5.3 Index compilation and performance

Index compositions were formulated from the best performing metrics in each metric category. The metrics were combined by scoring each on the 0 to 100 scale and then averaging the scores. Each index alternative was then evaluated for discrimination efficiency and other measures of representativeness and sensitivity. Index formulations were created and evaluated in two ways: automatic all-subsets modeling and manual metric substitutions.

The all-subsets analysis allowed consideration of a plethora of diverse index compositions that simply could not be computed by hand. Nineteen candidate metrics were selected for inclusion in index trials based on DE, *Z*-score, and professional opinion of the working group. An "all subsets" routine in R software (R Core Team 2020) was used to combine up to 10 metrics in multiple index trials. Each of the index alternatives was evaluated for performance using DE, *Z*-score, number of metric categories, and redundancy of component metrics. Those models including two or more correlated metrics (Spearman $|r| \geq 0.80$) were excluded from consideration. As many metric categories as practical were represented in the index alternatives so that signals of various stressor-response relationships would be integrated into the index. While several metrics should be included to represent biological integrity, redundant metrics can bias an index to show responses specific to certain stressors or taxonomic responses.

The metrics shown in Table 8 were included in the all-subsets analysis. The all-subsets model calculation and screening resulted in thousands of valid index combinations. Initially, the all-subsets analysis resulted in approximately 103,000 different index combinations. To identify the most sensitive, comprehensive, and practical index alternatives, the characteristics of the alternatives were screened for favorable characteristics such as high DEs and representation of multiple metric categories. Metrics with conceptual redundancy and unexplained response mechanisms were excluded. Habit metrics were not preferred because they did not have plainly understandable response mechanisms. To narrow down the long list of index alternatives, two reviewers (Ben Block (Tetra Tech) and James Meek (MassDEP)) were provided an Excel worksheet with results from the all-subsets analysis. The number of index alternatives was reduced to approximately twenty. The screening and exclusion criteria are summarized in Table 10. The resulting subset of index alternatives had similar performance statistics (Table 11), therefore, the final selection process involved subjective decisions on metric preference and performance.

The workgroup decided to pick indices with familiar metrics (composition, functional feeding group (FFG), richness, tolerance, and voltinism). Voltinism metrics were emphasized because they indicate ecosystem stability. Multivoltine taxa are short lived and have multiple generations per year. The presence/abundance of these taxa indicate a system that can experience more variability (e.g., flow) and potentially more disturbance overall. Semivoltine taxa require more than one year to complete their life cycle and thereby tend to require a more stable environment. The workgroup rationalized their choice based on empirical performance and ecological characteristics of the individual and combined metrics. They selected an index that was a top selection for both the MassDEP and SNEP projects. The final choice was Model 6_13784, which included six metrics (Tables 12 & 13).

Table 10. Reviewer screening and exclusion criteria for narrowing the list of index alternatives. Initially, the all-subsets model resulted in over 100,000 alternative index compositions.

Criteria #	Model Elimination Criteria (eliminated models with these criteria)	# Remaining models
1	Contains any Habit metrics	27388
2	Insect/Non-Insect Metrics > 1	22892
3	Contains both pt_EPT and pi_EPT	20995
4	Contains both nt_EPT and pi_EPT	19095
5	Contains both pt_tv_toler and pt_tv_intol	15940
6	Contains both pt_volt_semi and pt_volt_multi	11989
7	Contains no FFG metrics	8990
8	Contains no Tolerance metrics	7910
9	Number of Metrics < 5 OR > 7	5483
10	DE < 100	5032
11	*Z*-Score > -2.25	3740
12	Ref. q25 – Str. q75 < 18	3358
13	Ref. cv > 0.22	1366
14	Contains x_Beck and x_HBI	1224
15	Ref. q10 – Str. q90 < 3	861
16	Contains both nt_CruMol and pt_Amph	600
17	Contains both nt_ffg_pred and pt_ffg_col	425
18	Number of metric categories < 4	369
19	Number of Richness metrics > 2	249
20	Only Richness metrics are nt_CruMol, pt_Amph, or pt_NonIns	130
21	Contains no Composition metrics	95
22	Only Composition metric is pi_NonIns	84
23	Contains nt_CruMol or pt_Amph metrics	33

Evaluation of subsample size

After Model 6_13784 was selected, we performed an additional analysis on the full dataset to evaluate how much the IBI was affected by subsample size since some regional partners may lack sufficient resources to process 300-organisms (instead they may be limited to 200 or 100-count samples). Of particular interest was the effect on the two richness metrics (number of Plecoptera, Odonata, Ephemeroptera, Trichoptera (POET) taxa and number of predator taxa), since the number of taxa found in samples generally decreases with a decrease in the number of individuals collected (Gotelli and Graves 1996). With this consideration in mind, the working group wanted to explore: 1) the magnitude that subsample size affected the two richness metrics vs. the percent taxa versions of those metrics; and 2) if the percent taxa POET and predator metrics were substituted into IBI model 6_13784, did the alternative IBI perform equally well or better (as measured by DE, Z-score, and coefficient of variation (cv)] when using 300, 200, or 100-count samples). Ideally, the working group wanted to select an IBI that not only performed well in both the SNEP and MA/SNEP datasets but also performed well in 100, 200,