A Bayesian Forward Simulation Approach to Establishing a Realistic Prior Model for Complex Geometrical Objects

by

Yizheng Wang

Department of Statistical Science
Duke University

Date: ___________________________
Approved:

____
Alan E. Gelfand, Supervisor

____
Alexander Volfovsky

____
Peter David Hoff

Dissertation submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2018

Abstract

A Bayesian Forward Simulation Approach to Establishing a Realistic Prior Model for Complex Geometrical Objects

by

Yizheng Wang

Department of Statistical Science
Duke University

Date: ____________________
Approved:

____
Alan E. Gelfand, Supervisor

____
Alexander Volfovsky

____
Peter David Hoff

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2018

Abstract

Geology is a descriptive science making itself hard to provide quantification. We develop a Bayesian forward simulation approach to formulate a realistic prior model for geological images using the Approximate Bayesian computation (ABC) method. In other words, our approach aims to select a set of representative images from a larger list of complex geometrical objects and provide a probability distribution on it. This allows geologists to start contributing their perspectives to the specification of a realistic prior model. We examine the proposed ABC approach on an experimental Delta dataset and show that, on the basis of selected representative images, the nature of the variability of the Delta can be statistically reproduced by means of the IQSIM, a state-of-the-art multiple-point geostatistical (MPS) simulation algorithm. The results demonstrate that the proposed approach may have a broader spectrum of application. In addition, two different choices for the *size* of the prior, i.e., the number of representative images are compared and discussed.

For my parents, Alan and Jef

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I offer my deepest gratitude to my advisor, Alan Gelfand, for his perpetual support and guidance. I am so glad that I was fortunate enough to have Alan as my advisor and this thesis would probably not have existed without his supervision. His diligence, scientific curiosity and rigor have been an absolute inspiration to me, and I look forward to continue working together as colleagues and friends for many more years to come.

Next, I extend my sincerest thanks to Jef Caers, for his invaluable support and encouragement to this project and collaboartion. Even though I have received training in both petroleum engineering and Bayesian statistics, without Jef and his scholarly work, I most surly would have been lost at sea — I was so confused and had no idea how to integrate these two subjects. Most importantly, I have discovered my real passion that is to learn and explore in this new and exciting field. For that I am forever thankful.

I would also like to thank Peter Hoff and Alexander Volfovksy for serving as my committee members. They have a lager impact on my experience at Duke and their enthusiasm in the subjects and attractive styles of teaching stimulated my interests in Bayesian statistics.

Last but not least, I do not know if it is possible to convey in words the sense of my greatest appreciation, but I want to take this opportunity to say thank you to my parents for their unconditional love. I am blessed to call them as mother and father and able to talk with them about anything that I am going through. My ultimate goal in my life is one day

they will be proud to say “that is my son"!

Don’t ask what the world needs. Ask what makes you alive, and go do it. Because what the world needs is people who have come alive. – Howard Thurman

1

Introduction

Periods of unprecedented development and prosperity in the history of mankind result from ways of using energy. Energy is a such significant component in our everyday lives. The subsurface system offers many energy sources (Scheidt et al., 2018), such as geothermal heat and fossil fuels like natural gas and oil. However, with the distinct heterogenous nature of the geological depositions, it makes exploring subsurface systems become a high-risk field. Due to this inherent degree of uncertainty, making predictions in practical applications such as, for example, groundwater management, contaminant remediation and oil/gas production, turns out to be a challenge task. From the point of view adopted here, Bayesian inference is the most effective approach for addressing questions of uncertainty. In real field applications, lots of scholarly work tends to employ traditional models such as multivariate Gaussian distributions or processes to model the geological variability in the subsurface (Mariethoz and Caers, 2014). However, the geological variability is too complex to be captured by these well-known standard distributions, which suggests these models may be unrealistic.

Bayesians theorize that inferential gains can be found by combining data with subject matter expertise characterized by the concept of prior. Therefore, a common problem in geological sciences under a Bayesian framework is to formulate a realistic prior model. The goal of this work is to create a vehicle by which the geologist can start to contribute “information" to this prior. Here, information refers to 1) knowledge in terms of interpretation of site-specific data and 2) knowledge regarding the way processes create depositional systems, say a fluvial system. In multiple-point geostatistics, a training image is used to quantify spatial variability, and geologists believe it is appropriate to use training images as representative sets of a system (Scheidt et al., 2016; Renard and Mariethoz, 2014; Mariethoz and Caers, 2014). There are two ways currently allowing us to study the process of geological deposition: conducting an experiment in a laboratory or simulating through powerful computer models. Either way will render a set of highly dependent samples/realizations, and hence such representation carries too much redundancy. Thus, a general question arises: how to select a representative set that is much smaller than the larger set but still contain as much information as possible?

Although considerable research has been devoted to developing better and faster MPS simulation algorithms, rather less attention has been paid to addressing the above question. In existing work, the way to select a set of representative images among the larger set is primarily based on the clustering technique. With the training image selection methods mentioned in Scheidt et al. (2016), the nature of the variability of the delta can be statistically reproduced by means of the process model. However, these methods are not able to assign probabilities to their selections and hence such selection should not be considered as a prior model. In this thesis, we propose a Bayesian forward simulation approach by taking advantage of the Approximate Bayesian computation (ABC) method to establishing a realistic prior model that is qualified from both geological and statistical perspectives. In other words, the proposed approach can offer 1) a set of representative images and 2) a probability model over it.

The thesis is organized as follows. In chapter 2, I briefly introduce Approximation Bayesian computation (ABC) and multiple-point geostatistics (MPS). In chapter 3, I describe the proposed ABC model in detail and how it can be used in a more general context. In the next chapter, an experimental Delta dataset is introduced and examined under the proposed approach. We also include the results in this chapter. A sensitivity check regarding the number of training images is shown in chapter 5. The thesis ends with the main conclusions of this work.

2

Statistical and Geological Methods

2.1 Approximate Bayesian computation (ABC)

Approximate Bayesian computation (ABC) is a commonly used inference method that allows us to bypass the specification of an intractable likelihood when dealing with complex model. Its name was first proposed by Beaumont et al. (2002) in which they use this technique in population genetics problems to approximate the posterior distribution without explicitly involving likelihood calculations. But, the fundamental and essential idea of ABC can be traced back to a 1984 paper by Rubin et al.. In that paper, he stated:“Bayesian statistics and Monte Carlo methods are ideally suited to the task of passing many models over one dataset" which is viewed as a source of inspiration for the ABC approach (Marin et al., 2012). Independently of Rubin’s work, Tavaré et al. (1997) proposed a rejection algorithm to estimate the likelihood of coalescence time given the number of segregating sites. This rejection scheme was considered as the first ABC algorithm including ideas which comprise contemporary ABC (Olson, 2016) and was treated as a special case of an acceptance-rejection scheme (Robert, 2004; Marin et al., 2012). The generalized rejection algorithm is illustrated in Algorithm 1.

Algorithm 1: Generalized rejection algorithm

Input : Observed data Y_{obs},Model $\mathcal{P}$,prior distribution $\pi(\theta)$
Output: N samples from the posterior $\pi(\theta|Y_{obs})$

$V \leftarrow \{\}$;
for $i \leftarrow 1$ **to** N **do**
  repeat
    generate $\theta' \sim \pi(\theta)$;
    drawn y from $\mathcal{P}$ *with* θ' as parameter ;
  until $y = y_{obs}$;
  append θ' *to* V;
return V

Here, θ refers to the unknown variable and model $\mathcal{P}$ can be treated as likelihood. However, this algorithm has very limited potential to be applicable to the problem of collecting samples from the posterior distribution since a simulated data is very unlikely to be exactly same as the observed data, corresponding to a very low acceptance rate. This is particularly the case for continuous outcomes since $p(y = y_{obs}) \approx 0$, implying this algorithm is practically infeasible (Marin et al., 2012). In order to fix this problem, there is a modified algorithm that is laid out in algorithm 2, and most of this work is built on top of it. Rather than accepting the simulated data in terms of the exact equality, the simulated data can be accepted as long as it is close enough to the observed data. Mathematically, we can introduce a closeness metric, $d(\cdot,\cdot)$, and a user defined tolerance, ϵ. Simulated data, y, can be accepted when $d(y, y_{obs}) \leqslant \epsilon$. The choice for ϵ depends on the specific application context. It is worth noting that even though a large value of ϵ results in a higher acceptance rate, it will cause low accuracy of the simulated posterior distribution to the true posterior distribution. Theoretically, a simulated data is sampled from the true posterior distribution when $\epsilon = 0$. When ϵ is approaching to ∞, simulated data comes from the prior.

ABC has appeared in the literature now for over a decade and been well developed in the last couple years. There are numerous different algorithms and modified versions in the

Algorithm 2: Generalized rejection algorithm

Input : Observed data Y_{obs},Model $\mathcal{P}$,prior distribution $\pi(\theta)$, tolerance ϵ
Output: N samples from the posterior $\pi(\theta|Y_{obs})$

$V \leftarrow \{\}$;
for $i \leftarrow 1$ **to** N **do**
  repeat
    generate $\theta' \sim \pi(\theta)$;
    drawn y from $\mathcal{P}$ *with* θ' as parameter ;
  until $d(y, y_{obs}) \leqslant \epsilon$;
  append θ' *to* V;
return V

ABC literature, which makes it have broader application potential.

2.2 Multiple-point geostatistics (MPS)

Multiple-point geostatistics (abbr. MPS) was invented by Andre Journel in 1993 and he has also nurtured its development since then (Renard and Mariethoz, 2014). It aims to reproduce and learn the physical patterns of given training images (abbr. TIs) that are physical realities and used to quantify spatial variation (Mariethoz and Caers, 2014). On the basis of this “learning", it will then be able to create some stochastic realizations that incorporate essential patterns and some trends of the study problem (Scheidt et al., 2016). In terms of investigating trends, we essentially return to the forecasting problems. To address this, most engineers are naturally inclined to prioritize the physical sciences. They concentrate their efforts on the understanding of theoretical ideas by focusing on complex partial differential equations without emphasizing the measurement of uncertainty. Geological science is a high-risk field due to the inherent degree of uncertainty that is present, making its measurement vital. On the other hand, statistical modeling approaches can effectively address questions of uncertainty but they take physical realism into consideration less easily. “MPS was primordially born out of a need to address the insure of lack of physical realism as well as the lack of control in the simulated fields in traditional modeling" (Mariethoz and Caers, 2014). In general, MPS uses non-parametric statistics and treats TIs as the vehicle to model complex patterns. “Multiple" creates a contrast with traditional two point statistics, that is, covariance or variograms to mimic physical reality (Renard and Mariethoz, 2014).

Numerous MPS simulation algorithms exist or under development. The simulator used in this thesis is called IQSIM, a process-based algorithm, which is the current state-of-the-art simulation algorithm in terms of computation time. It takes advantage of a new image quilting technique that bypasses ad-hoc weighting of auxiliary variables by introducing a new probabilistic data aggregation method (Hoffimann et al., 2017).

3

Methodology

3.1 Proposed ABC model

Let $\tilde{X}$ denote for a set of training image that contains x_i, $i = 1, ..., r$, where hyper-parameter r stands for the number of training images in each set. The optimal choice for r can be tuned through the experiments. In order to select these $\tilde{X}$'s, some prior information will be needed. A discrete uniform distribution will be imposed on them. Let $\pi(\tilde{X})$ denote the uniform prior on these $\tilde{X}$'s. Denote a dynamic sequence of the observed data as $\tilde{Y}_{obs} = y_{j,obs}$, $j = 1, 2, ..., n$. In this work, we assume that the true data geometry mechanism is specified by an approximate process model as follows:

$$f(\tilde{Y}_{obs}|\tilde{X})\pi(\tilde{X})$$

where $\tilde{X}$ can be drawn from $\pi(\tilde{X})$, and $\tilde{Y}_{obs}$ can be sampled from $f(\tilde{Y}_{obs}|\tilde{X})$, a "black box", using the IQSIM algorithm.

From the Bayesian perspective, the goal is to get the posterior $\pi(\tilde{X}|\tilde{Y}_{obs})$. Since the "likelihood" under this setting is a "black box", we can get the posterior by using the ABC method to collect N posterior observations. Whenever the prior put mass $\frac{1}{r}$ on each x_i, the posterior will put mass p_i on x_i, where p_i is the proportion of the N posterior observations that gave the value x_i.

$$p_i \propto \sum_{s=1}^{N} \mathbb{I}_{x_s=x_i}$$

The ABC framework is formulated as follows. We draw x^*'s from $\pi(\tilde{X})$, and y^*'s will then be drawn from $f(\tilde{Y}_{obs}|x^*)$ using IQSIM. If the distance between y^* and y_{obs} is less than ϵ, then we accept x^* as a draw from $\pi(\tilde{X}|\tilde{Y}_{obs})$. Here the distance between y^* and y_{obs} needs to be defined. Let $m(y_{obs})$ be the $n \times n$ matrix of MHD [1] for y_{obs}. Similarly, let $m(y^*)$ be $n \times n$ matrix of MHD for y^*. Then, the distance from y_{obs} and y^* can be obtained as:

$$d(m(y_{obs}), m(y^*))$$

the distance between the two matrices. An empirical QQ-plot is employed to compare two matrices. $d(m(y_{obs}), m(y^*))$ is defined to be the sum absolute distance away from all of the points on the QQ-plot to quantify the distance.[2] To select ϵ, which corresponds to an acceptance rate, the distribution of distances under the model is needed. So, we need to draw many x's from $\pi(\tilde{X})$, then a $\tilde{y}$ given each x and create $m(\tilde{y})$. If a histogram of the distance $d(m(\tilde{y}), m(\tilde{y\prime}))$ for each pair of $\tilde{y}$ and $\tilde{y\prime}$ is created, then a specific quantile of the histogram of distance can be found and this corresponding value of distance can be used as ϵ.

3.2 Generalized framework

For a given dynamic sequence of images, it can be split time sequentially into r groups. Each group has approximately equal number of images. If we could find one training image in each group that can represent the corresponding group very well, then given

[1] Modified Hausdoff Distance, abbreviated throughout this work as MHD. It is a geometric measure to evaluate the closeness between spatial objects (Scheidt et al., 2016).

[2] The distance criterion is intended to be this variability matching criterion since that is the objective for the realistic prior specification. It can be changed to other measurements depending on a specific application context.

Algorithm 3: Proposed ABC Model

Input : Observed data $\tilde{Y}_{obs}$, quantile $\mathcal{Q}$, prior distribution $\pi(X)$
Output: An array V of N samples from the posterior $\pi(X|\tilde{Y}_{obs})$

$R \leftarrow \{\}; T \leftarrow \{\}$;
for $i \leftarrow 1$ **to** N **do**
    generate $X \sim \pi(X)$;
    drawn y^* from $f(\tilde{Y}_{obs}|X)$ using IQSIM;
    append X to T;
    append $m(y^*)$ to R;
$D \leftarrow \{\}; V \leftarrow \{\}$;
for $j \leftarrow 1$ **to** $N - 1$ **do**
    for k ←*j+1* **to** N **do**
        $l \leftarrow d(R[j], R[k])$;
        append l to D;
Create a histogram for all elements in D;
$\epsilon \leftarrow$ the distance at position $\mathcal{Q}$ of the histogram;
for $i \leftarrow 1$ **to** N **do**
    if $d(R[i], m(\tilde{y}_{obs})) \leqslant \epsilon$ **then** append $T[i]$ to V;
return V

these r training images, a set of representative images, MPS will be able to reproduce well the variability of the original dataset. Since the ultimate goal of this work is to provide a probability model on a set of representative images, if there are too many combinations of training images across different groups, the posterior distribution will end up being very flat. Therefore, we divide our time between *exploring* training images in each group that we think might be good with *exploiting* training images that we know are good. Specifically, in the exploring stage, we implement the proposed ABC framework with p samples. Each sample contains r training images that are randomly selected in each group. We can then obtain a histogram over the training images in each group based on the accepted samples.

In each group, the two training images with the highest counts will be selected. Now, a revised (informative) prior which is uniform on 2^r sets of r groups can be considered.

We make draws from this new prior and again run the proposed ABC framework to collect samples. With only 2^r possible values, we can get a posterior that differs from the prior and provide a probability model on sets of representative images.

Algorithm 4: Generalized Framework

Input : Observed data $\tilde{Y}_{obs}$, proposed ABC model $\mathcal{M}$, quantile $\mathcal{Q}$, group number r

Output: A probability model on a set of representative images

get $\pi(X)$ after dividing $\tilde{Y}_{obs}$ into r groups ;

$P \leftarrow$ run $\mathcal{M}$ based on the $\pi(X)$;

Create a histogram for all samples of P in terms of group;

$\pi'(X) \leftarrow$ pick the two training images with the highest counts from each group ;

$P' \leftarrow$ run $\mathcal{M}$ based on the $\pi'(X)$;

return probabilities for sets of training images in P'

4

Example and Results

4.1 The Experimental Delta Dataset

The data set used in this work came from the Fan-delta experiment that was performed in Tulane University Delta Basin with 4.2m long, 2.8m wide and 0.65m deep dimensionality. The experimental basin is used to describe autogenic organization and build physical stratigraphy of a delta experiencing constant external boundary conditions that include water supply, sediment discharge and subsidence rates. Since basin sedimentation is controlled by both internal dynamics of depositional systems and external boundary conditions (Wang et al., 2011), if boundary conditions are held constant, the intrinsic variability is therefore the only reason to affect the observed variability in deposition (Scheidt et al., 2016).

In addition, the topographic evolution for this experiment was recorded by time-lapse photos, 1 image every 1 minute. The total run hours were 100 hours, but in this work all analyses were based on 136 binary (wet-dry) images that represent two hours, during hour 80-82, of experimental run time. Fig. 4.1 shows some examples of the experiment.

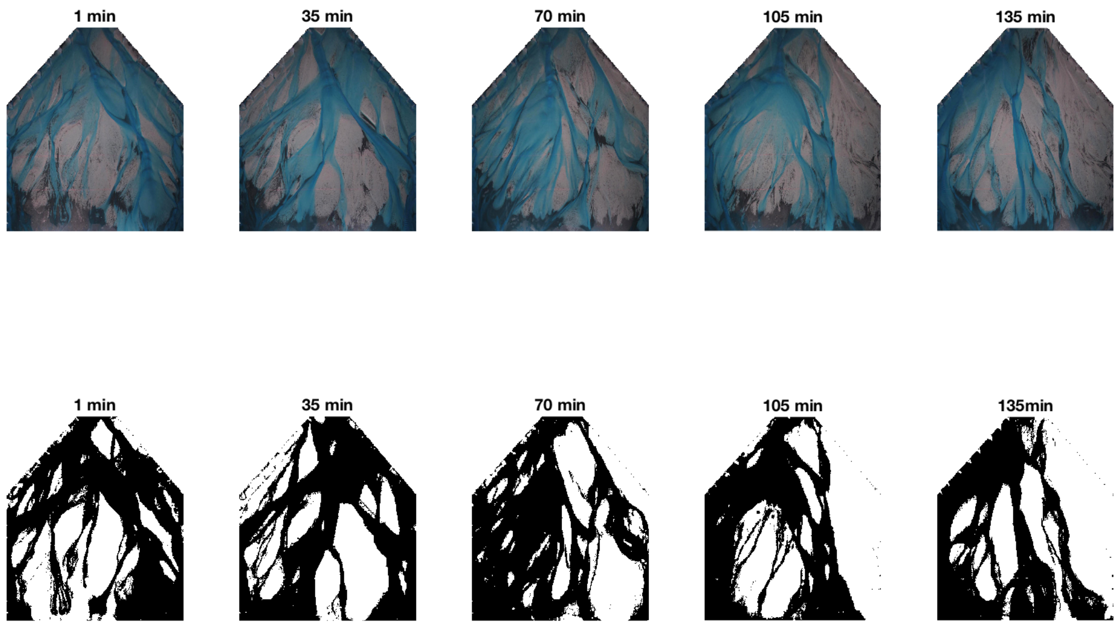

FIGURE 4.1: Example images from the Fan-delta experiment. Top: original experimental photograph; Bottom: binary representation of the images

4.2 Results

The whole dataset (136 binary snapshots) was evenly divided into 6 groups in time-sequential. Therefore, each group had approximate 23 images. In the exploring stage, we implemented the proposed ABC model with 15,000 samples (p). Each sample included 6 training images that were randomly selected from each group ($\pi(\tilde{X})$). ϵ was set to be the distance that was corresponding to the 0.2 quantile of the histogram of pairwise distances. It turned out totally 478 samples were accepted, and the histogram over the 23 training images in each group was then obtained based on these accepted samples. The histograms are presented in Fig. 4.2.

The two training images with highest counts were selected in each group, which are shown in Fig. 4.3. Then, a revised (or informative) prior which is uniform on 2^6 sets of six groups was considered. In this exploiting stage, we carried out the proposed ABC model with 30,000 samples and followed the generalized framework in order to obtain a posterior distribution, $\pi(\tilde{X}|\tilde{Y}_{obs})$. It is shown in Table 4.1.

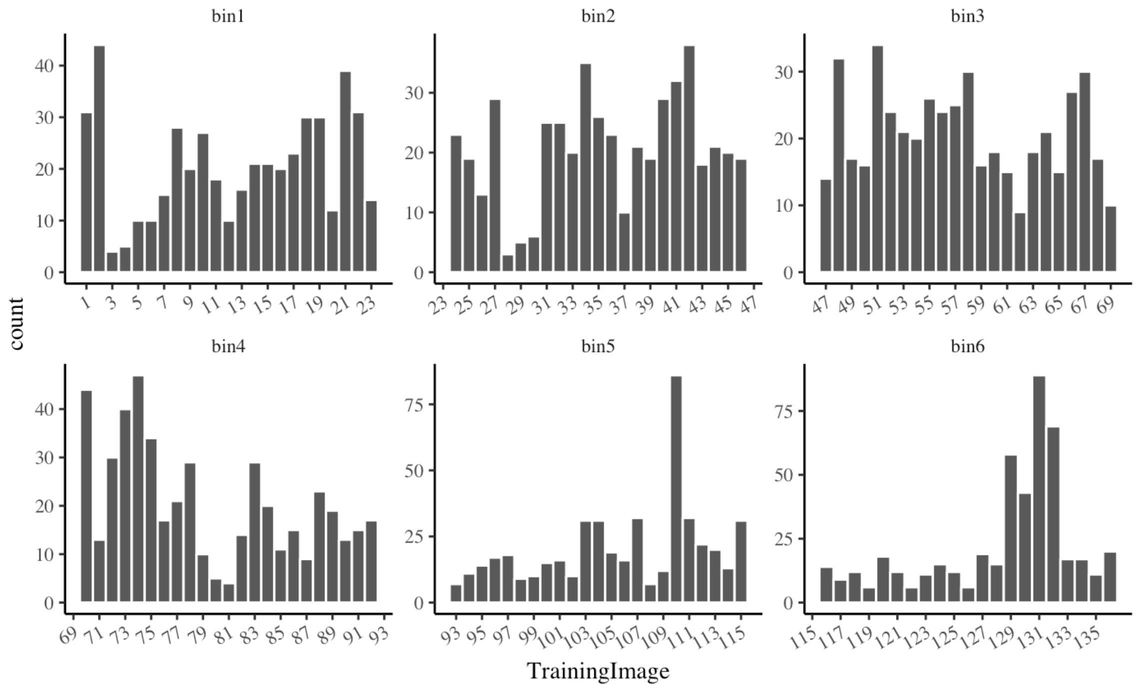

FIGURE 4.2: Histograms over all possible choices for 6 bins

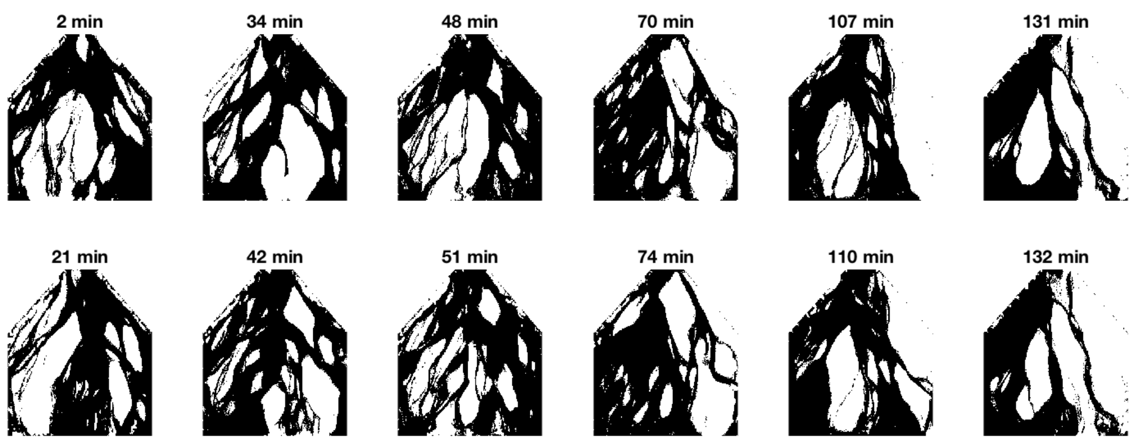

FIGURE 4.3: Selected TI based on the proposed ABC model for each group (column wise)

Table 4.1: Probabilities for all possible sets of representative images

#	Training Image Pair	Probability	#	Training Image Pair	Probability
1	2|34|48|70|107|131|	0.51%	15	21|34|51|70|107|129|	6.70%
2	21|42|48|74|107|129|	3.09%	16	21|34|51|70|110|131|	1.03%
3	21|42|51|74|107|131|	4.12%	17	2|34|51|70|107|129|	0.51%
4	21|42|48|70|110|131|	0.51%	18	21|42|51|70|107|129|	8.24%
5	21|34|51|70|107|131|	6.70%	19	21|34|51|70|110|129|	1.03%
6	2|42|48|70|107|131|	0.51%	20	21|34|51|74|107|131|	1.54%
7	2|42|51|70|107|129|	0.51%	21	21|42|51|70|107|131|	8.24%
8	21|42|48|70|107|129|	2.57%	22	21|42|48|70|110|129|	3.09%
9	21|34|48|74|107|129|	3.09%	23	21|42|48|74|107|131|	7.21%
10	21|34|48|74|107|131|	2.57%	24	21|42|51|70|110|131|	2.57%
11	2|42|51|70|107|131|	0.51%	25	21|34|48|70|110|131|	1.54%
12	21|34|48|70|107|129|	6.18%	26	21|42|51|70|110|129|	2.57%
13	21|34|51|74|107|129|	0.51%	27	21|42|51|74|107|129|	6.18%
14	21|42|48|70|107|131|	11.3%	28	21|34|48|70|107|131|	6.70%

Only 28 (out of $2^6 = 64$ possible) scenarios appeared in the posteriors (Table 4.1). In addition, we can get posterior predictive distribution, $f(\tilde{y}|y_{obs})$ indirectly using a Monte Carlo procedure: sample x from $\pi(\tilde{X}|y_{obs})$ and then draw $\tilde{y}$ from $f(\tilde{Y}|X)$. However, we focus on the question of whether the proposed framework can yield a proper selection of training images. Hopefully, MPS can effectively reproduce the variability of the observed data based on a high posterior set of training images. We calculated the MHD distances for both 136 images from the experiment and 136 generated MPS realizations given the high posterior sets of training images. Two empirical QQ-plots of them are presented in Fig. 4.4.

Suppose, instead, we use the TIs with zero posterior probability but still in this revised prior space. Then, the QQ-plots of MHD distances deviate from the diagonal red-dashed

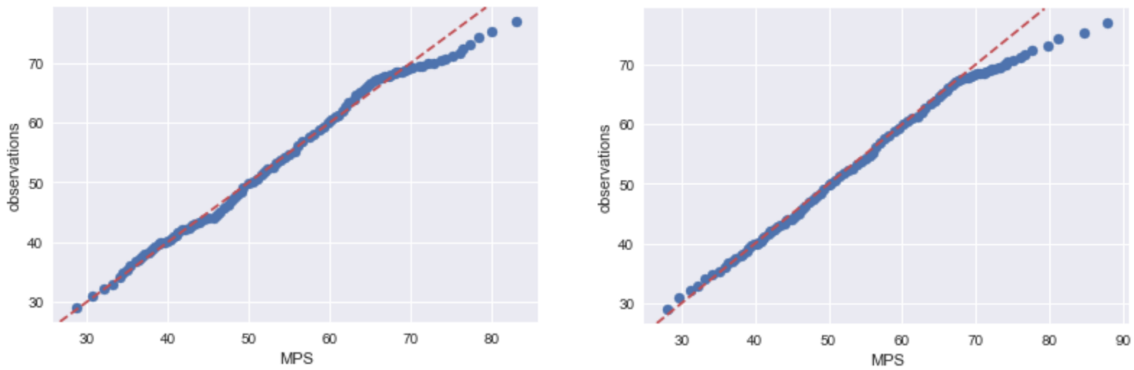

FIGURE 4.4: QQ-plot of the MHD distances of the observed data vs. MHD distances of the MPS realizations given the high posterior sets of TIs using IQSIM.

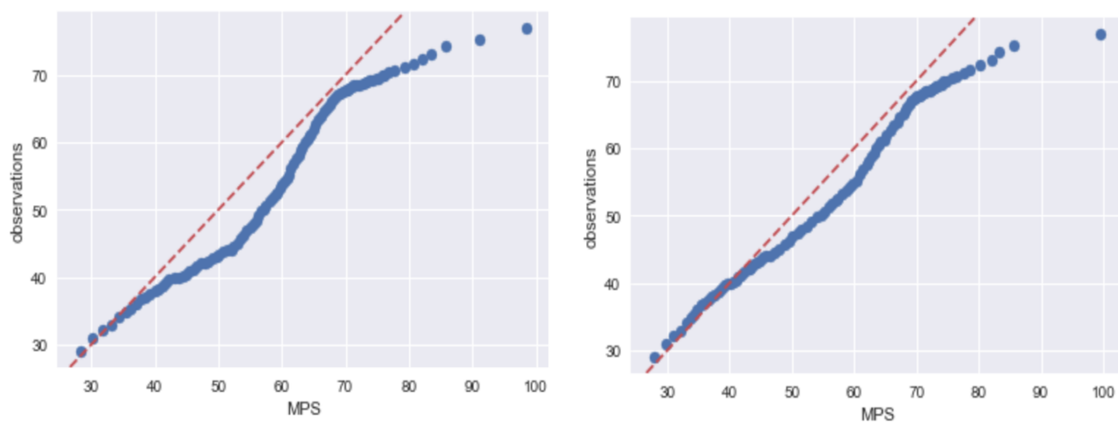

FIGURE 4.5: QQ-plot of the MHD distances of the observed data vs. MHD distances of the MPS realizations given the low posterior sets of TIs using IQSIM.

line far away. See Fig. 4.5. This suggests the MPS fails to reproduce the variability of the observed data given a low posterior set of training images. In other words, we expect that the distribution $\pi(d(m(\hat{y}), m(y_{obs}))|\hat{X})$ has a much smaller mean and is more informative than the distribution $\pi(d(m(\hat{y}), m(y_{obs}))|X_{null})$ by comparing their means and variances, where $\hat{X}$ refers to the posterior mode, $m(y)$ is the 136×136 MHD matrix associated with a given y, and X_{null} is a set of TIs with zero posterior probability. As shown in Fig. 4.6, this expectation is satisfied. It implies that the proposed ABC model is able to select a good set of TIs, and MPS can reproduce the variability of observed data based on this selection.

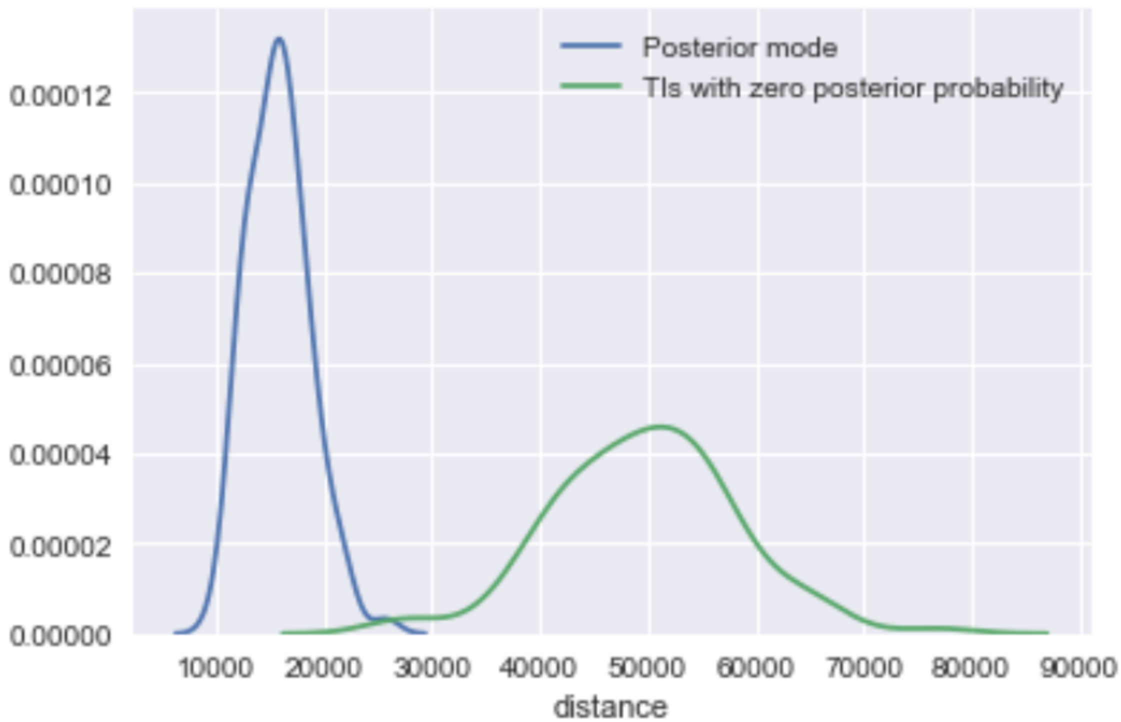

FIGURE 4.6: Comparing two distributions that are conditional on the highest posterior set of TIs and the set of TIs with zero posterior probability, respectively.

As a more general perspective, using a simulation framework, we want to know whether the observed data are plausible under a proposed model by comparing it with the samples that are generated from the model. In particular, we could find other statistics of interest for the data. Then, we could see whether or not a particular statistic calculated from the observations is in the tail of the distribution of this statistic under the model by comparing the value of this statistic with the values from each of the samples from the model. Here, for illustration, we choose the median distance of MHD matrix as the statistic of interest. By doing this Monte Carlo test, shown in Fig. 4.7, it turns out MPS does surprisingly well in representing the observed data on the basis of the highest posterior set of TIs that was selected by the proposed ABC model.

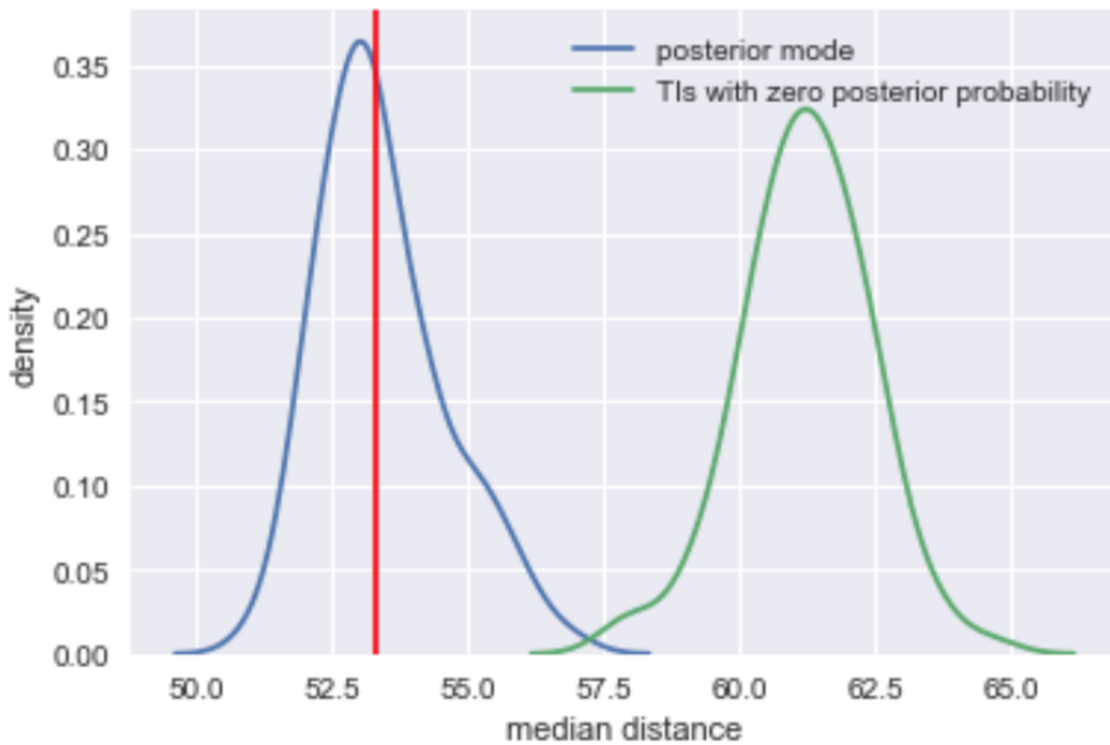

FIGURE 4.7: Monte Carlo test with median distance of MHD matrix as the statistic of interest. Red verticle line is the median distance for the flume data.

5

Sensitivities

As mentioned in the previous section, the true geometry mechanism is assumed to be *approximately* given by $f(\hat{Y}_{obs}|\tilde{X})\pi(\tilde{X})$. Therefore, the question arises: whether or not the number of TIs exerts an influence on the variability of the MPS realizations? To address this, we followed the generalized framework with $r = 8$ and obtained the posterior distribution under this new setting. Again, we chose the posterior mode, $\hat{X}_8$, and then got the distribution, $\pi(d(m(\hat{y}), m(y_{obs}))|\hat{X}_8)$. We compared this one with the distribution, $\pi(d(m(\hat{y}), m(y_{obs}))|\hat{X}_6)$ in terms of their means and variances, as shown in Fig. 5.1. MPS with 8 proper TIs performs better: realizations are closer to the observed data as measured by MHD but have a larger variance.

It is worth pointing out even though a larger number of TIs reults in a more accurate representation, it will make the realistic prior model be more complex. It is important for the model to be “simple" in order to generalize from the observations. Essentially, we are backing to the problem of a trade-off between bias and variance.

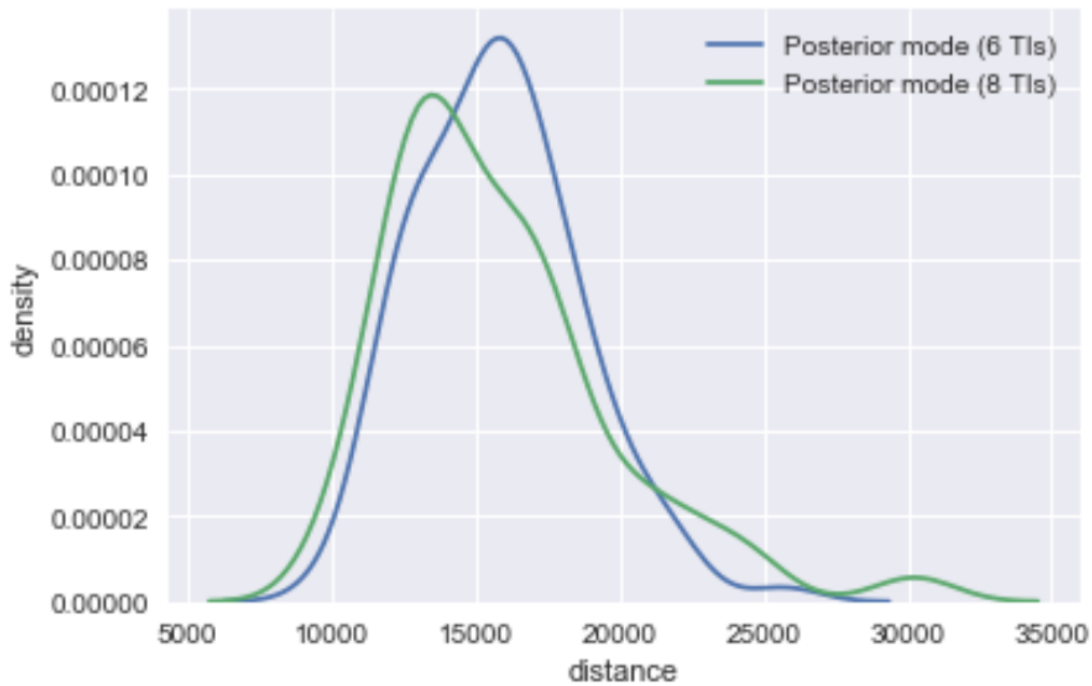

FIGURE 5.1: Sensitivitity check in terms of the number of TIs

6

Conclusion and Discussion

The proposed ABC model succeeds in selecting an appropriate representative images and providing a distribution model over it. The posterior distribution of this work will then be treated as the geological realistic prior model. We assess posterior distributions, $\pi(\tilde{X}|y_{obs})$, on the training images as well as predictive posterior distribution, $f(\tilde{y}|y_{obs})$, on the realizations, which allows a better understanding of generated geological realistic prior uncertainty. We believe that the proposed approach has a broader spectrum of application where a set of representative images is required from a larger list of objects with non-random variation in shape.

In addition, we use the ordered set of distances (QQ-plots) as the way of summarizing the pair-wise distance matrix, which may seem inappropriate from a statistical perspective since the observed images are viewed as a sequence over time. In other words, under the current setting, the temporal dependence in distance can not be considered. This is because the sampling procedure with image quilting that is used in IQSIM is blind to time: realizations generated from the simulation algorithm based on a single TI are exchangeable. Once we have a time-dependent process based model in the future, a natural improvement of this is to compute temporal dependence measures for both observed data and generated samples in order to make comparison.

All in all, a uniform prior is converted to an informative realistic prior which results from creating an ABC posterior. Essentially, this proposed ABC approach is a stochastic simulation method since model testing is verified via simulation to assess the plausibility of the observed data under the implemented model (Chen and Chen, 2017; Vasishth and Broe, 2010). In the forward simulation literature, most simulation algorithms look primarily for a single “good" set of inputs that meets a defined distance criterion over the input space. The approach that we proposed in this work puts a prior on a suitable subset of input space and converts it to a posterior providing a distribution over inputs. From this perspective, when applicable, it can be viewed as an alternative to more common stochastic search algorithms and can applied to the full range of forward simulation applications.

Bibliography

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian computation in population genetics," *Genetics*, 162, 2025–2035.

Blum, M. G. (2010), "Approximate Bayesian computation: a nonparametric perspective," *Journal of the American Statistical Association*, 105, 1178–1187.

Blum, M. G. and François, O. (2010), "Non-linear regression models for Approximate Bayesian Computation," *Statistics and Computing*, 20, 63–73.

Caers, J. (2011), *Modeling uncertainty in the earth sciences*, John Wiley & Sons.

Chen, D.-G. and Chen, J. D. (2017), *Monte-Carlo simulation-based statistical modeling*, Springer.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010), "Approximate Bayesian computation (ABC) in practice," *Trends in ecology & evolution*, 25, 410–418.

Didelot, X., Everitt, R. G., Johansen, A. M., Lawson, D. J., et al. (2011), "Likelihood-free estimation of model evidence," *Bayesian analysis*, 6, 49–76.

Dubuisson, M.-P. and Jain, A. K. (1994), "A modified Hausdorff distance for object matching," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, pp. 566–568, IEEE.

Emery, X. and Lantuéjoul, C. (2014), "Can a training image be a substitute for a random field model?" *Mathematical Geosciences*, 46, 133–147.

Fearnhead, P. and Prangle, D. (2010), "Semi-automatic approximate Bayesian computation," *Arxiv preprint arXiv*, 1004, 70.

Feyen, L. and Caers, J. (2006), "Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations," *Advances in Water Resources*, 29, 912–929.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F., Taly, J.-F., et al. (2009), "ABC likelihood-free methods for model choice in Gibbs random fields," *Bayesian Analysis*, 4, 317–335.

Hoff, P. D. (2009), *A first course in Bayesian statistical methods*, Springer Science & Business Media.

Hoffimann, J., Scheidt, C., Barfod, A., and Caers, J. (2017), "Stochastic simulation by image quilting of process-based geological models," *Computers & Geosciences*, 106, 18–32.

KIM, W., PETTER, A., STRAUB, K., and MOHRIG, D. (2014), "Investigating the autogenic process response to allogenic forcing," *From Depositional Systems to Sedimentary Successions on the Norwegian Continental Margin*, pp. 127–138.

Mariethoz, G. and Caers, J. (2014), *Multiple-point geostatistics: stochastic modeling with training images*, John Wiley & Sons.

Marin, J.-M. and Robert, C. P. (2009), "Importance sampling methods for Bayesian discrimination between embedded models," *arXiv preprint arXiv:0910.2325*.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012), "Approximate Bayesian computational methods," *Statistics and Computing*, 22, 1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003), "Markov chain Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences*, 100, 15324–15328.

Mohamed, L., Christie, M. A., Demyanov, V., et al. (2011), "History matching and uncertainty quantification: multiobjective particle swarm optimisation approach," in *SPE EUROPEC/EAGE annual conference and exhibition*, Society of Petroleum Engineers.

Olson, B. (2016), "Stochastic weather generation with approximate Bayesian computation," Ph.D. thesis, University of Colorado at Boulder.

Renard, P. and Mariethoz, G. (2014), "Special issue on 20 years of multiple-point statistics: part 1," .

Robert, C. (2007), *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media.

Robert, C. P. (2004), *Monte carlo methods*, Wiley Online Library.

Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. (2011), "Lack of confidence in approximate Bayesian computation model choice," *Proceedings of the National Academy of Sciences*, 108, 15112–15117.

Rubin, D. B. et al. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics*, 12, 1151–1172.

Scheidt, C., Fernandes, A. M., Paola, C., and Caers, J. (2016), "Quantifying natural delta variability using a multiple-point geostatistics prior uncertainty model," *Journal of Geophysical Research: Earth Surface*, 121, 1800–1818.

Scheidt, C., Li, L., and Caers, J. (2018), *Quantifying Uncertainty in Subsurface Systems*, Geophysical Monograph Series, Wiley.

Sheets, B., Hickson, T., and Paola, C. (2002), "Assembling the stratigraphic record: Depositional patterns and time-scales in an experimental alluvial basin," *Basin research*, 14, 287–301.

Strebelle, S. (2002), "Conditional simulation of complex geological structures using multiple-point statistics," *Mathematical geology*, 34, 1–21.

Tan, X., Tahmasebi, P., and Caers, J. (2014), "Comparing training-image based algorithms using an analysis of distance," *Mathematical Geosciences*, 46, 149–169.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997), "Inferring coalescence times from DNA sequence data," *Genetics*, 145, 505–518.

Turner, B. M. and Van Zandt, T. (2012), "A tutorial on approximate Bayesian computation," *Journal of Mathematical Psychology*, 56, 69–85.

Vasishth, S. and Broe, M. (2010), *The foundations of statistics: A simulation-based approach*, Springer Science & Business Media.

Wang, Y., Straub, K. M., and Hajek, E. A. (2011), "Scale-dependent compensational stacking: an estimate of autogenic time scales in channelized sedimentary deposits," *Geology*, 39, 811–814.