Hybrid modeling and prediction of oyster norovirus outbreaks

Shima Shamkhali Chenar and Zhiqiang Deng

ABSTRACT

This paper presents a hybrid model for predicting oyster norovirus outbreaks by combining the Artificial Neural Networks (ANNs) and Principal Component Analysis (PCA) methods and using the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite remote-sensing data. Specifically, 10 years (2007–2016) of cloud-free MODIS Aqua data for water leaving reflectance and environmental data were extracted from the center of each oyster harvest area. Then, the PCA was utilized to compress the size of the MODIS Aqua data. An ANN model was trained using the first 4 years of the data from 2007 to 2010 and validated using the additional 6 years of independent datasets collected from 2011 to 2016. Results indicated that the hybrid PCA-ANN model was capable of reproducing the 10 years of historical oyster norovirus outbreaks along the Northern Gulf of Mexico coast with a sensitivity of 72.7% and specificity of 99.9%, respectively, demonstrating the efficacy of the hybrid model.

Key words | ANN model, norovirus, outbreak, oyster, PCA, remote sensing

Shima Shamkhali Chenar
Zhiqiang Deng (corresponding author)
Department of Civil and Environmental Engineering,
Louisiana State University,
Baton Rouge, LA 70803,
USA
E-mail: *zdeng@lsu.edu*

HIGHLIGHTS

- A hybrid model is presented for the prediction of oyster norovirus outbreaks.
- The model is based on Artificial Neural Networks and Principal Component Analysis.
- The model greatly expands the spatial coverage of oyster safety monitoring programs.
- The model expands the water quality monitoring frequency from 1 month to 1 day.
- The model input data are satellite remote-sensing data that are freely available.

INTRODUCTION

Norovirus is a highly contagious virus with a low infectious dose, high infectivity, and efficient transmission in natural and human-made environments. Norovirus may concentrate in oysters, which are filter feeders and pump a large volume of water along with particles (such as norovirus) in water through their tissues if the oyster growing water is contaminated with sewage due to overland runoff, combined sewer overflow, and failing septic systems (Wang & Deng 2016). The consumption of raw oysters, therefore, may cause human norovirus infections and even large-scale outbreaks (Siebenga *et al.* 2009). Norovirus outbreaks due to the consumption of contaminated oysters resulted in closures of oyster harvest areas worldwide (David *et al.* 2007; McIntyre *et al.* 2017). For example, norovirus outbreaks in late 2016 and early 2017, in British Colombia, Canada, sickened more than 100 people and caused closures of 13 oyster farms (McIntyre *et al.* 2017). Consequently, it is vital

doi: 10.2166/wh.2021.251

to develop norovirus prediction models for informing decisions and implementing management interventions.

There is substantial evidence in the scientific literature that various environmental factors, such as water temperature, gage height, sunlight, salinity, wind, and precipitation, influence the survival and persistence of norovirus in the aquatic environment (Chenar & Deng 2017a, 2017b). While predicting viral contamination in the open natural environment (particularly coastal waters) is challenging due to unknown sources, complex marine environment (dilution and currents), and data scarcity (Pommepuy *et al.* 2005), machine learning-based models have shown promise in predicting norovirus outbreaks in oyster harvest areas (Wang & Deng 2016). Chenar & Deng (2018a) presented a Genetic Programming (GP)-based model for identifying the environmental conditions that trigger oyster norovirus outbreaks and predicting outbreaks in the Gulf of Mexico using the six environmental indicators, including water temperature, gage height, solar radiation, salinity, rainfall, and wind. Findings indicated that the GP-based model provided an efficient and effective tool for predicting potential oyster norovirus outbreaks. Chenar & Deng (2018b) also created an artificial intelligence (AI)-based forecasting model, called ANN-2Day model with a 2-day lead time, using the same six environmental predictors. The ANN-2Day model can forecast where, when, and under what environmental conditions oyster norovirus outbreaks occur. While the GP and ANN models are effective tools for predicting oyster norovirus outbreaks, applications of the models require daily data of the environmental parameters for all oysters harvest areas, making the *in situ* data collection challenging and even unlikely as the required data are not available daily for all oyster harvest areas along the U.S. Gulf coast. In fact, complete environmental datasets are available only in the areas highlighted in Supplementary Material, Figure S1, including Louisiana oyster harvest areas 1, 2, 3, 6, 7, 12, 13, 14, 15, 16, 18, 20, 24, 26, 27, 28, 29, 30, and Copano Bay and San Antonio Bay in Texas. The scarcity of *in situ* environmental data makes the *in situ* data-based management of oyster safety challenging.

Satellite remote sensing can be a useful technology for monitoring large areas like oyster harvest areas where field measurements are mostly not available. The data derived from NASA (National Aeronautics and Space Administration) EOS Terra/Aqua (MODIS (Moderate Resolution Imaging Spectroradiometer)) sensors (Tatem *et al.* 2004) can provide a robust and cost-effective approach for long-term monitoring of oyster harvest areas. Although norovirus cannot be sensed directly, remote-sensing data can be utilized to determine the environmental indicators of oyster norovirus. In fact, various MODIS spectral bands and their ratios have been widely used to quantify environmental indicators such as chlorophyll-a, colored dissolved organic matters (CDOM), Secchi disk depth, total phosphorus, water temperature, salinity, and gage height (Wang *et al.* 2005; Menken *et al.* 2006; Wu *et al.* 2009; Qing *et al.* 2013; Morozov *et al.* 2015). Wang & Deng (2017, 2018a, 2018b) presented a number of remote-sensing algorithms for retrieving sea surface salinity, sea surface temperature, and gage height using MODIS Aqua data, making it possible to predict oyster norovirus outbreaks using satellite remote-sensing data. In combination with an AI approach MODIS Aqua data, which reflect the characteristics of environmental parameters controlling norovirus epidemics in water, can be a useful data source for predicting human health risks associated with viral outbreaks.

The overall goal of this study is to demonstrate the efficacy of predicting oyster norovirus outbreaks by synergistically combining the AI-based modeling technique and satellite remote-sensing data directly from the MODIS sensor aboard the NASA Aqua spacecraft. To that end, a hybrid Principle Component Analysis (PCA) and Artificial Neural Network (ANN) modeling approach was developed in this study to establish a predictive relationship between oyster norovirus outbreaks and MODIS Aqua data from ocean color bands.

MATERIALS AND METHODS

Data collection and processing

Two types of time series data were collected, including epidemiological data for historical oyster norovirus outbreaks and remote-sensing data that reflect the characteristics of environmental parameters controlling epidemics. NASA launched the MODIS as a key instrument aboard the Terra and Aqua satellites. MODIS plays a vital role in

providing critical data necessary to monitor global changes to assist policymakers in protecting the environment (https://modis.gsfc.nasa.gov/about/). Terra passes from north to south across the equator at 10:30 AM local time, while Aqua passes south to north over the equator at 1:30 PM local time (https://modis.gsfc.nasa.gov/), making it possible for Aqua to capture more cloud-free images with high radiometric sensitivity (12 bit). Aqua has two spectral bands at a resolution of 250 m (red and near-infrared), five bands at 500-m resolution (blue, green, near-infrared, and mid-infrared), and 29 specialized bands at 1,000-m resolution (consisting of nine bands designed for ocean color applications, and thermal infrared bands for surface temperature measurement) (https://modis.gsfc.nasa.gov/about/specifications.php).

In this study, fully corrected or processed MODIS Aqua products, including Ocean Color and Sea Surface Temperature from 2007 to 2016 in which historical norovirus outbreak reports were available, were downloaded from the Ocean Color WEB (https://oceancolor.gsfc.nasa.gov/). The SeaWiFS Data Analysis System (SeaDAS) was used to extract spectral bands and other geophysical data (Table 1) from the center of each oyster harvest area (Supplementary Material, Figure S1) as oyster safety is managed on an area-by-area basis without considering the variability in the environmental parameters within individual oyster harvest areas due to the lack of detailed ground-truth data. The Calculate Geometry tool in ArcGIS was used to find the centroid of all oyster harvest areas for each cloud-free day

Table 1 | Spectral bands and geophysical data extracted from MODIS Aqua level 2 products

No.	Parameters	Norovirus indicators	Supporting References
1	Band 1	Gage height	Wang & Deng (2018b)
2	Band 3	Salinity	Wang & Deng (2018a)
3	Band 4	Gage height	Qing *et al.* (2013) and Wang & Deng (2018b)
4	Band 8	Gage height	Urquhart *et al.* (2012), Qing *et al.* (2013), and Wang & Deng (2018b)
5	Band 9	Salinity	Urquhart *et al.* (2012), Qing *et al.* (2013), and Wang & Deng (2018a)
6	Band 10	Gage height	Urquhart *et al.* (2012), Qing *et al.* (2013), and Wang & Deng (2018b)
7	Band 11	Salinity	Wang & Deng (2018a)
8	Band 12	Gage height	Wang & Deng (2017)
9	Band 13	Salinity	Urquhart *et al.* (2012), Qing *et al.* (2013), and Wang & Deng (2018a)
10	Band 14	Gage height	Wang & Deng (2018b)
11	Aerosol optical thickness at 869 nm	Solar radiation	Lee *et al.* (2013) and Chen *et al.* (2014)
12	Aerosol angstrom exponent		
13	Diffuse attenuation coefficient at 490 nm		
14	Instantaneous photosynthetically available radiation		
15	Normalized fluorescence line-height		
16	Particulate organic carbon (POC)	Indirect relationship with the food chain of oyster	
17	Chlorophyll-a concentration		
18	Sea surface temperature (SST)	Temperature	Wang *et al.* (2005), Handcock *et al.* (2006), Chipman *et al.* (2009), Gholizadeh *et al.* (2016), and Wang & Deng (2017)
19	Longitude	Gage height Salinity Temperature	Wang & Deng (2017, 2018a, 2018b)
20	Latitude	Gage height Salinity Temperature	Wang & Deng (2017, 2018a, 2018b)

from 2007 to 2016. In total, 6,454 data points were extracted for the model development, 3,281 for training and 3,173 for the testing phase (Supplementary Material, Table S4). In addition, 20 independent parameters containing spectral and geophysical data were extracted from each image. To develop a remote-sensing based prediction model, the possible associations between MODIS data and environmental indicators controlling oyster norovirus outbreaks were investigated and summarized in Table 1. Furthermore, spectral band ratios were produced as input variables to reduce atmospheric, and air–water surface influences on the remotely sensed signal (Dekker & Peters 1993; Lillesand *et al.* 2014). Band ratios are the most common form of strengthening the spectral differences between bands and reducing the effects of topography. Supplementary Material, Table S1 shows the 45 produced band ratios. To minimize the systematic difference among spectral bands and other geophysical parameters, all data associated with 65 input variables were normalized using Equation (1) so that all normalized data (x') vary in the same range of 0–1 based on historical maximum and minimum values of individual environmental data from 2007 to 2016. To that end, individual parameters (x) were ranked in an ascending order from 2007 to 2016, and the smallest and largest values were reported as historical minimum ($x_{\min}$) and maximum ($x_{\max}$) values. Historical norovirus outbreak records, associated with the consumption of raw oysters, were obtained from various online data sources (such as Louisiana Morbidity Reports released by Louisiana Department of Health (http://www.dhh.la.gov/index.cfm/newsroom/category/126)) and presented in Supplementary Material, Table S2.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

Principal Component Analysis

One of the important steps involved in the AI-based model development, called 'feature reduction', is to reduce the dimensionality of data with neural networks (Kecman 2005). Feature extraction is one of the approaches to achieve feature reduction that transforms, linearly or nonlinearly, the original set of features into a reduced one (Ivosev *et al.* 2008; Van *et al.* 2009). The Principal Component Analysis (PCA) is a popular multivariate statistical technique that transforms a number of correlated inputs into a smaller number of variables called principal components (PCs) (Richardson 2009). The purpose of applying PCA in this study is to extract the information from the MODIS Aqua data and compress the size of the dataset by keeping this important information. PCA computes new independent and linear compounds of input variables, called PCs, which are used instead of original input variables (Richardson 2009). The first PC is required to have the largest possible variance, while the second PC is computed under the constraint of being orthogonal to the first PC and to have the largest possible inertia (Abdi & Williams 2010). The other PCs are calculated likewise. PCs can be defined as:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \tag{2}$$

where Y_i represents PCs, a_i is related to eigenvectors (the coefficients for the formation of PCs) and X_i is input variables.

To assess the suitability of the data for PCA analysis, Kaiser–Meyer–Olkin (KMO) and Bartlett's tests were performed. The KMO index is a measure of sampling adequacy varying from 0 to 1 with values greater than 0.70 being considered suitable for PCA analysis (Budaev 2010). Bartlett's test of sphericity checks if there is a redundancy between variables that can be summarized with some components. When the test is significant ($p < 0.05$), it can be assumed that the correlation matrix is significantly different from an identity matrix and PCA can be applied (Williams *et al.* 2010). Cattell's screen test was used to determine the number of factors to retain. The test involves a visual examination of the eigenvalue plot in which only the components, forming the elbow before the slope of the graph goes from steep to flat, are kept (Budaev 2010). In addition, Varimax orthogonal rotation, which is the most popular rotation method (Kaiser 1958), was specified in this study. The statistical software package IBM SPSS was used to implement this technique.

ANN model

The AI-based ANNs method has been increasingly applied in classification, clustering, prediction, and many other areas (Du 2010; Khashei & Bijari 2012; Castellani 2013; Wang & Deng 2016, 2018a, 2018b; Chenar & Deng 2017). ANN has also been proven to be an effective tool for describing nonlinear relationships between norovirus outbreaks and environmental variables in coastal waters (Wang & Deng 2016; Chenar & Deng 2017). In this study, a three-layer feed-forward neural network with back-propagation learning was constructed to develop a model for predicting oyster norovirus outbreaks using MODIS data. The architecture of the ANN model includes an input layer (receiving model input data), a single hidden layer (consisting of 20 neurons), and an output layer (showing the norovirus outbreak risk in any oyster harvest area for a given day). The number of hidden neurons depends on the training pattern, and there is no general rule for it in ANN (Ahmad *et al.* 2017). The neuron number of 20 was selected through a trial and error procedure by the re-training model using 5, 10, 15, 20, and more neurons. It was found that the neuron numbers higher than 20 would not significantly improve the model performance in terms of true positives, false positive, false negative, and true negative rates. Therefore, 20 neurons were chosen to reduce the model complexity. The ANN model training and testing processes were performed using the Neural Network Toolbox in the MATLAB program. The ANN model was created using three subsets of the available data, including the training dataset to adjust the weights, the validation dataset to measure network generalization, and the testing dataset to assess the performance. This data-partitioning step was conducted to evaluate the ability of the model to generalize through the comparison of predictions with the remaining data that were not used in the training process. Specifically, the normalized datasets from 2007 to 2010 were randomly split into three groups for training (accounting for 60% of the datasets), validation (20%), and testing (20%). Random data division improves model performance and prevents overfitting (Palani *et al.* 2008). The best-trained ANN model was identified based on the performance of top-ranked models in forecasting oyster norovirus outbreaks. The ANN model predictions were then compared with confirmed historical norovirus outbreaks to determine a threshold value for model-predicted norovirus outbreak risks that were consistently associated with confirmed outbreaks. Moreover, the cross-validation was performed to assess the predictive ability of the model using the independent data collected from 2011 to 2016 that were excluded from the model development phase. Finally, the overall performance of the model was evaluated using true positive and true negative rates.

A hybrid PCA-ANN prediction model was created by using the outputs of candidate ANN models as inputs of the optimum PCA model that accurately predicted all historical oyster norovirus outbreaks with a minimum number of false outbreaks. Specifically, the 15 PCs were initially used as input variables of ANN models, producing 100 candidate ANN models (outputs Y1–Y100) (Supplementary Material, Figure S2). The 100 initially trained ANN models were then used as inputs to generate 50 refined ANN models (Z1–Z50). Finally, the 50 refined ANN models were utilized as the model input variables to develop the hybrid PCA-ANN model for predicting norovirus outbreaks (NOV). Supplementary Material, Figure S2 summarizes the steps involved in the development of the hybrid PCA-ANN prediction model.

The overall performance of the hybrid PCA-ANN model was evaluated using a confusion matrix. The number of days with reported outbreaks and the number of days without reported outbreaks were counted from 2007 to 2016. The days, on which the model predicted risk higher than the threshold but no outbreaks were reported, were labeled as false positives and the days, on which norovirus outbreaks were reported but the model predicted risk lower than the threshold, were treated as false negatives. Finally, true positive and negative rates and positive and negative predictive values were calculated.

RESULTS

Principle Component Analysis

Supplementary Material, Table S3 summarizes the results from the KMO and Bartlett tests. It can be seen from the table that the KMO statistic is 0.88 (close to 1), confirming the capability of using the PCA method in reducing the

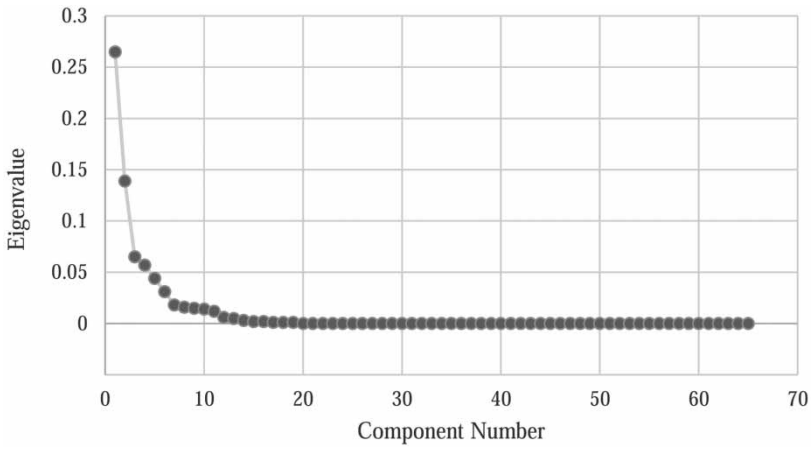

Figure 1 | Cattell's screen test plot.

number of input variables. According to Bartlett's test, the significance level is 0 ($p < 0.05$), indicating that the dataset is suitable for reduction and PCA can be useful for the dataset. The first 15 PCs were selected as inputs of the ANN model based on Cattell's screen test (Figure 1) and the cumulative variance proportion explained by the components. In fact, the interpretation of a screen test plot is subjective, requiring the judgment of the researcher to select the number of optimum PCs in such a way that it would adequately describe the input variable characteristics (Thompson 2004; Tabachnick & Fidell 2007). Table 2 presents eigenvalues, variance proportion, and cumulative variance proportion for the first 15 PCs. According to the table, it is clear that the first 15 PCs (PC1–PC15) represent 96.59% of the total variance proportion of input variables. According to the eigenvector values, the band ratios, involving spectral bands 8, 9, and 13, have the significant effects on the PC1 that described more than 53% of the variance proportion of the input variables. According to Table 1, bands 8, 9, and 13 were utilized to retrieve salinity and gage height from MODIS satellite data. Furthermore, band 9, which has the most strong impact on the second component (PC2), explains more than 15% of the variable variance. Likewise, PC3 is affected by the normalized fluorescence line-height, which could be considered as a solar radiation indicator, as Normalized Fluorescence Line-Height is a measure of the solar stimulated chlorophyll-a fluorescence.

Table 2 | Descriptive statistics and total variance of the created PCs

Components	Eigenvalue	Variance proportion (%)	Cumulative variance proportion (%)
PC1	34.51	53.09	53.09
PC2	10.29	15.84	68.93
PC3	4.47	6.87	75.80
PC4	2.07	3.19	78.99
PC5	1.79	2.75	81.74
PC6	1.63	2.51	84.25
PC7	1.43	2.20	86.45
PC8	1.21	1.85	88.30
PC9	1.19	1.83	90.13
PC10	0.84	1.29	91.42
PC11	0.82	1.26	92.67
PC12	0.76	1.16	93.84
PC13	0.70	1.08	94.91
PC14	0.61	0.94	95.85
PC15	0.48	0.74	96.59

Hybrid model development

The hybrid PCA-ANN model was selected from the top five ranked candidate models in terms of true positive and negative rates. A model-based threshold risk of 0.5, which consistently predicted the reported outbreaks, was selected by comparing predicted risks of norovirus outbreaks based on the ANN model with the occurrence of observed epidemics. Figure 2 shows the comparison between the PCA-ANN model-predicted risks of norovirus outbreaks and the

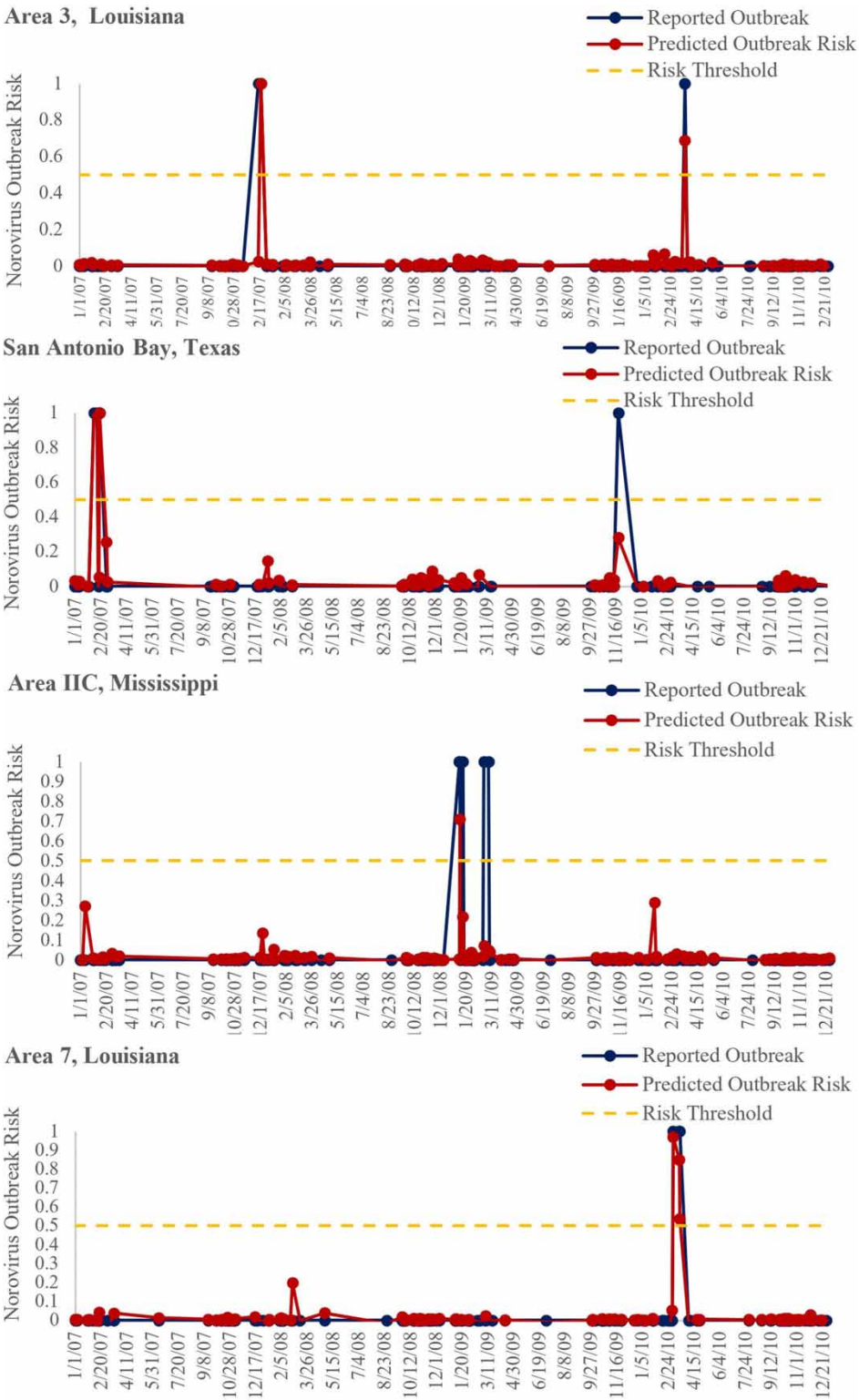

Figure 2 | Time series plots showing the norovirus outbreak risks predicted with the PCA-ANN model and the confirmed norovirus outbreaks in oyster harvest areas along the Northern Gulf of Mexico from 2007 to 2010 in Louisiana Areas 3, 7, Mississippi Area IIC, and Texas San Antonio Bay.

observed norovirus outbreaks in oyster harvest areas along the U.S. Gulf of Mexico coast.

The PCA-ANN model predicted two clusters of norovirus outbreaks in Louisiana Area 3 for 12 December 2007 and 3 March 2010 with risks of 1 and 0.69, respectively. Furthermore, Figure 2 shows daily risks of norovirus outbreaks in the San Antonio Bay oyster harvest area where two oyster norovirus outbreaks were reported for 1–24 February 2007 and 16–25 November 2009, respectively. The PCA-ANN model was capable of reproducing the first confirmed outbreaks for 7, 16, and 18 February 2007 with the risk of 1, but the model did not predict the second reported outbreak in November 2009. The PCA-ANN model was capable of reproducing one out of the two confirmed outbreaks for January 8 with the risk of 0.71 in the Mississippi oyster harvesting Area IIC where two oyster norovirus outbreaks were reported in January and February 2009 (Supplementary Material, Table S2). Figure 2 demonstrates that the PCA-ANN model predicted multiple norovirus outbreaks for 6, 18, and 25 March 2010 with the risks higher than the threshold of 0.5, which were consistent with the inferred norovirus outbreak period of 6 and 24 March 2010 in Louisiana Area 7. There were no reported norovirus outbreaks from 2007 to 2010 in areas 1, 2, 10, 11, 12, 14, 15, 19, 21, 24, 25, 28, and 29 and the PCA-ANN model did not produce any false outbreaks for those areas (Figure 2).

To validate the performance of the PCA-ANN model, the additional 6 years of environmental and epidemiological data from 2011 to 2016, which were excluded from the model development, were employed for cross-validation of the model. Figure 3 illustrates the cross-validation results of the PCA-ANN model with the 6 years of independent data collected from Louisiana Areas 23 and 30 and Texas oyster harvest area in Copano Bay. The PCA-ANN model predicted the risk of 0.76 for 24 April 2012, during the reported outbreak period, as illustrated in Figure 3 for Louisiana Area 23. It can be seen from Figure 3 that the model did not predict the reported norovirus outbreak in Area 30 that was closed on 4 January 2013 after 12 people were infected by norovirus due to eating raw oysters harvested from this area between 28 December 2012 and 4 January 2013. There was a reported norovirus outbreak between 26 December 2013 and 9 January 2014 in Copano Bay, Texas. The PCA-ANN model predicted a norovirus outbreak with the risk of 0.61 for 3 January 2014 during the reported outbreak period, as shown in Figure 3. The PCA-ANN model also predicted two unconfirmed norovirus outbreaks for this area for 2 November 2015 and 13 January 2016. Therefore, these unconfirmed norovirus outbreaks could be a true outbreak that was not reported or a false cluster of outbreaks due to the absence of viruses or a source of fecal contamination in the environment. The model was further validated for areas 1, 2, 10, 11, 12, 14, 15, 19, 21, 24, 25, 28, and 29 where no confirmed outbreaks occurred from 2011 to 2016. The PCA-ANN model did not produce any false outbreaks for these areas.

In terms of the model performance, the total number of true positives, false positive, false negative, and true negative were 8, 1, 3, and 10,304, respectively. The true positive and negative rates of model predictions were 72.7 and 99.9%, and positive and negative predictive values were 88.9 and 99.9%, respectively, demonstrating the ability of the model to predict the outbreaks. The relatively low true positive rate may be caused by the fact that some important environmental predictors, such as rainfall and wind (Chenar & Deng 2018a, 2018b), are not included in the new model.

DISCUSSION

While an oyster norovirus outbreak may theoretically occur at any risk level, our findings, which are based on a comparison of the hybrid PCA-ANN model predictions with reported norovirus outbreaks, indicate that reported historical oyster norovirus outbreaks in the human population were consistently associated with the risk range of 0.5–1.0. Therefore, the risk of 0.5 was selected as the risk threshold value for oyster norovirus outbreaks. It should be noted that the oyster norovirus outbreaks predicted with the hybrid PCA-ANN model only refer to the norovirus outbreaks associated with the consumption of contaminated oysters. The predicted oyster norovirus outbreaks do not include the human norovirus outbreaks caused by the spread of norovirus from person to person. It should also be pointed out that a lower norovirus outbreak risk (such as 0.55) predicted with the hybrid PCA-ANN model might cause a lower number of human infections. However, a quantitative relationship between the model-predicted

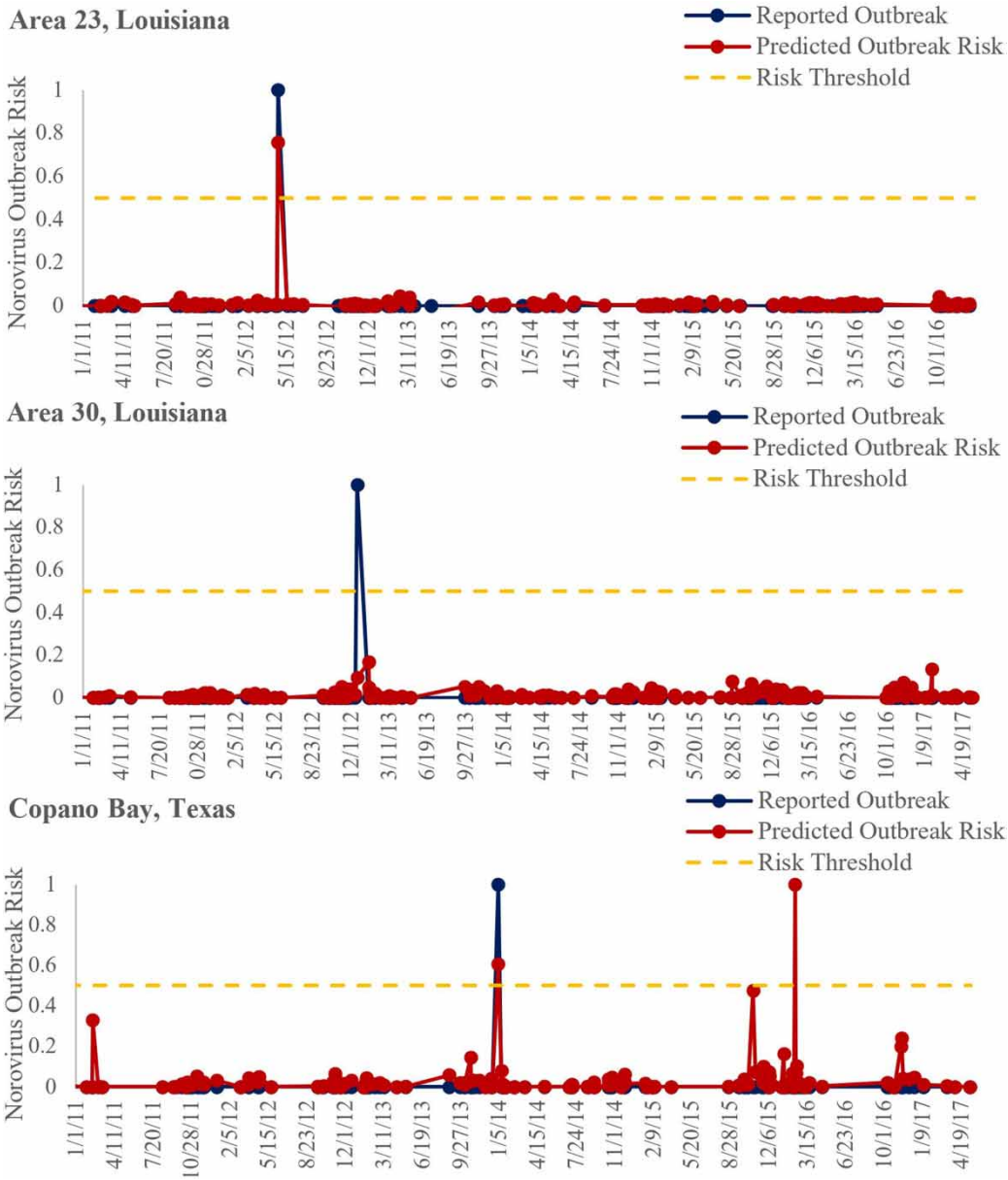

Figure 3 | Cross-validation time series plots showing the norovirus outbreak risks predicted with the PCA-ANN model and the confirmed norovirus outbreaks for the period of 2010–2016 in Louisiana Areas 23, 30, and Texas Copano Bay.

outbreak risk and the number of human infections is not available at this time due to the lack of detailed data for the norovirus concentration in oysters and associated epidemiological data. As a result, a model prediction can only be interpreted as an outbreak if the predicted risk is in the range of 0.50–1.00 or no outbreak if the predicted risk is in the range of 0.00–0.49. More efforts are needed to link a model-predicted risk level to a human infection level.

The application of the PCA method reduced the number of input variables of the ANN model from 65 to 15 without eliminating useful information, reducing model training time, and improving model performance. The majority of remote-sensing band ratios were loaded on the first component, demonstrating the efficiency of using band ratios in the model development. The PCA method was effective in terms of reducing and identifying the input variables for the

ANN model due to the lack of knowledge of the nonlinear relationship between an oyster norovirus outbreak and its environmental drivers described with MODIS band data in this study. Further studies in close consultations with norovirus epidemiologists and caseworkers are needed to improve understanding of the correlation between remotely sensed data and oyster norovirus outbreaks in coastal waters.

As compared with the nowcasting and forecasting models developed previously by the authors' research group (Wang & Deng 2016; Chenar & Deng 2018a, 2018b), a unique advantage and important feature of the hybrid PCA-ANN remote-sensing-based model is its capability of making daily prediction of the oyster norovirus outbreak risks for all oyster harvesting areas along the U.S. Gulf coast as long as the MODIS ocean color data are available, enabling managers to make timely decisions on oyster safety due to the expansion of the spatial coverage and temporal frequency of oyster safety monitoring. Specifically, the hybrid PCA-ANN model can be employed to predict norovirus outbreaks as long as MODIS Aqua satellite data are available. The PCA-ANN model expands the spatial coverage of oyster safety monitoring to larger geographical areas where complete *in situ* environmental datasets are not available. Due to the daily prediction, the PCA-ANN model also serves as an early warning system since oyster norovirus outbreaks in the human population generally occur 1–2 weeks (including the incubation period of 12–48 h for norovirus-associated gastroenteritis) after norovirus-contaminated oysters are harvested (Wang & Deng 2016). Hence, the hybrid PCA-ANN and remote-sensing-based model can be used as an effective modeling framework for improving the management of oyster safety and public health by ensuring that oysters are safe to harvest. Figure 4 shows how the remote-sensing data-based PCA-ANN model could be applied in combination with other *in situ* data-based models, developed by the authors' research group, for prediction of oyster norovirus outbreaks, and how the model predictions could be utilized to implement management interventions, such as field sampling, laboratory analysis, and oyster bed closures, for protecting public health while minimizing oyster ground closures.

Since a model may produce false predictions or alerts, as shown in Figures 2 and 3 (particularly San Antonio Bay in December 2009, Mississippi Area IIC in March 2009, and Louisiana Area 30 in December 2012), and the false alert-based closure of an oyster harvest area might cause substantial economic damage to oyster farmers, a sound-science-based procedure (Figure 4) should be followed in using the model predictions. Depending on the type of available data, the PCA-ANN model or the models based on *in situ* measurement of environmental data could be run to predict daily risks of oyster norovirus outbreaks. If a model-predicted risk of norovirus outbreak exceeds a predefined outbreak risk threshold, oyster harvesting in the implicated area should be temporarily suspended and sampling should be conducted to confirm or deny the model prediction. Specifically, the following procedure should be followed, whenever a model-predicted risk of norovirus outbreak exceeds a predefined outbreak risk threshold to prevent oyster norovirus outbreaks:

1. Water and oyster samples should be collected from the oyster harvest areas where the model predicts a potential outbreak;
2. Laboratory analyses of the water and oyster samples should be conducted to verify the presence or absence of norovirus in oysters; and
3. Implicated harvest areas should be closed if the laboratory results confirm the presence of norovirus in oysters.

This is the first major study on predicting potential oyster norovirus contamination in coastal waters by using satellite

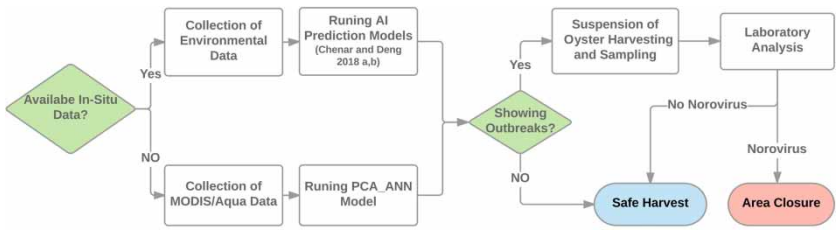

Figure 4 | The recommended management intervention procedure for oyster norovirus outbreaks.

remote-sensing data. While this study focused on the oyster harvest areas along the U.S. Gulf of Mexico coast, the remote-sensing-based hybrid modeling approach could be potentially applied to other regions and countries due to the global coverage of the NASA MODIS satellite data. Therefore, this study could be expanded in the future toward the development of a global-scale hybrid modeling system for providing daily predictions of norovirus outbreak risks in any oyster harvest areas, serving as a global-scale early alert system for potential oyster norovirus outbreaks and thus protecting public health. The model enables oyster management authorities to expand the monitoring and prediction of norovirus outbreak risks from areas where monitoring data are accessible to other oyster harvest areas where monitoring stations are not available. Although this study provided convincing evidence that predicting oyster norovirus outbreaks using satellite remote-sensing data are possible, field sampling-based validation of model prediction is always needed. A major limitation of the PCA-ANN model application is that the model may not work during the cloud cover days when the remote-sensing data, specifically MODIS Aqua data, are not available. Also, it appears from Figures 2 and 3 that the PCA-ANN model works well for Louisiana oyster harvest areas as the model was developed primarily using the data from the Louisiana oyster harvest areas. However, the model tends to produce more false predictions for other oyster harvest areas beyond Louisiana, although the model still works for the other areas with a lower efficacy. It is, therefore, recommended that the modeling and intervention procedure shown in Figure 4 be followed, by using both the *in situ* data and the remote-sensing data, to protect public health even during cloudy days. While the application of the model to other norovirus concerned oyster harvest areas like those in British Columbia may produce a higher rate of false predictions and thus require additional testing of the model with local data, the modeling methodology can be adopted for other norovirus concerned oyster farms to monitor any potential oyster safety threats.

CONCLUSIONS

This paper presents a hybrid monitoring and modeling approach for monitoring and predicting potential risks of oyster norovirus outbreaks in the human population. In terms of monitoring, high-resolution satellite remote-sensing data covering all oyster harvesting areas were utilized to provide daily data needed in modeling. In terms of modeling, a hybrid PCA and ANN model was created for the daily prediction of oyster norovirus outbreak risks and demonstrated using the data from the Northern Gulf of Mexico that is prone to epidemics. Based on the results of this paper, the following conclusions can be drawn:

- A remote-sensing-based PCA-ANN model is an effective tool for predicting oyster norovirus outbreaks. Specifically, the PCA-ANN model achieved the true positive and negative rates of 72.7 and 99.9%, respectively, in predicting reported oyster norovirus outbreaks, demonstrating the efficacy of the model in predicting outbreaks, and minimizing the adverse impact of outbreaks on human health and the shellfish industry.
- The PCA-ANN model greatly expands the spatial coverage of oyster safety monitoring programs from limited areas where *in situ* monitoring data are available to all oyster harvest areas including areas where *in situ* monitoring data are not available.
- The PCA-ANN model also greatly expands the temporal frequency of water quality monitoring regularly conducted by oyster safety monitoring programs from 1 month to 1 day, enabling oyster safety monitoring programs to make daily decisions on oyster safety and respond effectively to any potential oyster safety threats by implementing management interventions.
- Field validation is always needed, particularly when the PCA-ANN model predicts a high risk of an oyster norovirus outbreak, to confirm the presence of model-predicted norovirus outbreak in oyster harvest areas.

ACKNOWLEDGEMENTS

The material is based upon work supported by the U.S. NASA (National Aeronautics and Space Administration: award No. 80NSSC20M0216) and the Louisiana Board of Regents (LEQSF(2020-23)-Phase3-14). We appreciate the insightful comments from anonymous reviewers and particularly those from Charmaine Enns, Eleni Galanis,

Natalie Prystajecky, Jackie Plamondon, Lorraine McIntyre, and Theresa Burns of the British Columbia Center for Disease Control and Prevention, Canada, illuminating the final form of this paper.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Abdi, H. & Williams, L. J. 2010 Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2** (4), 433–459.

Ahmad, M. W., Mourshed, M. & Rezgui, Y. 2017 Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings* **147**, 77–89.

Budaev, S. V. 2010 Using principal components and factor analysis in animal behaviour research: caveats and guidelines. *Ethology* **116** (5), 472–480.

Castellani, M. 2013 Evolutionary generation of neural network classifiers – an empirical comparison. *Neurocomputing* **99**, 214–229.

Chen, M., Zhuang, Q. & He, Y. 2014 An efficient method of estimating downward solar radiation based on the MODIS observations for the use of land surface modeling. *Remote Sensing* **6** (8), 7136–7157.

Chenar, S. S. & Deng, Z. 2017a Environmental indicators for human norovirus outbreaks. *International Journal of Environmental Health Research* **27** (1), 40–51.

Chenar, S. S. & Deng, Z. 2017b Environmental indicators of oyster norovirus outbreaks in coastal waters. *Marine Environmental Research* **130**, 275–281.

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

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

Chipman, J. W., Olmanson, L. G. & Gitelson, A. A. 2009 *Remote Sensing Methods for Lake Management: A Guide for Resource Managers and Decision-Makers*. North American Lake Management Society.

David, S. T., McIntyre, L., MacDougall, L., Kelly, D., Liem, S., Schallié, K., McNabb, A., Houde, A., Mueller, P., Ward, P. & Trottier, Y. L. 2007 An outbreak of norovirus caused by consumption of oysters from geographically dispersed harvest sites, British Columbia, Canada, 2004. *Foodborne Pathogens and Disease* **4** (3), 349–358.

Dekker, A. & Peters, S. 1993 The use of the Thematic Mapper for the analysis of eutrophic lakes: a case study in the Netherlands. *International Journal of Remote Sensing* **14** (5), 799–821.

Du, K.-L. 2010 Clustering: a neural network approach. *Neural Networks* **23** (1), 89–107.

Gholizadeh, M. H., Melesse, A. M. & Reddi, L. 2016 A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors* **16** (8), 1298.

Handcock, R., Gillespie, A., Cherkauer, K., Kay, J., Burges, S. & Kampf, S. 2006 Accuracy and uncertainty of thermal-infrared remote sensing of stream temperatures at multiple spatial scales. *Remote Sensing of Environment* **100** (4), 427–440.

Ivosev, G., Burton, L. & Bonner, R. 2008 Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry* **80** (13), 4933–4944.

Kaiser, H. F. 1958 The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23** (3), 187–200.

Kecman, V. 2005 Support vector machines – an introduction. In: *Support Vector Machines: Theory and Applications* (L. Wang, ed.), Springer, Berlin, pp. 1–47, https://doi.org/10.1007/b95439.

Khashei, M. & Bijari, M. 2012 Hybridization of the probabilistic neural networks with feed-forward neural networks for forecasting. *Engineering Applications of Artificial Intelligence* **25** (6), 1277–1288.

Lee, Z., Hu, C., Shang, S., Du, K., Lewis, M., Arnone, R. & Brewin, R. 2013 Penetration of UV-visible solar radiation in the global oceans: insights from ocean color remote sensing. *Journal of Geophysical Research: Oceans* **118** (9), 4241–4255.

Lillesand, T., Kiefer, R. W. & Chipman, J. 2014 *Remote Sensing and Image Interpretation*. John Wiley & Sons, Hoboken.

McIntyre, L., Galanis, E., Prystajecky, N. & Kosatsky, T. 2017 BC oysters and norovirus: hundreds of cases in months with an 'r'. *BC Medical Journal* **59** (6), 326–327.

Menken, K. D., Brezonik, P. L. & Bauer, M. E. 2006 Influence of chlorophyll and colored dissolved organic matter (CDOM) on lake reflectance spectra: implications for measuring lake properties by remote sensing. *Lake and Reservoir Management* **22** (3), 179–190.

Morozov, E., Kondrik, D., Fedorova, A., Pozdnyakov, D., Tang, D. & Pettersson, L. 2015 A spaceborne assessment of cyclone impacts on Barents Sea surface temperature and chlorophyll. *International Journal of Remote Sensing* **36** (7), 1921–1941.

Palani, S., Liong, S.-Y. & Tkalich, P. 2008 An ANN application for water quality forecasting. *Marine Pollution Bulletin* **56**, 1586–1597.

Pommepuy, M., Hervio-Heath, D., Caprais, M.-P., Gourmelon, M., Le Saux, J.-C. & Le Guyader, F. 2005 Fecal contamination in coastal areas: an engineering approach. In: *Oceans and Health: Pathogens in the Marine Environment* (S. Belkin & R. R. Colwell, eds). Springer, Boston, pp. 331–359.

Qing, S., Zhang, J., Cui, T. & Bao, Y. 2013 Retrieval of sea surface salinity with MERIS and MODIS data in the Bohai Sea. *Remote Sensing of Environment* **136**, 117–125.

Richardson, M. 2009 *Principal Component Analysis*. Available from: http://people maths ox ac uk/richardsonm/SignalProcPCA pdf (accessed 3 May 2013). Aleš Hladnik Dr, Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales hladnik@ ntf uni-lj si.

Siebenga, J. J., Vennema, H., Zheng, D.-P., Vinjé, J., Lee, B. E., Pang, X.-L., Ho, E. C., Lim, W., Choudekar, A. & Broor, S. 2009 Norovirus illness is a global problem: emergence and spread of norovirus GII.4 variants, 2001–2007. *Journal of Infectious Diseases* **200** (5), 802–812.

Tabachnick, B. G. & Fidell, L. S. 2007 *Using Multivariate Statistics*. Pearson Education. Inc, Boston, MC.

Tatem, A. J., Goetz, S. J. & Hay, S. I. 2004 Terra and Aqua: new data for epidemiology and public health. *International Journal of Applied Earth Observation and Geoinformation* **6** (1), 33–46.

Thompson, B. 2004 *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. American Psychological Association.

Urquhart, E. A., Zaitchik, B. F., Hoffman, M. J., Guikema, S. D. & Geiger, E. F. 2012 Remotely sensed estimates of surface salinity in the Chesapeake Bay: a statistical approach. *Remote Sensing of Environment* **123**, 522–531.

Van Der Maaten, L., Postma, E. & Van den Herik, J. 2009 Dimensionality reduction: a comparative. *Journal of Machine Learning Research* **10** (66–71), 13.

Wang, J. & Deng, Z. 2016 Modeling and prediction of oyster norovirus outbreaks along Gulf of Mexico coast. *Environmental Health Perspectives* **124** (5), 627.

Wang, J. & Deng, Z. 2017 Development of MODIS data-based algorithm for retrieving sea surface temperature in coastal waters. *Environmental Monitoring and Assessment* **189** (6), 286.

Wang, J. & Deng, Z. 2018a Development of a MODIS data based algorithm for retrieving nearshore sea surface salinity along the Northern Gulf of Mexico coast. *International Journal of Remote Sensing* **39** (11), 3497–3511.

Wang, J. & Deng, Z. 2018b Development of a MODIS data-based algorithm for retrieving gage height in nearshore waters along the Louisiana Gulf coast. *Journal of Coastal Research* **34** (1), 220–228. https://doi.org/10.2112/JCOASTRES-D-16-00161.1.

Wang, K., Wan, Z., Wang, P., Sparrow, M., Liu, J., Zhou, X. & Haginoya, S. 2005 Estimation of surface long wave radiation and broadband emissivity using Moderate Resolution Imaging Spectroradiometer (MODIS) land surface temperature/emissivity products. *Journal of Geophysical Research: Atmospheres* **110**, D11109.

Williams, B., Onsman, A. & Brown, T. 2010 Exploratory factor analysis: a five-step guide for novices. *Australasian Journal of Paramedicine* **8** (3), 990399. https://doi.org/10.33151/ajp.8.3.93.

Wu, M., Zhang, W., Wang, X. & Luo, D. 2009 Application of MODIS satellite data in monitoring water quality parameters of Chaohu Lake in China. *Environmental Monitoring and Assessment* **148** (1), 255–264.

First received 1 November 2020; accepted in revised form 18 February 2021. Available online 12 March 2021