Predicting faecal indicator organisms in coastal waters using a complex nonlinear artificial intelligence model

Man-Yue Lam[1] and Reza Ahmadian[2]

[1]School of Engineering, Cardiff University, Cardiff CF24 3AA, UK. Email: lamM7@cardiff.ac.uk
[2]School of Engineering, Cardiff University, Cardiff CF24 3AA

ABSTRACT

High levels of faecal indicator organisms (FIOs) at bathing water sites can cause disease and impose threats to public health. There is a need for predicting FIO levels to inform the public and reduce exposure. Data-driven models are one of the main tools being considered as predictive models. However, identifying the main inputs of the data-driven models is a major challenge in developing FIO predictor models. This paper develops a data-driven model for FIO concentration prediction based on a limited number of critical input variables. Essential variables were identified with be a combination of Gamma test and GA (Gamma-GA-test). Artificial Neural Networks (ANNs) and linear regression models were developed using these two variable identification approaches for comparison. The models were applied to a case study, and it was found that the model using the Gamma-GA test has a high potential to predict FIO levels more accurately, although this requires further investigation with different case studies. A correlation analysis was required prior to the variable identification approaches in this study. The need of this analysis highlights the significance of understanding the waterbody and the data set in the development and application of data-driven models. Models using a Gamma-GA test were more capable of predicting extreme (high) FIO concentrations, making a Gamma-GA test more suitable for a bathing water quality early warning system. The importance of nonlinearity in such predictive models was also demonstrated by the better performance of nonlinear ANN models compared

to linear regression models regardless of the variable identification approaches used. This paper highlights the importance of nonlinearity in bathing water quality prediction and encourages further utilisation of nonlinear models for this application.

INTRODUCTION

Water-borne pathogens in waterbodies cause illnesses such as gastrointestinal infection, eye infection, skin complaints, and nose and throat infections (Pruss 1998;Pandey et al. 2014). Faecal indicator organisms (FIO), e.g. *E Coli* and *Enterococci*, are commonly used to indicate the level of pathogens in waterbodies (Dufour 1984;Pandey et al. 2014). In Europe, the European Union (EU) revised Bathing Water Directive (rBWD) (European Commission 2006) requires member states to monitor at least the concentrations of two FIO species in designated bathing waters for compliance. The rBWD recognises short-term occasional pollution and includes provisions for discounting compliance requirements when there is a predictive and warning system to alarm the public of impending poor water quality. Traditionally, FIO concentrations in bathing water samples are determined by culture-based methods. These methods require a minimum of 18-24 hour (USEPA 2010) laboratory assay. However, FIO concentrations change continuously (Boehm et al. 2002; Whitman et al. 2004; Kim et al. 2004; King et al. 2021), causing culture-based warning systems to give outdated water quality alerts. More rapid FIO analysis methods such as quantitative polymerize chain reaction (qPCR) can determine FIO concentrations in less than 6 hours, but these methods require significant up-front investments and trained personnel (Zhang et al. 2018) and still cannot be used as a predictive tool. While field sampling and analysis are important, they do not provide predictions to impending bathing water quality.

Two- and three-dimensional hydro-environmental models are commonly applied to assess FIO concentrations in waterbodies. These models numerically solve the mass and momentum equations of fluids as well as the fate and transport of FIOs, including decay and interaction with sediment. These models have been applied in a wide range of studies (e.g. Lee and Qu 2004; Lin et al. 2008; Schippmann et al. 2013; Huang et al. 2017; Abu-Bakar et al. 2017) to provide relatively accurate predictions of the spatial and temporal concentration distributions of FIOs. Nevertheless, these models require detailed knowledge of flow and FIOs at the boundary of the modelling domain, which are generally very expensive and time consuming to acquire. Moreover, such models are usually computationally demanding and require a long run time even on modern computers. Therefore, using such models in real-time as a part of early warning systems is not practical.

Data-driven models are promising alternatives in providing timely predictions of FIOs for bathing water quality warning systems due to their lower computational requirements. Such models utilize data obtained by environmental sensors to predict FIO concentrations in bathing waters. The public could then be warned about occasions with high FIO concentrations. The development of such data-driven models requires identifying FIO predictive variables and establishing relationships between these variables and FIO concentrations. These two steps have been previously conducted mainly by stepwise multi-linear regression (MLR) (e.g. Crowther et al. 2001; Nevers and Whitman 2005; Gonzalez et al. 2012; Wyer et al. 2013b; Gonzalez and Noble 2014). In a stepwise MLR, variables are included or excluded in a linear regression equation in a stepwise manner. Such inclusion or exclusion is decided by the influence of the input variables on estimating the target variables through linear regression analysis. This linear approach does not account for possible nonlinear relationships which could affect predicting extremes. A promising nonlinear approach is Artificial Neural Network (ANN), in which the predictive variables and FIO concentrations are linked by simplified yet nonlinear network-like models (Masters 1993; Garrett 1994; Russell and Norvig 2010). ANN has been applied to predict FIO concentrations in several studies e.g. Jin and Englande 2006; He and He 2008; Zhang et al. 2012; Thoe et al. 2015; Zhang et al. 2018. ANN models have shown better performance in predicting extreme FIO concentration (both high and low) (Zhang et al. 2012) and give higher sensitivity to poor water quality events (Thoe et al. 2015) compared to stepwise MLR. However, ANN models cannot identify predictive variables from a data set; manual selection of predictive variables is highly dependent on the users' judgement. This could be challenging in data-rich sites which will become more common as a result of enhancements in sensor and implementation of digital environments. While stepwise MLR can be conducted to identify predictive variables prior to ANN, stepwise MLR is not capable of identifying nonlinear relationships between variables and FIO concentrations.

A nonlinear alternative method to identify predictive variables is Gamma test. Gamma test determines the significance of a set of input variables in predicting the target data, e.g. FIO concentrations, by quantifying the residue variance that cannot be explained by any smooth nonlinear models (Jones 2004). Gamma test does not require an assumed nonlinear function relating input variables and target data a priori. On the other hand, Gamma test does not determine the nonlinear model itself. To identify predictive variables in a data set, Gamma tests can be applied to each possible combination of variables and choose the best combination to be the one that gives the smallest residue variance. However, for data sets containing large number of variables, searching the entire input combination space requires many Gamma test computations and large computational power. A Genetic-Algorithm (GA) model can be utilised to circumvent the need for such high computational power (Jones 2004). Although a cross validation approach (Stone 1974) may be applied for nonlinear variable identification, the approach usually requires *a priori* regression equations or network architectures and can be computation intensive for large data sets because of the increased number of networks needed for cross-comparison.

This paper develops a data-driven model for FIO concentration prediction based on a limited number of critical input variables. Essential variables were identified with be a combination of Gamma test and GA (Gamma-GA-test). This approach was compared with the stepwise MLR, which has been commonly applied in water quality prediction (Crowther et al. 2001; Nevers and Whitman 2005; Wyer et al. 2013b; Thoe and Lee 2014). From the variables identified by Gamma-GA test and stepwise MLR, ANN and linear regression models were developed and evaluated. These techniques were applied to a data-rich test site, namely Swansea Bay, UK, where a significant amount of FIO and environmental data were collected. This is the first time Gamma test has been applied to identify FIO predictive variables at bathing water sites. Gamma test has been applied in Kashefipour et al. 2005 and Lin et al. 2008 in bathing water quality modelling, but the test was not used for variable identification. The test has also been applied in Choubin and Malekian 2017 and Ghaderi et al. 2019 for variable identification, but their focuses were not bathing water quality. The improvements in predicting FIO concentrations using the complex model proposed in this study at the case study site are highlighted.

METHODOLOGY

The first step in developing data-driven models is identifying the input variables. However, there may be dependencies among the available variables. To identify linearly independent variables amongst all available variables, a singular value decomposition-based collinearity analysis was conducted. The output variables from the analysis became the input variables for the Gamma-GA tests or stepwise MLRs. ANN models were developed from predictive variables identified with Gamma-GA tests (GG-ANN) and stepwise MLRs (SL-ANN) respectively. Linear regression models were also developed from the identified variables (GG-Linear and SL-Linear models). Fig. 1 shows a flow chart of the modelling approach in this paper and the following section gives further details to the aforementioned tests and models.

Collinearity analysis

Linear correlation may exist among the variables within the measured data set. This is referred to as collinearity (Belsley et al. 1980). Correlation analysis was conducted in this paper to remove redundant variables. The correlation coefficients between variable pairs were computed. When the correlation coefficient was high (e.g. > 0.6), one of the variables was removed; the variable to remove was selected by mechanistic-process-based judgement. Correlation analysis was also conducted to determine the lag-time required for the time-lagged variables not to have a high correlation with the original un-lagged variables.

Gamma test and Genetic Algorithm

Gamma test determines the part of the variance of target data which cannot be accounted for by any smooth nonlinear models. Nevertheless, Gamma test does not determine the model itself. The Gamma test is briefly explained below but further details can be found in Stefansson et al. 1997, Evans and Jones 2002, and Jones 2004. Consider a data set of input variables (the independent environmental variables selected by the collinearity test in this paper) $\mathbf{X}$ and target data (FIO concentrations in this paper) **y**:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1M} & x_{2M} & \cdots & x_{NM} \end{bmatrix} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} \tag{2}$$

where x_{ji} is the j-th input variable at time i ($1 \leq i \leq M$); y_i is the target data at time i. The row vectors of matrix **X**, that is:

$$\mathbf{x}_i = \begin{bmatrix} x_{1i}, x_{2i}, \ldots, x_{Ni} \end{bmatrix} \tag{3}$$

where superscript T denotes matrix transpose, are the environmental variable data points measured at time instant i ($1 \leq i \leq M$). Assume that $\mathbf{x}_i$ and y_i are related as:

$$y_i = f(\mathbf{x}_i) + r \tag{4}$$

where f is a nonlinear and smooth function; r is a random variable (i.e. noise that is excluded from the input-target relationship). We define an imaginary data point $\mathbf{x}'_i$ near $\mathbf{x}_i$ and:

$$\gamma = \frac{1}{2M} \sum_{i=1}^{M} [y(\mathbf{x}'_i) - y(\mathbf{x}_i)]^2 \tag{5}$$

substitute Eq. 4 into Eq. 5 and consider the Taylor expansion $f(\mathbf{x}'_i) = f(\mathbf{x}_i) + (\mathbf{x}'_i - \mathbf{x}_i) \cdot \nabla f + O(|\mathbf{x}'_i - \mathbf{x}_i|^2)$, where ∇ is the gradient operator. Eq. 5 becomes:

$$\gamma = \frac{1}{2M}\sum_{i=1}^{M}[(\mathbf{x}'_i - \mathbf{x}_i)\cdot\nabla f]^2 + \frac{1}{2M}\sum_{i=1}^{M}(r_2 - r_1)^2 \tag{6}$$

where r_1 and r_2 are two realizations of the random variable r corresponding to $y(\mathbf{x}_i)$ and $y(\mathbf{x}'_i)$ respectively. It is obvious from Eq. 6 that if $\mathbf{x}'_i \to \mathbf{x}_i, \gamma \to \frac{1}{2M}\sum_{i=1}^{M}(r_2 - r_1)^2 = Var(r)$ where $Var(r)$ is the variance of r in probability. Note that $Var(r)$ is obtained without knowing the expression of f.

The data point $\mathbf{x}'_i$ that is arbitrarily close to $\mathbf{x}_i$ does not exist in the measured data set $\mathbf{X}$. An approach to estimate the first term in the right-hand side in Eq. 6 is required to obtain $Var(r)$ from γ. In Gamma test, $\mathbf{x}'_i$ is replaced by $\mathbf{x}_{N[i,k]}$, the k-th nearest data point to $\mathbf{x}_i$. For example, $\mathbf{x}_{N[i,k=1]}$ and $\mathbf{x}_{N[i,k=2]}$ are the nearest and the second nearest points to $\mathbf{x}_i$. With this replacement, Eq. 6 becomes:

$$\gamma(k) = A(k)\delta(k) + \Gamma \tag{7}$$

where

$$\delta(k) = \frac{1}{M}\sum_{i=1}^{M}|\mathbf{x}_{N[i,k]} - \mathbf{x}_i|^2 \tag{8}$$

$$\gamma(k) = \frac{1}{2M}\sum_{i=1}^{M}(y_{N[i,k]} - y_i)^2 \tag{9}$$

$$A(k) = \frac{\sum_{i=1}^{M}[(\mathbf{x}_{N[i,k]} - \mathbf{x}_i)\cdot\nabla f]^2}{\sum_{i=1}^{M}|\mathbf{x}_{N[i,k]} - \mathbf{x}_i|^2} \tag{10}$$

$$\Gamma = \frac{1}{2M}\sum_{i=1}^{M}(r_2 - r_1)^2 = Var(r) \tag{11}$$

where $|\cdot|$ denotes Euclidean distance; $y_{N[i,k]}$ is the target value associated with $\mathbf{x}_{N[i,k]}$ (note that $y_{N[i,k]}$ is not necessarily the k-th nearest point to y_i in the data set). Evans and Jones 2002 showed that $A(k) = A$ is a constant given M is sufficiently large. Eq. 7 becomes:

$$\gamma(k) = A\delta(k) + \Gamma \tag{12}$$

A and Γ in Eq. 12 can be obtained by conducting linear regression with $\gamma(k)$ and $\delta(k)$ computed from the k-th nearest points to $\mathbf{x}_i$ for $i = 1, \ldots, M$ and $k = 1, \ldots, p$ where p is the maximum value of k used. In this study, $p = 10$ as suggested by Jones 2004. $\mathbf{x}_{N[i,k]}$ for each $\mathbf{x}_i$ in $\gamma(k)$ and $\delta(k)$ are obtained by an efficient k-dimensional tree approach (computational time in the order of $MlogM$) in Bentley 1975.

Gamma tests need to be applied to all possible combinations of input variables in order to identify the strongest predictive variables from all available data. This is the combination which gives the lowest absolute value of Γ (i.e. the combination of input variables that gives the smallest noise variance). However, this approach is computationally demanding for large data sets; the number of possible combinations for a data set of m variables is $2^m - 1$. To circumvent the need for high computational power, the variable selection problem is expressed as a minimization problem which is solved using GA (Jones 2004). The combination of input variables that minimizes $|\Gamma|$ is selected as the solution. Combinations of input variables in the GA model are represented by a binary vector of length N (a mask) in which the inclusion or exclusion of a variable is indicated by 1 or 0, respectively. In this work, the GA function "ga" in MATLAB Global Optimization Toolbox (Mathworks 2020b) is used. Readers are also referred to Deb 2000 and Deep et al. 2009 for the detail of the GA approach.

M-test (Jones 2004) can be used to determine the minimum required length of the input and target data set, the value of M in Eq. 1. This value of M is also the minimum data length required for model training if a nonlinear model is applied to the data set. In an M-test, Gamma tests are conducted sequentially with progressively increasing M. The computed $|\Gamma|$ is plotted against data length. The minimum required M is the value of M beyond which $|\Gamma|$ becomes constant. M-test results are expected to be different when the ordering of the data is different. In this work, the data order was randomly generated and three different realizations (Realization 1, 2 and 3) of data in different order was tested to reduce reliance on the order of the data.

Artificial Neural Network (ANN)

Feedforward back-propagation ANN models were used to predict Faecal Indicator Organism concentrations, including *Enterococci* and *E Coli*, based on the predictive variables selected by the Gamma-GA tests. Each network consists of an input layer, an output layer, and one hidden layer. One hidden layer suffices in this case since Masters 1993 showed that networks with one hidden layer are generally capable of approximating most underlying functions. The authors also tested GG-ANN and SL-ANN networks with two hidden layers and no significant improvement in performance was obtained compared to one-hidden-layer networks. The number of nodes required in the hidden layer was determined by experimentation to give the best results without overfitting.

The networks were trained and validated with the ANN function "train" in MATLAB deep learning toolbox (Mathworks 2020a). To retain a portion of the data for model validation and to avoid over-training, the data set was divided to three sets, namely the training, validation, and testing sets. For each of the three realizations in the M-test, the data were grouped into the three sets according to the data order (i.e. the 1st to nth data were put into the training set; the $n+1$-th to m-th data were put into the validation set; the $m+1$-th to the end of the data were put into the testing set; $n < m$). The performance function used in this study was the mean squared error (MSE) between model outputs and target data. Training of a network was stopped when no further improvement in MSE for the validation data set can be achieved after six iterations. While this method avoids parameter (weights and bias) over-training, it does not avoid over-training due to over-complicated network architecture and redundant predictive variables. For each network, 300 training runs with random initial weights were conducted and the network that provided the minimum MSE was chosen. (Iyer and Rhinehart 1999) showed that the network obtained from this approach has a 95% confidence level that its MSE is within the lowest 1.0%.

Linear regression models with their predictive variables selected by the Gamma-GA tests (GG-Linear models) and stepwise MLR models (SL-Linear models) were also developed to assess ANN model performance.

MODEL APPLICATION

The model was applied to Swansea Bay, located on the north of the Bristol Channel in the South West of the UK, as shown in Fig. 2a. Along the bay are two sandy beaches with bathing water status: the Swansea Beach and the Aberavon Beach. Potential sources of FIO in the Bay are the discharges from rivers, streams, surface water drains, three offshore outfalls from wastewater treatment works, and transport by currents from sources outside of the Bay. Large amount of data were collected as a part of the previous Smart Coast Sustainable Communities (SCSC) research project (Wyer et al. 2013b, Wyer et al. 2018) which alongside the variety of the sources make the bay an ideal case study for data-driven modelling. The stream and drain discharges are generally low ($< 1m^3/s$); River Tawe, Clyne, Neath, and Afan have relatively high flow rates ($> 5m^3/s$). The water is well mixed in the Severn Estuary and Bristol Channel (Uncles 1981; Evans et al. 1990; Ahmadian et al. 2013). FIO concentrations in the beaches are governed by the sources and the hydrodynamics in the Bay (Ahmadian et al. 2013).

The concentrations of two FIO species, namely *E Coli* and *Enterococci*, were sampled at various sources and receptors at high frequency, i.e. intervals of 15-30 min, in year 2011 and 2012. In Swansea Beach, the large tidal range (exceeding 10 m) and sloping beach results in a tidal flat exposed up to 1500 m from shore during high spring tides. The large extent of the tidal flats makes single point FIO concentration measurement impossible. In the data collection scheme, FIO concentrations were measured along a sampling transect consisting of Designated Sampling Points (DSPs) in Swansea Bay, as shown in Fig. 2a. Fig. 2b shows the DSPs in the sampling transect in the 2011 bathing season. Environmental variables such as stream flow, tidal, meteorological, and water quality data at the locations are also shown in Fig. 2a. The samples were collected in sterile 1 L containers (Aurora Scientific) and stored in a refrigerator before analysis. The samples were then analysed for intestinal *Enterococci* and *E Coli* with standard membrane filtration techniques and analysed for turbidity with a bench turbidity meter (Hannah Instruments LP2000). Salinity was also measured with a conductivity meter (Mettler Toledo SevenGo). Fig. 2a also shows the sampling locations of the environmental variables in the SCSC project. Tide level and the flow rates at river

Tawe, Neath and Afan were measured by the existing gauges in the hydrometric monitoring network operated by Natural Resources Wales (NRW), the official natural resource management organisation in Wales. The water depths and velocities at the five smaller streams were measured by pressure transducers (OTT Orpheus Mini) and electro-magnetic velocity meters (Sensa RC2) respectively. Global radiation, temperature, relative humidity, rainfall, and wind speed were measured at the meteorological station. Global radiation was measured by a pyranometer (Skye Instruments SKS 1110); air temperature and relative humidity were measured by a sensor (Rotronic HygroClip2 HC2-S3); rainfall was measured with a tipping bucket rain gauge (Met One Instruments 370C 20.3 cm aperture, 0.2 mm tip); Wind speed and direction were measured with an anemometer (Gill Instruments WindSonic). Offshore wastewater discharge volumes were also measured in the SCSC project but were not included as potential model outputs because the tracer study conducted as a part of the SCSC project (Ahmadian et al. 2013) and the two-dimensional TELEMAC hydrodynamic simulation conducted by the authors suggested that they are not important for FIO concentrations at the DSPs compared to other FIO sources (not shown).

Table 1 and 2 summarises the data set used in this paper. The target variables were *Enterococci* and *E Coli* concentrations measured at the Bathing Water Designated Sampling Points (DSPs) during a bathing season, namely 22 June to 28 September, 2011. The input data set included 16 environmental variables measured in the same bathing season as shown in Table 2. The range of values of different input and target variables were significantly different, as shown in Table 1 and 2, due to the large number of factors that affects bacteria concentrations. In order to ensure consistency between data and reduce the impact of variation ranges on the model, all the data have been normalized to the range of 0-1 using the following equation:

$$x_{j,nor} = \frac{x_j - x_{j,min}}{x_{j,max} - x_{j,min}} \tag{13}$$

where x_j and $x_{j,nor}$ are the un-normalized and normalized j-th variable; $x_{j,max}$ and $x_{j.min}$ are the maximum and minimum of the time series x_j, before data processing and model training. Logarithmic transformation was applied to the variables that have relatively high skewness to transform them from a lognormal-like distribution to a normal-like distribution. This transformation is necessary because stepwise-MLR models assume normally distributed data (skewness=0). If the data has a significant skewness, the stepwise-MLR variable inclusion/exclusion procedures may not be suitable and subsequently the model would not result in good validation. Collinearities among variables were identified and redundant variables were removed. In order to build memory of the past conditions, e.g. solar radiation or rain prior to the simulation, and time required for transport of bacteria across the bay, which could significantly affect the concentration of bacteria, in the data-driven model, time-lagged variables were also considered as the input variables. Correlation analysis was conducted to identify collinear variables and determine and lag time that does not cause additional collinearity issues.

RESULTS AND DISCUSSION

Input variable selection

From the correlation analysis, 23 input variables were identified as shown in the column "Variables identified from the correlation analysis" in Table 3. Only one single representative stream flow, the flow of the Tawe River, was selected since flows at different streams with no time lag were found highly correlated. Such a high correlation could be explained by the small size of catchment associated with each stream, which means all streams are influenced with similar weather, and particularly rainfall, patterns. Burton et al. 2013 reported that the spatial correlation of rainfall at the site remains higher than 0.5 for two points that are 100 km apart, implying that the rainfall is correlated within a 100 x 100 = 10,000 km^2 area. This area is larger than the sum of the watersheds of three major rivers discharging to the Bay (506.4 km^2), namely the watershed for River Tawe (227.7 km^2), River Neath (190.9 km^2), and River Afan (87.8 km^2). Turbidity and salinity were also found highly correlated with the streamflow and thus eliminated from the data set. This is consistent with the idea in Thoe and Lee 2014 that salinity reflects the mixing between riverine freshwater and the ambient sea water. The correlation analysis also shows high correlations between a time-lagged variable and the same variable with no time-lag if the time lag is not sufficiently long as expected. The correlation between the no time-lag and time-lagged streamflow remains high (>

0.6) for a lag time from 0.25-36 hours; only the streamflow with 10-hour lag was selected following hydrodynamic model results (Lam and Ahmadian 2022). For other variables, the minimum time-lag from the FIO data was 2 hours to render the AI model predictive. Additional time-lags were applied to these variables at suitable time intervals such that the correlations between time-lagged and unlagged variables were not significant (< 0.6). The time intervals determined by correlation analysis were as follows: 2 hours for tides; 2 hours for radiation; 6 hours for humidity; 0.25-hour intervals for rainfall; 4 hours for temperature; 4 hours for wind speed N; 8 hours for wind speed E. Table 3 shows that the time interval for rainfall is greater than 0.25-hour; it is because rainfall was not expected to have an immediate effect on FIO concentrations from physical process point of view.

Table 3 shows the predictive variables identified by the Gamma-GA tests and stepwise MLRs. For consistency of comparison between the methods and prevent over-paramatization, both Gamma-GA tests and stepwise MLRs were constrained to choose a maximum of eight variables. Ideally, an interpretability analysis of the variables identified by Gamma-GA tests and stepwise MLRs is desirable to assess the performance of Gamma-GA tests, but such comparison requires a priori knowledge about the relative importance of the variables in the site, which is not available to date. Nevertheless, the physical plausibility of the selected variables is discussed as follows. Tide level was selected to be an important variable by both Gamma-GA tests and stepwise-MLR models. The results are consistent with the fact that tides were shown important to the flow in Swansea Bay (Ahmadian et al. 2013) as well as FIO concentrations (Lam and Ahmadian 2022). Tides were also identified as an important variable in other data-driven models for other nearshore coastal waters (Crowther et al. 2001; Nevers and Whitman 2005; He and He 2008; Zhang et al. 2012). Wind was also shown important for FIO concentration by both predictive variable identification methods. It is consistent with the stepwise-MLR results in Wyer et al. 2013b. Streamflow, as a known FIO source (e.g. Wyer et al. 2010; Wyer et al. 2013a; Lam and Ahmadian 2022), was included by only Gamma-GA model for *Enterococci* but the variable was not included for other tests. This is attributed to the small spatial and temporal scale (in a watershed of about 500 km^2 and sampling interval of 30 minutes) of the site. In this study, flow rates of different rivers under a time lag of less than 36 hours are highly correlated and one representative streamflow (River Tawe) at one particular time lag (10 hours) was selected. Information concerning the exact riverine FIO sources for the measured FIO concentration was lost. In summary, Gamma-GA test can identify predictive variables that are consistent with the literature.

M-test

M-test was conducted for variables identified by the Gamma-GA tests and stepwise MLRs to determine the data length needed for model training. Fig. 3 shows that the Gamma-GA tests selected variables that achieve lower (i.e. better) $|\Gamma|$ compared to the stepwise MLRs given a sufficiently long data (e.g. beyond 500 data points) which means that the data length for model training should be greater than 200 for Gamma-GA test to give better results compared with stepwise MLR. Following the M-test, the ratio of data points in training, validation and testing sets is 0.6:0.2:0.2, giving 949 x 0.6 = 571 data points in the training set, which satisfies the minimum of 200 data points imposed by the M-test. The mean and standard deviation values of the training, validation, and testing data sets are checked to be approximately comparable in all three realizations.

ANN model results

Selection of number of nodes

Fig. 4 shows a typical relationship between MSE and number of hidden layer node for *Enterococci*. SL-ANN models reached lower MSEs when there were few nodes in the hidden layer. As there were more hidden layer nodes ($>$10), GG-ANN models achieved better performance. For *E Coli*, Fig. 5 shows that GG-ANN and SL-ANN models gave similar MSEs. The fact that GG-ANN and SL-ANN models gave similar MSEs does not conflict with the M-test results. Although the M-test results suggested that the Gamma-GA tests identified variables had the potential to achieve lower MSEs, M-test does not specify the nonlinear model that gives such results. It is possible that other nonlinear models other than ANN give better results in comparison to the Gamma-GA, however, this is out of the scope of this study. For further comparison between models developed from the Gamma-GA tests and stepwise MLRs, networks with 1 to 50 nodes in the hidden layer were tested and the number of nodes in the hidden layer was selected based on providing the lowest validation MSE. The MSE resulting from different networks with a different number of nodes in the hidden layer is illustrated in Fig. 4 and 5.

Mean squared error (MSE) and R^2

Fig. 6 and 7 show the comparison between GG-ANN model results and target FIO concentrations for training, validation, and test data sets, as well as all data. Table 4 and 5 show the comparison between GG-ANN, SL-ANN, GG-Linear, and SL-Linear models. For most ANN models, the optimal SL-ANN models consisted of fewer hidden layer nodes compared to GG-ANN models, which is consistent with the section "Selection of number of nodes". GG-ANN and SL-ANN models gave better MSE and R^2 than GG-Linear and SL-Linear models. This shows the capacity of nonlinear models in capturing inherent nonlinear relationships between variables and FIO concentrations. GG-ANN models gave better training, validation, and testing results than SL-ANN models for *Enterococci*, but SL-ANN models gave better validation and testing results for *E Coli*. The better GG-ANN performance for *Enterococci* can be explained by the fact that GG-ANN better captures extreme FIO concentrations as illustrated in section "Performance Table". This GG-ANN property helps the models perform better for *Enterococci* because there are more extreme values for the data series of *Enterococci* (17.8% of the data was below 0.1 or above 0.9) compared to *E Coli* (8.9% of the data was below 0.1 or above 0.9). The MSE of SL-Linear models was better than the one of the GG-Linear models, verifying the fact that stepwise MLR chooses variables that optimize linear model performance compared to Gamma-GA test.

Performance tables

The ability to identify the most hazardous circumstances, namely poor water quality conditions, are particularly important when a real-time predictive model is used as an early warning system. The EU rBWD (European Commission 2006) considers the water quality in a bathing site "poor" if the 90-percentile FIO concentration in the log-normal distribution obtained from the last assessment period (usually the last four bathing seasons) exceeds a given threshold. The threshold is 185 cfu/100 mL for *Enterococci* and 500 cfu/100 mL for *E Coli*. In this study, the use of 90-percentile values is not sensible since water quality is being predicted at a 30-minute interval. To test the models' ability to identify poor water quality events, individual *Enterococci* and *E Coli* concentration values were compared to the 185 cfu/100 mL and 500 cfu/100 mL thresholds, respectively. Fig. 8 shows the performance tables of the data-driven models in correctly predicting poor water quality under the EU rBWD classification for the testing sets. In this context, sensitivity and specificity are defined as:

$$Sensitivity = \frac{Correctly \quad predicted \quad poor \quad water \quad quality}{Observed \quad poor \quad water \quad quality} \tag{14}$$

$$Specificity = \frac{Correctly \quad predicted \quad not \quad poor \quad water \quad quality}{Observed \quad not \quad poor \quad water \quad quality} \tag{15}$$

To explain, sensitivity represents the likelihood that a poor water quality event is correctly predicted. Specificity represents the likelihood that a "not poor" water quality event is correctly predicted. It can be alternatively interpreted as a minus false alarm rate. Being consistent with the result that the ANN models gave better MSE and R^2 values, the ANN models gave significantly higher sensitivity than the linear regression models: 24-62% and 0%-14% respectively. The observation that nonlinear models give more accurate FIO predictions is consistent with Thoe et al. 2015. The results are also consistent with Zhang et al. 2012 that ANNs capture extreme FIO values better than linear regression models. Sensitivities of GG-ANN models were higher than the ones for SL-ANN models for both *Enterococci* and *E Coli*, despite the MSEs for SL-ANN models being lower than the ones for GG-ANN models for *E Coli*. It suggests that GG-ANN models better capture high FIO concentrations compared to SL-ANN models. The sensitivity of SL-Linear models was better than the one of GG-Linear models as expected from the MSE and R^2 results. Specificities for all the models tested were always higher than 90%. This is probably due to the small proportion of poor water quality events during the study period.

Sensitivity was higher for *Enterococci* concentrations than for *E Coli* in Fig. 8 for GGANN models. The same was observed in the performance tables for the entire data set (949 data points). An explanation can be given from the probability distribution of the *Enterococci* and *E Coli* data.

A Chi-square test showed that both data follow lognormal distributions with a confidence level of above 95% if very small values (<3 cfu/100 mL) were removed. From the respective probability distribution functions, the exceedance probability of the *Enterococci* threshold (15.9%) was higher than the one of *E Coli* (5.5%) from the entire data set. With a lower exceedance probability, models that better capture extreme values are required to achieve a better sensitivity for *E Coli* compared to *Enterococci* concentrations.

DISCUSSION

Gamma test is a promising tool to identify input data for a data-driven model because it is nonlinear, and it does not require a regression equation a priori. Nevertheless, these advantages do not imply that Gamma tests can be applied with no knowledge about the waterbody or the data to which the test is applied. In this paper, a correlation analysis was conducted prior to the Gamma-GA tests to remove highly correlated candidate predictive variables.

The GG-ANN and SL-ANN models fitted better to the measured FIO concentrations and captured better the extreme FIO concentrations compared to GG-Linear and SL-Linear models. This is consistent with Zhang et al. 2012 for FIO concentrations and Keiner and Yan 1998 for chlorophyll-a and suspended sediments. While Zhang et al. 2012 arrived at this conclusion by comparing 15-variable ANN models to 5 or 6-variable linear regression models, this study confirms their findings with the same number of explanatory variables used in the ANN and linear models. This demonstrates the importance of including nonlinearity in capturing high FIO concentrations. The effect of nonlinearity of ANN is also reflected in higher sensitivities of GG-ANN and SL-ANN models compared to GG-Linear and SL-Linear models. Comparing GG-ANN and SL-ANN models, GG-ANN models gave better results for *Enterococci* for all training, validation and testing sets and most of the training sets for *E Coli*. While SL-ANN models gave better testing results for *E Coli*, GG-ANN models gave higher sensitivities for both *Enterococci* and *E Coli*, showing that Gamma-GA models select variables that capture better extreme FIO concentrations compared to the stepwise MLR. The results suggest that GG-ANN model is more suitable for bathing water quality warning applications in which predicting high FIO concentrations is the major concern.

This paper presented a GG-ANN model training and validation framework which is generally applicable to different sites, although a new GG-ANN model development is required for every new study. Once the GG-ANN model is developed, it can discern critical parameters from redundant parameters for water quality prediction and to keep the sampling cost of running the model in real-time limited by only measuring critical parameters. The data used in this paper had a very short (0.5 hr) sampling interval, while the commonly used sampling intervals are usually in the order of days (He and He 2008; Zhang et al. 2012; Thoe et al. 2015; Zhang et al. 2018). The models used in this paper generally gave lower R^2 than the daily predictions of water quality in the literature. This highlights the difficulty of short-term water quality prediction; further study of the effect of time-scale on prediction accuracy is needed. While Zhang et al. 2018 attempted to predict water quality with ANN at different time-scales, their results were not conclusive. Nevertheless, this paper highlights the potential for a combination of Gamma-GA test and a nonlinear predictive model to give timely bathing water quality prediction. This can be used as early warning systems on impending poor water quality together with real-time environmental sensors.

CONCLUSION

This paper develops a data-driven model for FIO concentration prediction with only limited number of critical and without unnecessary input variables. The performance of Gamma-GA test as a tool for predictive variable identification of *Enterococci* and *E Coli* at interval of 30 minutes was evaluated. ANN and linear regression models were developed from the variables identified from Gamma-GA test and stepwise MLR for comparison. The GG-ANN models gave better results for *Enterococci* for all training, validation and testing sets and most of the training sets for *E Coli*. The results also demonstrated the potential for Gamma-GA test to identify variables that give a better model compared to stepwise MLR. While SL-ANN models usually gave better MSE and R^2 for testing results of *E Coli*, GG-ANN model was better in identifying events of poor water quality. This illustrates the merit of nonlinear variable identification approach – the variables identified are more capable of predicting high FIO concentrations. Therefore, GG-ANN model is more suitable for bathing water warning applications in which predicting high FIO concentrations is the major concern. For the two variable identification approaches, ANN models were better than linear regression models in terms of MSE, R^2 as well as sensitivity. This result again highlighted the importance of including nonlinear effects in prediction models.

In conclusion, this paper demonstrated the potential of combining Gamma-GA test and ANN to predict bathing water quality. Prior to the variable identification tests, a correlation analysis was conducted to remove redundant variables in the data set. The need of such an analysis illustrates the importance of understanding the data set in the development and application of data-driven models.

DATA AVAILABILITY STATEMENT

Some data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request (The MATLAB codes for Gamma-GA tests and the ANN models and the model outputs). The bacteria and environmental data were collected by Aberystwyth University, Swansea City Council and Natural Resources Wales (Environmental Protection and Regulatory Authority in Wales) and unfortunately the authors have not been authorized to share the data set.

ACKNOWLEDGEMENTS

The study is carried out as a part of the EERES4WATER (Promoting Energy-Water nexus resource efficiency through renewable energy and energy efficiency) project, which is co-financed by the Interreg Atlantic Area Programme through the European Regional Development Fund under EAPA 1058/2018.

NOTATION

The following symbols are used in this paper:

A = the slope for the Gamma test regression equation;

$f(\cdot)$ = the nonlinear smooth function relating $\mathbf{x}_i$ and y_i;

M = the total number of time instants of the data;

r = random variable (noise);

$\mathbf{X}$ = the linearly independent environmental variables;

$\mathbf{x}_i$ = the linearly independent environmental variables at time instant i (i.e. the data point at i);

$\mathbf{x}'_i$ = the imaginary environmental variables at time instant i;

$\mathbf{x}_{N[i,k]}$ = the k-th nearest data point to $\mathbf{x}_i$;

x_j = the j-th environmental variable;

$x_{j,nor}$ = the normalized x_j;

$x_{j,max}$ = the maximum value of variable x_j;

$x_{j,min}$ = the minimum value of variable x_j;

$\mathbf{y}$ = target data (FIO concentrations);

$y_{N[i,k]}$ = the target data value associated with $\mathbf{x}_{N[i,k]}$;

γ = an estimate of variability among $\mathbf{x}_i$, $1 \leq i \leq M$;

Γ = the intercept for the Gamma test regression;

δ = an estimate of variability among y_i, $1 \leq i \leq M$

REFERENCES

Abu-Bakar, A., Ahmadian, R., and Falconer, R. A. (2017). “Modelling the transport and decay processes of microbial tracers in a macro-tidal estuary.” *Water Res.*, 123, 802–824.

Ahmadian, R., Bomminayuni, S., Falconer, R., and Stoesser, T. (2013). “Numerical modelling of flow and faecal indicator organism transport at swansea bay, uk.” *A report from the interreg 4a smart coasts – sustainable communities project*.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley and Sons, USA.

Bentley, J. (1975). “Multidimensional binary search trees used for associative search.” *Comm ACM*, 18, 309–517.

Boehm, A. B., Grant, S. B., Kim, J. H., Mowbray, S. L., McGee, C. D., Clark, C. D., Foley, D. M., and Wellman, D. E. (2002). “Decadal and shorter period variability of surf zone water quality at huntington beach, california.” *Environ. Sci. Technol.*, 36(18), 3885–3892.

Burton, A., Glenis, V., Jones, M. R., and Kilsby, C. G. (2013). “Models of daily rainfall cross-correlation for the united kingdom.” *Environmental Modelling and Software*, 49, 22–33.

Choubin, B. and Malekian, A. (2017). “Combined gamma and m-test-based ann and arima models for groundwater fluctuation forecasting in semiarid regions.” *Environ. Earth Sci.*, 76, 538.

Crowther, J., Kay, D., and Wyer, M. D. (2001). “Relationships between microbial water quality and environmental conditions in coastal recreational waters: The fylde coast, uk.” *Water Res.*, 35(17), 4029–4038.

Deb, K. (2000). “An efficient constraint handling method for genetic algorithms.” *Comput. Methods Appl. Mech. Engrg.*, 186, 311–338.

Deep, K., Singh, K. P., Kansal, M. L., and Mohan, C. (2009). “A real coded genetic algorithm for solving integer and mixed integer optimization problems.” *Applied Mathematics and Computation*, 212(2), 505–518.

Dufour, A. P. (1984). “Bacterial indicators of recreational water quality.” *Canadian journal of public health*, 75(1), 49–56.

European Commission (2006). “Directive 2006/7/EC of the european parliament and of the council of 15 february 2006 concerning the management of bathing water quality and repealing directive 76/160/EEC.” *OJEU*, 64, 37–51.

Evans, D. and Jones, A. J. (2002). “A proof of the gamma test.” *Proc R Soc Lond A*, 458, 2759–2799.

Evans, G. P., Mollowney, B. M., and Spoel, N. C. (1990). “Two-dimensional modelling of the bristol channel, uk.” *Proceedings of the conference on estuarine and coastal modelling*, S. ML, ed., 331–340.

Garrett, J. J. (1994). “Where and why artificial neural networks are applicable in civil engineering.”

J Comput. Civil Eng., ASCE, 8(2), 129–130.

Ghaderi, K., Motamedvaziri, B., Vafakhah, M., and Dehghani, A. A. (2019). “Regional flood frequency modeling: a comparative study among several data-driven models.” *Arab J Geosci*, 12, 588.

Gonzalez, R. A., Conn, K. E., Crosswell, J. R., and Noble, R. T. (2012). “Application of empirical predictive modeling using conventional and alternative fecal indicator bacteria in eastern north carolina waters.” *Water Res.*, 46, 5871–5882.

Gonzalez, R. A. and Noble, R. T. (2014). “Comparisons of statistical models to predict fecal indicator bacteria concentrations enumerated by qpcr- and culture-based methods.” *Water Res.*, 48, 296–305.

He, L. M. and He, Z. L. (2008). “Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern california, usa.” *Water Res.*, 42, 2563–2573.

Huang, G., Falconer, R. A., and Lin, B. (2017). “Integrated hydro-bacterial modelling for predicting bathing water quality.” *Estuarine, Coastal and Shelf Science*, 188, 145–155.

Iyer, M. S. and Rhinehart, R. R. (1999). “A method to determine the required number of neural-network training repetitions.” *IEEE T. Neural Networ.*, 10(2), 427–432.

Jin, G. and Englande, A. J. (2006). “Prediction of swimmability in a brackish water body.” *Management of Environmental Quality*, 17(2), 197–208.

Jones, A. J. (2004). “New tools in non-linear modelling and prediction.” *CMS*, 1, 109–149.

Kashefipour, S. M., Lin, B., and Falconer, R. A. (2005). “Neural networks for predicting seawater bacterial levels.” *P. I. Civil Eng-Wat M.*, 158(3), 111–118.

Keiner, L. E. and Yan, X. (1998). “A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery.” *Remote Sens. Environ.*, 66, 153–165.

Kim, J. H., Grant, S. B., McGee, C. D., Sanders, B. F., and Largier, J. L. (2004). “Locating sources of surf zone pollution: A mass budget analysis of fecal indicator bacteria at huntington beach, california.” *Environ. Sci. Technol.*, 38(9), 2626–2636.

King, J., Ahmadian, R., and Falconer, R. (2021). “Hydro-epidemiological modelling of bacterial transport and decay in nearshore coastal waters.” *Water Res.*, 196, 117049.

Lam, M. Y. and Ahmadian, R. (2022). “Numerical source-receptor connectivity study in nearshore coastal waters.” *Proceedings of the 39th IAHR World Congress, Guadiana, Spain, 19-24 Jun.*

Lee, J. and Qu, B. (2004). “Hydrodynamic tracking of the massive spring 1998 red tide in hong kong.” *J. Environ. Eng., ASCE*, 130(5), 535–550.

Lin, B., Syed, M., and Falconer, R. A. (2008). “Predicting faecal indicator levels in estuarine receiving waters - an integrated hydrodynamic and ann modelling approach.” *Environ. Modell. Softw.*, 23, 729–740.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. Academic, San Diego, CA.

Mathworks (2020a). *MATLAB Deep learning toolbox user’s guide*. Massachusetts, USA.

Mathworks (2020b). *MATLAB Global optimization toolbox user’s guide*. Massachusetts, USA.

Nevers, M. B. and Whitman, R. L. (2005). “Nowcast modeling of escherichia coli concentrations at multiple urban beaches of southern lake michigan.” *Water Res.*, 39, 5250–5260.

Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S., and Singh, V. (2014). “Contamination of water resources by pathogenic bacteria.” *AMB Express*, 4(51), 5250–5260.

Pruss, A. (1998). “Review of epidemiological studies on health effects from exposure to recreational water.” *International Journal of Epidemiology*, 27, 1–9.

Russell, S. J. and Norvig, P. (2010). *Artificial intelligence - a modern approach*. Pearson, England, 3rd edition edition.

Schippmann, B., Schernewhki, G., and Grawe, U. (2013). “Escherichia coli pollution in a baltic sea lagoon: A model-based source and spatial risk assessment.” *International Journal of Hygiene and Environmental Health*, 216, 408–420.

Stefansson, A., Koncar, N., and Jones, A. J. (1997). “A note on the gamma test.” *Neural Comput and Applic*, 5, 131–133.

Stone, M. (1974). “Cross-validatory choice and assessment of statistical predictions.” *J R Stat Soc Series B*, 36, 117–147.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M. L., and Boehm, A. B. (2015). "Sunny with a chance of gastroenteritis: Predicting swimmer risk at california beaches." *Environ. Sci. Technol.*, 49, 423–431.

Thoe, W. and Lee, J. H. W. (2014). "Daily forecasting of hong kong beach water quality by multiple linear regression models." *J Environ Eng, ASCE*, 140(2), 04013007.

Uncles, R. J. (1981). "A numerical simulation of the vertical and horizontal m2 tide in the bristol channel and comparisons with observed data." *Limnology and Oceanography*, 26, 571–577.

USEPA (2010). "Predictive tools for beach notification, vol i: Review and technical protocol." *Report No. EPA-823-R-10-003*, USEPA.

Whitman, R. L., Nevers, M. B., Korinek, G. C., and Byappanahalli, M. N. (2004). "Solar and temporal effects on escherichia coli concentration at a lake michigan swimming beach." *Appl. Environ. Microbial.*, 70(7), 4276–4285.

Wyer, M. D., Kay, D., Morgan, H., Naylor, S., Clark, S., Govier, P., Watkins, J., Davies, C., Francis, C., Jones, J., Palmer, C., and Kay, C. (2013a). "Faecal indicator source connectivity for inputs to swansea bay, south wales, uk, a report from the interreg 4a smart coasts-sustainable communities project." *Report no.*

Wyer, M. D., Kay, D., Morgan, H., Naylor, S., Clark, S., Watkins, J., Davies, C. M., Francis, C., Osborn, H., and Bennett, S. (2018). "Within-day variability in microbial concentrations at a uk designated bathing water: Implications for regulatory monitoring and the application of predictive modelling based on historical compliance data." *Water Res. X*, 1, 100006.

Wyer, M. D., Kay, D., Morgan, H., Naylor, S., Govier, P., Clark, S., Watkins, J., Davies, C. M., Francis, C., Osborn, H., and Bennett, S. (2013b). "Statistical modelling of faecal indicator organisms at a marine bathing water site: results of an intensive study at swansea bay, uk: A report from the interreg 4a smart coasts – sustainable communities project." *Report no.*

Wyer, M. D., Kay, D., Watkins, J., Davies, C., Kay, C., Thomas, R., Porter, J., Stapleton, C. M., and Moore, H. (2010). "Evaluating short-term changes in recreational water quality during a hydrograph event using a combination of microbial tracers, environmental microbiology, microbial source tracking and hydrological techniques: A case study in southwest wales, uk." *Water Res.*, 44(16), 4783–4795.

Zhang, J., Qiu, H., Li, X., Niu, J., Nevers, M. B., Hu, X., and Phanikumar, M. S. (2018). "Real-time nowcasting of microbiological water quality at recreational beaches: a wavelet and artificial neural network based hybrid modelling approach." *Environ. Sci. Technol.*, 52, 8446–8455.

Zhang, Z., Deng, Z., and Rusch, K. A. (2012). "Development of predictive models for determining enterococci levels at gulf coast beaches." *Water Res.*, 46(2), 465–474.

List of Tables

TABLE 1. Measured FIO concentrations during 22 June-28 September, 2011. Ln denotes natural logarithm.

	Variables	Ln transformation	Range after transformation	
			min	max
FIO Data:	*E Coli* (cfu/100 mL)	Yes	1.10	8.04
	Enterococci (cfu/100 mL)	Yes	1.10	8.37

TABLE 2. Measured environmental variables during 22 June-28 September, 2011. Ln denotes natural logarithm.

	Variables	Ln transformation	Range after transformation	
			min	max
Stream flow data:	Washinghouse Brook (m^3/s)	Yes	-4.71	-0.03
	Brockhole Stream (m^3/s)	Yes	-5.30	-1.94
	Clyne River (m^3/s)	Yes	-2.42	1.98
	Brynmill Stream (m^3/s)	Yes	-4.20	1.33
	River Tawe (m^3/s)	Yes	0.843	5.15
	River Neath (m^3/s)	Yes	0.642	5.03
	River Afan (m^3/s)	Yes	0.298	4.23
Tidal data:	Normalized Tide level at Mumbles (–)	No	-0.499	0.483
Meteorological data:	Global radiation (W/m^2)	Yes	-1.97	6.86
	Temperature (^oC)	No	8.92	23.2
	Relative humidity (%)	No	34.3	99.0
	Rainfall (mm)	Yes	-13.8	0.588
	Wind speed to the North (m/s)	No	-11.6	4.14
	Wind speed to the East (m/s)	No	-5.90	6.64
Water quality data:	Turbidity (NTU)	Yes	0.843	4.97
	Salinity (ppt)	No	1.90	153

TABLE 3. Predictive variables selected by the Gamma-GA tests and stepwise MLRs

Variables identified from the correlation analysis	*Enterococci*		*E Coli*	
	Gamma-GA test	Stepwise Linear model	Gamma-GA test	Stepwise Linear model
Streamflow [lag 10 h]	**1**	0	0	**1**
Mumbles Level [lag 2 h]	0	0	0	0
Mumbles Level [lag 4 h]	**1**	0	**1**	**1**
Mumbles Level [lag 6 h]	**1**	**1**	**1**	0
Global Radiation [lag 2 h]	0	**1**	0	**1**
Global Radiation [lag 4 h]	0	0	**1**	0
Global Radiation [lag 6 h]	**1**	0	0	0
Temperature [lag 2 h]	**1**	0	0	**1**
Temperature [lag 6 h]	**1**	0	**1**	0
Relative Humidity [lag 2 h]	0	**1**	0	**1**
Relative Humidity [lag 8 h]	**1**	0	**1**	0
Cum. of Rain [lag 2 h]	0	0	0	0
Cum. of Rain [lag 3 h]	0	0	0	0
Cum. of Rain [lag 4 h]	0	0	0	0
Cum. of Rain [lag 6 h]	0	0	0	0
Cum. of Rain [lag 8 h]	0	0	0	0
Cum. of Rain [lag 10 h]	0	0	0	0
Cum. of Rain [lag 12 h]	0	**1**	0	0
Wind Speed N [lag 2 h]	0	**1**	**1**	**1**
Wind Speed N [lag 6 h]	0	0	**1**	0
Wind Speed N [lag 10 h]	0	**1**	0	**1**
Wind Speed E [lag 2 h]	**1**	**1**	**1**	**1**
Wind Speed E [lag 10 h]	0	**1**	0	0

"1" denotes selected and "0" denotes not-selected.

TABLE 4. MSE and unadjusted R^2 between computed and measured *Enterococci* concentrations. The best results for each Realization are bolded.

Realization 1							
	Hidden layer node no.	MSE			R^2		
		Training	Validation	Testing	Training	Validation	Testing
GG-ANN	35	**0.0074**	**0.0157**	**0.0214**	**0.8369**	**0.6654**	**0.5361**
SL-ANN	11	0.0210	0.0232	0.0260	0.5400	0.5079	0.4357
GG-Linear	N/A	0.0357		0.0399	0.2224		0.1348
SL-Linear	N/A	0.0311		0.0328	0.3235		0.2883
Realization 2							
	Hidden layer node no.	MSE			R^2		
		Training	Validation	Testing	Training	Validation	Testing
GG-ANN	38	**0.0134**	**0.0172**	**0.0227**	**0.7177**	**0.6025**	**0.4993**
SL-ANN	6	0.0257	0.0194	0.0295	0.4542	0.5518	0.3611
GG-Linear	N/A	0.0368		0.0352	0.2021		0.2246
SL-Linear	N/A	0.0312		0.0322	0.3229		0.2895
Realization 3							
	Hidden layer node no.	MSE			R^2		
		Training	Validation	Testing	Training	Validation	Testing
GG-ANN	40	**0.0071**	**0.0188**	**0.0199**	**0.8292**	**0.6457**	**0.6156**
SL-ANN	12	0.0192	0.0243	0.0225	0.5385	0.5418	0.5699
GG-Linear	N/A	0.0359		0.0393	0.1944		0.2403
SL-Linear	N/A	0.0320		0.0293	0.2812		0.4337

TABLE 5. MSE and unadjusted R^2 between computed and measured *E Coli* concentrations. The best results for each Realization are bolded.

Realization 1

	Hidden layer node no.	MSE			R^2		
		Training	Validation	Testing	Training	Validation	Testing
GG-ANN	25	**0.0067**	0.0159	0.0214	**0.8396**	0.6077	0.5312
SL-ANN	18	0.0096	**0.0154**	**0.0174**	0.7705	**0.6198**	**0.6177**
GG-Linear	N/A	0.0325		0.0342	0.2166		0.2482
SL-Linear	N/A	0.0297		0.0305	0.2838		0.3290

Realization 2

	Hidden layer node no.	MSE			R^2		
		Training	Validation	Testing	Training	Validation	Testing
GG-ANN	13	0.0119	0.0178	0.0205	0.7208	0.5939	0.4801
SL-ANN	18	**0.0113**	**0.0135**	**0.0188**	**0.7370**	**0.6914**	**0.5221**
GG-Linear	N/A	0.0331		0.0315	0.2294		0.1996
SL-Linear	N/A	0.0299		0.0295	0.3041		0.2484

Realization 3

	Hidden layer node no.	MSE			R^2		
		Training	Validation	Testing	Training	Validation	Testing
GG-ANN	25	**0.0073**	**0.0146**	0.0196	**0.8066**	**0.6747**	0.6337
SL-ANN	27	0.0091	0.0155	**0.0186**	0.7582	0.6539	**0.6501**
GG-Linear	N/A	0.0321		0.0363	0.1873		0.3181
SL-Linear	N/A	0.0297		0.0310	0.2501		0.4167

List of Figures

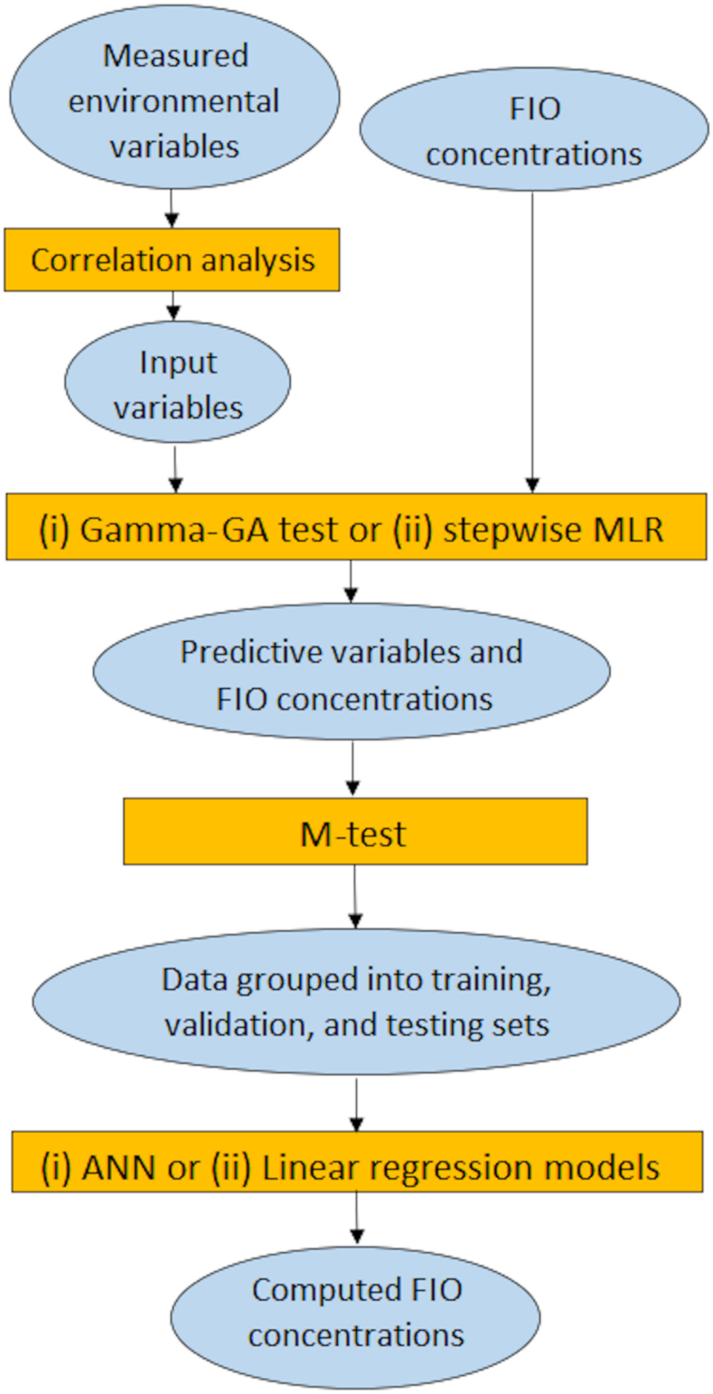
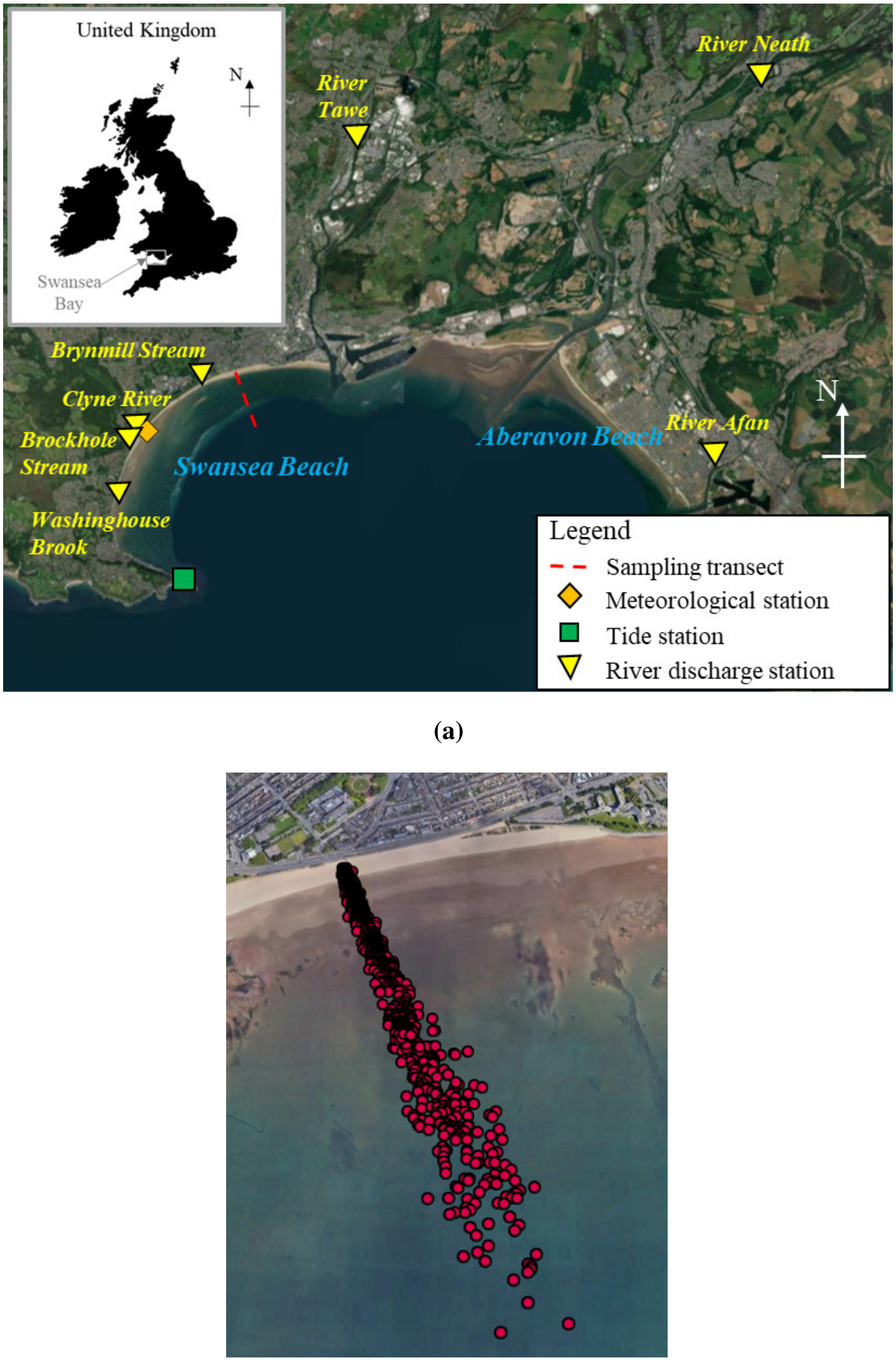

Measured environmental variables
FIO concentrations
Correlation analysis
Input variables
(i) Gamma-GA test or (ii) stepwise MLR
Predictive variables and FIO concentrations
M-test
Data grouped into training, validation, and testing sets
(i) ANN or (ii) Linear regression models
Computed FIO concentrations

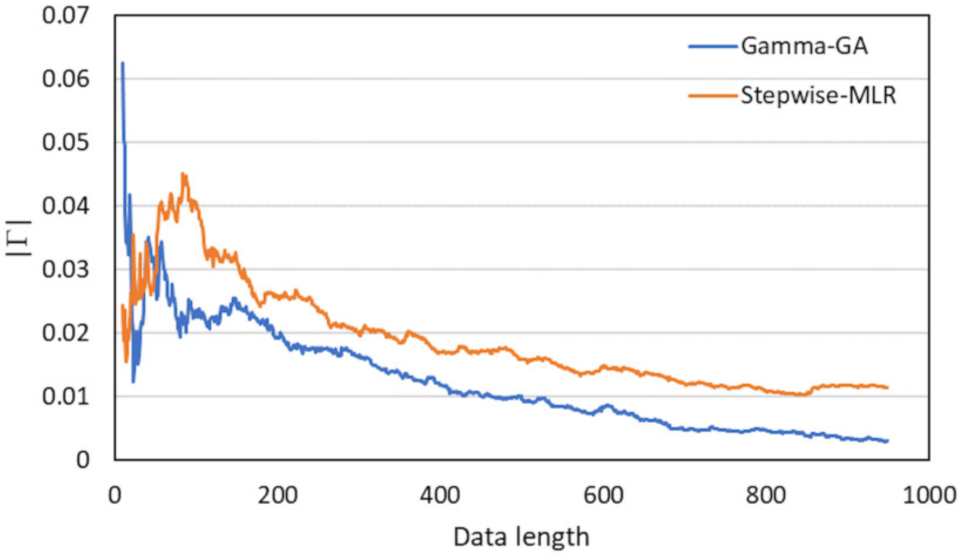
Fig. 1. Flow chart for the modelling approach.



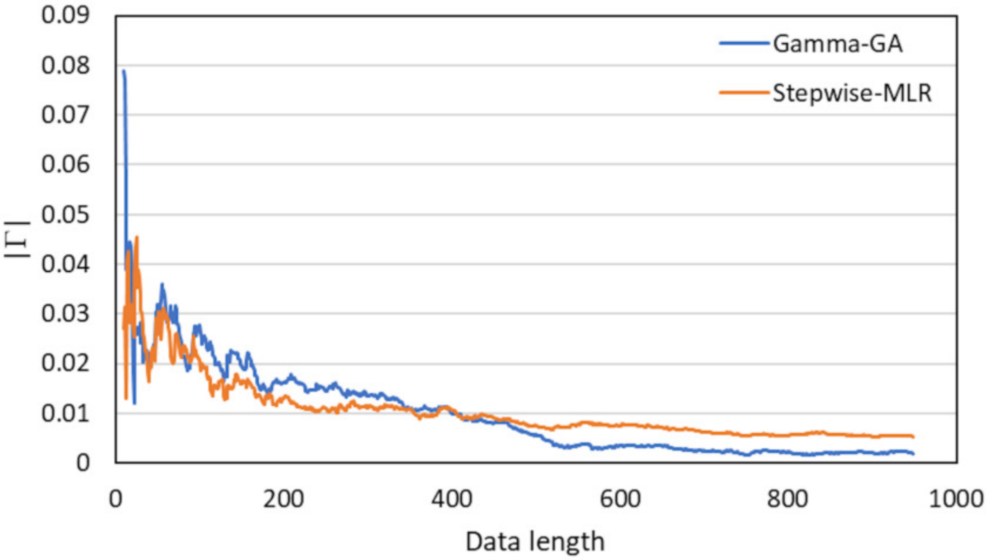
(a)

(b)

Fig. 2. (2a) Site layout and key sampling locations in Swansea Bay, UK. (2b) A close-up of Designated Sampling Points (DSPs, in purple) in the sampling transect in Figure 2a in the 2011 bathing season at 30-minute intervals from 07:00 to 16:00. Base map: (2a) Esri; (2b) Google Earth.

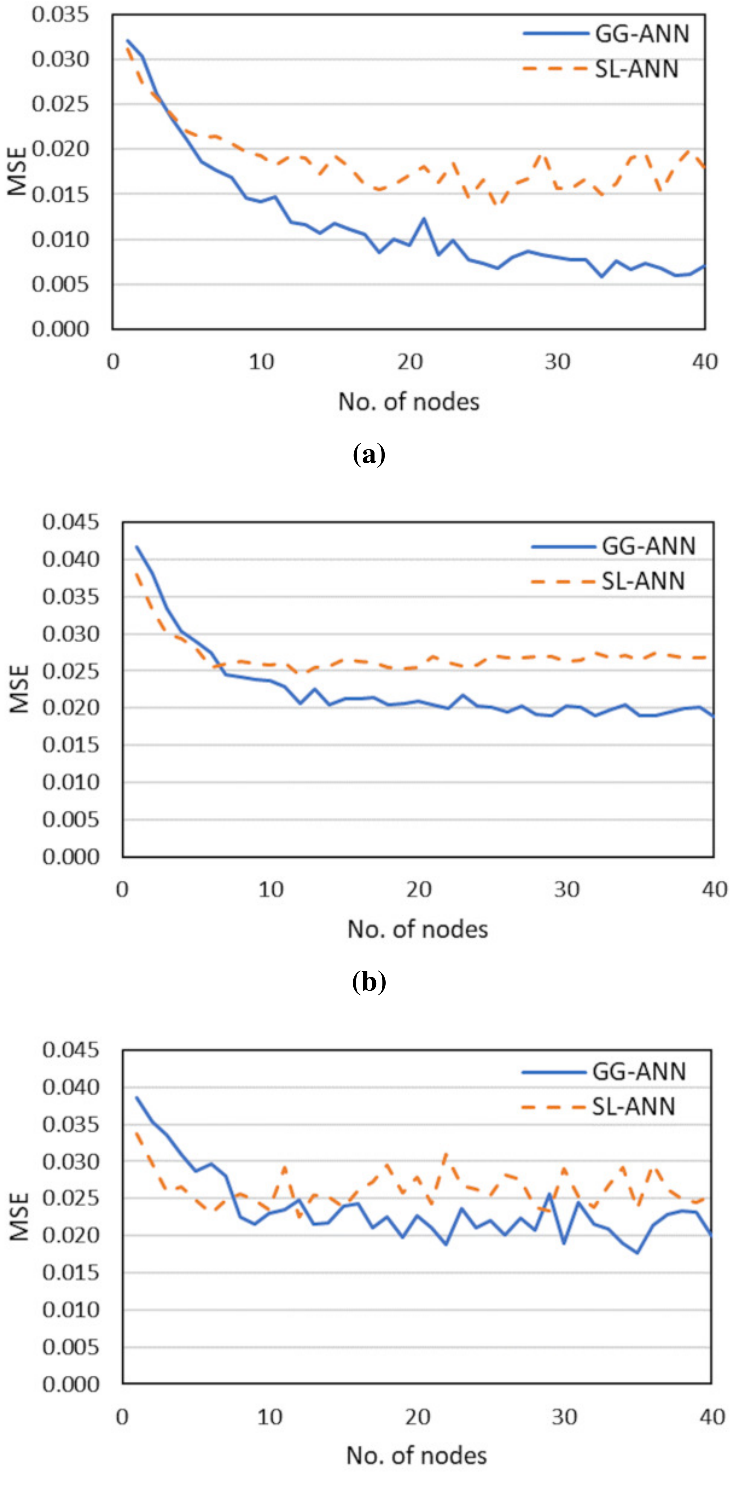


(a)



(b)

Fig. 3. M-test results for (3a) *Enterococci*, Realization 3 and (3b) *E Coli*, Realization 1.

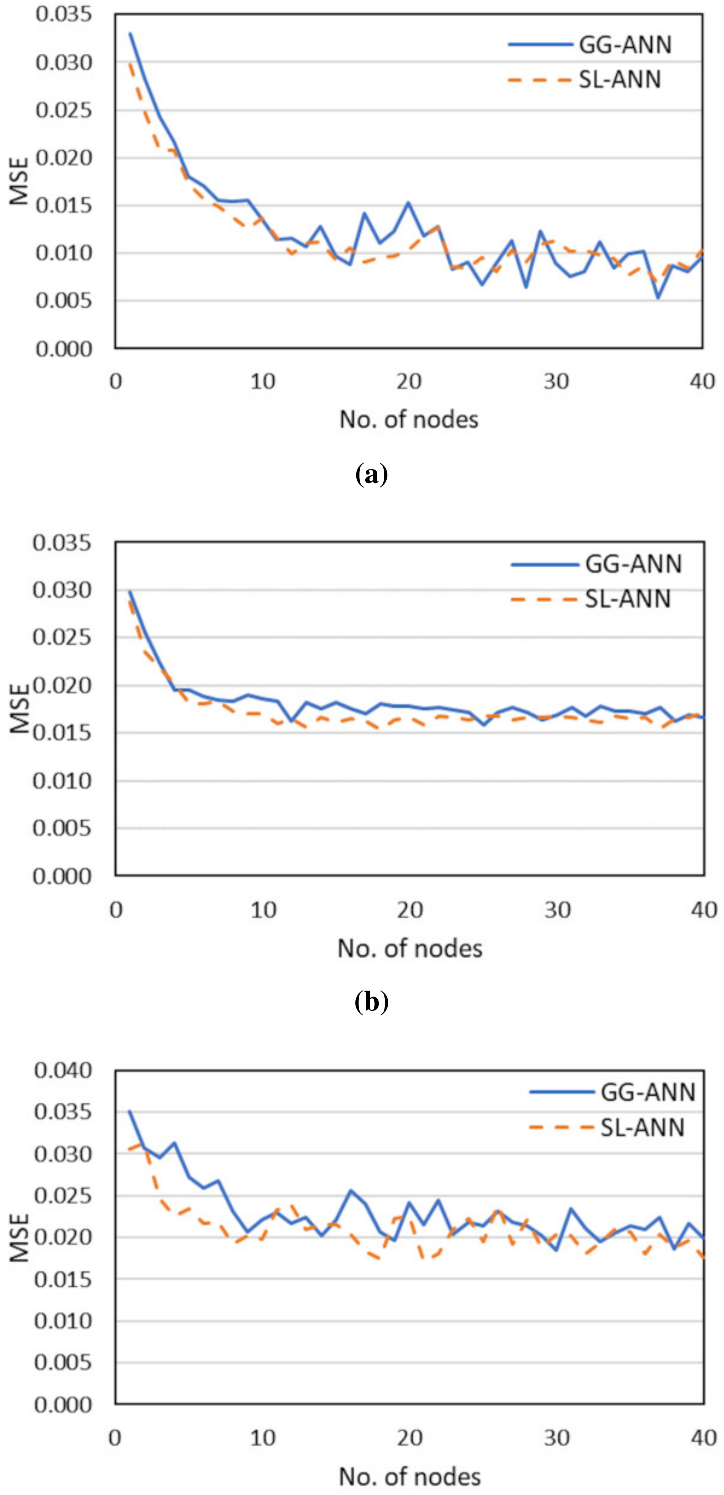


(a)

(b)

(c)

Fig. 4. Mean squared errors (MSEs) for the (4a) training, (4b) validation, and (4c) testing sets of GG-ANN and SL-ANN models versus number of hidden layer nodes for *Enterococci*, Realization 3.

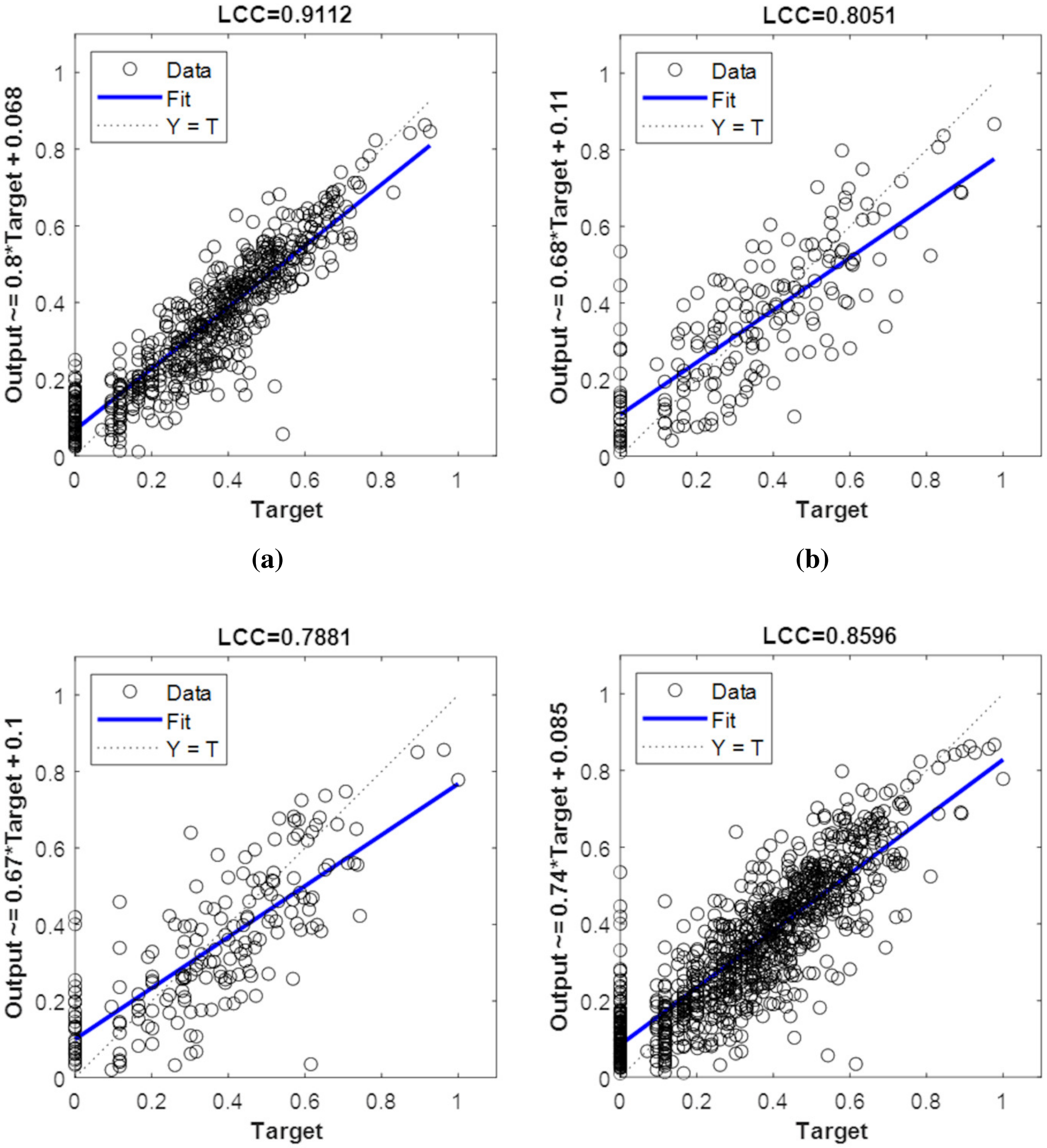


(a)

(b)

(c)

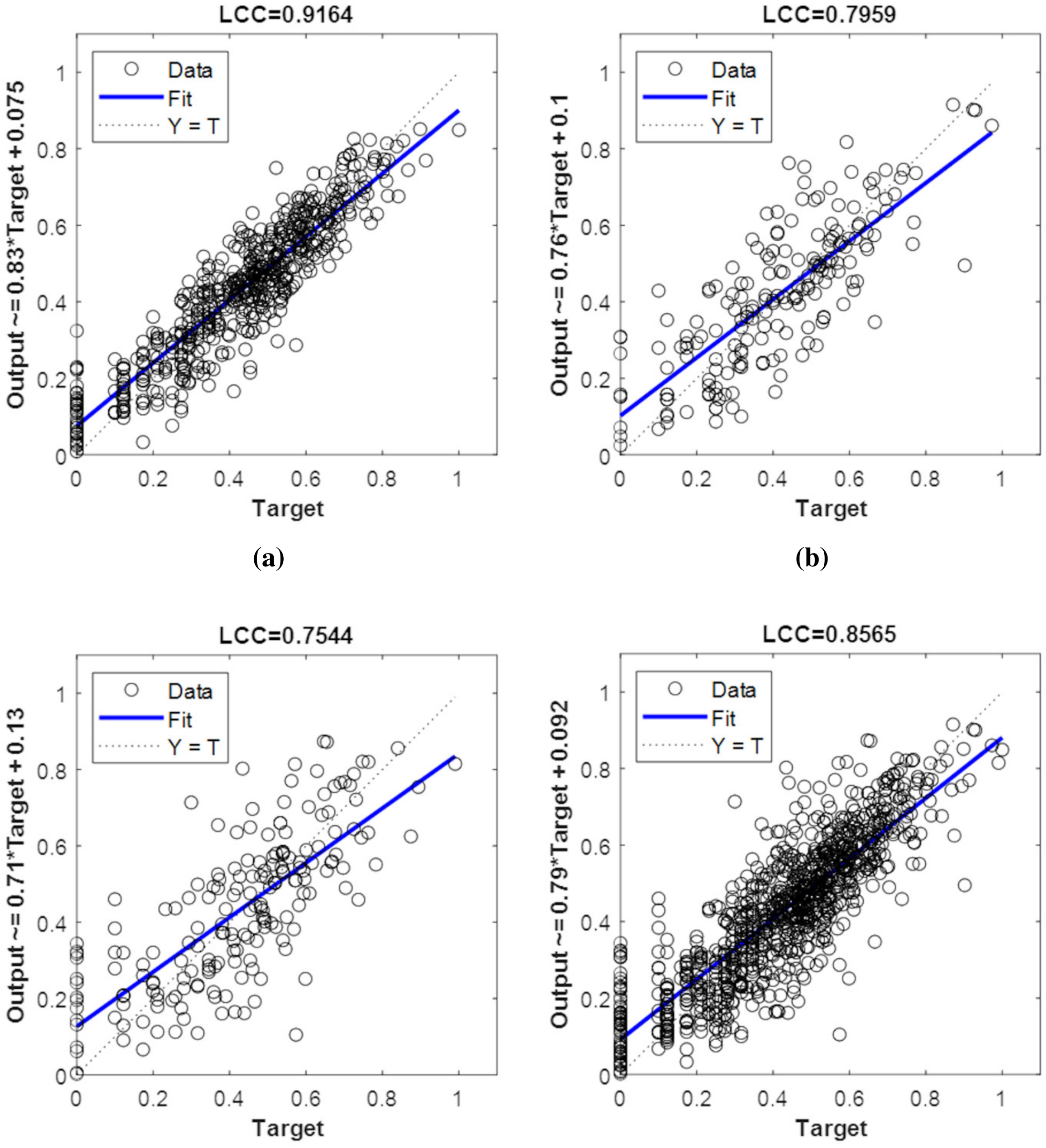
Fig. 5. Mean squared errors (MSEs) for the (5a) training, (5b) validation, and (5c) testing sets of GG-ANN and SL-ANN models versus number of hidden layer nodes for *E Coli*, Realization 1.



(a) **(b)**

(c) **(d)**

Fig. 6. Regressions between target *Enterococci* concentrations and GG-ANN model outputs for (6a) training; (6b) validation; (6c) testing; and (6d) all data sets, Realization 3. LCC: Linear correlation coefficient.



(a) **(b)**

(c) **(d)**

Fig. 7. Regressions between target *E Coli* concentrations and GG-ANN model outputs for (7a) training; (7b) validation; (7c) testing and (7d) all data sets, Realization 1. LCC: Linear correlation coefficient.

		Observed		
		Not poor	Poor	
Predicted	Not poor	XX	XX	XX% ← Predicted Value -
	Poor	XX	XX	XX% ← Predicted Value +
		XX%	XX%	XX% ← Overall Accuracy
		↑ Specificity (Illustration)	↑ Sensitivity	

		Observed		
		Not poor	Poor	
Predicted	Not poor	146	11	93%
	Poor	14	18	56%
		91%	62%	87%

(a)

		Observed		
		Not poor	Poor	
Predicted	Not poor	170	6	97%
	Poor	8	5	38%
		96%	45%	93%

(b)

		Observed		
		Not poor	Poor	
Predicted	Not poor	155	22	88%
	Poor	5	7	58%
		97%	24%	86%

(c)

		Observed		
		Not poor	Poor	
Predicted	Not poor	172	8	96%
	Poor	6	3	33%
		97%	27%	93%

(d)

		Observed		
		Not poor	Poor	
Predicted	Not poor	160	29	85%
	Poor	0	0	N/A
		100%	0%	85%

(e)

		Observed		
		Not poor	Poor	
Predicted	Not poor	178	11	94%
	Poor	0	0	N/A
		100%	0%	94%

(f)

		Observed		
		Not poor	Poor	
Predicted	Not poor	157	25	86%
	Poor	3	4	57%
		98%	14%	85%

(g)

		Observed		
		Not poor	Poor	
Predicted	Not poor	178	10	95%
	Poor	0	1	100%
		100%	9%	95%

(h)

Fig. 8. Performance tables for data-driven models; (left) *Enterococci*, Realization 1, testing sets; (right) *E Coli*, Realization 3, testing sets. (8a, 8b) GG-ANN models; (8c, 8d) SL-ANN models; (8e, 8f) GG-Linear models; (8g, 8h) SL-Linear models.