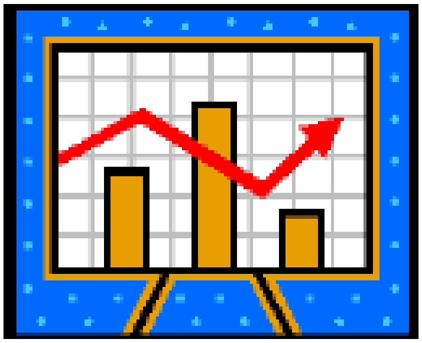
The Cumulative Frequency Diagram Method for Determining Water Quality Attainment

Report of the Chesapeake Bay Program STAC Panel to Review of Chesapeake Bay Program Analytical Tools



9 October 2006

Panel Members:

David Secor, Chair (Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science)
Mary Christman (Dept. of Statistics, University of Florida)
Frank Curriero (Departments of Environmental Health Sciences and Biostatistics, Johns Hopkins Bloomberg School of Public Health)
David Jasinski (University of Maryland Center for Environmental Science)
Elgin Perry (Statistics Consultant)
Steven Preston (US Geological Survey, Annapolis)
Ken Reckhow (Dept. Environmental Sciences & Policy Nicholas School of the Environment and Earth Sciences, Duke University)
Mark Trice (Maryland Department of Natural Resources)

STAC Publication 06-003

About the Scientific and Technical Advisory Committee

The Scientific and Technical Advisory Committee (STAC) provides scientific and technical guidance to the Chesapeake Bay Program on measures to restore and protect the Chesapeake Bay. As an advisory committee, STAC reports periodically to the Implementation Committee and annually to the Executive Council. Since it's creation in December 1984, STAC has worked to enhance scientific communication and outreach throughout the Chesapeake Bay watershed and beyond. STAC provides scientific and technical advice in various ways, including (1) technical reports and papers, (2) discussion groups, (3) assistance in organizing merit reviews of CBP programs and projects, (4) technical conferences and workshops, and (5) service by STAC members on CBP subcommittees and workgroups. In addition, STAC has the mechanisms in place that will allow STAC to hold meetings, workshops, and reviews in rapid response to CBP subcommittee and workgroup requests for scientific and technical input. This will allow STAC to provide the CBP subcommittees and workgroups with information and support needed as specific issues arise while working towards meeting the goals outlined in the *Chesapeake 2000* agreement. STAC also acts proactively to bring the most recent scientific information to the Bay Program and its partners. For additional information about STAC, please visit the STAC website at www.chesapeake.org/stac.

Publication Date:
October 2006

Publication Number:
06-003

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

STAC Administrative Support Provided by:
Chesapeake Research Consortium, Inc.
645 Contees Wharf Road
Edgewater, MD 21037
Telephone: 410-798-1283; 301-261-4500
Fax: 410-798-0816
http://www.chesapeake.org

The Cumulative Frequency Diagram Method for Determining Water Quality Attainment

Report of the Chesapeake Bay Program STAC Panel to Review of Chesapeake Bay Program Analytical Tools

9 October 2006

Panel Members:

David Secor, Chair (Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science)
Mary Christman (Dept. of Statistics, University of Florida)
Frank Curriero (Departments of Environmental Health Sciences and Biostatistics, Johns Hopkins Bloomberg School of Public Health)
David Jasinski (University of Maryland Center for Environmental Science)
Elgin Perry (Statistics Consultant)
Steven Preston (US Geological Survey, Annapolis)
Ken Reckhow (Dept. Environmental Sciences & Policy Nicholas School of the Environment and Earth Sciences, Duke University)
Mark Trice (Maryland Department of Natural Resources)

STAC Publication 06-003

Executive Summary

Background and Issues

In accordance with the Chesapeake 2000 Agreement, the Chesapeake Bay Program has recently implemented important modifications to (1) ambient water quality criteria for living resources and, (2) the procedures to determine attainment of those criteria. A novel statistical tool for attainment, termed the Cumulative Frequency Diagram (CFD) approach, was developed as a substantial revision of previous attainment procedures, which relied upon a simple statistical summary of observed samples. The approach was viewed as advantageous in its capacity to represent degrees of attainment in both time and space. In particular, it was recognized that the CFD could represent spatial data in a synoptic way: data that is extensively collected across diverse platforms by the Chesapeake Bay Program Water Quality Monitoring Program. Because the CFD approach is new to Bay Program applications, underlying statistical properties need to be fully established. Such properties are critical if the CFD approach is to be used to rigorously define regional attainments in the Chesapeake Bay.

In Fall 2005, the Chesapeake Bay Program Scientific, Technical and Advisory Committee charged our working group to provide review and recommendations on the CFD attainment approach. As terms of reference we used guidelines of Best Available Science recently published by the American Fisheries Society and the Estuarine Research Federation. Statistical issues that we reviewed included,

1. What are the specific analytical/statistical steps entailed in constructing CFD attainment curves and how are CFDs currently implemented? (Section 2)
2. How rigorous is the spatial interpolation process that feeds into the CFD approach? Would alternative spatial modeling procedures (e.g., kriging) substantially improve estimation of water quality attainment? (Section 3)
3. What are the specific analytical/statistical steps entailed in constructing CFD reference curves? (Section 4)
4. What are the statistical properties of CFD curves? How does sampling density, levels of attainment, and spatial covariance affect the shape of CFD curves? What procedures are reliable for estimating error bounds for CFD curves? (Section 5)
5. From a statistical viewpoint, does the CFD approach qualify as best available science? (Section 6)
6. What are the most important remaining issues and what course of directed research will lead to a more statistically rigorous CFD approach over the next three years? (Section 7)

The central element of our work was a series of exercises on simulated datasets undertaken by Dr. Perry to better evaluate 1) sample densities in time and space, 2) varying levels of attainment, and 3) varying degrees of spatial and temporal covariance. Further, trials of spatial modeling on fixed station Chesapeake Bay water quality data by Dr.s Christman and Curriero were conducted to begin to evaluate spatial modeling procedures. These exercises, literature review and discussions leading to consensus opinion are the basis of our findings. In August

2006, the working group supplied preliminary findings and related text for use in the 2006 CBP Addendum to Ambient Water Quality Criteria that is now under review.

Findings

1. **The CFD approach is feasible and efficient in representing water quality attainment.**

   The CFD approach can effectively represent the spatial and temporal dimensions of water quality data to support inferences on whether regions within the Chesapeake Bay attain or exceed water quality standards. The CFD approach is innovative but could support general application in water quality attainment assessments in the Chesapeake Bay and elsewhere. The CFD approach meshes well within the Chesapeake Bay Program's monitoring and assessment approaches, which have important conceptual underpinnings (e.g., segments defined by designated uses).

   In accepting the CFD as the best available approach for using time-space data, the panel contrasted it with the previous method and those sustained by other jurisdictions. The previous method used by the Chesapeake Bay Program, similar to the approaches used in other states, was simply based on EPA assessment guidance in which all samples in a given spatial area were compiled and attainment was assumed as long as > 10% of the samples did not exceed the standard. In this past approach all samples were assumed to be fully representative of the specified space and time and were simply combined as if they were random samples from a uniform population. This approach was necessary at the time because the technology was not available for a more rigorous approach. But it neglected spatial and temporal patterns that are known to exist in the standards measures. The CFD approach was designed to better characterize those spatial and temporal patterns and weight samples according to the amount of space or time that they actually represent.

2. **CFD curves are influenced by sampling density and spatial and temporal covariance. These effects merit additional research. Conditional simulation offers a productive means to further discover underlying statistical properties and to construct confidence bounds on CFD curves, but further directed analyses are needed to test the feasibility of this modeling approach.**

   The panel finds that the CFD approach in its current form is feasible, but that additional research is needed to further refine and strengthen it as a statistical tool. The CFD builds on important statistical theory related to the cumulative distribution function and as such, its statistical properties can be simulated and deduced. Through conditional simulation exercises, we have also shown that it is feasible to construct confidence ellipses that support inferences related to threshold curves or other tests of spatial and temporal compliance. Work remains to be done in understanding fundamental properties of how the CFD represents likely covariances of attainment in time and space and how temporal and spatial correlations interact with sample size effects. Further, more work is needed in analyzing biases across different types of designated use segments. The panel expects that a two-three year time frame of directed research and development will be required to identify and measure these sources of bias and imprecision in support of attainment determinations.

3. **The success of the CFD-based assessment will be dependent upon decision rules related to CFD reference curves. For valid comparisons, both reference and attainment CFDs should be underlain by similar sampling densities and spatial covariance structures.**

   CFD reference curves represent desired segment-designated use water quality outcomes and reflect sources of acceptable natural variability. The reference and attainment curves follow the same general approach in derivation: water quality data collection, spatial interpolation, comparison to biologically-based water quality criteria, and combination of space-time attainment data through a CFD. Therefore, the biological reference curve allows for implementation of threshold uncertainty as long as the reference curve is sampled similarly to the attainment curve. Therefore, we advise that similar sample densities are used in the derivation of attainment and reference curves. As this is not always feasible, analytical methods are needed in the future to equally weight sampling densities between attainment and reference curves.

4. **In comparison with the current IDW spatial interpolation method, kriging represents a more robust method and was needed in our investigations on how spatial covariance affects CFD statistical inferences. Still, the IDW approach may sufficiently represent water quality data in many instances and lead to accurate estimation of attainment. A suggested strategy is to use a mix of IDW and kriging dependent upon situations where attainment was grossly exceeded or clearly met (IDW) versus more-or-less "borderline" cases (kriging).**

   The current modeling approach for obtaining predicted attainment values in space is Inverse Distance Weighting (IDW), a non-statistical spatial interpolator that uses the observed data to calculate a weighted average as a predicted value for each location on the prediction grid. IDW has several advantages. It is a spatial interpolator and in general such methods have been shown to provide good prediction maps. In addition, it is easy to implement and automate because it does not require any decision points during an interpolation session. IDW also has a major disadvantage – it is not a statistical method that can account for sampling error.

   Kriging is also a weighted average but first uses the data to estimate the weights to provide statistically optimal spatial predictions. As a recognized class of statistical methods with many years of dedicated research into model selection and estimation, kriging is designed to permit inferences from sampled data in the presence of uncertainty. Thus the quantity and distribution of the sample data are reflected in those inferences. Indeed, the panel's initial trials on the role of spatial sources of error in the CFD have depended upon the ability to propagate kriging interpolation uncertainty through the CFD process in generating confidence intervals of attainment.

In comparison to IDW, kriging is more sophisticated but requires greater expertise in implementation. Kriging is available in commercial statistical software and also in the free open source R Statistical Computing Environment, and requires geostatistical expertise and programming skills for those software packages. Segment by segment variogram estimation and subsequent procedures would require substantial expert supervision and decision-making. Thus, this approach is not conducive to automation. On the other hand, there may be CBP applications where the decision on attainment is clearly not influenced to any substantial degree by the method of spatial interpolation. One suggested strategy is to use a mix of IDW and kriging - dependent upon situations where attainment was grossly exceeded or clearly met (IDW) versus more-or-less “borderline” cases (kriging).

5. **More intensive spatial and temporal monitoring of water quality will improve the CFD approach but will require further investigations on the influence of spatial and temporal covariance structures on the shape of the CFD curve. This issue is relevant in bringing 3-dimensional interpolations and continuous monitoring streams into the CFD approach.**

   In the near future, the panel sees that the CFD approach is particularly powerful when linked to continuous spatial data streams made available through the cruise-track monitoring program, and the promise of continuous temporal data through further deployment of remote sensing platforms in the Chesapeake Bay (Chesapeake Bay Observing System: http://www.cbos.org/). These data sets will support greater precision and accuracy in both threshold and attainment determinations made through the CFD approach but will require directed investigations into how data covary over different intervals of time and space. Further, there may be important space-time interactions that confound the CFD attainment procedure.

   Some of the assessments for the Bay such as that for dissolved oxygen require three dimensional interpolation, but the field of three dimensional interpolation is not as highly developed as that of two dimensional interpolation. Kriging can be advantageously applied in that it can use information from the data to develop direction dependent weighted interpolations (anisotropy). Kriging can include covariates like depth. Options for implementing 3-D interpolation include: custom IDW software, custom kriging software using GMS routines, or custom kriging software using the R-package.

Recommendations

The panel identified critical research tasks that need resolution in the near future. The following is a list of critical aspects of that needed research. These research tasks appear roughly in order of priority. However, it must be recognized that it is difficult to formulate as set of tasks that can proceed with complete independence. For example, research on task 1 may show that the ability to conditionally simulate the water quality surface is critical to resolving the sample size bias issue. This discovery might eliminate IDW as a choice of interpolation under task 3. The Panel has made significant progress on several of these research tasks and CBP is encouraged to implement continued study in a way that maintains the momentum established by our panel.

Task

1. Effects of Sampling Design on CFD Results

(a) Continue simulation work to evaluate CFD bias reduction via conditional simulation.
(b) Investigate conditional simulation for interpolation methods other than kriging - this may lead to more simulation work.
(c) Implement and apply interpolation with condition simulation on CBP data.

2. Statistical inference framework for the CFD

(a) Conduct confidence interval coverage experiments.
(b) Investigate confidence interval methods for non-kriging interpolation methods.
(c) Implement and evaluate confidence interval procedures.

3. Choice of Interpolation Method

(a) Implement a file system and software utilizing kriging interpolation for CBP data.
(b) Compare interpolations and CFDs based on kriging and inverse distance weighting (IDW).
(c) Investigate nonparametric interpolation methods such as LOESS and spline approaches.

4. Three-Dimensional Interpolation

(a) Implement 2-D kriging in layers to compare to current approach of 2-D IDW in layers.
(b) Conduct studies of 3-D anisotrophy in CBP data.
(c) Investigate software for full 3-D interpolation.

5. High Density Temporal Data

(a) Develop methods to use these data to improve temporal aspect of CFD implementation.
(b) Investigate feasibility of 4-Dimensional interpolation.

Table of Contents

1. Introduction

In June 2000, Chesapeake Bay Program (CBP) partners adopted the Chesapeake 2000 agreement (http://www.chesapeakebay.net/agreement.htm), a strategic plan that calls for defining the water quality conditions necessary to protect aquatic living resources. These water quality conditions are being defined through the development of Chesapeake Bay specific water quality criteria for dissolved oxygen, water clarity, and chlorophyll_a to be implemented as state water quality standards by 2005. One element of the newly defined standards is an assessment tool that addresses the spatial and temporal variability of these water quality measures in establishing compliance. This tool has become known as the Cumulative Frequency Diagram (CFD).

The (CFD) was first proposed as an assessment tool by Paul Jacobson, of Langhei Ecology (www.LangheiEcology.com). At that time Dr. Jacobson was consulting with the Chesapeake Bay Program as a member of the Tidal Monitoring Network Redesign Team. Within this group, the CFD concept gained immediate recognition and support as a novel approach that permitted independent modeling of the time and space dimensions of the continuous domain that underlies Chesapeake Bay water quality parameters. In addition, because preparation of the CFD uses spatial interpolation, the approach can allow integration of data collected on different spatial scales such as fixed station data and cruise track data.

While the benefits of the CFD approach has been recognized (U.S. EPA 2003) and the the CBP has begun implementation of the approach for certain water quality parameters and segments of the Chesapeake Bay, investigations of the statistical properties revealed that the underlying shape parameters of the CFD were sensitive not only to rates of compliance but also to sampling design elements such as sample density. The novelty of the approach coupled with concerns about its statistical validity motivated the Chesapeake Bay Program to request that its Scientific and Technical Advisory Committee (http://www.chesapeake.org/stac/) empanel a group with expertise in criteria assessment, spatial data interpolation, and statistics to assess the scientific defensibility of the CFD. Here we report the findings of this panel.

The primary goal of this panel is to provide an initial scientific review of the CFD compliance approach. This review addresses a wide range of issues including: bias and statistical rigor, uncertainty, practical implementation issues, and formulation of reference curves. Because of the novelty of the CFD approach, the panel has endeavored to research and explain the properties of the CFD and spatial modeling upon which the CFD approach depends to provide a basis for this evaluation. These activities are beyond the scope of the typical review. However, because so little is known about the CFD, it was necessary to expand the knowledge base.

The report is organized into 7 sections. In Section 2 of this report we present the CFD approach as a series of steps, each of which needs to be considered carefully in evaluating its statistical properties. Spatial interpolation is a critical but the most statistically nuanced step in the CFD approach. Spatial interpolation of water quality data in the CBP has to date received little statistical review. In Section 3 we evaluate alternative geostatistical methods as they pertain to the CFD approach. The CFD approach is an attainment procedure, which depends upon statistical comparison between attainment and reference curves. In Section 4, we present alternative types of references curves and discuss statistical properties of each. In Section 5 the statistical properties of CFD curves (applicable to both attainment and reference curves) is elucidated through a series of conditional simulation trials.

In addition to this primary charge, the panel is sensitive to the fact that the CFD will be employed in the enforcement of water quality standards. Use as a regulatory tool imposes a standard of credibility, which we review in Section 6. We use here “best available science” and “best science” criteria to evaluate the overall validity and feasibility of the CFD approach, following guidelines established by the American Fisheries Society and Estuarine Research Federation (Sullivan et al. 2006). These follow other similar criteria (e.g., The Daubert Criteria (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993) and include:

1. A clear statement of objective
2. A conceptual model, which is a framework for characterizing systems, sating assumptions, making predictions, and testing hypotheses.
3. A good experimental design and a standardized method for collecting data.
4. Statistical rigor and sound logic for analysis and interpretation.
5. Clear documentation of methods, results, and conclusions
6. Peer review.

The panel has made progress in better understanding statistical properties of the CFD approach and overall, we recommend it as a feasible approach and one that qualifies under most criteria for best available science. Still, we believe that our efforts should only represent the beginning of a longer term effort to (1) Use simulations and other means to support statistical comparisons of CFD curves; and (2) Support the CBP’s efforts to model water quality data with sufficient rigor in both spatial and temporal dimensions. Research and implementation recommendations follow in Section 7

2.0 Background
2.1 The CFD assessment approach.

The water quality criteria assessment methodology currently proposed by the E.P.A. Chesapeake Bay Program (CBP) involves the use of a Cumulative Frequency Diagram (CFD) curve. This curve is represented in a two dimensional plane of percent time and percent space. This document briefly discusses the reasoning that lead to the development of this assessment tool. The proposed algorithm for estimating the CFD is given and illustrated with small data sets. Some properties and unresolved issues regarding the use of the CFD are briefly discussed. In Section 5, simulation studies explore in greater specificity the multiple issues related to error and bias in the CFD approach.

Reasoning behind the CFD Approach

The CFD assessment methodology evolved from a need to allow for variability in water quality parameters due to unusual events. For the water quality parameter to be assessed, a threshold criterion is established for which it is determined that water quality that exceeds this threshold is in a degraded state (For simplicity, we will speak of exceeding the threshold as representing degradation, even though for some water quality constituents such as dissolved oxygen, it is falling below a threshold that constitutes degradation). Because all water quality parameters are inherently variable in space and time, it is unlikely that a healthy bay will remain below the threshold in all places at all times. In the spatial dimension, there will be small regions that persistently exceed the threshold due to poor flushing or other natural conditions. It is recognized by CBP that these small regions of degraded condition should not lead to a degraded assessment for the segment surrounding this small region. Similar logic applies in the temporal dimension. For a short period of time, water quality in a large proportion of a segment may exceed the threshold, but if this condition is short lived and the segment quickly returns to a healthy state, this does not represent an impairment of the designated use of the segment. Recognition that ephemeral exceedances of the threshold in both time and space do not represent persistent impairment of the segment leads to an assessment methodology that will allow these conditions to be classed as acceptable while conditions of persistent and wide spread impaired condition will be flagged as unacceptable. The assessment methodology should first ask how much of the segment (for simplicity, a spatial assessment unit is called a segment, but more detail is given on spatial assessment units in Section 2) is not in compliance with the criteria (percent of space) for every point in time. In a second step the process should ask how often (percent of time) is a segment out of compliance by more than a fixed percent of space. The results from these queries can be presented in graphical form where percent of time is plotted against percent of space (Figure 2.1). It is arbitrary to treat space first and time second. A similar diagram could be obtained by first computing percent noncompliance in time and then considering the cumulative distribution of percent time over space.

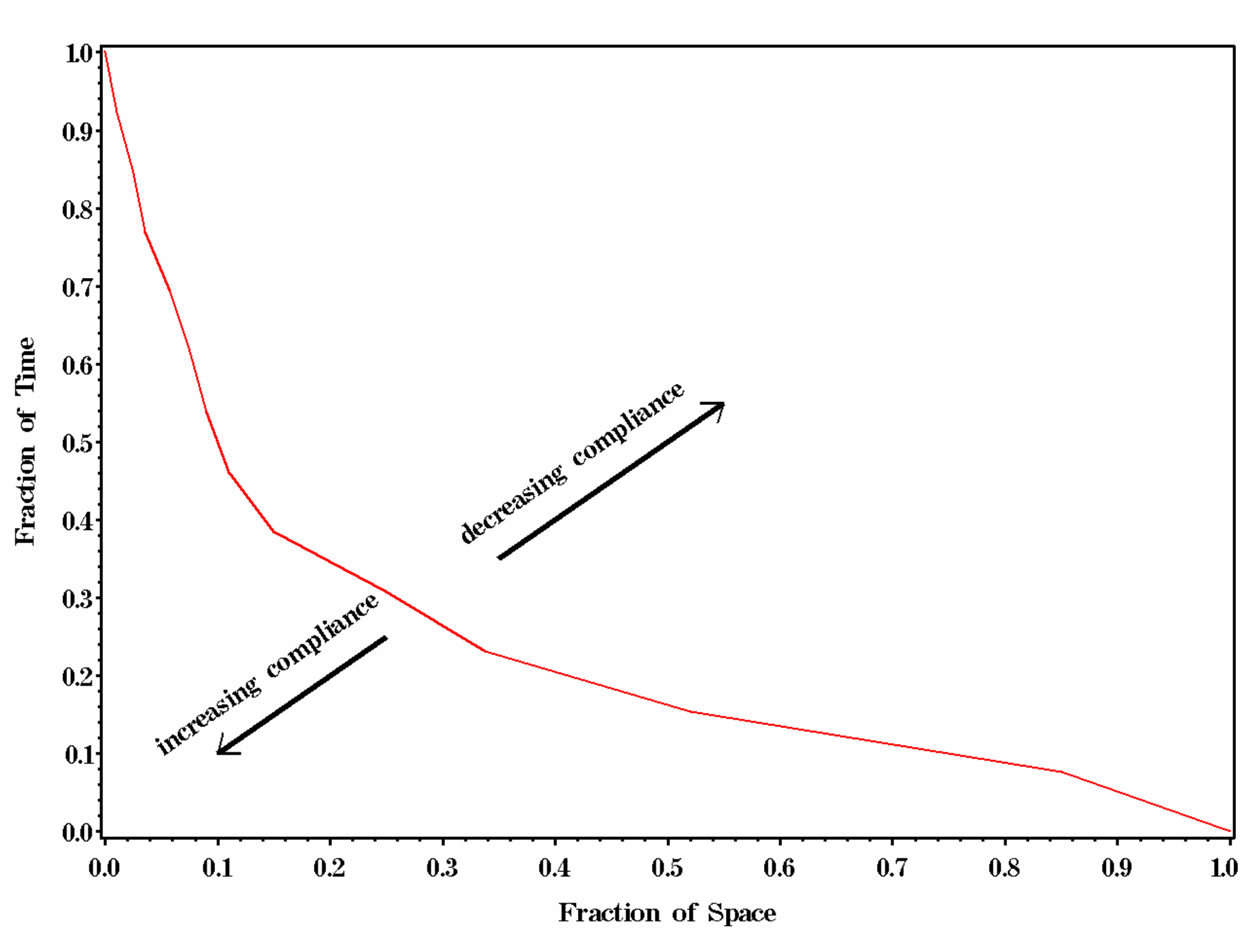


Figure 2.1 Illustration of CFD for 12 dates

If a segment is generally in compliance with the criterion, then one expects a high frequency of dates where the percent out of compliance is low. In this case, the CFD should descend rapidly from the upper left corner and pass not too far from the lower left corner and then proceed to the lower right corner. The trace in Figure 2.1 shows the typical hyperbolic shape of the CFD. The closer the CFD passes to the origin (lower left corner), the better the compliance of the segment being assessed. As the CFD moves away from the origin, a higher frequency of large percents of space out of compliance is indicated.

Formulating an Estimate of the CFD.

The algorithm developed by CBP for estimating the CFD is most easily described as a series of steps. These steps are given in bullet form to provide a frame work for the overall approach. The quickly defined framework is followed by a simple example. This in turn is followed by more detailed discussion of each step.

The steps:

1. Collect data from a spatial network of locations on a series of dates in a three year assessment period .

2. For each date, interpolate the data for the entire system (e.g. mainstem bay) to obtain estimates of water quality in a grid of interpolation cells.

3. For each interpolation cell assess whether or not the criterion is exceeded.

4. For each assessment unit (e.g. segment), compute the percentage of interpolator cells that exceed the criterion as an estimate of the percent of area that exceeds the criterion.

5. Rank the percent of area estimates for the set of all sample days in the assessment period from largest to smallest and sequentially assign to these ranked percents a value that estimates percent of time.

6. Plot the paired percent of time and percent of area data on a graph with percent of area on the abscissa and percent of time on the ordinate. The resulting curve is the Cumulative Frequency Diagram.

7. Compare the CFD from a segment being assessed to a reference CFD. If at any point the assessment CFD exceeds the reference CFD, that is, a given level of spatial noncompliance occurs more often than is allowed, then the segment is listed as failing to meet it's designated use.

Simple Numerical CFD Example:

For this example, assume a segment for which the interpolation grid is 4 cells by 4 cells. In reality, the number of grid cells is much larger. Also let data be collected on 5 dates. Typically data would be monthly for a total of 36 dates. Let the criterion threshold for this fictitious water quality parameter be 3. In what follows, you will find an illustration of the steps of computing the CFD for these simplified constraints. The three columns of the next page show the first three steps. Column 1 shows fictional data for five dates for five fixed locations in a 2 dimensional grid. Column 2 shows a fictional interpolation of these data to cover the entire grid. Column 3 shows the compliance status of each cell in the grid where 1 indicates noncompliance and 0 indicates compliance.

Step 1. Collect data at known locations.

date 1

3			3
		5	
2			1

date2

1			1
		3	
1			1

date3

4			2
		2	
1			1

date4

1			4
		2	
4			1

date5

1			3
		2	
1			1

Step 2. Interpolate the data to grid cells.

date 1

3	4	5	3
4	4	5	2
3	3	4	1
2	3	3	1

date2

1	2	3	1
2	2	3	2
1	3	2	1
1	1	1	1

date3

4	3	2	2
3	2	2	1
2	2	1	1
1	1	1	1

date4

1	2	3	4
2	2	2	3
3	3	2	1
4	3	1	1

date5

1	2	3	3
2	2	2	2
1	1	1	1
1	1	1	1

Step 3. Determine compliance status of each cell.

date 1

1	1	1	1
1	1	1	0
1	1	1	0
0	1	1	0

date2

0	0	1	0
0	0	1	0
0	1	0	0
0	0	0	0

date3

1	1	0	0
1	0	0	0
0	0	0	0
0	0	0	0

date4

0	0	1	1
0	0	0	1
1	1	0	0
1	1	0	0

date5

0	0	1	1
0	0	0	0
0	0	0	0
0	0	0	0

Step 4: Percent compliance by date.

sample date	percent space
date 1	75.00%
date 2	18.75%
date 3	18.75%
date 4	43.75%
date 5	12.50%

Step 5. Rank the percent of space values and assign percent of time = (100*R/(M+1.0)), where R is rank and M is total number of dates.

sample date	ranked percent space	cumulative percent time
date 1	75.00%	16.67
date 4	43.75%	33.33
date 2	18.75%	50.00
date 3	18.75%	66.67
date 5	12.50%	83.33

Steps 6 and 7: The plot of the CFD and the comparison to the reference curve are shown in Figure 2.2. For this hypothetical case the assessment area would be judged in noncompliance. For a percent area of 18.75, the allowable frequency on the reference curve is about 53%. That is, 18.75% of the segment area should not be out of compliance more that 53% of the time. For date 3, the estimated frequency of 18.75% noncompliance is 66.67%. Thus the frequency of 18.75% of space out of compliance is in excess of the 53% allowed. The reference curve is exceeded for dates 4 and 1 as well. Note: in this cumulative distribution framework, the actual date is not relevant. One should not infer that noncompliance occurred on that date if the data point associated with a date falls above the reference. Date is being used here as a label for each coordinate pair.

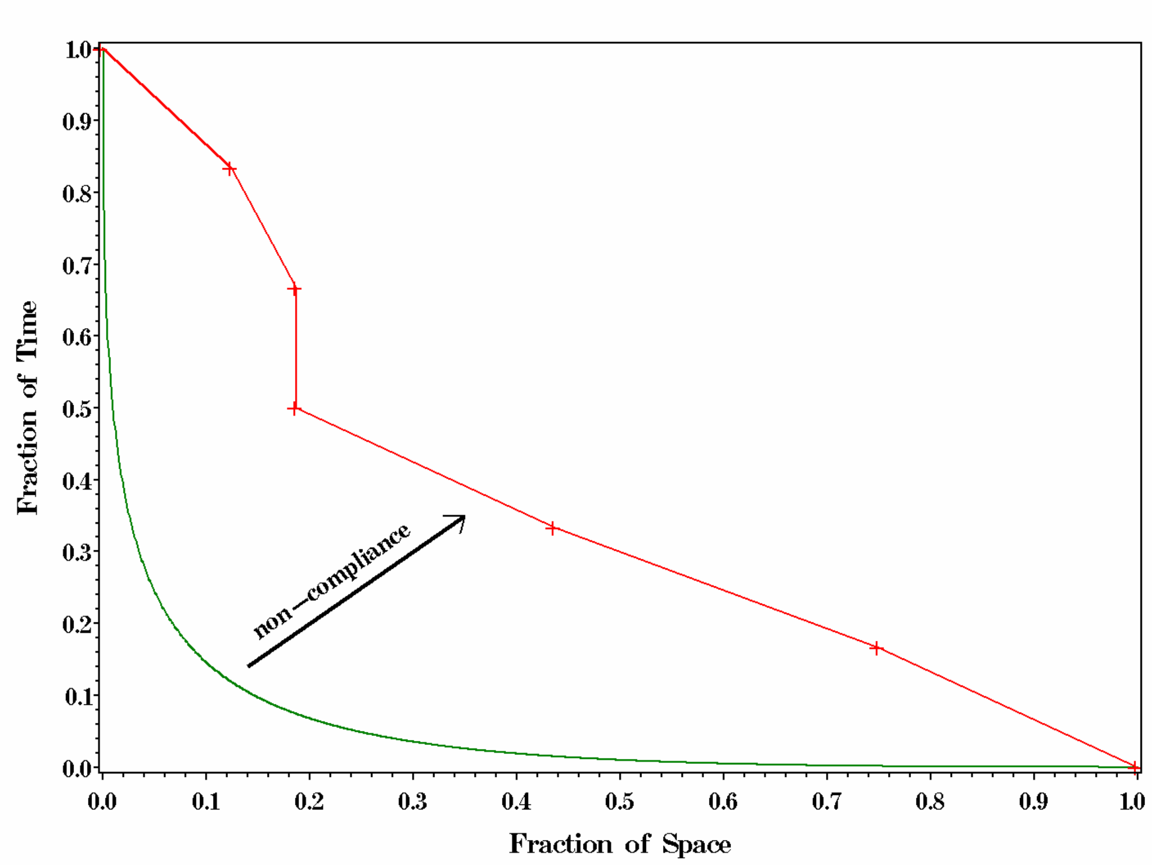


Figure 2.2 Graphical representation of CFD from the above example (red, '+') with hypothetical reference curve (green, smooth).

Defining the CFD Ideal

As defined above, the CFD is a data driven formulation. But the data used to formulate the CFD are a sample of points taken from a population. Defining the CFD becomes complex when one considers the many different levels for which it might be defined. At one level, the CFD might be defined based on the true state of a segment. Imagine that the state of a segment could be frozen for sufficient time to permit deployment of an analog sampler (that is one that measures water quality continuously rather than in discrete samples) to assess the percent of area out of compliance at that instant. Now stretch that imagination one step further to relax the condition that the segment be frozen and allow that these analog measurements of percent of area out of compliance be determined continuously in time. With this information, a determination of the CFD for the true state of the segment is possible. While the information needed to construct the ideal CFD is not obtainable, it is important to ask how well the CFD based on obtainable data represents this ideal (see also Section 5). Is a data driven CFD consistent for the ideal CFD in the statistical sense? Loosely speaking, consistency implies that the data driven CFD should get closer to the ideal CFD as more data are used. Is the data driven

CFD unbiased for the ideal CFD? Unbiasedness implies that even with small amounts of data, the data driven CFD on average covers the ideal CFD.

One might argue that if both the assessment CFD and the reference CFD are data driven, then it is not important for the CFD to approximate the ideal. Even so, it is important to understand the behavior of the CFD as a function of samples size and the relative temporal and spatial contributions to the variance in the water quality parameter. If the curve changes shape as a more data are used, this could result in unfair comparisons between assessment and reference regions. In Section 4, statistical properties for both types of reference curves are evaluated further.

Defining Reference Curves

Two approaches to defining the reference curve are being considered. One is a biologically based definition. The idea is to identify appropriate reference regions with healthy biological indicators and compute the reference CFD for these regions. For example, healthy benthic IBI scores might be used as indicators of adequate bottom dissolved oxygen. Thus after stratifying by salinity zone and perhaps other factors, a series of dissolved oxygen reference CDF curves could be computed from the existing 20+ year monitoring data base. When it is not possible to establish a reference condition some more arbitrary device must be employed. Alternatives are discussed in Section 4.0.

Discussion of Each Step

Step 1 - data collection. One of the advantages of the CFD approach is that it will accommodate a variety of input data and still arrive at the same assessment endpoint. Data collection methods currently in place include: fix station data, cruise track data, continuous monitor data, aircraft flight path data, and satellite imagery data. Because of the interpolation step, all of these data can be used (and potentially combined) with varying degrees of success to estimate the total spatial (to the limit of interpolator pixel size) distribution of a water quality constituent. As noted above, one could construct this process by reversing the roles of time and space. That is, first interpolate over time and then build a cumulative distribution in space. In theory it is an abitrary choice to first standardize the data over space by interpolation and then construct the cumulative distribution in time. However, in practice, there is a greater diversity of sampling designs over space and therefore it is the sampling in the spatial dimension more than the temporal that creates many types of data that must be forced to a common currency.

Step 2 - interpolation. Interpolation is the step that puts data collected at various spatial intensities on a common footing. On the one hand, this is advantageous because data collected at many spatial intensities are available for the assessment process. On the other hand, it can be misleading to accept interpolated surfaces from different data sources as equivalent without qualifying each interpolation with a measure of the estimation error that is associated with each type of data. Clearly an interpolation based on hundreds of points per segment (such as cruise track data) will more accurately reflect the true noncompliance percent when compared to an interpolation based on two or three points per segment (such a fixed station data). Of the various types of interpolation algorithms available, the method proposed for this assessment is kriging. Kriging offers the best available approach for the estimation error associated with interpolation.

Step 3 - pointwise compliance. Determining the percent of compliance of each cell from each interpolation would seem to be a simple step. If the estimated value for a cell exceeds the criterion then that cell is out of compliance.

While interpolation allows for a standardization of many types of data, pointwise compliance allows for standardization of many criteria. Because compliance is determined at points in time and space, it is possible to vary the compliance criteria in time and space. If different levels of a water quality constituent are acceptable in different seasons, then the criterion can vary by season. It is possible to implement different criteria over space for a segment that bridges oligohaline and mesohaline salinity regimes. It would even be possible to let the criterion be a continuous function of some ancillary variable such as temperature or salinity. All that is required is that the final determination be yes or no for each interpolator cell.

Even the simplicity of this concept becomes diminished when issues of interpolation error are considered. Consider the assessment of two interpolator cells from an interpolation based on cruise track data. One cell near the cruise track has an estimated value is 4 and a standard error of 0.1. A second cell far from the cruise track has an estimated value of 4 and a standard error of 1.0. If the criterion were 3.0, it is fairly certain that the first cell represents exceedance. It is much less certain that the second cell represents exceedance. In the simple assessment of non-compliance, they count the same.

Step 4 - percent non-compliance in space. Computing a percentage should also be a simple step. The estimate is simply 100 times the number of cells out of compliance divided by the total number of cells. As a rule, the uncertainty of a binary process can be modeled using a binomial distribution. However, the issue of uncertainty described for step 3 propagates into computing the percent of compliance for a segment. Add to that the fact that estimated values for interpolator cells have a complex dependence structure which rules out a simple binomial model and the rules governing the uncertainty of this step are also complex. The number of interpolator cells, N, is relatively constant and under an independent binomial model the variance of the proportion of cells not in compliance, p, would be (p)(1-p)/N. Intuitively, one expects the variance of p to decrease as the number of data points that feeds the interpolation increases. This expectation has been confirmed by simulation, but the mathematical tools for modeling this propagation of error are yet to be developed.

Step 5 - Percent of time. While the percent of space coordinate of the CFD has simple interpretation of the percent of the segment out of compliance on a given date, the percent of time coordinate is not simply the percent of time out of compliance at a given point. Instead the percent of time coordinate has an interpretation similar to that of a cumulative distribution function. The percent of time coordinate is the percent of time that the associated spatial percent of noncompliance is exceeded. For example, if the (percent space, percent time) coordinates for a point on the CFD are (90,10), one would say that the spatial percent of noncompliance is greater than or equal to 90% about 10% of the time.

This step is very similar to computing an empirical distribution function which is an estimator of a cumulative distribution function. Because of this similarity, one immediately thinks of statistical inference tools associated with empirical distribution functions, such as the Kolmogorov-Smirnov, Shapiro-Wilk, Anderson-Darling, or Cramer-von Mises, as candidates for inference about the CFD. These procedures model uncertainty as a function of sample size only; in this case the number of sample dates. The fact that it does not incorporate the uncertainty discussed the previous steps seems unsatisfactory.

A quick review of probability plotting will reveal several methods on estimating the percent of time coordinate in step 5. Formulae found in the literature include: (R/N), (R - 0.5) / (N - 1). and (R - 0.375) / (N + 0.5), where R is rank and N is sample size. These generally fall in to a family of given by (R - A)/(N - 2A + 1) for various values of A. They are approximately equal, but the choice should be fixed for a rule.

6. Plotting the CFD. Even the plotting of the points is subject to variation, although these variations are somewhat minor compared to the larger issue of assessing the uncertainty of the assessment curve. The simple approach used in the figures above is to connect the points by line segments. In the statistical literature, it is more common to use a step function. If the graph represents an empirical distribution function, each horizontal line segment is closed on the left and open on the right. Because the CFD is an inversion of an EDF it would be appropriate for these line segments to be closed on the right and open on the left.

7. Comparing the Curves. It is at the point of comparing the assessment curve to the reference curve that the issue of uncertainty becomes most important. From the preceding discussion it is clear that uncertainty in the assessment curve is an accumulation of uncertainty generated in and propagated through the preceding 6 steps. If the reference curve is biologically based, it is derived under the same system of error propagation. Developing the statistical algorithms to quantify this uncertainty is challenging.

Even if the uncertainty can be properly quantified, the issue of who gets the benefit of doubt due to this uncertainty is a difficult question to resolve. This is a broad sweeping issue regarding uncertainty in the regulatory process, not a problem specific to the CFD approach. None-the-less, it must be dealt with here as well as elsewhere. One option is to require that the assessment curve be significantly above the reference curve to establish noncompliance. This option protects the regulated party from being deemed out of compliance due to random effects, but if assessment CFD curves are not accurately determined, it could lead to poor protection of environmental health and designated uses. A second option is to require that the assessment curve be significantly below the reference curve to establish compliance. This results in strong protection of the environmental resource, but could lead to the regulated party implementing expensive management actions that are not necessary. Some compromise between these extremes is needed. The simplest compromise is to ignore variability and just compare the assessment curve to the reference curve. As long as unbiased estimation is implemented for both the assessment curve and the reference curve, this third option will result in roughly equal numbers of false positive (declaring noncompliance when in fact compliance exists) and false negative (declaring compliance when in fact noncompliance exists) results. This offers a balanced approach, but there is no mechanism to motivate a reduction of these false positive and false negative errors

2.2 Data Available and Current Methods

Overview of Types of Data Available

The Chesapeake Bay monitoring program routinely monitors 19 directly measured water quality paramenters at 49 stations in the mainstem Bay and 96 stations in the tidal tributaries. The Water Quality Monitoring Program began in June 1984 with stations sampled once each month during the colder late fall and winter months and twice each month in the warmer months. A refinement in 1995 reduced the number of mainstem monitoring cruises to 14 per year. "Special" cruises may be added to record unique weather events. The collecting organizations coordinate the sampling times of their respective stations, so that data for each sampling event, or "cruise", represents a synoptic picture of the Bay at that point in time. At each station, a hydrographic profile is made (including water temperature, salinity, and dissolved oxygen) at approximately 1 to 2 meter intervals. Water samples for chemical analysis (e.g., nutrients and chlorophyll) are collected at the surface and bottom, and at two additional depths depending on the existence and location of a pycnocline (region(s) of density discontinuity in the water column). Correlative data on sea state and climate are also collected.

In addition, Chesapeake Bay Program partner organizations Maryland Department of Natural Resources and the Virginia Institute of Marine Science have recently begun monitoring using a technology known as data flow. DATAFLOW is a system of shipboard water quality probes that measure spatial position, water depth, water temperature, salinity, dissolved oxygen, turbidity (clarity of the water), and chlorophyll (indicator of plankton concentrations) from a flow-through stream of water collected near the water body’s surface. This system allows data to be collected rapidly (approximately every 4 seconds) and while the boat is traveling at speeds up to 20 knots.

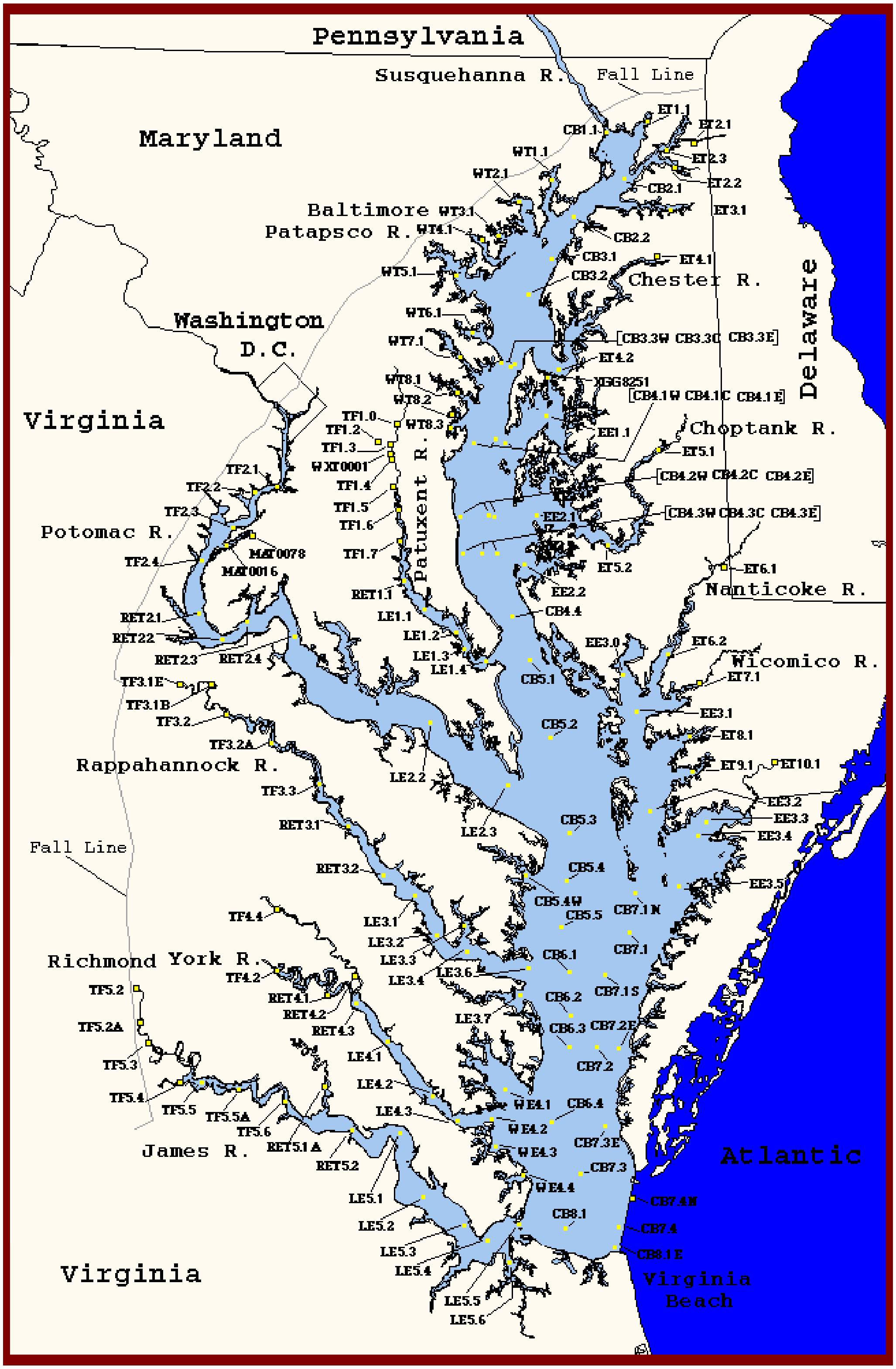


Figure 2.3. Map of the tidal water quality monitoring stations

In 2005, the MDDNR Water Quality Mapping Program covered 16 Chesapeake Bay, Coastal Bay and Tributary systems. The St. Mary's, Patuxent, West, Rhode, South, Middle, Bush, Gunpowder, Chester, Eastern Bay, Miles/Wye, Little Choptank, Chicamacomico and Transquaking Rivers will be mapped, as well as Fishing Bay and the Maryland Coastal Bays. In Virginia, dataflow data are available for the Piankatank, York, Pamunkey and Mataponi Rivers.

Beginning in 1990, Chlorophyll-a concentrations were measured over the mainstem Chesapeake using aircraft remote sensing. From 1990-1995, the instrument used for this study was the Ocean Data Acquisition System (ODAS) which had three radiometers measuring water leaving radiance at 460, 490 and 520 nm. In 1996, an additional instrument was added, the SeaWiFS Aircraft Simulator (SAS II). SAS II has sensors at seen wavebands which improves detection of Chlorophyll in highly turbid areas. Since 1990, 25-30 flights per year have been made during the most productive times of year.

The data described above and additional information can be obtained from: www.chesapekebay.net mddnr.chesapeakebay.net/eyesonthebay/index.cfm

www2.vims.edu/vecos/

Description of the current nearest neighbor/IDW interpolator

The current Chesapeake Bay Interpolator is a cell-based interpolator. Water quality predictions for each cell location are computed by averaging the nearest “n” neighboring water quality measurements, where “n” is normally 4, but this number is adjustable. Each neighbor included in the average is weighted by the inverse of the square of Euclidean distance to the prediction cell (IDW).   Cell size in the Chesapeake Bay was chosen to be 1km (east- west) x 1km (north-south) x 1m (vertical), with columns of cells extending from surface to the bottom of the water column, thus representing the 3-dimensional volume as a group of equal sized cells extending throughout the volume. The tributaries are represented by various sized cells depending on the geometry of the tributary, since the narrow upstream portions of the rivers require smaller cells to accurately model the river’s dimensions. This configuration results in a total of 51,839 cells by depth for the mainstem Chesapeake Bay (segments CB1TF-CB8PH), and a total of 238,669 cells by depth for all 77 segments which comprise the mainstem Bay and tidal tributaries.

The Chesapeake Bay Interpolator is unique in the way it computes values in 3 dimensions. The interpolator code is optimized to compute concentration values, which closely reflect the physics of stratified water bodies, such as Chesapeake Bay. The Bay is very shallow compared to its width or length; hence water quality varies much more vertically than horizontally. The Chesapeake Bay Interpolator uses a vertical filter to select the vertical range of data that are used in each calculation. For instance, to compute a model cell value at 5m deep, monitoring data at 5m deep are preferred. If fewer than n (typically 4) monitoring data values are found at the preferred depth, the depth window is widened to search up to d (normally +/-2m) meters above and below the preferred depth, with the window being widened in 0.5m increments until n monitoring values have been found for the computation. The smallest acceptable n value is selectable by the user. If fewer than n values are located, a missing value (normally a –9) is calculated for that cell. A second search radius filter is implemented to limit the horizontal distance of monitoring data from the cell being computed. Data points outside the radius selected by the user (normally 25,000m) are excluded from calculation. This filter is included so that only data that are near the location being interpolated are used.

In this version of the Interpolator, Segment and Region filters have been added. Segments are geographic limits for the interpolator model. For instance, the Main Bay is composed of 8 segments (CB1TF, CB2OH, …,CB8PH). The tributaries are composed of 77 additional segments, using the CBP 2003 segmentation. These segments divide the Bay into geographic areas that have somewhat homogeneous environmental conditions. This segmentation also provides a means for reporting results on a segment basis, which can show more localized changes compared to the whole Bay ecosystem.

Segment and bathymetry information use by the interpolator is stored in auxiliary files. Segment information allows the interpolator to report results on a segment basis which can show more localized changes compared to the whole Bay ecosystem. These segment and bathymetry files have been created for the main bay and all of the larger tributaries. The CBP segmentation scheme was replicated in these files by partitioning and coding the interpolator cells that fall within each segment.

The interpolator also identifies the geographic boundary that limits which monitoring station data are included in interpolation for a given segment through a region file. Use of data regions ensures that the interpolator does not “reach across land” to obtain data from an adjacent river which would give erroneous results. By using data regions, each segment of cells can be computed from their individual subset of monitoring data. Each adjacent data region should overlap by some amount so that there is a continuous gradient, and not a seam, across segment boundaries.

Current Implementation of CFD

The Chesapeake Bay Program has initiated implementation of the CFD as an assessment tool. The Criteria Assessment Protocols (CAP) workgroup was formed in the fall of 2005 to develop detailed procedures for implementing criteria assessment. This workgroup has developed and implemented procedures that use the CFD process and conducted a CFD evaluation of dissolved oxygen for many designated assessment units.

The CFD methodology was first applied in the Chesapeake Bay for the most recent listing cycle which was completed in the Spring of 2006 and was based on data collected over the period 2002 through 2004. CFDs were developed and utilized primarily for the dissolved oxygen (DO) open- and deep-water monthly mean criteria because there were insufficient data collected to assess the higher-frequency DO criteria components. The clarity criteria were not assessed based on the CFD because there were few systems in which there was sufficient data for an assessment. Chlorophyll criteria were not available from the Chlorophyll criteria team in time to implement a chlorophyll assessment.

In general, the CFD analysis indicated that most of the Bay waters failed one or more of the open-water or deep-water DO criteria components. However, there were also many tributaries in which all of the DO criteria assessed indicated attainment. Thus in this initial application, the CFD method did appear to distinguish between impaired and unimpaired systems in a manner that is consistent with the expectations of the many stakeholders in the CAP workgroup.

In the 2006 application of the assessment methodology, there were many details that required resolution in order to fully implement the methodology. Procedures generally followed the theoretical description as described in Section 2.1, but some details were modified to address unforeseen complications. The following describes some of those details.

In general, data were obtained from the CBP CIMS data base and parameters included date, location, depth, salinity, temperature and the water quality parameter being assessed. Some State data were also incorporated and those data were obtained directly from the relevant State. Once all the data were compiled, they were assigned to a time period based on the sample date. Fixed-station data are normally collected during a monitoring cruise that covers the entire tidal Chesapeake Bay over several days. However, in order to provide a “snapshot” in water quality, the data collected within a cruise are assumed to be contemporaneous in order to perform a single spatial interpolation. For any data not associated with a cruise, a cruise number is assigned representing the closest cruise in time to the collection of each datum. Co-located data points in the same cruise were averaged.

The assessment procedure requires assessment over large areas rather than at points in space. Spatial interpolation using the CBP IDW interpolator was performed for each water-quality criteria parameter for each cruise. Clarity and surface chlorophyll were interpolated in the two horizontal dimensions using inverse distance squared weighting. Dissolved oxygen was first linearly interpolated in the vertical dimension within each column of data beginning at 0.5 meters and continuing at one meter intervals, not to exceed the deepest observation in that column. Each depth was then interpolated horizontally using inverse distance squared weighting. Data regions were specified for each segment in order to prevent the interpolation algorithm from using data points in neighboring tributaries.

Designated uses in the Chesapeake Bay are defined vertically in order separate stable water layers that have differing criteria levels for dissolved oxygen. The surface layer (open water) is that layer defined to be above the pycnocline and thus exposed to the atmosphere. The middle layer (deep water) is defined to be the layer between the upper and lower pycnocline. And the lower layer (deep channel) is defined to be the layer below the pycnocline. Given that the pycnocline is dynamic and moves up and down with each monitoring cruise, the designated use of each grid cell must also be defined based on the available data for each cruise.

The pycnocline is defined by the water density gradient over depth. Temperature and salinity are used to calculate density, which in turn is used to calculate pycnocline boundaries. Density is calculated using the method described in: *Algorithms for Computation of Fundamental Properties of Seawater* (Endorsed by UNESCO/SCOR/

ICES/IAPSO Joint Panel on Oceanographic Tables and Standards and SCOR Working Group 51. Fofonoff, N P; Millard, R C Jr. UNESCO technical papers in marine science. Paris , no. 44, pp. 53. 1983). For each column of temperature and salinity data, the existence of the upper and lower pycnocline boundary is determined by looking for the shallowest robust vertical change in density of 0.1 kg/m3/m for the upper boundary and deepest change of 0.2 kg/m3/m for the lower boundary. To be considered robust, the density gradient must not reverse direction at the next measurement and must be accompanied by a change in salinity, not just temperature.

The depths to the upper pycnocline boundary, where detected, and the fraction of the water column below the lower boundary are interpolated in two dimensions. If no lower boundary was detected the fraction was considered to be zero. The depth to the upper pycnocline boundary tends to be stable across horizontal space and so spatial definition of that boundary using interpolation generally worked well. However, interpolation of the lower boundary is more complicated because the results can conflict with the upper boundary definition or with the actual bathymetry of the Bay. As a result, interpolation of the lower boundary was performed based on “fraction of water column depth”. In that way, the constraints of the upper pycnocline boundary definition and the actual depth were imposed and errors related to boundary conflicts were eliminated.

Assessments were performed based on criteria specific averaging periods. The instantaneous assessment for deep channel dissolved oxygen was evaluated using the individual cruise interpolations. All monthly assessments were based on monthly averages of interpolated data sets. To calculate the monthly averages, each interpolated cruise within a month was averaged on a point-by-point basis. Generally, there were 2 cruises per month in the warmer months and 1 cruise per month in the cooler months. Spatial violation rates are calculated for each temporally aggregated interpolation in an assessment period. For example, for a three-year summer open-water dissolved oxygen assessment, the twelve monthly average interpolations representing the four summer months over three years were used.

3. Protocol for Interpolating Water Quality

The CFD approach uses the proportion of space in attainment in any given month estimated using an approach based on a statistical model. The current method uses data collected in a specific month at a set of sampling locations within the segment of interest to estimate the parameters of the model. The estimated model is then used to interpolate likely values at unsampled locations, specifically at a set of prediction locations arranged in a grid over the segment. The predictions thus obtained are used to calculate the proportion of space in compliance that month. The current estimation procedure for obtaining predicted values is Inverse Distance Weighting (IDW), a non-statistical spatial interpolator that uses the observed data to calculate a weighted average as a predicted value for each location on the prediction grid. The method calculates the weight associated with a given observation as the inverse of the square of the distance between the prediction location and the observation.

The panel considered several interpolation methods in addition to IDW. Of these, kriging methods emerged as a principal alternative approach for populating the grid of prediction locations. Non-parametric methods were also considered. These include Loess regression or cubic spline methods. These approaches could be advantageous in that they are statistical methods that provide levels of error, but panel analyses and deliberations have been insufficient to provide definitive statements on this class of methods. Table 3.2 which appears in Section 3.3 summarizes our determinations.

3.1 Kriging Overview

Kriging is a spatial interpolation technique that arose out of the field of geostatistics, a subfield of statistics that deals with the analysis of spatial data. Kriging and the field of geostatistics has been employed in a wide variety of environmental applications and is generally accepted as a method for performing statistically optimal spatial interpolations (Cressie 1991, Schabenberger and Gotway 2004, Diggle and Ribeiro 2006). Applications of kriging in water related research can be found in (Kitanidis 1997, Wang and Liu 2005,Ouyang et al. 2006). References on kriging methodology, geostatistics, and their related statistical development can be found in (Cressie 1991, Diggle et al. 1998, Schabenberger and Gotway 2004, Diggle and Ribeiro 2006).

Kriging can equivalently be formulated in terms of a general linear regression model

$$\mathrm{Y}\,(\boldsymbol{s}) = \beta_0 + \beta_1\,\mathrm{X}_1(\boldsymbol{s}) \cdots + \beta_\mathrm{p}\,\mathrm{X}_\mathrm{p}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \qquad (1)$$

with s representing a generic spatial location vector (usually 2-D) assumed to vary continuously over some domain of interest, $\mathrm{Y}(\boldsymbol{s})$ the outcome of interest measured at $\boldsymbol{s}$, $\mathrm{X}_1(\boldsymbol{s}), \ldots ,\mathrm{X}_\mathrm{p}(\boldsymbol{s})$ potential covariates indexed by location s, and their associated regression effects $\beta_1, \ldots , \beta_\mathrm{p}$. Note that covariates must be known at every prediction location. The elements of the spatial vector $\boldsymbol{s}$ can be used as covariates for modeling spatial trends. On the other hand water quality measures such as salinity which may have a strong association with the outcome of interest, is of limited value as a covariate because it is not known at all prediction locations. The uncertainty in this regression relationship is modeled with the random error term $\varepsilon(\boldsymbol{s})$ assumed to have zero mean and constant variance. Spatial data like the type sampled in the Chesapeake Bay water-quality criteria assessments often exhibit a property known as (positive) spatial dependence, observations closer together are more similar than those further away. This property is accounted for in model (1) by allowing $\varepsilon(\boldsymbol{s})$ to have a spatial correlation structure.

Some further specifics on $\varepsilon(\boldsymbol{s})$ are warranted. Common distributional assumptions on $\varepsilon(\boldsymbol{s})$ include normality or log-normality, although kriging can be performed based on other statistical distributions and data transformations (Christenson et al. 2001). The spatial correlation in $\varepsilon(\boldsymbol{s})$ is represented by positive definite functions. These functions can be assumed isotropic where correlation decay depends just on distance, or anisotropic where correlation decay depends on distance and direction. Variograms are another special type of mathematical function closely related to spatial correlation functions that can and are more often used to represent spatial correlation. For purposes here and in many kriging applications, variograms and spatial correlation functions provide equivalent representations of spatial structure. For consistency in what follows only the term variogram will be used in discussions of spatial structure.

While there is considerable flexibility in implementing the error structure of a kriging model, it is possible to generalize somewhat with respect to the error structure of Chesapeake Bay water quality data. Of the three water quality parameters being assessed, chlorophyll and clarity measures tend to follow the log-normal distribution and dissolved oxygen is reasonably approximated by the normal distribution. The horizontal decay rate of spatial correlation does not tend to be directionally dependent. Thus if the bay is viewed as a composite of horizontal layers, isotropic variograms are appropriate for kriging each layer. In a vertical direction, water quality can change rapidly and thus spatial correlation can decay over a short distance. A 3-D interpolation procedure would benefit from use of an anisotropic variogram in order to differentiate the vertical correlation decay from the horizontal correlation decay.

Note, in the literature model (1) is referred to as a universal kriging model. When covariates (the X's) are not considered to influence interpolation of Y the right hand side of model (1) contains just the constant term β_0 and ε(s). The resulting model is referred to as the ordinary kriging model. When the spatial structure (variogram) for model (1) is known, statistically optimal predictions for the variable Y at unsampled locations (outside of estimation of possible regression effects) can be derived using standard statistical principles. The optimality criteria results in spatial predictions that are linear in the data, statistically unbiased, and minimize mean squared prediction error, hence referred to as best linear unbiased predictions (BLUPs). The minimized mean squared prediction error is also taken as a measure of prediction uncertainty. In practice, however, spatial structure of the data is unknown, the estimation of which via the variogram function is cornerstone to kriging applications.

To demonstrate let $\{y(s_1), \ldots, y(s_n)\}$ represent a set of spatial data, for example a water-quality parameter such as dissolved oxygen sampled at a set of n spatial locations $s_1, \ldots, s_n$. Assume this data to be a realization of the ordinary kriging version of model (1). The first step in kriging is variogram estimation. There are several methods available, method of moments and statistical likelihood based being two of the more common, all of which though are based on the sample data {y(s$_1$), . . . , y(s$_n$)}. Without going into detail, this process ends with a chosen variogram function and its parameter estimation, describing the shape and strength (rate of decay) of spatial correlation. There is also a determination, again based on the sampled data, of whether the spatial structure is isotropic or anisotropic. The estimated variogram is then assumed known and kriged interpolations and their interpolated uncertainty are computationally straight forward to generate at numerous locations where data were not observed. Accounting for uncertainty in variogram parameter estimation has commonly been explored using Bayesian methods (Diggle and Ribeiro 2006).

3.2 IDW Overview

The inverse distance weighting method that is currently used in the CFD approach has already been described. Hence, this section provides a short review of IDW's technical details and a comparison of IDW to alternative interpolation methods.

The IDW method is essentially a deterministic, non-statistical approach to interpolating a two or three dimensional space. As a result it lacks statistical rigor so that estimates of the prediction errors are not calculable without additional assumptions. Similar to kriging, IDW predicts a value ($\hat{Y}$) at an unobserved site, say at location s_0, using a weighted average of the N nearest observed neighbors (N specified by the modeler):

$$\hat{Y}(s_0) = \sum_{i=1}^{N} w_i Y(s_i)$$

where the weights, w_i, are inversely related to the distance between locations s_0 and s_i

$$w_i = \frac{d(s_0, s_i)^{-2}}{\sum_{j=1}^{N} d(s_0, s_j)^{-2}},$$

$d(s_0, s_i)$ is the Euclidean distance between locations s_0 and s_i, and the denominator of the weight is to ensure that the weights sum to 1. The IDW is an exact interpolator in that the predicted values for observed locations are the observed values and the maximum and minimum values of the interpolated surface can occur only at observed sites.

Recent research has compared IDW to other interpolation techniques, most notably variations in kriging (Table 3.1). The authors found that in some cases kriging was at least as good an interpolator as IDW and in some instances better. The non-parametric techniques (splines and similar methods) were not as precise as kriging and IDW. The method used for comparison in virtually all of the research was some variant of cross-validation, a method where some data are kept aside and not used in the model estimation phase and then using the resulting model to predict values for the data kept aside. The predicted and observed values are then compared and a statistic is calculated that summarizes the differences between the two sets of values (observed and predicted).

None of these studies used datasets with highly irregular edges such as are found in the Chesapeake Bay nor did they use any distance metric other than Euclidean distance. Whether one method is preferable to another in these more difficult situations remains unexplored.

One final and important issue with IDW is that, as currently used, IDW is a deterministic method which makes no assumptions as to the probability distribution of the data being interpolated. Hence, it does not allow for estimating prediction errors, i.e. it does not allow for the possibility of random variation at interpolation sites. A simple question is whether IDW can be recast in the kriging framework given the similarity in prediction method (weighted average) and hence can a method be found to estimate prediction errors? The short answer is no – the distance function used by IDW, which is an implicit assumption about the autocorrelation function in the spatial field, does not meet the assumptions required for development of a valid variance-covariance matrix describing the spatial covariance. As a result, IDW cannot be modified to take advantage of the statistical knowledge that has been developed for geostatistical analyses such as kriging. This does not imply that other approaches to estimating prediction error are also not possible.

A non-parametric approach for estimating variance was proposed (Tomczak, 1998) in which jack-knifing was used to provide error estimates. 95% confidence intervals for the mean were calculated and then compared to the actual observed values. Not surprisingly, only 65% of the data were captured within their associated confidence interval. The method appears to have been misapplied – the jackknifing method as used estimates the standard error of the mean assuming independent observations. As a result, the confidence interval is not capturing the effect of the spatial dependencies nor is it based on the fact that we are predicting a value for the unobserved site rather than estimating a mean. The development described by Tomczak (1998) should be explored further and other alternatives such as block bootstrapping for variance estimation as well.

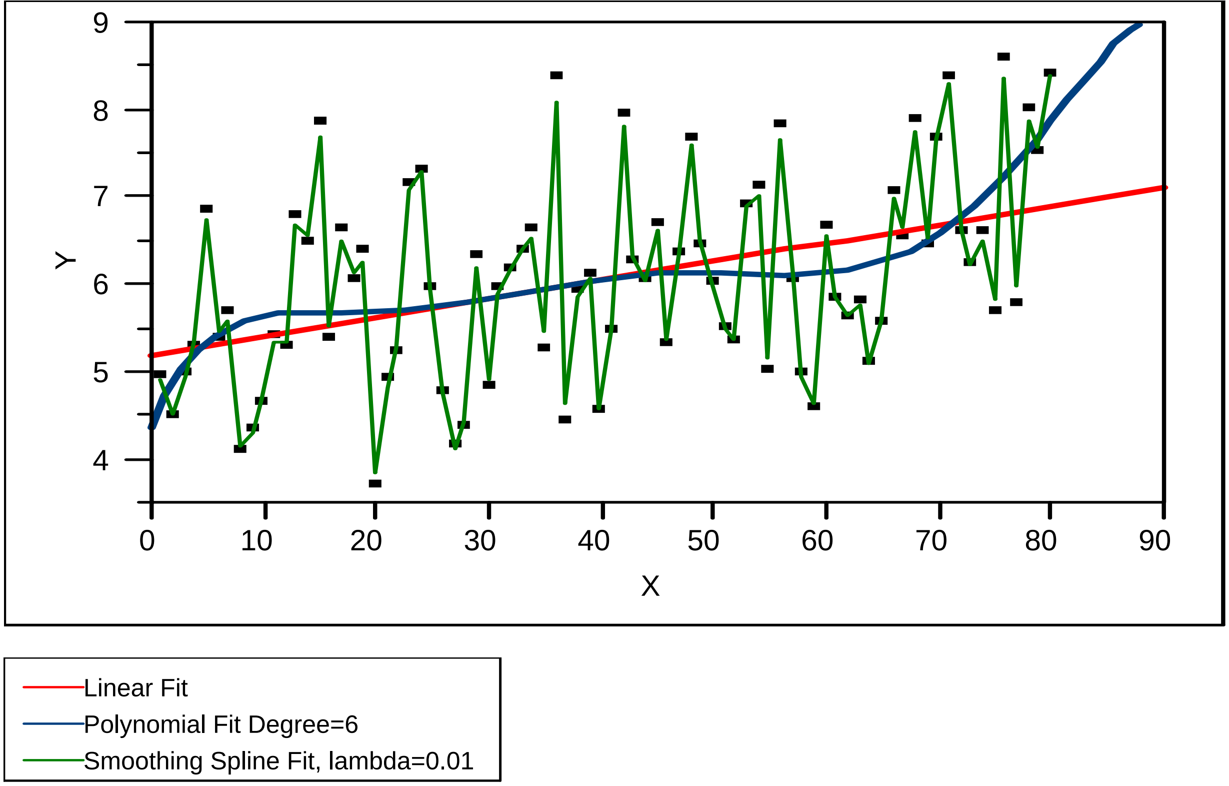
Table 3.1. A short list of recent articles comparing the precision of IDW to a subset of other possible interpolation methods.

Authors	**Methods Compared**	**Variables Manipulated**	**Conclusions**
Kravchenko (2003)	Inverse Distance Weighting (IDW), Ordinary Kriging (OK)	spatial structure and sample grid spacing	IDW better than OK unless sample sizes were fairly large
Dille, et al. (2002)	IDW, OK, Minimum Surface Curvature (MC), Multiquadric Radial Basis Function (MUL)	neighborhood size, spatial structure, power coefficient in IDW, sample grid spacing, quadrat size	No interpolator appears to be more precise than another. Sample grid spacing and quadrat size were deemed more important.
Valley, et al. (2005)	IDW, OK, Non-parametric Detrend + Splines	spatial structure, sample size, quadrat size	OK tended to be more precise but IDW was very similar
Lloyd (2005)	moving window Regression (MWR), IDW, OK, simple kriging with locally varying mean (SKlm), kriging with external drift (KED)	spatial structure, sample size	KED and OK best
Reinstorf, et al. (2005)	IDW, OK, KED + deterministic chemical transport models	single dataset was analyzed	OK best
Zimmerman, et al. (1999)	2 types of IDW, UK, OK	spatial structure, sampling pattern, population variance	UK and OK better than IDW

3.3 Non-parametric Interpolation Methods

There are many variations on spatial interpolation in addition to kriging and IDW. See Cressie (1989) for a review. The committee did not have sufficient time to compare all models, but CBP in encouraged to continue this research. One promising category of models are for interpolation based on non-parametric methods that do not rely on measuring and accounting for spatial autocorrelation. All of the non-parametric approaches would be based on the assumption that the autocorrelation observed in the data is due to unobserved explanatory variables and hence alternative modeling approaches are not unreasonable. The particular set we mention are the regression type analyses with the locational indices (northings, eastings) used as explanatory variables. Examples include generalized additive models (Hastie and Tibshirani, 1990), high-order polynomials (Kutner, Nachtsheim, Neter, and Li, 2004), splines (Wahba, 1990), and locally weighted regression (“loess” or “lowess”, Cleveland and Devlin, 1988). In some kriging and IDW methods, large-scale trend is modeled relatively smoothly using locational indices and local smaller-scale variation is modeled using the estimated autocorrelation in conjunction with the values of the variable at nearby observed sites. The nonparametric methods replace estimation of the local variation based on correlation functions with models of the large-scale trend that are less smooth and more responsive to the spatial variation in the observed data. A visual demonstration is given in Figure 3.1 which shows a one-dimensional dataset with Y as the variable to be predicted and X as the location along the one dimensional axis. For example, X could be distance from the mouth of a river and Y could be chlorophyll a concentration.

Figure 3.1. Bivariate fit of Y By X. Straight line is a linear large-scale trend fit (R^2 = 0.19); the moderately wavy line around the straight line is a 6th-order polynomial fit (X enters the model as X, X^2, X^3, …, and X^6; R^2 = 0.25); and the jagged line is a spline fit with a very small bandwidth (neighborhood used in local estimation at each X; R^2 = 0.90).



One advantage of these approaches is that each of the methods has extensive statistical research into estimation of model parameters as well as standard errors for those parameters and for predictions at interpolation sites. Another is that the main modeling decisions are related to bandwidth selection or degree order of polynomial to fit. These decisions can be automated by developing rules for roughness of fit based on reduction in MSE as compared to modeling a straight line (in X). Disadvantages are the same as for kriging, all model estimation is data dependent which means that the spatial configuration and number of sampling sites has a direct influence on the predictions and their error estimates. In addition, a study done by Laslett (1994) comparing kriging and splines indicated that the two methods are similar in predictive power but for certain sampling regimes kriging performs better. We recommend more study since the non-parametric approaches would be easier to implement than kriging.

3.3 Comparison of Methods

The following describes some of the benefits and potential limitations of kriging in regards to CBP application with some comparisons to the IDW approach towards spatial interpolation outlined in the previous section. Nonparametric methods are not sufficiently developed to include in this comparison. A primary benefit of the kriging methodology compared to IDW is that it is a statistical technique. As such the field of statistics (including kriging) is designed to make inference from sampled data in the presence of uncertainty and the quantity and quality of the sample data are reflected in those inferences. However, kriging is a less than routine type of statistical analysis and requires a certain level of statistical expertise to carry out the process. The short description on variogram estimation provided above merely introduces this involved and often complicated step. This requirement for informed decision making limits the degree to which kriging can be automated and still maintain its flexibility and optimal properties.

Further issues regarding kriging and CBP applications are listed below.

- Kriging is flexible in that it is based on an estimate of the strength of spatial dependence in the data (variogram). Kriging can consider direction dependent weighted interpolations (anisotropy) and can include covariates (universal kriging) to potentially influence interpolations, either simple trends in easting and northing coordinates or water related measures such as sea surface temperature measured by satellite.
- A key feature of a statistical technique like kriging is that a measure of uncertainty (called the kriged prediction variance) is generated along with kriged interpolations. Research has been initiated (i.e., conditional simulation) to propagate this interpolation uncertainty through the CFD process for generating confidence intervals for estimates of attainment.
- Kriging can be applied in situations where the data are sparse, as in CBP fixed station data, or densely sampled, as in CBP shallow water monitoring. Kriged and IDW spatial interpolations may very well produce near identical results for these two extreme scenarios. However it is the kriging approach that provides a statistical model, the uncertainty of which is influenced by the quantity and quality of data. Knowledge of interpolation uncertainty is crucial for discriminating the improved water quality assessment obtained from densely sampled networks relative to sparsely sampled networks.

As alluded to earlier kriging is an advanced statistical technique and like all such techniques should be carried out by well trained statistician(s) with experience in spatial or geostatistical methodology and experience analyzing water quality data. Assessing model fits (of the variogram and regression model) and kriging accuracy via cross validation and/or likelihood based criteria should be employed routinely.

To further exemplify this point consider kriging the densely sampled shallow water monitoring data which is generated by the DATAFLOW sampling. In addition to the other technical complexities mentioned within, this spatial sampling design may raise other issues not immediately recognized by untrained users (Deutsch 1984).

For kriging in CBP applications one potential methodological drawback is the issue of non-Euclidean distance (Curriero 2006). Current kriging methodology only allows the use of the straight line Euclidean distance as the measure of proximity. However, the irregular waterways in the Chesapeake Bay system may very well suggest other non-standard measures of distance. For example, the spatial design of the fixed station data including those in the Bay mainstem and tidal tributaries. The straight line Euclidean distance may very well intersect land particularly in regions containing convoluted shorelines. There has been research initiated on this topic (Curriero 2006, Jensen et al. 2006, Ver Hoef et al. 2007), however, results are not yet ready for universal use.

Three dimensional interpolations (including depth as the third dimension) are potentially required for CBP applications. The IDW and kriging methodologies, mathematically speaking, certainly extend to three dimensions. However the rapid change of water quality over depth would lead to significant anisotropies in the application three dimensional kriging that would complicate this approach far more than the application of IDW. On the other hand, a simplistic implementation of IDW that does not recognize the rapid decay of covariance over depth would inappropriately reach across the pycnocline when choosing nearest neighbors. Clearly the special properties of water quality in a highly stratified bay require innovation for 3-dimensional interpolations. Another approach would be to apply universal kriging where a third dimension (depth) is used as a covariate. The use of depth as an independent variable is motivated by the observation that often water quality exhibits a predictable trend over depth as for example the trend of DO decreasing with increasing depth. To include depth as a covariate, model (1) would be written as

$Y(\mathrm{s}) = \beta_0 + \beta_1 \mathrm{Depth}(\mathrm{s}) + \varepsilon(\mathrm{s})$:

A third approach to interpolation in three dimensions is to implement 2-D interpolation in layers. Note that the IDW interpolator currently implemented by CBP (Section 2.2) employs a layered strategy by severely restricting (+/- 2m) the vertical distance that may be searched for nearest neighbors. A similar strategy could be implemented using 2-D kriging to interpolate the layers. Which of these approaches is best suited to 3-D interpolation for the bay will depend on the data available and assumptions related to vertical structure. Full 3-D kriging interpolation treats the 3rd dimension as a spatial dimension in the error term $\varepsilon(\boldsymbol{s})$. The covariate approach requires that the change over depth be a predictable trend. Interpolation in layers assumes that covariance decays so rapidly over depth that it is adequate to treat the layers as independent entities. Data sufficiency requirements increase for all approaches when considering three dimensional interpolations. When data are sparse, again a statistical based approach like kriging allows this to be reflected in prediction uncertainty.

In many applications, attainment or lack of attainment will be so extreme that the assessment end point is clear even without optimizing the error estimation of the CFD. In these extreme cases, IDW or kriging simplified for automation could be sufficient to support the attainment ruling without precise quantification of estimation uncertainty. For these cases, the customized IDW algorithm that is currently implemented by CBP provides a tool with which to begin testing the CFD assessment procedure, but kriging simplified for automation may offer some advantages. Kriging can be simplified for automation by fixing the variogram model to one mathematical form, say exponential, for all applications. With the variogram model fixed, kriging becomes like IDW in assuming the same mathematical form for the spatial dependence for all cases, but it is more flexible than IDW in that the rate of spatial correlation decay could be allowed to vary among applications. In addition, the simplified kriging opens the door for conditional simulation, with potential benefits that are discussed in Section 5. While a simplified kriging algorithm offers some advantages, there are also some potential drawbacks. Because variogram estimation typically entails use of an iterative procedure such as maximum likelihood or non-linear least squares, there is the potential that lack of convergence of these algorithms would be problematic for an automated implementation of kriging.

In terms of computing, IDW is available in commercial GIS software, requiring GIS skills for application. Kriging is available in commercial statistical software and also in the free open source R Statistical Computing Environment (R Development Core Team 2005, Ribeiro and Diggle 2001) and requires programming skills for those software packages.

In summary, kriging is more sophisticated than IDW, but requires greater expertise during implementation to fully exploit its full benefit. Table 3.2 provides a comparison of the capabilities of assessments based simply on: 1) percent of samples, 2) spatial interpolation based on IDW and 3) spatial interpolation based on kriging.

Table 3.2 – Comparison of the capabilities of methods available for interpreting data collected for Chesapeake Bay water-quality criteria assessment.

Attributes	**Sample-based**	**IDW**	**Kriging**
Provides Spatial Prediction	Yes	Yes	Yes
Provides Prediction Uncertainty	No	not routine	Yes
Uncertainty for CFD	No	No	Yes
Deal with Anisotropy	No	Possible, but not routine	Yes
Can Include Cruise Track/Fly over	No	Yes	Yes
Feasibility of 3 dimensional interpolations	No	Yes	Possible, but not routine
Feasibility of mainstem-tributary interpolations	No	Yes	Possible
Inclusion of covariates to improve prediction	No	No	Yes
Predictions of non-linear functions of predicted attainment surfaces P(y>c)	No	No	Yes
Level of Sophistication	Lowest	Low	Very High
Automation	Yes	Yes	Possible, but not routine

4.0 CFD Reference Curves

There are several approaches to defining reference curves that are proposed for use in the CFD assessment methodology. One is a biologically based definition and other approaches are based on an arbitrary allowable frequency (see Section 2). Here we review these options in greater detail.

4.1. Biological Reference Curves

The idea behind biological reference curves is to identify regions of the Bay that have healthy biological indicators and are thus considered to be in attainment of their designated use. CFDs would be developed for these areas in the same way that CFDs would be developed elsewhere, but those curves developed for healthy areas would be considered "reference" curves. For example, healthy benthic IBI scores might be used as indicators of adequate bottom dissolved oxygen.

The success of the CFD-based assessment will be dependent upon decision rules related to the biological reference curves. These curves represent desired segment-designated use water quality outcomes and reflect sources of acceptable natural variability. The reference and attainment curves follow the same general approach in derivation – water quality data collection, spatial interpolation, comparison to biologically-based water quality criteria, and combination of space-time attainment data through a CFD. Therefore, the biological reference curve allows for implementation of threshold uncertainty as long as the reference curve is sampled similarly to the attainment curve. Bias and uncertainty are driven in CFD curves by sample densities in time and space. Therefore, we advise that similar sample densities are used in the derivation of attainment and reference curves. As this is not always feasible, analytical methods are needed in the future to equally weight sampling densities between attainment and reference curves.

4.2. CBP Default Reference Curve

In some cases, the development of biologically-based reference curve is not possible due to lack of data describing the health of the relevant species. In such cases, a more arbitrary approach is required since better information is not available. EPA recommends the use of a default curve in cases where a biologically-based one is not available. That default curve is defined by these properties:

1. symmetric about the 1:1 line,
2. hyperbolic,
3. total area = 0.1, and
4. pass through (1,0) and (0,1)

(see EPA, 2003; page 174). The equation that describes this figure is defined by the equation:

```
(x+b)*(y+b) = a
```

Where: $b = 0.0429945$

$a = b^2 + b$

This reference curve is illustrated in Figure 4.1 by the blue curve.

An alternative default reference curve might be formulated by extending the arbitrary allowance of 10% exceedance into the two dimensional framework of the CFD.
The criterion threshold is a value that should be rarely exceeded by a population at healthy levels. When the population is unidimensional, say concentration in a point source effluent, then one can obtain this upper threshold based on the simple distribution of values in a healthy population (Figure 4.2). The ninetieth percentile of this distribution might be chosen as the criterion threshold. Thus in this example, 10% noncompliance is allowed because this level of noncompliance is expected in a healthy population. A standard technique for estimating distribution percentiles is to assume a mathematical form for the distribution, e.g., the normal distribution, and to estimate the percentile as some number of standard deviations above the mean. The 90th percentile of the normal distribution is 1.2815 standard deviations above the mean.

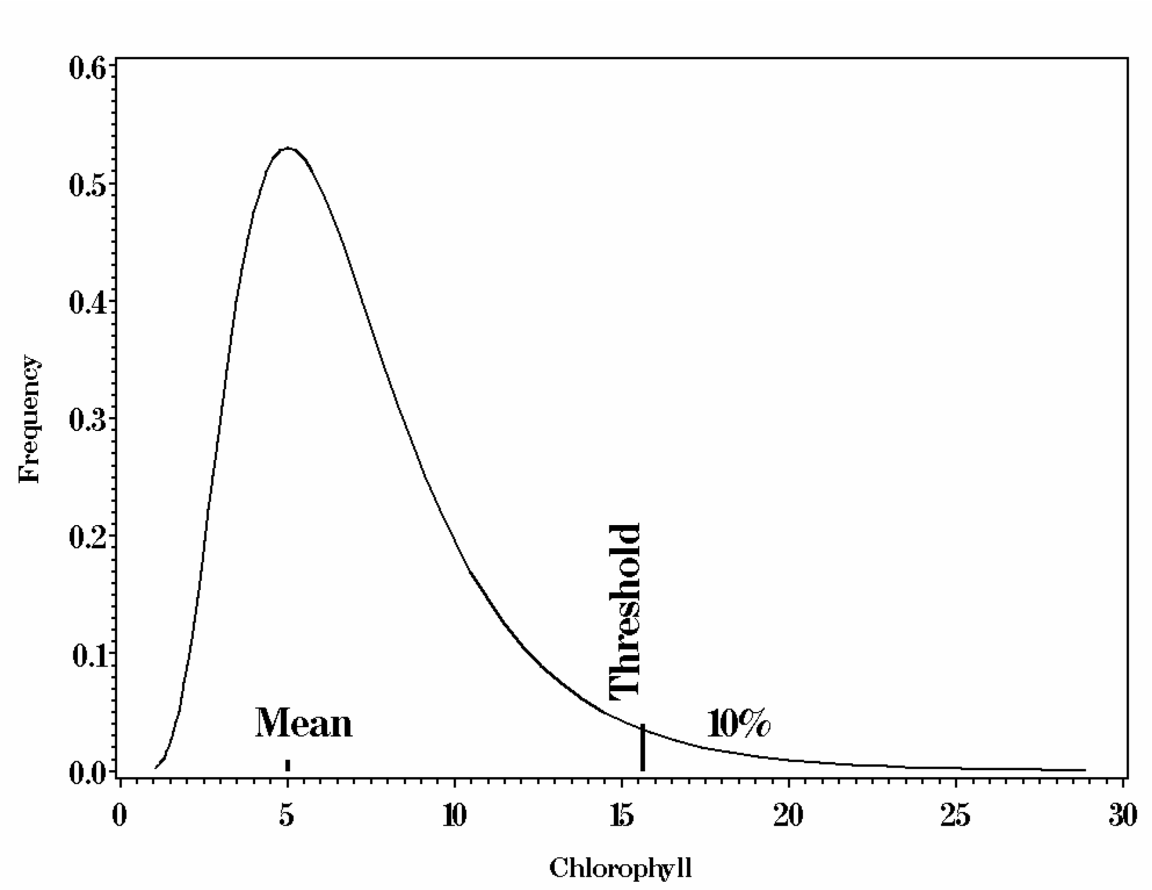


Figure 4.2. Hypothetical lognormal distribution that might be typical of Chlorophyll. The figure illustrates the relation of the geometric mean and the criterion threshold set at the 90th percentile.

When regulating populations that are distributed in both space and time, this simple concept for regulating noncompliance must be extended to account for the variability in each dimension. While there is some added complexity in the mathematics, the fundamental concept remains the same: That is, to set the criterion threshold at a certain distance above the mean so that exceedance of that threshold will be rare in a healthy population. In this case, the distance by which the threshold must exceed the mean is a function of both the spatial and temporal variance components as described below.

To establish these criteria thresholds for populations with two components of variance, assume the simple model:

$$Y_i(\boldsymbol{s}_j) = \mu + \alpha_i + \beta_i(\boldsymbol{s}_j)$$

where:

μ is the desired mean level of chlorophyll (in log space)
α_i is a random term for variation over time with variance σ^2_α ,
$\beta_i(\boldsymbol{s}_j)$ is a random term for variation over space with variance σ^2_β
$Y_i(\boldsymbol{s}_j)$ is a water quality constituent measured at time i and location $\boldsymbol{s}_j$.

The variance of x_{ij} is $\sigma^2_\alpha + \sigma^2_\beta = \sigma^2$. The standard dev of x_{is} is sqrt(σ^2) = σ. It is common to allow an overall 10% exceedance rate without declaring an assessment unit out of compliance. We would expect 10% of the x_{is} to fall above $u + 1.2815*\sigma$ where 1.2815 is the 90th percentile of the standard normal distribution. Thus (assuming normality) a population with spatial and temporal variance characterized by σ^2_α and σ^2_β that has a mean that is $1.2815*\sigma$ below the threshold criterion should have an exceedance rate of 10% over space and time. Note that the reference curve is determined by the ratio $\sigma^2_\alpha/\sigma^2_\beta$ and the distance in standard deviations of the mean from the threshold. The actual values of the variance components, the mean, and the threshold, are not important as long as the relationships hold. Thus as long as the variance ratio is consistent, and mean to threshold distance is a fixed number of standard deviations, the same reference curve will serve for all seasons and regions.

Letting chlorophyll observed in the decade of the 1960's serve as a reference population, the parameters in Table 4.1 can be used to construct this reference curve based on the variance ratio and the mean to threshold distance given in the table. The ratio $\sigma^2_\alpha/\sigma^2_\beta$ is computed as the ratio of the temporal variance term and the spatial variance term. The mean to threshold distance is computed to be 1.2815σ for all regions and seasons. Based on there parameters, a reference curve for chlorophyll can be derived (green curve, figure 4.1). For comparison a reference curve based on a variance ratio of 1.0 (red curve, Figure 4.1) and the standard CBP reference curve (blue curve, Figure 4.1) are also shown.

Table 4.1. Chlorophyll criteria derived by computing and upper threshold based on predicted means for mid-flow1960's chlorphyll data.

Season	Salinity Zone	Mean Log(chl)	GMmean (chl)	Temporal Variance	Spatial Variance	Std Dev log(chl)	Threshold Criterion log(chl)	Threshold Criterion (chl)
Spring	OH	0.7684	5.87	0.0233	0.0658	0.2985	1.2594	18.17
Summer	OH	1.1693	14.77	0.0233	0.0658	0.2985	1.6603	45.74
Spring	MH	0.4137	2.59	0.0233	0.0658	0.2985	0.9047	8.03
Summer	MH	0.8626	7.29	0.0233	0.0658	0.2985	1.3536	22.58
Spring	PH	0.1386	1.38	0.0233	0.0658	0.2985	0.6296	4.26
Summer	PH	0.218	1.65	0.0233	0.0658	0.2985	0.709	5.12

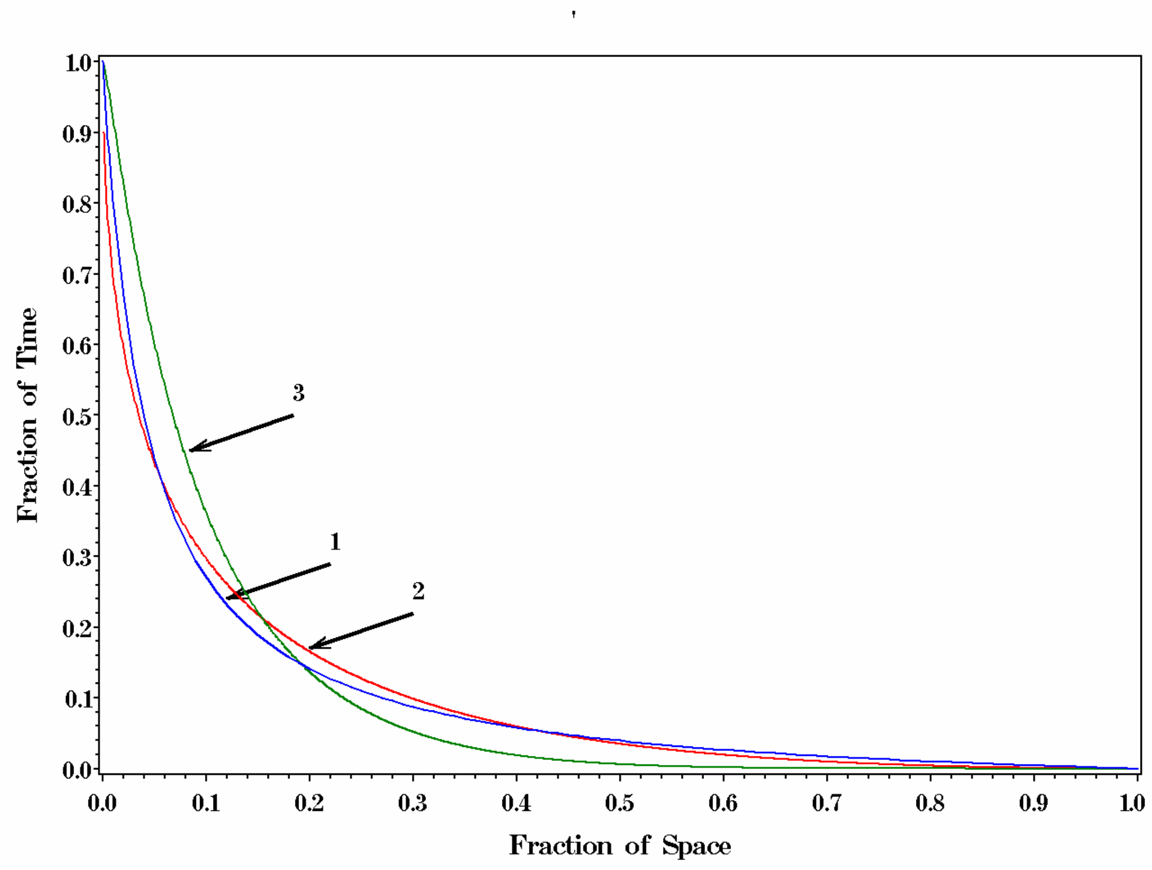


Figure 4.1. Illustrations of three reference curves: 1) the standard CBP reference curve derived to cover 10% of the percent space by percent time plane (blue); 2) a reference curve based on 10% exceedance frequency and a temporal-spatial variance ratio of 1.0(red); and 3) a reference curve based on 10% exceedance frequency and a temporal-spatial variance derived from chlorophyll data(green).

Relative to the standard reference curves, the curve based on the observed variance ratio for chlorophyll is more restrictive of events where large portions of the population are out of compliance. For example, the CBP standard reference (blue) would allow 40% of area to exceed the criterion threshold up to about 6% of the time. The proposed chlorophyll reference curve (green) would restrict occurrences of 40% of area out of compliance to about 2% of the time. Conversely, the proposed curve (green) allows a higher frequency of events where a small percentage of space in out of compliance. For example, 10% of space is allowed out of compliance 36% of the time under the proposed curve and 27% of the time under the standard curve.

While there is mathematical and statistical logic underpinning this proposed chlorophyll reference curve, it is important to remember that it is based on parametric models and simplifying assumptions. It is recommended that validation exercises be performed to insure that the general shape of CFD curves generated from data collected in near reference conditions is approximated by the proposed curve.

4.3 Accommodating Seasonality in Reference Curves

The degree of acceptable exceedance can vary with season. For example, benthos are less tolerant of hypoxia in warmer water temperatures. In addition, the threshold criterion may never be exceeded in some seasons and frequently be exceeded in others. By combining seasons, the acuteness of a specific seasonal exceedence is diluted by data from the acceptable season(s). To some extent, seasonal differences can be accommodated by changing the threshold criterion among seasons. However, there may still be a need to develop separate reference curves by season.

5.0 Review CFD Statistical Properties Including Bias, Precision, and Inference.

The CFD as an assessment tool is a relatively new and unstudied concept. Its close relationship to the empirical distribution function does give some insight on the mathematical behavior of the CFD. In this section we review some of the properties of the CFD and discuss the complications that arise from these properties when the CFD is used as an assessment tool. After defining the population which determines the CFD, we go on to discuss the currently proposed sampling and estimation scheme, sources of error in the estimation scheme, and problems that result from these. The goal is to succinctly define these problems and elucidate possible solutions. This section will cover: the behavior of the CFD as a function of temporal and spatial variance, methods for construction CFD reference curves, the influence of sampling and estimation variance on the CFD shape, and feasible methods for developing statistical inference tools.

5.1 Review of CFD Properties

With any statistical application, it is important to distinguish between the true descriptive model underlying the population being sampled and the estimate of this model derived from the data collected in a sample. As described above, the CFD has a data driven definition where the CFD is constructed based on a sample from a population for some water quality parameter. This population is a continuous random process over space and time.

In order to quantify the statistical properties of the CFD, the CFD is defined in terms of a population of experimental units. This approach is a discrete approximation of the continuous random process in both time and space. However, the estimation scheme involves interpolation to discrete units in a spatial dimension and discrete days in the temporal dimension. To facilitate an understanding of the relation of the estimator to the true population, it seems reasonable to use a discrete approximation as the model for the true population.

5.2 Defining the CFD Ideal

The population will be defined as having different sizes of experimental units in much the way we think of a population that gives rise to a nested design or repeated measures design. The Chesapeake Bay will be partitioned into segments. Assessment will be done for each segment based on a three year record of the segment. Thus a three year period for the segment defines the entire population that will be partitioned into experimental units. The continuous time dimension is partitioned into days to form the primary units which are the state of a segment for a day. Call this a **Segment-Day**. Let there be **M** segment-days in the assessment period (typically 3 x 365). The continuous spatial dimension is partitioned into **N** 3-dimensional **cells** (may range from hundreds to thousands). The state of each cell for a day will be a unit nested within the segment-day. The attribute of interest will be a measure of water quality for each cell for a day. Examples might be the mean level of Chlorophyll-a in the cell for one day or the minimum of dissolved oxygen in the cell during the day. Let **Y** be a random variable for the attribute of interest and consider the following model

$$Y_i(\boldsymbol{s}_j) = \mu + \alpha_i + \beta_i(\boldsymbol{s}_j) \qquad \textbf{Eqn 5.1.1.1}$$

the vector $\underline{\alpha}$ will be assumed to have expectation $\underline{\mathbf{0}}$ and variance Σ_α and
each vector $\underline{\boldsymbol{\beta}_\mathbf{i}}$ will be assumed to have expectation $\underline{\mathbf{0}}$ and variance $\Sigma_{\beta i.}$
i is the ordinal index for days and
s is a vector valued ordinal for spatial location.

Under this model, Σ_α defines the correlation over time at the segment-day level and $\Sigma_{\beta i}$ defines correlation over space that occurs cell to cell within a day.

Let $\mathbf{C}_i(\boldsymbol{s}_j)$ be a collection of threshold limits that define the acceptable criterion for the measured attribute. If $\mathbf{Y}_i(\boldsymbol{s}_j)$ exceeds $\mathbf{C}_i(\boldsymbol{s}_j)$ in a cell, that cell is called degraded. The criterion is allowed to vary in both time and space so that in theory each $\mathbf{Y}_i(\boldsymbol{s}_j)$ might be compared to a unique $\mathbf{C}_i(\boldsymbol{s}_j)$.. It may vary over time because different levels of **Y** may be acceptable in different seasons. It may vary over space because different levels of **Y** may be acceptable in different salinity regimes so that even within a segment, **C** may be a function of salinity. As a rule, it is anticipated that $\mathbf{C}_i(\boldsymbol{s}_j)$ will be constant for regions of space and time such as salinity zones and seasons.

Now convert the measured attribute $\mathbf{Y}_i(\boldsymbol{s}_j)$ to a Boolean response as follows

$$\mathbf{TY}_i(\boldsymbol{s}_j) = \mathbf{I}(\mathbf{Y}_i(\boldsymbol{s}_j) > \mathbf{C}_i(\boldsymbol{s}_j)) \quad \begin{cases} = 1 \text{ if } \mathbf{Y}_i(\boldsymbol{s}_j) > \mathbf{C}_i(\boldsymbol{s}_j) \\ = 0 \text{ otherwise} \end{cases} \qquad \textbf{Eqn 5.1.1.2}$$

Thus **TY** takes the value 1 when **Y** exceeds the threshold defined by **C**. Using **TY**, we summarize the state of a segment on one day as the fraction of that segment that is out of compliance

$$P_i = (1/N)\sum_{j=1}^{N} TY_i(s_j) \qquad \textbf{Eqn 5.1.1.3}$$

The CFD that we wish to estimate is one minus the cumulative distribution function of the P_i's. If $P_{(i)}$ represents the ordered values of the P_i's for any assessment period, then let

$$G(p) = (1/M)\sum_{i=1}^{M} I(P_{(i)} \geq p) \qquad \textbf{Eqn 5.1.1.4}$$

G defines the CFD that if it were known would be used for an exact assessment. The cumulative distribution function is determined by the mean and variance of the ideal population. This population is defined with a spatial variance component and a temporal variance component. The final CFD shows the cumulative percent of time that a certain percent of space is below the criterion threshold. If the CFD shows that water quality in a segment is beyond the threshold for too much space and too much time, then the segment is classified as impaired.

For one assessment period, **G** can be considered exact as defined above, but recognize that even this is only one observation of the many possible observations of **G** that could result from sampling different assessment periods.

Assume for simplicity that **Y** is normal. If Σ_α were 0 so that **Y** had constant expectation over time and if Σ_β were of the form $\sigma^2\mathbf{I}$ then each cell on each day would have constant probability of exceeding a constant value of **C** given by $\mathbf{1 - \Phi(C)}$ where Φ is the normal cumulative density function. In this greatly simplified scenario, $\mathbf{P_i}$ would be the outcome of **N** independent Bernoulli trials. The ideal CFD would be the cumulative distribution function of **M** outcomes of a binomial random variable with **N** trials. If we allow Σ_β to have positive off diagonal elements**,** then the Bernoulli trials become dependent (i.e. adjacent cells are more likely to either both exceed or both meet the standard than distant cells). This should make the distribution of the $\mathbf{P_i}$ more variable than under the independent binomial model, but the expectation of $\mathbf{P_i}$ would be constant over time. If we relax the assumption that Σ_α is 0, then the expectation of the $\mathbf{P_i}$ would vary over time which would increase the variability of the $\mathbf{P_i}$ even more.

Under the simplifying assumptions of independence, constant mean, and constant variance, it is possible to obtain an analytical formulation for the CFD based on the parameters of **Eqn 5.1.1.1**. However, when the more realistic time dependent, space dependent model with seasonal nonstationarity is considered, an analytical formulation is not tractable. The lack of an analytical formulation for this estimator under realistic dependence assumptions, e.g. non-trivial Σ_α and Σ_β, points toward computer intensive simulation techniques to develop statistical inference procedures for this problem. None-the-less, it is interesting to consider the behavior of the CFD under the simplified model.

5.3 CFD Behavior under a Simplified Model

In what follows, the behavior of the CFD under various parameter formulations for Equation 5.1.1.1 are presented in graphical form. There are four parameters involved: μ the population mean, σ_t the temporal variance, σ_s the spatial variance, and **C** the criterion threshold. In the examples that follow, three of these parameters are held constant and the fourth is varied to illustrate the effect of the varied parameter.

In this exercise, the parameters of Equation 5.1.1.1 are simplified as follows: $\Sigma_\alpha = \sigma_t \mathbf{I}$ and $\Sigma_\beta = \sigma_s \mathbf{I}$**,** where **I** is the identity matrix. Thus in both the temporal and spatial dimensions, independence and constant variance is assumed.

Example 1. Example 1 considers the effect of changing the population mean on the shape of the CFD.

Table 5.1. Parameter values and color key for the family of curves shown in Figure 5.1.

μ	σ_t	σ_s	c	color	curve number
5	1	1	5	Red	1
4	1	1	5	Orange	2
3	1	1	5	Brown	3
2	1	1	5	Green	4
1	1	1	5	Blue	5

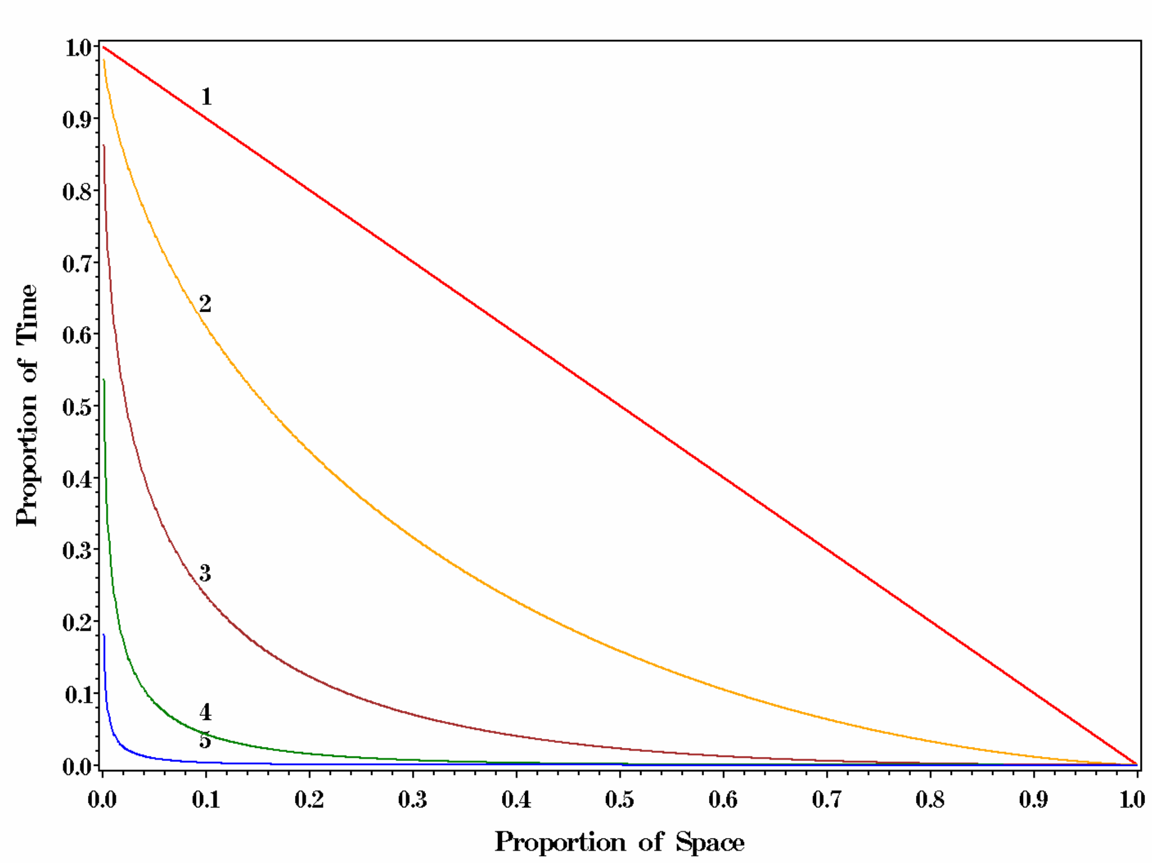


Figure 5.1. A family of curves illustrating the behavior of the CFD as the population mean decreases from the criterion threshold. The parameter values for each curve and the corresponding color are given in the following Table 5.1

Note that when the population mean is equal to the criterion threshold, the CFD is a diagonal line from upper left to lower right (Figure 5.1, red). This is largely an artifact of using symmetric distributions, the normal, for both the time and space variance components. That is, when the population median is equal to the criterion threshold, we expect an average of 50% noncompliance over time and we expect the exceed 50% noncompliance 50% of the time.

As the overall population mean decreases from the criterion threshold, the family of curves tends to move from the diagonal line toward the lower left corner. Thus a reference population, which should have a small probability of exceeding the criterion threshold might have a shape similar to the green curve. This illustrates the importance of the shape of the CFD in measuring compliance. A CFD from a highly compliant population will tend to hug to lower left corner similar to the blue and green curves. As the population mean approaches the criterion threshold, the CFD approaches the red line. If the population mean were to exceed the criterion threshold, the CFD would tend toward the upper right corner.

Example 2. Example 2 considers the effect of changing the temporal variance on the shape of the CFD. Note that the population mean is held constant at 3 which corresponds to the yellow curve of the preceding example.

Table 5.2. Parameter values and color key for the family of curves shown in Figure 5.2.

μ	σ_t	σ_s	c	color	curve number
3	1	1	5	Red	1
3	2	1	5	Orange	2
3	3	1	5	Brown	3
3	4	1	5	Green	4
3	5	1	5	Blue	5

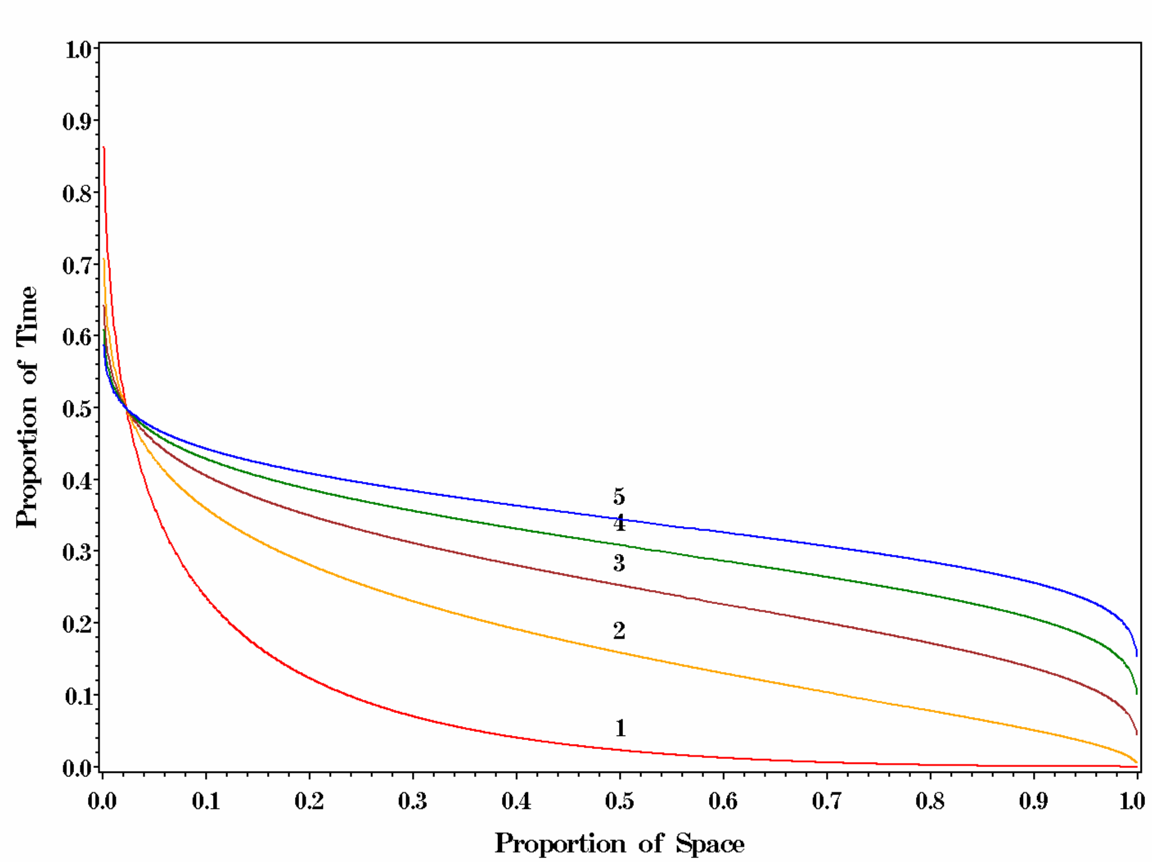


Figure 5.2. A family of curves illustrating the behavior of the CFD as the temporal population variance increases. The parameter values for each curve and the corresponding color are given in Table 5.2. Note that the red curve here has the same parameters as the yellow curve of Figure 5.2.

As temporal variance increases, the frequency of large proportions of space going out of compliance increases (Figure 5.2, lower right). Conversely, the frequency of small proportions of space out of compliance (i.e. large proportions of space being in compliance) decreases (Figure 5.2., upper left). That is, shifting the daily mean either down or up tends to shift the entire segment toward or away from compliance.

In preparing water clarity CFDs for reference areas defined by having successful SAV beds, it is not unusual to find a curve shape similar to Figure 5.2 orange or yellow curves. This pattern suggests that SAV is tolerant of ephemeral events of spatially broad degraded water clarity. If water clarity is persistently degraded over portions of the area, SAV may be impaired.

Example 3. Example 3 considers the effect of changing the spatial variance on the shape of the CFD. Again the population mean is held constant at 3 which corresponds to the yellow curve of the first example.

Table 5.3. Parameter values and color key for the family of curves shown in Figure 5.3.

μ	σ_t	σ_s	c	color	curve number
3	1	1	5	Red	1
3	1	2	5	Orange	2
3	1	3	5	Brown	3
3	1	4	5	Green	4
3	1	5	5	Blue	5

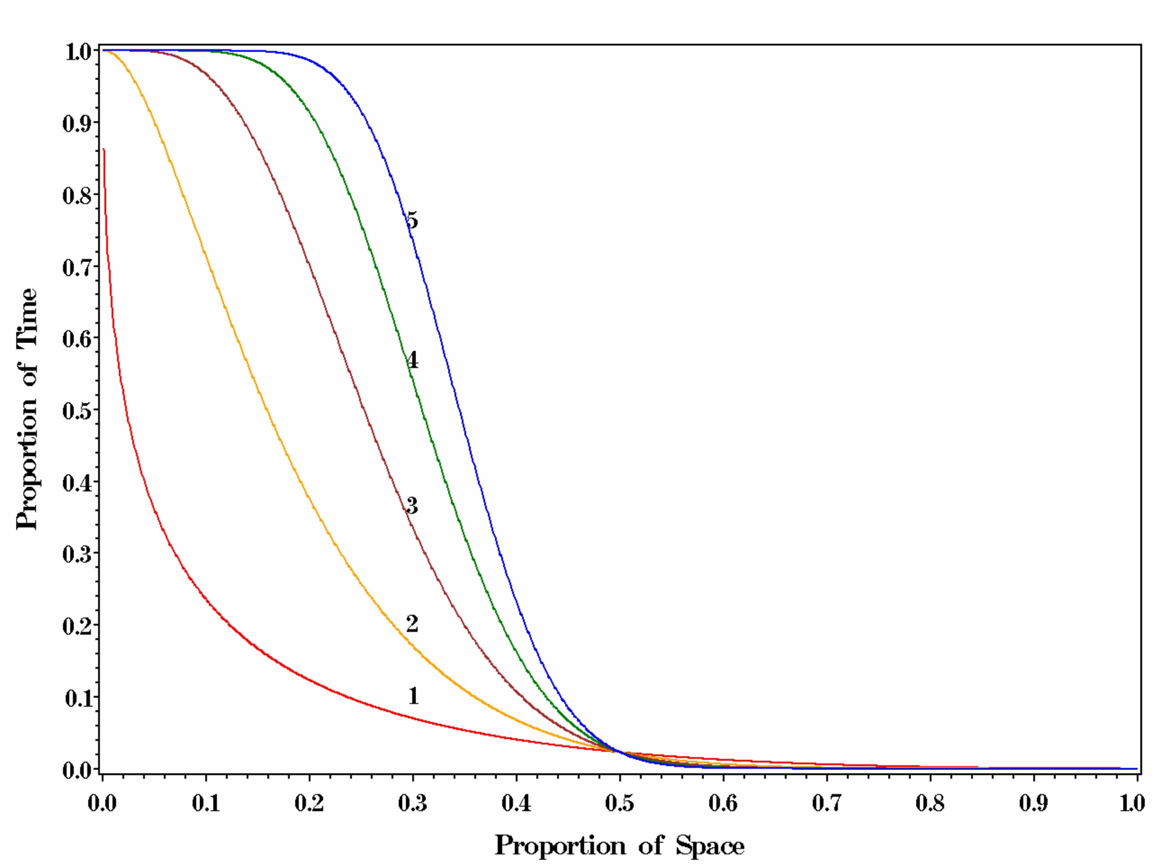


Figure 5.3. A family of curves illustrating the behavior of the CFD as the spatial population variance increases. The parameter values for each curve and the corresponding color are given in Table 5.3.

Increasing the spatial variance results in a family of curves that is complementary to those that follow an increase in temporal variance. Increasing spatial variance results in a higher frequency of small proportions being out of compliance. It is not so much an all-or-nothing phenomenon.

Example 4. Example 4 considers the effect of changing both temporal and spatial variance on the shape of the CFD.

Table 5.4. Parameter values and color key for the family of curves shown in Figure 5.4.

μ	σ_t	σ_s	c	color	curve number
3	1	1	5	Red	1
3	2	2	5	Orange	2
3	3	3	5	Brown	3
3	4	4	5	Green	4
3	5	5	5	Blue	5

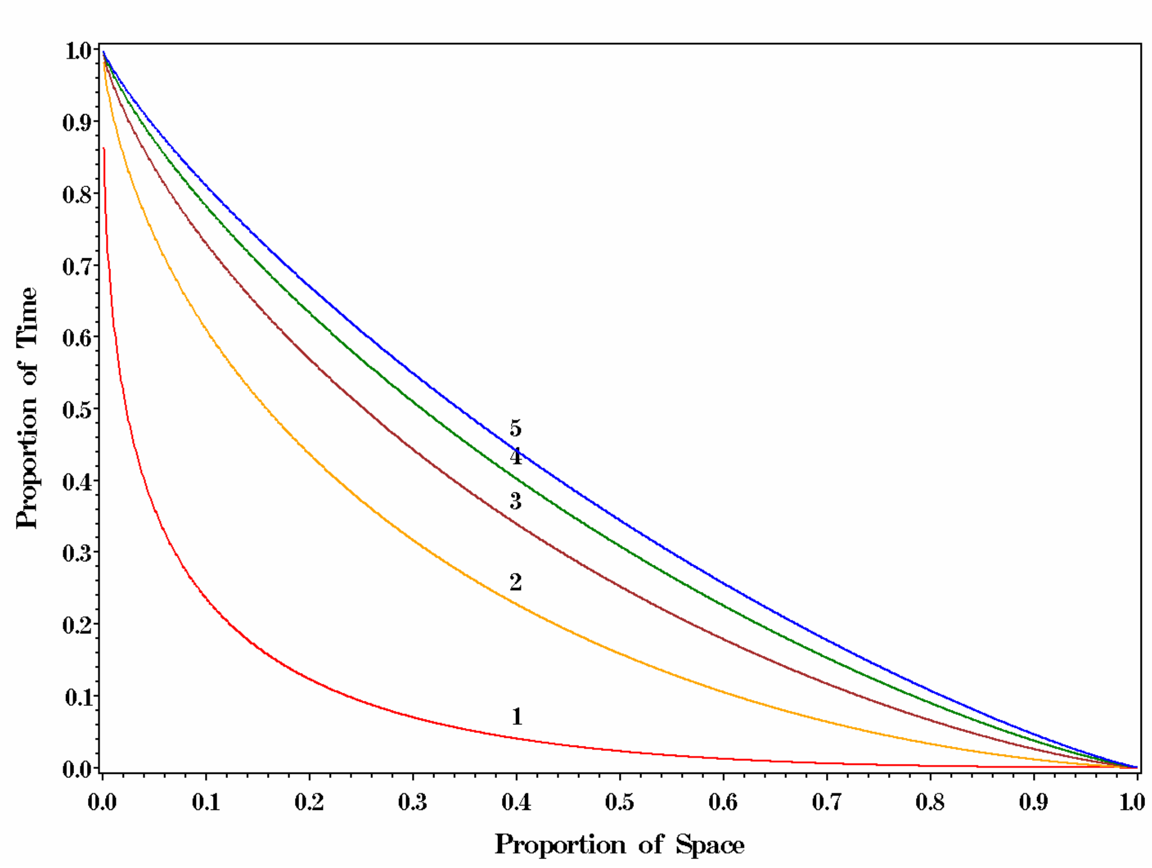


Figure 5.4. A family of curves illustrating the behavior of the CFD as both temporal and spatial variance increases. The parameter values for each curve and the corresponding color are given in Table 5.4.

Increasing the spatial and temporal variance together has the opposite effect of decreasing the population mean. The CFD tends to move in a direction of noncompliance. Thus compliance as measured by the CFD depends on the relative values of the population mean, the temporal and spatial variance, and the criterion threshold. Increasing the population mean has the same effect as decreasing the criterion threshold. Increasing population variance has the same effect as increasing the mean or decreasing the criterion threshold. In a sense, the CFD is measuring the distance between the population mean and the criterion threshold in units of variance analogous to a simple t-test. A nuance introduced here that has no analogy in the t-test is that the ratio of spatial to temporal variance controls the symmetry of the curve.

5.4 Uncertainty and Bias

In Section 5.1., it was shown that the shape of the CFD is a critical element to determining compliance. Thus it is important that this shape be primarily determined by the state of compliance of a segment and not be influenced by factors not relating to the status of compliance. Because the CFD is constructed based on data that are a sample from the whole, it is clear that some uncertainty in the CFD will result. In addition, the CFD is a function of the empirical distribution function (EDF) of fraction of space in compliance. The shape of this EDF is determined by the mean and variance of the sample. Thus any factor, such as sample size, that affects the precision of the fraction of space estimate, will affect the shape of the CFD. In this section we review the effect of noncompliance factors on the shape of the CFD.

Sample Size and Shape

As noted, because the CFD is a function of the EDF of estimates of "fraction of space", any factor affecting the precision of the estimate of fraction of space in exceedance will affect the shape of the CFD. In particular, the number of samples used for each p-hat (% exceedence) will affect precision. For a given segment, this fraction will be estimated more accurately if twelve samples are used to form the interpolated surface rather than six. Because of unknown spatial dependence in the data, it is difficult to analytically quantify the magnitude of this sample size effect. Therefore simulation analysis was employed to address this issue.

Numerous simulation tests were performed. These begin with a simulation of structurally simple data that have no temporal or seasonal trend and progress to simulated data that mimic the temporal and spatial structure of observed data. Because the results from this latter simulation are most relevant, these are the results that are presented and discussed.

Simulation Experiment

Simulated data were created to mimic the properties of surface chlorophyll in the Patuxent estuary. Data were created to fill a 5 by 60 cell grid which approximates the long and thin nature of an estuary. These data have mean zero and a spatial variance-covariance structure chosen to approximate the spatial variance-covariance structure of cruise-track chlorophyll observed in the Patuxent estuary. Thirty six grids of data were simulated to represent 36 months in a three year assessment period. The temporal and spatial trends were added to the simulated data by adding in means computed for each month and river kilometer during the period Jan 1, 1991to Dec 31, 1993. Simulated data were created using the "grf" function of the Geostatistical Package "geoR" of the R-package.

After the full population of data was simulated for 3 year assessment period, a sampling experiment was conducted to assess the effect of sample size on the shape of the CFD.

First, as a benchmark, a CFD was computed using all of the simulated data. To simulate the effect of sampling, a sample of fixed size was randomly selected from each the 36 5x60 grids of data. Using these samples, kriging (krige.conv function of geoR) was used to populate each monthly grid with estimates. These estimated chlorophyll surfaces were used to compute an estimate of the CFD which was graphically compared to the benchmark (Figure 5.5). For a fixed sample size, the process was repeated until it was clear whether the differences between the benchmark CFD and the estimated CFDs were due to variance or bias. To assess the effect of sample size, the process was repeated for several sample sizes.

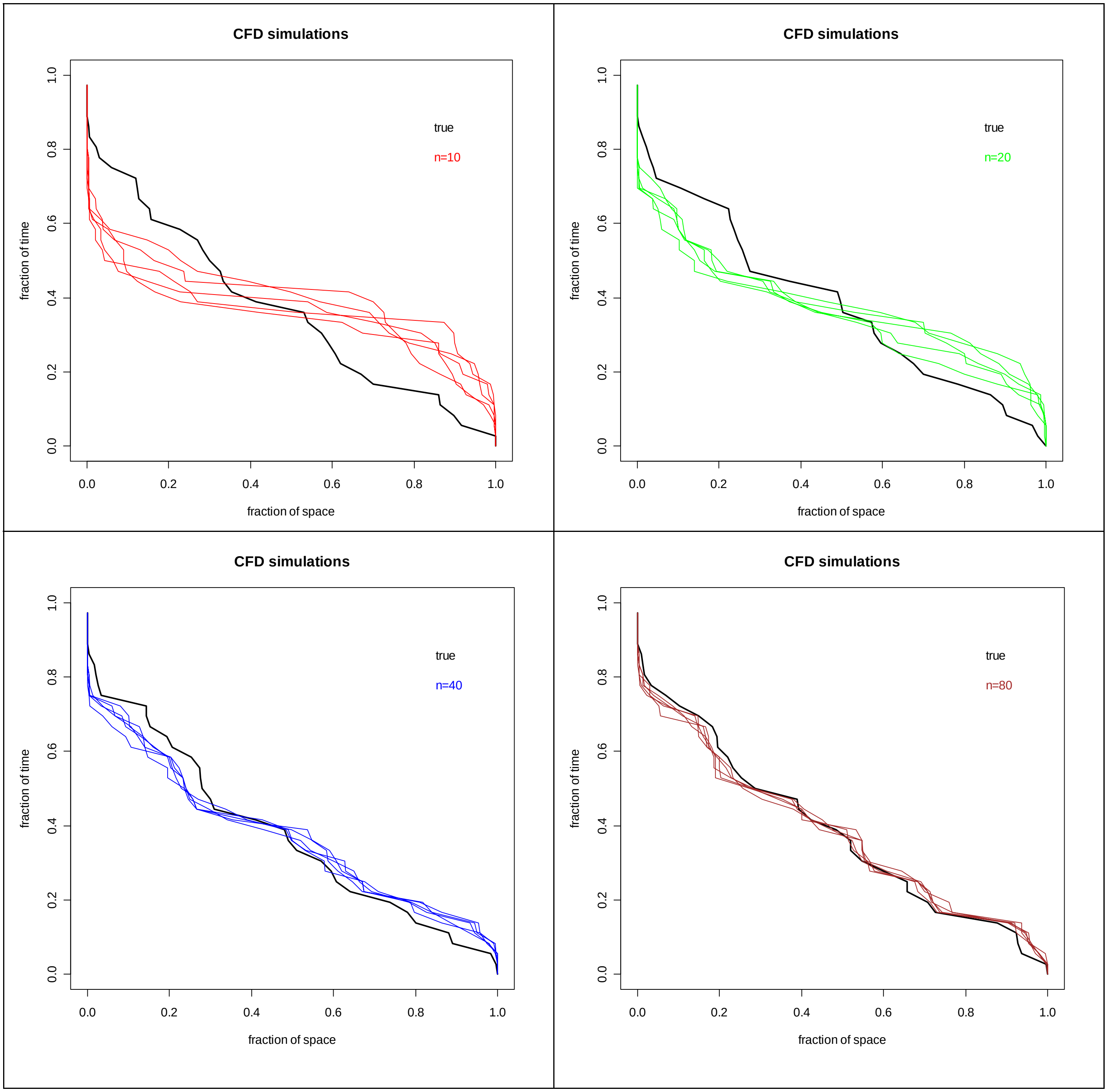


Figure 5.5 Illustration of the effect of sample size (n) on the shape of the CFD for sample sizes 10, 20, 40, and 80.

The effect of sample size on the shape of the CFD is consistent with expectations based on the relation of the CFD to the empirical distribution function (Figure 5.5). As sample size decreases, the variance of the estimated values of fraction of space increases. This increase in variance results in the estimated CFD being to the left of the true curve for low values of fraction of space and to the right of the true curve for high values of fraction of space. This assessment has been repeated many times, varying the threshold criterion, systematic vs. random sampling, the level of variability in the simulated data, and so on. This sample size effect persists for every case where realistic estimation is employed.

Sampling Scale and Shape

As shown above (Figures 5.2-5.4) the shape of the CFD is a function of the ratio of temporal and spatial variance. To the extent that the ratio of these variance components in the data represent the true state of nature, this is acceptable. However, under a model with strong spatial and temporal dependence, the ratio of these variance components might be influenced by the scale of sampling in the spatial and temporal dimensions. For example, samples collected far apart in time might reflect higher variance than samples collected close in time. If the ratio of temporal and spatial variance is influenced by the density of sampling in each dimension, then experimental design will have an effect on the asymmetry of the CFD estimate.

5.5 Confidence Bounds and Statistical Inference

An investigation into the use of conditional simulation to obtain confidence bounds for the CFD showed that not only is this a promising technique for statistical inference, but also has potential in correcting bias associated with sample size effects that has been identified as a central problem in implementing the CFD approach. Correcting the bias of the CFD due to the sample size effect is important in obtaining confidence bounds on the CFD that cover the true CFD for a segment. Because bias correction is an important first step, this aspect of the conditional simulation experiments will be discussed first. Conditional simulation will then be evaluated in its efficacy in obtaining confidence intervals.

This section first outlines the basic concept of conditional simulation and provides an algorithm that employs conditional simulation to estimate confidence bounds for the CFD. The results of this experiment support the potential of conditional simulation for correcting the sample size bias. A heuristic discussion of the mechanism underlying this adjustment for sample size effect is presented with the hope of motivating additional analytical investigation of this effect.

Conditional simulation (Journel, 1974; Gotway, 1994) is a geostastical term for simulating a population conditional on information observed in a sample. In the case of kriging, a sample from a spatial population is used to estimate the variogram and mean for the population. The conditional simulation procedure generates a field of simulated values conditioned on the estimated mean and variogram from the sample. To the extent that the estimated mean and variogram approximate the true mean and variogram and the assumed distribution is a reasonable model for the true distribution, repeated simulations of this virtual population will represent the variability typical of the true population. It follows that statistics computed from the conditionally simulated fields will represent the expected variability of statistics from the true distribution. The CFD is a graphical representation of ordered statistics of percent compliance over time and it is a reasonable to assume that repeated conditional simulations will lead to effective confidence bounds for the CFD.

Conditional Simulation Methods

In the computation of the CFD, conditional simulation is implemented at the interpolation step for each month. Interpolation produces an estimate of the spatial surface of the target parameter. From that estimate of the surface is obtained an estimate of the percent of noncompliance. Using conditional simulation, the surface can be reconstructed 1000 times. From the 1000 simulated surfaces are computed 1000 estimates of the proportion of noncompliance. When this is repeated for each month for say 36 months, the result is an array of 1000 sets of 36 values of the proportion of noncompliance. Each of the 1000 sets of 36 can then be ranked from largest to smallest to compute a CFD in the usual way which results in 1000 CFD estimates. The variability among these 1000 CFDs can be used to estimate confidence intervals.

To evaluate this concept, the following simulation experiment was conducted

1) The first step is to simulate a population that will be considered the "true" population for this exercise. A grid of dimensions 5x60 is populated using an exponential spatial variance model with variogram parameters set to (0.00625026, 2.67393446). These variogram parameters were estimated from Patuxent cruise track chlorophyll data. This grid is populated 36 times to represent 36 months. The mean and variogram are held constant for the 36 simulations to create a simplistic case with no seasonal or spatial trend. Using this set of data, the CFD is computed in the usual way and this is considered the "true" CFD.

2) A sample of size 40 is selected from each of the 36 simulations at random locations on the grid. Ordinary kriging is used to estimate the spatial surface for each simulation and from these 36 estimates of the monthly spatial surfaces, a CFD is computed. This is called the 'estimated' CFD.

3) For each of the kriged monthly surfaces, 1000 conditional surfaces are simulated based upon the mean and variogram estimated from the sample data. The Cholesky decomposition is used to reconstruct the covariance structure indicated by the estimated variogram. The conditionally simulated surfaces were processed to yield 1000 estimates of the proportion of noncompliance. The 1000x36 noncompliance values are used to compute 1000 CFDs, which are called the population of "conditionally simulated" CFDs.

4) Each "rank position" of the monthly ordered proportions of noncompliance has 1000 values in this simulated population. To assess variability in the simulated population, graphs of the miniumum, the 2.5th percentile, the 50th percentile, the 97.5th percentile, and the maximum at each rank position are plotted to illustrate a 95% confidence envelop for the CFD (Figure 5.6).

To test this procedure under various conditions, this basic simulation exercise was repeated varying the sample size and adding temporal and spatial trend to the simulation of the "true" population to reflect conditions more similar to real populations.

Conditional Simulation Results

The results of this simulation exercise are presented graphically. In Figure 5..1 the black line represents the CFD computed for the true population computed from the original data. The red line is the estimated CFD computed from kriging estimates based on samples from the true population. The brown lines represent the min and max of the 1000 conditionally simulated CFDs. The green lines represent the 2.5 and 97.5 percentiles of the 1000 conditionally simulated CFDs, which is the proposed 95 percent confidence interval. The blue curve is the median of the 1000 CFD curves.

Bias Assessment

The results in Figure 5.6 are unusual in several respects. First note that the red curve shows the typical sample size bias for the CFD as described above (n=40). Relative to the true CFD (black) the estimated CFD is below the black line for half the curve and above the black line for the remainder. The first unusual feature is that the distribution of the conditionally simulated CFD curves is not centered on estimated CFD. In fact the estimated CFD is not completely within the bounds (min, max) of the conditionally simulated population. A surprising feature is that the median of the simulated population tracks fairly well with the true CFD (black). It is clear that the simulated CFD population is estimating something other than what is estimated by the estimated CFD (red). At the same time, it appears that the median of the simulated population is a good estimator of the true CFD and the proposed confidence bands (green) is reasonable confidence envelop about the true CFD.

What follows is a heuristic explanation for why CFD computed from conditional simulations might be a better estimator of the true CFD than a CFD computed from the kriging estimator. Additional analyses test whether this property might hold in general or is an artifact of the simple conditions (no spatial or temporal trend) under which this experiment was performed.

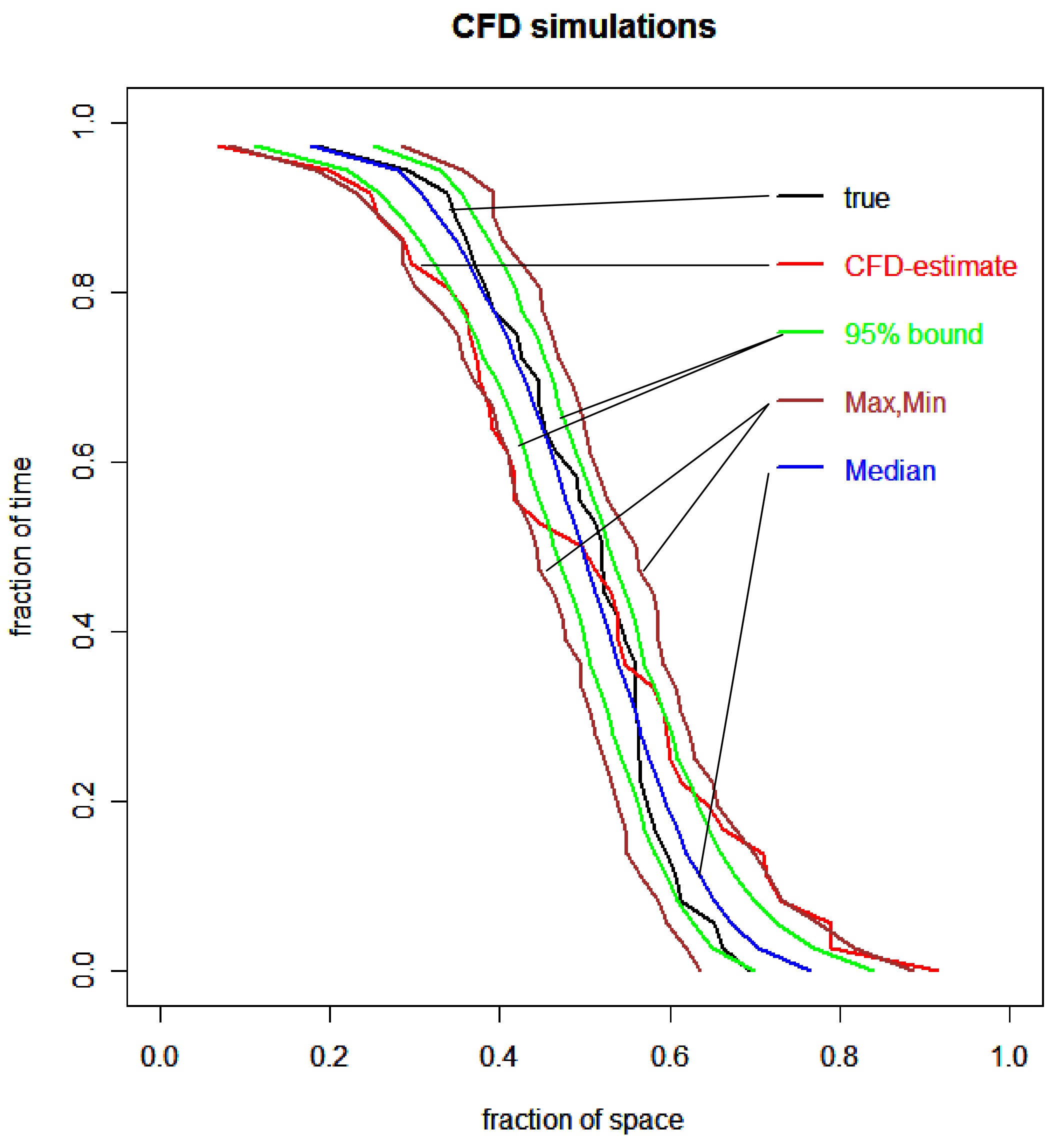


Figure 5.6. Confidence bounds computed based on quantiles of fraction of space computed on conditionally simulated surface estimates using variogram estimates from data. The base simulation has spatial correlation and no spatial or temporal trend. Sample size is 40.

In prior discussions we have noted that the CFD is the inverse of the CDF of the population of p's where p is fraction of space out of compliance with the criterion threshold. It is the variance of the p's that determines the steepness of the CFD: the smaller the variance, the steeper the CFD. In real applications, estimates of the p's have two important variance components. One variance component comes from true variance over time in the parameter being assessed. Another variance component comes from imperfect estimates due to sampling variability. In the base simulation with no spatial or temporal trend in the data, it is this second source of variance that controls the shape of the CFD.

Because the variance of the p's is critical to the shape of the CFD, consider the variance of p's computed from three sources in the experiment outlined above: 1) the true data, 2) a krig estimate based on a sample from the true data, and 3) conditionally simulated data based on a krig estimate of 2). To enhance our understanding of this comparison, the variance of the p's are discussed for two cases for each source. The first case assumes complete independence in the base simulation and does not use interpolation to estimate proportion of area out of compliance. This simplification allows us to easily infer the behavior of the CFD using analytical methods. The second case introduces an unknown spatial dependence in the base simulation and uses interpolated data to estimate the proportion of area out of compliance. These additional complexities make it difficult to implement analytical inference but conclusions may still be inferred by analogy to the simple independent case.

Consider the sequence of sources where the base simulations are generated under the simple constraints of constant mean, constant variance and the errors for each cell of the grid that are independent. For this case the exceedance probability is:

$$\mathrm{p} = 1 - \Phi((\mathrm{x_s} - \mu - C)/\sigma)$$

where : C is the criterion threshold,

$\mathrm{x_s}$ is the data at location s,

μ is the mean used in the simulation,

σ is the variance used in the simulation, and

Φ is the standard normal Cumulative Distribution Function.

The distribution of the true p's computed from all 300 cells of the 5x60 simulation grid would behave like that of a independent binomial with N=300 with a variance of (p(1-p)/300). From these independent data draw a sample of size 40. Using only the proportion of the sample that is out of compliance to estimate the p's, the distribution of the p's would be that of a independent binomial with N = 40 and variance (p(1-p)/40). Clearly the p's estimated from the sample of 40 have much larger variance than p's from the base simulation with 300 cells. Thus the true CFD computed using data from 300 cells will be steeper than the sample CFD computed from 40 data points. This pattern is illustrated by comparing the true CFD (black curve) and the estimated CFD (red curve) in Figure 5..1. This increase in the variance of the p's due to small sample size is the kernel of the sample size problem with the CFD. Now consider the behavior of p's computed from conditional simulations based on the sample. Compute $\bar{x}$ and s as estimates of ③ and ⑨ from the sample of 40 in the usual way. The conditional simulation is done by populating the 5x60 grid with data from a normal distribution with mean $\bar{x}_i$ and variance s^2_i. The exceedance probability for these simulated data for the ith month is

$$p'_i = 1 - \Phi((\mathrm{xs}_\mathrm{s} - \bar{x}_i - \mathrm{C})/\mathrm{s}_i)$$

where : xs_s is simulated data at location s
$\bar{x}_i$ is the estimated mean used in the conditional simulation, and
s_i is the estimated standard deviation used in the conditional simulation.

If the p' were constant over months, the variance of the p's estimated by conditional simulation would be (p'(1-p')/300). The sample size component of this variance has been standardized to 300 which is the same as the sample size component of the true p's, but the variability of conditionally simulated p's will be greater than that of true p's because estimates of $\bar{x}_i$ and s^2_i will vary over months. The parameter p and it's estimate p' will be close if $\bar{x}$ and s are close to ③ and ⑨. In the simple case with constant mean and independent errors, the CFD estimated by conditional simulation will better approximate the true CFD because both are based on binomial distributions with the same N and approximately the same p.

Now consider the same sequence of distributions where the assumption of independence is relaxed and interpolation of the data is used to estimate the proportion of noncompliance. The introduction of spatial covariance in the base simulation changes distribution of the true p's to a dependent binomial. The dependent binomial will have variance similar to an independent binomial with N < 300. Sample size that approximates the variance of the dependent binomial is termed Nb. The variance of the p's estimated from spatially dependent data is approximated by (p(1-p)/Nb) where Nb < 300 and thus the CFD from the independent case will be steeper than from the dependent case. The degree to which Nb is less than N will depend on the strength of the spatial correlation.

Next consider the effect of dependent data and interpolation on the distribution of the p's. When we interpolate the sample of 40 onto the grid of 300, the interpolated surface is smooth relative to the original data (compare green and red in Figure 5.2). Because of this increased dependence in the krig estimates, the estimates of p computed from the interpolated data behave more like binomial data with N=Ns (the sample size) than like binomial data with N=Nb (the number of grid cells). Because Ns is smaller than Nb, the variance of the population of p's computed from interpolated data will be greater. The greater variance explains why the red line in Figure 5.1 is much flatter than the black line.

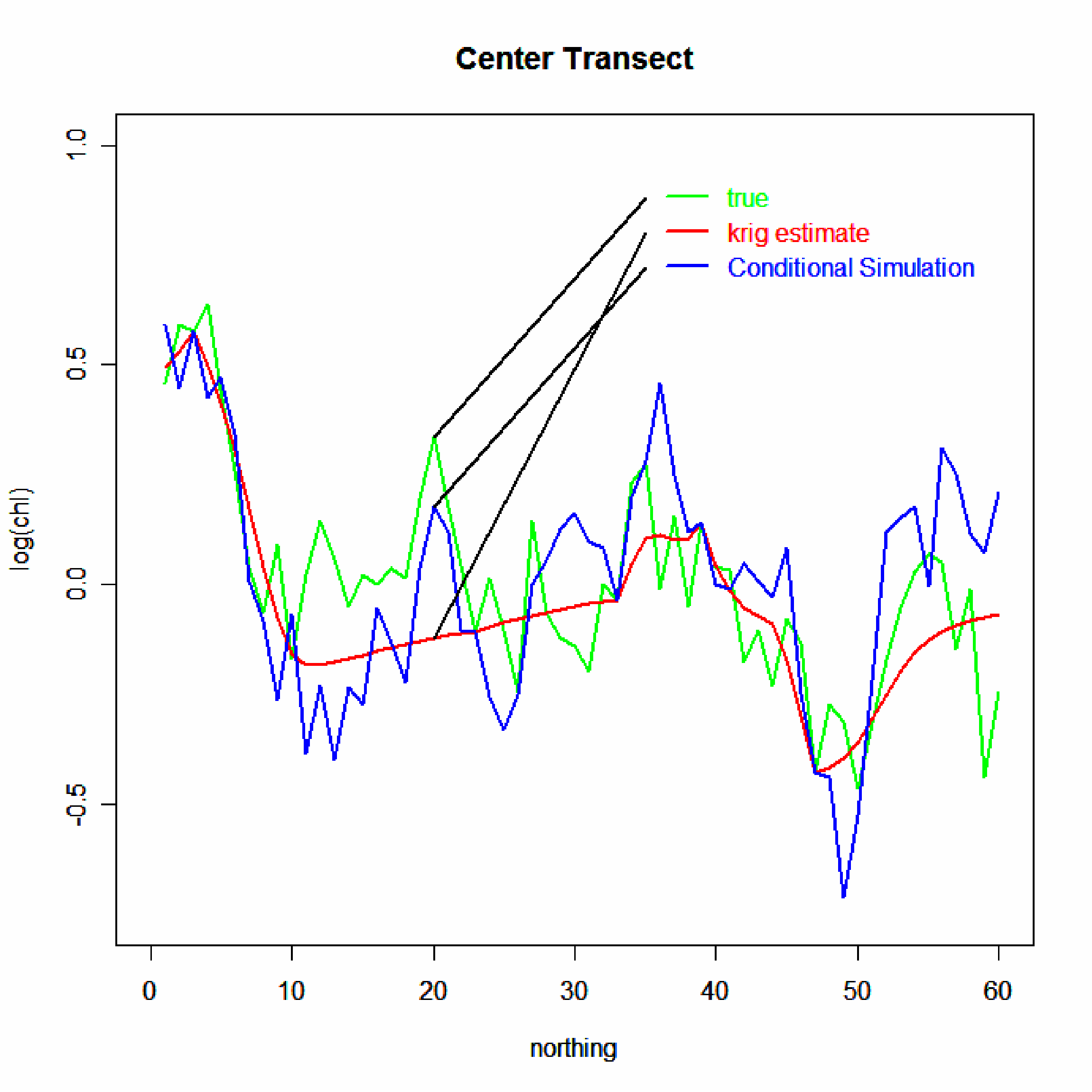


Figure 5.7. Simulated chlorophyll data, kriging estimates based on a sample of the simulated data, and conditionally simulated data where the simulation is conditioned on the data used obtain the kriging estimates.

Finally consider the effect of conditional simulation on the distribution of the p's. When data are conditionally simulated and the mean and variogram estimated from the sampled data are accurate, then the character of the simulated data will be similar to that of the true data (compare the green and blue in Figure 5.7). Like the simple independent case, the population of p's computed from the conditionally simulated data will have a binomial variance that is similar to a binomial with sample size Nb. The simulation experiment shows that the CFD computed from these conditionally simulated p's will have a shape similar to the true CFD. This effect is illustrated in Figure 5.6 where the median of the conditionally simulated CFDs (blue line) is more similar to the true CFD (black line) than is the CFD estimate based on kriging (red line). Additional analytical work is needed to formalize the heuristic concepts presented here, but this finding indicates a productive direction in developing statistical inference procedures in the CFD approach.

Confidence Intervals

The most successful technique for computing confidence bounds for the CFD were obtained using conditional simulation based on kriging interpolation of the sample data. The 95% confidence bands (green lines, Figure 5.6) are well centered over the true CFD (black line) for the simplistic case where the true data have spatial dependence but no spatial or temporal trends. When these simplistic assumptions are relaxed (Figure 5.8) and the true data are simulated to have spatial dependence and temporal and spatial trends similar to chlorophyll data from the Patuxent estuary, the confidence bands cover the true CFD in this case as well. Experiments that varied the sample size also produced confidence bands with good coverage.

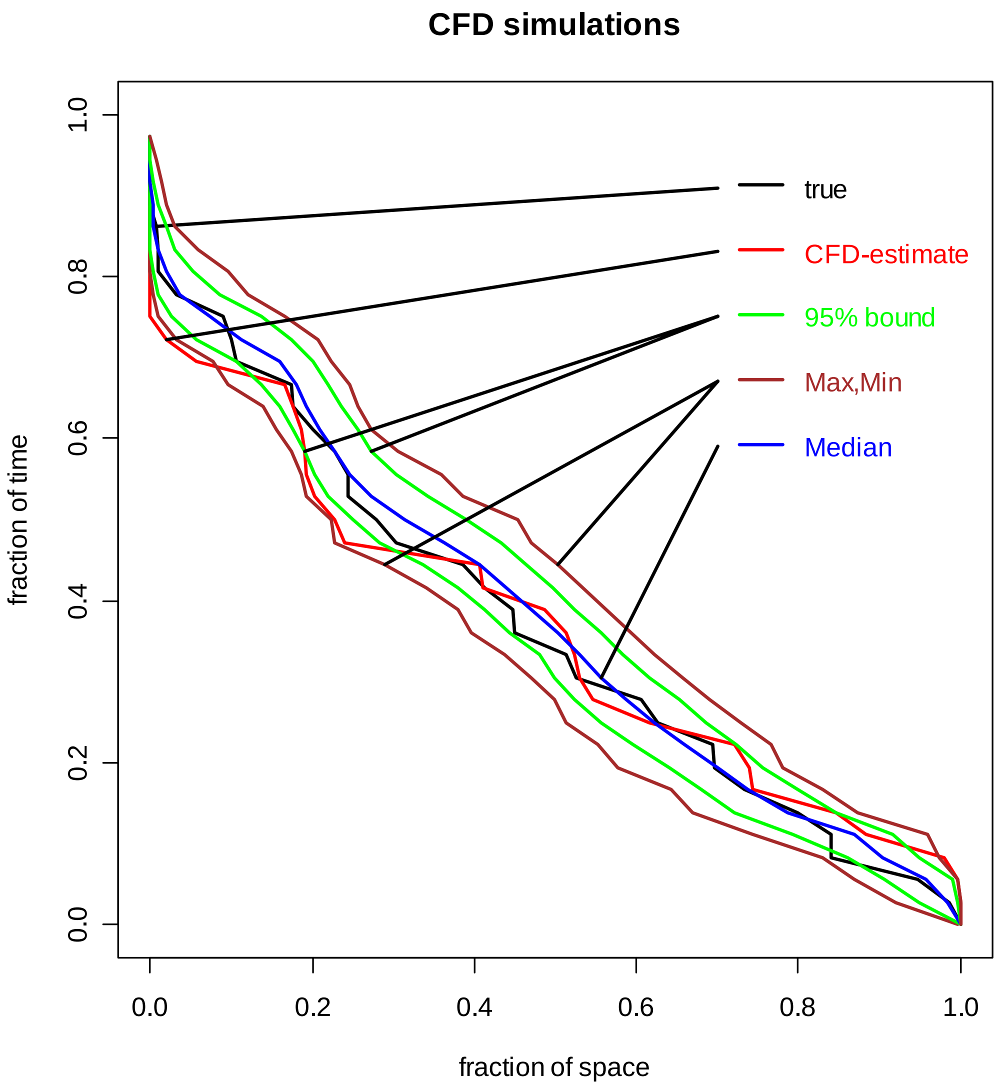


Figure 5.8. Confidence bounds based on quantiles of fraction of space computed on conditionally simulated surface estimates using variogram estimates from data. The base simulation has spatial and temporal trend estimated from Patuxent data. Sample size is 40

Additional evaluation of the confidence band procedure should include a series of confidence band coverage experiments to assess the true coverage rate in comparison to the nominal coverage rate (e.g. 95% in this example). This series of experiments should be conducted with simulated data where the simulations are designed to produce data with properties similar to the three primary assessment water quality parameters.

6.0 Findings – Scientific Acceptance of CFD Compliance Approach

6.1. CFD Approach as Best Available Science

This report represents an initial expert review of the CFD compliance approach. In addition the panel undertook simulation tests on the effects of 1) sample densities in time and space, 2) varying levels of attainment, and 3) varying degrees of spatial and temporal covariance. Further, trials of spatial modeling on fixed station Chesapeake Bay water quality data were conducted to begin to evaluate spatial modeling procedures. Based upon review of underlying theory, initial statistical assessments, and implementation feasibility, the panel finds that the CFD approach currently represents best available science in its application to water quality attainment determinations in the Chesapeake Bay. Using criteria for Best Science and Best Available Science developed by the American Fisheries Society and the Estuarine Research Federation (Sullivan et al. 2006), we list relevant attributes of the CFD approach (Table 6.1).

The CFD builds on important statistical theory related to the cumulative distribution function and as such, its statistical properties can be simulated and deduced. We have also shown that it is feasible to construct confidence ellipses that support inferences related to threshold curves or other tests of spatial and temporal compliance. Work remains to be done in understanding fundamental properties of how the CFD represents likely covariances of attainment in time and space and how temporal and spatial correlations interact with sample size effects. Further, more work is needed in analyzing biases across regions and designated use segments. The panel expects that a two-three year time frame of directed research and development will be required to identify and measure these sources of bias and imprecision in support of attainment determinations.

Through simulations of the CFD approach, it is feasible to analyze bias and error for both temporal and spatial sources of attainment variability. In particular, conditional simulations merit additional investigation as a relatively unbiased approach for supporting statistical comparisons among CFD curves. Much work remains to be done in understanding fundamental properties of how the CFD represents likely covariances of attainment in time and space. Still, the panel finds the approach feasible: one which merits additional development, testing, and application. Indeed, the CFD approach is beginning to attract scientific and management attention outside the Chesapeake Bay community.

As shown by analyses in previous sections, the approach can efficiently combine spatial and temporal data to support inferences on whether regions within the Chesapeake Bay attain or exceed water quality standards. On the other hand, we recognize substantial bias and imprecision can occur due to small sample size, non-independence in temporal trends, and inadequate spatial interpolations. More work is needed in analyzing these biases across regions and designated use segments. Further, the old saw of needing more samples cannot be ignored. In particular, the panel is optimistic in the application of continuous spatial data streams made available through the cruise-track monitoring program, and the promise of continuous temporal data through further deployment of remote sensing platforms in the Chesapeake Bay (CBOS web site, etc). These data sets will support greater precision and accuracy in both threshold and attainment determinations made through the CFD approach.

In classifying the CFD approach as best available science, we seek to make several important distinctions (Table 6.1). First, the CFD approach is a scientifically based approach based upon its clear purpose, conceptual and design framework, empirical procedures, documentation, and intent to develop rigorous statistical and review procedures (Sullivan et al. 2006, Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993). That the approach permits evaluation of uncertainty also supports its classification as best available science (Christman 2006). On the other hand, we do not believe that the CFD approach yet constitutes best science. Here, further analyses of underlying statistical properties of the approach (including sampling design and interpolation elements) and vetting by outside experts is needed. Indeed, although the CFD approach is beginning to get featured in scientific venues, it has not yet been reviewed as part of the scientific literature. The panel sees this as an overdue next step for necessary for its acceptance, further development, evaluation, and application.

The panel contrasted the CFD approach with existing state and jurisdictional water quality criteria and attainment procedures that are based strictly upon the observed sample, where site selection is not based upon probability sampling, inferences are not based upon error structure, and monitoring does not involve a scientifically rational design. Indeed, standard practice for assessing compliance with water quality criteria throughout the US is to sample monthly at a fixed set of stations and make judgments about compliance strictly from those samples. Sampling stations are typically located for convenience (e.g., bridge overpasses), there is reluctance to re-evaluate and change location (so as to maintain a time series at a fixed point), and no consideration is given to representativeness of the sample for the space/time not sampled. Thus the previous method used by the Chesapeake Bay Program, similar to the approaches used in other states, was simply based on EPA assessment guidance in which all samples in a given spatial area were compiled and attainment was assumed as long as > 10% of the samples did not exceed the standard. In this past approach all samples were assumed to be fully representative of the specified space and time and were simply combined as if they were random samples from a uniform population. This approach was necessary at the time because the technology was not available for a more rigorous approach. But it neglected spatial and temporal patterns that are known to exist in the standards measures. The CFD approach was designed to better characterize those spatial and temporal patterns and weight samples according to the amount of space or time that they actually represent.

Table 6.1. Evaluation of CFD approach as Best Science or Best Available Science according to AFS/ERF "Defining and Implementing Best Available Science for Fisheries and Environmental Science, Policy, and Management" (Sullivan et al. 2006).

Attribute	**Best Science**	**Best Available Science**	**Current State of Development of CFD Approach**
Clear Objective	YES	YES	Using biological response standards, combine available water quality in time and space to determine levels of attainment of Bay segments.
Conceptual Model	YES	YES	1. Bay divided into functional classifications – "Designated Uses." 2. Reference curves establish biologically relevant threshold levels for attainment. 3. CFD combines and weights equally temporal and spatial sources of water quality variability.
Experimental Design	NO	YES	1. Bay segments are quasi-stratified for water quality data collection. 2. Stratification of water quality data by designated units does not yet occur. 3. Seasonal assessment of water quality attainment through spatial interpolation and the CFD approach is feasible but incompletely developed.
Statistical Rigor	NO	YES	1. Procedures for quantifying uncertainty associated with sampling design, spatial interpolation and CFD approach are feasible but incompletely developed. 2. Procedures for interpolating water quality data are feasible but incompletely developed, particularly for 3-D interpolations of dissolved oxygen. 3. Procedures for testing inferences related to the CFD curve are feasible but incompletely developed.
Clear Documentation	YES	YES	CFD approach, water quality sampling design, and current interpolation procedures well documented in Chesapeake Bay Program Reports and on website.
Peer Review	NO	YES	1. CFD approach and sampling design upon which it is based has not been peer-reviewed in the scientific literature. 2. This report comprises the first external review by scientists with statistical expertise. 3. Grey literature reports produced by CBP received expert and stakeholder input.

6.2 The CFD approach and peer review

The panel views the CFD approach as innovative, one that has general application in water quality attainment assessments, but scientific acceptance of the approach will require that it is subjected to more extensive and targeted peer-review in the scientific literature. Because the CFD is a regulatory tool, it is particularly important that the approach is effectively communicated to the scientific community at large, for general acceptance but more critically for the sustained research and development that the CFD, as a nascent approach, requires. As highlighted elsewhere, bias and imprecision that can occur due to small sample densities, non-independence in temporal trends, and inadequate spatial interpolations. Such work is novel and should elicit interest among biostatisticians as it addresses questions of both fundamental and applied consequence.

Although, continued working groups, involvement through STAC of expert biostatisticians, and related reports such as this one will remain important in scientific acceptance of the CFD approach, the panel recommends immediate attention in subjecting the CFD to traditional peer review. One or several review papers should be submitted by CFD principals that lay out the theory, general approach and lists emergent scientific issues to stimulate other scientists to begin to address such issues. Several such papers might be appropriate given potential interest by biostatisticians and environmental and regulatory scientists. Scientific interest will also be garnered by public and stakeholder interest. The CFD approach here presents a challenge as it is complex in explanation. Still with careful diagrams and examples, a brochure on the CFD approach should be extremely useful in getting uninitiated scientists and stakeholders on the same page.

6.3. Biological Reference Curves

The success of the CFD-based assessment will be dependent upon decision rules related to the biological reference curves. These curves represent desired segment-designated use water quality outcomes and reflect sources of acceptable natural variability. The reference and attainment curves follow the same general approach in derivation – water quality data collection, spatial interpolation, comparison to biologically-based water quality criteria, and combination of space-time attainment data through a CFD. Therefore, the biological reference curve allows for implementation of threshold uncertainty as long as the reference curve is sampled similarly to the attainment curve. Bias and uncertainty are driven in CFD curves by sample densities in time and space. Therefore, we advise that similar sample densities are used in the derivation of attainment and reference curves. As this is not always feasible, analytical methods are needed in the future to equally weight sampling densities between attainment and reference curves.

Conceptually, the CFD approach builds on the underlying view that water quality criteria are surrogates for Designated Uses (regions that define ecosystem function). Implicit is a bottom up model based upon eutrophication, which is expected to diminish the designated use. Reference curves represent thresholds related to the functioning of designated use regions. Therefore, choice of reference regions or periods and sampling design in developing reference curve is critical to the implementation of a scientifically-rigorous CFD approach.   Choice of such regions is beyond the scope of this review, but we emphasize several relevant statistical issues in developing reference curves in Section 4.

7.0 Recommendations for Future Evaluation and Refinement of the CFD Assessment Methodology

As part of its conclusions, the STAC CFD Review Panel identified critical remaining issues that need resolution in the near future. The following is a list of critical aspects of that needed research. These research tasks appear roughly in order of priority. However, it must be recognized that it is difficult to formulate as set of tasks that can proceed with complete independence. For example, research on task 1 may show that the ability to conditionally simulate the water quality surface is critical to resolving the sample size bias issue. This discovery might eliminate IDW as a choice of interpolation under task 3. The Panel has made significant progress on several of these research tasks and CBP is encouraged to implement continued study in a way that maintains the momentum established by this research group (Table 7.1.).

1. <u>Effects of Sampling Design on CFD Results</u> - The CFD is a special case of an unbiased estimator for a cumulative distribution function of a population. Like the cumulative distribution function, the CFD is a function of the mean and the variance of the population being assessed. And the better the mean and variance are characterized with sample data, the more accurate the shape of the CFD will be. As the sampling density increases, the estimated CFD begins to approach the true CFD. However, if the sampling density is low, then sampling error could become important and there is potential that it could affect the shape of the CFD and ultimately the accuracy of the compliance assessment. Furthermore the potential for the sample size to affect the shape could create a compliance assessment bias if the reference curve and assessment curve are based on different sampling densities. Conditional simulation methods developed by STAC panel members showed promise toward resolving these issues and mitigating potential biases caused by differences in sample size.

2. <u>Statistical inference framework for the CFD</u> - It is important in a regulatory process to differentiate an exceedance that is small and might have resulted from chance variability from those that are large and indicative of an inherent problem. This differentiation will require mathematical tools to quantify the variability in the CFD that occurs as a result of sampling. The STAC panel made progress on this issue by demonstrating a confidence interval procedure based on conditional simulation associated with kriging. It remains to be assessed whether or not confidence intervals produced by this algorithm perform at the nominal level of coverage, fore example, does a nominally 95% CFD confidence interval cover the true CFD 95% of the time.

3. <u>Choice of Interpolation Method</u> - The STAC panel considered several interpolation methods and outlined the features of each. Those features illustrate tradeoffs between ease of implementation and maximizing the information garnered from the data. Further work is needed to compare the features to the requirements of wide-scale implementation of assessment procedures and formulate a plan for tractable implementation that results in credible assessments.

One strategy is to implement easily performed analysis (e.g. IDW) as a screening tool to identify cases where compliance / non-compliance is clear, and then implement more labor intensive methods (e.g. kriging) for cases where compliance is more difficult to resolve. One difficulty with implementing a full comparison of methods is that implementation of each method requires considerable work in terms of setting up file systems, interfacing software and data, and coupling the considerable bathymetry data of the bay. Thus it would be prudent to narrow the choices based on theoretical considerations where possible.

4. Three-Dimensional Interpolation - Assessments of the dissolved oxygen criteria require three-dimensional interpolation. However, the field of three-dimensional interpolation is not as highly developed as that of two-dimensional interpolation. While the mathematics of each method extend easily to three dimensions, there are relatively few examples of 3-D interpolation available in the literature and issues such as data density requirements for reliable results are not well studied. Efforts are needed to further evaluate research in three-dimensional interpolation and seek additional outside scientific input and review with the goal of implementing the best available technology for this aspect of criteria assessment. One of the first efforts under this task is a study of the 3-D variance stucture of the data to be interpolated. A short term option is to implement the optimal 2-D interpolator in layers as is done with the current IDW interpolator.

5. High Density Temporal Data - As currently formulated, assessment for most of the open-waters of the Bay are based on “snapshots” in time of the spatial extent of criteria exceedence estimated via interpolation. Data collected for use in interpolation are actually spaced over multiple days due to the large expanse over which sampling must be conducted. It is clear that technology is becoming available that will produce high density data in both space and time. Interpolation should accommodate data that are collected densely in space. However, it is unclear how the CFD process will accommodate data that are high density in time. Further work is needed to evaluate methods to fully utilize the temporally intensive data that is currently being collected.

The panel discussed several mechanisms for the CBP to make progress on challenging tasks ahead (Table 7.1). We recommend that a review panel oversee the tasks over the next 3-5 year time frame. This panel would periodically review trials and other products conducted by individual external scientists (academic scientists or consultants) and existing teams of CBP scientists (e.g., the Criteria Assessment Protocols (CAP) workgroup). Tasks 1 and 2 are most immediate and critical and we recommend that these tasks by contracted out to external scientists, exploiting state-of-the-art approaches and knowledge. Task 3 could be conducted through CAP or other group of CBP scientists. Task 4 and 5 are less immediate but again will require substantial expertise and innovation and may be most efficiently accomplished by scientific expertise outside the immediate CBP community.

Table 7.1. Research Tasks, examples of specific subtasks, and suggested time frame for continued CFD research.

Task	Schedule
1. Effects of Sampling Design on CFD Results (a) Continue simulation work to evaluate CFD bias reduction via conditional simulation. (b) Investigate conditional simulation for interpolation methods other than kriging - this may lead to more simulation work. (c) Implement and apply interpolation with condition simulation on CBP data.	2006-2008
2. Statistical inference framework for the CFD (a) Implement and evaluate confidence interval procedures. (b) Conduct confidence interval coverage experiments. (c) Investigate confidence interval methods for non-kriging interpolation methods. (c) Implement and evaluate confidence interval procedures.	2006-2008
3. Choice of Interpolation Method (a) continue to investigate other more nonparametric interpolation methods (e.g. loess and splines). (b) implement a file system and software utilizing the “best” interpolation for CBP data. (b) compare interpolations and CFD's based on IDW and “best” method.	2006-2008
4. Three-Dimensional Interpolation (a) Implement 2-D kriging in layers to compare to current approach of 2-D IDW in layers. (b) Conduct studies of 3-D anisotrophy in CBP data. (c) Investigate software for full 3-D interpolation. Examples of options include: custom IDW software, custom kriging software using GMS routines, custom kriging software using the R-package, or some other off the shelf product.	2007-2009
5. High Density Temporal Data (a) Develop methods to use these data to improve temporal aspect of CFD in current implementation. (b) Investigate feasibility of 4-dimensional interpolation.	2008-2010

REFERENCES

Christensen OF, Diggle PJ, Ribeiro PJ. 2001. Analysing positive-valued spatial data: the transformed Gaussian model. In: Monestiez P, Allard D, and Froidevaux, editors. GeoENV III - Geostatistics for environmental applications. Quantitative Geology and Geostatistics. Dordrecht (Netherlands): Kluwer Academic Publishers. 11:287-298.

Christman MC. 2006. The characterization and incorporation of uncertainty in fisheries management, In Fisheries Ecosystem Planning for Chesapeake Bay. Bethesda (MD): American Fisheries Society. (In press).

Cressie N. 1989. The Many Faces of Spatial Prediction. Mathematical Geology 1:163-176.

Cressie N. 1991. Statistics for Spatial Data. New York: Wiley. 928 p.

Curriero FC. 2006. On the Use of Non-Euclidean Distance Measures in Geostatistics. Mathematical Geology (in press).

Daubert v. Merrell Dow Pharmaceuticals. Inc. 1993. 509 U. S. Supreme Court. 579.

Deutsch CV. 1984. Kriging with Strings of Data. Mathematical Geology. 26:623-638.

Diggle PJ, Tawn JA, Moyeed RA. 1998. Model Based Geostatistics (with Discussion). Applied Statistics 47:299-350.

Diggle PJ, Ribeiro PJ. 2006. Model-based Geostatistics. New York: Springer. 230 p.

Dille JA. 2003. How good is your weed map? A comparison of spatial interpolators. Weed Science 51:44-55.

Gotway, CA. 1994. The Use of Conditional Simulation in Nuclear Waste-Site Performance Assessment. Technometrics 36:2:129-141.

Hastie TJ, Tibshirani RJ. 1990. Generalized Additive Models. New York: Chapman and Hall p335.

Jensen OP, Christman MC, Miller TJ. 2006. Landscape-based geostatistics: A case study of the distribution of blue crab in Chesapeake Bay. Environmetrics 17:605-621.

Journel A. 1974. Geostatistics for conditional simulation of ore bodies. Economic Geology 69:673-687.

Kitanidis PK. 1997. Introduction to Geostatistics: Applications in Hydrogeology. New York: Cambridge University Press. 271 p.

Kravchenko AN. 2003. Influence of Spatial Structure on Accuracy of Interpolation Methods. Soil Sci. Soc. Am. J. 67:1564-1571.

Kutner MH, Nachtsheim CJ, Neter J, Li W. 2004. Applied Linear Statistical Models, 5th edition. Boston: McGraw-Hill. ??? p

Laslett GM. 1994. Kriging and splines: an empirical comparison of their predictive performance in some applications. *JASA* 89:391-400.

Lloyd CD. 2005. Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. Journal of Hydrology 308:128-150.

Ouyang Y, Zhang JE, Ou LT. 2006. Temporal and spatial distributions of sediment total organic carbon in an estuary river. J. Environmental Quality. 35:93-100

Reinstorf F,Binder M, Schirmer M, Grimm-Strele J, Walther W. 2005. Comparative assessment of regionalization methods of monitored atmospheric deposition loads. Atmospheric Environment 39:3661-3674.

Ribeiro PJ, Diggle PJ. 2001. geoR: A package for geostatistical analysis. R News 1:2:15-18.

Schabenberger O, Gotway CA. 2005. Statistical Methods for Spatial Data Analysis. Boca Raton, FL: Chapman and Hall/CRC Press. 512 p.

Spokas K, Graff C, Morcet M, Aran C. 2003. Implications of the spatial variability of landfill emission rates on geospatial analyses. Waste Management 23:599-607.

Sullivan PJ, Acheson J, Angermeier PL, Faast T, Flemma J, Jones CM, Knudsen EE, Minello TJ, Secor DH, Wunderlich R, Zanatell BA. 2006. Defining and implementing best available science for fisheries and environmental science, policy and management. Bethesda, Md: American Fisheries Society and Port Republic, Md: Estuarine Research Federation. Port Republic, Maryland. (available: www.fisheries.org/AFSmontana/AFS.ERF.BestScience.pdf)

Tomczak, M. 1998. Spatial interpolation and its uncertainty using automated anisotropic inverse distance weighting (IDW) – cross-validation/jackknife approach. J. Geog. Infor. and Decision Analysis 2:18-30.

[EPA] Environmental Protection Agency (US). 2003. Technical support documentation for identification of Chesapeake Bay designated uses and attainability. Annapolis (MD): EPA. 177 p. EPA 903-R-03-002. (Available: http://www.chesapeakebay.net/search/pubs.htm)

Valley RD,Drake MT,Anderson CS. 2005. Evaluation of alternative interpolation techniques for the mapping of remotely-sensed submersed vegetation abundance. Aquatic Botany 81:13-25.

Ver Hoef JM, Peterson, E, and Theobald, D, 2007, Spatial Statistical Models that Use Flow and Stream Distance, Environmental and Ecological Statistics (in press).

Wahba, G. 1990. Spline Models for Observational Data. Philadelphia (PA): Society for Industrial and Applied Mathematics. 169 p.

Wang XJ, Liu RM. 2005. Spatial analysis and eutrophication assessment for chlorophyll a in Taihu Lake. *Environmental Monitoring and Assessment* 101:167-174.

Zimmerman D., Pavlik C, Ruggles A, Armstrong MP. 1999. An experimental comparison of ordinary and universal kriging and inverse distance weighting. Mathematical Geol. 31:375-390.