GEOPHYSICS, VOL. 87, NO. 5 (SEPTEMBER-OCTOBER 2022); P. M163–M177, 16 FIGS., 3 TABLES.
10.1190/GEO2021-0665.1

A Bayesian multivariate model using Hamiltonian Monte Carlo inference to estimate total organic carbon content in shale

Shib Sankar Ganguli[1], Mohamed Mehdi Kadri[2], Akash Debnath[1], and Souvik Sen[3]

ABSTRACT

The prediction of total organic carbon (TOC) content using geophysical logs is one of the key steps in shale reservoir characterization. Various empirical relations have previously been used for the estimation of TOC content from well-logs; however, uncertainty quantification in the model estimation is often ignored while performing TOC estimation in a deterministic framework. We introduce the problem of TOC estimation in a Bayesian setting with the goal of enhancing the TOC content prediction together with the quantification of the uncertainty in the model prediction. To signify the uncertainty, we draw random samples of model parameters from the posterior distribution by realizing multidimensional stochastic processes within the Hamiltonian Monte Carlo algorithm. The posterior model for the variables that influence TOC estimation is conditioned on the available well-log observations and is further defined by a priori and likelihood distributions. We demonstrate examples of applications of this approach to estimate the TOC content on two real field data sets from the well-known Devonian Duvernay shale of Western Canada and the Silurian shale of the Ahnet Basin. The accuracy in the estimation is arbitrated by comparing the prediction results with those obtained using the two most widely used empirical models. Corroborating the results by the laboratory-measured TOC contents demonstrate that the Bayesian approach offers a more reliable and better confidence in predictions when compared with the empirical models, as it provides additional information on the prediction uncertainty. Finally, the implications of the present approach are derived in terms of depositional environments to characterize the high TOC content zone in the studied organic shale formations.

INTRODUCTION

Total organic carbon (TOC) content is one of the key parameters in the geophysical characterization of shale oil/gas reservoir assessment that has an explicit impact on the source-rock quality and shale oil/gas-in-place estimations (Jarvie et al., 2007; Vernik and Milovac, 2011; El Sharawy and Gaafar, 2012). In practice, this variable can be reliably obtained from the analysis of cutting and core samples in a limited number using Rock-Eval pyrolysis in the laboratory, which is, however expensive, and burdensome. As a matter of fact, TOC data are sparsely spread, offering a hindrance to the investigation. To overcome this, many empirical relations and mathematical equations have been developed and widely applied to predict TOC values from wireline well-log variables. In general, the presence of organic matter in shale shows a unique response on wireline log variables. Tixier and Curtis (1967) propose to use density log to identify source-rock signals on wireline log based on a linear volumetric average of density log. Supernaw et al. (1978) and Fertl and Rieke (1980) derive a mathematical relationship based on the gamma-ray (GR) spectral well-log data to identify the shale resource potential by estimating TOC, and subsequently, this approach also has been applied by several workers (Lüning and Kolonic, 2003; Wang et al., 2014; Renchun et al., 2015). Schmoker and Hester (1983) propose a model that predicts TOC based on the reciprocal of the density log. Even though this method entails a small quantity of input bulk density data, it may not be useful in some cases, especially when the bulk density is influenced by the reservoir or geologic variables. Meyer and Nederlof (1984) establish a qualitative technique involving a relation-

Manuscript received by the Editor 8 October 2021; revised manuscript received 26 April 2022; published ahead of production 17 May 2022; published online 4 July 2022.
[1]CSIR-National Geophysical Research Institute, Hyderabad, Telangana, India. E-mail: shibg@ngri.res.in (corresponding author); akashdebnath836@gmail.com.
[2]Université Kasdi Merbah Ouargla, Laboratoire de Géologie du Sahara, Ouargla, Algeria. E-mail: kadri.univ@gmail.com.
[3]Geologix Limited, Mumbai, Maharashtra, India. E-mail: souvikseniitb@gmail.com.

ship between resistivity, sonic, and bulk density logs to identify source rocks from wireline logs. Mendelzon and Toksoz (1985) present a multivariate regression model with a high coefficient of determination to identify quantitative relationships between the wireline logs and core-derived TOC values. Passey et al. (1990) propose the δ log R method by overlaying porosity logs (e.g., sonic, neutron, and density), and resistivity log, which is the most widely used technique over the last three decades. However, one needs to be cautious while selecting the log baseline manually because it may vary significantly from well-to-well and across the formations, often resulting in ambiguous TOC estimation. Carpentier et al. (1991) introduce the CARBon Organic LOG (CARBOLOG) method in which in situ organic matter content is estimated by blending physical properties with sonic transit time and resistivity data. Huang and Williamson (1996) present a data-driven approach, namely artificial neural network modeling to infer an accurate relationship between TOC content and well-log variables. Kamali and Mirshady (2004) develop a hybrid method based on the δ log R method and neuro-fuzzy approaches to predict the rich intervals with high TOC content. Jacobi et al. (2008) establish a method for TOC estimation based on the discrepancy between grain density and inorganic grain density, which was useful to distinguish between source rocks and nonsource rocks. Further, numerous researchers revisit the δ log R method and devise various improved methods for the prediction of TOC content (Pan et al., 2009; Bakhtiar et al., 2011; Liu et al., 2015; Hu et al., 2016; Zhu et al., 2019). With the advent of intelligent systems and machine learning techniques and recognition of their potential, these approaches have been recently applied to analyze the relationship between TOC content and geophysical well-log variables (Khoshnoodkia et al., 2011; Tan et al., 2013; Zhao et al., 2015; Verma et al., 2016; Yu et al., 2017; Bai and Tan, 2020). TOC estimation using data-driven approaches is complex as these are mostly based on nonlinear relationships that involve multiple parameters. Consequently, simple, or multivariate regression methods are often preferred for organic content assessment in shales.

Bayes' theorem (Bayes, 1763) and Bayesian data analysis are widely known and uncomplicated. In recent decades, the Bayesian approach has gained popularity in the geophysical application for exploration and reservoir studies (e.g., Buland and Omre, 2003; Larsen et al., 2006; Rimstad et al., 2012; Sen and Biswas, 2017; Grana, 2020). Doyen (1988) and Bortoli et al. (1993) apply geostatistical methods, for the first time, to infer petrophysical properties from seismic data. A brief synopsis of uncertainty quantification, stochastic behavior, and Bayesian inverse theory for geophysical applications under various statistical assumptions is covered in Sen and Stoffa (1996), Scales and Tenorio (2001), Mosegaard and Tarantola (2002), and Tarantola (2005). Buland and Omre (2003) bring forward an analytical solution, based on the Bayesian approach, related to the inverse problem of linearized amplitude variation with offset. Eidsvik et al. (2004) cast the prediction of facies and pore-fluid distributions in the Bayesian framework with a statistical rock-physics model to generate probability maps for the identification of promising plays within the reservoir. Recently, Grana (2020) presents Bayesian petroelastic inversion by realizing the posterior distribution as a summation of contributions from all of the likelihood functions of plausible models based on multiple prior models.

Even though research efforts to establish a relationship between the TOC content and well-log variables have been on the rise, using Bayesian theory for TOC estimation has not been explored systematically, except by Qian et al. (2019) and Deng et al. (2020). However, these two studies propose a Bayesian-inference-based inversion scheme to estimate TOC along with other petrophysical properties (e.g., porosity, water saturation, and brittleness volume), and they are not directly related to demonstrating the relationship between shale TOC content and well-log variables. This caters a good opportunity to strengthen and endorse our efforts in using the Bayesian approach to predict the TOC content, and also assess the associated uncertainty in estimation for prospective shale plays. In this study, we outline and demonstrate the application of the Bayesian approach combined with the Hamiltonian Monte Carlo (HMC) inference method to estimate shale TOC contents incorporating prior knowledge gathered from various geophysical log variables, namely GR, sonic, bulk density, thorium (Th), and uranium (U) logs. We then apply the method to predict the TOC contents through two examples: (1) Devonian Duvernay Formation, Western Canada and (2) Silurian shale of the Ahnet Basin, which are used as field case studies. Further, to demonstrate the efficacy of the present approach, we compare the results from the Bayesian setting using HMC inference with those from the widely used conventional methods for TOC estimation and corroborate the results by laboratory-measured TOC contents.

THEORY

Bayes' theorem

Bayesian analysis is a rigorous way to get information about the probability of the model parameters (known as "prior"), which is essentially the blend of predictions about the unknown model parameters and information learned about the same from the data (Buland and Omre, 2003; Rimstad et al., 2012; McElreath, 2016). Simply put, it is a method to interpret evidence in the context of prior experience that might be related to the event. For instance, if the risk of shaking due to earthquakes is known to impact severely the loose unconsolidated sediments, Bayes' theorem permits the risk to a particular area of known unconsolidated sedimentation to be evaluated more precisely than merely assuming that the particular case is typical of the population as a whole. In general, the process acts in a nonlinear iterative fashion and can be summarized by three key steps, as given next.

1) Based on some data and expectations on how the data being generated, we develop models (mostly crude approximation) through the information about probability distributions.
2) Then, we use a Bayesian approach to augment data to the models for deriving logical consequences from blending the data and our speculations.
3) Finally, we criticize the developed models by examining whether the model makes logical predictions adopting various criteria including data, experience, and occasionally even comparing with other relevant models.

Mathematically, a Bayesian approach that combines priors and data can be expressed as

$$p(\mathbf{m}|\mathbf{d})p(\mathbf{d}) = p(\mathbf{d}|\mathbf{m})p(\mathbf{m}), \tag{1}$$

where p, $\mathbf{m}$, and $\mathbf{d}$ are probability, model variables, and measured data, respectively; $p(\mathbf{m}|\mathbf{d})$ is the posterior distribution of the model

$\mathbf{m}$ given the measured data $\mathbf{d}$; and $p(\mathbf{d}|\mathbf{m})$ is the conditional probability (read as the likelihood of observing the data for a given model) of the data $\mathbf{d}$ given the model variables as $\mathbf{m}$. It is vital to choose a probability distribution that best describes the measured data. It is to note that the information about $p(\mathbf{d})$, in many applications, does not need to be quantified because the Bayesian framework takes the advantage of the theorem that the posterior distribution is proportional to the likelihood times the prior, given $p(\mathbf{d})$ as a normalizing constant (equation 1).

The prior $p(\mathbf{m})$ should express what information about the model parameters can be gathered before seeing the data $\mathbf{d}$. Conceptually, prior defines a statistical model for prior information about the model $\mathbf{m}$ expressed statistically by the probability density function (PDF). In this case, we assume it to be any valid PDF that defines the a priori knowledge about TOC before any core data are investigated. We define the parameters of the prior model as "$\boldsymbol{\pi}$" that control the prior knowledge about the model, which are mostly fixed in Bayesian analysis. If these parameters $\boldsymbol{\pi}$ are not fixed, then we recast equation 1 as

$$p(\mathbf{m}|\mathbf{d},\boldsymbol{\pi})p(\mathbf{d}|\boldsymbol{\pi}) = p(\mathbf{d}|\mathbf{m},\boldsymbol{\pi})p(\mathbf{m}|\boldsymbol{\pi}). \tag{2}$$

The likelihood is defined by the plausibility of the data given the model $\mathbf{m}$. It comprises forward models that signify the physical relation between model and data including the uncertainty. In practice, likelihood takes the form of a normal or Gaussian distribution. In this case, the likelihood can be expressed as $p(\mathbf{D}|\mathbf{m})$ that represents a link between the TOC distribution and well-log variables (GR, sonic, etc.) $\mathbf{D}$ in the following form:

$$p(\mathbf{D}|\mathbf{m}) = \prod_t p(\mathbf{D_t}|\mathbf{m_t}). \tag{3}$$

The posterior is the outcome from the Bayesian analysis and echoes all that we learn about the problem. In general, it is the balance between likelihood and prior, represented by the probability distribution of model parameters within the model and not a single value. Often, the posterior is treated as the updated probability of an event before considering (new) data. Subsequently, posterior from one study can be used as the prior for a new analysis. In such a way, the best model can be obtained by choosing the model, which has the highest posterior PDF. The posterior distribution for the prior is represented by equation 2 in a parameter space $\omega_{\boldsymbol{\pi}}$ can be obtained by

$$p(\mathbf{m}|\mathbf{d}) = \int_{\omega_{\boldsymbol{\pi}}} p(\mathbf{m}|\mathbf{d},\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{d})d\boldsymbol{\pi}. \tag{4}$$

For more details about the Bayesian theorem, readers can refer to Sivia and Skilling (2006) and McElreath (2016).

Model inference using HMC

HMC is an extensive and successful Markov chain Monte Carlo (MCMC) method used to get a sequence of random samples that converge to being disseminated according to a target probability distribution for which direct sampling is strenuous. Instead of depending on fragile heuristics (guess-and-check-strategy), the HMC method is built upon a rich theoretical basis that suggests obtaining information about the geometry of the unmapped areas of the typical set (see, e.g., Mackay, 2003; Betancourt and Girolami, 2013; Sen and Biswas, 2017). Compared with the Metropolis-Hastings algorithm, the HMC method lessens the correlation among successive sampled states by offering transfers to distant states, which nurture a high probability mass due to approximate energy preserving properties of the simulated Hamiltonian dynamic (Duane et al., 1987; Neal and Radford, 2011; Sen and Stoffa, 2013). Subsequently, a state of convergence can be achieved where successive simulations will be similar to drawing samples from the posterior distribution of the model that we wish to estimate. This makes the algorithm to be inimitably suited for the well-behaved target distributions in the high-dimensional problems of applied interest.

HMC is a fixed-dimensional MCMC method that augments the target state space with an auxiliary momentum variable $\mathbf{P}$ and follows Hamiltonian dynamics to make proposals for the Metropolis algorithm in the following form:

$$p(\mathbf{m_i},\mathbf{P}) = p(\mathbf{m_i}|\mathbf{P})p(\mathbf{m_i}). \tag{5}$$

Because the HMC method has been discussed in detail by numerous authors (Mackay, 2003; Neal and Radford, 2011; Betancourt and Girolami, 2013; Sen and Biswas, 2017), we prefer not to repeat it here. Nevertheless, brief detail on the formulation of the method is provided in Appendix A. Because there is no analytical solution for Hamilton's equation, the Hamiltonian dynamics are generally approximated in a discrete-time "t." This discretization is achieved by the leapfrog algorithm, which adapts discrete-time steps (e.g., "Δt") conserving the two most vital properties of the Hamiltonian dynamics, "reversibility" and "volume preservation." The leapfrog integration begins by drawing a fresh momentum term self-reliantly on the previous momentum value or parameter values $\mathbf{m_i}$, and can be expressed as follows:

$$\begin{aligned}
\mathbf{P}\left(t+\frac{\Delta t}{2}\right) &= \mathbf{P}(t) - \frac{\Delta t}{2}\frac{\partial f(\mathbf{m_i})}{\partial \mathbf{m_i}}(\mathbf{m_i}(t)),\\
\mathbf{m_i}(t+\Delta t) &= \mathbf{m_i}(t) + \mathbf{P}\Delta t\left(t+\frac{\Delta t}{2}\right),\\
\mathbf{P}(t+\Delta t) &= \mathbf{P}\left(t+\frac{\Delta t}{2}\right) - \frac{\Delta t}{2}\frac{\partial f(\mathbf{m_i})}{\partial \mathbf{m_i}}(\mathbf{m_i}(t+\Delta t)).
\end{aligned} \tag{6}$$

In the leapfrog method, for one integration step, we start by simulating the momentum $\mathbf{P}$ for the current state using a Gaussian distribution $(0,\sum)$ by $(\Delta t)/2$ time units. Then, we perform a full-step (Δt) simulation for $\mathbf{m_i}$ using the updated values of momentum $\mathbf{P}$ and position variables sequentially. Subsequently, we perform the simulation for momentum dynamics for the remaining half-step $(\Delta t)/2$ with the intent that the momentum and model perturbations can attain the full-time steps. In such a way, the proposal state $(\mathbf{m_i^*},\mathbf{P^*})$ can be reached from an initial state $(\mathbf{m_i},\mathbf{P})$ via L steps (total number of integration steps) of step size Δt with a total of "$L\Delta t$" time. Note that the leapfrog method could not conserve $H(\mathbf{m_i},\mathbf{P})$ precisely due to the integration error of the order of $(\Delta t)^3$ per step and $(\Delta t)^2$ globally. Hence, to account for such numerical errors, a Metropolis correction (acceptance) step becomes necessary to ensure proper sampling. The probability of accepting the pro-

posal $(\mathbf{m}_\mathbf{i}^*, \mathbf{P}^*)$ as the next state of the Markov chain from the transition state $(\mathbf{m_i}, \mathbf{P})$ is

$$p_\theta = \min\{1, \exp(H(\mathbf{m_i}, \mathbf{P}) - H(\mathbf{m}_\mathbf{i}^*, \mathbf{P}^*))\}. \quad (7)$$

If the proposal is not accepted, then the previous parameter value is returned for the subsequent draw and used to initialize the next iteration. It is noteworthy that the choice of L and Δt can greatly contribute to obtaining the optimized acceptance rate; hence, they need to be tuned properly (not too high nor too low). To tune these parameters, the no-U-turn sampler (NUTS) plays a significant role, which is discussed in Appendix B for the sake of completeness. In general, NUTS is an extension by regulating L value in each iteration automatically to avoid the requirement of knowledge of L and random-walk behavior. This makes it efficient to perform sampling of the posteriors in a much faster way. The efficient employment of NUTS depends on the acceptance probability. This class of samplers can spontaneously choose a step size that attains an acceptance probability of 0.6, which is optimum.

The Bayesian model

The aim of the present study is to predict the TOC profile together with the uncertainty in predictions of model variables, and the probability distribution of the predicted TOC profile. To solve a geophysical inverse problem dealing with measured data $\mathbf{d}$, model parameters $\mathbf{m}$, and error ϵ, we suppose that the set of physical relations $\mathbf{G}$ (i.e., the operator that generates the data for known parameters) is already identified (Buland and Omre, 2003; Grana, 2020). The forward problem can be expressed as

$$\mathbf{d} = \mathbf{G}(\mathbf{m}) + \epsilon. \quad (8)$$

Because the interest of the present study lies in uncertainty quantification, we emphasize probabilistic methods, namely on Bayesian regression method, intending to compute the posterior distribution of the model parameters given the observations (see equation 1). To be specific, our interest is to estimate the TOC values against depth using the prior(s) based on the knowledge obtained from available geophysical logs. Hence, in a probabilistic framework, the model is a multivariate regression, and the predicted TOC with an expected value φ can be expressed as

$$d_{\mathrm{TOC}} \sim B(\varphi = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5, \sigma = \epsilon), \quad (9)$$

where β_0 is the intercept, $\beta_{i=1,2,..,5}$ are the coefficients for the covariate x_i, and $\sigma(=\epsilon)$ denotes the observation error. Note that d_{TOC} observed as the Gaussian distribution with a mean of φ and standard deviation of σ. To develop the Bayesian framework, we must assign a prior distribution to the unknown model parameters. For a Bayesian regression model such as equation 18, a reasonable generic preference for regression parameters would be a normal distribution because these variables can be positive and negative. Therefore, we can write

$$\begin{aligned} &\beta_0 \sim B(\varphi_{\beta_0}, \sigma_{\beta_0}), \\ &\beta_{i=1,2,..,5} \sim B(\varphi_{\beta_i}, \sigma_{\beta_i}), \\ &\sigma \sim |B(0, \sigma_\epsilon)|. \end{aligned} \quad (10)$$

For the sake of completeness, we provide the Kruschke diagrams for easy representation and interpretation of the deterministic variables (e.g., φ, as represented by "=") and stochastic variables such as β_0, $\beta_{i=1,2,..,5}$, and ϵ, as denoted by "~" (Figure 1).

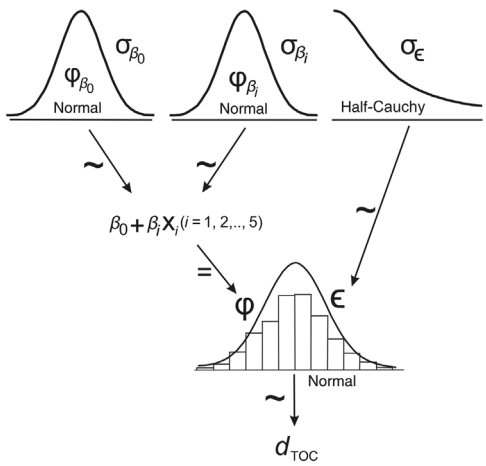

Figure 1. Schematic diagram representing the probabilistic formulation of the Bayesian regression model. Here, arrows marked by ~ signify stochastic dependency, and those marked by = represent deterministic dependency.

APPLICATIONS

Example 1: Experiment on the Devonian Duvernay Formation, Western Canada

To demonstrate the applicability of our approach in the first example, we choose the wireline log data together with core-measured TOC values from well A, representing an organic shale from the Devonian Duvernay Formation, Western Canada. The data are from a well-known study reported by Passey et al. (1990). The well-log data comprise GR, resistivity, compensated acoustic log (DT), compensated density (RHOB), and compensated neutron log (CNL), which are the inputs to the Bayesian model with priors. The target zone is 2249–2435 m, representing the reservoir intervals of the studied formation. Here, we define the Bayesian model by the stochastic random variables with normal prior PDFs for the regression coefficients with a standard deviation of 10, and for the standard deviation of the observed data as a half-normal distribution. Further, we consider the sampling distributions of the outcomes in the core-measured TOC, representing the data likelihood to be normal distribution. Note that the parameters for the likelihood of the model are not static values, unlike for the priors, rather a combination of deterministic and stochastic objects. Next, we obtain the posterior approximates for the unknown parameters in the model by drawing samples using the HMC method with the NUTS. We sample two chains in parallel that allow 1000 tunes for each chain to meet its steady-state and then generate an additional 3000 samples from all parts of the posterior distribution (or

fully exploring the distribution), considering that the chain has converged after 1000 samples. To ensure that the independent chains are converged on the same space, although they are initiated from randomly selected places, we realize trace plots, which are illustrated in Figure 2. We notice that the curves within the trace plots are freely meandering around, indicating a good mixing. The left column of the trace plot characterizes the marginal posteriors of each stochastic parameter required to build the regression model, as a smoothed histogram, whereas the right column represents the samples of the HMC chains plotted in sequential order (Figure 2). Table 1 summarizes the corresponding descriptive statistics from the analysis. To ascertain if the chains for HMC samples are converged, we analyze the trace plots in conjunction with the evaluation of the Gelman-Rubin convergence criteria. In general, it is a way of checking whether the Markov chains converge on the same posterior, represented by a ratio of the concerted variance of values covering entire chains to the average variance of each chain. Institutively, the value of the Gelman-Rubin statistic (R_hat) should be one if all of the chains for the given parameter converge in the same space. If there is more variability across chains and within chains, then it suggests that they are frozen in their own various local spaces, and the variance of the values across chains will be higher than that of each chain. In this example, the R_hat value is equal to 1.0 for all of the variables used in the model, which suggests that there is no variability between the chains and within chains and that the chains are converged successfully.

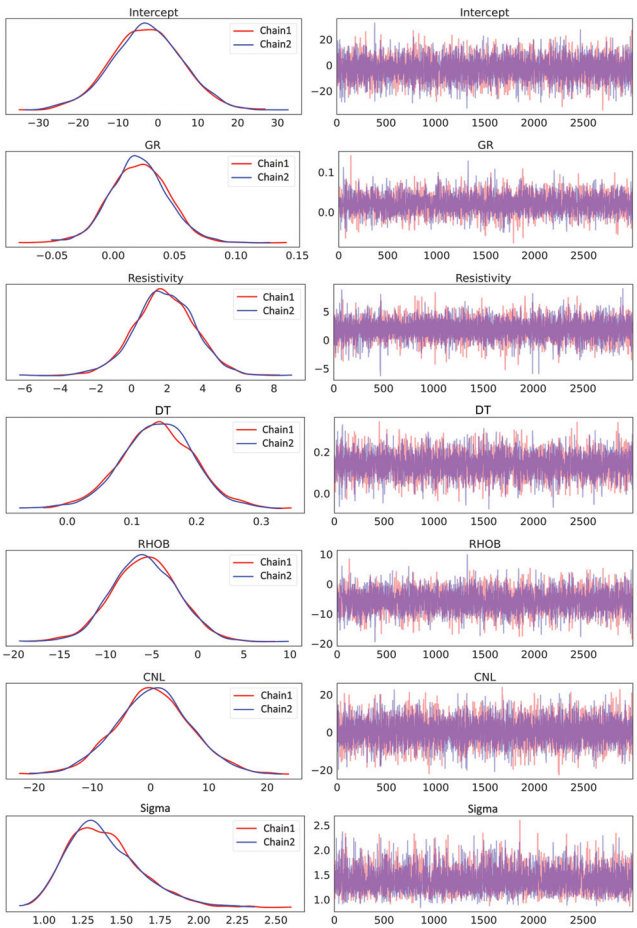

Figure 2. Trace plots of samples of HMC chains in sequence (right column) and marginal posteriors of each stochastic parameter (left column) for the multivariate regression model for well A.

To demonstrate the effectiveness of the proposed approach for TOC prediction, we compare the regression results with those obtained from the widely used conventional techniques such as the GR-based approach and the modified Schmoker and Hester (1983) model (henceforth, modified SH model). In comparison, the Bayesian model predicted TOC contents are better than those of GR-based and modified SH models, as can be seen from the crossplots in Figure 3. By contrast, the correlation coefficient (i.e., R^2, a numerical measure of a linear relationship between variables) of the Bayesian predicted model is highest, approximately 0.802, when compared with that of both of the conventional methods for TOC estimation, as listed in Table 2. Note that in the case of perfect prediction in which predicted TOC values perfectly match the core-measured TOC values, R^2 is equal to 1.0. The mean absolute error (MAE) and root-mean-square (rms) error for the Bayesian model are approximately 0.937 wt% and 1.187 wt%, respectively, which are significantly lower than those values of both of the conventional methods (Table 2). We thus find that the Bayesian model is more flexible as it shows the highest accuracy among all three methods, which is evident from the example case.

Example 2: A case study from the Silurian shale, Ahnet Basin

In example 2, we focus on validating the present approach for TOC estimation on real data collected from a well located in the Ahnet Basin, targeting the Silurian shale. This north–south-trending basin is situated in the west-central part of the southern Algerian Sahara and hosts a nearly 3000 m thick Paleozoic sequence of Cambrian to Carboniferous age (Logan and Duddy, 1998). A major regional flooding event during the Silurian deposited a thick transgressive marine shale comprising sapropelic and mixed (types I and II) kerogen (Makhous et al., 1997). The Silurian shale in the Ahnet Basin is characterized by 20%–50% clay and 15%–51% quartz with 11%–39% pyrite and minor carbonate associations (Figure 4). The X-ray diffraction data indicate that illite and chlorites are the dom-

Table 1. A statistical summary of the posterior estimates of the Bayesian model parameters for the studied well, well A.

Coefficients	Posterior mean	HPD_3%	HPD_97%	ESS	R_hat
Intercept	−2.768	−19.551	13.890	3427	1.0
GR	0.021	−0.023	0.065	3006	1.0
Resistivity	1.856	−1.309	5.158	4755	1.0
DT	0.139	0.028	0.243	3310	1.0
RHOB	−5.727	−12.152	1.101	3222	1.0
CNL	0.456	−12.524	12.459	4767	1.0
Sigma	1.390	0.989	1.846	3573	1.0

Note: ESS, effective sample size; HDI, highest posterior density; R_hat, Gelman-Rubin statistic.

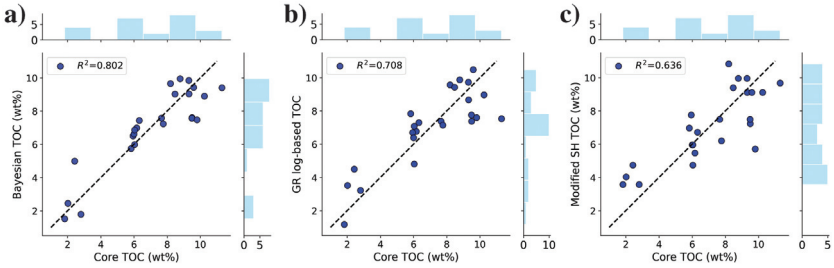

Figure 3. Comparison of the core-derived TOC contents with predicted TOC contents based on (a) the Bayesian model using HMC inference, (b) GR-based model, and (c) modified SH model for the well A example. The correlation coefficients of laboratory-measured TOC contents and the Bayesian predicted, GR log precited, and modified SH predicted TOC contents are approximately 0.802, 0.708, and 0.636, respectively.

Table 2. Comparison of the Bayesian predicted and other conventional methods predicted TOC contents for the well A, example 1.

Methods	Prediction accuracy indicators		
	rms error (wt%)	MAE (wt%)	Correlation coefficient (R^2)
Bayesian model using HMC inference	1.185	0.938	0.802
GR-based model	1.441	1.210	0.708
Modified SH model	1.610	1.301	0.636

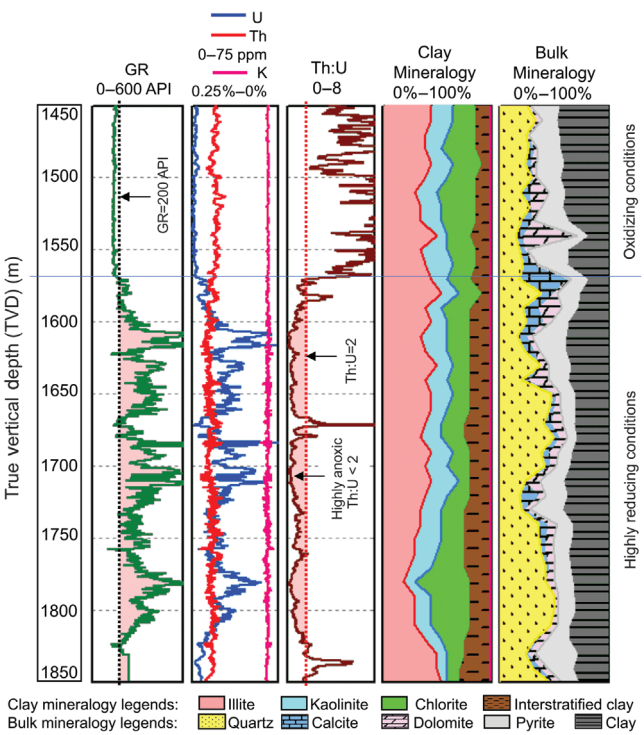

Figure 4. Characterization of the Silurian shale properties from the Ahnet Basin from wireline logs and mineralogy data. The lower part of the shale is exhibiting a more anoxic environment.

inant clay minerals in this formation. A geophysical log suite of this well comprises natural GR, RHOB, compressional sonic (DTCO), and natural GR spectrometry logs including uranium and thorium. The shale section can be distinctly divided into two sections based on the wireline log responses. The upper shale has a consistent approximately 150 API GR value with 2–5 parts per million (ppm) U and 15–21 ppm Th (Figure 4). A Th:U ratio of >2 is indicative of the oxidizing environment in the upper shale. In contrast, the lower Silurian shale reveals very high GR (>200 API), U ≥ 15 ppm, and Th < 15 ppm (Figure 4). The entire lower shale exhibits Th/U < 2, which corresponds to a highly anoxic condition (Carvalho et al., 2011).

TOC estimation based on the Bayesian approach using HMC inference

We corroborate the findings with the TOC contents derived from a total of 25 core plug samples, as obtained from drilling through the target Silurian shale formation. First, we perform the necessary data preprocessing before using the well-logs. For example, we ensure the density log readings by checking the caliper readings to identify the washout sections, if any, which can potentially underestimate the formation bulk-density properties. Figure 5 illustrates the crossplots between various log variables and target TOC content, together with the histograms of each variable along the diagonal. Overall, we observe a linear relationship in the distributions between each variable; therefore, the application of the standard Bayesian approach to the data is feasible. Moreover, with careful observation of the univariate histograms, we realize that the trend of TOC versus GR and TOC versus U log distributions is quite alike, yet these distributions exhibit a large variability (Figure 5).

We define the model that we want to estimate for this case example by considering the formula given as follows:

$$d_{\mathrm{TOC}} \sim B(\varphi = \beta_0 + \beta_1 \mathrm{GR} + \beta_2 \mathrm{DTCO} + \beta_3 \mathrm{RHOB} + \beta_4 \mathrm{U} + \beta_5 \mathrm{Th}, \sigma = 10), \quad (11)$$

Here, we assume that the prior distributions for the regression coefficients dealing with the geophysical log inputs are Gaussian (i.e., normal) and the prior number of modes is finite, given the restricted range of magnitudes of the well-log parameters against depth. Also, we assume that the prior distribution has a zero mean and a half-normal distribution for the constant standard deviation value of 10. Similar to the case of example 1, we repeat the same workflow to draw samples from the posterior by realizing multidimensional stochastic processes within the HMC-based MCMC algorithm, meaning that when simulated will converge. We allow the algorithm, in which NUTS is the assigned step method, to draw samples from the posterior to resemble the posterior for each model parameter. Consequently, the posterior distribution will have one dimension for each model parameter and the distribution of samples can be used to infer a series of possible estimates or point estimates by considering the mean of all samples. To set the tunable parameters of the HMC algorithm, NUTS adapts several self-tuning strat-

egies. NUTS necessitates setting a scaling matrix parameter to a judicious value to make the sampling competent. If the selection of scaling parameters is poor then NUTS slows down considerably, occasionally almost stopping it entirely. We run NUTS by setting the parameters $\delta = 0.6$ and $\gamma = 0.05$ with a target acceptance of 0.9. For the initial guess, the step size is 0.25. The step-size adaptation parameter for dual averaging to meet time reversibility is set to 0.75. We initiate with eight maximum tree depths throughout the tuning phase of sampling and end up with the trajectories when the maximum tree depth reaches a value of 10. Such proper parameterization sets NUTS to reasonable values based on the variance attained during the tuning process. In this experiment, we sample two chains parallel that consider approximately 1000 steps for each chain to attain a state of convergence, and subsequently sample for another 2000 steps to produce samples from the posterior distribution of the model we wish to estimate. The extra draw that the sampling algorithm considers during the tuning phase is to allow the Markov chain to sample from a sensible model space in the distribution. In practice, these initial draws are rejected and not included in the sample of the final model space. The warm-up phase for NUTS is 1000 iterations. The models sampled after the warm-up can be used to characterize the target posterior distribution. After the warm-up phase, in total, we perform approximately 6000 successive simulations to obtain the marginal posterior distributions. The main results comprise the predicted TOC model along with the uncertainty in the prediction.

To deduce if the sampler made suitable HMC sample draws from the posterior distributions, we anticipate seeing a series of chains running over the parameter space in the trace plots. A review of the trace plots, as graphically depicted in Figure 5, indicates that the marginal posterior distributions of each stochastic variable are unimodal and bilaterally symmetrical. Note that the model result is not a unique single value but rather a distribution for the parameters. Trace plots are extremely informative and exemplify stationarity (i.e., no odd patterns) and excellent convergence (Figure 6). This implies that the samples hold on within the posterior distribution across iterations and chains, and they are not strolling outside the posterior distribution. If the chains are strikingly different, it would imply poor convergence. The convergence of chains helps us understand whether we have samples from the posterior distribution. Equally important is to ensure that we have enough required samples that are accurate and stable (Martin, 2018). Ideally speaking, if we draw 6000 samples then we should have 6000 unique information about the posterior distribution. However, regrettably, we get less information than that because the draws are correlated, which is defined by autocorrelation. It is important to desire autocorrelation to be small as that suggests we require shorter chains for a proper representation of the posterior distribution. Effective sample size (ESS) for each parameter is defined as the number of independent samples that we have from the posterior after autocorrelations in the chains are accounted for. The ESS obtained during the case study analysis signifies a reasonable value (mostly greater than 1700 samples) for each model parameter to result in stable estimates. This also can be verified by analyzing the mean estimate of each parameter. We conclude that the HMC algorithm has done a reasonable job of estimating the required parameters for the regression model (Table 3). In addition, we can infer the relationship of the target TOC contents with the various parameters of interest. For example, the physical properties such as bulk density, U, and Th contents are negatively correlated with the TOC contents in organic shale in this experiment. Even though the mean characterizes the most probable value of a model parameter, it is beneficial to review the uncertainty associated with it so as not to set excess confidence in a noisy estimate. Consequently, it is a general practice to calculate the highest posterior density (HPD) and credible interval (CDI) estimate of the posterior distribution. HPD is the shortest interval that contains a given portion of the posterior probability density. Table 3 (third and fourth columns) represents HPD intervals of parameters of interest for this case example while outlining the summary statistics. Note that the HPD and confidence intervals in frequentist statistics are two different entities, and HPD is represented by a value of 94%. For example, according to our model, the HPD of the intercept signifies that there is a 94% probability that the parameter "intercept" for the estimation of TOC content lies between 0.087 and 31.207 (Figure 7). Similar inferences can be drawn for other parameters of interest for this Bayesian framework. In ad-

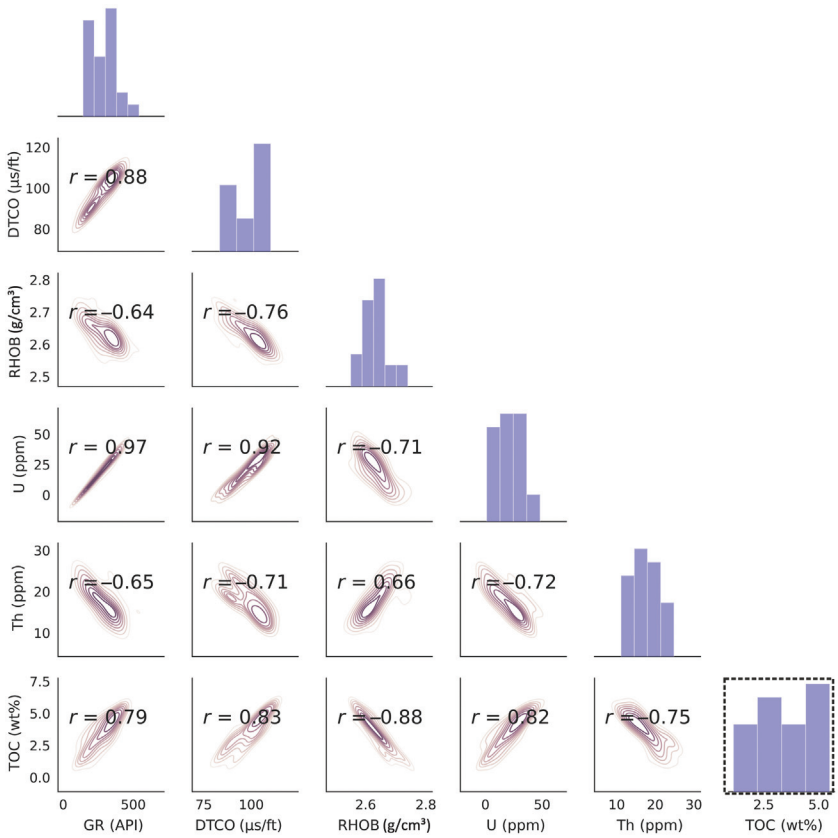

Figure 5. Crossplots among five well-log variables, i.e., GR, DTCO, RHOB, U and Th contents, and target TOC content measured from core samples in the laboratory. The dashed black square displays the distribution of samples in TOC content.

dition to the visual inspection of trace plots, we analyze the Gelman-Rubin convergence criterion. It offers a valuable and quantitative guide for convergence quality. Table 3 indicates that all of the stochastic parameters have R_hat values equal to 1.0. Therefore, the distributions in the two chains are indistinguishable and the Markov chains are said to have converged for the marginal parameter distributions.

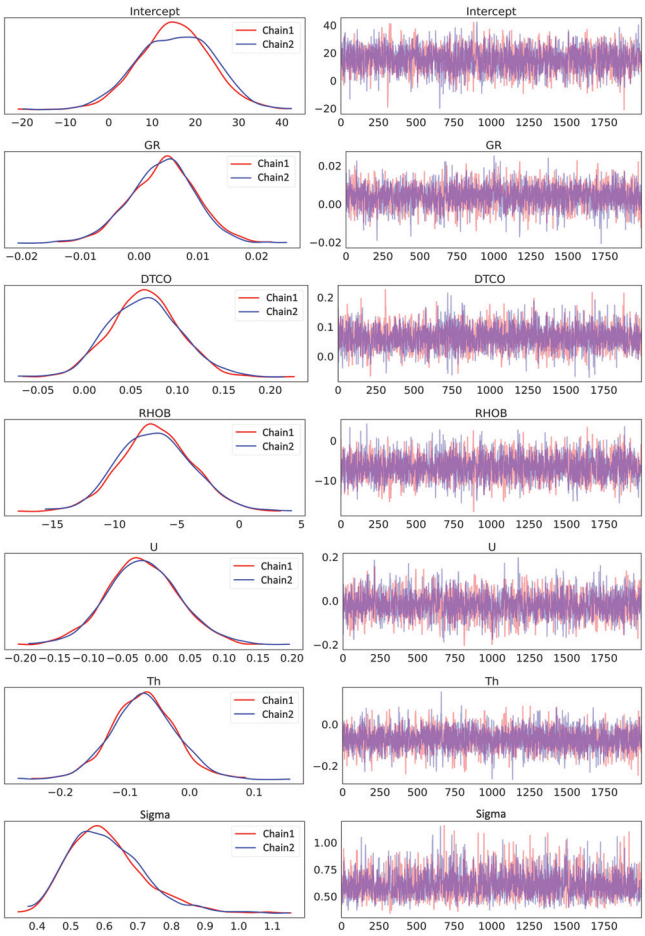

Figure 6. Representation of marginal posteriors of each stochastic parameter (left column) and trace plots of individual samples of HMC chains in sequence (right column) for the multivariate Bayesian model for the case study (example 2) data.

Table 3. A statistical summary of the posterior estimates of the Bayesian model parameters for the studied well, well A.

Coefficients	Posterior mean	HPD_3%	HPD_97%	ESS	R_hat
Intercept	15.289	0.087	31.207	1980	1.0
GR	0.004	−0.007	0.015	1972	1.0
DTCO	0.065	−0.005	0.136	2436	1.0
RHOB	−6.699	−12.118	−1.127	2127	1.0
U	−0.019	−0.125	0.084	1765	1.0
Th	−0.072	−0.167	0.021	2469	1.0
Sigma	0.607	0.411	0.814	1674	1.0

Note: ESS, effective sample size; HDI, highest posterior density; R_hat, Gelman-Rubin statistic.

To evaluate the model variable effects on the TOC estimation, we explore the model space by varying one parameter at a time while keeping the others constant. This mechanism assesses the model across the range of parameter values along with pooled values for the numerous samples. We can have a different set of parameters from the trace as it draws every single time. By this, we can analyze the effect of a single model variable together with the associated uncertainty in the model estimation. We believe that each nonquery variable is at the median value. Figure 8 demonstrates the effect of changing model variables one at a time on the posterior predictive of the TOC contents. We notice that the uncertainty associated with the variables GR, DTCO, and U contents during the model estimation appears to be higher (extreme in sideways, but less so toward the mean) when compared with the other two model variables such as RHOB and Th contents. This implies that, among many physical properties, information about RHOB and Th contents is vital to the TOC content estimation because the predictive posterior distributions are more uniform for such variables and follow a certain trend. Hence, the uncertainty in the model estimation should be less while considering RHOB and Th contents among other variables. These observations are specific to our model concerning this field study. Nevertheless, we do not rule out the possibility of discerning the influence of other model variables on TOC estimation for some other field data.

To locate unseen parameter values that lie within a specific subjective probability, it is common practice to summarize the predictions in an interval estimate akin to a confidence interval in the frequentist approach. These intervals are known as CDIs because the approximations provide a certain amount of probability (credibility) of the parameter along with the upper and lower bounds. Any parameter value within the intervals has the highest probability density than any other point outside. For example, a 95% CDI suggests two limits for the middle area in which 95% of the posterior distribution will fall. The two shaded portions in Figure 9 illustrate the 50% CDI (cyan shaded area) and 95% CDI (sky-blue shaded area) for all of the variables for the model estimates. These CDIs offer a more intuitive and valuable decision in interpretation, espe-

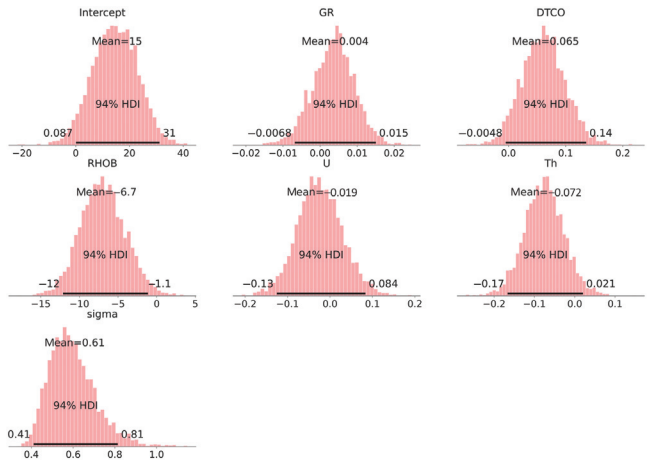

Figure 7. Histograms showing the posterior distribution (HPD) of the model parameters of interest for the case study (example 2) data.

cially dealing with uncertainties. We have seen that the relationship between predicted TOC contents and each variable is different, which is obvious because every parameter signifies different physical properties and may not directly influence the TOC content in organic shale. Nevertheless, the observed TOC contents fall within the predicted ranges including the 50% and 95% CDI limits, except for a few points for some variables such as GR, DTCO, and Th contents (Figures 9a, 9b, and 9e). The uncertainty in the prediction is large when a few variables crossed a certain range of values. For example, after a certain GR value (i.e., 380 API), the uncertainty in estimation rises as evidenced by the fact that the width of 95% CDI has increased. As a whole, it appears that the model is a useful representation of the data. To make an informed assessment of the TOC contents, it is critical to comprehend not only how the average TOC value is related to a certain depth or parameter, but how much the parameter or specific depth is expected to vary around that average. Prediction intervals are beneficial in such cases, and the Bayesian framework ensures that it is straightforward to acquire these intervals.

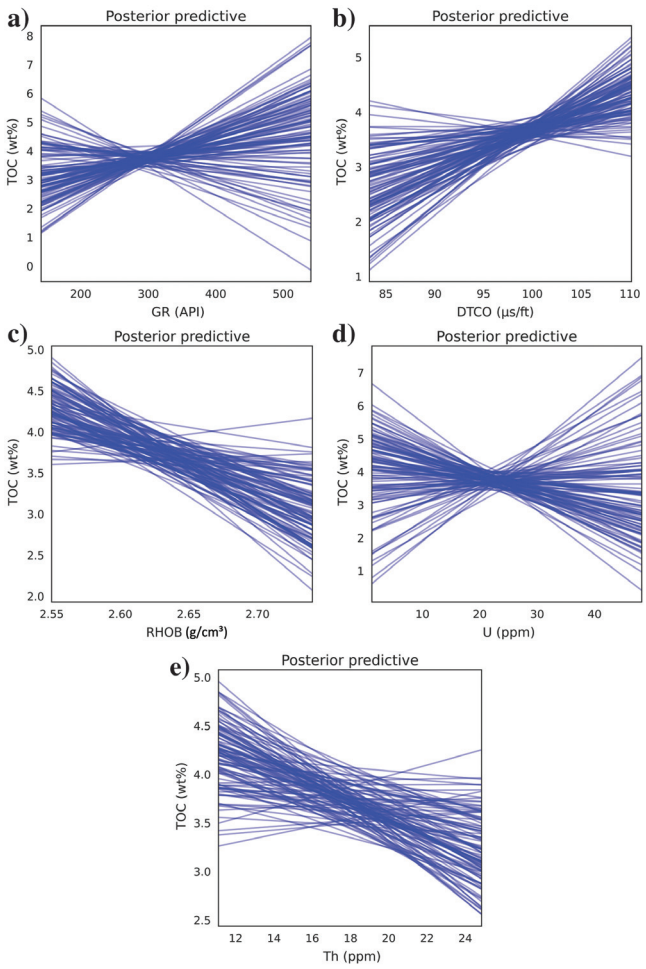

Figure 8. Represents posterior predictive model variable effect and associated uncertainty in the model estimates. The variables used are (a) GR (API), (b) DTCO (µs/ft), (c) RHOB (g/cm^3), (d) U contents (ppm), and (e) Th contents (ppm), respectively.

Model validation

Experiencing accepted convergence and determining that we now have a judicious number of samples, we can take measures to use the simulation results to make inferences from the model. This is an additional useful way to ensure the convergence of the HMC method and validate the model. The procedure is known as a posterior predictive check (PPC) in which data are generated from the model using parameters drawn from the posterior distribution and then compared with the posterior predictions of the model. In such a way, PPC associates two sources of uncertainty; one is the parameters uncertainty apprehended by the posterior and another is sampling uncertainty captured by likelihood. PPC supports considering a different model for a better explanation of data if reasonable discrepancies are observed between the model and the predicted data. Nonetheless, it is quite difficult to comment on how many alternative models one can try. The model drawn should be tractable and meaningful to describe the data. We draw a total of 800 random samples of parameters from the observed model. Next, it will draw 4200 random numbers from a normal distribution for each sample. Figure 10

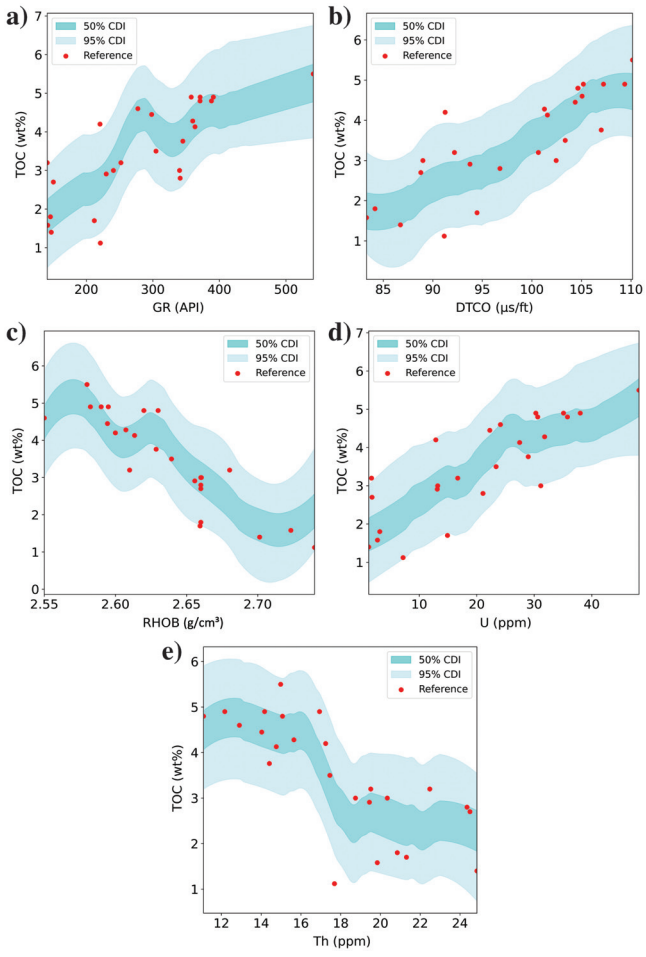

Figure 9. Scatter plots showing posterior predicted TOC contents against each variable: (a) GR (API), (b) DTCO (µs/ft), (c) RHOB (g/cm^3), (d) U contents (ppm), and (e) Th contents (ppm), respectively. The sky-blue shaded zone is 95% CDI, and the cyan shaded zone is 50% CDI. The red dots define the reference TOC contents.

depicts the posterior predictive mean plot of the measured data together with the calculated mean from each one of the 800 posterior predictive samples. It is easy to see that the inferred mean ranges simulated from the model with a choice of parameter values appear to have well described the core-derived mean TOC value of 3.48 wt%. This suggests that the proposed model aptly represents the observed data and that the sampler is converged properly.

To further validate the model with confidence, we compare the model with typically another one or more standard models that vary from the main model in some ways. Even though the convergence criteria are fulfilled, it is always better to compare (probably through cross validation) the main model with 2C or 3C mixture models. For any model, the aim is to ensure how well a model represents the true distribution over the model parameters and measured data. Nonetheless, it is difficult to know this, so surrogates are used to assess the average fit of a few data considering the model trained on the remaining data or approximate the whole data duly corrected for the model's pliability from its parameters. We use the former information criteria to assess the developed model's capability of capturing the true TOC content distribution obtained from the real field case. We notice that the TOC contents from the simulated model have a good agreement with the laboratory-measured values (Figure 11).

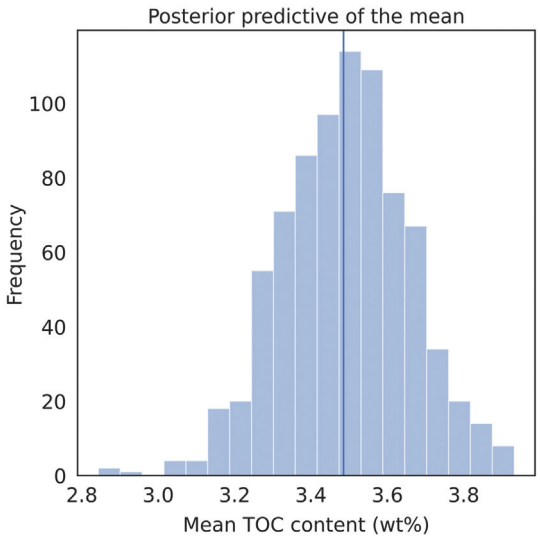

Figure 10. Comparison of the posterior predictive mean with that of core-derived TOC content (represented by a light blue line) to validate the model.

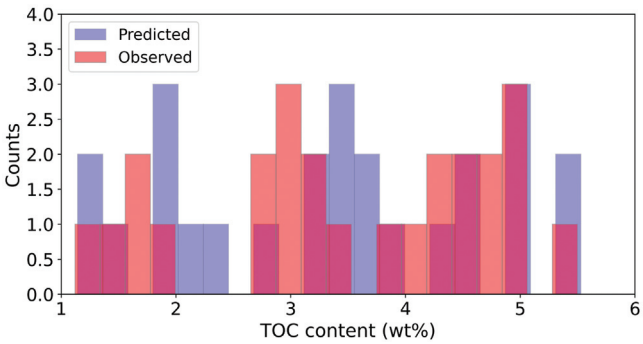

Figure 11. Representation of the posterior predictive samples chosen randomly to simulate data from the model together with the measured TOC data at a certain depth.

In addition, we test the model by predicting a new data point and then use it to corroborate the mean of the measured TOC content. For this, we choose a normal distribution of estimated outputs by multiplying model parameters with data values, and the mean of each parameter from the trace is considered to aid as the best approximation of the parameter. Then, to predict a new data point, we substitute the value of the parameters and obtain the PDF for the TOC contents. As evidenced in Figure 12, the mean estimate (i.e., 1.6 wt%) lines up well with the core-measured TOC value (i.e., 1.51 wt%). However, a wide estimated interval prevails in this case, which led to slight uncertainty in estimation, which could have been avoided if we had enough core-derived TOC values. By and large, in comparison to the known TOC value of 1.51 wt%, the probabilistic model performs well, with a peak probability at 1.49 wt%.

Figure 13 depicts a quantitative assessment of the Bayesian approach inferred by the HMC method for the estimation of TOC contents. We notice that data are more clustered along the 45° line, meaning that the inferred TOC contents match reasonably well with the reference TOC values obtained from the laboratory. Ultimately, the model yields a correlation coefficient, R^2 value of 0.836 among the predicted results, and core experiment results. The MAE and rms error are approximately 0.380 wt% and 0.505 wt%, respectively.

A more intuitive way to interpret the results can be established by assessing the prediction uncertainty and variability of the predictions concerning the ensemble mean. As said previously, we use the randomly drawn 800 samples from the posterior distribution to construct the ensemble of the predicted models. In practice, the uncertainty will be relatively small if the ensemble mean matches the reference TOC contents sensibly and vice versa. Also, we perform sensitivity analysis on several models in the ensemble and conclude that at least 800 models are essential for the posterior distribution to attain stable statistics. In fact, the predicted standard

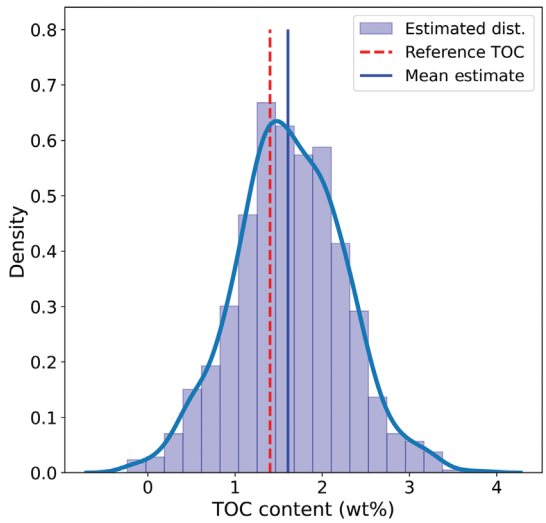

Figure 12. Posterior probability density with an overlain histogram for the estimated TOC content distribution while using the model to predict new data.

deviation converges for 800 ensembles of realizations; with less than 200 models, the ensemble subsides after a few iterations. Note that the computation performed assumes that ensembles follow a Gaussian distribution, and we are approximating the posterior distribution through ensemble realizations. Figure 14 reveals the average expected value of the TOC contents at specific depth intervals and the 50% and 95% CDIs for the expected value (shaded area). It is noticeable that the ensemble mean agrees well with the reference model, i.e., core-measured TOC contents. In addition, both the CDIs, i.e., 50% CDI (cyan shaded area) and 95% CDI (sky-blue shaded area), vary within a narrow range, indicating that the estimation uncertainty is not very large. Most of the data are within the 50% CDI intervals, except for a few depth points that fall within the 95% CDI. The TOC content is relatively less (approximately 1.4 wt %) within the depth range between 1480 and 1550 m, then proceeded to increase, leading to increased TOC content (as high as approximately 5.5 wt%) for deeper depth intervals (1550–1840 m). Figure 15 illustrates the ensemble mean together with 800 ensemble predictions and reference TOC values obtained from the core during the laboratory experiment. A visual inspection of the ensemble predicted results indicates that the posterior mean is consistent with the individual runs, and that is corroborated by the reference TOC contents (Figure 15a). The probability of TOC content within the first depth intervals (i.e., 1480–1680 m) appears to be more as compared with the deeper depth intervals ranging between 1690 and 1840 m.

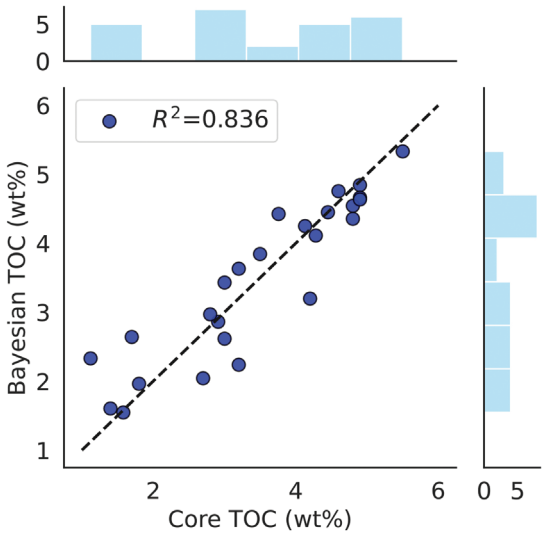

Figure 13. Crossplot showing the comparison of the core-derived TOC contents with the predicted TOC contents based on the Bayesian model using HMC inference.

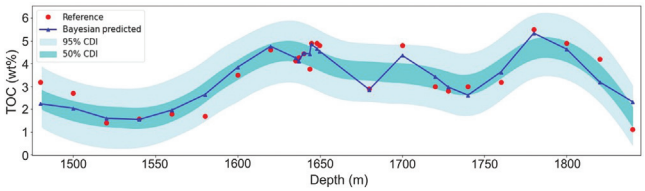

Figure 14. Results showing the ensemble mean (blue curve) with 50% (cyan shaded zone) and 95% CDI (sky-blue shaded zone) from 800 predictions using the HMC-based Bayesian approach. The red dots signify the reference TOC values as obtained from the laboratory experiment.

TOC estimation based on conventional methods

We follow the modified Schmoker and Hester (1983) method to calculate the TOC content from the density (RHOB) log. By plotting the reciprocal of bulk density (1/RHOB) as a function of the measured TOC, the following relationship can be established (Figure 16a)

$$\mathrm{TOC} = 170.52 * \left(\frac{1}{\mathrm{RHOB}}\right) - 61.22. \tag{12}$$

The response of the GR provides a positive correlation with the TOC values (Figure 16b), which can be expressed as

$$\mathrm{TOC} = 0.0094 * \mathrm{GR} + 0.8627. \tag{13}$$

Figure 16c shows the results from the conventional methods, in which we can see that the estimated TOC values reasonably agree with the laboratory-measured TOC values. However, the correlation coefficient (0.780) for the conventional approach is significantly smaller than that of the corresponding Bayesian model (0.836).

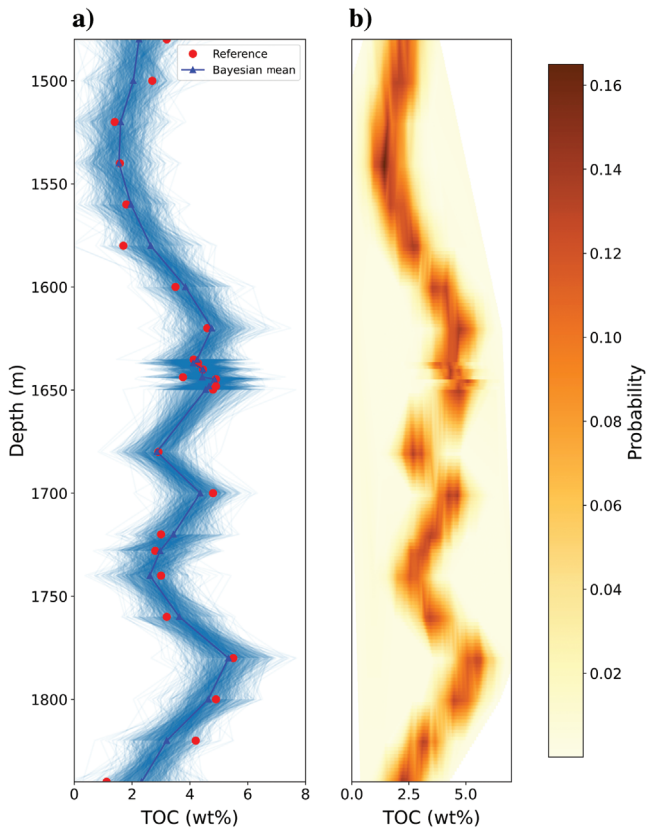

Figure 15. Results from ensemble predictions, from left to right: (a) the ensemble predictive samples with the ensemble mean and laboratory-measured TOC data and (b) the posterior probability at each depth interval.

Also, we observe that the anoxic lower Silurian shale has RHOB < 2.67 g/cm^3 and GR > 200 API, indicating this shale to be organically richer than the upper shale section.

DISCUSSION

The workflow for estimating TOC contents presented here makes use of a Bayesian setting inferred ensembles that account for the uncertainty of the model estimates. The present approach focuses on the HMC-based MCMC algorithm applied to approximate the posterior distributions for the unknown model variables, which are often problematic and might not be analytically tractable. In general, the implementation of the proposed approach for the TOC content estimation is not complicated and can be applied easily to estimate other parameters of interest as well. In case of limited available samples, as the present study reports, the Bayesian inference is a better method to develop models owing to its tendency to provide reasonable estimates with few data points using judicious prior(s). Several studies indicate unequivocally that the HMC approach in the Bayesian setting is very much dependent on prior information and measured data points; this study is no exception (Buland and Omre, 2003; Eidsvik et al., 2004; Grana, 2020; Feng et al., 2021). In the Bayesian setting, the results are interpreted intuitively by establishing a preliminary estimate and improving it further as we gathered more information about it.

Due to subsurface heterogeneities and limited data to capture those heterogeneities, numerous sources of uncertainty arise. It can be quantified by using probabilistic methods in which prior knowledge can be integrated with the information from the data to infer the posterior distribution. The uncertainty can be characterized in various forms including probability distributions, confidence intervals, and measures of variability. In general, the uncertainty inferred from posterior distributions exclusively reckons for the uncertainty in the measured data and the physical relation between data and model parameters (Sen and Stoffa, 1996; Scales and Tenorio, 2001; Mosegaard and Tarantola, 2002). Predicting uncertainty is critical to decision-making in any reservoir characterization including shale oil/gas reservoirs, and this should be based not only on the most likely model but also on the uncertainty estimates.

The present study assesses the prediction uncertainty by an ensemble of predicted models. We start the approach by exploring approximately 200 ensemble realizations from the posterior distribution and finally realize up to 800 ensembles of models because a limited number of ensembles are not very useful to assess the uncertainty. The results show that the estimated uncertainty is not too high as most of the measured data points fall within the 50% and 95% CDI, and the predicted mean follows the trend of the real data. Moreover, the present approach could quantify the prediction uncertainty of the TOC content at each depth. To give an instance, in our example 2, the estimated uncertainty based on the standard deviation for the depth range 1520–1680 m is narrow (the TOC content is predicted more assuredly) when compared with the depth range 1700–1840 m (Figures 14 and 15). We speculate that the variation in the prediction is due to a lack of representative data and can be improved by refining prior models or incorporating more core-derived TOC data obtained from the laboratory experiment (Feng et al., 2021). This suggests that the present approach for TOC estimation still has few limitations, and the uncertainty in the model prediction can be improved by incorporating sets of multiple prior models involving various facies profiles (Grana, 2020).

Conventional methods make use of one or two geophysical log variables to infer the TOC content in organic shale, and thus could not truly capture TOC characteristics in the complex shale reservoirs, more particularly in low TOC-dominated shale reservoirs. The application to both the example case studies demonstrates that the TOC content estimation in the Bayesian setting is far better than the conventional methods. The benefits of the present approach are twofold: predicted model with uncertainty estimates, and better accuracy in the TOC estimation.

Furthermore, note that the Devonian Duvernay Formation (Passey et al., 1990) contains 6%–10% TOC between 2249 and 2435 m intervals, and the Silurian shale of the Ahnet Basin (second example used in this study) is characterized by relatively low TOC values (<6%). The proposed Bayesian multivariate model provides satisfactory results in the studied shales with varying TOC ranges from the two continents. The studied Silurian Ahnet shale indicates a change of redox condition from bottom to top; the lower shale interval between 1570 and 1850 m appears to be deposited during the early transgressive phase in a highly reduced and anoxic environment, whereas the upper shale between 1450 and 1570 m might have deposited at a subsequent stage of the transgression in a relatively oxic environment. The change from anoxic to the oxic environment is reflected in the TOC content of the Silurian shale, as the anoxic conditions

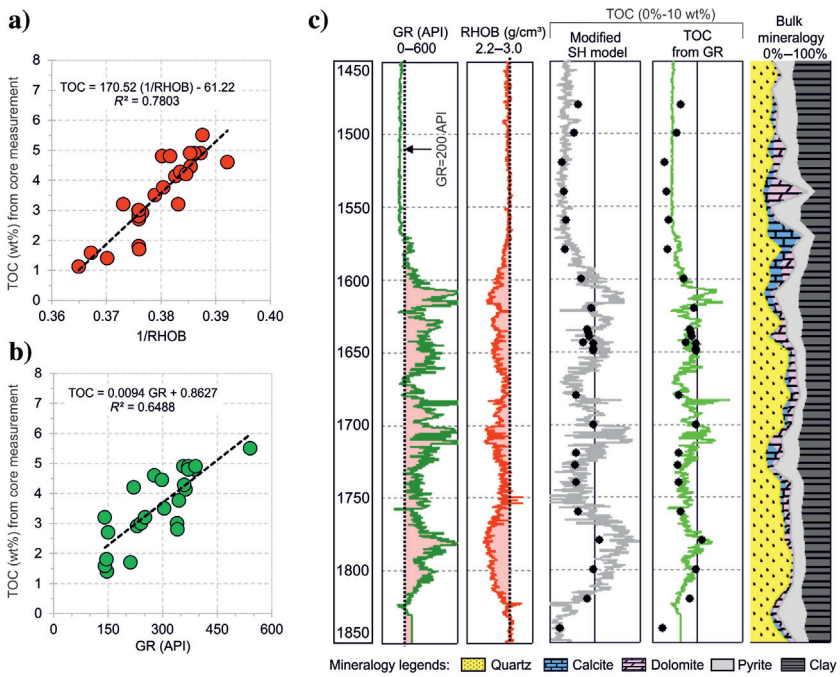

Figure 16. Relationship of (a) reciprocal of bulk density following modified SH model and (b) GR logs with measured TOC in the Silurian shale, and (c) estimation of density and GR-derived TOC values. Black dots indicate laboratory-measured TOC values using the cores.

favored higher organic matter preservation, yielding higher TOC in the lower Silurian shale. The Bayesian TOC estimation of the entire 400 m of the Silurian shale exhibits good results irrespective of the shift in the depositional environment within the same formation. However, it seems that the present approach predicts the TOC of the oxic upper Silurian shale more confidently than its lower counterpart, although sufficient data availability might be a probable reason for such variation in the TOC prediction.

CONCLUSION

Through this study, we have explored the possibility of the application of the Bayesian framework in predicting the TOC content in organic shales and quantifying the associated uncertainty in a regression process. Although the Bayesian approach is applied more than ever in many geophysical applications, the application of the HMC-based MCMC algorithm for TOC matter estimation has not been reported elsewhere. A suite of available geophysical wireline logs is used to demonstrate the applicability of the present approach and the results are then compared with those from the two most widely used conventional methods (GR-based and modified Schmoker methods). The two case examples of different geology demonstrate that the proposed approach is computationally efficient and accurately predicts the TOC contents together with model uncertainty. Specific conclusions drawn from this study are given as follows:

1) The posterior distributions of the unknown model parameters of interest are obtained by drawing samples using a multidimensional stochastic process based on the HMC method with the NUTS. We have sampled two chains parallelly to realize a total of 6000 samples that finally met the convergence criteria.
2) The Bayesian inference helps to quantity prediction uncertainty within the model by generating ensembles of models randomly drawn from the posterior. The predicted mean closely follows the core-measured TOC values as obtained from the laboratory experiment, and that lies within the CDIs.
3) The Bayesian setting performs well with reasonable statistics, duly corroborated by the laboratory-measured TOC values, as compared with the conventional methods for the TOC content estimation.
4) The proposed model is proven to be better suited to predict the high TOC contents with higher accuracy in the lower Silurian shale deposited mainly during the anoxic conditions (the early transgressive phase) irrespective of the shift in the depositional environment within the same formation.

ACKNOWLEDGMENTS

S. S. Ganguli is grateful to the director, CSIR-NGRI for all support and for providing consent to publish the results (reference no. NGRI/Lib/2022/Pub-16). This study is financially supported by the DST Inspire Faculty research grant (grant no. DST/INSPIRE/04/2016/000174). We would like to offer our deepest gratitude to J. Etgen, M. Pervukhina, Y. Liu, and the anonymous reviewers for their fruitful suggestions that significantly improved the manuscript.

DATA AND MATERIALS AVAILABILITY

Part of the data associated with this research are available and can be obtained by contacting the corresponding author.

APPENDIX A

GOVERNING EQUATIONS OF HAMILTONIAN DYNAMICS

In Hamiltonian dynamics, the total Hamiltonian energy, i.e., the summation of potential and kinetic energy, is estimated for the present state and an object's motion can be expressed concerning its location through model $\mathbf{m_i}$ and momentum $\mathbf{P}$ at a particular time (Mackay, 2003; Neal and Radford, 2011; Betancourt and Girolami, 2013). For a system following Hamiltonian dynamics, the Hamiltonian energy function can be expressed in terms of the object's potential energy, i.e., $V(\mathbf{m_i})$ and kinetic energy as $T(\mathbf{P})$, in the following manner:

$$H(\mathbf{m_i}, \mathbf{P}) = V(\mathbf{m_i}) + T(\mathbf{P}). \tag{A-1}$$

In practice, while assessing a random variable $\mathbf{m_i}$ with a PDF $f(\mathbf{m_i})$ in the HMC method, an auxiliary density is established that does not depend on the parameters $\mathbf{m_i}$, and follows a multivariate normal distribution: $f(\mathbf{P}) \sim \text{Multinormal}(0, \sum)$, where $\sum$is the covariance matrix. The joint density function of $f(\mathbf{m_i}, \mathbf{P})$ can be expressed as

$$\begin{aligned} f(\mathbf{m_i}, \mathbf{P}) &= \exp[\log f(\mathbf{m_i}) + \log f(\mathbf{P})] \\ &\propto \exp\left(\log f(\mathbf{m_i}) - \frac{1}{2}\mathbf{P}^T \sum^{-1} \mathbf{P}\right). \end{aligned} \tag{A-2}$$

For a given physical system, because of the decomposition of the joint density, we recast the Hamiltonian in equation A-2 as

$$f(\mathbf{m_i}, \mathbf{P}) = \exp[-U(\mathbf{m_i}) - K(\mathbf{P})] = \exp\{-H(\mathbf{m_i}, \mathbf{P})\}. \tag{A-3}$$

HMC draws samples from this joint distribution $(\mathbf{m_i}, \mathbf{P})$, which ultimately led to producing samples from the target distribution by selecting only $\mathbf{m_i}$ and the new momentum is evolved by following Hamilton's equations:

$$\frac{d\mathbf{m_i}}{dt} = +\frac{\partial H}{\partial \mathbf{P}} = +\frac{\partial T}{\partial \mathbf{P}}, \quad \frac{d\mathbf{P}}{dt} = -\frac{\partial H}{\partial \mathbf{m_i}} = -\frac{\partial T}{\partial \mathbf{m_i}} - \frac{\partial V}{\partial \mathbf{m_i}}. \tag{A-4}$$

As per the Hamiltonian dynamics, the samples are transferred while preserving the total energy of the system. Hence, the momentum density is independent of the target density, which means that $\partial T/\partial \mathbf{m_i} = 0$, yielding a new momentum evolved by following Hamilton's equations:

$$\frac{d\mathbf{m_i}}{dt} = +\frac{\partial H}{\partial \mathbf{P}}, \quad \frac{d\mathbf{P}}{dt} = -\frac{\partial V}{\partial \mathbf{m_i}}. \tag{A-5}$$

The vectors $\mathbf{m}$ and $\mathbf{P}$ can be combined into another vector form $\mathbf{Z} = (\mathbf{m}, \mathbf{P})$ in two dimensions and the Hamiltonian stated previously can take the following form:

$$\frac{d\mathbf{Z}}{dt} = J\nabla H(\mathbf{Z}), \tag{A-6}$$

where ∇H defines the gradient of H and $J = \begin{bmatrix} 0_{D\times D} & I_{D\times D} \\ -I_{D\times D} & 0_{D\times D} \end{bmatrix}$ is the 2D × 2D matrix involving identity and zero matrices, respectively.

APPENDIX B

NUTS IN HMC

In this section, we will outline how NUTS helps to tune the parameters for getting an optimized acceptance rate. NUTS directs the Hamiltonian dynamics forward and backward in time in a random manner until a U-turn criterion is met, that is,

$$\frac{\partial X}{\partial t} = \frac{\partial}{\partial t}\left(\frac{1}{2}(\mathbf{m}_\mathbf{i}^* - \mathbf{m}_\mathbf{i})^T(\mathbf{m}_\mathbf{i}^* - \mathbf{m}_\mathbf{i})\right) = (\mathbf{m}_\mathbf{i}^* - \mathbf{m}_\mathbf{i})^T\mathbf{P} < 0, \tag{B-1}$$

where X is half the squared distance between the current position $\mathbf{m}_\mathbf{i}^*$ and the initial position $\mathbf{m}_\mathbf{i}$ at each leapfrog step. If the preceding condition is satisfied, then a random point from the path is selected for the MCMC sample and the same process is continued from that new point. Nevertheless, this approach does not guarantee time-reversibility or convergence to the accurate distribution. NUTS overcomes this issue by adopting a recursive algorithm that preserves reversibility by introducing a slice variable "v" with conditional distribution:

$$p(v|\mathbf{m}_\mathbf{i}, \mathbf{P}) = \text{Uniform}\left(v; \left[0, \exp\left\{f(\mathbf{m}_\mathbf{i}) - \frac{1}{2}\mathbf{P}^T\sum^{-1}\mathbf{P}\right\}\right]\right). \tag{B-2}$$

Subsequently, we fix v and sample $\mathbf{m}_\mathbf{i}$ uniformly from the horizontal sliced region $\mathbb{R}$, as defined by $\mathbb{R} = \{\mathbf{m}_\mathbf{i} : v \le \varphi(\mathbf{m}_\mathbf{i})\}$, where $\varphi(\mathbf{m}_\mathbf{i})$ is the kernel of $f(\mathbf{m}_\mathbf{i})$. It is, however, challenging to find the bounds of $\mathbb{R}$, which can be solved by producing a finite set of all $(\mathbf{m}_\mathbf{i}, \mathbf{P})$ by repeatedly doubling its size (Neal and Radford, 2011) until the endpoints are outside $\mathbb{R}$. The doubling process yields by arbitrarily selecting forward and backward leapfrog steps to meet time reversibility. Ultimately, doubling stopped when for one of these subtrees, the states $\mathbf{m}_\mathbf{i}^-$, $\mathbf{P}^-$ and $\mathbf{m}_\mathbf{i}^+$, $\mathbf{P}^+$ linked with the leftmost and rightmost leaves of that subtree satisfies the following:

$$(\mathbf{m}_\mathbf{i}^+ - \mathbf{m}_\mathbf{i}^-)^T\mathbf{P}^- < 0 \quad \text{or} \quad \mathbf{m}_\mathbf{i}^- - (\mathbf{m}_\mathbf{i}^+)^T\mathbf{P}^+ < 0. \tag{B-3}$$

Following is the process of tuning Δt for the kth iteration of a Markov chain:

$$\log((\Delta t)_{k+1}) \leftarrow \alpha - \frac{\sqrt{k}}{\gamma}\frac{1}{k+k_0}\sum_{j=1}^{k}(\delta - p_\theta)_j,$$
$$\log(\bar{\Delta}t_{k+1}) \leftarrow \rho_k \log((\Delta t)_{k+1}) + (1-\rho_k)\log(\bar{\Delta}t_k),$$
$$(\Delta t)_{k+1} \leftarrow \bar{\Delta}t_{k+1}. \tag{B-4}$$

where p_θ and δ represent the actual acceptance probability and desired average acceptance probability, respectively; and $\gamma(>0)$ defines a free parameter that controls the volume of shrinkage toward α, which is a spontaneously chosen point wherein the iterated $(\Delta t)_k$ shrinks, respectively. The parameter k_0 also is a free parameter that dampens early exploration. It is recommended to set $\alpha = \log(10\Delta t_1)$ and $\delta \approx 0.6$ to save computation and reach the condition for $\alpha \to \delta$ (Hoffman and Gelman, 2014). In practice, a substitute statistic to Metropolis acceptance probability should be defined because NUTS selects $(\mathbf{m}_\mathbf{i}^*, \mathbf{P}^*)$ from several candidates and not only one. In such a case, the acceptance probability for each iteration can be written as

$$p_\theta = \frac{1}{|\mathcal{B}_k|}\sum_{\mathbf{m}_\mathbf{i},\mathbf{P}\in\mathcal{B}_k}\min[1, p(\mathbf{m}_\mathbf{i}^k, \mathbf{P}^k)/p(\mathbf{m}_\mathbf{i}^{k-1}, \mathbf{P}^{k,0})]. \tag{B-5}$$

Here, $\mathbf{m}_\mathbf{i}^k$ and $\mathbf{P}^k$ are the proposals, $\mathbf{m}_\mathbf{i}^{k-1}$ and $\mathbf{P}^{k,0}$ are initial values, and $\mathcal{B}_k$ represents the set of all states being explored by the algorithm during the final doubling process of the Markov chain. In each iteration, NUTS constructs a tree node based on subtrees ($k = 0, 1, 2, \ldots$) generated by a recursive process, such that a subtree is generated with 2^k nodes in the same iteration as the previous subtree, but in a random direction. For a detailed understanding of the NUTS including flowcharts, one can refer to Hoffman and Gelman (2014).

REFERENCES

Bai, Y., and M. Tan, 2020, Dynamic committee machine with fuzzy-c-means clustering for total organic carbon content prediction from wireline logs: Computers & Geosciences, **146**, 104626, doi: 10.1016/j.cageo.2020.104626.

Bakhtiar, H. A., A. Telmadarreie, M. Shayesteh, M. H. H. Fard, H. Talebi, and Z. Shirband, 2011, Estimating total organic carbon content and source rock evaluation, applying delta logR and neural network methods: Ahwaz and Marun Oilfields, SW of Iran: Petroleum Science and Technology, **29**, 1691–1704, doi: 10.1080/10916461003620495.

Bayes, T., 1763, An essay towards solving a problem in the doctrine of chances: Philosophical Transactions of the Royal Society of London, **53**, 370–418 (reprinted with biographical note by G. A. Barnard, in Biometrika, 45, 293–315), doi: 10.1098/rstl.1763.0053.

Betancourt, M., and M. Girolami, 2013, Hamiltonian Monte Carlo for hierarchical models: arXiv:1312.0906.

Bortoli, L. J., F. Alabert, A. Haas, and A. Journel, 1993, Constraining stochastic images to seismic data, *in* A. Soares, ed., Geostatistics TROIA'92, quantitative geology and geostatistics: Springer, 5, 325–337.

Buland, A., and H. Omre, 2003, Bayesian linearized AVO inversion: Geophysics, **68**, 185–198, doi: 10.1190/1.1543206.

Carpentier, B., A. Y. Huc, and G. Bessereau, 1991, Wireline logging and source rocks — Estimations of organic carbon content by CARBOLOG method: The Log Analyst, **32**, 279–297.

Carvalho, C., R. M. Anjos, R. Veiga, and K. Macario, 2011, Application of radiometric analysis in the study of provenance and transport processes of Brazilian coastal sediments: Journal of Environmental Radioactivity, **102**, 185–192, doi: 10.1016/j.jenvrad.2010.11.011.

Deng, T., J. Ambía, and C. Torres-Verdín, 2020, Multi-well interpretation of wireline logs and core data in the Eagle Ford Shale using Bayesian Inversion: Unconventional Resources Technology Conference, SEG, Global Meeting Abstracts, 4331–4340.

Doyen, P. M., 1988, Porosity from seismic data: A geostatistical approach: Geophysics, **53**, 1263–1275, doi: 10.1190/1.1442404.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth, 1987, Hybrid Monte Carlo: Physics Letters B, **195**, 216–222, doi: 10.1016/0370-2693(87)91197-X.
Eidsvik, J., P. Avseth, H. Omre, T. Mukerji, and G. Mavko, 2004, Stochastic reservoir characterization using prestack seismic data: Geophysics, **69**, 978–993, doi: 10.1190/1.1778241.
El Sharawy, M. S., and G. R. Gaafar, 2012, Application of well log analysis for source rock evaluation in the Duwi Formation, Southern Gulf of Suez, Egypt: Journal of Applied Geophysics, **80**, 129–143, doi: 10.1016/j.jappgeo.2011.12.005.
Feng, R., D. Grana, and N. Balling, 2021, Variational inference in Bayesian neural network for well log prediction: Geophysics, **86**, no. 3, M91–M99, doi: 10.1190/geo2020-0609.1.
Fertl, W. H., and H. H. Rieke, 1980, Gamma ray spectral evaluation techniques identify fractures shale reservoirs and source rock characteristics: Journal of Petroleum Technology, **32**, 2053–2062, doi: 10.2118/8454-PA.
Grana, D., 2020, Bayesian petroelastic inversion with multiple prior models: Geophysics, **85**, no. 5, M57–M71, doi: 10.1190/geo2019-0625.1.
Hoffman, M. D., and A. Gelman, 2014, The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo: Journal of Machine Learning Research, **15**, 1593–623.
Hu, H. T., R. Su, C. Liu, and L. W. Meng, 2016, The method and application of using generalized-ΔLgR technology to predict the organic carbon content of continental deep source rocks: Natural Gas Geoscience, **27**, 145–155.
Huang, Z., and M. A. Williamson, 1996, Artificial neural network modelling as an aid to source rock characterization: Marine & Petroleum Geology, **13**, 277–290, doi: 10.1016/0264-8172(95)00062-3.
Jacobi, D., M. Gladkikh, B. Lecompte, G. Hursan, F. Mendez, J. Longo, S. Ong, M. Bratovich, G. Patton, and P. Shoemaker, 2008, Integrated petrophysical evaluation of shale gas reservoirs: Canadian International Petroleum Conference/SPE Gas Technology Symposium Joint Conference, SPE-114925-MS, P23.
Jarvie, D. M., R. J. Hill, T. E. Ruble, and R. M. Pollastro, 2007, Unconventional shale-gas systems: The Mississippian Barnett Shale of north-central Texas as one model for thermogenic shale-gas assessment: AAPG Bulletin, **91**, 475–499, doi: 10.1306/12190606068.
Kamali, M. R., and A. A. Mirshady, 2004, Total organic carbon content determined from well logs using ΔlogR and neuro fuzzy techniques: Journal of Petroleum Science and Engineering, **45**, 141–148, doi: 10.1016/j.petrol.2004.08.005.
Khoshnoodkia, M., H. Mohseni, O. Rahmani, and A. Mohammadi, 2011, TOC determination of Gadvan Formation in South Pars Gas field: Using artificial intelligent systems and geochemical data: Journal of Petroleum Science and Engineering, **78**, 119–130, doi: 10.1016/j.petrol.2011.05.010.
Larsen, A. L., M. Ulvmoen, H. Omre, and A. Buland, 2006, Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model: Geophysics, **71**, no. 5, R69–R78, doi: 10.1190/1.2245469.
Liu, C., C. H. Yin, and S. F. Lu, 2015, Predicting key parameters for variable-coefficient Δlog*R* logging technique and its application in source rocks evaluation: Natural Gas Geoscience, **26**, 1925–1931.
Logan, P., and I. Duddy, 1998, An investigation of the thermal history of the Ahnet and Reggane Basins, Central Algeria, and the consequences for hydrocarbon generation and accumulation: Geological Society, London, Special Publications, 132, 131–155.
Lüning, S., and S. Kolonic, 2003, Uranium spectral gamma-ray response as a proxy for organic richness in black shales: Applicability and limitations: Journal of Petroleum Geology, **26**, 153–174, doi: 10.1111/j.1747-5457.2003.tb00023.x.
Mackay, D. J., 2003, Information theory, inference, and learning algorithms: Cambridge University Press.
Makhous, M., Y. Galushkin, and N. Lopatin, 1997, Burial history and kinetic modeling for hydrocarbon generation, part II, applying the GALO model to Saharan basins: AAPG Bulletin, **81**, 1679–1699, doi: 10.1306/3B05C41A-172A-11D7-8645000102C1865D.
Martin, O., 2018, Bayesian analysis with python: Packt Publishing Ltd.
McElreath, R., 2016, Statistical rethinking: A Bayesian course with examples in R and Stan: CRC Press.
Mendelzon, J. D., and M. N. Toksoz, 1985, Source rock characterization using multivariate analysis of log data: Presented at the 26th Annual Logging Symposium, SPWLA.
Meyer, B. L., and M. H. Nederlof, 1984, Identification of source rocks on wireline logs by density/resistivity and sonic transit time/resistivity cross plots: AAPG Bulletin, **68**, 121–129, doi: 10.1306/AD4609E0-16F7-11D7-8645000102C1865D.
Mosegaard, K., and A. Tarantola, 2002, Probabilistic approach to inverse problems, *in* W. H. K. Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger, eds., International handbook of earthquake & engineering seismology, Part A: Academic Press, 237–265.
Neal, R. M., and M. Radford, 2011, MCMC using Hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. L. Jones, X.-L. Meng, and A. Tarantola, eds., Handbook of Markov chain Monte Carlo: Chapman & Hall/CRC, 2, 113–162.
Pan, R. F., Y. Wu, and Z. Song, 2009, Geochemical parameters for shale gas exploration and basic methods for well logging analysis: China Petroleum Exploration, **13**, 6–9.
Passey, Q. R., S. Creaney, J. B. Kulla, F. J. Moretti, and J. D. Stroud, 1990, Practical model for organic richness from porosity and resistivity logs: AAPG Bulletin, **74**, 1777–1794, doi: 10.1306/0C9B25C9-1710-11D7-8645000102C1865D.
Qian, K., J. Ning, X. Liu, and Y. Zhang, 2019, A rock physics driven Bayesian inversion for TOC in the Fuling shale gas reservoir: Marine & Petroleum Geology, **102**, 886–898, doi: 10.1016/j.marpetgeo.2019.01.011.
Renchun, H., W. Yan, C. Sijie, L. Shuai, and C. Li, 2015, Selection of logging-based TOC calculation methods for shale reservoirs: A case study of the Jiaoshiba shale gas field in the Sichuan Basin: Natural Gas Industry B, **2**, 155–161, doi: 10.1016/j.ngib.2015.07.004.
Rimstad, K., P. Avseth, and H. Omre, 2012, Hierarchical Bayesian lithology/fluid prediction: A North Sea case study: Geophysics, **77**, no. 2, B69–B85, doi: 10.1190/geo2011-0202.1.
Scales, J. A., and L. Tenorio, 2001, Prior information and uncertainty in inverse problems: Geophysics, **66**, 389–397, doi: 10.1190/1.1444930.
Schmoker, J. W., and T. C. Hester, 1983, Organic carbon in Bakken Formation, United States portion of Williston Basin: AAPG Bulletin, **67**, 2165–2174, doi: 10.1306/AD460931-16F7-11D7-8645000102C1865D.
Sen, M. K., and R. Biswas, 2017, Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm: Geophysics, **82**, no. 3, R119–Z23, doi: 10.1190/geo2016-0010.1.
Sen, M. K., and P. L. Stoffa, 1996, Bayesian inference, Gibbs sampler and uncertainty estimation in geophysical inversion: Geophysical Prospecting, **44**, 313–350, doi: 10.1111/j.1365-2478.1996.tb00152.x.
Sen, M. K., and P. L. Stoffa, 2013, Global optimization methods in geophysical inversion: Cambridge University Press.
Sivia, D., and J. Skilling, 2006, Data analysis: A Bayesian tutorial: Oxford University Press.
Supernaw, I. R., A. D. Mccoy, and A. J. Lind, 1978, Method for in situ evaluation of the source rock potential of earth formations: U. S. Patent US4071755 A.
Tan, M., Q. Liu, and S. Zhang, 2013, A dynamic adaptive radial basis function approach for total organic carbon content prediction in organic shale: Geophysics, **78**, no. 6, D445–D459, doi: 10.1190/geo2013-0154.1.
Tarantola, A., 2005, Inverse problem theory and methods for model parameter estimation: SIAM.
Tixier, M. P., and M. R. Curtis, 1967, Oil shale yield predicted from well logs: Drilling and Production — 7th World Petroleum Congress, Elsevier, WPC-12271.
Verma, S., T. Zhao, K. J. Marfurt, and D. Devegowda, 2016, Estimation of total organic carbon and brittleness volume: Interpretation, **4**, no. 3, T373–T385, doi: 10.1190/INT-2015-0166.1.
Vernik, L., and J. Milovac, 2011, Rock physics of organic shales: The Leading Edge, **30**, 318–323, doi: 10.1190/1.3567263.
Wang, G., T. R. Carr, Y. Ju, and C. Li, 2014, Identifying organic-rich Marcellus Shale lithofacies by support vector machine classifier in the Appalachian basin: Computers & Geosciences, **64**, 52–60, doi: 10.1016/j.cageo.2013.12.002.
Yu, H., R. Rezaee, Z. Wang, T. Han, Y. Zhang, M. Arif, and L. Johnson, 2017, A new method for TOC estimation in tight shale gas reservoirs: International Journal of Coal Geology, **179**, 269–277, doi: 10.1016/j.coal.2017.06.011.
Zhao, T., S. Verma, D. Devegowda, and V. Jayaram, 2015, TOC estimation in the Barnett Shale from triple combo logs using support vector machine: 85th Annual International Meeting, SEG, Expanded Abstracts, 791–796, doi: 10.1190/segam2015-5922788.1.
Zhu, L. Q., C. M. Zhang, S. Zhang, X. Q. Zhou, and W. N. Liu, 2019, An improved method for evaluating the TOC content of a shale formation using the dual-difference Δlog*R* method: Marine and Petroleum Geology, **102**, 800–816, doi: 10.1016/j.marpetgeo.2019.01.031.

Biographies and photographs of the authors are not available.