Comparison of Adaptive Neuro-Fuzzy Inference System (ANFIS) and Gaussian Process for Machine Learning (GPML) Algorithms for the Prediction of Norovirus Concentration in Drinking Water Supply

Hadi Mohammed[1], Ibrahim A. Hameed[2], Razak Seidu[1]

[1]Water and Environmental Engineering Group, Department of Marine Operations and Civil Engineering, Faculty of Engineering

[2]Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology (NTNU)
Postboks 1517, NO-6025 Ålesund, Norway

{hadi.mohammed, ibib, rase}@ntnu.no

Abstract. Monitoring of Norovirus in drinking water supply is a complicated, rather expensive, process. Norovirus represent a leading cause of acute gastroenteritis in most developed countries. Modeling of general microbial occurrence in drinking water is a very active field of study and provides reliable information for predicting microbial risks in drinking water. In this work, adaptive neuro-fuzzy inference system (ANFIS) and Gaussian Process for Machine Learning (GPML) are proposed as predicting models for the total number of Norovirus in raw surface water in terms of water quality parameters such as water pH, turbidity, conductivity, temperature and rain. The predictive models were based on data from Nødre Romrike Vannverk water treatment plant in Oslo, Norway. Based on the model performance indices used in this study, the GPML model showed comparable accuracy to the ANFIS model. However, the ANFIS model generally demonstrated more superior prediction ability of the number of Norovirus in drinking water, with lower MSE and MAE values relative to the GPML model. In addition, the ability of the ANFIS model to explain potential effects of interactions among the water quality variables on the number of Norovirus in the raw water makes the technique more efficient for use in water quality modeling.

Keywords. ***GPML, ANFIS, Norovirus, water pH, turbidity, water conductivity, temperature, conductivity, rain***

1 Introduction

Norovirus is increasingly recognized as a leading cause of non-bacterial gastrointestinal infections, and is a major cause of waterborne disease outbreaks worldwide [1, 2]. The virus is reported as the most common cause of diarrheal disease globally with an estimated economic burden of $4.2 billion in direct health system costs worldwide [3]. Moreover, enteric Noroviruses have very low infectious dose and like other viruses, their small size, low inactivation rates and the inability to culture them make their removal difficult and their susceptibility to disinfection largely remain unknown [4-6]. Further, results of some studies suggest that the viruses may be resistant to wastewater treatment since effluents are not completely devoid of the them [7 - 9], resulting in source water contamination when effluents of wastewater treatment plants are discharged to surface waterbodies [10-12].

Although recent molecular methods have improved the detection, identification and characterization of Norovirus in the environment and clinical samples, widespread emergence of the virus still present a challenge to the detection technique [13]. These complicate the health risks associated with the use of drinking and recreational water from contaminated water resources. Mitigating the morbidity and mortality associated with waterborne infections of Norovirus accordingly require proactive measures to augment rather costly monitoring exercises that assess microbial quality of drinking water sources. The dynamics of Norovirus occurrence in raw water sources depends on a complex interaction of different variables, including environmental factors (e.g. temperature, rainfall etc.) [15]. There is very little understanding of the environmental factors that significantly trigger the occurrence of Norovirus [15]. Environmental factors such as rainfall and temperature are associated with increased concentrations of indicator pathogen in surface water and are noted as potential predictors of increased source water pathogen concentration [16-17]. As with many water treatment plants worldwide, Norwegian water treatment plants (WTPs) do not monitor raw water sources for specific pathogenic organisms (including Norovirus) due to cost considerations.

Limitations regarding the lack of source water microbial quality data needed for microbial risk assessment have necessitated the use of mathematical models to predict the occurrence of pathogenic organisms in raw water sources. Predictive models, based on environmental and water quality parameters have been widely applied to improve the accuracy of raw water quality assessments, in order to assist watershed managers in making informed decisions regarding the protection of public health. Physically based techniques such as hydrodynamic models have been used to monitor microorganism generation, fate and transport in surface water sources [18-19].

Whereas data-driven techniques such as regression analysis have widely been used, other artificial intelligence (AI) techniques such as artificial neural network and adaptive neuro-fuzzy inference system (ANFIS) are recently being applied in predicting the concentration of microbial organisms in water sources [15,20]. In a recent study in Norway, Peterson et al (2016) applied a quantitative microbial risk approach to model the concentration of Norovirus in surface water based on *E. coli* and *C. perfringens* concentrations with assumptions regarding the source of fecal contamination [14].

Although poor correlations were found between the pathogen and indicator data, the approach provided an insight into potential level of Norovirus contamination in a typical surface water body in Norway [22]. In this paper, Adaptive Neuro-Fuzzy Inference System (ANFIS) is built to predict the concentration of Norovirus in the raw water source of the Nødre Romrike Water Treatment Plant in Oslo, based on measured rainfall in the catchment of the water supply system and water quality parameters such as water temperature, turbidity, conductivity and pH. In addition, Gaussian Processes for Machine Learning (GPML) modeling approach is used on the same dataset to predict the count of Norovirus in the raw water. The performances of the two modeling approaches are then compared using mean square prediction errors, mean absolute errors and the coefficient of multiple determination, R2 values. The paper is organized as follows: ANFIS model is presented in Section II. In Section III, the data and modelling approaches are presented. Results are presented in Section IV. In Section V, concluding remarks are drawn and suggestions for future work are presented.

2 Methodology

2.1 Dataset

This study built a model to accurately predict the concentration of Norovirus in relation to precipitation and physico-chemical characteristics of the raw water source of the Nødre Romrike Vannverk (NRV) water treatment plant in Oslo, Norway. NRV is one of the largest water treatment plants in Norway supplying 42000 m^3 of drinking water to six municipalities in Norway [16]. The plant depends on raw water from Glomma River, the largest river in Norway. Data used in the study are based on raw water samples at the intake of NRV from January 2011 to April 2012, covering the four main seasons in Norway. Sampling and analysis of raw water for Norovirus (GI and GII) was conducted under an EU project (VISK) [17]. For a detailed description of the analysis, recovery and assessment of Norovirus concentration in the raw water readers are referred to Grøndahl-Rosado et al. (2014) [18]. Data on precipitation was collected at a weather station located within the catchment of the raw-water intake while physical and chemical parameters data of the raw water were drawn from NRV database. The main physical and chemical parameters accounted for are water temperature (^oC), turbidity (NTU/mL), conductivity (µS/cm), rain (mm/day) and pH.

2.2 ANFIS model

ANFIS is a well-known artificial intelligence technique that has been used in hydrological processes [18]. With respect to water quality monitoring, the technique has been widely used to model treatment processes, estimation of concentrations of disinfection byproducts as well as other water quality indices of groundwater [22 - 24]. By analyzing mapping relationships between input and output data, ANFIS optimizes the distribution of membership functions by using a hybrid learning algorithm consists of a combination of least-squares and back-propagation gradient descent algorithm [18]. In this paper two membership functions are assigned for each input variable and therefore

the ANFIS model will generate 64 rules (i.e., 26 rules). The proposed ANFIS (Figure 1) has six inputs; pH, turbidity, conductivity, rain, temperature and seasonality and one output, the concentration of Norovirus. Each input is represented by two fuzzy sets, and the output by a first-order polynomial of the inputs. The ANFIS extracts *n* rules mapping the inputs to the output from the input/output dataset. A typical Sugeno-fuzzy rule can be expressed in the following form:

$$\begin{aligned} Ri: IF\ & x_1\ is\ A_{1,j} \\ & AND\ x_2\ is\ A_{1,j} \\ & \vdots \\ & AND\ x_m\ is\ A_{m,j} \\ & THEN\ y_i = f_i(x_1, x_2, ..., x_m) \end{aligned} \tag{1}$$

where $x1$, $x2$, ..., x_m are the input variables, $A_{1,j}$, $A_{2,j}$, ..., $A_{m,j}$ are fuzzy sets or fuzzy labels used to fuzzifiy each input, y_i (i.e., Norovirus count of rule *i*) is either a constant or a linear function of the input variables of the model. When y_i is constant, a zero-order Sugeno fuzzy model is obtained in which the consequent of a rule is specified by a singleton. When y_i is a first-order polynomial of the inputs, the consequent of a rule is a polynomial that takes the form:

$$y_i = k_{i0} + k_{i1}x_1 + k_{i2}x_2 + \ldots + k_{im}x_m \tag{2}$$

A first-order Sugeno fuzzy model is obtained where k_{i0}, k_{i1}, k_{i2},…, and k_{im} are a set of parameters specified for rule *i* [25]. An ANFIS model is normally represented by a six-layer feed-forward neural network representing the architecture of a first-order Sugeno fuzzy model. The first layer is called input layer. Neurons in this layer simply pass external crisp signals to the second layer. The second layer is called the fuzzification layer. Neurons in this layer perform fuzzification. Fuzzification neurons have a bell-shaped activation function specified as:

$$y_i^{(2)} = \frac{1}{1 + \left(\frac{x_i^{(2)} - a_i}{c_i} \right)^{2b_i}} \tag{3}$$

where x_i is the input and y_i is the output of layer 2; a_i, b_i and c_i are the parameters that control, respectively, the center, width and slope of the bell activation function of neuron *i*. The third layer is called the rule layer. Each neuron in this layer corresponds to a single Sugeno-type fuzzy rule, as it is shown by Eq. (1). A rule neuron receives inputs from the respective fuzzification neurons and calculates the firing strength of the rule it represents. In an ANFIS, the product operator is used to evaluate the conjunction of the rule antecedents:

$$y_i^{(3)} = \prod_{j=1}^{k} x_{ij}^{(3)} = \mu_{i1} \times \mu_{i2} \times \cdots \times \mu_{ik} \tag{4}$$

where x_{ij} is the input from neuron j in layer 2 to neuron i in layer 3, and y_i is the output of layer 3.

Layer 4 is called the *normalization layer*. Each neuron in this layer receives inputs from all neurons in the rule layer, and calculates the normalized firing strength of a given rule. The normalized firing strength is the ratio of the firing strength of a given rule to the sum of the firing strengths of all rules. It represents the contribution of a given rule to the final result. The output of neuron i in layer 4 is obtained as:

$$y_i^{(4)} = \frac{x_{ij}^{(4)}}{\sum_{j=1}^{n} x_{ij}^{(4)}} = \frac{\mu_i}{\sum_{j=1}^{n} \mu_j} = \overline{\mu}_i \tag{5}$$

where x_{ij} is the input from neuron j in layer 3 to neuron i in layer 4, and y_i is the output of layer 4, and n is the total number of fuzzy rules. Layer 5 is called the defuzzification layer. Each neuron in this layer is connected to the respected normalization neuron in the normalization layer, and also receives initial inputs; x_1, x_2,…, x_m. A defuzzificaiton neuron calculates the weighted consequent value of a given rule as:

$$\begin{aligned} y_i^{(5)} &= x_i^{(5)} \left(k_{i0} + k_{i1} x_1 + \ldots + k_{im} x_m \right) \\ &= \overline{\mu}_i \left(k_{i0} + k_{i1} x_1 + \ldots + k_{im} x_m \right) \end{aligned} \tag{6}$$

where x_i is the input and y_i is the output of neuron i in layer 5, k_{i0}, k_{i1}, …, and k_{im} is a set of consequent parameters of rule i, defined by Eq. (2). Layer 6 is finally represented by a single summation neuron. This neuron calculates the sum of outputs of all defuzzification neurons and produces the overall ANFIS output y.

It is not necessary to have any prior knowledge of rule consequent parameters for an ANFIS to deal with the problem. The parameters of the consequent polynomials are initialized to zero values. ANFIS uses a hybrid-learning algorithm that combines the least-squares estimator and the gradient descent method to learn parameters of the consequent polynomials and to tune the parameters of the membership functions. The only prior information required from the user is the number of membership functions required to fuzzify each input variable. The universe of discourse of each input variable is divided equally between its respective membership functions to find its centers. The widths and slopes are set to allow sufficiently overlapping between the respective functions.

In the ANFIS training algorithm, each epoch is composed from a forward pass and a backward pass. In the forward pass, a training set of input patterns is applied to the ANFIS, neuron outputs are calculated on the layer-by-layer basis, and the least-squares estimator identifies rule consequent parameters. In the backward pass, the

error signals are propagated back and the back-propagation algorithm is used to update/tune the membership functions parameters of the rule antecedents.

3 GPML model

Gaussian processes (GPs) [32] have convenient properties for many modeling tasks in machine learning and statistics. They can be used to specify distributions over functions without having to commit to a specific functional form. Applications range from regression over classification to reinforcement learning, spatial models, survival and other time series models. A GP is specified by a mean function and a covariance function. These functions are mostly difficult to specify fully a priori, and typically they are given in terms of hyperparameters, that is, parameters which have to be inferred. Any functions can be used as a mean and a covariance functions. The mean function is usually defined to be zero. Several covariance functions have been used in literature. However, a squared exponential (SE) is usually used as a predominant choice. Another source of difficulty is the likelihood function. For Gaussian likelihoods, inference is analytically tractable; however, in many tasks, Gaussian likelihoods are not appropriate, and approximate inference methods such as Expectation Propagation (EP), Laplace's approximation (LA) and variational bounds (VB) become necessary [33].

GPML, in this paper, is implemented using GPML toolbox [34]. The GPML toolbox provides a wide range of functionality for GP inference and prediction. It is designed to simplify the process of constructing GP models and make it easy to extend where a library of various mean, covariance and likelihood functions as well as inference methods is provided.

4 Modelling Approach

4.1 ANFIS Model

In this paper, various ANFIS models with various settings are used to predicted count of Norovirus in raw water in terms of a set of input variables: water pH, water turbidity, water conductivity, rain, water temperature, and finally seasonal effect (i.e., winter time).The model building process for ANFIS consists of the following five steps: 1). Selection of the input and the output data for training ANFIS model (data set). 2). Normalization of the input and the output data attributes. 3). Training of the normalized data using a hybrid-learning algorithm; 4).Testing the goodness of fit of the model; and 5. Comparing the predicted output with the desired/target output. Each of these steps are presented as follows:

4.2 Selection of the input and output data for tarning ANFIS model

In this paper, an ANFIS model is developed to accurately predict the concentration of Norovirus in terms of precipitation and physico-chemical characteristics of the raw water source of the Nødre Romrike Vannverk (NRV) water treatment plant in Oslo, Norway. NRV is one of the largest water treatment plants in Norway supplying 42000 m3 of drinking water to six municipalities [19]. The plant depends on the raw water from Glomma River, the largest river in Norway. Data used in the study are based on raw water samples at the intake of NRV in the period from January 2011 to April 2012, covering the four main seasons in Norway. Sampling and analysis of raw water for Norovirus (GI and GII) was conducted under an EU project (VISK) [27]. For a detailed description of the analysis, recovery and assessment of Norovirus concentration in the raw water interested readers are advised to refer to Grøndahl- Rosado et al. (2014) [28]. Data on precipitation was collected at a weather station located within the catchment of the raw- water intake (e-Klima) while physical and chemical parameters data of the raw water were drawn from NRV database. The main physical and chemical parameters accounted for were water temperature (oC), water turbidity (NTU/mL), water conductivity (μS/cm), rainfall (mm/day), and water pH. A total of 156 data samples are used in this study.

4.3 Data normalization

The input and the output data obtained are measured on different scales and therefore have to be normalized using mean and standard deviation to a notionally common scale. First the mean, x, and standard deviation, σ_x , of all the data variables individually were calculated. The values for each parameter were then normalized using the equation:

$$x_n = (x - \overline{x}) / \sigma_x \tag{7}$$

4.4 Training of input data

After obtaining the normalized data, the next step is to train the input data using proposed ANFIS . ANFIS model uses a hybrid-learning algorithm that combines the least- squares estimator and the gradient descent method to learn parameters of the consequent polynomials and to tune the parameters of the membership functions. The algorithm, by default, takes only 70 percent of the input data for training. So out of 156 samples only 109 are taken for training and these are selected randomly from the set of data. The rest 47 samples are kept for validation and testing.

4.5 Testing and validation

Testing is done after the training of the data is complete and the error is below the tolerance levels. 30 % of the input data are used for testing and validation in both cases (i.e., 47 samples).

5 GPLM Model

The proposed GPML model has six inputs; water pH, turbidity, conductivity, rain, temperature and seasonality. The GPML toolbox in MATLAB is used [24] for constructing a GP predictive model for Norovirus using the aforementioned six descriptive features. GPs are used to formalize and update knowledge about distribution over functions. To set up a GP model, a mean function with an initial value of mean=0 is chosen. A squared exponential covariance function with a hyper-parameters $\psi = \{1, 2\}$ is used. A Gaussian likelihood function with hyperparameter (i.e., Gaussian noise with variance ρ) where $\rho = \{1\}$ is used. An expectation propagation (EP) approximate inference algorithm used.

5.1 Comparisons of actual data and predicted data

After the testing is done, the ANFIS and the GPML models are saved. The mean absolute error (MAE) and mean- squared error (MSE) between actual and predicted outputs and the coefficient of determination, R^2 are used as performance indices of the models' accuracy. A graph is plotted between the actual output and the predicted output so that a comparison can be easily made.

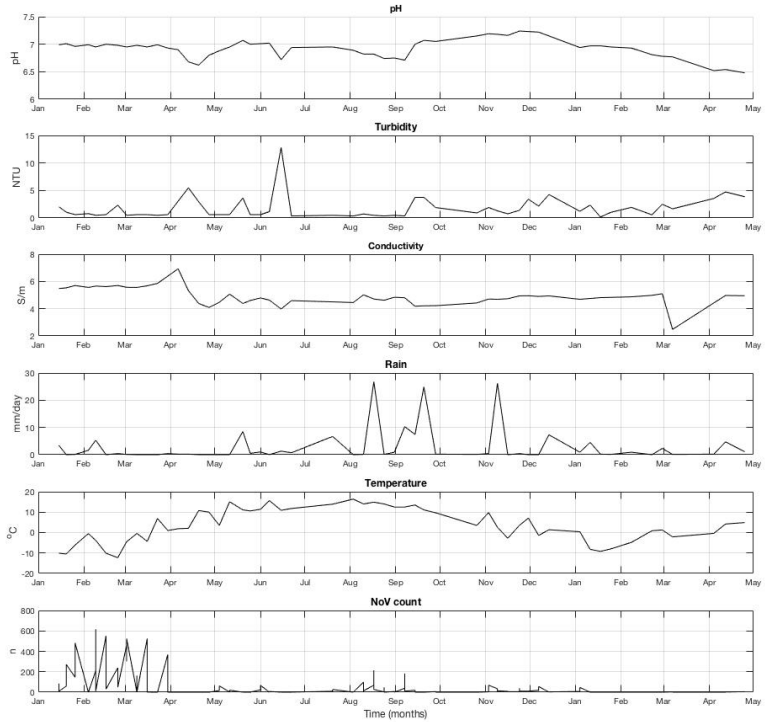

Fig. 1. Raw data from NRV in the period from middle of January 2011 until end of April 2012.

6 Results

Fig. 1 shows the distribution of the water quality parameters and their influence on Norovirus concentration. The measured Norovirus concentrations had large variations. While variations in the water pH and electrical conductivity remain low, the range of variations in measured values of water temperature, rainfall and turbidity remained high. For the months in which rainfall over the study area was high (mostly between July and December), the water temperature was continuously high. However, the turbidity level in the raw water reached its peak of 13 NTU in in the middle of June, just prior to the onset of elevated rainfall in mid-July. High Norovirus concentrations over the sixteen-month study period, the observed Norovirus concentrations were very high between January and April of 2011 (500 particles per liter), with intense variations. Subsequently however, few Norovirus particles are observed intermittently.

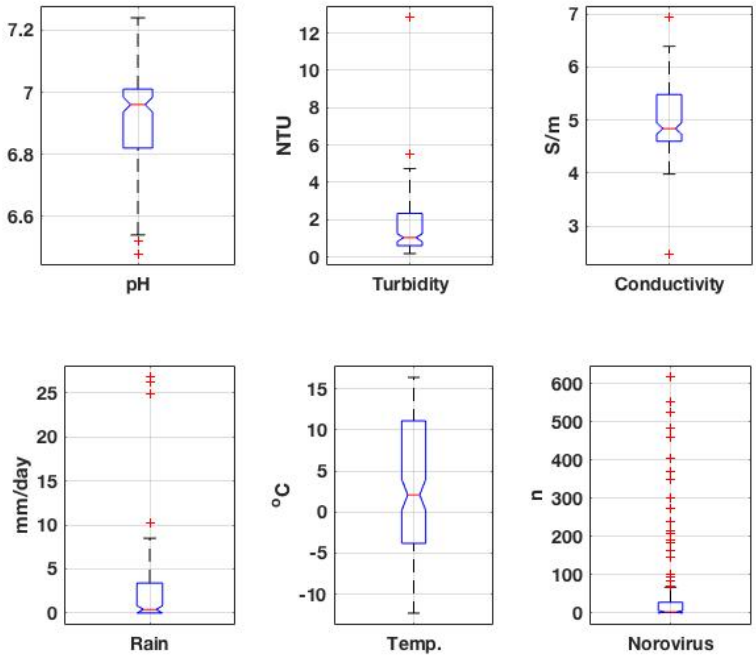

Fig. 2. Boxplot of the raw data

Fig. 2 shows the box plot of the raw water quality parameters showing the medians, minimum, and maximum values of each individual variable. The correlation coefficient between the dependent (i.e., count of Norovirus) variable and independent variables (i.e., model inputs) are shown in Table 1. From the table, it can be concluded that the observed number of Norovirus is not significantly correlated with raw water pH, rainfall and season type (i.e., winter season).

Table 1. Correlation and dependance between dependant and independent variables

Parameter	Correlation	P-value
pH	0.0383	0.6346
Turbidity	**-0.2084**	**0.0090**
Conductivity	**0.3908**	**0.0000**
Rain	-0.0426	0.5975
Temperature	**-0.3048**	**0.0001**
Winter season	0.1057	0.1889

Table 2. $ANFIS_{6,2}$ generalized fuzzy rules after 250 epochs of training

Rule	Rule's description
1	**IF** *pH* is low **AND** *turbidity* is low **AND** *conductivity* is low **AND** *rain* is low **AND** temperature is low **AND** winter is low **THEN** *y*=0.43+0.16*pH*-0.23*turbidity*+0.06conductivity-0.39rain+0.40temperature-0.56winter
2	**IF** *pH* is low **AND** *turbidity* is low **AND** *conductivity* is low **AND** *rain* is low **AND** temperature is low **AND** winter is high **THEN** *y*=0.54*pH*-0.25*turbidity*+0.36conductivity+0.60rain+0.09temperature+0.07winter
…	**…**
64	**IF** *pH* is high **AND** *turbidity* is high **AND** *conductivity* is high **AND** *rain* is high **AND** temperature is high **AND** winter is high **THEN** *y*=0.92+0.91*pH*-0.06*turbidity*+0.51conductivity-0.13rain+0.96temperature-0.69winter

6.1 Prediction of NoV using GPML

The response of the GPML model is plotted against the normalized measured Norovirus concentration in the raw water, as it is shown in Fig. 3 (upper) while prediction error is shown in Fig. 3 (bottom). The mean absolute and mean squared prediction errors are 0.3614 and 0.4179, respectively.

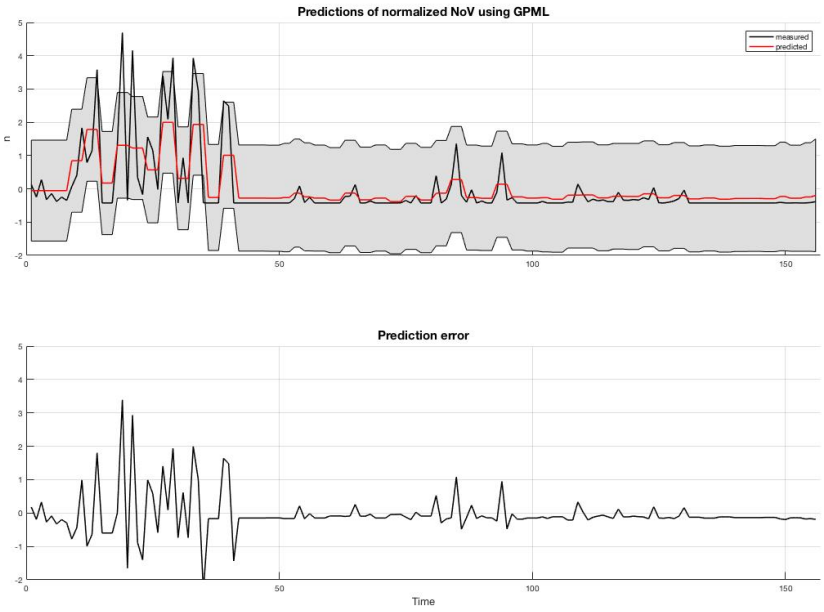

Fig. 3. Response of GPML model in red plotted against the normalized measured Norovirus concentration in raw water source while the marginal likelihood is shown in gray color (upper) and prediction error (bottom).

6.2 Prediction of NoV using ANFIS

Various settings of ANFIS model are used for predicting NoV count as follows:

6.2.1 6 inputs 2 MFs ANFIS model (ANFIS6,2).

$ANFIS_{6,2}$ has six inputs and 2 MFs namely *low* and *high* to fuzzy each crisp input; pH, turbidity, conductivity, rainfall, temperature, and time accounting for seasonality and one output, the concentration of Norovirus.

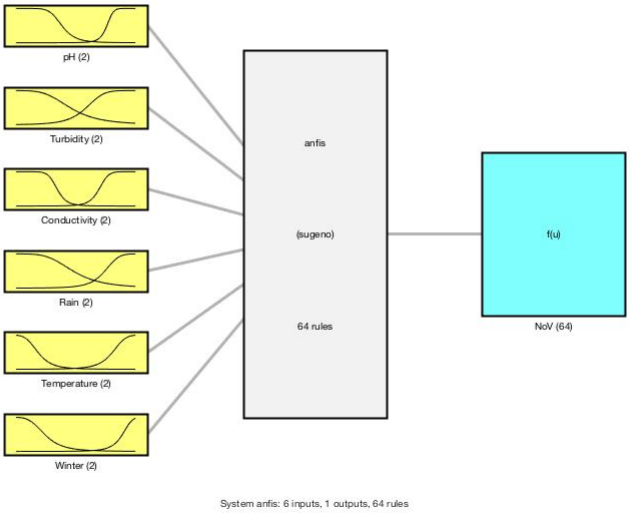

Fig. 4. ANFIS6,2 for predicting Norovirus count, y, in terms of pH, turbidity, etc. as the model inputs.

In this model, each input is represented by two fuzzy sets, and the output is represented as a first-order polynomial of the inputs. The ANFIS extracts $r=2^6=64$ rules mapping the inputs to the output from the input/output observations. The proposed $ANFIS_{6,2}$ is shown in Fig. 4. Initial and final MFs for the input variables are shown in Fig. 5 and 6, respectively, where the best result was obtained after 250 epochs with a MSE=0.3567 for the normalized output. A sample of the generated rules is shown in Table II.

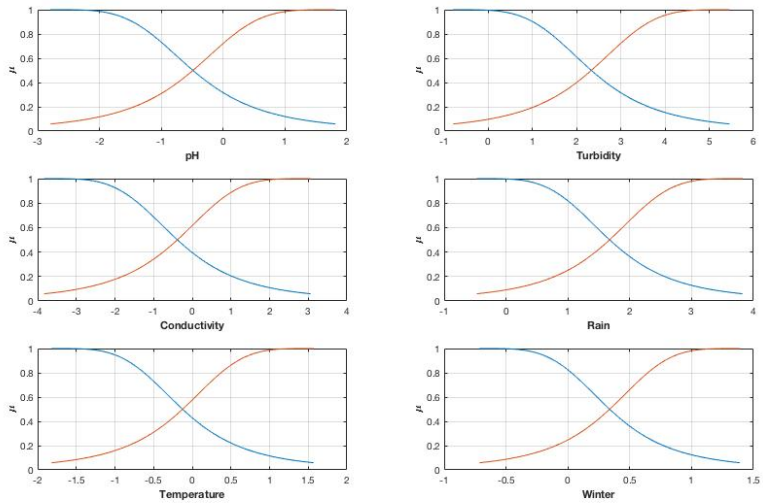

Fig. 5. Initial MFs of the proposed ANFIS6,2 model (2 MFs are used for fuzzifying each input).

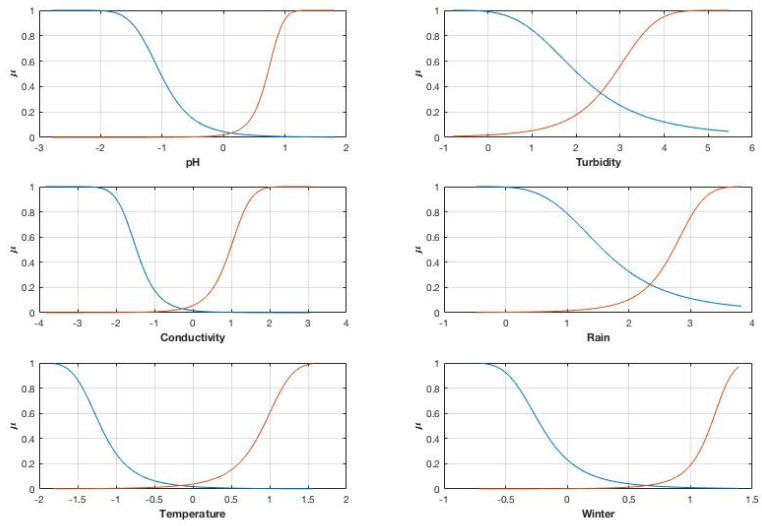

Fig. 6. Final MFs of the proposed ANFIS6,2 model after 250 epochs of training.

To assess the predictability of the proposed model, the response of the $ANFIS_{6,2}$ model is plotted against the normalized measured Norovirus concentration in the raw water, as it is shown in Fig. 7 (upper) while prediction error is shown in Fig. 7 (bottom). The model was capable of adequately predicting periods in which counts of Norovirus were observed in the raw water as well as periods of no counts. More importantly, the model efficiently predicted periods of intense variations in the counts of the virus in raw water (the first three months of the study period). This is a necessary information for the optimization of water treatment processes in order to prevent potential waterborne illnesses. In addition, to examine how interactions among the water quality variables affect the level of Norovirus in the raw water, surface view of the input-output mapping were generated as shown in Fig. 8. It is evident from these plots that the influence of each water quality parameter on the virus differs when it interacts with a different variable. For instance, while high pH at high temperature are associated with increased counts of the virus in the raw water, elevated water turbidity occurring at high pH results in lower counts. Similarly, high turbidity result in increasing the level of the virus when conductivity is also high. However, for the same turbidity level, increasing temperature results in lower counts of the virus in the raw water. Finally, interactions among certain pairs of variables (such as between turbidity and pH, turbidity and conductivity) appear to have higher impact on the number of Norovirus than interactions among other pairs (eg, conductivity and pH).

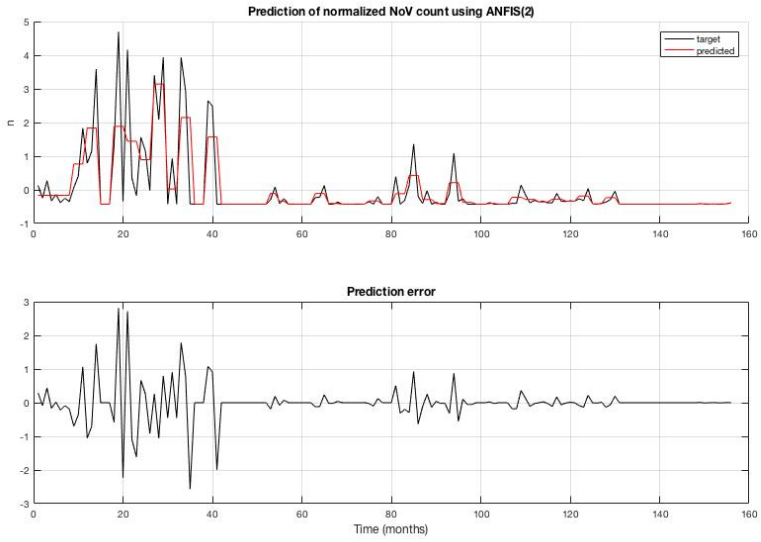

Fig. 7. Response of ANFIS6,2 model and its mean squared prediction error.

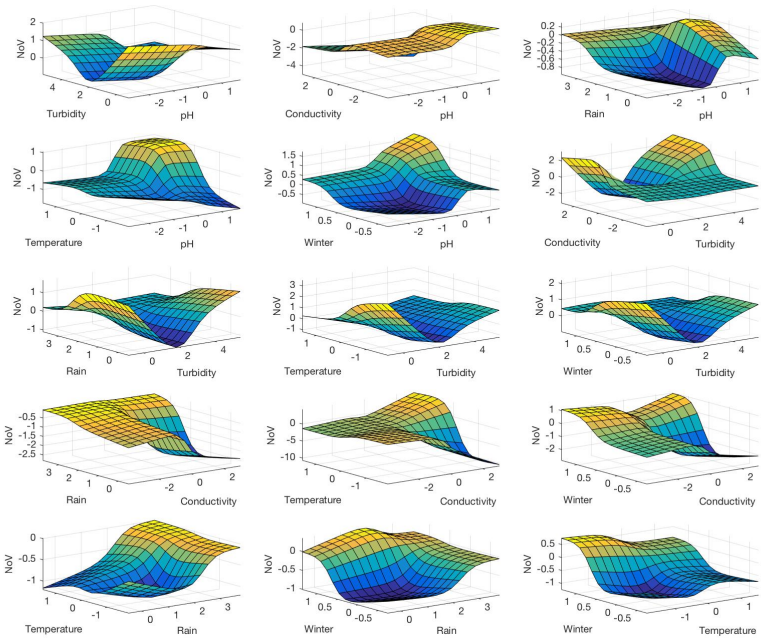

Fig. 8. Surface view of the generated $ANFIS_{6,2}$ fuzzy relations showing the effect of each input variable to the count of the predicted NoV.

6.2.2 3 **inputs** 2 MFs ANFIS model (ANFIS3,2,COR)

A reduced set of the input features; turbidity, conductivity, and temperature are used to train this model as they are more relevant to the concentration of Norovirus, as it was shown in the correlation coefficients in Table I and 2 MFs namely *low* and *high* to fuzzy each crisp input is developed. The proposed $ANFIS_{3,2,COR}$ model is shown in Fig. 8. After 250 epochs, the model was able to extract 2^3=8 rules with prediction error MSE =0.3735. The structure of the developed $ANFIS_{3,2,COR}$ model using two MFs and 8 rules is shown in Fig. 9.

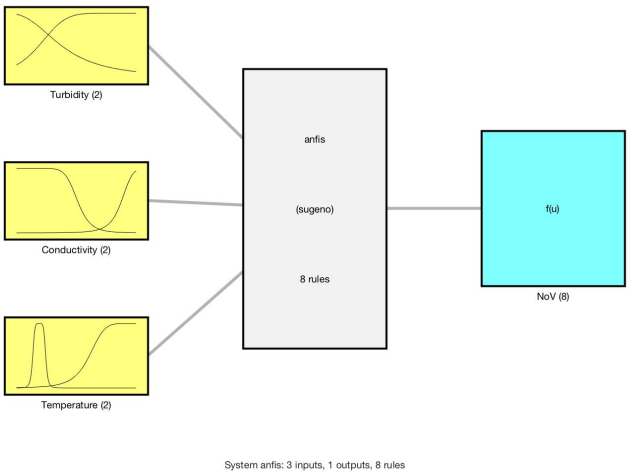

Fig. 9. ANFIS3,2,COR for predicting Norovirus count, Nov, in terms of turbidity, conductivity and temperature.

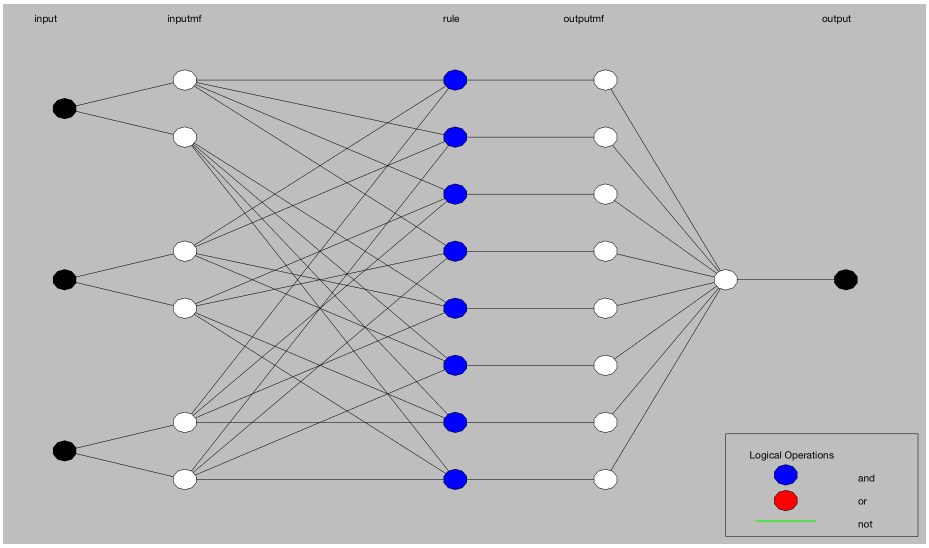

Fig. 10. ANFIS model structure for three inputs and 2 MFs each.

The model predictions versus target values are shown in Fig. 11 (upper) while prediction error is shown in Fig. 11 (bottom). Surface view of the input-output mapping of this model is shown in Fig. 12. Here, the effects of the interactions among the water quality parameters found to be more significantly correlated with the count of Norovirus (turbidity, conductivity, temperature) are more distinct. For instance, a sharp

drop in the count of the virus resulting from a combined increases in water turbidity (> 2 NTU) and conductivity (> 1 µS/cm) can be seen in Fig. 12.

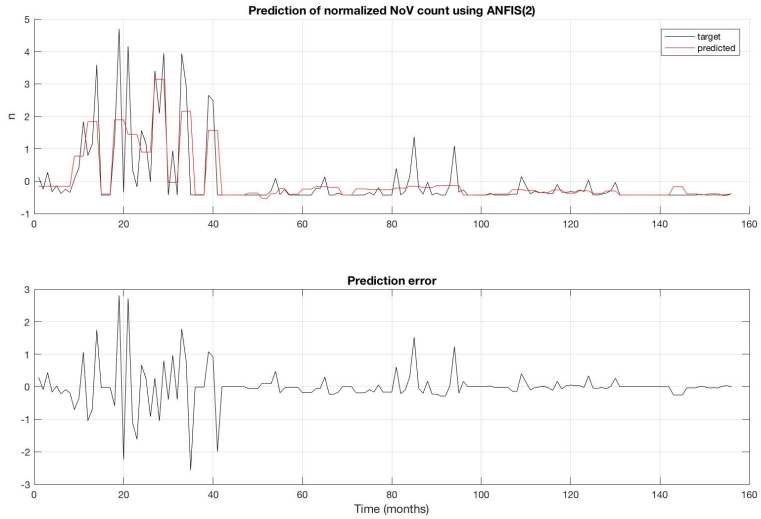

Fig. 11. Response of $ANFIS_{3,2,COR}$ model and the mean squared error.

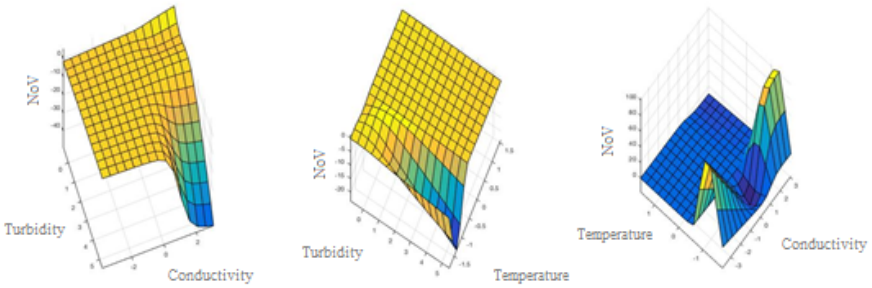

Fig. 12. Surface view of the generated $ANFIS_{3,2,COR}$ fuzzy relations showing the effect of each input variable to the count of the predicted NoV.

6.2.3 3 inputs 2MFs ANFIS model (ANFIS3,2,PCA)

A reduced model using three input features obtained using principal component analysis (PCA) and 2 MFs namely *low* and *high* to fuzzy each crisp input is developed. After 250 epochs, the model was able to extract 2^3=8 rules with prediction error MSE

=0.3623. The model predictions versus target values are shown in Fig. 13 (upper) while prediction error is shown in Fig. 13 (bottom). Surface view of the input-output mapping of this model is shown in Fig. 14.

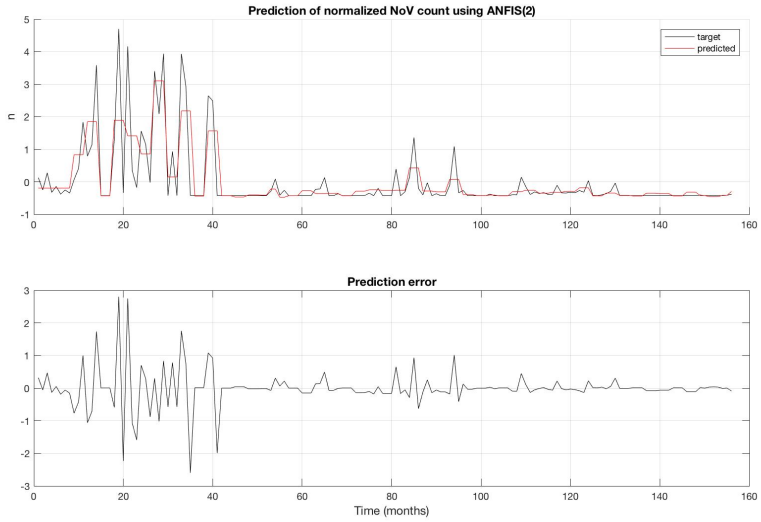

Fig. 13. Response of $ANFIS_{3,2,PCA}$ model and the mean squared error.

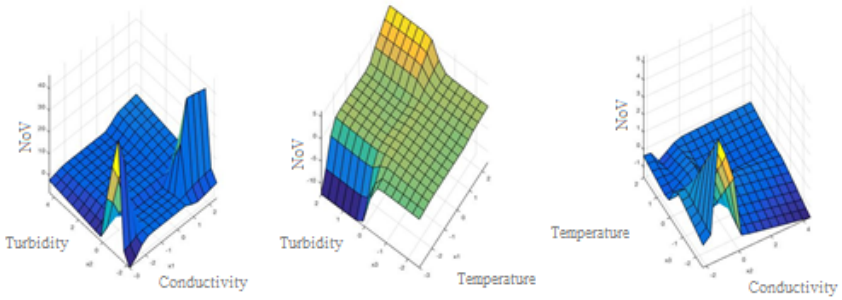

Fig. 14. Surface view of the generated $ANFIS_{3,2,PCA}$ fuzzy relations showing the effect of each input variable to the count of the predicted NoV.

6.2.4 6 inputs 3 MFs ANFIS model (ANFIS6,3)

An ANFIS model with 6 inputs and 3 MFs, namely, *low*, *medium*, and *high* for fuzziying each input variable is developed. After 250 epochs, the model generated 36=729 rules and a prediction error MSE=0.3567.

6.3 Model comparison

In this study, both the GPML and ANFIS models of carious structures were trained on the entire dataset and tested for predicting the concentration of Norovirous in raw water source using climate and water quality parameters that are measured at water treatment plants in real time. The performances of both models were compared by the mean absolute error (MAE), the mean squared error (MSE), and R^2 criteria are shown in Tables 3 and 4. The MAE indicates how close the predictions are to the measured values which is given by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right| \tag{8}$$

As seen in Equation (8), the mean absolute error can be defined as the average of absolute errors; the absolute error given by $|e_i| = |f_i - y_i|$, where f_i is the prediction and y_i the true value. It should be noted that in MAE, all the individual errors have equal weight in the average, making it a linear score. In order to have a reliable statistical comparison between the mathematical models, both the MAE and MSE can be used together to ascertain the variation in errors in a given set of predictions. Calculation of MSE involves squaring the difference between the predicted and corrsponding observed values, and averaging it over the sample size. This can be written as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}e_i^2 \tag{9}$$

MSE has a quadratic error rule, where the errors are squared before being averaged. As a result, a relatively high weight is given to large errors. This could be useful when large errors are undesirable in a statistical model. From table 3 it can be deduced that for the Gaussian model the MSE is slighly higher as compaed to ANFIS. Another measure of goodness-of-fit of the model is the R^2 criteria. R^2 is a proportion of variance '*explained*' by the model. Higher values are indicative that the predictive model fits the data in a better way. By definition, R^2 is the proportional measure of variance of one variable that can be predicted from the other variable. Thus ideally the values of R^2 to approach one is always desirable. However, a high R^2 tells you that the curve came very close to the points but in reality it does not always indicate the model quality. From Table 4, both Gaussian and ANFIS models have similar R^2 values which indicate that in both modeling techniques, the prediction capability is similar. However, using the R^2 criteria in conjunction with the MAE and MSE, it can be fairly deduced that the Gaussian and ANFIS models can be accurately used for the prediction of norvirous concentration in raw water source.

Table 3. Performance comparison of GPML and ANFIS models for the test data set.

Model	Inputs	MFs	Rules	MAE	MSE
GPML	-	-	-	0.3614	0.4179
$ANFIS_{6,3}$	6	3	729	0.2704	0.3567
$ANFIS_{6,2}$	6	2	64	0.2704	0.3567
$ANFIS_{3,2,PCA}$	6	2	8	0.2917	0.3623
$ANFIS_{3,2,COR}$	3	2	8	0.2990	0.3735
	3				

Table 4. R^2 criteria comparison of GPML and ANFIS models.

Modelling technique	R^2
GPML	0.7802
$ANFIS_{6,3}$	0.8006
$ANFIS_{6,2}$	0.8006
$ANFIS_{3,2,PCA}$	0.7971
$ANFIS_{3,2,COR}$	0.7900

7 Conclusions

ANFIS and GPML models for prediction of the counts of Norovirus in raw water have been developed and the predictive abilities of the two models compared. Both modeling approaches have demonstrated adequate performances in their predictions during the model testing stage. In terms of the mean square error and mean absolute error values of the model predictions, the ANFIS models showed considerable accuracy relative to the GPML model. However, the computed R^2 values of the models indicate that no distinct disparity exists between the model performances. However, certain features of the ANFIS modelling approach make it more efficient for application. For instance, the advantages of ANFIS compared to other black box models include: (1) it combines in a transparent manner the linguistic representations of fuzzy logic and the learning capabilities of artificial neural networks, (2) it provides an automated approach for rule generation and parameter optimization procedure that simplifies the complex process of model development, and finally (3) it creates a transparent solution that is expected to offer useful insights into the physical processes involved in the modeling process and therefore help its end user the understand why certain values are obtained.

PCA in this paper is used to provide a reduced set of input features where the most influential three variables to NoV count are chosen. Although the ANFIS model trained using this reduced set of variables is able to provide better predictions, it lacks the ability to provide a meaning to the input-output mapping compared to the ANFIS model trained using the reduced set obtained from correlation coefficients table. It is also noted that using more than 2 MFs does not have any significant effect on the predictability of the produced ANFIS model. For future work, new models and new data sets will be used for providing more accurate predictions of NoV count in terms of climate and water quality parameters.

8 Acknowledgements

The authors wish to thank the managers of the Nødre Romrike Water Treatment Plant in Oslo for the provision of required data. Thanks to Ricardo Rosado and Mette Myrmel for providing the Norovirus data. This work is part of the project KLIMAFORSK funded by the Research Council of Norway (Project No: 244147/E10). The authors would like to express their sincere thanks to the editor and anonymous reviewers for their suggestions and comments to improve the quality of the paper.

9 References

1. S. M. Ahmed, A. J. Hall, A. E. Robinson, L. Verhoef, P. Premkumar and U. D. Parashar, "Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis," *The Lancet Infectious Diseases,* vol. 14, no. 8, 2014, pp. 725-730.
2. G. M. Brion, T. R. Neelakantan and S. Lingireddy, "Using neural networks to predict peak Cryptosporidium concentrations," Journal of American Water Works Association (AWWA), vol. 93, no. 1, 2001, pp. 99-105.
3. S. M. Bartsch, B. A. Lopman, S. Ozawa, A. J. Hall and B. Y. Lee, "Global Economic Burden of Norovirus Gastroenteritis," PLoS ONE, vol. 11, no. 4, 2016, e0151219. doi: 10.1371/journal.pone.0151219.
4. Z. Altintas, M. Gittens, J. Pocock, and I. E. Tothill, "Biosensors for waterborne viruses: Detection and removal," Biochimie, vol. 115, 2015, pp. 144-154.
5. K. R. Wigginton, and T. Kohn, "Virus disinfection mechanisms: the role of virus composition, structure, and function," Current opinion in virology, vol. 2, no. 1, 2012, pp. 84-89.
6. I. Xagoraraki, Z. Yin and Z. Svambayev, "Fate of viruses in water systems," Journal of Environmental Engineering, vol. 140, no. 7, 2014, pp. 04014020-18.
7. A. K. da Silva, J. C. Le Saux, S., Parnaudeau, M. Pommepuy, M. Elimelech, & F. S. Le Guyader, (2007). Evaluation of removal of noroviruses during wastewater treatment, using real-time reverse transcription-PCR: different behaviors of genogroups I and II. *Applied and Environmental Microbiology*, *73*(24), 7891-7897.
8. M. A. Laverick, A. P. Wyn-Jones, & M. J. Carter, (2004). Quantitative RT-PCR for the enumeration of noroviruses (Norwalk-like viruses) in water and sewage. *Letters in Applied Microbiology*, *39*(2), 127-136.
9. T. Westrell, P. Teunis, H. van den Berg, W. Lodder, H. Ketelaars, T. A. Stenström, & A. M. de Roda Husman, (2006). Short-and long-term variations of norovirus concentrations in the Meuse river during a 2-year study period. *Water research*, *40*(14), 2613-2620.
10. M. Barrett, K. Fitzhenry, V. O'Flaherty, W. Dore, S. Keaveney, M. Cormican, ... & E. Clifford, (2016). Detection, fate and inactivation of pathogenic norovirus employing settlement and UV treatment in wastewater treatment facilities. *Science of the Total Environment*, *568*, 1026-1036.
11. W. J. Lodder, & A. M. de Roda Husman, (2005). Presence of noroviruses and other enteric viruses in sewage and surface waters in The Netherlands. *Applied and Environmental microbiology*, *71*(3), 1453-1461.
12. Y. Ueki, D. Sano, T. Watanabe, K. Akiyama, & T. Omura, (2005). Norovirus pathway in water environment estimated by genetic analysis of strains from patients of gastroenteritis, sewage, treated wastewater, river water and oysters. *Water research*, *39*(18), 4271-4280.
13. Y. Chen, "Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression," PLOS ONE, 2016, http://dx.doi.org/10.1371/journal.pone.0146865.
14. A. Lermontov, L. Yokoyama, M. Lermontov M. and M. A. S. Machado, "River quality analysis using fuzzy water quality index," Riberia do Iguape river watershed, Brazil. Ecol. Indic. 9, 2009, pp. 1188-1197.
15. T. Andreas, B. Olof and F. Bertil, "Precipitation Effects on Microbial Pollution in a River: Lag Structures and Seasonal Effect Modification," PLoS One. 2014; 9(5): e98546.
16. L. D. Bruggink and J. A. Marshall, "Norovirus epidemics are linked to two distinct sets of controlling factors," International Journal of Infectious Diseases, vol. 13, 2009, pp. e125-e126.
17. E. Sokolova, T. J. R. Pettersson, O. Bergstedt and M. Hermansson, "Hydrodynamic modelling of the microbial water quality in a drinking water source as input for risk reduction management," Journal of Hydrology, vol. 497, 2013, pp. 15–23.

18. Y. Icaga, “Fuzzy evaluation of water quality classification,” Ecological Indicators. 7, 2007, pp. 710-718.
19. S. R. Petterson, T. A. Stenstrom and J. Ottoson, “A theoretical approach to using faecal indicator data to model norovirus concentration in surface water for QMRA: Glomma River, Norway,” Water Research, vol. 91, 2016, pp. 31e37.
20. J. A. Marshall and L. D. Bruggink, “The dynamics of norovirus outbreak epidemics: recent insights,” International Journal of Environmental Research and Public Health, vol 8, no. 4, 2011, pp. 1141–1149, doi: 10.3390/ijerph8041141.
21. D.C.S. Bisht and A. Jangid, “Discharge modelling using adaptive neuro-fuzzy inference system,” International Journal of Advanced Science and Technology, vol. 31, 2011, pp. 99–114.
22. S. Chowdhury, P. Champagne, & P. J. McLellan, (2009). Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. *Science of the Total Environment*, *407*(14), 4189-4206.
23. S. Heddam, A. Bermad, & N. Dechemi, (2012). ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. *Environmental monitoring and assessment*, *184*(4), 1953-1971.
24. M. Sahu, S. S.Mahapatra, H. B. Sahu, & R. K. Patel (2011). Prediction of water quality index using neuro fuzzy inference system. *Water Quality, Exposure and Health*, *3*(3-4), 175-191.
25. J. S. R. Jang, “ANFIS: adaptive network-based fuzzy inference systems,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, 1993, pp. 665–685.
26. M. Negnevitsky, “Artifial intelligence: a guide to intelligent systems,” Pearson, 3rd Ed., pp. 277–285.
27. VISK. Retrived from http://www.norskvann.no/ and http://www.nrva.no/ in October 2016. Retrived from http://visk.nu/ in October 2016.
28. R. C. Grøndahl-Rosado, I. Tryland, M. Myrmel, K. J. Aanes and L. J. Robertson, “Detection of Microbial Pathogens and Indicators in Sewage Effluent and River Water During the Temporary Interruption of a Wastewater Treatment Plant,” Water Quality, Exposure and Health, vol. 4, no. 3, 2014, pp. 155–159.
29. H. Chen, & Y. Hu, (2016). Molecular Diagnostic Methods for Detection and Characterization of Human Noroviruses. *The Open Microbiology Journal*, *10*(1).
30. Retrieved from http://www.norskvann.no/ and http://www.nrva.no/ in October 2016.
31. Retrieved from http://visk.nu/ in October 2016.
32. C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006.
33. H. Nickisch and C. E. Rasmussen. Approximations for binary Gaussian process classification. Journal of Machine Learning Research, 9:2035–2078, 10 2008.
34. C. E. Rasmussen and C. K. I. Williams. GPML Matlab code version 4.0 http://gaussianprocess.org/gpml/code/matlab/doc/index.html.