Optimisation of Prediction-Driven Monitoring Programs

Rui Hugman[1], Cecile Coulon[2], Johanna Zwinger[3], Eduardo de Sousa[4]

[1]*INTERA Inc, 8000-254 Faro, Portugal, rhugman@intera.com, ORCID 0000-0003-0891-3886*

[2]*INTERA SAS, 69760 Limonest, France, ccoulon@intera.com, ORCID 0000-0001-9722-9976*

[3]*INTERA Inc, 6000 Perth, WA, Australia, jzwinger@intera.com, ORCID 0000-0003-0891-3886*

[4]*INTERA Inc, 6000 Perth, WA, Australia, edesousa@intera.com*

Abstract

Data acquisition optimisation in a decision-support modelling context is demonstrated. Groundwater model forecasts can accrue substantial uncertainty. Whilst assimilating field data can reduce this uncertainty, data collection can be expensive. In this paper an approach for optimal data collection that minimizes costs, maximizes informational value, and supports long-term resource sustainability is demonstrated. Using Ensemble Variance Analysis within a multi-objective optimisation framework, cost-effective monitoring locations for a real-world site are identified. Outcomes are a set of monitoring configurations which provide the optimal trade-off between cost and uncertainty reduction. From these, a monitoring program is selected that achieves 90% of possible uncertainty reduction at 30% of total cost.

Keywords: Monitoring, optimisation, uncertainty, decision-support, modelling, groundwater

Introduction

Effective groundwater management at mine sites often relies on numerical modelling to support decision-making, including designing dewatering systems, securing water supplies for mining operations, and evaluating potential environmental impacts of mining activities. Groundwater models, being simplifications of reality, are inherently uncertain, particularly given the scarcity of information about real-world system properties and stresses. This uncertainty can typically be reduced by assimilating data from system state measurements. However, data collection and monitoring programs can be costly, and from a decision-support perspective, not all data holds equal value.

Value of data in model-based decision-making is proportional to its capacity to reduce the uncertainty of model predictions. Data Worth Analysis (DWA) provides a systematic approach to evaluating the potential of new data to achieve this objective. Data worth based on linear analysis are well established in the groundwater modelling literature (e.g., Dausman *et al.* 2010; Fienen *et al.* 2010). However, they are limited by the assumption of linearity between changes in model parameters and predictions and come with the computational cost of filling out a Jacobian (e.g., sensitivity) matrix. The latter become particularly prohibitive when using high-dimensional parameterisation schemes, required to express hydrogeological heterogeneity and uncertainty.

More recently, He *et al.* (2018) introduced the Ensemble Variance Analysis (EVA) approach to assess data worth. EVA operates under the assumption that forecasted and measured values jointly follow a multi-Gaussian distribution. And that to estimate the change in forecast uncertainty, it is not necessary to know the value of future measured data, only the covariance between

the forecast and the measured data. Using this assumption, the variance (i.e., the uncertainty) and covariance (i.e., how knowing about one value changes uncertainty in another value) of model outputs can be estimated from an ensemble of simulations without requiring prior knowledge of the measured values. This ensemble of model outputs is generated by running a model many times with different samples of plausible parameter values. Each sample is referred to as a realisation.

From a practical perspective, EVA offers a significant advantage over linear methods: it relies on ensembles of model outputs rather than finite difference derivatives, making its computational cost independent of the number of model parameters and removing the assumption of linearity. This can reduce the number of model runs required for DWA from the order of a few thousands to a few hundred.

The current paper discusses the application of EVA to optimise a monitoring network at an undisclosed mine site. Measured data is used for history matching a decision-support groundwater model. The model is used to support management of extraction wells, with potential effects on several environmental receptors. As data collection is expensive, the objective is to rationalize the monitoring network to ensure high-quality forecasts whilst minimizing cost. Although we focus here on groundwater level data, this same approach is readily extendible to any datatype that can be employed to inform a model.

Methods

In summary, a numerical groundwater model is constructed to simulate predictions of management interest, as well as potential as-of-yet uncollected data from the monitoring network. This model is simulated many times with different parameter realisations. The combination of parameter realisations is referred to as an "ensemble". The simulated outputs from the ensemble of models are used to calculate the co-variance between predictions of interest and potential new data. This enables calculation of the expected predictive uncertainty, if the as-of-yet uncollected data is collected.

Subsequently, multi-objective optimisation is undertaken by calculating the expected predictive uncertainty many times, assuming different combinations of collected data, searching for the combination of monitoring locations that maximize uncertainty reduction at the minimum cost. Note this does not require re-running the numerical model, only the EVA calculations which have a low computational cost.

Site and Numerical Modelling

Due to confidentiality reasons, details of the site cannot be disclosed. However, this should not detract from the approach and outcomes. The mine site is located in an arid area and relies on a well-field to maintain water supply throughout the project's lifespan. The well-field extracts groundwater from a paleochannel aquifer, overlain by low-permeability calcretes and a phreatic aquifer. Annual recharge is low and sustainability of the well-field yield relies on storage and lateral inflow to the paleochannel. Management of the well-field is additionally constrained by needing to ensure that nearby environmental receptors are not affected by drawdown, and that the confined paleochannel is not desaturated.

A numerical model for the site is used to support management of the well-field and forecast sustainable yields. The model simulates historical and future project lifespan. History matching is undertaken for the historical period using an iterative ensemble smoother (IES), as implemented in the open-source software PESTPP-IES (White *et al.* 2018). IES provide computationally efficient approaches to condition model parameters to measured data. Information from available field data is assimilated (including hydraulic heads, site characterisation tests and other soft data). Hydraulic properties and unknown stresses (e.g., recharge, poorly documented extraction rates, external boundary conditions) are represented with a high-dimension parameterisation scheme to express spatial heterogeneity. Following history matching, model forecasts are made with the ensemble of models, providing quantified uncertainty of the predictions of interest.

Ensemble Variance Analysis

EVA operates on the assumption of a multivariate Gaussian relationship between the observation data and the prediction. It quantifies the expected reduction in uncertainty using covariance information derived from a set of simulations (i.e., an ensemble of model outputs). Thus, to conduct EVA, an ensemble of model-simulated outputs is required. This ensemble must include simulated outputs for both potential future data and the forecasts of interest. Conveniently, when using PESTPP-IES, an ensemble of model outputs is a byproduct of predictive uncertainty analysis.

Let the vector composed of the sub-vectors $\boldsymbol{s}$ and $\boldsymbol{d}$ denote outputs generated by a model $\boldsymbol{Z}$ simulated with uncertain parameters represented by the vector $\boldsymbol{k}$.

$$\begin{bmatrix} \boldsymbol{s} \\ \boldsymbol{d} \end{bmatrix} = \mathbf{Z}(\boldsymbol{k}) \qquad (1)$$

Vector $\boldsymbol{s}$ contains model outputs that correspond to predictions of interest. The vector $\boldsymbol{d}$ contains simulated outputs that correspond to as-of-yet uncollected data.

If the model $\boldsymbol{Z}$ is simulated many times, each time with a different sample of $\boldsymbol{k}$, the ensemble of model outputs can be collected into a matrix from which the covariance between predictions and as-of-yet uncollected data can be calculated as:

$$\boldsymbol{C}\left(\begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{s} \end{bmatrix}\right) = \begin{bmatrix} C_{ss} & C_{sd} \\ C_{ds} & C_{dd} \end{bmatrix} \qquad (2)$$

From the above, and assuming a multi-Gaussian distribution between a prediction s and measured data *d*, the expected posterior variance of the prediction can be calculated as:

$$\sigma^2_{s|d} = \sigma^2_s - C_{sd} C^{-1}_{dd} C_{ds} \qquad (3)$$

Under the assumption of multi-Gaussian distribution, posterior variance of the prediction s is independent of the value of measured data *d*. Expected variance is the average variance of s given a value of $\boldsymbol{d}$, across the ensemble. Thus, it provides a conservative lower bound of expected uncertainty reduction. Different values of s can be computed assuming different combinations of d to assess their relative value in reducing predictive uncertainty. As these calculations are computationally cheap, it becomes feasible to wrap them within global optimizer algorithms, as described below.

Monitoring Network Optimisation

Multi-objective optimisation was undertaken employing particle swarm optimisation (PSO; Kennedy and Eberhart 1995)—a population-based, stochastic search algorithm inspired by natural swarm behaviour—and the NSGA-II algorithm (Deb *et al.* 2002), which uses fast non-dominated sorting to efficiently handle trade-offs among multiple objectives. The workflow is implemented using the open-source software PESTPP-MOU (White *et al.* 2022).

Optimisation objectives are quantities that the optimisation algorithm aims to minimise or maximise. The monitoring optimisation was formulated as a two-objective optimisation:

1. maximise total uncertainty reduction, and
2. minimise cost.

For the case described herein, we consider the total uncertainty reduction as an aggregate of the "percentage uncertainty reduction" across all forecasts. It is calculated by summing percentage uncertainty reductions across all forecasts. For this case, we simply aim to minimize aggregate uncertainty. However, more complex formulations of the objective function are possible, such as aiming to achieve a minimum variance for a given prediction.

Cost is calculated as the total number of samples from all sites, multiplied by the average cost per sample. This value does not account for variable costs, such as distance travelled between sites, as they were not available – however, it could. For this site an assumption of one sample per year was made. However, more complex parameterisations are possible, such as optimising for sampling frequency and duration.

The optimisation algorithm explores the solution space by testing different combinations of monitoring locations, calculating their cost and assessing their worth at reducing uncertainty. Outcomes are combinations that provide the maximum uncertainty reduction for a given cost (or a minimal cost for a given uncertainty reduction).

Results

The optimal trade-off between cost and uncertainty reduction is shown in Fig. 1. Referred to as the "pareto front", this curve represents the set of optimal solutions in a multi-objective optimisation problem, where no objective can be improved without worsening at least one other. In other words, at any point along the curve, it is impossible to reduce uncertainty without incurring greater cost (and vice versa).

The y-axis on Fig. 1 is scaled to percentage of the maximum possible uncertainty reduction achievable by collecting data from all available sites. Outcomes show that approximately 85% of possible uncertainty reduction can be achieved with around 20% of the monitoring locations. This represents a substantial saving in terms of cost. As the number of sites in the monitoring network increases, there are diminishing returns. Most of the information gains are achieved from a small portion of the monitored sites.

For comparison, the outcomes for an expert-knowledge (i.e., "manual") designed monitoring plan are displayed in Fig. 1. The proposed plan provides sub-optimal uncertainty reduction. In other words, uncertainty could be reduced substantially further for cheaper.

It remains incumbent on the decision-maker to determine the acceptable trade-off. This will always be case specific. For the site, a desired uncertainty reduction of at least 90% of achievable was specified by the decision-maker. The proposed monitoring configuration was reduced to 65 sites, approximately 30% of the total projected cost, assuming an average cost per sample over the project lifespan.

Conclusions

Data worth optimisation using EVA was employed to optimize data collection to inform prediction-driven modelling. This approach provided a sub-set of monitoring locations that would provide 90% of the information content of the entire network at 30% of the cost. Furthermore, the optimised monitoring locations achieved better results than previously proposed "expert knowledge" derived locations, both in terms of cost and uncertainty reduction.

The approach described herein provides an effective and computationally cheap approach to inform the design of data acquisition programs within a prediction-driven modelling context. Although here it is employed for an existing network and targeting groundwater level measurements, the approach is readily extendible to other contexts and datatypes.

Acknowledgements

The authors thank all co-organisers for hosting the IMWA 2025 Conference and the reviewers and editors for providing constructive comments on this text.

References

Dausman AM, Doherty J, Langevin CD,Sukop MC (2010) Quantifying data worth toward reducing predictive uncertainty. Groundwater, 48: 729–740. https://doi.org/10.1111/j.1745-6584.2010.00679.x

Deb K, Pratap A, Agarwal S, Meyarivan, TAMT (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2), 182-197. https://doi.org/10.1109/4235.996017

Fienen MN, Doherty J, Hunt RJ, Reeves HW (2010) Using pre-diction uncertainty analysis to design hydrologic monitoring networks: Example applications from the Great Lakes water availability pilot project. Technical Report. US Geological Survey.

He J, Sarma P, Bhark E, Tanaka S, Chen B, Wen X, Jaira K (2018) Quantifying expected uncertainty reduction and value of information using ensemble-variance analysis. SPE J. 23 428–448. doi: https://doi.org/10.2118/182609-PA

Kennedy J, Eberhart R (1995) Particle swarm optimization. *In Proceedings of ICNN95-international conference on neural networks* (Vol. 4, pp. 1942–1948). IEEE. http://dx.doi.org/10.1109/ICNN.1995.488968.

White JT, Fienen MN, Doherty J (2016) pyEMU: a python framework for environmental model uncertainty analysis, version .01: U.S. Geological Survey software release, https://dx.doi.org/10.5066/F75D8Q01.

White JT (2018) A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environmental Modelling & Software,* 109, 191–201.

White JT, Hunt RJ, Fienen MN, Doherty J (2020) Approaches to Highly Parameterized Inversion: PEST++ Version 5, a Software Suite for Parameter Estimation, Uncertainty Analysis, Management Optimization and Sensitivity Analysis: *U.S. Geological Survey Techniques and Methods* 7C26., https://doi.org/10.3133/tm7C26.

White JT, Knowling MJ, Fienen MN, Siade A, Rea O, Martinez G (2022). A model-independent tool for evolutionary constrained multi-objective optimization under uncertainty. *Environmental Modelling & Software*, 149, 105316.