Geophys. J. Int. (2005) doi: 10.1111/j.1365-246X.2005.02573.x

Importance reweighting reduces dependence on temperature in Gibbs samplers: an application to the coseismic geodetic inverse problem

Benjamin A. Brooks* and L. Neil Frazer

School of Ocean and Earth Science and Technology, University of Hawaii, 1680 *East-West Rd., Honolulu, HI* 96822, *USA*

Accepted 2004 December 22. Received 2004 November 10; in original form 2003 December 8

SUMMARY

We employ importance reweighting to extend Gibbs sampling (GS) to a larger class of unnormalized, multidimensional probability functions and to reduce the dependence of the results on critical temperature T_*, which is sometimes poorly known. Instead of sampling at T_*, we sample at several sampling temperatures, T_S, in an interval centred on an estimate of T_*, correcting the results for each temperature to $T = 1$. Convergence is verified in part by agreement of marginal posterior distributions obtained at different T_s. For the coseismic geodetic problem, experiments with synthetic data suggest that optimal sampling temperature varies inversely with the signal-to-noise ratio (SNR): as signal strength increases, optimal sampling temperature decreases. Inversion of surface displacement data from the 1994 Northridge earthquake confirms coseismic source parameters from other methods, while providing extra information in the form of properly scaled marginal posterior probability density functions.

Key words: coseismic deformation, earthquake source, inversion, Gibbs sampling, heat bath, Monte Carlo.

1 INTRODUCTION

Non-linear inversion methods combined with ever-increasing computational power now allow workers to address challenging numerical problems from across the geophysical spectrum (Cervelli *et al.* 2001; Chapman & Jaschke 2001; Dosso & Wilmut 2002; Mosegaard & Sambridge 2002). One widely used such class of tools is based on Monte Carlo sampling (Kirkpatrick *et al.* 1983; Mosegaard & Sambridge 2002), which collects pseudo-random samples from multidimensional parameter space as a proxy for the true posterior probability density of the problem, $\sigma(\mathbf{m})$, where $\mathbf{m}$ is a vector of parameters. Often, the goal of an inversion using a sampling-based technique is not simply to obtain an optimal solution, but rather to estimate posterior densities for their value as indicators of resolution and uncertainty. In practice, however, inversion techniques usually depend on a parameter that controls the sampling process; thus, it is important to understand the relationship between this control parameter and sampling results. Here, we examine a popular method, Gibbs sampling (GS), and correct for the dependence of this technique on a control parameter known as temperature. In the larger context of Markov chain Monte Carlo (MCMC) algorithms, our method is a way of modifying the stationary distribution of the chain along lines first suggested by Jennison (1993) to improve mixing, i.e. to improve the probability of transitions between modes of $\sigma(\mathbf{m})$ separated by regions of very low probability. Gilks & Roberts (1996) review alternative ways to speed up mixing.

The heat bath (HB) algorithm (Creutz 1980; Rebbi 1984; Basu & Frazer 1990; Chapman & Jaschke 2001) is an algorithm for GS: it samples from the Gibbs–Boltzmann probability distribution

$$p_j(\mathbf{m}) = \frac{\mathrm{e}^{-E_j/T}}{\sum_{j'=1}^{N_S} \mathrm{e}^{-E_{j'}/T}}, \tag{1}$$

where $\mathbf{m}$, referred to here as the model, is a vector of unknown parameters, E_j is the free energy of the jth of N_S states and T is temperature. For use below, note that the denominator in eq. (1) is called the partition function, Z_T, and that $p_j(\mathbf{m})$ can be approximated by a continuous probability density function $p(\mathbf{m})$, so that Z_T is given by the integral $\int d\mathbf{m} \mathrm{e}^{-E(\mathbf{m})/T}$. In the HB algorithm, one begins at a high temperature and each parameter, $\mathbf{m}_j$, is visited in sequence. During a visit to the jth parameter, the values for the other parameters are held fixed and the system energy, E_j, is calculated for each allowed value of $\mathbf{m}_j$. These energies are used to generate a Gibbs–Boltzmann distribution for parameter j, from which a new state for that parameter is chosen by sampling without rejection. In statistics, this is known as sampling from the conditionals $p(\mathbf{m}_j|\mathbf{m}_{-j})$. After each parameter in the system has been visited once, in a cycle called a sweep, the temperature T is lowered by a small amount and each parameter is visited again. When the desired sampling temperature is reached, samples are collected for later use, but samples obtained during the cooling process are not used.

Ideally, to sample from an arbitrary $\sigma(\mathbf{m})$ one simply defines E in eq. (1) by $E(\mathbf{m}) = -\ln \sigma(\mathbf{m})$, then samples at $T = 1$. As the

*Contact Information: School of Ocean and Earth Science and Technology, University of Hawaii, 1680 East-West Rd. POST Suite #602, Honolulu, HI 96822, USA. E-mail: bbrooks@soest.hawaii.edu

number of sweeps becomes infinite, the equilibrium distribution of the system becomes the Gibbs–Boltzmann distribution (Geman & Geman 1984; Rothman 1986) for $T = 1$. In practice, before beginning to sample at $T = 1$, GS moves to the sampling temperature via a cooling schedule whose characteristics affect the performance of the associated inversion method (Rothman 1985). If cooled too rapidly, the system will not be in equilibrium and sampling will be biased, but cooling too slowly is computationally inefficient. (Statisticians who use GS refer to this period as the burn-in time.) A common practice is to cool very slowly from a high temperature to T_*, the critical temperature at which a phase change occurs (Basu & Frazer 1990). At T_*, low-E models are preferred, but the system is still warm enough for the sampler to escape from local energy minima. Below T_*, the system is at least partly frozen and successive samples tend to have the same value. In other words, if $T_\star < 1$, samples taken at $T = 1$ are a good proxy for $\sigma(\mathbf{m})$; but if $T_* > 1$, samples at $T = 1$ may cluster near one mode of $\sigma(\mathbf{m})$ and an impractical amount of time is needed to gather enough samples to represent $\sigma(\mathbf{m})$ accurately. Briefly, our strategy will be to sample at T_*, recognizing that such samples are actually samples from $\sigma(\mathbf{m})^{1/T_*}$, and then use those samples from $\sigma(\mathbf{m})^{1/T_*}$ to calculate expectations with respect to the true $\sigma(\mathbf{m})$. The mathematical basis of the method is known as importance reweighting (Gilks & Roberts 1996).

2 INVERSION

Without invoking Bayes rule, Tarantola & Valette (1982) give one of the clearest expositions of what has, in recent years, come to be called Bayesian inversion. Their fundamental result is

$$\sigma(\mathbf{m}, \mathbf{d}) = \frac{\theta(\mathbf{m}, \mathbf{d})\rho(\mathbf{m}, \mathbf{d})}{\mu(\mathbf{m}, \mathbf{d})}, \tag{2}$$

in which $\mathbf{m}$ is a set of unknown parameters, $\mathbf{d}$ is the data, and σ, θ, ρ and μ are respectively the posterior, theory, prior and non-informative joint probability densities of $\mathbf{m}$ and $\mathbf{d}$. (Duijndam 1988, gives a formal derivation of the Tarantola–Valette relation from Bayes rule, though not for densities). Integration over the data space then gives the posterior marginal for $\mathbf{m}$ in the form

$$\sigma(\mathbf{m}) = \int d\mathbf{d}\sigma(\mathbf{m}, \mathbf{d}). \tag{3}$$

In almost all applications, the Tarantola–Valette relation is simplified in a number of steps. First, one writes

$$\theta(\mathbf{m}, \mathbf{d}) = \theta(\mathbf{d}|\mathbf{m})\mu(\mathbf{m}) = \delta\left[\mathbf{d} - \mathbf{g}(\mathbf{m})\right]\mu(\mathbf{m}) \tag{4}$$

in which δ is the delta function in the data space, $\mathbf{g}(\mathbf{m})$ is an algorithm for calculating $\mathbf{d}$ from $\mathbf{m}$ and $\mu(\mathbf{m})$ is the non-informative density of $\mathbf{m}$. Secondly, one writes

$$\rho(\mathbf{m}, d) = \rho(\mathbf{d})\rho(\mathbf{m}) = n(\mathbf{d}_o - \mathbf{d})\rho(\mathbf{m}), \tag{5}$$

in which n is the noise process of the measuring instrument(s), d_o is the vector of measurements and $\rho(\mathbf{m})$ is the prior density for $\mathbf{m}$. Finally, one writes

$$\mu(\mathbf{m}, \mathbf{d}) = \mu(\mathbf{d})\mu(\mathbf{m}) \tag{6}$$

and substitution of eqs (4)–(6) into eq. (3) gives the final result

$$\begin{aligned}\sigma(\mathbf{m}) &= \rho(\mathbf{m}) \int d\mathbf{d} \frac{\delta[\mathbf{d} - \mathbf{g}(\mathbf{m})]n(\mathbf{d}_o - \mathbf{d})}{\mu(\mathbf{d})} \\ &\propto \rho(\mathbf{m})n[\mathbf{d}_o - \mathbf{g}(\mathbf{m})]\end{aligned} \tag{7}$$

in which, in the second step, the non-informative density of the unknown true data has been assumed to be constant. A quick, formal derivation of the last relation is to ignore the distinction between the measurements and the unknown true data (e.g. to ignore the difference between gravimeter readings and the unknown true gravity) and use conditional probabilities to write $p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})p(\mathbf{m})$. One then identifies $p(\mathbf{m}|\mathbf{d})$ with $\sigma(\mathbf{m})$, $p(\mathbf{d}|\mathbf{m})$ with the likelihood function $n[\mathbf{d} - \mathbf{g}(\mathbf{m})]$ and $p(\mathbf{m})$ with the prior $\rho(\mathbf{m})$.

In GS, one works with an energy function $E(\mathbf{m}) = -\ln \sigma(\mathbf{m})$. As outlined above, $\sigma(\mathbf{m})$ is typically the unnormalized product of a likelihood function $n[\mathbf{d} - \mathbf{g}(\mathbf{m})]$ and a prior, $\rho(\mathbf{m})$. We saw that the likelihood function requires assumptions about the noise process contaminating the data. The examples in this paper assume $\mathbf{d} = \mathbf{g}(\mathbf{m}) + n$ where $\mathbf{d}$ is the data vector, $\mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M)^T$ is the model vector, the function g is non-linear and possibly many-to-one, and n is a Gaussian noise process. In consequence of the Gaussian assumption, the log likelihood is a weighted sum of squares and the energy function has the form

$$E(\mathbf{m}) = [\mathbf{d} - \mathbf{g}(\mathbf{m})]^T\mathbf{C}^{-1}[\mathbf{d} - \mathbf{g}(\mathbf{m})] - \ln \rho(\mathbf{m}) \tag{8}$$

in which $\mathbf{C}$ is the data covariance matrix. Our examples assume no prior information about $\mathbf{m}$ is available and that the geometry of the model space is Euclidean; in such cases, the prior is said to be flat and the term $\ln \rho(\mathbf{m})$ is omitted from the energy.

To apply GS, we find T_* using the method of Basu & Frazer (1990). Briefly, T_* is estimated with a series of fixed-temperature short runs for approximately five starting models. A short run consists of say, 25 sweeps over the parameter space, far fewer sweeps than for an actual inversion. The energy of the final model from each short run is thus a function of a short-run starting model and short-run temperature: $E(\mathbf{m}_\mathrm{s}, T)$, say. This $E(\mathbf{m}_\mathrm{s}, T)$, is averaged over a starting model, and T_* is taken to be the temperature that minimizes the averaged E or, if there is no minimum, the temperature at which averaged E begins to increase. The short-run method works because at high temperatures the GS will not seek low-energy states and at low temperatures the GS tends to freeze into the nearest local minimum. At T_*, the GS is biased toward low-energy states, but it is mobile enough to escape local minima. The short-run method is not the only method for determining T_*; see, for example, Chapman & Jaschke (2001) and Cervelli *et al.* (2001).

It is useful to first consider the case where T_* is less than 1, so the correction developed below is not needed. When T_* is less than 1, we cool slowly to $T = 1$ and then sample at $T = 1$. The samples obtained by using GS in this way are samples from $\sigma(\mathbf{m}) = \mathrm{e}^{-E(\mathbf{m})}/Z_1$. To estimate the expected value of a function $f(\mathbf{m})$, we write

$$\langle f \rangle = \int d\mathbf{m} f(\mathbf{m}) \frac{\mathrm{e}^{-E(\mathbf{m})}}{Z_1} \approx \frac{1}{N_\mathrm{S}} \sum_{k=1}^{N_\mathrm{S}} f\left[\mathbf{m}^{(k)}\right], \tag{9}$$

in which $\int d\mathbf{m}$ is an abbreviation for $\int dm_1 \int dm_2 \ldots \int dm_M$ and $\{\mathbf{m}^{(k)}\}_{k=1}^{N_\mathrm{S}}$ are the samples at $T = 1$. A marginal probability density such as $\sigma_j(\mathrm{m}_j = \tilde{\mathrm{m}}_j)$ is estimated with eq. (2) by choosing $f(\mathbf{m})$ to be a boxcar function that takes the value $1/\varepsilon$ for $\mathrm{m}_j \in [\tilde{\mathrm{m}}_j - \varepsilon/2, \tilde{\mathrm{m}}_j + \varepsilon/2]$ and vanishes elsewhere. The result is $\sigma_j(\mathrm{m}_j = \tilde{\mathrm{m}}_j) \approx N_l/(\varepsilon N_\mathrm{S})$ where N_l is the number of samples for which $\mathrm{m}_j^{(l)} \in [\tilde{\mathrm{m}}_j - \varepsilon/2, \tilde{\mathrm{m}}_j + \varepsilon/2]$. Higher order marginal probability densities are estimated in a similar manner: for example, $\sigma_{jk}(\mathrm{m}_j = \tilde{\mathrm{m}}_j, \mathrm{m}_k = \tilde{\mathrm{m}}_k)$ is estimated by $N_l/(\varepsilon^2 N_\mathrm{S})$ where N_l is now the number of samples $\mathbf{m}^{(l)}$ for which $\mathrm{m}_j^{(l)} \in [\tilde{\mathrm{m}}_j - \varepsilon/2, \tilde{\mathrm{m}}_j + \varepsilon/2]$ and $\mathrm{m}_k^{(l)} \in [\tilde{\mathrm{m}}_j - \varepsilon/2, \tilde{\mathrm{m}}_j + \varepsilon/2]$.

Now consider the more difficult case where T_* is greater than 1. As we cannot sample directly at $T = 1$, we cool to T_* and then

sample. Our estimate of $\langle f \rangle$ is obtained as

$$\begin{aligned}\langle f \rangle &= \int dm\, f(\mathbf{m}) \frac{\mathrm{e}^{-E(\mathbf{m})}}{Z_1} \\ &= \frac{Z_*}{Z_1} \int dm\, f(\mathbf{m}) \mathrm{e}^{-(1-1/T_*)E(\mathbf{m})} \frac{\mathrm{e}^{-E(\mathbf{m})/T_*}}{Z_*} \\ &\approx \frac{Z_*}{Z_1} \frac{1}{N_\mathrm{s}} \sum_{k=1}^{N_\mathrm{S}} f[\mathbf{m}^{(k)}] \mathrm{e}^{-(1-1/T_*)E[\mathbf{m}^{(k)}]}, \end{aligned} \tag{10}$$

in which the samples $\{\mathbf{m}^{(k)}\}_{k=1}^{N_\mathrm{S}}$ are now from $\mathrm{e}^{-E(\mathbf{m})/T_*}/Z_*$. The term reweighting arises from the fact that in eq. (10) each $f(\mathbf{m}^{(k)})$ now has the weight $Z_* Z_1^{-1} N_\mathrm{s}^{-1} \mathrm{e}^{-(1-1/T_*)E[\mathbf{m}^{(k)}]}$, whereas in eq. (2) it had the weight N_s^{-1}. To evaluate the ratio Z_*/Z_1, recall that $Z_1 = \int dm \mathrm{e}^{-E(\mathbf{m})}$. Dividing both sides by Z_* and rearranging gives

$$\begin{aligned}\frac{Z_1}{Z_*} &= \int dm \mathrm{e}^{-E(\mathbf{m})(1-1/T_*)} \frac{\mathrm{e}^{-E(\mathbf{m})/T_*}}{Z_*} \\ &\approx \frac{1}{N_\mathrm{s}} \sum_{k=1}^{N_\mathrm{S}} \mathrm{e}^{-E[\mathbf{m}^{(k)}](1-1/T_*)}, \end{aligned} \tag{11}$$

in which the samples $\{\mathbf{m}^{(k)}\}_{k=1}^{N_\mathrm{S}}$ are the same samples from $\mathrm{e}^{-E(\boldsymbol{m})/T_*}/Z_*$ used in the estimate of $\langle f \rangle$. When $f(\mathbf{m})$ is a marginal probability such as $\sigma_j(\mathbf{m}_j = \tilde{\mathbf{m}}_j)$ there is, in principle, no need to compute the ratio Z_1/Z_* because it is the same for every $\tilde{\mathbf{m}}_j$. However, it is useful to compute Z_1/Z_* anyway, as its convergence with increasing N_S confirms that T_* was well chosen.

For a given inverse problem, how likely is it that T_* is greater than unity? The answer is rooted in the relationship between T_* and data variance, ν^2. Consider a thought experiment in which data from an event are measured simultaneously by less precise and more precise equipment. The more precise measurements will yield data with variances ν^2/k^2 where $k > 1$. If the squared residual, $[\mathbf{d} - g(\mathbf{m})]^\mathrm{T} \mathbf{C}^{-1} [\mathbf{d} - g(\mathbf{m})]$ is used for $E(\mathbf{m})$ in eq. (1), then a factor k^2 will enter the numerator of the exponent. The physics of the melt has not changed, so when k increases (data variance decreases), T_* will increase. This reveals an interesting irony associated with GS-based inversion: when the signal-to-noise ratio (SNR) is higher (i.e. the data variances are smaller and the data are qualitatively better), T_* is more likely to be greater than unity and an uncorrected GS procedure is less likely to be sampling from the true distribution, $\sigma(\mathbf{m})$.

3 SIGNAL-TO-NOISE RATIO, CRITICAL TEMPERATURE AND SAMPLING RANGE

To illustrate our development and better understand when the SNR might be high enough so that $T_* > 1$, we consider a well-known problem from geophysics, the coseismic geodetic inverse problem (CGIP; Ward & Barrientos 1986; Fig. 1a). The CGIP relates the displacements, $\mathbf{d}$, of monuments at the surface of the earth to the source parameters, $\mathbf{m}$, of an earthquake through $g(\mathbf{m})$, a kinematic model of the faulting process such as a dislocation in an elastic half-space (Steketee 1958; Okada 1985). Depending on the depth and size of the earthquake, surface displacements typically range from metres to millimetres and can be measured with a variety of techniques including leveling, Global Positioning System (GPS) geodesy and synthetic aperture radar interferometry (InSAR) (Reilinger *et al.* 2000). Generally the model vector $\mathbf{m}$ contains nine parameters describing the location, orientation and relative displacement of the dislocation approximating the earthquake source (e.g. Hudnut *et al.* 1996).

Fig. 1(a) shows the surface displacement field from a synthetic earthquake as a result of slip on a buried fault plane approximated as a dislocation in an elastic half-space. The error ellipses associated with each displacement vector are geometric representations of the covariance matrix associated with the displacement errors, which were drawn from Gaussians. We assume negligible correlation between the east, north and up components of the displacement; thus, the covariance matrices are diagonal, and the error ellipses have their major and minor axes aligned parallel to the coordinate axes.

Fig. 1(b) shows the determination of T_* from a series of HB short runs (Basu & Frazer 1990) with three different signal-to-noise ratios (SNRs). For a given displacement field, we define the SNR as the median value of the ratio $d_i/V_i^{1/3}$ for each displacement measurement, where d_i is the magnitude of the ith displacement and V_i is the volume of its associated error ellipsoid. The three cases share the same basic displacement field shown in Fig. 1(a): the SNR is assigned by scaling the Gaussian covariances from Fig. 1(a) and the displacements have been perturbed by samples from their associated Gaussians. Fig. 1(b) shows the inverse relationship described above between the SNR and T_*; when the SNR increases, T_* exceeds unity and the procedure given above (i.e. eqs 10 and 11) is necessary to sample from $\sigma(\mathbf{m})$. Although we use the CGIP as a demonstration, the SNR is, of course, a concept independent of the specific inverse problem and we expect its relationship with T_* to be consistent with this example. Evidently, when the SNR is higher, T_* is more likely to be greater than unity (Fig. 1b). There is thus an important caveat associated with the procedures given above for $T_* > 1$: they make it easier than ever to compute dubious posterior distributions from incorrectly underestimated data variances.

Fig. 1(c) shows the full HB inversion and marginal density functions for the middle SNR case with $T_* \approx 4$ (middle curve in Fig. 1b). The light-grey histograms show marginal distributions estimated from samples at T_* without our correction to $T = 1$; the black histograms are the marginals corrected using eqs (10) and (11). The uncorrected marginals centre on the true parameter values, but with much wider distributions than when the correction is used. Note that this tightening of the marginal distributions is a result of the proper reweighting of samples in the formula for the marginals.

As previously stated, samples at T_* are from $\sigma(\mathbf{m})^{1/T_*}$, a smoother distribution than $\sigma(\mathbf{m})$ if $T_* > 1$. In high dimensional parameter spaces there is a higher possibility of missing volumes of significant probability (or low energy, E), so how do we ensure that our samples at T_* do not alias $\sigma(\mathbf{m})$? Aliasing is, of course, a problem inherent in all sampling-based methods, but our implementation of GS reduces this risk by using the technique of Chapman & Jaschke (2001): instead of sampling from a regular grid with fixed parameter values, we sample from a randomized fuzzy grid that allows sweeping over a continuous range of parameter values without much extra computation.

As our method involves sampling from the flattened density $\sigma(\mathbf{m})^{1/T_*}$, one might ask why not sample from the uniform distribution (i.e. let T go to infinity) and then reweight? This does not work because for a fixed number of samples the variance of the estimate given by eq. (10) increases with T, and as $T \to \infty$ a prohibitively large number of samples is needed to lower the variance. This is illustrated in Fig. 2, which shows uncorrected and corrected posterior probability distributions at various T values for representative parameters (width and strike) from the middle SNR case in the synthetic examples of Fig. 1. At values of $T > T_*$, the corrected distributions are significantly tighter or more peaky than at $T_* = 4$

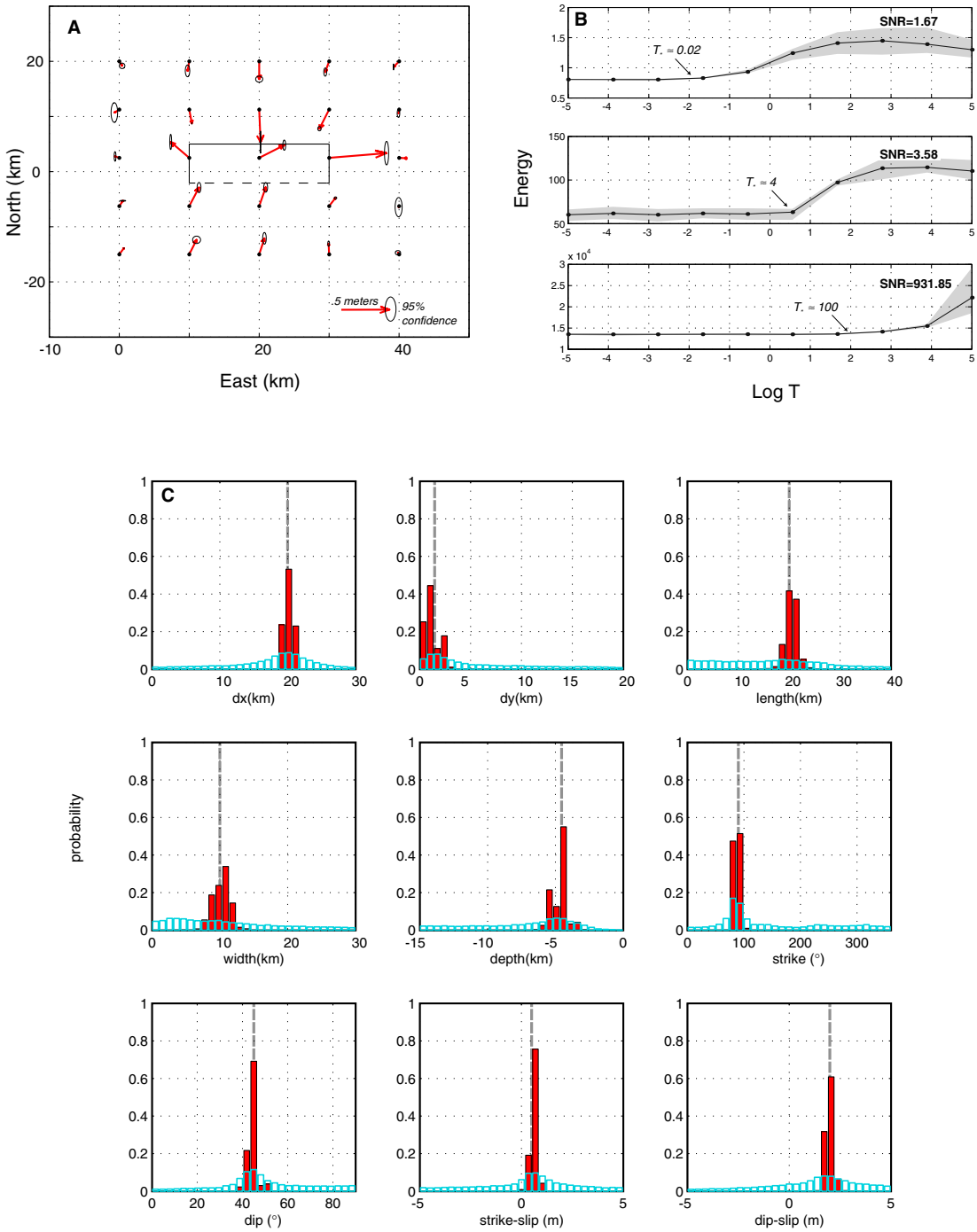

Figure 1. (a) The displacement field (black vectors) for a synthetic earthquake with synthetic errors (ellipses). The dislocation used to create the field (following Okada 1985) is shown as a black rectangle with the down-dip edge dashed. (b) Short-run plots used in determining critical temperature following Basu & Frazer (1990) for three signal-to-noise ratios (SNRs). In addition to plotting average energy for each temperature (solid line), the minimum and maximum energy for each short run is indicated by the shaded envelope. The higher signal-to-noise ratio (SNR) data have critical temperatures (T_*) exceeding unity. (c) Marginal probability distributions for the nine dislocation parameters estimated in the coseismic geodetic inversion problem (CGIP) for the middle SNR case (SNR = 3.58). The grey dashed line is the actual value, the open histograms are the uncorrected Gibbs Sampling (GS) results and the black filled histograms are the corrected GS results.

(Fig. 2) and the peaks are not necessarily centred on the true values. This effect is not related to binning as it is noticed even when the number of bins is increased. It is related, however, to the fact that high-T samples are from a distribution closer to uniform (in Fig. 2, compare the uncorrected marginals at $T_* = 10$ with those at $T_* = 4$) so a greater component of randomness is introduced. Often one of the models has a slightly lower energy than all of the other, very poor models and so in eq. (10) it is favoured over the other models. Additionally, it is important to recall here that while we display marginal distributions for each parameter, E is calculated for the ensemble of parameter values, so the randomness introduced by higher T sampling can manifest in any of the multiple parameters

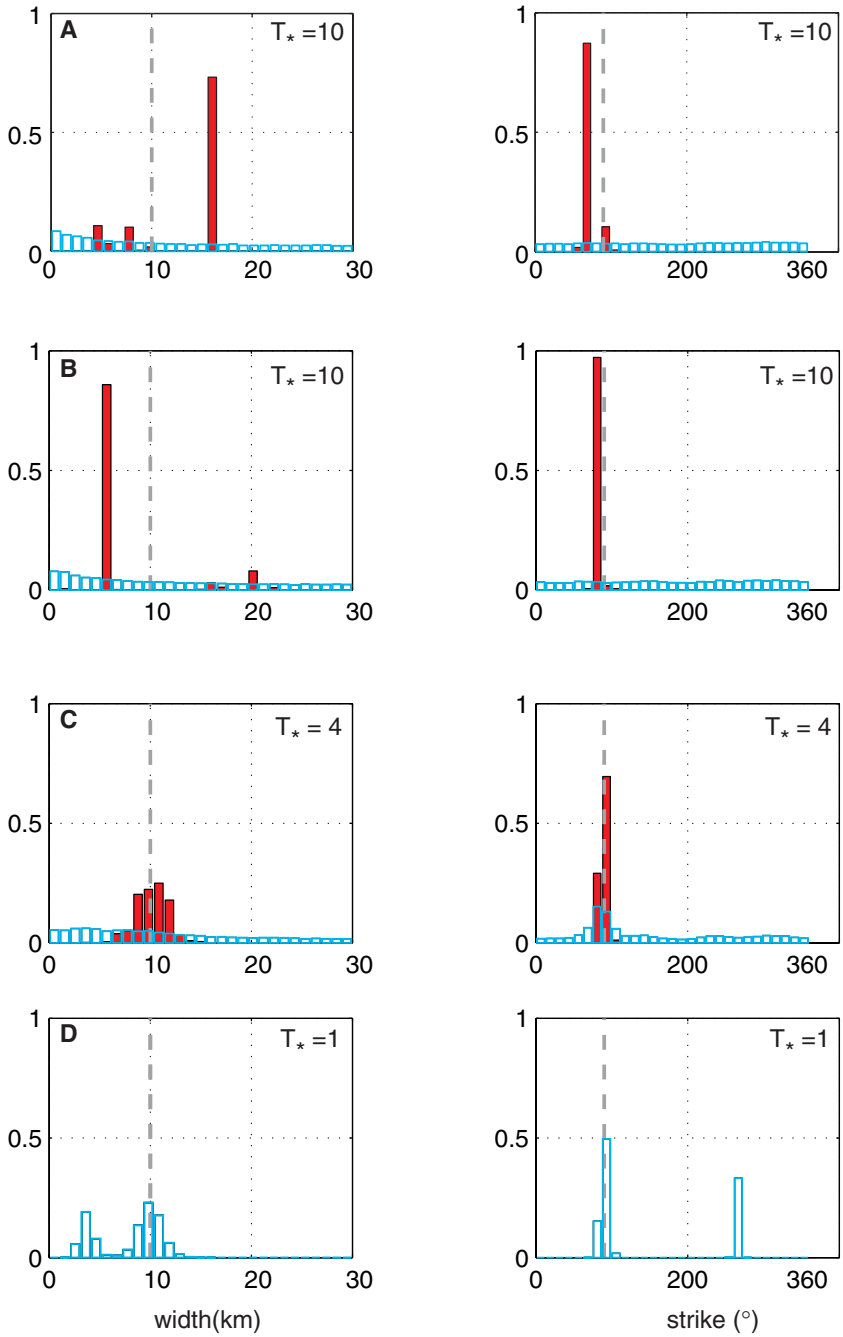

Figure 2. Posterior marginal distributions from Gibbs Sampling (GS) inversion for the width and strike parameters of synthetic data sets at decreasing values of T_*; as in Fig. 1, the grey dashed line is the actual value, the open histograms is the uncorrected GS results and the black filled histograms are the corrected GS results. (a) $T_* = 10$. The peaky and discontinuous nature of the corrected distributions are a result of sampling at too high a T. (b) $T_* = 10$ with a different random seed than (a). The change in position of the corrected histograms reveals that it is too random for confidence in the corrected results. (c) $T_* = 4$. A more appropriate value at which to sample. (d) $T_* = 1$. The bimodal distributions indicate the influence of sampling from local mimima.

of a problem. Fig. 2(b) illustrates this with two entirely independent GS runs on the same data set. Notice that the spurious peaks are in different spots for the different runs (each consisting of a different set of random numbers). Conversely, samples at the most appropriate sampling temperature, $T_* = 4$ (Fig. 2c), give marginals with modes centred on the true parameter values. At $T_* = 1$ (Fig. 2d), the sampled distribution is strongly bimodal with only one mode centred on the true parameter value and the other mode centred on a local minimum. The difference between the samples at $T_* = 4$ and $T_* = 1$ illustrates the well-known risk of sampling at too low a T.

For real data examples, where determining a unique T_* might be difficult, our experience with the synthetic examples motivates an approach in which we discard the notion of sampling at a single temperature and instead sample at various sampling temperatures T_s in parallel. In this way, we also mitigate errors associated with determination of a single critical temperature (e.g. Cervelli *et al.* 2001). This general approach is made possible because of the reweighting correction and the rapidly decreasing cost of computational power. We suggest the following algorithm.

(i) From the short-run results (Fig. 1b), select a range of sampling temperatures, T_s.

(ii) For each T_s, run the full GS/HB algorithm for a fixed number of iterations and compare the marginal distributions obtained by reweighting each run to $T = 1$. If T_s is too high or low, then the corrected marginals will either be too peaky or overly influenced by local minima. The most appropriate marginals will show little change from one value of T_s to another.

4 APPLICATION TO THE NORTHRIDGE EARTHQUAKE

We apply our method to the 1994 Northridge California (M_w6.7) earthquake, a well-studied recent event for which a high-quality coseismic GPS displacement data set is freely available (Hudnut *et al.* 1996; Shen *et al.* 1996; Wald *et al.* 1996; Shearer *et al.* 2003; Fig. 3a). The Northridge event is a good test case because the results from any inversion can be compared to the planar distribution of multiple aftershocks thought to indicate the seismically ruptured fault plane (Mori *et al.* 1995). As they were not available for this study, the east-up and north-up velocity correlations were assumed to be zero; non-zero east-up and north-up correlations might give tighter marginal distributions. Fig. 3(b) shows a series of short runs for densely spaced T values. As expected for a real data example, the short runs do not distinguish an obvious temperature at which to sample, rather there is a range (the dark grey region in Fig. 3b) over which it is appropriate to examine the results from a complete GS/HB inversion. Accordingly, on multiple processors we simultaneously run the full GS/HB inversion for sampling temperatures in the range $1 < T_s < \sim 20$.

Fig. 4 shows that towards the higher end of the T_s range ($\sim 8 < T_s < 18$) representative posterior distributions are either too peaky or their characteristics (i.e. shape, position, smoothness) change significantly with changes in T_s. Towards the lower end of the range ($T_s = 2$), the modes of the width and strike distributions are markedly different from those at slightly higher values of T_s, suggesting that the Gibbs sampler was caught in a local minimum. At slightly higher values, between $\sim T_s = 3$ and 5, the character of the marginal distributions changes little with T_s and we take this as an indication that proper T_s has been found. We quantify this convergence by plotting the integrated differences in marginals between successive T_s values (Fig. 4b); the plots show that the differences are minimum at $3 < T_s < 5$ and then rapidly increase at lower values of T_s where the sampler appears to have been stuck in a local minimum.

Fig. 5 shows the marginal distributions for all parameters for the full inversion with corrected and uncorrected results sampled at $T_s = 5$, the upper end of the range in which distribution differences are smallest. Results from $T_s = 3$ or 4 are similar, but when there is a choice, higher T_s values are preferable to avoid the possibility of the Gibbs sampler becoming stuck in a local minimum. (If T_s is too high, however, then the variance of the marginals increases, as discussed above.) For the Northridge data, our cooling schedule consisted of 15 000 samples and close to four million function evaluations, so efficiency is a consideration. This number of samples appears to be adequate as can be seen in Fig. 6, which shows the convergence of all parameters. For all of the parameters, somewhere near ~5000 samples values are needed for convergence.

The peaks of the corrected distributions from our analysis (Fig. 5) are in good agreement with those estimated independently by Hudnut *et al.* (1996) using a random cost inversion method (Berg 1993). With our approach, however, there is additional

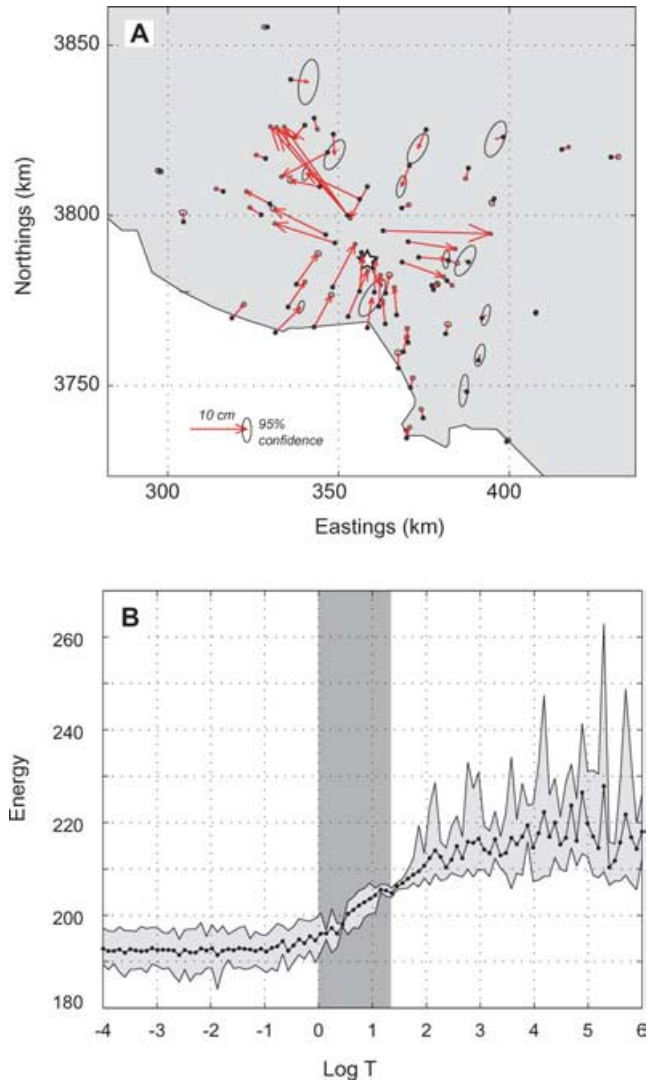

Figure 3. (a) Global Positioning System (GPS) displacement field for the Northridge earthquake from Hudnut *et al.* (1996). The white star is the epicentre of the event. (b) Average short-run energies used in determining sampling temperature range, T_s. The grey shaded area highlights T_s.

important information in the shapes of the individual distributions. For instance, certain parameters (*dx*, *dy*, strike) are better resolved than others (length, width, dip-slip value) and the dip-slip marginal is seen to have a density skewed about its mode of ~3.

Our analysis of the Northridge event confirms the utility of the GS/HB method for a real data example. Moreover, a troubling aspect of GS inversion has been that in many problems some components, m_j, clearly froze before others (Chapman & Jaschke 2001); thus, critical temperature was often a range rather than a point. The dependence on T_* of posterior distributions similar to those of Fig. 3(e) thus made inversion results somewhat subjective. Our method removes one source of subjectivity by making it easier for a Gibbs sampler to, in effect, sample from an arbitrary $\sigma(\mathbf{m})$, although subjectivity still remains from the choice of $\sigma(\mathbf{m})$ itself.

5 CONCLUSIONS

Our numerical experiments suggest that, by running at various temperatures and correcting each run to the common temperature $T = 1$, it is possible to find an optimal sampling temperature that

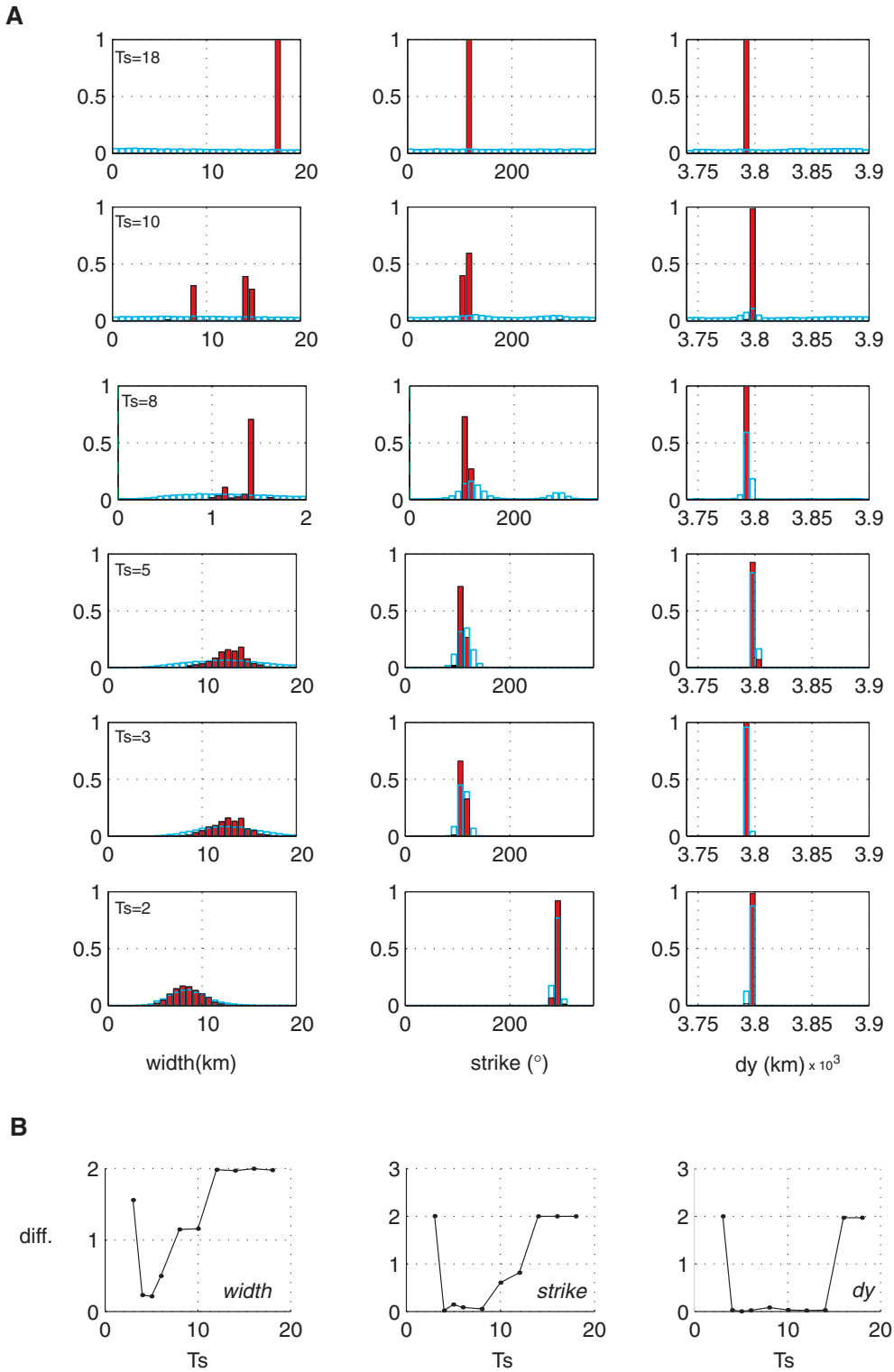

Figure 4. (a) Posterior marginal distributions for Gibbs Sampling (GS) runs on the Northridge data set at representative T_S values from within the sampling range in Fig. 3(b). The results converge to a stable solution on $\sim 3 < T_S < 5$. (b) Plot of *diff* versus T_S, where *diff* is defined as the integrated difference in posterior marginal distributions between successive T_S values (Fig. 4). Low *diff* values mean the successive marginal distributions exhibit little change in character with change in T_S.

is high enough for mixing but low enough that the variance of the marginals is acceptable. Convergence to a stable solution is evaluated *a posteriori* through comparison of marginal distributions obtained at different temperatures. GS can now be used to sample probability functions for which T_* is greater than unity, although it should be kept in mind that the variance of estimates increases with sampling temperature, so longer runs may be needed for such functions, and convergence should be carefully checked. This should facilitate quantitative intercomparisons from inversions based on different data sets. Our inversion of surface displacement data from

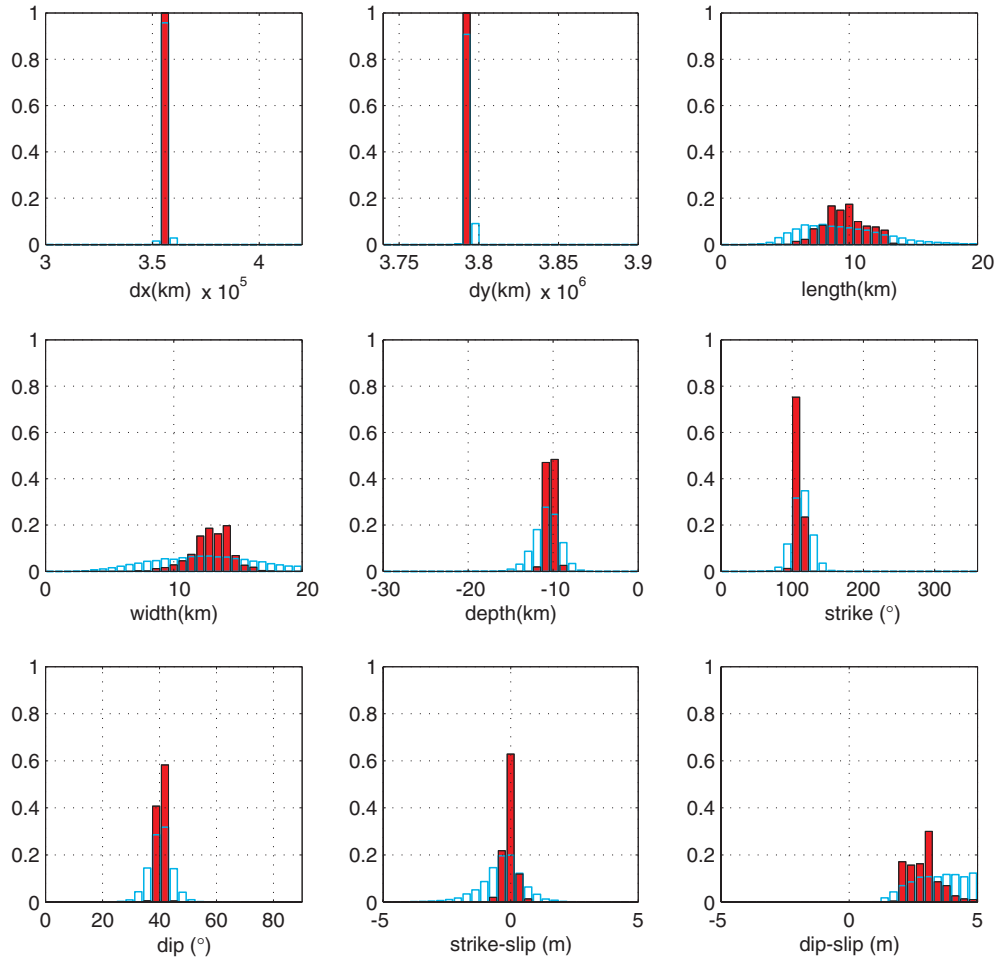

Figure 5. (a) Marginal probability distributions for the nine dislocation parameters estimated for the Northridge data set at $T_s = 5$. The grey dashed line is the actual value, the open histograms are the uncorrected Gibbs Sampling (GS) results and the black filled histograms are the corrected GS results.

the 1994 Northridge earthquake confirms coseismic source parameters from other methods while providing extra information in the form of properly scaled marginal distribution functions.

ACKNOWLEDGMENTS

BAB acknowledges a National Science Foundation grant EAR 9615393, and both BAB and LNF acknowledge the Donors of the Petroleum Research Fund administered by the American Chemical Society. We thank Ken Hudnut for providing a copy of the Northridge GPS data set, and also Janet Becker, Cecily Wolfe and Mike Bevis for discussion and for reviewing of a draft of this paper.

REFERENCES

Basu, A. & Frazer, L.N., 1990. Rapid Determination of the Critical Temperature in Simulated Annealing Inversion, *Science,* **249,** 1409–1412.

Berg, B.A., 1993. Locating global minima in optimization problems by a random-cost approach, *Nature,* **361,** 708–710.

Cervelli, P., Murray, M.H., Segall, P., Aoki, Y. & Kato, T., 2001. Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan, *J. geophys. Res.,* **106,** 11 217–11 237.

Chapman, N.R. & Jaschke, L., 2001. Freeze bath inversion for estimation of geoacoustic parameters, in, *Inverse Problems in Underwater Acoustics,* pp. 15–35, eds Taroudakis, M.I. & Makrakis, G., Springer-Verlag, New York.

Creutz, M., 1980. Monte-Carlo study of quantized SU(2) gauge theory, *Phys. Rev., D.,* **21,** 2308–2315.

Dosso, S. & Wilmut, M.J., 2002. Quantifying data information content in geoacoustic inversion, *IEEE J. Ocean. Eng.,* **27,** 296–304.

Duijndam, A.J.W., 1988. Bayesian estimation in seismic inversion. Part 1: Principles, *Geophys. Prospect.,* **36,** 878–898.

Geman, S. & Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **6,** 721–741.

Gilks, W.R. & Roberts, G.O., 1996. Strategies for improving MCMC, in, *Markov Chain Monte Carlo in Practice,* pp. 89–114, eds Gilks, W.R., Richardson, S.W. & Spiegelhalter, D.J., Chapman and Hall, New York.

Hudnut, K.W. *et al.*, 1996. Co-Seismic Displacements of the 1994 Northridge, California, Earthquake, *Bull. seism. Soc. Am.,* **86,** S19–S36.

Jennison, C., 1993. Discussion on the meaning of the Gibbs sampler and other Markov chain Monte Carlo methods, *J. R. Stat. Soc., B.,* **55,** 53–102.

Kirkpatrick, S., Gelatt, C.D.J. & Vecchi, M.P., 1983. Optimization by simulated annealing, *Science,* **220,** 671–680.

Mori, J., Wald, D.J. & Wesson, R.L., 1995. Overlapping fault planes of the 1971 San Fernando and 1994 Northridge, California earthquake, *Geophys. Res. Lett.,* **22,** 1033–1036.

Mosegaard, K. & Sambridge, M., 2002. Monte Carlo analysis of inverse problems, *Inverse Problems,* **18,** 29–54.

Okada, Y., 1985. Surface deformation due to shear and tensile faults in a half-space, *Bull. seism. Soc. Am.,* **75,** 1135–1154.

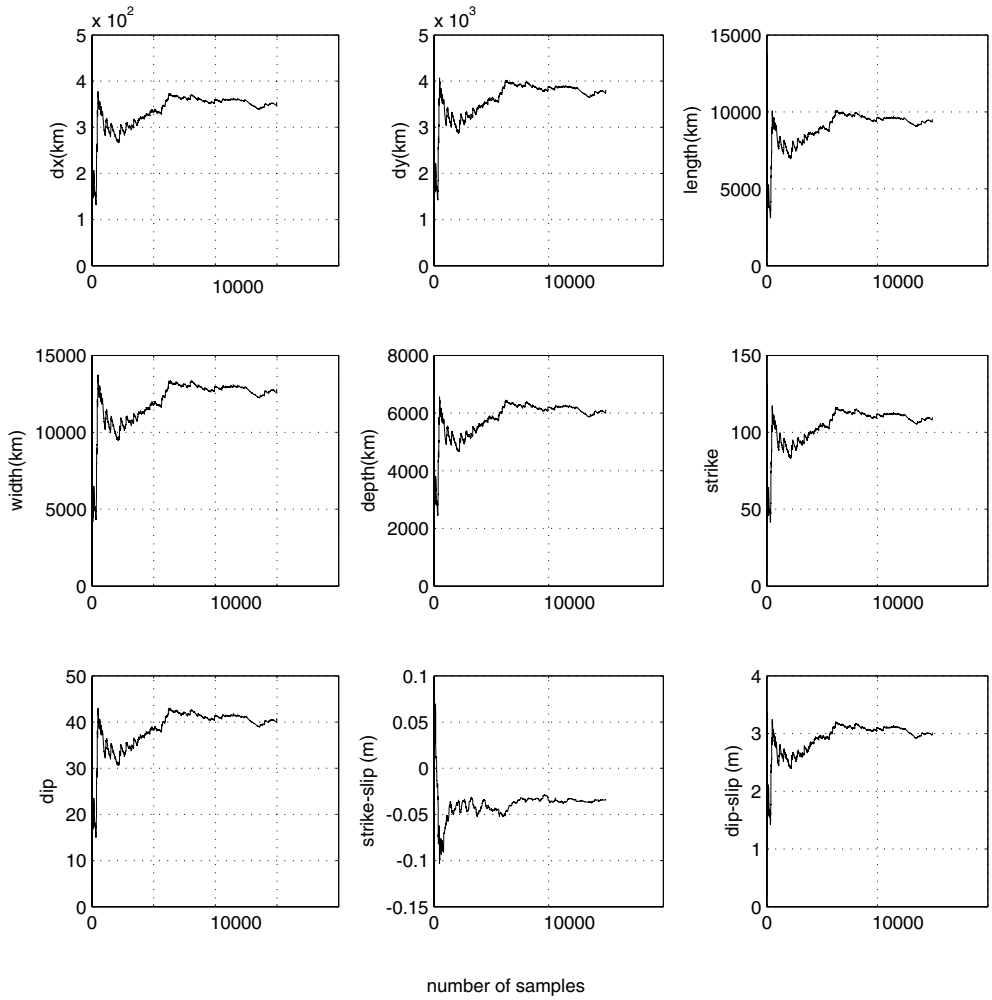

Figure 6. Plots of sample number versus parameter value for each parameter from Fig. 5. When the number of samples exceeds ∼5000, the parameter values converge towards a stable value.

Rebbi, C., 1984. Monte Carlo calculations in lattice gauge theories, in *Applications of the Monte Carlo Method in Statistical Physics,* pp. 277–298, ed. Binder, K., Springer-Verlag, New York

Reilinger, R.E. *et al.*, 2000. Coseismic and Postseismic Fault Slip for the 17 August 1999, M = 7.5, Izmit, Turkey, Earthquake, *Science,* **289**, 1519–1524.

Rothman, D.H., 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation, *Geophysics,* **50,** 2784–2796.

Rothman, D.H, 1986. Automatic estimation of large residual statics corrections, *Geophysics,* **51,** 332–346.

Shearer, P.M., Hardebeck, J.L., Astiz, L. & Richards-Dinger, K.B., 2003. Analysis of similar event clusters in aftershocks of the 1994 Northridge, California, earthquake, *J. geophys. Res.,* **108,** 2035–2049.

Shen, Z.-K., Ge, B.X., Jackson, D.D., Potter, D., Cline, M., Sung, L.-Y., Teng, T.-L.E. & Aki, K.E., 1996. Northridge earthquake rupture models based on the Global Positioning System measurements, *Bull. seism. Soc. Am.,* **86,** S37–S48.

Steketee, J.A., 1958. Some geophysical applications of the elasticity theory of dislocations, *Can. J. Phys.,* **36,** 1168–1198.

Tarantola, A. & Valette, B., 1982. Inverse problems-quest for information, *J. Geophys.,* **50,** 159–170.

Wald, D.J., Heaton, T.H., Hudnut, K.W., Teng, T.-L.E. & Aki, K.E., 1996. The slip history of the 1994 Northridge, California, earthquake determined from strong-motion, teleseismic, GPS, and leveling data, *Bull. seism. Soc. Am.,* **86,** S49–S70.

Ward, S.N. & Barrientos, S.E., 1986. An inversion for slip distribution and fault shape from geodetic observations of the 1983, Borah Peak, Idaho, earthquake, *J. geophys. Res.,* **91,** 4909–4919.