Adv. Geosci., 64, 23–26, 2024
https://doi.org/10.5194/adgeo-64-23-2024

Advances in Geosciences
Open Access
EGU

An Efficient Surrogate-based Multi-objective Optimisation Framework with Novel Sampling Strategy for Sustainable Island Groundwater Management

Weijiang Yu[1], **Domenico Baù**[1], **Alex S. Mayer**[2], **and Mohammadali Geranmehr**[1]

[1]Department of Civil and Structural Engineering, University of Sheffield, Sheffield, S10 2TN, UK
[2]Department of Civil Engineering, University of Texas at El Paso, El Paso, 79968, USA

Correspondence: Weijiang Yu (weijiang.yu@sheffield.ac.uk)

Received: 5 May 2024 – Revised: 9 July 2024 – Accepted: 15 July 2024 – Published: 22 August 2024

Abstract. In groundwater pumping optimization (GPO), offline-trained data-driven surrogates can be used to replace numerical-intensive simulators in order to save computing time. The traditional offline training approach involves building surrogates prior to optimization, fitting training datasets that cover the input space uniformly or randomly, which can prove inefficient due to the potential oversampling of low-gradient areas and under-sampling of high-gradient areas. This study proposes an offline machine-learning (ML) algorithm that ranks candidate training points by scoring them based on their distance to the closest training point and on the local gradient of the surrogate estimate and then choosing the highest-rank point. This method is applied to develop surrogates for solving a two-objective GPO problem formulated on a three-dimensional (3D) island aquifer, using hydrogeological conditions representative of San Salvador Island, Bahamas. The objectives are to minimise the supply cost (f_{OC}) resulting from groundwater pumping and desalination and maximise fresh groundwater supply (Q_p), subject to constraints on seawater intrusion (SWI) control expressed in terms of aquifer drawdown Δs at pumping locations and aquifer salt mass increase ΔSM. Gaussian Process (GP) is the technique applied to construct surrogates of objectives and constraints, alongside the estimation of uncertainties. Using GP models, it is possible to estimate the probability of "Pareto optimality" for each pumping scheme by Monte Carlo simulation. Pareto optimal pumping schemes (POPS) are then characterized by a probability of occurrence, which can be verified by numerical simulation. The GP training strategy's effectiveness in generating POPS is compared to traditional training approaches, showing that such a strategy can efficiently identify reliable POPS.

1 Introduction

Limited by recharge rates and land area, freshwater in island aquifers is typically in the shape of a lens, with a thickness between a few meters to a few tens of meters, which makes it particularly vulnerable to SWI (Kourakos and Mantoglou, 2015; Gulley et al., 2016; Coulon et al., 2022). Island aquifers are often the main freshwater source for local communities, making groundwater pumping inevitable. Yu et al. (2023) have shown that while pumping from the island center at shallow depths is cost-effective for meeting demand, it increases the risk of seawater intrusion (SWI), highlighting a conflict between supply costs and SWI control.

GPO is usually adopted to investigate the trade off between sustainability and/or SWI control against cost. GPO is formulated by optimizing management objectives subject to constraints, both of which are typically functions of decision variables (DVs) and state variables (SVs). Applying the simulation-optimization (SO) method to derive POPS may incur huge computational costs due to repeated calls to numerically-intensive simulators. To reduce this cost, fast data-driven surrogates can be employed instead, forming a surrogate-based SO framework. Traditionally, these surrogates are built by "offline" training, that is, prior to the optimization, by fitting training datasets that cover the DV space either uniformly or randomly. Traditional training methods may result in inefficiency due to oversampling low-

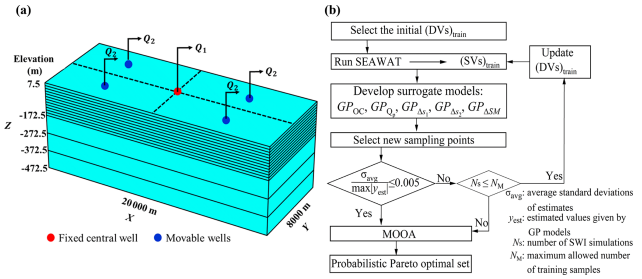

Figure 1. **(a)** The conceptualization of the island aquifer domain. **(b)** Offline MOOA framework based on the proposed strategy.

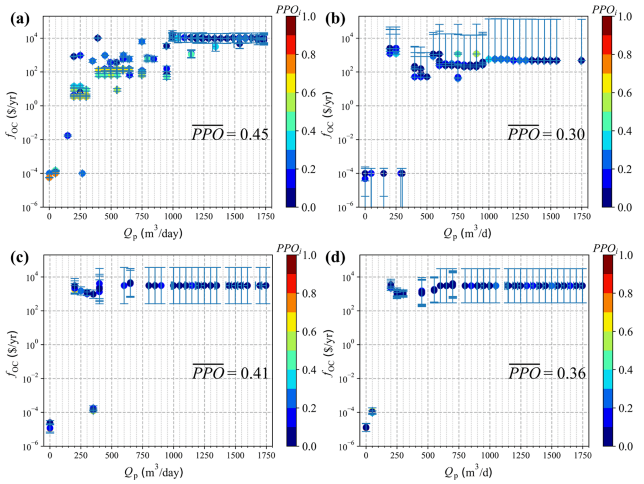

Figure 2. PO sets of rate of freshwater produced Q_p vs. supply cost f_{OC} along with values of the probability of Pareto-optimality and estimation uncertainty under training scenarios **(a)** T1, **(b)** T2, **(c)** T3 and **(d)** T4.

gradient and under-sampling high-gradient areas of the DV space. We introduce an offline ML algorithm for selecting training points, applied in an island aquifer multi-objective GPO. This efficient sampling strategy helps develop surrogate models to quickly identify reliable POPS and promote a sustainable coastal hydrogeological environment.

2 Methods

2.1 Study area

An illustrative rectangular aquifer, loosely representing the lens-shaped freshwater system of San Salvador Island (Bahamas) (Gulley et al., 2016), is chosen as a test site. This aquifer, depicted in Fig. 1a, is discretized into a 3D finite-difference grid of 60 000 cells. The four lateral sides are at a constant salt concentration of 35 g L^{-1} and a constant head of 0 m, representing seaward boundary conditions. Hydrogeological details specific to the study area can be found in Yu et al. (2023). It is important to note that this model, though computationally substantial, does not fully account for real-world site conditions, such as the presence of surface water bodies. This simplification approach aims to provide qualitative insights applicable to similar island aquifer settings.

2.2 SWI modelling and surrogate-based GPO

Pumping is operated by five wells, one at the island centre, and the other four positioned symmetrically relative to the "central" well, located at $(0, 0)$. The central well is defined by two DVs, the pumping depth D_1 and the pumping rate Q_1. The pumping depth D_2, the pumping rate Q_2, and the coordinates ($\pm X_2$,$\pm Y_2$), are the DVs for the other four wells (Fig. 1a).

SEAWAT (Langevin et al., 2008) is used to model aquifer response under different pump schemes and construct training data. GP with a square exponential covariance (kernel) function is used to build surrogates (Rasmussen and Williams, 2006) of objectives and constraints in the 6D DV space. GP_{OC}, GP_{Q_p}, $GP_{\Delta s_1}$, $GP_{\Delta s_2}$ and $GP_{\Delta SM}$ denote the GPs for estimating f_{OC}, Q_p, Δs_1, Δs_2, and ΔSM, respectively. The GP training relies on an offline ML training strategy, in which an initial set of training points is progressively reinforced by adding new points, selected by identifying the DV set with a maximum value of a score function $F = 0.5 \cdot R\,(\mathrm{dc}) + 0.5 \cdot R\,(\nabla G)$. R is a "normalization" operator ($R\,(\blacksquare) \in [0,1]$), "dc" is the distance of a candidate point from the closest training point, and ∇G is the local gradient based on GP model estimates. Figure 1b shows a flow chart of the offline multi-objective optimisation (MOOA) framework based on the proposed strategy.

For any DV set, GP models enable the estimation of f_{OC}, Q_p, Δs_1, Δs_2 and ΔSM and quantify the uncertainty on it. The problem solution is represented by Pareto optimal (PO) sets of "non-dominated" solutions, each of which is characterized by a probability of Pareto optimality PPO_i, estimated by Monte Carlo simulation with a sample size of 100. The reliability of the derived POPS is assessed by means of the average probability of Pareto optimality of a trade-off set ($\overline{PPO}$), and the normalized root mean square errors (NRMSE), εOC, εQ_P, $\varepsilon \Delta s_1$, $\varepsilon \Delta s_2$ and $\varepsilon \Delta$SM, which quantify the difference between GP model predictions and their "true" values calculated by SEAWAT simulations over the 30 PO solutions with higher PPO.

3 Results and Discussion

Preliminary tests (not presented here) indicate that the performance of the offline training approach depends on the number of allowed training points N_M. In these tests, $\overline{PPO}$ is seen to increase with N_M, as larger training datasets enhance surrogate estimates, but gains become marginal for N_M approaching 400. We thus investigate the case, denoted as T1, in which $N_M = 400$, against the case T2 in which 700 train-

Table 1. Normalized root mean square errors of POPS in the training scenarios T1–T4.

Training	εOC	εQ_P	$\varepsilon \Delta s_1$	$\varepsilon \Delta s_2$	$\varepsilon \Delta$SM
T1	0.18	0.31	0.00	0.00	0.00
T2	0.32	0.44	0.21	0.35	0.00
T3	0.42	0.41	0.00	0.30	0.00
T4	0.40	0.39	0.00	0.47	0.00

ing points are uniformly distributed within the DV space, and two other cases T3 and T4, which consist of two different realizations of 1000 training points randomly distributed across the DV space.

Figure 2 presents the PO sets obtained in (a) T1, (b) T2, (c) T3 and (c) T4. Each point is plotted with a colour depending on PPO_i (see scale), and horizontal and vertical error bars indicative of the surrogate uncertainty. POPS in T1 are characterized by a remarkably smaller uncertainty than those in T2, T3 and T4, indicating that the former are more robust and more reliable than the latter. $\overline{PPO}$ in T1 is significantly higher than in T2, T3 and T4, suggesting that the training approach is more effective, other than computationally less intensive, than traditional methods. NMRSE values reported in Table 1 confirm the better reliability of results obtained in T1.

4 Conclusions

The performance of a novel surrogate sampling strategy, using a novel ML algorithm for training point selection, was assessed for an island multi-objective GPO problem. The findings revealed that, compared to traditional training approaches, the proposed strategy can generate more reliable POPS and consume a lower computational cost. For the formulated GPO problem, the results in Fig. 2a indicate that supply cost f_{OC} increases gradually with a fresh groundwater supply Q_p up to $950\,m^3\,d^{-1}$. Beyond this point, f_{OC} increases sharply, suggesting that $950\,m^3\,d^{-1}$ is a sustainability threshold for balancing economic viability and SWI control.

Code availability. Island aquifer response to the pumping activities was simulated using version 4 of the SEAWAT groundwater software (https://www.usgs.gov/software/seawat-computer, United States Geological Survey, 2012). Scikit-learn, a publicly available Python library, was employed to train and develop Gaussian Process models (https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html, Pedregosa et al., 2011). The Monte Carlo stochastic runs, identification of new sampling points at each iteration, and plotting in this paper were carried out using various Python libraries, including pandas, numpy, and matplotlib. These libraries can be accessed through official Python websites: https://www.python.org/ (Python, 2024).

Data availability. The underlying research data for this study are the pumping patterns and their corresponding management objective and constraint values. Data on aquifer response to the pumping activities can be obtained by simulation using SEAWAT, and calculations of the management objective and constraint values can refer to https://doi.org/10.1029/2023WR034798 (Yu et al., 2023). Research data are also available on request from the authors.

Author contributions. WY was responsible for the conceptualization, methodology, formal analysis, software, visualization, writing of the original draft, and review and editing of the manuscript. DB contributed to the conceptualization, methodology, formal analysis, and supervision, as well as the review and editing of the manuscript. ASM and MG were involved in the supervision and review of the writing. All authors have read and agreed to the published version of the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Special issue statement. This article is part of the special issue "Groundwater management in the context of global change: integrating innovative approaches (EGU2024 HS8.2.1 session)". It is a result of the EGU General Assembly 2024, Vienna, Austria, 14–19 April 2024.

Acknowledgements. The authors sincerely thank the Engineering and Physical Sciences Research Council (UK) and the National Science Foundation (USA) for their financial support, with grant numbers EP/T018542/1 and 1903405, respectively. The authors also would like to thank the editor, and two anonymous reviewers for their kind feedback and insightful comments, which helped improve the clarity of this paper.

Financial support. This research has been supported by the Engineering and Physical Sciences Research Council (grant no. EP/T018542/1) and the National Science Foundation (grant no. 1903405).

Review statement. This paper was edited by Estanislao Pujades and reviewed by two anonymous referees.

References

Coulon, C., Lemieux, J. M., Pryet, A., Bayer, P., Young, N. L., and Molson, J.: Pumping optimization under uncertainty in an island freshwater lens using a sharp-interface seawater intrusion model, Water Resour. Res., 58, e2021WR031793, https://doi.org/10.1029/2021WR031, 2022.

Gulley, J. D., Mayer, A. S., Martin, J. B., and Bedekar, V.: Sea level rise and inundation of island interiors: Assessing impacts of lake formation and evaporation on water resources in arid climates, Geophys. Res. Lett., 43, 9712–9719, https://doi.org/10.1002/2016GL070667, 2016.

Kourakos, G. and Mantoglou, A.: An efficient simulation-optimization coupling for management of coastal aquifers, Hydrogeol. J., 23, 1167–1179, https://doi.org/10.1007/s10040-015-1293-7, 2015.

Langevin, C. D., Thorne Jr., D. T., Dausman, A. M., Sukop, M. C., and Guo, W.: SEAWAT Version 4: A computer program for simulation of multi-species solute and heat transport, in: U.S. Geological Survey Techniques and Methods Book 6 (p. 39), Chapter A22, https://doi.org/10.3133/tm6A22, 2008.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (last access: 19 August 2024), 2011.

Python: python, https://www.python.org/ (last access: 19 August 2024), 2024.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, the MIT Press, ISBN 026218253X, https://doi.org/10.7551/mitpress/3206.001.0001, 2006.

United States Geological Survey: SEAWAT: A computer program for simulation of three-dimensional variable-density ground-water flow and transport, U.S. Geological Survey Software Release [software], https://www.usgs.gov/software/seawat-computer, 2012.

Yu, W., Baù, D., Mayer, A. S., Mancewicz, L., and Geranmehr, M.: Investigating the impact of seawater intrusion on the operation cost of groundwater supply in island aquifers, Water Resour. Res., 59, e2023WR034798, https://doi.org/10.1029/2023WR034798, 2023.