Feb. 3, 2025

Draft Methodology Update for Assessing Freshwater Biocriteria

Technical Support Document

For development of the Water Quality Status Report and List of Impaired Waters

This document was prepared by
Oregon Department of Environmental Quality
WQ Assessment Program
700 NE Multnomah Street, Portland Oregon, 97232

Authors:
Lesley Merrick
Shannon Hubler
Kaegan Scully-Engelmeyer
Travis Pritchard

Contact: Lesley Merrick
Lesley.merrick@deq.oregon.gov
Phone: 971-323-7228
www.oregon.gov/deq

State of Oregon
DEQ Department of Environmental Quality

Translation or other formats

Español | 한국어 | **繁體中文** | **Русский** | **Tiếng Việt** | العربية

800-452-4011 | TTY: 711 | deqinfo@deq.oregon.gov

Non-discrimination statement

DEQ does not discriminate on the basis of race, color, national origin, disability, age or sex in administration of its programs or activities. Visit DEQ's Civil Rights and Environmental Justice page.

Executive Summary

The Federal Clean Water Act requires Oregon to report on the quality of its surface waters every two years. The Oregon Department of Environmental Quality assesses surface waters to determine if they contain parameters at levels that exceed protective water quality standards. The result of these analyses and conclusions is called the "Integrated Report" because it combines the requirements of Clean Water Act section 305(b) to develop a status report and the section 303(d) requirement to develop a list of impaired waters. DEQ's Integrated Report Assessment Methodology contains the "decision rules" DEQ will use to compare data and information to existing water quality standards in the development of the Integrated Report. This document contains DEQ's proposed revisions to the Assessment Methodology used to interpret Oregon's narrative biocriteria standard (340-041-0011) in freshwater for use in the 2026 Integrated Report.

Under the Clean Water Act framework, detrimental changes in biological integrity are considered a form of pollution that should be included in the assessment of water quality status for a waterbody. EPA guidance recommends using measurable components of an ecological system, including macroinvertebrate assemblages, as indicators of aquatic life beneficial use support. DEQ has assessed macroinvertebrate assemblage data to interpret Oregon's narrative biocriteria for freshwater since 2010. This update represents the largest revision to Oregon's freshwater bioassessment methodology for the Integrated Report.

The primary changes described in this update are: (1) an update to the previous approach to define Reference condition, (2) the refinement of DEQ's existing Observed vs. Expected index, (3) the introduction of two new assessment indices, (4) minimum sample size requirements, (5) assessment benchmarks linked to ecological condition, and (6) the adoption of a hybrid assessment framework that utilizes multiple lines of evidence. These updates were designed to increase confidence in waterbody assessment conclusions and categorical determinations based on biological data.

Contents

Introduction

Under the Clean Water Act, states, territories, and authorized Tribes are required to report to the U.S. Environmental Protection Agency about the status of designated beneficial uses of their waters and to identify waterbodies where water quality impacts are affecting those uses. The result of these analyses and conclusions is called the "Integrated Report" because it combines the requirements of Clean Water Act section 305(b) to develop a status report and the section 303(d) requirement to develop a list of impaired waters. DEQ's Integrated Report represents the state's most comprehensive evaluation of water quality data and information to determine whether Oregon's waters are fully supporting beneficial uses, such as fish and aquatic life, drinking water or water contact recreation.

The stated goal of the Clean Water Act is to "restore and maintain the physical, chemical, and biological integrity of the Nation's waters." EPA guidance recommends using biological community assessments as an indicator for aquatic life beneficial use support. Biological assessment is the quantitative measure of the biological condition of the resident aquatic community. When bioassessment tools indicate detrimental changes in resident biological communities, it is considered a form of pollution to be regulated under the CWA. Oregon's narrative biocriteria standard states "*Waters of the State must be of sufficient quality to support aquatic species without detrimental changes in the resident biological communities*" (OAR 340-041-0011).

This document provides the rationale of proposed updates to DEQ's existing Integrated Report assessment methodology used to interpret and assess the narrative biocriteria water quality standard.

Background

Implementing the narrative biocriteria criteria

DEQ has been using freshwater benthic macroinvertebrate assemblage biological assessment tools as a direct measurement of aquatic life beneficial use support since the 2010 Integrated Report. The foundation of the existing assessment methodology is the application of a multivariate predictive model, where the observed macroinvertebrate assemblage at a stream site is compared to the assemblage predicted to occur if the site were truly in "Reference" (least disturbed) condition (Hubler 2008). If the number of observed taxa (O) is equal the number of expected Reference taxa (E), the O/E ratio is 1. For sites with ratios less than 1.0, the value can be expressed as a percentage of Reference "taxa loss", or percent reduction in native biodiversity.

For the assessment, DEQ's Assessment Methodology uses benchmark values for percent taxa loss to indicate where and when deviations from Reference conditions and loss of native taxa are detrimental to macroinvertebrate assemblages and impair aquatic life use support in the waterbody. The benchmark values were derived using the lower tenth percentile of taxa loss scores at Reference sites. When the average taxa loss value for an assessment unit or monitoring location (depending on assessment unit type) exceeds this taxa loss benchmark the waterbody was placed in Category 5 and included on Oregon's 303(d) list of impaired waters.

History of Biocriteria Assessment in Oregon

2010

In 2010, DEQ developed a biocriteria assessment methodology to identify impaired waterbodies based on the condition of macroinvertebrate assemblages. DEQ selected 10th and 25th percentiles of Reference conditions as the benchmarks for aquatic life use support to be consistent with previous biological assessments. This methodology did not use Category 5 to add waterbodies with impaired biological condition to the 303(d) list, but rather identified them as Category 3C: Impairing Pollutant Unknown.

In its review of the 2010 Integrated Report, EPA agreed with DEQ's determination that 321 stream segments were impaired for biocriteria based on its use of the assessment methodology. However, EPA disagreed with DEQ that these waterbodies should not be put on Oregon's 303(d) list and that a TMDL could not be developed because the pollutants were unknown. EPA disapproved DEQ's decision to not include these 321 impaired segments on Oregon's 303(d) list, based on the determination that they are impaired for "pollutant unknown" and subsequently EPA included these waterbodies as Category 5 in the 2010 Integrated Report. In its approval letter, EPA stated its expectation that DEQ will "… include all biological impairments in Category 5 …".

2012

Based on EPA's approval rationale for Oregon's 2010 Integrated Report, DEQ modified its biocriteria methodology for the 2012 report to identify biocriteria impairments as Category 5: Impaired and a TMDL is needed.

2018/2020

From 2014 to 2020, DEQ underwent a large Integrated Report improvement project, which included updates to the 2012 biocriteria assessment methodology. The need for technical peer review on substantial updates to the Integrated Report assessment methodology was established by the legislature through ORS 468.B.039 in 2015. As a result, in 2017 DEQ convened

a panel of bioassessment experts to solicit independent scientific and technical input regarding the biocriteria impairment assessment benchmarks. DEQ summarized the peer review responses and documented the key findings of the panel as:

(1) DEQ's biocriteria benchmarks are valid and are similar to benchmarks used in other states.
(2) Use of two impairment benchmarks is more technically defensible than use of a single benchmark and may more accurately inform management decisions.
(3) Moving forward, DEQ should:
    a. Attempt to relate impairment benchmarks to ecological condition, rather than relying solely on statistically-based benchmarks.
    b. Improve Reference validation datasets to independently assess model accuracy.
    c. Improve estimates of error rates or repeatability.

Based on the panel recommendations, DEQ committed to making minor revisions to the biocriteria assessment methodology for the 2018/2020 Integrated Report, while also working on more in-depth revisions. The panel identified an area of concern related to the inherent variability and repeatability of macroinvertebrate sampling. To address this, DEQ added an additional assessment benchmark defined by the lower 5th percentile of taxa loss at Reference sites for waterbodies with only one macroinvertebrate sample. Waterbodies with two or more samples retained the 10th percentile assessment benchmark.

2024

Following the 2018/2020 assessment methodology update, DEQ began working on revising the biocriteria methodology based on the suggestions identified in the 2018 peer review. In October of 2024, DEQ convened a scientific peer review panel, which included several original members of the 2018 panel, to review the methodology updates. DEQ provided the 2024 peer review panel the first draft of this Technical Support document detailing the proposed assessment methodology updates, decision-making rationale, and other supporting information. The primary updates described in the first draft were: (1) an update to the previous approach to define reference condition, (2) refinement of DEQ's existing Observed vs. Expected index, (3) the introduction of two new assessment indices: Multi metric Index (MMI) and Biological Condition Gradient (BCG), (4) a new approach to derive ecologically relevant assessment benchmarks based on the relationship between the individual indexes and the BCG, and (5) the adoption of a hybrid assessment framework that uses multiple lines of evidence.

DEQ received constructive feedback from the panel on all aspects of the proposed methodology, with general consensus on a few recommended significant changes to the

approach presented in the first draft. The proposed assessment methodology detailed in this document has been revised in response to comments received by the 2024 panel.

New Bioassessment Tools

DEQ has been working on revising its macroinvertebrate based bioassessment tools for the last decade. Revisions have included: migrating macroinvertebrate data to AWQMS (DEQ's water quality database), an updated Reference Condition Approach, a new RIVPACS (River Invertebrate Prediction and Classification System)-type O/E predictive index, a multi metric index (MMI), and incorporation of a multistate Biological Condition Gradient (BCG) index.

The three bioassessment indexes developed by DEQ use biological macroinvertebrate data as direct indicators of fish and aquatic life use support. They are useful in combination because they each represent various aspects of biotic integrity. RIVPACS-type O/E models represent native taxa richness expected under least disturbed (Reference) conditions, and low O/E values can be considered loss of native taxa richness (Hawkins et al. 2000, Hawkins 2009). MMIs use a combination of metrics — each representing a different aspect of biotic integrity — such as tolerance to pollution, taxa richness, or functional feeding groups (Hawkins et al. 2009, Mazor et al. 2016). BCGs are based on a narrative gradient of stream conditions from fully natural to highly disturbed, with quantitative rules that must be fully met to be classified into a certain BCG-level (Paul et al. 2020). The rules vary by each BCG-level, represented by metrics of community composition such as richness, diversity and tolerance.

There are also considerable differences in construction methods for each index. O/E models use only Reference sites (least disturbed by human activities) in predicting expected taxa richness. MMIs use both Reference and Most Disturbed sites, with metrics chosen based on which ones show the greatest differences between Reference and Most Disturbed sites. To develop the new RIVPACS O/E and MMI indexes, DEQ worked closely with experts from Utah State University's National Aquatic Monitoring Center, who have experience developing bioassessment indexes at national, regional, and state scales. NAMC staff provided R-code for developing models, which was adapted to fit DEQ's data systems.

BCGs are developed with samples across the entire range of disturbances used in index construction, without relying exclusively on Least or Most Disturbed sites. In addition, human disturbance is factored into the process by assigning macroinvertebrate taxa to attribute levels indicative of tolerance (or sensitivity) to human disturbances. A panel of experts is asked to classify sites into narrative BCG classes, based entirely on the composition of macroinvertebrate samples, followed by developing a set of rules to classify sites into levels based on the panelists results. The BCG model (Stamp 2022) was developed in concert with federal, state, and local

experts in bioassessment. The project was led by experts from Tetra Tech with experience building BCGs in a variety of settings and for multiple aquatic assemblages. The final BCG was developed from datasets covering the western portions of both Oregon and Washington. For Oregon specifically, this included samples collected in the following Omernik Level III ecoregions: Coast Range, Willamette Valley, and Cascades.

This document describes how to apply the O/E and MMI indexes to assess Oregon's biocriteria water quality standard, as well as use of the BCG in relating O/E and MMI results to ecological conditions. It is beyond the scope of this document to fully describe the Reference Condition Approach and the various index development methods. Further information on the model development can be found in the technical appendices.

Updated Reference Condition Approach

The definition of Reference conditions is an integral part of bioassessment models because it establishes one end of the spectrum of the biological condition. Since 2014, DEQ has undertaken multiple efforts to bring its Reference Condition Approach up to date, incrementally improving methods for determining 'least disturbed' Reference conditions, as well as defining "Most Disturbed" conditions. (For complete details on Reference and Most Disturbed classes, see: ODEQ 2022). Examples of RCA improvements include automating watershed delineations, extensive review of sampling locations and rectifying errors associated with digitized stream layers, sourcing candidate Reference sites from other agencies and organizations, and adding quality assurance steps to verify disturbance status. An additional and significant update to the RCA is the application of disturbance thresholds in the GIS screening process equally across the state instead of the ecoregion level, as was practiced in the previous approach. Candidate Reference and Most Disturbed sites were selected based on scores across a suite of disturbance metrics before moving through a verification process to determine the final group of sites in each disturbance class. Reference and Most Disturbed sites selected using the updated RCA were used to develop the new O/E and MMI models. Macroinvertebrate assemblage data was not used to define disturbance status, thus avoiding circularity in index construction and application.

Performance of the Bioassessment Indexes

DEQ's new MMI and updated O/E bioassessment indexes were compared to a set of performance metrics to identify how well they represent the ability to measure biological integrity (Table 1). These performance metrics were derived from conversations with experts at the NAMC, as well as from California's Department of Fish and Wildlife in developing their own O/E and MMI indexes (Mazor et al. 2016). Because the MMI and O/E indexes are on different

scales, for direct comparisons between the two models, the MMI is re-scaled by dividing all MMI scores by the mean of Reference calibration MMI scores. (For clarity: "calibration" = samples used to build the model, "validation" = samples in the same disturbance class that were not used to build the models but rather to provide a sense of how well the models may work to assess novel stream reaches.)

Accuracy: Both the RIVPACS O/E and MMI indexes had Reference means ~ 1.0, substantially greater than mean scores of Most Disturbed sites (0.73 and 0.63, respectively (**Table 1**)). Reference validation samples were close with an O/E mean of 0.97 and a MMI mean of 1.0.

Precision: Precision was measured as the standard deviation of O/E and MMI scores at calibration samples used to build the indexes, as well as for calibration sites with multiple samples. The O/E model Reference calibration SD was 0.16, while the MMI was slightly more precise with a SD of 0.14. Both of these values are slightly lower than reported for the CAFW indexes (Mazor et al. 2016), and similar to other published models in the western U.S. (Hargett et al. 2007, Hubler 2008). The SD of MMI and O/E scores at Most Disturbed sites was nearly 2 times that observed for Reference sites.

Responsiveness: DEQ tested the responsiveness of the new indexes by performing a t-test on Reference calibration and Most Disturbed calibration scores, using the same set of Most Disturbed samples for each index, even though Most Disturbed samples are not included in RIVPACS-type O/E models. Both indexes showed highly significant ($p < 2.2 \times 10^{-16}$) difference between Reference scores and Most Disturbed scores. T-scores for comparisons of Reference and Most Disturbed O/E (t = -13.3) and MMI scores (t = -12.6) were higher than for any single macroinvertebrate summary metric (t = -11.8 for both # of Trichoptera taxa and # rheophilic taxa).

Sensitivity: To measure sensitivity, DEQ compared O/E and MMI scores at Most Disturbed sites to the 10^{th} percentile of Reference calibration scores. Both indexes had > 50% of Most Disturbed sites with index scores less than the Reference 10^{th} percentile, with MMI (66%) showing slightly higher sensitivity than O/E (55%). Higher sensitivity for MMI is expected, since index calibration is designed to distinguish between Reference and Most Disturbed.

Bias: DEQ tested bias in the predictive models by running Reference-calibration O/E and MMI scores through Random Forest (RF) models based on StreamCat (Hill et al. 2016) natural predictors. If a substantial amount of Reference score variation was explained by these natural gradients, the models were considered biased and potentially inaccurate across the landscape. For example, a high value for "percent variability explained" in RF model output, with elevation and precipitation showing high variable importance values, it might reasonably be concluded an index performed differently for low vs higher elevations and precipitations. RF model results

showed no bias in our index predictions to natural gradients, with low percent variability explained across both indexes.

Table 1. Model performances and evaluation criteria for the newly developed RIVPACS-type O/E and MMI bioassessment indexes. Because the two indexes are on different scales, re-scaled values are provided for the MMI to allow for direct comparisons to the O/E index. (MMI was re-scaled by dividing MMI scores by the Reference calibration mean MMI score.)

Performance Metric	Bioassessment Index	
	O/E v2.0	**MMI v1.0**
Reference 10th percentile	0.79	0.81
ACCURACY: • **Reference means ~ 1.0 (when MMI re-scaled)** • **~ 90% of Reference sites have scores > 10th percentile of calibration Reference site scores**		
Reference calibration Mean	1.02	1.01
Most Disturbed calibration Mean	n/a	0.63
% of Reference Validation > 10th percentile of Reference Calibration	85%	89%
PRECISION: Scores are similar when measured under similar settings • **low Standard Deviation of Reference calibration scores** • **low Standard Deviation of Reference calibration scores with multiple samples**		
Reference calibration standard deviation	0.16	0.14
Reference repeat calibration samples standard deviation	0.15	0.13
Most Disturbed calibration standard deviation	n/a	0.34
RESPONSIVENESS: Scores change in response to human activity gradients • **Highly significant difference (t-statistic) between Reference and Most Disturbed scores**		
t-value (Reference vs Most Disturbed)	-13.3	-12.6
SENSITIVITY: Scores indicate poor condition at sites with high human disturbance • **> 50% of Most Disturbed scores are < 10th percentile of Reference calibration sites**		
% Most Disturbed sites below 10th percentile of Reference	60%	66%
BIAS: Reference scores are minimally influenced by natural gradients • **Low variance explained by Random Forest models of Reference scores and natural predictors**		
% variance in Reference calibration scores explained by natural predictors	-0.12%	-26.7%

Developing Assessment Benchmarks

A key component to this assessment methodology update is to interpret the narrative biocriteria using assessment benchmarks linked to ecological condition. The consensus of the 2018 peer review panel was that DEQ's existing biocriteria benchmarks were valid, derived from standard and acceptable methods, based soundly on a statistical distribution approach and similar to methods employed by other states. However, they noted the lack of a direct linkage between the benchmarks and descriptions of ecological condition.

For this reason, DEQ is proposing to align the narrative language in the BCG levels to the statistically derived assessment benchmarks for attainment and non-attainment (impairment) of the narrative biocriteria. The rationale for this decision is based on the intended purpose of the BCG, which is built on a narrative backbone of changes in structural and functional characteristics of stream biota as they degrade in response to human disturbance. DEQ will continue to derive the assessment benchmarks from the statistical distribution of index scores at Reference sites used to build the indexes; then DEQ will use the corresponding BCG level to explain the biological consequences of index values below the assessment benchmarks.

Assessment Benchmark Values

DEQ will continue to use statistical-based benchmarks at the 10th and 25th percentiles of O/E index values for Reference calibration samples used to build the models. Therefore, the O/E assessment benchmark for impairment is 0.79, which equates to the 10th percentile of Reference calibration samples. The O/E attainment benchmark of 0.91 equates to the 25th percentile of Reference calibration samples used to build the model (Figure 1). For comparison, statistical-based benchmarks used in DEQ's previous O/E indexes (O/E v1.0, Hubler 2008) were slightly lower for impairment in

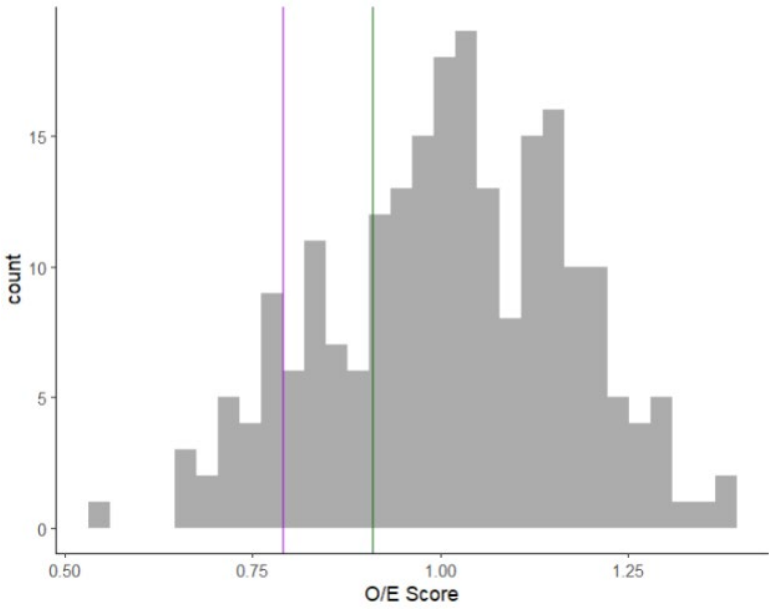

Figure 1. Histogram of the distribution of O/E index values for Reference calibration samples used to build the model. Purple vertical line represents the 10th percentile value of 0.79 and the green vertical line represents the 25th percentile at 0.91.

the WC+CP model (0.78) and 0.06 higher for the MWCF model (0.85), while the attainment benchmarks were essentially equivalent across all models.

Similarly, the MMI assessment benchmark for impairment is 0.81, which equates to the 10th percentile of Reference calibration samples. The MMI attainment benchmark of 0.90 equates to the 25th percentile of Reference calibration samples (Figure 2).

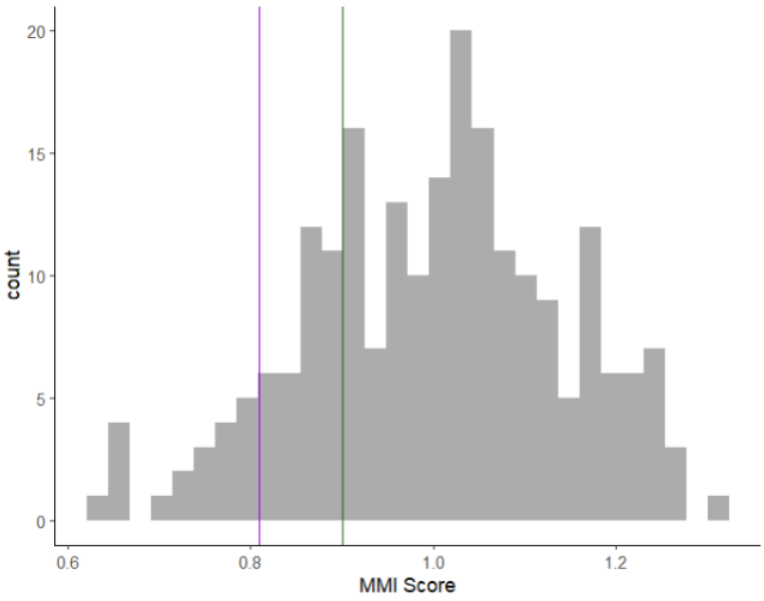

Figure 2. Histogram of the distribution of MMI values for Reference calibration samples used to build the model. Purple vertical line represents the 10th percentile value of 0.81 and the green vertical line represents the 25th percentile at 0.90.

Linking assessment benchmarks to ecological condition

DEQ's OAR 340-041-0002 defines "Ecological Integrity" as "the summation of chemical, physical, and biological integrity capable of supporting and maintaining a balanced, integrated, adaptive community of organisms having a species composition, diversity, and functional organization comparable to that of the natural habitat of the region." In DEQ's previous freshwater biocriteria methodology, taxa loss (derived from O/E v1.0) was the key attribute used to describe changes in macroinvertebrate assemblages. A critique of this method was that there are aspects of ecological integrity, particularly ecological function, that are not well described by taxa loss alone. The use of two models in this methodology update, that together describe five aspects of macroinvertebrate assemblage composition and structure, is an important improvement on DEQ's previous approach.

While O/E and MMI models are derived using biologically based metrics to describe ecological condition in a variety of ways, it is challenging to interpret the resulting numeric output in ecological terms. The BCG is a conceptual model describing how ecological attributes change in response to human disturbance (Stamp, 2022). Though fundamentally qualitative, the stressor

response curve generated in the construction of the BCG offers a practical way to interpret the numeric assessment benchmarks for O/E and MMI models in ecological terms. Linking the O/E and MMI indexes to the BCG narrative helps identify the changes in ecological condition the attainment and impairment benchmarks represent. Put another way, relating O/E and MMI scores to the BCG can provide information on the ecological consequences (what is lost) by falling below a certain benchmark. To demonstrate this, O/E and MMI sample scores across all disturbance gradients were plotted along with their associated BCG categories between 2-6, each of which represents a narrative description of ecological condition (Figure 3).

Attaining macroinvertebrate samples (> 25th Reference percentiles) most commonly are associated with the following ecological conditions:

- **BCG level 2**: Minimal changes in structure of the biotic community and minimal changes in ecosystem function. A diverse community reflects high habitat complexity and resilience; regionally endemic and highly sensitive taxa are maintained with some changes in biomass and/or abundance; sensitive taxa contribute a high proportion of the total taxa richness and individuals; ecosystem functions are fully maintained within the range of natural variability.
- **BCG level 3**: Evident changes in structure of the biotic community from reduced habitat complexity and resilience, with minimal changes in ecosystem function. Diminished biodiversity due to loss of some regionally endemic and highly sensitive taxa. Some shift towards dominance by common, widespread, less sensitive taxa. Intermediate sensitive taxa are still common and abundant, and ecosystem functions are fully maintained.

Conversely, impaired macroinvertebrate samples (≤ 10th Reference percentiles) are most commonly associated with one of these ecological conditions:

- **BCG level 5**: Major changes in structure of the biotic community and moderate changes in ecosystem function. Sensitive taxa are markedly diminished; conspicuously unbalanced distribution of major groups from that expected; organism condition shows signs of physiological stress; system function shows reduced complexity and redundancy; increased build-up or export of unused materials.
- **BCG level 6**: Extreme changes in structure and ecosystem function; wholesale changes in taxonomic composition; extreme alterations from normal densities.

Samples that fall between the attainment and impairment classes most commonly are associated with the following ecological conditions:

- **BCG level 4**: Moderate changes in structure of the biotic community and minimal changes in ecosystem function. Moderate changes in structure due to replacement of some intermediate sensitive taxa by more tolerant taxa, but reproducing populations of

some sensitive taxa are maintained; overall balanced distribution of all expected major groups; ecosystem functions largely maintained through redundant attributes.

As pointed out by the 2024 peer review panel, the current state of the BCG index shows the need for more refinement, especially to reduce the overlap in O/E and MMIs scores among high quality (Levels 2 and 3) and low quality (Levels 5 and 6) samples. That said, the relationships of the O/E and MMI indexes show a pattern of reduced ecological conditions below the impairment benchmarks, as related to the current BCG index. In other words, a ~ 20% loss of expected reference taxa (O/E < 0.79) or a ~ 20% reduction in expected MMI score (MMI < 0.81), represents major or extreme changes in the macroinvertebrate assemblage structure and function.

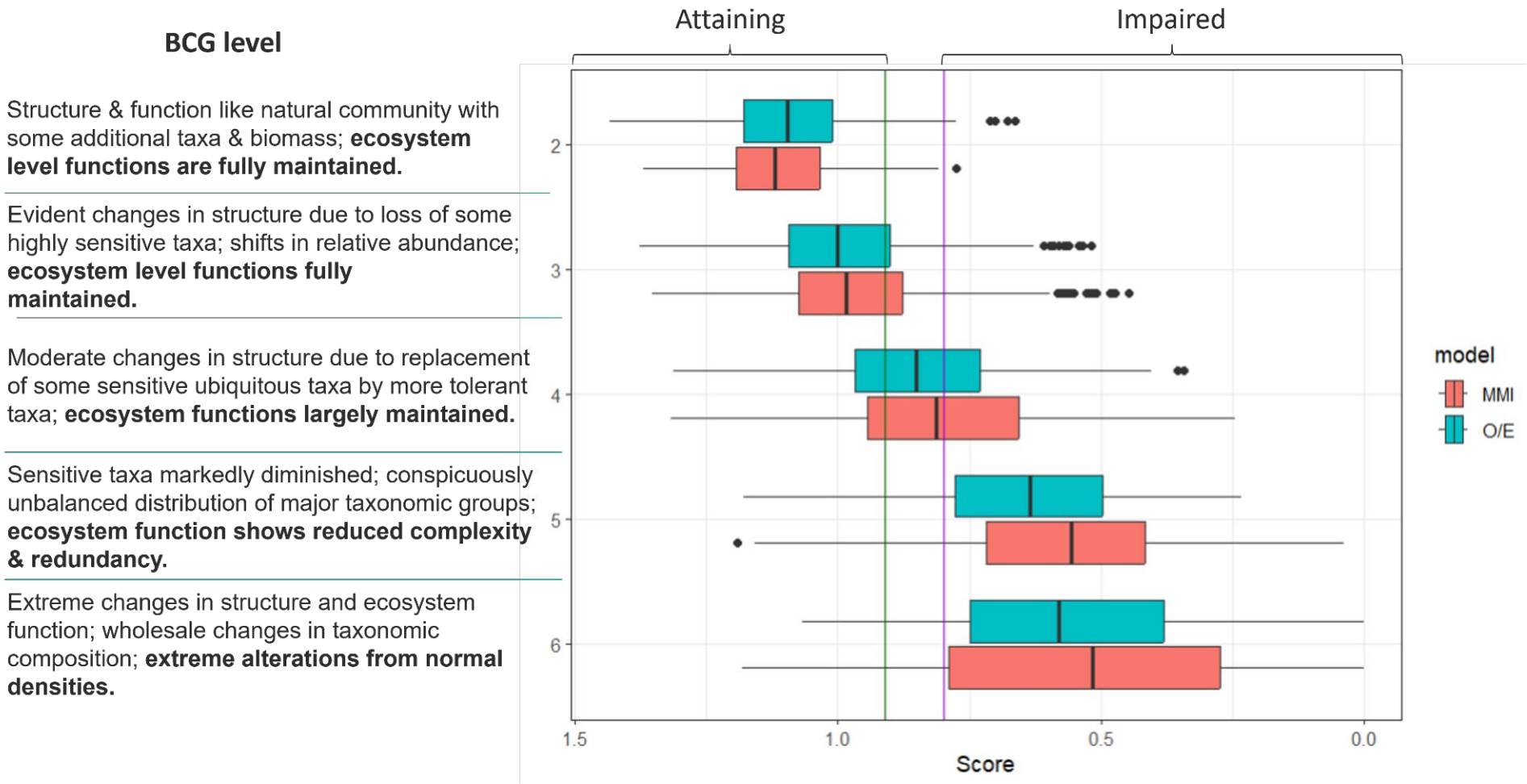

Figure 3. Box plots of O/E and MMI scores for all sites in DEQ's AWQMS database linked to the narrative levels of the Biological Condition Gradient index for western Oregon.

Hybrid framework

DEQ's marine biocriteria assessment methodology uses a hybrid approach to combine multiple lines of evidence to determine waterbody status. DEQ is proposing to use a similar approach here to relate the O/E and MMI assessment benchmarks as shown in Figure 4. In this approach, the O/E and MMI assessment benchmarks are used in conjunction to classify waterbody status. DEQ's previous biocriteria assessment methodology relied on a single macroinvertebrate index of biotic condition (O/E v1.0), which represents a single aspect of macroinvertebrate assemblage conditions (loss of expected Reference taxa richness). Incorporating the MMI adds an index that assesses conditions of four other aspects of assemblage composition and structure, providing a broader assessment of aquatic life use support. In addition, the hybrid approach requires attainment and impairment decisions to be confirmed by both indices, which reduces the potential error in assessment conclusions based on an individual index.

When both MMI and O/E index values are above the attaining assessment benchmarks, the waterbody will be classified as attaining (Category 2). When both index values are below the impairment assessment benchmark, the waterbody will be classified as impaired (Category 5). In cases when at least one index value falls between the assessment benchmarks but neither falls below the impairment benchmark, the waterbody will be classified in "Category 3C: insufficient data; non-Reference condition: Biocriteria scores differ from Reference condition, but are not classified as impaired". This sub-category was created in the 2020 IR to identify waterbodies with minimally disturbed biological condition from those units that are on the cusp of impairment, or in what is commonly considered "fair" condition in bioassessment. In cases where one index value falls below the assessment benchmark for impairment and the other index value does not, DEQ will classify the waterbody as "Category 3B: insufficient data; potential concern" and will prioritize the waterbody for follow up monitoring. Those waterbodies that lack sufficient data quality or are outside of the index calibration area will be assigned "Category 3: Insufficient data to determine whether a designated use is supported".

- Category 5 – Both indexes indicate impairment
- Category 2 – Both indexes indicate attainment
- Category 3C: insufficient data; non-Reference condition, but not impaired – This category is used to classify waterbodies that are closer to attainment than impairment
- Category 3B: Potential Concern – This category is used to classify waterbodies where one model indicates impairment and the other does not. It is used to indicate uncertainty in the assessment and recommendation for follow up monitoring

<table>
<tr><th colspan="2" rowspan="2">Hybrid Assessment Framework</th><th colspan="3">O/E v2.0</th></tr>
<tr><th>Biological impact & deviation from natural background ≤ 0.79</th><th>Differs from reference, but not impaired >0.79 and ≤ 0.91</th><th>Biological impact does not deviate from natural background > 0.91</th></tr>
<tr><td rowspan="3">MMI v1.0</td><td>Biological impact & deviation from natural background MMI score ≤0.81</td><td>Impaired – Cat 5</td><td colspan="2">Category 3B – Potential Concern</td></tr>
<tr><td>Differs from reference, but not impaired >0.81 and ≤ 0.90</td><td rowspan="2">Category 3B – Potential Concern</td><td rowspan="2">Cat 3C – differs from reference, but is not impaired</td><td></td></tr>
<tr><td>Biological impact does not deviate from natural background > 0.90</td><td>Cat 2 - Attaining</td></tr>
</table>

Figure 4. Hybrid assessment framework for freshwater biocriteria that combines multiple lines of evidence to determine assessment category.

Analysis on minimum sample size and error rate

Analyses on the use of a single sample and Type I and II error rates were identified by the 2018 and 2024 peer review panels as important areas of consideration in DEQ's biocriteria assessment methodology. Those topics were explored in this proposed methodology update.

Use of a single sample for assessment

There is uncertainty inherent in applying bioassessment tools to determine impairment of the biocriteria WQS because true exceedance of the criteria cannot be known, but rather must be determined based on a representative sample or collection of samples. In the assessment of other parameters in the five-year IR data window, DEQ typically uses a minimum sample size of five to determine attainment status of waterbody for conventional pollutants which are easier to obtain and two for toxics pollutants which are more costly to obtain. For biocriteria assessments that use biological data as a direct measure of beneficial use support, DEQ has previously based these decisions on a single sample, with the justification being that biological samples are representative of stream conditions over time (i.e., biological assemblage composition is largely considered integrative of stressors and disturbances). To continue this approach, DEQ explored the consistency of assessment conclusions for monitoring locations with multiple samples.

Sample variability in bioassessment for the Integrated Report can be divided into three groups: sampling and analysis variability (e.g. differences in conclusions from same day results based on field and lab quality control duplicates), seasonal variability (e.g. variability at a site throughout the year), and annual variability (e.g. variability in samples collected from the same site in different years). To investigate consistency of category assignments, DEQ looked at these three sources of variability at monitoring locations with two samples in a five-year window. DEQ found that on average seventy percent of the sample pairs were assigned consistent IR Categories of 2, 5 or 3 (both C and B). Additionally, less than one percent of the sample pairs shifted between impairment and attainment (Table 2). The low rates of assessment error between attainment and impairment may support the use of a single sample for assessment. However, based on the rates of consistent category assignment in this analysis coupled with the recommendations from both the 2018 and 2024 peer review panels, DEQ is now proposing to require two or more samples in a five-year period to classify waterbodies as either impaired or attaining.

Table 2. Analysis of rate of consistency in Integrated Report Category assignment for two samples collected at the same location with different amounts of time between sample collection.

Amount of time separating 2 samples from the same location	n	Percent Consistent Category Assignment (5, 2 or 3)	Shift between Attaining and Impairment
Same Day	191	73%	0.50%
Same season	42	71%	0%
1-5 years	52	65%	0.50%

Balancing assessment error

A key consideration in developing assessment methodologies for the Integrated Report is understanding how Type I and Type II errors factor into categorical determination. In the case of 303(d) assessment, the determination of impairment and/or attainment can be thought of as a management problem, wherein Type I error (false positive) occurs when an attaining assessment unit would be misidentified as impaired (Category 5), and Type II error (false negative) occurs when an impaired assessment unit would be falsely identified as attaining (Category 2). From a management perspective, Type I errors can be costly and resource intensive to address within the Clean Water Act framework, while Type II errors may result in a failure to fully protect the beneficial use. Traditionally, Type I and II error rates are quantified when evaluating a binary decision such as attainment vs. impairment. In this case, there are two key elements of DEQ's new proposed assessment methodology that reduce instances of false positives and false negatives in assessment conclusions but complicate the quantitative reporting and balancing of traditional Type I and II error rates.

To estimate error, "true" waterbody condition must be assumed at a subset of sites. False positive (Type I) error in this case was estimated by exploring instances where a waterbody is identified as impaired when in fact it is attaining. The distribution of MMI and O/E index values for Reference sites were used to calculate false positive error because of the high degree of confidence in the ecological integrity of those sites (ODEQ 2022). False negative (Type II) error in this case was estimated by exploring instances where a site is identified as attaining, when in fact it is impaired. A subset of "highly degraded" sites within the Most Disturbed site class were used to estimate Type II error. Highly degraded sites were identified based on a combination of two factors: (1) high scores in key reference screening disturbance metrics and (2) independent measures of impairment. Three key disturbance metrics identified in DEQ's updated Reference Condition Approach were urban development, mines, and gravel mines. High scores in any of those metrics qualified sites for the "highly degraded" subset so long as it also met the independent measure of impairment. Independent measure of impairment was defined by sites that were identified as impaired in at least one independent aquatic life use assessment (temperature, pH, DO or toxics) according DEQ's 2024 Integrated Report. A subset of 85 "highly degraded" sites were identified in this way and used to estimate Type II error. Assessment error is considered in this section by examining two elements of DEQ's methodology: (1) the two benchmark approach and (2) the hybrid assessment framework.

The two benchmark approach

In adopting two assessment benchmarks for each index, one benchmark is identified to determine impairment of the biocriteria and another to determine attainment. This creates a middle ground between the two benchmarks, similar to the "fair" category that is commonly described when reporting on bioassessment condition estimates. The fair category in this case describes sites that fit into DEQ's "Category 3C: insufficient data; non-Reference condition" (Figure 4). The inclusion of Category 3C in the assessment allows for a more refined management response, as sites identified in this way are likely good candidates for follow up monitoring and/or restoration efforts to improve conditions. From a technical standpoint this third decision category complicates the traditional Type I and Type II error concept, as it introduces the possibility of correctly identifying an attaining site as not being impaired, but not correctly identifying it as attaining. Regardless of this complicating factor, a simple way to explore assessment error rates for each index using the two benchmarks is examine false positive and false negative assessment errors.

Type I error rates for each index were calculated based on the proportion of the Reference site distribution that falls below the impairment benchmark (7.1% for O/E and 3.6% for MMI) (Table 3). Type II error was calculated based on the proportion of the "highly degraded" site distribution that falls above the attainment benchmark (15.6% for O/E and 10.6% for MMI) (Table 3). The error rates reported here are specific to each index and do not reflect error in final assessment conclusions, which are based on the hybrid framework combining both indexes. These results show that the MMI has lower error rates (~ 1/3 lower) than the O/E model for both types of error. It should also be noted that error rates are not equivalent between false positives and negatives for both models, with false negatives occurring more than twice as frequently as false positives. These error rates also help illustrate the benefit of adopting separate benchmarks for attainment and impairment, as the middle ground between the two benchmarks serves as a buffer between management conclusions (Figure 5).

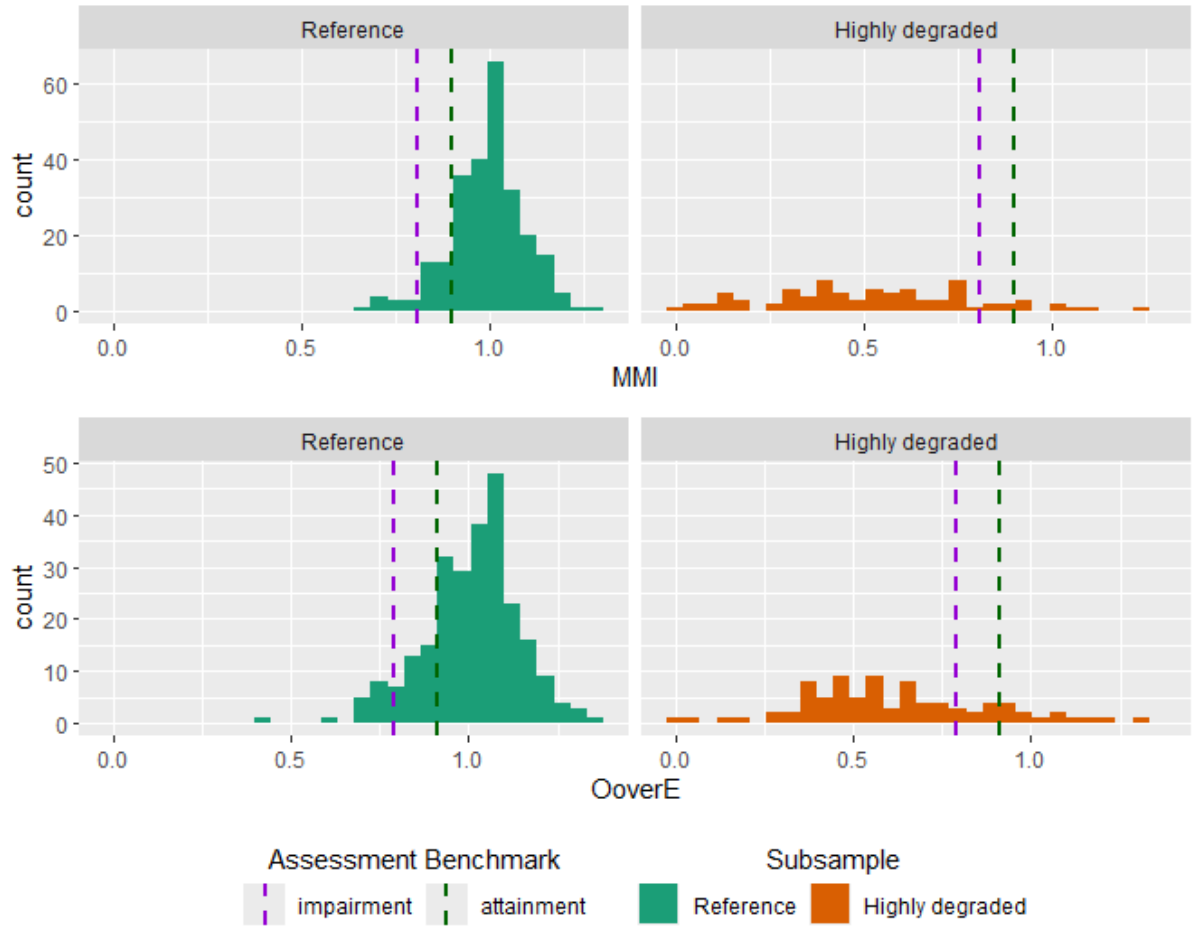

Figure 5. Samples from reference sites to the left of the impairment benchmarks (purple lines) determine Type I error rates for each index. Samples from highly degraded sites to the right of the attainment benchmarks (green lines) determine type II error rates for each index.

The hybrid assessment framework

The effect of the hybrid framework on Type I and Type II error was explored by comparing the categorical determinations of each model individually to those based on the hybrid assessment framework (Figure 4). Reference and highly degraded subsamples are again used here to illustrate how the use of two lines of evidence (two indexes together) reduces instances of false positives (Type I) and false negatives (Type II) in assessment conclusions. Table 3 illustrates how assessment conclusions based on the hybrid framework result in the lowest rates of false positives (2%) when compared with each model individually (MMI: 3.6% & O/E: 7.1%). Additionally, the conclusions from the hybrid framework result in the lowest false negative rate (8.2%) when compared to conclusions from each model individually (MMI: 10.6% & O/E: 15.3%). The low rate of false positive error demonstrated in this analysis of the hybrid framework is encouraging, given the management implications associated with identification of impaired waters.

Table 3: False positive and false negative rates in assessment error are calculated based on instances of reference sites being classified as impaired (Category 5) and highly degraded sites being classified as attaining (Category 2). The effect of assessment type on categorical determination and resultant error rates are illustrated below.

Subset of sites	n	Type of Error	Assessment type	Categories				Assessment Error Rate
				5	3B	3C	2	
Reference	253	False Positive (type I error)	O/E model alone	18	N/A	31	204	7.1%
			MMI model alone	9	N/A	31	213	3.6%
			Hybrid framework	5	17	42	189	2.0%
Highly degraded	85	False Negative (type II error)	O/E model alone	64	N/A	8	13	15.3%
			MMI model alone	72	N/A	4	9	10.6%
			Hybrid framework	61	14	3	7	8.2%

Per the recommendations from the 2018 and 2024 peer review panels, DEQ considered the effects of sample size and error rates on assessment conclusions. After reviewing repeated sampling over multiple timeframes (same day, same season, and five-year window) together with reviewer recommendations, DEQ is proposing to increase the minimum sample size for assessment to two samples. This will reduce the influence of sampling variability on the assessment process and increase confidence in assessment decisions. To review error rates of this assessment methodology, DEQ analyzed assessment false positive and false negative error rates as they relate to the two benchmark approach and hybrid assessment framework. Reference sites misidentified as impaired were used to estimate false positive rates (Type I error) and a subset of "highly degraded" sites misidentified as attaining were used to estimate false

negative rates (Type II error). DEQ examined the effect of two assessment benchmarks in terms of error rates at the index level, and the effect of the hybrid framework in terms of error rates of final categorical outcomes. Based on these reviews, DEQ believes that identifying two assessment benchmarks (impairment and attainment) and the use of the hybrid assessment framework are practical and effective approaches to minimize the likelihood of mis-identifying impaired sites as attaining and attaining sites as impaired.

Updated Assessment Methodology

Water quality standards

340-041-0011 Biocriteria - Waters of the State must be of sufficient quality to support aquatic species without detrimental changes in the resident biological communities.

Assessment methodology – freshwater

Detrimental changes in resident biological communities are a form of pollution.[1,2] EPA guidance recommends using biological community assessments as an indicator for aquatic life beneficial use support.[3] DEQ uses the method described here to implement Oregon's narrative standard for biocriteria in freshwater by assessing the conditions in biological communities. However, the assessment methodology does not identify specific pollutants as potential causes of impairment, which is outside of the scope of the methodology. EPA guidance recommends listing waters with aquatic use impairments as Category 5: 303(d) even if the pollutant is not known.[4]

This method is based on biological community information for freshwater macroinvertebrates at Reference sites throughout Oregon. Freshwater macroinvertebrates include insects, crustaceans, snails, clams, worms, mites, etc. DEQ used updated procedures to identify sites that are least disturbed by anthropogenic activities and uses these sites as Reference sites. DEQ's updated biological assessment tools use information from these Reference and Most Disturbed sites coupled with environmental predictors to set expectations of intact benthic communities in the

[1] Federal Water Pollution Act Section 502(19) (33 U.S.C 1362) (Clean Water Act)
[2] Oregon Administrative Rules 340-041-0002(39)
[3] US EPA, July 29, 2005, Guidance for 2006 Assessment, Listing and Reporting Requirements Pursuant to Sections 303(d), 305(b) and 314 of the Clean Water Act, page 41.
[4] US EPA, July 29, 2005, Guidance for 2006 Assessment, Listing and Reporting Requirements Pursuant to Sections 303(d), 305(b) and 314 of the Clean Water Act, page 60.

waters.[5] The method applies numeric benchmarks to evaluate the integrity of aquatic biological communities.

Data evaluation

DEQ will use two indexes to assess biological integrity of macroinvertebrate communities in smaller wadeable streams. First, DEQ's updated Observed over Expected (O/E) predictive model uses natural environment predictors (not influenced by human disturbance) to determine the most appropriate reference sites to set expected taxa at a sample site. The reporting index O/E, which is the ratio of **observed** taxa at sample site to taxa **expected** if the site was in reference condition, represents loss of native expected reference taxa richness. Second, DEQ's new predictive Multi Metric Index (MMI) uses Reference and Most Disturbed sites to characterize ecological structure and function based on four individual metrics used in the index. Both models cover all of Oregon except for the Northern Basin and Range ecoregion 80 in the southeast corner of the state due to lack of applicable Reference sites. Waterbodies in this region will be assessed by best professional judgement until a bioassessment tool for this region is developed.

DEQ will use numerical assessment benchmarks for each index to interpret the narrative biocriteria water quality standard. The benchmarks are statistical-based percentiles of index values for Reference calibration samples used to build the models, with the 10^{th} percentile representing the impairment benchmark and the 25^{th} representing the attainment benchmark. The benchmarks for the two indexes will be used as multiple lines of evidence via the hybrid assessment framework (Figure 4).

Data requirements

For DEQ to evaluate data using this assessment methodology, the data must meet the following specifications and data quality requirements:

- At least two samples are available at the assessment unit level for river and stream units or monitoring location level for watershed assessment units;
- Samples must be collected during or after 1998 to be comparable to the Reference site data used to build the models;
- Samples must be collected within the model index period of June 1 through October 15;

[5] Stoddard, J. L., Larsen, D. P., Hawkins, C. P., Johnson, R. K., & Norris, R. H. (2006). Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological applications, 16(4), 1267-1276. https://doi.org/10.1890/1051-0761(2006)016[1267:SEFTEC]2.0.CO;2

- Samples must be collected using standard field methods and identified to appropriate taxonomic levels as described in the DEQ Mode of Operations Manual or equivalent protocols used throughout the Pacific Northwest; [6]
- Samples are collected from wadeable streams;
- Samples are collected from riffle habitats or using transect methods (multi-habitat samples that have been shown to be equivalent to riffle samples);
- Samples must contain a total abundance greater than 300 organisms;
- Waterbodies must be similar enough to the Reference population (outliers excluded from routine use of O/E and MMI indexes include all large (non-wadeable) rivers, sites in ecoregion 80 and glacier runoff dominated streams). This may be determined by best professional judgment of assessment and biological monitoring staff.

Data from macroinvertebrate samples collected by organizations other than DEQ may be considered for the assessment and will be evaluated using the DEQ's bioassessment tools, if all DEQ data quality objectives, file formats, and taxonomic consistency are acceptable.

An average index value will be calculated for all valid samples in an assessment unit or monitoring location in a watershed assessment unit for the period of record. The average value will be used for comparison to the applicable assessment benchmark.

Assignment of assessment category

Following the assessment flowchart (Figure 6), index scores will be compared to assessment benchmarks and assigned categories based on the hybrid framework.

Category 5: water quality limited, TMDL needed (303(d) list)

Waterbodies with two or more macroinvertebrate samples that meet the data requirements and the average MMI and O/E index values are less than or equal to both the impairment assessment benchmark values.

Category 4: water quality limited, TMDL not needed

Where DEQ has information relating specific pollutants to impaired biological conditions in the waterbody, a TMDL can be developed. Where data are available for specific pollutants identified as causing detrimental changes to biological communities, and TMDLs have been approved with load allocations for all the pollutants, the waterbody will be placed in Category 4A if no additional TMDLs are needed. Waterbodies will also be placed in Category 4C for biological

[6] DEQ, 2009, Mode of Operations Manual, Version 3.2, DEQ03-LAB-0036-SOP

criteria if adequate information is available to indicate that detrimental changes to biological communities are due to pollution and not a pollutant.

Category 3C: insufficient data; non-Reference condition

Assessment units identified as Category 3C: Potential Concern refer to assessment units that are neither impaired nor equivalent to Reference conditions and may reflect minimal disturbance. These are likely to be the sites that may be the easiest to reverse the impairment through restoration and best management practices in the watershed.

Waterbodies will be assigned this category when the average of at least one index value falls between the assessment benchmarks for that index but neither falls below the impairment benchmark.

Category 3B: insufficient data; potential concern

The average of one index value is less than or equal to the impairment assessment benchmark and the average of the other index value is greater than the impairment benchmark. Waterbodies in this Category will be recommended for additional monitoring due to the uncertainty to attainment status.

Category 3: insufficient data to determine whether a designated use is supported

Waterbodies with just one sample, are not adequately represented by the population of Reference sites, or have low counts less than 300 total abundance.

Category 2: Attaining

Waterbodies with macroinvertebrate sampling data that meets the data requirements and both MMI and O/E index values are greater than the attainment assessment benchmarks.

Delisting – new data

For the 2026 IR cycle, DEQ will assess all the high-quality macroinvertebrate data in AWQMS going back to 2000 with this updated methodology. Existing biocriteria impairments will be removed from the 303(d) list when the assessment units are assigned Category 2, 3C or 3B with the rationale of new assessment methodology applied (Attains Code: WQS_NEW_ASMT_METHOD).

For previous 303(d) listings based on a single sample, DEQ will not delist if both of the index values are below the 5th percentile of Reference calibration samples used in model development (O/E = 0.75 and MMI = 0.77). For previous 303(d) listings based on a single sample that are

eligible for delisting, the impairments will be removed from the 303(d) list and placed in Category 3 (insufficient data) until data requirements are met for assessment using the updated methodology.

After the 2026 IR cycle, waterbodies may be delisted for biocriteria based on multiple site sampling events showing results that are attaining benchmarks. A minimum of two samples in different years within the most recent five-year time-period must be collected in the same sampling season and assessment unit (or waterbody for watershed type units), with all samples showing results that attain appropriate benchmarks. These waterbodies will be placed in Category 2: Attaining.

Other approaches to assess biological integrity in freshwater

While this methodology is DEQ's preferred approach and provides the most robust and contemporary method for assessing biological integrity in smaller, wadeable streams and rivers, other approaches may be appropriate and used for specific cases and datasets. For example, in studies examining the effects in non-wadeable rivers and/or of point-sources, study designs may look at upstream-downstream changes in macroinvertebrate community composition and function.

Assessment Flowchart

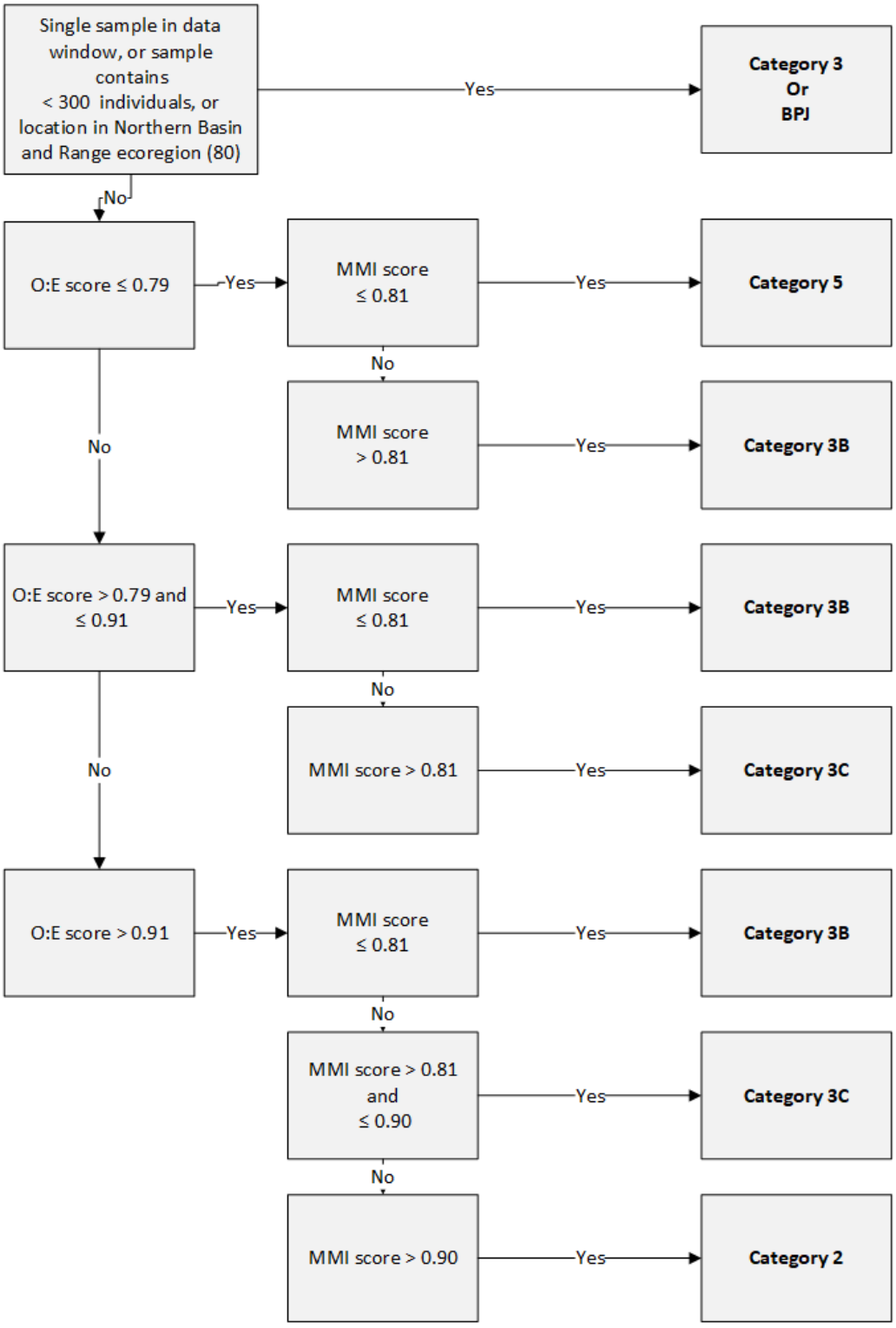

Figure 6. Assessment flowchart for the 2026 freshwater biocriteria update.

Conclusion and future directions

DEQ is proposing an updated assessment methodology for interpreting the narrative biocriteria water quality standard. The updated method uses the new and improved bioassessment tools developed at the DEQ laboratory, relates assessment benchmarks with ecological function, uses a hybrid assessment framework to incorporate multiple lines of evidence, and increased the minimum sample size needed to make an attainment decision. All of these updates will increase the confidence in using biological data to assess the narrative biocriteria water quality standard in freshwater streams that are adequately represented by the population of streams in DEQ's updated Reference Condition Approach.

For waterbodies that are not represented by the population of streams in Reference Condition, it is entirely appropriate for different bioassessment tools and approaches to be used to validate or refute a biocriteria listing. However, DEQ reserves the right to review the assessment tool for methodological and statistical rigor and may or may not approve of its use in making an impairment determination. In addition, DEQ is authorized to use other methods of evaluation to assess organism condition, or other ecosystem attributes relevant to biocriteria, so long as natural background conditions can be established to determine whether an impact is taking place outside of natural ecosystem variability. Future work may include expanding the interpretation of the narrative biocriteria to include Northern Basin and Range ecoregion, regions of the state, larger rivers, lakes and estuaries for use in future assessment cycles.

Literature Cited

Hargett, E.G., ZumBerge, J.R., Hawkins, C.P. and Olson, J.R., 2007. Development of a RIVPACS-type predictive model for bioassessment of wadeable streams in Wyoming. Ecological indicators, 7(4), pp.807-826.

Hawkins, Charles P., Richard H. Norris, James N. Hogue, and Jack W. Feminella. 2000. "Development and evaluation of predictive models for measuring the biological integrity of streams." *Ecological applications* 10, no. 5 (2000): 1456-1477.

Hawkins, Charles P. 2009. "Revised invertebrate RIVPACS model and O/E index for assessing the biological condition of Colorado streams." *Prepared by Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed Sciences, Utah State University for Colorado Department of Public Health and Environment, Water Quality Control Division–Monitoring Unit*.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R. and Thornbrugh, D.J., 2016. The Stream-Catchment (StreamCat) Dataset: A database of watershed metrics for the conterminous United States. JAWRA Journal of the American Water Resources Association, 52(1), pp.120-128.

Hubler, S.L., 2008. PREDATOR: Development and use of RIVPACS-type macroinvertebrate models to assess the biotic condition of wadeable Oregon streams. Oregon Department of Environmental Quality, Laboratory Division, Watershed Assessment Section.

Hubler, S., Stamp, J., Sullivan, S.P., Fernandez, M., Larson, C., Macneale, K., Wisseman, R.W., Plotnikoff, R. and Bierwagen, B., 2024. Improved thermal preferences and a stressor index derived from modeled stream temperatures and regional taxonomic standards for freshwater macroinvertebrates of the Pacific Northwest, USA. *Ecological Indicators*, *160*, p.111869.

Mazor, Raphael D., Andrew C. Rehn, Peter R. Ode, Mark Engeln, Kenneth C. Schiff, Eric D. Stein, David J. Gillett, David B. Herbst, and Charles P. Hawkins. 2016. "Bioassessment in complex environments: designing an index for consistent meaning in different settings." Freshwater Science 35, no. 1: 249-271.

ODEQ. 2022. Oregon DEQ's Reference Condition Approach (RCA – 2020): An updated approach to defining least disturbed areas. Document ID: DEQ22-LAB-0047-TR

Paul, Michael J., Ben Jessup, Larry R. Brown, James L. Carter, Marco Cantonati, Donald F. Charles, Jeroen Gerritsen et al. 2020. "Characterizing benthic macroinvertebrate and algal biological condition gradient models for California wadeable Streams, USA." Ecological Indicators 117: 106618.

Stamp, J. 2022. Calibration of the Biological Condition Gradient (BCG) for Macroinvertebrate Assemblages in Freshwater Wadeable Streams in the Pacific Northwest Maritime Region of Oregon and Washington. Prepared for US EPA Office of Water and US EPA Region 10.

Vander Laan, Jacob J., and Charles P. Hawkins. 2014. "Enhancing the performance and interpretation of freshwater biological indices: an application in arid zone streams." Ecological Indicators 36: 470-482.

Technical Appendix A. Index development

This appendix describes how DEQ applied standardized procedures to develop Oregon specific RIVPACs and MMI predictive models.

RIVPACS-type O/E index

DEQ's most recent bioassessment model, PREDATOR, was in fact a RIVPACS-type O/E index. To avoid confusion, DEQ has decided to use "O/E" in referring to the updated O/E index and will no longer use "PREDATOR" to refer to the O/E index.

Step 1. Define the Reference site population to set expectations

RIVPACS-type methods rely exclusively on Reference sites for developing O/E models. DEQ screened available macroinvertebrate samples from the pool of available Reference sites, allowing only samples that met the following criteria:

- Time: sampled from 1997 onwards (this represented significant changes in sampling and sorting protocols)
- Habitat: riffle or transect samples
- Index period: samples must have been collected between June and October
- Abundance: total abundance greater than 200 individuals
    - DEQ used 25 (out of 221) samples with less than 300 individuals (the typical target).
    - These lower count reference samples were used because these sites were located in regions that would otherwise be less represented in the final indexes.
    - DEQ believe this tradeoff in spatial representation outweighs the potential in reduced modeling performance due to lower counts. In fact, O/E and MMI scores in these lower count samples showed similar results to higher count samples.
- A single sample from each site:
    - preference was given to samples used in previous O/E indexes (formerly PREDATOR, now "O/E v1.0") to maintain a record of historical Reference conditions in the face of changing climate
    - For newly identified Reference sites, older samples were preferred if they met the above criteria

Following the screens above resulted in 316 samples from unique Reference sites. However, many Reference sites were not spatially distinct, in that they were associated with other Reference sites on the same stream. To avoid biasing model predictions towards individual streams, DEQ allowed for only one Reference site per stream, unless the sites were located in distinct Assessment Units as defined the IR Assessment Methodology. Combined, all of these screens resulted in 265 unique Reference sites for initial O/E modeling.

Step 2. Define environmental predictors

To allow for the greatest ease of applying these models across organizations and entities, DEQ chose to use predictor data from widely available databases associated with digitized stream networks. All predictors were selected from the USEPA StreamCat database (Hill et al 2016) and the NHD Plus. DEQ limited our initial set of StreamCat predictors to only those metrics considered to represent natural gradients. We primarily relied on Watershed-scale metrics from StreamCat. However, for sites that did not fall on the NHD medium-resolution stream layer (thus without a COMID and unable to be associated with StreamCat), a COMID was manually assigned to the site based on the nearest COMID and we selected Catchment-scale metrics. Stream slope estimates were sourced from the NHD. Predictor metrics that were incomplete or highly correlated with other predictor metrics ($r > 0.89$) were dropped.

Step 3. Random Forest modeling

- Assign macroinvertebrate data to Operational Taxonomic Units (OTUs)
    - No ambiguous taxa allowed
    - That is, consistently rolled taxonomic identifications up to a common level (e.g., all species within a genus were changed to the genus) or dropped less-resolved taxa (e.g., family level IDs were dropped if genus or species within the family IDs were retained).
- Randomly subsample to 300 count
- Only Reference samples with at least 200 count were used to build the models (n = 221)
- Clustering—based on Bray Curtis dissimilarity between Reference macroinvertebrate samples
    - Associate samples with biologically similar groups
    - Rule: must have at least 10 Reference sites within each group
- RF models – using all predictors
- RF models – using reduced set of predictors
    - Variable Importance Plots
- Final model selection
    - Mean ref O/E ~ 1.0

- Lowest standard deviation in ref O/E
- Greatest difference between modeled and null SDs
- Closest SD to replicate sampling error SD

DEQ used standard RIVPACS model development methods and R-code provided by Utah State University. Macroinvertebrate data preparations included assigning all taxonomic identifications to Operational Taxonomic Units, so that ambiguous taxa were excluded; then samples were randomly subsampled to a maximum of 300 count. Only Reference sites with at least 200 count were retained for modeling. Modeling used cluster analysis and Random Forest modeling. Clustering, using Bray-Curtis dissimilarity, was based on macroinvertebrate data only, with a requirement for at least 10 Reference samples within each Reference group. Radom Forest models were used as a means to identify Reference group membership probabilities, based on a set of predictors representing only natural gradients. DEQ first used all predictors and identified the most important set of predictors in identifying group membership. Then a Random Forest model was built using this reduced set of important predictor variables. Successful reduced models showed minimal reductions in predictive power of the full models, while protecting against over-fitting. The final model chosen was based on the following criteria (in order of importance):

1. mean Reference sample O/E ~ 1.0
2. lowest standard deviation in Reference O/E
3. greatest difference between modeled and null-model (no predictors) standard deviations
4. Reference standard deviation closest to replicate sampling standard deviation

During the model exploration phase, several Reference sites acting as outliers (low O/E) were identified. As a result, 44 Reference sites were dropped due to the following causes:

- Northern Basin and Range ecoregion: just as in model building for RIVPACS 1.0 (PREDATOR), inclusion of sites from Southeast Oregon resulted in poor overall model performance, due to low total richness in this region. Mean "E" (expected taxa richness) for these sites was 5 taxa, making accurate predictions difficult.
- Glacially influenced sites: these sites routinely result in low O/E values, suggesting they are not modeled accurately. This has implications on applying O/E models to these stream-types.
- Poor taxonomic resolution: samples with high levels of individuals identified to low-resolution taxonomy (e.g., family or order) were dropped because most OTUs were at the genus or species levels.

The final model building phase included 221 unique Reference sites and 44 candidate predictors. The final model chosen had the following specifications:

Table A-4 Final RIVPAC 2.0 model specifications.

	Full model	Reduced model	Null model
# of Reference groups	8	8	0
# of predictors	44	6	0
Mean O/E	1.03	1.01	1.00
Standard Deviation	0.161	0.162	0.173
Replicate sampling error	0.131	0.128	0.141

The six predictors chosen for the final model included:

- TMAX8110 (30-year Average Annual Normal Maximum Air Temperature)
- ELEV (Mean Elevation)
- MWST_mean08.14 (Mean Winter Stream Temp, averaged across 2008 – 2014)
- CLAY (Mean % clay content of soils)
- PRECIP8110 (30-year Mean Annual Precipitation)

Step 4. Model Comparison

DEQ compared O/E results from this recent version (RIVPACS v2.0) to our previous model (PREDATOR, hereafter O/E v1.0). It was expected that O/E would be relatively similar between the two models, given that a substantial portion of the samples were used in both versions of the model. Moderately disturbed and Most disturbed sites closely followed the 1:1 line for v1.0 and v2.0 O/E values (Figure A-7). Reference showed general agreement between v1.0 and v2.0 O/E values; however, low Reference O/E scores were noticeably higher in v2.0, possibly suggesting improved predictions for the Reference population.

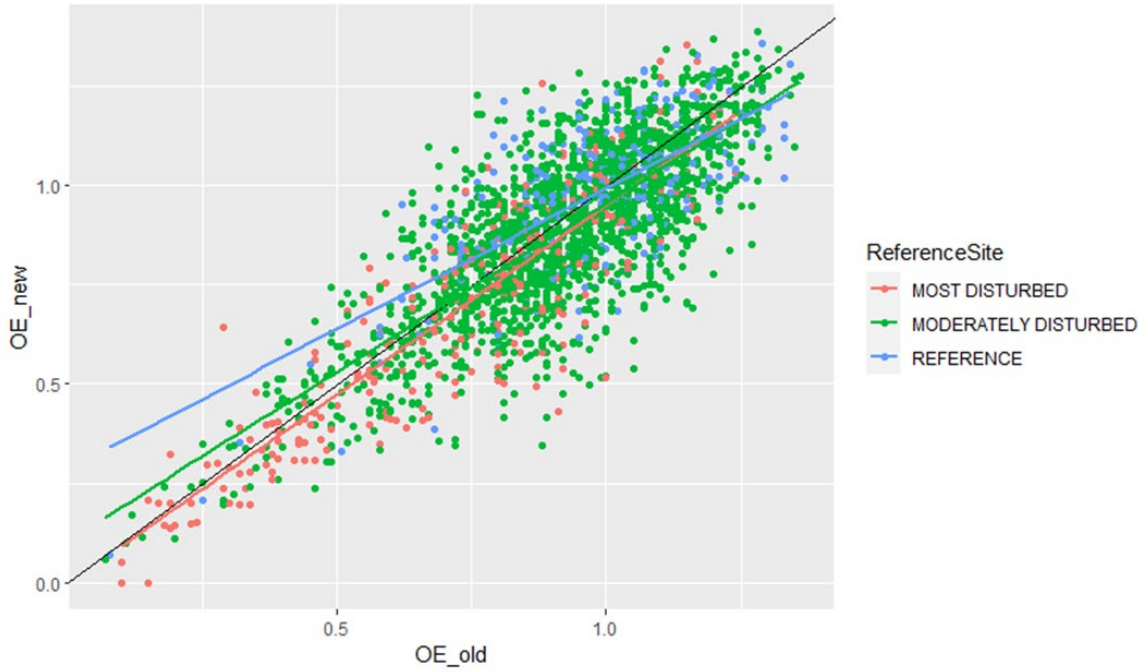

Figure A-7. Comparison of RIVPACs v1.0 and v2.0.

DEQ also compared O/E values from RIVPACS v2.0 across disturbance classes and Level II ecoregions (Figure A-8). Reference O/E scores in the Marine Western Coastal Forest (MWCF) and Western Cordillera (WC) were centered around 1.0 with smaller deviations than observed in other disturbance classes. Most disturbed sites tended to show lower O/E values in these two ecoregions than compared to other disturbance classes. O/E values in the Cold Desserts ecoregion were substantially lower than observed in the other regions—even for Reference sites. As such, DEQ does not recommend applying the RIVPACS O/E model to SE OR sites.

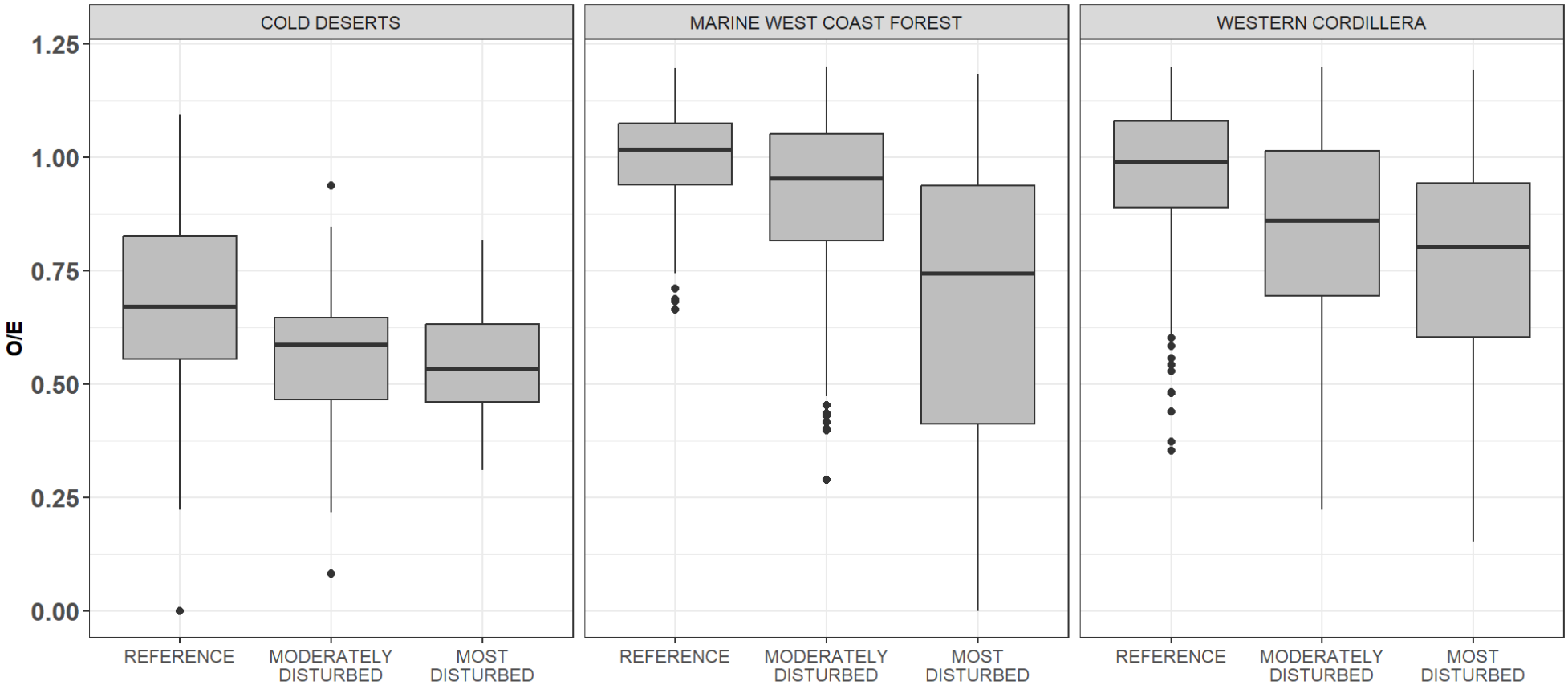

Figure A-8. Box plots of RIVPACS v2.0 index values at different ecoregions.

Multi Metric Index

With the goal of using multiple lines of evidence in assessing freshwater biocriteria, DEQ developed its first predictive macroinvertebrate MMI. This section describes how DEQ adapted the MMI models for use in Oregon, using standard model development procedures (Vander Laan and Hawkins 2014, Mazor et al. 2016).

Step 1. Define Reference and Most Disturbed sites

Unlike RIVPACS models, which are based exclusively on Reference sites, MMIs use both Reference and Most Disturbed sites. DEQ used the same 221 Reference sites used to build the RIVPACS v2.0 model. Additionally, 158 Most Disturbed sites were used.

Step 2. Calculate sample metrics

Metrics were calculated for each sample using the BiomonTools R-package, provided by Tetra Tech. Ten metrics had zero values, so they were dropped, leaving 474 candidate metrics.

Step 3. Define environmental predictor to set expectations

DEQ used the same 44 StreamCat and NHD metrics for initial MMI modeling as defined for the RIVPACS v2.0 model. Similarly, Catchment-scale StreamCat metrics were again used for sites without COMIDs.

Step 3. Modeling

DEQ followed standard MMI modeling techniques, with guidance from USU NAMC staff on certain modeling choices. Prior to modeling, all macroinvertebrate samples were randomly subsampled to a 300 count target. We performed Random Forest (RF) modeling on all 474 metrics, retaining modeled metrics only for those models that explained 10% or more of the variability in metric values. For metrics with poor models (< 10% variability explained), we moved forward with the unmodeled metrics or those derived from the sample. We calculated residual metric values (observed metric – expected metric). T-tests were used to determine the level of significance between metric values for Reference and Most Disturbed populations. Principal Components Analysis (PCA) was used as a form of correlations analysis to see which metrics were most related to each other.

The final MMI was chosen as such: First, the absolute PCA values were used to rank metrics with the strongest associations to each PCA axis. We selected the top metrics for each PCA axis with a threshold of 0.7 or more for PCAs 1-4, or 0.6 for PCA 5. Next, we compared t-values for each of the top metrics with each axis. Optimally, we would use the metrics with the highest t-value for

each PCA axis, but we also had the goal of ecological independence for each of the final metrics chosen. For example, in one of our two candidate final models, the top metrics (highest t-value) in PCA 4 and PCA 5 were taxonomy-based metrics. In this case, we chose a tolerance-based metric for PCA 4 and a taxonomy-based metric for PCA 5, based on the smaller difference in t-values for the top two metrics in PCA 4. DEQ calculated MMI scores for the two candidate MMI models by re-scaling each metric to values between 0 – 1.0, then averaging across each of the chosen metrics. The final MMI model was chosen using t-tests of MMI scores to determine which model best discriminated between Reference and Most Disturbed sites.

The final MMI selected included four RF-modeled metrics: % intolerant taxa, # of rheophilic taxa, % cold-cool water taxa, and % EPT individuals—without Hydropsychidae included (Table A-2). Mean MMI scores for Reference sites was 0.73, while the mean MMI scores for Most Disturbed sites was 0.46. To check for potential bias in the modeled MMI, we ran a RF model of the final MMI scores against a suite of 28 natural (not influenced by human activities) StreamCat predictors, using Reference sites only. The results of this test showed no natural bias in Reference site MMI scores (-27% variance explained). Figure A-9

Table A-5. MMI metric random forest model results.

Metric	Metric description	Metric source	Variability explained by RF	Predictors
pt_tv_intol	% intolerant taxa	HBI	36%	TMAX8110, CLAY, OM, KFFACT, PRECIP8110
nt_habitat_rheo	# rheophilic taxa (preferring fast-flowing water)		18%	TMAX8110, ELEV, OM
pt_ti_stenocold_cold_cool	% taxa with thermal preferences for very cold to cool water	Hubler et al. 2024	34%	MSST_mean08.14, KFFACT, TMAX8110, PERM, CLAY, AREASQKM, SLOPE
pi_EPTNoHydro	% individuals from Ephemeroptera, Plecoptera, and Trichoptera (minus Hydropsychidae)		12%	PRECIP8110, KFFACT, CLAY, P205

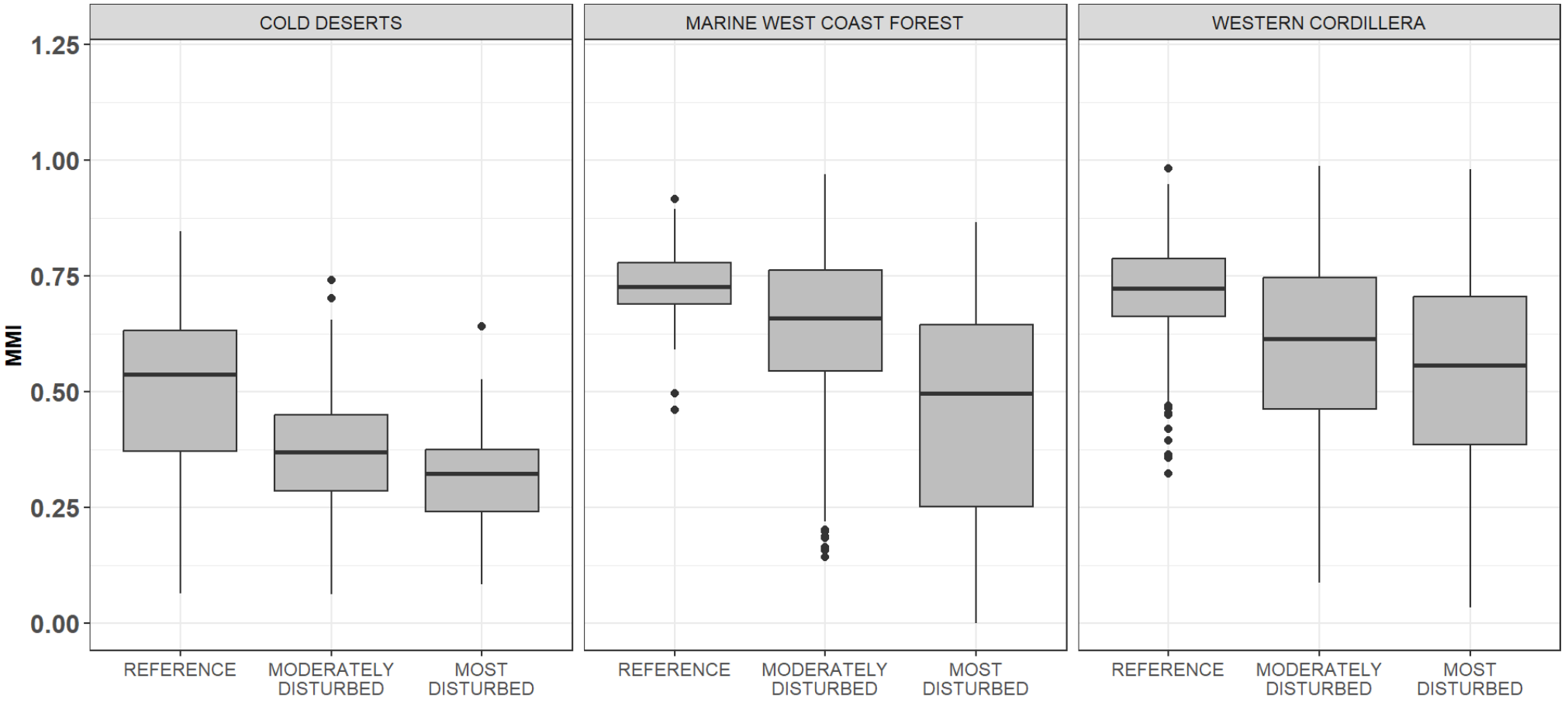

Figure A-9. MMI model box plot of Reference and Most Disturbed sites