Parallel tempering for strongly nonlinear geoacoustic inversion

Stan E. Dosso[a)]
School of Earth and Ocean Sciences University of Victoria, Victoria, British Columbia V8W 3P6, Canada

Charles W. Holland
The Pennsylvania State University, Applied Research Laboratory, State College, Pennsylvania 16804-0030

Malcolm Sambridge
Research School of Earth Sciences, The Australian National University, Canberra, ACT 0200, Australia

(Received 21 April 2012; revised 28 July 2012; accepted 19 September 2012)

This paper applies parallel tempering within a Bayesian formulation for strongly nonlinear geoacoustic inverse problems. Bayesian geoacoustic inversion consists of sampling the posterior probability density (PPD) of seabed parameters to estimate integral properties, such as marginal probability distributions, based on ocean acoustic data and prior information. This sampling is usually carried out using the Markov-chain Monte Carlo method of Metropolis–Hastings sampling. However, standard sampling methods can be very inefficient for strongly nonlinear problems involving multi-modal PPDs with the potential to miss important regions of the parameter space and to significantly underestimate parameter uncertainties. Parallel tempering achieves efficient/effective sampling of challenging parameter spaces with the ability to transition freely between multiple PPD modes by running parallel Markov chains at a series of increasing sampling temperatures with probabilistic interchanges between chains. The approach is illustrated for inversion of (simulated) acoustic reverberation data for which the PPD is highly multi-modal. While Metropolis–Hastings sampling gives poor results even with very large sample sizes, parallel tempering provides efficient, convergent sampling of the PPD. Methods to enhance the efficiency of parallel tempering are also considered. © *2012 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4757639]

PACS number(s): 43.30.Pc, 43.60.Pt [ZHM] Pages: 3030–3040

I. INTRODUCTION

Knowledge of seabed geoacoustic properties in shallow-water environments is important for a variety of sonar, geophysical, and geotechnical applications. Because direct measurements (e.g., coring) can be difficult and expensive, inferring *in situ* information about seabed model parameters from the inversion of ocean acoustic data has received much attention, e.g., Refs. 1–21. Although earlier inversion approaches sought only optimal parameter estimates by minimizing data misfit using nonlinear (numerical) optimization,[1–7] more recently, Bayesian inversion has been applied to provide quantitative uncertainty analysis for a variety of geoacoustic problems, including the inversion of full-field,[8–14] reflection,[15–17] dispersion,[18,19] and reverberation[20,21] data.

In Bayesian inversion,[22–24] model parameters are considered random variables constrained by measured data and prior information with the goal of estimating integral properties of the multi-dimensional posterior probability density (PPD), such as marginal probability distributions. For nonlinear problems, such as geoacoustic inversion, numerical methods must be applied to estimate these integrals. In particular, the Markov-chain Monte Carlo (MCMC) method of Metropolis–Hastings sampling (MHS)[23,24] has been applied in virtually all Bayesian geoacoustic inversions to date. In MHS, a Markov chain is constructed that samples the PPD by repeatedly perturbing the model parameters and applying a probabilistic condition (the Metropolis criterion) to accept/reject perturbations. However, MHS can be very inefficient for strongly nonlinear problems where the PPD includes multiple isolated regions of high probability (multiple modes). MHS is generally poor at transitioning between modes and prone to become trapped in local high-probability regions, potentially missing important regions of the parameter space and underestimating parameter uncertainties. Multi-modal PPDs have been observed in the inversion of full-field,[8–10,13] reflection,[15] dispersion,[19] and reverberation[20] data and may, in fact, be more prevalent than indicated in the literature given the limitations of MHS which is typically applied to these problems.

The method of parallel tempering[25–27] provides an alternative MCMC approach that can be much more efficient/effective than MHS for problems involving multi-modal PPDs. Parallel tempering is based on running multiple Markov chains at a series of sampling temperatures $T \geq 1$, such that for $T > 1$, lower-probability models are accepted (according to the Metropolis criterion) to more widely search the parameter space and increase the possibility of bridging isolated modes. The Metropolis criterion is also applied to provide probabilistic interchange between chains at different temperatures, thereby providing low-temperature chains access to all regions of the space. Because chains at $T > 1$ provide biased sampling, usually only the $T = 1$ chain is used for integral

a)Author to whom correspondence should be addressed. Electronic mail: sdoss@uvic.ca

 0001-4966/2012/132(5)/3030/11/$30.00

estimates. However, schemes to improve parallel-tempering efficiency by weighting samples collected at $T > 1$ to remove the bias[28] and by running multiple chains at each temperature with the number of chains decreasing with T (because higher-temperature chains mix faster) are also considered here.

This paper compares MHS and parallel tempering for the geoacoustic inversion of (simulated) reverberation data; however, the approach is applicable to any nonlinear inverse problem. The reverberation inversion is shown to involve a highly multi-modal PPD, although the multi-modality might not be detected with limited sampling using MHS. Parallel tempering effectively samples the multiple modes and is found to be orders of magnitude more efficient than MHS, providing a convergent solution in a reasonable sample size while MHS is not close to convergence even for an extremely large sample.

II. THEORY AND ALGORITHMS

A. Bayesian inversion

This section describes a Bayesian approach to geoacoustic inversion;[9,14] more general treatments of Bayesian theory can be found elsewhere.[22–24] Let $\mathbf{d}$ be a vector of N acoustic data and let $\mathbf{m}$ be a vector of M seabed model parameters. In a Bayesian approach, these quantities are considered random variables that obey Bayes' rule

$$P(\mathbf{m}|\mathbf{d}) = \frac{P(\mathbf{d}|\mathbf{m})P(\mathbf{m})}{P(\mathbf{d})}. \tag{1}$$

In Eq. (1), $P(\mathbf{m})$ is the prior probability density, representing the available parameter information independent of the data. $P(\mathbf{d}|\mathbf{m})$ is the conditional probability of $\mathbf{d}$ given $\mathbf{m}$. Interpreted as a function of $\mathbf{d}$, this term represents the residual error distribution. However, when $\mathbf{d}$ represents the (fixed) observed data, the term is interpreted as the likelihood of $\mathbf{m}$, which can generally be written $L(\mathbf{m}) \propto \exp[-E(\mathbf{m})]$, where $E(\mathbf{m})$ is the data misfit function (considered in the following text). $P(\mathbf{m}|\mathbf{d})$ represents the state of information for $\mathbf{m}$ given data and prior and is known as the PPD. Equation (1) can be written

$$P(\mathbf{m}|\mathbf{d}) = \frac{\exp[-\phi(\mathbf{m})]}{Z}, \tag{2}$$

where a generalized misfit, combining data and prior, is defined

$$\phi(\mathbf{m}) = E(\mathbf{m}) - \log_e P(\mathbf{m}), \tag{3}$$

and the normalization term (referred to as the partition function) is given by

$$Z = \int \exp[-\phi(\mathbf{m})]\mathrm{d}\mathbf{m}. \tag{4}$$

To interpret the PPD for multi-dimensional problems generally requires estimating properties defining parameter values, uncertainties, and inter-relationships, such as the maximum *a posteriori* (MAP) model parameters, mean model parameters, model covariance matrix, and marginal probability distributions, defined, respectively, as

$$\hat{\mathbf{m}} = \arg\max_{\mathbf{m}} P(\mathbf{m}|\mathbf{d}) = \arg\min_{\mathbf{m}} \phi(\mathbf{m}), \tag{5}$$

$$\bar{\mathbf{m}} = \int \mathbf{m}P(\mathbf{m}|\mathbf{d})d\mathbf{m}, \tag{6}$$

$$\mathbf{C}_m = \int (\mathbf{m} - \bar{\mathbf{m}})(\mathbf{m} - \bar{\mathbf{m}})^T P(\mathbf{m}|\mathbf{d})d\mathbf{m}, \tag{7}$$

$$P(m_i|\mathbf{d}) = \int \delta(m_i - m_i')P(\mathbf{m}'|\mathbf{d})d\mathbf{m}', \tag{8}$$

with higher-dimensional marginals defined in a manner similar to Eq. (8). For nonlinear inverse problems general solutions to Eqs. (5) to (8) do not exist, and numerical optimization and integration (sampling) methods are required.

This paper considers data with uncorrelated Gaussian-distributed random errors of (unknown) standard deviation σ, leading to likelihood function

$$L(\mathbf{m}, \sigma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{|\mathbf{d} - \mathbf{d}(\mathbf{m})|^2}{2\sigma^2}\right]. \tag{9}$$

A maximum-likelihood estimate for σ can be determined by setting $\partial L/\partial\sigma = 0$, leading to

$$\hat{\sigma}(\mathbf{m}) = [|\mathbf{d} - \mathbf{d}(\mathbf{m})|^2 / N]^{1/2}, \tag{10}$$

which, when substituted into Eq. (9) with additive constants neglected, leads to misfit (negative log-likelihood) function[8,11]

$$E(\mathbf{m}) = N \log_e |\mathbf{d} - \mathbf{d}(\mathbf{m})|. \tag{11}$$

Equation (11) treats σ is an implicit unknown parameter in the inversion, automatically applying the maximum-likelihood estimate corresponding to the current model parameters $\mathbf{m}$ when optimizing or sampling over $\mathbf{m}$ [i.e., $\hat{\sigma}(\mathbf{m})$ varies as $\mathbf{m}$ varies in the inversion]. The $\hat{\sigma}$ values sampled in this manner can be computed using Eq. (10). In practical applications where the form of the error distribution may not be known independently, an assumption of Gaussian errors should be validated *a posteriori* by examining data residuals. This procedure has been carried out for measured reverberation data and validated the Gaussian assumption for the data set considered.[20]

B. Metropolis–Hastings sampling

MCMC methods[23,24] are generally used to sample the PPD to evaluate integrals such as Eqs. (6) to (8), which are of the general form

$$I = \int f(\mathbf{m})P(\mathbf{m}|\mathbf{d})d\mathbf{m}. \tag{12}$$

The common MCMC approach in Bayesian geoacoustic inversion is MHS, which draws samples by perturbing the model $\mathbf{m}_i \rightarrow \mathbf{m}_{i+1}$ and accepting the perturbed model with probability

$$p = \min\left\{1, \frac{P(\mathbf{m}_{i+1}|\mathbf{d})}{P(\mathbf{m}_i|\mathbf{d})}\frac{Q(\mathbf{m}_i;\mathbf{m}_{i+1})}{Q(\mathbf{m}_{i+1};\mathbf{m}_i)}\right\}, \tag{13}$$

where $Q(\mathbf{m}_{i+1};\mathbf{m}_i)$ represents the proposal distribution used to generate $\mathbf{m}_{i+1}$ given $\mathbf{m}_i$. If the perturbed model is not accepted, the current state is repeated: $\mathbf{m}_{i+1} = \mathbf{m}_i$. If the proposal density is symmetric, $Q(\mathbf{m}_i;\mathbf{m}_{i+1}) = Q(\mathbf{m}_{i+1};\mathbf{m}_i)$, the acceptance condition simplifies to the Metropolis criterion

$$p = \min\{1, \exp[\phi(\mathbf{m}_i) - \phi(\mathbf{m}_{i+1})]\}. \tag{14}$$

Drawing K models with MHS, the integral can be estimated

$$I \approx \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{m}_k). \tag{15}$$

An advantage of MHS is that PPD sampling does not require calculating the partition function Z, Eq. (4), as it cancels out in the acceptance criterion.

The (symmetric) proposal distribution used here applies Gaussian-distributed perturbations in a principal-component (rotated) parameter space, which minimizes inter-parameter correlations.[3,29–31] The rotation matrix and perturbation sizes (principal-component standard deviations) are computed from an eigenvalue decomposition of the model covariance matrix. The covariance matrix, Eq. (7), is estimated from an initial burn-in sampling, which is itself initiated using a linearized covariance estimate[31] (burn-in samples are discarded).

C. Parallel tempering

Strongly nonlinear problems in which the PPD involves multiple isolated modes in high dimensions can be particularly challenging for MHS. The method of parallel tempering[25–27] provides an effective alternative approach for such problems based on running a series of parallel interacting Markov chains at sampling temperatures $T = 1$ and $T > 1$ to sample from distributions proportional to $\exp[-\phi(\mathbf{m})/T]$ by generalizing the Metropolis criterion to

$$p = \min\{1, \exp[(\phi(\mathbf{m}_i) - \phi(\mathbf{m}_{i+1}))/T]\}. \tag{16}$$

According to Eq. (16), for $T > 1$ MHS samples from $P(\mathbf{m}|\mathbf{d})^{1/T}$, and lower-probability models are accepted. Hence high-T chains provide wider sampling of the parameter space and the possibility of bridging isolated modes, while low-T chains provide more precise local sampling but are prone to become trapped in localized regions of the space. Parallel tempering improves sampling by providing interchange between chains at different temperatures. Including higher-T chains ensures that the lower-T chains can access all regions of the space while still providing efficient local sampling, resulting in a robust ensemble sampler. Because chains at $T > 1$ provide biased PPD sampling (sample from the PPD raised to the power $1/T$ rather than the PPD), usually only the samples collected by the $T = 1$ chain are retained although an alternative is also considered in this paper.

To derive the interchange probability in parallel tempering,[25,26] consider independent Markov chains at temperatures T_i and T_j with current models $\mathbf{m}_i$ and $\mathbf{m}_j$, respectively. The probability of this joint configuration is equal to the product of the probabilities of the constituent configurations, which is proportional to $\exp[-\phi(\mathbf{m}_i)/T_i - \phi(\mathbf{m}_j)/T_j]$. At a given Monte Carlo step, the joint system can be updated by interchanging the two models, or, equivalently, by interchanging temperatures. The probability of the joint configuration after interchange is $\exp[-\phi(\mathbf{m}_j)/T_i - \phi(\mathbf{m}_i)/T_j]$. Hence according to the Metropolis criterion, this interchange is accepted with probability

$$\begin{aligned} p &= \min\left\{1, \frac{\exp[-\phi(\mathbf{m}_j)/T_i - \phi(\mathbf{m}_i)/T_j]}{\exp[-\phi(\mathbf{m}_i)/T_i - \phi(\mathbf{m}_j)/T_j]}\right\} \\ &= \min\{1, \exp[(\phi(\mathbf{m}_i) - \phi(\mathbf{m}_j))(1/T_i - 1/T_j)]\}. \end{aligned} \tag{17}$$

As in MHS, the partition functions at each temperature cancel out in the acceptance criterion of parallel tempering and need not be computed.

Employing N_c Markov chains in parallel tempering requires of the order of N_c times more computational effort than a single unit-temperature chain. However, for strongly nonlinear problems, the improved mixing due to interchanges between chains at different temperatures can more than compensate for this computational overhead, providing more efficient overall sampling. An efficient parallel tempering algorithm requires an appropriate choice of the number of chains and sampling temperatures of those chains. Common practice is to use a lowest temperature of $T = 1$ and a highest temperature such that the chain does not become trapped in locally optimal regions of the parameter space but is still sensitive to the structure of the PPD. Sampling temperatures are usually chosen to increase logarithmically such that a reasonable acceptance rate (e.g., 25% between adjacent temperatures[26]) is achieved for proposed interchanges. Interchanges can be attempted between adjacent temperatures or between randomly selected temperatures. Interchange attempts are not computationally expensive as they do not require solving the forward problem (i.e., computing acoustic data for a given model) because predicted data for both chains have already been computed in accepting the current models.

This paper considers two approaches to attempt to improve efficiency in parallel tempering. The first is based on the observation that because high-T chains inherently mix more rapidly than low-T chains, it may be advantageous to apply proportionally less computational effort at higher temperatures. One way to accomplish this is to include multiple parallel chains at each temperature and to decrease the number of chains logarithmically with increasing T. For example, Sec. III considers parallel tempering using 16, 8, 4, and 2 chains at temperatures of $T = 1$, 2, 4, and 8, respectively. In this case, 16 of 30 chains sample at $T = 1$ rather than 1 of 4 chains if only a single chain was included at each temperature.

The second attempt to increase efficiency is based on weighting the samples collected at temperatures $T > 1$ to remove the sampling bias,[28,31] and including these in PPD integral estimates. For example, in the case referred to in the

preceding text, if the $T=2$ chains are included, 24/30 chains contribute to the estimate. To derive the appropriate weighting,[28] Eqs. (2) and (12) are combined to write the PPD integral

$$I = \int f(\mathbf{m}) \frac{\exp[-\phi(\mathbf{m})]}{Z} d\mathbf{m}. \tag{18}$$

This result (i.e., an exact integral) is unchanged by multiplying the integrand by $1 = \exp[-\phi(\mathbf{m})/T]/Z_T \times Z_T/\exp[-\phi(\mathbf{m})/T]$, leading to

$$I = \frac{Z_T}{Z} \int f(\mathbf{m}) \exp[-\phi(\mathbf{m})(1-1/T)] \times \frac{\exp[-\phi(\mathbf{m})/T]}{Z_T} d\mathbf{m}, \tag{19}$$

where

$$Z_T = \int \exp[-\phi(\mathbf{m})/T] d\mathbf{m}. \tag{20}$$

According to Eq. (19), drawing K models from $\exp[-\phi(\mathbf{m})/T]/Z_T$ via MHS at temperature T provides an unbiased integral estimate[28]

$$I \approx \sum_{k=1}^{K} f(\mathbf{m}_k) \exp[-\phi(\mathbf{m}_k)(1-1/T)] / \sum_{k=1}^{K} \exp[-\phi(\mathbf{m}_k)(1-1/T)]. \tag{21}$$

Defining weights

$$a_k = \exp[-\phi(\mathbf{m}_k)(1-1/T)], \tag{22}$$

this result can be written concisely as a weighted average

$$I \approx \frac{1}{\sum_{k=1}^{K} a_k} \sum_{k=1}^{K} a_k f(\mathbf{m}_k). \tag{23}$$

For $T=1$, Eq. (23) simplifies to standard MHS, while for $T \to \infty$, it becomes uniform Monte Carlo integration, representing wide, but highly inefficient, sampling. Note that the denominator of Eq. (23) simply represents normalization; for computation of marginal distributions, this can be replaced with the normalization requirement of unit area.

Using the sample weighting of Eq. (21), chains sampled at $T>1$ provide unbiased information. The requirement is that the sample summation at each temperature must be normalized before being averaged with normalized results at other temperatures. This means that only chains that have converged to reasonably stable integral estimates should be included in the average; high-temperature chains that are undersampled (after weighting) can degrade results.

III. REVERBERATION INVERSION

A. Inversion scenario

This section illustrates and compares the various sampling approaches described in the preceding text for Bayesian geoacoustic inversion of reverberation data. The scenario considered here is based on simulated (noisy) reverberation data. Simulation provides a number of advantages for such comparisons in that an appropriate model parameterization is known, and the error statistics are also known and controlled. Hence characteristics of the inversion, such as PPD multi-modality, arise solely from the physics of the forward problem and are not an artifact of a poor choice of parameterization or unaccounted-for sources of error.

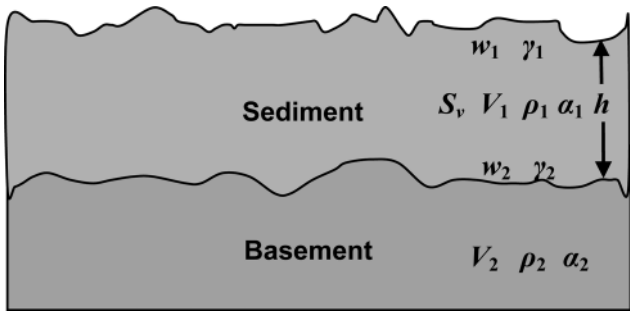

FIG. 1. Schematic diagram of the two-layer seabed model indicating unknown parameters.

The (range-independent) seabed model assumed for the reverberation inversion problem is illustrated in Fig. 1. The seabed is represented by an upper sediment layer of thickness h, sound velocity V_1, density ρ_1, and attenuation α_1, overlying a semi-infinite basement with corresponding parameters V_2, ρ_2, and α_2. Acoustic backscatter occurs at rough interfaces at the top and bottom of the sediment layer and at heterogeneities within the volume of the sediments. The spatial roughness of the upper and lower interfaces is assumed to be isotropic and characterized by a two-dimensional power-law spectrum[32]

$$R_i(\mathbf{k}) = \frac{w_i}{|\mathbf{k}|^{\gamma^i}}, \tag{24}$$

where $\mathbf{k}$ is the horizontal wave vector, w_i is the spectral strength, and γ_i is the spectral exponent with $i=1, 2$ corresponding to the upper and lower interfaces, respectively. The volume-scattering intensity cross-section[33] for the sediments is represented by S_V.

TABLE I. True values, prior bounds, and MAP estimate for seabed parameters.

Parameter and Units	True value	Prior bounds	MAP
h (m)	5	[0, 10]	4.09
V_1 (m/s)	1470	[1450, 1650]	1451
V_2 (m/s)	1660	[1500, 1800]	1579
ρ_1 (g/cm^3)	1.4	[1.2, 2.0]	1.29
ρ_2 (g/cm^3)	1.8	[1.4, 2.2]	1.54
α_1 (dB/m/kHz)	0.05	[0.01, 0.5]	0.078
α_2 (dB/m/kHz)	0.1	[0.01, 0.5]	0.192
γ_1	3	[2.5, 3.9]	2.60
γ_2	3	[2.5, 3.9]	3.44
w_1 (m$^{4-\gamma_1}$)	0.02	[10^{-8}, 0.1]	0.023
w_2 (m$^{4-\gamma_2}$)	0.02	[10^{-8}, 0.1]	0.082
S_V (m^{-3})	10^{-6}	[10^{-9}, 10^{-4}]	4.7×10^{-7}

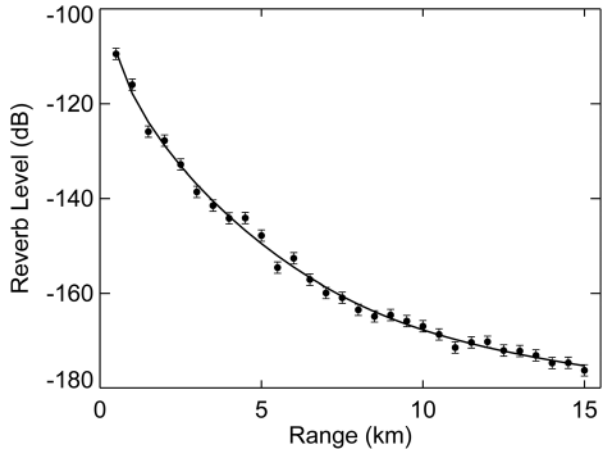

FIG. 2. Simulated noisy reverberation data (circles) and modeled data (solid line) computed for the MAP model. Error bars indicate the maximum-likelihood standard-deviation estimate [Eq. (10) evaluated at the MAP model].

The forward model used to compute reverberation data assumes an iso-velocity water column with the lower boundary defined in terms of the plane-wave reflection coefficient for a layered seabed, including effects of beam displacement. Reverberation is predicted using energy flux theory[34] with independent scattering mechanisms assumed for the top and bottom sediment interfaces and for the sediment volume. Scattering[35] is modeled using the Born approximation, applying perturbation theory for interface scattering and uncorrelated point scatterers for volume scattering.

The particular case considered here involves a 100-m water column of sound velocity 1512 m/s and density 1.03 g/cm^3 (water-column properties are assumed known). The true parameter values for the seabed model and the bounds for uniform prior distributions used in inversion are given in Table I. Note that wide prior bounds are chosen that limit parameter values to physically reasonable values. Simulated reverberation data are computed at a frequency of 2000 Hz for $N = 30$ ranges from 0.5 to 15 km, assuming a signal pulse-length of 0.1 s and an idealized broadside beam width of 10° (no sidelobes) for a towed horizontal receiver array. Gaussian-distributed random errors of standard deviation $\sigma = 1$ dB are added to form the noisy data set shown in Fig. 2 (previous inversions of measured reverberation data indicated Gaussian errors of this magnitude according to statistical residual tests[20]). The fit to the data and estimated standard deviation computed for the MAP model, determined using using numerical optimization (an adaptive hybrid algorithm combining simulated annealing with the downhill simplex method[7]) are also shown. The parameter values for the MAP model are given in Table I.

B. Inversion results

This section considers three approaches to sampling the PPD of the geoacoustic model for the reverberation data in Fig. 2. The first approach is standard MHS based on one Markov chain at sampling temperature $T = 1$ as described in Sec. II B. The second approach is parallel tempering (Sec. II C) based on one Markov chain at each of four temperatures $T = 1$, 2, 4, and 8, with samples from the $T = 1$ chain (i.e., 1 of 4 chains) applied to PPD integration. For brevity, this approach is referred to as PT1. The third approach is

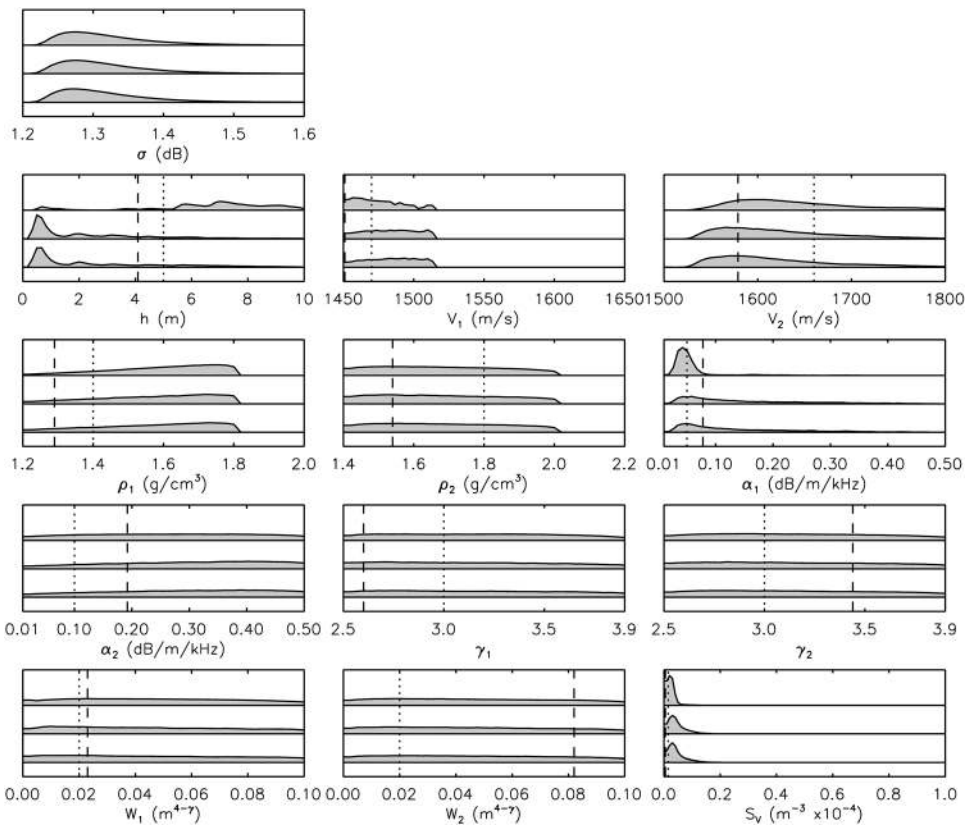

FIG. 3. Marginal probability distributions. Top, middle, and bottom distributions in each panel correspond to results from MHS, PT1, and PT2, respectively. MHS results used 2×10^6 samples; PT1 and PT2 each used 10^6 total samples (all chains). True parameter values are indicated by dotted lines, MAP estimates by dashed lines. Note that plot bounds correspond to prior bounds, and each panel is normalized independently for display purposes.

parallel tempering based on 16, 8, 4, and 2 chains at $T = 1$, 2, 4, and 8, respectively, with samples from the $T = 1$ and $T = 2$ chains (24/30 chains) applied to marginal estimates [the $T = 2$ samples are weighted according to Eq. (21)]. This approach is referred to as PT2. Both parallel-tempering schemes apply interchanges between sampling temperatures selected at random with an average acceptance rate of approximately 20%. For a consistent comparison, the burn-in sampling (12 000 samples) for all three approaches was initiated at the MAP estimate (Table I).

Results for the three inversion approaches are shown in Fig. 3 in terms of marginal probability distributions for the data standard deviation, σ, and for the 12 seabed parameters. The similarity in distributions for σ indicates all three approaches sample models with the same level of misfit. Considering the seabed parameters, two immediate observations can be made from Fig. 3. First, only a subset of the parameters are meaningfully resolved by the reverberation data (it is still worth including all parameters in the inversion, as the Bayesian approach integrates over all parameters, accounting for their sensitivity). Second, the inversion results for the two parallel-tempering approaches are essentially identical while the results for MHS differ significantly for some of the sensitive parameters (e.g., h, V_1, α_1). The results in Fig. 3 are based on large sample sizes: MHS used 2×10^6 samples, while the two parallel-tempering approaches each used total sample sizes of 10^6 (including all chains). It is worth emphasizing here that the goal of Bayesian inversion is to properly represent parameter probabilities over the entire parameter space; the fact that MHS marginals for some parameters (α_1 and S_V) in Fig. 3 appear more concentrated near their true values does not represent a superior solution but rather is an indication that MHS was trapped and did not properly sample the space (i.e., all regions with good misfit were not visited).

To examine the three inversion approaches in more detail, it is useful to consider joint marginal probability distributions for selected parameters at various stages in the sampling. Figure 4 shows joint marginals for combinations of h, V_1, and α_1 as computed using MHS for five sample sizes ranging from 10^5 to 2×10^6 (sample size increases with row from top to bottom). The joint marginals for h and V_1 (left column) indicate a strongly nonlinear inverse problem involving multiple PPD modes and correlated parameters. MHS makes infrequent transitions between modes, such that only one mode is visited in the first 10^5 samples (top row); if the inversion was limited to this sample size, the multi-modality of the problem would not be detected. New

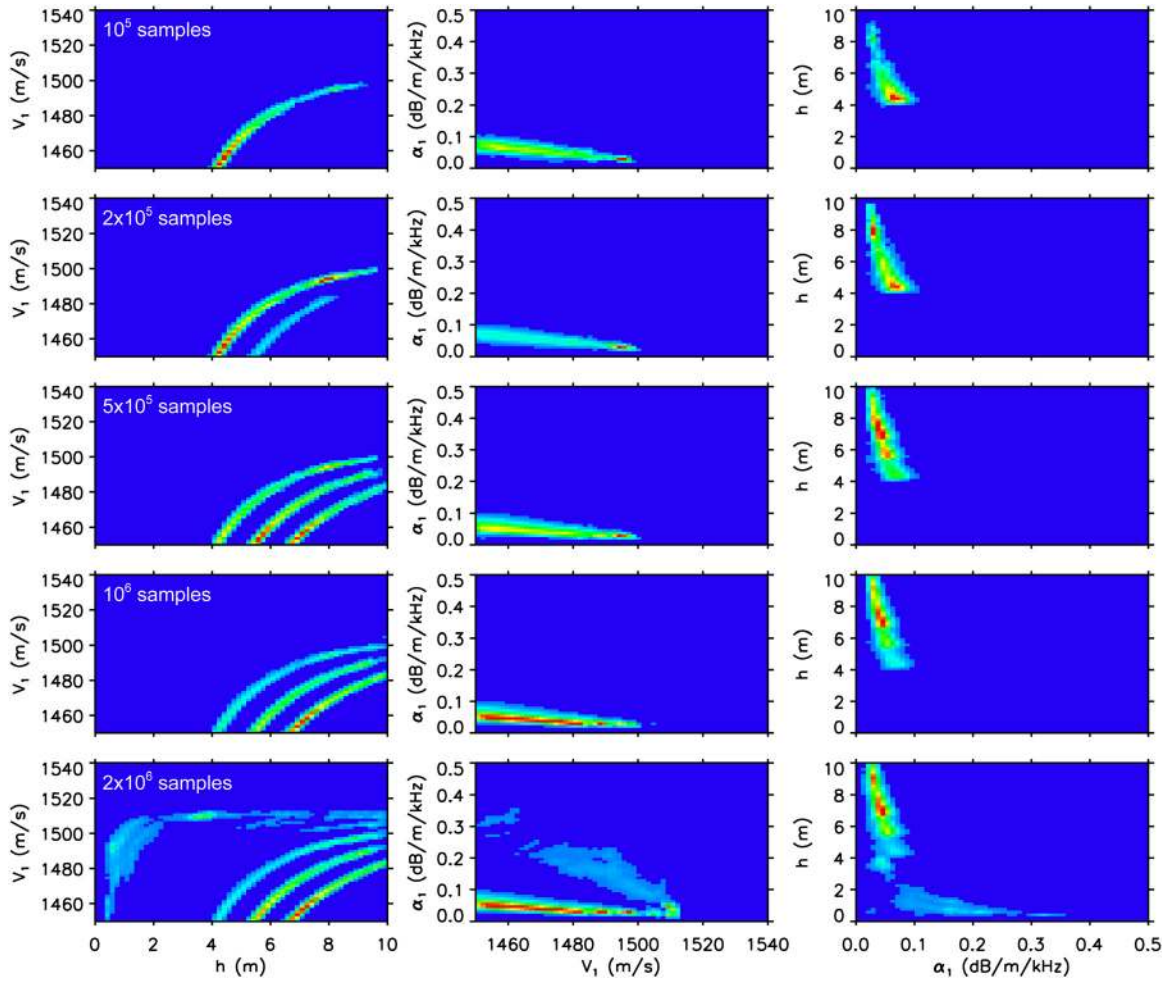

FIG. 4. (Color online) Joint marginal probability distributions for selected parameters computed using MHS for five different numbers of samples corresponding to the five rows (total number of samples indicated on the left panel of each row). Note that the plot width for V_1 is less than the prior bounds. Lighter/warmer shading indicates high probability while darker/cooler shading indicates low probability. Each panel is normalized independently.

modes are gradually visited as the sample size increases, with four modes detected at a practical sample-size limit of 2×10^6 (bottom row), representing ~1 week of computation time on a 2.4-GHz desktop PC. The joint marginals for V_1 and α_1 (center column) and for h and α_1 (right column) are relatively consistent and confined to small regions of the parameter space for up to 10^6 samples (top four rows), then begin to extend into wider regions by 2×10^6 samples (bottom row) as a new mode is visited. From the behavior in Fig. 4, it appears unlikely that all modes have been visited even with 2×10^6 samples, and it is clear the sampling is not close to convergence (typically requiring many visits to all modes).

Figure 5 shows joint marginal PPDs for the same parameters computed using PT1 for four total sample sizes (all chains) from 10^5 to 10^6 (in each case only samples from the $T = 1$ chain, representing one-quarter of the total samples, are used in the marginal estimates). The first row of this figure shows that the multi-modality of the joint marginals are mapped out far better with 10^5 total samples using parallel tempering than with 2×10^6 samples using MHS (Fig. 4, bottom row). In particular, in Fig. 5 at 10^5 samples, the joint marginal for h and V_1 already extends to more modes and the joint marginals for V_1 and α_1 and for h and α_1 extend over wider regions of the parameter space. This structure is not apparent with MHS in Fig. 4 for sample sizes up to 20 times larger. Although it appears all modes have been visited using 10^5 total samples in the top row of Fig. 5 (in comparison to results for larger samples in lower rows), the sampling is sparse in that the marginals are computed from only the 25 000 samples of the $T = 1$ chain. The lower rows in Fig. 5 show how the marginal distributions fill in with larger sample sizes.

Figure 6 shows joint marginal PPDs computed using PT2. The marginals generally resemble those shown in Fig. 5 for PT1, with all modes in the h-V_1 marginal and the full extent of the V_1-α_1 and α_1-h marginals sampled at an early stage. However, the marginal distributions fill in somewhat faster in Fig. 6 as a higher proportion of the total samples contribute (24/30 in Fig. 6 versus 1/4 in Fig. 5). While the difference between results for the two parallel tempering schemes is far smaller than the difference between MHS and parallel-tempering, approaches such as PT2 that make use of a higher proportion of samples may be worthwhile. The close agreement between marginal distributions for the two parallel-tempering schemes is strong evidence that the sampling has converged. Further, the fact that these is little practical difference in results between 10^5 and 10^6 samples in Fig. 6 indicates that extremely large sample sizes are not required to map the complicated multi-modal PPD structure using parallel tempering.

Because samples collected for $T > 1$ are included in PT2, the sampling at non-unity temperatures should be considered. Figure 7 shows two-dimensional histograms of the raw (unweighted) samples collected at temperatures $T = 1$, 2, 4, and 8 for a total of 2×10^5 samples (all chains) in each case. As T increases, the sampled structure broadens (less probable models are accepted) until at $T = 8$, the underlying PPD structure is barely discernible, illustrating how high-T chains can bridge isolated modes at the expense of

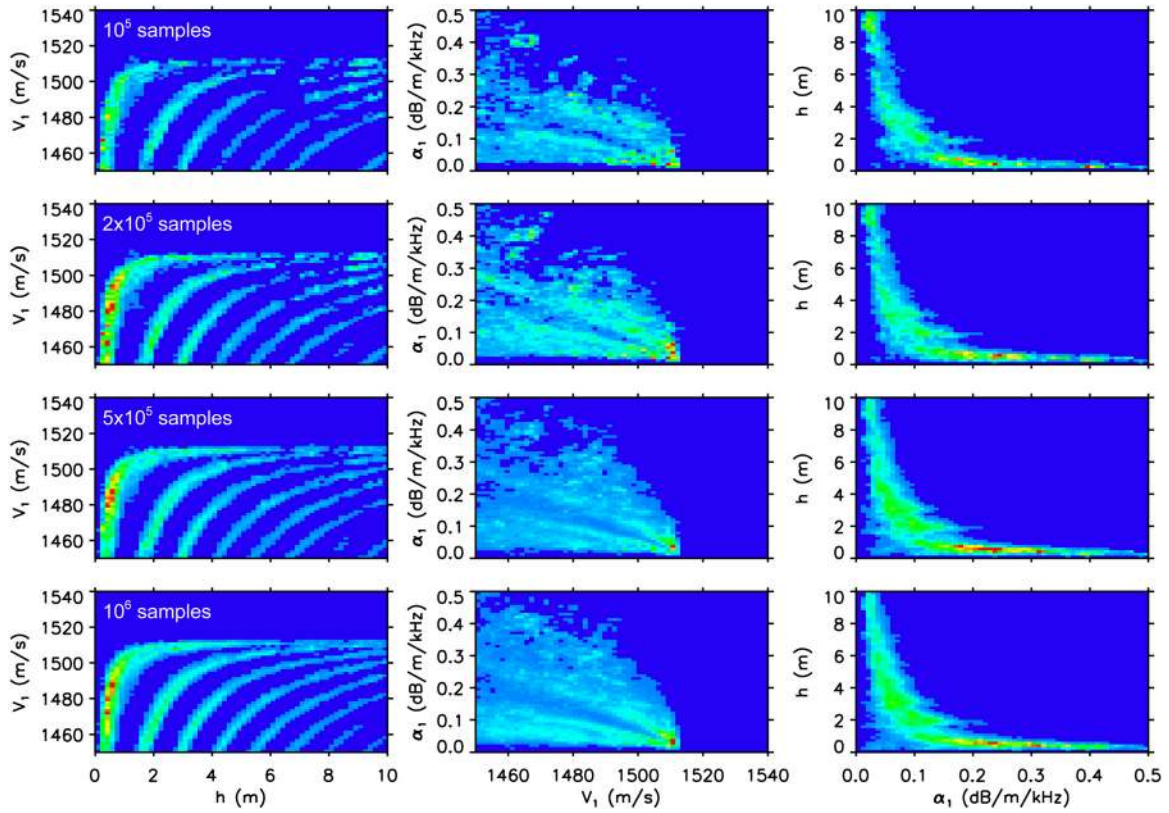

FIG. 5. (Color online) Joint marginal probability distributions for selected parameters computed using PT1. Total number of samples (all chains) indicated on left panels.

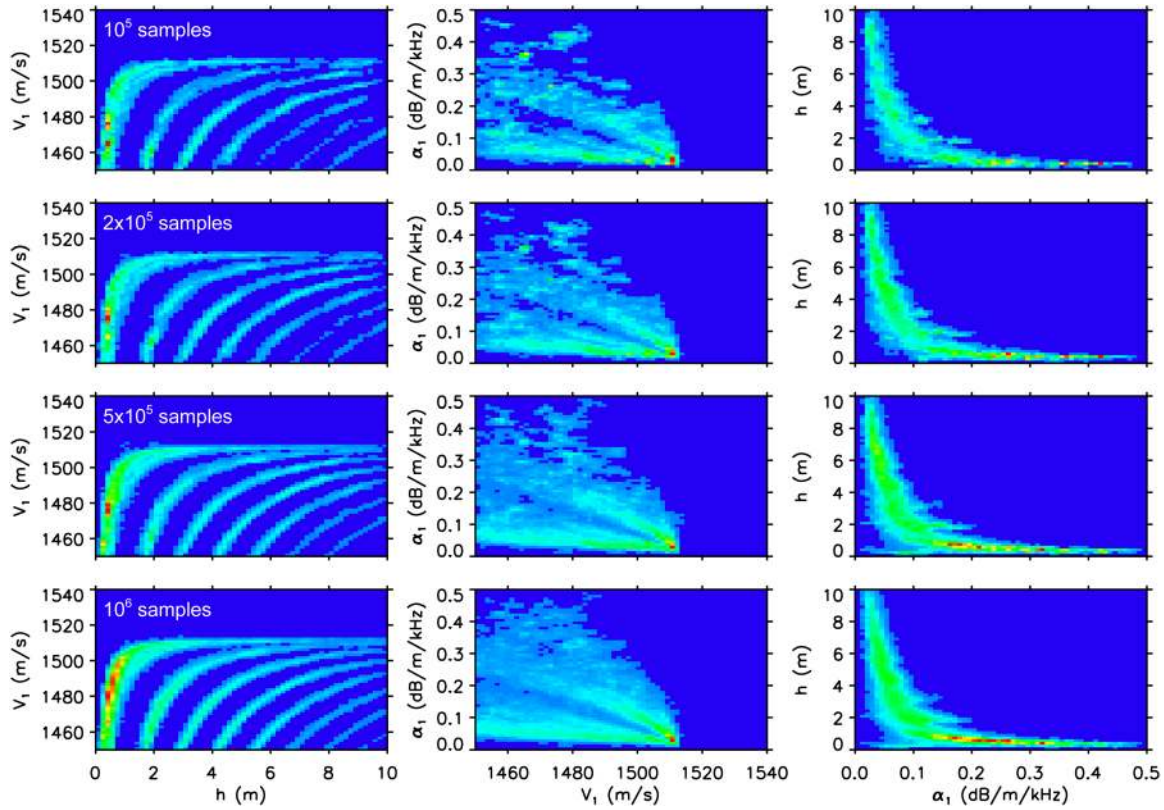

FIG. 6. (Color online) Joint marginal probability distributions for selected parameters computed using PT2. Total number of samples (all chains) indicated on left panels.

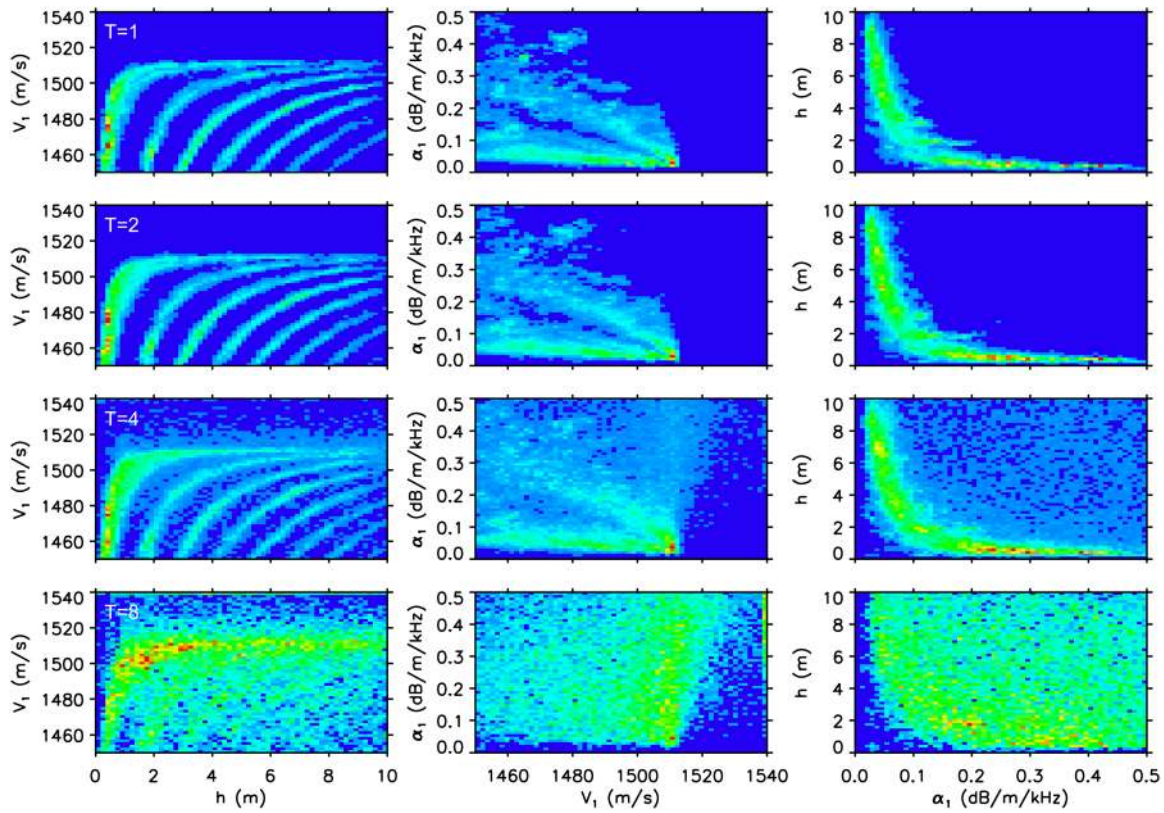

FIG. 7. (Color online) Histograms of samples for selected parameters collected at various temperatures using PT2 (samples not weighted). Sampling temperatures indicated on left panels. Total number of samples (all chains) in each case is 2×10^5.

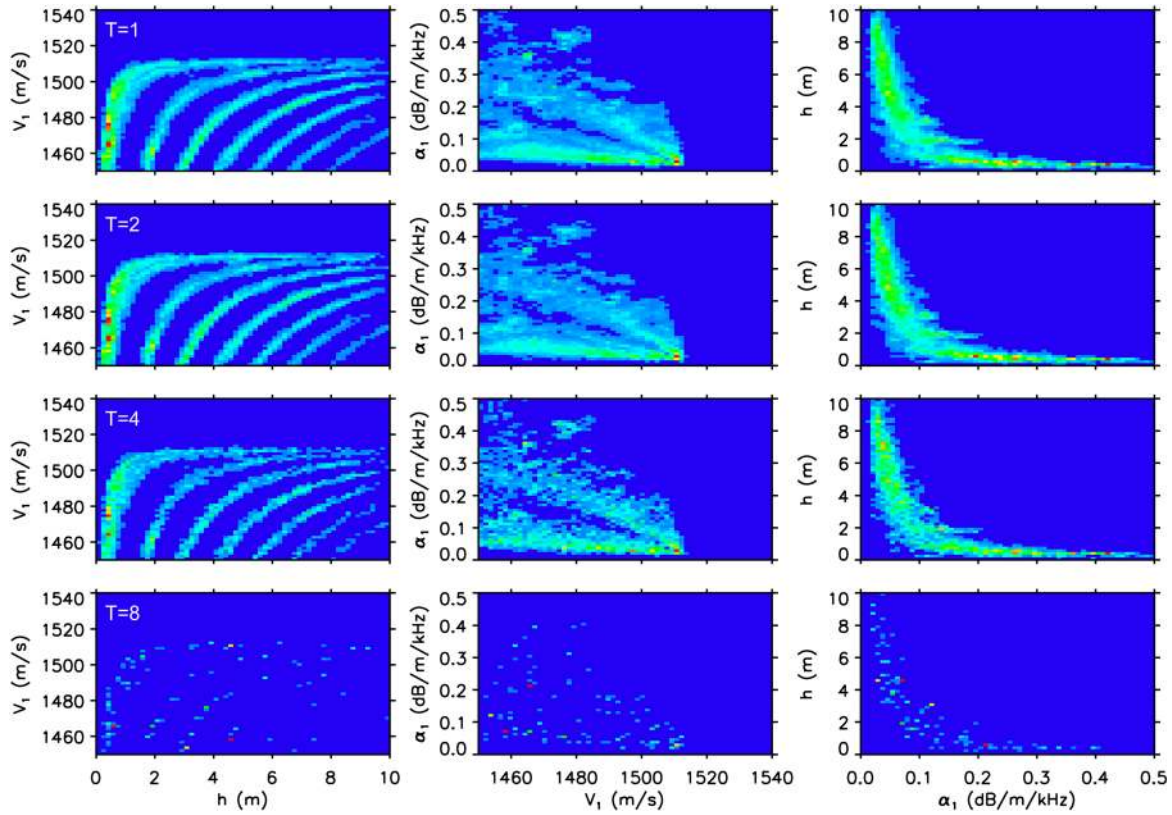

FIG. 8. (Color online) Joint marginal probability distributions for selected parameters computed from Markov chains at various temperatures using PT2 [same samples as Fig. 7, but weighted according to Eq. (21)]. Sampling temperatures indicated on left panels. Total number of samples in each case (all chains) is 2×10^5.

inefficient, biased sampling. Figure 8 shows joint marginal PPDs computed from the samples in Fig. 7 by applying the weighting of Eq. (23) The marginal distributions at each temperature represent the true PPD structure but become increasingly under-sampled at higher temperatures. This plot shows that the $T = 1$ and 2 chains give virtually identical results, indicating sufficient convergence to include in the marginal estimates in PT2. The chains at $T = 4$ differs slightly from the lower temperature chains (appears slightly under-sampled), while the $T = 8$ chain is strongly under-sampled; hence, results at these two higher temperatures are not included in the PT2 results.

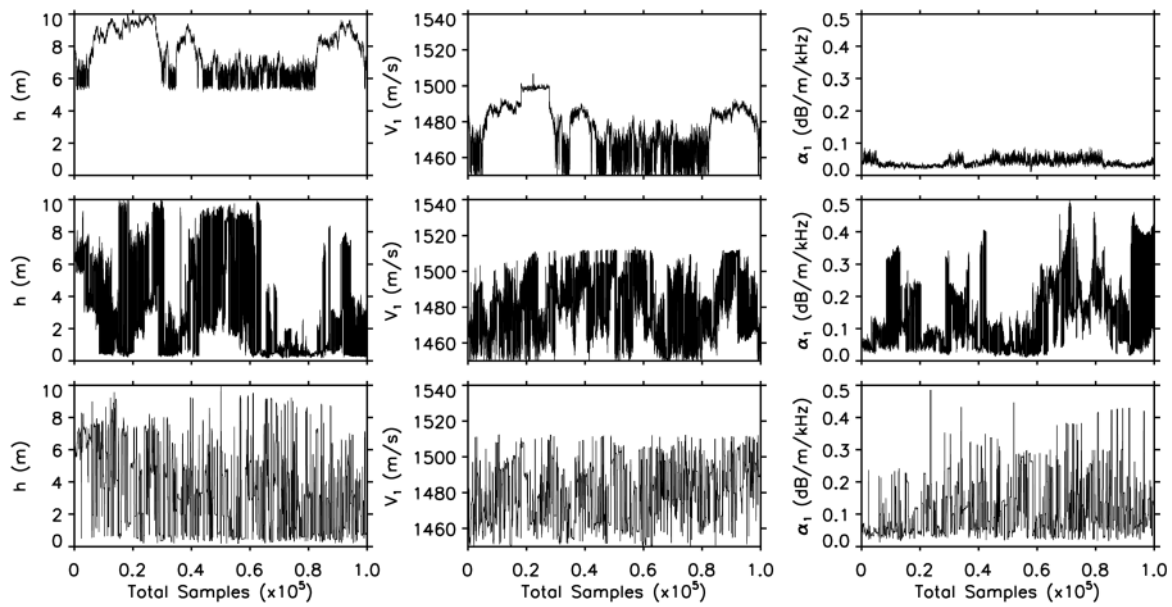

FIG. 9. Chain mixing at $T = 1$ for MHS (top row), PT1 (middle row), and PT2 (bottom row, one chain shown). Horizontal axes indicate the total number of samples considered by the algorithm (i.e., chains in middle and bottom rows include 1/4 and 1/30 as many samples as the top row, respectively).

As a final comparison of the three sampling approaches, Fig. 9 considers Markov-chain mixing (i.e., how rapidly the chain wanders the parameter space and converges to a stationary distribution). The top row of Fig. 9 shows the first 10^5 samples for sediment parameters h, V_1, and α_1 collected using MHS. MHS moves slowly in small steps over the parameter space with particularly small variations in α_1. No transitions between PPD modes occur in this section of the sampling (upper row of Fig. 4). The middle row of Fig. 9 shows chain mixing for PT1. Although the algorithm considered 10^5 total samples, only the 2.5×10^4 samples of the $T = 1$ chain are plotted. Nonetheless the chain mixes far better than the MHS chain with many transitions between modes and wide sampling of all parameters. Finally, the bottom row of Fig. 9 shows one of the $T = 1$ chains for PT2, consisting of 3333 samples. Even for this small sample, the chain mixes very well, jumping continuously between modes and varying rapidly for all parameters.

IV. SUMMARY AND DISCUSSION

This paper considered parameter sampling approaches in a Bayesian formulation of strongly nonlinear geoacoustic inverse problems. It was shown that the common practice of Metropolis–Hastings sampling can be impractically slow for problems in which the posterior probability density includes multiple modes. Further, multiple modes can go undetected with insufficient sampling, which can result in high-probability regions of the parameter space being neglected and parameter uncertainties under-estimated. Parallel tempering, which applies multiple Markov chains at sampling temperatures $T \geq 1$ with probabilistic interchange between chains, was shown to greatly improve sampling of multimodal PPDs.

Two parallel tempering schemes were considered: PT1, based on one Markov chain per temperature with the samples collected by the $T = 1$ chain applied to estimate PPD integrals, and PT2, based on multiple chains per temperature with the number of chains decreasing with T (because high-T chains mix faster), and including (weighted) samples collected at $T > 1$, provided these chains have converged sufficiently. PT2 has the advantage that a higher proportion of the chains are used in estimating PPD integrals than PT1 (24/30 versus 1/4 in the case considered here).

MHS and the two parallel-tempering schemes were illustrated and compared for an inversion scenario of estimating seabed geoacoustic parameters from shallow-water reverberation data that was characterized by a highly multimodal PPD. The rigorous quantification of convergence rates for nonlinear sampling algorithms is a difficult problem and is not pursued in this paper. Nonetheless it is clear from the inversion results here that parallel tempering can be orders of magnitude more efficient than standard MHS for difficult multi-modal problems, providing convergent results with a reasonable sample size to problems which MHS cannot address in a practical sense. Finally, PT2 appeared to provide a modest qualitative improvement in sampling over PT1 for the case considered here.

[1]M. D. Collins, W. A. Kuperman, and H. Schmidt, "Nonlinear inversion for ocean-bottom properties," J. Acoust. Soc. Am. **92**, 2770–2783 (1992).
[2]S. E. Dosso, M. L. Yeremey, J. M. Ozard, and N. R. Chapman, "Estimation of ocean-bottom properties by matched-field inversion of acoustic field data," IEEE J. Ocean. Eng. **18**, 232–239 (1993).
[3]M. D. Collins and L. Fishman, "Efficient navigation of parameter landscapes," J. Acoust. Soc. Am. **98**, 1637–1644 (1995).
[4]P. Gerstoft, "Inversion of acoustic data using a combination of genetic algorithms and the Gauss-Newton approach," J. Acoust. Soc. Am. **97**, 2181–2190 (1995).
[5]M. Siderius, P. Gerstoft, and P. Nielsen, "Broadband geoacoustic inversion from sparse data using genetic algorithms," J. Comp. Acoust. **6**, 117–134 (1998).
[6]P. Gerstoft and Z.-H. Michalopoulou, "Global optimization in matched field inversion," in *Proceedings of the 4th European Conference on Underwater Acoustics* (Italian National Research Council, Rome, 2000), pp. 27–32.
[7]S. E. Dosso, M. J. Wilmut, and A.-L. S. Lapinski, "An adaptive-hybrid algorithm for geoacoustic inversion," IEEE J. Ocean. Eng. **26**, 324–336 (2001).
[8]P. Gerstoft and C. F. Mecklenbräuker, "Ocean acoustic inversion with estimation of *a posteriori* probability distributions," J. Acoust. Soc. Am. **104**, 808–819 (1998).
[9]S. E. Dosso and P. L. Nielsen, "Quantifying uncertainty in geoacoustic inversion. II: Application to broadband, shallow-water data," J. Acoust. Soc. Am. **111**, 143–159 (2002).
[10]D. J. Battle, P. Gerstoft, W. S. Hodgkiss, W. A. Kuperman, and P. L. Nielsen, "Bayesian model selection applied to self-noise geoacoustic inversion," J. Acoust. Soc. Am. **116**, 2043–2056 (2004).
[11]S. E. Dosso and M. J. Wilmut, "Data uncertainty estimation in matched-field geoacoustic inversion," IEEE J. Ocean. Eng. **31**, 470–479 (2006).
[12]D. Tollefson, S. E. Dosso, and M. J. Wilmut, "Matched-field geoacoustic inversion with a horizontal array and low-level source," J. Acoust. Soc. Am. **120**, 221–230 (2006).
[13]C. Yardim, P. Gerstoft, and W. S. Hodgkiss, "Geoacoustic and source tracking using particle filtering: Experimental results," J. Acoust. Soc. Am. **128**, 75–87 (2010).
[14]S. E. Dosso and J. Dettmer, "Bayesian matched-field geoacoustic inversion," Inverse Probl. **27**, 1–23 (2011).
[15]M. Riedel, S. E. Dosso, and L. Beran, "Uncertainty estimation for amplitude variation with offset (AVO) inversion," Geophysics **68**, 1485–1496 (2003).
[16]S. E. Dosso and C. W. Holland, "Geoacoustic uncertainties from viscoelastic inversion of seabed reflection data," IEEE J. Ocean. Eng. **31**, 657–670 (2006).
[17]J. Dettmer, S. E. Dosso, and C. W. Holland, "Uncertainty estimation in seismo-acoustic reflection travel-time inversion," J. Acoust. Soc. Am. **122**, 161–176 (2007).
[18]J. Dettmer, S. E. Dosso, and J. C. Osler, "Bayesian evidence computation for model selection in nonlinear geoacoustic inference problems," J. Acoust. Soc. Am. **128**, 3406–3415 (2010).
[19]H. Dong and S. E. Dosso, "Bayesian inversion of interface-wave dispersion for seabed shear-wave velocity profiles," IEEE J. Ocean. Eng. **36**, 1–11 (2011).
[20]S. E. Dosso, P. L. Nielsen, and C. H. Harrison, "Bayesian inversion of reverberation and propagation data for geoacoustic and scattering parameters," J. Acoust. Soc. Am. **125**, 2867–2880 (2009).
[21]S. Ivansson, "Parameter reduction and Laplace distributed data errors in connection with Bayesian reverberation inversion," in *Proceedings of the 9th European Conference on Underwater Acoustics* (Istanbul, 2010), pp. 543–551.
[22]A. Tarantola, *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation* (Elsevier, Amsterdam, 1987), pp. 1–583.
[23]W. R. Gilks, S. Richardson, and G. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman and Hall, London, 1996), pp. 1–486.
[24]M. Sambridge and K. Mosengaard, "Monte Carlo methods in geophysical inverse problems," Rev. Geophys. **40**, 3-1–3-29 (2002).
[25]C. J. Geyer, "Markov chain Monte Carlo maximum likelihood," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (New York, 1991), pp. 156–163.
[26]D. J. Earl and M. W. Deem, "Parallel tempering: Theory, applications, and new perspectives," Phys. Chem. Chem. Phys. **7**, 3910–3916 (2005).
[27]A. Jasra, D. A. Stephens, and C. C. Holmes, "On population-based simulation for static inference," Stat. Comp. **17**, 263–279 (2007).

[28]B. F. Brooks and L. N. Frazer, "Importance reweighting reduces dependence on temperature in Gibbs samplers: An application to the inverse coseismic geodetic problem," Geophys. J. Int. **161**, 12–20 (2005).
[29]B. Nolte and L. N. Frazer, "Vertical seismic profile inversion with genetic algorithms," Geophys. J. Int. **117**, 162–179 (1992).
[30]A. Malinverno, "Parsimonious Bayesian Markov chain Monte Carlo inversion in nonlinear geophysical problems," Geophys. J. Int. **151**, 675–688 (2002).
[31]S. E. Dosso and M. J. Wilmut, "Uncertainty estimation in simultaneous Bayesian tracking and environmental inversion," J. Acoust. Soc. Am. **124**, 82–97 (2008).
[32]D. R. Jackson, R. I. Odom, M. I. Boyd, and A. N. Ivankin, "A geoacoustic bottom interaction model (GABIM)," IEEE J. Ocean. Eng. **35**, 603–617 (2010).
[33]A. Ivankin, "A unified approach to volume and roughness scattering," J. Acoust. Soc. Am. **103**, 827–837 (1999).
[34]J. X. Zhou, "The analytic method of angular power spectrum, range and depth dependence of the echo-to-reverberation ratio in shallow water sound field," Acta Acust. **5**, 86–99 (1980).
[35]J. E. Moe and D. R. Jackson, "First-order perturbation solution for rough surface scattering cross section including the effects of gradient," J. Acoust. Soc. Am. **96**, 1748–1754 (1994).