# Uncertainty Quantification and Sensitivity Analysis of Reservoir Forecasts with Machine Learning

Jihoon Park (jhpark3@stanford.edu)

## Introduction

Successful development of oil/gas fields highly depends on informed decision making process. Such a decision making includes buying/selling the field, where/when to drill wells, how much oil should be produced per day and so on. Due to limited information about the subsurface a reservoir engineer builds multiple models (e.g. Monte Carlo simulations) and run flow simulations (reservoir simulations) to quantify uncertainty (UQ) of responses (e.g. barrels of oil produced for 20 years). Sensitivity analysis (SA) is performed in tandem with UQ which is to quantify how response uncertainty is apportioned to each uncertain model parameter. Knowledge obtained from SA is used in reducing complexity of models by fixing non-influential parameters, data assimilation and risk analysis.

One of challenges of UQ/SA is that a reservoir simulation is computationally expensive. For realistic UQ/SA a sufficient number of samples must be generated from prior distributions of model parameters and flow simulations need to be run to obtain responses. Nevertheless, even a single simulation may take several hours to days to complete. Another challenge is that a response is high dimensional (spatiotemporal) and a lot of SA method assumes that a response is univariate. In this project, functional PCA (FPCA) is applied to reduce the dimensionality of responses (time-series curve at each well, see Fig. 2) in order to overcome the difficulties stated above. Next, a regression model is built by taking uncertain model parameters as predictors (Table 1) and PCA components as reduced responses. For a regression algorithm, boosting with regression model is utilized. Then Global Sensitivity Analysis (GSA) is carried out to quantify the effect of each parameter on the response. GSA is known as a robust SA method for nonlinear system but it has been limitedly applied to reservoir forecasts due to large computations. It is demonstrated that the proposed method can achieve UQ/SA of multivariate reservoir forecasts with high computational efficiency.

## Related Work

Current practices of UQ/SA of reservoir forecasts focus on proxy modeling coupled with design of experiments (Zubarev et al., 2009; Jati et al., 2015; Arinkoola et al., 2015). They allow us to decrease the number of full forward simulations significantly but they have limitations. First, they assume that the response varies smoothly and it may not be true if a model parameter has a stochastically varying component (Caers, 2009). The case study in the project is a good example for this and it will be discussed later. Second if a behavior of reservoir is highly nonlinear, the number of models may not be sufficient.

When it comes to SA, local sensitivity analysis (LSA, Morris, 1991) is widely used in reservoir engineering (Dubey et al. 2016; Xiao-Hu et al., 2012). LSA computes sensitivities by perturbing each parameter ‘one at a time’ while other parameters are fixed. LSA is fast and offers easy interpretation but sensitivities are not valid when a behavior of system is highly nonlinear due to interactions between parameters. This necessitates the application of GSA which varies parameters globally over their domains of uncertainty (Sobol, 2001; Saltelli et al., 2008). One of drawbacks of GSA is its computational cost and it is often attacked by building a surrogate model of a flow simulator. Another disadvantage of GSA is that it assumes the response is univariate. As a result a response of interest is often confined to scalar value such as cumulative oil produced or net present value at a certain time. Nevertheless, reservoir responses are basically spatiotemporal. In order to tackle this problem some approaches compute sensitivities at all data points and sensitivities are represented as function of time (Helton et al., 2006; Herman et al., 2013). They are not only computationally intensive but also offers redundant information (Campbell, 2006). This is because sensitivities at time step $t_n$ is strongly correlated with the previous time step $t_{n-1}$. Therefore the goal of the research is to perform UQ/SA of multivariate responses with high computational efficiency.

## Dataset and Features

Dataset used in this project is based on a field scale oil reservoir located in central northern Libya (Ahlbrandt, 2001, see Fig. 1 for geological model). There are 10 uncertain parameters (features) and their parametric distributions are given in Table 1. In order to obtain training set, Latin hypercube sampling is performed to generate 1,000 ( $N = 1,000$ ) reservoir models ( $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ ) and responses $\{\mathbf{y}_{original}^{(i)}\}_{i=1}^{N}$ are obtained from a flow simulator (streamline simulation). Here, $\mathbf{y}_{original}^{(i)}$ is oil production rate (barrel/day) at three producing wells (therefore function of time and space), see Fig. 2.

Next, FPCA is applied to $\mathbf{y}_{original}^{(i)}$ . This includes decomposing $\mathbf{y}$ into linear combination of basis functions (Eq. 1) and applying PCA to the coefficients. (Ramsay, 2006; Grujic et al., 2015). B-spline is chosen as the basis function and principle components (PC) are taken as reduced response $\mathbf{y}_j^{(i)} (j = 1, \cdots, N_{PC})$ . Here, $N_{PC}$ is the number of PC taken and 7 PCs are taken because they explain more than 99% of total variance, see Fig. 3. 1st and 2nd PCs are displayed in Fig. 4. We can observe that there are three distinct groups of models which are consistent with Fig. 2. This is an example of abrupt changes in responses discussed in the previous section. In sum, a predictor is a sample of uncertain parameters generated from Monte Carlo simulations and responses are PCs.

$$\mathbf{y}_{original}^{(i)}(t) \approx \sum_{i=1}^{N_b} c_i \phi_i(t) \quad (t \in (1, T)) \tag{1}$$

where $\phi_i(t)$ : i-th basis function, $c_i$ : Coefficient of i-th basis function, $N_b$ : Number of basis function.

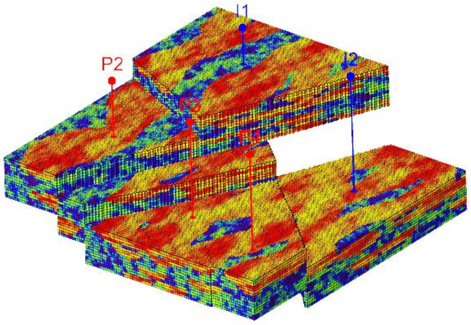


*Fig. 1 Reservoir model for the case study*

*Table 1 List of uncertain parameters (TM: Transmissibility Multiplier)*

| Num | Parameters | Abbrev. | Distribution |
|---|---|---|---|
| 1 | Oil-water contact | owc | U[-1076,-1061] |
| 2 | TM of fault 1 | mflt1 | U[0,1] |
| 3 | TM of fault 2 | mflt2 | U[0,1] |
| 4 | TM of fault 3 | mflt3 | U[0,1] |
| 5 | TM of fault 4 | mflt4 | U[0,1] |
| 6 | Residual oil saturation | sor | N[0.2, 0.05$^2$] |
| 7 | Connate water saturation | swc | N[0.2, 0.05$^2$] |
| 8 | Oil viscosity | oilvis | N[10, 2$^2$] |
| 9 | Corey exponent of oil | oilexp | N[3, 0.25$^2$] |
| 10 | Corey exponent of water | watexp | N[2, 0.1$^2$] |

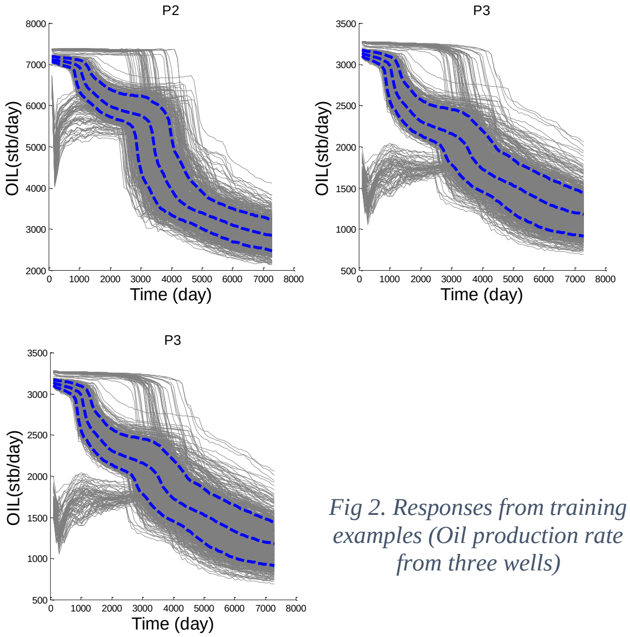


*Fig 2. Responses from training examples (Oil production rate from three wells)*

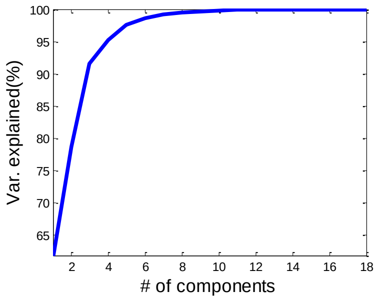


*Fig. 3 Cumulative variance explained*

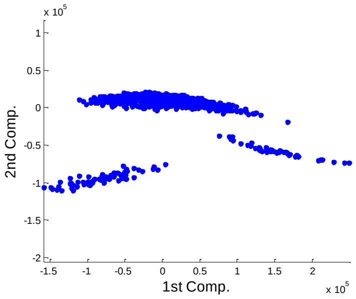


*Fig. 4 First two PCs*

## Methods

As a regression method, boosting with regression tree is utilized. Tree-based method or CART (Classification And Regression Tree) stratifies or segments the predictor space into a number of simple regions (Breiman et al., 1984; Hastie et al., 2008). For prediction a hypothesis decides the region which the given input belongs to and the average of response is provided as predicted value (regression). Suppose that we divide parameter spaces into $M$ distinctive regions $R_1, \cdots R_J$ . The goal is to find those regions that minimize the residual sum of square (RSS) in Eq. 2 (James et al., 2013).

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y^{(i)}-\hat{y}_{R_j})^2 \qquad (2)$$

where $y^{(i)}$ : i-th observation, $\hat{y}_{R_j}$ : mean of $y^{(i)}$ that belongs to $R_j$ .

In boosting the addition of elementary basis functions constitute a hypothesis (Hastie et al., 2008). The basis functions are shallow trees (stump) in this analysis. Loss function $L$ is taken as squared sum of errors. The number of trees is determined by cross validation (Fig. 5 and 6, left).

The algorithm for boosting with trees is offered from Hastie et al. (2008). The predictive rule is that if an input $\mathbf{x}$ is assigned to region $R_j$ a tree predicts the constant $\gamma_j$ for that region. This can be formulated as:

$$T(x;\theta)=\sum_{j=1}^{J}\gamma_j I(x\in R_j)$$

with parameters $\Theta=\{R_j,\gamma_j\}_1^J$ . $\Theta$ can be found by minimizing the empirical risk

$$\hat{\Theta}=\arg\min\sum_{j=1}^{J}\sum_{x^{(i)}\in R_j}L(y^{(i)},\gamma_j)$$

Constants $\gamma_j$ can be easily determined (usually mean). To obtain $R_j$ , the following suboptimal solution is computed from greedy-top down approach.

$$\hat{\Theta}=\arg\min_{\Theta}\sum_{i=1}^{m}\tilde{L}(y^{(i)},T(x^{(i)},\Theta))$$

Then boosted tree can be written as the sum of each tree (stump in this analysis).

$$h_\theta(x)=\sum_{k=1}^{N_{tree}}T(x,\Theta_k)$$

Taking the forward stagewise procedure, the following equation is solved at each iteration.

$$\hat{\Theta}_k=\arg\min_{\Theta_k}\sum_{i=1}^{m}L(y^{(i)},h_\theta^{(k-1)}(x^{(i)})+T(x^{(i)};\Theta_k))$$

Loss function is taken as the squared error loss as

$$L(y^{(i)},h_\theta^{(k-1)}(x^{(i)})+T(x^{(i)};\Theta_k))$$
$$=(y^{(i)}-h_\theta^{(k-1)}(x^{(i)})+T(x^{(i)};\Theta_k))^2$$

When a regression model is trained with trees the algorithm split the tree at the location where RSS is minimized. If a predictor is influential on response, the split will lead to large reduction of RSS. Therefore the sum of the reduction of RSS indicates variable importance. This corresponds to the concept of sensitivity and the reason why boosting with regression trees is chosen in the project is because the result of GSA can be validated by comparison.

GSA computes sensitivities from Monte Carlo sampling and it is based on the decomposition of variance of response (Sobol, 2001; Satelli et al., 2008). The basic idea is that if a model parameter is sensitive it will contribute to large portion of the variance of response. Two types of sensitivities are computed – first order sensitivity index $S_j$ and total effect index $S_{Tj}$. $S_i$ quantifies the main effect of single parameter $X_j$ on response $Y \in R$ without any interaction. Total effect index $S_{Tj}$ quantifies the effect of $X_j$ on $Y \in R$ including all related interactions. Two effects can be written as Eq. 3 and 4, respectively.

$$S_j = V[E(Y \mid X_j)] / V(Y) \tag{3}$$

$$S_{Tj} = 1 - V[E(Y \mid X_{\sim j})] / V(Y) \tag{4}$$

GSA has been utilized in various field of science and engineering since it yields robust and consistent sensitivities for nonlinear systems. The drawback of GSA is its computational expense. If we have $m$ samples with $n$ parameters, the number of forward runs required to estimate sensitivities is $m(n+2)$.

## Results and Discussion

Seven hypotheses are fitted since they explain more than 99% of total variance (Fig. 3). Fig. 5 and 6 (right) show the scatter plots between training example and predicted values with correlation coefficient (1st and 3rd PC). It is observed that the hypothesis shows significantly low training errors. To avoid overfitting the number of trees in boosting is determined by cross validation.

Next, another 1,000 samples (test set) are generated from the distribution specified in Table 1. This is to validate whether the fitted model offers valid uncertainty range for test sets. Because PC loadings were saved in the previous step, it is possible to compute the coefficients of basis functions in Eq. 1. As a result, curves of time series can be reconstructed and Fig. 7 shows the result. In order to visualize the uncertainty, P10, P50, and P90 of data at each time step are displayed (blue for training, red for test sets). It is observed that reconstructed curves show close approximation to original uncertainty, meaning UQ with the proposed method is valid.

With regression model obtained, both first order sensitivity and total effect indices are computed, see Fig. 8 (sensitivities of 1st PC). The sample size required to make sensitivities converge is approximately 50,000. Because there are 10 predictors the number of total simulations required is 600,000 ($m(n+2)$). In this example both indices do not show large difference which means that interactions are not significant. It is computationally infeasible if a full flow simulator is used. Nevertheless, with regressions the computation takes less than 30 mins.

The results show that 'big hitters' are *owc* and *oilvis* followed by *sor*. In order to verify the result of GSA, variable importance from trees is also computed, see Fig. 9 (top: 1st PC, bottom: 3rd PC). We can observe that variable importance from trees are consistent with GSA. This proves that the proposed method offers valid sensitivities from GSA with high computational efficiency.

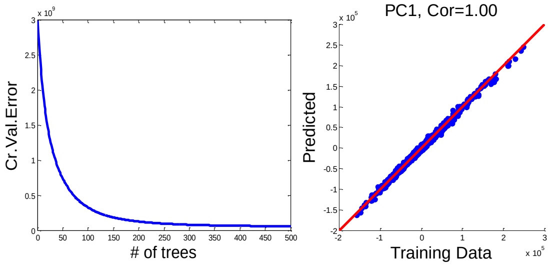


*Fig. 5 Cross validation errors with number of trees (left) and training data vs. predicted values (right) for 1st PC.*

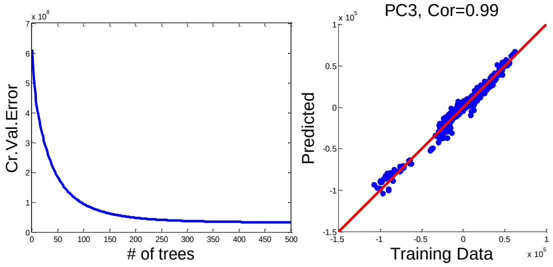


*Fig. 6 Cross validation errors with number of trees (left) and training data vs. predicted values (right) for 3rd PC.*

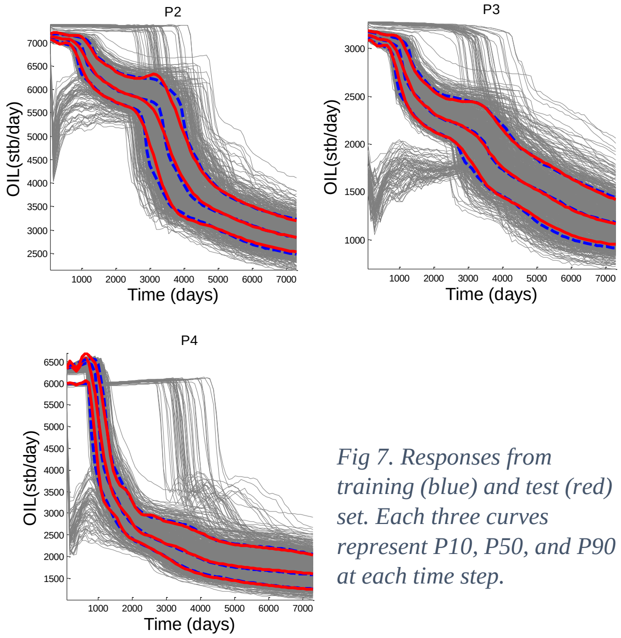


*Fig 7. Responses from training (blue) and test (red) set. Each three curves represent P10, P50, and P90 at each time step.*

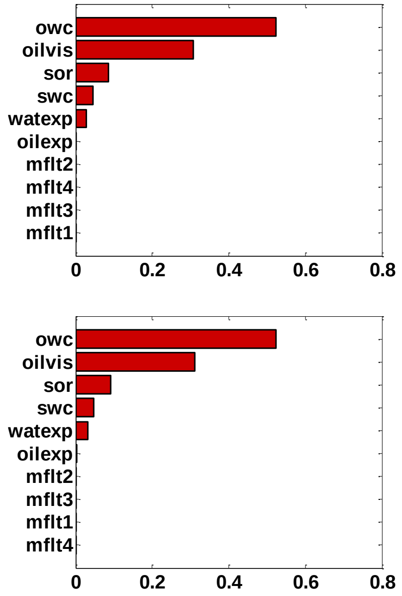
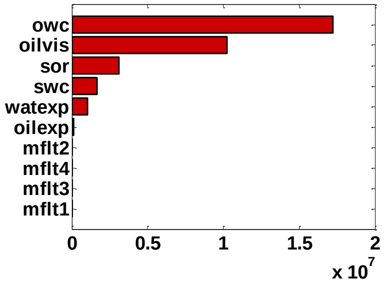


*Fig 8. First order sensitivity (top) and total effect (bottom) indices.*



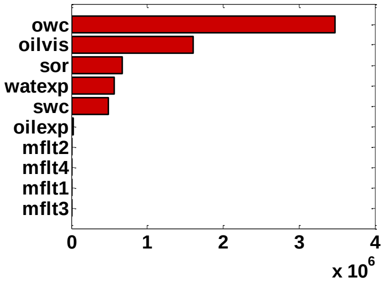


*Fig. 9 Variable importance for 1st PC (top) and 3rd PC (bottom)*

## Conclusion

In the project, the workflow for UQ/SA of reservoir forecasts with machine learning is proposed. By applying functional PCA to the response, high dimensionality of responses is reduced. Regression modeling by boosting with regression trees is performed to build a proxy flow simulator. It is demonstrated that the proposed method gives close approximation to full flow simulator with much faster computations. Regressors obtained from the analysis is used for GSA which requires a large amount of forward simulations. The results are validated by comparing with variable importances from trees. It is proved that the proposed method offers valid sensitivities with high computational efficiency.

## Future work

In the project, separate regression analyses are performed depending on the number of PCs. If a lot of PCs are needed to account for the variance this would be cumbersome. Therefore for the future work multivariate regressions will be applied to tackle this challenge.

## References

Ahlbrandt, T.S., 2001. The Sirte Basin Province of Libya: Sirte-Zelten Total Petroleum System. US Department of the Interior, U.S. Geological Survey Bulletin 2202–F.

Arinkoola, A.O. and Ogbe, D.O., 2015. Examination of Experimental Designs and Response Surface Methods for Uncertainty Analysis of Production Forecast: A Niger Delta Case Study. *Journal of Petroleum Engineering*, *2015*.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A., 1984. *Classification and regression trees*. CRC press.

Caers, J., 2011. *Modeling uncertainty in the Earth Sciences*. John Wiley & Sons.Dubey, P., Okpere, A., Sanni, G. and Onyeukwu, I., 2016, December. A Cost Effective, Fit-for-Purpose Single Well Producer-Injector Completion Strategy for Improved Recovery of Oil: Case Study in Niger Delta. In *SPE/AAPG Africa Energy and Technology Conference*. Society of Petroleum Engineers.

Grujic, O., Da Silva, C. and Caers, J., 2015, September. Functional Approach to Data Mining, Forecasting, and Uncertainty Quantification in Unconventional Reservoirs. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Hastie, T., Tibshirani, R., Friedman, J., 2008. *The Elements of. Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Helton, J.C., Johnson, J.D., Sallaberry, C.J. and Storlie, C.B., 2006. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, *91*(10), pp.1175-1209.

Herman, J.D., Reed, P.M. and Wagener, T., 2013. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resources Research*, *49*(3), pp.1400-1414.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning*. Springer.
Zubarev, D.I., 2009, January. Pros and cons of applying proxy-models as a substitute for full reservoir simulations. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Jati, N., Rahman, F., Kurniawan, H., Sari, Z.F. and Puspasari, S., 2015, October. Design of Experiment and Statistical Approach to Optimize New Zone Behind Pipe Opportunity: North Roger Block Case Study. In *SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition*. Society of Petroleum Engineers.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2), pp. 161-174.

Ramsay, J.O., 2006. *Functional data analysis*. John Wiley & Sons, Inc.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S., 2008. *Global sensitivity analysis: the primer*. John Wiley & Sons.

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, *55*(1), pp.271-280.

Xiao-Hu, D. and Hui-Qing, L., 2012. Investigation of the features about steam breakthrough in heavy oil reservoirs during steam injection. *Open Petroleum Engineering Journal*, *5*, pp.1-6.