2020

Predicting Escherichia coli loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern

Olufemi P. Abimbola

Aaron R. Mittelstet

Tiffany Messer

Elaine D. Berry

Shannon L. Bartelt-Hunt

See next page for additional authors

Authors

Olufemi P. Abimbola, Aaron R. Mittelstet, Tiffany Messer, Elaine D. Berry, Shannon L. Bartelt-Hunt, and Samuel Hansen

Science of the Total Environment 722 (2020) 137894

Contents lists available at ScienceDirect

Science of the Total Environment

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

Predicting *Escherichia coli* loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern

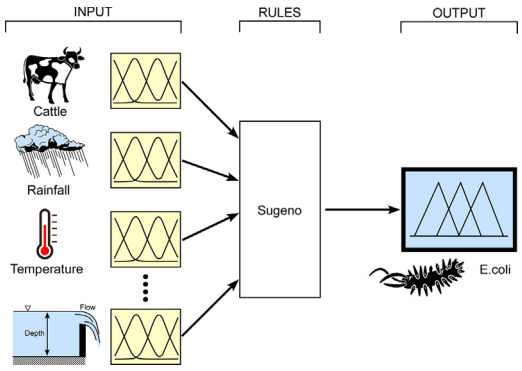
Olufemi P. Abimbola [a], Aaron R. Mittelstet [a,*], Tiffany L. Messer [a,b], Elaine D. Berry [c], Shannon L. Bartelt-Hunt [d], Samuel P. Hansen [a]

[a] Department of Biological Systems Engineering, University of Nebraska-Lincoln, 223 L. W. Chase Hall, Lincoln, NE 68583-0726, United States
[b] Conservation and Survey Division, School of Natural Resources, University of Nebraska-Lincoln, 101 Hardin Hall, 3310 Holdrege Street, Lincoln, NE 68583-0996, United States
[c] USDA Meat Animal Research Center, P.O. BOX 166, (State Spur 18D)/USDA-ARS-PA-MARC, Clay Center, NE 68933, United States
[d] Department of Civil and Environmental Engineering, University of Nebraska-Lincoln, 1110 S. 67th St., Omaha, NE 68182-0178, United States

HIGHLIGHTS

- Samples collected for six storm events and analyzed for *E. coli* concentration.
- *E. coli* modeled using machine learning at two cascading dams.
- Hydro-climatic variables and grazing density most important model parameters
- ANFIS models with FCM resulted in lowest errors of 0.17 logMPN/;100 mL and R^2 of 0.98.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 14 January 2020
Received in revised form 6 March 2020
Accepted 11 March 2020
Available online 12 March 2020

Editor: Jay Gan

Keywords:
E. coli prediction
Feature selection
Grazing pattern
Animal density estimation
Machine learning

ABSTRACT

Accurate prediction of *Escherichia coli* contamination in surface waters is challenging due to considerable uncertainty in the physical, chemical and biological variables that control *E. coli* occurrence and sources in surface waters. This study proposes a novel approach by integrating hydro-climatic variables as well as animal density and grazing pattern in the feature selection modeling phase to increase *E. coli* prediction accuracy for two cascading dams at the US Meat Animal Research Center (USMARC), Nebraska. Predictive models were developed using regression techniques and an artificial neural network (ANN). Two adaptive neuro-fuzzy inference system (ANFIS) structures including subtractive clustering and fuzzy c-means (FCM) clustering were also used to develop models for predicting *E. coli*. The performances of the predictive models were evaluated and compared using root mean squared log error (RMSLE). Cross-validation and model performance results indicated that although the majority of models predicted *E. coli* accurately, ANFIS models resulted in fewer errors compared to the other models. The ANFIS models have the potential to be used to predict *E. coli* concentration for intervention plans and monitoring programs for cascading dams, and to implement effective best management practices for grazing and irrigation during the growing season.

* Corresponding author.
E-mail address: amittelstet2@unl.edu (A.R. Mittelstet).

https://doi.org/10.1016/j.scitotenv.2020.137894
0048-9697/

1. Introduction

Microbiological impairment of surface waters has a major impact on the quality of human life. Water that is contaminated with fecal material is a common source of transmission of many pathogens that cause human and animal disease. Because *E. coli* is ubiquitous in the intestines of mammals and birds, its detection is considered to indicate fecal contamination. *E. coli* has been a common source identifier in microbial source tracking methods (Pachepsky and Shelton, 2011). In agricultural ecosystems, runoff from livestock pastures, as well as improper or over-application of manure, are common non-point sources of *E. coli* to surface waters and aquifers. *E. coli* contamination is a major concern near dams and reservoirs, in both agricultural and urban aquatic ecosystems, because of its implications on public health and food safety (Efting et al., 2011). Although the microorganisms are usually expected to have a low survival rate outside of the host organism (Zaleski et al., 2005), water resources have often been found to be contaminated (Unc and Goss, 2004). Furthermore, one of the limitations of current *E. coli* monitoring methods is the requirement for water samples to be collected, cultured, and incubated for several hours before colony growth is visible, and results are usually not available until the next day (Whitman et al., 2003). By the time the results are available, *E. coli* levels may have changed significantly. Thus, there is a need for faster methods for predicting *E. coli* concentrations.

In order to assess and manage natural water systems effectively, simulation models are often employed for predicting *E. coli* fate and transport. Prediction of *E. coli* fate in surface waters is complicated by the physical (e.g., temperature, UV light), chemical (e.g., pH, nutrients, sulfate, and nitrate), and biological (competing microflora, chlorophyll) factors and processes involved, which impede the development of useful and accurate predictive models (Flint, 1987; Sjogren and Gibson, 1981; Lessard and Sieburth, 1983; Robakis et al., 1983; Noguchi et al., 1997; Nevers and Whitman, 2005). Since it is almost impossible for any model to account for all these factors and heterogeneity involved in *E. coli* fate and transport, care should be taken not to generalize the results. For agricultural non-point source pollution, livestock waste deposition both on land and in streams is not well defined in terms of spatio-temporal patterns of loading, and concentrations of *E. coli* in livestock waste and manure vary widely.

The relationship between *E. coli* loads and fate and transport factors becomes more complex with the addition of flow rate (Whitman et al., 2004; McKergow and Davies-Colley, 2009). Whereas Vidon et al. (2008) observed *E. coli* loads were significantly higher at high flows compared to low flows, McKergow and Davies-Colley (2009) reported *E. coli* peak loads always preceded discharge and turbidity peaks even though both had similar timings. Thus, there is clearly a nonlinear relationship between *E. coli* and both flow and turbidity. *E. coli* in surface waters are associated with sediment, which influences their transport characteristics (Jamieson et al., 2005). Models that do not account for resuspension and deposition usually capture spatial trends successfully, but they tend to be incapable of explaining changes in concentrations in water during and after storm events (Hellweger and Masopust, 2008). Even when resuspension is incorporated into models, there is still a high level of uncertainty involved in predicting the amount of *E. coli* that has been resuspended. In most studies, the resuspension rate is either specified (Petersen et al., 2009) or expressed primarily as a function of flow (Tian et al., 2002; Collins and Rutherford, 2004).

Understanding the relative importance and the relationships among physical, chemical and biological variables is required to strengthen development of increasingly detailed models for predicting *E. coli* fate and transport in dams and other water bodies. However, for practical water-quality monitoring designs in dams, and in order to better inform environmental decision-making, it is important that predictive models are developed using variables that can be easily measured.

Several process-based models have been developed that use mass conservation principles (Baffaut and Benson, 2003; Coffey et al., 2007) and complex mechanistic and empirical relationships to predict *E. coli* loads in surface waters at different scales (Arnold and Fohrer, 2005; Pachepsky et al., 2006; Benham et al., 2006). However, the effectiveness of these models is limited due to excessively complex mechanistic relationships among input variables. The approximation and simplification of input parameters describing transport processes often results in high uncertainties in *E. coli* load estimations. Other models have been developed that use statistical and machine learning algorithms for predicting *E. coli* loads using variables such as water quality, meteorological, and hydrodynamic data. Regression methods have been used to predict *E. coli* levels (Brooks et al., 2016; Gonzalez et al., 2012; Nevers and Whitman, 2005, 2011; Shively et al., 2016). Nevers and Whitman (2005) used multiple linear regression to predict *E. coli* loads using turbidity, wave height, and lake chlorophyll for individual beaches of southern Lake Michigan, while Brooks et al. (2016) predicted *E. coli* concentrations at seven beaches in Wisconsin by applying multiple regression models. Linear mixed effects (LME) models were used to predict *E. coli* levels at Lake Michigan beaches (Jones et al., 2013). Park et al. (2018) evaluated and compared the performance of artificial neural network (ANN) and support vector regression (SVR) for predicting the concentration of *E. coli* at two recreational beaches.

Although different models are applicable for different surface water systems, such as reservoirs and freshwater lakes (Jin et al., 2003; Hipsey et al., 2008), streams and rivers (Medema and Schijven, 2001), as well as coastal lagoons and estuaries (Steets and Holden, 2003; McCorquodale et al., 2004), it is difficult for users to confidently implement these models since most of the physical, chemical and biological input variables cannot be easily measured. For this study, in order to predict *E. coli* concentrations at the outlets of two cascading dams, there was a need to use easily measured hydrometeorological variables (e.g. air temperature, water temperature, rainfall, water depth, and flow) as well as animal management variables (e.g. pasture utilization and animal density) that control its occurrence and sources. The objectives of this study were (i) to develop models to predict *E. coli* concentrations in cascading dams using regression, ANN and ANFIS by selecting and transforming the "most important" features (input variables) and (ii) to evaluate and compare the prediction accuracy of the machine learning models.

2. Materials and methods

2.1. Study area

This study was conducted at the U.S. Meat Animal Research Center (USMARC) near Clay Center, Nebraska, during summer and fall of 2018. During World War II, the site was used for the production and storage of ammunition, which led to groundwater contamination. A groundwater remediation plan was developed and implemented by the U.S. Army Corps of Engineers (USACE) in order to treat the contaminated groundwater water for agricultural reuse (USACE, 2010). The remediation plan, which involved the installation of abstraction wells and a water treatment facility, started operation in April 2013. The wells continuously remove and treat groundwater at a rate of 14,000 l/min throughout the year. The groundwater is then discharged as surface water into an existing stream at the Dischargee Well (DW), which flows 11.3 km through the USMARC property to an 81-ha reservoir (Fig. 1).

Nine cascading dams or grade control structures (GCS) restricted the flow of water across the site in order to store water for irrigation, suppress floods by preventing erosion from high-flow storm events, and recharge the underlying aquifer through percolation of the treated water. Five of nine GCSs (#1, 2, 4, 5, and 6) had the capability to control discharge by adding or removing stop logs, with a maximum of four stop logs per GCS. Each stop log is 1.2 m long and 0.3 m high.

Except at GCS4 which usually has no stop logs installed, in a normal spring, one or two stop logs were usually installed at the remaining four

logged GCSs (#1, 2, 5, and 6) once the reservoir below GCS9 was full (Fig. 2). However, because the winter and spring of 2018 were dry, only two stop logs were installed (Table 1). From September 1 to 4, 2018, USMARC received approximately 94 mm of rain (which was more than the long-term average of 62 mm for the month of September, and after also having 30% higher than normal rainfall in June, July, and August) which filled the reservoir. Due to the heavy rains in early September and risk of flooding, all the stop logs were installed between September 4 and 5, even at GCS4.

2.2. Hydrologic monitoring

For this study, hydrologic monitoring was conducted at the outflow from the groundwater treatment system (discharge well (DW)) and the first two GCSs downstream of DW (GCS1 and GCS2) using portable surface water samplers to collect water for determination of *E. coli* concentrations moving through the cascading dams during the summer and fall of 2018 (ISCO, Teledyne, Lincoln, NE, USA). A sampler was installed at the DW, the first grade control structure (GCS1) and the second grade control structure (GCS2) from March 2018 through October 2018. The sampler installation included a pressure sensor for recording water depth every five minutes. To supplement the ISCO depth measurements, additional HOBO U20L water level loggers (Onset HOBO, Bourne, MA, USA) were installed at DW, GCS1 and GCS2 to record water depth every fifteen minutes. Throughout the study period, the flowrate at DW was taken directly from the recorded flowmeter from the treatment well pump. Flow rates at GCS1 and GCS2 were calculated using the Kindsvater-Carter equation for suppressed rectangular, sharp-crested weir:

$$Q = \left[\left(0.4000\left(\frac{H}{P}\right) + 3.220\right)(L - 0.003)(H + 0.003)^{\frac{3}{2}}\right] \times 0.028316847 \quad (1)$$

where Q = flowrate (m^3/s), H = water level (ft), P = height of the weir (ft) and L = length of the weir crest (ft).

2.3. Water quality monitoring and analysis

Water temperature at the sampling stations were measured with HOBO U20L loggers since *E. coli* survival rates vary based on water temperature (Blaustein et al., 2013; Jamieson et al., 2004). An ISCO sampler and rain gauge were installed at each study site. Each sampler was configured to activate sampling based on rainfall for six storm events during the study period. Water sampling was set to begin immediately after the rainfall rate reached 0.254 cm hr^{-1}. In order to catch the first flush of *E. coli* through the weirs at GCS1 and GCS2, the first six samples were taken at a 30-minute interval, while the remaining six samples were taken at a rate of 1 sample/h to measure *E. coli* concentrations once flow returned to baseflow. A total of 84 samples were collected at each site: 72 samples from six storm events and an additional 12 samples from a non-storm event.

To ensure accurate determination of bacteria levels, the samples were collected and analyzed within 24 h of each rainfall event. *E. coli* concentrations were determined with the IDEXX Colilert® reagent and 97-well Quanti-Tray®/2000 analysis. This method provides results within 18 h, instead of 48–72 h in previous analytical methods (Sartory and Vandevenne, 2009).

2.4. Pasture management and cattle grazing

In order to manage pasture forage at the USMARC facility, cattle were rotated on 790 individual pastures. Detailed grazing records used for this study included daily information on forage type, number and type of cattle grazing, and the number of days each group stayed in each pasture for the entire study period. Given most contamination was typically observed in streams with unrestricted cattle access (Bragina et al., 2017), pasture locations and grazing dates were used to identify potential cattle interactions with the stream during the studied storm events.

The number of pastures was narrowed down to include only those pastures that drained into the cascading dams and/or were within 50 m of the streams (Fig. 3). This proximity limit was based on the assumption that a higher likelihood of *E. coli* delivery and contamination occurred when cattle were in close proximity to the stream (Berry et al., 2015). Of all the pastures within 50 m of the streams, forty-one pastures drained into the stream above GCS1 while forty-six pastures drained into the stream above GCS2.

To account for the difference in animal weights, the number of Animal Units (AUs) for each pasture was determined based on the number of head of grazing cattle present. AUs are used as a basis for standardizing and expressing stocking rates based on metabolic bodyweight and development, with one AU defined as one 454-kg cow with or without

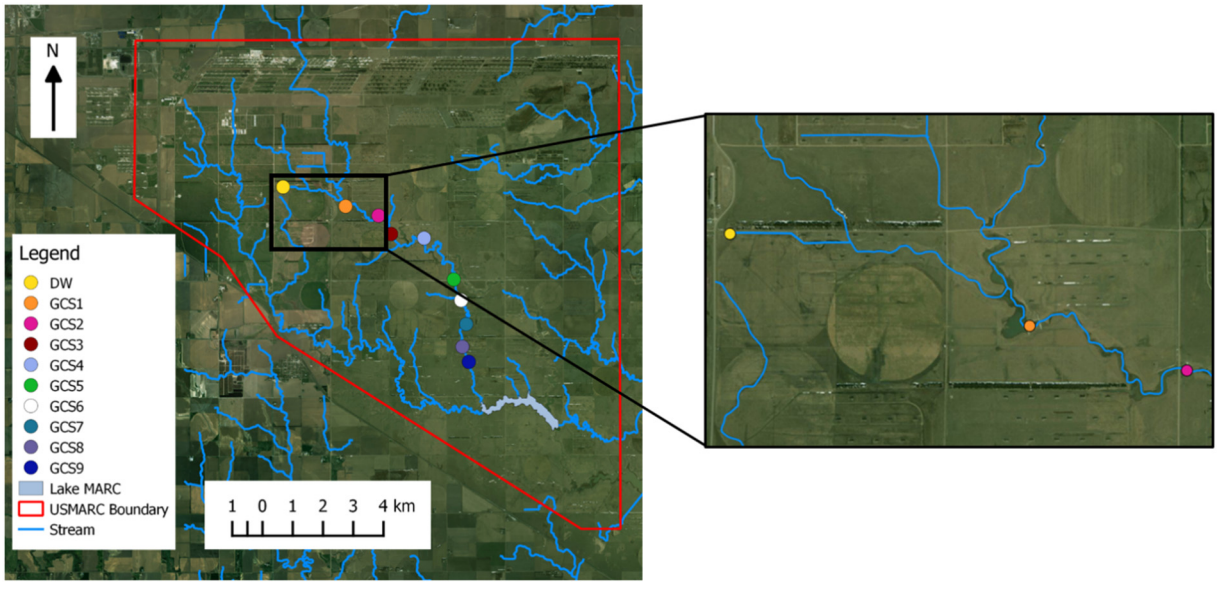


Fig. 1. The study site within the U.S. Meat Animal Research Center in Nebraska, USA. GCS represents grade control structure. DW = Discharge Well; GCS = Grade Control Structure.

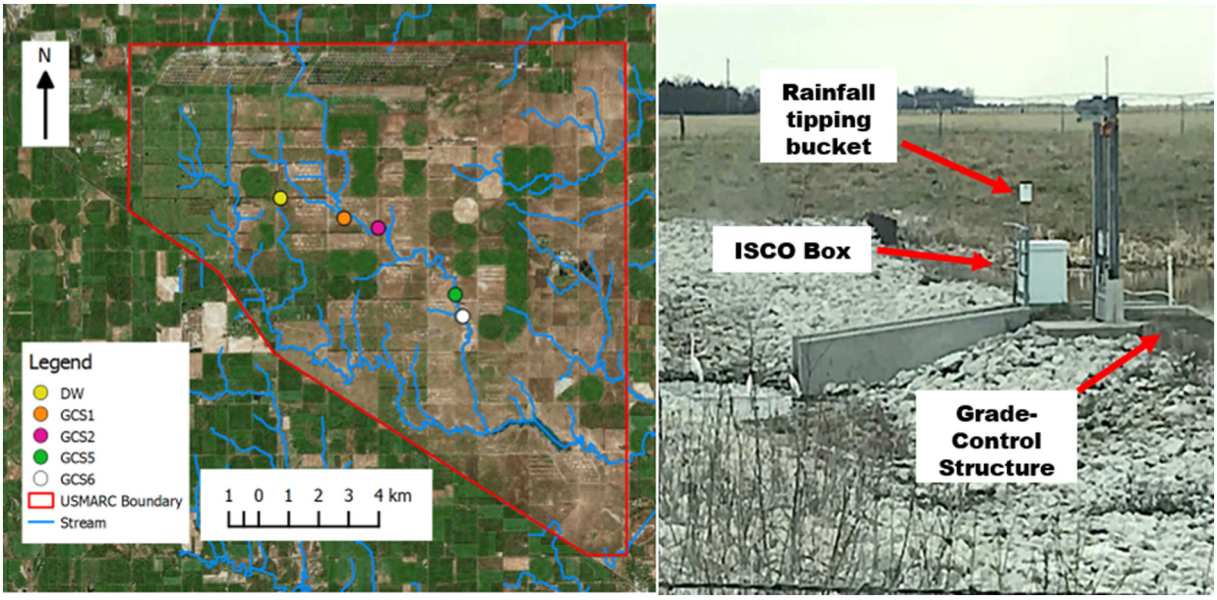


Fig. 2. USMARC facility with (*left*) GCS locations having stop logs; (*right*) ISCO 6712 water sampler setup at GCS1 and GCS2. DW = Discharge Well; GCS = Grade Control Structure (Taken from Hansen et al., 2020).

her unweaned calf. They also help normalize other factors that are related to the number of head of grazing cattle (Manske, 1998).

2.5. Dimensional reduction and feature selection

Identifying the sources of *E. coli* contamination of dams in an agricultural area requires a clear understanding of the influence of the various variables that influence *E. coli* fate once it enters the waterways. However in machine learning, as the dimensionality (number of variables or features) of the data increases, the amount of data required to make reliable and accurate predictions increases exponentially (Hira and Gillies, 2015). A common approach to the problem of high-dimensional datasets is "reduction of dimensionality". This means simplifying the understanding of data by searching for a projection of the data onto a smaller number of predictor variables (or features) which preserves the information as much as possible. Large datasets with the "large *p*, small *n*" problem (where *p* is the number of features and *n* is the number of samples) are susceptible to overfitting. An overfitted model often mistakes small fluctuations for important variance in the data, which may lead to prediction errors (Lever et al., 2016). Our study is typical of this type of small sample problem, where only six storm events were captured and each data point (water sample) had many features. To overcome this problem in *E. coli* prediction, it was important to find a method to reduce the number of features considered for the model.

Principal component analysis (PCA) (Abdi and Williams, 2010; Razmkhah et al., 2010; Zhang et al., 2012) and feature selection (Seo et al., 2014; Asghari and Nasseri, 2014; Hira and Gillies, 2015) are two techniques often used to minimize the number of features used in predictive models. PCA reduces the dimensionality of data while retaining most of the variation in the dataset (Ringnér, 2008; Jolliffe, 2002). Depending on the selection method, feature selection adds features that are significantly important or removes features that are redundant.

Although PCA makes the direct visualization of high dimensional datasets possible since humans can only comprehend three dimensions, it also makes the dataset difficult to interpret as it only outputs linear combinations of the features. Thus, the strength of PCA in giving visual representation of the dominant patterns in a dataset was coupled with feature selection in this study.

2.5.1. Hypothesizing based on prior knowledge

Developing a simple model for a complex system requires prior knowledge and understanding of the processes and features (variables) controlling the system. To select the "most important" features, we first hypothesized that *E. coli* concentration at the outlet of a dam was a

Table 1
Status of stop log installation at each grade control structure on the indicated date during the study.

GCS no./date	5-23-2018	9-4-2018	9-5-2018	11-8-2018	11-13-2018
GCS1	2	4	4	4	3
GCS2	2	2	4	4	3

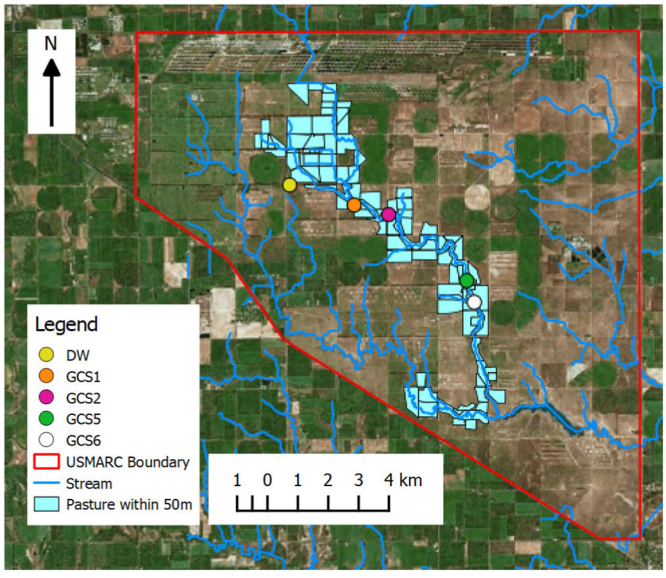


Fig. 3. USMARC grade-control structures (GCS) and pastures within 50 m of stream system.

function of recent (due to runoff, fecal inputs and stream sediments) and past (due to resuspension of *E. coli* stored in dam sediments) storm events, as well as past AUs (within the proximity limit), flow rate and temperature. This hypothesis was based on studies that have attempted to determine the influence of hydrometeorology and cattle grazing practices on *E. coli* concentrations within watersheds (Wagner et al., 2012; Derlet et al., 2012; Hansen et al., 2020; Larsen et al., 1994; Hancock et al., 1994). Although a small fraction of the *E. coli* in fecal material may remain viable for a grazing season or longer at a site (Buckhouse and Gifford, 1976), there is still a potential for contamination long after the cattle have been rotated from the site (Larsen et al., 1994). However, before contamination can be measured at a dam outlet, bacteria in fecal material have to reach a stream (upstream of the dam) by either direct deposit or by overland transport in surface runoff events. Larsen et al. (1994) observed that the contamination of surface waters from *E. coli* and other fecal bacteria depended on the size and number of cattle, distance of the cattle and their fecal deposits from water bodies, characteristics of the fecal deposition site, and the viability of bacteria from the time of deposition to surface runoff events. In a recent study at USMARC, Hansen et al. (2020) found that *E. coli* concentrations had a strong correlation with increasing accumulation of cattle (i.e. by adding the total number of cattle within each pasture for each day) on the pastures throughout the grazing season. Similar to previous studies (Wagner et al., 2012; Derlet et al., 2012) focusing on cattle, the study by Hansen et al. (2020) found a strong correlation when cattle were present on pastures adjacent to the stream on the day of rainfall events.

2.5.2. Selecting the "most important" features

Although there are many features that control *E. coli* fate at a dam outlet, there is a need to avoid over-parameterization when developing predictive models. In order to extract the most important information from the features, PCA was first used to analyze these features, which were inter-correlated in general. The goal was to express them as a set of new orthogonal variables (principal components) that allow visual assessment of similarities and differences between samples and determine whether samples can be grouped by displaying them as points in maps. Using a few components, each sample can be represented by relatively few "most important" features instead of by values for many features.

After PCA analysis, the "most important" features were selected based on the statistical dependence of the log-transformed *E. coli* concentration on all potential features. Forward stepwise selection was chosen because it is a widely used feature selection method based on sequential forward selection (Ruan et al., 2019; Ouali et al., 2017). It involves starting with no features in the model, testing the addition of each feature using a chosen model fit criterion (e.g. residual sum of squares, Akaike Information Criterion), adding the feature (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.

A description of all the features (independent variables) and target (dependent variable) prior to feature selection and model development is shown in Table 2. Although weighted averages of rain gauges were used at GCS1 and GCS2, rainfall values were also weighted on a 3-day basis such that the cumulative rainfall 1 day before the sampling time accounted for 20%, the cumulative rainfall between 1 day and 2 days to the sampling time accounted for 60%, and the cumulative rainfall between 2 days and 3 days to the sampling time accounted for 20% (see *Wtdrain* on Table 2).

2.6. Statistical analysis and model development

E. coli sample data were $\log_{10}$-transformed before developing the machine learning models since concentration values ranged over three orders of magnitude. In addition, although features related to rainfall and AU were numerical, but because they were not continuous, they were converted into categorical features using three or four bins.

Machine learning algorithms such as multiple linear regression, regression trees, decision tree ensembles, support vector regression, Gaussian process regression, and artificial neural network (ANN) were used to analyze datasets in MATLAB 2019b. Adaptive neuro-fuzzy inference system (ANFIS) models were also developed using two clustering methods (subtractive and fuzzy c-means).

The multiple linear regression (MLR) attempts to model the relationship between two or more features and a target by fitting a linear equation to observed data. Every value of a feature is associated with a value of the target. An MLR equation with k features (predictor variables) $X_1, X_2, \ldots, X_k$ and a target (dependent variable) Y', can be written as follows:

$$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (2)$$

where Y' is the estimated target, β_0 is the intercept which is a constant value, and β_i $(i = 1, 2, \ldots, k)$ are the regression coefficients which assign the effects of the features X_i on the target. For MLR, we used four model types: regular (features only), interaction, robust and stepwise linear models. A regular MLR fits a linear equation using Eq. (2). An interaction MLR model includes features and the two-way interaction between them, while a robust linear model returns a $(p + 1)$-by-1 vector β of coefficient estimates for a robust MLR. By default in MATLAB, the algorithm uses iteratively reweighted least squares with a bisquare weighting function.

A regression tree (RT) builds regression models in the form of a tree structure where each internal node of the tree represents a test of one of the features used for prediction. The topmost node in a tree which corresponds to the best feature is called root node. RT tests whether the value of a numeric feature is less than or greater than a threshold value stored at the node, or whether the value of a Boolean feature is true. It breaks down a dataset into smaller subsets such that there is a corresponding or associated subtree for each possible test outcome. Each leaf node in the tree stores the values that satisfy all the tests

Table 2
Description of the features and target used in the development of the models.

	Variable	Description of variables
Features	Atemp	Air temperature on the sampling time (°C)
	Wtemp	Water temperature on the sampling time (°C)
	Wdepth	Water depth at the outlet on the sampling time (m)
	Flow	Discharge through the dam weir (m^3/s)
	RainDay1	Cumulative rainfall 1 day to the sampling time (mm)
	RainDay2	Cumulative rainfall between 1 day and 2 days to the sampling time (mm)
	RainDay3	Cumulative rainfall between 2 days and 3 days to the sampling time (mm)
	CumRain2	Cumulative rainfall 2 days to the sampling time (mm)
	CumRain3	Cumulative rainfall 3 days to the sampling time (mm)
	Wtdrain	Weighted[a] rainfall on 3-day basis (mm)
	AU0sum	Total AUs within 50 m of the stream on the day of sampling (AU)
	AU0dens[b]	Animal density within 50 m of the stream on the day of sampling (AU/ha)
	AU1sum	Total AUs within 50 m of the stream 1 day before the sampling day (AU)
	AU1dens	Animal density within 50 m of the stream 1 day before the sampling day (AU/ha)
	AU2sum	Total AUs within 50 m of the stream 2 days before the sampling day (AU)
	AU2dens	Animal density within 50 m of the stream 2 days before the sampling day (AU/ha)
Target	*E. coli*	*Escherichia coli* concentration (MPN[c]/100 mL)

[a] (0.2* RainDay1) + (0.6* RainDay2) + (0.2* RainDay3).
[b] Animal density was calculated by dividing Total AUs within 50 m of the stream by the sum of pasture hectares.
[c] MPN, most probable number.

between the root node and that leaf node. The RT prediction algorithm navigates the tree structure by applying the node tests to the features, starting with the test at the root node, and continuing on to the subtree selected by the test (Dale et al., 2010). For RT modeling in this study, a fine tree, a medium tree and a coarse tree with minimum leaf sizes of 4, 12 and 36 respectively were used according to MATLAB settings.

In addition to using individual RT algorithms, we investigated decision tree ensembles (DTE) for *E. coli* prediction. The DTE is a method that functions by combining many RTs to produce better predictive performance than using a single RT. The main principle behind the DTE model is that a group of weak RTs are combined to form a strong model. Two ensemble techniques were used in this study: bagged trees and boosted trees. Whereas in bagged trees, the prediction made by an ensemble is obtained by combining the predictions made by individual RTs (taking bootstrap samples of dataset with replacement) using averaging, on the other hand, boosted trees use all the data to train each RT but with weights assigned in order to take a weighted average of their predictions.

Support vector machine regression (SVR) is a nonparametric technique that relies on kernel functions. Smola and Schölkopf (2004) and Awad and Khanna (2015) provided a detailed description of SVR. Linear, quadratic, cubic, and Gaussian kernel functions were used in this study. The Gaussian process regression (GPR) is also a non-parametric method that uses a measure of similarity between samples (kernel function) to predict the value for an unseen sample from training data. It defines a distribution over functions which can be used for Bayesian regression. Detailed description of GPR was provided by Rasmussen (2004). For GPR, exponential, squared exponential, rational quadratic, and matern 5/2 kernel functions were used in this study.

ANNs are mathematical models consisting of a network of computation nodes called neurons with established connections between them (Sattari et al., 2017). An advantage of ANN is that it does not require any a priori assumptions about the relationships between features and targets as well as the functions to be used (Wu et al., 2013). For ANN in this study, one hidden layer with both five and ten neurons was tested. An alternative method to ANN is fuzzy logic which can generate models by integrating expert knowledge and available measurements for a system by using a set of easily understandable rules in the form of a fuzzy inference system (FIS) (Zadeh, 1965). ANFIS is one of the most successful methods which integrates fuzzy logic and ANN to give better performance of predictive models especially when dealing with complex systems (Sattari et al., 2017; Rudnick et al., 2015; Naderloo et al., 2012). Five separate layers are used to describe an ANFIS model structure, and it usually requires division of features and target data into rule patches (Guillaume, 2001). The first layer is the fuzzification layer; the second layer is the rule base layer; the third layer is for normalizing the membership functions; the fourth and fifth layers are the defuzzification and summation layers, respectively (Jang, 1993).

A number of clustering methods such as fuzzy c-means (FCM) (Bezdek, 1981), subtractive clustering (Yager and Filev, 1994), and grid partitioning (Giotis and Giannakoglou, 1998) can be used to get membership functions when creating a FIS. These clustering methods allow the grouping of features into groups with each group having similar properties that help to discern the correlation between the data thus simplifying the prediction process (Benmouiza and Cheknane, 2018). For each clustering method, two different FIS models (Mamdani-type FIS and Sugeno-type FIS) have been developed (Nayak et al., 2013). In order to obtain a small number of fuzzy rules due to the relatively small sample size in this study, ANFIS with subtractive clustering (radii of influence of 0.4 and 0.8) and FCM clustering were applied in this study using MATLAB (MathWorks Inc. Product 2018a).

2.6.1. Model performance evaluation

To develop the predictive models, the dataset was randomly divided into a training dataset (80% of the total data) and a test dataset (20% of the total data). With five-fold cross-validation, four folds (80%) were used for training and the last fold (20%) was used for testing. For one run, this process was repeated five times, leaving one different fold for evaluation each time. For the results to be valid, the performance of each model was averaged on thirty runs. The coefficient of determination (R^2) and root mean squared log error (*RMSLE*) statistics were used for comparing the performance of the different algorithms (Eqs. (3) and (4), respectively). *RMSLE* was chosen instead of the commonly used root mean squared error (*RMSE*) since the *E. coli* concentrations were log-transformed due to the presence of high concentration values. These outliers can increase the error to a very high value. *RMSE* value increases in magnitude if the scale of error increases, whereas *RMSLE* only considers the relative error between predicted and actual values, and the scale of the error is nullified by the log-transformation.

Furthermore, *RMSLE* penalizes underestimation more than overestimation. This is especially useful in our study where the underestimation of the target variable (*E. coli* concentration) is not acceptable but overestimation can be tolerated. For example, if our predictive models overestimate *E. coli* concentration, a water-quality monitoring manager can quickly provide timely information for making a same-day dam or grazing notification decision, and this slight overestimation is acceptable. However, the problem arises when the predicted *E. coli* concentration is less than the actual concentration. In this case, the manager is more

Table 3
Descriptive statistics of all features and target at GCS1 and GCS2.

Variable	Unit	GCS1				GCS2			
		Min	Max	Mean	Standard deviation	Min	Max	Mean	Standard deviation
Atemp	°C	4.4	24.3	12.3	6.4	4.4	24.4	11.6	6.5
Wtemp	°C	9.4	27.1	16.2	6.2	8.5	27.4	15.9	6.9
Wdepth	m	0.7	1.2	1.0	0.2	0.8	1.2	1.0	0.2
Flow	m^3/s	0.0	0.2	0.1	0.1	0.0	0.2	0.1	0.1
RainDay1	mm	1.7	40.7	11.2	12.1	1.6	43.9	10.3	11.3
RainDay2	mm	0.0	40.7	6.7	10.6	0.0	43.9	6.8	11.5
RainDay3	mm	0.0	26.7	4.9	9.7	0.0	26.3	5.1	9.8
CumRain2	mm	1.7	57.9	17.9	19.1	1.6	62.6	17.1	19.3
CumRain3	mm	1.7	71.0	22.8	21.7	1.6	76.7	22.3	22.6
Wtdrain	mm	0.3	30.5	7.2	8.3	0.3	32.9	7.2	8.9
AU0sum	AU	0.0	846.8	566.7	301.0	0.0	1112.5	589.6	361.7
AU0dens	AU/ha	0.0	18.2	8.5	5.1	0.0	18.2	8.4	5.3
AU1sum	AU	0.0	1090.7	671.9	368.3	0.0	1112.5	726.6	382.2
AU1dens	AU/ha	0.0	18.2	10.3	5.9	0.0	18.2	10.2	5.6
AU2sum	AU	0.0	1109.9	635.1	402.9	0.0	1401.4	722.5	440.0
AU2dens	AU/ha	0.0	18.2	9.4	5.9	0.0	18.2	9.9	5.7
Log_{10}*E. coli*	Log(MPN/100 mL)	0.7	3.4	1.8	0.8	0.4	3.4	1.8	1.0

likely to assume all is fine, and as a result, the problem will go uncorrected.

$$R^2 = \frac{\sum_i (\hat{x}_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \tag{3}$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} [\log(x_i + 1) - \log(\hat{x}_i + 1)]^2} \tag{4}$$

where x_i is the observed *E. coli* concentration (most probable number (MPN)/100 mL), $\hat{x}_i$ is the predicted *E. coli* concentration (MPN/100 mL), $\overline{x}$ is the mean of the observed *E. coli* concentration (MPN/100 mL), and n is the total number of samples considered.

2.6.2. Statistical significance testing

Since this study compared different machine learning algorithms on a single domain, paired *t*-tests were conducted to determine if the RMSE were significantly different. This was an important step because the paired t-tests helped us understand the degree to which the RMSE results represent the general behavior of the algorithms. A summary of model evaluation and the description of the paired t-test can be found in Japkowicz and Shah (2011). To check the validity of the results, the performance of each algorithm was averaged based on thirty runs.

3. Results and discussion

For the six storm events used in this study, all except two samples at the DW were below the detection limit and were treated as 0.5 MPN/100 mL. These two samples fell between 1 and 2 MPN/100 mL thus

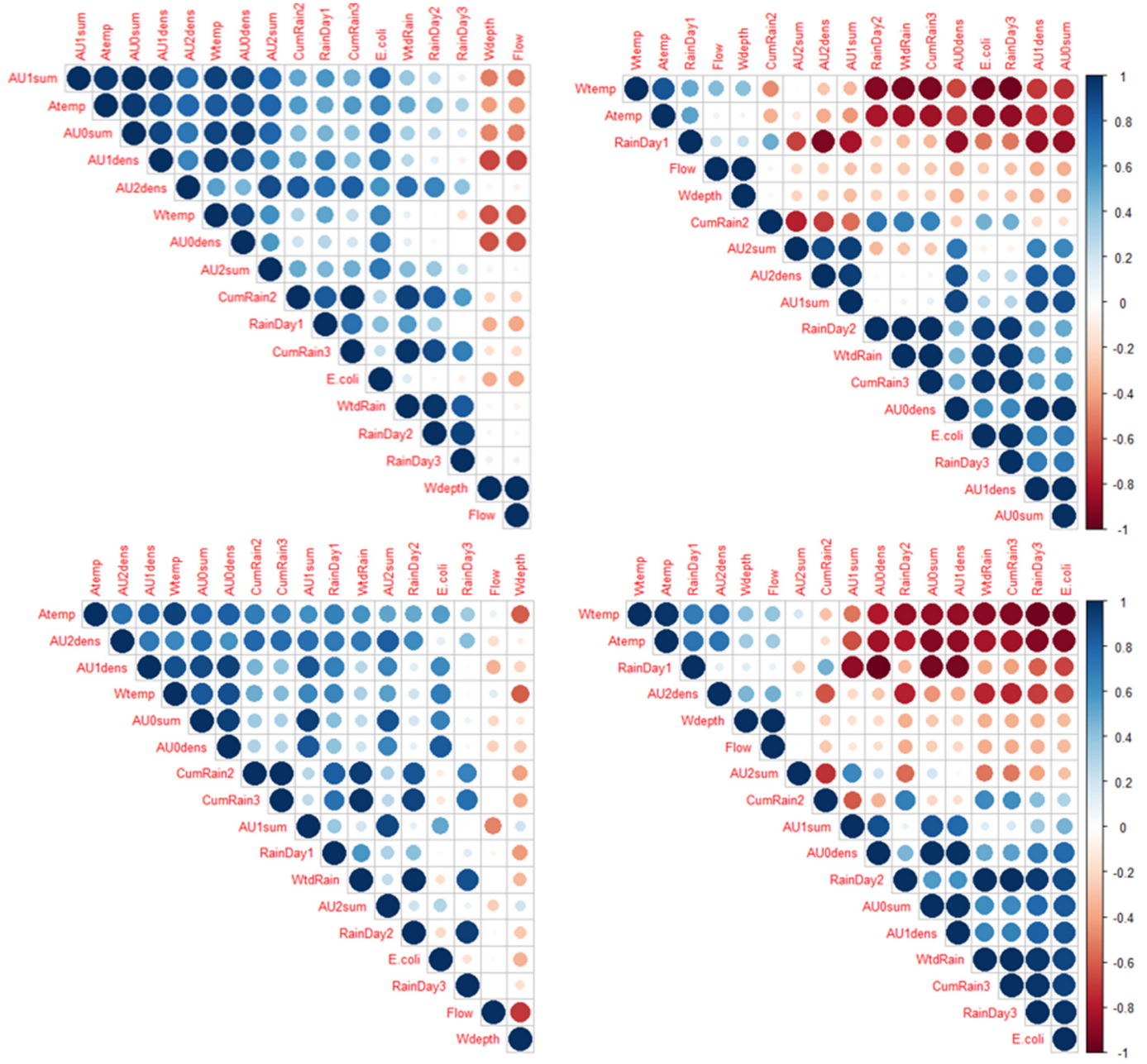


Fig. 4. Correlation matrix for (*top*) GCS1 and (*bottom*) GCS2; (*left*) with two stop logs in, and (*right*) with four stop logs in. Features and target are arranged according to first principal component.

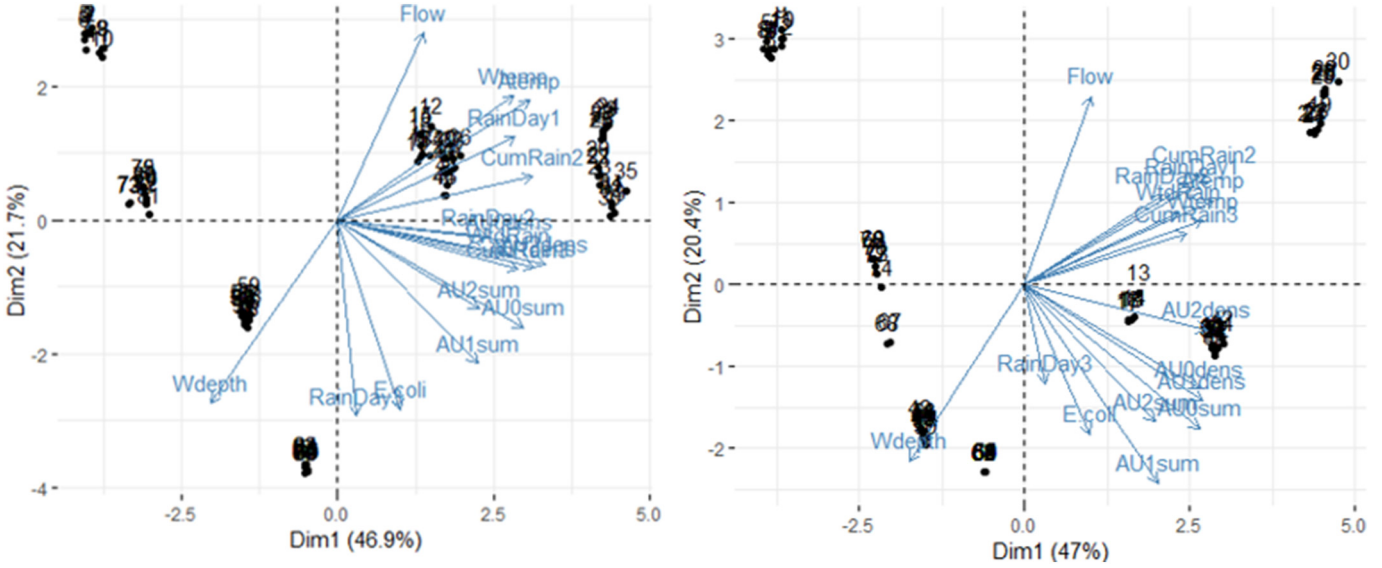


Fig. 5. PCA-Biplots of features with aggregated datasets for (*left*) GCS1 and (*right*) GCS2. PC1 (Dim1) and PC2 (Dim2) are the principal components along x-axis and y-axis respectively.

indicating that DW rarely recorded any detectable *E. coli* because the treated groundwater was its only source of water. While most of the samples fell within the countable *E. coli* range at GCS1 and GCS2, 8.5% and 19.2% of the samples were above the detection limit respectively and treated as the maximum countable 2419.6 MPN/100 mL.

3.1. Descriptive statistics

Descriptive statistics of the features and target are shown in Table 3. The log-transformed mean and maximum *E. coli* concentrations at GCS1 and GCS2 were the same. The log-transformed minimum *E. coli* concentration at GCS1 was slightly higher than that of GCS2. Except for the target and the features related to pasture management (AU sum and density), the remaining features have similar distributions for both GCS1 and GCS2.

Fig. 4 shows the magnitude of coefficient of correlation (r) among the features and target studied at GCS1 and GCS2 (when two and four stop logs were put in), with features/target arranged according to first principal component (PC1). At GCS1, when two stop logs were installed, the *E. coli* concentrations had stronger positive correlations ($r > 0.60$) with AU and temperature features than with rainfall features at p-value <0.05. Conversely, flow and water depth showed negative correlation with *E. coli* concentrations with two stop logs installed. When four stop logs were installed, *E. coli* concentrations was positively correlated with AU and rainfall features while negatively correlated with flow, water depth and temperature features.

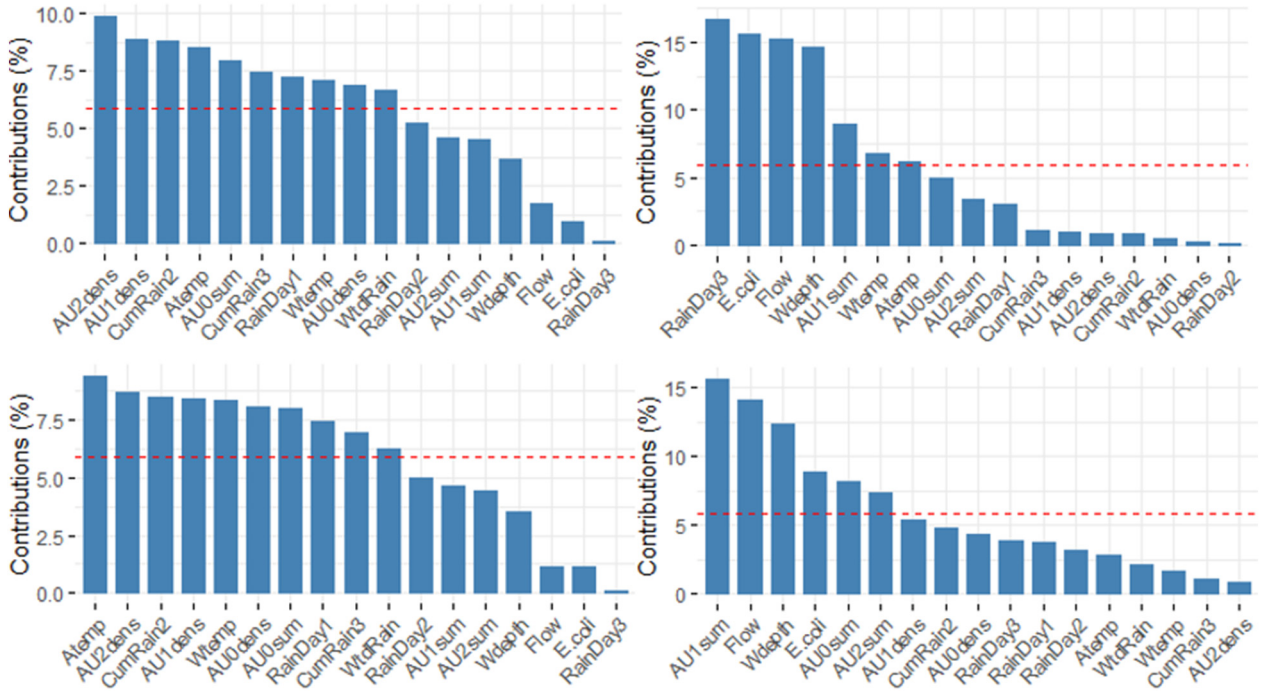


Fig. 6. Contribution of features to (*left*) PC1, and (*right*) PC2 with aggregated datasets at (*top*) GCS1 and (*bottom*) GCS2. The reference red dashed lines correspond to the expected value (5.88%) if the contributions were uniform.

Table 4
Most important features for predicting *E. coli*, ordered by information gain based on forward selection.

GCS1			GCS2		
Variable	RSS	AIC value	Variable	RSS	AIC value
RainDay3	29.2	−80.8	AU0sum	46.0	−37.1
AU1dens	18.0	−118.2	AU2sum	19.2	−103.2
RainDay2	12.1	−149.0	Atemp	7.1	−178.8
AU2dens	10.7	−157.3	AU1sum	5.1	−202.7
AU1sum	8.3	−175.5	AU0dens	4.2	−216.2
AU2sum	7.4	−182.8	RainDay3	3.9	−220.7
Wtemp	5.8	−200.8	AU2dens	3.1	−235.7
Wdepth	4.9	−212.9	AU1dens	2.7	−244.0
Flow	4.1	−225.3			
AU0sum	4.0	−225.5			
AU0dens	3.6	−231.9			

Similarly, at GCS2 when two stop logs were installed, *E. coli* concentrations had a strong, positive correlation with AU and temperature features at p-value <0.05. Water depth was the only feature that was negatively correlated with *E. coli* concentration with two stop logs. When four stop logs were installed, *E. coli* concentrations resulted in a strong negative correlation with temperature features as well as flow and water depth, whereas in general, *E. coli* concentrations showed strong positive correlation with most rainfall and AU features.

With aggregated datasets for both two and four stop logs, PCA-biplots were constructed for both GCS1 and GCS2 (Fig. 5). PC1 is labeled as Dim 1 while Dim 2 is the second principal component (PC2). At GCS1 and GCS2, the first two principal components explain 68.6% and 67.4% of total variations respectively. The seven clusters show the samples collected during the six storm events and those collected in March 2018 (at the beginning of this study) before it started raining in the spring. The contributions of the features and target to the first two principal components are shown in the Scree plots with a reference dashed line, which corresponds to the expected value (5.9%) if the contributions of the seventeen features and target were uniform (Fig. 6). At both GCS1 and GCS2, the first ten features that contribute most to PC1 were the same, although not in the same order of contribution. At GCS1, RainDay3 had the highest loading on the PC2, whereas at GCS2, AU1sum had the highest loading on PC2. The common major contributors to PC2 at both locations were *E. coli*, Flow, Wdepth, and AU1sum.

3.2. Feature performance

Table 4 presents the results of the forward selection of the "most important" features based on the residual sum of squares (RSS) and Akaike information criterion (AIC). Based on the RSS and AIC values only, eleven features were selected as input variables to predict *E. coli* at GCS1, while eight features were chosen to predict *E. coli* at GCS2. Except for Atemp that was selected at GCS2, as well as RainDay2, Wtemp, Wdepth, and Flow that were selected at GCS1, the same seven features were common to both GCSs. However, combining the results of forward selection as well as the contributions of the features to both PC1 and PC2, twelve features from the union of the two "feature sets" were eventually selected as input variables for predicting *E. coli* concentration at both GCS1 and GCS2.

3.3. Model performance

Table 5 shows the performance of the "best subset models" for each machine-learning algorithm at GCS1 and GCS2. For both locations, we tested the effect of using the aforementioned algorithms (five regression types, ANN and ANFIS) for training and testing while varying the components of each algorithm and using the twelve selected features as input variables. Of all the five regression algorithms, the MLR model had the best performance for GCS1 ($RMSLE = 0.21$) while the SVM model had the best performance for GCS2 ($RMSLE = 0.22$). For ANN, we varied the number of hidden neurons starting with five, and then ten. We found that the performance of using either five or ten neurons was almost the same and there was no improvement in model performance when compared to the best regression models for each GCS (Table 5). For ANFIS, although the number of epochs was not as important as the prediction error, 5, 10, 20, 50, and 500 epochs were tried for both subtractive and FCM clustering methods in order to avoid overfitting. It was observed that 10 epochs was sufficient as higher epochs did not significantly increase model performance.

For subtractive clustering method, we again varied the radius of influence, starting with 0.8, and then 0.4. There was no significant difference between subtractive clustering and previous models when both radii were used at GCS1. At GCS2, we found lower performance of subtractive clustering irrespective of the radius. For FCM clustering method, we tested five and seven rules and found that seven rules performed relatively better than five rules at GCS1 while the converse was true at GCS2 (Table 5). On the average, we found a significant improvement in performance with ANFIS FCM algorithms, with up to approximately 12% and 36% reductions in error for GCS1 and GCS2 respectively.

Comparison of all algorithms showed that better *E. coli* concentration predictions were obtained at both locations using ANFIS than regression models and ANN. Although ANFIS and ANN algorithms are both based on neural networks, one of the major limitations of ANN is its lack of explanatory power, often referred to as the "black box problem" (Dastorani et al., 2010). ANFIS eliminates some of these limitations by integrating both neural networks and fuzzy logic principles. The superiority of ANFIS over ANN modeling approach has been well established by Nayak et al. (2004), Dastorani et al. (2010), Talebizadeh and Moridnejad (2011), Emamgholizadeh et al. (2014), and Luo et al.

Table 5
Comparison of "best subset models" during model training and testing phases at GCS1 and GCS2.

Model	Components	GCS1			GCS2		
		R^2	RMSLE (logMPN/100 mL)		R^2	RMSLE (logMPN/100 mL)	
		Training	Training	Testing	Training	Training	Testing
MLR	Linear	0.93	0.20	0.21	0.95	0.22	0.24
DT	Fine tree; minimum leaf size = 4	0.90	0.24	0.26	0.94	0.25	0.28
DTE	Boosted trees; minimum leaf size = 8; number of learners = 30	0.90	0.24	0.25	0.93	0.26	0.27
SVM	Kernel function = Gaussian; Kernel scale = 0.87	0.93	0.21	0.22	0.96	0.21	0.22
GPR	Kernel function = exponential; basis function = constant	0.94	0.19	0.23	0.96	0.21	0.24
ANN	Number of hidden neurons = 5; Levenberg-Marquardt fitting	0.97	0.18	0.22	0.97	0.22	0.22
ANN	Number of hidden neurons = 10; Levenberg-Marquardt fitting	0.97	0.18	0.20	0.98	0.21	0.21
ANFIS subtractive	Number of epochs = 10, radius = 0.8; FIS type = Sugeno	0.99	0.09	0.20	0.98	0.22	0.31
ANFIS subtractive	Number of epochs = 10, radius = 0.4; FIS type = Sugeno	0.99	0.10	0.22	0.99	0.16	0.31
ANFIS FCM	Number of epochs = 10; number of rules = 5; FIS type = Sugeno	0.98	0.15	0.18	0.98	0.17	0.16
ANFIS FCM	Number of epochs = 10; number of rules = 7; FIS type = Sugeno	0.99	0.11	0.18	0.99	0.16	0.15

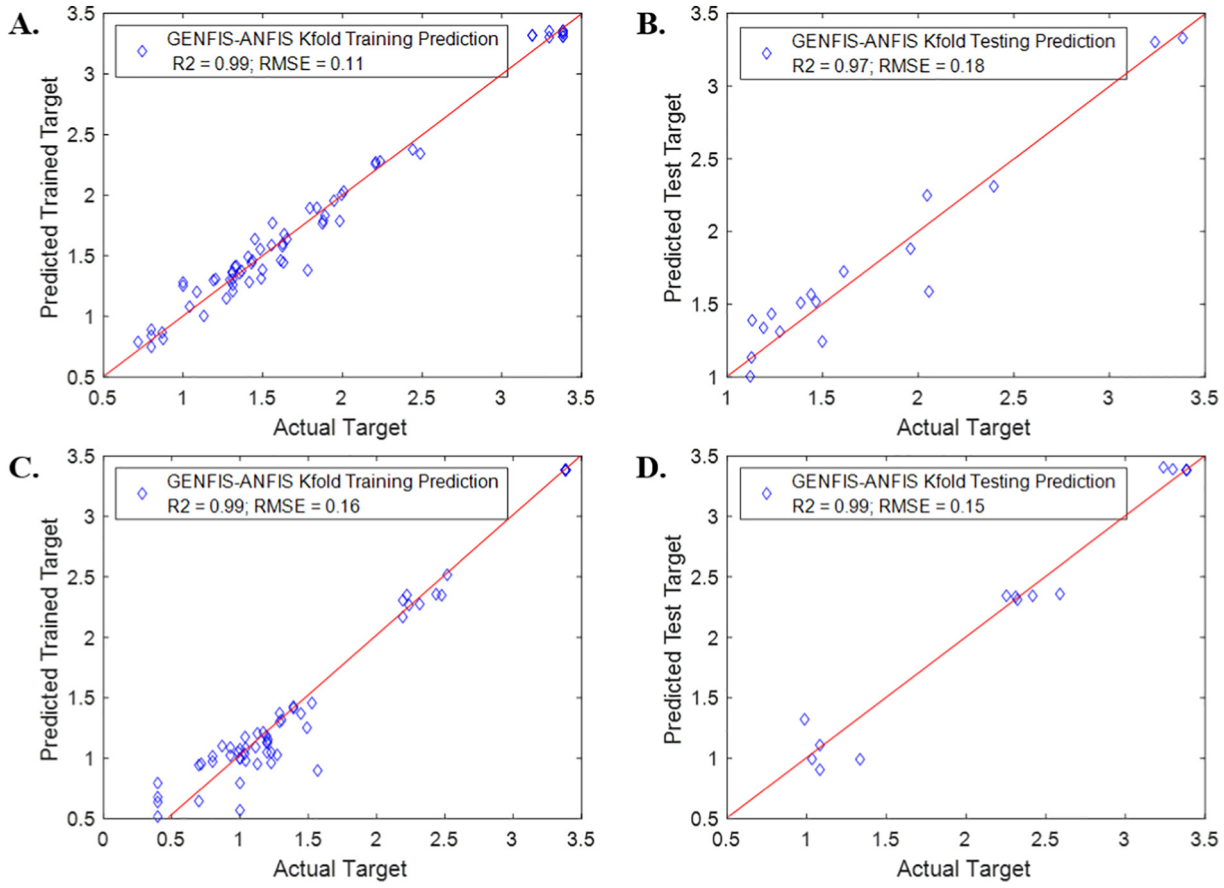


Fig. 7. Actual versus predicted *E. coli* concentrations (log cfu/100 mL) at GCS1 (A,B) and GCS2 (C,D) using ANFIS FCM clustering with 7 rules; (*left*) training set, and (*right*) testing set.

(2019) in various fields of ecohydrology. Scatter plots between actual and predicted *E. coli* concentrations at GCS1 and GCS2 using ANFIS with subtractive clustering are shown in Fig. 7.

4. Conclusions

We have demonstrated the application of different machine-learning algorithms for *E. coli* concentration prediction at two cascading dams (GCS1 and GCS2). A major finding of this study was the integration of hydrometeorology, animal density, and grazing pattern in a unique way to extract and select the most important features used for developing and validating the models. These features included those that were newly developed in this work, which are less explanatory individually, but can contribute to *E. coli* prediction accuracy and performance. We observed that only twelve out of the sixteen features carry most of the information for predicting *E. coli* concentration. Specifically, the number of animals close to the streams, grazing density and cumulative rainfall between two and three days to the sampling time were the most informative features. The integration of features provides an important foundation for future work on *E. coli* prediction at the nine cascading GCSs at the USMARC facility, and other dams and surface waters in other areas. Despite the fact that it is almost impossible for any model to account for all the processes and heterogeneity involved in *E. coli* transport in dams, our results show that machine-learning algorithms, provided with good extraction and selection of features, provide potential tools for predicting *E. coli* transport through dams. As more samples are taken at different times of the year during high and low flows (within and outside storm events), and curation of data associated with all the important features is done, the set of available training data will grow. New features can be incorporated and tested in combination with existing features. Further, novel prediction algorithms have the potential to be implemented and tested.

The ANFIS models we have developed provide good estimates of *E. coli* concentrations and have the ability to be modified by the users based on their preferences for accuracy and precision. However, since the models were developed using data for our study area, the level of uncertainty in applying our models or methods to another dam would depend on the knowledge of the study area, data quality, and a thorough understanding of all the processes involved and features used in modeling.

CRediT authorship contribution statement

Olufemi P. Abimbola: Methodology, Formal analysis, Writing - original draft.**Aaron R. Mittelstet:**Conceptualization, Methodology, Funding acquisition, Writing - review & editing, Project administration.**Tiffany L. Messer:**Conceptualization, Funding acquisition, Writing - review & editing.**Elaine D. Berry:**Investigation, Writing - review & editing, Resources.**Shannon L. Bartelt-Hunt:**Writing - review & editing.**Samuel P. Hansen:**Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is based on research that was supported by the Nebraska Agricultural Experiment Station with funding from the State of

Nebraska in collaboration with the Agricultural Research Service, U.S. Meat Animal Research Center, U.S. Department of Agriculture and the U.S. Department of Agriculture - National Institute of Food and Agriculture (Hatch project NEB-21-177). The authors also thank Alan Boldt and Shannon Ostdiek for their technical assistance.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Computational Statistics 2, 433–459.

Arnold, J.G., Fohrer, N., 2005. SWAT2000: current capabilities and research opportunities in applied watershed modelling. Hydrol. Process. 19 (3), 563–572.

Asghari, K., Nasseri, M., 2014. Spatial rainfall prediction using optimal features selection approaches. Hydrol. Res. 46 (3), 343–355. https://doi.org/10.2166/nh.2014.178.

Awad, M., Khanna, R., 2015. Support vector regression. Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_4.

Baffaut, C., Benson, V.W., 2003. A bacterial TMDL for shoal creek using SWAT modeling and DNA source tracking. Total Maximum Daily Load (TMDL) Environmental Regulations-II Proceedings of the Conference. ASAE, St. Joseph, MI ASAE Publication No. 701P1503.

Benham, B.L., Baffaut, C., Zeckoski, R.W., Mankin, K.R., Pachepsky, Y.A., Sadeghi, A.M., Brannan, K.M., Soupir, M.L., Habersack, M.J., 2006. Modeling bacteria fate and transport in watersheds to support TMDLs. Trans. ASABE 49, 987–1002.

Benmouiza, K., Cheknane, A., 2018. Clustered ANFIS network using fuzzy c-means, subtractive clustering, and grid partitioning for hourly solar radiation forecasting. Theor. Appl. Climatol., 1–13 https://doi.org/10.1007/s00704-018-2576-4.

Berry, E.D., Wells, J.E., Bono, J.L., Woodbury, B.L., Kalchayanand, N., Norman, K.N., Suslow, T.V., López-Velasco, G., Millner, P.D., 2015. Effect of proximity to a cattle feedlot on *Escherichia coli* O157:H7 contamination of leafy greens and evaluation of the potential for airborne transmission. Appl. Environ. Microbiol. 81, 1101–1110. https://doi.org/10.1128/AEM.02998-14.

Bezdek, J.C., 1981. Pattern Recognition With Fuzzy Objective Function Algorithms. Springer US, Boston.

Blaustein, R.A., Pachepsky, Y., Hill, R.L., Shelton, D.R., Whelan, G., 2013. *Escherichia coli* survival in waters: temperature dependence. Water Res. 47, 569–578.

Bragina, L., Sherlock, O., van Rossum, A.J., Jennings, E., 2017. Cattle exclusion using fencing reduces *Escherichia coli* (*E. coli*) level in stream sediment reservoirs in northeast Ireland. Agric. Ecosyst. Environ. 239, 349–358. https://doi.org/10.1016/j.agee.2017.01.021.

Brooks, W., Corsi, S., Fienen, M., Carvin, R., 2016. Predicting recreational water quality advisories: a comparison of statistical methods. Environ. Model. Softw. 76, 81–94. https://doi.org/10.1016/j.envsoft.2015.10.012.

Buckhouse, J.C., Gifford, G.E., 1976. Water quality implications of cattle grazing on a semiarid watershed in southeastern Utah. J. Range Manag. 29, 109–113.

Coffey, R., Cummins, E., Cormican, M., Flaherty, V.O., Kelly, S., 2007. Microbial exposure assessment of waterborne pathogens. Hum. Ecol. Risk. Assess. 13, 1313–1351.

Collins, R., Rutherford, K., 2004. Modelling bacterial water quality in streams draining pastoral land. Water Res. 38 (3), 700–712.

Dale, J.M., Popescu, L., Karp, P.D., 2010. Machine learning methods for metabolic pathway prediction. BMC Bioinformatics 11, 15 (http://www.biomedcentral.com/1471-2105/11/15).

Dastorani, M.T., Moghadamnia, A., Piri, J., Rico-Ramirez, M., 2010. Environ. Monit. Assess. 166, 421–434. https://doi.org/10.1007/s10661-009-1012-8.

Derlet, R.W., Richards, J.R., Tanaka, L.L., Hayden, C., Ger, K.A., Goldman, C.R., 2012. Impact of summer cattle grazing on the Sierra Nevada watershed: aquatic algae and bacteria. J. Environ. Public Health 2012, 1–7. https://doi.org/10.1155/2012/760108.

Efting, A.A., Snow, D.D., Fritz, S.C., 2011. Cyanobacteria and microcystin in the Nebraska (USA) Sand Hills Lakes before and after modern agriculture. J. Paleolimnol. 46, 17–27. https://doi.org/10.1007/s10933-011-9511-3.

Emamgholizadeh, S., Moslemi, K., Karami, G., 2014. Prediction the groundwater level of Bastam Plain (Iran) by Artificial Neural Network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). Water Resour. Manag. 28, 5433–5446. https://doi.org/10.1007/s11269-014-0810-0.

Flint, K.P., 1987. The long-term survival of *Escherichia coli* in river water. J. Appl. Bacteriol. 63, 261–270.

Giotis, A.P., Giannakoglou, K.C., 1998. An unstructured grid partitioning method based on genetic algorithms. Adv. Eng. Softw. 29, 129–138.

Gonzalez, R.A., Conn, K.E., Crosswell, J.R., Noble, R.T., 2012. Application of empirical predictive modeling using conventional and alternative fecal indicator bacteria in eastern North Carolina waters. Water Res. 46 (18), 5871–5882. https://doi.org/10.1016/j.watres.2012.07.050.

Guillaume, S., 2001. Designing fuzzy inference systems from data: an interpretability-oriented review. Fuzzy Sys. IEEE Trans. 9, 426–443.

Hancock, D., Besser, T., Kinsel, M., Tarr, P., Rice, D., Paros, M., 1994. The prevalence of *Escherichia coli* O157.H7 in dairy and beef cattle in Washington State. Epidemiol. Infect. 113 (2), 199–207. https://doi.org/10.1017/S0950268800051633.

Hansen, S., Messer, T., Mittelstet, A., Berry, E., Bartelt-Hunt, S., Abimbola, O., 2020. Escherichia coli concentrations in waters of a reservoir system impacted by cattle and migratory waterfowl. Sci. Total Environ. 705, 135607. https://doi.org/10.1016/j.scitotenv.2019.135607.

Hellweger, F.L., Masopust, P., 2008. Investigating the fate and transport of *Escherichia coli* in the Charles River, Boston, using high-resolution observation and modeling. J. Am. Water Resour. Assoc. 44 (2), 509–522.

Hipsey, M.R., Antenucci, J.P., Brookes, J.D., 2008. A generic, process-based model of microbial pollution in aquatic systems. Water Resour. Res. 44, W07408. https://doi.org/10.1029/2007WR006395.

Hira, Z.M., Gillies, D.F., 2015. A review of feature selection and feature extraction methods applied on microarray data. Adv. Bioinforma. 2015, 198363 13 pages. https://doi.org/10.1155/2015/198363.

Jamieson, R., Joy, D.M., Lee, H., Kostaschuk, R., Gordon, R., 2005. Transport and deposition of sediment-associated *Escherichia coli* in natural streams. Water Res. 39 (12), 2665–2675.

Jamieson, R.C., Joy, D.M., Lee, H., Kostaschuk, R., Gordon, R.J., 2004. Persistence of enteric bacteria in alluvial streams. Eng. Sci. 3, 203–212.

Jang, J.S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. Sys. Man. Cybern. IEEE Trans. 23, 665–685.

Japkowicz, N., Shah, M., 2011. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press.

Jin, G., Englande, A.J., Liu, A., 2003. A preliminary study on coastal water quality monitoring and modeling. J. Environ. Sci. Health A38, 493–509.

Jolliffe, I.T., 2002. Principal Component Analysis. 2nd edition. Springer, New York.

Jones, R.M., Liu, L., Dorevitch, S., 2013. Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. Environ. Monit. Assess. 185 (3), 2355–2366. https://doi.org/10.1007/s10661-012-2716-8.

Larsen, R.E., Miner, J.R., Buckhouse, J.C., Moore, J.A., 1994. Water-quality benefits of having cattle manure deposited away from streams. Bioresour. Technol. 48, 113–118. https://doi.org/10.1016/0960-8524(94)90197-X.

Lessard, E.J., Sieburth, J.M., 1983. Survival of natural sewage populations of enteric bacteria in diffusion and batch chambers in the marine-environment. Appl. Environ. Microbiol. 45, 950–959.

Lever, J., Krzywinski, M., Altman, N., 2016. Points of significance: model selection and overfitting. Nat. Methods 13, 703–704.

Luo, W., Zhu, S., Wu, S., Dai, J., 2019. Comparing artificial intelligence techniques for chlorophyll-a prediction in US lakes. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-019-06360-y.

Manske, L.L., 1998. Animal Unit Equivalent for Beef Cattle Based on Metabolic Weight. North Dakota State University Dickinson Research Extension Service, Fargo ND, pp. 1–3.

McCorquodale, J.A., Georgiou, I., Carnelos, S., Englande, A.J., 2004. Modeling coliforms in storm water plumes. J. Environ. Eng. Sci. 3, 419–431.

McKergow, L.A., Davies-Colley, R.J., 2009. Stormflow dynamics and loads of Escherichia coli in a large mixed land use catchment. Hydrol. Process. 24 (3), 276–289. https://doi.org/10.1002/hyp.7480.

Medema, G.J., Schijven, J.F., 2001. Modelling the sewage discharge and dispersion of Cryptosporidium and Giardia in surface water. Water Res. 35, 4307–4316.

Naderloo, L., Alimardani, R., Omid, M., Sarmadian, F., Javadikia, P., Torabi, M.Y., Alimardani, F., 2012. Application of ANFIS to predict crop yield based on different energy inputs. Measurement 45 (6), 1406–1413.

Nayak, G.K., Narayanan, S.J., Paramasivam, I., 2013. Development and comparative analysis of fuzzy inference systems for predicting customer buying behavior. Int. Jour. Eng. Tech. 5 (5), 4093–4108.

Nayak, P., Sudheer, K., Rangan, D., Ramasastri, K., 2004. A neuro-fuzzy computing technique for modeling hydrological time series. J. Hydrol. 291 (1–2), 52–66.

Nevers, B.M., Whitman, R.L., 2005. Nowcast modeling of *Escherichia coli* concentrations at multiple urban beaches of southern Lake Michigan. Water Res. 39 (20), 5250–5260. https://doi.org/10.1016/j.watres.

Nevers, M.B., Whitman, R.L., 2011. Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches. Water Res. 45 (4), 1659–1668. https://doi.org/10.1016/j.watres.2010.12.010.

Noguchi, K., Nakajima, H., Aono, R., 1997. Effects of oxygen and nitrate on growth of *Escherichia coli* and *Pseudomonas aeruginosa* in the presence of organic solvents. Extremophiles 1, 193–198.

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2017. Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. J. Adv. Model. Earth Sy. 9 (2), 1292–1306. https://doi.org/10.1002/2016MS000830.

Pachepsky, Y.A., Shelton, D.R., 2011. *Escherichia coli* and fecal coliforms in freshwater and estuarine sediments. Crit. Rev. Environ. Sci. Technol. 41, 1067–1110. https://doi.org/10.1080/10643380903392718.

Pachepsky, Y.A., Sadeghi, A.M., Bradford, S.A., Shelton, D.R., Guber, A.K., Dao, T.H., 2006. Transport and fate of manure-borne pathogens: modeling perspective. Agric. Water Manag. 86, 81–92.

Park, Y., Kim, M., Pachepsky, Y., Choi, S.H., Cho, J.G., Jeon, J., Cho, K.H., 2018. Development of a nowcasting system using machine learning approaches to predict fecal contamination levels at recreational beaches in Korea. J. Environ. Qual. 47, 1094–1102. https://doi.org/10.2134/jeq2017.11.0425.

Petersen, C.M., Rifai, H.S., Stein, R., 2009. Bacteria load estimator spreadsheet tool for modeling spatial *Escherichia coli* loads to an urban bayou. J. Environ. Eng. 135 (4), 203–217.

Rasmussen, C.E., 2004. Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (Eds.), Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, p. 3176.

Razmkhah, H., Abrishamchi, A., Torkian, A., 2010. Evaluation of spatial and temporal variation in water quality by pattern recognition techniques: a case study on Jajrood River (Tehran, Iran). J. Environ. Manag. 91, 852–860.

Ringnér, M., 2008. What is principal component analysis? Nat. Biotechnol. 26 (3), 303–304.

Robakis, N., Cenatiempo, Y., Meza-Basso, L., Brot, N., Weissbach, H., 1983. A coupled DNA-directed in vitro system to study gene expression based on di- and tripeptide formation. Methods Enzymol. 101, 690–706.

Ruan, X., Huang, J., Williams, D., Harker, K., Gergel, S., 2019. High spatial resolution landscape indicators show promise in explaining water quality in urban streams. Ecol. Indic. 103, 321–330. https://doi.org/10.1016/j.ecolind.2019.03.013.

Rudnick, D.R., Sharma, V., Meyer, G.E., Irmak, S., 2015. Using fuzzy logic to predict and evaluate the magnitude and distribution of precipitation on rainfed maize and soybean yields in Nebraska. Trans. ASABE 58 (5), 1215–1229. https://doi.org/10.13031/trans.58.10831.

Sartory, D.P., Vandevenne, C.A., 2009. Improved Methods for Detecting *E. coli* and Coliforms in Drinking Water: AFNOR Validation of Colilert-18/Quanti-Tray (Alella, Spain).

Sattari, M.T., Dodangeh, E., Abraham, J., 2017. Estimation of daily soil temperature via data mining techniques in semi-arid climate conditions. Earth Sci. Res. J. 21 (2), 85–93.

Seo, J.H., Lee, Y.H., Kim, Y.H., 2014. Feature selection for very short-term heavy rainfall prediction using evolutionary computation. Adv. Meteorol. 2014, 203545 15 pages. https://doi.org/10.1155/2014/203545.

Shively, D.A., Nevers, M.B., Breitenbach, C., Phanikumar, M.S., Przybyla-Kelly, K., Spoljaric, A.M., Whitman, R.L., 2016. Prototypic automated continuous recreational water quality monitoring of nine Chicago beaches. J. Environ. Manag. 166, 285–293. https://doi.org/10.1016/j.jenvman.2015.10.011.

Sjogren, R.E., Gibson, M.J., 1981. Bacterial survival in a dilute environment. Appl. Environ. Microbiol. 41, 1331–1336.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Steets, B.M., Holden, P.A., 2003. A mechanistic model of runoff associated fecal coliform fate and transport through a coastal lagoon. Water Res. 37, 589–608.

Talebizadeh, M., Moridnejad, A., 2011. Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ANN and ANFIS models. Expert Syst. Appl. 38, 4126–4135.

Tian, Y.Q., Gong, P., Radke, J.D., Scarborough, J., 2002. Spatial and temporal modeling of microbial contaminants on grazing farmlands. J. Environ. Qual. 31 (3), 860–869.

Unc, A., Goss, M., 2004. Transport of bacteria from manure and protection of water resources. Appl. Soil Ecol. 25 (1), 1–18. https://doi.org/10.1016/j.apsoil.2003.08.007.

USACE, 2010. Record of decision, sitewide groundwater, former naval ammunition depot, Hastings, Nebraska. URL. https://semspub.epa.gov/work/HQ/189068.pdf (Accessed on 07/01/2019).

Vidon, P., Tedesco, L.P., Wilson, J., Campbell, M.A., Casey, L.R., Gray, M., 2008. Direct and indirect hydrological controls on concentration and loading in midwestern streams. J. Environ. Qual. 37 (5), 1761–1768. https://doi.org/10.2134/jeq2007.0311.

Wagner, K.L., Redmon, L.A., Gentry, T.J., Harmel, R.D., 2012. Assessment of cattle grazing effect on *E. coli* runoff. Trans. ASABE 55, 2111–2122.

Whitman, R.L., Shively, D.A., Pawlik, H., Nevers, M.B., Byappanahalli, M.N., 2003. Occurrence of *Escherichia coli* and enterococci in Cladophora (Chlorophyta) in nearshore water and beach sand of Lake Michigan. Appl. Environ. Microbiol. 69 (8), 4714–4719. https://doi.org/10.1128/AEM.69.8.4714-4719.2003.

Whitman, R.L., Nevers, M.B., Korinek, G.C., Byappanahalli, M.N., 2004. Solar and temporal effects on *Escherichia coli* concentration at a Lake Michigan swimming beach. Appl. Environ. Microbiol. 70, 4276–4285.

Wu, W., Tang, X.P., Guo, N.J., Yang, C., Liu, H.B., Shang, Y.F., 2013. Spatiotemporal modeling of monthly soil temperature using artificial neural networks. Theor. Appl. Climatol. 113, 481–494.

Yager, R.R., Filev, D.P., 1994. Generation of fuzzy rules by mountain clustering. J Intell Fuzzy Syst Appl Eng Technol 2, 209–219.

Zadeh, L.A., 1965. Fuzzy sets. Inf. Control. 8, 338–353.

Zaleski, K.J., Josephson, K.L., Gerba, C.P., Pepper, I.L., 2005. Survival, growth, and regrowth of enteric indicator and pathogenic bacteria in biosolids, compost, soil, and land applied biosolids. J. Residuals Sci. Technol. 2 (1), 49–63.

Zhang, B., Song, X., Zhang, Y., Han, D., Tang, C., Yu, Y., Ma, Y., 2012. Hydrochemical characteristics and water quality assessment of surface water and groundwater in Songnen plain, Northeast China. Water Res. 46, 2737–2748.