Environ Monit Assess (2023) 195:229
https://doi.org/10.1007/s10661-022-10836-9

Deep learning-based ensemble modeling of *Vibrio parahaemolyticus* concentration in marine environment

Peyman Namadi · Zhiqiang Deng

Received: 5 May 2022 / Accepted: 10 December 2022 / Published online: 24 December 2022

Abstract *Vibrio parahaemolyticus* (*V.p*) is a marine pathogenic bacterium that poses a high risk to human health and shellfish industry, yet an effective regional-scale nowcasting model for managing the risk remains lacking. This study presents the first regional-scale model for nowcasting the level of *V.p* in oysters in the marine environment by developing an ensemble modeling approach. The ensemble modeling approach involves the integration of genetic programming (GP) and deep artificial neural networks (DNN)-based modeling. The new approach was demonstrated by developing three GP-DNN ensemble models for predicting the *V.p* level in North Carolina, New Hampshire, and the combined region. Specifically, GP was employed to establish nonlinear functions between the *V.p* level and antecedent conditions of environmental variables. The nonlinear GP functions and current conditions of individual environmental variables were then utilized as inputs into a DNN model, forming a GP-DNN ensemble model. Modeling results indicated that the GP-DNN ensemble models were capable of predicting the *V.p* level with the correlation coefficient of 0.91, 0.90, and 0.80 for North Carolina, New Hampshire, and the combined region, respectively, demonstrating the impact of distinct environmental conditions in the local areas on accuracy of the combined regional-scale model. Sensitivity analysis results showed that sea surface temperature and sea surface salinity are the two most important environmental predictors for the abundance of *V.p* in oysters, followed by water level, pH, chlorophyll-*a*, and turbidity. The findings suggested that the GP-DNN ensemble models could be utilized as effective predictive tools for mitigating the *V.p* risk.

Keywords Artificial neural networks · Genetic programming · Antecedent environmental conditions · Temperature · Salinity · Oysters

Introduction

Vibrio parahaemolyticus (*V.p*) is a marine pathogenic bacterium endemic to estuarine waters. *V.p* may cause acute gastroenteritis in humans, and it is one of the leading causes of seafood-borne illness in the USA and across the globe (Daniels et al., 2000; Fernandez-Piquer et al., 2011; Zimmerman et al., 2007). According to the Centers for Disease Control and Prevention (CDC), the annual number of reported human infections caused by *V.p* in the USA increased by 440% over the 14-year period of 2000–2014 (www.cdc.gov/vibrio/surveillance.html). The increasing trend in *V.p* infections poses a growing public health risk to shellfish consumers and beachgoers. Modeling efforts have been made in the development of predictive tools for managing and mitigating the risk.

P. Namadi · Z. Deng (✉)
Department of Civil and Environmental Engineering,
Louisiana State University, Baton Rouge, LA 70803, USA
e-mail: zdeng@lsu.edu

Urquhart (2015) presented several data-driven models, including generalized linear modeling (GLM), generalized additive modeling (GAM), and random forest (RF) models, for predicting the likelihood of *Vibrio* occurrence as well as abundance in the Chesapeake Bay by using sea surface salinity (SSS) and sea surface temperature (SST). Results indicated that a hybrid approach involving GAM for classification and RF for regression exhibited a good accuracy for predicting the abundance of *V.p*, as indicated by a mean absolute error of 5.8 cells. Urquhart et al. (2016) developed a nowcasting model that can be used to estimate the likelihood of *V.p* (presence/absence) in oysters in New Hampshire. They used SST, SSS, and chlorophyll-*a* as environmental parameters for the prediction of the likelihood of *V.p* presence in oysters. Their results showed the true positive rate (TPR) of 0.52, true negative rate (TNR) of 0.91, and Matthews correlation coefficient (MCC) of 0.46. Froelich et al. (2015) proposed a predictive model for the abundance of *V.p* bacteria in oysters and water samples in North Carolina by using a linear regression analysis. Results showed that their model can predict the level of *V.p* in water and oysters with the R-squared accuracy of 0.48 and 0.47, respectively. Paranjpye et al. (2015) presented a multiple linear regression method for predicting the level of *V.p* in the Pacific Northwest (Washington) water by using ten biotic and abiotic environmental variables.

In spite of the scientific advances in the development of local models, there are no regional-scale models that could be applied generally to different regions for providing near real-time predictions. Therefore, there is a need to develop regional-scale or even global-scale models. By using SST and SSS as model input variables, Namadi and Deng (2021) presented a series of random forest-based regional-scale forecasting models with the lead-time ranging from 1 to 4 days for predicting the *V.p* abundance in oysters. While the forecasting models are useful for planning purposes, nowcasting models are also needed to provide near real-time predictions. The need for regional-scale nowcasting models motivates this paper. Specifically, this paper is intended to fill the knowledge gap in regional-scale nowcasting models by using advanced machine learning and particularly deep-learning methods.

Artificial neural networks (ANNs) and genetic programming (GP) have proven to be effective modeling methods particularly for describing nonlinear relationships (He & He, 2008; Sætrom et al., 2005; Wang & Deng, 2019). Chenar and Deng (2018a) presented a GP-based model for predicting daily risks of oyster norovirus outbreaks along the Northern Gulf of Mexico coast using environmental indicators including SST, gage height, SSS, solar radiation, rainfall, and wind. Prediction results showed that the GP model was capable of predicting oyster norovirus outbreaks with the true positive and negative rates of 78.53% and 88.82%, respectively, demonstrating the efficacy of the GP model. Similarly, Chenar and Deng (2018b) proposed an ANN model with a 2-day lead time for forecasting norovirus outbreaks by utilizing epidemiological and environmental data. The ANN model was capable of forecasting norovirus outbreaks with the positive and negative predictive values of 76.82% and 100%, respectively. ANN models also have been successfully applied for prediction of *fecal coliform* concentrations in oyster-harvesting areas along the Louisiana Gulf coast (Wang & Deng, 2019).

The overall goal of this study was to present an effective and efficient predictive tool for nowcasting the *V.p* abundance in oysters and thereby mitigating the risk of *V.p* infection to the human health and oyster industry. To that end, the specific objectives of this paper were (1) to identify the functional relationships between the *V.p* abundance in oysters and individual environmental predictors (such as SST, SSS, water level) by means of GP modeling and (2) to develop ensemble models for nowcasting the *V.p* abundance by integrating the GP-based functional relationships and the Deep learning-based Artificial Neural Networks (DNN), producing three GP-DNN ensemble models for three study areas. The scientific significance of these models is that the GP-DNN models can be utilized to explore some important research and management questions including but not limited to: What are the major environmental drivers for *V.p* abundance? Where is the high-risk area of *V.p* contamination to oysters? How will the climate change (particularly the global warming) impact the temporal variation and spatial distribution of *V.p*? This paper provides insightful discussion on how the GP-DNN models can be employed to address the important questions. Answers to these questions provide a scientific basis for the implementation of management intervention for mitigation of potential *V.p* contamination and infection risks, protecting public health.

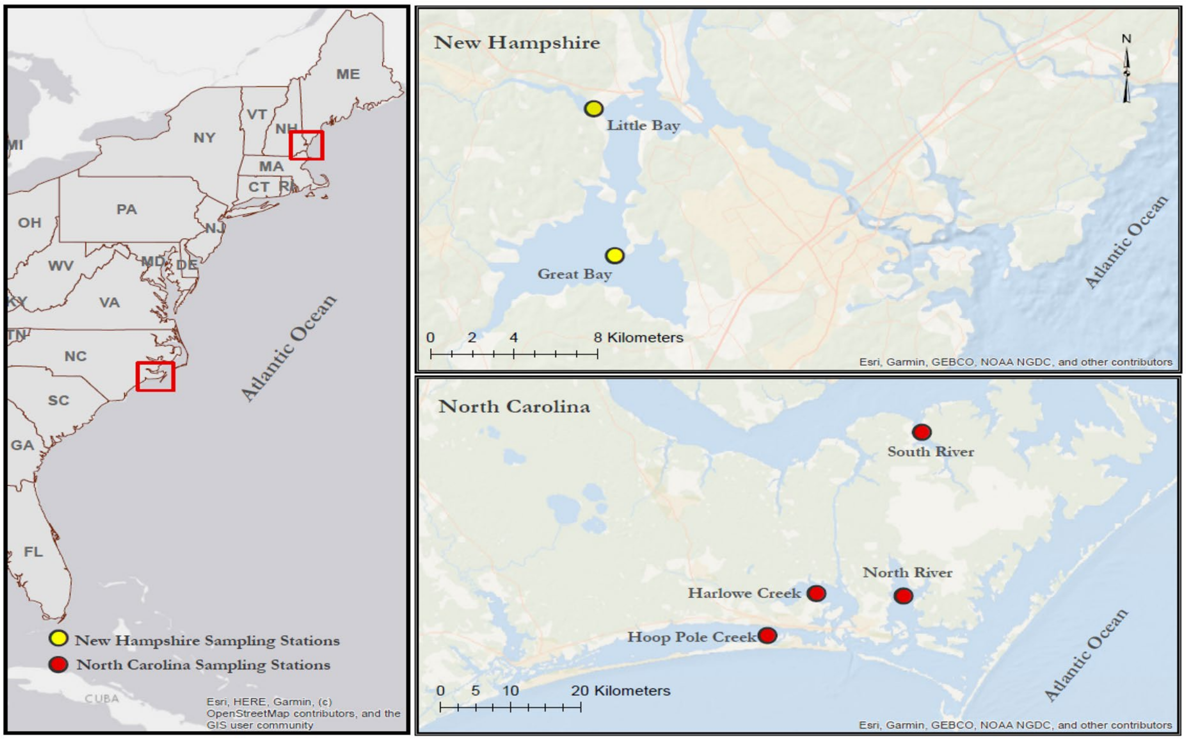
Materials and methods

Study areas

This study focuses on shellfish-harvesting areas in two environmentally distinct study areas, including North Carolina and New Hampshire (Fig. 1). Specifically, oysters (*Crassostrea virginica*) were collected from six harvesting sites along the east coast of the United States (USA), including two sites (Oyster River and Nannie Island) in the US state of New Hampshire and four sites (Harlowe Creek, Hoop Pole Creek, North River, and South River) located along the eastern North Carolina coast. Environmental conditions in these two areas are different due to the difference in their latitudes and climate conditions. The average SST in New Hampshire is lower than those of the sampling sites in North Carolina due to the difference in their latitudes (17.78 °C vs. 19.6 °C). Moreover, the two sampling sites in the New Hampshire located within the Gulf of Main watershed with more freshwater inflow that produces a relatively low SSS range in comparison with that in the North Carolina sampling sites (22.22 ppt vs. 25.8 ppt). Furthermore, the average values of pH and chlorophyll-*a* in North Carolina samples were higher than the corresponding values in New Hampshire.

Data collection and processing

A modeling study conducted by the authors (Namadi & Deng, 2021) found that the *V.p* abundance in the marine environment is affected by environmental indicators including SST, SSS, pH, chlorophyll *a*, and turbidity, respectively. Their study also found that the *V.p* abundance in oysters is controlled by antecedent environmental conditions, while the antecedent environmental conditions can be described with time-lagged variables of each environmental indicator with the time lag of 1–30 days. The time lag of 1–30 days was considered in this study due to the fact that environmental conditions in the marine environment are strongly influenced by tides that cause the periodic (occurring at regular intervals (about 30 days))



Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

Esri, Garmin, GEBCO, NOAA NGDC, and other contributors

Esri, Garmin, GEBCO, NOAA NGDC, and other contributors

Fig. 1 Study area map where yellow circles show the sampling points in New Hampshire, and red circles show the sampling points in the North Carolina

variations (rise and fall) in the surface water level of marine waters (Zhang et al., 2012). Therefore, field data for the five environmental indicators along with the water level and associated time-lagged variables are needed for the development of a predictive model for *V.p*. To that end, two types of data were collected from the literature, including the historical data on *V.p* concentration in oysters and environmental data. Specifically, the data for *V.p* concentration in oysters in Oyster River and Nannie Island (Great Bay, NH) and the data for the five environmental indicators were obtained from the supplementary data of Urquhart et al. (2016) for the period of 2007–2013 (122 samples). Samples described in the supplementary data were collected bi-weekly during warm seasons and monthly during cold seasons (Urquhart et al., 2016). Likewise, the *V.p* concentration and environmental data from 2013 to 2015 (104 samples) for the four oyster-harvesting sites in North Carolina were obtained from the supplementary data sheet of Williams et al. (2017). The data from a total of 226 samples collected from 2007 to 2015 at six oyster-harvesting sites along the US east coast were used in this paper. In addition to the 226 datasets from the literature, new data describing antecedent environmental conditions (up to one month prior to each field sampling day) for chlorophyll-*a*, turbidity, and pH at North Carolina sites were collected from various online sources for the days without field samples in the same period of 2007–2015. The data for SST, salinity, and water levels were obtained from the National Oceanic and Atmospheric Administration Tides and Currents (NOAA Tides & Currents) stations located in the Bogue Sound (https://tidesandcurrents.noaa.gov). The water level is the average of daily water surface elevation measured above mean lower low water (MLLW). Chlorophyll-*a* data for North Carolina were downloaded from the NOAA part of system-wide monitoring program (SWMP) from 2013 to 2015.

All data were then normalized by using feature scaling (unity-based normalization) so that each environmental predictor varies only in the range of 0–1 to eliminate regional effects of datum. Daily SST, water level, and SSS data were reorganized into time series ensembles of individual predictors to describe antecedent environmental conditions. The potential time series ensemble of each environmental predictor consisted of a finite number of time-lagged variables involving daily average and covering the antecedent period of 1–30 days as the gravitational attraction of the moon and the sun to the Earth affects tides on a monthly basis in the marine environment.

Genetic programming-based modeling

Genetic programming (GP) is an evolutionary algorithm based on Darwinian theories of natural selection and survival of the fittest. It operates on the parse tree structure composed of function and terminal sets and approximates the solution that best describes the input–output relationship. A significant benefit of using GP is that a model developed with GP does not need to make explicit assumptions about the functional form of relationship (Chenar & Deng, 2018b). GP, similar to other machine learning techniques, demonstrates its ability to handle dynamic and nonlinear data, especially when the process-based models are not available or underlying physical relationships are not fully understood (Muttil & Chua, 2006). Previous studies demonstrated the capability of GP in solving many practical problems such as modeling and predicting of norovirus outbreaks (Chenar & Deng, 2018b), rainfall-runoff (Mehr & Nourani, 2018), harmful algal blooms (Sivapragasam et al., 2010), and uncertainty analysis of model-estimated longitudinal and lateral dispersion coefficients in open channels (Najafzadeh et al., 2021).

GP was applied in this paper first to identify functional relationships between the log-transformed *V.p* level and individual environmental predictors, including SST, SSS, and water level, for the study areas (North Carolina, New Hampshire, and the combined region) under the impact of antecedent environmental conditions. To that end, the daily average for antecedent or time-lagged SST, SSS, and water level was utilized to determine the optimal functional relationships between the *V.p* level and the daily averages of individual environmental predictors. The GP was then utilized to produce an initial population of randomly generated programs (equations), calculate their fitness, and subsequently select the programs for reproduction and recombination to form a new population. The mixture of arithmetic operators ($+, -, \times, \div$) and mathematical functions (sin, cos, log) was selected to form a set of equations for each generation iteration. The fitness measures, such as mean absolute error ($MAE = (\sum_{i=1}^{n} |\hat{y}_i - y_i|)/n$, where $\hat{y}_i$= a mode-predicted value and y_i = an observed value), mean squared error ($MSE = (\sum_{i=1}^{n} (\hat{y}_i - y_i)^2)/n$), and root mean square error ($RMSE = \sqrt{(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2)/n}$), must be appropriately chosen to ensure that the individual programs, which best fit the data, are selected from the initial population. Specifically, the programs that best fit the data were selected to exchange part of the information between them for producing optimal model equations through "crossover" and "mutation" processes that mimic the natural reproduction process. The maximum tree size of 35 was used as the termination criterion. The adopted fitness measure was the mean absolute error (MAE). Table 1 presents the control parameters that were used in this study.

In addition to the functions describing antecedent environmental conditions, functional relationships between the *V.p* level and the current environmental conditions, described with current SST, SSS, the deviation of salinity from the optimum SSS of 28 ppt (S_{op}=|SSS -28|), pH, turbidity, and chlorophyll-*a*, were also constructed for North Carolina, New Hampshire, and the combined region. GPLAB toolbox, developed by Silva and Almeida (2003) in MATLAB, was utilized for the construction of the functions. The functional relationships were then employed as input variables (functions) for ANN modeling.

Artificial intelligence-based modeling

Artificial intelligence-based neural networks (ANNs) have shown promise in modeling nonlinear relationships between target variables and predictor variables (Noori et al., 2016; Zhang et al., 2015). Since the relationship between the *V.p* abundance and environmental predictors is complicated and nonlinear, nonlinear ANNs are best suited for modeling the *V.p* abundance.

A typical ANN model consists of three primary layers, including an input layer, a hidden layer, and an output layer. A deep artificial neural network (deep ANN) was utilized in this study to predict the *V.p* abundance in oysters. A critical difference between a basic ANN (shallow neural network) and a deep ANN is the number of hidden layers. More hidden layers allow deep ANN models to have multiple processing layers to learn representations of data with multiple levels of abstraction (LeCun et al., 2015). The adaptive gradient descent back-propagation algorithm was utilized for model training. The advantage of the adaptive algorithm over the basic gradient descent algorithm is the variable learning rate. In the adaptive gradient descent (g), the model determines the output with the initial neural network and compares it with observational data to calculate the error. In the next step, the model uses the present learning rate (0.01) for the second iteration to find new weights and biases. Then, the model calculates the new error rate and compares it with the previous error rate. If the new error increases, the new learning rate (α) would be decreased by a default ratio (γ: the ratio used in this study was 0.7). If the new error decreases, the learning rate would be increased by another default ratio (the ratio to increase the learning rate in this study is 1.05). Equation 1 shows the weight (W) update equation where $\alpha_{k+1} = \gamma\alpha_k$ and $\gamma = 0.7$ if the new error is higher than the previous one; otherwise, $\gamma = 1.05$.

$$W_{k+1} = W_k - \alpha_{k+1} g_k \tag{1}$$

The deep ANN model presented in this study consists of five layers, including an input layer, three hidden layers, and an output layer. The input layer involves seven variables including current SST, SSS, chlorophyll-*a*, and four GP functions for SST($f(T_i)$), SSS($f(S_i)$), water level ($f(W_i)$), and all current environmental predictors ($f(All)$). The number of neurons in hidden layers and their activation functions were determined through a trial-and-error process until the highest prediction accuracy was obtained. Based on the final design of the deep ANN, the first hidden layer consists of 40 hidden neurons, and the activation function is $a = tansig(n)$. The second hidden layer consists of 40 hidden neurons, and the activation function is $a = satlin(n)$. The third hidden layer consists of 20 hidden neurons, and the activation function is $a = logsig(n)$. Figure 2 shows the architecture of the Deep ANN.

Table 1 Values of GP control parameters

Parameter	Value
Population size	600
Generation	600
Maximum tree size	35
Crossover rate	0.75
Mutation rate	0.25

Eight years of data collected from North Carolina and New Hampshire were divided into two groups, including a development data group used for model development (60% of all data) and an independent validation data group (40% of all data). The development data group was further divided into three subgroups, including training, validation, and testing. Table 2 shows the model development dataset (training, validation, and testing) and the independent validation dataset used for the two local models and the regional model. Each deep ANN model was continuously trained until the highest model performance was achieved. The model performance was measured using two statistical metrics including the correlation coefficient and mean squared error (MSE). The top models of the high performance with the development dataset were further tested with the independent validation dataset. The model of the best overall performance with both groups of data (dependent and independent datasets) was finally selected as the deep ANN model. In fact, three deep ANN models were created. Specifically, one regional-scale deep ANN model was presented for prediction of the *V.p* level in the US east coast. Two local deep ANN models of better performance were also proposed for nowcasting the *V.p* level in the North Carolina coast and the New Hampshire coast, respectively.

Several performance metrics were utilized to illustrate the classification ability of a model as a binary (presence/absence) classifier, including the true positive rate (TPR), true negative rate (TNR), accuracy (ACC), and Matthews correlation coefficient (MCC). Equations 2, 3, 4, and 5 show definitions for the TPR, TNR, ACC, and MCC, respectively (Matthews, 1975), where TP is the number of true positive (presence) predictions that a model produced correctly; TN is the number of true negative (absence) predictions that a model produced correctly; FP is the number of false positive (presence) predictions that a model made, and FN is the number of false negative (absence) predictions that a model produced.

Table 2 The duration of the dependent dataset and independent dataset for three models

Model	Model development		Validation	
	Start	End	Start	End
North Carolina	2/4/2013	4/25/2014	5/1/2014	10/16/2015
New Hampshire	6/27/2007	6/28/2010	7/14/2010	12/5/2013
Regional model	6/27/2007	5/22/2013	5/28/2013	10/16/2015

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$TNR = \frac{TN}{TN + FP} \tag{3}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Sensitivity analysis

Sensitivity analysis was conducted to examine the effects of different input variables on the model-predicted *V.p* level and thereby to identify key variables that control the occurrence and concentration of *V.p* in oysters. Two different methods were applied for the sensitivity analysis,

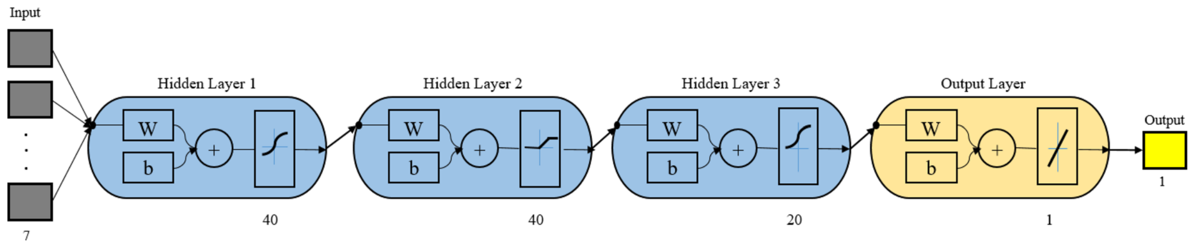


Fig. 2 Deep artificial neural network architecture

including a local sensitivity analysis method and a global sensitivity analysis method (Chenar & Deng, 2018b).

In terms of the local sensitivity analysis, the one-at-a-time method (OAT) was used. Specifically, values of individual model input variables, including SST, SSS, chlorophyll-*a*, turbidity, pH, and water level, were changed by ± 10% to ± 50% from their mean values, and then, the corresponding percent changes in the model-predicted *V.p* level were recorded and compared graphically. In terms of the global sensitivity analysis, the perturb method was employed (Chenar & Deng, 2018a). Specifically, the model-predicted *V.p* levels before and after the perturbation were calculated, and then, the mean squared errors (MSE) in the model-predicted *V.p* levels, produced by perturbations to the input variables, were compared graphically. A 50% perturbation was utilized for each input variable in this study.

Results and discussion

Genetic programming functions

Three GP-evolved equations, including $f(T_i)$,$f(S_i)$, and $f(W_i)$ describing the best functions for antecedent SST, SSS, and water level, are presented in Eqs, respectively, for the two study areas and their combined region. Parameters involved in each GP function are time-lagged predictors, demonstrating the cumulative effect of extended favorable environmental conditions on the level of *V.p* in oysters. Table 3 lists the correlation coefficients between the log-transformed *V.p* level and individual environmental predictors (variables) as well as the GP-evolved functions for the predictors.

Results indicate that the GP models created with the time-lagged variables improve the correlation between the *V.p* level and SST, SSS, and water level by 13%, 84%, and 1075%, respectively, in the regional model. Another important statistical metric for a prediction model is mean squared error (MSE) that demonstrates the model performance. The incorporation of the GP functions for SST, SSS, and water level into the regional model reduced the MSE by 29%, 23%, and 36%, respectively, in comparison with the regional model involving only current environmental predictors. Table 3 shows the improvement percentages of three models.

Table 3 Correlation coefficients between *V.p* and GP functions and current predictors

Parameter name	Model	SST	*f*(*T*)	Improvement	Salinity	*f*(*S*)	Improvement	Water level	*f*(*W*)	Improvement
Correlation with log(*V.p*)	North Carolina	0.57	0.67	18%	0.13	0.55	323%	0.1	0.63	530%
	New Hampshire	0.51	0.63	24%	0.27	0.42	56%	0.06	0.33	450%
	Regional model	0.54	0.61	13%	0.25	0.46	84%	0.04	0.47	1075%
Mean square deviation (MSD)	North Carolina	1.62	0.89	45%	1.99	1.13	43%	1.95	0.93	52%
	New Hampshire	0.97	0.66	32%	1.08	0.97	10%	1.7	1.07	37%
	Regional model	1.22	0.86	29%	1.36	1.05	23%	1.72	1.1	36%

North Carolina:

$$\begin{aligned} f\left(T_i\right) =&\left(T_8 + T_4\right) \times T_4 \times [T_{10} + Sin\left[4\left(T_{11} + T_6\right)\right. \\ &\left.\times T_{11}^3 \times T_6 \times Sin\left(2T_8^2 \times \left(T_{24} + T_8\right)\right)\right] \end{aligned} \tag{6}$$

$$f\left(S_i\right) = \frac{S_{17} \times S_3 \times S_{10} \times Cos\left(\frac{S_{15}}{S_{26}}\right) \times Cos\left(\frac{S_5}{S_{25}}\right)}{Cos(Cos\left(\frac{S_6}{S_{26}}\right) \times \frac{S_4}{Cos\left(\frac{S_{11}}{S_{24}}\right)} \times \frac{S_{24}}{S_{21} \times S_{29}}} \tag{7}$$

$$\begin{aligned} f\left(W_i\right) =&W_{30} + Cos\left(Log\left(W_{26}\right)\right) \\ &+ \left[W_{30} + W_6 + Cos\left(\frac{W_{16}}{W_5}\right)\right] \\ &\times W_{25} + Cos\left(Log\left(W_{18}\right)\right) \\ &+ Cos\left[Cos\left(Log\left(W_{16} \times LogCos\left(W_7\right)\right)\right)\right. \\ &\left.+ W_{16} + W_1\right] \end{aligned} \tag{8}$$

New Hampshire:

$$\begin{aligned} f\left(T_i\right) =& [Cos\left(T_3 \times T_{30} - T_6\right) + T_1] \\ &\times T_3^2 \times (T_1 + T_{30}) \times Cos(Log\left(2T_{20}^2\right)) \\ &\times (T_3 \times CosT_{20} + T_{30}) \times T_1^2 \end{aligned} \tag{9}$$

$$\begin{aligned} f(S_i) =& (Cos(Log(Log(S_{25} - S_{29}))) + Cos(S_{31})) \\ &\times Cos(S_{17} \times Log(Log(S_{31})) \times S_1^2 \\ &\times (Cos(Log(Log(S_{25} - S_{29}))) + S_1) \\ &\times (Cos(Log(S_{31})) + S_{11}) \end{aligned} \tag{10}$$

$$\begin{aligned} f(W_i) =& (Cos((2W_4) + 2W_{11}) \\ &\times Cos(W_{14} + Cos(Log(W_1) + W_{16}) \\ &\times (2W_{16} + Cos(Log(W_7)) \\ &\times Cos(W_4 + 2W_{16})) \times W_{11} \end{aligned} \tag{11}$$

Regional model:

$$\begin{aligned} f(T_i) =& [[Cos(\mathrm{Log}(T_{30} \times Cos(Log(T_{19} - T_{24}))))] \\ &\times T_{29} \times T_4 + T_3 \times Sin(T_7 \times (T_{17}^2 + T_7 \\ &+ T_7 \times T_{17}))] \times 2T_3^2 \end{aligned} \tag{12}$$

$$\begin{aligned} f(S_i) =& [Cos(Log[[S_1 + S_{31} \times Cos(Log[S_{12} \\ &\times Log[(S_{27} + S_{14}) \times S_4 \times S_1]]])] \\ &\times S_1^3] \times S_1 + S_1)] \times \frac{S_{20}}{S_2 \times S_{23}} \end{aligned} \tag{13}$$

$$\begin{aligned} f(W_i) =& (Cos(Log(W_{16})) \times [Cos(W_7 \times W_4) \\ &+ Cos(W_9 + Log(W_3) - W_{15}) \\ &\times Cos(Log(W_3 \times W_{12} + Log(W_4)))] \\ &\times (Cos(Log(W_4)) + Cos(Log(W_{16}))) \end{aligned} \tag{14}$$

In addition to the three GP functions of time-lagged environmental predictors for each model, GP was also applied in the development of prediction models without considering the time-lagged effects or variables by using current environmental predictors (SST, optimum SSS, pH, and turbidity). Equations (15), (16), (17) show the best GP models that were finally selected for North Carolina, New Hampshire, and the combined region, respectively, describing the explicit functional relationships between log (*V.p*) and the environmental predictors. The parameters T, S, S_{op}, pH, Tur, and Chl denote SST, SSS, optimum SSS (|SSS-28|), pH, turbidity, and chlorophyll-*a*, respectively. The correlation coefficient between the log (*V.p*) level and a GP-evolved equation is 0.70 for North Carolina, 0.71 for New Hampshire, and 0.66 for the combined region, respectively.

North Carolina:

$$f(All) = \frac{|T - pH|}{pH + \left|\frac{pH \times (T - S + Tur) \times (T - S + \frac{Chl}{Chl - Tur})}{(T - Chl) \times (pH - Chl) \times Tur}\right|} \tag{15}$$

New Hampshire:

$$\begin{aligned} f(All) =& [[|LogLogLog(S_{op} - Tur)| + S_{op} - Tur] \\ &\times T + Chlo + S_{op}] \times pH \times T \\ &\times (S + Tur + T) \times (S + T + pH \times T) \end{aligned} \tag{16}$$

Regional-scale model:

$$\begin{aligned} f(All) =& Cos(Log(T)) \times [(T + pH) \times T \\ &+ T \times Cos(Log(S \times T + (Tur + T) \\ &\times Cos(Log(T^2 \times Cos(Log(S)))))) + T \\ &\times pH \times Cos(Log(Log(S_{op}))) \times Cos(Log(S))] \end{aligned} \tag{17}$$

Artificial intelligence-based modeling

The best-trained deep ANN model for each study area was finally selected as the prediction model for the *V.p* level in oysters. Input parameters of the deep ANN (DNN-GP) models include $f(T_i)$,$f(S_i)$, $\left(W_i\right)$, $f(All)$, SST, SSS, and chlorophyll-*a*. Figures 3 and 4 show the performance of the DNN-GP models with the model development and independent validation datasets for North Carolina, New Hampshire, and the combined region, respectively. Specifically, Fig. 3A shows the comparison between the log(*V.p*) abundance predicted with the DNN-GP model and observed in four sampling locations in North Carolina oyster-harvesting

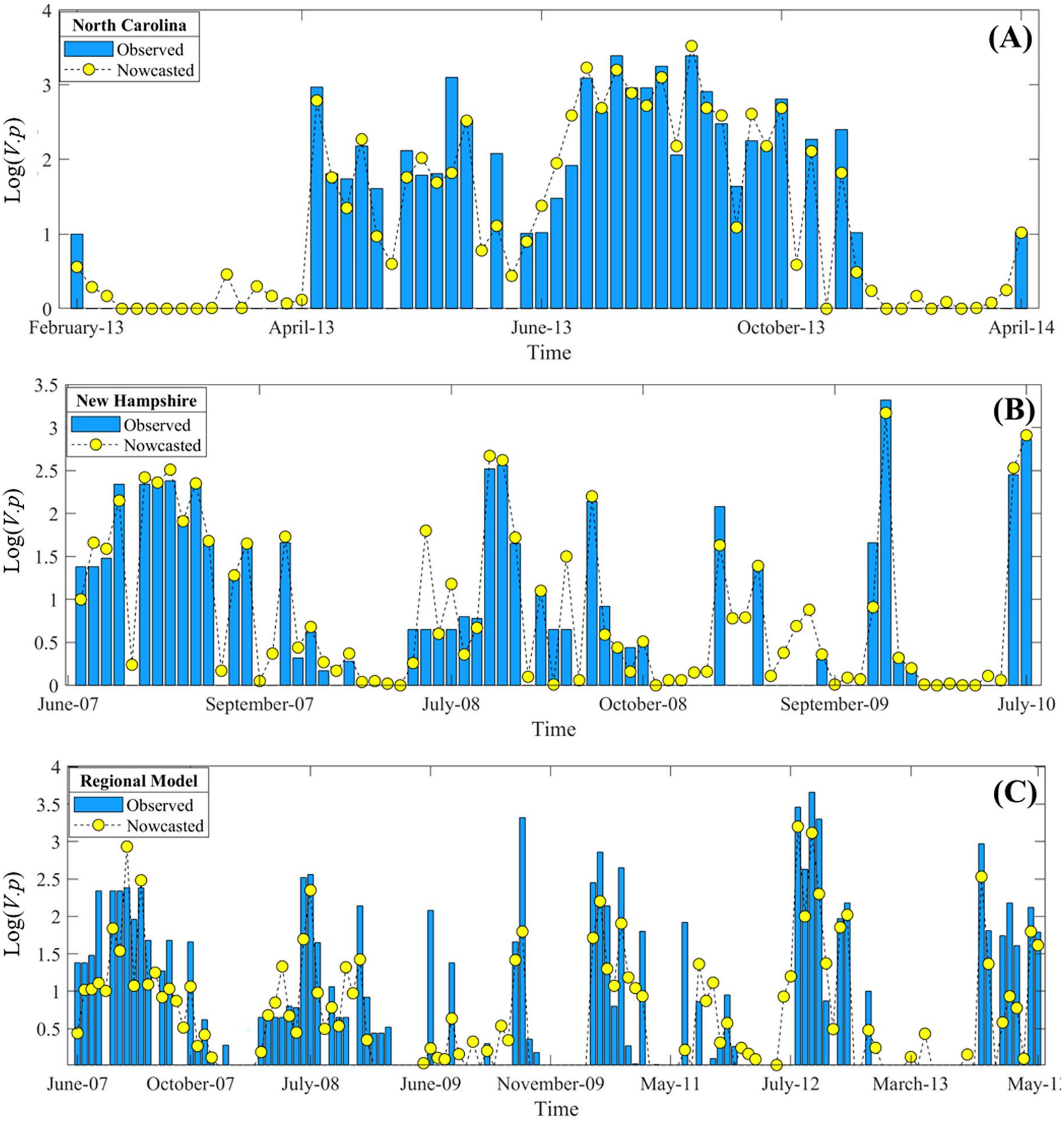


Fig. 3 Comparison between the log-transformed *V.p* levels predicted by the DNN-GP models with the model development dataset and observed in North Carolina **A**, New Hampshire **B**, and the entire region **C**, respectively

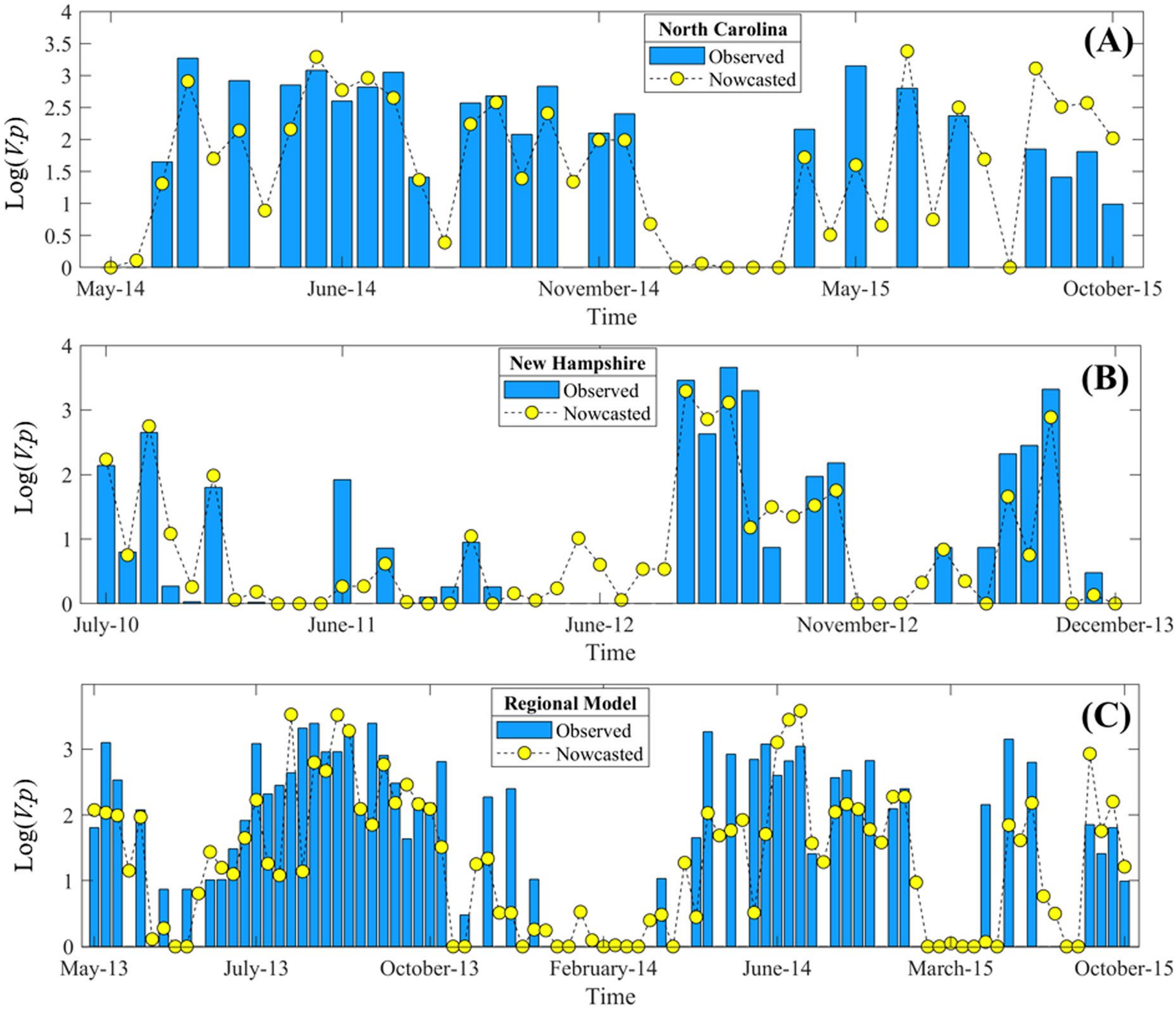


Fig. 4 Comparison between the log-transformed *V.p* levels predicted by the DNN-GP models with the independent validation dataset and observed in North Carolina (**A**), New Hampshire (**B**), and the entire region (**C**), respectively

areas. The correlation coefficient between predicted *V.p* levels and observed data is 0.96, and MSE for this model is 0.12. The performance metrics demonstrated the high efficacy of the model in predicting the *V.p* level in the model development phase. Similarly, Fig. 3B displays the performance of the New Hampshire model. The correlation coefficient of 0.94 and MSE of 0.10 confirms the very good performance of the New Hampshire model. Likewise, Fig. 3C provides a comparison between the log-transformed *V.p* levels predicted with the DNN-GP model and observed during the 6 years from June 27, 2007, to May 22, 2013, in New Hampshire and North Carolina oyster-harvesting areas. The correlation coefficient of 0.85 and MSE of 0.3 demonstrates the high efficacy of the model in predicting the *V.p* level in the model development phase at the regional scale. Figure 4 shows the performance of the DNN-GP models with the independent validation dataset. Table 4 summarizes the performance metrics for the models in terms of the correlation coefficient and MSE. The metrics indicate that the overall correlation coefficients are 0.91, 0.90, and 0.80 for North Carolina, New Hampshire, and the combined region, respectively. Furthermore, the MSE is 0.27, 0.21, and 0.48 for North Carolina, New Hampshire, and the combined region, respectively. It is clear that

Table 4 Performance metrics of three models

Performance of model	Model	Model development	Validation	100% of the dataset
Correlation	North Carolina	0.96	0.84	0.91
	New Hampshire	0.94	0.85	0.9
	Regional model	0.85	0.74	0.8
MSE	North Carolina	0.12	0.5	0.27
	New Hampshire	0.1	0.37	0.21
	Regional model	0.3	0.74	0.48

the performance of the North Carolina model and that of the New Hampshire model are high and comparable. The performance of the combined regional model is impacted by the distinct environmental and climatic conditions in North Carolina and New Hampshire, as stated in the previous section. The model performance results suggest that site-specific models may serve as tools for site-specific management intervention due to their high accuracy, while the combined regional-scale model may serve as a general model that may be applied to other regions as well, demonstrating the importance of both the local and the regional-scale models to the management of potential *V.p* risk to the general public.

The predictive DNN-GP models can also be employed for the purpose of classification in terms of whether the model-predicted *V.p* level in oysters meets the seafood safety standard defined by the National Shellfish Sanitation Program (NSSP) Guide (FDA, 2017). According to the NSSP Guide, the safety standard for the *V.p* level in oysters is 30 MPN/gram, while any *V.p* level lower than 30 MPN/gram is considered to be undetectable or safe. This threshold level was used to convert the model prediction of *V.p* level in oysters into a binary classification (presence or absence). That is, a model predicts the presence of *V.p* in oysters if the predicted cell count ≥ 30 MPN/gram and the absence if the cell count < 30 MPN/gram.

Table 5 summarizes the performance metrics for the three models. The receiver operator characteristic curve (ROC) (Fawcett, 2006) was also used to visualize the performance of a binary classifier model, where the area under the curve (AUC) of 1 indicates a perfect performance, while the area of 0.5 means that the model is pointless. Specifically, the area under the curve is 0.730 for the regional-scale model, 0.973 for the North Carolina model, and 0.970 for New Hampshire model, as shown in Fig. 5.

Sensitivity analysis

The result from the local sensitivity analysis method is shown in Fig. 6. The vertical axis represents percent changes in SST, SSS, water level, turbidity, chlorophyll-*a*, and pH from their corresponding mean values, while the horizontal axis indicates how the model responds to the changes in individual predictors in terms of percent change in the model-predicted *V.p* level. It is clear from Fig. 6 that SST is by far the most effective predictor as a 50% decrease in SST reduces the model-predicted *V.p* level by 57%. SSS is the second influential parameter to the model-predicted *V.p* level, followed by water level, pH, and chlorophyll-*a* concentration. Turbidity is the least important parameter to the *V.p* prediction.

Figure 7 shows the global sensitivity analysis result from the perturb method. The figure indicates that temperature is again the most important environmental indicator for the *V.p* level with the highest MSE value of 1.15, followed by SSS (0.81), water level (0.78), pH (0.74), chlorophyll-*a* (0.66), and turbidity (0.64). It is clear that SST and SSS are the two most important environmental indicators, and chlorophyll-*a* and turbidity are the two least important indicators for *V.p* levels. The results from two different sensitivity analysis methods are consistent, confirming that

Table 5 Assessment metrics of three *V.p* models

Model assessment	New Hampshire	North Carolina	Regional model
TPR	0.88	0.88	0.68
TNR	0.96	0.93	0.96
ACC	0.93	0.90	0.86
MCC	0.83	0.81	0.69

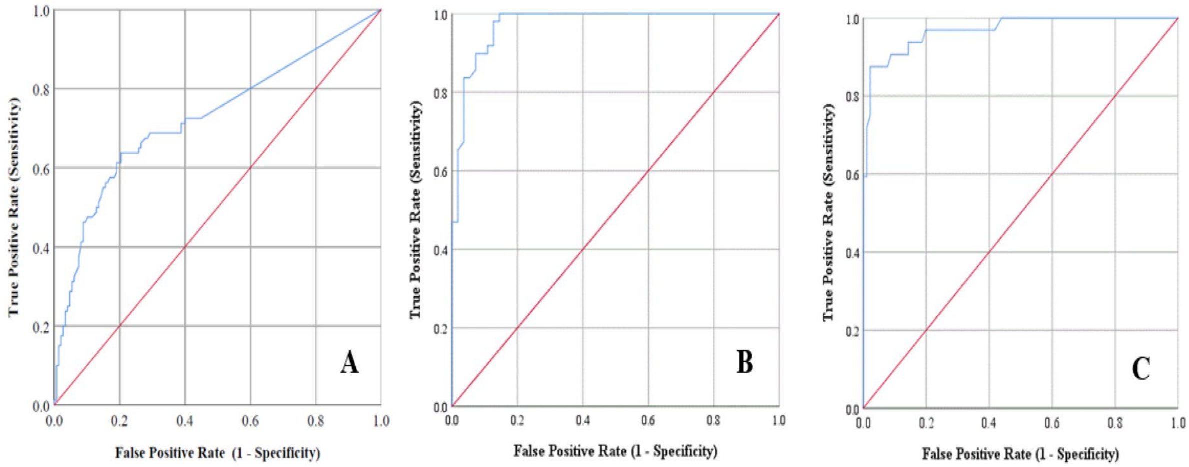


Fig. 5 Receiver operating characteristic (ROC) curve for the *V.p* prediction models, including the regional model **A**, North Carolina model **B**, and New Hampshire model **C**

Fig. 6 Local sensitivity analysis result showing the relative importance of six environmental indicators to model-predicted *V.p* level

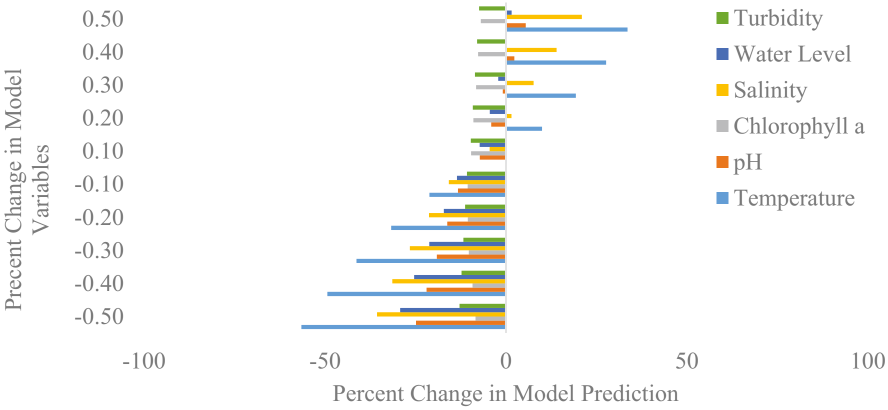
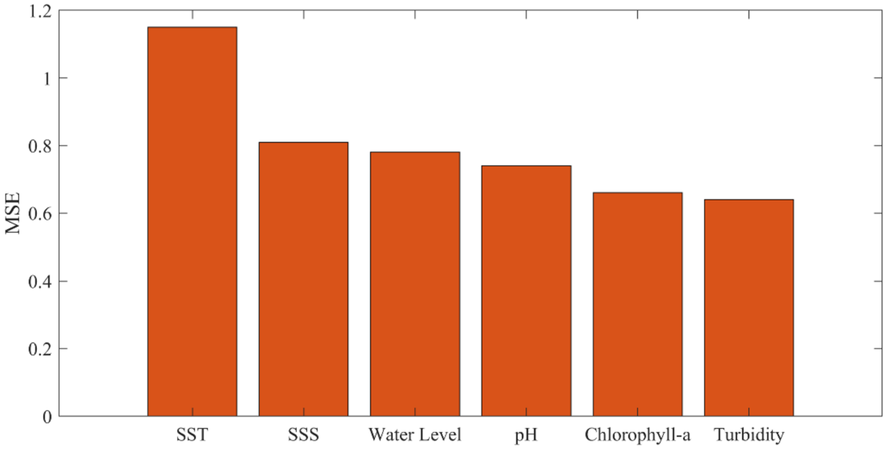


Fig. 7 The global sensitivity analysis result showing the relative importance of six independent environmental indicators to model-predicted *V.p* level based on the perturb method



the *V.p* prevalence in oysters is affected not only by SST and SSS but also by other environmental factors (particularly water level and pH).

Discussion

While increasing *Vibrio* infection cases has been reported (https://www.cdc.gov/vibrio/index.html), our ability to predict the *V.p* presence and infection remains limited due to the lack of deep understanding of environmental predictors for the *V.p* abundance and thus the lack of an effective, predictive, and particularly regional-scale model. The scientific significance of the GP-DNN models is that these models can be utilized to explore some important research and management questions including but not limited to: What are the major environmental drivers for *V.p* abundance? Where is the high-risk area of *V.p* contamination to oysters? How will the climate change (particularly the global warming) impact the temporal variation and spatial distribution of *V.p*?

In terms of the major environmental drivers for *V.p* abundance, Figs. 6 and 7 clearly indicate that SST is by far the most important environmental driver or predictor for *V.p* abundance, followed by SSS. While the results are consistent with those from previous studies (https://products.coastalscience.noaa.gov/vibrioforecast/gom/default.aspx; Namadi & Deng, 2021; FDA, 2005), this study also identified the importance of other environmental predictors (including water level, pH, chlorophyll-*a*, and turbidity) and particularly their time series ensembles, demonstrating the importance of antecedent environmental conditions to *V.p* abundance and thus leading to the development of the hybrid DNN-GP models.

In terms of the high-risk area of *V.p* contamination to oysters, Figs. 4 and 5 illustrate that the hybrid DNN-GP models are not only capable to predict the *V.p* concentration but also the risk of *V.p* contamination to oysters at any oyster-harvesting areas such as those shown in Fig. 1. If the model predictions show a risk of *V.p* contamination to oysters or a high *V.p* concentration exceeding the threshold level (FDA, 2005) in a particular oyster-harvesting area, management interventions should be implemented by testing and potentially closing this area to shellfish harvesting (Chenar & Deng, 2021).

In terms of climate change impact, Figs. 6 and 7 clearly indicate that the model-predicted *V.p* concentration is most sensitive to temperature changes, while the global warming is causing the increase in sea surface temperature. Spatially, a higher SST means that some relatively cold regions like Alaska, where *V.p* infections were rarely reported, may experience more and more frequent *V.p* infections in the future due to global warming. Therefore, the global warming-induced temperature rise will expand not only the spatial extent of *V.p* infections but also the time span of *V.p* infections from summer only to spring, summer, and even autumn.

The hybrid DNN-GP models provide new insights into the cumulative and time-lagged effect of six environmental predictors on the *V.p* abundance in terms of how the antecedent conditions of the environmental predictors or time-lagged environmental predictors affect the *V.p* abundance. As a result, the hybrid DNN-GP models are capable of achieving improved performance in predicting the *V.p* abundance compared with some other models. For instance, the hybrid DNN-GP model for New Hampshire is capable of achieving the TPR, TNR, and MCC of 0.88, 0.96, and 0.83, respectively, while the generalized linear models (GLM) developed by Urquhart et al. (2016) using the same datasets achieved significantly lower TPR, TNR, and MCC values of 0.52, 0.91, and 0.46, respectively, for the same oyster-harvesting areas. The hybrid DNN-GP model for the combined region is characterized by the area under the receiver operating characteristic curve of 0.83, 0.81, and 0.69 for the training, testing, and validation datasets, respectively. The TPR, TNR, and ACC of the regional hybrid DNN-GP model are 0.68, 0.96, and 0.86, respectively, while the generalized additive model (GAM) proposed by Urquhart et al. (2016) achieved the TPR, TNR, and ACC of 0.48, 0.76, and 0.65 that are much lower than those of the hybrid DNN-GP model, demonstrating the efficacy of the regional-scale DNN-GP model in predicting *V.p* levels.

The results from this study demonstrate the importance of both current and time-lagged environmental variables to the prediction of the *V.p* concentration in oysters. The results also suggest that the time-lagged variables provide an independent and, in some cases, superior predictive power compared to current variables. Specifically, the SST on previous 3–30 days, SSS on previous 1–31 days, and water level on previous 3–16 days control the current *V.p* level in oysters. The findings provide a theoretical basis for further improvements of predictive tools for the management and particularly intervention of *Vibrio* infections.

Overall, this paper is unique in terms of presenting an effective regional-scale DNN-GP model that performs best among existing models and thus can be employed to address the important questions with the highest accuracy. Accurate answers to these questions provide a scientific basis for the implementation of management intervention for mitigation of potential *V.p* contamination and infection risks, protecting public health.

Conclusions

Three hybrid DNN-GP models, including two local models and one regional-scale model, were created within the MATLAB computing environment for predicting *V.p* concentration in oysters in the marine environment. Based on the findings from the model predictions and sensitivity analysis, the following conclusions can be drawn:

1. The integration of deep artificial neural networks and genetic programming provides an effective approach to identifying important environmental predictors and particularly their time series ensembles describing the cumulative effect of antecedent environmental conditions on *V.p* concentration in oysters in the marine environment.
2. The antecedent environmental conditions control *V.p* concentration in oysters. Specifically, the SST on previous 3–30 days, SSS on previous 1–31 days, and water level on previous 3–16 days control the current *V.p* concentration in oysters. The genetic programming-based nonlinear functional relationships (time series ensembles) for antecedent SST, SSS, and water level improved model performance in terms of reducing MSE and increasing the correlation coefficient in comparison with current environmental predictors.
3. In addition to SST and SSS that are commonly included in existing models, water level, pH, chlorophyll-*a*, and turbidity are also important environmental predictors for *V.p* concentration in oysters.
4. The hybrid DNN-GP models are particularly useful in terms of their classification capability as binary (presence/absence) classifier. The application of the hybrid models would allow management interventions and thereby potentially reduce the risk of *V.p* infection to human health.

Author contribution PN performed data collection, model development and validation, first draft writing, and graph construction under the direction of ZD who designed the methodology, edited the current version of the manuscript, and applied for funding. All authors read and approved the final manuscript.

Funding This work was supported by the US NASA (National Aeronautics and Space Administration: award number 80NSSC20M0216) and the Louisiana Board of Regents (LEQSF(2020–23)-Phase3-14).

Data availability All data used in this work are available from the literature.

Code availability The codes used in this research are available from the authors upon request.

Declarations

Competing interests The authors declare no competing interests.

References

Chenar, S. S., & Deng, Z. (2018a). Development of artificial intelligence approach to forecasting oyster norovirus outbreaks along Gulf of Mexico coast. *Environment International, 111*, 212–223.

Chenar, S. S., & Deng, Z. (2018b). Development of genetic programming-based model for predicting oyster norovirus outbreak risks. *Water Research, 128*, 20–37.

Chenar, S. S., & Deng, Z. (2021). Hybrid modeling and prediction of oyster norovirus outbreaks. *Journal of Water and Health, 19*(2), 254–266. https://doi.org/10.2166/wh.2021.251

Daniels, N. A., MacKinnon, L., Bishop, R., Altekruse, S., Ray, B., Hammond, R. M., Thompson, S., Wilson, S., Bean, N. H., & Griffin, P. M. (2000). *Vibrio parahaemolyticus* infections in the United States, 1973–1998. *Journal of Infectious Diseases, 181*, 1661–1666.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fernandez-Piquer, J., Bowman, J. P., Ross, T., & Tamplin, M. L. (2011). Predictive models for the effect of storage temperature on *Vibrio parahaemolyticus* viability and counts of total viable bacteria in Pacific oysters (*Crassostrea* *gigas*). *Applied and Environmental Microbiology, 77*(24), 8687–8695.

Food and Drug Administration (FDA). (2017). National shellfish sanitation program guide for the control of molluscan shellfish–2017 revision. *Center for Food Safety and Applied Nutrition,* Food and Drug Administration, U.S. Department of Health and Human Services.

Food and Drug Administration (FDA). (2005). Quantitative risk assessment on the public health impact of pathogenic *Vibrio parahaemolyticus* in raw oysters. *Center for Food Safety and Applied Nutrition*, Food and Drug Administration, U.S. Department of Health and Human Services.

Froelich, B., Ayrapetyan, M., Fowler, P., Oliver, J. D., & Noble, R. J. (2015). Development of a matrix tool for the prediction of *Vibrio* species in oysters harvested from North Carolina. *Applied and Environmental Microbiology, 81*(3), 1111–1119.

He, L., & He, Z. (2008). Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA. *Water Research, 42*, 2563–2573.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure, 405*(2), 442–451.

Mehr, A. D., & Nourani, V. (2018). Season algorithm-multigene genetic programming: A new approach for rainfall-runoff modeling. *Water Resources Management, 32*(8), 2665–2679.

Muttil, N., & Chau, K.-W. (2006). Neural network and genetic programming for modeling coastal algal blooms. *International Journal of Environment and Pollution, 28*(3–4), 223–238.

Najafzadeh, M., Noori, R., Afroozi, D., Ghiasi, B., Hosseini-Moghari, S. -M., Mirchi, A., Haghighi, A. T., & Kløve, B. (2021). A comprehensive uncertainty analysis of model-estimated longitudinal and lateral dispersion coefficients in open channels. *Journal of hydrology*, *603* (Part A), 126850, https://www.sciencedirect.com/science/article/pii/S0022169421009008?via%3Dihub

Namadi, H. P., & Deng, Z. (2021). Modeling and forecasting *Vibrio Parahaemolyticus* concentrations in oysters. *Water Research, 189*, 116638. https://doi.org/10.1016/j.watres.2020.116638

Noori, R., Deng, Z., Kiaghadi, A., & Kachoosangid, F. T. (2016). How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers?". *Journal of Hydraulic Engineering, 142*(1), 04015039. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001062

Paranjpye, R. N., Nilsson, W. B., Liermann, M., Hilborn, E. D., George, B. J., Li, Q., Bill, B. D., Trainer, V. L., Strom, M. S., & Sandifer, P. A. (2015). Environmental influences on the seasonal distribution of *Vibrio parahaemolyticus* in the Pacific Northwest of the USA. *FEMS Microbiology Ecology, 91*(12), fiv121, https://academic.oup.com/femsec/article/91/12/fiv121/2467317

Sætrom, P., Sneve, R., Kristiansen, K. I., Snøve, O., Grünfeld, T., Rognes, T., & Seeberg, E. (2005). Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. *Nucleic Acids Research, 33*(10), 3263–3270. https://doi.org/10.1093/nar/gki644

Silva, S., & Almeida, J. (2003). Gplab-a genetic programming toolbox for Matlab. *Book Gplab-a genetic programming toolbox for Matlab*, City: Citeseer.

Sivapragasam, C., Muttil, N., Muthukumar, S., & Arun, V. M. (2010). *Prediction of Algal Blooms Using Genetic Programming., 60*(10), 1849–1855. https://doi.org/10.1016/j.marpolbul.2010.05.020

Urquhart, E. A., Zaitchik, B., Guikema, S., Haley, B., Taviani, E., Chen, A., Brown, M., Huq, A., & Colwell, R. (2015). Use of environmental parameters to model pathogenic *Vibrios* in the Chesapeake Bay. *Journal of Environmental Informatics, 26*(1), 1–13. https://doi.org/10.3808/jei.201500307

Urquhart, E. A., Jones, S. H., Jong, W. Y., Schuster, B. M., Marcinkiewicz, A. L., Whistler, C. A., & Cooper, V. S. (2016). Environmental conditions associated with elevated Vibrio parahaemolyticus concentrations in Great Bay Estuary. *New Hampshire. Plos One, 11*, e0155018. https://doi.org/10.1371/journal.pone.0155018

Wang, J., & Deng, Z. (2019). Modeling and predicting *fecal coliform* bacteria levels in oyster harvest waters along Louisiana Gulf coast. *Ecological Indicators, 101*, 212–220. https://doi.org/10.1016/j.ecolind.2019.01.013

Williams, T. C., Froelich, B. A., Phippen, B., Fowler, P., Noble, R. T., & Oliver, J. D. (2017). Different abundance and correlational patterns exist between total and presumed pathogenic *V. vulnificus* and *V. parahaemolyticus* in shellfish and waters along the North Carolina coast. *FEMS Microbiology Ecology, 93*(6), fix071, https://doi.org/10.1093/femsec/fix071.

Zhang, Z., Deng, Z., Rusch, K. A., & Walker, N. D. (2015). Modeling system for predicting enterococci levels at Holly Beach. *Marine Environmental Research, 109*, 140–147. https://doi.org/10.1016/j.marenvres.2015.07.003

Zhang, Z., Deng, Z., & Rusch, K. A. (2012). Development of predictive models for determining *Enterococci* levels at Gulf Coast beaches. *Water Research, 46*(2), 465–474. https://doi.org/10.1016/j.watres.2011.11.027

Zimmerman, A., DePaola, A., Bowers, J., Krantz, J., Nordstrom, J., Johnson, C., & Grimes, D. (2007). Variability of total and pathogenic *Vibrio parahaemolyticus* densities in northern Gulf of Mexico water and oysters. *Applied and Environmental Microbiology, 73*(23), 7589–7596. https://doi.org/10.1128/AEM.01700-07

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.