Fisheries and Oceans Canada | Pêches et Océans Canada
Ecosystems and Oceans Science | Sciences des écosystèmes et des océans

Canadian Science Advisory Secretariat
Pacific Region **Science Response 2018/053**

EVALUATION OF THE REFERENCE CONDITION APPROACH FOR YUKON PLACER MINING MONITORING

Context

Since 2008, Fisheries and Oceans Canada's Fisheries Protection Program (DFO FPP) and its partners, including the Yukon Government, rely on the Adaptive Management Framework (YPAHWG 2008a) implemented by the Yukon Placer Secretariat to manage gold placer mining activity in the Yukon. Founded on principles of adaptive management and incorporating a risk-based approach to decision making, the Fish Habitat Management System (FHMS) is intended to balance the objectives of a sustainable Yukon placer mining industry with the conservation and protection of fish and fish habitat.

A set of protocols has been designed to guide the FHMS. These are the Aquatic Health Monitoring Protocol (Yukon Placer Aquatic Health Working Group [YPAHWG] 2008b), the Water Quality Objectives Monitoring Protocol and the Economic Health Monitoring Protocol. DFO and the Yukon Government's Department of Environment are responsible for implementing the Aquatic Health Monitoring Protocol designed to assess how effective the FHMS is for maintaining aquatic health for fish and fish habitat, and to generate monitoring results which will be used in the adaptive management framework assessment and adjustment phases. Yukon Government's Department of Energy Mines and Resources are responsible for the other two protocols.

As part of the FHMS, in the 2000's the Reference Condition Approach (RCA) was selected by Fisheries and Oceans Canada and the Yukon Government, in consultation with First Nations and industry, to help assess and monitor aquatic health. This RCA uses benthic invertebrates as an ecosystem indicator of aquatic health, and aquatic health as a surrogate for the health of fish and fish habitat. The Canadian Aquatic Biomonitoring Network (CABIN) was chosen as the biomonitoring program. CABIN provides standardized sampling methods, the sampling protocol, and the data warehouse used to store and analyze the Yukon placer RCA dataset (Reynoldson and Bailey 2013, 2014[1]; Reynoldson et al. 2016). In an effort to improve its reliability, the CABIN model has undergone several revisions, the most recent in 2013. DFO FPP has recently identified concerns about the reliability of the 2013 Yukon model (herein referred to as the 2013 Yukon CABIN model). Given that the RCA results are used to inform adaptive management and to make regulatory decisions under the *Fisheries Act*, it is important that with the RCA findings, DFO FPP is able to confidently determine if the aquatic health of streams exposed to placer mining activity is being maintained or improved over time.

DFO Fisheries Protection Program has requested that DFO Science Branch evaluate the suitability of the Reference Condition Approach, and provide guidance regarding the adequacy of RCA for informing regulatory decisions for placer mining in the Yukon. The advice arising from this Canadian Science Advisory Secretariat (CSAS) Science Response Process (SRP) will

[1] Reynoldson, T.B. and Bailey, J.L. 2014. Reference model supporting documentation for CABIN analytical tools. Unpublished memo to the Yukon Placer Secretariat.

be used to inform DFO FPP on the effectiveness of the RCA model in detecting changes in aquatic health in streams exposed to placer mining activity.

This Science Response Report results from the Science Response Process of July 2018 on the Evaluation of the Reference Condition Approach for Yukon Placer Mining.

Background

Gold placer mining, which is the extraction of gold from alluvial deposits, poses risks to aquatic habitats as a result of the discharge of suspended sediments into watercourses during mining, and the alteration of sediment and geomorphic processes caused by the disturbance of riparian and valley slopes both during and after mining. The Yukon Placer regulatory system is designed to minimize these risks through the establishment of water quality objectives and sediment discharge standards for mines, and guidelines for channel and valley reclamation. Some aspects of the regulatory requirements are set out in watershed-based *Fisheries Act* authorizations.

Benthic macroinvertebrates (BMI) are present in all streams, and their abundance and species composition is sensitive to anthropogenic stressors such as elevated suspended sediment levels or pollution (Seakem Group Ltd., 1992; Mathers et al. 2017). BMI are also relatively simple to sample and analyze, and as a result they have become the most common tool for assessing stream health.

Benthic macroinvertebrates are widely accepted as an indicator of the effects of sediment and sedimentation on stream ecosystems. A growing body of research has identified mechanisms that cause changes in the abundance and composition of invertebrate communities with increased sediment in rivers and streams (e.g., Jones et al. 2012); the presence of causal linkages supports the use of invertebrate metrics to monitor sediment effects. Earlier studies in Alaska and Yukon show invertebrate abundance and diversity are affected by suspended sediment and habitat disturbance associated with placer mining (Seakem 1992).

The Yukon Placer Secretariat has chosen to use BMI to monitor the status of streams exposed to placer gold mining, and uses a study design based on the Reference Condition Approach and the multivariate mode of analysis employed by the Canadian Aquatic Biomonitoring Network called BEAST (BEnthic Assessment of SedimenT). RCA evaluates the status of a site by comparing the BMI community at that site to a suite of reference sites that are not exposed to the stressor or activity of interest. If the exposed site lies outside the range observed for the reference sites then there is evidence for an effect of the stressor.

The **multivariate** method of analysis used in the CABIN program is entirely empirical in that it uses statistics to describe differences in abundance and diversity of BMI of reference and test sites. No causal mechanism is attached to those differences. In contrast, multimetric analyses, which is the dominant approach to bioassessment in the US, EU and increasingly in the UK, use metrics that take advantage of *a priori* knowledge of the response of invertebrate taxa to particular stressors (e.g., Turley et al. 2016). For example, a metric might be the abundance of taxa that are known to be sensitive to a stressor of interest (e.g., sediment), and there is an expectation that metric scores for sites or streams would be related to the magnitude of the exposure of the site to that stressor.

To implement the CABIN approach a reference dataset that consists of single samples taken from streams throughout the Yukon that are not impacted by mining was assembled from datasets from a variety of sampling programs conducted from 2004 to present. Counts of invertebrates (resolved to family) in each sample are then subject to an ordination and clustering analysis that serves to group samples based on their similarity. The Bray-Curtis

metric is used, which is a measure of similarity based on counts of taxa common to pairs of samples. Raw abundance data were used in the 2013 Yukon CABIN model. For the Yukon analysis, samples are classified into five groups based on similarity among samples using cluster analysis (e.g., Strachan and Reynoldson 2014). A series of environmental or habitat variables are then used in a discriminant function analysis (DFA) to develop a predictive model that can be used to assign test (exposed) samples to one of the reference groups based on the match of habitat variables.

Once test samples are assigned to one of the groups a comparison is made between the BMI community at the test site and reference sites belonging to the same group. The status of the test site is evaluated by computing the probability that the sample belongs to that reference group based on its position in a two-dimensional ordination. The location of the test sample relative to probability ellipses around the reference data is used to assess status. Under the adaptive management protocol sites outside the 90% ellipse are considered out of reference.

Analysis and Response

The analysis is based on four objectives defined in the terms of reference.

1. Is the RCA suitable for evaluating the effects of placer mining activity on the aquatic health of fish and fish habitat?

CABIN's BEAST (BEnthic Assessment of SedimenT) model in relation to the AHMP objectives.

The objectives of the Aquatic Health Monitoring Protocol (AHMP) are to evaluate individual streams and watersheds to determine whether their aquatic health differs from streams unaffected by placer activity, and to be able to track the health of individual sites and watersheds over time. There is also a desire to compare results from aquatic health monitoring to water quality objectives and monitoring results.

The adaptive management framework integrates information from the invertebrate monitoring, water quality monitoring and economic analysis to evaluate the performance of watershed authorizations over 3-5 year time frames.

The RCA as implemented in the CABIN approach is designed to evaluate individual sites (which are in fact individual samples) against a collection of reference sites. It has the potential to meet some of AHMP objectives with respect to the assessment of individual sites. There is no direct linkage between the output of the multivariate procedure and impacts associated with placer mining or with water quality objectives.

Methods to simultaneously evaluate multiple samples that would allow an assessment at broader spatial and temporal scales as described in the adaptive management framework have not been developed.

Strengths and limitations of the current approach.

Strengths

The primary strength of the BEAST/CABIN approach to assessing stream health is that it is an entirely empirical approach that circumvents the need to explicitly define "health" in the context of invertebrate communities, or the need to develop indicators or metrics of change that are expected to respond to the presence of stressors. Test sites are simply evaluated on the basis of their similarity or difference relative to the reference sites whereas some metric-based approaches use a subjective assessment of sensitivity to stressors and could result in inconsistent results.

The CABIN sampling protocol is simple and straightforward to implement and the Yukon Reference dataset can be augmented by a variety of users, and it can be used for multiple applications. The breadth of sampling that has been conducted in the Yukon does provide a fuller accounting of the natural variation in BMI communities than what would be obtained from a more localized control-impact study design. The existing CABIN database is an efficient means to manage the data and make them available to a variety of users or applications.

Limitations:

Potential shortcomings of the BEAST approach have prompted this review; these are briefly listed below and are expanded in the subsequent sections.

The approach deviates from standard statistical practise. Invertebrate samples are highly variable in their composition. Although taxon richness may be adequately captured with a single large sample, counts of individual taxa can be highly variable, due to chance or environmental effects within the year, as well as fine-scale spatial variation. The practise of taking a single sample at a single site in each river without consistent interannual and spatial replication reduces the power of the analysis and renders the results vulnerable to sampling error. This may be accentuated through the use of untransformed abundance data for the analysis; in many studies a variance-stabilizing transformation is employed prior to analysis to reduce the influence of extreme values. The sampling design is a challenge for the Adaptive Management Framework since the standard object of analysis in BEAST (a single sample) differs from management objectives that are assessed at the scale of the watershed authorizations, and the analysis of trends at 3-5 year time scales. A hierarchical sampling scheme that replicates samples within sites, sites within larger spatial units (e.g., streams, regions), and repeats sampling in time permits the data to be analyzed in a framework that can account for the various sources of variation and will likely produce more reliable estimates of real changes in the BMI communities, particularly at the stream and watershed scales.

Grouping of reference sites is influenced by chance events and sampling error. In the absence of a hierarchical sampling program, the variable nature of the data means that the grouping of individual sites is based on a combination of common environment or habitat influences and chance events that affect the composition of individual samples. Samples that are dominated by particular taxa, such as Chironomidae, will tend to group together, regardless of whether this is due to random sampling variation, or habitat or spatial factors. Because samples are not replicated there is no way to know if chance variation is contributing to the model being "overfit" to random variation. Overfitting results in poor performance in subsequent analyses as is suggested by the Type I error rates outlined below.

There is no biological interpretation of "out of reference". The BEAST results from individual samples from exposed sites are compared to the ordination of samples from reference sites of the same group, and the test sample is assigned to a probability band based on the empirical distribution of the reference samples. The probability band is a measure of how different the sample is from the reference samples but inference about the biological significance of the difference is not possible within CABIN.

The assessment protocol is inadequate for spatial or trend analysis. The dichotomous test for individual samples (i.e., being in or out of reference) is an inefficient means to perform assessments at larger spatial or temporal scales. Repeat sampling at individual sites shows that both the group assignment and the test results vary from year-to-year, which may be a result of a lack of replication in the data and the potential overfitting of the groups to habitat data. Failure to account for spatial and temporal variation in the reference set and the potential for group assignments to vary annually as a result of variable habitat metrics (e.g., flow) being used in the

predictive model casts doubt on whether the RCA approach will be useful for assessing trends at the 3-5 year interval proposed in the adaptive management framework.

Similarly, no rigorous method exists to evaluate condition at a watershed or larger scale. The dichotomous test is performed on individual samples but no approach has been developed to combine samples or analytical results at a larger spatial scale. An additional challenge results from the potential for different samples (sites) within a watershed to be assigned to different reference groups, potentially resulting in inconsistent test results from the effects of a common stressor.

2. Performance of the current Yukon 2013 CABIN RCA Model

Model error rates and other applications

There are two types of errors that can result from the application of the BEAST model for classifications of individual samples/sites, and they are referred to as Type I and Type II errors (Table 1).

Table 1. Type I and II errors for the BEAST model classifications.

		True state of site	
		Reference	**Out-of-reference**
Model prediction	**Reference**	Correct	Type II error
	Out-of-reference	Type I error	Correct

Type I errors result from reference sites being classed as out-of-reference because their ordination scores fall outside of the space of most of the other reference sites. In a BEAST model the assessment of individual samples is based on where, in ordination space, the sample falls relative to percentiles of the ellipse surrounding the reference sample ordination scores. For example, the 90th percentile has been used as the decision rule to identify sites that are "moderately out of reference" (Strachan and Reynoldson 2014). However, this means that 10% of true reference sites could be classified as out-of-reference. In the typical BEAST application, the Type I error rate is set by the user, based on the decision rule that is deemed appropriate for the application.

Type I errors in the BEAST model can be evaluated by partitioning reference samples into fitting and validation datasets. The model is developed using the fitting samples, and the reliability of the model can be tested by running the validation sites through the model. The Type I error rate for the validation set is directly related to the percentile decision threshold used to delineate reference/out-of-reference sites. That is, if the 90th percentile is used to classify out of reference sites, the Type I error rate should be 10%. Two recent publications contain such evaluations for the Yukon data: in the first, Strachan and Reynoldson (2014) found Type I error rates of 0.53, whereas Reynoldson et al. (2016) obtained a value of 0.14, both higher than the expected value of 0.10. No compelling reasons were offered by the authors of the validation studies for the increased rate of Type I errors, but the difference in error rates between the two studies is noteworthy. When the BEAST model was applied to two other invertebrate datasets the Type I error rates also exceeded the expected values (Strachan and Reynoldson 2014). One possible explanation is model overfitting, which is caused by the model being fit to random error in the dataset. If the model is overfit, then deterioration in performance is expected when the model is applied to new data. Unfortunately there is insufficient information on the fitting procedures and no measures of model fit are available for review. It is encouraging that the Type I error rates from the most recent BEAST model are more similar to expected values.

Type II errors occur when samples from impacted locations are incorrectly classified as being in reference condition. Rates of Type II errors from a BEAST model are analyzed with a procedure analogous to that used by statistical power analysis. Power analysis estimates the probability that a sample with pre-defined effect size will be distinguished from a sample or samples from a control or reference group, taking into account the variability in the data and the Type I error rate. In the context of multivariate RCA this type of analysis is difficult to conduct because the effect size is not readily formulated from the multivariate statistics as would be the case if native metrics were used, such as total abundance or taxon richness. For example, if the assessment metric was the number of taxa, an analysis of Type II errors might ask: what is the probability of detecting a 50% decline in the number of taxa relative to reference conditions? To circumvent this problem, BEAST researchers have artificially manipulated samples so that they are different from reference samples through the reduction or elimination of taxa thought to be sensitive to a particular stressor. Variation in the severity of this manipulation is used to simulate different effect sizes. For example, Bailey et al. (2014) used an existing classification of individual taxa for tolerance to stress caused by eutrophication and reduced or eliminated sensitive taxa to simulate 3 levels of impairment. In the context of placer mining, Reynoldson et al. (2016) used correlations between taxa abundance and stream sediment metrics in the Yukon database to score taxa for sensitivity to placer impacts, and then created 3 simulated levels of impairment to evaluate Type II errors. Both schemes for simulating impacts have a scientific basis but no evidence is provided to support whether changes in taxon richness or abundance are within the realm of what is likely in the field.

Using simulated impairment data, Type II error rates range from 36-55% for the most severe impairment across the two studies. Error rates are 70-80% for the mild impairment data. These error rates are based on the use of the 90th percentile of the reference data as the cutoff for determining whether a site is in reference. It is difficult to compare results across studies as the reference dataset, BEAST model, and the simulated impairment algorithms were different. Bailey et al. (2014) compared RCA analytical approaches and found the BEAST model had higher Type II error rates than most other methods, based on results from the earlier Yukon dataset and model.

High rates of Type II errors are the result of significant overlap between ordination scores for the reference and simulated impaired datasets. This outcome could be the result of the simulated impairment resulting in relatively small “effect sizes”, however, without an explicit expectation for the effects of placer mining it is difficult to determine whether the simulated impairment is realistic. In the analysis of Reynoldson et al. (2016) the moderate effect size resulted in a decrease in abundance of 35% and no loss in richness, whereas the largest impairment reduced abundance by 44% and taxon richness was reduced by 8%. Reductions in abundance of this magnitude have been observed in streams exposed to placer mining (Seakem Group Ltd.1992).

High rates of Type II errors are also caused by the inherent variability in the data that results in a broad cloud of ordination scores for reference sites and for simulated impaired data. That variability is a result of both natural variation that is not controlled by the habitat variables, and the details of the sampling design particularly with respect to the lack of replication at the site, across sites, and through time.

Trade-offs between error rates and regulatory decisions

The significance of the error rates depends on the consequences to managers of those errors. In the case of Yukon placer management, there are no direct management implications for Type I errors resulting from reference sites that are outside the probability ellipse used to define sites that are in reference. In this case the decision criteria for Type I errors can potentially be

lowered, to improve power. Lower values (i.e., from 90 to 75th percentile) will cause more test sites to be assessed as out of reference and will cause the Type II error rates to decrease (which is observed in the studies using simulated test sites). However, it does mean there could be considerable overlap between characteristics of some of the test sites that are considered out of reference, and those reference sites that are in the outer bands of the ordination space. For example, choosing the 75th percentile as a decision rule means that 25% of reference sites have characteristics similar to some of test sites that will be considered out of reference. The only way to improve this situation is to reduce the overlap in distribution between reference and test sites by reducing sampling variation in the field, or by using a sampling design and analytical approach that takes into account the various sources of uncertainty in the data.

Alternative modeling approaches and options for improving model error rates.

The Type II error rate is a function of the effect size, decision criteria that define the Type I error rate, and the variability in the data. In this context effect size is the minimum change (the difference from control or reference conditions) that the sampling and analytical methods should be able to reliably detect. Effect size is usually defined by biological or management criteria and is usually independent of sampling or analytical considerations. Improvements to Type II errors rates can be achieved by reducing the variability in the data (e.g., increasing replication), or by improving the analytical approaches for the management of that variability. To reduce variability in the data, changes to the sampling design and field sampling protocols may be required. Replication in space and time will improve the precision of the analysis, particularly when the analytical methods take advantage of replication to partition the various sources of uncertainty. Alternative analytical approaches may be better suited to modelling the variability in the data which should lead to more precise estimates of possible effects of exposure to placer mining. This could potentially lead to a decrease in both Type I and Type II error rates.

A statistically robust method for analyzing differences in communities in a multivariate framework is available with tools such as PERMANOVA, which is a multivariate adaptation of Analysis of Variance (Anderson and Walsh 2013). These methods require a structured, replicated and preferably balanced sampling design, and could be used to test for differences in communities using samples organized into reference and test groups, while accounting for factors such as watershed, or environmental variables related to habitat conditions. Time (year) could also be entered in the analysis to evaluate changes over time. Similar to BEAST these methods rely on dichotomous hypothesis tests and do not yield outputs that are easily communicated to non-technical audiences.

An alternative to multivariate analyses are multimetric methods that rely on indices to estimate the strength of stressor effects on aquatic health. Methods are currently being developed for sediment-related impacts based on the relative abundance of sediment-sensitive taxa (Turley et al. 2016). Indices from these approaches can be analyzed in mixed-model or Bayesian frameworks that account for spatial and temporal variation as well as important habitat factors.

3. Sampling design and protocol

Spatial and temporal variation of existing benthic macroinvertebrate data

The Aquatic Health Monitoring Protocol sampling design consists of taking a single sample at a designated site, within a stream, and within a year (Figure 1). Between 0 and 64 reference sites have been sampled each year and a total of 372 samples from 286 locations have been sampled from 2004 to 2017. Sampling new and existing reference sites takes place to update the spatial and temporal representation of reference conditions. However, to include new reference samples in the analysis the multivariate model must be rerun to produce a new model. Currently there are 33 reference samples collected since 2012 that cannot be used until

a new multivariate model is built. Test sites are sampled in response to the flux in placer development, recent placer activities in authorized watersheds, and unacceptable monitoring results for test sites. While some sites have been sampled multiple times (i.e., multiple years) (reference sites range from one to three years; test sites one to seven years), most have been sampled once.

Variation in the BMI communities was assessed to identify the distribution of variance across sampling scales and where additional sampling effort might improve the precision of the estimated effects of placer mining on BMI communities. This analysis used all of the reference data for BMI communities however the lack of replication made it difficult to compare all sampling scales in a single analysis. For example, analyses that examined both spatial and temporal components of the surveys were not informative because annual replication (samples taken over multiple years) was confounded by the lack of within-site sample replication. That is, year-to-year differences in invertebrate communities within a site will be due to a combination of sampling variation (single sample) as well as natural variation across years. Consequently, temporal and spatial variation of reference sites was assessed using two approaches. For temporal variation, qualitative assessments of raw data for individual sites were summarized. Spatial variation was summarized using conditional mixed-effects models to partition variance at each spatial level (scale) (Appendix 1).

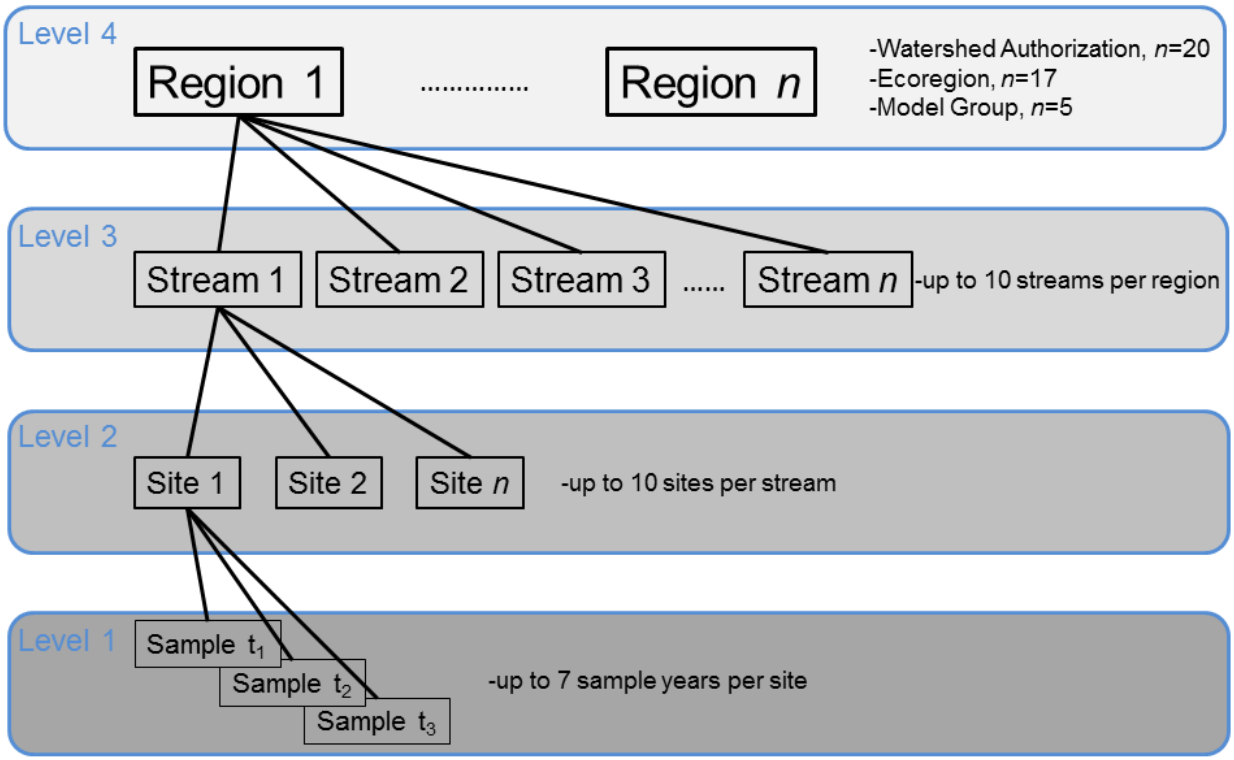

Figure 1. The Yukon Aquatic Health Monitoring Protocol sampling design. Level 1 is the scale at which samples are taken in the field. Only one sample per site is taken in year t. Samples may be taken across years but are often separated by multiple years. Level 2 shows that multiple sites may be sampled within stream n, and Level 3 shows that multiple streams may be sampled within region n. Streams can be organized into three regional grouping structures: 1) Watershed Authorization, which aligns with areas covered by Fisheries Act Authorizations for placer mining; 2) Ecoregion, which describes areas that share natural communities; and 3) Model Group, which organizes sites according to the output of the multivariate 2013 Yukon CABIN model, based on their BMI communities rather than geographic location.

Temporal variation

Temporal variation in BMI communities can be high. Shifts in abundance and the proportion of taxa among years within a site are evident in the BMI data from reference sites (Figure 2). Total abundances were dominated by few taxa; about 90% of the total abundance across all reference sites is only represented by 11% of the taxa. A shift from one dominant taxa to another (e.g., Figure 2C), or a large increase in one or more taxa that previously had low abundance occurred at a number of sites (e.g., Figure 2F and H). Other sites showed relatively similar taxonomic distribution among years (e.g., Figure 2E and G). At some sites, up to 60% of the families are observed in only one of the two years sampled resulting in a >65% increase in family richness from one sample year to the next (2007 and 2014; Figure 2D). Total abundance can also change dramatically among years. For instance, at site YPS-445, 97% fewer individuals were counted in 2010 than in 2016 (Figure 2A). Changes in abundance of this magnitude were not uncommon in reference sites. For context, the simulation analysis of Reynoldson et al. (2016) that was used to evaluate the performance of the multivariate method set the largest effect size to be a 44% decline in abundance and an 8% decrease in taxa richness. The year-to-year, sample-to-sample variation, or both, within a site could be much larger than these assumed effect sizes. The reference data suggest that natural and/or sampling variability could swamp the assumed effect size and may contribute significantly to the high Type II errors from the BEAST model. In contrast, MacDonald and Cote (2014) evaluate the impacts of urbanization on macroinvertebrate community composition over from 2006 to 2011 and found that the temporal variation was small relative to change that occurred at impacted sites. It is difficult to assess the temporal variation on a time scale relevant to placer management because many of the comparisons among years are based on intermittent sampling at time scales greater than the 3-5 year time period (e.g., Figure 2D) suggested for adaptive management.

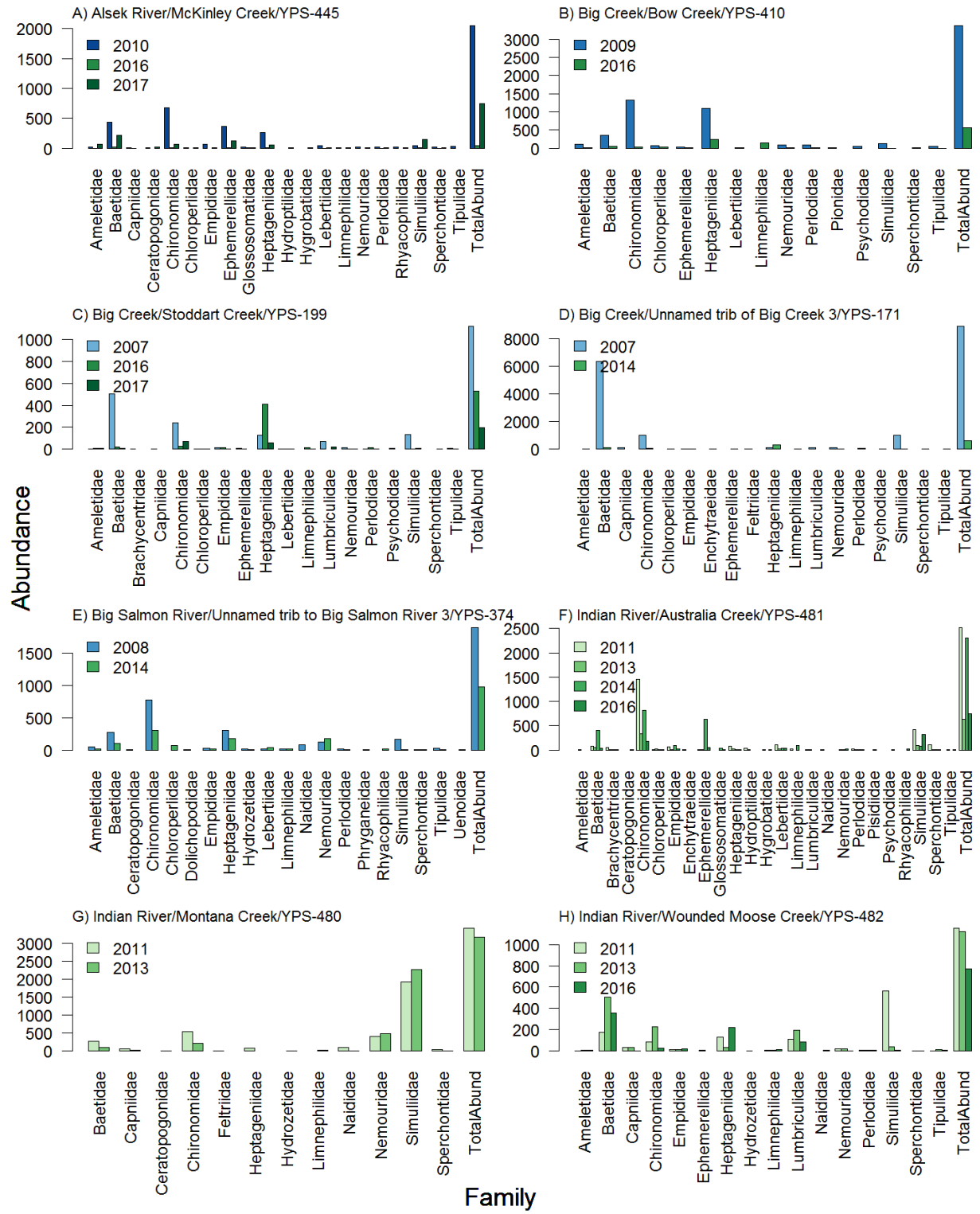

Figure 2. The abundance by family and total abundance for a selection of reference sites with multiple sample years (sites are the first 8 based on the alphabetical order of watershed names). Bars with shades of blue indicate years in the first half of the monitoring program (2004-2010) and bars with shades of green indicate years in the second half (2011-2017). Names in the top right corner are the watershed name stream name, and site code.

Spatial variation

Spatial variation was evaluated across stream and watershed scales with data from reference sites from the Yukon CABIN database. Commonly used BMI community response metrics were computed and the variance of each was partitioned into the different spatial sampling levels (and residual variance) using conditional null mixed-effects models (Nakagawa et al. 2017, and see Appendix 1 for methods). Partitioning variance to the appropriate spatial scale provides an indication of where spatial structure is important and where greater sampling effort would provide more precise estimates of the BMI communities (Harrison 2015).

Sample variation

Typically, replicate samples are taken within a site and year (site-year) for BMI surveys. Replication provides information about the repeatability of the sampling protocol and sampler and accounts for fine-scale spatial variation in abundance. Assessments that rely on a single sample will incorporate unnecessary error due to chance sampling events. This error will increase where there is fine-scale variability in habitat and BMI communities or biological events that can lead to dramatic changes in the communities, such as mass hatching events observed in BMI. Replicate samples within a site-year were not collected as part of the AHMP meaning the variation among samples at a site within a year is unknown. While it is acknowledged that replicate samples increase cost and time for collecting and processing samples, the lack of replication likely increases random noise in the data and lowers classification accuracy when determining reference groups and assigning test sites to those groups.

Earlier studies show that the repeatability of BMI samples collected using the CABIN protocol can vary among response metrics, sites, site type (e.g., reference vs. test) and ecoregion. Overall, repeatability was high (i.e., low coefficient of variation, CV) but for some sites BMI community metrics repeatability was low (e.g., CV up to 102%) (Strachan et al. 2009). Among the most variable metrics was abundance (e.g., CV: mean=25%; range=1-42%). Results also varied among ecoregions (Rosenberg et al. 1999) and whether samples were from reference sites or test sites (Sylvestre et al. 2005). Abundance-based metrics (e.g., total abundance) were more variable than taxa-presence metrics (e.g., taxa richness). This high variability led to greater classification errors produced by the BEAST model (which is abundance based) compared to RIVPACS or AUSRIVAS models (presence-absence based) when they were applied to data from the Fraser River Basin (Rosenberg et al. 1999). Transformations could be used to reduce the variability in metrics such as abundance. Replicate samples taken within and among years for a given site would be beneficial in identifying if improvements to the sampling protocol are needed. While this wouldn't allow for replication to be used in models it would increase consistency in sample composition and better represent the site's BMI community, without substantially increasing costs.

Site variation

Site variance is the variability in samples among sites nested within streams and streams nested within Watershed Authorizations or Ecoregions. On average, 6.7% and 8.3% of the variance was at the site scale when Watershed Authorization and Ecoregion were used as grouping variables, respectively (Watershed Authorization range: < 0.01% to 16.5%; Ecoregion range: <0.01 to 22.1%) (Table A2). This means that the site scale accounted for very little variance in BMI response metrics when compared to the variation associated with other levels of sampling (Figure 1). More specifically, the variance at the site scale for family richness and abundance was extremely low suggesting that there was large variation among site replicates (Table 2A and B) (i.e., variation among samples or years). Some variance was observed at the site scale for the EPT metric compared to the variance observed at the stream and region scales. There is likely considerable uncertainty about the component of variance due to site as

relatively few sites had more than one sampling event from which estimates of the variance could be made. Variance among sites could not be estimated when the data were structured by the 2013 BEAST model groupings as reference sites sampled after 2012 were not assigned to a group, thus reducing the sample size and preventing the model from converging on a solution (Table 2C).

Stream variation

Stream variance is the variability in samples among streams nested within Watershed Authorization, Ecoregion or BEAST model group. Variance at the stream scale differed among BMI metrics however it was greater than the variance at the site scale (Table 2B). On average, 13% of the total variance was explained at the stream scale (range: <0.01 to 40.5%) (Table A2), and there was little difference in the mean variance if the data were grouped by ecoregion, watershed authorization or BEAST model group (Table A2). Variance was highest for family richness and total abundance at the stream scale when Watershed Authorization or Ecoregion were used to group samples. This suggests that the much of the variation in family richness and total abundance is structured at the stream scale and sites within streams are more similar.

Region variation

Region variance is the variability in samples within Watershed Authorization, Ecoregion or CABIN model group. The amount of variance at the region scale was also highly variable among BMI metrics. On average, 12.6% and 13.5% of the variance was associated with Watershed Authorization and Ecoregion (Table A2). When sites were grouped by BEAST model groups, however, more of the variance was associated with the regional level (mean: 26.4%, range: 0 to 92.7). In particular, nearly all of the variance in total abundance was associated with the BEAST model groups (Table 2C). This highlights that abundance is an important driver of the model groups generated using the BEAST model. Diversity may play some role in determining model groups from reference sites and assignment of test sites to the model groups, but it is likely much less compared to the role of abundance.

Variation summary

While the distribution of variance is variable among BMI metrics and across scales there appears to be some general spatial patterns. When samples are grouped by Watershed Authorization or Ecoregion, considerable variation remains at the lowest (residual) level. In terms of sampling, this means that more replicates at-a-site (within and among years) should be collected as opposed to more sites within river or rivers within regions. The distribution of variability among levels is similar to that found by Li et al. (2001) for a suite of small Oregon streams. Modifications to the sampling design to incorporate a stratified sampling protocol may be necessary to capture the variation in BMI communities at larger spatial scales (e.g., among regions) and should produce more precise estimates of effects required to characterize watershed health at the regional scale. However, this sampling strategy will only be useful if complemented by an analysis that can incorporate spatial and temporal variation. Advice on sampling designs is found in Stevens (2002) and Foster et al. (2017) describe ways to incorporate legacy data into redesigned sampling programs.

Table 2. Variance components calculated from conditional null mixed-effects models for 3 community metrics and three regional groupings (See Appendix for methods). Samples have been organized into three different groupings that correspond to: 1) watershed authorizations associated with Fisheries Act Authorizations for placer mining, 2) ecoregions which are areas containing similar natural communities, and 3) model groups which are made up of sites with similar BMI communities and used in the 2013 Yukon BEAST model. The percent of variance in BMI metrics associated with the three spatial scales and the unstructured residual variance are presented in columns. Because there is no replication of samples within a site and year, the Site scale includes inter-annual variation as well as sample variation at a site. Blue shading (superscript 'b') indicates the sampling level with the lowest variance and grey shading (superscript 'g') indicates sampling levels with the highest variance across each row. Family richness is the total number of families in the sample, total abundance is the sum of all BMI, and %EPT abundance is the percentage of total abundance represented by EPT taxa (orders Ephemeroptera, Plecoptera, and Trichoptera).

Grouping	Response	Site	Stream	Group	Residual Variance
A) Watershed Authorizations	Family richness	<0.01 [b]	22.2	38.2 [g]	39.6
	Total abundance	<0.01 [b]	37.5 [g]	15.1	47.4
	% EPT abundance	6.7	11.8 [g]	1.9 [b]	79.6
B) Ecoregions	Family richness	<0.01 [b]	26.3	34.9	38.8
	Total abundance	<0.01 [b]	40.5 [g]	18.2	41.3
	% EPT abundance	5.4	12.7 [g]	2.6 [b]	79.4
C) Model Groups	Family richness	-	33.5 [b]	34.5 [g]	32
	Total abundance	-	0.7 [b]	92.7 [g]	6.6
	% EPT abundance	-	11.0 [b]	12.4 [g]	76.7

Replication and the BEAST approach

For the multivariate BEAST approach, sampling more reference sites over more years will increase variability in reference conditions. This will make it difficult to assign test sites and/or determine test site condition. To incorporate additional reference data, the BEAST model must be rerun, resulting in new reference groups, conditions, and a new predictive model for assigning test sites to reference groups. With each model revision the prediction accuracy tends to decrease, and is similar to a pattern observed in other regions with increasingly large networks (Reynoldson and Bailey 2013). This may be due to:

1. more complex networks and broader reference conditions make it difficult to discriminate affected sites, or
2. the initial model could have included spurious relationships and by incorporating more sites the spurious relationships weakened reducing the current model's performance.

For alternative analytical approaches that can properly incorporate the spatial and temporal variation, increasing sampling effort will lead to more precise estimates of the possible effects of placer mining on stream health.

Adaptive management

Adaptive management in the context of watershed-level authorizations is a complex task. Adaptive management involves sampling of test sites over time and reassessing the status of exposed sites within the area defined by each Fisheries Act authorization. Changes in status in the authorization area will inform potential management actions. The sampling design in its current form is insufficient for adaptive management for the following reasons:

1. too few test sites have been reassessed annually and high temporal variation and sample variation within a site-year could lead to the inaccurate assessment of changes in the health of a test site,
2. the BEAST model cannot assess trends in site or watershed status, and
3. the output of the multivariate procedure that is used to designate the health status of a site is not linked to the impacts associated with placer mining at the watershed scale (Reynoldson et al. 2016).

In the absence of a quantitative approach for assessing changes in stream or watershed health status, qualitative assessments of temporal trends can be used. However, determining the change in status from one year to another is complicated due to the difficulty in determining what degree of change warrants management action, particularly when the output used to assess change in the BMI community is in ordination space and is difficult to interpret.

Impacts associated with placer mining have not been linked to the aquatic health of a site as assessed by the multivariate BEAST model. The relationship between how far a site is out of reference and the degree of impact associated with placer mining such as the increase in suspended sediment rates is unknown. Therefore, an out-of-reference assessment provides no direct information about the impact of a placer mine. If the CABIN approach is continued, the relationship between the aquatic health assessments and placer activities should be established.

Sampling error and bias

Sampling error is unavoidable but should be quantified and where possible minimized. Large scale and long-term biomonitoring programs are prone to large sampling error due to the application of standardized protocols to a range of habitat conditions, multiple data collectors, and the employment of rapid techniques which in some cases can compromise the quality of the sample collected. However, sampling error can be mitigated by careful selection of indicators, consistent interpretation and application of sampling protocols and the collection of replicate samples. Specifically, the potential sources of sampling error for the AHMP include:

1. variation among samplers,
2. rapid assessment methodology of CABIN, and
3. variability in stream conditions.

Sampler variation

Large-scale and long-term monitoring programs often require many people to participate in sample collection. Turnover in sampling crews, difference in experience levels, and protocols interpreted and applied differently among samplers can increase sample variation. While this source of variation is unavoidable, protocols that standardize application and limit interpretation can mitigate the magnitude of sampling error. Kick-net samples, where the sampler disturbs the substrate upstream of a kick-net, collecting the disturbed invertebrates that float downstream, typically are less repeatable than more standardized sampling methods such as Hess and Surber samplers because kick-net samples do not sample a standardized area. Furthermore, repeatability of kick-net sampling could be related to stream size, whereby repeatability is lower in larger streams compared to smaller streams. However, the drawback of the Hess and Surber sampling methods is that they are limited to shallow moderately flowing habitats such as riffles and runs and inappropriate for slow moving pools or glides. The amount of error in the Yukon data that is associated with sampler variation is unknown and cannot currently be assessed.

Rapid assessment methods and the CABIN protocol

The CABIN sampling protocol is designed to provide practitioners with rapid and standardized methods for sampling BMI in the field. It uses kick-net sampling methods as described above whereby the sampler disturbs the substrate upstream of a kick-net, collects the disturbed invertebrates that float downstream, and continues to walk upstream throughout the habitat for a three minute period. Local habitat data are also collected at each site visit. The duration of sampling events, the number of replicates, and the habitat coverage have been evaluated for kick-net samples. More replicates of shorter kick-net samples are favoured over fewer longer samples (Feeley et al. 2012). Targeted riffle sampling was less precise than reach-wide sampling although the difference was small (Rehn et al. 2007). Gerth and Herlihy (2006) sampled high and low gradient streams and found that in high gradient streams the sampling of riffles yielded similar results to reach-wide sampling although reach-wide samples included more taxa. In low gradient systems where riffles are uncommon, the two approaches give different results and reach-wide sampling is preferred (Blocksom et al. 2008). Although the CABIN protocols have been modified to increase consistency in sample procedures, protocols designed to reducing time and costs may result in less precise samples. The lack of replication decreases sampling time in the field and lab but the significance of sampling variation cannot be assessed or accounted for.

Large range and variability of habitat conditions

A large number of habitat measures are taken at each sampling location. Among the reference sites sampled, there is a large range of habitat conditions as the protocol is applied to very small headwater streams and large mainstem rivers. For example, the range in average channel depth is 0.06 to 1.21 m (mean = 0.31 m) and wetted width is 0.8 to 117 m (mean = 8.3 m) among reference samples. Some habitat variables will not vary significantly over time but others, especially those influenced with annual variations in flow, will vary with each sample. Thus, habitat conditions for the same site can change such that the site is assigned to a different reference group each time it is sampled. This complicates the assessment of trends in stream health particularly if stream health is assumed to be relatively intransient (i.e., does not vary with annual or seasonal changes in environmental conditions).

4. Predictor variables

Predictor variables in the 2013 Yukon CABIN model

Predictor variables are used to assign test sites to reference groups in a two stage process. First, habitat characteristics are related to the reference groups using discriminate function analysis. All 84 habitat variables are competed in a model selection process that identifies the set of variables that best describes reference groups, although this process seems to be entangled with the process that determines the reference groups (see below). Once a final model is selected it is used to assign a test site to a reference group using habitat characteristics. For each test site, the model predicts the probability of the test site being within one of the five reference groups. The uncertainty of assigning the reference group to a test site adds to the uncertainty of the assessment of stress.

Utility of current predictor variables

The probability of assignment of test sites to a single reference group defined in the 2013 Yukon CABIN model is low, which suggests that assignment of a site to the optimal reference group is not always obvious. The low predictive performance of the model could be due to the inclusion of predictor variables with low predictive power and unsuitable model building practices that lead to low statistical power. Specifically, there is a strong reliance on variables with unclear

biological rationale while other key variables shown to have strong relationships with BMI communities are missing from the analysis. In addition, the 2013 Yukon CABIN model overlooks standard statistical practices used for predictive model development. These concerns are outlined below.

Biological rationale

Of the 84 predictor variables that are collected, 14 remained in the final 2013 Yukon CABIN model. Variable selection was a data-driven approach with little biological rationale provided for the final set of variables. Most variables considered in the analysis lack clear biological linkages to BMI communities. While climate and landscape structure undoubtedly play a role in shaping BMI communities, it is difficult understand the role each predictor variable plays without a hypothesis-driven approach describing the basic ecology of BMI communities in the Yukon. For example, it is unclear how the percent cover of Bryoids (mosses, lichens) in a watershed influences the BMI community, particularly when the average cover of terrestrial Bryoids constitutes on average less than 0.5% of the land cover. Similarly, climate variables such as metrics of monthly average precipitation are likely to drive hydraulic patterns but the mechanism by which BMI communities are affected is not specified. Decisions about the inclusion of variables in the initial model building process could benefit from a hypothesis-driven approach that is based on known relationships between predictor variables and BMI communities. The use of predictor variables with no clear biological relationship with BMI metrics will also lead to overfitting as spurious correlations will provide inaccurate predictions and will eventually break down in the absence of true relationships with BMI responses. No measures of the predictive power of the habitat variables have been provided.

Some variables included in the final model are a function of local environmental conditions, potentially including the stressor of interest, placer mining. This means that the assignment of a test site to one of the groups could be sensitive to local conditions at the time of sampling. Variables such as water quantity (represented by velocity, water depth and wetted channel width) may be linked to BMI communities and are included in the final model but could be directly affected by local events such as rainstorms or mining. Alternative variables such as stream channel slope would provide a much better representation of the BMI communities given the scale of measurement. It appears these data have been collected but are not available for all sites in the CABIN database and there appear to be many erroneous values in the current database (e.g., % channel slope > 10).

Selection of reference groups and predictor variables are two separate steps in the RCA approach and the CABIN protocol. However, the authors of the 2013 Yukon CABIN model suggest that the final groupings were determined in part by how well the habitat predictors can explain the difference among groups (Reynoldson and Bailey 2013). The evaluation of predictor variables and group selection based on biological assemblages should be decoupled. The development of biological grouping should have a strong biological basis. Biases introduced by spurious correlations with predictor variables will lead to inappropriate groupings and misclassifications of test sites. If the habitat predictors do not perform well then the most important habitat features may not be included in the model rather than the issue being poorly formed groups. This again highlights the benefits of taking a hypothesis-driven approach to variable selection when using these empirical modeling methods.

Standard statistical practices

There is extreme collinearity among predictor variables used to build and in the final 2013 Yukon CABIN model. Correlations (r-values) among the predictor variables included in the final model were up to 0.99. The inclusion of highly correlated variables will increase Type II errors by increasing the standard errors of parameter estimates. This is indicative of a model that has

been over fit (e.g., Zuur et al. 2010). Collinearity among variables should be examined prior to model development and redundant variables should be removed from the model building procedure.

Unstructured model selection techniques can lead to overfitting or under-fitting and result in poor predictive power. The 2013 Yukon CABIN model uses forward and backward stepwise discriminant function analyses. The procedure for selecting the best model (i.e., set of predictors) is not well documented and appears to have elements of subjectivity. There are many useful papers outlining commonly-used statistical methods available for reducing variable redundancy and selecting the best model (e.g. Zuur et al. 2010). Consideration should be given to alternative approaches to RCA analyses as highlighted in Bailey et al. (2014).

Conclusions

Benthic macroinvertebrates are a useful tool for evaluating impacts to stream condition but invertebrate populations are often highly variable in space and time. Thus the sampling design, analytical procedures, and management application should account for this variation to maximize the utility of the approach.

Results of this review suggest there are significant challenges in using the CABIN program for evaluating the effects of placer activity on the aquatic health of fish and fish habitat, particularly for the objectives of the Adaptive Management Framework. Under the CABIN protocol collection of both reference and test site samples has used an *ad-hoc* approach that has resulted in a database that is not consistently replicated in space and time. The BEAST model is based on the Bray-Curtis metric that is strongly influenced by raw count data; a highly variable measure that can span multiple orders of magnitude among samples. Without replicated sampling the extent to which sampling variation and the lack of replication contributes to the inconsistent performance of the models is difficult to determine. Further, no method exists within the BEAST approach to simultaneously analyse multiple samples as is required to conduct analyses at the stream or Watershed Authorization level, or to evaluate trends in stream condition as is envisioned in the Adaptive Management Protocol. No remedies to these shortcomings are apparent under current sampling and analytical protocols.

It is recommended that consideration be given to some modifications to the current sampling programs and additional analyses to support the long term goal of developing a more statistically defensible approach that meets the needs of the monitoring protocol and the adaptive management framework. To meet this goal a series of recommendations for future work are outlined below. The recommendations are intended as a sequential series of activities, as the first steps involve decisions or provide information that will inform latter ones.

Recommendations

Refinements to the Yukon CABIN approach

1. Some of the shortcomings with the current modeling approach that are outlined in this review could be addressed with a re-analysis and reformulation of the current model. This would include incorporation of more recent reference data, statistical transformation of abundance data (e.g., log-transform) to improve model performance, and a revised approach to the selection of habitat variables. However, none of these improvements address the larger issue of adapting the BEAST approach for use in a watershed-level analysis as outlined in the Adaptive Management Framework.

Improvements to the sampling protocol

2. As an initial step in the development of a design that can be used for the Adaptive Management Framework, a pilot study that uses a spatially and temporally balanced sampling design that includes replication to better estimate the components of variation in time and space should be implemented.
3. Based on the pilot study, the feasibility of moving the entire Aquatic Health Monitoring program to a more spatially and temporally balanced sampling design that will meet the objectives of the Adaptive Management Framework (using site-specific assessments to evaluate impacts at the watershed authorization scale as well as estimate trends in watershed health) should be evaluated. Such an approach should take advantage of the existing dataset, but would use a new sampling design such as the one outlined in Recommendation #2.

Consider other analytical approaches

4. Recent developments that use multimetric indices and benchmarks for impact assessments, as well as existing information on the effects of stressors such as sediment on invertebrate communities should be used to explore the use of multimetric measures to meet the objectives of the Adaptive Management Framework. This could be conducted on the existing Yukon dataset or the pilot study outlined in Recommendation #2.
5. Multivariate or other modeling approaches that can incorporate spatial and temporal variation in the macroinvertebrate dataset should be evaluated. Model development could begin at any time but evaluating model performance might be most useful after the spatially balanced design outlined in Recommendation #2 has been implemented.

Maintain the CABIN legacy

6. If possible, a revised program should continue to use the CABIN field protocol for collecting benthic invertebrate samples and use CABIN as the repository for sample data so that the reference dataset can be archived and made available for this and other uses in the Yukon Territory.

Contributors

Contributor	Affiliation
Mike Bradford	DFO Science, Pacific Region, Author
Doug Braun	DFO Science, Pacific Region, Author
Lisa Christensen	DFO Science, Pacific Region, Centre for Science Advice Pacific
Dave Cote	DFO Science, Newfoundland Region, Reviewer
Jennifer Harding	DFO FPP, Pacific Region, Client
Jeska Gagnon	DFO FPP, Pacific Region, Client
Nathan Ferguson	DFO FPP, Pacific Region, Client

Approved by

Carmel Lowe
Regional Director
Science Branch, Pacific Region
Fisheries and Oceans Canada

October 2, 2018

Sources of Information

Anderson, M.J. and Walsh, D.C.I. 2013. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? Ecological Monographs. 83:557-574.

Bailey, R.C., Linke, S., and Yates, A.G. 2014. Bioassessment of freshwater ecosystems using the Reference Condition Approach: Comparing established and new methods with common data sets. Freshwater Biology. 33: 1204-1211.

Blocksom, K.A., Autrey, B.C., Passmore, M., and Reynolds, L. 2008. A comparison of single and multiple habitat protocols for collecting macroinvertebrates in wadeable streams. Journal of the American Water Resources Association. 44: 577-593

Feeley, H.B., Woods, M., Baars, J., and Kelly-Quinn, M. 2012. Refining a kick sampling strategy for the bioassessment of benthic macroinvertebrates in headwater streams. Hydrobiologia. 683: 53-68

Foster, S.D., Hosack, G.R., Lawrence, E., Przeslawski, R., Hedge, P., Caley, M.J., Barrett, N.S., Williams, A., Li, J., Lynch, T., Dambacher, J.M., Sweatman, H.P.A., Hayes, K.R. 2017. Spatially-balanced designs that incorporate legacy sites. Methods in Ecology and Evolution 8: 1433–1442.

Gerth, W.J., and Herlihy, A.T. 2006. Effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. Journal of the North American Benthological Society. 25: 501-512.

Harrison X.A. 2015. Using observation-level random effects to model overdispersion in count data in ecology and evolution. Peerj 2, e616. (doi:10.7717/peerj.616).

Jones, J.I., Murphy, J.F., Collins, A.L., Sear, D.A., Naden, P.S., and Armitage, P.D. 2012. The impact of fine sediment on macroinvertebrates. River Research and Applications. 28:1055-1071.

Li, J., Herlihy, A., Gerth, W., Kaufman, P., Gregory, S., Urquhart, S., and Larsen, D.P. 2001. Variability in stream macroinvertebrates at multiple spatial scales. Freshwater Biology. 46:87-97.

MacDonald, A.J. and Cote, D. 2014. Temporal variability of benthic invertebrate communities at reference sites in eastern Newfoundland and its significance in long-term ecological monitoring, Journal of Freshwater Ecology, 29:2, 201-211

Mathers, K. L., Rice, S. P., and Wood, P. J. 2017. Temporal effects of enhanced fine sediment loading on macroinvertebrate community structure and functional traits. Science of the Total Environment. 599-600: 513-522.

Nakagawa, S., Johnson, P.C.D., and Schielzeth, H. 2017. The coefficient of determination R2 and intra-class correlation coefficients from generalized linear mixed-effects models revisited and expanded. Journal of the Royal Society Interface. 14: 20170213. (Accessed October 1, 2018).

Rehn, A.C., Ode, P.R. and C.P. Hawkins. 2007. Comparisons of targeted-riffle and reach-wide benthic invertebrate samples: implications for data sharing in stream condition assessments. Journal of the North American Benthological Society 26:15-31.

Reynoldson, T.B. and Bailey, J.L. 2013. Revision of the Yukon CABIN Invertebrate Bioassessment Model using 2004-12 Reference Site Data. Prepared for Yukon Placer Mining Secretariat by GHOST Environmental.

Reynoldson, T.B., Bailey, R.C., and Bailey, J.C. 2016. A review of the Yukon Placer Mining Aquatic Health Monitoring Protocol implementation. Prepared for Yukon Placer Mining Secretariat by GHOST Environmental.

Rosenberg, D.M., Reynoldson T.B., and Resh, V.H. 1999. Establishing reference conditions for benthic invertebrate monitoring in the Fraser River Catchment, British Columbia, Canada. Environment Canada.

Seakem Group Ltd. 1992. Yukon Placer Mining Study. Volume 1. Prepared for the Yukon Placer Mining Implementation Review Committee. Sidney, British Columbia.

Snijders, T., and Bosker, R. 2012. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage, London.

Stevens D.L. 2002. Sample design and statistical analysis methods for the integrated biological and physical monitoring of Oregon streams. Oregon Department of Fish and Wildlife, Report Number OPSW-ODFW-2002-07.

Strachan, S.A., and T.B. Reynoldson. 2014. Performance of the standard CABIN method: comparison of BEAST models and error rates to detect simulated degradation from multiple data sets. Freshwater Science. 33:1225-1237.

Strachan, S., Ryan, A., McDermott, H., and MacKinlay, C. 2009. Benthic invertebrate and water quality assessment of the Quinsam River Watershed in British Columbia 2001-2006. Environment Canada.

Sylvestre, S., Fluegel, M., and Tuominen, T. 2005 Benthic invertebrate assessment of streams in the Georgia Basin using the reference condition approach: Expansion of the Fraser River invertebrate monitoring program 1998-2002. Environment Canada.

Turley, M.D., Bilotta, G.S., Chadd, R.P., Extence, C.A., Brazier, R.E., Burnside, N.G., Pickwell, A.G.G. 2016. A sediment-specific family-level biomonitoring tool to identify the impacts of fine sediment in temperate rivers and streams. Ecological Indicators. 70:151-165.

Yukon Placer Aquatic Health Working Group (YPAHWG). 2008a. Adaptive management framework. (Accessed October 1, 2018)

Yukon Placer Aquatic Health Working Group (YPAHWG). 2008b. Aquatic health monitoring protocol. (Accessed October 1, 2018)

Zuur, A.F., Ieno, E.N., and Elphick, C.S. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution. 1: 2-14.

Appendix: Variance Components Methods

The variance components estimate the proportion of total variation that can explained at a given sampling level or scale. They can also be thought of as the strength of correlation between samples within a sampling level (Snijders and Bosker 2012). Variance components or Intraclass correlations are similar to Pearson's correlation coefficient (when used in a linear model) and is often used to determine appropriate sample sizes at different levels of sampling. Comparing the amount of variance in the response variable into specified sample scales highlights sampling trade-offs such as the number of samples within a site vs. the number of sites required to achieve a specified level of power. Conditional null mixed-effects models were constructed where the only fixed parameter was the global mean intercept and random effects included site, stream, and region. A linear form of the model was used when Gaussian error distribution was assumed and a generalized model with binomial error for proportion metrics. The conditional null model with Gaussian error was:

$$Y_{i,j,k,y} = \beta_0 + \alpha_{j,k,y} + \gamma_{j,y} + \delta_y + \varepsilon_{i,j,k,y}$$
$$\alpha \sim N(\sigma^2),$$
$$\gamma \sim N(\sigma^2),$$
$$\delta \sim N(\sigma^2),$$
$$\varepsilon \sim N(\sigma^2)$$

Where $Y_{i,j,k,y}$is the *i*'th measurement (i.e., sample) from site *j*, stream *k* within region *y*, $\alpha_{j,k,y}$ is the site within stream, region *j,k,y* random effect, $\gamma_{k,y}$ is stream within the region *k,y* random effect, δ_y is the region *y* random effect, β_0 is the global intercept, and $\varepsilon_{i,j,k,y}$ is the residual variance. All random effects and the residual variance are assumed to have a mean of zero and are normally distributed. The calculated intraclass correlation metrics are outlined in Table A1 and the results in Table 2 in the main text. Briefly, variance terms for random effects and residual error were used to determine the relative proportion of the total variance explained by one or more sample levels. For models with binomial error structure, an additional observation-level variance term was included (See Nakagawa et al. 2017 for details)

Table A1. Variance components or intraclass correlation metrics, equations, and descriptions for different sampling scales of the Yukon AHMP monitoring program. Variance parameters are Site/Stream/Region = σ_δ^2, Stream/Region = σ_α^2, Region = σ_γ^2, and Residual = σ_ε^2. For models with binomial error structure, an additional observation-level variance term was added to the denominator of each equation (Nakagawa et al. 2017). Three sets of metrics were calculated where region was the watershed authorization area, ecoregion, or the model groups from the 2013 Yukon CABIN model.

Variance Scale	Equation	Description
$Var_{Site/River/Region}$	$\frac{\sigma_\delta^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\varepsilon^2}$	The $Var_{Site/Stream/Region}$ describes the variance explained by grouping samples by site.
$Var_{Stream/Region}$	$\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\varepsilon^2}$	The $Var_{Stream/Region}$ describes the variance explained by grouping samples by stream.
Var_{Region}	$\frac{\sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\gamma^2 + \sigma_\delta^2 + \sigma_\varepsilon^2}$	The Var_{Region} describes the variance explained by grouping samples by region.

Samples were organized into three different regional structures including:

1. watershed authorization areas, which are geographical areas that correspond to the application of Placer Mine *Fisheries Act* Authorizations (*n*=20);
2. ecoregions, which are geographical areas that contain distinct natural communities – multiple ecoregions can span a Watershed Authorization Area (*n*=17);
3. model groups, which are the groups of reference sites with similar benthic invertebrate assemblages determined by the BEAST multivariate analyses used in the 2013 Yukon CABIN analysis (see the Background Section) and are not geographically based (*n*=5).

Table A2. Variance components calculated from conditional null mixed-effects models for 9 community metrics and three regional groupings (See Appendix for methods). Samples have been organized into three different regional groupings that correspond to: 1) watershed authorizations which are the areas pertaining to a Fisheries Act Authorization, 2) ecoregions which are areas containing similar natural communities, and 3) model groups which are made up of sites with similar BMI communities and used in the 2013 Yukon CABIN model. The amount of variance in BMI metrics associated with the three spatial scales and the unstructured residual variance are presented in columns. Because there is no replication of samples within a site and year, the Site scale includes inter-annual variation as well as sample variation at a site. Blue shading (superscript 'b') indicates the sampling level with the lowest variance and grey shading (superscript 'g') indicates sampling levels with the highest variance across each row.

Grouping	Response	Site	Stream	Group	Residual Variance
A) Watershed Authorizations	Simpson Index	16.5[g]	4.8[b]	13.0	65.7
	Shannon Index	10.1	9.2[b]	25.6[g]	55.1
	Family Richness	0.0[b]	22.2	38.2[g]	39.6
	Pielou's Evenness	13.9	15.6[g]	5.3[b]	65.2
	Total Abundance	0.0[b]	37.5[g]	15.1	47.4
	% EPT abundance	6.7	11.8[g]	1.9[b]	79.6
	% Sensitive abundance	3.8[g]	2.2[b]	2.5	91.6
	% Tolerant abundance	2.6	5.7[g]	0.0[b]	91.7
	Mean % variance	**6.7**	**13.6**	**12.7**	**67.0**
B) Ecoregions	Simpson Index	16.2	1.4[b]	17.7[g]	64.6
	Shannon Index	12.7	6.5[b]	24.0[g]	56.7
	Family Richness	0.0[b]	26.3	34.9[g]	38.8
	Pielou's Evenness	22.1[g]	5.8	7.2	65.0
	Total Abundance	0.0[b]	40.5[g]	18.2	41.3
	% EPT abundance	5.4	12.7[g]	2.6[b]	79.4
	% Sensitive abundance	4.2[g]	0.0[b]	3.7	92.1
	% Tolerant abundance	5.1[g]	0.0	0.0	94.9
	Mean % variance	**8.2**	**11.6**	**13.5**	**66.6**
C) Model Groups	Simpson Index		19.9[b]	21.0[g]	59.1
	Shannon Index		17.3[b]	24.0[g]	58.6
	Family Richness		33.5[b]	34.5[g]	32.0
	Pielou's Evenness		16.2[b]	21.5[g]	62.3
	Total Abundance		0.7[b]	92.7[g]	6.6
	% EPT abundance		11.0[b]	12.4[g]	76.7
	% Sensitive abundance		6.0[g]	5.2[b]	88.8
	% Tolerant abundance		5.1[g]	0.0[b]	94.9
	Mean % variance		**13.7**	**26.4**	**59.9**

This Report is Available from the

Centre for Science Advice
Pacific Region
Fisheries and Oceans Canada
3190 Hammond Bay Road
Nanaimo, BC V9T 6N7

Telephone: (250) 756-7208
E-Mail: csap@dfo-mpo.gc.ca
Internet address: www.dfo-mpo.gc.ca/csas-sccs/

ISSN 1919-3769

Correct Citation for this Publication:

DFO. 2019. Evaluation of the reference condition approach for Yukon placer mining monitoring. DFO Can. Sci. Advis. Sec. Sci. Resp. 2018/053.

Aussi disponible en français :

MPO. 2019. Évaluation de l'approche des conditions de référence pour la surveillance des activités d'exploitation des placers du Yukon. Secr. can. de consult. sci. du MPO, Rép. des Sci. 2018/053.