

Assessment and redevelopment of visual survey expansion factors and a summary of current interpolation methods for Fraser River Sockeye (*Oncorhynchus nerka*) spawner estimates

Michael Arbeider, Brian Leaf, Tanya Vivian, Sean Everitt, Marissa Glavas, Brit van Amerom, Daniel Doutaz, Kaitlyn Dionne, A. Scott Decker

Fraser and Interior Area Stock Assessment
Pacific Region
Fisheries and Oceans Canada
985 McGill Pl.
Kamloops, BC V2C 6X6

2026

**Canadian Technical Report of
Fisheries and Aquatic Sciences 3767**



Fisheries and Oceans
Canada

Pêches et Océans
Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. Il n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques*.

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of
Fisheries and Aquatic Sciences 3767

2026

Assessment and redevelopment of visual survey expansion factors and a summary of
current interpolation methods for Fraser River Sockeye (*Oncorhynchus nerka*) spawner
estimates

by

Michael Arbeider¹, Brian Leaf¹, Tanya Vivian¹, Sean Everitt¹, Marissa Glavas²,
Brit van Amerom¹, Daniel Doutaz¹, Kaitlyn Dionne², A. Scott Decker¹

¹Fraser and Interior Area Stock Assessment
Pacific Region
Fisheries and Oceans Canada
985 McGill Pl.
Kamloops, BC V2C 6X6

²Fraser and Interior Area Stock Assessment
Pacific Region
Fisheries and Oceans Canada
100 Annacis Pkwy Unit 3
Delta, BC V3M 6A2

© His Majesty the King in Right of Canada, as represented by the Minister of the Department of Fisheries and Oceans Canada, 2026

This work is licensed under the [Open Government Licence](#)

Cat. No. Fs 97-6/3767E-PDF ISBN 978-0-660-99278-5 ISSN 1488-5379

<https://doi.org/10.60825/1kyf-kv54>

Correct citation for this publication:

Arbeider, M., Leaf, B., Vivian, T., Everitt, S., Glavas, M., van Amerom, B., Doutaz, D., Dionne, K., Decker, A.S. 2026. Assessment and redevelopment of visual survey expansion factors and a summary of current interpolation methods for Fraser River Sockeye (*Oncorhynchus nerka*) spawner estimates. Can. Tech. Rep. Fish. Aquat. Sci. 3767: vi + 119 p.
<https://doi.org/10.60825/1kyf-kv54>

Table of Contents

ABSTRACT	v
RESUME	vi
1. Background.....	1
1.1 Visual Estimates and Fraser Sockeye Salmon.....	1
2. Methods.....	5
2.1 Study Area.....	5
2.2 Data.....	5
2.2.1. High-accuracy and Visual Methods	8
2.3 Updated Expansion Methods.....	12
2.3.1 Approach	12
2.3.2 Model Validation and Selection	17
2.3.3 Application to 2024 Data and Additional Decision Rules based on Expert Advice	20
2.4 Other Expansions, Infills, and Data Considerations	20
2.4.1 Biological Interpolations and Infills	20
2.4.2 Fence Counts.....	21
2.4.3 Hydroacoustic Estimates.....	21
2.4.4 Mark-Recapture Estimates.....	22
2.4.5 Lake Spawning Estimates	23
2.4.6 North Thompson & Other Outliers	24
2.4.7 Prior Specialized Peak Live Expansions	25
3. Results.....	25
3.1 Updated Expansion Approach	25
3.1.1. Candidate Models and Model Ranks.....	25
3.1.2. Expanded Estimates	27
3.1.3. Application to 2024 Data & Additional Expansion Rules based on Expert Advice.....	33
4. Discussion	39
5. Conclusion	42
6. Acknowledgements.....	45
7. References	46
Appendix A. Data, Code, and Summary Tables	52
Appendix B. Model Specifications, Diagnostics, and Parameter Results.....	55
Mean-of-Ratios.....	55
Ratio-of-Means.....	59

Inverse Regression.....	63
Linear and Generalized Linear Regression.....	67
Generalized (1).....	69
Size-Aggregate (2).....	74
Individual (5)	80
Early Stuart Specialized (6).....	81
Discharge Specialized (6)	84
Appendix C. Rank-Set Dashboards and Test Results	90
Appendix D. Other Expansions, Infills, And Data Considerations	112
Appendix E. Figures Used in Out-of-Sample Range Model Implementation Discussions	114

ABSTRACT

Arbeider, M., Leaf, B., Vivian, T., Everitt, S., Glavas, M., van Amerom, B., Doutaz, D., Dionne, K., Decker, A.S. 2026. Assessment and redevelopment of visual survey expansion factors and a summary of current interpolation methods for Fraser River Sockeye (*Oncorhynchus nerka*) spawner estimates. Can. Tech. Rep. Fish. Aquat. Sci. 3767: vi + 119 p. <https://doi.org/10.60825/1kyf-kv54>

This document provides a systematic assessment and update of expansion factors and overview of interpolation methods for estimating Fraser River Sockeye (*Oncorhynchus nerka*) spawning abundance (escapements). Accurate spawner data are fundamental for Pacific salmon management and conservation, impacting implementation of policies and agreements like the *Species at Risk Act*, *Wild Salmon Policy*, and *Pacific Salmon Treaty*. Historically, visual survey methods were prone to bias and uncertain accuracy due to application of a generalized expansion factor of 1.8 for Fraser River Sockeye. Spawner estimates were altered post-hoc to account for expected biases through a Run-Size-Adjustment process facilitated by the Fraser River Panel Technical Committee; which identified a need for this reassessment. This report introduces a methodical approach to select system-characteristic-specific expansion methods, using a paired-estimates database (n=162) and ranking models' performance. The updated approach reduced bias (mean-percent-error) from -13.3% to -0.5% and improved accuracy (mean-absolute-percent-error) from 30.2% to 24.5%. Precision of estimates was overestimated. Standardized application rules to account for caveats within the framework were developed from test application to the 2024 data. Alternative approaches or adjustments may be required when systems exceed the characteristics represented in the database or where the system characteristics consistently resulted in bias or low accuracy.

RÉSUMÉ

Arbeider, M., Leaf, B., Vivian, T., Everitt, S., Glavas, M., van Amerom, B., Doutaz, D., Dionne, K., Decker, A.S. 2026. Assessment and redevelopment of visual survey expansion factors and a summary of current interpolation methods for Fraser River Sockeye (*Oncorhynchus nerka*) spawner estimates. Can. Tech. Rep. Fish. Aquat. Sci. 3767: vi + 119 p. <https://doi.org/10.60825/1kyf-kv54>

Ce document présente une évaluation systématique et une mise à jour des facteurs d'expansion ainsi qu'un aperçu des méthodes d'interpolation pour estimer l'abondance des géniteurs (montaisons) du saumon rouge (*Oncorhynchus nerka*) du fleuve Fraser. Des données exactes sur l'abondance des géniteurs sont essentielles à la gestion et à la conservation du saumon du Pacifique car elles ont une incidence sur la mise en œuvre de la Loi sur les espèces en péril, la politique concernant le saumon sauvage ainsi que le Traité sur le saumon du Pacifique. Historiquement, les méthodes de recensement visuel étaient sujettes aux biais et à une précision incertaine en raison de l'application d'un facteur d'expansion généralisé de 1,8 pour le saumon rouge du fleuve Fraser. Pour tenir compte des biais attendus, les estimations des géniteurs ont été modifiées avec un processus d'ajustement post hoc de la taille des remontées, facilité par le comité technique du conseil du fleuve Fraser qui a identifié le besoin de réévaluer les facteurs d'expansion. Ce rapport présente une approche rigoureuse pour sélectionner des méthodes d'expansion adaptées aux caractéristiques des rivières recensées à l'aide d'une base de données d'estimations appariées (n = 162) et en classant les modèles selon leur performance. Cette nouvelle approche a permis de réduire le biais (erreur moyenne en pourcentage) de -13,3 % à -0,5 % et d'améliorer l'exactitude (erreur absolue moyenne en pourcentage) de 30,2 % à 24,5 %. La précision des estimations était surestimée. L'approche a été testée sur les données collectées en 2024, permettant d'élaborer des règles d'application standardisées tenant compte des limitations. Des approches alternatives ou des ajustements pourront être nécessaires si les rivières recensées présentent des caractéristiques différentes de celles représentées dans la base de données, ou si ces caractéristiques entraînent systématiquement un biais ou une faible précision.

1. BACKGROUND

1.1 VISUAL ESTIMATES AND FRASER SOCKEYE SALMON

The management and conservation of Pacific salmon (*Oncorhynchus* spp.) requires accurate and precise information on salmon populations (DFO 2005, PSC 2019). Pacific salmon are an anadromous, semelparous species (excluding *O. mykiss*; which is sometimes referred to as a salmon or anadromous trout) that reproduce (spawn) in freshwater systems ranging from small streams to very large rivers and lakes. Estimating their spawning abundance (also referred to as escapement) in fresh water is integral to many domestic and international policies and treaties such as the *Species at Risk Act* that manages populations as Designatable Units (DU) (DFO 2007), the *Wild Salmon Policy* that manages populations as Conservation Units (CUs) (DFO 2005), and the *Pacific Salmon Treaty* that manages populations as Management Units (MUs) (PSC 2019). Sockeye salmon (*O. nerka*) hold ecological, cultural, and economic value and changes in their spawning abundance directly impact conservation status, harvest allocations, and rebuilding plans. The number of spawners and trends in their abundance are foundational metrics in determining population status (Holt 2009) and spawner estimates are integral to classical salmon stock-recruitment analyses that are often used in determining management reference points (Hilborn and Walters 1992, Holt and Bradford 2011). Spawner estimates of higher precision and accuracy allow for more definitive population assessments and lower uncertainty for fisheries management decisions. A managers' confidence in scientific advice may also lay along a spectrum where high-accuracy-abundance estimate types provide greater certainty in science-informed decisions than low-accuracy-abundance estimate types or indices. For example, in review of previous *Wild Salmon Policy* integrated status assessments, it was found that the final conservation status was given a higher confidence rating when metrics that used high accuracy and high precision abundance were available (Pestal *et al.* 2023, DFO 2024). Higher-accuracy data can result in more discrete management reference points and the ability to detect trends more reliably, which can result in clearer management decisions. Methodologies that provide high quality (defined here as having high accuracy and precision) data however, are often costly or logistically restrictive (Parsons and Skalski 2010).

Accounting for bias and variability inherent in salmon abundance indices with strategic analysis is one way to increase their accuracy with relatively minimal cost. In spawning ground enumeration for salmon, one of the most common methods employed is the peak count visual estimate ((Parsons and Skalski 2010). In peak count enumeration methods, observers visually count live and dead salmon around a population's historical average spawning dates. Raw peak counts are normally biased low because observer efficiency is imperfect and there are often other biological factors that make counting all of the salmon that spawn over the course of an entire season impossible to do in a single day (Symons and Waldichuck 1984, Andrew and Webb 1987, Farwell *et al.* 1999, Bailey *et al.* 2000). These visually based estimates are typically expanded using various analytical methods to reduce the persistent negative bias so that they may be treated as closer to a true-abundance population estimate (i.e. absolute abundance) than an index (i.e. relative abundance). The primary methods for estimating expansion factors require the pairing of high-quality estimates and visual estimates (hereafter referred to as paired data points), which may produce region or water body specific expansion methods (Parsons and Skalski 2010). One shortcoming of expansion methods is that the estimates of precision are either not robust, not reported, or reported qualitatively as "low-precision" (Parsons and Skalski

2010). They are also often difficult to generalize to the diversity of water bodies (or *systems*, which can include streams, rivers, or lake shores) and conditions experienced by observers and require large quantities of data to account for system variability.

Historically in the Fraser River watershed, expansion factors have been developed for specific species and/or stream conditions. There are different expansion factors across and often within species and may even vary annually. For example, there is a general Chinook salmon (*O. tshawytscha*) program expansion factor for peak counts done with aerial surveys (Farwell et al. 1999; Bailey et al. 2000), but there are also system-specific expansion factors for indicator populations that have been identified by the Chinook Technical Committee (CTC) of the Pacific Salmon Commission (PSC). The system-specific expansions are treated as true-abundance and high quality information and the general expansion is treated as an index but recognized to be fairly accurate (Parken et al. 2003; JCTC 2023). The system-specific expansion factors also create an estimate of precision associated with the expanded estimate that meets the precision guidelines set out by the CTC. The Fraser River Area Coho salmon (*O. kisutch*) program expansion factors may be modified annually based on peak-count data that is part of an Area-Under-the-Curve estimate (Holt and Cox 2008); which can account for annual variation in observer ability or conditions (*unpublished*, M. Arbeider pers comm). For Fraser River Sockeye salmon (*O. nerka*), two general expansion factors are commonly used for visual-based methods: one for lake surveys and another for stream surveys. Fraser River Sockeye salmon expansions also include a cumulative dead count to their peak live count when available, which accounts for more spawning behavior variability than simple peak counts. Certain population groups have water body or year-specific paired data points that account for area variability allowing for the use of a more specialized expansion factor, but these cases are more rare.

The focus of this report is to assess and update expansion factors for visual-based peak-count methods used to estimate Sockeye salmon abundance in streams, rivers, and for some lake shores (subsequently referred to collectively as *systems*) in the Fraser River watershed. The methods used to estimate Fraser River Sockeye vary based on a two-tiered predicted-abundance-based system; where visual methods are primarily used to enumerate populations with low expected returns while higher accuracy and precision methods are used for populations with higher expected returns (Schubert and Houtman 2007). The delineation of a *low* spawner abundance has changed through time. Originally, a low predicted abundance was defined as <25,000, but since 2004, it was changed to <75,000 due to changes in management mandates and constraints from funding and logistics (Welch *et al.* 2007). Some assessments deviate from this guideline and implement high accuracy methods on low abundances due to conservation concerns (Selbie *et al.* 2022), mandates that demand high-quality information, or because visual assessments have been deemed unviable. Conversely, there are occasions when high-accuracy methods face operational restrictions and visual-based estimates must be used as substitutes. Overall, the tiered system has generally resulted in MUs and CUs having different proportions of their aggregate abundance being estimated using visual-based estimates (Table 1) due to the variability in their stream and population characteristics. This difference in the proportion of the population being estimated with visually-based estimates has also influenced where effort was expended in collecting paired data-points for the purpose of estimating expansion factors.

Table 1. The mean proportion of the total spawner abundance estimated using visual-based methods for Fraser River Sockeye salmon Management Units in two periods. The “<25K” period data ranges from 1938-2003 while the “<75K” period ranges 2004-2023. The periods differ between their definitions of a “low abundance” population within the two-tiered enumeration method planning process.

Management Unit	<25K & prior 2004			<75K & 2004 and after		
	Mean	Minimum	Maximum	Mean	Minimum	Maximum
Early Stuart	77%	9%	100%	92%	49%	100%
Early Summer	28%	0%	100%	40%	11%	83%
Summer	12%	0%	98%	10%	3%	31%
Late	10%	0%	87%	28%	6%	73%

The existing, default expansion factor for Sockeye in the Fraser River has been flagged as problematic by DFO and the PSC through review in their joint Fraser River Panel Technical Committee (Patterson et al. in press). The default expansion factor is 1.8 and based on a few paired samples from small sized, clear (i.e. not turbid nor tannic) streams (Woodey 1984, Schubert and Houtman 2007). This expansion factor is multiplied with the sum of the single largest live count and cumulative dead counts up to the date of the largest (peak) live count to formulate the expanded estimate. It is currently used as the default expansion factor for visual surveys of Sockeye Salmon across the Fraser River watershed with some exceptions, such as Early Stuart systems and a handful of other systems. Prior analysis of subsequent paired-count data from a wider variety of systems has indicated that the 1.8 expansion factor is not always appropriate and often results in negatively biased estimates (Welch et al. 2007), particularly when there are higher abundances (e.g. >25,000) or water clarity is turbid or tannic. Since 2009, assumed biases have been addressed in the Fraser Panel Technical Committee Run-Size-Adjustment (RSA) process; however, there is unquantified uncertainty associated with the adjustments made to estimates nor validation of the magnitude of the corrections (Robinson et al. in pre).

One objective of this report is to create a methodical approach to identify what affects expansion factors and various methods to account for them. The two general elements that affect visual counting methods are observer efficiency and salmon behaviour (Holt and Cox 2008). Observer efficiency is impacted by a range of environmental factors that may affect an observer’s ability to see individual salmon in the water such as water clarity, flow, stream width/size, brightness (how sunny it is), glare, wind ripple, shadowing or canopy encroachment, or the presence of deep pools and large woody debris (Bevan 1961, Jones et al. 1998, Parken et al. 2004). An observer’s survey experience and overall ability to detect fish may also affect observer efficiency, but this is usually accounted for by training observers and only selecting experienced and consistent observers for larger or more difficult counts. Additionally, the method of observation (e.g. by foot, raft, or helicopter) affects an observer’s ability to observe salmon due to the angles of observation or total area being observed in a single instance (Shardlow et al. 1987, Dangel and Jones 1988). The other element, fish behaviour, can move an expected peak of spawn date out of a planned survey window or interact with observer efficiency. Arrival timing and behaviour in salmon can vary annually, which can result in unpredictable abundance pulses that are incongruent with planned survey schedules (Hilborn et al. 1999, Korman et al. 2002, Holt and Cox 2008). Arrival timing can be a simple shift in the peak spawning period or it may be a protraction of the spawning, which results in no evident peak live count but a gradual or

steady increase in carcasses. These variations are usually only observed if a large number of surveys are executed (Holt and Cox 2008), which can undermine the cost savings of a visual-based method. Other behaviours like holding in deep pools or lakes, or under large woody debris and undercut banks can also change the number of individual fish observed or an observer's efficiency in seeing them. Spawning density may also affect observer's efficiency (Murdoch *et al.* 2019), where an increase in density can increase individual detection errors due to salmon crowding and overlapping. Pre-season planning aims to prepare surveyors for these inevitabilities that may impact visual counts. Variability is buffered through the occurrence of multiple surveys, comparisons from multiple observers, and flexibility in survey dates to avoid poor survey conditions among other methods. However, there are also various analytical approaches to account for variability in expansion factors and reduce biases caused by certain conditions (Parsons and Skalski 2010); as long as there are data to support the approach.

Part of this report's primary purpose is to create a standardized approach for selecting the most appropriate expansion factor or method given a system's conditions based on retrospective analysis. Here, we define "most appropriate" as the selection of a method that results in the least amount of bias and error relative to a high-quality estimate on average across an existing time series of paired data points. An ancillary objective is to develop methods where an estimate's precision can be explicitly reported instead of relying on a qualitative labels such as "low-precision" or "high-precision." Within Pacific Salmon management, one-step-ahead approaches are often implemented for selecting models to reduce error in abundance forecasts (MacDonald and Grant 2012, Velez-Espino *et al.* 2019, Liu and Xu 2024). The retrospective forecasts (i.e. re-running models with data up to Year_{*i*} and forecasting to Year_{*i*+1} up to the current year) are used to estimate a suite of retrospective performance metrics, whose combined rank is used to select which model performed best historically (Velez-Espino *et al.* 2019). In the case of the visual expansion database, a leave-one-out (LOO) approach was deemed more appropriate since the datapoints are not part of a timeseries and sample sizes are low, but performance metrics can be calculated similarly as with the forecast approaches (Folkes *et al.* 2018). The database may require aggregating or separating the data for model inputs however, depending on availability of covariates or other shared or exclusive aspects of each pair of data. A model's inputs may then limit which systems it can be applied to; which requires an amended approach in the model selection process relative to the forecast approach.

As a secondary goal to updating visual expansion methods, this report will also document and describe other scenarios where incomplete survey information requires expansion, extrapolation, interpolation, or infilling. These scenarios include instances where data was intended to be recorded, but was not due to equipment malfunction, interference, or misalignment, etc. Other enumeration methods also rely on sub-sampling a population and interpolating or extrapolating variables from the sub-sample to the rest of the population. Then there are cases of system-specific approaches, such as estimating North Thompson Sockeye, that are external to the paired-points database that require documentation. Summarizing these methods will improve upon the posterity of Fraser River and Interior Area Sockeye estimation methods.

The objectives of this report are to i) summarize and publish the expansion paired-points database; ii) develop a suite of updated expansion factors and models that reduce bias on average and estimate precision; iii) create a standardized model selection framework to apply the most appropriate expansion factors and models based on a system's characteristics; iv)

verify the model selection algorithm and outputs using recent data and garner expert-driven advice to establish standardized rules for application to systems whose counts or characteristics exceed the paired-points database; and iv) summarize other expansion or interpolation methods currently in use by the Fraser and Interior Area Sockeye Stock Assessment group.

2. METHODS

2.1 STUDY AREA

The expansion analysis uses data collected in the Fraser River watershed. The Fraser River basin has a catchment size of roughly 230,000 km², making it the largest river system entirely in British Columbia and the fourth largest in Canada by discharge. The mainstem flows approximately 1,400 km from its headwaters in the Canadian Rocky Mountains near the British Columbia-Alberta border to its outlet in the Strait of Georgia; initially flowing northwest before making a sharp turn south near Prince George, and continuing its southward course until reaching Hope, where it veers west towards the Pacific Ocean. The Fraser River basin is both geographically and climatically diverse as it spans numerous biogeoclimatic zones, including alpine and mountainous regions, dry interior forests, steep canyons, grasslands and wet coastal forests, resulting in highly variable precipitation and temperature regimes throughout (Schnorbus *et al.* 2010). The watershed is predominantly snowmelt driven with a mean annual discharge of around 3,600 m³/s and is fed by several large tributaries including the Nechako, Chilcotin, Quesnel, Thompson and Harrison rivers. This hydrological regime varies across sub-basins however, with low-elevation regions near the coast being more heavily dominated by rainfall in the fall and winter months (Martins *et al.* 2023). This diverse set of environmental conditions is reflected in conditions encountered during spawning ground assessments of Fraser River Sockeye.

2.2 DATA

The Fraser River watershed supports 24 current (or “verified”) Fraser River Sockeye CUs. Escapement assessments are conducted on an annual basis for all 24 CUs, and 14 of the CUs have representation within the paired-count dataset. Several years of funding from the Pacific Salmon Commission Southern Boundary Endowment Fund resulted in these paired data-points being collected to cover a greater diversity of system conditions than what was available historically (Welch and Leaf 2019). Low sample sizes however, still result in unrobust expansion factors using the traditional methods (Welch and Leaf 2019). Part of this standardization approach will propose guidelines to define the levels of robustness for implementing expansion factor methods. These results may also inform planning of future data collection for the paired-points database; particularly those that may not currently have good representation or result in low accuracy estimates.

There are 162 high-accuracy/visual-count data pairs in this analysis, but due to historical priorities and contexts, samples are not evenly distributed between CUs, through time, or across magnitudes in abundance (Appendix A). Three streams from the Takla-Trembleur-Estu (Early Stuart) CU and one stream, the Stellako River, from the Francois-Fraser-S CU comprise about half of the dataset ($n = 83$). Multiple unpublished (and some published) assessments of the quantity and quality of the paired-count data have been made over time (Welch and Leaf 2019). These prior assessments identified major differences among systems that influenced observer

efficiency and average expansion factors between high-accuracy estimates and visual estimates.

Two cases of pseudo-replication were considered in this data set, but the ability to address them was limited. Some paired-data points share the same high-accuracy estimate in the case that both a ground and aerial survey were completed in the same year, in conjunction with the high-accuracy method. The other case involved repeated observations from a single stream (e.g. Stellako $n = 35$). A common method to account for pseudo-replication in data is with a random effect (Hurlbert 1984, Millar and Anderson 2004). Within this dataset, the level of the random effect would have to be at the system, year, or system-year level; however, many systems and years (and particularly system-years) do not have sample sizes greater than one or two. The limited sample sizes would result in difficulties for the model differentiating between the random effect and the general residual error (Maas and Hox 2004). In practice, the models may not-converge or would be unstable. In the first case, the inclusion of the survey method covariate would be expected to explain variation between expansion factors derived from the two visual methods paired with the same high-accuracy estimate. In the case of systems with greater sample sizes, it must be taken as a caveat that they will have a disproportionately large influence on expansion factor or parameter estimation in general models. However, individual-system based models were also explored as a way to investigate both their effects on covariate parameters as well as to investigate if individual-system models were robust enough to be applied.

There are many covariates that have been collected for each visual survey (Appendix A). In many instances, not all covariates were collected, or they were not retrievable for all historical data. Differences in the availability of paired-data points with covariate data was one aspect that guided model creation. For example, not all systems have a monitoring station to measure discharge; therefore, only a small portion of the 162 data points are linked to a discharge covariate. This resulted in different sets of paired-data points, with different sample sizes, that either shared or excluded certain covariates. The expectancy of a covariate's retrievability and its correlation with other covariates in a set of data were used to determine if it would be included in a given model's development. Some unique covariates also determined specialized model development in the case where the more unique covariate could describe variation in observer efficiency and expansion factors more directly, whereas more general covariates, or easily retrievable covariates, were used in developing models that could be applied to most historical data scenarios.

There are seven covariates (categories bolded; levels italicized) that are expected to be retrievable for most historical and future data. The visual **survey method** is standardized as either *ground* or *aerial*. *Foot*, *raft* and *boat* are grouped as *ground* surveys due to the similarity in angle of observation to the water surface relative to *aerial* surveys. There are exceptions where a survey is a combination of *ground* and *aerial* due to logistical accessibility constraints; in these cases, the portion of the survey that accounts for most of the raw count is used to assign the survey method (e.g. if 50.1+% of the count was from the ground portion of the survey, the survey is assigned as ground). **Stream size** is a constant based on the average individual stream morphologies: wetted width and depth (see ReadMe_for_calibration_database.xlsx for more details). Previously, five stream categories were used (*very small*, *small*, *medium*, *large*, *x-large*), but due to similarities from a survey perspective (e.g., the entire wetted width of both *very small* and *small* streams can be readily

viewed by a single surveyor) and low sample sizes, the categories of *very small* (< 5 m wetted width) and *small* (5-10 m wetted width) were combined into a single category of *small* (< 10 m wetted width). Stream sizes generally correlate with stream order but there is some overlap in orders based on the actual important characteristics (width and depth) of the survey area within the total length of the stream (*small* = <5, *medium* = 5-6, *large* = 5-6, *x-large* = 6-7). Further analysis indicated that additional pooling of stream sizes into **size aggregates** would also be necessary to increase sample sizes and the robustness of models. The two size aggregates were identified from preliminary and prior unpublished analysis; *small-medium* (SM) and *large-x-large* (LXL). **Substrate shade** has three categories (*light*, *medium*, *dark*) and is not expected to vary in an individual stream throughout the time series. **Canopy cover** (*low*, *medium*, *high*) was originally proposed as a covariate, but significant forest fires have altered canopy cover in some watersheds and historical data delineating those changes may not be available. For this reason, canopy cover was often excluded from the analysis. **Large woody debris** was also thought to be relatively stable through time but may not be; this covariate was also avoided unless highly uncorrelated with other variables (correlation coefficient < 0.40) within a set of data. Finally, **water clarity** (*clear*, *turbid/tannic*) was also originally proposed as a stable or commonly retrievable covariate that had good correlations with differences between expansion factors; however, review of historical data and the paired-datapoints dataset indicates that this covariate may not be as retrievable or interpretable as desired. The definition of water clarity from the unpublished reports indicates that the categories are based on another covariate, **water visibility**; which is a record of the observers' perception of how far they can see into the water. Within the paired-points dataset, the water clarity categories do not align with the data in the water visibility field; which leads to water clarity being redefined. Water visibility was also not always retrievable. Water clarity will normally be relatively stable annually within a system, but in some years there may be an environmental effect such as a mud slide or glacial melt that impacts the water clarity. Annual variations must be captured in comments or through exploring the water visibility annually within a system. The result is that the water clarity covariate was not treated as consistently retrievable but more likely retrievable than not. *Turbid*, *partially turbid*, and *tannic* were all originally recorded as separate levels, but due to their low sample sizes and the expectation that the direction of their effects relative to the *clear* water clarity, they were grouped together as a single level *turbid/tannic* (sometimes written simply as *turbid* when reporting results).

There are two covariates that are specialized and have only been collected in a subset of systems that were explored in this analysis. **Discharge** was one variable predicted to be negatively correlated with observer efficiency because increased discharge often correlates with deeper water, wider channels, or increased turbidity. Daily discharge estimates were sourced from the HYDAT database, a national archive of Canadian hydrometric data managed by Environment and Climate Change Canada with the R package tidyhydat (Albers 2017). Not all systems have a water station and not all years where water stations were present have corresponding data, which results in lower coverage in the database and lower retrievability. **Protraction indices**, metrics of population-level spawning duration, were previously explored for Early Stuart systems, where visual survey frequency is generally higher and more consistent than other areas. The protraction indices explored here used the total cumulative carcasses recovered (i.e. including those past the peak count date) to help identify if there was a protracted spawning period. The protraction indices are predicted to be positively correlated with

expansion factor values, with the hypothesis that the peak count did not capture all spawners because there was still a large proportion outside of the spawning area.

There are additional covariates in the paired-database, but they were either highly correlated with other covariates or their sample sizes were too few to have been used to create a covariate-exclusive set of data. They included but are not limited to: system stability class (*stable, volatile*), percent bank full category (*0-25, 25-50, 50-75, 75-100*), brightness of lighting (*dark, medium, bright, full*), percent cloud cover category (*0, 25, 50, 75, 100*), observer-perceived fish visibility category (*low, medium, high*), meters of water depth visibility category (*0.25-0.5, 1-3, 3-bottom*), subjective observer efficiency (average %).

2.2.1. High-accuracy and Visual Methods

High-accuracy methods are defined as those with a target coefficient of variation (CV) of less than 15 % and low to zero bias. There are occurrences in the database where a high-accuracy estimate's CV is > 15 %, but the estimates were deemed more valuable to be included in the analysis than detrimental to parameter estimates. There are three methods used to produce high-accuracy estimates in this database: fence counts, sonar, and mark-recapture.

Fence Counts & Complete Census

Fence counts are a high-accuracy complete census method where Sockeye are intercepted during their migration into terminal spawning grounds using enumeration weirs sometimes in combination with fish traps (Schubert and Houtman 2007). Typically, weirs span the entire width of the channel and have a removable gate that allows for passage. When Sockeye are present, the gate is removed and fish are counted either manually or by video surveillance as they move through the fence. The escapement estimate is the net upstream passage. Sockeye travelling downstream through the fence are subtracted from upstream Sockeye passage to prevent double counting Sockeye that may travel up and down through the fence over time. Spawned out sockeye, identified by wear from redd building or competition for mates, are not subtracted from the count as they are assumed to have successfully spawned and are not likely to return upstream. If fish are already present in the system when the fence is installed, the escapement estimate will include the fence count and an expanded live count of upstream areas conducted the day of installation. Similarly, if any spawning is observed below the fence, a live count of spawners is conducted at peak of spawn and added to the fence estimate. It is assumed that species identification errors are minimal when fish are counted manually or using video surveillance. Fences are susceptible to species identification errors and negative bias under turbid conditions due to poor visibility; however, they are typically operated in systems where Sockeye are either the only salmon present or the dominant species.

Sockeye are also enumerated by complete census methods in artificial spawning channels via enumeration fences and/or complete carcass removal. Carcasses are either physically removed from the channel or bisected to ensure that they are not double counted. Here, it is assumed that every fish that entered the channel to spawn was tallied during channel surveys after they perished (i.e. no adjustment for predator removal or carcass scavenging).

Hydroacoustic Estimates (SONAR)

Since 2006, the Fraser Sockeye Stock Assessment Program has increasingly adopted the use of hydroacoustic (SONAR) technology as a high-accuracy enumeration method, which offers an alternative to mark-recapture studies and enumeration weirs (Cronkite *et al.* 2006, Holmes *et al.* 2006, Lilja *et al.* 2008). Multi-beam SONAR technology produces -an underwater image of

individual fish as they pass into terminal spawning areas. The system can be applied under almost all water clarity conditions, such as low light or high turbidity. DIDSON or ARIS units (manufactured by Sound Metrics in Bellevue, WA) are installed at fixed locations downstream of most spawning grounds for the target population. A partial weir is often used to redirect fish passage, so they converge and migrate in front of the SONAR to ensure they are captured in the image. A single- or dual-bank setup is deployed depending on stream width; one sonar unit is operated from one bank and ensonifies the entire opening (up to 15 m using high-frequency imaging), or two sonar units are operated, one from each bank, to ensonify a larger stream width. One assumption is that very few Sockeye migrate in the middle of rivers due to their strong shore-oriented migrating behaviour to save energy (Hughes 2004). In dual-bank setups, units are deployed at the same river kilometre to mitigate double counting possible by Sockeye crossing to an opposite bank. All potential blind spots and gaps in the SONAR imaging are assessed during installation, therefore the assumption of this method is that all passage is captured by the SONAR image. As light is not a constraint, SONAR units are operated 24 hours a day over the duration of the project. SONAR files are sampled by manually counting and recording all upstream and downstream fish passage, where net upstream passage is used as the estimate of escapement.

For most systems enumerated by SONAR, escapement is typically estimated using only hydroacoustic data, though there are some exceptions. If any spawning is observed below the hydroacoustic site, a live count of spawners is conducted at peak of spawn and added to the hydroacoustic estimate (typically represents <1% of the estimate).

The benefits of using hydroacoustic methods are (i) lower operating costs due to reduced labour requirements relative to mark-recapture, (ii) greater precision compared to mark-recapture methods, and (iii) less susceptible to failure during high water events in comparison to enumeration fences (Lilja *et al.* 2008). A main limitation of using hydroacoustic technology is the initial capital cost of units and the availability of a suitable site. Hydroacoustic sites should ideally be straight, single channels with non-turbulent flow where there is active migration (i.e. minimal milling nor spawning) by the target population (Enzenhofer and Cronkite 2000). The site should also have moderate substrate size with no boulders and a planar streambed profile to reduce the likelihood of blind spots in the SONAR image. Additionally, hydroacoustic technology is not recommended for systems where similar-sized species overlap in migration timing, as the risk of species misidentification can be substantial without additional data such as that from a test fishery. More detailed descriptions of field methods, the stratified sampling design, and underlying statistics used to generate escapement estimates are provided by (Enzenhofer and Cronkite 2000, Cronkite *et al.* 2006, Holmes *et al.* 2006, Lilja *et al.* 2008, Xie and Martens 2014).

Mark-recapture Estimates

The mark-recapture technique is another high-accuracy method used to enumerate Sockeye populations with a high expected return. This enumeration method estimates population size by marking a random sample of fish, releasing them back into the stream to redistribute amongst the population and then implementing an appropriate recovery effort to obtain a second random sample of both marked and unmarked fish. The proportion of marked fish in the second sample is assumed to reflect the proportion of marked fish in the entire population.

Sockeye are typically captured using seine nets as they arrive and hold in terminal areas and are marked by applying Petersen Disk or spaghetti tag just beneath their dorsal fin. Carcass

recovery surveys cover historical spawning areas throughout the spawning and die-off period and have a main focus to assess the rate of marked and unmarked fish, while also recording tag status. Tag application and recovery data are used to generate Sockeye escapement estimates using the Petersen estimator, or more recently, stratified mark-recapture estimators (Darroch 1961, Schwarz and Taylor 1998, Schubert and Houtman 2007).

The accuracy of mark-recapture estimates is dependent on how effectively the following assumptions are met: (i) the population is closed (i.e. no immigration or emigration from the system); (ii) tag application has no effect on fish behaviour; (iii) tags are not lost; (iv) tags are always detected during recovery; and (v) all fish have equal tagging and recovery opportunities (Ricker 1975, Seber 1982). Fraser Sockeye mark-recapture studies follow a study design with specific procedures to meet these assumptions. Variations in spawning behaviour and arrival timing between sexes are common in Sockeye populations (Healey and Heard 1984, Quinn 1999). To avoid breaching the assumption of equal tagging and recovery opportunities, male, female, and jack samples are typically analyzed separately, and mark-recapture estimates are generated for each group separately. In addition, tags are applied as long as fresh fish continue entering the system. The tag application portion of the program targets major holding areas to ensure all fish have some probability of being tagged, therefore, tagging may be conducted at a single site or at multiple locations downstream, upstream, or throughout the spawning area. Additionally, a net mesh size that is non sex-selective based on body size is used to ensure males, females, and jacks have an equal opportunity for capture. Stress that may be induced by tag application is reduced by following holding time limits (45 or 60 minutes) and monitoring fish condition while they are being held in nets. A secondary mark, either an operculum punch or second external tag, is applied to a portion of or all fish to ensure marked fish are easily identified during recovery and to account for tag loss. Carcass recovery surveys are conducted regularly throughout historical spawning areas and will continue as long as carcasses remain available for recovery. Re-pitch surveys are conducted periodically, where carcasses that have previously been recovered and recorded as unmarked, are examined for failed recognition of markings. Prior to producing an escapement estimate, the tag application and carcass recovery data are reviewed to examine whether the assumptions have been met and to identify and minimize sources of bias.

There are various procedures used to assess mark-recapture data for bias associated with each of the five assumptions, these are outlined in greater detail by Schubert (2007) and discussed in section 2.4.4 Mark-Recapture Estimates. Bias test results are used to inform whether a stratified estimate or Petersen estimator should be used. Stratified Petersen Analysis Software (SPAS) is used to generate escapement estimates with the stratified datasets and the unstratified data (Arnason *et al.* 1996, Schwarz 2013).

Visual Estimate Methods

Visual assessment methods are currently the most widely used enumeration method for individual Sockeye streams throughout the Fraser River watershed and are typically considered a low-precision, low-accuracy method type. For visual assessments, observers count both live and dead salmon throughout a population's historical spawning area over the entire duration of the spawning and die-off period when possible. The objective is to capture the peak of spawn or at least the peak live count, when the highest proportion of that year's escapement is present on the spawning grounds. To achieve this, live counts are systematically conducted before, during, and after peak spawning to ensure full coverage of the spawning and die-off period. The peak of

spawn is identified retrospectively as the survey date with the highest proportion of spawning salmon observed during a live count. The peak live count however, may sometimes also include a high proportion of holding Sockeye if the actual peak spawn date is missed or spawning is protracted. This peak live count date serves as a reference point and is used in subsequent escapement analyses.

Visual methods include ground surveys conducted by foot, raft, or boat, and aerial surveys conducted by helicopter. Carcasses are also counted to account for the portion of fish that spawned and died prior to the peak live count date; they are included in counts to maximize the proportion of the escapement available for inclusion in the estimate. When carcasses are recovered and counted during ground surveys, they are marked externally (bisected) to prevent double counting in subsequent surveys. In most cases, escapements are estimated by expanding the peak live count and cumulative dead count (the count of all carcasses recovered up to and including the date of the peak live count) by the 1.8 or a stream-specific expansion factor (Cousens *et al.* 1982, Welch *et al.* 2017). For systems enumerated by aerial surveys only, carcasses cannot be marked during surveys to indicate they have been previously counted, therefore, the estimate is derived from the expanded peak live count and carcasses counted only during that survey. There are opportunistic ground or raft surveys that may also occur where only the carcasses counted during the single live count are available, but these are often low-abundance systems.

The number of visual surveys conducted varies by system and target population but are typically conducted once per week throughout the historical spawning period. Survey frequency may be greater for populations with low expected returns or high priority as greater effort translates to a greater likelihood of capturing peak spawner abundance. Survey frequency may be lower for streams that are remote and difficult to access, or that are larger in channel width and are therefore difficult to enumerate from the ground. When remote or large streams can only be surveyed effectively by helicopter, survey frequency is often limited to a total of 1-3 surveys to reduce flight costs. When only one aerial survey is able to be conducted, it is timed to coincide with the expected peak abundance, which is based on historical run timing for that system, or to align with observations of spawn timing in nearby populations within the same CU. There are several factors to consider when using visual methods to estimate escapement. It is assumed that survey timing and frequency is sufficient to capture peak of spawn and that the peak count is not only well-defined in the dataset but is also a sufficient representation of the peak spawner abundance.

Historically, the expansion factor of 1.8 was assumed to be appropriate for use in most Fraser Sockeye streams. Although the same expansion factor is applied to peak counts obtained from either ground or air, the two methods have differences in potential sources of error that often cannot be accounted for. Sources of error in both may include observer inexperience, which can lead to the underestimation of live counts; carcass removal due to predation or high-water events, which can reduce carcass counts; and environmental conditions that reduce fish visibility, such as large deep pools, fluctuating water levels, light conditions, water clarity, weather conditions, and whether observers use polarized glasses). Additionally the sources of error that exist for aerial counts include observers' position in the helicopter, fluctuating flight speed, stability and altitude from the ground, the helicopter's orientation in relation to the stream, and habitat type (e.g. fish holding in large pools are often more difficult to count from the air than from ground). Aerial counts also lack a count of cumulative dead since aerial surveys

are often limited in number and there is no way to assess if a carcass was counted during a previous survey. Ground counts are generally considered to be more accurate than aerial counts because observers can move more slowly and at a constant rate, and the view from the river bank is generally better due to a closer proximity to fish. The addition of a cumulative dead count can also improve the accuracy of ground surveys relative to aerial surveys. Exceptions include extremely wide systems where the entire channel cannot be easily seen from the ground.

2.3 UPDATED EXPANSION METHODS

Model development and selection criteria were established for assessing expansion factors of visual count data for Fraser Sockeye. Key considerations included sample sizes, covariate availability, and system similarities. To account for these differences, models were organized into a six-level specialization structure based on increasingly shared system characteristics and separated by survey method type (ground or aerial) within each level. Data were subset based on system similarities and some models were developed based on only these subsets of data while others used the whole database. Four expansion methods were explored, each with different strengths and weaknesses; mean-of-ratios, ratios-of-means, inverse regression, and calibration regression. The model selection process aimed to identify models that were most effective in reducing bias and improving expansion accuracy for a given subset of data. A leave-one-out (LOO) cross validation approach was used to calculate model performance statistics to select the most appropriate model. Models were then ranked by performance for a given subset of data, and a set of decision rules were applied to determine which model was the most suitable for a given system-year based on its similarities within a data subset (Figure 2).

The updated expansion factors were then tested using the 2024 preliminary data. During this testing, it was determined that additional decision rules were required for system-years whose visual counts fell outside of the observed calibration dataset range (the training data range). Model behaviour was explored outside of the dataset's range to help determine a standardized approach for these scenarios.

2.3.1 Approach

An important aspect in the analysis of this dataset is the sample sizes associated with individual systems, specific covariates, and overlap or similarities between systems and covariates. Generally, the robustness of parameter estimates increases with increasing sample sizes. Similarly, the generalizability of models also increases with larger sample sizes and with more diverse and representative samples. For example, models that use a single system's paired-data set would likely only be applicable to that system, or systems that are highly similar. Models were categorized on a scale of more specialized to more general for the purposes of model development. Specialized models often have smaller sample sizes but use more system-specific covariates or data transformations. Generalized models use larger subsets (larger sample sizes) of the dataset and include more widely available covariates. In this report, the level of specialization is referred to with a general descriptor and numerical value. For example, a model that was developed with a set of datapoints where stream size was aggregated and the data was subset to a level such as *Large-X-Large*, its **set-level** (denoted as **s** in subsequent equations) would be reported as **Size-Aggregate (2)**, where the 2 indicates the level of specialization. Here, the most general level starts at 1, with increasing values indicating increasing levels of specialization. The spectrum of generalized to specialized models and

varying sample sizes is an important facet of candidate model development, ranking, and the model selection process.

Set-levels were defined based on previous unpublished report results, covariate availability, and the factors that were predicted to have the greatest influence on differences in expansion factors. The difference between aerial and ground survey methods is predicted to be one of the most influential factors affecting expansion factors. Every set-level is divided between aerial and ground survey methods:

- **Generalized (1)** models are fit with all data points in the database and generally with retrievable covariates. Size was predicted to be another key factor in determining differences in expansion factors but sample sizes at the larger levels were low; therefore, aggregates were created based on clustering of prior results.
- **Size-Aggregate (2)** models are fit with data that are subset to either the *Large-X-Large* size aggregate or *Small-Medium* size aggregate.
- **Size (3)** models are fit using data that is subset to each size category.
- **Size-Clarity (4)** models are fit with data that are subset further based on the water clarity factor. The water clarity factor further reduces sample sizes within many subsets of data of this set-level and not all possible size-clarity combinations are expected to have data to support them.
- **Individual – Simple (5)** models are fit with only data from an individual system.
- **Individual – Specialized (6)** are models developed using data from multiple systems with specialized covariates but due to the subsequent model selection framework, their results are compared at the individual system level. Only some system-years may have the specialized covariate within an individual system; thus samples sizes within the Individual – Specialized (6) subsets may be smaller than the prior set-level.

It is expected that the sample sizes of each data subset will either be less than or equal to the previous set-level as the set-level specialization increases. Reductions in sample size will be from fewer systems having shared characteristics or fewer system-years with the necessary specialized covariates (Figure 1).

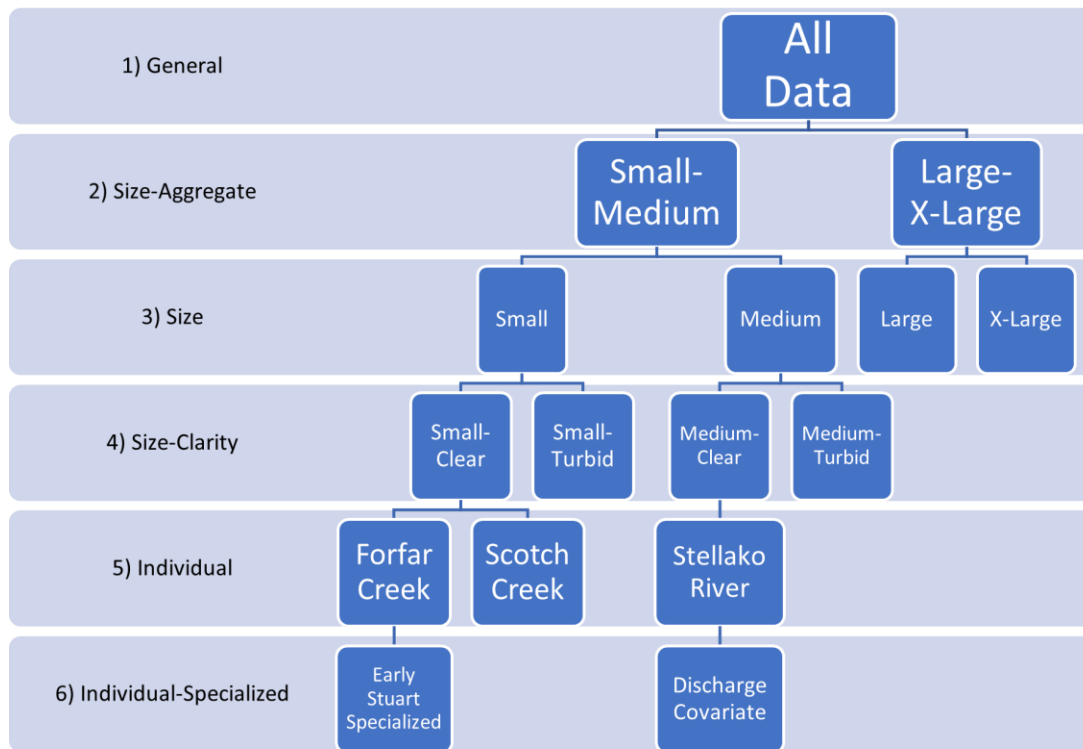


Figure 1. Example of the Set-Level framework. Sample sizes (n) generally decrease with increased set-level specialization. Note that not all branches are shown and that there are matching set structures for both Ground and Aerial survey methods.

Within the spectrum of model covariate or data set specialization, various model types were explored for their different features and assessed for inclusion in the candidate model set. Parsons and Skalski's (Table 3; 2010) review was used to identify peak-count expansion methods. Most methods either had a poor ranking for estimating variance or had unreasonable assumptions with regards to data requirements (Parsons and Skalski 2010). The methods investigated include: **means-of-ratios**, **ratios-of-means**, **inverse prediction**, and **calibration regression** (See Appendix B for more detailed methods and results). The means-of-ratios estimator and variance estimate were both ranked as poor. The historical 1.8 expansion factor method used the same means-of-ratios approach, therefore could be compared directly with updated expansion ratios and set-levels. The ratio-of-means variance estimate was also ranked as poor, but the estimator itself is less susceptible to bias and outliers than the means-of-ratios estimator (Cochran 1977, Falcy *et al.* 2016). Inverse prediction addresses a break in assumptions by standard calibration regression, but its application has few examples to build from, and it is less versatile. The calibration regression breaks the assumption that its independent variables are known without error (Parsons and Skalski 2010) because the visual count error is inherently unknown. Calibration regression was the most versatile method and therefore both generalized and specialized models were developed in its framework. Each of these models have different formulations, assumptions, and diagnostics that were used to assess their potential as candidate models for model selection (

Appendix B. Model Specifications, Diagnostics, and Parameter Results). Some models explored were excluded in the final candidate model set and the model selection process due to poor fits or low sample sizes used in their fits. Regardless of a model's selection success, its exploration allowed for comparisons in expansion factors and identification of where more data may be needed to support its use or development.

The primary objective of the model selection process is to reduce the average bias and improve accuracy in expanded estimates by identifying the best model for a system-year. A common approach to assessing model predictive ability is to use cross validation and calculate specific performance metrics that can be used to measure a model's performance relative to its specific goal (Armstrong 1985, Yates *et al.* 2023). The varying sample sizes used to fit models within different set-levels created a unique situation for comparing model performance between models that may use different subsets of data in their development, and particularly with low sample sizes ($n < 200$). The LOO method, sometimes referred to as jackknifing or n-way cross validation, is recommended over k-fold (e.g. 10-fold) or other cross validation methods when sample sizes are low (Armstrong 1985, Folkes *et al.* 2018). The LOO approach also produces performance statistics for individual system-years within each model; therefore, a system-year will have performance statistics that can be compared across every model it was used in. The model selection process is designed to subset these performance results so that comparisons are always made using the same system-year groups and therefore the same sample size of performance results.

The model comparisons went through a step-wise process, starting at the set-level of **Individual – Specialized (6)** and sequentially progressing to the most **Generalized (1)** level. Performance metrics were selected (3) to align with the goals of reducing bias and improving accuracy. Performance metrics' results were calculated for every system-year for each model it was used in. The performance results from models at every set-level are then subset to the system-years that are shared at the most specialized set-level in the comparison, and an average model performance is calculated for each model. The subset of performance results and their comparison are defined as a **rank-set**, where rank refers to an individual model's average performance relative to other models in a specific set-level (i.e. includes model results from all other set-levels but subset to only the data points from the most specialized level being compared). Only models that have results for every shared system-year in the subset of data at the set-level being compared are included. At the most specialized level, all more generalized model results would be included because there should be a complete overlap in system-years used in every other model. As the model comparisons are made at more generalized set-levels, sample sizes of system-years will generally increase but the number of models being compared will decrease as specialized models are excluded for not being fit with the same data points (Figure 1). During the final comparison step, to determine if a model should be used for system-years with similar characteristics and available covariates, an algorithm asks three questions.

The three questions in the algorithm (Figure 2) are designed to determine if a model is robust and expected to create the desired results for any data that it may be applied to consistently. The questions are: i) "is the top ranked model from the most specialized level of those being compared;" ii) "is the sample size greater than or equal to a threshold;" and iii) "is the mean CV greater than or equal to a threshold." The first question is designed to identify the generalizability of the model and its outcome determines the threshold in question two. If a top model is specialized, then the assumption is that it is representative of the conditions experienced by the systems in that set-level, suggesting that diverse data are not needed to

capture the behaviour of expansion factors. Conversely, a specialized model should only be applied to systems that share the same characteristics as those used to fit the model. If the top model was designed at a more generalized level, then it may not be specific to the characteristics of the streams in the given set-level but could still capture the general behaviour well if the sample size was large enough. In the second question, the threshold is set to 10 if the top model in a rank-set is from the most specialized level, and it is set to 20 if it is from a more generalized level. The sample size of 20 is a minimum sample size rule-of-thumb for detecting differences in categorical variables (e.g. in a chi-square test) (Wilson VanVoorhis and Morgan 2007). The reduction to 10 samples for more specialized levels was primarily done due to the low sample sizes in the dataset and assumes that there is lower variability in expansion factors, requiring less power to detect difference, and results in reduced generalizability. These rules-of-thumb and assumptions could benefit from simulation testing in future work. The CV threshold for question three is set at 50% with the understanding that these estimates will still be treated as low-moderate precision but that now an estimate of precision may be available for them. A lower precision model may be flagged to identify where additional data points may be needed to improve the parameter estimation or to indicate that an alternative model may be required. If the top model in a rank-set passes the thresholds in questions two and three, then it is applied to system-years that have the same characteristics and the same covariates available. The selection process moves to the next set-level if either threshold is not met or covariates are unavailable, indicating that more data are required to support model development at that level of specialization.

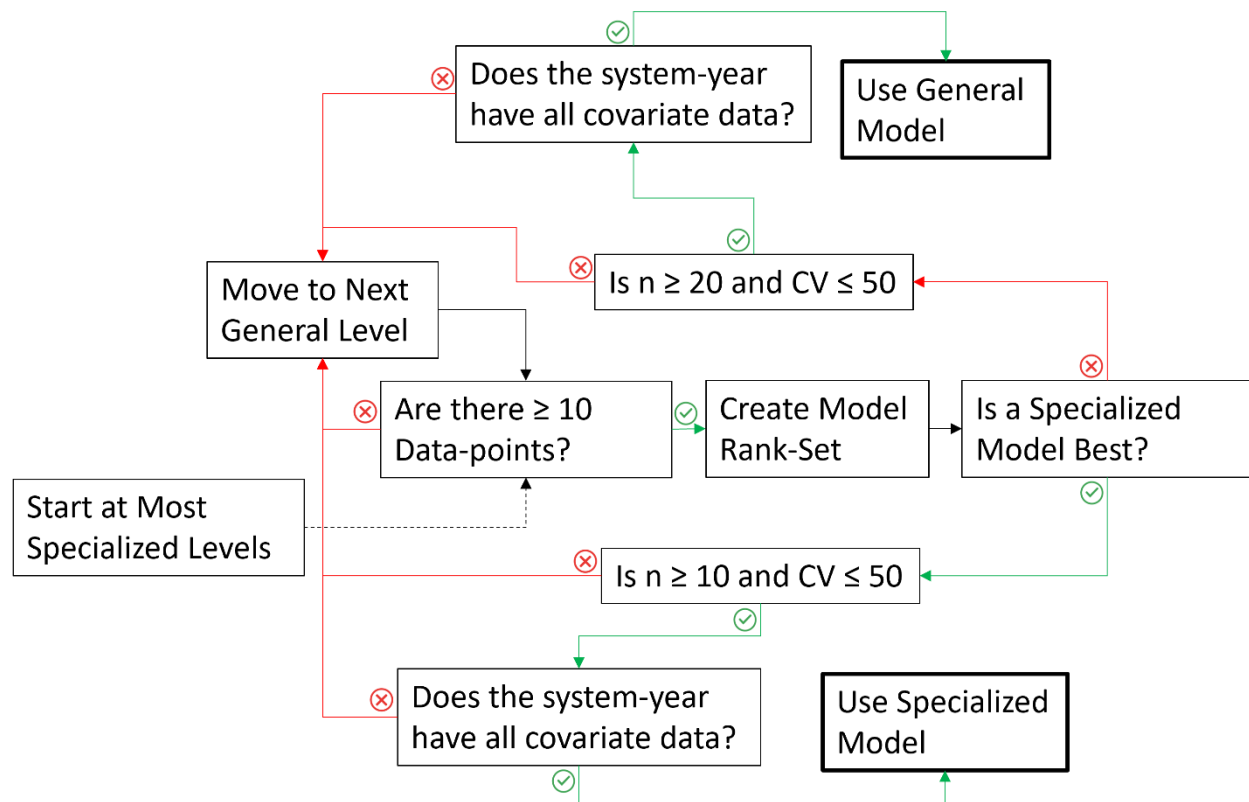


Figure 2. Algorithm pathway and decision points used for model selection based on system characteristics and covariate availability.

2.3.2 Model Validation and Selection

Model performance of candidate models was evaluated using metrics calculated by leave-one-out (LOO) cross validation (Folkes *et al.* 2018, Yates *et al.* 2023). For every unique system-year data point within each model and set-level, the model was refit without the data point and then used to predict an expanded estimate and its variance for that data point:

(1)

$$\widehat{EN}_{msi} = \text{predict}(\text{model}_{s-i}|C_i)$$

$$\sigma^2_{msi} = \text{variance}(\text{model}_{s-i}|C_i)$$

Where $\widehat{EN}_{msi}$ is the expanded estimate given a visual count (**C**) from a system-year (**i**) from a model (**m**) from a set-level (**s**) that was refit without that system-year (**-i**). If a model operates in a log-transformed space, the expanded estimate is transformed back into real space but if the variance (σ^2_{msi}) is calculated in log-transformed space, it remains in the log-transformed space for estimating the coefficient of variation in the next steps. Performance results were calculated for each leave-one-out prediction and subsequently for sets of predictions based on the shared rank-set system-years. This is like the one-step-ahead retrospective analyses used in the Fraser River Sockeye forecast analyses (e.g. Folkes *et al.* 2018, Dionne and Liu 2024) with the exception that there is no temporal component to the data considered. The error (ε_{si}), the percent error (δ_{si}), the absolute percent error ($|\delta|_{si}$), the squared error (ε^2_{si}), and a coefficient of variation (CV) were all calculated for each system-year (**i**) within each model (**m**) within a set-level (**s**):

(2)

$$\varepsilon_{msi} = \widehat{EN}_{msi} - \hat{N}_i$$

$$\delta_{msi} = \frac{(\widehat{EN}_{msi} - \hat{N}_i)}{\hat{N}_i}$$

$$|\delta|_{msi} = \left| \frac{(\widehat{EN}_{msi} - \hat{N}_i)}{\hat{N}_i} \right|$$

$$\varepsilon^2_{msi} = (\widehat{EN}_{msi} - \hat{N}_i)^2$$

$$CV_{msi} \begin{cases} \sqrt{e^{\sigma^2_{msi}} - 1}, & \text{if } m \text{ is in log transformed space} \\ \frac{\sqrt{\sigma^2_{msi}}}{\widehat{EN}_{msi}}, & \text{if } m \text{ is in real space} \end{cases}$$

A negative error and negative percent error indicates that the expanded estimate was biased low relative to the high-accuracy estimate for that system-year ($\hat{N}_i$) and vice-versa. For a subset of LOO model predictions determined by the rank-set, the absolute mean percent error (|MPE|), the mean absolute percent error (MAPE), the root of the mean squared error (RMSE), the Pearson correlation coefficient between the predicted and known values (R), and the mean CV were calculated:

$$\begin{aligned}
|MPE|_{ms} &= \left| \frac{\sum_i^k \delta_{msi}}{k} \right| \\
MAPE_{ms} &= \frac{\sum_i^k |\delta|_{msi}}{k} \\
RMSE_{ms} &= \sqrt{\frac{\sum_i^k \varepsilon_{msi}^2}{k}} \\
R_{ms} &= \frac{1}{k-1} \sum_i^k \left(\frac{\widehat{EN}_{msi} - \widehat{EN}_{ms}}{s_{EN_{ms}}} \right) \left(\frac{\widehat{N}_{msi} - \widehat{N}_{ms}}{s_{N_{ms}}} \right) \\
CV_{ms} &= \frac{\sum_i^k CV_{msi}}{k}
\end{aligned}$$

Where s is the standard deviations of either the expanded estimates ($\widehat{EN}_{ms}$) or the high-accuracy estimates ($\widehat{N}_{ms}$). The sample size (k) varies for a given model based on which set level it is being compared within. These metrics were selected based on the primary objectives of reducing the average bias in expanded estimates and improving their overall accuracy (Hyndman and Koehler 2006). The absolute MPE metric is an indication of the average magnitude of the model's bias and thus used in the model selection steps where the objective is to reduce the bias regardless of direction. The raw MPE is used later as a general summary statistic to include the average directional bias of data points but not used in the model selection steps. The MAPE and RMSE metrics are both indicators of a model's accuracy, with the RMSE being more sensitive to detecting outliers (i.e. highly inaccurate predictions in this case) due to the errors being squared before they are summed and averaged (Armstrong and Fildes 1995, Armstrong 2001). The Pearson correlation coefficient (R^2) metric measures the linear correlation between the predicted and known estimates, indicating how much variation in the known estimates is described by the predicted estimates. R^2 is assumed to be positive for all model cases as the models were developed to correlate the predictions with known values, however, the degree of correlation may vary. Mean absolute scaled error (Hyndman and Koehler 2006), scaled to the results of a 1.8 expansion, was considered as well as excluding R^2 from the set of performance metrics but the combination of the four selected performance metrics created the most consistent results (Appendix A. Data, Code, and Summary Tables). The average CV is an indication of the robustness of the model results and is used as a decision point in the model selected algorithm after average ranks are created for each model within a rank-set.

Lower values in each metric represent a better performance result, with the exception of the R^2 ; however, the scale of the results may differ based on the system-years included in each rank-set. Performance metrics were standardized within a rank-set and averaged across all system-years for each model in a rank-set to make the results comparable across all rank-sets (Folkes *et al.* 2018):

$$Standardized\ Metric_{jms} = \begin{cases} \frac{Metric_{jms} - \min(Metric_{js})}{\max(Metric_{js}) - \min(Metric_{js})}, & \text{except for } R_{ms} \\ \frac{(Metric_{jms}) - \max(Metric_{js})}{\min(Metric_{js}) - \max(Metric_{js})}, & \text{only for } R_{ms} \end{cases}$$

$$Model_{ms}\ Rank = \frac{\sum_j^l Standardized\ Metric_{jms}}{l}$$

Each metric (j out of l) was standardized between 0 and 1 within a rank-set (s) so that 0 represented the model (m) that scored the best in that metric and 1 was the worst. The average of all performance metrics was then taken for each model within a rank-set. Therefore, the result is a standardized model rank that describes how well a model did within a specific rank-set but on a scale that is comparable across rank-sets. The model rank closest to 0 was identified as the top model within a rank-set and entered the model selection algorithm.

The decisions of the model selection algorithm determine which models are to be applied to existing, historical, and future visual counts to estimate their expanded estimates and associated standard errors. System-years follow the format of checking if a top model was selected by the algorithm at the most specialized set level that they belong to first. If no top model was identified or the system-year does not have data for a covariate, then the next more general set-level model results are checked for a successful top model. This process is repeated until a set-level subset is identified to have a top model or the top Generalized (1) model rank-set results are used.

With the set-level subset identified, the selected expansion factors are applied to the data in the database for comparison between the 1.8 expansions and the updated expanded estimates. The expansion estimates in the comparison do not use the cross validation results; instead, the estimates are the model fit results. The same performance metrics from Equations 2 and 3 are calculated for comparison with the expansions made with the 1.8 expansion factor. Three abundance bins are presented to represent the historical and recent breakpoints in abundance categories at 25,000 and 75,000. Results were further compared within each abundance bin between clear and turbid/tannic water clarities because water clarity is a common factor used to apply corrections in the RSA process.

General linear models (i.e. linear regressions, ANOVAs, and T-tests) were used to test if there was a significant difference between the results of the performance metrics between the updated methods and the historical 1.8 expansion. Tests were done to detect differences within the three abundance bins (1-24,999, 25,000-74,999, 75,000+) and two water clarities (clear, turbid/tannic). T-tests were used to determine if MPE was not significantly different than 0. Lastly, temporal trends in MPE and MAPE were tested using linear regression and likelihood ratio tests were used to determine if the slope parameter was significant. The significance level was set to $\alpha = 0.05$ but adjusted with a Bonferroni correction (Bonferroni 1936, Jean Dunn 1961) to α of α/m , where m is the number of tests applied. Correcting α for multiple singular comparisons reduces the probability of committing Type I errors (i.e. rejecting H_0 when H_0 is

true) (Cabin and Mitchell 2000). The number and characteristics of expanded estimates with greater than 50% absolute percent error are also flagged and discussed. All statistical analysis were completed using the R software (R Core Team 2024) and documented in supplemental materials (Appendix A. Data, Code, and Summary Tables).

2.3.3 Application to 2024 Data and Additional Decision Rules based on Expert Advice

The models selected through the model validation and selection process were then applied to the 2024 data and compared against both the default estimates and the run-size-adjusted estimates. This application highlighted a need to apply additional decision rules for the algorithm and further model investigation. The original model selection was done using only the data from within the database, which are not representative of all of the combinations of system-characteristics nor the range of visual counts experienced by observers. The updated expansion algorithm was first applied to the 2024 data without regard for whether the counts were within the observed range of the calibration database. The percent change in the expanded estimates relative to the default and RSA estimates was used to highlight where possible issues may reside in the framework.

Two unique data subsets were developed to explore model behaviour for scenarios where visual counts fall either below or above the range of visual counts in the database. The behaviour of low abundance data was explored using a subset of datapoints where the visual counts were less than 25,000 and only included data from *small* and *medium* sized systems because they are the source of most low abundance counts. The behaviour of high abundance data was explored using a subset of datapoints with visual counts greater than 75,000. Rank-sets were made for both aerial and ground surveys like in the regular selection process as well as combining the methods since the sample sizes were low for some of the subsets. Model fits were explored by plotting the data relative to their different parameter combinations. Final model specifications were made using expert opinion considering the objective to reduce post-season run-size adjustments. This consideration equates to focusing on models, particularly for low abundance, that incorporate standardized parameters for both method and water clarity (two of the most common reasons for run-size adjustments).

2.4 OTHER EXPANSIONS, INFILLS, AND DATA CONSIDERATIONS

Following is a general summary of scenarios and procedures for expansions or extrapolations, infills or interpolations, or other data considerations.

2.4.1 Biological Interpolations and Infills

Spawning success and sex-specific estimates are derived from carcass recovery samples collected on the spawning grounds and sometimes require borrowing information from adjacent systems. Spawning success is assessed by examining female carcasses for the presence of retained eggs. Sex-specific estimates are based on the observed ratio of male-female-jack carcasses recovered and are assumed to represent the sex ratio of the entire spawning population. Depending on the size of the population, a minimum of 20-50 carcass samples are desired for an adequate sample size. If the carcass recovery sample is insufficient, or does not have suitable spatial and temporal representation, Fraser Sockeye Stock Assessment follows a set of CU-specific rules for approximating spawning success and sex ratios. These rules are outlined in "Fraser Sockeye Stock Assessment Summary Rules for Approximating Sex Ratio

and Spawning Success for Spawning Sites and Populations (CUs) with Insufficient Carcass Recovery Samples” found in Appendix D.

For each escapement assessment, a subsample of the carcasses recovered have scale and otolith samples collected for age determination by the Pacific Salmon Commission (PSC) ageing laboratory. The sockeye associated with these samples are initially assigned adult or jack status according to their fork length (≥ 45 cm adult, ≤ 44.9 cm jack) while being sampled in the field. These scale and otolith samples are later analyzed in the laboratory to confirm age. The lab-confirmed age results are then used to correct any misclassifications in the carcass recovery samples. The adjustments to sex ratio and age classifications are then applied to the expanded escapement estimate.

2.4.2 Fence Counts

Spawning success and sex-specific estimates from systems enumerated using fence counts are generally obtained from the sex ratio observed in carcass recoveries based on the guidelines presented in 2.4.1 Biological Infills. In some systems, the sex of fish is recorded at the time of passage through the fence. When this occurs, corrections to the sex identified at the time of passage are made using carcass recovery data or other relevant data. For example, broodstock collected at the fence are assigned a sex at capture. When these broodstock are later spawned and sacrificed, the sex is confirmed. This information is then used to correct the sex ratio determined at the fence.

One potential risk to fence operations is a breach due to high water levels or equipment malfunction. When this occurs, the fence remains non-operational for a period of time, resulting in a gap in data. To address these data gaps, interpolating can be used to account for short periods when counts could not be completed, however this method is not used when significant portions of the migration are missed. When the breach period is 24 hours or less, the missing data are interpolated using the average of the net upstream counts for the same hour on the day before and the day after each missing hour. If the period of missing data exceeds 24 hours, a curve is fitted to the migration data from several days prior to and following the missing period. The curve is then used to estimate the missing migration data.

2.4.3 Hydroacoustic Estimates

Hydroacoustic estimates of net upstream migration are obtained by expanding net upstream counts from hydroacoustic file recordings to cover the full migration period. Each hour of migration is typically estimated from a 20-minute sub-sample. The net upstream count is equal to the upstream count minus the downstream count in each 20-minute file. The net upstream count for each 20-minute period is then multiplied by 3 to expand the count for the hour. If fewer than 20 minutes were recorded, the net upstream count is divided by the number of recorded minutes and then multiplied by 20, before being expanded for the hour. Lilja et al. (2008) determined that the 20-minute sampling strategy is sufficient to estimate escapement with 95% confidence intervals of $\pm 10\%$ or better.

Hydroacoustic files are periodically missing due to equipment failure, misalignment, or other reasons. When the period of missing files is 24 hours or less, the missing data are interpolated using the average of the net upstream counts for the same hour on the day before and the day after each missing file(s). If the period of missing data exceeds 24 hours, a curve is fitted to the migration data from several days prior to and following the missing period. The curve is then used to estimate the missing migration data. In cases where there is a SONAR operating on

both banks and only one of the banks is missing data, the passage from the bank with the working SONAR is multiplied by the other bank's average migration proportion to interpolate missing data.

Effort is made to ensure that hydroacoustic operations capture the full period of migration. There are situations, however, where the start or end of migration may not be captured. In such cases, a curve is fit to the migration to determine the rate of growth or decay, and this curve is used to extrapolate for the missing data. This method is rarely used for the beginning of the migration because water levels normally facilitate weir and sonar installation before Sockeye arrive in their terminal spawning areas. When the method is used for tail-end extrapolation, it is often for less than 1% of the total abundance.

Lake systems and rivers with multiple tributaries assessed using hydroacoustic technology generate a system-wide estimate of all Sockeye spawning in areas upstream of the hydroacoustic installation. The tributaries and lakeshore spawning areas are visually surveyed throughout the spawning period to generate spatial estimates of spawner abundance. The relative proportions of these visual estimates, calculated against the total SONAR count for the system, are then used to estimate the individual spawning populations within the system-wide estimate. Similarly to fence counts, if there is a portion of spawning that occurs downstream of hydroacoustic installations, that area is visually estimated and added to the hydroacoustic estimate to provide a complete system estimate.

2.4.4 Mark-Recapture Estimates

Adjustments are made to mark-recapture data to help meet the five main assumptions of a Petersen mark-recapture study.

The population closure assumption is addressed by examining recovery data and adjusting for tag emigration. When fish that have been tagged leave the mark-recapture study area, an emigration rate estimate is applied to account for this movement. Emigrant tags, which are those that have left the study area, are removed from the mark-recapture data based on the type of emigration (referred to internally as Type 1 or Type 5). Type 1 Emigrants: These tags are recovered in areas outside the regular recovery cycle and are not consistently recovered over time. The ratio of Type 1 emigrant tags recovered to the number of untagged carcasses in each emigrant area is used to estimate the number of unrecovered tags that emigrated to that area. This total is then subtracted from the total number of tags used in the study. Type 5 Emigrants: These tags are recovered in spawning streams that not considered as part of the mark-recapture system but are regularly surveyed throughout the arrival and spawning periods. Individuals from these streams may occasionally migrate through the target population but are not considered representative of the target population. The ratio of Type 5 emigrant tags recovered to the number of untagged carcasses in each emigrant area is used to estimate the number of unrecovered tags that emigrated to that area. The timing of Type 5 emigration is known from proportional recovery data over time, and the estimated total number of Type 5 emigrant tags is proportionately removed through time from the sample of tags applied in the study.

The tag loss assumption is addressed by recording the number of fish that were initially double-marked (primary and secondary mark), but that were recovered with only one mark present. If tags are lost after application, an estimated tag loss rate is applied to account for these losses. A subsample of all tagged Sockeye also receive a secondary mark. By comparing the

secondary mark status of recovered tags to their status at application, we can estimate the rates of primary, secondary, and double tag loss. This tag loss rate is then applied to the recovery sample to estimate the number of additional tags that would have been recovered if no tag loss had occurred. These additional tags are added to the total number of tags recovered prior to calculating the mark-recapture estimate.

Assessing the effect of tag application on fish behaviour is done using chi-square tests that examine whether the recovery probability of tagged fish is affected by: (i) how long a fish was held in the seine net prior to being tagged, (ii) how many times a tagged fish was recaptured during tag application, and (iii) how vigorously a fish swam away after being tagged and released. The results of these tests are used to determine whether individual tagged fish should be removed from the tag application and recovery samples.

To assess missed tags (i.e. misidentification of recovery status), a re-pitch survey is conducted on a subsample of recovered carcasses throughout the recovery period. Re-pitch follows the same procedures as a carcass dead-pitch survey, with the exception that chopped carcasses (i.e. those that had already been sampled) are examined. The number of tags found during re-pitch surveys is divided by the total number of carcasses examined during re-pitch. This rate of missed tags is then applied to the untagged recovery sample to estimate the total number of tags misidentified as untagged during recovery. These estimated tags are then added to the number of recovered tags before calculating the final estimate.

Sex classification at the time of tagging is also corrected based on sex confirmed during recovery. Sex is more accurately determined during carcass recovery because internal sex organs can be directly observed. The number of tagged carcasses recovered with a recovery sex different than its sex assigned at application is used to calculate a sex misidentification rate according to the formula in Appendix D. Other Expansions, Infills, And Data Considerations. This rate is then used to estimate a correction to be applied to the application sample.

Equal tagging and recovery probability is assessed by stratifying the application and recovery data by sex, both temporally and spatially, and using chi-square tests to determine whether there are significant differences in tag incidence and recovery rates. Application bias refers to the non-proportional application of tags and insufficient mixing across tagging areas. It is assessed by stratifying the recovery sample and comparing mark incidence (the proportion of tagged carcasses) between strata. Similarly, recovery bias refers to the non-proportional recovery of carcasses and insufficient mixing across recovery areas; it is assessed by stratifying the application sample and comparing recovery rates (the proportion of tagged fish that were recovered) between strata. Bias test results are used to inform whether a stratified estimate or Petersen estimator should be used. Stratified Petersen Analysis Software (SPAS) is used to generate escapement estimates with the stratified datasets and the unstratified data (Arnason *et al.* 1996, Schwarz 2013). Akaike information criterion (corrected for small sample sizes) values are used to select the best model, either one of the various stratified estimates or the simple Petersen using unstratified data.

2.4.5 Lake Spawning Estimates

In lake systems where spawning occurs in areas with high turbidity or at depths where visual live counts are ineffective, escapement is estimated using a catch-per-unit-effort approach based off carcass recovery. This method assumes that each completed survey will recover 5% of the total spawning population. To estimate total escapement in these systems, the total

number of carcasses recovered is divided by the total number of surveys completed, and the result is then divided by 5% (Andrew and Webb 1987).

2.4.6 North Thompson & Other Outliers

North Thompson River spawners (part of the Kamloops-S CU) have been difficult to enumerate through time due to the river's x-large size and high turbidity. The only available method is helicopter counts (higher precision methods are not currently viable for this system); however, these estimates are assumed to be significantly biased low. When counting conditions are sub-optimal, a visual-based estimate may only be a loose indication of presence/absence and requires a substantial run size adjustment to the spawning escapement (Robinson et al. in press). Various infilling methods or terminal run-size estimate methods have been applied to the North Thompson through time in years where counting conditions are poor (i.e. <30% assumed observer efficiency), but most recently, a ratio approach has been implemented.

Within the Fraser and Interior Area, it is common practice across species to estimate whole-system abundance by infilling based on the average relative proportion of each population to its CU (English *et al.* 2007). This method assumes that populations within a CU share similar life history characteristics, environmental conditions, recruitment, and fisheries pressures, and therefore spawn in a relatively stable distribution between the systems within the CU. The system requiring infilling is estimated by expanding known system's abundance to become representative of the total CU abundance, and then allocating abundance to the unknown systems based on their historical average proportion they contribute to the CU. In the case of the Kamloops-S CU, there is only one other primary spawning area: Raft River. The Raft River however, is medium sized and generally clear, making it better suited to visual assessments than the North Thompson. Since there are only two systems in the CU, the current approach takes the ratio-of-means of the two system escapements over the most recent 20 years and applies it to the current Raft River estimate to infill the North Thompson estimate for that year (used in 2023 and 2024).

A major caveat of the recent North Thompson ratio-of-means method is that the historical estimates use a mixture of methods and that there is a lack of high-quality estimates of the North Thompson. It is recommended to complete an in depth review of historical North Thompson estimates and methodologies and update the time series with a consistent approach.

The 2023 Nahatlatch-ES estimate was compromised due to a wildfire that blocked access to the spawning ground assessment areas. While the river was visually surveyed via helicopter under optimal counting conditions, lake spawners were unable to be enumerated due to the system's glacial water (the lake expansion method is only appropriate when surveyed by boat over multiple surveys). Similar to the approach listed above for the North Thompson River, a ratio-of-means of the Nahatlatch River and Nahatlatch Lake estimates over the previous 20 years was used for infilling.

Other populations that do not have an appropriate proxy spawning area are not able to be estimated using this infilling method and are therefore assigned an estimate of NA (e.g., Taseko 2019).

2.4.7 Prior Specialized Peak Live Expansions

The upper Bowron River (primary spawning site for the Bowron-ES CU) and Gluske, Forfar, and O'Ne-Ell creeks (major contributors to the Takla-Trembleur-ES CU) have specific expansion factors that are used annually to expand the visual-based estimates.

The upper Bowron River expansion factor (2.8) was derived from a single fence count paired to a helicopter survey in 1995 and is intended to reflect the unique migration behaviour of Bowron Sockeye. Peak spawning data shows that Bowron Sockeye often exhibit a protracted (>14 day) spawning period that is likely an adaptation to the slow, meandering river with little-to-no cover from predators. Two or three helicopter surveys are conducted each year and often yield similar live counts across all surveys, which suggests fresh fish are entering the system over time and there is no well-defined 'peak count'.

The Gluske, Forfar, and O'Ne-Ell Creek expansion factors (1.59, 1.69, 1.71, respectively) were derived from ground surveys on a four-day cycle paired with fence counts between 1990-2009 (unpublished, analogous results as in Table 6). They used the mean-of-ratios method. These represent some of the most data-rich systems for all Fraser Sockeye; as such, the stream-specific expansions are used for the three creeks listed above and the average of these calibration points (1.65) has been applied to all other Early Stuart creeks that are surveyed on a four-day cycle. The remaining Early Stuart spawning sites that are either enumerated on a seven-day cycle or opportunistically are given the standard expansion factor (1.8) for Fraser Sockeye populations in *small* sized, clear streams. The individual and aggregate mean-of-ratios expansion factors were updated in this exercise.

3. RESULTS

3.1 UPDATED EXPANSION APPROACH

Initial model development created 45 candidate expansion factors/models, which were compared across 36 rank-sets. A total of 18 rank-sets passed the model selection algorithm's robustness thresholds. Nine models were selected for application to the calibration data set for comparison with the default 1.8 expansion, as well as for future and historical application.

The updated model application resulted in generally improved performance metrics. The magnitude of bias (MPE) was nearly always improved, regardless of how the data were subset and investigated. Accuracy (MAPE and RMSE) was somewhat improved but not in all organizational levels of the data. Some outliers were shared between the default approach and the updated expansion, indicating that no improvements were made in explaining their variability. *X-Large aerially* surveyed systems produced the greatest number of outliers relative to their occurrence in the database. Estimate precision was overestimated, resulting in paired high accuracy estimates not being captured by the error intervals of updated expansion estimates.

3.1.1. Candidate Models and Model Ranks

A total of 45 expansion factors or expansion models made it through their respective model diagnostics and inclusion criteria (Appendix B. Model Specifications, Diagnostics, and Parameter Results; Appendix C. Rank-Set Dashboards and Test Results Table 24). The ratio-of-means expansion factor development was limited to the Size-Clarity (4) set-level and

updating the Early Stuart Individual (5) subsets because prior review ranked the method as poor and easily biased (Parsons and Skalski 2010); therefore, they were developed only using subsets of data that were expected to have the most similarities. Three Size-Clarity (4) means-of-ratios models passed their inclusion criteria of producing average CVs less than 50% and having sample sizes of at least 20 in addition to the three Early Stuart Individual (5) models. Ratio-of-Means expansion factors were developed at the Generalized (1) through Size-Clarity (4) set-levels. The Ground surveyed *large-x-large* set at the Size-Aggregate (2) level and any more specialized set of those two size categories had fewer than 20 samples and were excluded. Within the Size-Clarity (4) level, no sets with a *turbid/tannic* clarity had greater than 20 samples and were excluded as well. A total of 13 Ratio-of-Means expansion factors were added to the candidate model set. Inverse regression models were developed at the Size-Aggregate (2) and Individual (5) levels, either separated by survey method or by combining the data from both methods. Only the *ground* surveyed *large-x-large* size subset had fewer than 10 data points and was excluded from the Size-Aggregate (2) results, leaving a total of five inverse regression models at that level. Seven Individual (5) inverse regression models were added, which included Stellako River models for both survey method types. Linear regression models primarily took the form of generalized linear regression (GLM) and were created at the Generalized (1), Size-Aggregate (2), Individual (5), and Specialized (6) levels. Two Generalized (1) GLM models were added to the candidate set; one with a water clarity covariate and one without. Two Size-Aggregate (2) models were added; one per size-aggregate level, with only the *small-medium* aggregate model including a covariate for water clarity. Seven Individual (5) GLM models were added, for the same seven system-methods as with the inverse regression models. Three Specialized (6) regression models were added, one applicable to Early Stuart systems and two applicable to systems that have discharge data. One of the discharge models was specific to Stellako. Therefore, a total of 5 Generalized (2), 10 Size-Aggregate (2), 4 Size (3), 6 Size-Clarity (4), 17 Individual (5), and 3 Specialized (6) models were passed through to the model ranking algorithm.

A total of 18 rank-sets passed the robustness thresholds in the model ranking algorithm (Appendix C), at least one from every set-level. Seven rank-sets were passed at the Specialized (6) set-level; however, the two method subsets per discharge model created duplicate results because Stellako River was the only unique system with greater than 10 data points and had complete overlap of data-pairs in both models. Therefore, there are only five unique rank-sets at the Specialized (6) set-level; two for Stellako discharge models (one result per survey method) and three rank-sets for the three Early Stuart systems that comprised the Early Stuart protraction index regression (all ground surveys). Only one rank-set passed at the Individual (5) set-level, which was for Stellako River ground surveys; however, the model selected was from a more general level (Size-Aggregate (2)). Three rank-sets passed at the Size-Clarity (4) set-level, which were for *clear* water clarities in the *small (ground)* and *medium (ground and aerial)* size categories. One Size-Clarity (4) rank-set resulted in the selection of a Size-Clarity (4) model while the other two rank-sets resulted in more general model selections from the Generalized (1) models. Four rank-sets passed at the Size (3) set-level, which were for *small (ground)*, *medium (ground and aerial)*, and *x-large (aerial)* sizes. All four Size (3) rank-sets resulted in more general model selections from the Generalized (1) models. Three rank-sets passed at the Size-Aggregate (2) set-level, all except for the *large-x-large (ground)* subset, which only had 8 samples. All three Size-Aggregate (2) rank-sets resulted in Generalized (1) model selections. Both survey-type subsets at the Generalized (1) level selected GLMs over Ratio-of-Means

models. Overall, generalized models were selected relative to the rank-set's level with the exception of the Specialized (6) rank-sets and one Size-Clarity (4) rank-set.

Nine out of the 45 candidate models were selected within the 18 rank-sets that passed the robustness thresholds, with some models being selected multiple times. All three Specialized (6) models were selected. The general Discharge model (6) was selected for Stellako River *aerial* surveys, whereas the Stellako-specific Discharge model (6) was selected for Stellako River *ground* surveys. The Early Stuart Protraction model (6) was selected for the three *ground* surveyed Early Stuart systems with greater than 10 data points (O'Ne-El Creek, Forfar Creek, and Gluske Creek). One Size-Clarity (4) model was selected, the mean-of-ratios model developed for the Small-Clear-Ground survey subset. One Size-Aggregate (2) was selected, the Small-Medium GLM (2) for Stellako River *ground* surveys when a discharge covariate is not available (rank-set level 5). Four Generalized (1) models were selected across eleven rank-sets. The Generalized (1) GLM with no Clarity covariate (Minimal GLM) model (1) was selected seven times across rank-set levels 1-4: *ground* and *aerial medium-sized clear* surveys (set-level 4), *aerial* and *ground medium-sized* surveys (set-level 3), *aerial* and *ground small-medium-sized* Aggregate surveys (set-level 2), and all sizes *aerial* surveys (set-level 1). The Generalized (1) GLM with a Clarity covariate (Clarity GLM) model (1) was selected twice at two set-levels: *ground small* surveys (set-level 3), and all sizes Ground surveys (set-level 1). The Minimal GLM model was ranked as the next best model in the case of all sizes *ground* surveys if no clarity covariate is available. The *aerial* method Generalized Ratio of Means model (1) was selected once for *aerially* surveyed *large-x-large* aggregate (set-level 2). The combined method Generalized Ratio of Means model (1) was selected once for the *aerially* surveyed *x-large* systems (set-level 3). The Minimal GLM model had the highest occurrence in robust rank-sets as a top model ($n = 7$) and the highest occurrence within the top five results of all rank-sets ($n = 22$).

3.1.2. Expanded Estimates

The updated expansion approach resulted in better average summary statistics in most cases, but did depend in part on how the data were organized. The data were first organized by high-accuracy abundance ranges (bins) that correlate with historical and current definitions of "low abundances" where a visual survey estimate would be planned as the primary estimate method or not (breaks at 25,000 and 75,000). Next, each abundance bin was further split between water clarities (*clear* vs *turbid/tannic*) in addition to reporting on the aggregation of the two water clarities. Along with a summary of all data pairs, this created nine levels of organization that results were investigated (Table 2, Figure 3). Across all data pairs (no subsetting), the average of each metric improved relative to if the default expansion was applied. Overall (all data points combined) MPE improved by 12.8% to -0.5% ($p < 0.0025$), and MAPE decreased by 5.7% to 24.5% ($p < 0.0025$) and the RMSE and correlation coefficient improved by minor amounts. There were only five out of nine other levels of data organization where all four metrics improved after organizing the data at finer subsets; however, the updated estimates were always improved in at least two out of the four metrics in the remaining four subsets.

The T-Test α -level was adjusted to 0.0025 due to the large number of comparisons per test type ($n = 20$ per question, including the overall summary comparisons), which determined the significance of average MPE and MAPE results relative to the default expansion or zero (no bias or error), respectively. Adjusting the α -level did result in five changes to the frequentist significance measure (Appendix C; Table 25). The MPE metric was generally improved upon,

with only the lowest abundance [0-25,000) bin's average at both the combined level ($p = 0.541$) and the *clear* water clarity subset ($p = 0.016$) not being statistically different than the default (Appendix C; Table 25). The updated and default MPE metrics in the lowest abundance bins' were not statistically different than zero MPE (all $p > 0.1$). All other updated MPE metrics were not significantly different than zero; whereas, the only other default MPE metric not significantly different than zero was in the middle abundance [25,000-75,000) bin with *clear* water clarity ($p < 0.025$). The MAPE metric was only statistically better than the default in one instance other than the overall average, which was the largest abundance [75,000+] bin at the *clear* subset ($p < 0.025$). The MAPE would have been significantly different at the unadjusted α -level of 0.05 at the lowest abundance [0-25,000) in the *clear* ($p = 0.003$) and combined ($p = 0.003$) clarities and at the largest abundance [75,000+] bin in the combined ($p = 0.009$) subset. The MAPE was larger in the updated series only in the [25,000-75,000) *clear* bin, but not significantly different than the historical MAPE ($p = 0.628$). The MAPE was only near zero in the lowest abundance [0-25,000) bin for turbid/tannic systems ($p = 0.021$) but the sample size was only seven. The updated RMSE was larger in three instances: the [25,000-75,000) *clear* and combined bins, and [75,000+] *turbid/tannic* bin (no statistical test applied). The updated correlation coefficient was lower in three instances: the *turbid/tannic* bins in each abundance bin: [0-25,000), [25,000-75,000), and [75,000+] (no statistical test applied). Overall, bias (MPE) was always improved but the general accuracy of estimates (MAPE) was not always improved relative to the default 1.8 expansion method; although, not significantly worse in any levels of data organization.

Table 2. Comparison of the mean percent error (MPE), mean absolute percent error (MAPE), root-mean squared error (RMSE) and correlation coefficient performance metrics between updated and default estimate types of Fraser Sockeye expansion methods. Comparisons are organized by abundance bins based on historical definitions that determined where visual estimate methods would be planned (<25,000 prior 2004 and < 75,000 in 2004 and after). Comparisons are also further made within abundance bins between water clarities (clear and turbid/tannic) in addition to combining those levels (combined water clarity). ¹Denotes that the updated estimate's average metric is significantly different than the default estimate type's metric results. ²Denotes that the updated estimate's average metric is not significantly different than zero. ³Denotes that the default estimate's average metric is not significantly different than zero.

Abundance Bin	Water Clarity	n	Average Abundance	Estimate Type	MPE	MAPE	RMSE	Correlation
[0-25,000)	Clear	53	6,553	default	³ 10	20.6	2,113	92.8
				updated	² 3.0	13.8	1,887	93.8
	Turbid/Tannic	7	13,135	default	³ -22.6	30.7	4,316	85.6
				updated	^{1,2} 13.7	² 23.4	4,491	84.9
	Combined	60	7,321	default	³ 6.2	21.7	2,473	90.7
				updated	² 4.3	14.9	2,345	93.0
[25,000-75,000)	Clear	30	43,327	default	³ -16.3	30.6	19,505	17.7
				updated	^{1,2} 6.7	33.1	21,321	36.1
	Turbid/Tannic	13	50,290	default	-37.6	40.7	22,511	72.6
				updated	^{1,2} -6.3	29.3	20,863	71.5
	Combined	43	45,432	default	-22.7	33.7	20,461	41.7
				updated	^{1,2} 2.8	32.0	21,184	51.3
[75,000+]	Clear	45	173,938	default	-21.6	34.0	84,619	83.0
				updated	^{1,2} -6.7	¹ 25.1	55,712	92.6
	Turbid/Tannic	14	223,004	default	-41.5	43.2	94,889	89.3
				updated	^{1,2} -11.5	40.2	125,576	87.2
	Combined	59	185,580	default	-26.4	36.1	87,165	83.7
				updated	^{1,2} -7.8	28.7	78,162	89.2
All	Combined	162	82,358	default	-13.3	30.2	53,670	90.8
				updated	^{1,2} -0.5	¹ 24.5	48,437	92.6

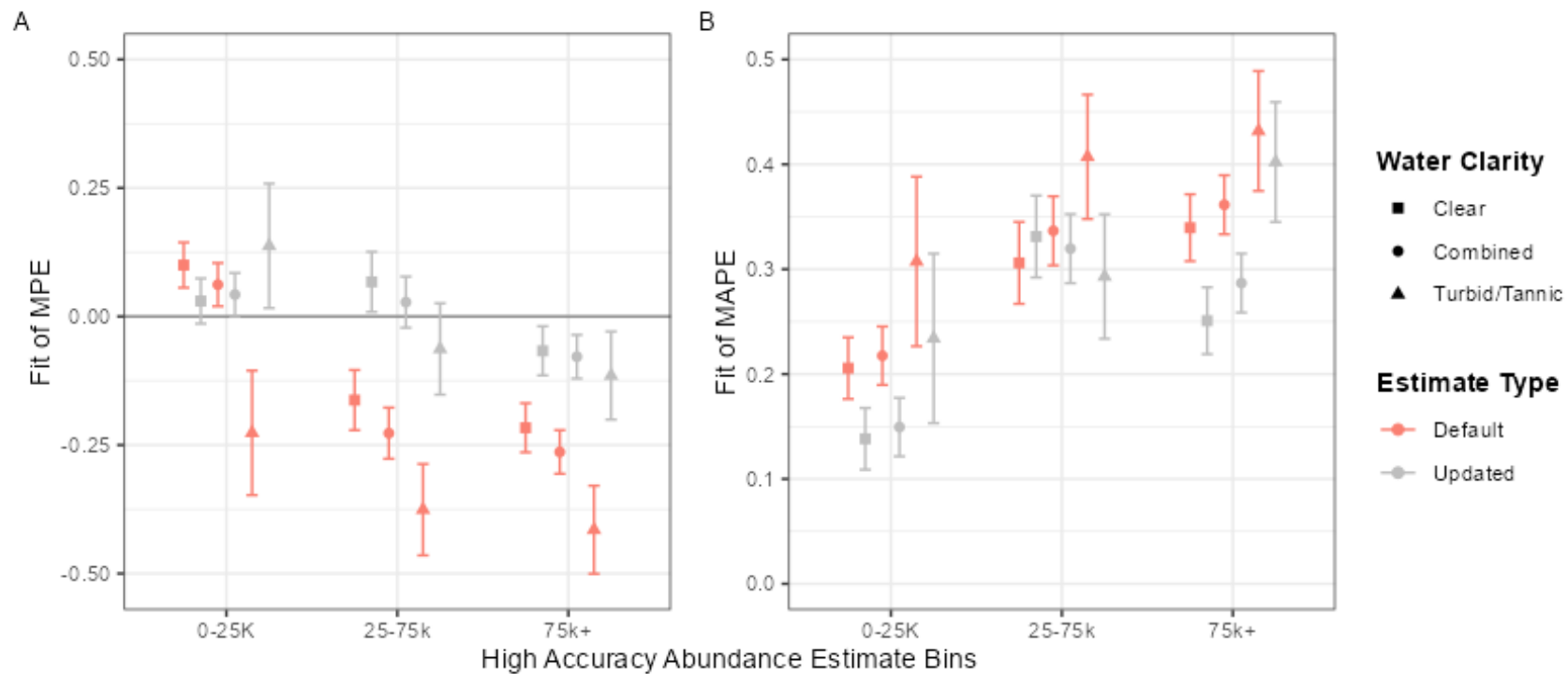


Figure 3. Analysis-of-variance model fits and their standard errors describing the (A) average mean percent error (MPE, which relates to average model bias) and (B) mean absolute percent error (MAPE, which relates to average model accuracy) between the default expansion estimates (red) and the updated estimates (grey) relative to three high accuracy abundance bins (0-24,999, 25,000-74,999, and 75,000+) and water clarity bins (clear-squares, turbid/tannic-triangles, and combined-circles). Sample sizes (n) can be found for each level of data organization in Table 3.

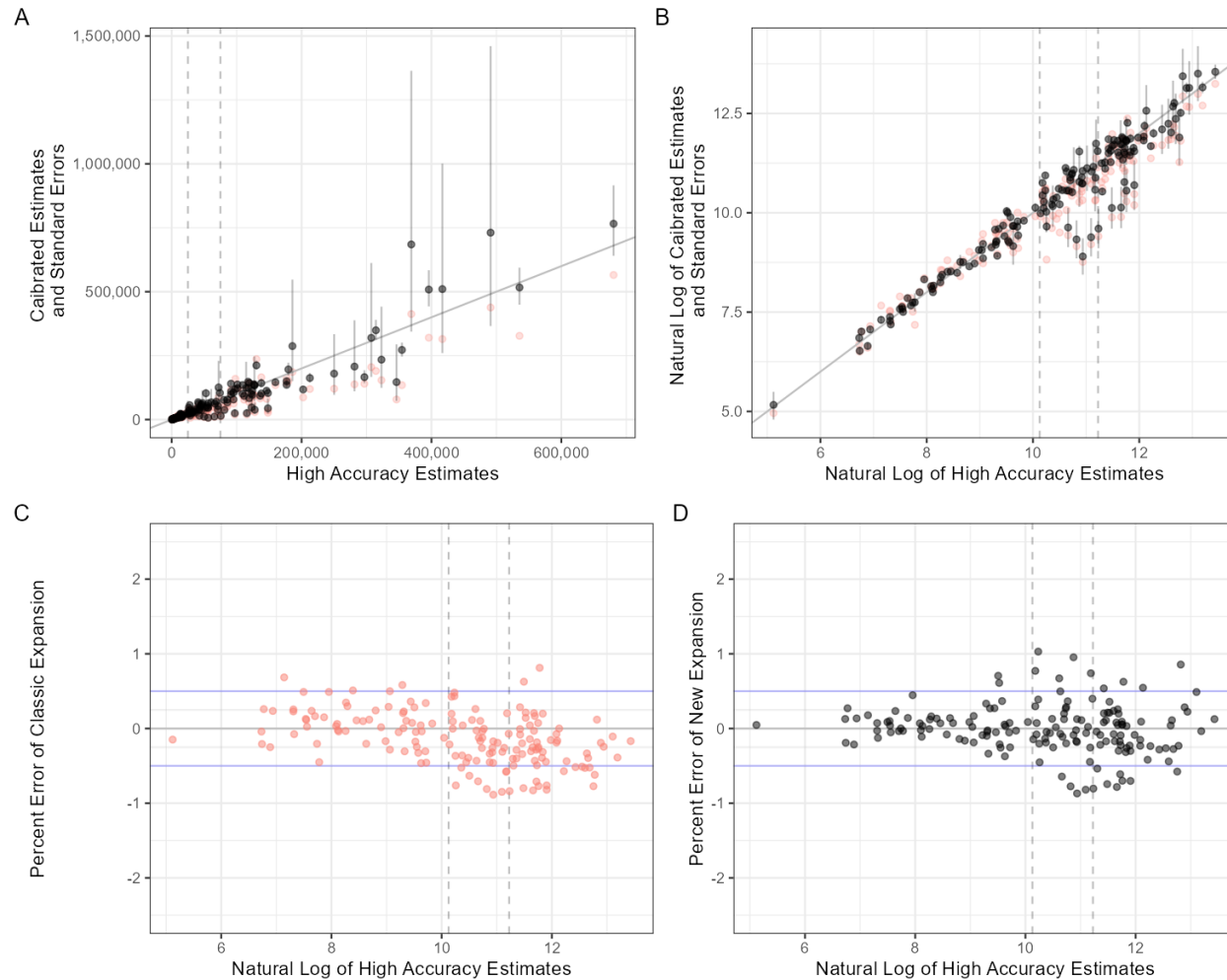


Figure 4. Comparison of default 1.8 expansion factor estimates (red) and updated expanded estimates (black points with 1 standard error whiskers in A & B) from the calibration database ($n = 162$). A) Expanded estimates relative to their high-accuracy estimates in untransformed space. B) Expanded estimates relative to their high-accuracy estimates in natural logarithmic space. C) Percent error of historical expanded estimates relative to their high-accuracy estimates in natural logarithmic space. D) Percent error of updated expanded estimates relative to their high-accuracy estimates in natural logarithmic space. A-D) Solid black line indicates a perfect expansion such that the expanded estimate equals the high-accuracy estimate; dashed lines indicate 25,000 and 75,000, respectively. C-D) Blue lines indicate $\pm 50\%$ relative to 0% error; points above or below the blue lines are described as outliers.

Temporal trends in MPE and MAPE were tested using linear regression and likelihood ratio tests and comparisons made between the default expansion and updated expansion (Appendix C; Table 26). Significance at the α -level was adjusted to 0.00625 due to the multiple comparisons ($n = 8$), which determined the significance of covariates describing linear temporal trends. Likelihood ratio tests identified that addition of a temporal trend parameter to MPE and the addition of an interaction with abundance bins were significant for the default expansion ($p < 0.00625$ in both) but were not significant for the updated expansion ($p = 0.052$ and 0.222 , respectively). The adjusted R^2 for the simple temporal linear model for the default method's MPE was 8.9% and the linear model that included an interaction with abundance bins had an adjusted R^2 of 18.2%. The MPE model for the default expansion with interaction terms had a slightly negative slopes, with the most negative slope occurring at the [0-25,000) bin (-0.010 MPE/year) and less negative slopes for the [25,000-75,000) bin (-0.004 MPE/year) and [75,000+] bin (-0.004 MPE/year). The likelihood ratio tests identified that the addition of a temporal trend parameter to MAPE was significant for both the default expansion and updated expansion ($p < 0.00625$ in both) but neither were improved by the addition of an interaction with abundance bins ($p = 0.107$ and 0.029 , respectively). The adjusted R^2 s for the simple temporal linear models for MAPE were 7.7% for the default expansions and 8.4% for the updated expansions. The temporal trends for both the default and updated expansions were positive, 0.007 MAPE/year and 0.007 MAPE/year, respectively. No model had an R^2 greater than 30%, indicating the models did not explain a large portion of the variance in the data. Generally, there was a small negative temporal trend in MPE in the default expansion method and both the default and updated expansions had small positive temporal trends in MAPE.

Twenty-three outliers ($\text{MAPE} > 0.5$; Figure 4) spanned across multiple levels of data organization in the updated estimates that occurred at different frequencies relative to their frequency in the database. Eighty-seven percent of the outliers were from *aerial* surveys, which is higher relative to their occurrence in the database (*aerial* surveys = 43%); meaning that *ground* survey outlier occurrence (13%) was relatively lower than the database (57%). Sixty-five percent of outliers were from *clear* systems, which is lower relative to their occurrence in the database (*clear* surveys = 79%); meaning *turbid/tannic* system outliers (35%) were disproportionally higher (21%). *Large* (13%) and *x-large* (52%) stream size outliers were higher relative to their occurrence in the database (*large* = 6%, *x-large* = 17 %); whereas *small* (4%) and *medium* (30%) stream size outliers were lower relative to their occurrence in the database (*small* = 36%, *medium* = 40%). *Aerial* surveys of *x-large* systems were generally the least accurate expanded estimates in both water clarities with a general tendency for all surveys that were *turbid/tannic* to be at a higher likelihood of being an outlier.

Six models out of the total nine models used in the updated expansion created outliers. The Early Stuart Specialized (6) Regression, Small-Clear-Ground (4) Mean-of-Ratios, and Small-Medium Aggregate (2) GLM did not produce any outliers. The General Both-Methods (1) ratio-of-means model had the largest discrepancy in estimate outlier occurrence (52%) but had a low application occurrence (17%) in the dataset. The General Both-Methods (1) ratio-of-means model was solely applied to *aerially* surveyed *x-large* systems for both water clarities. The remaining models produced outliers at more similar frequencies to their model application occurrence.

In the updated expansion, there were eleven outlying system-years that were different than the original 28 default outliers and twelve that remained the same (43%). Of the twelve overlapping

outliers, nine were from *aerially* surveyed *x-large* system-years (Harrison River 2015-2017 and 2020,2021; Tachie River 1994, 2008, and 2013; and Little River 2022). The three other overlapping outliers were also from *aerially* surveyed system-years (Adams River (Lower) 2011 and 2022, and Stellako River 2022). The eleven new outliers from the updated expansion were mostly (82%) from *aerial* surveys as well but represented all size classes except for *large*.

All models were developed to also create an estimate of precision or, inversely, uncertainty (Figure 4). The representativeness of an estimate's precision was investigated relative to its range containing the paired high-accuracy estimate. Thirty percent ($n = 49$) of expanded estimate's standard error intervals contained the high accuracy estimate. The average CV of estimates was higher ($CV = 36.9\%$) for estimates whose standard error interval contained the high-accuracy estimate relative to those whose errors did not ($CV = 19.6\%$). Fifty percent ($n = 81$) of expanded estimate's 95% confidence intervals contained the high-accuracy estimate. The average CV of estimates was higher ($CV = 32.1\%$) for estimates whose 95% confidence interval contained the high-accuracy estimate relative to those whose errors did not ($CV = 17.6\%$). Outliers ($n = 23$) had a lower proportion of high-accuracy estimates captured within the standard error interval (13%, $n = 3$) and 95% confidence interval (26%, $n = 6$). Overall, uncertainty was underestimated relative to the models' abilities to capture the high-accuracy estimates within their estimates of precision.

3.1.3. Application to 2024 Data & Additional Expansion Rules based on Expert Advice

The model selection and application framework was verified by subject matter experts (primarily the author group) using the recent 2024 data and estimates to identify potential issues of the methods. The application of the updated expansion process to 2024 data resulted in 53 estimates that can be compared to the default estimates that used the previous expansion factor (1.8 or Early-Stuart-specific) or run-size-adjusted estimates. The 2024 default aggregate estimate for the 53 sites totaled 99,425. The run-size-adjusted estimates totaled 126,039, a 27% increase from the default process. The total using the updated expansion algorithm was 174,797, which is 76% greater than the default total and 39% greater than the RSA total. The number of updated estimates greater than their default outnumbered those that decreased ($n = 49$ and 4, respectively). The average percent difference was 308%, much greater than the average MPE observed in the calibration dataset (-13.3%, Table 2). The largest discrepancy between the updated expanded estimate and the default was 5591%, which was deemed an illogical result by subject matter experts. This discrepancy in estimates highlights an underlying issue of applying the updated expansion algorithm without additional consideration of model behaviour at their extremes.

Through verification and discussion, the discrepancies were determined to be primarily from the 2024 data exceeding the range of the training data. The 2024 visual counts were compared to the range of observed counts in the calibration dataset (i.e. the training data), including a 10% buffer with the assumption that model behaviour just outside the training range should still be reasonable (as determined by the author group visually investigating the relationships). The models apply covariates across multiple factors; therefore, the ranges can be grouped similarly to the range of generalized to specialized set-levels. At the most general level of all data, sub-setting by method only, 60.4% ($n = 32$) of the 2024 visual counts were below 90% of the minimum observation of the training dataset counts (Table 3, Table 5,, Figure 10). At set-level 4 (sub-setting by method, size, and clarity), 2 observations had no representation within the training dataset, and of the remaining 51, 84.3% ($n=43$) of visual counts were below 90% of the

minimum observation. No 2024 visual count was larger than the training data range but instances of this are known to occur in the historical series.

Table 3. Sample sizes and 10% buffered ranges of visual-count data for size-method sets. Note that the minimum is a 10% buffer lower and the maximum is a 10% buffer higher.

Method	Size	n	Buffered Minimum	Buffered Maximum
Aerial	Small	1	20,410	24,940
Aerial	Medium	38	2,230	144,480
Aerial	Large	3	13,370	48,030
Aerial	X-Large	27	3,200	267,800
Ground	Small	58	70	26,050
Ground	Medium	27	4,860	101,090
Ground	Large	7	17,460	345,560
Ground	X-Large	1	47,790	58,410

The regression models all produced the largest positive discrepancies between the default and updated expanded results. Large discrepancies near the origin are due to regression intercepts along the centered and scaled visual count axis varying based on system size, with few sizes having observations less than 10,000 and even fewer less than 1,000. These results guided the additional considerations for application and proposed rules of the updated expansion approach.

Rules for Out-of-Sample-Range Visual Counts

Different approaches were taken by experts to determine which models to apply between low (less than buffered training data) or high (greater than buffered range) abundance visual counts. Most focus was given to low abundance counts since they are most frequent and because high abundance counts will typically have a high-accuracy method planned for implementation and their presence in the historical dataset is low (M. Arbeider personal communication). The survey method and water clarity characteristics were also taken into account during the model development and selection for low abundance counts because standardizing parameters used to estimate their effects will reduce the frequency of post-season run-size adjustments.

Few models included covariates or had specific expansion factors for both survey method and water clarity within the low abundance rank-set. Both regression models, the Small-Medium GLM (2) and Clarity GLM (1), ranked better than the mean-of-ratios models that were developed using the low abundance data subset (Appendix E. Figures Used in Out-of-Sample Range Model Implementation Discussions). However, it is generally not recommended to predict or extrapolate with GLM models outside of their training data range (Bolker 2007). This statistical guideline was particularly poignant when investigating how the expansion factors derived from these regression models behaved neared the origin (Figure 5, Figure 6). The regression-based expansion factors were asymptotic nearing 0 because a logarithm taken of a value less than 1 goes towards negative infinity at 0. This behaviour was also not dampened nearing low values for most stream sizes because the centered intercepts were greater than abundances of 10,000 with the exception of *small* streams that were surveyed from the ground. Various data manipulation exercises were discussed to constrain the behaviour of these models but it was

eventually determined that this effort introduced too much subjectivity and bending of the model as it was designed. The regression-based models and all models that did not have parameters for method or water clarity were therefore removed from comparison, leaving only the mean-of-ratios expansion factors that were developed using the low abundance subset data (Table 6).

The low-abundance mean-of-ratios models ranked moderately in model comparison but otherwise had the desired qualities of the experts. Generally, most models' raw performance metrics were quite similar (Appendix E. Figures Used in Out-of-Sample Range Model Implementation Discussions). The mean-absolute-percent-error in the *aerial* subset was the only metric that the mean-of-ratios model was the worst in relation to all other models; however the difference between it and the top-ranked model was only 7.1%. The average relative-model-rank within the *ground* subset was nearly in a four-way tie for second overall after a top-ranked model that did not account for water clarity. The simplicity of the method is comparable to the default method but includes the added expectation and observation of increasing the expansion under *turbid/tannic* conditions and when a count was done using *aerial* methods. This method also does not suffer any effects of asymptotic behaviour as the expansion factor itself is fixed. Therefore, this straight-forward approach of using the low abundance mean-of-ratios expansion factors was selected for low-abundance visual counts.

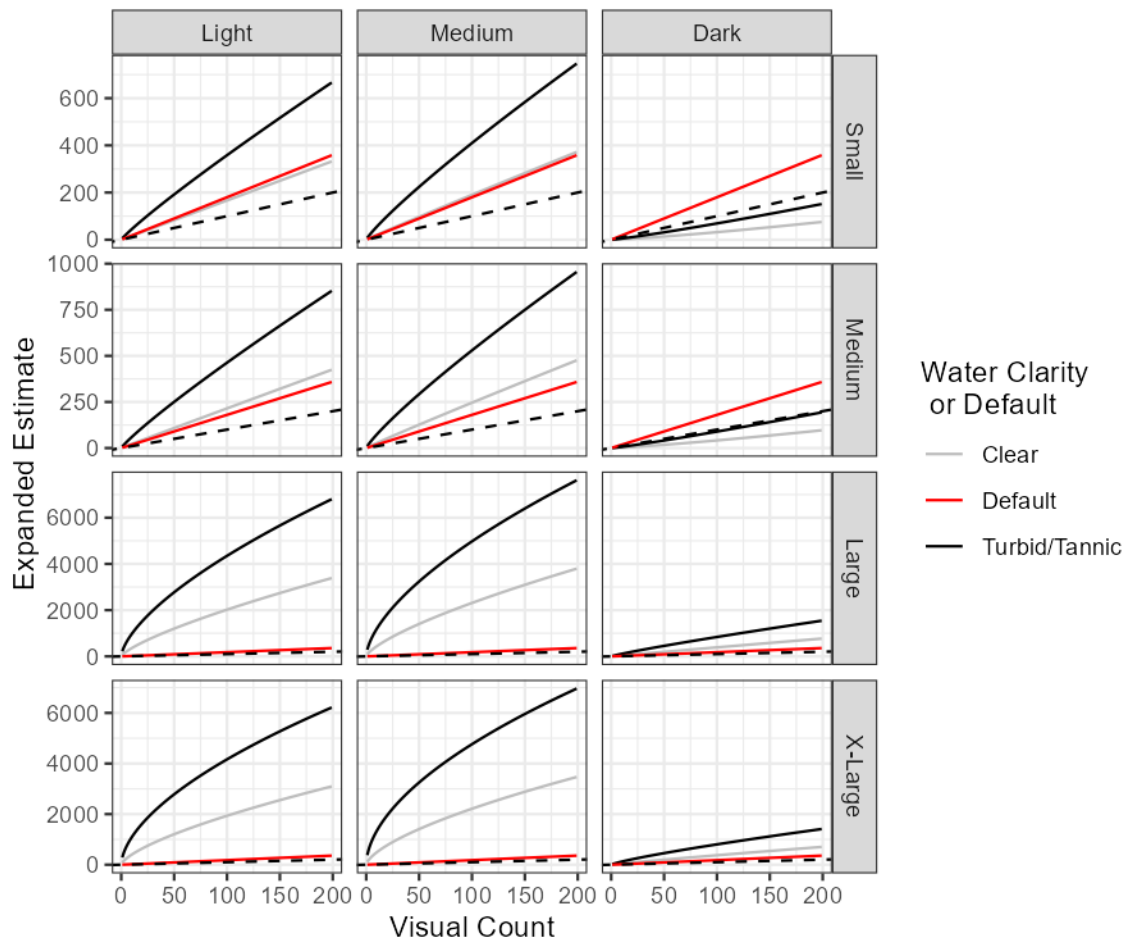


Figure 5. Example of Generalized (1) Clarity Generalized Linear Model predicted expanded estimates relative to their visual count at low abundances. Columns are for Substrate Shade (Light, Medium, Dark), Rows are system Size

(Small, Medium, Large, X-Large), and colours refer to if the survey had water clarity of Clear (grey) or Turbid/Tannic (black). The other lines are the default 1.8 (red) and a 1:1 (dashed black). The survey method is fixed to Ground.

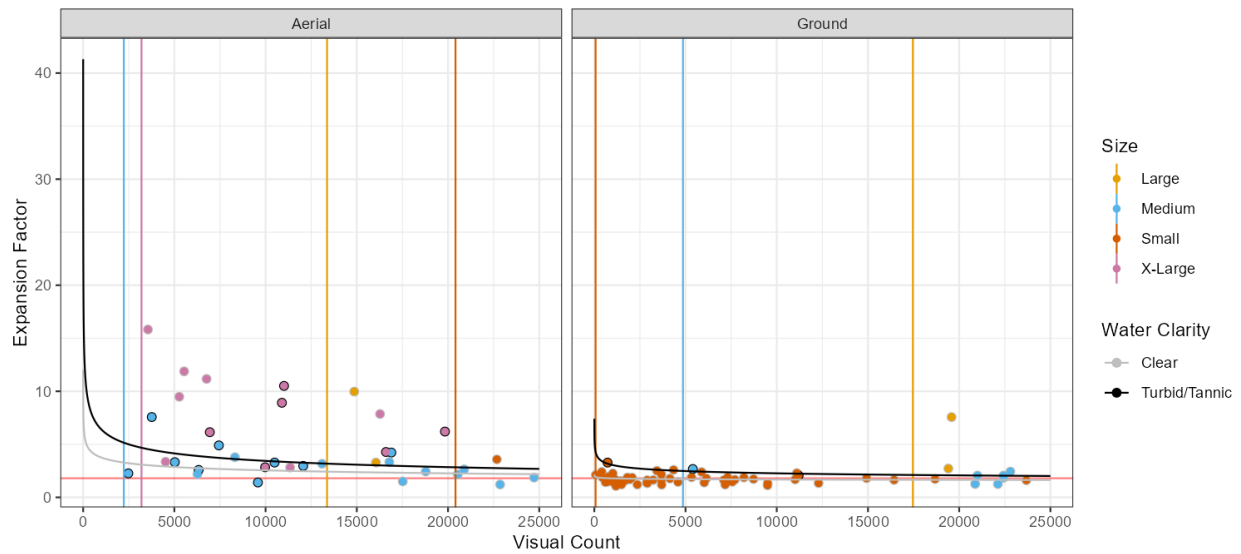


Figure 6. Expansion factors calculated from the model fits of the Generalized (1) Clarity Generalized Linear Model relative to their visual count. The model fit is for Small Systems with Medium Substrate Shade. Columns are for Survey Method (Aerial and Ground), line fit colours refer to if the survey had water clarity of Clear (grey) or Turbid/Tannic (black), and the dot and vertical line colours are sizes (yellow, blue, red, and pink for large, medium, small, and x-large, respectively). The vertical lines are the buffered minimum values from Table 3. The red horizontal line is the default 1.8 expansion factor.

One exception was made for low-abundance out-of-sample visual counts. This exception is for Early Stuart systems that are *small* and *clear*. *Small clear* samples in the database are mostly from Early Stuart systems and have the lowest visual count in the database (79). The model behaviour near the origin was not atypical or asymptotic. Historically, Early Stuart systems have also used an Early-Stuart-specific ratio-of-means due to the high abundance of data. Due to this robust sample, and the high sampling intensity of Early Stuart systems, we decided to make an exception within the algorithm to apply the Early Stuart specific GLM that incorporates a spawning protraction index instead of using the generic low-abundance ratio-of-means expansion factors.

The Minimal GLM model was ranked the highest for data subset greater than 75,000 in both methods separately (Appendix E) and when combined but had similar behavioral issues when predicting outside of the training data ranges. The Clarity GLM behaviour was also investigated during consideration of a model that included a water clarity covariate but the *turbid/tannic* fit dropped below the *clear* fit at higher abundances. That behaviour did not align with the behaviour within the training data range nor the hypothesized relationship. The behaviour of the Minimal GLM was also highly variable based on size (Figure 7, Figure 8) and was determined that too many data modifications would be required to constrain results into what was deemed reasonable by experts. Therefore, the top ranked ratio models were selected for high abundance counts; which were the *ground* method Generalized Ratio of Means model (1) for ground surveys and the *combined* method Generalized Ratio of Means model (1) for aerial surveys.

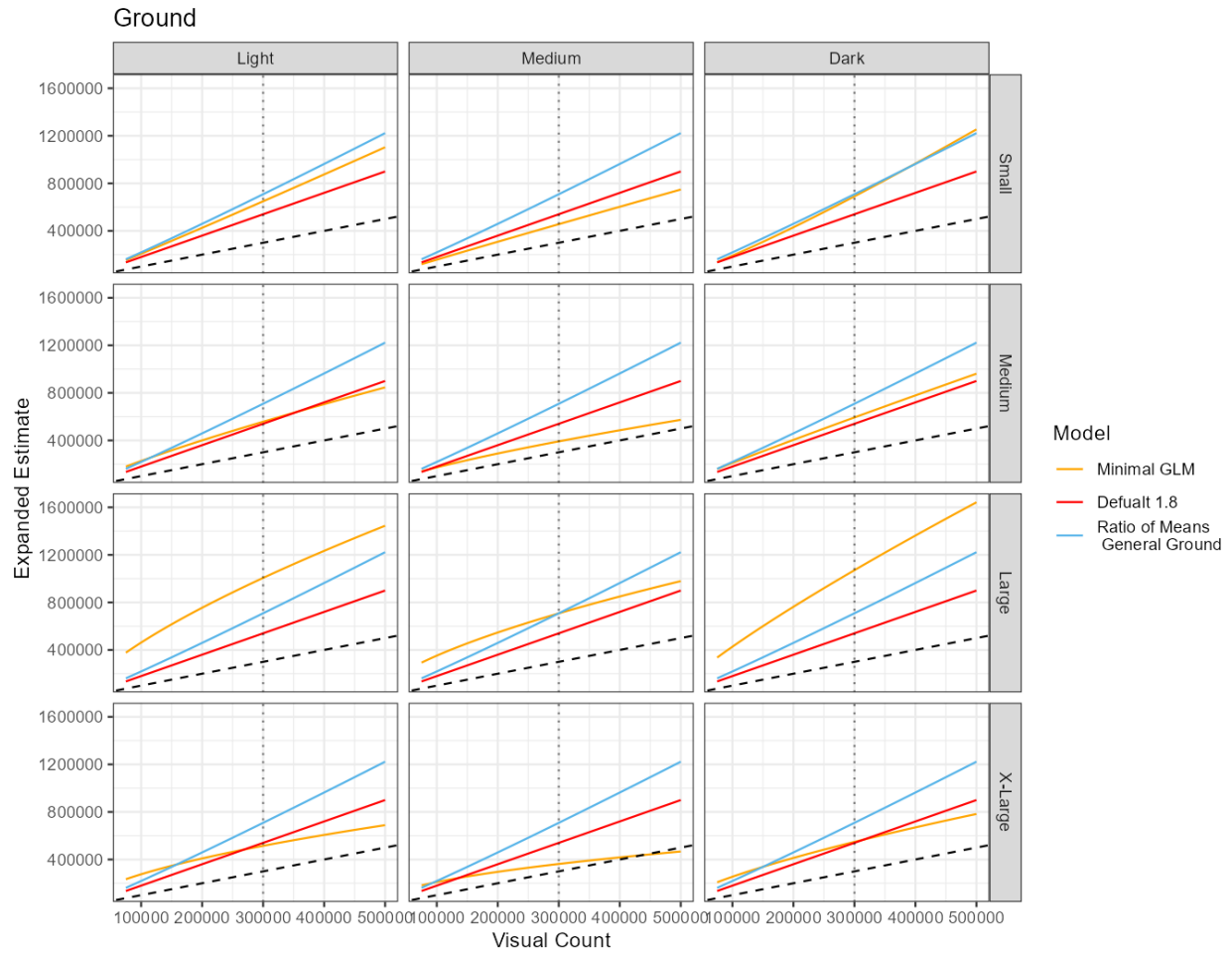


Figure 7. Model predicted expanded estimates relative to their visual count. Columns are for Substrate Shade, Rows are system Size. The lines are Minimal GLM (orange), the default 1.8 (red), the Generalized Ground-only (1) Ratio-of-Means (blue) model fits, and a 1:1 (dashed black). The survey method is fixed to Ground.

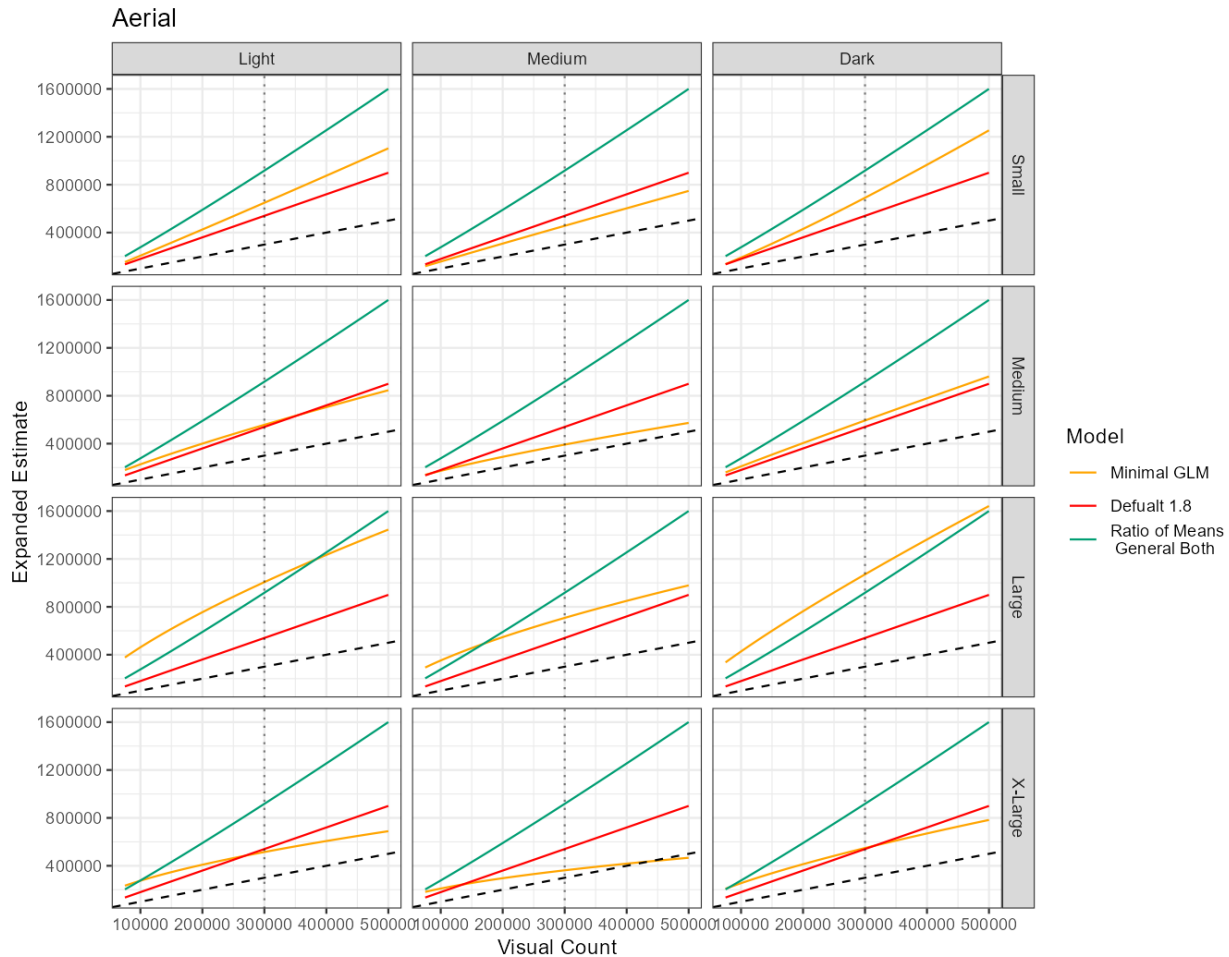


Figure 8. Model predicted expanded estimates relative to their visual count. Columns are for Substrate Shade, Rows are system Size. The lines are Minimal GLM (orange), the default 1.8 (red), the Generalized Both (1) Ratio-of-Means (green) model fits, and a 1:1 (dashed black). The survey method is fixed to Aerial.

High abundance visual counts were also given one different decision rule relative to the low abundance counts. The size-aggregate set-level (2) was chosen to define upper range boundary over the size set-level (3) (which was chosen for the lower range boundary) because the *large* and *x-large* model selection in the algorithm was already at the Size-Aggregate (2) set-level and high abundances are expected more at larger size classes than the smaller classes, which also shared more trends in their observations.

Application to 2024 Data Including Additional Rules

The new aggregate escapement estimate for all spawning sites assessed using the peak count method and updated expansion algorithm in 2024 was 132,198. The new total was 33.0% greater than the default total (99,425) and 4.9% greater than the RSA total (126,039). The average percent differences between the new expanded estimates and the default and RSA estimates across individual sites were 26.0% and 3.2%, respectively. The largest discrepancy between an updated estimate (63) and its default (27) was from Forfar Creek (133% different), which had no RSA. Forfar is an Early Stuart system, and the recovery data suggested a longer than average spawning protraction, which resulted in a larger than default expansion. The next largest discrepancy between an updated expanded estimate (352) and its paired default (151)

and RSA estimate (181) at an individual site was 133% and 94%, respectively. The RSA was made for turbidity and the model applied in the updated estimate (Clarity GLM) included a covariate to account for turbidity. Analytical comparison is difficult to make for many of the other low abundance expansions that had the out-of-sample model applied and resulted in a positive percent difference. There is no way to validate either the default method nor the updated methods with data for estimates lower than the observed database characteristics. The benefit of the updated method is that it can more explicitly convey the uncertainty of estimates made using it and applies standardized expansions based on average observations between factors such as survey method and water clarity instead of relying on arbitrary post-hoc adjustments.

4. DISCUSSION

Accurate and precise information on Pacific salmon populations is crucial for effective management and conservation. This report establishes new methods to improve estimates of spawning Sockeye Salmon in the Fraser River and describes where inaccuracies may be present in historical and future visual-based estimates. On average, the updated approach reduced bias (mean-percent-error) from -13.3% to -0.5% and improved accuracy (mean-absolute-percent-error) from 30.2% to 24.5%. Estimates of precision can also accompany visual expansions now, however, the analysis suggests that they are overconfident and do not always contain the true value. Additionally, a systematic framework was developed to allow for future model development and model selection criteria. Investigating outliers and the application to 2024 data highlight the caveats of this approach, the general limitations of visual methods, and considerations for future applications. Finally, other instances of data interpolating or expanding were summarized in this report, which will allow continued reassessment of methods and foster understanding within the Fraser River and Interior Area and across the DFO.

Bias was pervasive in the calibration database when the default 1.8 expansion factor was applied to all estimates (MPE = -13.3%). Estimates were considerably underestimated on average, with the exception of estimates less than 25,000 in *clear* (water clarity) survey conditions. On average, bias was slightly positive (overestimates) in *clear* surveys when abundance was less than 25,000 when using the default expansion factor, but not significantly so. Negative bias increased as abundance increased and was exacerbated under *turbid/ tannic* water clarities. It is important to note that the default expansion factor was not always applied historically, and since 2009, post-hoc corrections have been attempted through the PSC Fraser Panel Technical Committee run-size-adjustment process (Robinson et al. accepted). Application of the updated expansion models will provide a more consistent approach to dealing with perceived and confirmed bias in visually-based expanded estimates, both through possible historical updates to estimates and in all future application.

The updated expansion approach greatly reduced bias in all scenarios where bias was present. The magnitude of bias was always reduced across abundance levels and water clarities relative to the default model application. Although bias was not reduced to zero on average in most instances, it was also not statistically different than zero, indicating that the updated approach generally achieved the objective of reducing bias. The bias was only similar to the default at abundances less than 25,000 in *clear* surveys; however, both the default and updated bias estimates were also not significantly different than zero. Within abundances between 25,000-74,999, bias changed from very negative using the default to slightly positive or slightly negative with the updated expansion for *clear* and *turbid/tannic* surveys, respectively. Bias remained

negative for all survey conditions for abundances greater than 75,000, but it was greatly reduced relative to the default 1.8 expansion. The direction of bias for *turbid/tannic* water clarities was more variable, which may be indicative of their susceptibility to other conditions that compound impacts to observer efficiency. For example, low light conditions may reduce observer efficiency by four times in *turbid/tannic* water but only reduce efficiency by two times in *clear* water, resulting in more contrast in observer efficiencies within *turbid/tannic* surveys. Contrasts in observer efficiency in *turbid/tannic* water surveys may also provide partial explanation of the average accuracy of results.

The updated expansion approach resulted in modest improvements to the average accuracy of estimates. Accuracy was improved overall and at abundances below 25,000 and above 75,000 but remained equivalent for middling abundances. When considering water clarity, accuracy was not improved significantly in turbid surveys at any abundance. *Clear* survey conditions in the middling abundances (25,000-74,999) was also the only data subset to have an average accuracy lower than the default method, albeit not significantly different. Similar to bias, the accuracy magnitude was greater and more variable for *turbid/tannic* surveys within each abundance subset. Variability in precision and accuracy metrics could be due to the lower sample sizes of *turbid/tannic* surveys in the database. The low sample sizes forced a grouping of *turbid* and *tannic* systems prior to analysis, despite them likely having different average effects on observer efficiency.

Low sample sizes create two considerations when interpreting the results. Averages and their precision created with smaller samples are more susceptible to outliers and variability, and so are models that aim to explain that variability. Low sample sizes reduced the ability to include more covariates or interaction effects during model development. Low samples of covariates are partially due to their inconsistent collection and partially due to naturally low occurrences where paired-data have been implemented historically. The separation of *turbid* and *tannic* as separate factors has not been possible due to their low sample sizes. Their grouping is due to the assumption that they would have the same direction of bias relative to *clear* streams but it is hypothesized that they would not have the same magnitude of effect. Identification of low accuracy outliers within the updated expansion framework may help guide future data collection, both with regards to adding calibration data pairs as well as for planning survey methods to produce more accurate results relative to a stream's characteristics.

Two system characteristics resulted in higher proportions of outliers than their occurrence in the database. Most low accuracy outliers were from *aerial* surveys even though they comprised just less than half of the database. There are likely multiple factors contributing to the variability in *aerial* expansion factors that may not affect *ground* surveys. Logistics may be one driver of *aerial* survey observer variability. The timing of *aerial* surveys are limited by a number of factors from pilot availability to weather conditions. These factors may result in the peak-of-spawn being missed or having to settle for a suboptimal weather day and observer conditions. Repeat flights are also costly and limited, making it difficult to determine whether the peak-of-spawn was captured. Cumulative dead is also often not estimated during aerial surveys and only dead at the time of the peak count is available, which can result in an underestimate of carcasses if predation or river flushing is high. *Aerial* survey observer efficiency is also confounded by the second stream characteristic that resulted in a higher occurrence of outliers, which was stream sizes from the *large-x-large* aggregate. *Large* and *x-large* streams are not logistically feasible to survey with ground or raft methods in many cases. Stream width, depth, and length can make streams impossible to survey from a single raft or ground point, which requires an *aerial*

assessment method to be applied. Multiple helicopter passes are sometimes necessary to enumerate the full wetted width, which can increase the risk of double counting or missing fish due to their movement. *Large* and *x-large* streams are deeper on average or have an increased occurrence of deep pools where sockeye can be stacked up in high densities, making enumeration of individuals difficult. Carcass enumeration also becomes difficult as carcasses sink and collect in slower deeper pools. Collecting enough additional data pairs of aerially surveyed, *large* or *x-large* streams may be difficult to describe the variability of expansion factors required to create accurate estimates due to numerous compounding interactions.

Further consideration of caveats of the expansion framework come from testing with the 2024 data. Most of the visual count data in 2024 fell outside the observations in the calibration dataset, which is a possible reason why there were initially large discrepancies between the 2024 run-size-adjustment estimates and the updated expanded estimates. The regression methods resulted in the largest discrepancies; which may be expected considering the lack of data near the origin. Many of the 2024 visual counts were lower than the observations in the calibration database by at least one order of magnitude, meaning that the expansion methods were applied to data that were outside of their observed (i.e. training) range. Unfortunately, low peak counts would likely be the norm and not the exception for historical and future counts (Patterson et al. in press). It is difficult to find resources towards collecting paired data points on such low abundances under current methodological allocation policies (see introduction). In the interim, while little data exists for low abundance counts, expert driven decision rules were developed for how to apply the new models and tested with the 2024 data. The rules determine which model to use at low or high abundances as well as when certain parameters must be altered to align with realistic expectations and expert-observations. These standardized guidelines resulted in what were deemed acceptable expansions by experts while accounting for some of the major system characteristics that affect observer efficiency such as water clarity and survey method.

Low-abundance ratio-of-means models that were subset by method and water clarity were selected for instances where visual counts fell below the observed calibration dataset range. One additional rule and consideration was also created to ensure realistic or at least more precautionary expansions due to lack of robust data at low abundances. Fish behaviour at low abundances was considered and used to create a decision rule that could allow the system expert to change a system's water clarity from *turbid/tannic* to *clear*. The conditions for this change to be applied are that the spawning needs to occur in sections of the river where visibility is not impeded by turbidity and that mostly spawners are observed during the survey. The hypothesis is that even if the stream is *partially turbid* or *tannic* on average, that observer efficiency will not be impeded by it at lower densities because there is less competition for spawning grounds or less need to hold in deeper pools. These rules and considerations should allow biologists to approach expansions in a standardized way but still allows for some expert opinion to inform expansions in a precautionary manner.

An additional solution to the low count caveats is to aggregate counts of systems that share similar characteristics and then expand the aggregate. By aggregating data, individual systems would be treated more like reach (survey) breaks that are part of a larger system. The total count may be more likely to fall within the range of the observed calibration data and its application would align with the model performance assessments. The expanded estimate would then be allocated based on the proportion of the count each system contributed. This idea was first considered with regards to lake shore counts, which are sometimes already

treated as reaches within a larger lake system. It may be reasonable to consider streams within a CU as constituents of a larger total count for an aggregated system. Aggregating systems may also gain the benefits of the law of large numbers and changing the focus to a coarser scale (Levin 1992, Schindler *et al.* 2015), where observer efficiencies that are above or below the average on different streams counteract each other and align more with the average trends described by the expansion models.

For instances where visual counts were greater than the calibration dataset range, the Generalized (1) Ratio-of-Means Ground-Only and Both models were selected for *ground* and *aerial* counts, respectively. Higher abundance observations in the calibration dataset were primarily from *large* and *x-large* sized systems and generally lacked contrast in the water clarity covariate. Fewer rules were discussed for expansions that would be larger than the training set because their occurrence should be limited and there were fewer than 15 instances historically that had expansions applied that resulted in estimates greater than 200,000. Expert judgement may still be required for scenarios where a visual count is greater than the range of the calibration dataset, particularly because these methods do not account for water clarity.

There are additional areas where future research could be targeted to evaluate this calibration research. The calibration framework is a novel application of the leave-one-out cross validation approach through the use of the set-level framework. Simulation testing could be used to test the robustness thresholds used in the algorithm but larger sample sizes at each set-level may also be required to increase thresholds without losing specialized levels. The inherent nature of the set-level approach could also be tested more extensively as well, to identify if there are other intrinsic flaws not apparent through this application. Comparative methods such as decision-tree analysis or machine learning (Ripley 2007, Pestal *et al.* 2023) could offer alternative frameworks. Additionally, alternative models' within the framework could be explored, such as measurement-error models, to account for the variance in the high accuracy estimates. Or alternatively, performance could be compared across data subsets that only use fence counts (that have $CV = 0$) relative to those that use SONAR or mark-recapture ($CV > 0$) to determine if there are differences in overall accuracy.

5. CONCLUSION

The updated expansion framework will reduce the amount of post-hoc adjustments in visual estimates required due to concerns of persistent bias that have typically occurred during the run-size-adjustment process. The average bias was improved in all subsets of the calibration database results. The accuracy of visual expansions was not greatly improved, highlighting the persistent lower accuracy of peak-count-based methods. Estimates of precision can also be reported, but the confidence interval estimates were still unrepresentative (over-confident) and did not contain the true abundance estimates 50% of the time. The application of the updated expansion framework will require users to consider various scenarios but can follow a standardized decision pathway during visual count expansion to account for its caveats:

1. The updated expansion framework should be applied to peak visual counts if the count is contained within the range of the dataset plus or minus a 10% buffer on the maximum and minimum values, respectively (Figure 9).
 - Post-analysis corrections based on the average bias (Table 2) should be considered under the following count characteristics:
 - *Aerially* surveyed counts (low accuracy);

- *Large* and *x-large* systems (low accuracy);
 - Expansions with results between 25,000-75,000 where a water clarity covariate was not used in the model (*clear* has positive bias and *turbid/tannic* has negative bias);
 - Any surveys larger than 75,000 (persistent slightly negative bias);
 - Survey conditions that have behavioral or environmental anomalies present such as protracted spawning or holding behaviors, absence of a peak-of-spawn observation, or atypical carcass flushing rates (e.g. due to high scavenging rates or high water events).
 - Alternative methods should be considered for systems that are not represented by the systems in the calibration database (e.g. the North Thompson River mainstem).
2. If the visual count is less than 90% of the minimum training data range of its method-size subset:
- Apply the method-clarity specific Low Abundance Mean-of-Ratios expansion factor.
 - If the system was *partially turbid* or *tannic* on average, consider if the abundance was distributed more in shallow water where the water clarity was more similar to *clear* visibility. Change the water clarity to *clear* based on expert judgement.
 - If the system was affected by any of the aspects listed under bullet 1, consider applying changes or run-size-adjustments based on expert judgement.
3. If the visual count was greater than 110% of the maximum training data range of its method-size-aggregate subset:
- Apply the Ground or Combined Generalized (1) Ratio-of-Means if the survey was *ground* or *aerial*, respectively.
 - Additional expert judgement may be required.

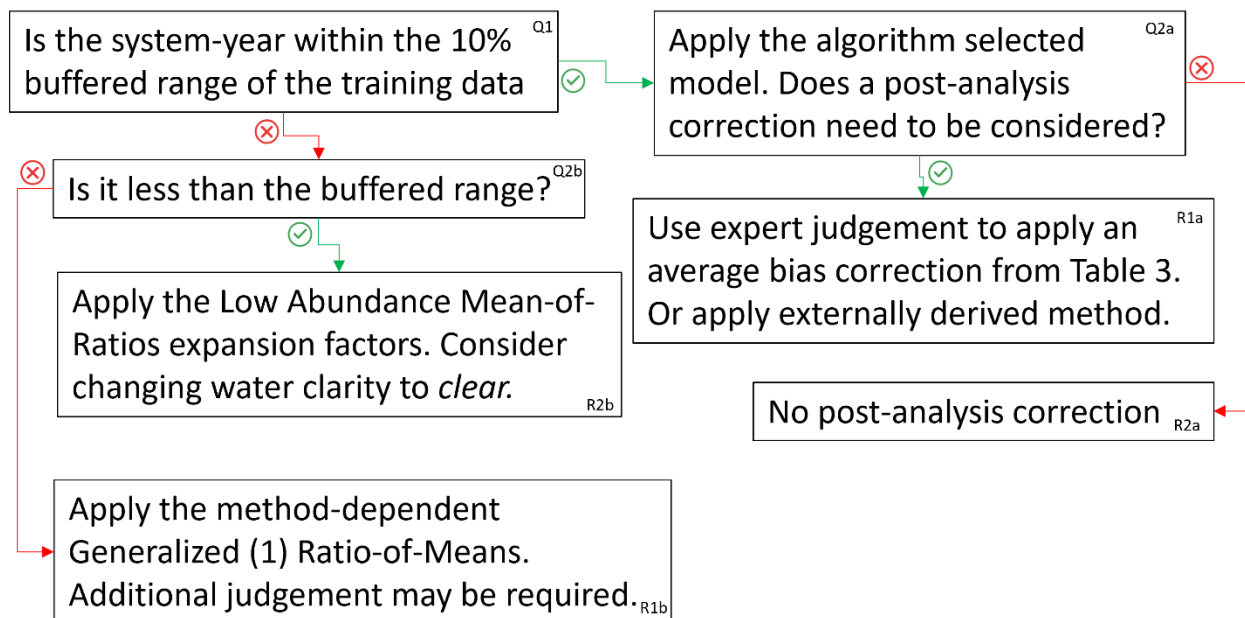


Figure 9. Expansion factor and adjustment decision pathway.

Increasing sample sizes in the calibration database has the potential to improve on the accuracy for some of the noted caveats. Validation of low counts (e.g. < 2,500) in *medium* and *larger* sizes would likely result in the greatest increase in certainty-per-system annually with potential cumulative impacts to the total magnitude of Sockeye estimated by visual methods. However, despite the cumulative impacts, the annual, MU-level impacts would likely be minimal due to the small proportions these counts add to the total annual abundances. The coefficients of variation for expansions from low counts were higher than the method's average in the calibration database (had the lowest precision) due to the method being applied to data outside of the observed range of the calibration dataset. Even the addition of low count calibration pairs for *small* size classes (e.g. 1-500 for *ground* counts and < 2,500 for *aerial* counts) would benefit measures of certainty (assuming there is a reliable relationship). Although the MU-level impacts may be minimal, status assessments of DUs and CUs with low abundances will remain uncertain when visual methods are used and remain unvalidated by observations.

There may also be so much variability within some characteristic sets that achieving high accuracy results will require an incredibly large sample size (e.g. *aerially* surveyed *large* and *x-large* systems) and the incorporation of additional or more specific covariates. For example, increases in *turbid* and *tannic* samples at all stream sizes and methods would benefit the database as their results were the most uncertain and variable. Observers also report that visibility between *turbid* and *tannic* systems is quite different, despite both being worse than *clear* systems on average. Quantifying mechanisms that may interact with observer efficiency in *turbid* and *tannic* environments is a knowledge gap, and having great enough sample sizes to test covariates to measure them is potentially beneficial. Controlled experiments on systems with variable turbidity or *tannic* water may also be an avenue to collect information on the effect of observer efficiency from interactions like cloud cover, brightness, glare, etc.

When considering updating historical estimates, there may be multiple considerations on how to prioritize them. Two possible ways to prioritize them could be based on those that may have the largest-per-estimate correction or to select counts that fall within the calibration training scope. Estimates greater than 75,000 made with the default expansion had the largest and most consistent negative bias and had the greatest improvements to bias and accuracy through the updated framework (with accuracy in *turbid/tannic* surveys being an exception of having no improvement). However, *the* largest estimates may also be outside of the observed training data, and may require additional discussion. Estimates between 25,000-75,000 could be prioritized next based on their relative improvements to bias and magnitude of correction, followed by <25,000. Adjusting estimates from visual counts outside of the observed training data carries inherent uncertainty and may also increase the amount of effort by area experts in determining if covariates need to be fixed based on the algorithm's suggestions. Data management of all downstream uses of the data should also be considered as multiple organizations and user groups exist and may be impacted by edits to escapement estimates (COSEWIC 2017, Grant et al. 2020, Doutaz et al. 2023, Robinson et al. in press).

This expansion framework was developed to allow for continuous improvement. After an initial review and update to historical expanded estimates, it is recommended to do an impact assessment of further updates because they may have more costly time commitments than what their impact is on Fraser Sockeye conservation and fisheries management. It is recommended to conduct sensitivity analysis on underlying relationships that use spawner estimates and affect fisheries management decisions to determine the threshold of a change in estimates required for editing historical records. It is recommended to publish subsequent

technical report documents when changes to historical estimates that use this updated expansion framework occur or when updates to the framework are completed. Whereas for all future visual count expansions on Fraser Sockeye, it is recommended to implement this updated expansion framework.

6. ACKNOWLEDGEMENTS

We would like to thank and acknowledge the decades of data collection done by countless DFO staff, First Nations, and other partner groups; and for the forethought of prior biologists for implementing the collection of this data. Prior unpublished reports and analysis were completed by DFO staff Brian Leaf, Paul Welch, Keri Benner, Andrew Grant, Scott Decker, Mike Hawkshaw, and Katie Davidson. Funding for targeted calibration work was provided by the PSC Southern Boundary Enhancement and Restoration Fund. General operations where calibration data points were opportunistically collected were funded by DFO Science. Funding for the authorship of this report was partially provided through the Pacific Salmon Strategy Initiative and DFO Science. We acknowledge that salmon enumeration data was collected in the traditional territories of many different First Nations throughout British Columbia but that this assessment was completed only through the lens of settlers and western science.

7. REFERENCES

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *In*: B.N. Petrov and F. Csáki, eds. *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR*. Budapest: Akadémiai Kiadó, 267–281.
- Albers, S., 2017. tidyhydat: Extract and Tidy Canadian Hydrometric Data. *The Journal of Open Source Software*, 2 (20), 511.
- Andrew, J.H. and Webb, T.M., 1987. *Review and assessment of adult sockeye salmon enumeration programs on the Fraser River*. Vancouver, BC.
- Armstrong, J.S., 1985. *Long-range forecasting: from crystal ball to computer*. Second Edition. Toronto: John Wiley & Sons.
- Armstrong, J.S., 2001. Evaluation Forecasting Methods. *In*: J.S. Armstrong, ed. *Principles of Forecasting: A handbook for Researchers and Practitioners*. Boston: Kluwer Academic Publishers.
- Armstrong, J.S. and Fildes, R., 1995. On the Selection of Error Measures for Comparisons Among Forecasting Methods. *Journal of Forecasting*, 14, 67–71.
- Arnason, A.N., Kirby, C.W., Schwarz, C.J., and Irvine, J.R., 1996. Computer analysis of data from stratified mark-recovery experiments for estimation of salmon escapements and other populations. *Can. Tech. Rep. Fish. Aquat. Sci.*, 2106, vi + 37 p.
- Bailey, R.E., Parken, C.K., Irvine, J.R., Rosenberger, B., and Farwell, M.K., 2000. Evaluation of utility of aerial overflight based estimates versus mark-recapture estimates of chinook salmon escapement to the Nicola River, B.C. *Can. Sci. Adv. Sec. Res. Doc.*, (2000/152).
- Bartoń, K., 2024. _MuMIn: Multi-Model Inference_ [online]. *R package version 1.48.4*. Available from: <https://CRAN.R-project.org/package=MuMIn> [Accessed 13 Apr 2025].
- Bevan, D.E., 1961. Variability in Aerial Counts of Spawning Salmon. *Journal of Fisheries Research Board of Canada*, 18 (3).
- Bolker, B.M., 2007. *Ecological Models and Data in R*. Princeton: Princeton University Press.
- Bonferroni, C., 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commerciali di firenze*, 8, 3–62.
- Burnham, K.P. and Anderson, D.R., 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33 (2), 261–304.
- Cabin, R.J. and Mitchell, R.J., 2000. To Bonferroni or Not to Bonferroni: When and How Are the Questions. *Bulletin of the Ecological Society of America*, 81 (3), 246–248.
- Cochran, W.G., 1977. *Sampling Techniques*. New York: John Wiley & Sons.
- COSEWIC, 2017. *COSEWIC assessment and status report on the Sockeye Salmon *Oncorhynchus nerka*, 24 Designatable Units in the Fraser River Drainage Basin, in Canada*. Committee on the Status of Endangered Wildlife In Canada.

- Cousens, B.F., Thomas, G.A., Swann, C.G., and Healey, M.C., 1982. A review of salmon escapement estimation techniques. *Can. Tech. Rep. Fish. Aquat. Sci. No. 1108*.
- Cronkite, G.M.W., Enzenhofer, H.J., Ridley, T., Holmes, J., Lilja, J., and Benner, K., 2006. Use of high-frequency imagine sonar to estimate adult sockeye salmon escapement in the Horesfly River, British Columbia. *Can. Tech. Rep. Fish Aqu. Sci.*, 2647.
- Dangel, J.R. and Jones, J.D., 1988. *Southeast Alaska Pink Salmon Total Escapement and Stream Life Studies, 1987*. Juneau, Alaska.
- Darroch, A.J.N., 1961. The Two-Sample Capture-Recapture Census when Tagging and Sampling are Stratified. *Biometrika*, 48 (3), 241–260.
- DFO, 2005. *Canada's Policy for Conservation of Wild Pacific Salmon*. Fisheries and Oceans Canada. Canada.
- DFO, 2007. Revised protocol for conducting recovery potential assessments. *DFO Can. Sci. Advis. Sec. Sci. Advis. Rep.*, (2007/039).
- DFO, 2024. Rapid status approximations for Pacific salmon derived from integrated status assessments under DFO's Wild Salmon Policy. *DFO Can. Sci. Advis. Sec. Sci. Resp.*, 004, 42.
- Dionne, K. and Liu, Q., 2024. Pre-season Run Size Forecasts for Fraser River Sockeye (*Oncorhynchus nerka*) Salmon and Pink Salmon (*Oncorhynchus gorbuscha*) in 2023. *Can. Tech. Rep. Fish. Aquat. Sci.*, 3634, iv + 120 p.
- Doutaz, D., Huang, A., Decker, S., and Vivian, T., 2023. *Recovery Potential Assessment for Fraser River Sockeye Salmon (Oncorhynchus nerka), Nine Designatable Units Part 2: Biology, Habitat, Threats, Mitigations and Allowable Harm - Elements 1-11, 14, 16-18, 22*.
- English, K.K., Bailey, R.E., and Robichaud, D., 2007. Assessment of Chinook returns to the Fraser River watershed using run reconstruction techniques 1982-4. *DFO Can. Sci. Advis. Sec. Res. Doc*, 2007/020, vi + 76.
- Enzenhofer, H.J. and Cronkite, G., 2000. Fixed Location Hydroacoustic Estimation of Fish Migration in the Riverine Environment: An Operational Manual. *Can. Tech. Rep. Fish. Aquat. Sci.*, (2312), 46p.
- Falcy, M.R., McCormick, J.L., and Miller, S.A., 2016. Proxies in practice: Calibration and validation of multiple indices of animal abundance. *Journal of Fish and Wildlife Management*, 7 (1), 117–128.
- Farwell, M.K., Bailey, R.E., and Rosenberger, B., 1999. Enumeration of the 1995 Nicola River Chinook Salmon Escapement. *Can. Manuscr. Rep. Fish. Aquat. Sci.*, 2491, 44.
- Folkes, M., Thomson, R.E., and Hourston, R.A.S., 2018. Evaluating Models to Forecast Return Timing and Diversion Rate of Fraser Sockeye Salmon. *DFO Can. Sci. Advis. Sec. Res. Doc.*, 2017/021, vi + 220p.
- Gelman, A. and Hill, J., 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

- Grant, S.C.H., Holt, C.A., Pestal, G., David, B.M., and MacDonald, B.L., 2020. The 2017 Fraser Sockeye Salmon (*Oncorhynchus nerka*) Integrated Biological Status Re-Assessments Under the Wild Salmon Policy Using Standardized Metrics and Expert Judgment. *DFO Can. Sci. Advis. Sec. Res. Doc.*, (2020/035), vii + 221.
- Greenwell, B.M. and Shubert Kabban, C.M., 2014. investr: An R Package for Inverse Estimation. *The R Journal.*, 6 (1), 90–100.
- Groot, C. and Margolis, L., 1991. *Pacific Salmon Life Histories*. Vancouver: UBC Press.
- Healey, M.C. and Heard, W.R., 1984. Inter- and Intra- Population Variation in the Fecundity of Chinook (*Oncorhynchus tshawytscha*) and its Relevance to Life History Theory. *Canadian Journal of Fisheries and Aquatic Sciences*, 41, 476483.
- Hilborn, R., Bue, B.G., and Sharr, S., 1999. Estimating spawning escapements from periodic counts: A comparison of methods. *Canadian Journal of Fisheries and Aquatic Sciences*, 56 (5), 888–896.
- Hilborn, R. and Walters, C.J., 1992. *Quantitative Fisheries Stock Assessment: Choice, Dynamics, & Uncertainty*. Springer Science.
- Holmes, J.A., Cronkite, G.M.W., Enzenhofer, H.J., and Mulligan, T.J., 2006. Accuracy and precision of fish-count data from a 'dual-frequency identification sonar' (DIDSON) imaging system. *ICES Journal of Marine Science*, 63 (3), 543–555.
- Holt, C.A., 2009. Evaluation of Benchmarks for Conservation Units in Canada's Wild Salmon Policy: Technical Documentation. *DFO Can. Sci. Advis. Sec. Res. Doc.*, 3848/059, x+50.
- Holt, C.A. and Bradford, M.J., 2011. Evaluating Benchmarks of Population Status for Pacific Salmon. *North American Journal of Fisheries Management*, 31 (2), 363–378.
- Holt, K.R. and Cox, S.P., 2008. Evaluation of visual survey methods for monitoring Pacific salmon (*Oncorhynchus* spp.) escapement in relation to conservation guidelines. *Canadian Journal of Fisheries and Aquatic Sciences*, 65 (2), 212–226.
- Hughes, N.F., 2004. The wave-drag hypothesis: An explanation for size-based lateral segregation during the upstream migration of salmonids. *Canadian Journal of Fisheries and Aquatic Sciences*, 61, 103–109.
- Hurlbert, S.H., 1984. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs*, 54 (2), 187–211.
- Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22 (4), 679–688.
- Jean Dunn, O., 1961. Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56 (293), 52–64.
- Jones, E.L., Quinn, T.J., and Van Alen, B.W., 1998. Observer Accuracy and Precision in Aerial and Foot Survey Counts of Pink Salmon in a Southeast Alaska Stream. *North American Journal of Fisheries Management*, 18 (4), 832–846.

- Korman, J., Ahrens, R.N.M., Higgins, P.S., and Walters, C.J., 2002. Effects of observer efficiency, arrival timing, and survey life on estimates of escapement for steelhead trout (*Oncorhynchus mykiss*) derived from repeat mark-recapture experiments. *Canadian Journal of Fisheries and Aquatic Sciences*, 59 (7), 1116–1131.
- Levin, S.A., 1992. The problem of pattern and scale in ecology. *Ecology*, 73 (6), 1943–1967.
- Lilja, J., Ridley, T., Cronkite, G.M.W., Enzenhofer, H.J., and Holmes, J.A., 2008. Optimizing sampling effort within a systematic design for estimating abundant escapement of sockeye salmon (*Oncorhynchus nerka*) in their natal river. *Fisheries Research*, 90, 118–127.
- Liu, Q. and Xu, Y., 2024. Pre-season Run Size Forecast for Fraser River Sockeye (*Oncorhynchus nerka*) Salmon in 2022. *Can. Tech. Rep. Fish. Aquat. Sci.*, v + 340p.
- Maas, C.J.M. and Hox, J.J., 2004. Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58 (2), 127–137.
- MacDonald, B.L. and Grant, S.C.H., 2012. Pre-season run size forecasts for Fraser River Sockeye salmon (*Oncorhynchus nerka*) in 2012. *DFO Can. Sci. Advis. Sec. Res. Doc.*, 2012/011, v + 64p.
- Martins, E.G., Déry, S.J., and Patterson, D.A., 2023. Fraser River Basin. In: M.D. DeLong, T.D. Jardine, A.C. Benke, and C.E. Cushing, eds. *Rivers of North America*. Elsevier, 650–674.
- Millar, R.B. and Anderson, M.J., 2004. Remedies for pseudoreplication. *Fisheries Research*, 70, 397–407.
- Murdoch, A.R., Frady, C.H., Hughes, M.S., and See, K., 2019. Estimating population size and observation bias for spring Chinook Salmon. *Conservation Science and Practice*, 2019;1:e120, 1–12.
- Parken, C.K., Bailey, R.E., and Irvine, J.R., 2004. Incorporating Uncertainty into Area-under-the-Curve and Peak Count Salmon Escapement Estimation. *North American Journal of Fisheries Management*, 23 (1), 78–90.
- Parsons, A.L. and Skalski, J.R., 2010. Quantitative assessment of salmonid escapement techniques. *Reviews in Fisheries Science*, 18 (4), 301–314.
- Patterson, D.A., Michielsens, C.G.J., Latham, S.J., Robinson, K.A., Decker, A.S., Hawkshaw, M.A., and Martens, F.J., in press. Recommendations for Improvements and Use of Stock-Specific Run Size Estimates for Fraser River Sockeye Salmon. *DFO Can. Sci. Advis. Sec. Res. Doc.*
- Pestal, G., Macdonald, B.L., Grant, S.C.H., and Holt, C.A., 2023. State of The Salmon: Rapid status assessment approach for Pacific salmon under Canada's Wild Salmon Policy. *Can. Tech. Rep. Fish. Aquat. Sci.*, 3570, xiv+200p.
- PSC, 2019. *Treaty Between the Government of Canada and the Government of the United States of America Concerning Pacific Salmon*. Pacific Salmon Commission. Canada, USA.
- Quinn, T.P., 1999. Variation in Pacific salmon reproductive behaviour associated with species, sex, and levels of competition, (136), 179–204.

- R Core Team, 2024. R: A language and environment for statistical computing.
- Ricker, W.E., 1975. Computation and Interpretation of Biological Statistics of Fish Populations. *Bulletin of the Fisheries Research Board of Canada*, 191, xviii + 382p.
- Ripley, B.D., 2007. *Pattern recognition and neural networks*. Cambridge University Press.
- Robinson, K.A., Patterson, D.A., Michielsens, C.G.J., Latham, S.J., Decker, A.S., Hawkshaw, M.A., and Martens, F.J., in press. Determination of stock-specific run size estimates for Fraser River Sockeye Salmon. *DFO Can. Sci. Advis. Sec. Res. Doc.*
- Royston, J.P., 1982. Algorithm AS 181: The W Test for Normality. *Applied Statistics*, 31 (2), 176.
- Schindler, D.E., Armstrong, J.B., and Reed, T.E., 2015. The portfolio concept in ecology and evolution. *Frontiers in Ecology and the Environment*, 13 (5), 257–263.
- Schnorbus, M., Bennet, K., and Werner, A., 2010. *Quantifying the Water Resource Impacts of Mountain Pine Beetle and Associated Salvage Harvest Operations across a Range of Watershed Scales: Hydrologic modelling of the Fraser River Basin*. Victoria, British Columbia.
- Schubert, N.D., 2007. Estimating the 1995 Fraser River Sockeye Salmon (*Oncorhynchus nerka*) Escapement. *Can. Tech. Rep. Fish. Aquatic. Sci.*, 2737, ix + 71p.
- Schubert, N.D. and Houtman, R., 2007. Estimating the 1998 Fraser River Sockeye Salmon (*Oncorhynchus nerka*) Escapement, with Special Reference to the Effect of Migration Stress on Estimation Accuracy. *Can. Tech. Rep. Fish. Aquat. Sci.*, 2732, ix + 121p.
- Schwarz, C.J., 2013. SPAS-2 - Version 2014-01-01 User Manual and Sample Analysis.
- Schwarz, C.J. and Taylor, C.G., 1998. Use of the stratified-Petersen estimator in fisheries management: estimating the number of pink salmon (*Oncorhynchus gorbuscha*) spawners in the Fraser River. *Canadian Journal of Fisheries and Aquatic Sciences*, 55 (2), 281–296.
- Seber, G.A.F., 1982. *The estimation of animal abundance*. 2nd ed. Griffin, London: The Blackburn Press.
- Selbie, D.T., Korman, J., Pon, L.B., and Bradford, M.J., 2022. Recovery potential assessment for the endangered Cultus Lake sockeye salmon (*Oncorhynchus nerka*) (2019). *DFO Can. Sci. Advis. Sec. Res. Doc.*, (2022/051), vi + 99.
- Shardlow, T., Hilborn, R., and Lightly, D., 1987. Components Analysis of Instream Escapement Methods for Pacific Salmon (*Oncorhynchus* spp.). *Canadian Journal of Fisheries and Aquatic Sciences*, 44, 10311037.
- Symons, P.E.K. and Waldichuck, M., 1984. Proceedings of the workshop on stream indexing for salmon escapement, West Vancouver, B.C. 2-3 February 1984. *Can. Tech. Rep. Fish. Aquat. Sci.*, xv + 258p.
- Velez-Espino, L.A., Parken, C.K., Clemons, E.R., Peterson, R., Ryding, K., Folkes, M., and Pestal, G., 2019. *ForecastR: tools to automate forecasting procedures for salmonid terminal run and escapement*. Vancouver, BC.

- Venables, W.N. and Ripley, B.D., 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer.
- Welch, P., Benner, K., Decker, S., and Leaf, B., 2017. *Calibration of visual assessment methods for Fraser River Sockeye salmon (Oncorhynchus nerka) - Year 8*. Report prepared by Fisheries and Oceans Canada for the Pacific Salmon Commission, Southern Boundary Restoration and Enhancement Fund, Vancouver, BC.
- Welch, P., Benner, K., and Leaf, B., 2007. *Calibration of Assessment Methods for Fraser River Sockeye Salmon (Oncorhynchus nerka) Spawning Populations (25,000 to 75,000) in the Horsefly, Stellako and Adams River Systems*. Vancouver, BC.
- Welch, P. and Leaf, B., 2019. Calibration of Visual Assessment Methods for Fraser River Sockeye Salmon (*Oncorhynchus nerka*) - Year 10. *Pacific Salmon Commission, Southern Boundary Restoration and Enhancement Fund*, 32.
- Wilson VanVoorhis, C.R. and Morgan, B.L., 2007. Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology*, 3 (2), 43–50.
- Woodey, J., 1984. Escapement estimation in the management of Fraser River sockeye salmon. In: P.E.K. Symonds and M. Waldichuck, eds. *Proceedings of the workshop on stream indexing for salmon escapement estimation, West Vancouver, B.C., 2-3 February 1984*. 121–132.
- Xie, Y. and Martens, F.J., 2014. An Empirical Approach for Estimating the Precision of Hydroacoustic Fish Counts by Systematic Hourly Sampling. *North American Journal of Fisheries Management*, 34 (3), 535–545.
- Yates, L.A., Aandahl, Z., Richards, S.A., and Brook, B.W., 2023. Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, 93.
- Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M., 2009. *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

APPENDIX A. DATA, CODE, AND SUMMARY TABLES

Data and code for model development, diagnostics, selection, application, and comparison are available on the Git Hub repository:

[Fraser-River-Sockeye-Expansion-Factors-Tech-Report-Supplement](#)

Table 4. Sample sizes and date ranges of systems and their Conservation Units.

Conservation Unit	Stream Name	Sample Size	Earliest Record	Recent Record
Bowron-ES	Bowron River	1	1995	1995
Chilliwack-ES	Chilliwack River (Upper)	3	2012	2020
Francois-Fraser-ES	Nadina River	5	2016	2022
Francois-Fraser-S	Stellako River	35	1989	2022
Harrison (U/S)-L	Weaver Creek	1	1996	1996
Harrison River (River-Type)	Harrison River	10	2009	2022
Kamloops-ES	Raft River	1	2000	2000
Lillooet-Harrison-L	Birkenhead River	3	2013	2016
North Barriere - ES	Barriere River (Upper)	4	1988	1996
Quesnel-S	Mitchell River	3	1994	2010
Quesnel-S	Horsefly River	2	2007	2010
Shuswap Complex-L	Adams River (Lower)	10	1994	2022
Shuswap Complex-L	Little River	4	2006	2022
Shuswap Complex-L	Eagle River	1	2018	2018
Shuswap-ES	Seymour River	7	1994	2018
Shuswap-ES	Eagle River	2	1998	2014
Shuswap-ES	Adams River (Upper)	1	2000	2000
Shuswap-ES	Scotch Creek	1	2018	2018
Takla-Trembleur-Estu	Gluske Creek	18	1990	2009
Takla-Trembleur-Estu	Forfar Creek	17	1990	2009
Takla-Trembleur-Estu	O' Ne-Ell Creek	13	1991	2004
Takla-Trembleur-Estu	Dust Creek	2	2001	2002
Takla-Trembleur-Estu	Crow Creek	1	1996	1996
Takla-Trembleur-Estu	Narrows Creek	1	1996	1996
Takla-Trembleur-Estu	Paula Creek	1	1996	1996
Takla-Trembleur-Stuart-S	Tachie River	12	1994	2021
Takla-Trembleur-Stuart-S	Middle River	2	1997	2005
Takla-Trembleur-Stuart-S	Kuzkwa River	1	2005	2005

Table 5. Sample sizes and 10% buffered ranges of visual-count data for size-clarity-substrate shade-method sets. Note that the minimum is a 10% buffer lower and the maximum is a 10% buffer higher.

Size	Water Clarity	Method	Substrate Shade	n	Median	Minimum	Maximum
Small	Clear	Aerial	Medium	1	22,677	22,677	22,677
Small	Clear	Ground	Light	18	1,621	79	8,726
Small	Clear	Ground	Medium	38	3,686	372	23,681
Small	Turbid Tannic	Ground	Medium	2	5,964	728	11,200
Medium	Clear	Aerial	Light	3	38,405	18,777	83,495
Medium	Clear	Aerial	Medium	22	36,250	6,267	131,349
Medium	Clear	Ground	Medium	25	44,997	20,874	91,897
Medium	Turbid Tannic	Aerial	Dark	5	6,333	2,477	77,872
Medium	Turbid Tannic	Aerial	Light	7	16,893	3,761	66,368
Medium	Turbid Tannic	Aerial	Medium	1	10,493	10,493	10,493
Medium	Turbid Tannic	Ground	Medium	2	15,633	5,399	25,866
Large	Clear	Aerial	Medium	3	16,050	14,860	43,660
Large	Clear	Ground	Medium	7	105,654	19,405	314,148
X-Large	Clear	Aerial	Light	7	5,536	3,552	67,090
X-Large	Clear	Aerial	Medium	3	51,480	16,274	175,000
X-Large	Clear	Ground	Medium	1	53,100	53,100	53,100
X-Large	Turbid Tannic	Aerial	Dark	12	27,797	6,943	243,450
X-Large	Turbid Tannic	Aerial	Light	3	16,600	11,012	113,905
X-Large	Turbid Tannic	Aerial	Medium	2	62,408	48,315	76,500

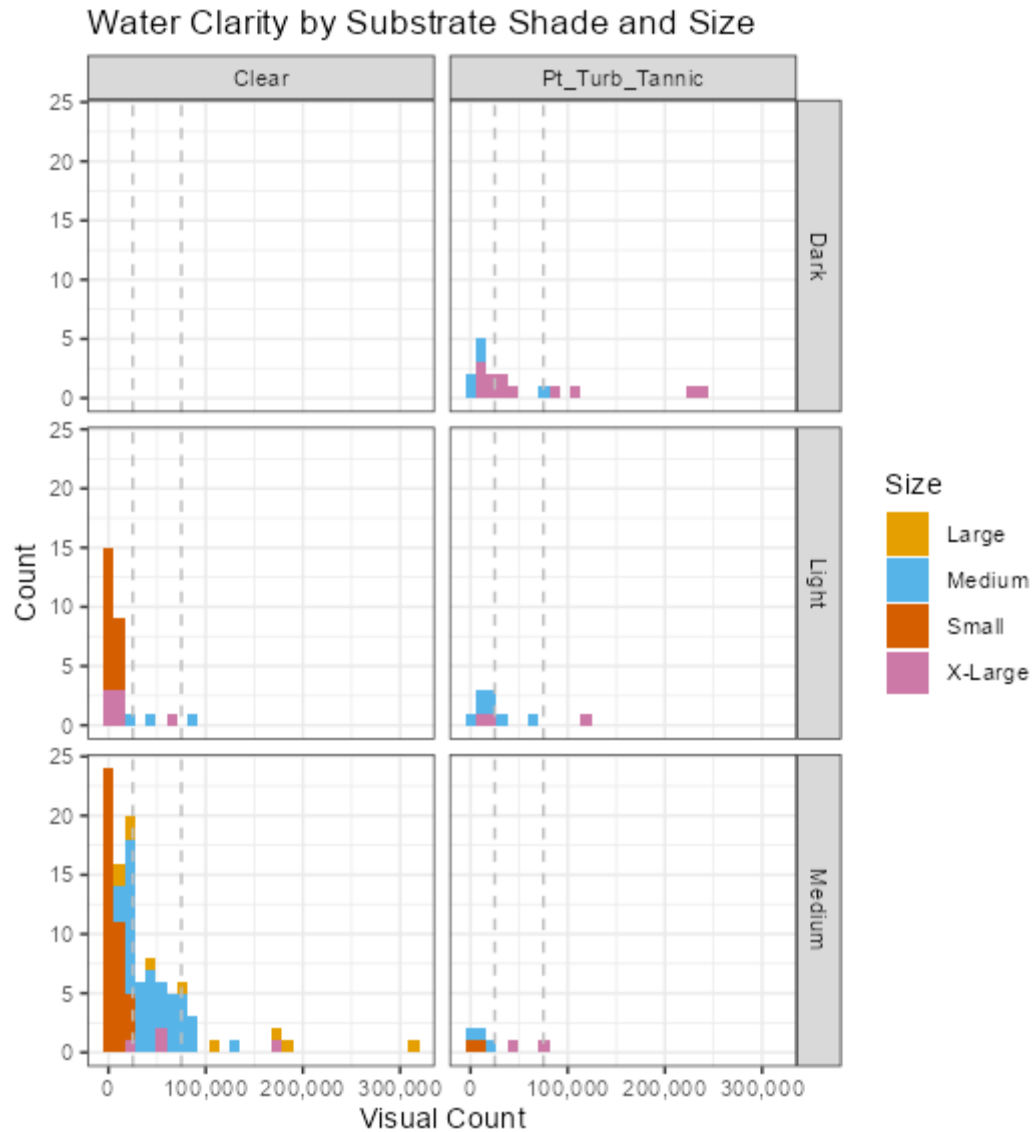


Figure 10. Histogram showing the distribution of visual counts by their characteristics. Columns are water clarity, rows are substrate shade, and colour is size (large – yellow, medium – blue, small – orange, x-large pink). Dashed vertical lines indicate 25,000 and 75,000, respectively, in each figure.

APPENDIX B. MODEL SPECIFICATIONS, DIAGNOSTICS, AND PARAMETER RESULTS

Each model has a base framework, set of equations, and assumptions. Next, varying sets (i.e. subsets) of data-pairs are identified to fit the models, depending on the set-level or covariates of interest. Models then require investigating their fit to the data as well as covariate or model selection from an initial global model. Outlying data points are highlighted but are kept in the sets during parameter estimation to acknowledge their contribution to the uncertainty of expansion methods.

MEAN-OF-RATIOS

The mean-of-ratios method applied here is described in McPherson et al. (2000; Table 3). First, individual ratios are estimated for each system-year data pair in the database:

(5)

$$\hat{\pi}_i = \hat{N}_i C_i^{-1}$$

Where $\hat{\pi}_i$ is the estimated expansion factor for system-year i , $\hat{N}_i$ is the high-accuracy estimate for system-year i , and C_i is the visual count for system-year i . The approximate variance for an expansion factor is:

(6)

$$v(\hat{\pi}_i) = v(\hat{N}_i) C_i^{-2}$$

The estimated variance, v , of an individual expansion factor is based on the variance of the high-accuracy estimate, $\hat{N}_i$, and the visual count, C_i , whose error is unknown and unable to be incorporated. Next, the mean expansion factor, $\bar{\pi}_s$, is calculated based on a set, s , of system-years that contain a range of sample sizes, k , from more specialized groups to the generalized whole database:

(7)

$$\bar{\pi}_s = \frac{\sum_{i=1}^k \hat{\pi}_i}{k}$$

The estimated variance of an expansion factor for a given set combines the variance across the individual expansion factors and the sum of their individual variances.

(8)

$$v(\bar{\pi}_s) = \frac{\sum_{i=1}^k (\hat{\pi}_i - \bar{\pi}_s)^2}{k - 1} + \frac{\sum_{i=1}^k v(\hat{\pi}_i)}{k}$$

It is assumed that this is a minimum estimate of variance (biased low) because the unknown observer error in the visual counts is not accounted for. The expanded estimate for the specialized set, $\widehat{EN}_{si}$, is then simply the product of the visual count for the system-year within that set, C_{si} , and the average set expansion factor, $\bar{\pi}_s$:

(9)

$$\widehat{EN}_{si} = C_{si} \bar{\pi}_s$$

The variance of the expanded estimate is then dependent on the set's average expansion factor's variance:

(10)

$$v(\widehat{EN}_{si}) = C_{si}^2 v(\bar{\pi}_s)$$

The mean-of-ratios estimator method is the only method Parsons and Skalski (2010) ranked as “poor” out of the methods tested here. Therefore, to reduce the final total of models available for the model selection step, originally, only two levels of specialization were selected for model development. The Size-Clarity (4) specialized level was selected due to prior unpublished reports investigating at this level and preliminary analysis identifying that there were visible contrasts between these levels. River size and water clarity are also commonly reported by field observers as one of the primary drivers of a system's counting difficulty. Individual (5) level models were also updated for the three Early Stuart systems that had pre-established expansion factors that used the same ratio-of-means method in their development (2.4.7 Prior Specialized Peak Live Expansions). Two combined Early Stuart mean-of-ratios estimates were also made for exemplary purposes, one that included the three highly sampled systems (Gluske, Forfar, and O'Ne-ell) and one that also included all other small-clear systems. Finally, an irregular data subset was made for low abundance, out-of-sample, visual counts. The low abundance subset only included data from *small* and *medium* sized streams with visual counts less than 25,000. *Small* and *medium* were selected because they were deemed to be the most representative of most systems that would have a low abundance counted as well as conserved the expected contrast between *clear* and *turbid/tannic* observer efficiency.

Sample size (k) and CV were used as filters to limit the number of Size-Clarity models to those with relatively robust results. Larger sample sizes should reduce the effects of outliers and the CV should identify the general uniformity of the ratios in a given set. For a Size-Clarity model to make it to the candidate model ranking stage, the CV must be < 50% and the sample size have a minimum of 20. These rules were applied as general rules of thumb to indicate that the expansion factor was relatively stable and would not be overly affected by outlying data-pairs. The rules were not applied to the Early Stuart Individual (5) set-level models because they are from systems with inherently lower variability in their survey conditions. The rules were also not applied to the low abundance subset because the model selection process for out-of-sample data was more driven by expert decisions than the algorithm process. Outlying expansion factors within a set were defined as those that were 1.5 times greater or less than the 75th and 25th quartiles, respectively.

There were five Size-Clarity sets that had outlying expansion factor estimates. The small-clear ground surveyed expansion factors had two upper outliers ($\hat{\pi} > 2.50$): Barriere River (Upper) in 1995 ($\hat{\pi} = 2.59$) and O'Ne-Ell Creek in 1992 ($\hat{\pi} = 2.51$). The small-clear ground surveyed expansion factors had the smallest CV and a relatively small interquartile range, which is one statistical reason that these expansion factors lay outside of the outlier cutoff by a small difference (<0.1). Most other set's outliers were much greater or smaller than the outlier cut-off (difference > 0.5). The medium-clear ground surveyed expansion factors had one upper outlier ($\hat{\pi} > 3.29$): Stellako River in 2010 ($\hat{\pi} = 4.20$). The large-clear ground surveyed expansion

factors had one upper outlier ($\hat{\pi} > 5.93$): Adams River (Lower) in 2011 ($\hat{\pi} = 7.57$). The medium-turbid/tannic aerial surveyed expansion factors had one upper outlier ($\hat{\pi} > 5.08$) and one lower outlier ($\hat{\pi} < 1.69$): Eagle River in 1998 ($\hat{\pi} = 7.57$) and Nadina River in 2020 ($\hat{\pi} = 1.40$), respectively. The x-large-turbid/tannic aerial surveyed expansion factors had two upper outliers ($\hat{\pi} > 7.67$): Tachie River in 2013 ($\hat{\pi} = 8.92$) and Harrison River in 2015 ($\hat{\pi} = 10.51$). There were no unifying features such as year/observers or environmental records that were shared between all outliers. Out of the eight outliers, four noted the use of only one observer's counts in the visual method (O'Ne-Ell Creek 1992, Stellako River 2010, Adams River (Lower) 2011, and Nadina River 2020). A single observer could indicate either the survey was being used to train the second counter or only one counter was present. Relying on only one counter may increase the potential for miscounts, transcription errors, or other errors. Unfortunately, the number of observers is not always available in historical records, nor is the observer's experience. Additionally, some non-outlying expansion factor estimates also only used a single observer's counts. Harrison River 2015 had a notable field comment of "species distinguishing was challenging;" however, two other years of Harrison River aerial surveys also commented on difficulty due to species overlaps. There were no consistent reasons for the outliers across sets.

The mean of ratio estimates differed across the Size-Clarity sets but only three passed the robustness filters (Figure 11, Table 6). Of the ground survey methods, only small-clear and medium-clear systems had greater than 20 samples and were added to the candidate model set. Small-clear and medium-clear ground surveyed expansion factors ($\bar{\pi} = 1.70$ and 1.99, respectively) were the closest to the original default 1.8 expansion factor estimate. Small-clear had the largest number of samples and smallest CV (21.0%). The sample sizes for ground surveyed small and medium-turbid/tannic sets were both only two each, but their average was in the expected direction of requiring a larger expansion factor, indicating lower observer efficiency. Of the aerial survey method sets, only medium-clear had greater than 20 samples and passed the CV threshold (CV = 31.2%). The aerially surveyed medium turbid/tannic set sample size was only 13, but its average expansion factor was also in the expected direction of being greater than its clear counterpart. Of the size sets with samples in both clear and turbid/tannic, only the aerially surveyed X-large set did not follow the emerging pattern of turbid/tannic systems requiring a larger expansion factor. The aerially surveyed X-large sets also had the most variability in expansion factors and largest CV, indicating that the simple separation by Size-Clarity was not enough to account for variability in observer efficiency in X-large systems. Overall, there was an emergent pattern of requiring relatively larger expansion factors in sets that were in turbid/tannic water within a size-method, but small samples sizes and low precision limited the applicability of the mean-of-ratios models developed with only three Size-Clarity models being added to the candidate set.

The three Early Stuart Individual (5) mean-of-ratios estimates differed slightly from the previously estimated expansion factors (Table 6). Forfar Creek expansion was 0.01 lower, Gluske Creek was 0.04 higher, and O'Ne-Ell Creek was 0.02 higher than the previous estimates. All of their CVs were below 30%, with Forfar Creek and Gluske Creeks CVs being below 20%. The Early Stuart average expansion factors were both 1.68, which is 0.03 higher than the historical 1.65 averaged expansion factor. Only the Individual (5) models were added to the candidate model set.

Table 6. Expansion factors (π) for Size-Clarity (4) sets by Method for Fraser River Sockeye estimated using the Mean-of-Ratios method. Early Stuart Individual (5) sets are also presented at the bottom of the table. CV is the Coefficient of variation of the expansion factor and k is the sample size.

Method	Size or System	Water Clarity	k	Average Abundance	π	Min	Max	Standard Deviation	CV
Aerial	Small	Clear	1	81,169	3.58	-	-	-	-
Aerial	Small	Turbid/Tannic	0	-	-	-	-	-	-
Aerial	Medium	Clear	25	87,586	2.31	0.99	3.78	0.72	31.2
Aerial	Medium	Turbid/Tannic	13	74,792	3.55	1.40	7.57	1.65	46.6
Aerial	Large	Clear	3	182,364	7.06	3.28	9.97	3.49	49.4
Aerial	Large	Turbid/Tannic	0	-	-	-	-	-	-
Aerial	X-Large	Clear	10	121,683	7.10	2.38	15.83	4.89	68.8
Aerial	X-Large	Turbid/Tannic	17	164,673	3.81	1.50	10.51	2.69	70.6
Ground	Small	Clear	56	8,363	1.70	1.07	2.59	0.36	21.0
Ground	Small	Turbid/Tannic	2	12,710	2.67	2.06	3.28	0.86	32.4
Ground	Medium	Clear	25	94,483	1.99	1.11	4.2	0.71	35.5
Ground	Medium	Turbid/Tannic	2	35,306	2.42	2.17	2.67	0.74	30.4
Ground	Large	Clear	7	354,543	3.62	2.17	7.57	1.98	54.6
Ground	Large	Turbid/Tannic	0	-	-	-	-	-	-
Ground	X-Large	Clear	1	127,386	2.40	-	-	-	-
Ground	X-Large	Turbid/Tannic	0	-	-	-	-	-	-
Early Stuart Specific (5)									
Ground	Forfar Creek		17	7,099	1.68	1.19	2.4	0.32	18.9
Ground	Gluske Creek		18	5,301	1.63	1.07	2.11	0.28	16.9
Ground	O' Ne-Ell Creek		13	9,937	1.73	1.14	2.51	0.46	26.6
Ground	Forfar, Gluske, O'Ne-Ell		48	7,193	1.68	1.07	2.51	0.34	20.5
Ground	All Early Stuart - Clear		51	6,935	1.68	1.07	2.51	0.35	20.8
Low Abundance Specific									
Aerial	Clear		11	42,672	2.53	1.22	3.78	0.88	34.7
Aerial	Turbid/Tannic		9	28,731	3.61	1.4	7.57	1.99	55.1
Ground	Clear		62	11,430	1.71	1.07	2.59	0.37	21.5
Ground	Turbid/Tannic		3	13,280	2.67	2.06	3.28	0.8	30

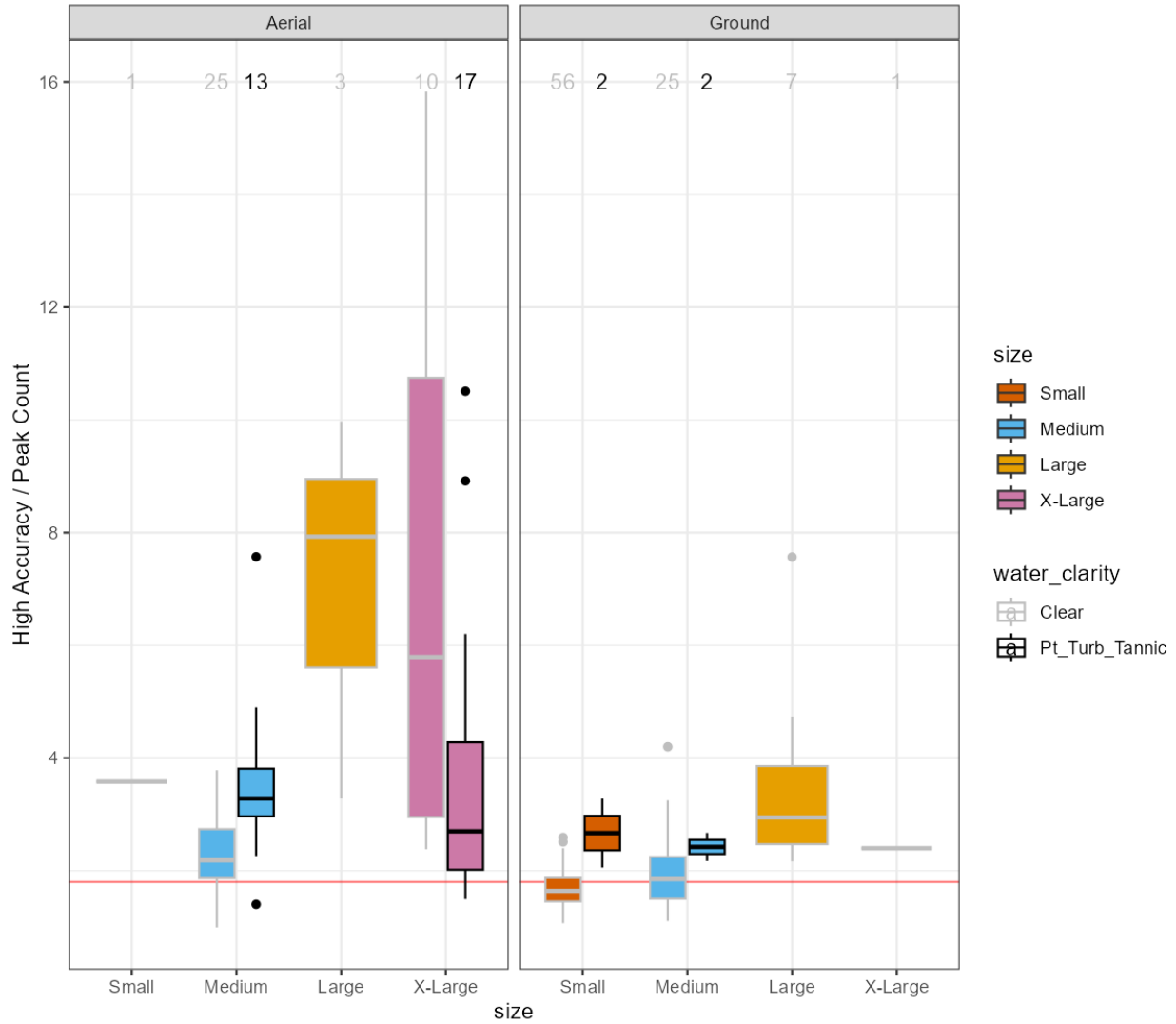


Figure 11. Boxplots of the ratio of high-accuracy estimates relative to visually based peak counts from Aerial (left panel) or Ground (right panel) survey methods for small (burnt-orange), medium (blue), large (dark-yellow), and x-large (pink) systems separated by if the water clarity was clear (grey outline) or turbid/tannic (black outline). The horizontal red line is 1.8, the old default expansion factor. The numbers at the top indicate the sample sizes. The hinges of the boxes display the 25-75 percentile range with the median denoted between them. The whiskers indicate 1.5 times the interquartile range above or below the hinges, whereas dots are ratios that extend that range (outliers).

RATIO-OF-MEANS

The ratio-of-means method applied here is described in Cochran (1977) and Eberhart and Simmons (1987). The expansion factor is defined as (sharing similar notation as in the Mean-of-Ratios section):

(11)

$$\hat{R}_s = \sum_i^k \hat{N}_i / \sum_i^k C_i$$

The sum of the numerator and denominator for a given set are taken to estimate the ratio ($\hat{R}_s$). This is the same as taking the means, but since both numerator and denominator would be divided by the same sample size (k), they cancel out.

Bias may be introduced into ratio-of-means estimates when the numerator and denominator of the ratio are not normally distributed (Cochran 1977). The distribution of high-accuracy estimates (numerator) and visual counts (denominator) was tested using a Shapiro-Wilk test (shapiro.test in R; R core Team 2024; Shapiro and Wilk 1965; Royston 1982). The natural logarithm transformation resulted in the closest to normal for each of the distributions. Therefore, the equation actually applied was:

(12)

$$\hat{R}_s = \sum_i^k \hat{x}_i / \sum_i^k y_i$$

Where $\hat{x}_i = \ln(\hat{N}_i)$ and $y_i = \ln(C_i)$. The natural log of the expanded estimate for the set, $\ln(\widehat{EN}_{si})$, is then simply the product of the visual count for the specific system-year within that set, C_{si} , and the set-specific expansion factor, $\hat{R}_s$:

(13)

$$\ln(\widehat{EN}_{si}) = y_{si} \hat{R}_s \quad \text{or} \\ \widehat{EN}_{si} = e^{y_{si} \hat{R}_s}$$

The approximate variance of the expansion factor used the Taylor Series approximation, or “Delta-Method”, for estimating the variance of a ratio described in Elandt-Johnson and Johnson (1980):

(14)

$$v(\hat{R}_s) = \frac{v(\bar{x}_s)}{\bar{y}_s^2} + \frac{2\bar{x}_s\bar{y}_s \text{Cov}(\hat{x}, y)}{\bar{y}_s^3} + \frac{\bar{x}_s^2 \text{var}(\bar{y}_s)}{\bar{y}_s^4}$$

Where $\bar{x}_s$ and $\bar{y}_s$ are the means of $\hat{x}_{i,...,k}$ and $y_{i,...,k}$ within a set (s), respectively. The Taylor Series approximation was used over the variance estimators described in Cochran (1977) because there were not easily definable values for their N or n' (equations 6.13 or 12.72, Cochran 1977) in this data set. The estimated variance of the mean of the natural logarithm of the high-accuracy estimates, $\text{var}(\bar{x}_s)$, combines the variance across the individual expansion factors and the sum of their individual variances; similarly to equation (8) of the mean-of-ratios estimate.

(15)

$$v(\bar{x}_s) = \frac{\sum_{i=1}^k (\hat{x}_i - \bar{x}_s)^2}{k - 1} + \frac{\sum_{i=1}^k v(\hat{N}_i) / \hat{N}_i^2}{k^2}$$

The variance of the expanded estimate within a set was then:

$$v(\ln(\widehat{EN}_{si})) = y_{si}^2 v(\widehat{R}_s)$$

The CVs were estimated in the predictor space of the fit (natural-logarithm) as well as the error intervals before transforming back to normal space.

The ratio-of-means estimator was ranked as “fair” by Parsons and Skalski (2010) and the variance estimate as “poor”, but generally as more robust and less biased than the means-of-ratios method. Cochran (1977) had several recommendations around creating more robust estimates, including large sample sizes ($k > 30$) and that the CVs of the individual numerator and denominator means were $< 10\%$. Large sample sizes at most levels of specialization would preclude many sets from having a ratio-of-means estimator applied; therefore the rule-of-thumb for sample size was relaxed to $k > 20$. Transforming the numerator and denominator data with the natural-log generally reduced their CVs, but this rule was not applied in excluding which ratios would be included in the candidate model ranking set.

Four levels of specialization were selected for ratios-of-means model set development; of which, twelve expansion factors were added to the candidate model selection set (Table 7). The specialization levels used to create sets were the Size-Clarity (4), Size (3), Size-Aggregate (2), and Generalized (1). The ground surveyed Large-X-Large size-aggregate set ($k=8$) and any more specialized set of those two size categories had fewer than 20 samples. Within the Size-Clarity-level sets, no sets with a turbid/tannic clarity had greater than 20 samples.

Table 7. Ratio-of-means expansion factors and their errors by method, set, size, and water clarity. Only sample sizes greater than 3 are presented here. CV is the Coefficient of variation of the average expanded estimate for a given set. Note that the expansion factors are applied in a logarithmic space.

Method	Set	Size	Water Clarity	Expansion Factor	Sample Size	Standard Error	CV
Both	Generalized (1)	-	-	1.089	162	0.056	5.6
Aerial	Generalized (1)	All	-	1.113	69	0.065	6.5
Ground	Generalized (1)	All	-	1.068	93	0.035	3.5
Aerial	Size-Aggregate (2)	Large-X-Large	-	1.139	30	0.081	8.1
Ground	Size-Aggregate (2)	Large-X-Large	-	1.102	8	0.044	4.4
Aerial	Size-Aggregate (2)	Small-Medium	-	1.093	39	0.043	4.3
Ground	Size-Aggregate (2)	Small-Medium	-	1.064	85	0.031	3.1
Ground	Size (3)	Large	-	1.105	7	0.047	4.7
Aerial	Size (3)	Medium	-	1.092	38	0.043	4.3
Ground	Size (3)	Medium	-	1.062	27	0.031	3.1
Ground	Size (3)	Small	-	1.065	58	0.031	3.1
Aerial	Size (3)	X-Large	-	1.134	27	0.081	8.1
Ground	Size-Clarity (4)	Large	Clear	1.105	7	0.047	4.7
Aerial	Size-Clarity (4)	Medium	Clear	1.076	25	0.034	3.4
Aerial	Size-Clarity (4)	Medium	Turbid/Tannic	1.126	13	0.044	4.4
Ground	Size-Clarity (4)	Medium	Clear	1.060	25	0.03	3
Ground	Size-Clarity (4)	Small	Clear	1.063	56	0.029	2.9
Aerial	Size-Clarity (4)	X-Large	Clear	1.181	10	0.096	9.6
Aerial	Size-Clarity (4)	X-Large	Turbid/Tannic	1.110	17	0.065	6.5

INVERSE REGRESSION

The inverse regression method applied here is described in Kleiven et al. (2023). Similarly to the ratio-of-means method, the high-accuracy estimate and visual count were transformed with a natural logarithm where $\hat{x}_i = \ln(\hat{N}_i)$ and $y_i = \ln(C_i)$, respectively. Transforming the data prior to applying the linear regression was done because the current analytical tools available did not support generalized linear model formations. The general inverse regression formula follows a simple linear regression:

(17)

$$y_s = \beta_{0,s} + \beta_{1,s}\hat{x}_s + \varepsilon_s$$

$$\varepsilon_s = N(0, \sigma^2)$$

Where the set, s , determines the set for which the data, intercept ($\beta_{0,s}$), slope ($\beta_{1,s}$), and error (ε_s) belong to. The error is assumed to be normally distributed with homogenous variance (σ^2). The natural logarithmic transformation provided a better fit for the assumption that the error was normally distributed and homogenous. From the simple linear regression parameters, an expanded estimate is derived by solving for $\hat{x}_s$:

(18)

$$\ln(\widehat{EN}_{si}) = \frac{y_{i,s} + \beta_{0,s}}{\beta_{1,s}} \quad \text{or}$$

$$\widehat{EN}_{si} = e^{\frac{y_{i,s} + \beta_{0,s}}{\beta_{1,s}}}$$

The R-package *investr* (Greenwell and Shubert Kabban 2014) was used to estimate the Wald 95% prediction interval and standard error (see code, *01-03-Inverse-Regression.Rmd*).

The inverse regression estimator was ranked as “good” by Parsons and Skalski (2010) and the variance estimate as “good”, but generally has not been applied in many cases. The benefit of the inverse regression estimator is that its assumptions are more readily met than a regular regression method and should more accurately represent the magnitude of the variance of an estimate. However, due to its limited prior application and descriptions, the complexity of inverse regression methods used here was limited to simple formulations that included no covariates.

Instead of covariates, the reliance of different sets of data were used to explore if there were differences between survey method and size-aggregates. Models developed were at the Size-Aggregate Level 2 or at the Individual Level 5. The Size-Aggregate sets were broken out by method but a combined model was also tested to be used as a null comparison for a total of six level two models. Individual models were all built within a method. Only Individual systems with greater than five samples were explored, resulting in eight Individual (5) models. Therefore, a total of 14 inverse regression models were developed.

The coefficients of determination (R^2) and response residuals were investigated to determine if a model would be added to the candidate model set. All models had R^2 greater than 60. Two had R^2 less than 70: Large-X-Large Aerial and Harrison River Aerial. Three had R^2 less than 80 (but greater than 70): Stellako River Aerial, Stellako River Ground, and Tachie River Aerial. Of the Size-Aggregate Level 2 models, the Large-X-Large Ground residuals were the most

heterogenous and its generally small sample size ($k=8$) indicated that the fit may not be very robust. The Individual (5) model residuals were generally not homogeneous with maybe the exceptions of Forfar, Gluske, and O'Ne-Ell ground methods (the Early Stuart systems). Gluske did have one relatively small fitted value that had the potential to have high leverage. Adams River (Lower) and Harrison River had the smallest sample sizes ($k<11$) and most evident patterns in their residuals. Overall, models with sample sizes less than 10 had the worst fits; therefore, only inverse regression models with sample sizes greater than 10 will be added to the candidate model set at this time.

Table 8. Inverse Regression parameter estimates and coefficient of variation (R^2) by method, set, and size. Only models with samples sizes (k) greater than 5 are presented here.

Method	Set	Size or System	k	Intercept	Int. Error	Slope	Slp. Error	R^2
Both	Size-Aggregate (2)	Large-X-Large	38	-3.02	1.33	1.14	0.11	73.91
Aerial	Size-Aggregate (2)	Large-X-Large	30	-2.77	1.69	1.12	0.15	67.65
Ground	Size-Aggregate (2)	Large-X-Large	8	-2.37	2.51	1.1	0.2	83.15
Both	Size-Aggregate (2)	Small-Medium	124	0.07	0.18	0.92	0.02	95.37
Aerial	Size-Aggregate (2)	Small-Medium	39	-0.88	0.81	0.99	0.07	83.16
Ground	Size-Aggregate (2)	Small-Medium	85	-0.22	0.16	0.96	0.02	97.46
Ground	Individual (5)	Adams River (Lower)	7	-3.14	2.75	1.16	0.22	84.77
Ground	Individual (5)	Forfar Creek	17	-0.53	0.45	1	0.05	96.11
Ground	Individual (5)	Gluske Creek	18	-0.6	0.29	1.02	0.04	98.02
Aerial	Individual (5)	Harrison River	10	-2.46	3.03	1.06	0.27	65.76
Ground	Individual (5)	O' Ne-Ell Creek	13	-0.15	0.72	0.96	0.08	92.63
Aerial	Individual (5)	Stellako River	14	0.3	1.75	0.91	0.15	74.37
Ground	Individual (5)	Stellako River	21	1.95	1.26	0.77	0.11	71.79
Aerial	Individual (5)	Tachie River	12	-2.78	2.22	1.14	0.19	78.3

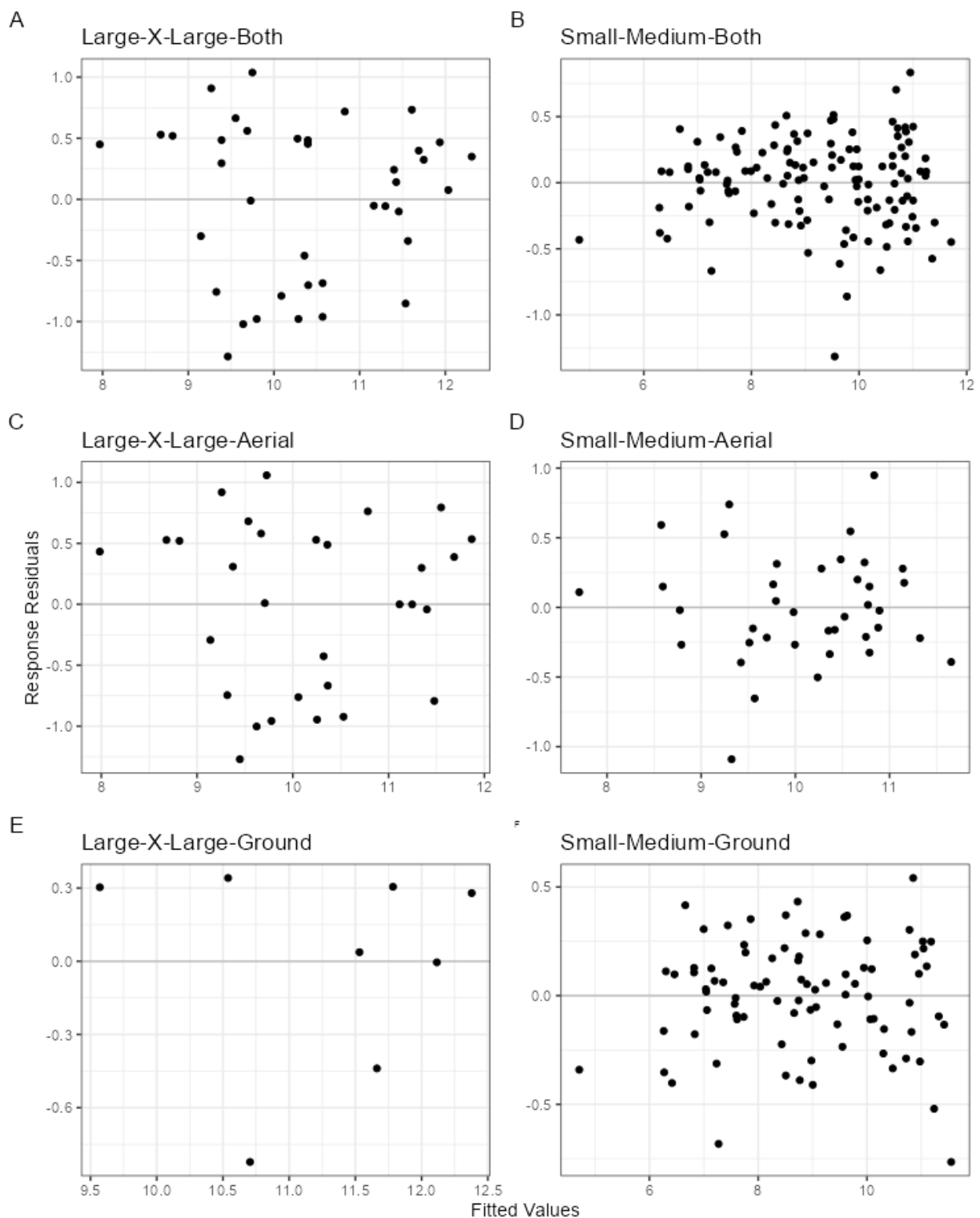


Figure 12. Response residuals compared to the fitted values of the inverse regression models for sets in the Size-Aggregate Level (2); Large-X-Large Both methods (A), Small-Medium Both methods (B), Large-X-Large Aerial method (C), Small-Medium Aerial method (D), Large-X-Large Ground method (E), and Small-Medium Ground method (F). The grey horizontal line indicates 0.

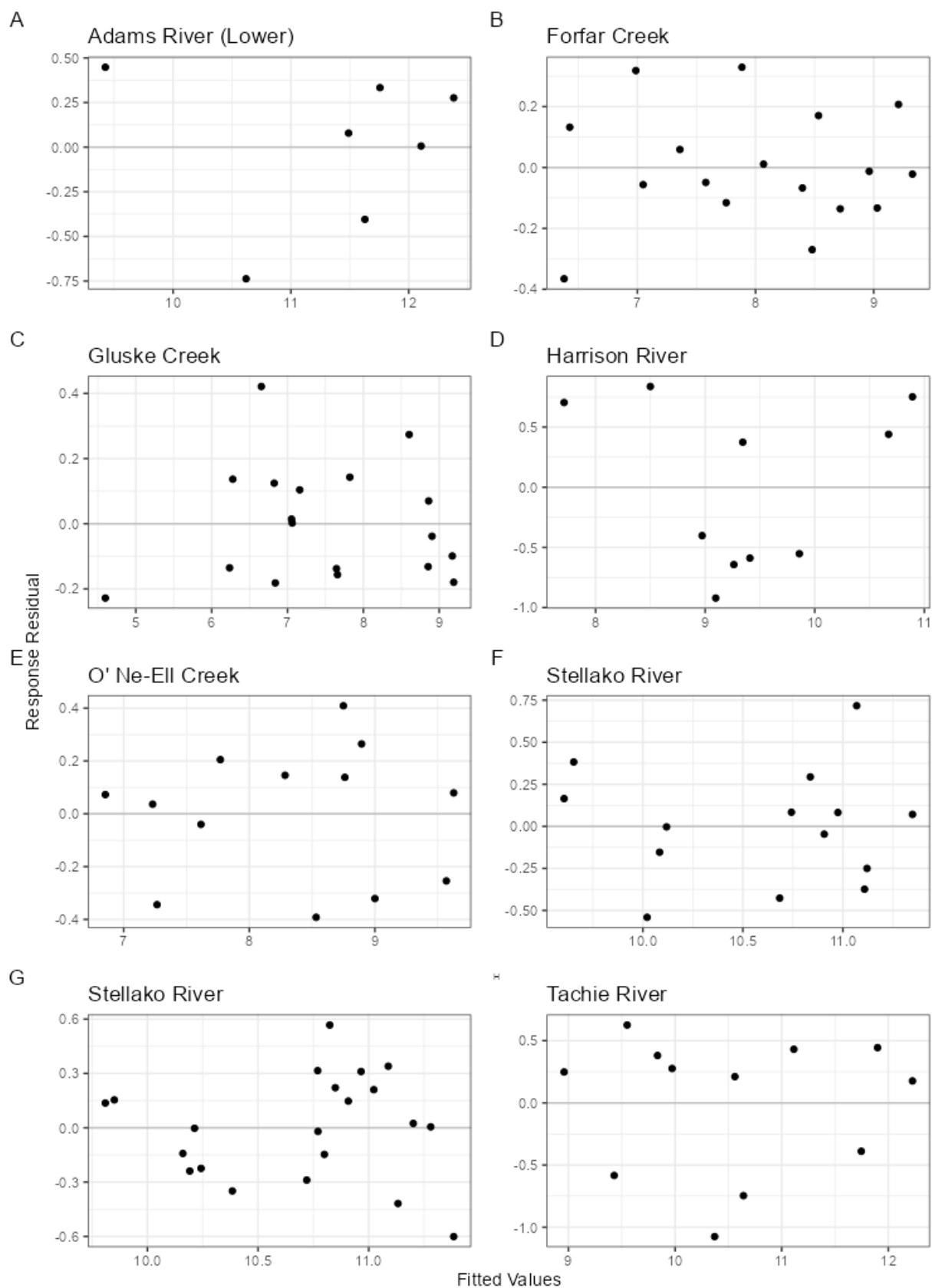


Figure 13. Response residuals compared to the fitted values of the inverse regression models for sets in the Individual Level (5); Adams River (lower) Ground method (A), Forfar Creek Ground method (B), Gluske Creek Ground method (C), Harrison River Aerial method (D), O'Ne-El Creek Ground method (E), Stellako River Aerial method (F), Stellako River Ground method (G), and Tachie River Aerial method (H).

LINEAR AND GENERALIZED LINEAR REGRESSION

Some linear and generalized linear regression model assumptions may not be met in this use, but the versatility of regression models made them a valuable tool to investigate. Simple linear regression assumptions are of: i) normality and ii) homoscedasticity of residuals; iii) fixed explanatory variables; iv) independence between observations; and v) correct model specifications (Zuur *et al.* 2009). Generalized linear approaches are often designed to deal with violations in assumptions i) and ii) by applying transformations through a predictor function family and link, and their assumptions can be verified by checking residuals. The visual count is *not* a fixed explanatory variable because it is known to be inaccurate, violating assumption iii); however, the bias in the visual count is assumed to be in the same direction for all observers and the variability between and within observers is assumed to be normally distributed around a central tendency. Given these considerations, the violation in assumption iii) is not thought to negatively affect average parameter estimation but uncertainty may be underrepresented. Lack of independence was discussed in section 2.2.1, but it should be of note that systems with larger sample sizes in the database will have a larger effect on parameter estimation than systems with smaller sample sizes. The versatility of regression models allows for many specifications and comparisons between specifications; which required a model development and selection framework.

Model development and selection for inclusion in the candidate set followed several steps. The first step was identification of the set level being investigated and if the data was being subset. Next, the family of the error distribution of the response data was considered. In most cases, since the response variable is a count, the Poisson and Negative Binomial families with a log link were investigated. Simple models were fit with each error family using just the visual count as an explanatory variable. The residual deviance was compared against the degrees of freedom within each model and the models were compared against each other with AICc. Residual deviance ratios should be close to 1 and the lower AICc indicates the better fit model (Zuur *et al.* 2009). Once an error family was selected, covariates were selected. Covariates were first filtered based on those available for all data in the subset. Numerical data, such as visual counts, were centered and standardized by one standard deviation. Next, correlation coefficients were compared to identify covariates with high collinearity or multiple collinearity. High collinearity here is defined as a Pearson's correlation coefficient greater than 0.8 but moderately high flags were also placed on pairs greater than 0.6. A Full Global model was then fit with all covariates and no interaction terms and the variance inflation factors (VIFs) were compared across covariates where a VIF greater than 5 was a flag. Covariates were then reduced by excluding those with high or the highest VIF or most instances of collinearity.

Models were then developed using all of or some of the covariates that were screened in the prior step and often included interaction terms, which are an important consideration of these model's logic. Analogous to the ratio-of-means, the visual count is hypothesized to have the largest effect in determining the expanded estimate. The visual count parameter is akin to the ratio-of-means expansion factor; however, it will differ due to the additional estimation of an intercept and other factors like the error family or link functions. Most of the other covariates in this dataset are categorical; which would typically only change the intercept when paired with

their own parameters. Only changing the intercept in this regression formation is the same as adding or subtracting a deterministic number of fish based on conditions; which is not biologically logical. The more likely mechanism is that the overall expansion factor is changed because the observer efficiency has been impacted by a certain set of conditions. Specifying covariates as having an interaction with the visual count will effectively act as if the expansion factor is being modified. This biological reasoning is why more complex models were selected over more parsimonious models during model development in the subsequent steps.

Models made using the reduced set of covariates were compared using AICc (Akaike 1973, Burnham and Anderson 2004). Lower AICc values indicate a better model fit. If a model could not be fit with all covariates having an interaction with visual count (e.g. due to a singularity/non-convergence), then multiple models would be created using further subsets of covariates and their interaction terms. The model with the lowest AICc of various combinations of covariates was then dredged using the dredge function in R (Bartoń 2024) to quickly produce all iterations of the model parameterization. A more parsimonious model (i.e. one with fewer covariates included) was only selected if the largest model (the one with all covariates and interaction terms included) was 2 AICc units greater than the lowest scoring model (“top” model). Within 2 AICc units of the top model, models that included interactions were preferred because the inclusion of the interactions aligned with the mechanistic reasoning of the models. After completing each of those steps, then a model would be selected for inclusion in the candidate model set.

During model development, all models were selected to use a negative binomial error distribution (Gelman and Hill 2007, Zuur *et al.* 2009) and were fit with the glm.nb function in R (Venables and Ripley 2002) following the formation:

(19)

$$\begin{aligned}\hat{N}_i &\sim NB(\mu_i, \kappa) \\ E(\hat{N}_i) &= \widehat{EN}_i = \mu_i \quad \text{and} \quad var(\hat{N}_i) = \mu_i + \frac{\mu_i^2}{\kappa} \\ \mu_i &= e^{\eta(X_{il}, \dots, X_{iq})} \quad \text{or} \quad \ln(\mu_i) = \eta(X_{il}, \dots, X_{iq})\end{aligned}$$

Where the high-accuracy estimates ($\hat{N}_i$) are fit with a negative binomial distribution where the expected expanded estimate ($E(\hat{N}_i) = \widehat{EN}_i$) is the mean (μ_i) and the variance is equal to the mean corrected for overdispersion ($\frac{\mu_i^2}{\kappa}$). The negative binomial family also uses a logarithmic-link (natural log) in the predictor function (η), which ensures that the fitted values are non-negative. Subsequent diagnostics and results will be reported in the space of the predictor function. Model fits were investigated by inspecting Pearson residuals. Outliers and those with high leverage were noted but not excluded in this analysis to incorporate their effect on the uncertainty of the parameter estimates.

Models were initially developed at three set levels: Generalized (1), Size-Aggregate (2), and Individual (5). Results of the models that made it to the candidate model set are presented here. Other model selection steps can be found in the code repository in Appendix A. Individual-Specialized (6) models were also developed using regression and will be presented in more detail subsequently.

Generalized (1)

The first Generalized (1) model developed included the covariates: visual count, water clarity, substrate shade, system size, and visual count method. The model was fit with interaction terms between visual count and every other covariate and dredged to produce all possible combinations of covariates. The full model with all terms was within 2 AICc units of the top model (Table 9) and there were minimal patterns in its residuals (Figure 14); therefore, it was selected for use in the candidate model set. Only Stellako River 2010-Ground had a Pearson residual considered an outlier (absolute value > 3). There were 16 system-years flagged as having potentially high leverage; 13 of which were from aerial surveys.

This model will be referred to as the Generalized (1) – Clarity GLM to differentiate it with the next model that excluded water clarity for scenarios where that covariate was not retrievable. All parameter estimates did not cross the zero-effect line when considering their standard error, with the exception of the slope interaction effect of the size-class X-large (Table 10, Figure 15). Size and substrate shade had the most model weight of models within 2 AICc units from the top and their parameter effects were generally larger than water clarity and survey method (Table 9, Table 10).

Table 9. Covariate interaction selection results of the Generalized (1) Clarity GLM developed within 4 delta AICc. C represents the interaction with visual count. + indicates the inclusion of the covariate. Note that all models included the non-interaction terms for every covariate and an intercept but are not shown in this table.*

C*Method	C*Size	C*Substrate	C*Clarity	AICc	ΔAICc
	+	+	+	3458.791	0.00
	+	+		3459.113	0.32
+	+	+		3459.34	0.55
+	+	+	+	3459.487	0.70
	+			3459.772	0.98
+	+			3461.416	2.62
	+		+	3462.159	3.37

Generalized (1) - Clarity GLM

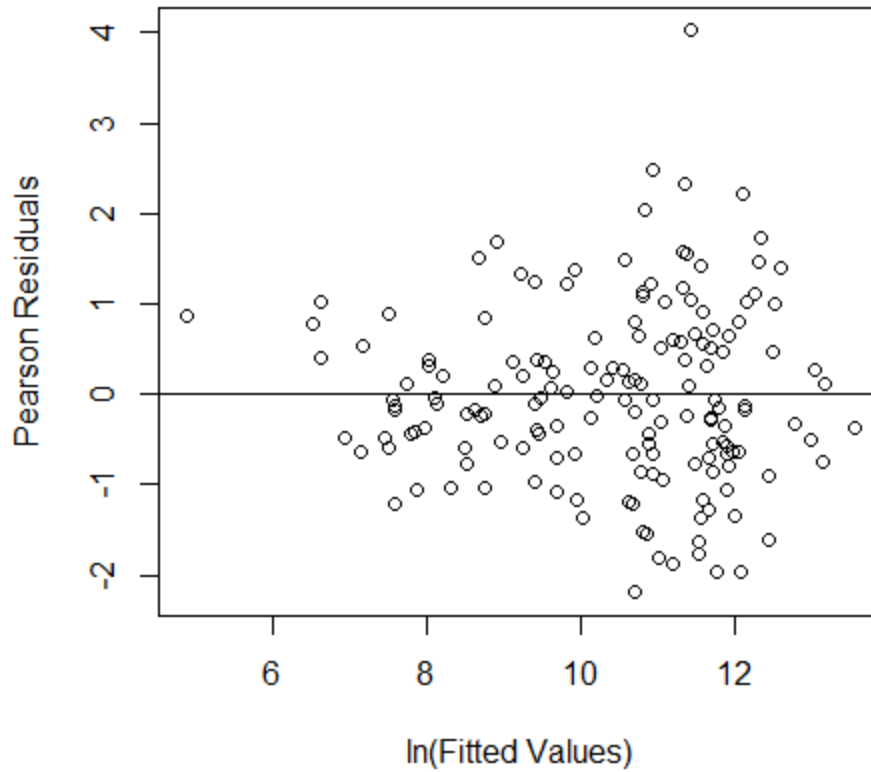


Figure 14. Pearson Residuals of the Generalized (1) Clarity GLM.

Table 10. Parameter estimates for Generalized (1) Clarity GLM, sorted by effect size. The intercept case is for: method = aerial, size = large, substrate = dark, clarity = clear.

Parameter	Estimate	Standard Error
(Intercept)	11.158	0.205
Scaled Logarithm of Visual Count	1.309	0.222
Size Small	-1.265	0.181
Size Medium	-1.080	0.176
Substrate Shade Light	0.497	0.116
Substrate Shade Medium	0.483	0.127
Count*Substrate Medium Interaction	-0.411	0.138
Count*Size Small Interaction	0.391	0.155
Survey Method Ground	-0.377	0.121
Size X-Large	-0.374	0.188
Count*Size Medium Interaction	0.368	0.179
Count*Substrate Light Interaction	-0.364	0.140
Water Clarity Turbid Tannic	0.272	0.094

Count*Survey Method Ground Interaction	0.219	0.157
Count*Water Clarity Turbid Tannic Interaction	-0.157	0.107
Count*Size X-Large Interaction	-0.105	0.203

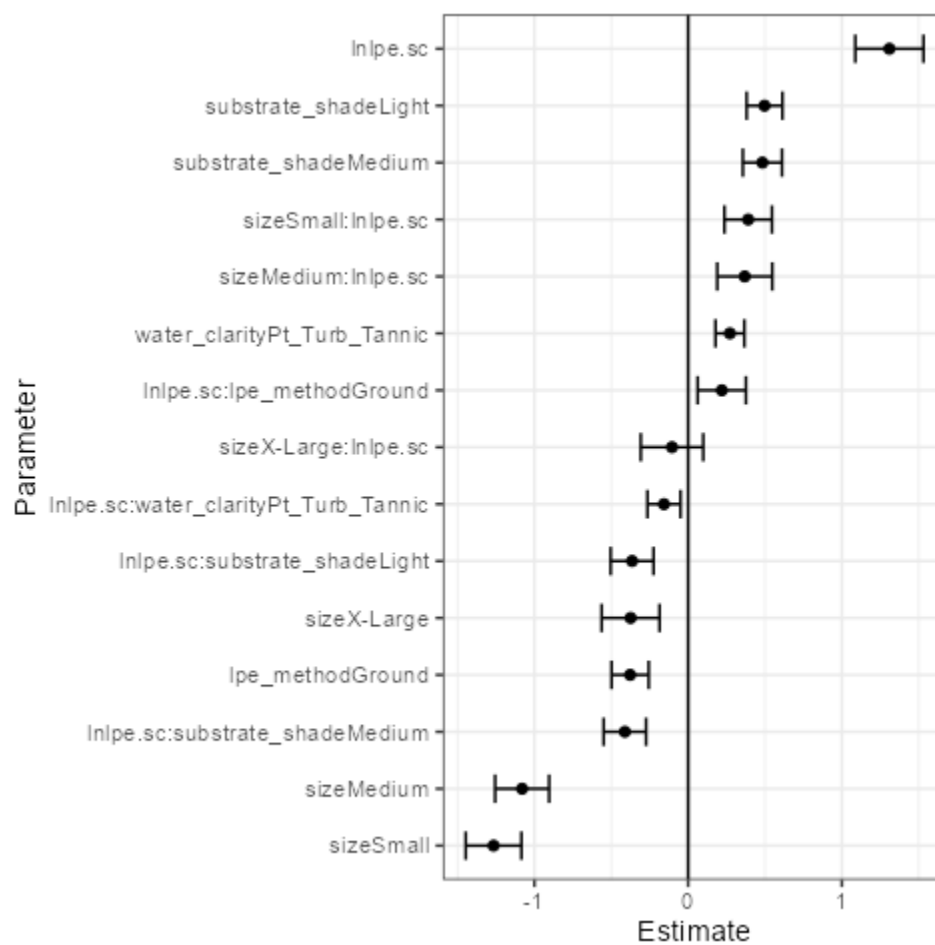


Figure 15. Parameter estimates with their standard errors for the Generalized (1) Clarity GLM. Excludes intercept, which is a large effect (11.16). The intercept case is for: method = aerial, size = large, substrate = dark, clarity = clear.

Since the Generalized (1) – Clarity GLM included a covariate that may not always be retrievable (water clarity), an additional model was developed at this set level that only included retrievable or annually-stable covariates. Water clarity was removed as a covariate to only include covariates that are highly retrievable. To further increase the generalizability of the model, survey method was also dropped as a covariate. Survey method was chosen to be dropped since it held the least amount of model weight in the prior comparison and generally had lower effect sizes than the average effect size of the covariates size and substrate shade (Table 9, Table 10). The model was fit with interaction terms between visual count and size and substrate shade. The model with all terms was the top model (Table 11) and there were minimal patterns in its residuals (Figure 16); therefore, it was selected for use in the candidate model set. Stellako River 2010-Ground survey remained an outlier in this model in addition to Scotch Creek 2018-Aerial survey. There were 13 potentially high-leverage system years; 10 of which were from aerial surveys.

This model will be referred to as the Generalized (1) – Minimal GLM. All parameter estimates did not cross the zero-effect line when considering their standard error, with the exception of the slope interaction effect of the size-class Medium (Table 12, Figure 17). The direction and overall magnitude of parameter estimates did not change greatly between the two Generalized (1) models.

Table 11. Covariate selection results of the Generalized (1) Minimal GLM developed within 4 delta AICc. C* represents the interaction with visual count. + indicates the inclusion of the covariate. Note that all models included non-interaction terms for visual count and an intercept but are not shown in this table.

Size	Substrate Shade	C*Size	C*Substrate Shade	AICc	Δ AICc
+	+	+	+	3471.83	0.00
+	+	+		3475.84	4.00
+		+		3476.25	4.42

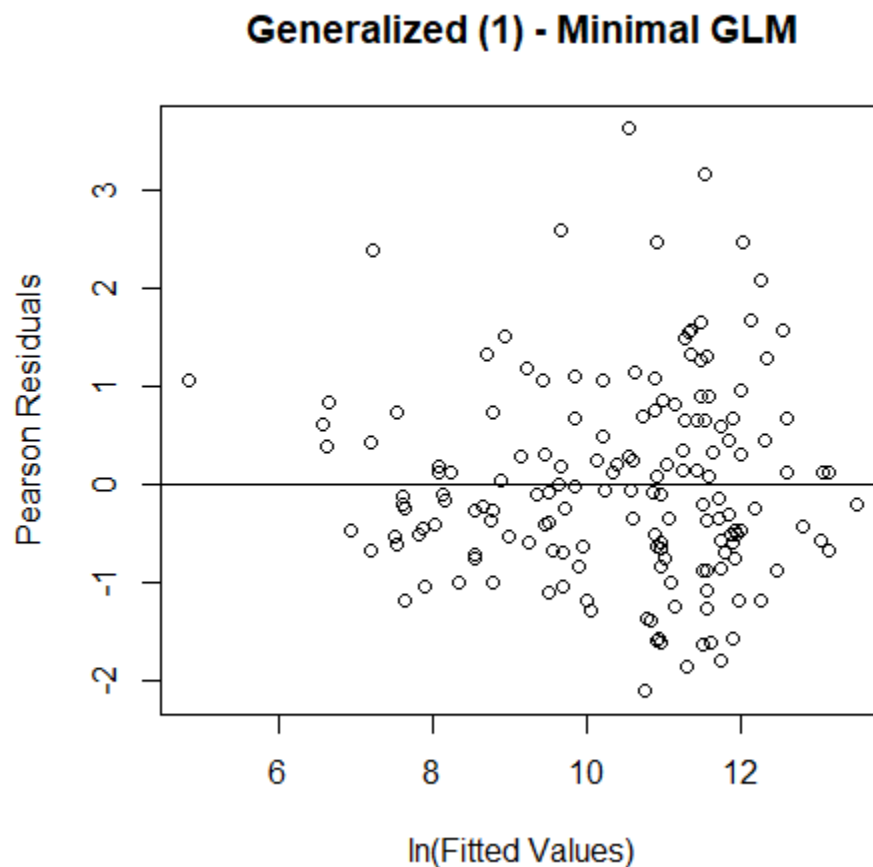


Figure 16. Pearson residuals of the Generalized (1) Minimal GLM.

Table 12. Parameter estimates for Generalized (1) Minimal GLM, sorted by effect size. The intercept case is for: size = large, substrate = dark.

Parameter	Estimate	Standard Error
(Intercept)	11.271	0.202
Size - Small	-1.473	0.183
Scaled Logarithm of Visual Count	1.296	0.189
Size Medium	-0.929	0.181
Count*Size Small Interaction	0.513	0.159
Substrate Shade Light	0.337	0.112
Count*Substrate Shade Medium Interaction	-0.311	0.121
Size X-Large	-0.236	0.192
Count*Size X-Large Interaction	-0.215	0.175
Substrate Shade Medium	0.212	0.111
Count*Substrate Shade Light Interaction	-0.198	0.128
Count*Size Medium Interaction	0.168	0.165

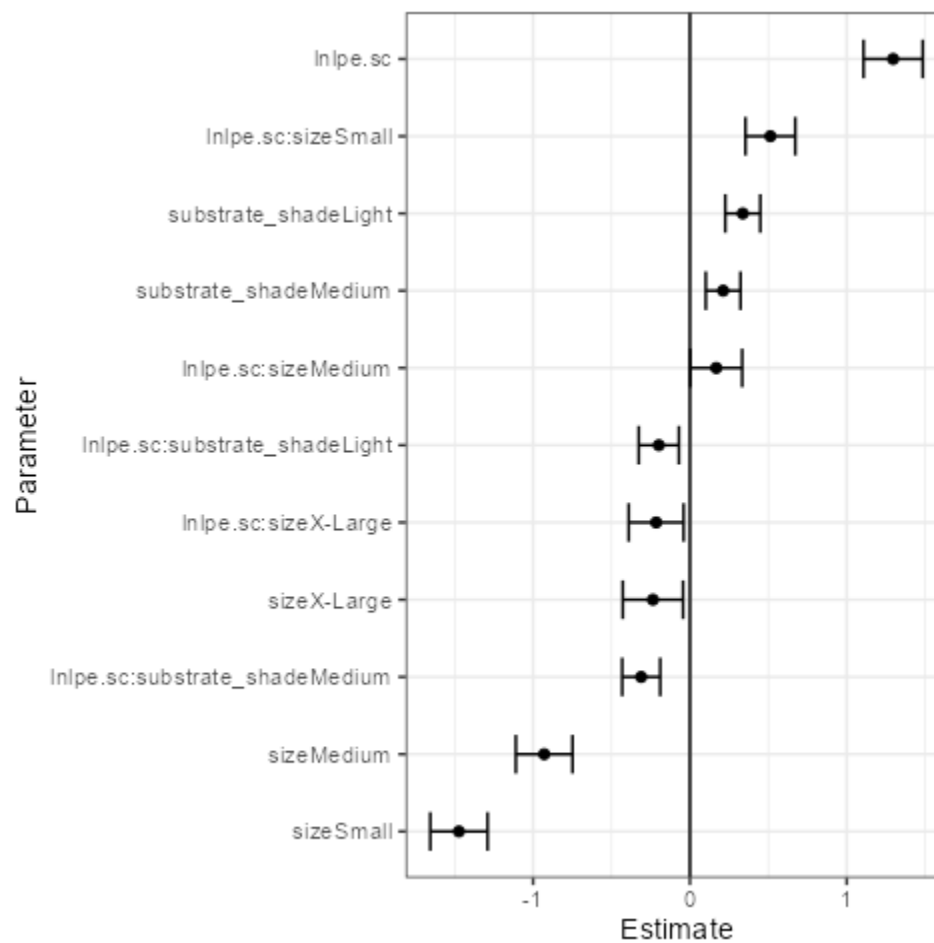


Figure 17. Parameter estimates with their standard errors for Generalized (1) Minimal GLM. Excludes the intercept, which is a large effect (11.27). The intercept case is for: size = large, substrate = dark.

Size-Aggregate (2)

There are two subsets of data at the Size-Aggregate (2) set-level. The covariates screened in to the small-medium subset of data model development included: visual count, water clarity, substrate shade, system stability, and visual count method. The model was fit with interaction terms between visual count and every other covariate and dredged. The full model with all terms was not within 2 AICc units of the top model (Table 13). Within 2 AICc units of the top model, the model that dropped both the system stability interaction and intercept effects was selected for diagnostics. There were minimal patterns in its residuals (Figure 18); therefore, it was selected for use in the candidate model set. This model will be referred to as the Size Aggregate (2) – Small-Medium Model. Only Stellako River 2010-Ground had a Pearson residual considered an outlier (absolute value > 3); which also had high leverage (Cook's distance > 3*average distance). There were seven other system-years with potentially high leverage; three of which were from Nadina River aerial surveys, two from Eagle River aerial surveys, and second Stellako River survey but instead from the aerial method. All parameter estimates did not cross the zero-effect line when considering their standard error (Table 14, Figure 19). Substrate parameter effects were generally the largest after the effect of visual count (Figure 19).

Table 13. Covariate selection results of the Size-Aggregate (2) model developed with the subset of small-medium data within 4 delta AICc. C* represents the interaction with visual count. + indicates the inclusion of the covariate. Note that all models included non-interaction terms for visual count and an intercept but are not shown in this table.

Stability	Clarity	C*Method	C*Substrate	C*Stability	C*Clarity	AICc	ΔAICc
	+	+	+			2490.4	0
+	+	+	+			2491.2	0.82
	+	+	+		+	2491.7	1.25
+	+	+	+		+	2492.4	1.98
+	+	+	+	+		2493.5	3.15

Size Aggregate (2) - Small-Medium Model

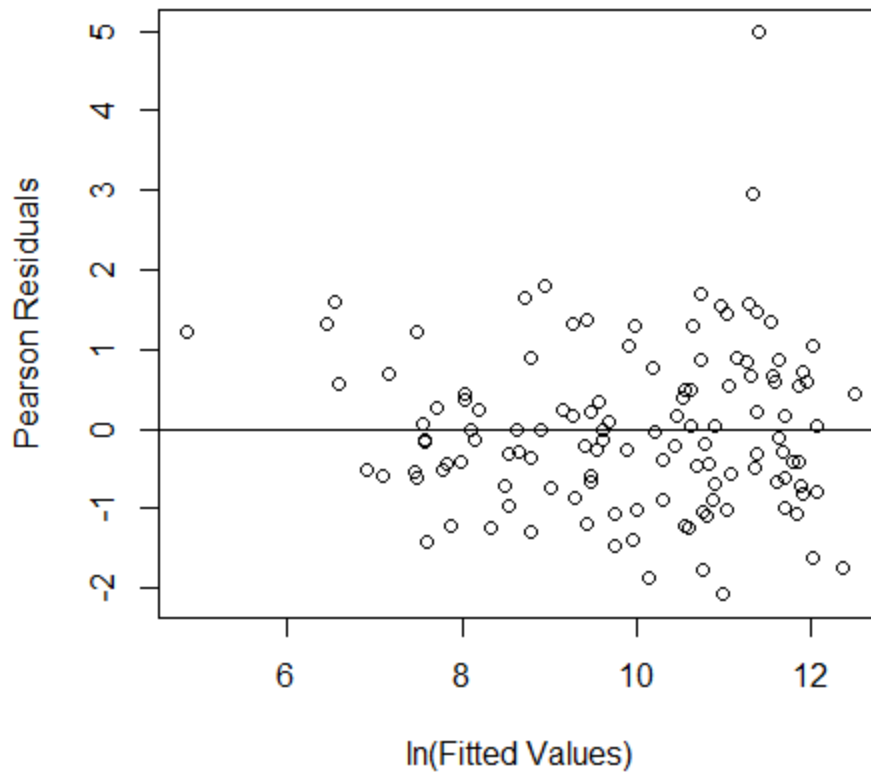


Figure 18. Pearson residuals of the Size Aggregate (2) Small-Medium Model.

Table 14. Parameter estimates for the Size Aggregate (2) Small-Medium Model, sorted by effect size. The intercept case is for: method = aerial, clarity = clear, substrate = dark.

Parameter	Estimate	Standard Error
(Intercept)	9.81	0.15
Scaled Logarithm of Visual Count	1.83	0.18
Count*Substrate Shade Light Interaction	-0.62	0.19
Count*Substrate Shade Medium Interaction	-0.62	0.19
Method Ground	-0.52	0.08
Substrate Shade Medium	0.52	0.15
Substrate Shade Light	0.48	0.14
Clarity Turbid Tannic	0.39	0.09
Count*Method Ground Interaction	0.35	0.09
Count*Clarity Pt_ Turb_ Tannic Interaction	-0.11	0.11

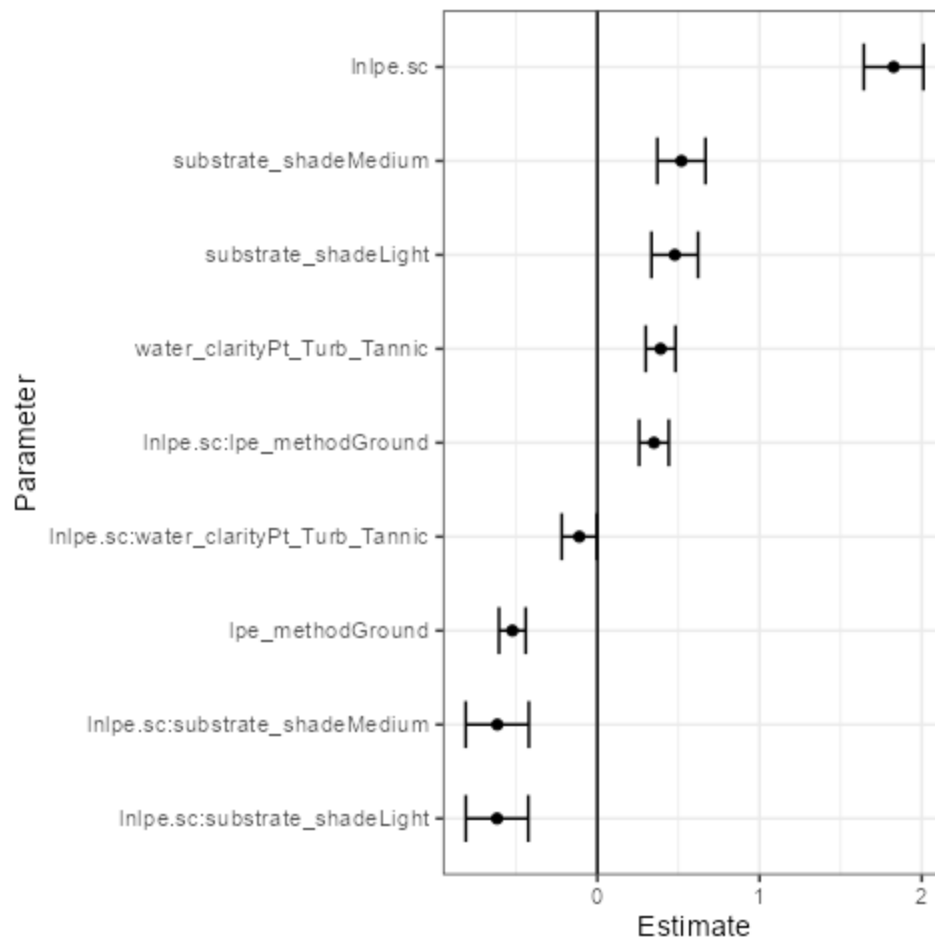


Figure 19. Parameter estimates with their standard errors for the Size Aggregate (2) Small-Medium Model. Excludes intercept, which is a large effect (9.81). The intercept case is for: method = aerial, clarity = clear, substrate = dark.

The large-x-large subset of data had fewer samples and a limited number of covariates with heterogeneous factors. System size and survey method had the highest correlation coefficient (0.72) but neither had the largest VIF. The covariates screened in to the large-x-large subset of data model development included: visual count, water clarity, substrate shade, system size, and visual count method. The model was fit with interaction terms between visual count and every other covariate and dredged. No model within 2 AICc units of the top model had interaction terms (Table 15). The top three models either had size, water clarity, or both size and water clarity covariates, respectively. The model that contained both, in addition to visual count, was selected instead of model averaging because the results would be similar. There was a patterns in the model's residuals; the residuals reduced in spread the larger the fitted estimate (Figure 20), thus the model was rejected. Investigating the data further resulted in subsetting to aerially surveyed data because it was most of the data in the aggregate ($n = 30$ aerial surveys relative to $n = 8$ for ground surveys). This reduced the variability in covariates even more. An interactions model was developed only using the covariates water clarity and system size and dredged. The interaction with system size was within the top 2 AICc but the interaction with water clarity was greater than 2 AICc from the top (Table 16). The residuals of the model with only system size and it's interaction with visual count also had a better distribution (Figure 21), although still slightly tapering at higher fits. There were no outliers and only the Adams River

(Lower)-2018 Ground survey had potentially high leverage. The model was selected to be included in the candidate model set and will be referred to as the Size-Aggregate (1) – Large-X-Large Aerial Model. All parameter estimates did not cross the zero-effect line when considering their standard error (Table 17, Figure 22).

Table 15. Covariate selection results of the Size-Aggregate (2) model developed with the subset of large-x-large data within 3 delta AICc. C represents the interaction with visual count. + indicates the inclusion of the covariate. No model was selected from this suite.*

Method	Size	Substrate	Clarity	C* Method	C* Size	C* Substrate	C* Clarity	AICc	ΔAICc
	+							947.87	0
			+					948.22	0.35
								949.44	1.58
	+		+					949.66	1.8
+	+							949.71	1.84
		+						950.17	2.31
	+				+			950.28	2.42
+			+					950.75	2.89
	+	+						950.85	2.98

Size Aggregate (2) - Large-X-Large Model

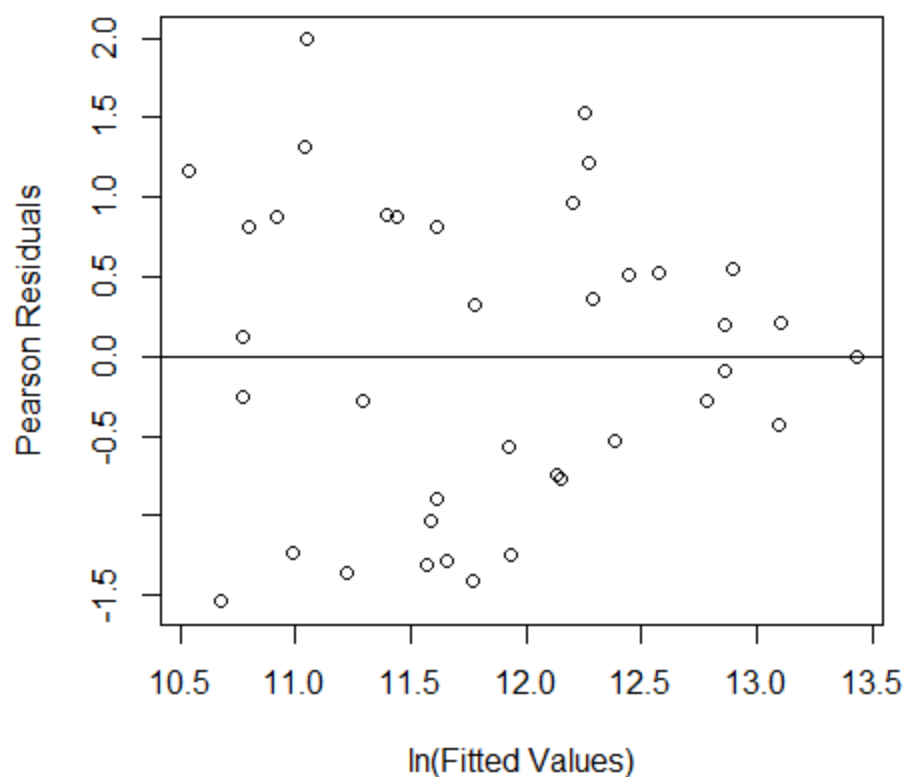


Figure 20. Pearson residuals of the Size Aggregate (2) Large-X-Large (both methods) Model. This model was rejected due to the conical shape of the residuals.

Table 16. Covariate selection results of the Size-Aggregate (2) Large-X-Large model developed with the subset of aerial large-x-large data within 3 delta AICc. C* represents the interaction with visual count. + indicates the inclusion of the covariate.

Size	Clarity	C*Size	C*Clarity	AICc	Δ AICc
+				743.00	0
	+			743.10	0.11
				743.20	0.20
+		+		744.70	1.71
+	+			744.75	1.76
	+		+	745.99	3.00

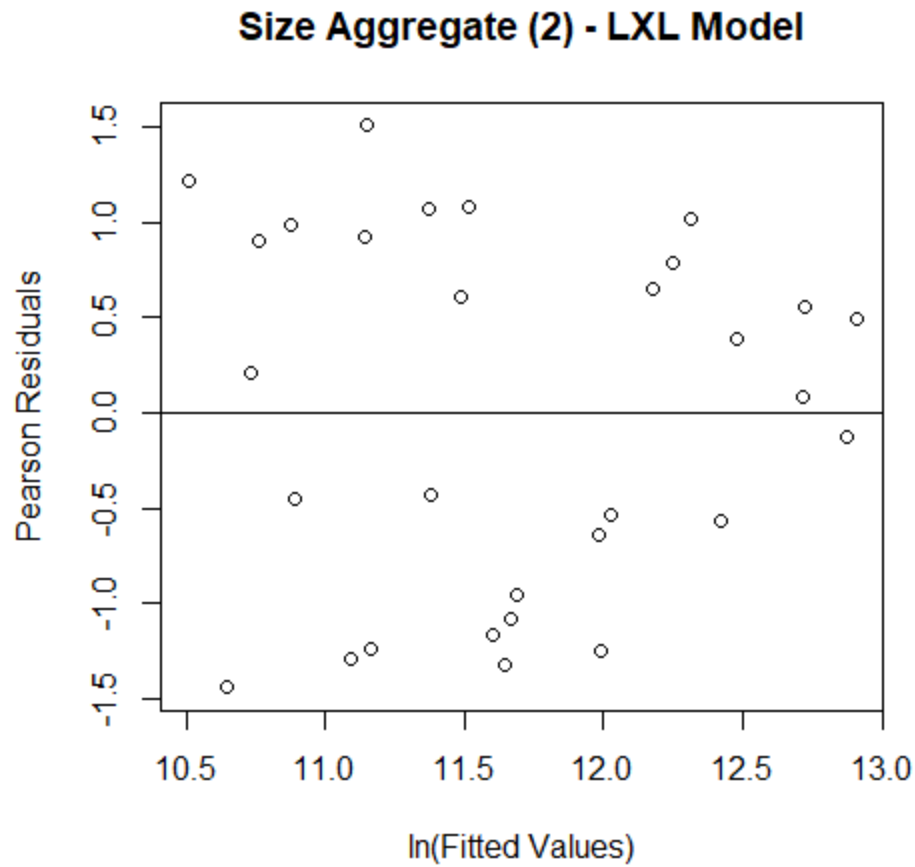


Figure 21. Pearson residuals of the Size Aggregate (2) Large-X-Large Aerial Model.

Table 17. Parameter estimates for the Size Aggregate (2) Large-X-Large Aerial Model, sorted by effect size. The intercept case is for: size = large.

Parameter	Estimate	Standard Error
(Intercept)	12.13	0.27
Scaled Logarithm of Visual Count	1.35	0.62
Count*Size X-Large	-0.66	0.63
Size X-Large	-0.50	0.28

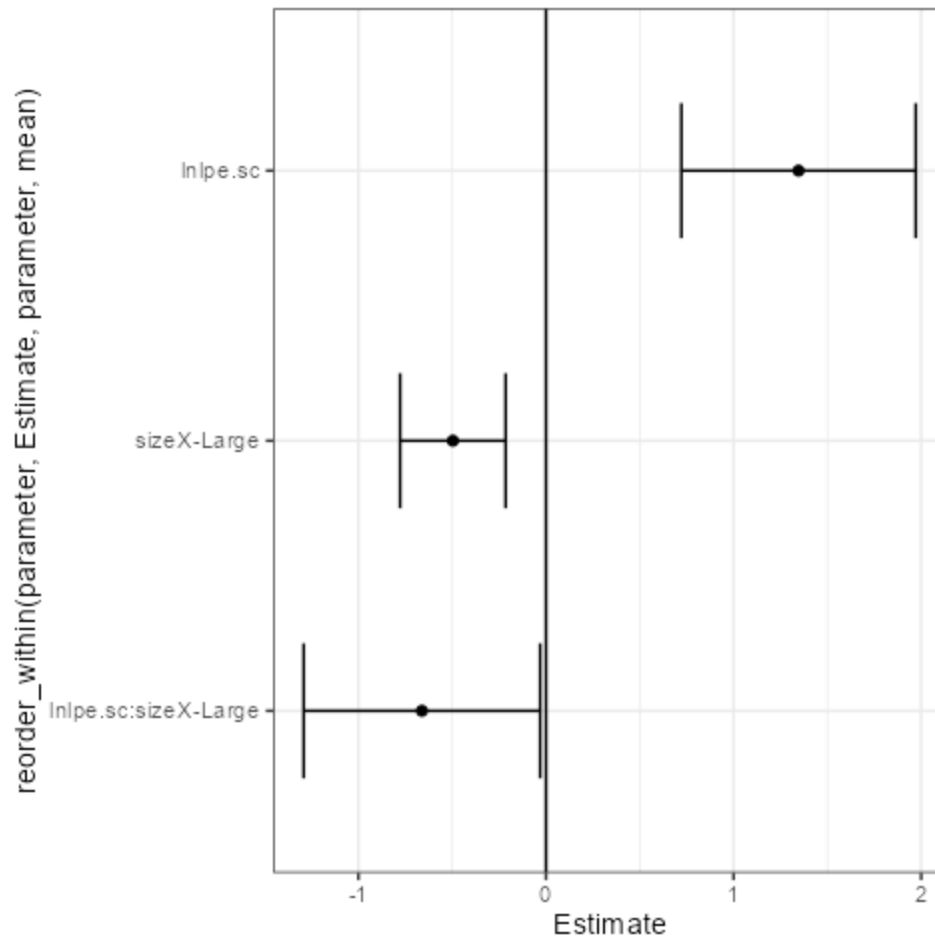


Figure 22. Figure of parameter estimates with their standard errors for the Size Aggregate (2) Large-X-Large Aerial Model. Excludes intercept, which is a large effect (12.13). The intercept case is for: size = large.

Individual (5)

No covariates were applied to individual data subsets other than the visual count because sample sizes were small, resulting in candidate model selection purely relying on residuals to determine their inclusion. Most Individual (5) models had sparse data and some amount of heteroscedasticity in their residuals (Figure 23). Only Stellako-Ground had one Pearson residual considered an outlier from the year 2010 (absolute value > 3); which also had high leverage (Cook's distance > 3*average distance). All Individual (5) models had one data point with high leverage with the exception of O' Ne-Ell Creek-Ground, which had no high leverage points. Individual (5) models were not ideal fits, but it was difficult to determine which should be excluded from the candidate model set. Stellako-Ground had the most defined upside-down U-shape. All are currently included in the candidate models except for Adams River (Lower) due to its sample size <10, but other diagnostics or rules may be needed to determine exclusion in the future.

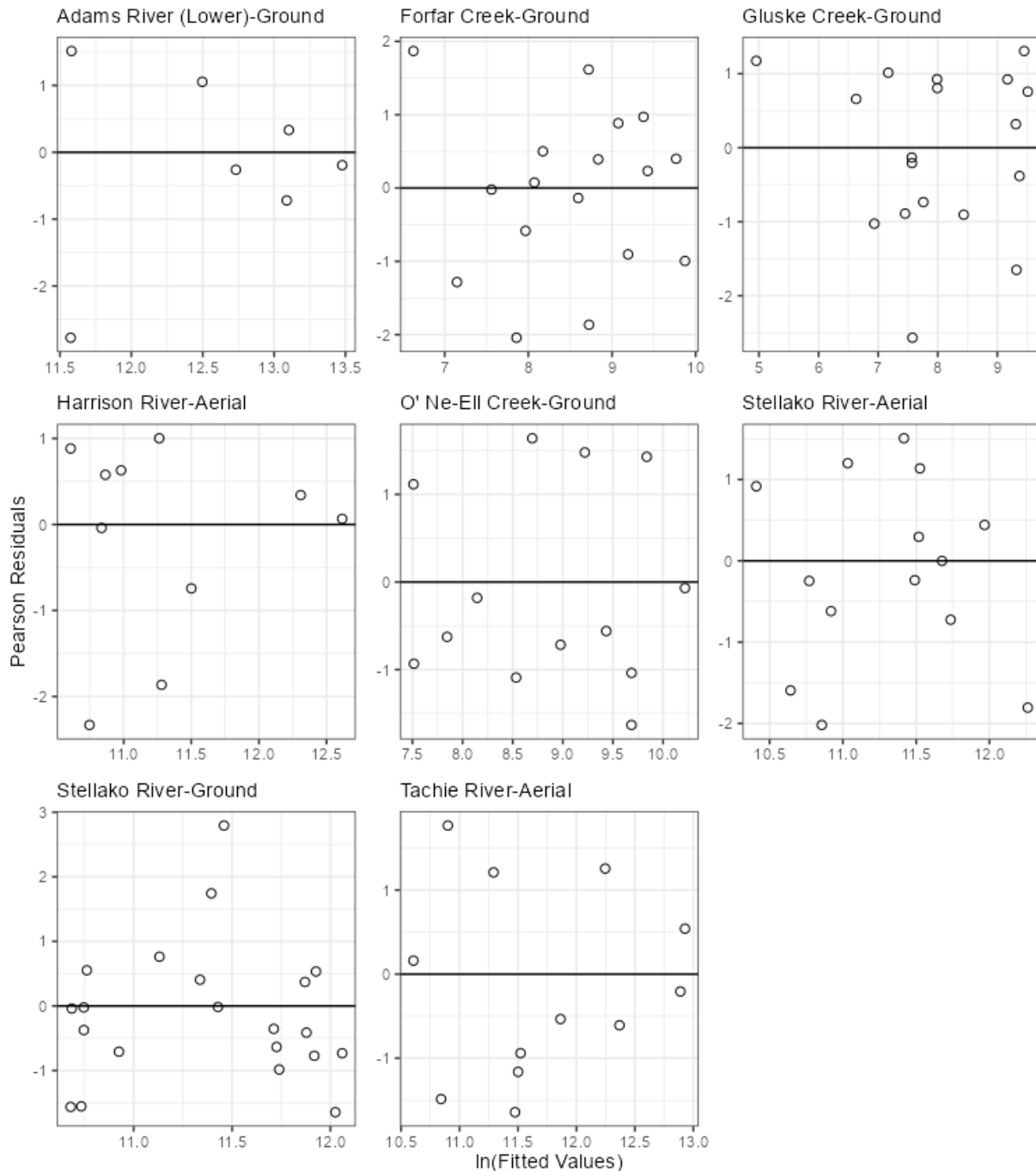


Figure 23. Individual (5) Model Pearson residuals. Each plot is associated with system-survey method in its title.

Early Stuart Specialized (6)

Early Stuart has the highest average proportion of the total spawner abundance estimated using visual-based methods in the Fraser River; more than double of any other MUs (Table 1), and was the target of historical expansion factor development. Early Stuart has been the focus of paired data collection because of the reliance of visual methods to enumerate its systems. The three systems used in the Early Stuart Specialized (6) model development comprise 28% of the

total database. The model development follows the same general framework as the prior models in this section with the difference of specialized covariate development, data transformations instead of the use of generalized linear models, and inclusion of a random effect (mixed model).

The models make use of the high survey effort implemented at Early Stuart systems by creating specialized covariates that can only be calculated from systems that have extended and systematic carcass recovery and enumeration efforts. The covariates attempt to address the variability in spawning behaviour of Sockeye. On average, spawn timing follows a normal or log-normal distribution, where there is a peak in spawner abundance followed by a peak in die-off or carcass abundance (Groot and Margolis 1991). However, variability in the temporal distribution of spawning activity may result in a protracted spawning window with no or multiple peaks in spawning abundance and a gradual accumulation of carcasses. When there is protracted spawning, with no obvious peak or multiple peaks, the traditional peak-of-spawn methods often become biased low due to observers not being able to enumerate spawners and carcasses effectively.

Three covariates were investigated to address this issue: total cumulative carcasses at the end of the visual surveys, and two protraction indices that used the total cumulative carcasses in different ratios. The first ratio was the cumulative carcasses at the peak of spawn relative to the total cumulative, which was developed to indicate if there was a lengthy die-off period after the peak of spawn count. A value near zero indicates that spawning, and inversely die-off, was highly protracted while values closer to one indicate an abrupt end to spawning after the peak of spawn. The second ratio was the total cumulative carcasses relative to the peak live count, which was developed to indicate if the peak count captured or missed most of the spawning abundance. Values below one or closer to one indicate that the peak live count captured most of the spawning abundance while larger values indicate that there was not a defined peak in spawning or arrival. Each of these covariates can only be calculated if survey effort extends past the peak of spawn so that the variability in spawning behaviour can be recorded.

Data transformations were explored instead of using generalized linear models after testing the distributions of the data with the Shapiro-Wilk normality test (Royston 1982, R Core Team 2024). The system's abundances are generally low (median = 4,194; range 167-27,010) and did not follow a log-normal distribution like the entire database. Normality was achieved in most variables (high-accuracy estimate, visual count, total cumulative carcasses, and the 2nd ratio) using a cube-root transformation and the 1st ratio received a natural-log transformation. All variables (except for the response variable) were also then centered and scaled to make the parameter estimates comparable. All of the Early Stuart systems in the data subset share the same size, clarity, and stability classes as well were all surveyed with ground methods. The only categorical covariates explored were substrate shade and large woody debris. The largest Pearson correlations were with the total cumulative carcasses covariate, so it was removed from further model selection.

The model development and selection of the final candidate model followed the same process of dredging a global model with all covariates coded as interaction with visual count; however, system was included as a random intercept. A random intercept of system will account for the pseudoreplication of data pairs within each of the three Early Stuart systems that may also have small inherent differences not described by the rest of the covariates. Inclusion of the random effect will also allow for better estimation of the uncertainty in other parameters' estimates (Zuur *et al.* 2009). When fitting and dredging the global model, the model was fit with maximum-likelihood.

Within the top 2 AICc, only the 2nd ratio of cumulative dead over the peak live count had an interaction term present (Table 18). Therefore, only visual count, the cumulative dead over peak alive ratio, and their interaction were included in the model; which will be referred to as the Early Stuart Specialized (6) model. The model was refit with restricted maximum likelihood before investigating residuals and parameter estimates (Zuur *et al.* 2009). The residuals were relatively homoscedastic with no outliers (Figure 24). There were five potentially high leverage system-years; four of which belonged to O'Ne-ell Creek surveys.

Table 18. Covariate selection results of the Specialized (6) Early Stuart Protraction model developed with the subset of Early Stuart data within 3 delta AICc. C* represents the interaction with visual count, which is cube-root transformed in this model. + indicates the inclusion of the covariate. Note that all models included an intercept but are not shown in this table. LWD = the presence of large woody debris.

LWD	Substrate	C* Substrate	Protraction 2	C* Protraction 2	Protraction 1	AICc	ΔAICc
	+		+	+		127.30	0
			+	+		129.13	1.83
	+		+	+	+	129.48	2.18
	+	+	+	+		129.86	2.55
+	+		+	+		129.87	2.57
+			+	+		130.13	2.82

Early Stuart Specialized (6) Model

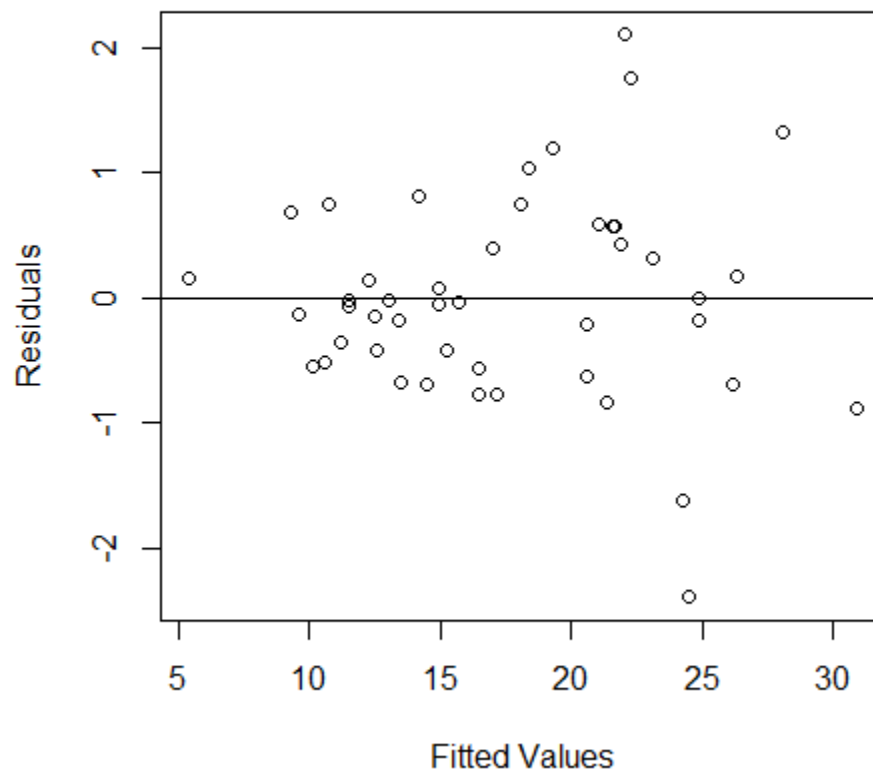


Figure 24. Residuals of the Specialized (6) Early Stuart Protraction Model.

Table 19. Parameter estimates for Specialized (6) Early Stuart Protraction model, sorted by effect size

Parameter	Estimate	Standard Error
(Intercept)	17.370	0.212
Scaled Cube-Root Count	5.377	0.142
Scaled Cube-Root Protraction Index 2	0.980	0.143
Count*Protraction Index 2 Interaction	0.394	0.130

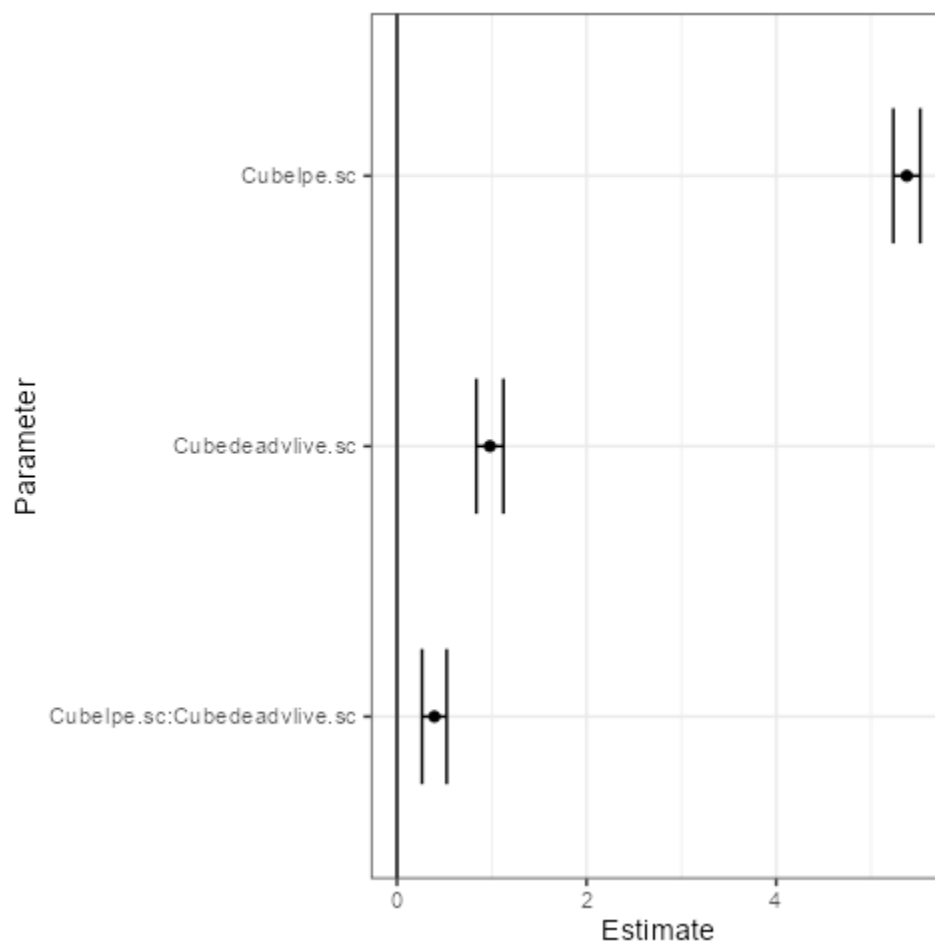


Figure 25. Parameter estimates with their standard for the Specialized (6) Early Stuart Protraction Model. Excludes intercept, which is a large effect (17.37).

Discharge Specialized (6)

Discharge had been previously suggested in unpublished reports as an index for water depth and overall visibility that would affect observer efficiency. Sixty-three data-pairs have an associated discharge recorded for them. Stellako River comprised a third of the data ($n = 23$) and a separate model was also developed using it. The discharge values are sourced from hydrometric data from the HYDAT database, a national archive of Canadian hydrometric data managed by Environment and Climate Change Canada with the R package tidyhydats (Albers 2017). The model development follows the same general framework as the prior models in this section with the difference of specialized covariate development.

The proposed mechanism for how discharge affects observer efficiency is that higher flows and therefore deeper depths reduce the ability to see into the water and count fish. Therefore, similarly to all prior covariates, the expectation is that there should be an interaction between a discharge covariate and the visual count. Changes in discharge and their impacts, however, are relative to the average discharge of individual streams. Thus, the discharge on the date of the visual count was centered and scaled by one standard deviation relative to the discharge experienced by the systems during the historical spawning window. The average and standard deviation of discharge was taken for each stream based on the last 20 years (2004-2023). In this way, all discharge values shared the same scale across different streams.

Data transformations were explored after testing the distributions of the data with the Shapiro-Wilk normality test (Royston 1982, R Core Team 2024). The systems' abundances did follow a log-normal distribution like the larger database. Normality was achieved in high-accuracy estimates and visual counts using a natural-log transformation. Visual counts were also centered and scaled, based on the data included in the discharge, and Stellako-specific discharge subsets, to make the parameter estimates comparable. The log-link was used in the GLMs to account for the log-normal distribution in the high-accuracy counts and the expectation that their error would be log-normally distributed.

In the all data discharge model, stability, water clarity, and canopy cover all had high correlation coefficients along with size; however, size was determined to be the most important due to its parameter estimates in prior models and the other three were removed before inclusion in the global model. Substrate shade was the only other covariate with diversity and a low correlation coefficient.

The model development and selection of the final candidate model followed the same process of dredging a global model with all covariates coded as interaction with visual count. Ideally, a random intercept of system would account for the pseudoreplication of data pairs within each system that may have multiple samples, but multiple systems only had one or less than five samples and models did not converge when including a random effect for system. Stellako River consisted of the largest sample and so a caveat of the all-systems discharge model is that Stellako River will have a disproportionately large effect on the parameter results.

Discharge was not present as a covariate within 2 AICc and an interaction between discharge and visual count was not present within 4 AICc (Table 20). The one model within 4 AICc that included discharge as a covariate but with no interaction was still investigated. This model is likely inherently fraught and the mechanistic interpretation would be that discharge increases the number of fish present, regardless of how many fish were counted. The residuals were relatively homoscedastic within the middle range of the fitted values but narrow at the extremes, indicating possible issues with the fit (Figure 26, Figure 24). Two data points were flagged as being possible outliers and three data points were flagged as having high leverage. The model is not recommended for application but was included in the candidate set for comparison purposes.

Table 20. Covariate selection results of the Specialized (6) All-Systems Discharge model developed with the subset all data with a discharge recorded within 3 delta AICc. C represents the interaction with visual count. + indicates the inclusion of the covariate.*

Discharge	C	Method	Size	Substrate	C*Discharge	C*Method	C*Size	AICc	ΔAICc
	+		+	+				1507.58	0.00
	+		+	+			+	1508.98	1.40
	+	+	+	+				1509.84	2.26
+	+		+	+				1509.96	2.38

+	+	+	+	+	1510.96	3.38
+	+	+	+	+	1511.50	3.92

Specialized (6) - Discharge All Systems Model

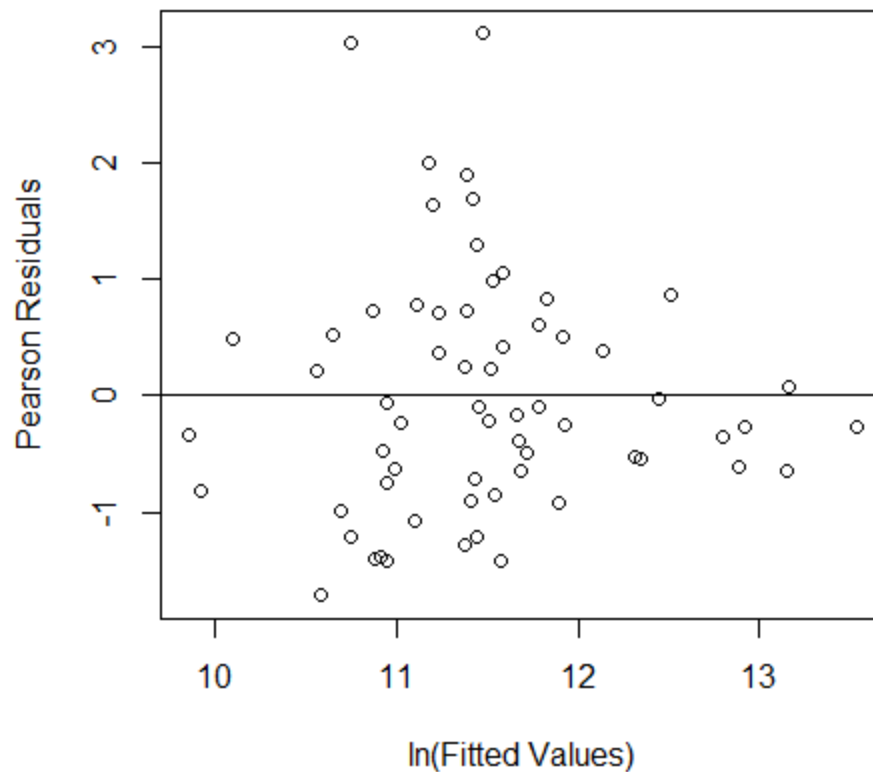


Figure 26. Pearson residuals of the Specialized (6) All-Systems Discharge Model.

Table 21. Parameter estimates for Specialized (6) All-Systems Discharge Model, sorted by effect size. The intercept case is for: size = large, substrate = dark.

Parameter	Estimate	Standard Error
(Intercept)	11.967	0.2573
Size Medium	-0.750	0.1351
Scaled Visual Count	0.702	0.0493
Substrate Shade Light	0.460	0.1696
Size X-Large	-0.402	0.2476
Substrate Shade Medium	0.014	0.2293
Scaled Discharge	0.003	0.0062

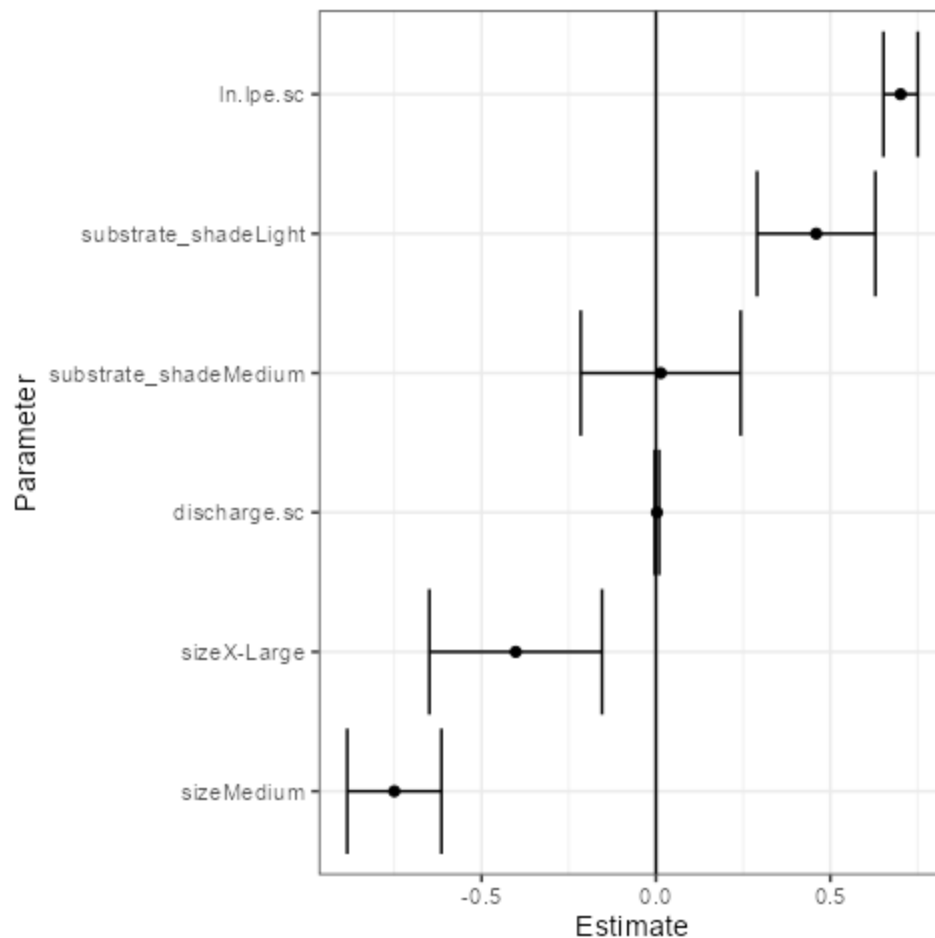


Figure 27. Parameter estimates with their standard errors for Specialized (6) All-Systems Discharge Model. Excludes intercept, which is a large effect (11.99). The intercept case is for: size = large, substrate = dark.

In the Stellako River only data discharge model covariate selection only the visual count, survey method, and discharge were considered because the other covariates were either the same every year (e.g. size, substrate, water clarity) or did not have complete coverage (e.g. cloud cover, brightness, etc.). When considering the distribution of the Stellako River visual count data, the Shapiro-Wilks test indicated that a square root transformation would be better at normalizing the data but that a natural log transformation was on the borderline of normal (i.e. $p = 0.049$). Two simple models using scaled root-transformed and log-transformed visual counts were compared using AICc, and the log-transformed variable was better by greater than four AICc, thus the log-transformation was used on the visual counts before centering and scaling it.

During data exploration, an oddity was also discovered around the relationship between the visual counts and the high accuracy estimates. When plotting the high accuracy estimates against the visual counts, it was evident that there was a plateauing of the high precision estimates despite the visual counts increasing. This lead to trying the addition of a quadratic to the visual counts as well as interacting the quadratic with each other covariate.

Discharge, the quadratic of visual counts, and their interaction terms were in the models within 4 AICc (Table 22). The most parsimonious model excluded visual method and the interaction between discharge and visual counts was only in one model. The most parsimonious model that

included discharge, visual counts, the quadratic of visual counts, and interaction terms between discharge and the quadratic of visual counts was selected. The residuals were relatively homoscedastic (Figure 28). No data points were flagged as being possible outliers but one data point was flagged as having high leverage. The model was included in the candidate set.

Table 22. Covariate selection results of the Specialized (6) Stellako Discharge model developed within 4 delta AICc. C* represents the interaction with visual count. + indicates the inclusion of the covariate. Note that all models included an intercept but are not shown.

Discharge	C	I(C^2)	Method	C* Discharge	I(C)^2* Discharge	C* Method	I(C^2): Method	AICc	ΔAICc
+	+	+			+			675.12	0.00
+	+	+	+		+			678.48	3.36
+	+	+		+	+			678.55	3.43

Specialized (6) - Stellako Discharge Model

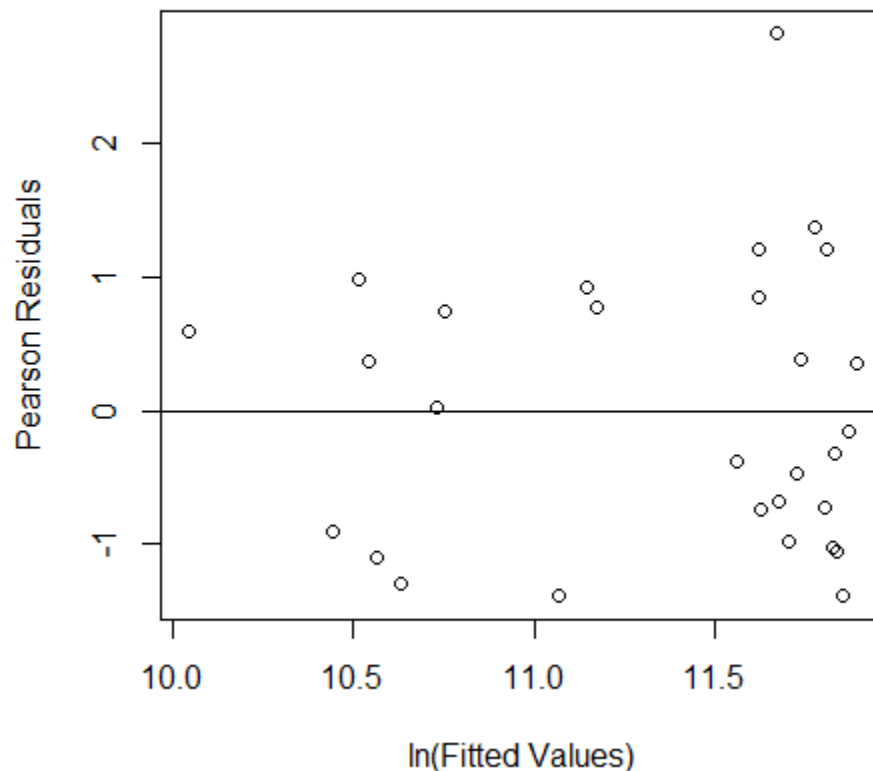


Figure 28. Pearson residuals of the Specialized (6) Stellako Discharge Model.

Table 23. Parameter estimates for Specialized (6) Stellako Discharge Model, sorted by effect size.

Parameter	Estimate	Standard Error
(Intercept)	11.536	0.0743
Scaled Logarithm of Visual Count	0.572	0.0560

$\ln(\text{Scaled Logarithm of Visual Count}^2)$	-0.201	0.0620
Scaled Discharge* $\ln(\text{Scaled Logarithm of Visual Count}^2)$		
Interaction	0.062	0.0201
Scaled Discharge	-0.015	0.0319

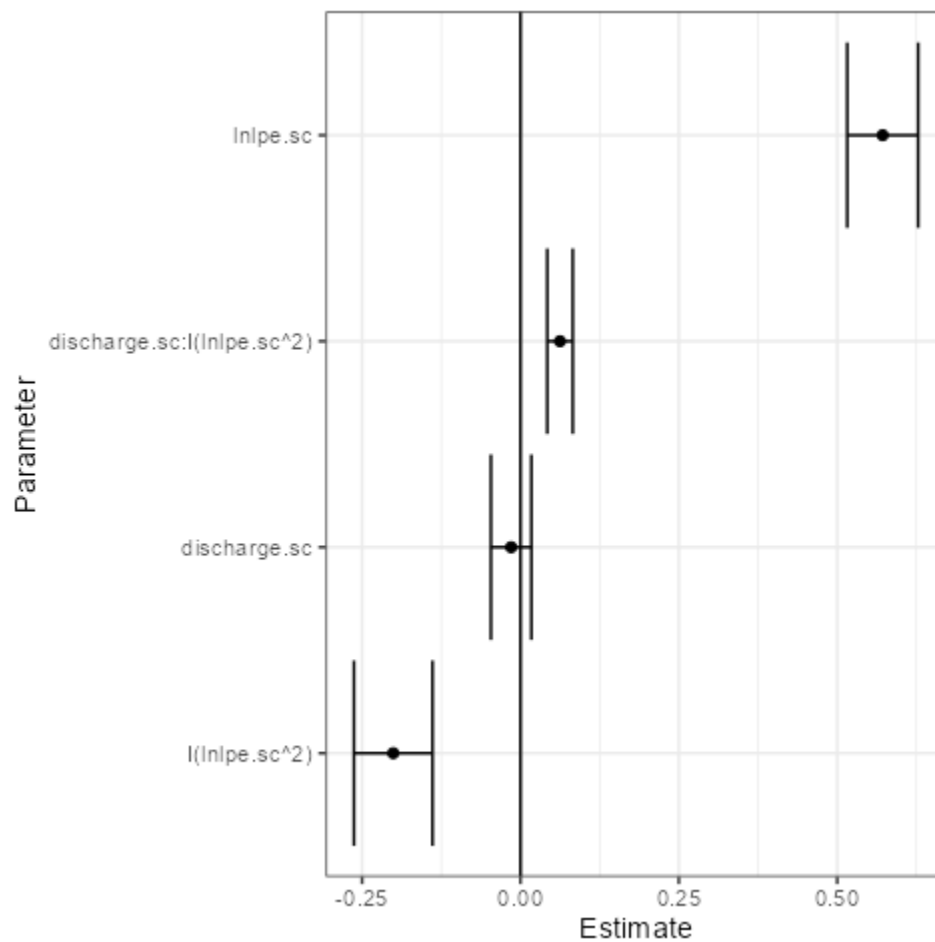


Figure 29. Parameter estimates with their standard errors for the Specialized (6) Stellako Discharge Model. Excludes intercept, which is a large effect (11.54).

APPENDIX C. RANK-SET DASHBOARDS AND TEST RESULTS

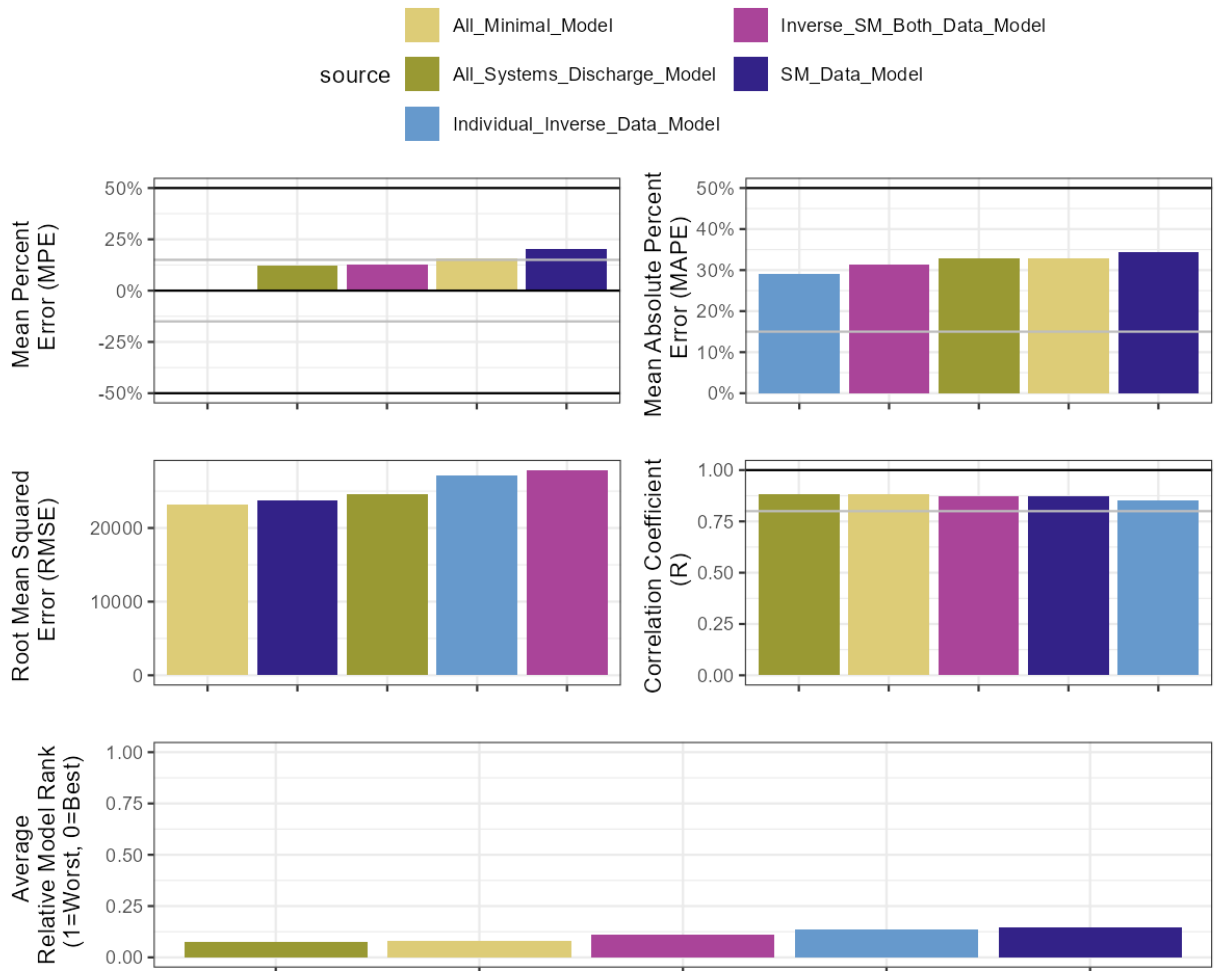
Table 24. List of Candidate Models included in the algorithm for comparison.

Expansion Method	Survey Method	Set-Level	Subset	Sample Size
Mean-of-Ratios	Aerial	Size-Clarity (4)	Medium-Clear	25
Mean-of-Ratios	Ground	Size-Clarity (4)	Small-Clear	56
Mean-of-Ratios	Ground	Size-Clarity (4)	Medium-Clear	25
Mean-of-Ratios	Ground	Individual (5)	Forfar Creek	17
Mean-of-Ratios	Ground	Individual (5)	Gluske Creek	18
Mean-of-Ratios	Ground	Individual (5)	O'Ne-Ell Creek	13
Ratio-of-Means	Both	Generalized (1)	All	162
Ratio-of-Means	Aerial	Generalized (1)	All	69
Ratio-of-Means	Ground	Generalized (1)	All	93
Ratio-of-Means	Aerial	Size-Aggregate (2)	Large-X-Large	30
Ratio-of-Means	Aerial	Size-Aggregate (2)	Small-Medium	39
Ratio-of-Means	Ground	Size-Aggregate (2)	Small-Medium	85
Ratio-of-Means	Aerial	Size (3)	Medium	38
Ratio-of-Means	Ground	Size (3)	Medium	27
Ratio-of-Means	Ground	Size (3)	Small	58
Ratio-of-Means	Aerial	Size (3)	X-Large	27
Ratio-of-Means	Aerial	Size-Clarity (4)	Medium-Clear	25
Ratio-of-Means	Ground	Size-Clarity (4)	Medium-Clear	25
Ratio-of-Means	Ground	Size-Clarity (4)	Small-Clear	56
Inverse Regression	Both	Size-Aggregate (2)	Large-X-Large	38
Inverse Regression	Aerial	Size-Aggregate (2)	Large-X-Large	30
Inverse Regression	Both	Size-Aggregate (2)	Small-Medium	124
Inverse Regression	Aerial	Size-Aggregate (2)	Small-Medium	39
Inverse Regression	Ground	Size-Aggregate (2)	Small-Medium	85
Inverse Regression	Ground	Individual (5)	Forfar Creek	17
Inverse Regression	Ground	Individual (5)	Gluske Creek	18
Inverse Regression	Aerial	Individual (5)	Harrison River	10
Inverse Regression	Ground	Individual (5)	O' Ne-Ell Creek	13
Inverse Regression	Aerial	Individual (5)	Stellako River	14
Inverse Regression	Ground	Individual (5)	Stellako River	21
Inverse Regression	Aerial	Individual (5)	Tachie River	12
Generalized Regression	Both	Generalized (1)	All w/ Water Clarity Covariate	162
Generalized Regression	Both	Generalized (1)	All w/ Minimal Covariates	162
Generalized Regression	Aerial	Size-Aggregate (2)	Large-X-Large	38
Generalized Regression	Both	Size-Aggregate (2)	Small-Medium	124
Generalized Regression	Ground	Individual (5)	Forfar Creek	17

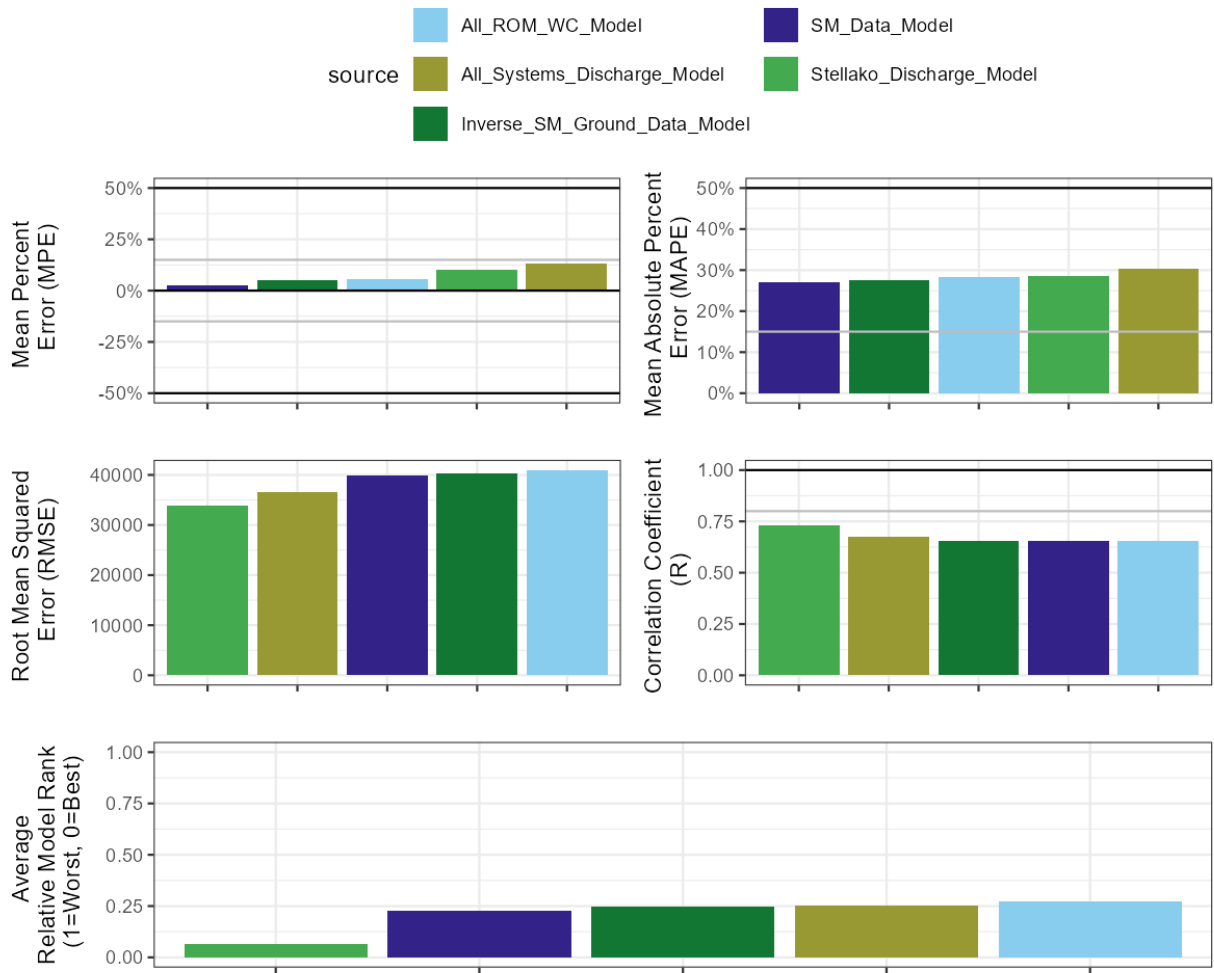
Generalized Regression	Ground	Individual (5)	Gluske Creek	18
Generalized Regression	Aerial	Individual (5)	Harrison River	10
Generalized Regression	Ground	Individual (5)	O' Ne-Ell Creek	13
Generalized Regression	Aerial	Individual (5)	Stellako River	14
Generalized Regression	Ground	Individual (5)	Stellako River	21
Generalized Regression	Aerial	Individual (5)	Tachie River	12
Linear Regression	Ground	Specialized (6)	Early Stuart	46
Generalized Regression	Both	Specialized (6)	All - Discharge	66
Generalized Regression	Both	Specialized (6)	Stellako - Discharge	29

Rank-set dashboards are presented only for those that passed the algorithm robustness rules. Each dashboard shares the same panel structure, so figure captions will only be descriptive of the rank-set level. Each dashboard has a title, legend, and five bar plots. The title indicates the numerical specialization level, the general name of the rank-set (e.g. specialized model name followed by the system name it is applied to, or the Size if it is a set-level 3), the survey method type (ground or aerial), and the sample size in the rank-set (in square brackets). The legend indicates the various models that were compared in the rank-set. The models are subset to only the top five models within each rank-set, so some dashboards do not show all model results. The bar plots are the performance metrics (MPE, MAPE, RMSE, Correlation Coefficient, and Average Relative Model Rank), sorted by descending model score within that rank. Note that MPE is plotted but the absolute MPE determined the relative rank for that metric.

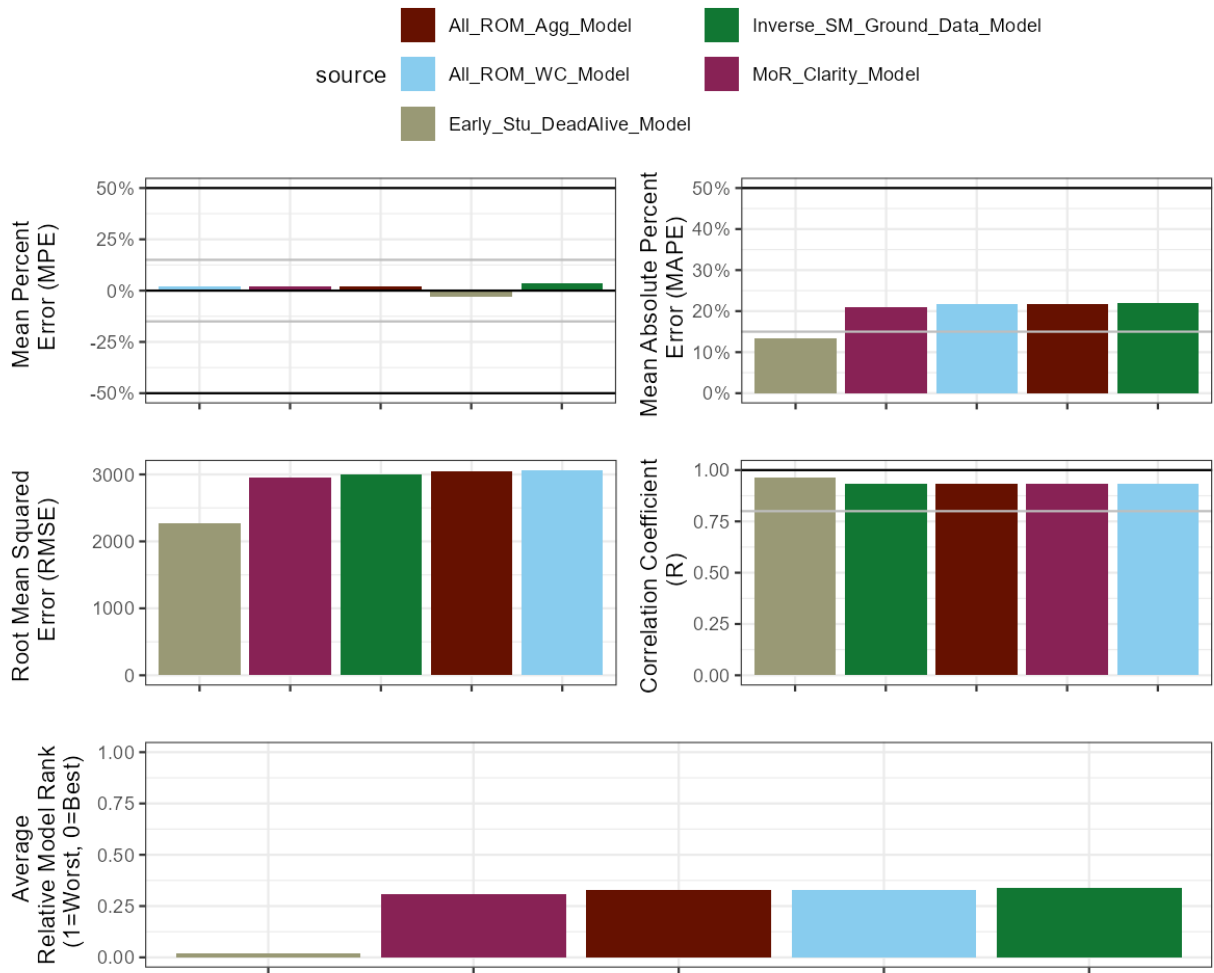
Model Performance Results for 06_Stellako_Discharge_Model - Stellako River - Aerial Set [11]



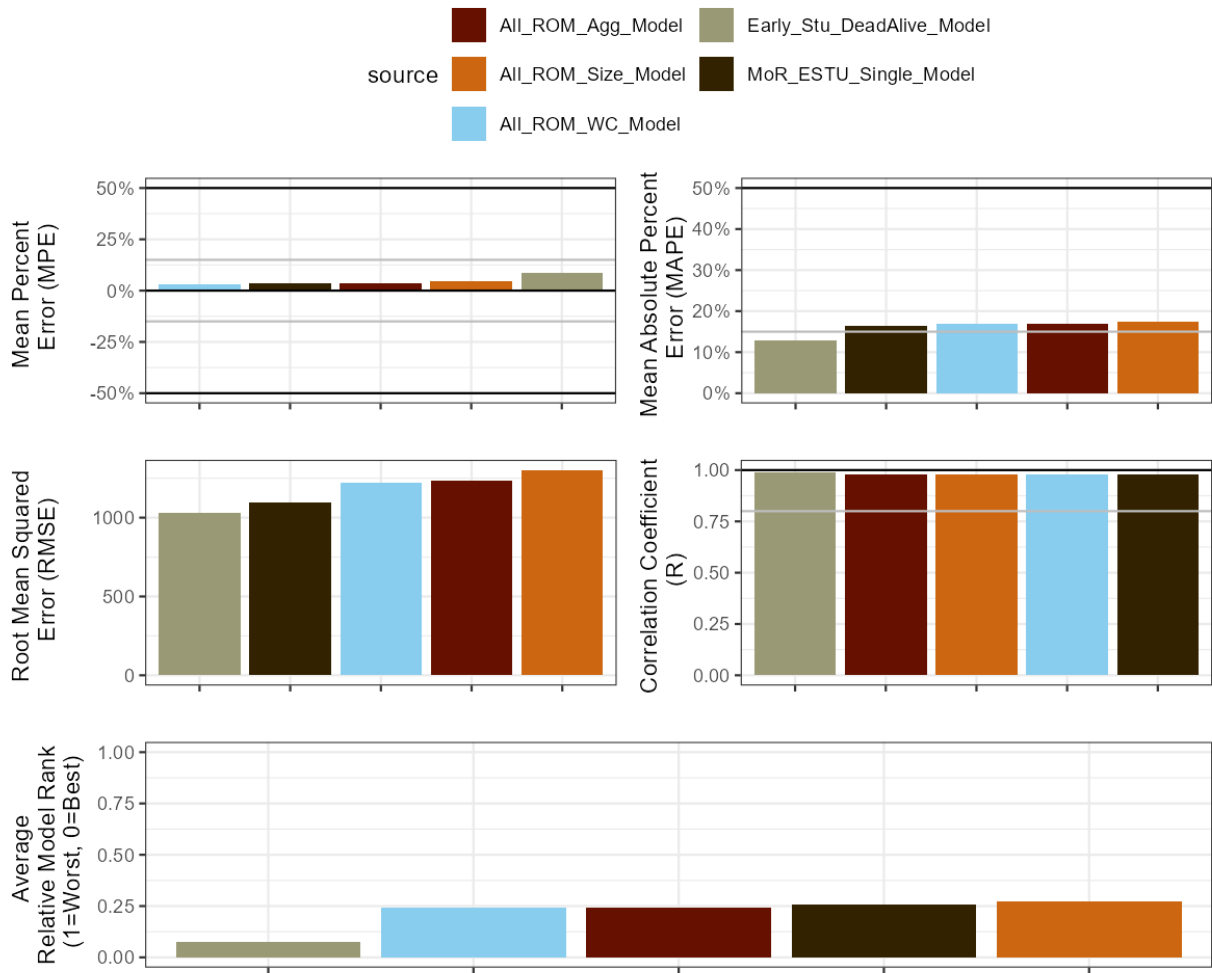
Model Performance Results for 06_Stellako_Discharge_Model - Stellako River - Ground Set [18]



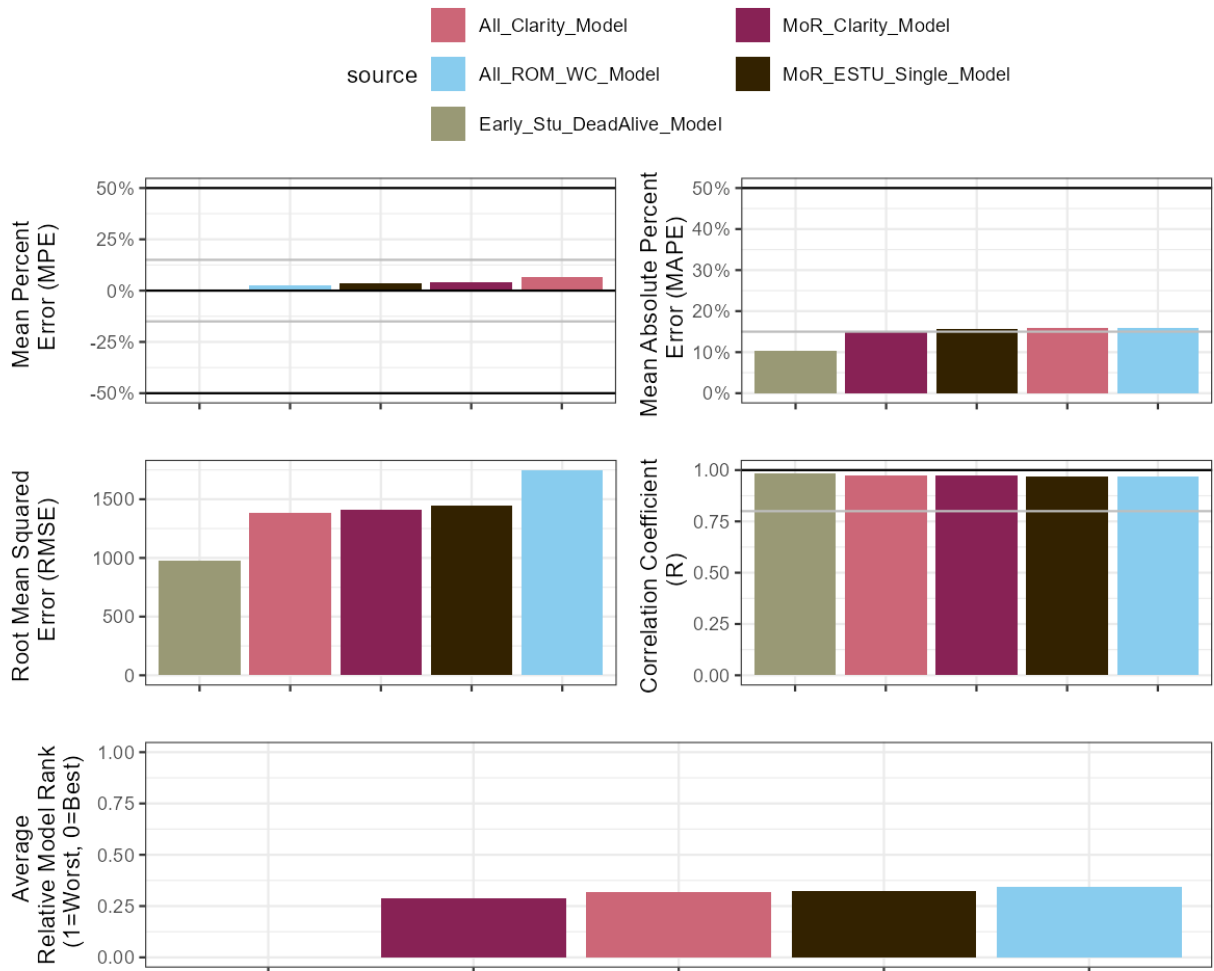
Model Performance Results for 06_Early_Stu_DeadAlive_Model - O' Ne-El Creek - Ground Set [12]



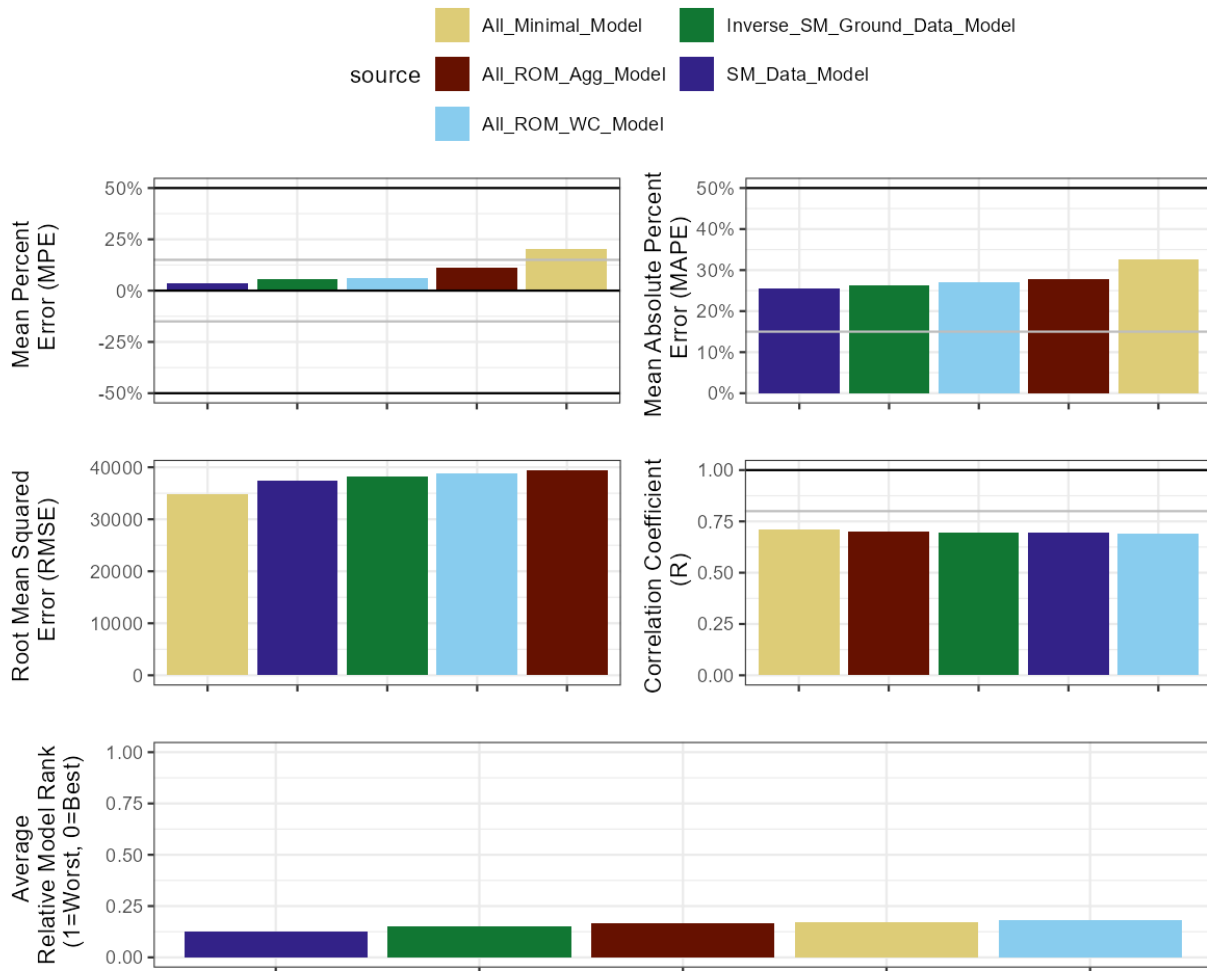
Model Performance Results for
06_Early_Stu_DeadAlive_Model - Gluske Creek - Ground Set [17]



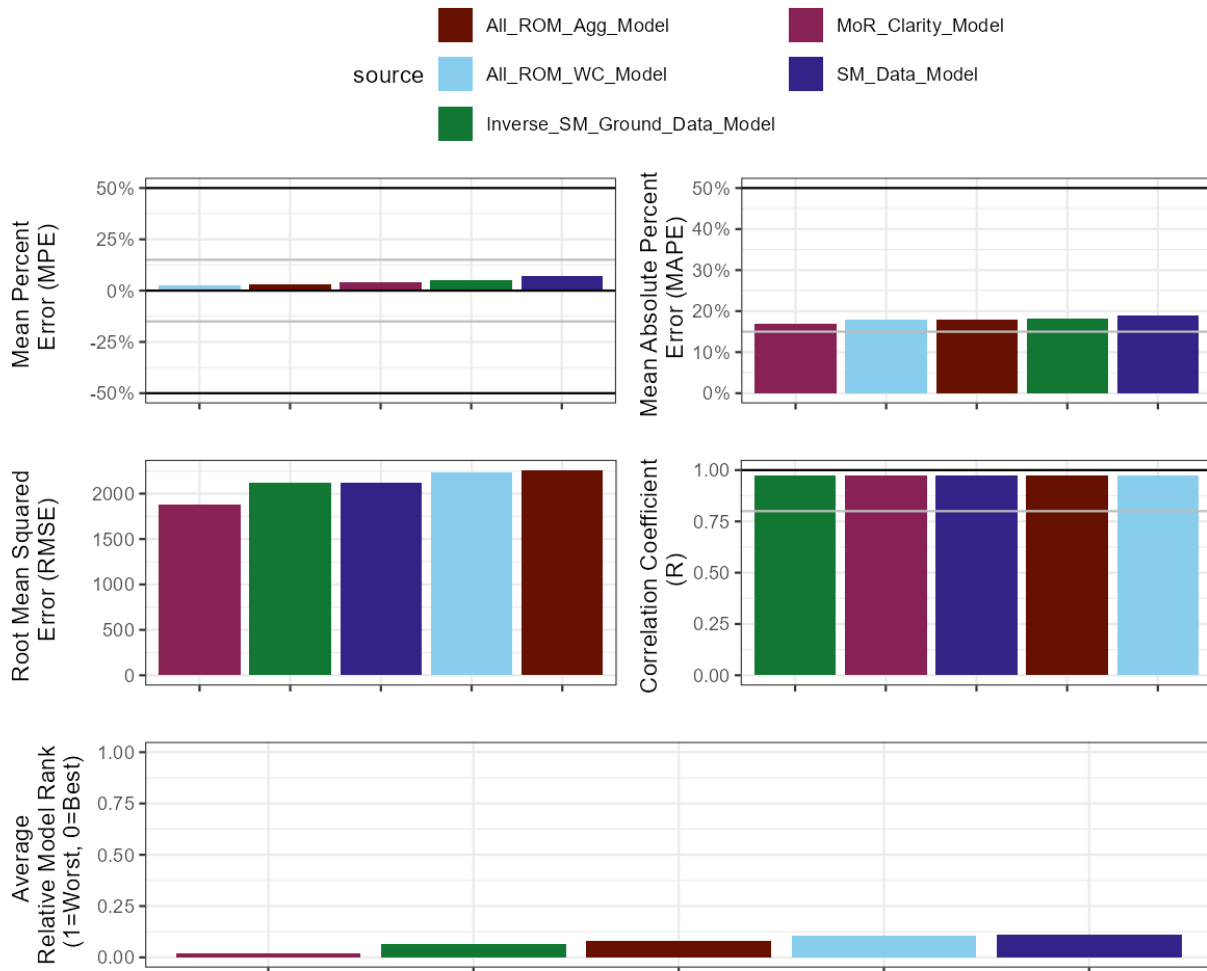
Model Performance Results for 06_Early_Stu_DeadAlive_Model - Forfar Creek - Ground Set [17]



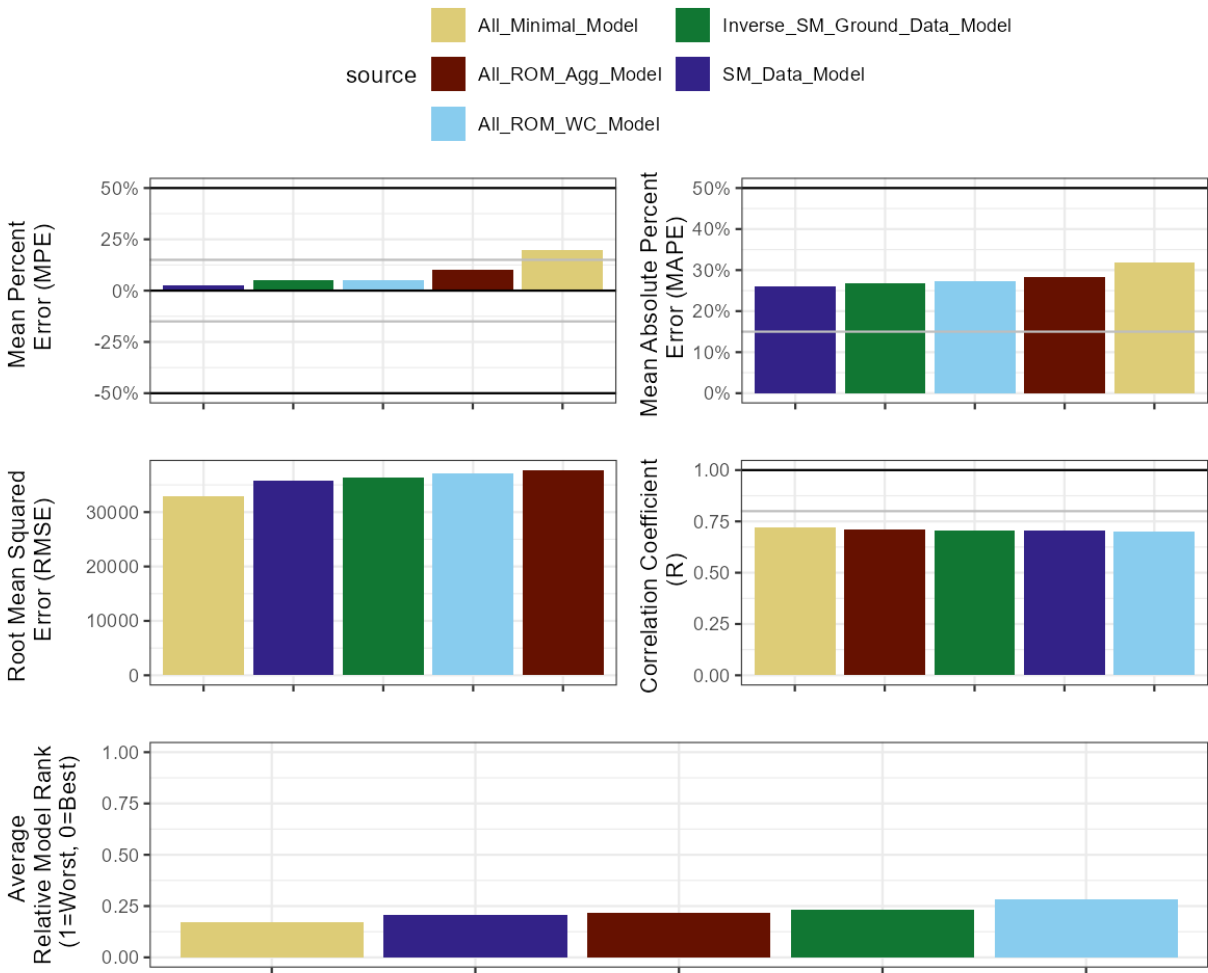
Model Performance Results for 05_Individual - Stellako River - Ground Set [21]



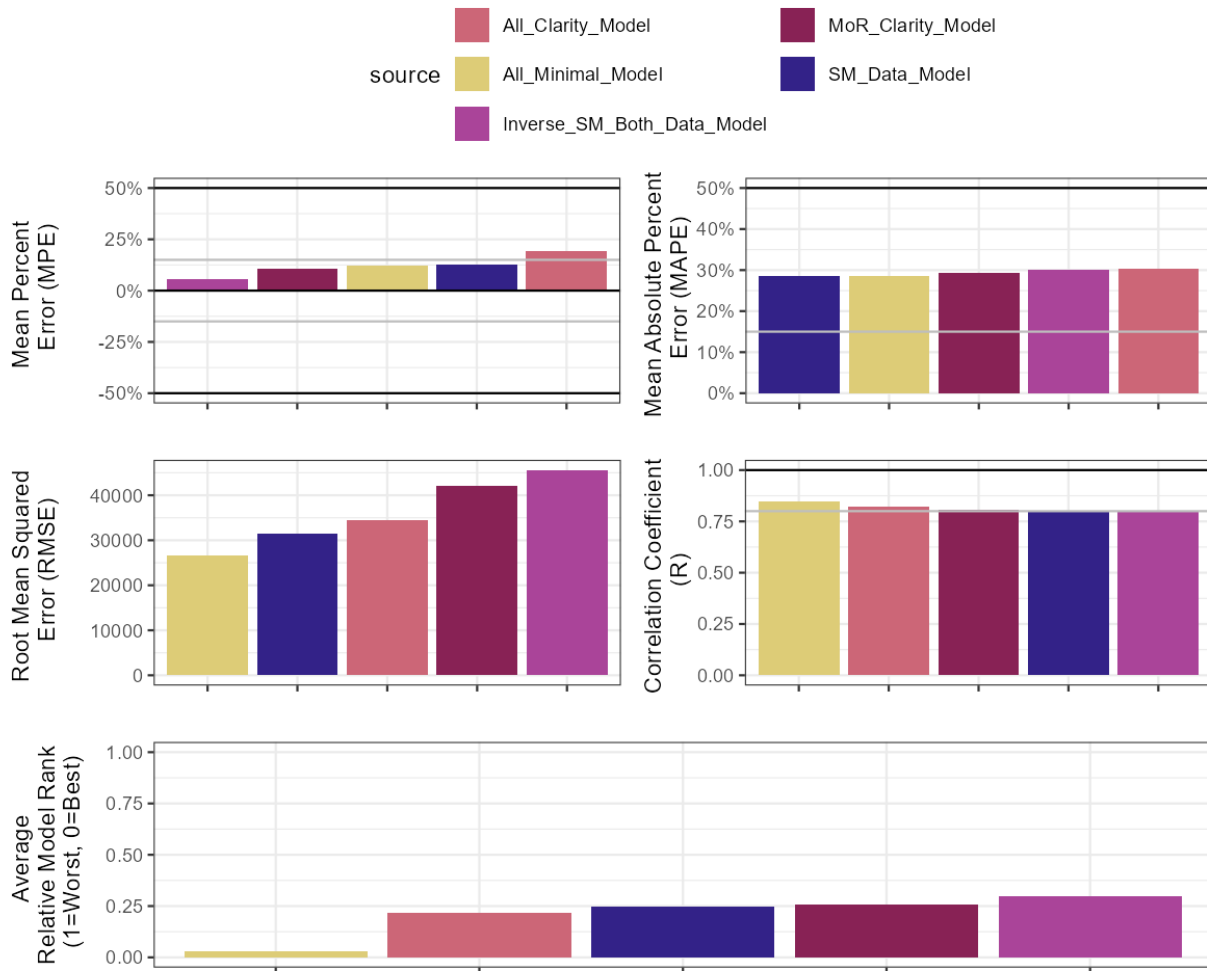
Model Performance Results for 04_SizeClarity_Specific - Small_Clear - Ground Set [56]



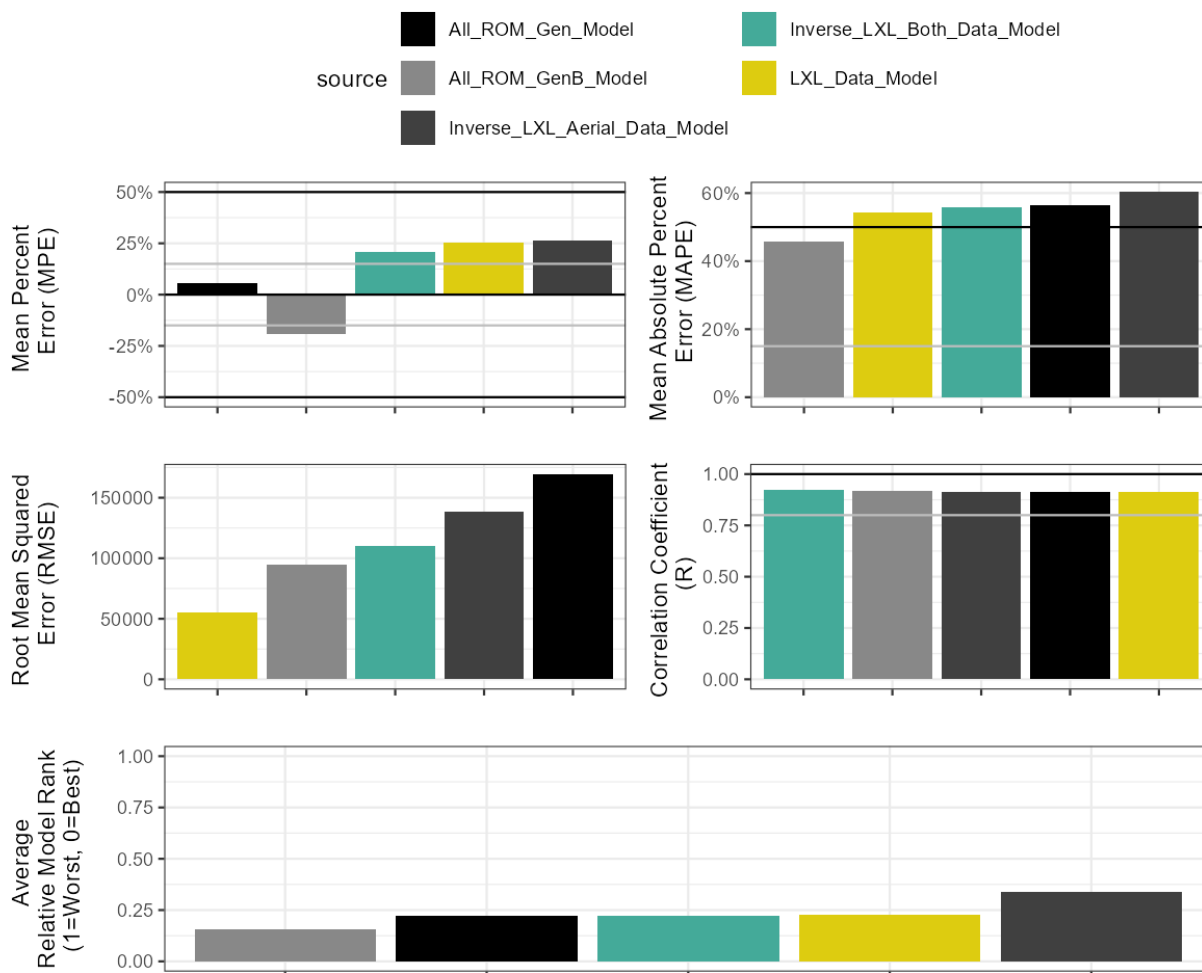
Model Performance Results for
04_SizeClarity_Specific - Medium_Clear - Ground Set [25]



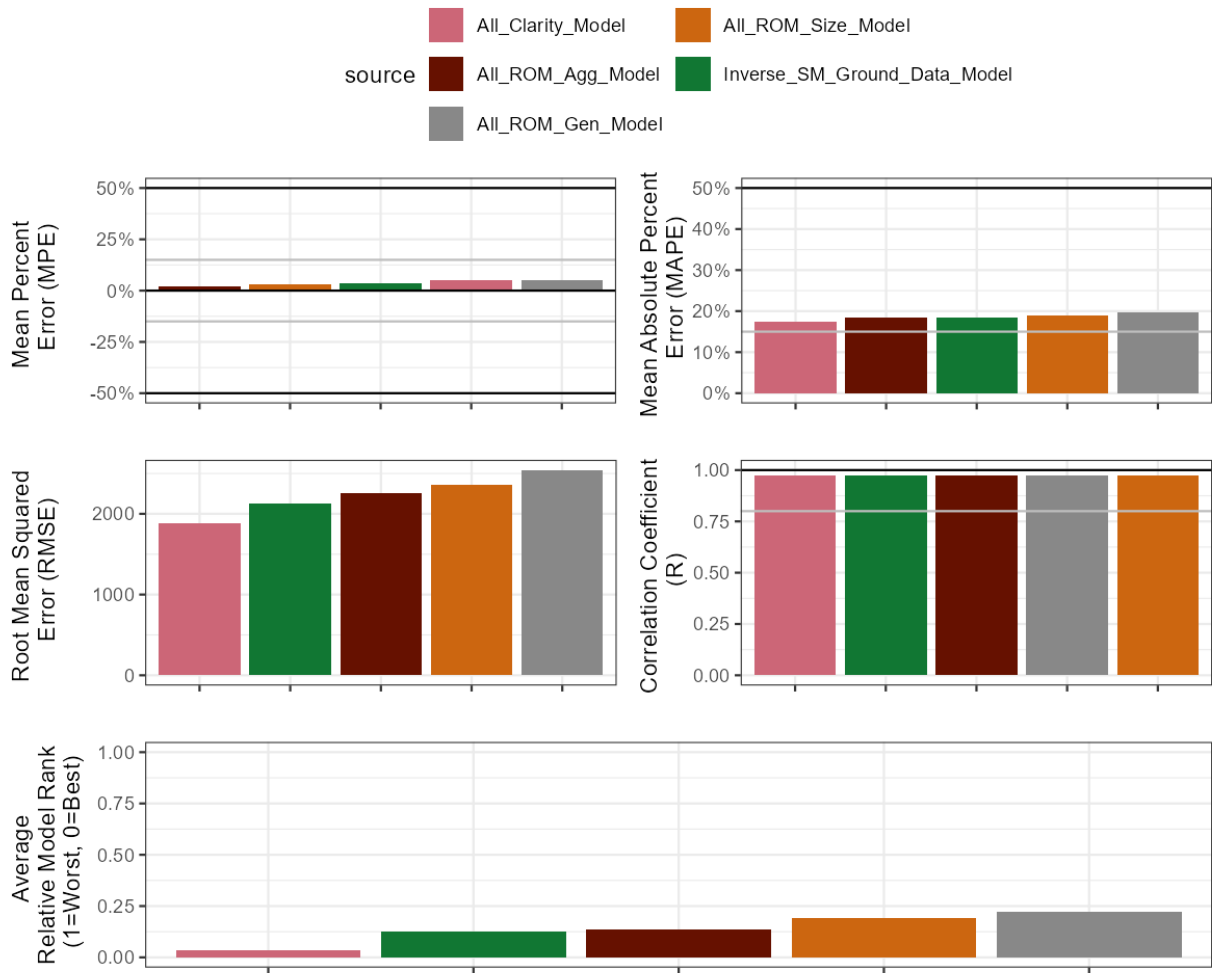
Model Performance Results for 04_SizeClarity_Specific - Medium_Clear - Aerial Set [25]



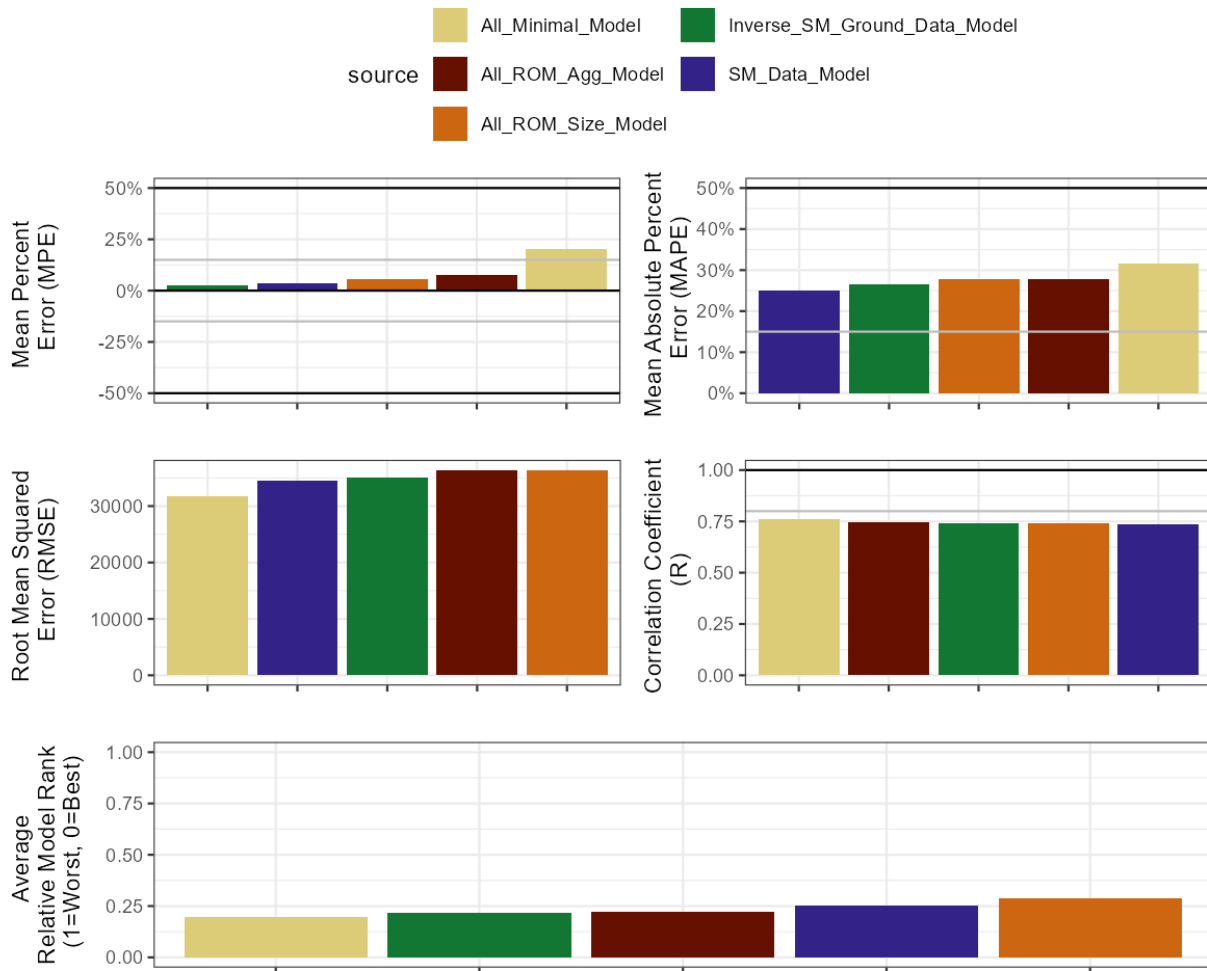
Model Performance Results for 03_Size_Specific - X-Large - Aerial Set [27]



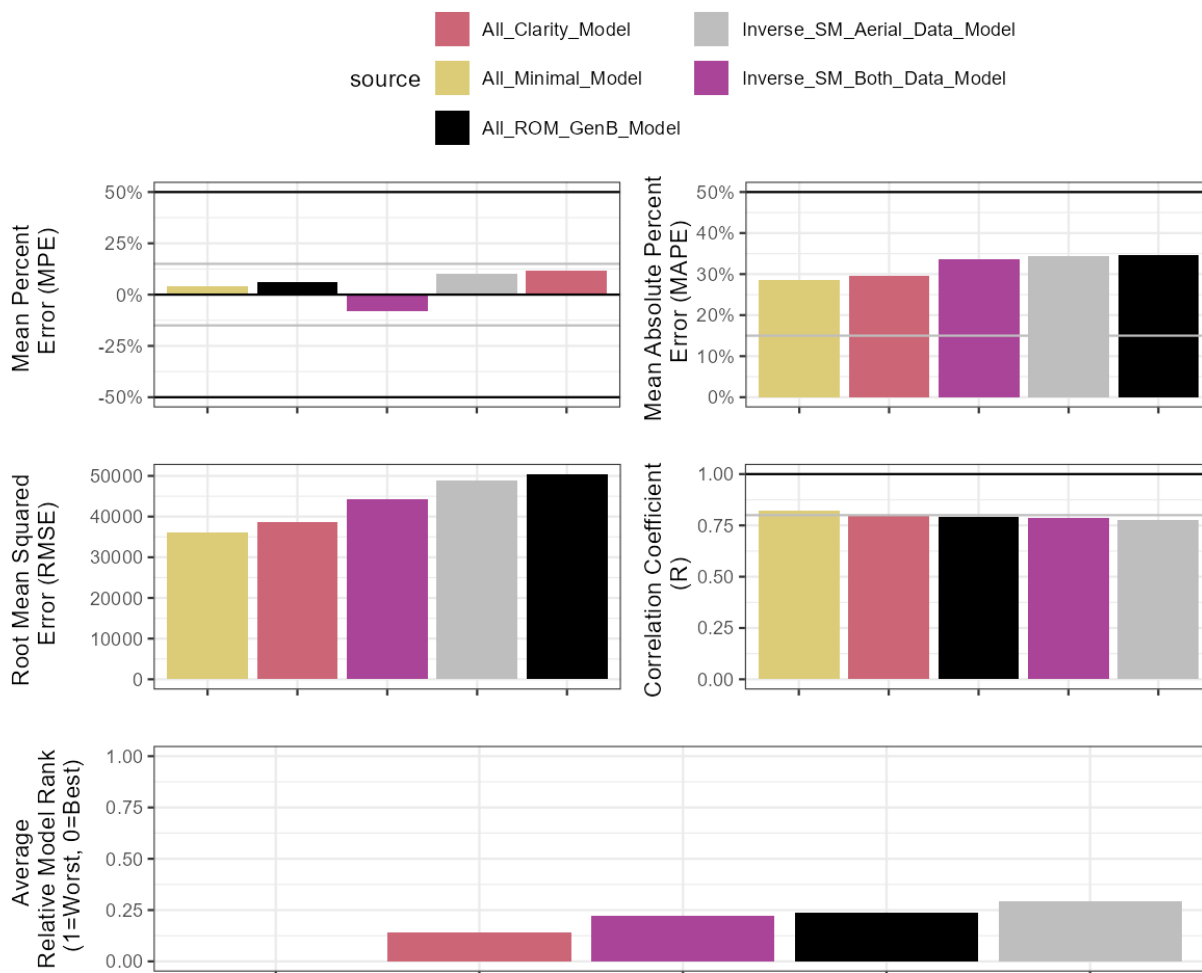
Model Performance Results for
03_Size_Specific - Small - Ground Set [58]



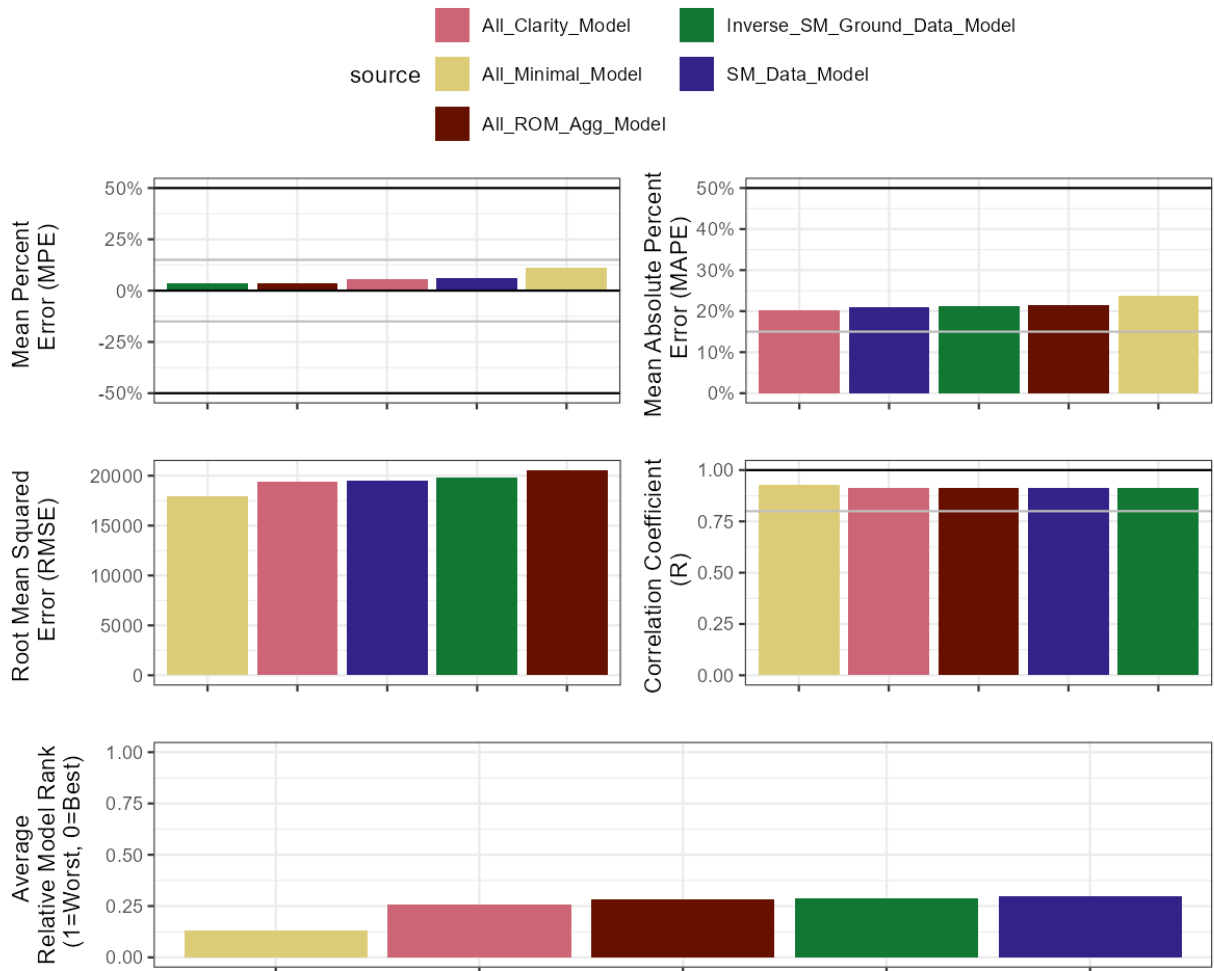
Model Performance Results for 03_Size_Specific - Medium - Ground Set [27]



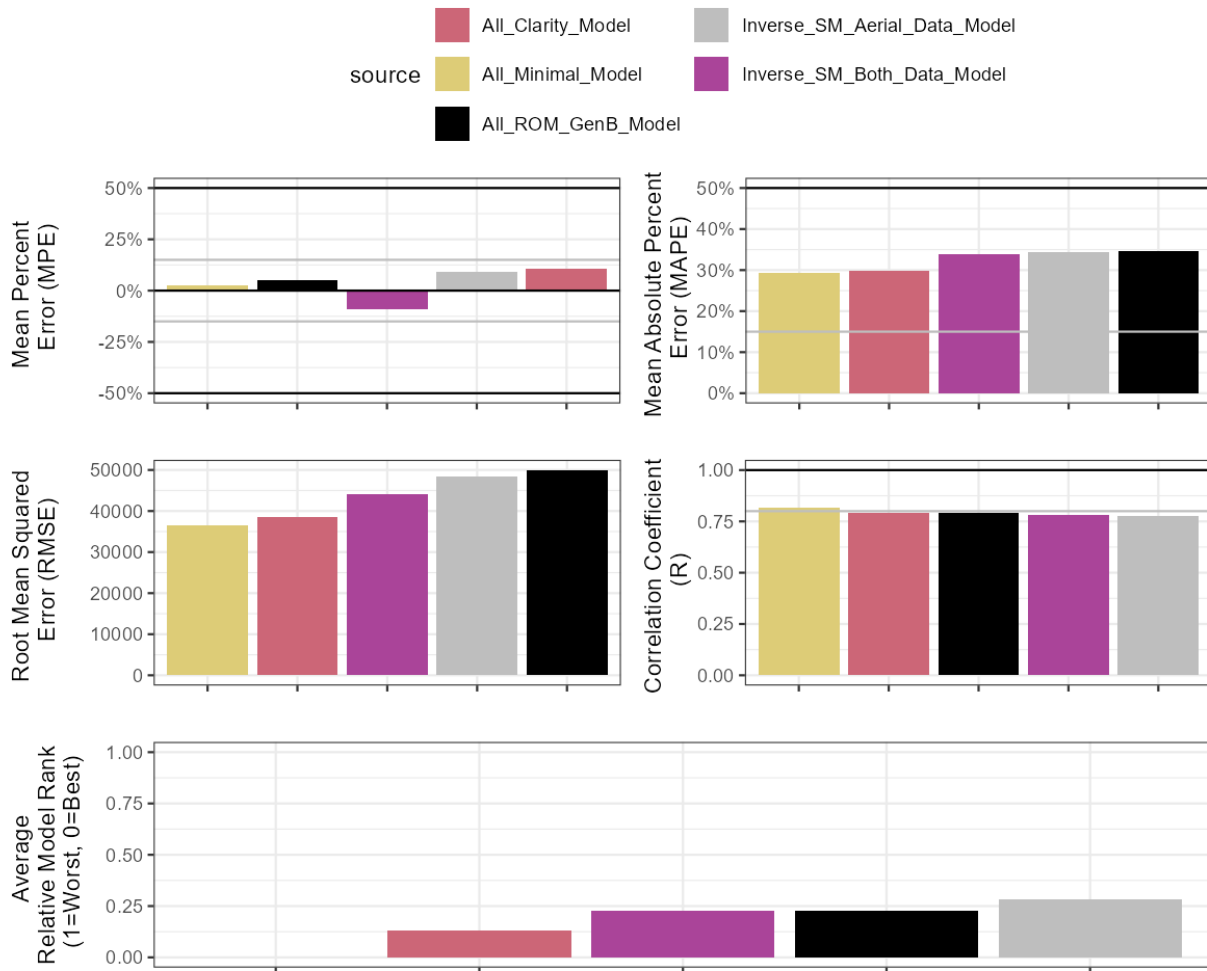
Model Performance Results for 03_Size_Specific - Medium - Aerial Set [38]



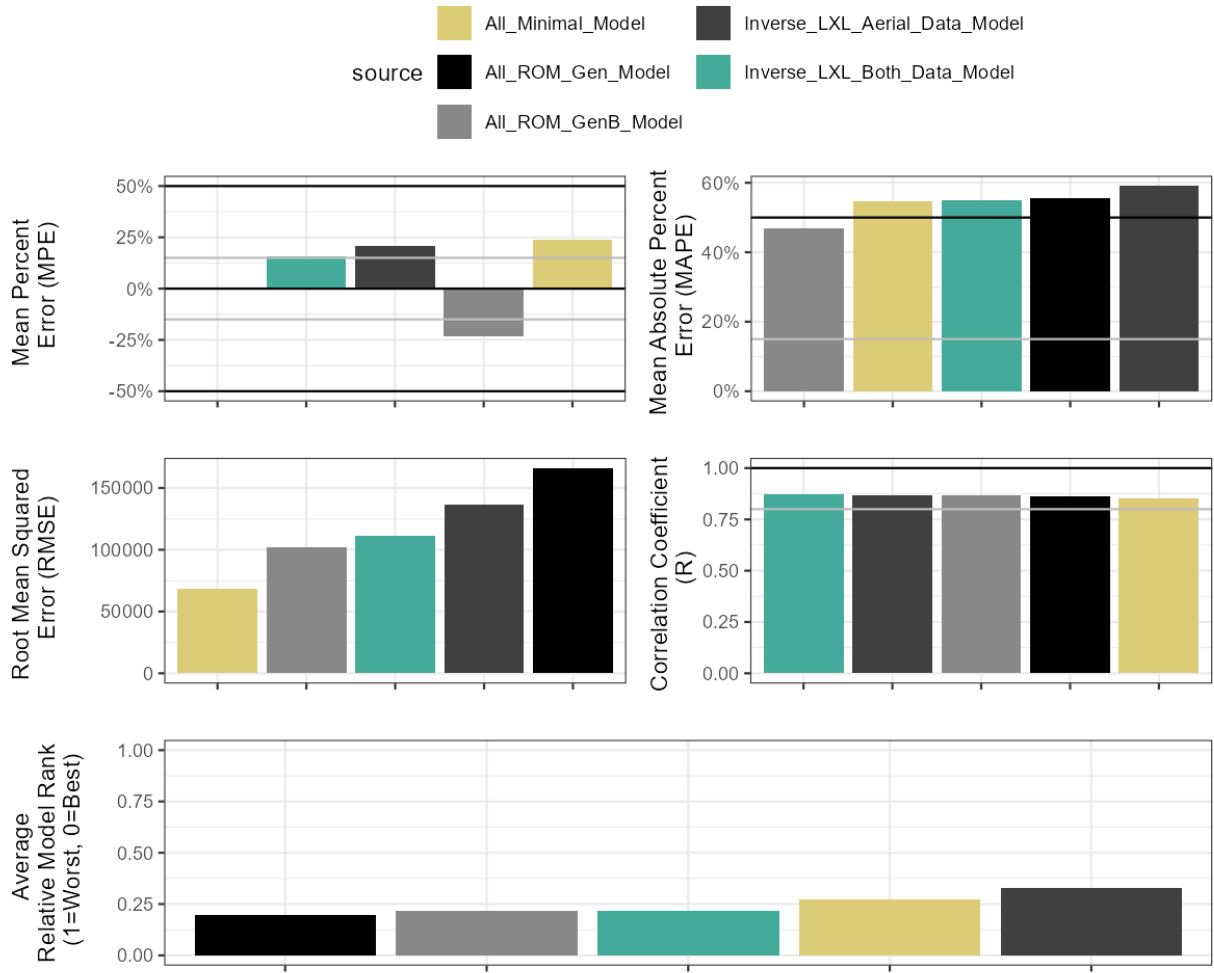
Model Performance Results for 02_AggregateSize_General - SM - Ground Set [85]



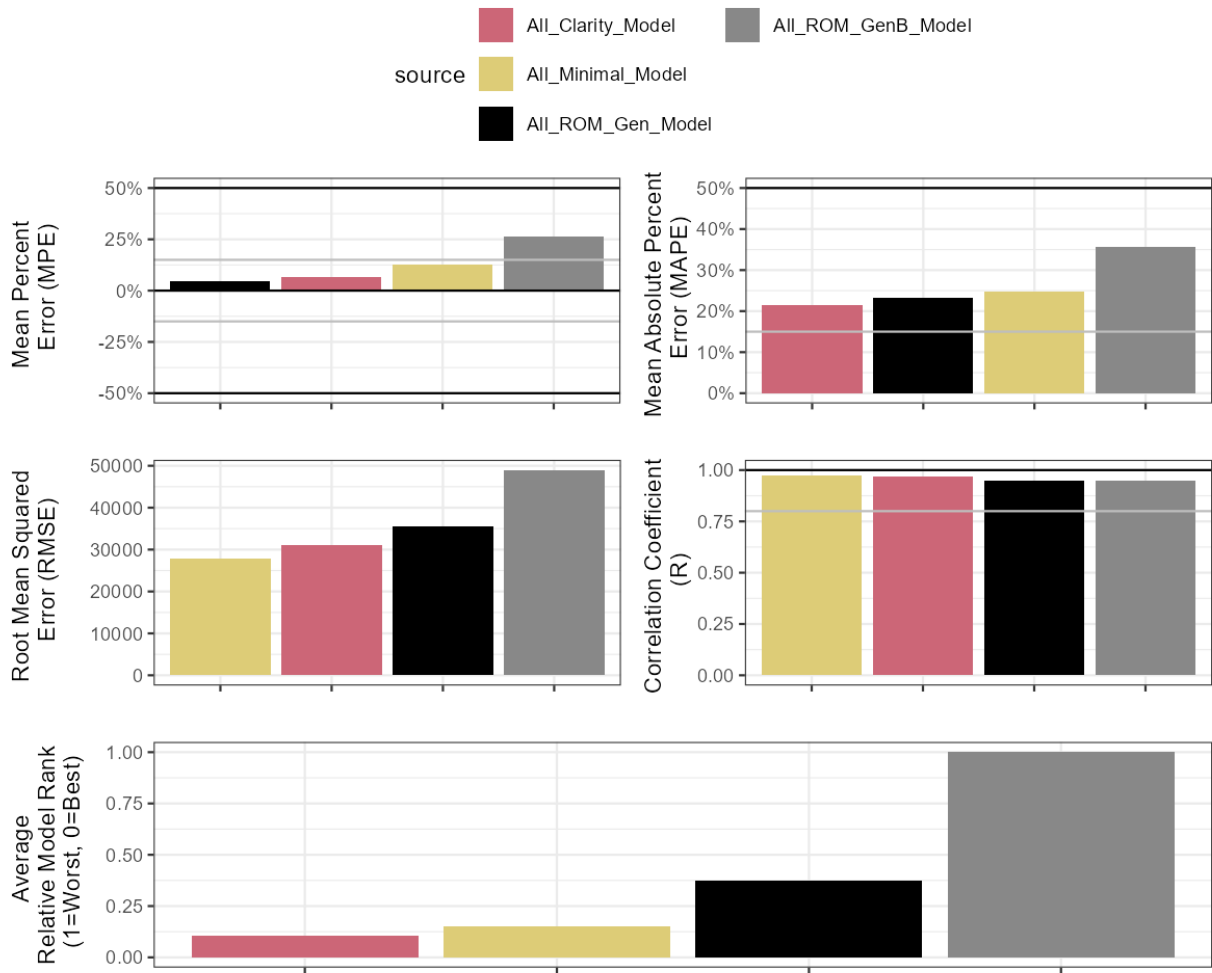
Model Performance Results for 02_AggregateSize_General - SM - Aerial Set [39]



Model Performance Results for 02_AggregateSize_General - LXL - Aerial Set [30]



Model Performance Results for
01_All_General - All - Ground Set [93]



Model Performance Results for
01_All_General - All - Aerial Set [69]

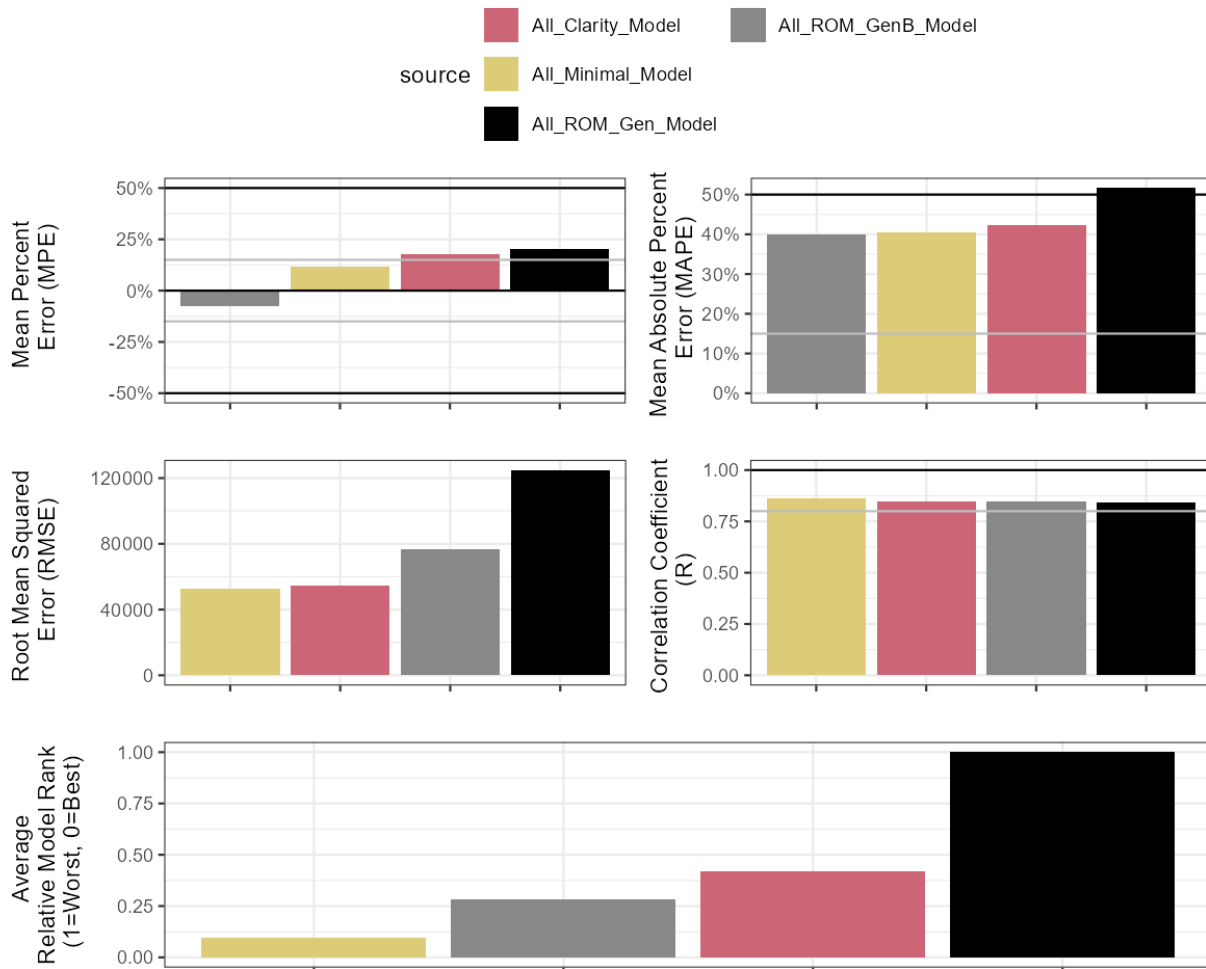


Table 25. T-test p-value results of the mean percent error (MPE) and mean absolute percent error (MAPE) comparisons between: the updated and default expansion results and the updated or default metric results relative to them being zero (a perfect expansion). Comparisons are organized by abundance bins based on historical definitions that determined where visual estimate methods would be planned (<25,000 prior 2004 and < 75,000 in 2004 and after). Comparisons are also further made within abundance bins between water clarities (clear and turbid/tannic) in addition to combining those levels (combined water clarity). *n* is the sample size.

Abundance Bin	Water Clarity	n	Metric	Updated Relative to Default	Updated Relative to Zero	Default Relative to Zero
[0-25,000)	Clear	53	MPE	0.016	0.236	0.0043
			MAPE	0.003	< 0.0025*	< 0.0025*
	Turbid/Tannic	7	MPE	< 0.0025*	0.293	0.0573
			MAPE	0.495	0.0208	< 0.0025*
	Combined	60	MPE	0.541	0.107	0.0754
			MAPE	0.003	< 0.0025*	< 0.0025*

[25,000-75,000)	Clear	30	MPE	< 0.0025*	0.423	0.0183
			MAPE	0.628	< 0.0025*	< 0.0025*
	Turbid/Tannic	13	MPE	< 0.0025*	0.553	< 0.0025*
			MAPE	0.164	< 0.0025*	< 0.0025*
	Combined	43	MPE	< 0.0025*	0.674	< 0.0025*
			MAPE	0.697	< 0.0025*	< 0.0025*
[75,000+]	Clear	45	MPE	< 0.0025*	0.159	< 0.0025*
			MAPE	< 0.0025*	< 0.0025*	< 0.0025*
	Turbid/Tannic	14	MPE	< 0.0025*	0.393	< 0.0025*
			MAPE	0.748	< 0.0025*	< 0.0025*
	Combined	59	MPE	< 0.0025*	0.0982	< 0.0025*
			MAPE	0.009	< 0.0025*	< 0.0025*
All	Combined	162	MPE	< 0.0025*	0.841	< 0.0025*
			MAPE	< 0.0025*	< 0.0025*	< 0.0025*

Table 26. Likelihood Ratio Test (LRT) results for temporal trend regressions for mean percent error (MPE) and mean absolute percent error (MAPE). Slope and standard error are either reported for the singular temporal trend term or by abundance bin (0-24,999; 25,000-74,999; and 75,000+) if the model included an interaction with abundance. LRT p -value < 0.001 indicates that the slope or interaction effect was a significant addition to the model.

Metric	Expansion Method	Model	LRT p-value	R ²	Slope (0-25; 25-75; 75+)	SE (0-25; 25-75; 75+)
MPE	Default	Temporal General	< 0.001	0.089	-0.011	0.003
MPE	Default	Interaction w/Abundance	< 0.001	0.182	-0.01; -0.004; -0.004	0.006; 0.009; 0.009
MPE	Updated	Temporal General	0.052	0.017	-0.005	0.003
MPE	Updated	Interaction w/Abundance	0.222	0.027	0.002; -0.001; -0.01	0.006; 0.01; 0.01
MAPE	Default	Temporal General	< 0.001	0.077	0.007	0.002
MAPE	Default	Interaction w/Abundance	0.107	0.097	0.001; 0.007; 0.005	0.004; 0.006; 0.006
MAPE	Updated	Temporal General	< 0.001	0.084	0.007	0.002
MAPE	Updated	Interaction w/Abundance	0.029	0.121	0.005; 0.008; 0.003	0.004; 0.006; 0.006

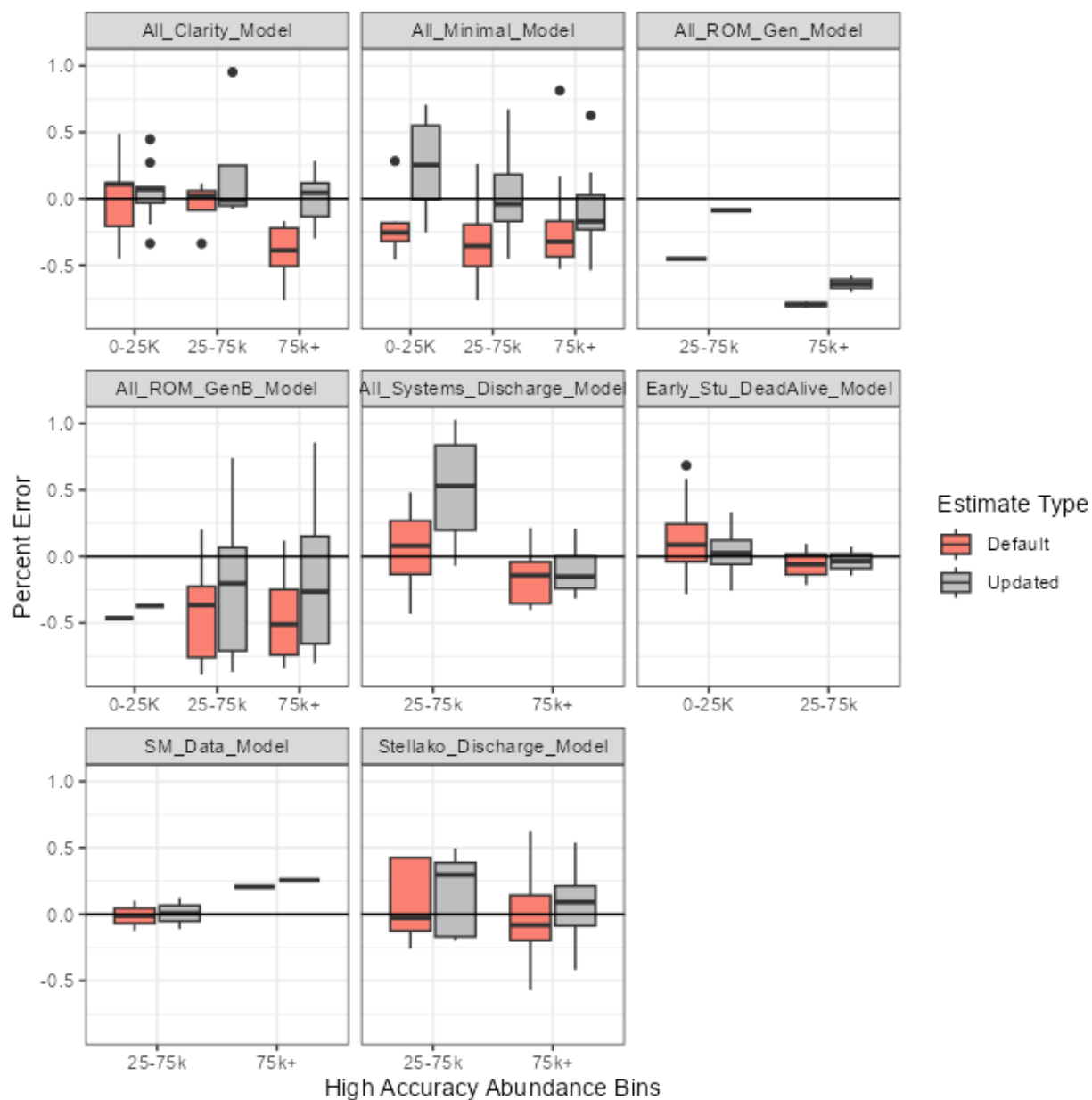


Figure 30. Boxplots of the percent error of the updated (grey) and default (salmon) expansions within the calibration dataset after the model selection and application, separated by models and high accuracy abundance bins.

APPENDIX D. OTHER EXPANSIONS, INFILLS, AND DATA CONSIDERATIONS

Fraser Sockeye Stock Assessment applies summary rules for approximating sex ratio and spawning success for spawning sites and populations (CUs) with insufficient carcass recovery samples. Each year, sex id and female spawning success data collected from each spawning site are reviewed to evaluate spatial and temporal representation and whether the sample size is adequate. Minimum sample sizes range 20-50 depending on spawner abundance and how extensive the spawning area is. Other factors considered include high water events, high turbidity, survey frequency or predators reducing the availability of carcasses for recovery. Spawning site-specific estimates of sex ratio and spawn success are always preferred and aggregate estimations are only used due sparse data.

The infill rules depend on if there are distinct geographical sub-aggregates within a CU or if they are disaggregated or single sites. Sub-aggregate estimates of sex ratio and spawning success are calculated as abundance-weighted means for CUs that have distinct geographical sub-aggregates. Depending on the abundance within the CU in a given year, sub-aggregates may need to be further aggregated to provide sufficient deadpitch data. For example, large escapements (and greater distribution of spawners across spawning sites) to the Quesnel-S CU may justify using all five aggregates listed below (Table 27) to estimate a data deficient spawning site. Conversely, during a low abundance year there may be too few carcasses recovered in any single pre-defined aggregate, therefore a Quesnel-S total aggregate would be applied. When a high accuracy program is run that estimates sex explicitly (e.g. mark-recapture), the sex-ratio from the high accuracy method is used for the CU because it will be less biased than deadpitch samples. For all spawning sites where values for an aggregate cannot be substituted (i.e., the sample is insufficient for the CU aggregate itself), sex ratio and spawning success values are based on long term averages or naive assumptions (i.e., 50% males, 50% females, 0% jacks; 100% female spawning success) in data limited CUs.

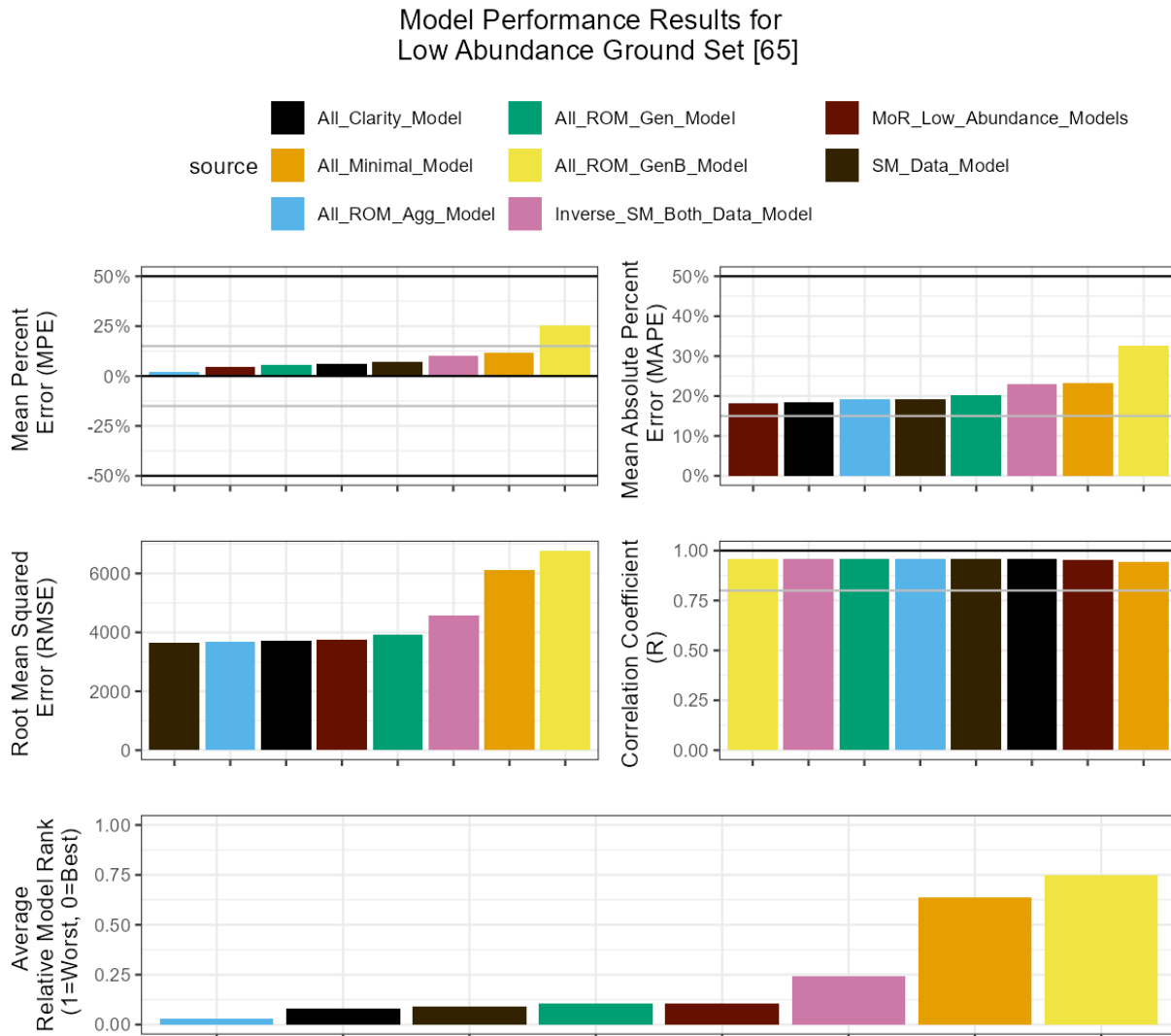
Table 27. Sub-aggregate groupings within Conservation Units and Timing Groups used for infilling biological data.

Timing Group	Conservation Unit	Sub-Aggregate
Summer	Quesnel-S	Mitchell River System; Horsefly River System; Quesnel Lake Shoreline; Quesnel Lake North Arm Tributaries Quesnel Lake East Arm Tributaries
Late	Shuswap-L	Eagle River system; Lower Shuswap River System; Shuswap Lake Shoreline; Adams River / Adams Lake.
Early Summer	Shuswap-ES	Adams River / Adams Lake; Eagle River; Seymour River System.
Early Stuart	Takla-Trembleur- EStu	Takla Lake Tributaries; Trembleur Lake Tributaries; Middle River Tributaries .
Summer	Takla-Trembleur- Stuart-S	Middle River System; Tachie River System; Stuart Lake Tributaries.

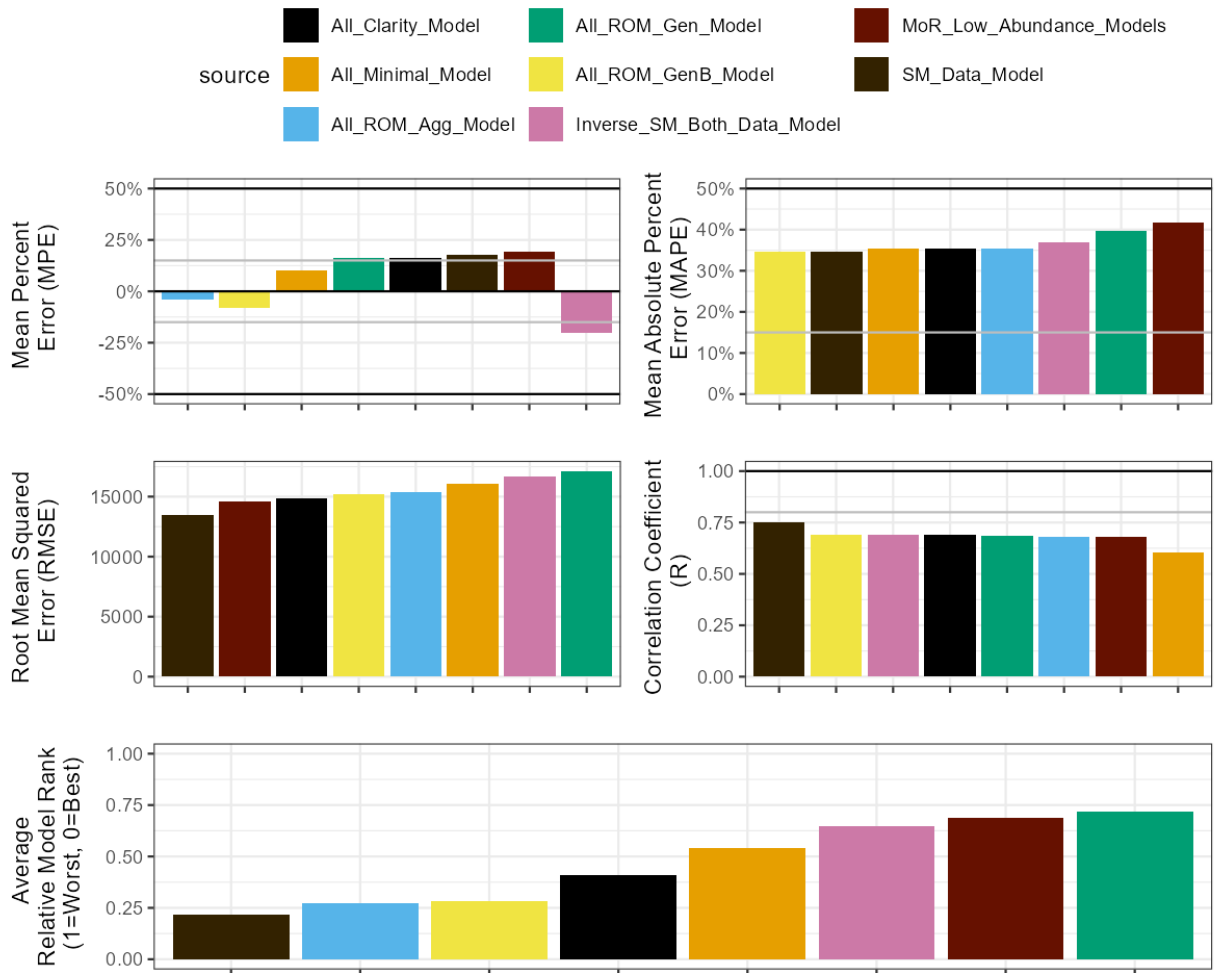
Summer	Chilko-ES/S	Chilko Lake North Shoreline Chilko Lake South Shoreline Chilko River
--------	-------------	--

APPENDIX E. FIGURES USED IN OUT-OF-SAMPLE RANGE MODEL IMPLEMENTATION DISCUSSIONS

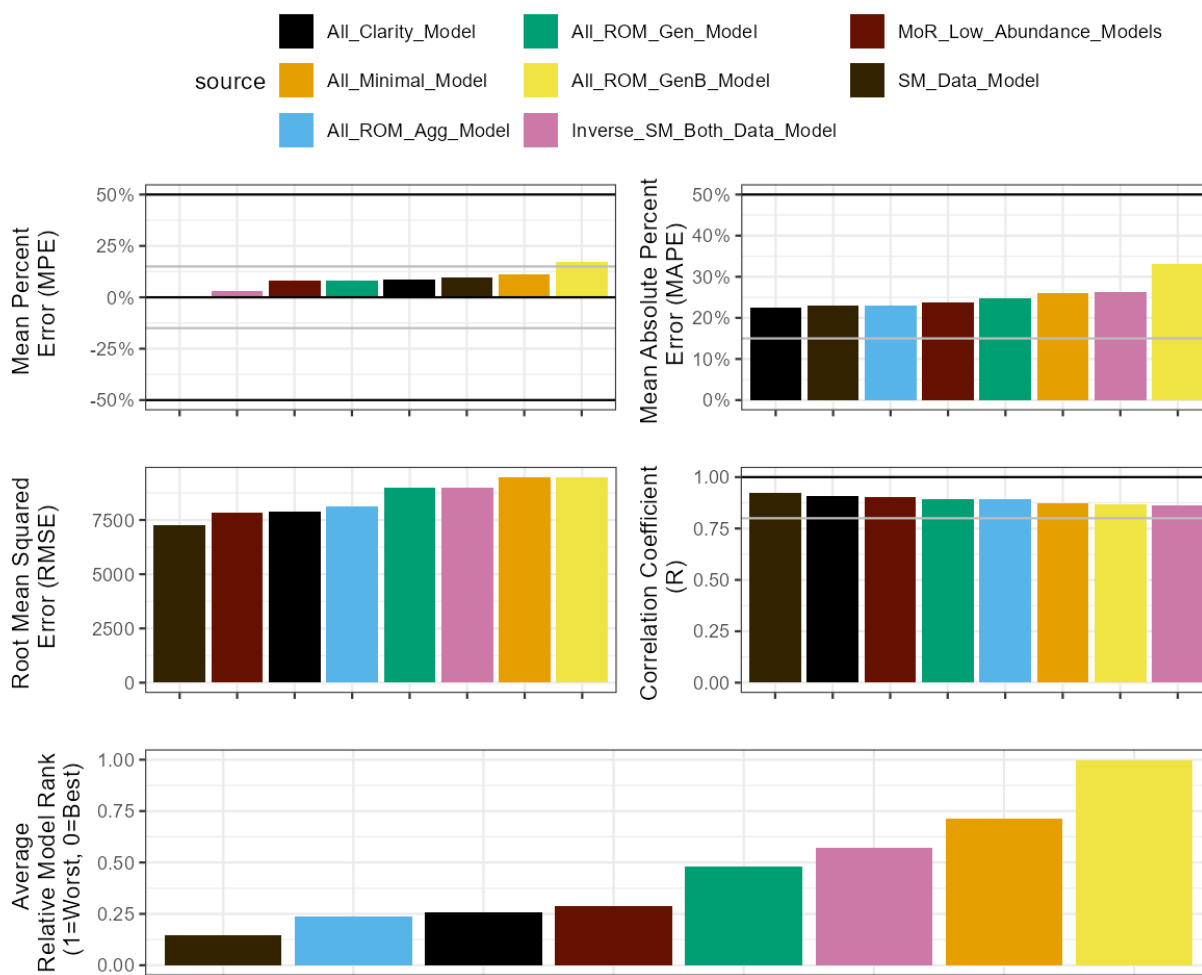
Each dashboard shares the same panel structure, so figure captions will only be descriptive of the rank-set level. Each dashboard has a title, legend, five bar plots, and one scatter plot. The title indicates the numerical specialization level, the general name of the rank-set (e.g. specialized model name followed by the system name it is applied to, or the Size if it is a set-level 3), the survey method type (ground or aerial), and the sample size in the rank-set (in square brackets). The legend indicates the various models that were compared in the rank-set. The models are subset to only the top five models within each rank-set, so some dashboards do not show all model results. The bar plots are the performance metrics (MPE, MAPE, RMSE, Correlation Coefficient, and Average Relative Model Rank), sorted by descending model score within that rank. Note that MPE is plotted but the absolute MPE determined the relative rank for that metric.



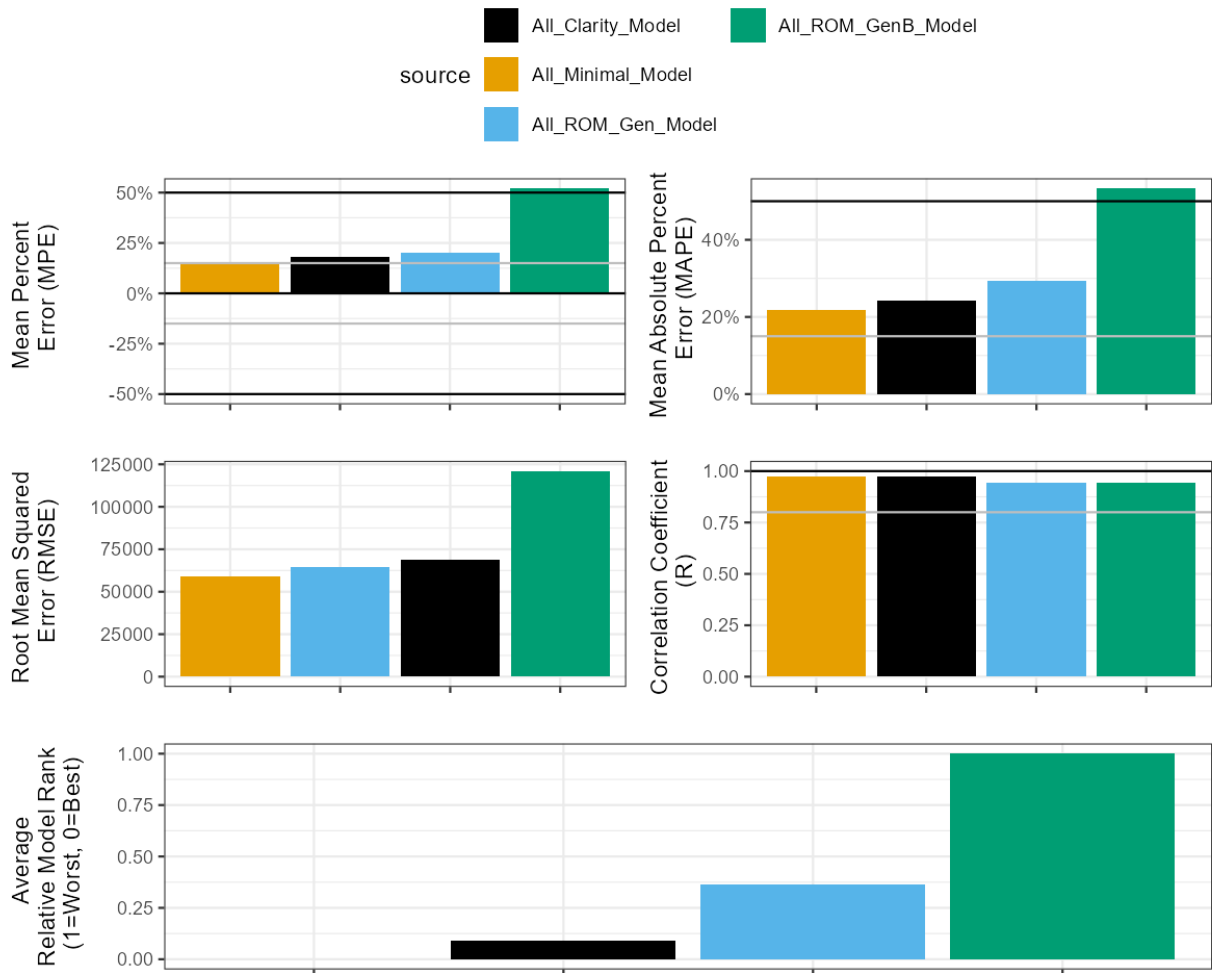
Model Performance Results for Low Abundance Aerial Set [20]



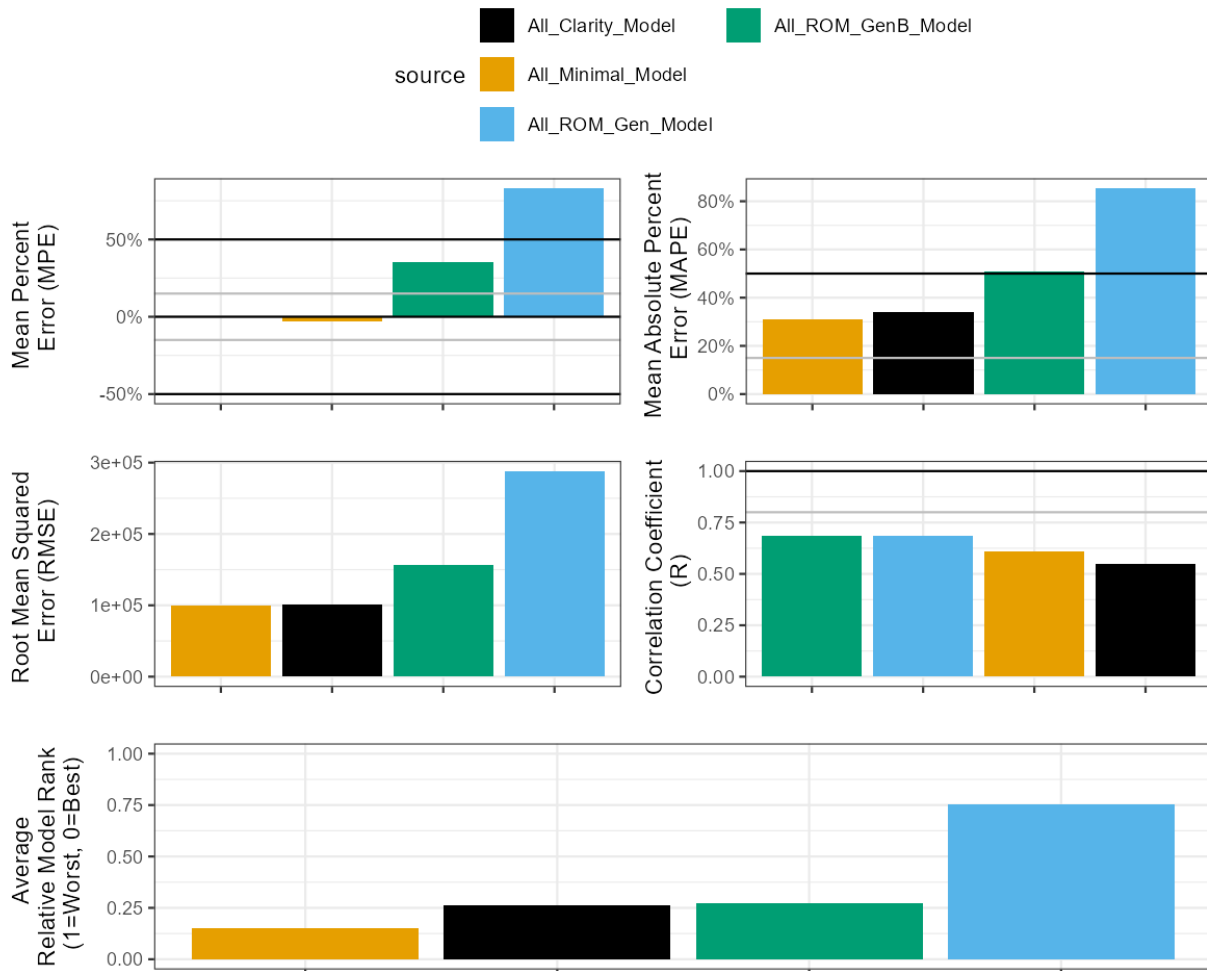
Model Performance Results for Low Abundance Both Set [85]



Model Performance Results for High Abundance Ground Set [11]



Model Performance Results for High Abundance Aerial Set [11]



Model Performance Results for High Abundance Both Set [22]

