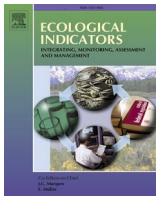
Ecological Indicators 140 (2022) 109022

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Original Articles

A water quality barometer for Chesapeake Bay: Assessing spatial and temporal patterns using long-term monitoring data

A.R. Zahran [a], Q. Zhang [b], P. Tango [c], E.P. Smith [d,*]

[a] Math and Statistics Department, Utah State University, 3900 Old Main Hill, Logan, UT 84322, USA
[b] University of Maryland Center for Environmental Science / U.S. Environmental Protection Agency Chesapeake Bay Program, 1750 Forest Drive, Suite 130, Annapolis, MD 21401, USA
[c] U.S. Geological Survey / U.S. Environmental Protection Agency Chesapeake Bay Program, 1750 Forest Drive, Suite 130, Annapolis, MD 21401, USA
[d] Department of Statistics, Virginia Tech, Blacksburg, VA 24060, USA

ARTICLE INFO

Keywords:
Water quality criteria
Dissolved oxygen
Frequency
Magnitude
Duration
Bootstrap
Numerical criteria
Designated use
R shiny

ABSTRACT

This paper develops a barometer that indexes water quality in the Chesapeake Bay and summarizes quality over spatial regions and temporal periods. The barometer has a basis in risk assessment and hydrology, and is a function of three different metrics of water quality relative to numerical criteria: relative frequency of criterion attainment; magnitude of deviation from a numerical criterion; and duration of criterion attainment. Metrics associated with these features are calculated at the station level, allowing flexibility for simultaneously evaluating multiple stressors, different designated uses, and physical characteristics of the water. The barometer score is then created as a geometric mean of the three metrics. The water quality barometer (WQB) station scores may be spatially aggregated to report habitat scores across a spectrum of spatial resolutions (e.g., management segment, tidal subsystem, or the whole tidal bay). Dissolved oxygen measurements in the Chesapeake Bay collected during summer seasons of 1985 to 2020 are used to evaluate water quality. The WQB score and its bootstrapped confidence interval are reported at the station, segment, tidal subsystem and whole tidal bay levels. Notably, water quality interpreted through application of the WQB with dissolved oxygen concentration data and averaged over the 29-year period of record is good (i.e. protects aquatic living resources) in tributaries such as the James River, Rappahannock River and others; but is not as good in areas such as the Upper Tributaries and the York River. Recent summaries indicate that while the water quality is improving in much of the bay and its tidal tributaries, however, there is an indication of decline in quality in the period 20182020, especially in the upper regions of the Bay. The barometer is designed around using the time series data produced by the Chesapeake Bay Programs annual monitoring strategy; the approach has application to other large water bodies with large scale monitoring programs with extended time series or for integrating information from environmental sensor systems.

1. Introduction

To assess water quality of large aquatic systems, many water quality variables (characteristics) such as dissolved oxygen (DO), nutrients, chlorophyll-*a* and water clarity are often monitored. For regulatory purposes, many of these characteristics are often assessed relative to specific standards or numerical criteria designed to protect living resources, aesthetics, recreational use, or water supplies for human health. Water is potentially considered impaired or of lower quality with respect to the characteristics when exceeding criteria thresholds that support definitions of water quality standards. The standards may be different for different habitats based on the protection target as defined through a designated use (DU) that may vary over space, time and water characteristics. Different approaches have been used to evaluate and summarize water quality, including: i) indicator organisms such as bacteria (Leight, Crump and Hood, 2018), fish (Karr, 1981), macrobenthos (Weisberg et al., 1997) and submerged aquatic vegetation (SAV) (Moore et al., 2000, Orth et al., 2010), ii) environmental report cards (Williams et al., 2010), iii) weighted and un-weighted average of attainment/compliance indicators (Carlson, 1977, Smith et al., 2001; USEPA, 1997, USEPA, 2003a, USEPA, 2007; USEPA, 2017; Zhang et al., 2018a; Zhang et al., 2018b), iv) principal components and factor analysis (Mustapha

* Corresponding author.
E-mail address: epsmith@vt.edu (E.P. Smith).

https://doi.org/10.1016/j.ecolind.2022.109022
Received 6 April 2022; Received in revised form 26 May 2022; Accepted 28 May 2022
Available online 13 June 2022

et al., 2013), v) fuzzy and machine learning approaches (Kung et al., 1992; Lu et al., 1999; Mujumdar and Sasidumar, 2002,; Ostad-Ali-Askari, Shayannejad, and Ghorbanizadeh-Kharazi, 2017, Langendorf et al., 2021) and vi) acceptance sampling by variables (Smith et al., 2003).

Many large aquatic systems have long-term fixed-station monitoring programs and this leads to the need for summarization of water quality and new indices. The Chesapeake Bay (CB), for example, has benefited from a long-term tidal water quality monitoring program (Fig. 1) (Tango and Batiuk, 2016). Although the program has generated a high-quality set of data for evaluating water quality, understanding changes in water quality over different spatial scales and how this relates to living resources is complicated as the sampling region is divided into multiple segments of different surface area (USEPA, 2005). Each segment may have different numbers of habitats that may change with season as has been reflected in criteria supporting definitions for five designated uses (DUs) (USEPA, 2003a). Several water quality indices are available for estimating attainment of water quality standards. Hernandez et al. (2020) developed a multi-metric indicator of CB water quality, which is a segment-by-DU-surface-area-weighted average of segment-by-DU combinations estimated to be in attainment. Calculation is done on individual water quality characteristics, where for each water quality characteristic of interest, applicable segment criterion attainment measures are calculated for each DU separately and then a DU-segment-surface-area weighted average of all segments estimated to be in attainment over all of these water characteristics with its specified DUs is calculated. The authors recognize the need to consider duration and magnitude of exposure and build this into the numerical criteria using ideas from Batiuk et al. (2009). Using this multi-metric indicator, Zhang et al. (2018a) evaluated CB water quality for every three-year-rolling window during 1985–2016. They used the cumulative frequency diagram (CFD) approach coupled with applying a rule set to account for information gaps with unassessed criteria to estimate attainment status for segments and their degree of non-attainment (USEPA, 2003a, 2017; STAC, 2006, Batiuk et al., 2009; Zhang et al., 2018a; Hernandez et al., 2020). This approach estimates compliance over space and time relative to expectation. A three-year rolling window is used in the calculation to provide an adequate sample size for evaluation of attainment and reduce the influence of episodic weather events (Batiuk et al., 2009).

Although the above approaches are very useful for summarizing bay health from a complex set of measurements the methods primarily evaluate water quality in terms of relative frequency of compliance or estimated attainment of water quality standards. To better protect aquatic living resources and reduce risk to their success in survival, growth and reproduction, direct consideration of the magnitude of exceedance and the duration that a habitat is in criterion exceedance may be useful. An index that could directly capture these habitat features may be of more appeal as it integrates these different aspects of compliance. Such an approach has its basis in the toxicology and risk assessment literature (Diamond et al., 2005; Brooks et al., 2003). For

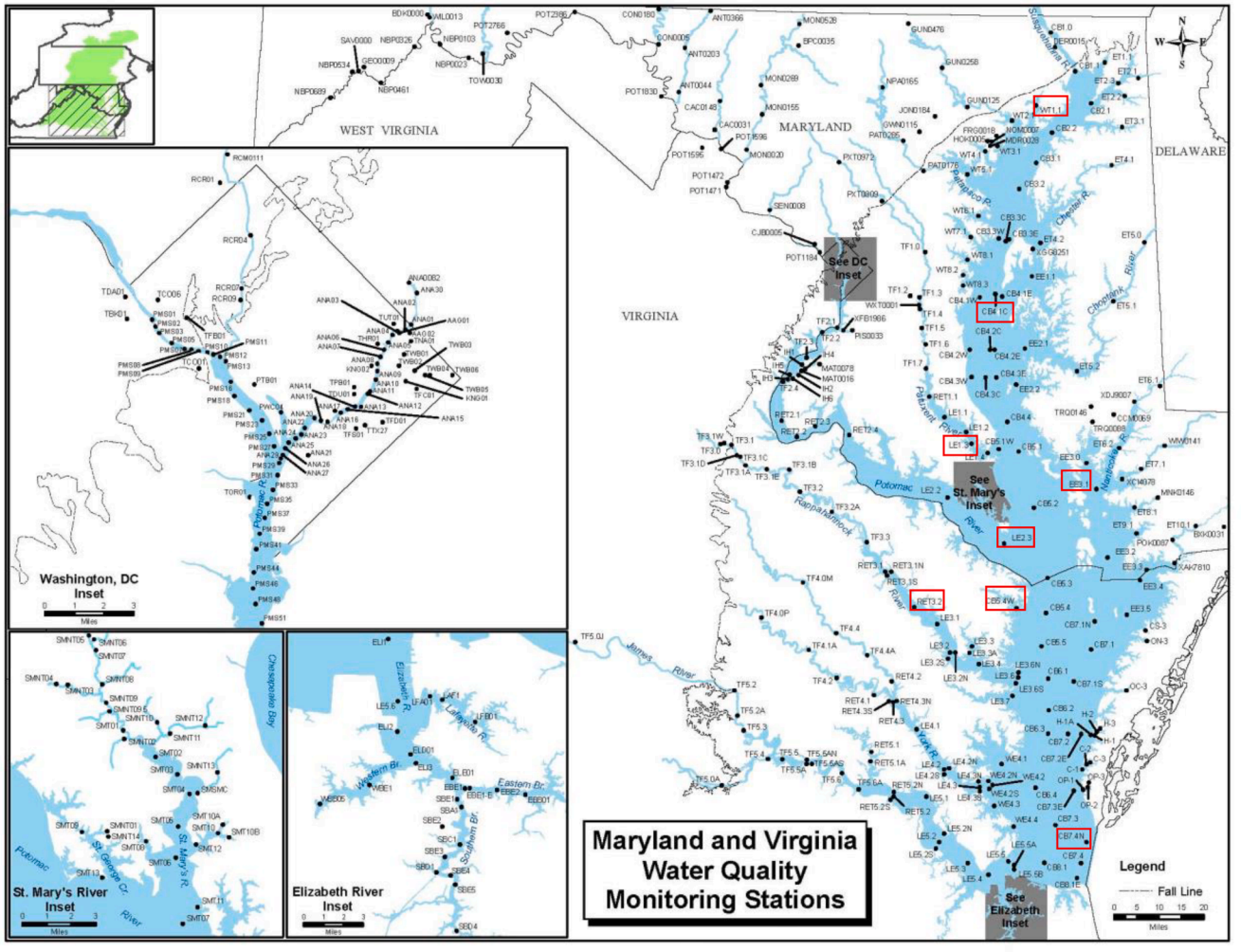


Fig. 1. Chesapeake Bay tidal water-quality monitoring stations. Source: https://www.chesapeakebay.net/images/maps/cbp. Stations in red boxes are the eight stations selected to show its WQB-plot in Section 4.

example, in exposure assessment, frequency, magnitude and duration of exposure to an agent are measured, and are known as exposure factors (Diamond et al., 2005; Gray et al., 2009, NRC, 2007). These factors are used to calculate exposure estimates such as cumulative risk to growth and survival. A related approach is the intensity–duration–frequency (IDF) curves used in water resources management (Sun et al., 2019). Such an index including these factors might be useful as a summary measure of quality as well as an exposure measure or indicator that might be useful in evaluating the effect of water quality on living resources.

In this paper, we propose a new index: a Water Quality Barometer (WQB) that is based on these three metrics, i.e., frequency, magnitude and duration. As numerical criteria are available for Chesapeake Bay tidal waters to suggest thresholds for effects for living resources (Batiuk

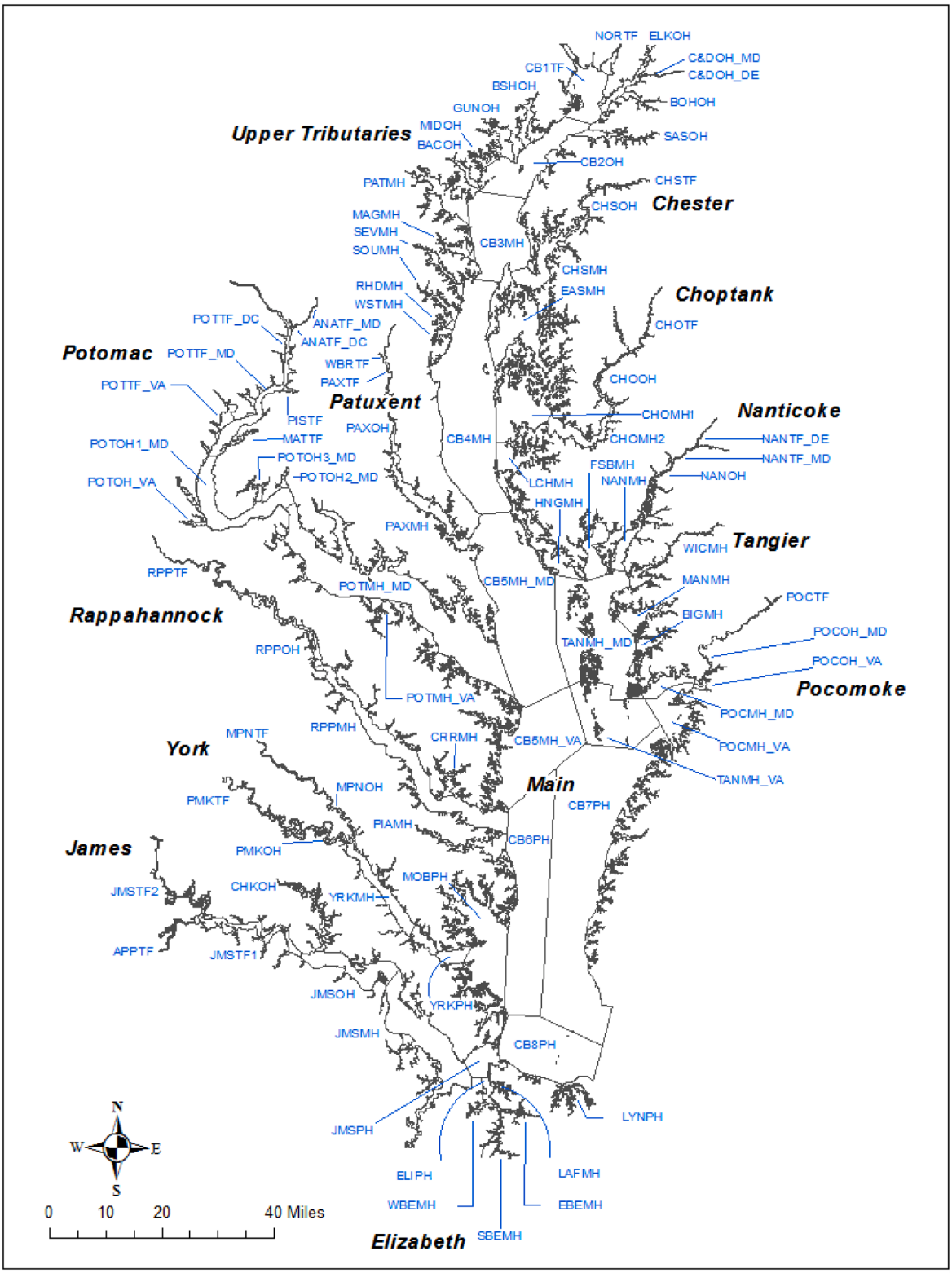


Fig. 2. Chesapeake Bay segments and tidal systems.

et al., 2009, Tango and Batiuk, 2013I), these will be used to form the basis of relative measures. This approach is different than a range of other approaches by simultaneously stressing three important features of habitat exposure impacting living resources versus singular condition assessments. The WQB differs from its closest indicator relative, the Chesapeake Bay Multimetric Water Quality Indicator (Hernandez et al., 2020) that emphasizes criterion stressor magnitude, duration, and return frequency in the numerical criteria using concepts focused on absolute measurements discussed in Batiuk et al. (2009). The new index focuses on the three metrics directly, rather than through the criteria, and features two properties: i) it is calculated at the station level, level one in the measurement hierarchy, and ii) it utilizes the multivariate nature of the data across the different DUs, if multiple DUs are considered. The numerical criteria used in this Chesapeake Bay case study of the WQB are considered as thresholds protective of living resource survival, growth and reproduction conditions in each habitat (USEPA, 2003a), hence the barometer may be viewed as a risk exposure indicator and complementary to the Hernandez et al. (2020) approach.

For the period 1985–2020, the proposed barometer is calculated at the station level (Fig. 1) for each three-year-rolling window using DO concentration measures. Habitats (water units defined through DUs) are assigned to observations based on temperature and salinity data used to compute pycnocline boundaries when present that separate Open Water (OW) from Deep Water (DW) and Deep Channel (DC) DUs. These station WQBs can be aggregated to any specific higher aggregation level of interest, e.g., segment, tidal sub-system or the whole bay.

The paper is structured as follow: Section 2 describes the Chesapeake Bay data set and sampling program. The mathematical details of the metrics and WQB are presented in Section 3. Section 4 presents the results of applying the method to the bay data at the station, management segment, tidal sub-system and whole bay levels over the period from 1985 to 2020. There is also a results comparison with the attainment deficit method of Zhang et al. (2018). Additional figures and analysis are presented in the supplemental material. Some conclusions and further ideas are in Section 5.

2. Data

The Chesapeake Bay is the largest estuary in the United States and its watershed extends to six different states: New York, Pennsylvania, Delaware, Maryland, Virginia and West Virginia, as well as Washington, D.C. Its tidal waters are divided into 92 tidal segments and 13 tidal systems (Fig. 2). The width of the CB ranges between 8 and 48 km (km) and its mainstem is about 300 km long. In the open bay, the depth has an average of 8.4 m (m) with a maximum depth of 53 m. The CB surface area is 11,500 km^2 including open bay and the tributaries (Cerco and Cole, 1993).

Since 1984, the EPA Chesapeake Bay Program (CBP) has monitored the CB water quality at over 150 fixed-site stations with a common protocol of collecting a vertical profile of water quality measurements at each station (USEPA, 2010 section 5, Tango and Batiuk, 2016). Observations of some water quality characteristics, such as DO, chlorophyll-*a*, water clarity, water temperature and salinity are collected at the station level (Fig. 1). Such data suppost asessment summarization at the subbaybay or whole bay level.

Tidal bay segmentation has varied by study and application. For examples, IAN-Ecocheck provides annual reports of water quality conditions for 15 bay regions (https://ecoreportcard.org/report-cards/chesapeake-bay/bay-health/). Llanso et al. (2003) used 67 segments while Llanso et al. (2009) used 85 segments to evaluate bay health based on the benthic macroinvertebrate community integrity measures. Lefcheck et al. (2018) divided the bay into 120 subestuaries with 112 station-units for evaluating SAV trends and factors influencing SAV populations. Since 1984, the CBP partnership has used various versions of a basic segmentation scheme ranging from 78 to 104 segments to organize data collection, analysis and reporting (USEPA, 2004, 2005, 2008). Segments were developed based on hydrodynamics, chemistry, bathymetry and biology. Stations within segments are considered to represent the same set of conditions. Currently, the bay is divided into 92 segments (USEPA, 2008, Fig. 2). Each segment is monitored for compliance with the specific characteristic standards based on a set of criteria for DO, water clarity/SAV and chlorophyll-a applicable across seasons in each DU. Five DUs are identified for the bay: Migratory fish Spawning and Nursery (MSN), Open-Water fish and shellfish (OW), Deep-Water seasonal fish and shellfish (DW), Deep-Channel seasonal refuge (DC), and Shallow Water (SW); (USEPA, 2003b, 2017; Zhang et al., 2018a). Each tidal segment has up to five DUs. All stations inside a segment have the same maximum number of DUs (M_{seg}). A table that identifies the applicable DUs for each segment can be found in USEPA (2017, Appendix F, pp 111–115).

Water-quality monitoring data for the period of 1985–2020 are downloadable from the CBP Data Hub (https://www.chesapeakebay.net/what/downloads/cbp_water_quality_database_1984_present). The data set represents information on DO, temperature and salinity for more than the 150 fixed stations, however, there are a considerable number of missing values with regard to the time frame of sampling. Various decisions were made to try to provide summaries for as many stations as possible over the time period. These decisions are discussed in section 3.6.

3. Methods

The study focuses on application of the WQB for the summer season (June 1-September 30) DO (mg O_2/L) conditions in the tidal Chesapeake Bay and its tidal tributaries. There are one to three applicable habitats relative to summer DO depending on location, stratification, pycnocline formation and pycnocline boundaries defined by vertical water column density structure (USEPA, 2003a, USEPA, 2008). The designated uses (DUs) associated with the habitat layers that may exist for this parameter and season are OW, DW and DC. Data are identified by date, location, depths and layers in the water column. Hence, any observation is defined by its position in a three-dimensional plane defined by a temporal dimension (month, year), a spatial dimension (station, segment, tidal subsystem) and a vertical dimension (depth, layer), and has an associated DU. Station is viewed as the basic sampling unit and measurements within the station for a given month as subsamples.

Let $y_{yr,mo,ts,seg,st,la,dp}$ be a characteristic-measurement of interest described with respect to these three dimensions; i.e., within a specific temporal dimension indexed by year (*yr*) and month (*mo)* and spatial dimension indexed by tidal subsystem (*ts*), segment (*seg*) and station (*st*) measurements are made in a vertical-dimension indexed by layer (*la*) and depth (*dp*). Summer season DO concentration data are combined for 3-year continuous periods when producing a habitat status assessment consistent with a protocol used by the CBP on the data set (Hernandez et al., 2020). In the study period of interest (1985–2020), using an annual time step, there are 34 three-year-rolling windows (*rw*). A station sufficiently identifies its tidal subsystem and segment, therefore, for ease of notation, the subscripts of these two components (*ts* and *seg*) are suppressed. The average of the measurements at each combination of the three dimensions (vertical, spatial, temporal) is calculated, $\bar{y}_{yr,mo,ts,seg,st,la,dp}$, call it $x_{yr,mo,st,la,dp}$. From the data, three metrics are developed and scaled to be within (0,1).

3.1. Magnitude metric

The magnitude metric is intended to measure how far above or below the measurements are from each applicable numerical criterion. We define a distance-based magnitude metric(*MAG*) that compares measurements exceeding the applicable criterion (Bad) with measurements not exceeding the applicable criterion (Good) as follows:

$$MAG_{rw,st} = \frac{1 + magn}{2}, 0 \le MAG \le 1 \quad (1)$$

The *magn* function is the ratio of the difference of two distances relative to its sum (details for the *magn* formula are in Appendix A), so that it ranges between −1 and 1. These two distances are in the opposite directions; one is in the quality direction ('good' distance) and the other one is in the impaired direction ('bad' distance). If $magn \ge 0$, then the 'good' observations that satisfy the criterion (i.e., those with positive deviations) are greater than the "bad' ones that do not satisfy the applicable criterion (i.e., those with negative deviations).

Figs. 3a and 3b illustrates two scenarios of two distances. Fig. 3a depicts these two distances for a hypothetical example where the good distance dominates the bad one. Here $M_{seg} = 2$, $L_1 = 3$, $L_2 = 3.5$, $\bar{\bar{x}}_{rw,st}^{DU_1,G} = 3.9$, $\bar{\bar{x}}_{rw,st}^{DU_1,B} = 2.7$, $\bar{\bar{x}}_{rw,st}^{DU_2,G} = 4.5$ and $\bar{\bar{x}}_{rw,st}^{DU_2,B} = 2.9$. The good distance (green) is larger than the bad distance (red), hence $magn > 0$. Fig. 3b shows another hypothetical example, where $M_{seg} = 2$, $L_1 = 3$, $L_2 = 3.5$, $\bar{\bar{x}}_{rw,st}^{DU_1,G} = 3.9$, $\bar{\bar{x}}_{rw,st}^{DU_1B} = 1.7$, $\bar{\bar{x}}_{rw,st}^{DU_2,G} = 4.5$ and $\bar{\bar{x}}_{rw,st}^{DU_2,B} = 0.7$. In this case, the bad distance is larger than the good one, resulting in $magn < 0$. This function is a multivariate function that can capture whether the net distance is in the desired direction (positive or good) reflecting no violation of numerical criterion, or in the undesired direction (negative or bad) reflecting violation of numerical criterion across all the station's DUs. The magnitude metric (*MAG*) scales the *magn* values to be between 0 and 1. Hence, values of *MAG* greater than or equal to 0.5 reflect favorable status.

3.2. Frequency metric

We define the frequency metric (*PROP*) as:

$$PROP_{rw,st} = \frac{\sum_1^{M_{seg}} prop_{wind,st}^{DU_i}}{M_{seg}}, 0 \le PROP \le 1 \quad (2)$$

where $prop_{rw,st}^{DU_i} = AVE(p_{yr,mo,st}^{DU_i})$ is the three-year rolling-window attainment proportion (compliance proportion) of dissolved oxygen measurements for a specific station for the ith DU during a specific rolling window, i.e., the proportion of measurements from the total within a period that are meeting or exceeding the critical threshold criterion. *AVE* is average function and $p_{yr,mo,st}^{DU_i}$ is the monthly average of the attainment indicator for a measurement *x*:

$$p_{yr,mo,st}^{DU_i} = \sum I_{yr,mo,st,la,dp}^{DU_i}(x) \quad (3)$$

where

$$I_{yr,mo,st,la,dp}^{DU_i}(x) = \begin{cases} 1 & if \ \ x \ge L_i \\ 0 & otherwise \end{cases}.$$

The frequency metric (*PROP*) is basically the average of the station DUs window attainment proportions. Congruently with the raw score method, the station is considered not in violation of numerical criterion for the ith DU at a specific rolling window (*rw*) if $prop_{rw,st}^{DU_i} \ge 0.9$. Note that the compliance proportion cutoff point is the complement of the raw score cutoff point. However, for the frequency metric (*PROP*), a value greater than or equal to 0.9 does not necessarily imply that the station does not exceed the numerical criterion for all its defined DUs. It indicates only that the station does not exceed the numerical criterion for at least most of its DUs. Nevertheless, as *PROP* approaches one, the station approaches being in favorable condition for every DU.

3.3. Duration metric

By definition, station criterion duration in a certain season is the maximum number of consecutive months not exceeding a numerical criterion during this season of a specific year. We define the duration metric (*DUR*) as:

$$DUR_{rw,st} = \frac{\sum_1^{M_{seg}} dur_{rw,st}^{DU_i}}{M_{seg}}, 0 \le DUR \le 1, \quad (4)$$

where $dur_{rw,st}^{DU_i} = \frac{\sum_{j=1}^{3} dura_{rw,st,j}^{DU_i}}{3*(m_{season})}$ is the relative total attainment duration of a specific station for the ith DU during a specific three-year-rolling window, $dura_{rw,st,j}^{DU_i}$ is the duration in year j of the specified window and m_{season} is the total number of months in the studied season, i.e., if season is the whole year then $m_{season} = 12$, if season is summer (Jun-Sep) then $m_{season} = 4$.

If $dur_{rw,st}^{DU_i} > 0.5$, then the station is not in violation of the numerical criterion for most of the months during the specified window, we consider this value as a quality boundary.

The duration metric (*DUR*) is basically the average of the station relative duration over its applicable DUs. A value greater than 0.5 for this metric does not guarantee that this quality boundary is met in all the

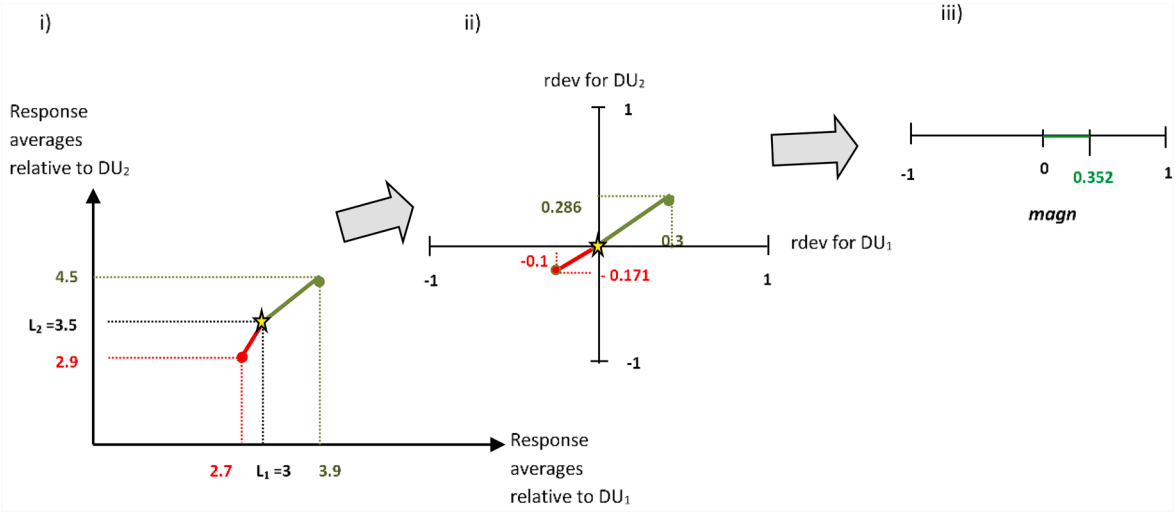


Fig. 3a. A hypothetical station not in violation of a numerical criterion with two DUs assuming Lower threshold (L_i), mapped in three different planes: i) response average plane, ii) relative deviation (*rdev*) plane, iii) magnitu function (*magn*); where red line is bad distance, green line is good distance and star is criterion-defined threshold point (L_1, L_2).

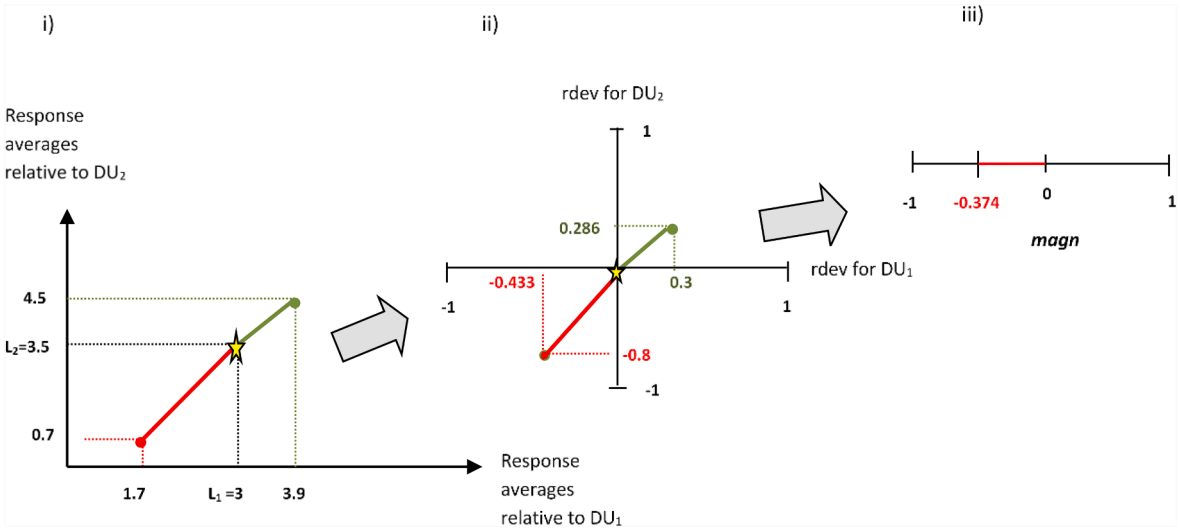


Fig. 3b. A hypothetical station in violation of a numerical criterion with two DUs assuming lower threshold (L_i), mapped in three different planes: i) response average plane, ii) relative deviation (*rdev*) plane, iii) *magn* function; where red line is bad distance, green line is good distance and star is threshold point (L_1, L_2).

DUs, though. Nevertheless, as *DUR* approaches one, the station approaches compliance for every DU in all months in the specified rolling window.

3.4. Water quality barometer

Similar to the idea of Barnes et al. (2007), we define WQB at the station level for every three-year-rolling window as the geometric mean of the above three metrics, i.e.,

$$WQB = \sqrt[3]{(MAG)(PROP)(DUR)}, 0 \leq WQB \leq 1 \quad (5)$$

As WQB values increase the station water quality increases. Using the quality-boundaries of the three metrics, a quality-boundary for WQB is 0.61 ($= \sqrt[3]{(0.5)(0.9)(0.5)}$), thus "good" habitat will have a value of WQB $\geq$ 0.61. However, a value greater than or equal to 0.61 does not necessarily imply that each metric satisfied its quality boundary. Nevertheless, as WQB approaches one, at least two indicators satisfy their quality boundaries. For example, during rolling window 2012–2014, station LE2.3 had a WQB value of 0.637, where MAG = 0.752, DUR = 0.555 and PROP = 0.618. Whereas during rolling window 2016–2018, station WT1.1 had a WQB of 0.897 with PROP = 0.944, MAG = 0.918 and DUR = 0.833 (Fig. 4). To further evaluate the

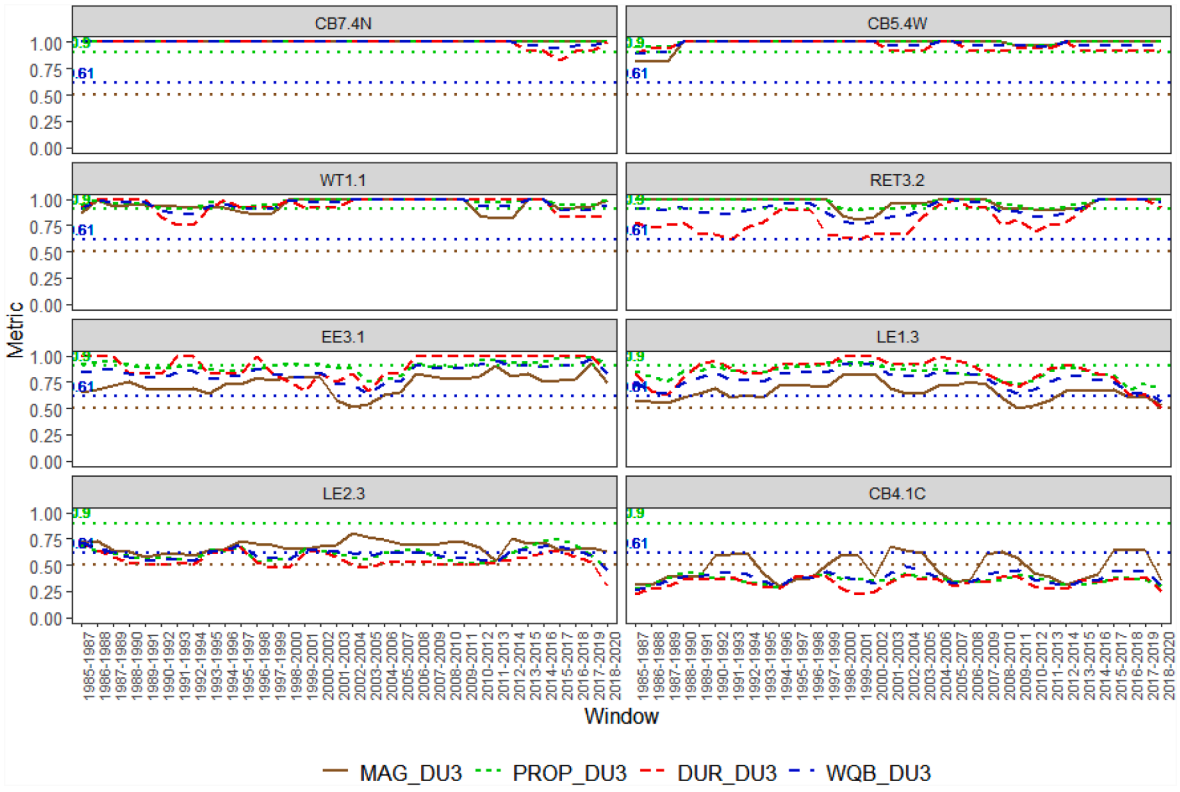


Fig. 4. WQB from a representative sample of stations for applicable combined DUs (DU3) with its three metrics: DUR, MAG, PROP. Dotted lines represent different thresholds.

barometer, confidence intervals for the summaries over the 34 time windows are calculated using a time-series based bootstrap method (R package meboot, Vinod and López-de-Lacalle, 2009). Lastly, we note that although the barometer is developed for multiple DUs, it could be created also for just a single DU.

3.5. Water quality barometer at higher levels of aggregation

Once the metrics and associated barometer are calculated at each station level (level-one unit), the metrics may be evaluated or averaged to any higher level of aggregation, such as segment, tidal sub-system or the whole bay to create the barometer at the higher level of interest. There are several possible approaches to aggregation such as combining the station level barometers, combining (averaging) the station level metrics or using weighted combinations. Here the approach aggregated the station-level metrics using the hierarchy principle (i.e., to get the bay level, we aggregate station results first to segment and then to tidal subsystem and finally to the tidal bay level).

3.6. Data decisions

In this application to CB, the DO data measured during summers of 1985–2020 are summarized. Three DUs (OW, DW and DC) are considered. Hence, the number of months is at most 4, the maximum value of M_{seg} is three and there are 34 rolling windows. Data are thus representative of a monthly summary.

Numerical criteria used to assess DO differ subject to the DU of interest (Batiuk et al., 2009, USEPA, 2003a; USEPA, 2003b, USEPA, 2017; Zhang et al., 2018a; Hernandez et al., 2020). For the above three DUs, the umbrella criterion assumption is used to address assessment status when multiple criteria are used to define health status. For DW and OW, the 30-days standards mean criterion of DO concentrations is used (L_{OW} = 5.5 mg/l with 0–0.5 ppt salinity, L_{OW} = 5 mg/l with greater than 0.5 ppt salinity, L_{DW} = 3 mg/l) to estimate attainment of suitable habitat health where multiple criteria are simultaneously applicable to assess attainment but data are unavailable to evaluate all criteria; while for DC, the instantaneous minimum of DO is used (L_{DC} = 1 mg/l). A measurement below the DO criterion is viewed as representative of potentially unhealthy water quality. In this application, WQB is calculated for the combined applicable DUs (DU3) and for each individual applicable DU (OW, DW, DC).

Although there are a large number of stations in the database, many of these stations do not have complete data. To obtain a more homogeneous number of samples, a number of stations were omitted, and imputation was used to fill in some of the gaps in the data. At the station level, 181 stations have at least one DO measurement in each of the 34 rolling windows; these stations are located in 73 segments. Six segments and 49 stations could be added in the analysis if we consider stations which have measurements in at least 27 rolling windows. For the remaining 13 segments, there are not enough data measurements to adequately represent them without the excessive use of imputation techniques.

In our analysis, we considered the 230 stations that have DO measurements at least in 27 rolling windows, which are located in 79 segments. Exponentially weighted moving average (EWMA) imputation and/or borrowing-from-nearest-station-in-same-segment techniques are used to impute the missing windows (Hamzah et al., 2020). Results are provided for station, segment, tidal subsystem levels as well as the entire bay. All calculations were carried out in SPSS 25 and R version 3.6.3.

4. Results and discussion

The CB water-quality monitoring data for the period of 1985–2020 are used to demonstrate the use of WQB, for individual DU as well as for the combined DUs, at every level of the data hierarchy: station, segment, tidal subsystem and bay level. The similarities in the WQB across DUs are investigated. Percent relative changes in WQB (individual as well as the combined barometer) of 2014–2016 and 2018–2020 compared to the initial window (1985–1987) are investigated. Finally, results from WQB are compared to the attainment deficit (AD) measure published by Zhang et al. (2018b). A R-shiny app that displays the WQB and its elements over time periods is available at https://birchtree.shinyapps.io/CBBarometer/ for station level visualization. Additional displays, especially for segment and higher levels of aggregation are presented in the supplemental material.

4.1. Station level

The frequency, magnitude, and duration metrics underpinning the WQB indicator are first calculated at the station level for summer season DO concentration data in continuous 3-year blocks (e.g., 1985, 1986, 1987) using annual time steps (e.g., 1985–87, 1986–88, etc.). Because there are a large number of stations, a representative subset is graphically displayed in Fig. 4. For each station, a plot that represents the WQB and its three metrics during the 34 rolling windows is produced (WQB-plot). Each plot has three reference lines: i) 0.9, the frequency metric boundary for achievement, ii) 0.5, the magnitude and duration metric quality-boundary, and iii) 0.61, the WQB quality-boundary.

To illustrate some of the patterns in the data, the WQB values were rank ordered, then grouped into eight classes and a representative pattern from each group graphed. Fig. 4 shows the WQB-plot for eight stations chosen according to the values of WQB of the combined applicable DUs (DU3), where the average rank of the 230-station WQB values over the 34 rolling windows is calculated (rank 1 being the best). Then these 230 average ranks are arranged into eight classes; each with approximately 29 stations. From each class, one station is chosen. The eight selected stations are marked in Fig. 1 with a red box. These eight stations are ordered in Fig. 4 (row wise), such that the first panel shows the selected station from the first class (the best 29 stations on average) while the last panel shows the selected station from the last class (the worst 29 stations on average). The stations are selected from these classes, such that they vary in the number of applicable DUs.

During all three-year rolling windows, stations from the first four rank-classes feature high WQB values (>0.8) as well as high values in all its metrics, where none of the metrics violates its quality boundary (*CB7.4N, CB5.4W, WT1.1 and RET3.2*). *DUR* is the most volatile metric in these classes. Stations EE3.1 and LE1.3 (from the 5th and 6th rank classes) still feature good WQB (>0.61; its quality boundary), with the exception of the last window of LE1.3. The three metrics exhibit fluctuations, where in some windows *PROP* violates its quality boundary (0.9). For LE2.3, which represents the seventh class, violations of the quality boundary of the frequency metric are more prominent, leading WQB to violate its 0.61 quality-boundary in some windows. The other two metrics (*DUR* & *MAG*) generally do not violate their 0.5 quality-boundary. In the last rank-class (the worst stations), *DUR* and *MAG* begin to fall below their quality boundary, and hence WQB begins to fall below its quality boundary in all windows (station CB4.1C). The number of applicable DUs in the station does not appear to affect WQB. There are stations with three DUs that have high WQB (*CB5.4W and RET3.2*), while there are others of the same number of DUs with very low WQB (*CB4.1C*).

Station CB4.1C, located in the mid-bay region of the bay's mainstem, is the worst performer among these chosen stations. This region is known for its chronic annual hypoxia conditions (Hagy et al., 2004, Testa et al., 2017) where summertime hypoxia has been documented for over 80 years (Testa et al., 2017, CBP, 1991).

For these eight stations, the WQB is created for each applicable DU separately. Fig. 5 shows these individual WQBs as well as the WQB of the combined applicable DUs (DU3).*WQ* for DU3 is an integration of the individual DU WQB. Note that for stations with only one applicable DU, its WQB for DU3 is just the WQB for that DU.

Apart from station RET3.2 (Fig. 5), the relationships among these

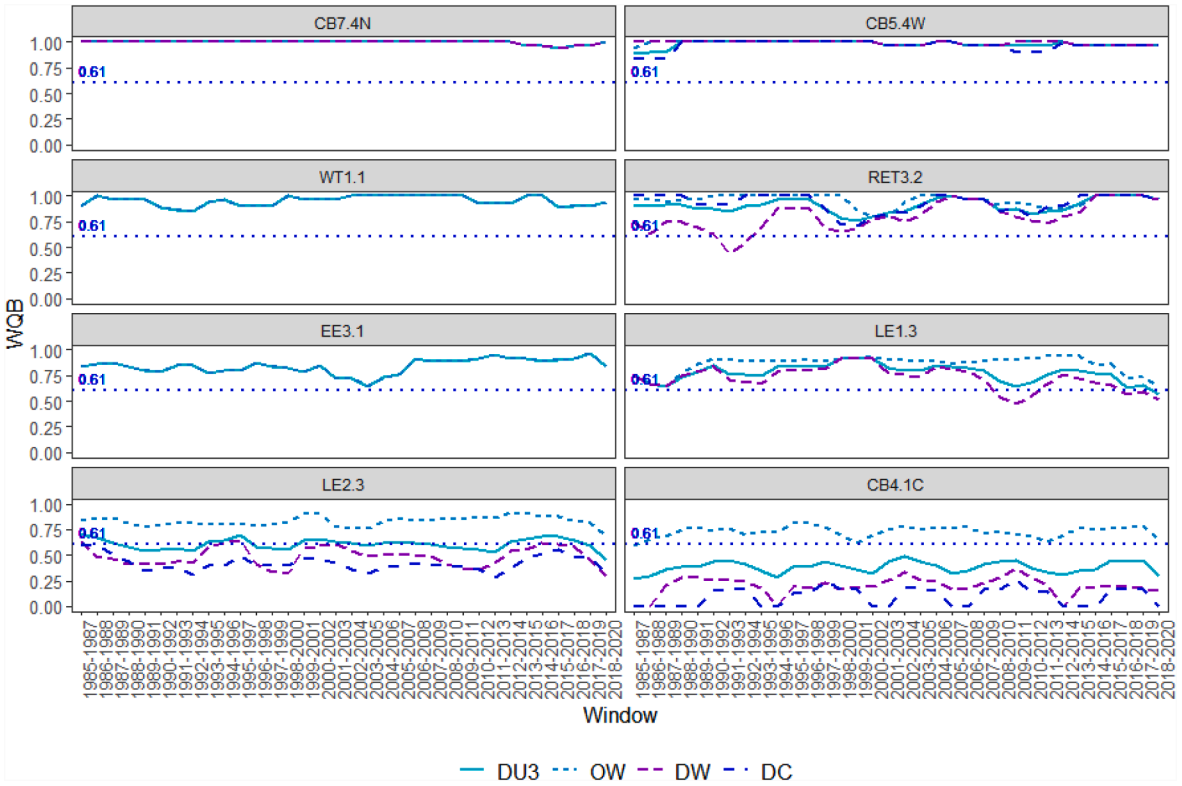


Fig. 5. WQB at the station level for combined applicable DUs (DU3) and individual DU: open water (OW), deep water (DW), and deep channel (DC) in selected stations.

four *WQBs* satisfy the following: *WQB* for *DC* $\leq$ *WQB* for *DW* $\leq$ *WQB* for *DU3* $\leq$ *WQB* for *OW*.

The distributions of *WQB* for DU3 and each individual DU during the study period are shown in Fig. 6.

Fig. 7 depicts the WQB for DU3 with its 95% confidence intervals. Here the maximum entropy bootstrap method (R package meboot with B = 1000) is used to create confidence intervals for WQB during the 34 rolling windows in the period of study.

4.2. Segment level

Evaluation of the WQB at the segment level is important as segments are useful in TMDL analysis and management. Over the last 34 rolling windows from 1985 to 1987 to 2018–2020, WQB at the segment level for DU3 is indicative of good quality (WQB $>$ 0.61) with the exception of main bay segment CB4MH. As measured by the barometer, water quality for segments in the middle part of the mainstem (CB4MH, CB5MH_MD, CB5MH_VA) is worse than water quality of the segments in its northern and southern regions. Among the tidal tributary systems, the James, Rappahannock, Chester and Choptank segments are of high quality (WQB $>$ 0.7), while most of the York segments are of lesser quality (0.4 $<$ WQB $<$ 0.5) (Fig. 8).

Fig. 9 shows the relative change in the WQB at the segment level for each DU separately as well as for DU3 in three different windows: Zhang's et al. (2018b) last window (2014–2016) and the most recent two windows (2017–2019 and 2018–2020) compared to the initial window (1985–1987). In terms of WQB of DU3 and OW, most segments in window 2014–2016 are better off than in the initial windows. In window 2018–2020, most segments are worse relative to the initial window. This pattern remains the same for DW and DC: most segments with applicable DW or DC are better off in 2014–2016 compared to the initial window; by 2018–2020, most DW and DC segments are worse than they were in the initial window (Fig. 9).

In terms of WQB for DU3 in 2014–2016, 53 segments improved compared to the initial window, with 41 segments having percent relative change of at most 25%, nine segments increasing by 25% to 75% and three increasing by more than 100%. WBEMH is the most improved with a percent relative change of 136%. 22 of the 26 degraded segments indicated further degradation, i.e., had negative relative change, of at most 25% relative to the initial value. POCTF is the worst performer with a percent relative change of −100%. In 2018–2020, 41 segments improved compared to the initial window, with 33 segments improving by at most 25%, seven segments improving by 25% to 75%, with only one segment (SBEMH) improving by 77%. 27 of the 38 degraded segments degraded by at most 25% of the initial value. POCTF is again the worst performer with a percent relative change of −100% (see supplemental material for graphical displays).

In 2014–2016, 54 segments improved compared to the initial window with respect to WQB of OW. Among them, 43 improved by at most 25%, nine segments with percent relative change between 25% and 75%, and two segments with more than 75% relative change percent. WBEMH shows the largest improvement with a percent relative change of 136%. 22 of the 25 degraded segments have percent relative change of at most −25%. POCTF is the worst performer with a percent relative change of −100%. In 2018–2020, only 42 segments show improvement compared to the initial window, with 32 of them improving by at most 25%, nine segments show percent relative change between 25% and 75%, and one segment improved by more than 75%. This segment is SBEMH with a percent relative change of 85%. 28 of the 37 degraded segments have percent relative change of at most −25%. The worst segment is still POCTF.

In terms of WQB for DW, ten segments show improvement in 2014–2016 relative to the initial window. PATMH shows the largest improvement in window 2014–2016 with a percent relative change of more than 150%, while EASMH is the worst segment with −44% relative change. In window 2018–2020, only four segments show improvement. PATMH is still the best performer but with a lower percent relative change of 100%, while SEVMH is the worst segment with −55% relative

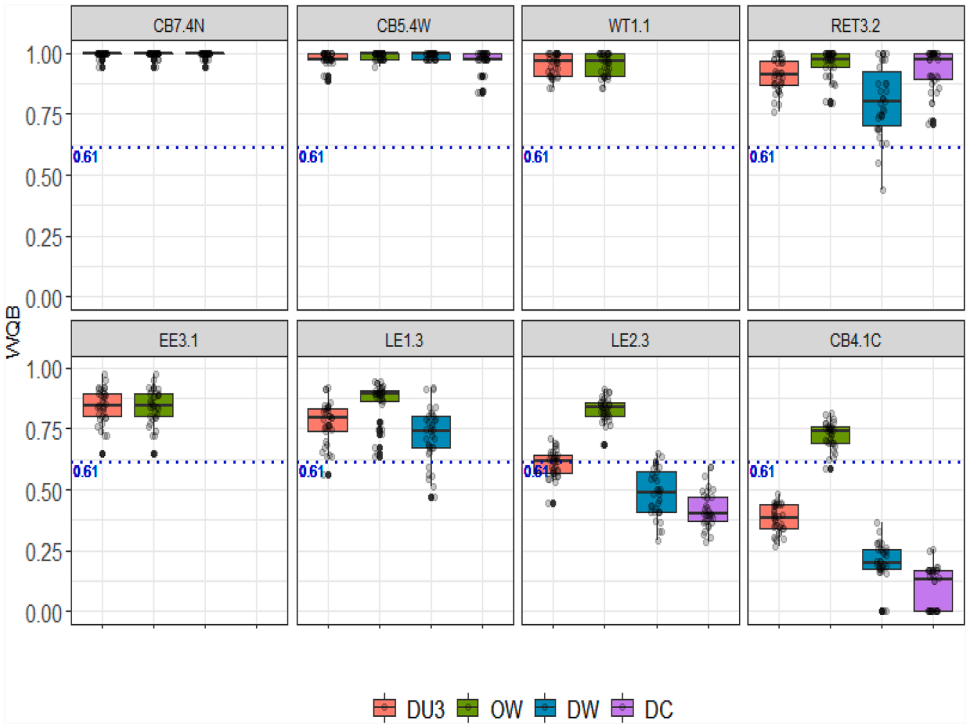


Fig. 6. Distribution of WQB at the station level for applicable combined DUs (DU3) individual DU: open water (OW), deep water (DW), and deep channel (DC) in selected stations during the 34 rolling windows in 1985–2020.

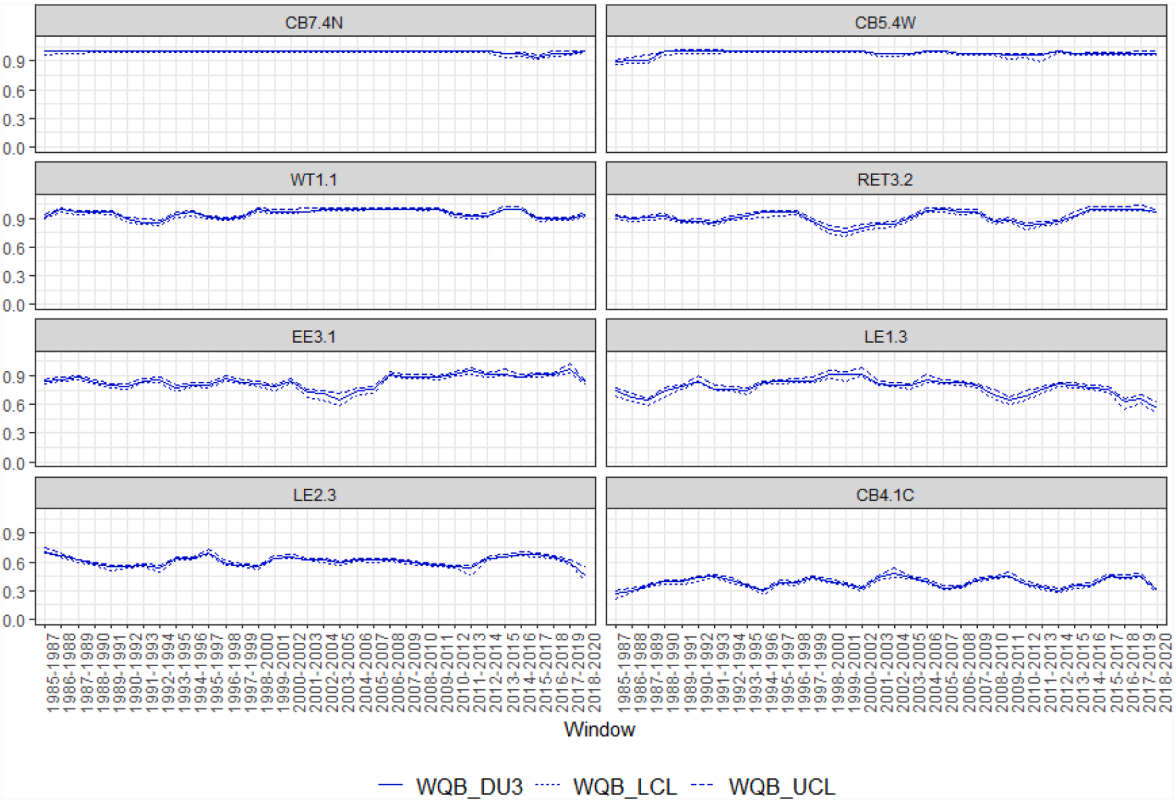


Fig. 7. Sample set of DU3-WQBs at the station level with different designated uses with bootstrapped 95% confidence interval. (LCL = Lower confidence limit, UCL = Upper confidence limit).

change.

With regard to the WQB for DC, five segments show improvement in window 2014–2016 compared to the initial window. CB5MH_VA shows the greatest improvement with a percent relative change of almost 19%, while CHSMH shows the largest degradation with a percent relative change (decrease) of 20%. In 2018–2020, only two segments improved

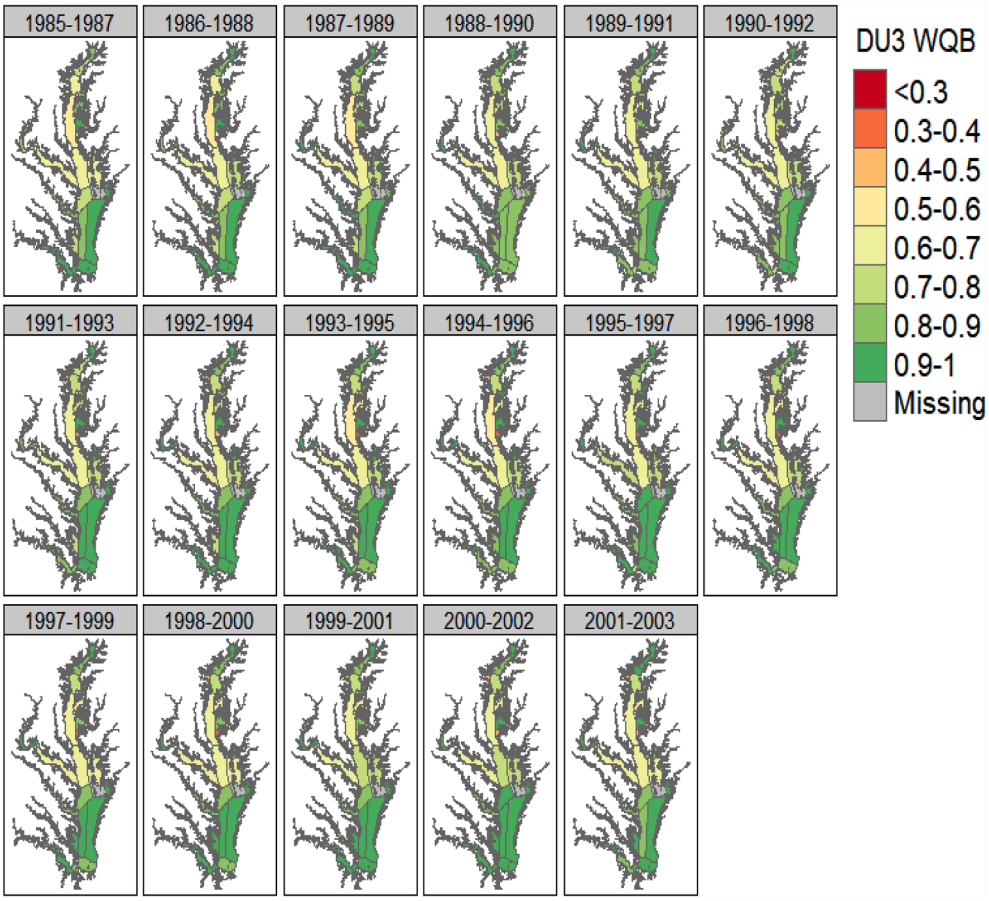


Fig. 8. WQB for Chesapeake Bay aggregated at the segment level during all three-year-rolling windows in 1985–2003. WQB for Chesapeake Bay aggregated at the segment level during all three-year-rolling windows in 2002–2020.

compared to the initial window: CB5MH_VA and POTMH_MD; where CB5MH_VA still being the best segment with a percent relative change of 9%, while PATMH shows the largest degradation with a percent relative change of −48%.

To relatively order segments in terms of water quality during the study period, segments are ranked according to its WQB value in each window separately; then a 95% confidence interval for the average of segment ranks across all windows is calculated. This process is done for the combined version of WQB (DU3), as well as *WQB* for the individual DUs (OW, DW, DC). Among the DC segments, RPPMH of the Rappahannock River is the best segment on average, while PATMH of the Upper Tributaries is the worst. Among the DW segments, CB7PH is the best segment on average, while SOUMH of the Upper Tributaries is the worst. Among the OW segments, C&DOH_MD of the Upper Tributaries is the best one, while POCTF of the Pocomoke is the worst although this is not unexpected as the Pocomoke is a natural blackwater system with historically low DO. These segments remain the best and worst ones in terms of the ranked WQB for DU3, however note that the second worst segment is changed to PMKOH of the York.

4.3. Tidal subsystem level

To evaluate quality at the tidal subsystem level, the WQB for the three designated uses and the combined barometer (OW, DW, DC and DU3) are aggregated to the tidal subsystem level and boxplots are used to display the 34 rolling windows in Fig. 10. Note that not all systems have all four designations. For tidal systems with DW applicable segments, Chester is the best on average across all the 34 rolling windows, while the Upper Tributaries is the worst on average. For tidal systems with DC applicable segments, Rappahannock is the best, and the Upper Tributaries remains the worst one. In terms of WQB for DU3 and for OW, James is the best, while the York is the worst (Fig. 10).

In 2014–2016, 9 tidal systems improved compared to the initial window with respect to *WQB* of DU3, among them 7 improved by at most 20%. Elizabeth is the most improved tidal system (almost 60%). Pocomoke is the worst performer with a percent relative change of almost −10% followed by York with a percent relative change of −9.5%. In 2018–2020, only 5 tidal systems show improvement compared to the initial window, with only one tidal system (Elizabeth River) improving by more than 40%. York and Pocomoke tidal system remain the worst with percent relative changes of −16% and −15%, respectively. This performance/degradation status of the tidal systems remains the same in terms of WQB of OW.

For tidal systems with DW applicable DU, four tidal systems show improvement in 2014–2016 relative to the initial window. Elizabeth shows again the largest improvement with a percent relative change of almost 95%, while Chester is the worst with a degradation of 20%. In window 2018–2020, only Elizabeth is still showing the largest improvement of 70%, while Chester remains the worst performer with a degradation of 48%.

For tidal systems with DC applicable DU, three tidal systems show improvement in window 2014–2016 compared to the initial window, where the mainstem is the best performer with a percent relative change of 10%, while Chester is again the worst performer with a percent relative change of −16%. In window 2018–2020, all tidal systems show degradation with Potomac having the least degradation and the Upper Tributary having the largest degradation, of 0.31% and 39%, respectively.

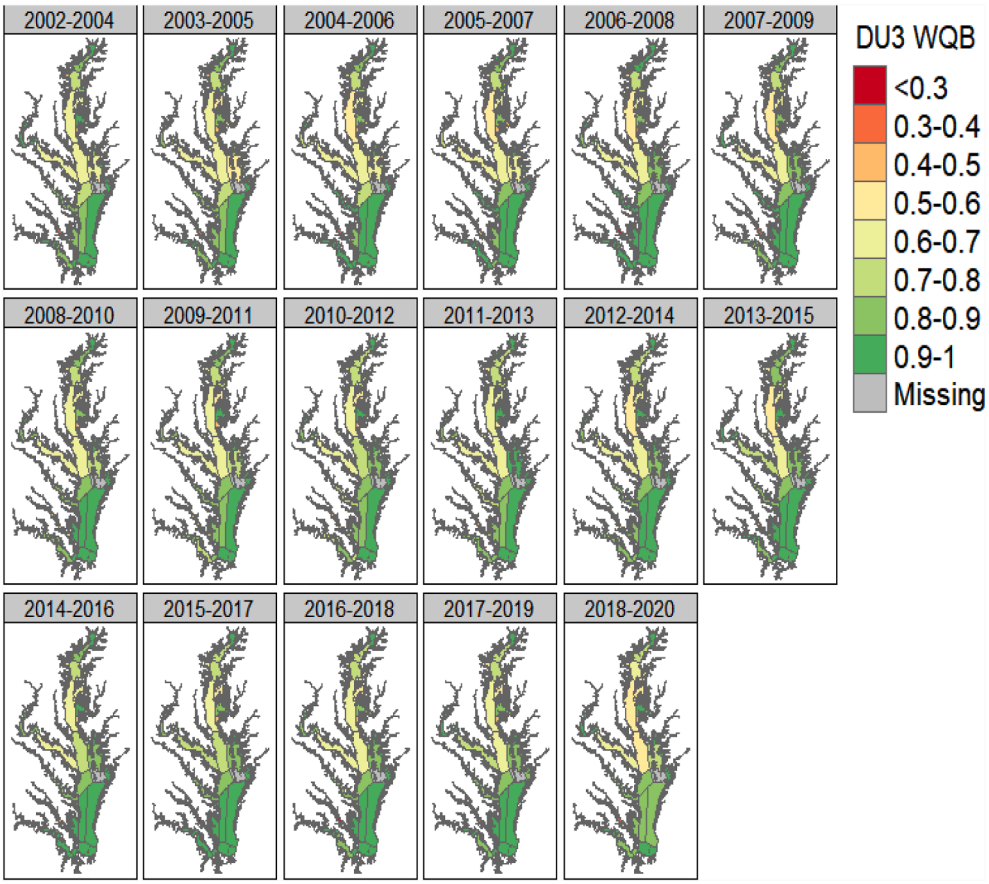


Fig. 8. (*continued*).

4.4. Bay level

During 1985–2020, water quality of the whole bay is generally high (WQB for DU3 $\geq$ 0.74). The lowest values of the barometer were during rolling windows 1989–1991 (WQB = 0.742), 2003–2005 (WQB = 0.752), 2011–2013 (WQB = 0.759) and 2018–2020 (WQB = 0.743). Those periods (windows) align with major weather events affecting bay water quality. For example, major storm events that delivered significant precipitation and subsequent high flows enriched with nutrient and sediment occurred in 2003 (Hurricane Isabel), 2004 (Hurricane Ivan) and 2011 (Hurricane Irene and Tropical Storm Lee) (Zhang et al, 2018a). *WQB* of DU3 records its highest value ($\cong$ 0.83) in window 1992–1994 (Fig. 11).

A decreasing behavior is noticeable in the WQB in the period after 2015–2017 for DC, while it begins even earlier (2012–2014) in WQB for DW. *WQB* for OW does not violate the 0.61 quality boundary in any window, while WQB for DW violates it in five windows (1997–1999 and the most recent four windows). In contrast, the DC WQB could not reach this boundary in any window except 1987–1989, where it barely exceeds the 0.61 boundary (Fig. 12).

Fig. 12 shows the WQB distributions for DU3 and each one of the individual DUs during the 34 rolling windows. Consistent with the results above, bay level *WQB* for DC is the poorest while bay level WQB for OW is the best.

The 95% bootstrapped confidence intervals for WQB for DU3 and each one of the individual DUs during the 34 rolling windows are presented in Fig. 13.

4.5. Comparing WQB to attainment deficit of individual DUs

The attainment deficit (AD) metric for CB is an alternative measure of habitat status introduced by Zhang et al. (2018b) for the CB that focuses on frequency of attainment for DO. It is useful to compare results based on this frequency approach to the WQB approach. Zhang et al. (2018b) analyzed CB-summer data for DO with three designated uses (OW, DW, DC). The AD is created at the segment level and then is aggregated to higher levels. In their analysis they compared the last window in their analysis (2014–2016) to the initial window (1985–1987) for four aggregation levels, namely, the whole bay (total), different DUs (OW, DW, DC), different salinity zones (tidal freshwater or TF, polyhaline or PH, mesohaline or MH, oligohaline or OH) and tidal subsystems. We consider the same levels and the two window comparison to compare results for the WQB to the AD indicators of water quality condition For three aggregation levels (total bay, salinity zones and tidal subsystems), the combined WQB (DU3) is used, while the WQB at the individual DUs is used for the DU level.

The status of segments with DC in window 2014–2016 is slightly worse than it was in the initial window. For the other DUs and the whole bay level, water quality conditions are slightly better in window 2014–2016 than they were in the initial window. This is consistent with the AD results, except for DC, where its AD was better in 2014–2016 than it was in the initial. In addition, the WQB ranks them as: OW > total bay > DW > DC, while AD ranks them as: OW > DW > total bay > DC. The WQB tends to group the OW and total bay together and different from the DW and DC while the AD groups OW, DW and total bay together.

With regard to the salinity zones, the WQB and AD results are similar; as both reveal that conditions for each zone are better in 2014–2016 than they were in the initial window. However, the WQB indicates a bit more variation in the zones and orders the zones in 2014–2016 as PH > OH > TF > MH, while AD orders them as: PH > TF > OH > MH. For the tidal systems, five tidal systems: Chester, Choptank, Rappahannock,

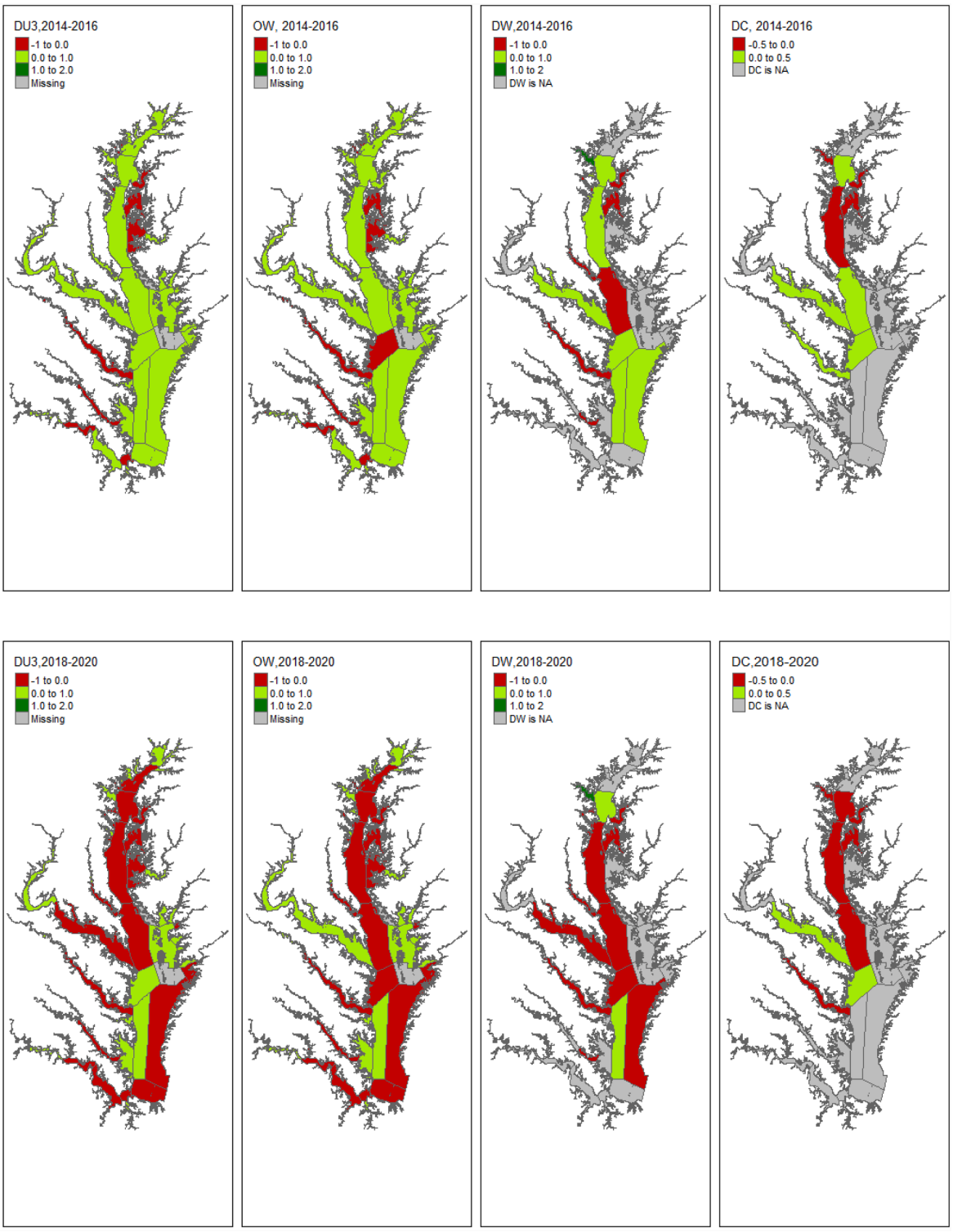


Fig. 9. Relative change in the WQB for different designated uses (DU3, OW, DW, DC) aggregated to the segment level compared to the initial window 1985–1987. Upper panels are for 2014–2016; lower panels are for 2018–2020. Note the ranges are different for the deep channel.

Pocomoke and York are showing degradation associated with DO based on the WQB. In terms of AD, however, only three tidal systems are viewed as degrading: York, Tangier, and Pocomoke. While AD and WQB criteria coincide in their evaluation for most of the tidal systems, they contradict on a few of them, e.g., Tangier, Chester, Choptank, and Rappahannock. In general, while the two measures indicate some similarity, the WQB values are more variable than the AD values and provide different rank ordering for water quality of the segments and tidal subsystems (graphical displays are in the supplemental material).

5. Conclusions, implications and applicability

This paper proposes a new water quality barometer that combines the frequency, duration and magnitude of one or more water quality parameters relative to critical threshold criteria into a single measure. These three metrics measure aspects of water quality important to

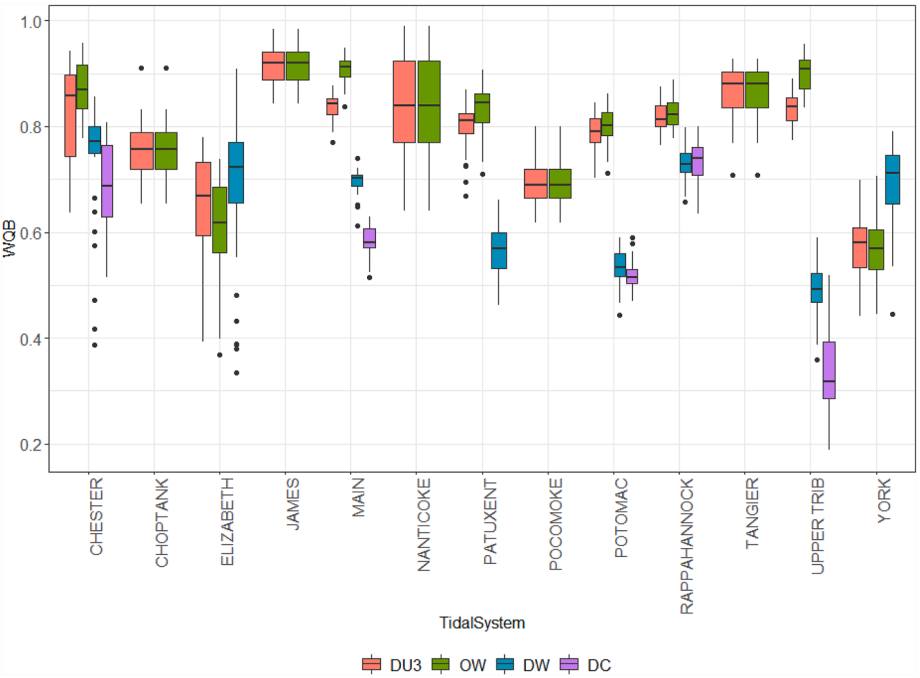


Fig. 10. Distribution of WQB at tidal system level for combined applicable DUs (DU3) and individual DU: open water (OW), deep water (DW), and deep channel (DC) during the 34 rolling windows in 1985–2020. Boxplots represent medians, quartiles and extreme values.

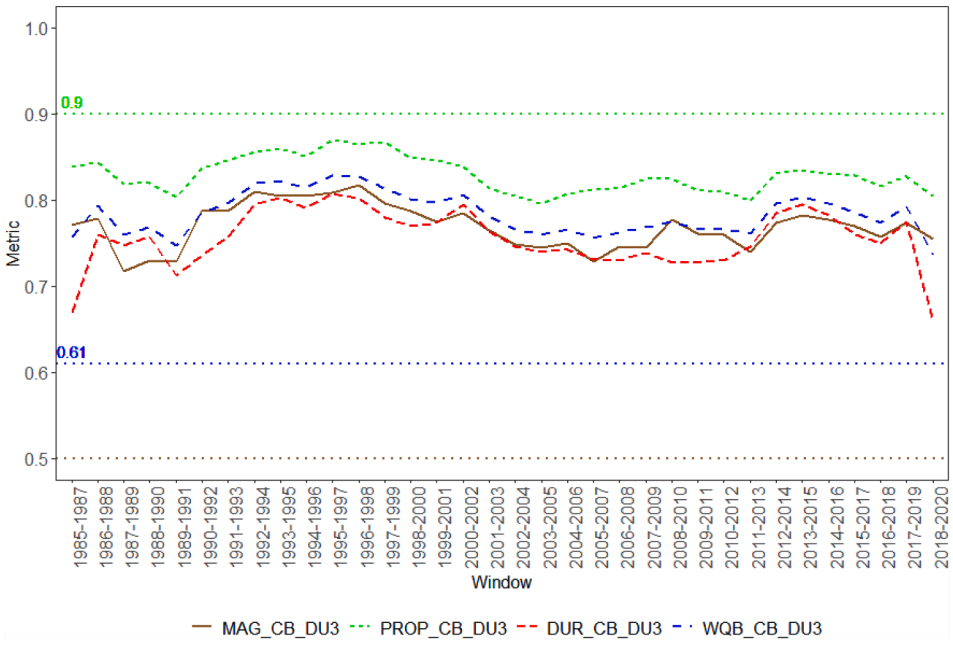


Fig. 11. WQB at CB level for the applicable combined DUs (DU3) with its three metrics: DUR, MAG, PROP.

survival, growth and reproductive life history functions in living resources and represent risk factors expressed by habitat health. Definitions of these metrics are presented mathematically in the context of data associated with the CB long-term water quality monitoring program. The barometer is an indicator developed as the geometric mean of frequency, duration and magnitude and is calculated at the station level. Besides value as a summary measure, the barometer may be useful in the analysis of aquatic living resources and their relationship with cumulative conditions and specific features of habitats and their water quality.

When applied to CB dissolved oxygen measurements, the metrics and resulting WQB indicator capture water quality variability across the spectrum of bay health conditions from high stress low oxygen conditions to low stress well oxygenated waters. Assessments were made across all the defined designated uses applicable for stations, the smallest spatial unit of measurement. Once the three metrics and resulting WQB are calculated, they can be aggregated to higher levels of interest such as segment, river, tidal subsystem or the whole bay to provide scale-specific insights on habitat quality. Comparing the WQB behavior to another indicator, the AD indicator, shows that there is generally agreement however there are some notable differences. Although, the WQB is not an EPA approved indicator, the work provided

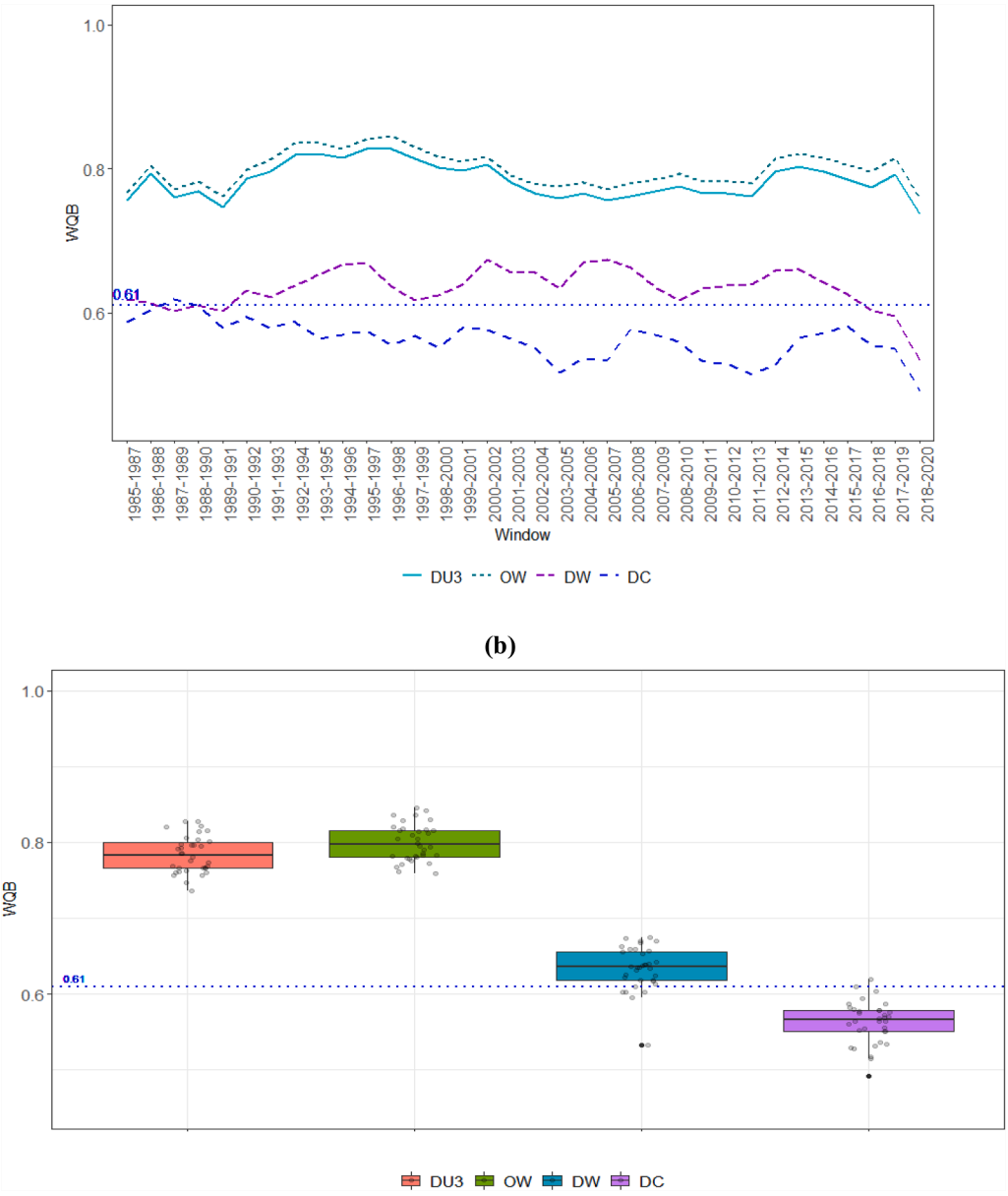


Fig. 12. WQB at the bay level for combined applicable DUs (DU3) and individual DU: open watefr (OW), deep water (DW), and deep channel (DC); a) time plot representation b) boxplot representation with jittered points.wj]

here demonstrates its value and utility to provide complementary information regarding the status and trends of DO habitat conditions in CB at diverse scales of interest for time and space. This information is critical to the Chesapeake Bay Program partnership for understanding the dynamics of the bay ecosystem and for assessing the effectiveness of management initiatives aimed toward CB restoration. More broadly, the WQB can be easily applied to other large water bodies with large scale monitoring programs or for integrating information from environmental sensor systems. Our application of this barometer to CB provides an example where long-term water quality monitoring data and a science-based index approach can be combined to evaluate complex ecosystems.

The geometric mean was selected as a way to combine the three metrics. Other approaches are possible such as the arithmetic mean however, we note that the geometric mean is better than the arithmetic mean in cases where large variations/fluctuations among the components occur or components are not independent of each other. Also, the geometric mean is much less sensitive to outliers (Hirzel and Arlettaz, 2003). In developing the WQB, the same weight was given to each metric in the formula of the geometric mean however, different weights could be utilized as well if desired (Barnes et al., 2007). In addition, we used cutoffs of 0.5 for the magnitude and duration metrics and 0.9 for the frequency metric. While these cutoffs have some justification as thresholds between good and poor water quality conditions, other cutoffs could be developed based on empirical and theoretical evidence.

There are also potential opportunities for variations of the WQB as a summarization metric of risk exposure. The numerical criteria based on designated uses leads to the cutoff criteria for our application. Other criteria might be useful for ecological applications. For example, if interest is in a specific fish species, one might consider DO concentration cutoffs based on behavioral response to habitat conditions, growth rate impairment or mortality rates if these cutoffs are available. For adult striped bass (*Morone saxatilis*), for example, a range of 3–4 mg/l might be used for avoidance behavior and <2 mg/l for mortality (Lipton and Hicks, 2003); a different barometer cutoff might be used for juvenile or young-of-the-year. The resulting indicators might be useful not just for summarization but also in computer modeling or regression analysis of

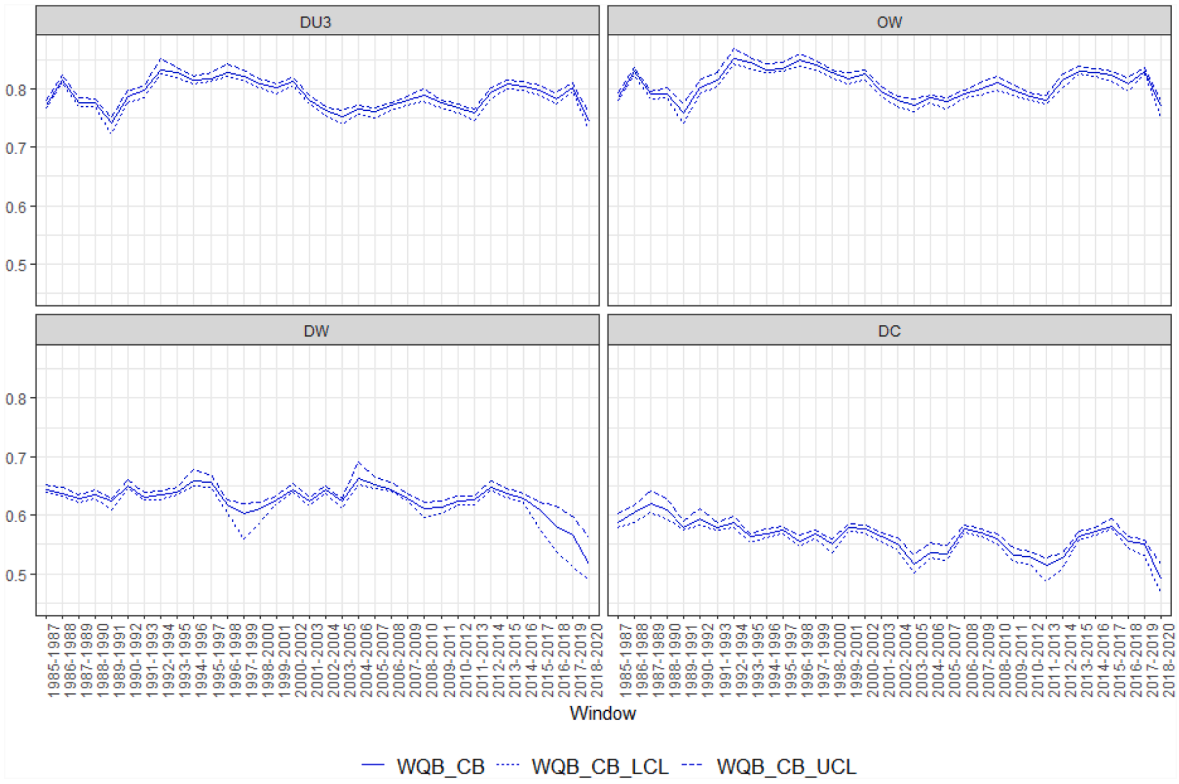


Fig. 13. WQB at CB level with bootstrapped 95% Confidence interval for all uses (DU3, OW, DW and DC).

living resources as an alternative to approaches that predict the effects of exposure. The function used for calculating magnitude assumes a linear effect; this can be adjusted, for example, one might consider an "S" shaped curve for cases where there are lower and upper limits to effects. Another variation that might be useful is to change the numerical criteria based on other environmental factors. For example, temperature might affect habitat suitability for endangered Atlantic sturgeon and a barometer could easily be developed to visualize potential thermal stress associated with winter mortality (Markin and Secor, 2020). This approach would be consistent with the use of temperature in the setting of the numerical criterion. With shortnose sturgeon, for example, EPA (2003) specifies "at temperatures stressful to shortnose sturgeon (>29 °C), a 4.3 mg liter^{-1} instantaneous minimum criteria should apply". There are potential limitations to the WQB as well as other summarizations of CB water quality. Although there are a relatively large number of stations that provide DO data, the number of samples within a year for a station is not large. A three-year moving window was used to increase sample size and remove the effects of episodic weather events. Near continuous measurements using sensor systems would increase the number of annual measurements (and hence temporal variability). Segment information requires sampling over space and is also possibly limited. Also, while the sampling program for the bay is well-supported and effective, it is based on fixed locations rather than random locations and hence is oriented towards trend assessments rather than status (compliance) assessment. Despite possible limitations, the barometer should give researchers a different view of water quality that may be used to help understand, quantify and explain estuarine habitat dynamics, and identify important relationships between water quality and the health of bay aquatic living resources.

CRediT authorship contribution statement

A.R. Zahran: Conceptualization, Investigation, Writing – review & editing. **Q. Zhang:** Writing – review & editing. **P. Tango:** Conceptualization, Writing – review & editing. **E.P. Smith:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Breck Sullivan and Leslie Hager-Smith for comments on an earlier version of the manuscript and appreciate the comments from three reviewers. We also thank Zhaoying Wei for making the segments map. The Chesapeake Bay Program's monitoring team is acknowledged for collecting, validating, and maintaining the long-term water-quality monitoring data. We thank Virginia Tech's Open Access Subvention Fund for financial support to help cover page charges. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Appendix A. Details for *magn* function of the *MAG* indicator

$$magn = \frac{\sqrt{\sum_{i=1}^{M_{seg}} \left(Rdev^{G}{}_{i}\right)^{2}} - \sqrt{\sum_{i=1}^{M_{seg}} \left(Rdev^{B}{}_{i}\right)^{2}}}{\sqrt{\sum_{i=1}^{M_{seg}} \left(Rdev^{G}{}_{i}\right)^{2}} + \sqrt{\sum_{i=1}^{M_{seg}} \left(Rdev^{B}{}_{i}\right)^{2}}} \quad \text{(A1)}$$

i = 1,2,…., M_{seg},

M_{seg} is the number of DUs applicable for the segment.

$Rdev_i^B = \frac{\bar{\bar{x}}_{rw,st}^{DU_i,B} - L_i}{L_i} = \frac{dev_i}{L_i}$, measures the average bad-deviation from DU_i lower standards for a particular site/station during a specific window r*w*,

$\bar{\bar{x}}_{rw,st}^{DU_i,B} = \frac{\sum_1^{n_{rw,st}^{DU_i,B}} \bar{x}_{yr,mo,st}^{DU_i,B}}{n_{rw,st}^{DU_i,B}}$, is the window average of bad measurements with respect to DU_i for a particular station or spatial unit,

$\bar{x}_{yr,mo,st}^{DU_i,B} = \frac{\sum_1^{n_{st,yr,mo}^{DU_i,B}} x_{yr,mo,st,la,dp}^{DU_i,B}}{n_{yr,mo,st}^{DU_i,B}}$ *for* $x_{yr,mo,st,la,dp}^{DU_i} < L_i$, is the monthly average[1] of bad measurements with respect to DU_i for a particular station in a specific year,

$Rdev_i^G = \frac{\bar{\bar{x}}_{rw,st}^{DU_i,G} - L_i}{L_i} = \frac{dev_i}{L_i}$, measures the window average good-deviation from DU_i lower criterion for a particular site during a specific window *rw*,

$\bar{\bar{x}}_{rw,st}^{DU_i,G} = \frac{\sum_1^{n_{rw,st}^{DU_i,G}} \bar{x}_{yr,mo,st}^{DU_i,G}}{n_{rw,st}^{DU_i,G}}$, is the average of good measurements with respect to DU_i for a particular station or spatial unit,

$\bar{x}_{yr,mo,st}^{DU_i,G} = \frac{\sum_{la,dp} x_{yr,mo,st,la,dp}^{DU_i}}{n_{yr,mo,st}^{DU_i,G}}$ *for* $x_{yr,mo,st,la,dp}^{DU_i} \geq L_i$, is the monthly average of good measurements with respect to DU_i for a particular station in a specific year,

L_i = (lower) numerical criteria for DU_i,

B: bad, and G: good,

$n_{yr,mo,st}^{DU_i,B}$ = number of observations in month *mo* of year *yr* satisfying $x_{yr,mo,st,la,dp}^{DU_i} < L_i$ (i.e., the number of the station's bad observations in a specific month and year) associated with DU_i,

$n_{rw,st}^{DU_i,B}$ = number of bad observations in a particular station during the specific rolling window,

$n_{yr,mo,st}^{DU_i,G}$ = number of observations in month *mo* of year *yr* satisfying $x_{yr,mo,st,la,dp}^{DU_i} \geq L_i$ (i.e., the number of station' good observations in a specific month and year) associated with DU_i,

$n_{rw,st}^{DU_i,G}$ = number of good observations in a particular station during the specific rolling window.

In case of upper numerical criteria (U_i), the relative deviations are defined as following:

$$Rdev_i^G = \frac{U_i - \bar{\bar{x}}_{rw,st}^{DU_i,G}}{U_i} = \frac{dev_i}{U_i},$$

$$\bar{\bar{x}}_{rw,st}^{DU_i,G} = \frac{\sum_1^{n_{rw,st}^{DU_i,G}} \bar{x}_{yr,mo,st}^{DU_i,G}}{n_{rw,st}^{DU_i,G}},$$

$$\bar{x}_{yr,mo,st}^{DU_i,G} = \frac{\sum_{la,dp}^{n_{st,yr,mo}^{DU_i,G}} x_{yr,mo,st,la,dp}^{DU_i}}{n_{yr,mo,st}^{DU_i,G}} \quad for \quad x_{yr,mo,st,la,dp}^{DU_i} \leq U_i$$

$$Rdev_i^B = \frac{U_i - \bar{\bar{x}}_{rw,st}^{DU_i,B}}{U_i} = \frac{dev_i}{U_i},$$

$$\bar{\bar{x}}_{rw,st}^{DU_i,B} = \frac{\sum_1^{n_{rw,st}^{DU_i,B}} \bar{x}_{yr,mo,st}^{DU_i,B}}{n_{rw,st}^{DU_i,B}}, \text{ and.}$$

$$\bar{x}_{yr,mo,st}^{DU_i,B} = \frac{\sum_{la,dp}^{n_{st,yr,mo}^{DU_i,B}} x_{yr,mo,st,la,dp}^{DU_i,B}}{n_{yr,mo,st}^{DU_i,B}} \quad for \quad x_{yr,mo,st,la,dp}^{DU_i} > U_i$$

Scaling the deviation, dev_i, by the respective criterion makes it scale free and unifies the measuring unit across all DUs standards.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2022.109022.

References

Barnes, T.K., Volety, A.K., Chartier, K., Mazzotti, F.J., Pearlstine, L., 2007. A habitat suitability index model for the Eastern Oyster (*Crassostrea virginica),* a tool for restoration of the Caloosahatchee estuary, Florida. J. Shellfish Res. 26, 949–959.

Batiuk, R.A., Breitburg, D.L., Diaz, R.J., Cronin, T.M., Secor, D.H., Thursby, G., 2009. Derivation of habitat-specific dissolved oxygen criteria for Chesapeake Bay and its tidal tributaries. J. Exp. Marine Biol. Ecol., 381, S204–S215.

Brooks, B.W., Foran, C.M., Richards, S.M., Weston, J., Turner, P.K., Stanley, J.K., Solomon, K.R., Slatteru, M.S., LaPoint, T.W., 2003. Aquatic ecotoxicology of fluoxetine. Toxicol. Lett. 142, 169–183.

Carlson, R.E., 1977. A trophic state index for lakes. Limnol. Oceanogr. 22, 361–369.

CBP 1991. Dissolved oxygen trends in the Chesapeake Bay (1984-1990). Computer Sciences Corporation under contract to the USEPA (Contract No.68-WO-0043).

Cerco, F.C., Cole, T.M., 1993. Three-dimensional eutrophication model of Chesapeake Bay. J. Environ. Eng. 119 (6), 1006–1025.

[1] Some stations may have only one bad (or good) measurement in a month, where its value would be still the monthly average.

Diamond, J., Bowersox, M., Latimer, H., Barbour, C., Bearr, J., Butcher, J., 2005. Effects of pulsed contaminant exposures on early life stages of the Fathead Minnow. Arch. Environ. Contam. Toxicol. 49, 511–519. https://doi.org/10.1007/s00244-005-7023-8.

Gray, T.R., LaGasse, L.L., Smith, L.M., Derauf, C., Grant, P., Shah, R., Arria, A.M., Della Grotta, S.A., Strauss, A., Haning, W.F., Lester, B.M., Huestis, M.A., 2009. Identification of prenatal amphetamines exposure by maternal interview and meconium toxicology in the Infant Development, Environment and Lifestyle (IDEAL) study. Drug Monitor 31 (6), 769–775. https://doi.org/10.1097/FTD.0b013e3181bb438e. PMID: 19935364; PMCID: PMC2784609.

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

Hamzah, F.B., Mohd Hamzah, F., Mohd Razali, S.F., Jaafar, O., Abdul Jamil, N., 2020. Imputation methods for recovering streamflow observation: a methodological review. Cogent Environ. Sci. 6 (1), 1745133.

Hernandez, A., Tango, P., Batiuk, R., 2020. Development of the multi-metric water quality indicator. Environ. Manage. Assess. 192, 94–110. https://doi.org/10.1007/s10661-019-7969-z.

Hirzel, A.H., Arlettaz, R., 2003. Modeling habitat suitability for complex species distributions by environmental-distance geometric mean. Environ. Manage. 32, 614–623.

Karr, J.R., 1981. Assessment of biotic integrity using fish communities. Fisheries 6, 21–27.

Kung, H., Ying, L., Liu, Y.C., 1992. A complementary tool to water quality index: fuzzy clustering Analysis. Water Resour. Bull. 28 (3), 525–533.

Langendorf, R.E., Lyubchich, V., Testa, J.M., Zhang, Q., 2021. Inferring controls on dissolved oxygen criterion attainment in the Chesapeake Bay. ACS ES&T Water 1, 1665–1675. https://doi.org/10.1021/acsestwater.0c00307.

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, Batiuk. R.A., 2018. Long term nutrient reductions lead to an unprecedented recovery in a temperature coastal region. PNAS, 115 (14) 3658-3662.

Leight, A.K., Crump, B.C., Hood, R.R., 2018. Assessment of fecal indicator bacteria and potential pathogen co-occurrence at a shellfish growing area. Front. Microbiol. 9, 384. https://doi.org/10.3389/fmicb.2018.00384.

Lipton, D., Hicks, R., 2003. The cost of stress: low dissolved oxygen and economic benefits of recreational striped bass (*Morone saxatilis*) fishing in the Patuxent River. Estuaries 26, 310–315. https://doi.org/10.1007/BF02695969.

Llanso, R.J., Dauer, D.M., Volstad, J.H., Scott, L.C., 2003. Application of the benthic index of biotic integrity to environmental monitoring in Chesapeake Bay. Environ. Monit. Assess. 81, 163–174.

Llanso, R.J., Dauer, D.M., Volstad, J.H., 2009. Assessing ecological integrity for impaired waters decisions in Chesapeake Bay. Mar. Pollut. Bull. 59 (1–3), 48–53.

Lu, R.S., Lo, S.L., Hu, J.Y., 1999. Analysis of reservoir water quality using fuzzy synthetic evaluation. Stoch. Env. Res. Risk Assess. 13, 327–336.

Markin, E.L., Secor, D.H., 2020. Growth of juvenile Atlantic Sturgeon (*Acipenser oxyrinchus oxyrinchus*) in response to dual-season spawning and latitudinal thermal regimes. U.S. National Marine Fisheries. Service Fish. Bull. 118, 74–86.

Moore, K.A., Wilcox, D.J., Orth, R.J., 2000. Analysis of the abundance of submersed aquatic vegetation communities in the Chesapeake Bay. Estuaries 23, 115–127.

Mujumdar, P.P., Sasidumar, K., 2002. A fuzzy risk approach for seasonal water quality management of a river system. Water Resour. Res. 38 (1), 1–9.

Mustapha, A., Aris, A.Z., Juahir, H., Ramli, M.F., Kura, N.U., 2013. River water quality assessment using environmetric techniques: case study of Jakara River Basin. Environ. Sci. Pollut. Res. 20, 5630–5644.

National Research Council 2007. Applications of toxicogenomic technologies to predictive toxicology and risk assessment, Appendix C: Overview of Risk Assessment. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10201/.

Orth, R.J., Williams, M.R., Marion, S.R., Wilcox, D.J., Carruthers, T.J.B., Moore, K.A., Kemp, W.M., Dennison, W.C., Rybicki, N., Bergstrom, P., Batiuk, R.A., 2010. Long term trends in submerged aquatic vegetation, (SAV) in Chesapeake Bay, USA, related to water quality. Estuaries Coasts 33, 1144–1163.

Ostad-Ali-Askari, K., Shayannejad, M., Ghorbanizadeh-Kharazi, H., 2017. Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran. KSCE J. Civil Eng. 21, 134–140. https://doi.org/10.1007/s12205-016-0572-8.

Scientific and Technical Advisory Committee -STAC 2006. The cumulative frequency diagram method for determining water quality attainment: report of the Chesapeake Bay Program STAC panel to review of Chesapeake Bay Program analytical tools. STAC Publication, 06-003. Chesapeake Bay Program Scientific and Technical Advisory Committee. Chesapeake Research Consortium, Edgewater, MD.

Smith, E.P., Zahran, A.R., Mahmoud, M.A., Ye, K., 2003. Evaluation of water quality using acceptance sampling by variables. Environmetrics 14, 373–386.

Smith, E.P., Ye, K., Hughes, C., Shabman, L., 2001. Statistical assessment of violations of water quality standards under Section 303(d) of the Clean Water Act. Environ. Sci. Technol. 35, 606–612.

Sun, Y., Wendi, D., Kim, D.E., et al., 2019. Deriving intensity–duration–frequency (IDF) curves using downscaled in situ rainfall assimilated with remote sensing data. Geoscience Lett. 6, 17. https://doi.org/10.1186/s40562-019-0147-x.

Tango, P.J., Batiuk, R.A., 2013. Deriving Chesapeake bay water quality standards. J. Am. Water Resour. Assoc. 1–18 https://doi.org/10.1111/jawr.12108.

Tango, P.J., Batiuk, R.A., 2016. Chesapeake Bay recovery and factors affecting trends: long-term monitoring, indicators, and insights. Reg. Stud. Marine Sci. 4 (2016), 12–20.

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

USEPA 2003a. Ambient water quality criteria for dissolved oxygen, water clarity and Chlorophyll-a for the Chesapeake Bay and its tidal tributaries. USEPA Region III Chesapeake Bay Program Office EPA 903-R-03-002, Annapolis, Maryland.

USEPA 2003b. Technical support document for identification of Chesapeake Bay designated uses and attainability. USEPA Region III Chesapeake Bay Program Office EPA 903-R-03-004, Annapolis, Maryland.

USEPA 2004. Chesapeake Bay Program Analytical Segmentation Scheme: Revisions, decisions, and rationales. 1983-2003. October 2004. USEPA Region III Chesapeake Bay Program Office EPA 903-R-04-008, Annapolis, Maryland.

USEPA 2005. Chesapeake Bay Program Analytical Segmentation Scheme: Revisions, decisions, and rationales. 1983-2003. 2005 Addendum, December 2005. USEPA Region III Chesapeake Bay Program Office EPA 903-R-05-004, Annapolis, Maryland.

USEPA 2007. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll-a for the Chesapeake Bay and Its Tidal Tributaries: Chlorophyll-a Addendum. USEPA Region III Chesapeake Bay Program Office EPA 903-R-07-005, Annapolis, Maryland.

USEPA 2008. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries – 2008 Technical Support for Criteria Assessment Protocols Addendum. September 2008. EPA 903-R-08-001. Region III Chesapeake Bay Program Office, Annapolis, MD.

USEPA. 2010. Chesapeake Bay total maximum daily load for nitrogen, phosphorus and sediment. U.S. Environmental Protection Agency, Region 3 Chesapeake Bay Program Office, Annapolis, MD.

USEPA 2017. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll-a for the Chesapeake Bay and Its Tidal Tributaries: 2017 Addendum. USEPA Region III Chesapeake Bay Program Office EPA 903-R-17-002, Annapolis, Maryland.

Vinod, H.D., López-de-Lacalle, J., 2009. Maximum entropy bootstrap for time series: the meboot R package. J. Stat. Softw. 29 (1), 1–19.

Weisberg, S.B., Ranasinghe, J.A., Dauer, D.M., Schaffner, L.C., Diaz, R.J., Frithsen, J.B., 1997. An estuarine benthic index of biotic integrity (B-IBI) for the Chesapeake Bay. Estuaries 20, 149–158.

Williams, M.R., Longstaff, B.J., Wicks, E.C., Carruthers, T.J.B., Florkowski, L.N., 2010. Ecological report cards: integrating indicators into report cards. Chapter 6. In: Longstaff, B.J., Carruthers, T.J.B., Dennison, W.C., Lookingbill, T.R., Hawkey, J.M., Thomas, J.E., Wicks, E.C., Woerner, J. (Eds.), Integrating and Applying Science: A Handbook for Effective Coastal Ecosystem Assessment. IAN Press, Cambridge, Maryland.

Zhang, Q., Murphy, R., Tian, R., Forsyth, M.K., Trentacoste, E.M., Keisman, J., Tango, P. J., 2018a. Chesapeake Bay's water quality condition has been recovering: insights from a multimetric indicator assessment of thirty years of tidal monitoring data. Sci. Total Environ. 1617–1625.

Zhang, Q, Tango, P.J., Murphy, R.R., Forsyth, M.K., Tian, R., Keisman, J., Trentacoste, E. M., 2018b. Chesapeake Bay dissolved oxygen criterion attainment deficit: three decades of temporal and spatial patterns. Front. Marine Sci., 21 https://doi.org/10.3389/fmars.2018.00422.