LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-646027

# Bayesian inversion and optimization of geothermal reservoirs using multivariate adaptive regression spline

M. Chen, R. Mellors, A. Tompson

November 11, 2013

Applied Energy

## Disclaimer

# An efficient Bayesian inversion of a geothermal prospect using a multivariate adaptive regression spline method

Mingjie Chen*, Andrew F.B. Tompson, and Robert J. Mellors

Atmospheric, Earth and Energy Division, Lawrence Livermore National Laboratory, Livermore, CA, USA

*Corresponding author:

Mingjie Chen, PhD
Lawrence Livermore National Laboratory
P.O. Box 808, L-223
Livermore, California 94551
1-925-423-5004 (office)
1-925-423-0153 (fax)
cmj1014@gmail.com

## Abstract

In this study, an efficient Bayesian framework equipped with a Multivariate Adaptive Regression Spline (MARS) technique is developed to alleviate computational burdens encountered in a conventional Bayesian inversion of a geothermal prospect. Fast MARS models are developed from training dataset generated by CPU-intensive hydrothermal models and used as surrogate of high-fidelity physical models in Markov Chain Monte Carlo (MCMC) sampling. This Bayesian inference with MARS-enabled MCMC method is used to reduce prior estimates of uncertainty in structural or characteristic hydrothermal flow parameters of the model to posterior distributions. A geothermal prospect near Superstition Mountain in Imperial County of California in USA is used to illustrate the proposed framework and demonstrate the computational efficiency of MARS-based Bayesian inversion. The developed MARS models are also used to efficiently drive calculation of Sobol' total sensitivity indices. Only top sensitive parameters are included in Bayesian inference to further improve the computational efficiency of inversion. Sensitivity analysis also confirms that water circulation through high permeable structures, rather than heat conduction through impermeable granite, is the primary heat transfer method. The presented framework is demonstrated an efficient tool to update knowledge of geothermal prospects by inversing field data. Although only thermal data is used in this study, other type of data, such as flow and transport observations, can be jointly used in this method for underground hydrocarbon reservoirs.

*Keywords*: Geothermal prospect; inversion; surrogate; uncertainty; sensitivity

## 1. Introduction

Quantitative, model-based prediction of geothermal reservoir behavior requires knowledge of both structural and parametric hydrothermal characteristics of the reservoir system. These include the location, size, and shape of important hydrogeologic flow units and faults, their associated permeability and thermal conductivity, as well as bounding temperature and fluid pressure and saturation conditions. These elements are difficult to fully characterize in the subsurface prior to reservoir development and are, at best, not completely known. This uncertainty will limit the accuracy of reservoir model predictions and reduce the reliability of any type of geothermal production or design that is based upon such predictive capabilities. Inverse methods are often utilized to better identify and estimate uncertain flow properties or system characteristics by matching field observations to corresponding predictive model simulations (Carrera et al., 2005; Hill and Tiedeman, 2007; Tompson et al., 2013; Tonkin and Doherty, 2009), yielding a model with improved accuracy and reduced uncertainties. Depending on the specific techniques used, inverse modeling may be subject to mathematical limitations or require a large number of intensive forward model simulations in order to be successful.

It is widely recognized that incomplete knowledge of the underground fluid reservoir may introduce considerable uncertainty into modeling analyses of such systems and can typically lead to ill-posed nonlinear inverse problems with multiple solutions (Carrera, 1988; de Marsily et al, 1999; Oliver et al., 2008). The idea of "Bayesian inference" has been demonstrated to provide an effective inverse framework which addresses the issues of ill-posedness and non-uniqueness by incorporating prior information and presenting the inversed solution in terms of probability distribution (e.g., Cui et al., 2011; Fu and

Gomez-Hernandez, 2009; Oliver et al., 1997; Tarantola, 2004). In most cases, the posterior densities are difficult to sample directly from the analytical forms of distribution. Markov Chain Monte Carlo (MCMC) approaches provide a more practical simulation method for sampling from target distributions as a means to approximating posterior distributions for parameters or quantities of interest (e.g., Efendiev et al., 2005). However, for many types of large-scale modeling problems, including geothermal prospect characterization, MCMC simulations may become computationally prohibitive because of the large number of uncertain parameters that need to be considered (curse of dimensionality), the absolute size and discretization of the modeling domain of interest, and/or the complexity of the flow physics involved in each forward calculation (Smith and Marshall, 2008; Tompson et al., 2013). Although significant advancements in MCMC sampling efficiency have been made recently (Liu et al., 2000; Mariethoz et al., 2010; Vrugt et al., 2009), the computational burden may still be unaffordable for large-scale high-resolution numerical simulation problems.

Two promising ways to address this challenge include (1) the use of up-front parameter sensitivity analyses to identify the most important parameters pertinent to an analysis of interest, prior to any formal inversion effort, and (2) the development of cheaper "surrogate" forward models for application in the MCMC sampling. Surrogate models attempt to replicate the behavior of complex models with simpler models using fewer but sensitive parameters. Sensitivity is a measure of the contribution of an input variable to the total variances of an output variable. In this study, the sensitivity analysis is used as a "screening" analysis to help focus the stochastic inversion work on

parameters, properties, or characteristics that would appear to be most important, thus focusing computational efforts where they will be most beneficial.

A surrogate model is meant to provide a fast approximation of a high-fidelity physical model calculation. Applications of surrogate modeling techniques in hydrology have been extensively studied in recent years (Razavi et al., 2012). A variety of approximation techniques haven been tried, such as polynomials (e.g., Fen et al., 2009), radial basis functions (RBFs) (e.g., Regis and Shoemaker, 2007), kriging (e.g., Simpson and Mistree, 2001), support vector machines (SVMs) (Zhang et al., 2009), artificial neural networks (ANNs) (Behzadian et al., 2009; Dowla and Rogers, 2003), and sparse grid interpolation (Zeng et al., 2012). However, few studies have applied the multivariate adaptive regression spline (MARS) technique, a non-parametric approximation method developed by Friedman (1991). In a thorough review of surrogate modeling applications in water resources, none of 48 applications reviewed by Razavi (2012) has used the MARS technique. In a recent application of surrogates in optimizing the process of hydraulic fracturing, Chen et al. (2013) compared MARS to a suite of polynomial models including different number of input parameters and in various orders, and surrogate model approximated by MARS was demonstrated to have the best predictive performance. Motivated by this successful application of MARS in optimization, this study employs this function approximation technique to construct a surrogate of hydrothermal flow model for use in applying a Bayesian inversion algorithm to a geothermal prospect.

## 2. Methodology

The proposed methodology involves identification of important, yet uncertain parameters for the problem of interest, sampling in high-dimensional parameter space,

development of numerical hydrothermal flow models, evaluation of response functions, construction and validation of MARS models, performing global sensitivity analyses, and coupled execution of Bayesian inference with MARS models.

## 2.1. Inversion framework

As illustrated in Figure 1, the inversion proceeds as follows:

1. A set of *M* design variables pertinent to the problem are identified and associated with statistical distributions descriptive of their priori uncertainty. In this case, these variables pertain to a series of parametric and structural characteristics of the hydrothermal flow Case Study described below. Their uncertainty, as indicated by their distributions and ranges, are representative of prior knowledge about the target geothermal field and formation.
2. Each of these variables are sampled *N* times from their probability distributions using a Latin Hypercube (LH) method (McKay et al., 1979). These *N* data realizations are used to drive *N* hydrothermal flow model realizations to (in our case below) a natural steady state and their corresponding responses (in our case below, temperatures at a finite set of observation locations) are evaluated from the simulated results.
3. The *N* design variable-response pairs (shaded in Figure 1) are then used as a training and validation dataset to construct a surrogate MARS model. To make best use of limited dataset, leave-one-out cross validation (LOOCV) method is applied to validate the fitted MARS model (Picard and Cook, 1984). Given *N* input samples, a surrogate MARS model is constructed *N* times, each time leaving out one of the input sample from training, and using the omitted sample to test the model.

4. The importance of the design variables in the hydrothermal model system are ranked according to Sobol' total sensitivity indices (Sobol', 1993, 2003), obtained from a global sensitivity analysis of response surface to input design variables. The Sobol' method computes and decomposes the variances of response into fractions attributed to each input (first-order indices) variable and their interactions (second- or higher-order indices), and hence the fractions (sensitivity indices) measure contribution of each input variable to variances of response variable (Chen et al., 2013). Only sensitive variables are included in Bayesian inference.
5. The surrogate model is utilized in place of the hydrothermal flow model within a Bayesian inversion scheme (Tompson et al., 2013), from which posterior distributions of the sensitive variables are inferred from comparisons with observed temperature data. The posterior data distributions represent a subset of the prior data that lead to solutions most consistent with the observed temperature data.

## 2.2. Hydrothermal model

The mass balance equation for transient hydrothermal water flow in saturated porous media is considered in this study can be written as:

$$\partial(\phi\rho)/\partial t = -\nabla\cdot(\phi\rho\mathbf{v}) + Q, \quad (1)$$

where $Q$ is the source/sink term, $\phi$ is the porosity, $\rho$ is the fluid density , and $\mathbf{v}$ is the velocity, which can be computed from Darcy's law: $\phi\mathbf{v} = -\mathbf{k}(\nabla P + \rho g\nabla \mathrm{z})/\mu$, where $\mathbf{k}$ is the permeability tensor, $P$ is the fluid pressure, $\mu$ is the fluid viscosity, $g$ is the gravitational constant, and $z$ is an elevation above some datum. In the current implementation, the principal axes of the permeability tensor are assumed to be aligned with the xyz coordinate system so that $\mathbf{k}$ is diagonal ($k_x$, $k_y$, $k_z$). Permeability of each

geological unit is assumed isotropic ($k_x = k_y = k_z$) so we use K instead of **k** to represent permeability below this point. Note that the density and viscosity will, in general, be temperature ($T$) and pressure ($P$) dependent.

In addition to the mass balance equation, the energy balance equation governing heat transfer can be expresses as:

$$\partial\left[\phi\rho u + (1-\phi)\rho_s C_{ps}(T - T_r)\right]/\partial t = \nabla \cdot [\phi\rho HK\nabla P/\mu + K_T \cdot \nabla \mathrm{T}] + Q_T, \quad (2)$$

where $\phi$ is the rock porosity, $u$ is the water internal energy, $\rho_s$ is the rock density, $C_{ps}$ is the rock heat capacity, $T_r$ is a reference temperature, $H$ is the water enthalpy, $K_T$ is the averaged thermal conductivity of both water an rock, and $Q_T$ is the heat source/sink term.

The above mass and energy balance partial differential equations are discretized and solved numerically in NUFT (Nonisothermal Unsaturated-saturated Flow and Transport), a code developed in Lawrence Livermore National Laboratory and applied successfully in many models simulating mass and heat transfer (Nitao, 1998). NUFT is used to build hydrothermal models in this study. Fault size, temperature $T$ at reservoir bottom boundary, rock permeability $K$ and thermal conductivity $K_T$ of each geologic unit are assumed key uncertain reservoir properties for heat transfer, and hence selected as design variables for inversion stage. All the hydrothermal model simulations run to steady state under natural condition.

## 2.3. Multivariate adaptive regression spline (MARS)

Multivariate adaptive regression spline represents a nonparametric technique which adaptively develops local models in local regions for flexible regression modeling of high dimensional data. A MARS model can be expressed as

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{k} a_i B_i(\mathbf{x}), \quad (3)$$

where $\mathbf{x} \in \mathcal{R}^m$, and $\mathcal{R}^m$ is the *m*-dimensional domain of interest. *k* and $a_i$ are the number and coefficients of associated basis functions $B_i(\mathbf{x})$ given by

$$B_i(\mathbf{x}) = \begin{cases} 1, & i = 1 \\ \prod_{j=1}^{J_i} \left[ S_{ji} \cdot \left( x_{v(j,i)} - t_{ji} \right) \right]_+, & i = 2,3,\ldots \end{cases} \quad (4)$$

where $(\cdot)_+ = \max(0, \cdot)$, $J_i$ is the interaction order of basis $B_i$, that is, the number of variables included in the basis function, $S_{ji} = \pm 1$ is the sign indicators, $v(j,i)$ is the index of the design variable $x$ which is split on knots $t_{ji}$. For example, suppose a basis function is given by $B_i = (x_3 + 2.5)_+[-(x_5 - 3.3)]_+$. Apparently, the interaction order is 2, so $J_i = 2$. The sign indicators $S_{1i} = 1, S_{2i} = -1$. The index for the design variable are $v(1,i) = 3, v(2,i) = 5$, and the knots are given by $t_{1i} = -2.5, t_{2i} = 3.3$.

The first and second derivatives are enforced to match on the boundaries of adjacent regions to ensure continuity between local models. Once the number of locations of knots (points at the boundaries) is adaptively chosen based on the response function changes, the coefficients $a_i$ and basis functions $B_i(\mathbf{x})$ can be examined and determined. Comprehensive illustration of the MARS algorithm can be found in Friedman (1991). Compared to other popular techniques, the use of MARS is limited to automatic engineering design applications (e.g., Sudjianto et al., 1998) and has seldom been reported in hydrothermal literature. The superiority of MARS over other high dimensional regression methods appears to be accuracy and reduction in computational cost of fitting process (Chen et al., 2013; Jin et al., 2001).

## 2.4. MARS-enabled Bayesian inference

The purpose of Bayesian inference is to update the beliefs about uncertain parameters by combining information from the prior distribution and the measurements through the calculation of the posterior distribution. Assuming **x** is the vector formed by design variables to be inversed, **y** is the measurements, Bayes' theorem relates the posterior distribution $p(\mathbf{x}|\mathbf{y})$ to the product of the conditional probability of the measurements $p(\mathbf{y}|\mathbf{x})$ and the prior probability $p(\mathbf{x})$ of design variables as follows:

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y}), \qquad (5)$$

where marginal distribution $p(\mathbf{y}) = \int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}$ is an integral, which doesn't provide any additional information about posterior distribution and can be seen as a normalized constant. $p(\mathbf{x})$ represents uncertainty prior to any knowledge of measurements, and is assumed uniformly distributed within an appropriate range in this study. Hence, the posterior $p(\mathbf{x}|\mathbf{y})$ is proportional only to $p(\mathbf{y}|\mathbf{x})$.

$p(\mathbf{y}|\mathbf{x})$ is also called likelihood function, which quantifies the degree of fit between predictions and measurements. The likelihood can be calculated by forwarding hydrothermal models with the given design variables to steady state, at which the errors between predicted and measured temperatures at observed locations can be included

$$\boldsymbol{\varepsilon} = \mathbf{y} - f(\mathbf{x}), \qquad (6)$$

where $\boldsymbol{\varepsilon}$ is the errors, $f(\mathbf{x})$ is the predictions from NUFT models. The smaller the errors are, the higher the likelihood is. By assuming $\boldsymbol{\varepsilon}$ follows multiple dimensional Gaussian distribution with zero mean and known covariance matrix **C**, the likelihood can be expressed as (Zeng et al., 2012):

$$p(\mathbf{y}|\mathbf{x}) = \exp(-\boldsymbol{\varepsilon}^T\, \mathbf{C}^{-1}\, \boldsymbol{\varepsilon}/2)/\left[(2\pi)^{n/2}|\mathbf{C}|^{1/2}\right], \qquad (7)$$

where $n$ is the number of measurements, $|\mathbf{C}|$ is the determinant of $\mathbf{C}$. In the study, MARS model $\hat{f}(\mathbf{x})$ is used as surrogate of NUFT hydrothermal model $f(\mathbf{x})$ in calculating this Gaussian likelihood function. In this way, the posterior can be obtained without running expensive NUFT models during MCMC sampling, thus accelerating Bayesian inversion significantly.

### 2.5. Implementation

The proposed framework is written in Python by incorporating hydrothermal NUFT models and various numerical codes from PSUADE suite (Tong, 2009), including LH sampling, MARS approximation, Sobol' method, and MCMC algorithm.

## 3. Case studies and discussions

To illustrate and demonstrate the proposed approach, a geothermal prospect at Superstition Mountain in California is chosen as the example study owing to its data availability of geological stratigraphy and borehole temperature logs from Navy geothermal program (Figure 2) (Bjornstad et al., 2006; Tiedeman et al., 2011).

### 3.1. Three-dimensional model development for Superstition Mountain

A three-dimensional geologic model built with digital elevation and layer horizon data in Tompson et al. (2008) are used to conceptualize the geologic structure of Superstition Mountain (Figure 2c). Geophysical and drilling logs from the three boreholes near Superstition Mountain provide additional information to refine the model (Figure 2b). The prospect is bounded on the southwest by granite basement and sedimentary layers to the northeast. A major active fault, the Superstition Mountain fault (SMF), lies near the prospect. The study by Layman Energy Associates (2012) supported

a hypothesis that one or more of the high-permeable principal or cross faults serve as the vertical pathways for hot water flow from deep zones to shallow aquifers through low permeable granite zone. This water circulation is believed to be the cause of elevated temperatures observed in three NAFEC boreholes and nearby shallow temperature surveys. A recent hydrothermal model developed using NUFT for the prospect (Mellors et al., 2013; Tompson et al., 2013), which contained a vertical conjugate fault (CF) transverse to SMF and extending to northeast through NAFEC-3, predicted temperature profiles closely matching temperature logs of three NAFEC boreholes.

This NUFT model domain is adapted in our model as the core region enclosed by a larger far field domain (Figure 3a). The X direction in the core domain is parallel with the CF, while the Y direction is parallel with the SMF. The core model domain used here extends along the X axis 6.5 km to the northeast of the SMF and is restricted to a 1.5 km width in the Y direction, with the center of the left boundary intersecting CF at right angle (Figure 2b). The core domain extends vertically downwards 3.2 km from the ground surface and is discretized into 100 m cubic grid blocks. The current system is considered to be fully saturated throughout the domain. In the future it will be upgraded to more representative partially saturation conditions in the shallower sediments. The larger far-field domain incorporates reduced grid resolution beyond its core as a means to control computational costs. Representation of geological structures in the model grid is simplified from the geologic model. As shown in vertical X-Z section crossing the center of left boundary, where CF is located, the five geologic units crossing the core domain section are sequenced from the bottom as a fractured, low permeability Granite, a permeable sandstone layer Ti, and alluvial sediment layers Tp2, Tp1, Qb, along a

downward slope in X direction (Figure 3b). The left boundary is consistent with SMF at X=0 m, and the 100-m thick permeable vertical CF is normal to the left boundary with uncertain length and height (Figure 3c). Pressure and temperature are specified at the top boundaries to represent the average atmospheric conditions. High temperature is fixed at the bottom boundary to mimic the geothermal heat source, with a lower fixed temperature at the ground surface. Groundwater is allowed to flow from the left (X = 0) to the right (X = 6.5km) sides of the as a result of specified pressure conditions that reproduce a small hydraulic gradient in this area, and is also allowed to enter the bottom of the domain as a result of another fixed pressure condition. Depending on permeability conditions, such inflows may support the generation of hydrothermal inflows that may circulate and exit the right side of the boundary. No flow conditions are maintained along the Y faces of the domain.

Because this is considered a "natural" flow system, the hydrothermal models are used to develop a steady-state flow and temperature solution by running them in a transient mode from provisional initial conditions for one million years. The predicted temperatures at steady state time are compared to measurements at observing locations along three NAFEC boreholes during Bayesian inversion.

### 3.2. MARS models construction

Following the procedure outlined in Figure 1 and Section 2.1, the inversion starts with the identification of a set of parameters to be treated uncertain (design variables). The differential equation governing geothermal heat transfer (Eq. 2) indicates formation permeability $K$ and thermal conductivity $K_T$ will be crucial properties controlling hot groundwater circulation and heat conduction respectively. Thus the two properties for

292 faults, Granite and four sediment formations, totally 12 variables, are included as design
293 variables. The values of temperature fixed at the bottom boundary, which represent the
294 strength of heat source, are partially unknown (certainly warmer than the surface
295 temperature) but may affect steady state temperature distribution across the model
296 domain, and hence are included as design variables too. In addition, the length and height
297 of CF, which is anchored at the SMF at low left corner, are also considered random. The
298 CF unit is considered to be a more permeable feature able to support geothermal
299 circulation into shallower zones if sufficient flow connectivity exists. These 15 design
300 variables, along with their lower and upper bounds, are listed in Table 1. The log-
301 transformed permeabilities and all the other variables are assumed to follow uniformly
302 random distribution across their indicated ranges. A total of 1500 input samples are
303 drawn from the 15-dimensional parameter space using the LH method, with each sample
304 vector containing, as components, 15 values of the design variables. The 15 component
305 values of each sample vector, together with other fixed parameters, are written into the
306 input file of NUFT model for simulation. The temperatures at 23 observation locations
307 (red circles in Figure 4) along the three NAFEC boreholes, obtained from the output of
308 the NUFT model, are used as the response values. The 1500 NUFT model simulations,
309 specifically the 1500 sets of input vectors and 1500 sets of output response values, are
310 used as both a training and a validation dataset to construct the MARS models. Each
311 MARS model $\hat{f}(\mathbf{x})$ consists of 100 basis functions $B_i(\mathbf{x})$, each with 10 orders of
312 interactions $J_i$ (10 design variables $x$). It should be noted that a well-fitted MARS model
313 does not necessarily mean that it will have good performance for prediction due to over-
314 fitting issue, and hence it has to be validated before the use for prediction. The predictive

315 performance of MARS models in this study is measured by LOOCV method (Chen et al.,
316 2013; Picard and Cook, 1984). The quality of the MARS models can be illustrated by
317 scatter plot comparing the response values simulated by the surrogate model versus those
318 simulated by the physical NUFT model, based upon the 1500 samples. As shown in
319 Figure 5, the *R*-squared values obtained for the scatter plots in both the fitting and
320 validation steps of the MARS model (with mean response values) are 0.979 and 0.959
321 respectively, suggesting a good predictive ability of the well-fitted MARS model.

322 3.3. Global sensitivity analysis

323 Sensitivity of model responses to the design variables values can be efficiently
324 calculated using MARS models. The Sobol' total sensitivity indices (SI) for 15 variables
325 are listed in Table 1 and visualized in Figure 6 (Sobol' 2003). As expected, the
326 dimensional characteristics of CF, i.e., height and length, rank as the top two (1 and 2)
327 sensitive variables for defining the temperature distribution in the aquifer where NAFEC
328 boreholes are located (depth < 1000 m), while the permeability of the CF and Ti units,
329 which represent primary groundwater circulation pathways, are moderately sensitive (SI >
330 0.1). The low sensitivity (SI < 0.05) of the specified temperature at bottom boundary
331 (heat source of the model) indicates that its variations between 125 $^{o}$C and 225$^{o}$C lead to
332 little change of the temperature values in shallow aquifers. This finding demonstrates that
333 the efficiency of heat transfer is more important than heat storage for a geothermal field.
334 Low SIs (< 0.05) associated with the thermal conductivity of both granite and the Ti units
335 reveal that the heat conduction is a minor mechanism of heat transfer, compared to
336 groundwater convection, through these two formations in the Superstition Mountain
337 geothermal prospect. It is not surprising that the granite permeability is insensitive (SI <

0.01), given that its value ranges between $10^{-19}$ and $10^{-17}$ $m^2$, which can be considered effectively impermeable, as compared to the crossing CF permeability ($10^{-14}$ – $10^{-12}$ $m^2$). Neither the permeability nor the thermal conductivity of the upper sediment aquifers (Tp2, Tp1, and Qb) is sensitive for temperature around NAFEC boreholes, which is could be considered a potential geothermal production area. It is reasonable since these formations are not the primary groundwater circulation pathways. Overall, the sensitivity quantification and associated ranking for the hydrothermal model system of the geothermal prospect demonstrates that groundwater circulation is the primary mechanism of heat transfer in the field, consistent with previous studies for this area. Reducing the uncertainty of those most sensitive properties, which is critical for potential geothermal reservoir development and management, is a critical priority of exploration investment. In addition to those expensive geophysical surveying approaches, Bayesian inversion equipped with fast MARS models is applied to achieve better knowledge of these important properties from the temperature observations shown as red circles in Figure 4.

### 3.4. Bayesian inversion with MARS-enabled MCMC

The MARS surrogate model was used to enable a MCMC-based Bayesian inversion process using the prior probability density functions (PDFs) shown in Table 1 for the six top sensitive design variables identified in Figure 6. The MCMC procedure starts with burn-in phase in which 10,000 MARS model simulations are employed. During the following phase of creating posteriors, the chain converges after MARS model calls amount to three sequential sample increments, with each 10,000 in size for convergence check. The total MARS model runs for the complete MCMC, therefore, is 40,000 in this case of Bayesian inversion, which cost about 5 minutes of computing time, while an

equivalent NUFT model simulation on a scalar machine take around 10 minutes averagely. Compared to NUFT-based inversion, the MARS-based approach is projected to improve the inversing efficiency by 10/5*40,000= 80,000 times. Although NUFT model simulations were, in fact, used to support the inversions described in Tompson et al (2013) on a similar model domain as the core domain in this study, they were accomplished using a parallel implementation of NUFT and exploited the naturally parallel benefits of conducting multiple MCMC simulation chains. That said, the power of the MARS method can be most effectively exploited when, for example, a larger scale and higher resolution model grid is used for more realistic, variably-saturated hydrothermal flow simulations, a configuration that the NUFT-only platform cannot currently address in an efficient, cost effective manner.

Among the six posterior PDFs are shown in Figure 7, the two least sensitive variables, bottom temperature and granite thermal conductivity, are almost equally likely in their ranges, suggesting that little additional knowledge is gained from prior information by Bayesian inference due to their low identifiability. Among the other four variables, CF height is identified as 3200 m at its upper bound that results in best matches with observation data with a highest probability of 0.45, more than twice the magnitude of the second highest probability. This result strongly suggests CF fault penetrates the entire granite zone vertically. Figure 7b indicates that CF length should be 1100m in order to best match the data, with the highest probability of 0.33. The probabilities of lengths larger than 1100 m are much higher than those for smaller values, indicating CF should be long enough in order to maintain contact with (and support fluid flows into) the shallower and permeable Ti formation (Figure 3b). This finding makes sense since a

384 contiguous connection through a permeable CF and Ti supports a viable groundwater
385 circulation pathway to convective heat transfer to the observation wells. The log-
386 transformed permeabilities of the CF and Ti units have highest probability of 0.138,
387 0.146 at values of -13.16 and -13.44 $m^2$ respectively. While the possible CF permeability
388 clusters in the mean value of its range, Ti permeability is prone to higher values within
389 the range. The comparison between simulated results by NUFT model using the
390 parameter set with highest probability and measured temperatures along three NAFEC
391 boreholes shows a good match in Figure 4. The corresponding temperature distributions
392 on vertical slice consistent with CF, and horizontal slice at depth of 800 m, are contoured
393 in Figure 3b and 3c, respectively.

394 **4. Conclusions**

395 In this study, an efficient Bayesian inference framework equipped with multivariate
396 adaptive regression spline (MARS) method has been used to reduce geological
397 uncertainties associated with evaluation of a geothermal prospect. Fast surrogate models
398 for hydrothermal flow were constructed by a MARS-based approach for use in a
399 Bayesian MCMC inversion procedure. Computational efficiencies gained in this process
400 (over traditional high-fidelity hydrothermal simulation codes) suggest that more complex
401 aspects of the system can be ultimately addressed, certainly when the costs of physical
402 models becomes too unwieldy. In addition, Sobol' total sensitivity indices for each design
403 variable can also be efficiently calculated using a MARS model instead of a higher-
404 fidelity code. Insensitive variables were screened out of inverse process, enabling
405 Bayesian inference to be conducted that much more efficiently. Owing to the data
406 availability, a geothermal prospect near Superstition Mountain was chosen as the pilot

site to test the efficiency and validity of this method. It was demonstrated that MARS-enable Bayesian inference entailing 40 thousands model runs can be accomplished in 5 minutes, while an individual high-fidelity model (NUFT) run can cost around 10 minutes.

Future work will be focused on adapting the MARS technique to more realistic problems that incorporate larger and higher-resolution domains, or variably saturated flow conditions, aspects that could not have been effectively addressed with high-fidelity hydrothermal models. In addition, the MARS technique can be further utilized in subsequent optimization calculations that may be associated with the design and engineering of a geothermal production operation. In this case, an optimization phase involving hundreds or even thousands of objective function evaluations of reservoir performance under various design configurations could be more readily conducted using a MARS-based simulation approach. Higher-fidelity hydrothermal models incorporating transient source/sink term will cost much more computational time than those in natural condition in this work. This type of operation will also be better constrained, of course, once a viable hydrothermal model of an undisturbed prospect is achieved from as inversion process as described above. Preliminary numerical experiments show that a single geothermal production model simulating up to 1000 years of operation lasts about 5 hours. While only temperature data are used in study to demonstrate the developed method, various data sources are possible to be jointly inversed by extending the current Bayesian framework.

## Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. We

would like to thank DOE GTO office for supporting this project under award DE-EE24675.

We also appreciate the Navy geothermal program for providing data.

## References

Behzadian K, Kapelan Z, Savic D, Ardeshir A. Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks. Environ Model Software 2009;24(4):530-41.

Bjornstad S, Hall B, Unruh J, Richards-Dinger K. Geothermal Resource Exploration, NAF El Centro- Superstition Mountain Area, Imperial Valley, California. Geoth Res Council Trans 2006;30.

Carrera J. State of the art of the inverse problem applied to the flow and solute transport problems, Groundwater Flow and Quality Modeling, NATO ASI Ser 1988;224:549-83.

Carrera J, Alcolea A, Medina A, Hidalgo J, Luit J. Inverse problem in hydrogeology. Hydrogeol J 2005;13:206-22. doi:10.1007/s10040-004-0404-7.

Chen M, Sun Y, Fu P, Carrigan CR, Lu Z, Tong C, Buscheck TA. Surrogate-based optimization of hydraulic fracturing in pre-existing fracture networks. Comput & Geosciences 2013;58:69-79. doi:10.1016/j.cageo.2013.05.006.

Cui T, Fox C, O'Sullivan MJ. Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. Water Resour Res 2011;47:W10521. doi:10.1029/2010WR010352.

de Marsily G, Delhomme J, Delay F, Buoro A. 40 years of inverse problem in hydrogeology. CR Acad Sci Ser II A Earth Planet Sci 1999;329(2):73-87.

Dowla FU, Rogers LL. Solving problems in environmental engineering and geosciences with artificial neural networks. Cambridge, MA:MIT Press;2003.

Efendiev Y, Datta-Gupta A, Ginting V, Ma X, Mallick B. An efficient two-stage Markov chain Monte Carlo method for dynamic data integration, Water Resour Res 2011;41:W12423. doi:10.1029/2004WR003764.

Fen CS, Chan CC, Cheng HC. Assessing a response surface-based optimization approach for soil vapor extraction system design. J Water Resour Plann Manage 2009;135(3): 198-207.

Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991;19(1):1-67.

Fu J, Gomez-Hernandez J. Uncerainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov Chain Monte Carlo method. J Hydrol 2009;364:328-41.

Hill MC, Tiedeman CR. Effective groundwater model calibration, with analysis of data, sensitivities, predictions, and uncertainty. New York: John Wiley; 2007, 480 pp.

Jin R, Chen W, Simpson TW. Comparative studies of metamodelling techniques under multiple modelling criteria. Struct Multidisc Optim 2001;23:1-13.

Layman Energy Associates, Inc. Superstition Mountain geothermal project. http://laymanenergy.com/Superstition-Mountain.html; 2012.

Liu JS, Liang FM, Wong WH. The use of multiple-try method and local optimization in Metropolis sampling. J Am Stat Assoc 2000;95(449):121-34.

McKay M, Beckman R, Conover W. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21(2):239-45.

Mariethoz G, Renard P, Caers J. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resour Res 2010;46:W11530. doi:10.1029/2010WR009274.

Mellors RJ, Ramirez AL, Tompson AFB, Chen M, Yang X, Dyer KM et al. Stochastic joint inversion of a geothermal prospect. 38th Workshop on Geothermal Reservoir Engineering Stanford University, Palo Alto, CA, February11–13; 2013.

Nitao JJ. Reference manual for the NUFT flow and transport code, version 2.0, Technical Report UCRL-MA-130651. Lawrence Livermore National Laboratory, Livermore, CA; 1998.

Oliver D, Cunha L, Reynolds A. Markov chain Monte Carlo methods for conditioning a log permeability field to pressure data, Math Geosci 1997;29:61-91.

Oliver D, Reynolds A, Liu N. Inverse theory for petroleum reservoir characterization and history matching. Cambridge University Press; 2008.

Picard RR, Cook RD. Cross-validation of regression models. J Am Stat Assoc 1984;79:575-83.

Razavi S, Tolson BA, Burn DH. Review of surrogate modeling in water resources. Water Resour Res 2012;48:W07401. doi:10.1029/2011WR011527.

Regis RG, Shoemaker CA. A stochastic radial basis function method for the global optimization of expensive functions. INFORMS J Comput 2007;19(4):497-509.

Smith TJ, Marshall LA. Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques. Water Resour Res 2008;44:W00B05. doi:10.1029/2007WR006705.

Simpson TW, Mistree F. Kriging models for global approximation in simulation-based multidisciplinary design optimization, AIAA J 2001;39(12):2233-41.

Sobol' IM. Sensitivity estimates for non-linear mathematical models. Math Modeling Comput Exp 1993;4:407-14.

Sobol' IM. Theorems and examples on high dimensional model representation. Reliab Eng Syst Saf 2003;79(2):187-93.

Sudjianto A, Juneja L, Agrawal A, Vora M. Computer aided reliability and robustness assessment. Int J Rel Qual Saf Eng 1998;5:181-93. doi: 10.1142/S0218539398000182.

Tarantola A. Inverse problem theory and method for model parameter estimation. Philadelphia, PA: SIAM; 2004.

Tiedeman A, Bjornstad S, Alm S, Frazier L, Meade D, Page C et al. Intermediate depth drilling and geophysical logging results at superstition mountain, Naval Air Facility El Centro, California. Geoth Res Council Trans 2011;35:1037-44.

Tompson AFB, Demir, Z, Moran, J, Mason, D, Wagoner, J. Kollet, S, Mansoor, K, and McKereghan, P, Groundwater availability within the Salton Sea Basin: Final report, Lawrence Livermore National Laboratory, Livermore, CA; LLNL-TR-400426; 2008.

Tompson AFB, Mellors, RJ, Ramirez, A, Chen, M, Dyer, K, Yang, X, Wagoner, J, and Trainor-Guitton, W. Evaluation of A Geothermal Prospect Using A Stochastic Joint Inversion Modeling Procedure, Proceedings for the Geothermal Resources Council 37th Annual Meeting, 29 Sept–2 Oct, 2013, Las Vegas, NV, USA; 2013.

Tong C. PSUADE User's Manual (Version 1.2.0). Lawrence Livermore National Laboratory, Livermore, CA; LLNL-SM-407882; 2009.

Tonkin M, Doherty J. Calibration-constrained Monte Carlo analysis of highly-parameterized models using subspace techniques. Water Resour Res 2009;45(12):W00B10. doi:10.1029/2007WR006678.

Vrugt JA, ter Braak CJF, Diks CGH, Higdon D, Robinson BA, Hyman JM. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int J Nonlin Sci Numer Simul 2009;10(3):273-90.

Zhang XS, Srinivasan R, Van Liew M. Approximating SWAT model using artificial neural network and support vector machine. J Am Water Resour Assoc 2009;45(2):460-74.

Zeng L, Shi L, Zhang D, Wu L. A sparse grid based Bayesian method for contaminant source idenficiation. Adv Water Resour 2012;37:1-9.

Tables and Figures

Table 1. Ranges of input design variables in constructing MARS models. The importance of inputs are ranked according to Sobol' total sensitivity indices for average temperatures along the three NAFEC boreholes

| Input parameter set | Min | Max | Indices | Rank |
|---|---|---|---|---|
| Fault height (m) | 100 | 3200 | 0.712 | 1 |
| Fault length (m) | 100 | 3200 | 0.406 | 2 |
| Fault log permeability ($m^2$) | -14 | -12 | 0.119 | 3 |
| Ti log permeability ($m^2$) | -15 | -13 | 0.107 | 4 |
| Bottom boundary temperature ($^oC$) | 125 | 225 | 0.0378 | 5 |
| Granite thermal conductivity (W/m-C) | 0.1 | 4.0 | 0.0286 | 6 |
| Ti thermal conductivity (W/m-C) | 0.1 | 4.0 | 0.0139 | 7 |
| Granite log permeability ($m^2$) | -19 | -17 | 0.007 | 8 |
| Fault thermal conductivity (W/m-C) | 0.1 | 4.0 | 0.0037 | 7 |
| Tp1 thermal conductivity (W/m-C) | 0.1 | 4.0 | 0.0033 | 10 |
| Tp2 log permeability ($m^2$) | -15 | -13 | 0.0 | 11 |
| Tp2 thermal conductivity (W/m-C) | 0.1 | 4.0 | 0.0 | 12 |
| Tp1 log permeability ($m^2$) | -15 | -13 | 0.0 | 13 |
| Qb thermal conductivity (W/m-C) | 0.1 | 4.0 | 0.0 | 14 |
| Qb log permeability ($m^2$) | -15 | -13 | 0.0 | 15 |

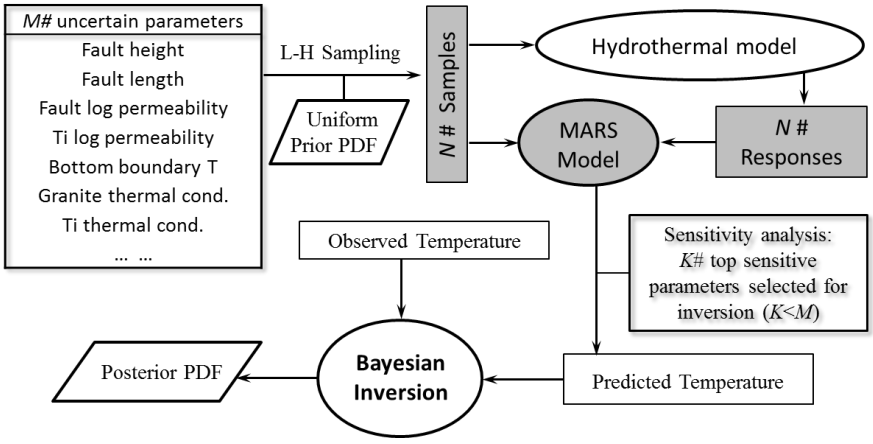


Figure 1. Schematic diagram of the MARS-based Bayesian inversion framework. The gray-shaded boxes indicate the construction of the training dataset used to develop the MARS surrogate model. The full list of design variables is shown in Table 1. The Bayesian Inversion within the oval is conducted with MCMC simulation using the MARS surrogate model.

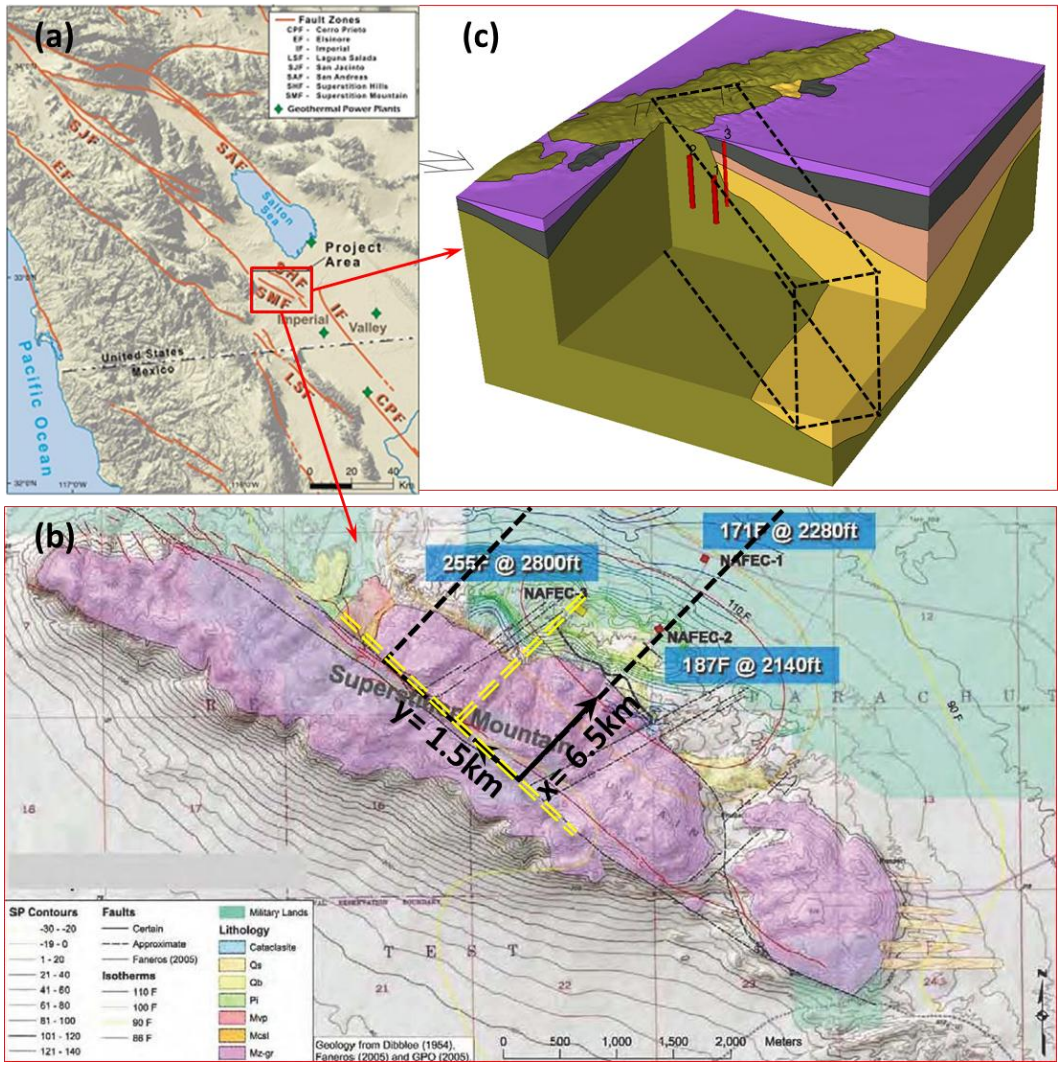


Figure 2. Superstition mountain geothermal prospect. (a) Location in Imperial County, California, USA (Bjornastad et al., 2006); (b) Surface geology and three NAFEC boreholes. Black and yellow dashed lines show the areal projection of the core domain and faults of hydrothermal models (Figure 3). Adapted from Tiedeman et al. (2011); (c) Geological model looking from the Northeast, and showing (from bottom) the granite basement, sandstone Ti, and sedimentary layers Tp2, Tp1, Qb (Figure3b). Dashed box outlines the 3D core domain. The three boreholes are illustrated with red tubes.

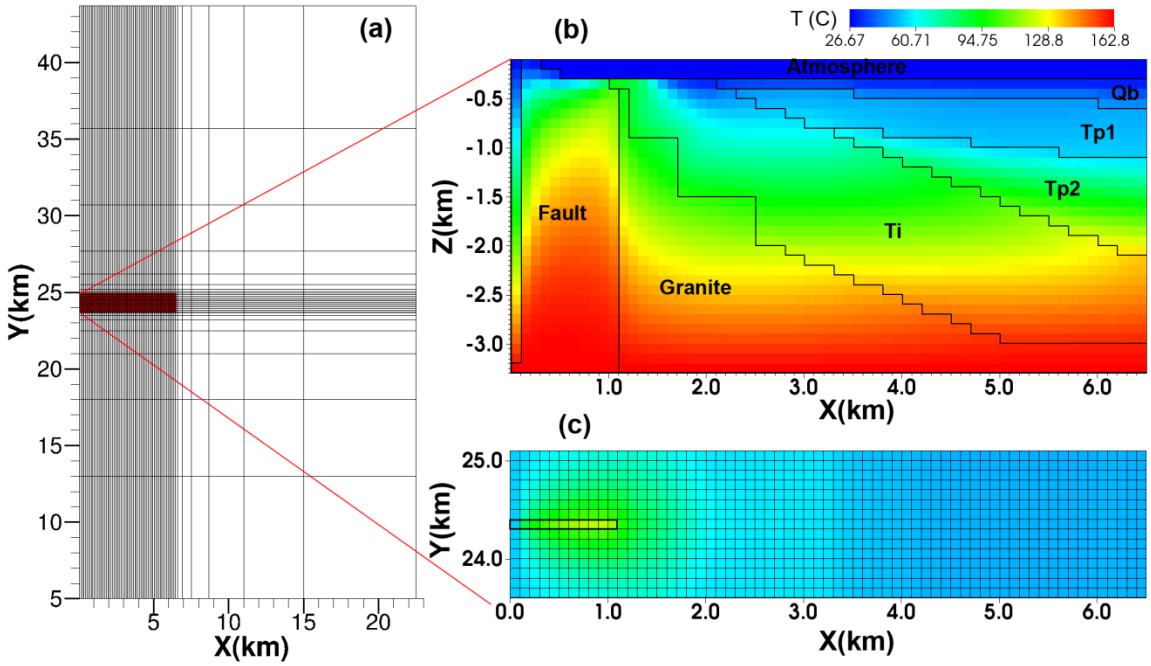


Figure 3. Hydrothermal model domain showing (a) Far field and core area plan view (red shaded area); (b) Vertical slice of the core model domain at Y=28km, where conjugate fault is located. The fault height and length, and the temperature distribution are corresponding to the input parameter set with the highest probability inferred from Bayesian inversion (Figure 7); (c) Horizontal slice of the core model domain at Z = 800m.

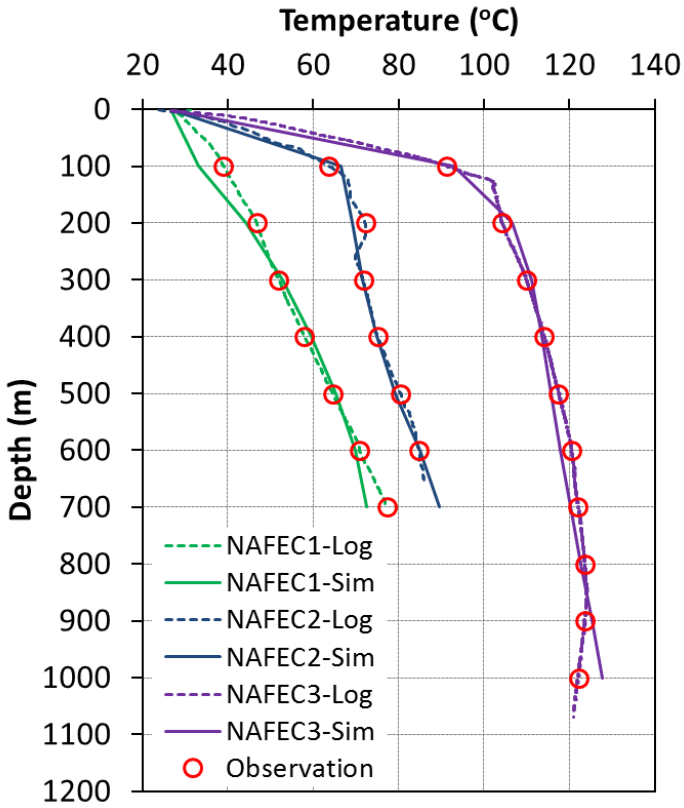

Figure 4. Measured and simulated temperature profiles along the three "NAFEC" boreholes (Tiedeman et al., 2001). The parameter set obtained from inversion with highest probability is used in simulation. The red circle marks indicate the discrete locations along the measured data curves used as observations in the stochastic inversion process.

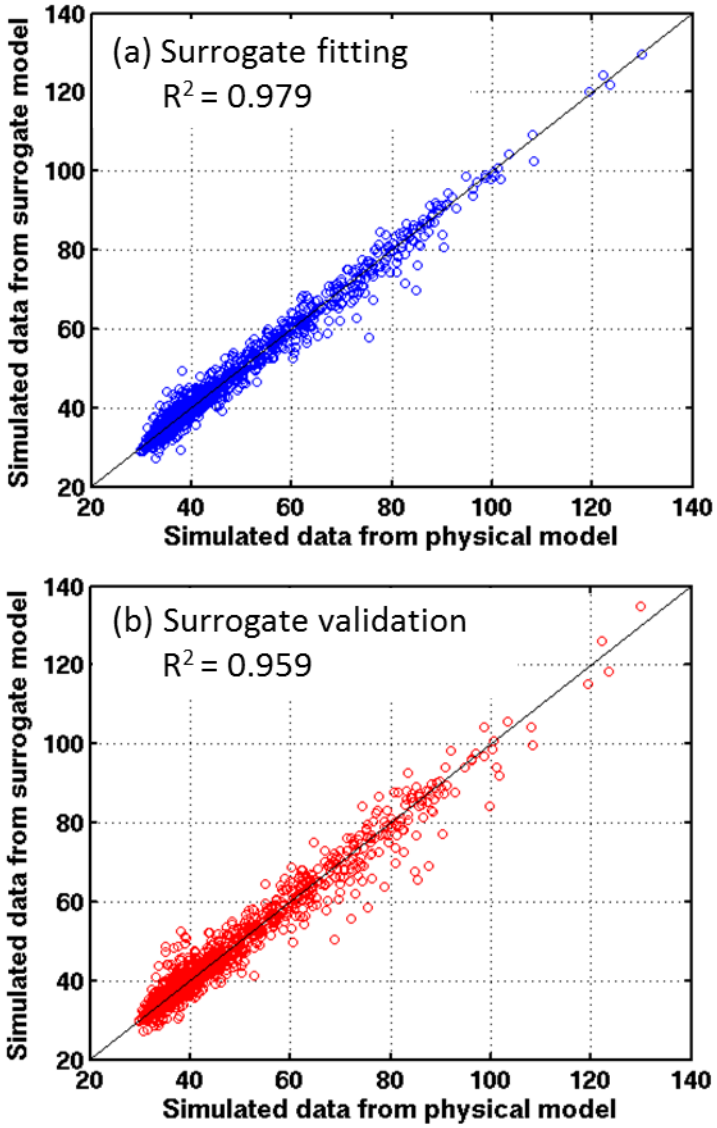


Figure 5. Scatter plots of mean temperature in the three observation wells obtained from 1500 surrogate and physical model simulations. Plot (a) corresponds to the surrogate model fitting step, while plot (b) corresponds to the surrogate model –cross-validation step.

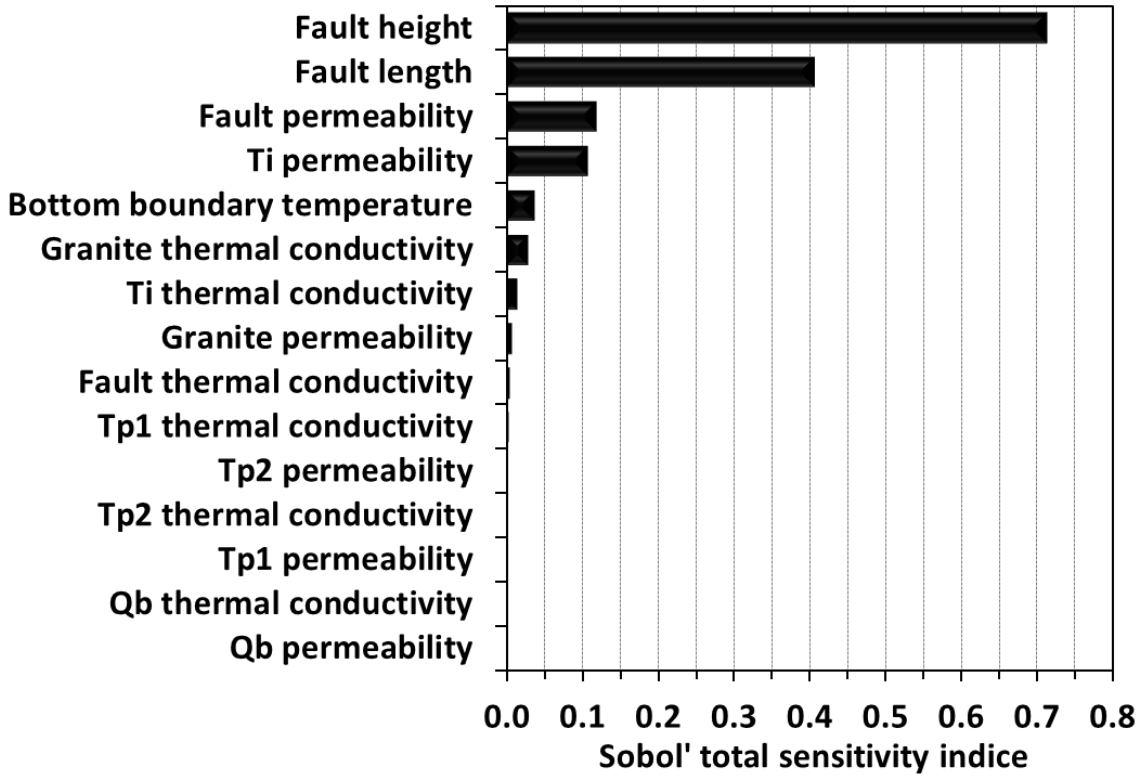


Figure 6. Parameters ranking according to the sensitivity of mean temperature along three boreholes to the 15 hydrothermal parameters (Table 1). The sensitivity is measured by Sobol' total indice.

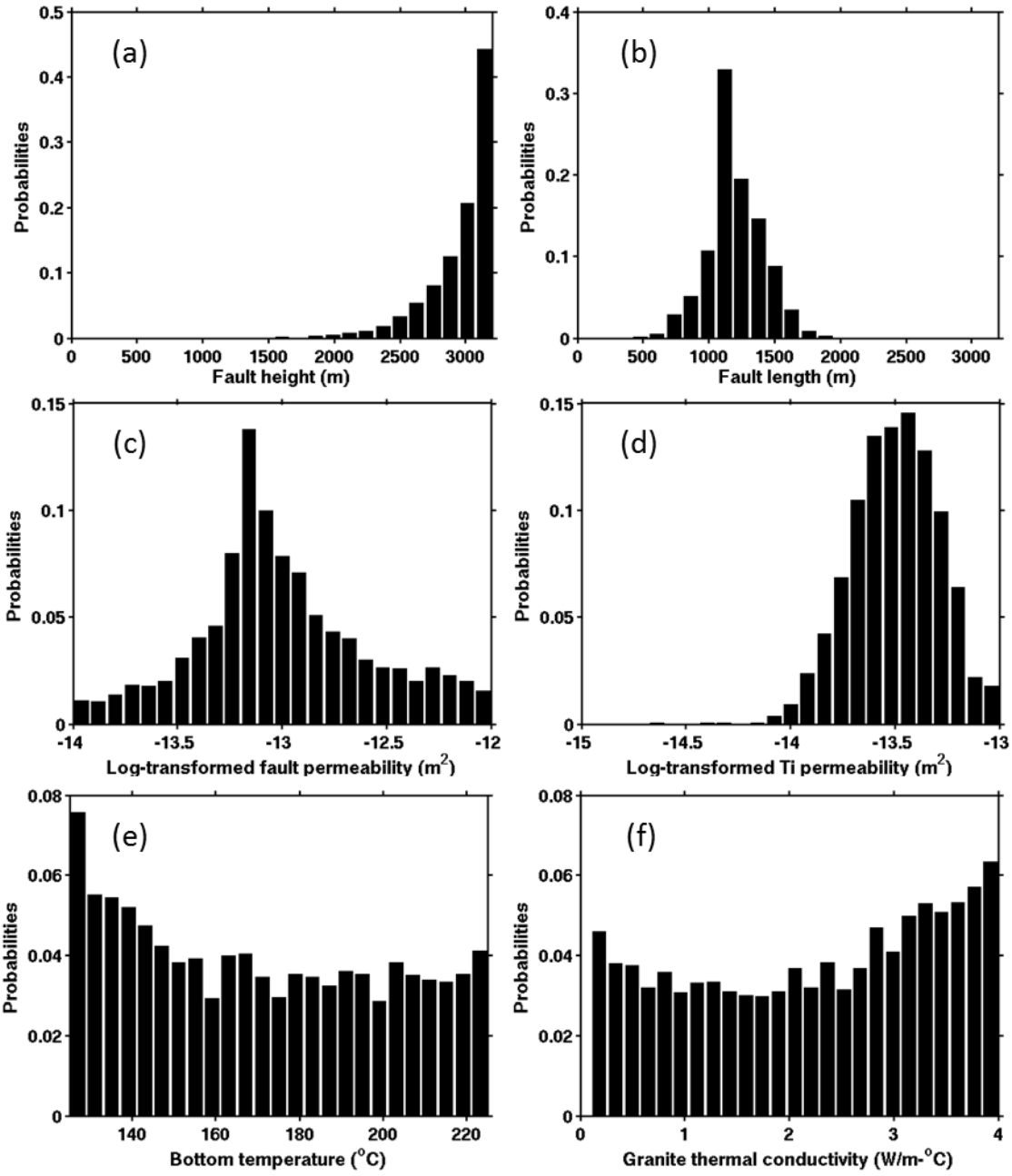


Figure 7. Posterior probability density function (PDF) of the six most sensitive parameters: (a) fault height, (b) fault length, (c) fault permeability, (d) Ti permeability, (e) bottom boundary temperature, and (f) granite thermal conductivity. Note the prior probability of each parameter is uniformly distributed within its range.