Groundwater Level Monitoring Networks Design with Machine Learning Methods

Sadaf Teimoori (sadaf.teimoori@wayne.edu)
Wayne State University https://orcid.org/0000-0001-5599-9817
Mohammad Hessam Olya
Wayne State University
Carol Miller
Wayne State University

Article

Keywords:

Posted Date: August 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1874647/v1

Groundwater Level Monitoring Networks Design with Machine Learning Methods

Sadaf Teimoori[1*], Mohammad Hessam Olya[2†], Carol J. Miller[1‡]

[1]*Department of Civil and Environmental Engineering—College of Engineering, Wayne State University, 5050 Anthony Wayne Drive, Detroit, MI 48202, USA*
[2]*Department of Industrial and Systems Engineering—College of Engineering, Wayne State University, 4815 Fourth Street, Detroit, MI 48202, USA*

Collecting and analyzing groundwater data are essential to evaluate the regional groundwater flow patterns and quality under existing and future temporal/spatial hydraulic stressors. These data also provide researchers with the information required to calibrate the groundwater models and study the interaction between surface water and groundwater resources. In general, frequent and well-organized groundwater monitoring improves the comprehension of groundwater systems and enhances groundwater resource management. Both surface water and groundwater quality monitoring networks have been extensively studied and presented in literature reviews. In contrast, there is much less literature focused on groundwater level monitoring networks. In many regions, groundwater level monitoring networks are limited and do not provide sufficient data for decision-making purposes. This gap in data availability is due to multiple reasons but includes financial constraints and the limited interest in those areas that rely more heavily on surface water resources for human consumption and industrial purposes. Here, we introduce methods employing K-means clustering and/or Relevance Vector Machine to design an optimal groundwater level monitoring network. The machine learning algorithms utilize the hydrogeological datasets obtained from the initial groundwater MODFLOW models and consider the uncertainties in the aquifer characterization through stochastic simulations. The result of this research is three groundwater level monitoring networks which the optimal one is selected based on the minimum modeling error. The network configurations are demonstrated in terms of the number and location of the observation wells. The proposed monitoring network improves the procedure of groundwater modeling and significantly reduces modeling errors.

[*]Corresponding author.
†h.olya@wayne.edu
‡ab1421@wayne.edu

There is no doubt that groundwater is one of the critical hydrogeological elements worldwide. It provides water for wells, wetlands, streams, and lakes, supports ecological and industrial systems, and functions as a medium for contaminant transport. Groundwater quantity and quality determine the value and reliability of groundwater for freshwater in various sectors. Degraded quantity and/or quality of groundwater can negatively affect environmental and human health. Consequently, monitoring groundwater resources is essential to recognize existing and future quantity and quality issues caused by natural and/or anthropogenic events and develop the best approaches for groundwater protection/management. Groundwater monitoring networks are clustered or distributed sampling points for characterizing groundwater quantity and quality. The measurements may be made in-situ (as is typical for quantity characterization, through hydraulic head measurements) or remotely (as is more typical for quality measurements, through transporting samples to an off-site analytical lab). Monitoring networks are most efficient when the sampling point locations and frequency are optimized. However, detailed attention to monitoring well designs is typically constrained due to the cost and difficulty of subsurface investigations. Limited field measurements exacerbate uncertainties in groundwater assessment and challenge the implementation of groundwater modeling, research, and management at various locations and times. A regular and well-organized field monitoring network provides initial information for the efficient development of groundwater models to better understand and predict groundwater conditions. Groundwater data is frequently gathered from the monitoring network and stored in the cloud for future use. The collected data has value as "raw" discrete data for analysis of temporal and spatial trends, as well as value as input to groundwater models for the development of a more continuous depiction of groundwater quality and quantity and prediction of response to different stressors as guided by researchers, engineers, and managers.

Evaluating groundwater quantity and the water table status plays a significant role in managing water resources. The design of underground infrastructure relies heavily on knowledge of groundwater geometric features – principally, groundwater table and thickness of the saturated unit. Groundwater, regardless of its depth, can also transfer a wide variety of contaminants to locations far from the pollution source. Some pollutants like volatile organic compounds can even migrate vertically, passing through the unsaturated zone and reaching the land surface [1]. Removing these environmental contaminations from groundwater and soil is usually expensive and challenging. Understanding the groundwater quantity and movement provides initial information

on transport pathways. In addition, it helps us recognize and assess the potential for groundwater to become contaminated due to urban development (human activities) at or near the land surface.

The design and assessment of Groundwater Level Monitoring Networks (GLMNs) have received less attention than other monitoring systems (such as surface water and air systems) in the environmental literature. The first GLMN was founded in England/Wales in 1845 [2]. In the United States, however, the construction of groundwater monitoring networks traces back to 1923 in New Jersey and 1925 in Pennsylvania, with more nationwide applications beginning in the late 1960's [2,3]. The earlier GLMN designs were based on hydrogeological approaches. After the 1980s, researchers started applying geostatistical techniques to design groundwater monitoring networks as the statistical method advanced in water resources applications. Since 2010, multi-objective/hybrid analyses have also been used for monitoring network design, which helped researchers accommodate multi-criteria groundwater monitoring networks [4,5]. There are two classes of GLMNs in terms of design objectives: (1) networks of characterization wells, which contain all new observation wells; and (2) networks of long-term monitoring wells, which are a subset selection of (many) existing observation wells (OBWs) for frequently observation practices [6]. The historical background and objectives of GLMNs in different countries are summarized by Singh and Katpatal [4]. They classified GLMNs based on their needs and purposes and compared their design approaches based on the design principles, sampling frequency, input parameters, etc. Their study outlined the present and possible future trends in groundwater monitoring network research, emphasizing the significance of the efficient design of GLMNs in many countries.

While the machine learning approach has extensively been employed in groundwater modeling [7-11], using machine learning techniques in groundwater monitoring network design is a quite new research area in recent decades. Asefa, et al. [6] presented a new ML technique known as Support Vector Machines (SVMs) based on Statistical Learning Theory (STL). They used their method to design a long-term GLMN and reduce redundant wells without changes in estimation error for the Water Resources Inventory Area (WRIA), Washington. This method is efficient and time-saving since the monitoring site location is selected based on the importance of the collected information, not searching on various configurations of the monitoring scheme. Ammar, et al. [12] described a method based on the sparse Bayesian Learning approach called Relevance Vector Machine (RVM), predicting uncertainties in the dataset and the model parameters. By applying

this technique to an existing monitoring network of the West Bank Palestine aquifers, the authors noticed that 68% of existing wells are redundant in the network. Khader and McKee [13] combined Monte Carlo simulations with the RVM technique to identify the uncertainties in hydraulic conductivities and recharge values and design the groundwater monitoring network in the Eocene Aquifer, Palestine. The results showed that the RVM method can reduce the number of wells and consequent monitoring costs while increasing the groundwater monitoring accuracy. Lal and Datta [14] developed an adaptive strategy for coastal groundwater resource management and designed an optimal monitoring network using the unsupervised learning method (K-mean clustering) in Bonriki aquifer at Kiribati island at the central Pacific Ocean. The objective function of their proposed network ensures monitoring of the groundwater quality in highly contaminated areas of the case study.

This research proposes a novel technique that combines groundwater flow modeling with stochastic models and machine learning techniques to design the optimal GLMN and apply it to the Metro Detroit area as the pilot example. The GLMN problem is solved through supervised and unsupervised learning approaches, including Relevance Vector Machine (RVM) and K-means clustering. The application of this method will lead to the identification of the optimum number of sampling sites and their locations within the study area.

Results

This section covers the results obtained from groundwater flow modeling, stochastic simulations, and GLMN design using clustering and regression approaches. The groundwater flow model and stochastic simulations provide inputs for machine learning algorithms to design GLMN for the study area. This section also presents the modeling results for existing OBWs and proposed GLMNs.

Groundwater Flow (Existing OBWs) Figure 1-a presents the created 3D model of groundwater with the grid and all applied data for the Metro Detroit area. As seen in Figure 1-a, the thickness of the model layer gradually decreases from the northwest to the southeast due to the increase of ground surface elevation toward the northwest. The density of groundwater wells shows groundwater extraction is more significant in the northwest of the region due to the higher depth of groundwater. The storm-sewer network, as one of the urban area elements, is mostly expanded in the city of Detroit in the southeast of the region. The initial groundwater model is also provided in Figure 1-b, showing a general groundwater flow direction of northwest to east, southeast, and

south of Metro Detroit, discharging to the Detroit River and Lake St. Clair. It is worth noting that groundwater may move in other directions at the local scale, while the present analysis focuses on the regional scale. The groundwater is found in lower elevations/depths in the southeast, confirming the site investigations and literature review. Groundwater also has greater depth in the northwest and is found at higher elevations compared to the southeast.

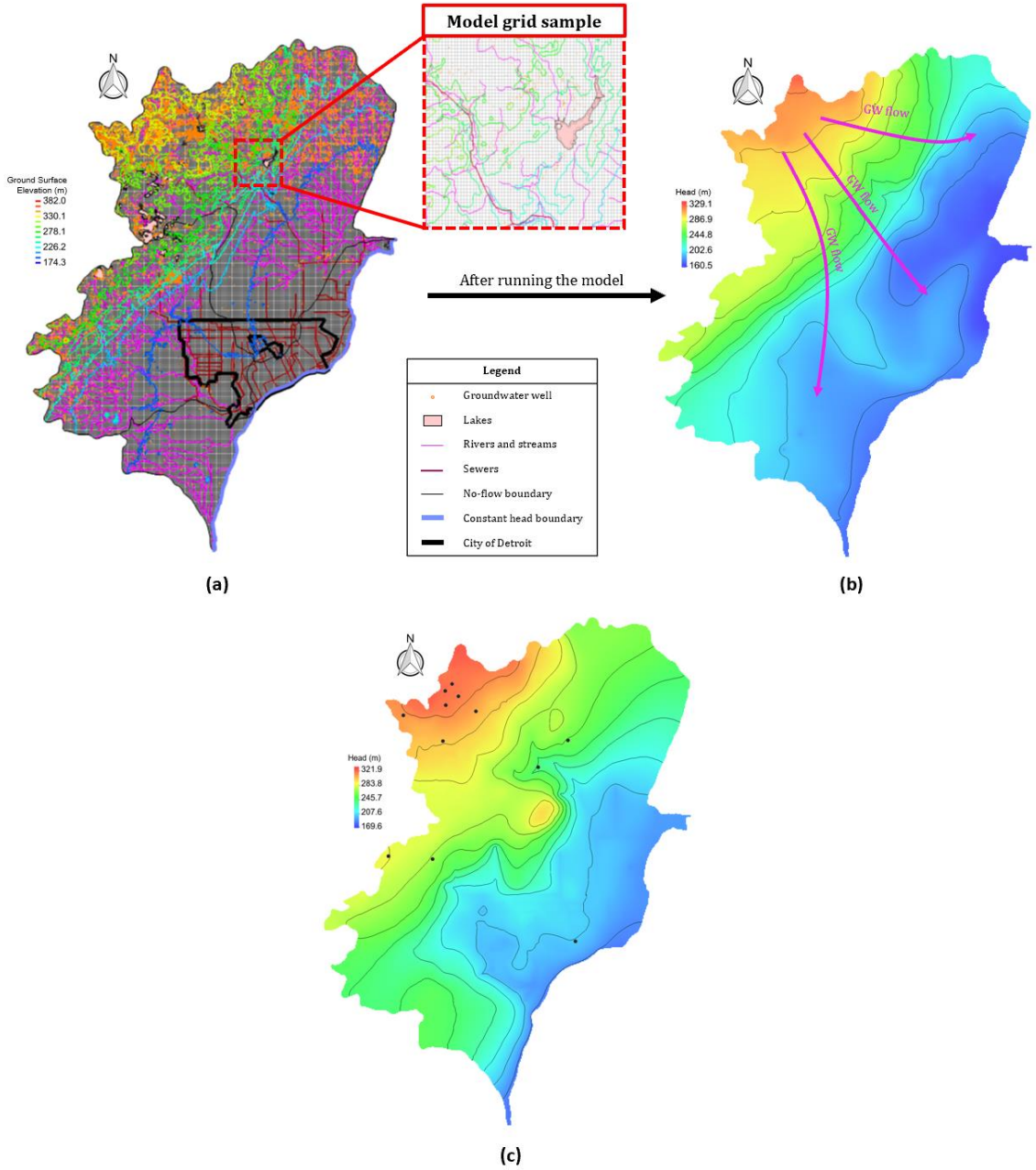
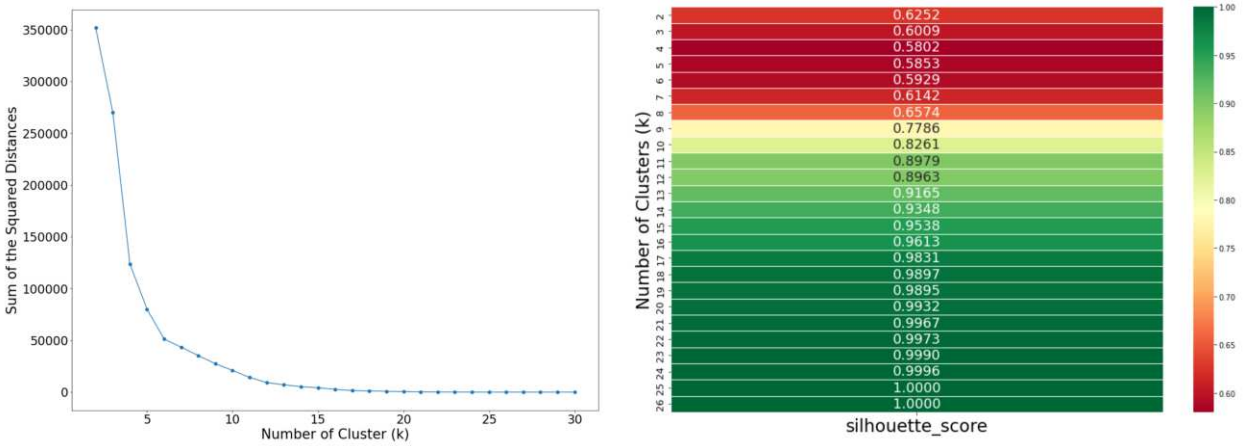


Figure 1: **Groundwater MODFLOW Model (a) Inputs, (b) Initial Result, and (c) Calibrated Results with Existing OBWs, Observed Data for November 14-16, 2017**

Although insufficient groundwater observation data caused some difficulties in the model calibration of the study area, we calibrated the model to the best of our efforts through parameter estimation (PEST) for hydraulic conductivity and recharge rate. For calibration, we used the observed groundwater head data collected on November 14-16, 2017, for 12 existing OBWs within the study area. Figure 1-c presents a calibrated model of groundwater flow, including the existing OBWs for the study area. It shows the general flow direction is similar to the initial results; however, groundwater behaves differently in the middle and south regions of the model and flows in other directions. In addition, the calibrated model shows that the maximum and minimum head levels are ~169 (m) to ~322 m, respectively. However, in the initial model, the groundwater head ranges from ~160 (m) to ~329 m. The PEST calibration running time is 9 hours and 14 minutes.

K-means Clustering Approach Since groundwater monitoring wells cannot be installed in every model cell, clustering the hydrogeological features of the study area enables us to reduce or even optimize the number of wells in a GLMN. We used the K-means algorithm to cluster all model cells based on the groundwater-related characteristics and standard deviation of groundwater head levels from stochastic simulations.

We used the Elbow and Silhouette methods to define the optimal number of clusters for the K-means algorithm. Figure 2 presents the results of the abovementioned methods in detecting the optimal number of clusters. Figure 2-a shows that the Elbow method results in a smooth graph in which the optimal number of clusters is ambiguous [15] due to the spatial proximity between input data. Since the elbow point in the Elbow method is not clearly visible, the Silhouette method is employed to determine the best optimal number of clusters in this research. Figure 2-b presents the results of the Silhouette method, illustrating the Silhouette scores for each cluster number. Considering the groundwater characteristics of Metro Detroit, we predefined 25 clusters in the K-means algorithm as the optimal number of clusters since its Silhouette score is the first highest score.



(a) (b)

Figure 2: **Methods for Finding Optimal Number of Clusters (a) Elbow Method, and (b) Silhouette Method**

After defining 25 as the cluster number in the algorithm, the Metro Detroit hydrogeological datasets are clustered based on their similarity through the K-means algorithm. Figure 3-a shows these clusters with different colors and presents the location of their centroids (red dots) within the study area in the Python 3.10 language Jupiter Notebook environment. The location of cluster centroids is assumed as the location of groundwater monitoring wells. As seen in Figure 3-a, the K-means clustering generates a uniform GLMN compared to the existing OBWs. Therefore, the K-means clustering GLMN covers the entire study area to ensure the availability of observation data used for modeling and/or management purposes.

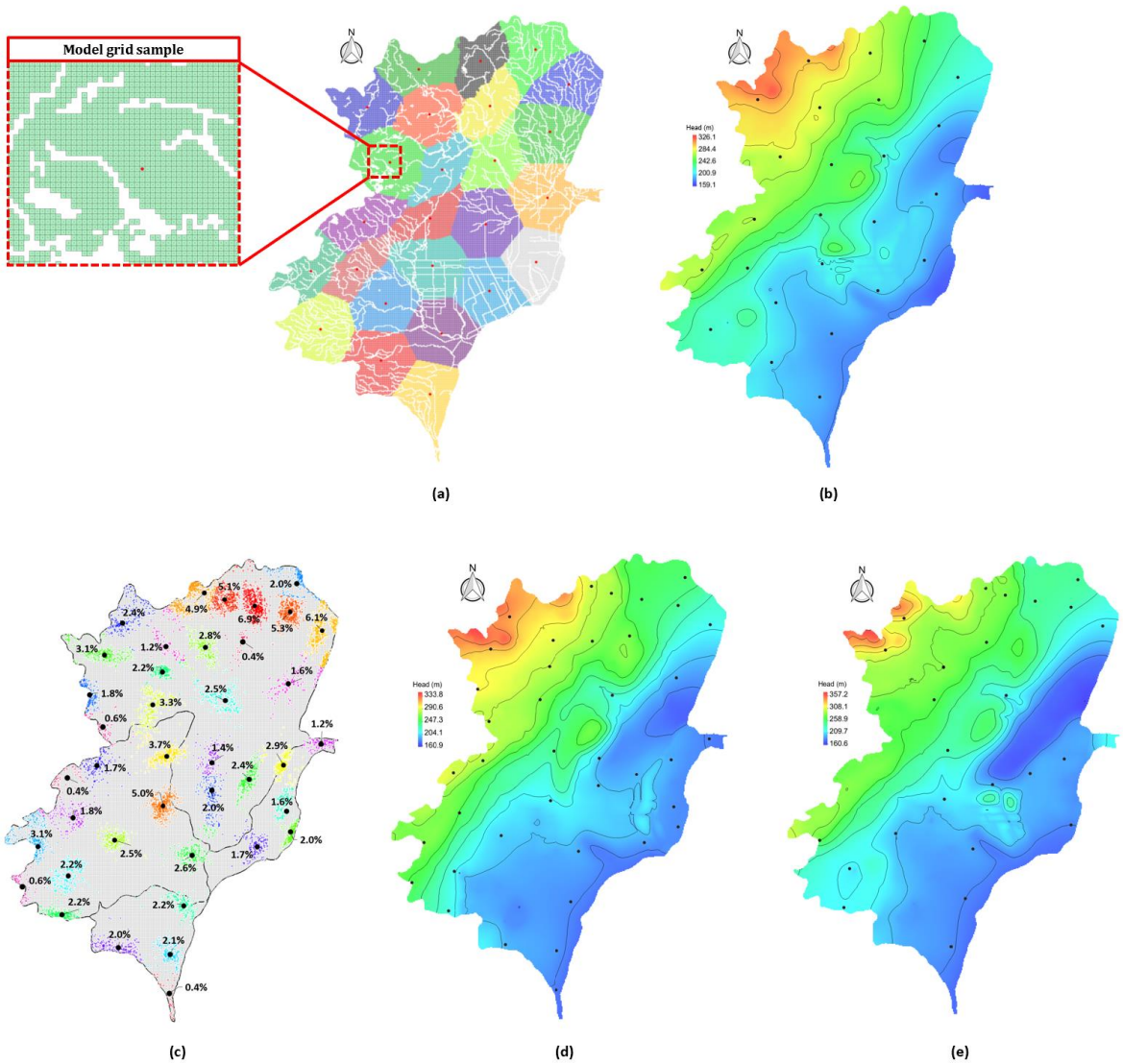


Figure 3: **(a) K-Means Clusters and Their Centroids Assumed as the Location of Groundwater Monitoring Wells, (b) Groundwater MODFLOW Model with GLMN Design by K-means Clustering Approach, (c) Distribution of RVs and Probability of Occurrence Presented at Cluster Centroids, (d) Groundwater MODFLOW Model with GLMN Design by RVM Approach, and (e) Groundwater MODFLOW Model with Modified GLMN (RVM Approach)**

As described in the method section of this paper, the cluster centroids found by the K-means algorithm are applied to the initial groundwater model of the study area as the observation points. To evaluate the performance of the K-means GLMN in groundwater modeling, we interpolated groundwater levels of existing observed values on November 14-16, 2017, for these cluster centroids. Figure 3-b presents the calibrated groundwater model for the GLMN designed

through the K-means algorithm. As seen in Figure 3-b, groundwater head level changes from ~159 (m) to ~326 (m), which, on average, is 11.9 (m) less than the calibrated groundwater levels in Figure 1-c. In addition, the general groundwater flow is similar to the calibrated model; however, in the middle and south regions, groundwater level changes are more minor. The initial model error of the model is 19121, which has been reduced to 32.4 during the PEST calibration run for 4 hours and 9 minutes.

RVM Approach This subsection presents the GLMN designed through the RVM algorithm and its results. Figure 3-c shows the final populated Relevance Vectors (RVs) in each run of the RVM algorithm for training datasets. Although training datasets are randomly selected from stochastic models, the RVM algorithm identified RVs at frequent locations at each training run. The populated RVs are categorized into 40 clusters, and the center of each cluster is assumed as the best location to install the observation well of a GLMN. As seen in Figure 3-c, the cluster centroids are irregularly distributed within the study area, with the highest probability of RVs in the north.

Similar to the K-means approach, the cluster centroids obtained from the RVM approach are assumed as the observation wells of GLMN and applied to the groundwater model. The MODFLOW model is run with these new proposed observed wells. The modeling result for GLMN designed through the RVM approach is shown in Figure 3-d. The groundwater model shows head levels with a range of ~160 (m) to ~334 m. The general groundwater flow remains the same, with minor differences at some locations. The initial model error of the model is 26699 due to the increased number of observation wells, which has been reduced to 1082 during the PEST calibration run for 6 hours and 58 minutes.

Modified RVM Since the final model error is still high in the RVM approach, we modified the designed GLMN and selected the first 25 observation wells with the highest probability of Relevance Vectors (RVs) occurrence. Figure 3-e shows the calibrated groundwater model with 25 observation wells of GLMN designed by RVM approach. The groundwater model shows head levels with a range of ~160 (m) at northwest to ~357 (m) at southwest. The initial model error is 22047 due to the increased number of observation wells, which has been reduced to 1.26 during the PEST calibration run for 4 hours and 36 minutes.

Discussion

This section presents the discussion of obtained results. The existing OBWs have irregularly been placed within the Metro Detroit area and collect groundwater data at infrequent

intervals. K-means and RVM algorithms recognize 25 and 40 centroids, respectively, as the optimized number of observation wells in a GLMN. These proposed locations for observation wells fill up the gap in groundwater data and help to improve the understanding of hydrogeological events as they cover the areas where groundwater data do not exist. In the RVM network, 40 observation well would obviously increase the total cost of GLMN design, data collection efforts, and future maintenance costs. Table 1 shows the modeling/calibration performance of existing OBWs and proposed GLMNs. As seen in Table 1, the groundwater model with the K-means network has fewer model errors than existing OBWs and RVM network. In addition to the cost issues explained above, the RVM network has the highest model errors among all models. Therefore, the RVM approach may not be considered the best approach to design GLMNs. To compare the RVM network with the K-means network, we reduced the number of wells in the RVM network to 25 wells based on the RV frequencies. The model with the modified RVM network shows the lowest errors among all other configurations, suggesting the best option to design a GLMN.

Table 1: **Performance of Groundwater Modeling for Different Configurations of GLMN**

	Existing OBWs	K-means Network	RVM Network	RVM Network (modified)
Mean Error	-0.0484	0.0402	-0.0605	0.0069
Mean Absolute Error	0.3347	0.1406	0.8553	0.0083
Root Mean Squared Error (RMSE)	0.5199	0.2309	1.1951	0.0147

Figure 4 is presented to show the model error at five iterations of the model run with different configurations of GLMN. The model with existing OBWs starts with the model error of 13912, abnormally increases to 130635, and gradually decreases to 115.1 in five iterations. The model with K-means networks also has a model error of 19121 due to more observation wells in the network compared to the existing OBWs, and ends up with a model error of 32.4 in five iterations. The model with the RVM network is not as successful as existing OBWs and/or K-means approach in decreasing model error during PEST calibration. However, the model with the modified RVM network shows a significant decrease in the model error from 22047 to 1.3. Therefore, the modified RVM network leads to normal PEST runs, significantly decreasing model errors.

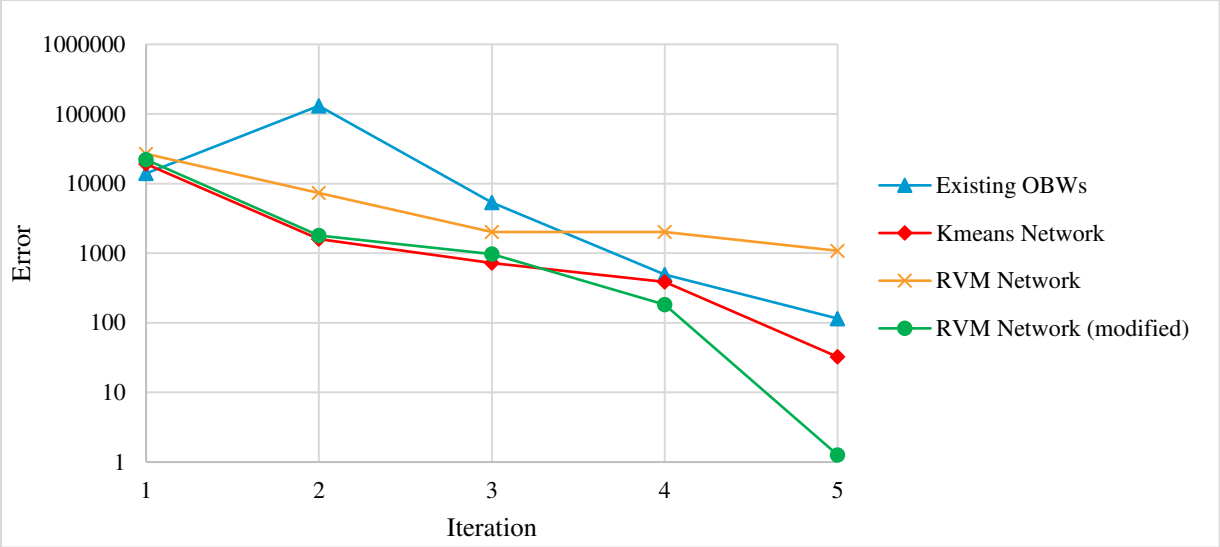


Figure 4: **Model Error vs. Iteration for Different Configurations of GLMN**

Among all models run in this research, the model with the K-means network has the least total running time (4 hours and 9 minutes), followed by the model with the modified RVM network completing the model run in 4 hours and 36 minutes. Although the RVM approach suggests more observation wells in the monitoring network than the existing OBWs, the model runs in 6 hours and 58 minutes. Comparing the model with existing OBWs, K-means, modified RVM, and RVM networks decrease the total model run by %55, %50, and %25, respectively. Therefore, all proposed GLMNs in this research lead to time-saving modeling and accelerate the total model/calibration run time.

Furthermore, as reported in the previous section, all models show similar results of general groundwater flow direction with local-scale differences at some locations, confirming the literature investigations on groundwater within the Metro Detroit area. The groundwater generally flows from northeast to southeast and finally discharges to the Detroit River. The groundwater level ranges in each model are presented in Table 2. The model with existing OBWs shows the least groundwater level changes, while the modified RVM network has the most groundwater level changes, offering a wider range of hydraulic gradients to groundwater flow. While the top elevations are constant in all models, the 95th percentile of head levels in models with proposed GLMNs remains below the 95th percentile of top elevation values.

Table 2: **Groundwater Minimum, Maximum, and 95th Percentile of Head Levels in Existing OBWs and Proposed GLMNs**

Value \ Network	Existing OBWs	K-means Network	RVM Network	RVM Network (modified)
Minimum Head (m)	169.1	159.1	160.9	160.6

Maximum Head (m)	321.9	326.1	333.8	357.2
Max. Head - Min. Head	152.8	167	172.9	196.6
95th Percentile of Head Levels	316.8	312.9	312.8	313.5

Conclusions

This study proposed methods to find the location of the optimized number of observation wells and put them together as GLMNs. The groundwater data obtained from the networks developed in this study can provide sufficient information for groundwater modeling and further resource management. The proposed methods utilize the initial data compiled from initial groundwater modeling efforts, consider stochastic simulations to estimate the uncertainties in hydraulic conductivity and recharge rate datasets, and apply supervised and unsupervised learning algorithms to all available hydrogeological data. Metro Detroit area is selected as the research pilot study area due to the insufficient hydrogeological studies and lack of observation wells. This research proposes GLMNs designed through two machine learning methods and results in three configurations of K-means, RVM, and modified RVM. The performance of all three networks is demonstrated using the modeling/calibration process and compared based on the model RMSE for each monitoring configuration. Since the proposed methods need an initial sense of groundwater behavior, the research is limited to the study areas with minimum head-level data observed. Therefore, this work cannot be applied to regions with no head-level investigation/observation history. Besides all other groundwater modeling assumptions explained in the method section, the methods have only been applied/verified in a steady state due to the lack of transient data within the study area. Designing GLMNs considering transient state requires considerable computation time and memory due to the complexity of stochastic simulations but is recommended for future works.

Finally, the significant contribution of this research is the ability of the proposed methods to improve the groundwater models and calibration process. Throughout this work, all groundwater models show that groundwater generally flows in a northwest-southeast direction within the study area, confirming the groundwater flow direction of the region investigated by the literature. In terms of observation well number, K-means and modified RVM networks have 25 observation wells, which is 15 observation wells less than the RVM configuration. Therefore, given identical installation/maintenance costs for each observation well, K-means and modified RVM approaches suggest more economical options than the RVM approach. All in all, K-means and the modified RVM approaches lead to the time-saving modeling and decrease

PEST/calibration time by %55 and %50, respectively. Modified RVM configuration also results in the least amount of modeling RMSE and is highly recommended for the optimal design of GLMNs.

Methods

In this section, the framework of the proposed methodology is clarified. Figure 5 is the flowchart of the proposed method, illustrating the three phases in designing a GLMN. Phase I starts with obtaining a group of physical, hydrogeological, and boundary condition datasets to create the groundwater model and characterize the groundwater flow system for the study area. In Phase II, a series of stochastic models is applied to measure the groundwater model uncertainties. In Phase III, the number and location of observation wells are determined using two algorithms of the machine learning method, i.e., K-means clustering and RVM. For each proposed GLMN, a MODFLOW model is generated with equal attributes except for different configurations of observation wells. In order to run the MODFLOW models, it is essential to assign observed values to any OBWs in the region. Therefore, an interpolated ArcGIS raster of head levels in the existing OBWs is generated to assume/estimate the observed values for the new OBWs in the proposed GLMNs. Finally, the proposed networks are compared in terms of the model Mean Squared Error (MSE) to develop the optimal GLMN network. The MSE is calculated based on the observed values (estimated from the interpolation raster) and the calculated values.

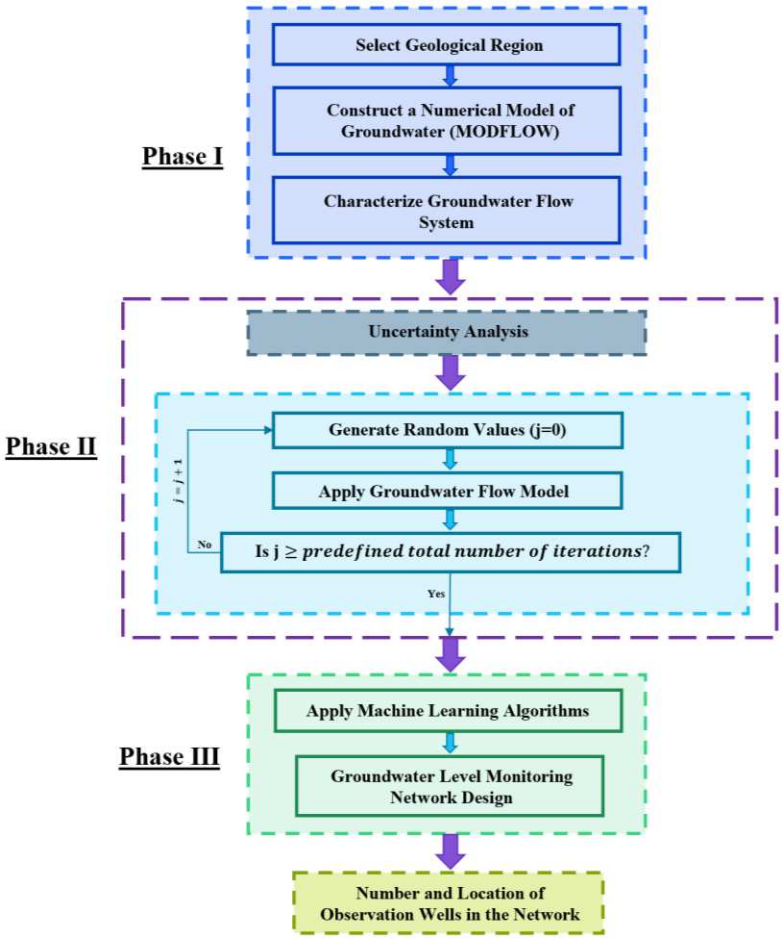


Figure 5: **The General Flowchart of the Proposed Method**

Phase I: Groundwater Flow Model – MODFLOW In water resources studies, a conceptual groundwater model is always necessary to reveal the data gap and provide a general understanding of existing problems in the site, groundwater flow direction, discharge, and depth to groundwater [16,17]. In phase I, a groundwater model is developed using USGS MODFLOW-2005 [18] from a licensed Groundwater Modeling System (GMS) package. USGS MODFOLW-2005 is a newer version of the MODFLOW family, which is able to manage the internal data. USGS MODFLOW is a worldwide standard three-dimensional finite-difference groundwater modeling tool. It is used for simulating flow in the saturated zone by solving a flow equation combining Darcy's law and the principle of conservation of mass. It also can simulate steady and transient flow in various conditions of layers, including confined, unconfined, or a combination of both. In USGS MODFLOW, the region is divided into cells, defining a grid for solving groundwater equations. It accommodates different boundary conditions, including wells, drains, areal recharge, evapotranspiration, rivers, and lakes. There are many successful applications of USGS

MODFLOW numerical models to simulate the subsurface processes in previous studies, including Wang, et al. [19], Panagopoulos [20], Bailey, et al. [21], Hughes, et al. [22]. Therefore, USGS MODFLOW serves as an appropriate modeling tool for assessing groundwater flow for this project.

The initial step of groundwater modeling is collecting and processing data. The groundwater model calibration is the process in which the initial model input parameters are re-estimated to improve the match between predicted and observed values of the dependent variable within the predefined criteria. This process requires sufficient field measurements of the dependent variable within the study area.

Phase II: Uncertainty-Stochastic Analysis In this work, the stochastic simulations generate variability for hydrogeological inputs. The stochastic simulations are implemented in MODFLOW-2005, with the parameter zonation method using Latin Hypercube Sampling (LHS). The LHS method efficiently seeks the probability proportion of each parameter to generate a sample of size N from the X variables. The total probability is defined by standard deviation, mean, and lower and upper bounds. The LHS achieves the same model confidence level reflecting the real underlying distribution through fewer runs and samples than the simple random sampling and Monte Carlo methods [23-26]. Like all sampling methods, LHS generates the probability distribution and different scenarios for groundwater flow movement. Furthermore, the models created from LHS data are used in the machine learning algorithm as training datasets.

This research deals with model uncertainties by implementing 512 stochastic simulations based on equal plausible candidate realizations. These uncertainties are basically caused by data deficiency of groundwater variables, including hydraulic conductivity and recharge rate. First, the range of each variable is divided into 512 non-overlapping intervals based on equal probability size. Next, at each simulation, one random value is selected for each variable with respect to the probability distribution. Finally, the 512 values thus obtained for each variable are randomly paired with the 512 values of other variables.

Phase III: Machine Learning Algorithms Since we will design a first-ever GLMN (characteristic network), the ideal (and hypothetical) scenario is that any possible cell on the groundwater model can be suggested as a potential location for a groundwater OBW. Although this ideal scenario can cover the whole study area for observation and provide detailed information on groundwater head levels from numerous locations, it is neither a feasible nor efficient approach

in the regional study of groundwater resources due to the limited budgetary allocations. Even if there are no budgetary limits for monitoring projects, OBWs should be installed at the scattered points determined by the hydrogeological properties of the region to avoid redundancy in collected data. A key feature of optimal GLMN design is to locate a permissible number of OBWs at locations suitable for collecting useful, non-redundant, and reliable groundwater head data [14,27]. Therefore, we carefully consider the location of all potential OBWs to cover the entire study area through the machine learning algorithms (i.e., clustering and regression). This research utilizes K-means clustering and RVM methods and investigates the results for each approach. Furthermore, a subset of the potential OBWs acquired from the RVM approach is selected as the optimal OBWs and compared with other proposed GLMNs. The machine learning algorithm input consists of the cell features calculated in groundwater models, including longitude and latitude of cells, top and bottom elevations, average groundwater level and standard deviation obtained from stochastic simulations, hydraulic conductivity, recharge rate, and evapotranspiration rate. The cells representing rivers and sewer systems or flagged as dried cells once or more in stochastic simulations are also excluded from the inputs to prevent the algorithms from considering those cells in the calculations and proposing them as OBWs.

Unsupervised Learning: K-means Clustering Among the available clustering algorithms, the K-means algorithm offers an efficient and popular clustering method [28]. A detailed explanation of the K-means clustering methodology is presented by Bandyopadhyay and Maulik [29]. In this project, we use the K-means algorithm to cluster all model cells based on their hydrogeological features and find the potential OBW locations from the entire study area in the model. In the K-means clustering, we aim to group model cells while maximizing cell similarity in each group and minimizing cell similarity in different clusters. The main idea of using K-means clustering is to categorize the set of cell centers into k disjoint clusters.

The number of clusters (k) in the K-means algorithm must be determined in advance. There are many methods to find the best optimal number of clusters, including the Elbow and Silhouette methods [15,30-32]. The Elbow method is the oldest visual method to determine the number of clusters which calculates the sum of squared distances between data points. As the cluster number increases, the sum of squared distances dramatically drops to reach a point, i.e., elbow, after which

the sum of squared distances reaches a plateau. The optimal number of clusters is where the elbow is visually distinguishable.

The Silhouette mothed compares within-cluster distances and between-cluster distances and scores their differences for each number of clusters. The Silhouette score ranges from -1 to 1, with larger (more positive) values indicating the more optimal the clustering number is. The Silhouette score of 1 means the clusters are dense and well-separated. The K-means algorithm is run with the cluster number (k) determined through either of these methods. The final solution offers one centroid for each cluster after the algorithm converges. The closest model cell center to each K-means centroid is proposed as a potential location for OBW installation in a GLMN.

Supervised Learning: Relevance Vector Machine (RVM) The Relevance Vector Machine (RVM) method introduced by Tipping [33] is a statistical Basiyan framework to model sparse solutions for regression and classification problems. Given a training dataset in form of input-target pairs, the algorithm creates the conditional Gaussian distribution of the target values based on the input vectors. A detailed description of the RVM algorithm can be found in Tipping [33] and Fletcher [34].

In the present research, the RVM algorithm obtains input values, including latitude, longitude, and the average groundwater heads in each cell. The target values are the standard deviation of groundwater head levels for all stochastic simulations. The results provide the location and number of RVs for each application of the algorithm. RVs represent a subset of the datasets whose properties are very important to the correct implementation of the algorithm. The overall result of RVs from 75 runs of the algorithm would show consistency in selecting the locations and number of RVs. All RVs are compiled together and grouped based on their distances from each other. Next, the density of each group is calculated and expressed in proportion to the overall density of RVs. This proportion shows the occurrence probability assigned to the center of each RV group. The center of each group is assumed as a potential location for OBW installations in a GLMN. The occurrence probability at each location enables us to prioritize the OBW installation based on the financial limitations of every project.

This proposed method is applied to the Metro Detroit region as a pilot example of the approach utility. The Metro Detroit area is located between the latitudes 42.171417° to 42.810517° N, and longitudes 82.791759° to 83.540047° W, within Southeast Michigan. The study location

area is 3850.32 km2, encompassing four major watersheds, including Clinton, Lake St. Clair, Rouge, and Ecorse Creek. Figure 6 provides the study area geography. No previous study designed a groundwater monitoring network within the Metro Detroit area to the best of our knowledge. Due to the limited observation wells that spatially and/or temporally collect data [35], groundwater modeling and calibration efforts are highly difficult and time-consuming within the study area. The model boundary conditions include general head boundary and no-flow boundary. The model grid is set up in on layer with 109,909 active cells with a size of ~187 (m) by ~187 (m). The minimum ground surface elevation is ~174 (m) (NAVD88) in the southeast, while the maximum elevation is up to ~382 (m) (NAVD88) at some locations in the northwest. For more clarification, the 95th percentile of top elevation values is 315.5 (m). The bottom elevation ranges from ~144 (m) (NAVD88) to ~282 (m) (NAVD88). All simulations are conducted in steady-state since the transient datasets are unavailable for the defining data sets.

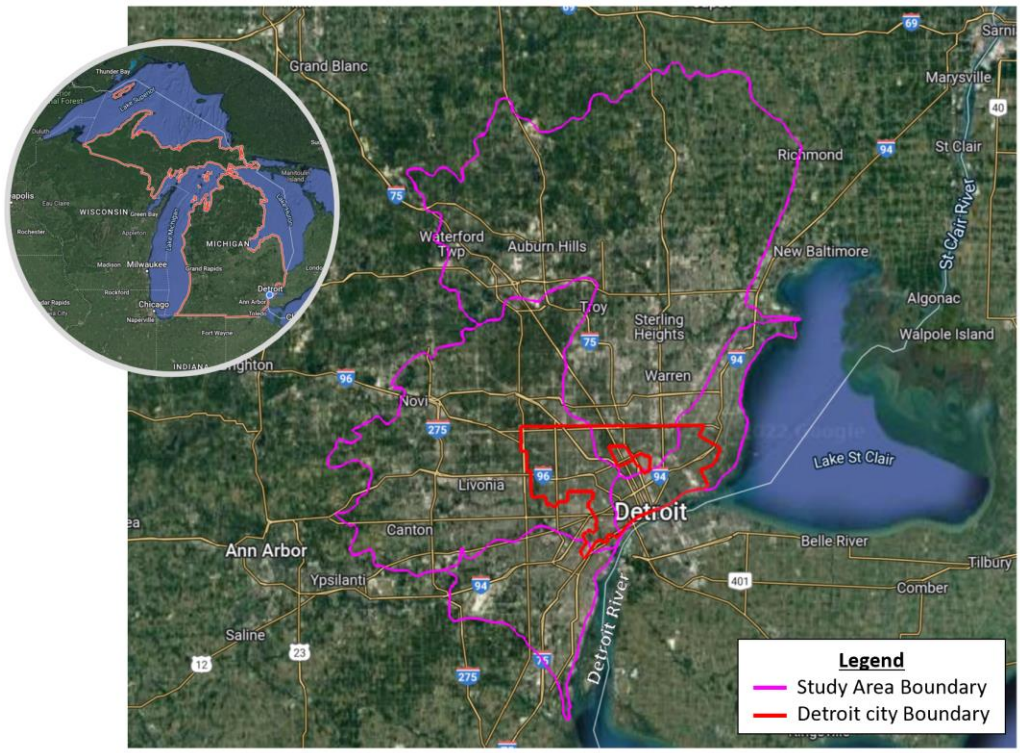


Figure 6: **Geographic Location of Metro Detroit Area (Obtained from Google Maps)**

Data Availability

For Metro Detroit, the input datasets are obtained from multiple sources. The hydraulic conductivity values are obtained from the borehole logs dataset [36,37]. The average annual values of

evapotranspiration and precipitation are collected through the USGS [38] data sources. The head stages of the main rivers within the study area, i.e., Clinton River and Rouge River, come from USGS [38] datasets to simulate the groundwater recharge through surface waters. The data for the 47 inland lakes with an area greater than 0.3 km^2 are also collected from the Michigan Department of Natural Resources [39]. Within the Metro Detroit area, there are 12,866 active pumping wells that are mostly located in the northwestern portion of the region and categorized into five groups: industrial, household, irrigation, public supply, and commercial/institutional wells. Their data are gathered from USGS groundwater datasets for Michigan and applied to the groundwater model.

Code Availability

All GMS-MODFLOW models and Python 3.10 programming codes used for the data analysis in this research are available from the corresponding author upon a reasonable request.

References

1 Miller, C. J. *et al.* A Review of Volatile Organic Compound Contamination in Post-Industrial Urban Centers: Reproductive Health Implications Using a Detroit Lens. *Int. J. Environ. Res. Public Health* **17**, 8755, doi:10.3390/ijerph17238755 (2020).

2 Jousma, G. & Roelofsen, F. World-wide inventory on groundwater monitoring. *Report nr. GP* **1** (2004).

3 Holmes, S. L. History of Water Resources activities of the US Geological Survey (water fact sheet). Report No. 2331-1258, (US Geological Survey, 1985).

4 Singh, C. K. & Katpatal, Y. B. A Review of the Historical Background, Needs, Design approaches and Future Challenges in Groundwater Level Monitoring Networks. *Journal of Engineering Science & Technology Review* **13**, doi:10.25103/jestr.132.18 (2020).

5 Zhou, Y., Dong, D., Liu, J. & Li, W. Upgrading a regional groundwater level monitoring network for Beijing Plain, China. *Geoscience Frontiers* **4**, 127-138, doi:10.1016/j.gsf.2012.03.008 (2013).

6 Asefa, T., Kemblowski, M. W., Urroz, G., McKee, M. & Khalil, A. Support vectors–based groundwater head observation networks design. *Water Resources Research* **40**, doi:10.1029/2004WR003304 (2004).

7 Chen, C., He, W., Zhou, H., Xue, Y. & Zhu, M. A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Scientific Reports* **10**, 1-13, doi:10.1038/s41598-020-60698-9 (2020).

8 El Bilali, A., Taleb, A. & Brouziyne, Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agricultural Water Management* **245**, 106625 (2021).

9 Jaafarzadeh, M. S., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H. R. & Rouhani, H. Groundwater recharge potential zonation using an ensemble of machine learning and bivariate statistical models. *Scientific Reports* **11**, 1-18 (2021).

10 Podgorski, J. & Berg, M. Global threat of arsenic in groundwater. *Science* **368**, 845-850, doi:10.25103/jestr.132.18 (2020).

11 Xu, T., Valocchi, A. J., Choi, J. & Amir, E. Use of machine learning methods to reduce predictive error of groundwater models. *Groundwater* **52**, 448-460, doi:10.1111/gwat.12061 (2014).

12 Ammar, K., Khalil, A., McKee, M. & Kaluarachchi, J. Bayesian deduction for redundancy detection in groundwater quality monitoring networks. *Water resources research* **44**, doi:10.1029/2006WR005616 (2008).

13 Khader, A. I. & McKee, M. Use of a relevance vector machine for groundwater quality monitoring network design under uncertainty. *Environmental modelling & software* **57**, 115-126, doi:10.1016/j.envsoft.2014.02.015 (2014).
14 Lal, A. & Datta, B. Application of Monitoring Network Design and Feedback Information for Adaptive Management of Coastal Groundwater Resources. *International Journal of Environmental Research and Public Health* **16**, 4365, doi:10.3390/ijerph16224365 (2019).
15 Kodinariya, T. M. & Makwana, P. R. Review on determining number of Cluster in K-Means Clustering. *International Journal* **1**, 90-95 (2013).
16 Jousma, G. *et al.* Guideline on: Groundwater monitoring for general reference purposes. *International Groundwater Resources Assessment Centre (IGRAC), Utrecht* (2006).
17 Navy, U. Guide to optimal groundwater monitoring. (Technical report, Naval Facilities Engineering Service Center, 2000. 2.1, 2000).
18 Harbaugh, A. W. *MODFLOW-2005, the US Geological Survey modular ground-water model: the ground-water flow process*. (US Department of the Interior, US Geological Survey Reston, VA, 2005).
19 Wang, S. *et al.* Application of MODFLOW and geographic information system to groundwater flow simulation in North China Plain, China. *Environmental Geology* **55**, 1449-1462, doi:10.1007/s00254-007-1095-x (2008).
20 Panagopoulos, G. Application of MODFLOW for simulating groundwater flow in the Trifilia karst aquifer, Greece. *Environmental Earth Sciences* **67**, 1877-1889, doi:10.1007/s12665-012-1630-2 (2012).
21 Bailey, R. T., Park, S., Bieger, K., Arnold, J. G. & Allen, P. M. Enhancing SWAT+ simulation of groundwater flow and groundwater-surface water interactions using MODFLOW routines. *Environmental Modelling & Software* **126**, 104660, doi:10.1016/j.envsoft.2020.104660 (2020).
22 Hughes, J. D., Langevin, C. D. & White, J. T. MODFLOW-based coupled surface water routing and groundwater-flow simulation. *Groundwater* **53**, 452-463 (2015).
23 Aquaveo. *GMS:Stochastic Modeling*, <https://www.xmswiki.com/wiki/GMS:Stochastic_Modeling> (2005).
24 Atangana, A. in *Fractional Operators with Constant and Variable Order with Application to Geo-Hydrology* (ed Abdon Atangana) 15-47 (Academic Press, 2018).
25 El Mezouary, L. & El Mansouri, B. in *E3S Web of Conferences.* 04008 (EDP Sciences).
26 Helton, J. C. & Davis, F. J. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety* **81**, 23-69 (2003).
27 Khader, A. I. Value of Information in Design of Groundwater Quality Monitoring Network Under Uncertainty. (2012).
28 Žalik, K. R. An efficient k′-means clustering algorithm. *Pattern Recognition Letters* **29**, 1385-1391, doi:10.1016/j.patrec.2008.02.014 (2008).
29 Bandyopadhyay, S. & Maulik, U. An evolutionary technique based on K-means algorithm for optimal clustering in RN. *Information Sciences* **146**, 221-237, doi:10.1016/S0020-0255(02)00208-6 (2002).
30 Cui, M. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance* **1**, 5-8 (2020).
31 Nainggolan, R., Perangin-angin, R., Simarmata, E. & Tarigan, A. F. in *Journal of Physics: Conference Series.* 012015 (IOP Publishing).
32 Syakur, M., Khotimah, B., Rochman, E. & Satoto, B. D. in *IOP conference series: materials science and engineering.* 012017 (IOP Publishing).
33 Tipping, M. E. in *S. A. Solla, T. K. Leen, and K. R. Muller, editors, Advances in Neural Information Processing Systems 12* Vol. 1 652-658 (MIT Press, 2000).
34 Fletcher, T. Relevance vector machines explained. *University College London: London, UK.* (2010).

35 Teimoori, S., O'Leary, B. F. & Miller, C. J. Modeling Shallow Urban Groundwater at Regional and Local Scales: A Case Study in Detroit, MI. *Water* **13**, 1515 (2021).
36 Department of Environmental Quality. *Michigan Groundwater Maps*, <https://www.egr.msu.edu/igw/GWIM%20Figure%20Webpage/> (2003).
37 Wellogic System. *Department of Environmental Quality (DEQ), State of Michigan's statewide groundwater database*, <https://secure1.state.mi.us/wellogic/Login.aspx?ReturnUrl=%2fwellogic%2fdefault.aspx> (2019).
38 USGS. *National Water Information System data available on the World Wide Web (USGS Water Data for the Nation), United States Geological Survey*, <https://waterdata.usgs.gov/mi/nwis> (2020).
39 Michigan Department of Natural Resources. *DNR Open Data*, <https://gis-midnr.opendata.arcgis.com/search?collection=Dataset> (2019).

Acknowledgments

This work is financially supported by Wayne State University through the Office of the Vice President for Research and by Healthy Urban Waters through the Fred A. and Barbara M. Erb Family Foundation. In addition, the authors gratefully acknowledge Dr. Yongli Zhang (Wager), Dr. Timothy M. Dittrich, and Dr. Shirley A. Papuga from Wayne State University for supporting this research and providing feedback.

Author contributions

S.T. and C.J.M. conceptualized the study. S.T. and M.O. contributed to the methodology and study design. S.T. retrieved/preprocessed data, wrote the software code, performed validation, formal analysis, investigation, interpretation of the results, visualization, and wrote the original draft. C.J.M. supervised the work and managed financial support. All authors contributed to the review/revision of the paper.

Competing Interest The authors declare no competing interests

Additional Information

Correspondence and requests for materials should be addressed to S.T.