Effort optimization of bottom trawl subsampling of fish for length and age composition: tradeoffs between stock assessment input sample size and survey workforce health and efficiency

Pete Hulson1,\*, Benjamin Williams1, Meaghan Bryan2, Jason Conner3, and Matthew Siskey4

1 Auke Bay Laboratories, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 17109 Point Lena Loop Rd., Juneau, AK 99801  
2 Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115  
3 Resource Assessment and Conservation Engineering Division, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115  
4 School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA

\* Correspondence: [Pete Hulson <[pete.hulson@noaa.gov](mailto:pete.hulson@noaa.gov)>](mailto:pete.hulson@noaa.gov)

# Abstract

*Why are you doing this? [context and aim]*

*What did you do?* [*methods*](#methods)

*What did you find? [core results – say something useful – no motherhood statements or deference to the main text!]*

*What does this mean? [interpretation in context]*

*What is it good for? [application]*

# Introduction

Many integrated fishery stock assessments rely on estimates of fish population demographic composition information in the form of length composition (Punt et al. 2013) or as a derived quantity from length composition expanded to age composition (Maunder and Piner 2015). The most reliable age and length composition estimates are based on observations from fishery-independent surveys because these platforms generally avoid the sample selection bias inherent in directed commercial fisheries (Council 1998). Past analyses have focused on the statistical efficiency of composition estimates for data weighting and applications to likelihood functions (e.g., Thorson 2014, Francis 2017, Thorson et al. 2017). It is recognized in sampling theory that increasing the number of samples increases the precision of the estimate, and from this perspective, one should require the maximum number of samples per each survey observation. However, to increase the precision of age or length composition estimates it’s not just the magnitude of sampling that is influential, but also the manner in which the samples are collected (Siskey et al. 2023). Further, each fish sampled on a survey has an immediate physical labor cost (e.g., at-sea time management issues), a cumulative health cost (e.g., repetitive stress injuries), downstream labor costs (e.g., staff that read otoliths to determine age), and budget costs (e.g., the monetary cost to perform a fisheries-independent survey).

It has become commonly understood that sampling to determine age and length composition on fishery-independent or fishery-dependent platforms can be influenced by intra-haul correlation, or, that samples taken from a school of fish that are very similar to each other in size and/or age may not be representative of the overall population’s size and/or age distribution (e.g., Pennington et al. 2000). In order to evaluate and identify the level of intra-haul correlation, and how this can be accounted for in stock assessment, the concept of ‘effective sample size’ has been developed, in which the effective sample size is smaller than the actual sample size and reflects the increase in uncertainty that is due to the intra-haul correlation displayed by the species that are sampled (e.g., McAllister and Ianelli 1997). A number of studies have used effective sample size to evaluate the impact on assessment results including whether effective sample size can be estimated by the assessment model (Hulson et al. 2012, Francis 2017, Thorson et al. 2017) and more recently as a tool to evaluate the implications of changes to sampling methodologies and the subsequent influence on assessment model results (Siskey et al. 2023). Further, methods have been developed that mimic the sampling design for length and age composition that estimate effective sample size from collected samples trough bootstrap techniques that are external to an assessment model (Stewart and Hamel 2014). Overall, the use of effective sample size has become a universal method to implement uncertainty caused by overdispersion of samples in length and age composition data.

Within a fishery-independent survey, thousands of samples are obtained for both age and length composition across multiple species. For example, the National Oceanic and Atmospheric Administration’s (NOAA’s) Alaska Fisheries Science Center (AFSC) is responsible for the execution of several fisheries-independent bottom trawl surveys (Stauffer 2004), spanning most of the continental shelf in Alaskan waters south of the Bering Strait, including the eastern Bering Sea (EBS), the Aleutian Islands (AI), and the Gulf of Alaska (GOA). The quantitative time series of comprehensively cataloging the biota encountered at each sampling station within these surveys began in 1982 with the adoption of standardized trawling protocols (Lauth et al. 2019), but trawl observations were made as early as 1955 (Zimmermann et al. 2009). Within the quantitative time series, AFSC scientists have routinely collected observations of fish length and age composition, consisting of complete or random subsamples of fish within each trawl. There are a number of unintended consequences associated with intense sampling of species within a fishery-independent survey. The length measurement process is repeated over hundreds of thousands of times in a given survey each year, while collecting otoliths is repeated several thousands of times, representing a daunting amount of work for the field scientists working on research vessels. Further, a large portion of the length frequency data that is collected is subsequently sampled for sex determinations in order to support stock assessments that are sex-specific, which involves a substantial increase in handling and data collection effort. Each year, this work flow results in acute and repetitive stress injuries, some requiring medical interventions and claims to the U.S. Office of Workers’ Compensation Programs. Another consequence of the intensity of this work are unrecoverable errors in the observed data; as fatigue or injuries accumulate during the course of field work, so do data collection errors (e.g. failure to properly encode length measurements with the correct species or sex or incorrect identification of sex), despite extensive *in situ* quality assurance protocols. Besides the budgetary constraints that are directly associated with operationalizing a fishery-independent survey, there are downstream fiscal costs associated with composition collections, in particular, requiring properly trained staff to read otoliths collected across a number of different species.

While maximizing the number of samples collected for age and length composition is desirable from a statistical viewpoint, it is becoming increasingly the case that optimizing collections that balance the statistical quality of the data with the health of the workforce and budgetary constraints is more broadly desirable. And in this case, optimizing collections is more often associated with decreasing collection efforts from historical levels. From the stock assessment perspective, the primary concern when reducing the sampling of length frequency data (whether sex-specific or not) and age composition data collections is the impact on the uncertainty in subsequent age and length composition data. The use of effective sample size, particularity when evaluating sampling strategies, is a useful tool to determine the consequences of changing sampling effort. An additional consideration when evaluating the consequences of optimizing sampling effort is the impacts on a species-by-species basis to determine whether there are life-history characteristics that minimize the impacts of reduction in sampling effort in one species compared to another.

In this study, using data collected by the AFSC bottom trawl surveys across the EBS, AI, and GOA we evaluate the consequences of reductions in length frequency collections (including sex-specific length frequency) coupled with reductions in age composition sampling. Using effective sample size as the primary statistic to evaluate uncertainty in the age and length composition data that is derived from the bottom trawl surveys, we evaluated the impact of reducing sampling to answer three questions 1) what is the impact of reducing sampling in length frequency sampling (including sex determination sampling) and the subsequent expansion to age composition on the uncertainty in the composition data? 2) Is there a point of diminishing returns as we increase the number of age and length samples? And, 3) are there life-history characteristics that mitigate, or exaggerate, the consequences of reductions in age or length composition sampling?

# Methods

## Computing length and age composition from bottom trawl survey data

Data collection for each AFSC groundfish bottom trawl survey is described in respective NOAA Technical Memorandums (EBS: Lauth et al. 2019, AI: von Szalay et al. 2017, GOA: von Szalay and Raring 2018). Fundamental methods of length sample collection are generally synchronized between these surveys, with species-specific exceptions for minimum length sub-sample size. Observations of sex-specific length distributions for designated species encountered within a catch sample are collected by 1) sorting the trawl sample by species, 2) weighing each species in aggregate, 3) obtaining a random sub-sample of target sample size, 4) sorting the sub-sample by sex (each fish is cut with a scalpel, gonads are identified and placed in a sex-specific receptacle), and 5) measuring and recording each fish length (each fish is placed on a measuring board, length is identified and the species, sex and measurement are recorded on a computer). To facilitate age estimation, individual fish are processed at sea to record sex, length and weight and to remove sagittal otoliths that are returned the AFSC Age and Growth laboratory for age determination. Survey age sampling protocols are specific by fish species and follow 1 of 2 paradigms: 1) a stratified collection that is distributed over both the spatial frame of the stratification scheme and the expected size range of a species; or 2) a small subsample (3-6 fish, depending on species) collected randomly per trawl. The protocol for some species has changed over the time series, which has followed a trend of transitioning from protocol 1) to protocol 2).

Length frequency samples collected by the AFSC bottom trawl surveys are expanded by area-swept type catch-per-unit-effort (CPUE) and stratum area to obtain estimates of population abundance-at-length (i.e., design-based expansion). In a design-based expansion process, this is often referred to as the ‘first stage expansion’ and is a common method to obtain population estimates at length from area-swept survey data (Miller and Skalski 2006, Ailloud and Hoenig 2019). Population abundance-at-length are computed for three sex categories (males, females, and unsexed) at the stratum level, which are then summed across strata to obtain the population abundance-at-length for the management-scale region (i.e., EBS, AI, or GOA). Strata are defined as regions which have similar bathymetric characterstics, in particular, depth ranges, and population abundance-at-length with strata can also be summed to any sub-region level. Age-length-keys (ALKs) generated from the age-length paired observations within a survey are then applied to estimated abundance-at-length to provide an estimate of abundance-at-age (e.g., Quinn and Deriso 1999). Currently, no species age composition are estimated using depth or area stratified ALKs, rather, all observations are pooled for the entirety of the survey area each survey year. The specific methods AFSC uses to expand length and age samples to abundance are shown in Hulson et al. (in review). The stocks selected for this analysis are shown in Table @ref(tab:species\_sample)), and were selected because all are assessed by AFSC with statistical catch-at-age models and have corresponding expanded age and/or length composition estimates from the respective bottom trawl surveys. Data from the recent three survey years for each survey were used in this analysis in order to show results that reflect consequences of sub-sampling that were most applicable to the current status of the stocks. The average length and age sample sizes from the most recent three survey years for the species selected by survey are also shown in Table @ref(tab:species\_samples)).

## Simulation-Bootstrap framework

To evaluate the effect of reductions in sampling length frequency and age collections we developed a bootstrap-simulation framework that 1) allows for reductions in the historical number of length frequencies and age specimen data collected and 2) conducts the first (length) and second (age) stage expansion process for each bootstrap replicate of length frequency and age specimen information to generate length and age composition [*(****ref?****) - do we cite the R-package here? or the tech-memo (****Hulson2023?****)?*]. We used the historical length frequency and age specimen data that were collected from the AFSC bottom trawl surveys to evaluate the impact of reduced sampling.  
The bootstrap-simulation framework is composed of a suite of nested resampling protocols. Bootstrap resampling was performed either with replacement (wr) or without replacement (wor) depending on the needs of a particular protocol. Functions to run the sampling protocols were developed in a compartmentalized manner to provide for substantial flexibility in exploring desired resampling protocols. The order of operations (Figure @ref(fig:bs\_flows) *ben - could we add in a part that sub-samples ages too to this figure?*) has the following schedule:

1. Resample hauls (wr) from the set of hauls with associated catch per unit effort (in numbers).
2. Within the resampled hauls from step 1, resample the observed lengths (wr).
3. From the resampled lengths in step 2, within each haul subset the lengths (wor) at pre-determined subsampling level.
4. With the resampled and subsampled lenth frequency data in step 3, calculate sex-specific population abundance-at-length.
5. Within the resampled hauls from step 1, resample the observed ages (wr).
6. From the resampled ages in step 5, for the total ages sampled across hauls subset the ages (wor) at pre-determined subsampling level.
7. With the resampled and subsampled age data in step 6 and the sex-specific population abundance-at-length in step 4, calculate sex-specific population abundance-at-age.

The core of the bootstrap-simulation function (steps 3 and 6 above) is designed to explore reductions in the sample size of lengths that are collected on a per haul basis, as well as in the aggregated sample size for ages. In this bootstrap-simulation, the number of lengths (whether total or sex-specific) in a given haul must be less than or equal to the desired sample size *x* determined in step 2. In step 3, when the number of resampled lengths from step 2 in a haul is less than *x* then is used directly, if then a random draw of lengths is taken without replacement in step 3. We set the subsampling level for length frequency at numbers per haul in order to evaluate the consequences of reductions in length sampling at the haul level. Alternatively, to subsample ages we set the proportion of the total number of ages sampled in step 6 in order to evaluate the consequences of reductions in overall age sampling, as by design at the per haul bases there are a small number of ages sampled. The bootstrap-simulation then repeated steps 1-7 iteratively for each length and age subsample level, providing iterated population abundance-at-length and age that was then compared to the historical (the full sample without any resampling of data) population abundance-at-length and age determined by the bottom trawl surveys.

The length subsample levels that we evaluated from the historical length frequency collections were 50, 100, 150, 200, and 250 samples per haul. The age subsample levels that we evaluated from the historical age specimen collections were 25%, 50%, 75%, and 90% of the total number of ages collected in any given survey year. We also ran the bootstrap-simulation for the historical number of length frequency and age specimen collections without subsetting in order to compare the base level uncertainty to the increase in uncertainty gained through sub-sampling. We ran the bootstrap-simulation for 500 iterations, which was a level for which the variability in population abundance at length results had stabilized. The bootstrap-simulation was developed in R (R Core Team 2022) and is available via GitHub as an R package (<https://github.com/BenWilliams-NOAA/swo>).

## Computing effective and input sample size

We used two performance metrics to quantify potential changes in uncertainty of length and age compositions (i.e., population abundance-at-length and age) owing to reduced sampling of length frequency and age specimen information. First, effective sample size was used to compare each bootstrap replicate of length and age composition to the original length and age composition calculated from the complete historical dataset. Second, input sample size was used to summarize uncertainty across bootstrap replicates as the harmonic mean of effective sample size replicates.

Effective sample size, as introduced by McAllister and Ianelli (1997), is a statistic that can evaluate the level of intra-haul correlation in composition samples that are collected on a survey (whether from age or length frequency collections). It is also a statistic that can evaluate the amount of uncertainty in an estimated composition compared to an observed composition. Effective sample size is given by:

where is the estimated proportion for category-*c* (which can be either age or length or any other arbitrary category across which proportions are computed) and is the observed proportion. It can be interpreted by a higher ESS indicates less uncertainty in the composition estimates, while lower ESS indicates more uncertainty.

In this bootstrap-simulation the underlying length and age composition derived from the historical bottom trawl surveys was treated as the observed proportions in equation (1). For each iteration of the bootstrap simulation for a determined sub-sample size or proportion we computed an estimated proportion () that was then compared to the underlying historical sex-specific length and age composition (the effective sample size for the total length and age composition, as the sum of population abundance at length, was also computed). Thus, across each iteration of the bootstrap simulation we computed an effective sample size that indicated the amount of increased uncertainty that was caused by sub-sampling length frequency and age specimen data. To summarize effective sample size across iterations we used the harmonic mean, which has been shown to reduce bias in recovering the true sample size in simulations for a multinomial distribution. Due to this reduction in bias the harmonic mean has also been recommended to determine the ‘input sample size’ that is used in stock assessment models to fit compositional data (Stewart and Hamel 2014). Herein, when we use the term ‘effective sample size’ we are referring to the effective sample sizes that were computed for each iteration of the bootstrap-simulation, when we use the term ‘input sample size’ we are referring to the harmonic mean of the iterated effective sample sizes.

## Evaluating life-history relationships to consequences of subsampling

[add text here describing what was done to compare reductions to life history]

# Results

[length results for selected species, including samples saved, ess and iss reductions - Pete and Ben/Meaghan]

[age results, including ess and iss reductions - Pete and Ben/Meaghan]

# Discussion

[Summary para of main results]

[Cost-benefit of precision compared to survey injuries]

[Still thinking on other paras]

# Acknowledgments

We thank *reviewer1* and *reviewer2* for their helpful reviews of this manuscript.

# Citations

Ailloud, L.E., and Hoenig, J.M. 2019. A general theory of age-length keys: Combining the forward and inverse keys to estimate age composition from incomplete data. ICES Journal of Marine Science 76(6): 1515–1523.

Council, N.R. 1998. Improving fish stock assessments. The National Academies Press, Washington, DC. doi: [10.17226/5951](https://doi.org/10.17226/5951).

Francis, R.I.C.C. 2017. Revisting data weighting in fisheries stock assessment models. Fisheries Research 192: 5–15.

Hulson, P.-J.F., Hanselman, D.H., and Quinn II, T.J. 2012. Determining effective sample size in integrated age-structured assessment models. ICES Journal of Marine Science 69: 281–292.

Hulson, P.-J.F., Williams, B., Siskey, M., Bryan, M., and Conner, J. in review. Bottom trawl survey age and length composition input sample sizes for stocks assessed with statistical catch-at-age assessment models at the Alaska Fisheries Science Center. U.S. Department of Commerce. NOAA Technical Memorandum.

Lauth, R.R., Dawson, E.J., and Conner, J. 2019. Results of the 2017 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna. U.S. Department of Commerce. NOAA Technical Memorandum NMFS-AFSC-396: 260 p.

Maunder, M.N., and Piner, K.R. 2015. Contemporary fisheries stock assessment: Many issues still remain. ICES Journal of Marine Science 72(1): 7–18.

McAllister, M.K., and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences 54(2): 284–300.

Miller, T.J., and Skalski, J.R. 2006. Integrating design- and model-based inference to estimate length and age composition in north pacific longline catches. Canadian Journal of Fisheries and Aquatic Sciences 63(5): 1092–1114.

Pennington, M., Burmeister, L.M., and Hjellvik, V. 2000. Assessing the precision of frequency distributions estimated from trawl-survey sampes. Fishery Bulletin 100(1): 74–80.

Punt, A.E., Huang, T., and Maunder, M.N. 2013. Review of integrated size-structured models for stock assessment of hard-to-age crustacean and mollusc species. ICES Journal of Marine Science 70(1): 16–33.

Quinn, T., and Deriso, R. 1999. QUantitative fish dynamics. Oxford University Press, New York, NY.

R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from <https://www.R-project.org/>.

Siskey, M.R., Punt, A.E., Hulson, P.-J.F., Bryan, M.D., Ianelli, J.N., and Thorson, J.T. 2023. The estimated impact of changes to otolith field-sampling and ageing effort on stock assessment inputs, outputs, and catch advice. Canadian Journal of Fisheries and Aquatic Sciences 80(1): 115–131.

Stauffer, G. 2004. NOAA Protocols for Groundfish Bottom Trawl Surveys of the Nation’s Fishery Resources. U.S. Department of Commerce. NOAA Technical Memorandum NMFS-F/SPO-65: 205 p.

Stewart, I.J., and Hamel, O.S. 2014. Bootstrapping of sample sizes for length-or age-composition data used in stock assessments. Canadian Journal of Fisheries and Aquatic Sciences 71(4): 581–588.

Thorson, J.T. 2014. Standardizing compositional data for stock assessment. ICES Journal of Marine Science 71(5): 1117–1128.

Thorson, J.T., Johnson, K.F., Methot, R.D., and Taylor, I.G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research 192: 84–93.

von Szalay, P.G., and Raring, N.W. 2018. Data Report: 2017 Gulf of Alaska bottom trawl survey. U.S. Department of Commerce. NOAA Technical Memorandum NMFS-AFSC-374: 260 p.

von Szalay, P.G., Raring, N.W., Rooper, C.N., and A, L.E. 2017. Data Report: 2016 Aleutian Islands bottom trawl survey. U.S. Department of Commerce. NOAA Technical Memorandum NMFS-AFSC-349: 161 p.

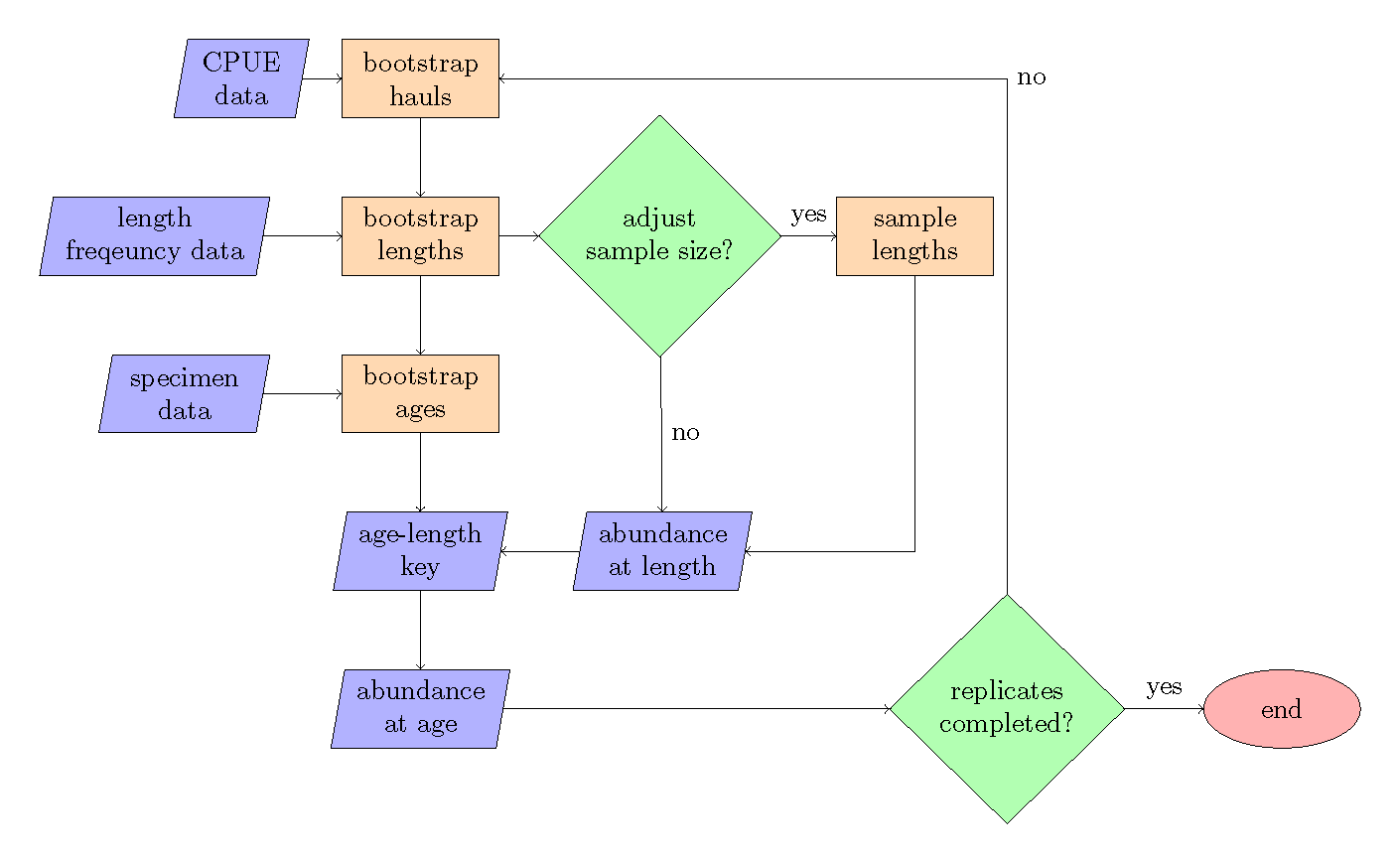
Zimmermann, M., Dew, C.B., and Malley, B.A. 2009. History of Alaska red king crab, Paralithodes camtschaticus, bottom trawl surveys, 1940-61. Marine Fisheries Review 71(1): 1.

# Tables

(#tab:species\_sample)Average length and age (shown in parentheses) samples from the most recent three AFSC bottom trawl surveys by region for the species evaluated in the bootstrap-simulation for reduction in length and age collections.

| Stock | Scientific name | AI | EBS | GOA |
| --- | --- | --- | --- | --- |
| Alaska plaice | Pleuronectes quadrituberculatus | – | 10,486 (327) | – |
| arrowtooth flounder | Atheresthes stomias | 9,868 (460) | 14,928 (702) | 36,842 (597) |
| Atka mackerel | Pleurogrammus monopterygius | 7,888 (610) | – | – |
| Dover sole | Microstomus pacificus | – | – | 3,020 (224) |
| flathead sole | Hippoglossoides elassodon | 4,602 (0) | 22,356 (810) | 15,233 (225) |
| Greenland turbot | Reinhardtius hippoglossoides | – | 266 (131) | – |
| Kamchatka flounder | Atheresthes evermanni | 3,526 (0) | 2,454 (460) | – |
| northern rock sole | Lepidopsetta polyxystra | 10,754 (0) | 24,698 (584) | 3,685 (197) |
| northern rockfish | Sebastes polyspinis | 14,928 (571) | – | 2,298 (447) |
| Pacific cod | Gadus macrocephalus | 5,723 (604) | 11,477 (1,449) | 3,452 (464) |
| Pacific ocean perch | Sebastes alutus | 3,2491 (1,019) | – | 23,319 (1,150) |
| rex sole | Glyptocephalus | – | – | 12,878 (412) |
| southern rock sole | Lepidopsetta billineta | – | – | 7,190 (730) |
| walleye pollock | Gadus chalcogrammus | 10,806 (791) | 4,9,544 (1600) | 16,772 (1,075) |
| yellowfin sole | Limanda aspera | – | 28,108 (845) | – |

# Figures



(#fig:bs\_flows)Bootstrap-simulation flow chart, the steps refer to the order of operations as described in the *Bootstrap-simulation framework* section.