STUDENT SUMMER INTERNSHIP TECHNICAL REPORT

pyLEnM Update: A Machine Learning and Data Analysis Python Package for Long-Term Soil and Groundwater Monitoring

DOE-FIU SCIENCE & TECHNOLOGY
WORKFORCE DEVELOPMENT PROGRAM

Date submitted:

December 10, 2021

Principal Investigators:

Aurelien Meray (DOE Fellow Student)
Florida International University

Haruko Wainwright, Ph.D. (Mentor)
Lawrence Berkeley National Laboratory

Ravi Gudavalli, Ph.D. (Program Manager)
Florida International University

Leonel Lagos, Ph.D., PMP® (Program Director)
Florida International University

Submitted to:

U.S. Department of Energy
Office of Environmental Management
Under Cooperative Agreement # DE-EM0005213

Applied Research Center
FLORIDA INTERNATIONAL UNIVERSITY

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

This research work has been supported by the DOE-FIU Science & Technology Workforce Development Initiative, an innovative program developed by the U.S. Department of Energy's Office of Environmental Management (DOE-EM) and Florida International University's Applied Research Center (FIU-ARC). During the summer of 2020, a DOE Fellow intern, Aurelien Meray, spent 10 weeks doing a summer internship at Lawrence Berkeley National Laboratory under the supervision and guidance of Research Scientist, Dr. Haruko Wainwright. The intern's project was initiated on June 1, 2021, and continued through August 6, 2021, with the objective of continuing the development of a pyLEnM package to support DOE-EM's Advanced Long-Term Monitoring Systems (ALTEMIS) project in effectively analyzing soil and groundwater monitoring datasets.

Recent technological advancements in geophysics, in-situ groundwater sensors, satellite-based remote sensing, reactive transport modeling, and artificial intelligence (AI), have led to great potential in establishing a new paradigm of long-term monitoring systems for contaminated groundwater sites with improved effectiveness and reliability. In situ sensors have proven to be a strong alternative to traditional groundwater sampling and laboratory analysis, especially when it comes to monitoring master variables, which are frequently leading indicators of plume movement change. Despite these advances, there are still issues to be solved, such as where to install additional sensors, determine which in situ variables provide the most information, and how to successfully anticipate plume movement using contaminant concentration estimations. The work described in this manuscript involves the development of the python package pyLEnM, a suite of machine learning algorithms to analyze monitoring datasets effectively. A lot of emphasis was placed on the development of an advanced spatial interpolation algorithm that uses a combination of regression and kriging techniques to accurately estimate contaminant plumes. Lastly, a sensor placement algorithm, which is built on top of the spatial interpolation method, was created to effectively select locations from a set of existing wells to maximize the capture of critical information for predictive modeling for new sensors.

1. INTRODUCTION

The main objective of this research is to continue the development of the pyLEnM package, which was initiated in mid-2020, which uses data science and machine learning to aid in the analysis of soil and groundwater data as part of the Artificial Intelligence effort for the Advanced Long-Term Monitoring Systems (ALTEMIS) project. The research builds on top of the previously developed functions to bring more complex functionality. A major focus was on spatial interpolation for estimating a contaminant plume along with a sensor placement algorithm for identifying key well locations to place new sensors. Lastly, several other functions for data transformation and visualization were developed during this internship. Once again, the historical dataset from the Savanah River Site (SRS) F-Area was used to validate the python package. All the available pyLEnM functions to date are shown in Figure 1 below.

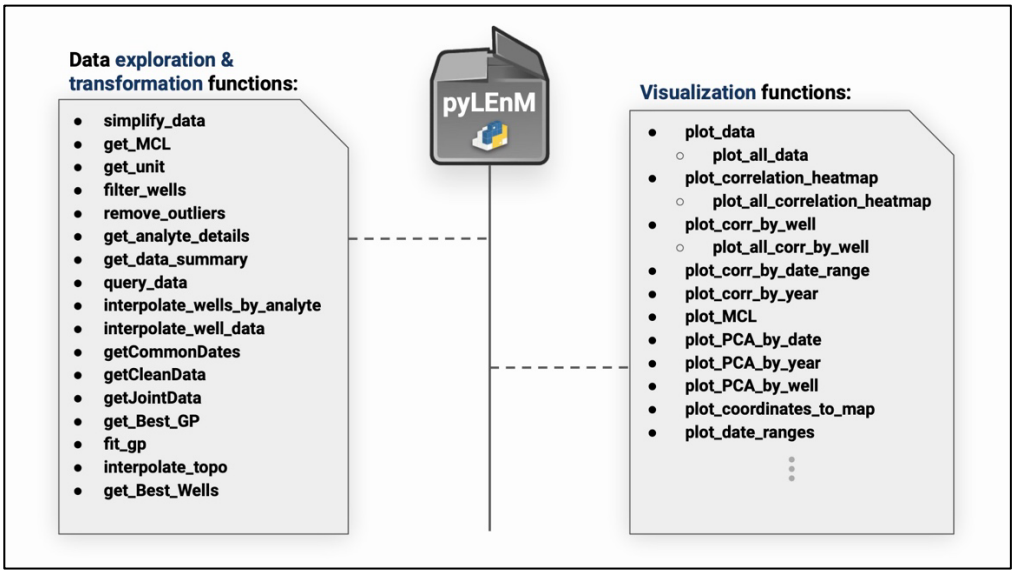

Figure 1. Overview of pyLEnM functions.

2. RESEARCH DESCRIPTION

2.1. Spatial Estimation

When monitoring contaminated sites, it is important to identify the plume boundary to ensure its containment as well as track its movement. This emphasizes the significance of using an accurate and realistic spatial estimation approach. Typically, there are a limited number of wells/sensors that collect analyte samples at a fixed location which are then interpolated to the surrounding areas. The Gaussian Process (GP) is a popular interpolation method which assumes a multivariate normal distribution and only uses data from the given points. Our goal for this research was to develop an approach which can further improve the traditional GP method.

Our approach provides the estimation algorithm with more than just the values at each well location. Using NASA's public digital elevation dataset (SRTM) at 30-meter resolution, along with other features such as terrain slope, flow accumulation and the distance from the basin, the algorithm used these predictors to generate a more accurate map of the plume. We establish a relationship between given predictors and the concentrations at the wells using various regression models of choice. We provide the user with four models to choose from which are linear, lasso, ridge, and random forest regression. The several options are intended to accommodate varied data trends; some may follow linear trends while others do not. A regression model is first trained on the coordinates, the additional predictor(s) provided, and the concentration at each well and predicted on the entire site coordinates. Additionally, a GP model is trained on the well coordinates and the residual concentrations provided by the regression model and then predicted on the site coordinates. The final estimation is calculated by adding the two predictions produced by the regression and GP models. The main idea is that running a GP on the residuals should capture factors missed by the regression part. A visual description on how the spatial estimation algorithm works is shown in Figure 2.

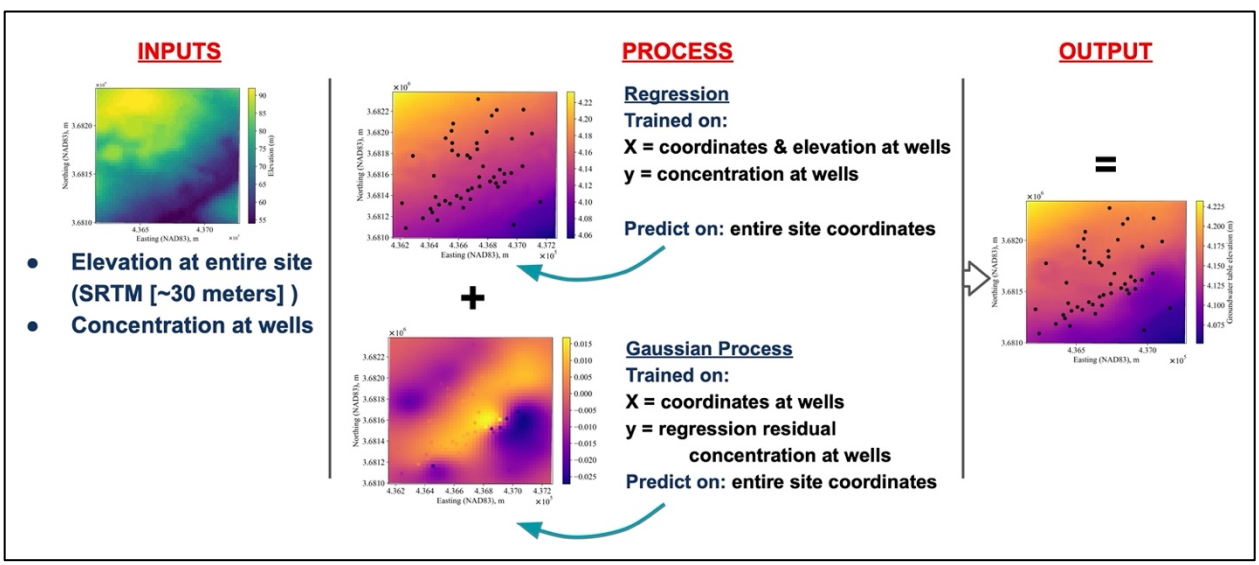

Figure 2. Spatial estimation process.

For this research, we ran the spatial estimation on both the water table (WT) and tritium concentrations. The quality of the estimation was measured both in terms of mean squared error (MSE) and the R^2. Furthermore, due to the limited number of well data available, the algorithm was tested twice: once utilizing all the data for the fitting process and again using leave-one-out cross validation (LOOCV). These results are discussed in Section 3.

2.2. Sensor Placement Optimization

Another focus of the research was on developing a sensor placement algorithm to strategically choose a subset of existing well locations to place new sensors. The main idea was to find a configuration of a limited number of sensors that can spatially estimate with close to identical resolution to a map estimated on all possible sensors (reference field). Algorithmically, the main idea of the optimization was to minimize the overall error between the high-quality reference field and the spatially interpolated map with the 15-20 subset of wells. The MSE was used as the error metric for this problem. The selection of the subset of well locations was performed using the algorithm described below.

Individual time-step sensor placement selection algorithm (pseudocode) description:

- **Input:**
  - List of all available well locations to choose from
  - [max_wells]: Maximum number of wells to select
  - List of starting wells (if any)
- **Algorithm:**
  1. If there is a list of starting wells, they will be chosen first as the optimal locations.
  2. This process is run for ([max_wells] - # of starting wells) number of times:
     - **a.** For each of the available wells remaining:
       - **i.** A GP is created using the current well + the starting wells + previously selected wells.
       - **ii.** The mean squared error (MSE) is calculated between the GP from **Step 2.a.i.** and the Ground truth GP.
     - **b.** The well combination that contributes the least MSE from **Step 2.a.** is chosen and added to the list of selected wells.
- **Output:**
  - List of selected well names in the order from most optimal to least optimal.

2.3. Divers Functions

A variety of other functions were also created during the internship. One of the functions is a visualization for viewing a particular analyte. This visual can depict a lot of meaningful information about the analyte. Figure 3 is an example of the visual plotted for tritium at the F-Area. The time series for each well is plotted vertically where we can determine when the first and last sample was recorded. Additionally, the concentrations can be seen by matching the colored points with the color bar on the right-hand side. Lastly, potential trends can be identified since the wells are ordered by increasing distance from the center of the basin. In certain cases, like with the WT, we can construe that the further away the wells are from the basin, the lower the concentration is.

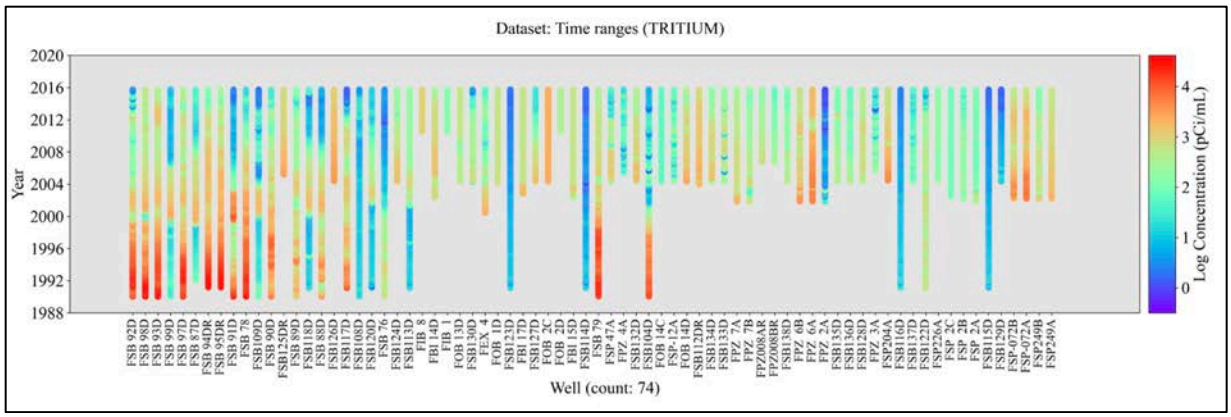

Figure 3. Time series visual for tritium.

The F-Area historical dataset was created from scientists manually collecting water samples at well locations. The days that samples were collected are spread out over days, sometimes weeks. For analysis, this poses a challenge since samples are timestamped differently. For machine learning algorithms data needs to be spaced in equal intervals to be processed. To solve this issue the *getJoinData* function was created which returns a new data frame with the index being time ranges instead of individual timestamps. The only parameters the function requires are the analytes to use and the how many days to look forward and backward from a specific date specified by a lag value. For example, a lag of 2 means to bucket information with a range of 5 days (2 past days + current day + 2 future days). An example of how the function works is shown in Figure 4. Overall, this method is good at preserving information that is otherwise lost if dates do not coincidently line up together.

	Well 1	Well 2	Well 3	Well 4	Well 5
Time 4			3.00	4.00	
Time 5	4.00	2.00			
Time 6					3.00
Time 7			3.50		
Time 8					
Time 9	5.00				2.00

RAW data

	Well 1	Well 2	Well 3	Well 4	Well 5
Time 2 - Time 6	4.00	2.00	3.00	4.00	3.00
Time 3 - Time 7	4.00	2.00	3.25	4.00	3.00
Time 4 - Time 8	4.00	2.00	3.25	4.00	3.00
Time 5 - Time 9	4.50	2.00	3.50	4.00	2.50
Time 6 - Time 10	5.00	2.00	3.50	4.00	2.50
Time 7 - Time 11	5.00	2.00	3.50	4.00	2.00

Bucketing + Interpolate

Lag = +/- 2 days

Figure 4. Example of the transformation performed by the getJointData function.

The *getJointData* function was implemented into many other functions of pyLEnM, such as the correlation and the PCA plots, to preserve more data and therefore get more accurate results.

2.4. Demonstration Notebooks

Creating demonstration notebooks was an important part of the internship. The idea was to create interactive notebooks that can help other people get started with the pyLEnM package by providing

them with a code walk through. Three notebooks were created: the first one is the Basics, the second is the unsupervised learning functions and the third notebook is the supervised learning and well optimization. The pyLEnM Basics notebook covers how to properly ingest the dataset, how to get information on all the functions such as the parameters and descriptions, along with exploration functions. The second notebook covers functions like correlation analysis and principial component analysis (PCA). Lastly, the final notebook covers spatial estimation where the example using WT is used along with an example of the sensor placement optimization function. Screenshots of the 3 notebooks can be seen in Figure 5.

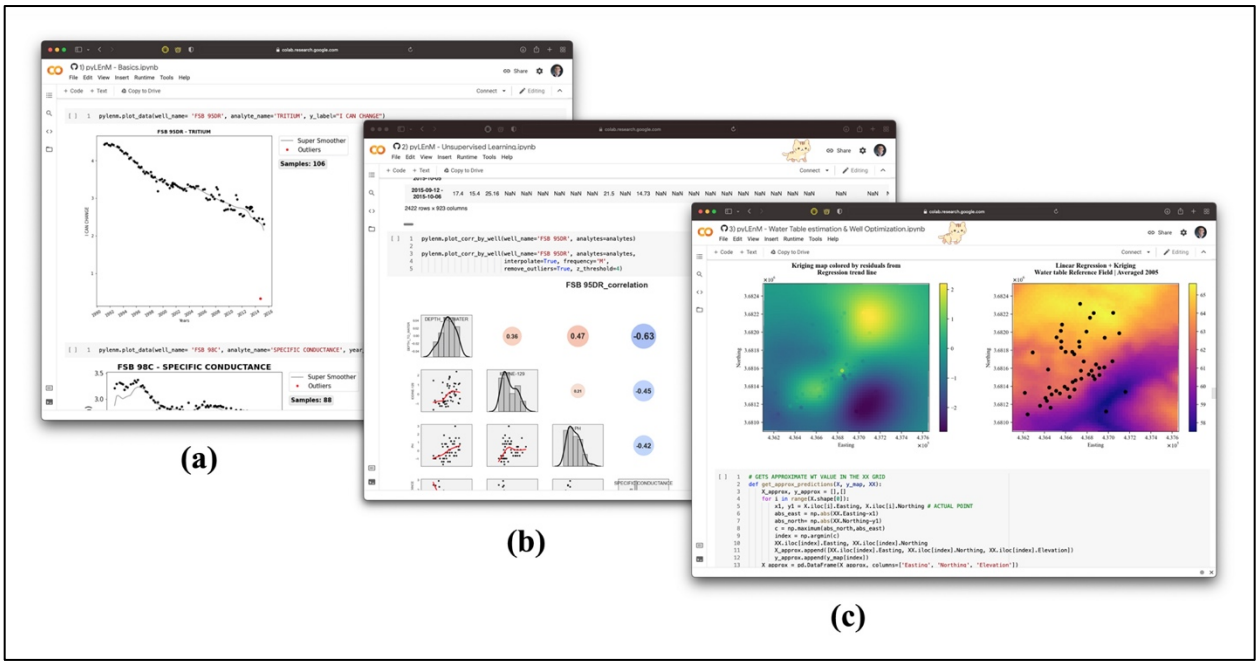

Figure 5. Demonstration Interactive Python Notebooks (IPYNB)
(a) Basics, (b) Unsupervised learning, (c) Spatial estimation and well optimization.

3. RESULTS AND ANALYSIS

3.1. Spatial Estimation

To evaluate the best possible model for estimating the WT and tritium plume, a script was devised to evaluate the data with different models and predictors. The script was executed with two variations, one for the fitting process and the other using the LOOCV. The top 3 models for each variation and analyte are shown in Table 1 below.

Table 1. Top Results for the WT and Tritium Spatial Estimation

	Fitting Process Results				LOOCV Results			
	Model	Features	MSE	R^2	Model	Features	MSE	R^2
Water Table	RF + GP	Easting, Northing	2.30E-07	0.9983	RF + GP	Easting, Northing, Elevation	1.80E-05	0.8663
	Lasso + GP	Easting, Northing, Elevation, Slope	2.67E-07	0.9981	RF + GP	Easting, Northing, Elevation, Slope, Flow Accumulation	1.90E-05	0.8646
	Ridge + GP	Easting, Northing, Elevation, Slope	2.83E-07	0.9980	RF + GP	Easting, Northing, Elevation, Slope	1.90E-05	0.8602
	GP		5.92E-07	0.9957	GP		2.40E-05	0.8272
Tritium	Lasso + GP	Easting, Northing, Elevation, Slope, Flow Accumulation	3.01E-03	0.9959	Linear + GP	Easting, Northing, Elevation, dist_to_basin	4.05E-01	0.4456
	Lasso + GP	Easting, Northing, Elevation, Slope	3.01E-03	0.9959	Ridge + GP	Easting, Northing, Elevation, dist_to_basin	4.06E-01	0.4444
	Ridge + GP	Easting, Northing, Elevation, Slope	4.77E-03	0.9935	Linear + GP	Easting, Northing, Elevation, Slope, dist_to_basin	4.24E-01	0.4190
	GP		3.04E-01	0.5839	GP		4.65E-01	0.3628

For the WT, both in the fitting process and the LOOCV, the results are very good. The regular GP already performed well with an R^2 of 0.9957, so our best 3 models did not improve significantly from the baseline. On the other hand, during the LOOCV evaluation, the R^2 increased by about 5% from 0.8272 to 0.8686. The predicted map using the Lasso regression method is shown in Figure 6b. The tritium fitting process does very well but tends to overfit the data. This would explain the discrepancy between the best fitting process result (R^2: 0.9959) and the best LOOCV result (R^2: 0.4457). Although the LOOCV result is relatively poor, we still get an improvement of about 22% from the GP baseline. The predicted tritium maps are shown in Figure 6c and Figure 6d, best fitting process model and best LOOCV model respectively. Although the fitting process model has a better accuracy, the resulting map does not quite resemble the expected output. That said, the best LOOCV model, despite having a poor accuracy, has a more realistic depiction of the plume with the proper boundary.

3.2. Sensor Placement Optimization

For the purpose of demonstrating the algorithm, the average 2015 WT values were used. Figure 7 depicts the iterations of the algorithm as it selects optimal wells one at a time. There is a significant decrease in error during the selection of the first 5 wells after which it slows down. Notice how overall, at each iteration, the MSE decreases as the number of selected wells increase. Even though the interest is to select about 15-20 locations, we set the maximum number of wells parameter to 30 to see the evolution of the spatial estimation. Looking at the ith row of Figure 7, the detail of the maps also increases as a function of the number of selected wells.

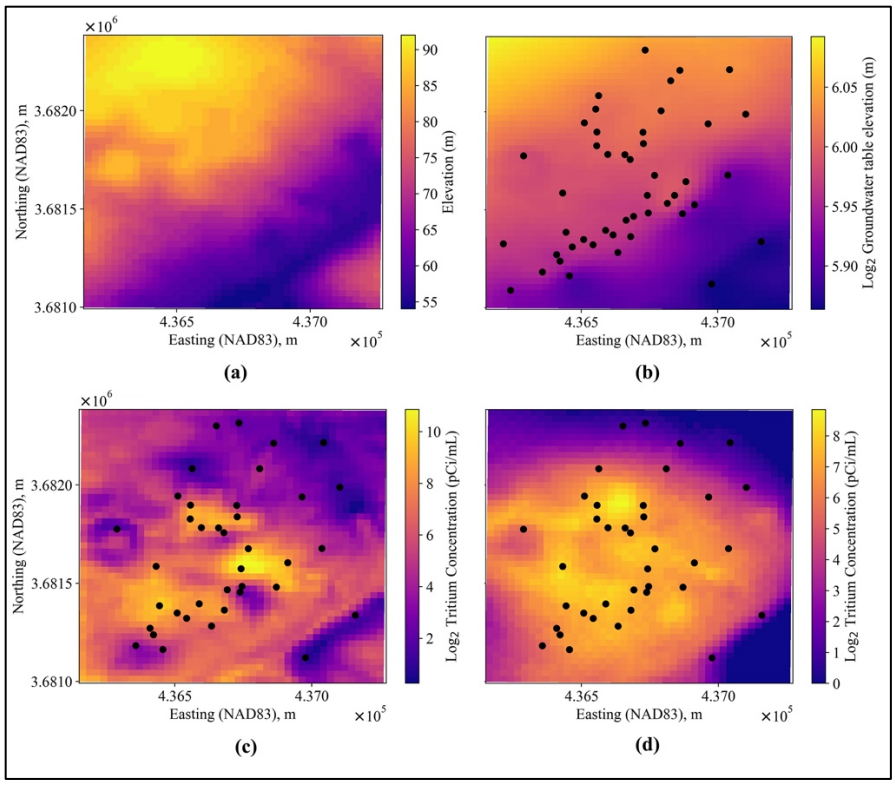

Figure 6. Spatial estimation results (a) SRTM elevation heatmap, (b) Best fitting process WT spatial estimation map, (c) Best fitting process tritium estimation map, (d) Best tritium LOOCV estimation map.

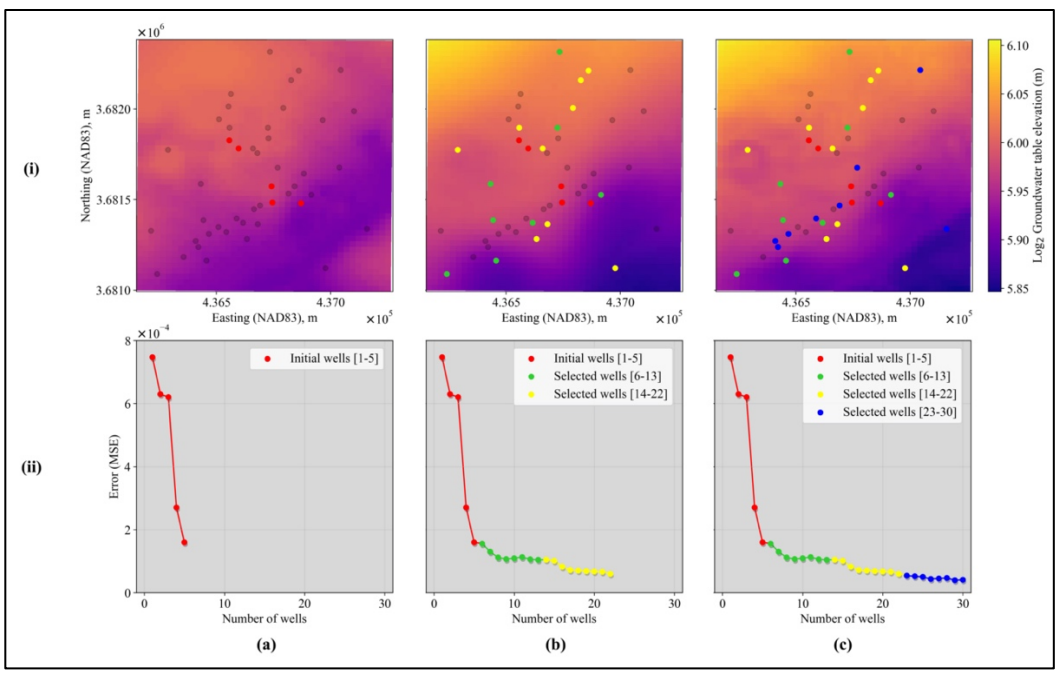

Figure 7. Sensor placement optimization on averaged 2015 WT data.

4. CONCLUSION

The internship was a great learning experience with many successful results. Mainly, an effective spatial estimation algorithm was devised to estimate contaminant plumes at the SRS F-Area. In addition, an easy-to-use sensor placement optimization algorithm was introduced which can be used for variety of different purposes. The functions and the respective results on the demonstrated dataset are a great addition to the pyLEnM package. In the near future, we hope to see this python package be used to evaluate additional contaminated datasets and solve other challenges.

5. REFERENCES

[1] Denham, Miles E., et al. "A New Paradigm for Long Term Monitoring at the F-Area Seepage Basins, Savannah River Site." 2019, DOI: 10.2172/1504623.

[2] Schmidt, Franziska, et al. "In Situ Monitoring of Groundwater Contamination Using the Kalman Filter." *Environmental Science & Technology*, vol. 52, no. 13, 2018, pp. 7418–7425., DOI: 10.1021/acs.est.8b00017.

[3] Wainwright, Haruko, et al. "Objective." *ALTEMIS*, 19 Aug. 2020, altemis.lbl.gov/about/.