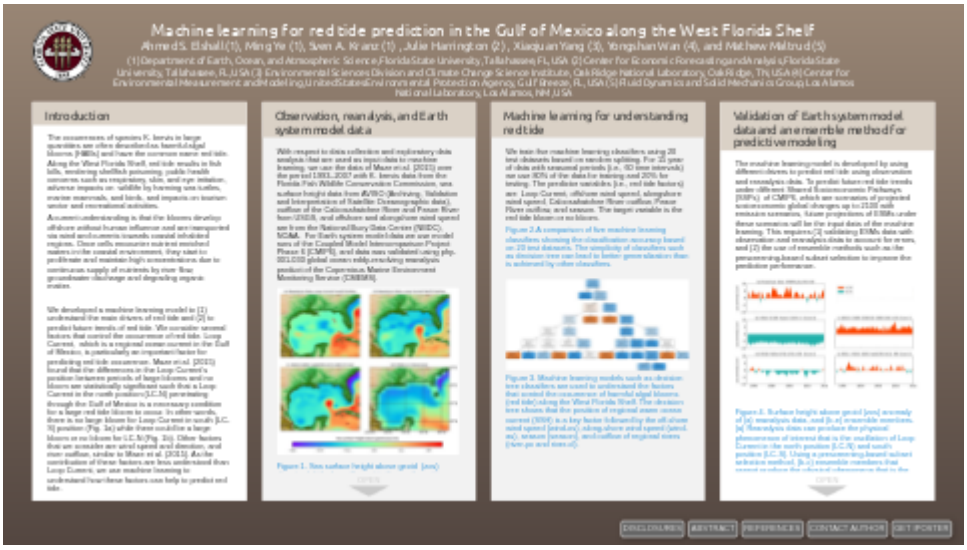
Machine learning for red tide prediction in the Gulf of Mexico along the West Florida Shelf



Ahmed S. Elshall (1), Ming Ye (1), Sven A. Kranz (1) , Julie Harrington (2) , Xiaojuan Yang (3), Yongshan Wan (4), and Mathew Maltrud (5)

(1) Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA (2) Center for Economic Forecasting and Analysis, Florida State University, Tallahassee, FL, USA (3) Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA (4) Center for Environmental Measurement and Modeling, United States Environmental Protection Agency, Gulf Breeze, FL, USA (5) Fluid Dynamics and Solid Mechanics Group, Los Alamos National Laboratory, Los Alamos, NM, USA

PRESENTED AT:

INTRODUCTION

The occurrences of species K. brevis in large quantities are often described as harmful algal blooms (HABs) and have the common name red tide. Along the West Florida Shelf, red tide results in fish kills, rendering shellfish poisoning, public health concerns such as respiratory, skin, and eye irritation, adverse impacts on wildlife by harming sea turtles, marine mammals, and birds, and impacts on tourism sector and recreational activities.

A current understanding is that the blooms develop offshore without human influence and are transported via wind and currents towards coastal inhabited regions. Once cells encounter nutrient enriched waters in the coastal environment, they start to proliferate and maintain high concentrations due to continuous supply of nutrients by river flow, groundwater discharge and degrading organic matter.

We developed a machine learning model to (1) understand the main drivers of red tide and (2) to predict future trends of red tide. We consider several factors that control the occurrence of red tide. Loop Current, which is a regional ocean current in the Gulf of Mexico, is particularly an important factor for predicting red tide occurrence. Maze et al. (2015) found that the differences in the Loop Current's position between periods of large blooms and no bloom are statistically significant such that a Loop Current in the north position (LC-N) penetrating through the Gulf of Mexico is a necessary condition for a large red tide bloom to occur. In other words, there is no large bloom for Loop Current in south (LC-S) position (Fig. 1a) while there could be a large bloom or no bloom for LC-N (Fig. 1b). Other factors that we consider are wind speed and direction, and river outflow, similar to Maze et al. (2015). As the contribution of these factors are less understood than Loop Current, we use machine learning to understand how these factors can help to predict red tide.

OBSERVATION, REANALYSIS, AND EARTH SYSTEM MODEL DATA

With respect to data collection and exploratory data analysis that are used as input data to machine learning, we use the data of Maze et al. (2015) over the period 1993–2007 with K. brevis data from the Florida Fish Wildlife Conservation Commission, sea surface height data from AVISO (Archiving, Validation and Interpretation of Satellite Oceanographic data), outflow of the Caloosahatchee River and Peace River from USGS, and offshore and alongshore wind speed are from the National Buoy Data Center (NBDC), NOAA. For Earth system model data we use model runs of the Coupled Model Intercomparison Project Phase 6 (CMIP6), and data was validated using phy-001-030 global ocean eddy-resolving reanalysis product of the Copernicus Marine Environment Monitoring Service (CMEMS).

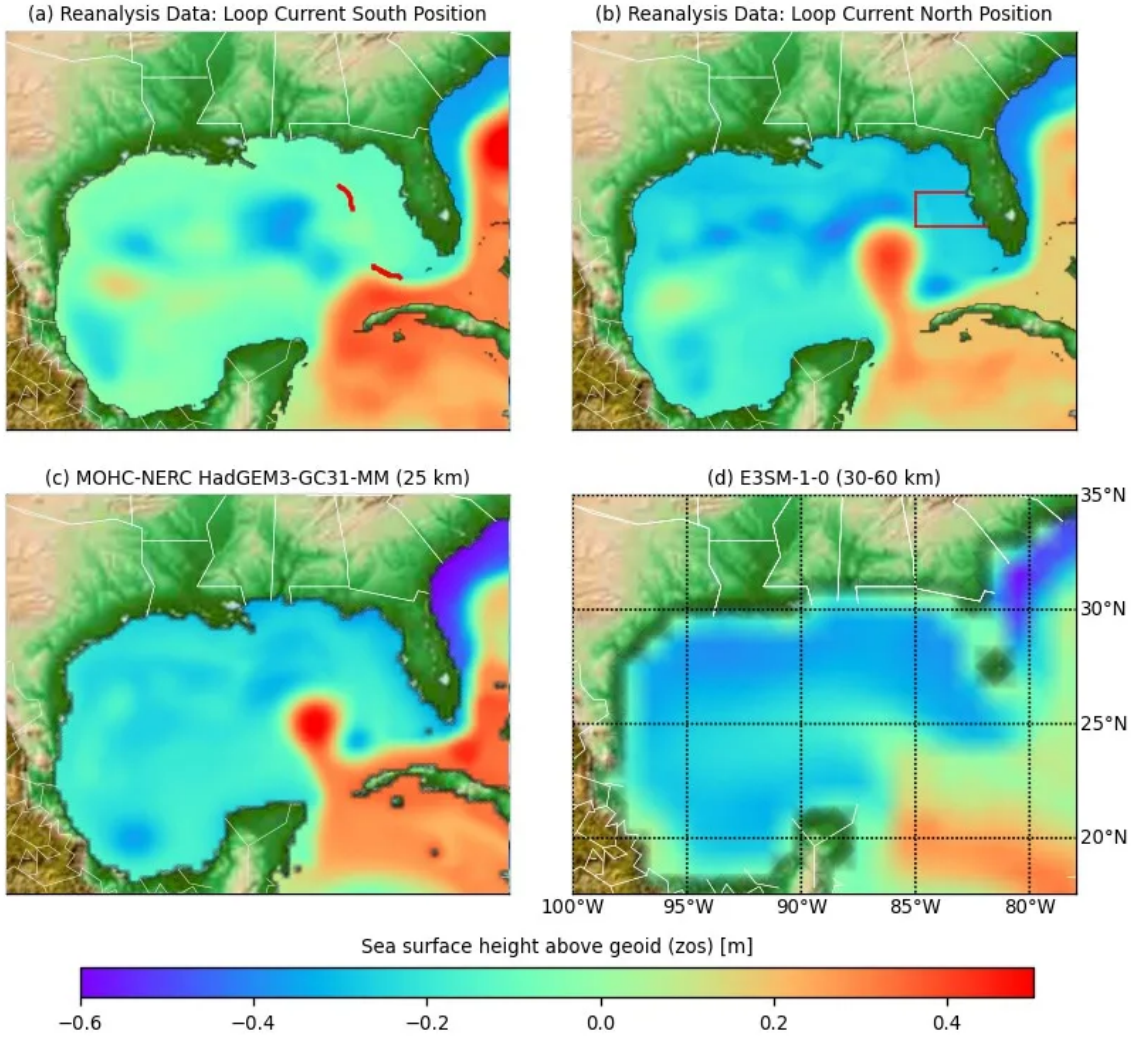


Figure 1. Sea surface height above geoid (zos) showing (a) the Loop Current in the south position (LS-C) in 2010-03 for reanalysis data, and the Loop Current in the north position (LC-N) in 2010-06 for (b) reanalysis data, (c) a high-resolution Earth system model (ESM), and (d) a standard-resolution ESM. Two red segments along the 300m isobath in (a) are used to determine Loop Current position (i.e., LC-N and LC-S). The red box of (b) shows the study area where red tide blooms are considered by this study and Maze et al. (2015). Figs. (c-d) confirms our prior knowledge that standard-resolution models cannot simulate the Loop current, unlike high-resolution ESMs.

MACHINE LEARNING FOR UNDERSTANDING RED TIDE

For 15 year of data with seasonal periods (i.e., 60 time intervals) we use 80% of the data for training and 20% for testing. We trained 12 machine learning classifiers using 200 test datasets based on random splitting with the condition that the ratio of large-bloom to no-bloom events of any test dataset is greater than or equal 0.25, since the original dataset has a large-bloom to no-bloom ratio of 0.27. The predictor variables (i.e., red tide factors) are Loop Current, offshore wind speed, alongshore wind speed, Caloosahatchee River outflow, Peace River outflow, and season. The target variable is the red tide large-bloom or no-bloom event.

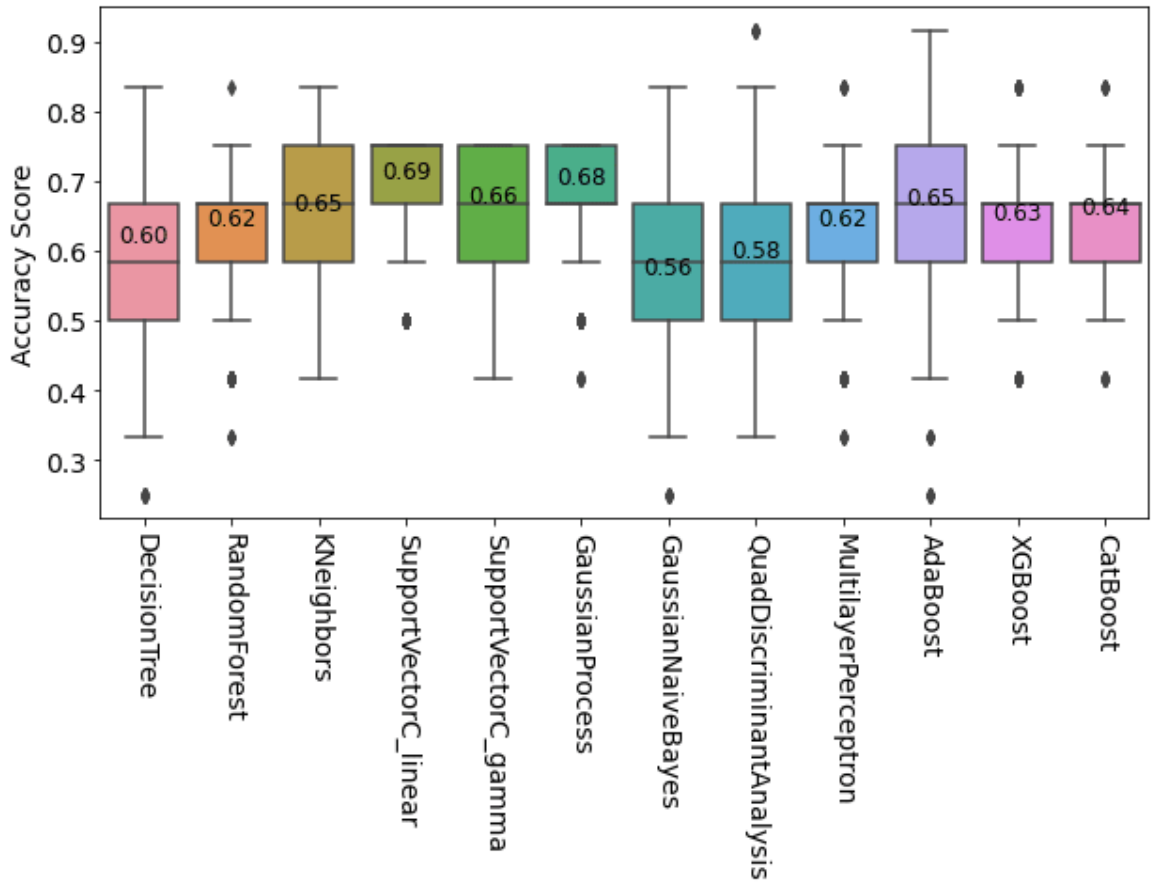


Figure 2.A comparison of 12 machine learning classifiers showing the classification accuracy based on 200 test datasets each with 5 repeated runs. The mean accuracy of each classifer is annotated on the figure. The simplicity of classifiers such as decision tree and random forst can lead to better generalization than other classifiers. The ensemble methods (e.g., adaBoost, XGBoost, and CatBoost) do not necessarily lead to better performance, and SVC methods (e.g., SupportVectorC_linear, and SupportVectorC_gamma) gain advantage through finding linear separation.

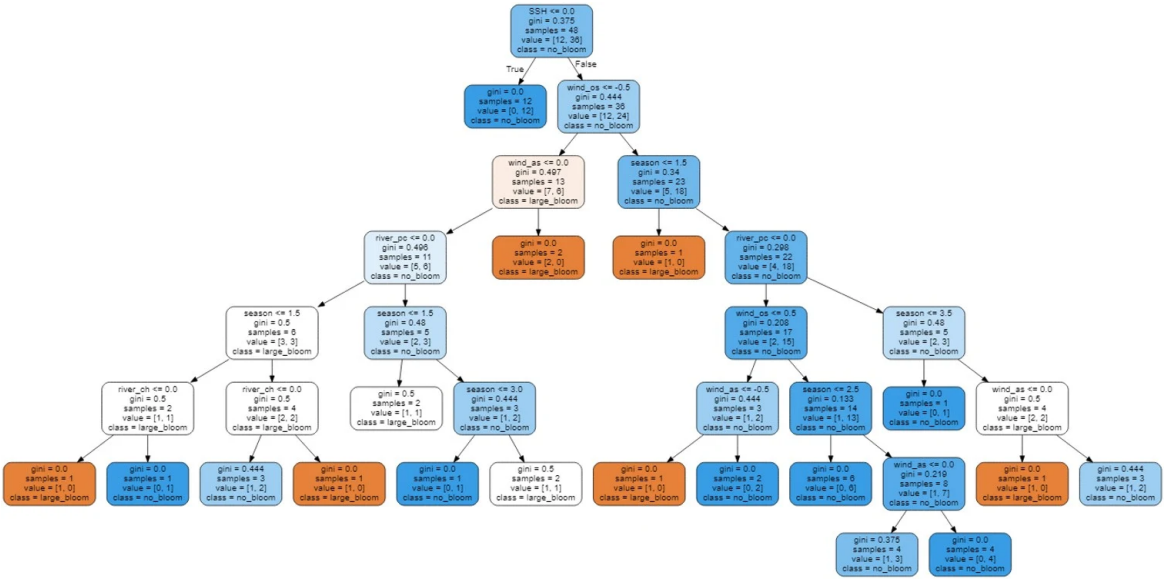


Figure 3. Machine learning models such as decision tree classifiers are used to understand the factors that control the occurrence of harmful algal blooms (red tide) along the West Florida Shelf. The decision tree shows that the position of regional warm ocean current (SSH) is a key factor followed by the off-shore wind speed (wind-os), along-shore wind speed (wind-as), season (season), and outflow of regional rivers (river-pc and river-cl).

VALIDATION OF EARTH SYSTEM MODEL DATA AND AN ENSEMBLE METHOD FOR PREDICTIVE MODELING

The machine learning model is developed by using different drivers to predict red tide using observation and reanalysis data. To predict future red tide trends under different Shared Socioeconomic Pathways (SSPs) of CMIP6, which are scenarios of projected socioeconomic global changes up to 2100 with emission scenarios, future projections of ESMs under these scenarios will be the input data of the machine learning. This requires (1) validating ESMs data with observation and reanalysis data to account for errors, and (2) the use of ensemble methods such as the prescreening-based subset selection to improve the predictive performance.

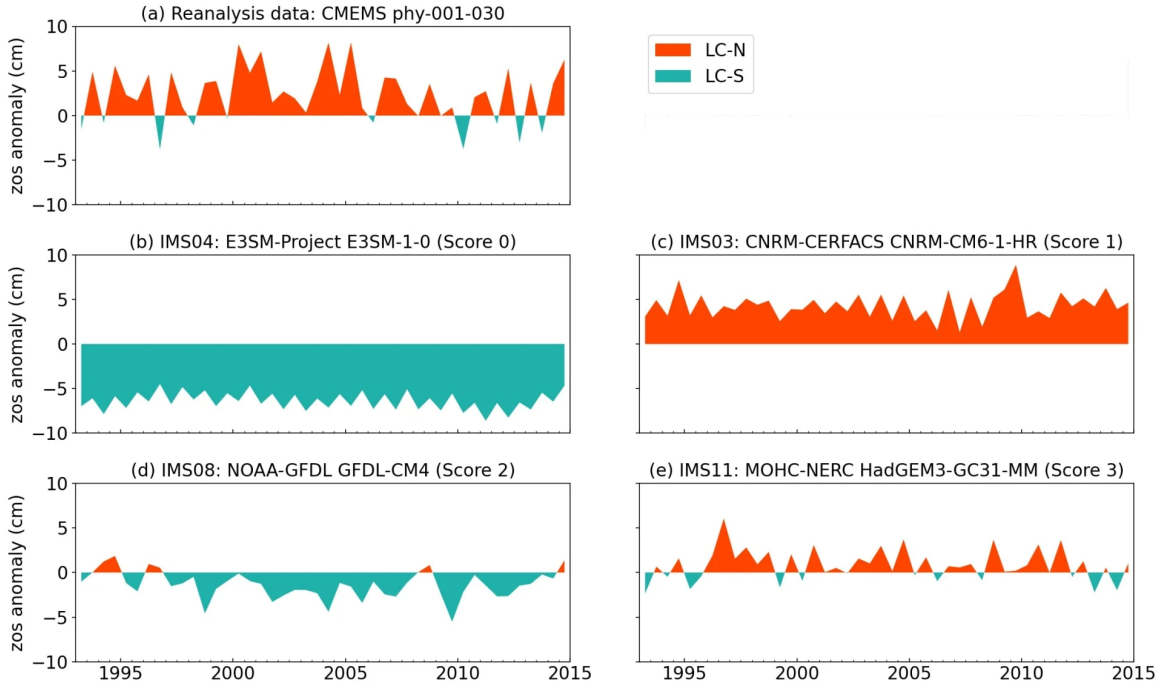


Figure 4. Surface height above geoid (zos) anomaly of (a) reanalysis data, and (b-e) ensemble members. (a) Reanalysis data can produce the physical phenomenon of interest that is the oscillation of Loop Current in the north position (LC-N) and south position (LC-S). Using a prescreening-based subset selection method, (b-c) ensemble members that cannot produce the physical phenomena that is the oscillation of LC-N and LC-S get score of zero and one; (d) ensemble members that can resolve LC-N and LC-S with inaccurate frequency get a score two; and (e) ensemble members that can resolve LC-N and LC-S with frequency similar to reanalysis data get a score 3.

Based on the score of each ensemble member different ensembles can be formed to improve predictive performance.

Figure 5. Temporal match of large bloom/no bloom with Loop Current positions given by (a) observation reanalysis, and (b) by simulations of high resolution ESMs for multi-model ensemble average. Positive and negative bars indicate LC-N and LC-S, respectively. The figure shows that ensemble methods such as prescreening-based subset selection are critical for improving predictive performance. For example, ensemble SME3210 (b) without any model selection and no prior knowledge has less predictive performance in terms for reproducing the LC-N and LC-S trends of the reanalysis data (a). In comparison, ensembles with subset selection based on prescreeing scores (c-e), which exclude members with zero score, show similar Loop current frequency and oscillation trends similar to reanalysis data (a)

DISCLOSURES

This work is funded by NSF Award #1939994. The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

ABSTRACT

The objective of this study is to understand relations between multiple physical and environmental factors and red tide, which is a common name for harmful algal blooms occurring along coastal regions worldwide. Large concentrations of Karenia brevis, a toxic mixotrophic dinoflagellate, make up the red tide along the West Florida Shelf (WFS) in the Gulf of Mexico. Besides being toxic, red tide causes unpleasant odor and scenery, which result in multiple environmental and socioeconomic impacts and public health issues. Understanding the physical and biogeochemical processes that control the occurrence of red tide is important for studying the impact of climate change on red tide frequency, and accordingly assessing the future environmental and socioeconomic impacts of red tide under different mitigation techniques and climate scenarios. We use observation and reanalysis data in the WFS to train machine learning (ML) models to predict red tide, as a classification problem of large bloom or no bloom. We develop the ML model using seasonal input data of Peace River and Caloosahatchee River outflow, alongshore and offshore wind speed, and Loop Current position. The Loop Current, which is a warm ocean current that enters and loops through the Gulf of Mexico before exiting to join the Gulf Stream, can be detected from sea surface height. In addition to the observation and reanalysis data, these variables can be simulated by the Earth system models (ESMs) of the Coupled Model Intercomparison Project Phase 6 (CMIP6), especially by the high-resolution models of the High Resolution Model Intercomparison Project (HighResMIP) of CMIP6. This is needed to understand the frequency and future trends of red tide under different Shared Socioeconomic Pathways (SSPs) of CMIP6. In this preliminary study, we investigate the impact of different choices regarding ML model selection and training dataset on the accuracy of red tide prediction, and the physical interpretation of the results. We also discuss the validation of ESMs data for predictive modeling, and ensemble methods for improving predictive performance. The study provides several insights that can be useful for predicting the future trends of red tide under SSPs using CMIP6 data.

REFERENCES

Maze, G., Olascoaga, M. J., & Brand, L. (2015). Historical analysis of environmental conditions during Florida Red Tide. *Harmful Algae*, *50*, 1–7. https://doi.org/10.1016/j.hal.2015.10.003 (https://www.zotero.org/google-docs/?PKSWRw)