APPENDIX 1

SPECTRAL CASTING VALIDATION

PART 1. VALIDATION OF THE SPECTRAL CASTING METHOD FOR APPLICATION TO ASSESSING SHORT DURATION CHESAPEAKE BAY DISSOLVED OXYGEN WATER QUALITY CRITERIA ATTAINMENT

USEPA (2003) and USEPA (2007) criteria publications recommended the further development and eventual application of a statistical technique called spectral analysis for assessing short-duration dissolved oxygen criteria. However, as both publications stated, further validation of this tool was required in order to demonstrate its adequacy for DO criteria assessment. Thus evaluation and validation of spectral analysis techniques comprised an essential component of the Umbrella Criteria Assessment Team effort.

The updated and validated spectral analysis method described here is referred to as **spectral casting**. Spectral casting uses spectral analysis as an interpolation device to create a synthetic high frequency data set for monitored locations where high frequency data are not available (Perry Appendix 10, 11). The **sending site** is a monitoring location with high frequency data (e.g., 5-, 15-, 30-min intervals) which might be measured by an automated DO sensor (e.g. on a buoy). The **receiving site** is a monitoring location with low frequency data (e.g., 1 or 2 measures/month) such as one of the CBP long-term water quality monitoring fixed station network sites. Vertical profiles of high frequency synthetic data for the receiving site are formulated by combining the low frequency data variability signal from the receiving site with the high frequency data variability signal for the sending site (Box 1).

Box 1: Spectral Casting Definitions

Definitions related to Spectral Casting

(Elgin Perry 2010).

Umbrella Criterion: the most protective criterion. When compliance with one criterion insures compliance at others, the one (i.e. most protective) is termed "the umbrella" criterion.

Spectral Casting: In the early 1990's it was proposed (Neerchall, 1994) that spectral analysis could be used to create a synthetic high frequency data set at a monitoring location with only low frequency data. Because the technique involves transporting the high frequency signals in observations from one monitoring location to a nearby location, Perry proposes we call this technique 'Spectral Casting' - an analogy to casting with a spinning rod.

Sending Site: Spectral casting involves combining the high frequency data variability signal from one site with the low frequency data variability signal of a second monitoring site. The result is a synthetic high frequency data record for the low frequency-monitoring-only site. Because the high frequency signal is being transported from one site to another, Perry proposes the site that generates the high frequency signal the 'Sending Site'.

Receiving Site: Following up on the preceding definition, the low frequency site that receives the high frequency signal is called the 'Receiving Site'.

High/Low Frequency Criterion: High (Low) frequency criteria are criteria which require high (low) frequency data for water quality criteria assessment.

Description of Method:

Spectral Casting uses spectral analysis as an interpolation device to create a synthetic high frequency data set for locations where high frequency data are not available. The sending site is a location with high frequency data (e.g., 5-, 15-, 30-min intervals) which might be taken by an automated DO sensor on a buoy. The receiving site is a location with low frequency data (e.g., 1 or 2 x/month) which might be one of the fixed station monitoring locations of the CBP fixed station network. High Frequency synthetic data for the receiving site are formulated by combining the low frequency signal from the receiving location with the high frequency signal for the sending location. The computation begins by using a Fast Fourier Transform (FFT) to obtain a spectral decomposition at both locations (equations 1 and 2). High frequency terms are trimmed from equation 1 (frequency < 1 month) to leave a smooth function that interpolates the long term means (Figure 1) for the receiving site. The mean term and low frequency terms (frequency > 1 month) are trimmed from equation 2 to leave just short-term variation about zero (Figure 2) at the sending site. These two trimmed series are summed to form the synthetic data which track the long term mean for receiving site and reflect the short term variability of the sending site (Figure 3). The high frequency criteria for the receiving site are then assessed using the high frequency synthetic data.

$$x_t^r = \bar{x}^r + \sum_{k=1}^{rn/2} \left\{ a_k^{rt} \cos(2\pi f_k^r t) + b_k^r \sin(2\pi f_k^r t) \right\} \qquad \text{Eqn. 1}$$

where:

r is a superscript indicating receiving,

x_t^r are observations in the long term time series, t= 1,2, . . . rn

$\overline{x}^r$ is the mean over the rn observations

a_k^r, b_k^r are spectral coefficients estimated by the FFT

f_k^r are Fourier frequencies

$$x_t^s = \overline{x}^s + \sum_{k=1}^{sn/2} \left\{ a_k^s \cos(2\pi f_k^s t) + b_k^s \sin(2\pi f_k^s t) \right\} \qquad \text{Eqn. 2}$$

where terms are defined analogously to Eqn. 1 with s indicating sending site.

Umbrella Assessment Project

Various projects are underway to assess the Umbrella concept. As a rule, we are seeking criteria which require only low frequency data that will serve as an umbrella for criteria that require high frequency data. One method of assessment is to use spectral casting to generate synthetic high frequency data for the fixed station monitoring network. These synthetic high frequency data can then be processed through existing computer software to complete a CFD assessment of both the low frequency and high frequency criteria. A comparison of the assessment of the two criteria is a test of the umbrella concept.

Steps for the spectral casting project:

1. Develop a method for assigning sending sites to receiving sites.

2. Use results of Step 1 above to create synthetic data for receiving sites in test. If the sending site is observed at 15 minute intervals then a 3-year synthetic data set has 3x365x24x4= 105,120 observations. This takes 1-2 minutes per site x depth so for the 50+ stations in the main bay fixed station network, this will represent 8-10 hours of computation.

3. After populating the high frequency time series, these data are interpolated for spatial coverage. This will entail 105,120 interpolations which at 15 seconds per interpolation (the reported execution speed of the VB interpolator) will require 438 hours of computation and require an enormous amount of disk storage. The disk storage problem could likely be solved by summarizing for step 4. as we go.

4. The last step (which encapsulates all steps of CFD assessment) entails using the results of step 3 to compute the percent of space and percent of time criteria exceedances for the CFD assessment. A comparison of CFD results for high and low frequency criteria serve to test the umbrella concept.

This is a validation of the spectral casting method. Two near continuous data sets from the buoy data files are selected (Figure 1, Table 1).

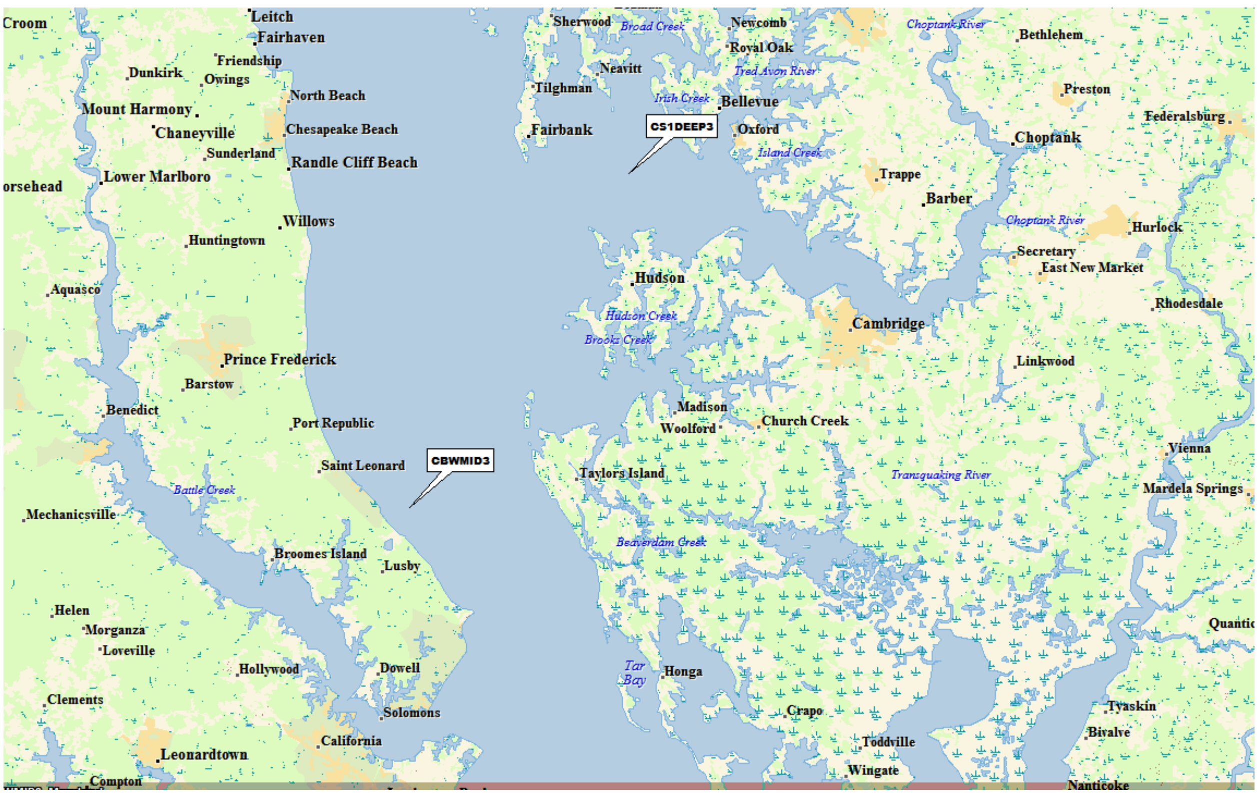


Figure 1. Location of send and receive sites.

Table 1. Properties of send and receive sites used in this test of spectral casting.

DataName	source	Lat	Long	Sensor Depth	Total Depth	Time Interval	Begin Date	End Date
CBWMID3	SANFORD	38.4517	76.4333	6	11	5	12-Aug-87	9-Sep-87
CE1DEEP3	SANFORD	38.5542	76.3967	19	23	15	12-Aug-87	9-Sep-87
CE2DEEP3	SANFORD	38.6458	76.3097	13.1	16	5	12-Aug-87	9-Sep-87
CS1DEEP3	SANFORD	38.6625	76.2567	6	8	12	12-Aug-87	7-Sep-87

CSDEEP3 is the sending site, CBWMID3 is the receiving site.

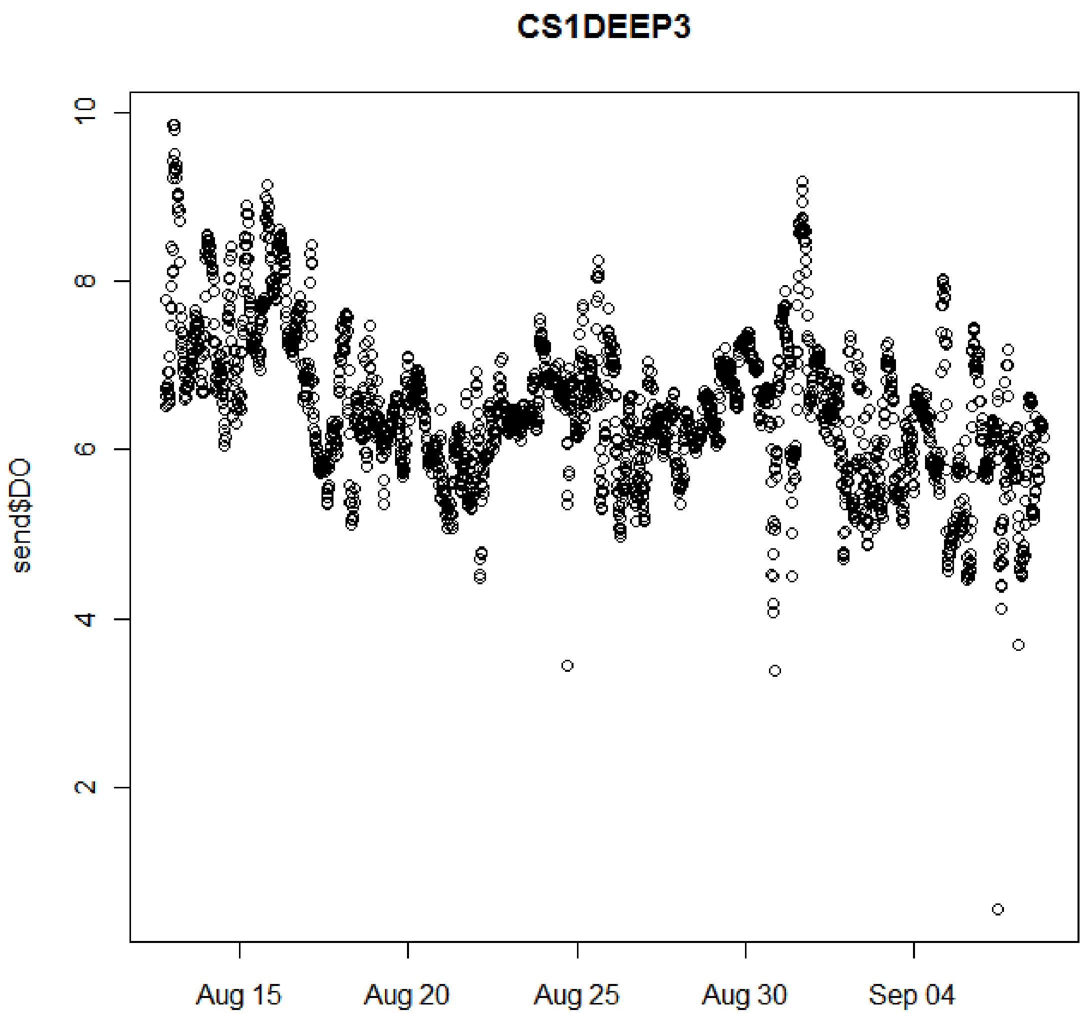


Figure 2. Time series of DO data for sending site.

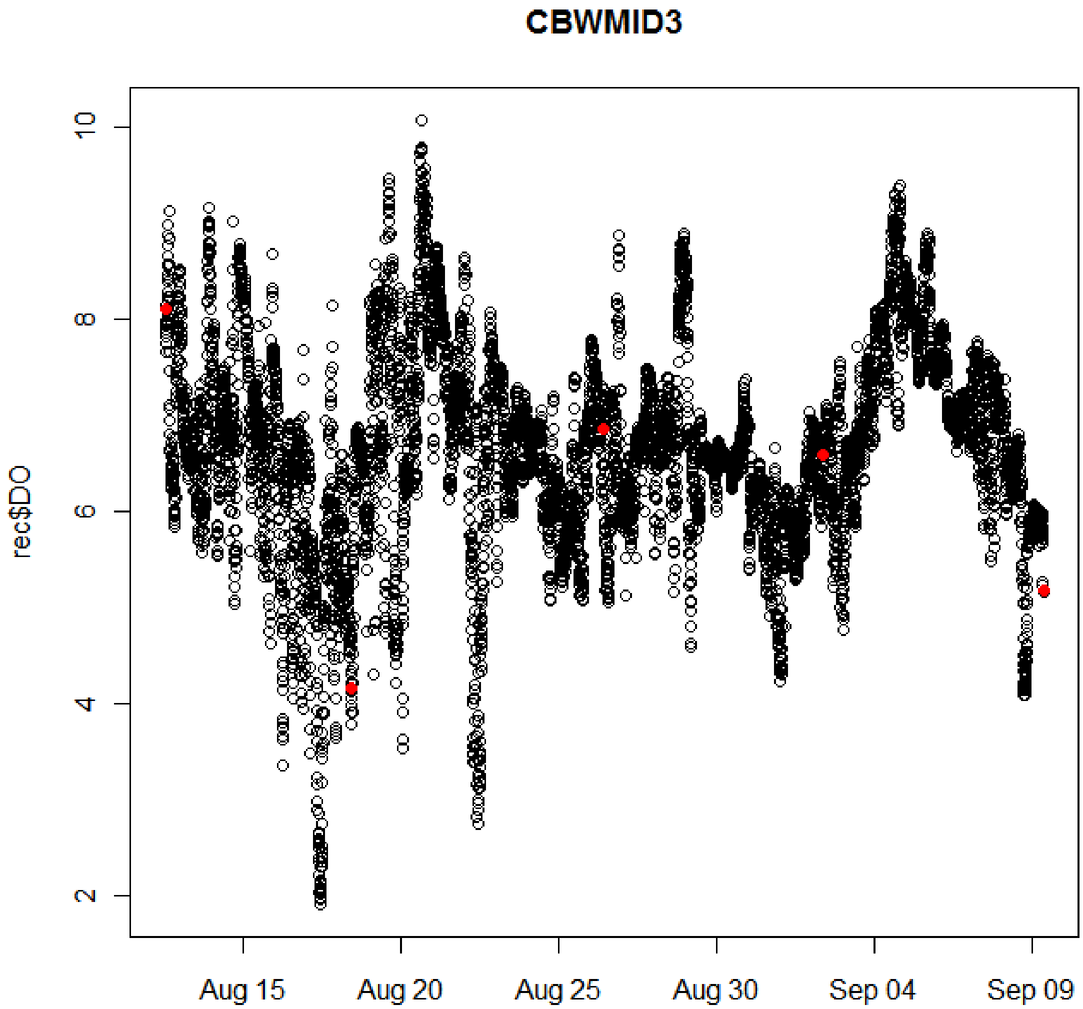


Figure 3. Time series of DO data for receiving site. Red dots are subsample for low frequency interpolation.

The low frequency sub-sample was chosen as the beginning and end + points once a week. The once a week points were chosen as day-time (morning) and about the same time of day.

1987-08-12 13:20:00

1987-08-18 10:20:00

1987-08-26 11:00:00

1987-09-02 09:50:00

1987-09-09 09:20:00

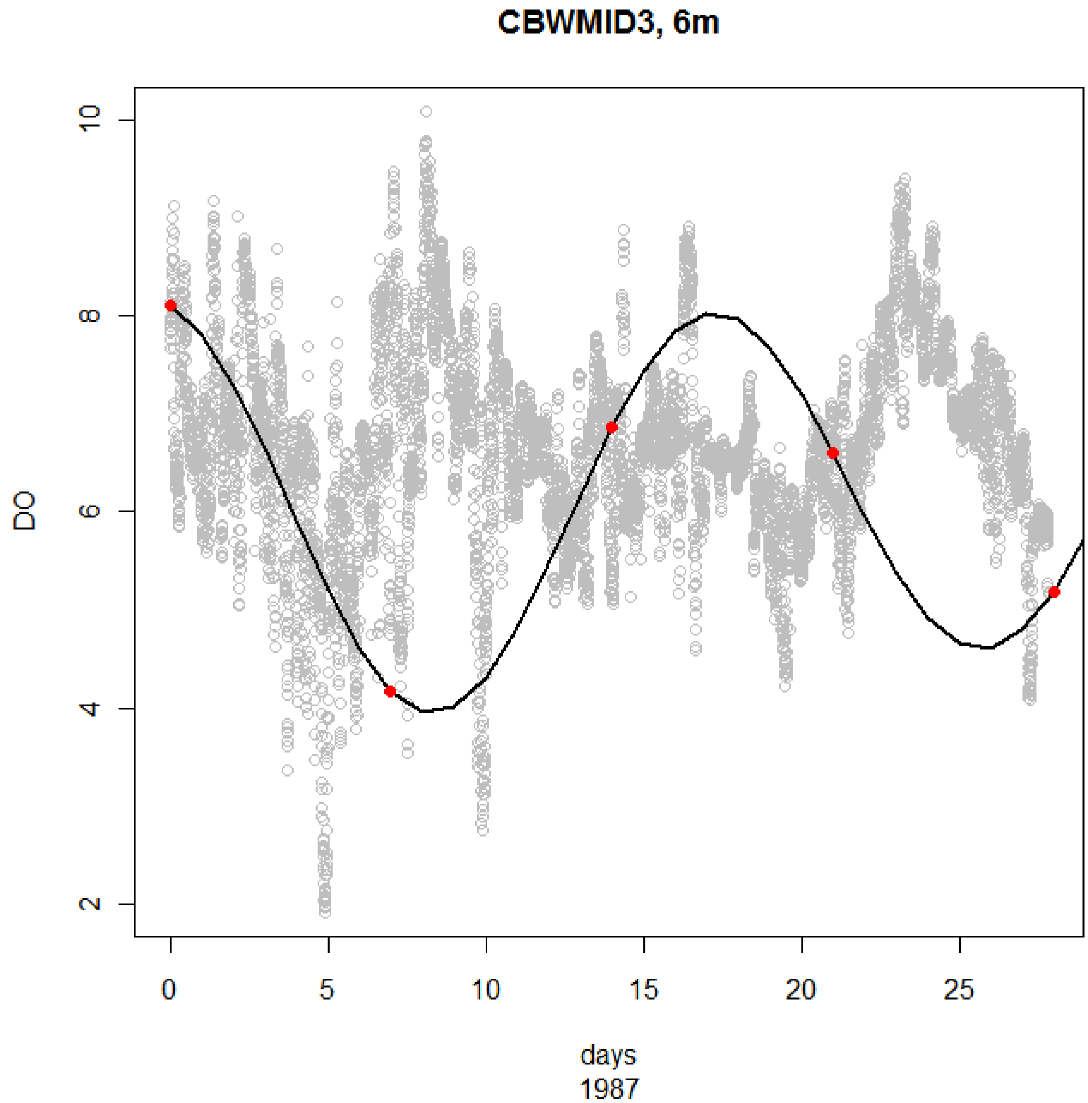


Figure 4. Time series of DO data for receiving site. Red dots are subsample for low frequency interpolation. Black line is interpolating function.

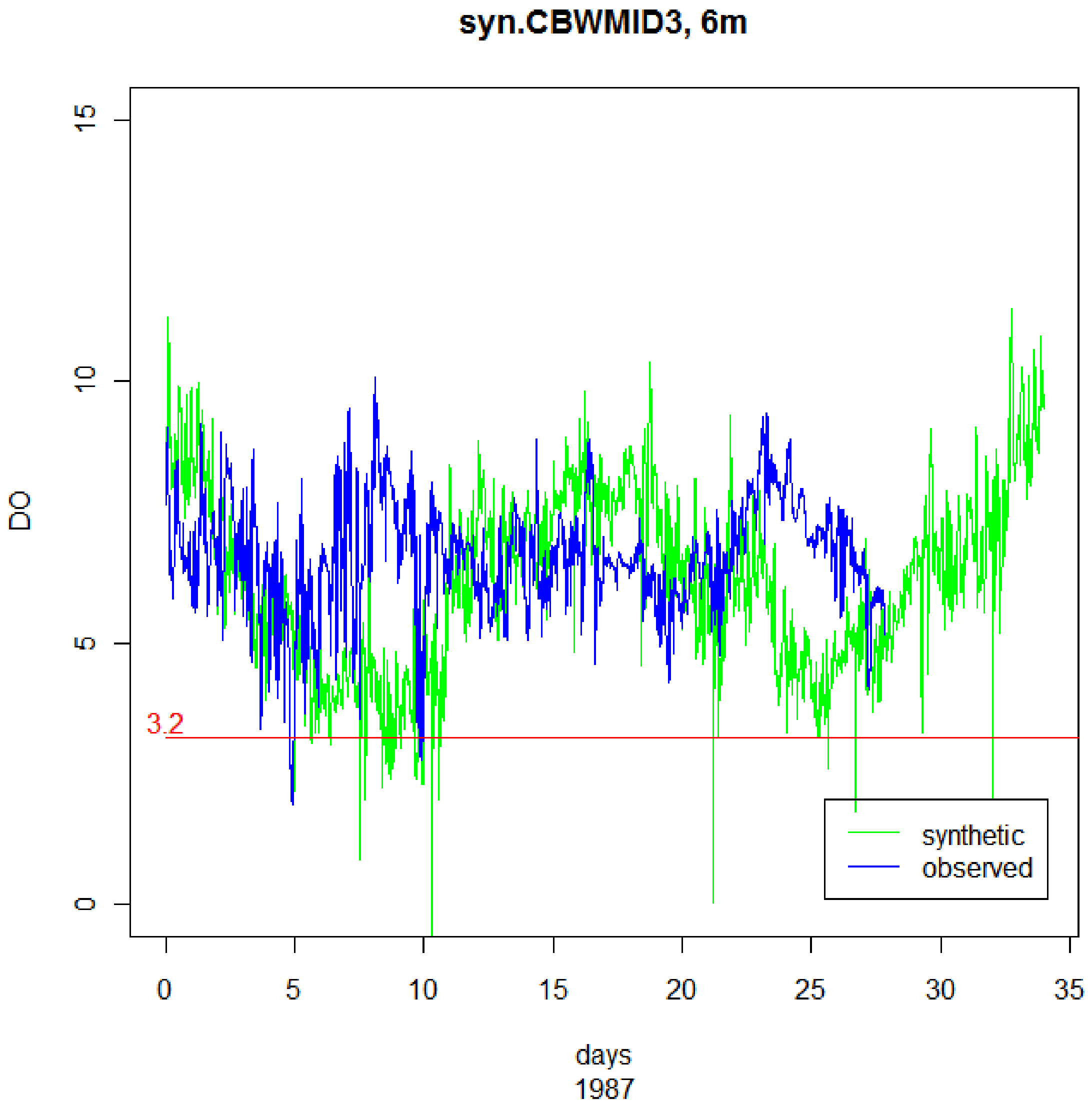


Figure 5. Time series of observed (blue) and synthetic (green) data at the receiving site.

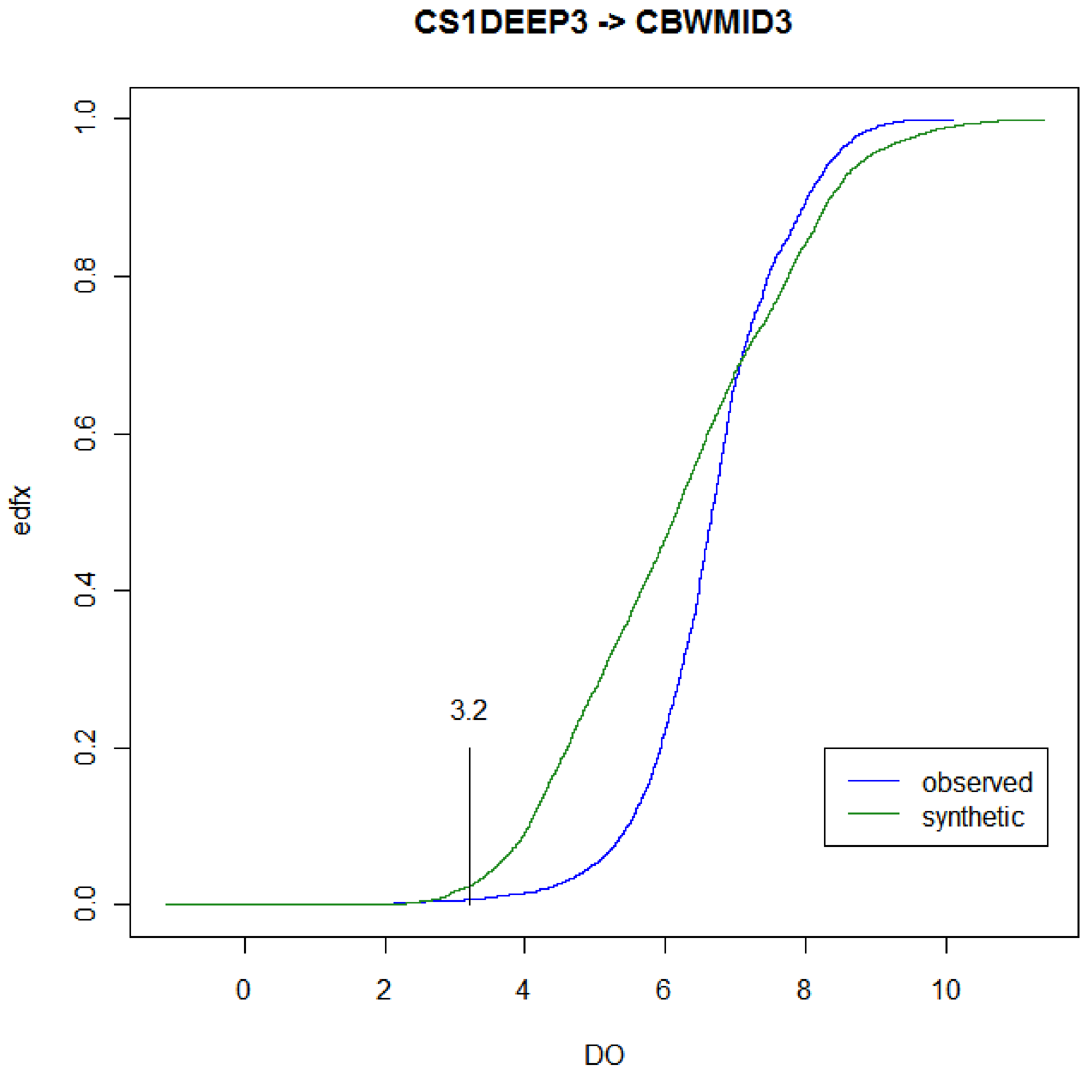


Figure 6. Empirical Cumulative Distribution Function for observed and synthetic data at receiving site.

- **Synthetic data show slightly greater violations of the Instantaneous Minimum criterion (3.2) than observed data.**
- **If the curves were shifted to the left (i.e. lower mean DO) the difference in violation rate could be substantial.**
- **Overall, the synthetic data have greater variability than the observed, but this greater variability did not originate at the sending site.**

var(rec$DO) = 1.089469,	**var(rec.syn$DO) = 2.787914**	**var(send$DO) = 0.7394112**
sd(rec$DO) = 1.043776	**sd(rec.syn$DO) = 1.669705**	**sd(send$DO) = 0.8598902**

PART 2. ASSESSMENT OF RELATIVE NOISE CONTRIBUTION OF LOW FREQUENCY SUBSAMPLING VERSUS HIGH FREQUENCY CASTING BASED ON SHALLOW WATER DATA WITH TESTS OF DIFFERENT INTERPOLATION APPROACHES. (From Elgin Perry 2/22/2011)

In my last discussion, I reviewed the relative contribution to uncertainty from two sources in the spectral casting process for data from the Southern Bay. In the first step of the spectral casting method a low frequency sample is interpolated in the time domain to estimate central tendency over time. In the second part of the process, short term - high frequency variability is borrowed from a sending site in an effort to fill in the extremes of variability around the estimated central tendency. Each step of the process will cause the percent of violations in the synthetic data to deviate from the percent of violations in the true DO time series. Using a long term - high frequency record near VIMS as a receiving site and casting from 10 different sites with 3 days of high frequency data from around the southern Bay, produced results that showed greater uncertainty due to the selection of the low frequency sample than due to the selection of casting site.

Here I present more results using this same approach on shallow water data.

Brief recap of the methods.

To assess which step contributes the greatest uncertainty, we perform a validation excercise. In the validation, we subsample a high frequency time series to create a low frequency subsample that is the receiving site of the spectral cast. The low frequency subsample is interpolated to as if it were low frequency time series from the fixed station data. Using Fourier analysis, the high frequency variation from a sending site is cast in and convoluted with the low frequency interpolation to form synthetic data. The validation step is to compare the percent violation in the synthetic data to the true percent violation in the original high frequency time series. Two variations on this validation exercise allow us to differentiate uncertainty due to low frequency sampling and uncertainty due to casting.

Step 1 is to examine the variability due to low frequency sample selection. To examine this, two ConMon sites are selected. One will serve as the receiving site and the other as the sending site. For one iteration of the spectral casting process, a low frequency (once every two weeks) sample is selected from the receiving site. This low frequency sample is interpolated using a Fourier series to estimate the central tendency of the synthetic data. Spectral Analysis is used to capture the high frequency signal from a subset of the sending site which is super-imposed on the central tendency to create the synthetic data. The synthetic data are compared to the observed high frequency data at the receiving site. This spectral casting process is repeated iteratively using a different low frequency subsample for each iteration and using the same high frequency signal. The variability among these iterations measures uncertainty due to the selection of the low frequency subsample.

Step 2 is to examine the variability due to choosing different sending data. In this exercise, we hold the low frequency subsample constant and study the variation contributed by different sending data. Unlike the assessment of buoy data in the southern bay where different sending sites were used to assess variability due to casting, for these shallow water data, a single sending site is used and data from this site are broken into two-week segments to create multiple sending data sets. Thus we hold the subset selected as the low frequency sample constant and use multiple temporal subsets from the sending site to created variability due to casting selection.

Results:

The two sites (yellow squares in Figure 1.) used for this exercise are Maryland ConMon Locations 'XBF7904' near St. George's Island and 'XCD5599' in Breton Bay (Figure 1).   Both locations were sampled at 15 minute intervals in the years 2006-2009.

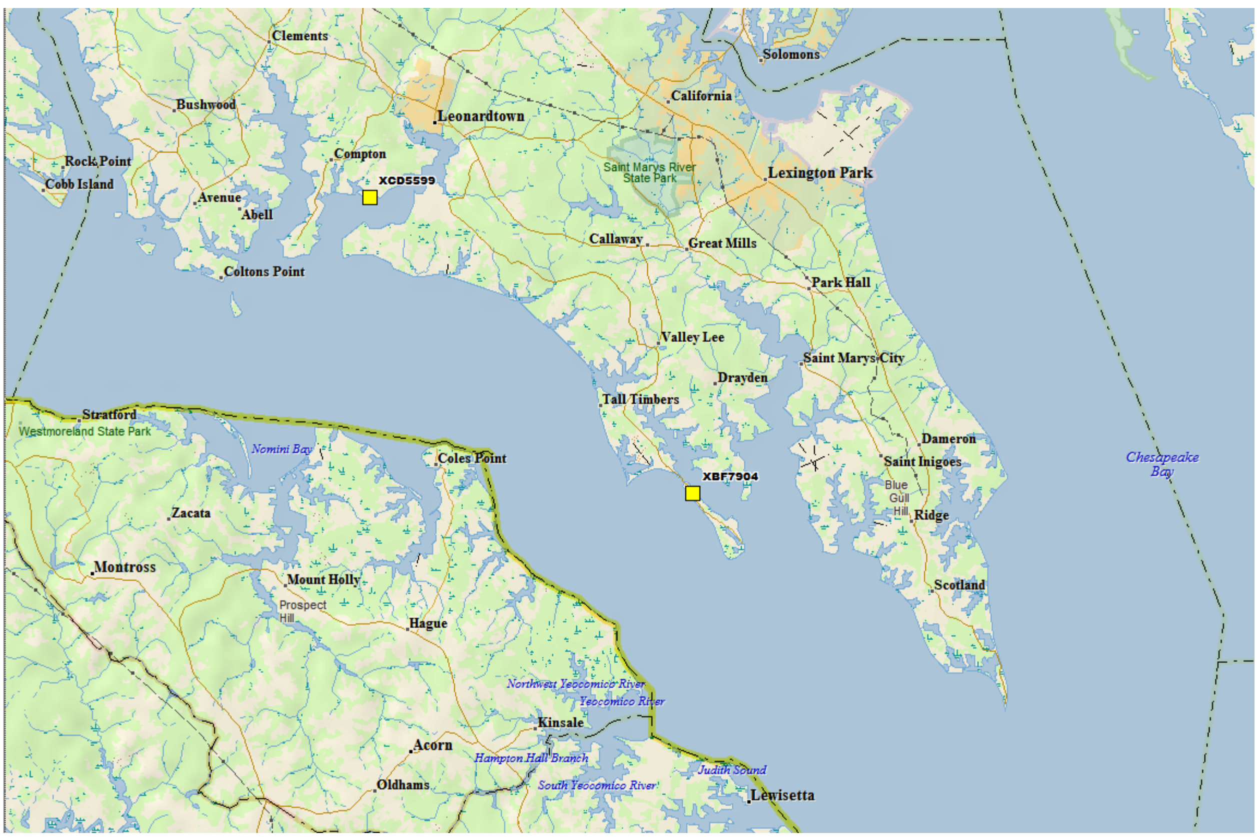


Figure 1. Location of two sites used to compare uncertainty from low frequency sampling vs. uncertainty from casting.

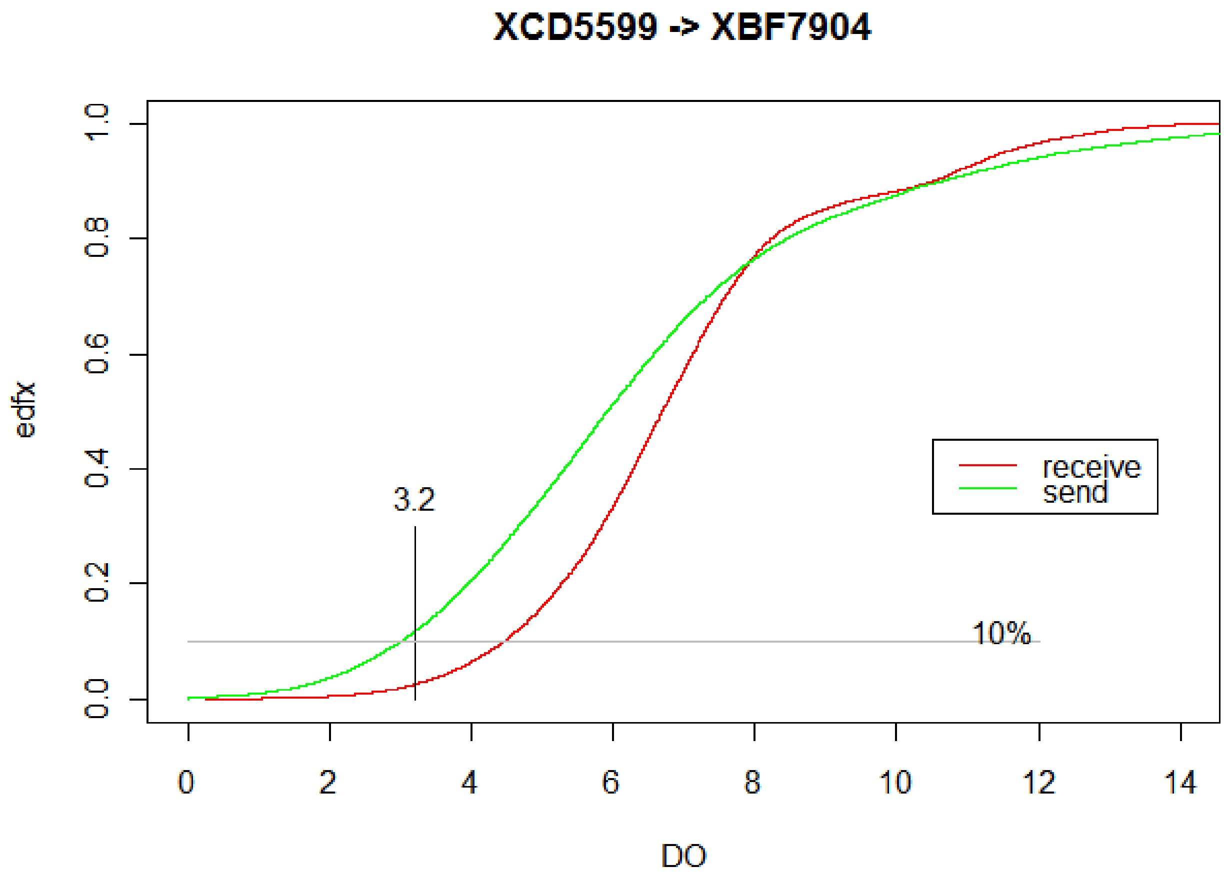


Figure 2. Cumulative empirical distribution functions for sending and receiving sites.

Over the three year period, the sending site has a greater proportion of violations of the instantaneous minimum criterion than the receiving site (Figure 2.). In addition, the sending site show greater variability in DO than does the receiving site. The greater variability at the sending site is mostly in the form of more low values than at the recieving site.

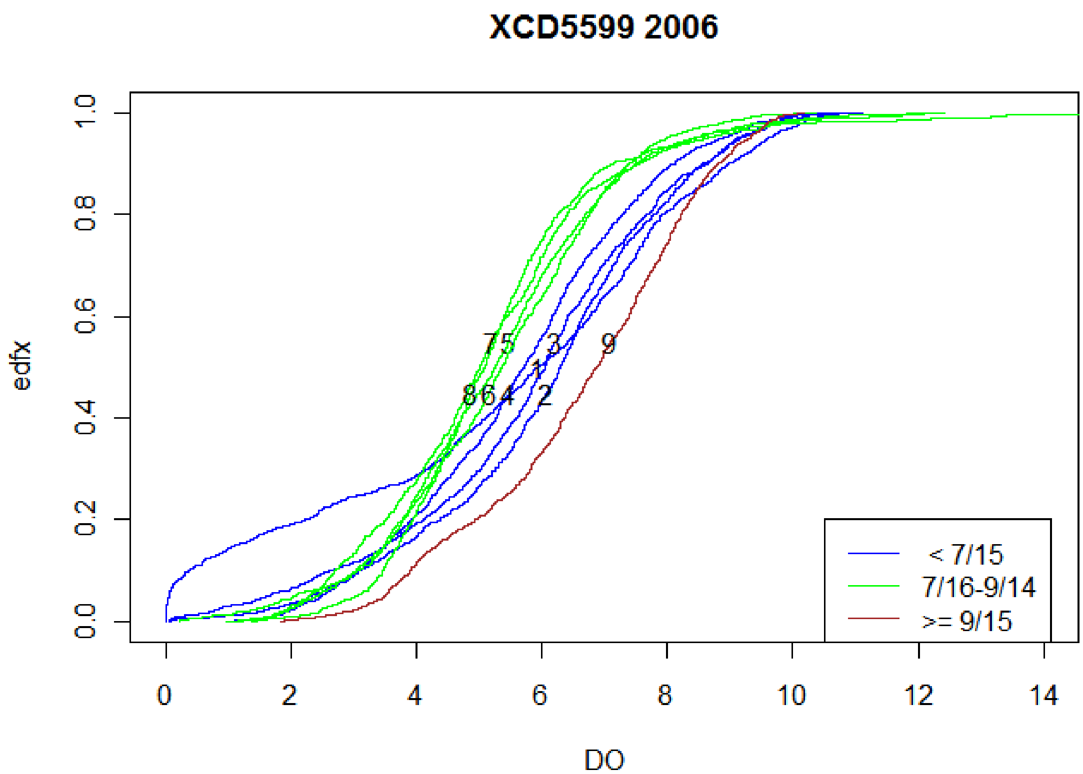


Figure 3. The cumulative empirical distribution function of raw data for each sending data set for 2006.

To create multiple sending data sets, for each of 3 years of data, the summer time series was broken into two week intervals. To get so concept of the variability among these different sending data sets, the ECDF for each one is plotted both in raw form (Figure 3.) and after the long term signal is removed leaving only the high frequency variability (Figure 4.). The 2006 data show a seasonal pattern of DO being relatively high through mid July (blue curves), low in late July and August (green curves) and returning to relatively high in fall (brown curve).

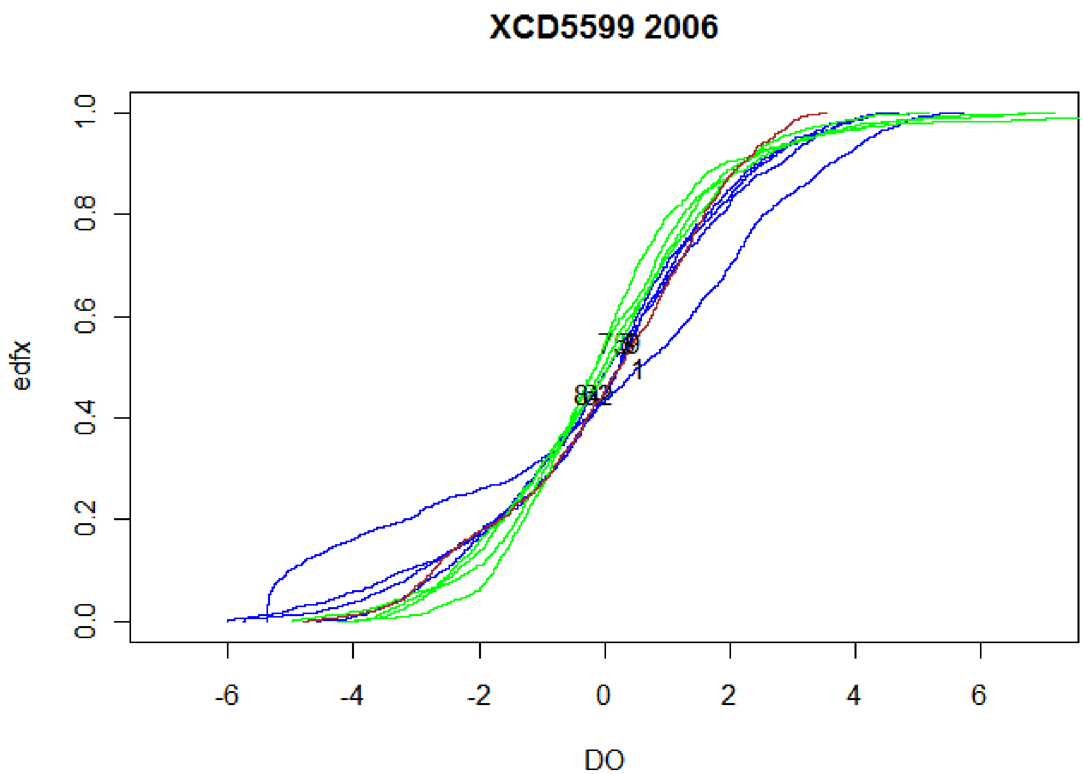


Figure 4. The cumulative empirical distribution function of data adjusted for central tendency leaving just high frequency variability for each sending data set for 2006.

The short-term variability for each two-week period (Figure 4.) shows a much tighter cluster than the raw data. One curve (labeled 1 in Figures 3 and 4) show a marked deviation from the remainder. It is the first 2-week period of the 2006 time series and has a mix of very high and very low DO values. Looking at a time series plot (Figure 5.) for this two-week period reveals a period low DO that occurred early in the season. While the data are unsusual, it does appear to be a valid data record.

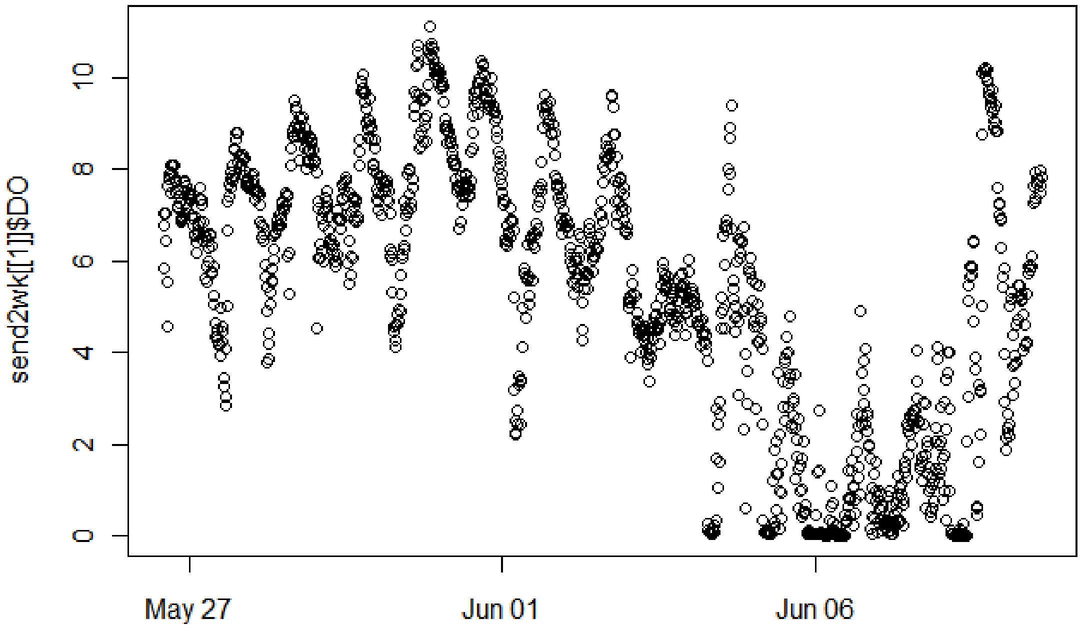


Figure 5. Time series plot of DO from the 1st two-week period in 2006 at the sending site.

This exercise of comparing the variability among cedf's for raw data and cedf's for centered data is repeated for the years 2007 (Figure 6.) and 2008 (Figure 7.).

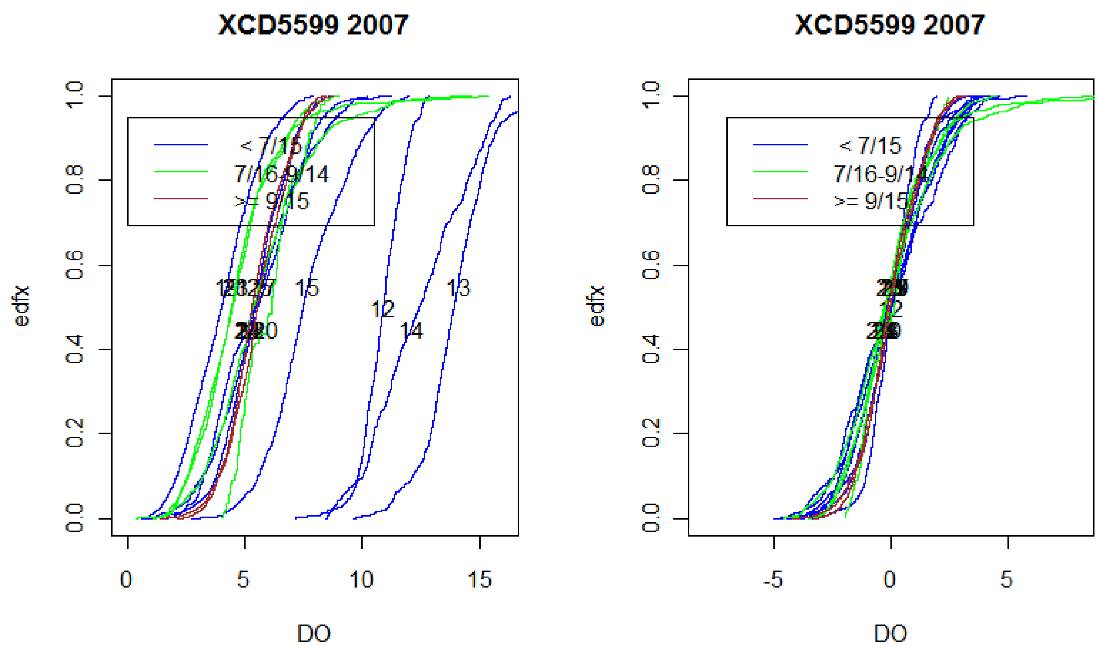


Figure 6. Comparison raw data (left) and centered data (right) for 2007.

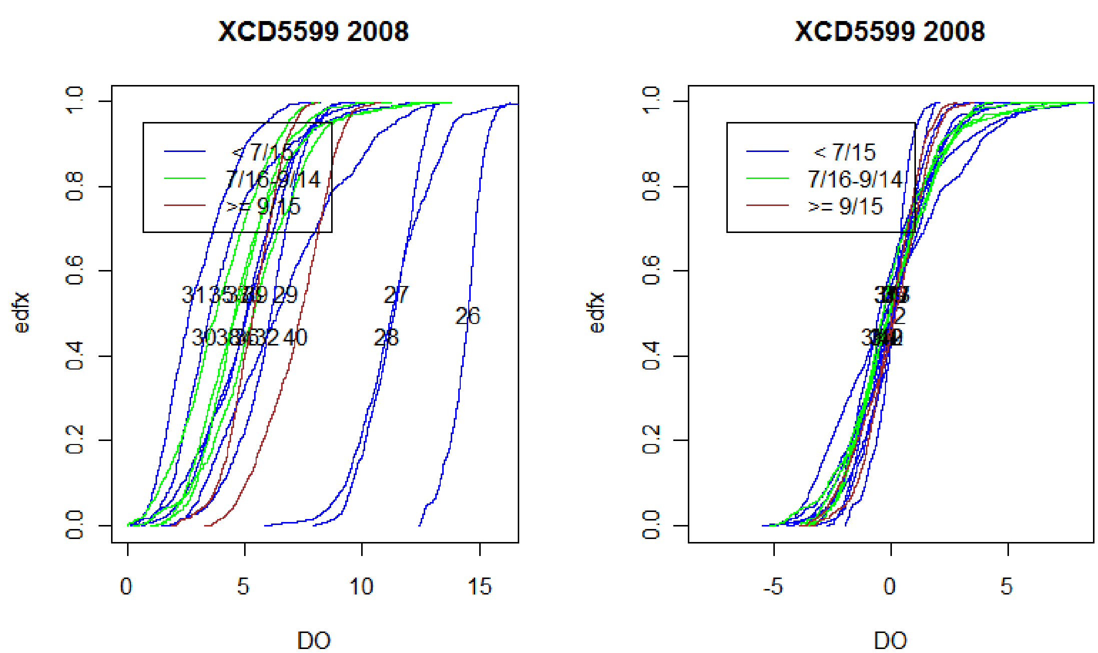


Figure 7. Comparison raw data (left) and centered data (right) for 2008.

The seasonal pattern in the raw data differs among the three years, but the centered data shows a fairly tight cluster in each case.

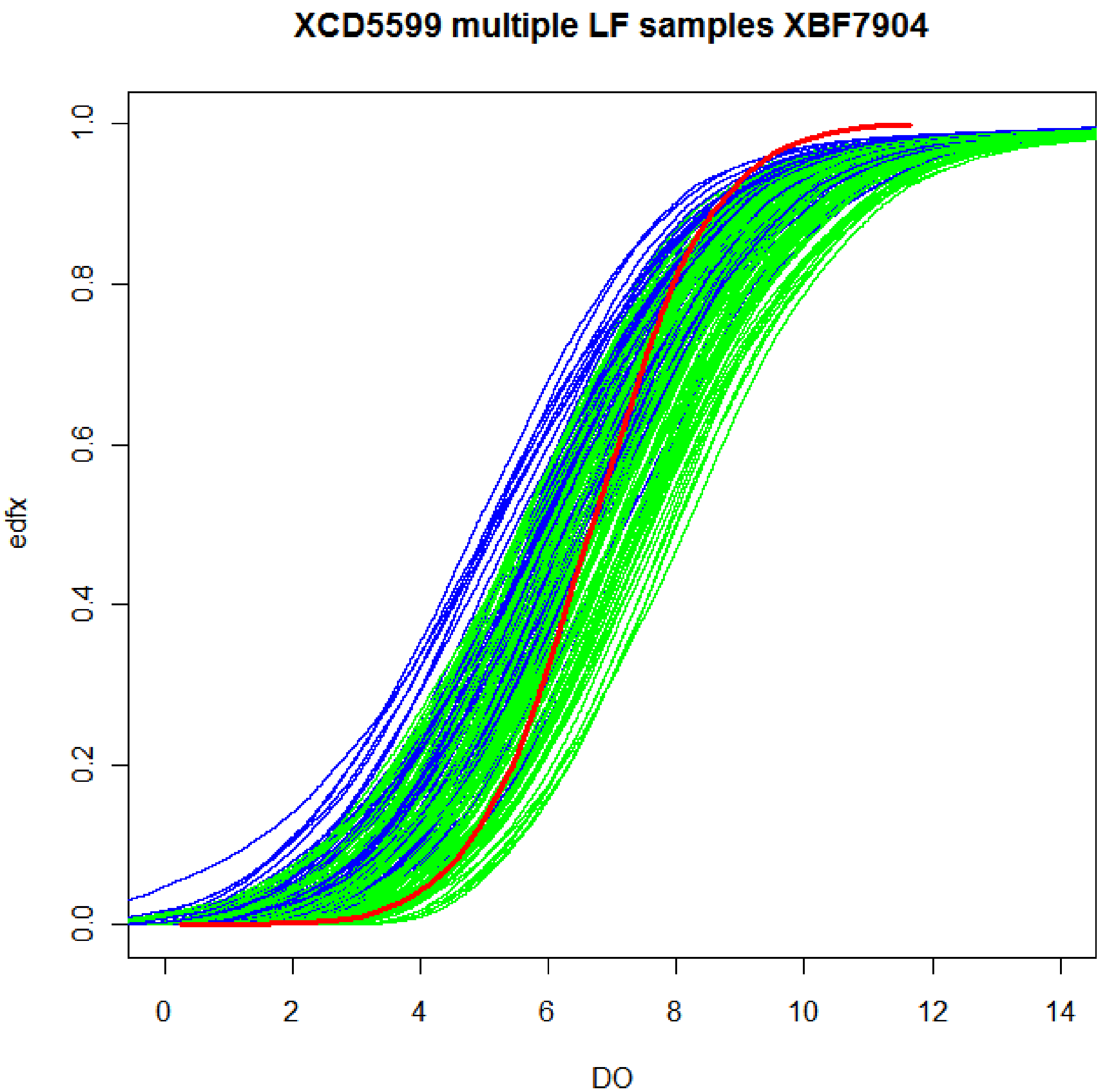


Figure 8. Variation due to multiple low-frequency samples from the receiving site with Fourier Series interpolation. The sending data set is held constant at one two-week interval. Blue curves synthetic data based on a series of night samples. Green curves are from a series of day samples. The red curve is the receiving site high frequency data.

The results of step 1, holding the sending data constant and varying the low frequency subsample (Figure 8.) show considerable variability. In general the synthetic data show greater failure of the instantaneous minimum relative to the observed data. However, it is clear that by collecting low frequency data only during the day is possible to obtain synthetic data that are anti-conservative.

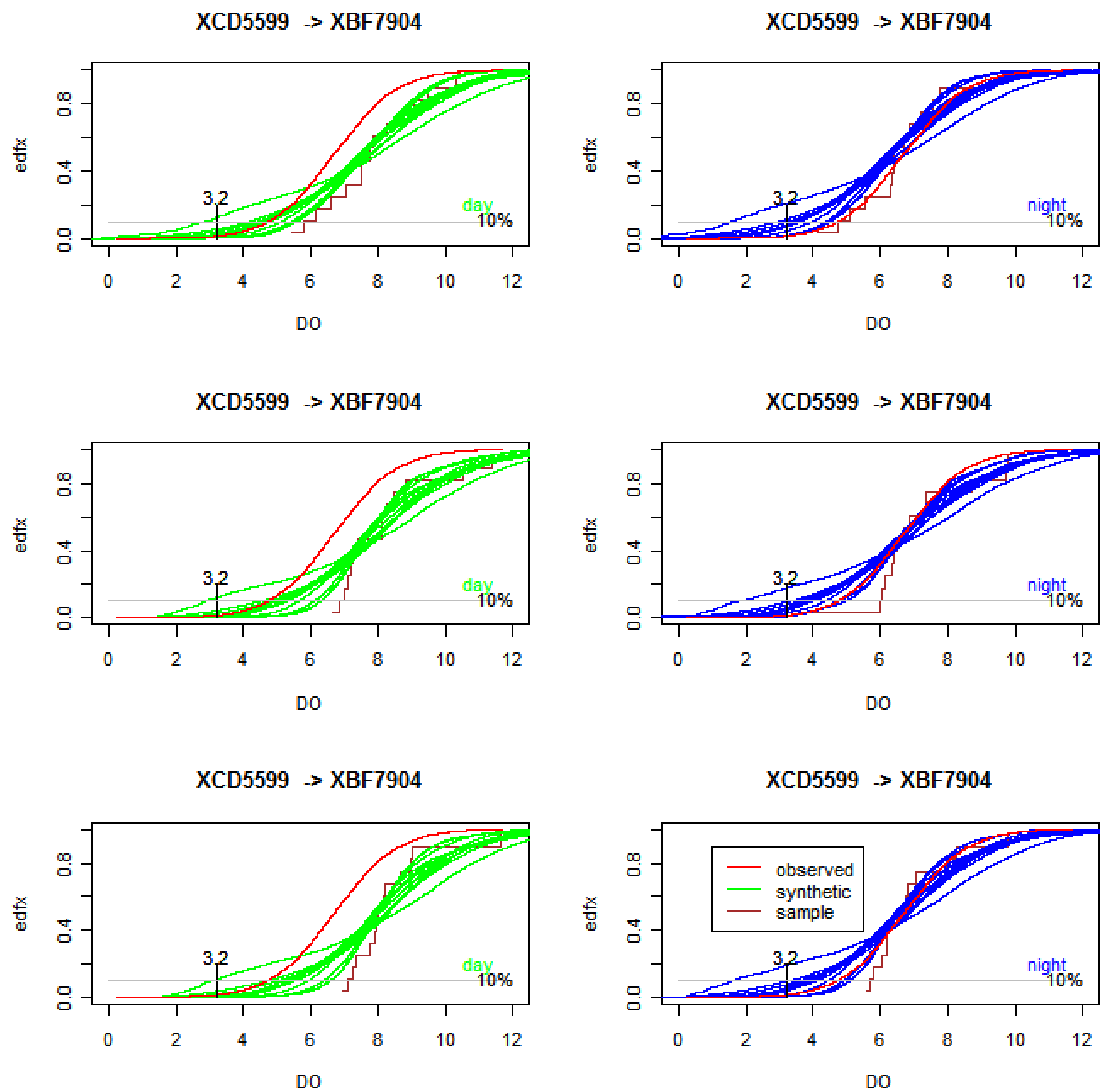


Figure 9. In each panel are the original receiving high frequency data (red) , the low-frequency sample (brown) and 12 synthetic data sets (blue or green) base on 12 sending data sets. Green curves use low-frequency data collected during the day. Blue curves use low frequency data collected at night. Each panel is for a different set of low-frequency data.

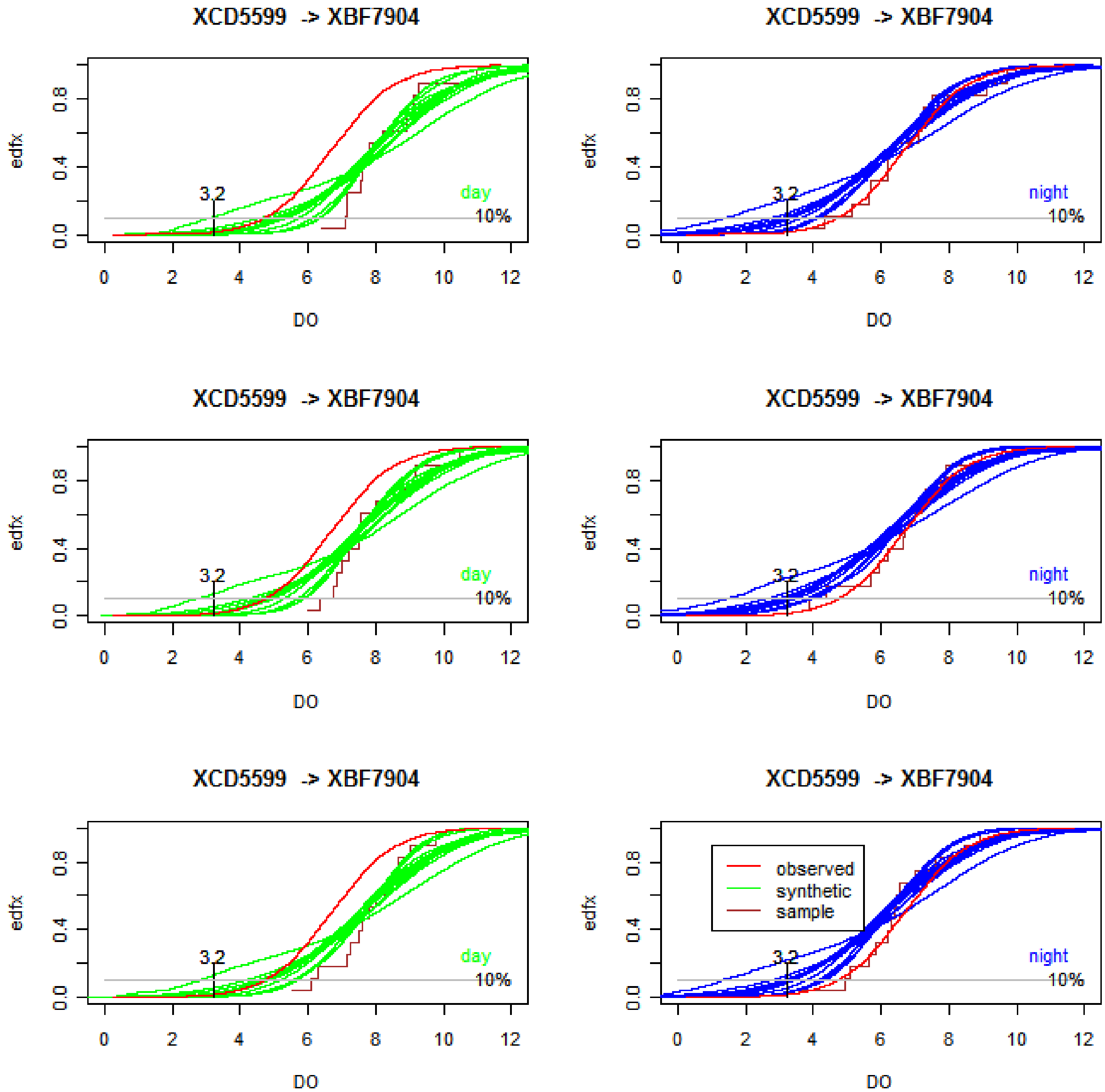


Figure 10. This is a continuation of results in Figure 9.

Each panel of this figure shows 12 synthetic data sets (green or blue curves) based on 12 sending data sets superimposed on 1 low frequency sample (brown step curve) as compared to the original high frequency receiving data (red). Moving from panel to panel changes the low frequency sample (brown) that was drawn from the high frequency data (red). It is clear that daytime sampling leads to a positive bias in the estimated status of DO while night sampling results in negative bias. It is also clear that deviations among the 12 Castings tend to be less than the deviations of the low frequency sample from the true CDF.

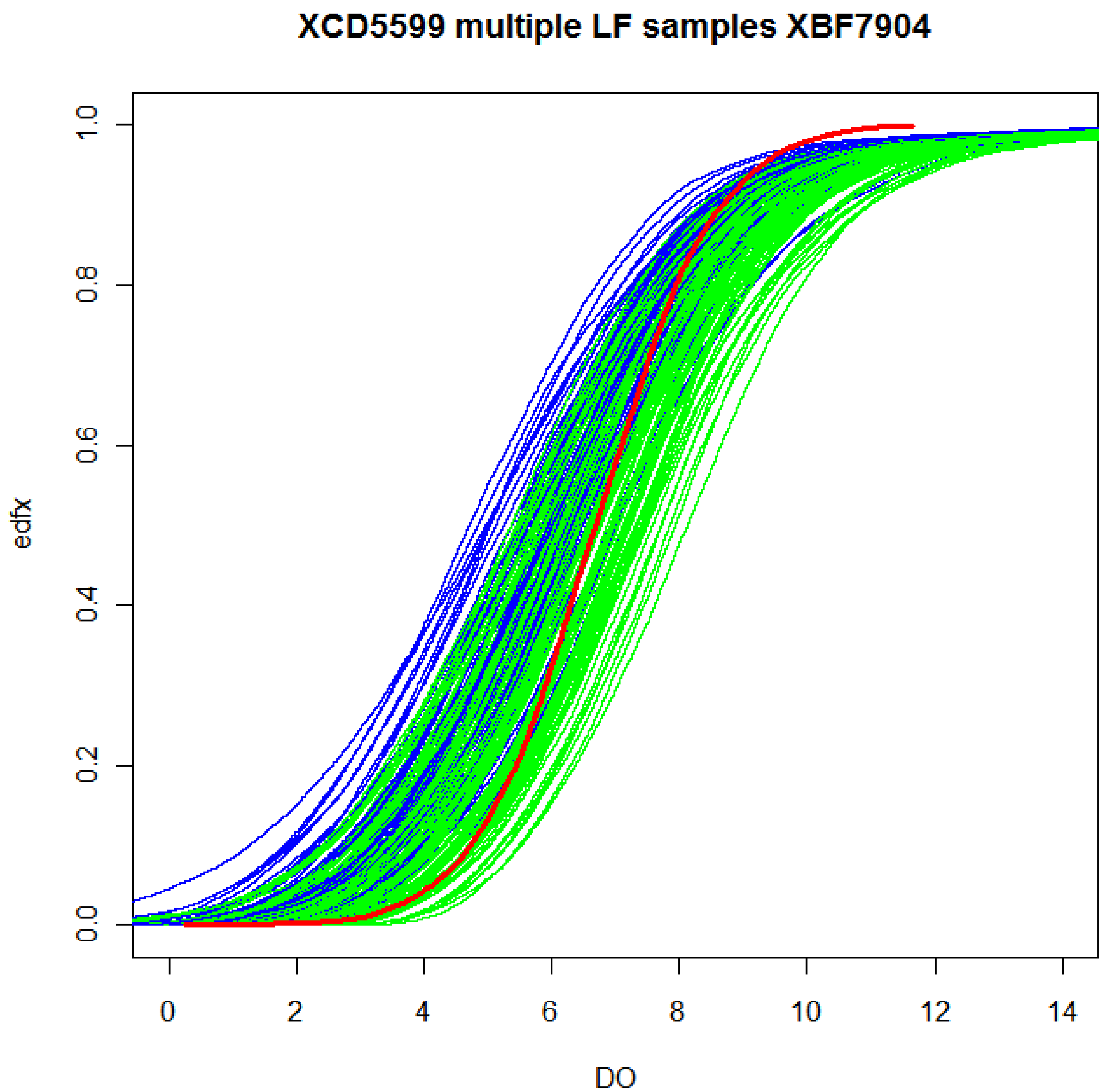


Figure 11. This figure shows the variation due to different low-frequency samples with Cubic Spline interpolation. The sending data set is held constant at one two-week interval.

Comparing Cubic Spline interpolation (Figure 11.) with Fourier Series interpolation (Figure 8.) for the low-frequency component of the synthetic data appears to make little difference.

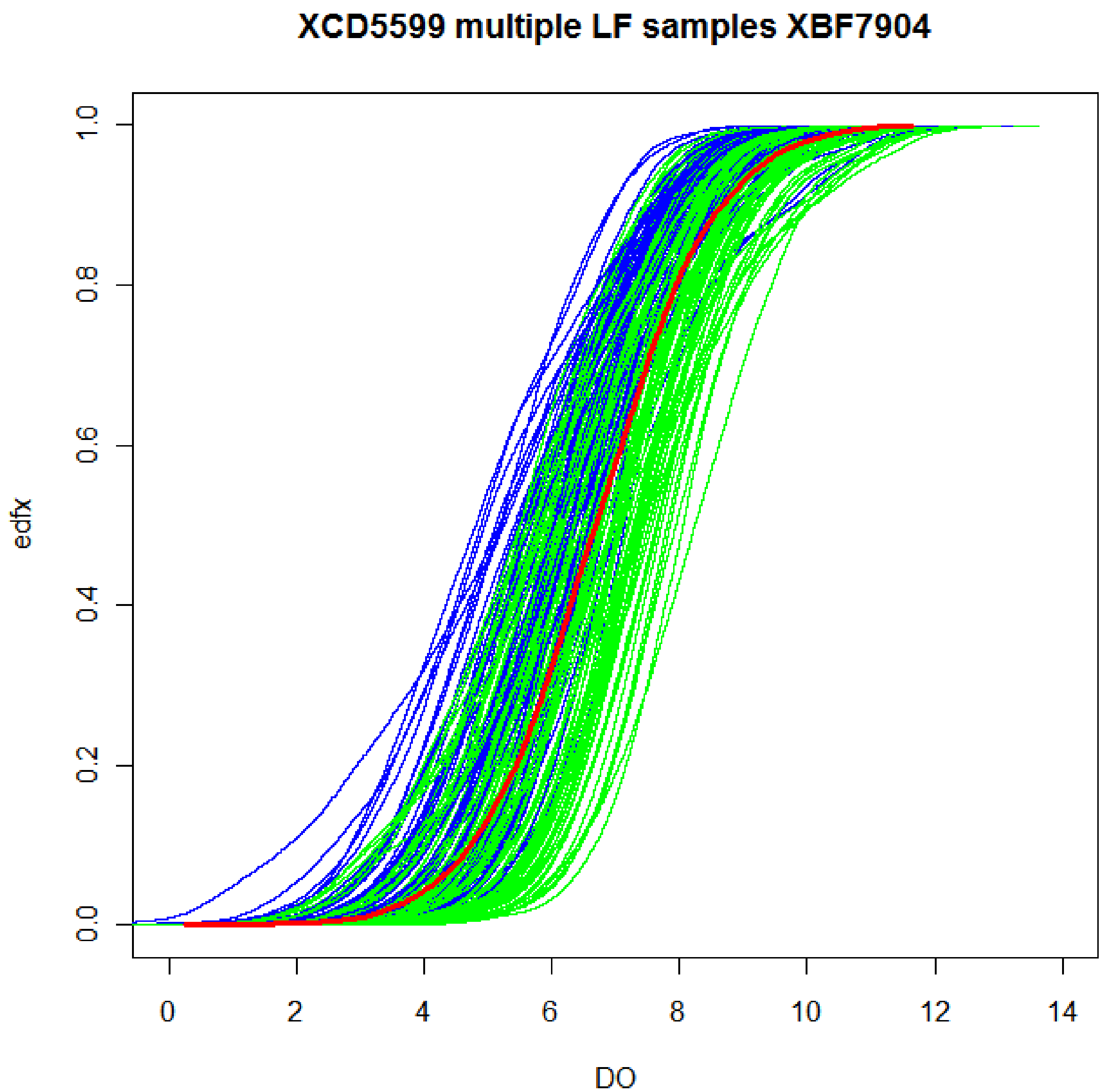


Figure 12. This figure shows the variation due to different low-frequency samples with Simple Linear interpolation. The sending data set is held constant at one two-week interval.

Comparing linear interpolation (Figure 12.) to Cubic Spline interpolation (Figure 11.) or Fourier Series interpolation (Figure 8.), it appears that linear interpolation yields a better fit.

APPENDIX 2

DISSOLVED OXYGEN (DO) CRITERIA ATTAINMENT ANALYSIS FOR SHALLOW WATER HABITATS USING CONTINUOUS MONITORING DATA SETS.

[UMCES Technical Series No. TS-606-10-CBL]. Boynton, W.R., E.M. Bailey, L.A. Wainger and A.F. Drohan. 2010. In: Ecosystem Processes Component (EPC). Maryland Chesapeake Bay Water Quality Monitoring Program, Level 1 report No. 27. Jul. 1984 - Dec. 2009. Ref. No. [UMCES] CBL 10-098.

Part 1. Dissolved Oxygen (DO) Criteria Attainment Analysis for Shallow Water Habitats Using ConMon Data Sets

W.R. Boynton, E.M. Bailey, L.A. Wainger and A.F. Drohan

Introduction

Until the last decade, water quality monitoring in Chesapeake Bay and tributary rivers was largely based on monthly or bi-monthly sampling at fixed stations located over the deeper (channel) portions of these systems. Such a design had many benefits, especially those related to developing seasonal to inter-annual scale indices of water quality status and trends. However, as in virtually all environmental science activities, a single measurement scheme is not adequate for addressing all questions. Thus, about a decade ago, a new program was initiated, first on a pilot-scale basis, to add measurements of water quality for shallow near-shore habitats. Concern for SAV habitat quality was a prime consideration in developing this program.

The program was named ConMon to indicate the near-Continuous Monitoring feature of this activity. The program used in-situ sensor systems (YSI Sondes) programmed to take measurements of a suite of water quality variables every 15 minutes. Included in the water quality suite was water temperature, salinity, pH, DO, turbidity and chlorophyll-a. In most instances ConMon sites were active from April – October (the SAV growing season) and in most cases sites remained active for three years. In a few cases, sites have remained active for up to 9 years, thus serving as long-term or sentinel sites. To place this sampling intensity in perspective, at a typical main channel site about 16 measurements of water quality variables were collected per year. In contrast, at a ConMon site about 20,500 measurements per year are obtained, an intensity of measurement about three orders of magnitude higher than traditional monitoring.

There have been about 60 sites in the Maryland Bay and Maryland Coastal Bays where ConMon data have been collected. The program is continuing although at somewhat fewer sites than in the recent past. The considerable spatial extent (encompassing sites with water quality varying from quite good to very poor) of these data sets allows for comparative analyses wherein it is likely that relationships between near-shore water quality and management actions can be found.

There are several prime uses of ConMon data sets. First, they have been used as a guide in selecting and monitoring SAV habitat restoration sites. Second, these data have "opened our eyes" to a new scale of hypoxia, namely diel-scale hypoxia wherein DO concentrations can reach critically low levels at night (and especially in the immediate post-dawn hours). Third, these data can be used to make estimates of community production and respiration, both of which are fundamental ecosystem features known to be related to nutrient loading rates. Fourth, these data can be used in DO criteria assessments for shallow open water sites (EPA 2010).

It is the last ConMon use that is the focus of this chapter. In an earlier portion of this report the strategy and details of making DO criteria assessments using ConMon data have been described. In this section we provide examples of DO criteria % non-attainment for three sites in the Bay system. It remains unclear as to which of several approaches best captures meaningful DO non-attainment; we present results of all approached in this section. There is a STAC-sponsored DO criteria workshop scheduled for the fall of 2010 and we will participate in this workshop and hopefully arrive at a consensus approach for computing this metric.

Methods

Continuous monitoring data was obtained from Maryland Department of Natural Resources Tidewater Ecosystems Assessment division in electronic (.txt file) format (dnr_cmon_sonde_2001-08). This file contained all the collected ConMon data from 2001 to 2008. A SAS® (www.sas.com) program was written to allow selection of dissolved oxygen data by station and year. The program then calculated 6 different methods/averages (Table 4-1) and gave each occurrence of dissolved oxygen (instantaneous or averaged) a score of 1 if lower than the criteria and a score of 0 if equal to or above (based on Chesapeake Bay Program guidelines and discussions with MDDNR and TWMAW). Criteria were chosen prior to selecting specific stations and we selected the higher DO value to make this analysis more "conservative."

Table 1. Calculation methods, file names, descriptions and criteria used for DO criteria % non-attainment calculations.

Calculation Method	SAS Filename	Description	DO Criteria
Instantaneous	doyyyST_allcrit	Uses every available data point (~every 15 minutes per annual data set).	4 mg L^{-1}
Daily Mean	doyyyST_daycrit	Takes the mean DO for all measurements over the course of 24 hours. No data point is reused in the calculation.	4 mg L^{-1}
7 Day Moving Average	doyyyyST_wkcrit	Takes the mean DO for a 7 day chunk of data moving forward 15 minutes for each iteration. Measurements are reused.	4 mg L^{-1}
1 Average per 7 Days	doyyyyST_1perwk	Takes a mean DO for all measurements over the course of 7 days. No data point is reused in the calculation.	4 mg L^{-1}
30 Day Moving Average	doyyyyST_moncrit	Takes the mean DO for a 30 day chunk of data moving forward 15 minutes for each iteration. Measurements are reused.	5 mg L^{-1}
1 Average per 30 Days	doyyyyST_1pmo	Takes a mean DO for all measurements over the course of 30 days. No data point is reused in the calculation.	5 mg L^{-1}

Exact criteria values will be refined in FY2011 in consultation with MDDNR for each specific station and month. SAS output files were named DO(dissolved oxygen), yyyy (year), ST (two-letter station code), underscore followed by an identifier for the calculation method used. Percent non-attainment was calculated as: (sum of the non-attainment score)/(total # of observations) * 100. Percent non-attainment was calculated for the entire available annual dataset, June-August and July.

Results from Selected Sites

Estimates of DO % non-attainment have been developed for three sites in the Bay system. The first site was St George’s Island (XBF7904), located in a small embayment of the lower portion of the Potomac River estuary. This site was chosen for initial analysis because water quality at this site is relatively good compared to many other Maryland tributary sites. Water quality here was good enough for this site to be selected for SAV restoration work. ConMon data are available for this site for the years 2006-2008. The second site selected was Sycamore Point (XHH3851), located in the upper portion of the Corsica River estuary. Multi-year monitoring of this site indicates poor to very poor water quality and there are indications from Dataflow mapping that some water quality conditions have been deteriorating further in recent years. Data for the years 2005-2008 were available for this analysis. The third site was the Fort McHenry site

(XIE5748) located in the Patapsco River estuary, adjacent to the city of Baltimore, MD. This site was selected because it is an urban site and because there is a considerable ConMon record available from this site.

Low Impact Site (St. George's Island, Lower Potomac River: XBF7904)

Results of DO % non-attainment are summarized for the St George's Island site (2006-2008) in Table 4-1 and Figures 4-1 to 4-3. For each year, 6 different averaging schemes were employed; these have been described earlier in this chapter. Across the top of Table 4-2 a simple average DO concentration was calculated for three time periods, including: 1) the whole year (Jan-Dec); 2) summer period (Jun – Aug) ; and 3) just the month of July. Further to the right in Table 1 DO % non-attainments were calculated for each time period using all 6 averaging schemes. Several patterns are readily evident.

First, % non-attainment consistently increases with smaller time period evaluations. For example, during 2006, the "All Data" computation indicated 4% non-attainment for the whole year evaluation, 8% for the summer evaluation and 10% for the July evaluation. At this site, the July evaluation for all % non-attainment approaches was the highest and this was also true for all three years evaluated. It is interesting to note that hypoxia/anoxia in the mainstem Bay reaches a maximum in July of most years since the monitoring program began in 1985. It may be that the single most critical water quality month is July in most years. Further analysis will clarify this issue.

Second, it is not completely clear which of the averaging techniques provides the most sensitive metric of DO non-attainment. For data collected during 2006 and 2007 it appears that the "All Data" approach detected more non-attainments than any other approach (i.e., it was the most protective). However, during 2008 the same pattern did not emerge. In fact, some counter-intuitive results emerged. The highest July % non-attainment emerged from the 30 day moving average approach, a considerably larger % non-attainment than that obtained from all other approaches, including the "All Data" approach. The fact that the 30 day average had a higher criteria threshold (5 mg/l vs 4 mg/l for other averaging schemes) probably played into this result. Based on results from this single site, it appears that the 7-day moving average and the 1 average per 30 days did not detect DO non-attainment as frequently as did other averaging schemes.

Another way of visualizing these computations is shown as a sequence of three box and whisker plots (Figures 4-1–3; 2006, 2007 and 2008, respectively). In these figures data for the entire annual ConMon data set were included (whole year). What is clear in these diagrams is that the mean of the full data set were always above criteria thresholds (5 and 4 mg L^{-1}). However, instances of non-attainment were most frequently observed using the “all Data”, daily mean and, to a lesser extent, the 7-day moving average approaches. The final three computation methods detected no criteria violations during 2006 (Figure 4-1), only a few during 2007 (Figure 4-2) and a few more during 2008 (Figure 4 3), the year with the poorest water quality.

Table 2. A summary of DO % non-attainment estimates from the St George’s Island ConMon site for the period 2006-2008. The various methods of computing % DO non-attainment were described in the methods section of this chapter. The “whole year” columns used data for the period April-October. Other calculation periods are as indicated in the table.

Perhaps the strongest “take-home” messages from analyses at this site is that DO criteria violations occur even at sites with relatively good water quality and that substantial inter-annual variability exists relative to DO non-attainments…some years are clearly better than others. To a

St. George’s Island (XBF7904)

Year	Method	Available Annual Dataset Mean	June through August Mean	July Mean	Available Annual Dataset % Non-Attainment	June through August % Non-Attainment	July % Non-Attainment
2006	Instantaneous	6.69	5.78	5.68	4	8	10
	Daily Mean				1	2	3
	7 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 7 Days				0	0	0
	30 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 30 Days				0	0	0
2007	Instantaneous	7.05	5.73	5.35	5	9	17
	Daily Mean				2	4	13
	7 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 7 Days				0	0	0
	30 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 30 Days				0	0	0
2008	Instantaneous	7.11	5.33	5.07	10	21	27
	Daily Mean				4	9	17
	7 Day Moving Average (15 min. increment)				1	1	4
	1 Average per 7 Days				4	8	25
	30 Day Moving Average (15 min. increment)				12	25	40
	1 Average per 30 Days				0	0	0

large degree this finding is consistent with findings using the historical Cory data set collected from 1964-1969 in the Patuxent River estuary.

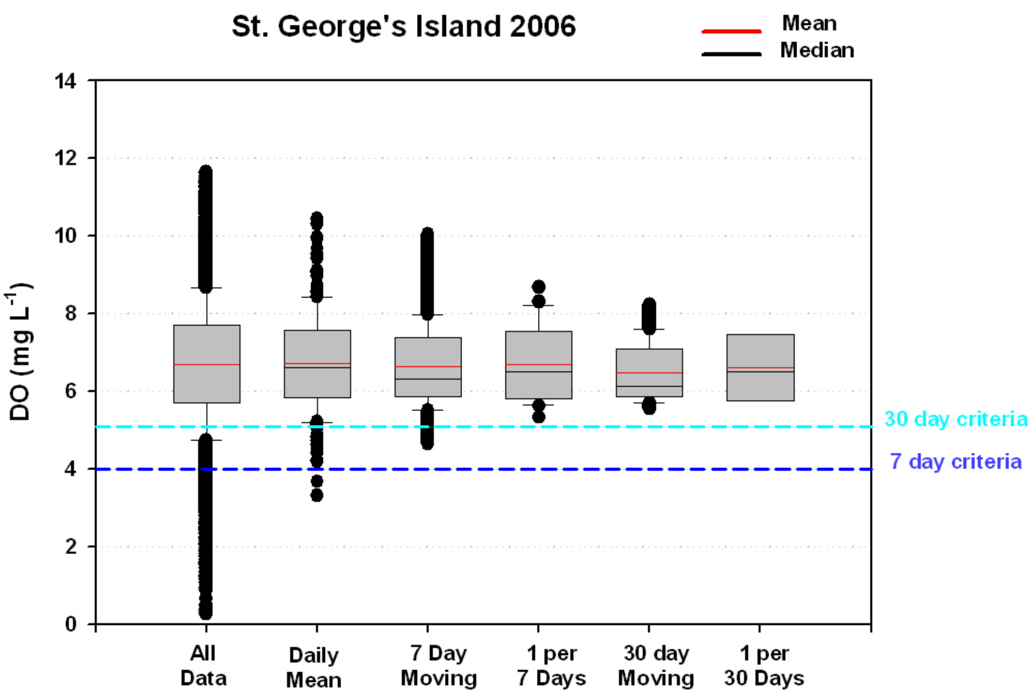


Figure 4-1.

Box and whisker plots of DO concentration based on data collected at the St. George's Island ConMon site in the lower Potomac River estuary during 2006.

The categories indicated on the x-axis were described in the Method section of this chapter. The two horizontal lines indicate DO criteria concentrations for open water sites.

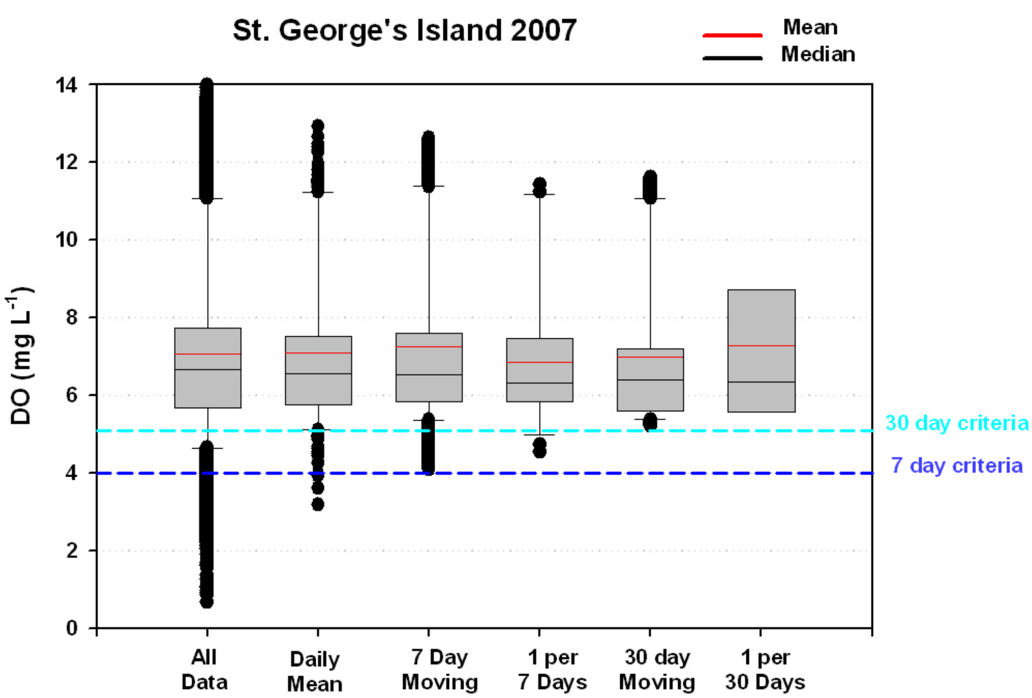


Figure 4-2.

Box and whisker plots of DO concentration based on data collected at the St. George's Island

ConMon site in the lower Potomac River estuary during 2007.

The categories indicated on the x-axis were described in the Method section of this chapter. The two horizontal lines indicate DO criteria concentrations for open water sites.

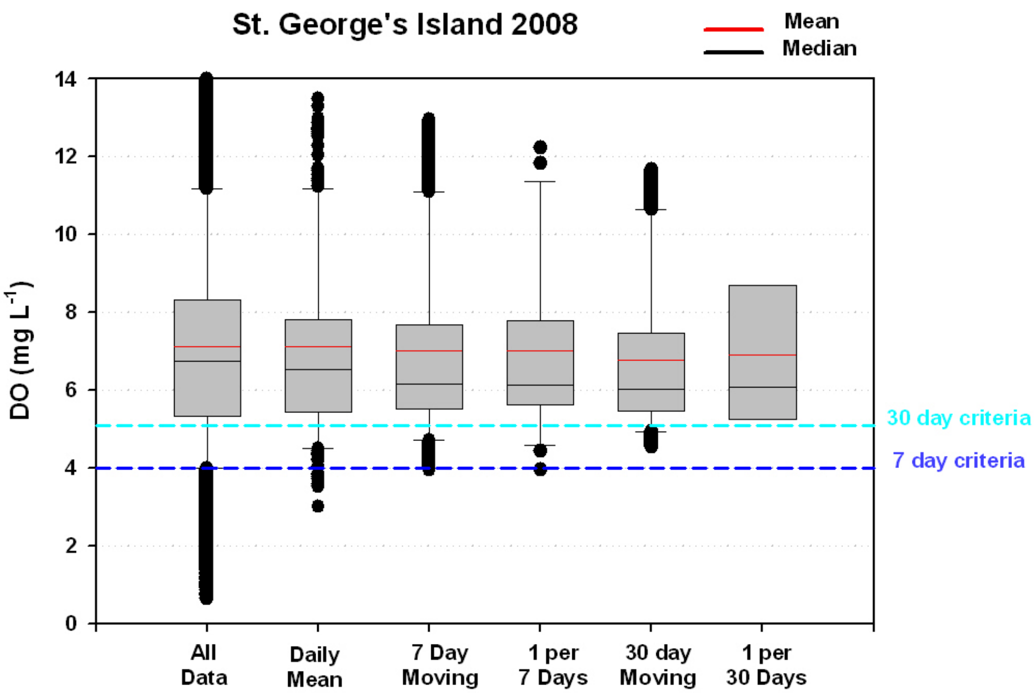


Figure 3. Box and whisker plots of DO concentration based on data collected at the St. George's Island ConMon site in the lower Potomac River estuary during 2008. The categories indicated on the x-axis were described in the Method section of this chapter. The two horizontal lines indicate DO criteria concentrations for open water sites.

High Impact Site (Sycamore Point, Upper Corsica River: XHH3851)

The Sycamore Point site in the upper portion of the Corsica River estuary is heavily impacted by nutrient additions, mainly from the agriculturally dominated watershed (Boynton *et al.* 2009). Results from % DO non-attainment for this site are summarized in Table 4-3. Several important points emerge. First, there were far higher % non-attainment rates observed at this site than at the St. George's Island site, as expected. The St. George's Island site is relatively "clean" compared the Sycamore Point site. In addition, the Sycamore Point site has far higher % non-attainment results than found in the historical data from the Cory ConMon site operated in the 1960s. Thus, it appears that there is considerable range in results consistent with our general impressions of water quality.

As at the previous site, there was not a clear result concerning the metric that might be adopted for general use in criteria attainment or non-attainment. For example, the All Data and the Daily Mean approaches tended to detect the highest failure rates. But, this was not always the case. During 2006 both the 30 day moving average and the 1 average per 30 days produced failure rates higher than the previously mentioned metrics. It may well be that the differences in criteria threshold values (4 versus 5 mg O_2 L^{-1}) were that cause of this result. However, data from both 2005 and 2008 do not support this conclusion.

Table 3. A summary of DO % non-attainment estimates from the Corsica River, Sycamore Point (XHH3851) ConMon site for the period 2005-2008. The various methods of computing % DO non-attainment were described in the methods section of this chapter. The "whole year" columns used data for the period April-

Sycamore Point (XHH3851)

Year	Method	Available Annual Dataset Mean	June through August Mean	July Mean	Available Annual Dataset % Non-Attainment	June through August % Non-Attainment	July % Non-Attainment
2005	Instantaneous	8.05	5.55	5.51	16	36	39
	Daily Mean				12	25	32
	7 Day Moving Average (15 min. increment)				3	11	16
	1 Average per 7 Days				3	11	0
	30 Day Moving Average (15 min. increment)				3	8	22
	1 Average per 30 Days				0	0	0
2006	Instantaneous	9.10	4.96	5.40	12	37	27
	Daily Mean				8	28	14
	7 Day Moving Average (15 min. increment)				10	36	17
	1 Average per 7 Days				6	29	0
	30 Day Moving Average (15 min. increment)				13	45	29
	1 Average per 30 Days				11	100	ND
2007	Instantaneous	8.57	4.93	5.76	16	47	35
	Daily Mean				13	41	30
	7 Day Moving Average (15 min. increment)				9	29	23
	1 Average per 7 Days				3	9	0
	30 Day Moving Average (15 min. increment)				18	56	6
	1 Average per 30 Days				25	100	ND
2008	Instantaneous	10.03	5.95	5.22	10	29	39
	Daily Mean				6	16	29
	7 Day Moving Average (15 min. increment)				1	4	12
	1 Average per 7 Days				0	0	0
	30 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 30 Days				0	0	0

December. Other calculation periods are as indicated in the table.

The time span considered in these evaluations also needs consideration. Without exception, the "Whole Year" computations of % non-attainment were lowest and therefore likely the least protective. When compared to the June-August % non-attainment rates the whole year rates were 2 to 3 times less frequent. However, July alone non-attainment rates were not always higher than those computed from a longer summer period (June – August). We had originally suspected that the July alone computations would yield the highest % non-attainment rates because investigations of hypoxia in deeper waters indicates this month to consistently have the most severe hypoxia. That turns out not to be the case. Of the 24 comparisons that can be made (6 computation schemes for each year and four years of data), 13 times % non-attainment was greater using the June-August data set while on 7 occasions the July only data set yielded higher % non-attainment results (4 cases of zero non-attainment were not included).

Urban Site (Fort McHenry, Patapsco River: XIE5748)

A summary of DO % non-attainment at the urban, Ft. McHenry site is presented in Table 4-4. Here again, results tended to follow many of the patterns seen at the others sites. First, there was substantial inter-annual variability. During 2004 the maximum DO % non-attainment was detected using the instantaneous metric (23%) and four of the remaining five metrics detected no failing DO conditions. During 2007, the instantaneous DO % non-attainment rate was much larger for all time periods (24-39%) and some small failure rates were found with the other metrics. Finally, it is now reasonably clear simple averages (left portion of table; pink background) are not sufficient to detect DO % non-attainment rates. At these relatively shallow sites (<2 m) DO variations on a daily basis can be severe because, in part, the effects of sediment

Fort Mc Henry (XIE5748)

Year	Method	Available Annual Data Set Mean	June through August Mean	July Mean	Available Annual Data Set % Non-Attainment	June through August % Non-Attainment	July % Non-Attainment
2004	Instantaneous	7.09	6.17	5.65	13	18	23
	Daily Mean				6	10	13
	7 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 7 Days				0	0	0
	30 Day Moving Average (15 min. increment)				0	0	0
	1 Average per 30 Days				0	0	0
2007	Instantaneous	6.85	5.44	5.52	24	39	34
	Daily Mean				18	30	29
	7 Day Moving Average (15 min. increment)				7	7	0
	1 Average per 7 Days				10	9	0
	30 Day Moving Average (15 min. increment)				1	2	0
	1 Average per 30 Days				0	0	0

respiration can be large and result in strong DO depressions, especially during the late night and early morning hours. The instantaneous metric appears to capture these events at this site better than any of the other metrics.

Table 4. A summary of DO % non-attainment estimates from the Fort McHenry (XIE5748) ConMon site in the Patapsco River for the period 2004 and 2007. The various methods of computing % DO non-attainment

were described in the methods section of this chapter. The "whole year" columns used data for the period April-November. Other calculation periods are as indicated in the table.

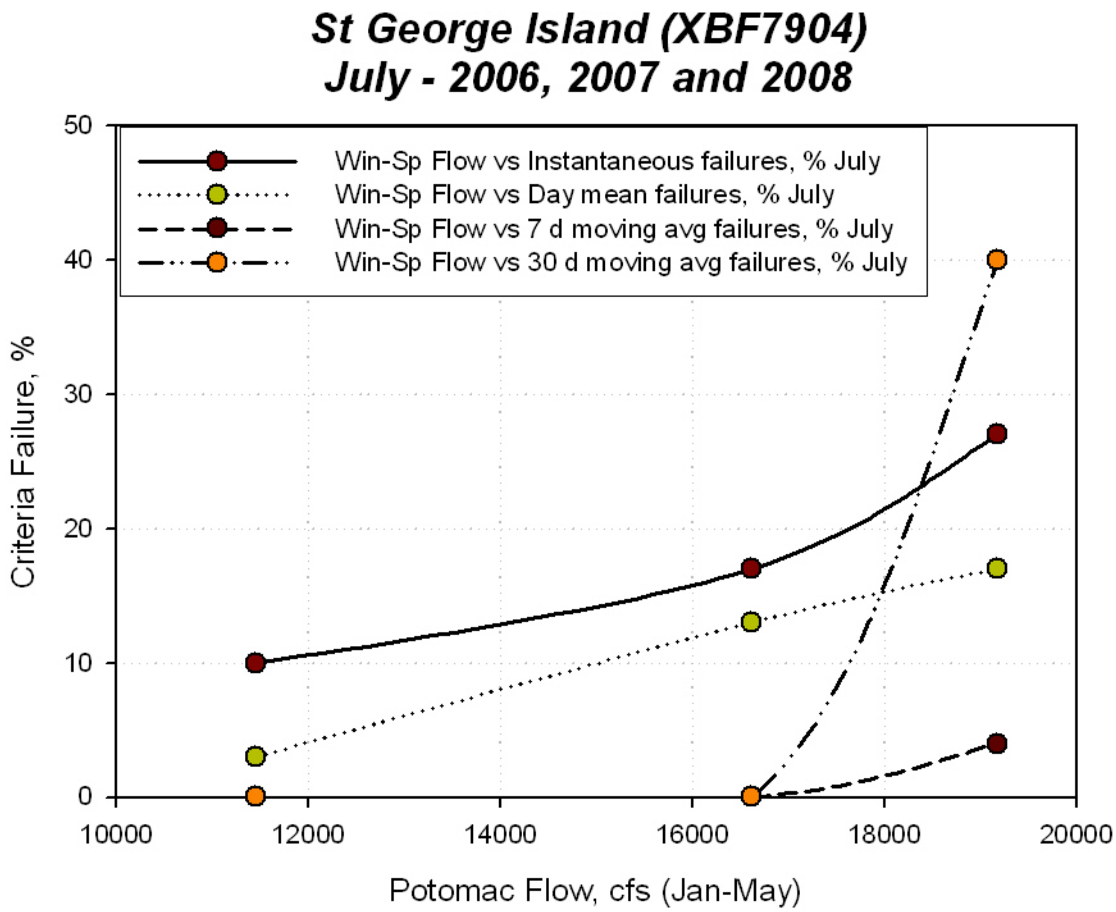
Relating DO Criteria % Non-Attainment to Other Water Quality Variables

One major goal of this work is to simply compute rates of % DO criteria non-attainment for shallow areas of the open water zone. As with many ecological issues, this one turns out to be not so simple. There are a variety of ways to compute this metric and it remains to be seen which might be the most appropriate method. There is also the issue of merging the DO criteria assessment associated with ConMon based data sets collected in shallow waters relative to open water assessments made with the traditional, low frequency monitoring data. It remains unclear as to just how this will be accomplished.

Finally, since there are not ConMon sites at all locations in the Bay and tributary rivers it would be useful to have some simple water quality variable(s) that could be used as a surrogate for data collected at a ConMon site. It would also be useful to link, in some quantitative fashion, % DO non-attainment results to other ecosystem features to explain the apparent large degree of inter-annual variability observed at some stations.

We are at early stages of this effort. However, data collected at the St George's Island ConMon site can serve as an example of future, and more thorough, efforts to link criteria results with management actions and general understanding. The % DO non-attainment results (developed using 4 different approaches) computed from 2006-2008 ConMon data were plotted as a function of Potomac River flow (Figure 4-4). In this analysis, two metrics of % DO non-attainment increased in a near-linear fashion as a function of river flow. Two other DO % non-attainment metrics remained very low until river flow was quite high at which point one increased slightly while the other exhibited a very large increase, threshold-like in nature. In this simple case the conceptual model supporting this analysis is based on the fact that river flow adds both freshwater (and buoyancy) as well as sediments and nutrients to these systems. Nutrients, in turn, tend to support higher rates of primary production. Organic matter resulting from this nutrient-stimulated production can cause increased respiration rates (utilization of DO) by the heterotrophic community. The net result, in this example, would be higher DO% non-attainment rates. We expect to continue this effort using a variety of water quality variables in addition to freshwater flow and nutrient loading rates. Variables such as TN, TP and chlorophyll-a concentration will be considered in an effort to better understand and predict levels of inter-annual variability of DO % non-attainment rates.

Figure 4. A multiple scatter plot of July DO % criteria non-attainment as a function of Potomac River flow (Jan-May flow period). Different DO % non-attainment calculation methods are indicated on the diagram.



References

Boynton, W.R., J.M. Testa and W.M. Kemp. 2009. An Ecological Assessment of the Corsica River Estuary and Watershed: Scientific Advice for Future Water Quality Management: Final Report to Maryland Department of Natural Resources. Ref. No. [UMCES]CBL 09-117. [UMCES Technical Series No. TS-587-09-CBL].

USEPA. 2010. Ambient water quality criteria for dissolved oxygen, water clarity and chlorophyll-a for the Chesapeake Bay and its tidal tributaries: 2010 technical support for criteria assessment protocols addendum. EPA 903-R-10_002 CBP/TRS 301-10. May 2010.

PART 2. Shallow Water, High Frequency Measurements and the 30 Day Mean Umbrella Approach: Two Preliminary Computations

W. R. Boynton, E. M. Bailey, M. Hall and E. Perry

Background: The traditional water quality monitoring program, in place since 1984, obtains water quality measurements once or twice a month during daylight periods, generally between 0800 and 1500 hours, at many stations in the mainstem Bay and tributary rivers. These stations, for many good reasons, are generally located over the deeper channel areas of the Bay and tributary rivers. Over a decade ago, pilot studies were conducted to see if these main channel (or off-shore) measurements accurately represented water quality conditions in the shallow waters (< 2 m depth) of the Bay and tributary rivers. The answers to this question ranged, in some instances, from generally yes to, in other instances, generally no. Other evaluations concluded that deep versus shallow water quality conditions were inconsistent and little in the way of firm generalities could be developed. In addition, there was a continuing focus in the Bay Program on SAV restoration and these communities were, of course, centered in these shallow water habitats. With these questions and goals in mind, two programs were added to the monitoring program, one focused on obtaining a much finer spatial scale data set of water quality conditions with particular attention paid to actual or potential shallow water SAV habitat (dataflow program) and the other to obtaining both long-term (3 years or more) and temporally detailed (15 minute intervals) water quality measurements in shallow waters (ConMon program). This section deals with ConMon data.

At the present time a very large data base of ConMon measurements has been generated. Some 98 different sites have had ConMon measurements (for at least 3 years) in the Maryland portion of the Bay and tributary rivers and others have been made in Virginia. One of the central findings to emerge from this data set is that there is yet another temporal scale of hypoxia in the Bay in addition to the seasonal scale hypoxia chronic to the deeper portions of the Bay. ConMon data often indicate a diel-scale (24 hour period) of hypoxia, severe at some locations, wherein dissolved oxygen concentrations drop to low levels during the hours of darkness and sometimes reach dangerously low concentrations at and just after sunrise. Qualitative inspection of these data indicate that the most severe diel-scale hypoxia is observed at sites experiencing severe eutrophication. More quantitative analyses of diel-scale hypoxia related to nutrient conditions are in progress.

Given the above observations it became apparent that ConMon data would be especially useful in at least two ways: 1) these data could be used in a variety of ways to assess trends in water quality conditions in shallow waters and SAV habitat and 2) time high frequency nature of these measurements could be used to directly evaluate surface water DO criteria attainment or failure. The latter of these items is addressed in this section with two different approaches.

ConMon Measurements and the 30-day mean: How protective is it?: The general arguments concerning application of the 30 day DO mean as a protective DO standard have been fully discussed earlier. Here we provide some sample analyses wherein for a variety of shallow water ConMon sites the summertime (Jun-Aug) 30 day mean is directly compared to the rate of DO criteria (instantaneous criteria) failure (Figure 1). In this example ConMon data were assembled from nine different locations, ranging from those having severe water quality issues to those having relatively good water quality conditions. The procedure for computing the DO means, percent failure rates and criteria values are provided in Table 1. In all there were 104 months of data included in this analysis. Several issues are apparent. First, when the 30 day mean is below the 30 day criteria value (DO < 5.0 mg/l) the rate of instantaneous DO criteria (<3.2 mg/l) failure rare is often quite high (>25%). In this case, both results signal a DO criteria failure. However, there were approximately 22 months (of 104) where the 30 day mean DO criteria was satisfied but the instantaneous criteria was not satisfied. Similar analyses have been conducted by C. Buchanan (see section X of this report) focused on the ConMon sites along the Potomac River estuary and similar results were obtained. We also conducted this same type of analysis but used the 7 day failure rate and in that case the 30 day mean was more protective but not completely protective. As a part of the Maryland Chesapeake Bay Biomonitoring Program we will continue to use ConMon data and make these computations for additional sites in the Maryland Bay. At this point it seems safe to tentatively conclude that for shallow water areas the 30 day mean is not protective of short-term DO criteria in many instances during summer periods (Jun – Sep).

The Issue of Duration of Low DO Conditions: For the formal DO criteria analysis there are both temporal and spatial considerations. In this analysis, using ConMon data, we are only considering the temporal aspect. However, in the formal analysis there is recognition in both the temporal and spatial domains that there needs to be some degree of "forgivness" of criteria violations and this seems appropriate given the very dynamic nature of estuarine systems. In our analysis of DO conditions in the Patuxent estuary during the 1960s, a period before this system underwent severe eutrophication, there were times (not very frequent) when surface DO criteria were violated. Thus, if a single violation was all it took to fail DO criteria, we would likely always have DO failures in most places for most time periods. That being the case, a 10% failure buffer has been adopted. However, this buffer needs to be considered in the light of just how the 10% acceptable violation rate is distributed in time.

Consider for a minute the breathing rate of a human as an analog of this problem. If we inhale once every 6 seconds we take 10 breaths per minute, 600 hundred breaths per hour and 14,400 per day. If we were to skip 10% of those breaths at a rate of 1 in every 10 breaths we would be fine…maybe a bit inconvenienced, but basically fine. However, if we were to skip all 10% at one time we would be dead…quite the difference.

We have examined the issue of DO criteria violation rate duration at a selection of Maryland ConMon sites and will continue to examine additional sites for the next several months. The data used for dissolved oxygen criteria failure duration calculations was extracted from the 2001 to 2008 Maryland Department of Natural Resources Continuous Monitoring database (www.eyesonthebay.net) provided by Ben Cole (MDDNR). The file was in .txt format and imported into SAS® 9.2 (http://www.sas.com/). Data found to have error codes (http://mddnr.chesapeakebay.net/eyesonthebay/documents/SWM_QAPP_2010_2011_FINALDraft1.pdf) were removed prior to analysis. For this exercise the duration of time a measurement of dissolved oxygen was found below 3.2 mg L^{-1} (instantaneous criteria) and separately for 5.0 mg L^{-1} (30 day mean criteria) was calculated for the period of record at a Con Mon station. Con Mon measurements are made up of a dissolved oxygen reading taken every 15 minutes so each increment of duration of failure is 15 minutes. A duration sequence of failure was calculated as a series of continuous 15 minute intervals where the measured dissolved oxygen value was below the chosen criteria. If a measurement time stamp exceeded 40 minutes (to allow some variance in time stamp intervals due to data sonde set up) or changed dates (data sonde was removed or unavailable for some period of time) the duration sequence was reset to start again. Total duration of dissolved oxygen failure for a sequence was the sum of the 15 minute intervals. In this early version of the duration calculator, failures are terminated at the end of each 24 hour period. We know that in some cases the failure duration continues into the next day. The calculator needs to be up-graded to address this issue as well as several other problems. So, at present the calculator provides a minimum estimate of DO failure duration.

Examples of DO criteria failure duration for two criteria levels are provided in Table 2. We selected sites exposed to very severe eutrophication (Bishopville Prong in the MD Coastal Bays), reasonably good water quality conditions (St. George Island), a tidal freshwater site in an enriched estuary (Jug Bay; Patuxent River) and a mesohaline site exposed to open waters (Pin Oak; Patuxent River). As expected, at the site with severe eutrophication there were many criteria failures and criteria failure durations ranged from 12 to 24 hours (likely longer than this). At the less impacted sites, DO criteria failures were of shorter durations, especially for the instantaneous criteria (< 3.2 mg/l). At the higher DO criteria (< 5.0 mg/l) duration of failures remained long, often up to 24 hours. This evaluation is in early stages and some refinements have already been suggested. The point we make here is that it does not appear that DO failure rates are evenly distributed in time and this needs to be further evaluated to be certain that DO criteria values are as protective as they were intended to be.

Appendix 3

Conditional Probability

30-day mean vs. 7-day mean

Elgin Perry

8-27-2010

The following work summarizes some additional analyses conducted with the Potomac Continuous Monitoring data to address the question of whether the 30-day mean criterion serves as an umbrella for the 7-day mean criterion. The results here seem to confirm that it would be a rare situation where the 30-day mean would be satisfied and the 7-day mean would be violated more than 10% of the time. However, this does not seem to be a broad umbrella in that the margin of protection is not great.

Methods:

The method employed is based on the simple-minded approach (Figure 1) that if the variability of the 7-day mean about the 30-day mean has a standard deviation less that 0.7805, then we can expect that the 7-day criterion will be violated less than ten percent of the time if the 30-day criterion is met.

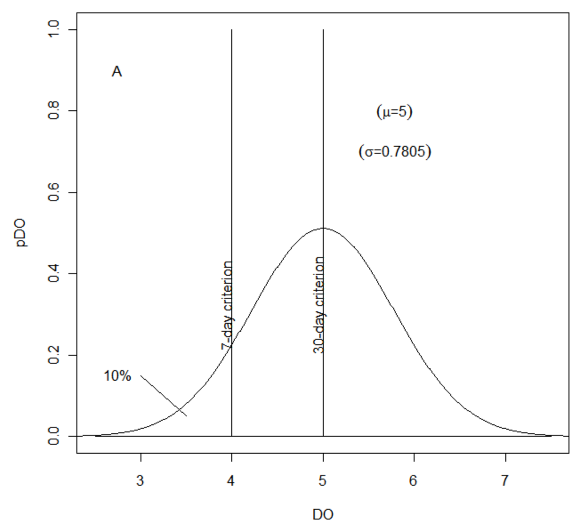


Figure 13. Illustration of the level of variability of the 7-day mean about the 30-day mean that results in up to 10 % violations of the 7-day mean criterion when the 30-day mean criterion is met.

To use this approach, an estimate of the standard deviation of the 7-day mean about the 30-day mean is needed. To estimate this quantity, I used data from the Potomac ConMon locations (Table 1, Figure 2).

Table 1. Names, locations, and years of Continuous Monitor data used.

location	Latitude	Longitude	years
Occoquan	38.64038	-77.219416	2007-2009
Pohick Creek	38.67591	-77.16641	2007-2009
Potomac Creek	38.3436	-77.30485	2007-2009
Monroe Bay	38.23197	-76.96372	2007-2009
Nomini Bay	38.1316	-76.71759	2007-2009
Yeocomico River	38.02878	-76.55184	2007-2009
Fenwick	38.66993333	-77.11513333	2004-2008
Piscataway Creek	38.70156667	-77.02593333	2004-2008
Mattawoman Creek	38.55925	-77.1887	2004-2008

Figure 2. Locations of Potomac ConMon data collection sites used for this analysis.

Beginning with the first collection day for each year at each location, blocks of 30 days were created to represent months. Partial months at the end of each collection year were counted as a month. Similarly, weeks were created by starting with the first collection day of each year and counting off blocks of 7 days. With these definitions, monthly means were computed as the arithmetic average of DO for each month. Weekly means were computed as the arithmetic average of DO for the intersection of month and week. Thus a week that bridges across two months would have it's data divided by month and a weekly mean computed for each part. Weekly means and Monthly means were merged by month and a residual computed by subtracting the monthly mean from each weekly mean computed within that month. Various analyses were conducted on these residuals to assess the variability of weekly means about the monthly mean.

Graphical analyses were used to assess the uniformity of variation over other factors. Distribution functions and quantile estimation was used to estimate the rate of violation of the 7-day criterion given that the 30-day criterion was satisfied.

Results:

First I report a number of graphical assessments:

The basic distributional assessment of the residuals (Figure 3) shows that they are reasonably symmetric and centered about zero .   The distributions is heavy tailed compared to the normal distribution in the extreme tails suggesting that there are weeks that have a greater deviation (both high and low) than would be expected for a normal distribution. The central part of the distribution seems to follow the normal distribution closely.

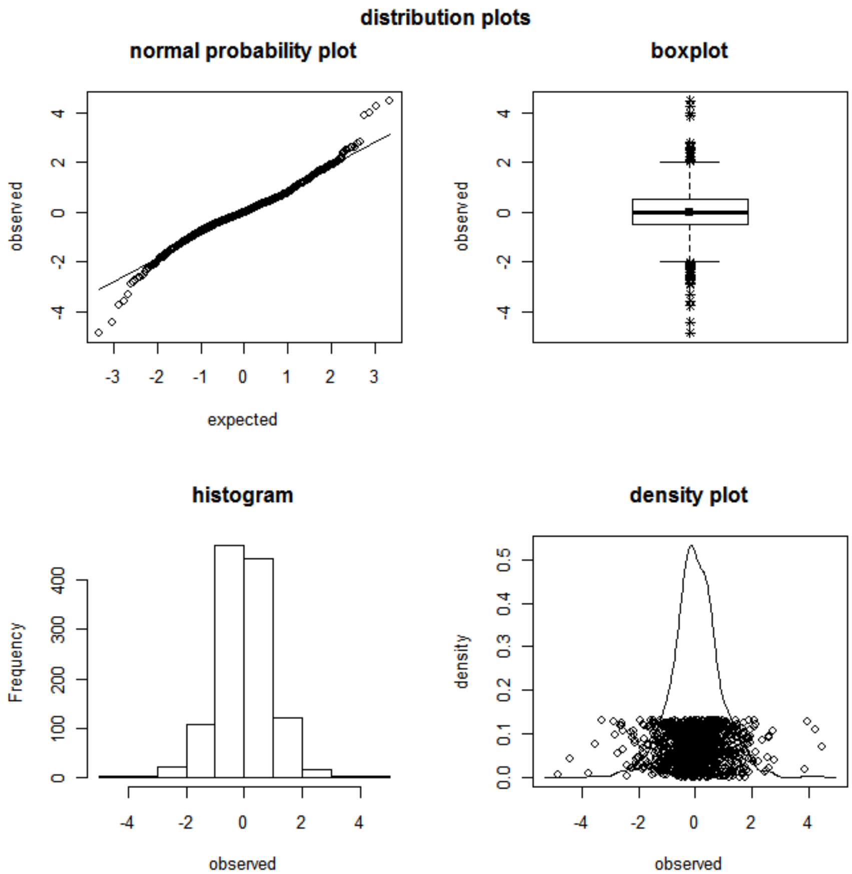


Figure 14. Basic distributional properties of the residuals.

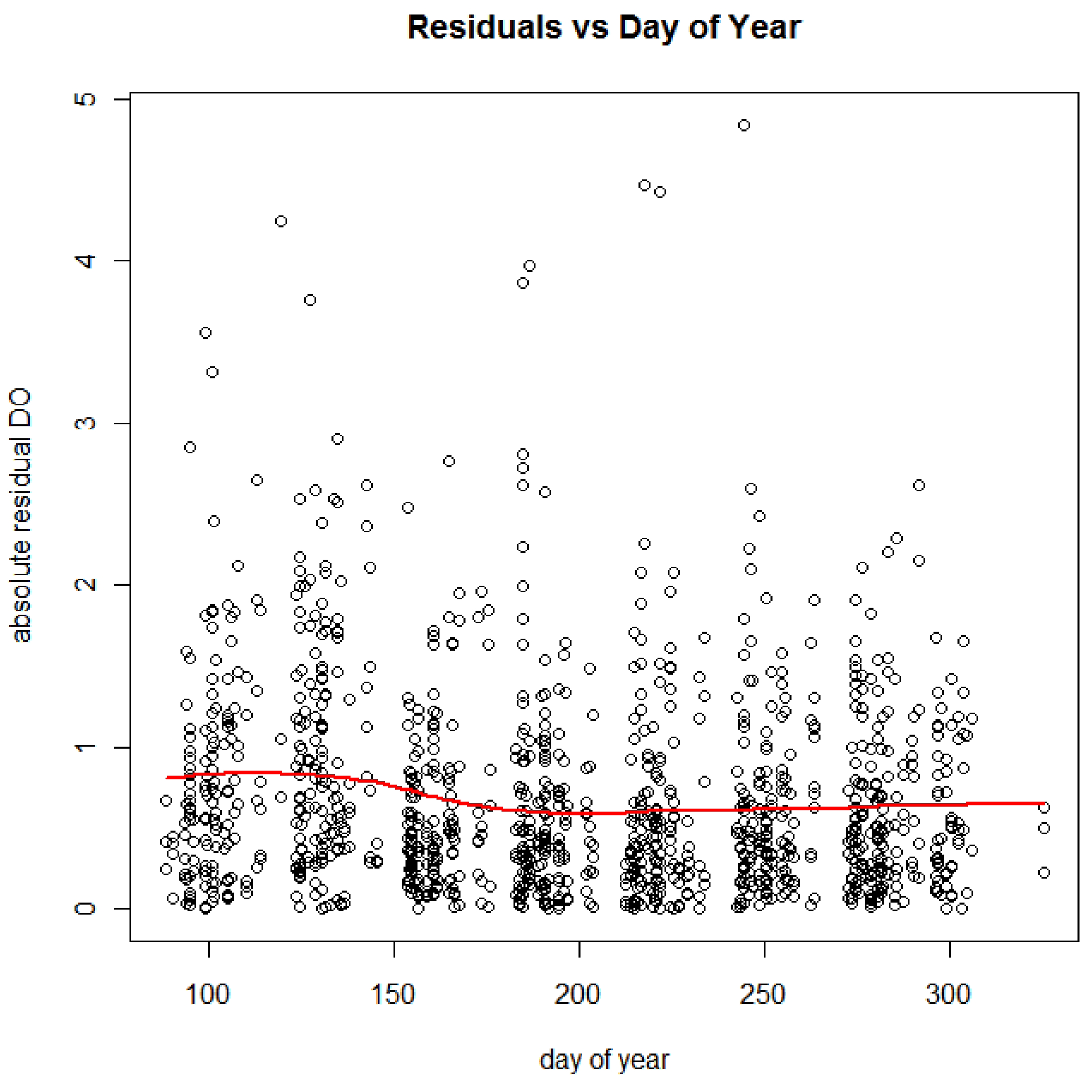


Figure 15. Assessment of seasonal trend in variability.

There is evidence of higher variability in spring than in summer and fall (Figure 4).

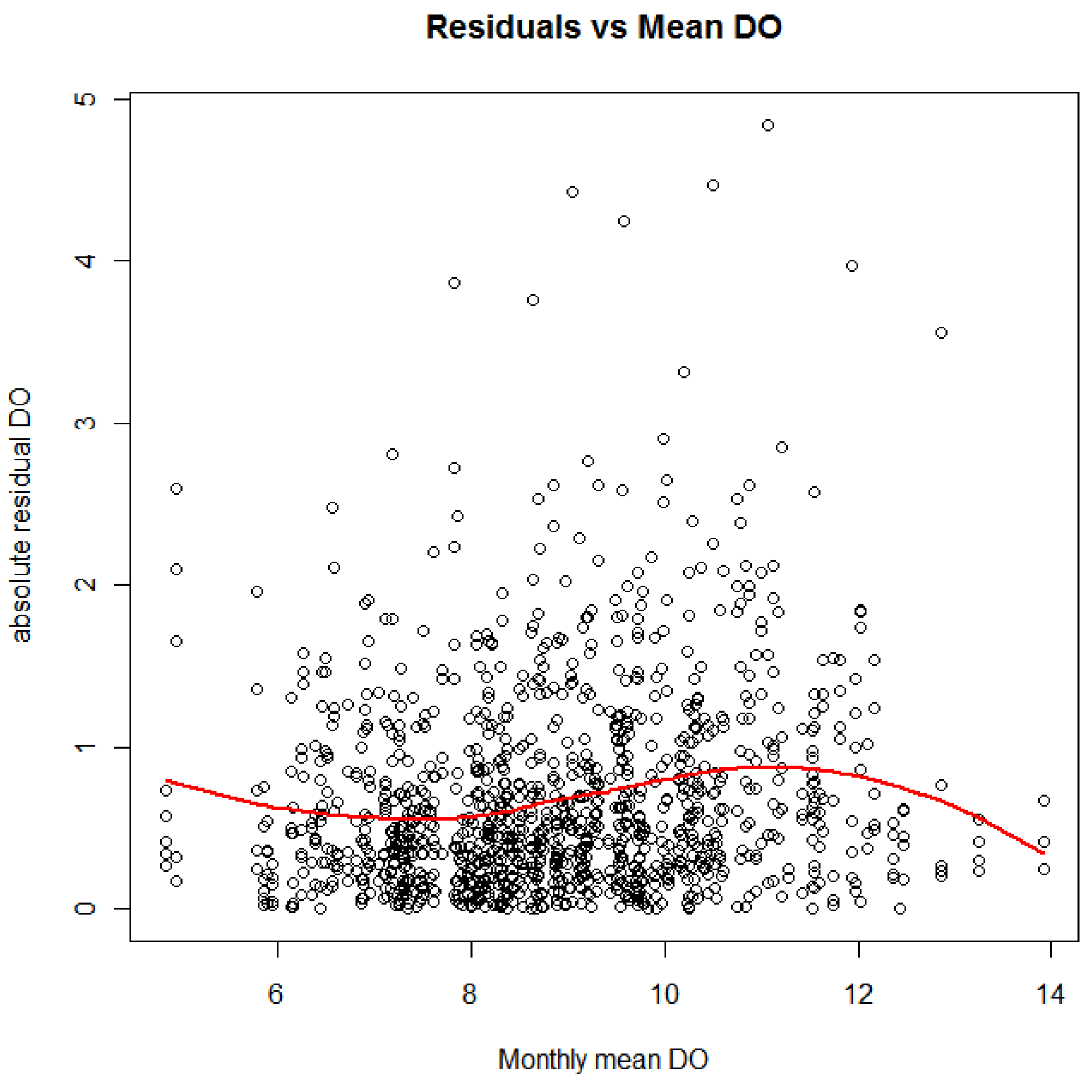


Figure 16. Trend of variability with mean DO.

There is evidence that the variability increases with increasing mean DO (Figure 5.). It is likely that the seasonal trend and the trend with the Mean DO are the same trend because there is a seasonal trend in mean DO.

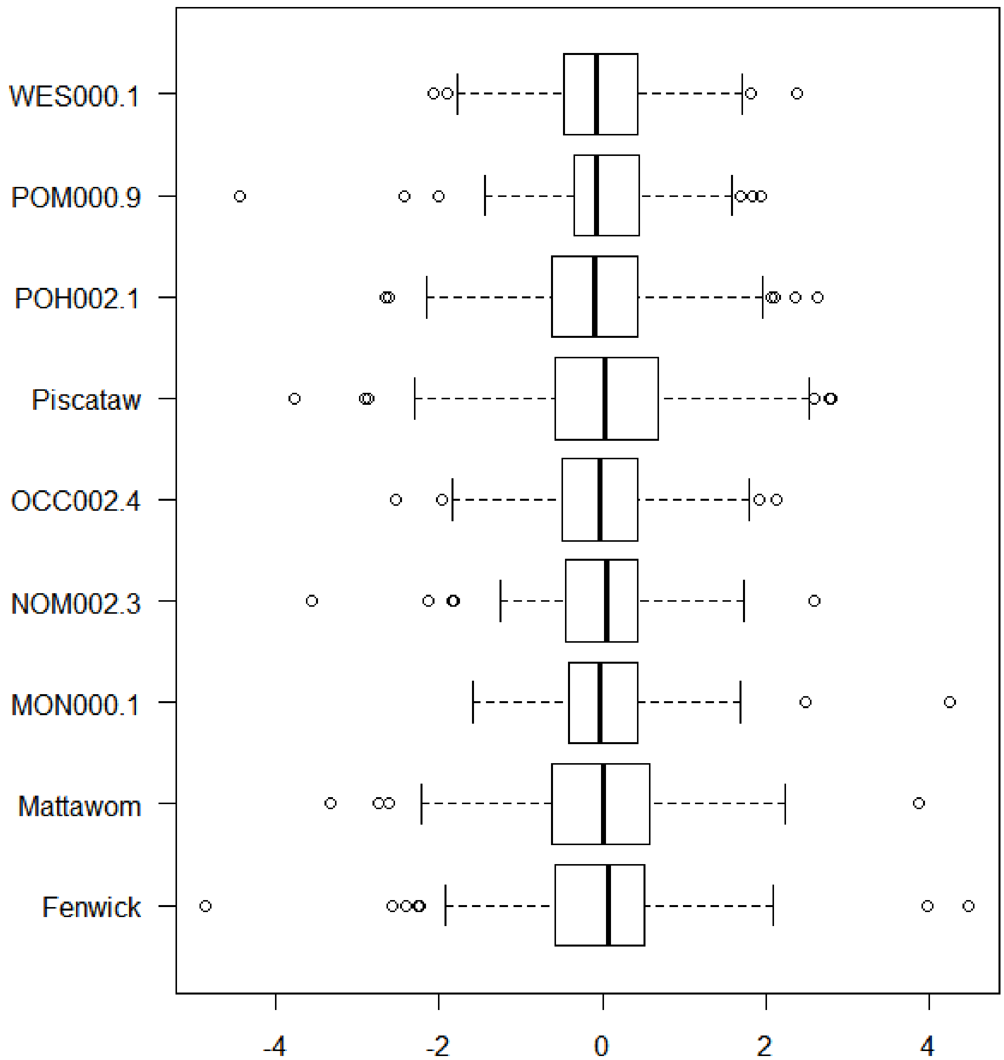


Figure 17. Box and Whisker plots of residuals by sampling location.

There is little evidence of change in variability with location except that Maryland locations appear to have greater variability than Virginia locations.

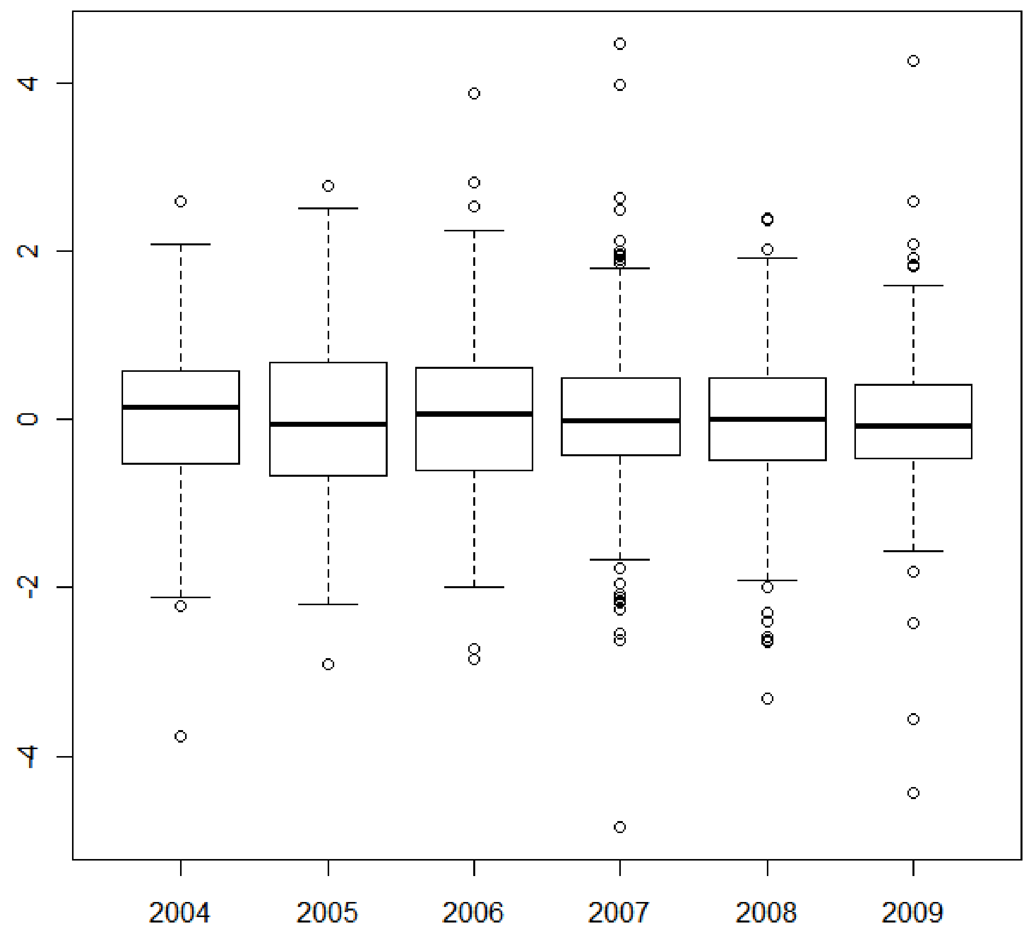


Figure 18. Box and Whisker plots of residuals by year.

There appears to be a time trend in variability (Figure 7.) with a decrease in variability occurring in between 2006 and 2007. However, recall from the collection dates (table 1) that only Maryland collected data prior to 2007 and thus this is the state trend from a different view. The pattern of difference in variability between states persists when the data are subsetted to just 2007-2009 for Mattawoman and Fenwick, but variability at Piscataway is more comparable to Va. locations for this time period (figure not shown).

If the standard deviation of weekly mean about monthly mean is estimated for all data, the value is 0.9648719 which is slightly above the value of 0.7805 which would insure that there would be less than 10% violation of the 7 day criterion is the 30-day criterion were satisfied (Table 2., column 2). However, if the 30-day mean is increased to just 5.3, then we would expect fewer than 10% violation of the 7-day mean. Thus it seems that if the 30-day mean were hovering between 5.0 and 5.3 for an extended period, then there might be greater than 10% violation of the 7-day criterion when the 30-day criterion is satisfied. This circumstance would seem to be a rare event.

Recall that there is evidence that variability increases with the 30-day mean DO (Figure 5.). It is reasonable to exclude the variability associated with high DO because when the 30-day mean DO is high, then it is not hovering in that region close to the criterion which we would expect to also observe violations of the 7-day criterion. To exclude the variability associated with high DO, a subset of the data was created that included only weeks associated the 30-day mean DO of less than 8.0 (the minimum value for the 30-day mean is 4.848). Using this subset of the data, the standard deviation was estimated as 0.8439. This remains slightly larger than the 0.7805 which would insure that the 30-day criterion is an umbrella for the 7-day criterion, but with this, the 30-day mean DO need only be greater than 5.1 to insure fewer than 10% violation of the 7-day criterion (Table 2., column 3).

Note that in low salinity waters where the 30-day criterion is 5.5, then we would expect only 6% or 4% violations of the 7-day criterion for the two estimates of standard deviation (Table 2, row 6). Thus it seems reasonable to conclude that the 30-day mean is an effective umbrella for low salinity. These probability estimates have been made using an assumption that the weekly residuals are approximately normally distributed.

Table 2. Probability of violating 7-day mean criterion as a function of 30-day mean DO for two levels of 7-day mean variability.

30-day mean DO	Prob(|sd=0.9649)	Prob(|sd=0.8439)
5.0	0.1500	0.1180
5.1	0.1271	0.0962
5.2	0.1068	0.0775
5.3	0.0889	0.0617
5.4	0.0734	0.0486
5.5	0.0600	0.0377
5.6	0.0486	0.0290
5.7	0.0390	0.0220
5.8	0.0311	0.0165
5.9	0.0245	0.0122
6.0	0.0191	0.0089

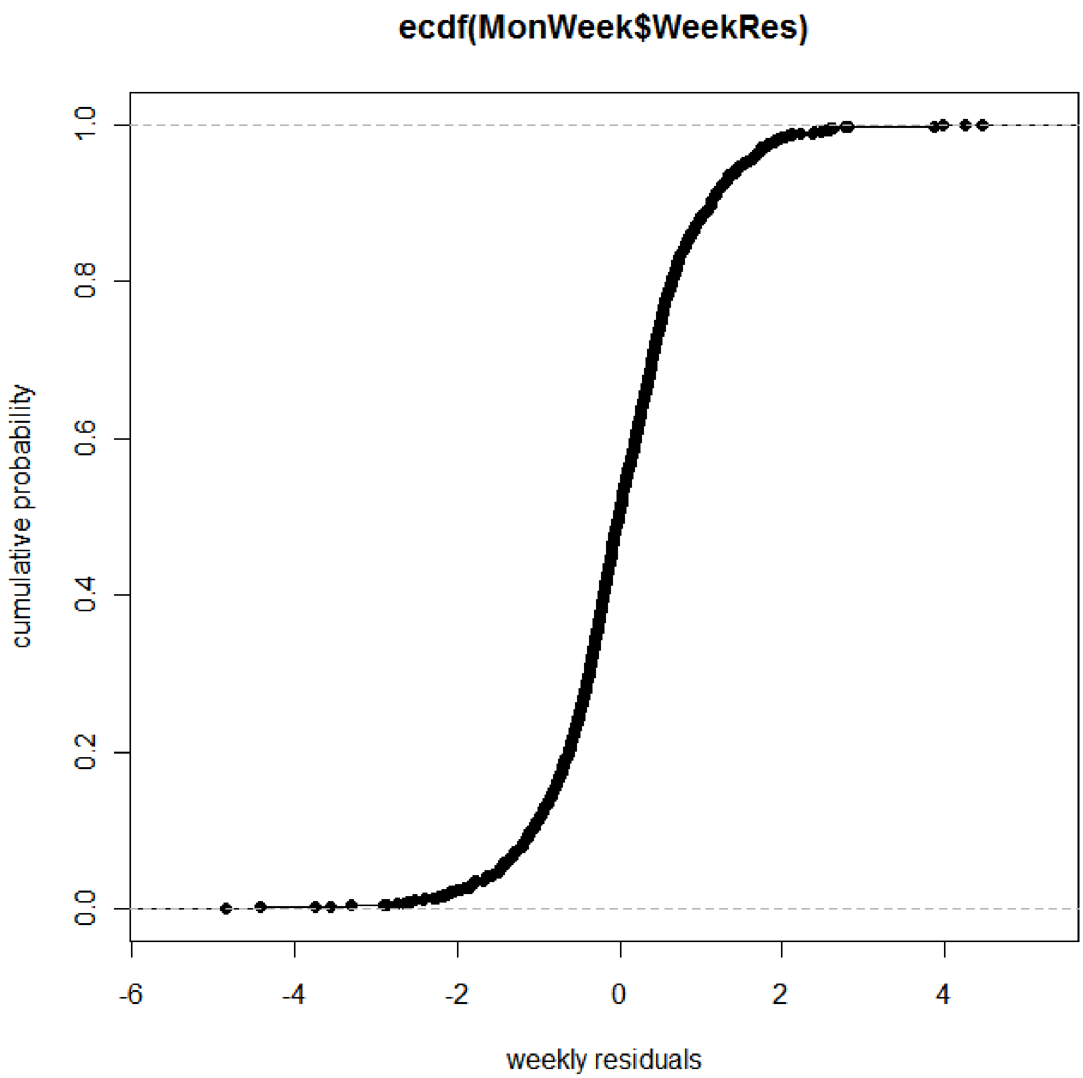


Figure 19. Empirical Cumulative Distribution Function (ECDF) of the set of weekly residuals about the monthly mean.

As a check on the reliability of the normality assumption, we compare the important quantiles of the normal distribution to those of the empirical cumulative distribution function (Figure 8.). The 10th percentile of a normal distribution with mean zero and standard deviation = **0.9649** is

-1.236533. The 10th percentile of the empirical cumulative distribution function is -1.076785. This suggest that the observed deviations at the 10th percentile are less than predicted by the normal distribution which implies that the probabilities in Table 2 are somewhat higher than might be realized. The difference between the 30-day criterion and the 7-day criterion is -1.0 and a quantile of -1.0 corresponds to between 11 and 12 percent (Table 3., column 2) based on the ECDF. This suggests that if the 30-day mean criterion were satisfied exactly, there would be on average 11-12 percent violations of the 7-day criterion. Of course typically the 30-day mean would be satisfied by some margin and if that margin were as little as 0.14 then on average there would be fewer than ten percent violations of the 7-day criterion.

Table 3. Percentiles and corresponding quantiles from the empirical cumulative distribution functions for all weekly residuals (column 2) and weekly residuals given that the 30-day mean is less than 8.0.

percentile	quantile all Weekly residuals	quantile given mean DO < 8
13%	-0.9160	-0.7321
12%	-0.9574	-0.7545
11%	-1.0329	-0.8156
10%	-1.0768	-0.8448
9%	-1.1404	-0.9186
8%	-1.2011	-0.9633
7%	-1.3006	-1.0411
6%	-1.3921	-1.1210
5%	-1.4802	-1.1903

These same calculations are done using the ECDF of weekly residuals given that the 30-day mean is less then 8.0. These results (Table 3, column 3) suggest that if the 30-day criterion is satisfied exactly, there would be on average 7-8 % violations of the 7-day criterion.

Conclusion:

These results suggest that we would only see greater than 10% violations of the 7-day criterion given that they 30-day criterion is met if the 30-day mean were hovering at or just above the 30-day criterion. Because the 30-day mean rarely exhibits this behavior, it seems safe to conclude that in most cases the 30-day criterion acts as an umbrella for the 7-day criterion. However, the umbrella does not seem to be a broad one. Slight increases in the variation of the weekly mean about the monthly mean without corresponding increases in the monthly mean could start to increase the violation rate for the 7-day criterion to above 10 percent.

APPENDIX 4

A TEST OF THE "UMBRELLA CRITERIA" CONCEPT IN TIDAL POTOMAC RIVER SHALLOW WATERS

Claire Buchanan - ICPRB

Dissolved oxygen criteria for migratory fish spawning and nursery uses and open water fish and shellfish uses are presently applied to shallow waters designated for bay grass use on the assumption that DO conditions are not sufficiently different to warrant criteria for shallow waters. High frequency DO data are suggesting this assumption may be faulty and the "umbrella criteria" that perform in open waters may not necessarily work in shallow waters. The validity of the "umbrella criteria" concept in shallow waters was tested with high frequency data collected at 19 nearshore sites in the tidal Potomac River (**Figure 1**). Specifically, does attainment of the 30-day mean DO criteria protect against failure of the 7-day mean DO criteria, and do both of these criteria protect against failure of the instantaneous minimum DO criteria in shallow waters?

Over 1.1 million dissolved oxygen records were collected by Maryland and Virginia at 19 Potomac shallow water stations during spring, summer, and autumn. Most stations were monitored for two or three years between 2004 and 2008; four were monitored all five years. Sondes were positioned at median depths of 0.2 – 3.0 m below the surface and reading made at 15-minute intervals. The 30-day, 7-day, and instantaneous minimum criteria were applied to each site -year-season subset of CMON data to address the "umbrella criteria" question.

Means derived from high frequency (CMON) data are very close to the true mean for a given location while those derived from low frequency, or point sample, data can diverge from the true mean. To evaluate the sensitivity of low frequency data estimates of the 30-day mean to 7-day and instantaneous DO criteria failures, 30-day means were computed both from the available low frequency (calibration) data and high frequency data. Guidelines for computing and applying the 30-day mean, 7-day mean and instantaneous minimum criteria were established to ensure consistency and avoid artifactual results (**Table 1**). The CMON data, the data analysis methods, and some of the results and conclusions are described in more detail in Buchanan (2009).

Figure 2 shows the frequency per month of failing the 7-day mean criteria, plotted against the 30-day mean derived from high frequency (CMON) data and low frequency (calibration) data. **Figure 3** shows the frequency per month of failing the instantaneous minimum criteria, plotted against the 30-day mean derived from high frequency (CMON) data and low frequency (calibration) data. **Figure 4** shows the frequency per 7-day period of failing the instantaneous minimum criteria, plotted against the 7-day mean. Spring results are separated from summer and autumn results and tidal fresh (TF) and oligohaline (OH) results are separated from mesohaline (MH) results because DO criteria differ according to season and salinity zone (see **Table 1**). In each graph, the inverted triangle indicates the criteria applicable to the metric on the x-axis. Points to the right of the inverted triangle are achieving that metric's criteria. **Figure 5** compares monthly means derived from the low and high frequency data. **Figure 6** delineates the threshold for failure of the instantaneous minimum criteria as a function of diel or weekly magnitude of change in DO and daily or weekly mean DO.

<u>Findings</u>:

- *Depending on allowable exceedances*, the 30-day mean criteria applied to the low frequency (calibration) data could be considered protective of the 7-day mean criteria. The 30-day mean criteria applied to the high frequency (CMON) data show similar results and support this finding.
- If the allowable exceedance dictates, for example, that only 1 month of the year can have failures of the 7-day mean criteria, then the 30-day mean criteria is *not* protective of the 7-day mean criteria in Piscataway Creek (2004, 2005), Piney Point (2006), St. Mary's River (2008), and Breton Bay (2008).
- The 30-day mean and 7-day mean criteria are *not* protective of the instantaneous minimum DO criteria, regardless of whether they are derived from low or high frequency data.
- Monthly means derived from low frequency (calibration) data are inaccurate estimates of the true mean, and also appear to be biased in some season-salinity zones.
- Meeting the instantaneous minimum criteria is a largely function of the daily mean DO *and* the diel magnitude of change in DO and the trajectory these parameters take over time. If the diel magnitude is large and the mean is relatively low, the probability of failing the instantaneous minimum criteria is high. If the diel magnitude is small and the mean relatively high and stable, the probability of failing the instantaneous minimum criteria is very low.
- Data points representing weeks with low frequencies (>0% - 1%) of failing the instantaneous minimum criteria provide a linear boundary that separates days or weeks achieving the instantaneous minimum criteria from those failing the criteria. This linear boundary quantitatively describes the relationship between mean DO and magnitude of change and the thresholds of instantaneous minimum criteria failure on daily and weekly scales.
- The 7-day mean DO criterion of 6 mg/liter in spring migratory and spawning reaches is only protective of the instantaneous minimum criteria when the weekly magnitude of

change in DO is less than ~6 mg/liter. The 7-day mean DO criterion of 4 mg/liter is only protective of the instantaneous minimum criteria when the weekly magnitude of change is less than ~2 mg/liter, a phenomenon that the tidal Potomac embayments and river flanks never experienced in the spring, summer or fall between 2004 and 2008.

- Daily or weekly DO means of ~10 mg/liter are protective of the instantaneous minimum DO criteria in almost all circumstances. DO means below 10 mg/liter are only protective of the criteria if their magnitudes of cyclic change in DO (diel, weekly) are proportionately smaller, i.e. below the boundary lines indicated in **Figure 5**.

Buchanan, C. 2009. An Analysis of Continuous Monitoring Data Collected in Tidal Potomac Embayments and River Flanks. ICPRB Report 09-3, 56 pgs. Available online at: http://www.potomacriver.org/cms/publicationspdf/ICPRB09-03.pdf

Table 1. Metrics, criteria, and computation guidelines.

Metric and Criteria	CMON data	Calibration data
30-day mean • ≥5.5 mg/liter, TF year-round • ≥5.0 mg/liter, OH & MH year-round	all available data averaged by month (not exactly 30 days); months with less than 20 days of uninterrupted DO recordings not included	samples from 0.5-1.5 m depths only, all available data averaged by day, then averaged by month
7-day mean • ≥6.0 mg/liter, TF&OH Feb1-May 31 • ≥4.0 mg/liter, all salzones Jun1-Jan 31 and MH Feb1-May 31	all available data between midnight on the first day and midnight on the 7th day averaged; means were calculated a) from rolling 7-day averages advanced in 1-day steps or b) from sequential 7-day periods with uninterrupted data records (method noted in figures); weeks with less than 6 days of uninterrupted DO recordings were excluded	n/a
Instantaneous minimum • ≥5 mg/liter, TF & OH Feb1-May 31 • ≥3.2 mg/liter @ ≤29^{o}C and ≥4.3 mg/liter @ >29^{o}C, all salzones Jun1-Jan31 and MH Feb1-May 31	the frequency of observations failing the criteria each day, week, and month; excluded: days with n<95 records, weeks with n<576 records (6 days), months with n<27 days of uninterrupted DO recordings	n/a

Figure 1. Tidal Potomac River "continuous monitoring" (CMON) stations, 2004-2008. Data for 3 of the 22 stations were not included in the analysis: the two stations in the District of Columbia and one in Neabsco Creek. Data for the two District of Columbia stations were available but had not been QA/QC'ed by the provider when the analysis was performed. Data for Neabsco Creek were only collected in the summer of 2006.

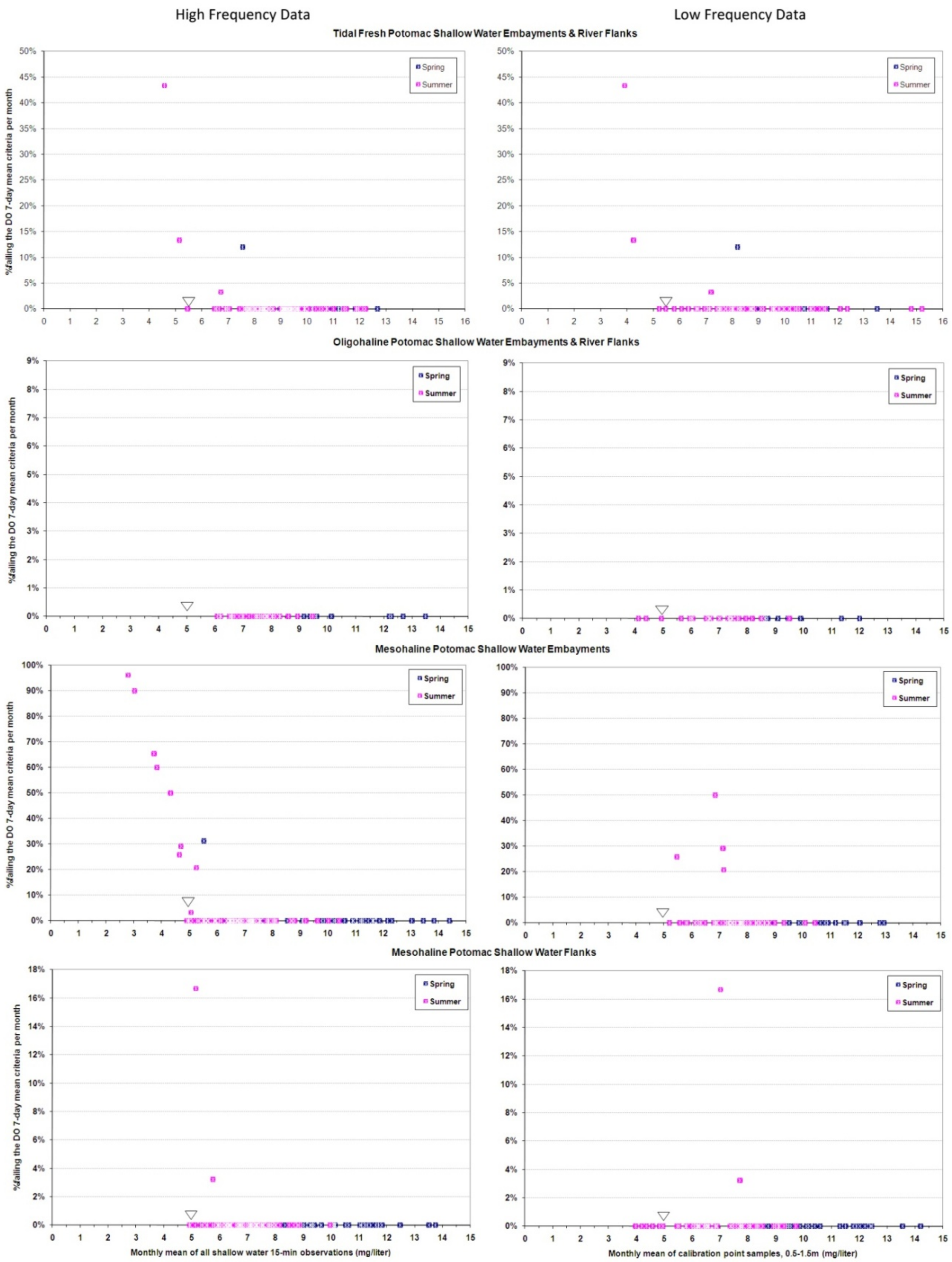


Figure 2. The frequency per month that rolling 7-day periods (1-day step) fail the 7-day mean DO criteria, plotted against the corresponding 30-day mean DO. Overall, 16 of the 415 months (3.86%) represented in the 20 tidal Potomac shallow water stations between 2004 and 2008 had one or more 7-day periods failing the 7-day mean criteria. Approximately half the failures occur in months where the 30-day mean criteria are met. Note: not all of the failing months are apparent in the low frequency data because point samples were not available for all months.

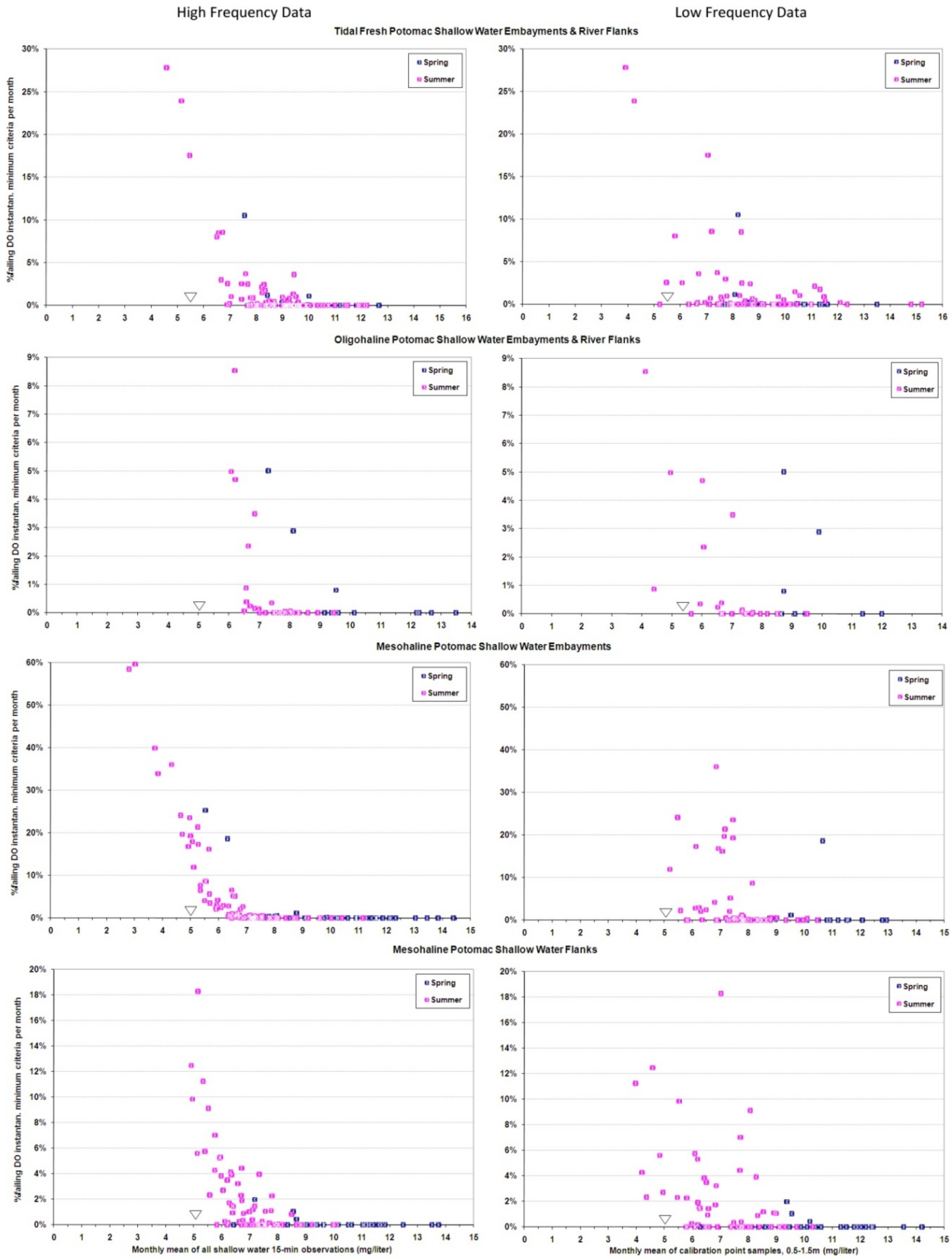


Figure 3. The frequency per month that rolling 7-day periods (1-day step) fail the 7-day mean DO criteria, plotted against the corresponding 30-day mean DO. Overall, 175 of the 415 months (42.2%) represented in the 20 tidal Potomac shallow water stations between 2004 and 2008 had failures of the instantaneous minimum DO criteria. Most instantaneous minimum criteria failures occurred in months where the 30-day mean criteria are met.

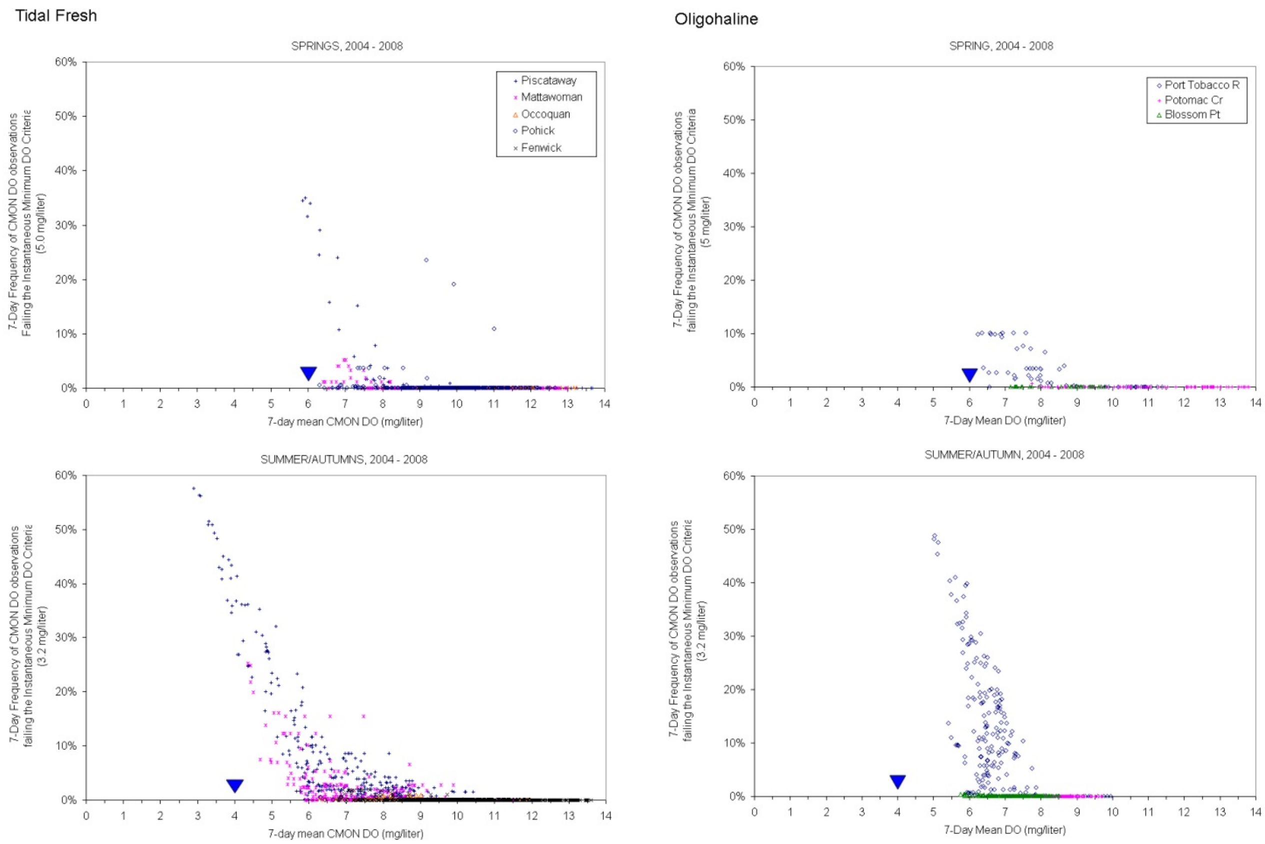


Figure 4. The frequency per 7-day period of failing the instantaneous minimum criteria, plotted against the corresponding 7-day mean in tidal fresh and oligohaline salinities. Frequencies were calculated on rolling 7-day periods (1-day step). Most instantaneous minimum criteria failures occurred in 7-day periods where the 7-day mean criteria are met.

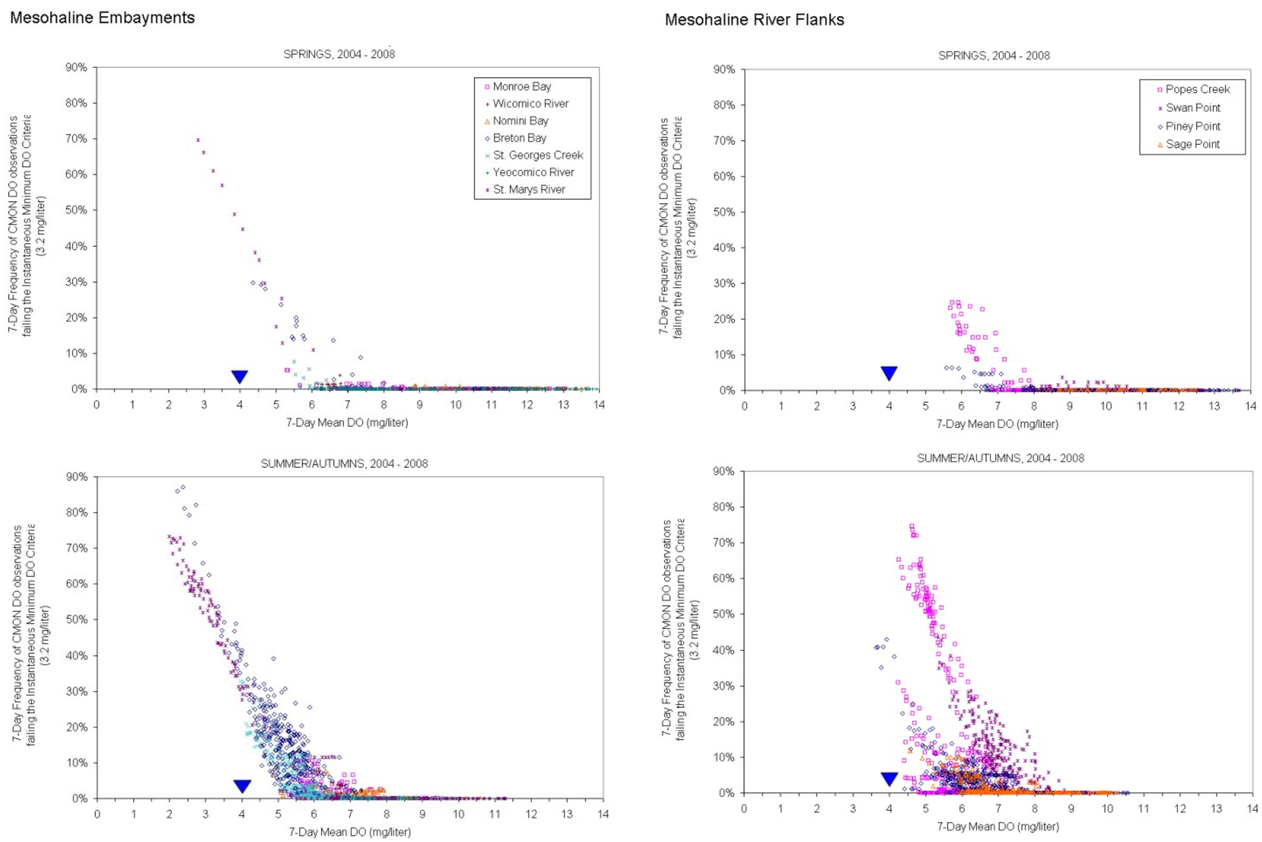


Figure 4 (cont.). The frequency per 7-day period of failing the instantaneous minimum criteria, plotted against the corresponding 7-day mean in mesohaline salinities. Frequencies were calculated on rolling 7-day periods (1-day step). Most instantaneous minimum criteria failures occurred in 7-day periods where the 7-day mean criteria are met.

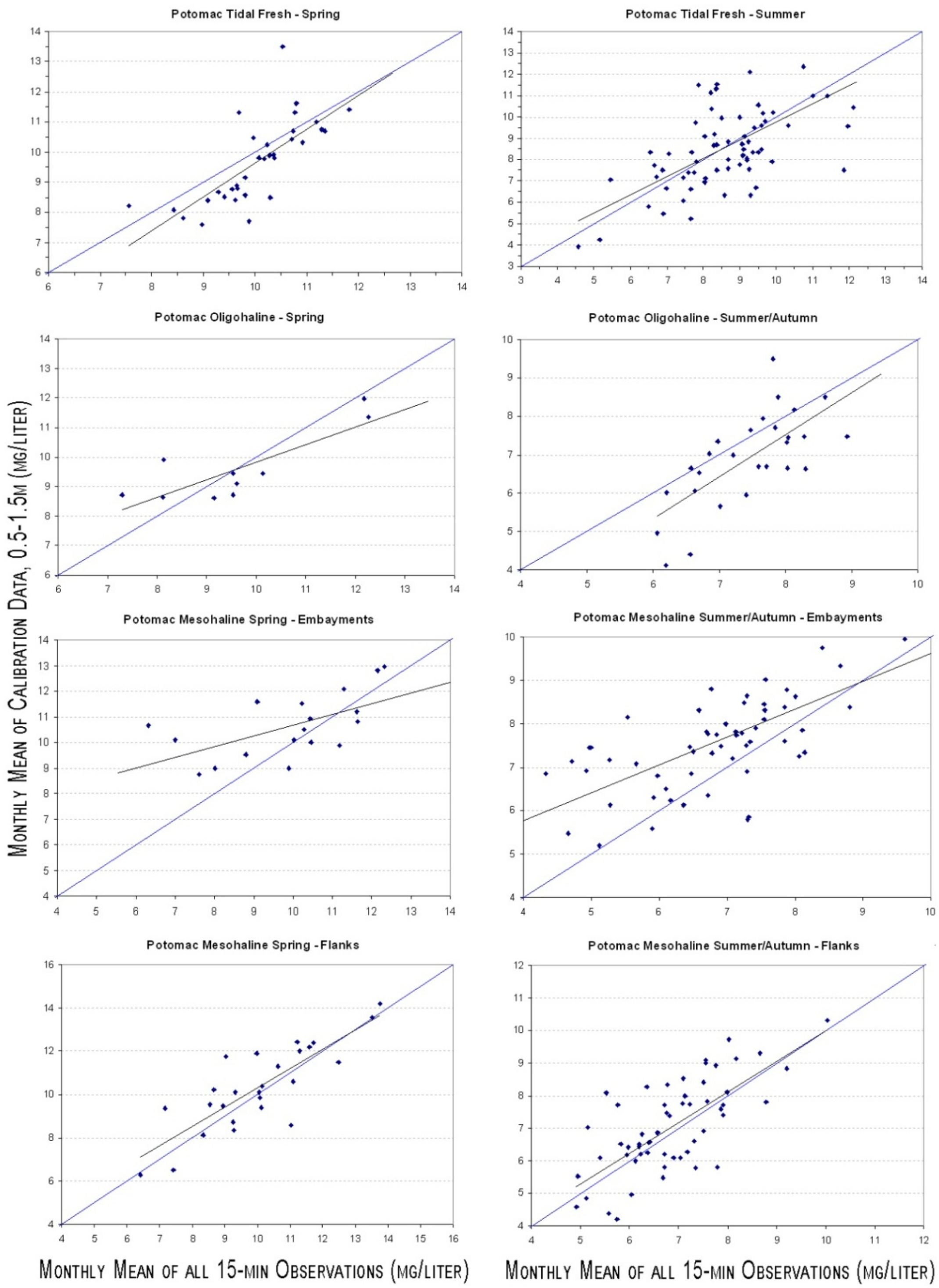


Figure 5. Paired comparison of the 30-day mean DO estimates calculated from low frequency (calibration) and high frequency (CMON) data. Blue line represents the 1:1 relationship, and it is assumed that the means derived from high frequency data (x-axis) are very close to the true mean. There are large differences between the paired 30-day means in all seasons and salinity zones, indicating inaccuracy on the part of the means derived from the low frequency data. Further, low frequency means appear to be biased downward in TF spring and OH summer/autumn, and biased upward in MH embayments.

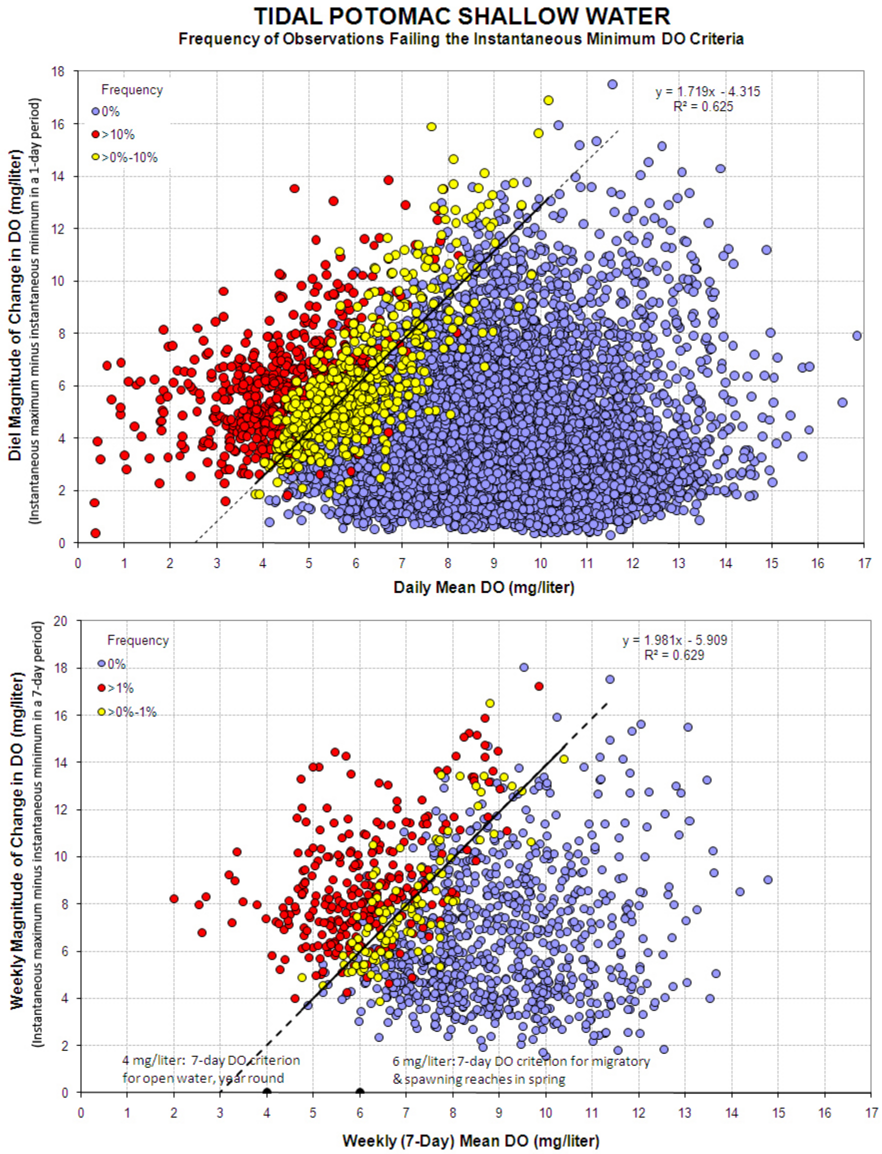


Figure 6. Attainment of the instantaneous minimum DO criteria as a function of daily (top) or weekly (bottom) mean DO and magnitude of change in DO. All tidal Potomac CMON data meeting the guidelines in Table 1 are included in each panel regardless of season or salinity zone. Top, n= 9,879 days; bottom, n= 668 weeks. The 7-day mean DO is calculated on sequential 7-day periods. The 7-day mean DO criteria presently applied to shallow waters are indicated on the x-axis in the bottom panel.

APPENDIX 5

A parametric simulation approach to assessing the Umbrella Concept for the instantaneous minimum criteria.

Elgin Perry 3/10/2011

High frequency samples of DO at fixed locations show that there is considerable serial dependence or autocorrelation in these DO time series. This lack of independence makes it difficult to analytically compute the probability that an instantaneous criteria will be violated when an umbrella criterion (e.g. weekly or monthly mean) is satisfied. Here we develop and show results from a simulation approach to addressing this question.

The basic approach of the simulation is to generate time series that have properties similar to observed DO time series. The data used for this exercise are the open water buoy data compiled by Olson. In these data, time series that are more than 1 week in length were parsed into 1 week time series. A simple AR(2) model that included structural terms for the mean, linear trend, and diel cycle was fitted to each of these time series using Proc AutoReg in SAS. Each fitting results in a vector of 7 parameters

b_int - the intercept which reflects the mean because other covariates are centered.

b_cday - linear trend term for the week fitted as a coefficient of centered day.

b_sin, b_cos - coefficients for diel trend fitted to trig-transformed time

b_ar1,b_ar2 - autoregressive terms at lags 1 and 2

mse - residual mean square error

These parameter estimates were obtained for each 1-week time series to yield 251 sets of parameters. These 251 vector observations were analyzed by Multivariate Analysis of Variance (MANOVA) using Proc GLM in SAS. The model included terms for Month, Total Water Depth, Sensor Depth, Latitude and Longitude. Some results from this overall model are presented.

For the simulation, only data from CB4 in the surface layer (sensor < 10 m depth) were used. A MANOVA model which included terms for Month, Total Water Depth, and Sensor Depth. Coefficients from this model were used to estimate a mean predicted value for the parameter vector which seeded the parametric simulation.   A multivariate normal random number generator (R-package) was used to generate 1000 realizations using this mean vector and the Variance-Covariance matrix estimate from the MANOVA. Each of these 1000 realization of the parameter vector were passed to a function which estimated a 1-week time series based on the simulated parameter vector values. The percent of violations of the instantaneous minimum criterion (3.2 ppt) were tabulated yielding 1000 estimates of this percentage. The range and frequency of these percentages are compared for various mean vectors associated with different conditions specified by different values of the independent variables in the MANOVA model.

Results:

When examining data from all buoy locations, in a multivariate sense, all of these terms are statistically significant (Table 1).

Table 1. Manova test results for dependent vector

(b_int,b_cday,b_sin,b_cos,b_AR1,b_AR2,mse).

Source	Pillai's Trace	Pr > F
month	0.2895	0.0191
TotDep	0.1018	0.0007
SampDep	0.2063	<.0001
lat	0.0592	0.0451
long	0.2102	<.0001

Table 2 shows which independent variables appeared to have an effect on which dependent variables.

Table 3. P-values for each manova term and for each dependent variable.

Source	b_int	b_cday	b_sin	b_cos	b_AR1	b_AR2	mse
month	0.0861	0.9041	0.3811	0.4845	0.0130	0.0909	0.1277
TotDep	<.0001	0.4168	0.9888	0.7560	0.1728	0.2066	0.1374
SampDep	<.0001	0.4214	0.0381	0.5415	0.1808	0.2711	0.0331

lat	0.2065	0.3651	0.2688	0.0563	0.9958	0.2387	0.1713
long	0.7956	0.0432	0.9265	0.9906	<.0001	0.2204	0.0290

Table 4. Coefficient estimates for covariates.

Source	b_int	b_cday	b_sin	b_cos	b_AR1	b_AR2	mse
TotDep	0.2224	0.0060	0.0001	-0.0031	-0.0106	0.0080	0.0148
SampDep	-0.4079	-0.0072	0.0309	-0.0074	0.0125	-0.0083	-0.0255
lat	-0.2449	0.0244	-0.0496	0.0703	0.0001	0.0271	0.0493
long	0.1058	-0.1157	0.0087	-0.0009	-0.3149	0.0595	0.1666

Notes from tables 3 and 4.

DO seems to improve as water depth increases.

DO degrades as sensor depth increases.

AR1 terms are stronger in the western bay

mse decreases with sensor depth

Table 5. Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r| DF = 239 .

	b_Int	b_cday	b_sin	b_cos	b_AR1	b_AR2	MSE
b_Int	1.000000	-.052225	-.116969	0.113032	0.252967	-.225183	-.078779
		0.4206	0.0705	0.0805	<.0001	0.0004	0.2240
b_cday	-.052225	1.000000	0.128183	-.019640	0.083167	-.026105	-.132840
	0.4206		0.0473	0.7621	0.1992	0.6874	0.0398
b_sin	-.116969	0.128183	1.000000	-.074374	-.296165	0.205687	0.020856
	0.0705	0.0473		0.2511	<.0001	0.0014	0.7479
b_cos	0.113032	-.019640	-.074374	1.000000	0.095132	-.089933	-.185441
	0.0805	0.7621	0.2511		0.1417	0.1649	0.0039
b_AR1	0.252967	0.083167	-.296165	0.095132	1.000000	-.816881	-.297462
	<.0001	0.1992	<.0001	0.1417		<.0001	<.0001
b_AR2	-.225183	-.026105	0.205687	-.089933	-.816881	1.000000	0.264092
	0.0004	0.6874	0.0014	0.1649	<.0001		<.0001
MSE	-.078779	-.132840	0.020856	-.185441	-.297462	0.264092	1.000000
	0.2240	0.0398	0.7479	0.0039	<.0001	<.0001	

Notes on Table 5.

Strongest correlation is among parameters that model the error process. The autoregressive terms b_AR1 and b_AR2 have an inverse dependence. The mse term is correlated with the AR terms and with b_cos and b_cday.

There is little correlation among terms that model the mean (i.e. b_int, b_cday, b_sin, b_cos) add tables and notes on just CB4 MANOVA.

CB4 - violation results-

Using the manova model for CB4 we can obtain a predicted value of the time series parameter vector as a function of month, water depth, and sensor depth. In this simuluation I have started with a choice of month, water depth, and sensor depth for which the mean DO is just greater than the 30 day mean criterion of 5.0.

The independent variable vector that yields this prediction is

May	Jun	Jul	Aug	Sep	Oct	WaterDepth	SensorDepth
0	0	1	0	0	0	10	6

for which the predicted vector of time series parameters is

b_Int	b_cday	b_sin	b_cos	b_AR1	b_AR2	mse
5.0058	-0.0493	-0.4072	-0.0527	0.9333	-0.0319	0.3164

This predicted vector and the estimated Variance-Covariance matrix is used to seed a multivariate normal random number generator that creates 1000 realizations of the time series parameter vector. A one week time series 15 minute observations is generated for each realization. The b_Int term of this predicted vector is the weekly mean of the one week time series. Based on the 15 minute observations, the percent of observations below the instantaneous minimum criterion is computed. The umbrella concept is assessed by comparing the true monthly mean (5.0058), the simulated weekly means (b_Int) in the 1000 realizations, and the violation rates of the instantaneous minimum.

By changing the SensorDepth of the independent variable vector, the longterm mean can be adjusted to assess the effect of this parameter on the relationship among the three criteria assessments. Thus by raising the sensor depth from 6m to 3m the mean DO is increased from 5.0058 to 7.0082 (Table ?). The time series parameters for diel signal and the mse term increase as well. The linear trend term and the AR terms remain fairly constant.

Sensor Depth	b_Int	b_cday	b_sin	b_cos	b_AR1	b_AR2	mse
6	5.0058	-0.0493	-0.4072	-0.0527	0.9333	-0.0319	0.3164
5	5.6733	-0.0476	-0.5114	0.0094	0.9328	-0.0294	0.4112
4	6.3408	-0.0460	-0.6156	0.0714	0.9324	-0.0268	0.5060
3	7.0082	-0.0443	-0.7198	0.1335	0.9320	-0.0243	0.6008

Results at 6.0

To compare violation rates of the 7-day criterion and instantaneous criterion I

crosstabulate cases where the 7 day mean < 4.0 against cases where the violation rate of the instantaneous minimum exceeds 10% in each 1 week time series.

Sensor Depth = 6 mean DO = 5.0058	7-day mean > 4.0	7-day mean < 4.0	marginal failure instantaneous minimum
failure Instantaneous minimum < 10%	520 62.35%	4 2.41%	524 52.4%
failure Instantaneous minimum > 10%	314 37.65%	162 97.59%	476 47.6%
marginal for failure of 7-day mean	834	166	1000

Sensor Depth = 5 mean DO= 5.6733	7-day mean > 4.0	7-day mean < 4.0	marginal failure instantaneous minimum
failure Instantaneous minimum < 10%	671 71.01%	4 7.27%	675 67.5%
failure Instantaneous minimum > 10%	274 28.99%	51 92.73%	325 32.5%
marginal for failure of 7-day mean	945	55	1000

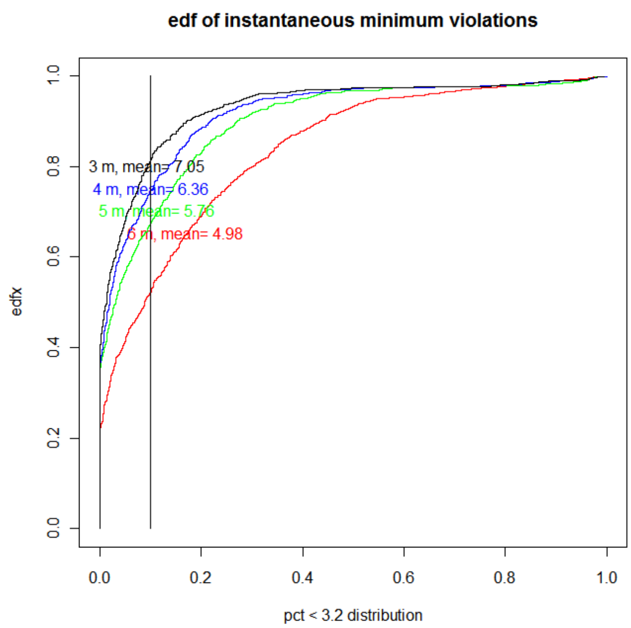
Sensor Depth = 4 mean = 6.3408	7-day mean > 4.0	7-day mean < 4.0	marginal failure instantaneous minimum
failure Instantaneous minimum < 10%	747 75.84%	0 0%	747 74.7%
failure Instantaneous minimum > 10%	238 24.16%	15 100%	253 25.3%
marginal for failure of 7-day mean	985	15	1000

Sensor Depth = 3 mean = 7.0082	7-day mean > 4.0	7-day mean < 4.0	marginal failure instantaneous minimum
failure Instantaneous minimum < 10%	815 81.91%	0 0%	815 81.5%
failure Instantaneous minimum > 10%	180 18.09%	5 100%	185 18.5%
marginal for failure of 7-day mean	995	5	1000

sensor depth	6	5	4	3
Monthly Mean DO	5.0058	5.6732	6.3407	7.0082
7 day criterion failure rate	16.6%	5.5%	1.5%	0.5%
rate of instantaneous criterion > 10%	47.6%	32.5%	25.3%	18.5%

When the long term mean DO is at a 'just passing' level, the simulation predicts that the 7-day mean criterion will be violated about 16.6% of weeks . If the long term mean DO increases to 5.7 then we expected fewer than 5.5% weeks with failure of the 7-day criterion. Thus if the 30-day mean criterion is satisfied, it is quite likely that violations of the 7-day mean criterion will be satisfied unless the 30 day mean hovers in the 'just passing' zone for an extended period.

Looking the violations of the instantaneous minimum is not so encouraging. When the long term mean is 'just satisfied', the simulation predicts that the instantaneous minimum criterion exceedance rate will exceed 10% in about 47% of weeks. Even when the long term mean DO is 7, the simulation predicts 18.5% of weeks will have an instantaneous minimum criterion exceedance rate in excess of 10%.



edf of instantaneous minimum violations
3 m, mean= 7.05
4 m, mean= 6.36
5 m, mean= 5.76
6 m, mean= 4.98
edfx
pct < 3.2 distribution
0.0
0.2
0.4
0.6
0.8
1.0

APPENDIX 6

ASSESSMENT OF SAMPLING VARIABILITY ON THE ABILITY OF THE 30-DAY MEAN DO CRITERION TO SERVE AS AN UMBRELLA FOR THE 7-DAY MEAN CRITERION.

Elgin Perry
eperry@chesapeake.net
Feb 26, 2013

Introduction

A sub-committee of TMAW reviewed the variability of the weekly mean DO about the monthly mean DO and concluded that in general, that if the 30-day DO criterion is satisfied by the monthly mean, then there is less than a 10% chance that the 7-day DO criterion will be violated by the weekly mean. This conclusion is based on having very accurate estimates of both the monthly mean and the weekly mean derived from near continuous high frequency observations of DO. However, in many parts of the bay, the monthly mean is estimated from as few as one to two point observations per month. Because the uncertainty of a monthly mean of two observations is much greater than the uncertainty of a monthly mean from near continuous data, it is reasonable to expect that effectiveness of the umbrella effect of the 30-day criterion for the 7-day criterion will diminish when the low sample size mean is employed. In this study, we examine the additional uncertainty that is created by the use of small sample size and further evaluate the consequences of this uncertainty for the umbrella concept.

Methods

This study evaluates the additional uncertainty from low sample sizes by resampling from near continuous records in a manner that simulates the twice monthly sampling of routine cruises. The near continuous DO records used are from the Potomac ConMon studies. For each calendar month in the May through September period of each record, a random day between 1 and 15 was chosen as the first sampling day of the month. To get a second sampling day, a random increment from 10 to 16 was generated and added to the first. In the event that there was no data on this second day, then the last day of the month with data was used. For each selected day, a random selection from the roughly 24 observations taken between 9:00 a.m. and 3:00 p.m. was chosen as the point estimate. These two estimates were summed and divided by 2 to obtain the monthly mean estimate. This simulation was repeated 20 times to obtain 20 monthly mean estimates for each station and month.

Months were calendar months, and weeks were designated as sequential weeks beginning January 1st of each year. Weekly means were computed for each unique combination of month and week. Thus if a month terminus divided a week, then the week was divided at this point and the resulting partial weeks were assigned to the two months. Deviations of the weekly means about the monthly mean were computed as (weekly mean DO – monthly mean DO) for weeks that occur within a month. In all cases, the weekly mean DO was computed as the mean of all high frequency observations within a week and is referred to as the near true weekly mean. The monthly mean was computed two ways. A near true estimate of the monthly mean uses all observations in the near continuous record; a small sample estimate of the monthly mean uses only two observations as described by the resampling methods above. The root mean square error (rmse) was computed across months, years, and stations for both the near true deviations and the small sample deviations. These root mean square estimates quantify the standard deviation of the weekly mean about the monthly mean for both the near true case and the small sample estimate case. The increase in the rmse for small sample case relative to the near true case illustrates the loss of precision in estimating the monthly mean by small samples.

Using these estimates of standard deviation and assuming a normal distribution for these deviations, we estimate the probability that the weekly mean is less than 4.0, the 7-day mean criterion, while the monthly mean is 5.0, the 30 day mean criterion. This probability is a measure of the efficacy of the 30-day criterion as an umbrella for the 7-day criterion.

Results:

Descriptive statistics for the true weekly deviations and the small sample deviations show a negative bias of small sample deviations relative to the true deviations (Table 1). This shows that the resampled monthly means which use daytime data only tend to be biased high, but on average this effect is not large. The range of the mean of the deviations over the resampling experiments is (-0.3428 -0.0133). The variability of the small sample deviations is much greater than that of the near true deviations. The true deviations have a rmse very close to 1.0 while the rmse from the small sample deviations always exceeds 1.6 and in one case exceeds 1.9 indicating a 60 to 90 percent increase in variability.

Summary of Monthly Mean DO vs. Weekly Mean DO resampling study results
Mon Feb 25 16:33:58 2013

Table 1.0 Summary of comparing weekly DO means to Monthly DO means for 'true' means and monthly means from 20 small sample resampling experiments.

Simu-lation	sample size	mean	rmse	minimum	q25	Median	q75	maximum
true	833	0.0017	1.005	-4.18	-0.4816	0.0125	0.4828	3.2042
1	833	-0.1344	1.6578	-5.1447	-0.9944	-0.0542	0.8052	4.9893
2	833	-0.0247	1.6903	-5.6588	-0.8519	0.0165	0.8543	4.4843
3	833	-0.2745	1.7132	-6.684	-1.1194	-0.1775	0.6852	4.4353
4	833	-0.2187	1.8037	-7.9388	-0.9968	-0.0879	0.7284	5.3265
5	833	-0.1723	1.8766	-8.2638	-0.9699	-0.0726	0.8603	4.9031
6	833	-0.0666	1.6177	-5.379	-0.8897	-0.0173	0.7745	4.6073
7	833	-0.2252	1.7196	-6.8519	-1.066	-0.2264	0.6948	5.3679
8	833	-0.0133	1.6054	-5.4517	-0.7627	0.0211	0.8046	5.1295
9	833	-0.3428	1.7471	-6.3008	-1.1947	-0.2999	0.5542	4.3745
10	833	-0.1639	1.7156	-7.3597	-1.0652	-0.1465	0.8385	4.7042
11	833	-0.0948	1.7555	-5.7288	-1.0169	-0.0054	0.8369	5.0334
12	833	-0.2193	1.9286	-7.2316	-1.0929	-0.0793	0.7621	5.5595
13	833	-0.2014	1.692	-6.5302	-1.0818	-0.0624	0.7351	5.1557
14	833	-0.1747	1.6198	-6.2597	-1.063	-0.1254	0.8021	3.9682
15	833	-0.1424	1.7216	-6.3428	-1.0468	-0.1171	0.8693	4.8051
16	833	-0.1055	1.7055	-6.114	-1.0153	0.0278	0.9094	4.3039
17	833	-0.1663	1.8126	-6.424	-1.1035	-0.107	0.7703	4.6611
18	833	-0.2157	1.8397	-6.3407	-1.1281	-0.1486	0.8262	5.2234
19	833	-0.0624	1.7048	-5.3103	-1.0217	-0.0165	0.8549	4.7011
20	833	-0.2306	1.7493	-8.2242	-1.1226	-0.1713	0.7209	4.2593

The distribution of the true weekly deviations tends to follow the normal distribution closely for the bulk of the observations (Figure 1.0). However, there is a small percentage of outliers at both the upper end and the lower end of the distribution that are more extreme than are expected for the normal distribution. Because of this heavy tailed feature of the true deviations, when the normal distribution is used to compute probabilities for this problem, these probabilities may be a slight underestimate of the true probabilities. There appear to be 10 to 15 extreme outliers in the lower tail of the distribution and thus the probability bias may be 1.2 to 1.8 percent.

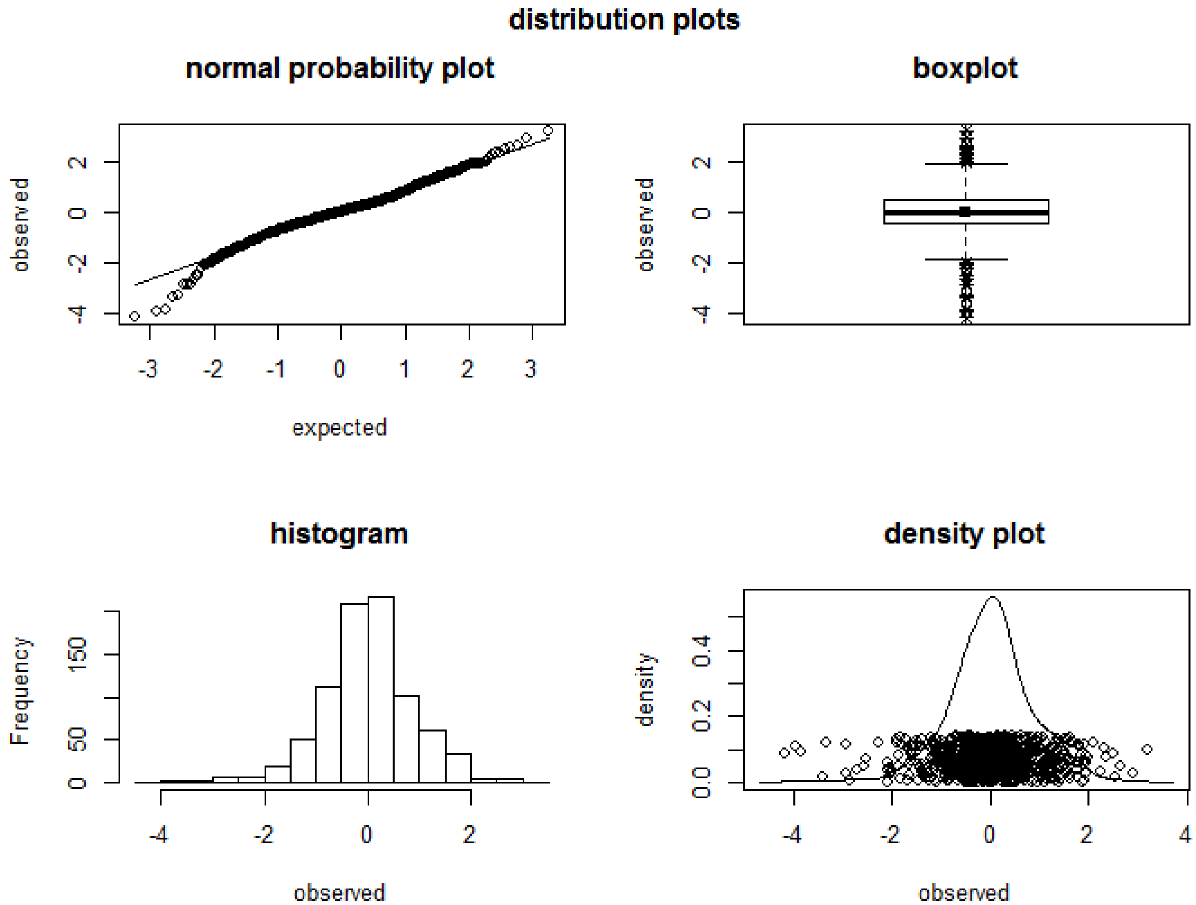


Figure 20.0 Distribution plots for the true weekly deviations.

The weekly deviations computed using the small sample monthly mean estimates appear to fit the normal distribution better than the true week deviations (Figure 2.0). The variability of deviations in the small sample experiment is clearly greater than variability for the true deviations. Compare for example the frequency of observation where the weekly mean is greater than 2 units below the monthly mean between Figures 1 and 2.

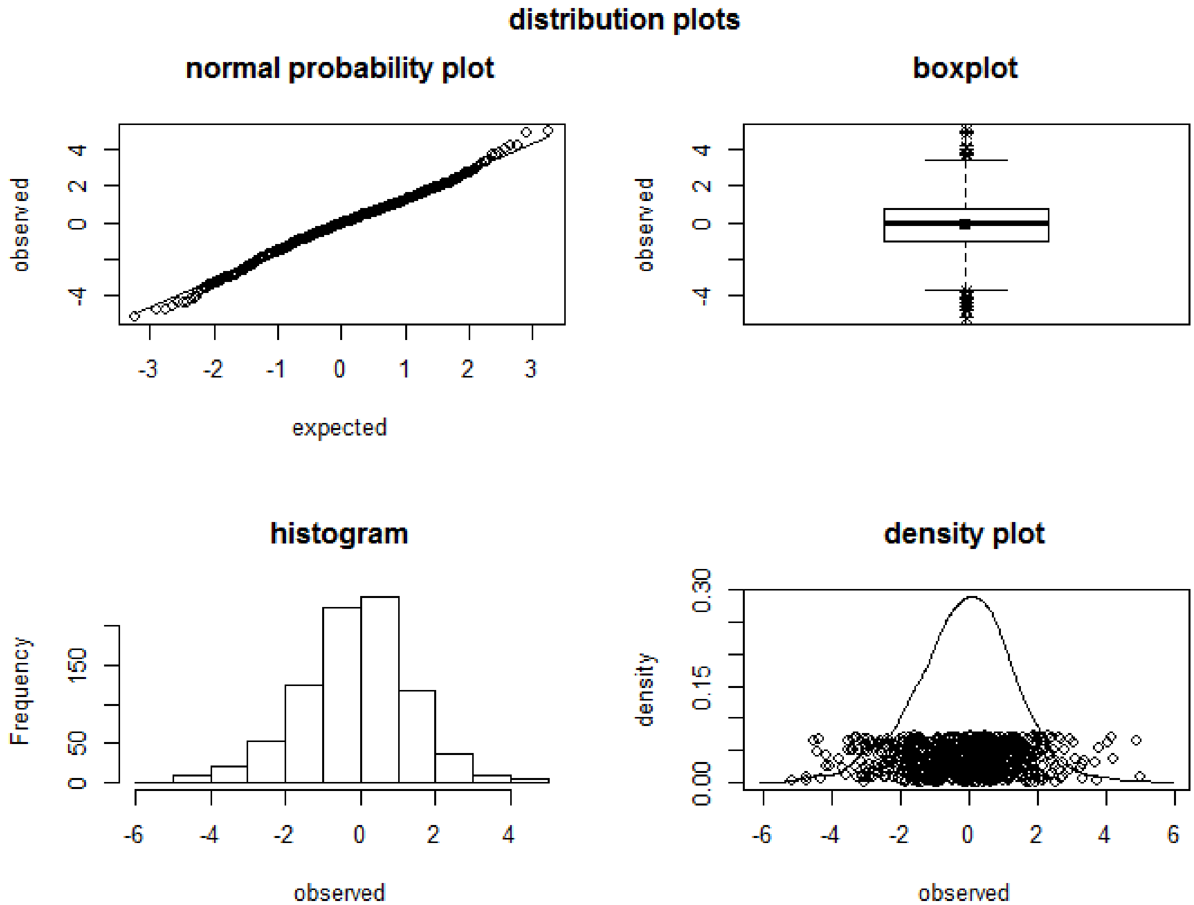


Figure 21. Distribution plots for weekly deviations computed for the first resampling experiment.

Discussion

If the 30-day criterion is to serve as an umbrella for the 7-day criterion, we would like to show that if the 30-day criterion is satisfied, there is a small probability that the 7-day criterion is violated. If we accept less than 10% as an acceptable risk of wrongly concluding that the 7-day criterion is satisfied when it is in fact violated, then this is satisfied when the standard deviation of the weekly mean from the monthly mean is .7805 or smaller. At this level of variability in the weekly deviations, excursions of the weekly mean below the 7-day criterion of 4.0 while the monthly mean is at the 30-day criterion of 5.0 would be about 10% (Figure 3). This scenario would be strong evidence that the 30-day criterion is an effective umbrella for the 7-day criterion.

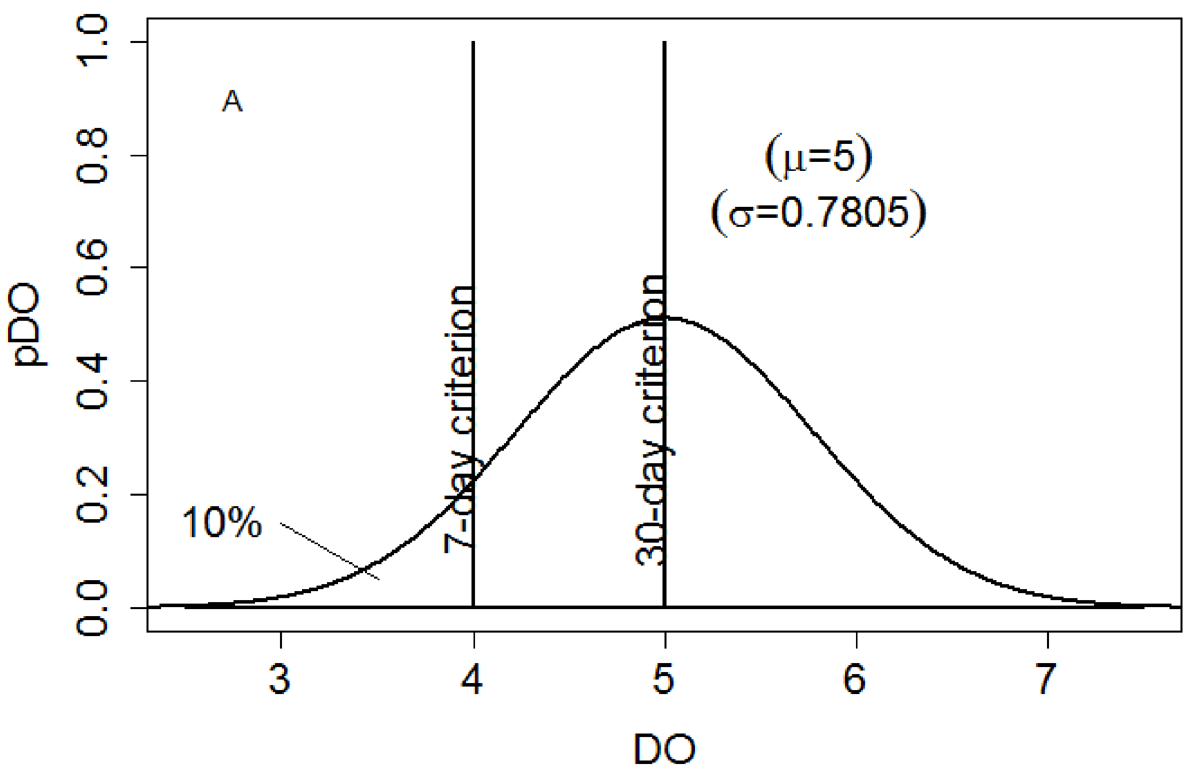


Figure 22. Illustration of the 30-day criterion as an umbrella for the 7-day criterion.

In the Potomac ConMon data, the standard deviation of the weekly mean from the monthly mean exceeds this ideal 0.7805 value and is estimated to be 1.005 or very close to 1.0. At this level of variability, the risk of violating the 7-day criterion when the 30-day criterion is satisfied exactly is about 16% (table 2., Figure 4.) However, increasing the monthly mean to 5.285 again brings the risk of violations of the 7-day criterion to an acceptable level of 10%. Because it is unlikely that the monthly mean will hover in this narrow window of (5.0, 5.285) for an extended time then it seems reasonable to consider that the 7-day criterion is satisfied if the 30-day criterion is satisfied. This evidence is one supporting fact for the TMAW conclusion that the 30-day criterion is an effective umbrella for the 7-day criterion. It is important to recognize that this conclusion assumes that both the true monthly mean and the true weekly mean can be estimated with great precision. This high level of precision is obtained here by using a near continuous record of DO.

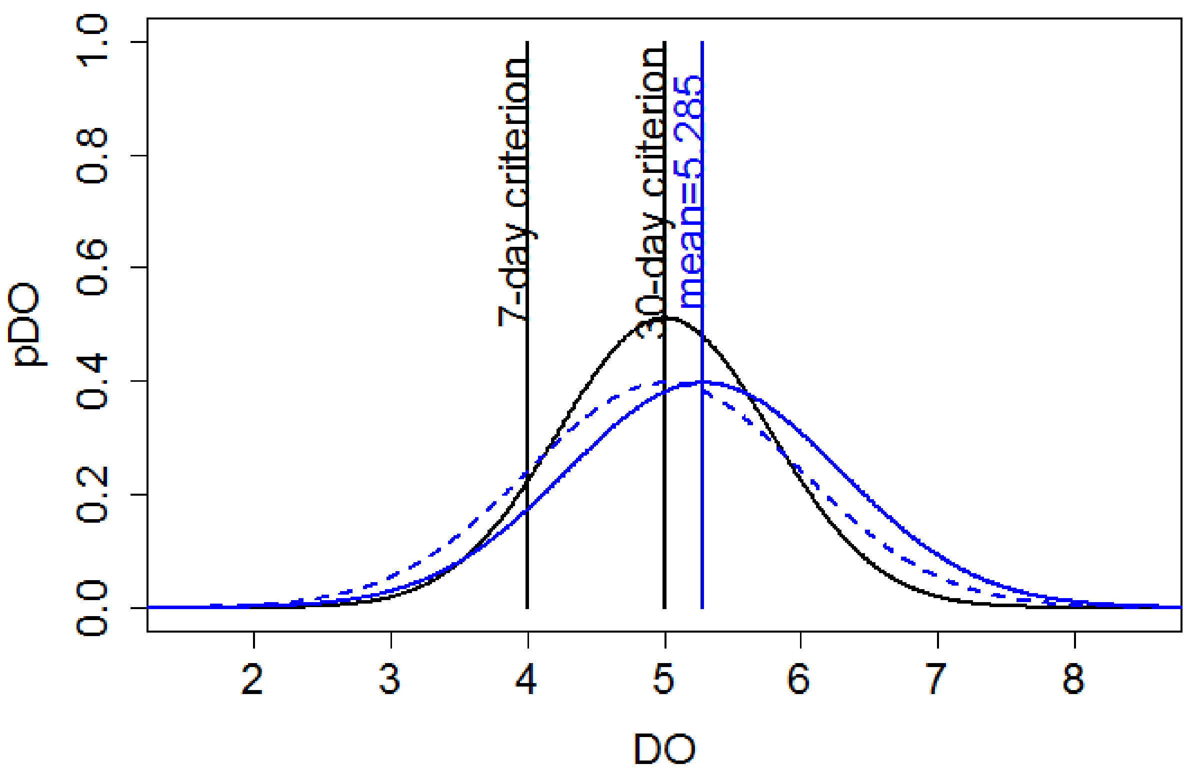


Figure 23. Illustration of the shift in the monthly mean required to meet 10% risk tolerance for the 7-day criterion when the weekly mean deviation is 1.005.

When the monthly mean is estimated by a sample size of two observations, then the variability of the deviations between the monthly mean estimate and the weekly means increases by 60 to 90 percent (table 3.0, Figure 5). At this higher level of variability, satisfying the 30-day criterion exactly results in a 28% risk of violating the 7-day criterion. Estimates of the monthly mean would have to exceed a threshold of 6.22 to insure that the risk of violating the 7-day criterion is 10% or less.

Table 2.0 Estimates of risk of violating the 7-day criterion given the monthly mean estimate (column 1) and four levels of sampling variation (columns 2-5). Column 1 assumes near true weekly deviations, column 2 assumes variation the average of 20 small sample estimates, column 3 assumes variation at the minimum of 20 small sample estimates and column 4 assumes variation at the maximum of 20 small sample estimtes.

	Risk of violating 7-day criterion			
Monthly Mean DO		SD=1.7358[2]	SD=1.6054[3]	SD=1.9287[4]
5.0	0.1598	0.2822	0.2666	0.3020
5.1	0.1368	0.2631	0.2466	0.2842
5.2	0.1162	0.2446	0.2273	0.2669
5.3	0.0979	0.2269	0.2090	0.2501
5.4	0.0818	0.2099	0.1915	0.2339
5.5	0.0677	0.1937	0.1750	0.2183
5.6	0.0556	0.1783	0.1594	0.2033
5.7	0.0453	0.1636	0.1448	0.1890
5.8	0.0366	0.1498	0.1311	0.1753
5.9	0.0293	0.1368	0.1183	0.1622
6.0	0.0232	0.1246	0.1064	0.1498
6.1	0.0183	0.1131	0.0954	0.1381
6.2	0.0142	0.1024	0.0852	0.1269
6.3	0.0110	0.0925	0.0759	0.1165
6.4	0.0084	0.0833	0.0674	0.1066
6.5	0.0064	0.0748	0.0597	0.0974

1 standard deviation of true weekly mean from true monthly mean
2 standard deviation base on pooling 20 resampling estimates
3 standard deviation based on minimum of 20 resampling estimates
4 standard deviation based on maximum of 20 resampling estimates

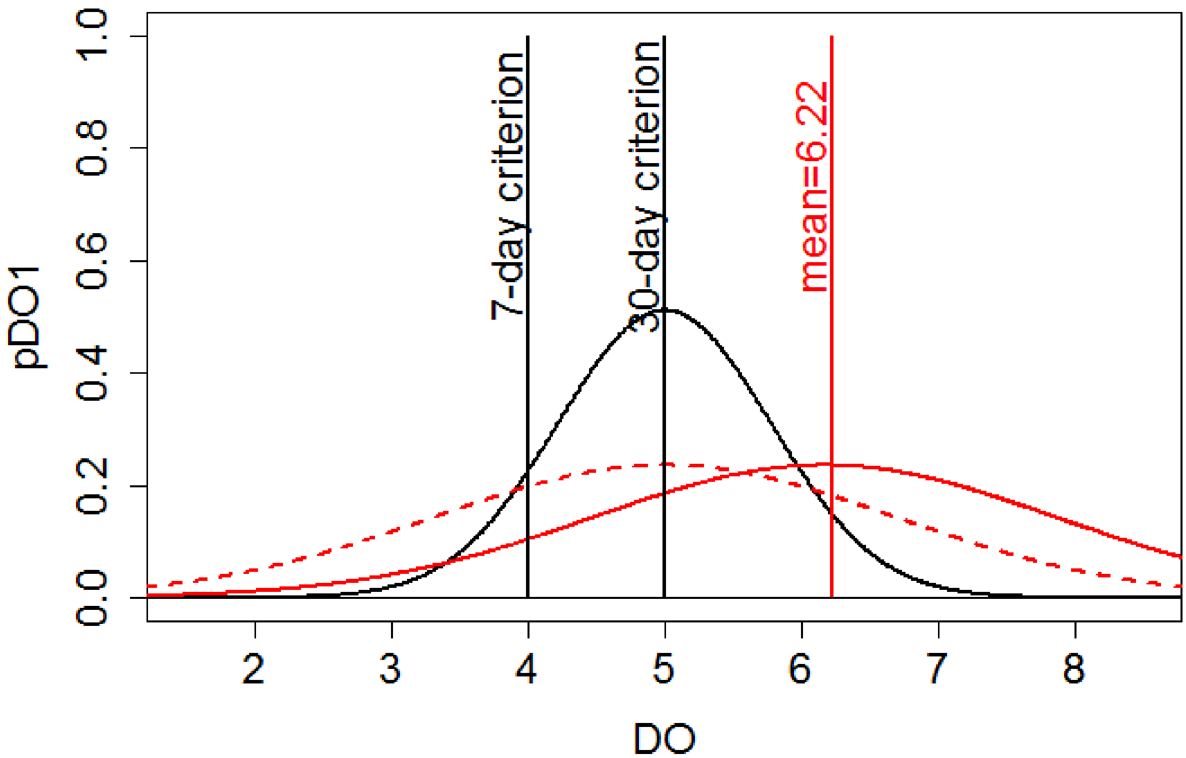


Figure 24. Illustration of the shift (from red dashed line to red solid line) in the monthly mean required to meet 10% risk tolerance for the 7-day criterion when the weekly mean deviation is at the small sample level of 1.74.

Conclusion:

Based on the evidence that the monthly mean threshold of 6.22 which insures that violations of the 7-day criterion is less than 10% is far above the 30-day criterion of 5.0, the 30-day criterion is not an umbrella for the 7-day criterion when the monthly mean is estimated by a sample size as small as two.