Data Analysis of 200 West Pump-and-Treat System using Supervised Machine Learning Modeling

Rohan Shanbhag[1,2], Xuehang Song[1,*], Mark Rockhold[1], Marinko Karanovic[3], Matt Tonkin[3], Inci Demirkanli[1], and Rob Mackley[1]

[1]Pacific Northwest National Laboratory, [2]Florida International University, [3]S.S. Papadopulos and Associates Inc

Background

- This study focuses on the 200 West Area in Hanford Site's Central Plateau, where plutonium recovery at the Plutonium Finishing Plant led to subsurface contamination. A Pump-and-Treat (P&T) system has been operating since 2012 to extract contaminants like CCl_4, nitrate, and Tc-99.
- A significant challenge in groundwater remediation is the uncertainty in the distribution of underground contamination due to limited groundwater sampling data.
- A decade's worth of historical P&T operations in 200 West P&T has yielded a wealth of operational data, including pumping rates and contamination mass recovery. This information could potentially provide valuable insights for enhancing site characterization and understanding plume distribution. (Figure 1).

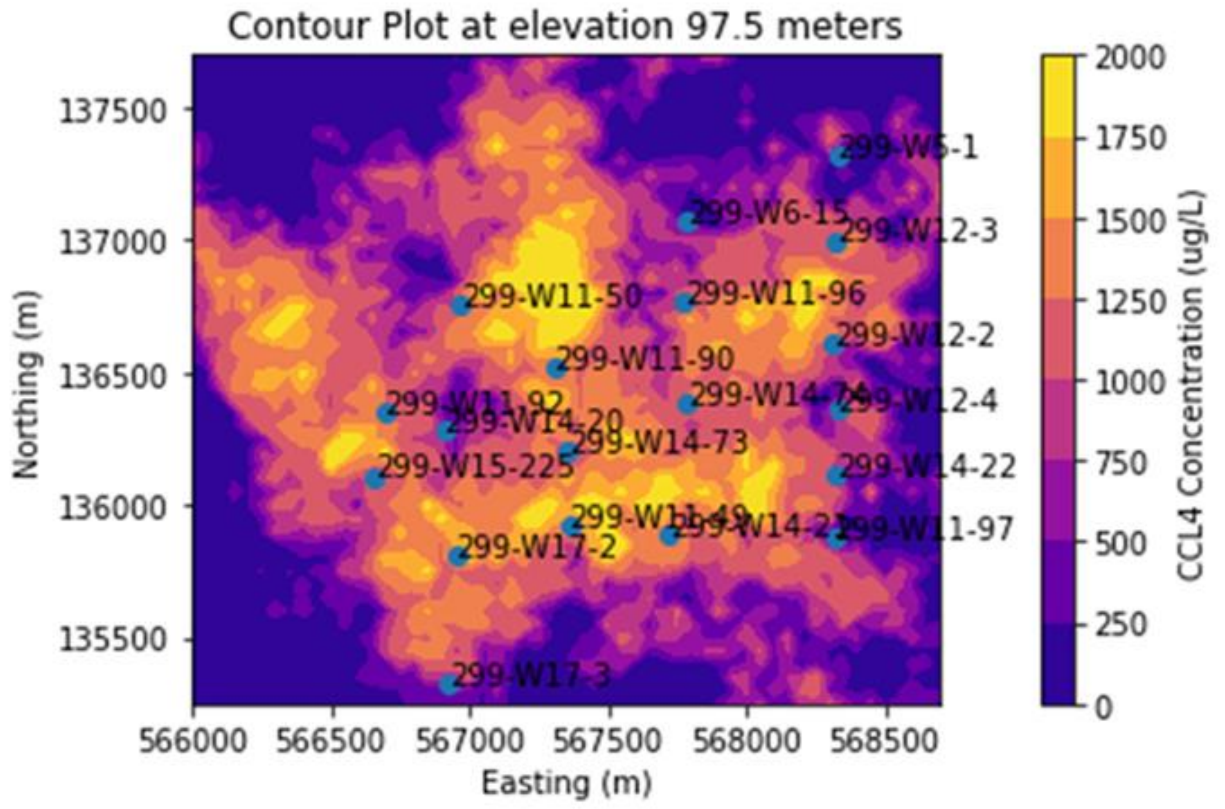


Figure 1. One geostatistical realization of CCl_4 distribution at certain elevation (97.5 m)

Objective

- The research aimed to create supervised machine learning models correlating potential CCl_4 plume distributions with actual well performance data from P&T operations (2012-2020).
- Machine learning models were then used to identify which plume realization best explains the spatial variability of well performances.

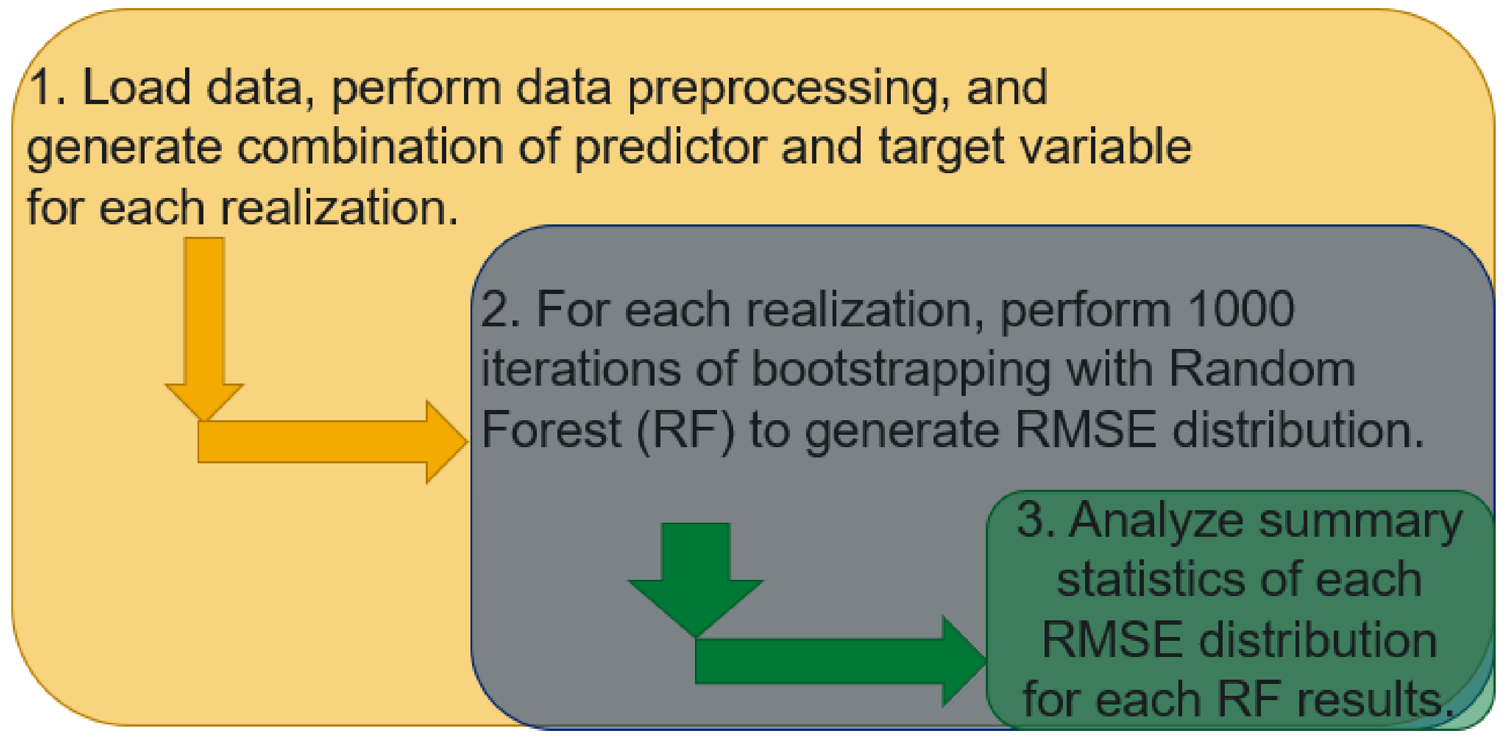
Methodology

- Predictor variables were derived from extraction well pumping rates (temporal) and existing geostatistical realization plume data (spatial). Target variables were based on the CCl_4 mass and/or concentration time series from these wells. (Table 1)

Predictor Variables	Target Variables
• Mean Q, standard-deviation of Q, minimum and maximum Q, percentiles (25th, 50th, and 75th) of Q, and total years of operation. • SSA (trend, periodicity) data of Q. • Total mass of contamination within certain radii (50 meters, 100 meters, 150 meters, and 200 meters) of a well. • Concentration-weighted distance between all points in the plume and each well.	• Mean CCl_4 concentration and mass. • Standard-deviation of CCl_4 concentration and mass. • Minimum/maximum CCl_4 concentration and mass. • Percentiles (25th, 50th, and 75th) of CCl_4 concentration and mass. • SSA (trend, periodicity) data of CCl_4 concentration and mass.

Table 1. List of predictor / target variables

- Random Forest, an ensemble-based supervised machine learning method, was chosen for its ability to effectively prevent model overfitting. (Figure 2).
- Bootstrapping was employed to create multiple datasets from limited well location. (Figure 2).



1. Load data, perform data preprocessing, and generate combination of predictor and target variable for each realization.

2. For each realization, perform 1000 iterations of bootstrapping with Random Forest (RF) to generate RMSE distribution.

3. Analyze summary statistics of each RMSE distribution for each RF results.

Figure 2. Machine Learning model framework design

Results

- The primary error metric used to evaluate how well the model performed was RMSE between predicted and actual values CCl_4 mass.
- The training results of the 2nd set of predictors and target combinations outperform other models. (Figure 3).

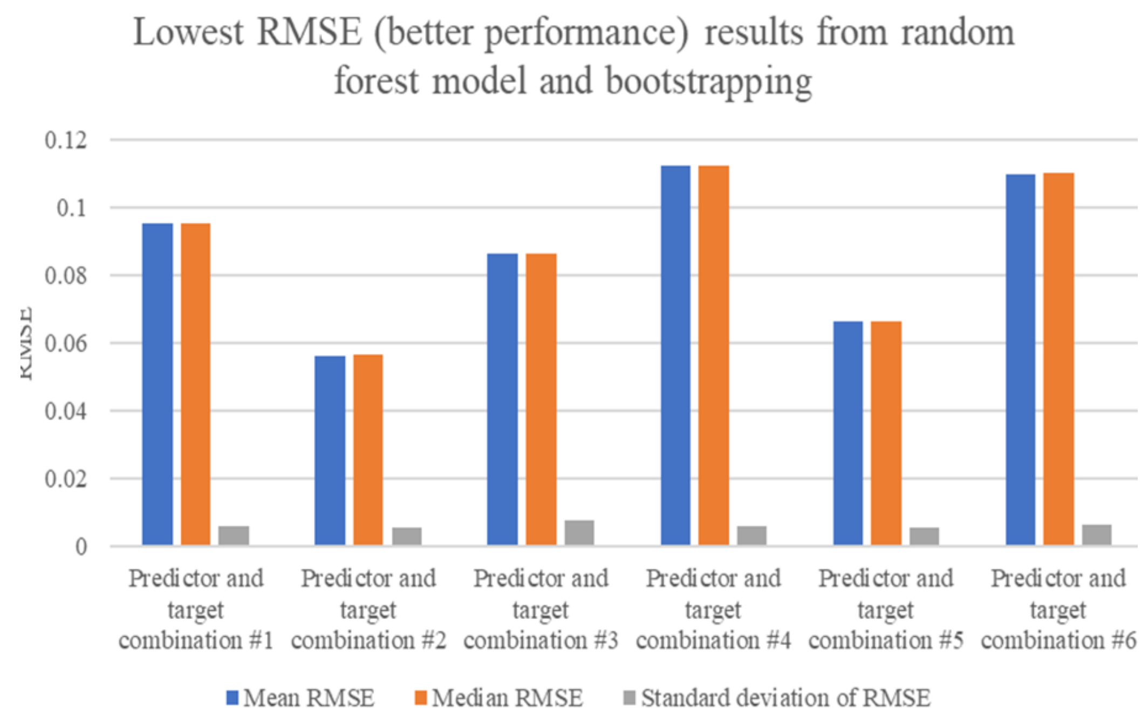


Figure 3. Random forest model with bootstrapping results for all combinations of predictor and target variables with lowest RMSE

- The best machine model was then used to determine the best geostatistical realization of plumes. (Figure 4).

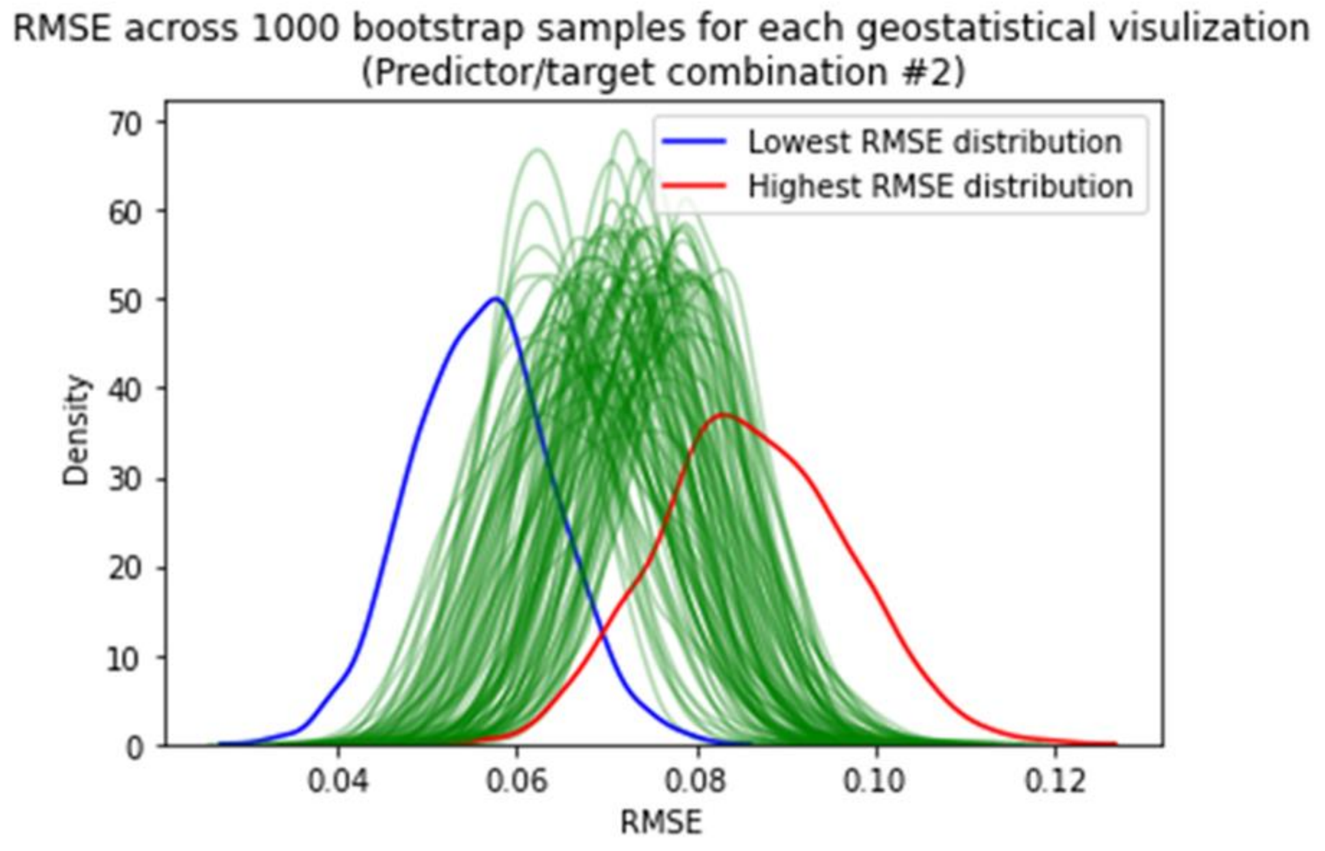


Figure 4: Distribution of RMSE errors across 100 geostatistical realizations using original well extraction rate and plume distribution data as predictors with CCl_4 mass as target (Combination #2)

Discussion

- The optimal predictor combination for machine learning model performance was the original well extraction rate and CCl_4 spatial plume distribution data.
- This study presents a novel data-driven method for evaluating contaminant plume geostatistical realizations using well pumping data.
- The evaluation does not require numerical modeling, making it a fast method for preliminary site assessment that can be applied to different sites.
- The results contribute to the ongoing P&T optimization work in the 200 West Area and to the broader remediation studies of the Hanford Site.

Acknowledgements:

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program.

I would also like to acknowledge the Deep Vadose Zone – Applied Field Research Initiative at Pacific Northwest National Laboratory and the U.S. Department of Energy Richland Operations Office.