A COMPREHENSIVE SIMULATION STUDY OF A CLASS OF ANALYSIS METHODS FOR PAIRED-TOW COMPARATIVE FISHING EXPERIMENTS

Yihao Yin and Hugues P. Benoît

Fisheries and Oceans Canada
Bedford Institute of Oceanography
Dartmouth, NS
B2Y 4A2

2022

Canadian Technical Report of
Fisheries and Aquatic Sciences 3466

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. II 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 3466

2022

A COMPREHENSIVE SIMULATION STUDY OF A CLASS OF ANALYSIS METHODS FOR PAIRED-TOW COMPARATIVE FISHING EXPERIMENTS

by

Yihao Yin[1] and Hugues P. Benoît[2]

[1] Fisheries and Oceans Canada
Bedford Institute of Oceanography
Dartmouth, NS
B2Y 4A2
[2] Fisheries and Oceans Canada
Maurice Lamontagne Institute
Mont-Joli, QC
G5H 3Z4

Cat. No. Fs97-6/3466E-PDF ISBN 978-0-660-42186-5 ISSN 1488-5379

Correct citation for this publication:

Yin, Y. and Benoît, H.P. 2022. A Comprehensive Simulation Study of A Class of Analysis Methods for Paired-Tow Comparative Fishing Experiments. Can. Tech. Rep. Fish. Aquat. Sci. 3466: vi + 99 p.

CONTENTS

ABSTRACT

Yin, Y. and Benoît, H.P. 2022. A Comprehensive Simulation Study of A Class of Analysis Methods for Paired-Tow Comparative Fishing Experiments. Can. Tech. Rep. Fish. Aquat. Sci. 3466: vi + 99 p.

Bottom-trawl surveys provide fishery-independent information key to assessments for groundfish and some shellfish stocks worldwide. When a change in survey protocol is necessary, comparative fishing experiments are often conducted for a required calibration of catchability to maintain consistency. These experiments often involve paired trawling of the former and replacement vessel, gear and protocol. Fisheries and Oceans Canada (DFO) is undertaking comparative fishing in each of its six Atlantic bottom-trawl surveys in 2021 and 2022 to calibrate two new offshore fisheries survey vessels that will replace two retiring longstanding vessels. A suite of existing statistical models are proposed for the analysis of the resulting data. This report presents a comprehensive simulation study to test the performance of these models and evaluate their statistical qualities. This includes a realistic simulation based on resampling of data from past DFO Maritimes region bottom-trawl surveys and additional simulations designed to assess specific issues associated with the sampling design for comparative fishing experiments and validation of the simulation study itself. Results from the simulations are thoroughly investigated and discussed. Throughout we provide recommendations for the planning of comparative fishing experiments and guidance for the analyses in practice.

RÉSUMÉ

Yin, Y. and Benoît, H.P. 2022. A Comprehensive Simulation Study of A Class of Analysis Methods for Paired-Tow Comparative Fishing Experiments. Can. Tech. Rep. Fish. Aquat. Sci. 3466: vi + 99 p.

Les relevés au chalut de fond fournissent des renseignements indépendants de la pêche essentiels aux évaluations de stock des poissons de fond et certains stocks de crustacés et mollusques à plusieurs endroits au monde. Lorsqu'un changement de protocole de relevé est nécessaire, des expériences de pêche comparatives sont souvent menées pour un étalonnage de la capturabilité afin de maintenir la cohérence des séries temporelles. Ces expériences impliquent souvent le chalutage en paire de l'ancien et du navire de remplacement, et de l'engin de pêche et du protocole propre à chacun. Pêches et Océans Canada (MPO) entreprend une pêche comparative dans chacun de ses six relevés au chalut de fond de l'Atlantique en 2021 et 2022 pour calibrer deux nouveaux navires de relevés halieutiques hauturiers qui remplaceront deux navires de longue date qui seront retirés. Une série de modèles statistiques existants est proposée pour l'analyse des données résultantes. Ce rapport présente une étude de simulation complète pour tester les performances de ces modèles et évaluer leurs qualités statistiques. Cela comprend une simulation réaliste basée sur le rééchantillonnage des données des relevés au chalut de fond antérieurs du MPO dans la région des Maritimes et des simulations supplémentaires conçues pour évaluer les problèmes spécifiques associés au plan d'échantillonnage pour les expériences de pêche comparatives et la validation de l'étude de simulation elle-même. Les résultats des simulations sont minutieusement étudiés et discutés. Tout au long, nous fournissons des recommandations pour la planification de expériences de pêche comparatives et conseils pour les analyses en pratique.

1 Introduction

Bottom-trawl surveys provide key inputs to stock assessments for groundfish and some shellfish stocks worldwide. These surveys can produce annual indices of abundance that are proportional to stock size, provided that the proportionality constant, often called catchability, does not change over time. If this consistency is not achieved via proper sampling design and standardization, then there is a risk that changes in abundance will be confounded with changes in catchability. Maintaining consistency in survey protocols, and the survey vessel and gear, is key to maintaining constant catchability. However, periodically it becomes necessary or desirable to change one or more of these aspects and calibration experiments are required to estimate adjustments for possible changes in catchability. The most common and effective form of these experiments is comparative fishing, which usually involves paired trawling of the former and replacement vessel/gear/protocol as close together as safety permits. This design minimizes the difference in fish densities sampled by the trawls, such that differences in catches over replicates of paired-trawl sampling will reflect the difference in catchability. Fisheries and Oceans Canada (DFO) is undertaking comparative fishing in each of its six Atlantic bottom-trawl surveys in 2021 and 2022 to calibrate two new offshore fisheries survey vessels that will replace two retiring longstanding vessels. In some surveys, the change in vessel will also be accompanied by a change in survey trawl and survey procedures (e.g., tow duration), and the joint effect of all of these factors on relative catchability should be reflected in results of the comparative fishing experiments.

There are numerous analytical approaches available for estimating relative catchability from comparative fishing data. The sophistication and flexibility of these methods has improved considerably since DFO first began undertaking comparative fishing (e.g., Fanning 1985; Koeller and Smith 1983), and even since the last major round of experiments that took place in Atlantic Canada in 2004-2005 (e.g., Benoît 2006; Cadigan et al. 2006; Bourdages et al. 2007). Notably, recent methods involve the use of random effects to account for key sources of variability and non-parametric modeling of fish length-dependent catchability functions (Miller 2013; Thygesen et al. 2019; Cadigan et al. submitted). In principle, these methods provide enhanced robustness to variability in comparative fishing data and better modeling of key factors affecting catchability. However, these methods and their predecessors have been subjected to limited simulation testing necessary to confirm robustness and statistical efficiency. Some studies have employed forms of self-testing, where data are simulated from the model and the ability of the model to recover parameters is assessed (e.g., Cadigan and Dowden 2010; Thygesen et al. 2019) or somewhat simple parametric simulations aimed at choosing model options (e.g., Lewy et al. 2004; Cadigan et al. submitted). However, none of the modeling approaches to our knowledge have been subjected to extensive testing involving structured independent simulation. This is required to evaluate the robustness of models to mis-specification, which is an important consideration given that the “true” process that generates catches in a comparative fishing experiment is unknown and a matching statistical model cannot simply be chosen.

In this report, we present results of independent simulation testing of a suite of statistical models of varying complexity for comparative fishing data (Miller 2013). These models were notably applied in the analysis of extensive comparative fishing that took place in the northeast USA in the 2000s (Miller et al. 2010) and appear well suited to a variety of comparative fishing data. We begin by describing the estimation models including their mathematical development, the

model fitting (parameter estimation) procedure, and the model selection method. We then motivate and describe a realistic simulation framework based on resampling of data from past Maritimes research vessel (RV) bottom-trawl survey and a set of metrics designed to evaluate the performance of these estimation models. The majority of the report covers this suite of simulation tests. Following the interpretation of results from this simulation study, we provide a more detailed discussion on the differences between the candidate models, and an overview of general analytical methods for comparative fishing paired catches, in order to guide analysis in practice. The final sections of the report address specific issues associated with the sampling design for comparative fishing experiments, and a brief exploration of parametric alternatives for simulation testing which led to us adopting a non-parametric simulation approach. Throughout we provide recommendations for the planning of comparative fishing experiments and the analyses of the resulting data.

2 Estimation Models

A suite of models are presented in this section that are proposed for analysis of paired catches and estimation of relative catch efficiency between the pair of gears in the comparative fishing experiment. These models are developed and described in detail in Miller (2013).

2.1 Binomial Models

To estimate the relative catch efficiency between a pair of gears (for simplicity, "gear" in this study refers to a vessel-gear combination with a specific survey protocol) in the comparative fishing, we assume the expected catch from gear g ($g \in \{A, B\}$) at length L and at station i is

$$E[C_{gi}(L)] = q_{gi}(L)D_{gi}(L)f_{gi}, \tag{1}$$

where $q_{gi}(L)$ is the catchability of gear g, $D_{gi}(L)$ is the underlying population density sampled by gear g, and f_{gi} is a standardization term which usually includes the swept area of a tow and if applicable, the proportion of sub-sampling for size measurement on-board. In a binomial model, catch from gear A at station i conditioning on the combined catch from both gears within this station, $C_i(L) = C_{Ai}(L) + C_{Bi}(L)$, is binomial-distributed

$$C_{Ai}(L) \sim BI(C_i(L), p_{Ai}(L)), \tag{2}$$

where $p_{Ai}(L)$ is the expected proportion of catch from gear A. Paired tows are assumed to sample the same underlying density, $D_{Ai}(L) = D_{Bi}(L) = D_i(L)$, hence the logit-probability of catch by gear A is

$$logit(p_{Ai}(L)) = log(\frac{E[C_{Ai}(L)]}{E[C_{Bi}(L)]}) = log(\rho_i(L)) + o_i. \tag{3}$$

This gives the conversion factor, $\rho_i(L)$, as the ratio of catchabilities between gear A and B at length L and at station i,

$$\rho_i(L) = q_{Ai}(L)/q_{Bi}(L), \tag{4}$$

where $o_i = log(f_{Ai}/f_{Bi})$ is an offset that can be derived from known standardization terms of the survey tows.

For a length-based conversion factor, we consider a smooth length effect based on a regression spline approximation,

$$log(\rho(L)) = \sum_{k=0}^{K} \beta_k X_k(L) = \mathbf{X}^T \beta, \tag{5}$$

where β are the spline coefficient parameters and are estimated, $\mathbf{X}$, or $\{X_k(L), k = 0, 1, \cdots, K\}$, are a set of smoothing basis functions, and K is the dimension of the basis which controls the number of coefficient parameters and is usually pre-defined. In this study, we use the cubic spline smoother (Friedman et al. 2001), and the basis functions and penalty matrices can be defined correspondingly.

The estimation of a cubic spline smoother is based on the penalized sum of squares smoothing objective but in practice this is usually replaced by a penalized likelihood objective (Green and Silverman 1993),

$$\mathcal{L}(\beta, \lambda) = f(\mathbf{Y}|\mathbf{X}, \beta) e^{-\frac{\lambda}{2}\beta^T \mathbf{S} \beta}, \tag{6}$$

where $\mathcal{L}$ is the likelihood objective function, $f(\mathbf{Y}|\mathbf{X}, \beta)$ is the joint probability function of the survey data $\mathbf{Y}$ conditional on the basis functions and coefficient parameters, $\mathbf{S}$ is the penalty matrix defined by the smoother and the dimension of the basis, and λ is the smoothness parameter. This smoothness parameter is estimated by maximum likelihood along with other model parameters but may be sensitive to the data and in such cases it can be determined by other criteria such as generalized cross-validation (Wood 2000).

The penalized maximum likelihood smoother can also be re-parameterized into a mixed effects model (Verbyla et al. 1999; Wood 2017),

$$log(\rho_i(L)) = \mathbf{X}_f^T \beta_f + \mathbf{X}_r^T \mathbf{b}, \tag{7}$$

where β_f are fixed effects and $\mathbf{b}$ are random effects. $\mathbf{X}_f$ and $\mathbf{X}_r$ are transformed from the basis functions $\mathbf{X}$ and an eigen-decomposition of the penalty matrix $\mathbf{S}$, $\mathbf{X}_f = \mathbf{U}_f^T \mathbf{X}$ and $\mathbf{X}_r = \mathbf{U}_r^T \mathbf{X}$, where $\mathbf{U}_f$ and $\mathbf{U}_r$ are the eigenvectors that correspond to the zero and positive eigenvalues of $\mathbf{S}$. The random effects $b \sim \mathrm{N}(0, \mathbf{D}_+^{-1}/\lambda)$ where D_+ is the diagonal matrix of the positive eigenvalues of S. In the mixed effects model representation of the cubic spline smoother, the number of fixed effects is 2 and the number of random effects is bounded by $K - 2$. Smoothing effects are transformed into shrinkage of random effects in the fitting of random deviations, and can be integrated into complex mixed effects models commonly used in fisheries science (Thorson and Minto 2015).

Additional random effects can be incorporated into the mixed effects model to address variations in the relative catch efficiency among stations,

$$log(\rho_i(L)) = \mathbf{X}_f^T (\beta_f + \delta_i) + \mathbf{X}_r^T (\mathbf{b} + \epsilon_i), \tag{8}$$

where $\delta_i \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\epsilon_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}_+^{-1}/\xi)$. From a similar re-parameterization of the cubic spline smoother, these random effects allow for deviations of the length-based conversion at each station. $\mathbf{\Sigma}$ is the covariance matrix of the random effects corresponding to the random deviations δ_i and contains three parameters. ξ controls the degree of smoothness of the random smoother and the smoother at each station can differ.

A summary of the above binomial mixed model is as follows,

$$\begin{cases} C_{Ai}(L) \sim BI(C_i(L), p_{Ai}(L)) \\ C_i(L) = C_{Ai}(L) + C_{Bi}(L) \\ logit(p_{Ai}(L)) = log(\rho_i(L)) + o_i \\ log(\rho_i(L)) = \mathbf{X}_f^T(\beta_f + \delta_i) + \mathbf{X}_r^T(\mathbf{b} + \epsilon_i). \end{cases} \tag{9}$$

The model is estimated via maximum marginal likelihood and the marginal likelihood is obtained by integrating out random effects,

$$\mathcal{L}(\beta_f, \mathbf{\Sigma}, \lambda, \xi) = \int \left(\prod_{i=1}^{m} \int \int f(\mathbf{Y}_i | \mathbf{X}_f, \mathbf{X}_r, \beta_f, \mathbf{b}, \delta_i, \epsilon_i) f(\delta_i | \mathbf{\Sigma}) f(\epsilon_i | \xi) \mathrm{d}\delta_i \mathrm{d}\epsilon_i \right) f(\mathbf{b} | \lambda) \mathrm{d}\mathbf{b}. \tag{10}$$

The binomial mixed model can take various assumptions on the smoother and station variation to accommodate different underlying density of a species and data limitations especially in length measurements. A set of binomial models are presented in Table 1.

2.2 Beta-Binomial Models

The binomial assumption of the catch can be extended to a beta-binomial distribution to explain over-dispersion at the stations,

$$C_{Ai}(l) \sim BB(C_i(L), p_{Ai}(L), \phi_i(L)). \tag{11}$$

The beta-binomial distribution is a compound of the binomial distribution and a beta distribution. More specifically, it assumes a beta-distributed random effect in the expected proportion of catch from gear A across stations. As a result, the expected catch by gear A has a variance of

$$var(C_{Ai}) = C_i p_i (1 - p_i) \frac{\phi_i + C_i}{\phi_i + 1}, \tag{12}$$

where ϕ is the variance parameter that is estimated to capture the over-dispersion of a binomial variable, or extra-binomial variation.

The same smoothing length effect can be applied to this over-dispersion parameter,

$$log(\phi_i(L)) = \mathbf{X}_f^T \gamma + \mathbf{X}_r^T \mathbf{g}, \tag{13}$$

where γ are fixed effects and $\mathbf{g}$ are random effects, $\mathbf{g} \sim \mathrm{N}(0, \mathbf{D}_+^{-1}/\tau)$. This length effect in the variance parameter models heterogeneity and can improve model projection of uncertainty.

A summary of the beta-binomial mixed model is as follows,

$$\begin{cases} C_{Ai}(L) \sim BB(C_i(L), p_{Ai}(L), \phi_i(L)) \\ C_i(L) = C_{Ai}(L) + C_{Bi}(L) \\ logit(p_{Ai}(L)) = log(\rho_i(L)) + o_i \\ log(\rho_i(L)) = \mathbf{X}_f^T(\beta_f + \delta_i) + \mathbf{X}_r^T(\mathbf{b} + \epsilon_i) \\ log(\phi_i(L)) = \mathbf{X}_f^T \gamma + \mathbf{X}_r^T \mathbf{g}. \end{cases} \tag{14}$$

The marginal likelihood is

$$\mathcal{L}(\beta_f, \gamma, \mathbf{\Sigma}, \lambda, \xi, \tau) = \int\int \left(\prod_{i=1}^{m} \int\int f(\mathbf{Y}_i|\mathbf{X}_f, \mathbf{X}_r, \beta_f, \mathbf{b}, \gamma, \mathbf{g}, \delta_i, \epsilon_i) f(\delta_i|\mathbf{\Sigma}) f(\epsilon_i|\xi) \mathrm{d}\delta_i \mathrm{d}\epsilon_i \right) f(\mathbf{b}|\lambda) f(\mathbf{g}|\tau) \mathrm{d}\mathbf{b} \mathrm{d}\mathbf{g}. \tag{15}$$

Likewise, various smoothing assumptions can be applied to the variance parameter. Table 2 presents a set of beta-binomial mixed models.

2.3 Model Implementation and Selection

In the analysis of a comparative fishing experiment, the conversion factor can be developed for a user-specified length range. The length range is usually selected as the minimum to the maximum observed length from the comparative fishing experiment. Rare and extremely large or small individuals may be excluded from the analysis to avoid possible disproportional impact from these outliers.

The binomial and beta-binomial models in Tables 1 and 2 are implemented in `TMB` (Kristensen et al. 2015) which are compiled into objective functions and subsequently optimized in R (R Core Team 2020). The basis functions for the cubic smoothing spline and the corresponding penalty matrices are generated using the R package `mgcv` (Wood 2011) based on 10 equally-spaced knots ($K = 9$) within the pre-specified length range depending on the species and the comparative fishing survey. `TMB` automatically calculates a standard error for the maximum likelihood estimation of the conversion factor via the delta method (Kristensen et al. 2015) and this can be used to develop a confidence interval for inference.

There are in total 13 candidate models for estimating the conversion factor. The best model for each species and each comparative fishing survey is selected by AIC (Akaike information criterion) to maximize model fitting, while avoiding over-fitting of more complicated models especially in cases without adequate data. In practice, the estimated relative catch efficiency (i.e., conversion factor) from all converged models should be compared along with the sample proportions to provide a more appropriate interpretation of results.

3 Method for the Simulation Study

The survey catch of a species depends on the population density at the sampling site and the selectivity of the gear. Simulation of a comparative fishing experiment should consider both the true relative catch efficiency (a.k.a., conversion factor) and the underlying species population density. Assumptions for the relative catch efficiency can be either direct, with a parametric form for the conversion factor, or indirect, from assumed catchabilities for each of the two gears. The effect of fish size (length effect) and potential variations related to station locations and resulting from small scale variation in density (station effect) should be considered, as gear selectivity naturally varies according to the size of fish, and conditions at a specific tow location may also affect gear performance. The underlying population density is more difficult to simulate accurately as a species can distribute unevenly in both space and time and population levels also vary greatly for different sizes or age classes. In addition, the sampling process that

generates catch numbers given population density and gear selectivity, i.e., via bottom trawl, can be highly unpredictable and the type of variations is difficult to categorize. For these reasons, using explicit assumptions for the underlying population density and catch distribution during the sampling process is likely to lead the simulation model to mis-represent reality. In order to generate realistic catch numbers for paired comparative fishing tows and to minimize possible mis-specification due to lack of understanding of the actual catch process, we simulate survey catches based on a resampling of past Maritimes RV surveys.

A description of methods for the simulation study is detailed in this section. We designed a variety of simulation scenarios to reflect those that might be encountered during the forthcoming comparative fishing experiments. Each scenario consists of a data simulation step, a model estimation step using the proposed suite of models in Section 2, and after iterations of this simulation-estimation cycle, an assessment step. The assessment includes a set of metrics to summarize the simulated datasets in a scenario, and a set of performance measures for evaluating model estimation results. The process of the simulation study is illustrated in Figure 1.

Sections 6 and 7 include additional simulation studies for selected discussion topics, e.g., related to sample size and catch distribution. The additional simulation studies use similar methods as described in this section but these studies explore alternative simulation assumptions that are designed for specific purposes. Specific methodological details associated with those particular simulation studies are described in the relevant sections to avoid confusion with the general methodology presented here for the main analyses.

3.1 Simulation Step

3.1.1 Species Considered

In this simulation study, we consider catch characteristics for five archetypal species chosen to explore a variety of scenarios representing different population levels, spatial distributions, and size compositions (Maritimes region species codes and latin names in parenthesis):

- Silver hake (14, *Merluccius bilinearis*): abundant in catches over the survey area
- Haddock (11, *Melanogrammus aeglefinus*): abundant catches of both juvenile and adult size classes
- Redfish (23, *Sebastes spp.*): moderate catch quantities with limited spatial distribution
- Atlantic cod (10, *Gadus morhua*): infrequent small catches, sparsely distributed
- Winter skate (204, *Leucoraja ocellata*): extremely low population level with sporadic catches

3.1.2 Assumptions for Relative Catch Efficiency

Four assumptions are included in this study for generating the true relative catch efficiency (referred to as ρ-assumptions hereafter). We denote the old gear by A and the replacement

gear by B in a paired-tow comparative fishing experiment. The relative catch efficiency between gears A and B is denoted as $\rho(L)$ and equivalently, the proportion of catch by gear A within the pair of gears is $\mu(L)$ (see Equation 3). Below is a detailed description of these assumptions:

1. The first assumption is designed as a control group where the two gears have the same catchability. Hence, the true relative catch efficiency is 1 and invariant with respect to length and station,
$$\begin{cases} \mu(L) = 0.5 \\ \rho(L) = 1. \end{cases} \tag{16}$$

2. In the second assumption, two logistic size-selectivity curves, $q_A(L)$ and $q_B(L)$, are assumed for the two gears featuring different midpoints and steepness parameters (Hilborn and Walters 2013). The midpoints for gears A and B are set at $1/2$ and $1/3$, respectively, of the pre-specified maximum possible length (L_s^{max}) depending on the species (s), so that gear B has the ability to catch more small-sized fish. The relative catch efficiency is subsequently derived as the ratio of gear selectivities. This assumption simulates a case where there is a contrast in the length effect between juveniles and adults as observed in some species in past comparative fishing experiments, e.g., redfish in Miller (2013).
$$\begin{cases} q_A(L) = 1/(1 + \exp(-0.2(L - L_s^{max}/3))) \\ q_B(L) = 1/(1 + \exp(-0.1(L - L_s^{max}/2))) \\ \mu(L) = q_A(L)/(q_A(L) + q_B(L)) \\ \rho(L) = q_A(L)/q_B(L). \end{cases} \tag{17}$$

3. Instead of an explicit assumption for gear selectivity, the third scenario directly simulates the relative catch efficiency with a parametric function of length using the exponential model from Bourdages et al. (2007). This scenario represents a simple, yet realistic, case where the relative catch efficiency is a monotonic function with an asymptote toward large sizes, as could occur when the pair of gears are of the same type but with different mesh sizes. Parameters a_s, b_s and c_s (s denotes species) in the exponential model control the relative size-selectivity. For simulation, we use values extracted from model estimates for the 2004-2005 northern Gulf of St. Lawrence comparative fishing analysis (Bourdages et al. 2007) for each simulated species in this study (Table 3). Note that Bourdages et al. (2007) did not provide estimates for Atlantic cod; in this case, we use the parameter values for haddock considering similarity in the two species' length range.
$$\begin{cases} \mu(L) = 1 - 1/(\exp(a_s + b_s \exp(-c_s L)) + 1) \\ \rho(L) = \mu(L)/(1 - \mu(L)). \end{cases} \tag{18}$$

4. A more intricate fourth assumption takes into account the station-level variation. This variation may arise when gear performance is affected by conditions at the tow location such as bottom type. This effect can also compensate for within-pair differences in their sampled underlying fish population density due to fine-scale heterogeneity in distribution. In this scenario, the linear model from Bourdages et al. (2007) is used for $\mu(L)$ but includes a random deviation for variations relating to the station. The deviation is represented in the model parameters ϵ_i and ν_i (i denotes station), each simulated from a lognormal distribution. The lognormal distribution was chosen because it is a heavy-tailed statistical

distribution often used for generating samples with occasional extreme observations. Consequently, divergence from the true relative catch efficiency can be significantly large at some stations.

$$\begin{cases} \mu_i(L) = 1 - 1/(\exp(0.03L \cdot \epsilon_i - 2 \cdot \upsilon_i) + 1) \\ \rho_i(L) = \mu_i(L)/(1 - \mu_i(L)) \\ \epsilon_i \sim LN(0 - 0.2^2/2, 0.2^2) \\ \upsilon_i \sim LN(0 - 0.2^2/2, 0.2^2). \end{cases} \tag{19}$$

Figure 2 illustrates the simulated relative catch efficiency as a function of length for each of the ρ-assumptions 1, 2 and 4 within a length range of 10-100 cm. The third ρ-assumption simulates differently for each species, hence the simulated relative catch efficiency functions are presented separately in Figure 3 and for each species.

3.1.3 Assumptions for Paired Catches

The underlying population density and the catch numbers resulting from the sampling process can be difficult to simulate explicitly. However, the comparative fishing experiment follows standard sampling protocols designed to minimize systematic differences between tows in a pair. Most importantly, the two gears fish simultaneously and in parallel, separated by the shortest distance considered practical and safe. It is reasonable to assume the pair of gears sample a similar underlying population density, albeit with some random errors. This facilitates simulation of paired catches in relative terms without the need to specify the exact underlying population density in the survey area. In addition, we can circumvent the assumption of parametric catch distributions by utilizing historical survey catches in this area. Catch distributions (e.g., Poisson, negative binomial distributions) are often used in population abundance models to explain dispersion within a spatial region (or any defined stratification), but their ability to accurately capture catch variations is limited and sometimes problematic (Smith 1990; Greene 1994; Ver Hoef and Boveng 2007). In the simulation of paired comparative fishing tows, expected catch numbers by gear A is generated by resampling from catches by recent DFO RV survey in the Maritimes region (in 2015-2017) such that A is representative of CCGS *Alfred Needler* with a Western IIA trawl (denoted as Needler-WIIA hereafter). Expected catch by gear B is subsequently derived by applying the assumed relative catch efficiency. The resampling technique also retains length continuity in survey catches within the same tow. This continuity is a result of interaction between gear catchability, population size composition, and measurement errors; the level of continuity can otherwise be difficult to quantify accurately in a parametric simulation.

To provide a pool for resampling, we used data from the Maritimes bottom-trawl survey for 2015-2017, which includes 48 strata and 633 stations (number of stations for each stratum ranges between 5-30). The swept area for a tow is $0.04\ km^2$ on average with a standard deviation of $0.003\ km^2$. For the majority of tows and species, catch numbers are counted; however, when catches are large, numbers are estimated from sub-sampling, where a portion of fish is weighed and counted, then scaled against total catch weight to derive an estimate of a total catch number. Survey catches for the five species considered are presented in Figures 4-8 in order to show their levels and among-tow variation within the survey area. Length-specific mean catch number

per tow (mnpt) is also calculated for each species (Figure 9) to demonstrate their overall catch-length compositions from the Maritimes RV survey and for comparison with simulated datasets.

We explain the simulation step, or the data generating process, with an example of Atlantic Cod. Atlantic Cod caught in the Maritimes bottom-trawl survey during 2015-2017 ranged from 3 to 90 cm (lengths were recorded at an interval of 1 cm). For each station, survey catches are standardized by the sub-sampling ratio and swept area first; this sample catch-at-length is smoothed with a 5-point moving average to reduce sampling errors and then assumed to constitute the expected catch-at-length for gear A at that location, $E[C_A(L)]$. Assuming gear B samples the same underlying density, the expected catch-at-length by gear B at the same location is derived by applying the simulated true relative catch efficiency,

$$E[C_B(L)] = E[C_A(L)]/\rho(L). \tag{20}$$

This creates a pool of available stations for each stratum where a pair of tows can be potentially sampled by the gears according to a simulated comparative fishing station sampling design. For simplicity, we assume that the catch is completely enumerated and therefore do not simulate additional error associated with sub-sampling. For each selected station and for each gear, a random swept area is simulated based on a Gaussian distribution with a mean of $0.04\ km^2$ and standard deviation of $0.003\ km^2$, and a random multiplicative sampling error is assumed for generating catch numbers based on expected catch-at-length:

$$\begin{cases} C_{gi}(L) = E[C_{gi}(L)] \cdot f_{gi} \cdot \eta_{gi}(L) \\ f_{gi} \sim N(0.04, 0.003) \\ \eta_{gi}(L) \sim LN(1, \sqrt{log(\widehat{CV}^2 + 1)}). \end{cases} \tag{21}$$

where $C_{gi}(L)$ is the simulated catch of a species at station i by gear g and measured at length L, f_{gi} contains the standardization terms and here includes the swept area, and η_{gi} is the random sampling error, which is length-dependent and simulated based on an estimated coefficient of variation (CV, defined as the ratio between sample standard deviation and sample mean) from the Maritimes survey data (the estimated CV profile is denoted by $\widehat{CV}$).

The sampling error is inherent to every fisheries survey (Krause et al. 2002) and can be influenced by various factors related to the gear itself, the sampled species, etc. The lognormal distribution is suitable for modeling the sampling error in fisheries datasets (Dick 2004) and is used in this study for simulating the sampling error. This distribution includes two parameters, the mean and the variance. We assume there is no systemic bias in the sampling process, i.e., $E[\eta_{gi}(L)] = 1$, and we estimate the coefficient of variation from the Maritimes bottom-trawl survey for simulation of realistic variance parameters for these species. The CV is estimated as a function of length for each species. In order to improve estimation of the CV profile, the sample CV is calculated from survey catches for each year independently and then averaged over the years; assuming the survey sampling process is consistent in each year, this increases the sample size for CV estimation. The sample CV is then smoothed over length to increase precision of estimated CV (see Figure 10 for estimated CV profiles for the species considered). Eventually, the variance parameter can be derived from $\widehat{CV}$ (Equation 21). The CV informs the level of variation to be expected during the bottom-trawl sampling process and here effectively subsumes both observation error and process error caused by variability in gear performance resulting from factors such as differences in the personnel operating it and effects of bottom

habitat. More accurate estimation for the CV should consider strata differences; however, this was not possible for these species due to insufficient number of stations and survey catches. In addition, data insufficiency was extreme for some species and consequently, even length-dependency in $\widehat{CV}$ cannot be properly estimated for winter skate.

3.2 Iteration Process

There are five species and four assumptions for the relative catch efficiency (i.e., ρ-assumption) considered for the simulations, yielding 20 simulation scenarios. For convenience of cross-referencing within the document, the scenarios are labeled as "species code - ρ-assumption code", where "species code" includes 10 (Atlantic cod), 11 (haddock), 14 (silver hake), 23 (redfish), and 204 (winter skate), and "ρ-assumption" includes the four listed assumptions for the true relative catch efficiency, e.g. scenario 10-1 simulates a species population density based on Atlantic Cod in the Maritimes summer survey and a comparative fishing experiment between two gears with the same catchability.

For each simulated data series, we fit the 13 candidate models to estimate the relative catch efficiency. Iterations are retained in which the simulation generates a minimum of five catches for each gear, noting that in practice, cases with fewer than this number of catches would be rejected for analysis a priori due to data sparseness. In each retained iteration, the best model is selected by the lowest AIC among all models that properly converge (see details of estimation models in Section 2). In practice, an analyst would likely compare estimates of relative catch efficiency from all converged models instead of relying on a single metric such as AIC (see discussion in Section 5), but this was not tractable in a simulation setting.

The estimated relative catch efficiency from the best model is used for performance assessment of the suite of models. The simulation step and estimation step together form a simulation-estimation cycle (Figure 1); this cycle is iterated 100 times for each scenario. Iterations with successful and proper model convergence are retained (optimizations of models are checked for their maximum gradients and hessian matrices during the estimation step). Thereafter convergence rates of all iterations are reported for each scenario, and residual bias and variance are summarized from retained iterations. The residuals are investigated in order to provide a comprehensive review of model performance under all simulation scenarios.

3.3 Performance Measures

3.3.1 Summary of Simulated Data

We designed the following metrics to summarize the level of "difficulty" in terms of recovering the true relative catchability when the estimation models are applied to the simulated data, allowing for a comparison between scenarios. These metrics are calculated for each iteration and their median statistics among the 100 iterations are reported. The median is used to represent "average" characteristics of repeated simulations instead of the mean in order to avoid disproportional impacts from potential extreme cases.

- Magnitude of the length effect in the true relative catch efficiency:

$$LE(\rho) = S_L(\rho(L)). \tag{22}$$

$S_L(\cdot)$ denotes the calculation of standard deviation over the length dimension. The magnitude of the length effect is assessed as the standard deviation of $\rho(L)$ for its divergence from a constant, or length-invariant conversion factor. For the 4th ρ-assumption, the standard deviation is summarized for $\rho(L)$ prior to its compounding with station variation.

- Magnitude of the station effect in the true relative catch efficiency:

$$SE(\rho) = M_L(S_i(\rho_i(L)). \tag{23}$$

$M_L(\cdot)$ denotes the calculation of mean over the length dimension. Similar to the length effect, the station effect is the standard deviation of $\rho_i(L)$ among stations and across length bins. The station effect only applies to the 4th ρ-assumption and is zero for other three assumptions of true relative catch efficiency.

- Linearity of the underlying relative catch efficiency:

$$Lin(\rho) = M_L(|\mu(L) - lm(L)|). \tag{24}$$

In each scenario, a least squares regression line is fit to the true $\mu(L)$, i.e., expected proportion of catch by A; the absolute deviation of $\mu(L)$ from the optimized linear model $lm(L)$ is used to indicate its similarity to a linear function of length. The proposed binomial and beta-binomial models both use the cubic smoothing technique to estimate the proportion as a smooth function over length. Correspondingly this metric is designed to measure the complexity of $\mu(L)$ in its "shape". A higher value indicates more divergence in $\mu(L)$ from a straight line, or equivalently, a divergence in $\rho(L)$ from an exponential function. In the 4th ρ-assumption, $\mu_i(L)$ also varies according to station; for any scenarios based on this assumption, $Lin(\rho)$ is calculated for each station first, and then averaged across stations.

- Encounter Probability:

$$P(enc) = \frac{n(enc(i, L))}{n(i) \cdot n(L)}. \tag{25}$$

Here, $enc(i, L)$ denotes encounter of a species of length L at station i. Encounter is defined when there is positive catch of a species (by either gear within a pair of tows). $n(\cdot)$ denotes the count function, e.g., $n(i)$ is the number of stations and $n(L)$ is the number of length bins. The encounter probability is approximated by the sample proportion of encounters, calculated as the ratio between the number of encounters across all lengths and stations and the total number of combinations of length bins and stations. $P(enc)$ represents the encounter probability during a simulated comparative fishing experiment and measures the presence of a species within the survey area in each simulation case.

- Number of effective stations:

$$EN(stn) = n(\{i : \sum_L C_i(L) \geq 3\}), \tag{26}$$

where $C_i(L)$ is the combined catch for each pair of tows, $C_i(L) = C_{Ai}(L) + C_{Bi}(L)$. An effective station is defined here somewhat arbitrarily as a station where there is catch in a minimum of three length bins. In the simulation study, two stations are sampled for each of the 48 strata so there are in total 96 stations. The number of effective stations evaluates the spatial presence distribution of a species.

- Median positive catch:

$$Med(C) = median(\{C_i(L) : C_i(L) > 0\}). \tag{27}$$

The median positive catch is defined as the median of positive catch numbers across all length bins and stations. This metric assesses the positive catch rate, and in combination with the encounter probability, describes a species' underlying population density in the a simulated dataset.

- Pair difference:

$$\begin{cases} PD(C) = M_{i,L}(\{d(A,B) : d(A,B) > 0\}) \\ d(A,B) = \frac{|C_{Ai}(L) - \rho(L) \cdot C_{Bi}(L)|}{\frac{1}{2}(C_{Ai}(L) + \rho(L) \cdot C_{Bi}(L))}, \end{cases} \tag{28}$$

where $d(A,B)$ calculates the absolute difference between the pairs of catches after adjusting for the true relative catch efficiency between the gears, and $PD(C)$ is the average of differences in paired catches across stations and length bins. $PD(C)$ evaluates the difference between catches sampled by the pair of gears beyond different catchability, and is an indicator of random deviations of simulation realizations from the simulated truth.

Factors that facilitate model estimation are a) sufficient and informative catches in the simulated data, b) a smooth conversion with minimal station variation and length variation, c) an abundant species population with high encounter probability and positive catch rate across its length range, and d) a sampling process with small observation errors.

3.3.2 Summary of Estimation Results

The objective of the comparative fishing analysis is to derive a calibration between the pair of gears. The structural parameters defining the statistical models, i.e., the fixed and random parameters in Equations 9 and 14 excluding $\rho(L)$, are auxiliary in the simulation study. Evaluation of model estimation results is thus focused alone on the estimated relative catch efficiency between these gears, i.e., $\rho(L)$. The estimation residual is defined as

$$e(\rho(L)) = \rho(L) - \hat{\rho}(L), \tag{29}$$

where $\hat{\rho}(L)$ is the estimated quantity and $e(\rho(L))$ denotes its estimation errors. Note that for simulation scenarios based on ρ-assumption 4, estimation residuals are derived for the overall relative catch efficiency rather than station-specific, $\rho_i(L)$, since in practice it is the overall relative efficiency that is used as conversion factor to adjust survey data once a gear change has occurred.

In this study, each simulation scenario includes 100 iterations (or cycles). During each simulation-estimation cycle, estimation residuals for $\rho(L)$ are recorded for each of the 13 candidate models

that properly converged (see Section 5 for a detailed model comparison); however, the summary of the 100 iterations in a simulation scenario is based on residuals from the AIC-selected best models. More specifically, for each scenario, residuals from successful iterations are retained, upon which the performance measures defined in this section are calculated in order to present a comprehensive, yet concise, comparison of estimation results and model performances. The Monte-Carlo method of repeated simulation-estimation cycles can result in a vast amount of model output, and given the large numbers of simulation scenarios and candidate models, it is necessary to condense the results for a direct comparison among scenarios and models.

The residuals $e(\rho)$ are length-dependent; to summarize residuals over the length dimension, a weighted-mean technique is applied to integrate length bins (of 1 cm) based on a generic fish length-weight conversion function defined as follows,

$$w(L) = \beta(L) \cdot \frac{C(L) \cdot L^3}{\sum_L C(L)}, \tag{30}$$

where $C(L) = \sum_i C_i(L)$ is the total catch at length L from all stations in the experiment (i.e., during one simulation iteration), and the scale parameters $\beta(L)$ are computed and used to standardize the set of weighted-mean coefficients $w(L)$ such that

$$\sum_L w(L) = 1. \tag{31}$$

The weighted-mean technique takes into account both the length-catch composition in a simulated dataset and a simplified isometric relationship between the length and weight of a species. This allows the weighted average of residuals to balance risks associated with estimation loss from different sizes, while placing more weight on adult sizes that are often of greater interest for abundance indices and stock assessment. Admittedly, this technique may not be effective for species without sufficient catches covering an essential length range, e.g., simulated data for winter skate generally are sporadic catches; for such species, the un-summarized length-specific residuals are more informative. Nevertheless, the technique is applied to all considered species as any other summary methods are subject to the same issue.

The performance measures for model estimation results are defined as follows by summarizing the estimation residuals for the relative catch efficiency. We use k ($k = 1, 2, \cdots, 100$) to denote the 100 simulation iterations (iterations are conducted independently of each other and without order).

- Weighted residual mean:

$$\bar{e}(\rho) = \frac{1}{n(k)} \sum_k \sum_L w_k(L) e_k(\rho(L)). \tag{32}$$

  Residuals are summarized across length dimension with the set of weighted-mean coefficients, $w(L)$, and then averaged across 100 simulation iterations for assessment of overall prediction accuracy. $n(\cdot)$ denotes the count function, e.g., $n(k)$ is the number of iterations, which is 100 in this study. The application of a conversion factor to calibrate a survey index is a linear convolution process involving $\rho(L)$. Hence, this measure can be interpreted as the potential loss (either under-estimation or over-estimation) of total catch weight due to inaccurate gear conversion, given the assumed length-composition and length-weight relationship.

- Weighted residual standard deviation:

$$s(\rho) = S_k\left(\sum_L w_k(L)e_k(\rho(L))\right). \tag{33}$$

We define $S_k(\cdot)$ as the calculation of the standard deviation of a quantity over dimension k, or in this case, the standard deviation of the weighted-average residuals across 100 simulations. $s(\rho)$ measures how consistently the models perform over repeated simulations, and is an approximation of modeling uncertainty, or precision.

- Probability of a loss greater than 20%:

$$P.20 = \frac{n(\{k : \sum_L w_k(L)|\rho(L) - \hat{\rho}_k(L)| < 0.2\})}{n(k)}, \tag{34}$$

where $n(k)$ is the total number of simulation iterations, i.e., 100 in this study. The loss is defined as the weighed average of the absolute residuals and approximates the potential loss of total catch biomass when applying the conversion factor to calibrate the survey index. The numerator contains the number of iterations where estimated relative catch efficiency would result in an over-estimation or under-estimation of total catch weight within 20%.

- Probability of a loss greater than 50%:

$$P.50 = \frac{n(\{k : \sum_L w_k(L)|\rho(L) - \hat{\rho}_k(L)| < 0.5\})}{n(k)}. \tag{35}$$

Similar to $P.20$, $P.50$ is the proportion of simulation iterations where the model estimated conversion factors would resulted in less than 50% over-estimation or under-estimation.

4 Summary of Results

4.1 Simulated Data

For each of the 20 simulated scenarios, there are 100 simulation iterations, and in each iteration, 96 stations are sampled by the pair of gears A and B. An example of simulated dataset from each simulation scenario is presented in Figures 11-15. For consistency, we select the same underlying stations for the example (by assigning a random number seed of 1 in R), although the simulated data from each iteration can vary greatly for the same scenario due to large sampling variations assumed. We also calculate the mean number per tow (mnpt) for both gears in every simulation iteration for the 20 simulation scenarios in order to demonstrate overall size-compositions of simulated catches (Figures 16-20). Iterations based on the same Maritimes RV survey catches (Figure 9) can vary greatly, due to high sampling variability. The simulation study samples two stations in each stratum, unlike in the Maritimes RV survey where some strata included more stations than others. Depending on the spatial distribution, a species may have been sampled more frequently in strata where it is densely populated; since our calculation of mnpt treats each station equally, this may result in a discrepancy in the scale of mnpt between the simulated datasets and results from a proper analysis of the Maritimes survey. For example,

Atlantic cod had a patchy spatial distribution (majority of catches are within a small number of stations), so its discrepancy is the most pronounced among the five species: the simulated catches, when averaged per tow, are generally much smaller than averages from the Maritimes RV survey (Figures 9 and 16).

A summary of simulated datasets for each scenario using metrics defined in Section 3.3.1 is provided in Table 4. The metrics are in three categories to assess the true relative catch efficiency, simulated catch abundance levels and distribution, and differences between the comparative fishing pair, respectively.

The first ρ-assumption (i.e., assumption for true relative catch efficiency) is a control group where the two gears have the same catchability and there is no length effect or station effect in the conversion $\rho(L)$, hence $LE(\rho)$, $SE(\rho)$ and $Lin(\rho)$ are zeros. The second ρ-assumption has the strongest length effect among the four assumptions as well as shape complexity, indicated by large $LE(\rho)$ and $Lin(\rho)$. The third ρ-assumption is species-dependent but the simulated $\rho(L)$ generally has a simple shape due to the exponential model setting. The fourth ρ-assumption is the only assumption to include station effect, and while length effect can be significant, the shape of $\rho(L)$ is rather smooth giving a small $Lin(\rho)$.

Metrics that summarize catch numbers for each species are indicators of their underlying population abundance and density distribution in the simulation scenarios. Haddock and silver hake have a wide spatial distribution: on average of 100 simulations in each simulation scenario, over half of all stations result in positive catch ($EC(stn) > 48$) and their encounter probabilities are the highest among all five species ($P(enc) > 0.2$). These two species also have high population levels with large positive catches (catch number per km^2), $MED(C)$, and small pair differences, $PD(C)$. These are signs of sufficient and good quality data for model estimation. Redfish has a wide presence but is relatively less abundant in the survey area; pair difference is slightly higher than haddock and silver skate as well. Simulated catches for Atlantic cod indicate a low population level and sparse spatial distribution. For winter skate, catches are both sporadic and sparse across the survey area. Atlantic cod and winter skate both have the highest pair differences, where overall difference between the pair of gears can be over 1.5 times of the average between the combined catches ($PD(C) > 1.5$). Different ρ-assumptions also contribute to variations in simulated catch abundance and density when comparing simulation scenarios, and the interaction between ρ-assumption and species assumption is complicated. Among scenarios for the same species, the scenario based on the first ρ-assumption generates lower catches than the other three assumptions, as the other three ρ-assumptions mostly simulate a more efficient replacement gear that elevates catch numbers.

4.2 Model Convergence and Selection

The suite of models include five Binomial models and eight Beta-Binomial models; these candidate models are fit to 100 iterations in each simulation scenario. A summary of model convergence and model selection results for each simulation scenario is presented in Table 5 and Table 6, respectively. The number of iterations where each candidate model properly converged indicates estimability of a model given a simulation scenario. The number of iterations where each candidate model was selected as the best model (by lowest AIC) indicates suitability of a model for each scenario.

Model convergence is directly related to the quality of simulated datasets which in turn, depends on the conditions of simulation assumptions. In general, more complicated models require more sufficient data for successful estimation, or optimization of model parameters. The Binomial models have a high convergence rate for all species except winter skate which has extremely low abundance levels. In comparison, a Beta-Binomial model seeking to estimate an additional variance parameter than its Binomial model counterpart tends to result in lower convergence rates, e.g., BB0 vs. BI0, BB4 vs. BI3. However, BB6 and BB7 have slightly better convergence rates than their counterpart, BI4, especially for abundant species, possibly because the variance parameters in BB6 and BB7 can allow for more flexibility in the cubic smoothing of the mean parameter, giving these two models a small advantage to balance out their more complicated model structures. Within the same category of Binomial or Beta-Binomial models, convergence rates tend to decrease as number of model parameters (model complexity) increases; the most complicated models such as BI4, BB6 and BB7 are the least likely to properly convergence since length effects, station effects and variance are estimated simultaneously. In addition, for the same model, convergence is easier to attain for more abundance species due to more sufficient and informative catches, e.g., Scenario 10-1 vs. 11-1 at BB3.

Despite relatively higher convergence rates, the Binomial models rarely prevail over their Beta-Binomial counterparts in terms of model fit as assessed by AIC (Table 6). In simulation scenarios for winter skate, the Beta-Binomial models mostly fail to converge and the Binomial models are selected as the best models; for other species, the best models are predominantly Beta-Binomial models and at the same time, whenever a more complicated model converges successfully, it tends to result in lower AIC. This indicates the necessity to allow for extra-binomial variance in these simulation scenarios, and is a reasonable result considering the high sampling variability cast upon survey catches in the simulation step. The Binomial models have fixed variance and due to the bias-variance trade-off, are more likely to produce a “wavy” estimation for the relative catch efficiency (see Figure 21 for an illustration and Section 5 for a detailed discussion on model differences). Failure to account for the length effect or station effect can also have significant consequences. In practice, results from the 13 candidate models should be examined as part of model selection. Nevertheless, the simulation study with massive iterations require an objective and efficient model selection method and hence the single metric, i.e., AIC is used. A more detailed comparison between the Binomial and Beta-Binomial models is discussed in Section 5. Overall, BB5 is a balance between convergence and performance for more abundant species, while BI3 is a balanced model for scarce species.

4.3 Performance Assessment

Detailed estimation residuals of the best models are presented for each simulation scenario to assess performance of the suite of models (Figures 22-25 for Atlantic cod, Figures 26-29 for haddock, Figures 30-33 for silver hake, Figures 34-37 for redfish, and Figures 38-41 for winter skate). A summary of estimation residuals using performance measures defined in Section 3.3.2 is in Table 4, next to the summary statistics for the simulated datasets for these scenarios.

The control groups generally result in the lowest estimation bias (weighted-average error of repeated simulations, $\bar{e}(\rho)$) among the four scenarios for each of these species, respectively, i.e., Scenarios 10-1, 11-1, 14-1, 23-1 in Table 4; median estimation residuals are effectively zero

across length bins (Figures 22, 26, 30, 34, and 38). However, estimates are less precise for the control groups than other scenarios, as weighed-mean residual standard deviations, $s(\rho)$, are mostly higher (Table 4). This may be due to the assumption of equivalent gear efficiency in the control groups which generates lower catch levels contributing to higher estimation uncertainty, while the other three scenarios mostly assume more efficient replacement gears and result in more sufficient data when combining the pair of gears. Besides, even though the control groups do not include a length effect in the relative catch efficiency, the precision tends to attenuate at small and large sizes with widened dispersion resulting from the smaller catch numbers at these lengths.

Scenarios including a length effect in the true relative catch efficiency result in varying amounts of estimation error (Table 4, and Figures 23-25,Figures 27-29,Figures 31-33,Figures 35-37,Figures 39-41, for the five species, respectively). The estimation errors are primarily related to species abundance, as lower assumed densities and therefore catch numbers lead to smaller effective sample sizes. Haddock, redfish and silver hake in these scenarios feature more than 6% over-estimation in terms of approximated total biomass while Atlantic cod and winter skate tend to be under-estimated. More precisely, residuals vary considerably as a function of length and estimation quality of less abundant lengths are more impacted. In addition, estimation bias seems to propagate along length, possibly due to length smoothing in the models, as the length smoothing technique assumes continuity and smoothness for $\rho(L)$ resulting in autocorrelation among similar lengths. The propagation is more prominent for Beta-Binomial models than Binomial models since the Beta-Binomial models prefer simpler shapes (more smoothness) by allowing for flexible variance parameters. The station effect is also critical to model performance. Station variation adds noise to the data and affects both estimation accuracy and precision. Less abundant species are the most impacted; for example, Atlantic cod and winter skate both result in the lowest $P.20$ in scenarios 10-4 and 204-4, respectively, among the four scenarios for each species (Table 4).

Overall, species abundance and population density distribution are key to model performance since simulated catches directly depend on them. Abundant species such as haddock, silver hake and redfish have a consistent capped loss: all simulation scenarios have high quality estimations where estimated quantities lead to under 20% loss (either over-estimation or under-estimation). For less abundant species such as Atlantic cod and winter skate, performance is more sensitive to nuanced factors in the true relative catch efficiency such as length effects and station effects. In general, the length effect can be successfully recovered by these models with the smoothing technique, except when a strong length effect is expected and in combination with low population density, e.g. Scenarios 10-2 (Figure 23) and 204-2 (Figure 39). The station effect is more difficult to effectively account for, and these models tend to have reduced performance dealing with strong station variation, e.g., Scenarios 10-4 (Figure 25), 23-4 (Figure 37), and 204-4 (Figure 41). Overall, model performance is acceptable and in most scenarios can contain the loss within 50% (Table 4).

5 Model Comparison

The Monte Carlo simulation study requires an objective and simple-to-apply criterion for efficient model selection during the automated iteration process; for this purpose, the Akaike

information criterion (AIC) is used. Information-based model selection criteria such as the AIC, AICc (AIC with a correction for small sample sizes) and BIC (Bayesian Information Criterion) are designed to balance model fit and complexity by penalizing the number of parameters, or degree of freedom, that are estimated given the data (Friedman et al. 2001). With different penalties, these metrics have their respective advantages and disadvantages in performance, for example, for finite samples, BIC prefers simpler models while AIC tends to select more complex models (Friedman et al. 2001). Relying on a single form is unlikely to be successful universally (Brewer et al. 2016). In the present case of simulations, we used a single metric as it simplified the summarization of results; however, in practice, it is recommended to scrutinize and compare results from the suite of models in practice in addition to these metrics for proper model selection.

The 13 candidate models are progressive in complexity, with a variety of statistical assumptions for effects that potentially exist in the data that are related to over-dispersion, lengths, or stations etc. In this section, we compare these candidate models based on the simulation study in Sections 3 and 4. We examine their statistical assumptions in detail and implications for model performance with respect to different scenarios, in order to guide model selection in practice. The performance measures in Section 3.3.2 are applied to estimation residuals from each of these models and a summary of residuals are presented in Tables 7-18 for the three species Atlantic cod (10), silver hake (14) and winter skate (204), and the four ρ-assumptions for true relative catch efficiency in Section 3. Discussion in this section is focused on theses three species for simplicity, as results from the three species are sufficient for illustration of main differences in data properties against which the performance of different models is evaluated.

5.1 Length Effect

When the true relative catch efficiency does not contain any length effect, any estimation model that uses the length smoothing technique (BI2-BI4 and BB2-BB7) is more complicated than necessary. This can lead to over-fitting and inflated estimation errors. In Scenarios 10-1, 14-1 and 204-1, estimation bias is smaller for estimation models that assume no length-dependence, corresponding to the truth, than models that incorrectly assume length-dependence (Tables 7, 11 and 15). In these scenarios, AIC failed to select the simpler models even though they have better performance. However, the AIC-selected best models do not have significantly worse performance, either, since length smoothing can produce a flat conversion as well, despite more sensitivity to sampling variability. Moreover, because surveys in the Gulf and Maritimes region will involve a change in both vessel and gear, and because gear changes often involve a a change in length selectivity, it is not realistic to assume the length effect is trivial, and when the true relative catch efficiency does contain length effect, these simple models quickly deteriorate in performance (Tables 8-10, 12-14 and 16-18).

Among models which apply the smoothing technique, accurately capturing the length effect highly relies on data quality. When there are no significant catches within a certain length range, the smoothing technique must interpolate using information from catches in nearby length bins, hence missing any particular length effect within the range, e.g. Scenarios 10-2, 14-2 and 204-2 where length smoothing cannot capture the detailed transition in the V-shaped relative catch efficiency (Tables 8, 12 and 16, Figures 23, 31 and 39). In these scenarios, statistical models

simply can not compensate for data deficiency.

A practical approach for proper model selection that accommodates consideration of the length effect is to compare all model results; if model-estimated length effects are not significant statistically, we should select a simpler model regardless of AIC. This can be assessed by examining the confidence intervals, or more rigorously, by model comparison tests that calculate a p-value to indicate statistical significance of the length effect.

5.2 Station Effect

Implications of the station effect on model performance are less conclusive. The station effect in the estimation models assumes a particular form (Section 2) and when this form cannot explain the dispersion among stations, its incorporation may not lead to much improvement. For example, in Scenario 10-4, the true relative catch efficiency contains a station effect, but estimation models without a station effect perform better than models with one in terms of $P.20$, i.e., BI2 versus BI3 and BB2-BB3 versus BB4-BB7 (Table 10). In Scenario 14-4, however, BB5 and BB7 give more accurate estimates than BB2 (Table 14) with progressively more elaborate structure of a station effect. In practice, a station effect should be examined by comparing sample conversion among stations when there is adequate data (e.g., by individually calculating the sample proportion of catch of one gear from each station and then evaluating difference among stations).

5.3 Binomial versus Beta-Binomial

The suite of estimation models for the comparative fishing analysis in this study includes two categories, the Binomial models and the Beta-Binomial models, distinguished by a binomial versus beta-binomial distribution for the conditional catch as a random variable. The binomial distribution $\mathrm{BI}(n, p)$ has a fixed variance, i.e., $np(1-p)$, in relation to its mean, np, while the beta-binomial distribution $\mathrm{BB}(n, p, \phi)$ includes an explicit variance parameter, ϕ, that allows for larger variance given the same mean. This additional parameter also allows the beta-binomial distribution to take various shapes, e.g., bell-shape similar to the binomial distribution, J-shape or inverse J-shape, and even U-shape (the beta-binomial distribution can be equivalently characterized by two shape parameters in place of the mean and variance parameters). As a result, the beta-distribution is more flexible than the binomial distribution and usually more effective when there is clear over-dispersion in the random variable. For this reason, the beta-binomial distribution has been widely used for binomial variables in many fields when the binomial distribution is not adequate in modeling the dispersion structure, e.g., in ecology (Harrison 2015), in phytopathology (Hughes and Madden 1993) and in machine learning (Schuckers 2003).

Survey catch from bottom trawl surveys tends to have large variability and in paired tow comparative fishing analysis, it can be a necessity to account for extra variance compared to the simple binomial assumption. This is because when variance is not properly explained, the estimate for the mean parameter is subject to a large estimation uncertainty, and the estimated value in practice can potentially deviate greatly from the truth. To illustrate this circumstance, an

example in presented in Figure 21 based on one simulation case (the simulation iteration based on the random seed 1 and scenario 14-4 from Sections 3 and 4). In this example, the simulation casts large sampling variability onto simulated catches. Most Binomial models result in "wavy" shapes for the estimated relative catch efficiency - failure to explain the dispersion leads to huge swings in the estimates even though the confidence interval (estimation uncertainty) can cover the truth. The Beta-Binomial models, on the other hand, predict smooth curves for the relative catch efficiency since variance can be properly addressed in these models.

However, this variance parameter is attached to a specific assumption for the type of over-dispersion, i.e., a "prior" beta-distribution for the proportion of catch (Miller 2013). When there is model mis-specification, that is, a mis-match between the simulation model (or truth) and the estimation model, it could lead to an estimation bias. This is the case in Scenarios 14-2, 14-3, and 14-4 where there is a noticeable difference in $\bar{e}(\rho)$ between the Binomial models and their Beta-Binomial counterparts (Tables 12-14). Although the Binomial models have better estimation accuracy, the estimates contain higher estimation variance, indicated by $s(\rho)$, and consequently their overall performance, gauged by $P.20$ in particular, is slightly poorer compared to the Beta-Binomial models. Besides, if we examine these simulations case by case, the Beta-Binomial models tend to predict a more realistic and smoother shape for the relative catch efficiency. The estimation bias in Beta-Binomial models was also noticed by Cadigan and Bataineh (2012) in their analysis of comparative fishing data. However, Cadigan and Bataineh (2012) simulated negative binomial catches from a parametric distribution rather than empirically from past survey data and they did not consider length-dependency in the relative catch efficiency. As an expansion to Cadigan and Bataineh (2012), we included an additional simulation study in Section 7 where we simulated negative binomial catches with length-dependency in order to test performance of Binomial vs. Beta-Binomial models and to elaborate this potential estimation bias. In addition, Kim and Lee (2017) studied the estimation bias when fitting beta-binomial distributions to over-dispersed binomial variables in general and concluded that the beta-binomial distribution could fit better than the binomial distribution but recommended caution when there is significant evidence that it is not valid for the data. Our simulation study here supports a similar recommendation.

As for model selection, AIC tends to select a Beta-Binomial model over Binomial models, despite some loss in accuracy, again demonstrating the limitation of using a summarized model selection metric (see detailed discussion in the second paragraph of Section 5). The trade-off between bias and variance is a common topic in statistics and the choice of a balance point depends on risk tolerance in reality (Friedman et al. 2001; Ding et al. 2018).

5.4 Overview of Model Types

Aside from the suite of Binomial and Beta-Binomial models proposed in this study, there are various other methods for analysis of paired catches, e.g., Miller et al. (2010) provided a list of models for comparative fishing analysis, although all lacked consideration for the length effect. The spectrum of model complexity can relate to the bias-variance trade-off and here are some examples of model types and a brief discussion of their relative strengths and weaknesses (in the order of model complexity):

1. The simplest estimator for the relative catch efficiency is the ratio estimator, i.e., the sample mean of ratios between paired catches,

$$\hat{\rho}(L) = \frac{\sum_i C_{Ai}(L)}{\sum_i C_{Bi}(L)}, \tag{36}$$

where $\hat{\rho}(L)$ is the station-aggregated sample relative catch efficiency (sample conversion), $C_{Ai}(L)$ and $C_{Bi}(L)$ are the catches at station i and length L by gears A and B, respectively. The sample conversion is subject to wide uncertainty especially when sample size is small, and may not exist for some lengths due to lack of data. An improvement to the ratio estimator is to incorporate a non-parametric smoothing over length; however, this has limited improvements on precision since large sampling variability during bottom trawl surveys typically results in excessive sample variance. The ratio estimator is asymptotically unbiased and can be used as a benchmark for more complicated estimation models.

2. The station effect can be considered in the ratio estimator as well, for example, by fitting a parametric distribution to the station-wise sample ratios,

$$\begin{cases} \rho_i(L) = \dfrac{C_{Ai}(L)}{C_{Bi}(L)} \sim \mathrm{Dist} \\ \hat{\rho}(L) = mean(\mathrm{Dist}), \end{cases} \tag{37}$$

where $\rho_i(L)$ is the station-specific sample ratios and $\hat{\rho}(L)$ is the estimated overall conversion. Using a statistical distribution may better locate an estimate for $\hat{\rho}(L)$ as the mean parameter than simple averaging, but any improvement depends on suitability of the distributional assumption and sample size. Besides, in reality and for many species, there is insufficient catch when partitioned by length and station (curse of dimensionality). When the catch by B is zero, the sample ratio is undefined. In practice, it is possible to simply discount zeros samples from gear B and then use a censored or truncated distribution instead; this is necessary because a redefinition of the measurable set is required for the ratio statistic when there is a finite probability of zero catch by B (see Griffin 1992; Dietz and Böhning 2000; Johnson et al. 2005, for details about this practice). Consequently, the estimate can be highly biased and it is not recommend for the comparative fishing analysis of data limited species (Nemes et al. 2009).

3. Instead of a parametric distribution for the ratio, we can consider fitting to the catch numbers by the two gears separately, for example, with independent Poisson or independent negative binomial distributions,

$$\begin{cases} C_{Ai}(L) \sim \mathrm{Dist}_1 \\ C_{Bi}(L) \sim \mathrm{Dist}_2 \\ \hat{\rho}_i(L) = \dfrac{mean(\mathrm{Dist}_1)}{mean(\mathrm{Dist}_2)}, \end{cases} \tag{38}$$

This surmounts the issue in modeling the ratio distribution but introduces more model parameters, and the method is both sensitive to suitability of distribution and sample size, especially when length is considered. In theory, these methods will have the same MLE with the station-aggregated ratio estimator in Point 1.

4. Another alternative to modeling ratio distribution is by conditioning the catch of one gear on the combined catches from the pair of gears. The Binomial and Beta-Binomial models in this study fall into this category. Possible issues have been discussed extensively in this section and in addition, model mis-specification can also cause bias in the estimate, particularly in the Beta-Binomial models (see Section 7 for a detailed discussion).

5. The Binomial and Beta-Binomial models in this study use a non-parametric smoothing for the length effect. This technique typically does not introduce bias but can be sensitive to sampling variability when sample size is small; performance may decrease for low-abundance species, and for small and large sizes of a species. Besides, extrapolation outside of the modeled length range can be inappropriate since the technique is a local smoothing. Stronger assumptions can be made for length-dependency to improve these issues, e.g., a monotonic smoothing, a parametric form for the length effect, or based on autocorrelation among lengths (Bourdages et al. 2007; Thygesen et al. 2019; Cadigan et al. submitted). There can be alternative assumptions for differences among stations as well. However, as we have concluded from the simulation study, there are consequences associated with each statistical assumption. Therefore, justification for model building and vigorous validation are crucial for any additional or alternative assumptions.

In summary, the suite of Binomial and Beta-Binomial models cover a range of assumptions and have the flexibility to cover a diversity of situations. However, analysis of a specific species may be improved by adapting these models to include assumptions that suits the species, including those on length-selectivity, spatial stratification (either based on stations or spatial area), etc.

6 Sample Size

In this section, we explore the effect of sample size on model performance. Two topics are discussed: 1) equal sampling of each stratum, where increases in the number of stations sampled is uniform across strata, and 2) unequal sampling of each stratum, where some strata are sampled disproportionately to target a particular species, aiming at increasing the ratio of effective stations.

6.1 Number of Stations Per Stratum

To keep the additional simulation study succinct, we only test three situations of abundant, less abundant and scarce population levels, based on the three archetype species silver hake (14), Atlantic cod (10) and winter skate (204), respectively, as well as the four assumptions for true relative catch efficiency in Section 3. This gives 12 combinations of species and relative catch efficiency assumptions in total; each combination is expanded to three scenarios with different sample sizes, where the number of sampled stations per stratum are set to 1, 2 and 3. Except for the sample size, all other simulation settings remain the same as in Section 3. There are 36 scenarios in this section and each scenarios include 100 iterations; in each iteration, data are generated and models are fit to the data. For each scenario, the same data summary metrics and model performance measures are used for performance assessment, and estimation

residuals are compared among scenarios for different sample sizes. For convenience of reference, the scenarios in this section are named as "species code-ρ-assumption code-sample size code", where sample size code includes 1, 2 and 3, e.g., 10-1-1 denotes the scenario where population density is simulated to resemble Atlantic cod, relative catch efficiency is simulated as length-independent at 1, and a single station is sampled for each stratum in the comparative fishing experiment.

Table 19 summarizes the simulated datasets for each scenario. Metrics for the true relative catch efficiency remain the same when increasing the sample size, despite small variations in the station effects $SE(\rho)$ due to randomization. Effective stations increase roughly proportionally to the number of stations per stratum, while encounter probabilities and median catches are mostly invariant with respect to sample size, all as expected. For abundant species, pair differences, $PD(C)$, also remain similar for different sample sizes, but for scarce species such as Atlantic cod and winter skate, $PD(C)$ seems slightly elevated when number of stations per stratum decreases.

Estimation residuals for the scenarios are compared in groups where only sample size varies (Figures 42-53). Performance summaries are presented in Tables 20, 21, and 22 to assess the impact of sample size on estimation accuracy, precision, and overall performance, respectively. While precision consistently improves with more sampled stations per stratum for most scenarios, accuracy only improves in scenarios of more abundant species and without length effects (Tables 20 and 21). For most scenarios, the weighed-mean residuals do not indicate significant improvements with increasing sample size. This could be an estimation error that is difficult to improve by simply increasing sampled stations per stratum, as additional stations hardly address the key issue, i.e., a lack of population in certain sizes such as juveniles due to gear selectivity. However, it may also result from an inherent bias from the estimators that is related to model mis-specification, as in most cases, the Beta-Binomial models are selected as the best model and these tend to sacrifice some accuracy for precision (i.e., bias-variance trade-off; a more detailed exploration and discussion on this topic is in Section 5). Overall model performance measured by $P.20$ gives a mixed conclusion (Table 22): while three stations per stratum could improve estimation of relative catch efficiency for Atlantic cod in all scenarios, estimation quality for winter skate does not improve as much. Besides, when station variation is present, additional stations led to worse performance, possibly due to an overwhelming amount of noise (strong station variation) rather than useful information (few effective catches) within these sampled tows.

Considering that the number of iterations where models properly converged is small for winter skate, and hence these summarized statistics have less credibility than for abundant species (i.e., statistical power is low), we examine the residual plots in detail (Figure 50-53). Residuals for winter skate exhibit a slight but positive response to increased number of stations, which results in less erratic behavior and fewer cases of extreme deviations (residuals from each iteration in grey dotted lines appear slightly less dispersed for simulations with larger sample size). At the same time, when residuals for Atlantic cod are assessed over length, it seems that estimation for less abundant length bins does not benefit as much from larger sample sizes, e.g., the median residual deviations are unchanged for small lengths in Figures 43, and for large lengths in Figures 45. This is similar to the winter skate situation where indiscriminately increasing the number of stations failed to increase data quality.

Conclusion from the comparison of model performance for different species and in different length ranges of the same species emphasizes the dominant role of population density which has the strongest impact in data quality to support statistical analysis. Simply adding more stations, without specific additions of stations involving larger catches that provide a contrast in relative catch amounts between gears, does not improve performance. Thus for species with sporadic small catches, there is limited benefit to simply increasing the number of paired tows.

6.2 Targeted Stations

Section 6.1 demonstrated the more limited benefits of indiscriminately increasing sample size without regard to the likely catch densities of the species of interest. A better approach might be to target additional paired stations to areas in which a species (or a certain length range of a species) of interest is relatively abundant. Targeted stations can improve data quality theoretically by generating more effective catches. However, it can be difficult or even impossible to "target" given the need to optimize the experimental design with respect to numerous species that have different spatial distributions and because identifying high density areas for survey sampling can be difficult when a species is sparsely distributed and distribution is poorly characterized using recent surveys or when there are shifts in distribution among years. Based on results of the simulation study above, benefits of targeting stations depends on the species, the survey, and the "targeting" plan.

7 Parametric Simulation of Catch

The simulation method in Section 3 uses a resampling technique to generate catch numbers from a pool of catch-at-length established based on past Maritimes RV surveys. This method is both convenient and realistic, without projecting excessive assumptions on the characteristics of the sampling process. However, most traditional simulations employ a parametric model for survey catch with statistical distributions, such as the Poisson and negative binomial distributions. These distributions are prevalent in the modeling of survey catch or population abundance (Hilborn and Walters 2013; Venables and Dichmont 2004) where they are used to account for variations within a group of observations, e.g., catches stratified by time, space or other indices, but do not address the sampling process directly. In this section, we test these alternative methods that generate catch numbers in the comparative fishing experiment based on parametric catch models, and compare to the resampling method in Section 3.

7.1 Assumption for Population Density

The population density is required for simulation of catch based on parametric distributions. However, the spatial distribution of the absolute underlying population density is difficult to simulate explicitly as it is obscure in its nature. In this study, the available density to the sampling gears are considered instead, and only at a limited number of stations rather than spatially in the study area. The available density is the population density "seen" by a specific gear (i.e., post-gear selection). This facilitates simulation of paired catches as they can be considered

in relative terms for the two gears. We estimate the available density for gear A from recent Maritimes RV survey catches (during 2015-2017) and then derive the available density for gear B by applying the assumed relative catch efficiency. Assuming within-stratum homogeneity of a species distribution, the mean catch per tow is calculated for each stratum and each length bin, from which the available density for gear A at station i (in stratum s) is estimated by a 5-point moving average over length and denoted by

$$\mu_{Ai}(L) = q_{Ai}(L)D_i(L). \tag{39}$$

Here, $q_{Ai}(L)$ denotes catchability of gear A and may include station variation; D_i is the absolute population density at station i and is not estimated explicitly. The available density to gear B can be derived by applying the assumed true conversion $\rho_i(L)$,

$$\mu_{Bi}(L) = q_{Bi}(L)D_i(L) = q_{Ai}(L)D_i(L) \cdot \rho_i(L). \tag{40}$$

In reality, catches can vary greatly even for stations within the same stratum. To simulate this phenomenon, a random station variation δ_i is added to the available densities: $\mu A, i(L)\delta_i$ and $\mu B, i(L)\delta_i$. This station variation is the same for the pair of tows within the same station and does not contribute to pair difference. δ_i is realized from a lognormal distribution, and the median of δ_i is set to 1 so that half of the stations are expected to have elevated catches while the other half are depressed. This effect can create a contrast among stations even within the same stratum where homogeneity is assumed.

7.2 Assumption for Catch Distribution

We test the following three candidate assumptions of parametric catch distributions for comparison with the resampling method:

- $Pois$: Independent Poisson distributions are used for generating paired catches by gears A and B,

$$\begin{cases} C_{Ai}(L) \sim \text{Pois}(\mu_{Ai}(L)) \\ C_{Bi}(L) \sim \text{Pois}(\mu_{Bi}(L)), \end{cases} \tag{41}$$

  where $\mu_{Ai}(L)$ and $\mu_{Bi}(L)$ are the available densities for gears A and B, respectively, at station i and length L.

- NB: Independent negative binomial distributions are used for generating the pair of catches by gears A and B,

$$\begin{cases} C_{Ai}(L) \sim \text{NB}(\mu_{Ai}(L), \phi_{Ai}(L)) \\ C_{Bi}(L) \sim \text{NB}(\mu_{Bi}(L), \phi_{Bi}(L)), \end{cases} \tag{42}$$

  where $\phi_{Ai}(L)$ and $\phi_{Bi}(L)$ are the variance parameters of the negative binomial distributions and control the variance of realized catches.

- $NB - Beta$: A negative binomial distribution is used for generating the combined catch from the pair of gears and then, a Beta distribution is used for partitioning the combined catch

and assigning to the two gears,

$$\begin{cases} C_{Ai}(L) + C_{Bi}(L) \sim \mathrm{NB}(\mu_{Ai}(L) + \mu_{Bi}(L), \phi_i(L)) \\ \dfrac{C_{Ai}(L)}{C_{Ai}(L) + C_{Bi}(L)} \sim \mathrm{Beta}(\alpha_i(L)\mu_{Ai}(L), \alpha_i(L)\mu_{Bi}(L)), \end{cases} \tag{43}$$

where $\phi_i(L)$ controls the variance of realized combined catches, and parameters of the Beta distribution are formulated in such a way that expectation of proportion of catch by A is the same as the proportion of the available density to gear A, while parameter k controls the station variation in the proportion,

$$E[\frac{C_{Ai}(L)}{C_{Ai}(L) + C_{Bi}(L)}] = \frac{\mu_{Ai}(L)}{\mu_{Ai}(L) + \mu_{Bi}(L))}. \tag{44}$$

The Beta-distribution assumption for the proportion of catch by gear A within a pair matches the process assumed in the Beta-Binomial estimation models, hence there is no model mis-specification between the simulation step and the estimation steps. This is typical in statistical simulation studies that seek to recover model parameters or test model efficiency (Morris et al. 2019) and in fisheries is sometimes referred to as "self-testing". The $Pois$ assumption does not introduce mis-specification, either, since both the Binomial and Beta-Binomial models assume conditional binomial distribution for paired catches while the sum of two independent Poisson variables is still Poisson-distributed (Miller 2013). The use of negative binomial distributions for generating catches based on population density, and the resampling method with lognormal sampling errors in the simulation step is a mismatch with the Binomial and Beta-Binomial models in the estimation step.

Catch numbers by gears A and B at station i are generated by the catch distributions based on the density available to each gear at the station. For $Pois$ and NB assumptions, the process is straightforward: catches are realized from a Poisson or negative binomial distribution for each gear, respectively. Note that in the Poisson distribution, variance is equal to the mean; in the negative binomial distribution, an alternative parameterization can relate the variance (ϕ) to the mean (μ) as $\phi = \mu + \mu^2/k$, and we constrain the variance by fixing the parameter $k = 0.5$ for both gears A and B; for the $NB - Beta$ assumption, the combined catch at each station is realized from the negative binomial distribution and subsequently assigned to the two gears based on a random Beta-distribution, and the tuning parameters are specified as $k = 0.5$ and $\alpha = 0.5$.

For simplicity, the simulation study in this section only considers two scenarios based on two species, 10-4 for Atlantic cod and 14-4 for silver hake, to test contrasting cases of an abundant and a less abundant species, respectively. The simulation settings for the true relative catch efficiency is maintained the same as in Section 3, so here we only describe the simulation of catch numbers. For convenience of reference, the scenarios in this section are named as "species code-conversion code-catch assumption code", where catch assumption code includes "$Pois$", "NB", "$NB - Beta$", and "R" (for the resampling technique in Section 3), e.g., 10-4-NB.

7.3 Comparison of Catch Simulations

In all scenarios, two stations are sampled for each stratum with a total of 48 strata in the Maritimes region. Figures 54 and 55 present an example of the simulated datasets for each

simulation scenario (based on the same randomization setting, i.e., random seed is 1 in R). Similar metrics to summarize simulated datasets and measures to assess model performance are used in this section (Table 23).

Despite an additional station effect to the underlying population density, the two stations within each stratum tend to give similar patterns in catch-at-length, especially for scenarios 10-4-$Pois$ and 14-4-$Pois$ (Figures 54 and 55), while scenarios based on the resampling method do not exhibit such clear patterns and catches are more erratic (4th row of Figures 11, 13). The degree of station difference in the survey catch-at-length can be extremely high and difficult to simulate functionally with parametric models (Figures 4 and 6). In addition, number of effective stations is on average smaller for scenarios based on empirical simulation than scenarios based on parametric distributions (Table 23). These are indications that parametric simulation of catch does not fully capture certain characteristics of the survey sampling process, and suggests that the empirically-based simulations may provide a more realistic test base for the models.

As for model performance, scenarios with the $Pois$ and $NB-Beta$ assumptions result in negligible prediction bias; since both Binomial and Beta-Binomial models are correctly specified; this indicates adequate statistical properties of the models including identifiability and estimability (Table 23). Facing the lowest data quality (most complicated datasets) from the resampling method, estimation errors are higher than scenarios with the $Pois$ and $NB-Beta$ assumptions and on par with scenarios based on the NB assumption depending on the species: both residual bias and precision improves with population level for the resampling technique, i.e., from Atlantic cod to silver hake (Table 23). This is a more reasonable result than the NB simulation technique and is an indication that the resampling simulation technique may be more realistic. The suite of models do not perform well in scenarios with the NB simulation assumption; recovering relative catch efficiency from independently negative binomial distributed catches is ineffective even for the most abundant species, silver hake, where there is a consistent estimation bias even though estimation variance is small (Table 23). More specifically, this bias is related to the mis-specification of Beta-Binomial models, as AIC mostly selects a Beta-Binomial model over Binomial models in these simulation cases. Cadigan and Bataineh (2012) noticed a similar issue with the Beta-Binomial models when applied to paired catches simulated from independent negative binomial distributions although their simulation study did not include length-dependent relative catch efficiency. This issue is also a result of the bias-variance trade-off (see detailed discussion on model comparison between Binomial and Beta-Binomial models in Section 5).

To further investigate how model mis-specification (or mismatch between the simulation model and estimation model) impacts performance, we examine residuals from models BI3 and BB3 (Figures 56 and 57) and compare between scenarios with the NB simulation and scenarios using the resampling method. These two models have a balanced performance between convergence and fit among models in each category and are representative. In both scenarios, estimates from BB3 are smoother, more consistent, and less dispersed among 100 iterations. This can be explained by the additional variance parameter in the Beta-Binomial model that accounts for potential divergence in the relative catch efficiency from the truth. This comparison indicates that the Beta-Binomial model is not as effective as the Binomial model when dealing with independent negative binomial assumption of paired catches and may be inherently biased - the additional structural assumption in the Beta-Binomial models is specified as a Beta-prior for station variation in the relative catch efficiency, and hence when there is strong model mis-specification, i.e., divergence from this prior, model performance can be expected to deteriorate.

The Binomial models, on the other hand, do not model the variance explicitly, and thus are less subject to constraints in smoothness of the estimates. The fixed variance of the Binomial models can cause the model to "chase" the shape of the sample conversion when smoothing over length, and to result in more wiggly estimates in individual cases as was often the case in the simulation study in Section 3. This bias-variance trade-off demonstrates that all model assumption come with consequences, and the two model categories balance bias and variance differently.

The suite of models are sensitive to mis-specification and consequently the parametric NB simulation can lead to an inaccurate conclusion of model performance. We examine model performance under the resampling simulation scenarios: in scenarios 10-4-NB and 14-4-NB, BB3 results in small biases while BI3 does not have such biases; in scenario 14-4-R, BB3 resulted in similar residual bias with BI3, while in scenario 10-4-R, BB3 has lower estimation bias than BI3 (Figures 56 and 57). These results are reasonable as the suite of models are expected to have improved performance for more abundant species.

In conclusion, the resampling simulation is recommended, and model performance under the resampling simulation is satisfactory. A comprehensive simulation study should consider the mismatch between the simulation model and the estimation model, and assess the effect of potential model mis-specifications to avoid partial conclusions. One estimation model may perform better than another model for a specific type of data. Failure to account for mis-specification may produce a prejudiced and biased simulation study. In addition, statistical methods play an important role in ecological modeling, and modeling fisheries count data can be tricky (Elphick 2008). Model robustness to mis-specification should not be ignored since model assumptions come with consequences, and any additional assumption should be evaluated to achieve a desired balance between bias and variance in the estimator. We believe the resampling method is more suitable than the parametric NB assumption at simulating realistic paired catches for the comparative fishing experiment, and the proposed estimation method including the suite of models is recommended for analysis. However, in practice, it is highly recommended to use caution, for example, by assessing the data and comparing between models before automatic selection of the best model based on AIC.

8 Sampling Variability

Simulation of sampling error (as a measure of how sampled values differ from "truth") in Section 3 included two components. Variation in the swept area of a tow is simulated by a Gaussian distribution which is bi-directional and symmetrical. This variation may broadly represent gear performance deviations that apply to catches of all lengths of a species within the same tow. However, sampling variability of this component is generally small. The major component of sampling variability was simulated by a lognormal distribution to represent sampling errors in survey catches, i.e., their deviations from simulated true underlying available density. Simulation of sampling error in catches depended on the sample coefficient of variation in past RV surveys to be realistic and varied over length (depending on data sufficiency, see Section 3 for details). In practice, errors were generated from the lognormal distribution for each length bin independently; however, errors may be autocorrelated within lengths. The simulation study in Sections 3 did not account for autocorrelated sampling errors due to its ambiguous and abstruse nature.

Continuity in survey catches is expected despite large sampling variability in RV surveys; in fact, both the mean (relates to underlying available density) and variance (relates to coefficient of variation) demonstrated this continuity over length in past RV surveys (Figure 9 and Figure 10, respectively). As such, continuity within underlying density and coefficient of variation in survey catches were both considered in the simulation. However, the superimposed sampling errors are less conclusive. Anecdotally, one possibility for measurement error may come from human errors in counting fish, e.g., random factors in length measurements leading to variations in catch numbers. Mis-classification errors may contributed to negative autocorrelation in neighboring length bins but are impossible to prove or quantify solely based on past survey data.

In order to assess this sampling error, we assume that catch-length composition from a tow is an unbiased sampling of the underlying available density at this location; deviations of sampled catch-length composition from its moving average (i.e., anomalies sequence) may reflect the scale of sampling variability since the smoothed catch-length composition is an unbiased and non-parametric estimator. Given sufficient number of repeated sampling (i.e., stations), we calculated autocorrelation from these anomaly sequences (i.e., cross-correlation at lag one; equivalent to autocorrelation coefficient of an AR(1) model). Figure 58 illustrates distributions of this autocorrelation for each assessed species (where 5-point moving average was used for smoothed catch-length composition). Estimated autocorrelation coefficients were highly variable depending on the moving average and data sufficiency of the assessed species; for more abundant species haddock, silver skate and redfish (species 11, 14 and 23 in Figure 58, respectively), these coefficients mostly did not show conclusive evidence for a strong positive autocorrelation within length that is akin to survey catch-length composition .

The above assessment of sampling error autocorrelation is subject to high uncertainty but nevertheless could provide some guidance on simulation of such autocorrelation. We adapted the simulation study in Section 3 with these estimated coefficients. More specifically, to incorporate autocorrelation among lengths in the sampling process, the lognormal distributions in Equation 21 that applied to each length individually were changed to a multivariate lognormal distribution (MVLN),

$$\{\eta_{gi}(L), L \in LR\} \sim MVLN(1, \Sigma), \tag{45}$$

where $\eta_{gi}(L)$ is the sampling error for length L and gear g at station i, LR denotes the length range depending on the species, and Σ is the MVLN covariance matrix for sampling errors over all lengths. The covariance matrix was derived from sampling variances for each length bin (based on the sample coefficient of variation, same as in Equation 21) and a Toeplitz correlation matrix (based on an auto-regressive correlation model with coefficient ρ); covariance between errors at length L_m and L_n is defined by

$$cov[\eta_{gi}(L_m), \eta_{gi}(L_n)] = \Sigma_{m,n} = \rho^{|m-n|}\sqrt{log(\widehat{CV(L_m)}^2 + 1)}\sqrt{log(\widehat{CV(L_n)}^2 + 1)}. \tag{46}$$

For the simulation study in this section, ρ was resampled from estimated coefficients from each station in past RV surveys (Figure 58).

In this section, we tested autocorrelated sampling errors on an example species, silver hake, considering relatively more credible estimations of ρ based on abundant data. All procedures of simulation remain the same as in Section 3 except the generation of sampling errors in Equation 21 which was replaced by Equation 45. Simulation scenarios 14-1, 14-2, 14-3 and 14-4 as defined in Section 3 were modified accordingly in this section and coded as 14-1-AE,

14-2-AE, 14-3-AE and 14-4-AE ("AE" for autocorrelated errors). Results for these simulation scenarios are in Table 24.

Results for simulation scenarios 14-1-AE, 14-2-AE, 14-3-AE and 14-4-AE did not differ significantly from scenarios 14-1, 14-2, 14-3 and 14-4; simulated autocorrelation were mostly close to zero (see Figure 58, species 14), hence, these additional simulation scenarios did not introduce much difference in their generated datasets. This indicates that simulations in Section 3 and 4 were adequate. However, upon further investigation, a high positive autocorrelation (i.e., close to 1) would generally simulate more structured errors that lead to less accurate and more divergent estimates (e.g., higher $s(\rho)$, lower $P.20$): structured errors contribute to more uniform deviations between the true values and simulated samples, and since estimates from these binomial and beta-binomial models tend to follow the sample mean due to the smoothing technique, model performances can decrease consequently. This should be considered if there is conclusive evidence to support a high autocorrelation within sampling errors.

9 Conclusion

As a summary, this section provides a list of conclusions for the simulation studies in this document for the suite of Binomial and Beta-Binomial models that are proposed for analysis of paired catches from the comparative fishing experiment (Section 2):

1. A first simulation study was designed to assess performance of the analysis method under various scenarios (Section 3). These scenarios include four assumptions for the true underlying relative catch efficiency between paired gears and five archetypal species. Simulation of paired catches is based on resampling of data from past Maritimes RV survey and two stations are sampled for each of 48 strata. Results (Section 4) indicated that species abundance and population density are key factors to generating effective data during the comparative fishing survey and are impactful for model performance: estimated conversion factor for abundant species such as haddock, silver skate, and redfish can contain loss (over-estimation or under-estimation of converted total biomass) within 20%; for less abundance species such as Atlantic cod and Winter skate, estimation quality tend to be sensitive to nuanced factors due to data deficiency but can mostly contain loss within 50%.

2. The analysis method can efficiently recover length effects in the relative catch efficiency (due to different size-selectivity between the pair of gears) in most scenarios, but station effects in the relative catch efficiency (due to sampling variability at stations, fine-scale heterogeneity in distribution etc.) can significantly reduce model performance especially when compounded with low population density (Section 4).

3. The Beta-Binomial models tend to estimate a smoother shape in relative catch efficiency than the Binomial models. Depending on the nature of catch dispersion, the Beta-Binomial models may result in a small estimation bias; the Binomial models, on the other hand, often have more estimation uncertainty. The Beta-Binomial models have better overall performance considering both bias and variance. In practice, model selection should

be guided by meticulous comparison of estimation results from the suite of models (Section 5).

4. A second simulation study indicated that indiscriminately increasing sample size over the survey area may not result in estimation improvement (Section 6). Increasing sample size is only effective if the additional paired tows sample overall densities and length-specific densities that contrast with those covered by the original sampling, e.g., via targeted stations. However, tactical targeting may be difficult to implement in practice due to the need to optimize efficiency over many species and because specific high density locations may be difficult to identify for some species.

5. In a simulation study, model mis-specification (mis-match between the simulation model and estimation model) can lead to biased or partial results. A third simulation study compared different simulation methods and results indicated that the resampling technique is the most realistic, reaffirming conclusions from the first simulation study and validating the performance of the proposed analysis method (Section 7).

10 Acknowledgements

We would like to thank the reviewers, Noel Cadigan and David M. Keith, for their thoughtful comments and suggestions which have helped us improving this technical report.

11 References

Benoît, H.P. 2006. Standardizing the southern Gulf of St. Lawrence bottom-trawl survey time series: results of the 2004-2005 comparative fishing experiments and other recommendations for the analysis of the survey data. DFO Can. Sci. Adv. Sec. Res. Doc. 2006/008.

Bourdages, H., Savard, L., Archambault, D., and Valois, S. 2007. Results from the August 2004 and 2005 comparative fishing experiments in the northern Gulf of St. Lawrence between the CCGS Alfred Needler and the CCGS Teleost. Can. Tech. Rep. Fish. Aquat. Sci. 2750: ix + 57 p.

Brewer, M.J., Butler, A., and Cooksley, S.L. 2016. The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity. Methods in Ecology and Evolution **7**(6): 679–692.

Cadigan, N.G. and Bataineh, O. 2012. Inference about the ratio of means from negative binomial paired count data. Environmental and Ecological Statistics **19**(2): 269–293.

Cadigan, N.G., Yin, Y., and P. Benoît, H. A nonparametric-monotone regression model and robust estimation for paired-tow bottom-trawl survey comparative fishing data. Manuscript submitted for publication.

Cadigan, N.G. and Dowden, J.J. 2010. Statistical inference about the relative efficiency of a new survey protocol, based on paired-tow survey calibration data. Fishery Bulletin **108**(1).

Cadigan, N.G., Brodie, W., and Walsh, S.J. 2006. Relative efficiency of the Wilfred Templeman and Alfred Needler research vessels using a Campelen 1800 shrimp trawl in NAFO Subdivision 3Ps and Divisions 3LN. DFO Can. Sci. Adv. Sec. Res. Doc. 2006/085.

Dick, E. 2004. Beyond "lognormal versus gamma": discrimination among error distributions for generalized linear models. Fisheries Research **70**(2-3): 351–366.

Dietz, E. and Böhning, D. 2000. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis **34**(4): 441–459.

Ding, J., Tarokh, V., and Yang, Y. 2018. Model selection techniques: An overview. IEEE Signal Processing Magazine **35**(6): 16–34.

Elphick, C.S. 2008. How you count counts: the importance of methods research in applied ecology. Journal of Applied Ecology **45**(5): 1313–1320.

Fanning, L. 1985. Intercalibration of research survey results obtained by different vessels. Canadian Atlantic Fisheries Scientific Advisory Committee Research Document 85/3.

Friedman, J., Hastie, T., and Tibshirani, R. 2001. The elements of statistical learning, volume 1. Springer series in statistics New York.

Green, P.J. and Silverman, B.W. 1993. Nonparametric regression and generalized linear models: a roughness penalty approach. Crc Press.

Greene, W.H. 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.

Griffin, T.F. 1992. Distribution of the ratio of two Poisson random variables. Ph.D. thesis.

Harrison, X.A. 2015. A comparison of observation-level random effect and beta-binomial models for modelling overdispersion in binomial data in ecology & evolution. PeerJ **3**: e1114.

Hilborn, R. and Walters, C.J. 2013. Quantitative fisheries stock assessment: choice, dynamics and uncertainty. Springer Science & Business Media.

Hughes, G. and Madden, L. 1993. Using the beta-binomial distribution to describe aggregated patterns of disease incidence. Phytopathology (USA) .

Johnson, N.L., Kemp, A.W., and Kotz, S. 2005. Univariate discrete distributions, volume 444. John Wiley & Sons.

Kim, J. and Lee, J.H. 2017. The validation of a beta-binomial model for overdispersed binomial data. Communications in Statistics-Simulation and Computation **46**(2): 807–814.

Koeller, P. and Smith, S. 1983. Preliminary analysis of at cameron–lady hammond comparative fishing experiments 1979–1981. Canadian Atlantic Fisheries Scientific Advisory Committee Research Document 83/59.

Krause, A.E., Hayes, D.B., Bence, J.R., Madenjian, C.P., and Stedman, R.M. 2002. Measurement error associated with surveys of fish abundance in lake michigan. Journal of Great Lakes Research **28**(1): 44–51.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B. 2015. Tmb: automatic differentiation and Laplace approximation. arXiv preprint arXiv:1509.00660 .

Lewy, P., Nielsen, J.R., and Hovgård, H. 2004. Survey gear calibration independent of spatial fish distribution. Canadian Journal of Fisheries and Aquatic Sciences **61**(4): 636–647.

Miller, T.J. 2013. A comparison of hierarchical models for relative catch efficiency based on paired-gear data for us northwest atlantic fish stocks. Canadian Journal of Fisheries and Aquatic Sciences **70**(9): 1306–1316.

Miller, T.J., Das, C., Politis, P.J., Miller, A.S., Lucey, S.M., Legault, C.M., Brown, R.W., and Rago, P.J. 2010. Estimation of Albatross IV to Henry B. Bigelow calibration factors .

Morris, T.P., White, I.R., and Crowther, M.J. 2019. Using simulation studies to evaluate statistical methods. Statistics in medicine **38**(11): 2074–2102.

Nemes, S., Jonasson, J.M., Genell, A., and Steineck, G. 2009. Bias in odds ratios by logistic regression modelling and sample size. BMC Medical Research Methodology **9**(1): 1–5.

R Core Team 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schuckers, M.E. 2003. Using the beta-binomial distribution to assess performance of a biometric identification device. International Journal of Image and Graphics **3**(03): 523–529.

Smith, S.J. 1990. Use of statistical models for the estimation of abundance from groundfish trawl survey data. Canadian Journal of Fisheries and Aquatic Sciences **47**(5): 894–903.

Thorson, J.T. and Minto, C. 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES Journal of Marine Science **72**(5): 1245–1256.

Thygesen, U.H., Kristensen, K., Jansen, T., and Beyer, J.E. 2019. Intercalibration of survey methods using paired fishing operations and log-Gaussian cox processes. ICES Journal of Marine Science **76**(4): 1189–1199.

Venables, W.N. and Dichmont, C.M. 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fisheries research **70**(2-3): 319–337.

Ver Hoef, J.M. and Boveng, P.L. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology **88**(11): 2766–2772.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., and Welham, S.J. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. Journal of the Royal Statistical Society: Series C (Applied Statistics) **48**(3): 269–311.

Wood, S.N. 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **62**(2): 413–428.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(1): 3–36.

Wood, S.N. 2017. Generalized additive models: an introduction with R. CRC press.

12 Tables

Table 1. A set of binomial models with various assumptions on the length effect and station effect in the relative catch efficiency. A smoothing length effect can be considered and the station effect can be added to the intercept, without interaction with the length effect, or added to both the intercept and smoother to allow for interaction between the two effects.

Model	$log(\rho)$	Description
BI_0	β_0	no length effect, no station variation
BI_1	$\beta_0 + \delta_{0,i}$	no length effect, with random station variation on the intercept
BI_2	$\mathbf{X}_f^T \beta_f + \mathbf{X}_r^T \mathbf{b}$	smoothing length effect, without station variation
BI_3	$\mathbf{X}_f^T \beta_f + \mathbf{X}_r^T \mathbf{b} + \delta_{0,i}$	smoothing length effect, with random station variation on the intercept
BI_4	$\mathbf{X}_f^T (\beta_f + \delta_i) + \mathbf{X}_r^T (\mathbf{b} + \epsilon_i)$	smoothing length effect, with random station variation of the smoother

Table 2. A set of beta-binomial models with various assumptions on the length effect and station effect in the relative catch efficiency, and the length effect on the variance parameter. A smoothing length effect can be considered in both the conversion factor and the variance parameter. A possible station effect can be added to the intercept, without interaction with the length effect, or added to both the intercept and the smoother to allow for interaction between the two effects.

Model	$log(\rho)$	$log(\phi)$	Description
BB_0	β_0	γ_0	no length effect, no station variation, constant variance over length
BB_1	$\beta_0 + \delta_{0,i}$	γ_0	no length effect, random station variation on the intercept, constant variance over length
BB_2	$\mathbf{X}_f^T\beta_f + \mathbf{X}_r^T\mathbf{b}$	γ_0	smoothing length effect, no station variation, constant variance over length
BB_3	$\mathbf{X}_f^T\beta_f + \mathbf{X}_r^T\mathbf{b}$	$\mathbf{X}_f^T\gamma + \mathbf{X}_r^T\mathbf{g}$	smoothing length effect, no station variation, smoothing length effect in variance parameter
BB_4	$\mathbf{X}_f^T\beta_f + \mathbf{X}_r^T\mathbf{b} + \delta_{0,i}$	γ_0	smoothing length effect, random station variation on the intercept, constant variance over length
BB_5	$\mathbf{X}_f^T\beta_f + \mathbf{X}_r^T\mathbf{b} + \delta_{0,i}$	$\mathbf{X}_f^T\gamma + \mathbf{X}_r^T\mathbf{g}$	smoothing length effect, random station variation on the intercept, smoothing length effect in variance parameter
BB_6	$\mathbf{X}_f^T(\beta_f + \delta_i) + \mathbf{X}_r^T(\mathbf{b} + \epsilon_i)$	γ_0	smoothing length effect, random station variation of the smoother, constant variance over length
BB_7	$\mathbf{X}_f^T(\beta_f + \delta_i) + \mathbf{X}_r^T(\mathbf{b} + \epsilon_i)$	$\mathbf{X}_f^T\gamma + \mathbf{X}_r^T\mathbf{g}$	smoothing length effect, random station variation of the smoother, smoothing length effect in variance parameter

Table 3. Parameter values used in the exponential model for simulating the third assumption of true relative catch efficiency (Section 3.1.2). Values are extracted for each species from parameter estimations of the exponential model (Bourdages et al. 2007); Bourdages et al. (2007) did not provide an estimation for Atlantic Cod so the same model is assumed for this species as Haddock considering similar length range of the two species.

Species		Parameter		
Name	Code	a	b	c
Atlantic Cod	10	1.1003	4.8778	0.0594
Haddock	11	1.1003	4.8778	0.0594
Silver Hake	14	-4.0647	5.2651	0.0018
Redfish	23	0.7653	2.4198	0.2607
Winter Skate	204	1.6172	-3.8652	0.1199

Table 4. Result summary of the simulation study of the 20 simulation scenarios in Section 3, including a summary of simulated datasets and estimation results using the suite of Binomial and Beta-Binomial models. $LE(\rho)$, $SE(\rho)$ and $Lin(\rho)$ assess length effect, station effect and nonlinearity of the true conversion, respectively; $EN(stn)$, $P(enc)$, $MED(C)$, $PD(C)$ summarize number of effective stations, encounter probability, median positive catch, and pair differences of the simulated datasets, respectively; $\bar{e}(\rho)$, $s(\rho)$, $P.20$ and $P.50$ measure residual mean, standard deviation and proportions of loss of 20% and 50%, respectively. See Section 3 for detailed definition of the performance measures.

Scenario	Simulated Data Summary							Estimation Residuals Summary			
	Relative Catch Efficiency			Population Density			Pair	Weighted-Mean Residuals			
	$LE(\rho)$	$SE(\rho)$	$Lin(\rho)$	$EN(stn)$	$P(enc)$	$MED(C)$	$PD(C)$	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
10-1	0.00	0.00	0.00	8	0.02	4.00	1.73	-0.0261	0.1638	0.78	0.99
10-2	0.76	0.00	0.11	11	0.03	5.00	1.52	0.0675	0.0766	0.94	1.00
10-3	0.11	0.00	0.02	30	0.05	21.00	1.68	0.0866	0.0358	1.00	1.00
10-4	0.47	0.38	0.01	17	0.04	7.00	1.60	0.1295	0.1779	0.55	0.99
11-1	0.00	0.00	0.00	59	0.23	76.75	1.43	-0.0039	0.0444	1.00	1.00
11-2	0.48	0.00	0.08	60	0.24	100.00	1.27	-0.0668	0.0324	1.00	1.00
11-3	0.10	0.00	0.01	69	0.30	402.50	1.39	-0.0708	0.0145	1.00	1.00
11-4	0.23	0.20	0.01	65	0.27	157.75	1.32	-0.0607	0.0302	1.00	1.00
14-1	0.00	0.00	0.00	56	0.19	92.00	1.42	-0.0029	0.0472	1.00	1.00
14-2	0.43	0.00	0.08	59	0.20	114.00	1.30	-0.0744	0.0322	1.00	1.00
14-3	0.06	0.00	0.00	61	0.23	151.50	1.32	-0.0760	0.0231	1.00	1.00
14-4	0.22	0.22	0.01	62	0.23	196.75	1.34	-0.0711	0.0296	1.00	1.00
23-1	0.00	0.00	0.00	41	0.13	33.00	1.52	-0.0002	0.0708	1.00	1.00
23-2	0.27	0.00	0.05	42	0.14	37.00	1.40	-0.0956	0.0579	0.96	1.00
23-3	0.08	0.00	0.02	46	0.15	51.75	1.43	-0.1163	0.0465	0.97	1.00
23-4	0.13	0.14	0.00	50	0.16	72.75	1.46	-0.1126	0.0411	0.99	1.00
204-1	0.00	0.00	0.00	1	0.00	5.00	1.79	-0.2477	0.7507	0.56	0.86
204-2	0.62	0.00	0.11	2	0.01	5.00	1.50	0.1315	0.1724	0.67	0.99
204-3	0.09	0.00	0.02	3	0.01	9.00	1.72	0.1080	0.0449	0.99	1.00
204-4	0.46	0.41	0.01	2	0.01	6.75	1.52	0.1995	0.2119	0.42	0.95

Table 5. Summary of convergence results for the simulation study with the 20 simulation scenarios described in Section 3. Each scenario includes 100 iterations (only iterations are retained in which the simulation generated a minimum of five catches for each gear) and the 13 Binomial and Beta-Binomial models are fit in the estimation step within each iteration. The number of iterations where each model properly converged is provided. See Tables 1 and 2 for model specifications.

Scenario	BI0	BI1	BI2	BI3	BI4	BB0	BB1	BB2	BB3	BB4	BB5	BB6	BB7
10-1	100	100	100	90	1	98	93	93	67	82	62	48	9
10-2	100	100	100	100	0	99	98	97	89	96	69	57	13
10-3	100	100	100	98	9	100	96	91	80	89	72	14	9
10-4	100	100	99	100	7	100	96	98	88	92	84	17	20
11-1	100	100	100	100	0	100	100	99	97	64	62	22	10
11-2	100	100	100	100	3	100	100	100	100	100	79	39	11
11-3	100	100	98	100	11	100	100	100	94	98	94	22	23
11-4	100	100	99	100	9	100	100	99	92	97	94	5	29
14-1	100	100	100	100	6	100	100	100	100	55	74	39	14
14-2	100	100	100	100	3	100	100	99	100	97	99	43	26
14-3	100	100	100	100	7	100	100	100	98	82	95	35	15
14-4	100	100	98	100	8	100	100	99	100	100	100	8	31
23-1	100	100	100	100	11	100	100	99	100	67	75	44	14
23-2	100	100	100	99	10	100	100	99	99	95	96	57	17
23-3	100	100	100	100	14	100	100	94	100	90	98	45	18
23-4	100	100	98	100	23	100	100	99	100	96	98	10	18
204-1	71	71	71	61	0	28	9	8	0	4	0	0	0
204-2	76	76	76	63	0	44	20	14	0	3	0	1	0
204-3	76	76	75	62	1	44	41	9	0	11	0	1	0
204-4	81	81	80	71	1	49	32	14	1	3	1	1	0

Table 6. Summary of model selection results of the simulation study with the 20 simulation scenarios described in Section 3. Each scenario includes 100 iterations (only iterations are retained in which the simulation generated a minimum of five catches for each gear) and the 13 Binomial and Beta-Binomial models are fit in the estimation step within each iteration. The number of iterations where each model properly converged is provided. See Tables 1 and 2 for model specifications.

Scenario	BI0	BI1	BI2	BI3	BI4	BB0	BB1	BB2	BB3	BB4	BB5	BB6	BB7
10-1	0	3	0	12	1	0	0	0	2	17	11	46	8
10-2	0	0	0	5	0	0	1	2	3	10	18	48	13
10-3	0	0	3	10	7	0	0	2	14	6	36	13	9
10-4	0	0	3	6	1	0	0	14	15	14	19	12	16
11-1	0	0	0	0	0	0	17	0	3	15	38	17	10
11-2	0	0	0	0	0	0	2	0	0	17	37	33	11
11-3	0	0	0	0	0	0	0	0	0	6	58	16	20
11-4	0	0	0	0	0	0	0	0	5	5	59	2	29
14-1	0	0	0	0	0	0	10	0	0	12	35	29	14
14-2	0	0	0	0	0	0	1	0	0	10	37	28	24
14-3	0	0	0	0	0	0	2	0	0	21	34	31	12
14-4	0	0	0	0	0	0	0	0	0	26	41	4	29
23-1	0	0	0	0	0	0	9	0	0	10	31	37	13
23-2	0	0	0	0	0	0	0	0	0	8	34	44	14
23-3	0	0	0	0	0	0	1	1	0	16	32	36	14
23-4	0	0	0	0	0	0	0	4	0	40	30	9	17
204-1	0	0	9	55	0	0	0	3	0	4	0	0	0
204-2	0	2	24	44	0	0	0	3	0	2	0	1	0
204-3	0	1	36	33	1	0	0	3	0	1	0	1	0
204-4	0	1	36	36	1	0	1	2	1	2	0	1	0

Table 7. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 10-1. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.004	0.250	0.660	0.970
BI1	-0.014	0.175	0.750	0.990
BI2	-0.044	0.267	0.620	0.940
BI3	-0.111	0.415	0.644	0.933
BI4	-	-	-	-
BB0	-0.013	0.128	0.847	1.000
BB1	-0.006	0.129	0.839	1.000
BB2	-0.021	0.158	0.763	1.000
BB3	-0.016	0.129	0.836	1.000
BB4	-0.013	0.154	0.768	1.000
BB5	-0.018	0.134	0.839	1.000
BB6	-0.035	0.138	0.812	1.000
BB7	-	-	-	-
AIC	-0.026	0.164	0.780	0.990

Table 8. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 10-2. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.098	0.130	0.740	1.000
BI1	0.109	0.115	0.790	1.000
BI2	0.037	0.119	0.930	1.000
BI3	0.089	0.114	0.850	1.000
BI4	-	-	-	-
BB0	0.035	0.100	0.949	1.000
BB1	0.073	0.116	0.847	1.000
BB2	0.047	0.075	0.990	1.000
BB3	0.042	0.073	0.989	1.000
BB4	0.074	0.082	0.948	1.000
BB5	0.068	0.081	0.913	1.000
BB6	0.058	0.078	0.965	1.000
BB7	0.039	0.061	1.000	1.000
AIC	0.067	0.077	0.940	1.000

Table 9. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 10-3. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.184	0.061	0.460	1.000
BI1	0.220	0.018	0.130	1.000
BI2	0.063	0.055	0.990	1.000
BI3	0.112	0.032	1.000	1.000
BI4	-	-	-	-
BB0	0.139	0.037	0.960	1.000
BB1	0.201	0.026	0.479	1.000
BB2	0.057	0.040	1.000	1.000
BB3	0.050	0.043	1.000	1.000
BB4	0.100	0.031	1.000	1.000
BB5	0.088	0.032	1.000	1.000
BB6	0.095	0.036	1.000	1.000
BB7	-	-	-	-
AIC	0.087	0.036	1.000	1.000

Table 10. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 10-4. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.288	0.211	0.150	0.930
BI1	0.445	0.087	0.000	0.690
BI2	0.019	0.271	0.465	0.939
BI3	0.172	0.169	0.480	0.990
BI4	-	-	-	-
BB0	0.307	0.116	0.170	0.970
BB1	0.426	0.094	0.010	0.750
BB2	0.063	0.179	0.622	1.000
BB3	0.045	0.186	0.636	1.000
BB4	0.162	0.157	0.522	0.989
BB5	0.151	0.163	0.524	0.988
BB6	0.176	0.156	0.471	1.000
BB7	0.175	0.153	0.500	1.000
AIC	0.130	0.178	0.550	0.990

Table 11. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 14-1. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	-0.028	0.198	0.710	0.990
BI1	-0.015	0.101	0.950	1.000
BI2	-0.056	0.212	0.610	0.980
BI3	-0.047	0.135	0.870	0.990
BI4	-	-	-	-
BB0	-0.002	0.044	1.000	1.000
BB1	-0.002	0.044	1.000	1.000
BB2	-0.002	0.047	1.000	1.000
BB3	-0.002	0.046	1.000	1.000
BB4	-0.001	0.053	1.000	1.000
BB5	-0.005	0.048	1.000	1.000
BB6	-0.004	0.043	1.000	1.000
BB7	-0.004	0.045	1.000	1.000
AIC	-0.003	0.047	1.000	1.000

Table 12. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 14-2. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	-0.097	0.128	0.790	1.000
BI1	-0.053	0.066	0.980	1.000
BI2	-0.021	0.113	0.910	1.000
BI3	0.004	0.065	0.990	1.000
BI4	-	-	-	-
BB0	-0.149	0.037	0.910	1.000
BB1	-0.146	0.038	0.910	1.000
BB2	-0.074	0.031	1.000	1.000
BB3	-0.077	0.031	1.000	1.000
BB4	-0.071	0.030	1.000	1.000
BB5	-0.076	0.031	1.000	1.000
BB6	-0.068	0.032	1.000	1.000
BB7	-0.079	0.032	1.000	1.000
AIC	-0.074	0.032	1.000	1.000

Table 13. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 14-3. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.010	0.073	1.000	1.000
BI1	0.033	0.034	1.000	1.000
BI2	-0.006	0.080	1.000	1.000
BI3	0.021	0.044	1.000	1.000
BI4	-	-	-	-
BB0	-0.074	0.020	1.000	1.000
BB1	-0.072	0.020	1.000	1.000
BB2	-0.078	0.022	1.000	1.000
BB3	-0.081	0.024	1.000	1.000
BB4	-0.076	0.023	1.000	1.000
BB5	-0.081	0.023	1.000	1.000
BB6	-0.077	0.024	1.000	1.000
BB7	-0.084	0.027	1.000	1.000
AIC	-0.076	0.023	1.000	1.000

Table 14. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 14-4. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.067	0.060	0.990	1.000
BI1	0.058	0.030	1.000	1.000
BI2	0.026	0.064	0.990	1.000
BI3	0.029	0.041	1.000	1.000
BI4	-	-	-	-
BB0	-0.067	0.028	1.000	1.000
BB1	-0.054	0.029	1.000	1.000
BB2	-0.081	0.030	1.000	1.000
BB3	-0.087	0.030	1.000	1.000
BB4	-0.069	0.029	1.000	1.000
BB5	-0.074	0.029	1.000	1.000
BB6	-	-	-	-
BB7	-0.062	0.029	1.000	1.000
AIC	-0.071	0.030	1.000	1.000

Table 15. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 204-1. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	-0.064	0.215	0.648	0.958
BI1	-0.057	0.239	0.648	0.944
BI2	-0.248	0.751	0.563	0.859
BI3	-0.190	0.500	0.557	0.852
BI4	-	-	-	-
BB0	-0.075	0.195	0.643	0.964
BB1	-	-	-	-
BB2	-	-	-	-
BB3	-	-	-	-
BB4	-	-	-	-
BB5	-	-	-	-
BB6	-	-	-	-
BB7	-	-	-	-
AIC	-0.248	0.751	0.563	0.859

Table 16. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 204-2. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.178	0.094	0.539	1.000
BI1	0.232	0.139	0.421	0.947
BI2	0.127	0.171	0.671	0.987
BI3	0.194	0.138	0.603	0.968
BI4	-	-	-	-
BB0	0.186	0.101	0.523	1.000
BB1	0.243	0.142	0.300	0.950
BB2	0.121	0.089	0.786	1.000
BB3	-	-	-	-
BB4	-	-	-	-
BB5	-	-	-	-
BB6	-	-	-	-
BB7	-	-	-	-
AIC	0.132	0.172	0.671	0.987

Table 17. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 204-3. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.076	0.035	1.000	1.000
BI1	0.124	0.054	0.934	1.000
BI2	0.100	0.040	1.000	1.000
BI3	0.133	0.048	0.935	1.000
BI4	-	-	-	-
BB0	0.076	0.033	1.000	1.000
BB1	0.134	0.046	0.902	1.000
BB2	-	-	-	-
BB3	-	-	-	-
BB4	0.134	0.037	1.000	1.000
BB5	-	-	-	-
BB6	-	-	-	-
BB7	-	-	-	-
AIC	0.108	0.045	0.987	1.000

Table 18. Summary of estimation residuals from the 13 candidate Binomial, Beta-Binomial models and the AIC-selected best models for simulation scenario 204-4. See Section 3.3.2 for definition of the performance measures.

	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
BI0	0.275	0.159	0.284	0.938
BI1	0.346	0.177	0.210	0.802
BI2	0.188	0.205	0.438	0.963
BI3	0.271	0.204	0.296	0.887
BI4	-	-	-	-
BB0	0.256	0.151	0.286	0.959
BB1	0.356	0.175	0.188	0.812
BB2	0.218	0.174	0.500	0.929
BB3	-	-	-	-
BB4	-	-	-	-
BB5	-	-	-	-
BB6	-	-	-	-
BB7	-	-	-	-
AIC	0.200	0.212	0.420	0.951

Table 19. Summary of simulated datasets for the 36 simulation scenarios in Section 6. These scenarios are designed to assess the effect of sample sizes, e.g., scenarios 10-1-1, 10-1-2 and 10-1-3 use the same simulation settings based on Atlantic Cod and the first ρ-assumption but the number of stations per stratum are 1, 2 and 3, respectively. See Section 3 for details of the simulation process and definition of the summary metrics.

Scenario	Relative Catch Efficiency			Population Density			Pair
	$LE(\rho)$	$SE(\rho)$	$Lin(\rho)$	$EN(stn)$	$P(enc)$	$MED(C)$	$PD(C)$
10-1-1	0.00	0.00	0.00	4	0.02	4.00	1.76
10-1-2	0.00	0.00	0.00	8	0.02	4.00	1.73
10-1-3	0.00	0.00	0.00	12	0.02	4.00	1.71
10-2-1	0.76	0.00	0.11	6	0.03	5.00	1.53
10-2-2	0.76	0.00	0.11	11	0.03	5.00	1.52
10-2-3	0.76	0.00	0.11	17	0.03	4.75	1.51
10-3-1	0.11	0.00	0.02	15	0.05	24.50	1.70
10-3-2	0.11	0.00	0.02	30	0.05	21.00	1.68
10-3-3	0.11	0.00	0.02	45	0.05	21.25	1.68
10-4-1	0.46	0.36	0.01	9	0.04	8.00	1.63
10-4-2	0.47	0.38	0.01	17	0.04	7.00	1.60
10-4-3	0.47	0.38	0.01	26	0.04	7.00	1.58
14-1-1	0.00	0.00	0.00	84	0.19	96.00	1.42
14-1-2	0.00	0.00	0.00	56	0.19	92.00	1.42
14-1-3	0.00	0.00	0.00	84	0.19	96.00	1.42
14-2-1	0.43	0.00	0.08	30	0.21	118.50	1.29
14-2-2	0.43	0.00	0.08	59	0.20	114.00	1.30
14-2-3	0.43	0.00	0.08	88	0.20	113.25	1.29
14-3-1	0.06	0.00	0.00	31	0.23	153.75	1.31
14-3-2	0.06	0.00	0.00	61	0.23	151.50	1.32
14-3-3	0.06	0.00	0.00	91	0.22	151.75	1.32
14-4-1	0.22	0.21	0.01	31	0.23	187.00	1.34
14-4-2	0.22	0.22	0.01	62	0.23	196.75	1.34
14-4-3	0.22	0.22	0.01	94	0.23	189.50	1.33
204-1-1	0.00	0.00	0.00	0	0.00	4.00	2.00
204-1-2	0.00	0.00	0.00	1	0.00	5.00	1.79
204-1-3	0.00	0.00	0.00	1	0.00	3.25	1.70
204-2-1	0.62	0.00	0.11	1	0.01	5.00	1.71
204-2-2	0.62	0.00	0.11	2	0.01	5.00	1.50
204-2-3	0.62	0.00	0.11	2	0.01	4.75	1.44
204-3-1	0.09	0.00	0.02	1	0.01	10.00	1.90
204-3-2	0.09	0.00	0.02	3	0.01	9.00	1.72
204-3-3	0.09	0.00	0.02	5	0.01	10.00	1.72
204-4-1	0.46	0.39	0.01	1	0.01	6.00	1.63
204-4-2	0.46	0.41	0.01	2	0.01	6.75	1.52
204-4-3	0.46	0.41	0.01	2	0.01	6.00	1.51

Table 20. Summary of estimation results for the 36 simulation scenarios in Section 6: a comparison of estimation accuracy $\bar{e}(\rho)$, (see Section 3.3.2 for definition). These scenarios are designed to assess the effect of sample sizes, e.g., the first row includes the three scenarios 10-1-1, 10-1-2 and 10-1-3. See Section 6 for details of the simulation study.

Scenario	Sample Size		
	n=1	n=2	n=3
10-1	-0.0414	-0.0261	-0.0031
10-2	0.0538	0.0675	0.0595
10-3	0.0882	0.0866	0.0812
10-4	0.1027	0.1295	0.1467
14-1	0.0007	-0.0029	0.0006
14-2	-0.0840	-0.0744	-0.0765
14-3	-0.0762	-0.0760	-0.0802
14-4	-0.0755	-0.0711	-0.0779
204-1	-0.0640	-0.2477	-0.0672
204-2	0.1431	0.1315	0.1643
204-3	0.0976	0.1080	0.1138
204-4	-1.0954	0.1995	0.1957

Table 21. Summary of estimation results for the 36 simulation scenarios in Section 6: a comparison of estimation precision $s(\rho)$, (see Section 3.3.2 for definition). These scenarios are designed to assess the effect of sample sizes, e.g., the first row includes the three scenarios 10-1-1, 10-1-2 and 10-1-3. See Section 6 for details of the simulation study.

Scenario	Sample Size		
	n=1	n=2	n=3
10-1	0.2591	0.1638	0.1223
10-2	0.1404	0.0766	0.0658
10-3	0.0519	0.0358	0.0388
10-4	0.2870	0.1779	0.1524
14-1	0.0470	0.0472	0.0366
14-2	0.0469	0.0322	0.0286
14-3	0.0332	0.0231	0.0200
14-4	0.0380	0.0296	0.0233
204-1	0.4613	0.7507	0.3042
204-2	0.1324	0.1724	0.1092
204-3	0.0467	0.0449	0.0415
204-4	7.7243	0.2119	0.3548

Table 22. Summary of estimation results for the 36 simulation scenarios in Section 6: a comparison of estimation quality $P.20$, (capped estimation loss within 20%, see Section 3.3.2 for definition). These scenarios are designed to assess the effect of sample sizes, e.g., the first row includes the three scenarios 10-1-1, 10-1-2 and 10-1-3. See Section 6 for details of the simulation study.

Scenario	Sample Size		
	n=1	n=2	n=3
10-1	0.6500	0.7800	0.8800
10-2	0.8300	0.9400	0.9700
10-3	0.9900	1.0000	0.9900
10-4	0.5300	0.5500	0.6100
14-1	1.0000	1.0000	1.0000
14-2	1.0000	1.0000	1.0000
14-3	1.0000	1.0000	1.0000
14-4	1.0000	1.0000	1.0000
204-1	0.6667	0.5634	0.6374
204-2	0.6512	0.6711	0.6383
204-3	0.9767	0.9868	0.9681
204-4	0.4906	0.4198	0.3579

Table 23. Result summary of the simulation study in Section 7, including a summary of simulated datasets and estimation results using the suite of Binomial and Beta-Binomial models. These eight scenarios are designed to assess the impact of model mis-specification on estimation quality. See Section 3 for detailed definition of the performance measures.

Scenario	Simulated Data Summary							Estimation Residuals Summary			
	Relative Catch Efficiency			Population Density			Pair	Weighted-Mean Residuals			
	$LE(\rho)$	$SE(\rho)$	$Lin(\rho)$	$EN(stn)$	$P(enc)$	$MED(C)$	$PD(C)$	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
10-4-R	0.47	0.38	0.01	17	0.04	7.00	1.60	0.1295	0.1779	0.55	0.99
10-4-NB	0.47	0.38	0.01	19	0.04	7.00	1.70	-0.0621	0.1138	0.86	1.00
10-4-$Pois$	0.47	0.38	0.01	16	0.05	7.00	1.39	0.0249	0.1052	0.96	1.00
10-4-$NB-Beta$	0.47	0.38	0.01	20	0.04	7.00	1.75	0.0032	0.1494	0.95	0.99
14-4-R	0.22	0.22	0.01	62	0.23	196.75	1.34	-0.0711	0.0296	1.00	1.00
14-4-NB	0.22	0.21	0.01	76	0.24	102.50	1.61	-0.1496	0.0305	0.95	1.00
14-4-$Pois$	0.22	0.21	0.01	75	0.28	110.50	1.13	0.0089	0.0193	1.00	1.00
14-4-$NB-Beta$	0.22	0.21	0.01	77	0.21	98.25	1.76	0.0179	0.0262	1.00	1.00

Table 24. Result summary of the simulation study of the 4 simulation scenarios in Section 8, including a summary of simulated datasets and estimation results using the suite of Binomial and Beta-Binomial models. See Section 3 for detailed definition of the performance measures.

Scenario	Simulated Data Summary							Estimation Residuals Summary			
	Relative Catch Efficiency			Population Density			Pair	Weighted-Mean Residuals			
	$LE(\rho)$	$SE(\rho)$	$Lin(\rho)$	$EN(stn)$	$P(enc)$	$MED(C)$	$PD(C)$	$\bar{e}(\rho)$	$s(\rho)$	$P.20$	$P.50$
14-1	0.00	0.00	0.00	56	0.19	92.75	1.42	-0.0040	0.0449	1.00	1.00
14-2	0.43	0.00	0.08	59	0.21	114.00	1.29	-0.0709	0.0308	1.00	1.00
14-3	0.06	0.00	0.00	61	0.23	151.75	1.32	-0.0753	0.0231	1.00	1.00
14-4	0.22	0.22	0.01	63	0.23	196.50	1.34	-0.0730	0.0259	1.00	1.00

13 Figures

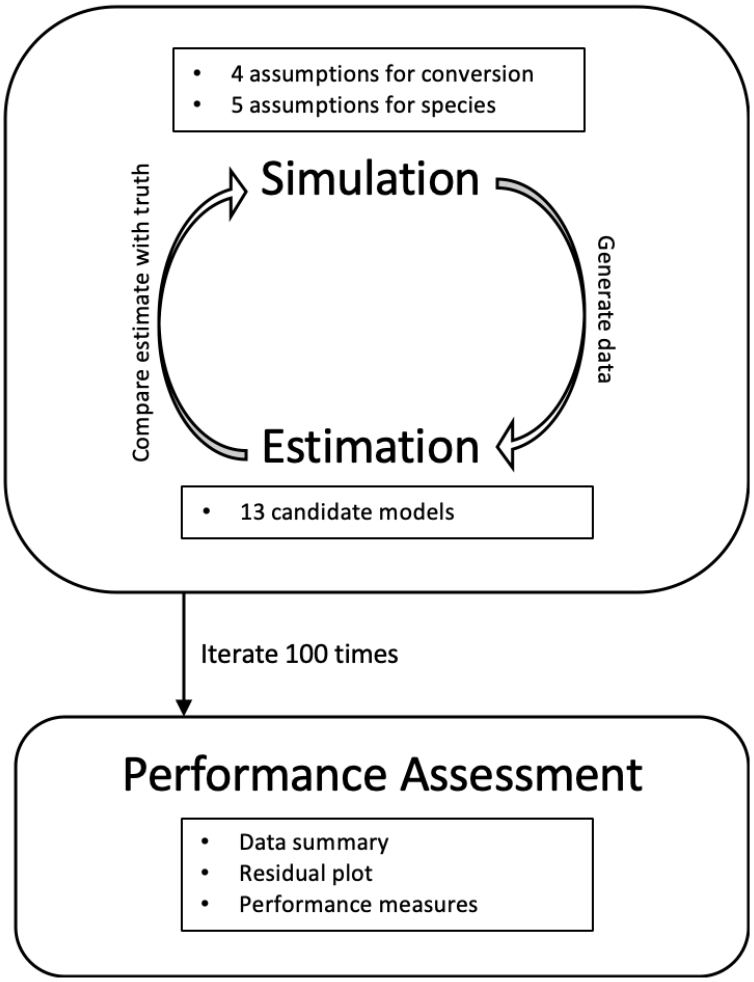

Figure 1. A schematic for the simulation study in Section 3. Each simulation scenario includes 100 iterations of the simulation cycle, where a dataset is simulated, then the 13 candidate models are estimated; performance of the suite of models is summarized for the 100 iterations for each scenario.

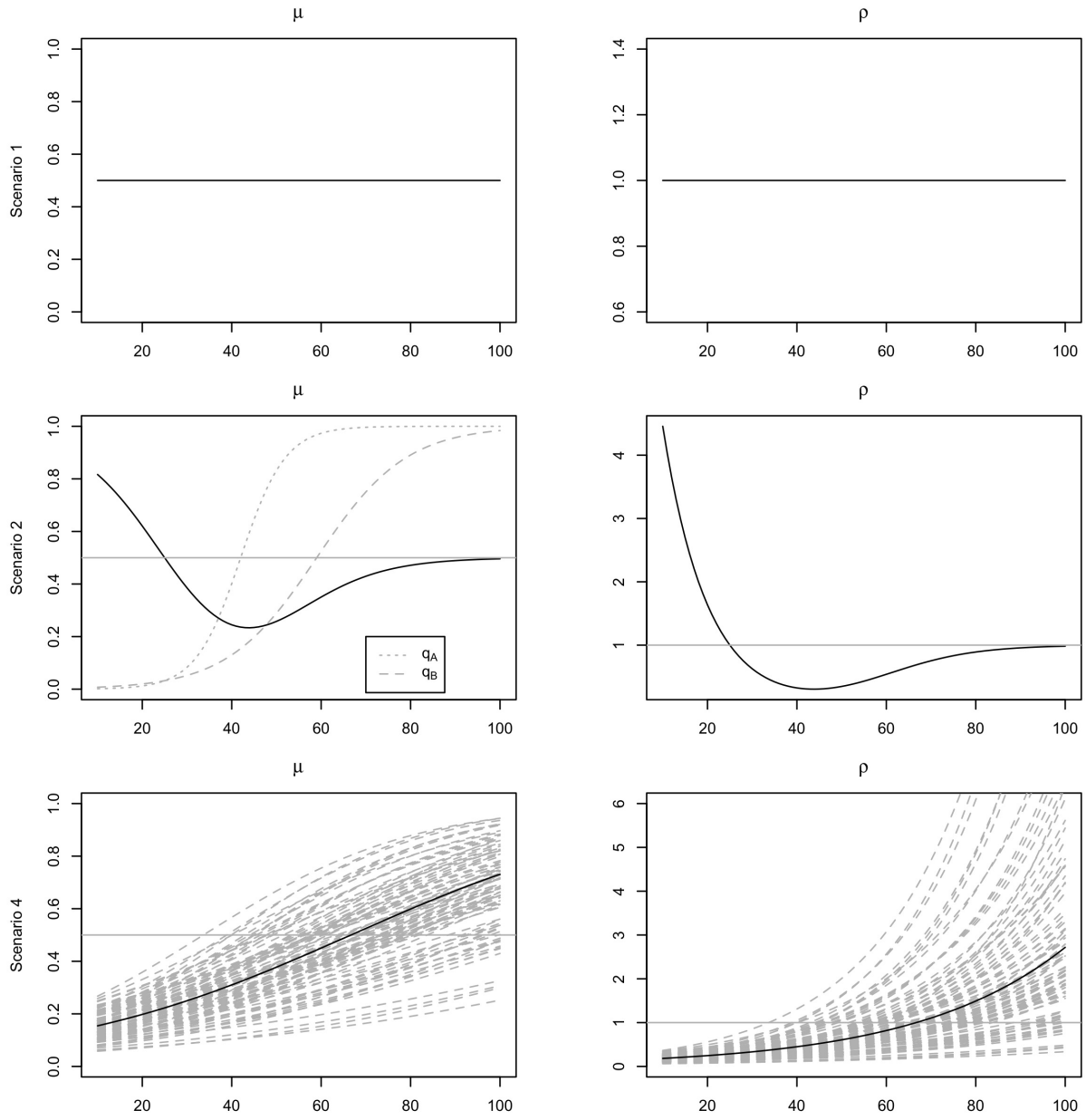

Figure 2. An illustration of ρ-assumptions 1, 2 and 4 (in rows 1, 2, and 3, respectively) with an example species within the length range 1-100 cm for the simulation study in Section 3. The left panels are the simulated proportion of catch by gear A within the pair of comparative fishing gears (μ, in black solid lines) and the right panels are the simulation relative catch efficiency between gears A and B (ρ, black solid lines). On the left panel in the second row, the dotted and segmented lines are simulated catchabilities for gears A and B, respectively. On the panels in the third row, segmented lines are the simulated μ and ρ for each station, respectively, and solid lines are the true μ and ρ before adding station variations.

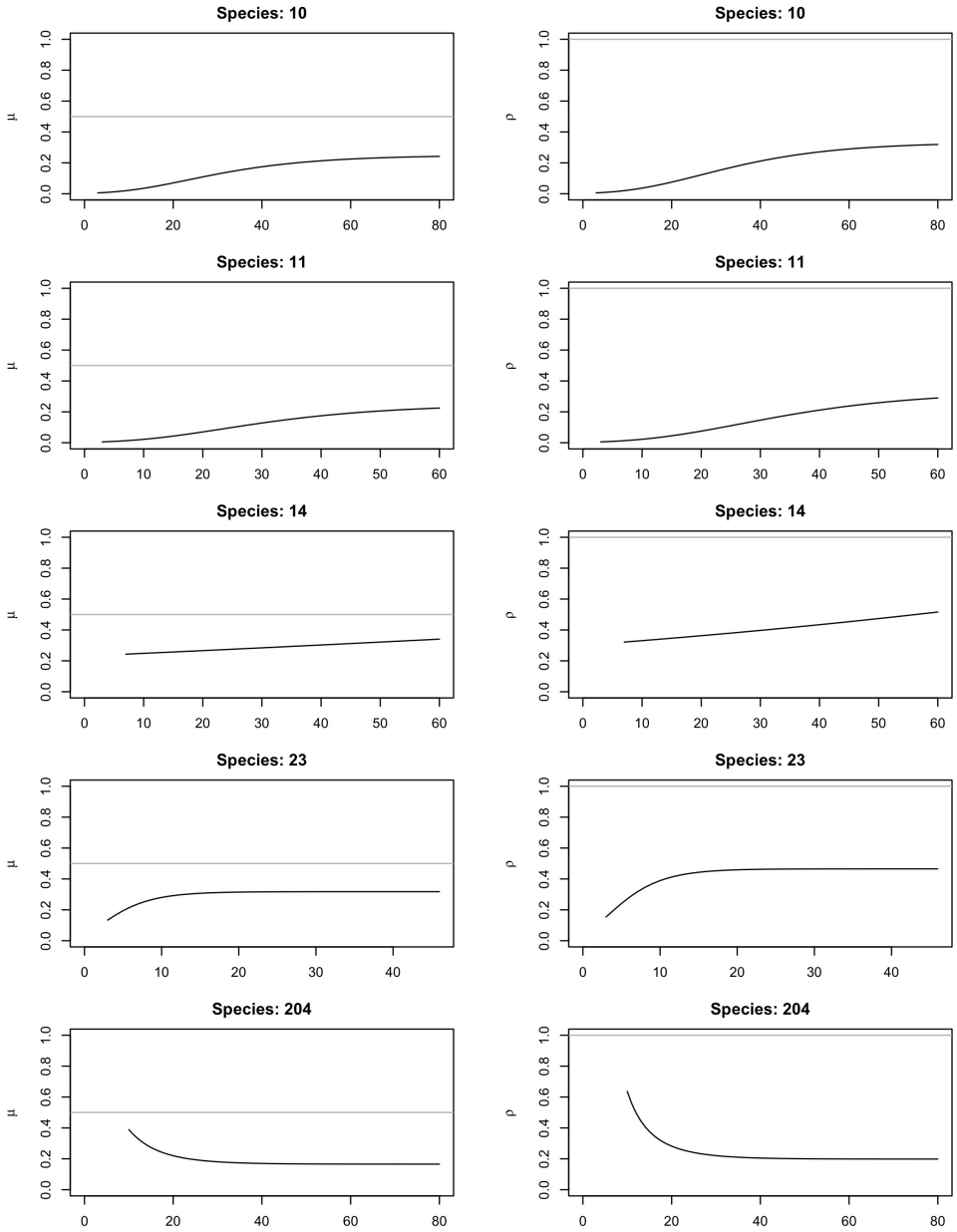

Figure 3. An illustration of ρ-assumptions 3 for the simulation study in Section 3. ρ-assumption 3 is species dependent and each row corresponds to each simulated species; the left panels are the simulated proportion of catch by gear A within the pair of comparative fishing gears (μ) and the right panels are the simulation relative catch efficiency between gears A and B (ρ).

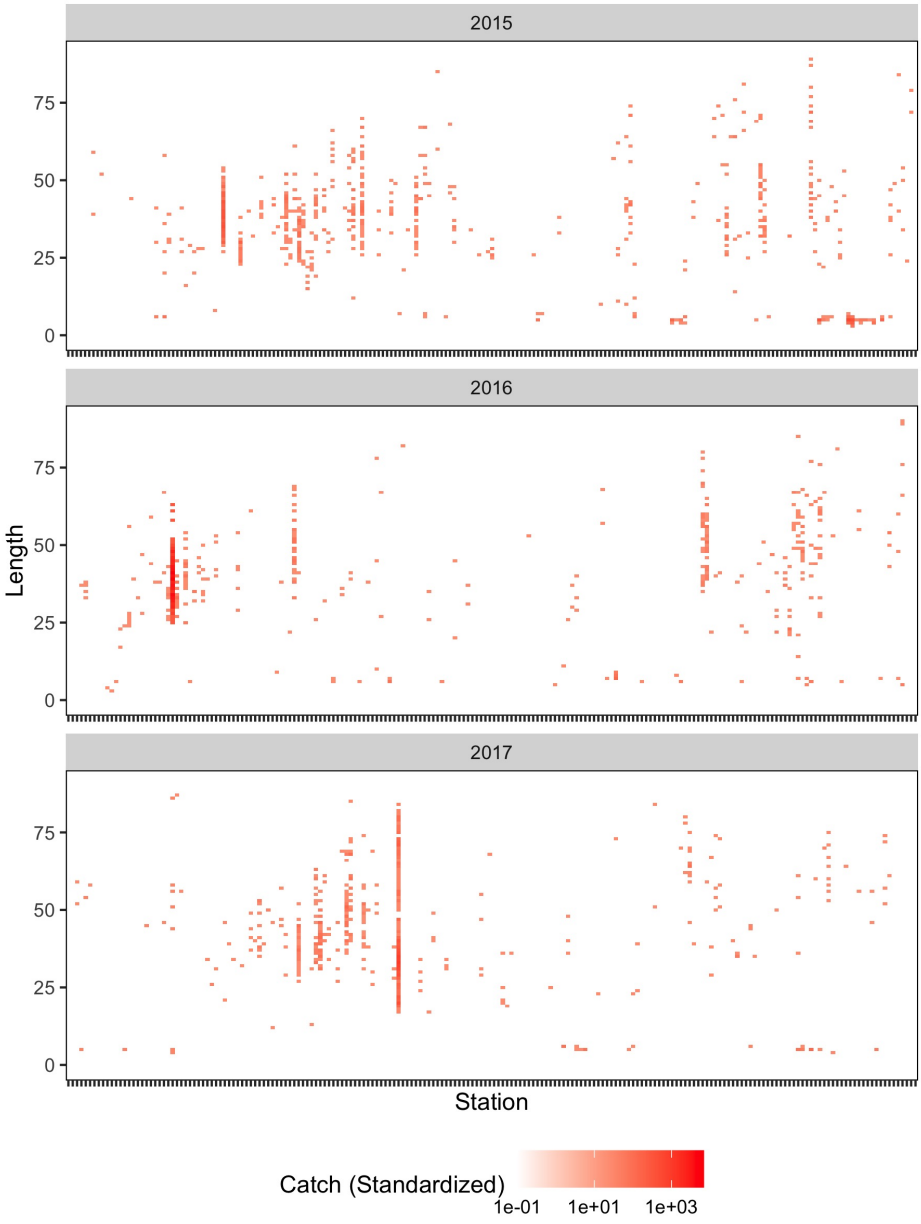

Figure 4. Catch numbers (effort-standardized) of species 10 from the past Maritimes surveys. Color intensity indicates catch at each length (cm) and station (for simplicity, station ID is not labelled). Stations are numbered sequentially according to the original order in which they were sampled.

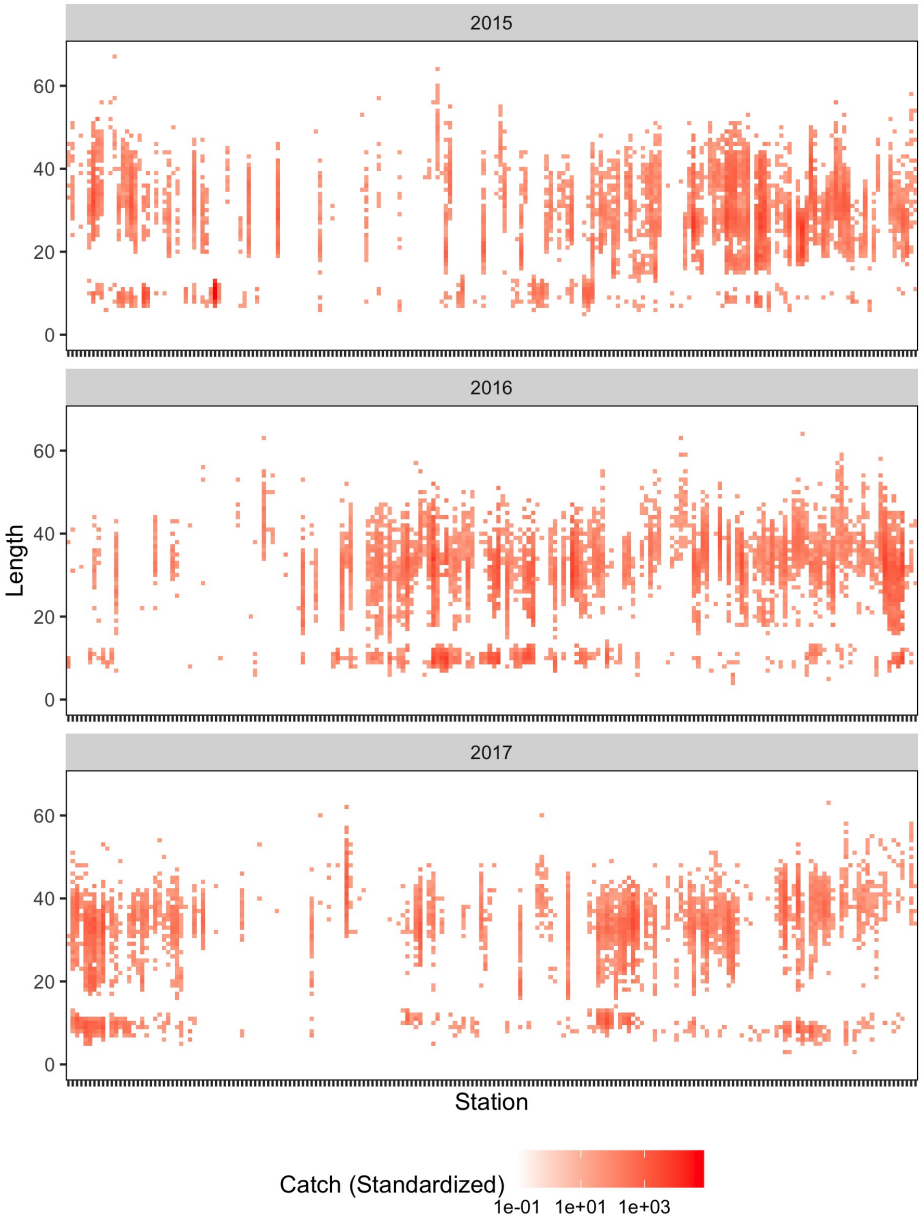

Figure 5. Catch numbers (effort-standardized) of species 11 from the past Maritimes surveys. Color intensity indicates catch at each length (cm) and station (for simplicity, station ID is not labelled). Stations are numbered sequentially according to the original order in which they were sampled.

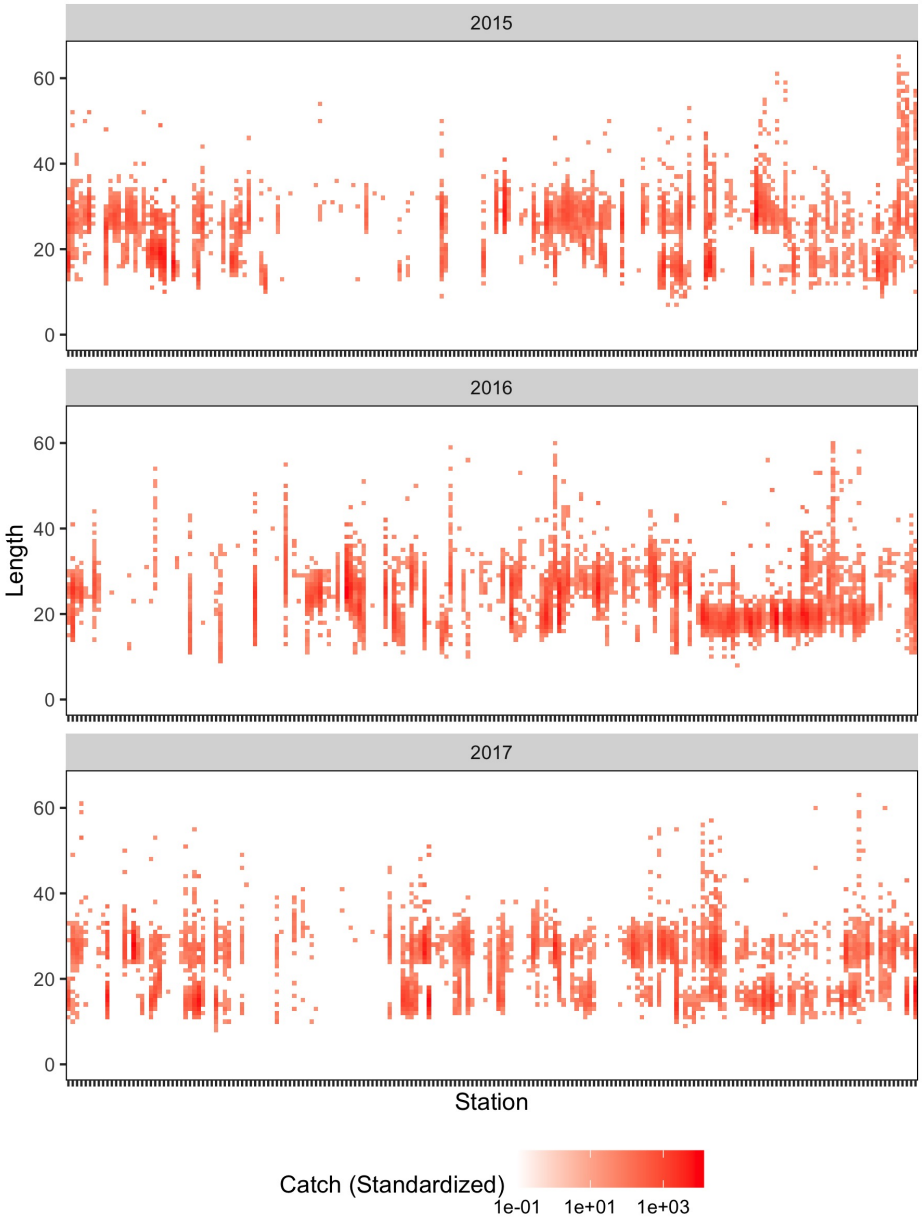

Figure 6. Catch numbers (effort-standardized) of species 14 from the past Maritimes surveys. Color intensity indicates catch at each length (cm) and station (for simplicity, station ID is not labelled). Stations are numbered sequentially according to the original order in which they were sampled.

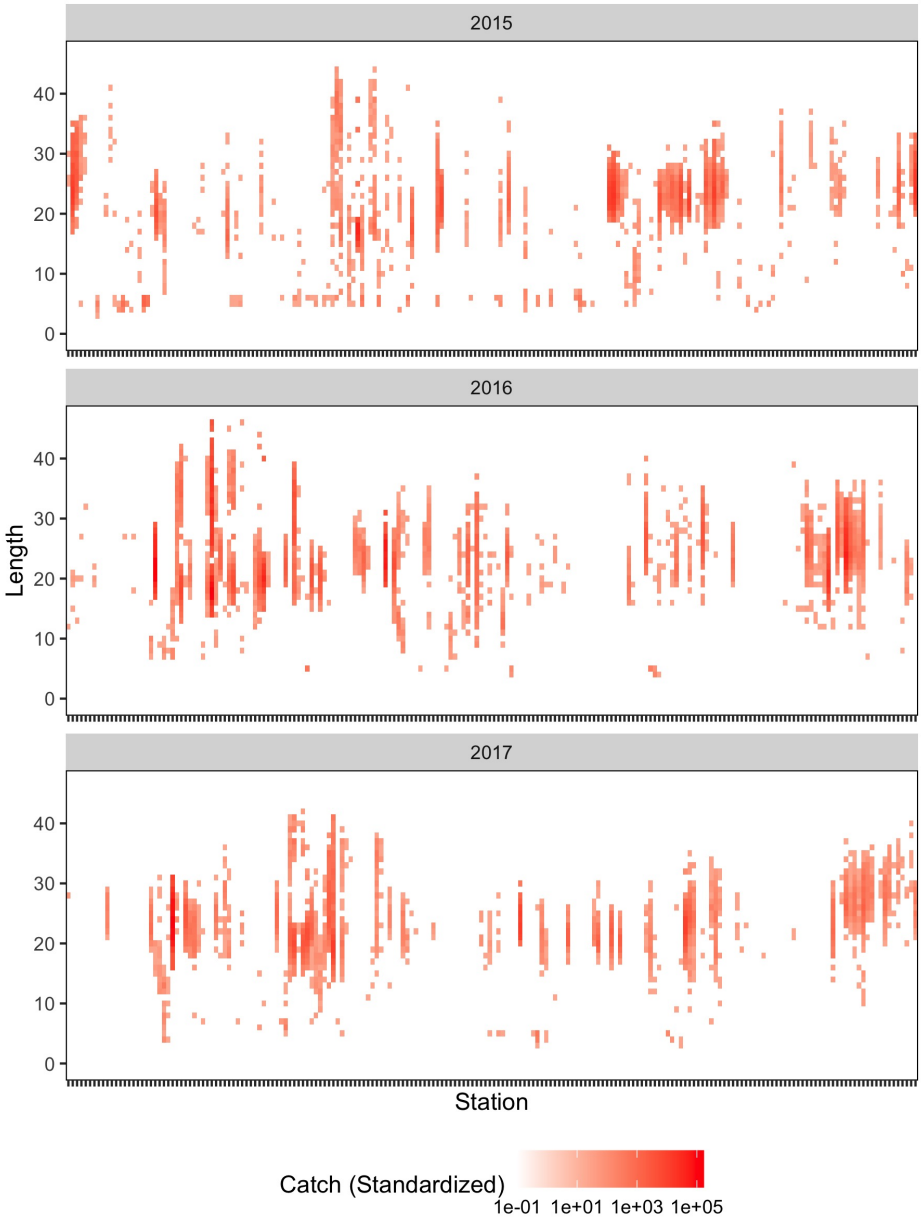

Figure 7. Catch numbers (effort-standardized) of species 23 from the past Maritimes surveys. Color intensity indicates catch at each length (cm) and station (for simplicity, station ID is not labelled). Stations are numbered sequentially according to the original order in which they were sampled.

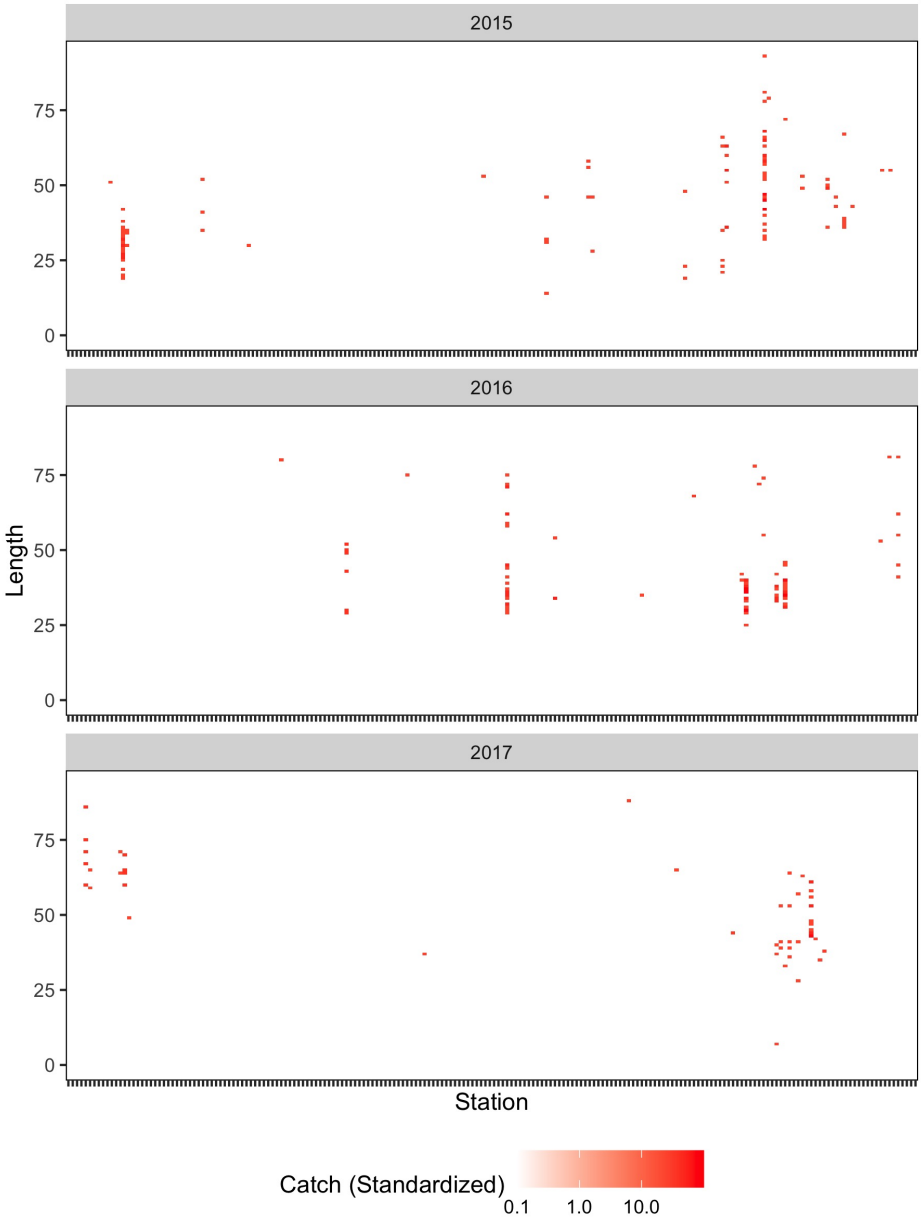

Figure 8. Catch numbers (effort-standardized) of species 204 from the past Maritimes surveys. Color intensity indicates catch at each length (cm) and station (for simplicity, station ID is not labelled). Stations are numbered sequentially according to the original order in which they were sampled.

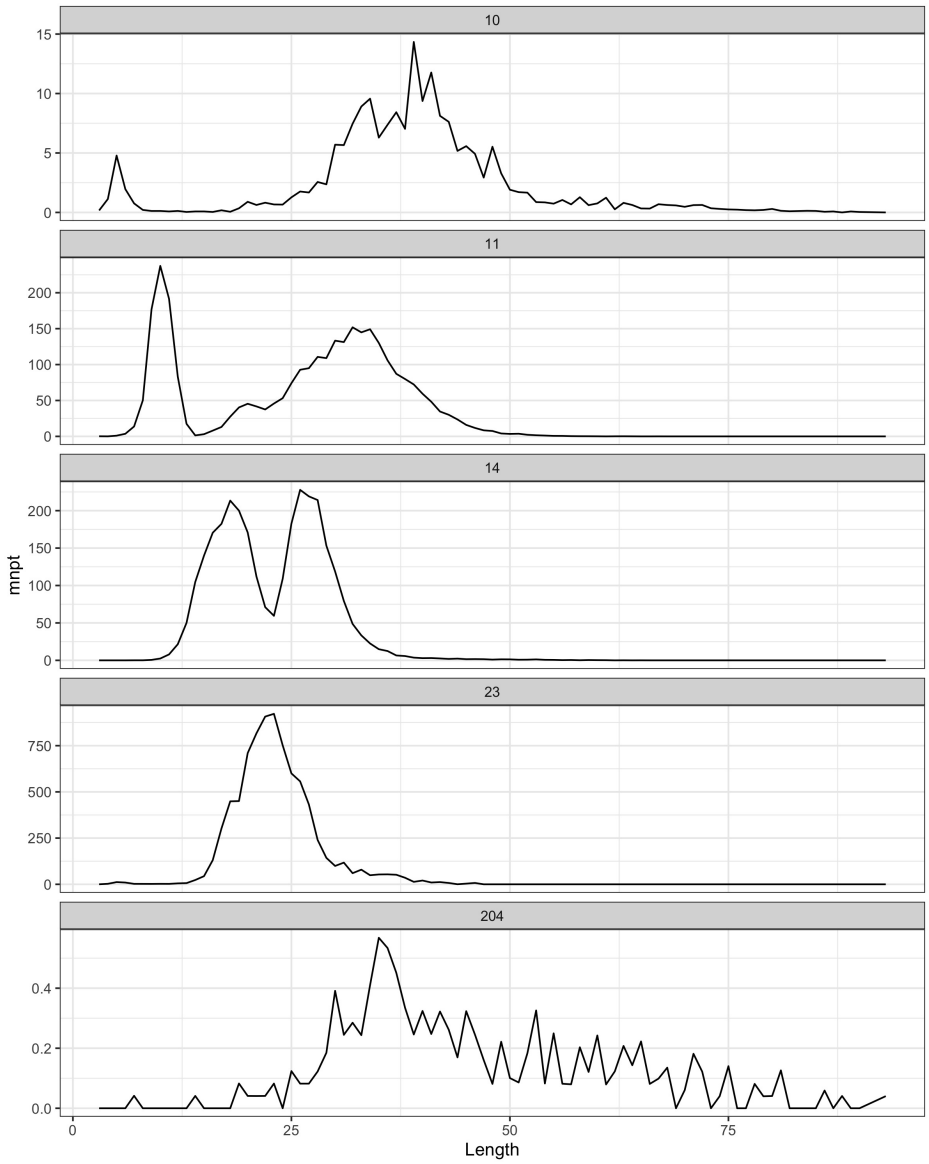

Figure 9. Mean catch number per tow (mnpt) for each species from past Maritimes survey (2015-2017).

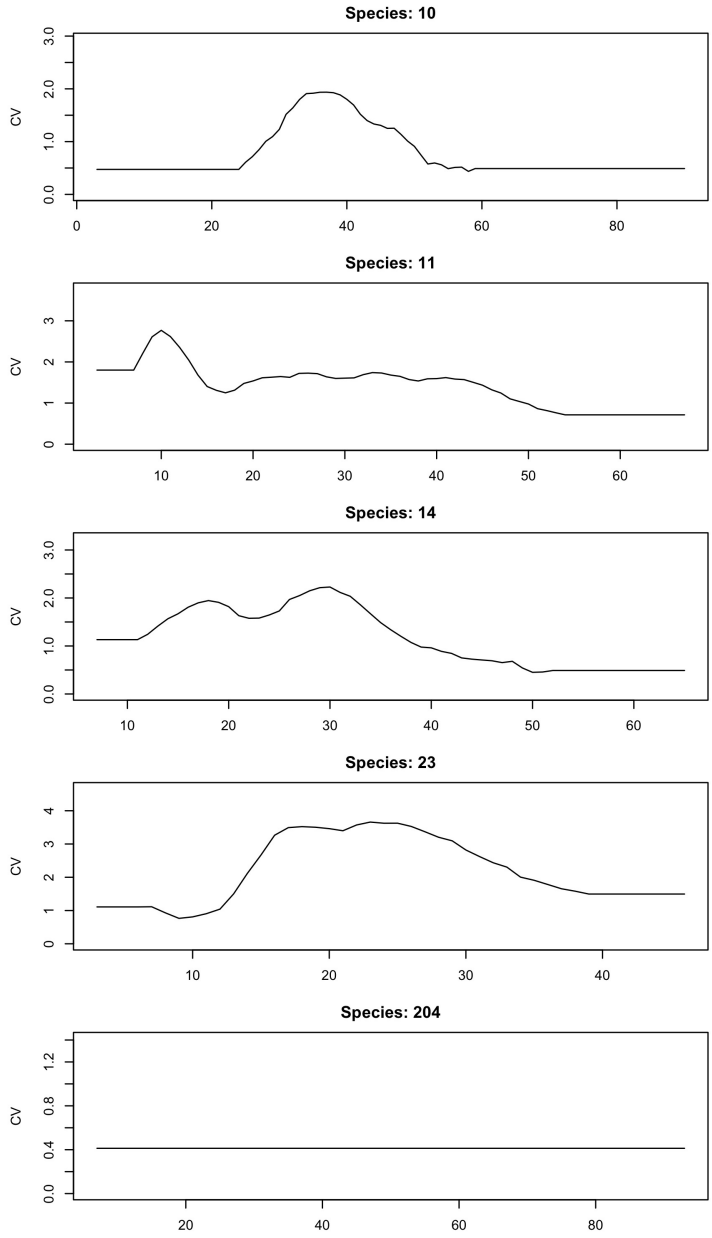

Figure 10. The coefficient of variation over length (CV profile) for each species from past Maritimes survey (2015-2017). See Section 3.1.3 for details of estimation of the CV profile.

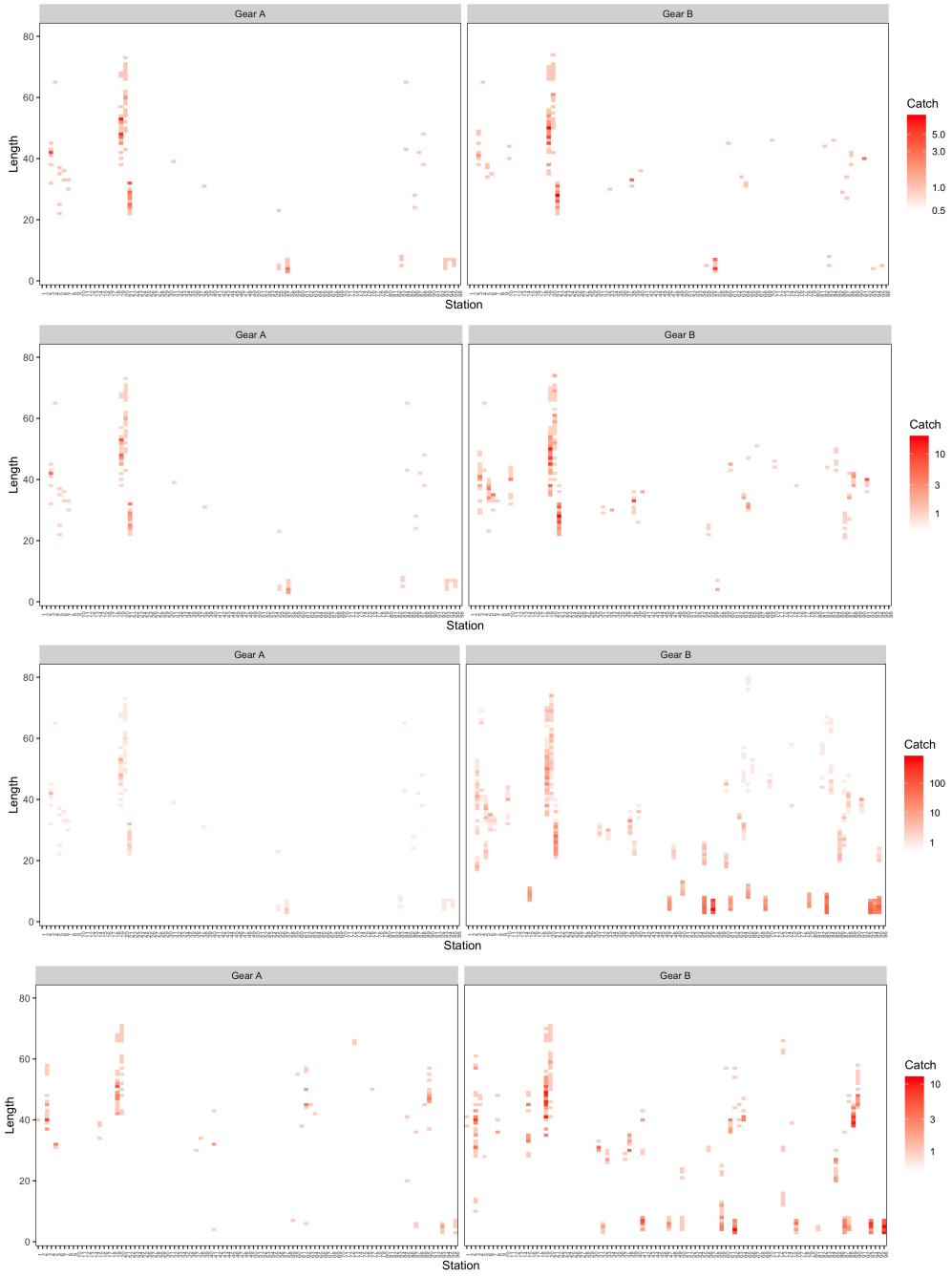

Figure 11. An example of simulated datasets for simulation scenarios (rows) 10-1, 10-2, 10-3, and 10-4, i.e., catch numbers for the pair of gears A and B (right and left columns, respectively) at each station and length (cm). The examples are generated using random seed 1 in R for each scenario.

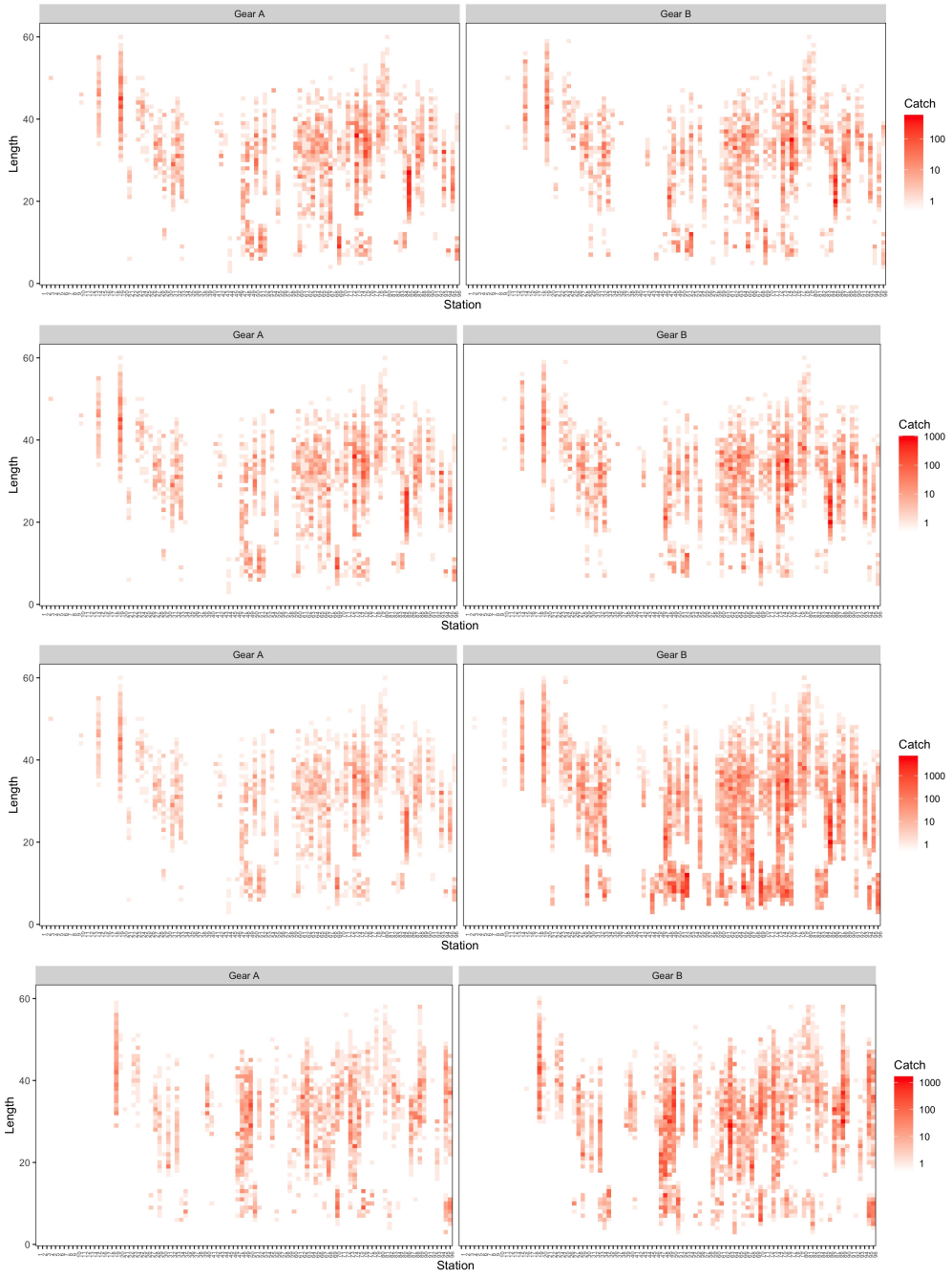

Figure 12. An example of simulated datasets for simulation scenarios (rows) 11-1, 11-2, 11-3, and 11-4, i.e., catch numbers for the pair of gears A and B (right and left columns, respectively) at each station and length (cm). The examples are generated using random seed 1 in R for each scenario.

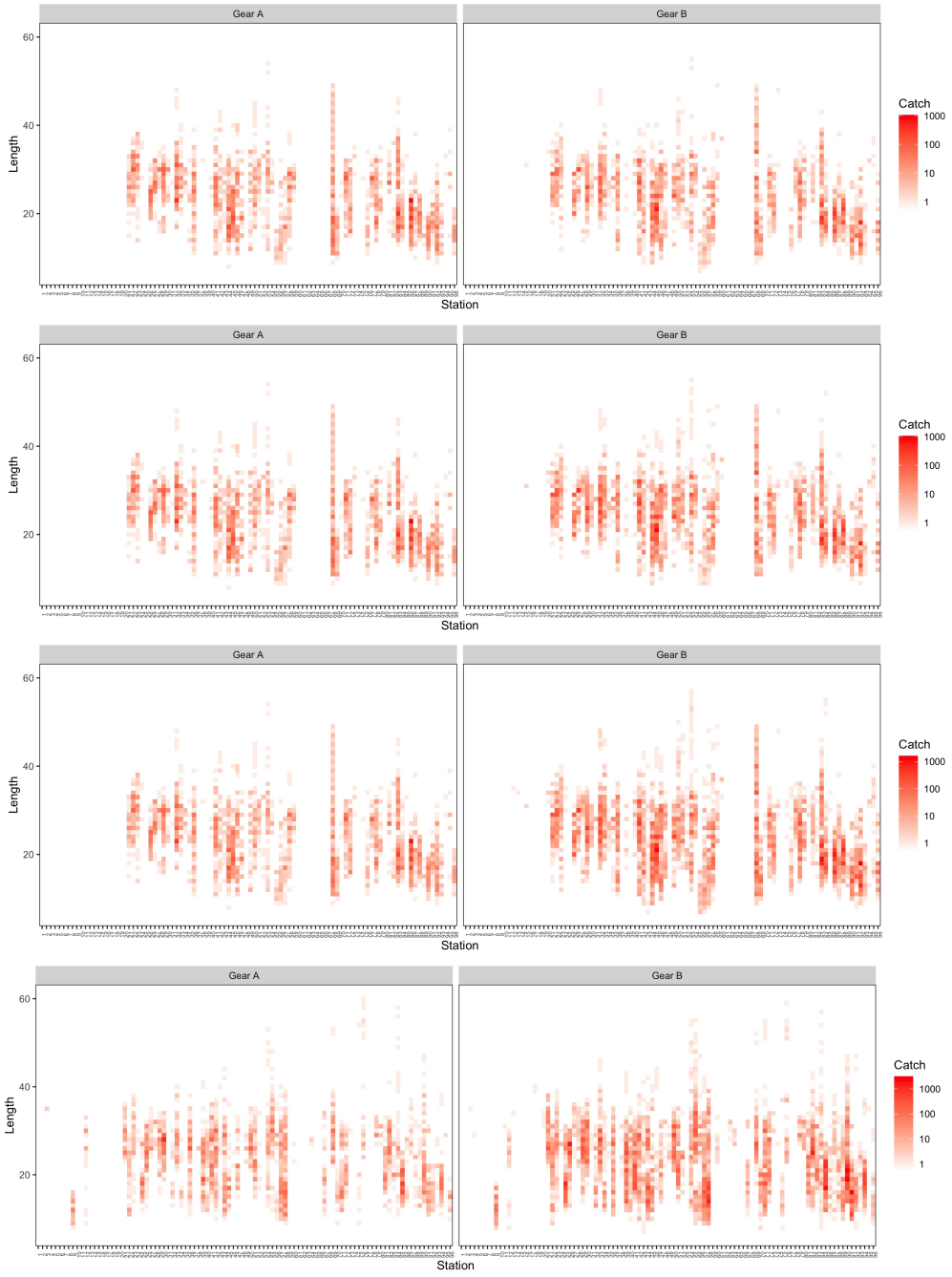

Figure 13. An example of simulated datasets for simulation scenarios (rows) 14-1, 14-2, 14-3, and 14-4, i.e., catch numbers for the pair of gears A and B (right and left columns, respectively) at each station and length (cm). The examples are generated using random seed 1 in R for each scenario.

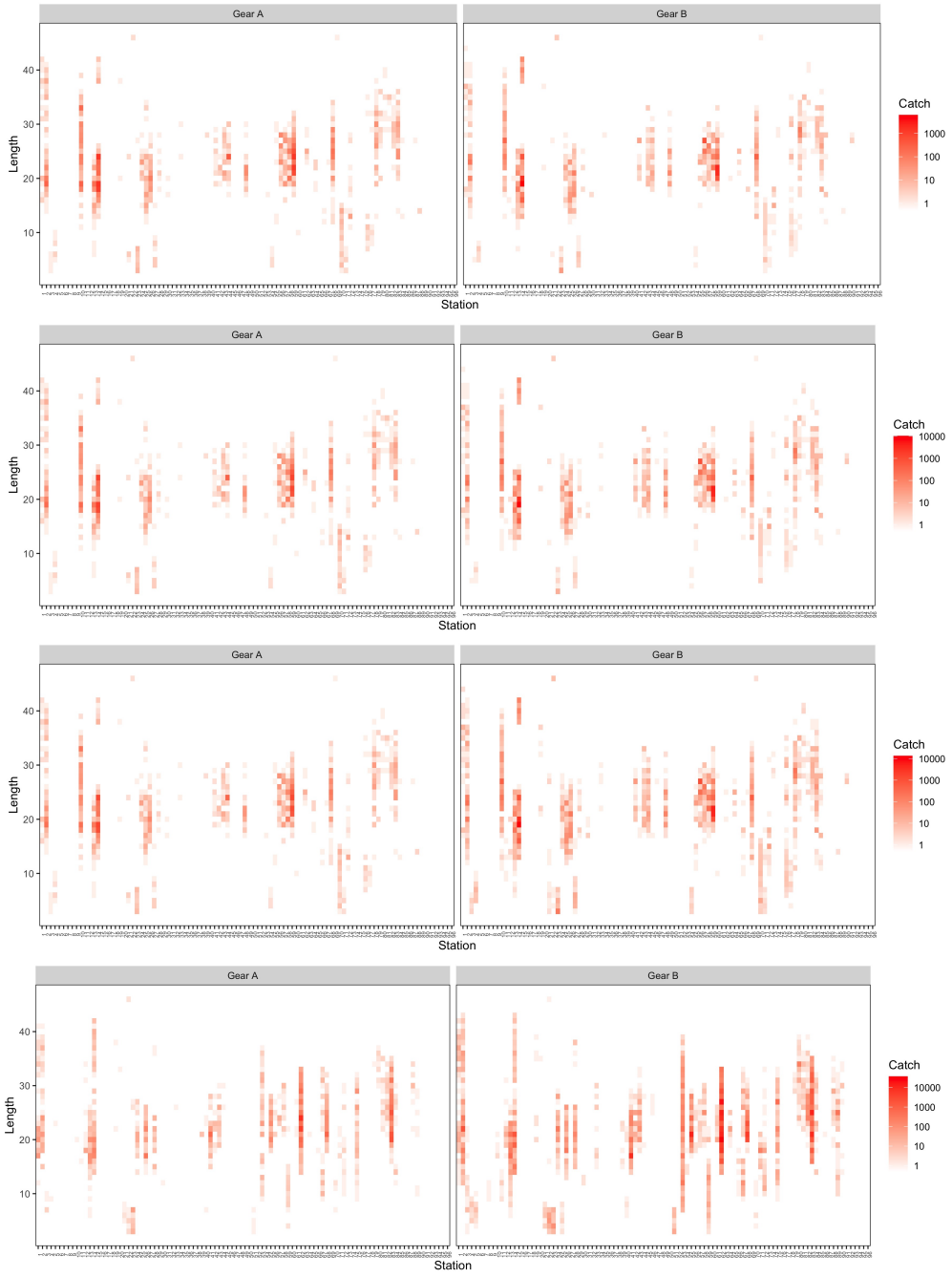

Figure 14. An example of simulated datasets for simulation scenarios (rows) 23-1, 23-2, 23-3, and 23-4, i.e., catch numbers for the pair of gears A and B (right and left columns, respectively) at each station and length (cm). The examples are generated using random seed 1 in R for each scenario.

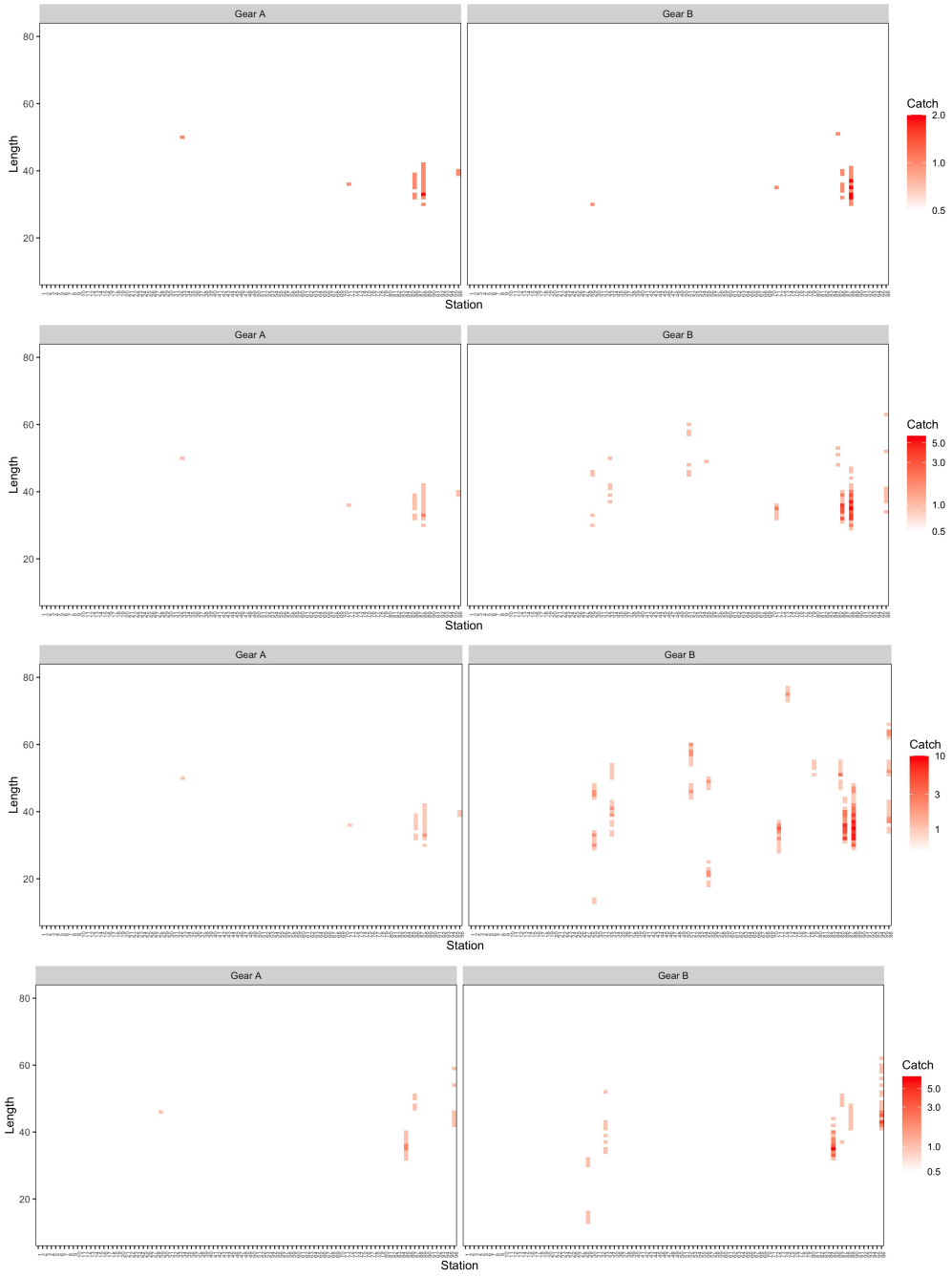

Figure 15. An example of simulated datasets for simulation scenarios (rows) 204-1, 204-2, 204-3, and 204-4, i.e., catch numbers for the pair of gears A and B (right and left columns, respectively) at each station and length (cm). The examples are generated using random seed 1 in R for each scenario.

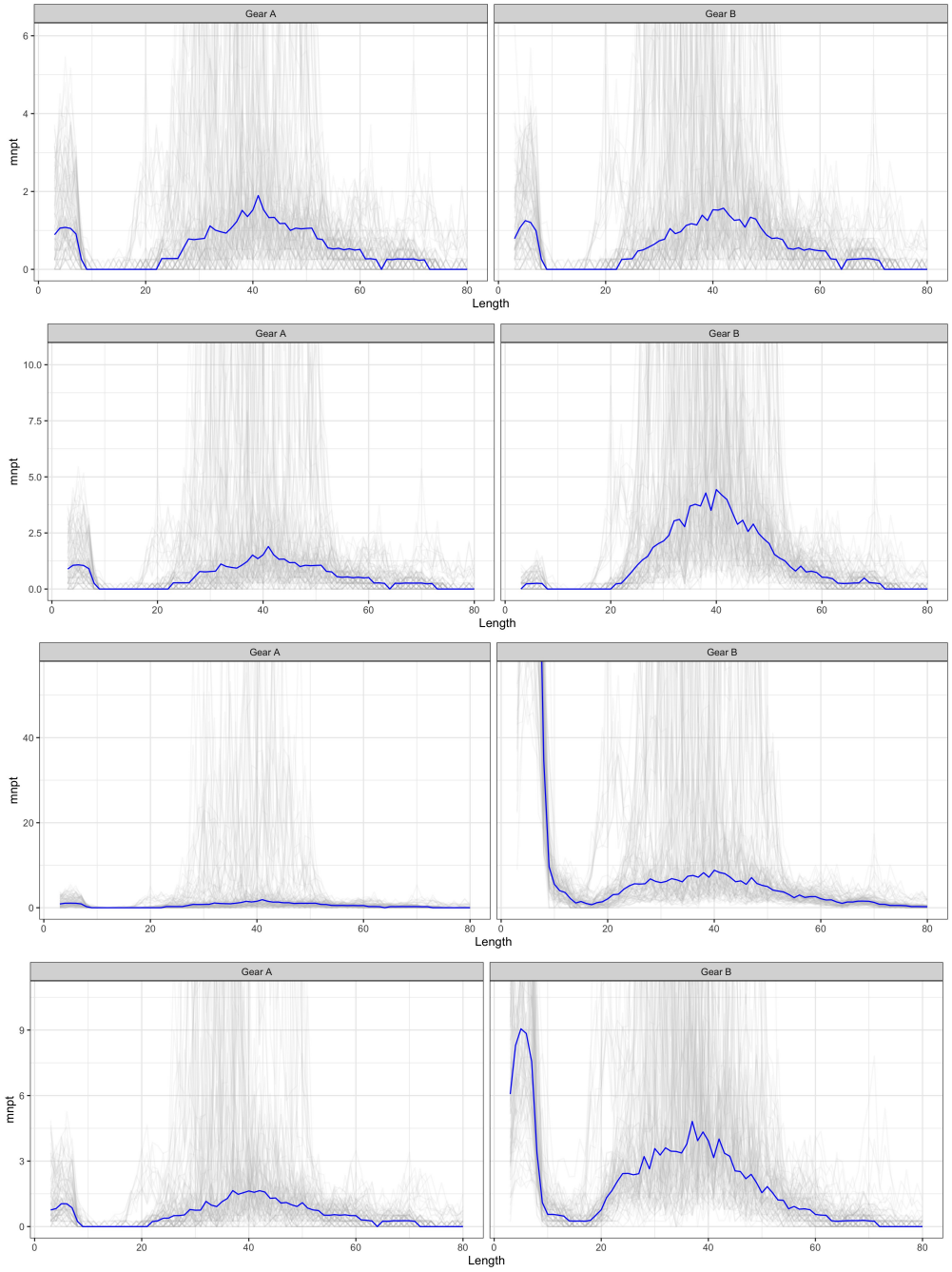

Figure 16. Mean catch number per tow (mnpt) calculated for each simulated dataset (gray line) and the median across 100 simulation datasets (blue line), for simulation scenarios 10-1, 10-2, 10-3, and 10-4. The lognormal distribution may generate extreme errors in some simulation iterations; the y-axis is limited by the 95% quantiles of the plotted data in each scenario.

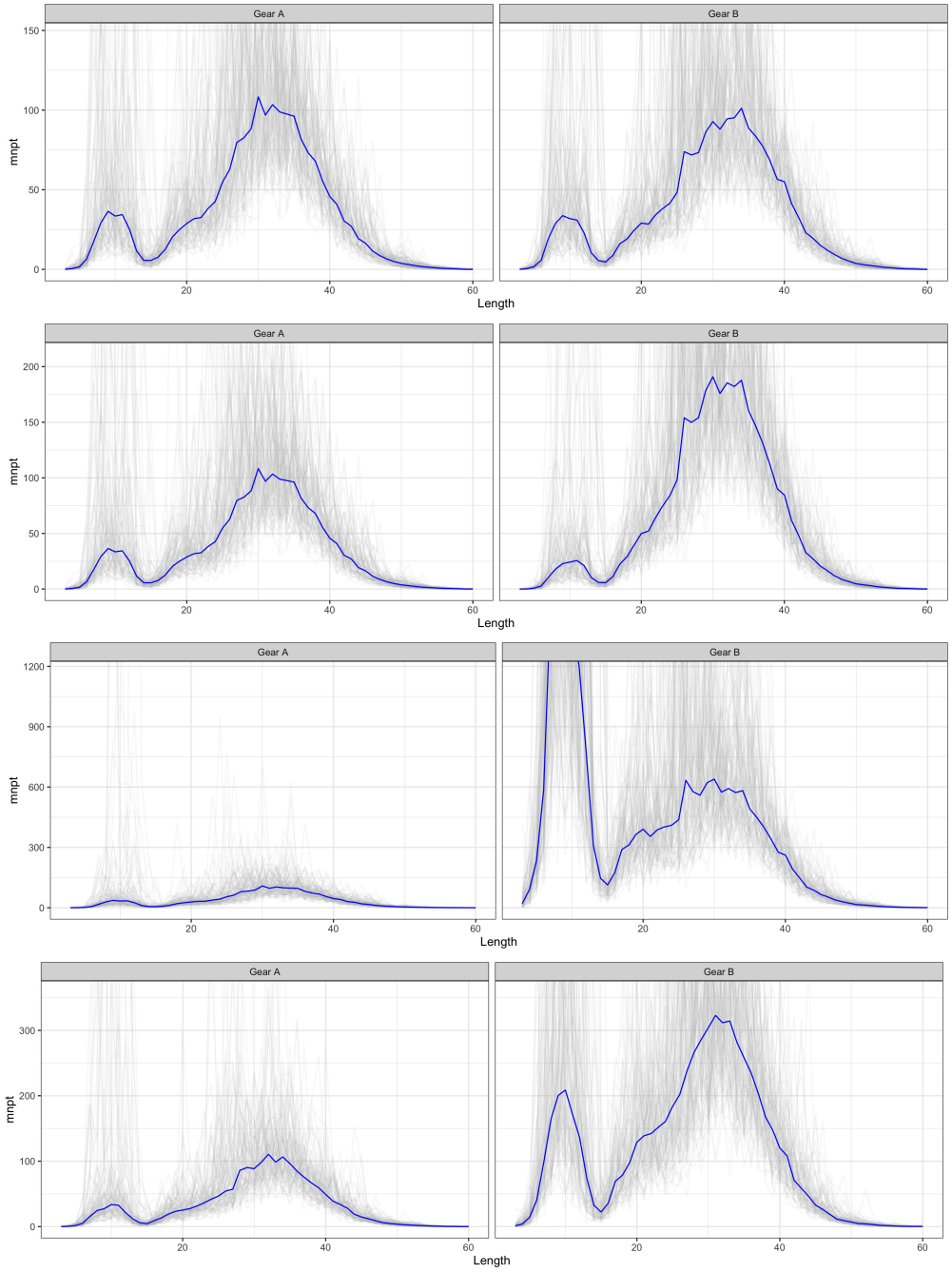

Figure 17. Mean catch number per tow (mnpt) calculated for each simulated dataset (gray line) and the median across 100 simulation datasets (blue line), for simulation scenarios 11-1, 11-2, 11-3, and 11-4. The lognormal distribution may generate extreme errors in some simulation iterations; the y-axis is limited by the 95% quantiles of the plotted data in each scenario.

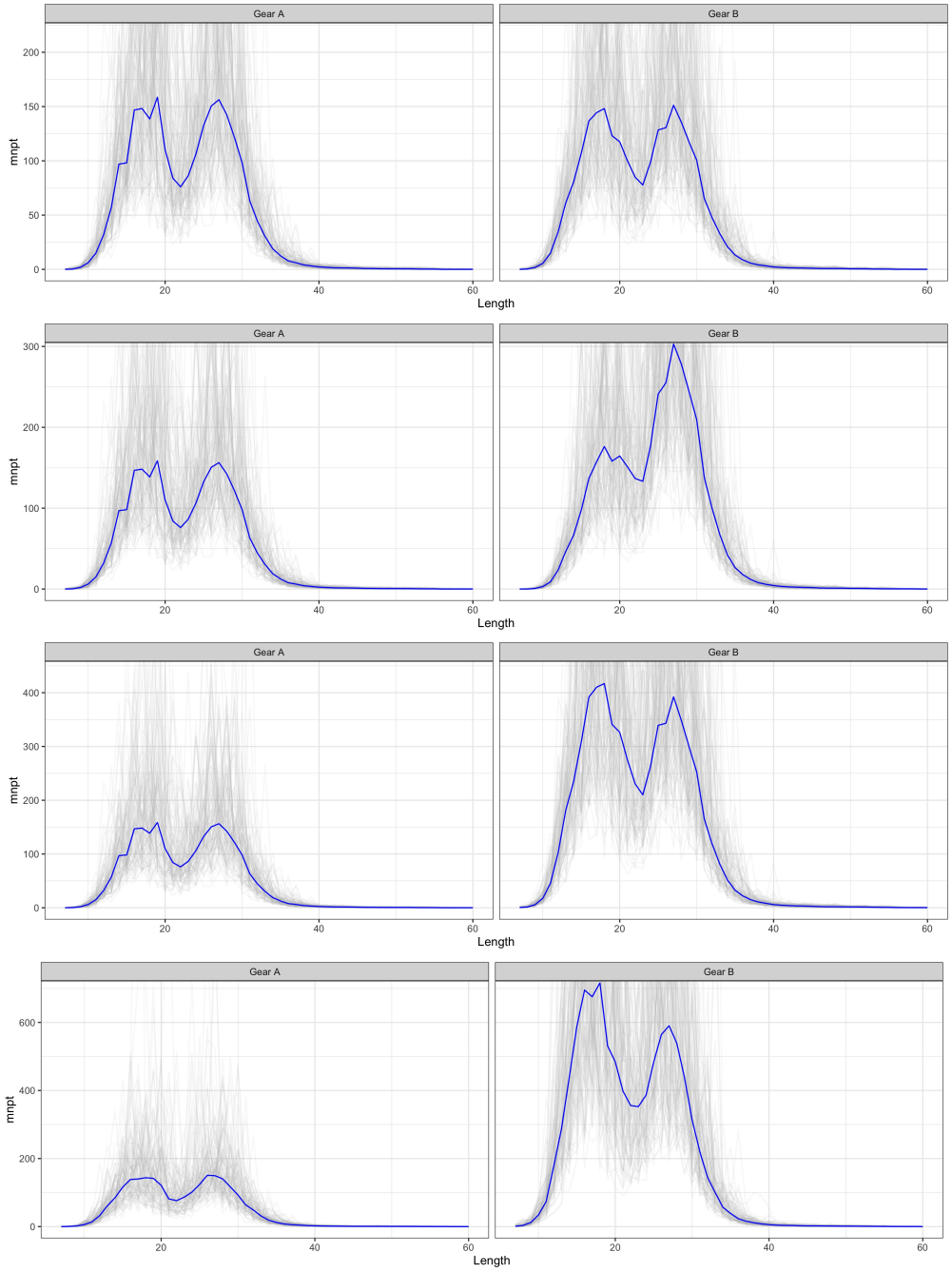

Figure 18. Mean catch number per tow (mnpt) calculated for each simulated dataset (gray line) and the median across 100 simulation datasets (blue line), for simulation scenarios 14-1, 14-2, 14-3, and 14-4. The lognormal distribution may generate extreme errors in some simulation iterations; the y-axis is limited by the 95% quantiles of the plotted data in each scenario.

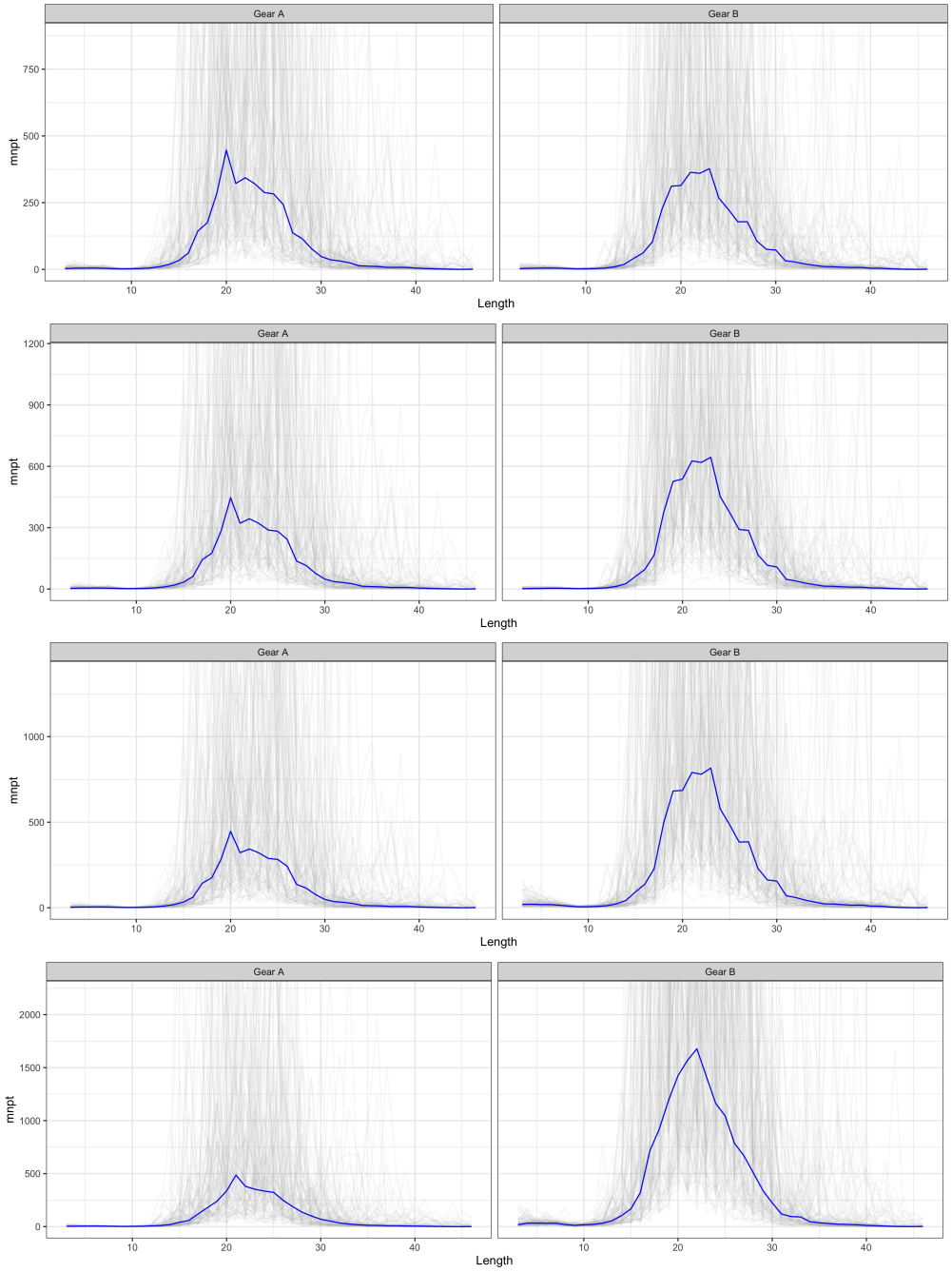

Figure 19. Mean catch number per tow (mnpt) calculated for each simulated dataset (gray line) and the median across 100 simulation datasets (blue line), for simulation scenarios 23-1, 23-2, 23-3, and 23-4. The lognormal distribution may generate extreme errors in some simulation iterations; the y-axis is limited by the 95% quantiles of the plotted data in each scenario.

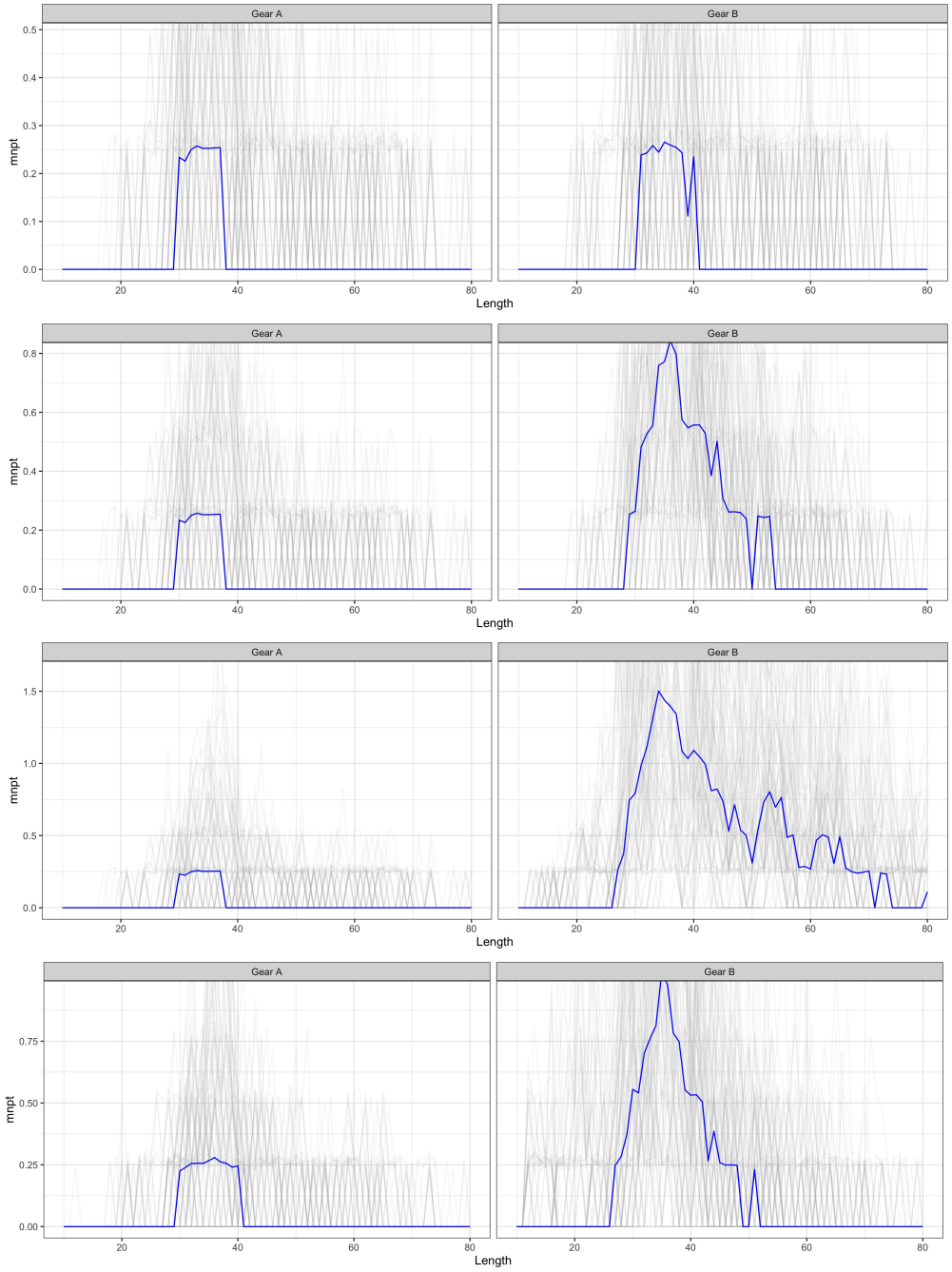

Figure 20. Mean catch number per tow (mnpt) calculated for each simulated dataset (gray line) and the median across 100 simulation datasets (blue line), for simulation scenarios 204-1, 204-2, 204-3, and 204-4. The lognormal distribution may generate extreme errors in some simulation iterations; the y-axis is limited by the 95% quantiles of the plotted data in each scenario.

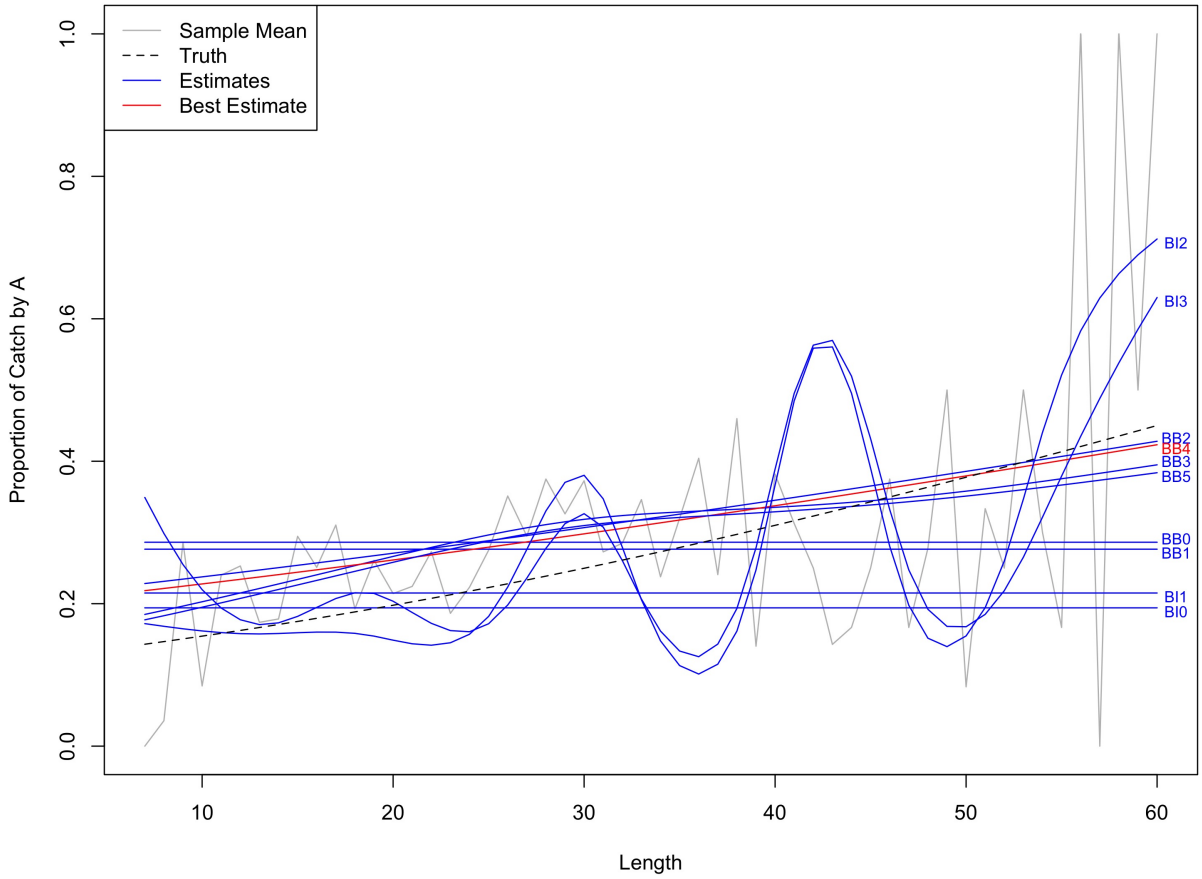

Figure 21. An example of estimated proportion of catch by gear A from each candidate model to illustrate the difference in bias-variance trade-off between the Binomial and Beta-Binomial models. The Binomial models tend to produce "wavy" estimations and the Beta-Binomial models tend to result in smoother shapes in the presence of strong sample variance. This example is from the simulation iteration based on the random seed 1 and scenario 14-4 in Sections 3 and 4). The segmented line is the simulated true proportion of catch by gear A; the gray line is the sample average proportion; the blues lines are estimated quantities from the candidate models and the red line is the estimate from the AIC-selected best model.

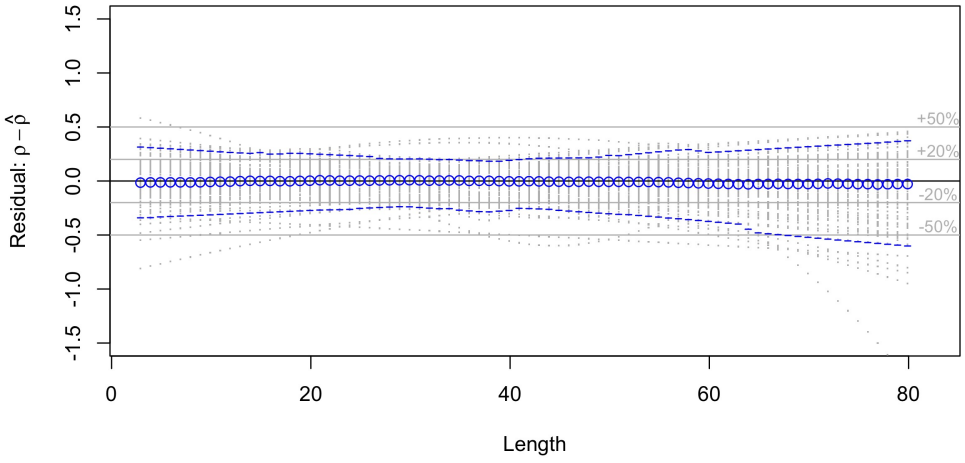

Figure 22. Estimation residuals for the relative catch efficiency over length in simulation scenario 10-1. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

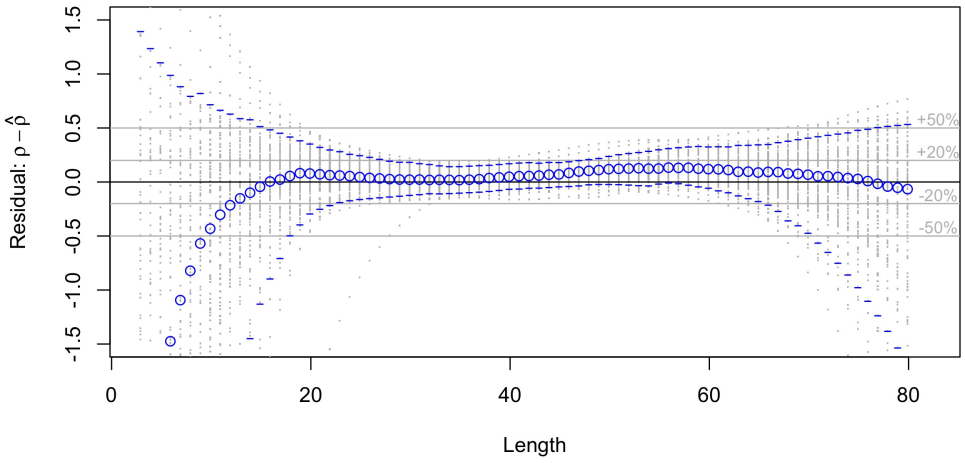

Figure 23. Estimation residuals for the relative catch efficiency over length in simulation scenario 10-2. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

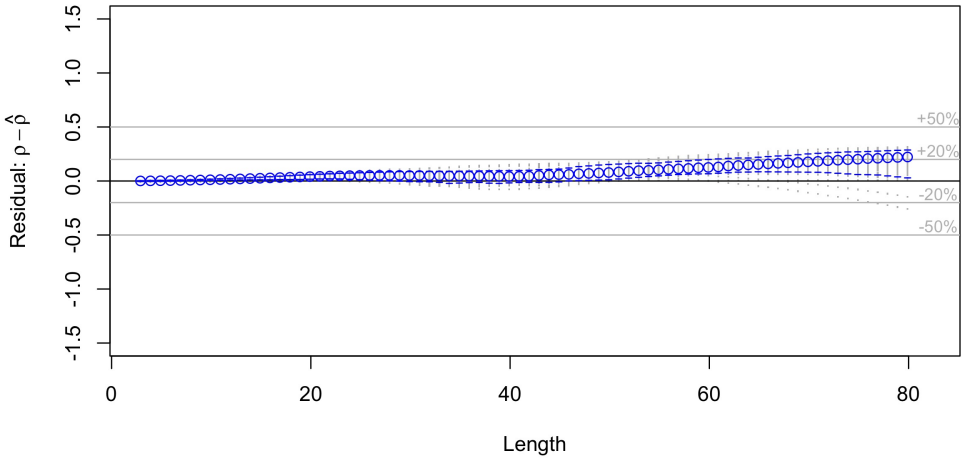

Figure 24. Estimation residuals for the relative catch efficiency over length in simulation scenario 10-3. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

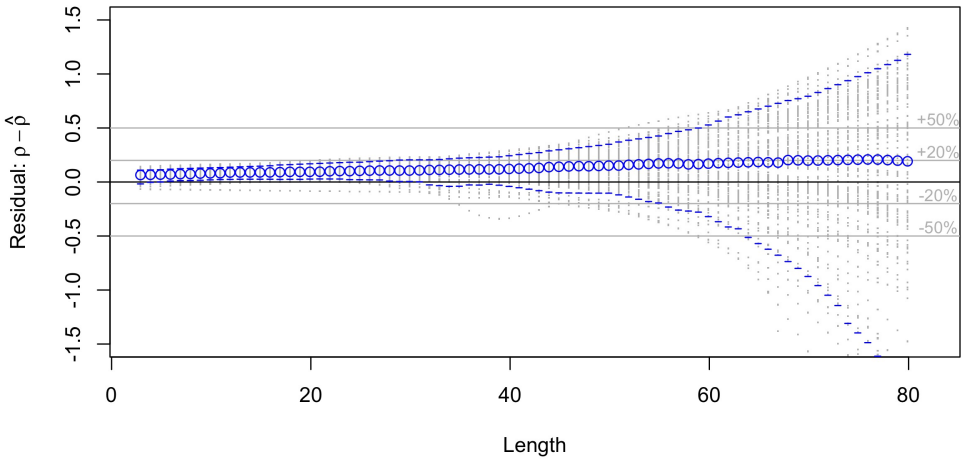

Figure 25. Estimation residuals for the relative catch efficiency over length in simulation scenario 10-4. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

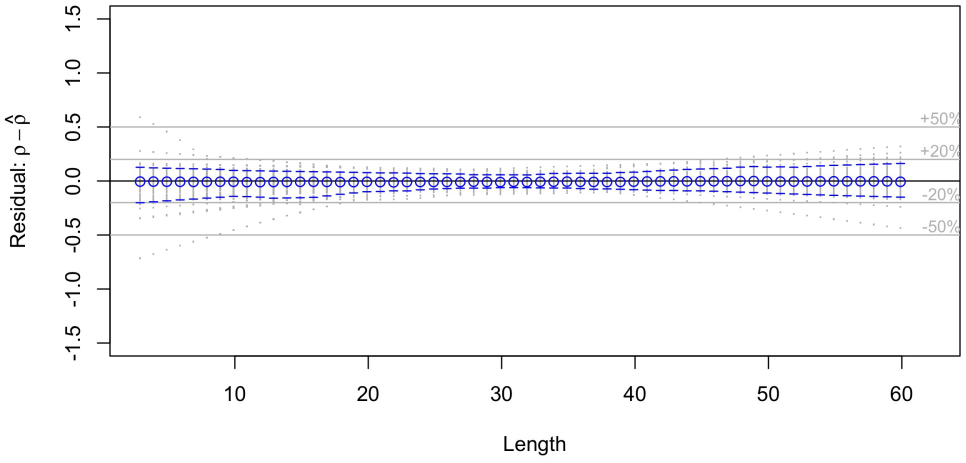

Figure 26. Estimation residuals for the relative catch efficiency over length in simulation scenario 11-1. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

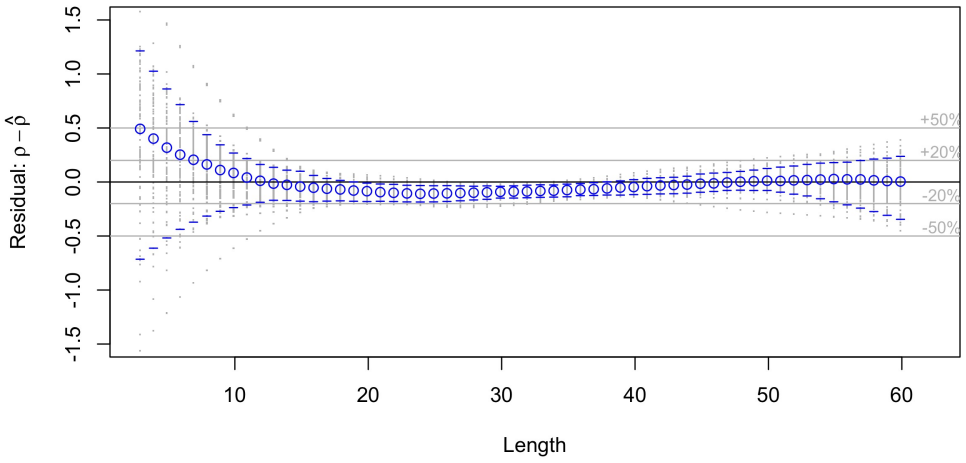

Figure 27. Estimation residuals for the relative catch efficiency over length in simulation scenario 11-2. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

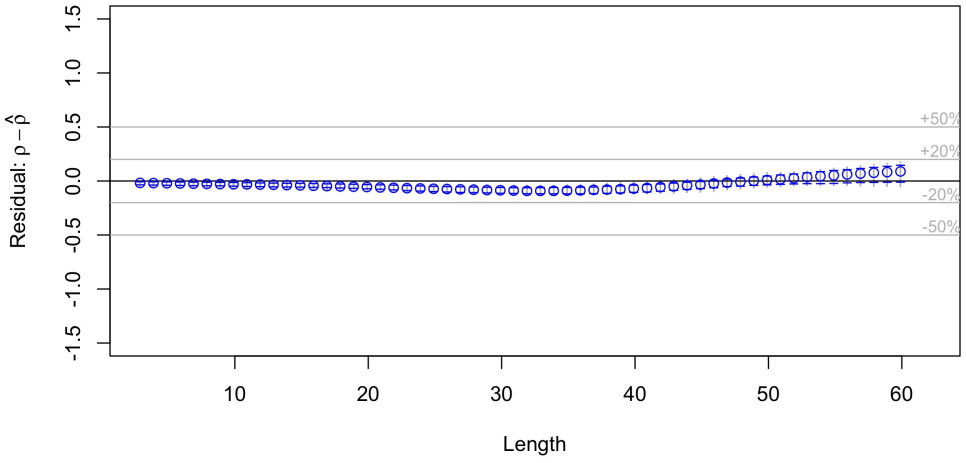

Figure 28. Estimation residuals for the relative catch efficiency over length in simulation scenario 11-3. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

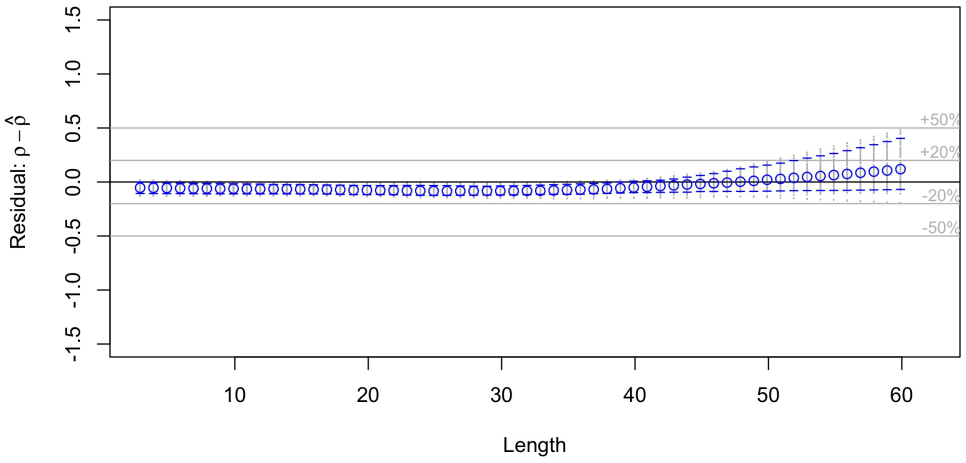

Figure 29. Estimation residuals for the relative catch efficiency over length in simulation scenario 11-4. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

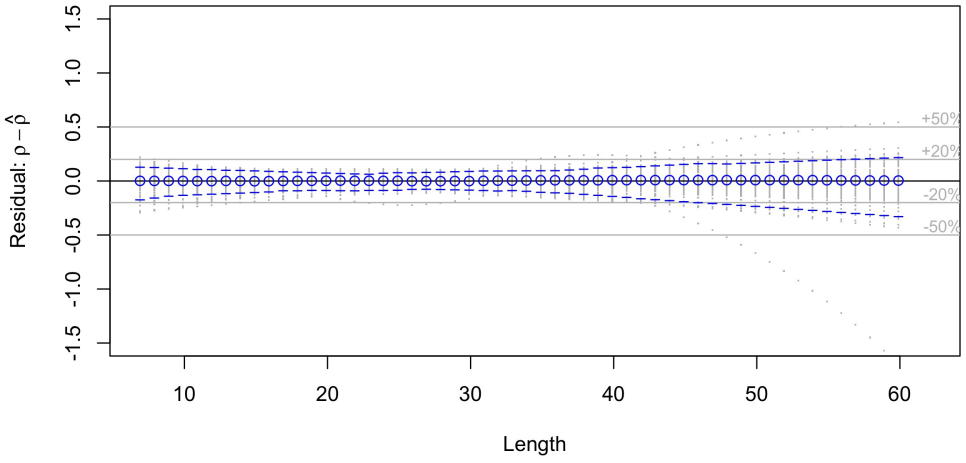

Figure 30. Estimation residuals for the relative catch efficiency over length in simulation scenario 14-1. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

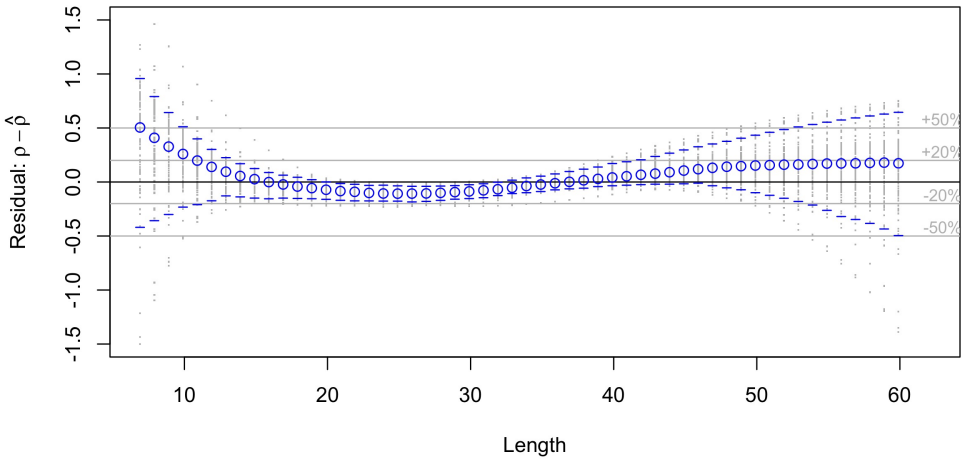

Figure 31. Estimation residuals for the relative catch efficiency over length in simulation scenario 14-2. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

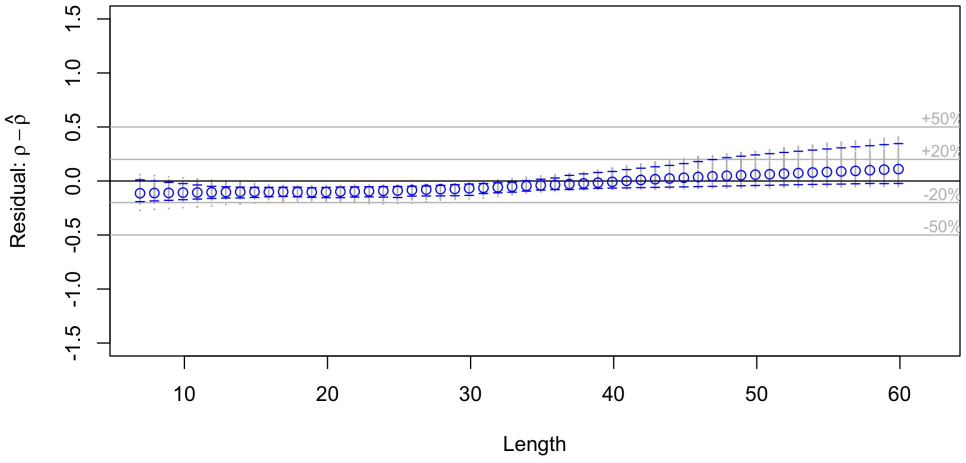

Figure 32. Estimation residuals for the relative catch efficiency over length in simulation scenario 14-3. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

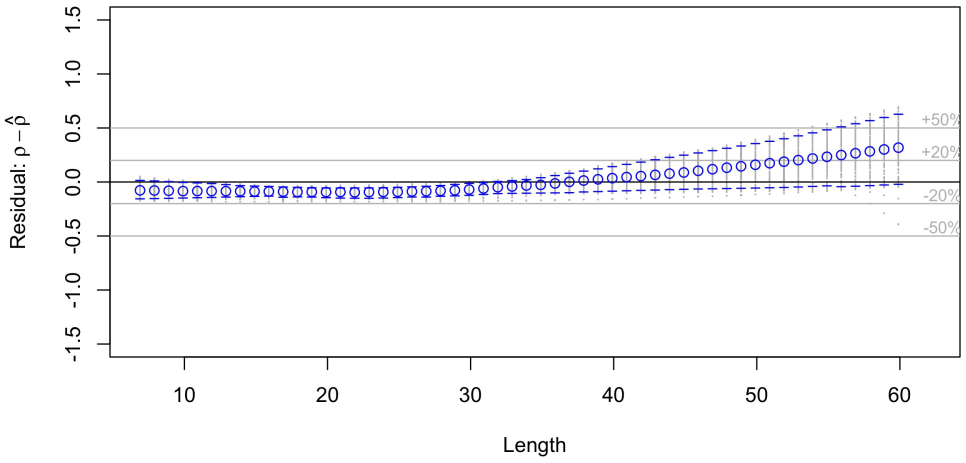

Figure 33. Estimation residuals for the relative catch efficiency over length in simulation scenario 14-4. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

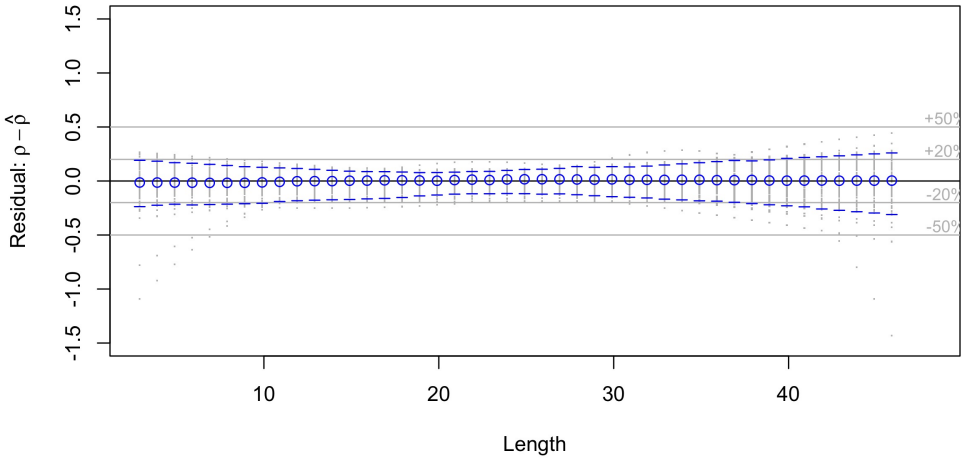

Figure 34. Estimation residuals for the relative catch efficiency over length in simulation scenario 23-1. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

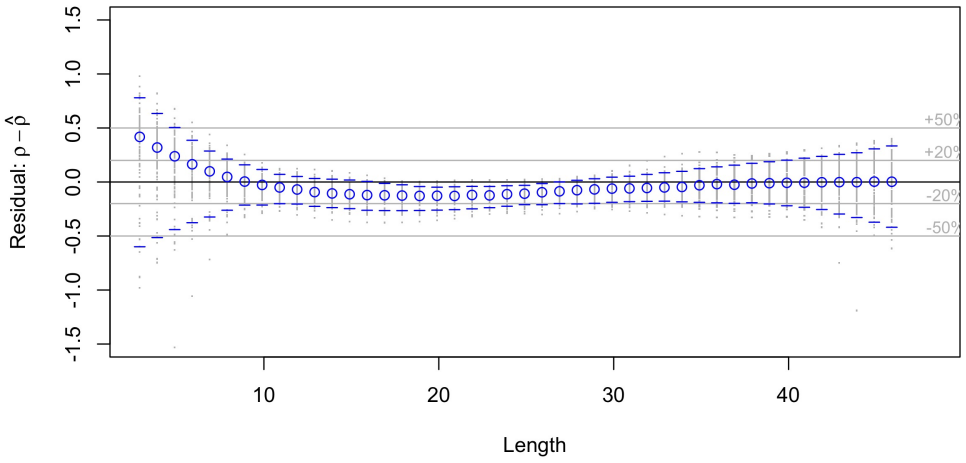

Figure 35. Estimation residuals for the relative catch efficiency over length in simulation scenario 23-2. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

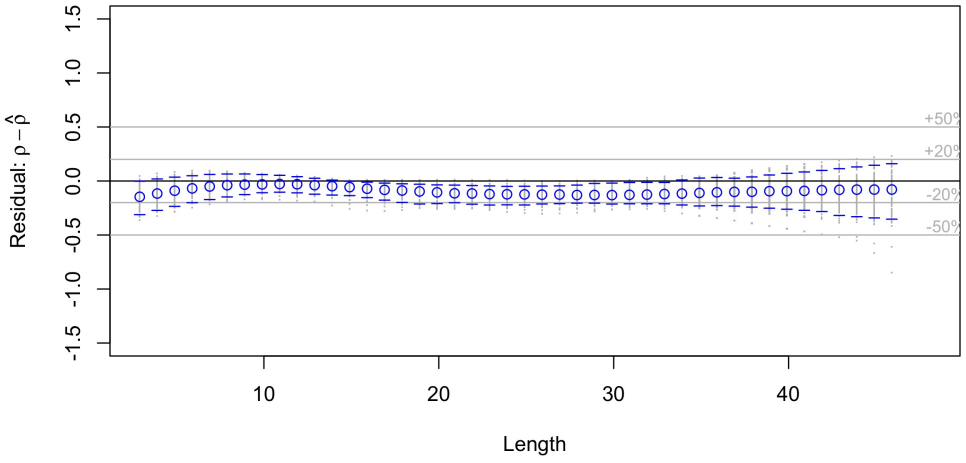

Figure 36. Estimation residuals for the relative catch efficiency over length in simulation scenario 23-3. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

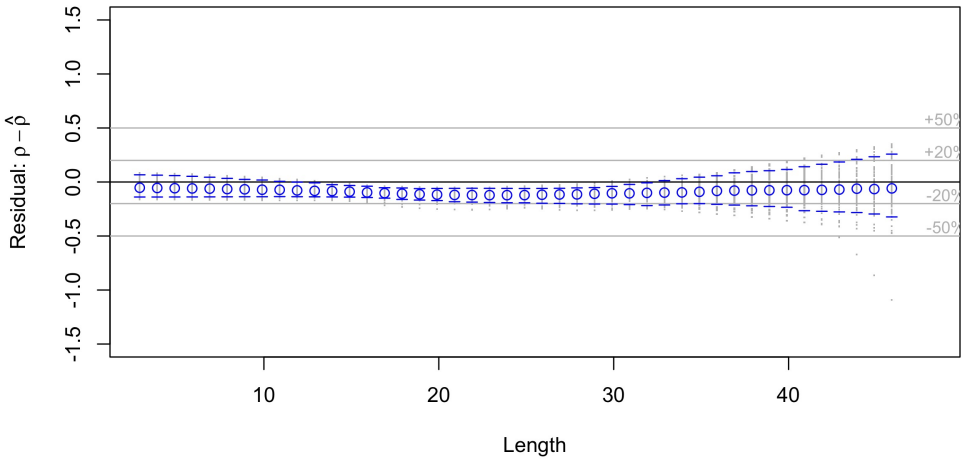

Figure 37. Estimation residuals for the relative catch efficiency over length in simulation scenario 23-4. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

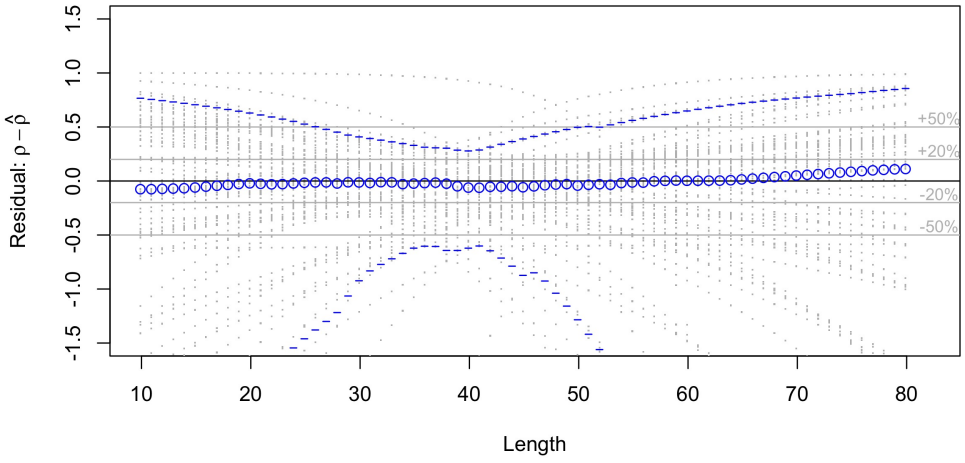

Figure 38. Estimation residuals for the relative catch efficiency over length in simulation scenario 204-1. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

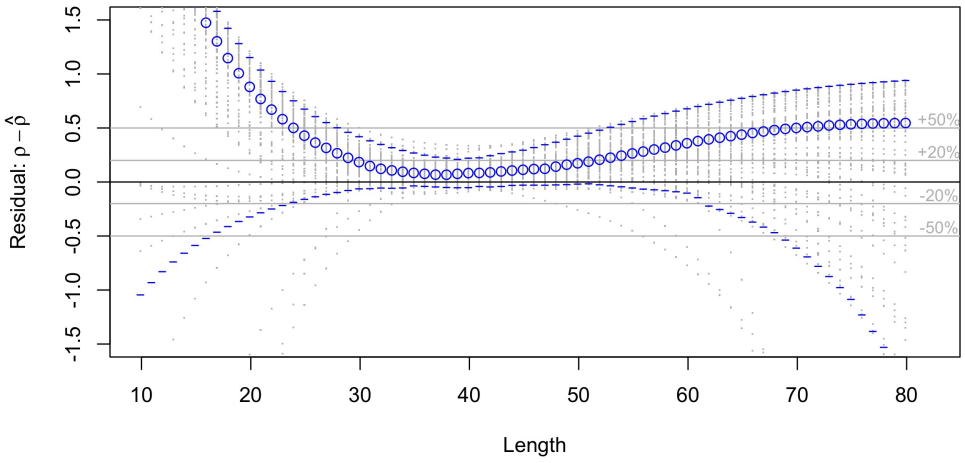

Figure 39. Estimation residuals for the relative catch efficiency over length in simulation scenario 204-2. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

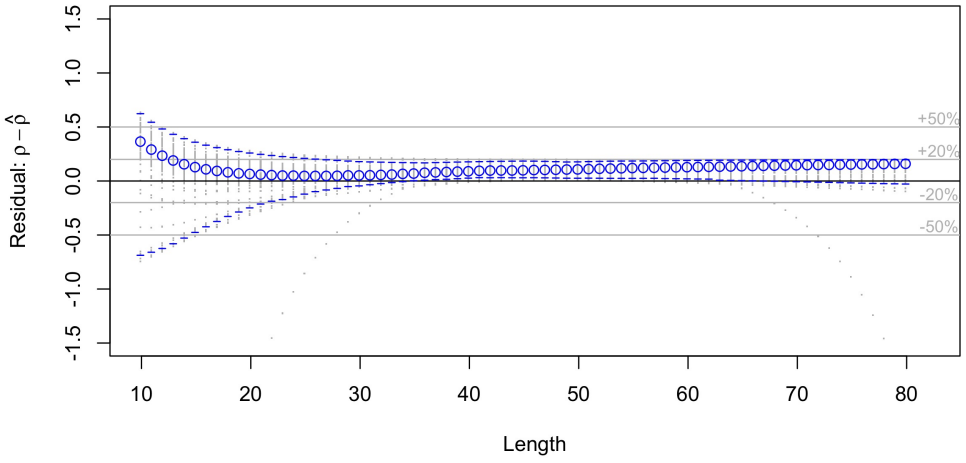

Figure 40. Estimation residuals for the relative catch efficiency over length in simulation scenario 204-3. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

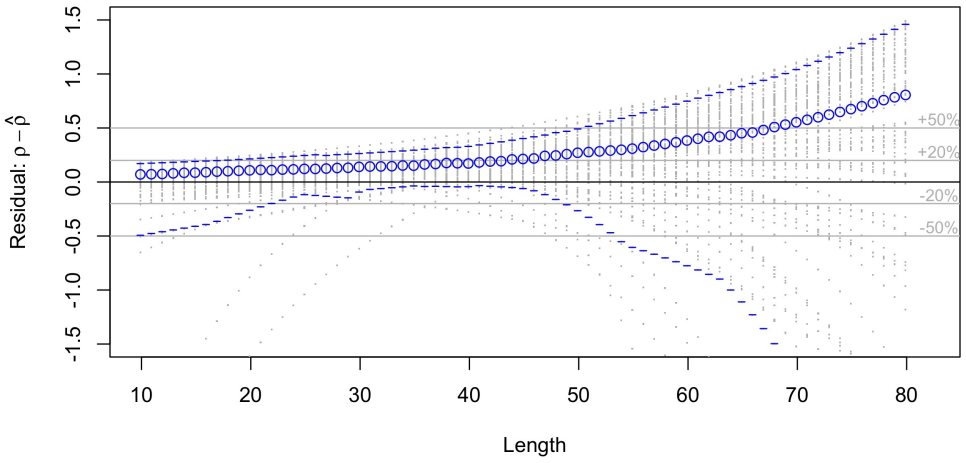

Figure 41. Estimation residuals for the relative catch efficiency over length in simulation scenario 204-4. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges.

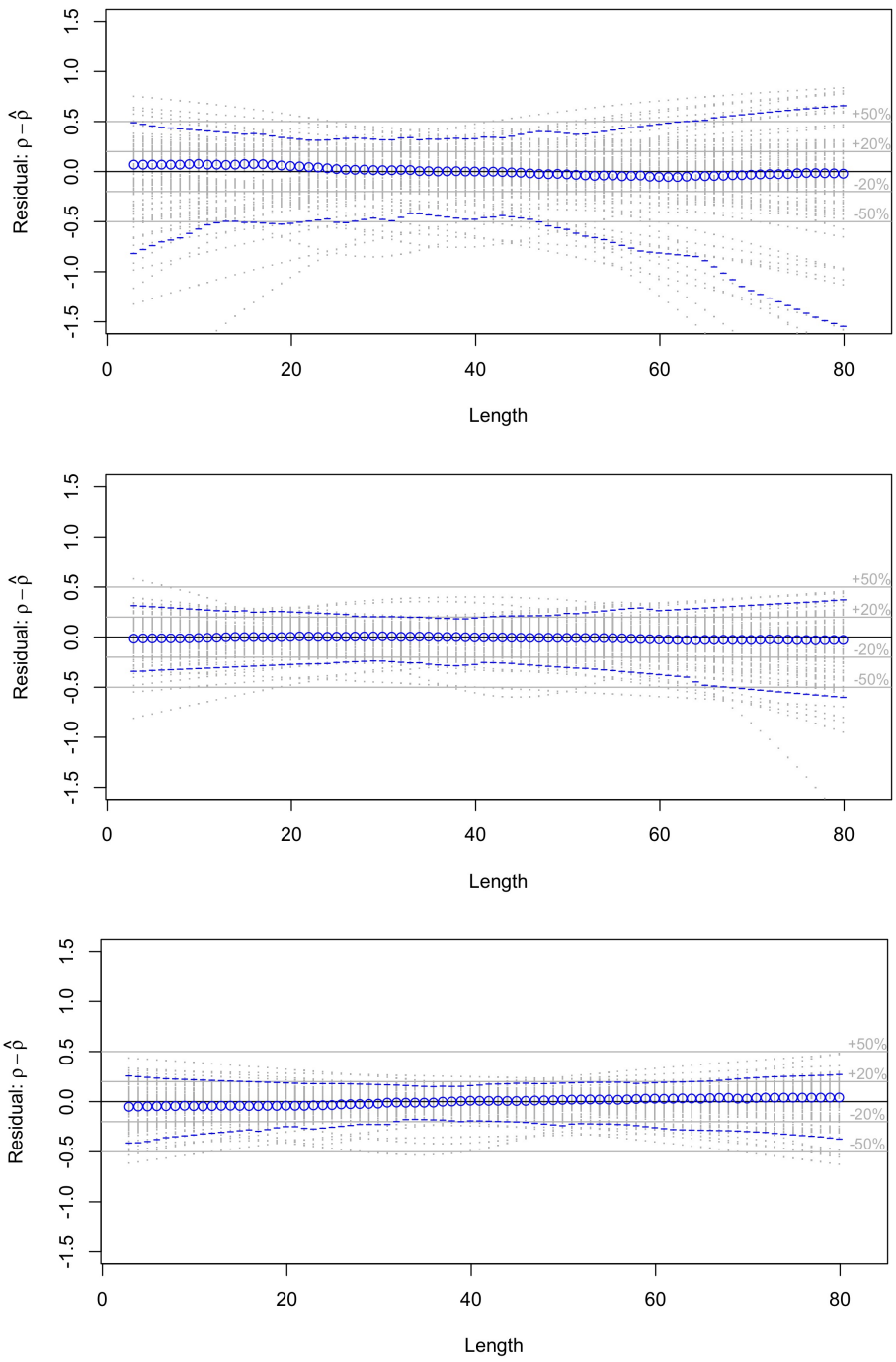

Figure 42. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 10-1-1 (top),10-1-2 (middle) and 10-1-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

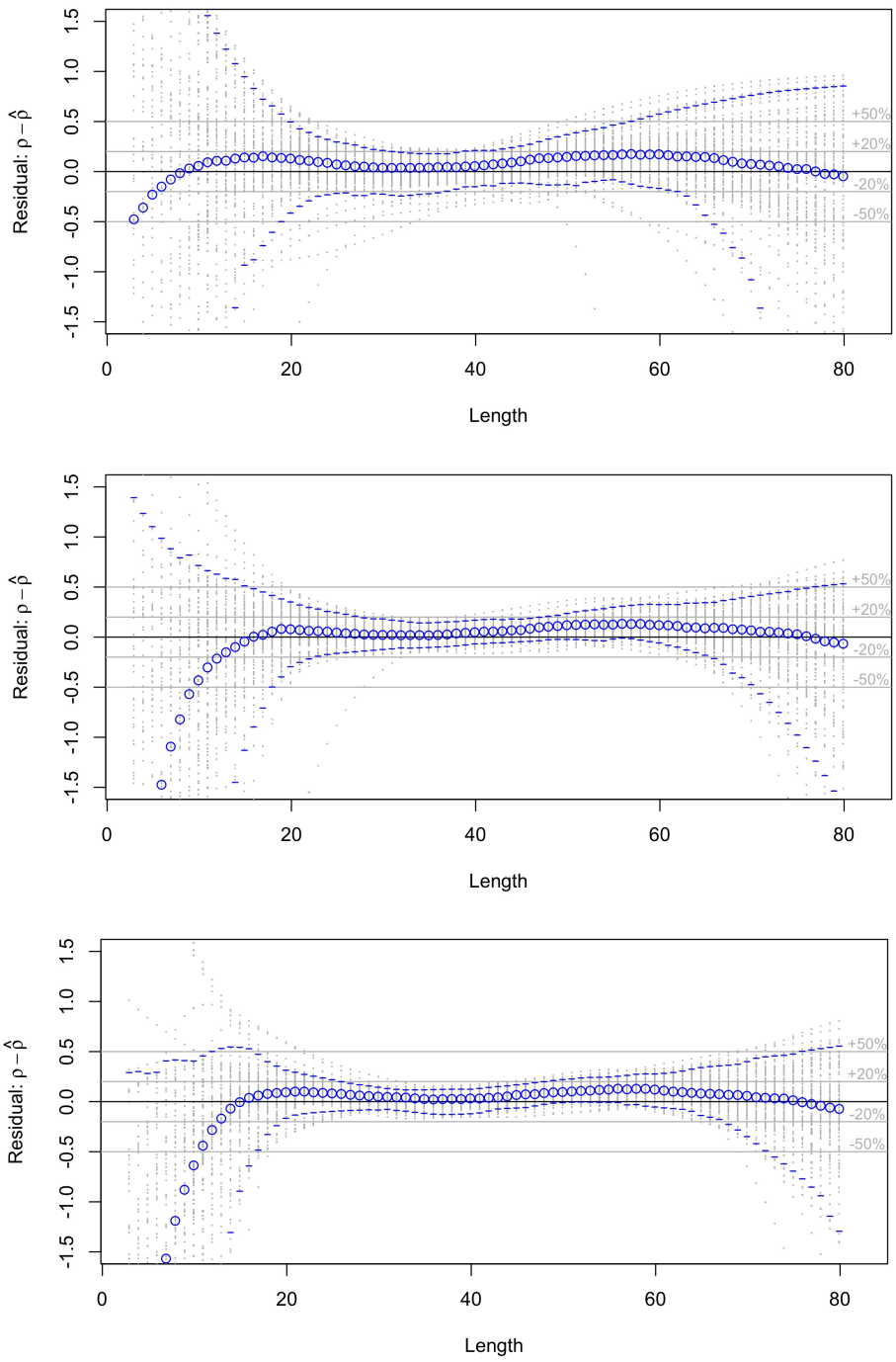

Figure 43. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 10-2-1 (top),10-2-2 (middle) and 10-2-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

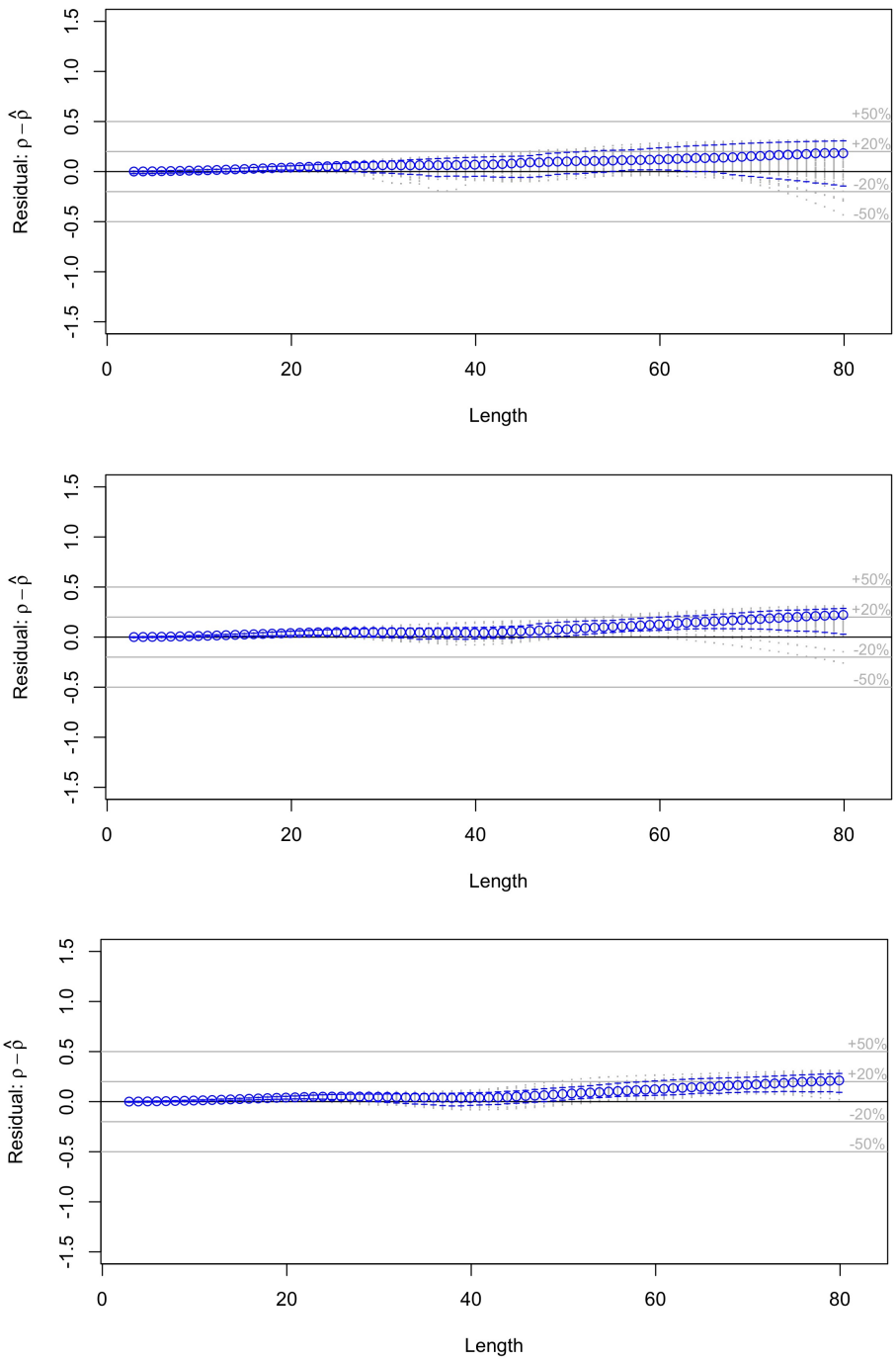

Figure 44. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 10-3-1 (top),10-3-2 (middle) and 10-3-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

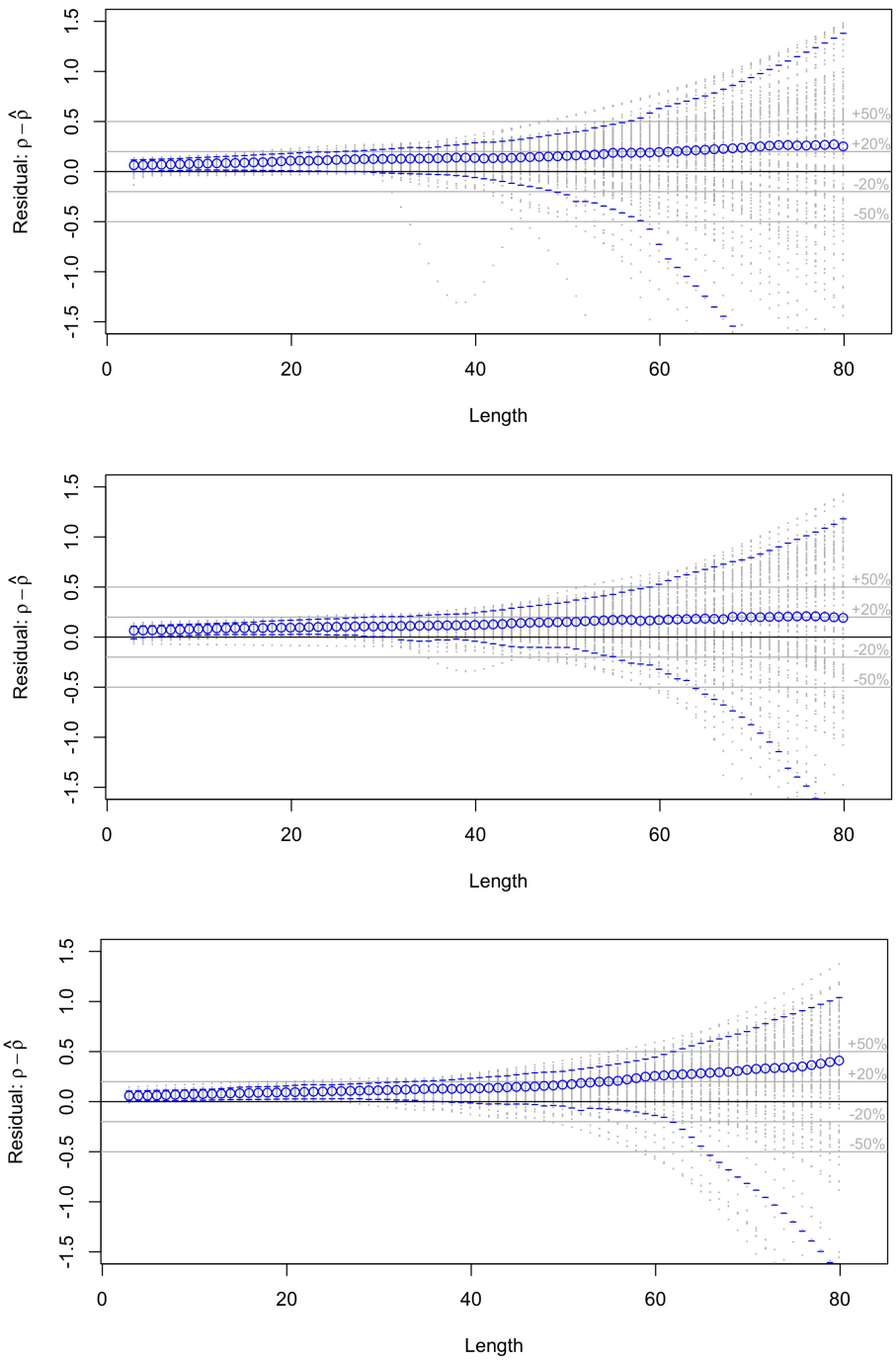

Figure 45. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 10-4-1 (top),10-4-2 (middle) and 10-4-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

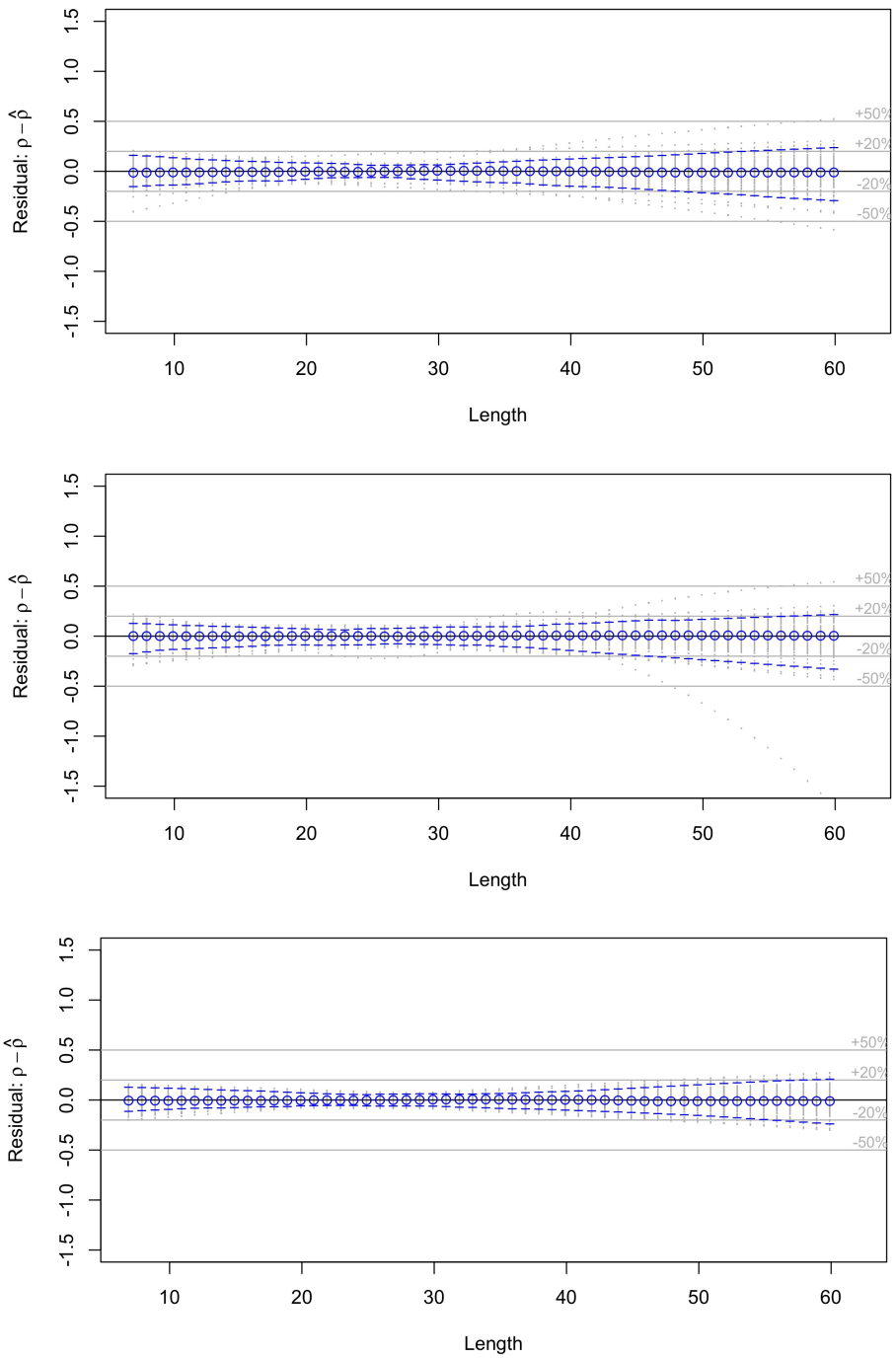

Figure 46. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 14-1-1 (top),14-1-2 (middle) and 14-1-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

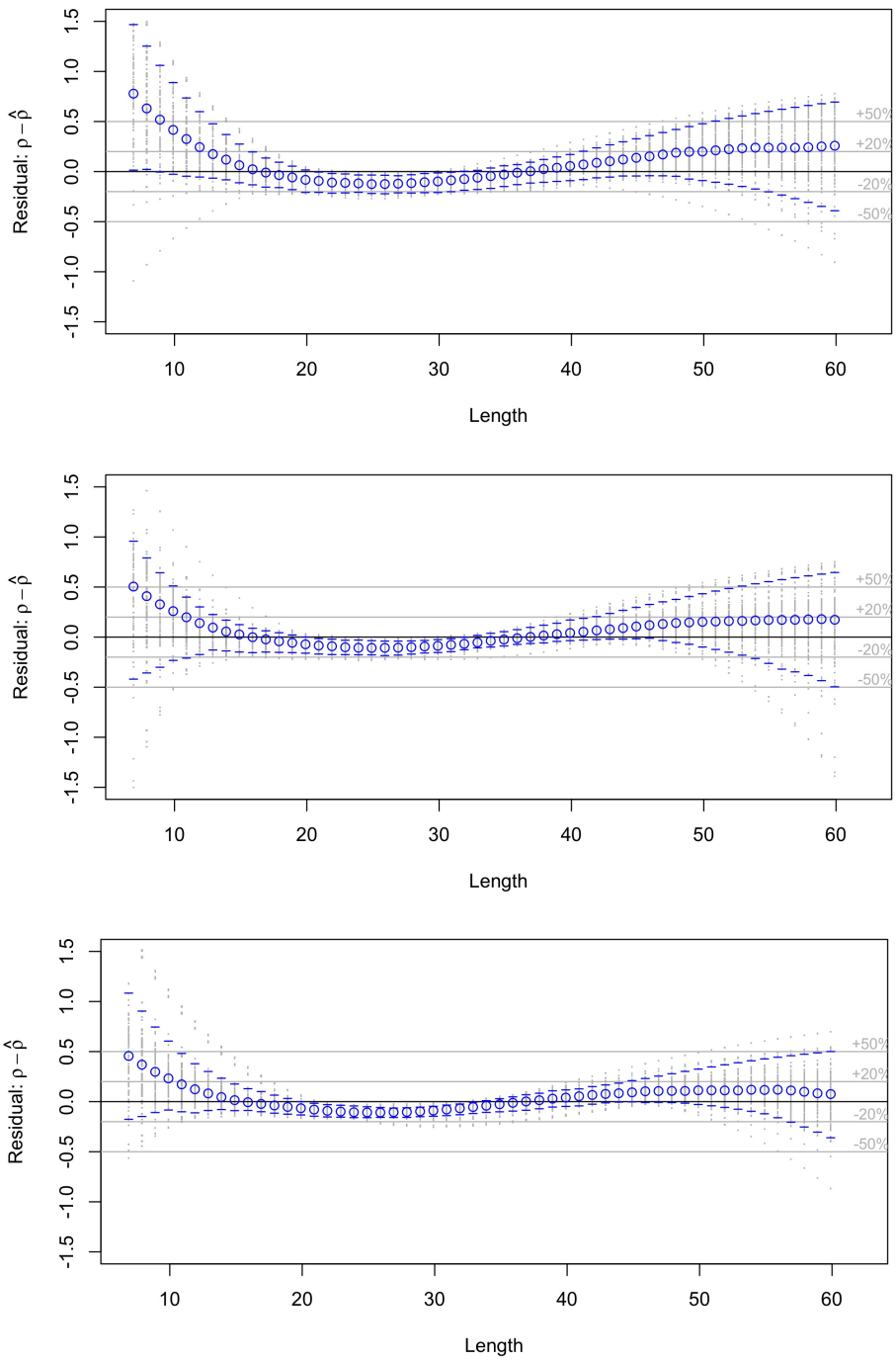

Figure 47. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 14-2-1 (top),14-2-2 (middle) and 14-2-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

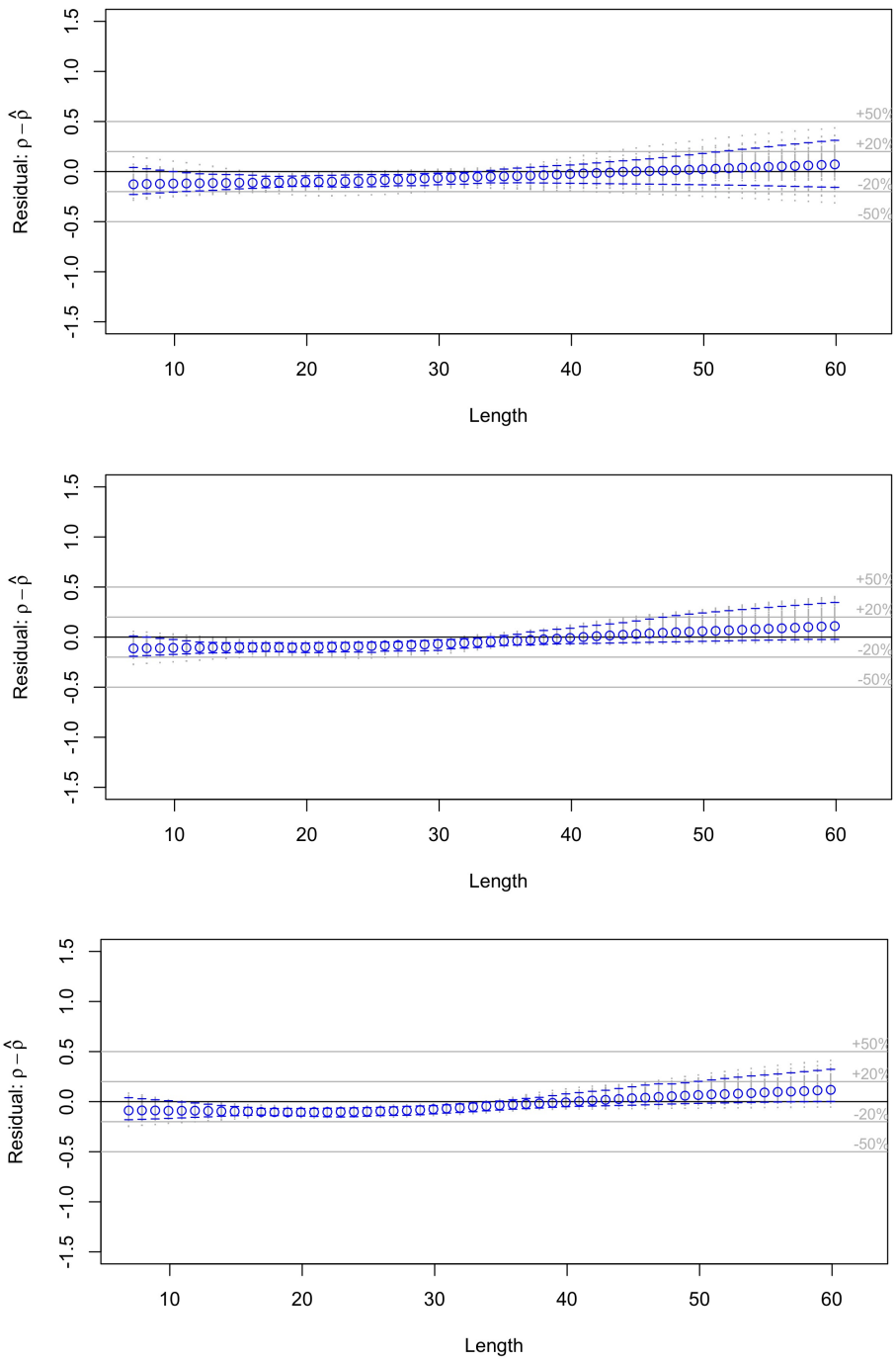

Figure 48. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 14-3-1 (top),14-3-2 (middle) and 14-3-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

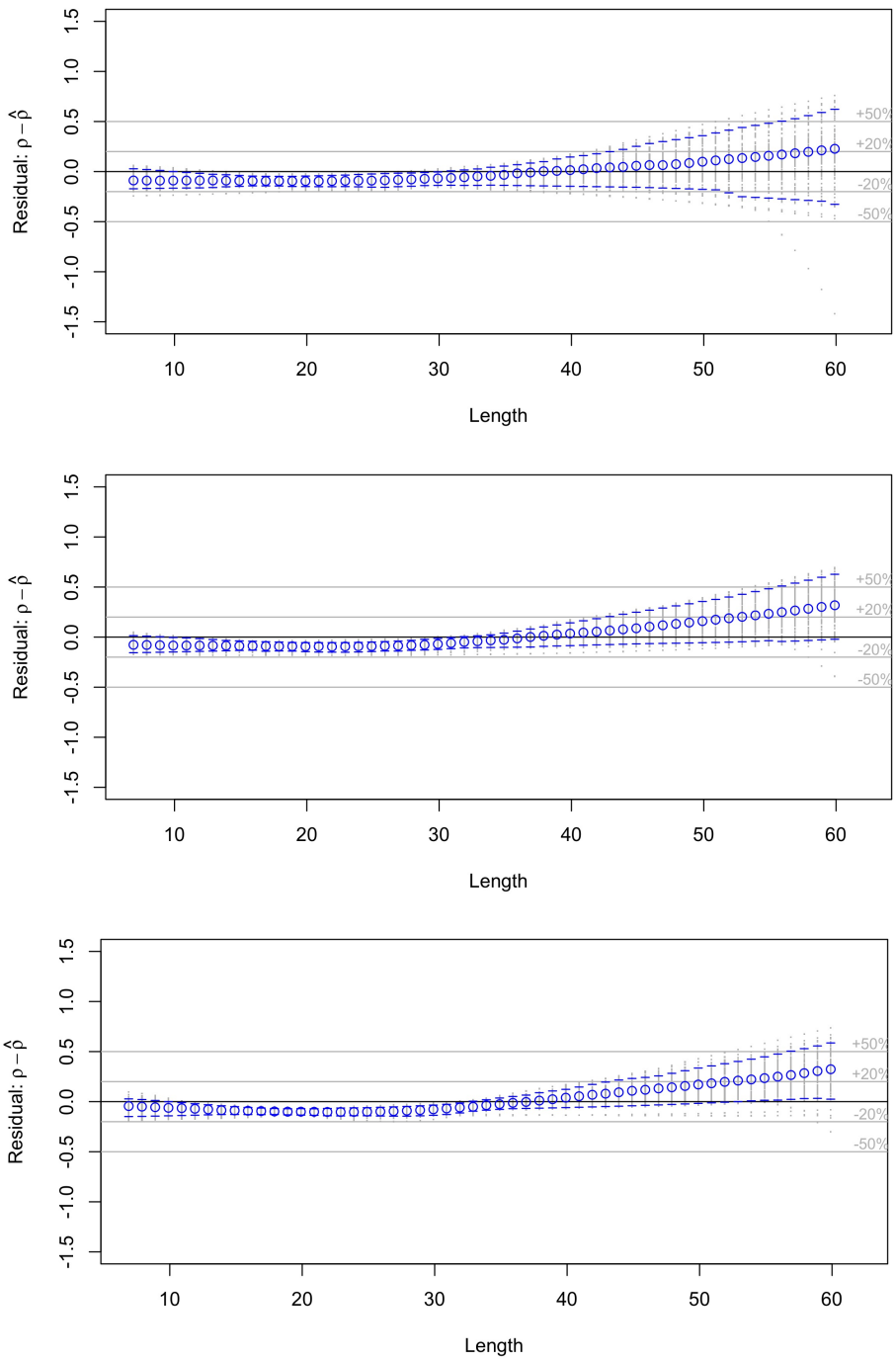

Figure 49. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 14-4-1 (top),14-4-2 (middle) and 14-4-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

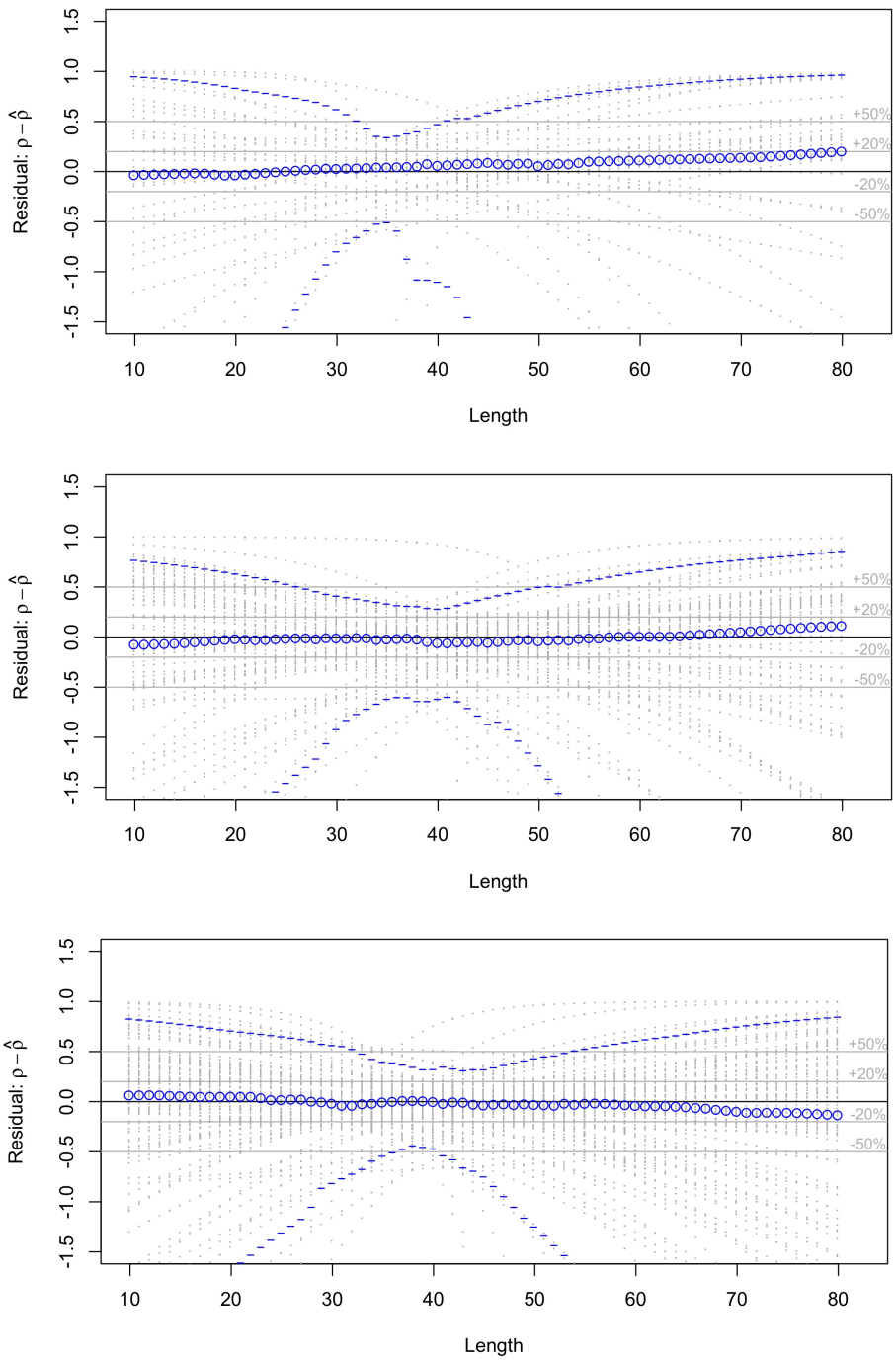

Figure 50. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 204-1-1 (top),204-1-2 (middle) and 204-1-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

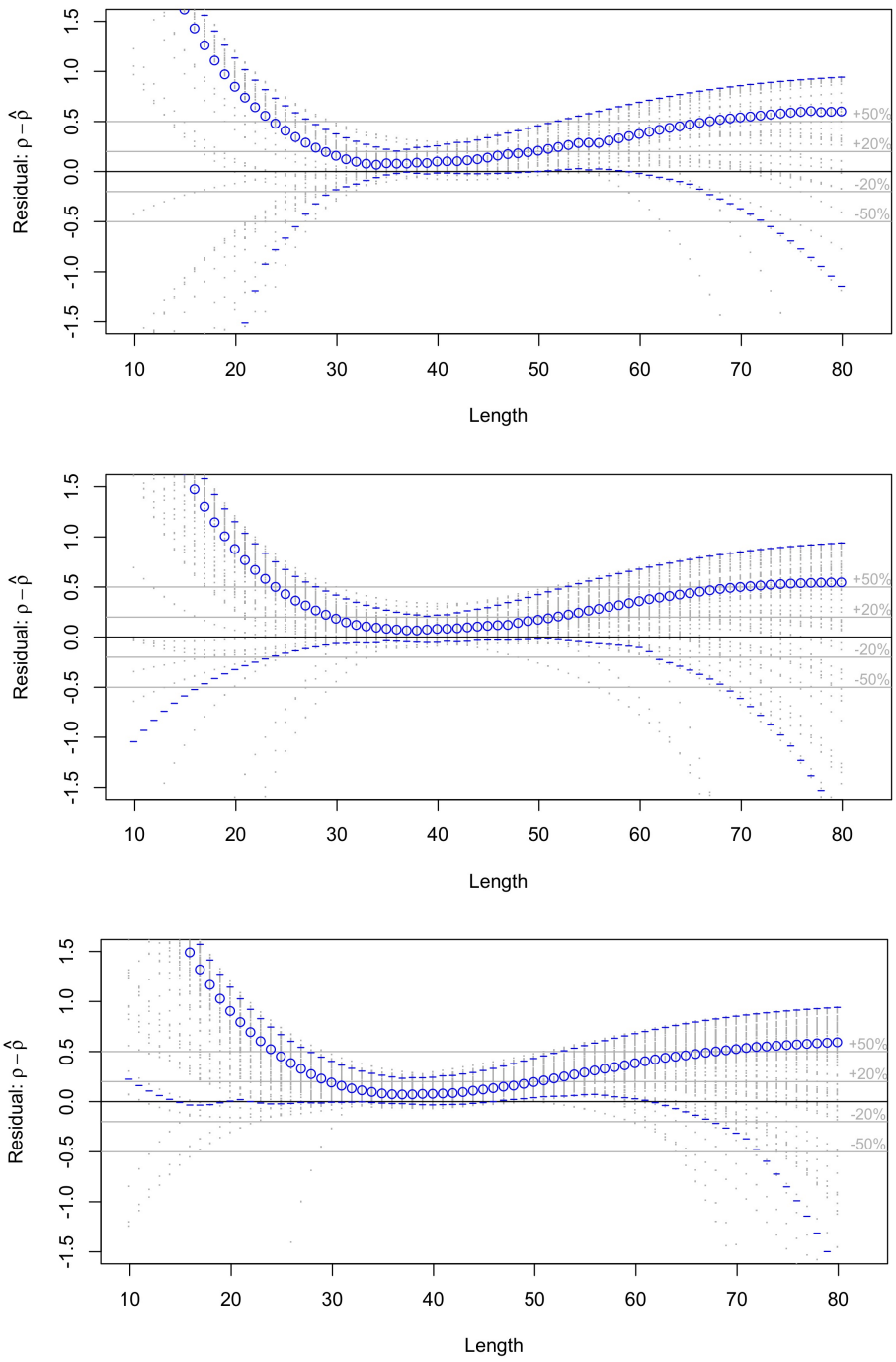

Figure 51. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 204-2-1 (top),204-2-2 (middle) and 204-2-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

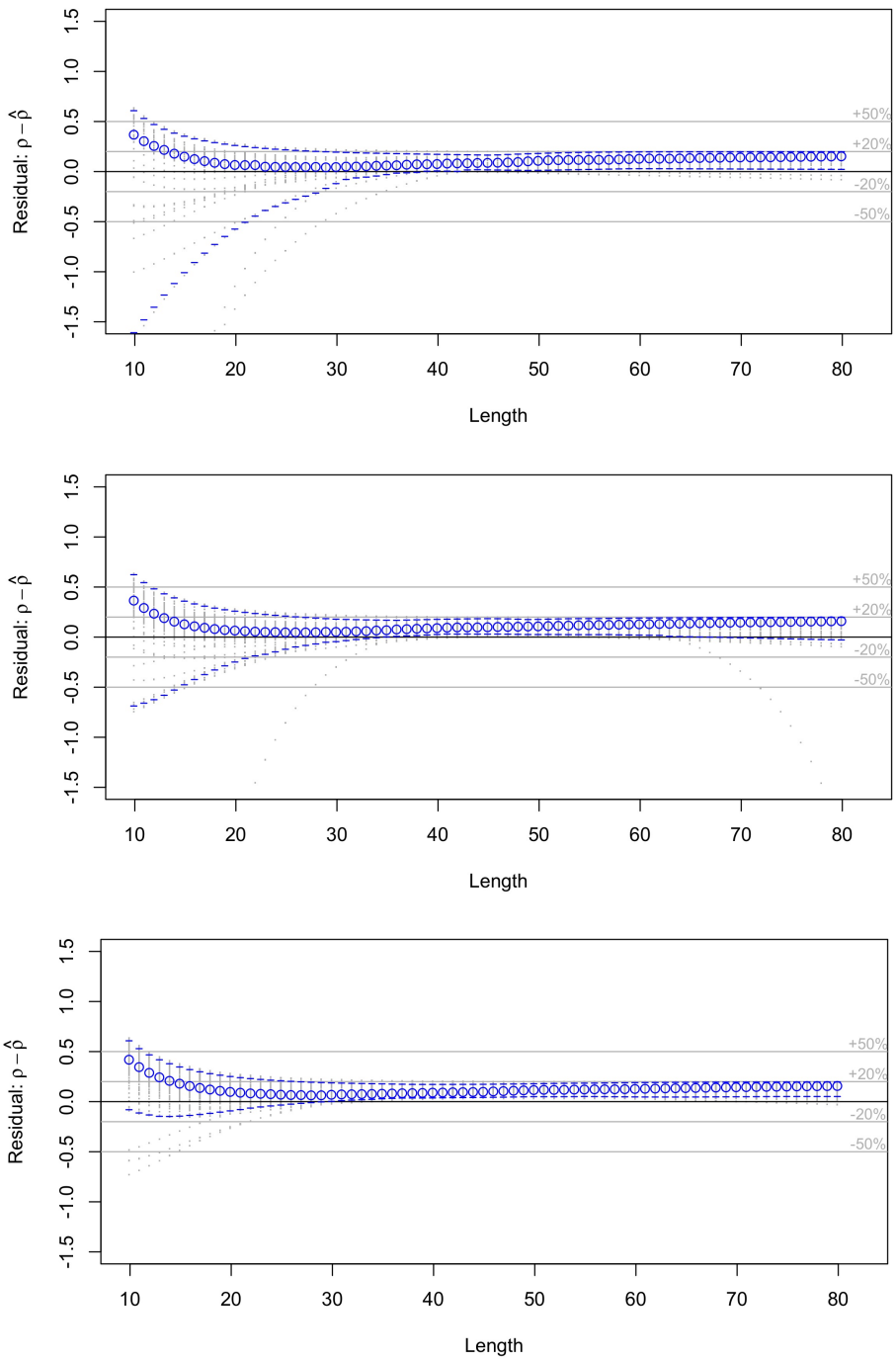

Figure 52. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 204-3-1 (top),204-3-2 (middle) and 204-3-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

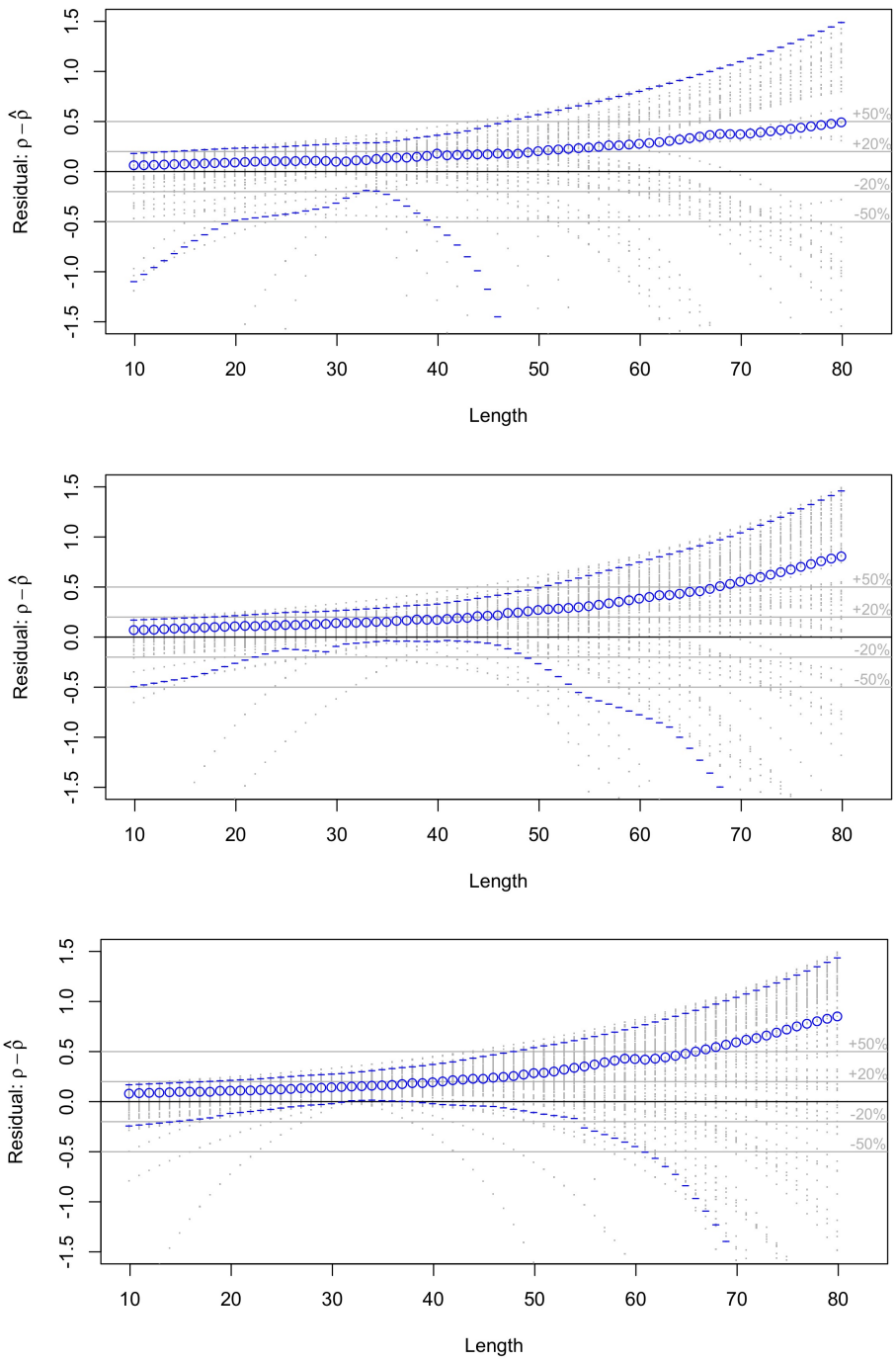

Figure 53. Estimation residuals for the relative catch efficiency over length from 100 iterations for simulation scenarios 204-4-1 (top),204-4-2 (middle) and 204-4-3 (bottom), where 1, 2 and 3 stations per stratum are sampled, respectively. Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 6 for the simulation study designed to assess the effect of sample sizes.

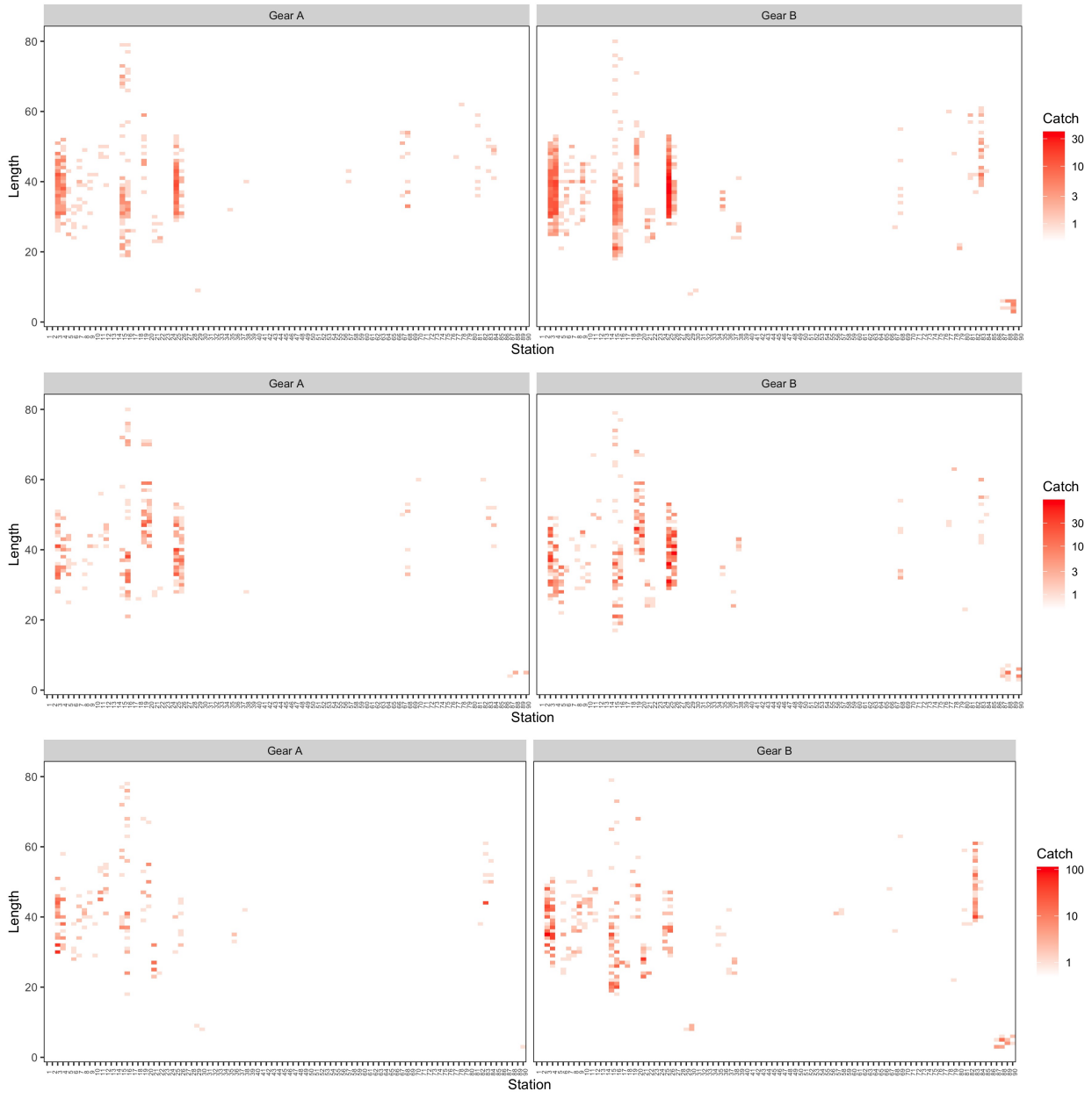

Figure 54. An example of simulate datasets for simulation scenarios 10-4-$Pois$, 10-4-NB and 10-4-$NB-Beta$, i.e., catch numbers for the pair of gears A and B at each station and length (cm). The examples are generated using random seed 1 in R for each scenario. See Section 7 for the simulation study designed to assess the effect of model mis-specification.

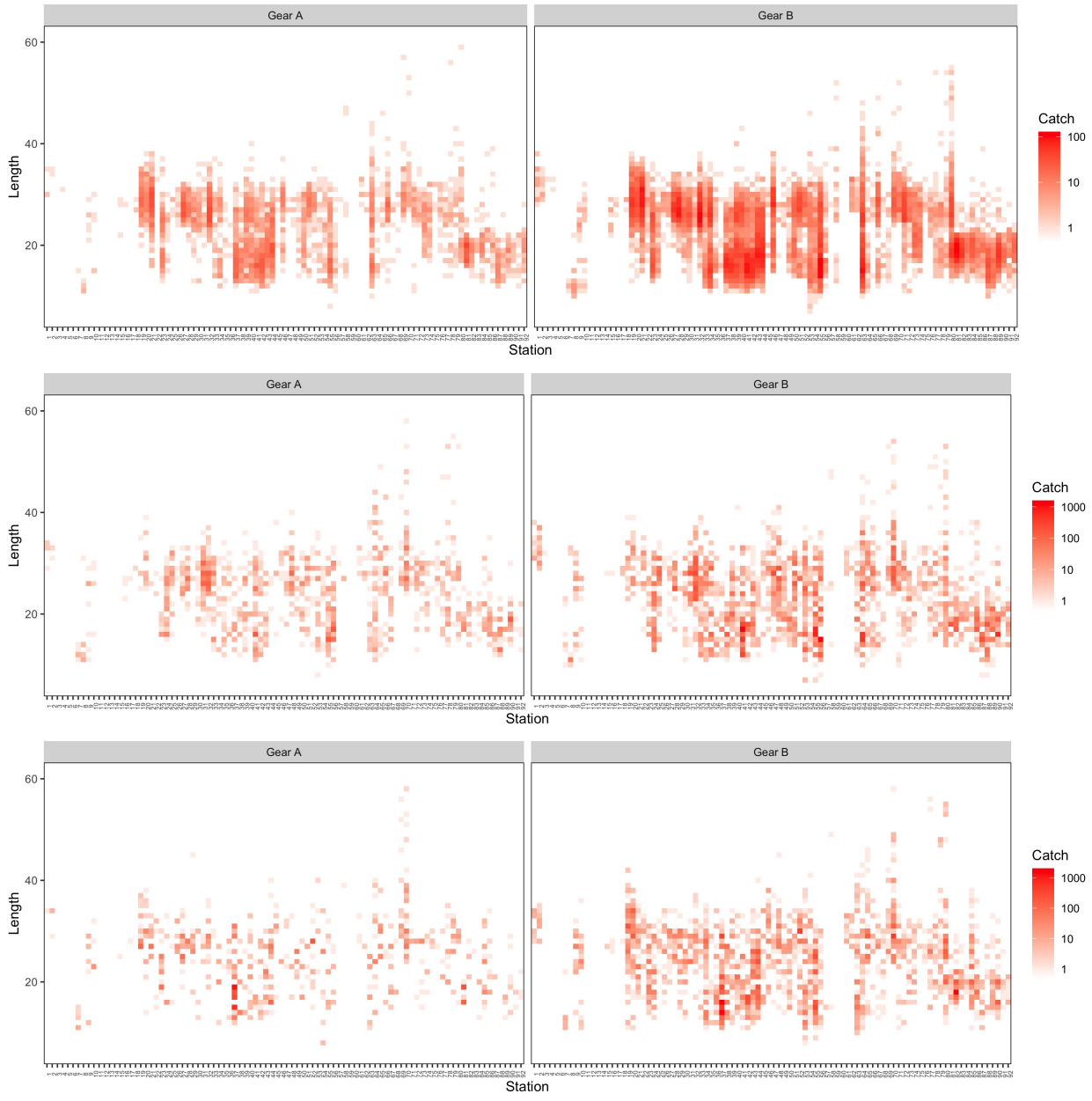

Figure 55. An example of simulate datasets for simulation scenarios 14-4-$Pois$, 14-4-NB and 14-4-$NB - Beta$, i.e., catch numbers for the pair of gears A and B at each station and length (cm). The examples are generated using random seed 1 in R for each scenario. See Section 7 for the simulation study designed to assess the effect of model mis-specification.

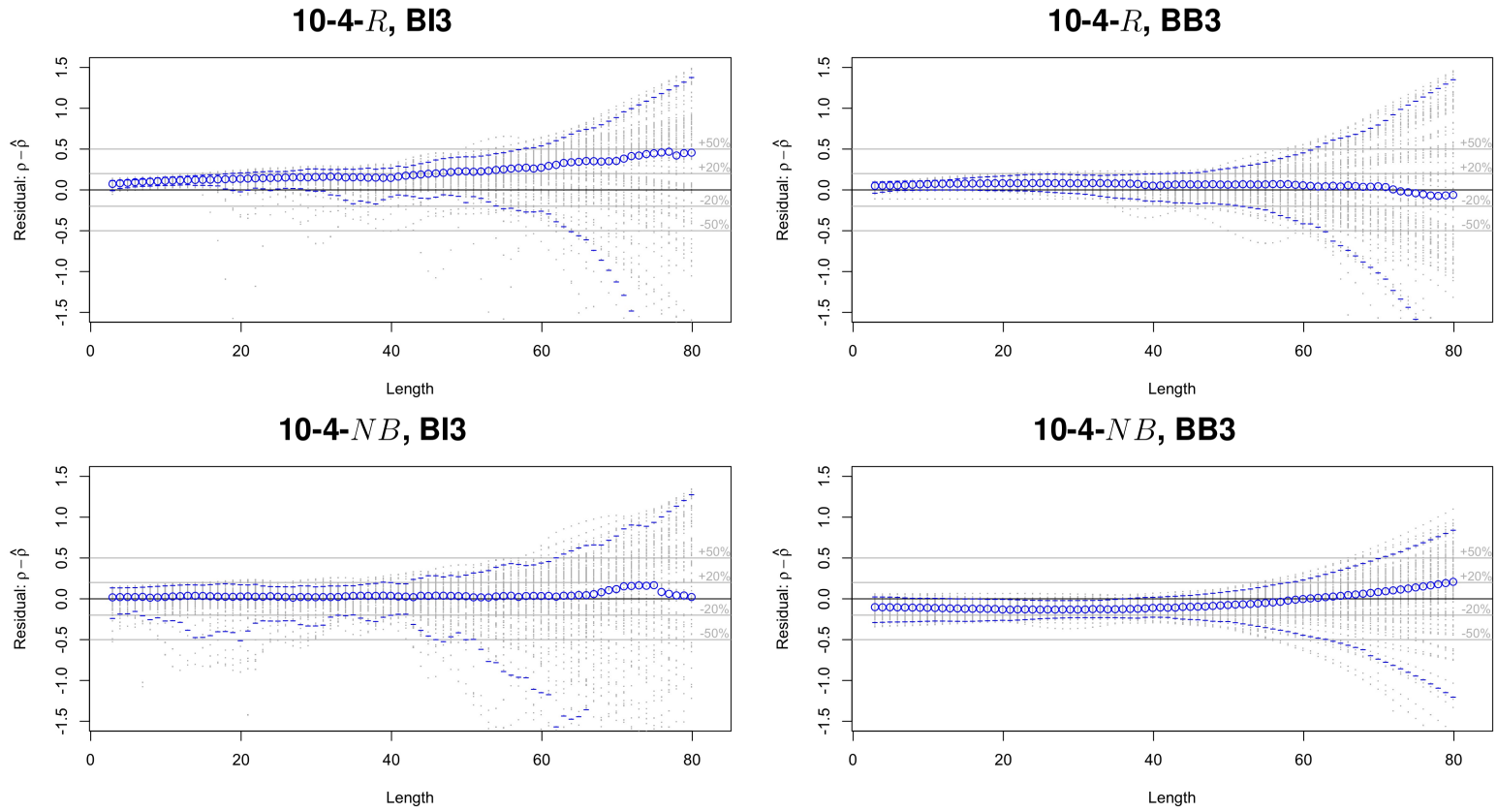

Figure 56. Comparison of estimation residuals: datasets are simulated from scenarios 10-4-R and 10-4-NB fit to estimation models BI3 and BB3 (see Section 7 for definition of scenario coding). Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 7 for the simulation study designed to assess the effect of model mis-specification.

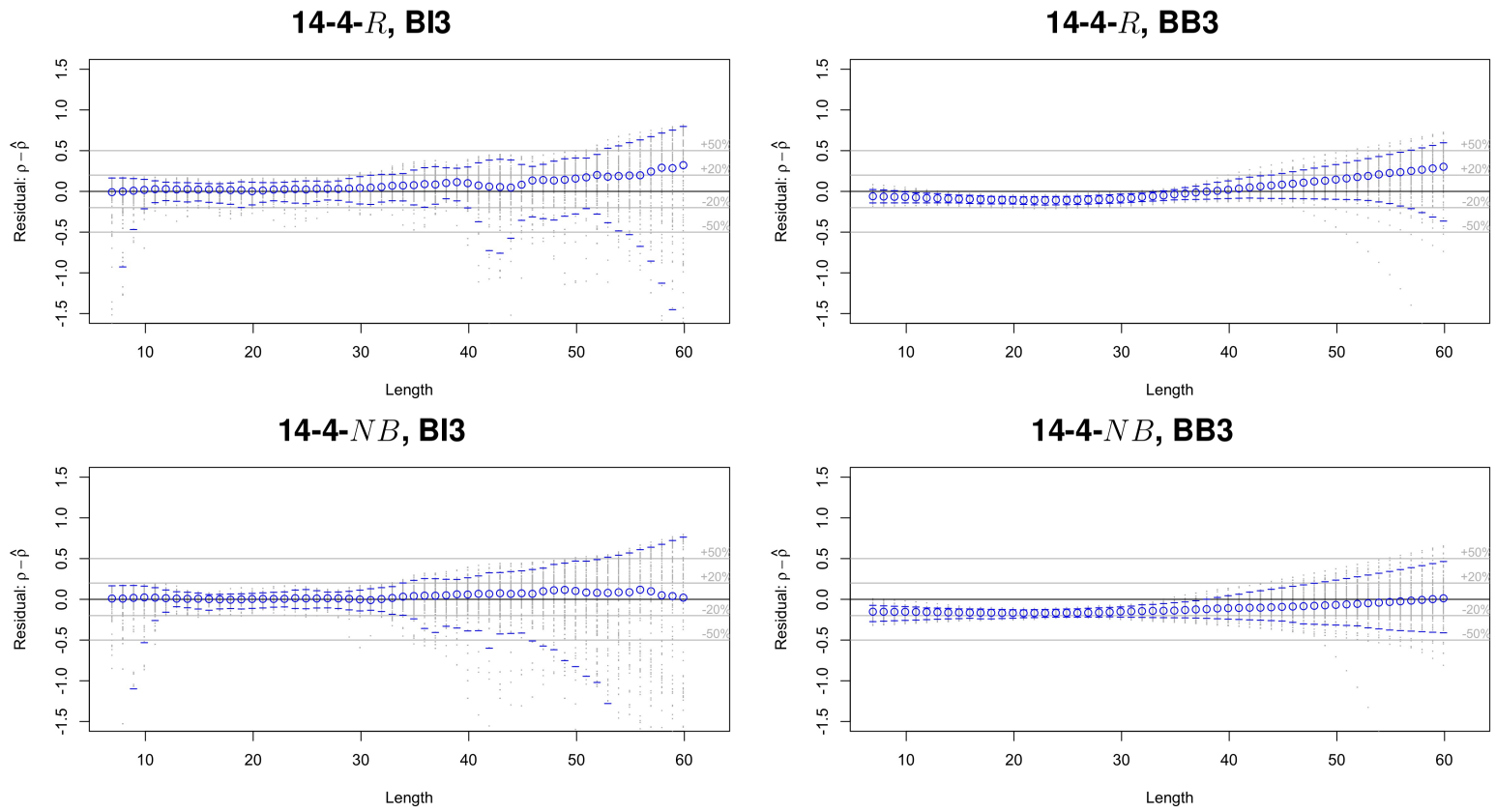

Figure 57. Comparison of estimation residuals: datasets are simulated from scenarios 14-4-R and 14-4-NB fit to estimation models BI3 and BB3 (see Section 7 for definition of scenario coding). Blue dotted lines are residuals from the selected model in each of the 100 simulation iterations, blue circles are the average residuals for the 100 iterations, blue segments are the 5% and 95% quantiles of the 100 iterations, and the gray horizontal lines indicate $\pm 20\%$ and $\pm 50\%$ estimation error ranges. See Section 7 for the simulation study designed to assess the effect of model mis-specification.

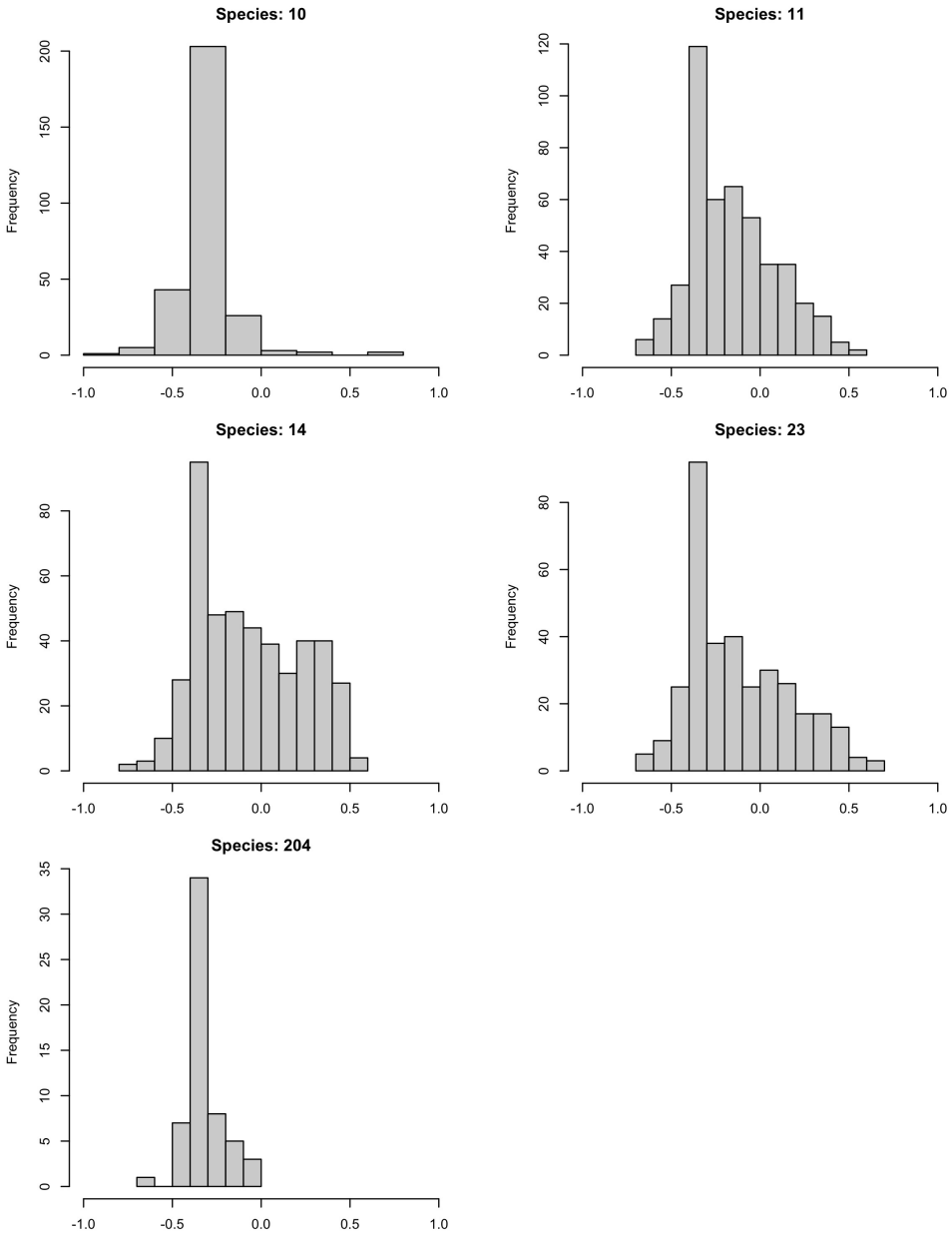

Figure 58. Estimated AR(1) autocorrelation coefficients for each species using anomaly sequences from each station (deviations of survey catch-length compositions from their 5-point moving averages).