An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model

Andrew C. Ross[a,b,*], Charles A. Stock[b]

[a] *Princeton University Program in Atmospheric and Oceanic Sciences, 300 Forrestal Road, Sayre Hall, Princeton, NJ 08540, USA*
[b] *NOAA Geophysical Fluid Dynamics Laboratory, Princeton University Forrestal Campus, 201 Forrestal Road, Princeton, NJ 08540, USA*

Abstract

Subseasonal to seasonal forecasts have the potential to be a useful tool for managing estuarine fisheries and water quality, and with increasing skill at forecasting conditions at these time scales in the atmosphere and open ocean, skillful forecasts of estuarine salinity, temperature, and biogeochemistry may be possible. In this study, we use a machine learning model to assess the predictability of column minimum dissolved oxygen in Chesapeake Bay at a monthly time scale. Compared to previous models for dissolved oxygen and hypoxia, our model has the advantages of resolving spatial variability and fitting more flexible relationships between dissolved oxygen and the predictor variables. Using a concise set of predictors with established relationships with dissolved oxygen, we find that dissolved oxygen in a given month can be skillfully predicted with knowledge of stratification and mean temperature during the same month. Furthermore, the predictions generated by the model are consistent with expectations from prior knowledge and basic physics. The model reveals that accurate knowledge or skillful forecasts of the vertical density gradient is the key to successful prediction of dissolved oxygen, and prediction skill disappears if stratification is only known at the beginning of the forecast. The lost skill cannot be recovered by replacing stratification as a predictor with variables that have a lagged correlation with stratification (such as river discharge); however, skill is obtainable in many cases if stratification can be forecast with an error of less than about 1 kg m^{-3}. Thus, future research on hypoxia forecasting should focus on understanding and forecasting variations in stratification over subseasonal time scales (between about two weeks and two months).

Keywords:
estuaries, dissolved oxygen, prediction, stratification, USA, Chesapeake Bay

1. Introduction

Chesapeake Bay, a coastal plain estuary located along the Mid-Atlantic Bight, experiences extensive hypoxia and anoxia in the summer following the delivery of nutrients by the spring freshet and the establishment of strong density stratification (Newcombe and Horne, 1938; Taft et al., 1980; Officer et al., 1984). Although there is some evidence that hypoxia has been an occasional feature of the bay for centuries (Karlsen et al., 2000), many studies have identified a dramatic increase in the extent and severity of hypoxia as a result of increased nutrient loading over the last century (Officer et al., 1984; Karlsen et al., 2000; Hagy et al., 2004; Murphy et al., 2011). Other estuaries and coastal systems worldwide exhibit similar increases in hypoxia, primarily as a result of increases in fertilizer runoff and other anthropogenic nutrient inputs (Diaz, 2001; Diaz and Rosenberg, 2008; Rabalais et al., 2010; Breitburg et al., 2018). In the future, climate change and sea-level rise have the potential to alter the intensity and frequency of hypoxia, both in Chesapeake Bay (Najjar et al., 2010; Irby et al., 2018) and globally (Rabalais et al., 2010).

Extensive regulations have been implemented to reduce pollutants in Chesapeake Bay, including nitrogen and phosphorus, with the goal of improving water quality and reducing hypoxia (Linker et al., 2013; Shenk and Linker, 2013). Recently, there has been some evidence that water clarity and dissolved oxygen concentrations have improved (Zhang et al., 2018) and that coverage of submerged aquatic vegetation has expanded (Gurbisz and Michael Kemp, 2014; Lefcheck et al., 2018). However, historically progress has been slow (Boesch, 2006) and currently less than half of the bay area meets all water quality goals (Zhang et al., 2018).

*Corresponding author
Email addresses: andrew.c.ross@noaa.gov (Andrew C. Ross), charles.stock@noaa.gov (Charles A. Stock)

While hypoxia and anoxia are nearly always present in some deep areas of Chesapeake Bay during the summer months, both the timing of hypoxia development and the spatial extent of hypoxia can vary dramatically (Hagy et al., 2004; Scully, 2016b). The susceptibility of the bay to hypoxia and the large interannual variability of hypoxia driven by weather and climate variability pose challenges for water quality and marine resource management (Boesch et al., 2001; Testa et al., 2017). Skillful forecasts of future weather and climate have the potential to improve the management of water quality and fisheries; for example, subseasonal to seasonal scale forecasts of temperature can improve the effectiveness of fisheries management (Hobday et al., 2016; Tommasi et al., 2017). Similarly, Huang and Smith (2011) show that accounting for hypoxia improves management of brown shrimp in the Neuse River Estuary; when hypoxia is more severe, the optimal opening date of the fishery is earlier in the year.

Statistical models have been developed for forecasting the volume of hypoxic water in Chesapeake Bay (Scavia et al., 2006; Liu et al., 2011; Murphy et al., 2011), and although these forecasts are regularly published online and have received attention from the media and general public (Testa et al., 2017), the forecasts are not currently considered in management of Chesapeake Bay water quality or fisheries. One key limitation is that these forecasts predict overall hypoxic volume and provide no information about the spatial distribution of hypoxia. Accounting for spatial variability is an important component of ecosystem based fisheries management (Marasco et al., 2007), and resolving spatial variability is particularly important in Chesapeake Bay because the bay straddles two states (Maryland and Virginia) and has been divided into five categories for regulation of dissolved oxygen and water quality (Batiuk et al., 2009). Additionally, although previous forecast models appear to have modest skill at predicting hypoxic volume, the models have not been thoroughly evaluated for predictive skill beyond the period of data used to fit the forecast models. Therefore, the development of skillful, spatially resolved subseasonal hypoxia forecasts is an essential step for aiding and improving management decisions.

In this study, we assess the predictability of dissolved oxygen at a monthly time scale for many locations in Chesapeake Bay by combining a simple mechanistic set of predictors with flexible machine learning methods. Our objectives are to explore the upper bounds of prediction skill (given perfect knowledge of the mechanistic drivers) and to identify key prediction bottlenecks. Previous forecasts of Chesapeake Bay hypoxia have relied on ordinary or multiple linear regression models (Murphy et al., 2011; Prasad et al., 2011; Testa et al., 2017) or on curves derived from idealized physical models (Scavia et al., 2006; Liu et al., 2011). Machine learning methods, however, have more flexibility to represent nonlinearity, spatial variability, and seasonal changes in the response of dissolved oxygen to predictor variables, thus providing an opportunity for new insights. Several studies have used machine learning methods to predict hypoxia and other biogeochemical and water quality parameters in other estuaries and coastal systems. Park et al. (2015) used regression trees to estimate chlorophyll a given contemporaneous observations of nutrients and water temperature; they found that the regression trees were capable of representing seasonal changes in which inputs were predictive of chlorophyll concentrations. Thoe et al. (2014) compared the ability of a classification tree, an artificial neural network, and three regression methods to predict the presence of fecal indicator bacteria at Santa Monica Beach; they obtained the best performance with the classification tree method. Coopersmith et al. (2010) used the k-nearest neighbor (KNN) algorithm to produce one-day forecasts of hypoxia in Corpus Christi Bay. Coopersmith et al. (2010) also considered the use of regression trees, but the performance of the regression trees was worse than KNN. Tamvakis et al. (2012) found that model trees produced superior predictions of contemporaneous chlorophyll a compared to an artificial neural network and multiple linear regression, and Muhling et al. (2018) used model trees to predict surface temperature and salinity in Chesapeake Bay using projected atmospheric conditions from an ensemble of global climate models as predictors.

To analyze the predictability of spatially resolved dissolved oxygen in Chesapeake Bay, we use a model tree method similar to Muhling et al. (2018). As Park et al. (2015) noted for regression trees, model trees are capable of representing seasonal changes in which inputs are predictive of the response variable; this is potentially useful in Chesapeake Bay because Scully (2016b) suggested that early summer hypoxia was driven primarily by biological processes and that physical influences on hypoxia became more important later in the summer. Also, as Muhling et al. (2018) noted, model trees are capable of extrapolating outside of the range of values in the training observations (although such extrapolations should be treated with caution); this is potentially useful for using the forecast model for scenario simulations to predict the effect of climate change or nutrient loading reductions on hypoxia. In Chesapeake Bay, model trees and similar methods may be more useful than time series methods, such as autoregressive models, because the inter-monthly autocorrelation of dissolved oxygen is low (Section 4.2).

A danger of machine learning methods is the temptation to include diverse predictors with dubious relationships to the variable being predicted. To avoid this, we focus on a distinct set of drivers that have established relationships with dissolved oxygen (Table 1). We begin by testing the predictability of dissolved oxygen under ideal conditions where we have perfect knowledge of the state of the mechanistic predictors in Table 1. Then, we reassess the skill when permutations of the predictors requiring forecasts—temperature, mean sea level and stratification—are only known at the beginning of the forecast period. This reveals stratification and, to a lesser degree, temperature, as key bottlenecks for forecasting hypoxia. We then discuss a) the accuracy of stratification forecasts required for skillful hypoxia forecasts, and b) the viability of replacing stratification as a predictor with a lagged relationship to river discharge.

2. Methods

To predict and forecast dissolved oxygen and hypoxia, we developed a machine learning model that uses a model tree to predict the monthly mean, column minimum dissolved oxygen concentration (hereafter referred to as just dissolved oxygen or DO) at a given location. We refer to this model as a "mechanistic" model because the choice of predictor variables in model was based on mechanisms that are known to influence DO in Chesapeake Bay. These predictor variables, the associated datasets, and the known connections to DO are summarized in Table 1. Based on common availability in all datasets, we used data from 1986 to 2017. These data were split into training and testing groups to fit and evaluate the model; the model was fit to the training dataset, which contained data from years 1986 to 2007, and the model was evaluated using the test dataset, which contained data for the last ten years of the record (2008 to 2017). The choice of years for training and testing does not have a substantial impact on the results; for example, using the first ten years of data as testing instead resulted in a similar model fit, and although there were some differences in skill, our conclusions would not be significantly changed.

Table 1: Variables used as inputs to the mechanistic dissolved oxygen model.

Abbreviation	**Input variable**	**Data source**	**Mechanism and references**
L5	TN load from Susq. River, total over previous 5 months	USGS	Phytoplankton, correlated with river discharge, estuarine circulation, and stratification.
W_{spring}	Mean wind along NE/SW axis, Feb-Apr	NDBC	Transport of phytoplankton biomass; Lee et al. (2013).
$\overline{\mathrm{T}}$	Column-mean temperature anomaly, forecast month	CBP	Solubility and oxygen sinks; Li et al. (2015); Li et al. (2016).
MSL	Mean sea level anomaly, forecast month	PSMSL	Vertical exchange time, estuarine circulation, potentially correlated with stratification; Hong and Shen (2012).
$\Delta\rho$	Vertical density difference anomaly, forecast month	CBP	Mixing.
M	Forecast month		
H	Forecast hour		
D	Profile bottom depth	CBP	
X	Longitude	CBP	
Y	Latitude	CBP	

2.1. Data sources and preprocessing

Vertical profiles of temperature, salinity, and dissolved oxygen were obtained from the Chesapeake Bay Program (CBP) Water Quality Database (Chesapeake Bay Program, 2018). All three variables were typically measured at 1 m intervals in each profile, and the measurements were typically taken bimonthly for each site during the warm season. We selected data only from sites that had frequent observations during May to September in the last 5 years of the training period (2003 to 2007) by requiring that a site have data for at least 20 of the 25 months in this time frame. We did not include sites that were located in the upper reaches of some tributaries and that never experience hypoxia (defined here as column minimum concentration below 2 mg L^{-1}), and we also did not include a cluster of sites in the Elizabeth River near Norfolk that have experienced hypoxia in the past. We assumed that variability in dissolved oxygen in these regions is driven by more localized factors, such as discharge from minor tributaries and point source pollution, compared to the bay mainstem factors considered herein.

For each vertical profile, we calculated the column mean temperature and the column minimum dissolved oxygen concentration. We also obtained density from the temperature and salinity profiles using the International Thermodynamic Equation Of Seawater—2010 (IOC, SCOR and IAPSO, 2010), and we calculated the density stratification as the difference between the density nearest the bottom and nearest the surface (so that a more positive value indicates a more stable density stratification).

We subtracted the climatological mean values from the CBP data to prevent the strong seasonal cycles of dissolved oxygen, temperature, and salinity from overwhelming the interannual variability that we seek to predict. To subtract the climatology from a variable y at a site i, we fit a generalized additive model (Hastie and Tibshirani, 1986; Wood, 2006) with a smooth seasonal cycle and a constant mean:

$$y_{ij} = s_i(DOY_j) + \beta_i + \epsilon_{ij}$$

where s_i is a cyclic cubic spline, DOY_j is the day of year of the j-th observation, β_i is the long-term mean, and ϵ_{ij} is an independent, normally-distributed residual. A separate model was fit for each variable and site using the training dataset. The models were used to predict climatological mean values for each observation in both the training and testing datasets, and the fitted climatological values were subtracted from the observations to produce anomalies. Finally, anomalies were averaged at sites with multiple observations in a given month to produce a time series of monthly anomaly values for each site.

We also calculated lagged values (the value from the previous month) of the mean temperature and density stratification anomalies. At each measurement site, all data (including non-lagged variables) were eliminated if there were no measurements during the previous month. After applying this restriction and the restrictions discussed previously, 126 unique locations remained in the database. A text file providing the names and coordinate information of these 126 locations is provided in the supporting information. The training dataset contained 11,810 vertical profiles, and the test dataset contained 4,936 profiles.

Data for the input of total nitrogen (TN) from the Susquehanna River were obtained from the United States Geological Survey (USGS) (Moyer and Blomquist, 2018). These data were produced by combining observations and the Weighted Regressions on Time, Discharge, and Season method (Hirsch et al., 2010). As input to the model, we used the total nitrogen loading summed over the previous five months. For a June hypoxia prediction, the previous five months are January through May, which matches the period used in other studies (Scavia et al., 2006; Liu et al., 2011; Murphy et al., 2011; Testa et al., 2017).

Observed wind speeds and directions were obtained from the National Data Buoy Center for Thomas Point, MD, a location in the upper Chesapeake Bay near Annapolis, MD. Winds were measured at 18 m above mean sea level. As a predictor in the models, we included mean wind speed along the northeast-southwest direction, averaged over February to April. Lee et al. (2013) suggested that winds along this axis influence the transportation of phytoplankton biomass. Because the Thomas Point station measured winds for only six days during the February to April period of 2010, the mean NE-SW wind for 2010 was determined from the value observed at Rappahannock Light, a station with similar anemometer elevation (16.9 m) located over water closer to the bay mouth. Other periods of missing data for the Thomas Point station were shorter, and the mean February-April wind was determined from all available data from the station.

Monthly mean sea level anomaly at Kiptopeke Beach was obtained from the Permanent Service for Mean Sea Level (Holgate et al., 2013). We chose this location because the data is available for the same time period as the other variables and contains less missing data than most other sites in the bay. Months that were missing in the dataset were imputed with linear interpolation.

2.2. Model for column minimum dissolved oxygen

The machine learning model for dissolved oxygen was built using a model tree (Quinlan, 1992) as implemented and extended by the Cubist package (Kuhn et al., 2018) for R (R Core Team, 2017). In the model tree method, the training data are iteratively partitioned into groups based on the values of the predictor variables, forming a tree that contains a node for each division of the data. A multiple linear regression model is developed for the data at each node of the tree, and the final predicted value is generated from a combination of the regressions along the path of the tree traversed for the given predictors (Quinlan, 1992; Kuhn et al., 2018). Model trees are controlled by a parameter for the number of "rules", which sets the maximum number of partitions of the data included in the model. Cubist allows the addition of "neighbors" to the model, in which case the prediction for a given set of predictors is adjusted by the difference between the actual and predicted values for a specified number of neighboring, similar predictors (Quinlan, 1993). Cubist also includes the option to use "committees", in which case the final prediction is an average of a specified number of model trees that iteratively attempt to balance errors produced by other trees (Kuhn et al., 2018).

We determined the approximate optimal value for each of the three parameters by searching a $4 \times 4 \times 4$ grid containing 25, 50, 100 and 200 rules; 1, 10, 25, and 50 committees; and 0, 1, 2, and 5 neighbors. Each of the 64 parameter sets was evaluated using 10-fold cross-validation, repeated 10 times, with the training dataset. The optimal set of parameters, which minimized the mean squared error of predicted DO over all stations, was 100 rules, 50 committees, and 0 neighbors.

2.3. Model evaluation

The model predictions of dissolved oxygen anomaly were compared with the observations by calculating the Pearson correlation coefficient, the mean bias, and the root mean square error for each site using predictions from the test period. After clustering the sites and calculating cluster mean dissolved oxygen (Section 2.5), we also created target diagrams (Jolliff et al., 2009), which split the root mean square error (RMSE) into two components: bias, and unbiased (centered) RMSE. These components are plotted on the vertical and horizontal axes, respectively, so that the total RMSE is equivalent to the distance from the origin of the target diagram. The plots are normalized by dividing by the RMSE of climatological forecasts during the training period, so that a total RMSE below 1 indicates skill relative to a forecast of the training period climatology. We evaluate skill relative to climatology rather than persistence because the inter-monthly autocorrelation of dissolved oxygen is low (Section 4.2).

2.4. Model sensitivity

To verify that the model tree is physically reasonable and to determine the effect of each input variable and how the model output is ultimately sensitive to the inputs, we visualized the effects of individual terms in the model using plots of individual conditional expectations (ICE) (Goldstein et al., 2015) and the average of the ICEs, known as partial dependence (Friedman, 2001). These plots are commonly used to visualize models where the functional form of the model is not easily interpretable. Following Goldstein et al. (2015), the partial dependence is $f_s = \mathbb{E}_{\mathbf{x}_c}\left[f(\mathbf{x}_s, \mathbf{x}_c)\right]$, where $\mathbf{x}$ is the matrix of predictor variables, s denotes a set of one or more predictor variables for which the partial dependence is calculated, and c is the compliment of this set (the remaining predictor variables). In other words, f_s gives the effect of the variables in s averaged over the other predictor variables. To calculate the partial dependence from the actual data and model, f_s is estimated as

$$\hat{f}_s = \frac{1}{N}\sum_{i=1}^{N}\hat{f}(\mathbf{x}_s, \mathbf{x}_{ci})$$

where $\hat{f}$ is the predicted value from the model and i denotes one of the N observations. To reduce computational costs, we calculated $\hat{f}_s$ for one variable at a time and for 41 evenly spaced values spanning the minimum and maximum values of s observed during the training period. Additionally, we plotted the individual conditional expectations, which are simply the N curves of $\hat{f}$. For a given plot, all curves were standardized by subtracting the value of the curve at the minimum value of s, so that every line originates at zero at the minimum value of s. This allows an easier comparison of the trajectories of the ICE curves as the value of s is increased.

We also calculated the importance of each predictor variable in the mechanistic model. For a given variable, the importance was determined as the percentage of the total number of splits and regressions in the tree where the variable was used (Kuhn et al., 2018). This provides a simple measure of how important each variable is; however, the output from a model tree is also determined by the coefficients in each regression model along the tree, and this is not captured by the importance metric.

2.5. Station clustering

To summarize the ability of the model to predict dissolved oxygen concentrations in different regions of the Bay, we grouped the CBP stations into eight clusters based on location and the percent of observations between May and September with hypoxia (or prevalence of hypoxia) (Figure 1). We first placed all stations where hypoxia never occurred into one cluster. Then, stations from the tributaries on the western side of the bay (Patuxent, Potomac, Rappahannock, and York Rivers) were assigned to clusters for their respective tributaries. Finally, stations in the mainstem (including eastern shore tributaries, which are shorter in length and have fewer stations than those on the western shore) were grouped into three clusters by applying k-means clustering to the latitude and prevalence of hypoxia over all months between May and September in the training period for each station. This neatly groups the stations into a "core hypoxic" region that experiences frequent hypoxia, an "upper bay" cluster that includes stations in the northern half of the bay that experience occasional hypoxia, and a "lower bay" cluster that includes stations in the southern half of the bay that also experience occasional hypoxia.

2.6. Assessing the potential for forecasts

The analyses described above assessed prediction with perfect knowledge of contemporaneous conditions. In a forecast setting, however, the values of essential predictors are not known precisely. We thus considered three experiments to assess the potential for skillful forecasts of future dissolved oxygen concentrations. First, we assessed whether the contemporaneous variables in the mechanistic model (mean temperature anomaly, stratification anomaly, and mean sea level) can be replaced with other variables that are known in advance. We fit this "lagged" model by replacing the contemporaneous variables in the mechanistic model with the values observed during the previous month.

Second, from the results of the core mechanistic prediction analysis (Section 3.2), we found that accurate knowledge of stratification is the key to skillful predictions of dissolved oxygen in Chesapeake Bay. We therefore fit a "correlated" model by replacing stratification as a predictor with lagged river discharge variables that have a correlation with stratification. For this model, daily streamflow for the Susquehanna River at

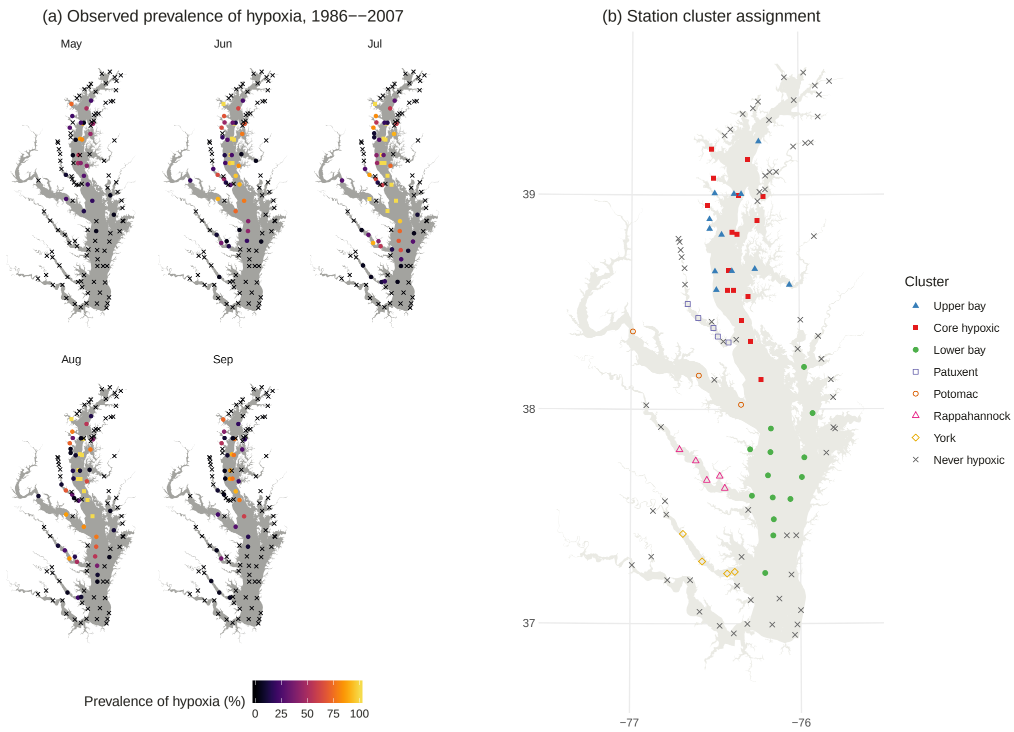


Figure 1: (a) Observed prevalence of hypoxia during the model training period. Black "x"s indicate points where hypoxia was never observed, and squares indicate points where hypoxia was always observed. Circles indicate values between these extremes. (b) Cluster assigned to each station based on geographical position and prevalence of hypoxia during May to September.

Conowingo, MD, the Potomac River near Washington, D.C., and the James River near Richmond, VA were obtained from the USGS. Together, these rivers represent nearly 80% of the typical freshwater discharge to the bay (Boicourt et al., 1999). The streamflow data were averaged monthly, and streamflow anomalies were calculated by subtracting the 1986 to 2007 means for each calendar month. Finally, lagged streamflow anomalies were calculated by taking a rolling average of the anomalies over the previous three months.

Lastly, to assess how accurate stratification forecasts need to be to support skillful hypoxia forecasts, we quantified the degradation of prediction skill in response to imperfect stratification forecasts with increasing levels of noise. We ran simulations where Gaussian random noise with zero mean and various levels of variance was added to the observed stratification during the test period. The simulations assumed perfect spatial error correlation (i.e. in a given simulation, year, and month, all locations have the same error). 100 simulations were conducted for each level of error variance. For each simulation, we used the mechanistic model to predict dissolved oxygen using the temperature, mean sea level, spring winds, and nitrogen loading from the test dataset along with the perturbed stratification data. Then, for each region and calendar month, we calculated the average RMSE over the 100 simulations for each level of variance.

3. Results

3.1. Dissolved oxygen hindcast with mechanistic predictors

With the stratification, mean temperature, and other values observed during the prediction month as inputs, the model tree produces skillful predictions of minimum dissolved oxygen anomalies during the test period. The model predictions have at least moderate correlation with the observations at the majority of sites: over all months, 54% of correlation coefficients are above 0.5 (Figure 2a). A few poor or negative correlations are found in central and lower bay along the thalweg. Except at a few stations, bias is low during the test period (Figure 2b). Over all sites and months, the predictions during the test period are essentially unbiased, with a mean bias of -0.07 mg L^{-1} and the 25th to 75th percentiles spanning -0.3 to 0.2 mg L^{-1}. The bias does tend to become more negative (i.e. model predictions are too low) as the months progress from May (mean bias 0.07 mg L^{-1}) to August (mean -0.2 mg L^{-1}), and several stations along the deep channel also have a large negative bias in September. Despite the biases, the overall model errors are reasonable, with predictions for 70% of all stations and months having lower RMSEs than climatological predictions (Figure 2c). RMSEs are generally low near the mouth of the bay and in some of the tributaries, with slightly higher errors present in the center of the bay. Despite low errors near the mouth of the bay, many points there are not skillful relative to climatology. This suggests that the interannual variation is low at these points, potentially as a result of exchange with saturated water from the shelf. Consistent with results from previous metrics, many points along the thalweg are also not skillful relative to climatology. In the tributaries, despite sometimes having higher RMSEs compared to average, most points are skillful relative to climatology. Overall, 65% of RMSEs are below 1 mg L^{-1}, and the median RMSE is 0.8 mg L^{-1}. To put these values in context, we have included a figure of the mean minimum DO concentration for each station and month in the Supporting Information (Figure S1).

When aggregated to cluster means, the model predictions are generally skillful compared to the training period climatology (Figure 3), as indicated by points inside the solid circles. Overall, skill is highest in June through August, when all regions have lower errors than the climatological reference forecast. Most of the model predictions have lower variances than the observations (indicated by points to the left of the origin). Because the model predictions still have reasonable correlation with the observations (Figures 2a and 3), the model predictions are essentially a smoothed representation of reality. The core hypoxic cluster has lower skill than other clusters due to both larger biases than in other regions and a failure to capture the weak variability of dissolved oxygen in this region. However, because severe hypoxia is nearly always present during the summer months in this region, the lower skill would have a limited impact on predicting the presence or absence of hypoxia. Predictions for the lower bay are skillful for May through August; however, skill declines significantly in September.

3.2. Predictor importance and sensitivity

On average, the model dissolved oxygen predictions are most sensitive to the vertical density stratification (Figure 4). Consistent with physical expectations, the marginal effect of increased stratification is to significantly reduce the concentration of dissolved

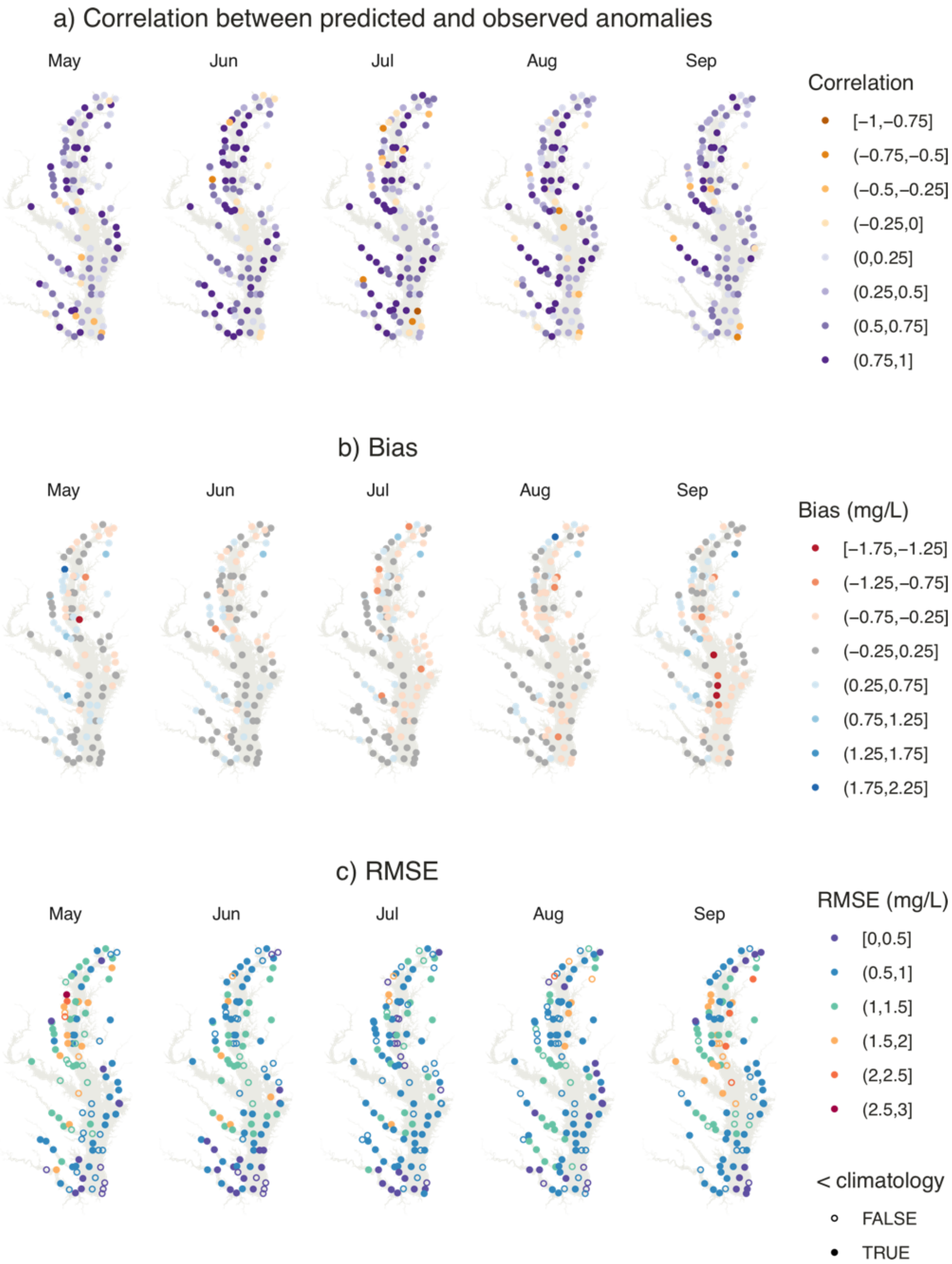


Figure 2: Skill of the main dissolved oxygen model at the station level: correlation coefficient (a), bias (b), and root mean square error (c). Solid points in panel (c) indicate lower errors than a climatological forecast.

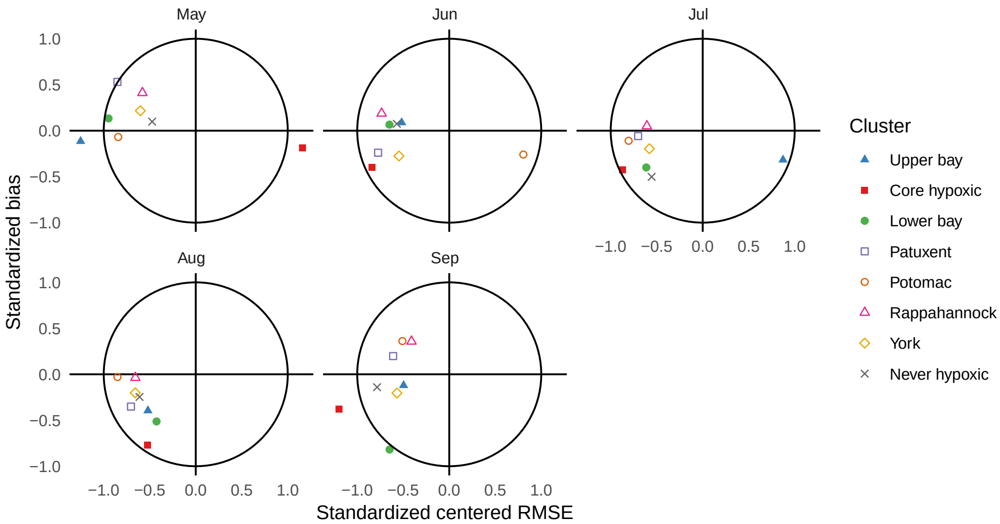


Figure 3: Target diagrams (Section 2.3) for cluster-mean predicted dissolved oxygen. Points inside the circle are considered skillful relative to the training period climatology. Points with a negative standardized centered RMSE have lower interannual variability than the observations.

oxygen. The ICE plots suggest that the marginal effect of stratification is stronger for some conditions or locations than others. A closer investigation showed that points where stratification has a large marginal effect in the model are typically shallow (not shown). This could be interpreted as an effect of the density gradient (an equal density difference over a shallower depth implies a higher, more stable density gradient) or a result of the lower variability of minimum dissolved oxygen in deeper regions.

Warmer water is modeled to have a lower dissolved oxygen concentration, which is consistent with the decreased oxygen solubility and increased biological activity associated with warmer water. Unlike stratification, the effect of temperature is not a strong function of depth. The remaining variables have relatively weak effects on dissolved oxygen on average, although the individual conditional expectations show a fair amount of variability and suggest that interactions with other variables are present. Mean sea level and nutrient loading have weak positive effects on DO on average, while stronger winds from the northeast (positive W_{spring}) have a weak negative effect. Although all three coordinate variables (depth, latitude, and longitude) have zero partial dependence because the model was fit to anomalies, the ICE plots reveal significant interactions with other variables, especially for depth and latitude. In addition to the already noted interaction between stratification and depth, interactions with latitude are not surprising: because Chesapeake Bay is roughly oriented along the north-south axis, most along-channel variations, including variations in tidal amplitude and mean salinity, can be described as functions of latitude. The ICE plots for both latitude and longitude also diverge around 38.5° and -76°, respectively. This region typically has both low dissolved oxygen and frequent hypoxia along the center channel and higher dissolved oxygen and infrequent hypoxia adjacent to the channel and in the Choptank River (Figure 1). The divergence in the ICE plots suggests that the model has learned the difference between these two regions.

The predictor importance metric (Figure 5), which is based on the percent of the splits and regressions in the model tree in which a given variable is used, is generally consistent with the sensitivities revealed in the ICE plots. In the mechanistic model, density stratification remains the single most important variable for predicting dissolved oxygen. Latitude and depth are the two most important coordinate variables. Mean

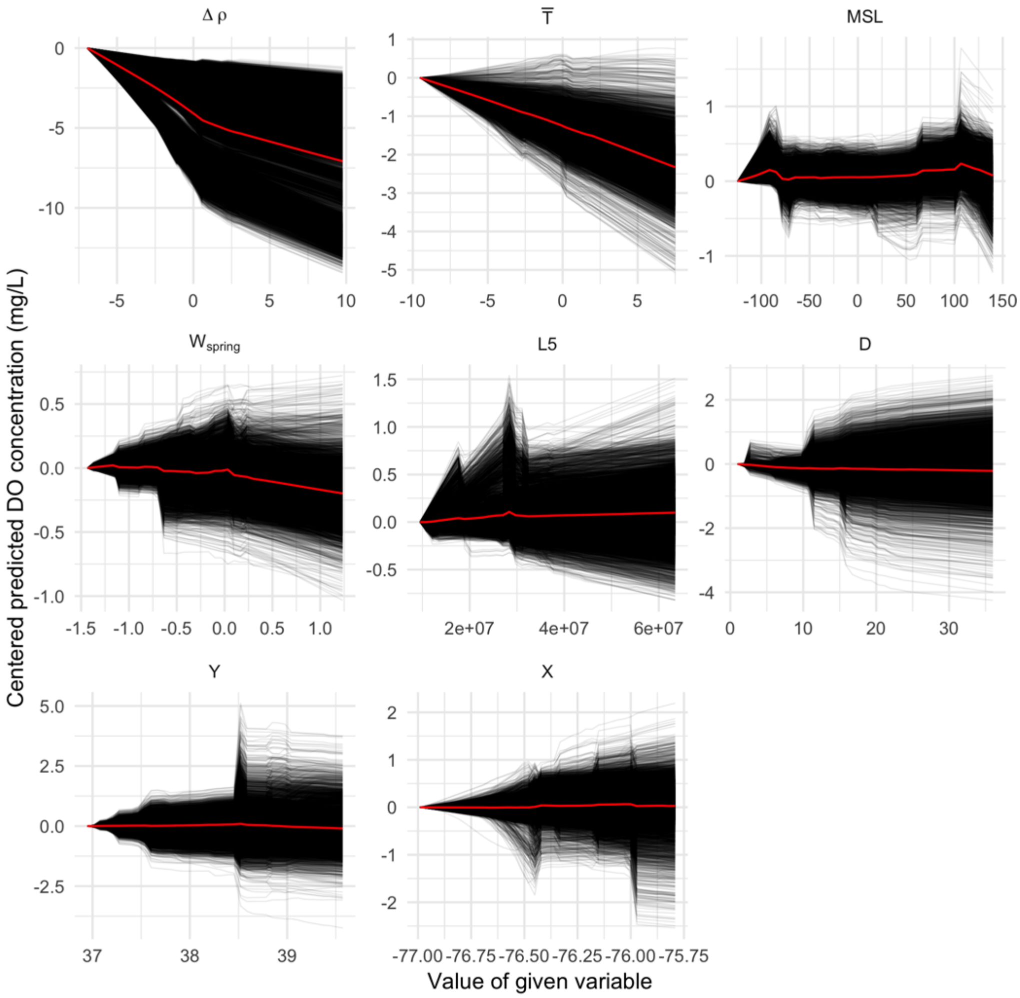


Figure 4: Individual conditional expectations (black lines) and partial dependence (red lines) for several of the predictors in the mechanistic model. Note that the y-axis for each plot is different.

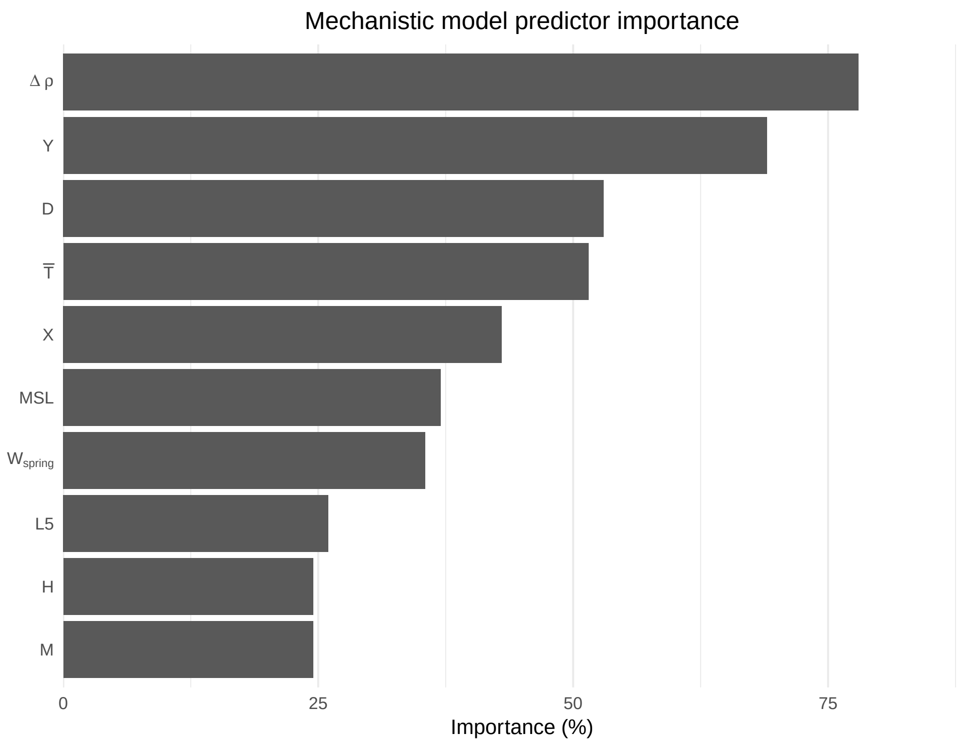


Figure 5: Importance of each variable (Section 2.4). Symbols are defined in Table 1.

temperature anomaly also appears in just over half of the splits and regressions, while sea level, winds, and nitrogen loading are relatively unimportant.

3.3. Limits of predictability

Because the mechanistic model results show that knowledge of stratification is the key to skillful prediction of dissolved oxygen, we consider several modifications to the model (detailed in Section 2.6) to explore the limits of predictability of DO and to potentially make the model useful in a forecast setting where stratification is not perfectly predictable. First, we create a "lagged" model by replacing all contemporaneous variables in the model (mean temperature anomaly, stratification anomaly, and mean sea level) with the values observed during the previous month. This model has significantly reduced skill compared to the mechanistic model (Figure 6); the predicted mean DO for all regions has a higher error than climatology in July, and errors in the remaining months are centered around climatology, with predictions in some regions having comparatively higher skill and predictions in other regions having lower skill. However, the lagged model does improve the mechanistic model prediction skill in a few cases, including in the upper bay and core hypoxic regions in May and in the core hypoxic region in September.

Second, in Figure 6, we test a "correlated" model by replacing the stratification predictor in the mechanistic model with discharge from three major rivers that have a lagged correlation with stratification. This model produces a modest improvement over the lagged model in many regions. The correlated model has some skill in many regions in May and September, and it improves on the mechanistic model predictions in the upper bay and core hypoxic regions in these months, suggesting there is some relationship between lagged river discharge and dissolved oxygen during the fringes of the hypoxia season. However, in nearly all regions during the main summer months, the mechanistic model performs significantly better.

Overall, neither the correlated model nor the lagged model appear to be viable replacements for the mechanistic model, with the possible exception of May and September in the core hypoxic and upper bay regions. This shows that stratification is the key to successful forecasts. In Figure 7, we examine how accurately stratification must be known to allow skillful DO forecasts. Results vary by month and region, but in general the standard deviation of stratification anomaly errors must be less than 1 kg m^{-3} for dissolved oxygen forecasts to be skillful in the majority of the regions (assuming the mean error is zero, i.e. the stratification forecasts are unbiased). Although seemingly small, this error is comparable to the interannual standard deviation of the stratification anomaly (Figure S2). Therefore, skillful dissolved oxygen forecasts would likely be possible if skillful forecasts of stratification were also possible. Predictions for DO in the upper bay and never hypoxic regions are more sensitive to errors in stratification than predictions in other regions; however, these results also have lower interannual variability of stratification, so the potential for predictability remains.

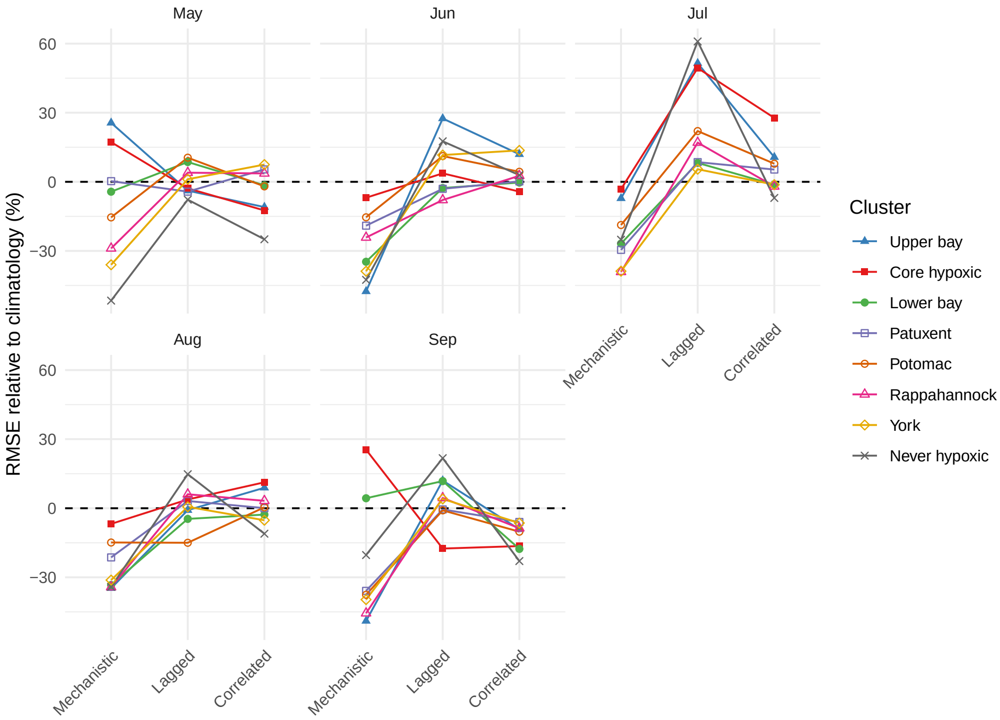


Figure 6: Root mean square error for cluster-mean dissolved oxygen during the test period. Error is normalized by the error of a prediction of climatological (training period) mean dissolved oxygen; negative values indicate errors that are lower than the climatological forecast errors. "Mechanistic" denotes predictions using the mechanistic model; "lagged" indicates predictions from a model where the contemporaneous variables in the mechanistic model are replaced with values observed in the previous month; "correlated" denotes predictions from a model similar to the mechanistic model but with the stratification anomaly replaced with correlated variables (lagged streamflow anomalies).

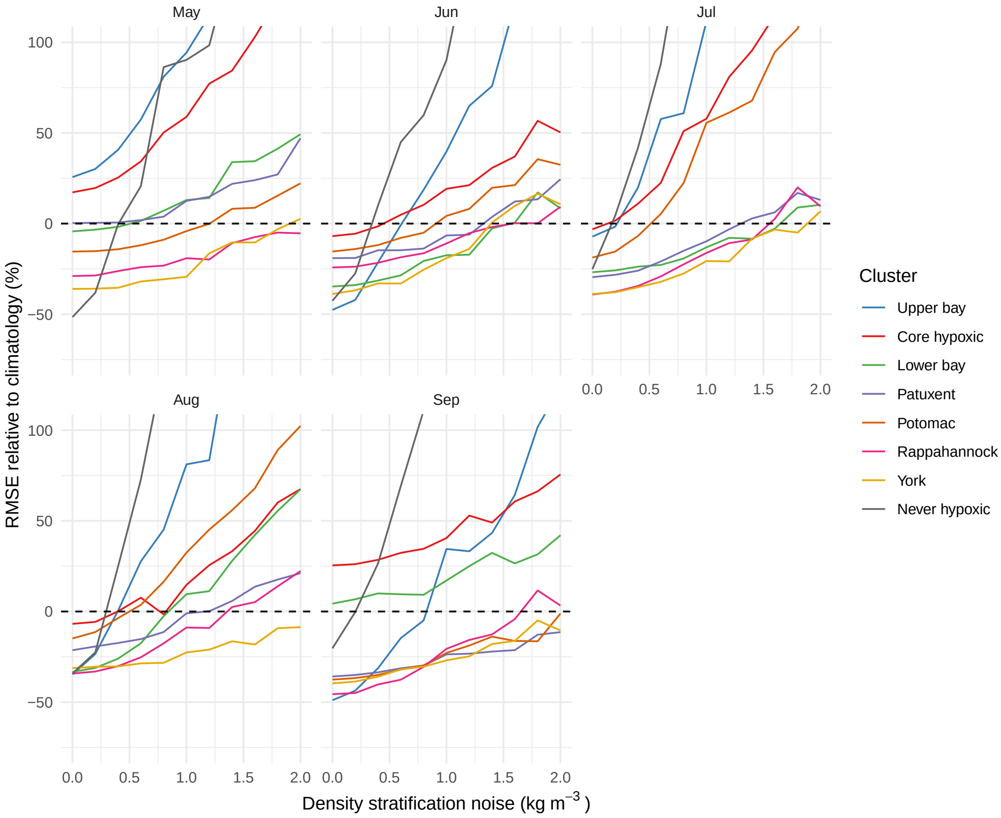


Figure 7: Root mean square error of dissolved oxygen predictions as a function of errors in the stratification anomaly input. RMSE is normalized by the error of a climatological forecast (identical to Figure 6). Stratification noise gives the standard deviation of random Gaussian errors added to the stratification predictor.

4. Discussion

4.1. Summary and comparison with previous studies

The mechanistic model used a concise set of predictor variables that were identified in previous studies as having a potential relationship with dissolved oxygen and hypoxia in Chesapeake Bay. Of the five time-varying variables in the model, we found that stratification and temperature had the largest influences on DO, while nutrient loading had the smallest effect. In this subsection, we summarize our findings on the effects of stratification, temperature, and nutrient loading and compare them with the results of previous studies. The comparison increases our confidence in our finding that stratification and temperature control the interannual variability of dissolved oxygen—particularly since our model, which was built on observations but with no prior assumptions about the form of the relationship between dissolved oxygen and the predictor variables, produced results that are broadly similar to other studies that have used different methods and assumptions.

4.1.1. Stratification is the strongest predictor of dissolved oxygen

The mechanistic model showed that, of the variables considered, stratification is most predictive of dissolved oxygen. This is in agreement with the numerical model results in Cerco and Noel (2013); they found that stratification was the only significant predictor of bottom DO in the deeper waters of Chesapeake Bay. Our result is also partially consistent with the study of observations by Murphy et al. (2011). Murphy et al. (2011) found that stratification had a larger influence than TN load on early July hypoxic and anoxic volumes. In late July, however, Murphy et al. (2011) found that stratification had a negligible influence on hypoxia and anoxia, but stratification during the previous period (early July) had about the same influence on anoxic volumes as TN load. These findings of a strong correlation between DO and stratification are in contrast to Wang et al. (2015), who found that variability in nutrient loading was primarily responsible for interannual variability of anoxic volume. However, Wang et al. (2015) compared anoxic volume over the main bay with stratification observed at a single site (CB4.1C), whereas we have compared stratification measured at each site with concurrent dissolved oxygen measurements. Compared to Wang et al. (2015) and the other cited studies, we have also considered dissolved oxygen concentrations over a broader area including the tributaries and shallow water monitoring stations.

4.1.2. Water temperature has a significant effect on dissolved oxygen

The model in this study identified a stronger and more consistent link between warmer water and lower dissolved oxygen concentrations than previous studies have. Wang et al. (2015) found a weak negative correlation between observed summer mean bottom water temperature and anoxic volume. On the other hand, Hagy et al. (2004) found a weak positive correlation between the date of anoxia onset and the spring mean bottom temperature. Also using observed data, Scully (2016b) found essentially no correlation between summer mean sea surface temperature at Thomas Point and bay-wide hypoxic volume; however, using model simulations, Scully (2016b) found a weak positive correlation between temperature and hypoxic volume.

A possible reason that our model identified a strong and consistent link between temperature and DO is that it used column mean water temperature, which is largely independent of density stratification, as a predictor rather than using surface or bottom temperature. Modeling studies that applied long-term perturbations to atmospheric temperatures, and therefore modified the column mean temperature, have found positive relationships between oxygen and temperature that are similar to this study. For example, Scully (2013) perturbed the seasonal cycle of atmospheric temperature, resulting in a 2 °C change in water temperature and a 25% larger hypoxic volume. Irby et al. (2018) analyzed climate change simulations and concluded that the decrease in bottom DO caused by temperature change will be greater than the changes in bottom DO caused by other climate changes. Irby et al. (2018) found that the effect of temperature on solubility was responsible for 65-85% of the total effect of temperature on DO. Using observed data, Wang et al. (2015) also identified a weak positive correlation between atmospheric temperature and anoxic volume.

A second possible reason for differences between our study and some of the cited previous studies is that we included data from the tributary and shallow water regions that other studies neglected. Muller et al. (2016) found that hypoxia in two smaller tributaries, the Severn and South Rivers, was driven by temperature and temperature stratification more than by salinity and salinity stratification. However, nearly all of the individual conditional expectations in Figure 4 show that increased temperature lowers DO concentration, so the effect of temperature is consistent across different stations and regions.

4.1.3. Nitrogen loading explains a small portion of recent oxygen variability

The mechanistic model produces only a weak sensitivity of dissolved oxygen to total nitrogen loading over the study period, which is consistent with previous studies. Hagy et al. (2004) fit a linear regression to predict July hypoxic volume from January to May nitrate loading; they obtained an R^2 value of 0.17. Murphy et al. (2011) fit linear regressions to predict hypoxic volume from January to May total nitrogen loading and obtained R^2 values of only 0.08 and 0.21 for early and late July hypoxic volume. With only a simple model for oxygen where the oxygen consumption rate is fixed and does not respond to nutrient loading and biological activity, numerical models are still capable of skillfully simulating interannual variability in dissolved oxygen and hypoxic volume (Scully, 2010, 2013, 2016b; Irby et al., 2016). Scully (2016b) noted that despite the lack of any response to nitrogen loading in the model, the model nevertheless produced a strong correlation between nitrogen loading and hypoxic volume, which Scully (2016b) attributed to the increased stratification caused by higher discharge.

It is important to note that although nitrogen loading has only a weak effect on dissolved oxygen in our model, this does not mean that efforts to reduce nitrogen loading to the bay are not worthwhile. First, of the ten predictor variables in the mechanistic model (Table 1), nitrogen loading is the only variable over which humans have some degree of control. Second, the recent interannual variability of nitrogen loading is small compared to the targeted reduction of over 40% (Cerco and Noel, 2013; Linker et al., 2013). In simple simulations using the mechanistic model with nitrogen loading uniformly reduced by 40% over the training period, predicted dissolved oxygen increased significantly, especially over the core hypoxic region (not shown).

4.2. Drivers of oxygen variability not captured by the model

The ability to predict dissolved oxygen using the model in this study is likely to be limited by short-term variability that is not captured in the model. Observations have shown that DO concentrations can fluctuate by several mg L^{-1} over time scales as short as 5 to 15 minutes (Breitburg, 1990; Sanford et al., 1990). These fluctuations are driven by several physical factors, including barotropic tides (Breitburg, 1990) and oscillations of the pycnocline caused by internal tides and winds (Sanford et al., 1990). The short time scales associated with these events, as well as the role of advection from nearby regions, make these fluctuations essentially unpredictable using the model in this study. Because the minimum dissolved oxygen and the stratification and temperature predictors are typically derived from the average of two vertical profiles per month for each measuring site, extreme short-term variability could have also obscured the effects of the predictors in the training and testing data.

Modeling studies (Scully, 2010; Li and Li, 2012) and observations (Scully, 2016a) have also shown the role of winds in driving oxygen variability over time scales of a few days. Some aspects of this variability could be captured in the mechanistic model; for example, stratification also responds to these wind events (Scully et al., 2005; Li and Li, 2011; Xie and Li, 2018). However, when we constructed models that replaced the stratification predictor with various combinations of wind speed and direction averaged over the forecast month, the models did not achieve significant skill at predicting dissolved oxygen. We did not examine skill using wind predictors aggregated over shorter time scales because these winds are essentially unpredictable more than a few days in advance.

An additional potential source of variability and predictability that would not be captured by the model in this study is persistence of dissolved oxygen concentrations from the previous month. However, the inter-monthly correlation of dissolved oxygen in Chesapeake Bay is typically low (Figure S3). Over all months and regions, the only correlation coefficient above 0.5 is between August and September DO in the lower bay region. There is some evidence for higher correlation between months near the beginning and end of the hypoxia season (May—June and August—September). However, even in these months the correlation coefficients are typically between 0.2 and 0.4, and in other months the coefficients are even lower. Not surprisingly, using the minimum DO concentration observed during the previous month as a predictor in the model did not increase the prediction skill.

4.3. Potential changes in the relationship between oxygen and predictor variables over time

The suitability of the machine learning model for predicting future conditions could be restricted by the potential for nonstationarity in the response of oxygen to the forcing variables. For example, some estimates have found that the amount of summer hypoxia produced for a given amount of spring nitrogen loading nearly doubled during the study period (Hagy et al., 2004; Testa and Kemp, 2012). Observations also indicate that hypoxic volumes are increasing in the early summer, but volumes are decreasing in the late summer and hypoxia is breaking up earlier (Murphy et al., 2011). Given the trends in temperature, mean sea level, stratification, and other physical forcings (Murphy et al., 2011; Du et al., 2018), identifying the cause of the nonstationarity has been challenging and several hypotheses have been proposed.

In simulations with numerical models, a trend towards earlier development of hypoxia is consistent with the effect of warmer water (Irby et al., 2018). In this case, it would be possible to capture this effect with the mechanistic model used here. Murphy et al. (2011) suggest that an increasing trend in the strength of stratification explains some of the nonstationarity in hypoxia, which would also be captured by the mechanistic model. However, Testa and Kemp (2012) and Testa et al. (2018) proposed that these trends are a result of changes in nitrogen cycling in the bay as a result of long term hypoxia, which would not be captured by the model used in this paper.

In the mechanistic model, biases became negative during the test period from May to August, especially in the core hypoxic and lower bay regions (Section 3.1). This is consistent with hypoxia breaking up earlier in the test period than during the training period, and suggests that the causes of the earlier breakup are not captured by the predictors included in the mechanistic model. Despite this potential nonstationarity, the model predictions were still skillful compared to climatology during the test period, which suggests that potential nonstationarity will not have a severe impact on model predictions for the near future. Furthermore, as additional observations are collected, the model can be adapted to any nonstationarity by including these observations and adding any variables that are discovered to be causing changes in dissolved oxygen concentrations.

4.4. Comparison of machine learning and other models

While it was not our objective to conduct a comprehensive intercomparison of different methods for modeling dissolved oxygen, in this section we briefly discuss what our work shows may be advantages and disadvantages of the modeling approach used in this study compared to both simpler linear regression models and more complex numerical models.

Compared to simpler linear regression models, model trees and other machine learning methods have a number of potential advantages. For example, the model trees used in this study were able to model dissolved oxygen in different months by including the calendar month as a predictor variable, which was used by the model tree algorithm as a criterion for dividing the data and fitting different regressions. Some studies using linear regression models have adopted a similar, but manual, approach by creating multiple models for different months (e.g., Testa et al. (2017)). Unlike linear regression models, model trees and many other methods are capable of fitting complex and nonlinear relationships between the predictors and the variable being predicted. These advantages can lead to improved prediction skill over linear regression; for example, when we ran simple experiments using a multiple linear regression model with the same predictors as the mechanistic model, the linear regression model had lower skill in the majority of cases. However, complex machine learning models do have disadvantages compared to linear regression. The complex models can be much less interpretable, and the larger number of parameters in the complex models requires the availability of more data for training.

Although machine learning models can be complex, they still have advantages over even more complex numerical biogeochemical models. One clear advantage is computational cost: once optimal parameters have been found using cross-validation (which takes a few hours on a quad core computer), the model tree used in this study can be trained and used to predict years of data in a few seconds. By comparison, we have used a 3D numerical model of Chesapeake Bay in other research that requires over an hour to simulate a single month using a similar computer. A second advantage is the fewer number of parameters and the simpler process for learning these parameters. One disadvantage is that numerical models, which are rooted in fundamental physical principles, are more reliable when extrapolating beyond the range of historically observed conditions (for example, when simulating the effects of climate change). Numerical models also provide predictions of multiple variables simultaneously and allow an easier understanding of the physical reasoning behind the predictions. Overall, the mechanistic model tree appears to have skill that is comparable to the skill that Irby et al. (2016) obtained in a comparison of hindcast simulations from coupled numerical biogeochemical models, although a more detailed comparison is needed.

5. Conclusions

We developed a machine learning model to forecast and predict spatially explicit minimum dissolved oxygen in Chesapeake Bay at monthly time scales. The model results show that accurate knowledge of density stratification is the key to skillful predictions of dissolved oxygen. We developed two alternative models that replaced density stratification with other predictor variables, and neither alternative model was skillful enough to be a viable replacement for the mechanistic model. This suggests that although the mechanistic model is capable of skillfully at predicting dissolved oxygen, accurate forecasts of stratification are necessary to use the mechanistic model to forecast future dissolved oxygen.

Even if machine learning models like the one used in this study are not capable of standing alone as forecast models, they have a number of potential uses, including serving as replacements for complex and expensive biogeochemical model components in a numerical ocean model capable of predicting stratification. With significantly reduced computational costs, additional numerical model ensembles can be run, which will likely increase the accuracy of both subseasonal forecasts and decadal scale climate simulations.

6. Acknowledgements

The authors thank Barbara Muhling for providing useful information about the model tree method and Xiao Liu and Fernando González Taboada for providing helpful reviews of this manuscript. We also thank three anonymous reviewers for providing reviews that improved this manuscript. This report was prepared by the authors under award NA14OAR4320106 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce and with funding from the NOAA Integrated Ecosystem Assessment program. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce. Declarations of interest: none.

References

Batiuk, R.A., Breitburg, D.L., Diaz, R.J., Cronin, T.M., Secor, D.H., Thursby, G., 2009. Derivation of habitat-specific dissolved oxygen criteria for Chesapeake Bay and its tidal tributaries. Journal of Experimental Marine Biology and Ecology 381, S204–S215. doi:10.1016/j.jembe.2009.07.023.

Boesch, D., Brinsfield, R.B., Magnien, R.E., 2001. Chesapeake Bay eutrophication. Journal of Environmental Quality 30, 303–320.

Boesch, D.F., 2006. Scientific requirements for ecosystem-based management in the restoration of Chesapeake Bay and Coastal Louisiana. Ecological Engineering 26, 6–26. doi:10.1016/j.ecoleng.2005.09.004.

Boicourt, W.C., Kuzmić, M., Hopkins, T.S., 1999. The inland sea: Circulation of Chesapeake Bay and the Northern Adriatic, in: Malone, T.C., Malej, A., Harding, L.W., Smodlaka, N., Turner, R.E. (Eds.), Ecosystems at the Land-Sea Margin: Drainage Basin to Coastal Sea. American Geophysical Union, Washington, D. C., pp. 81–129. doi:10.1029/ce055p0081.

Breitburg, D., Levin, L.A., Oschlies, A., Grégoire, M., Chavez, F.P., Conley, D.J., Garçon, V., Gilbert, D., Gutiérrez, D., Isensee, K., Jacinto, G.S., Limburg, K.E., Montes, I., Naqvi, S.W., Pitcher, G.C., Rabalais, N.N., Roman, M.R., Rose, K.A., Seibel, B.A., Telszewski, M., Yasuhara, M., Zhang, J., 2018. Declining oxygen in the global ocean and coastal waters. Science 359. doi:10.1126/science.aam7240.

Breitburg, D.L., 1990. Near-shore hypoxia in the Chesapeake Bay: Patterns and relationships among physical factors. Estuarine, Coastal and Shelf Science 30, 593–609. doi:10.1016/0272-7714(90)90095-9.

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 49, 1119–1133. doi:10.1111/jawr.12107.

Chesapeake Bay Program, 2018. CBP Water Quality Database (1984-present). Https://www.chesapeakebay.net/what/downloads/cbp_water_quality_database_1984_present.

Coopersmith, E.J., Minsker, B., Montagna, P., 2010. Understanding and forecasting hypoxia using machine learning algorithms. Journal of Hydroinformatics 13, 64. doi:10.2166/hydro.2010.015.

Diaz, R.J., 2001. Overview of Hypoxia around the World. Journal of Environment Quality 30, 275. doi:10.2134/jeq2001.302275x.

Diaz, R.J., Rosenberg, R., 2008. Spreading Dead Zones and Consequences for Marine Ecosystems. Science 321, 926–929. doi:10.1126/science.1156401.

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. doi:10.1016/j.scitotenv.2018.02.265.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29, 1189–1232.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics 24, 44–65. doi:10.1080/10618600.2014.907095.

Gurbisz, C., Michael Kemp, W., 2014. Unexpected resurgence of a large submersed plant bed in Chesapeake Bay: Analysis of time series data. Limnology and Oceanography 59, 482–494. doi:10.4319/lo.2014.59.2.0482.

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 4, 634–658. doi:10.1007/BF02907650.

Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. Statistical Science 1, 297–318. doi:10.1214/ss/1177013604.

Hirsch, R.M., Moyer, D.L., Archfield, S.A., 2010. Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. Journal of the American Water Resources Association 46, 857–880. doi:10.1111/j.1752-1688.2010.00482.x.

Hobday, A.J., Spillman, C.M., Paige Eveson, J., Hartog, J.R., 2016. Seasonal forecasting for decision support in marine fisheries and aquaculture. Fisheries Oceanography 25, 45–56. doi:10.1111/fog.12083.

Holgate, S.J., Matthews, A., Woodworth, P.L., Rickards, L.J., Tamisiea, M.E., Bradshaw, E., Foden, P.R., Gordon, K.M., Jevrejeva, S., Pugh, J., 2013. New Data Systems and Products at the Permanent Service for Mean Sea Level. Journal of Coastal Research 288, 493–504. doi:10.2112/JCOASTRES-D-12-00175.1.

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. doi:10.1016/j.ecss.2012.03.014.

Huang, L., Smith, M.D., 2011. Management of an annual fishery in the presence of ecological stress: The case of shrimp and hypoxia. Ecological Economics 70, 688–697. doi:10.1016/j.ecolecon.2010.11.003.

IOC, SCOR and IAPSO, 2010. The international thermodynamic equation of seawater—2010: Calculation and use of thermodynamic properties. Technical Report Manuals and Guides No. 56. Intergovernmental Oceanographic Commission.

Irby, I., Friedrichs, M.A., Da, F., Hinson, K., 2018. The competing impacts of climate change and nutrient reductions on dissolved oxygen in Chesapeake Bay. Biogeosciences 15, 2649–2668. doi:10.5194/bg-15-2649-2018.

Irby, I.D., Friedrichs, M.A., Friedrichs, C.T., Bever, A.J., Hood, R.R., Lanerolle, L.W., 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. doi:10.5194/bg-13-2011-2016.

Jolliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A.M., Helber, R., Arnone, R.A., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. Journal of Marine Systems 76, 64–82. doi:10.1016/j.jmarsys.2008.05.014.

Karlsen, A.W., Cronin, T.M., Ishman, S.E., Willard, D.A., Kerhin, R., Holmes, C.W., Marot, M., 2000. Historical Trends in Chesapeake Bay Dissolved Oxygen Based on Benthic Foraminifera from Sediment Cores. Estuaries 23, 488. doi:10.2307/1353141.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, R., 2018. Cubist: Rule- And Instance-Based Regression Modeling, R package version 0.2.1. Https://topepo.github.io/Cubist.

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, 683–696. doi:10.1007/s12237-013-9592-5.

Lefcheck, J.S., Orth, R.J., Dennison, W.C., Wilcox, D.J., Murphy, R.R., Keisman, J., Gurbisz, C., Hannam, M., Landry, J.B., Moore, K.A., Patrick, C.J., Testa, J., Weller, D.E., Batiuk, R.A., 2018. Long-term nutrient reductions lead to the unprecedented recovery of a temperate coastal region. Proceedings of the National Academy of Sciences , 201715798doi:10.1073/pnas.1715798115.

Li, M., Lee, Y.J., Testa, J.M., Li, Y., Ni, W., Kemp, W.M., Di Toro, D.M., 2016. What drives interannual variability of hypoxia in Chesapeake Bay: Climate forcing versus nutrient loading? Geophysical Research Letters 43, 2127–2134. doi:10.1002/2015GL067334.

Li, Y., Li, M., 2011. Effects of winds on stratification and circulation in a partially mixed estuary. Journal of Geophysical Research: Oceans 116. doi:10.1029/2010JC006893.

Li, Y., Li, M., 2012. Wind-driven lateral circulation in a stratified estuary and its effects on the along-channel flow. Journal of Geophysical Research: Oceans 117. doi:10.1029/2011JC007829.

Li, Y., Li, M., Kemp, W.M., 2015. A Budget Analysis of Bottom-Water Dissolved Oxygen in Chesapeake Bay. Estuaries and Coasts 38, 2132–2148. doi:10.1007/s12237-014-9928-9.

Linker, L.C., Batiuk, R.A., Shenk, G.W., Cerco, C.F., 2013. Development of the Chesapeake Bay watershed total maximum daily load allocation. Journal of the American Water Resources Association 49, 986–1006. doi:10.1111/jawr.12105.

Liu, Y., Arhonditsis, G.B., Stow, C.a., Scavia, D., 2011. Predicting the Hypoxic-Volume in Chesapeake Bay With the Streeter–Phelps Model : a Bayesian Approach. Journal Of The American Water Resources Association 1, 1348–1363.

Marasco, R.J., Goodman, D., Grimes, C.B., Lawson, P.W., Punt, A.E., Quinn II, T.J., 2007. Ecosystem-based fisheries management: some practical suggestions. Canadian Journal of Fisheries and Aquatic Sciences 64, 928–939. doi:10.1139/f07-062.

Moyer, D.L., Blomquist, J.D., 2018. Nitrogen, phosphorus, and suspended-sediment loads and trends measured at the Chesapeake Bay River Input Monitoring stations: Water years 1985-2017. doi:10.5066/P96NUK3Q.

Muhling, B.A., Gaitán, C.F., Stock, C.A., Saba, V.S., Tommasi, D., Dixon, K.W., 2018. Potential Salinity and Temperature Futures for the Chesapeake Bay Using a Statistical Downscaling Spatial Disaggregation Framework. Estuaries and Coasts 41, 349—-372. doi:10.1007/s12237-017-0280-8.

Muller, A.C., Muller, D.L., Muller, A., 2016. Resolving spatiotemporal characteristics of the seasonal hypoxia cycle in shallow estuarine environments of the Severn River and South River, MD, Chesapeake Bay, USA. Heliyon 2, e00157. doi:10.1016/j.heliyon.2016.e00157.

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. doi:10.1007/s12237-011-9413-7.

Najjar, R.G., Pyke, C.R., Adams, M.B., Breitburg, D., Hershner, C., Kemp, M., Howarth, R., Mulholland, M.R., Paolisso, M., Secor, D., Sellner, K., Wardrop, D., Wood, R., 2010. Potential climate-change impacts on the Chesapeake Bay. Estuarine, Coastal and Shelf Science 86, 1–20. doi:10.1016/j.ecss.2009.09.026.

Newcombe, C.L., Horne, W.A., 1938. Oxygen-poor waters of the Chesapeake Bay. Science 88, 80–81. doi:10.1126/science.88.2273.80.

Officer, C.B., Biggs, R.B., Taft, J.L., Cronin, L.E., Tyler, M.A., Boynton, W.R., 1984. Chesapeake Bay Anoxia: Origin, Development, and Significance. Science 223, 22–27. doi:10.1126/science.223.4631.22.

Park, Y., Pachepsky, Y.A., Cho, K.H., Jeon, D.J., Kim, J.H., 2015. Stressor-response modeling using the 2D water quality model and regression trees to predict chlorophyll-a in a reservoir system. Journal of Hydrology 529, 805–815. doi:10.1016/j.jhydrol.2015.09.002.

Prasad, M.B.K., Long, W., Zhang, X., Wood, R.J., Murtugudde, R., 2011. Predicting dissolved oxygen in the Chesapeake Bay: Applications and implications. Aquatic Sciences 73, 437–451. doi:10.1007/s00027-011-0191-x.

Quinlan, J.R., 1992. Learning with continuous classes. Machine Learning 92, 343–348. doi:10.1.1.34.885.

Quinlan, J.R., 1993. Combining Instance-Based and Model-Based Learning. Machine Learning 76, 236–243.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Rabalais, N.N., Díaz, R.J., Levin, L.A., Turner, R.E., Gilbert, D., Zhang, J., 2010. Dynamics and distribution of natural and human-caused hypoxia. Biogeosciences 7, 585–619. doi:10.5194/bg-7-585-2010.

Sanford, L.P., Sellner, K.G., Breitburg, D.L., 1990. Covariability of dissolved oxygen with physical processes in the summertime Chesapeake Bay. Journal of Marine Research 48, 567–590. doi:10.1357/002224090784984713.

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 doi:10.1007/BF02784292.

Scully, M.E., 2010. Wind Modulation of Dissolved Oxygen in Chesapeake Bay. Estuaries and Coasts 33, 1164–1175. doi:10.1007/s12237-010-9319-9.

Scully, M.E., 2013. Physical controls on hypoxia in Chesapeake Bay: A numerical modeling study. Journal of Geophysical Research: Oceans 118, 1239–1256. doi:10.1002/jgrc.20138.

Scully, M.E., 2016a. Mixing of dissolved oxygen in Chesapeake Bay driven by the interaction between wind-driven circulation and estuarine bathymetry. Journal of Geophysical Research: Oceans 121, 5639–5654. doi:10.1002/2016JC011924.

Scully, M.E., 2016b. 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. doi:10.1002/lno.10372.

Scully, M.E., Friedrichs, C., Brubaker, J., 2005. Control of estuarine stratification and mixing by wind-induced straining of the estuarine density field. Estuaries 28, 321–326. doi:10.1007/BF02693915.

Shenk, G.W., Linker, L.C., 2013. Development and application of the 2010 Chesapeake Bay Watershed total maximum daily load model. Journal of the American Water Resources Association 49, 1042–1056. doi:10.1111/jawr.12109.

Taft, J.L., Taylor, W.R., Hartwig, E.O., Loftus, R., 1980. Seasonal Oxygen Depletion in Chesapeake Bay. Estuaries 3, 242. doi:10.2307/1352079.

Tamvakis, A., Miritzis, J., Tsirtsis, G., Spyropoulou, A., Spatharis, S., 2012. Effects of meteorological forcing on coastal eutrophication: Modeling with model trees. Estuarine, Coastal and Shelf Science 115, 210–217. doi:10.1016/J.ECSS.2012.09.003.

Testa, J.M., Clark, J.B., Dennison, W.C., Donovan, E.C., Fisher, A.W., Ni, W., Parker, M., Scavia, D., Spitzer, S.E., Waldrop, A.M., Vargas, V.M.D., Ziegler, G., 2017. Ecological Forecasting and the Science of Hypoxia in Chesapeake Bay. BioScience 67, 614–626. doi:10.1093/biosci/bix048.

Testa, J.M., Kemp, W.M., 2012. Hypoxia-induced shifts in nitrogen and phosphorus cycling in Chesapeake Bay. Limnology and Oceanography 57, 835–850. doi:10.4319/lo.2012.57.3.0835.

Testa, J.M., Kemp, W.M., Boynton, W.R., 2018. Season-specific trends and linkages of nitrogen and oxygen cycles in Chesapeake Bay. Limnology and Oceanography doi:10.1002/lno.10823.

Thoe, W., Gold, M., Griesbach, A., Grimmer, M., Taggart, M., Boehm, A., 2014. Predicting water quality at Santa Monica Beach: Evaluation of five different models for public notification of unsafe swimming conditions. Water Research 67, 105–117. doi:10.1016/J.WATRES.2014.09.001.

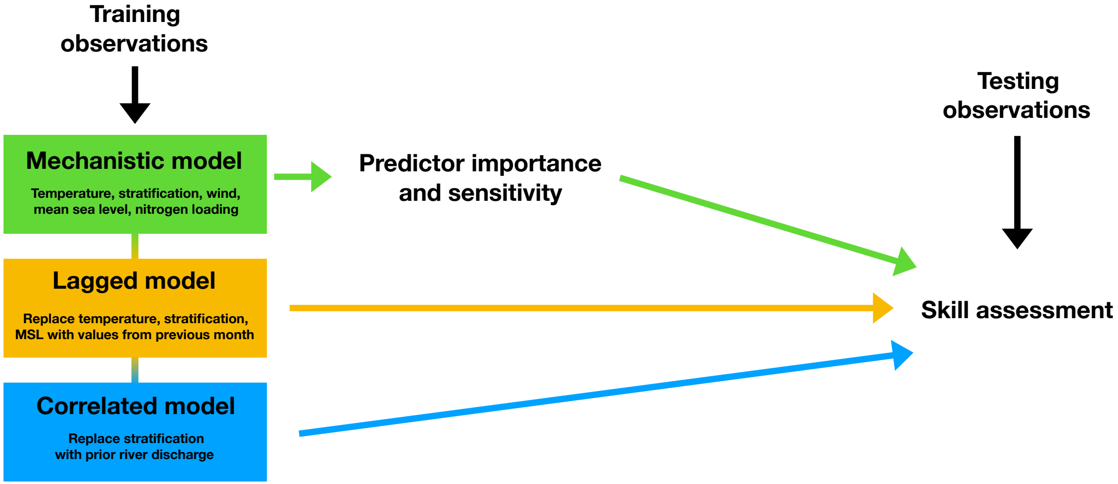
Tommasi, D., Stock, C.A., Pegion, K., Vecchi, G.A., Methot, R.D., Alexander, M.A., Checkley, D.M., 2017. Improved management of small pelagic fisheries through seasonal climate prediction. Ecological Applications doi:10.1002/eap.1458.

Wang, P., Wang, H., Linker, L., 2015. Relative Importance of Nutrient Load and Wind on Regulating Interannual Summer Hypoxia in the Chesapeake Bay. Estuaries and Coasts 38, 1048–1061. doi:10.1007/s12237-014-9867-5.

Wood, S.N., 2006. Generalized additive models: an introduction with R. Chapman and Hall/CRC, New York.

Xie, X., Li, M., 2018. Effects of Wind Straining on Estuarine Stratification: A Combined Observational and Modeling Study. Journal of Geophysical Research: Oceans 123, 2363–2380. doi:10.1002/2017JC013470.

Zhang, Q., Murphy, R.R., Tian, R., Forsyth, M.K., Trentacoste, E.M., Keisman, J., Tango, P.J., 2018. Chesapeake Bay's water quality condition has been recovering: Insights from a multimetric indicator assessment of thirty years of tidal monitoring data. Science of the Total Environment 637-638, 1617–1625. doi:10.1016/j.scitotenv.2018.05.025.


Training observations
Mechanistic model
Temperature, stratification, wind, mean sea level, nitrogen loading
Lagged model
Replace temperature, stratification, MSL with values from previous month
Correlated model
Replace stratification with prior river discharge
Predictor importance and sensitivity
Testing observations
Skill assessment