Exercise: Identification of Pollution Concentrations in a Shallow Groundwater Aquifer using Data-Driven Approaches

Roger Roca Miró
Legacy Contamination and Soil Remediation
TU Bergakademie Freiberg

March 6, 2025

1 Introduction

Groundwater contamination is a critical issue with significant environmental and public health implications. Accurate prediction of pollutant concentrations is essential for effective remediation and risk management. In this report, we apply the methodology of Taherdangkoo et al. [2] by adapting advanced neural network training algorithms—Levenberg-Marquardt (LM) and Bayesian Regularization (BR)—to a simplified 2D contamination model. Our approach simulates contaminant migration in a shallow groundwater aquifer and compares the performance of both algorithms while evaluating the influence of key hydrogeological parameters. This exercise demonstrates the potential of data-driven methods for assessing groundwater contamination and highlights the value of simplified models in environmental risk assessment.

2 Methodology

2.1 Conceptual Model and Boundary Conditions

The 2D generic contamination model domain spans 150 m in length and 100 m in width, and is set to be located at a depth of 30 m (depth being just relevant to make clear that it represents a shallow aquifer). Two leakage sources are positioned along the left boundary, while an observation well is placed on the right side of the domain. The key input parameters for the simplified model are permeability, *porosity*, and *layer length*, while the output is the contaminant concentration.

Groundwater flows from left to right under single-phase flow conditions, with the contaminant treated as a conservative tracer (i.e., sorption and degradation processes are neglected). Pressure head boundaries are applied laterally, and no-flow conditions are enforced at the top and bottom boundaries. All strata are assumed to be homogeneous and isotropic. The contaminant is allowed to leak into the aquifer over a simulation period of 150 hours.

Figure 1 presents a detailed diagram of the proposed model.

2.2 Neural Network Structures and Training Algorithms

A feedforward neural network with one hidden layer is employed to predict pollutant concentrations. The study investigates networks with the number of hidden neurons varying from 1 to 40. Key aspects of the network setup include:

- **Data Normalization and Division:** Input and target data are normalized to the range $[-1, 1]$ with the *mapminmax* function, since the used training algorithms perform better with data centered at zero [1, 2]. This process ensures all features contribute equally during network training and the normalization parameters (the original minimum and maximum values) are stored for later use in reversing the transformation on the network's predictions. The dataset is split into training (70%), validation (15%), and testing (15%) sets.
- **Network Architecture:** The hidden layer uses a hyperbolic tangent sigmoid (*tansig*) transfer function (chosen over the default *logsig* because of the $[-1, 1]$ range of the data after the normalization), while the output layer uses a linear (*purelin*) function.
- **Training Algorithms:**

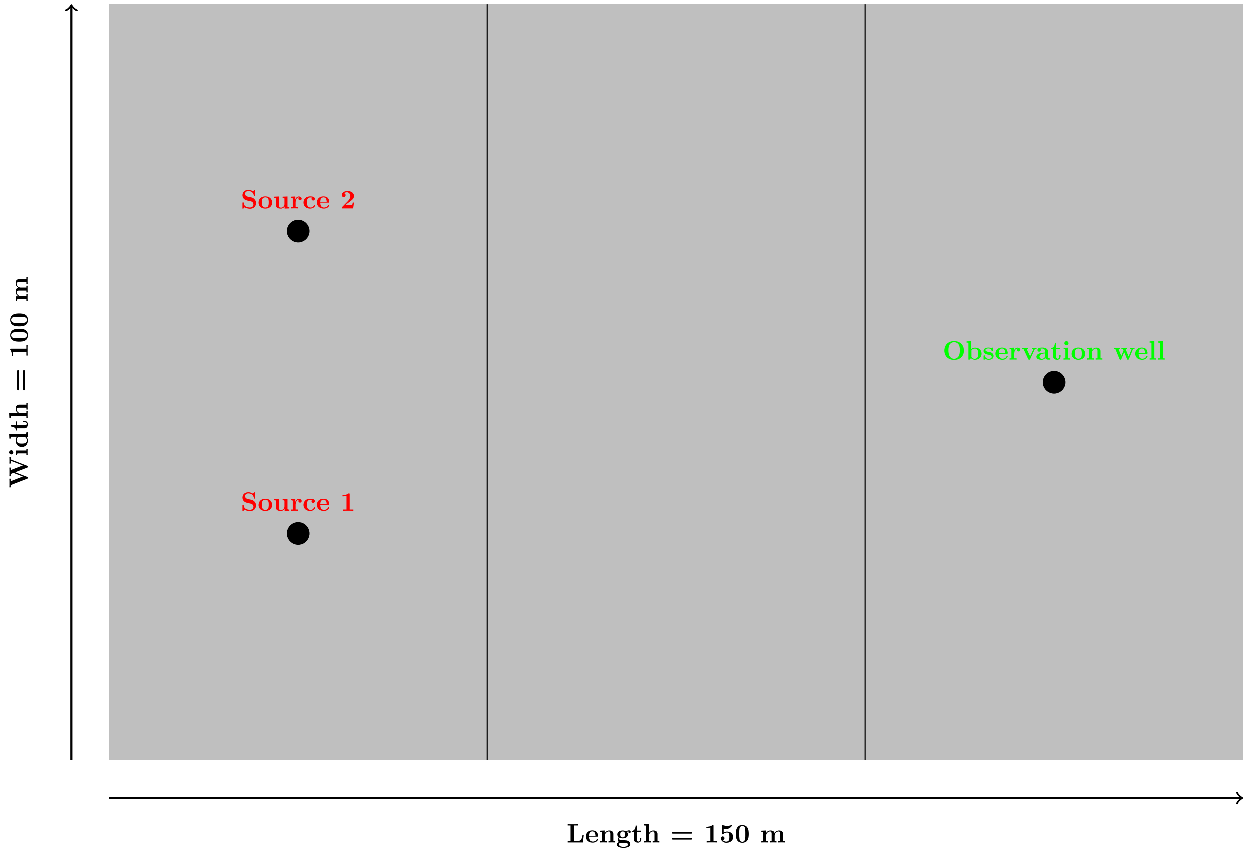

Figure 1: *2D generic contamination model to investigate flow and transport of contaminant plumes in a shallow aquifer. Modified in LaTeX from the WS 2023 Prof. Taherdangkoo slides.*

1. **Levenberg-Marquardt (LM):** The LM algorithm is used with the network configured to use a validation set to avoid overfitting. This algorithm is favored for its fast convergence and efficiency [2].
2. **Bayesian Regularization (BR):** The BR algorithm is applied combining the training and validation sets to enhance generalization. This is particularly advantageous when working with small datasets like ours, because BR minimizes the need for a separate validation phase and enhances the model's generalization ability [2].

- **Performance Metrics:** The networks are evaluated using the mean squared error (MSE) computed on training, validation, and testing datasets. Regression plots are generated to assess prediction accuracy, and a parameter importance analysis is conducted to determine the influence of input parameters (*permeability*, *porosity*, and *layer length*) on model performance.

  The idea behind the parameter performance analysis is to measure how much each input contributes to the network's performance. In the code, this is done by excluding one input at a time and then retraining the best selected network without that parameter. For each exclusion, the mean squared error (MSE) is computed on both the training and testing sets. These errors capture how much the network's performance degrades when that parameter is missing, they are averaged for all the parameters (each divided by the sum of all average impacts) to yield a relative importance metric (adding up to 1).

3 Results and Discussion

Both Levenberg-Marquardt (LM) and Bayesian Regularization (BR) algorithms demonstrated strong predictive capabilities, with the LM network achieving its best performance using 8 neurons (Training MSE: 3.293448e+04, Validation MSE: 4.731353e+04, Testing MSE: 7.251373e+04) and the BR network performing optimally with 27 neurons (Training MSE: 1.355082e+04, Testing MSE: 3.161464e+04), turning up to be the superior algorithm. *Figure 2* presents the performance comparison between the LM and BR algorithms in terms of MSE versus the number of neurons. The LM network shows separate curves for training, validation, and testing, with its best performance selected based on the validation error. In contrast, the BR network demonstrates a more robust performance without the advantage of avoiding the validation split.

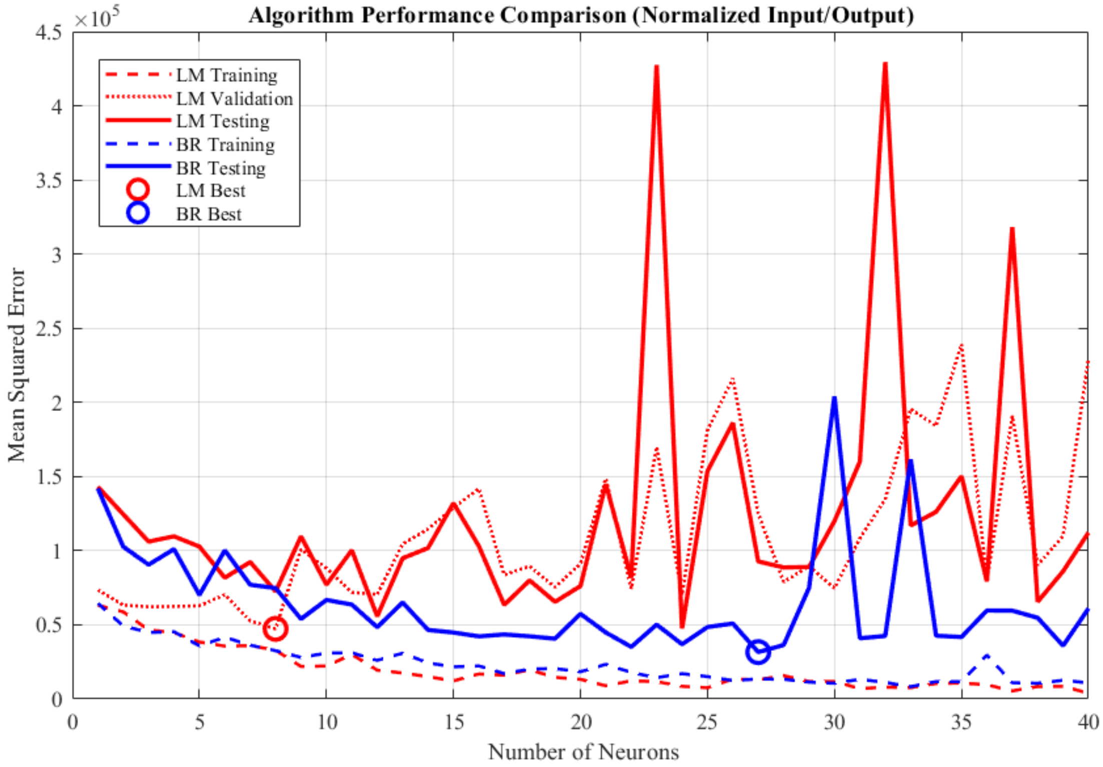

Figure 2: *Algorithm Performance Comparison: MSE vs. number of neurons for LM (training, validation, and testing) and BR (training and testing). Circular markers indicate the best performing networks for each algorithm.*

The regression analyses for the LM and BR networks are illustrated in *Figure 3* and *Figure 4*, respectively. For the LM network (*Figure 3*), scatter plots of true versus predicted values for training, validation, and testing datasets indicate a strong correlation along the 45-degree line. Similarly, the BR network (*Figure 4*) shows a better agreement between predicted and true pollutant concentrations, confirming the better suitability for this application.

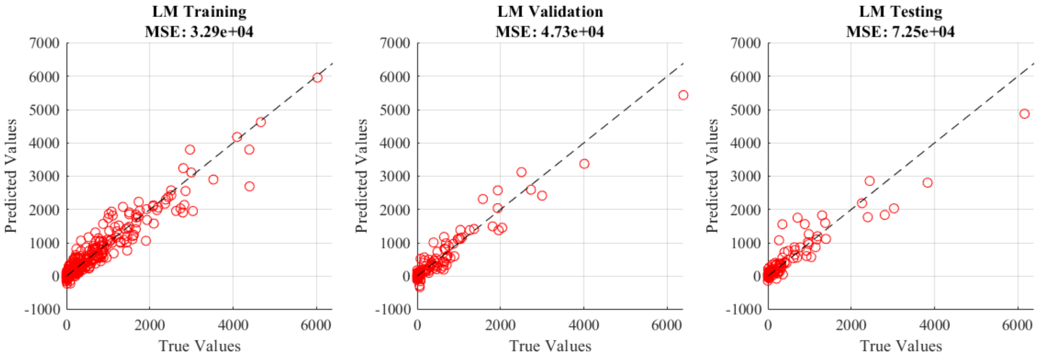

Figure 3: *LM Network Regression Analysis: Scatter plots for (a) training, (b) validation, and (c) testing datasets with corresponding MSE values.*

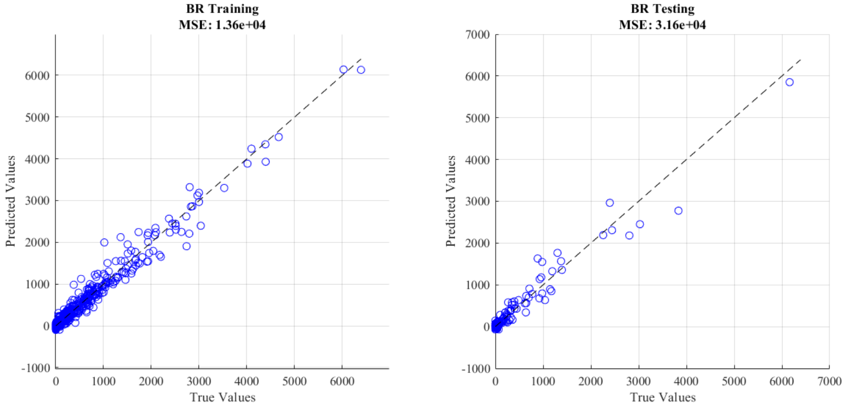

Figure 4: *BR Network Regression Analysis: Scatter plots for (a) training (combined with validation) and (b) testing datasets with corresponding MSE values.*

For the parameter importance analysis the bar chart in *Figure 5* reveals the greater contribution of *porosity* (0.510), over *permeability* (0.207), and *layer length* (0.283).

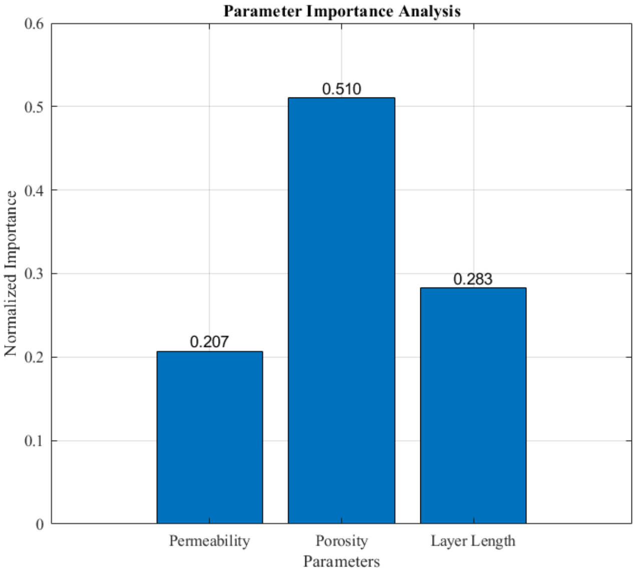

Figure 5: *Parameter Importance Analysis: Normalized importance of input parameters obtained by evaluating the impact of each parameter's exclusion on network performance.*

Overall, the results demonstrate that the BR algorithm is more effective in predicting pollutant concentrations, and confirm *porosity* as the most impactful parameter.

4 Conclusion

This study has demonstrated the application of data-driven neural network approaches for the identification of pollution concentrations in a shallow groundwater aquifer. A 2D generic contamination model was used to simulate contaminant transport under realistic boundary conditions. Neural networks with varying numbers of hidden neurons were trained using both Levenberg-Marquardt and Bayesian Regularization algorithms.

Both LM and BR networks provided accurate predictions, with the results making clear that the BR is the superior algorithm for the application. The LM network's performance was optimized by monitoring validation MSE, while the BR network benefited from an automatic regularization mechanism. Parameter importance analysis highlighted that the *porosity* is the main driver of pollution spread, while *permeability* and *layer length* can't be ignored.

References

[1] Guzman, Sandra M.; Paz, Joel O.; Tagert, Mary Love M. (2017): The Use of NARX Neural Networks to Forecast Daily Groundwater Levels. In Water Resour Manage 31 (5), pp. 1591–1603. DOI: 10.1007/s11269-017-1598-5.

[2] Taherdangkoo, R., Tatomir, A., Taherdangkoo, M., Qiu, P., & Sauter, M. (2020). Nonlinear autoregressive neural networks to predict hydraulic fracturing fluid leakage into shallow groundwater. Water, 12(3), 841. https://doi.org/10.3390/w12030841