Journal of Computational and Applied Mathematics 380 (2020) 112982

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Machine learning and transport simulations for groundwater anomaly detection

Jiangguo Liu [a,*], Jianli Gu [a], Huishu Li [b], Kenneth H. Carlson [b]

[a] *Department of Mathematics, Colorado State University, Fort Collins, CO 80523-1874, USA*
[b] *Department of Civil & Environmental Engineering, Colorado State University, Fort Collins, CO 80523, USA*

ARTICLE INFO

Article history:
Received 9 February 2019
Received in revised form 2 May 2020

Keywords:
Anomaly detection
Contaminant transport
Darcy flow
Groundwater
Machine learning
One-class support vector machine

ABSTRACT

This paper presents studies on modeling and algorithms for groundwater anomaly detection. Specifically, conductivity along with four other surrogates are used for identifying anomaly in groundwater, the one-class support vector machine (1-SVM) technique is utilized for model training, and real data from `Colorado Water Watch` is used for testing the model and algorithms. Design of code modules in `Python` is briefly discussed. Since groundwater contamination rarely happens in reality, we also use synthetic data from numerical simulations of flow and transport in porous media to test this model for anomaly detection.

1. Introduction

Recently the public has become increasingly concerned over groundwater quality due to unconventional oil and natural gas operations, which may raise the risk of groundwater contamination and methane emission. Unconventional oil and natural gas operations, especially hydraulic fracking [1], involve injection of massive hydraulic fluid to thousands of feet deep to extract oil or natural gas from unconventional formations (shale or sandstone). With the production of oil and/or natural gas, a large amount of "wastewater" (deep formation groundwater combined with chemical additives used in fracking) is generated and brought to the surface. During this process, although shallow groundwater aquifers are protected by several layers, "toxic" aqueous phase chemicals [2], petroleum produces, or methane [3] may leak to the drinking water aquifers due to casing failure.

Development of quantitative models and establishment of monitoring network for groundwater assessment will not only provide the baseline information but also detect contamination events in an early stage to avoid a wide spread of contamination. This relies on analysis of real-time data for groundwater and understanding of its transport mechanism.

There are various possible reasons for groundwater anomaly, for example, soil contamination, irrigation, natural disasters, season shifts, and failure in sensor maintenance. It may not be straightforward to characterize groundwater quality. However, the following five major surrogates are well accepted and used in many studies:

- Temperature; pH value; Conductivity;
- Oxidation–Reduction Potential (ORP);
- Dissolved Oxygen (DO).

[*] Corresponding author.
E-mail addresses: liu@math.colostate.edu (J. Liu), jianligu@hotmail.com (J. Gu), huishu@rams.colostate.edu (H. Li), kcarlson@engr.colostate.edu (K.H. Carlson).

https://doi.org/10.1016/j.cam.2020.112982

For anomaly detection, the majority of measurements or observations are *normal* but very few measurements are in *anomaly*. Groundwater anomaly refers to deviation of surrogate measurements from the normal tracks in a monitoring system, which indicate changes in groundwater quality. Due to complexity of subsurface aquatic environment, groundwater movement, and disturbance from human beings such as oil and gas exploration activities, groundwater anomaly patterns vary but appear in **four common types**: (a) *Single anomaly* contains only one anomaly observation at some time-stamp, the value of this observation appears to be obviously different from the normal observations; (b) *Continuous anomaly* contains a series of anomaly observations during a period of time; (c) *Baseline change* in observations are common in groundwater monitoring, these are usually caused by season shifts, changes in precipitation or stream charges; (d) *Fluctuation in observations* may also be classified as anomaly due to limited priori knowledge. These are also known as false negatives. See Fig. 1 for four illustrating examples.

There are many methodologies for detecting anomalies in sequential data or time series, e.g., Bayesian networks [4], decision trees [5], multivariate state estimation technique (MSET) [6], and support vector machines (SVMs) [7,8]. There have been efforts on applying machine learning algorithms including SVMs to data analysis for wastewater monitoring and processing [3,9,10]. However, as of our best knowledge, there been no reports yet on applying machine learning algorithms to real-time anomaly detection for groundwater. This paper presents our latest research results in this regard.

The research areas for groundwater quality monitoring, environmental monitoring, reservoir simulations, and anomaly detection are truly interdisciplinary and actively developing. Supervised and unsupervised machine learning methods were combined for groundwater quality assessment [11]; Deep learning methods were investigated for flush calculations in reservoir simulations [12]; Establishment of a realistic and public dataset with rare undesirable real events in oil wells was investigated in [13] so that it can be used as a benchmark for development of machine learning techniques for detecting and diagnosing undesirable events; Intelligent one-class techniques were developed in [14] for anomaly detection in an industrial facility to obtain the main material for wind generator blades production; Complex event processing software modules have been developed for intelligent environmental monitoring [15].

The results in this paper are from an interdisciplinary project that develops and validates a model for groundwater anomaly detection. Specifically,

(i) We use five surrogates (temperature, pH value, conductivity, oxidation–reduction potential, and dissolved oxygen) to monitor groundwater quality, especially potential contamination;
(ii) Machine learning algorithms based on 1-SVM are applied to groundwater anomaly detection;
(iii) The algorithm is implemented as `Python` code modules, which utilize the `Python` interface offered by `Libsvm`, a popular `C++` library for machine learning and data mining;
(iv) Real-time data from `Colorado Water Watch` [16] is used to test our model for groundwater anomaly detection;
(v) Since groundwater contamination rarely happens in reality and there is not much data available for testing, we test our detection model also by synthetic data from numerical simulations of contaminant transport, which obeys certain first principles in physics.

2. Basic concepts of SVMs and 1-SVM

Support vector machines are among the classical techniques for data classification and machine learning. This section briefly reviews the basic concepts of SVMs.

Data from observations (measurements or instrument readings) for certain phenomena, either numeric or nonnumeric, can be quantified in an input space. For example, for a human being, we may record data on gender, age, height, weight, etc. For groundwater, we collect data on temperature, pH value, conductivity, oxidation–reduction potential (ORP), and dissolved oxygen (DO). Clearly, measurements for these five surrogates for groundwater can be treated together as vectors in $\mathbb{R}^5$.

For data analytics, features are extracted from raw data. A set of numeric features can be formed as a feature vector. An inner product can be established on a feature space. Mappings from an input space to a feature space, especially kernel functions, are useful tools for data analytics.

One practical task in data analytics is data classification, including anomaly detection. For example, in drug discovery, one needs to separate active compounds from inactive compounds. For groundwater monitoring, one needs to separate anomaly status from normal status. Support vector machines are among the useful classical techniques for data classification.

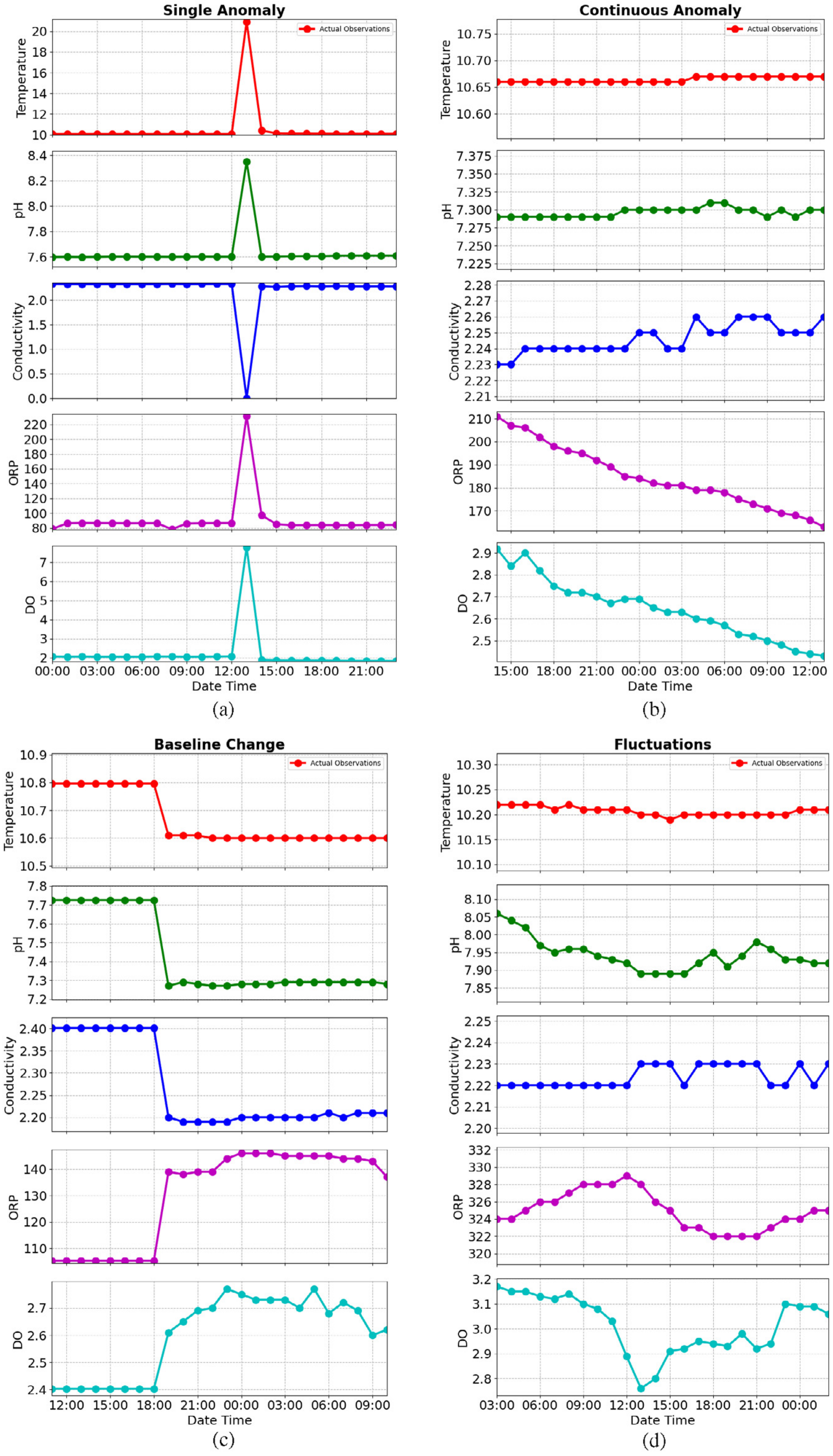
It is well accepted that SVM is a classifier or a supervised machine learning algorithm. We look for a separation hyperplane that can provide a binary classification. In other words, we seek $\mathbf{b} \in \mathbb{R}^m$, $c \in \mathbb{R}$, and consider

$$f(\mathbf{x}) = \mathbf{b} \cdot \mathbf{x} + c, \quad \forall \mathbf{x} \in \mathbb{R}^m. \tag{1}$$

Given data points $\mathbf{x}_j (1 \le j \le n)$ in $\mathbb{R}^m$ and their classification $y_j \in \{+1, -1\}$, we expect $f(\mathbf{x}_j)y_j \ge 1$. As shown in Fig. 2,

$$\mathbf{b} \cdot \mathbf{x} + c = 0$$

defines a hyperplane in $\mathbb{R}^m$. Accordingly, those vectors on the two parallel hyperplanes satisfying $f(\mathbf{x}_j)y_j = 1$ are called *supporting vectors*. Fig. 2 illustrates the main ideas of SVM.



(a)

(b)

(c)

(d)

Fig. 1. Four common types of groundwater anomaly: (a) single anomaly; (b) continuous anomaly; (c) baseline change; (d) fluctuation in observations.

SVM is essentially an optimization problem. We want a hyperplane with the maximal margin. That is, we seek $\mathbf{b} \in \mathbb{R}^m, c \in \mathbb{R}$ such that $\|\mathbf{b}\|$ is the smallest whereas $f(\mathbf{x}_j)y_j \geq 1$ for $1 \leq j \leq n$. Such an optimization problem can

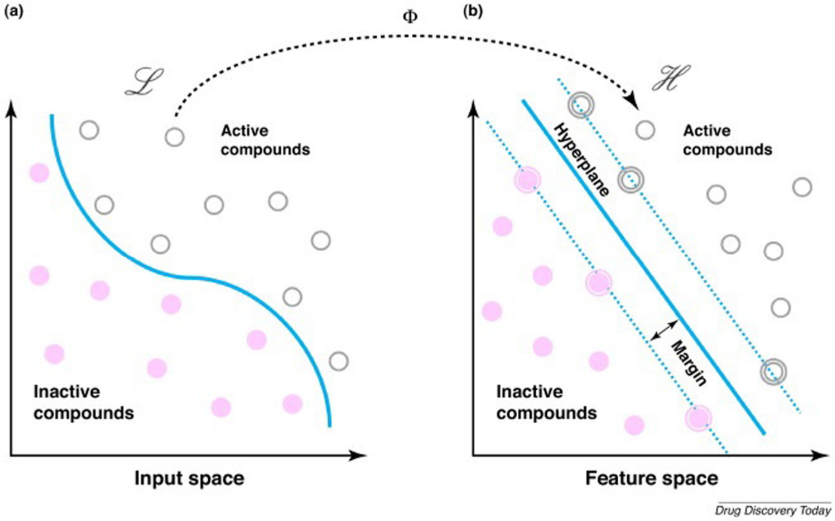


Fig. 2. SVM used in drug discovery: Separating active compounds from inactive ones.
Source: Lavecchia, *Drug Disc. Today* (2015) [17].

be formulated in the dual form. This further allows soft margin and slack variables for hard-to- or non-separable data. Kernel techniques are also widely used with SVM for nonlinear data.

Note that SVM is a supervised machine learning technique that requires a training dataset with two-class labels: $+1$ and -1. This allows it to fit a hyper-plane that maximally divides the two classes. However, the one-class SVM or 1-SVM is an unsupervised machine learning technique [18]. It enables training of a classifier by using only samples within the positive class, and check if a certain number of data points belong to the negative class. The idea is to define a hyper-boundary for the majority of the positive data points, whereas the minority of the data points staying outside the boundary are classified as outliers or anomalies. The minority portion is controlled by a parameter in the training module.

3. Applications of 1-SVM to groundwater anomaly detection

For groundwater anomaly detection investigated in this paper, the sequential multivariate data collected from the monitoring sites in the oil/gas fields is full of uncertainty due to unknown processes. This nature of data affects correctness of the conclusion drawn from the mathematical and/or statistical methods that require homogeneous data. However, our approach with 1-SVM can overcome this limitation, as the training algorithm classifies incoming data points in an incremental procedure, in other words, the machine learning technique is able to learn through time.

3.1. 1-SVM model for groundwater anomaly detection

The accuracy of a 1-SVM classifier highly depends on hyper-parameters ν and γ in the training stage (Fig. 3). To find the best ν and γ, the basic procedure contains selection of normal observations from historic data, and grid search based on k-folds cross validation.

Grid search performs exhaustive searching in a manually specified subset of hyper-parameter space. The specified parameter intervals should be investigated for performing valid searching usually conducted in log-space. The final outputs of grid search are the settings that achieve the best score of cross validation. To ensure that the optimized parameters do not overfit the training dataset, a tuning trick (regularization) comes into play. See [19] for more details.

Furthermore, `SciKit-Learn` is a popular machine learning package for Python, which is well documented and provides detailed descriptions on grid search. The interested reader is referred to https://scikit-learn.org/stable/

3.2. `Python` code modules

To validate the anomaly detection model presented in this paper, we have developed a `Python` code package. The package utilizes `Libsvm` (Version 3.11), a `C++` library for support vector machines, which has gained popularity for data classification and regression for machine learning and data mining. The library provides a `Python` interface.

Our code package contains two main modules: training and detection. These two modules work together through close interaction [19].

(i) The training module selects a classifier to be used by the detection module.

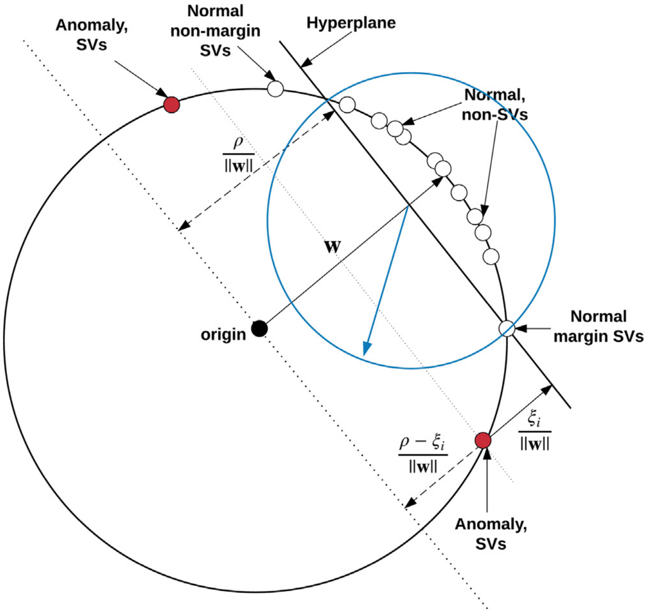


Fig. 3. An illustration for 1-SVM.

(ii) Detected (anomalous) observations from the detection module are fed back to the training module for updating the training dataset.

This ensures a dynamic evolvement of the 1-SVM classifier for (continuous) groundwater anomaly detection.

1. Training Module. A desired classifier needs to not only ensure the accuracy of classification, but also avoid overfitting. To obtain such a classifier in training, a pair of optimized hyper-parameters ν and γ need to be provided. This module uses an exhaustive grid search with k-folds cross validation to find the optimized hyper-parameters, and then use normal historic data to train such a classifier.

```
from libsvm.python import svmutil
def train(observations,labels,optimal_gamma,optimal_nu):
    # -s 2: one-class svm
    # -t 2: gaussian kernel
    # -n: nu
    # -g: gamma
    prob = svmutil.svm_problem(labels, observations)
    param = svmutil.svm_parameter("-s 2 -t 2 -n 0.1 -g 0.1")
    param.gamma, param.nu = optimal_gamma, optimal_nu
    model = svmutil.svm_train(prob, param)
    return model
```

2. Detection Module. Built on top of the trained classifier, this module is used to detect and label the state of new observations during the groundwater monitoring process. The output of detection is just a binary information: +1 for normal or −1 for anomaly. The sequential states form various data patterns in time-series, such as normal pattern, single-anomaly pattern, baseline change, and continuous-anomaly pattern.

```
from libsvm.python import svmutil
def predict(model, observation):
    # model: the trained model
    # observation: a single observation
    pred_label = svmutil.libsvm.svm_predict(model,observation)
```

Table 1
Five surrogates: Names, abbreviations, and their units.

Name	Abbreviation	Unit
Temperature	Temp	°C
pH	pH	1
Conductivity	Cond	s/m
Oxidation–reduction potential	ORP	mV
Dissolved oxygen	DO	mg/L

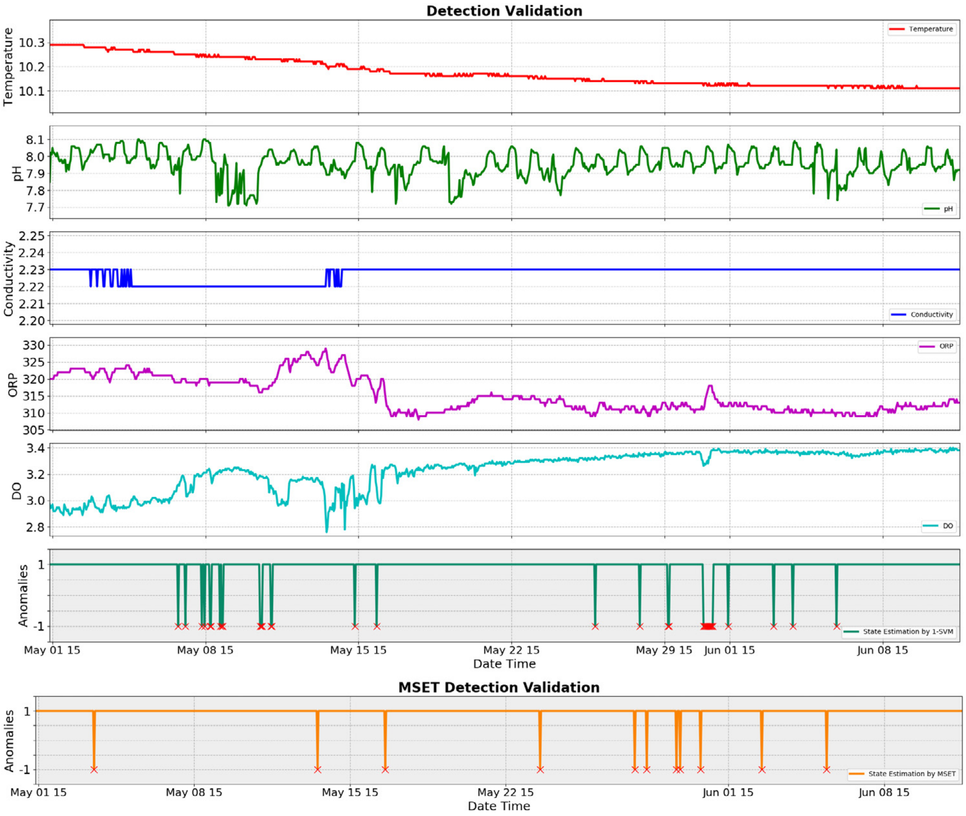


Fig. 4. *Row 1–5*: Real data collected by `Colorado Water Watch`; *Row 6*: Results of 1-SVM; *Row 7*: Results of MSET.

3.3. Results of groundwater anomaly detection by 1-SVM

In this subsection, we present some results on applying the 1-SVM model to groundwater anomaly detection. Real data from `Colorado Water Watch` is used. First, we list the abbreviations and units of the five surrogates in Table 1.

We use the 1-SVM classifier in real-time, that is, for the most recent observation: (1) If it is labeled as normal, then the observation is added to the training dataset; (2) Otherwise, it is labeled as anomaly.

In Fig. 4, Rows 1–5 show the plots of real data from `Colorado Water Watch` for the five surrogates: temperature, pH, conductivity, ORP, and DO. Row 6 shows the results of anomaly detection by 1-SVM. 35 cases of anomaly were reported, some correspond to noisy observations.

3.4. Comparison with Multivariate State Estimation Technique (MSET)

For purpose of comparison, this subsection briefly discusses the multivariate state estimation technique (MSET) for anomaly detection. MSET is an advanced pattern recognition technique originally developed by Argonne National Laboratory for fault monitoring in nuclear plant system, see [6] and references therein. This technique has also been used to detect anomaly in enterprise servers [20] and video streams [21], among many other applications.

Roughly speaking, MSET consists of two essential modules: estimation and anomaly detection [22]. Nonlinear and nonparametric techniques are used to train a group of weight values over some historic observations, then the sequential probability ratio test is used for examination of residual means and variances and thus anomaly detection.

MSET can also be applied to real-time groundwater anomaly detection as demonstrated in this paper and [19]. For each time-stamp t_j, the state of a new observation x_{obs} is estimated based on a training dataset, then a value "normal" $(+1)$ or "anomaly" (-1) is assigned. In general, MEST is less sensitive to signal fluctuations.

Shown in Fig. 4 last row is the result obtained from applying MSET. As a comparison, we see 35 anomaly cases were detected by 1-SVM, whereas 11 cases were detected by MSET. This shows 1-SVM maybe more sensitive than MSET, which may be beneficial for some applications.

4. Numerical methods for flow and transport simulations

Contaminant transport in groundwater is among many real-world problems that can be modeled as flow and transport in porous media. Although complicated, these processes obey certain physical principles such as conservation of mass and momentum, based on which, differential equations can be established. It is usually very difficult, if not impossible, to find exact (analytical) solutions for such equations. Numerical methods are inexpensive and practically useful tools. Some early work on numerical methods for flow and transport in porous media can be found in [23,24]. Work on the engineering and application perspectives can be found in [25]. Our recent paper [26] examined the coding and implementation perspectives and introduced an easy-to-use `Matlab` code package `DarcyLite`. [27] is another recent paper that investigated coupling of the Darcy and saturation equations for a two-phase model for subsurface flow.

4.1. Numerical approximation for Darcy flow

The Darcy equations for flow in porous media are prototyped as

$$\begin{cases} \mathbf{u} = -\mathbf{K}\nabla p, & \mathbf{x} \in \Omega, \\ \nabla \cdot \mathbf{u} = f, & \mathbf{x} \in \Omega, \\ p = p_D, & \mathbf{x} \in \Gamma^D, \\ \mathbf{u} \cdot \mathbf{n} = u_N, & \mathbf{x} \in \Gamma^N, \end{cases} \tag{2}$$

where $\Omega \subset \mathbb{R}^d (d = 2, 3)$ is a 2-dim or 3-dim bounded domain, $\mathbf{K}$ is the hydraulic conductivity (medium permeability divided by fluid viscosity), f is a discharge (source and sink), p is the unknown fluid pressure, $\mathbf{u}$ is the Darcy velocity, p_D is a Dirichlet boundary condition on Γ^D, u_N is a Neumann boundary condition on Γ^N, $\mathbf{n}$ is the unit outward normal vector on the domain boundary $\partial\Omega$ that has a nonoverlapping decomposition $\Gamma^D \cup \Gamma^N$.

There is a list of good numerical methods for solving the Darcy equations, e.g., the continuous Galerkin (CG) finite element methods with post-processing procedures [28], the discontinuous Galerkin (DG) finite element methods [29], the enriched Galerkin (EG) finite element methods [30], the weak Galerkin (WG) finite element methods [31,32], and the mixed finite element methods (MFEMs) [33,34]. All these methods possess two important properties: *local mass conservation* and *flux normal continuity*. A brief comparison of these methods and their `Matlab` implementations can be found in [26,35].

For some applications, the hydraulic conductivity and water-head elevation (pressure) are known geological data, and hence the Darcy velocity can be obtained via direct calculations from the 1st equation in (2). We shall discuss such a procedure on a uniform rectangular mesh in 2-dim by using the classical Raviart–Thomas spaces $RT_{[0]}$, see [31,34].

We consider a rectangular domain $\Omega = [x_a, x_b] \times [y_c, y_d]$ that is partitioned into a uniform rectangular mesh $\mathcal{E}_h$ with n_x, n_y being the numbers of partitions and $h_x = (x_b - x_a)/n_x$, $h_y = (y_d - y_c)/n_y$. For each rectangular volume in the mesh, its four edges are oriented as bottom, right, top, and left. Assume that

- The water-head elevation or pressure is a known constant on each volume: $p_{i,j}, i = 1, \ldots, n_x, \;\; j = 1, \ldots, n_y$;
- The hydraulic conductivity is also a known positive constant on each volume $K_{i,j}, i = 1, \ldots, n_x, \;\; j = 1, \ldots, n_y$.

We establish an approximation $\mathbf{u}_h$ to the Darcy velocity from the global Raviart–Thomas space $RT_{[0]}$. This is accomplished by considering approximation to the normal flux $\int_e (\mathbf{u}_h \cdot \mathbf{n}_e)$ for each single edge e in the mesh.

To illustrate the idea, we consider a vertical edge e shared by two volumes $E_{i,j}$ and $E_{i+1,j}$. Let the normal vector be $[1, 0]^T$. We only need to consider an approximation to the partial derivative $\partial_x p$ across this edge:

$$(\partial_x p)|_e \approx \frac{p_{i+1,j} - p_{i,j}}{h_x}.$$

The hydraulic conductivity takes two different values on these adjacent volumes. We consider their harmonic average for the shared edge e [31]:

$$K|_e = \frac{2K_{i,j}K_{i+1,j}}{K_{i,j} + K_{i+1,j}}. \tag{3}$$

The bulk normal flux across this vertical edge e is then

$$F_e = \int_e (\mathbf{u}_h \cdot \mathbf{n}_e) = \int_e \Big(-K|_e (\nabla p) \cdot \mathbf{n}_e \Big) = -K|_e \, \frac{p_{i+1,j} - p_{i,j}}{h_x} \, h_y. \tag{4}$$

For the element on the left, this edge is its 2nd edge, the bulk normal flux on this edge is F_e; For the element on the right, this edge is its 4th edge, the bulk normal flux is $-F_e$.

When this vertical edge is on the domain boundary, we set $F_e = 0$. Horizontal edges can be treated similarly. Utilizing these edgewise normal fluxes, one can recover a numerical velocity in the global $RT_{[0]}$ space [31,34].

4.2. Numerical approximation for concentration transport

The transport equation for contaminant concentration is prototyped as

$$\begin{cases} \partial_t c + \nabla \cdot (\mathbf{v}c - D\nabla c) = s(x,y,t), \quad (x,y) \in \Omega,\ t \in (0,T], \\ c(x,y,t) = 0, \quad (x,y) \in \partial\Omega, t \in (0,T], \\ c(x,y,0) = c_0(x,y), \quad (x,y) \in \Omega, \end{cases} \tag{5}$$

where $c(x, y, t)$ is the unknown concentration, $\mathbf{v} = \phi\mathbf{u}$ is the fluid velocity with ϕ being the porosity of the porous medium and $\mathbf{u}$ being the previously discussed Darcy velocity, $D > 0$ is the diffusion coefficient, s is the source term for the contaminant, $c_0(x, y)$ is an initial condition.

Various types of numerical solvers have been developed for the transport equation [26], e.g., upwinding finite difference/volume methods, discontinuous Galerkin and weak Galerkin finite element methods, characteristic tracking (ELLAM, Semi-Lagrangian) methods. Two main concerns are about (i) positivity of the numerical concentration; (ii) conditions for stability of numerical schemes that usually restrict the time step size.

Here in this paper, we focus on a convective transport problem, i.e., $D = 0$ (no diffusion):

$$\partial_t c + \nabla \cdot (\mathbf{v}c) = s. \tag{6}$$

Then the upwinding finite volume method can be applied [27].

Assume $\mathcal{E}_h$ is the same uniform rectangular mesh used in Section 4.1. Let $[0, T]$ be the time period with a uniform partition such that $\Delta t = T/N$ and $t_n = n\Delta t\,(n = 0, 1, \dots, N)$. At each time step t_n, we approximate the concentration by an elementwise constant

$$C_{i,j}^{(n)}, \quad i = 1, \dots, n_x, \quad j = 1, \dots, n_y.$$

On a typical finite volume E, Eq. (6) is rewritten in the integral form as, after applying Gauss divergence theorem,

$$\int_E \partial_t c + \int_{E^\partial} (\mathbf{v} \cdot \mathbf{n})c = \int_E s.$$

Based on an explicit Euler approximation, we have (using $|E|$ to denote the area of the finite volume)

$$\frac{C_E^{(n+1)} - C_E^{(n)}}{\Delta t}|E| + C^{(n)}\phi \int_{E^\partial} (\mathbf{u}_h \cdot \mathbf{n}) = \int_E s(\cdot, \cdot, t_n). \tag{7}$$

Now we apply the upwinding technique to handle the 2nd term on the left-hand side of the above equation. Note that for each edge e on E^∂, the edgewise bulk normal flux $F_e := \int_e (\mathbf{u}_h \cdot \mathbf{n}_e)$ has been computed in Eq. (4).

- If $F_e \geq 0$, then mass is *flowing out* this volume E across this edge and $C^{(n)}$ is chosen as $C_E^{(n)}$, that is, the concentration value in this volume;
- If $F_e < 0$, then mass is *flowing into* this volume across this edge from a neighboring volume, we choose $C^{(n)}$ as the concentration value in the neighboring volume.

This is an explicit time-marching scheme, the Courant–Friedrichs–Lewy (CFL) condition needs to be satisfied:

$$\Delta t \leq \frac{\min(h_x, h_y)}{\max_{E \in \mathcal{E}_h} |\mathbf{v}|}. \tag{8}$$

It can be verified that for each time step in the simulation, the total mass $\sum_{E\in\mathcal{E}_h} C_E^{(n)}|E| = \text{const}$.

5. Anomaly detection for synthetic data

Groundwater contamination does not happen often in reality. There is not much real data of contamination available for testing. However, once contamination happens in groundwater, both groundwater flow and contaminant transport follow certain physical principles that can be modeled by differential equations. Groundwater flow obeys the Darcy's law (2) for flow in porous media, whereas contaminant transport can be modeled by the time-dependent convection–diffusion equations (5)(6).

In this section, we conduct a study on applying the detection model (See Section 3) to synthetic data obtained from numerical simulations. In other words, we assume hypothetically that contaminant is leaked at certain subsurface sites and then transported in a reservoir, we check whether and how accurately our model can detect anomalies in groundwater due to such contamination.

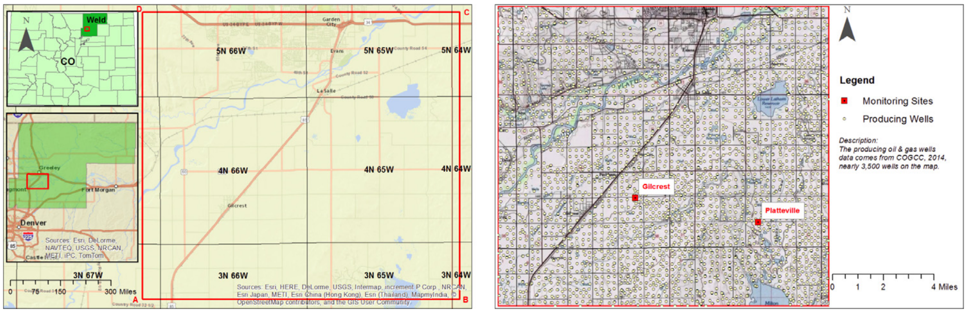

Weld
CO
Greeley
Fort Morgan
Denver
Sources: Esri, DeLorme, NAVTEQ, USGS, NRCAN, METI, iPC, TomTom
0 75 150 300 Miles
5N 66W
5N 65W
5N 64W
4N 66W
4N 65W
4N 64W
3N 67W
3N 66W
3N 65W
3N 64W
Gilcrest
La Salle
Sources: Esri, HERE, DeLorme, USGS, Intermap, increment P Corp., NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), MapmyIndia, © OpenStreetMap contributors, and the GIS User Community
A
B
C
D
N
Legend
Monitoring Sites
Producing Wells
Description:
The producing oil & gas wells data comes from COGCC, 2014, nearly 3,500 wells on the map.
Gilcrest
Platteville
0 1 2 4 Miles

Fig. 5. Maps for Wattenberg field (Colorado) for numerical simulations. *Left panel*: Map for the area; *Right panel*: Map for the wells in the 21 (km) by 19 (km) region. See [19] also.

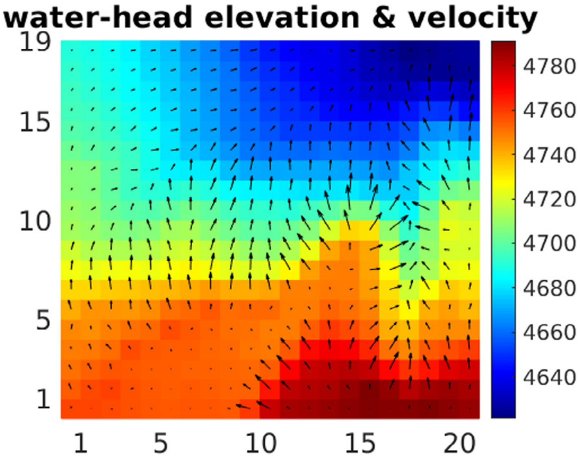


Fig. 6. Groundwater flow in Wattenberg field (a 21 km × 19 km region in northern Colorado) based on the geological data from [36] .

For chloride (`Cl`) transport in groundwater, listed below are some known facts.

- 20,000 (mg/L) is the typical concentration level in flowback water;
- The diffusion coefficient of chloride in groundwater is $\approx 2.030 \times 10^{-11} (\mathrm{m}^2/\mathrm{s})$;
- `Cl` concentration affects groundwater conductivity (`Cond`) linearly:

$$\mathrm{Cond} = 0.1084 * \mathrm{Cl}. \tag{9}$$

We consider a region in the Wattenberg field in northern Colorado that has geometric dimensions 21 km × 19 km. As shown in Fig. 5, there are about 3500 production wells in this region. There are also two monitoring sites: Gilcrest is located in the (9,7)-cell, Platteville is located at (18,6)-cell. We investigate chloride transport and its concentration in this region and correspondingly changes in groundwater conductivity, for which we test our anomaly detection model.

For the Darcy flow part, we use hydraulic conductivity data retrieved from `Colorado Decision Support Systems (CDSS)`, see Item (iii) in Acknowledgments for Data in Public Domains. The water-head elevation data is retrieved from `Colorado Geological Survey (CGS)`, see Item (iv) in Acknowledgments for Data in Public Domains. We assume the flow is confined in the aforementioned rectangular domain. Mathematically, we have $\Omega = [0, 21] \times [0, 19]$ and a no-flow boundary condition on $\partial\Omega$. Naturally, we consider a uniform rectangular mesh with $n_x = 21, n_y = 19$. The numerical method described in Section 4.1, particularly, formulas (3) and (4) are used to compute the edgewise normal fluxes and Darcy velocity. The profiles of velocity along with water-head elevation are presented in Fig. 6.

We assume hypothetically that chloride was initially found at two locations: one is located in the (8, 8)-cell, the other is located in the (19, 5)-cell. Then we utilize the numerical fluxes obtained from (4) and the upwinding finite volume scheme (7) to numerically simulate chloride transport in the Wattenberg field for a period of 6 years. The time step size used for simulations is $\Delta t = 1$ hour, which clearly satisfies the CFL condition. Shown in Fig. 7 are the profiles of chloride concentration simulated for the 0th (beginning) to the 5th year.

Assuming the initial leakage happened on 2016/01/01 at 00:00:00. The contaminant is then being transported in the groundwater. We also assume four other surrogates are kept as healthy constants for the time period of simulation. The values of the five surrogates are then fed into our anomaly detection model.

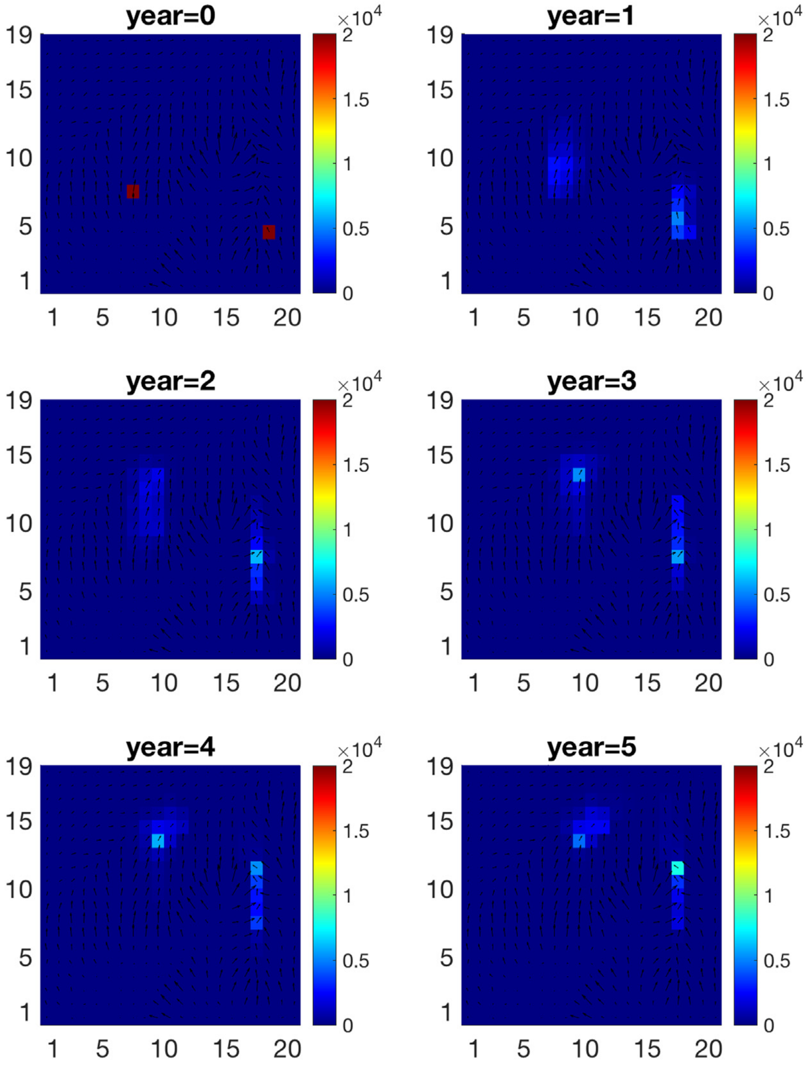


Fig. 7. Simulated chloride concentration for a period of 5 years.

Table 2
Chloride concentration & corresponding groundwater conductivity over 72 months.

t (months)	0	1	2	4	6	12	24	48	60	72
Chloride	21.00	262.93	817.15	2223.70	3489.56	4991.05	2901.77	318.02	91.40	26.97
Conductivity	2.28	28.50	88.58	241.05	378.27	541.03	314.55	34.47	9.91	2.92

- Fig. 7 shows one chunk of contaminant flows by the Platteville observation site (the one near the lower-right corner);
- Fig. 8 shows detection of anomaly as soon as in 6 days, since the contaminant is so close to the Platteville observation site located in the (18,6)-cell;
- Fig. 9 shows the increase in groundwater conductivity at the observation site for a period of about 12 months due to the increase in chloride concentration; Then the concentration level decreases but is still high for another 36 months.

Table 2 shows the chloride concentration values recorded for the Platteville observation site located in the (18,6)-cell at various time moments after the initial leakage. The table also shows the corresponding values of conductivity.

6. Concluding remarks

In this paper, we have presented studies on applying the technique of one-class support vector machine to groundwater anomaly detection with real-time data. To the best of our knowledge, this is the first ever attempt in such efforts. The real

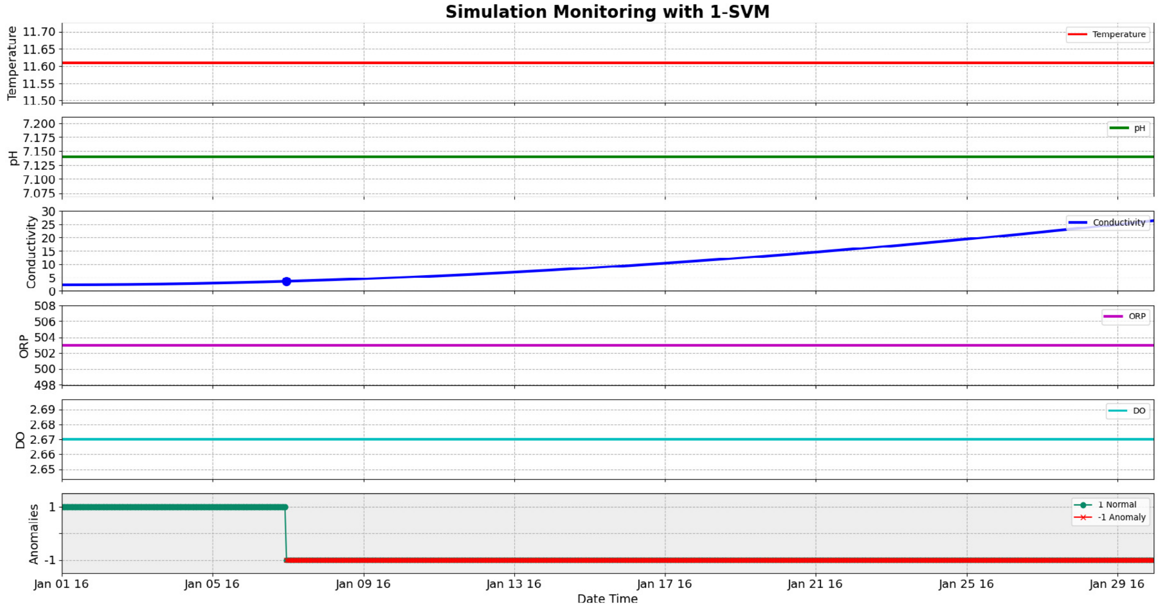


Fig. 8. Results of groundwater anomaly detection for the simulated contamination.

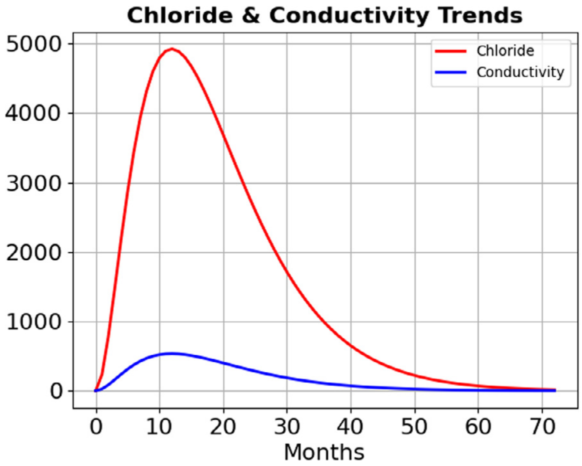


Fig. 9. Groundwater conductivity corresponding to the simulated chloride concentration for a period of 6 years.

data from Colorado Water Watch have been used for testing our model and code modules and demonstrated a good success rate. More results can be found in [19].

To further test our model and code for groundwater anomaly detection, we have also conducted experiments on synthetic data, which are obtained from numerical simulations of flow and transport in porous media. The Raviart–Thomas space and an upwinding finite volume scheme have been used to ensure local mass conservation and concentration positivity. These are regarded as novel applications of classical numerical methodology in an interdisciplinary research project that addresses realistic problems of public concern.

In addition to the use of 1-SVM, we have also used the multivariate state estimation technique (MSET) [6] for detecting anomaly in groundwater.

Further research can be pursued in several directions.

(i) Since methane is the major component of natural gas and it can be dissolved and transported with groundwater, correlation of methane concentration and groundwater anomaly is an interesting research topic [2].

(ii) Although there are four commonly known types of groundwater anomaly, most often there are a lot of normal measurements but very few anomaly measurements. As motivated by the research presented in [13], there should be efforts on collecting more real data of rare undesirable events in groundwater and analyzing their patterns.

(iii) More advanced machine learning algorithms [12] could be utilized for groundwater anomaly detection.
(iv) Data assimilation can be combined with anomaly detection for more advanced environmental monitoring.

Results from such efforts will be reported in our future work.

Acknowledgments for data in public domains

Some data used in our research for this paper are adopted from public domains, particularly, the following four domains/sites.

(i) `Colorado Water Watch`. This research project was led by Dr. Ken Carlson at Colorado State University. It provided real-time groundwater data collected at several monitoring sites in northern Colorado areas with frequent activities in oil and gas production:
https://source.colostate.edu/testing-the-water/
(ii) `Colorado Oil and Gas Information Systems (COGIS)` database maintained by Colorado Oil and Gas Conservation Commission (COGCC):
https://cogcc.state.co.us/data.html
(iii) `Colorado Decision Support Systems (CDSS)`, from which the hydraulic conductivity datasets were retrieved:
https://www.colorado.gov/pacific/cdss/division-1-south-platte
(iv) `Colorado Geological Survey (CGS)`, from which the groundwater elevation maps for the hydro-geologic characterization report of Gilcrest/LaSalle pilot project were derived:
https://coloradogeologicalsurvey.org/publications/gilcrest-lasalle-hydrogeologic-characterization/

Acknowledgments for Funding Support

J. Liu was partially supported by US National Science Foundation under grants DMS-1419077 and DMS-1819252. J. Gu was partially supported by US National Science Foundation under grant DMS-1419077.

References

[1] D.M. Jarvie, R.J. Hill, T.E. Ruble, R.M. Pollastro, Unconventional shale-gas systems: The Mississippian Barnett Shale of north-central texas as one model for thermogenic shale-gas assessment, Am. Assoc. Pet. Geol. Bull. 91 (4) (2007) 475–499.
[2] H. Li, K.H. Carlson, Distribution and origin of groundwater methane in the Wattenberg oil and gas field of northern Colorado, Environ. Sci. Technol. 48 (2014) 1484–1491.
[3] H. Li, J.-H. Son, K.H. Carlson, Concurrence of aqueous and gas phase contamination of groundwater in the wattenberg oil and gas field of northern colorado, Water Res. 88 (2016) 458–466.
[4] D.J. Hill, B.S. Minsker, E. Amir, Real-time Bayesian anomaly detection in streaming environmental data, Water Resour. Res. 45 (2009) W00D28.
[5] A. Muniyandi, R. Rajeswari, R. Rajaram, Network anomaly detection by cascading K-means clustering and C4.5 decision tree algorithm, Procedia Eng. 30 (2012) 174–182.
[6] R.M. Singer, K.C. Gross, J.P. Herzog, R.W. King, S. Wegerich, Model-Based Nuclear Power Plant Monitoring and Fault Detection: Theoretical Foundations, Technical Report, Argonne National Laboratory, 1997.
[7] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, NIPS'99 12 (1999) 582–588.
[8] I. Steinwart, D. Hush, C. Scovel, A classification framework for anomaly detection, J. Mach. Learn. Res. 6 (2005) 211–232.
[9] X. Qin, F. Gao, G. Chen, Wastewater quality monitoring system using sensor fusion and machine learning techniques, Water Res. 46 (2012) 1133–1144.
[10] H. Seshan, M.K. Goyal, M.W. Falk, S. Wuertz, Support vector regression model of wastewater bioreactor performance using microbial community diversity indices: Effect of stress and bioaugmentation, Water Res. 53 (2014) 282–296.
[11] R. Ratolojanahary, R.H. Ngouna, K. Medjaher, F. Dauriac, M. Sebilo, Groundwater quality assessment combining supervised and unsupervised methods, IFAC PapersOnLine 52–10 (2019) 340–345.
[12] Y. Li, T. Zhang, S. Sun, X. Gao, Accelerating flash calculation through deep learning methods, J. Comput. Phys. 394 (2019) 153–165.
[13] R.E.V. Vargas, C.J. Munaro, P.M. Ciarelli, A.G. Medeiros, B.G. do Amaral, D.C. Barrionuevo, J.C. de Araujo, J.L. Ribeiro, L.P. Magalhaes, A realistic and public dataset with rare undesirable real events in oil wells, J. Pet. Sci. Engrg. 181 (2019) 106223.
[14] E. Jove, J.L. Casteleiro-Roca, H. Quintián, J.A. Méndez-Pérez, J.L. Calvo-Rolle, Outlier generation and anomaly detection based on intelligent one-class techniques over a bicomponent mixing system, in: F. Martínez-Alvarez, A. Troncoso Lora, J.A. Saez Munoz, H. Quintian, E. Corchado (Eds.), 14th International Conference on Soft Computing Models in Industrial and Environmental Applications, Volume 950, Springer-Cham, 2019.
[15] A.Y. Sun, Z. Zhong, H. Jeong, Q. Yang, Building complex event processing capability for intelligent environmental monitoring, Environ. Model. Softw. 116 (2019) 1–6.
[16] H. Li, J.-H. Son, A. Hanif, J. Gu, A. Dhanasekar, K.H. Carlson, Colorado water watch: Real-time groundwater monitoring for possible contamination from oil and gas activities, J. Water Res. Prot. 9 (2017) 1660–1687.
[17] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discov. Today 20 (2015) 318–331.
[18] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. 13 (2001) 1443–1471.
[19] J. Gu, Mathematical Modeling of Groundwater Anomaly Detection (Master's thesis), Colorado State University, 2016.
[20] K. Whisnant, K.C. Gross, N. Lingurovska, Proactive fault monitoring in enterprise servers. in: Proceedings of the 2005 International Multiconference in Computer Science & Computer Engineering, 2005.
[21] K. Wang, J. Thompson, C. Peterson, M. Kirby, Identity maps and their extensions on parameter spaces: Applications to anomaly detection in video, in: 2015 Science and Information Conference (SAI), 2015, pp. 345–351.

[22] S. Cheng, M. Pecht, Multivariate state estimation technique for remaining useful life prediction of electronic products, Parameters 1 (2007) 26–32.

[23] Z. Chen, G. Huan, Y. Ma, Computational Methods for Multiphase Flows in Porous Media, SIAM, 2006.

[24] R.E. Ewing, The Mathematics of Reservoir Simulation, Volume 1 of Frontiers in Applied Mathematics, SIAM, 1983.

[25] C. Appelo, D. Postma, Geochemistry, Groundwater, and Pollution, second ed., CRC Press, 2010.

[26] J. Liu, F. Sadre-Marandi, Z. Wang, Darcylite: A Matlab toolbox for Darcy flow computation, Procedia Comput. Sci. 80 (2016) 1301–1312.

[27] V. Ginting, G. Lin, J. Liu, On application of the weak Galerkin finite element method to a two-phase model for subsurface flow, J. Sci. Comput. 66 (2016) 225–239.

[28] L. Bush, V. Ginting, On the application of the continuous Galerkin finite element method for conservation problems, SIAM J. Sci. Comput. 35 (2013) A2953–A2975.

[29] P. Bastian, B. Riviere, Superconvergence and $H(div)$ projection for discontinuous Galerkin methods, Internat. J. Numer. Methods Fluids 42 (2003) 1043–1057.

[30] S. Sun, J. Liu, A locally conservative finite element method based on piecewise constant enrichment of the continuous Galerkin method, SIAM J. Sci. Comput. 31 (2009) 2528–2548.

[31] G. Lin, J. Liu, L. Mu, X. Ye, Weak Galerkin finite element methdos for Darcy flow: Anistropy and heterogeneity, J. Comput. Phys. 276 (2014) 422–437.

[32] J. Liu, S. Tavener, Z. Wang, Lowest-order weak Galerkin finite element method for Darcy flow on convex polygonal meshes, SIAM J. Sci. Comput. 40 (2018) B1229–B1252.

[33] T. Arbogast, M. Correa, Two families of mixed finite elements on quadrilaterals of minimal dimension, SIAM J. Numer. Anal. 54 (2016) 3332–3356.

[34] F. Brezzi, M. Fortin, Mixed and Hybrid Finite Element Methods, Springer-Verlag, 1991.

[35] G. Lin, J. Liu, F. Sadre-Marandi, A comparative study on the weak Galerkin, discontinuous Galerkin, and mixed finite element methods, J. Comput. Appl. Math. 273 (2015) 346–362.

[36] P. Barkman, A. Horn, A. Moore, J. Pike, W. Curtiss, Gilcrest/LaSalle Pilot Project Hydrogeologic Characterization Report, Technical Report, Colorado Geological Survey, Golden, Colorado, 2014.