples n and the dimensionality d of the vector of parameters.

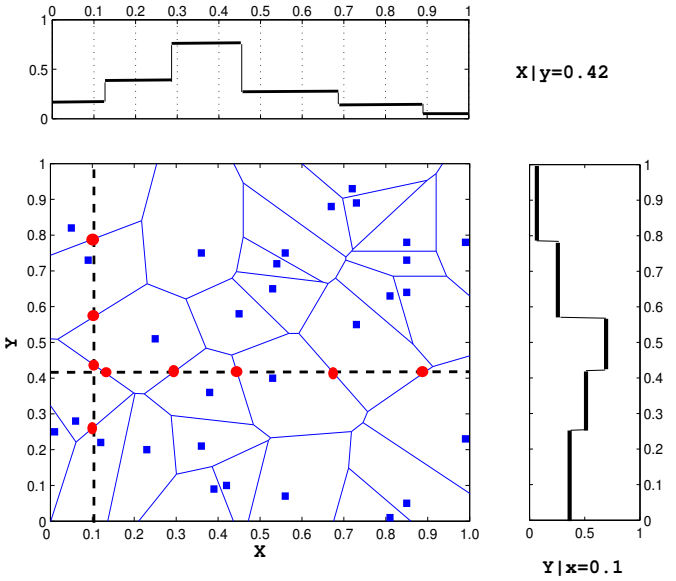

Figure 3: Voronoi tessellation in $2D$. A single Gibbs sampler step is performed. It is started with an arbitrary value $Y = 0.42$, then the approximate marginal density $P(X|Y = 0.42)$ is computed (Top Figure), then we sample a new point for X. With the new generated point $X = 0.1$, we can compute the approximate marginal density $P(Y|X = 0.1)$ (Right Figure). Initial samples are represented by squares. The circles are the intersection points between adjacent cells given a conditional (dotted) line.

2.2.1 STATISTICAL EXAMPLE

To test our non-parametric method, we use as a target function a mixture of bivariate Gaussian densities,

$$f(x,y) = pN\left(\begin{pmatrix} 4 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right) + (1-p)N\left(\begin{pmatrix} 8 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \qquad (1)$$

where the weight of the mixture is set to be equal to 0.7. Following the algorithm described in Section 2.1,

1. Initial Sampling Strategy: We use an Adaptive Metropolis scheme to simulate 200 samples from (1). Some specifications about the MCMC is that we do not use a burn-in period, and we update the covariance matrix of the proposal distribution every 50-th iteration since we are going to sample only two hundred points.

2. Voronoi Tessellation: Using the initial sample $(m_1, ..., m_{200})$, we construct the Voronoi tessellation and create the approximate PPD, $\hat{\pi}(m)$.

3. Sampling new points: Run a Gibbs sampler on the approximate PPD and extract $n^* = 500$ new points.

4. Regeneration: Add the new points to the initial sample and regenerate the surface .

5. Repeat steps (2)-(4) for 20 times to generate a sample of $10,000$ new points.

In Figure 4, (panel (a)) we look at the contour plots of the target distribution and the initial sample of 200 points obtained from the AM scheme. The regeneration of the surface (panels (b)-(c)) can be achieved by using sequentially a Gibbs sampler over the tessellation. To improve the sampling, we use a rotation in the direction of the sampling based on the eigenvalues computed from the sample covariance matrix obtained during the Adaptive Metropolis sampling. Due to the design of the regeneration of the Voronoi cells, regions with high density (small cells) tend to be more sampled than regions with low density (large cells), this explains the over sampling (panel (d)) of the maxima of the objective function (1). **It is worth noting that in this example, the initial sample provides a crude approximation of the mass of the two normal components. The new points and regeneration step help to refine the initial mass and to provide more samples in areas with high posterior mass. If in general the initial sample does not provide good guidance of where the posterior mass is majorly located, then adding new points could not provide good resolution to the PPD. Our approach is different to what typically is done when approximating response surfaces with regression or polynomial techniques where one tries to specify a design that fills the covariate space. In our case, the requirement of an ensemble of draws from the posterior distribution could become a critical step in achieving an efficient approximation of the PPD. The approach largely hinges on the accuracy of the voronoi-based estimate of the PPD given the initial sample. For further exploration, in Figure 5 we consider the same example as in Figure 4 but with the initial sample defined on a rectangular grid of the parameter space. The regeneration sampling recovers the mixture distribution. In this case, there were no issues dealing with edges on parameter boundaries since the initial sample is set in a re-**

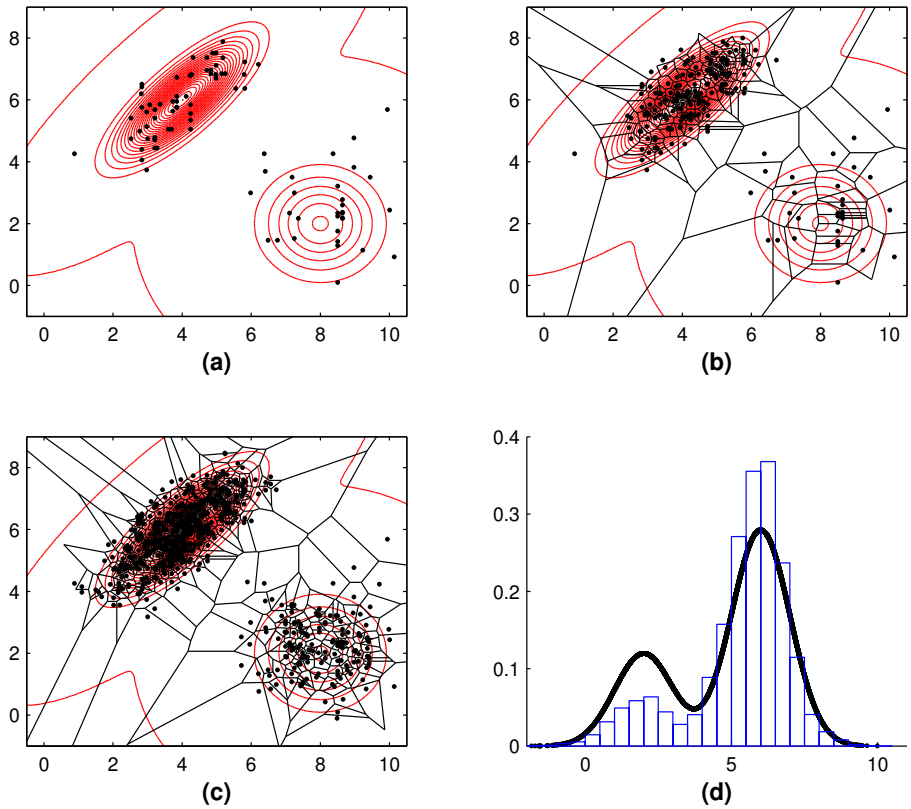

Figure 4: Target distribution $(pN(\mu_1, \Sigma_1) + (1-p)N(\mu_2, \Sigma_2)$ with $\mu_1 = (4, 6)'$, $\mu_2 = (8, 2)'$, Σ_1 is a matrix with elements, $\sigma_{11} = \sigma_{22} = 1$, and $\sigma_{12} = 0.8$; Σ_2 is an identity matrix, and $p = 0.7$ is the mixture weight.): (a) 200 Initial samples from Adaptive Metropolis; (b) Surface regenerated with 400 points (200 new); (c) Surface regenerated with 1000 points (800 new); (d) Real PDF (solid line) from the marginal $f(y) = 0.7N(6, 1) + 0.3N(2, 1)$ compared to histogram done with $10,000$ points obtained from hybrid regeneration.

gion that covers the probability mass almost completely. Further studies would be needed to address boundary issues for other types of PPD's.

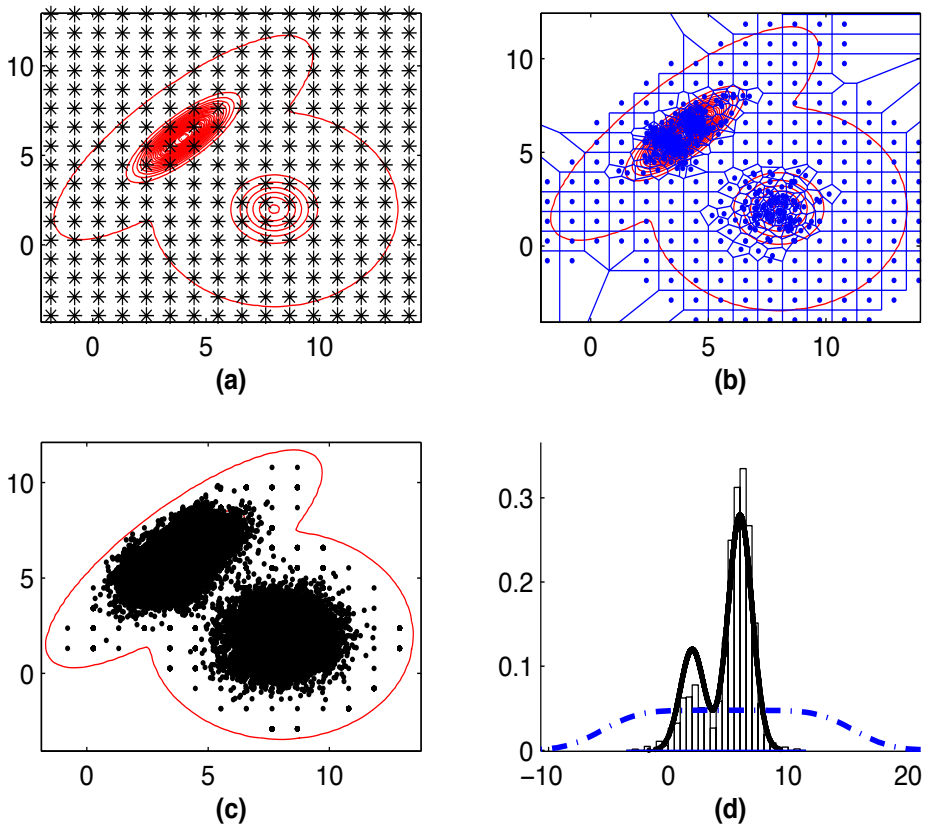

Figure 5: Target distribution $(pN(\mu_1, \Sigma_1) + (1-p)N(\mu_2, \Sigma_2)$ with $\mu_1 = (4, 6)'$, $\mu_2 = (8, 2)'$, Σ_1 is a matrix with elements, $\sigma_{11} = \sigma_{22} = 1$, and $\sigma_{12} = 0.8$; Σ_2 is an identity matrix, and $p = 0.7$ is the mixture weight.): (a) Grid of 400 initial points; (b) Voronoi tessellation generated with 600; (c) Scatter plot of $12,000$ new points; (d) Real PDF (solid line) from the marginal $f(y) = 0.7N(6, 1) + 0.3N(2, 1)$ compared to histogram done with $12,000$ points obtained from hybrid regeneration, and to a kernel density estimation by using the initial 400 points (dotted line).

3 CLIMATE MODEL CALIBRATION

In this Section, we consider a couple of climate model applications, the first climate model arises from Milankovitch (1941) theory of climate change, in which variations in the Earth's orbit cause climate variability through a local thermodynamic response to changes in insolation. The Earth's

orbital geometry parameters (obliquity, longitude of perihelion and eccentricity) are astronomical factors that influence the timing and intensity of the seasons. A second application in climate modeling is done by using few evaluations coming from NCAR's CAM model, where the parameters to consider are precipitation efficiency, and the environmental air entrainment rate. While dimensionality on these climate models may be low, the computational cost to evaluate them is quite high, this is the case of the CAM model experiments considered in this paper.

3.1 CLIMATE MODEL (Milankovitch)

Jackson and Broccoli (2003) take advantage of the short equilibration time (10 years) of an atmospheric general circulation model (AGCM), land surface model and a static mixed-layer ocean model, which includes a thermodynamic model of sea ice to derive the equilibrium climate response to accelerated variations in Earth's orbital configuration over the past 165,000 years. By fitting a time series of the evolution of each orbital component with the model output, they can estimate an amplitude of that component within the time series. This amplitude represents the sensitivity of the region and season to changes in that orbital component. The sensitivity of surface air temperature to obliquity forcing $A_{o,ijk}$ and precessional forcing $A_{p,ijk}$ can be defined for particular latitudes i, longitudes j, and seasons k. They represent the climate model's response to the seasonally and latitude varying changes in insolation for a given unit change in orbital parameter values. They are derived from an ordinary multiple least squares fitting procedure between modeled variations in climate found within a climate model integration of the past 165,000 yrs forced only by changes in Earth's orbital geometry and two basis functions representing the known temporal variations in obliquity and precession. In particular, the obliquity basis function $A_{o,ijk}\Phi'(t)$ consists of an unknown sensitivity $A_{o,ijk}$ and the time series of obliquity variations $\Phi'(t)$ over the past 165,000 years, where $\Phi'(t) = \Phi(t) - \Phi_o$, is the deviation of obliquity from its 165,000 yrs mean ($\Phi_o = 23.3515°$). The precessional basis function $A_{p,ijk}e(t)cos(\phi_{p,ijk} - \lambda(t))$ consists of an unknown sensitivity $A_{p,ijk}$, an unknown phase angle of response $\phi_{p,ijk}$, the time series of eccentricity $e(t)$, and the time series of the longitude of the perihelion $\lambda(t)$. The time series $e(t)$, $\lambda(t)$, and $\Phi'(t)$ are known from orbital mechanics and were used as input values in the AGCM which calculates the changes in insolation as a functionof latitude and season for each year of the experiment.

The multiple least squares fitting procedure provides estimates of $A_{o,ijk}$, $A_{p,ijk}$, and $\phi_{p,ijk}$ that best represent the climate model's response to the time evolving changes in orbital forcing. For instance, the variations in surface air temperature with respect to the 165,000 years annual mean for a given region and season $T_{ijk}(t)$ may be represented by,

$$T_{ijk}(t) = A_{o,ijk}\Phi'(t) + A_{p,ijk}e(t)cos(\phi_{p,ijk} - \lambda(t)) + R_{ijk}(t), \tag{2}$$

where $R_{ijk}(t)$ is a residual. The fitting procedure described above also allows one to construct a surrogate climate model using the estimated latitude, longitude, and seasonal obliquity and precessional forcing sensitivities. Villagran et al. (2008) give a comparison of the ability of the least squares fitting procedure with imposed time variations in Earth's orbital geometry to reproduce the AGCM's response to the annual mean air temperature in Antarctica averaged from 70° S to 90° S and separated into its obliquity, precessional, and residual components. This is done by averaging together the sensitivities of all latitude, longitude, and seasons for this region and estimating the response by imposing the changes in the obliquity and precessional components.

3.2 SURROGATE CLIMATE MODEL

The surrogate model is based on surface air temperature fields generated by an AGCM in its response to changes in three parameters specifying Earth's orbital geometry over the past 165,000 years. The response can be approximated in terms of obliquity and precession components by using the multiple least squares fitting procedure described in the previous section. We will use this surrogate climate model to test our non-parametric method as a follow-up study to Jackson et al. (2004), and Villagran et al. (2008), which consider the same surrogate model as in this paper but mostly compare Multiple Very Fast Simulated Annealing (MVFSA) against Adaptive Metropolis methods as sampling strategies for climate modeling. MVFSA is an heuristic method that attempts to strike a balance between optimization and sampling of the PPD from multiple starting points as described in Villagran et al. (2008).

We denote the Earth's orbital geometry parameters and their physical range as obliquity, $\Phi \in (22°, 25°)$, eccentricity, $e \in (0, 0.05)$ and longitude of perihelion, $\lambda \in (0°, 360°)$, therefore the dependency on t is omitted for the surrogate climate model. The observed data is a $3D$ array $d_{obs,ijk}$ which represents the observed surface temperature anomalies with respect to the long term 165,000 years mean at latitude i,

longitude j, and season k. The grid spacing is approximately 4.5° latitude by 7.5° longitude, then the latitude can take $I = 40$ different values, and the longitude $J = 48$. The season takes $K = 12$ values, which are selected days throughout the year. Each value of k would apply for that season and for all time t over the past 165,000 years. The observed data are simulated using $\Phi = 22.625$, $e = 0.043954$, and $\lambda = 75.93$, as ideal values for the climate model. We approximate the data using the relationship, $d_{obs,ijk} = g_{ijk}(m) + \eta_{ijk}$, where $\mathbf{m} = (\Phi, \lambda, e)$ is the vector of parameters, $\mathbf{g}$ represents the forward operator and it has the same dimensionality as $d_{obs,ijk}$. The definition of the function $\mathbf{g}$ is crucial since it is completely defined by the physical system. The term η_{ijk} is a Gaussian error with estimated variance given by B_{ijk}; this array represents the variance of the observations at each grid point. This variability comes from the 1,500 year integration of the model itself, but with the appropriate seasonal and climatological averages (i.e. 10 year means of particular seasons). Typically, in Earth science models the observational uncertainties are assumed as Gaussian, see Jackson et al. (2004), Tebaldi et al. (2005), and Lopez et al. (2006).

In this paper, the surface air temperature anomaly to a given change in the three parameters that define the Earth's orbital geometry is $\mathbf{g_{ijk}}(\mathbf{m})$. The surrogate climate model is defined as follows,

$$\mathbf{g_{ijk}}(\mathbf{m}) = \hat{A}_{o,ijk}\Phi' + e\hat{A}_{p,ijk}cos(\hat{\phi}_{p,ijk} - \lambda) + \hat{R}_{ijk}, \tag{3}$$

where $\hat{\phi}_{p,ijk}$ is the phase of the response to precessional forcing and $\hat{R}_{ijk}$ are the residuals averaged over time obtained from the AGCM in (2). This term is added to represent the effects of internal variability on 10 year seasonal means. Repeated experiments of the climate model will cycle through 1 of 150 possible values of $\hat{R}_{ijk}$ that come from a 1,500 year long control integration of the AGCM. $\hat{A}_{o,ijk}$ and $\hat{A}_{p,ijk}$ are the sensitivity of temperature to changes in obliquity and precession obtained using the time series fitting procedure in (2).

3.3 CALIBRATION RESULTS

An element considered in some geoscience models (Mu et al. (2003), Jackson et al. (2004), and Wang (2007)) to calibrate or measure the deviation generated from the observed data (d_{obs}) and the data generated from the model ($\mathbf{g}(\mathbf{m})$) is called the cost function. In general, the cost function can be represented as $E(m) = ||d_{obs} - g(m)||$, where $\mathbf{m}$ is any given vector of parameters of interest from the physical system, and the difference between the data and the model is given by a specific metric. The cost

Method	$\mathbf{E}(\mathbf{m}^*)$	$\Phi_{2.5\%}$	$\Phi_{97.5\%}$	$\lambda_{2.5\%}$	$\lambda_{97.5\%}$	$e_{2.5\%}$	$e_{97.5\%}$
SCAM	0.191	22.164	23.131	60.663	91.543	0.0312	0.0495
MVFSA	0.202	22.160	24.507	18.333	311.938	0.0089	0.0482
Hybrid	0.191	22.215	23.066	63.406	97.933	0.0340	0.0490

Table 1: Comparative estimation after 500 forward evaluations. $E(m^*)$ is the minimum of the cost function, Φ is the Obliquity, λ is the Longitude of Perihelion and e is the Eccentricity.

function can be defined in many ways. For instance, in the surrogate climate model considered here, the cost function is defined as,

$$E(m) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} B_{ijk}^{-1}(d_{obs,ijk} - g_{ijk}(m))^2, \tag{4}$$

where $\mathbf{m} = (\Phi, e, \lambda)$ is the vector of Earth's orbital geometry parameters (obliquity, eccentricity, and longitude of perihelion). On the climate model studied here there is just one field, surface air temperature anomalies, however we can have N different sets of observations, such as seasonal and annual mean surface air temperature, precipitation, winds, and clouds at different latitudes. It is proposed that, the likelihood function takes the form,

$$L(d_{obs}|m, S) \propto exp\{-SE(m)\}. \tag{5}$$

The parameter S is connected to B_{ijk} according to Jackson et al. (2004) as a scaling factor. S performs the function of weighing the significance of model-data differences.

We now compare three different computational techniques, MVFSA, Single Component Adaptive Metropolis (SCAM), and the hybrid non-parametric approach presented in this paper. Since one of the main concerns in climate modeling is the time spent in performing forward evaluations, it is prohibited to think about doing a typical MCMC simulation with thousands of iterations. Therefore, by suggestion of our climate expert, we allow only 500 model evaluations to be done in order to compare different strategies. In Table 1, we compare the uncertainty estimation of the parameters and the minimum values of the cost function. The performance of SCAM is remarkable since it does not only find a minimum cost with few forward evaluations, but it does also provide acceptable estimates of the 95% credible intervals of the orbital forcing parameters. The 95% credible intervals based on MVFSA

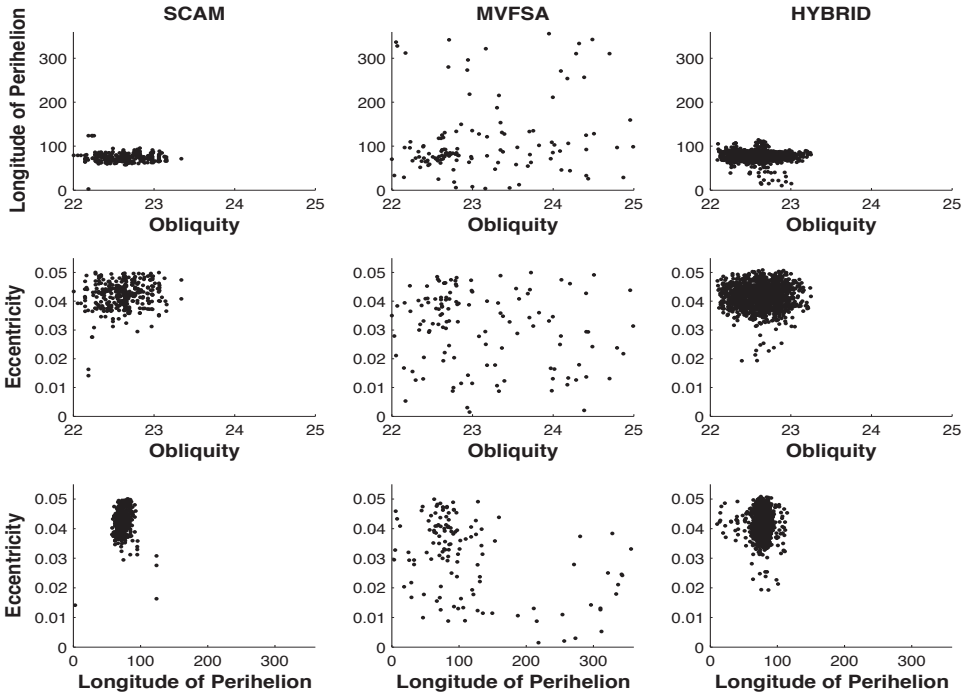

Figure 6: Bivariate scatter plots of orbital forcing parameters with just 500 forward evaluations. First column: SCAM. Second column: MVFSA. Third column: Hybrid regeneration.

are not informative at all because they practically cover the entire parameter space. Applying hybrid regeneration of the surface to the initial 500 forward evaluations, we generate $100,000$ new points that provided comparable credible intervals to the Adaptive Metropolis scheme. A particular remark to be made is that SCAM was used as initial sampling strategy for the hybrid non-parametric approach presented here because Villagran et al. (2008) demonstrated that it was a good choice for problems with parameter restrictions, which is the case of the climate model used in this paper. In addition, they showed that MVFSA produces biases in the tails since it uses multiple independent initial points along the parameters space.

In Figure 6, we look at the bivariate scatter plots of the samples of the different methods corresponding to each orbital parameter. No burn-in period was allowed neither for SCAM nor for MVFSA. The Adaptive Method concentrates around the optimum values with only few forward evaluations while MVFSA samples are dis-

tributed loosely all over the parameter space. The samples coming from the hybrid regeneration approach gather around the optimum values as SCAM does since the initial surface was based in the Adaptive sampling method.

In Figure 7, we can observe the regeneration of the surface at four stages of the process for longitude of perihelion and eccentricity. Just by using the initial 500 points sampled from the Adaptive Metropolis method, the hybrid non-parametric approach presented in this paper is able to rebuild the surface of the climate model parameters practically from scratch. Even though we transform the samples to the square [0,1] to improve visualization, it is not clear to observe at the first phase that there is any pattern in the shape of the cells. However, after some new samples are drawn and some regenerations of the surface are done, the distribution of the parameters becomes evident. These results were obtained by sampling 500 new points from each surface, and regenerating the surface 200 times.

In Figure 8, we first compare the estimation of the PPD for the obliquity parameter using only 500 **forward model evaluations.** If we use the samples coming from either SCAM and MVFSA to estimate the PPD for Obliquity and eccentricity, we can see there are some bumps even after having being used a kernel smoother. The hybrid method does a good job estimating the posterior densities by using only the 500 initial points and generating up to one hundred thousand new points. Even if we compare what would happen in the hypothetical case that we were able to evaluate a climate model $100,000$ times, we can see there are small differences in the densities using hybrid regeneration and the Adaptive Metropolis scheme. Being able to avoid thousands of expensive forward evaluations in the climate model itself is a huge difference in favor to the former scheme. The calibration can be done quite reasonably using the non-parametric approach since the optimum values (triangles) are covered by high density regions around them.

3.4 Community Atmospheric Model (CAM3.1)

Climatologists have an interest in making inferences about models that may take hours to days to execute a single iteration of a stochastic sampler. We consider the use of the technique presented in this paper useful, especially when forward evaluations have already been done and there is an interest in using them to calibrate the computer model. This is the case of the Community Atmospheric Model (CAM3.1) developed by the

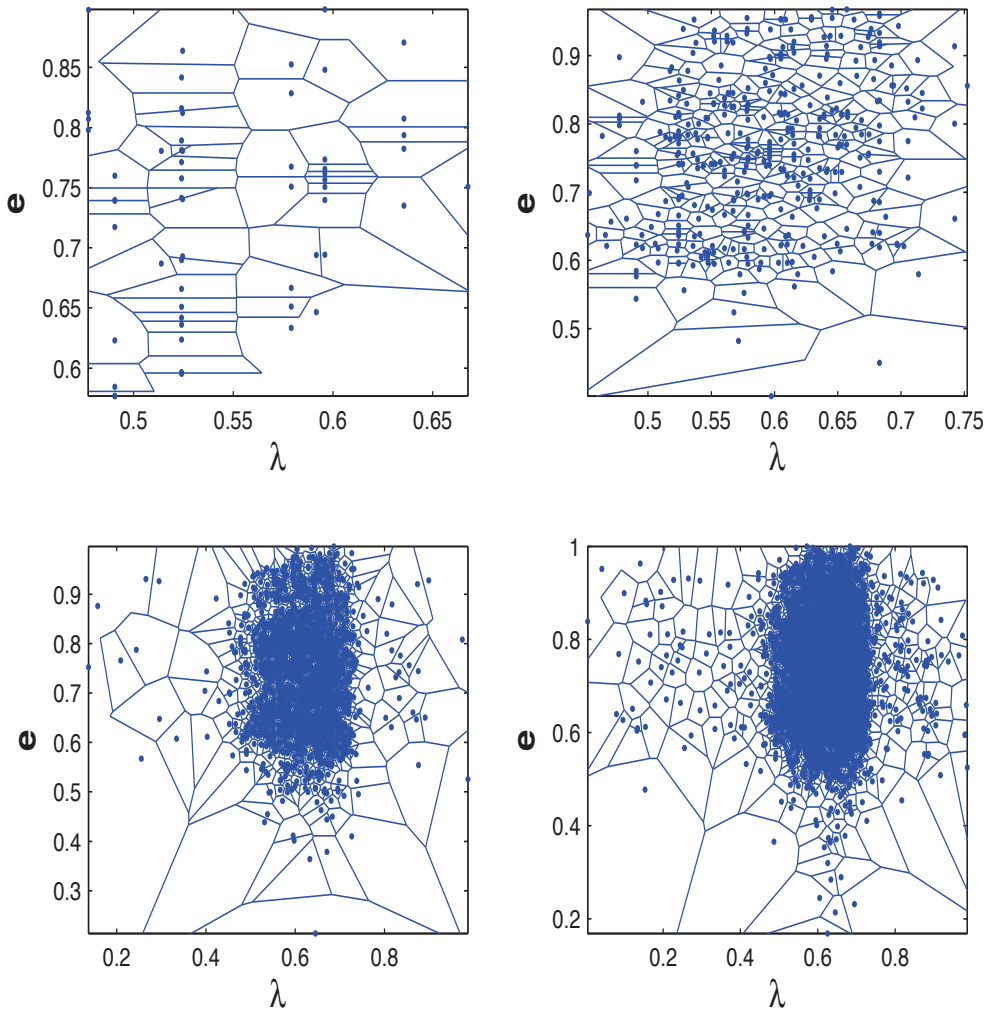

Figure 7: Longitude of Perihelion vs Eccentricity. Hybrid surface regeneration at four stages, from top left to bottom right: 500, 1000, 2500, and 5000 sampled points.

National Center for Atmospheric Research (NCAR). Using hybrid regeneration in the CAM3.1 imposes a new challenge, since the initial sample to construct the Voronoi tessellation would be from MVFSA instead of AM. This makes a difference since MVFSA provide samples over all the parameter space. This can lead to non-informative interval estimation (Table 1), and regeneration of cells that have extremely low density and that are not

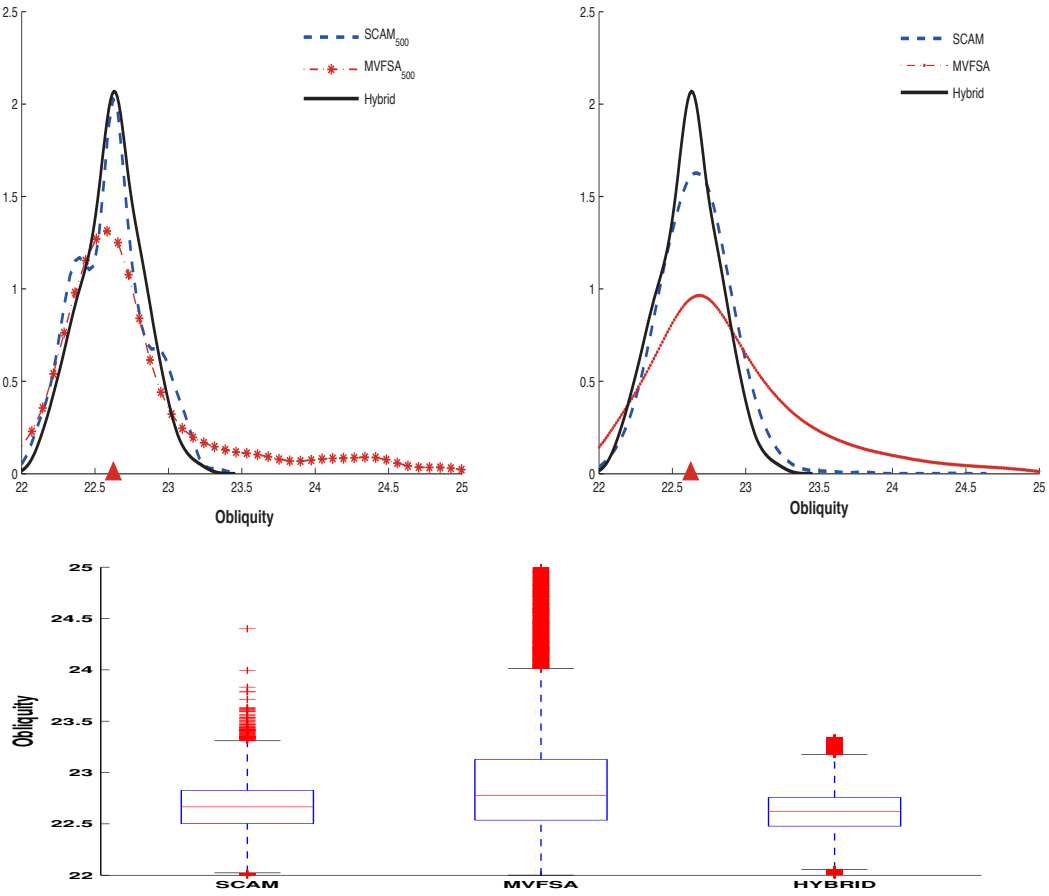

Figure 8: Comparison for Obliquity parameter. Top left: PPD estimation with 500 forward evaluations. Top right: PPD estimation with 100,000 forward evaluations for SCAM and MVFSA. Although Hybrid regeneration uses 100,000 samples, it only used 500 forward evaluations to estimate the PPD. Bottom: Box plots of the samples from different methods.

meaningful in terms of model calibration. Ongoing investigation is being done to make hybrid regeneration less dependent of the initial sample, and to provide ergodic properties of the method. It is also not clear that Gaussian process models provide better posterior approximations in these situations given that the climate response generated by CAM is highly multivariate. This climate model uses fourteen fields of response (Jackson et al. (2008)) and has six parameters, *RHMINL* (low cloud critical rela-

tive humidity), *RHMINH* (high cloud critical relative humidity), *ALFA* (initial cloud downdraft mass flux), *TAU* (consumption rate of CAPE), *ke* (environmental air entrainment rate), and *c0* (precipitation efficiency). Currently forward evaluations of this model are being performed by using MVFSA. Work is in progress to replace MVFSA by Adaptive Metropolis to reduce biases in the sampling and attempts of estimating parameters of such model with Gaussian processes had not been very successful. On the hand, we had implemented the methods of this paper to a situation where only two parameters of the CAM 3.1 (*ke* and *c0*) vary according to an 8×8 factorial design. A similar likelihood and cost function like the ones introduced in Section 3.3 was constructed based on the various response fields and seasons. Each response field provides a coarse spatial resolution that covers the tropics and poles of the Earth. Figure 9 shows the points where initial experiments were run and for which cost function values were computed. This 64 points are treated as the initial sample in our sampling scheme. The crosses are the sampled value after implementing step (1)-(4) of the algorithm and the contours represent a smooth version of the samples that correspond to the PPD. Figure 10 show the resulting marginal PPD distributions for both parameters *ke* and *c0*. Although the results in Jackson et al. (2008) vary six parameters at a time, the resulting posterior distribution reported there resembles the marginals obtained for Figure 10 from the hybrid-regeneration algorithm.

4 DISCUSSION

From a statistical standpoint, calibration in complex computer models has been dealt traditionally by proposing a statistical model or emulator that can avoid computational limitations. Emulators based on Gaussian process (Kennedy and O'Hagan (2000), Sansó et al. (2008), Christen and Sansó (2011)) assume that the model can be run at different levels of complexity, while Bliznyuk et al. (2008) propose a different approach by using a radial basis function approximation to the logarithm of the posterior density. The non-parametric method presented in this paper follows a completely different approach, a combination of a geometrical structure (Voronoi, 1908) and a nearest-neighbor approach drawn from geophysics (Sambridge, 1999), providing a non-standard way to avoid expensive evaluations in the computer model

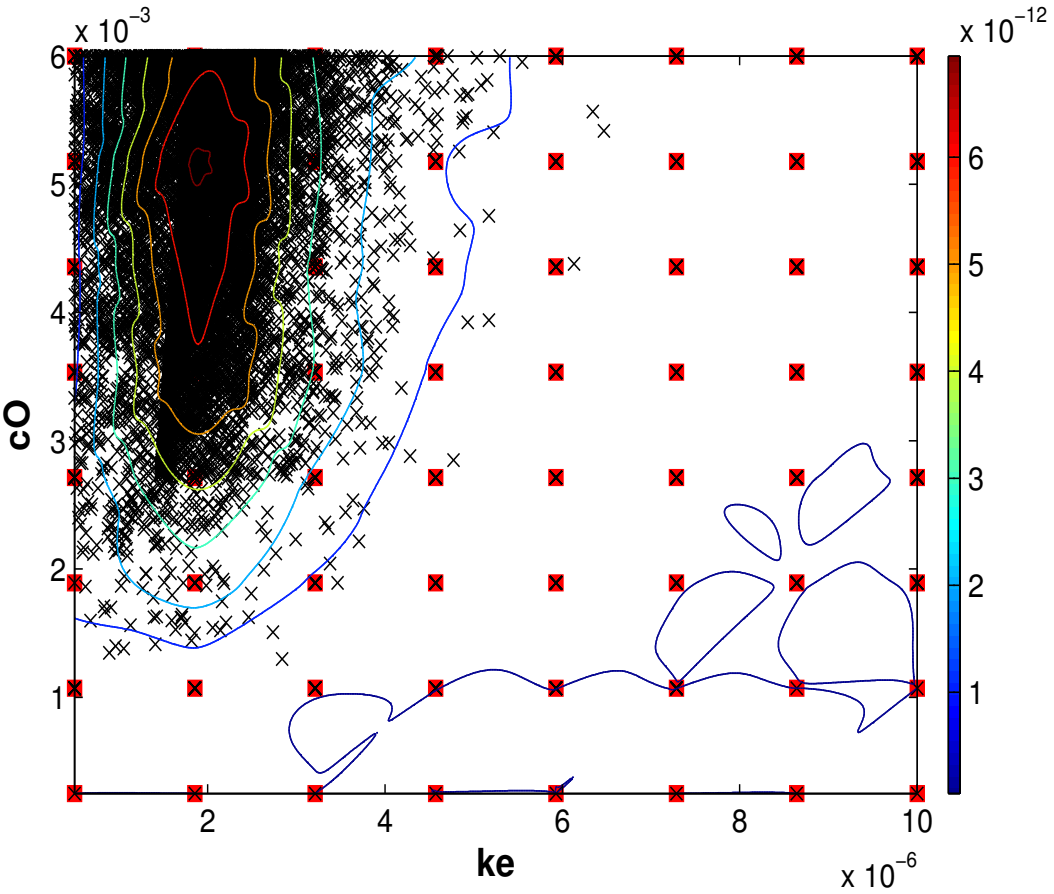

Figure 9: 8×8 factorial design for CAM 3.1 parameters *ke* and *c0*, samples for hybrid-regeneration algorithm and PPD contour plots.

itself. **Following the Kennedy et al. (2001) notation, a physical process z_i can be expressed as the sum of a real process $\xi(x)$ and an observational error ϵ, thus $z_i = \xi(x_i) + \epsilon_i$. Furthermore, $\xi(x) = \rho\eta(x, \theta) + \delta(x)$. In a GP model, a Gaussian Process prior distribution is assumed both in the computer code $\eta(\cdot)$ and the model inadequacy function $\delta(\cdot)$. In the climate model applications presented in this paper, no $\delta(x)$ is assumed, and the hybrid non-parametric procedure explore and estimate the parameters of the forward operator g, which is an estimator of the real process $\xi(\cdot)$. In contrast to these Gaussian process emulators, the geometrical structure studied here does not provide a direct estimate of model discrepancy error or a surface approximation to response. On the other hand, the Voronoi**

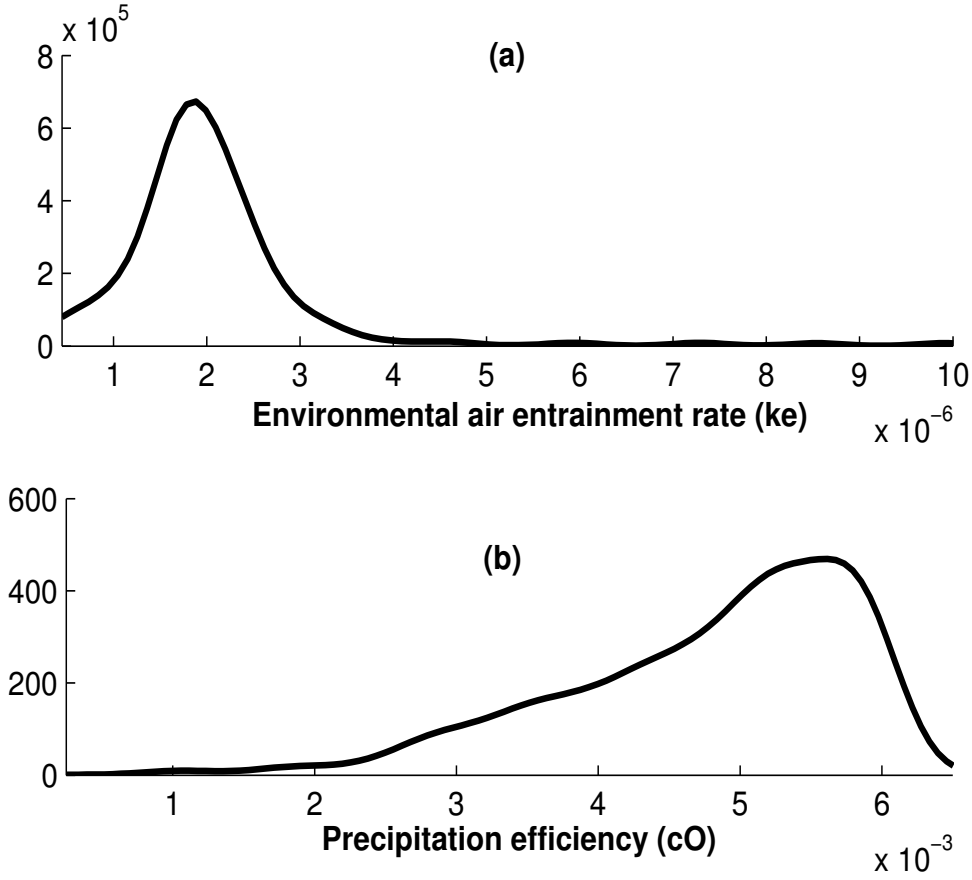

Figure 10: (a) Marginal PPD for *ke* parameter from hybrid regeneration. (b) Marginal PPD for *c0* parameter.

tessellation can provide a quick approximation to the posterior distribution of parameters when a Gaussian process model is not practical due to a high-dimensional response or parameter space.

By definition, the Voronoi tessellation defines a unique geometrical structure that completely covers the parameter space. In addition, the cells of the tessellation have a size proportional to the concentration or distribution of the cells over the entire space. These facts resemble the description of random vector and its joint probability density function over its support. Therefore, we have applied these ideas to formulate a non-parametric procedure. The algorithm can be depicted in few steps: (1) Propose a sampling scheme to obtain points from regions of the parameter space where there is high density, with this initial sample (2) construct the Voronoi tessellation. Using

this approximated surface, (3) draw a sample of new points; add the new points to the original sample and (4) rebuild the Voronoi tessellation. These steps can be repeated N times, resulting in an iterative procedure that will refine the surface by generating more new small cells (points with high density) than new big cells (points with low density).

The potential applications are broad and may prove invaluable for problems that are currently limited by computational requirements of the forward model. In this paper we have applied hybrid regeneration to a couple of AGCM climate models. **Sambridge (1999) has applied the nearest neighborhood approach to a problem with** 24 **parameters**. We have compared the results from our method against a long MCMC run (using either Adaptive Metropolis or MVFSA) and have concluded that hybrid regeneration provides a successful calibration given that the generation of new samples take place after a few number of forward evaluations in the computer climate models. The approach we have proposed allows to generate new points without the computational burden of additional model evaluations.

5 ACKNOWLEDGEMENTS

G. Huerta and A. Nosedal are supported by the Office of Science (BER), U.S. Department of Energy.

References

[1] Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P. (2008) "Bayesian Calibration and Uncertainty Analysis for Computationally Expensive Models Using Optimization and Radial Basis Function Approximation", *Journal of Computational and Graphical Statistics*, 17, 2, 270-294.

[2] Christen, A., and Sansó, B. (2011) "Advances in the Design of Gaussian Processes as Surrogate Models for Computer Experiments", *Communications in Statistics - Theory and Methods*, 40,24, 4467-4483.

[3] Craig, P.S., Goldstein, M., Rougier, J.C., and Seheult, A.H. (2001) "Bayesian Forecasting for Complex Systems Using Computer Simulators", *Journal of the American Statistical Association*, 96, 717-729.

[4] Craig, P.S., Goldstein, M., Seheult, A.H., and Smith, J.A. (1996) "Bayes Linear Strategies for History Matching of Hydrocarbon Reservoirs" *Bayesian Statistics 5*, 69-95.

[5] Goldstein, M. and Rougier, J. C. (2006) "Bayes linear calibrated prediction for complex systems", *Journal of the American Statistical Association*, 101, 1132-1143.

[6] Geman, S. and Geman, D., (1984) "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions PAMI-6*, 721-741.

[7] Haario, H., Laine, M., Lehtinen, M., Saksman, E. and Tamminen, J., (2004), "Markov Chain Monte Carlo methods for high dimensional inversion in remote sensing", *Journal of Royal Statistical Society*, 66, 591-607.

[8] Haario, H., Saksman, E., Tammimen, J., (2001) "An Adaptive Metropolis algorithm", *Bernoulli*, 7, 223-242.

[9] Hastings, W., (1970) "Monte Carlo sampling methods using Markov chains and their applications", *Biometrica*, 57, 97-109.

[10] Heitmann, K., Higdon, D., Nakhleh, C., and Habib, S. (2006) "Cosmic Calibration", *The Astrophysical Journal*, 646, L1-L4.

[11] Higdon, D., Lee, H., and Holloman, C. (2003), "Markov Chain Monte Carlo-Based Approaches for Inference in Computationally Intense Inverse Problems", *Bayesian Statistics 7*, 181-197.

[12] Jackson, C. and Broccoli, A. (2003) "Orbital forcing of Arctic climate: mechanisms of climate response and implications for continental glaciation", *Climate Dynamics*, 21, 539-557.

[13] Jackson, C.S., Sen, M.K., Huerta, G. Deng, Y. and Bowman, K.P. (2008) "Error Reduction and Convergence in Climate Prediction", *Journal of Climate*, 21, 24, 6698-6709.

[14] Jackson, C., Sen, M. and Stoffa, P. (2004) "An Efficient Stochastic Bayesian Approach to Optimal Parameter and Uncertainty Estimation for Climate Model Predictions", *Journal of Climate*, 17, 2828-2840.

[15] Kennedy, M., and O'Hagan, A. (2000) "Predicting the Output From a Complex Computer Code When Fast Approximations are Available", *Biometrika*, 87, 1-13.

[16] Lopez, A., Tebaldi, C., New, M., Stainforth, D., Allen, M. and Kettleborough, J. (2006) "Two Approaches to Quantifying Uncertainty in Global Temperature Changes", *Journal of Climate*, 19, 4785-4796.

[17] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M.N., Teller, A.H., and Teller, E., (1953) "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, 21, 1087-1091.

[18] Milankovitch, M. (1941) "Canon of insolation and the ice age problem", *Israel Program for Scientific Translations, Jerusalem.*

[19] Mu, Q., C. S. Jackson, P. L. Stoffa (2004) "A Multivariate EOF-based measure of climate model performance", *Journal of Geophysical Research*, 109.

[20] Okabe, A., Boots, B., Sugihara, K., and Chiu, S.N. (2000) "Spatial Tessellations: Concepts and Applications of Voronoi diagrams", *Wiley*, 2nd Edition.

[21] Sambridge, M. (1999) "Geophysical Inversion with a Neighbourhood Algorithm - I", *Geophysical Journal International*, 138, 479-494.

[22] Sansó, B., Forest, C.E. and Zantedeschi, D. (2008) "Inferring Climate System Properties Using a Computer Model (with discussion)", *Bayesian Analysis*, 3, 1, 1-62.

[23] Sen, M. and Stoffa, P., (1996) "Bayesian Inference, Gibbs sampler and uncertainty estimation in geophysical inversion", *Geophysical Prospecting*, 44, 313-350.

[24] Tebaldi, C., Smith, R., Nychka, D., and Mearns, L. (2005) "Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles", *Journal of Climate*, 18, 1524-1540.

[25] Villagran, A., Huerta, G., Jackson, C., and Sen, M. (2008) "Computational Methods for Parameter Estimation in Climate Models", *Bayesian Analysis*, 4, 823-850.

[26] Voronoi, M.G. (1908) "Nouvelles applications des parametres continus a la theorie des formes quadratiques", *J. Reine Angew. Math.*, 134, 198-287.

[27] Wang, F. (2007) "Investigating ENSO sensitivity to mean climate in an intermediate model using a novel statistical technique", *Geophysical Research Letters*, 34, L07705.