Author Manuscript

Article type : Article

Journal: Ecological Applications

Manuscript type: Articles

Running Head: Bayesian estuarine hypoxia forecasting

Advancing estuarine ecological forecasts: seasonal hypoxia in Chesapeake Bay

Donald Scavia[1,*], Isabella Bertani[2], Jeremy M. Testa[3], Aaron J. Bever[4], Joel D. Blomquist[5], Marjorie A.M. Friedrichs[6], Lewis C. Linker[7], Bruce D. Michael[8], Rebecca R. Murphy[2], Gary W. Shenk[9]

[1] School for Environment and Sustainability, University of Michigan, Ann Arbor, MI USA 48103

[2] University of Maryland Center for Environmental Science, Chesapeake Bay Program Office, Annapolis, MD USA 21403

[3] Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD USA 20688

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/EAP.2384

This article is protected by copyright. All rights reserved

[4] ANCHOR QEA, LLC, San Francisco, CA USA 94111

[5] Water Observing Systems Program, U.S. Geological Survey, Baltimore, MD USA 21228

[6] Virginia Institute of Marine Science, William & Mary, Gloucester Point, VA USA 23062

[7] U.S. EPA Chesapeake Bay Program Office, Annapolis, MD USA 21403

[8] Department of Natural Resources, Annapolis, MD USA 21401

[9] U.S. Geological Survey Chesapeake Bay Program Office, Annapolis, MD USA 21403

[*]Corresponding Author: scavia@umich.edu

Manuscript received 26 January 2021; revised 28 April 2021; accepted 26 May 2021.

ABSTRACT

Ecological forecasts are quantitative tools that can guide ecosystem management. The co-emergence of extensive environmental monitoring and quantitative frameworks allows for widespread development and continued improvement of ecological forecasting systems. We use a relatively simple estuarine hypoxia model to demonstrate advances in addressing some of the most critical challenges and opportunities of contemporary ecological forecasting, including predictive accuracy, uncertainty characterization, and management relevance. We explore the impacts of different combinations of forecast metrics, drivers, and driver time windows on predictive performance. We also incorporate multiple sets of state-variable observations from different sources and separately quantify model prediction error and measurement uncertainty through a flexible Bayesian hierarchical framework. Results illustrate the benefits of 1) adopting forecast metrics and drivers that strike an optimal balance between predictability and relevance to management, 2) incorporating multiple data sources in the calibration dataset to separate and propagate different sources of uncertainty, and 3) using the model in scenario mode to probabilistically evaluate the effects of alternative management decisions on future ecosystem state. In the Chesapeake Bay, the subject of this case study, we find that average summer or total

Author Manuscript

This article is protected by copyright. All rights reserved

annual hypoxia metrics are more predictable than monthly metrics and that measurement error represents an important source of uncertainty. Application of the model in scenario mode suggests that absent watershed management actions over the past decades, long-term average hypoxia would have increased by 7% compared to 1985. Conversely, the model projects that if management goals currently in place to restore the Bay are met, long-term average hypoxia would eventually decrease by 32% with respect to the mid-1980s.

KEY WORDS: Bayesian, Chesapeake Bay, Forecasts, Hypoxia

Author Manuscript

INTRODUCTION

Stakeholders, resource managers, and policy makers need to base their decisions on the best available knowledge of how natural resources are expected to respond to environmental and anthropogenic change. Making accurate and reliable quantitative ecological predictions is one of the key challenges faced by contemporary applied ecology (Carpenter 2002; Evans et al. 2013; Moquet et al. 2015). In response to this need, a growing number of efforts have advanced ecological forecasting (Coreau et al. 2009; Luo et al. 2011; Payne et al. 2017; Ross et al. 2020). Previously defined as "the process of predicting the state of ecosystems, ecosystem services, and natural capital, with fully specified uncertainties" (Clark et al. 2001), ecological forecasts seek to not only strengthen linkages between management questions and relevant research, but also to advance scientific knowledge of mechanisms underlying ecosystem dynamics (Testa et al. 2017a; Dietze et al. 2018).

While forecasts of atmospheric conditions have long been a feature of climate science and operational weather forecasting, ecological forecasts have been less frequently applied given the challenges of modeling ecological systems and limitations of adequate data (e.g., Petchey et al. 2015). Nonetheless, the potential for ecological forecasts to guide and improve management decisions has sparked interest beyond academic settings, with several government agencies investing resources and supporting initiatives to explore its development and application. The United States National Oceanographic and Atmospheric Administration (NOAA) has a long

This article is protected by copyright. All rights reserved

history of both experimental and operational forecasts in areas such as harmful algal blooms, hypoxia, fisheries, and pathogens (Valette-Silver and Scavia 2003; NOAA 2020), and other US agencies have sponsored similar efforts (Bradford et al. 2020; NASA 2020). A recently launched Ecological Forecasting Initiative represents the first broad effort to bring all these experiences together and foster the development of an interdisciplinary forecasting community (EFI 2020).

Despite growing interest and an increasing number of applications, there is currently no broad consensus on the ultimate predictability of ecological systems and the ability of models to generate reliable predictions to inform policy (Beckage et al. 2011; Schindler and Hillborn 2015). This may be partly because most ecological forecasting efforts are relatively recent and lack sufficiently long track records that build confidence. In addition, rigorous out-of-sample forecast skill assessment is not always performed (Johnson-Bice et al. 2020) either because forecasts are made over time frames (decades to centuries) that prevent timely comparisons with observed data (Dietze et al. 2018) or because protocols are not in place for regular forecast validation with new observations (White et al. 2019). Finally, although modeling approaches that quantify multiple sources of uncertainty are becoming increasingly common (Harwood and Stokes 2003; Clark 2005; Gimenez et al. 2014; Salon et al. 2019; Scavia et al. 2020c), a rigorous treatment of uncertainty is often missing (Dietze et al. 2018). This may result in overly confident forecasts that do not capture the full range of possible outcomes, thereby potentially leading to inadequate preparedness and loss of trust in models when observations fall outside of (underestimated) uncertainty bounds (Pappenberger and Beven 2006; Raftery 2016).

Models of oxygen dynamics date back a century or more (e.g., Streeter and Phelps 1925) and forecasts of hypoxia extent are perhaps one of the most established and mature examples of routine and operational ecological forecasts. Such forecasts for the Gulf of Mexico date back almost two decades (Scavia et al. 2003, 2017), followed in more recent years by similar efforts in other systems, such as the Chesapeake Bay (Scavia et al. 2006; Testa et al. 2017a; VIMS 2020b; Bever et al. 2021), Lake Erie (NOAA GLERL 2020), and the Neuse River Estuary (Katin et al. 2019; North Carolina Sea Grant 2020). Among these, the Chesapeake Bay has a 14-year transparent record of ecological forecast performance based on regular comparisons of predictions with out-of-sample observations (e.g., Scavia and Bertani 2020) and model validation (Evans and Scavia 2011). Since 2007, a statistical model that incorporates simple biophysical processes has been used to forecast mid-summer hypoxic volume (HV) in the Chesapeake Bay as

Author Manuscript

This article is protected by copyright. All rights reserved

a function of total nitrogen (TN) loads from the largest tributary to the Bay (Susquehanna River) (Scavia et al. 2006). Each year, the model`s forecast is assessed at the end of the season by comparing it to hypoxia observations made by monitoring agencies (Maryland DNR 2020; Scavia and Bertani 2020). Informed by this continuous validation and assessment process, the model has been revised over the years with a focus on improving performance and uncertainty characterization (Stow and Scavia 2009; Liu et al. 2011). Testa et al. (2017a) showed that these forecasts contributed substantially to public awareness and support for management actions in the Chesapeake Bay, in addition to helping advance fundamental understanding of ecological processes driving oxygen depletion in estuarine settings.

In this work, we build on the Chesapeake Bay hypoxia case study and present an enhanced version of the forecasting model that addresses some of the most critical challenges, opportunities, and best practices of contemporary ecological forecasting. These include identifying predictors and metrics of ecosystem state that improve model performance and management relevance, explicitly accounting for and propagating multiple sources of uncertainty, evaluating forecasting performance through hindcasting, and applying the model to answer management questions (Dietze et al. 2018; Harris et al. 2018; White et al. 2019; Carey et al. 2021). Guided by recent appreciation for the spatial distribution of nutrient sources that affect the Bay`s water quality, how loads have changed over time, and the complex intra-annual variability in hypoxia, we explore how model performance changes when different combinations of HV metrics, TN load sources, and TN load time windows are used as calibration inputs. We also take advantage of the model's flexible Bayesian framework to better characterize uncertainty by including multiple data sources (i.e., multiple sets of HV estimates) during calibration through a hierarchical approach that separates model prediction error and HV measurement error. Finally, we validate the model through hindcasting and showcase the use of the model for scenarios by predicting hypoxic conditions (with associated probability distributions) under alternative nutrient management scenarios routinely evaluated by the Chesapeake Bay Program (CBP), the Partnership that leads restoration efforts in the Bay.

METHODS

Historical context and management background

Author Manuscript

This article is protected by copyright. All rights reserved

Like many coastal ecosystems worldwide, water quality of the Chesapeake Bay, the largest estuary in the continental US, declined as a result of human activity over at least the last century (Kemp et al. 2005). Loss of submerged aquatic vegetation (Kemp et al. 2005), altered benthic macroinvertebrate production (Sturdivant et al. 2013), and extensive hypoxia (e.g., Hagy et al. 2004) are among the water quality impairments caused by elevated nutrient inputs, land use changes, and resource extraction. Extensive efforts have been in place to reduce nitrogen (*N*), phosphorus (*P*), and sediment (*S*) inputs since the 1980s, with the goal of improving water quality and reducing hypoxia (Linker et al. 2013; Shenk and Linker 2013). The United States Environmental Protection Agency (US EPA), working together with federal, state, local, and non-governmental partners, established a Total Maximum Daily Load (TMDL) in 2010 for *N*, *P*, and *S* (US EPA 2010). To meet the TMDL load reduction targets, state and local governments are responsible for developing Watershed Implementation Plans (WIPs) that describe needed management practices. Coincident with these efforts, which have also included point source decreases (Ator et al. 2020) and reductions in atmospheric nitrogen deposition (Eshleman et al. 2013; Da et al. 2018), water clarity and dissolved oxygen (DO) concentrations have improved some (Zhang et al. 2018) and submerged aquatic vegetation has expanded in some regions (Gurbisz and Kemp 2014; Lefcheck et al. 2018). However, progress has been slow (Boesch 2006) and currently less than half of the Bay area meets all water quality goals (Zhang et al. 2018).

One of the primary TMDL goals is raising DO concentrations to levels suitable for upper trophic levels (e.g., invertebrates, finfish). Low oxygen concentrations have contributed to decreased fish habitat, catch per unit effort (Buchheister et al. 2013), and blue crab harvests (Mistiaen et al. 2003), as well as reductions in production of benthic macroinvertebrates (Sturdivant et al. 2014) that serve as forage for many demersal fish. Although there is some evidence for recent improvements in DO in certain periods or when considering specific metrics (Murphy et al. 2011; Zhang et al. 2018), the overall annual volume of water with oxygen less than 2 mg/L (~63 mM) has changed little over the past 3-4 decades (Testa et al. 2018; Bever et al. 2018).

In support of nutrient control efforts, the CBP uses complex airshed, watershed, and water quality models (US EPA 2010) to determine oxygen concentration targets (Irby and Friedrichs 2019), but other predictive models have been used to both forecast and study oxygen dynamics

Author Manuscript

This article is protected by copyright. All rights reserved

(e.g., Testa et al. 2014; Irby et al. 2016, 2018; Da et al. 2018; Du et al. 2018; Moriarty et al. 2020), including the model presented here (Scavia et al. 2006).

Model overview

The model used here is an adaptation of the Streeter-Phelps model that simulates DO depletion in rivers downstream from a point source of organic matter (Streeter and Phelps 1925). It has been applied extensively to rivers and estuaries (Chapra 1997), as well as to the northern Gulf of Mexico (Scavia et al. 2003, 2004, 2006, 2017, 2020b) and the Chesapeake Bay (Scavia et al. 2006, 2019; Liu et al. 2011; Evans and Scavia 2011).

The model simulates subpycnocline DO concentration profiles along the mainstem of the Chesapeake Bay via subpycnocline net advection, organic matter decomposition and oxygen consumption, and oxygen flux from the surface layer. Assuming a correspondence between the measured extent of summer hypoxia and that which would be achieved at steady state, the steady state solution to the model is:

$$DO = DO_s - \frac{k_d BOD_u F}{k_r - k_d}\left(e^{-k_d \frac{x}{\nu}} - e^{-k_r \frac{x}{\nu}}\right) - D_i e^{-k_r \frac{x}{\nu}} \qquad \text{Eq. 1}$$

where DO = dissolved oxygen (mg/L), DO_s = oxygen saturation (mg/L), k_d = organic matter decay coefficient (1/day), k_r = reaeration coefficient (1/day), BOD_u = initial organic matter (mg/L), x = upstream distance (km), F = fraction of organic matter sinking below the pycnocline (unitless), D_i = initial oxygen deficit (mg/L), and ν = net advection (km/day). Because the reaeration coefficient k_r is known to vary with distance down estuary x, the model calculates $k_r = b_x K$, where b_x takes on different values over the length of the estuary that approximate the known spatial variation in k_r (Scavia et al. 2006; Evans and Scavia 2011) and K is a unitless scaling parameter estimated by the model. While ν represents river advection in the original Streeter-Phelps formulation, here it is a parameterization of the combined effects of horizontal transport and all ecological processes resulting in subsequent settling of organic matter from the surface. Therefore, it is a bulk parameter with no simple physical analog.

Nitrogen load is a surrogate for organic matter deposited below the pycnocline at the model origin (220 km down Bay from the Susquehanna River mouth), with model distance following

Author Manuscript

This article is protected by copyright. All rights reserved

the up-estuary flow of bottom water. Specifically, nitrogen load is converted to organic carbon (*C*) via the Redfield *C*:*N* ratio (106:16 or 5.67 g *C*/ g *N*), and then converted to BOD_u via the respiration ratio O_2:*C* (0.9, or 2.4 g O_2/ g *C*) (Scavia et al. 2006). In the original model, organic matter loading was assumed proportional to Jan-May Susquehanna River TN load; in this study additional load sources and time windows were tested (see below).

The Bay mainstem is divided into 137 1-km long segments and Eq. 1 is applied to estimate the steady state subpycnocline DO concentration at each segment *j* and in each year *i* (DO_{ij}). The overall length of the model-predicted hypoxic region in each year *i* (L_i) is then calculated by summing the lengths (l_{ij}) of all segments where DO_{ij} is less than 2 mg/L (Eqs. 2 and 3) and HV (V_i) is calculated from L_i using an empirical relationship (Eq. 4) derived from Chesapeake Bay measurements (Scavia et al. 2006):

$$L_i = \sum_{j=1}^{137} l_{ij} w_{ij} \qquad \text{Eq. 2}$$

$$w_{ij} = \begin{cases} 1, \; DO_{ij} < 2 \\ 0, \; DO_{ij} \geq 2 \end{cases} \qquad \text{Eq. 3}$$

$$V_i = 0.000391 \times L_i^2 \qquad \text{Eq. 4}$$

Other assumptions include: transport results from advection rather than longitudinal dispersion, subpycnocline oxygen consumption can be modeled as a first-order process proportional to organic matter concentration, oxygen flux across the pycnocline can be modeled as a first-order process proportional to the difference between surface and bottom layer oxygen concentrations, and subpycnocline organic matter oxygen demand is proportional to TN load. Tests of these assumptions, as well as calibration to average July subpycnocline oxygen concentration profiles and HVs from 1950 to 2003, have been described elsewhere (Scavia et al. 2006). Annual forecasts provided each spring since 2007 were shown to be rather robust (Scavia and Bertani 2020; Testa et al. 2017a).

Nitrogen load sources and time frames

We assembled TN loads from major tributaries and point sources downstream of the tributary monitoring stations (Figs. 1 and Appendix S1: Fig. S1) and tested various combinations of load

Author Manuscript

This article is protected by copyright. All rights reserved

sources and time frames as model drivers. Monthly TN loads estimated from 1985-2018 at stations located near the head of tide of nine major tributaries (Susquehanna, Potomac, James, Rappahannock, Appomattox, Pamunkey, Mattaponi, Patuxent, and Choptank) were from the United States Geological Survey (https://doi.org/10.5066/F7RR1X68). Estimates of TN loads from point sources located downstream of these tributaries were from the CBP (Chesapeake Bay Program 2017). Monthly point source loads are based on wastewater facility monthly flow and constituent concentration data submitted by the jurisdictions to the Integrated Compliance Information System National Pollutant Discharge Elimination System (ICIS-NPDES) and subsequently reviewed and quality checked by the CBP. On average, these nine tributaries and point sources make up approximately 77% of the 1990-2018 average annual TN load (calculated from https://www.chesapeakeprogress.com/?/clean-water/water-quality). We explored model performance using each of the following combinations of sources: Susquehanna alone, Potomac alone, Susquehanna + Potomac, Susquehanna + Potomac + point sources, all nine major tributaries, all nine major tributaries + point sources.

To evaluate the impact of different loading time frames on model performance, for each of the load source combinations described above, we calculated loads from the preceding year's October and each succeeding month through April (e.g., Oct-Apr, Nov-Apr, Dec-Apr, Jan-Apr, Feb-Apr, Mar-Apr, Apr), and then similar sequences through May, June, and July. We first screened candidate load windows by calculating the Pearson's correlation coefficient between HV metrics and different combinations of TN load windows × TN load sources. Initial explorations revealed that regardless of the TN load sources considered, load time windows ending in April or earlier never improved correlations compared to time windows that considered loads through May or later, so we only included time windows ending in May or later. In addition, correlations between HV metrics and TN loads in the Oct-Jul window were generally comparable to, or worse than, those obtained with Oct-May and Oct-Jun. Because of that, and considering that hypoxia forecasts are typically released in early June (i.e., before the July loads can be reliably predicted), we focused model calibration exercises on all possible sequential combinations of months in the Oct-May and Oct-Jun time windows.

Hypoxic volume metrics

This article is protected by copyright. All rights reserved

Author Manuscript

As part of the CBP`s long-term Water Quality Monitoring Program, Virginia and Maryland state agencies and partners have collected vertical profiles of DO since 1984 and made the data available through the CBP`s online data server (Chesapeake Bay Program 2020). Roughly 30-60 stations in the mainstem portion of the Bay are sampled semi-monthly in June through August and monthly throughout the remainder of the year, with vertical profiles collected at about 1-2 m vertical resolution. These data have been used by numerous groups to estimate the extent of hypoxia in the Chesapeake Bay (Bever et al. 2013, 2018; Zhou et al. 2014; Hagy et al. 2004; Murphy et al. 2011).

Previous versions of the model were calibrated to average July HV estimated through interpolation of DO measurements from a subset of the mainstem stations mentioned above by Hagy et al. (2004) and by Murphy et al. (2011) in more recent years (Scavia et al. 2019). The month of July was originally selected because that is when HV often reaches its seasonal maximum. However, retrospective assessments of forecast performance revealed consistent overprediction of July HV in years characterized by anomalous weather events (Testa et al. 2017a). In addition to that, different metrics may capture different aspects of an ecosystem`s status and metrics other than the seasonal maximum HV may be more relevant to stakeholders and decision makers depending on the specific ecological management target. For example, managers interested in assessing spawning habitat availability for a benthic species that tends to spawn in June would be more interested in average June HV. On the other hand, total annual HV may be the preferred metric when tracking watershed management progress over time, because it may be less sensitive to year-specific transient weather events and may better capture the cumulative effects of changes in nutrient loads over time. One of the goals of our analysis was thus to assess how model performance changed when different HV metrics were used as calibration endpoint to 1) identify which metrics may lead to improved forecasting performance and 2) provide stakeholders and managers with useful information on each metric`s predictability.

To compare model performance for different combinations of HV metrics, load sources, and load time frames while maintaining an interpolation method consistent with previous model versions, we used the updated time series (1985-2018) of HV estimates generated following Murphy et al. (2011). Murphy et al. (2011) apply two-dimensional (depth-length) ordinary kriging to DO observations collected during semi-monthly cruises at 21 stations along the main channel of the

This article is protected by copyright. All rights reserved

Bay. The interpolated DO profile estimated along the main channel for each cruise is assumed to remain constant across the mainstem and is extended laterally to estimate cruise-specific HV based on previously published cross-sectional volumes.

We tested six different HV metrics in the model's calibration (Figs. 2 and Appendix S1: Fig. S2): average of the two cruise-specific HVs for each month for June through September (km^3), average summer (defined as June-September) HV (km^3), and total annual HV (km^3*days). In cases when only a single cruise was available in a month (typically in September and sporadically in other months), that cruise`s value was taken as the monthly HV. Total annual HV was estimated by multiplying each cruise-specific HV by the number of days until the following cruise and then summing these values over each year (Bever et al. 2013).

Hypoxic volume interpolation methods

We considered two additional sets of HV estimates to investigate the influence of the interpolation methods on variability in HV estimates and model predictive uncertainty. We note that we use the terms "variability" and "model predictive uncertainty" to indicate, respectively, the range of variation of an outcome (e.g., HV) around its mean and the stochastic error component that estimates that variation within a model (e.g., the residual error term in a regression model) (Gelman and Hill 2007; Hofman et al. 2020). The different sets of HV estimates were generated using different subsets of DO profile stations as well as different interpolation methods. Zhou et al. (2014) performed universal kriging on cruise-specific DO profiles from approximately 40 stations located across the mainstem of the Bay. Bever et al. (2018) used the CBP volumetric inverse distance-squared interpolator (US EPA 2003) with DO profiles from a subset of 13 stations along the mainstem and in the lower Potomac River. Differences in cruise-specific HVs across these three methods (hereafter referred to as Murphy, Zhou, and Bever) are expected as a result of several factors, including differences in the interpolation approaches and relevant methodological choices (e.g., DO profile stations used), the bathymetry used in the interpolations, and the spatial extent over which interpolation was carried out.

Zhou et al. (2014) and Murphy et al. (2011) limited their spatial extent to the mainstem, while Bever et al. (2018) extended interpolations to the tributaries. To adjust for these differences

Author Manuscript

This article is protected by copyright. All rights reserved

while preserving the individual inter-annual variability, we scaled Murphy and Zhou HVs to Bever's using the average long-term ratio of mainstem-only HV to Bay-wide HV simulated by the CBP Water Quality and Sediment Transport Model (WQSTM). A comparison with long-term ratios of mainstem-only HV to Bay-wide HV calculated using HVs estimated by the CBP volumetric interpolator over the period 1985-2013 indicated that ratios estimated by the CBP WQSTM and the CBP interpolator are largely comparable (Appendix S1: Fig. S3). Because average ratios calculated for individual months and total annual HV did not differ substantially, we applied the total annual HV ratios to Zhou's and Murphy's monthly, average summer, and total annual HV metrics.

To quantify uncertainty due to HV estimation error and model prediction error separately, we used a hierarchical modeling approach to expand the original model formulation and simultaneously calibrate the model to the three sets of HV estimates (Obenour et al. 2014). The three individual HV estimates in each year *i* are modeled as arising from a normal distribution with mean y_i and standard deviation σ_{est} (Eq. 5). In this formulation, y_i represents the true, unknown HV in year *i* and is itself modeled as arising from a normal distribution with mean equal to the deterministic model prediction in year *i* as defined in Eqs. 1 and 4 ($\mathrm{V_i}$) and standard deviation σ_{res} (Eq. 6):

$$vol_{i,j} \sim Normal\left(y_i, \sigma_{est}^2\right) \qquad \text{Eq. 5}$$

$$y_i \sim Normal\left(V_i, \sigma_{res}^2\right) \qquad \text{Eq. 6}$$

where $\mathrm{vol}_{i,j}$ represents the HV estimate from method *j* (with *j*=1 for Murphy, *j*=2 for Bever, and *j*=3 for Zhou) in year *i* and the two stochastic terms σ_{est} and σ_{res} represent uncertainty deriving from HV estimation error and model prediction error, respectively.

Calibration and model skill assessment

The original model (Scavia et al. 2006) was a Monte Carlo implementation that accommodated potential variation in the bulk parameter ν. It was subsequently reformulated within a Bayesian framework (Evans and Scavia 2011; Liu et al. 2011) to account for uncertainty in additional parameters. In the present study, the model was calibrated under the range of conditions

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

described above using Bayesian fitting conducted with the software WinBUGS version 1.4.3 (Lunn et al. 2000; Gelman and Hill 2007) interfaced with R version 3.5.2 (R Core Team 2018) through the package R2WinBUGS version 2.1-21 (Sturtz et al. 2005). All model parameters were kept constant across years. The two parameters quantifying sources of uncertainty (σ_{est} and σ_{res}) are represented as precisions in WinBUGS (τ_{est} and τ_{res}, where $\tau = 1/\sigma^2$) and were assigned weak priors: τ_{est}, τ_{res} ~ Gamma(0.001, 0.001), while all other parameters were given the same priors used in the most recent model applications: *K*~Normal(0.6, 0.2)I[0, 1]; *F*~Normal(0.5, 0.5) I[0, 1]; k_d~Normal(0.11, 0.05)I[0,]; and ν~Normal(2.5, 0.77)I[0,], where the Gamma distribution is defined by the shape and rate parameters, the Normal distribution is defined by the mean and standard deviation, and 'I[]' denotes censoring to restrict values above 0 (I[0,]) or between 0 and 1 (I[0, 1]) (Evans and Scavia 2011; Liu et al. 2011). We ran four Markov Chain Monte Carlo chains with 5,000 iterations each and checked convergence by ensuring that $\hat{R}$<1.1 for all model parameters. We assessed how model performance changed when using multiple sets of HV estimates and different combinations of HV metrics, TN load sources, and TN load time windows using a combination of several metrics: the Nash-Sutcliffe Efficiency (NSE), the square of the correlation coefficient between observed and predicted values (r^2), the root mean square error (RMSE), the mean absolute error (MAE), and the residual standard error (RSTDE) (see Appendix S1 for a description of how each metric was calculated). Specifically, we evaluated all metrics simultaneously and assessed whether all metrics agreed in indicating which model performed best. By ensuring a high level of agreement among different metrics we aimed at providing a more comprehensive and robust assessment of the models` performance. When multiple sets of HV estimates were used in model calibration, all individual HV estimates from the different sets were used to calculate model performance metrics.

For the models exhibiting the best predictive performance according to the metrics defined above, we also computed the coverage of the 95% prediction intervals (i.e., the fraction of the observations that fell within the intervals) and the Continuous Ranked Probability Score (CRPS) (Matheson and Winkler 1976). The CRPS quantifies the error between the cumulative distribution function of a model`s prediction and that of the corresponding observed value, thereby providing an assessment of the calibration and sharpness of the predictive distributions (Gneiting and Katzfuss 2014). We used the R package scoringRules version 1.0.1 (Jordan et al. 2019) to calculate a CRPS value for each observation and then obtained a mean CRPS value for

This article is protected by copyright. All rights reserved

each model by averaging across all observations. We then calculated a CRPS skill score (Eq. 7) by comparing each model`s CRPS (CRPSmodel) with that of a respective benchmark null model (CRPSbenchmark) that does not have TN load as the predictor and thereby essentially corresponds to a constant-only model that predicts HV simply based on the historical long-term average (Pappenberger et al. 2015; Thomas et al. 2019):

$$CRPS\,skill\,score = 1 - \frac{CRPS_{model}}{CRPS_{benchmark}} \qquad \text{Eq. 7}$$

Because lower CRPS values indicate better performance, with zero corresponding to a perfect prediction, a CRPS skill score of 1 indicates a perfect prediction, values above zero indicate that a model is more skillful than its respective benchmark null model, and conversely values below zero indicate that a model performs worse than the benchmark.

Response curves and scenarios

Response curves were developed for the two best performing models by generating HV predictions, with 95% credible and prediction intervals, for a range of TN loads. The response curves were then used to estimate HVs for a set of alternative management scenarios routinely evaluated by the CBP:

- ***1985 FN*** and ***2018 FN*****:** Obtained by summing flow-normalized loads from all nine tributaries plus point sources in 1985 and 2018, respectively. Flow normalization (Hirsch et al. 2010) removes the influence of year-to-year variability in river flow, thereby providing an estimate of the amount of change in loads between 1985 and 2018 that may be attributed to changing nutrient sources, management actions, and other factors.
- ***2020 No Action*****:** Obtained by multiplying each tributary`s 1985 flow-normalized load by the ratio of 2020 No Action/1985 Progress Real Air scenario loads estimated for that tributary`s sub-watershed by the CBP partnership`s watershed model CAST (Chesapeake Bay Program 2017). Tributary loads were then summed together with point sources from the CAST 2020 No Action scenario. The 2020 No Action scenario estimates the long-term average loads expected given a constant 2020 land use, human and livestock populations, and cropping systems, but with management practices, point sources, septic loads, and atmospheric deposition set as if no actions had been taken to control nutrients since 1985.

Author Manuscript

This article is protected by copyright. All rights reserved

The 1985 Progress Real Air scenario estimates the long-term average loads expected from the watershed at each monitoring station given a constant 1985 land use, management practices, point sources, septic loads, cropping systems, livestock, and nutrient inputs of fertilizers, manure, N fixation, and atmospheric deposition.

- ***WIP3 Planning Targets*****:** Obtained by multiplying each tributary`s 2018 flow-normalized load by the ratio of Phase 3 Watershed Implementation Plan (WIP3) Planning Targets/2018 Progress Real Air scenario loads. Tributary loads were then summed together with point sources from the CAST WIP3 scenario. The WIP3 Planning Targets represent loads consistent with the Bay's TMDL (US EPA 2010) that are expected to achieve target water quality goals.
- ***WIP3 Actual*****:** In some cases, the WIP3s submitted by the states did not meet the WIP3 Planning Targets. WIP3 Actual was obtained by multiplying each tributary`s 2018 flow-normalized load by the ratio of the actual WIP3 plans submitted by the states/2018 Progress Real Air scenario loads estimated by CAST. Tributary loads were then summed together with point sources from the CAST WIP3 Actual scenario. The WIP3 Actual scenario estimates the long-term average loads expected if the WIP3s submitted by the states are completed, using modeled 2025 land use and population conditions. The 2018 Progress Real Air scenario is defined similarly to the 1985 Progress Real Air scenario defined above.

RESULTS

Total nitrogen loads and hypoxic volume metrics

Annual TN loads are dominated by the Susquehanna and Potomac rivers, followed by point sources that enter below the monitoring stations (Fig. 1). There was considerable inter-annual variability driven largely by precipitation. Highest loads occurred in especially wet years (e.g., 2003, 2004, 2011) and lowest loads in drier years (e.g., 1999-2002). Loads were typically highest in March and April, lowest in July and August, and most variable in September (Appendix S1: Fig. S1).

There was also substantial inter-annual variability in HV. The three interpolation methods showed relatively coherent patterns for total annual HV, summer average HV, and most of the

Author Manuscript

This article is protected by copyright. All rights reserved

individual months (Figs. 2 and Appendix S1: Fig. S2 and Table S1), with particularly large HV in 1998, 2003, and 2001, and relatively smaller volumes in 2001, 2002, and 2012. When averaged across the three sets of estimates, the smallest annual HV occurred in 2002 (557 ± 30 km^3*days) and the largest in 2003 (1235 ± 240 km^3*days). In most years HV peaked in July and declined between August and September, although there was substantial inter-annual seasonal variability and in some years the largest HVs occurred in June or August. The largest monthly HV was in July 2011. Using the coefficient of variation as an estimate of inter-annual variability, all three estimates exhibited substantially higher inter-annual variability in monthly HVs compared to summer average and total annual HV (Appendix S1: Table S1).

Model calibration

Based on general agreement among the performance metrics, the best fits (i.e., highest NSE, highest r^2, lowest RMSE, and lowest MAE) for total annual, summer average, and August HV were achieved when driven with Jan-Jun loads from all tributaries plus point sources (Table 1; Fig. 3). The June and July HV best fits were obtained with slightly different TN load sources and periods (Table 1), but their second-best models were also based on loads from all tributaries and point sources and were virtually indistinguishable from the best models' performance. Interestingly, models calibrated to only Susquehanna loads never ranked among the ten best-performing models for any of the HV metrics considered here. As an example, based on NSE the best performing models driven by TN loads from only the Susquehanna River explained 28% and 23% of the inter-annual variability in total annual and average July HV, respectively, compared to 52% and 29% obtained when using loads from all tributaries and point sources (Table 1). All models exhibited a CRPS skill score > 0, indicating that all models represented an improvement in performance compared to the respective null models, and the percentage of observations that fell within the 95% prediction intervals ranged between 94% and 100% (Table 1).

The highest model performances were obtained for average summer and total annual HV (Table 1). The monthly HV models performed better earlier in the season (e.g., June and July) compared to late summer (e.g., August and September), and the load time frames tested here had no predictive power for September HV.

Author Manuscript

This article is protected by copyright. All rights reserved

To more rigorously assess the performance of the overall best model (i.e., the one calibrated to total annual HV and driven by Jan-Jun loads from all tributaries and point sources), we generated blind forecasts for the years when regular forecasts were made (i.e., starting in 2007). To forecast each year, we calibrated the model using data up to the preceding year. This provides a more realistic estimate of how the model would perform when predicting outside of the calibration dataset. When run in this blind forecast mode, 100% of the left-out, post-2006 observations fell within the 95% prediction intervals and the CRPS skill score was equal to 0.14, indicating an improvement in performance compared to a corresponding null model run in blind forecast mode. Values of NSE indicated that the blind forecast total annual HV model explained 47% of the variability in HV when considering all years in the 2007-2018 window, and 58% of the variability in HV when excluding three years characterized by mid-summer disruptive weather events (2007, 2014, and 2018; Fig. 2). For comparison, when calibrated to only Susquehanna TN loads, the model explained 23% and 27% of the variability in total annual HV across all years and "normal" weather years, respectively.

Sources of uncertainty

When calibrating the best-performing models (i.e., average summer and total annual HV driven by Jan-Jun loads from all tributaries plus point sources) to three sets of HV estimates simultaneously, predictive performance (average summer: NSE = 0.39, r^2 = 0.52, RMSE = 1.11, MAE = 0.89; total annual: NSE = 0.50, r^2 = 0.60, RMSE = 136, MAE = 107) was comparable to that of the models calibrated using the same inputs but one set of HV estimates only (Table 1). Model prediction error (σ_{est}) and HV estimation error (σ_{res}) were similar, suggesting that the two sources of uncertainty are of comparable magnitude (Appendix S1: Table S2). The 95% prediction intervals accounting for parameter uncertainty, model prediction error, and HV estimation error contained the corresponding observed values 97% of the times for both models, and were on average 20% wider than those accounting for only parameter uncertainty and model prediction error (Fig. 4). The CRPS was equal to 75 km^3 (total annual HV) and 0.63 km^3 (average summer HV) while the CRPS skill score was equal to 0.26 (average summer HV) and 0.34 (total annual HV), indicating that the models performed better than the corresponding benchmark null models. Although model residuals did not show a clear trend over time, the ratio

Author Manuscript

This article is protected by copyright. All rights reserved

of total annual or summer average HV over the Jan-Jun TN load exhibited a significant positive trend using the two sets of HV estimates (Murphy and Bever) with complete records over 1985-2018 (Appendix S1: Fig. S4).

Author Manuscript

Response curves and scenarios

Parameters from the best models were used to construct HV-load response curves for summer average and total annual HV (Fig. 4). The best-estimate curve indicates that, based on flow-normalized loads, total annual HV declined on average from 930 km^3*days (95% credible interval, or CI: 840-1005 km^3*days) to 770 km^3*days (95% CI: 640-870 km^3*days) between 1985 and 2018 (Fig. 4a and Appendix S1: Table S2). These estimates are not meant to characterize HV in a specific year, but rather to quantify the change in HV predicted by the model between two given time periods over the long-term after averaging out the influence of inter-annual variability in TN loads due primarily to freshwater flow variability.

We also explored load reductions associated with specific management scenarios generated by the CBP Partnership`s watershed model CAST. The results suggest that had there been no point or nonpoint source management actions, long-term average HV would have increased to 995 km^3*days (95% CI: 910-1085 km^3*days) by 2020. The model also projects that if the TMDL is reached, long-term average HV would decrease to 635 km^3*days (95% CI: 440-785 km^3*days), or to 660 km^3*days (95% CI: 480-785 km^3*days) if the WIP3 Actual reductions are reached. This TMDL-based HV reduction represents 18% (95% CI: 10-32%) and 32% (95% CI: 22-49%) reduction from 2018 and 1985 flow-normalized conditions, respectively. Similar results were found for summer average HV (Appendix S1: Table S2).

For both total annual and summer average HV, TN load changes occurring at relatively high loads produce relatively small changes in HV. But, as loads decrease the curve's slope becomes steeper and the HV change per unit TN load increases, suggesting HV reductions may become more responsive as loads continue to decrease.

DISCUSSION

This article is protected by copyright. All rights reserved

Predictability of different HV metrics - Hypoxic extent metrics used for forecasts, scenarios, and reporting across several systems have often been estimates of summer maximum volume or area (e.g., Liu et al. 2011; Scavia et al. 2003, 2006, 2016, 2017; Testa et al. 2017a; Obenour et al. 2012, 2015; Rucinski et al. 2016; Bocaniov and Scavia 2016; Zhang et al. 2016; but see Katin et al. 2019; Del Giudice et al. 2020; Ross et al. 2020). However, these maxima are not necessarily representative of year-long conditions. For example, years with particularly large July HV, the metric historically used to forecast hypoxia in the Chesapeake Bay, do not always exhibit comparably large total annual HV and vice versa (Fig. 2; Bever et al. 2013; VIMS 2020b). Our results showed that summer average and total annual HV are considerably easier to predict than monthly HV (Table 1). This is largely because short-term meteorological events that increase vertical mixing and lateral advection of bottom water can temporarily decrease HV (Goodrich et al. 1987; Scully 2010a; Testa et al. 2017b). While these HV disruptions are often relatively short-lived, they increase variability at monthly scales and may lead to substantial overprediction on short time scales (Testa et al. 2017a). Similar disruptions of seasonal hypoxia occur in other systems (Turner et al. 2012; Bocaniov and Scavia 2016), leading to either incorporate weather-related drivers or to shift to hypoxia metrics that better integrate conditions throughout the year (Bever et al. 2013, 2018; Feng et al. 2012; Obenour et al. 2015; Matli et al. 2018, 2020).

In addition to being less sensitive to variability caused by episodic weather events, total annual HV better captures cumulative effects of year-to-year variability in nutrient loads, as illustrated by the largest improvement in performance when relating this metric to a more comprehensive estimate of total watershed loads (Table 1). Annual HV also has the benefit of incorporating climate change effects because it combines hypoxic volume and duration into one metric without being biased by climate-driven shifts in the timing or location of hypoxia (Irby et al. 2018). By representing a more integrated, annual-scale estimate of oxygen depletion, total annual HV may also capture a broader measure of living resource habitat limitation over the annual cycle.

However, monthly forecasts might be more informative if they capture more temporally dynamic representations of hypoxia severity within a year. Given the wide range of oxygen vulnerability among marine species (e.g., Vaquer-Sunyer and Duarte 2008), forecasts that quantify periods of both low and high hypoxia severity during a year may allow for more species-specific quantification of potential habitat loss and physiological stress. For example, many benthic invertebrates, which are an important forage base for finfish communities, can tolerate some

Author Manuscript

This article is protected by copyright. All rights reserved

degree of hypoxia (e.g., Modig and Olafsson 1998), while more severe hypoxia has more widespread ecosystem effects (Vaquer-Sunyer and Duarte 2008; Sturdivant et al. 2014). Thus, as some organisms may be able to tolerate modest and extensive hypoxia but cannot tolerate the most severe periods (Brady et al. 2009), it might be important to trade increased uncertainty for the shorter-term metric. Tradeoffs like this will likely play out in developing most ecological forecasts, where the chosen time frame for prediction is ultimately a function of the ecological target of interest and may include indices for both duration and spatial extent to represent the time-space integration of habitat availability.

Uncertainty characterization - Quantifying and communicating uncertainty is crucial when forecasts and scenarios are used for environmental decision making (Clark et al. 2001; Harwood and Stokes 2003; Irby and Friedrichs 2019). A rigorous and transparent characterization of forecast uncertainty enables stakeholders and policy makers to a) get a realistic picture of the current state of scientific knowledge of the process being predicted, b) quantitatively evaluate the risk associated with a range of possible future outcomes and make decisions accordingly, and c) prioritize future investments to fill knowledge gaps that are responsible for the largest sources of uncertainty (Pappenberger and Beven 2006; Dietze et al. 2018). The relative magnitude of different error sources provides useful insights on where to focus future research efforts to reduce forecast error (Obenour et al. 2014; Bertani et al. 2016; Del Giudice et al. 2020). The hierarchical approach demonstrated here provides a means to quantify multiple sources of uncertainty, including parameter uncertainty, model prediction error, and HV measurement error. While model predictive performance did not change when incorporating multiple sets of HV estimates, the separate characterization of measurement and prediction error led to wider, but more realistic, prediction intervals (Cressie et al. 2009). The ability to explicitly separate different sources of uncertainty also allowed us to develop different types of predictive intervals, depending on which types of uncertainty are of interest (Fig. 4; See "Management Application").

Reducing measurement error - We found that uncertainty associated with HV estimates is an important component of the overall predictive uncertainty (Fig. 4). As a result, efforts to improve HV estimates and reconcile differences across multiple sets of HV estimates have the potential to reduce forecast uncertainty. This is consistent with findings in other systems where a thorough analysis of uncertainty has revealed that accurately capturing temporal dynamics of complex

This article is protected by copyright. All rights reserved

Author Manuscript

ecological processes such as harmful algal blooms and hypoxia is still a major limitation to reducing forecast error (Del Giudice et al. 2020; Scavia et al. 2020c).

While few monitoring programs have the resources needed for the intensive monitoring required to accurately capture metrics such as algal and oxygen dynamics, advances in three-dimensional ecological modeling, space-time geostatistical estimation, and their fusion provide sophisticated interpolations of limited survey data. For example, as computational power increased and three-dimensional ecological models have become more sophisticated, they have been used to both provide insights into oxygen dynamics and integrate point estimates across time and space to generate continuous time series of hypoxia (Bever et al. 2013; Fennel et al. 2016; Katin et al. 2019). Geostatistical techniques are also being used to augment discrete monitoring data and generate enhanced estimates of algal blooms and hypoxia dynamics integrated over space and time with quantified uncertainty (Murphy et al. 2011; Obenour et al. 2013; Zhou et al. 2013, 2014; Matli et al. 2018; Fang et al. 2019). Matli et al. (2020) combined these two approaches by using output from a three-dimensional ecological model as covariates in their space-time geostatistical analysis for the Gulf of Mexico, reducing prediction uncertainty by 11-40% compared to using measurement alone. As these modeling and geostatistical approaches improve, together with the ever-increasing availability of high-frequency sensors and remote sensing products, the ability to expand beyond the limitations of traditional monitoring will allow for more integrative and accurate ecosystem metrics used in forecast and scenario development. The hierarchical framework presented here also allows for the estimation of separate measurement errors for sets of metrics that are known to be characterized by markedly different measurement uncertainty.

Reducing model error - Model error results from an incomplete deterministic representation of mechanisms and drivers. This type of uncertainty can be reduced through model improvements that include additional drivers and/or enhance the model`s ability to capture biophysical relationships. In our case, a better characterization of the load sources and replacing the calibration target with HV metrics that are less sensitive to short-term weather resulted in improved model performance (Table 1).

Considerable inter-annual HV variability remained unexplained (Table 1). This is expected because the relatively simple model does not include other drivers like climate-related variables

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

(Scully 2013; Li et al. 2016; Irby et al. 2018; Du et al. 2018). Models of intermediate complexity that combine the strengths of data assimilation with parsimonious ecological process-based representations have been effective in explaining additional variability in similar systems while retaining the ability to characterize uncertainty (Liu and Scavia 2010; Rucinski et al. 2014; Obenour et al. 2015; Del Giudice et al. 2020). However, adding drivers that help explain additional inter-annual variability but are not reliably forecast at seasonal time scales, as is often the case for weather-related variables, may add substantial uncertainty, or make the model less effective in forecast mode. All ecological forecast models will need to eventually strike a balance between the availability of driver forecasts, model performance, and parsimony.

Value of seasonal forecasts - Near-term seasonal forecasts benefit scientists and other stakeholders because they generate knowledge on external controls of ecosystems and permit the translation of that knowledge into a prediction with societal value (Testa et al. 2017a; Dietze et al. 2018). Seasonal forecasts relate causes and consequences of ecological conditions and can help raise public awareness of potential controls. Although the initial motivation for an ecological forecast may be to provide operational, quantitative information to support natural resource management, widely-communicated forecasts also engage audiences outside of the resource management community.

Public engagement can maintain motivation and build support for improving water quality. The release of seasonal hypoxia forecasts in Chesapeake Bay have facilitated that engagement (Scavia and Bertani 2020), along with periodic updates throughout the summer (Maryland DNR 2020), and end-of-year summaries of the yearly severity of hypoxia (VIMS 2020a). Testa et al. (2017a) showed that hypoxia-related media activity increased substantially following initiation of Chesapeake Bay hypoxia forecasts. Articles mentioning forecasts made up 41-56% of all articles related to Chesapeake Bay hypoxia between 2013 and 2015. Similarly, the Gulf of Mexico and Lake Erie annual forecasts each generate hundreds of local and national media reports, resulting in elevated awareness and support for action. Newsletters and websites that supplement the forecasts (e.g., Scavia and Bertani 2020; Rabalais 2020) draw attention to other issues associated with hypoxia, expand discussions around any unexpected factors causing the forecasts to fail, and provide platforms to assess new discoveries while allowing for continuous improvement of the forecast modeling tools.

This article is protected by copyright. All rights reserved

Author Manuscript

Our efforts also highlight how we can gain scientific insights by building and iteratively revisiting ecological forecast models (Dietze et al. 2018). By routinely evaluating our forecasts against observations and investigating the causes leading to model failure in specific years, we gained critical knowledge that guided refinements of HV metrics and relevant load sources. For example, overprediction of average July HV routinely observed in summers with anomalous weather events (Testa et al. 2017a) led to the exploration of HV metrics that would be less sensitive to transient weather conditions and would thus result in improved model performance (this study). This is only the last of a series of iterations that the model has gone through over the years as new data became available, more forecasts were made, and model performance could be re-assessed. For example, a re-evaluation of model performance with a longer forecasting record led to switching to a more parsimonious model formulation where all parameters are kept constant through time rather than allowed to vary over the years (Evans and Scavia 2011). That work also showed how model parameter values gradually changed and model accuracy and precision improved as individual years were progressively added to the calibration dataset. Results of that study indicated that gradual shifts in parameter estimates over time reflected an apparent increased sensitivity of the system to nutrient loads (Evans and Scavia 2011). Those findings led to the adoption of a moving-window calibration approach for a few years (2010-2014), which was abandoned in 2015 to return to a calibration based on the full dataset (Scavia and Bertani 2020) as new forecast performance indicated excessive sensitivity of the calibration window to years with anomalous weather. By continually updating model calibration as new data became available, we also found that the ratio of both summer average and total annual HV to spring TN load has been increasing in recent years (Appendix S1: Fig. S4). This is consistent with previous research that suggested the Bay became more susceptible to hypoxia over the past 35 years (Hagy et al. 2004; Kemp et al. 2005; Murphy et al. 2011). Persistent hypoxia despite N load reductions has been attributed to changes in wind forcing (Scully 2010b), altered spatial patterns of chlorophyll-*a* (Lee et al. 2013; Testa et al. 2018; Wang and Hood 2020), and warming (Du et al. 2018; Ni et al. 2020). These studies point to multiple compounding factors that may be counteracting nutrient reductions and offer hypotheses to test in future applications of our forecast model.

In addition, for cases where the same model is used for both seasonal forecasts and scenarios, the track records of the seasonal forecasts provide useful skill assessments and measures of

This article is protected by copyright. All rights reserved

confidence (e.g., Scavia and Bertani 2020; Scavia et al. 2020a,b; Testa et al. 2017a). Examples where the same model has been used for both seasonal and short-term forecasts and scenario planning include hypoxia in the Gulf of Mexico (Scavia et al. 2017), Chesapeake Bay (Irby and Friedrichs 2019, VIMS 2020b), and the Neuse River Estuary (Katin et al. 2019), and harmful algal blooms in Lake Erie (Scavia et al. 2016; Verhamme et al. 2016; Stumpf et al. 2016; Bertani et al. 2016).

Management scenario application - Unlike other ecological forecasts for the Gulf of Mexico and Lake Erie (GLWQA 2016; Task Force 2016), the original Chesapeake Bay model was not used to guide management decisions, primarily because it was driven only by Susquehanna River loads as opposed to watershed-wide loads. Our analyses demonstrated that driving the model with TN load from all major tributaries and point sources resulted in the best performance for the two metrics that best characterize the system's response to inter-annual variability in loads (Fig. 4). This not only corroborates the importance of watershed-wide load reduction strategies as expressed in the Chesapeake Bay TMDL (US EPA 2010), but also makes the revised model more suitable to evaluate those efforts. The Bay's water quality restoration targets are based on spatio-temporal patterns in DO concentrations rather than Bay-wide HV (US EPA 2010), and the resolution of this model prevents it from evaluating those targets directly. However, the model has been useful in tracking progress over time (Testa et al. 2017a). In addition, because the revised model is better connected to watershed-wide restoration efforts, it can now be used (e.g., Fig. 4) to explore how management actions have influenced hypoxia, how they may influence it in the future, and as an independent line of evidence to support results from the official suite of complex process-based models used by the CBP.

Being based on a steady-state solution, the model cannot predict how long it may take to achieve the mean HV expected under a specific management scenario. It is also important to note that scenario predictions may be conservative because our simple model does not account for future changes in biogeochemical processes such as in sediment oxygen demand. Changes in these processes would not influence seasonal forecasts because their impacts would have been accommodated during model calibration. However, such processes may change through time as a result of sustained load reductions. In the short- to mid-term, the accumulation of estuarine nutrients and organic matter is likely to result in a time lag between load reductions and

Author Manuscript

This article is protected by copyright. All rights reserved

detectable improvements in water quality (Jeppesen et al. 2005; Bocaniov and Scavia 2016); over the long term it is reasonable to expect that substantial and continued load reductions would eventually result in a decrease in oxygen consumption and specifically sediment oxygen demand (Smith and Matisoff 2008; Rucinski et al. 2014). This in turn may lead to additional reductions in HV, although there is substantial uncertainty on how and over what time frames these biogeochemical processes may respond to long-term management actions. Future model enhancements should address this limitation, for example by incorporating parsimonious parameterizations of oxygen consumption processes, similar to what has been done in other systems (Borsuk et al. 2001; Del Giudice et al. 2020; Obenour et al. 2015; Rucinski et al. 2014, 2016).

Another important consideration when using the model in scenario mode is that it was calibrated to a dataset in which inter-annual variability in loads is largely due to variation in precipitation and hydrology. On the other hand, decreases in loads due to management actions are expected to be mainly associated with decreases in constituent concentrations rather than changes in hydrology. Using the model in scenario mode thus assumes that the relationship between loads and HV observed over the calibration period would hold when changes in loads are due to changes in land management rather than changes in hydrology. Although this is a common underlying assumption of similar relatively simple models used both in forecasting and scenario mode (Obenour et al. 2014; Stumpf et al. 2016; Scavia et al. 2017), the inclusion of separate terms in the model for discharge and nutrient inputs would allow one to explore differences in the system's response to changes in loads due to different factors (Stumpf et al. 2012, Del Giudice et al. 2020).

Despite these limitations, some of the characteristics that make this model a useful complement to existing sophisticated three-dimensional hydrodynamic-biogeochemical models of the Chesapeake Bay include a) the ability to seamlessly and readily incorporate new data as they become available and routinely update model calibration in line with an adaptive management approach, b) the fast computation time, which makes it possible to easily evaluate large numbers of management scenarios, and c) the ability to rigorously characterize uncertainty and provide probabilistic predictions. Separating different sources of uncertainty is important because the target of management actions is typically the true, latent state of an ecosystem property (e.g., the

Author Manuscript

This article is protected by copyright. All rights reserved

true, unknown HV represented by y_i in Eq. 6), which is not affected by measurement error. The portion of the overall model predictive uncertainty that is due to HV measurement error can thus be removed when using the model to answer management questions, thereby leading to narrower prediction intervals (solid gray lines in Fig. 4). In addition to that, different error intervals are relevant to different management questions and uncertainty is generally lower when predicting a long-term average response compared to predictions for individual years (Fig. 4). In our case, when using the model to predict the expected long-term mean HV associated with a given management scenario, stochasticity associated with individual year variability (i.e., model prediction error) is not relevant because it does not influence the expected long-term mean response (Scavia et al. 2020c). However, this source of error should be considered when using the model in forecast mode to accommodate the additional uncertainty arising from forecasting HV in a specific year.

Forecasting best practices – There is increasing consensus among scientists as to what represent best practices that should be followed when producing, evaluating, and communicating ecological forecasts (Dietze et al. 2018; Harris et al. 2018; White et al. 2019; Carey et al. 2021). Some of those practices have been at the core of this work and we discussed their importance extensively in previous sections, including explicitly accounting for and propagating multiple sources of uncertainty, such as observation and process uncertainty, identifying better predictor variables that are expected to relate to the forecast endpoint, using the model to make both short- and long-term predictions to accommodate the time scales of management decisions while also using short-term forecasts to facilitate evaluation of model performance, and routinely assessing and updating the model with new data (Dietze et al. 2018; Harris et al. 2018; White et al. 2019). Our work also demonstrates the importance of several other proposed best practices. For example, the decrease in the best model`s predictive performance when run in blind forecast mode (NSE = 0.47) compared to full calibration mode (NSE = 0.52) confirms the importance of evaluating models through out-of-sample validation approaches, such as hindcasting, to avoid over-optimistic conclusions on forecasting performance (Dietze et al. 2018; Harris et al. 2018; White et al. 2019). We also showed that our model represents an improvement over a baseline model that assumes no changes over time and essentially predicts constant HV (Dietze et al. 2018; Harris et al. 2018; White et al. 2019). Finally, loads and DO measurements used to produce our forecasts are made publicly available within 2 and 6-10 months of collection,

Author Manuscript

respectively (Soroka and Blomquist 2020, Chesapeake Bay Program 2020), and past forecasts are archived publicly (Scavia et al. 2019) for retrospective assessment of performance (Dietze et al. 2018; Harris et al. 2018; White et al. 2019).

CONCLUSIONS

We presented an updated and revised version of a long-standing estuarine hypoxia forecasting model. Our revisions focused on some of the most critical challenges and opportunities faced by contemporary ecological forecasting models (Dietze et al. 2018), including a) the adoption of metrics of ecosystem state and anthropogenic pressure that strike an optimal balance between predictability and relevance for management purposes, b) the ability to incorporate multiple data sources within a (Bayesian hierarchical) framework that allows to rigorously separate and propagate different sources of uncertainty, and c) the ability to use the model in scenario mode to probabilistically evaluate the effect of alternative management decisions on future ecosystem state. The model`s relative simplicity facilitates an iterative process of model application, evaluation, and enhancement through regular incorporation of updated information and is part of what makes this tool a useful complement to more sophisticated process-based models. Finally, the basic formulation and minimal data needs (DO and TN are among the parameters routinely assessed in water quality monitoring programs) make forecast operations straightforward and transparent and the model itself readily adaptable to other estuarine systems facing similar anthropogenic pressures.

ACKNOWLEDGMENTS

The authors would like to thank Jessica Rigelman for assisting with and providing data on point source loads and management scenarios estimated by the Chesapeake Bay Program`s watershed model CAST. Richard Tian, Gopal Bhatt, Dave Montali, and Yu-Chen Wang provided numerous helpful discussions and suggestions during model development, analyses, and manuscript preparation. Scavia's contributions were supported in part by USEPA contract EP-C-17-046. Bertani and Murphy's contributions were supported by the USEPA (CBP Technical Support Grant No. 07-5-230480). Friedrichs' contribution to this research was funded by NOAA's National Center for Coastal Ocean Science under award NA16NOS4780207. Testa's contribution to this research was funded by the National Science Foundation (CBET-1360395)

Author Manuscript

This article is protected by copyright. All rights reserved

and the National Oceanographic and Atmospheric Administration (NA15NOS4780184). This is UMCES Contribution #XXXX and Ref. No. [UMCES] CBL 2XXX-XXX.

Author Manuscript

This article is protected by copyright. All rights reserved

REFERENCES

Ator, S.W., J.D. Blomquist, J.S. Webber, and J.G. Chanat. 2020. Factors driving nutrient trends in streams of the Chesapeake Bay watershed. *Journal of Environmental Quality* 49(4): 812-834.

Beckage, B., L.J. Gross, and S. Kauffman. 2011. The limits to prediction in ecological systems. *Ecosphere* 2(11): 125.

Bertani, I., D.R. Obenour, C.E. Steger, C.A. Stow, A.D. Gronewold, and D. Scavia 2016. Probabilistically assessing the role of nutrient loading in harmful algal bloom formation in western Lake Erie. *Journal of Great Lakes Research* 42: 1184-1192.

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

Bever, A.J., M.A.M. Friedrichs, C.T. Friedrichs, and M.E. Scully. 2018. Estimating hypoxic volume in the Chesapeake Bay using two continuously sampled oxygen profiles. *Journal of Geophysical Research: Oceans* 123: 6392-6407.

Bever, A.J., M.A.M. Friedrichs, and P. St-Laurent. 2021. Real-time environmental forecasts of the Chesapeake Bay: Model setup, improvements, and online visualization. *Environmental Modelling and Software* 140: 105036.

Bocaniov, S., and D. Scavia. 2016. Temporal and spatial dynamics of large lake hypoxia: Integrating statistical and three-dimensional dynamic models to enhance lake management criteria. *Water Resources Research* 52: 4247-4263.

Boesch, D.F. 2006. Scientific requirements for ecosystem-based management in the restoration of Chesapeake Bay and Coastal Louisiana. *Ecological Engineering* 26(1): 6–26.

Borsuk, M.E., D. Higdon, C. Stow, and K. Reckhow K. 2001. A Bayesian hierarchical model to predict benthic oxygen demand from organic matter loading in estuaries and coastal zones. *Ecological Modelling* 143: 165–181.

Brady, D.C., T.E. Targett, and D.M. Tuzzolino. 2009. Behavioral responses of juvenile weakfish (*Cynoscion regalis*) to diel-cycling hypoxia: swimming speed, angular correlation, expected

Author Manuscript

This article is protected by copyright. All rights reserved

displacement, and effects of hypoxia acclimation. *Canadian Journal of Fisheries and Aquatic Sciences* 66: 415-424.

Buchheister, A., C.F. Bonzek, J. Gartland, and R.J. Latour. 2013. Patterns and drivers of the demersal fish community of Chesapeake Bay. *Marine Ecology Progress Series* 481: 161–180.

Carey, C.C., W.M. Woelmer, M.E. Lofton, R.J. Figueiredo, B.J. Bookout, R.S. Corrigan, V. Daneshmand, A.G. Hounshell, D.W. Howard, A.S.L. Lewis, R.P. McClure, H.L. Wander, N.K. Ward, and R.Q. Thomas. 2021. Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. *Inland Waters* 1-14.

Carpenter, S.R. 2002. Ecological futures: building an ecology of the long now. *Ecology* 83: 2069-2083.

Chapra, S.C. 1997. Surface Water-Quality Modeling. McGraw-Hill, New York.

Chesapeake Bay Program. 2017. Chesapeake Assessment and Scenario Tool (CAST) Version 2017d. Chesapeake Bay Program Office. Accessed May 2020. https://cast.chesapeakebay.net/.

Chesapeake Bay Program. 2020. Chesapeake Bay Program Data Hub. Accessed April 2020. http://www.chesapeakebay.net/data.

Clark, J.S. 2005. Why environmental scientists are becoming Bayesians. *Ecology Letters* 8: 2-14.

Clark, J.S., S.R. Carpenter, M. Barber, S. Collins, A. Dobson, J.A. Foley, D.M. Lodge, M. Pascual, R.Jr. Pielke, W. Pizer, C. Pringle, W.V. Reid, K.A. Rose, O. Sala, W.H. Schlesinger, D.H. Wall, and D. Wear. 2001. Ecological forecasts: an emerging imperative. *Science* 293: 657–60.

Coreau, A., G. Pinay, J.D. Thompson, P.-O. Cheptou, and L. Mermet. 2009. The rise of research on futures in ecology: rebalancing scenarios and predictions. *Ecology Letters* 12: 1277–1286.

Cressie, N., C.A. Calder, J.S. Clark, J.M.V. Hoef, and C.K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* 19(3): 553-570.

Da, F., M.A.M. Friedrichs, and P. St-Laurent. 2018. Impacts of atmospheric nitrogen deposition and coastal nitrogen fluxes on oxygen concentrations in Chesapeake Bay. *Journal of Geophysical Research: Oceans* 123: 5004-5025.

Author Manuscript

This article is protected by copyright. All rights reserved

Del Giudice, D., V.R.R. Matli, and D.R. Obenour. 2020. Bayesian mechanistic modeling characterizes Gulf of Mexico hypoxia: 1968–2016 and future scenarios. *Ecological Applications* 30 (2): e02032.

Dietze, M.C., A. Fox, L.M. Beck-Johnson, J.L. Betancourt, M.B. Hooten, C.S. Jarnevich, T.H. Keitt, M.A. Kenney, C.M. Laney, L.G. Larsen, H.W. Loescher, C.K Lunch, B.C. Pijanowski, J.T. Randerson, E.K. Read, A.T. Tredennick, R. Vargas, K.C. Weathers, and E.P. White. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences* 115(7): 1424-1432.

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

EFI. 2020. Ecological Forecasting Initiative. Forecasts to understand, manage, and conserve ecosystems. Webpage. Accessed November 2020. https://ecoforecast.org.

Eshleman, K.N., R.D. Sabo, and K.M. Kline. 2013. Surface Water Quality Is Improving due to Declining Atmospheric N Deposition. *Environmental Science and Technology* 47(21): 12193–12200.

Evans, M.R., M. Bithell, S.J. Cornell, S.R.X. Dall, S. Dıaz, S. Emmott, B. Ernande, V. Grimm, D.J. Hodgson, S.L. Lewis, G.M. Mace, M. Morecroft, A. Moustakas, E. Murphy, T. Newbold, K.J. Norris, O. Petchey, M. Smith, J.M.J. Travis, and T.G. Benton. 2013 Predictive systems ecology. *Proceedings of the Royal Society B* 280: 20131452.

Evans, M.A., and D. Scavia 2011. Forecasting hypoxia in the Chesapeake Bay and Gulf of Mexico: Model accuracy, precision, and sensitivity to ecosystem change. *Environmental Research Letters* 6: 015001.

Fang, S., D. Del Giudice, D. Scavia, C.E. Binding, T.B. Bridgeman, J.D. Chaffin, M.A. Evans, J. Guinness, T.H. Johengen, and D.R. Obenour. 2019. A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent. *Science of the Total Environment* 695: 133776.

Author Manuscript

This article is protected by copyright. All rights reserved

Feng, Y., S.F. DiMarco, and G.A. Jackson. 2012. Relative role of wind forcing and riverine nutrient input on the extent of hypoxia in the northern Gulf of Mexico. *Geophysical Research Letters* 39: L09601, doi:10.1029/2012GL051192.

Fennel, K., A. Laurent, R. Hetland, D. Justic´, D.S. Ko, J. Lehrter, M. Murrell, L. Wang, L. Yu, and W. Zhang. 2016. Effects of model physics on hypoxia simulations for the northern Gulf of Mexico: A model intercomparison. *Journal of Geophysical Research: Oceans* 121: 5731–5750.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York.

Gimenez, O., S.T. Buckland, B.J.T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M-P. Etienne, R. Fewster, F. Gosselin, B. Merigot, P. Monestiez, J.M. Morales, F. Mortier, F. Munoz, O. Ovaskainen, S. Pavoine, R. Pradel, F.M. Schurr, L. Thomas, W. Thuiller, V. Trenkel, P. de Valpine, and E. Rexstad. 2014. Statistical ecology comes of age. *Biology Letters* 10(12): 20140698.

GLWQA. 2016. Great Lakes Water Quality Agreement. The United States and Canada adopt phosphorus load reduction targets to combat Lake Erie algal blooms. https://binational.net/2016/02/22/ finalptargets-ciblesfinalesdep/

Gneiting, T., and M. Katzfuss. 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application* 1: 125-151.

Goodrich, D.M., W.C. Boicourt, P. Hamilton, and D.W. Pritchard. 1987. Wind-induced destratification in Chesapeake Bay. *Journal of Physical Oceanography* 17(12): 2232-2240.

Gurbisz, C., and W.M. Kemp. 2014. Unexpected resurgence of a large submersed plant bed in Chesapeake Bay: analysis of time series data. *Limnology and Oceanography* 59(2): 482–494.

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

Harris, D.J., S.D. Taylor, and E.P. White. 2018. Forecasting biodiversity in breeding birds using best practices. *PeerJ* 6: e4278.

Author Manuscript

This article is protected by copyright. All rights reserved

Harwood, J., and K. Stokes. 2003. Coping with uncertainty in ecological advice: lessons from fisheries. *Trends in Ecology and Evolution* 18(12): 617-622.

Hirsch, R.M., D.L. Moyer, and S.A. Archfield. 2010. Weighted regression on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs. *Journal of the American Water Resources Association* 46: 857–880.

Hofman, J.M., D.G. Goldstein, and J. Hullman. 2020. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Irby, I.D., and M.A.M. Friedrichs. 2019. Evaluating confidence in the impact of regulatory nutrient reduction on Chesapeake Bay water quality. *Estuaries and Coasts* 42: 16-32.

Irby, I.D., M.A.M. Friedrichs, F. Da, and K.E. Hinson. 2018. The Competing Impacts of Climate Change and Nutrient Reductions on Dissolved Oxygen in Chesapeake Bay. *Biogeosciences* 15: 2649–2668.

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

Jeppesen, E., M. Søndergaard, J.P. Jensen, K.E. Havens, O. Anneville, L. Carvalho, M.F. Coveney, R. Deneke, M.T. Dokulil, B. Foy, D. Gerdeaux, S.E. Hampton, S. Hilt, K. Kangur, J. Kohler, E.H.H.R. Lammens, T.L. Lauridsen, M. Manca, M.R. Miracle, B. Moss, P. Noges, G. Persson, G. Phillips, R. Portielje, S. Romo, C.L. Schelske, D. Straile, I. Tatrai, E. Willen, and M. Winder. 2005. Lake responses to reduced nutrient loading - An analysis of contemporary long-term data from 35 case studies. *Freshwater Biology* 50: 1747–1771.

Johnson-Bice, S.M., J.M. Ferguson, J.D. Erb, T.D. Gable, and S.K. Windels. 2020. Ecological forecasts reveal limitations of common model selection methods: predicting changes in beaver colony densities. *Ecological Applications*: e02198.

Jordan, A., F. Krüger, and S. Lerch. 2019. Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software* 90(12): 1–37.

Author Manuscript

This article is protected by copyright. All rights reserved

Katin, A., D. Del Giudice, and D.R. Obenour. 2019. Modeling biophysical controls on hypoxia in a shallow estuary using a Bayesian mechanistic approach. *Environmental Modelling & Software* 120: 104491.

Kemp, W.M., W.R. Boynton, J.E. Adolf, D.F. Boesch, W.C. Boicourt, G. Brush, J.C. Cornwell, T.R. Fisher, P.M. Glibert, J.D. Hagy, L.W. Harding, E.D. Houde, D.G. Kimmel, W.D. Miller, R.I.E. Newell, M.R. Roman, E.M. Smith, J.C. Stevenson. 2005. Eutrophication of Chesapeake Bay: historical trends and ecological interactions. *Marine Ecology Progress Series* 303:1-29.

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

Lefcheck, J.S., R.J. Orth, W.C. Dennison, D.J. Wilcox, R.R. Murphy, J. Keisman, C. Gurbisz, M. Hannam, J.B. Landry, K.A. Moore, C.J. Patrick, J. Testa, D.E. Weller, and R.A. Batiuk. 2018. Long-term nutrient reductions lead to the unprecedented recovery of a temperate coastal region. *Proceedings of the National Academy of Sciences* 115(14): 3658-3662.

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

Linker, L.C., R.A. Batiuk, G.W. Shenk, and C.F. Cerco. 2013. Development of the Chesapeake Bay watershed total maximum daily load allocation. *Journal of the American Water Resources Association* 49(5): 986–1006.

Liu, Y., G.B. Arhonditsis, C.A. Stow, and D. Scavia. 2011. Predicting the hypoxic-volume in Chesapeake Bay with the Streeter Phelps model: a Bayesian approach. *Journal of the American Water Resources Association* 1(6): 1348–1363.

Liu, Y., and D. Scavia. 2010. Analysis of the Chesapeake Bay Hypoxia Regime Shift: Insights from Two Simple Mechanistic Models. *Estuaries and Coasts* 33: 629-639.

Lunn, D.J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10: 325–337.

Luo, Y.Q., K. Ogle, C. Tucker, S.F. Fei, C. Gao, S. LaDeau, J.S. Clark, and D.S. Schimel. 2011. Ecological forecasting and data assimilation in a data-rich era. *Ecological Applications* 21(5): 1429–1442.

Author Manuscript

This article is protected by copyright. All rights reserved

Maryland DNR. 2020. Chesapeake Bay Hypoxia Reports. Webpage. Accessed November 2020. https://dnr.maryland.gov/waters/bay/Pages/Hypoxia-Reports.aspx.

Matheson, J.E., and R.L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Science* 22(10): 1087-1096.

Matli, V.R.R, S. Fang, J. Guinness, N.N. Rabalais, J.K. Craig, and D.R. Obenour. 2018. Space-Time Geostatistical Assessment of Hypoxia in the Northern Gulf of Mexico *Environmental Science and Technology* 52: 12484−12493.

Matli, V.R.R., A. Laurent, K. Fennel, K. Craig, J. Krause, and D.R. Obenour. 2020. Fusion-Based Hypoxia Estimates: Combining Geostatistical and Mechanistic Models of Dissolved Oxygen Variability. *Environmental Science and Technology* 54: 13016−13025.

Mistiaen, J.A., I.E. Strand, and D. Lipton. 2003. Effects of environmental stress on blue crab (*Callinectes sapidus*) harvests in Chesapeake Bay tributaries. *Estuaries* 26(2): 316–322.

Modig, H., and E. Ólafsson. 1998. Responses of Baltic benthic invertebrates to hypoxic events. *Journal of Experimental Marine Biology and Ecology* 229: 133-148.

Moriarty, J.M., M.A.M. Friedrichs, and C.K. Harris. 2020. Seabed resuspension in the Chesapeake Bay: Implications for biogeochemical cycling and hypoxia. *Estuaries and Coasts*.

Mouquet, N., Y. Lagadeuc, V. Devictor, L. Doyen, A. Duputie, D. Eveillard, D. Faure, E. Garnier, O. Gimenez, P. Huneman, F. Jabot, P. Jarne, D. Joly, R. Julliard, S. Kefi, G. J. Kergoat, S. Lavorel, L. Le Gall, L. Meslin, S. Morand, X. Morin, H. Morlon, G. Pinay, R. Pradel, F. M. Schurr, W. Thuiller, and M. Loreau. 2015. Predictive ecology in a changing world. *Journal of Applied Ecology* 52: 1293-1310.

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

NASA. 2020. Ecological Forecasting. Strengthening Ecosystems. Webpage. Accessed November 2020. https://appliedsciences.nasa.gov/what-we-do/ecological-forecasting.

Ni, W., M. Li, A.C. Ross, and R.G. Najjar. 2020. Large Projected decline in dissolved oxygen in a eutrophic estuary due to climate change. *Journal of Geophysical Research: Oceans* 124: 8271–8289.

Author Manuscript

This article is protected by copyright. All rights reserved

NOAA. 2020. NOAA Ecological Forecasting. Predicting human health and coastal economies with early warnings. Webpage. Accessed November 2020. https://oceanservice.noaa.gov/ecoforecasting/

NOAA GLERL. 2020. Experimental Lake Erie Hypoxia Forecast. Webpage. Accessed November 2020. https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/hypoxiaWarningSystem.html.

North Carolina Sea Grant. 2020. Midsummer Neuse River Forecast Shows Greater Potential for Fish Kills. Webpage. Accessed November 2020. https://ncseagrant.ncsu.edu/currents/2020/06/midsummer-neuse-river-forecast-shows-greater-potential-for-fish-kills/.

Obenour, D.R., A.D. Gronewold, C.A. Stow, and D. Scavia. 2014. Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. *Water Resources Research* 50: 7847–7860.

Obenour D.R., A.M. Michalak, and D. Scavia. 2015. Assessing biophysical controls on Gulf of Mexico hypoxia through probabilistic modeling. *Ecological Applications* 25: 492–505.

Obenour, D.R., D. Scavia, N.N. Rabalais, R.E. Turner, and A.M. Michalak. 2013. Retrospective analysis of midsummer hypoxic area and volume in the northern Gulf of Mexico, 1985−2011. *Environmental Science and Technology* 47(17): 9808−9815.

Pappenberger, F., and K. J. Beven. 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resources Research* 42: W05302.

Pappenberger, F., M.-H. Ramos, H.L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon. 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology* 522: 697-713.

Payne, M.R., A.J. Hobday, B.R. MacKenzie, D. Tommasi, D.P. Dempsey, S. Fassler, A.C. Haynie, R. Ji, G. Liu, P.D. Lynch, D. Matei, A.K. Miesner, K.E. Mills, K.O. Strand, and E. Villarino. 2017. Lessons from the first generation of marine ecological forecast products. *Frontiers in Marine Science* 4: 289.

Petchey, O.L., M. Pontarp, T.M. Massie, S. Kéfi, A. Ozgul, M. Weilenmann, G.M. Palamara, F. Altermatt, B. Matthews, J.M. Levine, D.Z. Childs, B.J. McGill, M.E. Schaepman, B. Schmid, P.

Author Manuscript

This article is protected by copyright. All rights reserved

Spaak, A.P. Beckerman, F. Pennekamp, and I.S. Pearse. 2015. The ecological forecast horizon, and examples of its uses and determinants. *Ecology Letters* 18(7): 597-611.

Thomas, R.Q., R.J. Figueiredo, V. Daneshmand, B.J. Bookout, L.K. Puckett, and C.C. Carey. 2020. A near-term iterative forecasting system successfully predicts reservoir hydrodynamics and partitions uncertainty in real time. *Water Resources Research* 56(11): e2019WR026138.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Rabalais, N.N. 2020. Gulf of Mexico Hypoxia. https://gulfhypoxia.net/

Raftery, A.E. 2016. Use and communication of probabilistic forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9(6): 397-410.

Ross, A.C., C.A. Stock, K.W. Dixon, M.A.M. Friedrichs, R.R. Hood, M. Li, K. Pegion, V. Saba, and G.A. Vecchi. 2020. Estuarine forecasts at daily weather to subseasonal time scales. *Earth and Space Science* 7: e2020EA001179.

Rucinski, D., D. Scavia**,** J. DePinto, and D. Beletsky. 2014. Lake Erie's hypoxia response to nutrient loads and meteorological variability. *Journal of Great Lakes Research* 40(3): 151-161.

Rucinski, D., J. DePinto, D. Beletsky, and D. Scavia. 2016. Modeling hypoxia in the Central Basin of Lake Erie under potential phosphorus load reduction scenarios. *Journal of Great Lakes Research* 42: 1206-1211.

Scavia, D., N.N. Rabalais, R.E. Turner, D. Justic, and W. Wiseman Jr. 2003. Predicting the response of Gulf of Mexico Hypoxia to variations in Mississippi River Nitrogen Load. *Limnology and Oceanography* 48(3): 951-956.

Scavia, D., M.A. Evans, and D.R. Obenour. 2013. A scenario and forecast model for Gulf of Mexico hypoxic area and volume. *Environmental Science and Technology* 47:10423–10428.

Scavia, D., and I. Bertani. 2020. Chesapeake Bay Hypoxic Volume Forecasts. June 7, 2020. Available at: http://scavia.seas.umich.edu/wp-content/uploads/2020/10/2020-Chesapeake-Bay-forecast_EndOfSummer.pdf

Author Manuscript

This article is protected by copyright. All rights reserved

Scavia, D., Y-C.,Wang, and D.R. Obenour. 2020a. Lake Erie Harmful Algal Bloom Forecast. June 7, 2020. Available at: http://scavia.seas.umich.edu/wp-content/uploads/2020/07/2020-LakeErieBloomForecastRelease.pdf

Scavia, D., I. Bertani, C. Long, D.R. Obenour, and Y-C. Wang. 2020b. Gulf of Mexico Hypoxia Forecast. June 7, 2020. Available at: http://scavia.seas.umich.edu/wp-content/uploads/2020/08/2020-Gulf-of-Mexico-Hypoxic-Forecast.pdf

Scavia, D., I. Bertani, C. Long, and Y. Wang. 2019. Chesapeake Bay Hypoxic Volume Forecasts. June 7, 2019. Available at: http://scavia.seas.umich.edu/wp-content/uploads/2019/06/2019-Chesapeake-Bay-forecast.pdf

Scavia, D., I. Bertani, D.R. Obenour, R.E. Turner, D.R. Forrest, and A. Katin. 2017. Ensemble modeling informs hypoxia management in the northern Gulf of Mexico. *Proceedings of the National Academy of Sciences* 114: 8823-8828.

Scavia, D., J.V. DePinto, and I. Bertani. 2016. A Multi-model approach to evaluating target phosphorus loads for Lake Erie. *Journal of Great Lakes Research* 42: 1139-1150.

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

Scavia, D., Y-C. Wang, D.R. Obenour, A. Apostel, S.J. Basile, M.M. Kalcic, C.J. Kirchhoff, L. Miralha, R.L. Muenich, and A.L. Steiner. 2020c. Quantifying uncertainty cascading from climate, watershed, and lake models in harmful algal bloom predictions. *Science of the Total Environment*: 143487.

Schindler, D.E., and R. Hilborn. 2015. Prediction, precaution, and policy under global change. *Science* 347(6225): 953-954.

Scully, M.E. 2010a. Wind Modulation of Dissolved Oxygen in Chesapeake Bay. *Estuaries and Coasts* 33: 1164–1175.

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

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

Author Manuscript

This article is protected by copyright. All rights reserved

Shenk, G.W., and L.C. Linker. 2013. Development and application of the 2010 Chesapeake Bay Watershed total maximum daily load model. *Journal of the American Water Resources Association* 49 (5): 1042–1056.

Salon, S., G. Cossarini, G. Bolzon, L. Feudale, P. Lazzari, A. Teruzzi, C. Solidoro, A. Crise. 2019. Novel metrics based on Biogeochemical Argo data to improve the model uncertainty evaluation of the CMEMS Mediterranean marine ecosystem forecasts. Ocean Sci., 15, 997–1022, 2019 https://doi.org/10.5194/os-15-997-2019

Smith, D.A., and G. Matisoff. 2008. Sediment oxygen demand in the central basin of Lake Erie. *Journal of Great Lakes Research* 34(4): 731–744.

Soroka, A.M., and D.J. Blomquist. 2020. Nitrogen flux estimates in support of Chesapeake Bay Hypoxia and Anoxia forecasts, 1985-2020: U.S. Geological Survey data release, https://doi.org/10.5066/P9QU1DWS.

Stow, C.A., and D. Scavia. 2009. Modeling hypoxia in the Chesapeake Bay: ensemble estimation using a Bayesian hierarchical model. *Journal of Marine Systems* 76(1-2): 244-250.

Streeter, H.W., and E.B. Phelps. 1925. A Study in the Pollution and Natural Purification of the Ohio River, III Factors Concerning the Phenomena of Oxidation and Reaeration. US Public Health Service, Public Health Bulletin No. 146, Feb 1925 Reprinted by US PHEW, PHA 1958.

Sturdivant, S.K., M.J. Brush, and R.J. Diaz. 2013. Modeling the Effect of Hypoxia on Macrobenthos Production in the Lower Rappahannock River, Chesapeake Bay, USA. *Plos One* 8: e84140.

Sturdivant, S.K., R.J. Díaz, R.Llansó, and D.M. Dauer. 2014. Relationship between Hypoxia and Macrobenthic Production in Chesapeake Bay. *Estuaries and Coasts* 37 (5): 1219-1232.

Stumpf R.P., T.T. Wynne, D.B. Baker, G.L. Fahnenstiel. 2012. Interannual Variability of Cyanobacterial Blooms in Lake Erie. PLoS ONE 7(8): e42444. doi:10.1371/journal.pone.0042444

Stumpf, R.P., L.T. Johnson, T.T Wynne, and D.B. Baker. 2016. Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie. *Journal of Great Lakes Research* 42(6): 1174–1183.

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

Sturtz, S., U. Ligges, and A.E. Gelman. 2005. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software* 12 (3): 1–16.

Task Force. 2016. Mississippi River/Gulf of Mexico Watershed Nutrient Task Force. Looking forward: The strategy of the federal members of the Hypoxia Task Force (Mississippi River/Gulf of Mexico Watershed Nutrient Task Force, Washington, DC). Available at https://www.epa.gov/sites/production/files/2016-12/documents/federal_strategy_updates_12.2.16.pdf.

Testa, J.M., Y. Li, Y.J. Lee, M. Li, D.C. Brady, D.M.D. Toro, and W.M. Kemp. 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.

Testa, J.M., J.B Clark, W.C. Dennison, E.C. Donovan, A.W. Fisher, W. Ni, M. Parker, D. Scavia, S.E. Spitzer, A.M. Waldrop, V.M.D. Vargas, and G. Ziegler. 2017a. Ecological forecasting and the science of hypoxia in Chesapeake Bay. *Bioscience* 67 (7): 614–626.

Testa, J.M., Y. Li, Y.J. Lee, M. Li, D.C. Brady, D.M.D. Toro, and W.M. Kemp. 2017b. Modeling physical and biogeochemical controls on dissolved oxygen in Chesapeake Bay: Lessons learned from simple and complex approaches. In Modeling Coastal Hypoxia - Numerical Simulations of Patterns, Controls and Effects of Dissolved Oxygen Dynamics, ed. D. Justic, K. Rose, R. Hetland and K. Fennel. Cham, Switzerland: Springer.

Testa, J.M., R.R. Murphy, D.C. Brady, and W.M. Kemp. 2018. Nutrient- and Climate-Induced Shifts in the Phenology of Linked Biogeochemical Cycles in a Temperate Estuary. *Frontiers in Marine Science* 5: 114.

Turner, R.E., N.N. Rabalais, and D. Justić. 2012. Predicting summer hypoxia in the northern Gulf of Mexico: redux. *Marine Pollution Bulletin* 64: 319-324.

US EPA. 2003. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and its Tidal Tributaries Rep., 343 pp, U.S. Environmental Protection Agency Region III, Chesapeake Bay Program Office, Annapolis, MD.

This article is protected by copyright. All rights reserved

US EPA. 2010. Chesapeake Bay total maximum daily load for nitrogen, phosphorus and sediment. Available at: https://www.epa.gov/chesapeake-bay-tmdl/chesapeake-bay-tmdl-document

Valette-Silver, N. and D. Scavia. 2003. Ecological forecasting: New tools for coastal and marine ecosystem management. NOAA Technical Memorandum NOS NCCOS 1, 116 pp. http://scavia.seas.umich.edu/wp-content/uploads/2009/11/noaa_ecological_forecasting_book1.pdf

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

Verhamme, E., T. Redder, D. Schlea, J. Grush, J. Bratton, and J. DePinto. 2016. Development of the Western Lake Erie Ecosystem Model (WLEEM): application to connect phosphorus loads to cyanobacteria biomass. *Journal of Great Lakes Research* 42(6): 1193–1205.

VIMS. 2020a. Chesapeake Bay Dead-Zone Report Card. November 2020. Available at: https://www.vims.edu/research/topics/dead_zones/forecasts/report_card/index.php.

VIMS. 2020b. Chesapeake Bay Hypoxia Forecast. Webpage. Accessed November 2020. https://www.vims.edu/research/topics/dead_zones/forecasts/cbay/index.php.

Wang, J., and R.R. Hood. 2020. Modeling the origin of the particulate organic matter flux to the hypoxic zone of Chesapeake Bay in early summer. *Estuaries and Coasts* doi:10.1007/s12237-020-00806-0.

White, E.P., G.M. Yenni, S.D. Taylor, E.M. Christensen, E.K. Bledsoe, J.L. Simonis, and S.M. Ernest. 2019. Developing an automated iterative near-term forecasting system for an ecological study. *Methods in Ecology and Evolution* 10(3): 332-344.

WIP 2020. Chesapeake Bay Watershed Implementation Plans. Chesapeake Bay Program. https://www.chesapeakebay.net/what/programs/watershed_implementation. Accessed May 20 2020.

Zhang, H., L. Boegman, D. Scavia, and D.A. Culver. 2016. Spatial distributions of external and internal phosphorus loads in Lake Erie and their impacts on phytoplankton and water quality. *Journal of Great Lakes Research* 42: 1212-1227.

Author Manuscript

This article is protected by copyright. All rights reserved

Zhang, Q., R.R. Murphy, R. Tian, M.K. Forsyth, E.M. Trentacoste, J. Keisman, and P.J. Tango. 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.

Zhou, Y., D.R. Obenour, D. Scavia, T.H. Johengen, and A.M. Michalak. 2013. Spatial and temporal trends in Lake Erie hypoxia, 187−2007. *Environmental Science and Technology* 47(2): 899−905.

Zhou, Y., D. Scavia, and A.M. Michalak. 2014. Nutrient loading and meteorological conditions explain interannual variability of hypoxia in the Chesapeake Bay. *Limnology and Oceanography* 59: 373-374.

Author Manuscript

Tables

Table 1 - Best performing model for each HV metric. NSE = Nash-Sutcliffe Efficiency, r^2 = square of the correlation coefficient between observed and predicted values, RMSE = root mean square error, MAE = mean absolute error, RSTDE = residual standard error, Coverage = percentage of the observations used in calibration that fall within the 95% prediction intervals, CRPS = Continuous Ranked Probability Score, CRPS score = CRPS skill score (see text for definition), Sus = Susquehanna, Pot = Potomac, PS = point sources. Results for September HV not shown because no model resulted in NSE > 0. Three Average July models have the same NSE. For comparison, performance of the previous model version (driven by Jan-May Susquehanna River loads and predicting Average July HV) is also reported, together with performance of the two best models predicting Average July and Total Annual HV with Susquehanna loads only.

This article is protected by copyright. All rights reserved

HV metric	Load Sources	Load Period	NSE	r^2	RMSE	MAE	RSTDE	Coverage	CRPS	CRPS score
June	All tributaries	Mar-Jun	0.25	0.30	1.75	1.45	1.81	100%	1.02	0.12
July	Sus + Pot + PS	Oct-May	0.29	0.30	2.38	1.82	2.46	94%	1.35	0.20
July	Sus + Pot + PS	Nov-Jun	0.29	0.29	2.39	1.82	2.47	97%	1.35	0.19
July	All tributaries + PS	Nov-May	0.29	0.29	2.39	1.78	2.52	94%	1.36	0.19
August	All tributaries + PS	Jan-Jun	0.22	0.24	1.63	1.30	1.69	97%	0.93	0.20
Summer	All tributaries + PS	Jan-Jun	0.40	0.43	1.01	0.81	1.04	94%	0.57	0.26
Annual	All tributaries + PS	Jan-Jun	0.52	0.52	123	96	130	94%	68.12	0.36
July	Sus	Jan-May	0.14	0.18	2.62	2.08	2.68	97%	1.49	0.10
July	Sus	Dec-Jun	0.23	0.24	2.49	1.98	2.60	97%	1.42	0.14

Author Manuscript

This article is protected by copyright. All rights reserved

Annual	Sus	Jan-May	0.28	0.37	150	113	156	97%	82.17	0.22

Table 2 - Total annual and summer average HVs (mean and 95% credible intervals) predicted under different total nitrogen (TN) load scenarios. For details on each scenario see text.

Scenario	**Jan-Jun TN Load (kg/day)**	**Total Annual HV (95% CI) (km^3*days)**	**Summer Average HV (95% CI) (km^3)**
1985 FN	486713	930 (840-1005)	7.2 (6.5-7.8)
2018 FN	350360	770 (640-870)	5.9 (4.9-6.5)
2020 No Action	564932	995 (910-1085)	7.8 (7.2-8.4)
WIP3 Actual	285570	660 (480-785)	4.9 (3.8-5.9)
WIP3 Planning Targets	274250	635 (440-785)	4.7 (3.4-5.6)

FIGURE CAPTIONS

Fig. 1 - Annual total nitrogen (TN) loads from nine tributaries (Sus: Susquehanna; Rap: Rappahannock; Pot: Potomac; Pat: Patuxent; Pam: Pamunkey; Mat: Mattaponi; App: Appomattox; Jam: James; Cho: Choptank) and point sources downstream from the tributary monitoring stations (PS). Point source data for Jul-Sep 2018 are partial. Water year: Oct-Sep.

Fig. 2 - Average July (a) and total annual (b) hypoxic volumes (HVs) estimated using three different interpolation methods over 1985-2018. Zhou estimates are available only through 2010. Shaded areas mark years when weather events disrupted hypoxia shortly before the July cruises.

Fig. 3 - Observed vs. predicted total annual (a) and summer average (b) HV for the model calibrated to three sets of HV estimates simultaneously. The gray bars represent 95% predictive intervals accounting for model prediction error, HV measurement error, and parameter uncertainty. The 1:1 line is shown in black for reference.

Author Manuscript

This article is protected by copyright. All rights reserved

Fig. 4 - Response curves for total annual (a) and summer average (b) HV vs. average Jan-Jun load from all tributaries and point sources. The response curves were generated using models calibrated to three sets of HV estimates simultaneously (means of the three sets of estimates shown as circles for the years 1985-1994, squares for the years 1995-2004 and diamonds for the years 2005-2018). HV estimates are colored according to the corresponding average Jan-Jun flow from all tributaries. Shaded area: 95% credible intervals (accounting for parameter uncertainty); solid gray lines: 95% prediction intervals (accounting for parameter uncertainty and prediction error); dashed gray lines: 95% prediction intervals (accounting for parameter uncertainty, prediction error and HV estimation error). Dashed vertical and horizontal lines indicate the mean HV expected under different management scenarios after averaging out year-to-year variability in hydrology (see main text for a description of each scenario).

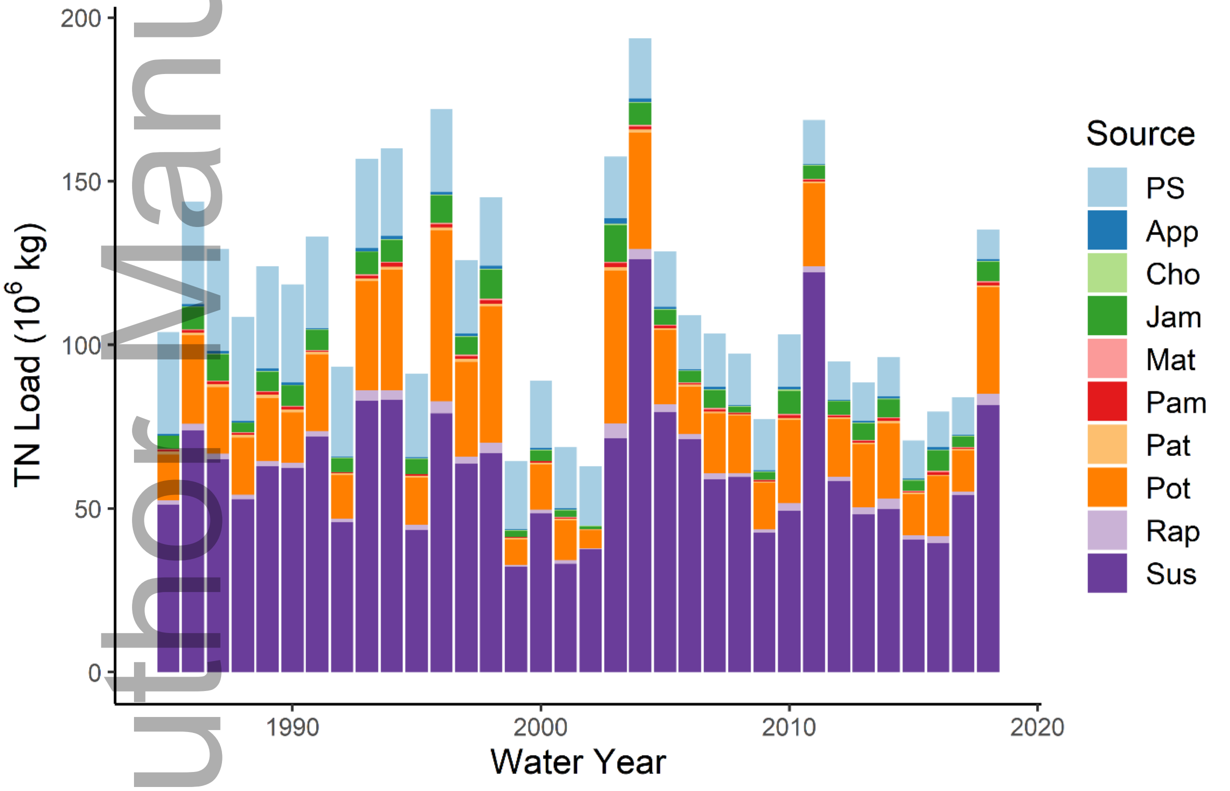

Fig. 1

Author Manuscript

This article is protected by copyright. All rights reserved

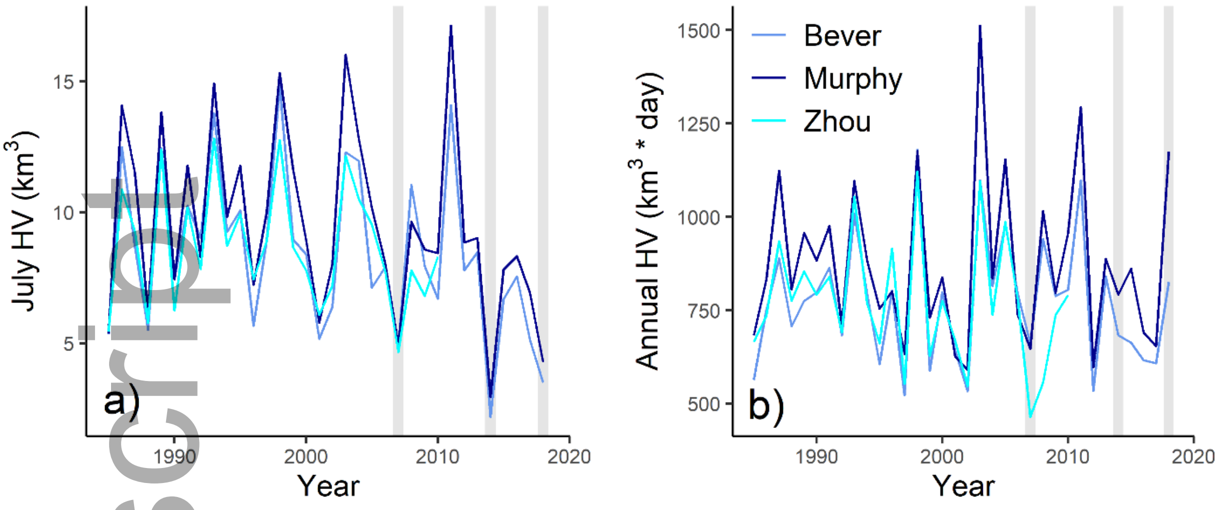

Fig. 2

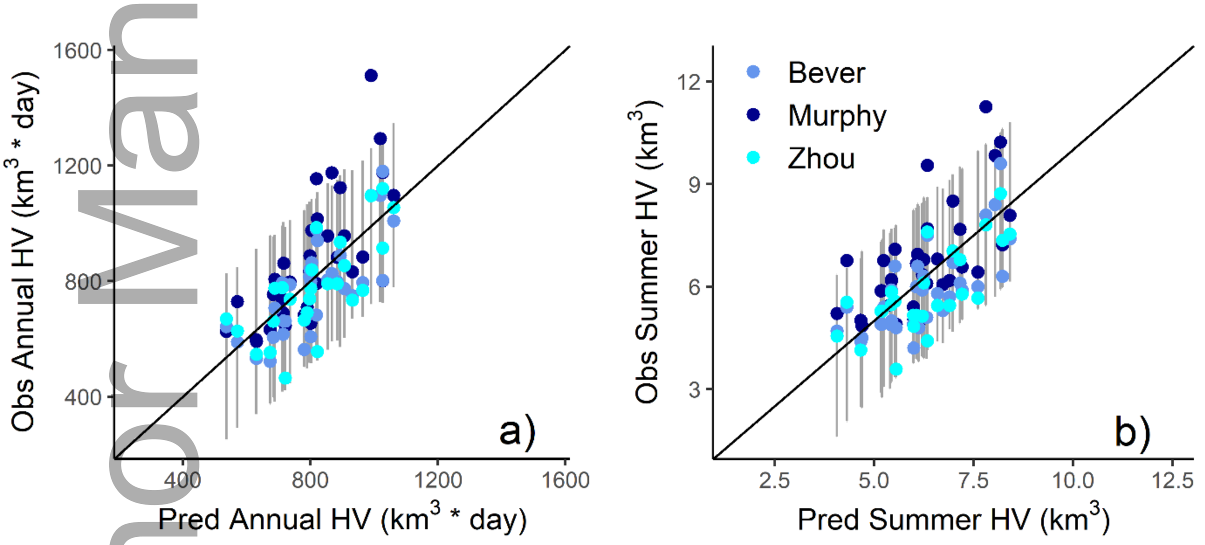

Fig. 3

Author Manuscript

This article is protected by copyright. All rights reserved

Author Manuscript

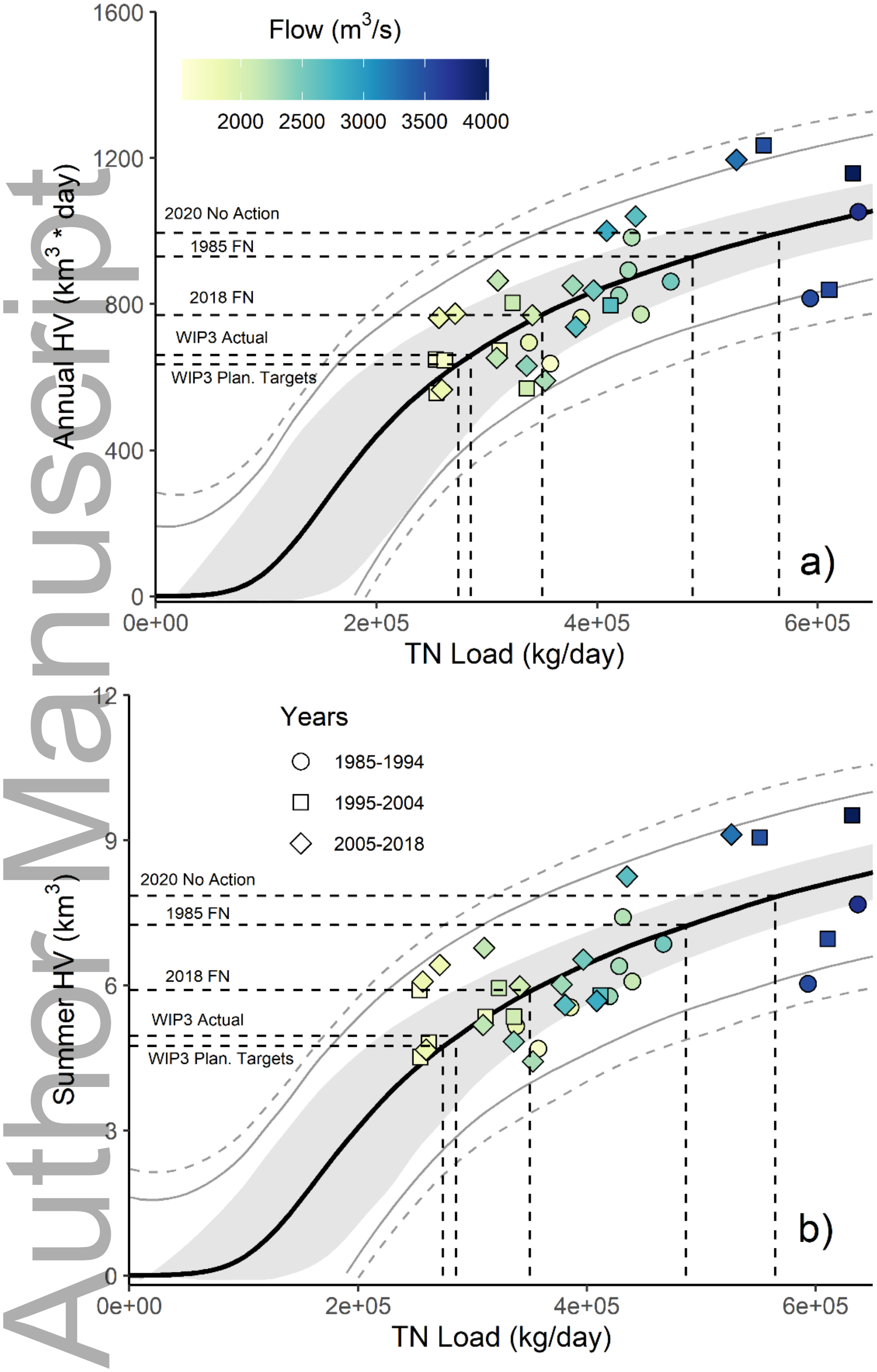

Fig. 4

This article is protected by copyright. All rights reserved