Structural geologic modeling as an inference problem: A Bayesian perspective

Miguel de la Varga¹ and J. Florian Wellmann²

Abstract

Structural geologic models are widely used to represent the spatial distribution of relevant geologic features. Several techniques exist to construct these models on the basis of different assumptions and different types of geologic observations. However, two problems are prevalent when constructing models: (1) observations and assumptions, and therefore also the constructed model, are subject to uncertainties and (2) additional information is often available, but it cannot be considered directly in the geologic modeling step — although it could be reduced to model uncertainties. The first problem has been addressed in recent work. Here we develop a conceptual approach to consider the second aspect: We combine uncertain prior information with geologically motivated likelihood functions in a Bayesian inference framework. The result is that we not only reduce uncertainties in the ensemble of generated models, but we also gain the potential to learn additional features about the model parameters. We develop an implementation of this concept in a probabilistic programming framework, in which we extend the functionality of a 3D implicit potential-field interpolation method with geologic likelihood functions. With schematic examples, we show how this combination leads to suites of models with reduced uncertainties and how it provides a deeper insight into parameter correlations. Furthermore, the integration into a hierarchical Bayesian model provides an insight into potential extensions of the method, for example, the interpolation functional itself, and other types of information, such as gravity or magnetic potential-field data. These aspects constitute promising paths for future research.

Introduction

Structural geologic models are constructed on the basis of observed (e.g., in boreholes or on outcrops) or inferred (e.g., from seismic data) geologic interface points and orientation measurements. Based on these data, relevant geologic surfaces and/or volumes can be modeled with a variety of methods (for recent overviews, see Caumon et al., 2009; Caumon, 2010; Jessell et al., 2014).

In this work, we consider geologic modeling methods based on interpolation functions (Mallet, 1992, 2004; Lajaunie et al., 1997; Carr et al., 2001; Hillier et al., 2014; Fouedjio et al., 2015), here referred to as a function of position \( \phi_k(x) \). Each of these methods has specific advantages and disadvantages, especially with respect to possible model automation and the types of data and geologic information that are incorporated in the modeling step, here considered as primary information \( k_j \). In addition to this primary information, each method requires the definition of additional parameters of the interpolation function \( \alpha_k \), for example, describing the variogram in a geostatistical interpolation method, as well as a topological description \( \beta_{ik} \), for example, representing the relationship between faults and surfaces. In combination, the process of obtaining a structural geologic model \( M \) at position \( \bar{x} \) can be described as a function \( f \) of all of these aspects:

\[
M = f(\bar{x}; \phi_i, k_j, \alpha_k, \beta_{ik}). \quad (1)
\]

The essential aspect here is that structural geologic modeling can be considered as a forward modeling step. So far, no modeling approach directly considers all types of data and geologic knowledge. It is common to construct a model, and then to evaluate manually or by visual inspection if a model is valid in the light of additional information, for example, with respect to expert knowledge about the geologic setting. For completeness, we refer to this additional information, which is not considered directly in the modeling step as auxiliary information \( y_m \).

In this work, we present a general approach that enables the consideration of additional information in...
geologic modeling, encapsulated in the form of likelihood functions. In essence, we do not consider geologic modeling as a forward modeling step, but as a Bayesian inference problem. We propose that this different viewpoint enables us to combine geologic modeling with additional information in a coherent manner, and that it provides a way to learn about parameter uncertainties and correlations as well as aspects of geologic and physical information.

Bayesian inference methods have been famous in the field of geophysics since the pioneering works of Tarantola and Valette (1982), Mosegaard and Tarantola (1995), and Sambridge and Mosegaard (2002). The motivation of our work presented here is the application of these concepts to geologic modeling. This approach is possible on the basis of recent developments in the field of structural geologic modeling, specifically the full automation of the modeling function $f$ for specific types of interpolation functions $\phi_i$ for complex 3D models, and the combination with the analysis of error propagation in geologic input parameters $k_i$ on the basis of Monte Carlo simulations (e.g., Jessell et al., 2010, 2014; Wellmann et al., 2010; Lindsay et al., 2012). We complement these approaches here with the consideration of the auxiliary geologic information $y_m$ in a nonparametric empirical Bayesian inference framework (Johns, 1957; Carlin and Louis, 1997), and we provide a full implementation of our method in a probabilistic programming environment with the open-source Python package PyMC (Patil et al., 2010).

In this paper, we first outline the general framework of Bayesian inference as a way to combine prior information with auxiliary information through likelihood functions, and explain how this concept relates to geologic data, information, and geologic modeling methods. As a motivation, we show a brief conceptual example of a 1D structural geologic inference case. We then show how the concept can be applied to an existing implicit geologic modeling technique, and how it enables us to consider additional information in the framework of full 3D structural geologic modeling with advanced implicit interpolation methods.

## Methods

### Bayesian inference

In recent years, Bayesian statistics increasingly have been applied as an effective tool to analyze complex systems with incomplete knowledge, mostly driven by successful developments in the fields of pattern recognition and machine learning (e.g., Louis, 1991; Vapnik and Vapnik, 1998; Challagulla et al., 2008), but also by many successful applications in geosciences (Tarantola and Valette, 1982; Mosegaard and Tarantola, 1995, 2002; Sambridge and Mosegaard, 2002; Tarantola, 2005; Guillen et al., 2008). The objective in Bayesian inference is to determine the posterior distribution $p(\theta | y)$ of a parameter set $\theta$, given prior distributions $p(\theta)$ and likelihood functions $p(y | \theta, \mathcal{M})$, which contain the information or observations $y$ related to the unknown parameters $\theta$. This consideration is encapsulated in the well-known Bayes equation:

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{\int p(y | \theta) p(\theta) d\theta}.$$  \hspace{1cm} (2)

Bayesian probabilistic models deem probabilities as degrees of belief. This means that our final goal is to quantify how certain our statement or model is given a specific amount of data. With this perspective, we specify the elements that make up Bayes’ theorem in the context of geologic modeling as follows.

1) **Mathematical forward model $\mathcal{M}$**: The mathematical model links parameters with observed data. Model selection is a choice (Matheron, 2012), and the complexity of the forward model needs to be balanced by available information and intended model objective; classical examples in this context are polynomial regression curves and the selection of a suitable polynomial degree. In the case of structural geologic modeling, these are typically interpolation functions (see below).

2) **Model parameters $\theta$**: These are the parameters that define the mathematical forward model $\phi_i$. A parameter can be either deterministic, if its value is exactly known, or stochastic, in the case that uncertainties about the exact value exist. In the latter case, these uncertainties are represented in the form of probability distributions. In our context, these can be any of the $\mathbf{\alpha}, k_i, \alpha_k$, and $\beta_e$ of equation 1.

3) **Observed data $y$**: These are any types of actual measurements or observations that can be compared with model outcomes. Typical examples are geophysical potential-field measurements (gravity and magnetics), but geologic observations can also be described in this form (see examples below).

4) **Likelihood functions $p(y | \theta, \mathcal{M})$**: These relate the previous components of model parameters and observed data; i.e., they define “the likelihood of the parameters given the data” (MacKay, 2003). In a mathematical sense, likelihood and probability function are essentially the same, except that the former is regarded as a function of the parameters $\theta$ and the latter as a function of the data $y$ (Patil et al., 2010).

Despite the apparent simplicity of Bayes’ axiom, the inference problem grows in complexity as more information is taken into account to define larger and larger uncertainty spaces. To outline the inference process, Gelman et al. (2014) propose the following sequence, here adapted to our specific scenario.

1) **Setting up a full probability model**: First, the complete joint probability space for the parameters is created. This can be seen as a multidimensional environment defined by every parameter $\theta$ that forms part of the model (for an example of a 2D space, see Figure 1a).
2) Conditioning on observed data: Afterward, the full prior probability model must be conditioned to the observations. The relation between parameters and data is given by a series of deterministic operations defined by our selected model $\mathcal{M}$. The result of this is evaluated against the likelihood function of the given observation. It is important to point out that not all the parameters necessarily must be related to all observations but any combination may be valid. Once all conditional probabilities are set, the Bayesian equation can be applied (equation 2). Given the likelihood of the parameters conditioned to the data $p(y|\theta, \mathcal{M})$, we calculate the posterior $p(\theta|y, \mathcal{M})$; i.e., we update our belief about the uncertain parameters $\theta$ in light of the information $y$.

3) In almost all practical cases, an analytical calculation of the denominator in equation 2 is not directly possible due to the intractability of the multidimensional space. This difficulty is partly responsible for the lack of wider use of Bayesian methods. In this paper, we apply the commonly used maximum a posteriori (MAP) methodology to obtain rapidly the most probable posteriors, as well as the broadly extended Markov chain Monte Carlo (MCMC) approaches to reach a full uncertainty analysis of the system (see the “Sampling methods” section).

4) Evaluation of the posterior model: The postprocessing analysis of the obtained posterior distribution depends on the intended use. We consider two different aspects: (1) the analysis of the posterior distributions of the parameters $\theta$ themselves with a Gaussian kernel density estimation (Scott and Sheather, 1985, as shown in Figure 1b) and (2) the analysis of the ensemble of generated models, applying measures of information entropy (Shannon, 1948), which we here use to analyze and visualize uncertainties in defined subsets of space, for example, in voxets of a 3D grid (see Wellmann and Regegnauer-Lieb, 2012; Wellmann, 2013).

**Sampling methods**

Up to this point, we describe the theoretical concept of treating geologic modeling as a Bayesian inference problem. However, the question is still how we obtain the posterior parameter distributions $p(\theta|y)$. As stated before, a closed-form solution only exists for simple and idealized problems. In practical cases, numerical methods are conveniently applied to obtain a solution. We describe here briefly two methods and how they can be useful in the context of geologic modeling: the MAP and the MCMC methods.

The choice of a suitable numerical method to obtain the posterior distribution depends on the requirements of the model. In the context of geologic modeling, it may be desirable to obtain a single model outcome that can be assumed to best match prior parameter distributions and the geologic likelihood functions. In such a case, the MAP method may be sufficient. However, we are often interested in determining a range of valid geologic models, as obtained with Monte Carlo methods (Hastings, 1970).

The successful use of the original Monte Carlo acceptance-rejection method (i.e., random sampling of the parametric space) to explore the target distribution depends highly on the posterior shape and the number of parameters. Probabilistic areas with long tails complicate the exploration of the parametric space, leading to high rates of rejection and therefore a drop in efficiency of the method. To improve the sampling process, a Markov chain scheme is adopted (Gilks, 2005). A simple analogy between the two approaches is playing golf and dropping the ball randomly in the field or hitting the ball sequentially with closed eyes from the last location.

The advantage of MCMC approaches is not only that they perform a more global search for an optimal solution but also that they generate rich sets of information that can be used for postprocessing analysis and the evaluation of model uncertainty. For completeness, both methods are briefly described below.

Figure 1. Example to visualize basic Bayesian concepts: (a) 2D joint probability space generated after two random parameters $\theta$ and (b) complementary methods for posterior representation. (1) Kernel entropy, (2) histogram, and (3) information entropy.
MAP method

The uncertainty space is proportional to priors ($\theta$) and likelihood $p(y|\theta, \mathcal{M})$:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \quad (3)$$

This means that to infer the most probable mode of our posterior, it is not necessary to explore the entire uncertainty space, as we only consider the maximum value and the evidence in the denominator can be neglected because it is constant given $y$, and therefore does not influence the maximum estimate:

$$\hat{\theta}_{\text{MAP}}(y) = \arg \max_{\theta} \left[ \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} \right] = \arg \max_{\theta} \left[ p(y|\theta)p(\theta) \right]. \quad (4)$$

MAP methods renounce the calculation of the normalization factor (i.e., the denominator in Bayes’ formula, equation 2) in favor of a faster simulation. There are several methods to obtain the MAP point, depending on the complexity of the statistical model. In this paper, the Nelder and Mead (1965) algorithm is used to perform the MAP optimization. It is based on a modification of the expectation-maximization algorithm, which approximates iteratively the desired target, i.e., the MAP probability.

In Bayesian statistics, the result is the whole posterior distribution over the model parameters. It is important to note that the solution of the MAP method only represents one solution without providing its associated probability. Depending on the size of the posterior, this may or may not lead to a sufficient representation of reality and can potentially provide a misleading result.

MCMC

In the previous sections, we have explained Bayes’ intuition in terms of the generation of a multidimensional space and its subsequent exploration. Following the analogy, to perform a successful exploration of the space, we face two major complications: (1) finding areas of high probability and (2) sampling these areas sufficiently to infer an accurate probability density.

Monte Carlo methods have been used extensively to perform difficult integration problems computationally (e.g., Boyle, 1977; Ljungberg et al., 2012; Rubinstein and Kroese, 2013). They are suitable for problems with high dimensionality because the method is not bounded by mathematical tractability.

All Monte Carlo processes are based on two concepts: (1) drawing a random sample from the target distribution $\theta$ and (2) the acceptance or rejection of this random sample according to a weight defined by the scaled up candidate density of the value. The variation of these two concepts derives on the different sampling methods.

Additionally, for large multidimensional spaces, a Markov chain approach is recommended. Its mathematical foundation is based on conditioning the future state $X_{t+1}$ uniquely on the current state $X_t$, so that

$$\Pr(X_{t+1} = x_{t+1}|X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \Pr(X_{t+1} = x_{t+1}|X_t = x_t). \quad (5)$$

The development of random MCMC walks has been a major field of research more than 50 years (Metropolis and Gibbs samplers; Metropolis et al., 1953; Geman and Geman, 1984) and still today experiences substantial progress (Hamiltonian and the no-u-turn sampler; Jasche and Kitaura, 2010; Homan and Gelman, 2014). For the realization of the case study, we make use of an adaptive Metropolis sampling method (Haario et al., 2001), which combines the traditional Metropolis-Hastings, with the capacity to tune step sizes continuously while approaching the target distribution.

Metropolis-Hastings methods are based on the creation of a candidate probability distribution $q(\theta|\theta^0)$. This distribution will determine the proposed draw of a sample at every iteration $t$. Within a Markov chain framework, candidates depend only on the previous iteration $\theta^{(t)}$ and an arbitrary value $\epsilon_t$, drawn according to the probability density of the candidate:

$$q(\theta|\theta^0) = p(\theta - \theta^0) = \theta^0 + \epsilon_t. \quad (6)$$

Usually, $\epsilon_t$ is a normal or student density, symmetric around zero. Once we generate the candidate $\theta'$, we need to define an acceptance ratio, which under these circumstances is given by equation 7:

$$a(\theta', \theta) = \frac{p(\theta')p(y|\theta')}{p(\theta)p(y|\theta)} = \pi(\theta') \pi(\theta), \quad (7)$$

where $\pi$ is the proportional posterior (equation 3). This acceptance ratio is then compared with a random value $u$ from a uniform distribution $U(0,1)$, following equation 8 to assure that this kernel transition to higher probability densities is not always enforced, thus enabling a better exploration of the whole probability space. This moves the chain state to high-density points in the distribution with high probability, and to low-density points with low probability:

$$\theta^{(t+1)} = \begin{cases} \theta' & \text{if } a(\theta', \theta) > U(0,1), \\ \theta^{(t)} & \text{otherwise}. \end{cases} \quad (8)$$

The last concept to take into consideration for a full understanding of the Metropolis methods is the choice of a suitable scale factor for a step size to determine a new candidate distribution. Large steps are highly exploratory, which decreases the acceptance rates, whereas smaller factors on the other side lead to higher acceptance rates but worse convergence and chain mixture speed.
The adaptive Metropolis-Hastings method that we use here generates a covariance matrix that is updated at every iteration, which gives as a result a gradual adaptation of the sampling size. Regardless of the random walk used to perform the MCMC methods, convergence diagnosis should be performed to at least assure the appropriate examination of uncertainty areas. In most cases, a common ad hoc method of plotting the posterior traces of different chains initialized in distant spatial points is used to ascertain that they show an asymptotic behavior over the last iterations. The plot of parameter traces is provided in supplementary Figure 1, which can be accessed online at s1.pdf. Additionally, there are some formal convergence diagnostics that can be calculated to obtain an analytical estimation. For the case we discuss, the time-series approach proposed by Geweke (1991) has been used:

\[ z = \frac{\bar{\theta}_a - \bar{\theta}_b}{\sqrt{\text{Var}(\theta_a) + \text{Var}(\theta_b)}}. \] (9)

This method compares early intervals \( a \) with late intervals \( b \), assigning a score \( z \) along the chain. Convergence of the chain should lead to \( z \) values that fall within \(-2 \) and \( 2 \). The convergence analysis using the Geweke measure is presented in supplementary Figure 2, which can be accessed online at s2.pdf.

**Structural geologic forward modeling**

A central step to our method is the application of a suitable geologic modeling method to perform the forward modeling step \( f \) in equation 1. A definite requirement for the implementation in a probabilistic framework is the choice of a suitable interpolation method \( \phi_i \), which enables a full automation of this modeling step, so that a model can be updated when relevant input parameters are changed. Implicit geologic modeling methods (e.g., Lajaunie et al., 1997; Mallet, 2004; Calcagno et al., 2008; Hillier et al., 2014; Mejía-Herrera et al., 2014) are well suited for this task, and developments in recent years have shown how they can be applied to automatic model reconstruction for the evaluation of error propagation (Cherpeau et al., 2010; Jessell et al., 2010; Wellmann et al., 2010; Lindsay et al., 2012).

As in our own previous work (Wellmann et al., 2010), we apply here an implicit geologic modeling method (Lajaunie et al., 1997; Calcagno et al., 2008), implemented in the software GeoModeller. This method is based on a global interpolation of a scalar field (denoted as potential field) \( T(x) \) for each topologically consistent sequence of geologic layers, where surfaces between geologic layers are represented as isovalue surfaces of this field for a value \( t_k \). For the interpolation of the field, a cokriging method is applied that considers: (1) contact points on geologic surfaces through increments of the potential field \( T(x_1) - T(x_2) \) and (2) orientation measurements anywhere in the geologic sequence as the partial derivative of the field \( \partial T(x)/\partial u_\beta \) in each direction \( u \). The relative contribution of increments and derivatives is defined by weights \( \mu_\alpha \) and \( \nu_\beta \). Combined, for a total number of \( M \) data points and \( N \) partial derivatives and an arbitrary origin \( x_0 \), the estimator then takes the following form (Calcagno et al., 2008):

\[
T(\bar{x}) - T(x_0) = \sum_{a=1}^{M} \mu_\alpha (T(\bar{x}_a) - T(x_0)) + \sum_{\beta=1}^{N} \nu_\beta \frac{\partial T}{\partial u_\beta}(\bar{x}_\beta),
\] (10)

where \( T \) is a random function with polynomial drift and a stationary covariance \( K(h) \). Practical considerations have shown that a cubic covariance model is a suitable choice (Aug, 2004; Chiles et al., 2004).

Fault surfaces are interpolated in the same way, and interactions between faults are described with fault-fault network relationships, and the influence of a fault on the geologic layers is accounted for by discontinuous drift functions in the cokriging system resulting in a discontinuous potential field. Finally, several types of interactions between consistent geologic layer sets, described with different potential fields, can be defined (for details, see Calcagno et al., 2008).

Following the general description of the modeling method in equation 1, the observations of surface contact and orientation measurements are the primary information \( k_j \). Additional parameters of the interpolation function \( a_\theta \) affect the drift function, as well as the covariance model. Finally, the \( \beta_\beta \) describe, in our case, the compartmentalization into blocks between fault zones with a consistent potential-field interpolation.

The applied method has originally been developed for geologic modeling in complex regions with sparse data, and as a method to efficiently test different geologic hypotheses (Calcagno et al., 2008). However, although the interpolation already allows direct integration of several geologic data types, additional types of observations (e.g., expected layer thickness), or additional aspects of geologic knowledge (e.g., about the type of a fault), are not taken into account. With our approach, we enable the consideration of these aspects, as we integrate them into a quantified evaluation of different hypotheses in the framework of Bayesian inference.

**Numerical implementation**

We implemented the theoretical concept described above by integrating the implicit geologic modeling into the probabilistic programming framework of PyMC 2. This package is a full-featured library to perform probabilistic programming with the programming language Python (for more information about scientific programming with Python, see, e.g., Langtangen, 2008). PyMC 2 provides numerous sampling methods and other analytical and estimation tools such as MAP. However, the major strength of the package is the hierarchical
model building framework. Variables of the probabilistic model (parameters, as well as observed data) are described as nodes in a Bayesian network (for detailed descriptions of Bayesian networks, see, Koller and Friedman, 2009). The relationship between nodes is described with parent and child nodes, where:

1) Parent nodes contain variables that influence other variables.
2) Child nodes contain variables that are affected by other variables; i.e., they are dependent on parent variables.

The elements that form the statistical model are represented in PyMC 2 by two types of variables, stochastic and deterministic. Deterministic variables are defined by mathematical functions that return a value given the parents values. These types of variables allow us to create complex mathematical relations between the stochastic variables. Following this scheme, a hierarchical graphical representation of the probabilistic model can be generated (see the “Results” section for an example). Stochastic variables, on the other hand, are used to either describe uncertain parameters $\theta$ or likelihood functions $p(y|\theta)$.

PyMC is fully object-oriented, and can be extended with its own object definitions inherited from the deterministic and stochastic class descriptions. In the case of structural geologic modeling as considered in this work, input parameters $\theta$ to the geologic modeling are described as stochastic variables, and the interpolation method itself is implemented as a deterministic function $\phi_i$, as described in equation 1.

The functionality of the interpolation technique that we use (implemented in the software GeoModeller) can be accessed through an automated programming interface. For simpler access and direct coupling with PyMC, we wrapped all relevant functions into Python functions, combined in the package pygeomod. See Appendix A for code availability.

Bayesian hierarchical network

As we have seen above, the incorporation of different sources of data into a single Bayesian network is done through the use of likelihood functions $p(y|\theta, M)$, which describe the certainty of the data. Therefore, the use of several likelihood functions leads to most scenarios, to several models $M_i$, which generate a hierarchical structure, where a certain number of parameters $\theta$ are related to one or several observed data sets $y$ through functions $f$ (e.g., the structural geologic modeling equation 1).

Below, we enumerate some of the aspects to consider in the geologic modeling and their specific function within the Bayesian network.

1) Surface contact points and orientation measurements can be used either as informative prior parameters $\theta$ or as observations $y$ (i.e., as an outcome of the modeling process). In an ideal case, the use of all of these elements as priors allows one to infer the most possible knowledge. However, due to computational issues for very high-dimensional conditions, it can be beneficial to reduce the number of priors. Doing so, we decrease the amount of knowledge inferred but not the information of the system or the accuracy of the calculation.

2) Geologic concepts can be integrated in the Bayesian model as likelihood functions, which constrain the probability of geologic features. Geologic concepts can be subdivided into two categories of models $M$: (1) those that directly affect a subset of the parameters and do not need the geologic model and (2) constraints related to the geometry of the model that require the calculation of the geologic interpolation (equation 1).

3) Additional geologic observations and measurements: The geostatistical interpolation method $\phi_i$ itself prescribes which type of geologic data ($k_j$ in equation 1) can be considered. However, additional geologic information is commonly available. For example, in the method applied in our case, we can only consider observations of contacts between two different geologic units, but not the observation of a unit itself — and this observation is often made in an outcrop. Further examples could be additional structural geologic features, such as fold axes or lineations. In the Bayesian framework, we can consider this type of observation in the form of a likelihood function, with the additional benefit of potentially assigning uncertainties to the observation.

4) Geophysical measurements: In addition to geologic observations, geophysical measurements can be integrated in the framework if a geophysical forward model can be constructed on the basis of the geologic model realization and additional rock-physical parameters. Classical examples are geophysical potential-field measurements of gravity and magnetics (Guillen et al., 2008; not shown here but to be explored in future work, see the “Discussion” section).

5) Interpolation parameters ($\alpha_k$ in equation 1) also can be considered as uncertain parameters $\theta$, with assigned probability distributions. Furthermore, it would potentially be possible to consider the interpolation function $\phi_i$ itself as uncertain, and to permit different types of interpolation functions in this framework (see the “Discussion” section).

Bayesian analysis for geologic modeling

We have seen the individual components that we combine into a probabilistic Bayesian inference framework. The important aspect to consider is that geologic data are defined in a large multidimensional space, and that a natural spatial correlation exists. Therefore, the selection of the interpolation method $\phi_i$ represents the core of the Bayesian network and establishes which type of observed data $y$ we can use.
We use here cokriging interpolation as the spatial correlation model. This choice carries with it the following considerations.

1) In general, we can assume that more data points will lead to better estimates at unsampled locations and to an improved understanding of the model, so that the number of parameters $\theta$ for a completely accurate description tends to infinity. This intuition is encapsulated in nonparametric Bayesian models. This raises the question of which model ($\phi_i$) and model parameters ($k_j, \alpha_k, \beta_\ell$) we must select to honor the data (see the “Discussion” section).

2) Because we describe a large uncertainty space with sparse data, we must make use of highly informative priors. The priors are ideally based on observation statistics or statistical models for the measurements. This dual perspective enriches the process and facilitates the description of complex models, and is often described as empirical Bayes.

We describe above the general considerations for the definition of structural geologic modeling as an inference problem. In the following, we show the application of these concepts to a simple and intuitive example, as well as the extension to a full 3D geologic model construction on the basis of the implicit potential-field approach.

**Results**

**Schematic example for the combination of information in geologic modeling**

We first present a simple schematic model to show how it is conceptually possible to combine a structural geologic modeling step with additional information. We consider a model of parallel geologic layers at depth and assume a simple modeling function $M$ that constructs a model on the basis of the distance of the layer positions (Figure 2a). The considered model has two parameters and we assume that the position of the

![Figure 2](attachment:image.png)

Figure 2. Schematic example of combined information in the context of geologic modeling: (a) prior distribution for the parameters identifying the depth of two layers, lower layer (sediment 2) with high uncertainty; (b) the likelihood function adding additional information about the expected thickness of the layer and yields, (c) combined with the prior distributions, the posterior distributions. It is visible how the uncertainty of the lower layer (sediment 2) is now reduced. In addition, the two parameters sediments (1 and 2) are now clearly correlated, as visible in the joint distribution plots (d and e).
top layer (sediment 1) can be estimated with a relatively high accuracy, but a second layer below it (sediment 2), at greater depth, is highly uncertain (e.g., because it is not well resolved by seismic data). We now assign probability distributions to the layer positions, reflecting the uncertainty about the exact position, and these take the form of our prior parameter distributions.

This scenario is depicted in Figure 2a, including the probability distributions to reflect parameter uncertainties. Samples from these distributions represent probable geologic models, and the lines then identify the top layers of two geologic units, sediments 1 and 2.

We now assume that we have additional information about the thickness of sediment 2. For example, we may have additional observations in the vicinity that lead to the assumption that this geologic layer has a thickness of approximately 180 m. Note that this information is not a part of the forward model, as it cannot be used as a parameter directly in the modeling step (equation 1), but as auxiliary information. We therefore encode it here in the form of a likelihood function and represent the uncertainty about this observation again in the form of a probability distribution, in this case, a normal distribution with a mean of 180 m and a standard deviation of 20 m (Figure 2b).

As a next step, we combine the prior information about the layer depth with the likelihood of the layer thickness and we perform an MCMC sampling from the posterior distribution to obtain possible layer positions and, given many samples, a histogram that approximates the posterior parameter distribution (Figure 2c). As expected, the additional information about the layer thickness leads to a reduced uncertainty in the geologic model.

In addition to performing a MCMC sampling of the posterior distribution, we can also produce an estimate of the MAP solution for this example. This method provides a single estimate for the geologic model, which combines the prior distribution and the additional information of the layer thickness. This point estimate is represented in Figure 2c with the dashed lines. It is obvious that this single estimate provides less information than the multiple models generated with the MCMC method, but that it can be a useful measure if a single “best-fit” model is required.

In addition to the analysis of the posterior distributions for single parameters, we also observe an interesting point in the joint probability distributions. Before performing the Bayesian inference step, the parameters were assumed to be independent, and this is represented in the joint probability distribution (Figure 2d). However, after adding the information about the layer thickness, the two parameters are highly correlated (Figure 2e). This result is in this case trivial, but it shows that we can learn additional features about the model parameters (the correlation of the layer positions) through the combination with the auxiliary information (the layer thickness).

**Application to a 3D geologic modeling study**

**Model set-up**

We proceed to a more complex geologic scenario and to the integration of the complete implicit geologic modeling technique as a more sophisticated interpolation method $\phi_i$ into our analysis. Consider the following scenario: We aim to create a 3D geologic model of a graben structure, with two normal faults offsetting a sedimentary sequence of four subparallel geologic layers (Figure 3a, note that the uppermost layer is transparent). This model is conceptually similar to the input model of the study in Wellmann et al. (2014). The major difference is that here we define geologic likelihood functions, and we implement the entire modeling scheme in the Bayesian inference framework described above.

The geologic parameters that we use to define the geologic model are shown in Figure 3b, and values are given in Appendix A (Table A-1). All sedimentary sequences are defined in a single geologic series, and therefore modeled with the same potential field. The two faults affect the entire sequence, and they are modeled as infinite surfaces. For the model interpolation step itself, we apply the standard settings in GeoModeler (isotropic covariance function, very low nugget effects for orientation data and surface contact points, drift of degree 1).

**Monte Carlo error propagation**

As a first step, we simply evaluate how parameter uncertainties propagate to the final geologic model through a Monte Carlo sampling method of the prior parameter space. We assume that all parameters are independent and follow a normal distribution, around a mean value identical to the initial data point, with increasing standard deviations for deeper layers (see Table A-1).

This error propagation step is conceptually identical to results shown previously (e.g., Jessell et al., 2010; Wellmann et al., 2010; Lindsay et al., 2012). In this sampling step, we generate a range of 600 geologic models. However, the representation of results for multiple 3D geologic models is not as easily possible as in the 1D case above (prior models in Figure 2a). We therefore use here the measure of information entropy (Shannon, 1948) to visualize spatial uncertainties (for more information on the application in this specific context, see Wellmann and Regenauer-Lieb, 2012; Wellmann, 2013). A 2D cross-section view of information entropy is given in Figure 3c. We observe that, as expected, uncertainties are highest where the effects of uncertain fault positions and uncertain layer depths intersect. The high values of information entropy ($H_{max} \approx 1.95 \approx \log_2(3) \approx 1.58$) show that, at least, four lithologic outcomes are possible in these regions of highest uncertainties.

**Bayesian inference with integration of additional information**

We now assume that additional information is available on the basis of (1) additional observations in the vicinity of the model and (2) the expected geologic setting in general. Simple mathematical relations between
Figure 3. Three-dimensional geologic modeling study: (a) Representation of model, consisting of two normal faults in a graben structure; all of the following representations are in an east–west section through the center of the model; (b) stochastic geologic model parameters and indicated uncertainties; (c) information entropy of prior model; (d) geologic likelihood concepts; (e) information entropy of posterior model ensemble; and (f) posterior analysis for parameter pairs, showing now a clear correlation.
the relative position of the parameters can describe the aforementioned auxiliary information in form of three models $\mathcal{M}$, embedded in a Bayesian network. The constraints are as follows (Figure 3d):

1) According to the conceptual geologic setting of a simple graben (without later reactivation), we would expect that the faults are normal faults.
2) The thickness of the geologic layers may be expected to fall within a certain range.
3) We may assume a certain offset on the fault.

We consider this additional information in the form of likelihood functions (Tables A-2 and A-3), which are added to the probabilistic model, and sampling is now performed with the MCMC approach described previously. This sampling method results now in an ensemble of models that considers the additional geologic information.

Analysis of generated models and posterior distributions

First, we analyze uncertainties in the ensemble of constrained models, as before using the measure of information entropy (Figure 3e). We observe that uncertainties in the resulting model are now clearly reduced. The reduction is especially visible in the range of the second layer. Uncertainties are mainly remaining around the expected positions of the layer interfaces, and these are clearly constrained by the additional information of the layer thickness. However, the maximum cell information entropy is still high ($H_{\text{max}} \approx 1.88$), indicating that high uncertainties partially remain in the model.

In addition to the generated models, posterior distributions provide information about the input parameters. As an example, Figure 3f shows clear correlations between the posterior joint distributions for three pairs of parameters. The positive correlation between sediment 1 center and sediment 2 center, as well as between sediment 1 center and basement center, is obvious. It is based on the consideration of layer thicknesses in the form of likelihood functions (similar to the previous example), and a deeper position of one surface therefore correlates to a deeper position of the other interface. Interesting also is the positive correlation between surface positions in the horsts and the graben, here shown for parameters sediment 2 left and sediment 2 center. This correlation is due to the definition of the fault offset in the form of likelihood functions. The derived correlations support the expected behavior in this simple case, but note that we learned this additional aspect about input parameters, as the prior distributions were again considered uncorrelated. This is an example of how correlation, clustering, and other postprocessing techniques can improve our understanding of geologic parameters and possibly also of the overall geologic setting.

---

**Figure 4.** Hierarchical Bayes’ model of 3D geologic model study. Ellipses represent our stochastic nodes: parameters probability density in light gray and likelihoods in dark gray. Triangles illustrate the deterministic values. Arrows point from parents to children ending up at observations $y$. 

---
**Exploration of hierarchical structure**

Probability density of parameters and likelihood functions are usually only conditioned to a limited set of neighbor parameters, and not to all other parameters in a model. This realization leads to the possibility of visualizing a probabilistic model in the form of a Bayesian network where every node in the model is influenced by its parents and in turn impacts on its children in a sequential manner for every model $M_i$ that form the Bayesian network. Figure 4 displays the hierarchical relations of our case study.

The hierarchical model is composed of 17 stochastic parameters $k_j$. The first set of parameters determine the interface depths $z$ of the respective layers (sediment 2, sediment 1, and basement and faults) and the second the orientation of the faults (azimuths and dips; Figure 4, light gray ellipses).

Thickness layers 1 and 2 (Figure 4, triangles) are deterministic nodes that derive from the difference of the interface values. Thickness layer 1: Likelihood and Thickness layer 2: Likelihood influence these deterministic nodes and subsequently all their respective children. Offset: Likelihood determines the amount of fault offset.

Finally, to honor the graben shape, offset below faults likelihood and offset negative likelihood are used (Figure 4, dark gray ellipses). The former ensures that the intersection of both faults remains above the model boundary, whereas the latter guarantees the downward direction of the offset.

Following our hierarchical scheme, the last step is the generation of the implicit geologic model. This procedure, from a Bayesian inference point, can be seen as a deterministic child of the stochastic parameters, as a single model realization is completely defined by the input parameter set. At this point, the conditional tree could be extended to further observations (Figure 4, additional observation). These likelihood functions $p(y|\theta, M)$ could capture all types of observations that can only be included on the basis of the final calculated model, for example, observations of rock types away from surface contacts (off-surface constraints). These additional types of information are described in the “Discussion” section.

**Discussion**

The results of the synthetic modeling studies show that it is indeed possible to consider geologic modeling as a Bayesian inference problem. We applied the concepts not only to a simple, conceptual case, but also to a full 3D implicit structural modeling method, implemented in a commercial software package (GeoModeler). The potential of the method derives from the fact that it provides a highly flexible way to consider parameter uncertainties, as well as a wide range of additional observations and measurements. The only basic requirement is that these additional aspects can be described in the form of likelihood functions, and that reasonable uncertainty ranges can be assigned. We showed how these aspects can be integrated into an existing probabilistic modeling framework (PyMC 2). This integration enables a flexible and powerful method for sampling from the posterior distribution, resulting not only in a suite of generated models that respect the information provided in the form of likelihood functions, but also in a rich data set that can be used to learn aspects of the considered parameters. By reducing the interpretation of the measurements, we are able to preserve large amounts of data. This facilitates the use of data mining and supervised machine learning approaches to improve the understanding of the interaction of different types of data. Overall, the consideration of structural modeling as an inference method provides, in our point of view, a valuable way toward a more complete consideration of knowledge in the context of structural geologic modeling.

We showed the application of our method to two simple model studies, in which we considered the parameters describing the depth of surface contacts as uncertain. In reference to our general modeling function in equation 1, these parameters belong to type $k_j$. A further step would be the inclusion of interpolation parameters $\alpha_i$. Depending on the mathematical interpolation model, these parameters can take different forms. In our case, using the implicit potential-field interpolation method based on cokriging, parameters could include anisotropies of the covariance model, or nugget effects of geologic parameters. A further extension of this approach could be to actually consider the interpolation method, the $\phi_i$ in equation 1, as a random variable, in the form of hyperparameters, which influence the choice of a particular interpolation functional. Wood and Curtis (2004) also introduce the idea of using hyperparameters to classify distinct geologic setting. Our framework may be used to expand this concept to more complex 3D scenarios. Furthermore, the topological description $\beta_i$, for example, in the form of fault interaction models, could be treated as uncertain. The investigation of the effect of these different types of uncertainties provides an interesting pathway for future research.

One major strength of the approach is its flexibility and the wide exploration of parameter space (Cambridge and Mosegaard, 2002), but this flexibility comes at a cost: the applied inference method requires a substantial number of forward simulations. In fact, the number of required forward calculations per inference step scales linearly with the number of parameters. One important aspect for the feasible application of the method is therefore that the forward modeling step itself is fast. In the presented modeling case, this is not a problem (model interpolation time is in the order of seconds on a conventional desktop computer), but could become more significant for larger models and more complex likelihood functions. A single MCMC chain cannot be parallelized; however, increasing computational requirements could, at least partly, be offset by suitable parallelization schemes for model interpolation and export.
In addition to potential speed-up of the forward modeling step, an improved sampling method could be applied. Recent developments in the direction of Hamiltonian MCMC methods provide a promising path (Jasche and Kitaura, 2010; Homan and Gelman, 2014). We expect to extend our described method in this direction in the near future, as these sampling algorithms are a fundamental part of the newest version of the probabilistic programming framework provided by PyMC 3. In addition, advances in developments of multilevel Monte Carlo (Cliffe et al., 2011) and multilevel MCMC (Dodwell et al., 2015) methods could be applied to reduce the computational cost for all calculations that require a full model export.

The approach to use a Bayesian framework for geologic model construction enables a rich problem formulation and combination of geologic and geophysical data in a consistent manner. Bayesian inference methods are already commonplace in geophysical inversions (Mosegaard and Tarantola, 1995; Tarantola, 2005; Guillen et al., 2008), and we encourage the use of these methods for the purpose of structural geologic modeling itself. However, a suitable method to extend our approach to geophysical observations is to combine it with a suitable geophysical forward simulation. One example is the calculation of the potential field of gravity on the basis of the calculated structural model, where rock densities are assigned according to the prevalent geologic lithologies. The misfit between calculated and observed gravity could then be used in the form of an additional likelihood function. Our first tests in this direction were successful (de la Varga et al., 2015), and we will investigate this possibility in more detail in the future.

A hallmark of the model inference approach is that we obtain an ensemble of parameter sets and, based on these parameter sets, a suite of geologic models that respect prior distributions and information from the likelihood functions. We presented the analysis of the generated posterior distribution with conventional methods (i.e., the comparison between prior and posterior parameter distributions as well as the analysis of parameter correlations). In the case of the simple 1D model, we presented multiple model results as line plots in a single figure (see Figure 2), and, in the more complex case of the 3D model, we used cell information entropy (Wellmann and Regenauer-Lieb, 2012) to visualize uncertainties (Figure 3c and 3e). Depending on the modeling context, several other methods to analyze the posterior distribution and the generated models could be suitable. Examples are spatial representations of estimated probabilities for specific geologic units, general geodiversity metrics (Lindsay et al., 2013), or the calculation of stratigraphic variability (Lindsay et al., 2012). In addition, correlations on the basis of information theory could be analyzed using the generated model ensemble (Wellmann, 2013). Overall, the presented approach generates a range of outcomes that are suitable for a rich variety of different posterior analysis methods.

Even though we suggest that the main advantage of this method is to obtain an ensemble of probable models, the approach also enables the construction of a single "best-fitting" model, when the generation of such a best-fit model may be required in specific circumstances. One convenient procedure to obtain such a single realization is to use the MAP approach, as briefly outlined above.

We described our approach in the general form of an arbitrary geologic modeling method (\(\phi_i\) in equation 1) that can be suitably automated. We want to highlight here the importance of the model selection, especially in a nonparametric Bayes framework. In the presented cases, the models are defined by the interpolation method and the number of parameters. The accuracy of the selection dictates the compatibility of the model with the data. Therefore, being able to describe the observations given their likelihood is an indispensable prerequisite of the selected model. However, the MCMC algorithm itself detects potential incompatibilities, leading to a zero-probability result or to the lack of convergence. Also notice that as the Bayesian networks grow in complexity, more than one model could relate parameters to different sets of observations that potentially provide a better description of our parameters, but also would make model selection more difficult.

In case of interpolation models, different methods also consider different types of input data and geologic information. For example, classical surface interpolation approaches consider contact points between significant geologic units (e.g., Caumon et al., 2009). Other methods, including many recent developments of implicit field interpolation techniques, enable the consideration of orientation measurements in the model construction step (e.g., Lajaunie et al., 1997; Mallet, 2004; Calcagno et al., 2008). Novel developments in this field.
are also able to consider off-surface information on geologic units (Hillier et al., 2014). In general, the consideration of the type of information in the form of initial parameters will differ from method to method — and other information may be included in the form of likelihood functions in the Bayesian inference framework.

In summary, we propose here a different way of thinking about geologic modeling by considering structural geologic modeling as a Bayesian inference process. Our exposition is motivated by the realization that geologic models are themselves always a representation of a specific geologic hypothesis, based on available data and additional geologic information, and interpolated in space by a mathematical algorithm. This consideration enables us to combine prior knowledge with additional information in the form of likelihood functions, and to include all aspects in a probabilistic framework.

Acknowledgments
J. F. Wellmann would like to acknowledge the support from the AICES Graduate School, funded by the German Research Foundation (DFG). We thank J. Niederau for his support with GeoModeller and Intrepid Geophysics for granting an academic license of the program for the purpose of our research.

Appendix A

Figures
Probability distribution of parameters in the 3D case
Prior and posterior probability distributions for parameters used in the 3D modeling scenario are represented as box plots in Figure A-1.

Prior distributions for 3D geologic modeling study
Prior parameter distributions for the 3D geologic modeling study are given in Table A-1. All prior distributions are described as normal distributions, with mean value and standard deviations given in Table A-1.

Likelihood table
Likelihood functions return probabilities given a value for a certain parameter \( \theta \). The probability density that defines the likelihood may follow any type of distribution from classical normal or student distributions to any combination of them. For our case study, we have made use of normal distribution for the addition of information about our interface depths (Table A-2), whereas to constrain the graben structure, we directly opt to define impossible events; i.e., \( p(x) = 0 \) (Table A-3).

Code availability
PyMC 2 is fully open source and can be obtained for Python under the domain https://pymc-devs.github.io/pymc/. Here, it is possible to find a fully descriptive

<table>
<thead>
<tr>
<th>Type</th>
<th>Observation Id</th>
<th>( \mu ) (m)</th>
<th>( \sigma ) (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
<td>Basement right</td>
<td>-650</td>
<td>80</td>
</tr>
<tr>
<td>Interface</td>
<td>Sediment 1 right</td>
<td>-500</td>
<td>40</td>
</tr>
<tr>
<td>Interface</td>
<td>Sediment 2 right</td>
<td>-350</td>
<td>30</td>
</tr>
<tr>
<td>Interface</td>
<td>Basement left</td>
<td>-650</td>
<td>80</td>
</tr>
<tr>
<td>Interface</td>
<td>Sediment 1 left</td>
<td>-500</td>
<td>40</td>
</tr>
<tr>
<td>Interface</td>
<td>Sediment 2 left</td>
<td>-350</td>
<td>30</td>
</tr>
<tr>
<td>Interface</td>
<td>Basement center</td>
<td>-850</td>
<td>80</td>
</tr>
<tr>
<td>Interface</td>
<td>Sediment 1 center</td>
<td>-700</td>
<td>40</td>
</tr>
<tr>
<td>Interface</td>
<td>Sediment 2 center</td>
<td>-550</td>
<td>30</td>
</tr>
<tr>
<td>Interface</td>
<td>Fault left top</td>
<td>-200</td>
<td>20</td>
</tr>
<tr>
<td>Interface</td>
<td>Fault left base</td>
<td>-900</td>
<td>80</td>
</tr>
<tr>
<td>Interface</td>
<td>Fault right top</td>
<td>-200</td>
<td>20</td>
</tr>
<tr>
<td>Interface</td>
<td>Fault right base</td>
<td>-900</td>
<td>80</td>
</tr>
<tr>
<td>Azimuth</td>
<td>Fault right</td>
<td>270</td>
<td>30</td>
</tr>
<tr>
<td>Azimuth</td>
<td>Fault left</td>
<td>90</td>
<td>30</td>
</tr>
<tr>
<td>Dip</td>
<td>Fault right</td>
<td>85</td>
<td>20</td>
</tr>
<tr>
<td>Dip</td>
<td>Fault left</td>
<td>85</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Likelihood Id</th>
<th>( \mu ) (m)</th>
<th>( \sigma ) (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thickness layer 1</td>
<td>180</td>
<td>20</td>
</tr>
<tr>
<td>Thickness layer 2</td>
<td>130</td>
<td>20</td>
</tr>
<tr>
<td>Offset</td>
<td>200</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observation Id</th>
<th>( p(\theta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset negative</td>
<td>( \begin{cases} 0 &amp; \text{if sedi 2 center &gt; sedi 2 left or right} \ 1 &amp; \text{otherwise} \end{cases} )</td>
</tr>
<tr>
<td>Offset below faults</td>
<td>( \begin{cases} 0 &amp; \text{if basement center &gt; (fault base left or right) - 100} \ 1 &amp; \text{otherwise} \end{cases} )</td>
</tr>
</tbody>
</table>
documentation to perform Bayesian analysis making use of this package.

The Python functions to access the functionality of the geologic modeling method are combined in a package, pygeomod. Code is available on github (https://github.com/flohorovicic/pygeomod). Additionally, there is a branch of the project that intends to improve the integration of PyMC 2 and pygeomod and document through IPython Notebooks the whole Bayesian Inference process (https://github.com/Leguark).

Please note that the code comes without any warranty and that it is (at the time of writing) in a development state. For further questions, please contact the authors.

References
Before coming to Germany, he studied mining engineer at Universidad de León and Universidad de Oviedo in Spain, specializing in mining and energy resources. In addition to his studies, he was involved in several other academic activities. He became a tutor for applied geoscience (M.S.) of “Numerical Reservoir Engineering: Geological Knowledge, Data and Models” and “Numerical Reservoir Engineering: Geophysical Process Simulation,” as well as assistant of “Geothermics — Practicals and Exercises.” Additionally, he participated as an assistant in the Exploratory Teaching Space project: “Interaktive Blended Learning Programmierübungen mit Jupyter Notebooks.” In addition, he had the opportunity to give talks in the “Deformation Mechanisms, Rheology and Tectonics” 2015, Aachen, on the subject: “Adding Geological Knowledge to Improve Uncertain Geological Models: A Bayesian Perspective”; and in the European 3D GeoModeller User Meeting 2015, Orleans, on the topic: “Automation with pygeomodeller and uncertainty studies.”

J. Florian Wellmann studied geology at the University of Tübingen with a focus on geophysics and applied geophysics, and performed theoretical and experimental work for his Diplom-thesis (M.S. equiv.) at the ETH Zürich. After his studies, he worked for two years for the German Antarctic Survey (AWI) as geophysicist and overwintering team member on the German Antarctic research station Neumayer II. Subsequently, his combined interest in geophysical and geologic methods brought him to the Institute of Geothermal Resource Assessment, and subsequently to the University of Western Australia to pursue postgraduate studies in the group of K. Regenauer-Lieb, where he investigated the influence of structural uncertainties on the dynamic stability in geothermal flow fields. He holds a position as junior professor in numerical reservoir engineering at the RWTH Aachen University, Germany, and is a young research group leader within the graduate school Aachen Institute for Advanced Study in Computational Engineering Sciences. The work in his group is focused on understanding the influence of uncertainties in subsurface property distributions, and the influence of these uncertainties on subsequent process simulations. His work has been applied to explore uncertainties in geothermal systems and carbon sequestration simulations, and his approaches to quantify and visualize uncertainties in 3D settings have been taken up in diverse areas, from modeling projects in geologic surveys to medical imaging. The interplay between geologic modeling and process simulations was also the subject of his interest in subsequent postdoc positions, at the CSIRO and UWA. He is an active member of multiple scientific and steering committees and, from September 2015, leading the coordination of the IDEA League joint program “Master of Applied Geophysics” at the RWTH Aachen.