Model Complexity: Initial Investigations

Orhun Aydin and Jef Caers

Stanford University

Abstract

Models that are computationally intensive are preferred over the ones that are simpler, in order to remain on the safe side in the decision making processes. Although simple models are generally avoided, they are more flexible than complex models, since they can be modified easier and quicker. In general practice, model complexity is studied to determine an optimal model size in order not to over-fit data at the same time representing the data sufficiently. Although model complexity is not studied extensively in the oil industry, there are numerous applications of model complexity in the areas of neural network design and model selection. Nevertheless, an important development in reservoir modeling making use of simpler models emerged with Top-Down Reservoir Modeling (TDRM). This pragmatic approach considers an overall workflow to incorporate reservoir uncertainty in model construction. In this paper a practical approach is taken to investigate model complexity in Earth models specifically. This is achieved by observing (dis)similarities in responses/outputs obtained from models by using smaller number of computationally intensive relations. Although simplifying Earth models is the purpose of this study, building over-complex or oversimplified models should avoided, in Albert Einstein's words "everything should be made as simple as possible but not simpler". In order to reach this, "simple enough" model, sensitivity of Earth models to the parameters that are used to create them should be known. For that purpose SGEMS is used to create earth models, with different inputs, such that effects of those inputs on the Earth models can be investigated. In order to observe effects of inputs on models, we propose to use information entropy functions by investigating the pattern variability at multiple scales. In this paper we provide some initial investigations into what seems to be a promising idea for future research.

Introduction

As easier fields are on the verge depletion and as the demand of energy increases everyday, energy companies are facing numerous challenges in modeling complicated settings in which they are forced to carry out their operations. Unfortunately, the first question that is tried to be answered by the modeler is about ways to deal with the complexities in the model that is created. The first question to be asked by the modeler should be about ways to ignore building overly complex models, better yet given the purpose of the study, how complex a model should be. The main trade-off in building a too complex and

a too simple model is about capturing features that are needed in the model for the stated decision purposes. In other words, a model should be complex enough so that it will be able to capture geological features that cannot be left out because they influence the decisions to be made, on the other hand it should be simple enough so that the model is manageable computationally and contains essential complexities. What do we mean by a model? In the case of energy industry, models represent fluid flow within a reservoir and models that include geological structures that make up a reservoir.

In the case of a geological model, a simple model can be thought as a homogeneous model, in which existing data is expressed in terms of continuous bodies, disregarding heterogeneities that are present within the reservoir and representing the whole reservoir by coarse grid cells. A complex model on the other hand, can be thought as a model that represents spatial variations in the shapes of the bodies present in the reservoir, bodies that are represented by irregular shapes instead of simple shapes such as ellipses and squares, together with models including local geological features such as compactions bands and small scale heterogeneities. Since these models are used to represent the geology for a given reservoir, it is crucial to capture needed features without overcomplicating the model beyond its purpose.

Despite the fact that, investigating model complexity can save the modeler a lot of time, together with decreasing computation needed to generate them as well as the response functions to be run on the model, model complexity is a topic which was not investigated in depth in the energy industry. BP's TDRM workflow (Williams et 2004) is such an example, but requires running flow simulations, which may be too expensive initially. Nevertheless, model complexity has various applications in the area of computer science and mathematics (Risannen,1989). In computationally demanding areas such as neural-network learning, model complexity is used to quantify the number of computational units needed to represent a given dataset (Kurkova, 2009). Numerous works have been done in neural network design (Floreen, Orponen, 1994), that investigates model complexity as a tool to avoid overfitting.

Model complexity is also used in mathematics, in model selection and especially modeling stochastic data. (Spiegelhalter et.al., 2002; Mackay,2003), describes the trade-off between fit to data obtained from a model and the number of free parameters needed to obtain that fit to data. In other words, the trade-off between using a relatively high number of parameters to build a model that fits the data closely and the one with relatively smaller number of parameters with a worse fit to data. In their paper they used Bayesian measures of fit and complexity (number of free parameters in this case) to compare models of arbitrary structure.

Besides mathematics and computer science, model complexity is evaluated in general by modelers from various disciplines. Since model complexity is an important issue in all of the modeling and simulation projects, advantages and disadvantages of model complexity is studied thoroughly by modelers. In their paper about simulation model complexity (Chwif et.al. 2000), Chwif et. al. state that main reasons for building complex models are unclear objectives and lack of understanding of the phenomenon to be modeled. Those two problems are generally the case in the energy industry as well, since there is

uncertainty and lack of understanding regarding geological models and starting to build those models without questioning the reasons to build those models.

In the energy industry, most of the decision are made on a coarse model which is an upscaled version of a finer model. In order to make decision over simplified models one want to make sure that these upscaled models represent sufficient complexity.

In this paper we focus on Earth modeling and investigate ways to determine the needed model complexity in Earth models in a simple and CPU effective way. Our approach is that of determining model complexity on a set of Earth models, not on a single model. Then we compute a so-called information entropy functions that captures the variability within this set. Determining which geological parameters determine most model complexity and hence should possibly preserved in coarser, i.e. less complex model representations. Thus we have two main type of parameters of investigation one being geological model parameters other being upscaling ration which is a decision parameter. An illustrative Boolean example is used to demonstrate our initial idea.

Methodology

Overview

A decision of upscaling of Earth models is made to decrease the computational time when applying a response function to those Earth models. However, upscaling models may change the response of Earth models.

Consider generating an Earth model. It requires several input parameters to generate such models. Parameters such as widths, lengths, variogram ranges, proportions, mean values are input into a geostatistical algorithm to generate reservoir models. These input parameters can be highly uncertain. Input parameters used to generate Earth models, together with upscaling ratio affects the response of Earth models. However, we do not know which parameters affect the response function the most. In order to explore the sensitivity of Earth model responses to given parameters a sensitivity study is needed. This sensitivity study will reveal how does the response of Earth models are changing with respect to an input parameter. Once an idea about sensitive parameters are obtained, relatively insensitive parameters may be used in their simplest form while keeping the sensitive parameters the same. We need to assess the impact of input parameters on a response function. In order to have input parameter combinations which are informative about how response changes with respect to a given parameter we use experimental design. Information entropy function is applied on the models generated with the parameter combinations obtained from experimental design. Finally sensitivity analysis is made on the models by using this response to detect the parameters that we can simplify.

Experimental design

At this initial part of the investigation we need a systematic way to determine combinations of input parameters to be used to generate Earth models. Experimental design is suitable for that purpose.

Effect of every parameter is captured by using Box-Behnken design without creating all possible combinations of input parameters, in the general class of 3*(k-p) design, Box-Behnken (Box, Behnken, 1960) is ideal for the purposes of this study. We will consider 3 states for each and every parameter, so we will have a low, a high and a medium value for each and every parameter with which we are going to build our earth models. By the use Box-Behnken design considering every possible combination of input parameters is avoided.

At this point of the study we have combinations of input parameters to be used for creating models of interest, so we will have the Earth models which are generated by using the parameter combinations using experimental design. In order to capture uncertainty for models which are created by the same input parameters we will have multiple realizations for every input parameter combination.

The next step is to apply a response function so that we can understand the effect of different input parameter combinations on the model response. In our study we have chosen information entropy to be the response function. Our reasons for that are elaborated in the next section.

Information entropy

A response function is needed to be used as a proxy for complexity of models. In this study, information entropy is applied to different Earth models. Which gives information about the geometrical pattern of a given Earth model.

In his 1948 paper Shannon defines information entropy function as the rate at which information is produced. Moreover, the entropy function H, gives the expected value of information. It also is a measure of the rate information is produced. Information entropy function is given by:

$H=\sum p_n * \log(1/p_n)$

Where, p_n is the probability of occurrence of an outcome of a random variable.

In order to have a continuous function for each Earth model, we need to establish a information entropy for an Earth model, which not only not a binary event, but also a multi-variate event as any Earth model is a gridded model with each grid cell possibly containing multiple variables. In this paper we limit ourselves to 2D/3D gridded models the basis for the comparison of Earth models and the sensitivity analysis to be carried out. Now lets take a look how the information entropy for different window sizes is calculated.

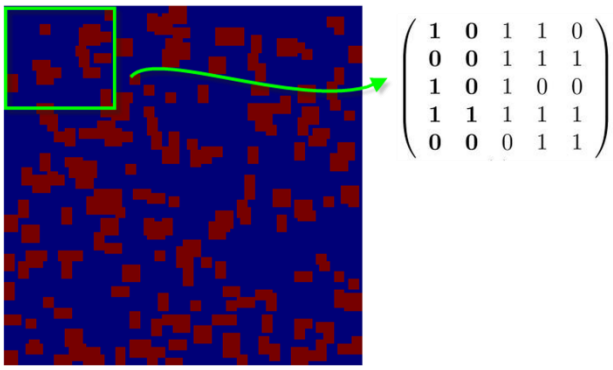

Figure 1: Obtaining Binary Matrix of the Earth Model for a Given Window Size

As seen in figure 1 for a pixel a neighborhood of pixels are considered. In this neighborhood, information entropy for the occurrence of the possible outcomes of pixel values are used to compute the information entropy. In our case the probability of occurrence of a pixel value is calculated by an image histogram.

Same procedure is carried out for different window sizes which in return gives a continuous function for each Earth model studied, just like the figure below:

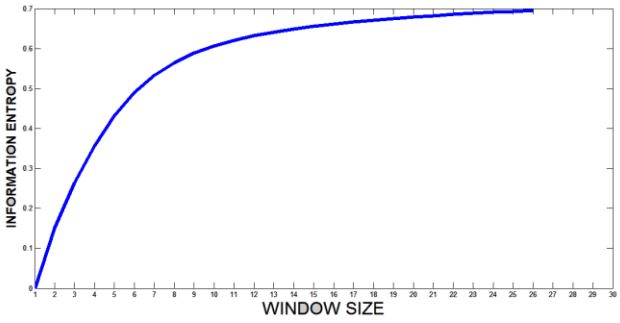

Figure 2:Information Entropy for an Earth Model for Different Window Sizes

As it can be seen in Figure 2, as window sizes increase we are capturing more and more patterns in the Earth model, which in return gives a constant information entropy after a certain window size is reached. For small window sizes since new patterns are observed with every increasing window size, information entropy increases relatively fast. Since for a given parameter set we are going to have multiple realizations, we need to consider how information entopy for all of them are changing, since they are going to represent the response of the parameter combinations used to make the model.

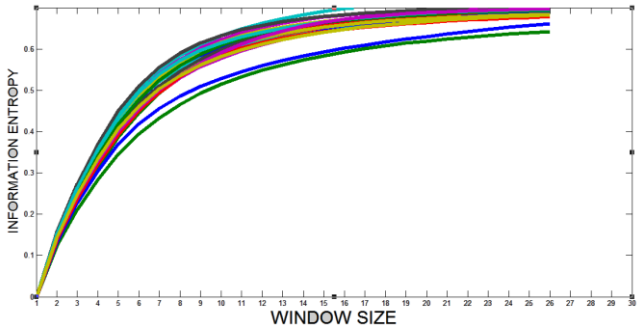

Figure 3: Information Entropy vs Window Size for 20 Realizations

As it can be seen in Figure 3, seperation for information entropy between different realizations are maximum for intermediate values of window size. In order to capture that, this range will be focused on by taking the average of information entropy for every realization at that range of window sizes. This will be carried out for all of the realizations generated by different input parameter combinations which brings us to a point where we have responses for our Earth models.

This concludes our response function discusssion. The next step is to determine parameters which have the highest influence on this response. This will be achieved by sensitivity analysis.

Sensitivity analysis

Sensitivity analysis is the last step in our discussion. We have obtained different earth models and for each and every Earth model we have calculated a response, which is average information entropy at given window sizes. Sensitivity analysis informs us about the effect of different input parameters on the model response. Most sensitive parameters are determined from this analysis. By this analysis we will be able to obtain parameters which affect the response of the earth models the most. This information will enable the modeler to have an idea about which parameters simplify and which parameters to keep as they are. Evidently, for parameters which have minor effects on the model response can be taken as simple as possible, in order not to waste any computational time on a parameter which is not going to change the result of simulations. In our case two sensitivity analyses will be carried out simultaneously. In the first part all of the parameters that we plan to include in our analysis will be used in the analysis. In the second part only the most sensitive parameters will be taken in the sensitivity analysis to determine how they react to different upscaling ratios. In order to do that we have to upscale our models and this is explained below.

Upscaling techniques

While models are upscaled to see the effect of the input parameters different scales the following scheme is used:

1-) Bicubic weighting is applied to decrease/increase the scale of the models. Bicubic averaging, outputs a pixel whose value is weighted average of pixels in 4-by-4 neighborhoods.

At this point we have an image with the desired scale however it is not a binary image. To convert it into a binary image Otsu's Method is used(Otsu, 1979).

2-) Otsu's Method is applied on a given intensity image to convert it into a binary image. With this method, a global threshold is computed. If the value of a pixel is smaller than this value it is assigned 0, if it is greater or equal than this value it is assigned 1.

It should be noted that this method preserves the proportion of different objects within a given image.

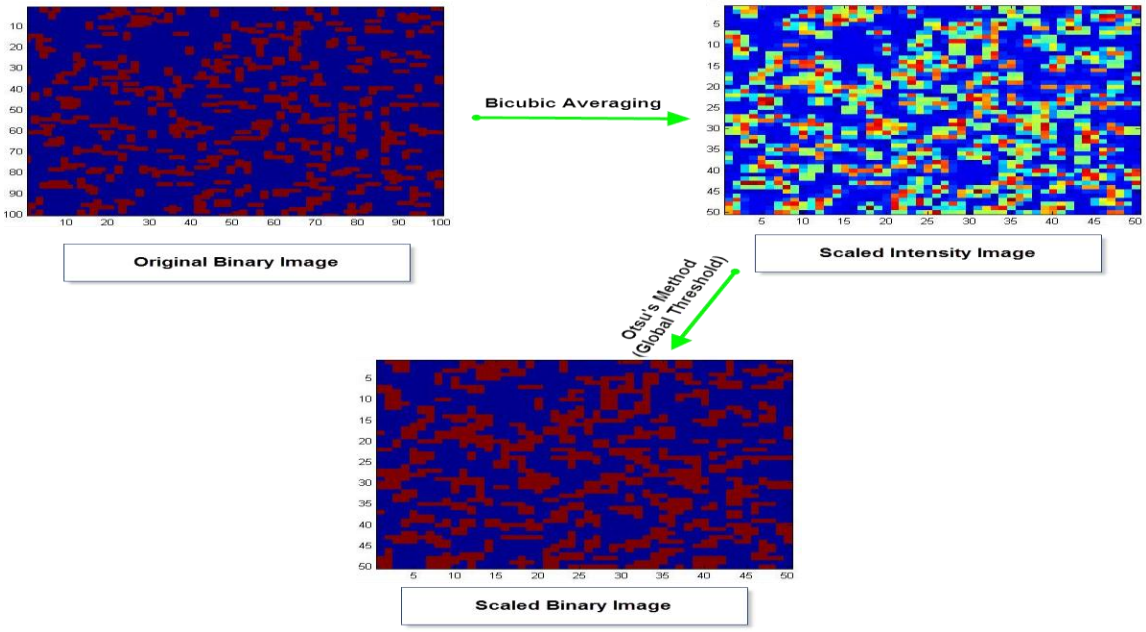

Figure 4: Methodology for Upscaling\Downscaling

Case study

For the case study, SGEMS is used to construct Earth models. 6 input parameters which are used to define objects in the Earth model are taken into consideration for the sensitivity analysis. For these 6 input parameters Box-Behnken design proposes 54 different combinations to be used in the generation of Earth models. Response function is calculated for all of these models and in the end sensitivity analysis is made in two phases. In the first phase of the sensitivity analysis all 6 input parameters are considered, after sensitivity analysis is carried out, most influencing input parameters are detected. Then a second round of sensitivity is made by following the same workflow, only the most sensitive parameters are included in the analysis in order to observe how they change with different upscaling ratios.

In the first stage all of the 54 Earth model sets (each containing 20 realizations) are used in the sensitivity analysis. After obtaining most sensitive input parameters from this initial set of Earth models, another sensitivity analysis is made just on the most sensitive parameters from the first phase of sensitivity analysis to observe how their sensitivity change with changing upscaling ratios. The following workflow is followed:

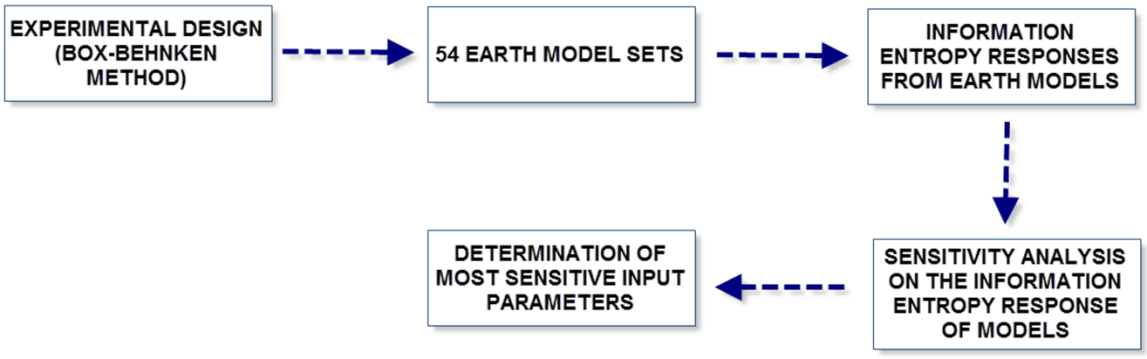

Figure 5: Workflow Followed

In summary, in the first round of sensitivity analysis, the workflow above is followed for all the parameters that we think are relevant. In the second round of sensitivity analysis, the workflow is again applied for the most sensitive parameters in the first round of sensitivity analysis.

The reader have came across with the terms all of the relevant parameters, in the section below we are defining which parameters are included and uncertainty associated with each of them.

Model of uncertainty

To define an Earth model objects are created and these objects are not known deterministically. Parameters in table 1 are the ones that are used for the first sensitivity analysis. It should be noticed that there are two groups of parameters. One being geological model parameters and the other is a decision parameter which is upscaling ratio.

Parameter Name	Low State	Medium State	High State
Interaction of Objects	No Overlap	Random Placement	Attached Objects
Trend	Trend in –X Direction	No Trend	Trend in +X Direction
Object Size	4X4	6X6	8X8
Proportion of Objects	0.20	0.30	0.40
PDF for Object Size	Triangular Distribution	Constant Distribution	Gaussian Distribution
Upscaling Ratio	0.2	1	2

Table 1: Input Parameters and Their Corresponding Values

As it can be seen from Tabel 1, 6 parameters are altered according to the states dictated by the experimental design. First five parameters in the table represents the uncertainty in the objects that we

create to construct Earth models. The last parameter, upscaling ratio, is the parameter for which we would like to see the sensitivity of other parameters on the response.

Sensitivity analysis results

Now the first round of the sensitivity analysis will be carried out. Sensitivity analysis carried out on the 54 Earth models with given data values are as follows:

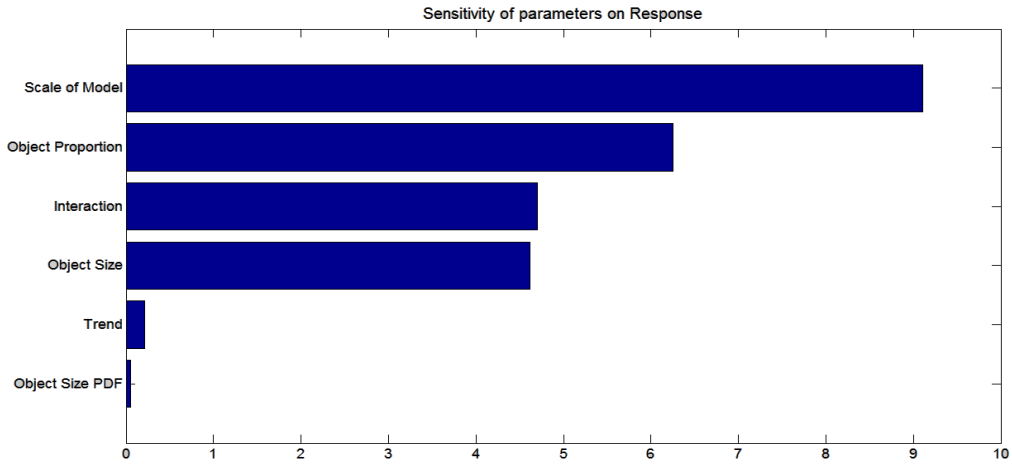

Figure 6: Sensitivity Analysis on Input Parameters

In this initial sensitivity analysis it is seen that spatial trends and object size pdfs are not affecting entropy response of Earth models. So, when building earth models they can be taken as simple as possible. In this case object size pdf can be taken as constant and trend can be set to no trend while creating Earth models.

Impact on upscaling

We now have came to the second round of the sensitivity study. After the initial study, several sensitivity analysis is made on proportion, interaction and object size for changing upscaling. We have designed another Box-Behnken experiment for 3 parameters. For that case a total 15 models are generated. Since we would like to see the change in these 3 parameters for different scales of models, we have generated those 15 sets for each and every upscaled model.

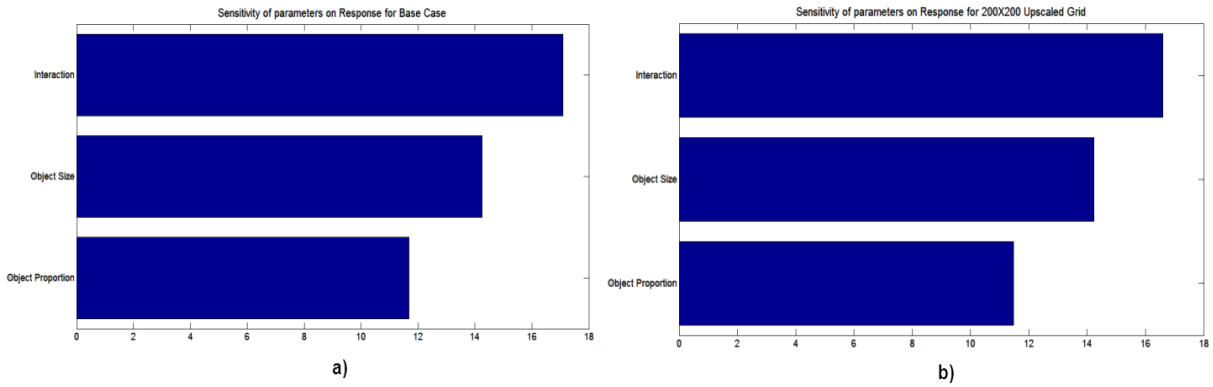

Figure 7: Sensitivity on Refined Input Parameters for a) 100X100 Grid b) 200X200 Grid(Scaled by 2)

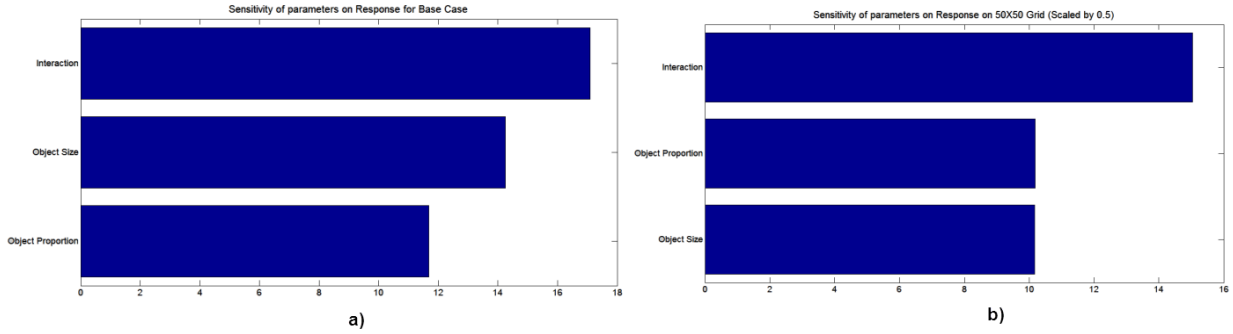

Figure 8: Sensitivity on Refined Input Parameters for a) 100X100 Grid b) 50X50 Grid(Scaled by 0.5)

As it is seen above object size loses its importance when models are upscaled to smaller proportions. Effect of object proportion increases a lot more when model sizes gets smaller and smaller. In the other case when models are made larger the effect of object proportion gets less important and object size becomes a more important parameter. However, in all of the cases interactions play an important role on the model response. They cannot be overlook to simplify models, so information about object interactions should always be preserved, however small alterations in object size and object proportion can give rise to tolerably small deviations from the original case.

Conclusions

We have applied our methodology to evaluate which parameters influence the complexity of the model. Complexity of a model is an important factor because if models are oversimplified decisions based on those model will be misleading. On the other hand, if too complex models are used, resources will be spent unnecessarily to make a decision.

We have seen in our object based modeling case, the most important factor to use is upscaling ratio and design parameters defining object geometry since they have the highest sensitivity for the response function that we have used. Placement of the objects is observed to have secondary importance.

The next step in our studies will be to evaluate if model complexity can be used to forecast a flow response.

References

Williams, G.J.J., Mansfield, M., MacDonald, D.G., Bush, M.D. (2004), Top-Down Reservoir Modelling, SPE

Rissanen, J. (1989). Stochastic complexity in statistical inquiry. World Scientific Series in Computer Science, 15, 03-07

Kainen, P., Kurkova, V., Sanguineti M.(2009), Complexity of Gaussian-radial-basis networks approximating smooth functions, 25: 63-74

Floreen, P., Orponen, P., (1994), Complexity Issues in Discrete Hopfield Networks

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002), Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64: 583–639

Mackay, D. (2003). Information theory, inference, and learning algorithms. Cambridge Press.

Chwif, L., Barretto M.R.P., Paul, R.J.: "On simulation model complexity," wsc, vol. 1, pp.449-455, 2000 Winter Simulation Conference (WSC'00) - Volume 1, 2000

Box, G., Behnken, D.: "Some new three level designs for the study of quantitative variables", Technometrics, Volume 2, pages 455–475, 1960

C.E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal, vol. 27, pp. 379–423, 623-656, July, October, 1948