Workshop on Visualisation in Environmental Sciences (EnvirVis) (2015)
A. Middel, G. Weber, and K. Rink (Editors)

Towards interactive definition of fast surrogate models for geochemical simulations using Visual Analysis

Janis Jatnieks[1], Mike Sips[1], Marco De Lucia[2] and Doris Dransch[1,3]

[1]GFZ German Research Centre for Geosciences, Section 1.5 Geoinformatics
[2]GFZ German Research Centre for Geosciences, Section 5.3 Hydrogeology
[3]Humboldt University of Berlin, Institute of Geography

Abstract

Geochemical models serve wide-ranging geoscientific applications for underground resource exploitation, aquifer remediation, gas storage and similar problems. Such models are time consuming to calculate. Replacing the full-capability simulation model, for each element of spatial discretization, with a fast surrogate is a promising acceleration approach for this problem. The balancing of speed and accuracy trade-off inherent to the surrogate modeling approach requires expert involvement and is best supported interactively. In this paper we argue that Visual Analysis has a prominent role to play for facilitating this process. It allows to involve expert knowledge regarding the specific characteristics of application scenario as part of the surrogate creation process. We describe the core problem of accelerating geochemical simulation by surrogate models and outline a strategy for approaching it with Visual Analysis.

Categories and Subject Descriptors (according to ACM CCS): J.2.8 [Computer Applications]: Physical Sciences and Engineering—Earth and Atmospheric Sciences I.6.5 [Simulation and Modeling]: Model Development—Modeling methodologies

1. Introduction

Geochemical simulation models are useful tools for understanding many geological and environmental processes involving chemical interactions between fluids and rocks [JH12]. Such models can be used for a wide spectrum of applications, such as aquifer remediation, nuclear waste confinement, extraction of mineral resources, evaluating the long-term evolution of underground CO_2 or natural gas storage, as well as geothermal systems [SAA*14].

Geochemical simulation models are used for simulating the transport of reactants because of fluid flow. This is called reactive transport and is crucial for studying the evolution of geochemical reactions in space and time. The spatial discretization of the geometry at a specific site is commonly represented by several elements. Reactive transport applications require the simulation model to be run for each element. Since the geochemical simulation models are complex, this results in high computational costs for reactive transport applications.

Reducing the computational cost of geochemical simulations is widely recognized as an important challenge in the application domain community [SAA*14, CKK10]. Many different strategies have been proposed for improving the run-times for such simulations. Among these acceleration approaches, surrogate modeling is acquiring increasing interest in the application domain community [ED15,CSM14]. The basic idea of surrogate modeling is to use a fast running approximation called surrogate model in place of a full-capability simulation model. The surrogate model is then used to compute chemical reactants for each geometry element.

The challenge with the surrogate modeling approach is balancing the trade-off between accuracy of the results and computational speed. Our close collaborations show that the inclusion of domain expert knowledge into the surrogate modeling approach could lead to better decisions about the appropriate speed/accuracy characteristics of surrogate models. We are convinced that this is a point where Visual Analysis can play a central role in pushing frontiers of an important domain application. The contribution of this paper is to

outline the key ingredients of a Visual Analysis strategy that is the result of frequent discussions with domain experts.

For each step of our strategy, we describe the important challenges for Visual Analysis that are interesting for a wider audience. The aim of this paper is to share our experience with the environmental visualization community. We do not describe a prototype implementation because we aim to address these challenges in continued research.

2. Problem statement

The main problem that we focus this paper on is the **reduction of run-times for geochemical simulations**. Exact speed benchmarks for geochemical simulation models warrant separate studies [CKK10], but to give a flavor of this problem, we provide an example from recent work of our collaborators. A groundwater flow model composed of 2950 elements without geochemical simulations runs for 2-3 hours. Adding coupled geochemical simulations with space and time dependencies increases the run-time of the simulations to several days. The model grid composed of 2950 spatial elements is already regarded as extremely coarse. For reference, the authority-required hydrodynamic simulation model geometry of the same site has 648420 elements. This level of detail is usually intractable for geochemical simulations [DLKK15]. For non-reactive hydrodynamic models the number of elements can easily range into millions [VBS*13].

This problem exists mainly because of two reasons. To support rock-fluid reactions under wide-ranging conditions requires considerable sophistication of the underlying simulation model. These reactions are contained in comprehensive databases [AP05]. Executing such a sophisticated model for many geometry elements dramatically increases simulation run times. The simulation usually produces many output variables (up to several hundred). This massively multivariate data from element-wise simulation model runs and the reuse of some outputs as parameters for nearby elements is the reason for the high computational costs of reactive transport simulations. Therefore, the reduction of run-times for geochemical simulations is of great interest for the application domain.

3. Approach and contribution

Instead of trying to simplify or accelerate an existing geochemical simulation model implementation, we contribute a Visual Analysis strategy for systematic creation of fast simplified models that are based on input-output data from the simulation model.

A promising way for accelerating geochemical simulations is using **surrogate models** in place of the simulation model. A surrogate model is a fast running approximation of the full-capability simulation model. It is a scientifically well established approach [Mül12, KM10]. In general, surrogate modeling can make the following contributions for geochemistry applications:

- using input and output data from the simulation model allows to exclude interdependencies in the simulation framework
- smoothing of numerical effects and missing results due to convergence problems

It became apparent that the construction of surrogate models from input-output relationships would consist of several important steps. An important advantage of Visual Analysis, that has also been recognized by our collaborators, is the ability to incorporate expert domain knowledge into surrogate modeling. This allows to facilitate better decisions about the accuracy/speed trade-off. However, the development of particular Visual Analysis approaches for this purpose requires a sound understanding of the challenges involved. Since we do not present prototype implementation, the contribution of this paper is to outline the important steps and discuss their associated challenges from a visualization perspective.

4. Related work

We are aware of the large and growing number of visulization contributions to the visualization areas of parameter space exploration [SHB*14], multivariate visualization [Cha06] and visualization of multifaceted scientific data [KH13]. An illustration of a very general pipeline for building surrogate models based on simulation model data is given in a recent survey by [SHB*14]. There is growing interest in applying such methods for speeding up complex simulations. Our focus in this paper is on understanding the opportunities of using Visual Analysis for the geochemical simulation domain. For this reason we focus on the domain-side developments and relate them to general Visual Analysis challenges.

Evidence of the large interest in surrogate modeling for geochemical applications is found in number of very recent publications that focus on speeding up geochemical simulation using surrogate modeling while using different terms in their communities, such as proxy models [JGL15], emulators [STD*12], meta-models [Roh14], reduced order models [Bac13] and response surface models [SOB13, KM10].

Different methods are used for constructing the surrogates in the existing literature. In [JGL15] the approach is called proxy modeling and employs functional principal component analysis for approximating the simulator response. The term meta-model is used in [Roh14] and the application is intended for acceleration of long-running flow simulators to enable global sensitivity analysis. The approach in [Bac13] employs non-linear regression with a quadratic polynomial for constructing the surrogate model from functional approximations of simulation results. Everything in this study is

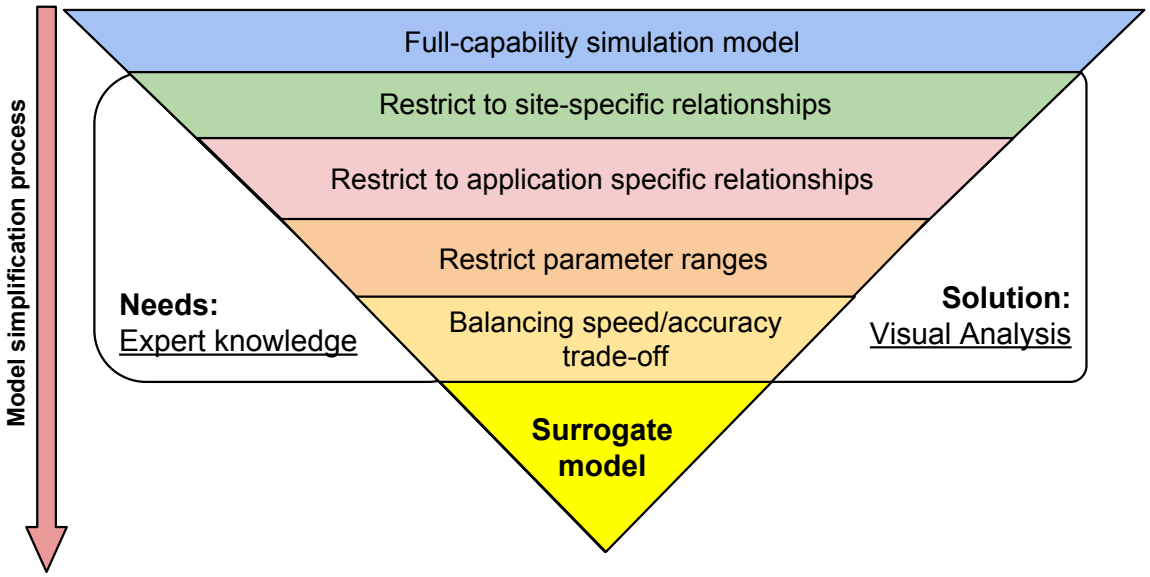

Figure 1: *Only site and application specific relationships and the appropriate parameter ranges are needed in the surrogate model. Approaching the process as Visual Analysis from the start and allowing the user to decide on the best speed/accuracy trade-off will allow to create the fastest approximation with acceptable accuracy.*

tailor-made for a specific site - Edwards Aquifer in south-central Texas. The Visual Analysis strategy in our paper is aimed at being able to systematically create surrogates from simulation model input and output data. The resulting visual approaches will enable creation of surrogates for different sites and application purposes. Reduced order model terminology is more commonly associated with finding more simplified mathematical formulations of complex numerical simulations [WVBHT12]. Genetic programming is used for the learning of surrogate model in [ED15]. The authors obtain accurate results in comparison to simulation model predictions at monitoring wells.

The response surface methods represent early ideas of surrogate modeling [KM10]. There are practical examples of classical response surface methodology utilized for solving modern geochemical problems [SOB13]. However, this becomes difficult above three dimensions and a typical geochemical simulation problem usually has many input parameters and even more output variables. Some recent contributions use the term "response surface" throughout their work [MS14, MKS14, Bac13]. However, this term is often used in abstract or statistical sense.

Surrogate models are used in many different application scenarios. Several recent studies aim to assist an optimization problem that needs to use results from geochemical simulations [CSM14, MS14, WVBHT12, STD*12]. The surrogate model approach in such studies is needed for speeding up the optimization task. In the visualization community surrogate models are often used for speeding up computationally expensive user-interaction steps. Tuner [TWSM*11] is designed for guiding the user to relevant parametrization for an image segmentation algorithm. In [BBP12] the surrogate model is used for computational steering of fluid dynamics simulations. HyperMoVal is a design that is focused on validation of surrogate models for an engineering application [PBK10].

Despite growing interest in applying surrogate model ideas for geochemistry applications, we are not aware of Visual Analysis approaches that support construction of surrogate models for geochemical reactive transport applications.

5. Visual Analysis strategy

We identified the main steps in the simplification process through a substantial collaboration with hydrogeology experts at GFZ German Research Center for Geosciences. Our collaboration has been established for more than one and a half year.

We started to analyze the operational aspects of using geochemical simulation models in the application domain. These operational aspects define the context that visual surrogate model creation approaches can take advantage of. These are:

1. geochemical simulations are often performed repeatedly for a specific study site
2. the site is associated with particular model parametrization

3. a particular application scenario is usually known beforehand and may require frequent changes to parametrization, even when used at the same site.

These operational aspects allowed us to distill the important steps of the simplification process. Figure 1 presents the important steps of simulation model simplification. These steps suggest a four step Visual Analysis process that captures the important parts of the observed expert work-flow. Each of these steps has additional complexities and challenges for interactive Visual Analysis.

Restrict to site-specific relationships. The usage for a simulation model is often tied to a particular study site (geographical model area). Model input parametrization is determined from the knowledge about the geology of the site and the prevalent environmental conditions, such as pressure, temperature, pH and others. There can also be additional human influences, such as operation of existing groundwater extraction wells, waste-water intrusions, industrial use of geothermal resources and others. Typically a number of input parameters (around 10-50) and output variables (up to several hundred) are involved in the simulation at each element and time step. Expert knowledge here is needed to identify only those input-output relationships that are meaningful for the particular site. Our collaborators would like to base this decision on their knowledge about the geological characteristics of their study sites.

This constitutes an important visualization challenge since it requires visualization to 1) tackle the exploration of many-to-many relationships to identify the important relationships in model input-output data and 2) provide an effective visual filter mechanism to support their inclusion in the surrogate model description.

Restrict to application specific relationships. Surrogate model is usually constructed for a particular application scenario. Therefore, it will be associated with a certain set of output values that are needed and a certain set of input parameters that control these output variables. For example, if a surrogate model is to be used for evaluating the feasibility of gas storage at a promising geological formation, then it needs to simulate the reactions for gases that could be of potential interest for storage in such formation. In contrast, a model for groundwater aquifer under an industrial area may need to include input-output relationships for rapid simulation of emergency spill scenarios.

Since some output variables may need to be re-used on the input parametrization side of nearby elements, the challenge here for Visual Analysis is to assist the expert user in balancing this consideration with the intended model application perspective and complexity.

Restrict parameter ranges. For each particular site and application scenario there is a limited set of meaningful input parameter ranges for the model. For example, temperature can be measured in a monitoring well at a certain depth and found to be in a limited range. If this is the case, then there is no need to incorporate all the possible model responses to temperatures outside this range. This requires a combinations of expert knowledge about the modeling area and a sound understanding of the model response to the sampled parameter ranges.

This presents several imporatant challenges for Visual Analysis as these considerations can be strongly tied to the surrogate model, the sampling strategy and the visual encoding of the relevant multi-parameter combinations.

Balancing the speed/accuracy trade-off. The various combinations of simplified input-output relationships, their functional approximation and relevant input parameter ranges can present a multitude of possible speed/accuracy trade-offs. Geochemical simulation results are often subject to high uncertainties, both on what is known about input parametrization and what can be validated in laboratory tests from the output variables [DHED12]. Many different formulations of surrogate models within these uncertainty bands may be possible. With expert involvement, some responses for relationships with high uncertainty on parameter or output side could be replaced by greatly simplified relationships while allowing overall model accuracy to stay appropriate. Expert understanding about the priorities of the scenarios for which the surrogate model is to be used can assist in better decisions about the most appropriate speed/accuracy trade-off.

Visual Analysis needs to deal with several challenges: 1) incorporating uncertainty visualization as part of the surrogate model definition process; 2) allowing interactive comparison of surrogate models with different speed/accuracy characteristics; 3) relating the various factors responsible for these characteristics so, that the surrogate model definition can be refined.

6. Conclusion

We present a four step strategy for Visual Analysis approach that can facilitate the creation of surrogate models for geochemical applications. We identify these four essential steps through a close collaboration with domain experts. Each of these steps require addressing difficult visualization challenges that are specific to the creation of surrogate models for geochemical simulations. The advantage of combining the surrogate modeling approach with Visual Analysis is the involvement of domain expert knowledge into the entire process. We look to develop practical prototypes that support this four step process. This will lead to Visual Analysis approaches that will push the frontier of the surrogate modeling approach for the application domain. These challenges are interesting for a wider environmental visualization audience because of the increasing applications of geochemical models and the growing interest in the surrogate modeling approach in the domain community.

References

[AP05] APPELO C. A. J., POSTMA D.: *Geochemistry, groundwater and pollution*. CRC press, 2005. 2

[Bac13] BACON D. H.: *Reduced-Order Model for the Geochemical Impacts of Carbon Dioxide, Brine and Trace Metal Leakage into an Unconfined, Oxidizing Carbonate Aquifer, Version 2.1*. Mar 2013. doi:10.2172/1166688. 2, 3

[BBP12] BUTNARU D., BUSE G., PFLUGER D.: A parallel and distributed surrogate model implementation for computational steering. In *Parallel and Distributed Computing (ISPDC), 2012 11th International Symposium on* (June 2012), pp. 203–210. doi:10.1109/ISPDC.2012.35. 3

[Cha06] CHAN W. W.-Y.: A survey on multivariate data visualization. Department of Computer Science and Engineering. Hong Kong University of Science and Technology. URL: http://www.cse.ust.hk/~wallacem/winchan/research/multivis-report-winnie.pdf. 2

[CKK10] CARRAYROU J., KERN M., KNABNER P.: Reactive transport benchmark of MoMaS. *Computational Geosciences 14*, 3 (2010), 385–392. doi:10.1007/s10596-009-9157-7. 1, 2

[CSM14] COZAD A., SAHINIDIS N. V., MILLER D. C.: Learning surrogate models for simulation-based optimization. *AIChE Journal 60*, 6 (2014), 2211–2227. doi:10.1002/aic.14418. 1, 3

[DHED12] DETHLEFSEN F., HAASE C., EBERT M., DAHMKE A.: Uncertainties of geochemical modeling during CO_2 sequestration applying batch equilibrium calculations. *Environmental Earth Sciences 65*, 4 (2012), 1105–1117. doi:10.1007/s12665-011-1360-x. 4

[DLKK15] DE LUCIA M., KEMPKA T., KÜHN M.: A coupling alternative to reactive transport simulations for long-term prediction of chemical reactions in heterogeneous CO_2 storage systems. *Geoscientific Model Development 8*, 2 (2015), 279–294. doi:10.5194/gmd-8-279-2015. 2

[ED15] ESFAHANI H. K., DATTA B.: Simulation of reactive geochemical transpor processes in contamined aquifers using surrogate models. *International Journal of GEOMATE 8*, 1 (2015). 1, 3

[JGL15] JOSSET L., GINSBOURGER D., LUNATI I.: Functional error modeling for uncertainty quantification in hydrogeology. *Water Resources Research* (2015). doi:10.1002/2014WR016028. 2

[JH12] JAMTVEIT B., HAMMER Ø.: Sculpting of rocks by reactive fluids. *Geochem Persp 1*, 3 (Jul 2012), 341–481. doi:10.7185/geochempersp.1.3. 1

[KH13] KEHRER J., HAUSER H.: Visualization and visual analysis of multifaceted scientific data: A survey. *Visualization and Computer Graphics, IEEE Transactions on 19*, 3 (2013), 495–513. 2

[KM10] KHURI A. I., MUKHOPADHYAY S.: Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 2 (2010), 128–149. 2, 3

[MKS14] MULLER J., KRITYAKIERNE T., SHOEMAKER C.: SO-MODS: Optimization for high dimensional computationally expensive multi-modal functions with surrogate search. In *Evolutionary Computation (CEC), 2014 IEEE Congress on* (July 2014), pp. 1092–1099. doi:10.1109/CEC.2014.6900599. 3

[MS14] MÜLLER J., SHOEMAKER C. A.: Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems. *Journal of Global Optimization 60*, 2 (2014), 123–144. doi:10.1007/s10898-014-0184-0. 3

[Mül12] MÜLLER J.: *Surrogate Model Algorithms for Computationally Expensive Black-Box Global Optimization Problems*. PhD thesis, Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology., 2012. URL: http://dspace.cc.tut.fi/dpub/bitstream/handle/123456789/21270/muller.pdf?sequence=3. 2

[PBK10] PIRINGER H., BERGER W., KRASSER J.: HyperMoVal: Interactive visual validation of regression models for real-time simulation. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 983–992. 3

[Roh14] ROHMER J.: Combining meta-modeling and categorical indicators for global sensitivity analysis of long-running flow simulators with spatially dependent inputs. *Computational Geosciences 18*, 2 (2014), 171–183. doi:10.1007/s10596-013-9391-x. 2

[SAA*14] STEEFEL C., APPELO C., ARORA B., JACQUES D., KALBACHER T., KOLDITZ O., LAGNEAU V., LICHTNER P., MAYER K., MEEUSSEN J., MOLINS S., MOULTON D., SHAO H., ŠIMŮNEK J., SPYCHER N., YABUSAKI S., YEH G.: Reactive transport codes for subsurface environmental simulation. *Computational Geosciences* (2014), 1–34. doi:10.1007/s10596-014-9443-x. 1

[SHB*14] SEDLMAIR M., HEINZL C., BRUCKNER S., PIRINGER H., MOLLER T.: Visual parameter space analysis: A conceptual framework. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 2161–2170. doi:10.1109/TVCG.2014.2346321. 2

[SOB13] SIMSEK E., ÖZDEMIR E., BEKER U.: Process optimization for arsenic adsorption onto natural zeolite incorporating metal oxides by response surface methodology. *Water, Air, & Soil Pollution 224*, 7 (2013). doi:10.1007/s11270-013-1614-1. 2, 3

[STD*12] SUN Y., TONG C., DUAN Q., BUSCHECK T., BLINK J.: Combining simulation and emulation for calibrating sequentially reactive transport systems. *Transport in porous media 92*, 2 (2012), 509–526. 2, 3

[TWSM*11] TORSNEY-WEIR T., SAAD A., MOLLER T., HEGE H.-C., WEBER B., VERBAVATZ J.-M.: Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (Dec. 2011), 1892–1901. doi:10.1109/TVCG.2011.248. 3

[VBS*13] VIRBULIS J., BETHERS U., SAKS T., SENNIKOVS J., TIMUHINS A.: Hydrogeological model of the Baltic Artesian Basin. *Hydrogeology Journal 21*, 4 (2013), 845–862. doi:10.1007/s10040-013-0970-7. 2

[WVBHT12] WISE J. N., VENTER G., BATTON-HUBERT M., TOUBOUL E.: Parameter optimisation in groundwater using proper orthogonal decomposition as a reduced modelling technique. In *5th International Conference from Scientific Computing to Computational Engineering (5th IC-SCCE)* (2012). 3