# Value of information analysis for subsurface energy resources applications*

Geetartha Dutta[1], Tapan Mukerji[1] and Jo Eidsvik[2]
1) Department of Energy Resources Engineering, Stanford University, USA
2) Department of Mathematical Sciences, NTNU, Norway

Corresponding author: Geetartha Dutta (gdutta@stanford.edu, 91-6001030958)

## Abstract

A computationally efficient method to estimate the value of information (VOI) in the context of subsurface energy resources applications is proposed. VOI is a decision analytic metric quantifying the incremental monetary value that would be created by collecting information prior to making a decision under uncertainty. The VOI has to be computed before collecting the information and can be used to justify its collection. Previous work on estimating the VOI of geophysical data has involved explicit approximation of the posterior distribution of reservoir properties given the data and then evaluating the prospect values for that posterior distribution of reservoir properties. Here, we propose to directly estimate the prospect values given the data by building a statistical relationship between them using regression and machine learning techniques. For a 2D reservoir case, the VOI of time-lapse seismic data has been evaluated in the context of spatial decision alternatives and spatial heterogeneity of reservoir properties. Different approaches are employed to regress the values on the data: Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) regress the values on the important linear combinations of the seismic data. Random forests regression (RFR) is employed to regress the values on a few features extracted from the seismic data. The uncertainty in the VOI estimation has been quantified using bootstrapping. Estimating VOI by simulation-regression is much less computationally expensive than other approaches that fully describe the posterior distribution. This method is flexible since it does not require rigid model specification of the posterior but rather fits conditional expectations non-parametrically from samples of values and data.

*Keywords: time-lapse seismic monitoring; reservoir development; value of information; simulation-regression.*

## 1. Introduction

Decisions involving energy resources are often characterized by high complexity, multiple objectives, uncertainties, long time frames and are capital-intensive in nature. Before making a critical decision, it might be worthwhile for the decision-maker to gather more information about the uncertainties. For example, in the case of subsurface energy resources, there usually exists a lot of uncertainty about the reservoir properties, as a result of which the future production of the reservoir is highly uncertain. Time-lapse seismic data is often acquired to reduce the uncertainty in the reservoir properties and thereby help in making better reservoir development decisions. It is particularly useful in identifying bypassed oil

*The short version of the paper was presented at ICAE2018, Aug 22-25, Hong Kong. This paper is a substantial extension of the short version of the conference paper.

and preventing early water breakthrough, and thus help in maximizing recovery and production efficiency, especially during the later stages of production. However, since time-lapse seismic data comes at a high cost, it is important to justify its cost by assessing the impact it could have on the decisions needed to be made. The value of information (VOI) is a decision-analytic metric that is suited for this purpose (Howard, 1966).

VOI is an estimate of the additional value that information brings to a decision situation. If the prospect values corresponding to the different decision alternatives are expressed in monetary units, the VOI gives a monetary estimate of the additional value of collecting information before making a decision. From a decision-analytic perspective, information is valuable not only if it reduces uncertainty, but only if the reduction in uncertainty helps in making better decisions and maximizing the value outcome. Time-lapse seismic data can reduce the uncertainty in reservoir properties like porosity, permeability and saturation, and thus help in making better reservoir development decisions such as well placement and control decisions. A VOI analysis of time-lapse seismic data can be performed in the context of a reservoir development decision to determine the highest amount that the decision-maker should be willing to pay to acquire the information.

Previous methods for VOI estimation of geophysical data have involved explicit approximation of the posterior distribution of reservoir properties given the data. A characteristic of VOI analysis of geophysical data is the spatial nature of the problem – the decision alternatives, the uncertain reservoir properties and the data proposed to be collected are all spatial variables (Eidsvik et al., 2016). In Houck (2007), a 1D model is used for VOI analysis of time-lapse seismic data and hence the spatial nature of the problem is not addressed. In Eidsvik et al. (2008), 2D models are used for VOI analysis of seismic amplitude data. However, the spatial dependence is modeled through simple Gaussian models which might not be very geologically realistic. In Trainor-Guitton et al. (2013), a VOI methodology is proposed for spatial problems incorporating complex geological spatial models simulated using multi-point geostatistics. To compute the VOI, they quantify the reliability of information or its likelihood by forward modeling the geophysical attribute from the prior models. In their workflow, they use the data to inform about the geological scenario rather than the spatial heterogeneity of the distribution of reservoir properties. In Barros et al. (2015), a VOI methodology for reservoir decisions is proposed which takes into account the stochastic variability in an ensemble of realizations.

The VOI methods discussed above are computationally very expensive as they require the approximation of posterior probabilities of reservoir properties for each possible data outcome. In this paper, a VOI methodology is proposed which seeks to directly estimate the conditional expectation of prospect values given the data outcomes, without approximating the posterior probabilities of reservoir properties (Eidsvik et al., 2017). This methodology of estimating VOI through simulation-regression has been used in the field of medicine (Heath et al., 2016; Strong et al., 2014). However, these applications do not have the spatial characteristics that are typical of VOI problems in subsurface energy applications.

This paper is organized as follows. In Section 2, a brief review of the applications of decision analysis (DA) and VOI in the energy industry is presented. The basic concepts and equations for VOI computation in typical subsurface energy applications are also provided in Section 2. Section 3 presents the simulation-regression methodology and provides a test of accuracy of the methodology by considering a univariate Gaussian problem. A discussion of how this methodology can be applied to high-dimensional problems is also presented. Section 4 illustrates the simulation-regression methodology using a VOI problem involving spatial heterogeneity of petroleum reservoir properties, high-dimensional data and spatial decision alternatives. Finally, in Section 5, some concluding remarks are provided.

# 2. Background

## 2.1. DA and VOI in the energy industry

DA has been extensively used in the energy industry since the 1960s. A detailed review of different DA techniques used in various application areas relating to energy and environmental modeling has been provided by Huang et al. (1995). It has been found that decision making under uncertainty, which consists of techniques like decision trees, influence diagrams and multiple attribute utility theory, is the most widely used class of DA techniques in energy and environmental modeling. The general application areas, in decreasing order of the number of publications, are energy planning and policy analysis, environmental control and management, technology choice and project appraisal, power-plant site selection, energy production and operation, and energy conservation. The type of energy with the highest number of applications is found to be electricity, followed by nuclear, oil and gas, and renewable energy. Most of these applications are in planning, policy and site selection. It has also been found that DA techniques, particularly influence diagrams and multiple attribute utility theory, have a lot of scope to be applied to energy and environmental projects or issues, as they are characterized by high complexity, multiple objectives, uncertainties, long time frames and capital-intensive nature.

Zhou et al. (2006) present an update to the survey by Huang et al. (1995). In the update, it has been noted that the number of publications on DA in energy and environmental modeling has almost tripled compared to that reported in Huang et al. (1995). It has been found that multiple criteria decision making techniques have increased in popularity. It has also been found that, in comparison to energy issues, energy-related environmental issues have increased in importance. Another significant observation made by Zhou et al. (2006) is that, after 1995, the share of applications of DA in renewable energy has increased while that in nuclear energy has decreased. A review of DA techniques applied to renewable and sustainable energy in the field of energy planning has been provided by Strantzali and Aravossis (2015). They observe that the number of publications involving DA in renewable energy has remarkably increased throughout the three decades from 1983 until 2014. The types of renewable energy analysed in the review include solar energy, wind energy, hydropower, biomass and geothermal energy. Recent applications of VOI analysis in renewable energy include Seyr and Muskulus (2016) who study the additional value of knowing failure rates in wind turbines for improved maintenance, Ødegård et al. (2017) who calculate the value of snow measurements in hydropower scheduling, and Witter et al. (2019) who analyze the value of learning subsurface scenarios for geothermal development.

## 2.2. VOI in high-dimensional problems

VOI problems in subsurface energy applications usually involve high-dimensional spatial variables. The VOI in the context of spatial decisions depends on three main factors: the prior uncertainty regarding the variables which affect the value outcomes of the decision, the decision situation comprising alternatives and prospect values, and the information reliability. In the case of time-lapse seismic data, reservoir properties such as facies, porosity, permeability, saturation, etc. affect the production and thereby the prospect values. These uncertain spatial variables are denoted by $\boldsymbol{x}$. Since $\boldsymbol{x}$ is very high dimensional, it is difficult to represent the distribution of $\boldsymbol{x}$ in analytical form, except for the special cases of multivariate Gaussian distribution (Eidsvik et al., 2016), or Gaussian mixture distributions. Therefore, the distribution of $\boldsymbol{x}$ is usually approximated by Monte Carlo sampling, thus representing the prior distribution of $\boldsymbol{x}$ as an ensemble of realizations $\boldsymbol{x}^1$, $\boldsymbol{x}^2$,…, $\boldsymbol{x}^B$. The decision alternatives are denoted by $\boldsymbol{a} = \{a_i : i = 1,2, \dots, n\}$, where the index $i$ is associated with spatial location.

The action or decision alternative $\boldsymbol{a}$ must be chosen from a set $A$ of all possible alternatives, i.e., $\boldsymbol{a} \in A$. The prospect values are functions of the particular realization and the decision alternative chosen, and are denoted by $v(\boldsymbol{x}, \boldsymbol{a})$.

Assuming a risk neutral decision maker, the prior value ($PV$) is defined as the maximum expected value over all the alternatives:

$$PV = max_{\boldsymbol{a} \in A} \left[ \int v(\boldsymbol{x}, \boldsymbol{a}) p(\boldsymbol{x}) d\boldsymbol{x} \right] \approx max_{\boldsymbol{a} \in A} \left[ \frac{1}{B} \sum_{b=1}^{B} v(\boldsymbol{x}^b, \boldsymbol{a}) \right] \tag{1}$$

where $p(\boldsymbol{x})$ is the prior probability distribution over $\boldsymbol{x}$.

If we have perfect information about what value the variable $\boldsymbol{x}$ would take, we would choose the optimal action for that value of $\boldsymbol{x}$. However, since the VOI calculation is done before actually collecting the data, the posterior value (PoV) with perfect information is computed by taking the expectation over all possible values of $\boldsymbol{x}$, as shown in Equation (2).

$$PoV(\boldsymbol{x}) = \int max_{\boldsymbol{a} \in A} \left[ v(\boldsymbol{x}, \boldsymbol{a}) \right] p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{B} \sum_{b=1}^{B} max_{\boldsymbol{a} \in A} \left[ v(\boldsymbol{x}^b, \boldsymbol{a}) \right] \tag{2}$$

Now, let us assume that time-lapse seismic data $\boldsymbol{y}$ is collected, which is tied indirectly to the reservoir properties by rock physics relations, such that $\boldsymbol{y} = f(\boldsymbol{x})$. In addition, the data $\boldsymbol{y}$ is at a lower resolution than the reservoir properties $\boldsymbol{x}$, and hence we cannot determine for certain the reservoir properties from the data, even if there is no noise in the data. For each realization $\boldsymbol{x}^b$, we can forward model the time-lapse seismic data $\boldsymbol{y}^b = f(\boldsymbol{x}^b)$. $\boldsymbol{y}^1, \boldsymbol{y}^2, \dots, \boldsymbol{y}^B$ represent the distribution of the data.

Then, the posterior value (PoV) with imperfect information is defined as:

$$PoV(\boldsymbol{y}) = \int max_{\boldsymbol{a} \in A} \left[ E(v(\boldsymbol{x}, \boldsymbol{a}) | \boldsymbol{y}) \right] p(\boldsymbol{y}) d\boldsymbol{y} \approx \frac{1}{B} \sum_{b=1}^{B} max_{\boldsymbol{a} \in A} E\left[ v(\boldsymbol{x}, \boldsymbol{a}) | \boldsymbol{y}^b \right] \tag{3}$$

where $p(\boldsymbol{y})$ is the marginal probability distribution over $\boldsymbol{y}$.

For a risk neutral decision maker, the VOI is given by the difference between the posterior value and the prior value (Bratvold et al., 2009):

$$VOI(\boldsymbol{y}) = PoV(\boldsymbol{y}) - PV \tag{4}$$

# 3. VOI computation by simulation-regression

## 3.1. Simulation-Regression methodology

In the simulation-regression methodology, the VOI is computed by simulating the model parameters, the data and the prospect values, and then regressing the prospect values on the data. The steps involved in this methodology are as follows:

a) Draw Monte Carlo samples of the model parameters ($\boldsymbol{x}^b$), like facies, porosity, permeability, etc., and generate the corresponding samples of data ($\boldsymbol{y}^b$) and prospect values ($v_{\boldsymbol{a}}^b$) for each alternative $\boldsymbol{a}$.

b) Regress the vector, $\boldsymbol{v_a}$, containing the samples of prospect values for each alternative $\boldsymbol{a}$ on the data matrix, $\boldsymbol{Y}$, the rows of which correspond to different observations and the columns to different data dimensions, to obtain the regression model $F_{\boldsymbol{a}}(\boldsymbol{Y})$.

c) Fit values using the regression model:

$\hat{v}_{\boldsymbol{a}}^b = F_{\boldsymbol{a}}(\boldsymbol{y}^b)$, which approximates the conditional expectation $E[v(\boldsymbol{x}, \boldsymbol{a})|\boldsymbol{y}^b]$.

d) The prior value is given by:

$PV = max_{\boldsymbol{a} \in \boldsymbol{A}} \left[\frac{1}{B} \sum_{b=1}^{B} v_{\boldsymbol{a}}^b\right]$

e) The posterior value with imperfect information is given by:

$PoV(\boldsymbol{Y}) = \frac{1}{B} \sum_{b=1}^{B} max_{\boldsymbol{a} \in \boldsymbol{A}} E[v(\boldsymbol{x}, \boldsymbol{a})|\boldsymbol{y}^b] \approx \frac{1}{B} \sum_{b=1}^{B} max_{\boldsymbol{a} \in \boldsymbol{A}} \hat{v}_{\boldsymbol{a}}^b$

f) Finally, the VOI is given by:

$VOI = PoV(\boldsymbol{Y}) - PV$

## 3.2. VOI in a univariate Gaussian problem

Considering a univariate Gaussian model, the accuracy of the VOI results obtained from simulation-regression is tested by comparing with the analytical VOI result for the Gaussian model. In this model, the uncertain variable of interest, $x$, is univariate and is distributed according to a standard normal distribution. The prospect value, $v$, is either equal to $x$ or 0, depending on the decision alternative chosen. The data, $y$, is measured with normally distributed noise with mean 0 and standard deviation $\tau$. Thus, we have

$$p(x) = N(0,1)$$

$$v = ax, a \in \{0,1\}$$

$$y = x + e$$

$$p(e) = N(0, \tau)$$

$$p(y) = N\left(0, \sqrt{1 + \tau^2}\right)$$

The decision alternative $a$ has to be chosen from $\{0,1\}$ such that the prospect value is maximized. The VOI for this decision problem is evaluated for noise standard deviation $\tau = 0.5$ as discussed next.

### 3.2.1. Analytical VOI solution

For this univariate Gaussian model, Equations (1) and (3) can be solved analytically to obtain the prior value and the posterior value respectively (Eidsvik et al., 2016).

Prior value, $PV = max\ [0, \mu]$ $\quad (\mu = \text{mean of } x)$

$= 0 \quad (\text{since } \mu = 0)$

Posterior value with information,

$$PoV(y) = \mu_w \theta \left(\frac{\mu_w}{r_w}\right) + r_w \phi \left(\frac{\mu_w}{r_w}\right)$$

where $\mu_w = 0$ and $r_w = \sqrt{\frac{1}{1+\tau^2}}$.

$\phi(z)$ denotes the PDF of the standard Gaussian and $\theta(z)$ denotes its CDF.

Then, the VOI is given by Equation (4):

$$VOI(y) = PoV(y) - PV$$

Using this analytical method, the VOI is found to be 0.3568 for $\tau = 0.5$.

### 3.2.2 VOI by simulation-regression

To compute the VOI by simulation-regression, three different regression techniques are used – linear regression, k nearest neighbors and cubic splines. The uncertainty due to Monte Carlo error in the VOI computations by different methods is tested by repeating the VOI computation 100 times for two different sample sizes (B = 100, 1000). The boxplot in Figure 1 represents the confidence intervals of the VOI, with the analytical VOI result represented by the horizontal line.

It is seen from the boxplot that the analytical result is captured by the simulation-regression results using all three regression techniques, and for both sample sizes. In addition, it is observed that with increasing sample size B, the uncertainty bounds get narrower, i.e. we can compute the VOI with more confidence.

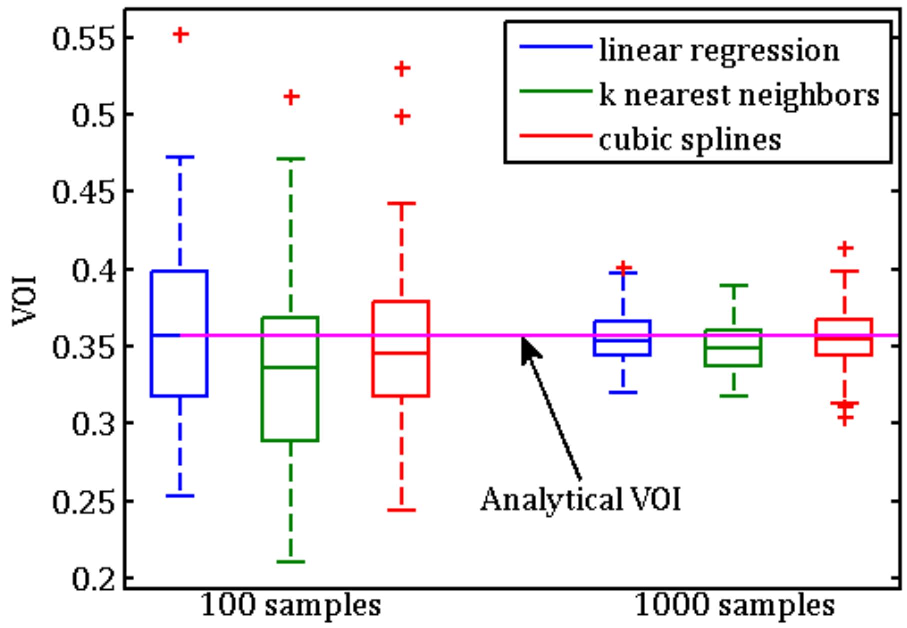

Figure 1: Boxplot representing confidence intervals of the VOI for sample sizes of 100 and 1000.

### 3.3. High-dimensional regressions

Since, typically in the case of time-lapse seismic data, the number of predictor variables (dimensions of the time-lapse seismic data) is much larger than the number of realizations, simple regression techniques like linear regression do not work. To perform the regression in this case, the dimensions of the data need to be reduced. In this work, two different approaches are employed:

a) Use Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) to regress the NPVs on a few components of the time-lapse seismic data.
b) Use Random Forests Regression (RFR) to regress the NPVs on a few predictive features extracted from the time-lapse seismic data.

PLSR is a regression technique which is frequently used for high dimensional regressions. Partial Least Squares (PLS) decomposes the data and the NPVs into score matrices and loading matrices such that the covariance between the scores is maximized (Rosipal & Kramer, 2006). PLSR assumes that the scores of the data and the NPVs are linearly related, such that the NPVs can be predicted using a linear function of the data scores. PCR is another regression technique that is appropriate for high dimensional regressions. One difference between PLSR and PCR is that PLSR takes into account both the data and the NPVs while computing the components, while PCR uses only the principal components of the data. In PCR, the NPVs are regressed on the principal components of the data using linear regression.

In some cases, it might be feasible to extract a few predictive features from the data and regress the NPVs on those features rather than the whole dataset. Non-linear regression techniques like RFR are suitable when there exists a highly non-linear relationship between the data features and the NPVs. Random forest is a type of ensemble learning technique which relies on an ensemble of decision trees to perform the regression. Individual decision trees typically have low bias, but high variance, and hence are not suitable for prediction. The idea of random forest is to greatly reduce the variance of trees, while keeping the bias almost the same, by averaging the predictions of a large number of uncorrelated trees (Hastie et al., 2009).

# 4. VOI for time-lapse seismic data in a 2D reservoir case

## 4.1. Decision problem definition

In this example, the VOI of time-lapse seismic data is evaluated in a 2D reservoir model, which has been producing for one year using one injector well and one producer well, in the context of a reservoir development decision with a finite number of alternatives. The reservoir development decision to be made at present consists of choosing one or two locations out of five possible locations for drilling the next producer well(s). This decision scenario is illustrated in Figure 2, with the locations of wells already drilled marked by gray circles, and those of prospective wells marked by black circles. Thus, the number of decision alternatives in this case is $C_1^5 + C_2^5 = 15$.

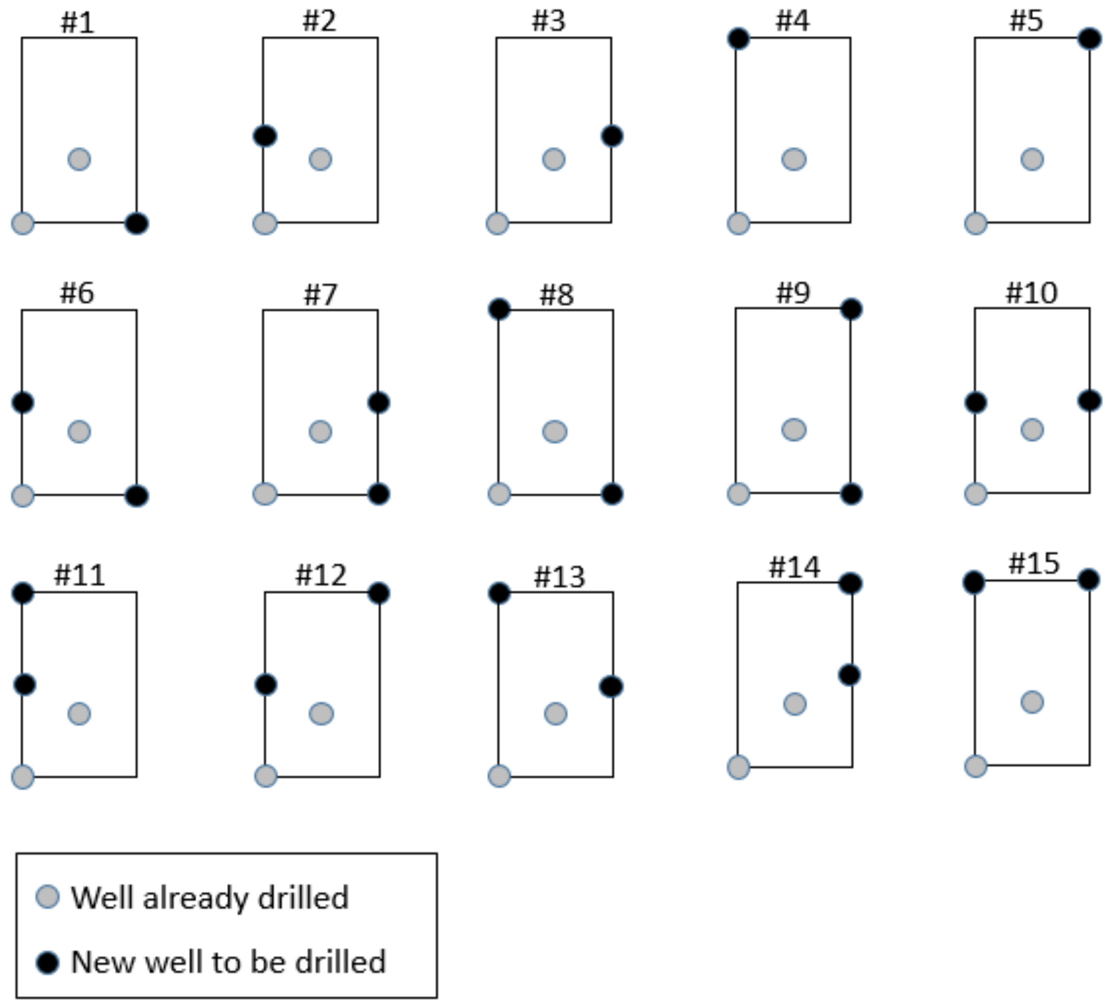


Figure 2: The well locations corresponding to the 15 drilling decision alternatives.

## 4.2. Modeling the data and the value outcomes

The uncertainty in reservoir properties that affects the value outcomes in this decision scenario is assumed to arise from stochastic variation within a channel geologic scenario. The facies realizations are first simulated using the multi-point geostatistical algorithm SNESIM (Strebelle, 2002) using a training image with two distinct facies – channel and floodplain. The porosity realizations are simulated conditioned to the facies using sequential Gaussian simulation. The permeability is then computed using the Kozeny-Carman equation (Mavko et al., 2009) which relates porosity to permeability. Figure 3 shows some facies, porosity and permeability realizations that are part of the prior set of realizations.

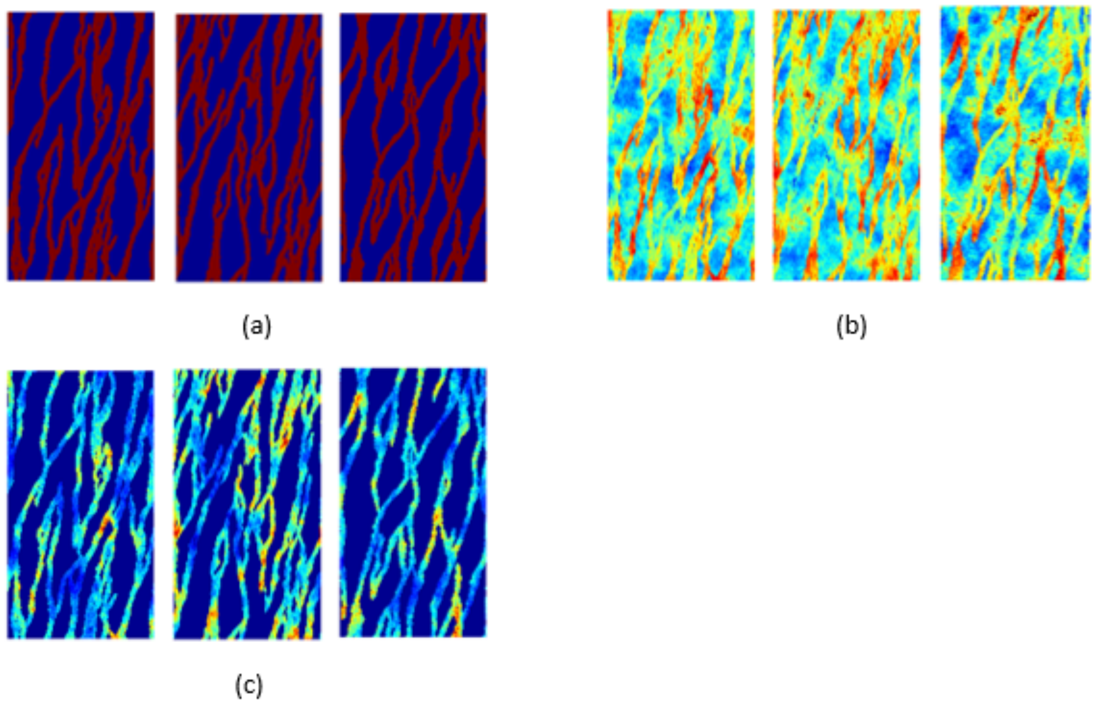

Figure 3: Three prior realizations each of (a) facies, (b) porosity, and (c) permeability.

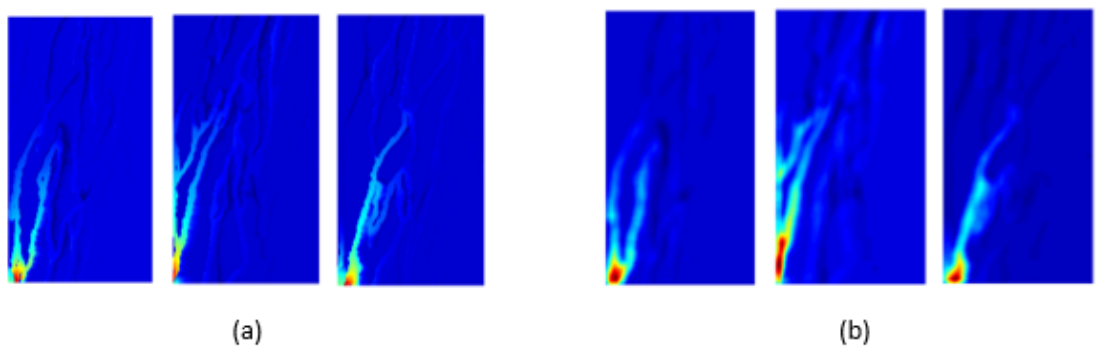

Figure 4: Change in AI at (a) the geostatistical scale, and at (b) the seismic scale for the three realizations shown in Figure 3.

To model the time-lapse seismic data, the acoustic impedance (AI) for each realization is modeled using the constant cement model (Avseth et al., 2000) and the Gassmann equations (Mavko et al., 2009) at both the initial time before production, and after one year of production. A moving mean filter is applied to the AI at the geostatistical scale to obtain the AI at the seismic scale, which approximates what can be obtained by inverting seismic data with a resolution of 60 m. The difference in the AI at the seismic scale between the two instances of time constitutes the time-lapse seismic signature. Figure 4 shows the AI at both the geostatistical scale and at the seismic scale for the three realizations shown in Figure 3.

Now, to compute the VOI using the simulation-regression method, the prospect value for each decision alternative corresponding to each realization has to be evaluated. The prospect values are given by the Net Present Value (NPV), which is a function of the oil production, the water production, the water injection and the cost of drilling wells. The NPV is calculated using the following equation:

$$V^b = \int_{t=0}^{T} \frac{q_o(t, \boldsymbol{x}^b) r_o - q_{wp}(t, \boldsymbol{x}^b) r_{wp} - q_{wi}(t, \boldsymbol{x}^b) r_{wi}}{(1+r)^t} dt - cost_{drill}$$

where $V^b$ = NPV for the $b^{th}$ realization, $\boldsymbol{x}^b = b^{th}$ realization of reservoir properties, $t$ = time, $T$ = producing life = 1 year, $q_o$ = oil production rate, $q_{wp}$ = water production rate, $q_{wi}$ = water injection rate, $r_o$ = price of oil produced = \$50/barrel, $r_{wp}$ = cost of water produced = \$5/barrel, $r_{wi}$ = cost of water injected = \$5/barrel, $r$ = discount rate = 10% per year, $cost_{drill}$ = drilling cost = \$100 million per well.

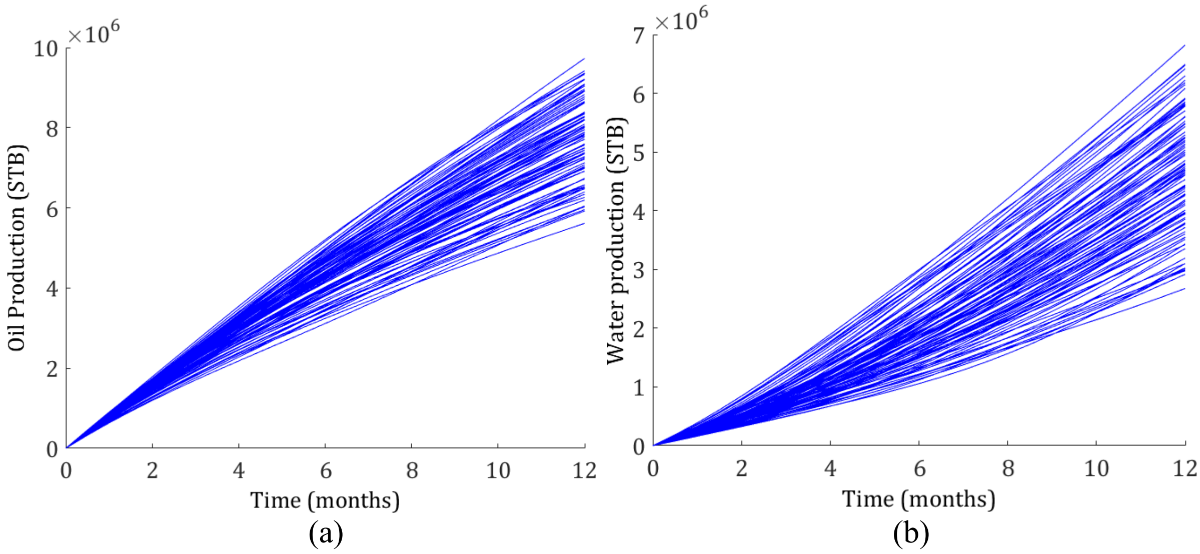


Figure 5: (a) The oil production profiles for all realizations for the first decision alternative, and (b) the water production profiles for all realizations for the first decision alternative.

Flow simulation is run on each realization for each decision alternative, and the NPVs calculated using the oil and water production rates. Figure 5 shows the oil and water production profiles for all realizations for the first decision alternative.

## 4.3. Value regression using PLSR and PCR

PLSR (Rosipal & Kramer, 2006) and PCR (Abdi, 2010) work well for high-dimensional data because they fit regression models in a reduced dimensional space by considering only a few components of the data. To determine the optimum number of PLSR and PCR components to use, leave-one-out cross validation is performed to evaluate the Predicted Residual Sum of Squares (PRESS) for varying number of components. The number of components that results in the minimum value of PRESS should be selected as the optimum number of components. Figure 6 shows the PRESS as a function of both PLSR and PCR components.

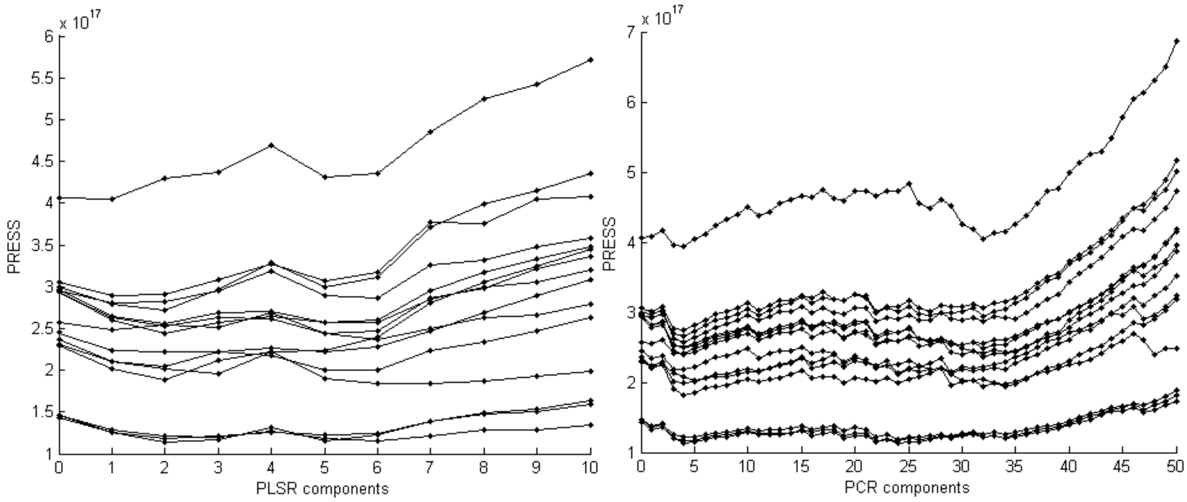


Figure 6: The PRESS as a function of the number of (a) PLSR components, and (b) PCR components. The different curves correspond to the 15 different decision alternatives.

From Figure 6, we see that for PLSR, the number of components that results in the minimum value of PRESS ranges from 2 to 6 for different decision alternatives. On the other hand, for PCR, most alternatives have a minimum around number of components = 30. So, PLSR models are built with number of components = {2,3,4,5,6}, and PCR models are built with number of components = {28,29,30,31,32}.

Since the number of realizations is much lower than the number of dimensions in the data, the uncertainty in the regression models is quite high, even for techniques like PLSR and PCR. Hence, bootstrapping is employed to estimate the uncertainty in the VOI computed using simulation-regression. The 10$^{th}$, 50$^{th}$ and 90$^{th}$ percentiles of the prior value are found to be \$295.0 million, \$300.4 million and \$308.1 million respectively. The corresponding percentiles of the posterior value with perfect information are found to be \$304.6 million, \$309.8 million and \$317.0 million respectively, and those of the value of perfect information to be \$6.7 million, \$9.2 million and \$12.1 million respectively.

The values for each bootstrap sample are fitted using the PLSR and the PCR models for that sample. Figure 7 shows the fitted values versus the observed values for one bootstrap sample using a PLSR model with 5 components and a PCR model with 30 components. It is seen that the fits are quite good, with a correlation coefficient of 0.94 for PLSR and 0.92 for PCR. The posterior value with imperfect information is then evaluated using the fitted values. The percentiles of the value of the imperfect time-lapse seismic data obtained from bootstrapping are shown in Table 1 for different number of PLSR and PCR components. It is seen that the VOI ranges from about \$2 million to about \$12 million overall. The VOI tends to increase with increasing number of components in both PLSR and PCR. Too few components could yield an overly smooth estimate of the conditional expected value, while too many components might overfit the conditional expected value to the value samples. Hence, cross-validation is performed to determine the optimum trade-off between underfitting and overfitting.

The uncertainty in the VOI estimates obtained from PLSR and PCR is quite high, as seen from the bootstrapping. This is because the number of dimensions in the time-lapse seismic data is much higher than the number of realizations. Therefore, in the next section, VOI estimation is performed by regressing the values on a few features extracted from the data.

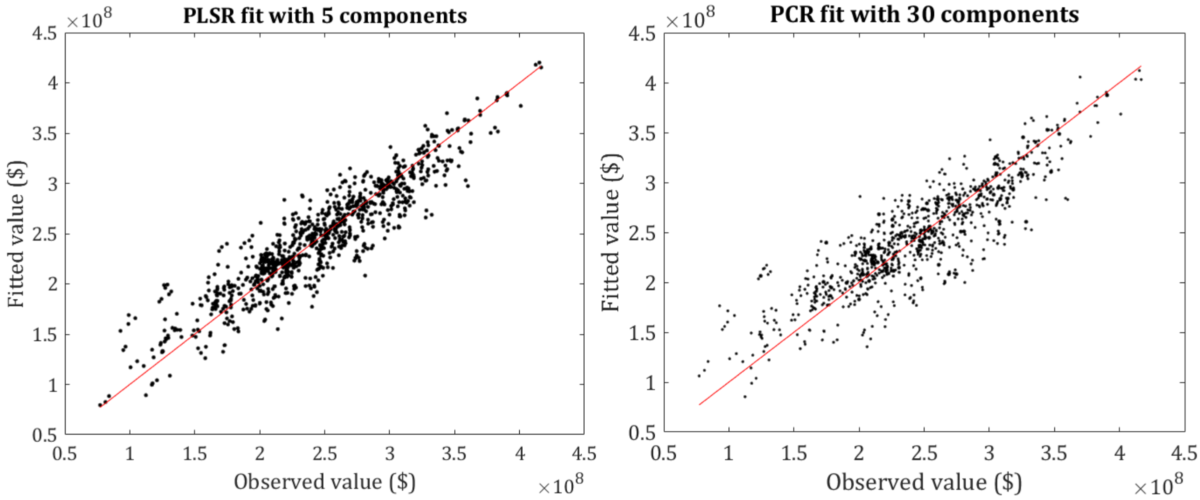


Figure 7: Plots of fitted values versus observed values for (a) a PLSR model with 5 components, and (b) a PCR model with 30 components. The correlation coefficient of the PLSR fit is 0.94 and that of the PCR fit is 0.92.

| **Regression technique** | **Number of components** | **VOI percentiles ($ million)** | | |
|---|---|---|---|---|
| | | **10th percentile** | **50th percentile** | **90th percentile** |
| PLSR | 2 | 2.1 | 4.6 | 7.8 |
| | 3 | 4.2 | 6.6 | 10.0 |
| | 4 | 4.6 | 8.2 | 10.9 |
| | 5 | 5.9 | 8.5 | 11.5 |
| | 6 | 6.7 | 9.0 | 11.8 |
| PCR | 28 | 4.8 | 7.3 | 10.2 |
| | 29 | 4.9 | 7.4 | 10.3 |
| | 30 | 5.1 | 7.5 | 10.5 |
| | 31 | 5.2 | 7.7 | 10.7 |
| | 32 | 5.3 | 7.8 | 10.8 |

Table 1: The VOI percentiles obtained by bootstrapping using PLSR and PCR models with different number of components.

## 4.4. Value regression on data features using random forests

The idea behind extracting features from the data is that we might not need the entire dataset to regress the values on, because most of the change in the AI happens only in a small region in the lower left corner of the AI change map as shown in Figure 8(a), and in most other regions the change in AI is almost zero. Six properties which characterize the region of high AI change in each realization are selected as features to regress the NPVs on. These properties are: area of the region, its perimeter, centroid, orientation of the ellipse that has the same second-moments as the region, and the mean and standard deviation of the change in AI in that region. To extract these features, the AI change maps are first converted to binary images by thresholding at the 90th percentile of the AI change. The 90th percentile is an arbitrary choice as the threshold, and we could have chosen any other reasonable threshold. Figure 8 shows a few AI change maps and the corresponding binary images obtained by thresholding. The six features are then computed using the regions of high AI change in the binary images.

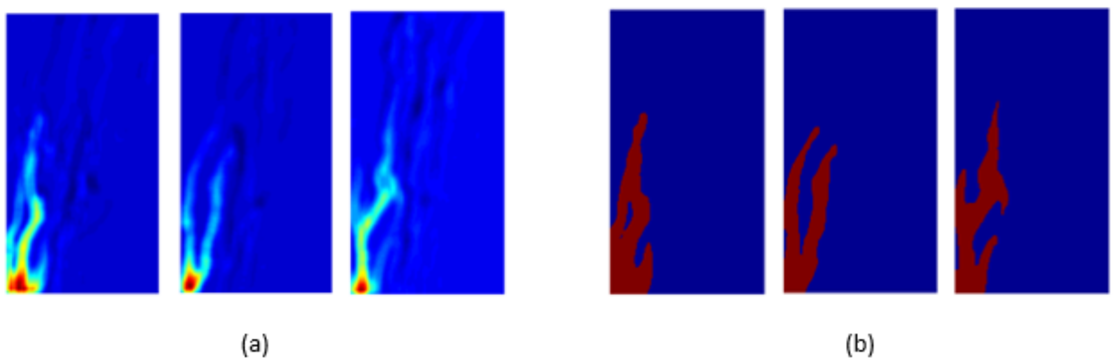


Figure 8: (a) Three AI change maps, and (b) their corresponding binary images containing a region of high AI change and a region of low AI change.

After extracting the features, they are used as regressors in a RFR model where the responses are the simulated NPVs. Since the NPVs are of the order of hundreds of millions of dollars, we might expect the uncertainty in the VOI estimate to be of the order of a few million dollars, even when using a very accurate regression model. Hence bootstrapping is employed to gauge the uncertainty in the VOI estimate. Figure 9 shows a plot of the fitted values using RFR versus the observed values for one bootstrap sample. It is seen that there is a high correlation between the fitted values and the observed values, with a correlation coefficient of about 0.98.

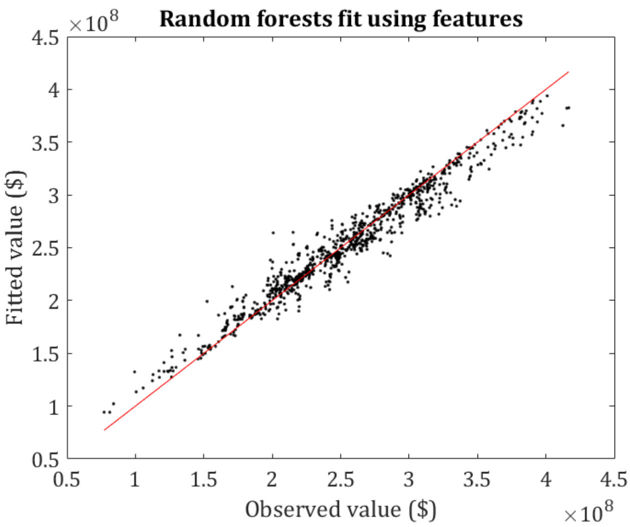


Figure 9: Plot of fitted values versus observed values obtained by a RFR model using features extracted from the data as regressors.

The fitted values from each bootstrap sample are then used to compute the posterior value with imperfect information for that sample. The estimates of the value of imperfect information thus computed from the bootstrap samples represent the uncertainty in the VOI. The 10$^{th}$, 50$^{th}$ and 90$^{th}$ percentiles of the VOI are found to be \$4.4 million, \$6.5 million and \$9.3 million respectively. This result is consistent with that obtained from PLSR and PCR in the first approach, as this VOI confidence interval is captured by those obtained from PLSR and PCR, which range from about \$2 million to about \$12 million. The VOI confidence interval obtained here is narrower than those obtained from PLSR and PCR because of the dimension reduction of the data by feature extraction. This results in a better fit which is manifested by the higher correlation coefficient between the fitted values and the observed values.

## 5. Conclusions

A computationally efficient and flexible simulation-regression methodology for VOI analysis in complex subsurface energy resources applications has been proposed. These applications are characterized by the spatial dependence of the uncertainty, as well as the spatial nature of information gathering and the decisions. The methodology has been illustrated using the case of VOI of time-lapse seismic data in the context of subsurface reservoir development. However, the methodology can be adapted to other spatial problems such as management of geothermal, hydroelectric energy, or geosequestration of $CO_2$ (see e.g. Eidsvik et al., 2016). In the example presented here, it is seen that the uncertainty in the estimated VOI for the high-dimensional spatial VOI problem is quite large, as shown by the bootstrap results. This is because the number of observations or realizations is very small compared to the number of predictor variables, and also because the values of the response variable, i.e. the NPV, are very high – of the order of hundreds of millions of dollars. In our case, the RFR model seems to give smaller uncertainties than PCR or PLSR, but the methods are comparable.

Even though uncertainties are large, the question of whether or not to collect the data before making the reservoir development decision might be easy to answer. In the case considered in this paper, it is advisable to collect the time-lapse seismic data if its cost is less than about $2 million. Estimating the VOI efficiently with a high degree of confidence using simulation-regression will lower the computational cost of VOI analysis a lot, especially for VOI problems involving sequential spatial decision-making and spatial data collection. Future studies on VOI in spatial decision problems would focus on sequential decisions as well as adaptive information gathering where the information gathering protocol is optimized as new information is obtained.

## References


Abdi, H., 2010, Partial least squares regression and projection on latent structure regression (PLS Regression), *WIREs Comp Stat*, v. 2, p. 97-106.

Avseth, P., Dvorkin, J., Mavko, G., & Rykkje, J., 2000, Rock physics diagnostic of North Sea sands: Link between microstructure and seismic properties, *Geophysical Research Letters*, v. 27, no. 17, p. 2761-2764.

Barros, E. G. D., Van den Hof, P. M. J., & Jansen, J. D., 2015, Value of information in closed-loop reservoir management, *Computational Geosciences*, v. 20, no. 3, p. 737-749.

Bratvold, R. B., Bickel, J. E., & Lohne, H. P., 2009, Value of information in the oil and gas industry: past, present, and future, *Society of Petroleum Engineers*, SPE 110378.

Eidsvik, J., Bhattacharjya, D., & Mukerji, T., 2008, Value of information of seismic amplitude and CSEM resistivity, *Geophysics*, 73(4), R59-R69.

Eidsvik, J., Mukerji, T., & Bhattacharjya, D., 2016, *Value of Information in the Earth Sciences*, Cambridge University Press.

Eidsvik, J., Dutta, G., Mukerji, T., & Bhattacharjya, D., 2017, Simulation-Regression Approximations for Value of Information Analysis of Geophysical Data, *Mathematical Geosciences*, v 49, no 4, 467-491.

Hastie, T., Tibshirani, R., & Friedman, J., 2009, *The Elements of Statistical Learning*, Springer-Verlag, 763 p.

Heath, A., Manolopoulou, I., & Baio, G., 2016, Estimating the expected value of partial perfect information in health economic evaluations using integrated nested Laplace approximation, *Stat in Med*, 35, 4264-4280.

Houck, R. T., 2007, Time-lapse seismic repeatability – How much is enough?, The Leading Edge, 26(7), 828-834.

Howard, R. A., 1966, Information value theory, *IEEE Transactions on Systems Science and Cybernetics*, v. 2, no. 1, p. 22-26.

Huang, J. P., Poh, K. L., & Ang, B. W., 1995, Decision analysis in energy and environmental modeling, *Energy*, v. 20, no. 9, p. 843-855.

Mavko, G., Mukerji, T., & Dvorkin, J., 2009, *The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media*, 2nd ed., Cambridge University Press.

Ødegård, H. L., Eidsvik, J., & Fleten, S. E., 2017, Value of information analysis of snow measurements for the scheduling of hydropower production, *Energy Systems*, 1-19.

Rosipal, R., & Kramer, N., 2006, Overview and Recent Advances in Partial Least Squares, *Subspace, Latent Structure and Feature Selection*, p. 34-51.

Seyr, H. & Muskulus, M., 2016, Value of information of repair times for offshore wind farm maintenance planning, *Journal of Physics: Conference Series*, v. 753, no. 9, IOP Publishing, 2016.

Strantzali, E., & Aravossis, K., 2016, Decision making in renewable energy investments: A review, *Renewable and Sustainable Energy Reviews*, 55, 885-898.

Strebelle, S., 2002, Conditional simulation of complex geological structures using multiple-point statistics, *Math Geol*, 34: 1-21.

Strong, M., Oakley, J., Brennan, A., 2014, Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: A nonparametric regression approach, *Med Dec Making*, 34, 311-326.

Trainor-Guitton, W. J., Mukerji, T., & Knight, R., 2013, A methodology for quantifying the value of spatial information for dynamic Earth problems, *Stoch Environ Res Risk Assess*, 27: 969-983.

Witter, J. B., Trainor-Guitton, W. J., & Siler, D. L., 2019, Uncertainty and risk evaluation during the exploration stage of geothermal development: A review. *Geothermics*, *78*, 233-242.

Zhou, P., Ang, B. W., & Poh, K. L., 2006, Decision analysis in energy and environmental modeling: An update, *Energy*, 31, 2604-2622.