4-24-2019

A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading

Jian Shen
Virginia Institute of Marine Science

Qubin Qin
Virginia Institute of Marine Science

Ya Wang

Mac Sisson
Virginia Institute of Marine Science

A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading

Jian Shen[1]*, Qubin Qin[1], Ya Wang[2], and Mac Sisson[1]

[1]Virginia Institute of Marine Science, College of William & Mary,

Gloucester Point, VA 23062, USA

[2]The Third Institution of Oceanography, Ministry of Natural Resources, Xiamen, China

*Corresponding Author:

Jian Shen

Virginia Institute of Marine Science

College of William & Mary

Gloucester Point, VA 23062

Email: shen@vims.edu

Phone: (804) 684-7359

Highlights:

1. A successful application of the Support vector machine (SVM) for algal bloom simulation provides new approach for predicting harmful algal bloom.
2. Combining Empirical Orthogonal Function and SVM enables simulations of algal blooms for the entire tidal freshwater region in one model.
3. Applying variable transformation is crucial for improving model predictive skill.
4. The data-driven model is capable of assessing algal blooms responding to changes of nutrients if it is trained appropriately.

Abstract

Algal blooms often occur in the tidal freshwater (TF) of the James River estuary, a tributary of the Chesapeake Bay. The timing of algal blooms correlates highly to a summer low-flow period when residence time is long and nutrients are available. Because of complex interactions between physical transport and algal dynamics, it is challenging to predict interannual variations of bloom correctly using a complex eutrophication model without having a high-resolution model grid to resolve complex geometry and an accurate estimate of nutrient loading to drive the model. In this study, an approach using long-term observational data (from 1990-2013) and the Support vector machine (LS-SVM) for simulating algal blooms was applied. The Empirical Orthogonal Function was used to reduce the data dimension that enables the algal bloom dynamics for the entire TF to be modeled by one model. The model results indicate that the data-driven model is capable of simulating interannual algal blooms with good predictive skills and is capable of forecasting algal blooms responding to the change of nutrient loadings and environmental conditions. This study provides a link between a conceptual model and a dynamic model, and demonstrates that the data-driven model is a good approach for simulating algal blooms in this complex environment of the James River. The method is very efficient and can be applied to other estuaries as well.

Keywords: Water quality model; Support vector machine; algal bloom simulation; tidal freshwater; James River.

1. Introduction

The tidal freshwater (TF) region is located in the upstream of an estuary where tidal forcings extend inland but beyond the limits of salinity intrusion. The TF ecosystem is highly influenced by freshwater discharge and the net transport is downstream, even as it experiences tidal fluctuations. The TF is often associated with complex geometry involving a meandering channel with irregular channel depth and cross-section. The interactions of complex geometry and the large fluctuation of freshwater discharge result in strong seasonal variations of dynamic conditions, which can alter pollutant transport and algal growth (Bukaveckas et al., 2017; Shen et al., 2016). Although the TF only accounts for a small portion of an estuary, it interfaces with the drainage basin and is sensitive to any perturbations occurring in the drainage basin. The seasonal and interannual changes of the retention time of pollutants can have a profound impact on the downstream estuary (Paerl, 2009). As it is located in a transition zone between river and estuary, the change of ecosystem in the TF is indicative of the changes in land use applications in the drainage area.

Algal blooms, including harmful algal blooms (HABs), often occur in the TF region (Seitzinger, 1991; Paerl et al., 2001; Bukaveckas et al., 2011). Algal community metrics have been used for a long time in the assessment of water quality conditions as indicators of biotic responses to environmental stressors such as eutrophication and acidification (Buchanan et al., 2005; Marshall et al., 2006). The United States Environmental Protection Agency (EPA) and Chesapeake Bay Program have developed specific Chlorophyll a (Chl-a) concentration criteria for the TF regions in the Chesapeake Bay (USEPA, 2010) and use numerical models to assess the impact of algal bloom on water quality (Cerco and Noel, 2004). As the algal growth is highly controlled by the nutrient inputs from the non-point sources, the Chl-a concentration criteria are often used to evaluate the efficiency of nutrient reduction in the drainage basins and estuary restoration.

Complex water quality models have been widely used to understand algal blooms in response to the flow and nutrient discharges and to determine nutrient loading reduction (Thomann and Mueller, 1987; Cerco and Noel, 2004; Shen, 2006). There are many applications for using 2D and 3D water quality models to simulate algae blooms (Wu and Xu, 2011; James, 2016; Kim et al., 2017; Jiang and Xia, 2017). However, it is always a challenge to calibrate a complex model to properly simulate an algal bloom because both physical transport and biological processes modulating algal biomass dynamics are highly variable under different residence times and biological timescales (Lucas et al., 2009; Qin and Shen, 2017). To simulate hydrodynamics well, a spatially fine resolution of the model grid is often required due to the complex geometry in TF portions of estuaries (e.g., Shen et al., 2016). Moreover, the accuracy of model simulation depends highly on model kinetic processes, while large uncertainties are always associated with the selection of model kinetic parameter values due to high correlations among these parameters and non-uniqueness of the parameter values (van Straten, 1983; Shen, 2006; Jiang et al., 2018). On the other hand, the accuracy of eutrophication models depends highly on the nutrient loadings from both point and nonpoint sources, which are often simulated by the watershed model (Shen et al., 2005; Riverson et al., 2013). As large uncertainties are associated with the watershed model as well, the linked watershed-receiving water quality modeling approach is often associated with a high level of uncertainty (Wu et al., 2006), which increases the difficulty for accurate simulation of an algal bloom.

In addition to the use of complex numerical models, statistical and empirical modeling approaches based on observational data have been applied to simulate algae, dissolved oxygen concentration (DO), and other pollutants in aquatic systems (Recknagel et al., 1997; Shen et al., 2008; Zhang et al., 2009; Rounds, 2002; Shen and Zhao, 2010; Xie et al., 2012; Kong et al., 2017). The empirical approaches have the advantage of providing a relationship between independent and dependent variables and estimating dependent variables according to the changes of a set of independent variables. Empirically based algorithms have been playing an increasingly important role in HAB modeling, providing an important link between conceptual and dynamical modeling approaches (McGillicuddy, 2010; Blauw et al., 2010; Anderson et al., 2010; Wang and Tang, 2010; Kong et al., 2017). As more and more observational data become available, many effective methods can be used to build empirical models, such as multiple variable regression and neural network. Recently, more sophisticated methods based on the field of machine learning have also been applied for water quality modeling in multiple ways (e.g. Recknagel, 2001; Muttil and Chau, 2006; Volf et al., 2011; Liang et al., 2015; Kong et al., 2017). Lui et al. (2007) used a vector autoregressive model to simulate algal blooms. Ribeiro and Torgo (2008) compared different methods for algal simulations, which showed that the support vector machine (SVM) has a good modeling skill. Xie et al. (2012) demonstrated the effective use of SVM for simulating freshwater algal bloom in a reservoir. Crisci et al. (2012) reviewed supervised machine learning used for ecological data. Recently, Park et al. (2015) developed an early warning protocol of algae using SVM in freshwater and reservoirs. Moe et al. (2016) applied Bayesian network technology to study cyanobacteria bloom in lakes. Kong et al. (2017) applied SVM to evaluate eutrophication statuses in coastal seas successfully. These studies indicate that the machine-learning approach is an effective tool for algal bloom simulations. However, most temporal variations of algal bloom simulations in the literature are limited to the simulation of algae at individual observation stations. When applying a statistical model or a machine-learning model to an entire region of an estuary with multiple observation stations, different models need to be created for the prediction at different stations, which could cause inconsistency in response to change of environmental conditions at different stations. Although the machine-learning has been applied to predict algal blooms, it is not well-studied if the model also is capable of responding to the change of watershed condition due to change of nutrient loadings as many model simulations are trained using input variables observed at the same station to be predicted (Xie et al., 2012; Park et al., 2015).

In this study, we investigated a data-driven modeling approach to simulate algal blooms in the James River. The James River estuary is a western tributary of the Chesapeake Bay. In the TF region of the James River, a bloom of cyanobacteria, a freshwater HAB species, often occurs in summer, and microcystin is often observed when the Chl-a concentration is high (Bukaveckas et al., 2018). The Chl-a distribution is strongly influenced by hydrodynamic conditions when the geometry changes from a narrow stream to a wider cross-section because of the limited mobility of algae. Bukaveckas et al. (2011) found that the location of the maximum of Chl-a concentration in the TF James River is determined in part by the natural geomorphic features of the channel. The transition from a riverine-type (narrow and deep) cross-sectional morphology to a broad channel with shallow lateral areas provides favorable light conditions for the algal growth. The residence time increases during the low-flow period, which coincides with the summer period of algal bloom (Shen et al., 2016; Qin and Shen, 2017). Consequently, the algal bloom occurs frequently during summer in this region. This is an ideal area for investigating the data-driven modeling approach.

The purpose of this study is to apply the learning machine technique to the TF portion of the James River to simulate the algal blooms using long-term monitoring data at multiple stations to provide new capability for harmful algal bloom prediction. The difference of the current modeling approach is that we applied Empirical Orthogonal Function (EOF) to separate Chl-a concentrations at multiple stations in the TF region to both spatial and temporal components. We simulated the principle temporal vectors based on the variation of environmental variables. Therefore, the entire TF region can be simulated by one model. We also used nutrient loadings and flow data at Full-Line as dependent variables in order to ensure the model to response to change of environmental variables. In addition, we conducted a transformation of variables and introduced combined new variables to improve the model prediction skill and ensure that the model would respond to the changes of nutrients discharged from the watershed. As a result, the model shows an improved predictive skill. The model sensitivity tests indicate that the model is suitable for investigating the responses in algal growth to changes in nutrient loadings when the model is trained appropriately. The approach is efficient in terms of model effort and can be applied to other estuaries.

2. Methods

2.1 Data collection

The tidal freshwater (TF) segment of the James River (salinity < 0.5 ppt) extends from the Fall Line (at Richmond, VA) to the downstream with a total length of approximately 115 km. The drainage area is 26,165 km^2, which is predominantly forested (about 71%) (Bukaveckas et al., 2017). The Virginia Department of Environmental Quality conducts monthly monitoring surveys in the James River. Observations of Chl-a concentration near the surface, together with temperature, DO, total suspended solid (TSS), total nitrogen (TN) and total phosphorus (TP), dissolved inorganic nitrogen (DIN), and phosphate (PO_4), are available from 1990-2013. The station locations are shown in Fig. 1. The stations located in the TF region include TF5-2, TF5-

2A, TF5-3, TF5-4, TF5-5, TF5-5A, and TF5-6 from upstream to downstream, and Station TF5-2 is located at the Fall Line near Richmond. The distribution of the Chl-a concentration along the James River mainstem is shown in Fig. 2. High Chl-a concentration is observed in the region with relatively wide cross-sections and the concentration decreases in both the upstream and downstream directions. The daily freshwater discharge is available at USGS freshwater gauge station (US02037500) near Richmond. The water quality station near the Fall Line is mainly controlled by the freshwater and nonpoint source discharges of nutrients and TSS. There is a statistically significant relationship between Chl-a concentration and discharge (Bukaveckas et al., 2017). Loadings of TN, TP, and TSS were estimated by multiplying daily flow and interpolated daily TN, TP, and TSS concentrations at the Fall Line. Hourly solar radiation was obtained from the Richmond International Airport. In this study, the data set of Chl-a concentration for all stations excluding TF5-2 from 1990 to 2013 was used, leading to the data size matrix of 274×6 in total. We excluded Station TF5-2 because it is located at the Fall Line and the measurements were used to estimate loadings for algae, TN, TP, and TSS for the nonpoint sources.

2.2 Variable transformation

Observations of Chl-a data and environmental variables used for the model were transferred first to improve the accuracy of the model prediction. The logarithmic transformation was applied to Chl-a data. Initial analysis showed that transforming discharge Q to $Q^{1/3}$ has a high correlation between flow and Chl-a concentration. A 5-day backward moving average flow prior to the date of observation was applied to the flow to account for the accumulating effect as the USGS flow station is located upstream of the study area, which provides better correlation between flow and log (Chl-a). An example of correlations of independent variables and log (Chl-a) at Station TF5-5 is listed in Table 1.

Both the observations of TN and TP at Fall Line were linearly interpolated and multiplied by flow to obtain daily loadings. Both TN and TP loadings are high in the spring and low in the summer, which are negatively correlated to high concentrations of Chl-a (Table 1). If we directly use it for the model, the Chl-a concentration will not respond correctly to the nutrient level. On the other hand, the summer high algal bloom depends not only on the total spring runoff of TN and TP, but also on the summer bottom fluxes of DIN and DIP from the bottom sediment due to the later winter and spring (February-May) deposition of organics and subsequent remineralization, which will be on the order of 100 days (Thomann and Mueller, 1987). To better reflect the real signal of TN and TP in summer when an algal bloom occurs, we first backward-average the loading (moving average) for a 120-day period prior to the date of Chl-a observation to obtain the accumulative effect of spring runoff. The 120-day moving average was determined based on the spring runoff period, time required for remineralization (Park et al., 1995), and the model performance. Both TN and TP daily loading were transferred to the new variables as follows:

$$TN_{new} = \frac{TN}{H_{TN}+TN}\theta^{T-20}$$

$$TP_{new} = \frac{TP}{H_{TP}+TP}\theta^{T-20} \qquad (1)$$

The approach is similar to the Monod-type nutrient limiting function applied in the water quality model (Thomann and Mueller, 1987; Park et al., 1995). By doing this transformation, the signal for high nutrients during spring was reduced. The correction of temperature, θ^{T-20}, was used to amplify the release of recycled nutrients in summer (θ = 1.03). Note that it is not a good approach to use temperature directly as an independent variable as it has the same annual cycle as algal blooms, which will be discussed more in the Discussion Section. We used the 75th percentile of loading values as the half-saturation coefficients for both TN and TP based on model test runs. With these changes, the respective correlations of Chl-a and TN and TP were improved (Table 1). In addition, as Chl-a concentration values were obtained on different dates for each month, a 15-day average of light was used for the model. Although the moving average of light did not show an improvement of the correlation, it did improve the model simulations. A detailed description of environmental variables used for model input and transforming are listed in Table 2.

2.3 Empirical orthogonal function analysis

There are many observation stations located in the estuary. A traditional approach for developing an empirical model is to develop a model for each observation station, which is not efficient and may not be consistent with changes of environmental conditions for each station. We applied the EOF method to reduce the data dimension and to be able to simulate the entire system based on the principle components of Chl-a data. The EOF method has often been applied to analyze complex data sets to understand the spatial and temporal patterns and distributions of state variables (Bergamino et al., 2007; Wang and Tang, 2010; Du et al., 2018). The purpose of using the EOF method in this study is to separate spatial variations and temporal variations of Chl-a based on principal components. Therefore, we can focus on the prediction of a few temporal vectors for all stations rather than develop a model for each station. The EOF analysis is based on the singular value decomposition method, which decomposes the data matrix F (log(Chl-a)) into the form:

$$F = SVD \qquad (2)$$

where S is the temporal vector of the matrix (274×6), D is an orthonormal matrix (6×6) of spatial vectors, and V is a diagonal matrix (6×6) storing the eigenvalues. Once we obtain the temporal variations for the principal components, the spatial variations at each station can be obtained based on Eq. (2).

2.4 Support vector machine LS-SVM

We used support vector machines (SVM) (Vapnik, 1999) for this study. SVM is a powerful learning machine for classification, and it can be applied to time-varying simulations. SVM has been first introduced within the context of statistical learning theory and structural risk minimization. The idea of SVM is to map the training data nonlinearly into a higher-dimensional feature space and then to construct a separating hyperplane with maximum margin there. LS-SVM, proposed by Suykens and Vandewalle (1999) and Suykens et al. (2002) is an extended version of the standard SVM. Different from the standard SVM, LS-SVM takes a squared loss function for the error variable and uses equality constraints instead of inequality constraints.

LS-SVM has been widely applied in fields of pattern recognition, classification and function estimation (Zhang et al., 2011). Recently, it was also combined with a water quality model to estimate model kinetic parameters (Liang et al., 2015; Park et al., 2015; Kong et al., 2017). Park et al. (2015) applied it successfully for predicting the eutrophication status in a coastal water.

The method is to estimate a function $f: R^N \rightarrow \{\pm 1\}$ using training data of N-dimension patterns x_i and class labels y_i, $(x_1, y_1), \ldots (x_l, y_l) \in R^N \times \{\pm 1\}$. Data can be mapped into the higher dimensional space via a nonlinear function φ (x) and:

$$y(x) = w^T \varphi(x) + b \quad (3)$$

where $w \in R^N$ and $b \in R$ are regression parameters to be determined. The following optimization is formed:

$$\min J(w,e) = \frac{1}{2} w^t w + \frac{1}{2}\gamma \sum_{k=1}^{N} e_k^2 \quad (4)$$

Subject to:

$$y(x) = w^T \varphi(x) + b + e_k, \qquad k = 1,2,\dots,N$$

The problem can be solved using non-linear optimization (Lagrangian method), and the LS-SVM model can be expressed as:

$$y(x) = \sum_{k=1}^{N} \alpha_k K(x_k\,, x) + b \quad (5)$$

where $\alpha = [a_1 a_2, \dots, a_N\ ]^T$ are the Lagrangian multipliers, and $K(x_k,x_l) = \varphi(x)'\varphi(x)$ is the kernel function. The linear SVM kernel is $K(x_k, x) = (x_k^T x + 1)^d$ and the RBF kernel is $K(x_k, x_l) = \exp\{-||x - x_k||_2^2/\sigma^2\}$, where σ is kernel parameters. Different kernels were tested and the RBF kernel was used for this study, which provides satisfactory performance.

We used flow, TSS, TN, TP, and Chl-a loadings at the Fall Line together with light and temperature as independent variables for the model. We first used the LS-SVM learning machine to conduct training for six temporal mode of eigenvectors obtained from Eq. 2 using the same independent variables. Although the 1st eigenvector has the highest contribution, the contribution of this vector to Chl-a concentration at each station is different. Therefore, LS-SVM was trained for each eigenvector. We used data from 1992-2005 for the model training because the algal concentration is much higher during the period from 1990-2002 and it decreases after 2002, the selection of data for training spanning both of these two periods was important. We compared the results using either date set of the first 14-year (1990-2003) or the data set of the last 14-year

(2000-2013) for model training to that of using 1990-2002 data set for training, the model has the best predictive skill using 1992-2005 data set. Adding more data for the training did not improve model performance much, which may cause over-fitting of the model.

Once the model was trained, the data from 1990, 1991, and 2006-2013 were used for verification. After having completed training and verification processes, the Chl-a concentration can be computed by combining three principal temporal and spatial eigenvectors at each station as follows:

$$\ln(Chl\ a(x_{t,i})) = \sum_{k=1}^{3} S(t,k)V(k,k)D(k,i), \quad i = 1,\ldots, 6 \qquad (6)$$

We also compared model predictions and observations of Chl-a at each station as verification. The sensitivity tests were also conducted to evaluate the model response to change of riverine loading. All components including data transformation, EOF analysis, and LS-SVM simulation were implemented in the Matlab. A detailed flow-chart of the procedure for machine learning is shown in Fig. 3.

3. Results

3.1 EOF analysis

The EOF results are listed in Table 3. The first 4 eigenvalues have a total contribution of 91%. Fig. 4 shows the spatial pattern of these stations based on the 1st and 2nd dominant modes. It can been seen that Stations TF5-2A and TF5-3 are close to each other in the lower right corner, while Stations TF5-4, TF5-5, and TF5-5A concentrate in the upper left corner, and Station TF5-6 is between these two groups. The upper tidal freshwater region, where Stations TF5-2A and TF5-3 are located, has the negative spatial value of the second mode, indicating that the change of eigenvector value is in the opposite direction as the downstream TF region. The correlations among parameters are similar to the distance between different stations with respect to the 1st mode (Fig. 4). High correlations exist between stations close to each other. The pattern of the distribution appears to be determined by similarities in geomorphology within and between TF segments. Stations TF5-2A and TF5-3 are located in a narrow upper TF (Fig. 1), where water moves fast and residence time is less than 5 days under the mean flow condition (Shen and Lin, 2006) and less algae can accumulate. Stations TF5-4, TF5-5, and TF5-5A, in contrast, are located in a wide segment with a relatively long residence time, which can create a favorable condition for algae to grow (Bukaveckas et al., 2011). Station TF5-6 is located downstream where the channel becomes narrow again and it can be influenced by nutrient loadings from the Fall Line and upstream transport of nutrients from the saline-water region due to estuarine circulation.

3.2 Simulation using LS-SVM

The model results for training and verification to fit eigenvectors for the first four modes are shown in Fig. 5. It can be seen that the model has the best skill for the first three modes with r^2 = 0.77, 0.60, and 0.36 ($p < 0.0001$), respectively. The performance decreases and varies for different modes. There is no predictive skill for the last three modes as they contribute minor contributions (Table 3) and are distributed randomly.

Using the first three modes of the EOF prediction, the Chl-a concentration can be computed using Eq. 6. The prediction results are shown in Fig. 6. The correlation (r^2) and the root-mean-square error (RMSR), mean error (ME) $(\sum(M-O)/n)$, absolute error (AE) $(\sum|M-O|/n)$, and model skill $SS = 1 - \frac{\sum(M-O)^2}{\sum(O-\bar{O})^2}$ are used to quantify the model performance, where M is model output, O is observations, and n is the number of observations. These statistics are commonly used for model skill assessment (Cerco and Noel, 1993; Allen et al., 2007;

Maréchal, 2004; Willmott, 1981). Statistical results for data used for training and prediction are listed in Table 4. It can be seen that the model prediction skill at each station is different. The skill for the model prediction is lower than the skill for the model training period. Performance levels are often categorized by SS as: > 0.65 excellent; 0.65–0.5 very good; 0.5–0.2 good; < 0.2 poor (e.g., Maréchal, 2004; Allen et al., 2007). The very good predictions are found at Stations TF5-4, TF5-5, and TF5-5a ($r^2 > 0.56$, SS > 0.5). The prediction skill decreases at Stations TF5-2A, TF5-3 and TF5-6. The worst station is TF5-2A in term of SS ($r^2 = 0.53$ and SS = 0.11) though the correlation coefficient is still high, suggesting that it is difficult to simulate high variations of algal blooms at this station. Both bias and the absolute difference between model training and prediction are on the same order. Based on model skill assessment statistics, overall, the model prediction skill is satisfactory based on the model skill assessment statistical measures (Maréchal, 2004; Allen et al., 2007). Compared to previously published applications of Chl-a simulations, the model skill is lower than that of Xie et al. (2012) based on correlations and mean errors. One of the reasons is that we only used seven environmental variables, while more independent variables at model station were used for training by Xie et al. (2012). The performance is comparable to Park et al. (2015) at most stations. Predictive skills at many stations are comparable to complex water quality models as well (e.g., Wu and Xu, 2011). The model simulation period is from 1990 to 2013, which covers both wet and dry seasons. Qin and Shen (2017) compared the interaction of biological and physical transport processes under different timescales and found that there is a good correlation between algal biomass and residence time under seasonal to annual scales in the TF portion of the James River. The inverse relationship between algal biomass and the flushing effect of physical transport in this area was successfully reflected by the model results, which show that the Chl-a concentration is lower during the high-flow period from 2003-2006 than during the 2000-2003 low-flow period.

The model results show a discrepancy in observations during summer when HABs exist (very high Chl-a concentration). It may be due to some factors (or variables) controlling HABs that are not exclusively included in the current model (e.g. competition of nutrients and light between species), as it is still not well-known why microcystin is often observed when Chl-a concentration is high (Bukaveckas et al., 2018). Chl-a observations are conducted monthly, which may be not insufficient for simulating microcystin. It appears that a high-frequency observation of Chl-a is needed to improve the model skill.

3.3 Response to nutrient reduction

The LS-SVM learning machine maps the training data nonlinearly into a higher-dimensional feature space and constructs a separating hyperplane with a maximum margin there. It then classifies new data based on the distance from the training data and separates these data into different classes. However, though the model prediction skill is satisfactory, the application of the model other than the prediction of Chl-a concentration may be limited as the model depends on training data. For example, it may be questionable if the model will respond to the changed nutrient reduction because the model may not be trained based on the underlying biological processes. However, with effective transformation of nutrient data (e.g. making model sensitive to low nutrients) and sufficient training data, the response of model to nutrients is feasible. To evaluate the reliability of the model application for nutrient reduction, it is useful to examine if the model responses to the changes of nutrients are reasonable. In this study, a model simulation was conducted by simultaneously reducing the loadings of TN and TP by 50%. After TN and TP are reduced from the watershed, the Chl-a concentration at the Fall Line will decrease proportionally by 50% as well. The model results compared with the baseline condition is shown in Fig. 7. The Chl-a concentration decreased correspondingly with reductions of TN and TP loadings. In the upper TF region, the reduction of Chl-a concentration is about 45%, lower than 50%. In the middle to lower TF, the reduction ranges from 36-41%. The reduction is about 36% at the downstream Station TF5-6. This comparison shows that the model response to the loading reduction is reasonable, which varies at different stations. More discussion of the model response to loading reduction will be presented in the Discussion Section.

4. Discussion

4.1 Contribution of EOF mode

The purpose of applying EOF analysis to the entire TF region is to use a single model to simulate algal blooms at different locations in the TF region rather than building a series of models at each station. The approach of applying EOF analysis has the potential to be applied to the entire estuary.

As shown in Table 3, the 1st mode accounts for about 62% of the variance using the matrix of data at the 7 stations. However, the contribution of the 1st mode to the variations in Chl-a concentration at each station is different. Fig. 8 shows examples of model simulations with respect to using different numbers of EOF modes at Stations TF5-3 and TF5-4, respectively. It can be seen that the 2nd and 3rd modes are important to improve the model prediction skill as well as the 1st mode at Station TF5-3, where the r^2 value improved from 0.54 to 0.76 and the RMSE value reduced from 7.24 to 5.75 μg/L. In contrast, the 1st mode has the dominant contribution to predict Chl-a concentration at Station TF5.4, while adding the 2nd and 3rd modes have much smaller contributions. The correlation r^2 value increases from 0.66 to 0.70 and the RMSE value decreases from 14.13 to 13.39 μg/L, suggesting that the contribution of each mode to different stations varies and the LS-SVM learning machine is able to fit Chl-a concentrations at different stations with the use of the same independent variables.

4.2 Variable transformation and model response to load reduction

For this study, we conducted variable transformations for TN and TP concentrations. If we directly use TN and TP loadings for the model, the Chl-a concentration will decrease because both TN and TP loadings are high in spring and low in summer. The model will also not respond to the nutrient reduction correctly as the Chl-a concentration will increase rather than decrease in summer.

Because algal blooms are highly temperature-dependent, including the temperature effect implicitly rather than using temperature itself as an independent variable is also important for the model. If we use temperature as an independent variable directly without transformation (Eq. 1), the model can simulate Chl-a concentration well with the same or improved skill for the model training (Table 5). However, the model response to nutrient reduction will be incorrect (Fig. 9). Compared to Fig. 7, the Chl-a concentrations increase at Station TF5-2A and the maximum reduction is less than 19%. Our approach, instead, is to apply a temperature correction to the nutrients. The approach is similar to the approach for nutrient limitation by using the Monod function for algal growth (Eq. 1), while the temperature modification is to amplify the effects of nutrient limitation and the benthic fluxes of nutrients from the bottom sediment in summer. The model is sensitive to the selection of the half-saturation nutrient value. We used 75th percentile values of TN and TP concentrations and the selection of the values are based on model performance.

4.3 Model limitation

The current model is built based on nonpoint source loadings of TN and TP and not explicitly expressed as DIN and DIP, and we did not include the point source loadings as independent variables as we assume they are close to constant based on the designed discharge flow without much seasonal variations, which the discharge maybe not always be constant. It is expected that the total reduction is lower than the reduction of DIN and DIP loadings, especially from point sources during the summer period. When time-varying point source data become available, especially including time-varying DIN and DIP loadings at downstream of Fall Line, the model response to nutrient loading reduction will be more accurate. It can be seen that the model can simulate interannual variations of algal blooms, but frequently under-estimate high bloom concentrations. As the cause of the HAB does not depends solely on hydrodynamic conditions and nutrients, the competition of nutrients between different algal species can also contribute to the variations. The model has no prediction skill for the last three modes of EOF indicating that some variations can be due to nonlinear and random effects. Occasionally, we can see that the Chl-a concentration increases while the nutrient concentration decreases. This is partially due to the non-linear behavior of algae. For example, as algal growth decreases, the light condition can be improved and nutrient may become available at the downstream in reality, and the Chl-a concentration can increase at some stations if it is light-limited. Therefore, a detailed evaluation is needed when applying the model to realistic simulations. Nevertheless, the model can be used to evaluate the impact of the non-point source of nutrients on algal blooms in the TF area.

5. Conclusions

An approach using long-term observational data and the LS-SVM learning machine for simulating algal bloom in the TF region of the James River estuary was conducted. The simulation period spanned from 1990-2013, which included both wet years and dry years. The

EOF method was introduced to reduce the data dimension that enables us to model the algal bloom in the entire TF region using only one model. The model simulated well seasonal and interannual variations of an algal bloom during the summer low-flow periods and the low Chl-a concentrations during a high-flow years. The model performance has a good modeling skill ($r^2 > 0.5$ and SS > 0.5) for most stations based on statistical measures. The results show that the bloom is highly modulated by the hydrodynamic condition. The model experiments with changes in nutrient loadings indicate that it has a correct response to nutrient loading reduction. Our modeling exercise indicates that an adequate data transformation is needed in order to use LS-SVM to adequately simulate an algal bloom and its response to loading changes.

As only nonpoint source nutrient loadings were included in the model, the algal bloom simulated can be considered as the response to the upstream nutrient loading. The model simulation results can be further improved if DIN, DIP, and additional parameters are included. This study demonstrates that the use of the LS-SVM learning machine is a good approach for simulating algal blooms in the complex environment of the TF portion of the James River with high efficiency, which can be applied to many other estuaries.

6. Acknowledgement

A portion of the funding for this project was provided by Virginia Department of Environmental Quality through the James River Water Quality Model Refinement project. We thanks editor and reviewers for their constructive comments and suggestions. This is Contribution No. 3808 of the Virginia Institute of Marine Science, College of William & Mary.

7. References

Allen, J.I., Somerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high‐resolution coupled hydrodynamic‐ecosystem models, *J. Mar. Syst.*, 64, 3–14.

Anderson, C.R., Sapiano, M., Prasad, M., Long, W., Tango, P.J., Brown, C., Murtugudde, R., 2010. Predicting potentially toxigenic Pseudo-nitzschia blooms in the Chesapeake Bay. *J. Marine Syst.* 83(3-4), 127-140.

Bergamino, N., Loiselle. S.A., C´ozar, A., Arduino M. Dattilo, A.M., Bracchini, L., Claudio Rossi, C. 2007. Examining the dynamics of phytoplankton biomass in Lake Tanganyika using Empirical Orthogonal Functions. *Ecological Modeling*, 204,156-162.

Blauw, A., Los, H., Huisman, J., Peperzak, L., 2010. Nuisance foam events and Phaeocystis globosa blooms in Dutch Coastal waters analyzed with fuzzy logic. *J Marine Syst.* 3(3):115-126.

Buchanan, C., Lacouture, R., Marshall, H.G., Olson, M., Johnson, J.M., 2005. Phytoplankton reference communities for Chesapeake Bay and its tidal tributaries. *Estuaries* 28: 138-159.

Bukaveckas, P.A., Barry, L.E., Beckwith, M.J., David, V., Lederer, B., 2011. Factors determining the location of the chlorophyll maximum and the fate of algal production within the tidal freshwater James River. *Estuarine, Coasts, 34*, 569–582.

Bukaveckas, P., Beck, R., Devore, D., Lee, W.M., 2017. Climatic variability and its role in regulating C, N and P retention in the James River Estuary. *Estuarine, Coastal and Shelf Science*, doi.org/10.1016/j.ecss.2017.10.004.

Bukaveckas, P.A., Franklin, R., Tassone, S., Trache, B, Egerton, T. 2018. Cyanobacteria and cyanotoxins at the river-estuarine transition. *Harmful Algae*, 76, 11–21

Cerco, C., Noel, M.R., 1993. Three-dimensional eutrophiciton model of Chesapeake Bay. *Journal of Environmental Engineering*, 1906-1025.

Cerco, C., Noel, M.R., 2004. Process-based primary production modeling of in Chesapeake Bay. *Marine Ecology Progress Series*, 282:45-58.

Crisci, C., Ghattas, B., Perera, Ghattas, 201. A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modeling* 240,113-122.

Du, J., Shen, J., Park, K., Wang, Y.P., Yu, X., 2018. Worsened physical condition due to climate change contributes to the increasing hypoxia in Chesapeake Bay. *Science of the Total Environment* 630, 707–717.

James, R. T., 2016. Recalibration of the Lake Okeechobee water quality model (LOWQM) to extreme hydro-meteorological events. *Ecological Modeling*, 325,71-83.

Jiang, L., Lia,Y., Zhaoa,X. Tillotsonc,M.R., Wangb, W., Zhang, S., Sarpong, L., Asmaab, Q., Pane,B 2018. Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China, *Ecological Modeling*, 375, 1-2.

Jiang, L., Xia, M. 2017. Wind effects on the spring phytoplankton dynamics in the middle reach of the Chesapeake Bay. *Ecological Modeling*, 363(10),68-80.

Kim, J. Lee, T., Seo, D. 2017. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model. *Ecological Modeling*, 266(24), 27-36.

Kong, X., Sun, Y., Su, R., Shi, X., 2017. Real-time eutrophication status evaluation of coastal waters using support vector machine with grid search algorithm. *Marine Pollution Bulletin* 119, 307-319.

Liang, S. Han, S., Sun, Z., 2015. Parameter optimization method for the water quality dynamic model based on data-driven theory. *Marine Pollution Bulletin* (98), 137-147.

Lucas, L.V., Thompson, J.K., Brown, L.R., 2009. Why are diverse relationship observed between phytoplankton biomass and transport time?, *Limnol. Oceanogr*., 54(1), 381–390.

Lui, G.C.S.,Li, W.L.,Leungb,K.M.Y., Lee, J.H.W., Jayawardenac,A.W. 2007. Modelling algal blooms using vector autoregressive modelwith exogenous variables and long memory filter. *Ecological Modeling*, 200, 130-138.

McGillicuddy Jr, D.J., 2010. Models of harmful algal blooms: conceptual, empirical, and numerical approaches. *J Mar Syst*. 83(3-4): 105–107. doi:10.1016/j.jmarsys.2010.06.008.

Maréchal, D., 2004, A soil-based approach to rainfall-runoff modeling in ungauged catchments for England and Wales, Ph.D. thesis, Cranfield Univ., Cranfield, U. K.

Marshall, H.G., Lacouture, R., Buchanan, C., Johnson, J.M., 2006. Phytoplankton assemblages associated with water quality and salinity regions in Chesapeake Bay, USA. *Estuarine, Coastal and Shelf Science* 69: 10-18.

Moe, S. J., Haande, S., Couture, R. 2016. Climate change, cyanobacteria blooms and ecological status of lakes: A Bayesian network approach. *Ecological Modeling*, 337, 330-347.

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

Paerl, H.W., Fulton III, R.S., Moisander, P.H., 2001. Harmful freshwater algal bloos, with an emphasis on cyanobacteria. *The Scientific World*, 1, 76-113. DOI 10.1100/tsw.2001.16.

Paerl, H.W., 2009. Controlling eutrophication along the freshwater–marine continuum: dual nutrient (N and P) reductions are essential. *Estuaries and Coasts* 32,592-601.

Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of the Total Environmental*, 202, 31-41.

Park, K., Kuo, A.Y., Shen, J., Hamrick, J.M., 1995. A Three-Dimensional Hydrodynamic Eutrophication Model (HEM-3D): Description of Water Quality and Sediment Process Submodels; Special Report in Applied Marine Science and Ocean Engineering No. 327; Virginia Institute of Marine Science: Gloucester Point, VA, USA, p. 102.

Qin, Q., Shen, J., 2017. The contribution of local and transport processes to phytoplankton biomass variability over different timescales in the Upper James River, Virginia. *Estuarine, Coastal and Shelf Science* 196, 123-133.

Recknagel, F. 2001. Applications of machine learning to ecological modelling. *Ecological Modelling* 146, 303–310

Recknagel, F., French, M., Harkonen, P., Yabunaka, K-I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* 96 (1997) 11-28.

Ribeiro, R., Torgo, L., 2008. A comparative study on predicting algae blooms in Douro River, Portugal. *Ecological Modelling* 212, 86–91.

Riverson, J., Coats, R., Costa-Cabral, M., Dettinger, M., Reuter, J., Sahoo, G., Schladow, G., 2013. Modeling the transport of nutrients and sediment loads into Lake Tahoe under projected climatic changes. *Climatic Change*. 116:35-50.Rounds, S.A., 2002. Development of a neural networks model for dissolved oxygen in the Tualatin Rier, Oregon. In Proceedings of the Second Federal Interagency Hydrologic Modeling Conference, Las Vegas, Nevada.

Seitzinger, S.P., 1991. The effect of pH on the release of phosphorus from Potomac estuary sediments: implications for blue-green algal blooms. *Estuary, Coastal and Shelf Science* 33(4), 409-418.

Shen, J., Wang, Y., Sisson, M., 2016. Development of the hydrodynamic model for long-term Simulation of water quality processes of the tidal James River, Virginia. Journal of Marine Science and Engineering. *J. Mar. Sci. Eng.* 2016, *4*(4), 82.

Shen, J., Zhao, Y., 2010. Combined Bayesian statistics and load duration curve method for bacteria nonpoint source loading estimation. *Water Research*, 44, 77-84.

Shen, J., Wang, T., Herman, J., Mason, P., Arnold, G.L., 2008. Hypoxia in a Coastal Embayment of the Chesapeake Bay: A Model Diagnostic Study of Oxygen Dynamics. *Estuaries and Coasts*, 31,652-663.

Shen, J., Lin, J., 2006. Modeling Study of the Influences of tide and stratification on age of water in the tidal James River. *Estuarine*, *Coastal and Shelf Science,* 68 (1-2): 101-112.

Shen, J., 2006. Optimal estimation of parameters for an estuarine eutrophication model. Ecological Modeling 191 (3–4), 521–537.

Shen, J., Parker, A., Riverson, J., 2005. A new approach for a Windows-based watershed modeling system based on a database-supporting architecture, *Environmental Modelling & Software* (20), 1127–1138.
Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. *Neural Process. Lett*. 9 (3), 293–300.

Suykens, J.A., Van Gestel, T., De Brabanter, J., et al., 2002. Least Squares Support Vector Machines. World Scientific (ISBN 981-238-151-1).

Thomann, R.V., Mueller, J., 1987. Principles of Surface Water Quality Modeling and Control. Harper Collins Publishers, New York, 644 pp.

USEPA, 2010. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries 2008 Technical Support for Criteria Assessment Protocols Addendum, EPA 903-R-10-002.

van Straten, G., 1983. Maximum likelihood estimation of parameters and uncertainty in phytoplankton models. In: Beck, M.B., van Straten, G. (Eds.), Uncertainty and Forecasting of Water Quality. Springer–Verlag, pp. 157–171.

Vapnik, V.N., 1999. An overview of statistical learning theory. *Neural Networks, IEEE Trans*. 10 (5), 988–999.

Volf, G., Atanasova, N., Kompare, B., Precali, R., Oani, N., 2011. Descriptive and prediction models of phytoplankton in the northern adriatic. *Ecological Modelling* 222, 2502–2511.

Wang, J., Tang, D., 2010. Winter phytoplankton bloom induced by subsurface upwelling and mixed layer entrainment southwest of Luzon Strait. *Journal of Marine Systems*, 83(3–4), 141-149.

Willmott, C.J., 1981. On the validation of models. *Physical Geography* 2, 184-194.

Wu, J., Zou, R., Yu., S.L., 2006. Uncertainty analysis for coupled watershed and water quality modeling systems. Journal of Water Resources Planning and Management 132 (5), 351–361.
Wilmott, C.J., 1981. On the validation of models. *Physical Geography* 2, 184-194.

Wu, G., Xu, Z. 2011. Prediction of algal blooming using EFDC model: Case study in the Daoxiang Lake. *Ecological Modeling*, 222, 1245-1252.

Xie, Z., Lou, I., Ung, W.K., Mok, K.M., 2012. Freshwater Algal Bloom Prediction by Support Vector Machine in Macau Storage Reservoirs. Hindawi Publishing Corporation. *Mathematical Problems in Engineering*, Volume 2012, Article 397473, 12 pages,doi:10.1155/2012/397473.

Zhang, C., Zhang, T., Yuan, C., 2011. Oil holdup prediction of oil-water two phase flow using thermal method based on multiwavelet transform and least squares support vector machine. *Expert Syst. Appl.* 38 (3), 1602–1610.

Zhang, X., Srinivasan, R., Liew, M.V., 2009. Approximating SWAT Model Using Artificial Neural Network and Support Vector Machine. *Journal of the American Water Resources Association (JAWRA)* 45(2):460-474. DOI: 10.1111/j.1752-1688.2009.00302.

Table 1. Correlation of selected independent variables and Chl-a concentration.

	Flow	TSS	TN	TP	light	$(light)^{1/2}$
Original data	-0.43	-0.26	-0.03	-0.01	0.44	0.42
Transformed	-0.65	-0.26	0.46	0.35	0.44	0.42

Table 2. List of Variables and transformation used for model input

Name	Variable	Transformation	Parameter values
Chlorophyll a (Chl-a)	Observations at each station (state variable)	logarithmic transformation for Chl-a at each station	
Chlorophyll a (Chl-a)	Observation at Full Line	Convert to loading (concentration ×flow×86400) (ug d^{-1})	
Flow (Q)	Daily observation at USGS flow station	Convert to $Q^{1/3}$, backward 5-day running average	
Temperature (T)	Observation at full line	θ^{T-20}	θ = 1.03
Suspended solid	Observation at full line	Convert to loading (concentration ×flow×86400) (g d^{-1})	
Total nitrogen (TN)	Observation at full line	Convert to loading (concentration ×flow×86400) (g d^{-1}), backward 120 moving average, and introduce new independent variable[1] $TN_{new} = \frac{TN}{H_{TN}+TN}\theta^{T-20}$	H_{TN} = 75th percentile of loading
Total phosphorus (TP)	Observation at full line	Convert to loading (concentration ×flow×86400) (g d^{-1}), backward 120 moving average, and introduce new independent variable[1] $TP_{new} = \frac{TP}{H_{TP}+TP}\theta^{T-20}$	H_{TP} = 75th percentile of loading
Solar radiation	Observation at full line	15-day average	

Table 3. Contribution of EOF modes.

mode	1	2	3	4	5	6
Eigenvalue	90.79	20.43	12.40	8.36	8.06	5.72
Contribution	62%	14%	9%	6%	6%	4%
Accumulative contribution	62%	76%	85%	91%	96%	100%

Table 4. A summary of model skill.

	RMSE		r^2		ME		AE		SS	
Station	Train.	Pred.	Train.	Pred.	Train.	Pred.	Train.	Pred.	Train.	Pred.
TF5-2A	3.00	6.79	0.67	0.53	-0.75	-0.79	1.62	2.73	0.76	0.11
TF5-3	4.34	6.93	0.76	0.67	-1.13	-0.59	2.31	3.17	0.69	0.29
TF5-4	13.88	12.87	0.70	0.58	-3.55	-3.77	7.56	8.79	0.58	0.54
TF5-5	14.61	14.21	0.72	0.71	-4.61	-3.66	9.06	9.29	0.62	0.50
TF5-5A	15.47	13.10	0.63	0.56	-4.53	-2.98	9.63	9.75	0.56	0.52
TF5-6	7.92	8.78	0.51	0.29	-2.34	-0.72	4.96	6.28	0.50	0.27

Table 5. A summary of model skill using temperature as an independent variable.

	RMSE		r^2		ME		AE		SS	
Station	Train.	Pred.	Train.	Pred.	Train.	Pred.	Train.	Pred.	Train.	Pred.
TF5-2A	2.22	7.24	0.66	0.45	-0.41	-0.16	1.15	3.21	0.87	-0.01
TF5-3	3.14	7.37	0.78	0.64	-0.68	0.24	1.75	3.58	0.84	0.20
TF5-4	12.79	13.86	0.68	0.50	-2.53	-3.79	6.70	9.67	0.65	0.47
TF5-5	12.65	14.29	0.74	0.65	-3.41	-3.99	7.85	9.54	0.72	0.50
TF5-5A	12.90	14.70	0.67	0.50	-3.38	-3.05	8.38	10.78	0.69	0.40
TF5-6	7.19	9.16	0.55	0.30	-1.90	-0.45	4.34	6.45	0.58	0.20

Figure Captions

Figure 1. Map of the tidal freshwater James River Estuary and the monthly monitoring locations in the mainstem.

Figure 2. The distribution of the Chl-a concentration in the log scale along the James River mainstem.

Figure 3. A flow-chart for simulation procedure.

Figure 4. Spatial pattern of EOF for each observation station.

Figure 5. Comparison of model simulation of temporal vectors for each of the first four EOF modes (data with red circles are used for training).

Figure 6. Comparison of model simulation and observations of Chl-a concentration (Black circles are observations, blue lines are training, and red lines are model predictions. Numbers show root-mean-square-error and r^2 for training data and model prediction inside brackets).

Figure 7. Comparison of model simulation with reduction of TN, TP, and Chl-a loadings by 50% simultaneously to the baseline condition (Black lines are baseline simulation and red lines are simulation with load reduction).

Figure 8. Comparison of contribution of each modes to the accurate prediction of Chl-a concentrations at Stations TF5-3 and TF5.4 (Black lines are observations, red lines are model simulations, and r^2 values are for training).

Figure 9. Comparison of model simulation with reduction of TN, TP, and Chl-a loadings by 50% simultaneously to the baseline condition using temperate as an independent variable (Black lines are baseline simulation and red lines are simulation with load reduction).

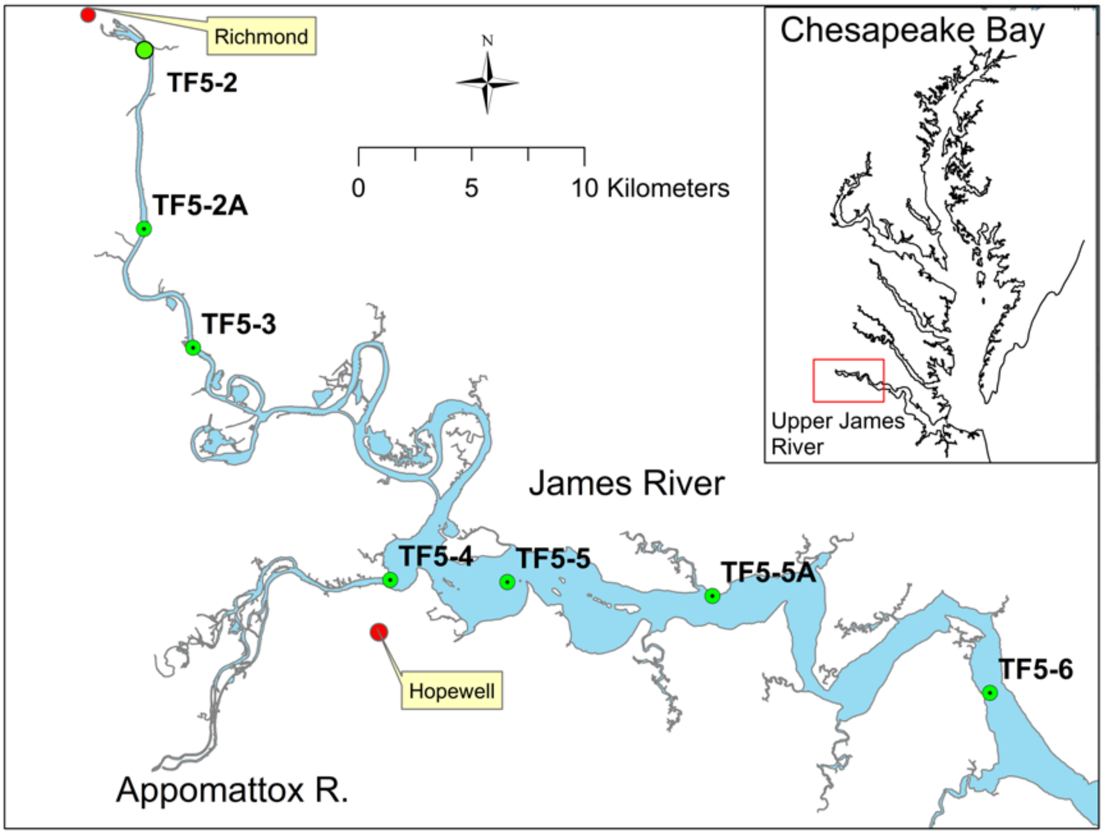


Figure 1. Map of the tidal freshwater James River Estuary and the monthly monitoring locations in the mainstem.

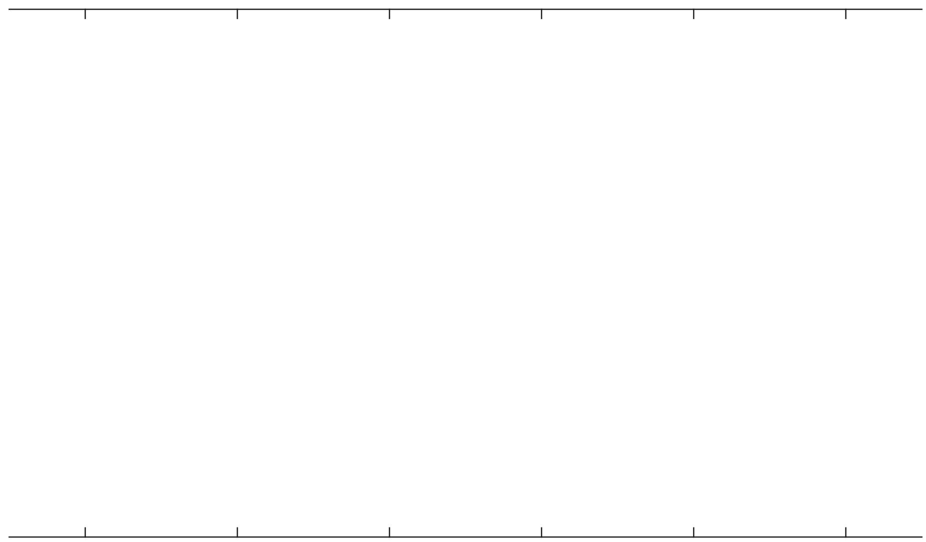


Figure 2. The distribution of the Chl-a concentration in the log scale along the James River mainstem.

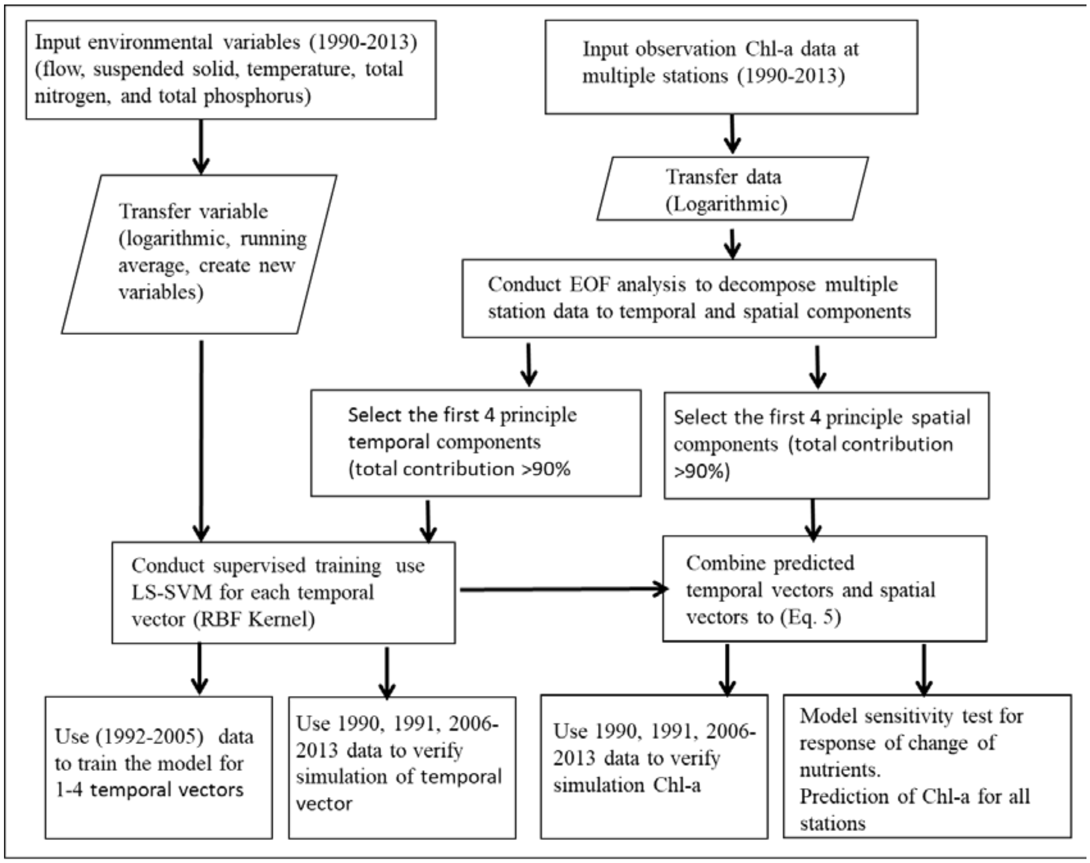


Figure 3. A flow-chart for simulation procedure.

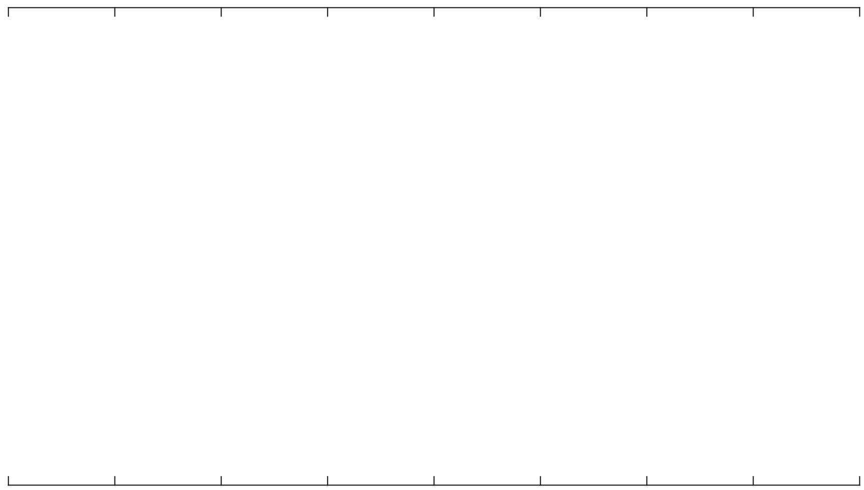


Figure 4. Spatial pattern of EOF for each observation station.

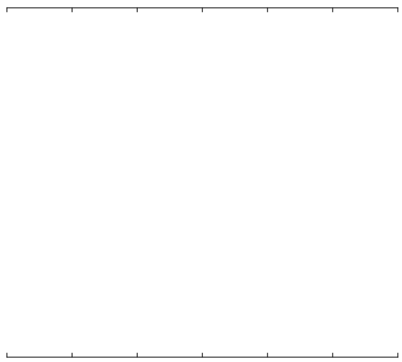


Figure 5. Comparison of model simulation of temporal vectors for each of the first four EOF modes (data with red circles are used for training).

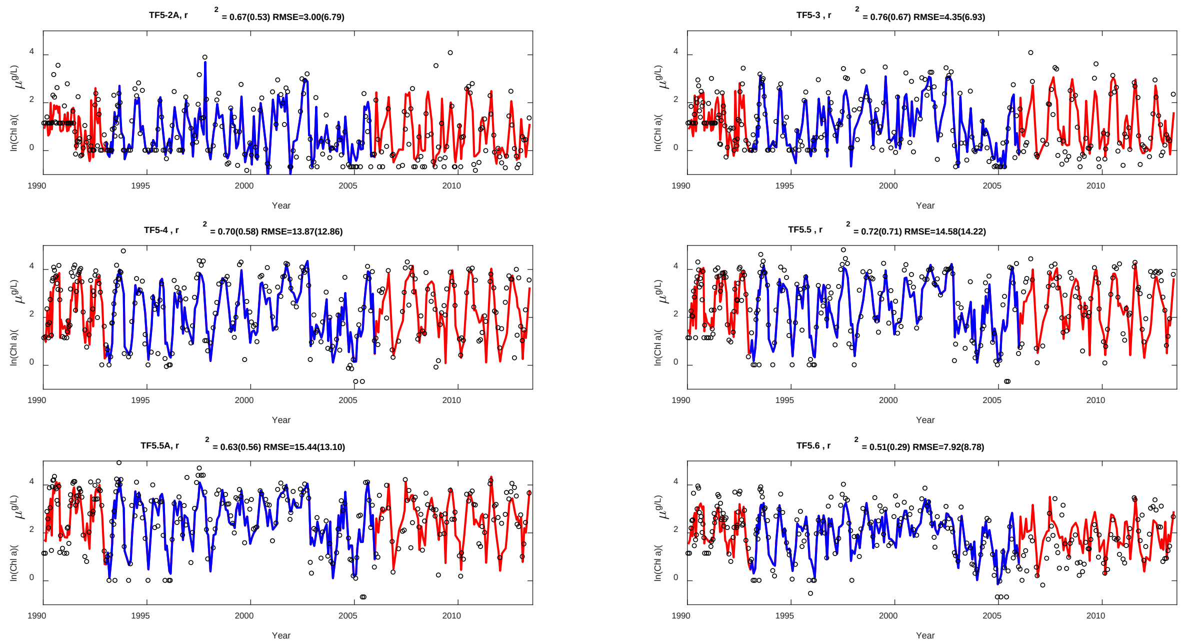


Figure 6. Comparison of model simulation and observations of Chl-a concentration (Black circles are observations, blue lines are training, and red lines are model predictions. Numbers show root-mean-square-error and r^2 is for training data and model prediction inside brackets).

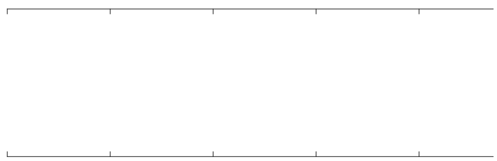
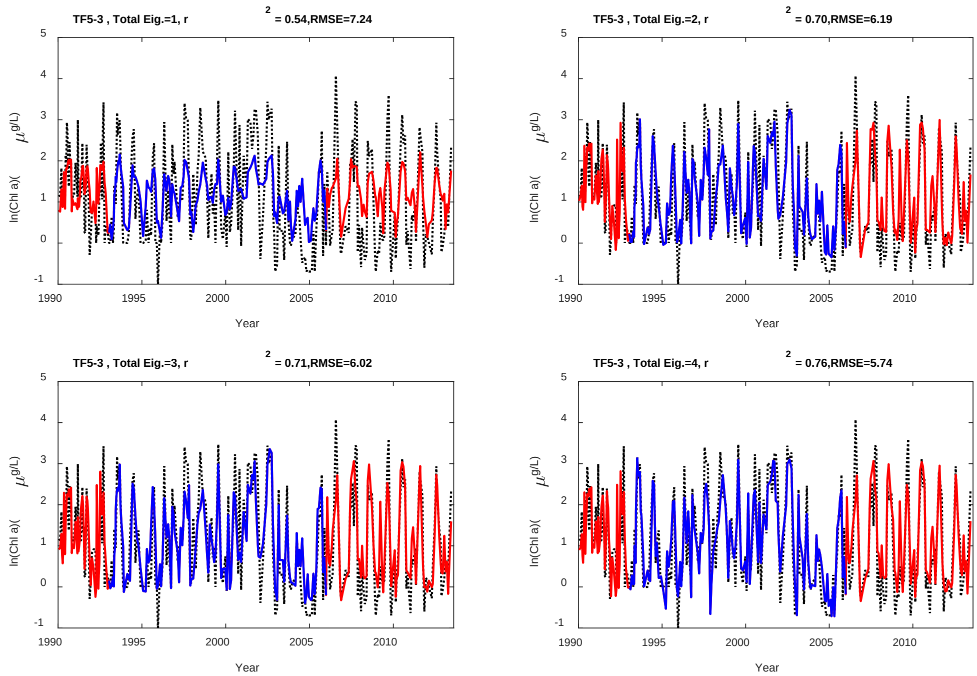


Figure 7. Comparison of model simulation with reduction of TN, TP, and Chl-a loadings by 50% simultaneously to the baseline condition (Black lines are baseline simulation and red lines are simulation with load reduction).



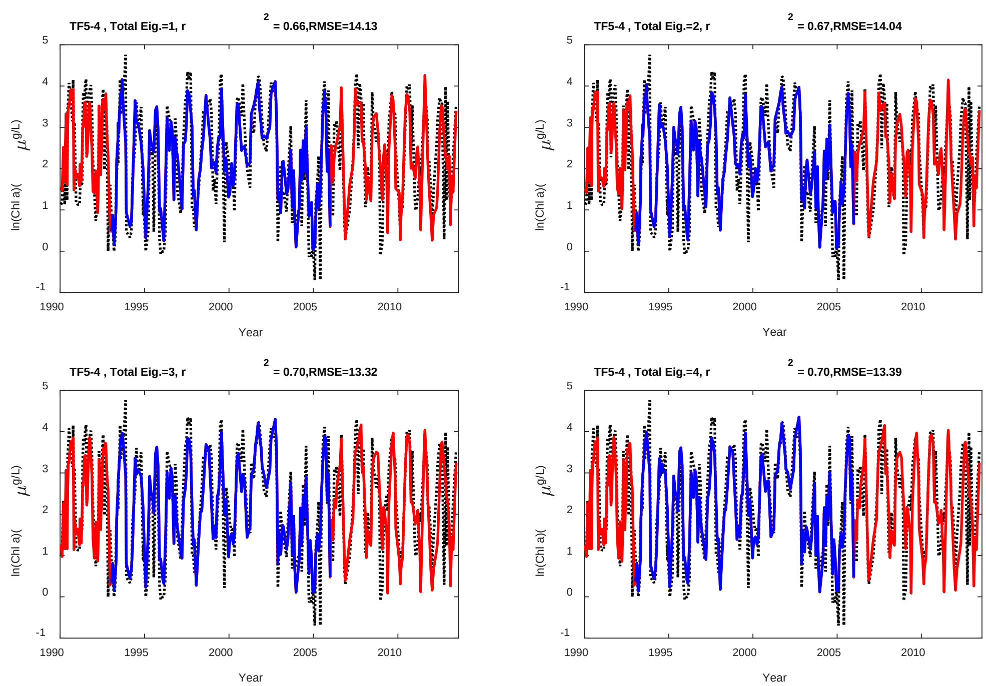


Figure 8. Comparison of contribution of each model to the accurate prediction of Chl-a concentrations at Stations TF5-3 and TF5-4 (Black lines are observations and red lines, model simulations, and r^2 values are for training).

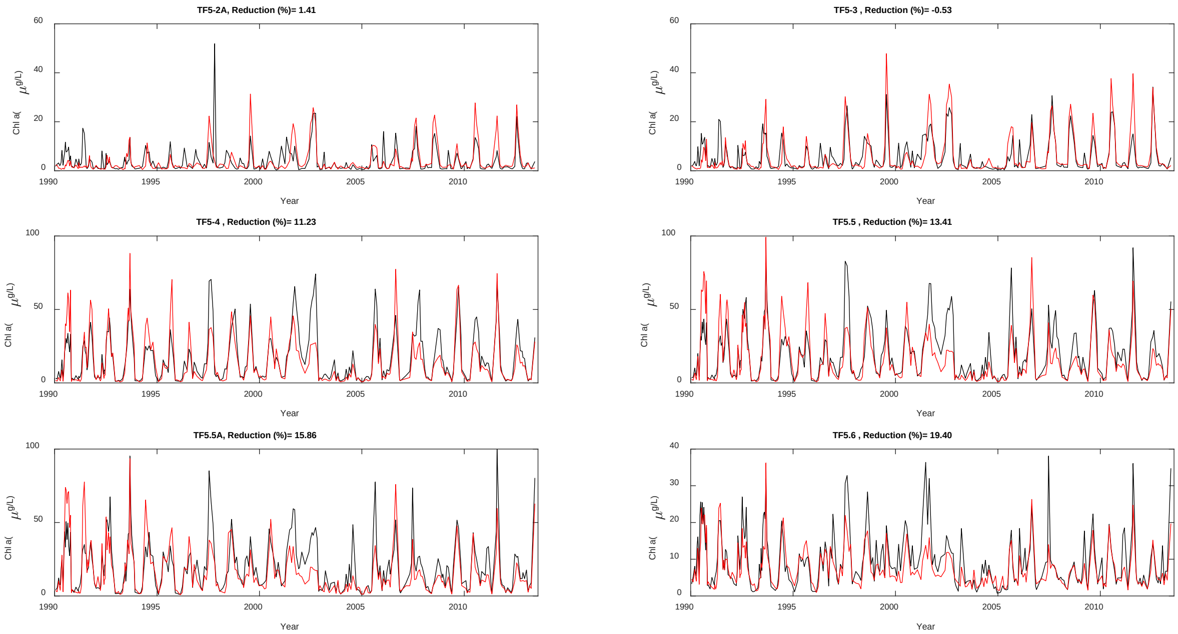


Figure 9. Comparison of model simulation with reduction of TN, TP, and Chl-a loadings by 50% simultaneously to the baseline condition using temperate as an independent variable (Black lines are baseline simulation and red lines are simulation with load reduction).