SIAM MPI23 Project Proposal—Model inversion for complex physical systems using low-dimensional surrogates

Pacific Northwest National Laboratory (PNNL) is enthusiastic about participating as an industry partner in the 2023 SIAM Mathematical Problems in Industry Workshop. We invite participants to collaborate with PNNL on developing data-driven, low-dimensional surrogate models of the observable response of complex physical systems and to use these surrogate models to calibrate parameterized computer models of the systems using field measurements. Calibrating computer models requires repeatedly querying the model for different inputs, which can be prohibitively expensive depending on the computational cost per query. This project aims to construct surrogate models of the observable response to substitute expensive model queries with cheap surrogate queries. To construct these surrogate models, we aim to exploit the low-dimensional structure of the map from model parameters to the observable response.

Industrial partner background. PNNL is a Department of Energy (DOE) research and development laboratory stewarded by the DOE Office of Science. The laboratory employs over 5,300 personnel across administrative offices and four research directorates—Physical & Computational Sciences, Energy & Environment, Earth & Biological Sciences, and National Security.

PNNL serves many missions and scientific endeavors, including

- Material science
- Nuclear and particle physics
- Weapons of mass effects
- Cybersecurity
- Energy grid
- Energy efficiency and storage
- Nuclear energy
- Nuclear nonproliferation
- Earth systems science
- Environmental management

among many others. Many of these applications focus on environmental science, including coastal and marine sciences and energy security.

Scientific background. Computer models of complex physical systems are often parameterized by one or many spatiotemporally heterogeneous scalar fields. Estimating these parameter fields is necessary to ensure the predictive capacity of these models. Directly measuring these parameter fields is often difficult or impossible, either because the measurement process is expensive and destructive or because the fields are effective quantities that cannot be observed directly. Therefore, it is customary to estimate these parameters by solving inverse problems in which we minimize the discrepancy between model predictions of the "observable response" of the physical system and field measurements of this response.

Inverse problems for complex physical systems are often ill-posed, that is, multiple parametric configurations can produce model predictions that match measurements up to observation error. Furthermore, they are

David Barajas-Solano, Mathematician, Pacific Northwest National Laboratory
(509) 375-3783, David.Barajas-Solano@pnnl.gov

often high-dimensional because representing heterogeneous fields requires a large number of degrees of freedom. Therefore, inverse problems are commonly formulated as model-constrained, regularized minimization problems, where regularization is introduced to address ill-posedness [2]. Solving these minimization problems requires repeatedly querying the model, which can be prohibitively expensive if the model's solvers are computationally costly. Consequently, it is customary to either reduce the dimensionality of the inverse problem or to replace model querying with a surrogate model of the observable response.

DOE is responsible for the Hanford Site, "one of the largest cleanup efforts in the world, managing the legacy of five decades of nuclear weapons production" [1]. Calibrating computer models of groundwater flow and contaminant transport at the Hanford Site is critical for evaluating remediation strategies and performing exposure assessments [3]. A network of sparsely-distributed observation wells (Figure 1) collects measurements of hydraulic pressure and tracer breakthrough curves from tracer experiments. These measurements can be used to estimate the heterogeneous parameter fields that govern groundwater flow and transport at the Hanford Site.

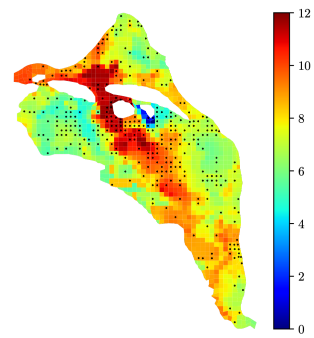

Figure 1: Log-hydraulic transmissivity field for a 2D confined aquifer model of the Hanford Site.

Project description. The observable response of complex physical systems as a function of the model parameters often exhibits a low-dimensional structure, which can be exploited to construct low-dimensional surrogate models of the observable response (e.g., [6, 5]). These surrogate models will allow us to simultaneously tackle the challenges of model reduction and computational cost. The goals of the project are:

- Identify a linear, low-dimensional embedding for the multivariate hydraulic pressure and tracer concentration response of a 2D model of the Hanford Site as a function of the hydraulic transmissivity.
- Use this low-dimensional embedding to construct a surrogate model of the observable response.
- Use the surrogate models to identify the hydraulic transmissivity field by solving a deterministic model inversion problem.

One way of identifying the low-dimensional structure of functions of many variables and constructing low-dimensional surrogate models is Basis Adaptation (BA) [5]. In BA, we construct a linear regression model of a scalar quantity of interest (QoI) as a function of the model parameters; this step is usually performed using sparse regression techniques. We then use this linear model to identify a linear transformation from parameter space to a single latent variable. Finally, we construct a nonlinear ridge function regressor for the QoI as a function of the latent variable.

BA is formulated for scalar QoIs, but it can be extended to multivariate QoIs by casting the associated linear regression problem as a "reduced-rank regression" (RRR) problem [4]. In RRR, we approximate the relation between a multivariate input and a multivariate output as a linear transformation represented by a rank-deficient matrix of coefficients. This transformation can be interpreted as a two-step procedure: (1) an orthogonal transformation from parameter space to a low-dimensional latent space, and (2) a transformation from latent space to the multivariate output. Once the transformation into latent space has been identified,

David Barajas-Solano, Mathematician, Pacific Northwest National Laboratory
(509) 375-3783, David.Barajas-Solano@pnnl.gov

we can train a nonlinear surrogate model for the multivariate QoI as a function of the latent variables.

Once a low-dimensional surrogate of the observable response has been constructed, the hydraulic transmissivity field can be estimated by solving a deterministic inverse problem. This inverse problem consists of finding the transmissivity field that minimizes the norm of the surrogate model's prediction minus measurements of the observable response. Such an inverse problem is ill-posed, and introducing a regularization penalty is required to obtain a unique solution. The regularization penalty encodes a priori information about the spatial distribution of hydraulic transmissivity, so it must be chosen carefully to ensure that the estimated field accurately approximates the conditions in the field. Some important questions to answer are:

- The BA procedure identifies the main direction of variability in parameter space for a scalar QoI. Can we identify the second most important direction of variability? The third most important? Is more information about the observable response necessary to identify additional directions?
- What is the sensitivity of BA and RRR to the size of the training dataset? Can it be improved with sparse regression techniques?
- What is a good regularizer choice for the Hanford Site transmissivity field?

Data description. The data to be used in this project consists of three datasets: (1) a generated set of pairs of model parameters and the corresponding simulated observable response, (2) the Hanford Site's 2D reference hydraulic transmissivity distribution, and (2) a set of synthetically generated noisy measurements of the observable response to be used for model inversion.

PNNL representative.

David A. Barajas-Solano
Mathematician, Computational Mathematics
Pacific Northwest National Laboratory
(509) 375-3783, David.Barajas-Solano@pnnl.gov

References

[1] Hanford Annual Site Environmental Report for Calendar Year 2021. DOE/RL-2022-08, 2022.
[2] D. A. Barajas-Solano and A. M. Tartakovsky. Approximate Bayesian model inversion for PDEs with heterogeneous and state-dependent coefficients. *J. Comput. Phys.*, 395:247–262, 2019. doi: 10.1016/j.jcp.2019.06.010.
[3] C. R. Cole, M. P. Bergeron, S. K. Wurstner, P. D. Thorne, S. Orr, and M. I. Mckinley. Transient Inverse Calibration of Hanford Site-Wide Groundwater Model to Hanford Operational Impacts-1943 to 1996. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA, United States, 2001.
[4] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *J. Multivar. Anal.*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.
[5] R. Tipireddy and R. Ghanem. Basis adaptation in homogeneous chaos spaces. *J. Comput. Phys.*, 259:304–317, 2014. doi: 10.1016/j.jcp.2013.12.009.
[6] R. K. Tripathy and I. Bilionis. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *J. Comput. Phys.*, 375:565–588, 2018. doi: 10.1016/j.jcp.2018.08.036.