REVIEW OF TRANCHE 1 MATERIAL SUBMITTED BY THE VANCOUVER-FRASER PORT AUTHORITY IN RESPECT OF ROBERTS BANK TERMINAL 2

The Roberts Bank Terminal 2 Independent Scientific Body

October 2024

INDEPENDENT SCIENTIFIC BODY MEMBERS

Chair: Mona Nemer, Chief Science Advisor of Canada

Kelly Munkittrick, Department of Biological Sciences, University of Calgary, Canada.

David M. Paterson, Scottish Oceans Institute, School of Biology, University of St Andrews, United Kingdom.

Margaret Rubega, Department of Ecology and Evolutionary Biology, University of Connecticut, U.S.A.

Hannah S. Wauchope, School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom.

Jean-Michel Weber, Professor Emeritus, Department of Biology, University of Ottawa

OFFICE OF THE CHIEF SCIENCE ADVISOR STAFF

C. Scott Findlay, Researcher in Residence

SECTION I: REVIEW APPROACH AND SUMMARY OF RECOMMENDATIONS

The ISB's Roberts Bank Terminal 2 follow-up program evaluation criteria

The Vancouver Fraser Port Authority has proposed the construction and operation of a new three-berth marine container terminal located at Roberts Bank in Delta, British Columbia (the Roberts Bank Terminal 2 (RBT2) Project). Potential adverse effects of the project included changes in the in-shore salinity regime resulting in changes in the abundance and quality of biofilm, a major food source for Western sandpiper. Pursuant to RBT2 Decision Statement Condition 10.4.1, an Independent Scientific Body (ISB) was established by the Chief Science Advisor of Canada to review the proponent's follow-up monitoring plans for salinity, biofilm and Western Sandpiper. This report provides a comprehensive review by the ISB of the first tranche of material submitted by the VFPA in response to the Decision Statement.

Under the *Impact Assessment Act*, a follow-up program "means a program for verifying the accuracy of the impact assessment of a designated project and determining the effectiveness of any mitigation measures"[1]. The RBT2 environmental assessment concluded that "… salinity changes resulting from the project will not adversely affect biofilm and migratory birds, including shorebirds."[2] These are then the predictions that the RBT2 follow-up program is supposed to verify. Doing so requires that it:

(1) Specify the adverse effects that will be the focus of monitoring and provides a compelling rationale for the selection of indicators of these effects. This means that the follow up program must specify, for example, which attributes of biofilm communities are of concern (quantity? quality? spatiotemporal distribution?) and what will be measured in the field or lab (e.g. what data will be collected that will allow inferences about project effects on, say biofilm quantity versus biofilm quality). This specificity is crucial for reducing ambiguity and misinterpretation.
(2) Includes quality assurance/quality control protocols for reducing potential sources of bias that may affect the accuracy of either field, laboratory or modelling data. Without such protocols, data may be biased, leading to biased estimates of project effects.
(3) Is likely to detect adverse effects on salinity, biofilm or Western sandpiper. If the follow-up program is such that only large and immediate project effects are detectable, the conclusion that there are (as predicted) no adverse effects is very likely to be wrong. The result may well then be that implemented mitigation measures will be insufficient to mitigate adverse effects.
(4) Provides for reasonable inference that any detected adverse effects are due to the project and not some other factor(s). For example, detected changes in biofilm in the intertidal area following RBT2 construction might be due to ongoing changes in the discharge of freshwater from the Fraser River as a result of climate change. The

[1] *Impact Assessment Act* S.C. 2019, c. 28, s 2.

[2] *Roberts Bank Terminal 2 Project: information request response Executive Summary (Sept. 24, 2*, p. 9, available at https://iaac-aeic.gc.ca/050/documents/p80054/141462E.pdf (accessed Sept. 10, 2024)

conclusion that the observed changes are due to the project would then be wrong. And if it *is* wrong, then additional mitigation measures based on this (erroneous) conclusion may well have no effect.

The VFPA's proposed follow-up program is based on a large body of research conducted in and around the RBT2 site, much of which informed the 2015 RBT2 environmental impact statement itself. Since then, the VFPA has continued to conduct research on a wide range of issues. This information should be used to design a follow-up program that satisfies (as much as possible) the above criteria. Hence, an important additional criterion is:

(5) to what extent has existing and historical information been used to inform the proposed follow-up program?

When reviewing a specific element of the proposed follow-up program, the ISB asked:

(1) Which of the 5 criteria are implicated?
(2) Based on the implicated criteria, does what is being proposed correspond to a higher or lower quality program?
(3) How could the element in question be improved to increase the quality of the program?

Constraints on design of follow-up programs

Experimental designs that permit strong inference about the cause of observed effects invariably involve deliberate experimental manipulation where the units of observation (subjects, sample plots, sites, etc.) are randomized to different treatments, sample size is large, treatments are carefully designed, and there is wide latitude in the choice of appropriate controls.

Such designs are simply not possible in the context of follow-up programs. In follow-up, one cannot randomize sites to "project" and "non-project" (i.e. control) treatments: the "project" site is fixed, and there is usually only one. The number of candidate control sites is limited by a host of biophysical factors. And investigators usually cannot modify or manipulate candidate sites to render them more appropriate as controls.

These limitations have at least two important consequences. First, they imply that *no* follow-up program *can* yield very strong (on an *absolute* scale) inference about project effects (see criterion 4 above). Nor is any follow-up program likely to have high (on an *absolute* scale) power (see criterion 3 above).

Second, because for any follow-up program, strong (on an absolute scale) inference is not possible, if an adverse effect is *not* detected, there will always be uncertainty associated with the conclusion that the project has had no adverse effect. And if an effect *is* detected, the conclusion that it is due to the project will also have associated uncertainty. This is an epistemic

fact of life for *any* follow-up program, one that not only has important implications for all conclusions based on inferences from follow-up, but perhaps more importantly, for regulatory decisions flowing from these conclusions.

Overall consistency with best practices in follow-up

The VFPA has clearly invested considerable effort in both developing and describing the proposed follow-up programs. With a few exceptions, these descriptions are detailed, comprehensive and clear: this comprehensiveness and clarity is welcome, as it (largely) liberates the ISB from the onerous and frustrating task of trying to figure out precisely what is being proposed.

Moreover, the proponent understands the importance of follow up designs that, irrespective of the actual results, have the potential to yield reasonable inference about adverse project effects (see criterion 4 above). In this regard, the decision to (wherever possible) implement Before-After/Control-Impact (BACI[3]) designs is consistent with best practices in effects follow-up.

The ISB also notes that in several instances, the proponent has taken an explicitly adaptive approach to sampling, whereby sampling tactics (including selection of sampling locations, determination of the number of sampling sites, etc.) is adapted based on results obtained either earlier in the field season or in previous years. This is, again, consistent with best practices in follow-up monitoring. The ISB encourages the VFPA to use this approach for all follow-up components where it is both relevant and feasible.

General recommendations

In section II of this report, the ISB provides a detailed review of each of the proposed follow-up study components. For each study component, the report includes a set of recommendations, distinguishing those whose individual implementation would, in the ISB's view, result in a comparatively large improvement in the proposed program (highlighted in red), from those which individually (though not necessarily cumulatively), would result in smaller improvements (highlighted in green). These recommendations fall into seven broad categories: (1) study design; (2) selection of measurement endpoints[4]; (3) field sampling strategy and methods; (4)

[3]A BACI design involves at least one control (C) and one impact (I) site, with monitoring of selected endpoints occurring both before (B) and after (A) the project is implemented. Project effects are then inferred from the difference between control and impact sites *after* implementation compared to the difference *before* implementation.

[4]An *assessment* endpoint is, in the impact assessment context, a *valued ecosystem component* (VEC) – for example, the salinity regime, the quality or quantity of biofilm, etc. A *measurement* endpoint is a measurable attribute of the system from which one infers the effect on the associated assessment endpoint – for

laboratory methods; (5) use of historical field data; (6) model validation; and (7) statistical analysis and inference.

Here we summarize a set of general recommendations that (a) if individually implemented, would - in the ISB's view- result in comparatively large improvements in the proposed follow-up program; and (b) apply to at least two study components, leaving study-specific recommendations to section II of the report. As virtually all the ISB's recommendations about field sampling and laboratory methods are study-specific, the summary does not include recommendations in these categories.

Study design

For the biofilm and Western sandpiper (WESA) follow-up designs, there is one designated *impact* area (the area in the immediate vicinity of RBT2) and either one (for biofilm) or two (for WESA) proposed *control* areas. The appropriateness of a control area is determined by the similarity between it and the impact area with respect to the factors that influence the selected endpoints (e.g. biofilm quantity or quality, sandpiper prey availability and quality). While the ISB recognizes that choices for potential control areas are limited, and (mostly) agree that the proposed areas are likely the best of the possible choices, the ISB nonetheless has concerns about their appropriateness.

> **Recommendation 1.** Factors known to affect selected measurement endpoints (e.g. salinity range, biofilm quantity or quality, etc.) should be identified for each study component, and these factors compared between the impact site and the selected control site(s). Based on these comparisons, the appropriateness of selected control areas for each study component should be comprehensively evaluated, as should the implications of this evaluation to (a) the choice of experimental design[5] and (b) inferences about project effects (or lack thereof).

Use of historical field data

Since 2012, the VFPA has collected extensive data on salinity regimes, biofilm distribution, quantity and quality, and WESA spatiotemporal distribution in and around the RBT2 impact site. But from the descriptions provided in the Tranche 1 material, the salinity, biofilm and WESA diet study components do not appear to make maximal use of these extensive historical data. Based on the descriptions provided, it appears that historical data could be employed to assess the

example, Polyunsaturated fatty acid content of biofilm is a measurement endpoint associated with the assessment endpoint biofilm quality.

[5] For example, if the comparison indicates that there is no sufficiently appropriate control area, then inferences about project effects based on a BACI design are more likely to be wrong than under an (appropriately designed) Before-After design.

empirical association between where and when sandpipers feed, the quantity and quality of biofilm and invertebrates in feeding areas, and the associated salinity regime.

> **Recommendation 2.** Historical field data on salinity, biofilm and Western sandpiper should be used to fit empirical models of the association between these three sets of endpoints. The estimated strength of these empirical associations should then be used to inform follow-up design and sampling strategy to increase the likelihood of detecting project-related effects on biofilm and Western sandpiper.

Use of models

Both the salinity and biofilm study components propose using model-generated data to assess project effects. Because of the absence of an appropriate empirical control area for salinity, the salinity model will be used to generate modelled data for salinity regimes in the absence of the project. These model data will then be compared with field data on salinity regimes after project implementation to infer project effects. In the case of biofilm, calibration models based on multispectral drone imagery will be used to estimate the quality and quantity of biofilm over the entire impact and control areas both before and after project implementation. In both cases, the validity of inferences about project effects depends critically on the accuracy and precision of the models employed:

> **Recommendation 3:** The empirical accuracy and precision of all models used to estimate selected measurement endpoints should be determined, and minimal empirical accuracy and precision thresholds for their application should be explicitly described and justified.

Statistical analysis and inference

In follow-up programs, conclusions about project effects are based on the statistical analysis of monitoring data. All such analysis involves fitting mathematical models to the monitoring data, with project effects usually being inferred from the magnitude of specific model terms. Consequently, which models are fitted, how they are fitted, and how projects effects are inferred from fitted models is crucial to understanding the validity of conclusions about project effects. Although the ISB recognizes that the appropriateness of a particular modelling approach depends in part on the data being analyzed, establishing a presumptive approach to modelling *before* the data are collected is nonetheless critical.

> **Recommendation 4.** For each study component, a clear and comprehensive description of all presumptive statistical models that will be used to infer project effects should be provided. This description should include the underlying rationale for the choice of models and the method(s) of fitting, as well as an explanation of how project effects will be inferred from fitted models. If the models that end up being used after data collection

differ from presumptive models, the underlying rationale for the change should be clearly explained and justified.

All of the proposed study components make reference to sample size (e.g. the number of sampling locations or times). In follow-up, sample size is important in part because of its implications to statistical power – the ability to detect a project effect of a certain magnitude. Follow-up designs with low power may fail to detect significant project effects, especially those that are comparatively small on an annual basis but accumulate over time, perhaps leading to the erroneous inference that the project has had no effect. In the standard approach to inference, sample size determination requires a specification of desired power and a minimal detectable effect size.[6] However, of all the proposed study components, only the WESA diet study is explicit about a minimal detection threshold (of 20% project-induced change in the proportion of biofilm in WESA diet).

Recommendation 5. For all selected salinity, biofilm and WESA endpoints, the assumptions on which estimates of minimal sample size for follow-up are based (i.e. desired power and minimal detectable relevant effect size) should be explicitly stated and justified.

Selection of measurement endpoints

For all of the proposed follow-up programs, multiple measurement endpoints are proposed. Although there are several advantages to using multiple endpoints, they may introduce problems, especially in interpreting the results. For example, if multiple endpoints are treated independently in statistical analysis, and the analysis results in detection of adverse effects for some but not all, what does one conclude? Moreover, if two selected endpoints are highly correlated, the resources devoted to sampling one or the other are largely wasted, as there is no additional information to be gained by monitoring both.

Recommendation 6. All selected measurement endpoints should be clearly and compellingly linked to one or more assessment endpoints, and correlations among multiple measurement endpoints and their implications should be explored by employing multivariate modelling approaches. Where follow-up results lead to different inferences about project effects for different measurement endpoints associated with the same assessment endpoint, how these inconsistent results will be interpreted should be explicitly stated.

[6] In the follow-up context, a minimal detectable effect size is the minimal size of a project effect one wants to detect with a specified probability. The smaller the minimal detectable effect size, and the greater the required probability of detection, the larger the sample size required.

Finally, to reduce ambiguity and uncertainty about predicted project effects, it is important that each study component have clear objectives and that predicted results for each selected endpoint be explicitly stated.

> **Recommendation 7**. Each study component should include clear, concise statements (1-2 sentences) that explicitly describe (1) the question(s) being addressed by the proposed study design; (2) how these questions follow from one or more elements of Decision Statement 10.4; and (3) the predicted changes in each selected measurement endpoint due to project implementation.

SECTION II: DETAILED REVIEW

1. REVIEW BACKGROUND AND CONTEXT

Among its findings, the Review Panel constituted for the Vancouver Fraser Port Authority (VFPA)'s proposed Roberts Bank Terminal 2 (RBT2) in Delta, British Columbia was not able to conclude with certainty whether the project would result in an adverse effect on polyunsaturated fatty acid production in biofilm, a potentially critical nutritional component of biofilm that is consumed by western sandpipers and other shorebirds during their migration stopovers at Roberts Bank. Consequently, the Panel was unable to conclude with reasonable confidence that the Project would or would not have a residual adverse effect on western sandpiper.

Under Condition 10.4 of the Roberts Bank Terminal 2 Decision Statement[7], the VFPA is required to:

> 10.4.1. identify monitoring parameters, methods, and thresholds and submit those monitoring parameter, methods, and thresholds to the Agency for review and advice by an independent scientific body established by the Agency. The thresholds must include thresholds beyond which a potential adverse environmental effect on biofilm or Western sandpiper (*Calidris mauri*) is likely to occur as a result of salinity changes caused by the Designated Project.
>
> 10.4.2. establish, prior to construction, baseline conditions, taking into account variability, for the parameters identified pursuant to condition 10.4.1.
>
> 10.4.3. monitor, for a minimum of 36 months immediately following construction of the east basin containment dyke of the marine terminal, the parameters identified pursuant to condition 10.4.1 and compare these against the thresholds established pursuant to condition 10.4.1.
>
> 10.4.4. submit the baseline conditions information established pursuant to condition 10.4.2 and the results of monitoring conducted pursuant to condition 10.4.3 to the Agency for review by the independent scientific body; and
>
> 10.4.5. if the monitoring pursuant to in 10.4.3 indicates that changes to the monitoring parameters attributable to the Designated Project are exceeding thresholds identified in condition 10.4.1, as confirmed by the independent scientific body, the Proponent shall develop and implement modified or additional mitigation measures to return the monitoring parameters below thresholds or to offset the effects. These modified or additional mitigation measures may include but are not limited to biofilm habitat creation or enhancement and infrastructure redesign or removal.

[7] Available at: https://iaac-aeic.gc.ca/050/documents/p80054/147356E.pdf (accessed July 8, 2024)

Pursuant to RBT2 Decision Statement Condition 10.4.1, an Independent Scientific Body (ISB) was established by the Chief Science Advisor of Canada on request from Impact Assessment Agency of Canada. The Terms of Reference[8] for the ISB were developed by the Office of the Chief Science Advisor of Canada with input from the Agency.

In this report, the ISB reviews the first tranche of material submitted by the proponent pursuant to Condition 10.4. This material comprises:

- Western Sandpiper (WESA) Follow Up Program - monitoring program design
- Appendix A: salinity study design
- Appendix B: biofilm availability study design
- Appendix C: WESA diet study design

Subsequent (Tranche 2) material that will be reviewed by the ISB includes:

- Appendix D: WESA Energy Intake Study Component Design
- Appendix E: WESA Foraging Distribution and Intensity Study Component Design
- Appendix F: WESA Abundance Study Component Design

The ISB will also review the WESA Adaptive Management Approach (AMA) that will be developed following the finalized design of the monitoring program and implementation of pre–construction data collection. The AMA will focus on characterization of the thresholds beyond which a potential adverse effect on biofilm or WESA is likely to occur as a result of changes in salinity caused by the project; a description of the adaptive management process that will be implemented if thresholds are exceeded; and potential adaptive management measures that may be implemented to mitigate any threshold exceedances.

As none of the Tranche 1 material includes information on adaptive management thresholds, the current review focuses on (a) the determination of baseline conditions, pursuant to Condition 10.4.2; and (b) the proponent's proposed set of monitoring parameters and methods for salinity, biofilm and WESA, pursuant to Condition 10.4.1.

2. FINDINGS

2.1 Interpretation of Decision Statement Condition 10.4

On August 24, 2020, the Minister of Environment and Climate Change Canada (ECCC) requested that the VFPA provide additional information regarding potential mitigating measures that would avoid or reduce several of the environmental effects described in the Environmental Impact Statement. As part of this request, the minister requested additional modelling of

[8] Available at: https://science.gc.ca/site/science/en/office-chief-science-advisor/independent-scientific-body-environmental-impact-roberts-bank-terminal-2-project/terms-reference-isb (Accessed July 22, 2024)

coastal geomorphology and salinity changes that may affect biofilm and migratory birds. The Proponent's response to information requested for biofilm and migratory birds was outlined in Information Request (IR) 2020-4: Biofilm and Effects to Migratory Birds (IR2020-4), and in Appendix IR2020-4 (IR2020-4-A). In its response, the VFPA concluded that:

> "The environmental impact statement conclusion that salinity changes resulting from the project will not adversely affect biofilm and migratory birds, including shorebirds, remains unchanged and continues to be supported by evidence showing that biofilm at Roberts Bank thrives and is abundant under variable salinity conditions."[9]

In its review of the VFPA's response to the information request, ECCC concluded:

> "Upon review of IR2020-4 and IR2020-4-A, ECCC's opinion… remains that the effects of the project, as designed, will likely be unmitigable and irreversible, resulting in an increased risk to the population viability of the Western Sandpiper species, in particular."

It seems clear then, that for ECCC the principal concern (i.e. the primary *assessment endpoint*) is WESA *population viability*.

VPFA's response to IR2020-4 states:

> "…the environmental impact statement conclusion that salinity changes resulting from the project will not adversely affect biofilm and migratory birds, including shorebirds, remains unchanged…"

But precisely *what* WESA ***assessment*** endpoint(s) are implied in this statement are unclear. Several obvious interpretations are:

(1) *whatever* adverse *proximate* effects the project may have on WESA (e.g. changes in diet, foraging behaviour, etc.), these effects would *not* pose a significant additional (i.e. above background) risk to the *population viability* of WESA.
(2) there will be no significant adverse proximate effects of the project on WESA, i.e. no significant changes in diet, foraging behaviour, etc.

These are different interpretations with important implications for follow-up and adaptive management. Under interpretation (1), the prediction concerns an adverse project effect on WESA population viability. Insofar as a follow-up program is supposed to verify predictions, it should therefore be designed to detect changes in measurement endpoints that are strongly predictive of WESA population viability.

[9] *Roberts Bank Terminal 2 Project: Information request response Executive Summary* (Sept. 24, 2021) p. 9.

Population viability may be estimated from fluctuations in population size over time or from estimated vital rates, neither of which the proposed follow-up program addresses. Nor is it clear that follow-up will continue for sufficiently long to allow for reliable estimates of project effects on population viability based on changes in either of these parameters. Hence under the interpretation that predictions concern WESA population viability, it appears that the proposed follow-up program cannot verify them.

By contrast, under the second interpretation, population viability may still be the ultimate concern, but the predictions strictly concern *the adverse proximate effects themselves*. The appropriate focus of follow-up are then measurement endpoints that are predictive of these proximate effects. Under this interpretation, the ISB's view is that many of the endpoints selected for the proposed follow-up program satisfy this criterion, in which case, the proposed program is – at least in principle – able to verify predictions.

For the purpose of the current review, the ISB has adopted the latter interpretation. However, the ISB notes that because the proposed follow-up program does not include measurement endpoints directly related to WESA population viability, any inferences to population viability will be subject to considerable – possibly large - extrapolation uncertainty.[10]

2.2. SALINITY COMPONENT STUDY

The salinity study is well described, with much of the detail required to evaluate the proposed follow-up program clearly presented. It is also clear that the proponent has invested considerable resources in salinity monitoring to date, collecting data that are critical not only to validating the Navier-Stokes (NS) salinity model, but in inferring project effects.

2.2.1. Study design

(1) Salinity field monitoring data will be available only for 2 of the 4 BACI cells (before-without project; after – with project); the other two cells (before-with project and after-without project) are based exclusively on *modelled* data. There is, therefore, no proposed empirical control.

This has several implications. First, because inferences about project effects based on the field data involve exclusively the "before-without" and "after-with" cells of the design matrix, for these data the study design is Before-After, a design with considerably lower inferential strength than a BACI design.

[10] This uncertainty might be reduced if, in the context of an adaptive management plan, exceedance of thresholds established *a priori* results in the design and implementation of one or more studies specifically designed to examine the relationship between project-induced effects on identified measurement endpoints and population fluctuations and/or vital rates. The ISB notes in passing that any such studies would require designs that are quite different than the ones currently proposed.

Second, for the NS model data there is no difference between "before without project" and "after without project", or between "before with project" and "after with project", if one assumes stationarity in the modelled processes affecting salinity over the time frame of interest. That is, the study design for the NS model data is fundamentally a Control-Impact (CI) design, again with lower inferential strength than a BACI design.

Third, as shown in Fig. 3.1. of the salinity component study, one can use the (a) "before-without" field monitoring data to evaluate the empirical accuracy and precision of the NS "without project" model; and (b) the "after-with" field monitoring data to evaluate the accuracy and precision of the NS "with project" model. But because there are no field data for "before-with" and "after-without", one cannot generate an independent *empirical* estimate of the accuracy and precision of the NS model data for these cells of the design matrix. The best one can do is assume stationarity in model accuracy and precision (so that, for example, empirical estimates of NS accuracy and precision based on "before without" and "after with" also apply to "after without" and "before with" respectively). So, the "with/without" ("CI") comparison using NS-modelled data is compromised by the relatively weak study design (CI versus BACI), and further compromised by an estimated accuracy and precision based purely on presumed - not validated - stationarity.

If the CI comparison using NS-modelled data yields only weak inference, why do it at all in the context of follow-up? The reason, the ISB suspects, is that the last 2 endpoints listed in Table 3.1 (which, unlike the first four, represent *extensive* rather than *intensive*[11] variables) cannot be empirically estimated from the salinity field data from the 11 stations. For these variables, one needs to estimate the salinity surface at a given time at < ha spatial resolution for the *entire impact area* (roughly 15 km^2), and 11 stations is woefully insufficient to do so. Consequently, the NS model is the only way to derive these estimates.

However, one cannot assume that the accuracy and precision derived from the Before-After empirical tests of the NS model apply to these extensive endpoints because the accuracy and precision evaluations concern point estimates of intensive variables at a small number of individual locations, not cumulative area estimates of extensive variables over a roughly 15 km^2 area. So even if the NS model has high empirical

[11] *Extensive* and *intensive* variables describe properties of a system that depend or do not depend respectively on the size of the system. For example, the *density* of an object is an intensive property, since density does not change with the size of the object (e.g. the density of a 1 kg ingot of gold is the same as the density of a 20 kg ingot). By contrast, the *mass* of an object is an extensive property: the mass of 1 L gold ingot is very different than the mass of a 10 L gold ingot. In the context of the proposed salinity study (see Table 3.1), daily salinity oscillations at a field station is an *intensive* variable, whereas the number of hectares in a key biofilm zone experiencing a specified daily salinity change is an *extensive* variable (Table 3.1). Similarly, in the proposed biofilm study (see Table 3.1), MPB biomass (in mg/m^2) is an intensive variable, whereas MPB biomass in a defined area (in tonnes) is an extensive variable.

accuracy and precision with respect to field data from the 11 stations, to infer that this is also the case for the last 2 endpoints listed in Table 3.1 over the entire 15 km^2 impact area would be, in the ISB's view, completely unwarranted.

Finally, the ISB notes that unlike the first 4 endpoints in Table 3.1, for which one has a defined number of stations, and hence, a defined sample size for spatial replicates, one cannot use NS – model estimates of the last two endpoints in Table 3.1 to infer project effects from fitted statistical models because sample size is arbitrary: for a modelled surface of area X (say, 15 km^2) with spatial mesh size ΔX (say, 1 ha), one can generate a sample with size anywhere between 1 and $X/\Delta X$ (say, 1500) And because sample size determines power, by choosing *N* to be small, one effectively eliminates any chance of detecting a project effect - however large - while by choosing *N* to be very large, even ecologically insignificant project effects may be detected. The ISB suspects that this is why the described analytical approach to the NS modelled data (see salinity component study, s. 4.3, p. 16) makes no reference to statistical models or inferences therefrom.

Recommendation 2.2.1.1. With respect to the last two (extensive) endpoints listed in Table 3.1, either:

(i) They are eliminated and the salinity follow-up focuses on the first four; or
(ii) They are retained with the results presented graphically similar to that proposed in s. 4.2.3 of the salinity component study. But in any such presentation, it should be explicitly stated that these "inferred changes" are, in fact, modelled changes whose empirical accuracy and precision cannot be directly estimated.

(2) The Brunswick impact area is predicted to show spatially variable project effects on salinity profiles, with the effect gradient running approximately from the northwest portion (where small positive changes in salinity and small expansions of daily salinity range are expected) to the southeast (where comparatively larger negative changes in salinity and reductions in daily range are predicted – see e.g. Figs. 3.4 and 3.7 of the salinity component study). For the purposes of verifying predictions then, salinity field sampling stations should be located in such a way as to provide a rigorous test of the predictive value of the salinity model during and after project build-out by ensuring that the sample sites cover the full (predicted) spatial gradient.

Recommendation 2.2.1.2. In the Brunswick impact area, salinity stations should be allocated to 4 different strata, with at least 3 stations (acting as triplicates) per strata: predicted salinity changes are (1) positive; (2) small negative; (3) moderate negative; (4) large negative. It is unclear whether the current set of 11 salinity stations satisfies this criterion: if not, the set of existing stations should be supplemented to ensure this criterion is satisfied.

2.2.2. Navier-Stokes (NS) salinity model validation

(1) It is unclear how precisely the NS-model will be validated (salinity study component s. 3.5.2.2, p. 15). From the provided description, it appears that for each day during the sandpiper migration period in a given year, there will be an empirical estimate of the salinity profile based on field monitoring of the first 4 endpoints listed in Table 3.1 as well as a modelled estimate at each station. Is this correct? If not, then what? Will the accuracy and precision analysis be stratified by station? By year? And how precisely will accuracy and precision be estimated? Linear regression of empirical measurements on modelled estimates, with accuracy determined by the slope and precision by the residual mean square error (or some such)? Finally, is there some threshold accuracy and precision below which the model would be considered unreliable?

Recommendation 2.2.2.1. An explicit and detailed explanation of how precisely the accuracy and precision of estimates of the first 4 endpoints listed in Table 3.1 will be determined should be provided.

(2) How will the uncertainty associated with estimates of the last 2 endpoints listed in Table 3.1 be characterized if they continue to be used (see Recommendation 2.2.1.1 above)? At present, the Appendix of the salinity component study provides very useful and detailed information on the NS-model and its predictions, but no information on how estimates of associated accuracy and precision will – or even *would* - be derived, given that (reliable) *empirical* estimates of model accuracy and precision for these endpoints would seem to be unobtainable under the proposed sampling design (see s. 2.2.1 (1) above).

Recommendation 2.2.2.2. If the last two modelled endpoints in Table 3.1 continue to be used, a detailed description of how the *empirical* accuracy and precision of the NS-model with respect to these endpoints will be determined should be provided.

2.2.3. Salinity measurement endpoints

The raw salinity field data are measurements of salinity at a given location (one of the 11 sampling stations) at a given time, i.e. a location-specific *time series*.

This has several implications. First, in principle, there are an infinite number of endpoints ("metrics") one can derive from such a series, of which four are proposed in Table 3.1. These lead to the obvious question: why these versus others?

In the ISB's view, the rationales given in Table 3.1 are not sufficiently compelling. For example, "useful for comparison to field data from other years" as a rationale for daily 5^{th} and 95^{th} percentile is weak: after all, for other years one would presumably also have the 10^{th} and 90^{th} percentile, or the 15^{th} and 85^{th} percentile, or …. The same weak rationale is used for short-

term salinity range, which has as an additional rationale that it constitutes a "validation metric for model performance" – an even weaker rationale since the NS model *raw* output is a salinity prediction at a given time and place, *not* a daily salinity *range* (which is a *derived* estimate based on model estimates throughout the day).

In the ISB's view, the choice of which attributes of the salinity measurement time series should be used to infer project effects should be based on two criteria: (1) the chosen attribute is clearly and directedly related to some explicit and relevant[12] hypothesis which can be tested using the salinity data (as would, for example, appear to be the case for daily salinity range); and/or (2) there is independent evidence (e.g. from other field or laboratory studies) of an empirical association between the salinity attribute in question and one or more selected biofilm quality or quantity measurement endpoints.

(1) *Short – term salinity range* (STSR): from the description, it is unclear exactly what this is. Is it the difference between the (NS model estimated) maximum and minimum salinity values over a 24-hr period?
(2) *Daily salinity oscillations* (DSO). Again, unclear how this is characterized, and how it differs from STSR. Is it the number of times during a 24-hr period where the salinity changes by more than 5 PSUs? If not, what is it? Why is the resolution set at 5 PSUs especially when for other endpoints, the resolution is 2 PSUs? Both the empirical and modelled data are at much finer resolution, so why impose an (arbitrary?) 5 PSU increment?
(3) *5 and 95% percentiles (5/95P) and 50% percentile during northward migration.* Percentile of what, precisely? The daily salinity distribution? And why these percentiles versus others? If these are percentiles of the daily salinity distribution, they will likely be correlated with each other and with STSR and DSO (depending on how they are defined), in which case these endpoints should not be considered independent (see also Section I, Recommendation 6)

> **Recommendation 2.2.3.1. All selected salinity endpoints should be unambiguously defined, with a clear and convincing rationale for each.**

2.2.4. Statistical analysis of salinity field monitoring data

(1) It is unclear from the description provided (s. 4.2 of the salinity study component) how whether salinity profiles "fall within the natural range of variability" will be determined. It appears that a two-variable regression model will be fitted to some sort of temporally averaged field data at each station, using data from before construction to fit the model. Without knowing how precisely the fitted regression model will be used to characterize "the range of natural variability", the ISB is unable to assess the validity of the proposed

[12] In this context, a "relevant" hypothesis is one whose truth (or otherwise) has demonstrably direct implications to the follow-up program objective of verifying predictions of the impact statement.

approach. But since the test for project effects is whether "after" data lie outside the normal range, how the normal range is characterized, what precisely is meant by "lie outside", and how whether the data lie outside this range will be determined, is critical information that is lacking in the current description.

Recommendation 2.2.4.1. A detailed description of how precisely the range of natural variation for salinity profiles will be determined based on the proposed regression approach and how whether an "after" profile "lies outside" this range will be determined should be provided.[13]

(2) From Table 3.1, the different field monitoring endpoints appear to be simply different attributes of the daily salinity distribution at a site. As such, one might expect to see some correlation among these attributes. The description is unclear on this issue, but it appears that the proposed analysis will treat each endpoint in Table 3.1 independently, thereby assuming that these different attributes are uncorrelated. In the ISB's view, a multivariate approach is indicated here, with the response (dependent) vector having elements corresponding to the different (intensive) salinity endpoints.

Recommendation 2.2.4.2. See Section I, Recommendation 6.

(3) The salinity data at individual stations will be temporally autocorrelated at several time-scales. Moreover, salinity at different stations may also be spatially autocorrelated – certainly this is suggested by the NS modelling results. From the description in s. 4.2 of the salinity component study, it does not appear that the suggested regression approach will consider temporal autocorrelation (see also (4) below). Spatial autocorrelation might be addressed by fitting separate models for each station, but one is then left with the possibility that the predictive value of the proposed regression model will differ among stations. What will one conclude if, for some subset of stations, "after" data lie outside

[13] Here the ISB might offer the following suggestion. Suppose that using some approach, one decides that a model M with associated error variance $E(M|B)$ provides the best fit to the "before without" salinity field data. One interpretation of the description given in s. 4.2 of the salinity study is that the *identical* model M will be used for (*not* fitted to) the "after" monitoring data, generating an "after" error variance $E(M|A)$. Then, the two error variances $E(M|B$ and $E(M|A)$ will be compared. If $E(M|A) > E(M|B)$, the inference is that there has been a project effect, and the greater the difference $(E(M|A) - E(M|B)$, the larger the inferred project effect. (This seems to us to be a – though by no means the only – reasonable approach, but from the description provided, it is unclear whether this is the intention.)

Adopting an hypothesis-testing approach to inference, one might then define a (one-tailed) null: H_0: $E(M|A) \leq E(M|B)$. The ISB notes, however, that this null is non-informative: indeed, it implicitly assumes the N-S model used to generate predictions about project effects is wrong because the null corresponds to *no* changes in salinity whereas the NS model predicts *some* (specific) changes. The point here is that because the null is non-informative, so is its rejection or acceptance. For us, this underscores the idea (see Recommendation 2.2.4 above) that here, as elsewhere, what matters for inferring project effects is the estimated "project – induced" effect size - in this case, $E(M|A) - E(M|B)$: the larger it is, the larger the inferred project effect. It follows that the focus of statistical modelling should be on generating effect size estimates, not on null hypothesis testing.

the natural range of variation, whereas for other stations, they do not? A better approach may well be to fit mixed models where station is modelled as a random effect, with a spatial autocorrelation term included.

(4) Based on the description in s. 4.2 of the salinity component study, it appears that there will be some sort of temporal "coarse graining" of the salinity, discharge and tide data, with medians or averages as well as ranges, derived for each time "segment". But it is unclear *why* such (temporal) coarse graining is required. The raw data are time series, so why not fit autoregressive models rather than simple regression models? This would eliminate the need for temporal coarse graining or, at the very least, reduce the temporal grain size (e.g. salinity data are collected every 5 minutes, so averaging over a 1 hr temporal grain size seems reasonable for autoregressive modelling). What is the rationale for not employing a seemingly more appropriate autoregressive approach to modelling rather than the proposed simple regression approach, especially since for such data, the proposed standard regression approach suffers from several deficiencies, including:

(a) Coarse graining at sufficiently large grain size may reduce the issue of temporal autocorrelation (see (3) above). But there is an associated cost: it also eliminates within-grain variation as a set of salinity values are replaced by a single value (e.g. a median), thereby reducing the error variance of any fitted model (since "unexplained" within-grain variation is automatically eliminated). If this within-grain variation is biologically important, coarse-graining will eliminate the possibility of detecting a project effect on within-grain time-scales.

(b) Different grain sizes will likely result in different regression model fits, which means differing amounts of unexplained error. If (i) this unexplained error is used for characterizing the "range of natural variation" (see (1) above); and (ii) exceedance of this natural range is being used to infer project effects; it follows that (iii) the choice of grain size may well affect inferences about project effects. This potential dependency of inferences about project effects on the choice of temporal grain size is, in the ISB's opinion, problematic. At the very least it demands a clear and compelling rationale for and justification of the choice of temporal grain size. A better approach would be to treat temporal grain size as an adjustable parameter and examine the sensitivity of inferences about project effects to changes in grain size, or fit autoregressive models where the issue of patterns on different time scales is addressed explicitly.

Recommendation 2.2.4.3. Statistical modelling to support conclusions about project effects should take explicit account of potential spatial and temporal autocorrelation.[14] Inferences about project effects should consider the potential effect of spatiotemporal grain size when models are fit to coarse-grained data.

[14] Various modelling tools exist for doing so. For example, the nlme package in R allows one to specify mixed models that include spatial and temporal autocorrelation effects.

2.2.5. Other issues

(1) For all years with field salinity data from the 11 stations, Navier -Stokes (NS) model precision and accuracy could be evaluated. Has this been done? If not, it should be, and the results of this analysis presented for each year for which salinity data are available. Of particular importance will be whether the combined inaccuracy and imprecision exceeds the expected project effects on salinity, that is, whether model noise is likely to obscure any project signal.

Recommendation 2.2.5.1. NS model precision and accuracy using historical salinity data should be evaluated and presented. This analysis should be used to assess overall accuracy and precision of the NS model, annual variation in accuracy and precision, and site-specific temporal variation in accuracy and precision.

(2) (i) Historical data provide detailed information on spatiotemporal variation in salinity in the Brunswick impact area based on the 11 salinity stations. Historical data (from 2016-2018) are also available for biofilm, in the immediate vicinity of 7 of these stations. These data would seem to allow for an investigation of the *empirical* association between various salinity endpoints (e.g. daily range) and biofilm endpoints for these years based on salinity and biofilm data from these stations and, more specifically, an estimate of the (empirical) predictive value of the former with respect to the latter.

(ii) From Fig. 3.3 of the salinity component study, it appears that virtually all of the existing salinity stations also have estimates of WESA use based on droppings. These data would seem to allow for an investigation of the empirical association between attributes of the salinity regime and WESA use. Since there are also 7 stations for which there are associated biofilm data, the empirical association between biofilm and WESA use could also be investigated.

(iii) As there are also modelled salinity data for each of the 11 stations, the analysis described in (i) and (ii) could also be done using modelled versus field data.[15]

The results of (i) – (iii) would be fitted empirical models that can be used to predict effects on biofilm and WESA resulting from predicted (i.e. NS-modelled) changes in the salinity regime. Equally importantly, the fitted models would provide quantitative estimates of predictive value and residual uncertainty (i.e. error variance).

[15] Many studies of salinity, biofilm and WESA at Roberts Bank have been conducted over the last decade, very few of which the ISB has reviewed. It is, therefore, entirely possible that the investigations described and recommended here (and perhaps elsewhere in this report) have already been done: if so, then these analyses should be presented.

Recommendation 2.2.5.2. Existing salinity (both measured and modelled), biofilm and WESA use data should be used to fit empirical models of the association between the three sets of endpoints. These models should then be used (1) to inform the salinity, biofilm and WESA sampling programs (e.g. to identify biologically important subranges for salinity, biofilm attributes or WESA use that have hitherto been inadequately sampled; and (2) make model predictions that can be directly compared to biofilm and WESA use data "after" project construction to assess the prospective predictive value of fitted models with respect to realized project effects.

2.3. BIOFILM STUDY

2.3.1. Study design (Biofilm quantification)

(1) The biofilm quantification study proposes using Westham Island as the control area:

(i) For Westham to be an appropriate control for biofilm, the factors influencing biofilm quality and quantity should be the same as in the Brunswick impact area, absent the project. The information provided in Table 3.2 of the WESA Monitoring Program Design document is, in the ISB's view, not sufficient to infer that such is the case. The factors affecting the salinity regime would seem to be different in the two areas based on the NS modelling results (see Appendix A of the salinity component study, especially Fig. 3.1 and 3.3), an inference supported by the fact that Westham was *not* chosen as a salinity control site. If the salinity regime is different, then it is entirely possible that relative to salinity, other factors affecting biofilm quality or quantity play a more important role in Westham than in Brunswick (or *vice versa*) even in the absence of the project, raising concerns about the appropriateness of Westham as a biofilm control area.

Recommendation 2.3.1.1. The appropriateness of Westham as a biofilm control areas should be assessed through a study explicitly designed to do so. The factors for which there is evidence of an effect on biofilm – including grain size, temperature, grazing pressure, exposure, and water clarity/turbidity - should be explicitly identified and compared between the impact site and the selected control site(s). Where no comparison with respect to an identified factor is done, this should be explicitly stated. Based on these comparisons, an evaluation of the appropriateness of the selected control site for the endpoints in question should be conducted, and the implications of this evaluation to inferences about detected project effects (or lack thereof) explicitly described.

(ii) If the designated control area for biofilm is considered unsuitable as a control site for salinity, it raises questions about the extent to which the proposed salinity and biofilm follow-up programs can be used to explore the hypothesized causal

relationship between salinity and biofilm. For example, what is the inference if (a) using the salinity Before-After design based on field monitoring data, a project effect is inferred but (b) based on the biofilm BACI design, no project effect is inferred? The obvious interpretation that the project has affected salinity but not biofilm is confounded by the fact that these two inferences are based on two different experimental designs with different inferential strength and the fact that in the BACI context, the power to detect ordinal rather than disordinal[16] interactions (on the basis of which project effects are inferred) is often low.

Recommendation 2.3.1.2. The proponent should clearly and explicitly explain how differences in inferential strength of the two designs will be taken into consideration in inferences about project effects on salinity versus biofilm.[17]

(iii) The follow-up program proposes a stratified random sampling design for the ground-truthed sampling, with 10 sites per strata in both the Brunswick impact and Westham control areas (so, 30 sites in total for each area). The strata are defined as areas of expected high, moderate and low biofilm density "based on multiple years of study at Roberts Bank".

Stratified random sampling would be indicated if it was predicted that the project would have different effects in, say, historically high biofilm density areas versus historically low biofilm density areas. In the case of salinity, the NS model clearly predicts different project effects across a spatial gradient in the Brunswick impact area, in which case, stratification by the predicted magnitude of the effect is indicated (see Recommendation 2.2.1.2). Is this also the case for biofilm?

Recommendation 2.3.1.3. An explicit rationale for the choice of a stratified random sampling design for biofilm should be provided. If, as is the case for salinity, project effects in the impact area are predicted to differ among sub-areas that historically have different biofilm densities, then (as suggested in the WESA diet component study, with stratification involving two subareas), stratified random sampling based on historical biofilm density should be deployed in both the control and impact areas using the same strata, in a manner that ensures adequate within-stratum spatial replication.

(2) Because the salinity regime during the spring migration period is largely determined by freshwater discharge from the Fraser River, under the proposed BACI design, the true

[16] For a clear description of ordinal and disordinal interactions, see "Interaction effects" at https://daniellakens.blogspot.com/2020/03/effect-sizes-and-power-for-interactions.html

[17] For salinity, biofilm and WESA, the proposed follow-up sampling designs on which project effects will be inferred differ in their *a priori* inferential strength and power. Such differences clearly have implications to inferences about the *comparative* effect of the project on salinity versus biofilm versus WESA. Any such inferences must therefore explicitly consider these differences in inferential strength and power.

independent sampling unit is year. Because comparatively few years will be sampled before, during and after project completion, power to detect project-induced changes in selected endpoints will likely be comparatively low. Power can be enhanced somewhat by increasing the precision of within-year estimates by increasing the number of within-season sampling events. Currently, few such events are planned - for example, the biofilm study proposes sampling only twice during the season – insufficient to improve on what will otherwise be low power.

The biofilm availability component study proposes to investigate project effects during the time of peak WESA northward migration. This implies that in the proponent's view, project effects on WESA do not accumulate over the migration period (i.e. from mid-April to mid-May,) and do not change migration patterns (e.g. decreases in biofilm or invertebrate quantity or quality resulting in prolonged stopover at Roberts Bank). Moreover, while the responses of biofilm to changes in salinity may be rapid, responses of the invertebrate community to changes in salinity may be slower. Hence it is possible that project-induced changes in the biofilm or invertebrate communities before or after the peak migration period may have (cumulative) effects on population vital rates (e.g. northward migration mortality and/or breeding ground fecundity). If this were the case, then one risks non-detection of project effects by sampling biofilm and invertebrate communities only during the peak migration period. Extending biofilm and invertebrate sampling to periods before and after peak migration not only increases the number of within-season sampling periods but increases the likelihood of detecting such lagged or cumulative effects.

Recommendation 2.3.1.4. Biofilm sampling should be extended to periods before and after peak migration and sampled at least 4 times over this period (one before peak, 2 during, one after peak).

(3) Historical data suggest that invertebrates make up about 50% of the diet of WESA. The ISB notes that although Decision Statement 10.4 specifically concerns the salinity-biofilm-WESA pathway, project-induced changes in biofilm due to salinity changes may have indirect compensatory or synergistic effects on invertebrate abundance or community composition, and hence, invertebrate-derived polyunsaturated fatty acids (PUFA). Although invertebrate sampling will occur in the WESA diet study component, it is not currently proposed in the biofilm study component. Integration of invertebrate sampling with the proposed biofilm study would not only provide critical information on this dietary source of PUFA but will also permit examination of the spatiotemporal associations between the two major dietary sources.

Recommendation 2.3.1.5. Invertebrate sampling should be incorporated into the proposed biofilm study component.

2.3.2. Drone multispectral imagery

(1) The proposal is to use drone multispectral imagery to estimate the biofilm measurement endpoints listed in Table 3.1. But the biofilm follow-up program is supposed to verify predicted effects on biofilm quality and quantity, *not* NDVI. NDVI is clearly a proxy indicator. Which leads to the question: in the context of the biofilm follow-up, why bother with it at all?

This is not clear to us from information provided in the biofilm component study Appendix, but there are several possibilities. One hearkens back to the rationale for the use of NS-modelling in the salinity component study (see s. 2.2.1. (1) above). Field biofilm sites can only be used to estimate biofilm quantity and quality at high spatial resolution, i.e. mg/m^2: such estimates represent *intensive* variables. They cannot of themselves be used to generate reliable estimates of *extensive* variables (e.g. MPB tonnes in each study area, as proposed in the biofilm study component Table 3.1) because of the high spatial variability referred to on p. 14-15: any sufficiently precise estimate of biofilm biomass at these larger spatial scales will require much higher sampling spatial resolution to permit reasonable spatial integration, i.e. many more than the proposed 30 stations/area. Using drone imagery and NDVI permits effectively infinitely fine spatial sampling granularity, but only of proxy data.

A second possibility also relates to the documented high spatial variation in biofilm quantity (and possibly quality). Given this variability, the ISB suspects than an estimate of average biofilm density in both the control and impact areas with sufficient precision to detect anything smaller than a large project effect would require a comparatively large number of sites – again, probably considerably more than the 30 per area that are proposed. Obtaining data from a large number of "virtual" sites via drone multispectral imagery is far less resource intensive than acquiring (and processing!) a larger number of field samples.

The historical data summary document indicates that a pilot project using drone multispectral imagery supported by ground-truthing was conducted in 2021. These data would seem to allow at least an initial estimate of the predictive power of the NDVI-MPB calibration curve. From the description, it is unclear whether Fig. 3.4 of the biofilm component study is derived from these pilot data but, if so, it strongly suggests that predictive value even for individual sampling locations is low. consequently, if a calibration curve like Fig. 3.4 were employed, estimates of the *intensive* endpoints listed in Table 3.1 (e.g. MPB density (mg/m^2) using NDVI would have considerable – possibly large - uncertainty. And this uncertainty will only increase for the *extensive* endpoints proposed in Table 3.1 since such estimates require integration over a large number of point estimates based on the NDVI calibration curve, each of which has an associated uncertainty. The problem here is isomorphic to the problem of using the NS-model to generate extensive salinity measurement endpoints (see section 2.2.1. above).

Finally, the ISB notes that the determination of the initial number of ground-truthed sites is apparently at least partially based on the conclusion "that a sample size of 30 biofilm ground-truthing samples/area is appropriate to calibrate the

spectral data to map and quantify MPB and biofilm nutritional components across the impact area and control area". While the ISB is unsure of what precisely this means, a sample size sufficient to "calibrate the spectral data" is, presumably, *not* the same as a sample size sufficient to achieve a desired power to detect project effects using a BACI design.

Recommendation 2.3.2.1(a). A clear and compelling justification for the drone/NDVI study, including a detailed analysis of the predictive value of NDVI for various biofilm components based on the data available to date, should be provided. A threshold predictive value below which NDVI is considered to be *not* a sufficiently good predictor of a specific biofilm component should be explicitly specified and justified.

Recommendation 2.3.2.1(b). If the predictive value of NDVI is below the specified threshold defined under Recommendation 2.3.2.1 (a): (i) the number and placement of biofilm field sampling sites in both the control and impact areas should be re-evaluated under the presumption that the critical issue is not calibration of the spectral map but rather detection of project effects on biofilm component using exclusively field data; (ii) utility of all extensive endpoints in Table 3.1 whose estimation is based on the NDVI – biofilm component calibration curve should be critically re-evaluated. If the decision is made to retain these endpoints, a detailed description of how their *empirical* accuracy and precision will be determined should be provided (see also Recommendations 2.2.1.1 and 2.2.2.2 above).

(2) The migration of diatoms in the surface of sediment is well-described and is known to affect NDVI measurement. NDVI is also influenced by various environmental conditions that may reduce the predictive value of the NDVI-biofilm calibration curve.

Recommendation 2.3.2.2. If the drone multispectral imaging study is retained, drone overflights should be aligned with the tidal cycle (mid- to low tide); if possible, the timing of flights with respect to tides should be roughly consistent between measurement dates and within a similar tidal window (e.g. 20 minutes either side of low tide), to minimize variation arising from these factors. Weather conditions at the time of overflight should also be recorded and used as potential covariates in fitting NDVI-biofilm calibration curves.

2.3.3. *Statistical analysis of biofilm data*

(1) As noted in the ISB's comments on the salinity study component with respect to NS-model data (see s. 2.2.1 above), it is, in the ISB's view, inappropriate to use NDVI model data to fit statistical models because sample size is arbitrary: for a given year, for a given time, one could in principle have an "observation" at every pixel. The problem does not disappear if one uses survey blocks within area as the sampling unit, because the

number of blocks is itself arbitrary (10 per area are proposed, but it could just as easily be 5, or 20, or …). Consequently, in the ISB's view, NDVI-derived data can only be used only to provide qualitative descriptions of project-induced changes in biofilm (see also Recommendation 2.2.1.1).

Recommendation 2.3.3.1(a). As was the case for the salinity study component, statistical analysis should be conducted only for the *intensive* biofilm endpoints listed in Table 3.1 using field data, not NDVI modelled data.

Recommendation 2.3.3.1(b). If the extensive endpoints listed in Table 3.1 are retained, inferences about project effects should be based on a comparison of estimates that explicitly considers their estimated *empirical* accuracy and precision (see Recommendation 2.2.2.2. above)

(2) Although the biofilm follow-up designs are presented as Before-After/Control-Impact (BACI), in several instances there are really 3 phases - before (B), early (E) construction, and late/after (L/A) construction/operations (see e.g. p. 9) - so not BACI, but BEL/ACI. In particular, although there may be a tendency to pool "E" and "L/A" data to increase sample size (and hence, increase power - all else being equal), there is also the possibility that doing so will reduce power if project effects are significantly lagged or accumulate over time, as larger "L/A" effects might be diluted by smaller "E" effects.

Recommendation 2.3.3.2. In fitting BACI models, the 3 phases (B, E, L/A) should, at least initially, be explicitly considered. Pooling of E and L/A data should be well-justified, not just statistically but biologically.

(3) The descriptions provided (see ss. 3.1.6.4 and 3.2.6.3 of the biofilm study component) suggest that conclusions about project effects will be based on hypothesis testing of fitted models. Such an approach necessarily gives rise to concerns about statistical power. The ISB's view is that information theoretic approaches to inference based on estimated effect sizes may be more appropriate, but an informed opinion on this issue will have to wait for the proponent's characterization of thresholds (expected in Tranche 2).

Recommendation 2.3.3.3: The use of hypothesis-testing versus information-theoretic approaches to inference about project effects should be explored.[18] A clear and convincing rationale should be provided for whatever approach is ultimately adopted.

[18] See, for example: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612111/

2.3.4. Biofilm characterization

For comments on biofilm characterization, see section 3.2 below.

2.3.5. Influence of environmental factors (including salinity) on biofilm-associated PUFA

2.3.5.1. Study design

(1) Westham *is* proposed as a control site for the study of the causal relationship between salinity fluctuations and PUFA but is *not* proposed as a control site for project-induced salinity changes. The implication is that although the determinants of the salinity regime in Westham *are different* from those in the Brunswick impact area even in the absence of the project (hence, Westham is not an appropriate control site for salinity), the effect of the salinity regime on biofilm PUFA in the two areas is the same. This is equivalent to the assumption that factors *other than salinity* (temperature, exposure, nutrients, etc.) that affect PUFA are the same in the two areas (if they are different, then Westham is not an appropriate control for inferring project-related effects on the relationship between salinity fluctuations and PUFA). What is the evidence that this is indeed the case?

Moreover, it is unclear to us why one needs – or even wants – a BACI design here. The hypothesis being tested – and the one on which, presumably, predictions about project -induced effects on PUFA are based – is that "… that large salinity changes (oscillations of >20 PSU within a tidal cycle) are required for biofilm to produce polyunsaturated fatty acids (PUFAs)" (salinity component study, Appendix A, p. 15). As stated, this hypothesis is *not* explicitly about project effects. Consequently, one doesn't need a BACI design to test it. All one needs are a sample of carefully selected sites which vary in their tidal cycle salinity range, either spatially or temporally (or both), and for which one has measures of biofilm quantity and PUFA. Indeed, sites should be selected to maximize the range of the predictor variable (salinity range), as this will maximize the precision of any fitted models. It may be that the 2 proposed sites in Westham are required to maximize this range. It may also be that focusing only on the 5 Brunswick sites close to shore will *not* maximize the range. The point is that to test the hypothesis, this is the criterion that should be used to choose (a) existing sites; and/or (b) select new ones, not whether " the selected salinity monitoring stations … overlap with areas supporting productive PUFA-rich biofilm that are heavily used by foraging WESA during northward migration", as a sample based predominantly on these stations may well reduce the range of variation in salinity range below what is required.

Recommendation 2.3.5.1. Sites to test the hypothesis linking salinity fluctuations and biofilm should be selected to maximize the range of tidal cycle salinity range values (TCSR).[19] At the same set of sites, biofilm sampling to determine PUFA should be

[19] One possible approach is to use historical data from the salinity stations to characterize empirically the relationship between tidal cycle salinity range (TCSR) and PUFA as well as determine spatial variation in this

conducted. These data should then be used to fit models that predict change in PUFA at each salinity station in the sample as a result of (predicted) project-induced salinity changes. These predictions should then be compared with observed changes during early and late construction/operation to evaluate the predictive value of the models.

3. WESA DIET STUDY COMPONENT

3.1. Study design

(1) Boundary Bay is *not* identified as a control area for either the salinity or biofilm study components. The implication is that - absent the project - the factors affecting WESA diet in Boundary Bay are the same as those affecting WESA diet in the Brunswick impact area, but for biofilm, they are not (else presumably Boundary would also be designated as a control area for biofilm). Thus, one is in the position of arguing that even though the drivers of biofilm quality and quantity are different in Boundary Bay and Brunswick, the effect of changes in biofilm quantity or quality on WESA diet are (more or less) the same in both. Though this is logically possible, it would not be the case if, for example, the factors affecting invertebrate abundance differed between the two. (The ISB notes that, based on historical data, Boundary Bay is much less used by WESA than Westham, and even less so in comparison to Brunswick. If use is related to invertebrate abundance, this might indicate that this assumption is invalid.)

(2) The addition of a second control area also introduces some complexities into the statistical analysis: in fitting a BACI model, one now has a categorical variable "Area" with 3 levels (impact, control 1, control 2) rather than two (control, impact), so that one now generates estimates of changes over time for the impact site and two control sites. Suppose that relative to the change over time at control 1 (say, Westham), the change at Brunswick is large, but negligible compared to control 2 (say, Boundary Bay): what is the conclusion?

Recommendation 3.1.1. The WESA diet study focus on the Westham control area and that the sampling effort that would otherwise be allocated to Boundary Bay be reallocated to Westham and Brunswick in a manner that provides the largest increase in power.

(3) The selection of Westham as a control for the dietary component study also gives us pause. Westham and Brunswick are immediately adjacent to one another, and it is

range among the 11 historical stations. Then determine whether the minimum of this range is sufficient to capture the largest predicted project-related reduction ("compression") in TCSR. If the historical salinity stations do not provide TCSRs that substantially exceed the entire range of predicted project-induced compressions, select locations for new salinity stations to ensure that this is the case: it may be that additional stations in Westham are required to ensure that locations with small TCSRs are included in the sample.

entirely possible that WESA move back and forth between the two on comparatively short time scales (e.g. minutes to hours). If so, fecal samples taken from either may well represent composites from both places. If the same birds are contributing to fecal samples in both areas, they are not independent samples. Both possibilities will reduce the likelihood of detecting project-induced differences between control and impact areas, the former by reducing potential differences between areas, the latter by inflating the true degrees of freedom for hypothesis-testing (i.e. power is overestimated).

Recommendation 3.1.2. The WESA follow-up should incorporate a movement study, using full time automated radio tracking of individuals tagged at each area (e.g., using Motus towers to detect movements of birds fitted with nanotags). Such a study is critical for determining the extent to which fecal samples from Westham and Brunswick are indeed independent. [20]

(4) It is also proposed that the impact area be subdivided into two sub-areas. The stated rationale is that "this division was made to allow separate assessment of WESA diets within each sub-area, given that the two sub-areas differ in the extent of predicted project-related changes to salinity and subsequently the potential to affect biofilm." (p. 8).

This suggests that within the Brunswick control area, a spatial gradient in project-induced biofilm effects is predicted, underscoring the importance of Recommendation 2.3.1.3. In this case, there are then 2 sets of biofilm sites: those based on stratification per Recommendation 2.3.1.3, and those based on where WESA are feeding. The former in principle allows for inferences about the impacts of the project on biofilm quantity and quality in general, the latter for inferences about project effects on biofilm quantity and quality where WESA are feeding. Since birds are expected to respond to small scale variation in habitat quality by moving to where habitat quality is highest, the difference – if any – between the distribution of biofilm endpoints in these two samples may permit inferences about project effects on WESA (biofilm) *habitat selection* (i.e. large differences in these two distributions suggest habitat selection, negligible differences no selection). Absent such a comparison, the proposed study may well underestimate project effects because of the potential for WESA to concentrate feeding in areas where food supply is the greatest, which may well be where project effects are the smallest.

[20] In Recommendation 3.1.1, the ISB recommends that Boundary Bay not be used as a control area for WESA. However, because Westham is immediately adjacent to Brunswick, implementation of Recommendation 3.1.2 may well indicate a frequency of movement between Westham and Brunswick that raises grave concerns about sample independence. On the other hand, because Boundary Bay is spatially disjunct from Brunswick, movement frequency between the two may be substantially less than between Westham and Brunswick, in which case, Boundary may well be a better control for WESA that Westham, Recommendation 3.1.1 notwithstanding. Any such determination will likely to require additional tagging at Boundary Bay. Finally, the ISB notes that if tagging indicates substantial lack of independence between Brunswick and Westham, and Westham and Boundary Bay, then there is, in the ISB's view, no appropriate control area for WESA, in which case inferences about project effects will be based on what amounts to a BEL/A design associated with sampling over time in Westham – with associated implications to inferential strength.

Recommendation 3.1.3. The proponent examines the difference between "background" biofilm quality and quantity estimated from the biofilm component study, and the "selected" biofilm quality and quantity based on the WESA diet study to estimate project effects on food availability and quality. (*N.B.* If, as suggested in Recommendation 2.3.1.5, invertebrate sampling is incorporated in the biofilm study, a similar analysis should be done for invertebrates.)

3.2. Laboratory methods

(1) Although HPLC can be used for fatty acid analysis, gas chromatography (GC) is the more common way to measure fatty acid composition and concentrations of a mixture (used across the board by the food industry, for example). HPLC is generally considered the method of choice to quantify important biofilm components like pigments (as proposed in the biofilm component study), but it is not the method of choice for fatty acid analyses. Overall, GC is more accurate, more readily available in laboratories in general, and less expensive.

Recommendation 3.2.1: GC should be employed for WESA diet fatty acid characterization.

(2) The parameters selected to assess the lipid components of the diet (total fat, total fatty acid concentration, and total PUFA concentration) are only weak indicators of diet quality. Standard GC (or HPLC) analyses conveniently yield concentrations for all individual fatty acids. Such information is crucial to assess diet quality and actual lipid composition. The availability of specific long chain (in the omega-3 and omega 6 families) PUFAs is particularly important for long-distance migrant birds that rely on them not only for energy, but for key physiological effects on endurance capacity.

Recommendation 3.2.2(a). Long chain PUFA of the omega 3 (in the omega-3 (or n-3) and omega 6 (or n-6) PUFA families should be quantified. Particular attention should be given to eicosapentaenoic acid (n-3 20:5 or EPA), docosahexaenoic acid (n-3 22:6 or DHA), and arachidonic acid (n-6 20:4 or ARA).

Recommendation 3.2.2(b). The ratio of total n-3 acids / total n-6 acids should be used as a measurement endpoint in diet analysis as it is an excellent indicator of dietary lipid quality for a migrant bird (note here that the optimal n-3/n-6 ratio for WESA will most likely be different than for other migrant bird species). This ratio should be estimated in biofilm, but also more importantly in the invertebrate WESA diet that may well provide the bulk of these critical PUFAs.

(3) Although the use of Bayesian mixing models such as MixSiar to determine diet composition has become standard practice in trophic ecology, these models are known to have several important limitations, including especially the sensitivity to uninformed

or incorrect DTDFs, especially for diet items implicating multiple trophic levels (e.g. invertebrates).

Recommendation 3.2.3. The rationale for DTDF selection should be explicitly stated, and ideally, should explore the sensitivity of inferences about project effects to DTDF uncertainty, and should be supported by independent BACI comparison of isotope variability.

(4) The WESA diet study states that "If sulfur content within WESA prey and droppings is sufficient for determination of $\delta^{34}S$ stable isotope signatures and this parameter (in addition to δ12C and δ15N) improves confidence in estimates of diet composition, models will be adapted relative to prior WESA diet studies." (p. 25). How will any improvements to confidence in estimates of diet composition be assessed?

Recommendation 3.2.4. How the extent to which incorporation of $\delta^{34}S$ "improves confidence" is determined, and what threshold of improvement will be used to decide whether relevant mixing models will be adapted to include $\delta^{34}S$ signatures, should be explicitly described.