Assessment Methodology for Determining Wadeable Stream Impairment Due to Excess Nitrogen and Phosphorus Levels

DECEMBER 2011

Prepared by:
Water Quality Planning Bureau, Water Quality Standards Section
Montana Department of Environmental Quality
Water Quality Planning Bureau
1520 E. Sixth Avenue
P.O. Box 200901
Helena, MT 59620-0901

WQPBMASTR-01

Prepared by:
Water Quality Planning Bureau
Michael Suplee, Ph.D., Water Quality Standards Section
Rosie Sada de Suplee, Monitoring and Assessment Section

Suggested citation: Suplee, M.W., and R. Sada de Suplee, 2011 Assessment Methodology for Determining Wadeable Stream Impairment Due to Excess Nitrogen and Phosphorus Levels. Helena, MT: Montana Dept. of Environmental Quality.

TABLE OF CONTENTS

Nitrogen and Phosphorus Impairment Assessment: Method Summary 1
1.0 Introduction 5
1.1 Scope of the Assessment Methodology 5
2.0 Recommendations for Isolating a Stream Reach for Assessment (the Sample Frame) 6
3.0 Assembling the Nutrient Assessment Data into a Decision Framework 7
3.1 Overview of Useful Parameters for Carrying out Nutrient Impairment Assessments 8
3.2 Details on the Use of Nitrogen and Phosphorus Concentration Data, and Other Measured Parameters, to Support Nutrient-Impairment Assessments 12
3.2.1. Nitrogen and Phosphorus Data 13
3.2.1.1 Minimum Sample Size for Nitrogen and Phosphorus Data 14
3.2.2 Minimum Sample Sizes for Other Parameters 15
3.2.3 Determine the Nutrient Most Likely to be Harming Use(s) 16
3.2.4 Examples of Nutrient-impairment Decisions 17
3.2.5 Overwhelming Evidence of Nutrient Impairment-All Regions 18
4.0 Nutrient Impairment Assessment Methodologies: Wadeable Streams in the Mountain and Transitional Region 19
4.1. Assessment of Benthic Algal Growth 20
4.2 Assessment Using Biometrics 20
5.0 Nutrient Impairment Assessment Methodologies: Wadeable Streams in the Plains Region 21
5.1 Assessment Using Biometrics 22
5.2 Assessment Using the Difference Between the Daily Maximum Dissolved Oxygen Concentration and the Daily Minimum Dissolved Oxygen Concentration (Delta) 22
5.2.1 Deploying a Continuous DO Monitoring Device in Wadeable Plains Streams 23
5.2.2 Instantaneous DO Monitoring in Wadeable Plains Streams 23
5.2.3 BOD_5 23
5.2.4 Algal and Macrophyte Indicators of Nutrient Enrichment in Wadeable Plains Streams 23
6.0 Acknowledgements 24
Appendix A. Statistical Considerations A-1
A.1.0 Introduction A-1
A.2.0 Sample Frame, Population, and Sampling Units A-1
A.3.0 Determining Sample Independence A-3
A.4.0 Selection of Inferential Statistical Tests, Confidence Levels, and Associated Decision Rules A-6
A.4.1 Rationale for Using Two Inferential Statistical Tests to Help Determine Compliance with Nutrient Standards A-6

A.4.2 Form of the Null Hypothesis, Alpha, Beta, Effect Size, and Critical Exceedance Rate A-7
A.4.2.1 Form of the Null Hypothesis for the Statistical Tests A-7
A.4.2.2 Alpha, Beta, Effect Size A-7
A.4.2.3 Critical Exceedance Rate A-8
A.5.0 Minimum Number of Nutrient Samples A-14
Appendix B. Details on Assessment Methodologies for Wadeable Mountain and Transitional Streams B-1
B.1.0 Assessment of Benthic Algal Growth B-1
B.1.1 Benthic Algal Chl *a* Levels and the Recreation Use B-1
B.1.2 Benthic Algal Chl *a* Levels Relative to Late-season Dissolved Oxygen Problems, and Potential Impacts to Fish and Associated Aquatic Life B-2
B.1.3 Benthic Algal AFDW Levels and Harm-to-use Thresholds B-5
B.1.4 Some Additional Considerations Regarding Benthic Algae Sampling B-5
B.2 Assessment Using Biometrics B-6
B.2.1 Biometrics Based on Diatom Algae B-6
B.2.2 Biometrics Based on Macroinvertebrates B-7
B.2.2.1 Hilsenhoff Biotic Index B-7
B.2.2.2 Interpreting the Macroinvertebrate Metric Correlations to Nutrients B-9
B.2.2.3 Patterns Observed Between HBI and Change-points, and Previously Used DEQ Thresholds B-10
Appendix C. Details on Assessment Methodologies for Wadeable Plains Streams C-1
C.1.0 Assessment Using Diatom Algae Biometrics C-1
C.2.0 Assessment Using the Difference Between the Daily Maximum Dissolved Oxygen Concentration and the Daily Minimum Dissolved Oxygen Concentration (Delta) C-2
C.2.1 Instantaneous DO Monitoring in Wadeable Plains Streams C-4
C.3.0 Biochemical Oxygen Demand C-5
Appendix D - References D-1

LIST OF TABLES

Table S-1. Non-wadeable river segments within the state of Montana 1
Table S-2. Start and Ending Dates for Three Seasons (Winter, Runoff and Growing), by Level III Ecoregion 2
Table 1-1. Non-wadeable river segments within the state of Montana 6
Table 3-1. Parameters in Streams that are Considered Useful in Assessing Nutrient Enrichment 8
Table 3-2. Three combinations of results, and the conclusions that can be drawn from them, using the parameters listed in Table 3-1. 11
Table 3-3. Start and Ending Dates for Three Seasons (Winter, Runoff and Growing), by Level III Ecoregion 13
Table 3-4. Simultaneous Review of Multiple Nutrients and Effect Variables 16
Table 4-1. Ecoregions (levels III and IV) in which Assessment Methods in this Section Best Apply 19
Table 5-1. Ecoregions (level III) in which Assessment Methods in this Section Best Apply 21
Table A3-1. Durbin-Watson Values for Time-series Collected Nutrient Samples at Selected Sites A-4
Table A4-1 Sites on the Clark Fork River (CFR) Not Exceeding the Maximum Benthic Algae Standard (Growing Season, 1998-2006) A-10
Table A4-2. Sites on the Clark Fork River (CFR) Consistently Exceeding the Maximum Benthic Algae Standard (Growing Season, 1998-2006). A-11
Table B1-1. Benthic Algae Density Measured at the High Dose Study Site, 2010 ..... B-4
Table B2-1. Nutrient Concentrations on the LOWESS Regression Line Corresponding to Specified Thresholds. B-10
Table C2-1. Long-term DO Monitoring Sites in Plains Streams C-3
Table C2-2. Statistics Associated with DO Delta Threshold C-4

LIST OF FIGURES

Figure 1-1. Map Showing Different Regions in which Different Assessment Methodologies Apply. 6
Figure 3-1. Flow-path for decision making using data parameters in Table 3-1. 10
Figure 3-2. Photographs of heavy, bank-to-bank and longitudinally continuous Cladophora growth 18
Figure 3-3. Massive *Cladophora* growth in the Clark Fork River, MT, 1984. 19
Figure A2-1. Four different stream reaches (shown by different colors), each representing 1 sampling frame (ADB stream segment) A-3
Figure A4-1. Least squares regression for TN exceedance rate vs. Max Summer Chl *a* (upper panel) and TP exceedance rate vs. Max Summer Chl *a* (lower panel), for ten Clark Fork River monitoring sites (1998-2009). A-14
Figure B1-1 Averages (Gray Dots) and 90% Confidence Bands for the Three Photos Bracketing the Acceptable-unacceptable Threshold, per Suplee et al. (2009) B-2
Figure B1-2. Temporal Changes in Dissolved Oxygen (DO) Levels in the High Dose Study Reach as Measured by Deployed YSI Instrument, 2010. ..... B-4
Figure B2-1. HBI metric vs. TP. B-8
Figure B2-2. HBI Macroinvertebrate Metric vs. Stream TN Concentration B-9

ACRONYMS

Acronym	Definition
ADB	Assessment database
AFDW	Ash Free Dry Weight
ARM	Administrative Rules of Montana
BACI	Before-After-Control-Impact
BACIP	Before After Control Impact Paired
BOD	Biochemical Oxygen Demand
BPJ	Best Professional Judgment
CFR	Clark Fork River
CV	Coefficient of Variation
DEQ	Department of Environmental Quality (Montana)
DO	Dissolved Oxygen
DPHHS	Department of Health and Human Services
EPA	Environmental Protection Agency (US)
HBI	Hilsenhoff Biotic Index
LOWESS	locally-weighted regression line
QAPP	Quality Assurance Project Plan
SAP	Sampling and Analysis Plan
SOD	sediment oxygen demand
SOP	Standard Operating Procedures
SRP	Soluble Reactive Phosphate
TKN	Total Kjeldahl Nitrogen
TMDL	Total Maximum Daily Load
TN	Total Nitrogen
TP	Total Phosphorus
WQPB	Water Quality Planning Bureau (DEQ)

NITROGEN AND PHOSPHORUS IMPAIRMENT ASSESSMENT: METHOD SUMMARY

The following Method Summary should provide sufficient detail for an assessor to undertake an assessment of nitrogen and phosphorus impacts in a wadeable stream. *You will probably still need to refer to details provided later in the document.* Large rivers are not addressed by this methodology; a list of large rivers to which these methods do not apply is shown below in **Table S-1**.

Table S-1. Non-wadeable river segments within the state of Montana

River Name	Segment Description
Big Horn River	Yellowtail Dam to mouth
Clark Fork River	Bitterroot River to state-line
Flathead River	Origin to mouth
Kootenai River	Libby Dam to state-line
Madison River	Ennis Lake to mouth
Missouri River	Origin to state-line
South Fork Flathead River	Hungry Horse Dam to mouth
Yellowstone River	State-line to state-line

Part 1: Defining the assessment reach

1. Compliance determinations described in this document are carried out on an assessment reach. Here we define an assessment reach as: *a wadeable stream segment listed in the Assessment Data Base (ADB;(Montana Department of Environmental Quality, Planning, Prevention and Assistance Division, 2009) and updates), or a sub-segment of an ADB stream segment.* A sampling unit within an assessment reach is defined *as: a sample collected from the assessment reach that is largely independent of other samples collected within the assessment reach and collected during the time when the numeric nutrient criteria apply.* Please consider the following:

    a. The aggregate of samples collected from an assessment reach should provide good overall representation of the assessment reach. Individual sites within the assessment reach that have known or suspected pollution problems should be sampled equitably along with sites where pollution problems are not suspected or are minimal or less pronounced. Do not just target the hotspots.

    b. Given the guidelines in 1a above, the assessor will have to judge if further stratification of the stream reach (i.e., create two or more sub-reaches) is warranted. If, for example, a relatively un-impacted upstream reach of an assessment reach can be isolated and its condition is substantially different from other downstream parts of the assessment reach, sub-segmenting may likely be justified. As a rule of thumb, it is better to lump than split reaches to avoid excessive sub-segmentation of streams and the consequential administrative and sampling requirements.

    c. Each sub-reach will have the same general data requirements (dataset minimums, tests, etc.) as the parent assessment reach would have had if it hadn't been divided.

    d. Samples should be collected when the criteria apply, during the ecoregion-specific Growing Season (**Table S-2**). However, a ten day window (plus/minus) on the Growing Season start

and end dates is acceptable in order to accommodate year-specific conditions (e.g., an early-ending spring runoff). Samples collected outside the Growing Season may be useful for other purposes (e.g., isolating load sources), but should not to be used for compliance determination for the Growing Season.

Table S-2. Start and Ending Dates for Three Seasons (Winter, Runoff and Growing), by Level III Ecoregion

Ecoregion Name	Start of Winter	End of Winter	Start of Runoff	End of Runoff	Start of Growing Season	End of Growing Season
Canadian Rockies	Oct.1	April 14	April 15	June 30	July 1	Sept. 30
Northern Rockies	Oct.1	March 31	April 1	June 30	July 1	Sept. 30
Idaho Batholith	Oct.1	April 14	April 15	June 30	July 1	Sept. 30
Middle Rockies	Oct.1	April 14	April 15	June 30	July 1	Sept. 30
Northwestern Glaciated Plains	Oct.1	March 14	March 15	June 15	June 16	Sept. 30
Northwestern Great Plains	Oct.1	Feb. 29	March 1	June 30	July 1	Sept. 30
Wyoming Basin	Oct.1	April 14	April 15	June 30	July 1	Sept. 30

2. Samples from within an assessment reach may generally be considered independent of other samples from the assessment reach if they meet or if you do the following:
   - Sites (or very short reaches functionally equivalent to sites) should be located at least 1 stream mile apart.
   - Sites may be placed < 1 mile apart on an assessment reach **if** there is a flowing tributary confluencing with the reach between the two sites.
   - Along an assessment reach, try to sample sites moving from downstream to upstream to avoid potentially re-sampling the same stream water.
   - Land use changes and land form changes should be considered and can be used to help define (1) breaks between assessment reaches and/or (2) additional sampling sites within an assessment reach
   - Samples collected at the same site should be collected about 30 days apart. <u>This does not apply to long-term or instantaneous measurement of dissolved oxygen.</u>

Part 2: Assessment Methodology

The following are recommended to determine nitrogen and phosphorus impacts in wadeable streams.

<u>For Mountainous and Transitional[1] Streams:</u> Assessment is carried out as a level I, level II process. If the level I results are inconclusive, move to the 2nd level[2].

<u>Level I:</u>

a. Collect, within the assessment reach, benthic algal Chl *a* and AFDW (Ash Free Dry Weight) from one or more sites[3] (following DEQ SOPs, including approved low-chlorophyll visual

[1] See **Table 4-1** later in this document for the list of ecoregions (levels III and IV) where this methodology applies.
[2] Nothing precludes the assessor from collecting, in a single sampling season, all data needed to carry out a level II assessment. Cases may arise (e.g.., land access issues) that may make this approach preferable.
[3] Treat each Chl *a* sampling event as an independent evaluation of use support Do not average together results from different sites within the assessment reach If Chl *a* is measured more than once at the same site, treat each sampling event as unique (NOT as temporal repeat measures)

estimation methods) for a minimum of three sampling events. A minimum of twelve or thirteen[4] independent nutrient samples should be collected within the same assessment reach. Use of diatom samples at level I are optional, but if the data exist (n ≥ 2 samples), they must be used in the assessment. Disperse sampling effort across sites as much as possible. The nutrient data are evaluated using the "Exact Binomial Test" and the "One-Sample Student's t-test for the Mean" which are housed in one of two Excel spreadsheets. If the assessment reach is a new, un-listed segment, use "MT-NoncomplianceTool.xls". If the assessment reach is already listed for a nutrient, use MT-ComplianceTool.xls". However, if a stream is currently listed for nitrogen but not for phosphorus, use the "MT-NoncomplianceTool.xls" to assess the phosphorus data.

b. In both spreadsheets, for either test, set alpha to 0.25 (25%) and the critical exceedance rate (p) to 0.2 (20%) in cells B5, B6. In the Binomial test the effect size (p2; gray zone) should be 0.15 (15%) and is set as a function of the exceedance rate. So, in "MT-NoncomplianceTool.xls" this means p2 should be set to 0.35 (cell B7), and in "MT-ComplianceTool.xls" p2 should be set to 0.05 (cell B7). If in the future DEQ decides that a lower exceedance rate (e.g., 10%) is needed, the gray zone will need to be adjusted accordingly.

c. Compliance with the nutrient criteria is determined via decision rules, which consider the Chl *a* and AFDW averages calculated for each sampling event, the results from the two nutrient statistical tests, and diatom metric results (if available). Go to the first tab of the Excel spreadsheet named "NtrntAssessFramework.xlsx". If the result is clear (assessment reach <u>is</u> or <u>is not</u> nutrient impaired), you are finished. If not, follow the instructions in the spreadsheet for level II assessment.

d. Most often, you will be assessing both N and P in an assessment reach. Consider the N and P results side-by-side; does is appear that one nutrient or the other is giving a clear signal (e.g., Binomial and T-test are both FAIL for Total Phosphorus (TP), but both PASS for Total Nitrogen (TN)? In this case, the best nutrient to list would be TP. Mixed results for both nutrients often will require a move to a level II assessment, and may lead to listing both N and P.

<u>Level II</u>:

a. Moving to level II often (not always) involves additional data collection, including more nutrient samples and benthic algal Chl *a*/AFDW samples. Level II data include both diatom and macroinvertebrate samples (at least two sampling events for each). **The exception to this is the Middle Rockies ecoregion, for which there are no validated diatom increaser metrics. In this ecoregion collect at least three macroinvertebrate samples.** As for level I, each sampling event for diatoms should be considered on its own merits (do not average results across sites, or across time at a site). In contrast, macroinvertebrate samples collected across time <u>at a site</u> should be averaged together; however, keep and assess data from <u>different</u> sites separate. When your dataset is ready, first pass data again through the

[4] Twelve independent nutrient samples for new, unlisted streams, and thirteen for streams already listed for nutrients on the 303(d) list. A nutrient sample is a type of nutrient (e.g., TP or TN) Sample minimums apply to *each* nutrient type. Smaller sample sizes may be justified; see **Section 3.2.2.1**, this document

Level I process using NtrntAssessFramework.xlsx. If results are clear, you are finished; if not, go to the 2nd tab.

b. Some data combinations at level II (2nd tab of the spreadsheet) will still lead to an unclear result. If this occurs, consult with your manager about how to proceed.

For Warm Water Plains Streams[5]*:* Assessment is carried out as a level I, level II process. If the level I results are inconclusive, move to the 2nd level[2].

Level I:

a. Determine, for at least 3 sampling events, the dissolved oxygen (DO) delta (i.e., the daily DO maximum minus the daily DO minimum). The daily minimum can be measured pre-dawn to 8:00 am, while the daily maximum usually occurs between 2:30 pm to 5:00 pm. Alternatively, collect a long-term DO dataset by deploying a YSI 6600 (or similar instrument) in at least one site; measure DO for at least 1 full day, with a 15-min time step. Even if you collect DO data with a deployed instrument, you still need a total of three sampling events (three days). **However, DO delta values need not be collected 30 days apart.** Also, collect within the assessment reach at least two diatom samples and a minimum of twelve or thirteen[4] nutrient samples. Disperse sampling effort across sites as much as possible. The nutrient data are evaluated using the tests "Exact Binomial Test" and the "One-Sample Student's t-test for the Mean" found in one of two Excel spreadsheets. If the assessment reach is a new, un-listed segment, use "MT'NoncomplianceTool.xls". If the assessment reach is already listed for a nutrient, use MT-ComplianceTool.xls". However, if a stream is currently listed for nitrogen but not for phosphorus, use the "MT-NoncomplianceTool.xls" to assess the phosphorus data.

b. See 1b above (in the Mountainous and transitional streams section) for instructions on setting test conditions in each Excel spreadsheet.

c. Compliance with the nutrient criteria is determined via decision rules, which consider together the results from the diatom metrics, the DO delta values, and the two statistical tests for nutrients. Go to the 3rd tab (plains level I) of the Excel spreadsheet "NtrntAssessFramework.xlsx". Long-term DO datasets require special consideration; see **Section 3.2.4**, scenario 2, and **Section 5.0** for details. If the result is clear (assessment reach is or is not nutrient impaired), you are finished. If not, follow the instructions in the spreadsheet for level II assessment.

d. Most often, you will be assessing both N and P in an assessment reach. Consider the N and P results side-by-side; does is appear that one nutrient or the other is giving a clear signal (e.g., Binomial and T-test are both FAIL for TN, but both PASS for TP)? In this case, the best nutrient to list would be TN. Mixed results for both nutrients often will require a move to a level II assessment, and may lead to listing both N and P.

[5] See **Table 5-1** later in this document for the list of ecoregions where this methodology applies.

Level II:

a. A level II assessment will often require additional data collection, including more nutrient and DO data, and (in some cases) biochemical oxygen demand (BOD_5) data. As for level I, each DO delta value should be considered on its own merits (do not average results across sites or across time), as is also the case for BOD_5 samples. When the data are ready, first pass the now-larger dataset back through the Level I assessment process. If results are clear, you are finished; if not, go to the appropriate scenario in the 4th tab (plains level II).

b. Some level II data combinations still lead to an unclear result. If this occurs, consult with your manager about how to proceed.

1.0 INTRODUCTION

The purpose of this document is to describe a framework for making decisions. Specifically, it defines a process by which one can determine if a wadeable stream is or is not impaired by nitrogen and phosphorus pollution (i.e., excess nutrients). The document covers a number of subjects including how to determine an appropriate sampling frame, which parameters are most useful for assessing nitrogen and phosphorus problems, how many samples are needed, how data are to be treated statistically, and how disparate data types are to be assembled in a final decision matrix.

In this document we have attempted to organize the information in a manner such that the users can locate what they needs quickly, and then read further for details only if they want to. The “why” discussions (i.e., why did we select a particular assessment parameter, why did we pick the impact threshold, etc.) are found in the appendices. We did this because we know that stream assessments are time-consuming undertakings, and therefore users will want to access the critical information easily.

1.1 SCOPE OF THE ASSESSMENT METHODOLOGY

Different assessment methods are recommended for different regions of the state (**Figure 1-1**). The assessment parameters that have been recommended for each region are the ones we believe are the most accurate and sensitive for determining nitrogen and phosphorus impacts for wadeable streams in those areas. For example, we recommend measuring dissolved oxygen in eastern Montana plains streams, but we have not recommended this approach for western Montana salmonid streams. This should not be construed to mean that DO concentrations in salmonid streams are never affected by nitrogen and phosphorus pollution, or that in any way this recommendation overrides existing DO standards for those state waters. Rather, we believe that in western Montana salmonid streams there are assessment tools other than DO that are more sensitive and will more readily detect nitrogen and phosphorous problems.

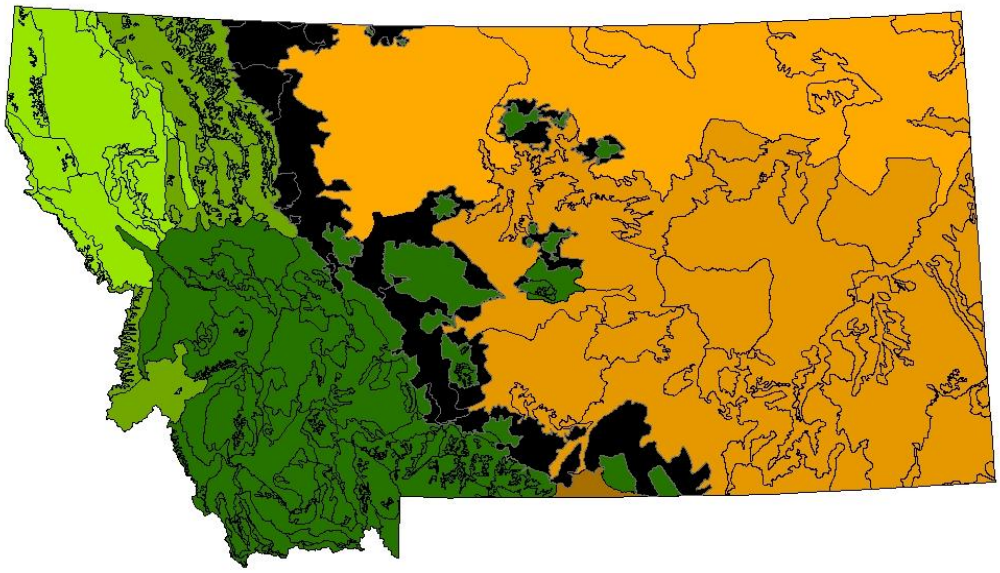


Figure 1-1. Map Showing Different Regions in which Different Assessment Methodologies Apply. The areas shown in shades of green and black comprise the mountain and transitional streams region. The areas shown in shades of brown comprise the eastern Montana plains region.

As mentioned above, methods in this document apply to wadeable streams. A DEQ workgroup spent considerable time working to define the break between wadeable streams and rivers and non-wadeable rivers, the results of which are presented in Flynn and Suplee (2010). The waterbodies that are not considered wadeable are provided below in **Table 1-1.**

Table 1-1. Non-wadeable river segments within the state of Montana

River Name	Segment Description
Big Horn River	Yellowtail Dam to mouth
Clark Fork River	Bitterroot River to state-line
Flathead River	Origin to mouth
Kootenai River	Libby Dam to state-line
Madison River	Ennis Lake to mouth
Missouri River	Origin to state-line
South Fork Flathead River	Hungry Horse Dam to mouth
Yellowstone River	State-line to state-line

2.0 Recommendations for Isolating a Stream Reach for Assessment (the Sample Frame)

Identifying and isolating an appropriate stream reach (sample frame) is the first required task. The following definitions are presented:

- **Sample Frame**: A wadeable[6] stream segment listed in the Assessment Data Base (ADB) (DEQ 2009, and updates) OR a sub-segment of an ADB stream segment. A segment such as this is referred to in this document as an "assessment reach".
- **Population**: All the water flowing through the assessment reach during the time period when the numeric nutrient criteria apply, and the surface area of the stream bottom over which the water flows.
- **Sampling Unit**: A sample collected from the assessment reach that is largely independent of other samples collected within the assessment reach and collected during the time when the numeric nutrient criteria apply.

A sampling frame must be representative of the population and, in stream assessment, this demands good judgment in the particular subject matter being studied. **Sections A.1** and **A.2** of **Appendix A** of this document, which is an updated and shortened version of an earlier Appendix H (Varghese and Cleland, 2008), contain a discussion on approaches to identifying assessment reaches.

The key idea presented in **Appendix A** is that each assessment reach that is assessed should be sufficiently homogenous that data collected from sites along the reach can be considered to represent the entire reach. To determine compliance with numeric nutrient criteria using statistical methods, it is important that (1) pollution sources generally be evenly dispersed along the reach, and (2) each sample is independent of the others. Following up on this idea, if an assessment reach appears to need further subdivision (e.g., into a reach above and a reach below a pollution point source), then each new assessment reach should generally be sampled with the same intensity (i.e., minimum sample size) as the parent reach would have been if it had not been subdivided. This will assure that the statistical rigor associated with specified sample-size minima (discussed below) is maintained. At the same time, as a general rule, it is better to lump than split to avoid unnecessary sampling and administrative work.

The need to create reasonably uniform assessment reaches is inherently in conflict with the need to "lump", the purpose of which is to keep stream reaches from being excessively subdivided (and all the additional work that entails). Judgment is needed on the part of the assessor to balance these two opposing factors and come up with an optimal sampling strategy for any given stream.

This process should be compatible with a randomized study of stream reaches as well as targeted, risk-assessment based approaches; again, the key point is that each assessment reach is sufficiently defined.

3.0 Assembling the Nutrient Assessment Data into a Decision Framework

Section 2.0 above discussed approaches used to identify appropriate assessment reaches. This section discusses how data that will have been collected from the assessment reach are to be assembled into a decision-making framework. The parameters and methods apply to wadeable streams. Non-wadeable waterbodies were listed back in **Table 1-1.**

[6] Wadeable streams are perennial as well as intermittent (ARM 17.30.602 [15]) streams in which large portions of the channel are wadeable during baseflow conditions. For the list of waterbody segments not considered wadeable (i.e., the large rivers), see **Table 1-1** above Derivation of the **Table 1-1** list is found in Flynn and Suplee (2010).

3.1 Overview of Useful Parameters for Carrying out Nutrient Impairment Assessments

Among the vast array of parameters that can be measured in a stream, we narrowed the list to those we believe are the best, readily-measured indicators of stream nitrogen and phosphorus enrichment (**Table 3-1**). Many of these parameters are discussed in Suplee *et al.* (2008), and are also discussed in detail in the appendices of this document.

Table 3-1. Parameters in Streams that are Considered Useful in Assessing Nutrient Enrichment

Parameter	How collected	Linkage to nutrient enrichment	Primary or secondary indicator*
Total nitrogen	Water sample	Total instream concentrations are indicative of the level of nutrients that are ultimately biologically available for autotrophic or heterotrophic uptake.	Primary
Total phosphorus	Water sample	Total instream concentrations are indicative of the level of nutrients that are ultimately biologically available for autotrophic or heterotrophic uptake.	Primary
Benthic algal biomass	Benthic samplings of stream bottom	Nutrients stimulate benthic algal growth in wadeable streams Benthic algal growth can develop to nuisance levels; nuisance algae level is known. Excess algal growth affects on DO have been documented.	Primary
Dissolved oxygen *delta* (daily max value minus the daily min value)	Instantaneous: By hand-held instrument, at dawn and in the late pm. Continuous monitoring: by deployed instrument	Nutrient enrichment stimulates autotrophic primary productivity and heterotrophic decomposition of organic material. Both of these in turn affect dissolved oxygen patterns in streams.	Primary
Diatom biometric (nutrient increaser taxa metric)	Benthic sampling of stream bottom	As primary producers, diatoms can be directly stimulated by increased availability of N and P. Diatom population structure has been found to vary in predictable ways with increasing nutrient enrichment.	Primary and Secondary
Macroinvertebrate biometric (Hilsenhoff Biotic Index, or HBI)	Kicknet sampling of stream bottom	A large number of macroinvertebrate taxa have been assigned a numeric value which represents each organism's tolerance to low dissolved oxygen/organic pollution. Resulting metric (HBI) found to significantly correlate to total nutrient concentrations in Montana streams.	Secondary
Biochemical oxygen demand (BOD_5)	Water sample; must be at laboratory within 48 hrs	High BOD can indicate presence of large quantities of dissolved and suspended organic matter, whose decomposition can produce a large DO demand. Can help determine if DO sags are caused by high primary productivity, high BOD, or both.	Secondary

Table 3-1. Parameters in Streams that are Considered Useful in Assessing Nutrient Enrichment

Parameter	How collected	Linkage to nutrient enrichment	Primary or secondary indicator*
Stream macrophyte species	By hand; field identification	Observed shift in dominance to a single macrophyte species in highly enriched prairie streams; loss of *Chara*.	Secondary

*Primary means the parameters is considered to be a very good indicator of nutrient enrichment. Secondary mean the parameter is considered a good indicator of nutrient enrichment, or helpful in identifying other factors affecting DO (e.g., BOD).

Note that **Table 3-1** contains physical and biological measurements. We support the long-held view in the WQPB that stream assessment is best carried out by looking at both data types together. The famous water-pollution biologist H.B Hynes said it best: “When the chemist and the biologist both work on the assessment of pollution they can discover much more together than either can alone” (Hynes, 1966).

The parameters in **Table 3-1** need to be arranged in a decision making framework in order to produce consistent decision outcomes (i.e., stream *is* impaired by nutrients, stream *is not* impaired by nutrients). **Figure 3-1** below outlines the process we recommend for this purpose.

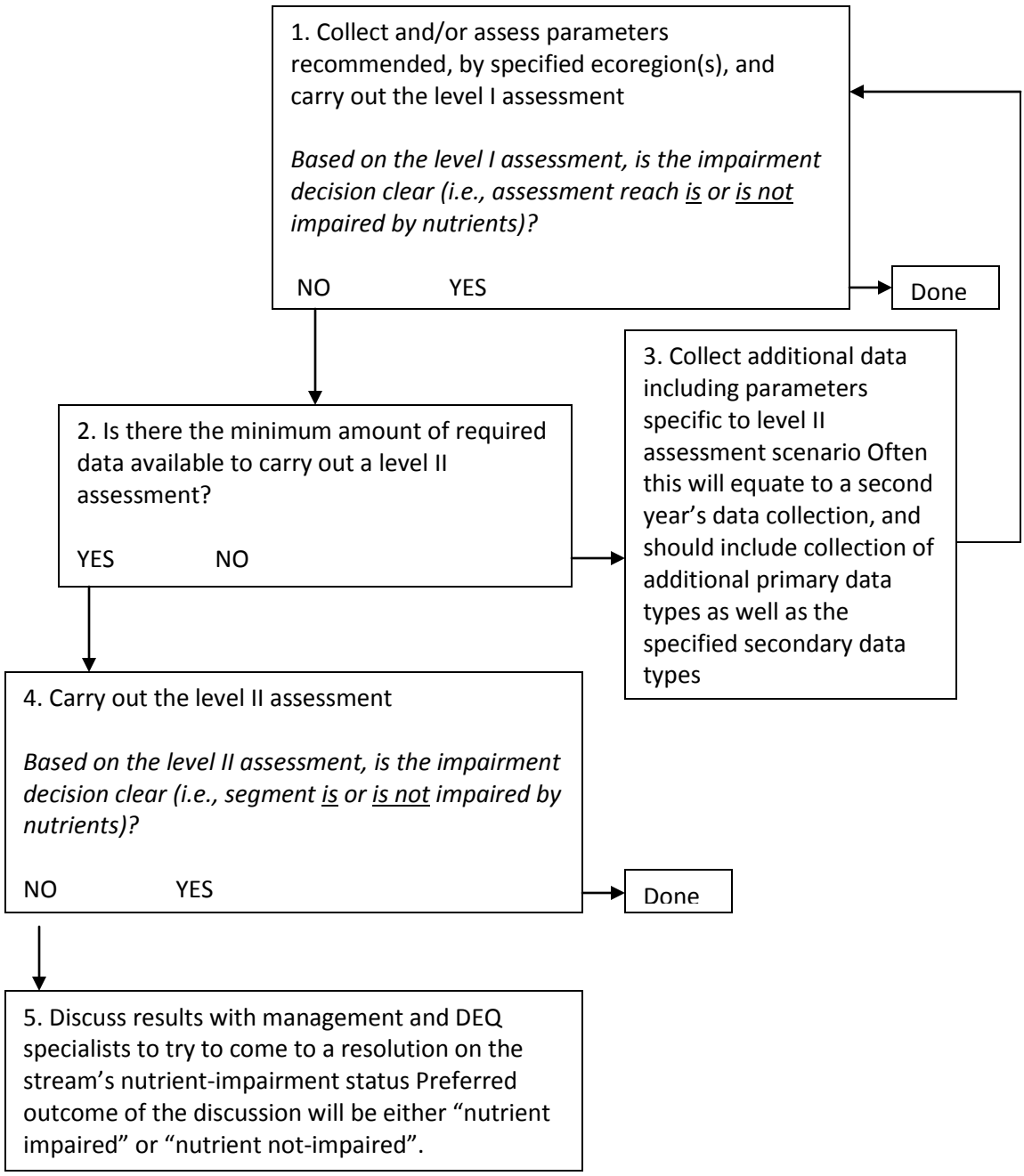


Figure 3-1. Flow-path for decision making using data parameters in Table 3-1.

If the level I assessment leads to an unclear decision, the assessor should then use the data (primary and secondary, if data sufficiency met) to carry out the level II assessment[7]. If a level I assessment is inconclusive and leads to a 2nd year of data collection, always pass the now-larger dataset back through the level I assessment matrix first. It may result that that the conclusion is now clear, without having to go to level II. **NOTE:** Nothing in the approach shown in **Figure 3-1** precludes an assessor from collecting,

[7] The approach shown in **Figure3-1** closely parallels the decision framework of EPA's CALM guidance (U.S. Environmental Protection Agency, 2002); see **Figure 3.2**, page 3-10 of that document.

in a single field season, all data needed to complete a level I and a level II assessment. Situations may arise (e.g., land access issues) where this approach is preferable.

As can be seen, one notable aspect of the approach in **Figure 3-1** is that the data we have labeled "secondary" (**Table 3-1**) are brought into the decision framework only after the primary data have lead to an unclear conclusion. In effect, secondary data are being held to the side until the primary data have been played out to their fullest. The approach attempts to keep data combination scenarios to a minimum and decision making as simple as is reasonable (Occam's razor; "plurality should not be posited without necessity")[8].

The different combinations of results that can occur have been assembled in an Excel spreadsheet (**NtrntAssessFramework.xlsx).** In the spreadsheet, the user identifies the unique combination of results from their assessment reach, and then derives a conclusion. For each combination of results, the spreadsheet provides an outcome (i.e., nutrient impaired, not-nutrient impaired, unclear), and an explanation as to what is likely going on in the stream's ecology. **Different parameter sets are used in different geographic regions, therefore the user must use the tabs for the region applicable to their stream.** Regional tabs are further subdivided to correspond to a level I or level II assessment, per the approach shown in **Figure 3-1**. As an example, three result combinations for the mountain and transitional region are given in **Table 3-2**.

Table 3-2. Three combinations of results, and the conclusions that can be drawn from them, using the parameters listed in Table 3-1.
All three examples apply to streams of the mountain and transitional region of the state, and are from a level I assessment.

Scenario	Nutrient Binomial Test	Nutrient T-test	Benthic Algae	Diatom Increaser Taxa-Probability of Impairment (OPTIONAL)*	Resulting Decision	Further Sampling?
1	PASS	PASS	≤120 mg Chl *a*/m^2 or ≤35 g AFDW/m^2	<51%	Waterbody <u>is not</u> nutrient impaired. All indications show that the stream is in compliance.	No

[8] We did this for two reasons. First, we believe that in most cases some types of data are inherently better for nutrient-enrichment assessment than others, and if the decision can be made using those data alone, the assessment will be simpler and less expensive. Second, it reduces the total number of data-combination outcomes and, in turn, the number of scenario-by-scenario conclusions about impairment that have to be made. To illustrate, for any given data type for which there is a dichotomous outcome (i.e., result is <u>above</u> or <u>below</u> some threshold), the number of possible permutations is 2 raised to the number of data types. Three data types result in (2^3) = 8 possible data-combinations, and one must consider what each unique combination of results is saying about nutrient impairment. Five data types considered together already results in 32 unique combinations, and so on If not all of the additional data are as useful as the previous, it becomes questionable whether the additional work, cost, and complexity are warranted.

Table 3-2. Three combinations of results, and the conclusions that can be drawn from them, using the parameters listed in Table 3-1.
All three examples apply to streams of the mountain and transitional region of the state, and are from a level I assessment.

Scenario	Nutrient Binomial Test	Nutrient T-test	Benthic Algae	Diatom Increaser Taxa-Probability of Impairment (OPTIONAL)*	Resulting Decision	Further Sampling?
12	PASS	FAIL	>120 mg Chl *a*/m^2 or >35 g AFDW/m^2	>51%	Waterbody is nutrient impaired. Non-compliance with the T-test suggests that pulsed nutrient loads are allowing high algae biomass to be maintained via luxury uptake. Diatoms confirm enrichment finding.	No
16	FAIL	FAIL	>120 mg Chl *a*/m^2 or >35 g AFDW/m^2	>51%	Waterbody is nutrient impaired. All indicators show that the stream is not in compliance.	No

*However, if the data minima are available for diatom metric category, they must be used in the decision framework

Subsequent sections will provide detail on which assessment parameters apply where, which statistical tools are to be applied to which parameters, etc. The important point to note here is that the combinations of results you will encounter have been accounted for in the spreadsheet tool (**NtrntAssessFramework.xlsx)**.

Returning to **Figure 3-1**, the "Discuss results with management and DEQ specialists" outcome occurs when the level II assessment has still not resulted in a clear conclusion. This resolution step was suggested by Mark Bostrom (DEQ Bureau Chief) as a way to come to a conclusion without ending up in an endless do-loop. Details of this process remain to be worked out; a potentially useful framework for carrying out the determination has been developed by EPA (Cormier, et al., 2000; Cormier and Suter, II, 2008).

3.2 Details on the Use of Nitrogen and Phosphorus Concentration Data, and Other Measured Parameters, to Support Nutrient-Impairment Assessments

As noted above, different groups of parameters best apply to particular regions of the state. The applicable list of parameters, their impact thresholds, and the delineation of the regions are provided in **Section 4.0** (mountain and transitional streams) and **Section 5.0** (plains streams). In order to maintain temporal independence to the best degree possible, samples collected sequential at a site should be collected about **30 days** after the previous sampling. (There are exceptions to this; see individual parameter list in **Section 3.2.2** below.) Spatial independence of sites within an assessment reach can generally be established by following these guidelines:

- Sites (or short reaches equivalent to sites) should be located a minimum of 1 stream mile apart.

- Sites may be placed < 1 mile apart along the assessment reach if there is a flowing tributary confluencing with the segment between the two sites.
- Try to collect water samples starting at the downstream end of the assessment reach moving upstream, to avoid re-sampling the same water.
- Land use changes and land form changes should be considered and can be used to help define additional sampling sites within an assessment reach.

See **Section A.3** in **Appendix A** for the derivation of these guidelines.

Numeric nutrient criteria apply during summer baseflow[9], also referred to as the growing season. Start and end dates for the growing season vary by ecoregion (Suplee, et al., 2008); see **Table 3-3** below. These dates should be adhered to for collection of the other parameters in **Table 3-1** as well. However, a ten day window (plus/minus) on the Growing Season start and end dates is acceptable, in order to accommodate year-specific conditions (e.g., an early-ending spring runoff). The assessor should use their best professional judgment when deciding if early or later sampling is warranted.

Table 3-3. Start and Ending Dates for Three Seasons (Winter, Runoff and Growing), by Level III Ecoregion

Ecoregion Name	Start of Winter	End of Winter	Start of Runoff	End of Runoff	Start of Growing Season	End of Growing Season
Canadian Rockies	Oct.1	April 14	April 15	June 30	July 1	Sept. 30
Northern Rockies	Oct.1	March 31	April 1	June 30	July 1	Sept. 30
Idaho Batholith	Oct.1	April 14	April 15	June 30	July 1	Sept. 30
Middle Rockies	Oct.1	April 14	April 15	June 30	July 1	Sept. 30
Northwestern Glaciated Plains	Oct.1	March 14	March 15	June 15	June 16	Sept. 30
Northwestern Great Plains	Oct.1	Feb. 29	March 1	June 30	July 1	Sept. 30
Wyoming Basin	Oct.1	April 14	April 15	June 30	July 1	Sept. 30

3.2.1. Nitrogen and Phosphorus Data

The nitrogen and phosphorus criteria are <u>not</u> presented in this document. Readers should refer to Suplee et al. (2008) and its addendums to locate the nutrient concentration values. DEQ anticipates that the nutrient criteria will be adopted by the Board of Environmental Review. After adoption they will be referred to as "base numeric nutrient criteria", and will be housed in a new DEQ circular (DEQ-12). Please use the most updated versions of the criteria in all assessments. Check with Standards on status.

Nutrient (TN, TP) concentration data from an assessment reach are to be assessed collectively, i.e., <u>all</u> nutrient data collected along the reach are to be assessed together, using statistical tests. We recommend two statistical testing procedures to evaluate the nutrient dataset; the Exact Binomial Test and the One-Sample Student's T-test for the Mean. The rationale for using two statistical tests is in **Appendix A**. The tests are in two Excel spreadsheets and their use is described below.

To use the statistical tests, do the following:

[9] Lakes generally require year-round nitrogen and phosphorus criteria if they are to be protected from cultural eutrophication This may in turn affect the time-of-application of nutrient standards in the near-field tributaries of those lakes.

- For new, un-listed stream segments, use the Excel spreadsheet tool named "MT-NoncomplianceTool.xls".
- For already-listed stream segments, use the Excel spreadsheet tool named "MT-ComplianceTool.xls"
- In both tools, for either test, set alpha to 0.25 (25%)[10]. For the Binomial set the critical exceedance rate to 0.2 (20%) in cell B6. The effect size (gray zone) should be 0.15 (15%) and is set as a function of the exceedance rate. So, in the "MT-NoncomplianceTool.xls" this means p2 should be set to 0.35 (i.e., enter 0.35 into cell B7), and in "MT-ComplianceTool.xls" p2 should be set to 0.05 (enter 0.05 in cell B7).

Both tests (Binomial, T-test) will produce a result (PASS, FAIL). For the Binomial, you need to compare the allowable number of exceedances shown by the test ("**e**s", found in column D) to the actual number of exceedances manifested by your dataset. For the T-test, you will need to enter the dataset into the spreadsheet along with the criterion concentration against which the data are being compared. If the assessment reach complies with a test, the result is PASS; if the assessment reach does not comply with a test, the result it FAIL.

Note: If a non 303(d)-listed nutrient species is the same element as a listed one (e.g., stream is listed for nitrate, but you are also assessing TN, and TN is not currently listed), use the "MT-ComplianceTool.xls" for the non-listed nutrient species as well.

3.2.1.1 Minimum Sample Size for Nitrogen and Phosphorus Data

In the vast majority of cases the assessor will be making the nutrient-impairment decision with a fairly small nutrient-concentration dataset (probably < 13 samples). Statistics derived from small datasets such as these are subject to a fair amount of uncertainty. For example, outcomes from the Binomial Test (compliant, non-compliant) will, for nutrient sample sizes around 13, have confidence levels of about 75% (i.e., alpha and beta error of about 25% each).

For assessment reaches, the target number of nutrient samples is 12 (new, un-listed stream segments) or 13 (already-listed stream segments). The rationale for these sample sizes is presented in **Appendix A**.

HOWEVER: Cases exist where a dataset smaller than 12 or 13 will provide a sufficiently clear result that further nutrient sampling is not warranted. At about 6 samples or less, beta error in the Binomial test can become unacceptably high (> 65%) and increasingly worse with smaller ***n.*** At 7-8 samples, however, there are cases where a certain number of exceedances would be extremely unlikely unless the stream's true exceedance rate was much in excess of 20%. Therefore, for sample sizes of 7, ≥ 4 exceedances can be considered FAIL for the Binomial test. If <4 exceedances are found, sampling should be resumed until the minimum of 12 or 13 is achieved. The T-test can also be used with 7 samples but its power is lower at this sample size. Please see the bullets in **Appendix A**, **Section A.5**.

Also, circumstances may arise where nutrient sampling that is planned over two field seasons may lead to a reduction in the necessary number of samples. For example, if at the end of year one ten (10) TN and TP samples have been collected from an assessment reach on an unlisted stream, and the number of exceedances in each dataset is one (1), it would not be necessary to collect the addition two samples

[10] Alpha, exceedance rate, and the gray zone can be changed via the input cells in the upper left hand corner of the spreadsheet.

(to achieve 12) the following year. This is because even if both of the subsequent samples were above the criteria, the decision outcome (assessment reach "attains") would not be altered. Assessors should consider these types of situations at the end of each field season in order to best optimize work and cost.

Important Caveat: When the nutrient-concentration dataset is large (as defined below), the nutrient impairment decision should be made using nutrient concentrations alone.

- Large nutrient dataset for already-listed segments: 90 samples in the assessment reach
- Large nutrient dataset for unlisted segments: 50 samples in the assessment reach.

The large sample sizes were determined using the Binomial distribution with an alpha of 0.05 (95% confidence level) and a balance between alpha and beta error (i.e., beta is also about 0.05).

If the large sample sizes listed above are available, the assessor should generally forgo the use of parameters other than total nitrogen and phosphorus (i.e., those in **Table 3-1**) in nutrient-assessment decision making. Nutrient concentrations alone can be used to assess standards compliance, via the Binomial test only.

3.2.2 Minimum Sample Sizes for Other Parameters

The remaining parameters in **Table 3-1** (with the exception of BOD_5; more on it below) are effect variables, i.e. they are affected by changes in nutrient concentrations. Sample size requirements for each parameter are summarized below. Each result, from a sampling event and for a parameter, should normally be considered on its own merits when using the decision spreadsheet (NtrntAssessFramework.xls) and completing the assessment. Exceptions to this exist; see below. The parameters applicable to specific regions (mountain and transitional streams vs. plains streams) and the impact thresholds associated with those parameters are given **Section 4.0** and **Section 5.0**. ***Important Note*: Within their region of application, parameters shown below are required in order to carry out a level I assessment (see Figure 3-1). If a parameter is only required for a level II assessment, this will be indicated.**

Benthic Algal Biomass Samples (Chl *a* and AFDW): At least three (3) sampling events for benthic algal biomass are to be carried out in the assessment reach. These may include approved visual estimation methods. If more than one site is established in the assessment reach, disperse sampling effort across the different sites. Otherwise, assure that about 30 days have passed before sampling again at the same site.

Diatom Samples (Nutrient Increaser Taxa Metric): At least two (2) diatom sampling events are to be carried out in the assessment reach. If more than one site is established in the assessment reach, disperse sampling effort across the different sites. Otherwise, assure that about 30 days have passed before sampling again at the same site. **Note:** Diatom samples are required at level I in the plains, but are only required for a **level II** assessment in the **mountain and transitional region.** However, since there is no validated diatom increaser metrics for the Middle Rockies ecoregion, they are not required collection there.

Macroinvertebrate Samples (HBI Metric): At least two (2) macroinvertebrate sampling events are to be carried out in the assessment reach, unless the site is in the Middle Rockies ecoregion, in which case at least three (3) sampling events are to be carried out. If more than one site is established in the assessment reach, disperse sampling effort across the different sites. **If only one site is present is the assessment reach**, **or if you decide to collect across-time samples at several sites**, collect the across-time samples at the site(s) approximately 30 days apart. The across-time HBI scores from a site should then be averaged prior to comparison to the threshold. If you only have one site, you will only have one HBI value for the assessment reach to compare to the threshold. **Note:** Macroinvertebrate samples are only required for a **level II** assessment, and only in the **mountain and transitional region**.

Sampling to Determine Dissolved Oxygen Delta: At least three (3) DO sampling events (i.e., days) are to be carried out in the assessment reach. If more than one site is established in the assessment reach, disperse sampling effort across the different sites. **DO deltas fluctuate rapidly, and therefore you do not need to wait 30 days to collect subsequent DO data at a site.** For example, collection at a site over 3 contiguous days is acceptable. The more DO deltas that can be collected in the assessment reach, the better.

BOD_5: At least three (3) BOD sampling events are to be carried out in the assessment reach. If more than one site is established in the assessment reach, disperse sampling effort across the different sites. Samples for the standard 5-day biochemical oxygen demand (BOD_5) test are similar to nutrient samples, in that they are a stream-water measurement that can change rapidly, and BOD's affect on DO varies according to other factors (e.g., wind mixing). **Note:** BOD_5 samples are only required for a **level II** assessment, and only in the **plains region**.

Observation Data for Macrophytes and Benthic Algae: The Fish Cover/Other form, i.e. the component of the form pertaining to macrophytes and benthic algal growth, is to be filled out in accompaniment with each benthic algal biomass and/or diatom sampling event. It should also be filled out at each site in the assessment reach at least once each summer. If across a summer growth conditions have notably changed at the site, fill it out again.

3.2.3 Determine the Nutrient Most Likely to be Harming Use(s)

Normally both N and P, and potentially different species of N (e.g., nitrate, TN), will be simultaneously evaluated within an assessment reach. Cases will arise where the harm-to-use signal from one element or the other is clearly stronger, which will help streamline the assessment determination and subsequent work (e.g., TMDL development). An example is provided below.

Table 3-4. Simultaneous Review of Multiple Nutrients and Effect Variables

Assessment Reach	Nutrient	n	Binomial	T-test	Diatom Increaser Taxa	Benthic Chl *a*
Fred Cr reach 1	NO_3+NO_2	14	PASS	PASS	Exceeds criteria	Exceeds criteria
Fred Cr reach 1	TN	14	PASS	PASS	Exceeds criteria	Exceeds criteria
Fred Cr reach 1	TP	14	FAIL	FAIL	Exceeds criteria	Exceeds criteria

Total P results in **Table 3-4**, when run through the assessment process in the "NtrntAssessFramework.xlsx" spreadsheet, result in a clear "nutrient impaired" decision, at level I. This is largely driven by the two FAILS for the statistical tests (i.e., TP concentrations were very elevated). But because each nutrient type is assessed separately and TN is PASS for both statistical tests, the TN outcome is considered unclear and would, as a result, move to level II. As a result, the TN-impairment determination would be driven only by the biotic response variables and not by TN (see scenario 10,

'Mountain and Transitional' tab, in NtrntAssessFramework.xlsx). However, the most succinct conclusion (again, applying Occam's razor) is that the problem in this assessment reach is related to P, not N, as can be seen in **Table 3-4**, and only P should be listed. Arranging and reviewing the data as shown in **Table 3-4** should prevent unnecessary chasing of vague results for a nutrient when clearly the alternate nutrient is the issue.

Cases will arise where both nutrients will give mixed results within the two statistical tests, and therefore neither nutrient is clearly the culprit. In such cases, in accordance with the final outcome of the weight-of-evidence assessment, N and P should probably both be listed as probable causes.

3.2.4 Examples of Nutrient-impairment Decisions

Below are 3 assessment reach examples and their outcomes, to demonstrate the process.

(**1**) The assessment reach is in western MT and has 6 sampling sites. Each site has been sampled 2 times for nutrient concentrations (TN, TP) and once for benthic Chl *a.* **Action*:*** The nutrient samples (n = 12) are assessed, by type (TN or TP), using the two statistical tools, which will result in PASS or FAIL for each test. Both TP tests are FAIL, but the TN tests are both PASS. Each of the six Chl *a* sampling events (each comprised of 11 replicates which have been reduced to a sampling-event average) are independently compared to the criteria. One of them exceeds 120 mg Chl *a*/m^2, so declare Chl *a* as 'Exceeds Criteria' for the assessment reach. The data suggest a TP problem but not a TN problem, per methods in **Section 3.2.3** above. TP is listed as the cause, and further data collection and assessment for TN is not necessary.

(**2**) The assessment reach is on an eastern MT plains stream. There are 3 sampling sites where nutrients have been sampled 4 times and DO has been continuously monitored by deployed instrument for two summer months, at one site. **Action*:*** The nutrient samples (n = 12) are assessed, by type (TN or TP), using the two statistical tools, resulting in PASS or FAIL for each test. Both TN tests are FAIL, TP tests are mixed (1 PASS, 1 FAIL). Daily DO deltas from the long-term DO dataset should be calculated and compared to the DO delta threshold of 5.3 mg/L. Because this is a long-term dataset, close attention should be paid to the percent of DO deltas exceeding the threshold; if >10%, DO would be declared 'Exceeds Criteria' for the assessment reach[11]. Both TN and TP are suspected (N much more strongly) and should both be listed.

(**3**) The assessment reach is in western MT and the assessment has gone to a 2nd year of data collection. There are four sites. Nutrients have been collected at the 4 sites three times, benthic Chl *a*/AFDW once at each site, macroinvertebrate samples have been collected at 3 sites once each, and twice at the 4th site, and diatom samples have been collected at all 4 sites two times each. **Action*:*** All data from both years are first routed through the level I decision framework. The nutrient samples (n = 12) are assessed, by type (TN or TP), using the two statistical tools, which will result in PASS or FAIL for each test. Both TP are FAIL, and both TN tests are PASS Each of the four Chl *a* and AFDW sampling events (each comprised of 11 replicates which have been reduced to a sampling-event average) are independently compared to the criteria. One of them exceeds 35 g/m^2 AFDW, which is sufficient to declare Chl *a*/AFDW as 'Exceeds Criteria' for the assessment reach. Each macroinvertebrate HBI metric score where there is only one observation per site (three sites) is considered independently. For the 4th site, the two temporally-

[11] If DO deltas >5.3 mg/L comprise < 10% of the dataset, consider if the site has a strong presence of macrophytes or not. If macrophytes are very common, the site could be declared as 'Meets Criteria'. If macrophytes are not common, it could be declared 'Does Not Meet Criteria'. Consult Standards Section for further assistance.

collected HBI scores are averaged. One of the preceding macroinvertebrate HBI scores is >4, thus 'Exceeds Criteria' would be declared for the macroinvertebrate category for the assessment reach If the diatom increaser taxa scores (all 8) were each <50% probability of impairment, then the diatom metric score would be declared as 'Meets Criteria' for the assessment reach. The assessment reach is found to be impaired for TP at level I without having to use the macroinvertebrate results. For TN, the assessment could move to level II assessment and this would show nutrient impairment; but that outcome is driven purely by biometrics, which are sensitive to both nutrients. The overall dataset, per methods in **Section 3.2.3** above, suggests a TP problem but not a TN problem. TP is listed, and further data collection and assessment for TN is not needed.

3.2.5 Overwhelming Evidence of Nutrient Impairment-All Regions

Some circumstances related to excess nutrient pollution are severe enough that a rigorous data collection effort is not required. Photo documentation will suffice. Below are listed conditions that can be considered overwhelming evidence; these apply equally to wadeable streams across the state. These conditions are likely to be intertwined with organic pollution problems.

- Fish kills involving massive growths of senescing algae mats. These mats may be attached to the bottom or floating. Dissolved oxygen levels at dawn will likely be less than 1 mg/L.

- Filamentous algal growth covering the entire bottom from bank to bank and extending continuously for a substantial longitudinal distance (> 150m). Use the photographs below (**Figure 3-2** and **3-3**) as a guide. Don't confuse these conditions with sporadic, longitudinally-patchy growths of heavy filamentous growth, in between which there is lighter algal growth. The latter are not extreme enough to warrant overwhelming evidence, and should be sampled/assessed per the method earlier described.

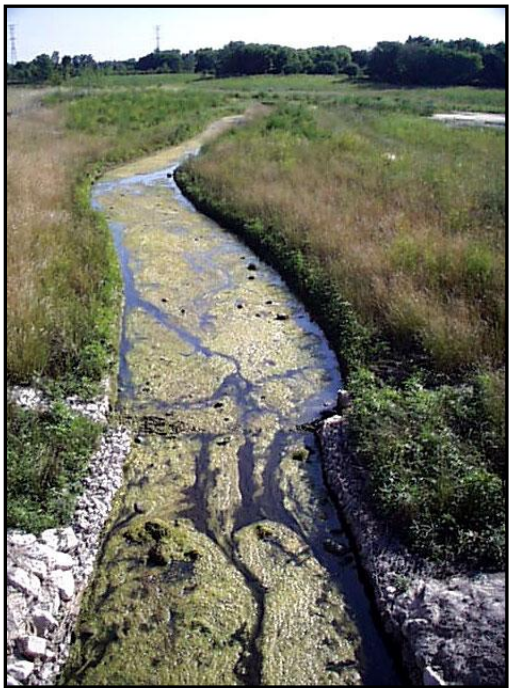


Figure 3-2. Photographs of heavy, bank-to-bank and longitudinally continuous Cladophora growth
Left photo is from (Sandgren, et al., 2004).

Photo courtesy Dr. Vicki Watson.

Figure 3-3. Massive *Cladophora* growth in the Clark Fork River, MT, 1984.
This nuisance alga is aptly named "blanket weed".

4.0 Nutrient Impairment Assessment Methodologies: Wadeable Streams in the Mountain and Transitional Region

The following subsections describe assessment methods best suited for use in mountainous streams and streams that transition between mountains and plains. Analysis shows that level IV and level III ecoregions are the most useful classification tool for defining nutrient zones (Varghese and Cleland, 2005), and nutrient criteria have been developed using ecoregions as the base zoning system (Suplee *et al.*, 2008). Consideration has also been given to the legal classification system for streams (B-1, C-3, etc.) which defines the streams' designated beneficial uses. There is a very high degree of correspondence between streams with salmonid fish among their beneficial uses (A-closed, A-1, B-1, B-2, C-1, C-2) and certain groups of ecoregions. Specifically, the mountainous level-III ecoregions (15, 16, 17, and 41) plus specified level-IV ecoregions along the Rocky Mountain front — Level IVs that are subunits of the level-III Northwestern Glaciated Plains (42) and Northwestern Great Plains (43) ecoregions — comprise a group well suited for assessment methodologies presented in this section. Four (4) additional level IV ecoregions (42l, 42n, 43o, 43t) that were not presented in Suplee *et al.* (2008) have been added to the group. These four level IV ecoregions are also transitional along the Rocky Mountain Front and comprise regions in which all or virtually all waterbodies are classified as supporting salmonid fishes among their beneficial uses. The regions are shown in **Table 4-1**.

Table 4-1. Ecoregions (levels III and IV) in which Assessment Methods in this Section Best Apply

Ecoregion Scale	Ecoregion Name	Ecoregion Number
Level III	Northern Rockies	15
Level III	Idaho Batholith	16
Level III	Middle Rockies	17
Level III	Canadian Rockies	41
Level IV	Sweetgrass Uplands	42l
Level IV	Milk River Pothole Upland	42n

Table 4-1. Ecoregions (levels III and IV) in which Assessment Methods in this Section Best Apply

Ecoregion Scale	Ecoregion Name	Ecoregion Number
Level IV	Rocky Mountain Front Foothill Potholes	42q
Level IV	Foothill Grassland	42r
Level IV	Unglaciated Montana High Plains	43o
Level IV	Non-calcareous Foothill Grassland	43s
Level IV	Shields-Smith Valleys	43t
Level IV	Limy Foothill Grassland	43u
Level IV	Pryor-Bighorn Foothills	43v

Note: The level IV ecoregion "Unglaciated Montana High Plains" (43o) has more than one polygon in Montana. Only the polygon located just south of Great Falls, MT should be considered part of the transitional streams group. Also, the level IV ecoregion "Foothill Grassland" (42r) has polygons associated with both the Middle Rockies *and* Canadian Rockies level III ecoregions. 42r polygons are associated with the level III ecoregion (either Middle Rockies or Canadian Rockies) against which they abut.

4.1. Assessment of Benthic Algal Growth

For wadeable streams, we recommend that site-average benthic algae densities of 120 mg Chl *a*/m^2 and 35 g AFDW/m^2 be used as thresholds (i.e., maximum allowable levels) to prevent impact to the fish and associated aquatic life uses (i.e., to maintain DO standards in DEQ-7), and the recreation use (ARM 17.30.637(1)(e)). Details on how these values were derived are in **Appendix B**.

Note: AFDW results from core samples should never be included in determining a site's average AFDW. The method measures organic material from the entire core sample, not just the surface where the algae are growing, and will therefore over-report AFDW.

Each sampling event result should be considered on its own merits when using the decision spreadsheet (NtrntAssessFramework.xlsx) and completing the assessment. That is, if 3 sampling events for benthic alga growth were undertaken and 1 of the Chl *a* averages exceeds the recommended threshold, then the conclusion for the assessment reach for the benthic algae category would be "exceeds 120 mg Chl *a*/m^2".

4.2 Assessment Using Biometrics

Biometrics based on diatom algae are stressor-specific (e.g., address nutrient pollution) and apply to specific regions. A diatom sample that indicates >51% probability of impairment by nutrients indicates the sample is from a site manifesting an excess nutrient problem. Details on how the diatom biometrics were developed and the thresholds derived are presented in the periphyton SOP (Montana Department of Environmental Quality, 2011b).

Always consider cautiously the results from samples collected very early and very late in the sampling season. Algae are a successional community, and if you sample too early, you will sample fewer 'pioneer' species and too late, you will start seeing the community as a whole die off - some taxa sooner than others. These changes can affect metric results.

Various biometrics based on macroinvertebrates were reviewed. We selected the Hilsenhoff Biotic Index (HBI) as the best tool for assessing nitrogen and phosphorus pollution problems. An HBI score of 4.0 should be used as the threshold (i.e., maximum allowable value) to prevent impact to fish and associated aquatic life uses. Details on how the biometrics were selected and the thresholds derived are presented in **Appendix B**.

Each sampling event result for a biometric should be considered on its own merits when using the decision spreadsheet (NtrntAssessFramework.xlsx) and completing the assessment. That is, if 2 sampling events for macroinvertebrates were undertaken and 1 of the results was an HBI score of 5.0, then the conclusion for the assessment reach for the macroinvertebrate category would be “>4.0” (i.e., exceeds).

5.0 Nutrient Impairment Assessment Methodologies: Wadeable Streams in the Plains Region

Table 5.1 below shows areas of the state in which the methods of this section best apply. Essentially, the methods apply to all of ecoregion 42 (Northwestern Glaciated Plains) and ecoregion 43 (Northwestern Great Plains) *except for* the level IV ecoregions along the Rocky Mountain Front which are being lumped with the mountainous ecoregions (see **Table 4-1**).

Table 5-1. Ecoregions (level III) in which Assessment Methods in this Section Best Apply
Some level IV ecoregions associated with the level IIIs shown are excluded; these are listed below each level III.

Ecoregion Scale	Ecoregion Name	Ecoregion Number
Level III	Northwestern Glaciated Plains	42
Level IV ecoregions of the Northwestern Glaciated Plains not in the Warm Water Fishery Class:		
Level IV	Sweetgrass Uplands	42l
Level IV	Milk River Pothole Upland	42n
Level IV	Rocky Mountain Front Foothill Potholes	42q
Level IV	Foothill Grassland	42r
Level III	Northwestern Great Plains	43
Level IV ecoregions of the Northwestern Great Plains not in the Warm Water Fishery Class:		
Level IV	Unglaciated Montana High Plains	43o
Level IV	Non-calcareous Foothill Grassland	43s
Level IV	Shields-Smith Valleys	43t
Level IV	Limy Foothill Grassland	43u
Level IV	Pryor-Bighorn Foothills	43v

Note: The level IV ecoregion “Unglaciated Montana High Plains” (43o) has more than one polygon in Montana. Only the polygon located just south of Great Falls, MT is excluded from the Warm Water Fishery Class.

5.1 Assessment Using Biometrics

Biometrics based on diatom algae are stressor-specific (e.g., address nutrient pollution) and apply to specific regions. A diatom sample that indicates **>51%** probability of impairment by nutrients indicates the sample is from a site manifesting a nutrient problem. Details on how the diatom biometrics were developed are presented in the periphyton SOP (Montana Department of Environmental Quality, 2011b).

Each biometric sampling event result should be considered on its own merits when using the decision spreadsheet (NtrntAssessFramework.xlsx) and completing the assessment. That is, if 2 sampling events for diatoms were undertaken and 1 of the results was "65% probability of impairment by nutrients", then the conclusion for the assessment reach for the diatom category would be ">51%," (i.e., exceeds).

Note: Diatom biometrics for the plains region have an inherently high false negative rate (62%; i.e., the chance that the metric declares a truly nutrient-impacted site as having no nutrient impact). This fact is given consideration, in that the resulting decisions in the spreadsheet (NtrntAssessFramework.xlsx) lean somewhat to the protective side.

Always consider cautiously the results from samples collected very early and very late in the sampling season. Algae are a successional community, and if you sample too early, you will sample fewer 'pioneer' species and too late, you will start seeing the community as a whole die off - some taxa sooner than others. These changes can affect metric results.

5.2 Assessment Using the Difference Between the Daily Maximum Dissolved Oxygen Concentration and the Daily Minimum Dissolved Oxygen Concentration (Delta)

We recommend that the magnitude of the daily DO concentration change (daily maximum minus the daily minimum, or *delta*) be used to assess plains streams. Elevated daily DO delta values indicate high productivity and the potential for DO standards exceedances (per DEQ-7) that would impact fish and aquatic life. We suggest that a DO delta of 5.3 mg/L be used as the threshold. **Assessors need not wait 30 days to take subsequent DO measurements at a site; each DO sampling event may be considered on its own merits.** Details on how the DO threshold was identified are provided in **Appendix C**.

Each DO sampling event result should be considered on its own merits when using the decision spreadsheet (NtrntAssessFramework.xlsx) and completing the assessment. That is, if 5 sampling events for DO delta were undertaken and 1 of the DO deltas exceeds the recommended threshold, then the conclusion for the assessment reach for the DO delta category would be "exceeds 5.3 mg/L". Further consideration may be needed if the data were collected long-term[12]. **Note:** Do deltas in the plains region have an inherently high false negative rate (63%; i.e., the chance that the DO deltas indicate that a truly

[12] If DO deltas >5.3 mg/L comprise < 10% of the dataset collected using an instrument deployed at least 14 days, consider if the site has a strong presence of macrophytes or not. If macrophytes are very common, the site could be declared as 'Meets Criteria'. If macrophytes are not common, it could be declared 'Does Not Meet Criteria'. Consult Standards Section for further assistance.

nutrient-impacted site has no nutrient impact). This fact is given consideration in that the resulting decisions in the spreadsheet (NtrntAssessFramework.xlsx) lean somewhat to the protective side.

5.2.1 Deploying a Continuous DO Monitoring Device in Wadeable Plains Streams

If a continuous monitoring device is to be deployed (e.g., YSI 6600 sonde), we recommend that at least one (1) full day of data be collected to properly calculate a daily DO delta. The day of deployment and the day of retrieval are usually truncated, and so at least one full day in between assures the necessary data are collected. We recommend a fifteen minute time step for monitoring, as that has worked well in our experience and provides good data resolution. Initial calibration, as well as drift from calibration— which is determined at the time the unit is retrieved— should be documented per the project's QAPP and or SAP.

5.2.2 Instantaneous DO Monitoring in Wadeable Plains Streams

Daily DO minimum and maximum concentrations each need to be obtained, and can be collected using a hand-held instrument. The daily DO minimum needs to be collected starting in the pre-dawn hours, up to as late as 8:00 am. The daily DO maximum will *usually* occur between 2:30 pm and 5:00 pm; observations can be collected every 15-30 minutes during this time to identify the highest value. Continue monitoring after 5:00 pm if observations are still climbing. Further details on how these time frames were identified are provided in **Appendix C**.

The YSI 85 instrument has a 50 reading, manual-entry memory which can be used for collecting DO maximums. With the unit on and the sensor properly deployed, depress the ENTER button for two seconds to record an instantaneous observation. Data may be downloaded later.

For the purpose of calculating DO delta, at least 3 DO sampling events (i.e., 3 different days) should be taken in each assessment reach; if collected at the same site, they do <u>not</u> need to be collected 30 days apart (e.g., 3 days in a row is OK).

5.2.3 BOD_5

We recommend that biochemical oxygen demand, or BOD_5 (also called just BOD), be used to assess plains streams at level II. We recommend a threshold of 8.0 mg/L be used. Each BOD sampling event result should be considered on its own merits when using the decision spreadsheet (NtrntAssessFramework.xlsx) and completing the assessment. That is, if 3 sampling events for BOD were undertaken and 1 of the BODs exceeds the recommended threshold, then the conclusion for the assessment reach for the BOD category would be ">8.0 mg/L" (exceeds).

5.2.4 Algal and Macrophyte Indicators of Nutrient Enrichment in Wadeable Plains Streams

The Fish Cover/other form (see periphyton SOP, (Montana Department of Environmental Quality, 2011b)) should be filled out when assessing plains streams. Although not required to fill out the form, we recommend that the dominant macrophytes be identified, which will help with your assessment back in the office using the information below.

In the Northwester Glaciated Plains ecoregion, streambed cover by filamentous algae should generally be less than 30% for a single sampling event and less than 25% for the summertime average (Suplee, 2004). These data can be collected visually using the Fish Cover/Other form, which is provided in the

periphyton SOP (Montana Department of Environmental Quality, 2011b). (Although a somewhat tangential issue, the presence of a healthy and widely distributed macrophyte community should be taken as indicative that the stream has a reasonable level of morphologic stability; stream instability has been found to be a major factor in controlling algae and macrophyte dynamics in prairie streams [Suplee, 2004]).

Throughout the plains region, attention should be paid to the types of macrophyte species present. We have observed that northern watermilfoil (simultaneously known as *Myriophyllum exalbescens, Myriophyllum sibiricum, and Myriophyllum spicatum L. var exalbescens* (Muenscher, 1944). DiTomaso and Healy (2003) is extremely common throughout the Northwestern Glaciated Plains and Northwestern Great Plains ecoregions, as has been observed by others (Klarich, 1982). However, in stream sites where high nutrient enrichment is occurring, we have observed northern watermilfoil's (and other macrophyte's) near-complete replacement by coontail (aka hornwort), *Ceratophyllum demersum*[13] Coontail is a rootless, free floating macrophyte—though it can anchor itself to bottom substrates via specialized buried stems—that can proliferate in streams which are being heavily loaded with nutrients (DiTomaso and Healy, 2003). In this it is similar to floating and benthic algae in that it relies on water-column nutrients for growth, because it does not take up nutrients from the sediment via roots, as other macrophytes do. Choking mats of coontail, or its presence to the exclusion of other macrophytes, should be taken as a strong indicator of nutrient over enrichment. Close-up and panoramic photos should be taken to record the extent of the problem, and aide identification of the plants in-office.

Finally, we documented during the Box Elder Creek dosing study that *Chara* spp. (commonly called stonewort or muskgrass) were greatly depressed in number in the nutrient-dosed reaches compared to the Control reach, and also compared to the pre-dosing period. *Chara* spp. are a branched form of algae, are an important component of natural aquatic ecosystems (DiTomaso and Healy, 2003), and are often associated with clean water.

6.0 ACKNOWLEDGEMENTS

We would like to thank Dr. Vicki Watson and her students for the many years of work she has carried out in support of DEQ projects. These projects (e.g., the reference streams project) have provided the basic data required to develop many of the guidelines in this document. We would also like to thank all of the DEQ staff and field crews who, over the years, have collected mounds of data that also support this work. Thanks to the Carter County Conservation District for the cooperation and support on the nutrient dosing study. Finally, we want to thank the contractors whose excellent work helped build the foundation on which this document stands: Arun Varghese and Josh Cleland (ICF International); Dr. Loren Bahls (*Hannaea*); Mark Teply (Cramer Fish Sciences); Wease Bollman (*Rhithron* Associates); and Dr. Jeroen Gerritsen and Dr. Lei Zheng (Tetra Tech).

[13] Coontail and watermilfoil can readily be distinguished in the field with a good macrophyte guidebook and a hand-held magnifying glass.

APPENDIX A. STATISTICAL CONSIDERATIONS

A.1.0 INTRODUCTION

The numeric nutrient criteria addressed in this appendix are not intended to be ideal standards, i.e., "no sample shall exceed" values. As such, appropriate inferential statistical tests, assumptions about stream sampling frames, etc. must be developed so that the criteria can be correctly applied. This appendix outlines these statistical considerations and provides rationales for the various approaches used. It also provides precautionary points where certain assumptions depart from more conservative statistical thinking, and discusses how improper sampling design has the potential to mislead a conclusion made about a stream's condition. The key issues addressed herein are:

- Sampling frames, populations, and sampling units for streams, and associated assumptions and precautions
- Consideration of what constitutes our best description of sample independence in streams (spatial and temporal), and associated assumptions and precautions
- Determination of appropriate critical exceedance rates for nutrients (nitrogen and phosphorus)
- Statistical testing procedures and accompanying decision rules
- Minimum sample sizes

A.2.0 SAMPLE FRAME, POPULATION, AND SAMPLING UNITS

All studies involving statistical evaluations of data require that a sample frame, population, and sampling unit be defined. Streams are particularly poor entities for establishing these parameters because streams are an interconnected network rather than discreet entities. Nevertheless, streams are the entities to be sampled so some effort must be made to segregate them into definable units. For the purposes of determining compliance with numeric nutrient criteria, we define the following:

- **Sample Frame:** A wadeable[14] stream segment listed in the Assessment Data Base (ADB) (DEQ 2009, and updates) OR a sub-segment of an ADB stream segment. These segments are referred to here as an "assessment reach".
- **Population:** All the water flowing through the assessment reach during the time period when the numeric nutrient criteria apply, and the surface area of the stream bottom over which the water flows.

[14] Wadeable streams are perennial and intermittent streams in which large portions of the channel are wadeable during baseflow conditions. For the list of waterbody segments not considered wadeable (i.e., the large rivers), see Flynn and Suplee (2010).

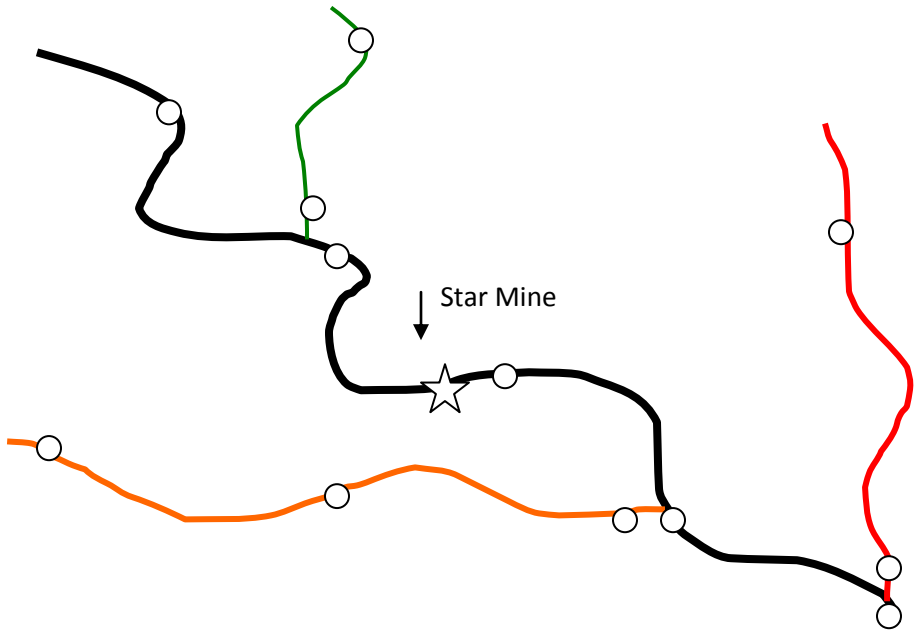
- **Sampling Unit:** A sample collected from the assessment reach that is largely independent of other samples collected within the assessment reach and collected during the time when the numeric nutrient criteria apply.

Assumptions: Each assessment reach (ADB segment or sub-segment) will be made up of a series of sampling sites, or a series of very short study reaches that are essentially sites (**Figure A2-1**). The minimum number of sites on an assessment reach is provided in DEQ SOPs (Montana Department of Environmental Quality, 2005). **Figure A2-1** illustrates the variety of ADB segments that may be found; segment lengths can vary tremendously. For purposes of determining compliance with numeric nutrient standards using statistical methods, it is usually assumed that (1) pollution sources are evenly dispersed along the reach, (2) sampling sites are randomly located along the reach, and (3) each sample is independent of the others (Spatial and temporal independence guidelines for sites are addressed in **Section A.3** below).

It some cases, ADB segments may have pollution problems (hotspots) concentrated only in a particular part of the stream, say, the last 5 stream miles. In such cases, it may not make sense to view the original ADB segment as the best possible sampling frame. That is, it would be better to further stratify the sample frame and, thus, the population of interest .This will prevent distortion of results caused by mixing together, for common analysis, data from the relatively un-impacted sub-segment with data from the impacted sub-segment. For example, in **Figure A2-1** it might be prudent to consider the sub-segment upstream of the Star Mine as a sampling frame apart from the sub-segment below the mine. Stratification is common in studies employing purely random sampling, where it is referred to as stratified random sampling (Cochran, 1977). Stratification allows maximal precision of estimates for minimal sampling effort (Norris, et al., 1992). **The assessor carrying out the analysis on an ADB segment will have to judge if further stratification is warranted.** If it is warranted, then sampling requirements, described above and further detailed below, would apply to *each* of the new sub-segments (aka assessment reaches), individually.

Precautionary Considerations: Pollution sources are rarely evenly-dispersed along stream segments, violating assumption 1 above. And purely random sampling is usually not practical due to stream access issues, etc. Targeting only the known or potential pollution "hotspots" — even within an assessment reach that has been broken out from a larger ADB segment — could over represent the hotspots and distort the statistical tests. Sampling and analysis plans (SAPs) should proceed with goal-oriented sampling (U.S.Environmental Protection Agency, 2000) that works towards striking an equitable balance between the number of hotspot sites and the number of un- or minimally impacted sites *within the defined assessment reach.* That is, the aggregate of collected samples should be representative (U.S. Environmental Protection Agency, 2002) of the assessment reach as a whole. Advanced knowledge and expertise of the field will be needed to accomplish this (Norris, et al., 1992), and modifications to the assessment reach boundaries can be made on-the-fly during field work, if deemed necessary. It is possible to sub-segment a stream reach to the point where, for a particular assessment reach, there really is little left but hotspots; if this is the case, then the hotspots *are* representative of the assessment reach. As a general rule, it is better to lump than split to avoid unnecessary sampling and administrative work. **The requirement to create reasonably uniform assessment reaches is inherently in conflict with the need to "lump" for the purpose of keeping assessments as simple as possible. Judgment is needed to balance these two opposed factors and come up with an optimal sampling strategy.**

Although this quasi-systematic approach is not a substitute for truly random sampling it will, if carried out properly, achieve good sample interspersion and representativeness. For further discussion of randomization *vs.* interspersion approaches, see page 196 of Hurlbert (1984).



Example sampling sites (hollow dots) are shown along each segment.

Figure A2-1. Four different stream reaches (shown by different colors), each representing 1 sampling frame (ADB stream segment)

A.3.0 Determining Sample Independence

According to definitions in Hurlbert (1984), much sampling carried out by DEQ on individual streams tends to violate spatial and temporal independence assumptions and results in pseudoreplication. For example, samples collected over time at a site can be serially correlated, which precludes temporal independence (Hurlbert, 1984). However, the statistical views advocated by Hurlbert are not universally supported; contrary opinions on the matter can be found in the literature (Stewart-Oaten and Murdock, 1986; Stewart-Oaten, et al., 1992; Osenberg, et al., 1994) and have led to what one journal referred to as a "healthy debate" (*Ecological Applications*, volume 4, No. 1, 1994). In general, more needs to be known about detection of non-independence and the frequency with which temporally independent samples can be collected (Underwood, 1994).

Time-series collected samples from a site may be used in inferential statistical testing, if used cautiously; this requires that one assumes that actual trends in time are identical in magnitude and direction for all the sites across the study (Norris and Georges, 1993). Osenberg *et al.* (1994) examine time-series serial correlation of physical and biological measurements in a BACI (Before-After-Control-Impact) study and conclude that, in the marine environment they study, sampling can occur at a site every 60 days without yielding substantial serial correlation.

DEQ recognizes the issue of temporal pseudoreplication, but also needs to be practical about the reality of sampling streams which, by their very nature, make collection of independent samples difficult In DEQ's reference project (Suplee, et al., 2005), 30 days has generally been used as a minimum time span between sampling events at a site to infer temporal independence of water samples. This time span was based on the experiential observation that, during the brief Montana summer, substantial changes in flow, temperature, and vegetation (both riparian and instream) occur from month to month, changes that would likely effect water quality. But Stewart-Oaten *et al.* (1986) recommend that the assumption of temporal independence be tested, rather than assumed. The Durbin-Watson test statistic is widely used to check for time-series serial correlation. Stream sites with monthly nutrient sampling during the summer were available in Montana, and some of these sites were tested using the Durbin-Watson statistic Results are shown in **Table A3-1** below.

Table A3-1. Durbin-Watson Values for Time-series Collected Nutrient Samples at Selected Sites

All Samples were Collected Approximately 30 Days apart Nutrients Showing Probable Time-series Serial Correlation (95% Confidence Level) are Highlighted.

					Nutrient	
Stream Site	**Months Sampled**	**Time Range**	**n**	**Total N**	**Total P**	**NO_{2+3}**
Rock Creek Site 2	June, July, Aug, Sept	2001-2004	12	1.18	1.43	2.3
Clark Fork R. at Deer Lodge (site 9)	July, Aug, Sept	1998-2006	25	1.81	1.78	1.68
Clark Fork R. above Little Blackfoot R. (site 10)	July, Aug, Sept	1998-2006	26	2.01	1.57	1.46
Clark Fork R. above Flathead R. (site 25)	July, Aug, Sept	1998-2006	26	1.76	1.21	2.08

In general, Durbin-Watson values around 2 mean there is no serial correlation, whereas values greater than approximately 2.5 or less than about 1.5 lead one to suspect negative or positive serial correlation, respectively (Neter, et al., 1989; Ott, 1993). What can be concluded from this limited analysis? Most nutrients did not show serial correlation, and one of the three that did is borderline cases (statistic =1.43, but power of test very low). Overall, it appears that serial correlation is present in nutrient samples collected a month apart, but the effect is very weak It is evident that 30-day separated water samples can provide a fairly high degree of independence for nutrients.

DEQ is aware that spatial independence is also a concern. Water flows from upstream to downstream, consequently influencing the spatial independence of downstream sampling sites. No generally applicable spatial minimums were found as of this writing. U.S.EPA guidance (USEPA, 2002) generally glosses over the topic of spatial independence in streams.

To address spatial independence, we tested a Montana dataset. We used the pre-dosing baseline data collected as part of the Box Elder Creek nutrient dosing study (Montana Department of Environmental Quality, 2010a). We found that total nutrient samples collected within hours of one another at two sites located 0.73 stream miles apart were not spatially correlated. We compared nutrient samples collected from the Low Dose site to those collected on the same day at the High Dose site which is 0.73 miles downstream. Box Elder Creek is perennial and was flowing during all sampling events. No tributary intervenes between the sites. Samples were collected within 1-2 hours of one another, during the summer index period. We only considered samples collected *prior to* nutrient dosing, as these are comparable to what one would encounter during routine stream sampling/assessment. Using the Rank von Neumann test (U.S.Environmental Protection Agency, 2006), we found that there was no serial correlation for total N or total P (i.e., we could not reject the null hypothesis "no serial correlation"), at an alpha of 0.05. There was serial correlation for Soluble Reactive Phosphate (SRP). We were unable to

assess soluble N as there were too many non-detects in the datasets, which led to too many rank-ties; too many rank-ties precludes proper statistical evaluation (Gilbert, 1987).

Spatial independence can therefore be established (albeit as rules of thumb) for total nutrients as a minimum of about 1 mile between two sites. Other factors leading to spatial independence include a tributary confluencing on a stream between two sampling sites, or if major land form or land use changes occur along the reach (Montana Department of Environmental Quality, 2007; Montana Department of Environmental Quality, 2011a).

Giving consideration to our findings, below are guidelines for establishing independence of samples collected within an assessment reach:

- Sites (or short reaches equivalent to sites) should be located a minimum of 1 stream mile apart.
- Sites may be placed < 1 mile apart along the assessment reach if there is a flowing tributary confluencing with the segment between the two sites.
- Try to collect water samples starting at the downstream end of the assessment reach moving upstream, to avoid re-sampling the same water.
- Land use changes and land form changes should be considered and can be used to help define (1) breaks between assessment reaches and/or (2) additional sampling sites within an assessment reach.
- Samples collected at the same site (or short reach) should be collected 30 days apart.

Total nutrient samples that meet the above conditions may generally be considered spatially and temporally independent for the purposes of determining compliance with the nutrient criteria. As such, they may be used in inferential statistical analyses and to make conclusions about the assessment reach in question.

Precautionary Considerations: The last bullet above (temporal independence resulting from approximate 30-day time spans) is <u>not</u> applicable for some bioassemblage samples (e.g., macroinvertebrates, fish). These organism populations operate on different (longer) time scales from water samples and diatoms and may show considerable year-to-year stability. Please see **Section 9.0** of Suplee (2004)) and Bramblett *et al.* (2005) for more details on temporal patterns of these biological assemblages. Diatom populations tend to shift quickly, within 1-5 weeks, in response to environmental changes (LaVoie, et al., 2008). Thus, this rate of change is sufficient to be able to consider diatom sampling events spaced 30 days apart as being largely independent of one another.

A.4.0 Selection of Inferential Statistical Tests, Confidence Levels, and Associated Decision Rules

A.4.1 Rationale for Using Two Inferential Statistical Tests to Help Determine Compliance with Nutrient Standards

Exhaustive reviews of the pros and cons of statistical tests available for determining compliance with numeric standards have already been published (U.S. Environmental Protection Agency, 2002). For brevity, rather than revisit all the detailed considerations put forward in those documents, recommendations are provided herein concerning what where judged to be the most applicable tests. These recommendations are then followed by a series of decision rules that allow the user to apply the tests in tandem. For purposes of compliance with numeric nutrient criteria, two tests should be used; the Exact Binomial Test and the One-Sample Student's t-test for the Mean.

- Exact Binomial Test: This test assumes data are dichotomous in nature (i.e., only two possible outcomes). For compliance with a criterion this reduces to (1) samples that exceed the criterion and (2) samples that do not exceed the criterion. If confidence levels, power, and exceedance rates (more on these below) are established upfront, minimum sample sizes can also be determined. The main disadvantage of the test is that it is blind to exceedance magnitude; that is, it takes no account of whether a sample exceeds the criterion by 1% or 1,000%.

- One-Sample Student's t-test for the Mean: This test does not assume the data take on a dichotomous relationship relative to the criterion. The test compares the mean of the samples in question to the criterion. The desired confidence levels in the test are established upfront. But unlike the Exact Binomial Test, it *is* greatly influenced by high values and outliers which can skew the dataset mean relative to the bulk of the other samples in the dataset. It is also influenced by the proportion of non-detects in the dataset[15].

The Exact Binomial Test is useful for determining sample sizes, and is not influenced by large numbers of non-detects in the dataset. In fact, if the magnitude of nutrient criterion exceedances was irrelevant, then the Exact Binomial Test could probably be used by itself. But this is not the case; one must consider the issue of luxury nutrient uptake by algae.

One of the main purposes of establishing nutrient criteria is to control excess algae growth and its effects on water quality. Many benthic and water-column algae have the ability to take up the non-limiting nutrient, be that N or P, in excess of immediate need and utilize it for growth later (luxury nutrient uptake; (Elrifi and Turpin, 1985; Portielje and Lijklema, 1994; Stevenson and Stoermer, 1982). If extracellular nutrient concentrations then decline in the water, growth can still be maintained on intracellular stores (Droop, 1973; Rhee, 1973). Therefore, pulsed loading events of nutrients to streams may allow algae to carry out luxury nutrient uptake which can sustain growth for several cell generations well after the pulse has ended.

[15] For the purposes of using the t-test, users should initially convert all non-detects to 50% of the reported detection limit (USEPA, 2006). If >> than 15% of the dataset will be affected, consult Standards Section.

Luxury nutrients uptake is a kinetics phenomenon dependent on the physiological condition of the algae, duration and magnitude of the nutrient pulse, etc.; complex factors not easily addressed by a simple t-test. But the t-test can help assess the *potential* for luxury nutrient uptake because pulsed loads of elevated nutrient concentrations, if captured during sampling, would increase the dataset mean and would show that mean water quality has exceeded the criterion; this is useful information not provided by the Exact Binomial Test.

Each test possesses strengths the other does not. Therefore, we recommended that the t-test be used in tandem with the Exact Binomial Test via a series of decision rules (**Section 3.0,** main document).

A.4.2 Form of the Null Hypothesis, Alpha, Beta, Effect Size, and Critical Exceedance Rate

All of the factors listed in **Section A.4.2's** title are interrelated and influence one another in statistical hypothesis testing. Again, rather than reiterate here the mass of discussion devoted to these topics already covered elsewhere ((U.S. Environmental Protection Agency, 2002; California Environmental Protection Agency State Water Resources Control Board, 2004), we will simply state what we concluded to be the best statistical parameters (form of null hypothesis, alpha, beta, etc.) associated with the two tests (Exact Binomial and One-Sample Student's t-test for the Mean), and provide further explanation where warranted.

A.4.2.1 Form of the Null Hypothesis for the Statistical Tests

<u>For Streams Already on the 303(d) List</u>:

- Null Hypothesis: Waterbody is not in compliance with numeric nutrient standards
- Alternative Hypothesis: Waterbody is in compliance with numeric nutrient standards

<u>For Streams Not on the 303(d) List</u>:

- Null Hypothesis: Waterbody is in compliance with numeric nutrient standards
- Alternative Hypothesis: Waterbody is not in compliance with numeric nutrient standards

In effect, this is a "guilty until proven innocent" approach for streams already considered to have water quality problems, and an "innocent until proven guilty" approach for newly assessed streams California uses the same approach (California Environmental Protection Agency State Water Resources Control Board, 2004).

A.4.2.2 Alpha, Beta, Effect Size

In statistical testing alpha, beta, effect size, and critical exceedance rate interact, and changes in one affect changes in the others. In environmental compliance work, there are strong arguments for attempting to balance type I (alpha) and type II (beta) errors; in doing so, it is important to consider the form of the null hypothesis and the implications for making one error or the other. Basically, each type of error has ramifications; one type of error leads to degradation of the environment, the other type of error leads to unnecessary expenditures on the part of the regulated entity. Working towards balancing type I and II errors is a process which inherently recognizes the consequences of each error type (California Environmental Protection Agency State Water Resources Control Board, 2004, Page 52, Appendix C; Mapstone, 1995, Page 178; Schroeter, et al., 1993; U.S. Environmental Protection Agency,

2002). *Given that* working towards balancing type I and type II errors is a valuable endeavor, here are general recommendations for the parameters to be input into statistical tests for nutrients in wadeable streams:

- Alpha should be about 0.25 or less (equates to ≥ 75% confidence level), depending on the form of the null hypothesis and its implications.

- Beta should be about 0.3 or less (equates to ≥ 70% power), and will vary according to the samples size (more on sample size minimum in **Section 5.0**).

- Effect size (gray zone) should be set at 0.15, per USEPA (2002).

In the statistical spreadsheet tools that accompany this technical appendix ("MT-NoncomplianceTool.xls" and "MT-ComplianceTool.xls"), one or the other file is used depending upon whether you are dealing with a new, unlisted stream (use "MT-NoncomplianceTool.xls") *vs.* a 303(d) listed stream (use "MT-ComplianceTool.xls"). You will be able to set alpha, critical exceedance rate (p), and effect size (p2) in the Exact Binomial Test in both of the files. The program will then return various sample sizes, their associated beta values, and the maximum number of exceedances allowed while still remaining in compliance with the criterion.

For the One-Sample Student's T-test, you must input alpha and the nutrient criterion in mg/L. The One-Sample Student's T-tests will then provide a result indicating if the statistical test can or cannot confirm the alternative hypothesis. (The alternative hypothesis will reverse, according to whether you are using the tool for a listed or for a new, unlisted stream).

A.4.2.3 Critical Exceedance Rate

Critical Exceedance Rate: *An estimate of the actual proportion of samples that exceed an applicable water quality criterion. When more than this proportion exceeds the criterion, the standard is not attained (i.e., stream is not in compliance with standard).*

Among the four statistical parameters critical to the Exact Binomial Test—alpha, beta, effect size, and exceedance rate—exceedance rate needs some kind of empirical ground-truthing to assure its validity. The implications of different alpha and beta errors can be understood relative to the form of the null hypothesis, while the effect size (gray zone) is not knowable *a priori*, and is therefore assumed; we recommend an effect size of 0.15 per EPA (U.S. Environmental Protection Agency, 2002). In contrast, an exceedance rate can be estimated using lines of reasoning, empirical evidence and literature values. The considerations used to estimate an exceedance rate for numeric nutrient standards were (**1**) recommended exceedance rates from EPA (U.S. Environmental Protection Agency, 2002) and (**2**) long-term benthic algae and nutrient relationships on the Clark Fork River, MT (Consideration (**1**) and (**2**) are further detailed below.). We recommend:

- A critical exceedance rate for compliance with numeric nutrient standards be set at 0.2 (20%)

Below are our two major considerations leading to the selection of the 20% exceedance rate.

(1) EPA recommends that, for a number of different polluting substances (e.g., fecal bacteria, conventional pollutants, toxic trace metals, etc.), criteria exceedance rates be set between 0.1 and 0.25

(10 to 25%) to protect beneficial uses (Environmental Research Laboratory-Duluth, 1997; U.S. Environmental Protection Agency, 2002).

(2) The analytical approach described in **2.1** below was undertaken in June 2008, and only considered Clark Fork River data through 2006. Subsequent data collection (through 2009) and a somewhat different approach to ascertaining an acceptable exceedance rate allowed us to update this analysis, as provided in **2.2.** Both analyses (that from 2008 [**2.1**], and the work done in 2011[**2.2**]) arrive at the same basic conclusion, and both are presented here. If readers are already familiar with the work in **2.1,** we recommend you skip to **2.2**.

(2.1)*The following analysis was completed in June 2008*.

Introduction: Numeric nutrient (TN and TP) and benthic algae (mg Chl *a*/m^2) standards have been in place on most of the Clark Fork River in Montana for about 6 years. A systematic collection of nutrient and algae data has been ongoing since 1998. At a number of sites both algae and nutrient data have been collected multiple times each year for nearly 10 years. These data lent themselves well to empirically deriving a numeric nutrient exceedance rate because some river sites almost always exceed the algae standards, while others do not. The question became:

> *Do sites on the Clark Fork River that routinely exceed the numeric algae standards exceed the river's established numeric nutrient (TN and TP) standards more frequently than sites that do not exceed the numeric algae standards?*

Benthic algae levels in excess of 150 mg Chl *a*/m^2 (maximum) are not to be exceeded during the summer (ARM 17.30.631). Maximum in this case does not refer to a single high repeat measure from a Clark Fork River site; it refers to the mean value of a series of repeat measures (n = 15 to 20) that are collected at a site *during a particular sampling event.* Clark Fork River sites are usually sampled several times throughout the summer. It has been noted for some years that, during the summer, some sites are usually above the algae standards, while others are not. TN and TP standards were established on the Clark Fork River (ARM 17.30.631) and, if ultimately met, should keep benthic algae below the nuisance threshold described above. However, an exceedance rate was never explicitly established in the regulations In carrying out the exceedance rate determination described herein, it is assumed that the magnitude of the TN and TP criteria on the Clark Fork River were accurately determined, and therefore any exceedance rate drawn from this analysis is meaningful.

Methods: Benthic algae and TN and TP concentration data where concurrently available for seven Clark Fork River sites from 1998-2006. Data were restricted to the time period June 30th to October 1st to generally comply with the summer growing season for this ecoregion (Suplee, et al., 2007)and the regulatory timeframe in ARM 17.30.631. Every benthic Chl *a* measurement from a site (n = 15-20 per sampling event) collected over time was treated as a repeat measure. This resulted in a grand total of 285 to 333 repeat measures of Chl *a* at each site for the period 1998-2006. A grand benthic Chl *a* mean was calculated for a site by averaging all the repeat measures collected between June 30th and Oct 1st for all available years. Nutrient data collected at the corresponding sites during the same time frames where similarly compiled. At each site nutrients were collected as a single grab sample and, as a consequence, there were fewer data (43 to 78 N or P samples per site). Total N data were not collected; however, Total Kjeldahl Nitrogen (TKN) and NO_{2+3} were. Therefore, for each site, individual Total N concentrations were calculated by summing the TKN and NO_{2+3} sample results collected simultaneously during a sampling event.

Next, the Clark Fork River TN and TP criteria concentrations were matched to their corresponding values in the nutrient cumulative frequency distributions for each site, and the associated percentile was recorded. For example, the TN criterion for the Clark Fork River is 0.3 mg/L, and it resulted that at site 9.0 (Clark Fork at Deer Lodge) 0.3 mg TN/L corresponded to the 23rd percentile of site 9.0's cumulative TN frequency distribution. This process was carried out for all 7 sites for both TN and TP. There is a break at the Blackfoot River confluence where the Clark Fork's upstream TP criterion (0.02 mg/L) differs from that below (0.039 mg/L); each TP criterion was applied as appropriate for a site's location along the river.

Results: Table A4-1 shows the results for 3 sites that, over the 1998-2006 time period, did not exceed the Clark Fork River's benthic algal biomass criteria. For this group of sites the nutrient criteria exceedance rate (both TN and TP) was, on average, about 8%. That is, nutrient samples whose concentrations exceed the standards occur only about 8% of the time at these sites. **Table A4-2** shows three sites that *did* exceed the benthic algae standard; for this group of sites, the nutrient criteria exceedance rate was, on average, about 58%. Sites in **Table A4-1** (did not exceed algae standard) had a range of exceedance rates (TN and TP) from 0.1%-24%, and sites in **Table A4-2** (exceed algae standard) had a range of exceedance rates from 27.7% to 88%. The remaining site examined (Site 12; Clark Fork River at Bonita), which is not presented in **Tables A4.1** or A**4.2**, had a mean algae density (144 mg Chl *a*/m^2) so close to the algae standard it was considered borderline. Site 12's exceedance rate was 30.8% for TN, 68% for TP.

Table A4-1 Sites on the Clark Fork River (CFR) Not Exceeding the Maximum Benthic Algae Standard (Growing Season, 1998-2006)

			Percentile in Site's Nutrient Frequency Distribution Matching CFR Standard		**Criteria Exceedance Rate (%)**	
Clark Fork River Site #	**Site Name**	**Long-term Benthic Algal Biomass (mg Chl *a*/m^2, growing season) Mean [median]**	**TN**	**TP**	**TN**	**TP**
15.5	Clark Fork above Missoula	96 [80]	90th	95th	10.2%	5.4%
22	Clark Fork at Huson	72 [52]	76th	96th	24.0%	3.8%
25	Clark Fork above Flathead	35 [20]	100th	99th	0.1%	1.5%
					Grand Mean:	**7.5%**
					Grand Median:	**4.6%**
					Maximum:	**24.0%**
					Minimum:	**0.1%**

Table A4-2. Sites on the Clark Fork River (CFR) Consistently Exceeding the Maximum Benthic Algae Standard (Growing Season, 1998-2006).

			Percentile in Site's Nutrient Frequency Distribution Matching CFR Standards		Criteria Exceedance Rate (%)	
Clark Fork River Site #	**Site Name**	**Long-term Benthic Algal Biomass (mg Chl *a*/m^2, growing season) Mean [median]**	**TN**	**TP**	**TN**	**TP**
9	Clark Fork at Deer Lodge	180 [147]	23rd	50th	77.0%	50.0%
10	Clark Fork above Little Blackfoot River	163 [117]	48th	12th	52.0%	88.0%
18	Cark Fork at Shuffields	197 [181]	50th	72nd	50.4%	27.7%
					Grand Mean:	**57.5%**
					Grand Median:	**51.2%**
					Maximum:	**88.0%**
					Minimum:	**27.7%**

Discussion: The main assumption of this analysis was that the magnitudes of the Clark Fork River nutrient criteria, which were established as standards for the river, are correct. That is, if the nutrient standards are achieved, then summertime algae levels should be kept below the established nuisance thresholds. It was assumed that, as has previously been shown, both N and P co-limit in the Clark Fork River (Lohman and Priscu, 1992); (Dodds, et al., 1997). It was further assumed that the algae standard (150 mg Chl *a*/m^2, site mean per sampling event) will protect beneficial uses. Regarding the later, research completed since the Clark Fork River standards were adopted in 2002 show that 150 mg Chl *a*/m^2 (site mean) is identified as a nuisance threshold by the Montana public majority (Suplee, et al., 2009). If all these assumptions hold true, then reasonable exceedance rates for the 9 year dataset can be derived and used as a case study. It would have been ideal to have a true population of data (rather than a subset of data for a single river over a specific time period) with which to carry out this analysis. But such data are not readily available, and the long-term dataset examined here will have to serve as a proxy.

Comparison of Clark Fork River sites 15.5, 22, and 25 (don't exceed algae standard; **Table A4-1**) *vs.* 9, 10, and 18 (do exceed algae standard; **Table A4-2**) show a clear separation in the consistency of compliance with the river's numeric nutrient standards. It is clear from **Table A4-2** that if the exceedance rate is about 50% then nuisance algae growth will almost certainly occur. But when the exceedance rate is ca. 5-10%, nuisance algae is unlikely to occur (**Table A4-1.**) For purposes of estimating a protective nutrient criteria exceedance rate, the range of exceedance rates from these site groups needs to be considered as well. Note that an exceedance rate *of as much as* 24% does not result in excess benthic algae in some cases (site 22; **Table A4-1**). On the other hand, notice that an exceedance rate *of as little as* 27.7% can result in non-compliance with the algae standard (site 18; **Table A4-2**). Thus, an exceedance rate around 25% probably represent a threshold; if about 25% of the dataset exceeds the nutrient criteria, then there are roughly equal odds that the site could have nuisance algae (or not). This is partially supported

by the fact that the single site with borderline algae conditions (site 12, Clark Fork River at Bonita; 144 mg Chl *a*/m^2) had a TN exceedance rate of 30.8%.

Conclusion: These analyses show that over a 9 year period (1998-2006) sites on the Clark Fork River that have consistently exceeded the nuisance algae standard (150 mg Chl *a*/m^2, summertime max) have TN and TP exceedance rates with a central tendency around 54%. On the other hand, sites that did not exceed the benthic algae standards had TP and TN exceedance rates with a central tendency around 6%. Within each group (sites that do not exceed algae standards, those that do; **Tables A4-1** and **A4-2**), individual sites had exceedance rates as high as or as low as about 25%. This suggests that 25% may be an exceedance rate threshold where the ability to assure compliance with the algae standard becomes tenuous. Given that about 50% is certainly too high of an exceedance rate and will not protect beneficial uses, approximately 10% is probably too restrictive, and 25% is borderline, it is recommended that a nutrient exceedance rate be set to 20%.

(2.2) 2011 Analysis.

The 12-year (1998-2009) nutrient and algae dataset for the Clark Fork River was very large, and was first reduced prior to statistical analyses. Data reduction followed the following general pattern: At any given site (e.g., CFRPO-12), for any given year (e.g., 2005), and for any given parameter (e.g., TP concentration), the data were reduced to a monthly mean for each summer month (June, July, August, or September). First, quality control duplicates collected on the same day were reduced to a mean (TN data was not analyzed directly until 2009 and so, for 1998-2008 data, TN is the sum of TKN and NO_{2+3} samples collected simultaneously during a sampling event). Next, the mean of all individual days where sampling occurred within a month was calculated, resulting in a monthly mean. Nutrient sampling effort varied considerably from site-to-site and from year-to-year, and we did not want heavily sampled months or years to be over-represented in the dataset in the final analysis. In the manner we reduced the data, therefore, each monthly value carries equivalent weight, with some summer months being better characterized (i.e., sampled more days) than others.

For benthic algae samples, up to 20 spatially-dispersed replicates were collected at a site during any given sampling event. Algae sampling events occurred only once a month. Thus, for a given site/year/month, the benthic algae mean calculated was the value used.

We next determined if each mean nutrient concentration, computed on a month-by-month basis, was above or below the Clark Fork River's applicable standards (TP or TN). This was only carried out for sites and times which had corresponding benthic algae samples. Then, we determined the proportion of months during a summer, at a site, that exceeded the river's nutrient criteria. For example, if a site in 2008 was sampled in June, July, August, and September, and June and August exceeded the TN standard, the TN exceedance rate for summer '08 would be 0.5 (50%). Each exceedance rate was then associated with its corresponding "Max Summer Chl *a*" value (exceedance rate as X, Max Summer Chl *a* as Y). Max Summer Chl *a* is the highest mean monthly Chl *a* value encountered during the summer at a site, per ARM 17.30.631. TN or TP data that were collected *after* the Max Summer Chl *a* event occurred were not included (e.g., if the Max Summer Chl *a* occurred in August, we did not include in the analysis the September TN or TP data for that site/year). Finally, least squares regressions (with 95% confidence intervals) were run for TN exceedance rate vs. Max Summer Chl *a* and TP exceedance rate vs. Max Summer Chl *a,* combining all sites and years together. The results are shown on the next page in **Figure A4-1.**

Regression statistics for both regressions were significant ($p << 0.01$). Using the line equations shown in **Figure A4-1**, 150 mg Chl *a*/m^2 (i.e., the maximum allowable benthic Chl *a* level for a summer; ARM 17.30.631) equates to a 26% exceedance rate of the TN standard and a 31% exceedance of the TP standard. The equivalent exceedance rates corresponding to the upper 95% confidence intervals (which are more conservative) are about 11% and about 5% for TN and TP, respectively.

These Clark Fork River data demonstrate that, across 10 sites with 12 year's worth of monitoring, there is a significant, definable relationship between benthic algal growth and the frequency of exceedance of the river's nutrient standards. That is, sites which frequently exceed the nutrient standards have higher levels of benthic algae. Sites that experience greater than about 25-30% exceedance of the nutrient standards will develop nuisance benthic algal growth, i.e., growth equal to or greater than 150 mg Chl *a*/m^2.

The analytical approach taken in 2008 (**2.1** above) was more coarse than what we have done here, in that it lumped all data by site and then looked to see how often that site—over the long haul—exceeded the nutrient standards. This analysis, in contrast, looks at each site and each summer as an individual event, and then collectively evaluates all the data together, regardless of location along the river (**Figure A4-1**). Interestingly, the overall results between the earlier analysis and the current one are largely the same, in spite of the different analytical approaches. If we continue to assume that the nutrient standards on the Clark Fork River are largely correct in magnitude, then this latest analysis indicates we would want to keep exceedance rates of the applicable nutrient standards between 5-31%, if we want to keep benthic algae below nuisance levels. Since these results correspond nicely to the earlier analysis, we continue to recommend that nutrient criteria exceedance rates be set at 20%.

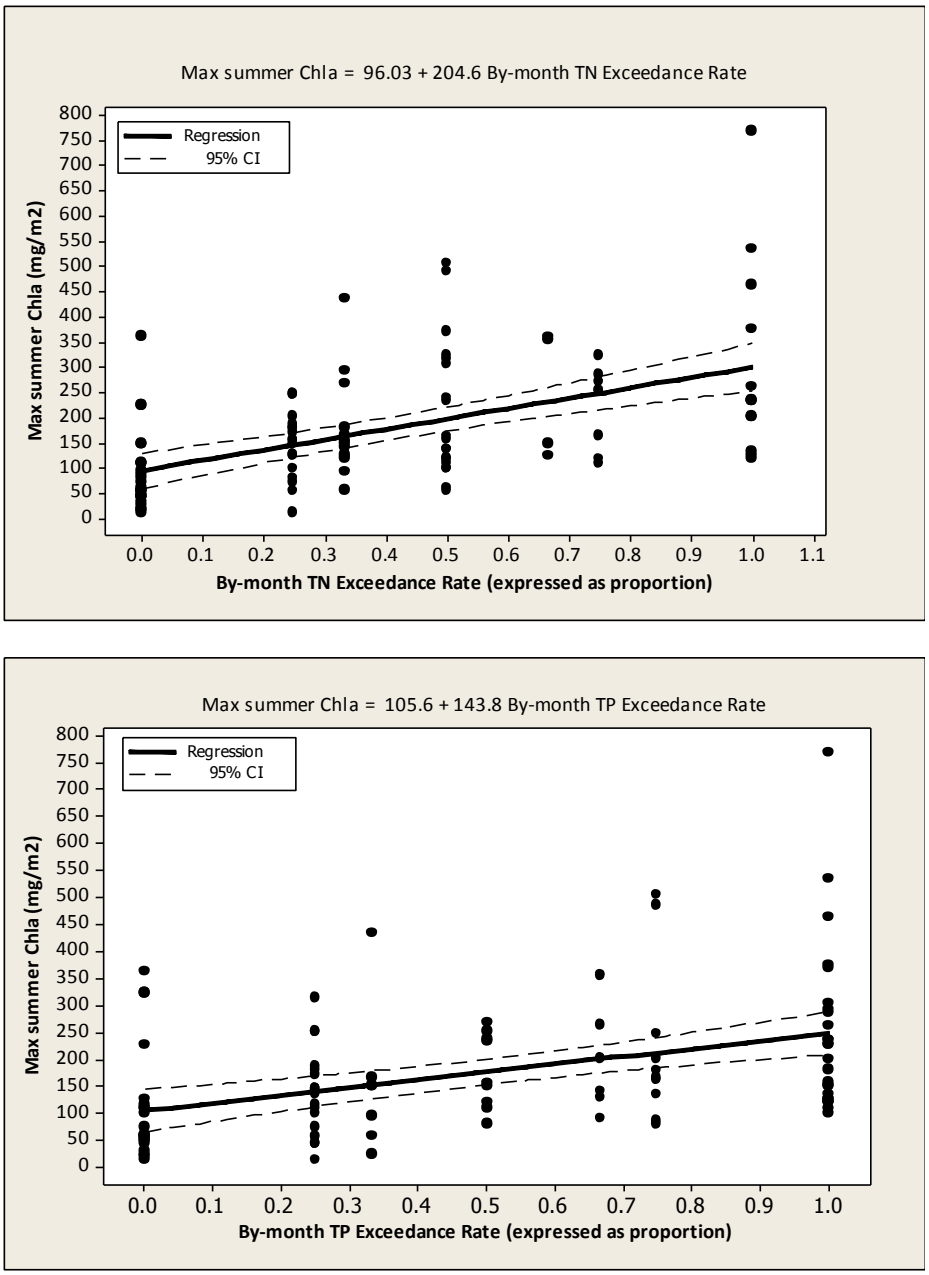


Figure A4-1. Least squares regression for TN exceedance rate vs. Max Summer Chl *a* (upper panel) and TP exceedance rate vs. Max Summer Chl *a* (lower panel), for ten Clark Fork River monitoring sites (1998-2009). Dotted lines are the 95% confidence intervals. Both regressions are significant ($p << 0.01$).

A.5.0 Minimum Number of Nutrient Samples

The final consideration is minimum nutrient sample size. A "nutrient" sample refers to a nutrient type, such as TP or TN Sample sizes apply to each nutrient type, and not to the total number of nutrient samples collected from a stream segment. So, if 7 TN and 7 TP samples were collected from a segment they would not represent 14 samples, but rather 7 of TN and 7 of TP. There is extensive discussion of determining appropriate sample size on a study-by-study basis in USEPA (2002). However, the

recommendations made here for determining compliance with the numeric nutrient standards are meant to apply generally to all Montana wadeable streams, mainly for purposes of 303(d) listing/delisting. *Please note that these sample size minimums do not apply to biological samples (e.g. benthic algae or diatoms) that may be collected concurrently with the nutrient samples.*

For unlisted streams, those for which the form of the null hypothesis is "complies with standard" (**Section A.4.2.1**), the implication for making a type II (beta) error is that a truly non-compliant stream segment would be incorrectly declared compliant. This is a scenario DEQ wants to minimize, and so the probability of such an outcome should be reduced well below 50%, i.e., well below that of a coin flip. The Exact Binomial Test in the accompanying spreadsheet tool can be used to estimate minimal sample sizes. In the test it can be seen that if alpha (type I) error is set to 0.25, exceedance rate (p) to 0.2, and effect size/gray zone (p2) to 0.15 (entered value = 0.35), then a Beta (type II) error of 0.35 is achieved with 12 samples. (Note in the spreadsheet that introducing lower and lower alpha values causes beta error to increase and, therefore, many more samples are needed to try to balance alpha and beta errors.) Twelve samples is about as low an **n** that can be used and still have roughly balanced (0.25 *vs*. 0.35) alpha and beta errors that are each well below 50%.

For listed streams, a similar approach is used. Listed streams are those for which the form of the null hypothesis is "does not comply with standard". In this scenario, the implication of making a type I (alpha) error is that a truly non-compliance stream segment is incorrectly declared compliant; again, this is a scenario. DEQ wants to minimize Setting the alpha to 0.25, exceedance rate (p) to 0.2, and effect size/gray zone (p2) to 0.15 (entered value = 0.05), a beta error of 0.14 (14%) can be achieved with 13 samples. This is a reasonable balance of type I and II errors, and provides a total sampling effort about the same as that for unlisted streams. Given these considerations, it is recommended that:

- For new, unlisted stream segments, a minimum of 12 independent samples for any given nutrient be collected for compliance determination.

- For 303(d)-listed stream segments that already have one 1 nutrient criteria exceedance for a given nutrient, a minimum of 13 independent samples (this total can include newly collected as well as previously collected samples) should be used for compliance determination.

- For listed streams with 13-18 total samples that already have 2 or more exceedances for a given nutrient, the default conclusion is that the stream segment has failed the Exact Binomial Test (no further sampling required at this time). Run the t-test as well and incorporate results in decision matrix.

- For listed streams that have > 18 samples for a given nutrient, set alpha to 0.25, exceedance rate to 0.2 and effect size to 0.15 in the Exact Binomial Test, and determine if the reach is (or is not) in compliance with the Exact Binomial Test. Carry out the same for the T-test.

- If a very large dataset (> 300 samples) is available for a particular stream, then lower type I (alpha) and type II (beta) error can be achieved with higher confidence in the results. Use the special feature of the Exact Binomial Test to help define these confidence levels. Confer with Standards Section if needed.

APPENDIX B. DETAILS ON ASSESSMENT METHODOLOGIES FOR WADEABLE MOUNTAIN AND TRANSITIONAL STREAMS

The following provides the rationales used for the selection of the assessment tools. It also provides the rationales for selected impact thresholds. This information is summarized in **Section 4.0** of the main document.

B.1.0 ASSESSMENT OF BENTHIC ALGAL GROWTH

An evaluation of statistical uncertainty in the site averages calculated using DEQ's standard procedure for collecting and analyzing benthic Chl *a* is detailed in **Appendix A** of the Chl *a* SOP (Montana Department of Environmental Quality, 2011c). To summarize: if a benthic Chl *a* sampling event has followed the SOP, DEQ is confident that—for a typical wadeable stream— at least 80% of the time the measured Chl *a* average calculated will be within ±30% of the true average. Given this known variability, decision points pertaining to benthic algae growth and harm-to-uses have been developed, and are further detailed in **Sections B.1.1** and **B.1.2** below

B.1.1 BENTHIC ALGAL CHL *A* LEVELS AND THE RECREATION USE

It is reasonable that, once a site's true average algal level exceeds about 150 mg Chl *a*/m^2, impairment to the recreational use has occurred. This is shown in **Figure B1-1**. But we also have to account for the uncertainty around the Chl *a* measurement. Shown are the three photographs that bracket the acceptable-unacceptable threshold, per Suplee *et al.* (2009) each with their interval widths (based on the 20 Chl *a* replicates associated with each photo) calculated at the 90% confidence level. Once algae levels have reached the lower bound of photo E (photo E's lower confidence bound = 169 mg Chl *a*/m^2), the acceptability threshold has already been exceeded. This is because the public majority finds the algae level shown in the photo to be highly undesirable. The gray zone in **Figure B1-1** represents the zone where public acceptability rapidly transitions from "OK" to "Not OK". Going forward, any measured average Chl *a* value DEQ believes could plausibly be as high as 165 mg Chl *a*/m^2 (in the gray zone, and at upper confidence bound of photo F, but still below the lower confidence bound of photo E) should be considered an exceedance.

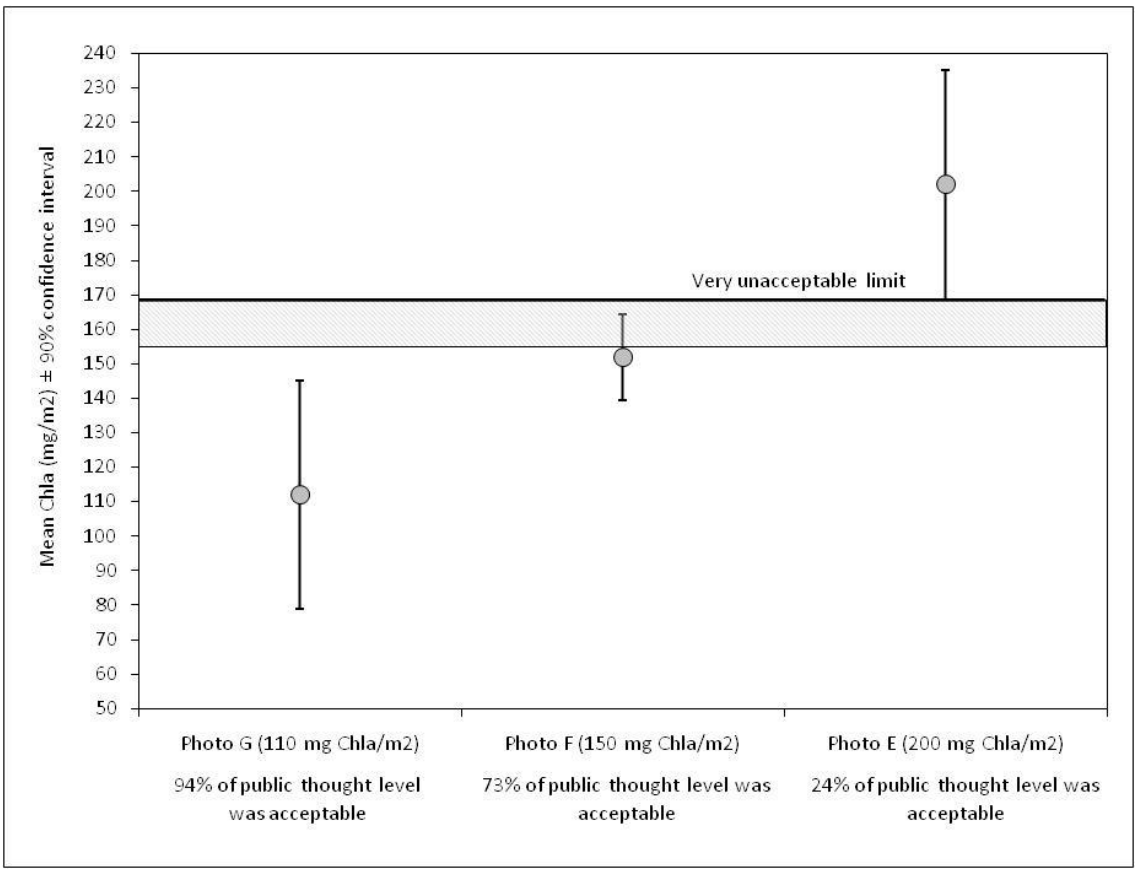


Figure B1-1 Averages (Gray Dots) and 90% Confidence Bands for the Three Photos Bracketing the Acceptable-unacceptable Threshold, per Suplee et al. (2009) The gray band shows the algae level range across which public opinion rapidly shifts from acceptable to unacceptable.

Returning to DEQ's algae-sampling protocol, the average that is calculated for any given sampling event has a definable interval width and, when the upper bound of that interval reaches about 165 mg Chl *a*/m^2, an impairment at that site should be considered to have occurred. Using the approach outlined in Appendix A of DEQ (Montana Department of Environmental Quality, 2011c) and given that ***n*** = 11, Coefficient of Variation (CV) = 73%, DEQ can be 80% certain that the true benthic algae average may be as high as 165 mg Chl *a*/m^2 when the measured average is ≥ 129 mg Chl *a*/m^2. Therefore, any sampling event for which the measured benthic algal Chl *a* average is **≥ 129 mg Chl *a*/m^2** should be considered an exceedance, and in violation of ARM 17.30.637(1)(e).

B.1.2 Benthic Algal Chl *a* Levels Relative to Late-season Dissolved Oxygen Problems, and Potential Impacts to Fish and Associated Aquatic Life

In 2009, DEQ commenced a BACIP (Before After Control Impact Paired) design, whole-stream nutrient addition study to better understand the exact way stream changes are manifested due to nitrogen and phosphorus pollution. We wanted to understand the relationship between these changes and stream

beneficial uses (Montana Department of Environmental Quality, 2010a)[16]. In summer 2010 soluble nitrogen and phosphorus were added to a reach (High Dose reach) of the stream, and major changes occurred as a result. One of the most interesting findings was the temporal manner in which stream dissolved oxygen (DO) problems occurred, and the relationship of those DO levels to measured benthic algae levels.

After nutrient additions began in early August 2010 and then continued throughout summer 2010, DO standards that protect fish (1 Day minima in DEQ-7;(Montana Department of Environmental Quality, 2010b)) were never exceeded — nor even approached —in the High Dose reach. This was true in spite of large daily DO swings (**Figure B1-2**)[17]. Relative to the upstream Control reach, DO increased dramatically during the day, but did not at night fall much below the Control reach values (**Figure B1-2**). Benthic algal production (growth) exceeded respiration throughout most of the summer and this, in conjunction with adequate re-aeration due to the stream's flow, likely prevented nighttime DO levels from dropping below the DO standards. However, in fall, the growing season's accumulated algal growth began to senesce en masse and we observed large amounts of decaying algae on the stream bottom in early October. The decaying algae induced a high oxygen demand which was concentrated near the bottom, in affect acting like a sediment oxygen demand (SOD), and which in turn led to exceedances of the DO standards. (The YSI in **Figure B1-2** was monitoring DO about 20 cm off the stream bottom, in a run.) The DO standards exceedances all occurred late in the season, after robust benthic algae growth had ended.

It should be noted that two other YSIs (one deployed upstream and another further downstream of the one in **Figure B1-2**, thus bracketing the High Dose reach) simultaneously recorded DO concentrations, none of which violated DO standards, even in October (data not shown). This was apparently due to longitudinal changes in stream morphology (e.g., width/depth relationships) and their affect on stream re-aeration, and dead algae accumulation on the bottom. We calculated that for DO to decline from what was measured by the YSI just upstream of the High Dose reach to that recorded in early October in **Figure B1-2** would require a SOD higher than any we could locate in the literature[18]. This suggests that DO concentrations were not uniform from stream surface to bottom, but rather, a bottom-to-surface DO gradient likely existed. Our analyses further indicated that DO was probably near zero near the bottom, and then near saturation at the surface. These findings suggest that DO problems of this nature can be both longitudinally and vertically patchy along the stream channel.

---

[16] Although this study was carried out in a C-3 warm water fishery stream, the stream's key characteristics relative to algal growth make it a reasonable comparison to western Montana gravel-bottom streams, as we have done here. The stream has a gravel dominated substrate, perennial flow of about 4-6 CFS in summer, a water surface slope of 0.4%, and riffles are common throughout, as are pools. The dominant filamentous algae that grew during the study was *Cladophora*, which is also commonly found in western MT streams.

[17] Routine QC checks of the instrument, including mid-project calibration, routine instrument cleaning, etc. was undertaken The low DO values measured are considered valid measurements.

[18] Stream water biochemical oxygen demand (BOD) samples were also collected in early October at the site. They were non-detect, thus the DO consumption had to be coming from the decaying algal material on the bottom.

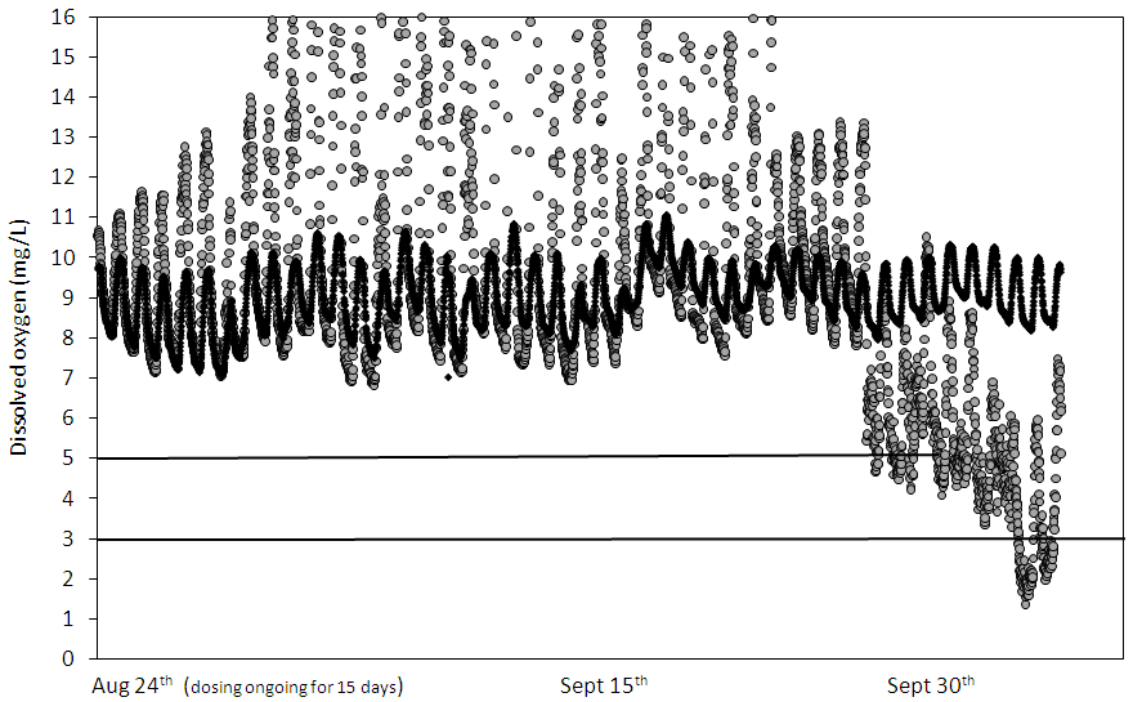


Figure B1-2. Temporal Changes in Dissolved Oxygen (DO) Levels in the High Dose Study Reach as Measured by Deployed YSI Instrument, 2010. Gray dots are DO observations at the High Dose Reach, collected at 15 min intervals. For comparison, the black line oscillating around 9 mg/L is the DO measured by another YSI in the upstream Control reach, which did not receive nutrient additions. The horizontal black lines are (upper) the adult fish and (lower) juvenile fish DO standards for this stream.

Returning to the High Dose reach's benthic-algal growth, benthic algae Chl *a* levels at the end of the growing season reached 127 mg Chl *a*/m^2 (**Table B1-1**). It is apparent from **Figure B1-2** that this level of benthic algae was sufficient to induce DO violations along the channel when the algae died and decomposed en masse. The implication of this finding is that the late-season average benthic algae we measured (127 mg Chl *a*/m^2) has the potential, in wadeable streams, to cause DO standards exceedances in the post-summer period, probably in late September or October. Although this study was carried out in a C-3 warm water fishery stream, the stream's late season water temperatures (which ranged from about 12-16 ^{0}C) are comparable to what is observed in typical western Montana gravel bottom streams at that time of the year. While it is true that water temperature strongly affects DO concentration, we would not expect western Montana streams manifesting similar algal densities to be able to compensate (i.e., maintain DO above standards) due to their having cooler water temperatures, as their temperatures are often about the same at that time of the year.

Table B1-1. Benthic Algae Density Measured at the High Dose Study Site, 2010

Sampling Date	Reach average benthic-algae density (mg Chl *a*/m^2)	Chl *a* replicates' CV(%)*	Reach average AFDW density (g/m^2)	AFDW replicates' CV(%)*
August 26, 2010	**111**	123	**26**	93
September 8, 2010	**116**	97	**34**	45
September 22, 2010	**87**	82	**37**	61
October 6, 2010	**127**	90	**33**	73

* 11 replicates were collected for each sampling event. Less than 11 were used to calculate reach average AFDW because core samples, if collected, are not included in the calculation of average reachwide AFDW density.

As mention at the beginning of **Section B.1**, there is a definable level of uncertainty around any given benthic Chl *a* average, but there is no way to know if the values in **Table B1-1** are at the low end, high end, or in the middle of that range. We'll assume here that the 127 mg Chl *a*/m^2 measured is, in fact, accurate. Thus, to be protective and assure that DO problems that could harm fish and associated aquatic life are prevented from occurring, we recommend that when a site's average benthic Chl *a* exceeds **120 mg Chl *a*/m^2** it is too high, and should therefore be considered an impact to fish and associated aquatic life.

B.1.3 Benthic Algal AFDW Levels and Harm-to-use Thresholds

Ash Free Dry Weight (AFDW) collected from natural stream-sediment surfaces is a useful measurement for estimating algal biomass. The laboratory method basically oxidizes and reports back the mass of all organic material in the sample (American Public Health Association, 1998). It is useful in that it provides an additional means of assessing accumulated algal biomass independent of Chl *a.* Chl *a* levels tend to be highest during peak growth, and then decline later as the Chl *a* molecules degrade as the algae senesce (Stevenson, et al., 1996). If an assessor samples a stream late in the season, they may find fairly low Chl *a* values in spite of the presence of a large biovolume of algal material. Thus, a site that may truly have an excess algae problem could potentially be assessed as unimpaired simply due to the fact that the samples were collected late in the season.

For this reason, we recommend that AFDW be determined for all samples when Chl *a* is collected. AFDW can be determined from the same sample in a subsequent analysis that follows the Chl *a* analysis Site average. AFDW can be determined from individual replicates, or as a weighted average.

Note: AFDW results from core samples should <u>never</u> be included in determining a site's average AFDW. The method measures organic material from the entire core sample, not just the surface where the algae are growing, and will therefore over-report AFDW.

DEQ has not collected AFDW using the 11-transect method long enough to be able to carry out the type of statistical uncertainty calculations used for Chl *a* (Appendix A of (Montana Department of Environmental Quality, 2011c). However, there are good estimates of what comprises too much algal AFDW. In Suplee et al. (2009), the threshold Chl *a* level of 150 mg/m^2 corresponds to 36 g AFDW/m^2 In New Zealand, extensive analysis of algal AFDW resulted in a recommendation of 35 g AFDW/m^2 as the maximum level for gravel/cobble streams, to protect recreation use (Biggs, 2000). Note in **Table B1-1** above that the late season AFDW corresponding to 127 mg Chl *a*/m^2 (the Chl *a* level linked to the late-season DO problems) is 33 g/m^2 Long -term monitoring in the Clark Fork River (1998-2009) shows that the average summer AFDW at sites that do not develop nuisance algae (i.e., they are consistently <150 mg Chl *a*/m^2) ranged from 17 to 48 g AFDW/m^2 (mean: 27 g AFDW/m^2). Given the values presented, we recommend that site average AFDW (i.e., mean of the 11 replicates collected at a site, replicates being only templates or hoops) should be no greater than **35 g AFDW/m^2.** This value should be protective of both fish and aquatic life and recreation uses.

B.1.4 Some Additional Considerations Regarding Benthic Algae Sampling

Recently, DEQ has instituted an economization practice that consolidates all hoop, core, or template samples from a sampling event together, so that only three (at most) Chl *a* samples need to be analyzed,

instead of eleven. While unquestionably thrifty, the ability to determine the replicates' variance has been lost. All of the Chl *a* confidence calculations discussed in **Section B.1.1** assume that a sampling event will manifest a typical replicate CV of 73%, but this is an assumption In cases where it is very important to truly know the replicates' CV, the replicates should each be analyzed separately..

Cases may also arise where an entity is not satisfied with the level of confidence or interval widths DEQ has presented here. Collecting 11 samples in a stream reach is already a time consuming and expensive procedure, and we consider the confidence level (80%) and interval width (± 30% of the mean) to be satisfactory for algae sampling. If an entity (regulated or otherwise) desires higher levels of precision, than it is our recommendation that the financial cost to achieve those levels fall to the entity.

If more precision is wanted, how many more algae samples should be collected? Long term sampling of benthic Chl *a* by Dr. Vicki Watson on the Clark Fork River shows that with about 20 replicates, one can be 90% confident that the measured average Chl *a* is within ± 20% of the true average. For benthic algae sampling, which is inherently noisy, this is a fairly high degree of confidence. For DEQ's wadeable stream method, this would involve placing 20 transects instead of 11 along a site, with algae collection occurring at each of the 20 transects using the systematic approach (R, L, C, repeat) described in the SOP.

B.2 Assessment Using Biometrics

DEQ has used diatom-algae assemblages and macroinvertebrate assemblages for many years to make assessment of stream water quality and condition. Some of these metrics are being incorporated into the process for assessing excess nitrogen and phosphorus pollution. Details of each are given in the **Sections B.2.1** and **B.2.2** below.

B.2.1 Biometrics Based on Diatom Algae

DEQ has been using benthic diatoms to assess water quality since the 1970s. Earlier approaches used diagnostic and descriptive biometrics based on quasi-universal ecological attributes of diatom species and observed structural characteristics of benthic diatom associations (Bahls, et al., 2008). The current approach (initiated in 2004) uses regional classification, stream reference sites, a priori knowledge of stressors in streams, and discriminant function analysis to identify "increaser" taxa that respond to specific stressors and in a predictable way (Teply and Bahls, 2006; Bahls, et al., 2008; Tepley and Bahls, 2005). The metrics were specifically developed to indicate the likelihood of nitrogen and phosphorus impairment, have been developed for many regions of the state, and can function properly in the presence of other major pollutants (Teply, 2010a; Teply, 2010b). Please see the periphyton SOP (Montana Department of Environmental Quality, 2011b) for details. Each sample will provide the probability of a nutrient problem, such as in this example:

> *This indicates that the sample represents a stream that has about a 65% percent probability of being impaired due to nutrients (nitrogen or phosphorus) under 303(d) guidelines. This probability is based on past evidence of taxa associated with nutrient-impaired streams in the Northern/Canadian Rockies Stream Group. Nutrient Increaser Taxa do not discriminate other causes of impairment and this result does not*

indicate whether the stream may or may not be impaired due to other causes.

Diatom nutrient-increaser metrics are available for the Northern and Canadian Rockies ecoregions, the Idaho Batholith ecoregion, and a series of level-IV ecoregions that predominate along the Rocky Mountain Front (i.e., mountain-to-plains transitional zones). **Note: There is currently no validated nutrient-increaser model for use in the Middle Rockies ecoregion.** As of this writing, a sample that indicates **>51%** probability of impairment by nutrients should be considered to indicate the sample is from a site with excess nutrient problems. Findings based on diatom samples are not, however, stand alone, and need to be incorporated with other data per the decision framework described in **Section 3.0** of the main document.

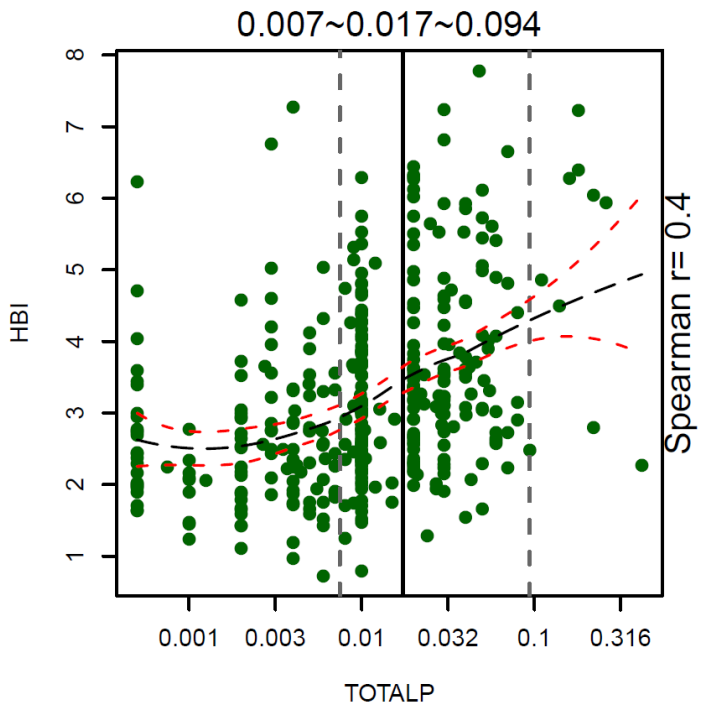
B.2.2 Biometrics Based on Macroinvertebrates

Benthic macroinvertebrates have been used for a long time as indicators of stream water quality (e.g., (Hilsenhoff, 1987; Barbour, et al., 1999). Recently, DEQ and EPA carried out a correlation analysis using Montana data to examine the relationship between stream nutrient concentrations and benthic macroinvertebrate metrics (Tetra Tech Inc., 2010). Among the metrics, one (Hilsenhoff Biotic Index, or HBI) is sufficiently well understood and showed a sufficiently patterned response to stream nutrient gradients in Montana's mountainous regions that we believe it can be used as a secondary response variable to help assess nutrient impacts. How the metric will be incorporated with other effect variables was discussed in **Section 3.0.** We here define a biological threshold for the HBI metric, giving consideration to the fact that almost all mountainous streams in Montana are to be maintained suitable for "growth and propagation of salmonid fishes and associated aquatic life" (A-Closed, A-1, B-1, C-1 classes), or "growth and marginal propagation of salmonid fishes and associated aquatic life" (B-2, C-2 classes).

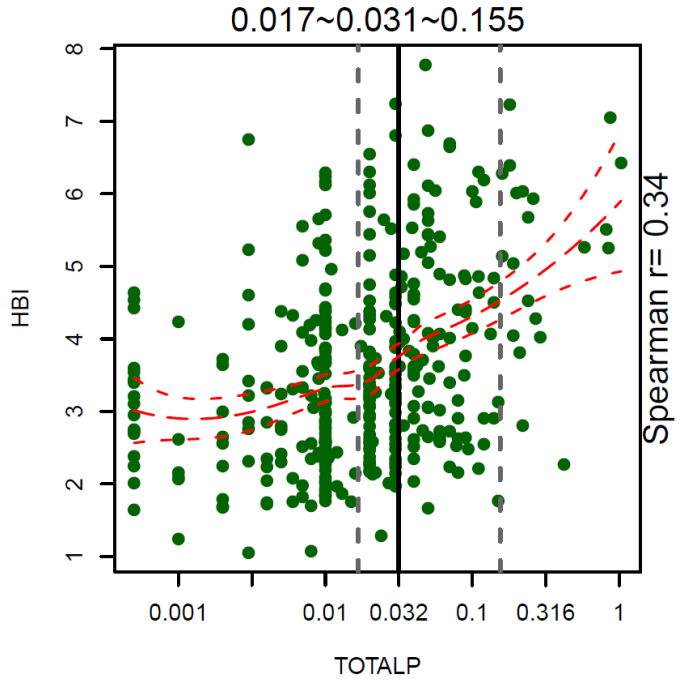
B.2.2.1 Hilsenhoff Biotic Index

HBI is based on tolerance values. A large number of macroinvertebrate taxa have been assigned a numeric value which represents the organism's tolerance to organic pollution (Barbour et al., 1999). HBI is then calculated as a weighted average tolerance value of all individuals in a sample. Higher index values indicate increasing tolerance to pollution.

Figure B2-1(A) shows the HBI *vs.* TP correlation in mountainous-region streams (Tetra Tech Inc., 2010). The data are from the "Mountains" site class (a.k.a. Mountains bioregion)(Montana Department of Environmental Quality, 2006). The Mountains Bioregion comprises stream sites whose catchments are mainly in the Middle Rockies, Canadian Rockies, Northern Rockies, and Idaho Batholith ecoregions and where elevation is greater than 1700 m, precipitation is greater than 700 mm/year, and annual mean daily maximum temperature is < 11°C. Also shown is the same data, but this time aggregated simply by level III ecoregion rather than bioregion (**Figure B2-1(B)**); note the very similar patterns. This indicates that ecoregions and bioregions work about equally well as geospatial frameworks to segregate macroinvertebrate data for the purpose of correlation to stream nutrient concentrations.



A. Mountains bioregion.



B Middle Rockies ecoregion.

Figure B2-1. HBI metric vs. TP.

The relationship between HBI and TN is shown below (**Figure B2-2**), aggregated by ecoregions (only Middle Rockies ecoregion shown). As for TP *vs*. HBI, there is a noisy but discernable (and significant) relationship between nutrients and the HBI score.

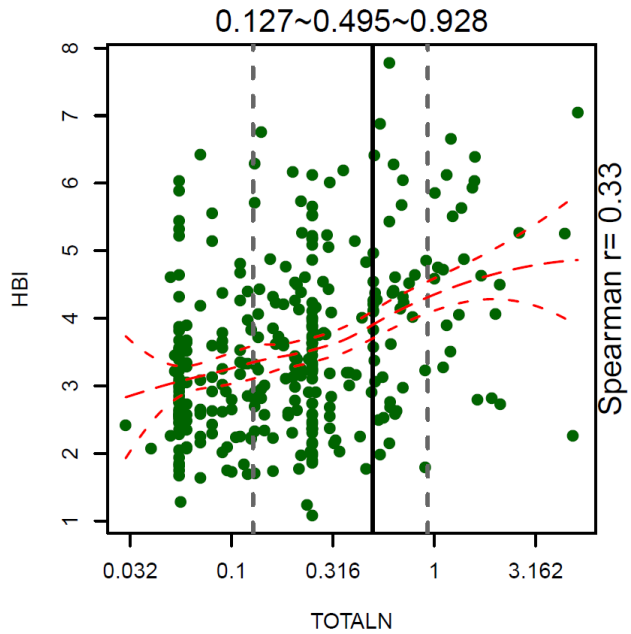


Figure B2-2. HBI Macroinvertebrate Metric vs. Stream TN Concentration Data are aggregated by ecoregions (Middle Rockies ecoregion shown).

Several components of **Figures B2-1 (A) and (B)** and **Figure B2-2** require explanation. Change-point analysis (Qian, et al., 2003) shows the statistically-derived point in the dataset where a shift, or threshold, in Y has occurred relative to X. In the figures, it is the vertical black line, bracketed to the right and left by its 90% confidence interval as dashed grey lines. This change-point, ± its 90% confidence interval, is shown as a numeric value at the top of each figure. The curving dashed lines running left to right are the locally-weighted regression line (LOWESS) and associated 90% confidence limit as dashed red lines. The Spearman's *rho* correlation value (Conover, 1999) between nutrient and metric is shown on the right side of each figure.

B.2.2.2 Interpreting the Macroinvertebrate Metric Correlations to Nutrients

Although **Figure B2-1** and **Figure B2-2** demonstrate significant correlations (parametric least-squares regression), they show large amounts of scatter. Numerous factors contribute to this scatter. For example, macroinvertebrates are separated from the direct effects of nutrient increases by one trophic level (i.e., nutrients directly influence aquatic plants and algae, and changes in plant species and biomass in turn influences macroinvertebrates). This is why they are considered secondary data, per the **Section 3.0** decision framework. In addition, environmental factors (natural and human-caused) other than nutrients influence macroinvertebrate populations, adding to the scatter in the relationships between the metrics and the nutrients. This is especially true for these data, which have been compiled over relatively large spatial areas and incorporate many different streams sampled over a long period of time (> 15 years). In spite of the scatter, there are patterns that can be discerned. Note, for example, that when TP is greater than 0.15 mg/L in **Figure B2-1(B)**, the likelihood of a stream having an HBI score <4 is very low.

The following section provides more detail on how **Figure B2-1** and **Figure B2-2** (and related data) were used to interpret the macroinvetebrate metrics relative to nutrients.

B.2.2.3 Patterns Observed Between HBI and Change-points, and Previously Used DEQ Thresholds

DEQ has in the past used thresholds for stream impairment using the HBI metric (Bukantis, 1998). Mountainous and intermountain valley regions used HBI values of 3.0 and 4.0, respectively, as the threshold values between full support and impairment of aquatic life (Bukantis, 1998). Hilsenhoff (1987) notes that the transition between Very Good water quality (slight organic pollution) and Good water quality (some organic pollution) is an HBI score of 4.5. How do the thresholds relate to the data seen in **Figures B2-1** and **Figure B2-2**? Results are shown below in **Table B2-1,** and include data from the Middle Rockies, the Mountains Bioregion, and the Low Valleys Bioregion.

Table B2-1. Nutrient Concentrations on the LOWESS Regression Line Corresponding to Specified Thresholds.
(Statistical Changepoint or Macroinvertebrate Metric Value) for **Figures B2-1, B2-2**, and Similar Graphs. Data are grouped by bioregions and ecoregions. Only Middle Rockies ecoregion shown due to insufficient data on other ecoregions.

			Corresponding Nutrient Concentration	
MT Bioregion	**Parameter**	**Metric Threshold Value**	**TP (mg/L)**	**TN (mg/l)**
Mountains bioregion	**Changepoint**		**0.02**	**0.46**
Mountains bioregion	HBI	4.0	0.07	1.15
Mountains bioregion	HBI	4.5	0.19	beyond graph
Low Valleys bioregion	**Changepoint**		**0.04**	**0.32**
Low Valleys bioregion	HBI	4.0	0.04	0.32
Low Valleys bioregion	HBI	4.5	0.19	1.35
Middle Rockies Ecoregion	**Changepoint**		**0.03**	**0.50**
Middle Rockies Ecoregion	HBI	4.0	0.04	0.55
Middle Rockies Ecoregion	HBI	4.5	0.16	1.7

One observation that can be made about **Table B2-1** is that, in any given region, HBI scores of 4.5 can correspond to nutrient concentrations much higher than the corresponding changepoint concentrations, whereas nutrient concentrations matching an HBI of 4.0 usually match fairly closely to the changepoint concentrations. These results suggest that an HBI score of 4.0 is a meaningful threshold relative to stream nutrient concentrations in western Montana streams. This conclusion stems from the fact that an HBI of 4.0 has previously been recommended as a threshold by DEQ for this region (Bukantis 1998), and that the data show a statistically-significant threshold (i.e., a change in biological structure relative to nutrients) at an HBI of 4.0. An HBI score of 4.0 is also meaningful from the perspective of water quality protection, as scores of 4.5 indicates transition into conditions where some organic pollution is already noted (Hilsenhoff, 1987). Thus, if elevated nutrients are suspected and an HBI score of >4 is encountered, there is a good chance that excess nutrients are causing the problem.

APPENDIX C. DETAILS ON ASSESSMENT METHODOLOGIES FOR WADEABLE PLAINS STREAMS

Benthic algae levels discussed in **Appendix B**, **Section B.1.1** are appropriate for the mountainous and transitional (mountain-to-plains) region of the western part of the state. Many eastern Montana plains streams are ecologically different from western Montana streams, and therefore the results from the public perception algae survey (Suplee et al., 2009) should probably not be universally applied to them. Montana plains streams often become intermittent, are generally low gradient, commonly have mud bottoms and are often turbid, and frequently have substantial macrophyte populations. It is not uncommon in these streams to see macrophytes intermixed with filamentous algae and floating masses of green algae; these conditions are even occasionally observed in plains streams minimally impacted by people (i.e., plains reference streams). These situations make measurement of benthic algal biomass in plains streams difficult and complicated. Further, our analysis shows that 4% of the sampling-event averages from plains reference streams have benthic algae >150 mg Chl *a*/m^2, whereas none of the sampling-event averages for benthic algae in western Montana reference streams even approach this value (e.g., the highest sampling-event average for a western Montana reference site was 76 mg Chl *a*/m^2). These findings, taken together, suggest that benthic algae measurement is probably not the best assessment tool for determining impairment for plains region streams. As such, we recommend that benthic algal biomass not be used to assess plains streams.

The following sections discuss the assessment tools we believe are more appropriate for assessing wadeable streams of the plains region.

C.1.0 ASSESSMENT USING DIATOM ALGAE BIOMETRICS

Nutrient-increaser diatom metrics have been developed for plains wadeable streams using the same methods as the diatom metrics presented in **Appendix B.** The metrics will indicate the likelihood of nitrogen and phosphorus impairment, and can function properly in the presence of other common pollutants such as sediment and metals (Teply, 2010a; Teply, 2010b). Please see DEQ's periphyton SOP (Montana Department of Environmental Quality, 2011b) for details. Each periphyton sample will provide the probability of a nutrient problem, such as in this example:

> *This indicates that the sample represents a stream that has about a 25% percent probability of being impaired due to nutrients (nitrogen or phosphorus) under 303(d) guidelines. This probability is based on past evidence of taxa associated with nutrient-impaired streams in the Warm-water Stream Group. Nutrient Increaser Taxa do not discriminate other causes of impairment and this result does not indicate whether the stream may or may not be impaired due to other causes.*

As of this writing, a sample that indicates **>51%** probability of impairment by nutrients should be considered to indicate the sample is from a site with excess nutrient problems. Findings based on diatom samples are not, however, stand alone, and need to be incorporated with other data per the decision framework described in **Section 3.0** of the main document.

C.2.0 Assessment Using the Difference Between the Daily Maximum Dissolved Oxygen Concentration and the Daily Minimum Dissolved Oxygen Concentration (Delta)

We initially considered using DEQ's DO standards for routine assessment of plains streams. But examination of long-term DO datasets, including those from the plains-stream dosing study (**Appendix B, Section B.1.2**), showed that DO standards are a fairly insensitive way to assess nutrient impacts in plains streams. We found only a low number of instances where streams that we know have excess nutrient impacts consistently violated the DO standards, and some nutrient-impacted stream sites never violated the DO standards at all (at least during summer and early fall).

We found that streams that have high daily DO **delta** (i.e., the daily maximum DO minus the daily minimum DO) may *eventually* manifest DO standards violations late in the year, when the algae die and decompose en masse (see **Appendix B**, **Section B.1.2**). To address this, the DO monitoring-period could be extended (e.g., to the end of October or early November), but this is not always practical given the unpredictable onset of winter and its affect on road access, retrieval of deployed instruments, etc. In lieu of extending the monitoring season, measurement of summer and early fall DO deltas has great potential as an assessment tool. Others have found that DO delta is related to harm to aquatic life. In Minnesota, strong positive correlations are found between the percent tolerant fish and the magnitude of the DO deltas. At DO deltas <4.5 mg/L, tolerant fish are usually <10% of the total fish population, but when DO deltas are > 4.5 mg/L tolerant fish become a substantial proportion of the population. Conversely, sensitive fish exhibit a wide range of values at DO deltas <4 mg/L, but above 4.5 mg/L they decline to 10% or less of the fish population(Minnesota Pollution Control Agency, 2010). The state of Minnesota is recommending that, for the northern plains regions at its southern end, measured DO deltas should not exceed 5.0 mg/L.

DO delta is also shown to be lower in reference streams. In Tennessee, the maximum DO delta value reported in a wadeable reference stream is 4.0 mg/L, whereas about 45% of impacted streams assessed have measured DO deltas greater than 4.0 mg/L (Arnwine and Sparks, 2003).

We calculated DO deltas for Montana plains <u>reference</u> streams. There were a total of 177 day's worth of delta values from the Box Elder Creek Control reach (Montana Department of Environmental Quality, 2010a) and the Little Beaver Creek Reference Site, collected in 2009 and 2010. 90% of the daily DO deltas from these reference sites were less than 5.3 mg/L. The single highest DO delta measured was 6.6 mg/L, from Little Beaver Cr, which is influenced by the presence of macrophytes. (Also, on about 10% of occasions, DO at the Little Beaver Cr reference site dropped just below the juvenile fish standard of 5.0 mg DO/L; however, it never got close to the adult fish DO standard of 3.0 mg/L.)

No fish data were collected contemporaneously with the reference site DO data, however fish populations have been evaluated in both of these streams (Bramblett, et al., 2005) at alternative reference sites (BoxElder_382_W and LittleBe_410_W) not far downstream. These alternative reference sites had among their fish populations substantial proportions of sensitive/intolerant species, especially Box Elder Creek in the Little Beaver Creek site 17% of the fish captured were considered sensitive/intolerant. Sensitive/intolerant species are typically the first species to disappear due to chemical and physical perturbations (e.g., low DO)(Barbour, et al., 1999). Assuming DO patterns at the alternative reference sites are roughly comparable to those which we monitored, the fish data suggest

that a healthy fishery is being maintained in spite of occasionally high DO deltas and occasional exceedances of the juvenile fish criterion[19].

We employed change-point analysis (Qian, et al., 2003) to help identify any DO delta thresholds. We had 11 plains-stream locations, in both reference (Suplee, et al., 2005) and non-reference condition, which had continuous instrument-measured DO data (**Table C2-1**). These sites comprised both perennial and intermittent streams. Delta values were calculated, resulting in over 550 days of DO delta values. Each location was assigned a rating (1 through 4) representing our BPJ assessment of how strongly it was impacted by nutrients, and these ratings were associated with the corresponding DO delta values. The ratings used were: 1 = no known nutrient impact; 2 = low nutrient impact; 3 = medium nutrient impact; 4 = high nutrient impact. In quite a few cases the ratings could be very accurately assigned, as some sites were reference sites and some were part of the nutrient dosing study (e.g., all DO deltas associated with the High-dose reach were assigned a rating of 4). Change-point analysis was then run on the dataset with delta values on the X axis and their corresponding rating scores on the Y A highly significant ($p << 0.001$) change-point was identified at 6.0 mg/L (the 90% confidence interval for the change-point was 5.5 mg/L to 6.6 mg/L). Essentially, analysis showed that in moving from sites rated 3 to sites rated 4, the magnitude of the DO deltas ramped up dramatically, with the threshold of this change occurring at 6.0 mg/L.

Table C2-1. Long-term DO Monitoring Sites in Plains Streams

Station ID	Continuous Data Time Range	DO observations time-step (min)
Y26BOXEC08-upstream	Aug 25 to Sept 30, 2010	15
Y26BOXEC08-downstream	Aug 24 to Sept 30, 2010	15
Y26BOXEC04	July 26 to Sept 26, 2009 AND July 19 to Oct 7, 2010	15
Y26BOXEC09-upstream	Aug 25 to Sept 30, 2010	15
Y26BOXEC09-downstream	Aug 11 to Sept 30, 2010	15
Y27LBVRC02	Aug 30 to Sept 25, 2008 AND Aug 29 to Oct 8, 2010	15
Y27LBVRC04	Aug 30 to Sept 24, 2008 AND July 29 to Oct 8, 2010	15
M22CTWDC03	July 22 to July 24, 2003	30
M22BSPRC10	Aug 17 to Aug 20, 2003	30
Y27LBVRC12	Aug 30 to Sept 25, 2008	15
Y27LBVRC01	July 28 to Sept 24, 2009	15

We then estimated false-positive and false-negative rates, and made comparisons to the reference data, using the datasets above. Data were aggregated to create two basic groups (ratings 1, 2 = nutrient un-impacted; ratings 3, 4 = nutrient impacted), and 65 observations were then randomly drawn for false positive/negative analysis (35 from the un-impacted group, 30 from impacted group). Results are shown in **C2-2** below.

[19] Dr. Robert Bramblett (MSU fishery biologist; personal communication, March 11, 2010) provided species counts and IBI scores for the two sites Box Elder Creek's score was quite good (77) Dr. Bramblett noted that the Little Beaver Creek site's IBI was being reduced (score 55) by the presence of northern pike; this non-native predatory fish plays a large role in reducing metric scores in the Bramblett IBI He noted that the site's habitat was simple and northern pike were crowded in with their prey, but otherwise, the Little Beaver Creek site seemed healthy.

Table C2-2. Statistics Associated with DO Delta Threshold

DO Delta Threshold	% of all reference-site deltas > threshold	Estimated false positive rate*	Estimated false negative rate†
≥ 6.0	3%	9%	77%
≥ 5.3	**10%**	**23%**	**63%**
≥ 5.0	15%	26%	60%
≥ 4.0	26%	54%	53%

* The probability that a truly un-impacted site is found to have a DO delta value great than the threshold.
† The probability that a truly impacted site is found to have a DO delta value less than the threshold

The change-point threshold (DO delta of 6.0 mg/L) is probably too high for assessment, as it has a particularly high false-negative rate (77%; **Table C2-2**). Other DO thresholds between 6.0 and 4.0 mg/L were also evaluated. A DO delta threshold of 4.0 mg/L provides good balance between alpha and beta error, but its ability to determine impact is no better than a coin flip. It also allows far too many of the deltas from the reference sites to be exceedances (in fact, almost all observations from one reference site are >4.0 mg/L). The reality is, sites with excess nutrient problems do not manifest high DO deltas every single day throughout the summer, due to the vagaries of clouds, weather, and wind, which means that as an assessment tool DO delta will inherently have high false-negative rates. We selected 5.3 mg/L as the threshold for these reasons:

- It has false positive and negative rates comparable to what was found for the diatom-based nutrient increaser metrics applicable to this region(Teply, 2010b)
- It is very close to the lower bound of the 90% confidence interval (i.e., DO delta of 5.5 mg/L) of the change-point, as determined from the change-point analysis
- It keeps the proportion of reference-site data exceeding the threshold to no more than 10%
- It is in fairly good agreement with the threshold recommended by Minnesota for their plains region (i.e., DO delta of 5.0 mg/L)to protect fish and aquatic life

C.2.1 Instantaneous DO Monitoring in Wadeable Plains Streams

DEQ assessment can continue to rely on instantaneous measurements of DO. The following guidelines are recommended for instantaneous DO data collection.

When to Measure, Minimum: Without question the best time to measure the lowest daily DO is at dawn. DO in streams and standing waters is usually at its daily low just before sunrise (e.g., (Odum, 1956; Teply, 2010b; Boyd, et al., 1978; Madenjian, et al., 1987; Quinn and Gilliland, 1989). DO measurements in streams at other times of the day usually cannot give a reliable estimation of the nighttime low, especially in plains streams, because numerous other factors (e.g., wind speed and direction, air temperature, the stream's sediment oxygen demand, presence/absence of aquatic macrophytes) play a role in the rate of DO decline per unit time. Simple models incorporating various environmental factors have been used to estimate dawn DO in aquaculture ponds (e.g., (Boyd, et al., 1978; Madenjian, et al., 1987), and more sophisticated models can simulate diel DO cycles in streams and rivers (e.g., QUAL2K; (Chapra, et al., 2008). However, numerous input variables are required to run these models, making these impractical approaches for routine stream DO assessment. Therefore, we recommend dawn DO measurements be taken. We examined a number of plains stream diel DO plots, including both reference and non-reference sites, and found that **the most appropriate time window for capturing the DO daily minima is between dawn (or pre-dawn) and 8:00 am.** This time frame should be adhered to when sampling during the summer growing season (June 16th to September 30th).

The assessor should make notes of weather conditions at the time (approximate wind speed and direction, cloud cover).

When to Measure, Maximum: The DO concentration maximum in flowing streams usually occur after solar noon, commonly around 4:00 pm. The combined effects of plant/algae respiration, plant/algae primary production (which peaks around solar noon), and flow and re-aeration influences on DO saturation result in the typical sinusoidal DO patterns observed each day (Odum, 1956; Chapra and Di Toro, 1991). We examined the continuous recordings of DO for plains stream (both flowing and intermittent), and found that most daily DO peaks occurred between 2:30 pm and 5:00 pm. Time of year appeared to have no discernable effect, and the exact timing of the DO peak seemed more influenced by local factors (probably clouds, and wind velocity and direction). **We recommend that monitoring for the DO maximum occur between 2:30 pm and 5:00 pm.** Measurements need not be taken continuously during that period; checking stream DO every 15-30 minutes should be sufficient to catch the peak. You may need to stay somewhat beyond 5:00 pm if the values are still climbing. DEQ's main hand-held instrument for DO measurement is the YSI 85. This instrument has a 50 reading, manual-entry memory which can be used for collecting daily DO maximums. Set the instrument up in situ and then leave it on between 2:30 pm and 5:00 pm; record readings every 15-30 minutes by depressing the ENTER button for two seconds. Data may be downloaded later.

For the purpose of calculating DO delta, at least 3 DO sampling events should be taken in each assessment reach. Temporal independence of DO measurements is not a concern, since DO delta can be quite variable on a day-to-day basis. Each sampling event DO delta can considered on its own merits. Therefore, there is no reason to wait for 30 days to collect a subsequent DO measurement at a site. The assessor may collect DO data each day while they are in the area. This will also help increase the number of sampling events collected from an assessment reach.

C.3.0 Biochemical Oxygen Demand

Biochemical oxygen demand is one of the oldest water quality assessment tools, first recommended for use by the English Royal Commission on Sewage Disposal in the early 1900s (Hynes, 1966). It is a standardized test carried out over 5 days that measures the amount of putrescible material in water, which consumes oxygen as it decomposes. It is also one of the required measurements for wastewater nationally under the National Secondary Treatment Regulations (40 CFR part 133).

Montana has no standard for ambient BOD_5 in streams (although wastewater facility effluent and mixing zones are held to BOD_5 requirements). Nevertheless, the following guidelines for BOD_5 are commonly followed in many parts of the world:

- 1-2 mg BOD_5/L: Very clean water, little biodegradable waste
- 3-5 mg BOD_5/L: Moderately clean water, some biodegradable waste
- 6-9 mg BOD_5/L: Many bacteria, much biodegradable matter
- ≥10 mg BOD_5/L: Very bad, large amounts of biodegradable wastes in the water

The method used at the DPHHS Environmental Laboratory currently has a detection limit of 4 mg BOD_5/L (which coincides with the Royal Commission's recommendation that 4 mg BOD_5/L not be exceeded; Hynes, 1966). In plains streams we have found that otherwise healthy streams (i.e., reference sites) can have values in the 6-9 range fairly often. We recommend a value of 8.0 mg/L as a threshold for concern in plains streams.

APPENDIX D - REFERENCES

American Public Health Association. 1998. Standard Methods for the Examination of Water and Wastewater. L. S. Clesceri, A. E. Greenberg, and A. D. Eaton (Eds.), 20th ed., Washington, DC: American Public Health Association.

Arnwine, D. H. and K. J. Sparks. 2003. Comparison of Nutrient Levels, Periphyton Densities and Diurnal Dissolved Oxygen Patterns in Impaired and Reference Quality Streams in Tennessee. Nashville, Tennessee: Tennessee Department of Environment and Conservation, Division of Water Pollution Control.

Bahls, Loren L., M. Tepley, Rosie Sada de Suplee, and Michael W. Suplee. 2008. Diatom Biocriteria Development and Water Quality Assessment in Montana: A Brief History and Status Report. *Diatom Research.* 23(2): 533-540.

Barbour, Michael T., Jeroen Gerritsen, Blaine D. Snyder, and James B. Stribling. 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates, and Fish: Second Edition. Washington, DC: United States Department of Environmental Protection, Office of Water. Report EPA 841-B-99-002.

Biggs, B. J. F. 2000. New Zealand Periphyton Guideline: Detecting, Monitoring and Managing Enrichment of Streams, Christchurch, New Zealand: NIWA. http://www.mfe.govt.nz/publications/water/nz-periphyton-guide-june00.html.

Boyd, C. E., R. P. Romaire, and E. Johnston. 1978. Predicting Early Morning Dissolved Oxygen Concentrations in Channel Catfish Ponds. *Transactions of the American Fisheries Society.* 107: 484-492.

Bramblett, Robert G., Thomas R. Johnson, Alexander V. Zale, and Daniel Heggem. 2005. Development and Evaluation of a Fish Assemblage Index of Biotic Integrity for Northwestern Great Plains Streams. *Transactions of the American Fisheries Society.* 134(3): 624-640.

Bukantis, Robert T. 1998. Rapid Bioassessment Macrointertebrate Protocols: Sampling and Sample Analysis SOPs: Working Draft. Helena, MT: Montana Department of Environmental Quality.

California Environmental Protection Agency State Water Resources Control Board. 2004. Functional Equivalent Document, Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List - Final.

Chapra, Steven C. and D. M. Di Toro. 1991. Delta Method for Estimating Primary Production, Respiration, and Reaeration in Streams. *Journal of Environmental Engineering.* 117(5): 640-655.

Chapra, Steven C., Gregory J. Pelletier, and Hua Tao. 2008. A Modeling Framework for Simulating River and Stream Water Quality, Version 2.1: Documentaion and Users Manual. Medford, MA: Civil and Environmental Engineering Department, Tufts University.

Cochran, W. G. 1977. Sampling Techniques, 3rd ed., New York: John Wiley and Sons.

Conover, W. J. 1999. Practical Nonparametric Statistics, 3rd. ed., New York: John Wiley & Sons.

Cormier, Susan, Susan Braen Norton, Glen W. Suter, II, and Donna K. Reed-Judkins. 2000. Stressor Identification Guidance Document. Washington, D.C.: U.S. Environmental Protection Agency, Office of Water, Office of Research and Development. Report EPA 822-B-00-025.

Cormier, Susan and G. W. Suter, II. 2008. A Framework for Fully Integrating Environmental Assessment. *Environmental Management.* 42: 543-556.

DiTomaso, J. M. and E. A. Healy. 2003. Aquatic and Riparian Weeds of the West. Oakland, CA: University of California Agriculture and Natural Resources. Report 3421.

Dodds, Walter K., V. H. Smith, and Bruce Zander. 1997. Developing Nutrient Targets to Control Benthic Chlorophyll Levels in Streams: A Case Study of the Clark Fork River. *Water Resources.* Vol. 31(no. 7): 1738-1750.

Droop, M. R. 1973. Some Thoughts on Nutrient Limitation in Algae. *Journal of Phycology.* 9: 264-272.

Elrifi, I. R. and D. H. Turpin. 1985. Steady-State Luxury Consumption and the Concept of Optimum Nutrient Ratios: A Study With Phosphate and Nitrate Limited *Selenastrum Minutum* (Chlorophyta). *Journal of Phycology.* 21: 592-602.

Environmental Research Laboratory-Duluth. 1997. Guidelines for Preparation of the Comprehensive State Water Quality Assessments (305(b) Reports) and Electronic Updates: Supplement. Washington, DC: Assessment and Watershed Protection Division, Office of Wetlands, Oceans, and Watershed, Office of Water, U.S. Environmental Protection Agency. Report EPA-841-B-97-002B.

Flynn, K. and Michael W. Suplee. 2010. Defining Large Rivers in Montana Using a Wadeability Index. Helena, MT: Montana Department of Environmental Quality. http://deq.mt.gov/wqinfo/Standards/default.mcpx.

Gilbert, R. O. 1987. Statistical Methods for Environmental Pollution Monitoring., New York: John Wiley & Sons, Inc.

Hilsenhoff, W. L. 1987. An Improved Biotic Index of Organic Stream Pollution. *Great Lakes Entomologist.* 20(1): 31-39.

Hurlbert, S. H. 1984. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs.* 54: 187-211.

Hynes, H. B. N. 1966. The Biology of Polluted Waters, 3rd ed., Liverpool: Liverpool University Press.

Klarich, Duane A. 1982. General Characteristics of Aquatic Macrophyte Associations in the Upper Poplar River Drainage of Northeastern Montana. Billings, MT: Montana Department of Health and Environmental Sciences.

LaVoie, I., S. Campeau, F. Darchambeau, G. Cabana, and P. J. Dillon. 2008. Are Diatoms Good Integrators of Temporal Variability in Stream Water Quality? *Freshwater Biology.* 53: 827-841.

Lohman, K. and John C. Priscu. 1992. Physiological Indicators of Nutrient Deficiency in *Cladophora* (Chlorophyta) in the Clark Fork of the Columbia River, Montana. *Journal of Phycology.* 28: 443-448.

Madenjian, C. P., G. L. Rogers, and A. W. Fast. 1987. Predicting Night Time Dissolved Oxygen Loss in Prawn Ponds of Hawaii: Part 1 - Evaluation of Traditional Methods. *Aquacultural Engineering.* 6: 191-208.

Mapstone, B. D. 1995. Scalable Decision Rules for Environmental Impact Studies: Effect Size, Type I, and Type II Errors. *Ecological Applications.* 5: 401-410.

Minnesota Pollution Control Agency. 2010. *Draft* Minnesota Nutrient Criteria Development for Rivers. Report wq-s6-08. http://www.pca/state.mn.us/index.php/water/water-permits-and-rules/water-rulemaking/proposed-water-quality-standards-rule-revision.html.

Montana Department of Environmental Quality. 2005. Field Procedures Manual For Water Quality Assessment Monitoring. Helena, MT: Montana Department of Environmental Quality, Water Quality Planning Bureau. Report WQPBWQM-020.

-----. 2006. Sample Collection, Sorting, and Taxonomic Identification of Benthic Macroinvertebrates. Helena, MT: Montana Department of Environmental Quality. Report WQPBWQM-009.

-----. 2007. Sampling and Analysis Plan, and Associated White Paper: Sampling Reference Sites to Produce a More Uniform Nutrient Dataset and Improve Montana's Wadeable Stream Nutrient Criteria.

-----. 2010a. Box Elder Creek Nutrient Addition Study:A Project to Provide Key Information for the Development of Nutrient Criteria in Montana Prairie Streams Quality Assurance Project Plan. Helena, MT: Water Quality Planning Bureau.

-----. 2010b. Circular DEQ-7, Montana Numeric Water Quality Standards. Helena, MT.

-----. 2011a. Field Procedures Manual for Water Quality Assessment Monitoring, Draft. Helena, MT: Montana Department of Environmental Quality. Report WQPBWQM-020.v.3.

-----. 2011b. Periphyton Standard Operating Procedure. Helena, MT: Montana Department of Environmental Quality. Report WQPVWQM-010.

-----. 2011c. Sample Collection and Laboratory Analysis of Chlorophyll-*a* Standard Operation Procedure, Revision 5. Helena, MT: Montana Department of Environmental Quality. Report WQPBWQM-011.

Montana Department of Environmental Quality, Planning, Prevention and Assistance Division. 2009. Montana 2008 Final Water Quality Integrated Report. Helena, MT: Montana Department of Environmental Quality, Planning, Prevention and Assistance Division, Water Quality Planning Bureau. Report WQPBDMSRPT-02-F.

Muenscher, W. C. 1944. Aquatic Plants of the United States: Cornell University Press.

Neter, John, W. Wasserman, and Michael H. Kutner. 1989. Applied Linear Regression Models, 2nd Edition ed., Homewood, IL: Irwin Press. Accessed 3/90.

Norris, R. H. and A. Georges. 1993. "Analysis and Interpretation of Benthic Macroinvertebrate Surveys," in *Freshwater Biomonitoring and Benthic Macroinvertebrates*, Rosenberg, D. M and Resh, V. H., (New York: Chapman and Hall)

Norris, R. H., E. P. McElravy, and V. H. Resh. 1992. "The Sampling Problem," in *The River Handbook*, Calow, P. and Petts, G. E., (Oxford, England: Blackwell Scientific Publications)

Odum, H. T. 1956. Primary Production in Flowing Waters. *Limnology and Oceanography.* 1: 102-117.

Osenberg, C. W., R. J. Schmitt, S. J. Holbrook, K. E. Abu-Saba, and A. R. Flegal. 1994. Detection of Environmental Impacts: Natural Variability, Effect Size, and Power Analysis. *Ecological Applications.* 4: 16-30.

Ott, R. L. 1993. An Introduction to Statistical Methods and Data Analysis, 4th Edition ed., Belmont, CA: Duxbury Press.

Portielje, R. and L. Lijklema. 1994. Kinetics of Luxury Uptake of Phosphate by Algae-Dominated Benthic Communities. *Hydrobiologia.* 275-276(1): 349-358.

Qian, S. S., R. S. King, and C. J. Richardson. 2003. Two Statistical Methods for the Detection of Environmental Thresholds. *Ecological Modelling.* 166: 87-97.

Quinn, J. M. and B. W. Gilliland. 1989. The Manawatu River Cleanup - Has It Worked? *Transactions of the Institute of Professional Engineers of New Zealand.* 16: 22-26.

Rhee, G. Y. 1973. A Continuous Culture Study of Phosphate Uptake, Growth Rate and Polyphosphate in *Scenedesmus Sp. Journal of Phycology.* 9: 495-506.

Sandgren, C. D., P. M. Engevold, S. Neerhof, and T. J. Ehlinger. 2004. Nuisance *Cladophora* in Urban Streams: Habitats, Seasonality, Morphology, Production, Nutrient Composition, Heavy Metals, Foodweb Bottleneck. In: Bootsma, Harvey A., Erika T. Jensen, Erica B. Young, and John A. Berges (eds.). Proceedings of a Workshop Held at the Great Lakes WATER Institute, University of Wisconsin-Milwaukee. *Cladophora* Research and Management in the Great Lakes. 43-56.

Schroeter, S. C., J. D. Dixon, J. Kastendiek, and R. O. Smith. 1993. Detecting the Ecological Effects of Environmental Impacts: A Case Study of Kelp Forest Invertebrates. *Ecological Applications.* 3: 331-350.

Stevenson, R. Jan, M. L. Bothwell, and R. L. Lowe. 1996. Algal Ecology, Freshwater Benthic Ecosystems: Academic Press.

Stevenson, R. J. and E. F. Stoermer. 1982. Luxury Consumption of Phosphorus by Five *Cladophora* Epiiphytes in Lake Huron. *Transactions of the American Microscopy Society.* 101: 151-161.

Stewart-Oaten, A., J. R. Bence, and C. W. Osenberg. 1992. Assessing Effects of Unreplicated Perturbations: No Simple Solutions. *Ecology.* 73: 1396-1404.

Stewart-Oaten, A. and W. W. Murdock. 1986. Environmental Impact Assessment: "Psuedoreplication" in Time? *Ecology.* 73: 929-940.

Suplee, Michael W. 2004. Wadeable Streams of Montana's Hi-Line Region : An Analysis of Their Nature and Condition With an Emphasis on Factors Affecting Aquatic Plant Communities and Recommendations to Prevent Nuisance Algae Conditions. Helena, MT: Montana Department of Environmental Quality, Water Quality Standards Section.

Suplee, Michael W., Rosie Sada de Suplee, David L. Feldman, and Tina Laidlaw. 2005. Identification and Assessment of Montana Reference Streams: A Follow-Up and Expansion of the 1992 Benchmark Biology Study. Helena, MT: Montana Department of Environmental Quality.

Suplee, Michael W., Arun Varghese, and Joshua Cleland. 2007. Developing Nutrient Criteria for Streams: An Evaluation of the Frequency Distribution Method. *Journal of the American Water Resources Association.* 43(2): 456-472.

Suplee, Michael W., V. Watson, M. Teply, and H. McKee. 2009. How Green Is Too Green? Public Opinion of What Constitutes Undesirable Algae Levels in Streams. *Journal of the American Water Resources Association.* 45: 123-140.

Suplee, Michael W., V. Watson, A. Varghese, and Joshua Cleland. 2008. Scientific and Technical Basis of the Numeric Nutrient Criteria for Montana's Wadeable Streams and Rivers. Helena, MT: MT DEQ Water Quality Planning Bureau.

Tepley, M. and Loren L. Bahls. 2005. Diatom Biocriteria for Montana Streams. Larix Systems,Inc. and *Hannaea*.

Teply, M. 2010a. Interpretation of Periphyton Samples From Montana Streams. Lacey, WA: Cramer Fish Sciences.

Teply, Mark. 2010b. Diatom Biocriteria for Montana Streams. Lacey, WA: Cramer Fish Sciences.

Teply, Mark E. and Loren L. Bahls. 2006. Diatom Biocriteria for Montana Streams: Middle Rockies Ecoregion. Helena, MT: Larix Systems, Inc.

Tetra Tech Inc. 2010. Analysis of Montana Nutrient and Biological Data for the Nutrient Scientific Technical Exchange Partnership Support (N-STEPS).

U.S. Environmental Protection Agency. 2002. Consolidated Assessment and Listing Methodology: Towards a Compendium of Best Practices. Washington, D.C.: U.S. Envrionmental Protection Agency.

U.S.Environmental Protection Agency. 2000. Nutrient Criteria Technical Guidance Manual, Rivers and Streams. Washington, DC: United States Environmental Protection Agency. Report EPA-822-B00-002. http://www.epa.gov/waterscience/criteria/nutrient/guidance/rivers/index.html.

-----. 2006. Data Quality Assessment: Statistical Methods for Practitioners. Washington, DC: United States Environmental Protection Agency, Office of Environmental Information. Report EPA/240/B-06/003.

Underwood, A. J. 1994. On Beyond BACI: Sampling Designs That Might Reliably Detect Environmental Change. *Ecological Applications.* 4: 3-15.

Varghese, A. and Joshua Cleland. 2008. Updated Statistical Analyses of Water Quality Data, Compliance Tools, and Changepoint Assessment for Montana Rivers and Streams. Fairfax, VA.

Varghese, Arun and Joshua Cleland. 2005. Seasonally Stratified Water Quality Analysis for Montana Rivers and Streams: Final Report. Fairfax, VA: ICF Consulting.