IMWA 2026
Jeongseon

Efficient Data Assimilation for Reactive Transport Models using Data-Space Inversion

Rui Hugman[1], Pablo Ortega[2] and Jeremy White[3]

[1]*INTERA Geosciences Pty Ltd, Faro, Portugal, rhugman@intera.com, ORCID 0000-0003-0891-3886*

[1]*INTERA Geosciences Pty Ltd, Perth, Australia, portega@intera.com*

[1]*INTERA Incorporated, Fort Collins, USA, jwhite@intera.com*

Abstract

Coupled groundwater reactive transport models are essential for water management but computationally expensive, limiting their use in ensemble-based uncertainty quantification and data assimilation. This paper demonstrates a Data-Space Inversion (DSI) surrogate modeling approach to overcome this constraint. DSI assimilates observations and conditions forecasts using statistical relationships from a prior ensemble, completely avoiding repeated full-model evaluations. Tested on a 3D groundwater system with nonlinear geochemistry (based on the DIZON pilot project), the DSI surrogate captured the physics-based model dynamics and predictive uncertainty while reducing computational costs by orders of magnitude. This approach enables rapid, probabilistic scenario evaluation, successful integration of high-fidelity models into efficient decision-support systems.

Keywords: Coupled flow and reactive transport, surrogate models, model emulation, decision support, uncertainty, data assimilation

Introduction

Coupled groundwater flow and reactive transport models (RTMs) are primary tools for assessing groundwater quality and evaluating operational risks in mine water management. However, integrating field observations into these models through data assimilation or parameter estimation is computationally demanding. Traditional inverse modelling requires iteratively evaluating the forward model to minimize the misfit between simulated outputs and field data. Because resolving non-linear geochemical reactions and multi-species transport across large three-dimensional domains is inherently resource-intensive, the repeated model executions required for parameter estimation and uncertainty quantification are frequently cost-prohibitive. While various surrogate modelling techniques exist to reduce this burden, preserving the non-linear spatiotemporal dynamics of reactive transport while maintaining a rigorous mathematical framework for uncertainty quantification remains a challenge.

This paper applies a Data-Space Inversion (DSI) (Sun and Durlofsky, 2017) surrogate modelling approach to bypass these computational constraints. DSI operates entirely within the data space, utilizing statistical relationships among model outputs to assimilate observations and condition forecasts. By constructing a surrogate from a prior ensemble of full-order model simulations, DSI removes the need for explicit parameter estimation. Consequently, no further evaluations of the forward model are required during the conditioning or forecasting phases, reducing the computational burden substantially.

We demonstrate the application of DSI within a prediction-driven decision support workflow using a three-dimensional synthetic groundwater system characterized by variable flow conditions, multiple reactive species, and nonlinear geochemical reactions. This approach provides a practical method for integrating high-fidelity RTMs into decision support systems where computational limits otherwise restrict standard inversion techniques.

Methods

We demonstrate a DSI-based surrogate modelling approach for data assimilation and predictive uncertainty analysis using a numerical model based on the DIZON (Deep well injection in Zuid-Oost Nederland) field study (Prommer and Stuyfzand, 2005). The following chapter presents the theoretical background for DSI, a summarized description of the study case, the setup of the coupled flow and reactive transport numerical model and the emulator-based data assimilation workflow.

Data Space Inversion

Initially introduced in reservoir engineering literature by Sun and Durlofsky (2017), DSI is a surrogate modelling technique that relies on the statistical properties of a model's output rather than its internal parameters. The core concept of DSI assumes that a full-order physics model M, driven by a set of uncertain input parameters k, produces a comprehensive output vector d. This output vector can be partitioned into two distinct categories: historical variables that can be directly observed (o) and future predictions required for decision-making (p):

$$\mathrm{d} = \mathrm{M(k)} = \begin{bmatrix} \mathrm{o} \\ \mathrm{p} \end{bmatrix}$$

To avoid the computationally expensive and highly dimensional process of traditional parameter estimation, DSI relies on a prior ensemble of n_{real} model realizations. By evaluating the model across a wide distribution of prior parameters, DSI establishes a direct, empirical correlation between the observable data o and the forecasts p. This approach is highly advantageous when dealing with multiple conceptual models, highly parameterized spatial structures, or numerical frameworks that are difficult to calibrate conventionally. This empirical relationship is mathematically captured by extracting a data covariance matrix, C_d from the prior ensemble of model simulated outputs. Then, Singular Value Decomposition (SVD) is employed to project the high-dimensional data space into a simplified latent-space representation:

$$\hat{\mathrm{d}} = \bar{\mathrm{d}} + \mathrm{C_d}^{1/2}\mathrm{x}$$

In this formulation, $\bar{d}$ is the ensemble mean, and x represents a vector of latent-space variables drawn from a standard normal distribution. Any new output vector $\hat{d}$ generated through this latent projection preserves the exact statistical moments of the original full-order model ensemble.

The fundamental benefit of this DSI formulation is its immense computational efficiency during the inversion process. Instead of iteratively running the complex numerical model, field observations are used to directly condition the latent-space vector x.This allows for the near-instantaneous generation of posterior predictive distributions, effectively eliminating the primary computational bottleneck of traditional history matching.

In summary, implementing a DSI surrogate based workflow requires:

- Generating an ensemble of training data by simulating a physics-based model with different random sets of model parameters (usually a few hundred model runs).
- Training the surrogate on the ensemble of model outputs.
- Calibrating the surrogate parameters.
- Obtain a posterior predictive probability distribution.

Case Study: DIZON Field Study

The foundation of our model emulation workflow is the DIZON (Deep well injection in Zuid-Oost Nederland) field study, a benchmark Aquifer Storage and Recovery (ASR) experiment originally detailed by Prommer and Stuyfzand (2005). The pilot project provides a well-documented, complex dataset of non-linear geochemical interactions triggered by artificial recharge of surface water to a deep, isolated groundwater system by means of direct injection, making it an interesting testbed for evaluating surrogate modelling efficiency.

The conceptual model involves the injection of oxic, pre-treated surface water into a deep, strictly anoxic aquifer. The native sedimentary matrix contains appreciable fractions of reactive pyrite (FeS_2) alongside minor sediment-bound organic matter. The injection of oxygen- and nitrate-rich water initiates a cascade of transient redox

reactions, dominated by pyrite oxidation. This primary reaction pathway consumes the injected oxidants and produces sulfate, acidity (H^+), and ferric iron precipitates ($Fe(OH)_3$). A critical feature of the DIZON site is the seasonal temperature variation of the injected surface water, which heavily dictates the reaction kinetics. Cold water injection suppresses reaction rates, allowing oxidants to penetrate further into the aquifer, whereas warm water drives rapid oxidant consumption immediately adjacent to the wellbore. Models must account for these temperature-dependent kinetics to accurately reproduce observed field-scale oxygen breakthrough and nitrate removal patterns.

Coupled Flow and Reactive Transport Model Setup

To demonstrate the application of DSI, the DIZON conceptual model was translated into a coupled groundwater flow and reactive transport model. Groundwater flow and transport is simulated using MODFLOW6 (Langevin *et al.* 2026), dynamically coupled to PHREEQC (Parkhurst and Appelo, 2013) using the open-source software *mf6rtm*. The numerical model couples three-dimensional groundwater flow, heat transport, and multicomponent geochemistry.

Geometry of the site is represented with a rectangular model domain approximately 120 by 300 m. A Voronoi mesh was employed to allow for refinement near injection/recovery wells. The model is divided into 12 layers, with layer elevations interpolated from borehole log data. Thirty-nine stress periods are simulated using daily time steps, for a total simulation period of 854 days. The model domain, grid and boundary conditions are shown in Fig 1.

The injection and recovery wells are placed on the southern boundary of the model reduce the overall computational demand under the assumption of radial flow. Injection and extraction was applied to model layers corresponding to screened intervals. Initial hydrogeological parameters are assigned. Horizontal and vertical conductivity were assigned a value of 1 m/d. The system is simulated as fully confined, with specific storage of 1e-4 m-1. The following parameters are assumed constant across all layers: effective porosity of 0.35; longitudinal dispersivity of 0.1; transversal horizontal and vertical dispersivity 0.01 and 0.001, respectively. The specific geochemical reaction networks explicitly included in the modelling are detailed in Prommer and Stuyfzand (2005).

FOM History Matching and Uncertainty Analysis

History matching was undertaken using the iterative ensemble smoother implemented in PEST++IES (White *et al.*, 2018). An initial ensemble of 201 samples of parameters from the prior probability distribution was generated. Each realization is comprised of randomly sampled, geostatistical correlated, model parameters, each of which reflects the conceptual understanding of the system. Each

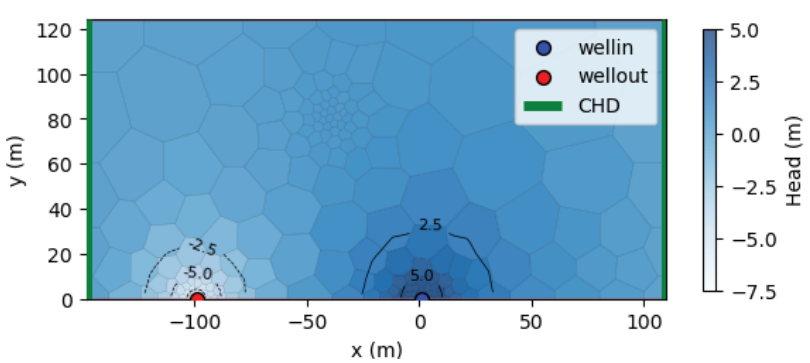

Figure 1 *Model domain, location of well boundary conditions and simulated spatial distribution of hydraulic head in layer 1after 50 days.*

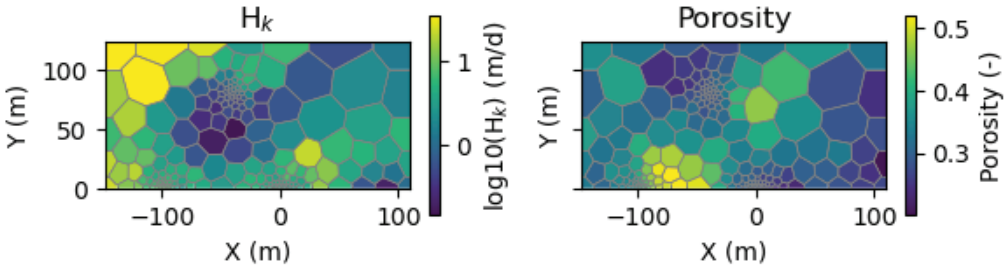

Figure 2 Example of spatial distribution of horizontal hydraulic conductivity (Hk) and porosity in model layer 6 for a single parameter realisation.

model layer has independent, spatially varying hydrogeological and geochemical parameters (e.g., hydraulic conductivity, specific storage, effective porosity), acknowledging horizontal and vertical heterogeneity and uncertainty. An example realisation of parameter spatial heterogeneity is shown in Fig.2.

A single realization was randomly selected as the "truth". Simulated values at observation wells using this parameter set were used as "measured" data during history matching. The realization was set aside and not included in the prior parameter ensemble or subsequent training data for DSI. Selected "measured" data for a few chemical species and physical parameters from the first 300 days of the simulation were used as calibration targets. The models' forecasting ability was evaluated by comparing simulated outputs to measured data during the rest of the study period.

As the synthetic model employed here is relatively fast to run (approximately 6 min), it is feasible to undertake history matching of the full-order model. Note that, in many real-world applications model computation time make this prohibitive. To compare outcomes and computational burden to the emulator-based workflow, we undertake history matching and predictive uncertainty quantification with both.

DSI History Matching and Uncertainty Analysis

A DSI surrogate model was trained on model output simulated using the same 200 parameter realisations from the FOM model prior Monte Carlo simulation. The surrogate algorithm is implemented in the open-source Python package *pyEMU*. Amongst other things, this package automates the configuration, training and set up of PEST++ for emulator-based workflows.

DSI latent parameters, x, were history matched using PEST++IES, with the same target observations obtained by simulating the "truth" model. As the computational cost is minimal (a few seconds), a large ensemble size is used (1000) to promote a more robust evaluation of predictive uncertainty.

Results

The synthetic model represents the introduction of pre-treated surface water into an aquifer, isolating how the injectate temperature and redox conditions (dissolved oxygen and nitrate) drive pyrite oxidation. The predictive variables of interest are the resulting spatial and temporal shifts in pH, iron, sulfate, and temperature. While this setup presents complex reactive transport dynamics, the primary focus of this study remains on evaluating the utility of DSI for data assimilation and uncertainty quantification. To this end, Fig 3 presents time series for selected variables, plotting the synthetic calibration targets against the prior and posterior ensembles generated by the full-order numerical model, as well as the rapidly generated DSI posterior ensemble.

Fig 3 displays time series across three distinct hydrodynamic zones: between the injection and extraction wells (wp1-f3), upgradient of the injection site (wp4-f5), and adjacent to the extraction well (pp1-f3). For all sites, data assimilation with both the full-order model and the DSI surrogate substantially reduces predictive uncertainty for SO4, TIC, and temperature. This variance

reduction is demonstrated by the tightly constrained posterior distributions compared to the wide variance of the prior ensemble.

Three iterations were employed for the full-order model history matching. This required a total of 1171 model runs, with an average run time of 8 min. Notably, approximately 70 model runs failed, likely due to numerical instability or exceeding run-time criteria. The wall-clock time that this process requires is dependent on the number of available CPUs. The 200 full-order model runs from the first iteration are used as training data for the DSI surrogate. These represent the only large compute cost of the DSI workflow. Surrogate-based history matching is accomplished in a few minutes on a high-end laptop.

Conclusions

The application of ensemble-based data assimilation to fully coupled reactive transport models (RTMs) is typically constrained by prohibitive computational costs. This study demonstrates the utility of Data-Space Inversion (DSI) as an efficient surrogate alternative to traditional parameter estimation. Using a three-dimensional synthetic model of aquifer injection and geochemical reaction networks related to pyrite oxidation, we evaluated the capacity of DSI to assimilate observations and generate conditional forecasts without requiring repeated forward model evaluations.

Results indicate that the DSI surrogate accurately reproduces the conditioning

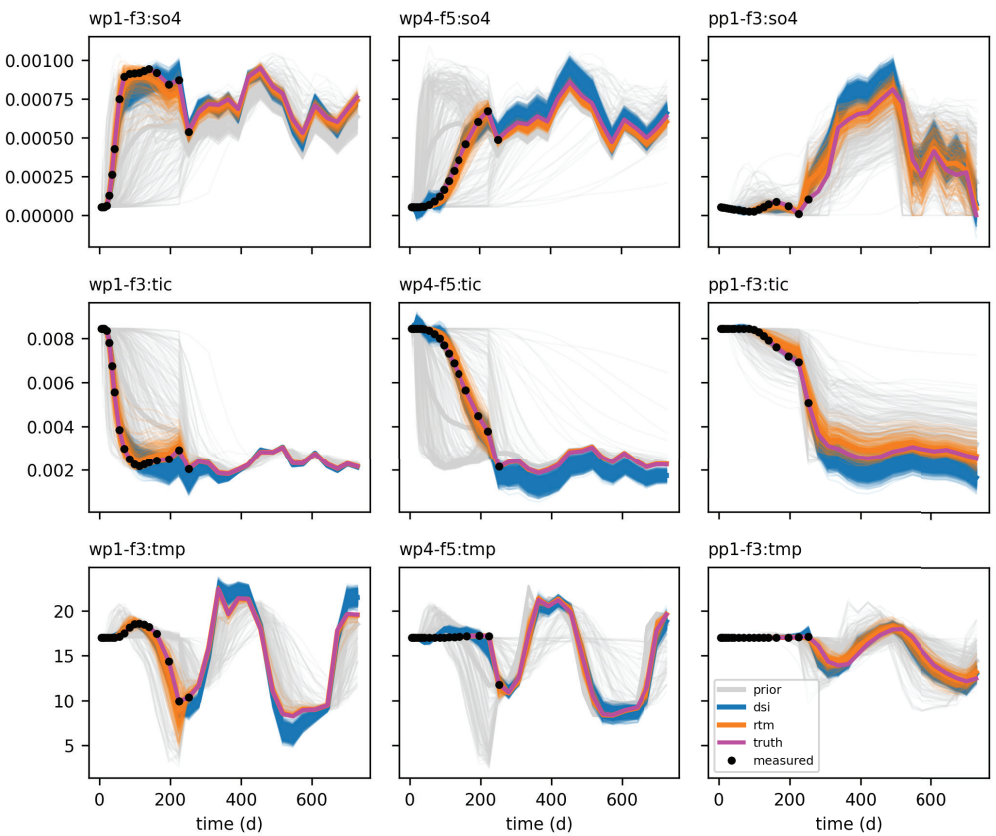

Figure 3 Time series of simulated and observed variables across three monitoring locations (columns) for sulfate (SO_4), total inorganic carbon (TIC), and temperature (rows). Each panel compares the reference synthetic truth and discrete calibration targets against three predictive ensembles: the unconditioned prior from the full reactive transport model (RTM), and the conditioned posteriors from the full RTM and the DSI surrogate.

behavior of the full-order model. Across distinct hydrodynamic zones, DSI successfully assimilated synthetic calibration targets and substantially reduced predictive uncertainty for key geochemical variables, including sulfate, total inorganic carbon, and temperature. By operating entirely in the data space, DSI bypassed the traditional computational bottlenecks associated with high-fidelity RTMs. Ultimately, this approach provides a practical framework for integrating complex geochemical models into routine uncertainty quantification, risk assessment, and decision support workflows for mine water management.

Acknowledgements

The authors thank the co-organisers for hosting the IMWA 2026 Conference. We thank the editors and reviewers for their constructive feedback.

References

Langevin CD, Hughes JD, Provost AM, Russcher MJ, Morway ED, Reno MJ, Bonelli WP, Niswonger RG, Panday S, Titus S, Merrick D, Banta ER (2026) MODFLOW6 Modular Hydrologic Model version 6.7.0: U.S. Geological Survey Software Release, 6 February 2026, DOI:10.5066/P1IJAXDZ

Parkhurst DL, Appelo CAJ, (2013) Description of input and examples for PHREEQC version 3–A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations: U.S. Geological Survey Techniques and Methods, book 6, chap. A43, 497 p.,

Prommer H, Stuyfzand P (2005) Identification of Temperature-Dependent Water Quality Changes during a Deep Well Injection Experiment in a Pyritic Aquifer. Environmental science & technology. 39. 2200-9. doi:10.1021/es0486768.

Sun W, Durlofsky L (2017) A new data-space inversion procedure for efficient uncertainty quantification in subsurface flow problems. Mathematical Geosciences, 49, 679–715. doi:10.1007/s11004-016-9672-8. URL doi:10.1007/s11004-016-9672-8

White JT (2018) A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. Environmental Modelling and Software, 109, 191–201. doi:10.1016/j.envsoft.2018.06.009.