Title

PyLEnM: A Machine Learning Framework for Long-Term Groundwater Contamination Monitoring Strategies

Journal

Environmental Science and Technology, 56(9)

ISSN

0013-936X

Authors

Meray, Aurelien O

Sturla, Savannah

Siddiquee, Masudur R

et al.

Publication Date

2022-05-03

DOI

10.1021/acs.est.1c07440

Peer reviewed

Environmental Science & Technology

pubs.acs.org/est

Article

PyLEnM: A Machine Learning Framework for Long-Term Groundwater Contamination Monitoring Strategies

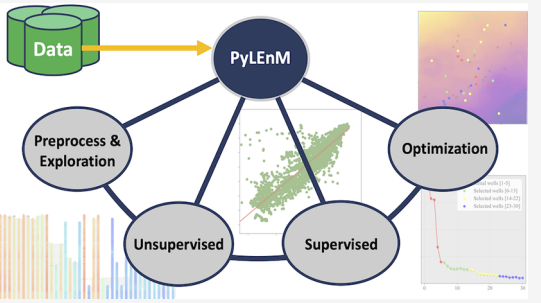
Aurelien O. Meray, Savannah Sturla, Masudur R. Siddiquee, Rebecca Serata, Sebastian Uhlemann, Hansell Gonzalez-Raymat, Miles Denham, Himanshu Upadhyay, Leonel E. Lagos, Carol Eddy-Dilek, and Haruko M. Wainwright*

Cite This: *Environ. Sci. Technol.* 2022, 56, 5973−5983

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this study, we have developed a comprehensive machine learning (ML) framework for long-term groundwater contamination monitoring as the Python package PyLEnM (Python for Long-term Environmental Monitoring). PyLEnM aims to establish the seamless data-to-ML pipeline with various utility functions, such as quality assurance and quality control (QA/QC), coincident/colocated data identification, the automated ingestion and processing of publicly available spatial data layers, and novel data summarization/visualization. The key ML innovations include (1) time series/multianalyte clustering to find the well groups that have similar groundwater dynamics and to inform spatial interpolation and well optimization, (2) the automated model selection and parameter tuning, comparing multiple regression models for spatial interpolation, (3) the proxy-based spatial interpolation method by including spatial data layers or in situ measurable variables as predictors for contaminant concentrations and groundwater levels, and (4) the new well optimization algorithm to identify the most effective subset of wells for maintaining the spatial interpolation ability for long-term monitoring. We demonstrate our methodology using the monitoring data at the Savannah River Site F-Area. Through this open-source PyLEnM package, we aim to improve the transparency of data analytics at contaminated sites, empowering concerned citizens as well as improving public relations.



KEYWORDS: *open-source package, machine learning, spatial estimation, sensor placement optimization, Gaussian process model, unsupervised learning, groundwater contamination*

INTRODCTION

Long-term monitoring is increasingly important for contaminated soil and groundwater sites.[1] It has been more than 40 years since the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) was passed to establish the Superfund sites in the US in 1980. Among the 1344 sites listed on the National Priorities List, cleanup has been completed at only 447 of them as of January 2022.[2] There is a growing recognition that current remediation technologies have limited effectiveness and that residual contaminants—at low levels but still above regulatory limits—are difficult to completely clean up. In response to this problem, sustainable remediation has emerged as a key concept to address such sites over the past decade.[3] Sustainable remediation considers net environmental impacts, including such side effects as waste production, noise/traffic/air pollution associated with heavy machinery and dump trucks, ecological disturbance, energy use, and greenhouse gas emission. It promotes the transition from intense soil removal and treatments to more sustainable, passive remediation approaches, as well as monitored natural attenuation (MNA). Longer institutional control and monitoring is often required, possibly for decades.

The objectives of long-term monitoring—different from initial characterization and remediation stages—are (1) to confirm the system stability and continuing reduction of contaminant and hazard levels, (2) to provide assurance to the public and prevent dissemination of false or misleading information, and (3) to detect changes or anomalies in contaminant mobility (if they occur) or discover any unexpected processes or events. In fact, there have been several examples in which long-term monitoring found that the contaminant concentrations were not decreasing as rapidly as originally predicted by models and led to improved conceptual models.[4] In contrast to emergency responses or site character-

Received: November 9, 2021
Revised: March 8, 2022
Accepted: March 21, 2022
Published: April 15, 2022

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

ization, long-term monitoring has to be carefully planned, considering cost, spatial coverage, and the priorities of the stakeholders. Historical data sets accumulated at the sites over years can greatly facilitate development of long-term monitoring strategies.

A variety of statistical and machine learning (ML) methods have been developed to discover hidden patterns and key factors in vast data sets and to improve groundwater monitoring or environmental contamination monitoring. The most common uses have been supervised learning to estimate the spatiotemporal distributions of contaminant concentrations or groundwater levels.[5,6] In addition, unsupervised learning approaches have been used to identify the correlations among different contaminant concentrations and/or in situ measurable parameters,[5] as well as to find the groups of wells that have similar groundwater dynamics.[7] At the same time, ML can augment or support decision making processes by compressing vast amounts of data into digestible information. One of the critical decision making steps for long-term monitoring is to determine the number of sufficient wells and their locations. There have been monitoring optimization algorithms based on spatial interpolation[8] as well as principal component analysis.[9,10]

The implementation of these methods to real-world applications is, however, still quite limited. MNA requires regular groundwater sampling at the wells, with regular frequency prescribed by regulators. However, such monitoring is often conducted mostly for compliance purposes; data are often simply archived without any analytics. The well locations are determined primarily based on expert judgments, including regulators' opinions. The challenge has been a lack of general pipelines from monitoring data to ML. Although there is commercial software available for groundwater monitoring and data visualization/analysis, their data analysis methods and their extensibility are often limited, without connection to recent advances in open-source ML libraries such as python scikit-learn.[11]

In this study, we aim to develop a framework to support long-term groundwater monitoring at the contaminated sites. Specifically, we seek to develop a python package, PyLEnM (Python for Long-term Environmental Monitoring), which defines the ML pipeline and workflow from data to ML through a collection of commonly used functions for monitoring data analysis. A particular focus is to extract critical information from historical data sets since MNA builds upon a large quantity of historical monitoring and characterization data. The novel aspects of our framework include (a) the new summarization/visualizations of spatiotemporal groundwater data, (b) flexible ways to find coincident and/or colocated data for developing a data-driven relationship, (c) the seamless integration of publicly available data such as surface elevation for creating predictors in ML, (d) the automated comparison/selection of multiple ML algorithms for spatial interpolation, (e) proxy-based spatiotemporal interpolation to integrate data-driven relationships for estimating groundwater table (WT) and contaminant concentrations, and (f) a new well-placement optimization algorithm.

The open-source package is based on the Jupyter iPython notebook, which can document the workflow from raw data to data analytics and visualization. Through this package, we aim to accelerate the process for developing new ML algorithms and functions for the monitoring community. We demonstrate this framework at the Savannah River Site (SRS) F-Area, where the historical data sets have been well-curated and archived. We make all the codes and data sets available for the community (Text S2). In addition, such public data can serve as benchmark data sets to develop and test different ML algorithms, ensuring the FAIR principle (findability, accessibility, interoperability, and reusability). The transparency of the monitoring data analytics workflow is particularly important for the contaminated sites with respect to public acceptance and assurance.

METHODOLOGY

Study Site and Demonstration Data Sets. In the SRS F-Area (Aiken, SC, USA), low-level radioactive waste from nuclear fuel reprocessing was discharged into three unlined seepage basins between 1955 and 1988.[5,12,13] Currently, an acidic plume, containing tritium (H-3), iodine-129 (I-129), uranium-238 (U-238), and nitrate, extends from the basins to about 600 m downgradient toward the local creek, Fourmile Branch. The main plume is located in the unconfined aquifer above a thin clay-rich layer. A pump-and-treat system was installed in 1997 and then replaced by passive remediation in 2004 using a hybrid funnel-and-gate system to inject alkaline solutions at the gates for neutralizing the acidic groundwater and enhancing the sequestration of cationic contaminants such as U-238.

The original data set used in this study was curated by the SRS containing over 400 analytes (including heavy metals, organic contaminants, and major cation/anion concentrations) from 1990 to 2015 (Tables S1 and S2). The groundwater sample collection and analysis were defined in the Resource Conservation and Recovery Act (RCRA) Permit at this site.[14] We demonstrate the PyLEnM capabilities with a subset of the F-Area data, including groundwater table levels, pH, specific conductance (SC), and tritium and uranium concentrations. Tritium is the contaminant that has been the main contributor to the radiological dose calculation,[1] while uranium has the largest mass among all the radionuclides.[12] The water table is a critical parameter, defining the hydrological boundary conditions and controlling plume migration. pH and SC are the in situ measurable parameters that can be measured continuously based on in situ sensors.[5]

PyLEnM Framework. The main components of the PyLEnM[15] workflow are designed to (1) facilitate data exploration through various data summarization and visualization processes, (2) identify the spatiotemporal patterns of covaried contaminant concentrations and groundwater table dynamics as well as identify groups of wells that behave similarly through unsupervised learning, (3) estimate the contaminant concentrations and groundwater table through supervised learning, and (4) optimize the selection of long-term monitoring wells among existing ones (Figures 1 and S1).

PyLEnM takes advantage of existing python packages (NumPy[16] and SciPy[17]) for scientific computing, Pandas[18] for data analysis and manipulation, scikit-learn[19] for ML, pyProj[20] for spatial projection, Matplotlib[21] and Seaborn[22] for statistical visualization, and ipyleaflet[23] for map visualization (Text S3). PyLEnM assumes a SQL or relational database with two tables: an analyte table for spatiotemporal data, storing monitoring data at different wells and times (including well names, date/time, concentrations, units, error range, and analyte names) (Table S1), and a well table for well information (such as their coordinates, surface elevations, screen depths, aquifer, and construction/decommission dates)

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

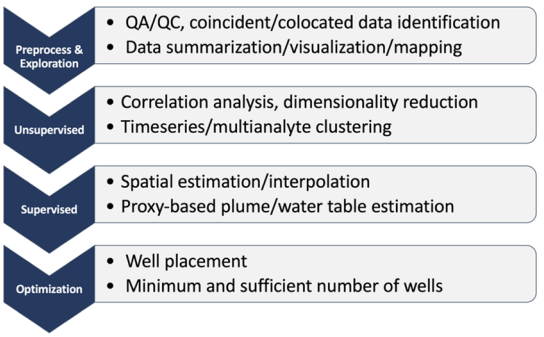


Figure 1. Flowchart of PyLEnM capabilities.

(Table S2). The well name acts as the SQL key (i.e., the unique identifier) between the two tables.

Exploratory Data Analysis. The basic PyLEnM functions include data summarization capabilities that provide the users with a swift overview of the spatiotemporal data and well information defined above, such as compiling the list of (1) wells available for each or selected analyte (get_analyte_details) and (2) analytes available for each well (get_well_analytes). These summary tables are also accompanied by the number of data points, the start and end dates, and the average and percentile values. In addition, filtering can be performed by the well name, date range, aquifer, and others in the same manner as that of the Pandas framework. In parallel, we have implemented several automated quality assurance and quality control (QA/QC) functions for time series data, including curve fitting (plot_data) and removal of outliers (remove_outliers). In the outlier removal, we can assign different fitting functions (e.g., Friedman's super smoother) and threshold values to identify outliers. PyLEnM also includes multiple visualization functions: time series plots with linear/nonlinear interpolation and the identification of outliers (Text S1 and Figure S2). In addition, there is a time range visualization functionality in which the start and end dates are plotted vertically by a unique well so as to identify the common sampling ranges and concentration changes.

Environmental data analytics begins by identifying the coincident/colocated data sets—that is, the different analytes at the same times and same wells—so that we can establish a data-driven relationship. Groundwater concentrations may not change so rapidly, so the data sets collected within a few days or longer may be considered coincident. In addition to standard gap filling and linear interpolation, we have created a function, getJointData, to identify coincident data with flexible time lags. The function takes the user-specified time lag (e.g., 1 week or 1 month) as a parameter and identifies the data points (from the different analytes and wells) that fall into each time period. This is important since a groundwater sampling campaign could take at least a few days or weeks. This process maximizes the integrity of the data prior to ML as well as avoids artifacts often created by gap filling.

Unsupervised Learning. Unsupervised learning generally consists of correlation analyses, dimensionality reduction (such as principal component analysis; PCA), and clustering. We have implemented the correlation analyses and PCA that were demonstrated in Schmidt et al. (2018) to identify the covariability among different analytes. PyLEnM quantifies the correlation between two time series with linear (Pearson) or nonlinear (Spearman or Kendall) correlation coefficients. The individual scatter plots embedded in the correlation plot can assist the user in determining which coefficient is the most appropriate. PCA compresses the correlations among multidimensional analytes into several principal components and facilitates the visualization of covaried analytes (Figure S4).

PyLEnM's unsupervised learning begins with the coincident data points (among different wells or different analytes) identified by the getJointData function or the colocated data points at the same well identified by querying well names. Clustering is then applied for identifying several groups of wells that have similar groundwater dynamics (Hastie et al., 2001). PyLEnM includes the *k*-means and hierarchical clustering methods and or distance measures or criteria (such as the Ward and complete linkage criteria in hierarchical clustering) that have been commonly used in environmental data analytics.[24,25] In addition to the PCA developed by Schmidt et al. (2018) to identify covaried multiple analytes at each well, we implemented the time series clustering,[26] which groups the wells according to the temporal dynamics of one analyte. The group of wells can then be mapped back in space to evaluate their spatial arrangement.

Supervised Learning. Supervised learning methods are used to estimate contaminant concentrations and groundwater elevation by interpolating between sparse wells. In contrast to common interpolation methods such as inverse distance-weighted interpolation, PyLEnM can accommodate known or site-specific predictors such as elevation, topographic metrics, and the distance to the source for further constraining the estimation. In particular, the algorithm ingests the publicly available surface elevation across the world (NASA SRTM Digital Elevation 30 m) through an application programming interface (API) and then computes topographic metrics [such as the topographic wetness index (TWI) and slope].

For the spatial interpolation, PyLEnM first builds a regression of sparse groundwater data as a function of these predictors using scikit-learn. The residual is interpolated based on the Gaussian process model (GPM), which captures the spatial correlations based on a covariance model such as the Matern covariance. PyLEnM also makes use of the GridSearchCV function in scikit-learn to optimize the covariance parameters. The regression performance is quantified based on the mean squared error (MSE) and R^2, both in the fitting process and in the leave-one-out cross validation (LOOCV). Compared to other cross-validation methods such as the k-fold cross validation, LOOCV is known to be effective at evaluating a model's performance with a limited number of data points, which is common in the environmental data sets.[28] PyLEnM automates this interpolation process, including parameter tuning, as well as the comparison of multiple supervised learning algorithms such as random forest (RF), Lasso regression, and Ridge regression, in addition to traditional linear regression methods.[27] This allows us to compare multiple algorithms and select the most appropriate one.

In addition, we developed an algorithm to estimate contaminant concentrations based on proxy in situ measurables (such as SC). This algorithm builds on the concept proposed by Schmidt et al. (2018), who estimated contaminant concentration time series based on in situ measurable SC and pH as proxies. The algorithm begins by building a regression of contaminant concentrations as a function of proxy variables. Assuming that the correlations are consistent over time and space, we use all the wells and time points for a particular contaminant. This regression results in the contaminant concentrations estimated at any given time

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

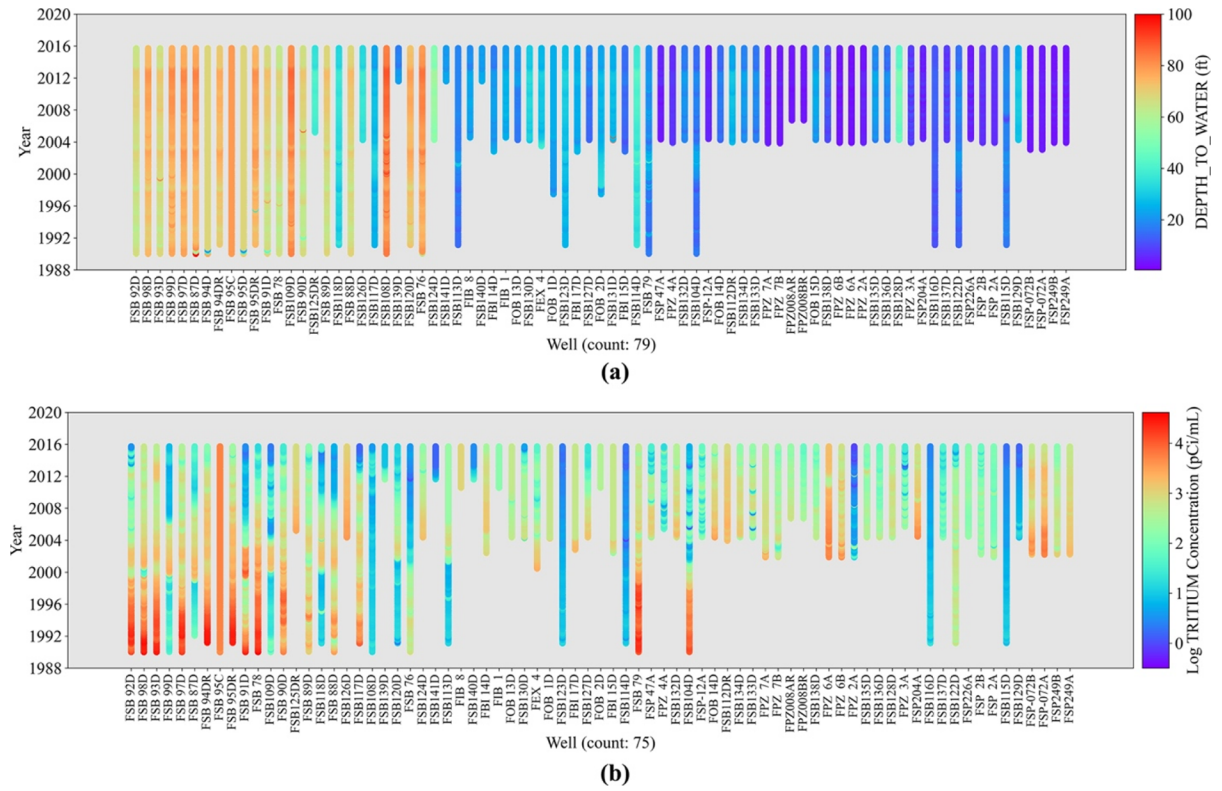


Figure 2. Time series and concentration visualization for (a) WT and (b) tritium sorted by the increasing well distance (left to right) from the center of the F-Area basin.

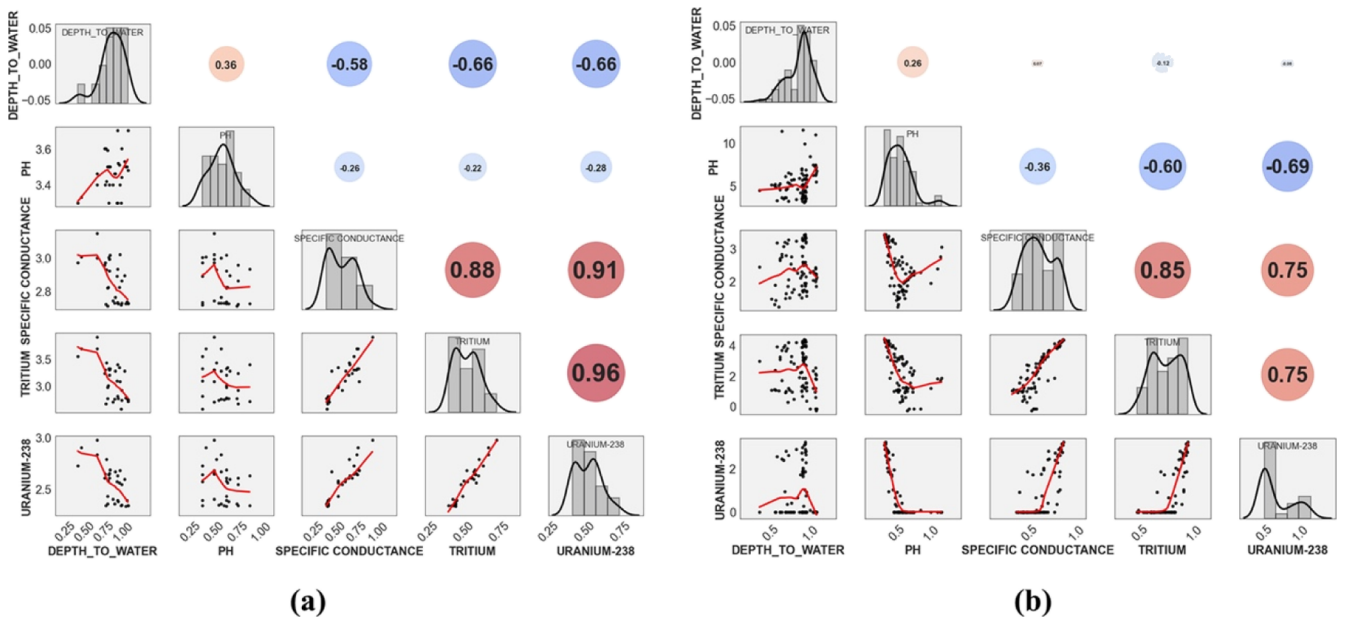


Figure 3. (a) Temporal correlation plot among analytes at FSB9SDR between 02/09/1993 and 07/30/2013 log concentrations. (b) Spatial correlation plot for all wells among analytes on 02/21/1993 with a lag of 12 days (02/09/1993 to 03/05/1993) log concentrations. The numbers in the circles on the upper diagonal are the Pearson correlation coefficients, and the size of the bubble represents the strength of the correlation. In addition, the red lines depict the pairwise data trend.

and at all the wells where the proxy variables are available. Finally, the same interpolation algorithm above is used to estimate the spatial distribution of contaminant concentrations over space.

Well Placement Optimization. The goal of well placement optimization is to capture the spatial heterogeneity of the plume or groundwater table with the fewest number of wells. We assume that the regression described above provides a reasonable spatiotemporal estimation as a reference or ground-truth field based on historical monitoring data. The algorithm builds on Sun et al. (2020), using a greedy approach[29] such that it selects one additional well at each iteration within the

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

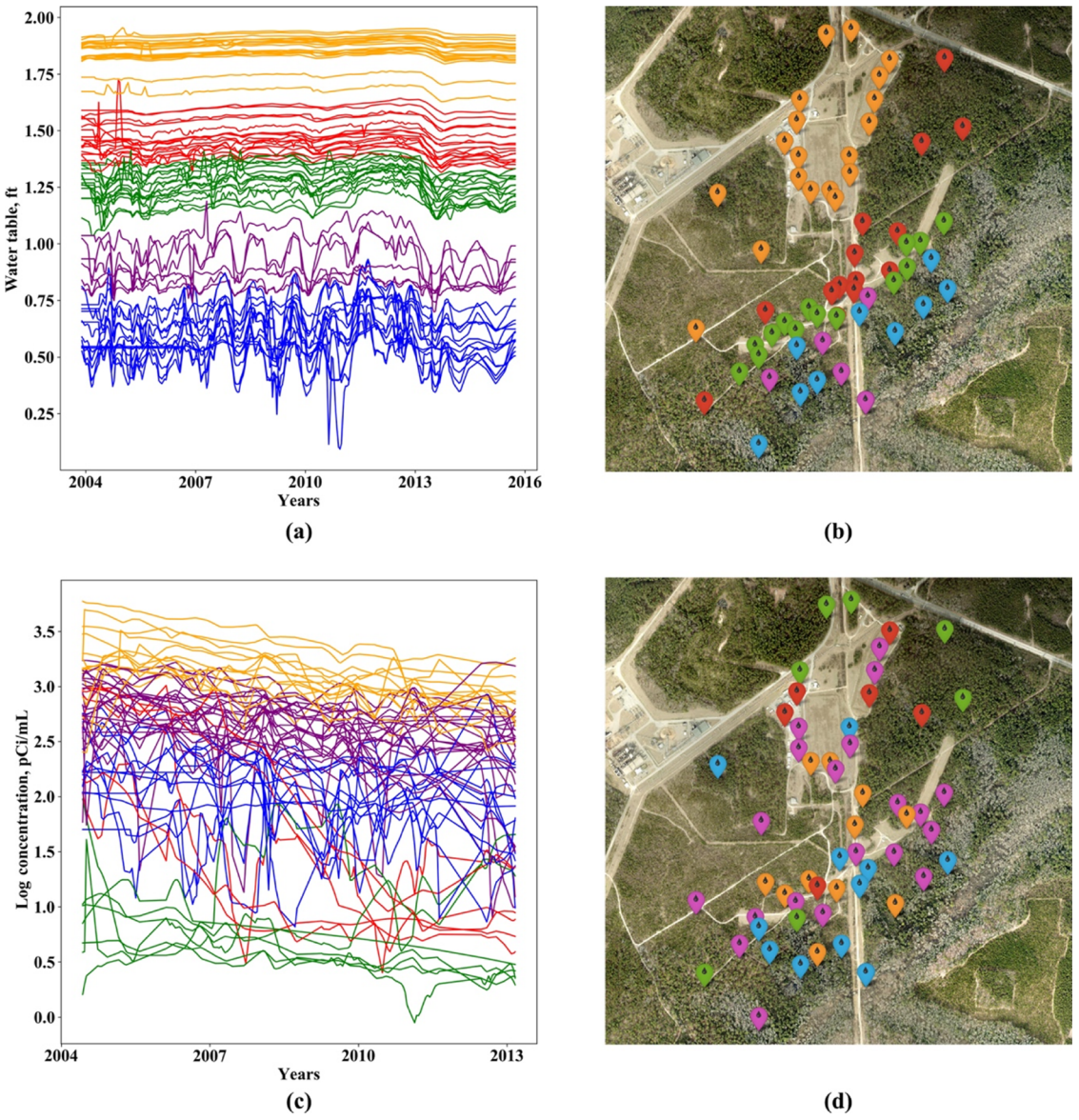


Figure 4. Time series clustering of (a,b) water table levels and (c,d) tritium concentrations. (a,c) show the time series and (b,d) show the well locations on the map according to their assigned cluster colors.

currently available monitoring wells. At each iteration, the algorithm performs spatial interpolation with every potential well location and selects the well that minimizes the MSE over all the pixels compared to the reference map. This process is repeated until the MSE converges or the MSE falls lower than the required threshold.

RESULTS

The data summary functions (get_data_summary and get_analyte_details) create the tables to concisely visualize the data availability (i.e., start/end dates and the number of samples) and summary statistics (mean and standard deviation) for the specified analytes at all the wells (Table S3) or each well (Table S4). Figure 2 demonstrates the new visualization tools, compressing the concentration time series at multiple wells as well as the data availability range. The wells are arranged according to the distance to the basin in this case, although the order of the wells can be specified by the users. This visualization facilitates identifying the disparity between the collected data where half of the wells started sampling in the mid-1990s, and the other half started in the mid-2000s. In addition, we can observe that the water table elevation is consistently higher in the upgradient wells near the source zone (Figure 2a), while the tritium concentration changes in time and space and is associated with a plume migration as a function of distances from the source (Figure 2b).

The correlation plots identify the covariability among the analytes spatially and temporally, particularly between the in situ variables (pH and SC) and contaminant concentrations (Figure 3). The correlations are generally consistent between the temporal variability at one well (Figures 3a and S3) and the spatial variability on a selected date (Figure 3b); the correlations are high among SC, tritium, and uranium concentrations, with the Pearson coefficients reaching as high as 0.96. In addition, the scatter plots show the nonlinear relationship between pH and the contaminant concentrations. At the same time, the water table depth is negatively correlated

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

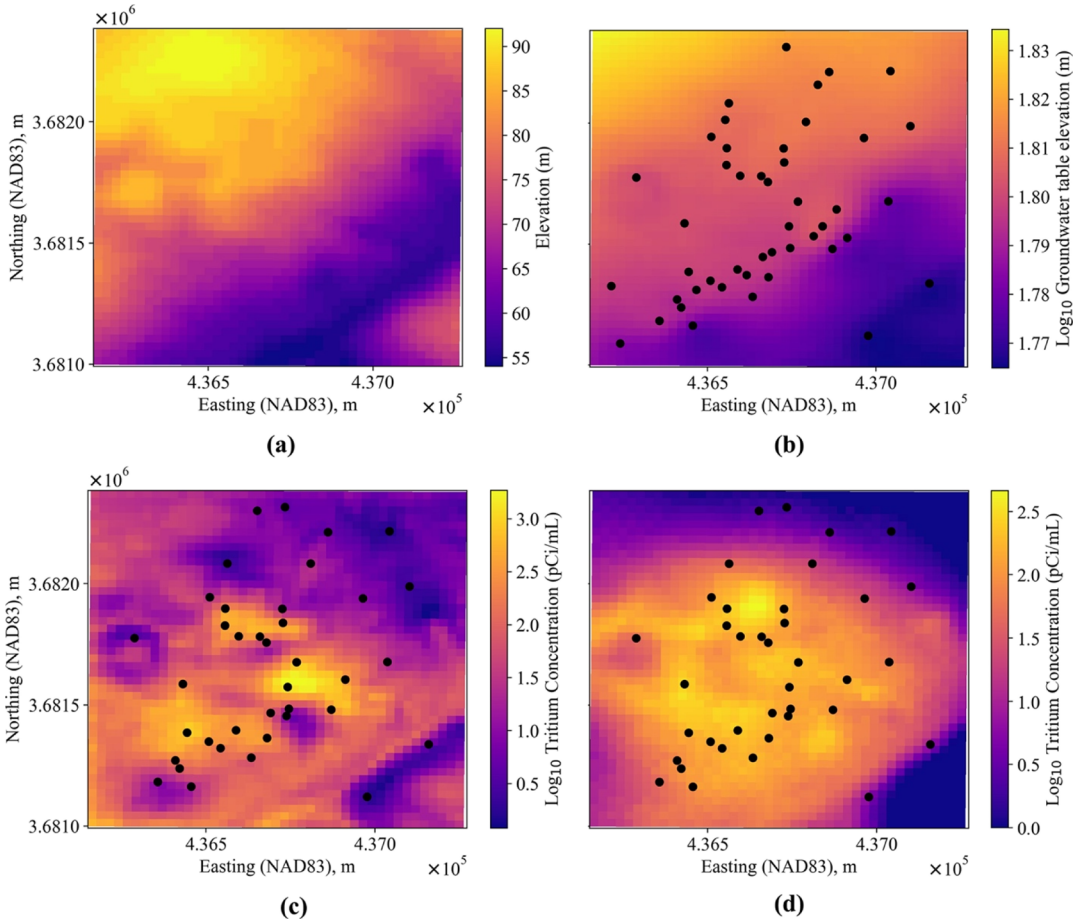


Figure 5. Supervised learning result: the spatiotemporal interpolation: (a) SRTM elevation heatmap across the F-Area, (b) water table elevation (the average 2015 values) estimated using the Lasso regression method, (c) tritium concentration map (the average 2015 values) using the Lasso regression method, and (d) tritium concentration map (the average 2015 values) using the Linear regression method.

Table 1. Top Results for the Spatial Estimation of Groundwater Table and Tritium Concentrations

	Fitting process results				LOOCV results			
	model	features	MSE	R^2	model	features	MSE	R^2
water table	RF + GP	easting, northing	2.30×10^{-7}	0.9983	RF + GP	easting, northing, elevation	1.80×10^{-5}	0.8663
	lasso + GP	easting, northing, elevation, slope	2.67×10^{-7}	0.9981	RF + GP	easting, northing, elevation, slope, flow accumulation	1.90×10^{-5}	0.8646
	ridge + GP	easting, northing, elevation, slope	2.83×10^{-7}	0.9980	RF + GP	easting, northing, elevation, slope	1.90×10^{-5}	0.8602
	GP		5.92×10^{-7}	0.9957	GP		2.40×10^{-5}	0.8272
tritium	lasso + GP	easting, northing, elevation, slope, flow accumulation	3.01×10^{-3}	0.9959	linear + GP	easting, northing, elevation, dist_to_basin	4.05×10^{-1}	0.4456
	lasso + GP	easting, northing, elevation, slope	3.01×10^{-3}	0.9959	ridge + GP	easting, northing, elevation, dist_to_basin	4.06×10^{-1}	0.4444
	ridge + GP	easting, northing, elevation, slope	4.77×10^{-3}	0.9935	linear + GP	easting, northing, elevation, slope, dist_to_basin	4.24×10^{-1}	0.4190
	GP		3.04×10^{-1}	0.5839	GP		4.65×10^{-1}	0.3628

with the contaminant concentrations temporally (Figure 3a), while the spatial correlation is not significant (Figure 3b).

In parallel, time series clustering based on the k-means clustering method (Figure 4) identifies the group of wells that have similar dynamics in the water table elevation and tritium concentration data. We identified the appropriate number of clusters as five using the elbow method (Figure S5). The water table is more variable spatially than temporally, with different wells having parallel lines (Figure 4a). There are five groups mapped to the actual locations, showing the correspondence to the topographic gradient (Figure 4b). The tritium concentrations have four clusters, mainly according to the concentration levels (Figure 4c). In the spatial map, the clusters are mapped as a function of the distance from the basin as well as the groundwater gradient, with one low-concentration group in the upgradient and periphery of the site and another high-concentration group near the basin (Figure 4d).

We then demonstrate supervised learning and spatial interpolation, using the water table elevation and tritium

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

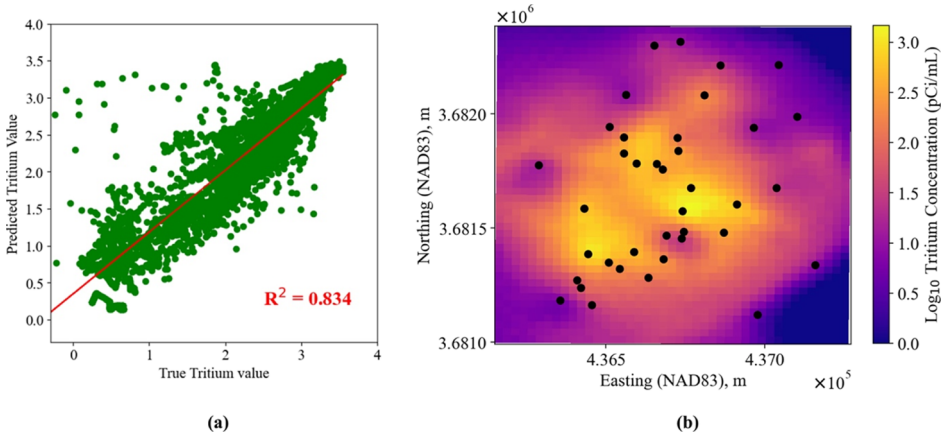


Figure 6. Supervised learning result: the proxy-based spatial interpolation of tritium contaminant concentration, (a) measured vs predicted tritium concentrations using the testing set, and (b) the estimated tritium concentration map with the well locations (the black circles). In (a), the red line represents the predicted values.

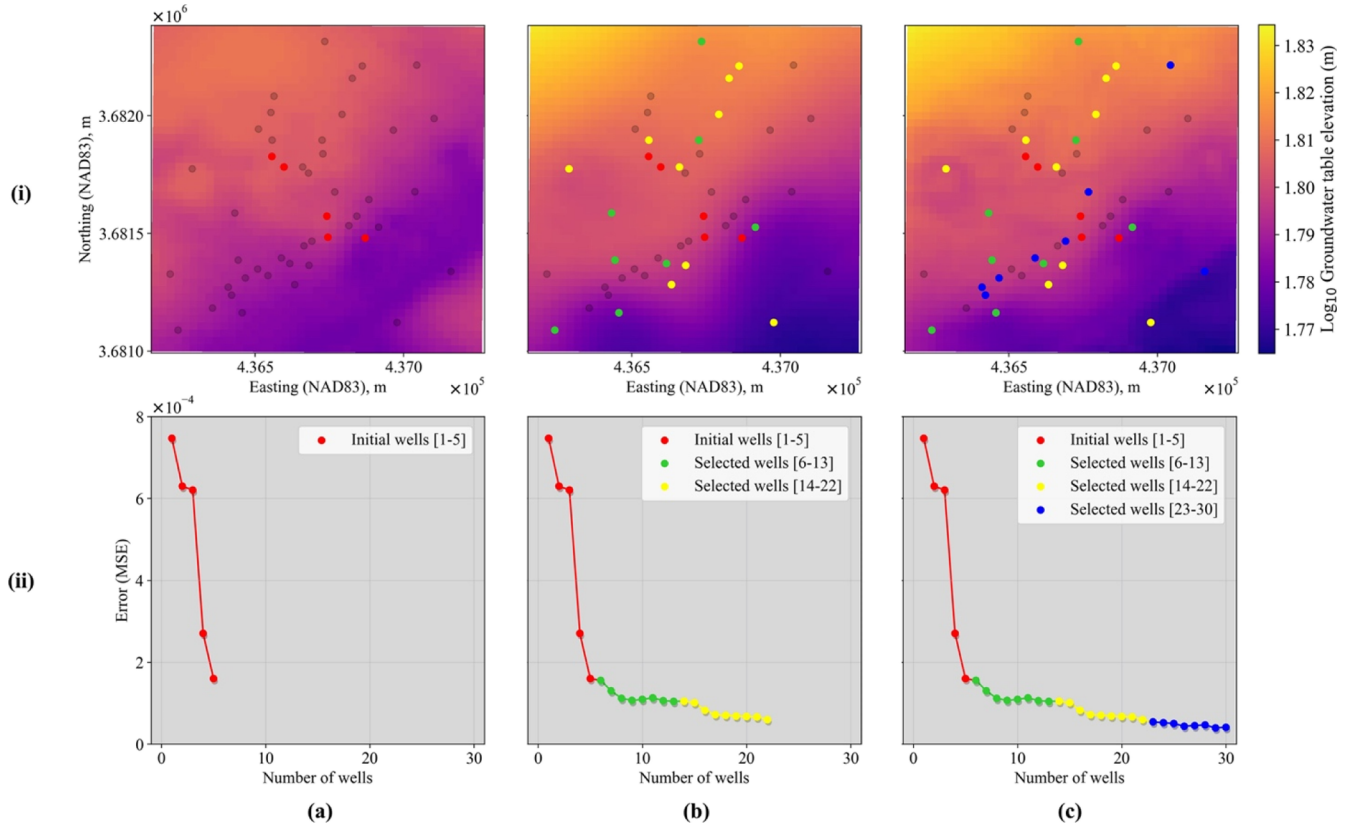


Figure 7. Reduced well configurations along with the estimated groundwater elevation (the average in 2015). The five starting wells are colored in red: "FSB 95DR", "FSB130D", "FSB 79", "FSB 97D", and "FSB126D". The colored circles, green, yellow, and blue, are the wells that are identified by the algorithm to best capture the water table spatial heterogeneity across the site. The bottom row shows the MSE as a function of the number of monitoring wells through the optimization for up to the first 5, 22, and 30 wells, respectively.

concentration averaged over 2015 (Figure 5). The estimated water table elevation shows the terrain following patterns (Figure 5b). Including the elevation (Figure 5a) and slope as predictors improves the performance of both fitting and LOOCV compared to the simple interpolation using the GPM (Tables 1 and S5). Among the multiple regression methods, RF shows the highest performance for the water table estimation (Figure S6), with an R^2 of 0.9983, although the Lasso regression (the second highest R^2) yielded the smoothest and most realistic map. For the tritium concentration, we included the distance to the source (i.e., the basin) as a predictor based on the clustering result (Figure 4c,d). Having the predictors is also effective for improving the predictive performance (Tables 1 and S6); the Lasso regression performed the best in fitting (Figure 5c) and the linear regression the best in LOOCV.

The proxy-based spatial estimation was performed to predict the tritium concentration map (the average within 2015) based on their spatiotemporal correlations to SC (Figure 3). First, the tritium concentrations at the wells over time were

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

predicted using the Ridge regression as a function of SC, excluding the 2015 data. Since the correlation was consistent in time and space (Figure 3), we could use the data from multiple wells and multiple time points. Different from the interpolation, a large number of data points (5852 individual samples) allowed us to reserve 20% as testing data. The regression performance showed an R^2 of 0.834 (Figure 6a) with the 20% testing data. This regression model was then used to predict the average tritium concentrations at the wells in 2015 with an R^2 of 0.799. Finally, we interpolated these tritium concentrations at the wells to create the plume map (Figure 6b). As can be seen, the proxy-based estimation slightly overestimates the center of the plume but accurately captures the plume boundary. The SC-based tritium estimation map produced an R^2 of 0.786 compared to the true tritium estimation (Figure 5d).

The monitoring well optimization was demonstrated using the annual averaged water table levels in 2015. The interpolated water table map using the Lasso regression method (Figure 5b) was used as the reference field. The five starting wells were selected according to the time series clustering results (Figure 4b). As the number of wells increases, the error decreases significantly for the first five wells (Figure 7a) that capture the multiple clusters associated with the water table gradient (Figure 4b). Then, there is a plateau between 6 and 13 wells (Figure 7b), which are located at the periphery of the site. The error continues to decrease slightly again from 14 to 22 wells, when wells are added mainly within the wetland zone. The error is further reduced from 23 to 30 wells (Figure 7c) when wells are added in between the wells already placed. The MSE converges between 20 and 30 wells, which appears to be sufficient to capture the spatial variability of the WT.

DISCUSSION

In this study, we have demonstrated an ML framework for supporting the long-term monitoring of groundwater contamination. Specifically, we developed an open-source python package to take advantage of the historical monitoring data sets typically accumulated during site characterization and remediation phases. Groundwater data are five-dimensional (5-D): well locations and screen intervals (3D), time, and multiple analytes. PyLEnM enables its users to explore this 5-D data set in many ways, such as using multiple time series of analyte concentrations at the same well over time or the same analyte concentrations across multiple wells at the same time. In particular, PyLEnM includes various preprocessing functions before ML, such as (1) QA/QC, (2) flexible coincident/colocated data identification to establish the data-driven relationship among different analytes and/or different wells, and (3) rapid data summarization and visualization to understand available data sets and to filter through the data sets. In addition, the key ML innovations in this package include (1) time series clustering to find the well groups that have similar groundwater dynamics and to inform spatial interpolation and well optimization, (2) the automated model selection and parameter tuning, comparing multiple regression models for spatial estimation/interpolation, (3) the proxy-based spatial interpolation method by including publicly available spatial data layers or in situ measurable variables as predictors for contaminant concentrations and groundwater levels, and (4) the new well optimization algorithm to identify the most effective subset of wells for maintaining the spatial interpolation ability for long-term monitoring.

Unsupervised learning enables us to identify key patterns in vast data sets such as the covariability among analytes in space and time. We extended the approach by Schmidt et al. (2018) that focused on the temporal correlations between in situ measurable variables and contaminant concentrations at each well. In this study, we found that the correlations between contaminant concentrations and in situ variables are consistent in time and space, although the correlation is linear with SC but nonlinear with pH. The correlation with SC results from the fact that total dissolved solids are dominated by nitrates, which are cocontaminants with tritium and uranium.[5] In addition, we found the time series correlations between the contaminant concentrations and groundwater table (depth to the water) such that the increasing groundwater table over time corresponds to lower concentrations. This is consistent with a modeling study,[1] showing that an increasing groundwater table typically leads to higher dilution.

We demonstrated the use of time series clustering, which has been increasingly used across various applications.[26] Rinderer et al. (2019) used hierarchical clustering to group wells with similar groundwater dynamics in order to map groundwater levels and their connectivity. Although the basic concept is the same, we have extended the approach to contaminant concentrations or any of the analytes in the data set. The results are useful for identifying similarly behaving wells, for identifying the dominant control on the spatial variability (such as the elevation for groundwater levels and the distance to the source for contaminant concentrations) and for selecting the initial set of wells for well optimization.

We have implemented comprehensive spatial interpolation algorithms for estimating groundwater table elevations and contaminant concentrations. Traditionally, simple interpolation (such as kriging or inverse-distance interpolation) has been used for such estimation.[6] PyLEnM allows us to find site-specific covariates or predictors such as elevation and topographic metrics, which provide additional constraints on estimation, significantly improving the estimation accuracy. Surface elevation has been known to be the main driver for groundwater elevation.[30,31] We have extended this approach by including different topographic metrics or the distance from the source for contaminant concentrations. Topographic information can be downloaded directly from a public database,[32] which makes our approach widely applicable to many surface aquifers.

In addition, we coupled standalone regression methods such as RF and linear regressions with the GPM. Although the GPM has been used before, the use of a grid search for covariance parameters adds an additional layer of automation that returns the most suitable covariance model for a given data set. In our case, we found that the Lasso and linear regression with the GPM yielded the best results when estimating both the water table and the tritium plume based on LOOCV. Among different ML methods, RF has become quite popular recently,[33,34] although Sekulić et al. (2020) found that ordinary kriging (OK, similar to the GPM) outperformed the other algorithms in terms of the mean absolute error (MAE). This is consistent with our results, in which the number of available data points (wells) was limited. Our automation of comparing multiple regression methods is powerful since the best models could be site-specific.[34]

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

Our results show that the estimation of contaminant concentrations is more challenging with a lower R^2 than the one of the water table. This is because the topography is a good predictor for the water table, which is aligned with the hydrology principle, while the distance to the source or the topography is not a sufficient predictor for the contaminant concentrations. This lack of predictive power is the reason why simpler regression methods (such as linear regressions rather than RF) were selected for the tritium concentration estimation (Table 1). Recent advances in physics-informed ML[35] could enable the integration of contaminant transport simulations (e.g., Xu et al. 2022)[36] to improve the contaminant concentration estimations in the future.

Furthermore, we demonstrated the proxy-based spatial estimation to predict contaminant concentrations based on in situ measurable parameters, by extending the temporal estimation proposed by Schmidt et al. (2018). We found that the spatiotemporal correlations between contaminant concentrations and SC are consistent in time and space at the SRS F-Area, which allows us to use historical data to predict the future concentrations. With in situ sensors and the internet of things (IoT) technologies that measure and transfer proxy variable data such as SC on a continuous basis, this would lead to spatially temporally continuous monitoring of contaminant concentrations, as well as detecting significant changes and anomalies.

PyLEnM includes a new monitoring well optimization algorithm to select the minimum and sufficient number of wells (among the existing ones) for capturing the spatiotemporal variability of the groundwater table and different analytes. Although there are other optimization methods available, they are primarily focused on representing the temporal behavior, using PCA.[9,10] Our approach, on the other hand, focuses on capturing spatial heterogeneity since the groundwater table and its gradient is important for plume mobility and direction. Compared to the algorithm in Sun et al. (2020), PyLEnM includes a more sophisticated algorithm for including multiple predictors, as well as for computing the overall estimation error at each added well, rather than adding a well at the highest error location. Although it might not be tractable to run the regression at each possible pixel, this approach is suitable for selecting a subset of existing wells, which is often the pressing need for long-term monitoring. If there is a need to select additional well locations, the original algorithm in Sun et al. (2020) is appropriate since it can select the pixel that is likely to have a large error locally rather than considering its effect on the overall interpolation error over all the pixels.

There are still limitations and challenges in PyLEnM that need to be resolved for broader applications. It assumes digitized and organized data sets in a defined format (i.e., the two tables). Data curation is an active area of research within ML and artificial intelligence such as digitizing data from existing papers or reports (e.g., Zavarin et al., 2022)[37] and managing an end-to-end data workflow from sensors/samples to data analysis.[38] These data curation and formatting technologies need to be integrated into PyLEnM. In addition, although PyLEnM offers the great flexibility to select different functions or their parameters, their appropriate choice is up to the users, and it may be site-specific. For example, a time-lag parameter to define the coincident data could be dependent on how fast groundwater conditions change at a particular site. To tackle these issues, we may expand the automated model and parameter selection performed for the spatial interpolation in this study (Table 1) to select parameters in other functions. At the same time, the correlations between contaminant concentrations and proxy variables may be site-specific or contaminant-specific. We plan to apply PyLEnM to other data sets and grow its user base to accumulate experiences on how to select appropriate models and parameters in different conditions.

We envision that this open-source framework should serve as a foundation that fosters ML development in the area of groundwater contamination research. Traditionally, the ML applications have been limited in groundwater contamination data due to the lack of quality data, with many gaps and anomalies embedded in the data. PyLEnM provides a variety of functions and tools to address this issue, cleaning up and formatting data sets so that they are ready for ML applications. In particular, the preprocessing, summarization, and visualization functions are powerful tools for not only understanding the working data set but also developing predictive ML. In addition, PyLEnM operates within the Google Colaboratory, connecting all the data sets seamlessly together through cloud computing. It also facilitates coupling of sparse groundwater data and publicly available spatial data layers (such as land cover types and remote sensing data) from python packages or the Google Earth Engine[39,40] in a seamless manner using an API, which would be particularly useful for regional-scale groundwater contamination[41] and naturally occurring contaminants.[42] At the same time, public trust and acceptance have been a difficult problem at contaminated sites. Through this open-source package and workflow from raw monitoring data to data processing and analysis, we envision that PyLEnM will play a critical role in improving the transparency of data analytics, as well as in empowering concerned citizens by enabling them to analyze data sets on their own.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c07440.

PyLEnM Functions, sample input data, summarization tables, QA/QC details, the interpolated correlation, PCA biplot of upper aquifer wells, links to the data and code repositories, links to the open-source tools used in PyLEnM, the elbow method for choosing the number of clusters, RF-based WT interpolation map, and the raw LOOCV results for WT and tritium estimation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Haruko M. Wainwright − *Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley 94704, United States; Department of Nuclear Science & Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*; Email: hmwainw@mit.edu

Authors

Aurelien O. Meray − *Applied Research Center, Florida International University, Miami, Florida 33174, United States;* orcid.org/0000-0001-8118-3176

Savannah Sturla − *Department of Environmental Science, Policy, and Management, University of California Berkeley, Berkeley, California 94709, United States*

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983

Masudur R. Siddiquee − *Applied Research Center, Florida International University, Miami, Florida 33174, United States*
Rebecca Serata − *Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, California 94709, United States*
Sebastian Uhlemann − *Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley 94704, United States*
Hansell Gonzalez-Raymat − *Savannah River National Laboratory, Aiken, South Carolina 29808, United States*
Miles Denham − *Panoramic Environmental Consulting, LLC, Aiken, South Carolina 29802, United States*
Himanshu Upadhyay − *Applied Research Center, Florida International University, Miami, Florida 33174, United States*
Leonel E. Lagos − *Applied Research Center, Florida International University, Miami, Florida 33174, United States*
Carol Eddy-Dilek − *Savannah River National Laboratory, Aiken, South Carolina 29808, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c07440

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was produced by Battelle Savannah River Alliance, LLC under Contract No. 89303321CEM000080 with the U.S. Department of Energy. Publisher acknowledges the U.S. Government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). This work was supported by the U.S. Department of Energy, Office of Environmental Management as a part of the Advanced Long-term Monitoring Systems (ALTEMIS) project, as well as by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research as part of the Lawrence Berkeley National Laboratory Science Focus Area, both under Award Number DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory. This research was also supported by the U.S. Department of Energy's Office of Environmental Management under Cooperative Agreement #DE-EM0005213 (PI Dr. Leonel Lagos).

■ REFERENCES

(1) Libera, A.; de Barros, F. P. J.; Faybishenko, B.; Eddy-Dilek, C.; Denham, M.; Lipnikov, K.; Moulton, D.; Maco, B.; Wainwright, H. Climate Change Impact on Residual Contaminants under Sustainable Remediation. *J. Contam. Hydrol.* **2019**, *226*, 103518.
(2) Superfund: National Priorities List (NPL). https://www.epa.gov/superfund/superfund-national-priorities-list-npl (accessed Feb 20, 2022).
(3) Ellis, D. E.; Hadley, P. W. *Sustainable Remediation White Paper Integrating Sustainable Principles, Practices, and Metrics into Remediation Projects*, 2009.
(4) Zachara, J. M.; Long, P. E.; Bargar, J.; Davis, J. A.; Fox, P.; Fredrickson, J. K.; Freshley, M. D.; Konopka, A. E.; Liu, C.; McKinley, J. P.; Rockhold, M. L.; Williams, K. H.; Yabusaki, S. B. Persistence of Uranium Groundwater Plumes: Contrasting Mechanisms at Two DOE Sites in the Groundwater-River Interaction Zone. *J. Contam. Hydrol.* **2013**, *147*, 45−72.
(5) Schmidt, F.; Wainwright, H. M.; Faybishenko, B.; Denham, M.; Eddy-Dilek, C. In Situ Monitoring of Groundwater Contamination Using the Kalman Filter. *Environ. Sci. Technol.* **2018**, *52*, 7418−7425.
(6) McLean, M. I.; Evers, L.; Bowman, A. W.; Bonte, M.; Jones, W. R. Statistical Modelling of Groundwater Contamination Monitoring Data: A Comparison of Spatial and Spatiotemporal Methods. *Sci. Total Environ.* **2019**, *652*, 1339−1346.
(7) Rinderer, M.; Meerveld, H. J.; McGlynn, B. L. From Points to Patterns: Using Groundwater Time Series Clustering to Investigate Subsurface Hydrological Connectivity and Runoff Source Area Dynamics. *Water Resour. Res.* **2019**, *55*, 5784−5806.
(8) Sun, D.; Wainwright, H. M.; Oroza, C. A.; Seki, A.; Mikami, S.; Takemiya, H.; Saito, K. Optimizing Long-Term Monitoring of Radiation Air-Dose Rates after the Fukushima Daiichi Nuclear Power Plant. *J. Environ. Radioact.* **2020**, *220*−*221*, 106281.
(9) Gangopadhyay, S.; das Gupta, A.; Nachabe, M. H. Evaluation of Ground Water Monitoring Network by Principal Component Analysis. *Groundwater* **2001**, *39*, 181−191.
(10) Khan, S.; Chen, H. F.; Rana, T. *Optimizing Ground Water Observation Networks in Irrigation Areas Using Principal Component Analysis*, 2008.
(11) 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.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(12) Bea, S. A.; Wainwright, H.; Spycher, N.; Faybishenko, B.; Hubbard, S. S.; Denham, M. E. Identifying Key Controls on the Behavior of an Acidic-U(VI) Plume in the Savannah River Site Using Reactive Transport Modeling. *J. Contam. Hydrol.* **2013**, *151*, 34−54.
(13) Wainwright, H. M.; Chen, J.; Sassen, D. S.; Hubbard, S. S. Bayesian Hierarchical Approach and Geophysical Data Sets for Estimation of Reactive Facies over Plume Scales. *Water Resour. Res.* **2014**, *50*, 4564−4584.
(14) *Renewal Application for a RCRA Part B Permit, Vol. I, Part 1, Section.1*; Savannah River Site: Aiken, SC, 2000.
(15) Meray, A.; Wainwright, H.; Upadhyay, H.; Siddiquee, M.; Sturla, S.; Patel, N. Pylenm. https://pypi.org/project/pylenm/ (accessed Jan 30, 2022).
(16) Berg, S.; Gommers, R.; Harris, C.; Hoyer, S.; Mendonça, M. W.; Pawson, I. NumPy. https://numpy.org/ (accessed Jan 31, 2022).
(17) Nelson, A.; Harris, C.; Baumgarten, C.; Carey, C. J. SciPy. https://scipy.org/ (accessed Jan 31, 2022).
(18) Abdalla, S.; Augspurger, T.; den Bossche, J. Pandas. https://pandas.pydata.org/ (accessed Jan 31, 2022).
(19) Cournapeau, D.; Blondel, M.; Brucher, M.; Buitinck, L. Scikit-Learn. https://scikit-learn.org/stable/ (accessed Jan 31, 2022).
(20) Snow, A. D.; Whitaker, J.; Cochran, M. Pyproj. https://pypi.org/project/pyproj/ (accessed Jan 31, 2022).
(21) Droettboom, M.; Caswell, T. Matplotlib: Visualization with python. https://matplotlib.org/ (accessed Jan 31, 2022).
(22) Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6*, 3021.
(23) Brochart, D.; Granger, B.; Grout, J. *Ipyleaflet: Interactive Maps in the Jupyter Notebook*; ipyleaflet, 2016.
(24) Devadoss, J.; Falco, N.; Dafflon, B.; Wu, Y.; Franklin, M.; Hermes, A.; Hinckley, E.-L. S.; Wainwright, H. Remote Sensing-Informed Zonation for Understanding Snow, Plant and Soil Moisture Dynamics within a Mountain Ecosystem. *Remote Sens.* **2020**, *12*, 2733.
(25) Wainwright, H. M.; Uhlemann, S.; Franklin, M.; Falco, N.; Bouskill, N. J.; Newcomer, M. E.; Dafflon, B.; Siirila-Woodburn, E. R.; Minsley, B. J.; Williams, K. H.; Hubbard, S. S. Watershed Zonation through Hillslope Clustering for Tractably Quantifying Above- and below-Ground Watershed Heterogeneity and Functions. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 429−444.
(26) Aghabozorgi, S.; Seyed Shirkhorshidi, A.; Ying Wah, T. Time-Series Clustering - A Decade Review. *Inf. Syst.* **2015**, *53*, 16−38.

(27) Trevor, H.; Robert, T.; Jerome, F. Overview of Supervised Learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2001; pp 9−39.
(28) Wong, T.-T. Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation. *Pattern Recognit.* **2015**, *48*, 2839−2846.
(29) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C. *Introduction to Algorithms*, 3rd ed.; The MIT Press, 2009.
(30) Naghibi, S. A.; Pourghasemi, H. R.; Dixon, B. GIS-Based Groundwater Potential Mapping Using Boosted Regression Tree, Classification and Regression Tree, and Random Forest Machine Learning Models in Iran. *Environ. Monit. Assess.* **2016**, *188*, 1−27.
(31) Prasad, P.; Loveson, V. J.; Kotha, M.; Yadav, R. Application of Machine Learning Techniques in Groundwater Potential Mapping along the West Coast of India. *GIsci. Remote Sens.* **2020**, *57*, 735−752.
(32) Amici, A. Elevation. https://pypi.org/project/elevation/.
(33) Mital, U.; Dwivedi, D.; Brown, J. B.; Faybishenko, B.; Painter, S. L.; Steefel, C. I. Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests. *Front. Water* **2020**, *2*, 20.
(34) Sekulić, A.; Kilibarda, M.; Heuvelink, G. B. M.; Nikolić, M.; Bajat, B. Random Forest Spatial Interpolation. *Remote Sens.* **2020**, *12*, 1687.
(35) Lavin, A.; Zenil, H.; Paige, B.; Krakauer, D.; Gottschlich, J.; Mattson, T.; Anandkumar, A.; Choudry, S.; Rocki, K.; Baydin, A. G.; Prunkl, C.; Paige, B.; Isayev, O.; Peterson, E.; McMahon, P. L.; Macke, J.; Cranmer, K.; Zhang, J.; Wainwright, H.; Hanuka, A.; Veloso, M.; Assefa, S.; Zheng, S.; Pfeffer, A. Simulation Intelligence: Towards a New Generation of Scientific Methods. **2021**, arXiv:2112.03235v1 (accessed Mar 30, 2022).
(36) Xu, Z.; Serata, R.; Wainwright, H.; Denham, M.; Molins, S.; Gonzalez-Raymat, H.; Lipnikov, K.; Moulton, J. D.; Eddy-Dilek, C. Reactive Transport Modeling for Supporting Climate Resilience at Groundwater Contamination Sites. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 755−773.
(37) Zavarin, M.; Chang, E.; Wainwright, H.; Parham, N.; Kaukuntla, R.; Zouabe, J.; Deinhart, A.; Genetti, V.; Shipman, S.; Bok, F.; Brendler, V. Community Data Mining Approach for Surface Complexation Database Development. *Environ. Sci. Technol.* **2022**, *56*, 2827−2838.
(38) Varadharajan, C.; Faybishenko, B.; Henderson, A.; Henderson, M.; Hendrix, V. C.; Hubbard, S. S.; Kakalia, Z.; Newman, A.; Potter, B.; Steltzer, H.; Versteeg, R.; Agarwal, D. A.; Williams, K. H.; Wilmer, C.; Wu, Y.; Brown, W.; Burrus, M.; Carroll, R. W. H.; Christianson, D. S.; Dafflon, B.; Dwivedi, D.; Enquist, B. J. Challenges in Building an End-to-End System for Acquisition, Management, and Integration of Diverse Data from Sensor Networks in Watersheds: Lessons from a Mountainous Community Observatory in East River, Colorado. *IEEE Access* **2019**, *7*, 182796−182813.
(39) Wainwright, H. M.; Steefel, C.; Trutner, S. D.; Henderson, A. N.; Nikolopoulos, E. I.; Wilmer, C. F.; Chadwick, K. D.; Falco, N.; Schaettle, K. B.; Brown, J. B.; Steltzer, H.; Williams, K. H.; Hubbard, S. S.; Enquist, B. J. Satellite-Derived Foresummer Drought Sensitivity of Plant Productivity in Rocky Mountain Headwater Catchments: Spatial Heterogeneity and Geological-Geomorphological Control. *Environ. Res. Lett.* **2020**, *15*, 084018.
(40) Google. Google Earth Engine. https://developers.google.com/earth-engine/apidocs (accessed Feb 20, 2022).
(41) Sajedi-Hosseini, F.; Malekian, A.; Choubin, B.; Rahmati, O.; Cipullo, S.; Coulon, F.; Pradhan, B. A Novel Machine Learning-Based Approach for the Risk Assessment of Nitrate Groundwater Contamination. *Sci. Total Environ.* **2018**, *644*, 954−962.
(42) Amini, M.; Abbaspour, K. C.; Berg, M.; Winkel, L.; Hug, S. J.; Hoehn, E.; Yang, H.; Johnson, C. A. Statistical Modeling of Global Geogenic Arsenic Contamination in Groundwater. *Environ. Sci. Technol.* **2008**, *42*, 3669−3675.

https://doi.org/10.1021/acs.est.1c07440
Environ. Sci. Technol. 2022, 56, 5973−5983