Value of Information Methodology for Dynamic, Spatial Earth Problems

Whitney J. Trainor

Program of Earth, Energy, and Environmental Science
Stanford University

Abstract

Spatial decisions regarding the sustainability of aquifers are difficult due to uncertainty in predictions made from hydrogeologic models. Specifically, the geologic heterogeneity is a large uncertainty that will determine the success of certain decision actions. The risk of making poor decisions due to this uncertainty may justify the collection of more information. The objective of this research is to develop a methodology for assessing the value of information (VOI) from spatial data for groundwater decisions. We borrow the VOI metric from the discipline of decision analysis and apply it to spatial problems. Both the prior geologic uncertainty and the information reliability must be quantified before the data is collected to estimate the VOI. The focus of this work is to 1) represent the uncertainty of the dynamic hydrogeologic response of the unknown subsurface, 2) provide a quantitative data reliability and 3) use both of these to propose a VOI workflow for spatial decisions and data. The dynamic response is the composite reaction to some external action (representing the decision action) and the geologic spatial heterogeneity. Geostatistical simulation and dynamic function are both used to actualize the data reliability. The geostatistical simulation requires a likelihood function describing the relationship of geophysical attributes to geologic indicators) Two different likelihood functions are used to test the sensitivity of the final VOI calculation to the degree of discrimination between the data attribute and the geologic indicator which influences the decision outcome. The geostatistical simulation creates Earth models conditioned to the synthetic geophysical data thereby representing possible interpretations that could be made if the information was purchased. The difference between the final VOI for the workflows using the two different likelihoods was minimal, thus more conditioned Earth models are needed to observe the influence of the likelihood function and also to achieve a pragmatic VOI calculation. In this proposed workflow, the intent is to capture the dynamic response that is important to the decision. Geostatistical simulation is used to represent the variability in the geophysical message and how it could resolve the static properties which influence the dynamic response. No forward modeling of geophysics is executed, but it could be included in the workflow.
1. Introduction

Worldwide, groundwater serves as an important resource for both agricultural purposes and human consumption. Many factors such as global climate change, increasing demand and contamination, threaten the quantity and quality of groundwater available. Public officials, groundwater managers and even private enterprise are being faced by decisions that will determine the sustainability of groundwater resources. Two examples of such decisions are managed aquifer recharge (to address overdrafts or store water for later recovery) or the removal of contaminant sources. Effective groundwater management requires numerous subsurface hydrogeologic models which represent our current assessment of uncertainty using all prior information, both qualitative and quantitative. The risks associated with poor decisions made based on these models alone could justify the cost of collecting additional data.

The field of decision analysis can be very useful for groundwater resource protection decisions. Decision analysis provides a probabilistic platform for engineers, scientists, and managers to communicate across expertise and provides tools for assigning a “dollar amount on uncertainty” [Howard, 1966]. The discipline of decision analysis conveniently provides a metric for the value of information (VOI) which is used to determine if a certain information source is worth its cost in terms of improving a specific decision goal. VOI can be applied to a variety of problems and decisions in fields such as engineering, business, and the natural sciences. Raiffa (1968) provides a thorough overview of both decision analysis and the value of information theory. Decision analysis and VOI has been widely applied to decisions involving engineering designs and tests, such as assessing the risk of failure for buildings in earthquakes, components of the space shuttle and offshore oil platforms [Cornell & Newmark, 1978, Paté-Cornell & Fischbeck, 1993, Paté-Cornell, 1990]. As demonstrated in those example studies, the performance statistics of these engineering apparatuses or components under certain conditions are readily available; therefore, applying the decision analysis framework for decisions involving engineering designs can be fairly straightforward and convenient. Additionally, the statistics on the accuracy of the tests or information sources that attempt to predict the performance of these designs or components are also available, as they are typically made repeatedly in controlled environments such as a laboratory or testing facility. These statistics are required to complete a VOI calculation as they provide a probabilistic relationship between the information message (the data) and the state variables of the decision (the specifications of the engineering design or component). This is often referred as the information’s reliability measurement.

Many challenges exist in applying this framework to spatial decisions pertaining to an unknown earth. Instead of predicting the performance of an engineer’s design
under different conditions, the desired prediction is the response of unknown subsurface—which can be very complex and poorly understood—due to some external action and the geologic heterogeneity. Geostatistical modeling can generate many possible representations of the subsurface, thus capturing the statistics of the “engineering design.” Variograms, or two-point statistics, are one way to describe the spatial correlation of a subsurface property, and they have been used to create many Earth models (or model realizations) through geostatistical simulation [Delhomme, 1979; Deutsch and Journel, 1997]. More recently multiple point statistics have been used in stochastic simulations. Hu and Chugunova (2008) give a review of multiple point geostatistics—the stochastic simulation of facies patterns in Earth models. Recent studies have demonstrated how this kind of geologic facies heterogeneity can dominate the hydrogeologic response of interest [Feyen and Caers, 2006; Ronayne et al, 2008]. It is the uncertainty of this hydrogeologic response that makes groundwater decisions difficult. Representing the heterogeneity as accurately as possible is important as it affects the decision actions that we take on the Earth, such as aquifer remediation [Wagner and Gorelick, 1987; Wagner and Gorelick, 1989; Freeze and Gorelick, 1999] Spatial heterogeneity must be included in the spatial decision-making process, which may include choosing between different locations for artificial recharge or contaminant-source removal.

Secondly, obtaining a meaningful measure of how spatial data will correctly identify geologic variables of interest (i.e. the information reliability) is not a trivial exercise, especially since this must be obtained before the proposed data is collected. VOI has been used in subsurface fluid flow decisions in petroleum engineering. Coopersmith et al. (2006) propose a qualitative approach to obtain the information reliability measurement. Using a modified Sherman-Kent language probability scale, they transfer language such as “certain,” “likely,” “unlikely,” and “doubtful” from expert interviews into discrete intervals of likelihoods. Bratvold et al. (2009) give a thorough review of the 30 VOI papers in the petroleum engineering literature from the past 44 years. Among other assessments and critiques, they note that only 13 of the 30 papers published address the issue of reliability, and 11 of these 13 used the subjective expert interview method.

VOI examples also exist in the hydrogeologic literature. Reichard and Evans (1989) present a framework for analyzing the role of groundwater monitoring in reducing exposure to contamination. The reliability here is the efficiency of detecting arsenic in the water (a scalar parameter), and no spatial dependence is included in the 1D contaminate transport modeling. Wagner et al (1992) compare four different strategies for representing uncertainty of hydraulic conductivities and solving for the optimal solution for a groundwater contamination decision. Therefore, the “information” is represented by the different deterministic and stochastic formulations and how their respective uncertainty measures affect the decisions made and resulting outcomes. The decision faced by groundwater managers in the example posed by Feyen and Gorelick (2005) is how to maximize the extraction of
groundwater (which is sold for profit) without violating ecological constraints (specific groundwater table levels). They consider how hydraulic conductivity information at certain locations may improve hydrogeologic model predictions and allow for an increase in water production while still observing the hydro-ecological balance. The aquifer is represented by many Earth models with a fixed mean, variance and covariance. The models with the hydraulic conductivity measurements are conditioned to these point location values. However, no particular measurement technique is specified, and thus, no analysis is made on the accuracy of a measurement, which would affect the value of information.

Spatial data such as from geophysical measurements may provide the kind of information needed to resolve the heterogeneity that affects the outcome of spatial decisions. The examples of VOI in the geophysics literature attempt to evaluate the geophysical phenomenon’s ability to resolve key geologic parameters. Houck (2004) uses VOI to evaluate whether better 3D seismic coverage is worth the extra survey costs. The seismic amplitude data is modeled using a 1D reservoir model to discern the reservoir thickness, porosity and fluid type, and the reliability of the amplitude data to detect these reservoir properties is modeled through a simulated calibrated data set (specifically linear regression is used to link the simulated data with the reservoir parameters). Houck and Pavlov (2006) use detection theory for “reconnaissance” controlled-source electromagnetics (CSEM) surveys for oil reservoir exploration. CSEM surveys with different configurations (different numbers of sources and receivers) are evaluated for their ability to detect economic and non-economic reservoirs (describing their absolute size). VOI is used to determine the survey design with the optimal value considering the cost of the survey and CSEM's ability to resolve reservoir size. Sensitivity maps are generated and used to determine the CSEM response to economic and uneconomic responses. For the sensitivity analysis, targets are limited to geometric shapes that are not representative of geologic features. Houck (2007) examines the worth of 4D data for two different reservoir conditions: one with and one without gas. The 4D “repeatability” is modeled through conditional probabilities that are generated by two different sets of triangle distributions of seismic attributes. Again, only 1D reservoir models are used. Eidsvik et al. (2008) introduce statistical rock physics and spatial dependence within a VOI methodology for the decision of whether or not to drill for oil. Spatial dependence is included in the porosity and saturation reservoir grids through a covariance model. At each of the grid locations, CSEM and seismic amplitude-versus-offset (AVO) data are drawn from likelihood models that represent the link between the reservoir properties and the geophysical attributes. Many data sets are generated (drawn), the posterior is approximated with a Gaussian, and then a posterior value is drawn from it. By drawing the geophysical information from a spatially correlated porosity and saturation field, this method attempts to preserve spatial dependence. However, while the decision is explicitly modeled, the spatial structure’s influence on the flow of oil is not taken into account. Finally, Bhattacharya et al. (2010) propose a decision-analytic approach to valuing
experiments performed in situations that naturally exhibit spatial dependence. They incorporate dependence by modeling the system as a Markov random field, and also include the effects of constraints on the decisions. Their methodology is illustrated with two examples related to conservation biology and reservoir exploration geophysics.

Even though studies have shown its effect, geologically realistic heterogeneity has not been included in the value of information analysis for spatial decisions and information. The examples in the geophysics literature do not include a flexible framework for including geologic realistic spatial heterogeneity. Yet, geophysical data has the potential to inform about spatial variability since it can provide more spatial coverage than core drillings, well logs or other spatially-limited data. Given these present shortcomings, the contributions of this work are three-fold. First, we present a flexible methodology for representing the prior uncertainty of spatial heterogeneity, which is used in the decision analysis framework. With this framework, the uncertainty of the dynamic hydrogeologic response (deemed important or related to the groundwater decision) can be captured. The second contribution addresses the need for a quantitative data reliability that includes the geologic heterogeneity and the dynamic response. The information reliability is achieved quantitatively through geostatistical simulation. Using a likelihood relationship between the observed geophysical attributes and the geologic indicators, Earth models conditioned to geophysical information represent how constraining the static properties can constrain the important dynamic response. In short, our methodology captures geophysical information’s potential to resolve the spatial heterogeneity and the subsequent composite hydrogeologic reaction of these spatial properties. The last contribution is the VOI calculation for spatial data which uses the first two contributions which are the prior and posterior uncertainty analyses.

As Keeney (1982) concisely describes, there are four characteristics of today’s decision problems that makes them, and consequently VOI studies, challenging: complicated structure, high stakes, no overall experts, and the need to justify decisions. Each of these characteristics will be described in terms of Earth science decisions, thereby reviewing the necessary notations and general concepts while introducing a synthetic demonstration case modeled after a European groundwater problem. This example will motivate and illustrate new decision challenges which require the concept and introduction of proxy decision variables. This case poses a significant problem for achieving a VOI metric as no information source measures the proxy variable directly or indirectly. For these proxy decision cases, a realistic VOI methodology for imperfect geophysical data is proposed by conditioning models to generated synthetic data that are derived from rock physics-based likelihoods.

2. Decision Analysis Review
2.1. Demonstration case setting

2.1.1. Complicated structure

Spatial decisions regarding the Earth vary from the frequently used “where to drill for oil,” to actions for protecting groundwater from contamination, and increasingly, to decisions regarding the subsurface storage of greenhouse gases such as CO₂. To demonstrate these decision problem characteristics, we will use an example which is inspired by a Northern European groundwater contamination decision case. The reader should distinguish this “demonstration case,” serving as illustration of a general methodology, from an actual case study since the main elements such as data and decision have been altered. Therefore, no real-world implications should be drawn from our analysis. In this example, we assume that a particular area relies solely on its groundwater sources to supply drinking water. Over the past several decades, the area aquifers have been compromised by surface-born contaminants due to urban growth and farming activities. Contamination will continue to be a threat until critical surface recharge locations are identified. This can only be successfully achieved if the hydraulically complex connections between the contaminant sources at the surface and the underlying aquifers are understood. Thus, the principle uncertainty is which surface locations act as recharge zones or entry points into the aquifer. The decision is to determine which contamination sources (e.g. farms) need to be removed to ensure a sustainable supply of drinking water.

Decision making in this context is difficult because of the uncertainty surrounding properties of the unknown Earth. Thus, the uncertainty regarding the unknown subsurface and how it will respond once the decision is made is the main facet of the “complicated structure” that we consider in this paper. Decision analysis manages the unknown parameters through probabilistic representation. Achieving the statistics of all the possible subsurface conditions is difficult, as the possible states to describe the Earth are tremendous. To address this, we represent uncertainty about the unknown subsurface by means of generating many Earth models. While any technique can be used, one can typically differentiate two types of uncertainty: “spatial uncertainty” (geostatistics) due to the limited amount of information to constrain such models and “model uncertainty” because the input to the spatial stochastic simulation algorithms is uncertain. The geological input parameters, represented by random variable Θ, are the dominant geologic features or characteristics of a particular locale that often control the outcome of the decision.

For the demonstration case, the geologic depositional system is interpreted as glacial buried valleys, which act as important subsurface groundwater resources in many countries in Northern Europe. Buried valleys are the result of the “waxing and waning of Pleistocene ice sheets” [BurVal Working Group, 2006]. These glacial valleys can be thought of as the primary level of uncertainty in the aquifer system structure. If largely filled with sand, the buried valley has potential for being a high
volume aquifer (reservoir). The superposition of three to five different generations of glaciation has been observed. Thus, glacial valleys from multiple generations cross-cut each other and can also appear to abruptly end as seen in Figure 1 [Jørgensen and Sanderson, 2006]. Figure 1 is not used in this study as a deterministic aquifer model. Instead, it is used as a concept for generating several training images in multiple-point geostatistical modeling and simulation [Strebelle, 2002; Caers, 2005].

Figure 1: Network of Buried Valleys; darker to lighter representing older to younger buried valley generations (Jørgensen & Sanderson, 2006)

Figure 2 depicts Θ as the glacial buried valley training images with two possible outcomes $\Theta \in \{\theta_1, \theta_2\}$ for the dimensions of the buried valley lithofacies. At the bottom of Figure 2, a few Earth models are shown that are generated from $\theta_i$ and a stochastic algorithm. Generally, any input geological parameter may have $N$ number of outcomes:

Equation 1

$$\Theta \in \{\theta_1, \ldots, \theta_i, \ldots, \theta_N\}$$

Ideally, experts (e.g. geologists) assign prior probabilities $Pr(\Theta = \theta_i)$ for these specific geologic input possibilities to occur. The spatial stochastic variation accounts for the spatial variability that may occur within any of the distinguishing qualities or features of $\theta_i$. As seen in Figure 2, many Earth models may be generated and represented by

Equation 2

$$z^{(t)}(\theta_i) \quad t = 1, \ldots, T$$
with \( T \) the total number of models. The ensemble of \( \mathbf{z}^{(t)}(\theta_j) \) Earth models captures the prior uncertainty regarding the subsurface properties of interest, which are all captured in the vector \( \mathbf{z} \).

![Diagram of Earth models creation process](image)

**Figure 2:** Schematic showing the two randomizations used to create the Earth models. First, the geologic input parameters are identified, then outcomes are drawn and lastly these outcomes are used as input in to stochastic algorithms to create each model.

### 2.1.2. High Stakes.

Much effort is made on creating realistic Earth Models such that the predictions of the decision alternatives (for example remove farm at location A or location B or at both locations A and B) reasonably capture the range of possible outcomes. The difference between these outcomes could represent severe environmental damage or the loss of millions of dollars. Accordingly, the decision makers must identify the possible alternatives to the decision, which are denoted by \( a=1,\ldots,A \), where \( A \) is total number of alternatives identified. These alternatives may have different outcomes due to the possible and unknown heterogeneity modeled with \( \mathbf{z}^{(t)}(\theta_j) \). We can imagine how removing contaminant sources at different locations with varying recharge properties will result in different aquifer quality protection. The function \( g_a \)
denotes the action taken on the Earth (such as the removal of contaminant sources). It models the predicted outcome of alternative \( a \) with unknown subsurface captured by the models \( Z^{(t)}(\theta_i) \). Lastly, in order to evaluate and compare the different alternatives, the outcome for each combination of subsurface possibility and alternative must be expressed in terms of value

\[
\nu_a^{(t)}(\theta_i) = g_a(Z^{(t)}(\theta_i)) \quad a = 1,\ldots,A \quad t = 1,\ldots,T
\]

where value can be in terms of monetary units ($), ecological health [Polasky and Solow, 2001] or some other appropriate utility.

### 2.1.3. No overall experts

So far, several disciplines have been introduced and are necessary for spatial Earth decisions: geology, geostatistics, modeling related to the physical, chemical or others processes of the decision action, and economics. An effective decision analysis requires the coordination of expertise from these fields. As identified by Howard (1966), a sensitivity analysis should be completed by the geologist(s) and modeler(s) to determine which uncertain parameters are most consequential and hence control the outcome of a decision. For this example, we assume this sensitivity analysis has identified the input geologic parameters \( \theta_i \) representing the different width, length and thickness dimensions of the buried valley lithofacies. The sensitivity of other parameters related to the economics or public policy constraints should be analyzed; however, we deem these considerations outside the scope of this work. Therefore, the sensitivity of values and costs of different decision alternative outcomes are ignored here.

### 2.1.4. Need to justify decisions

Both public and private decision makers will have to defend their Earth Science decisions, such as their choices of their prior uncertainty \( Z^{(t)}(\theta_i) \), their identification of the \( A \) alternatives, and their predicted outcomes \( \nu_a^{(t)} = g_a(Z^{(t)}(\theta_i)) \). Depending on the decision outcomes, environmental regulatory authorities and/or economic entities such as shareholders or supervisors will scrutinize the rigor of their decision analysis. During this regulatory or economic scrutiny, the decision makers may be asked why further information was not gathered to ensure a better result. Relevant information sources may be able to reveal which input geologic parameters \( \theta_i \) exist at a certain locale, and subsequently, the best decision for \( \Theta = \theta_i \) could be made.

### 2.2. Review of VOI for spatial problems
The decision of whether or not to purchase data is entirely dependent on the spatial decision represented in \( g_b \) and the perceived uncertainties. Note that VOI is calculated before any information or data is collected, as we are trying to decide if it is worth acquiring or purchasing. The formal definition of VOI is the difference between the value with data (also known as value with free experiment \( V_{FE} \)) and the prior value \( V_{prior} \) (or value without data):

\[
VOI = V_{FE} - V_{prior}.
\]

Consider \( V_a \) is the random variable describing the possible outcomes of value due to randomization of both input parameters and spatial variation of properties of \( Z(t)(\theta_i) \) and for a given action \( a \), then

\[
V_{prior} = \max_a E[V_a] \quad a = 1, \ldots, A.
\]

In other words, the prior value describes the best outcome given the \( A \) alternatives and the present uncertainty, represented by the expectation, assuming a risk-neutral decision maker [Raiffa, 1968]. We can estimate this expectation with our discrete, scalar values of Equation 3 and the prior probabilities \( \Pr(\Theta = \theta_i) \)

\[
V_{prior} = \max_a \left( \sum_{i=1}^{N} \Pr(\Theta = \theta_i) \frac{1}{T_{0_i}} \sum_{t=1}^{T_{0_i}} V_a(t) \right) \quad a = 1, \ldots, A
\]

where \( N \) is the number of outcomes from input geologic parameter \( \Theta \) and \( T_{0_i} \) represents the number of Earth models generated for each geologic parameter \( \theta_i \) (see Figure 2).

The value with the free experiment or information can be expressed as

\[
V_{FE} = E \left[ \max_a E[V_a | d] \right] \quad a = 1, \ldots, A
\]

where the vector \( d \) represents the synthetic or forward simulated data related to the proposed data source. Again, all the defined uncertainties, alternatives, and value outcomes are utilized, except now there’s one more expectation over the data \( d \). Recall that no data has been collected; instead we must simulate the possible datasets.
and evaluate how this data $d$ would influence the decision. If $d$ includes realistic errors that may be in the data, then $V_{FE}$ becomes the value of imperfect information $V_{II}$. Attaining a realistic $V_{II}$ involves an estimation of how accurate the data is in resolving certain Earth parameters that are relevant to the decision. This is also known as the reliability of the information.

In Trainor-Guitton et al. (2010), a general methodology (for any Earth model and data source) was outlined to attain a meaningful measure of the reliability of information in terms of how successful spatial data $d$ would correctly identify and distinguish the specified input geologic parameters $\Theta_i$. To achieve this, Trainor-Guitton et al. (2010) simulate the geophysical measurement on the Earth models $z(t)\{\Theta_j\}$ and made (automatic) interpretations on the resulting simulated data in terms of the original $\Theta_j$. All the resulting interpretations are described with the random variable $\Theta_{int}$. These interpretations can then be compared with the original $\Theta = \Theta_j$, and ultimately, the frequency at which each interpretation correctly or falsely identifies the input geologic parameter is obtained as the reliability of information

Equation 8
$$\Pr(\Theta_{int} = \Theta_j | \Theta = \Theta_j) \quad i, j = 1, ..., N.$$  

With this assumption, the random variable $\Theta_{int}$ replaces the expectation over $d$ in Equation 7 and uses the reliability measure to estimate the value with imperfect information $V_{II}$

Equation 9
$$V_{II} = \sum_{j=1}^{N} \left[ \Pr(\Theta_{int} = \Theta_j) \max_{a} \left( \sum_{i=1}^{N} \Pr(\Theta = \Theta_j | \Theta_{int} = \Theta_j) \frac{1}{T_{\Theta_j}} \sum_{t=1}^{T_{\Theta_j}} v_a(t)\{\Theta_i\} \right) \right] \quad a = 1, ..., A$$

In defining the reliability using Equation 8 and using it for VOI calculation in Equation 9, it is implicitly assumed that knowledge of $\Theta_j$ solves the decision problem deterministically; in other words, there is a direct link between the decision we make and the uncertainty on the input geologic parameter $\Theta_i$. This would mean in our demonstration case that knowing the buried valley dimensions would determine which farms should be removed.
3. **Methodology**

3.1. **The proxy decision variable: motivation**

In many situations, however, knowing $\theta_i$ does not resolve the decision problem deterministically. For our demonstration case, knowledge of the buried valley dimensions will not tell us precisely which farms to remove. The network of connected buried valleys is complex; “significant parts of the recharge area may therefore lie at relatively large distances from the valley [which represents the deep aquifer]” [Sandersen and Jørgensen, 2003]. Thus, contamination can be transported kilometers from its surficial entry point into a deep aquifer. Knowledge that a valley or non-valley exists at one particular location is not a thorough representation of possible risks of contamination or vulnerability. This illustrates the complex relationship between model characteristics (the heterogeneity as represented by alternative $\theta_i$’s) and the aquifer’s vulnerability to surface-born contaminants. In this paper, we introduce a new VOI calculation that relies on a so-called “proxy decision variable” which represents the geological heterogeneity locally as the summary of the subsurface heterogeneity and some dynamic response evaluated on it.

A “proxy decision variable” locally represents the geologic heterogeneity’s response to some applied stress or process. This variable is required when facing a decision relying on some dynamic response at a specific location within the Earth models. We will call these responses the dynamic simulation function, which may model physical, chemical, geomechanical, or any other processes. For our demonstration case, we will introduce a so-called “aquifer vulnerability” as the proxy decision variable, and the simulation of fluid flow in porous media is the dynamic simulation function. Defining this proxy decision variable will provide an indication of the decision outcome for a particular action at a particular location, hence removing the difficulty as to whether a polluting farm “connects” with an aquifer several kilometers away.

The next three sections describe the three key parts of our VOI methodology for dynamic, spatial decisions. First (Section 2.2), we will show how the proxy variable can be estimated for all the Earth models. This provides a range of proxy variable estimations that capture the uncertainty regarding the outcomes of location-specific decisions. Then in Section 2.3 we will address how to generate datasets utilizing a likelihood of geophysical data to inform lithology. Multiple Earth models conditioned to these synthetic datasets are achieve through geostatistical simulation. These represent the possible lithology interpretations that could be made from the data, and they are needed to calculate the value of imperfect information (Equation 7). Section 2.4 shows how the uncertainty of the proxy decision variable and the multiple conditioned models can be integrated into the VOI calculation framework. Explanation of details of the demonstration case will be give for these two steps.
Lastly, Section 3.0 applies the methodology to our demonstration case of aquifer vulnerability.

3.2. Estimating the proxy decision variable

We illustrate the estimation of the proxy decision variable by means of our demonstration case where the decision depends on aquifer vulnerability: the surficial locations where contaminants can enter into the aquifer [Thomsen, et. al, 2004]. However, this workflow can also be applied to other evaluations requiring proxy decision variables involving decisions depending on some form of connectivity: well-to-surface or well-to-well due to a fracture network [Kerrou, et. al, 2008; Karimi-Fard and Firoozabadi, 2001]. Similarly, infiltration potential defines surficial locations ability to serve as artificial recharge points [Stamos, et. al, 2002].

The proxy decision variable depends on the response of the Earth models to some dynamic simulation function which is denoted as $f$. In estimating the proxy decision variable, the output of $f$ may require some post-processing or interpretation. Specifically, we assume a single dynamic simulation function exists to transform each $z^{(t)}(\theta_i)$ into the proxy variable

Equation 10

$$
\nu^{(t)}_{\ell}(\theta_i) = f(z^{(t)}(\theta_i), \ell)
$$

Here $\ell$ denotes all the $L$ locations within each model $z^{(t)}(\theta_i)$ where the decision must be made. We must define $\ell$ because the goal is to determine the best decision alternative for different locations $\ell$ within $z^{(t)}(\theta_i)$. For our aquifer vulnerability study, $\ell$ will cover all possible surficial, recharge locations. The function $g_a$ represents the $A$ different actions or alternatives and their respective magnitudes that are deemed plausible. In our demonstration case, the two alternatives are to either remove possible contaminant sources (e.g. farms) or do nothing. For now we assume that action $g_a$ can only be taken at one position $\ell$ at a time. This restriction will be alleviated later.

The proxy variable $\nu^{(t)}_{\ell}$ has the information that will determine all the value outcomes at each location

Equation 11

$$
\nu^{(t)}_{\ell,a}(\theta_i) = g_a(\nu^{(t)}_{\ell}(\theta_i))
$$

With a proxy decision variable, we can determine the outcome of different
decision alternatives \( a \) at locations \( \ell \). And since these are obtained for all Earth
tools \( z^{(t)}(\theta_j) \), we have a range of the possible value outcomes for these local
decision actions, representing the uncertainty on the local decision outcome. The
location-specific values of Equation 11 will eventually be used in \( V_{\text{prior}} \) (Equation 6).
The details on how they will be utilized will be explained in Section 3.4.

### 3.3. Reliability for measuring proxy decision variables

#### 3.3.1. Reliability through geostatistical simulation

In VOI methodologies, a data reliability measure is a conditional probability of
the form

\[
\Pr(\text{what the data says | what the real world is})
\]

However, since the “real world” includes both the complex subsurface variation
as well as the proxy decision variable as a response to it, such reliability measure is
not trivial to explicitly determine by means of forward model runs [Bickel et al, 2006;
Houck and Pavlov, 2006; Houck, 2007]. Since we cannot explicitly derive the
reliability measures in terms of conditional probabilities, we propose a geostatistical
simulation approach for including data reliability into the VOI calculation based on
rock physics relationships. Geostatistical simulation is used to represent the
variability in the geophysical message and how it could resolve the static properties
which influence the dynamic response. No forward modeling of geophysics is
performed, but it could be included in the workflow. In this proposed workflow, the
intent is to capture the dynamic response that is important to the decision.

In summary, our methodology is as follows:

1) Use a likelihood function to generate a synthetic dataset from each prior
Earth model, where the likelihood describes the relationship between the
geophysical data and the geologic indicators.

2) From each of the synthetic data sets, derive a pre-posterior probability
distribution on how informative that synthetic dataset is about subsurface
lithology.

3) Use that pre-posterior probability to generate multiple, new Earth models
constrained to each of the synthetic data sets. These conditioned Earth
models represent the variability in what we could interpret from the data.

4) Create realizations of the proxy decision variable by applying the
dynamic simulation function \( f \) to the conditioned (interpreted) Earth
models.

These four steps are graphically demonstrated in Figure 3. This is a valid
approach to attain a VOI measure considering the nature of proxy decision variable
situations. By using rock physics relationships, synthetic data can be generated that contains local information, in our case whether a buried valley exists at that location or not. The local information has the potential (if the data is sufficiently reliable) to constrain the conditioned or interpreted Earth models, which in turn will influence the response to the dynamic simulation function. We can imagine the two extreme cases: the geophysical attribute either perfectly informs lithology or gives completely unreliable lithology information. With perfect information, these four steps will allow us to retrieve the true proxy variable, as all the actual valley & non-valley locations will be identified perfectly through the likelihood and soft probability. With the true proxy variable we can then make the best decisions for each location. Whereas completely unreliable data could generate conditioned lithology models that are very different from their respective prior models and from each other (as the pre-posterior is uninformative). Decisions made on proxy variables obtained through these inaccurately interpreted models, will not represent the best decisions made for the true proxy variable. This will be further validated in Section 2.4 with the expression for VII. Next, we will clarify each of the four steps in detail.

Figure 3: The 4 steps of obtaining a data reliability for proxy decision variables through geostatistical simulation. Step 1: Generate synthetic dataset from likelihood and prior model. Step 2: Generate soft probability cube for valley from dataset and information content. Step 3: Generate conditioned Earth
model with soft probability. Step 4: Obtain the conditioned proxy variable by applying the dynamic simulation function to the conditioned Earth model.

3.3.2. Creating synthetic data sets with rock physics

Rock physics relationships associate geologic indicators with geophysical attributes [Mavko, et. al, 1998]. Recall that all subsurface properties of interest can be captured in the vector $z$ of $z(t)(\theta_i)$. This could include geological indicators. Specifically for our demonstration case, the geologic input parameters $\theta_i$ represent different combinations of buried valley dimensions. We assume exclusivity between lithology and facies $(\theta_i)$: sand is always interpreted as buried valley and non-sand as non-valley. This assumption may be true for some geographic locations but not always [Sandersen and Jørgensen, 2003]. However, this assumption could be relaxed or accounted for by estimating a percentage of buried valleys that are filled with non-sand materials. But for this demonstration, an exclusivity between lithology and $\theta_i$ is represented in $z(t)(\theta_i)$.

Now we consider a link between the lithology and possible geophysical attributes. For this synthetic example, the geophysical information source being considered is transient or time-domain electromagnetic (TEM) data. TEM works with a transmitter loop that turns on and off a direct current to induce currents and fields into the subsurface; meanwhile, the larger receiver loop measures the changing magnetic field response of induced currents in the subsurface [Fitterman and Stewart, 1986; Christiansen, 2003]. The magnetic field response in time is then inverted into a layered model of electrical resistivity and thickness values [Auken et al, 2008]. The recovered electrical resistivity can be an indication of the lithology type as clay (non-sand) typically has an electrical resistivity less than 30 ohm-m, whereas sand is usually greater than 80 ohm-m.

The association between lithology ($litho$) and the electrical resistivity ($\rho$) may be described through an empirical relationship [Archie, 1942; de Lima and Sharma, 1990]. However, a probabilistic relationship is a more realistic description of what the indirect geophysical data can resolve and is typically modeled as a conditional probability relating the geologic indicators of the prior models (known) to the geophysical attributes (unknown). This conditional probability (a likelihood) can be obtained through forward models [Eidsvik et al, 2008], from some calibration dataset [Houck, 2004], from geologic analogs, or could be synthetically created. Two likelihoods of the form

\begin{equation}
Pr(P = \rho | Litho = litho) \quad litho = \{sand, non-sand\} \quad 0 < \rho < \infty
\end{equation}

were synthetically created for our demonstration case to describe two possible
relationships between electrical resistivity and lithology. To utilize Archie’s relation, information on porosity and saturation would be needed not just lithology, however we base our likelihoods off of a calibration of co-located inverted resistivity data and driller’s logs, which respectively have electrical resistivity and lithology information. The first one (shown in Figure 4) demonstrates a decent electrical resistivity contrast between sand (assumed here as valley) and non-sand (non-valley). Whereas the second data (Figure 5) has a less discriminating message about lithology as a greater overlap exists between the electrical resistivity and the two lithologies.

Figure 4: Synthetic data reliability describing a good contrast between electrical resistivity Y and given lithology X
Most importantly, we can create many synthetic datasets $d$ (which are in the form of the geophysical observables) using Monte Carlo sampling of Equation 12 (represented in Figure 4 and Figure 5). Creating many synthetic datasets $d$ in this way allows us to statistically represent the possible variation in the datasets. Namely, knowing the occurrence of lithology at a certain location within $z^{(t)}(\theta_j)$, an instance of electrical resistivity ($\rho$) can be drawn by Monte Carlo sampling of the cdf (cumulative distribution function) of either the top of Figure 4 (given that the considered location is sand/valley) or the bottom (given the considered location is non-sand/non-valley). Similarly, this can be performed with the cdf’s of Figure 5. This is repeated for all locations within the model $z^{(t)}(\theta_j)$ to generate the dataset:

$$d^{(t,i)} \quad i = 1, \ldots, N \quad t = 1, \ldots, T_{\theta_j}.$$  

Therefore, there is an electrical resistivity dataset $d^{(t,i)}$ that corresponds to each of the prior models $z^{(t)}(\theta_j)$. Figure 3 demonstrates this Step 1 and displays an example synthetic dataset of electrical resistivity that was generated using one prior model and the cdf’s versions of the pdf’s in Figure 4.
3.3.3. Deriving information content to create conditioned Earth Models

While data reliability models a conditional distribution of the form:

\[ \Pr(\text{what the data says} | \text{what the real world is}) \],

"information content" of a data source about the unknown Earth is of the form:

\[ \Pr(\text{what the real world is} | \text{what the data says}) \].

This conditional probability, also termed "pre-posterior" is important in generating new Earth models constrained to the synthetic data sets as outlined in our workflow. More specifically the pre-posterior probabilities we are after are of the form

\[ \text{Equation 14} \]

\[ \Pr(\text{Litho} = \text{litho} | P = \rho) \]

\[ \text{litho} = \{\text{sand, non-sand} \} \quad 0 \leq \rho \leq \infty. \]

Through Bayes Law, Equation 14 can be obtained from Equation 12. Using Equation 12 and Equation 14, we notice that perfect information would imply that an exclusive relationship exists between lithology and electrical resistivity, such that any draw from \( \Pr(\text{P} = \rho | \text{Litho} = \text{litho}) \) or \( \Pr(\text{Litho} = \text{litho} | \text{P} = \rho) \) would be 0 or 1.

We can create a lithology probability cube using each of the datasets \( \mathbf{d}^{(t_i)} \) to obtain a sand and non-sand probability from Equation 14

\[ \text{Equation 15} \]

\[ \mathbf{y}^{(t_i)}_{i = 1, \ldots, N} \quad t = 1, \ldots, T_{0_i}. \]

For our demonstration case, \( \mathbf{y}^{(t_i)} \) contains a probability of sand and non-sand (valley or non-valley) at each location within the dataset \( \mathbf{d}^{(t_i)} \), which is derived from the prior model \( \mathbf{z}^{(t_i)}(\theta_i) \) (see Step 2 of Figure 3). This is known as the "soft probability" for conditioning multiple-point realizations [Caers, 2005].

Step 3 involves creating multiple conditioned Earth models to this soft probability (see Step 3 of Figure 3). We let the \textit{snesim} algorithm (Single Normal Equation Simulation) within Stanford’s Geostatistical Earth Modeling Software (SGeMS) generate Earth models (realizations) of lithology

\[ \text{Equation 16} \]

\[ \mathbf{z}^{(t,w)}(\mathbf{y}^{(t_i)}, \theta_i) \]

\[ w = 1, \ldots, W \quad i = 1, \ldots, N \quad t = 1, \ldots, T_{0_i}. \]

Here \( w \) represents the number of realizations generated from the same soft probability cube. By generating several conditioned Earth models, we can capture the different possible Earth model interpretations that could be made from the data which
stems from the overlap in the likelihood. In addition to being conditioned to the soft probability \( \mathbf{y}^{(t,\ell)} \), these Earth models reflect the prior through the training images \( (\theta_i) \); furthermore, they could be constrained to any available hard data (see Caers et al, 2001 and Strebelle et al, 2003 for applications of this methodology to modeling sand/shale sequence in oil reservoirs). Depending on how discriminating the data is, these conditioned models \( \mathbf{z}^{(t,w)}(\mathbf{y}^{(t,\ell)},\theta_i) \) may be very different or very similar to the prior models \( \mathbf{z}^{(t)}(\theta_i) \) from which they are derived.

3.3.4. Creating data-constrained proxy decision variables

Finally, we arrive at Step 4: obtaining the proxy decision variables from the conditioned Earth model (see Step 4 of Figure 3). As with the prior models, the dynamic simulation function \( f \) must be applied to the new conditioned models \( \mathbf{z}^{(t,w)}(\mathbf{y}^{(t,\ell)},\theta_i) \) to get the conditioned or interpreted proxy variable

\[
\mathbf{s}^{(t,w)}_{\ell}(\mathbf{y}^{(t,\ell)},\theta_i) = f(\mathbf{z}^{(t,w)}(\mathbf{y}^{(t,\ell)},\theta_i)) \quad i=1,\ldots,N \quad t=1,\ldots,T_{\theta_i} \quad \ell=1,\ldots,L \quad w=1,\ldots,W
\]

The conditioned proxy variable \( \mathbf{s}^{(t,w)}_{\ell}(\mathbf{y}^{(t,\ell)},\theta_i) \) determines the outcome of the decision, which for the demonstration case is whether a farm is removed or not. Again, as in the prior proxy variable, we express the outcome of these decisions in terms of value:

\[
\mathbf{v}^{(t,w)}_{\alpha,a}(\mathbf{y}^{(t,\ell)},\theta_i) = g_{\alpha}(\mathbf{s}^{(t,w)}_{\ell}(\mathbf{y}^{(t,\ell)},\theta_i)) \quad \alpha=1,\ldots,A \quad i=1,\ldots,N \quad t=1,\ldots,T_{\theta_i} \quad \ell=1,\ldots,L \quad w=1,\ldots,W
\]

These individual values from the conditioned models are used to calculate the value of imperfect information \( V_{II} \), since they are derived from proxy variables that have been constrained by the data reliability measure. In the next section, we illustrate the application of these four steps (1-4) to the aquifer vulnerability demonstration case as well as present the complete VOI calculation methodology.

3.4. VOI Methodology summary

We now present the complete methodology, which allows us to assess the value of a geophysical technique that does not directly measure the decision variable. In particular, we need to compute the VOI components of \( V_{prior} \) and \( V_{II} \). We present a VOI workflow consisting of three parts. First, the generation of the prior models and calculation of \( V_{prior} \) is described. Second, we describe how the two synthetic
likelihood measures are utilized with the prior models to make many synthetic datasets. From these synthetic datasets, soft probability cubes are created from the datasets and used to generate conditioned Earth models that represent possible interpretations from the data. Third, the generated conditioned models are used to calculate $V_{II}$. Lastly in this section, the methodology will be generalized to allow several local actions to be taken simultaneously. We would like to emphasize that this paper is about the overall methodology and that some of the detailed components (e.g. tracer simulation, vulnerability definition) can be re-defined based on specific desires.

3.4.1. Prior model generation.

The details of Part A in the schematic of Figure 6, which depicts the entire proposed workflow, are described. The generation of the prior models $z^{(t)}(\theta_i)$ utilizes eighteen training images ($N=18$) to represent the model uncertainty of the buried valley length, width and thickness dimensions. All the training images are deemed equally probable: $\Pr(\Theta = \theta_i) = \frac{1}{N} = 0.055$. Within each of these $\theta_i$, ten binary facies Earth models are generated (examples shown at the bottom of Figure 2) using the snesim algorithm [Strebelle, 2002]. Each of the 180 models has $143 \times 91 \times 50$ grid cells with each cell dimension being $150\text{m} \times 150\text{m} \times 4\text{m}$.
Figure 6: Overall workflow of this VOI methodology. A) First the generation of the prior models and calculation of $V_{\text{prior}}$. B) Second, the reliability measure and how it is utilized with the prior models. C) The conditioned models are generated and used to calculate $V_{\text{II}}$.

3.4.2. Dynamic simulation $f$.

Once the prior Earth models $Z^{(t)}(\theta_i)$ are established, the dynamic simulation function $f$ is applied to achieve prior proxy variable $s^{(t)}(\theta_i)$. In order to define aquifer vulnerability, flow simulation is performed with tracer initially placed at all $L$ surficial locations. The permeability of two facies is assigned deterministically: valley (sand) is 1165 mD (2.8 ft/day or 9.8E-6 m/s) while non-valley (non-sand) is 1.1 mD (2.6E-3 ft/day or 9.2E-9 m/s). The simulation is run for twenty years with extraction and influx boundary conditions (representing pumping wells, precipitation and regional recharge).

3.4.3. Proxy decision variable.

We establish thresholds of the tracer concentration that will allow us to delineate which surface locations are major entry points into the aquifer. Thresholds are chosen to account for and remove situations where pooling occurs at the surface or
very insignificant amounts of tracer have reached into the aquifer. These concentration thresholds define continuous concentration bodies; Figure 7 shows an example of a 3D view of identified concentration bodies in unique colors and the topography as a semi-transparent surface. Locations where these concentration bodies intersect the surface are mapped as vulnerable. The volume of the concentration body represents the potential damage a contaminant could do if released at that surface location $\ell$. Therefore, the magnitude of vulnerability $s_{\ell}^{(t)}(\theta_i)$ at these surface intersections is equal to the volume of the concentration body. The conceptual 2D cross-section of Figure 8 demonstrates the effective entry points at the surface and the volume of concentration bodies. Figure 9 is an example vulnerability map. More refined vulnerability maps could be constructed including chemical and biological processes but this not the focus of this paper; instead the overall methodology is our contribution.

Figure 7: Tracer concentration thresholds are applied to flow simulation results such that each continuous tracer body is identified. Topography depicted in semi-transparent surface.
Figure 8: Conceptualization of how vulnerability maps (the proxy decision variable) are made from the tracer concentration bodies.

Figure 9: Example vulnerability map which indicates locations that serve as entry points into the aquifer. Their magnitude reflects the volume of aquifer each location can affect.

3.4.4. Decision action & $V_{\text{prior}}$.  

$s^t_k(\theta_j)$ directly determines the outcome of our decision $g_a$. The scalar value
results \( v_{t,a}^{(l)}(\theta_j) \) of Equation 11 can be used in the expression for \( V_{\text{prior}} \) of Equation 6 resulting in:

Equation 19

\[
V_{\text{prior}} = \sum_{\ell=1}^{L} \max_a \left( \sum_{j=1}^{N} \Pr(\Theta = \theta_j) \frac{1}{T_{\theta_j}} \sum_{t=1}^{T_{\theta_j}} v_{t,a}^{(l)}(\theta_j) \right) \quad a = 1, \ldots, A.
\]

This prior value includes a sum over the local values which determine the outcome of any action at location \( \ell \). This summation is outside \( \max_a \) because each decision at \( \ell \) can be made independently from other locations; recall that we want to allow an independent action \( a \) per each location \( \ell \). One could remove a farm (\( a=1 \)) at location \( \ell = 2 \) and not remove a farm (\( a=2 \)) at \( \ell = 1 \).

3.4.5. Reliability & synthetic data.

The second step of the workflow is represented in part B of Figure 6. The first of two tasks for Part B is to use the data likelihood (Equation 12) and the prior models to generate synthetic datasets \( d^{(l)} \) since we are considering purchasing information that we do not yet have. The synthetic data \( d^{(l)} \) is generated as described in Section 2.3. For each of our buried valley prior models, we generate one electrical resistivity dataset for a total of 180 datasets from the likelihood of Figure 4 and 180 datasets from the likelihood of Figure 5. These electrical resistivity datasets represent the information we could expect to collect, given our uncertainty of both the subsurface heterogeneity (represented with the prior models) and the TEM data to resolve lithology (represented through the two respective data likelihoods).

The second task of Part B uses the conditional probability of Equation 14 and datasets \( d^{(l)} \) to generate a soft probability cube \( y^{(l)} \). For each \( y^{(l)} \), two \( (W=2) \) new conditioned Earth models \( z^{(l,w)}\left(y^{(l)}, \theta_j \right) \) are generated. This may be considered the minimum of conditioned models that should be generated. More conditioned models will capture the possible variability due to the imperfect geophysical information message. However, again, our aim is developing the complete methodology. Ultimately, there are two sets of 360 conditioned models, each constrained to their respective synthetic datasets (represented through the soft probabilities \( y^{(l)} \)) and the spatial constraints of the original input geologic parameters \( \theta_j \) (through the training images).

**Conditioned models to attain VII.**

Third, we arrive at part C of Figure 6 which represents the part of the workflow
that achieves the value with imperfect information $V_{II}$. The conditioned models of our demonstration case are put through the same workflow as the prior models in order to achieve the aquifer vulnerability $s_{\ell}^{(t,w)}(y(t_i), \theta_i)$ (Equation 17). These results are ultimately used to obtain all the individual value outcomes $v_{\ell,a}^{(t,w)}(y(t_i), \theta_i)$ (Equation 18). At last, all the individual conditioned values are used to calculate the value with imperfect information

Equation 20

$$V_{II} = \sum_{\ell=1}^{L} \left[ \sum_{i=1}^{N} \Pr(\Theta = \theta_i) \frac{1}{T_{\theta_i}} \sum_{a=1}^{\Theta} \max \left( \frac{1}{W} \sum_{w=1}^{W} v_{\ell,a}^{(t,w)}(y(t), \theta_i) \right) \right] a=1,\ldots,A.$$ 

Chronologically, we first obtain the $W$ possible interpretations of the proxy variable $s_{\ell}^{(t,w)}(y(t_i), \theta_i)$ through the conditioned models before we make our decision. Therefore, we can choose the best decision action on average for those interpretations, represented by $\max_a$. Then the expected value from all the models for location $\ell$ is taken. Note that, unlike Equation 9, this $V_{II}$ is not weighted by probabilities derived from a reliability measure. Instead, the reliability is captured in imperfect simulated data $d(t_i)$. The conditioned models $z_{\ell}^{(t,w)}(y(t_i), \theta_i)$ account for the possible inaccuracies of the geophysical message to inform lithology. The imperfect data will have value if it can resolve the proxy variable and lead to decisions with a higher value output than the $V_{prior}$.

Notice that in the case of perfect information $v_{\ell,a}^{(t,w)}(y(t_i), \theta_i)$ will be equal to $v_{\ell,a}^{(t,w)}(y(t_i), \theta_i)$, as all prior models will be perfectly recovered through the data into $s_{\ell}^{(t,w)}(y(t_i), \theta_i)$. Therefore, we are assured that the best possible decision is made given our prior models. Whereas with data that has no information content, the interpreted or conditioned model will create a proxy variable that poorly represents the true Earth response and will be quite dissimilar from each other. Therefore decisions made on “inaccurate interpretations” will lead to lower outcomes on average (as the $W$ very different outcomes $v_{\ell,a}^{(t,w)}(y(t_i), \theta_i)$ will be averaged) and will have a lower best alternative compared to interpretations that are more similar to each other. Better quality data will ultimately lead to higher valued decision outcomes and consequently, a higher VOI.
3.4.6. Generalization

Finally, the restriction of each location \( \ell \) being considered independently is eliminated, such that several actions can be taken simultaneously at several \( \ell \) locations. We introduce variable \( c \) which represents different possible decision and spatial combinations for the multiple \( \ell \) locations, with \( K \) total combinations deemed or identified as possible and valuable. The notation of \( g_{c,a} \) where \( a = \{a_1, \ldots, a_L\} \) denotes which action is taken at which location, with \( \ell = 1, \ldots, L \) including all the \( L \) possible locations that a decision action \( a \) may be made. Suppose that there are three locations where it is possible to take a decision action \( (L=3) \), two unique combinations of the spatial and decision actions are considered \( (K=2) \) and there are two possible decision actions \( (A=2) \) such that \( a \in \{\text{remove, don't remove}\} \). The two possible decision-spatial combinations are deemed \( c=1: \{a=1, a=1, a=2\} \) and \( c=2: \{a=1, a=2, a=2\} \). The notation \( g_{c=1} \) would indicate that farms at \( \ell = 1 \) and \( \ell = 2 \) would be removed, and the farm at \( \ell = 3 \) would not be removed; whereas \( g_{c=2} \) indicates that only the farm at \( \ell = 1 \) would be removed. With the ability to apply combinations of different actions, \( V_{\text{prior}} \) now becomes

Equation 21

\[
V_{\text{prior}} = \max_{c,a} \left( \sum_{i=1}^N \Pr(\Theta = \theta_j) \frac{1}{T_{0,j}} \sum_{t=1}^{T_{0,j}} v_c^{(t)}(\theta_j) \right) \quad c=1,\ldots,K \quad a = \{a_1, \ldots, a_L\} \quad a = 1,\ldots,A
\]

such that the best we can do with the present uncertainty is the highest valued spatial-decision combination \( v_{c,a} \) on average over all \( \sum_{i=1}^N T_{0,j} \) Earth models. The value with imperfect information becomes

Equation 22

\[
V_{II} = \sum_{j=1}^N \Pr(\Theta = \theta_j) \frac{1}{T_{0,j}} \sum_{t=1}^{T_{0,j}} \max_c \left( \frac{1}{W} \sum_{w=1}^W v_c^{(t,w)}(y(t),\theta_j) \right) \quad c=1,\ldots,K \quad a = \{a_1, \ldots, a_L\} \quad a = 1,\ldots,A
\]

If the proposed information, represented synthetically with \( d^{(t)} \), can constrain the results of the dynamic simulation function and subsequently the proxy decision variable, then this imperfect information may have value. The degree of “constraining” is measured by estimating \( \text{VOI}_{\text{imperfect}} = V_{II} - V_{\text{prior}} \).
4. Application of VOI calculation to the demonstration case

To keep the example concise and clear, we assume the only possible decision alternatives are concerned with which contaminant sources (e.g. farm) should be removed or not. Thus there are only two alternatives (A=2) which are remove or don’t remove. With these two alternatives, there are four possible outcomes at each surface location \( \ell \) depending if its vulnerability

\[
\text{cost}^{(t)}_{\ell,a} = \begin{cases} 
1 & \text{if } s^{(t)}_{\ell} > 0, a=1 (\text{remove}) \\
1 & \text{if } s^{(t)}_{\ell} = 0, a=1 (\text{remove}) \\
2 & \text{if } s^{(t)}_{\ell} > 0, a=2 (\text{don’t remove}) \\
0 & \text{if } s^{(t)}_{\ell} = 0, a=2 (\text{don’t remove})
\end{cases}
\]

where we express these outcomes in nominal, unitless costs. It is more straightforward to use costs for these outcomes, but the value definition will be expressed in terms of negative costs (value = -cost). The first two cost outcomes, when a farm is removed regardless of its vulnerability, are deemed a unit cost (possibly representing the cost of farmer compensation to stop using chemicals). The third cost outcome (when farm is not removed on an effective entry point) is assigned a cost that is twice as much as compensation, representing the environmental consequences. The last cost outcome has no cost as no removal is taken at a location deemed not vulnerable. Table 1 contains the VOI imperfect results with these cost outcomes in column 1. The results for the two different reliabilities are shown in the two rows. As expected, the data generated with the reliability of Figure 5 has less value than that of Figure 4, which has a better electrical resistivity contrast between the two lithologies. These VOI imperfect results assume that the remove or not decision can be made independently at each of the \( L = 7,776 \) surface locations of the model. It is assumed that a farm exists at each one of these locations and therefore, as seen in Equation 20, a sum is made over all the \( L \) locations, which account for the VOI imperfect results being on the order of 50 whereas the costs are 1 and 2. Again, if only certain combinations of farm removal were deemed possible, this could be accounted for using Equation 21 and Equation 22.
Table 1: VOI imperfect Results for Aquifer Vulnerability Demonstration Case

<table>
<thead>
<tr>
<th>VOI imperfect = V_{ill} - V_{prior}</th>
<th>Costs = {1,1,2,0}</th>
<th>Costs scaled by vulnerability: b=0.5</th>
<th>Costs scaled by vulnerability: b=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Reliability 1 (Figure 4)</td>
<td>57.56</td>
<td>4,259.4</td>
<td>4,522.0</td>
</tr>
<tr>
<td>Data Reliability 2 (Figure 5)</td>
<td>51.655</td>
<td>4,210.7</td>
<td>4,426.4</td>
</tr>
</tbody>
</table>

Now, we consider costs that utilize the magnitude of the proxy decision variable of vulnerability. The vulnerability measure represents the magnitudes of the estimated damage to the groundwater resources. We can reward decisions to remove farms when at vulnerable locations by representing the costs savings with negative costs scaled to the vulnerability.

\[
\begin{align*}
\text{cost}_{t}^{(t)} & = \begin{cases} 
- b \; s_{t}^{(t)} & \text{if } s_{t}^{(t)} > 0, a = 1 \text{ (remove)} \\
1 & \text{if } s_{t}^{(t)} = 0, a = 1 \text{ (remove)} \\
s_{t}^{(t)} & \text{if } s_{t}^{(t)} > 0, a = 2 \text{ (don't remove)} \\
0 & \text{if } s_{t}^{(t)} = 0, a = 2 \text{ (don't remove)}
\end{cases}
\end{align*}
\]

The first cost outcome scales to the negative vulnerability through some scaling factor $b$. The third cost outcome (when a farm is not removed on an effective entry point) is equivalent to magnitude of damage done to the aquifer. The second and third columns of Table 1 contain the VOI imperfect results for this cost structure, where the second column scales the negative vulnerability by 0.5 and the third by 1. Generally, by weighting more by the vulnerability, the value of imperfect information increases. And again, as expected, “good” data (Figure 4) has a higher VOI for all three cost outcome structures (Table 1). We may expect to see this difference increase for $W>2$, as we would be better representing the uncertainty in the data likelihood. In all cases, the VOI imperfect was positive. Therefore, under the assumptions made in the demonstration case, it would be a sound decision to purchase the TEM data as long as the cost of acquisition was less than the VOI imperfect.
For this demonstration case, we make a significant simplifying assumption that all sand is valley facies, and that vulnerability is driven only by the buried valley structure (i.e. no sub-grid cell features exist that affect aquifer vulnerability). This lithology-facies assumption is extended into the geophysical reliability measure, such that we assume TEM’s ability to distinguish sand/clay lithologies means it also can exclusively identify valley/non-valley facies. Therefore, the two generated likelihoods do not describe how the geophysical data discerns the complex geologic input parameter $\theta_i$ (describing the dimensions of the buried valley) but the geologic indicator of lithology. This is because of the nature of the decision itself. The demonstration case requires local flow response information which is not deterministically captured in the $\theta_i$ alone.

5. Discussion & Conclusions

VOI can be viewed as the interaction of three components: the reliability of the information, the chance of making a poor or suboptimal decision, and the magnitude of the decision. Our methodology addresses the reliability of the data by proposing comprehensive prior Earth modeling schemes along with rock-physics data likelihood relationships. Since our focus is an estimation of VOI imperfect for situations necessitating proxy decision variables, the second and third VOI factors are outside of the scope of this paper. Here, we assume sensitivity studies have been performed in with the geologist and the modeler collaborating to establish satisfactory prior models that identify which subsurface elements are most impacting the decision. Also, the VOI sensitivity to other elements such as the costs and/or values assigned to different outcomes of the decision would need to be addressed. Nominal economic values have been chosen for costs. Realistic costs could be obtained through experts (e.g. economists), but this economic modeling is not the purpose of this example.

Although the methodology is presented using the aquifer vulnerability demonstration case, it is flexible to other Earth Science modeling schemes and decision problems. The geologic parameter $\Theta$ could also describe the reservoir volume or fault structure type which could affect oil production and development decisions. Other possible applications that could apply this methodology are enhanced oil recovery scenarios with water injection or a geomechanical model for subsidence problems or to carbon capture at different locations within the subsurface. We can imagine how the dynamic simulation function could be modeled with streamline simulation or a geomechanical response to achieve location specific information.

However, the methodology could be computationally expensive depending on the dynamic simulation function. The methodology requires that the dynamic response be simulated on all prior and conditioned models to capture the proxy decision variable. For certain applications, more conditioned models may be needed to
acheive a pragmatic VOI\textsubscript{imperfect} measure. For this demonstration case, it was already computationally expensive to run 360 3D flow simulations for 20 years each. Additionally, this methodology does not consider the errors that could result due to limitations related to the volume of support or depth of investigation of the geophysical technique. The generation of synthetic datasets only relies on the imperfect relationship between the geophysical attributes and geologic indicators; it does not incorporate how the geometry or physics of the measurement may distort the message as well. The synthetic data does not capture the actual process of geophysical data inversion to get earth resistivities from recorded electromagnetic fields.

The last assumption that different combinations of actions can be evaluated on the same proxy variable also depends on the dynamic simulation function and the proxy decision variable. In the case of aquifer vulnerability, the proxy variable will allow us to reasonably evaluate the outcome of these combinations (removal and non-removal). An example where the proxy variable may not reasonably evaluate different combinations of decision alternatives is when the alternatives represent different pumping or injection at different locations and/or at varying rates. The interaction of these responses may not be captured in the proxy variable and may need to be model explicitly.

For some Earth science decisions, knowledge of the geologic input parameters or “model uncertainty” will aid in decisions (e.g. don’t drill when $\theta_i = \text{small}$ (uneconomic) oil reservoir). This work addresses situations when we can not rely on knowledge or interpretation of the general heterogeneity $\theta_i$ to make a sound decision. We have defined a VOI methodology that accounts for such situations. Since the decision is based on the simulated dynamic response on the Earth models, we introduced the proxy decision variable which contains location-specific results of the dynamic response. In conjuction with the proxy decision variable, we proposed that rock physics relations with geostatistical simulation to achieve conditioned proxy decision variables. Finally we described how these two approaches together can achieve a value of imperfect information metric for geophysical data. Specifically, we have shown how this methodology applies to our demonstration case of aquifer vulnerability.

References


