2022

Data-Driven Approaches For Water Quality Modeling In Coastal Systems

Xin Yu
William & Mary - Virginia Institute of Marine Science, yuxin.uu@gmail.com

Data-driven approaches for water quality modeling in coastal systems

A Dissertation

Presented to

The Faculty of the School of Marine Science

The College of William & Mary

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

by

Xin Yu

January 2022

APPROVAL PAGE

This dissertation is submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

Xin Yu

Approved by the Committee, December 2021

Jian Shen, Ph.D.
Committee Chair / Advisor

Donna M. Bilkovic, Ph.D.

William G. Reay, Ph.D.

Harry V. Wang, Ph.D.

Kyeong Park, Ph.D.
Texas A&M University at Galveston
Galveston, Texas

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am grateful for many people who have supported me along the way in the pursuit of my PhD. First and foremost, I would like to thank my advisor Dr. Jian Shen for the guidance and wisdom he has provided to me during my time at VIMS. I also thank my committee members Donna Bilkovic, Kyeong Park, Willy Reay, and Harry Wang for their continuous support and valuable suggestions.

Thanks to the VIMS community and all my VIMS friends. I will always remember the great times we have had together in VIMS.

Last, but not least, I am especially grateful for the love and support of my family. To my mother and father, for their understanding and positive encouragement. To my husband, for his help and patience in every hurdle I have encountered. To my little daughter, for the endless joys she has brought to me.

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Water quality in coastal waters is of great socio-economic concern. Human activities along the coast have led to an increasing number of impaired waterbodies and degraded ecosystems. To manage water quality issues, accurately modeling coastal water quality is vitally important. One traditional way to model water quality is using numerical models. Despite great advances in hydrodynamic modeling over the past few decades, water quality simulation is still challenging as the performance of water quality model depends on how well the complex biogeochemical processes are parameterized.

While numerical models are the dominant tool for water quality modeling, there are increasing efforts in developing data-driven models in marine sciences. Several major challenges associated with data-driven models for coastal water quality are addressed in this dissertation. These challenges include difficulties in high-dimensional simulation, missing records in observational data, and uncertain watershed loadings.

A data-driven model for coastal water quality is introduced in this dissertation. The proposed model has three major components including (1) forcing transformation auto-selection, (2) empirical orthogonal functions (EOF), and (3) neural network. It uses EOF to extract principal components of the target variable and applies a neural network to simulate the temporal variations of nontrivial components. Different from previous empirical models, the approach is able to simulate three-dimensional variations of water quality variables and it does not use *in situ* measured physical conditions but only external forcings as model inputs.

The robustness of the model is verified with applications to predict temporal-spatial distributions of key water quality variables, including dissolved oxygen (DO) and Chlorophyll-a (Chl-a) concentration in Chesapeake Bay. Using a major portion of historical shipboard monthly measurements and corresponding external forcings for training, the model shows good performance in terms of predicting both seasonal and interannual variations for the testing period.

The model is also tested for high-resolution simulation using Visible Infrared Imaging Radiometer Suite (VIIRS) Chl-a data. The missing records in the satellite data are effectively interpolated by Data Interpolating Empirical Orthogonal Functions (DINEOF). An overall satisfactory model performance demonstrates that by combining DINEOF and machine learning, it is feasible to use data-driven models to predict high-resolution spatiotemporal variations of water quality variables in coastal waters.

Finally, to address the uncertainty in watershed loading, a typically important forcing for coastal water quality, an inverse method is introduced to estimate loading by combining observation and numerical model. In this method, an estuary is divided into multiple segments. Water and material fluxes between neighboring segments are computed from a set of linear equations derived from mass balance and the relationship between residence time and water fluxes. With sparse observational data, inversely estimated loadings agree well with loadings from a previously calibrated watershed model, demonstrating the reliability of the method.

Overall, this dissertation highlights the potential of data-driven model for coastal water quality simulation. With the rapidly accumulated observational data and quick advances in machine learning techniques, data-driven approaches have great potential for water quality modeling and environmental management in the future.

Data-driven approaches for water quality modeling in coastal systems

GENERAL INTRODUCTION

Water quality in coastal waters

Water quality in coastal waters is of great socioeconomic concern to the majority of nations in the world. Over the 198 countries, 154 have part of their territory adjacent to the coastal ocean. About 40% of the population in United States live within 100 km from the coastline as of 2014 (NOAA, 2021). Human society relies greatly on the coastal ocean for food, fuel, recreational activity, marine transportation, trade, and associated economic activities. Overuse of the coastal seas and estuaries has led to an increasing number of coastal waters with impaired water quality and degraded ecosystems. Managing the related water quality issues and restoring the impaired coastal ecosystem requires a lot of resources and efforts. The "health" condition of the coastal ocean is thus a major concern to almost every nation.

Extensive water quality issues emerge in the coastal ocean as a result of both anthropogenic activities and changing natural conditions. Deteriorated water quality conditions in coastal bays and oceans has been reported worldwide due to: increased nutrient load, land-use change, oil spills, wastewater discharge, and release of pollutants. The worsening condition has been reported for major water quality issues include eutrophication (Kemp et al., 2005), coastal hypoxia (Breitburg et al., 2018), ocean acidification (Portner, 2008), harmful algal bloom (Hallegraeff, 2003), bacterial infection (Stewart et al., 2008), and accumulation of microplastic (Wright et al., 2013). For instance, there have been significant changes in duration and intensity of hypoxia in estuaries and coastal waters during the past few decades, raising great concern from

environmental management agencies. Since 1950, more than 500 sites in coastal waters have reported hypoxic conditions, and fewer than 10% of these systems were known to have hypoxia before 1950 (Breitburg et al., 2018). The rapid spread of hypoxia worldwide is generally attributed to the increase of nutrient loading in past decades (Diaz and Rosenburg, 2008). In addition to anthropogenic activities, changing atmospheric conditions under the warming climate also likely worsens several major water quality problems. Recent studies suggest climate change (e.g., warming and changing wind field) and the resultant change in physical conditions (e.g., oxygen solubility and estuarine circulation) also contribute to worsened hypoxic conditions in lakes, estuaries, and coastal waters (Scully, 2010; Carstensen et al. 2014; Du et al., 2018). To understand how water quality responds to anthropogenic activities and changing oceanic-atmospheric conditions, accurately modeling water quality is critically important.

Numerical model and data-driven model

One traditional way to model water quality is using numerical model systems. A numerical model is a computer program that integrates primitive governing equations and empirical relationships to solve transport and mixing processes for a limited number of horizontally and vertically discrete grids in a finite domain. By parameterizing the linear and nonlinear relationships among different biogeochemical components and coupling the biogeochemical module with a hydrodynamic module, numerical model systems (e.g., CE-QUAL-ICM, Cerco, 1995; HEM3D, Park et al., 1995; FVCOM, Chen et al., 2003; ROMS, Shchepetkin and McWilliams, 2005; and SCHISM, Zhang et al., 2016) have been successfully applied in estuaries and coastal oceans for purposes of both process-based scientific research and loading-centered water quality management. When applying

a numerical model to a realistic coastal ocean with complex bathymetry and time-varying external forcings, substantial efforts are needed to first calibrate the model. Commonly, a model is calibrated by comparing the model outputs to observational data. The model validity is thus somehow subject to the availability and representativeness of the observational data at a limited number of monitoring stations or at the surface layer only (e.g., satellite data). When a model is well-calibrated, the numerical model could be useful for diagnostic analysis and water quality management.

Despite great advances in hydrodynamic over the past few decades, water quality modeling is still challenging. The challenges arise from several factors including but not limited to (1) the complexity of biogeochemical processes that regulate water quality variables in the water column and sediment (Fennel et al., 2006), (2) inherent errors associated with model structure and from the parameterization of these processes (van Straten, 1983), (3) the cascade of error in hydrodynamic simulation (Beck, 1987), (4) unavailability and uncertainty of loading data (e.g., for nutrient, sediment, and organic matter) (Boynton et al., 1995), (5) insufficient resolution of horizontal and vertical grids to resolve small scale mixing and advection, and (6) error in interpolated open boundary conditions particularly for boundaries in the open ocean where observational data are limited in both spatial coverage and temporal resolution. Improvements in water quality modeling will depend on the advances in our understanding of the biogeochemical processes, a lot of which are not yet fully explored or cannot be precisely parameterized with general governing equations as the underlying relationships are mostly site-specific.

A numerical model can incorporate “known” relationships, but there are still extensive “unknown” or poorly recognized processes in the ecosystem system. To

include these unknown processes, data-driven models are more suitable. A data-driven model is based on analyzing the data about a system, in particular finding connections between the system state variables (input, internal and output variables) without explicit knowledge of the physical and biological behaviors of the system (Solomatine et al., 2008). One major advantage of data-driven model is no requirement of fully understanding of the underlying dynamics, which makes it favorable for efficient forcasting and environmental management.

While numerical models are the dominant tool for water quality modeling, there are increasing efforts to develop data-driven models in marine sciences, because of rapidly accumulated data and quick advances in data-based methods including artificial intelligence, computational intelligence, data mining, soft computing, and machine learning (see Solomatine et al., 2008 for a review of data-driven models). The data here refers to not only traditional shipboard survey and monitoring data at gauge stations, but also data from remote sensing and reanalysis oceanic-atmospheric models, which provide high resolution and frequency data covering a large area. By assimilating observational data, global or regional reliable reanalysis oceanic-atmospheric models can be a reliable source for external forcings as used in extensive numerical model applications. In addition, rapid advances in machine learning, stimulated by an increasing demand from industry, lead to availibility of readily applicable packages in both commercial programming languages such as Matlab and open-sources languages such as Python and R. The rapid development of the machine learning techniques, along with an increasing number of applications, prompts the growing demand and applications in marine science research and water quality management. Specifically, they have been used for remote

sensing (e.g., Keiner and Yan, 1998; Vilas et al., 2011), water level prediction (Bajo and Umgiesser, 2010; Ren et al., 2020), rainfall-runoff processes (Hsu et al., 1995; Van et al., 2020), marsh classification (Morris et al., 2005), and algal bloom prediction (Muttil and Chau, 2006; Tian et al., 2017).

Challenges in data-driven model for coastal water quality

There are several major challenges when applying data-driven models to simulate coastal water quality. (1) First, previous studies tend to focus on the time series of water quality variable at one single location, which is a one-dimensional (1D) problem, and studies rarely target higher-dimensional problems. The spatial distribution, either horizontally or vertically, is also of great interest since water quality variables typically change in space. For instance, the hypoxic condition in Chesapeake Bay (Hagy et al., 2004) and the Northern Gulf of Mexico (Rabalais et al., 2002) varies interannually and its spatial extent is a key metric to determine the severity of the hypoxic condition. (2) Second, when predicting the water quality condition at a given location, previous studies frequently use *in situ* measurements of other parameters at the same locations as the predictor. For instance, Ross and Stock (2019) used salinity stratification to predict the bottom DO. (3) Third, relevant data may have substantial missing records (e.g., blank area in satellite data due to cloud cover). Malfunction of instruments is inevitable, which will lead to at least some degree of data gaps. (4) Finally, the input forcing may have uncertainty. Taking nitrogen loading as an example, its value is typically given by discharge and nutrient concentration measurements at an upper river monitoring station. Not even considering the inherent error in the measurement, there are still uncertainties imbedded since point sources and diffuse sources from watershed feeding into the

downstream are typically not included. In addition, estimation of atmospheric deposition of nitrogen, which accounts for about 7% of the total nitrogen load (Boynton et al., 1995), is also of great uncertainty.

Objectives of the dissertation

The overall objective of my dissertation is to address and discuss these challenges. I will propose a new data-driven approach to simulate 3D water quality by combining data dimension reduction method and artificial neural network. The model will be tested for different datasets, including long-term bi-monthly shipboard survey data and nearly daily high-resolution satellite data (with substantial gaps) in the Chesapeake Bay. Chesapeake Bay is chosen as an example coastal system to examine the model performance because there is comprehensive and available long-term data for both external forcing and target water quality variables. In addition, a method to estimate watershed loading by combining numerical modeling and *in situ* measurement will be introduced. Specific objectives of the enclosed four chapters are listed as follows.

Chapter 1: A Machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example.

Objectives: To introduce and assess the performance of a data-driven model developed to simulate both one-dimensional bulk index and three-dimensional variations of water quality variables in coastal ocean. This chapter uses the vertical and horizontal distributions of dissolved oxygen in the mainstem of Chesapeake Bay (3D problem) and hypoxic volume (1D problem) as examples.

Chapter 2: A data-driven approach to simulate the spatiotemporal variations of chlorophyll-a in Chesapeake Bay.

Objectives: To determine the feasibility of the data-driven model for a more challenging water quality variable (i.e., chlorophyll-a). Chlorophyll-a is known to vary in a much different way compared to other water quality variables (e.g., dissolved oxygen and salinity) that often change smoothly over time and space. The necessity of forcing transformation is also discussed.

Chapter 3: Chlorophyll-a in Chesapeake Bay based on VIIRS satellite data: spatiotemporal variabilities and prediction with machine learning.

Objectives: To address one key problem in a data-driven model – the missing data, which is especially common in remote sensing data. This chapter will discuss a promising way to interpolate the missing records in VIIRS satellite data. In addition, the feasibility of data-driven model for high-frequency simulation is examined.

Chapter 4: An inverse approach to estimate bacterial loading into an estuary by using field observations and residence time.

Objectives: To introduce a method to estimate the watershed loading using observations and modeled transport timescale. Loading is one major input forcing for forward modeling and yet typically known to have large uncertainties. The robustness of the method will be assessed.

REFERENCES

Bajo, M., & Umgiesser, G. (2010). Storm surge forecast through a combination of dynamic and neural network models. *Ocean Modelling*, *33*(1), 1–9.

Beck, M. B. (1987). Water quality modeling: a review of the analysis of uncertainty. *Water Resources Research*, *23*(8), 1393–1442.

Boynton, W. R., Garber, J. H., Summers, R., & Kemp, W. M. (1995). Inputs, transformations, and transport of nitrogen and phosphorus in Chesapeake Bay and selected tributaries. *Estuaries*, *18*(1), 285–314.

Breitburg, D., Levin, L. A., Oschlies, A., Grégoire, M., Chavez, F. P., Conley, D. J., et al. (2018). Declining oxygen in the global ocean and coastal waters. *Science*, *359*(6371), eaam7240.

Carstensen, J., Andersen, J. H., Gustafsson, B. G., & Conley, D. J. (2014). Deoxygenation of the Baltic Sea during the last century. *Proceedings of the National Academy of Sciences*, *111*(15), 5628–5633.

Cerco, C. F. 1995. Response of Chesapeake Bay to nutrient load reductions. *Journal of Environmental Engineering*, *121*, 549–557.

Chen, C., Liu, H., & Beardsley, R. C. (2003). An unstructured grid, finite-volume, three-dimensional, primitive equations ocean model: application to coastal ocean and estuaries. *Journal of Atmospheric and Oceanic Technology*, *20*(1), 159–186.

Diaz, R. J., & Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *Science*, *321*(5891), 926–929.

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.

Fennel, K., Wilkin, J., Levin, J., Moisan, J., O'Reilly, J., & Haidvogel, D. (2006). Nitrogen cycling in the Middle Atlantic Bight: results from a three-dimensional model and implications for the North Atlantic nitrogen budget. *Global Biogeochemical Cycles*, *20*(3).

Hagy, J. D., Boynton, W. R., Keefe, C. W., & Wood, K. V. (2004). Hypoxia in Chesapeake Bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. *Estuaries*, *27*(4), 634–658.

Hallegraeff, G. M., Anderson, D. M., & Cembella, A. D. (2004). *Manual on harmful marine microalgae*. Paris: UNESCO Publishing.

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*, *31*(10), 2517–2530.

Keiner, L. E., & Yan, X.-H. (1998). A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. *Remote Sensing of Environment*, *66*(2), 153–165.

Kemp, W. M., Boynton, W. R., Adolf, J. E., Boesch, D. F., Boicourt, W. C., Brush, G., et al. (2005). Eutrophication of Chesapeake Bay: historical trends and ecological interactions. *Marine Ecology Progress Series*, *303*, 1–29.

Morris, J. T., Porter, D., Neet, M., Noble, P. A., Schmidt, L., Lapine, L. A., & Jensen, J. R. (2005). Integrating LIDAR elevation data, multi-spectral imagery and neural

network modelling for marsh characterization. *International Journal of Remote Sensing*, *26*(23), 5221–5234.

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

NOAA, 2021. What percentage of the American population lives near the coast? https://oceanservice.noaa.gov/facts/population.html, accessed 10/3/2021

Park, K., Kuo, A., Shen, J., & Hamrick, J. (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, William & Mary.

Pörtner, H.-O. (2008). Ecosystem effects of ocean acidification in times of ocean warming: a physiologist's view. *Marine Ecology Progress Series*, *373*, 203–217.

Rabalais, N. N., Turner, R. E., & Wiseman, W. J. (2002). Gulf of Mexico hypoxia, a.k.a. "The Dead Zone." *Annual Review of Ecology and Systematics*, *33*(1), 235–263.

Ren, T., Liu, X., Niu, J., Lei, X., & Zhang, Z. (2020). Real-time water level prediction of cascaded channels based on multilayer perception and recurrent neural network. *Journal of Hydrology*, *585*, 124783.

Ross, A. C., & Stock, C. A. (2019). An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science*, *221*, 53–65.

Scully, M. E. (2010). The importance of climate variability to wind-driven modulation of hypoxia in Chesapeake Bay. *Journal of Physical Oceanography*, *40*(6), 1435–1440.

Shchepetkin, A. F., & McWilliams, J. C. (2005). The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modelling*, *9*(4), 347–404.

Solomatine, D., See, L. M., & Abrahart, R. J. (2008). Data-driven modelling: concepts, approaches and experiences. In R. J. Abrahart, L. M. See, & D. P. Solomatine (Eds.), *Practical Hydroinformatics* (Vol. 68, pp. 17–30). Berlin, Heidelberg: Springer Berlin Heidelberg.

Stewart, J. R., Gast, R. J., Fujioka, R. S., Solo-Gabriele, H. M., Meschke, J. S., Amaral-Zettler, L. A., et al. (2008). The coastal environment and human health: microbial indicators, pathogens, sentinels and reservoirs. *Environmental Health*, *7*(2), S3.

Tian, W., Liao, Z., & Zhang, J. (2017). An optimization of artificial neural network model for predicting chlorophyll dynamics. *Ecological Modelling*, *364*, 42–52.

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H., & Anh, D. T. (2020). Deep learning convolutional neural network in rainfall–runoff modelling. *Journal of Hydroinformatics*, *22*(3), 541–561.

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, pp. 157-171.

Vilas, L. G., Spyrakos, E., & Palenzuela, J. M. T. (2011). Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sensing of Environment*, *115*(2), 524–535.

Wright, S. L., Thompson, R. C., & Galloway, T. S. (2013). The physical impacts of microplastics on marine organisms: A review. *Environmental Pollution*, *178*, 483–492.

Zhang, Y. J., Ye, F., Stanev, E. V., & Grashorn, S. (2016). Seamless cross-scale modeling with SCHISM. *Ocean Modelling*, *102*, 64–81.

CHAPTER 1. A MACHINE-LEARNING-BASED MODEL FOR WATER QUALITY IN COASTAL WATERS, TAKING DISSOLVED OXYGEN AND HYPOXIA IN CHESAPEAKE BAY AS AN EXAMPLE

Published in *Water Resources Research* (2020, 56, doi: 10.1029/2020WR027227)

Abstract: Hypoxia is a big concern in coastal waters as it affects ecosystem health, fishery yield, and marine water resources. Accurately modeling coastal hypoxia is still very challenging even with the most advanced numerical models. A data-driven model for coastal water quality is proposed in this study and is applied to predict the temporal-spatial variations of dissolved oxygen (DO) and hypoxic condition in Chesapeake Bay, the largest estuary in United States with mean summer hypoxic zone extending about 150 km along its main axis. The proposed model has three major components including empirical orthogonal functions analysis, automatic selection of forcing transformation, and neural network training. It first uses empirical orthogonal functions to extract the principal components, then applies neural network to train models for the temporal variations of principal components, and finally reconstructs the three-dimensional temporal-spatial variations of the DO. Using the first 75% of the 32-year (1985-2016) dataset for training, the model shows good performance for the testing period (the remaining 25% dataset). Selection of forcings for the first mode points to the dominant role of streamflow in controlling interannual variability of bay-wide DO condition. Different from previous empirical models, the approach is able to simulate three-dimensional variations of water quality variables and it does not use *in situ* measured water quality variables but only external forcings as model inputs. Even though the

approach is used for the hypoxia problem in Chesapeake Bay, the methodology is readily applicable to other coastal systems that are systematically monitored.

Keywords: big-data analysis; EOF; neural network; machine-learning; hypoxic volume

1. INTRODUCTION

Hypoxia or low dissolved oxygen (DO) condition is one of the most critical environmental problems in coastal waters. DO concentration less than 2 mg/l is typically considered as the threshold for hypoxia, even though there are studies suggesting that the criterion should be region- and organism-specific (e.g., Vaquer-Sunyer and Duarte, 2008). Hypoxic conditions could cause mortality of aquatic organisms, change biogeochemical cycles, alter the ecosystem community, and reduce fishery yield (Diaz and Rosenburg, 2008). Well-known hypoxic zones include the Chesapeake Bay (Kemp et al., 2005), Northern Gulf of Mexico (Rabalais et al., 2002; Bianchi et al., 2010), Baltic Sea (Conley et al., 2002), and oxygen minimum zones in tropical oceans (Karstensen et al., 2008). There have been significant changes in the duration and intensity of hypoxia in estuaries and coastal waters during the past few decades, raising great concern from environmental management agencies. Since 1950, more than 500 sites in coastal waters have reported hypoxic conditions, and fewer than 10% of these systems were known to have hypoxia before 1950 (Breitburg et al., 2018). The rapid spread of hypoxia worldwide is generally attributed to the increase of nutrient loading in past decades (Diaz and Rosenburg, 2008). Recent studies suggest climate change (e.g., warming and changing wind field) and the resultant change in physical conditions (e.g., oxygen solubility and estuarine circulation) also contribute to the worsened hypoxic conditions in

lakes, estuaries, and coastal waters (Scully, 2010; Carstensen et al., 2014; Wilson et al. 2015; Du et al., 2018; Deng et al., 2018).

Chesapeake Bay, the largest estuary in the United States, was noted to have hypoxic conditions back to the 1930s (Newcombe and Horne, 1938) and has seen an increase of hypoxia over the past century (Hagy et al., 2004) (Fig. 1). Climatological condition at an upper bay station (CB3.3C) based on a 32-year record (1985-2016) shows hypoxia starts in late April and ends in middle September (Fig. 2) and summer hypoxic area covers about 150 km along the bay's main axis (Fig. 1c) . The seasonal hypoxia is generally believed to be caused by the seasonal growth-settling-decay cycle of phytoplankton and the significant seasonality in water column stratification and air temperature (Taft et al., 1980). Stimulated by the winter-spring pulse of freshwater and nutrient input, algae bloom in spring, resulting in a large amount of organic matter settling down to the bottom water in late spring. The subsequent intense DO consumption during the summer, together with stronger stratification and higher temperature, causes the imbalance of DO supply and consumption in the water column, leading to the depletion of bottom DO (Malone et al., 1986; Kemp et al., 2005; Shen et al., 2013).

Hosting one of the most productive ecosystems on Earth, estuaries and coastal waters have received a lot of attention and are thus the hot spots for environmental researches. Extensive studies have been carried out to understand how DO in coastal waters responds to external forcings with various timescales, such as tide, freshwater discharge, wind, and climate variations (e.g., Scully, 2010; Hong et al., 2012; Meier et al., 2012; Fennel and Testa, 2019). Multiple modeling approaches ranging from statistical to fully mechanistic models have been introduced to simulate and predict DO in

Chesapeake Bay, and have led to significant advances in understanding the controlling factors for the hypoxia variations. Murphy et al. (2011) used regression models and showed the early summer hypoxic volumes in Chesapeake Bay was significantly correlated with stratification strength. Using a cross-wavelet analysis, Muller and Muller (2015) built a neural network model to predict future hypoxic volume of the bay and revealed an antiphase relationship between southwesterly winds and hypoxic volume. With a 3D numerical model and assuming a constant DO consumption rate, Scully (2010) identified the important control of lateral circulation induced by wind on the bottom DO replenishment. Using a biogeochemistry model, Da et al. (2018) demonstrated the comparable importance of dissolved inorganic nitrogen inputs from the atmosphere and from the adjacent continental shelf. A recent study by Ross and Stock (2019) used a machine learning technique and found the column stratification as the strongest predictor for bottom DO. However, it is still very challenging to accurately simulate or predict DO in estuarine and coastal waters due to several reasons: (1) hydrodynamics in estuarine and coastal waters are probably among the most complex dynamic processes on Earth as they are affected by not only natural processes but also anthropogenic activities and they are sensitive to external perturbations such as flooding and storm events; (2) long-term and comprehensive dataset are still not readily available for most estuaries and coastal waters even though more field observations have become available recently; a comprehensive dataset for coastal water quality covers information of essential forcings (e.g., wind, air temperature, and nutrient fluxes) imposed on the land-water, air-water, and estuary-ocean interfaces of a given coastal water body; (3) low DO problems are caused by complex

biochemical processes that vary spatially and temporally. Seeking new approaches to advance current modeling of DO is thus of great interest.

With more available observation data and advances in machine-learning techniques, big-data modeling has been applied in a variety of fields. The new technique provides us an opportunity to improve simulation accuracy and further advance our understanding of environmental problems in estuarine and coastal waters. Data-driven or machine-learning-based modeling have already been applied for storm-surge prediction (Bajo and Umgiesser, 2010), rainfall-runoff processes (Hsu et al., 1995; Campolo et al., 1999), water level and flooding prediction (Campolo et al., 1999; Chang and Chen, 2003; Chen et al., 2012), satellite-data retrieval (Krasnopolsky, 2007; Keiner and Yan, 1998; Vilas et al., 2011), marsh classification (Morris et al., 2005), and algal bloom modeling (Recknagel, 2001; Muttil and Chau, 2006; Shen et al., 2019). Neural network application has been applied in Chesapeake Bay as early as 1996 by Scardi (1996) who trained an empirical model for primary phytoplankton production. One of the advantages of the data-driven model is its computational efficiency. It requires much less computational power compared to complex three-dimensional mechanic models. Therefore, the technique will likely provide us an efficient way for predicting water quality in estuarine and coastal waters.

Here, we propose a machine-learning-based data-driven model and examine its performance in simulating the DO condition and hypoxic volume in Chesapeake Bay. Systematical measurement of DO in the bay and continuous monitoring environmental data (e.g., wind, river discharge, air temperature, and nutrient fluxes) since 1984 were publically available. The comprehensive dataset in Chesapeake Bay makes the estuary a

perfect test site for developing data-driven models. We collected monitoring data of DO from the Chesapeake Bay Program (Fig. 1b) and forcing data from a variety of reliable and publically accessible sources to examine the applicability, robustness, and limitations of the data-driven model. The paper is organized as follows. Section 2 describes the method, including the framework of the data-driven model and data collections. Section 3 presents the simulation of hypoxic volume and spatial-temporal variations of DO. The robustness and limitations of the data-driven approach, will be discussed in Section 4, followed by a short summary.

2. METHODS

2. 1 Overall framework of a proposed data-driven model

The proposed data-driven model includes three major components: empirical orthogonal functions (EOF) analysis, automatic selection of forcing transformation (ASFT), and machine-learning (neural network) (Fig. 3). Observed values of the target variable, DO in this study, are first interpolated into a defined vertical grid. The observed target variable, together with forcing variables, will be first split into training and testing datasets at the beginning. We used the first 75% for the training and the remaining 25% for the testing (Fig. 4). In the training dataset of the target variable, the long-term mean is extracted and it will be used later to reconstruct the 3D structure. An EOF analysis is applied to decompose the spatial and temporal components of the target variable in the training dataset. For each major principal component extracted from EOF, a neural network model is trained with input forcings selected from the ASFT module. The ASFT is designed to search for the proper transformation of the input forcings. After the model

is trained, forcings from the remaining 25% testing dataset are transformed based on the transformation function determined by ASFT, and the transformed forcings are put into the trained neural network model to predict the major EOF modes for the testing period. The prediction of the target variable will be obtained as the sum of predicted temporal values multiplying with the corresponding spatial value (also known as the map) and the long-term mean value. The predicted value of target variable, as a function of space and time, will be calculated as

$$C(x,y,z,t) = C_o + \sum_{\text{mod}=1}^{N} Map_{\text{mod}}(x,y,z) \times PC_{\text{mod}}(t) \quad (1)$$

where C_o is the long-term mean extracted from the training dataset, *Map(x,y,z)* and *PC(t)* are the spatial and temporal values for a given mode, respectively. N is the number of principal components that are trained by the neural network model. In the following sections, the three major components of the proposed data-driven model will be explained in more detail.

2.2 EOF to reduce data dimension

By selecting principal modes based on EOF analysis, the data dimension can be efficiently reduced. The EOF analysis in this study is based on the singular value decomposition algorithm, which decomposes the data matrix F into the following form:

$$\mathbf{F}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} \quad (2)$$

where **U** is an orthogonal matrix of temporal vectors, **V** is an orthogonal matrix of spatial eigenvectors (referred to as maps), and **D** is a diagonal matrix of the eigenvalues. **U***__D__ will give the EOF time series. The data matrix **F** is arranged with first dimension

indicating spatial location and second dimension indicates time (i.e., **F**(i,j) with i indicate the location and j indicate the time). In this study, the data matrix **F** has a dimension of 800×288, representing the spatially varying values at 40 monitoring stations (monthly vertical profiles of DO at each station are interpolated vertically into 20 layers) over the 24 years (1985-2008) in the training dataset.

The major variance of the target variable can be well represented by major principal components. For the DO in Chesapeake Bay, our analysis below shows that the first EOF mode accounts for 87% of the total variance and the first five modes account for 93% of the variance. Instead of developing models for each station and each layer (40×20), we choose to fit five regression models for the first five modes to simulate the major variations of the target variable. This strategy will significantly enhance the computational efficiency without losing the major signals.

2.2 Forcing selection and transformation

As an essential part of the data-driven model, data of external forcings were carefully collected. Relevant forcings were selected including nutrient loading, river flow, air temperature, solar radiation, and wind speed and direction. River flow and wind are long known to regulate the stratification, estuarine circulation, and water exchange between ocean and estuaries (Hagy et al., 2004; Scully, 2010; Murphy, et al., 2011), while nutrient loading and solar radiation are generally regarded as the dominant factors controlling the algal growth. The input forcings are almost the same as required by a three-dimensional ecosystem model (e.g., Cerco and Noel, 2013). Forcing data were collected from various sources, including long-term monitoring programs and reliable reanalysis atmospheric model outputs. River flow and nutrient loadings of the large

tributaries, Susquehanna, Potomac, James, and Choptank Rivers, were extracted from USGS (https://www.usgs.gov/). As freshwater discharge from other small tributaries are highly correlated with these major rivers, freshwater and nutrients from the small tributaries are not included. Air temperature at Chesapeake Bay Bridge Tunnel station (CBBT, Station ID: 8638901) was extracted from a NOAA database (https://tidesandcurrents.noaa.gov/), with data gaps filled with measured values from a nearby station Cape Henry (Station ID: 8638999). For the wind data, instead of using a continuous monitoring data at a limited number of NOAA stations, we used the global ERA5 reanalyzed wind produced by European Centre for Medium-Range Weather Forecasts (ECMWF: https://www.ecmwf.int/), which has a full coverage of the entire Chesapeake Bay with a spatial resolution of 0.25 degrees (total of 33 grid points within the Chesapeake Bay are selected; Fig. S1 in the supporting information) and an hourly temporal resolution. Reanalysis wind field from reliable atmospheric models are widely used for 3D hydrodynamic and water quality models (e.g., Testa et al., 2014; Ye et al., 2018; Du et al., 2019). Taking the bay mouth station CBBT for instance, the observed wind is highly consistent with the ECMWF reanalysis wind (Fig. S2).

One feature that makes the model different from previous ones (e.g., Scardi et al., 1999; Shen et al., 2019) is that an auto-selection tool for forcing transformation is developed to find the suitable transformation for the model to account for underlying mechanisms with which water quality state variables respond to external forcings. Forcing transformation is a necessary preprocessing step for a data-driven model. Through forcing transformation, input forcings will be converted with the same temporal and spatial resolution, and, more importantly, some particular effects that are ubiquitous

in estuarine dynamics, such as time-delay effect and accumulative effect, can be included. For example, it is commonly agreed that summer hypoxia in Chesapeake Bay is attributable to January-May nutrient load instead of summer nutrient load (Murphy et al., 2011). The DO's responses to nutrient load are regulated by not only time-lag effect but also accumulative effect, which is physically meaningful as nutrients from upstream take months to reach middle or lower bay (Shen and Wang, 2007). In the proposed model, we include 7 types of time-lag, 13 types of accumulative average, and 8 types of transformation functions including log, exponential, Monod-type filter, and normalization (Table 1). In total, there are about 700 combinations of transformation. Furthermore, not all the input forcings are responsible for the DO variations and it is necessary to filter out the unnecessary forcings. The goal of ASFT module is to select the responsible forcings and search for the appropriate transformations for each selected forcing.

In the ASFT module, multiple linear regression is used to find forcings and transformations that can maximize the performance of the model to explain the target variable. Set the target variable as $\mathbf{Y}(n\times1)$ and the input forcing matrix $\mathbf{X}$ to be empty at the beginning. First, the coefficient of determination (R^2) between $\mathbf{Y}$ and all available forcing variables in all available transformations are computed. The forcing variable with transformation that gives the largest R^2 out of all possible forcings and transformations is selected as the first variable and stored in $\mathbf{X}$ (:, 1). ASFT adds the second forcing variable to $\mathbf{X}$ from the remaining forcings based on the R^2 from multiple linear regressions. Note that during the second round of forcing selection, there will be two forcings in $\mathbf{X}$, with the first one fixed and the second one chosen from all possible transformed value of the remaining forcings. The new variable will be selected only when the new R^2 is larger

than the previous R^2 by at least 0.005. This process will continue until increase of R^2 is less than 0.005. The process is to find a set of independent variables that has largest contribution to the target variable, while avoiding degrading the model performance when including a large number of variables with high covariance among forcings. The multiple linear regression may not be the best method as it only considers linear relationships, and could be further improved in future. For the current study, it works well for our problem. With advances in understanding of the underlying mechanisms, additional transformations can be further introduced.

2.4 Neural network

After selecting the forcings and corresponding transformations, neural network models are trained for the five primary EOF modes (Fig. 4). Artificial neural networks are computational models inspired by the functioning of the human brain (Scardi et al., 1999; Paliwal and Kumar, 2009). They are composed of a number of "neurons", the basic computational unit, which takes inputs (**x**) from other neurons or external sources, calculates the corresponding weight (**w**) for each input, sums the product of weights and input values (Σ**wx**), plus bias (b), and finally passes this value (b+ Σ**wx**) to an activation function (Fig. 4). A number of neurons constitute a hidden neural layer, and a network can have multiple hidden layers. Hyperbolic tangent sigmoid function is used as the activation function in this study, which converts the input value to a value ranging between -1 and 1. To train a neural network is to get the optimal weights and bias so that the cost function (i.e. the model error) approaches its minimum. Here we used the Levenberg-Marquardt backpropagation training function (Marquardt, 1963), which approaches second-order training speed without computing the Hessian matrix directly

and appears to be the fastest method for training moderate-sized feedforward neural network (Hagan and Menhaj, 1999). The Matlab Neural Network Toolbox (version 10.0) was used for this study. The training process will stop when any of the following conditions occurs: (1) number of epochs reach the defined maximum epochs (set to be 100); (2) cost function (mean square error) is minimized to the 0; (3) performance gradient falls below 1e-7. For details of the algorithm of Levenberg-Marquardt method, readers are referred to the help document in Matlab. The algorithm is widely recognized and well implemented in the Matlab software.

As mentioned in Section 2.1, the overall division of the full dataset in the proposed approach is configured as follows: the last 25% of data are reserved for the testing; the remaining 75% are used for the training. The neural network, internally, is also set to randomly divide the input dataset (i.e., the training dataset) into two independent portions, the training and validation portions, which accounts for 80% and 20% of the input data length (i.e., 60% and 15% of the full data records), respectively. By default, the toolbox randomly chooses the "train" and "validate" portions for each training, resulting in slightly different neural network parameters and thus different predictions for the testing period. To address the related uncertainties, we train the neural network model for 100 times for each principal component, use the ensemble mean of these models as the final prediction, and use the standard deviation of these models' predictions to quantify the uncertainties.

We use two hidden layers, with a neuron number of N and round up of N/2 for the first and second hidden layers, respectively, where N is the number of input forcings. The number of input forcings varies for different principal EOF components and is

determined in the forcing selection (see Section 2.3). Sensitivity tests (not shown) regarding the number of hidden layers and the number of neurons do not show significantly different results. It is believed the forcing selection is more important than the hidden layer configuration for water quality problems.

2.5 Model performance evaluation

Besides the common statistical measures including root mean square error (RMSE) and coefficient of determination (R^2), we also calculated model *skill* following Willmott (1981):

$$Skill = 1 - \frac{\sum \left| X_{\mathrm{mod}} - X_{obs} \right|^2}{\sum \left(\left| X_{\mathrm{mod}} - \overline{X_{obs}} \right| + \left| X_{obs} - \overline{X_{obs}} \right| \right)^2} \tag{3}$$

where X_{obs} and X_{mod} are the observed and modeled variables, respectively, with the overbar indicating the time average. *Skill* provides an index of model-data agreement, with a *skill* of 1 indicating perfect agreement and 0 indicating complete disagreement. *Skill* has been widely used to evaluate the performance of numerical models (e.g., Warner et al., 2005). While the R^2 indicates the model's capability of capturing the seasonal trend and interannual variations, and RMSE indicates the overall misfit between model and observation, *skill* can be regarded as a synthesis index to evaluate both the trend capturing and relative misfit.

3. RESULTS

3.1 EOF analysis

The first essential step in the proposed data-driven approach is to reduce the data-dimension. Through the EOF analysis and interpretation with our current knowledge of DO dynamics, we can understand the dominant processes controlling the DO variations. The spatial and temporal characteristics and the possible controlling mechanism of the first three EOF modes will be briefly described in the following paragraphs.

The first EOF mode is the dominant mode, accounting for 87% of the total variance. The map of the first mode is characterized with all positive values, meaning the changes in DO concentration are in phase among all the stations (Fig. 5a-b). The locations of higher values in the lower layer in middle to upper Bay correspond to the low DO region. The relative values in the map indicate the different magnitudes to which the DO concentration changes with time. For example, during the summer months when the temporal value of the first mode is negative, DO of the entire bay decreases from the long-term mean value and the maximum decrease occurs in the region with maximum map value (i.e., deep waters between 38N and 39N). The first mode has significant and obvious seasonality with the lowest temporal value in July, consistent with what is shown in Fig. 2. This is a combined result of seasonal variation in water temperature, stratification, organic matter abundance, and water column respiration. Since these environmental factors share a high covariance with each other, it is not easy to identify which processes dominate the first mode.

It is necessary to point out that the first EOF not only reflects the seasonal characteristics of the DO dynamics but also contains strong signals of interannual variability. A simple linear comparison between July values of the first EOF mode and hypoxic volume calculated by Bever et al. (2013) shows that 62% of the interannual variations of July hypoxic volume during 1985-2008 is explained by the first EOF mode (Fig. 5c).

The map of the second EOF mode is characterized with opposite values between surface and subsurface layer in both middle and upper bay regions, suggesting this mode is controlled by mechanisms that have a different impact on DO between surface and subsurface layers. The opposite values between surface and subsurface is remarkable in the middle and upper bay regions (Fig. 6b). The depth separating surface and subsurface is close to the long-term mean pycnocline depth, roughly at 10 m (Fig. 6b). A seasonal cycle is noticeable in temporal variation with maximum mean value in May and maximum variability in April (Fig. 6c); this pattern is similar to the chlorophyll-a concentration inside the bay (Fig. 6d). Spring algae bloom in Chesapeake Bay occurs as early as the end of winter and the spring bloom usually reaches its maximum intensity around April (Harding, 1994). As a result, the variability of chlorophyll-a concentration also has the largest variability around April (Du and Shen, 2015). Algal bloom has a distinctly different impact on bottom and surface DO, with increasing surface DO due to photosynthesis and decreasing bottom DO due to the decomposition of settled organic matter. Linear regression analysis between the second EOF mode and chlorophyll-a concentration in the previous month shows a strong correlation, with R^2 of 0.35. It suggests a one-month time-lag response of the second EOF mode to the chlorophyll-a

concentration. We will show later that using the data-driven model, the R^2 can be improved significantly to 0.95 for the second EOF mode when multiple forcings are used.

The third EOF mode is also characterized with opposite values between surface and subsurface but in the region of lower to middle bay, with maximum opposition around 37.8N where there is dramatic topographic change (Fig. 7b). What is unique in the third EOF mode is that it has two peaks in its temporal value (Fig. 7c). It is not clear what mechanisms are responsible to cause such unique spatial and temporal pattern. Not all EOF modes can be easily explained with our current knowledge. Sometimes, multiple mechanisms instead of a single one are responsible. In such cases, nonlinear models or machine learning techniques will be more suitable to explain the temporal variations. This is also why we try to use advanced deep-learning methods.

3.2 Simulation of dissolved oxygen

For the DO simulation, we trained the first five primary EOF modes. Little differences are found for the predicted DO in the testing period when the number of modes changes from five to nine. The first five modes account for 93% of the variance, while the first nine modes account for 95%. Including modes with minor contribution is believed unnecessary and may even introduce more noise to confound the major signals. From the training perspective, all of the five modes are well trained, with R^2 larger than 0.90 and *skill* over 0.95 (Fig. 8a-e). It is worthy to note that the uncertainties (indicated by the standard deviation of the 100 neural network models' predictions) are almost negligible for the first mode and increase for later modes.

The performance of the model is evaluated from multiple aspects. First, we compared the bottom DO at each station between modeled and observed values. Compared to surface DO that is merely controlled by the air-sea exchange and water temperature, the bottom DO is affected by many more factors (e.g., stratification, organic matter decay, and sediment oxygen demand) and usually more difficult to simulate. Taking three mainstem stations (CB3.3C, CB5.2, and CB6.1, representing the upper, middle, and lower bay) as an example, the RMSE ranges from 0.85 to 1.64 mg/l (Fig. 9). The peaks of bottom DO varied from year to year and their variabilities are captured by the model. For instance, the peak of bottom DO at CB3.3C in early 2012 was relatively small (less than 10 mg/l), compared to other years (Fig. 9a). In particular, there was a sharp decrease of bottom DO in the fall of 2011, which was believed to be caused by the large freshwater input due to Hurricane Irene and the subsequent Tropical Storm Lee (Ye et al., 2019). The sharp decrease was noticeable at station CB3.3C, which is clearly captured by the model. Additionally, we calculated the model *skill* for the anomaly (i.e., the difference from the seasonal cycle) to examine the model performance in reproducing the deviations from seasonal cycle. The anomaly *skill* can be up to 0.46 (e.g., at station CB6.1); clearly, more efforts are needed to improve the anomaly prediction such as including more modes. The relatively low *skill* in anomaly prediction is a trade-off when including the entire signal of the target variable during the training stage. If one would like to focus on the interannual variability (say July only), one can train a model for that given month only, which in turn may raise another issue concerning the limited length of the training data.

The model performance at all of the 40 stations is statistically illustrated by the Taylor diagram (Fig. 10). The correlation coefficients at the majority of the 40 stations from the proposed data-driven model concentrate in 0.90-0.95. Only one station has a correlation coefficient less than 0.8; this station (ET4.2) is located in Chester River, a small tributary discharging into the upper bay. It is likely that the bottom DO poorly-simulated at this station is more influenced by local processes. Overall, the model performance is comparable to previous deterministic 3D water quality models (e.g., Testa et al., 2014; Irby et al., 2016). For example, using a physical-biogeochemical model, Testa et al. (2014) show the correlation coefficient of 0.80-0.95 for the bottom DO at mainstem stations (see the Taylor diagram Fig. 6 in Testa et al., 2014). Irby et al. (2016) compared nine numerical models in simulating the bottom DO concentration, and, using a similar Taylor diagram (see Fig. 8 in Irby et al., 2016), they show that the correlation coefficient from different models ranges from 0.8 to 0.9. With respect to the variance, the data-driven model predictions are close to the observation, with a mean, minimum, and maximum normalized standard deviation of 0.99, 0.92, and 1.07 over the 40 stations. The root mean square deviation (the third axes in the Taylor diagram) is around 0.3, which is smaller than some numerical model simulations (e.g., Testa et al., 2014). It is worth noting that only model performance of the bottom DO is presented here as it is typically more difficult to simulate accurately compared to the surface DO.

In terms of the spatial variations, the model performance is acceptable. As expected, the marked seasonal variations of the spatial distribution are well predicted for the testing period (Fig. 11), despite the fact that the model seems to slightly underestimate the bottom DO. Furthermore, by comparing the spatial distribution of

summer DO in the testing period, it is obvious that the model captures the overall change of hypoxic area (Fig. 12). From year to year, the summer hypoxic area changed, with a minimum area in 2012 (2 mg/l contour starts from 38N) and a maximum area in 2011 (2 mg/l contour starts from 37.6N). Observation and model predictions are consistent in the distribution of 2 and 4 mg/l contours along the mainstem. It is expected that there is still noticeable bias especially concerning the severely hypoxic area (e.g., area with DO less than 1 mg/L), which is also a problem for previous models, either empirical or numerical ones (e.g., Testa et al., 2014; Cerco and Noel, 2013; Irby et al., 2016). One of the reasons is that there is no signal in DO when DO becomes zero, while forcing variables are still varying, resulting in high uncertainty at these anoxic regions.

It is worthy noting that in the proposed approach, the EOF spatial patterns extracted from the training dataset are assumed unchanged during the testing period. This assumption is valid if the underlying physical and biological processes have not undergone dramatic change, which is true except for some rare cases (e.g., when there is a regime shift) in which the covariance among different stations changes dramatically. In the case of DO in Chesapeake Bay, additional EOF analysis using the DO data in the testing period confirms the validity of the assumption. The spatial patterns for the first three EOF modes in the testing period (Fig. S3-5) are nearly identical to those in the training period (Fig. 5-7) despite the little differences, which can be attributed to the shorter records in the testing period. For practical application, both EOF analysis and training can be conducted dynamically as more data will become available in future to ensure the test period has the same spatial pattern as the training period.

3.3 Simulation of Hypoxic Volume

Similar methodology was applied to simulate the hypoxic volume in the Chesapeake Bay. Here we used the 28-year hypoxic volume time series (1985-2012) calculated by Bever et al. (2013) as the target variable. The hypoxic volume is a scalar index used to quantify overall DO condition in the Bay. Different from DO simulation, only one time series is used as the target variable and there is no need to use EOF analysis. The forcing selection and transformation and the model training algorithm applied are the same as for DO simulation. Input forcings include freshwater discharges and nutrient loads from major rivers, air temperature, and wind. For the training part, the model yields R^2 of 0.96 and RMSE=0.62 km^3 (Fig. 13).

We also conducted tests using other machine learning methods including multiple linear regression and decision tree with the same input forcing as used in the trained neural network models. Results show that the linear model (RMSE=1.46 km^3) is unable to capture the interannual variability (especially during summer months) and it frequently produces negative value during the winter months (Fig. 14). On the contrary, the decision tree (RMSE=1.35 km^3) and neural network (RMSE=1.29 km^3) methods seem to have better performance. Particularly, the decision tree almost perfectly predicts the zero values. While it is hard to determine which method is always the best, nonlinear models are usually regarded as more suitable than linear models when dealing with water quality variables. It is often a good strategy to test multiple data methods and examine the performance difference among them. Using the ensemble mean of results from multiple methods could be a good option to account for the uncertainties associated with choice of the machine learning method.

4. DISCUSSION

4.1 Robustness of data-driven model

Different from traditional empirical or numerical models, the data-driven model proposed in this study has shown advantages in several aspects. Firstly, the approach is purely based on reliable and systematic monitoring or reanalysis data, which avoids the uncertainties propagating and accumulating from base level of hydrodynamic models to higher level of water quality models. The error accumulation is a common issue for many deterministic model systems. With advanced numerical tools and data-assimilation techniques, performance of hydrodynamic simulation has been improved greatly over time. Nevertheless, errors still exist even in a high-resolution numerical hydrodynamic models with a high order of numerical schemes (e.g., Testa et al., 2014; Irby et al., 2016; Ye et al., 2019). Secondly, the proposed approach can simulate not only the temporal but also spatial variations by extracting the major components of the target water quality variable. This is very different from previous 1D empirical models (e.g., Scardi and Harding, 1999; Scarvia et al., 2006). Thirdly, the approach is highly computationally efficient, taking about 1 hour with one cpu to train the models for all the selected primary EOF modes. The high computational efficiency makes the data-driven model an efficient way for environmental research and management. Since the model uses external forcings as used by a traditional 3D water quality model, the data-driven approach can be used to predict water quality conditions under future climate scenarios in response to change in environmental forcing conditions, such as changes in temperature, wind, and river flow. The nonlinear influence of forcing conditions on water quality is implicitly inherited inside the data-driven model (Shen et al., 2019). However, it should be noted that such

predictions may only be possible when the training dataset has included the variations of specific forcing and that the trained model has correctly resolved the response of target variable to the given forcing. Finally, even though the approach is used for the DO problem in Chesapeake Bay, similar or same methodology is readily applicable to other coastal systems, as long as there is enough data to train a model with reasonable accuracy. How much data is enough? There is not a straight answer. The required data length depends on the problem to be studied and the timescale of major driving processes for both the biogeochemical and physical condition. The training dataset should at least cover multiple dominant periods during which major signals (e.g., seasonal cycle and interannual variations) are noticeable.

It is important to point out that there are a variety of data methods well developed for machine learning, ranging from the simplest linear regression to sophisticated nonlinear regression such asnearest neighbors, random tree, and support vector machine (Fig. S6). We tested these methods for the DO problem with the same input variables and target variable. Except for the K nearest neighbors, all other methods tested yield similar model performance, even though the neural network model appears to be slightly better than others (Fig. 15).

4.2 Lessons learned

When using the model, one has to be cautious when collecting and transforming forcing data. As described in section 2, in this study, only forcing data continuously monitored at boundaries of the systems are used. These boundaries can be the air-sea interface, river boundary, and estuary-ocean interface. No data within the water body is used. For instance, the nutrient concentration and water salinity at mainstem stations are

not used, even though they are available and believed to be highly correlated with the DO condition. Such protocol makes the proposed method distinctly different from many previous studies that used the *in situ* measurement to represent either the physical or biogeochemical conditions (e.g., Scavia et al., 2006; Ross and Stock, 2019). For instance, to estimate the hypoxic volume, Scavia et al. (2006) used the bottom-surface difference of DO to calculate the vertical DO flux. Scardi and Harding (1999) used *in situ* measured parameters including chlorophyll-a concentration, salinity, and photic zone depth to train a neural-network model of phytoplankton primary production. Avoiding using such type of forcings is crucial to make the data-driven model more capable for environmental management.

Another important lesson learned for the data-driven model is to choose appropriate transformations for each forcing. For coastal systems, especially large coastal systems, delayed response of water quality to change of external forcing is very common. For instance, Malone et al. (1988) suggest the maximum summer productivity in the mesohaline reach of Chesapeake Bay is caused by the recycled nutrient delivered to the system during the previous spring. From numerical experiments, Lee et al. (2013) suggests winter-spring wind could affect the summer hypoxia (i.e., with a time-lag of 4 months) by affecting the circulation pattern and thus the transport of spring bloom biomass along the mainstem and between shoal and deep channel. The time-lag shall likely differ in different regions of a coastal system and for different processes. It is usually subjective in determining how long the time-lag is. The ASFT tool proposed in this study will ease this effort and is particularly useful when training multiple models using a variety of forcings.

We found that wind is one of the important variables for hypoxic volume simulation and the wind speed is first selected forcing by the ASFT module (Table S6). Although wind forcing is considered to only have a short-term impact, wind forcing can be more influential than expected for interannual dynamics. Recent studies confirm the great control of wind on stratification, estuarine circulation, and thus hypoxic volume (Scully, 2010; Shen et al., 2013; Wilson et al., 2015; Du and Shen, 2016; Jiang and Xia, 2017; Du et al. 2018). Numerical experiments in previous studies (e.g., Hong and Shen, 2013; Scully, 2013) suggest the vertical mixing is less sensitive to the synoptic variability of river discharge but more sensitive to the wind condition. Scully (2013, 2016) finds that the mean summer wind speed is the single-most important physical variable contributing to the variations of hypoxic volume. Changes in physical condition induced by changing wind fields in the future should be considered for the environmental and water quality management in Chesapeake Bay.

The selection of forcing and transformation shall not be used directly to interpret the relative importance of one forcing over another because of high covariance among different forcings and the difficulty to isolate the influence of each forcing based on the correlation analysis. If two forcings are highly correlated, only the forcing with higher R^2 is selected, which is different from deterministic models. Nevertheless, the information of forcing selection is helpful to determine the relative complexity of each mode. For instance, only four forcings are selected for the first EOF mode (Table S1) while more than 20 forcings for second-fifth modes, suggesting higher complexity of dynamics responsible for these later modes.

Of particular interest is the forcing selection for the first EOF mode, the dominant mode containing not only the seasonal signal of bottom DO but also the interannual signal of bay-wide hypoxia (Fig. 5). The first mode is only related to air temperature, Susquehanna streamflow shifted by 40 days, and downwelling shortwave radiation shifted by 10 days (Table S1). Considering the consistency of seasonality among the first mode, air temperature, and solar radiation, it seems that the data model uses radiation and/or temperature to learn the seasonal cycle, and uses the streamflow in controlling interannual variability. Because nutrient loadings are highly correlated to streamflow, the information of nutrient loading is implicitly included. This is broadly consistent with other studies using either statistic or 3D numerical approaches (e.g., Hagy et al., 2004; Scavia et al., 2006; Testa et al., 2014).

4.3 Limitations of the data-driven model

It is important to acknowledge not only the advantages but also the limitations of a method. First, the accuracy of a trained model depends greatly on the length of data records. Even in a heavily monitored coastal system such as Chesapeake Bay, the length of data is still limited because of the monitoring frequency. One critical question is whether results of the bi-monthly surveys is representative for the monthly mean condition. By comparing the variability of measurement and simulation results of DO in Chesapeake Bay, Bever et al. (2013) demonstrated the monthly measurements need to be corrected in order to be spatially and temporally representative. Without enough temporal resolution, the EOF analysis of the existing dataset may be overwhelmed by minor modes, requiring more modes to be included in order to cover the major variance. Second, the model performance may be hampered by the unavailability of systematic

forcing. Monitoring instrument malfunction occurs frequently, especially after major extreme weather events such as hurricanes, which may result in a significant data gap. In these cases, reanalysis model results that assimilate observations could be good data sources. Taking the wind data for instance, instead of using the monitoring data, we used the outputs of reanalysis atmospheric models. Perhaps, the most concerning shortcoming is that data-driven models are not process-orientated, making them hard for process-based management and research. As many forcing variables are highly correlated with each other, the contribution of an individual forcing variable cannot be fully evaluated based on sensitivity tests although the model has a high predictive skill. However, it does not mean such models cannot be used for environmental management. Actually, there is an increasing need of using data-driven model for water quality management (e.g., Orougi et al., 2013; Chang et al. 2015; Shen et al., 2019; Ross and Stock, 2019). Shen et al. (2019) shows a data-driven model can be used for predicting the response of Chl-a to changes in nutrient loading if the appropriate parameter variables are used. Through model sensitivity tests by introducing different forcings and using different transformations, the data-driven model could be helpful in identifying possible important factors.

5. CONCLUSIONS

We present a data-driven model to efficiently and accurately simulate and predict DO conditions in estuarine and coastal waters. Different from previous statistical models that often use *in situ* measurements, the proposed approach relies purely on the external forcings, which make it more suitable for environmental assessment. Evaluation of the model performance suggests a high capability of data-driven model for water quality simulation. Based on a similar forcing dataset as required by a 3D numerical model, the

proposed approach predicts well the DO and hypoxic condition spatially and temporally. Even though the model is tested for the DO condition in Chesapeake Bay, the framework and methodology are readily transferable to any other coastal systems that are systematically monitored. Overall, this study provides a robust framework and methodology, upon which future research could be based.

With the quickly accumulated observational data and latest advances in machine-learning techniques, the data-driven model is a promising approach with high efficiency for water quality modeling and environmental management in the near future. In fact, there is increasing interest in using machine-learning techniques for water quality simulations (e.g., Shen et al., 2019; Ross and Stock, 2019). There are, however, still some questions that remain to be further explored. For example, how to transform the forcing and target variable appropriately? How to faithfully extract the principal component? And how to determine whether noncontinuous measurements are representative for monthly or daily conditions? Even though this study has attempted to answer them, the answers may vary depending on specific problems and characteristics of the given estuarine and coastal system.

ACKNOWLEDGEMENTS

We want to thank the three anonymous reviewers for their valuable and helpful comments. The long-term monitoring data for the dissolved oxygen in Chesapeake Bay are available at Chesapeake Bay Program (http://data.chesapeakebay.net/WaterQuality). Freshwater discharge data of major rivers are available United States Geological Survey (Susquehanna River at https://waterdata.usgs.gov/md/nwis/uv?site_no=01578310;

Potomac River at https://waterdata.usgs.gov/md/nwis/uv?site_no=01646500; James River at https://waterdata.usgs.gov/usa/nwis/uv?02037500; Choptank River at https://waterdata.usgs.gov/usa/nwis/uv?01491000). Data of air temperature at Chesapeake Bay Bridge Tunnel are available at NOAA (https://tidesandcurrents.noaa.gov/stationhome.html?id=8638863). Hourly reanalysis wind data are produced by European Centre for Medium-Range Weather Forecasts and available online https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. The code showing the framework of the data method can be found online https://zenodo.org/record/3973756. This is contribution NO. 3934 of Virginia Institute of Marine Science, the College of William and Mary.

REFERENCES

Bajo, M., & Umgiesser, G. (2010). Storm surge forecast through a combination of dynamic and neural network models. *Ocean Modelling*, *33*(1–2), 1–9.

Bever, A. J., Friedrichs, M. A. M., Friedrichs, C. T., Scully, M. E., & Lanerolle, L. W. J. (2013). Combining observations and numerical model results to improve estimates of hypoxic volume within the Chesapeake Bay, USA. *Journal of Geophysical Research C: Oceans*, *118*(10), 4924–4944.

Bianchi, T. S., DiMarco, S. F., Cowan, J. H., Hetland, R. D., Chapman, P., Day, J. W., & Allison, M. A. (2010). The science of hypoxia in the northern Gulf of Mexico: A review. *Science of the Total Environment*, *408*(7), 1471–1484.

Breitburg, D., Levin, L. A., Oschlies, A., Grégoire, et al. (2018). Declining oxygen in the global ocean and coastal waters. *Science*, *359*, eaam*7240*.

Campolo, M., Andreussi, P., & Soldati, A. (1999). River flood forecasting with a neural network model. *Water Resources Research*, *35*(4), 1191–1197.

Carstensen, J., Andersen, J. H., Gustafsson, B. G., & Conley, D. J. (2014). Deoxygenation of the Baltic Sea during the last century. *Proceedings of the National Academy of Sciences*, *111*(15), 5628–5633.

Cerco, C. F., & Noel, M. R. (2013). Twenty-one-year simulation of Chesapeake Bay water quality using the ce-qual-icm eutrophication model. *Journal of the American Water Resources Association*, 1-15.

Chang, F. J., & Chen, Y. C. (2003). Estuary water-stage forecasting by using radial basis function neural network. *Journal of Hydrology*, *270*, 158–166.

Chang, F., Tsai, Y., Chen, P., Coynel, A., & Vachaud, G. (2015). Modeling water quality in an urban river using hydrological factors - Data driven approaches. *Journal of Environmental Management*, *151*, 87–96.

Chen, C., Liu, H., & Beardsley, R. C. (2003). An Unstructured Grid, Finite-Volume, Three-Dimensional, Primitive Equations Ocean Model: Application to Coastal Ocean and Estuaries. *Journal of Atmospheric and Oceanic Technology*, *20*, 159–186.

Chen, W.B., Liu, W.C., Hsu, M.H., 2012. Comparison of ANN approach with 2D and 3D hydrodynamic models for simulating estuary water stage. *Advances in Engineering Software*, *45*, 69–79.

Conley, D. J., Humborg, C., Rahm, L., Savchuk, O. P., & Wulff, F. (2002). Hypoxia in the Baltic Sea and basin-scale changes in phosphorus biogeochemistry. *Environmental Science & Technology*, *36*(24), 5315–5320.

Da, F., Friedrichs, M. A. M., & St-Laurent, P. (2018). Impacts of Atmospheric Nitrogen Deposition and Coastal Nitrogen Fluxes on Oxygen Concentrations in Chesapeake Bay. *Journal of Geophysical Research : Oceans*, *123*, 5004–5025.

Deng, J., Paerl, H. W., Qin, B., Zhang, Y., Zhu, G., et al. (2018). Climatically-modulated decline in wind speed may strongly affect eutrophication in shallow lakes. *Science of the Total Environment*, *645*, 1361–1370.

Diaz, R. J., & Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *Science (New York, N.Y.)*, *321*(5891), 926–929.

Du, J., & Shen, J. (2015). Decoupling the influence of biological and physical processes on the dissoved oxygen in the chesapeake Bay. *Journal of Geophysical Research, 120,* 78–93.

Du, J. Shen, J. (2016). Water residence time in Chesapeake Bay from 1980-2012. *Journal of Marine Systems, 164,*101-111.

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.

Du, J., Park, K., Shen, J., Zhang, Y. J., Yu, X. et al. (2019). A hydrodynamic model for Galveston Bay and the shelf in the northern Gulf of Mexico. *Ocean Science, 15,* 951–966.

Fennel, K., & Testa, J. M. (2019). Biogeochemical controls on coastal hypoxia. *Annual Review of Marine Science, 11,* 105–103.

Hagan, M.T., and M. Menhaj (1999). Training feed-forward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks,* 5, 989–993.

Hagy, J. D., Boynton, W. R., Keefe, C. W., & Wood, K. V. (2004). Hypoxia in Chesapeake Bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. *Estuaries, 27*(4), 634–658.

Harding, L. W. (1994). Long-term trends in the distribution of phytoplankton in Chesapeake Bay: Roles of light, nutrients and streamflow. *Marine Ecology Progress Series, 104,* 267–291.

Hong, B., & Shen, J. (2012). Responses of estuarine salinity and transport processes to potential future sea-level rise in the Chesapeake Bay. *Estuarine, Coastal and Shelf Science*, *104–105*, 33–45.

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, *31*(10), 2517.

Irby, I. D., Friedrichs, M. A. M., Friedrichs, C. T., Bever, A. J., Hood, R. R., et al. (2016). Challenges associated with modeling low-oxygen waters in Chesapeake Bay : a multiple model comparison. *Biogeosciences*, *13*, 2011–2028.

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

Karstensen, J., Stramma, L., & Visbeck, M. (2008). Progress in Oceanography Oxygen minimum zones in the eastern tropical Atlantic and Pacific oceans. *Progress in Oceanography*, *77*, 331–350.

Keiner, L. E., & Yan, X. H. (1998). A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. *Remote Sensing of Environment*, *66*(2), 153–165.

Kemp, W. M., Boynton, W. R., Adolf, J. E., Boesch, D. F., Boicourt, W. C., Brush, G. et al. (2005). Eutrophication of Chesapeake Bay : historical trends and ecological interactions. *Marine Ecology Progress Series*, *303*, 1–29.

Krasnopolsky, V. M. (2007). Neural network emulations for complex multidimensional geophysical mappings: Applications of neural network techniques to atmospheric and oceanic satellite retrievals and numerical modeling. *Reviews of Geophysics*, *45*(3), 1–34.

Lee, Y. J., Boynton, W. R., Li, M., & Li, Y. (2013). Role of late winter–spring wind influencing summer hypoxia in Chesapeake Bay. *Estuaries and Coasts*, *36*(4), 683–696.

Malone, T., Kemp, W., Ducklow, H., Boynton, W., Tuttle, J., & Jonas, R. (1986). Lateral variation in the production and fate of phytoplankton in a partially stratified estuary . *Marine Ecology Progress Series*, *32*, 149–160.

Malone, T. C., Crocker, L. H., Pike, S. E., & Wendler, B. W. (1988). Phytoplankton production in a partially stratified. *Marine Ecology Progress Series*, *48*, 235–249.

Marquardt, D. (1963). An Algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11, 431–441.

Meier, H. E. M., Hordoir, R., Andersson, H. C., Dieterich, C., Eilola, K., et al. (2012). Modeling the combined impact of changing climate and changing nutrient loads on the Baltic Sea environment in an ensemble of transient simulations for 1961–2099. *Climate Dynamics*, *39*, 2421–2441.

Morris, J. T., Porter, D., Neet, M., Noble, P. A., Schmidt, L., Lapine, L. A., & Jensen, J. R. (2005). Integrating LIDAR elevation data, multi-spectral imagery and neural network modelling for marsh characterization. *International Journal of Remote Sensing*, *26*(23), 5221–5234.

Muller, A. C., & Muller, D. L. (2015). Forecasting future estuarine hypoxia using a wavelet based neural network model. *Ocean Modelling*, *96*, 314–323.

Murphy, R. R., Kemp, W. M., & Ball, W. P. (2011). Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. *Estuaries and Coasts*, *34*(6), 1293–1309.

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

Newcombe, C. L., and W. A. Horne (1938). Oxygen-poor waters of the Chesapeake Bay. *Science*, *88*, 80–81.

Orouji, H., Bozorg-haddad, O., & Fallah-mehdipour, E. (2013). Modeling of water quality parameters using data-driven models. *Journal of Environmental Engineering*, 139, 947-957.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, *36*(1), 2–17.

Rabalais, N. N., Turner, R. E., & Wiseman, W. J. (2002). Gulf of Mexico hypoxia, a.k.a. "the Dead Zone." *Annual Review of Ecology and Systematics*, *33*(1), 235–263.

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

Ross, A. C., & Stock, C. A. (2019). An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science*, *221*(December 2018), 53–65.

Scardi, M., & Harding, L. W. (1999). Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecological Modelling*, *120*(2–3), 213–223.

Scavia, D., Kelly, E. L. a, & Hagy, J. D. (2006). A simple model for forecasting the effects of nitrogen loads on Chesapeake Bay hypoxia. *Estuaries and Coasts*, *29*(4), 674–684.

Scully, M. E. (2010). Wind modulation of dissolved oxygen in Chesapeake Bay. *Estuaries and Coasts*, *33*, 1164–1175.

Scully, M. E. (2013). Physical controls on hypoxia in Chesapeake Bay: A numerical modeling study. *Journal of Geophysical Research: Oceans*, *118*(3), 1239–1256.

Scully, M. E. (2016). The contribution of physical processes to inter-annual variations of hypoxia in Chesapeake Bay : A 30-yr modeling study. *Limnology and Oceanography*, *61*, 2243–2260.

Shen, J., & Wang, H. V. (2007). Determining the age of water and long-term transport timescale of the Chesapeake Bay. *Estuarine, Coastal and Shelf Science*, *74*(4), 750–763.

Shen, J., Hong, B., & Kuo, A. Y. (2013). Using timescales to interpret dissolved oxygen distributions in the bottom waters of Chesapeake Bay. *Limnology and Oceanography*, *58*(6), 2237–2248.

Shen, J., Qin, Q., Wang, J., Sisson, M. 2019. A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading. *Ecological Modeling*, *398*, 44-54.

Taft, J. L., Taylor, W. R., Hartwig, E. O., and Loftus, R. (1980). Seasonal oxygen depletion in Chesapeake Bay. *Estuaries*, 3(4), 242–247.

Warner, J. C., Geyer, W. R., & Lerczak, J. A. (2005). Numerical modeling of an estuary: A comprehensive skill assessment. *Journal of Geophysical Research C: Oceans*, *110*(5), 1–13.

Testa, J. M., Li, Y., Lee, Y. J., Li, M., Brady, D. C., Di Toro, D. M., et al. (2014). Quantifying the effects of nutrient loading on dissolved O2 cycling and hypoxia in

Chesapeake Bay using a coupled hydrodynamic–biogeochemical model. *Journal of Marine Systems, 139*, 139–158.

Willmott, C. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194.

Wilson, R. E., Bratton, S. D., Wang, J., & Colle, B. A. (2015). Evidence for directional wind response in controlling inter-annual variations in duration and areal extent of summertime hypoxia in western Long Island Sound. *Estuaries and Coasts, 38*, 1735–1743.

Vaquer-Sunyer, R., & Duarte, C. M. (2008). Thresholds of hypoxia for marine biodiversity. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(40), 15452–15457.

Vilas, L. G., Spyrakos, E., & Palenzuela, J. M. T. (2011). Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sensing of Environment*, *115*(2), 524–535.

Ye, F., Zhang, Y. J., Wang, H. V, Friedrichs, M. A. M., Irby, I. D., Alteljevich, E., Valle-Levinson, A., Wang, Z., Huang, H., Shen, J., Du, J. (2018). A 3D unstructured-grid model for Chesapeake Bay : Importance of bathymetry. *Ocean Modelling*, *127*, 16–39. https://doi.org/10.1016/j.ocemod.2018.05.002

Ye, F., Zhang, Y. J., He, R., Wang, Z., Wang, H. V, & Du, J. (2019). Third-order WENO transport scheme for simulating the baroclinic eddying ocean on an unstructured grid. *Ocean Modelling*, *143*(April), 101466. https://doi.org/10.1016/j.ocemod.2019.101466

Table 1: A list of the transformation options in the data-driven model.

Transformation	**Sub-types**	**Formula**
Time-lag transformation	1-7	$\phi(t)=x(t-lag)$, with lag ranging within 0, 10, …60 days
Accumulative transformation	1-13	$\phi(t)=mean(x(\tau))$, where $\tau \in [t1-acc, t2]$, with acc ranging from 0 to 120. *t1* and *t2* are the beginning and end of each month; t1=t-15 and t2=t+15.
Regular transformation	1	$\phi=x$
	2	$\phi=\log(x)$
	3	$\phi=1/x$
	4	$\phi=\exp((x-mean(x))/std(x))$
	5	$\phi=x/(p50+x)$, also known as Monod-type filter
	6	$\phi=x/(p75+x)$
	7	$\phi=x/(p25+x)$
	8	$\phi=(x-mean(x))/std(x)$
Notes:	x: forcing variable ϕ:transformed forcing variable t: time std(x): the standard deviation of x min(x): the min value of x mean(x): the mean value of x P25, P50, P75: the 25, 50, and 75 percentile of x	

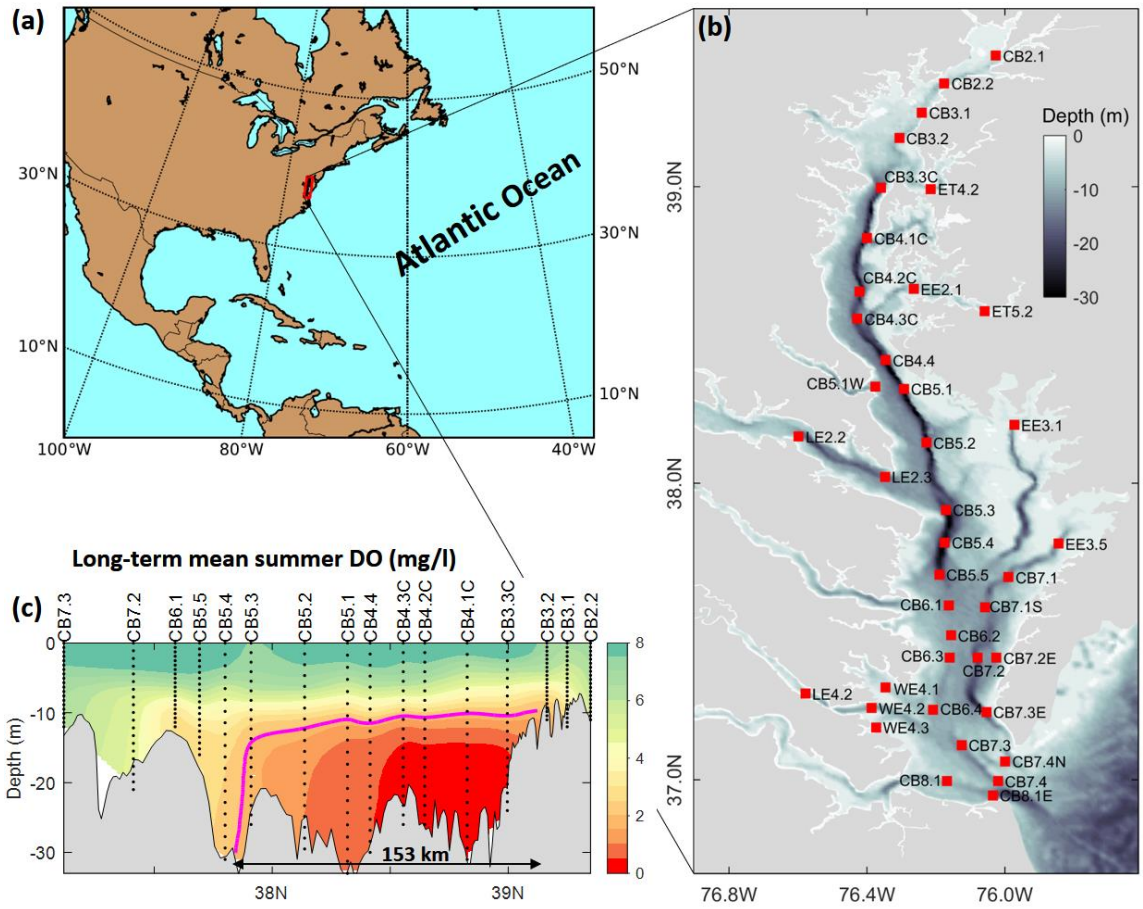

Figure 1: (a) Map of North America, with a red rectangle showing the location of Chesapeake Bay. (b) 40 long-term (1985-present) Chesapeake Bay Program monitoring stations. Data at these stations have less than 10% data gap and are used in this study. (c) The 32-year mean of the summer (Jun-August) DO concentration along the main axis of the bay, with the thick magenta line denoting the 2 mg/l (i.e., hypoxia threshold).

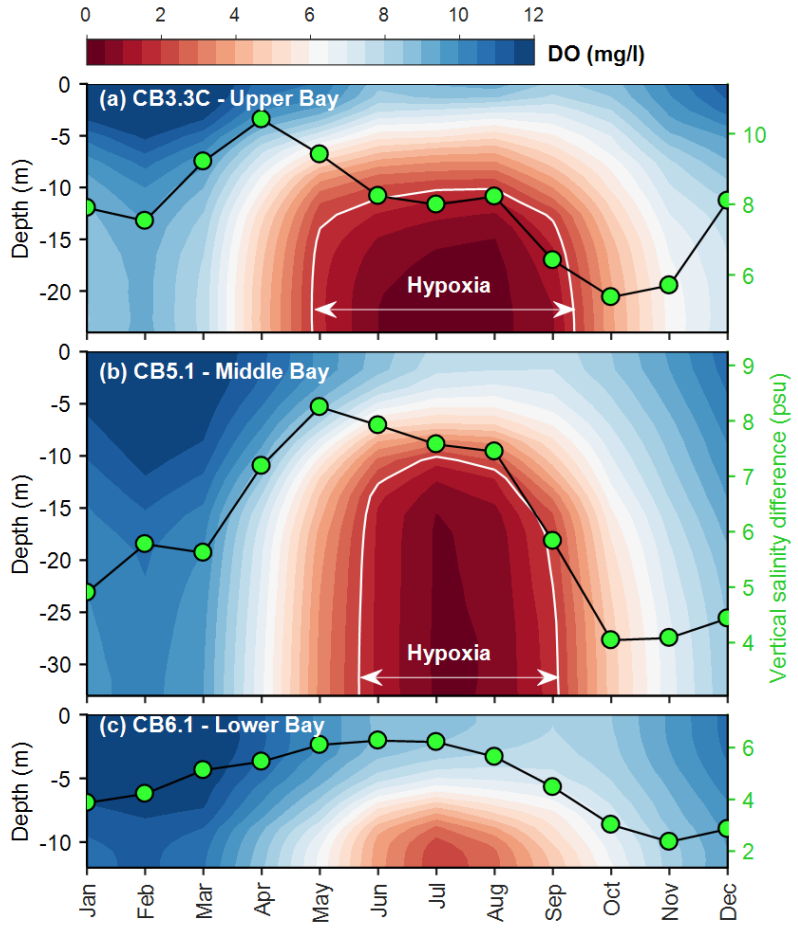

Figure 2: Seasonality of the vertical profile of dissolved oxygen concentration and stratification at selected three mainstem stations, representing the (a) upper bay, (b) middle bay, and (c) lower bay, respectively (see Fig. 1 for the location of these stations). White lines in each panel indicate the 2 mg/l contour line, while green dots show the seasonal variation of stratification characterized by the difference between bottom and surface salinity.

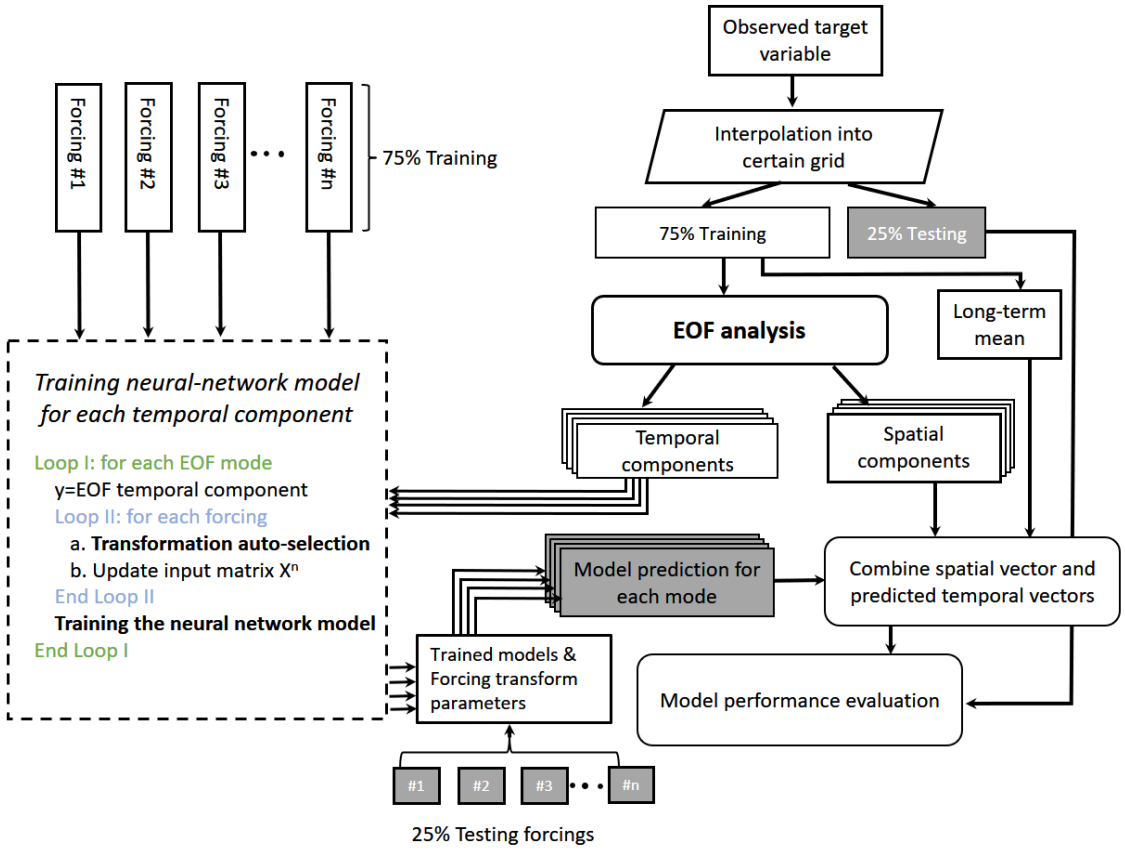

Figure 3: A sketch diagram showing the workflow of the data-driven approach.

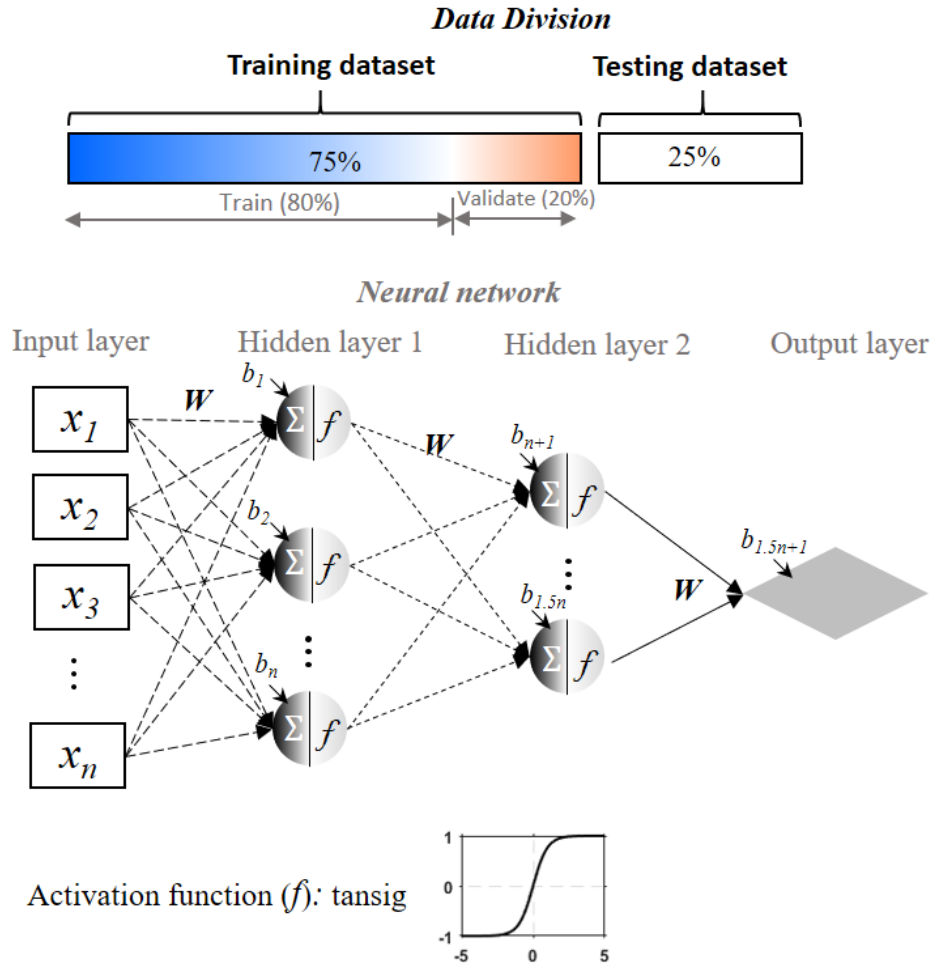

Figure 4: A sketch diagram showing the data-division scheme, the structure of a sample neural network, and the type of activation function used in the proposed data-driven model. For the data-division scheme, the full data is separated into training and testing sub-dataset; within the training dataset, 80% and 20% are used to "train" and "validate" by the neural network.

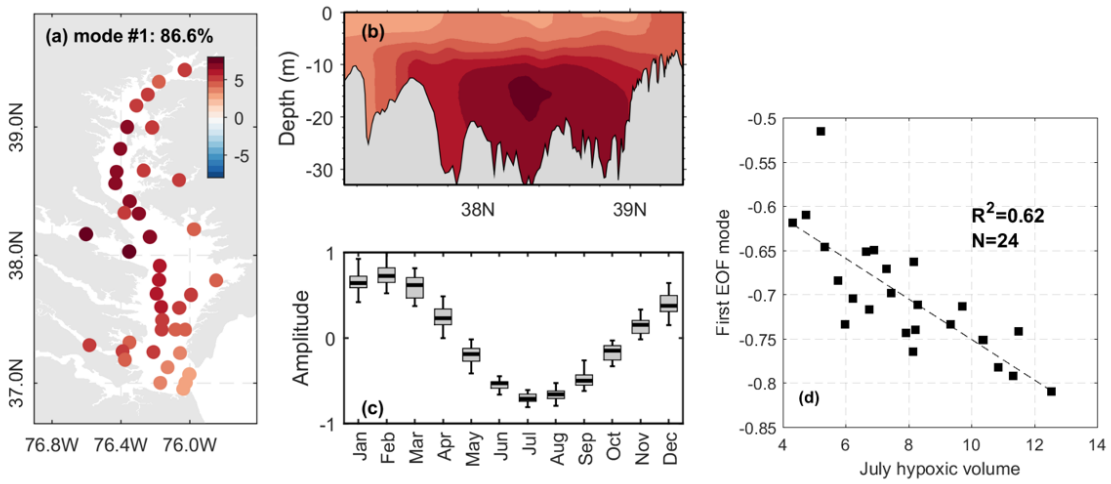

Figure 5: Characteristics of the first EOF mode that constitutes 86.6% of the total variance. (a) Horizontal distribution of bottom value; (b) the vertical distribution of the mode along the bay's main axis. Color dots in (a) and filled contours in (b) share the same color scale. (c) Box plots showing the seasonality of the time series. (d) Relationship between interannual variations of July hypoxic volume (data from Bever et al., 2013) and the July value of temporal variation of the first EOF mode.

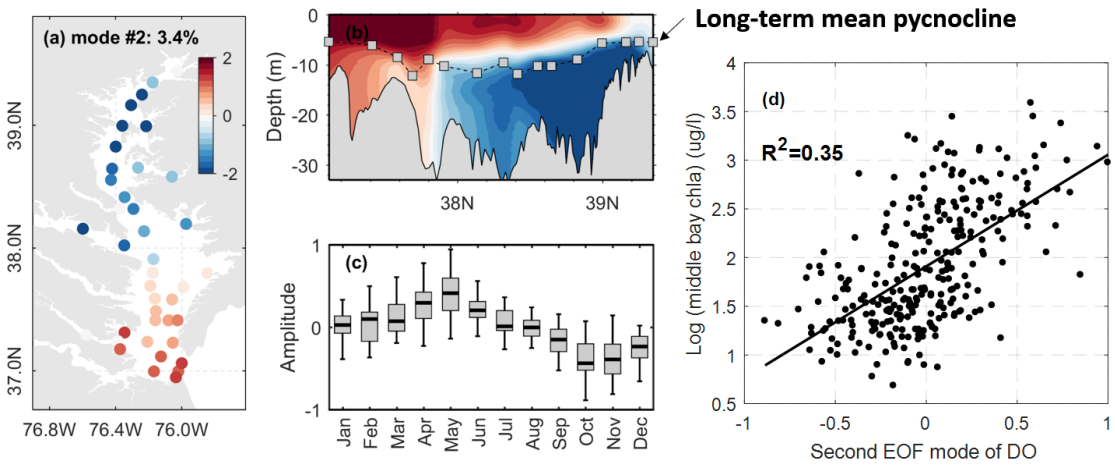

Figure 6: Characteristics of the second EOF mode. (a) Horizontal map of the bottom value. (b) Vertical distribution of second EOF mode along the mainstem, with gray solid rectangles denoting the long-term mean depth of pycnocline (determined as the depth where maximum salinity gradient occurs). (c) Seasonality of second EOF mode. (d) Relationship between chlorophyll-a averaged over middle bay stations, with each dot denoting the temporal value of second EOF mode at a given month and the logarithm of chlorophyll-a concentration in the previous month (i.e., with a one-month time-lag).

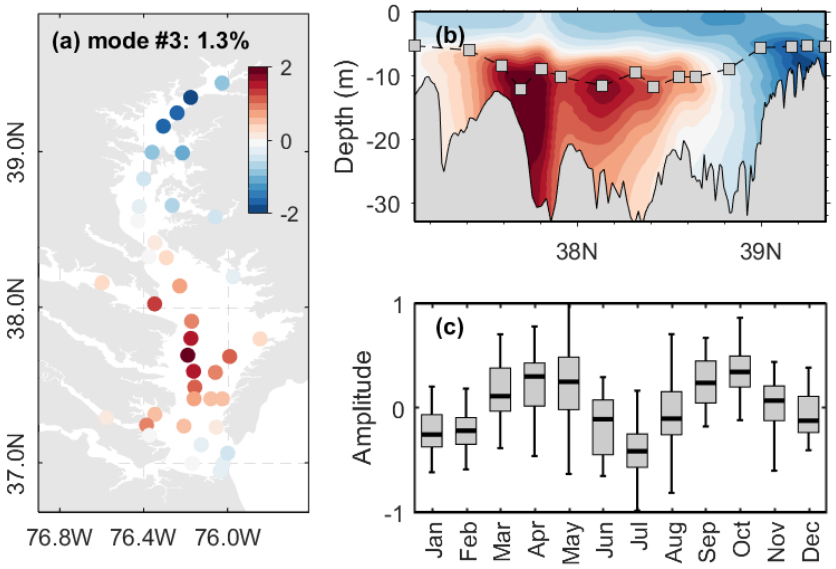

Figure 7: Characteristics of the third EOF mode (see caption of Fig. 5a-c).

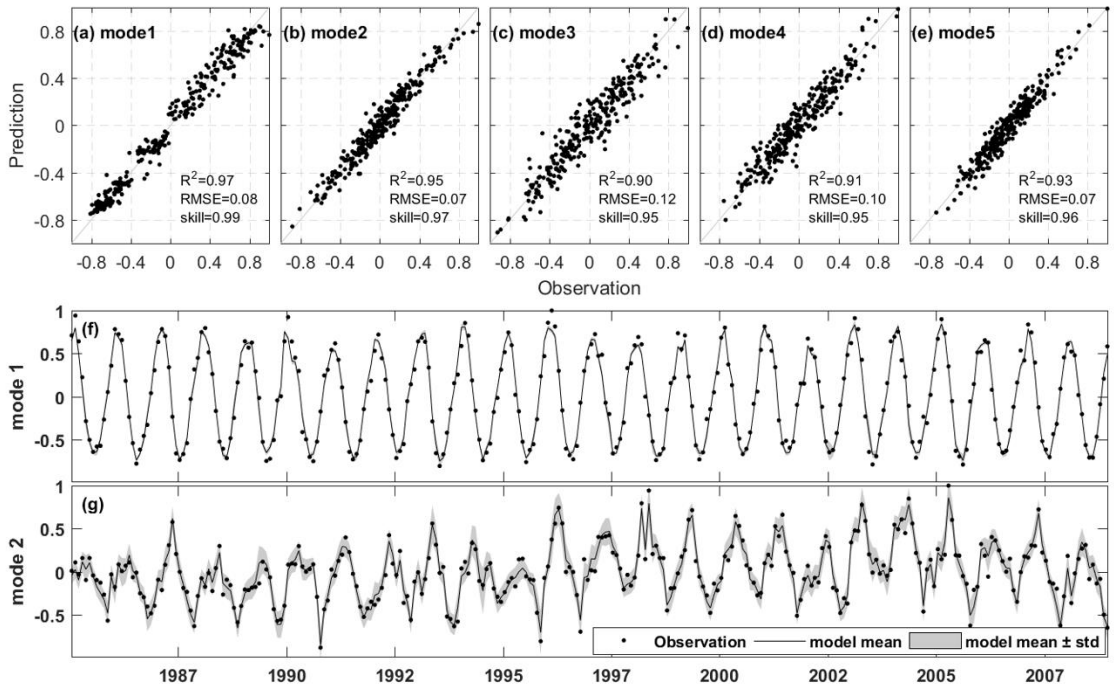

Figure 8: (a-e) Model training results for each of the first five EOF modes, with R^2, RMSE, and model skill shown in the bottom right of each panel. (f-g) Observed and modeled time series of first two EOF modes in the training dataset, with gray shade indicating the standard deviation of 100 times of neural network training.

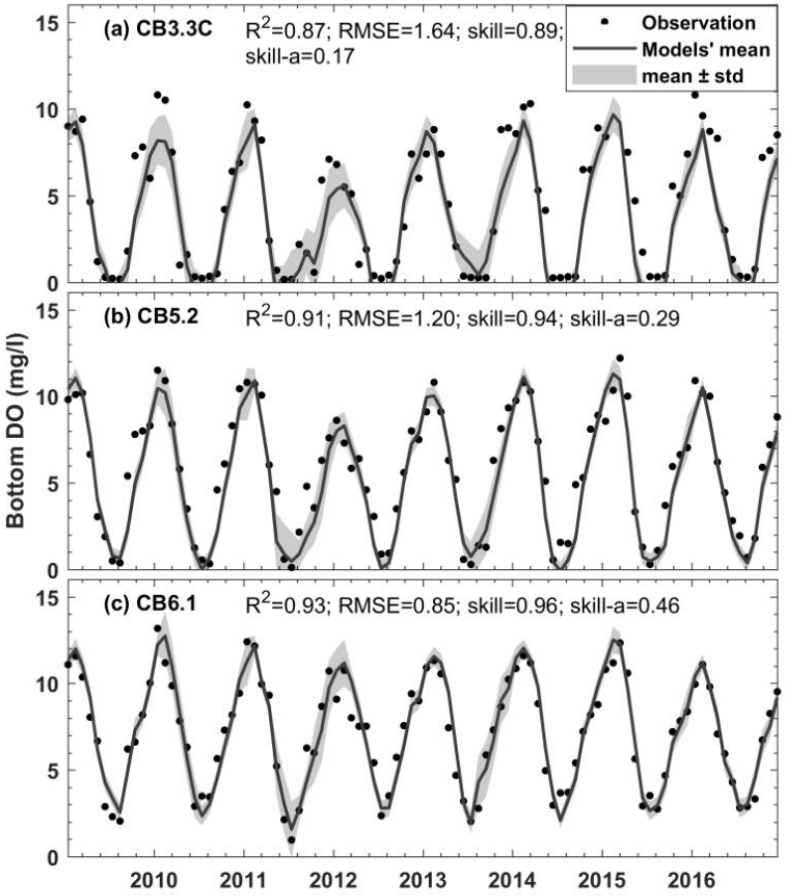

Figure 9: Observation (dot) and model prediction (line) of bottom DO at three selected mainstem stations, (a) CB3.3C, (b) CB5.2, and (c) CB6.1. Statistical indexes for the model performance are shown in the text on top. *Skill-a* indicates the *skill* for the anomaly prediction. The depth of the three stations are 25, 30, and 12m, respectively. Gray shade indicates the standard deviation of 100 times of neural network training.

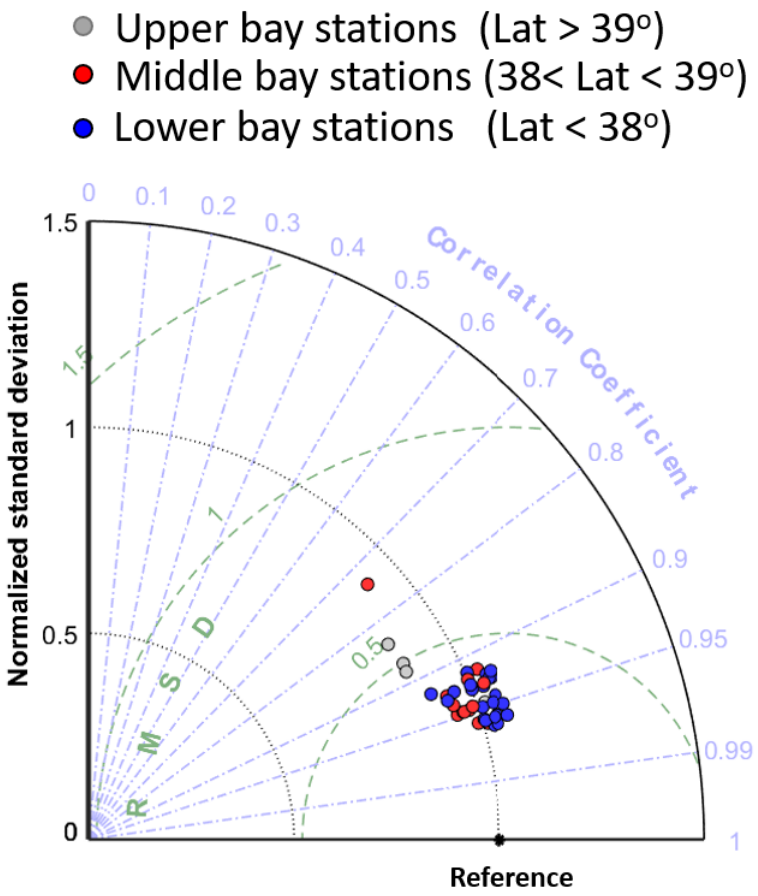

Figure 10: Taylor diagram illustrating the model performance for each of 40 stations, with different colors indicating stations in different regions. Only the 8-yr testing dataset (2009-2016) is used for the analysis. The radial distance from the origin is proportional to the ratio standard deviations; the azimuthal angle indicates the Pearson correlation coefficient; and the distance between each filled marker and the "reference" point indicates the centered root mean square deviation (RMSD).

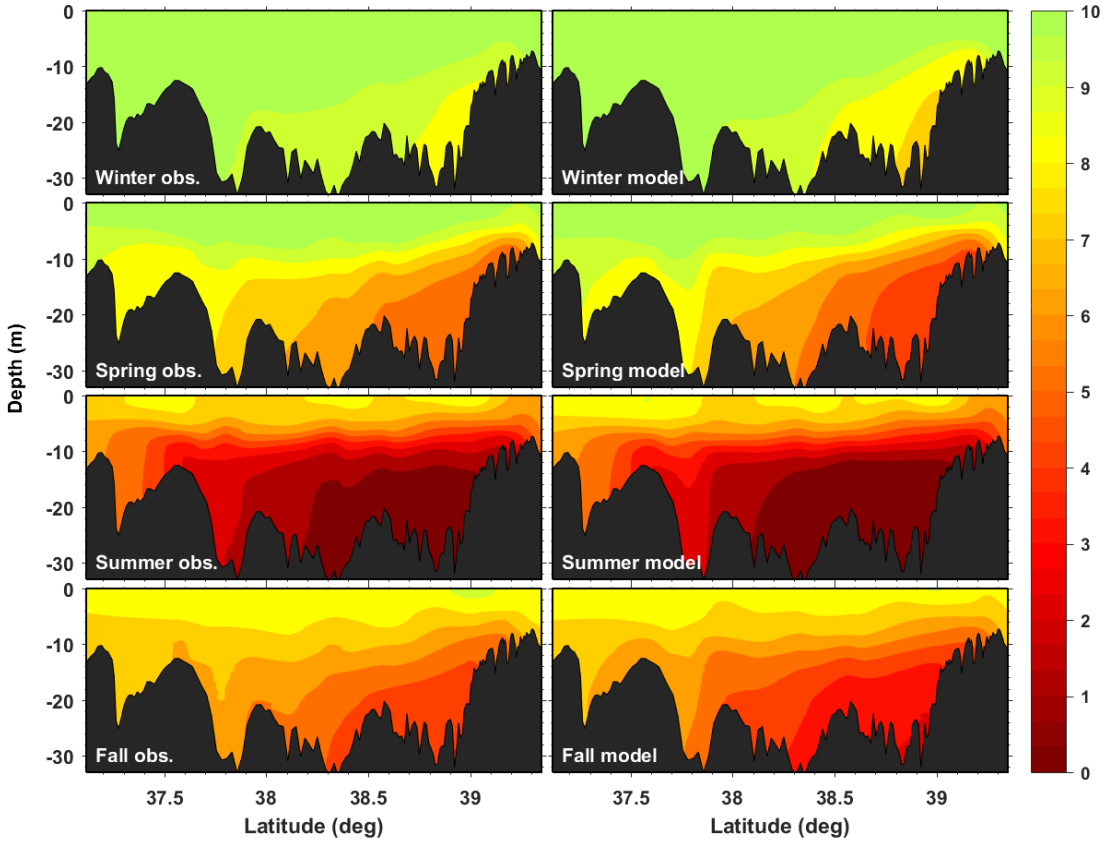

Figure 11: Spatial comparison between observed (left panels) and modeled (right panels) DO along the bay's mainstem.

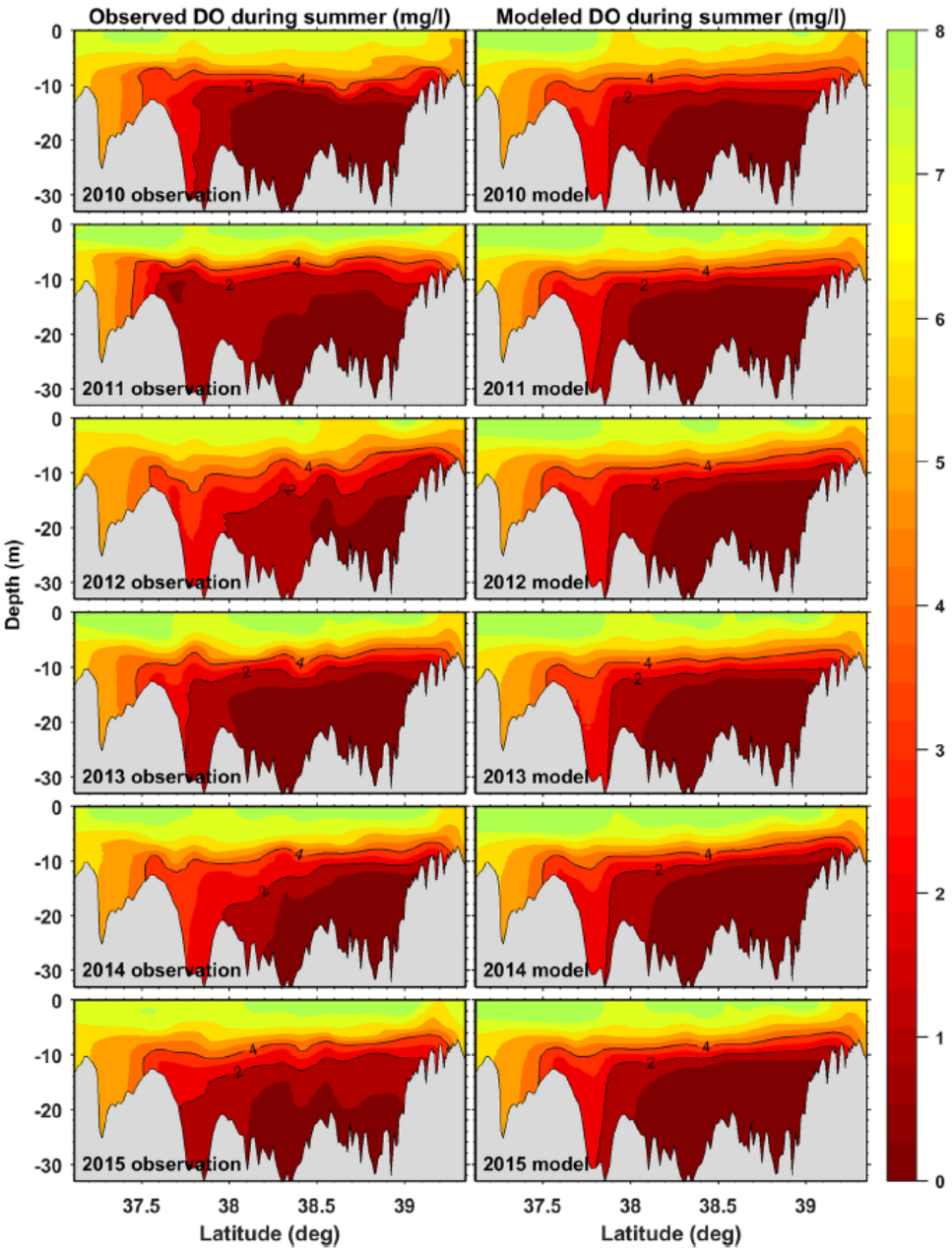

Figure 12: Vertical profile of the summer-mean DO concentration (averaged over June-August) along the bay's mainstem, with black lines denoting contour of 2 and 4 mg/l. Only testing dataset are shown here in order to demonstrate the model's capability in reproducing the spatial distribution when external forcings are provided.

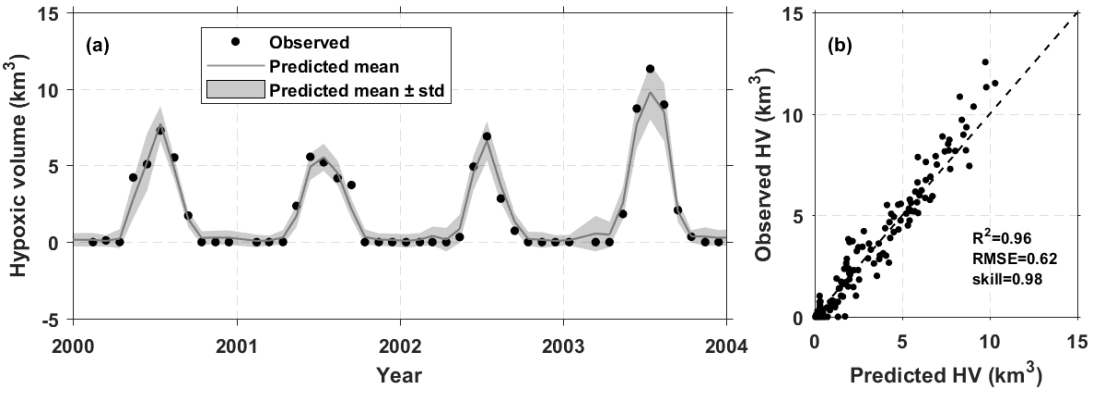

Figure 13: Training part for the hypoxic volume simulation. (a) shows the time series for the last four year in training period, while (b) shows the monthly data for the entire training period. Only training dataset of 1985-2004 is used for the training. In (a), the gray shading indicates the uncertainties calculated as the standard deviation of results from 100 times of neural network training.

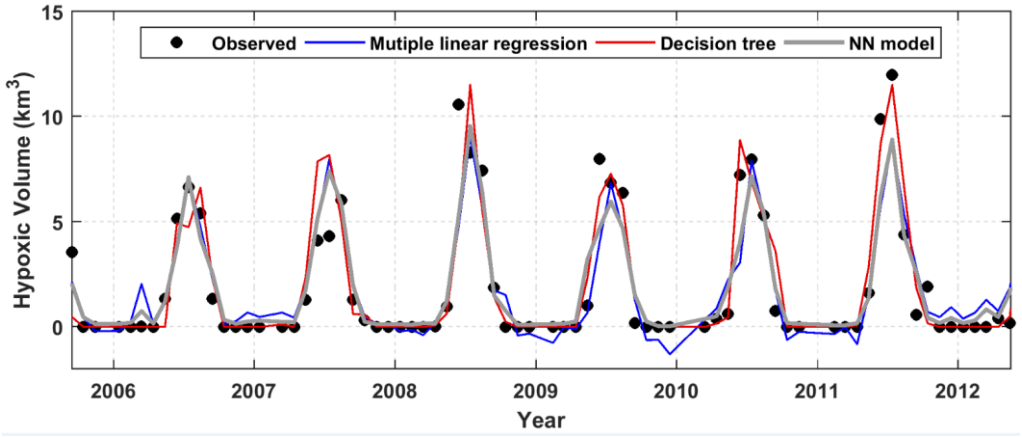

Figure 14: Testing part for the hypoxic volume simulation. Besides the neural network model, other methods including multiple linear regression, decision tree, and Bayesian regression are tested using the same input as in neural network.

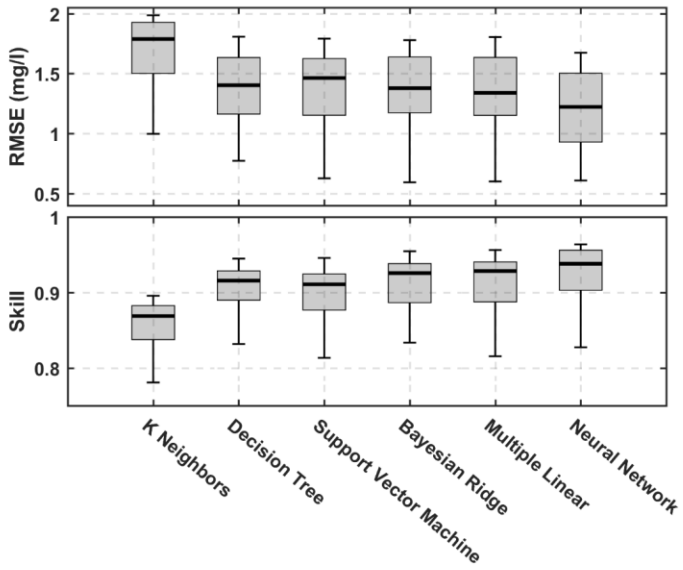

Figure 15: The model performance indicated by root mean square error (RMSE) and *skill* for the bottom DO. The gray bars and error bars indicate the mean and standard deviation of the performance over the 40 stations.

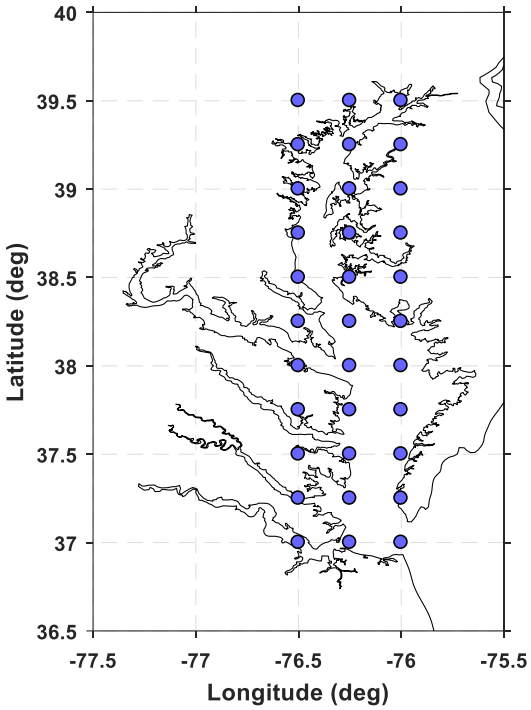

Figure S1: ECMWF ERA5 grid points (blue solid circles) used to get the bay-wide mean wind field. The ECMWF ERA5 global data has a spatial resolution of 0.25 degree for the wind field. The 33 grid points are within the longitude of [-76.5,-76] and latitude of [37, 39.5].

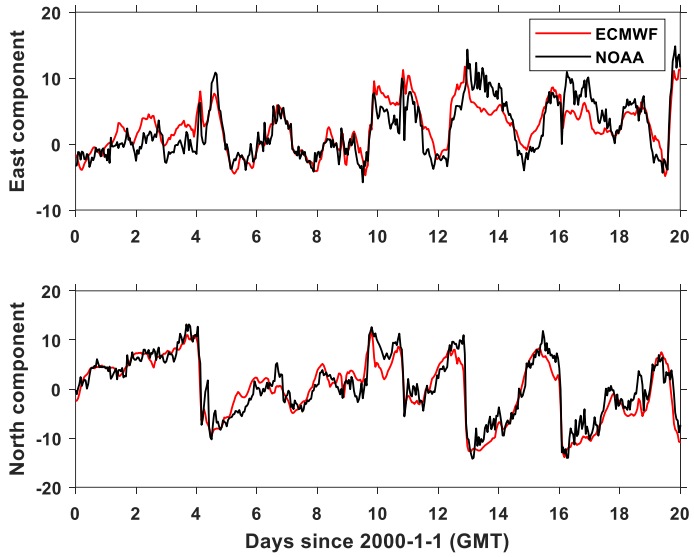

Figure S2: Wind data comparison between ECMWF reanalysis wind at (76W, 37N) and NOAA observations at CBBT. This figure shows that the reanalysis wind is overall consistent with the observation.

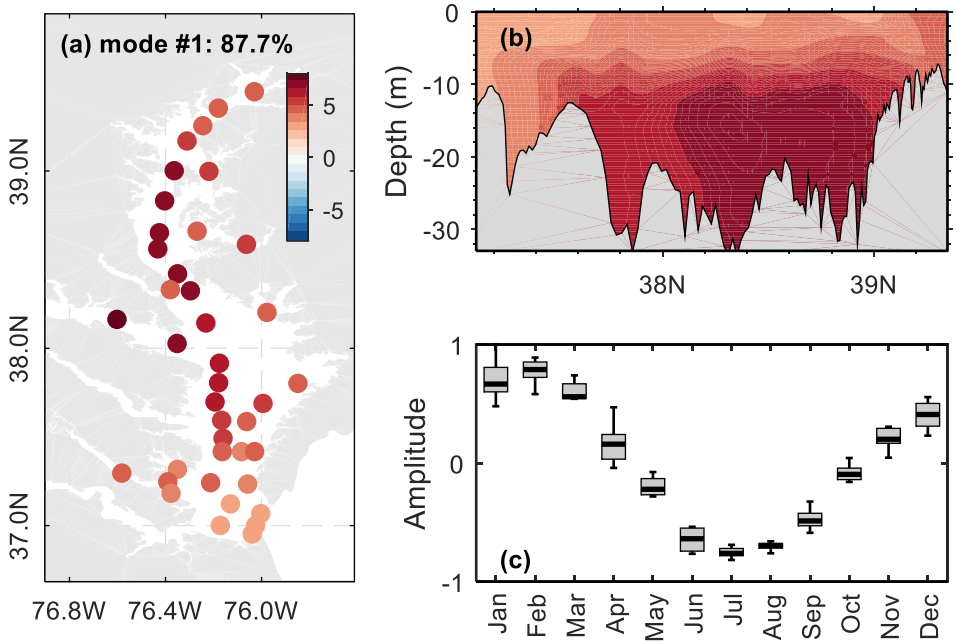

Figure S3: The spatial pattern and amplitude seasonality of the first EOF mode, based on the DO data in the testing period. (a) shows the spatial pattern of bottom DO and (b) shows the vertical distribution along the bay's mainstem. (a) and (b) share the same color scale.

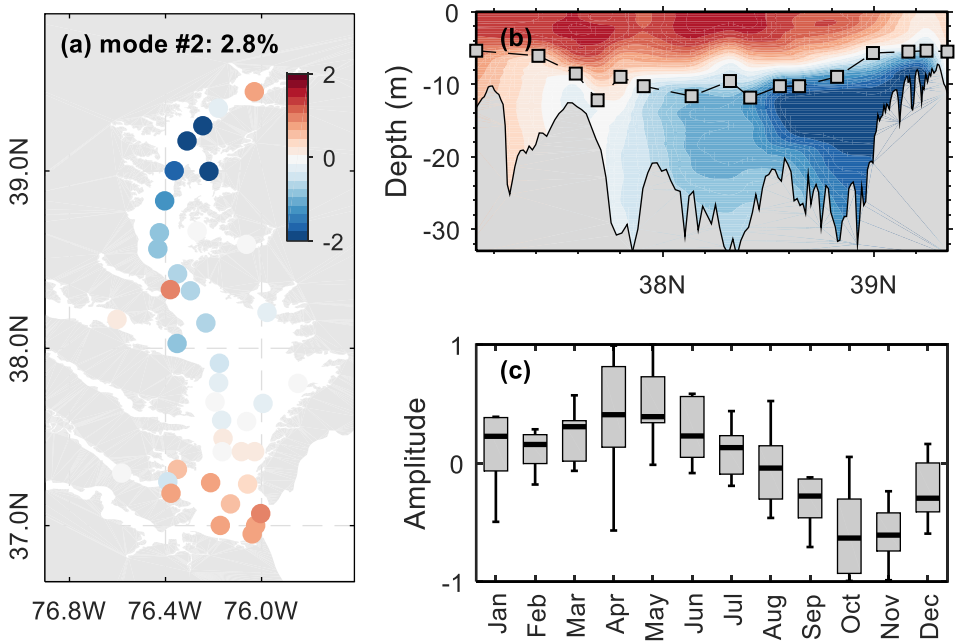

Figure S4: Same as Figure S3, but for the second EOF mode. Additionally shown in (b) is the long-term mean pycnocline depth based on salinity profile.

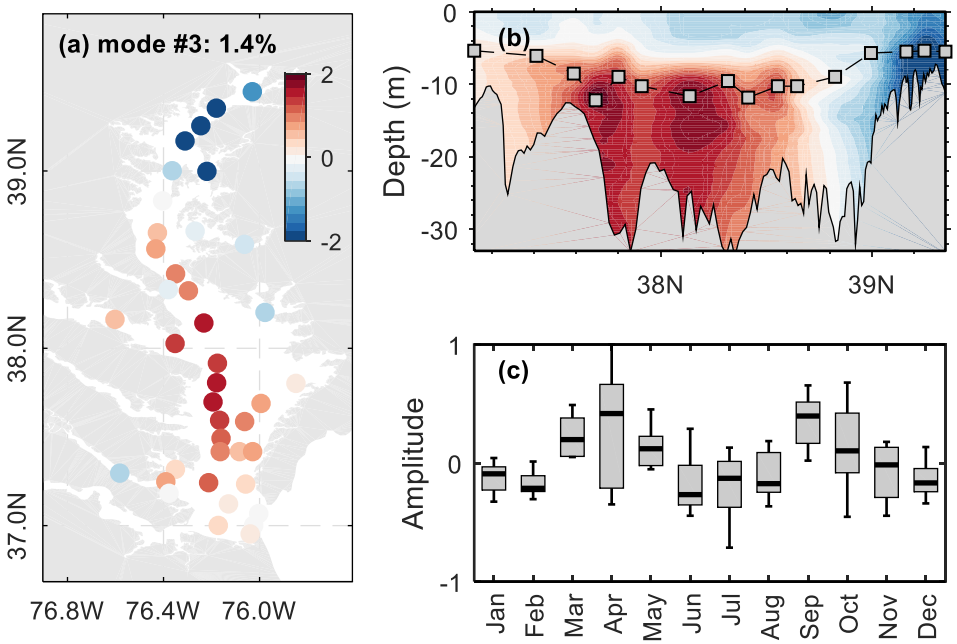

Figure S5: Same as Figure S3, but for the third EOF mode.

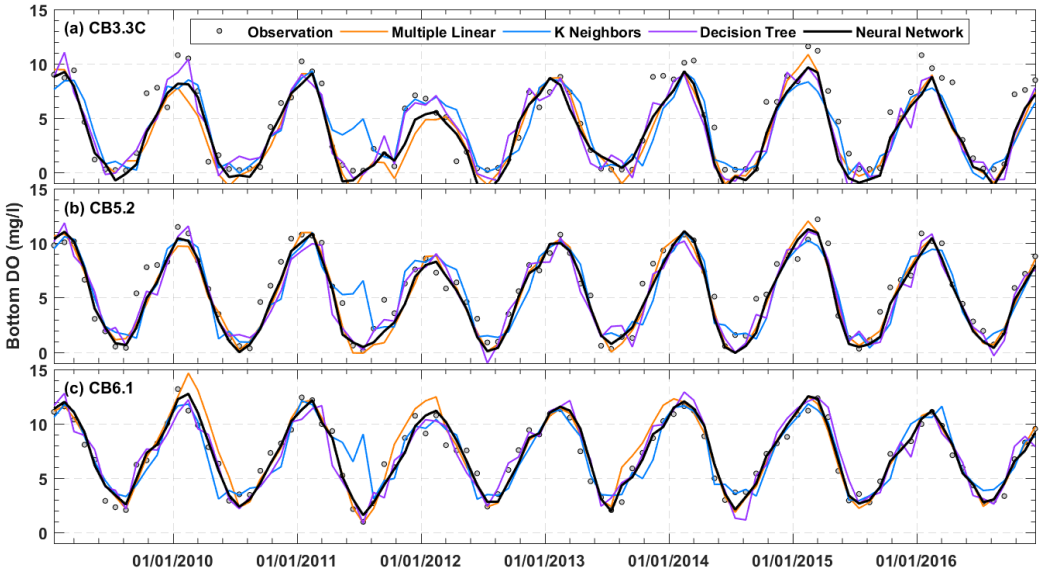

Figure S6: Observed bottom DO at three mainstem stations and the corresponding model results using different data methods. The Neural Network results are the ensemble mean over the 100 times of predictions using the 100 trained models.

Table S1: Selected forcing and transformation information for mode 01

Forcing	Transform function	Shifting (days)	Accumulation (days)
Air Temperature	1	0	15
Susquehanna flow	1	40	35
Downward short wave radiation flux	4	10	35

Table S2: Selected forcing and transformation information for mode 02

Forcing	Transform function	Shifting (days)	Accumulation (days)
Air temperature	1	50	85
Choptank TP loading	5	0	135
Choptank Sediment Loading	5	0	135
Susquehanna flow	1	20	35
Southerly wind intensity	4	10	15
Westerly wind speed	1	40	115
Potomac flow	4	30	85
Northerly wind intensity	3	30	15
Northerly wind hour (speed>4m/s)	1	30	85
James flow	6	30	135
Choptank TN loading	4	50	65
Susquehanna TN loading	2	10	135
Susquehanna TP loading	7	10	135
Easterly wind intensity	6	50	35
Easterly hour (speed>2m/s)	2	30	95
Westerly wind intensity	3	10	135
Northerly wind hour	4	10	15
Northerly wind hour (speed>2m/s)	8	20	15
Southerly wind hour (speed>4m/s)	4	60	55

Table S3: Selected forcing and transformation information for mode 03

Forcing	Transform function	Shifting (days)	Accumulation (days)
Easterly wind speed	1	60	65
Easterly wind hour (speed>2m/s)	3	0	45
James flow	1	40	105
Downward short wave radiation flux	3	10	15
Precipitation	6	0	15
Choptank TP loading	4	50	125
Westerly wind hour (speed>2m/s)	3	40	75
Westerly wind hour (speed>4m/s)	1	40	55
Westerly wind intensity	1	40	65
Susquehanna flow	6	0	15
Susquehanna TN loading	2	0	25
Susquehanna sediment loading	4	20	55
Easterly wind hour (speed>4m/s)	3	50	25
Choptank sediment loading	1	50	125
Susquehanna TP loading	1	0	55
Northerly wind speed	4	30	95
Northerly wind hour (speed>4m/s)	8	0	85
Northerly wind hour (speed>2m/s)	4	20	15
Southerly wind intensity	4	50	35
Southerly wind hour (speed>4m/s)	3	40	65
Southerly wind hour (speed>2m/s)	6	30	75
Southerly wind hour	1	10	85
Easterly wind hour	3	10	75
Westerly wind speed	8	50	55

Table S4: Selected forcing and transformation information for mode 04

Forcing	Transform function	Shifting (days)	Accumulation (days)
Susquehann sediment loading	3	0	75
Choptank TP loading	4	20	125
Susquehanna TP loading	1	0	45
Northerly wind hour	4	20	15
Northerly wind hour (speed>2m/s)	1	0	45
Wind speed	3	0	15
Susquehanna TN loading	7	0	25
Southerly wind hour (speed>2m/s)	4	30	25
Southerly wind hour	1	10	45
Downward short wave radiation flux	3	0	15
Westerly wind hour	6	60	115
Precipitation	1	40	15
Southerly wind hour (speed>4m/s)	1	10	45
Easterly wind speed	4	20	15
Choptank sediment loading	4	20	15
Choptank TN loading	8	50	25
Northerly wind hour (speed>4ms/s)	1	0	35
Northerly wind speed	4	0	25
James TP loading	7	0	135

Table S5: Selected forcing and transformation information for mode 05

Forcing	Transform function	Shifting (days)	Accumulation (days)
Susquehanna sediment loading	3	10	135
Southerly wind hour	1	60	135
Southerly wind speed	4	10	25
Westerly wind hour	3	40	15
Northerly wind hour	3	40	135
Northerly wind hour (speed>2m/s)	6	60	95
Northerly wind speed	4	10	15
Air temperature	4	0	15
Choptank TP loading	3	0	15
Wind speed	4	30	135
Precipitation	1	30	25
Potomac flow	4	20	55
James sediment loading	3	50	25
Westerly wind speed	4	40	95
Easterly wind hour (speed>4m/s)	1	50	135
Southerly wind intensity	3	30	105
Southerly wind hour (speed>4m/s)	6	50	45
Westerly wind hour (speed>4m/s)	6	30	95
Easterly wind speed	1	30	25
Easterly wind hour (speed>2m/s)	6	40	35
Westerly wind intensity	8	20	55
Choptank TN loading	4	50	15
Choptank sediment loading	1	50	95
Potomac TN loading	4	0	15

Table S6: Selected forcing and transformation information for hypoxic volume simulation

Forcing	Transform function	Shifting (days)	Accumulation (days)
Wind speed	1	0	15
Susquehanna river flow	1	60	55
Northerly wind hour	1	0	55
Westerly wind hour	1	0	15
Susquehanna TN loading	1	0	75
Southerly wind hour	1	60	95
Easterly wind hour	1	60	45

CHAPTER 2. A DATA-DRIVEN APPROACH TO SIMULATE THE SPATIOTEMPORAL VARIATIONS OF CHLOROPHYLL-A IN CHESAPEAKE BAY

Published in *Ocean Modelling* (2021, 159, 101748).

Abstract: Phytoplankton biomass, indicated by chlorophyll-a (Chl-a) concentration, is fundamentally important for aquatic ecosystems. Accurately simulating Chl-a is always challenging even when using state-of-the-art numerical models. We propose a data-driven modeling framework that combines Empirical Orthogonal Function (EOF) analysis and machine-learning technique to tackle this problem, using Chesapeake Bay as an example. Through the dimension reduction with EOF, the three-dimensional (3D) problem can be decomposed into multiple one-dimensional (1D) problems. The nonlinearity of these 1D problems will be modeled with machine learning using an artificial neural network. Model performance in terms of spatiotemporal Chl-a variations with both seasonal and interannual signals is evaluated. The model performance is comparable or higher than 3D numerical models previously applied in Chesapeake Bay. Sensitivity tests reveal the necessity of forcing transformations to improve the model predictive skill. Instead of manually applying a transformation for each input forcing variable, an auto-selection procedure is adopted to choose an appropriate transformation from a variety of transformation options. While it is unlikely the data-approach can replace the traditional numerical models, we argue that data-driven approaches provide a promising way for future studies in coastal and estuarine systems considering the fast accumulation of observational data.

Keywords: EOF; Neural network; water quality; chlorophyll-a simulation; estuaries

1. INTRODUCTION

Water quality in aquatic environments such as estuaries and coastal seas is of great public concern. Despite many efforts, it is still difficult to accurately predict water quality variables – e.g., dissolved oxygen and chlorophyll-a (Chl-a) concentration – because their spatiotemporal variations are not only subject to physical transport and mixing processes, but also significantly regulated by complex biogeochemical activities inside the water column and at the water–sediment interface (Beck, 1987; Arhonditsis and Brett, 2004; Fennel et al., 2006). Nonlinearities of biochemical processes are usually not easy to adequately describe using simple deterministic equations with a limited number of state variables. As a result, simulation and prediction of water quality variables have long been challenging even with the help of the most sophisticated numerical model systems. Extensive efforts have been made to identify the major controlling factors by examining the linear and nonlinear relationships between kinematic processes through analysis of *in situ* measurements or laboratory experiments (e.g., Kemp et al., 1997; Dauer et al., 2000). Such relationships are then parameterized into numerical model systems (e.g., HEM3D, Park et al., 1995; FVCOM, Chen et al., 2003; ROMS, Shchepetkin and McWilliams, 2005; and SCHISM, Zhang et al., 2016). These models have been successfully applied for water quality simulations in estuaries and coastal oceans (e.g., Fennel et al., 2006; Cerco and Noel, 2013; Testa et al., 2014; Yang et al., 2015). However, the accuracy of numerical model simulations greatly depends on the parameterization of the included kinetic processes, where large uncertainties always exist (van Straten, 1983; Shen, 2006; Jiang et al., 2018). Additionally, the accuracy of water

quality model simulation also suffers from the "cascading" or "propagation" of error originated in the hydrodynamic simulation (Beck, 1987). Errors generated during the hydrodynamic simulation will be passed to the water quality simulation, introducing additional uncertainties to the model accuracy in simulating water-quality state variables, because fundamental processes such as nutrient transport, algal growth, and oxygen distribution greatly depend on residence time, transport rate, and vertical mixing processes (Nixon et al., 1996; Lucas et al., 2009; Scully, 2010). To reduce the error, data assimilation has been widely used in forecast or reanalysis modeling (Ghil and Malaotte-rizzoli, 1991).

To manage the uncertainties associated with model structure, kinetic parameters, and error propagation, an alternative approach is to reduce error accumulation during the modeling process by relying, as much as possible, on a systematic observational dataset with reasonable temporal and spatial resolutions. Methods that rely purely on observational data are also referred to as "data-driven approaches" (Todorovski and Dzeroski, 2006; Shen et al., 2008; Yin et al., 2014; Yu et al., 2020). Data-driven approaches are not new and have been extensively used in industry (e.g., data mining and artificial intelligence) as well as in marine sciences (e.g., Anderson et al., 2010; Blauw et al., 2010; McGillicuddy, 2010; Wang and Tang, 2010; Kong et al., 2017). One data-driven approach is linear regression, perhaps the most widely used method in almost every research field. However, estuarine processes are complex and usually cannot be simply explained using linear relationships. To resolve nonlinear relationships, researchers often apply some transformations to independent variables (e.g., log and exponential transformation) and derive a variety of empirical formulas (Cohn et al., 1992;

Scardi and Harding, 1999; Attrill, 2002; Brush et al., 2002). Nevertheless, the improvement is limited for specific variables, such as Chl-a, which is known to be a function of various factors including nutrients, temperature, solar radiation, water clarity, and flushing rate (McCarthy et al., 1977; Harding et al., 1986; Cloern, 1999; Kemp et al., 2005). It is still challenging to simulate its spatiotemporal variations using a traditional empirical approach. In such cases, alternative methods are of great interest.

Availability of rapidly accumulated monitoring data including high-frequency *in situ* measurements, remote sensing, and reanalysis numerical modeling with data assimilation have paved the road for data-driven models. Recently, advanced data analysis methods have been developed quickly along with the increasing demand for big-data analysis. Combining these datasets and advanced methods may provide a new approach for predicting and understanding the variation of water quality conditions. With advances in monitoring techniques and the increasing availability of observational data, data-driven approaches are likely to have great future potential.

One example of advanced methods is neural network, which is a major component of our proposed data-driven approach. Neural networks are a widely used tool for empirical modeling in a complex system, especially useful when addressing nonlinear processes even if the underlying mechanisms are unknown or not fully understood (Scardi, 1996). Neural network models have been applied in the fields of classification, pattern recognition, and signal processing. Specifically, they have been used for remote sensing (e.g., Keiner and Yan, 1998; Vilas et al., 2011), water level prediction (Chang and Chen, 2003; Bajo and Umgiesser, 2010; Chen et al., 2012), rainfall-runoff processes (Hsu et al., 1995; Campolo et al., 1999), marsh classification (Morris et al., 2005), and

algal bloom prediction (Recknagel, 2001; Muttil and Chau, 2006). Neural network models were applied in Chesapeake Bay as early as 1996 by Scardi (1996) to train an empirical model for primary phytoplankton production. Following a similar approach, Scardi and Harding (1999) used *in situ* measurement of Chl-a, depth, light, and salinity conditions to predict the primary production rate in Chesapeake Bay. Muller and Muller (2015) used a wavelet-based neural network model to predict hypoxia volume in Chesapeake Bay based on the Oceanic Niño Index and river flow.

Most previous estuarine studies using data-driven models targeted a single time series regarding a bulk value or at a given location (e.g., Liang et al., 2015; Park et al., 2015; Kong et al., 2017), which is a one-dimensional (1D) simulation. Higher-dimensional modeling, however, is rarely reported in estuarine and coastal research. Theoretically, a higher-dimensional problem can be decomposed into a limited number of lower-dimensional problems, particularly for aquatic systems where materials are continuously exchanged vertically and horizontally. Constrained by estuarine circulations and regulated by the dilution process, water-quality variables (e.g., salinity, nutrient concentration, and dissolved oxygen) in an estuarine system typically share high covariance among different regions (Du et al., 2018). For a simple instance, salinity increases at the entrance of an estuary usually coinciding with an increase of salinity at the head of the estuary. It is therefore theoretically possible to decouple spatial patterns from temporal variations through a dimension-reduction analysis, such as empirical orthogonal functions (EOF), and to transform the three-dimensional (3D) problem into 1D or two-dimensional (2D) problems.

The purpose of this study is to examine the feasibility of a data-driven modeling approach by applying it to simulate spatial and temporal variations of Chl-a in Chesapeake Bay, the largest estuary in the US. Chl-a is one of the most fundamental state variables in determining the productivity and water quality of estuarine systems, and is also one of the most challenging subjects in water-quality modeling. Its spatial and temporal variations directly affect almost every aspect of biochemical processes for any given estuarine system. As one of the well-studied estuarine systems, Chesapeake Bay has been continuously monitored over several decades with monitoring stations covering a relatively large portion of the bay. Since 1985, monthly or bi-monthly surveys of major water-quality parameters including salinity, temperature, total nitrogen, dissolved oxygen, and Chl-a have been carried out by the Chesapeake Bay Program (CBP, https://www.chesapeakebay.net). Additional monitoring data including river flow, nutrient load, and air temperature are also made publicly available by the National Oceanic and Atmospheric Administration (NOAA) and the United States Geological Survey (USGS). These long-term and comprehensive datasets make Chesapeake Bay a perfect study site to evaluate a data-driven model.

The paper is organized as follows. Section 2 describes the collection of observational data, introduces the proposed data-driven model with focuses on its three major components. Section 3 presents the spatial and temporal pattern of observed Chl-a in Chesapeake Bay and shows the training and evaluation of the data-driven model. The necessity of including wind and forcing transformation, as well as the limitations and robustness of the data-driven model, is discussed in Section 4, followed by concluding remarks in Section 5.

2. METHODS

2.1 Data collection

Chesapeake Bay is a large, partially stratified estuary that extends about 320 km from the mouth of Susquehanna River to its entrance facing the Atlantic Ocean. Its water quality has been well monitored by the CBP. Thirty-five years (1985–2019) of the historical record of the target water-quality variable, specifically Chl-a, at the 16 mainstem stations was extracted from the CBP database (data available at https://www.chesapeakebay.net/what/downloads/cbp_water_quality_database_1984_present). The locations of the stations are shown in Fig. 1. Despite the spatial and temporal limitations on the sampling resolution, the available long-term monitoring dataset provided a reliable basis for analysis in many previous studies (e.g., Hagy et al., 2004; Kemp et al., 2005; Prasad et al., 2010; Murphy et al., 2011).

As an essential part of the data-driven model, external forcing data were carefully collected. Only relevant forcings were used, including nutrient loading, river flow, air temperature, and wind speed and direction (Table 1). River flow and wind are believed to regulate the stratification, estuarine circulation, and water exchange between ocean and estuary (Scully, 2010), while nutrient loading and air temperature are generally regarded as dominant factors controlling algal growth. Nevertheless, the combined influence from these forcings on Chl-a concentration can be complex (Harding, 1994). These forcing inputs are nearly the same as required by a 3D numerical model (e.g., Cerco and Noel, 2013). Forcing data were collected from a variety of reliable sources. River flow and nutrient loadings of the largest tributaries-Susquehanna, Potomac, James, and Choptank Rivers-were extracted from USGS (https://www.usgs.gov/). Air temperature at

Chesapeake Bay bridge-tunnel station (#8638901) was extracted from NOAA (https://tidesandcurrents.noaa.gov/), with data gaps being replaced with measurements at a nearby NOAA station, Cape Henry (#8638999). For the atmospheric data, instead of depending on measurements at a limited number of gauging stations, we used ERA5 reanalysis product provided by the European Centre for Medium-Range Weather Forecasts (ECMWF: https://www.ecmwf.int/), which cover the entire Chesapeake Bay with a spatial resolution of 0.25° and hourly temporal resolution.

2.2 Framework of the data model

To model not only the temporal but also the spatial variations, the Chl-a profile data with irregular vertical resolution was first converted into a gridded 3D dataset as a function of (x, z, t), where x is the horizontal location along the major axis of the bay, z is the depth, and t is time. Considering the high covariance among mainstem stations (i.e., water quality at one station usually changes in pace with nearby stations), an EOF analysis was conducted to decompose the spatial and temporal variations and thereby convert the 3D variable into several 2D maps and 1D time series (Fig. 2). The 1D time series could then be modeled with readily available statistical and machine-learning tools when provided proper input forcing variables.

The historical record of the target variable (i.e., Chla concentration) was first divided into two independent sub-datasets, namely training and testing datasets, with the former being used during the training process and the latter being used to evaluate the model performance. When training the model, it is important to keep the assumption that neither forcings nor the target variable in the testing period is known. Therefore, the training dataset instead of the full record was used for the EOF analysis.

Even though it is possible to train all EOF modes, it is essential to determine how many principal components are nontrivial and should be modeled. To distinguish interpretable signals from random noise, several methods have been frequently used, including the Kaiser–Guttman criterion, scree plot, broken-stick model, and total variance method (Jackson, 1993). Here we used the broken-stick model proposed by Frontier (1976), which suggests that components are interpretable when their eigenvalues exceed the corresponding value from a broken-stick distribution. The eigenvalue for the k^{th} component under the broken-stick model can be calculated as follows (Jackson, 1993):

$$b_k = \sum_{i=k}^{N} \frac{1}{i} \qquad (1)$$

where N is the total number of EOF components (e.g., 320 in this study). Only components with an eigenvalue larger than b_k will be selected.

For each selected EOF mode, the relationship between external forcings and temporal variations of each mode will be established with a neural network model. The following sections describe in detail the three major components of the proposed data-driven approach, i.e. EOF analysis, artificial neural network, and forcing transformation selection.

2.3 EOF analysis

EOF analysis is often used to study the principal components of a variable and how they change with time. An EOF uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Jolliffe, 2002). Depending on the covariance

between time series at different locations, the percentage of total variance accounted for by each mode varies.

Before performing EOF analysis, the observation profile data of Chl-a were averaged monthly and interpolated into 20 evenly distributed vertical layers at each station and the long-term mean value for each grid point was subtracted. In total, there were 320 sampling locations (16 stations × 20 layers) and 420 records over the period of 1985-2019. As a result, a data matrix **F** (420 × 320) was obtained, with each row representing a map for a given month and each column representing the time series of the variable for a given sampling location. Additionally, at each sampling location, the time series was normalized with its standard deviation.

The EOF analysis in this study was based on the singular value decomposition algorithm, which decomposes the normalized data matrix **F** into the following form:

$$\mathbf{F}=\mathbf{UDV}^{\mathrm{T}} \qquad (2)$$

where **U** is an orthogonal matrix (420 × 420) of temporal vectors, **V** is an orthogonal matrix (320× 320) of spatial vectors, and **D** is a diagonal matrix (420 × 320) storing the eigenvalues. The advantage of the EOF analysis is converting a 3D problem into multiple 1D problems. Instead of explaining the time series at all sampling locations, we only need to focus on a limited number of primary modes. The number of primary modes *p* to be modeled is determined using the broken-stick method (Eq. 1). Modes other than the first *p* modes are considered as noise. The noise-free Chl-a concentration (*C*) can then be calculated as follows:

$$C(x,t) = C_o(x) + \sum_{i=1}^{p} M_i(x) \times T_i(t) \qquad (3)$$

where C_o is the long-term mean, M is the map calculated as $\mathbf{V}\cdot\mathbf{D}^{\mathrm{T}}$ and is a function of the spatial location x, and T is the temporal value for a given mode as a function of time t. Both M and T are extracted from the EOF analysis. For prediction in the testing period, the predicted value of Chl-a (C) can be obtained similarly, with the maps kept unchanged but temporal components being replaced with predicted values ($\hat{T}$).

$$C(x,t) = C_o(x) + \sum_{i=1}^{p} M_i(x) \times T_i(t) \quad (4)$$

It can be very challenging to model the primary modes with simple linear regressions, as successes of regressions greatly depend on the linearity between forcing and responding variables. Considering many processes in nature are nonlinear and generally respond to a combination of multiple forcings, an artificial neural network was applied to model the temporal variations of the selected primary modes.

2.4 Artificial neural networks

Artificial neural networks are computational models inspired by the functioning of human brain (Scardi, 1996; Paliwal and Kumar, 2009). They are composed of numbers of "neurons", the basic computational unit that takes inputs (**x**) from other neurons or external sources, calculates the corresponding weight (**w**) for each input, sums the product of weights and input values (**Σwx**), plus bias (b), and finally passes this value (b+ **Σwx**) to an activation function. The outcome of the activation function is used as input for the next layer of neurons. Here the Levenberg-Marquardt backpropagation training function (Marquardt, 1963) was used, which approaches second-order training speed without computing the Hessian matrix directly and appears to be the fastest method for training moderate-sized feedforward neural networks (Hagan and Menhaj, 1999). The

Matlab Neural Network Toolbox (version 10.0) was used for this study. The training process will stop when any of the following conditions occurs: (1) number of epochs reach the defined maximum epochs (set to be 100); (2) cost function (mean square error) is minimized to the 0; and (3) performance gradient falls below 1e-7. For details of the algorithm of the Levenberg-Marquardt method, readers are referred to the help document in Matlab. The algorithm is widely recognized and well implemented in the Matlab toolbox.

It is worth noting that there are several uncertainty sources associated with the neural network and these uncertainties need to be considered when using the trained models for prediction. First, the neural network toolbox is set to randomly divide the input data (i.e., forcing data and target variable data for the training period) into "train-," "validate-," and "test-" parts. These three parts are set to account for 70%, 15%, and 15% of the input data, respectively. Second, weights and bias are commonly randomly initialized. Because of the randomness in data division and initialization of weights and bias, model predictions in both training and testing periods varies slightly from each training. To account for these uncertainties, we trained the neural network model 100 times for each principal component, used the ensemble mean of these models as the final prediction, and used the standard deviation of these models' predictions to quantify the uncertainties.

To improve model performance, multiple hidden layers of neurons are usually used. After testing the model by including different numbers of hidden layers, little change in model performance was found when more than two hidden layers were used. Therefore, two hidden layers were used in the model to obtain a balance of accuracy and

computational efficiency. The number of hidden neurons in each layer was automatically adjusted based on the number of input variables. The total numbers of neurons in the first and second layers were set as n and rounding of n/2, respectively, where n is the number of input variables.

2.5 Forcing transformation selections

One of the difficulties is selecting relevant forcings as input variables for the neural networks. Because each EOF mode usually relates to different underlying physical and biological processes or represents different regions, the response of each EOF mode to each input forcing varies. For example, a mode representing the upper estuary Chl-a variations will respond to flow and nutrient loading much faster than the mode that represents the lower estuary. A time-delay of response needs to be considered due to the transport processes of water and nutrients. Additionally, forcings sometimes affect estuarine dynamics through an accumulative manner. For instance, the growth of plankton in the bay is commonly believed to be linked to the loading over several prior months instead of a particular month (Scavia et al., 2006). To account for this effect, an accumulative moving average need to be applied for the forcings. In practice, we applied different lengths of accumulative average, including 30, 60, 90, 120, and 150 days, which are corresponding to different transport timescales for water moving from upstream to different regions of the bay (Shen and Wang, 2007).

Considering the unknown relationship between forcings and EOF modes, it is impractical to choose the transformations manually for each selected EOF mode. To this end, a process called “transformation auto-selection” is included in the method, in which forcings are transformed by multiple methods (e.g., accumulative average, time-delay,

and log transform, see Tables 2 and 3) and an appropriate transformation is selected automatically. The procedure included (1) transforming each input forcing variable with one of the functions listed in Table 2; (2) performing different moving averages for each transformed variable (Table 3); and (3) conducting different time-shifts for each transformed and moving averaged variable (Table 3). There are more than 200 combinations for each forcing, but only one type of transformation that improved and maximized model performance is selected. The target variable is set as **y** (n×1) and the input forcing matrix **X** is empty at the beginning. First, the coefficients of determination (R^2) between **y** and all available forcing variables in all available transformations were computed. The forcing variable with a transformation that gives the maximum R^2 is selected as the first input variable and stored in **X** (:, 1). The second forcing variable added to **X** is selected from the remaining forcings based on the R^2 from multiple linear regressions. The new forcing variable that most increased the R^2 is added. This process continued until R^2 does not increase by at least 0.005. This threshold is used to exclude forcings that made little contribution. Even though using linear regression analysis to determine the transformation is not the perfect option, such auto-selection gives an overall reasonable model performance (shown later). The selected forcing and transformation for each of the nine modes can be found in supplemental materials (Table S1–S9). More discussion on the necessity of transformation is presented in Section 4.3.

For each mode, input forcings are selected independently. There is no limit of the input forcings, but considering the availability of the data length and its relevance with Chl-a, we only chose those that are well known as influencing forcings according to previous numerical modeling studies (e.g., Cerco and Noel, 2013; Testa et al., 2014).

Several protocols need to be followed when choosing external forcings. First, no *in situ* measurements along the mainstem of the bay were used. It is not rational to model Chl-a if one can measure the *in situ* values of other environmental parameters such as salinity, water temperature, and total nitrogen. Instead, forcings that are regarded as "external" (e.g., wind, river discharge, and air temperature) were chosen (Table 1). Second, it is essential to ensure that the length of external forcings is equal or greater than the length of the target variable. Third, specific pre-processing may be needed for some input forcings. For example, wind hour was calculated to quantify the accumulative impact of wind blowing from a ±45° window of a given direction (e.g., easterly, westerly, northerly, and southerly) regardless of the wind speed. To account for wind speed, wind strength from each direction (unit, h m /s) was also computed as the product of wind hour and wind speed. In addition, we also calculated the wind hour when wind speed exceeded 2 or 4 m/s, considering that wind will be especially effective when its speed exceeds a certain value. Fourth, to be consistent with the temporal resolution of the target variable, all the forcing data were monthly averaged. For flow and nutrient loadings, not every river was used. Instead representative discharges and loadings from upstream discharge (Susquehanna River), large tributaries (e.g., Potomac and James Rivers), and small tributaries (e.g., Choptank River) were chosen.

2.6 Model performance index calculation

In addition to the common statistical measures, including root mean square error (RMSE) and correlation coefficient (R), model *skill* was also calculated following Willmott (1981):

$$Skill = 1 - \frac{\sum \left| X_{\mathrm{mod}} - X_{obs} \right|^2}{\sum \left(\left| X_{\mathrm{mod}} - \overline{X_{obs}} \right| + \left| X_{obs} - \overline{X_{obs}} \right| \right)^2} \qquad (5)$$

where X_{obs} and X_{mod} are the observed and modeled variables, respectively, with the overbar indicating the time average. *Skill* provides an index of model–data agreement, with a *skill* of 1 indicating perfect agreement and 0 indicating complete disagreement. The *skill* has been widely used to evaluate the performance of numerical models (e.g., Warner et al., 2005; Du et al., 2019). While R indicates the model's capability of capturing the seasonal or interannual trends and RMSE indicates the overall bias between model and observation, *skill* can be regarded as a synthesis index to evaluate both the trend capturing and relative bias.

Furthermore, the correlation coefficient and *skill* were also calculated for the anomaly to discount the influence of seasonality on the overall model performance. To obtain the anomalies, the seasonal cycle signal at each station and each layer based on the training dataset was subtracted for both the observed and predicted values in the testing period.

2.7 Sensitivity test

Wind plays an important role in water quality through modulating stratification, vertical mixing, longitudinal and lateral circulation, and water renewal in an estuary (Scully, 2010; Du and Shen, 2015). It is unknown whether wind will affect the intensity and extent of algal bloom. A sensitivity test with respect to the inclusion of wind force was conducted to answer this question by examining the difference between model performance with and without wind force.

Additional testing to illustrate the necessity of transformations was carried out by forcing the model with input variables that were not transformed. Even though input forcings were transformed by the activation function within the neural network, some important effects associated with estuarine dynamics such as time lagging and accumulative effect are not well included inside the neural network. Results from the sensitivity tests were compared to the base run (i.e., with wind and transformations) and their performance differences were illustrated by Taylor diagram (Taylor, 2001).

3. RESULTS

3.1 Long-term mean and EOF modes for Chl-a

It is worthwhile to describe the long-term mean because it provides the background condition upon which the distribution of Chl-a concentration varies spatially. The long-term mean Chl-a concentration ranged within 5–15 μg/l, with a higher value in the upper bay and near the surface layers (Fig. 3). The spatial pattern shares great similarities with the distribution of nitrogen concentration (Du and Shen, 2017), demonstrating the potential influence of nutrient concentration. Nitrogen limitation is evident in Chesapeake Bay, particularly during the summer, and the impact of nitrogen limitation increases from the upper bay to the lower bay (Fisher et al., 1992; Fisher et al., 1999). The vertical gradient (i.e., larger value at the surface and smaller value at the bottom) suggests a light limitation for algal growth at all locations. The maximum value occurred at 39°N, or CB3.3C, while the minimum value was observed at the bay mouth and the northernmost end of the bay. The maximum value occurred just below the downstream limit of the estuarine turbidity maximum zone, which is typically located

above 39.13N (Malpezzi et al., 2013). The location of the maximum Chl-a concentration suggests that the algal growth was likely regulated by both nutrients and light conditions. At 39N where nutrient levels were high and light conditions were sufficient, the accumulation of phytoplankton biomass was greatly favored (Keller et al., 2014). Roman et al. (2005), through a high-resolution sampling over 1995–2002, also found a persistent maximum in phytoplankton biomass occurring in the upper bay and attributed it to the physical and topographic discontinuities in this region.

3.2 Spatial and seasonal pattern of primary modes

More interesting is the spatial pattern and temporal variations of the primary EOF modes. The first four primary modes accounte for 45%, 17%, 8%, and 6% of the total variance, respectively. Due to the orthogonal nature, spatial patterns (hereafter referred to as maps) and time series differed from one another (Fig. 4). The first and second modes feature the spring algal bloom, but for different regions. The third and fourth modes seem to highlight the summer bloom. It is not the primary purpose of this study to distinguish the possible processes contributing to each mode. However, through an EOF analysis, one could identify the likely dominant processes accounting for the spatial and temporal variations, which will be further discussed in Section 4.1.

3.3 Determine the number of modes to explain

Using the broken-stick method, the number of interpretable components for Chl-a in Chesapeake Bay was determined to be nine, based on the fact that the first nine EOF modes have eigenvalues greater than that under the broken-stick distribution (Fig. 5). The residual values after subtracting the summation of the first nine modes and long-term mean from the original values are considered as noise, and they are not correlated among

different locations. The first nine EOF modes accounted for 88% of the total variance. Note that the number of primary components to be modeled is determined based on the training dataset and it is assumed this number also applied to the testing period.

3.4 Model training

The long-term dataset was separated into two parts: a training dataset (the first 80% of the dataset, i.e., 1985–2013) and a testing dataset (the last 20% of the dataset, i.e., 2014–2019). The training dataset was used to train the model, while the testing dataset was used to evaluate the prediction skill of the trained model. Using the neural network, each of the primary modes was well trained by various input forcings, with model *skill* exceeding 0.52 for all the modes (Fig. 6). We acknowledge that the model is not perfect for every mode as there are still variations not explained by the input forcings, for several following reasons. First, some nutrient sources were not included in the input forcings, such as the atmospheric deposition, point sources along the shoreline, and coastal ocean input. These sources could contribute to 30-40% of the total nitrogen load (Boynton et al., 1995). However, there were limited available data regarding these nutrient sources and therefore they were not included in the model. Another reason is the temporal limitation of the observation. The monthly or bi-monthly measurement might not be able to represent the monthly mean condition, particularly for regions with large temporal variabilities (Bever et al., 2013).

3.5 Model performance in testing period

The model predicts well the variability in different regions, i.e., larger variability in the upper bay and smaller variability in the lower bay (Fig. 7). For the subsurface Chl-a, the RMSE range is 7– 12 μg/l, with a larger RMSE in the upper bay. Correlation

coefficient (~ 0.8) and model *skill* (~ 0.8) are both high in the middle and upper bay stations, with exception of the upper-most station CB2.2. The model *skill* at the lower bay stations is relatively lower (0.14–0.72), partly because Chl-a in the lower bay is also controlled by coastal ocean conditions (e.g., nutrient concentration), which were not included in the model due to lack of observations. Recent studies (e.g., Du and Shen, 2017; Da et al., 2018) suggest the nutrient conditions in the open ocean can be influential, especially for the lower bay, as large bottom inflow can efficiently move the oceanic water into the bay. As nutrient concentration in the lower bay is relatively small, changes in nutrient level in the coastal ocean can thus have more impact in this region than the middle-upper bay, where riverine nutrients dominate. The lower performance in the lower bay can be also attributed to the decreasing variability of Chl-a when moving from the upper bay to the lower bay. For example, at the bay mouth station CB7.3, the concentration of Chl-a was typically below 10 μg/l, which was about one order less than the value in the middle and upper bay.

The seasonal patterns of Chl-a were also well-predicted and the pattern varied greatly among different seasons in terms of the magnitude and vertical gradient. Chl-a was much greater during winter and spring compared to summer and fall (Fig. 8). Interestingly, the vertical gradient followed an upward direction (i.e., decrease upward) during winter, weaker in spring, and downward in summer and fall. The upward gradient in winter and spring is unique in Chesapeake Bay, particularly during winter, whereas the vertical gradient during spring is weak except in the upper bay (>39°N). The exact underlying mechanism for such vertical distribution during winter is not well- known. It

could be caused by the combination of several processes, including settling of algae, landward bottom inflow, and/or lateral circulations.

Not only does the model predict well the seasonal cycle of Chl-a, it also reasonably predicts the interannual variability (Fig. 9). Of the five years (2014-2018) in the testing period, the observed spring Chl-a was characterized with a maximum value in the bottom of the upper bay for three years (2014–2016), suggesting effective bottom trapping near the turbidity maximum zone. Chl-a in the spring of 2016–2017 was much smaller and showed no marked bloom, likely caused by a lower nutrient level or a more dispersive hydrodynamic condition. There were significant interannual variabilities and the model is able to capture this interannual signal.

Overall, the model predictions are satisfactory for both spatial and temporal variations of Chl-a in Chesapeake Bay. The performance is comparable or better than the sophisticated 3D numerical models that have been applied for Chesapeake Bay (e.g., Cerco and Meyers, 2000; Li et al., 2009; Cerco and Noel, 2013; Testa et al., 2014; Yang et al., 2015). Using a physical–biogeochemical coupling model, Testa et al. (2014) simulated Chl-a with a mean correlation coefficient of 0.6 between model and observation. In another numerical modeling study (Feng et al., 2015), the correlation coefficient for Chl-a had a mean value of 0.3 when averaged over the mainstem monitoring stations. Irby et al. (2016) compared the performance of eight numerical models (all applied for Chesapeake Bay) in simulating salinity, temperature, dissolved oxygen, and Chl-a. They showed that all the numerical models had a low *skill* in predicting Chl-a compared to other water-quality variables. For the bottom Chl-a, the correlation coefficient from different models ranged within 0.1–0.6. In comparison, our

data-driven model had a correlation coefficient of about 0.75 when averaged over all 16 stations (Fig. 10). Furthermore, it seems all the numerical models greatly underestimate the variability of the bottom Chl-a, with the standard deviation about less than half of the true value (Testa et al. 2014; Feng et al., 2015; Irby et al. 2016). Even though most of these numerical models were designed for simulating the dissolved oxygen, there is no doubt that Chl-a simulation is extremely challenging even using the most advanced numerical model systems.

4. DISCUSSION

4.1 Possible mechanisms revealed by EOF analysis

It is of interest to understand the dominant processes that regulate the temporal and spatial distribution of a given water quality variable, and EOF analysis can serve as a useful tool for such a purpose. This strategy has been extensively used in many studies. For example, Scully (2016) performed an EOF analysis for the dissolved oxygen in Chesapeake Bay and attributed the second mode to the bathymetry discontinuity-induced convergence. Du et al. (2018) used EOF to identify the key external parameters including water temperature and Chl-a concentration for the first primary mode of dissolved oxygen in Chesapeake Bay.

Our analysis shows that the first and second modes, together, contributed to 64% of the total variance of Chl-a. The temporal values of both first and second EOF modes show clear seasonality, with peaks in April and March, respectively (Fig. 4b, d). The first mode is characterized with positive spatial value throughout the entire bay, meaning changes in Chl-a in this mode are in phase over the entire bay. Differently, the second

mode highlights a strong variability in the subsurface water of the upper bay. It is conceivable that the first and second modes accounted for the major seasonal bloom, particularly the spring bloom. The second mode leads the first one by about one month, suggesting that the spring bloom occurrs first in the upper bay with maximum bloom in March and then the algal bloom extends to the entire bay, reaching its maximum in April. Although the spring bloom has been commonly recognized in previous studies (e.g. Harding 1994; Keller et al., 2014), it is interesting to find the different timing of algal bloom in different regions.

It is worth pointing out the noticeable vertical difference in the spatial pattern of the first two EOF modes, both of which were characterized with a higher value at the bottom and a smaller value at the surface, indicating a larger variability (or more sensitive response) at the bottom than at the surface. Particularly, in the second mode, Chl-a concentration varies greatly at the bottom of the upper bay, compared to any other region. Such a spatial pattern is partially attributable to the sensitivity of sediment resuspension and accumulation at the water–sediment interface near the turbidity maximum zone. Keller et al. (2014) suggested that an efficient entrapment of phytoplankton and phytoplankton-derived organic matter occurs near the turbidity maximum zone. Contrary to our common understanding that algae generally concentrate in the upper column due to light attenuation, it is the subsurface layer that had a larger variability and this applied to the entire bay. Even after including the long-term mean value, the subsurface concentrations of Chl-a were generally larger than the surface concentration (Fig. 9).

4.2 Necessity to include wind

The sensitivity test without including the wind forcings shows that the model's performance is weakened. Without wind forcings, the overall performance is reduced for both the training and testing (Figs. 11and 12). Surface stress posed by the wind field can alter not only barotropic but also baroclinic processes. It affects Chl-a through several processes. First, wind could change the estuarine circulations, including both longitudinal and lateral circulations (Chen and Sanford, 2009; Scully, 2010; Li and Li, 2011). Change of circulation affects the along-bay and cross-bay water- and nutrient-exchange. Dispersion of the phytoplankton patches follows the water movement and circulations. A stronger downstream flow moves the surface water faster toward downstream and enhances the compensated bottom inflow, resulting in a smaller residence time and thus faster flushing (Shen and Wang, 2007; Du and Shen, 2016). A smaller residence time can inhibit algal blooms (Lucas et al., 2009; Qin and Shen, 2019). Second, wind forcing introduces external energy for vertical mixing. Stronger wind tends to enlarge the vertical mixing layer, whose thickness is known to affect the accumulation rate of biomass due to light attenuation and plankton biomass loss (e.g., due to predation or respiration) (Roman et al., 2005). Third, wind forcing can replenish the upper water column with the supply of nutrients from bottom layers after destratification (Miller et al., 2006). Such nutrient sources can be very important during summer, during which euphoric water is usually nutrient-limited after spring blooms (Fisher et al., 1992). Finally, wind-induced resuspension of plankton previously settled at the water–sediment interface can also directly affect Chl-a concentration in the water column. Schelske et al. (1995), based on

observations in Lake Apopka, Florida, showed that Chl-a concentrations >100 μg/l were highly correlated with wind speed primarily due to the resuspension of meroplankton.

Linear relationships between wind forcing and key hydrodynamic processes have been identified in previous studies (e.g., Scully, 2010). In practice, researchers usually decompose the wind into different directions. One of the most commonly used strategies is to respectively calculate wind hour, wind speed, and wind strength for southerly, northerly, westerly, and easterly components. This strategy was adopted in the wind data preprocessing for our Chesapeake Bay Chl-a application.

4.3 Necessity to apply transformations for the input forcings

Transformations are usually needed to obtain a better model performance, which applies to not only wind forcing but also other input forcings (Shen et al., 2019). Scardi et al. (1999) showed that their neural network-based model has performance enhancement when applying a log transformation to the input variables. However, performance enhancement does not necessarily occur for every case of applying a log transformation. For example, Maher and Eyre (2011) showed that a log transformation does not improve their model performance.

A sensitivity run without applying any transformation of the input forcings gave the poorest model performance, compared to the base run and sensitivity run without wind (Figs. 11and12). One noticeable change in the model performance is that the standard deviation is much larger than the other two runs (Fig. 11f).

Several processes warrant the necessity of including the transformation. First, it is known that nutrients from the primary diffusive source (i.e., Susquehanna River) take

months to be transported to the middle and lower portions of the bay (Shen and Wang, 2007). Even for the source from major ambient tributaries (e.g., Potomac, James, and Choptank Rivers), it also takes tens of days for nutrients released from the headwaters to merge into the mainstem bay. It is therefore essential to apply a time-shifting transformation for nutrient and freshwater loading. Not only the nutrients and freshwater loading, but any forcings influencing the transport processes might also need a time-shifting transformation. For example, a northeasterly wind in the winter is related to the summer hypoxia, and the underlying mechanism is believed to lie in the accumulation of organic matter in the lower bay due to the enhanced estuarine circulation under a northeasterly wind (Lee et al., 2013). Second, many input forcings change the physical–biological condition through an accumulation process. Multiple-month accumulative amounts of nutrients or freshwater load are generally considered to be more related with the spring algal bloom and summer hypoxic volume in Chesapeake Bay (Scavia et al., 2006).

However, it will be challenging and not practically sound to manually set transforming for each forcing and each mode, as there are too many forcings to be included and many modes to be explained. The effectiveness of any transformation depends on the input forcings and the relationship between the dependent and independent variables. Providing a series of transformation options and making the model select the best relevant one seems to be a good way to circumvent such obstacles. Therefore, we used an auto-selection procedure for the transformation step.

4.4 Robustness and limitation of the data-driven approach

The accuracy of the prediction and the success of the model rely greatly on the record length of both the forcing data and the measurement of target parameters. High integrity of the data in Chesapeake Bay is one of the reasons for the success of the model application. As the monitoring program continues, the length of records will increase steadily in the future, which will lead to more accurate model predictions. The increasing data volume in the future will make the data-driven approach more promising.

Nevertheless, it does not mean this approach is not suitable for a coastal system with measurements of limited duration. The limitation of time span can be compensated by a higher measurement frequency. Daily, hourly, or even minute-by-minute data can now be easily obtained through *in situ* measurements, remote sensing, or hindcast numerical results. For example, Chl-a concentration data from remote sensing (e.g., MODIS and VIIRS) is available daily. Even considering the cloud-induced lousy signal, a weekly average will give reasonable coverages. For each year, it is possible to get 52 weekly measurements per year, leading to data accumulation at a much faster pace and therefore a shorter time-span requirement when applying a data-driven approach. In coastal waters, the seasonal cycle usually plays an important role in physical and biological dynamics and therefore it is recommended to have data with length at least longer than a year.

It is noteworthy that the model performance is likely hampered by the temporal resolution of Chl-a measurements in Chesapeake Bay as well as the missing forcing data regarding the point source loading, coastal ocean input, and atmospheric deposition. Such types of limitations not only exist in our model but also for other 3D numerical models.

The limited temporal resolution of Chl-a *in situ* measurement may cause the proposed model to have difficulty in predicting the anomaly (i.e., the residual after subtracting the seasonal mean) of Chl-a in Chesapeake Bay. For the nine stations shown in Fig. 7, the anomaly correlation coefficient ranges within 0.2–0.6. The relatively low anomaly correlations do not necessarily indicate that the model is useless considering that Chl-a is extremely difficult to model in estuarine and coastal waters because Chl-a is affected by not only the fluid dynamics but also complex biogeochemical processes (Testa et al., 2014), for which uncertainties are relatively large.

Using similar inputs including river discharge, nutrient loading, and atmospheric forcing as required for a hydrodynamic-ecosystem model, the data-based model has a much better computational efficiency. The neural network itself does not consume too much time, but the filtering of the input forcings to obtain a set of relevant forcings and appropriate transformations does. The time consumed depends greatly on the number of input forcings and the number of transformations, as well as the length of records. The influence of the input forcings is usually nonlinear, and thus numerous transformations are recommended, including exponential, log, square, inverse, and time-shifting transformations (Table 2).

Perhaps the most crucial advantage of the data-driven approach is that it has less error accumulation. For numerical model systems, the success of the water-quality model greatly depends on the reliability of hydrodynamic model and the parameterization of various processes to obtain a numerical solution with discrete resolution in space and time. Inevitably, these errors tend to accumulate over time and from the hydrodynamic level to the upper ecosystem level. The data-driven approach proposed in this study could

be an alternative and efficient tool circumventing this type of error. The same concept can also be applied to some similar fields, such as dissolved oxygen, dissolved organic matter, and turbidity, considering the high similarities in their controlling mechanisms in an estuary.

5. CONCLUSIONS

This study introduced a data-driven approach for water-quality modeling by combining EOF analysis and neural network and converting the 3D problem into multiple 1D problems. The overall performance of this data-driven approach in modeling the spatial-temporal variations of Chl-a in Chesapeake Bay is comparable or even better than sophisticated 3D numerical models, all of which have difficulty in accurately modeling the variations of Chl-a. The approach could be useful as a tool to predict the response of an estuarine ecosystem to changes in environmental conditions.

We are not suggesting that the data-approach is better than the numerical models; however, we argue that data-driven approaches provide a promising way for future studies in coastal and estuarine systems considering the fast accumulation of data from a variety of monitoring programs. This study demonstrates that the dynamics of Chl-a, which is highly variable temporally and spatially and extremely challenging to simulate, can be reasonably predicted by a data-driven approach. Because of the nonlinear nature of ecosystem processes, modeling of water quality is always challenging, even with the help of sophisticated 3D numerical model systems. The approach presented here provides an alternative way to circumvent the error cascading in a numerical model and uncertainties induced by the not well-parameterized biogeochemical processes. By

combining EOF analysis and the neural network, this study highlights the potential of developing data-driven models for complex 3D problems in estuaries and coasts. With high-frequency monitoring data and rapidly advancing machine-learning techniques, the data-driven approaches are likely to be more frequently used in future estuarine studies.

ACKNOWLEDGMENTS

This work is supported by the Virginia Institute of Marine Science through a student scholarship. We sincerely thank Mac Sisson for proofreading. We thank anonymous reviewers for their constructive suggestions. We thank Jiabi Du for his help in the model development. This is contribution No. 3981 of the Virginia Institute of Marine Science, College of William and Mary.

REFERENCES

Anderson, D.M., Glibert, P.M., Burkholder, J.M. (2002). Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries, 25,* 704–726.

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. *Journal of Marine System, 83,* 127-140.

Arhonditsis, G.B., Brett, M.T. (2004). Evaluation of the current state of mechanistic aquatic biogeochemical modeling. *Marine Ecology Progress Series, 271,* 13–26.

Attrill, M.J. (2002). A testable linear model for diversity trends in estuaries. *Journal of Animal Ecology, 71,* 262–269.

Baird, D., Ulanowicz, R.E., Boynton, W.R. (1995). Seasonal nitrogen dynamics in Chesapeake Bay: a network approach. *Estuarine, Coastal and Shelf Science, 41,* 137–162.

Bajo, M., Umgiesser, G. (2010). Storm surge forecast through a combination of dynamic and neural network models. *Ocean Modelling, 33,* 1–9.

Beck, B. (1987). Water quality modeling: a review of the analysis of uncertainty. *Water Resource Research, 23,* 1393–1442.

Bever, A.J., Friedrichs, M.A.M., Friedrichs, C.T., Scully, M.E., Lanerolle, L.W.J. (2013). Combining observations and numerical model results to improve estimates of hypoxic volume within the Chesapeake Bay, USA. *Journal of Geophysical Research: Oceans, 118,* 1–21.

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

Boynton, W.R., Garber, J.H., Summers, R., Kemp, W.M. (1995). Inputs, transformations, and transport of nitrogen and phosphorus in Chesapeake Bay and selected tributaries. *Estuaries* 18, 285–314.

Brush, M.J., Brawley, J.W., Nixon, S.W., Kremer, J.N. (2002). Modeling phytoplankton production : Problems with the Eppley curve and an empirical alternative. *Marine Ecology Progress Series, 238,* 31–45.

Campolo, M., Andreussi, P., Soldati, A. (1999). River flood forecasting with a neural network model. *Water Resource Research, 35,* 1191–1197.

Cerco, F.C., Noel, M.R. (2013). Twenty-one-year simulation of Chesapeake Bay water quality using the CE-QUAL-ICM eutrophication model. *JAWRA Journal of the American Water Resources Association, 49,* 1119-1133

Cerco, C. F., Meyers, M. (2000). Tributary refinement to Chesapeake Bay model. *Journal of Environmental Engineering, 126,* 164–174.

Chang, F.J., Chen, Y.C. (2003). Estuary water-stage forecasting by using radial basis function neural network. *Journal of Hydrology, 270,* 158–166.

Chen, S.-N., Sanford, L.P. (2009). Axial wind effects on stratification and longitudinal salt transport in an idealized, partially mixed estuary. *Journal of Physical Oceanography, 39,* 1905–1920.

Chen, C., Liu, H., Beardsley, R.C. (2003). An unstructured grid, finite-volume, three-dimensional, primitive equations ocean model: Application to coastal ocean and estuaries. *Journal of Atmospheric and Oceanic Technology*, 20, 159–186.

Chen, W.B., Liu, W.C., Hsu, M.H. (2012). Comparison of ANN approach with 2D and 3D hydrodynamic models for simulating estuary water stage. *Advances in Engineering Software*, *45*, 69–79.

Cloern, J.E. (1999). The relative importance of light and nutrient limitation of phytoplankton growth : A simple index of coastal ecosystem sensitivity to nutrient enrichment. *Aquatic Ecololgy*, *33*, 3–16.

Cohn, A., Caulder, D.L., Gilroy, J., Zynjuk, L.D., Summers, R.M. (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay. *Water Resource Research*, *28*, 2353–2363.

Dauer, D.M., Weisberg, S.B., Ranasinghe, J.A. (2000). Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in Chesapeake Bay. *Estuaries*, *23*, 80-96.

Du, J., Shen, J. (2015). Decoupling the influence of biological and physical processes on the dissolved oxygen in the Chesapeake Bay. *Journal of Geophysical Research: Ocean*, 120, 78–93.

Du, J., Shen, J. (2016). Water residence time in Chesapeake Bay for 1980 - 2012. *Journal of Marine Systems*, *164*, 101–111.

Du, J., Shen, J. (2017). Transport of riverine material from multiple rivers in the Chesapeake Bay: important control of estuarine circulation on the material distribution. *J. Geophysical Research: Biogeosciences, 122,* 2998–3013.

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.

Du, J., Park, K., Shen, J., Zhang, Y.J., Yu, X., Ye, F., Wang, Z., Rabalais, N.N. (2019). A hydrodynamic model for Galveston Bay and the shelf in the northern Gulf of Mexico. *Ocean Science, 15,* 951-966.

Feng, Y., Friedrichs, M.A.M., Wilkin, J., Tian, H., Yang, Q., Hofmann, E.E., Wiggert, J.D, Hood, R.R. (2015). Chesapeake Bay nitrogen fluxes derived from a land-estuarine ocean biogeochemical modeling system: Model description, evaluation, and nitrogen budgets. *Journal of Geophysical Research: Biogeosciences, 120,* 1666–1695.

Fennel, K., Wilkin, J., Levin, J., Moisan, J., O'Reilly, J., Haidvogel, D. (2006). Nitrogen cycling in the Middle Atlantic Bight: Results from a three-dimensional model and implications for the North Atlantic nitrogen budget. *Global Biogeochemical Cycles,* 20, 1–14.

Fisher, T.R., Peele, E.R., Ammerman, J.W., Harding, L.W. (1992). Nutrient limitation of phytoplankton in Chesapeake Bay. *Marine Ecology Progress Series, 82,* 51–63.

Fisher, T.R., Harding, L.W., Stanley, D.W., Ward, L.G. (1988). Phytoplankton, nutrients, and turbidity in the Chesapeake, Delaware, and Hudson estuaries. *Estuarine, Coastal and Shelf Science, 27,* 61–93.

Fisher, T.R., Peele, E.R., Ammerman, J.W., Harding, L.W. (1992). Nutrient limitation of phytoplankton in Chesapeake Bay. *Marine Ecology Progress Series, 82,* 51–63.

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modd́le du bâton brisé. *Journal of Experimental Marine Biology and Ecology, 25,* 67-75.

Ghil, M., Malanotte-rizzoli, P. (1991). Data assimilation in meteorology and oceanography. *Advances in Geophysics, 33,* 141-266

González Vilas, L., Spyrakos, E., Torres Palenzuela, J.M. (2011). Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sensing of Environment,* 115, 524–535.

Hagan, M.T., Menhaj, M. (1999). Training feed-forward networks with the Marquardt algorithm. *IEEE Transation on Neural Network,* 5, 989–993.

Hagy, J.D., Boynton, W.R., Keefe, C.W., Wood, K.V. (2004). Hypoxia in Chesapeake Bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. *Estuaries, 27,* 634–658.

Harding, L.W., Meeson, B.W., Fisher, T.R. (1986). Phytoplankton production in two East Coast estuaries: Functions and patterns of carbon assimilation in Chesapeake and Delaware Bays. *Esturarine Coastal and Shelf Science, 23,* 773–806.

Harding, L.W. (1994). Long-term trends in the distribution of phytoplankton in Chesapeake Bay: roles of light, nutrients and streamflow. *Marine Ecology Progress Series, 104,* 267–291.

Hsu, K., Gupta, H.V., Sorooshian, S., (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resource Research, 31,* 2517.

Irby, I. D., Friedrichs, M. A. M., Friedrichs, C. T., Bever, A. J., Hood, R. R., Lanerolle, W. J., Li, M., Linker, L., Scully, M. E., Sellner, K., Shen, J., Testa, J., Wang, H., Wang, P., & Xia, M. (2016). Challenges associated with modeling low-oxygen waters in Chesapeake Bay : a multiple model comparison. *Biogeosciences*, *13*, 2011–2028.

Jackson, D.A. (1993). Stopping rules in principal components analysis : A comparison of heuristical and statistical approaches. *Ecology*, 74, 2204–2214.

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

Jiang, L., Li, Y., Zhao, X., Tillotson, M.R., Wang, W., Zhang, S., Sarpong, L., Asamaa, Q., Pan, B. (2018). Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China. *Ecological Modelling*, *375*, 1–12.

Jolliffe, I.T. (2002). Principal Component Analysis, Second Edition. Springer, pp. 1.

Keiner, L.E., Yan, X.H. (1998). A neural network model for estimating sea surface chlorophyll and sediments from thematic mapper imagery. *Remote Sensing of Environment*, *66*, 153–165.

Keller, D.P., Lee, D.Y., Hood, R.R. (2014). Turbidity maximum entrapment of phytoplankton in the Chesapeake Bay. Estuaries and Coasts 37, 279–298.

Kemp, W.M., Boynton, W.R., Adolf, J.E., Boesch, D.F., Boicourt, W.C., Brush, G. (2005). Eutrophication of Chesapeake Bay: Historical trends and ecological interactions. *Marine Ecology Progress Series*, *303*, 1–29.

Kemp, W.M., Smith, E.M., Marvin-DiPasquale, M., Boynton, W.R. (1997). Organic carbon balance and net ecosystem metabolism in Chesapeake Bay. *Marine Ecology Progress Series, 150*, 229–248.

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.

Lacouture, R., Haas, L.W., Wetzel, R.L. (1999). Spatial and temporal variation of resource limitation in Chesapeake Bay. *Marine Biology, 133*, 763–778.

Lee, C., Hu, C., Cannizzaro, J., Duan, H. (2013). Long-term distribution patterns of remotely sensed water quality parameters in Chesapeake Bay. *Estuarine, Coastal and Shelf Science, 128*, 93–103.

Li, Y., Li, M. (2011). Effects of winds on stratification and circulation in a partially mixed estuary. *Journal of Geophysical Research, 116*, C12012.

Li, M., Zhong, L., Harding, L.W. (2009). Sensitivity of plankton biomass and productivity variations in physical forcing and biological parameters in Chesapeake Bay. *Journal of Marine Research, 67*, 667-700.

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 relationships observed between phytoplankton biomass and transport time*? Limnology and Oceanography, 54*, 381–390.

Maher, D., Eyre, B.D. (2011). Benthic carbon metabolism in southeast Australian estuaries: Habitat importance, driving forces, and application of artificial neural network models. *Marine Ecology Progress Series*, 439, 97-115.

McGillicuddy Jr, D.J. (2010). Models of harmful algal blooms: conceptual, empirical, and numerical approaches. *Journal of Marine Systems*, *83*, 105–107.

Malpezzi, M.A., Sanford, L.P., Crump, B.C. (2013). Abundance and distribution of transparent exopolymer particles in the estuarine turbidity maximum of Chesapeake Bay. *Marine Ecology Progress Series*, *486*, 23–35.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, *11*, 431–441.

Mccarthy, J.J., Rowland, W., Taft, J.L. (1977). Nitrogenous nutrition of the plankton in the Chesapeake Bay. 1. Nutrient availability and phytoplankton preferences. *Limnology and Oceanography*, *22*, 996–2011.

Miller, W.D., Harding, L.W., Adolf, J.E. (2006). Hurricane Isabel generated an unusual fall bloom in Chesapeake Bay. *Geophysical Research Letters*, *33*, 2–5.

Morris, J.T., Porter, D., Neet, M., Noble, P.A., Schmidt, L., Lapine, L.A., Jensen, J.R. (2005). Integrating LIDAR elevation data, multi-spectral imagery and neural network modelling for marsh characterization. *International Journal of Remote Sensing*, *26*, 5221–5234.

Muller, A., Muller, D.L. (2015). Forecasting future estuarine hypoxia using a wavelet based neural network model. *Ocean Modelling*, 96, 314-323.

Murphy, R.R., Kemp, W.M., Ball, W.P. (2011). Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. *Estuaries and Coasts, 34*, 1293–1309.

Muttil, N., Chau, K.W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, *28*, 223.

Nixon, S.W., Ammerman, J.W., Atkinson, L.P., Berounsky, V.M., Billien G., Boicourt, W.C., Boynton, W.R., Church, T.M., Ditoro, D.M., Elmgren, R., Garber, J.H., Giblin, A.E., Jahnke, R.A., Owens, N.J.P., Pilson, M.E.Q., Seitzinger, S.P. (1996). The fate of nitrogen and phosphorus at the land-sea margin of the North Atlantic Ocean. *Biogeochemistry*, *35*, 141–180.

Paliwal, M., Kumar, U.A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, *36*, 2–17.

Park, K., Kuo, A.Y., Shen, J. (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, William & Mary.

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 Environment*, *202*, 31-41.

Prasad, M.B.K., Sapiano, M.R.P., Anderson, C.R., Long, W., Murtugudde, R. (2010). Long-term variability of nutrients and chlorophyll in the Chesapeake Bay: A retrospective analysis, 1985 – 2008. *Estuaries and Coasts*, *33*, 1128–1143.

Qin, Q., Shen, J. (2019). Pelagic contribution to gross primary production dynamics in shallow areas of York River, VA, U.S.A. *Limnology and Oceanography*, *64*, 1484-1499.

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

Roman, M., Zhang, X., Mcgilliard, C., Boicourt, W. (2005). Seasonal and annual variability in the spatial patterns of plankton biomass in Chesapeake Bay. *Limnology and Oceanography*, *50*, 480–492.

Scardi, M., Harding, L.W. (1999). Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecological Modelling*, *120*, 213–223.

Scardi, M. (1996). Artificial neural networks as empirical models of phytoplankton production. *Marine Ecology Progress Series*, *139*, 289–299.

Scavia, D., Kelly, E.L.A., Hagy, J.D. (2006). A simple model for forecasting the effects of nitrogen loads on Chesapeake Bay hypoxia. *Estuaries and Coasts*, *29*, 674–684.

Schelske, C.L., Carrick, H.J., Aldridge, F.J. (1995). Can wind-induced resuspension of meroplankton affect phytoplankton dynamics? *Journal of the North American Benthological Society*, *14*, 616–630.

Scully, M.E. (2010). Wind modulation of dissolved oxygen in Chesapeake Bay. *Estuaries and Coasts*, *33*, 1164–1175.

Scully, M.E. (2016). The contribution of physical processes to inter-annual variations of hypoxia in Chesapeake Bay: A 30-yr modeling study. *Limnology and Oceanography*, *61*, 2243–2260.

Shchepetkin, A.F., McWilliams, J.C. (2005). The regional oceanic modeling system (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modelling*, 9, 347–404.

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

Shen, J., Wang, H.V. (2007). Determining the age of water and long-term transport timescale of the Chesapeake Bay. *Estuarine, Coastal and Shelf Science*, *74*, 750–763.

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., Qin, Q., Wang, Y., Sisson, M. (2019). A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading. *Ecological Modelling*, *398*, 44–54.

Taylor, K.E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*, 7183–7192.

Testa, J. M., Li, Y., Lee, Y.J., Li, M., Brady, D.C., Di Toro, D.M., Kemp, W.M., Fitzpatrick, J.J. (2014). Quantifying the effects of nutrient loading on dissolved O2 cycling and hypoxia in Chesapeake Bay using a coupled hydrodynamic – biogeochemical model. *Journal of Marine Systems*, *139*, 139–158.

Todorovski, L., Džeroski, S. (2006). Integrating knowledge-driven and data-driven approaches to modeling. *Ecological Modelling*, *194*, 3–13.

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, pp. 157–171.

Vilas, L. G., Spyrakos, E., Palenzuela, J. M. T. (2011). Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sensing of Environment*, *115*, 524–535.

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*, 141-149.

Warner, J.C., Geyer, W.R., Lerczak, J.A. (2005). Numerical modeling of an estuary: A comprehensive skill assessment. *Journal of Geophysical Research: Oceans*, *110*, 1–13.

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

Xu, J., Long, W., Wiggert, J.D., Lanerolle, L.W.J., Brown, C.W., Murtugudde, R., Hood, R.R. (2012). Climate forcing and salinity variability in Chesapeake Bay, USA. *Estuaries and Coasts*, *35*, 237–261.

Yang, Q., Tian, H., Friedrichs, M.A.M.**,** Hopkinson, C.S., Lu, C., Najjar, R.G. (2015). Increased nitrogen export from eastern North America to the Atlantic Ocean due to climatic and anthropogenic changes during 1901-2008. *Journal of Geophysical Research: Biogeosciences*, *120*, 1046-1068.

Yin, S., Ding, S.X., Xie, X., Member, S., Luo, H. (2014). A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, *61*, 6418–6428.

Yu, X., Shen, J., Du, J. (2020). A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example. *Water Resource Research, 56*, e2020WR027227.

Zhang, Y. J., Ye, F., Stanev, E.V, Grashorn, S. (2016). Seamless cross-scale modeling with SCHISM. *Ocean Modelling*, *102*, 64–81.

Table 1: Input and output variables for the neural network

Variable	Data frequency	Source
Input forcing		
Air temperature	Hourly	NOAA tidal gauge station measurement at CBBT
River discharges from Susquehanna, Potomac, James, and Choptank Rivers	Daily	USGS
TN, TP loading for Susquehanna, Potomac, James, and Choptank Rivers	Monthly	USGS
Solar radiation	3-hourly	NCEP
Wind (speed, direction, wind hour)	hourly	ECMWF
Output variable		
Time series of each EOF mode	Monthly	

Table2: Eight transformations used for the data-driven model

Transformation	Formula
1	X = x
2	X = log(x)
3	X = 1/x
4	X = exp((x-mean(x))/std(x))
5	X = x/(p50+x)
6	X = x/(p75+x)
7	X = x/(p25+x)
8	X = (x-mean(x))/std(x)

std(x): the standard deviation of x
mean(x): the mean value of x
P25, P50, P75: the 25, 50, and 75 percentile of x

Table 3: A summary of transformations, shiftings, and average periods for each forcing

Forcing	Transformations	Shifting (days)	Average period (days)
Air Temperature	1, 2, 3, 4, 6	0–60	30–90
Flow	1–7	0–60	30–150
TN, TP	1–7	0–60	30–150
Solar	1–8	0–60	30–150
Wind	1–8, wind hour, wind strength	0–60	30–150

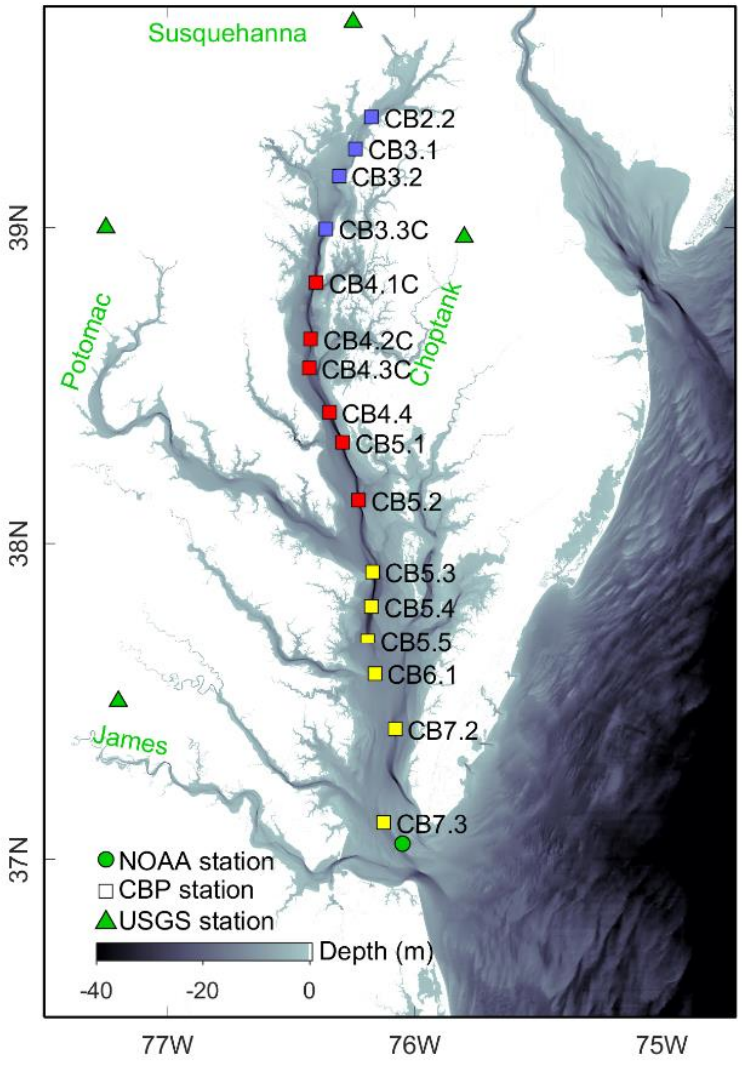

Figure 1: Map of Chesapeake Bay. Chesapeake Bay Program monitoring stations are marked with rectangles, with yellow, red, and purple colors for the lower, middle, and upper bay stations, respectively. Four major tributaries (i.e., James, Potomac, Susquehanna, and Choptank Rivers) are marked as text in the map, with their corresponding USGS gauge station marked with triangles. The NOAA gauge station at the bay mouth is marked with a solid circle. Bathymetry data is based on U.S. Coastal Relief Model generated by the National Geophysical Data Center (https://www.ngdc.noaa.gov).

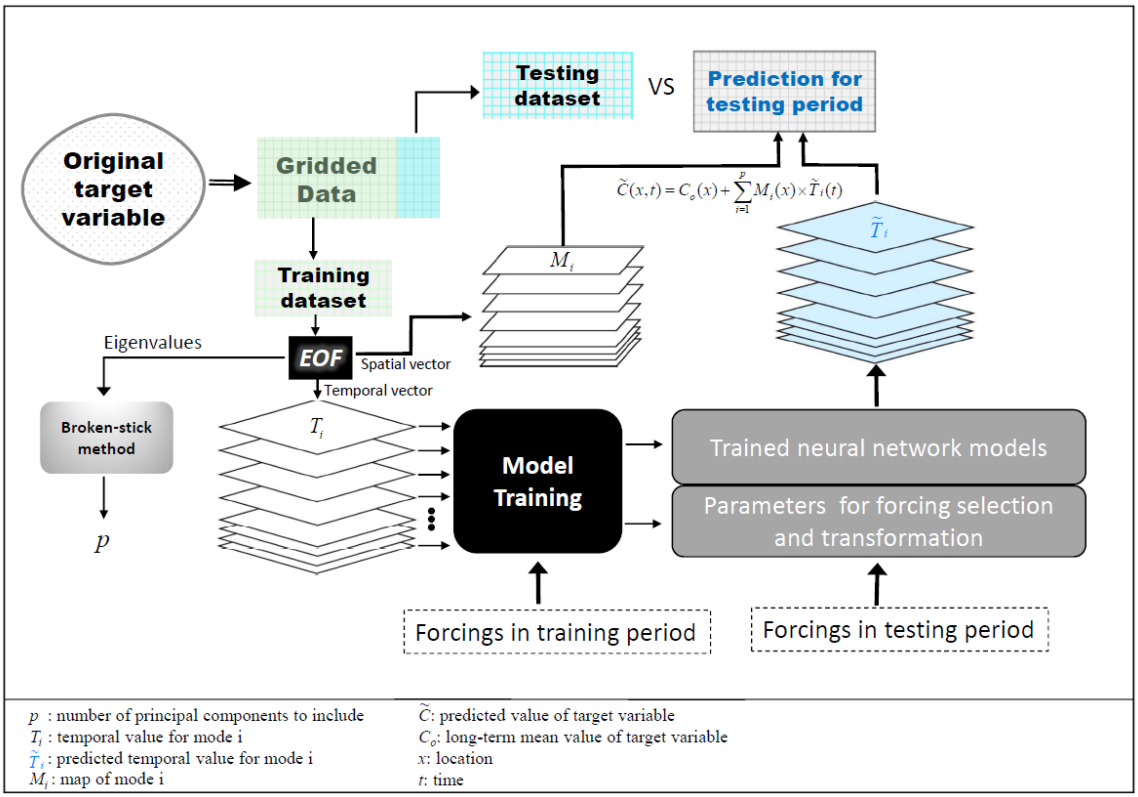

Figure 2: A diagram showing the framework of the proposed data-driven model.

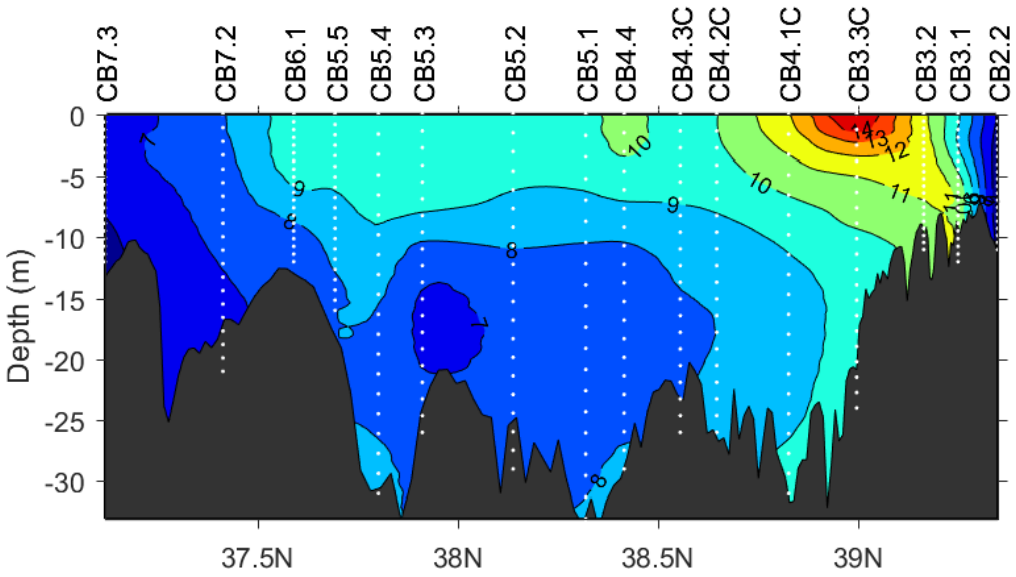

Figure 3: Long-term mean of Chl-a concentration (averaged over 1985-2019) along the mainstem of Chesapeake Bay. For each station, the observed vertical profiles are interpolated into 20 layers (white dots). The values are in the unit of μg/l.

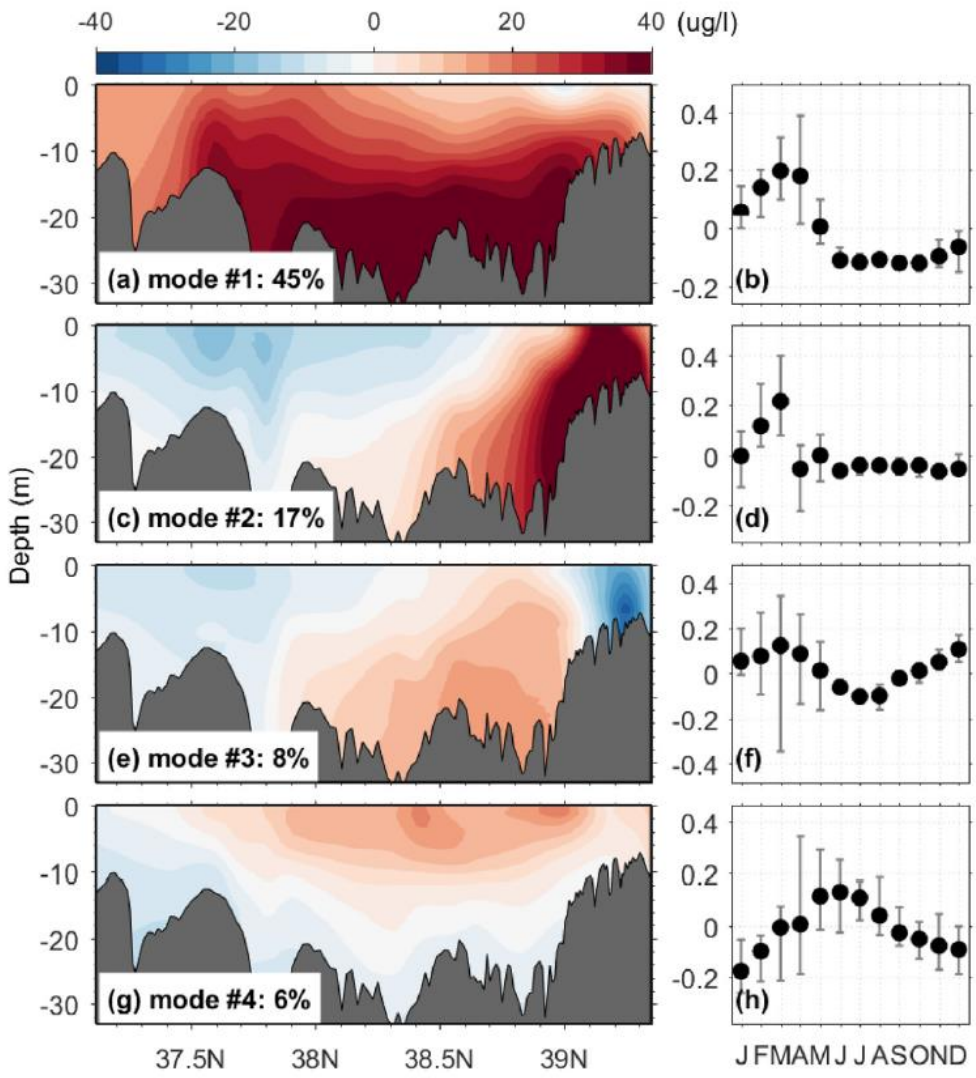

Figure 4: Spatial patterns (left panels) and seasonalities (right panels) of the first four EOF modes for Chl-a along the mainstem channel.

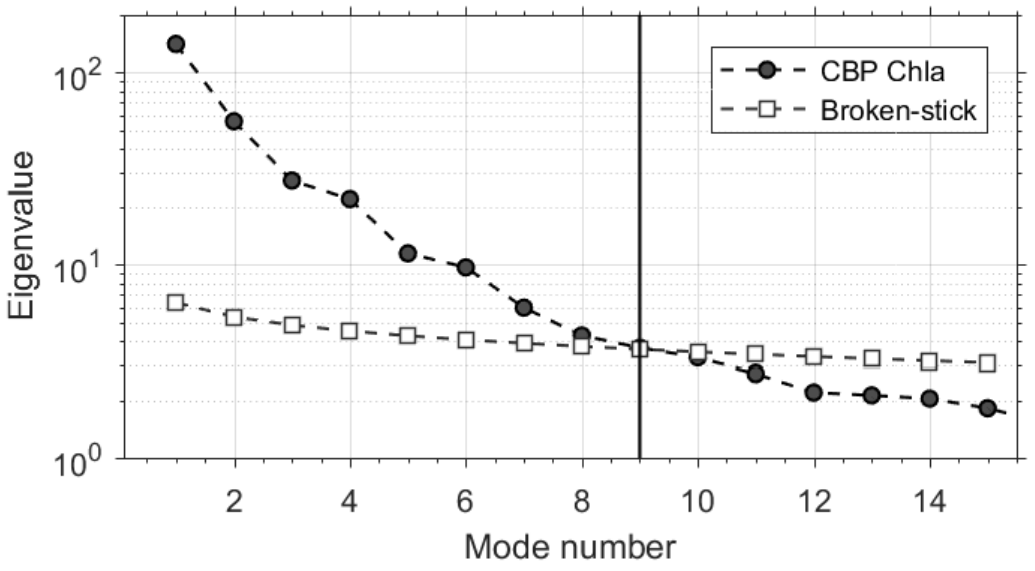

Figure 5: Eigenvalues from the EOF analysis of Chl-a data in training period (solid circles) and from the broken-stick distribution (squares).

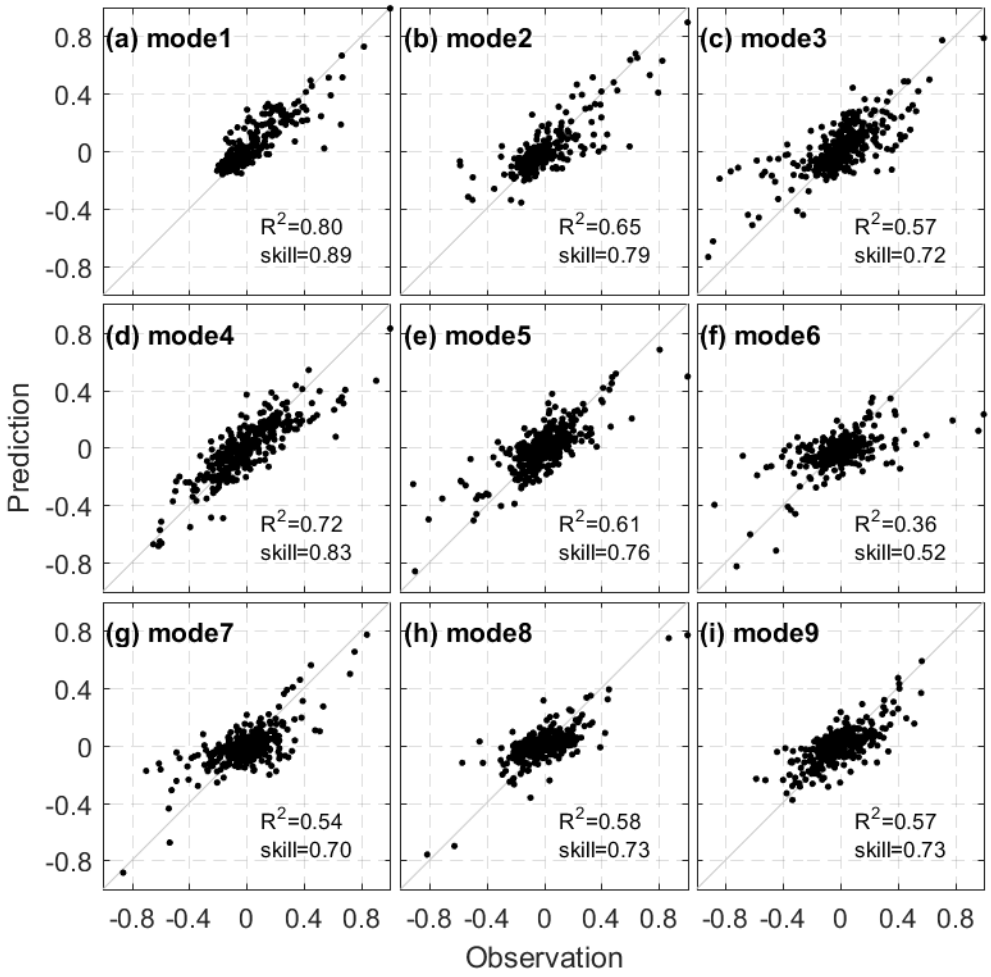

Figure 6: Scatterplots of predicted values against observed values for each mode, with R^2 and model *skill* shown in text. Only the training dataset was used for this plot.

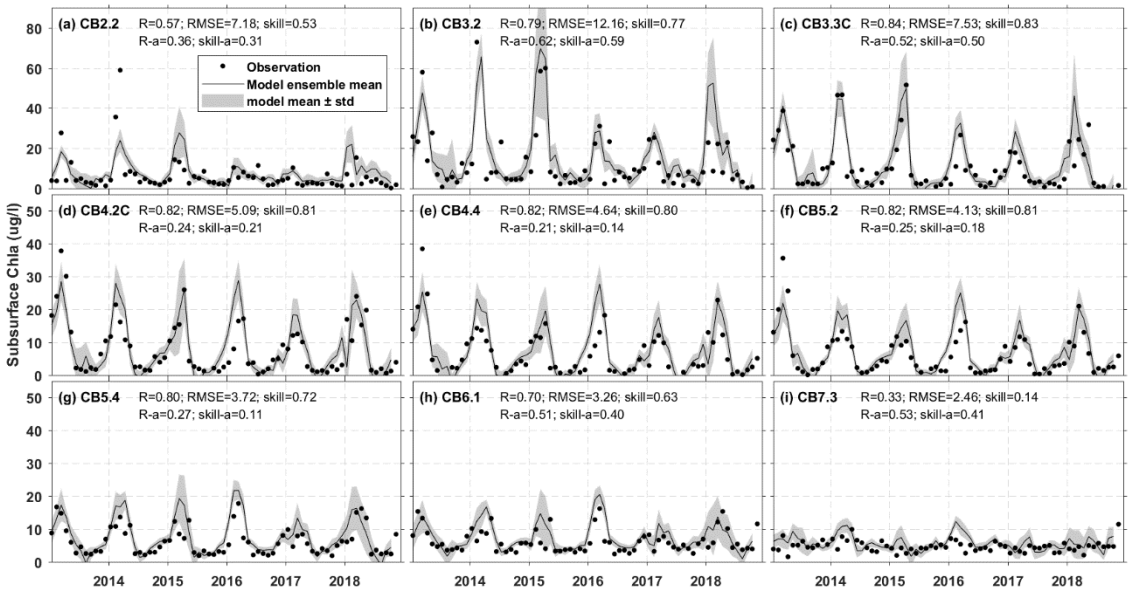

Figure 7: Predicted model results and the observed value of subsurface Chl-a at selected mainstem stations. Only the testing dataset was used for this plot. The gray shades indicate the uncertainties of model predictions; they denote the standard deviation of 100 neural network predictions. Correlation coefficient (R), root mean square error (RMSE), and model *skill* are shown in text. Also shown in text are the correlation coefficient and *skill* for the anomalies, denoted as R-a and *skill-a*. Missing prediction after September 2018 is because of the lack of nutrient load data.

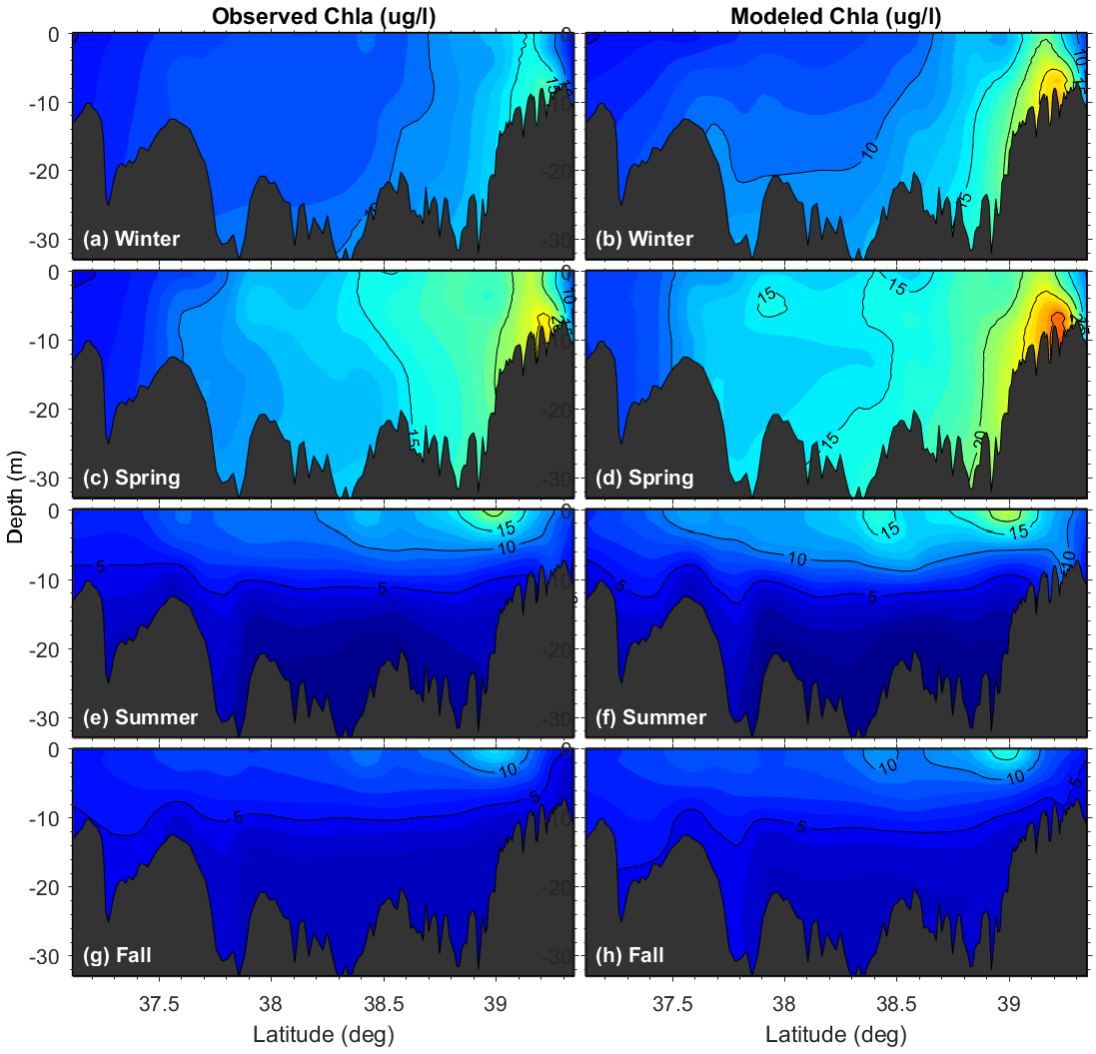

Figure 8: Seasonal pattern of Chl-a from observation and model prediction for the testing period. Note the difference in magnitude and spatial distribution of Chl-a (e.g.., very large Chl-a in the bottom during winter and spring, while smaller magnitude but larger in the surface during summer and fall).

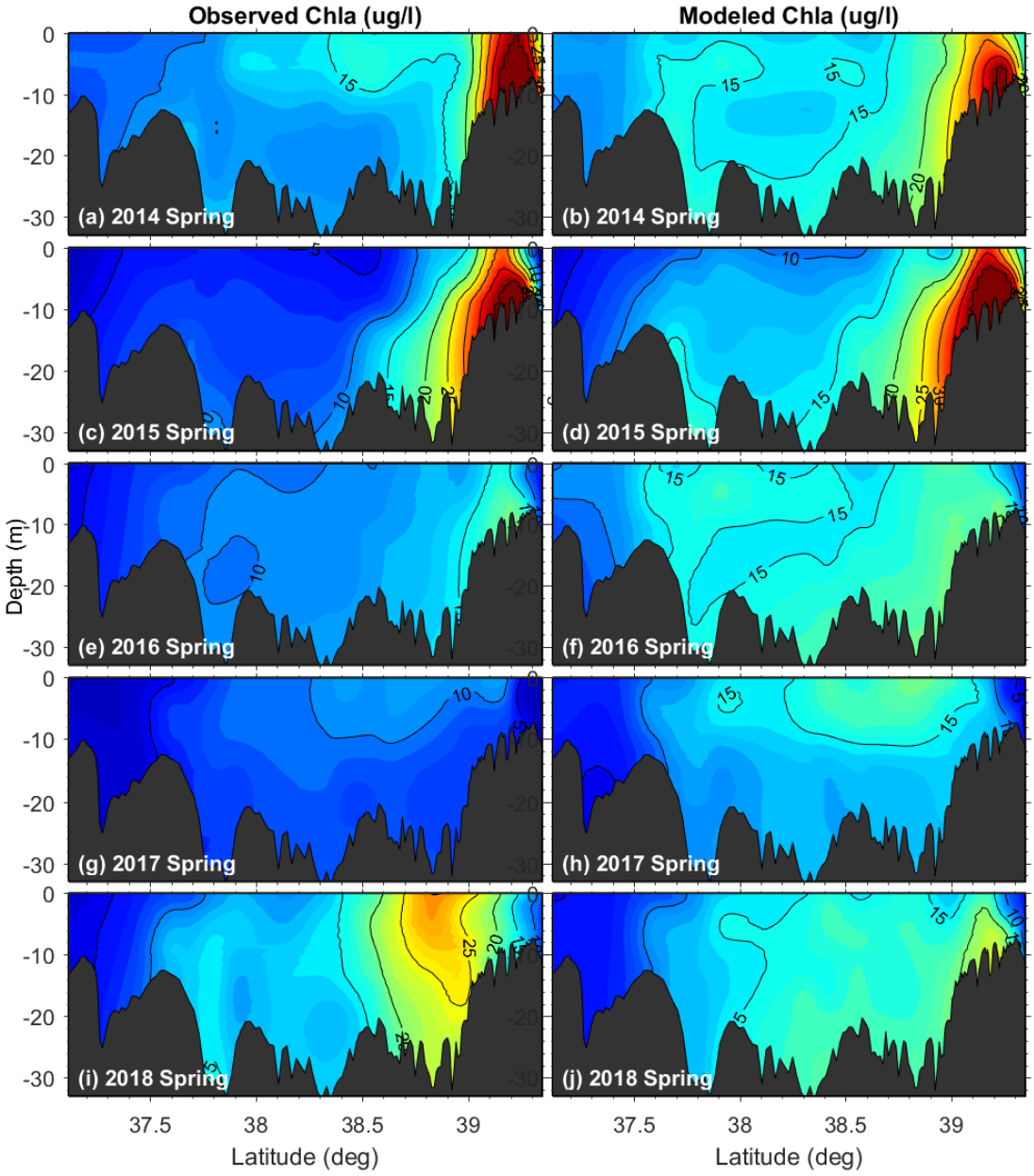

Figure 9: Comparison of spring Chl-a between observation and modeling results for the testing period.

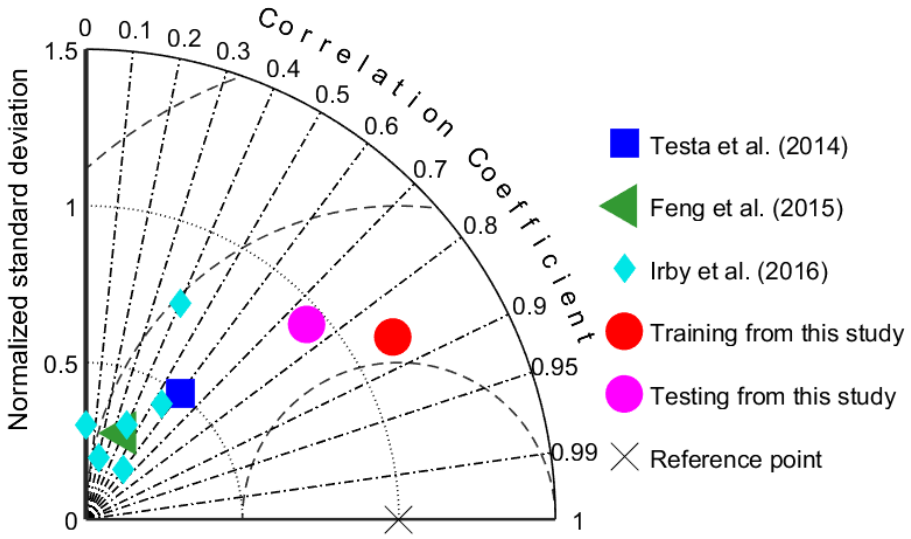

Figure 10: Taylor diagram showing performance in subsurface Chl-a from different models. The data from previous models are based on fig. 6 in Testa et al. (2014), fig. 5 in Feng et al. (2015), and fig. 8 in Irby et al. (2016). The radial distance from the origin is proportional to the ratio standard deviations; the azimuthal angle indicates the Pearson correlation coefficient; and the distance between the each filled marker and the "reference" point (marked with a cross) indicates the centered root mean square deviation.

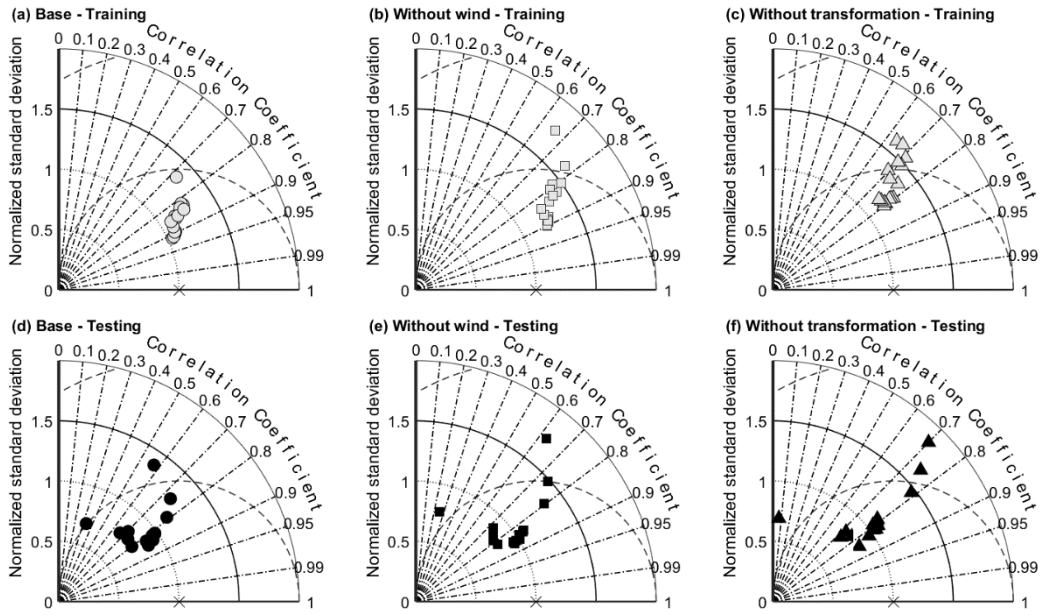

Figure 11: Taylor diagrams showing the model performance in simulating subsurface Chl-a at each of the 16 stations from base run and two sensitivity tests. Top and bottom panels for the training and testing dataset, respectively. (a, d) a base run with full forcings; (b, e) a run without wind; and (c, f) a run without performing transformation of input forcings.

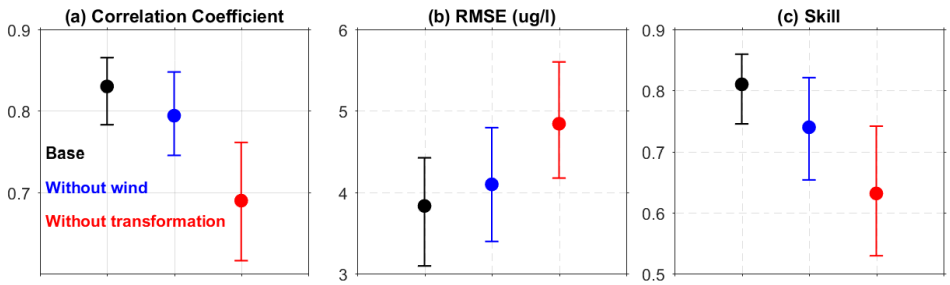

Figure 12: Model's training performance for different runs (indicated by different colors). The median values of a given statistic metric over 16 monitoring stations are shown with solid circles while the 25th and 75th percentiles are shown with error bars.

Table S1: Forcing and corresponding transformations used to model 1st principal component. See table 2 in the main article for details of transformation functions.

Forcing	Transform function	Shifting (days)	Accumulation (days)
Solar radiation	3	60	55
James flow	1	0	45
Potomac TP loading	3	50	95
Westerly wind speed	4	4	15
James TP loading	3	0	105
Potomac TN loading	4	0	105
Potomac flow	4	60	55
James sediment loading	6	0	75

Table S2: Forcing and corresponding transformations used to model 2nd principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
Air temperature	7	30	15
Susquehanna flow	8	10	15
Solar radiation	3	0	135
James TP loading	3	50	125
Choptank TN loading	4	50	35
Potomac TN loading	4	0	115
Susquehanna sediment loading	2	0	45

Table S3: Forcing and corresponding transformations used to model 3rd principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
Solar radiation	4	10	55
Choptank sediment loading	6	10	135
Susquehanna sediment loading	5	20	15
Potomac flow	4	50	15
Potomac sediment loading	1	0	25
Air temperature	7	10	25
Choptank TN loading	4	0	55
Northerly hour	1	5	15
Susquehanna flow	4	50	55
Susquehanna TP loading	4	40	135
Choptank TP loading	4	50	125
James flow	1	10	25
Potomac TP loading	3	20	15

Table S4: Forcing and corresponding transformations used to model 4th principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
Solar radiation	1	0	35
Potomac flow	4	40	85
Susquehanna TP loading	7	0	45
Air temperature	3	40	35
James TP loading	3	50	125
Potomac sediment loading	6	50	15
Susquehanna flow	1	0	45
Choptank TN loading	6	50	125
Choptank sediment loading	1	0	135
Potomac TN loading	6	50	125

Table S5: Forcing and corresponding transformations used to model 5th principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
James sediment loading	6	50	25
James TP loading	6	50	15
Easterly wind speed2_hour	4	1	15
Easterly wind hour	4	0	15
Northerly hour (speed>4m/s)	3	5	15
Southerly wind intensity	3	1	15
Susquehanna TP loading	4	40	135
Susquehanna flow	4	50	75
Southerly wind speed	2	5	15
Potomac TN loading	4	50	15
Potomac sediment loading	4	50	15
Choptank TN loading	4	50	35

Table S6: Forcing and corresponding transformations used to model 6th principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
James sediment loading	6	0	115
James flow	1	30	55
Potomac sediment loading	6	50	25
Susquehanna sediment loading	5	50	15
Susquehanna TP loading	5	50	15
Easterly wind hour (speed>4)	3	3	15
James TP loading	4	0	15

Table S7: Forcing and corresponding transformations used to model 7th principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
Choptank TN loading	4	0	75
Choptank sediment loading	1	50	15
Susquehanna sediment loading	4	40	135
Susquehanna flow	4	50	55
James sediment loading	1	50	15
James TP loading	7	50	65
Potomac TP loading	3	50	15
Potomac flow	4	50	15
Susquehanna TN loading	3	50	65
Susquehanna TP loading	2	50	15
James flow	6	40	15
James TN loading	2	20	125
Air temperature	7	0	135
Potomac sediment loading	4	50	15

Table S8: Forcing and corresponding transformations used to model 8th principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
James TP loading	1	20	15
Susquehanna flow	1	20	15
Susquehanna TN loading	1	0	55
Easterly wind speed	6	5	15
Potomac TP loading	4	20	35
Choptank TN loading	4	0	15
James sediment loading	1	20	25
Choptank TP loading	1	20	35
Susquehanna sediment loading	1	50	25
Easterly wind intensity	6	5	15

Table S9: Forcing and corresponding transformations used to model 9th principal component.

Forcing	Transform function	Shifting (days)	Accumulation (days)
Westerly wind hour (speed>4m/s)	3	5	15
James TN loading	1	20	65
Susquehanna flow	4	50	135
Susquehanna sediment loading	4	20	125
Air temperature	7	0	135
Choptank sediment loading	4	0	15
Potomac flow	4	10	135
James sediment loading	6	50	25
Choptank TN loading	4	50	25
Potomac TP loading	5	50	15

CHAPTER 3. CHLOROPHYLL-A IN CHESAPEAKE BAY BASED ON VIIRS SATELLITE DATA: SPATIOTEMPORAL VARIABILITIES AND PREDICTION WITH MACHINE LEARNING

Abstract: Chlorophyll-a concentration (Chl-a) is practically used to indicate the abundance of phytoplankton biomass, a fundamental component of aquatic ecosystems. Its spatiotemporal distribution is strongly related to the ecosystem dynamics. Compared to conventional low-frequency shipboard measurements at a limited number of sampling locations, satellite data provides a better coverage for a synoptic view of Chl-a variabilities, particularly in large coastal systems such as Chesapeake Bay. Here we use Visible Infrared Imaging Radiometer Suite (VIIRS) satellite data from 2011 to 2018 to analyze the Chl-a variability in Chesapeake Bay. The reliability of the satellite data is confirmed by its good agreement (R^2=0.56) with shipboard measurements. Analysis results show seasonality of Chl-a varies in different regions, with maxima occurring in spring for regions near mouth of major tributaries, winter near the bay entrance, and summer elsewhere. There are two seasonal peaks associated with spring and summer blooms. Data Interpolating Empirical Orthogonal Functions (DINEOF) is used to efficiently estimate the missing records. A machine-learning-based data-driven model is developed to simulate Chl-a distribution. Driven by external forcing including river discharge, nutrient loadings, solar radiation, wind, and air temperature, the data-driven model shows an overall satisfactory performance in reproducing the spatiotemporal variations of Chl-a, with a bay-wide averaged root mean square error of 1.85 ug/l. By combining DINEOF and machine learning, this study demonstrates the potential of

useing data-driven model to predict high-resolution spatiotemporal variations of water quality in coastal waters.

Keywords: remote sensing; machine learning; data-driven model; phytoplankton; Chesapeake Bay

1. INTRODUCTION

Phytoplankton biomass is a fundamental component of aquatic ecosystems and typically indicated by Chlorophyll-a (Chl-a) concentration. Chl-a is thus a key water quality index in estuarine and coastal waters. High Chl-a, mostly caused by excessive nutrient load, wastewater input, and warming climate (Harding et al., 2016), is closely related to coastal hypoxia (Kemp et al., 2005), loss of submerged aquatic vegetation (Orth et al., 2010), and blooms of toxic cyanobacteria (Tango and Butler, 2008). Chl-a can be measured or estimated in multiple ways, including *in situ* sampling, continuous fluorescence-based measurement, and remote sensing. These methods have different advantages and limitations in terms of data accuracy, spatial coverage and temporal resolution. *In situ* sampling and then laboratory analysis is still a typical routine for field surveys and probably the most precise way to measure Chl-a. However, this method is effort-intensive and time-consuming (Abbas et al., 2019); the resulting data often have poor spatial and temporal coverage. These limitations also apply to continuous fluorescence-based measurement at a fixed station. In contrast, remote sensing data, especially those satellite-based, provide a better spatial coverage and permit a more

synoptic assessment of larger-scale ocean dynamics. Drawbacks of satellite data are its accuracy and that only surface values are available.

Recent advances in remote sensing provide us valuable opportunities to examine the detailed spatial variabilities and understand the associated phytoplankton dynamics. Different from salinity or temperature that often smoothly changes over time and space in non-eddy coastal seas, Chl-a can be highly patchy (Martin et al., 2002), primarily because algae grow and aggregate in short timescale (days) and their concentration are sensitive to not only temperature, salinity, and nutrient, but also spatially varying flushing capacities (Lucas et al., 2009; Qin and Shen, 2021). As a result, the traditional shipboard measurements, conducted monthly or bimonthly at a limited number of stations as in prominent oceanic monitoring programs, could be insufficient to represent the mean condition of Chl-a over a long period (e.g., month) or a large area. Nevertheless, the "snapshot" data of Chl-a are still widely used to validate numerical models. Because of the high variability of Chl-a, numerical models often have poor performance in Chl-a simulation compared to other water quality state variables (e.g., Testa et al., 2014; Feng et al., 2015; Irby et al., 2016). Assessment of model performance is complicated by the fact that the horizontal discretion of space in a numerical model is much larger than the area represented by a monitoring station. For instance, numerical models (e.g., Testa et al., 2014; Du and Shen, 2016) applied to Chesapeake Bay have spatial resolution on the order of 1000 m. Spatially averaged Chl-a will be more suitable reference to compare with numerical simulations. Therefore, a bay-wide distribution of Chl-a, revealed from high-resolution satellite data, will be of great interest to the modeling community.

In this study, we analyze the spatial and temporal variations of Chl-a based on Visible Infrared Imaging Radiometer Suite (VIIRS) data (Zheng and DiGiacomo, 2017). This dataset has a nearly daily frequency and a spatial resolution of 750 m and the accuracy of Chl-a concentration has been proved to be superior than other satellite products (Zheng and DiGiacomo, 2017), such as Sea-Viewing Wide Field-of-View Sensor (SeaWiFS), Medium Resolution Imaging Spectrometer (MERIS), and Moderate Resolution Imaging Spectroradiometer (MOIDS). The data are available since late 2011. Reliability of the data is verified with existing long record of monthly shipboard measurements by Chesapeake Bay Program (https://www.chesapeakebay.net).

Inspired by rapid accumulation of satellite data and recent advances in machine learning technology for water quality simulation (Muller and Muller, 2015; Yu et al, 2020), we investigate the capability of machine-learning-based data-driven model for high resolution simulations. The data-driven model developed by Yu and Shen (2021) has shown a good performance in simulating the vertical and along-channel variability of Chl-a along the main axis of Chesapeake Bay. The purpose of this study is to further test the feasibility of applying the data-driven model to simulate Chl-a using high-resolution satellite data and to examine what temporal scales that the data model is able to simulate with reasonable accuracy. Compared to the simulation of Chl-a at a limited number of monitoring stations, simulating bay-wide Chl-a using satellite data will encounter a major challenge raised by its high variability in both space and time, as well as extensive gaps in satellite data. We will discuss a promising method and the necessity to fill these data gaps.

2. MATERIALS AND METHODS

2.1. *In situ* observational measurements

Long-term (1985-2019) shipboard measurements at 37 mainstem stations in Chesapeake Bay are collected to verify the satellite data and to examine the variability of Chl-a. Since 1985, water quality parameters including salinity, temperature, nutrient, and Chl-a have been measured by ship surveys carried out monthly or bi-monthly (data available at https://www.chesapeakebay.net). Locations of the stations are shown in Fig. 1a. Despite the spatial and temporal limitations on sampling resolution, this dataset has provided a reliable basis in previous studies (e.g., Hagy et al., 2004; Kemp et al., 2005; Murphy et al., 2011).

Chl-a at these mainstem stations were measured at a minimum of two layers, one at the surface and the other near the bottom. Only surface data are used to compare with satellite-derived Chl-a. To examine the variability of Chl-a, long-term mean, standard deviation, relative standard deviation, and seasonal mean are calculated for each station. The relative standard deviation is calculated as the standard deviation normalized by the mean.

2.2 Satellite data

Following Zheng and DiGiacomo (2017), phytoplankton component of the total light absorption coefficient of water is extracted using a generalized stacked-constraints model (GSCM). The total light absorption coefficient of water is derived from satellite remote-sensing reflectance using the Quasi-Analytical Algorithm (Lee et al., 2002). The GSCM has been applied to Chesapeake Bay and results show significant improved

accuracy of satellite-derived Chl-a. However, the cost is reduced data availability because GSCM provides no feasible solutions when the light absorption coefficient of water is subject to large errors in areas such as the highly turbid upper bay.

The VIIRS satellite data (2011-2018) have a spatial resolution of 750 m. Its high-resolution permits examination of the lateral variations across the bay's mainstem and within major tributaries (Fig. 1b-c). For this study, Chl-a is averaged over all available days within a 7-day interval and gridded into a prescribed mesh grid (resolution of 0.015×0.015 degree, from 75.5W to 77W in longitude and from 36.8N to 39.6N in latitude). A null value is assigned when there is no data within the 7–day period at any given grid. Comparison between *in situ* measurements and 7-day averaged satellite data (within a 3-day window of the observation date and within 750 m radius from the monitoring station) shows a good agreement, with R^2 of 0.56 and RMSE of 4.7 ug/l (Fig. 1d).

The extracted pixel-based Chl-a data have substantial gaps. Even after 7-day averaging, there are still areas with gap percentage greater than 50%. These areas are excluded from analysis. In total, there are 4813 grid points with gap percentage less than 50% in the study region (see Fig. 1a for the region of interest). On average, these selected grid points have about 20% of gaps. The gaps have noticeable seasonality, with persistent high percentage in specific months. For instance, in April, the gap is more than 30% over the study period (Fig. 2).

To simulate the spatiotemporal variations of Chl-a with a data-driven model, no data gaps are allowed (Yu and Shen, 2021). One choice is to remove grids that have data

gaps, which, however, will result in much less sampling points, leading to poor spatial coverage. Another choice is to interpolate the missing records. Interpolation is possible when the variable shares covariance either spatially or temporally. The method we use to interpolate the missing data is described below.

2.3 DINEOF

Data Interpolating Empirical Orthogonal Functions (DINEOF) is an EOF-based method to fill in missing data from geophysical fields (Beckers and Rixon, 2003). The procedure fills gaps by iteratively decomposing the data field via Singular Value Decomposition (SVD) until the best solution is found. This is done by progressively including more EOF modes in the truncated reconstruction until minimization of error converges between interpolated values and reference values.

A procedure is briefly described here, with the algorithm shown in Fig. 3, which depicts two major loops: one to get optimal estimation under a given number of modes; and the other to determine the optimal number of EOF modes. For detail descriptions, readers are referred to Beckers and Rixen (2003) and Alvera-Azcárate et al. (2016). Assume the original data matrix as **X**, with $\mathbf{X}_{ij}$ being the value at location i at moment j, and there are missing data points with (i,j) belonging to **I1.** A randomly selected subset of non-missing data with (i,j) belonging to **I2** is used as reference (1% of records are used in this study). First, the long-term mean is removed from the dataset. Second, for data points in **I1** and **I2**, zero values are assigned to generate a new Matrix $\mathbf{X_0}$. A SVD decomposition of matrix $\mathbf{X_0}$ gives the first estimate of the spatial and temporal eigen vectors **U** and **V** as well as their singular values **D** (the diagonal matrix).

$$\mathbf{UDV}^{*}= \mathbf{X_0} \qquad (1)$$

$\mathbf{X_0}$ will be updated with truncated reconstruction using the obtained EOF eigen vectors for both data points in **I1** and **I2**.

$$\left(\mathbf{X_0}\right)_{ij} = \left(\mathbf{U}_N \mathbf{D}_N \mathbf{V}_N^{*}\right)_{ij},\ (i,j) \in \mathbf{I1} \cup \mathbf{I2} \qquad (2)$$

The "truncated" here refers to the practice of using a limited number of EOF modes (N) instead of all modes. After step (2), root mean square error (RMSE) between true value and interpolated value in reference dataset (**I2**) will be calculated.

$$RMSE = \sqrt{\left(X_0 - X\right)^2 / m},\ (i,j) \in \mathbf{I2} \qquad (3)$$

By repeating step (1-3), the RMSE will decrease with more iterations. For a given number of modes, these steps will be repeated until the RMSE does not decrease by a predefined value (1×10^{-5} is used in this study). To determine the optimal number of modes (denoted by N) used in the truncated reconstruction, the above loop will start with N=1. When the RMSE decreases but by less than the prescribed criterion with additional iteration, N increases by 1. If the RMSE starts to increase after increasing N by 1, the entire procedure stops. The optimal number of modes is N-1.

Robustness of the method has been demonstrated by extensive applications for satellite data (e.g., Alvera-Azcárate et al., 2016; Hilborn and Costa, 2018; Yang et al., 2021). In this study, we run DINEOF with open-source language R (code available at http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF). For 7-day averaged satellite data, the optimal number of EOF modes is 16. The RMSE between reference records and

interpolated values decreases greatly when the number of modes increases from 1 to 9, while RMSE decreases negligibly after 10 modes (Fig. 4). An experiment is conducted to verify the interpolation. We randomly set 1% of records with null values and these records are independent dataset different from the internal reference records used during DINEOF. The interpolated data for the missing records is then compared to the true values (values before being masked as null). Results show DINEOF can estimate the missing records with acceptable error, with RMSE of 1.78 ug/l (Fig. 5). The error can be more than 10 ug/l when Chl-a is extremely large (Fig. 5). The gap-free data will be used to train and verify a data-driven model.

2.4 Data-driven model

A data-driven model introduced by Yu and Shen (2021) is applied to simulate the spatial and temporal variations of Chl-a. The data-driven model comprises three major components: empirical orthogonal function (EOF), artificial neural network, and forcing transformation auto-selection. EOF is applied to reduce the dimension of data by extracting the spatial pattern and temporal variations of principal components. The temporal variations of principal components will be simulated by artificial neural network. Distinguished from previous studies that use *in situ* measurements of other water quality parameters as inputs (e.g., Scardi and Harding, 1999; Soro et al, 2020), the data-driven model here uses only external forcings, include river flow and nutrient loadings from large tributaries (data from USGS, https://www.usgs.gov), air temperature (measured at Chesapeake Bay Bridge-Tunnel station; data from NOAA database https://tidesandcurrents.noaa.gov/), solar radiation and wind (in global ERA5 reanalyzed product from European Centre for Medium-Range Weather Forecasts,

https://www.ecmwf.int). Details of the model description and procedures can be found in Yu and Shen (2021). Simulating the satellite-derived Chl-a in this study is a follow-up of Yu and Shen (2021) to demonstrate the capability of the data-driven model to predict high-resolution, large-scale variations of Chl-a in coastal systems.

3. RESULTS

3.1 *In situ* monitoring data

The historical shipboard measurements of Chl-a show a similar pattern for mean, standard deviation, and relative standard deviation (Fig. 6). All of them have a higher value in the middle-upper bay, specifically the region between 38.5N-39N, and lower value in the upper-most region close to Susquehanna River outflow and in the lower bay. Low values in the upper-most region is likely caused by turbidity-induced light limitation (Harding et al., 1992; Zhang et al., 2021). In addition, there are noticeable lateral variabilities, with larger Chl-a at stations in shallow shoals rather than the deep channel. Furthermore, the pattern is persistent between the 35-year and the 8-year period from 2011-2018 (when satellite data are available) (Fig. 7).

It is worth to note the high variability at these upper bay stations, which may be attributed to the fluctuation of turbidity maximum zone induced by changing river discharge, wind forcing, and estuarine circulation (Sanford et al., 2001). When the turbidity maximum zone shifts northward, more stations will be exposed to a better light condition that favors the growth of phytoplankton. Stations within the turbidity maximum zone fluctuation area presumably have higher temporal variability of light condition and thus Chl-a concentration.

Another interesting fact is the spatial heterogeneity in Chl-a seasonality indicated by climatological monthly means. Different from the conventional understanding that spring bloom dominates in Chesapeake Bay, more than half of the monitoring stations (18 of the 37 stations) have a larger peak of Chl-a in summer months, while 13 stations have a larger peak of Chl-a in spring months (Fig. 6d). Near the bay entrance, Chl-a has a weak peak in late fall or winter. The spatially varying seasonality suggests that there are likely different limiting factors controlling algal bloom in different regions of the bay (Zhang et al., 2021). In particular, the summer bloom may be attributed to longer and stronger solar radiation, up-lift of pycnocline induced by surface heating and freshening following spring high flow, thinner surface mixing layer, and upward nutrient flux associated with seasonal hypoxia during summertime.

3.2 Satellite data

The 8-year mean of Chl-a satellite data show a similar spatial pattern as in the observation, but with clearer and detailed lateral distributions (Fig. 8). Mean Chl-a increases from the bay mouth to the upper bay, with noticeable larger values near both east and west shallow shoals compared to the deep channel (Fig. 8a). This lateral distribution is consistent with observation. However, some differences between the two datasets are also noticeable. For instance, the maximum mean Chl-a near 39N exceeds 20 ug/l in shipboard measurement, while not in the satellite data. This difference is not unexpected, since the model used to extract Chl-a from satellite data is known to be imperfect and the satellite data only cover a short period, not necessarily including extremely high or low Chl-a values. In addition, extreme high values of Chl-a are often

regarded as outliers when fitting the empirical conversion function with which Chl-a concentration is estimated from the light reflectance in remote sensing data.

Variability of Chl-a, indicated by RSTD, has two high-value regions, one in the upper bay and the other in the coastal ocean (Fig. 8c). A relatively high variability is found in the middle-bay right off the Potomac River mouth, which is not captured by the observational dataset primarily due to limited monitoring efforts in this region. The high variability of Chl-a in this region may be attributed to hydrodynamic interaction between the mainstem and tributaries, which is known to play a key role in exchange of soluble materials (e.g., dissolved oxygen) within the estuarine system (Kuo and Neilson, 1987). It is worth noting that the RSTD of satellite data is relatively smaller than that from observational data, largely because satellite data is averaged over a certain period (7-day) and over a certain area (~1.0 km in the case of gridded satellite data), while observation data are based on measurements once or twice each month and at a particular location.

With the high-resolution data, we can map the timing of algal bloom. For each grid, the climatological monthly mean is calculated and the month with peak mean Chl-a is determined. The timing of peak Chl-a exhibits interesting spatial patterns, largely consistent with the Chl-a observations (Fig. 8d). Spring peak occurs in the upper bay, inside the lower Potomac River, off the Potomac River mouth, and near the James River mouth. Summer peak occurs mainly in the upper bay (38-39N) and lower bay (37-37.5N). Spring-peak regions share one key factor: they are close to the mouth of major tributaries (e.g., Susquehanna, Potomac, Rappahannock, York, James, and Choptank Rivers). Note that the spring peak does not occur near Patuxent and Chester Rivers (the other two major tributaries), which coincides with the fact that these two rivers have the

smallest mean river discharge among major tributaries. The peak time seems to be regulated by discharge and corresponds to the time when riverine nutrients reach to the given location. For those spring bloom-dominated areas, the larger nutrient level in spring likely overwhelm the better light condition in summer. These areas may also experience summer bloom, but the summer Chl-a concentrations are typically smaller. Timing of bloom are possibly induced by transport processes and nutrient supply. It will take 100-300 days for riverine material discharged from Susquehanna River to reach the lower bay; the transport time varies depending on flow regimes and wind field (Shen and Wang, 2007).

The seasonality of satellite-derived Chl-a is characterized with clearly larger value in spring and summer, compared to the other two seasons (Fig. 9; Fig. 10). The overall along-bay and across-bay distribution seems persistent throughout the year. The spring bloom is a well-known ecological feature in Chesapeake Bay; it results from the stimulus of spring high flow after snowmelt in watersheds that feed the major rivers (Harding and Perry, 1997). The summer bloom is believed as a secondary bloom fueled by nutrients released from sediment and recycled nutrients in the water column (Malone et al., 1996; Kemp et al., 2005). We propose here that the summer bloom can be dominant for regions when the nutrient level is subject to prolonged transport.

The two-peak seasonal characteristic is more obvious from EOF analysis (Fig.11). The first EOF mode, accounting for 32% of the total variance, features positive spatial values throughout the entire bay, meaning that bay-wide Chl-a changes in phase. The mode has clear seasonality with two peaks, one in March and the other in July. A similar EOF analysis was conducted by Yu and Shen (2021) based on *in situ* Chl-a at 16

mainstem stations (no station at shallow shoals included). Their analysis shows that the first mode is featured with only one peak in March. The remarkable differences between these two EOF analyses suggest the important contribution of shallow shoals in the overall variance. It is likely that the summer peak will be more significant when including shallow shoals.

3.3 Simulation Chl-a with data-driven model

The temporal variations of a limited number of principal components are simulated by the data-driven model. The first step is to determine how many modes are non-trivial and worth to be included. DINEOF analysis shows that starting from mode 17, RMSE increases (Fig. 4). It suggests modes after 16 is largely noise, and unlikely to add a meaningful contribution to the overall variance. However, when using more than 10 modes, the RMSE decreases negligibly. Therefore, we include 10 modes in the data-driven model. These 10 modes account for 80% of the total variance. The spatial and seasonal variation for the first 4 modes can be found in Fig. 11. The first mode shows in-phase variations of Chl-a throughout the bay. High variations of Chl-a is located in the upper estuary and the variability gradually decreases toward lower bay. The second mode has out-of-phase variations between upper bay and middle-lower bay. The third mode features high variability in middle bay where the variation is opposed to both upper and lower bay. The fourth mode represents the variability in the upper bay. It can be expected that the response of each mode to external forcings will be different.

The temporal variations of each of the 10 modes are simulated using neural network that takes external forcings as inputs. For each mode, the datasets, including inputs and outputs, are divided into two sub-dataset, with the first 75% of records for

training and the other 25% for prediction. Considering the randomness in initialization of neural-network parameters, the outcome (or prediction) differs for each training even with the same inputs. To address this uncertainty, we train the model 50 times for each mode. The uncertainty is much larger than that in the dissolved oxygen simulation (Yu et al., 2020). Nonetheless, the model is well trained for the training period (Fig. 12).

Model performance for the testing period (2017-2018) is more meaningful, as input forcing and target variable in the testing period do not involve during training in any aspect. The overall spatial pattern of modeled Chl-a anomalies (relative to the long-term mean) and full signal agree well with that in satellite-observed Chl-a (see two example date representing spring and fall in Fig. 13). The seasonal variability is well captured, despite discrepancies in some regions. More importantly, the data-driven model reproduces the strong contrast of Chl-a concentration between the bay and the adjacent coastal ocean. As expected, the data-driven model, no matter how well-tuned, will not capture every detail of Chl-a's spatial variations. For instance, the model captures the spring bloom (e.g., in April of 2017), but the predicted magnitude is less than observation.

Comparisons of time series at selected six stations (Fig. 14) show that the model performance seems to decrease toward the upper bay. In the lower bay (e.g., station CB6.2 and CB7.3), the model well captures the seasonal and interannual variations of satellite Chl-a (Fig. 14). Model performance is also satisfactory in the middle bay. However, the model has difficulty reproducing the high variability in the upper bay (e.g., station CB3.3C). The spatial heterogeneity of model performance suggests that the upper bay Chl-a is less predictable compared to that in lower bay. Error of satellite data due to

high turbidity in the upper bay is likely a factor leading to the lower performance in that region.

The model performance varies spatially, as indicated by the root mean square error (RMSE) and the relative error (RE) between predicted and observed Chl-a. Calculated RE and RMSE show noticeable spatial heterogeneity, with the worst performance in the eastern shoals between 37.5N and 38.5N and in the upper bay. Averaged over all the grids, RMSE is around 1.85 ug/l (Fig. 15).

4. DISCUSSION

4.1 *In situ* observation vs satellite data

Comparison between *in situ* observation and satellite data has confirmed the reliability of the satellite data in terms of several aspects, including the lateral difference between shoals and the deep channel, longitudinal trend from the upper bay to lower bay, and the peak timing. However, we acknowledge that there are noticeable differences in the magnitude of temporal variability between these two different datasets. A major difference is in the bay-wide averaged standard deviation of monthly mean Chl-a, which is 8.8 ug/l in *in situ* measurements compared to 3.3 ug/l in satellite data. The smaller variability in satellite data may be attributed to three possible factors. (1) Satellite data are 7-day averaged, opposed to monthly (sometimes bi-monthly) sampling frequency in CBP data. (2) Each record of satellite data represents an average over a finite area; the area is determined by the horizontal resolution; for the VIIRS data, the area is 750 m × 750 m. (3) there is an inherent error when estimating Chl-a from satellite data, with extremely low or high values likely to be excluded when establishing the conversion

function (Zheng et al., 2015). If averaging over a finite area, the variability of true Chl-a tends to be smaller.

Both satellite data and *in situ* measurements demonstrate a remarkable difference between the deep channel and flanking shoals. The underlying mechanisms can be further explored if given more information of nutrient, salinity, and temperature vertical profiles in both shoals and deep channel. Even though it is beyond the scope of this study to uncover the exact mechanisms, it is worth noting the possible factors that may contribute to such a pattern. First, water in the deep channel is more dispersive and moving faster during each tidal cycle (Xiong et al., 2020), which is less favorable for phytoplankton to aggregate in the mainstem comparing to shoals. Tidal waters with stronger flushing capacity are known to have less chance for algal bloom (Lucas et al., 2009; Qin and Shen 2021). Second, the vertical mixing layer is about 10 m at the deep channel (Yu et al., 2020) while the majority of shoals is less than this depth. If with same light and nutrient supply, there is a larger chance that phytoplankton grows and accumulates to a thinner water column in shoals. In another word, the shoal water is less "dispersive" in vertical direction. Third, nutrient released from seafloor is more accessible to phytoplankton in shallow water because vertical mixing is much strong. In contrast, the persistent pycnocline at the deep channel will limit the vertical nutrient flux from bottom to surface.

4.2 Dealing with data gaps

There are several major challenges in high-resolution simulation of water quality variables in both data-driven and numerical models. Success of numerical model relies on accurate prediction of local hydrodynamics and comprehensive parameterization of biogeochemical processes, while for data-driven model, the integrity of data itself (both

input forcing and target variable) is particularly important. One problem for satellite Chl-a data is substantial data gaps, especially in coastal waters because of terrestrial substances (such as minerals and humus) that are optically significant but do not covary with phytoplankton (Zheng and DiGiacomo, 2017). Taking VIIRS Chl-a data in Chesapeake Bay as an example, gaps are still noticeable (~20%) even after a 7-day blending. Data gaps in coastal regions tend to appear more frequently in a specific season due to seasonal atmospheric and turbidity condition. For instance, the bay-wide satellite data show persistent high data gaps in April and May, with gap percentage frequently exceeding 30% (Fig. 2).

It has been shown that DINEOF provides a reasonable estimation of the missing data (Fig. 4). One major advantage of DINEOF is that the interpolation does not need any prior knowledge of the correlation between data points (Beckers and Rixen, 2003). The cost is that a large amount of data points and record length at each data points are needed, which is not an issue for satellite data. DINEOF is therefore suitable for interpolating missing records in satellite data. Furthermore, the auto-detection of optimal number of EOF mode provides an effective way to determine how many modes to simulate in the data-driven model. Broken-stick method (Yu and Shen, 2020) suggests 33 modes are needed, which is much more than the optimal number determined by DINEOF. The DINEOF-based optimal mode selection is more suitable for the data-driven model considering the gappy nature of the satellite data. It is worth noting that extreme high values from the original data are generally underestimated by DINEOF (Fig. 5), which also contributes to the less variability of satellite Chl-a (Fig. 8).

4.3 Higher frequency simulation

For water quality variables in coastal waters, the dominant frequency could be hourly, daily, or monthly depending on how quickly the underlying processes respond to external forcings. For instance, nutrient level in the lower bay typically responds to wind or river discharge-induced perturbations (with dominant frequency of several days) in a much slower manner compared to temperature or salinity. While nutrient level is a key factor controlling the growth of phytoplankton, variations of Chl-a concentration are subject to a variety of factors including water temperature, vertical mixing, horizontal dispersion, light condition, and abundance of predators. These factors have a wide spectrum of frequencies ranging from hourly to monthly and need to be considered when simulating Chl-a. Ideally, hourly or daily simulations are preferred when modelling Chl-a. However, several limiting factors make such high-frequency simulations extremely difficult. First, despite daily frequency, satellite data have large gaps if not blended over a period of multiple days. About 600 over the available ~2000 days have data gaps less than 50%, meaning only 30% of satellite images cover half of the bay. Second, some external forcings are not available in daily frequency. For instance, nutrient loading data used for the model is available from USGS on a monthly basis. Using monthly nutrient input to simulate daily Chl-a variations will likely result in poorer model performance.

We conducted two simulations using monthly and daily Chl-a data and compared the model performance to that using 7-day average. Missing records in Chl-a data for these two simulations are also interpolated using DINEOF. For daily data, we only use those images with data gap less than 50%. Same forcings and model settings are applied for these two simulations. The model has the best performance using monthly data while it has the worst performance using daily data (Fig. 16-17). Monthly simulation has the

smallest RMSE and RE when averaged over the domain of interest (Fig. 16). Specifically, RMSE in the upper bay is largely reduced compared to the 7-day simulation (Fig. 15). For daily simulation, RMSE is also high in the upper bay, but the RMSE is smaller in tributaries including James and Potomac River, which results in a slightly smaller overall RMSE when compared with the 7-day simulation (Fig. 17). However, the RE in daily simulation is extremely high in several regions including areas near the bay mouth, the upper bay, and eastern shoals at 38N. The overall model performance is shown in Fig. 18. RE decreases from daily to 7-day and to monthly. While the monthly simulation has the best model performance, the monthly mean will omit a major portion of Chl-a variability. Using 7-day average for Chl-a simulation is a balance of the model performance and data representativeness.

5. CONCLUSIONS

Analysis of recently available high-resolution satellite data demonstrates the spatial heterogeneity of mean Chl-a concentration, its variability, and the peak timing. Patterns revealed by satellite data are consistent with that of long-term monitoring data at a limited number of stations, suggesting the reliability of the satellite data. Both satellite data and long-term monitoring data show that:

1. Both mean and variability of Chl-a are higher in shallow shoals than in the deep channel.
2. There are two seasonal peaks of Chl-a in the upper and middle bay, with spring peak near major river outflow and summer peak elsewhere.

Compared to low-frequency shipboard measurements at a limited number of monitoring stations, we believe satellite-derived Chl-a is suitable to serve as a reliable reference for assessment of numerical models performance.

We also implemente DINEOF into the existing data-driven model to address the issue of substantial gaps in satellite data. Sensitivity tests confirm that DINEOF is an appropriate way to fill data gaps with reasonable accuracy. The model is applied to simulate high spatial resolution satellite-based Chl-a concentration. The overall model performance for predicting both spatial and temporal variation of Chl-a is satisfactory. This study demonstrates the feasibility of data-driven model for high-resolution water quality simulations.

REFERENCES

Abbas, M. M., Melesse, A. M., Scinto, L. J., & Rehage, J. S. (2019). Satellite estimation of chlorophyll-a using moderate resolution imaging spectroradiometer (MODIS) sensor in shallow coastal water bodies: Validation and Improvement. *Water*, *11*(8), 1621.

Alvera-Azcárate, A., Barth, A., Parard, G., & Beckers, J.M. (2016). Analysis of SMOS sea surface salinity data using DINEOF. *Remote Sensing of Environment*, *180*, 137–145.

Beckers, J. M., & Rixen, M. (2003). EOF calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, 20(12), 1839–1856.

Du, J., & Shen, J. (2016). Water residence time in Chesapeake Bay for 1980–2012. *Journal of Marine Systems*, *164*, 101–111.

Feng, Y., Friedrichs, M. A. M., Wilkin, J., Tian, H., Yang, Q., Hofmann, E. E., Wiggert, J. D., & Hood, R. R. (2015). Chesapeake Bay nitrogen fluxes derived from a land-estuarine ocean biogeochemical modeling system: Model description, evaluation, and nitrogen budgets. *Journal of Geophysical Research: Biogeosciences*, *120*(8), 1666–1695.

Hagy, J. D., Boynton, W. R., Keefe, C. W., & Wood, K. V. (2004). Hypoxia in Chesapeake Bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. *Estuaries*, *27*(4), 634–658.

Harding, L., & Perry, E. (1997). Long-term increase of phytoplankton biomass in Chesapeake Bay, 1950-1994. *Marine Ecology Progress Series*, *157*, 39–52.

Harding, L. W., Itsweire, E. C., & Esaias, W. E. (1992). Determination of phytoplankton chlorophyll concentrations in the Chesapeake Bay with aircraft remote sensing. *Remote Sensing of Environment*, *40*(2), 79–100.

Harding, J., Mallonee, M. E., Perry, E. S., Miller, W. D., Adolf, J. E., Gallegos, C. L., & Paerl, H. W. (2016). Variable climatic conditions dominate recent phytoplankton dynamics in Chesapeake Bay. Scientific Reports, 6(1), 23773.

Hilborn, A., & Costa, M. (2018). Applications of DINEOF to satellite-derived chlorophyll-a from a productive coastal region. *Remote Sensing*, *10*(9), 1449.

Irby, I. D., Friedrichs, M. A. M., Friedrichs, C. T., Bever, A. J., Hood, R. R., Lanerolle, L. W. J., et al. (2016). Challenges associated with modeling low-oxygen waters in Chesapeake Bay: a multiple model comparison. *Biogeosciences*, *13*(7), 2011–2028.

Kemp, W. M., Boynton, W. R., Adolf, J. E., Boesch, D. F., Boicourt, W. C., Brush, G., et al. (2005). Eutrophication of Chesapeake Bay: Historical trends and ecological interactions. *Marine Ecology Progress Series*, *303*, 1–29.

Kuo, A. Y., & Neilson, B. J. (1987). Hypoxia and salinity in Virginia estuaries. *Estuaries*, *10*(4), 277–283.

Lee, Z., Carder, K. L., & Arnone, R. A. (2002). Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters. *Applied Optics*, 41(27), 5755–5772.

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

Malone, T. C., Conley, D. J., Fisher, T. R., Glibert, P. M., Harding, L. W., & Sellner, K. G. (1996). Scales of nutrient-limited phytoplankton productivity in Chesapeake Bay. *Estuaries*, *19*(2), 371–385.

Martin, A. P., Richards, K. J., Bracco, A., & Provenzale, A. (2002). Patchy productivity in the open ocean. *Global Biogeochemical Cycles*, *16*(2), 9-1-9–9.

Muller, A., Muller, D.L., 2015. Forecasting future estuarine hypoxia using a wavelet based neural network model. *Ocean Modelling*, 96, 314-323

Murphy, R. R., Kemp, W. M., & Ball, W. P. (2011). Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. *Estuaries and Coasts*, *34*(6), 1293–1309.

Orth, R. J., Williams, M. R., Marion, S. R., Wilcox, D. J., Carruthers, T. J. B., Moore, K. A., et al. (2010). Long-term trends in submersed aquatic vegetation (SAV) in Chesapeake Bay, USA, related to water quality. *Estuaries and Coasts*, *33*(5), 1144–1163.

Qin, Q., Shen, J. (2021). Applying transport rate for quantifying local transport conditions in estuarine and coastal systems. *Journal of Marine Systems*, 218, 103542.

Sanford, L. P., Suttles, S. E., & Halka, J. P. (2001). Reconsidering the physics of the Chesapeake Bay estuarine turbidity maximum. *Estuaries*, *24*(5), 655–669.

Scardi, M., & Harding, L. W. (1999). Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling*, *120*(2), 213–223.

Shen, J., & Wang, H. V. (2007). Determining the age of water and long-term transport timescale of the Chesapeake Bay. *Estuarine, Coastal and Shelf Science*, *74*(4), 585–598.

Soro, M. P., Yao, K. M., Kouassi, N. L. B., Ouattara, A. A., & Diaco, T. (2020). Modeling the spatio-temporal evolution of Chlorophyll-a in three tropical rivers Comoé, Bandama, and Bia Rivers (Côte d'Ivoire) by artificial neural network. *Wetlands*, *40*(5), 939–956.

Tango, P. J., & Butler, W. (2008). Cyanotoxins in Tidal Waters of Chesapeake Bay. *Northeastern Naturalist*, *15*(3), 403–416.

Testa, J. M., Li, Y., Lee, Y. J., Li, M., Brady, D. C., Di Toro, D. M., et al. (2014). Quantifying the effects of nutrient loading on dissolved O2 cycling and hypoxia in Chesapeake Bay using a coupled hydrodynamic–biogeochemical model. *Journal of Marine Systems*, *139*, 139–158.

Xiong, J., Shen, J., Qin, Q., & Du, J. (2021). Water exchange and its relationships with external forcings and residence time in Chesapeake Bay. *Journal of Marine Systems*, 215, 103497.

Yang, M., Khan, F. A., Tian, H., & Liu, Q. (2021). Analysis of the monthly and spring-neap tidal variability of satellite chlorophyll-a and total suspended matter in a turbid coastal ocean using the DINEOF method. *Remote Sensing*, *13*(4), 632.

Yu, X., & Shen, J. (2021). A data-driven approach to simulate the spatiotemporal variations of chlorophyll-a in Chesapeake Bay. *Ocean Modelling*, *159*, 101748.

Yu, X., Shen, J., & Du, J. (2020). A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example. *Water Resources Research*, *56*(9), e2020WR027227.

Zhang, Q., Fisher, T.R., Trentacoste, E. M., Buchanan,C., Gustafson,A.B., Karrh, R., Murphy, R.R., Keisman, J., Wu, C., Tian, R., Testa, J.M., &Tango, T. (2021). Nutrient limitation of phytoplankton in Chesapeake Bay: Development of an empirical approach for water-quality management. *Water Research*, 188, 116407.

Zheng, G., & DiGiacomo, P. M. (2017). Remote sensing of chlorophyll-a in coastal waters based on the light absorption coefficient of phytoplankton. *Remote Sensing of Environment*, *201*, 331–341.

Zheng, G., Stramski, D., & DiGiacomo, P. M. (2015). A model for partitioning the light absorption coefficient of natural waters into phytoplankton, nonalgal particulate, and colored dissolved organic components: A case study for the Chesapeake Bay. *Journal of Geophysical Research: Oceans*, *120*(4), 2601–2621.

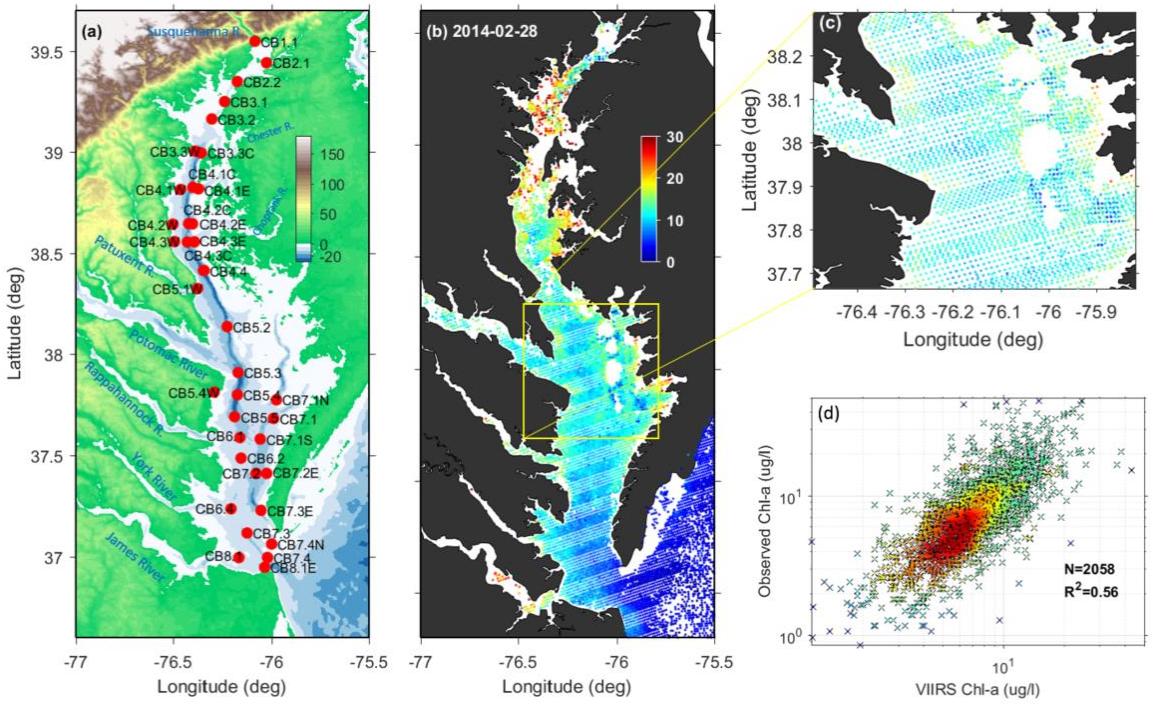

Figure 1: (a) Chesapeake Bay Program monitoring stations in the mainstem. (b) A sample snapshot of VIIRS Chl-a data (unit in ug/l) on Feb-28, 2014. (c) A zoom-in view of the satellite data. (d) Comparison between satellite data and shipboard measurements.

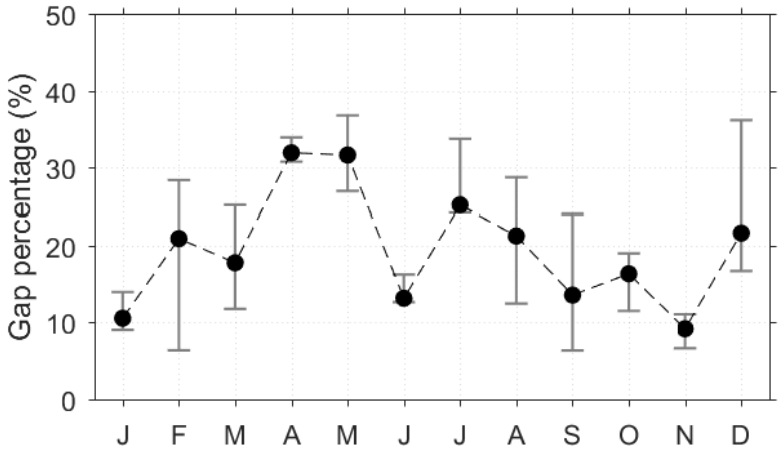

Figure 2: Gap percentage of the 7-day blended Chl-a satellite data. The gap percentage for every 7 days is calculated as the number of grids with valid value divided by the total number of grids (i.e., 4813).

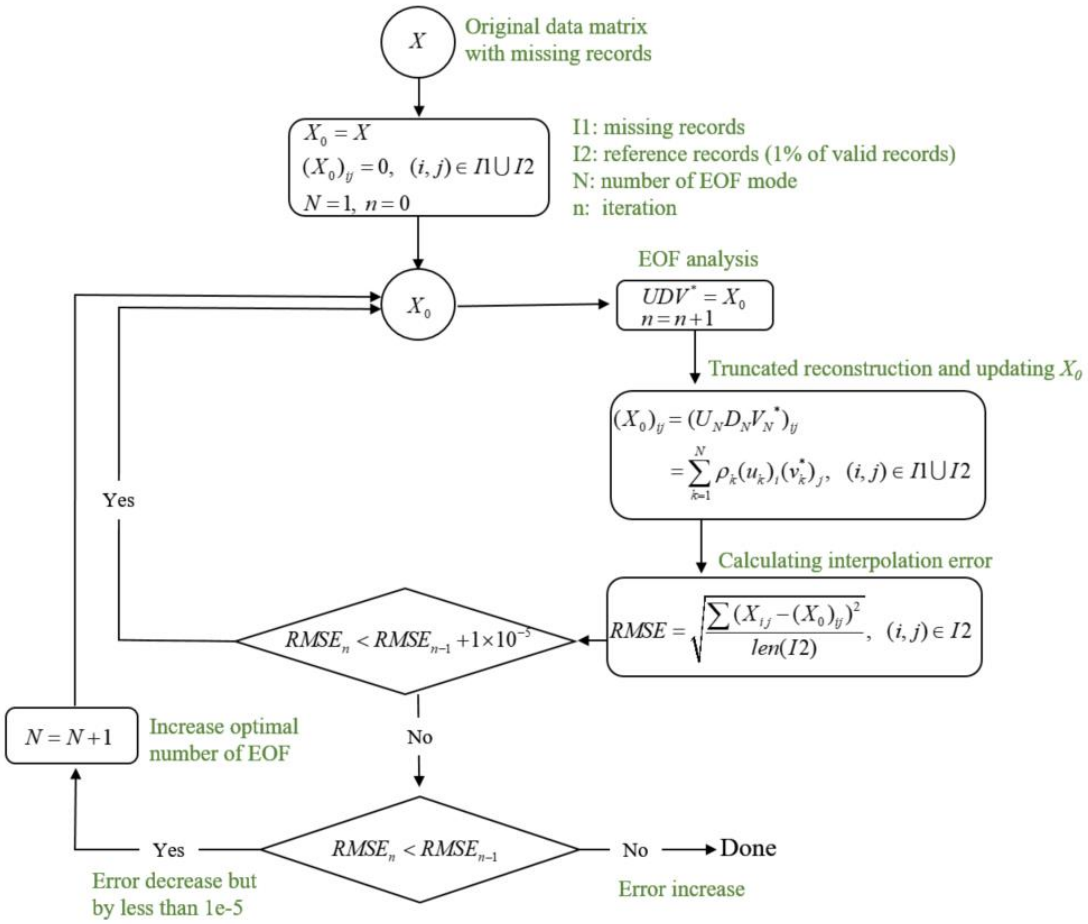

Figure 3: A diagram showing the algorithm of DINEOF. The algorithm is comprised of two loops, the inner loop to estimate the missing value with the given number of EOF modes and the other loop to determine the optimal number of EOF modes.

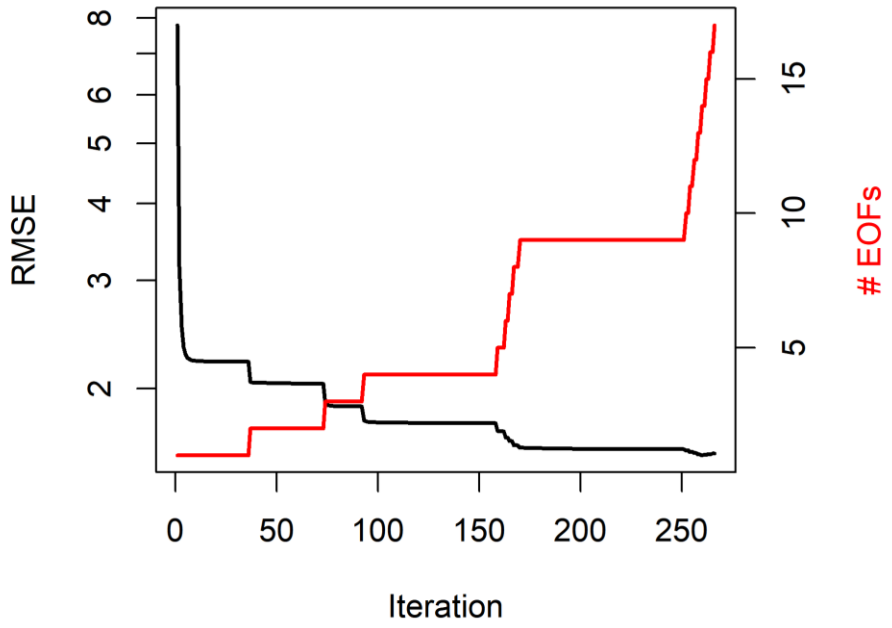

Figure 4: Determination of the necessary EOF modes to include in the data-driven model based on DINEOF analysis.

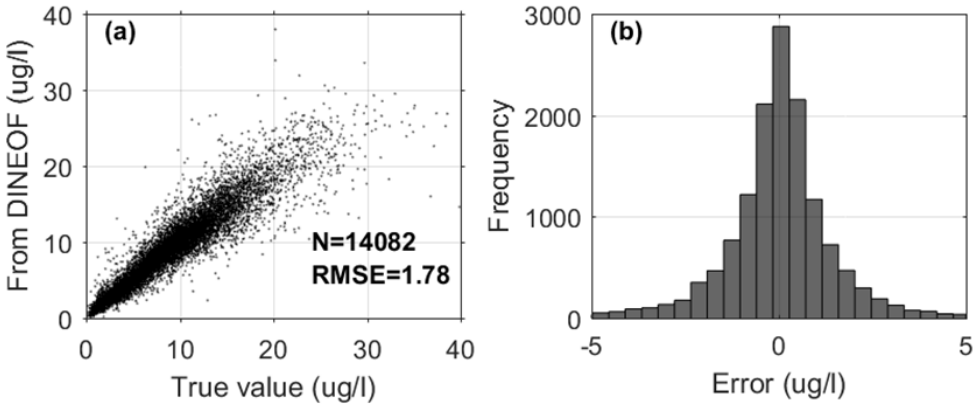

Figure 5: Performance of DINEOF in estimating the “missing” data. The “missing” data (i.e., the true value) are manually and randomly selected from the existing satellite data.

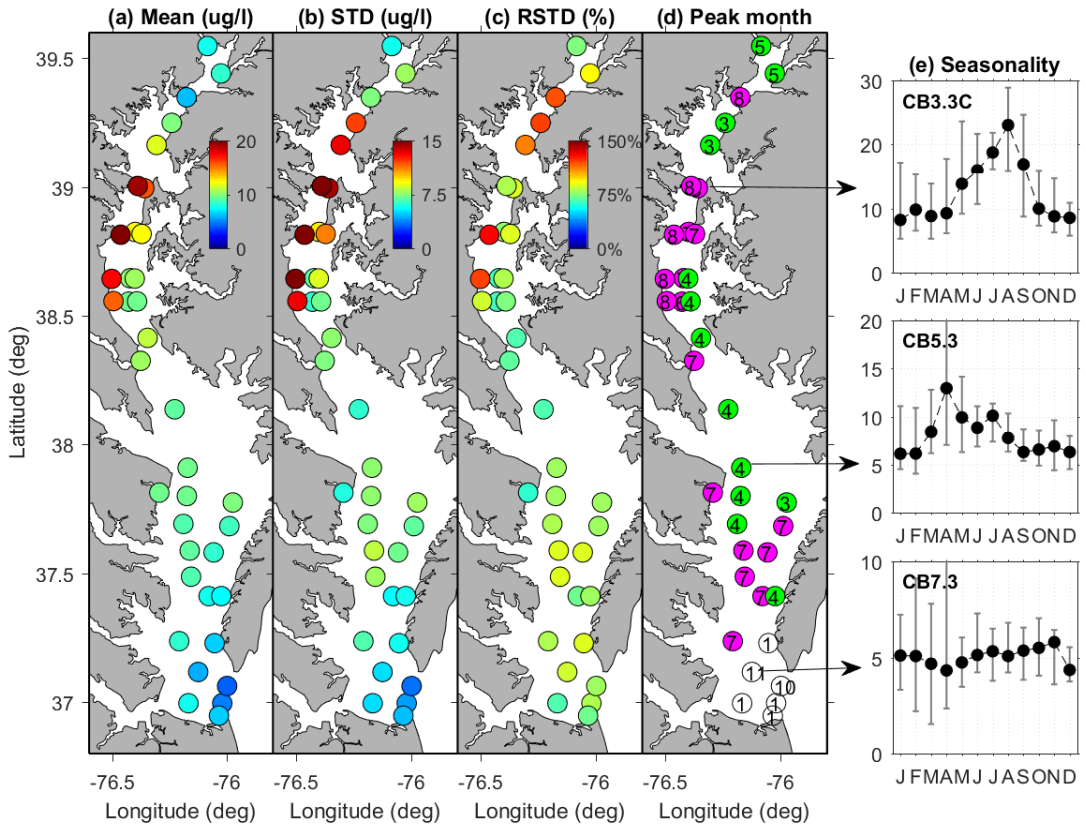

Figure 6: (a-d) Mean, standard deviation, relative standard deviation, and month of peak Chl-a at mainstem stations. (e) Three example stations showing different seasonality in different regions of the bay.

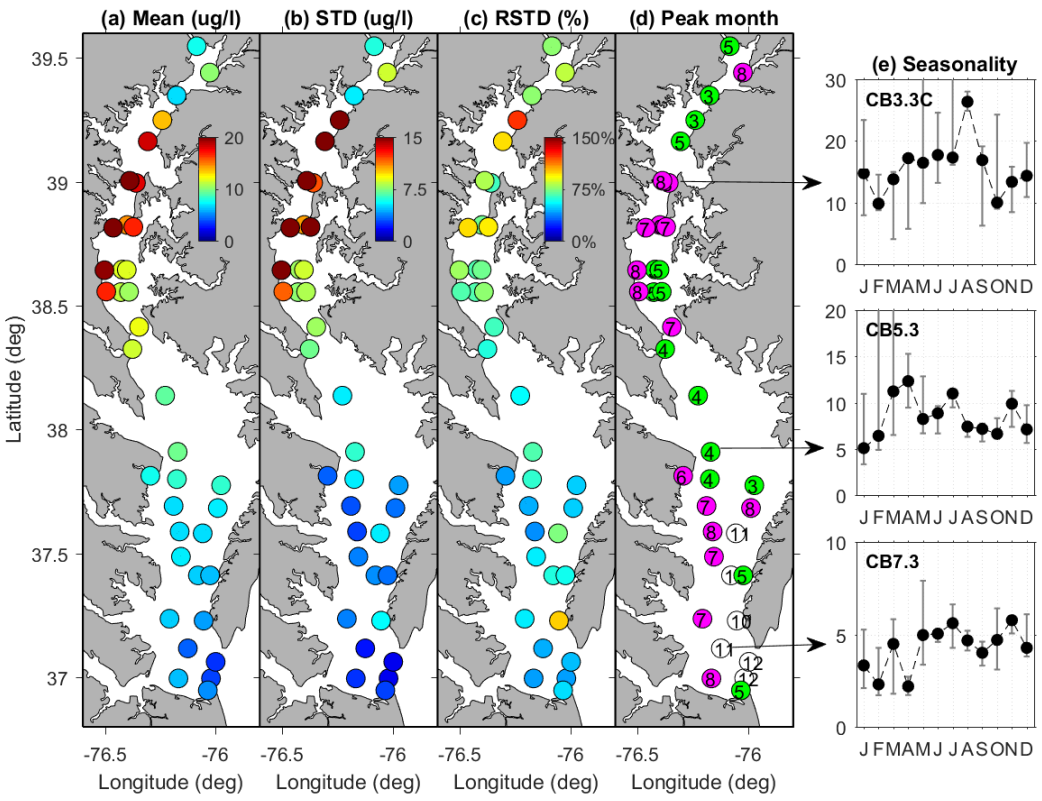

Figure 7: Same as Fig. 5 but using the observation data over a shorter period from 2012 to 2018, which is the time span of the satellite data.

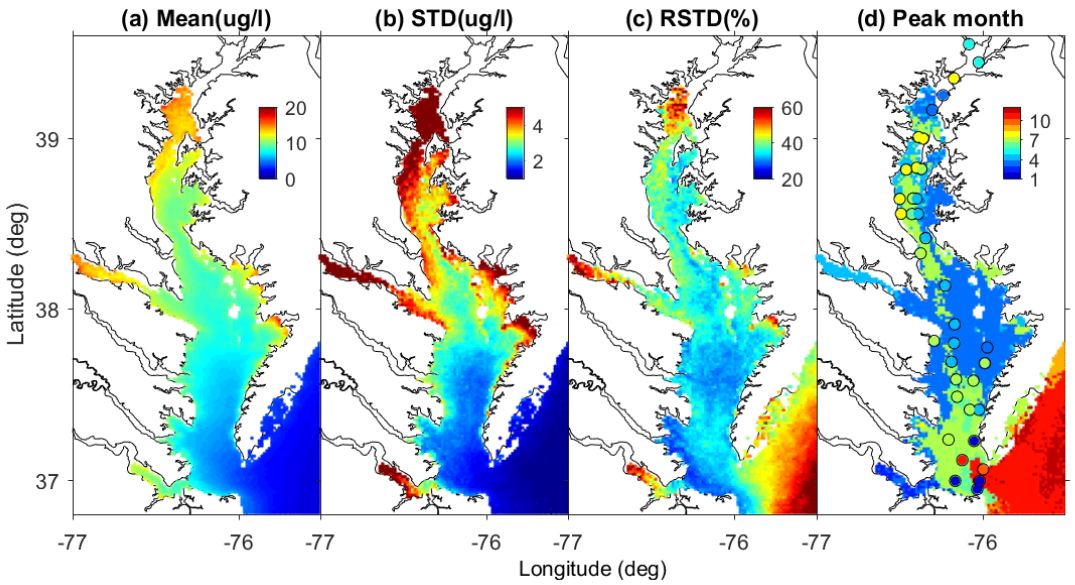

Figure 8: (a-d) Mean, standard deviation, relative standard deviation, and month of peak Chl-a revealed from the satellite data. In (d) the peak month based on shipboard measurements are shown with black circles filled with color.

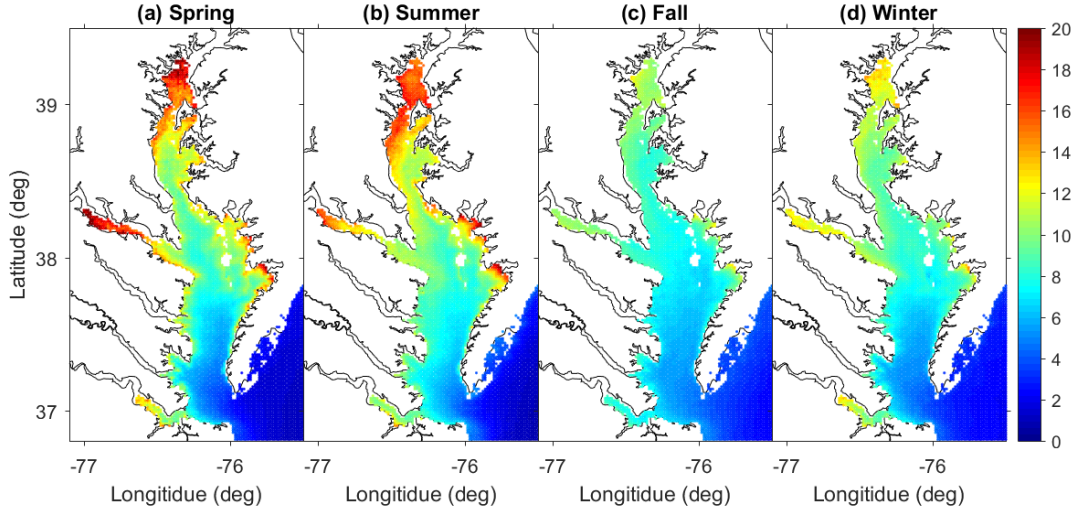

Figure 9: Seasonal mean of Chl-a from the satellite data.

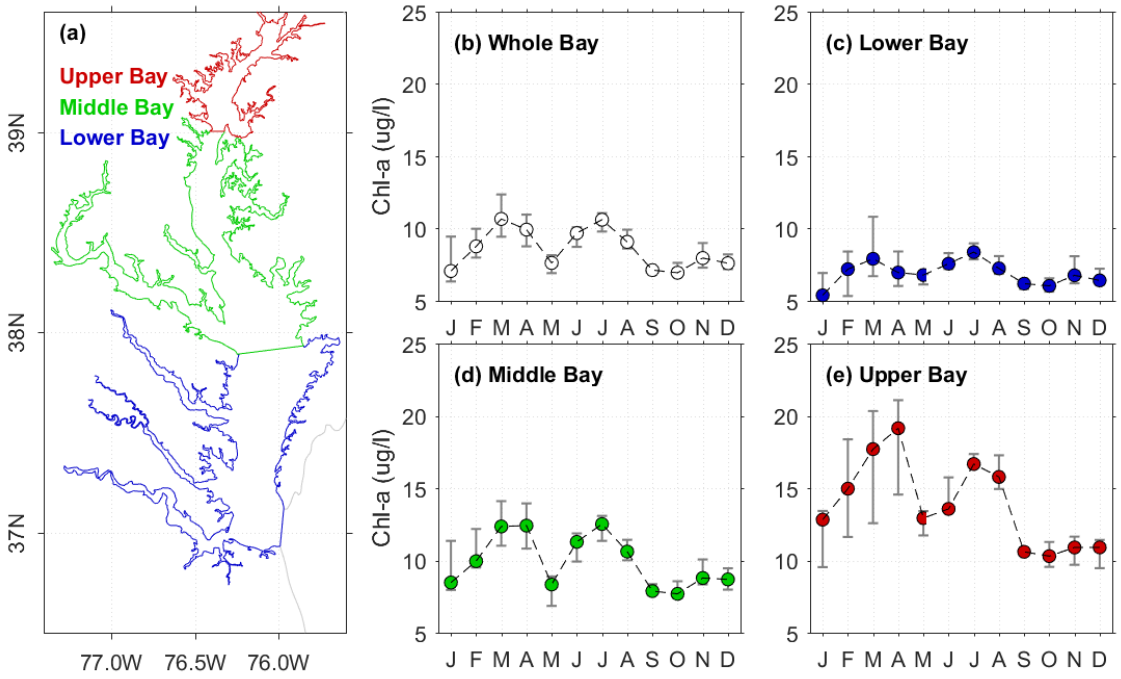

Figure 10: Seasonality of mean Chl a averaged over the entire bay and in different sub-regions of the bay, based on 2011-2018 satellite data. The different sub-regions are marked with colored polygons in the left panel.

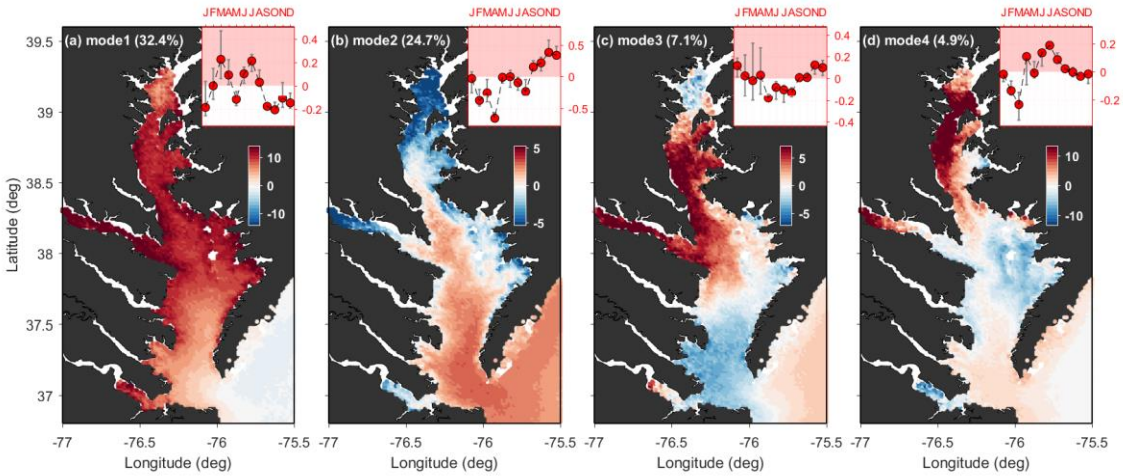

Figure 11: The spatial map and seasonal variations of the first 4 EOF modes. The up-right insets show the seasonality of the temporal variations, with error bars indicating the 25-75 percentiles and solid circles indicating the median value.

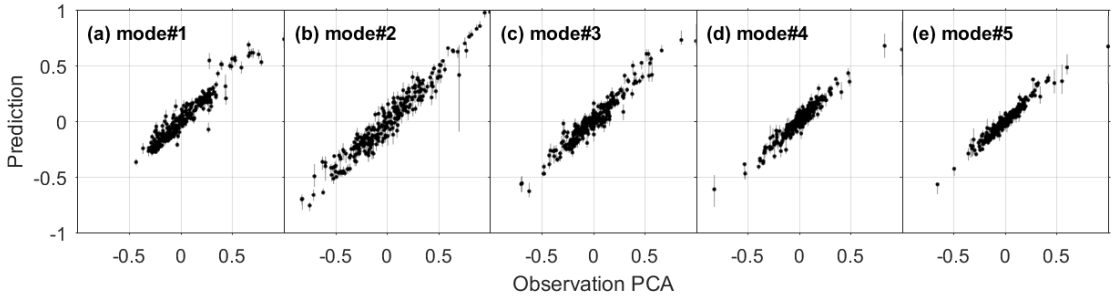

Figure 12: Model performance with the training dataset. The black dots are the median of the 50 predictions based on 50 neural network models. The error bars are for the 25 and 75 percentiles of the prediction; the error bars indicate the uncertainty associated with the neural network model.

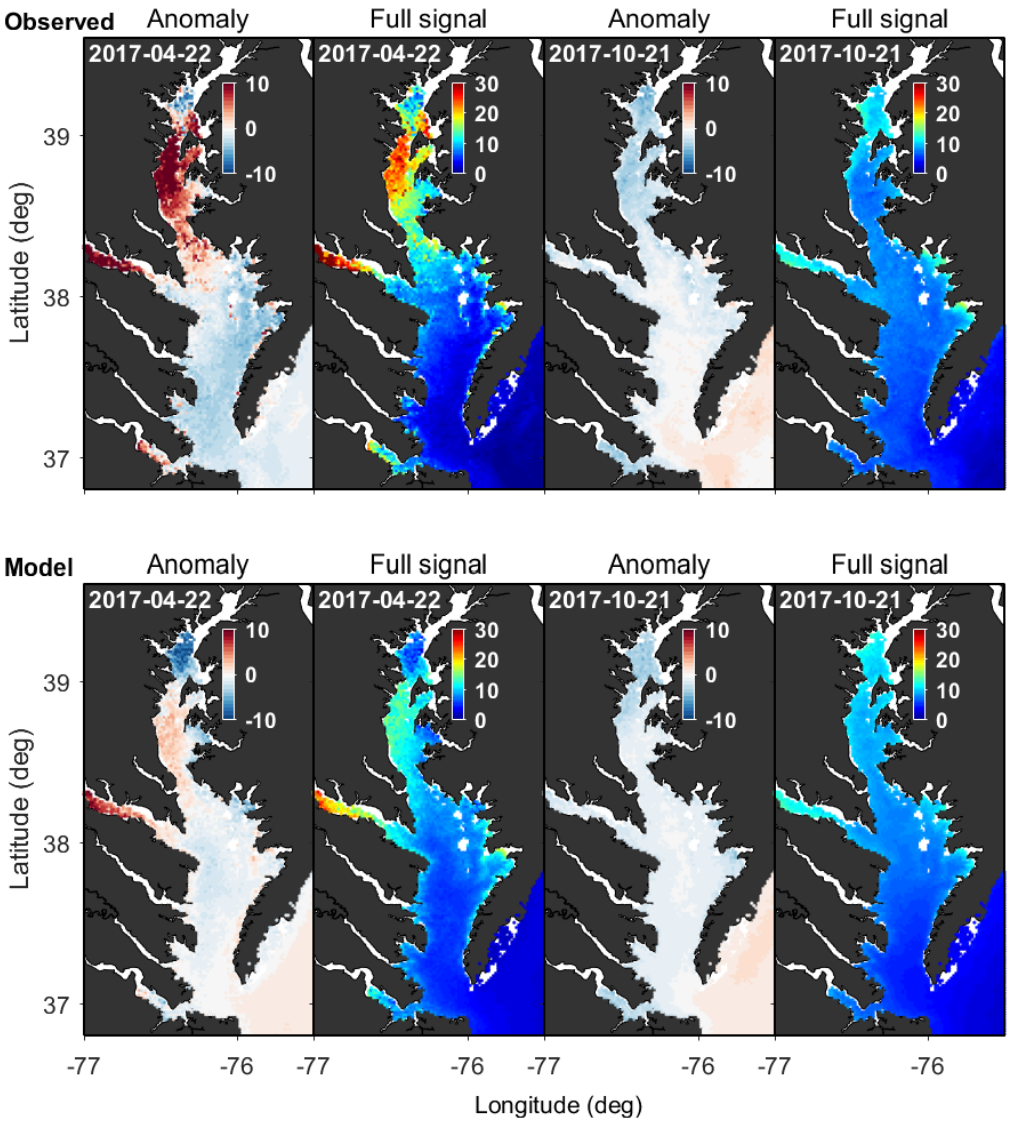

Figure 13: Comparison of satellite observed (upper panels) and data-driven model predicted (lower panels) Chl-a concentration at selected two dates during the testing period. The anomaly is the deviations from the long-term mean (see Fig. 8a for the long-term mean).

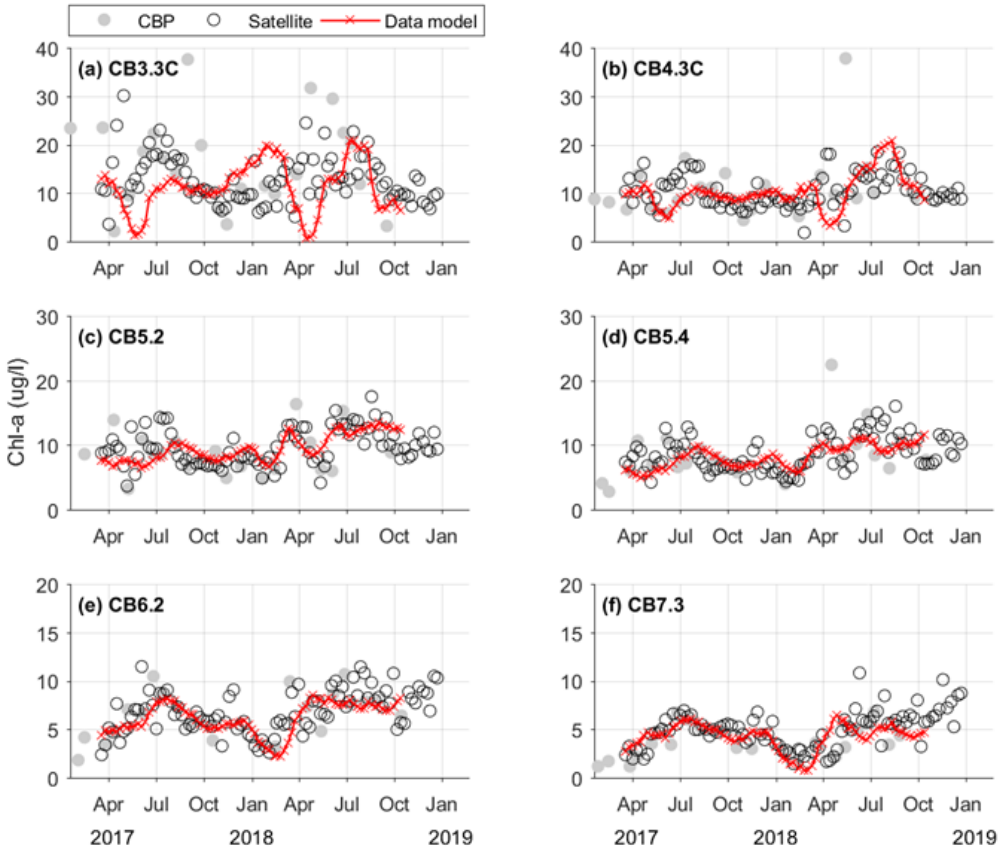

Figure 14: Comparison between model predicted Chl-a, satellite data, and *in situ* measurements from Chesapeake Bay Program (CBP).

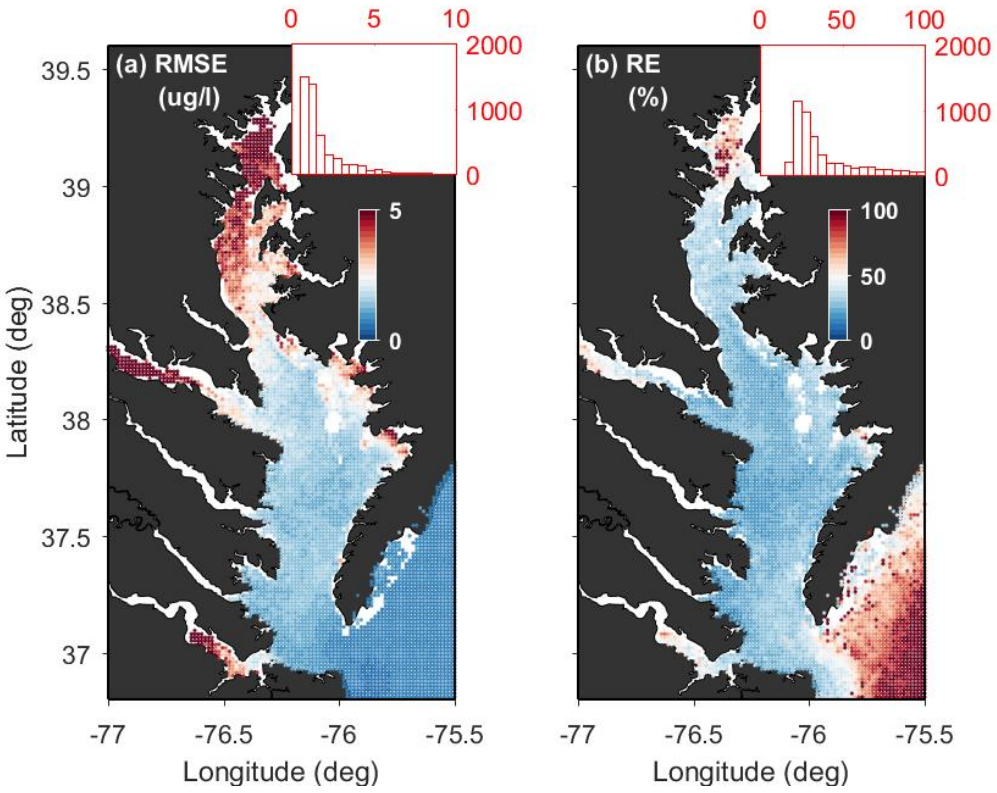

Figure 15: Performance of the data-driven model indicated by the calculated root mean square error (RMSE) and relative error (RE) for each grid point. The top-right insets are the histograms of RMSE (or RE) over the 4813 grids.

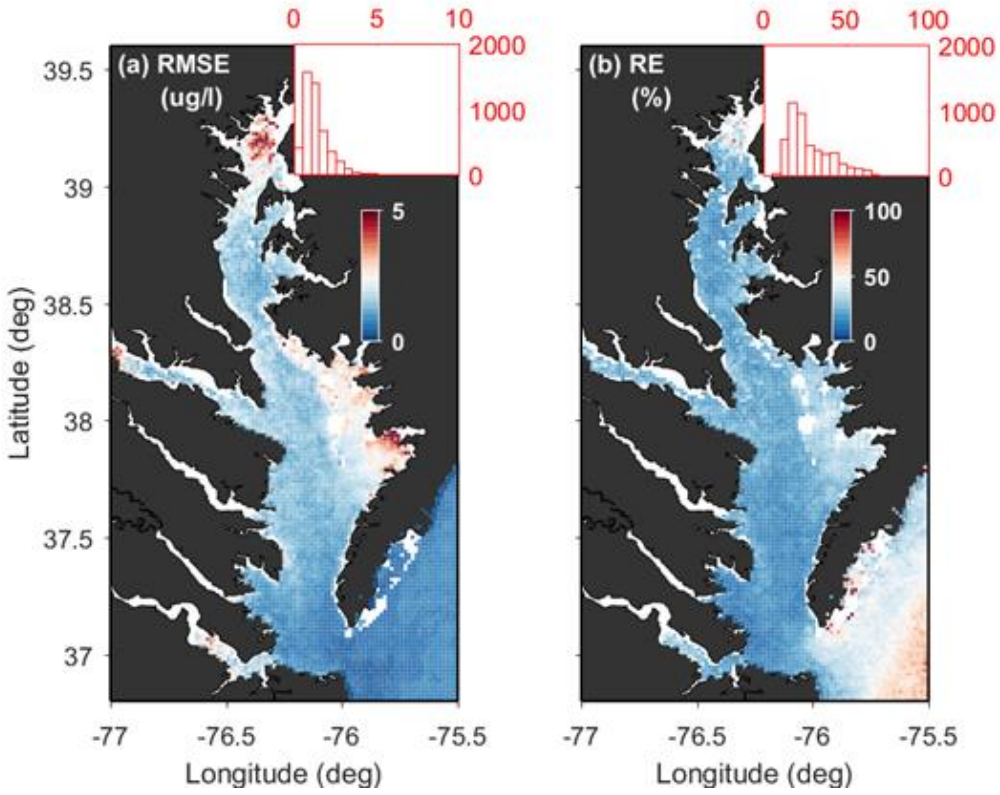

Figure 16: Same as figure 15, but for a simulation with monthly-averaged satellite data.

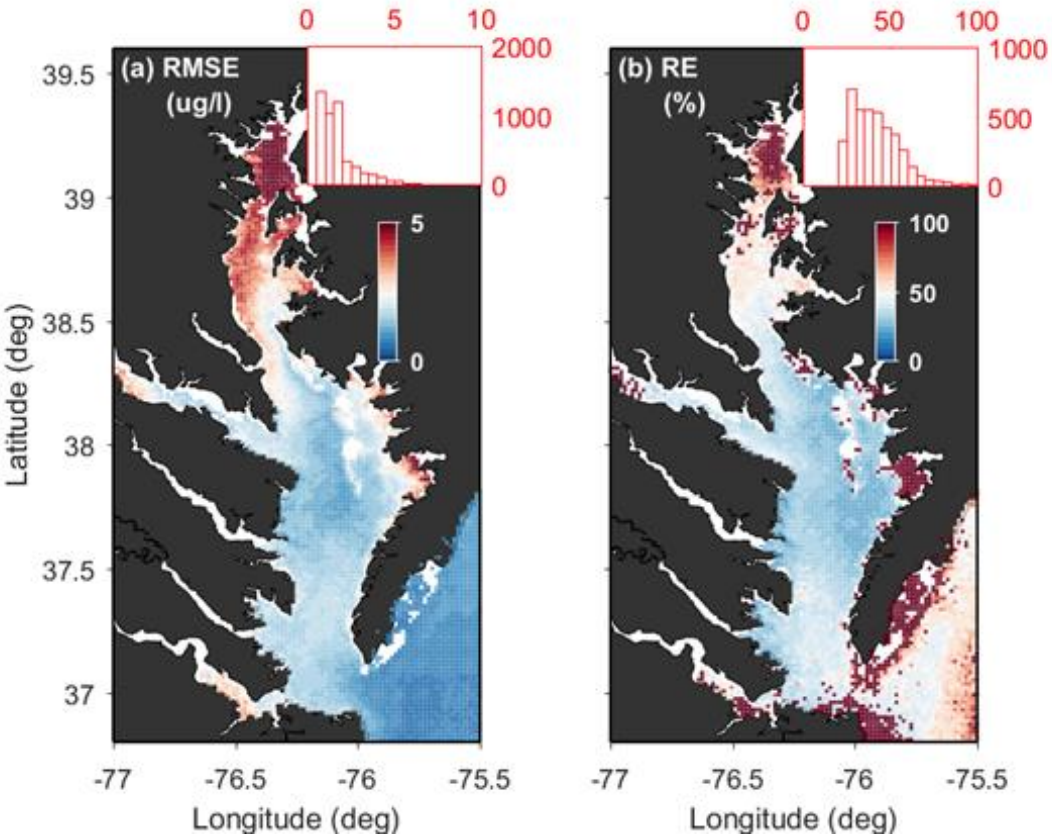

Figure 17: Same as fig. 15 but for simulation with daily satelllite data. Only those days with data gap <50% are used. The data gaps are interpolated with DINEOF.

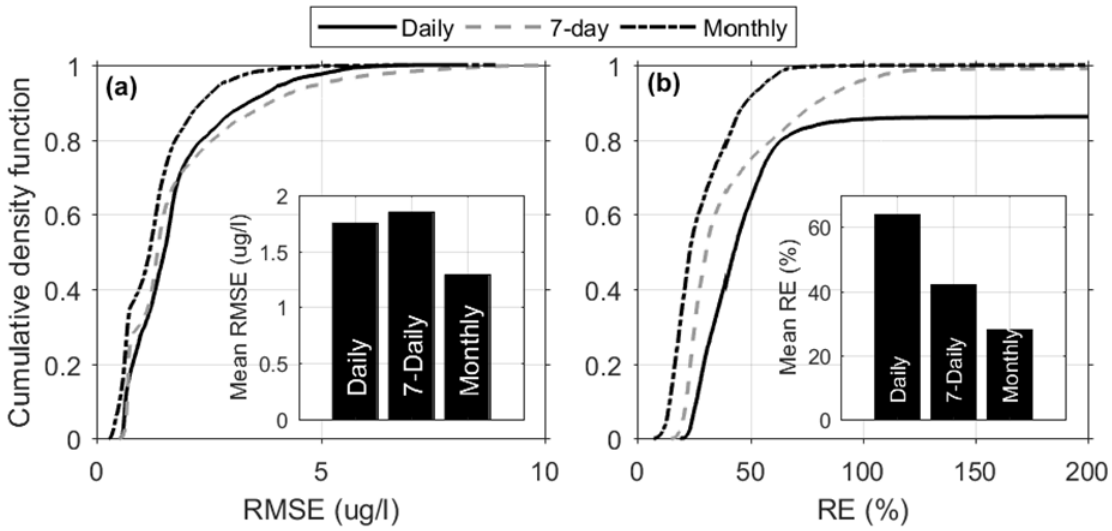

Figure 18: Comparison of model performance with cumulative density function plots for the three simulations (with daily, 7-day averaged, and monthly-averaged data).

CHAPTER 4. AN INVERSE APPROACH TO ESTIMATE BACTERIAL LOADING INTO AN ESTUARY BY USING FIELD OBSERVATIONS AND RESIDENCE TIME

Published in *Marine Environmental Research* (2021, 166, 105263)

Abstract: Pathogens, whose abundance is often measured by the concentration of fecal indicator bacterium, is listed as the top cause of waterbody impairments in the United States. An accurate estimation of the bacterial loading from watershed is thus fundamentally important for water quality management. Despite advances in watershed modeling, accurate estimation of bacterial load is still very challenging due to large uncertainties associated with bacterial sources, accumulation, and removal in the watershed. We introduce an inverse method using field-measuredbacterial concentrations and numerical model-calculated residence time to estimate the bacterial loading from the drainage basin. In this method, an estuary is divided into multiple segments. Water and bacterial fluxes between neighboring segments are computed from a set of linear equations derived based on mass balance equation and the relationship between residence time and water fluxes. Loading to each segment can then be estimated by combining the computed water fluxes and observed bacterial concentrations. The approach accounts for seasonal and interannual variations in hydrodynamics due to tide, river discharge, and estuarine circulations. The method was applied to Nassawadox Creek, a sub-estuary of Chesapeake Bay, where *fecal coliform* concentrations at 46 stations were continuously monitored. The method is verified by the high consistency between estimated loading and presumably known input loading in numerical experiments with either constant or time-

varying input loadings. With sparse observational data, the inversely estimated loadings agree well with the loadings from a previously calibrated watershed model, demonstrating the reliability of the method. The inverse approach can be used to cross-check the result of watershed models and assess changes in watershed condition. The method is also readily applicable to other types of materials, such as inorganic nutrients.

Keywords: residence time; bacterial modeling; loading assessment; fecal coliform; coastal water; inverse model

1. INTRODUCTION

Pathogen violation is one of the primary causes of water impairments in the United States. According to the United States Environmental Protection Agency (USEPA), a total of 9,874 water bodies was listed as impaired due to pathogen violation in 2016 (USEPA, 2019). As it is often infeasible to measure the abundance of all pathogens (e.g., viruses, bacteria, and protozoa) in the water due to their large diversity and low concentration (which means a large volume of water is required to measure the pathogen concentration), the abundance of indicator bacteria (e.g., *fecal coliform*, *Escherichia coli* [*E. coli*], and *Enterococci*) is practically used as water quality metrics (Brauwere et al., 2014; Jang et al., 2017). Fecal coliform (FC) is one of the commonly used indicator bacteria groups for evaluating microbiological water quality, especially in shellfish growing waters. Environmental agencies keep monitoring the concentrations of the indicator bacteria to ensure the quality of drinking water, waterbody's suitability for recreational activities and/or shellfish harvest. To mitigate the pathogen-induced impairment, total maximum daily load projects have been extensively conducted in the

past decades (e.g., He et al., 2007; Cho et al., 2016). Information regarding the source and abundance of indicator bacteria is, therefore, fundamentally important for environmental assessment and water quality management.

Despite the latest advances in watershed models and water quality models (e.g., Shen et al., 2005a; Arnold et al., 2012; Sobel et al., 2017; Plew et al., 2018), uncertainties in bacteria loading estimation are still a major factor limiting efficient water quality assessment. The uncertainties partially result from source diversity and the episodic nature of bacterial loading. Bacteria discharges are from point sources or non-point sources (also called diffuse sources). While point sources are relatively easy to identify and monitor under normal conditions, non-point sources are hard to accurately quantify (Brauwere et al., 2014). Point sources are permitted pollutant loads derived from individual sources (e.g., wastewater treatment plants) and discharged at designated locations. However, large uncertainties may also exist for point sources. For instance, FC concentration from wastewater treatment plants is generally controlled but unpredictable during storm events as a result of a shortened circuit of raw sewage (Passerat et al., 2011). Nonpoint sources are from various sources over a relatively large land area (e.g., manure application on agricultural land, and wildlife and pets' feces excretion), which are the dominant pollutant sources in most cases. Accumulation or removal of FC is controlled by the hydro-biogeochemical processes during the transport from land to receiving water. Land-deposited FC loading can be determined by a watershed model that incorporates the surface runoff, bacterial accumulation rate on land, land-use, and the density of contributors (mainly animals) on each land-use (e.g., Servais et al., 2007; Ferguson et al., 2008). Typically, the animal density and bacteria accumulation rate are

assumed constant for each land use. However, in reality, the density of the contributors is spatially and temporally variable, and hard to predict. In addition, direct input from waterfowls (e.g., geese, duck on the open water) is almost impossible to accurately quantify. To obtain a reasonable loading in many total maximum daily load projects, watershed models are usually calibrated through iterative trial-and-error learning based on the comparison between observations and model results (Shen et al., 2005b). Such processes are time-consuming and subjective. Therefore, an alternative and efficient way to estimate the bacterial loading is of great interest to water quality management.

Several alternative methods to estimate the loading by utilizing the monitoring data have been proposed by previous studies. Shen et al. (2006) treated the non-point source as parameters in a three-dimensional model and used a modified Gauss-Newton method for optimal estimation of the loads. By integrating the limited observational data, Bayesian methods have been used to estimate the unknown parameters in a watershed model or water quality model (e.g., Gronewold et al., 2009; Shen and Zhao, 2009; Shen and Zhao, 2010; Chen et al., 2012). These approaches are, however, rarely used to estimate long-term time-varying loadings, because computation cost is high using 3D model, while Bayesian methods are difficult to apply to estuaries with complex geometry and hydrodynamic transport processes. Shen et al. (2005b) used a tidal prism model to calculate the FC maximum daily load from the watershed to meet water quality criteria based on a trial-and-error method for a coastal embayment. Tidal prism model is a simple model compared to two-, three-dimensional hydrodynamic models and can be efficiently used for coastal embayments (Kuo et al., 2005). Basic assumptions when applying a tidal prism model are: (i) the tide rises and falls simultaneously throughout the system, (ii) the

system is in hydrodynamic equilibrium, and (iii) the segment with a length less than local tidal excursion is completely mixed during the high tide (Kuo et al., 2005). These assumptions are, however, usually invalid for coastal waters that are usually not well mixed. Strict segmentation is also required in a tidal-prism model, which limits its applicability to large estuaries with complex geometry.

Here, we present a new method to inversely estimate the bacterial loading based on observed bacterial concentrations and calculated residence time. The proposed method is derived based on bacterial and water mass balance, and the relationship between residence time and exchange fluxes. Its applicability is not limited by the strict assumptions as required in a tidal-prism model. The method is validated by an application for Nassawadox Creek, a sub-estuary of Chesapeake Bay, where FC concentrations at 46 stations have been monitored for multiple years. The new approach is applicable to other coastal systems, facilitating the local or regional water quality management. The advantages and disadvantages of the method are also discussed.

2. METHOD

2.1 Inverse method to estimate loading

When an estuary is divided into multiple segments, within a given segment *i*, the subtidal mass balance equation for a substance is:

$$\frac{dC_iV_i}{dt} = L_i + \sum_{j=neighbors} Q_{ji}C_j - \sum_{j=neighbors} Q_{ij}C_i - K_iV_iC_i \qquad (1)$$

where C is the concentration, V is the volume, Q is the flux, and K is the net removal rate (including die-off, settling, resuspension, etc.) and the subscript index denotes the segment number. Q_{ij} is outflux from segment i to neighboring segment j, and Q_{ji} is influx from neighboring segment j to segment i. From Equation (1), one can inversely estimate the loading L_i if given the value of C, V, Q, K. As C, V, and K can be obtained from measurement, the remaining question is how to obtain the water fluxes between segments.

The effective exchange fluxex between segments can be calculated from the mean residence time (τ), a bulk transport timescale characterizing the overall exchange between a domain of interest and the adjacent waters (Delhez, 2006; Du et al., 2018). The effective outflux form each segment i ($Q_{i,out}$) can be computed as the ratio of volume to the mean residence time.

$$Q_{i,out} = \sum_{j=neighbors} Q_{ij} = V_i / \tau_i \quad (2)$$

where $Q_{i,out}$ is the sum of outflux from segment i to all its neighboring segments. A smaller residence time corresponds to a larger $Q_{i,out}$ and stronger flushing. The equation resembles the water mass balance equation using the flushing time or e-folding time (Takeoka, 1984; Monsen et al., 2002). E-folding time is the time needed for material concentration to decrease to the e^{-1} (~0.37) of its initial value and is equal to the mean residence time for a well-mixed system. For a well-mixed system under a steady-state condition, Equation (2) can be derived (see Appendix A1). It is also verified for time-dependent problems (Xiong et al., 2021) and applicable for quasi-steady-state conditions, e.g., under which the hydrodynamics can be regarded as steady after removing high-

frequency components (e.g., tidal signal). Note that the flux is not necessarily equal to the surface outflux or bottom influx as in a typical two-layer estuarine circulation. Instead, they are effective exchange fluxes between neighboring segments. Depending on the structure of segment connectivity, $Q_{i,\,out}$ is comprised of one or multiple components. Taking the simple case in Fig. 1a as an example, for segment 2, $Q_{2,out}=Q_{21}+Q_{23}$. For the three boxes in Fig. 1a, there are four unknown Q (i.e., Q_{12}, Q_{21}, Q_{23}, and Q_{32}). To solve them, another two equations are needed, which can be derived from mass balance of water (i.e., the total influx equal to the total outflux).

$$Q_{i,in} = Q_{i,out} \qquad (3)$$

where $Q_{i,in}$ includes not only the flux between segments but also the river discharge,

$$Q_{i,in} = R_i + \sum_{j=neighbors} Q_{ji} \qquad (4)$$

When there are n segments, there will be (n-1)×2 unknown fluxes and (n-1)×2 equations. This applies to not only the case with aligned segments but also to cases with more complex segmentation configurations (e.g., a segment connected by a number of neighboring segments; Fig. 1b). The fluxes can be obtained by solving the multiple (2n-2) linear equations as long as the residence time and river discharge for each segments are known. These equations can be solved by matrix computation. The multiple linear equations can be expressed as,

$$AX = b \qquad (5)$$

where *X* is a vector of the unknown water fluxes, *A* is the matrix denoting the coefficient with values of either 1 or -1, and b is the right-hand side term comprised with values of either V/τ or river discharge.

After obtaining the fluxes, with known *C*, *K*, *V*, one can then calculate the loading based on Equation (1). The workflow for the application of this inverse method is illustrated in Fig. 2. To obtain the fluxes, a numerical hydrodynamic model is usually recommended to calculate the residence time, even though the residence time can also be calculated through other methods, such as geochemical tracer method (Bouchaou et al., 2008), tidal prism method (Sheldon and Alber, 2006), freshwater fraction method (Huang and Spaulding, 2002), isohaline-based salt exchange method (MacCready, 2011). Note that when using tidal prism-based methods, one can estiamte downstream flux at the segment's downstream boudanry, and the other fluxes can be obtratined if computation follows the order from upstream to downstream. The method and the workflow are straightforward, but cautions should be taken regarding the temporal variability of residence time and fluxes due to the combined effect of wind, river flow, and open boundary conditions. The uncertainty associated with the time-varying residence time will be further discussed in Section 4.1.

2.2 Estimate the net bacterial removal rate

One key parameter in the above method is the net removal rate, *K*. It depends on multiple factors including water temperature, salinity, suspended sediment concentration, and solar radiation. It can also be determined using observed spatial distribution of bacterial concentration from upstream to downstram and calculated transport timescale, following the method proposed in Du et al. (2020). Considering an exponentially

decreasing trend of bacterial concentration from its release location to downstream, K is calculated as

$$K = \frac{-\beta L}{\phi} \quad (6)$$

where the β is the slope between the logarithm of bacterial concentration and distance from the upstream, and ϕ is the transit time a water parcel needs to move over the distance of L from upstream to downstream. The transit time can be easily calculated as the freshwater age at a given downstream boundary, with age tracer being continuously released at the headwater of the river.

2.3 Application and verification of the method in a realistic estuary

To validate the method, we applied the method to a realistic estuary, Nassawadox Creek, a sub-estuary located on the eastern bank of lower Chesapeake Bay. It is 8 km long from its mouth to the head, with width varying from 300 m in the upper reach to 900 m in the lower reach and a drainage area of 76 km^2. The mainstem is joined by several smaller tributaries, forming a small estuarine system. Using the inverse modeling method, the loading discharging into each tributary can be estimated systematically; the application can thus serve as a good example to demonstrate the capability of the inverse method. More importantly, the Department of Shellfish Sanitation of Virginia has been monitoring the FC concentrations at 46 stations for almost every month since 1985. We used the data between 2007 and 2012. A total of 57 cruise surveys (nearly monthly) were conducted during this period. The FC concentrations in water samples collected during this period were measured based on the membrane filtration method with mTEC as the culture media, which improves the accuracy compared to the traditional most-probable-

number statistical method (Grant et al., 1997). This is an excellent dataset to test the proposed inverse approach.

The estuary was divided into 12 segments (Fig. 3b). We used a validated hydrodynamic model to compute the mean residence time for each segment (Du et al., 2020). To verify the inverse approach, three methods were used. First, we utilized the intermittent observation at the 46 stations inside the Nassawadox Creek and estimated the bacterial loading. The estimated loadings were compared with results from a previously-calibrated watershed model that was also applied for this estuary (Shen et al., 2005b). Second, a set of numerical experiments with constant FC loadings were conducted and we used the modeled FC output as "observed FC" to inversely estimate the loading, which was then compared with the presumably known loading (i.e., the constant loading input to the 3D numerical model). This approach is often used to validate an inverse model (Shen et al., 2006). Finally, the loading were calculated based on a realistic model run with temporally varying FC loading (from a watershed model). To mimic *in situ* sampling, bacterial concentration at the middle of eachsegment was extracted from model outputs for every 30 days. This experiment is to determine whether the inverse method is valid when estimating loading based on concentrations at a limited number of sampling stations by examining the deviations of the estimated loadings from the true loadings.

2.3.1 Hydrodynamic model

The Environmental Fluid Dynamics Computer Code (EFDC) (Hamrick, 1992) was used for this study to compute residence time and simulate FC transport in the Nassawadox Creek. EFDC is a general-purpose modeling package for simulating one-,

two-, and three-dimensional flow and transport in surface water systems including rivers, lakes, estuaries, reservoirs, wetlands, and coastal oceans.

For the Nassawadox Creek, a curvilinear orthogonal grid was used, with grid resolution ranging from 200 m at the open boundary to 20 m inside the creek. The model grids is carefully aligned with the shoreline (Fig. 3a). The bathymetry is based on the 3-arc second resolution Coastal Relief Model (https://www.ngdc.noaa.gov/mgg/coastal/crm.html, last access: January 1, 2021).

The model is driven by the river discharge enforced at the head of all the tributaries, tide at the open boundary, and reanalysis atmospheric forcing from NCEP (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.pressure.html, last access: January 1, 2021). The model has been well calibrated for hydrodynamics (Du et al., 2020).

2.3.2 Residence time calculation

For Nassawadox, we divided the entire domain into 12 segments, with one segment for each of the eight tributaries and four segments for the mainstem. Residence time for each segment was calculated based on the adjoint method proposed by Delhez (2006). Different from the traditional particle tracking method, the adjoint method (or backward method) takes into account the diffusion and can simulate the temporally and spatially varying residence time with one single model run, which is computationally efficient. The residence time, denoted by θ as a function of location x and time t, is calculated through

$$\frac{\partial\overline{\theta(t,x)}}{\partial t}+\delta_{\omega}(x)+v\cdot\nabla\overline{\theta(t,x)}+\nabla\cdot[\kappa\cdot\nabla\overline{\theta(t,x)}]=0 \qquad (7)$$

where v is the velocity vector, κ is the symmetric diffusion tensor, and

$$\delta_{\omega}(x)=\begin{cases}1 & if \quad x\in\omega\\ 0 & if \quad x\notin\omega\end{cases} \qquad (8)$$

where ω is the domain of interest (or the segments in this study). By specifying the ω, the EFDC model will allow computing the residence time for any defined segment. The adjoint method has been implemented into the EFDC model by Du and Shen (2016) and used to calculate the residence time in Chesapeake Bay in several previous studies (e.g., Du et al., 2017; Xiong et al., 2021). To calculate the residence time, we first run the hydrodynamic model and saved the half-hourly output of hydrodynamic fields including velocity, eddy diffusivity, surface elevation. Equation (7) was then computed backward from the end to the beginning of the simulation period based on the saved hydrodynamic output.

The distribution of 6-yr (2007-2012) mean residence time in each segment is shown in Fig. 3b. The mean residence time is high in the tributaries while low along the mainstem, with the smallest value near the mouth and the largest value at segments 7 and 8. Such distribution is related to the strength of the tidal current inside the bay, and the size of each segment. The magnitude of tidal current decreases greatly from the mouth toward upstream (not shown). The tidal current amplitude can be up to 0.35 m/s near the mouth while usually less than 0.05 m/s inside the tributaries. A stronger tidal current and tidal mixing near the mouth will lead to a faster flushing, resulting in a shorter residence

time. Consequently, the flux is large (on the order of 100 m^3/s) between mainstem boxes while much smaller (on the order of 1 m^3/s) between tributaries and the mainstem.

2.3.3 Experiments with hypothetical loading

To verify the inverse method, we treated model-calculated FC concentrations as observations. Eight numerical experiments using long-term mean constant river flow and constant FC loading (25, 50, 75, 100, 125, 150, 175, and 200% of long-term mean loading, respectively) were conducted. The net removal rate was set as 0.5 d^{-1} based on previous model calibration and observation data analysis (Du et al., 2020). For this experiment, the removal rate can be set as any number and it will not affect the loading calculation as long as both the 3D model and the inverse method use the same value of the removal rate.

Additional validation is conducted by comparing the realistic temporal varying daily "known loading" from the watershed model with the inversely estimated loading based on realistic model simulation results. The loading is inversely calculated for every 30 days, using the mean concentration and mean residence time averaged over the same period. Simulated vertical mean concentration at the middle of each segment was used as observed FC concentration (C) and the system was assumed to be under the dynamically quasi-steady-state with dC/dt=0. The inversely calculated loading (L_{inv}) was then compared with the known loading (L_o).

2.4 Method performance evaluation

Besides the common statistical measures including root mean square error (RMSE) and coefficient of determination (R^2), we also calculated *skill* following Willmott (1981):

$$Skill = 1 - \frac{\sum \left| X_{\mathrm{mod}} - X_{obs} \right|^2}{\sum \left(\left| X_{\mathrm{mod}} - \overline{X_{obs}} \right| + \left| X_{obs} - \overline{X_{obs}} \right| \right)^2} \tag{9}$$

where X_{obs} and X_{mod} are the observed and modeled variables, respectively, with the overbar indicating the time average. *Skill* provides an index of model-data agreement, with a skill of 1 indicating perfect agreement and 0 indicating complete disagreement. *Skill* has been widely used to evaluate the performance of numerical models (e.g., Warner et al., 2005) and data-driven models (e.g., Yu et al., 2020). While the R^2 indicates the model's capability of capturing the seasonal trend and interannual variations, and RMSE indicates the overall misfit between model and observation, *skill* can be regarded as a synthesis index to evaluate both the trend capturing and relative misfit.

3 RESULTS

3.1 Estimation of net removal rate

Based on the observational spatial distribution of bacterial concentration and calculated transit time, the net removal rate at three major tributaries (i.e., segment #5, 7, and 8) were separately calculated (using Equation (6)). These three tributaries were selected because there are multiple stations in these tributaries, which allow a statistically meaningful estimation of the parameter β. Note that not every cruise will result in a valid

estimation of *K*. When the logged bacterial concentrations do not follow a linearly decreasing trend from upstream to downstream, the estimation of *K* is deemed unreliable, as it indicates that there are likely additional unaccounted lateral sources (Du et al., 2020). There were a total of 50 estimations of *K* over the period of 2007-2012.

The estimated *K* varies temporally and follows roughly a normal distribution. The estimated *K* varies from 0.27 to 0.86 d^{-1}, with 25^{th} and 75^{th} percentiles of 0.43 and 0.63 d^{-1}, respectively. The estimated *K* falls into the normal range of FC die-off rate as measured in the laboratory (e.g., Bowie et al., 1985; Auer and Niehaus, 1993). A normal distribution, with a mean of 0.52 d^{-1} and a standard deviation of 0.14 d^{-1}, was used to fit the estimated *K* (Fig. 4). The normal distribution will be used to address the uncertainties in *K* when inversely estimate the loading. In the next section, we calculated the bacterial loading 100 times with 100 different *K* values based on this fitted normal distribution.

3.2 Method verification: comparison with a watershed model

The inverse method was first applied to estimate the loading based on observed bacterial concentration and calculated residence time. FC concentrations within each segment were averaged when applying the inverse method. Comparing the inversely estimated loading with the loading from a previously calibrated watershed model (Shen et al., 2005b) over the same period shows an overall good agreement between the two methods, in terms of their frequency distribution (Fig. 5), even though there are more occurrences of small loading from the watershed model compared to the inverse method. In Shen et al. (2005b), the watershed model for Nassawadox Creek was calibrated against the observations based on the best match between model results and observations to seek the least RMSE, which is a common practice for watershed model calibration. Using the

watershed model loading as a reference does not mean the watershed model is accurate. In the watershed model, there are also large uncertainties that are related to a variety of parameters or input data such as bacterial accumulation rate, land-use, and precipitation rate. The purpose of the comparison is to show that the inverse model can achieve a similar accuracy and frequency distribution as from a watershed modeling approach.

It is noticeable that, with different *K*, the estimated loading varies substantially (Fig. 5c). This is because the last term in Equation (1) dominates the bacterial budget in Nasawadox Creek where the tidal current is weak and exchange between neighboring segments is limited. The uncertainty associated with *K* and residence time will be further discussed in the discussion section.

3.3 Method verification: based on numerical experiments

A more accurate way to verify the inverse method is to conduct numerical simulation with "known" loadings (L_o) and a prescribed removal rate. The input loading can be constant or time-varying. We conducted tracer simulations with specified loading discharged from each subwatershed based on previous watershed model results. A constant *K* of 0.5 d^{-1} was used. The simulated tracer concentration and residence time were obtained to inversely calculate the input loading. For the constant loading scenarios, eight loadings were tested, including 25, 50, 75, 100, 125, 150, 175, and 200% of long-term mean loading. The inversely calculated loadings (L_{inv}) are nearly identical to the input loading for all major tributaries despite small bias (Fig. 6).

For the time-varying loading scenarios, the input loading varies over time, based on the outputs of watershed model. The 30-day mean tracer concentration at the middle

of each segment was used to mimic realistic sampling practice. The concentration was sampled at the same frequency as *in situ* measurements. It is shown that there isa good agreement between L_o and the ensemble mean of L_{inv} (Fig. 7). Because we used only one sample station for each segment and mean residence time over every 30 days while the watershed model loading varies every day, some differences between L_o and L_{inv} are expected. For instance, the estimated loading is biased from the input loading when the loading is high (Fig. 7). The discrepancies can be attributed to several factors. First, value at the middle of the segment was used as the segment-mean value. For realistic application, it is impractical to sample the entire segment and we sample in a way mimicking realistic practice. The sampled value may bias the true mean value of bacterial concentration in the given segment. Second, there may be a time-lag between the change of loading and the change of bacterial concentration. It may take days for water parcels to move from the head of a tributary to the mainstem (Du et al., 2020). As a result, a high value of bacterial concentration (which will lead to higher estimated loading) may not necessarily correspond to a high loading for the same given period. Nevertheless, the overall variations of L_o and L_{inv} match well with each other. The high value of *skill* and R^2, and low RMSE suggest the method is reliable.

4. DISCUSSION

Both realistic and idealized numerical experiments for Nassawadox Creek confirm the reliability of the proposed inverse method. Even though the initial motivation of introducing the method is to estimate bacterial loading, the method is readily applicable to other types of estuarine materials, such as dissolved inorganic nutrients. Like other methods to estimate watershed loadings (e.g., using watershed model), the

proposed method also has its limitations and its results are with uncertainties, which will be discussed in the following Sections.

4.1 Uncertainty associated with residence time and K

One of the challenges in assessing water quality conditions in estuaries arises from the model uncertainties. There are large uncertainties associated with unknown sources and time-varying decay rate of riverine materials (e.g., nutrients and bacteria) in the receiving waters. When using the inverse approach, the major uncertainties come from the decay rate and residence time. Residence time varies with time, due to changing transport processes regulated by time-varying tidal mixing, freshwater input, and atmospheric forcing. A smaller residence time indicates a larger flushing capability and will result in smaller bacterial concentrations given the same bacterial load. The temporal variability of the residence time over the monthly scale has already been considered in the above analysis, but residence time's daily variability was not included. Its impact will be shown in the following uncertainty analysis. The decay rate of bacteria varies and is affected by a variety of factors including temperature, irradiance, sediment settling, and resuspension. Among these, irradiance and temperature are generally considered the most important (Esham and Sizemore, 1998; Xu et al., 2002; Menon et al., 2003). It is commonly agreed that a higher temperature or irradiance tends to result in a larger die-off rate (e.g., Chigbu et al., 2005). Selection of K is important for the proposed inverse method.Using *in situ* observations to estimate K is a good approach to determine the range of K values and to evaluate the influence of uncertainties associated with the K. Nevertheless, the variability of decay rate for a specific region is usually much smaller than that of bacterial loading, which magnitude could vary in several orders.

To understand the relative influence of uncertainties associated with *K*, we calculated the loading inversely 100 times with different *K* values, with each *K* value randomly drawn from the fitted normal distribution. The frequency distribution for the estimated loadings with different *K* values are shown in Fig.6.

We also compared the influences of uncertainties associated with residence time and *K* on the estimated loadings. For each month, we calculated the loading using time-varying residence time and constant *K* to examine the influence of uncertainties associated with residence time; using constant residence time and a set of *K* values (randomly drawn from a normal distribution with a mean of 0.52 d^{-1} and standard deviation of 0.14 d^{-1}) to examine the influence of uncertainties associated with *K* value; and using varying residence time and varying *K* to examine the combined influences of uncertainties in *K* and residence time. For each month, there will be 120 residence time values (four values per day) and 100 *K* values to be tested. Results show that the influence of residence time-induced uncertainties is much smaller than of *K* induced uncertainties (Fig. 8). The combined uncertainties associated with *K* and residence time has nearly identical impact compared to *K* uncertainties alone (not shown). A larger range of *K* will certainly result in larger uncertainty. Since a standard deviation of 0.14 d^{-1} is a conservative value for *K* in natural waters (Bellair et al., 1977), it is fair to claim that uncertainties in the inversely calculated results can be mostly attributed to the *K* value instead of residence time for the Nassawadox Creek.

4.2 Broad applications of the inverse method

Using bacterial loading as an example, we show here that the proposed inverse method is an efficient alternative way to estimate watershed loading. This method takes advantage of the observations and the hydrodynamic model. The hydrodynamic models are nowadays relatively accurate to reproduce well the water movement and water exchange. The observations are assumed errorless although it may be not the case for bacteria concentrations as the bacterial loadings may vary dramatically over a short period. However, the error of measurements can be included in the uncertainty analysis.

The proposed method can be useful in multiple ways. First, it can be used to crosscheck the watershed model and thus calibrate the watershed model. Different from the traditional calibration method by back-forth running the watershed model and numerical model for the receiving water, which is tedious and time-consuming (Shen et al., 2006), one can use the inverse method to quickly adjust the spatially and seasonally varying parameters in watershed models. Using watershed models to estimate loadings suffers greatly from the uncertainties in animal distribution and seasonality of animal density (Jamieson et al., 2004; Jeong et al., 2019), and there is still no easy way to address this problem. While the inversely estimated loading also suffers from its limited temporal resolution, combining both two results is likely to give more reliable results.

Second, the inverse method can be used as an alternative way to estimate the loading if a watershed model is not available due to reasons such as lacking essential data of bacterial sources, spatially varying precipitation, domestic and wild animal density, and accurate land-use. Lacking watershed model input data (e.g., precipitation, animal density) is a common problem for bacterial pollution assessment. In such cases, the

inverse method could be a cost-effective and efficient option. With observational data at limited monitoring stations and an estimation of the residence time of a given waterbody, one can easily determine the bacterial loading and identify the bacterial sources.

As many remediations have been implemented in the watershed, the effectiveness of these efforts needs to be evaluated. This inverse method can be used to examine whether the watershed loading changes in the desired way following remediation efforts.

Lastly, the method can be used for other bacterial indicators (e.g., E. coli, and Enterococci) and riverine materials as well, such as total nitrogen (TN), phosphate, and heavy metal, etc. The mass transport for these materials follows the same rules as for the bacteria. The only difference is the removal rate. The removal rate of these materials differs from each other due to different sedimentation and biological uptaking. For instance, TN in the Chesapeake Bay has a removal rate of 0.001 d^{-1} on average (Dettman, 2001) and its net removal rate is related to the complicated sedimentation and plankton uptaking.

To use this approach, two independent datasets are needed: (1) the measured concentration of bacteria or other materials of interest, (2) the time-varying residence time for each segment. Constant residence time is acceptable when the residence time does not change dramatically. As shown in Section 4.1, for the Nassawadox Creek, the temporal variability of residence time has little impact on the final estimations of the loading because the water exchange in the system is more controlled by tidal exchange instead of freshwater inputs. For coastal bays where hydrodynamics are dominated by freshwater input and where freshwater input varies seasonally (e.g, Mobile Bay; Du et al.,

2018), the temporal variability of residence time cannot be ignored. For coastal bays with large volume and relatively stable sub-tidal exchange such as Chesapeake Bay where the two-layer subtidal exchange is prominent (Xiong et al., 2021), the residence time variability of shorter-timescale (e.g., daily) may be less important.

4.3 Limitations of the inverse method

Some restrictions shall be noted when applying the inverse method. The inverse method is based on the assumption that water or material in each segment is well-mixed. In reality, estuaries are usually not well-mixed due to the stratification induced by the riverine buoyancy. However, it doesn't mean the method is not suitable for the partially mixed or stratified estuaries as long as the fluxes and concentrations are representative of the entire segment. In these cases, multiple monitoring stations are needed to obtain a segment-mean value of tracer concentration.

It is also necessary to point out that residence time could vary with time. Residence time is a function of river flow, tide, and wind force, and therefore usually varies seasonally and interannually. Besides the adjoint method as used for the Nassawadox, residence time can be calculated by other methods, such as the tidal prism method (Kuo et al., 2005; Andutta, et al., 2014), freshwater fractional method (Knudsen, 1900), analytical method based on double diffusion equations (Choi and Lee, 2004), and isohaline based method (MacCready, 2011). We applied the adjoint method with a hydrodynamic model because of its efficiency as one single model run can provide spatial and temporal varying residence time with high accuracy.

The resolution of the results from the inverse method depends directly on the measurement accuracy and frequency. To obtain results with a higher temporal resolution, it is necessary to have continuous and higher frequency data.

5. CONCLUSIONS

We introduced a simple inverse method using observation data and residence time to estimate bacteria loading. We conducted a series of experiments to demonstrate the mothod is reliable. The method is computationally efficient and has the capability to estimate the long-term time-varying loadings for multiple connected estuarine segments. Using residence time instead of the tidal prism makes the method more suitable for estuarine water quality studies, as the tidal prism method is limited by its restricted segmentation scheme and high uncertainties associated with return ratio.

Regarding the accelerating change of land-use and great uncertainties in bacteria sources, bacterial concentration could vary significantly in an estuary and it may be not a good strategy to rely solely on land-use information and mean condition for time-sensitive environmental assessment. While the method has its limitations, it is shown to have high potential if given reliable residence time and continuous monitoring data with sufficient resolution.

ACKNOWLEDGMENTS

We would like to thank Anna Schlegel and Kristie Britt for providing resources and suggestions for the study. We thank Virginia Shellfish Sanitation for providing the monitoring data. This work is supported by Virginia Department of Environmental

Quality (award #16006). This is contribution No. 3986 of Virginia Institute of Marine Sciences, College of William and Mary.

APPENDIX A1:

Assuming (1) a well-mixed box with an initial tracer concentration C_0 and without additional internal source or sink, and (2) that an outflux Q_{out} is constant throughout the time, the tracer mass transport equation follows

$$\frac{dVC(t)}{dt} = -Q_{out}C(t) \qquad \text{(A1)}$$

The solution of A1 is

$$C = C_0 e^{-\frac{Q_{out}}{V}t} \qquad \text{(A2)}$$

The mean residence time (τ) can be expressed with the remnant function as (Takeoka, 1984),

$$\tau = \int_0^\infty r(t)dt \qquad \text{(A3)}$$

where r(t) is the ratio of material remaining inside the waterbody at time t to its initial total mass,

$$r(t) = \frac{C(t)}{C_0} \qquad \text{(A4)}$$

By combining A2-A4, the mean residence time can be solved as

$$\tau = \frac{V}{Q_{out}} \qquad \text{(A5)}$$

Therefore, Q_{out} can be regarded as the effective outflux corresponding to the mean residence time of the given box.

REFERENCES

Andutta, F.P., Ridd, P.V., Deleersnijder, E., Pradle, D. (2014). Contaminant exchange rates in estuaries – New formulae accounting for advection and dispersion. *Progress in Oceanography, 120*, 139-153.

Arnold, J. G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Harmel, R.D., van Griensven, A., Van Liew, M.W., Kannan, N., Jha, M. K. (2012). SWAT: model use, calibration, and validation. *Transaction of the ASABE* 55, 1491-1508.

Auer, M.T., Niehaus, S.L. (1993). Modeling fecal coliform bacteria-I. Field and laboratory determination of loss kinetics. *Water Research, 27*, 693–701.

Bellair, J.T., Parr-Smith, G.A., Wallis, I.G. (1977). Significance of diurnal variations in fecal coliform die-off rates in the design of ocean outfalls. *Journal of the Water Pollution Control Federation*, *49*, 2022-2030.

Bowie, G.L., Mills, W.B., Porcella, D.B., Campbell, C.L., Pagenkopf, J.R., Rupp G.L., Johnson, K.M., Chan, P.W.H., Gherini, S.A., Chamberlain, C.E. (1985). Rates, constants, and kinetics formulations in surface water quality modeling (2nd edition). EPA/600/3-85/040, Environmental Research Laboratory, U.S. Environmental Protection Agency, Athens, GA.

Bouchaou, L., Michelot, J.L., Vengosh, A., Hsissou, Y., Qurtobi, M., Gaye, C.B., Bullen, T.D., Zuppi, G.M. (2008). Application of multiple isotopic and geochemical tracers for investigation of recharge, salinization, and residence time of water in the Souss – Massa aquifer, southwest of Morocco. *Journal of Hydrology, 352*, 267–287.

Brauwere, A.De, Ouattara, N.K., Servais, P. (2014). Modeling fecal indicator bacteria concentrations in natural surface waters : A review. *Critical Reviews in Environmental Science and Technology*, *44*, 2380–2453.

Chen, D., Dahlgren, R.A., Shen, Y., Lu, J. (2012). A Bayesian approach for calculating variable total maximum daily loads and uncertainty assessment. *Science of the Total Environment*, *430*, 59–67.

Chigbu, P., Gordon, S., Strange, T.R. (2005). Fecal coliform bacteria disappearance rates in a north-central Gulf of Mexico estuary. *Estuarine, Coastal and Shelf Science*, *65*, 309–318.

Cho, K.H., Pachepsky, Y.A., Kim, M., Pyo, J., Park, M., Kim, Y.M., Kim, J.W., Kim, J.H. (2016). Modeling seasonal variability of fecal coliform in natural surface waters using the modified SWAT. *Journal of Hydrology*, *535*, 377–385.

Choi, K.W., Lee, J.H.W. (2004). Numerical determination of flushing time for stratified water bodies. *Journal of Marine Systems*, *50*, 263-281.

Delhez, E.J.M. (2006). Transient residence and exposure times. *Ocean Science*, *2*(1), 1–9.

Dettmann, E.H. (2001). Effect of water residence time on annual export and denitrification of nitrogen in estuaries: A model analysis. *Estuaries*, 24, 481–490.

Du, J., Park, K., Shen, J., Dzwonkowski, B., Yu, X., Yoon, B. Il (2018). Role of baroclinic processes on flushing characteristics in a highly stratified estuarine system, Mobile Bay, Alabama. *Journal of Geophysical Research: Oceans*, *123*, 1–20.

Du, J., Shen, J. (2016). Water residence time in Chesapeake Bay for 1980 – 2012. *Journal of Marine Systems*, *164*, 101–111.

Du, J., Shen, J., Bilkovic, D.M., Hershner, C.H., Sisson, M. (2017). A numerical modeling approach to predict the effect of a storm surge barrier on hydrodynamics and long-term transport processes in a partially mixed estuary. *Estuaries and Coasts*, *40*, 387-403

Du, J., Shen, J., Park, K., Yu, X., Ye, F., Qin, Q., Xiong, J., Chen, Y. (2020). Using observed bacteria concentration and modeled transit time under an analytical framework to estimate overall removal rate of fecal coliform in an estuary. arXiv preprint, arXiv:2001.07603. https://arxiv.org/abs/2001.07603

Esham, E.C., Sizemore, R.K. (1998). Evaluation of two techniques: mFC and mTEC for determining distributions of fecal pollution in small, North Carolina tidal creeks. *Water, Air, & Soil Pollution*, *106*, 179–197.

Ferguson, C.M., Charles, K., Deere, D.A., Ferguson, C.M., Charles, K., & Deere, D.A. (2008). Quantification of microbial sources in drinking-water catchments. *Critical Reviews in Environmental Science and Technology*, *39*, 1–40.

Grant, M.A. (1997). A new membrane filtration medium for simultaneous detection and enumeration of Escherichia coli and total coliforms. *Applied and Environmental Microbiology*, *63*, 3526-3530.

Gronewold, A.D., Qian, S.S., Wolpert, R.L., Reckhow, K.H. (2009). Calibrating and validating bacterial water quality models : A Bayesian approach. *Water Research*, *43*(10), 2688–2698.

Hamrick, J.M. (1992). A three-dimensional environmental fluid dynamics computer code: Theoretical and computational aspects. The College of William and Mary,

Virginia Institute of Marine Science, (Special Paper 317), 63. https://doi.org/10.21220/v5tt6c

He, L., Lu, J., Shi, W. (2007). Variability of fecal indicator bacteria in flowing and ponded waters in southern California : Implications for bacterial TMDL development and implementation. *Water Research*, *41*, 3132–3140.

Huang, W., Spaulding, M. (2002). Modelling residence-time response to freshwater input in Apalachicola Bay, Florida, USA. *Hydrological Processes*, *16*(15), 3051–3064.

Jamieson, R., Gordon, R., Joy, D., Lee, H. (2004). Assessing microbial pollution of rural surface waters: A review of current watershed scale modeling approaches. *Agriculture Water Management*, *70*, 1–17.

Jang, J., Hur, H., Sadowsky, M.J., Byappanahalli, M.N., Yan, T., Ishii, S. (2017). Environmental Escherichia coli : Ecology and public health implications - a review. *Journal of Applied Microbiology*, *123*, 570–581.

Jeong, J., Wagner, K., Flores, J.J., Cawthon, T., Her, Y., Osorio, J., Yen, H. (2019). Linking watershed modeling and bacterial source tracking to better assess E . coli sources. *Science of the Total Environment*, *648*, 164–175.

Knudsen, M. (1900). Ein Hydrographische Lehrsatz. *Annalen der Hydrographie und Marinen Meteorologie*, 28, 316-320.

Kuo, A. Y., Park, P., Kim, S.G., Lin, J. (2005). A tidal prism water quality model for small coastal basins. *Coastal Management*, 33, 101–117.

MacCready, P. (2011). Calculating estuarine exchange flow using isohaline coordinates. *Journal of Physical Oceanography*, *41*(6), 1116–1124.

Menon, P., Billen, G., Servais, P. (2003). Mortality rates of autochthonous and fecal bacteria in natural aquatic ecosystems. *Water Research*, *37*, 4151–4158.

Monsen, N.E., Cloern, J.E., Lucas, L.V., &Monismith, S.G. (2002). The use of flushing time, residence time, and age as transport time scales. *Limnology and Oceanography*, *47*(5), 1545–1553.

Passerat, J., Ouattara, N.K., Mouchel, J.M., Rocher, V., Servais, P. (2011). Impact of an intense combined sewer overflow event on the microbiological water quality of the Seine River. *Water Research*, *45*(2), 893–903.

Plew, D.R. (2018). Using simple dilution models to predict New Zealand estuarine water quality. *Estuaries and Coasts*, *41*, 1643–1659.

Servais, P., Garcia-armisen, T., George, I., Billen, G. (2007). Fecal bacteria in the rivers of the Seine drainage network (France): Sources , fate and modelling. *Science of the Total Environment*, *375*, 152–167.

Sheldon, J.E., Alber, M. (2006). The calculation of estuarine turnover times using freshwater fraction and tidal prism models: A critical evaluation. *Estuaries and Coasts*, *29*(1), 133–146.

Shen, J., Parker, P., Riverson, J. (2005a). A new approach for a windows-based watershed modeling system based on a database-supporting architecture. *Environmental Modelling Software*, 20, 1127-1138.

Shen, J., Sun, S., & Wang, T. (2005b). Development of the fecal coliform total maximum daily load using Loading Simulation Program C++ and tidal prism model in estuarine shellfish growing areas: A case study in the Nassawadox coastal

embayment, Virginia. *Journal of Environmental Science and Health, Part A, 40*(9), 1791–1807.

Shen, J., Jia, J., Sisson, G.M. (2006). Inverse estimation of nonpoint sources of fecal coliform for establishing allowable load for Wye River, Maryland. *Water Research, 40,* 3333–3342.

Shen, J., Zhao, Y., (2009). A Bayesian approach for estimating bacterial nonpoint source loading in an estuary with limited observations. *Journal of Environmental Science and Health, Part A, 44,* 1574–1584.

Shen, J., Zhao, Y. (2010). Combined Bayesian statistics and load duration curve method for bacteria nonpoint source loading estimation. *Water Research, 44*(1), 77–84.

Sobel, R.S., Rifai, H.S., Petersen, C.M. (2017). Integration of tidal prism model and HSPF for simulating indicator bacteria in coastal watersheds. *Estuarine, Coastal and Shelf Science, 196,* 248–257.

Takeoka, H. (1984). Fundamental concepts of exchange and transport time scales in a coastal sea. *Continental Shelf Research, 3*(3), 311–326.

USEPA. (2019). National summary of impaired waters and TMDL information. https://iaspub.epa.gov/waters10/attains_nation_cy.control?p_report_type=T (last accessed on January 1, 2021)

Warner, J.C., Geyer, W.R., Lerczak, J.A. (2005). Numerical modeling of an estuary: A comprehensive skill assessment. *Journal of Geophysical Research: Oceans, 110,* 1–13.

Willmott, C. (1981). On the validation of models. *Physical Geography, 2*(2), 184–194.

Xiong, J., Shen, J., Qin, Q., Du, J. (2021). Water exchange and its relationships with external forcings and residence time in Chesapeake Bay. *Journal of Marine Systems*, 215, 103497

Xu, P., Brissaud, F., Fazio, A. (2002). Non-steady-state modelling of fecal coliform removal in deep tertiary lagoons. *Water Research*, *36*, 3074–3082.

Yu, X.. Shen, J., Du, J. (2020). A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example. *Water Resource Research*, *56*, e2020WR027227

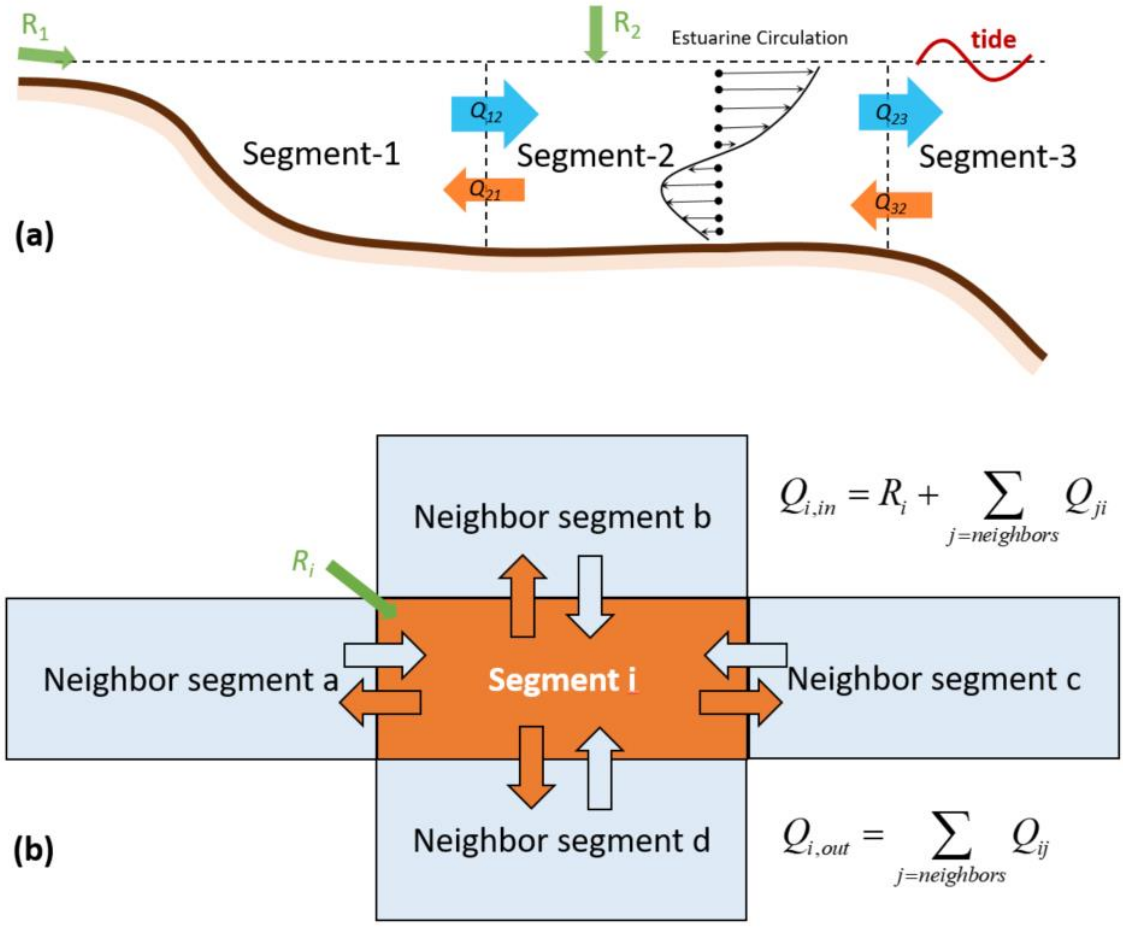

Figure 1: Sketch diagrams showing the water exchange between segments. (a) Sketch diagram showing the vertical velocity structure and water exchange between downstream and upstream. (b) Sketch diagram showing the water inflow and outflow for a given segment connected by a limited number of neighboring segments.

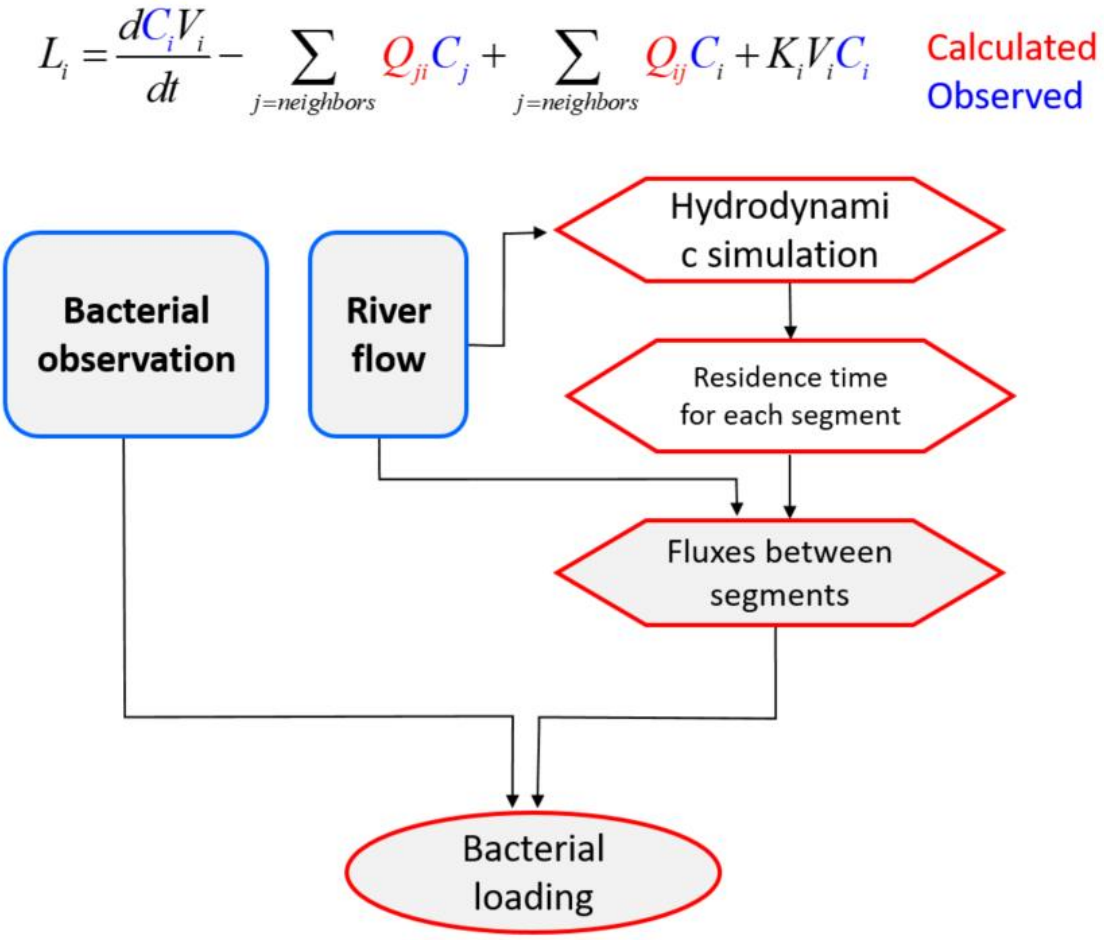

Figure 2: A sketch diagram showing the workflow of the inverse method application for an estuary.

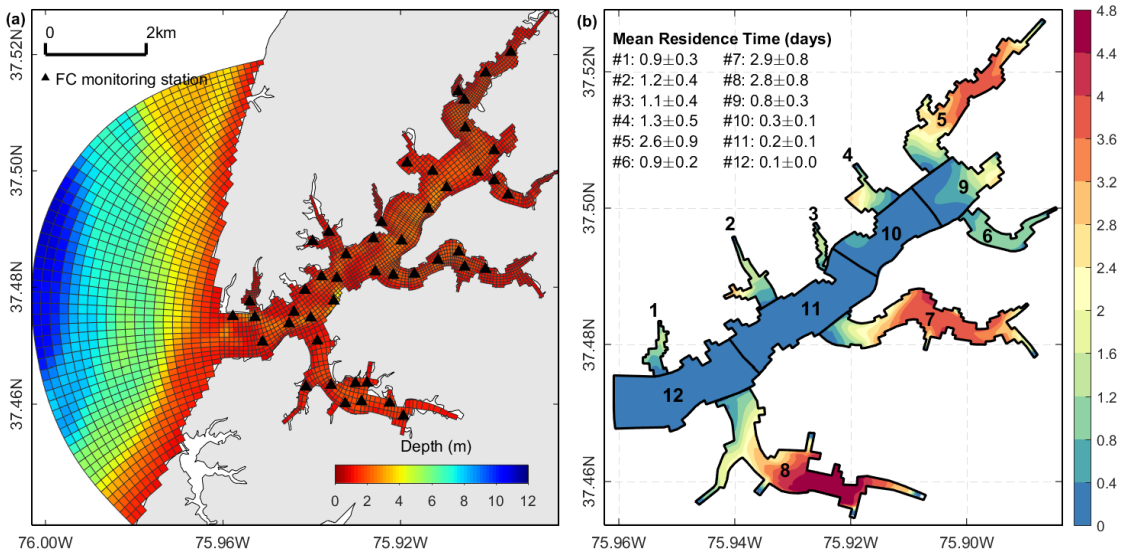

Figure 3: (a) Bacterial monitoring stations (black triangles) and numerical model grids, with filled color denoting the water depth. (b) The 6-yr mean residence time for the 12 segments, with the mean value and standard deviation shown with text in the top left.

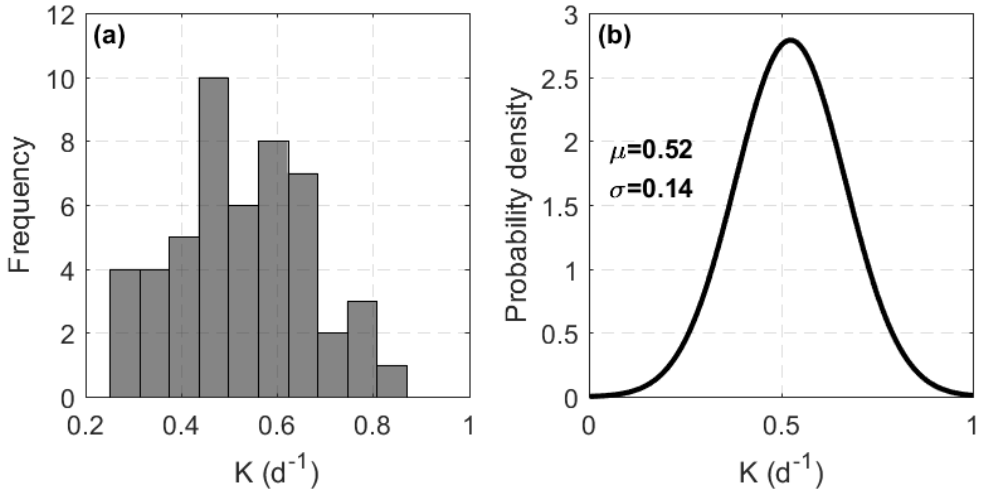

Figure 4: (a) Frequency distribution of estimated net removal rate of Fecal Coliform, K. (b) A normal distribution to fit the distribution of estimated K. The mean (μ) and standard deviation (σ) for the normal distribution are shown in text in the plot.

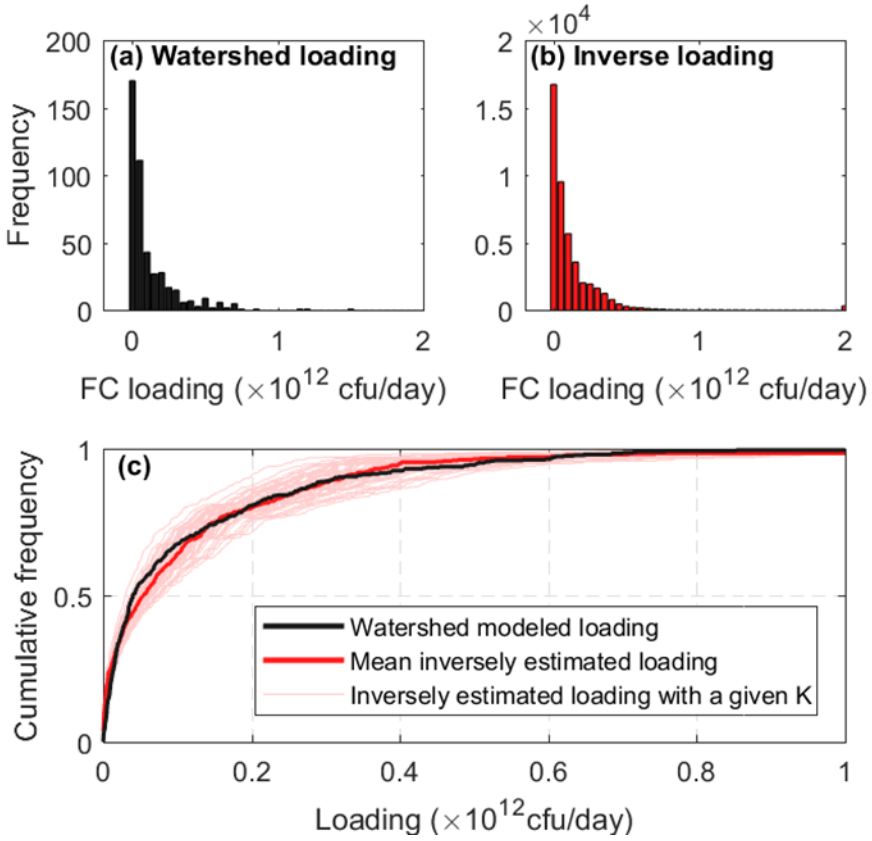

Figure 5: Comparison of the loading between inverse method and watershed model. (a) The histogram of watershed loading for all the 8 tributaries (total sample number 57×8=456). (b) The histogram for inversely estimated loading for all the 8 tributaries (total sample number 45600).The inverse method has been applied 100 times with randomly selected *K* values drawn from a normal distribution (μ=0.52, σ=0.14). (c) Cumulative distribution function comparison between watershed loading and inversely estimate loading based on different *K* value, with bold red line indicating the ensemble mean of the 100 times of calculation.

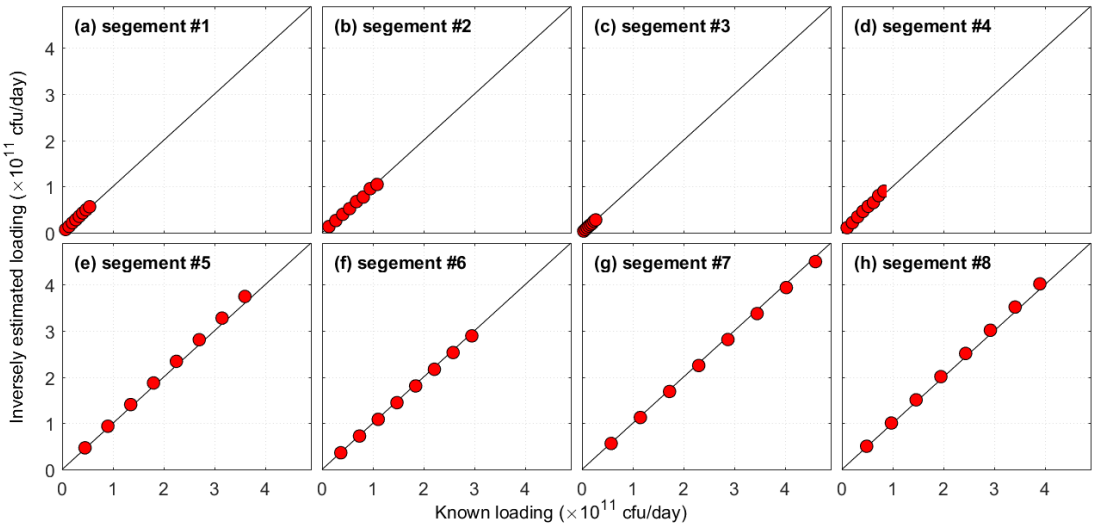

Figure 6: Model validation based on idealized numerical experiments with constant FC loadings for each of the eight tributaries (segment 1-8). For each numerical experiment, there is one estimated loading based on the mean bacterial concentration averaged over the entire simulation period.

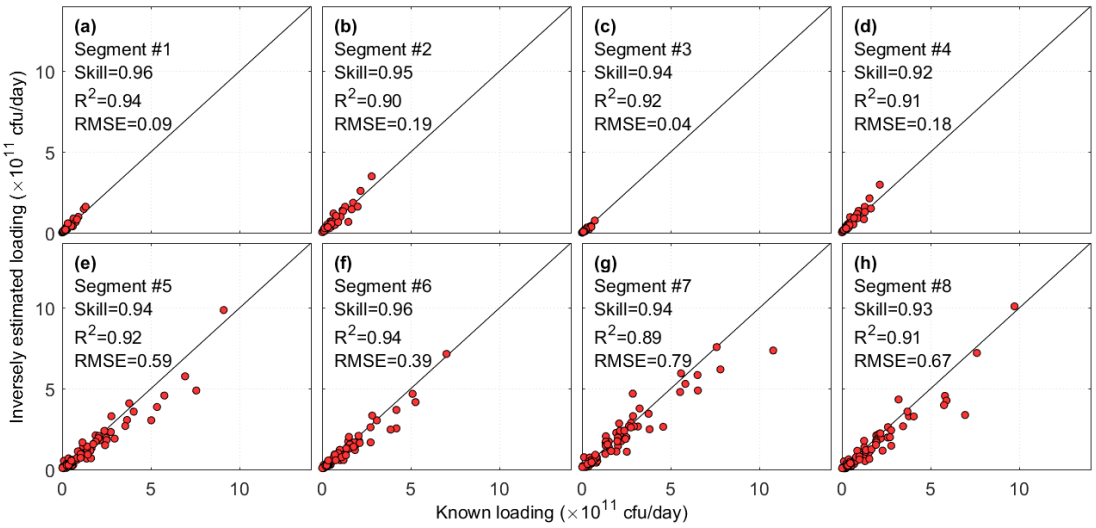

Figure 7: Model validation by comparing model input loading (a.k.a., known loading) and the inversely calculated loading based on model-calculated FC concentration for the eight tributaries (i.e., segment 1-8). Each data point represents a monthly mean value (total number of data points N=72). Statistical numbers are shown in text, including root mean square error (RMSE), R2, and *skill*.

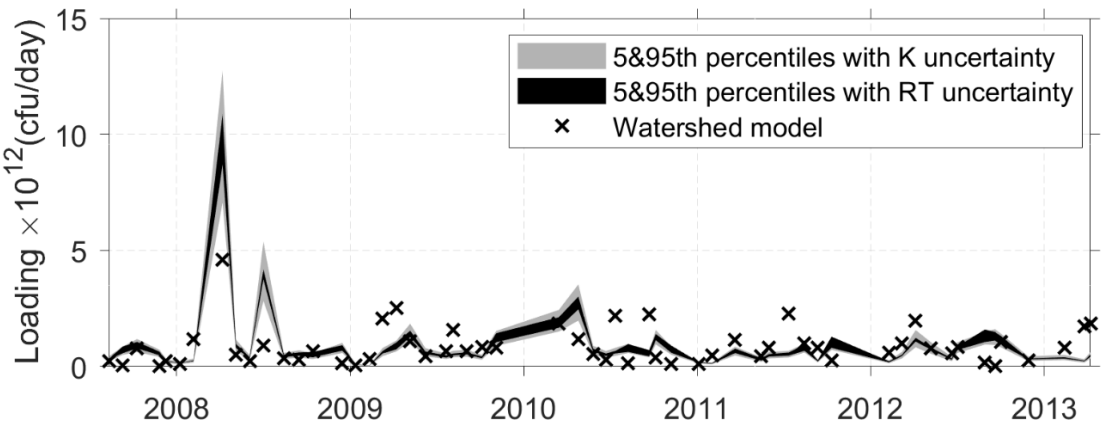

Figure 8: Total loadings discharging into the entire estuarine system and their uncertainties (represented with the 95% confidence) induced by removal rate (K) and residence time (RT). The time series here has the same frequency of field measurements (nearly monthly with some occasional gaps).

VITA

Xin Yu

Born in Shijiazhuang, Hebei Province, China. Graduated from Zanhuang High School, Zanhuang, Hebei Province in 2005. Earned a B.S. in Geography from Hebei Normal University, Shijiazhuang, Hebei Province, China in 2009. Earned a M.S. in Marine Geology from Nanjing University, Nanjing, Jiangsu, China in 2012. Entered the Ph.D. program at the Virginia Institute of Marine Science, College of William and Mary in August 2016 and was under the supervision of Dr. Jian Shen.