Inclusion of ageing error and growth variability in the estimation of age composition input sample size

Pete Hulson1,\*, and Benjamin Williams1

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

\* 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

Compositional information on age and length comprise critical data products used in statistical catch-at-age assessment models, as they facilitate the tracking of year classes and size-structure over time and improve our understanding of the population dynamics (Quinn and Deriso 1999). There are two primary sources for age and length composition data used in statistical catch-at-age models: fishery-independent sources, which include some level of randomized and standardized collection of samples in a non-targeted framework, and fishery-dependent sources, in which collection of age and length samples are also randomized at some level but are obtained from hauls that are not random but rather targeted within a specific fishery. Regardless of the source of age and length composition data, it is commonly accepted that overdispersion of collections are inherent to the data due to intra-haul correlation (e.g., **Pennington2000?**). The concept of ‘effective sample size’ has since been developed, that is smaller than the sample size collected, to reflect the increased uncertainty due to this overdispersion (e.g., McAllister and Ianelli 1997). Since modern statistical catch-at-age stock assessment models integrate multiple sources of data related to catch (e.g., fisheries catch-per-unit-effort, survey indices of abundance), life history (e.g., size-at-age, maturity-at-age, selectivity-at-age), and composition (e.g., length and age), it is imperative to consider the relative information content these data products provide to the model employed.

Because fisheries often depend on the periodic production of strong year-classes and subsequent recruitment into fishery catch, sampling efforts for age and length data, scaling of these data to the population level (‘compositional expansion’), and the weight assigned to these data products in assessment models are highly important in order to provide accurate advice for management. This is often handled through the use of data-weighting methods, checking the fit of compositional data in the model, and ensuring a good match between the variance of the data and the variance implied by the model (Francis 2017). The weight assigned to annual compositions (the ‘input sample size’) can follow a myriad of methods (e.g., fixed values as in Monnahan et al. (2021), number of samples or tows sampled upon as in Hulson et al. (2021) or Spencer and Ianelli (2022), bootstrapping compositions as in Stewart and Hamel (2014)). The primary consideration when assigning a weight to age or length composition data is to account for the potential variability and correlation in the sampling process. The method developed by Stewart and Hamel (2014) allows for resampling techniques to be employed at each level of the sampling design, and provides an objective avenue to determine input sample size that is smaller than the actual sample size based on the observation variability contained within the sampling process.

In addition to intra-haul correlation, for every fish species sampled for which age is capable of being determined from otoliths there is resulting variability in the ageing of the otolith, often called ‘ageing error’ (e.g., Punt et al. 2008). There are a number of factors that can influence the magnitude of ageing error (Nesslage et al. 2022), but inherent to obtaining ages from otoliths is variability in the age readings. To account for this source of variability, ageing laboratories regularly evaluate precision through obtaining multiple readings of the same otolith across different age readers (Morison et al. 2005). Several methods have been developed to account for ageing error in statistical catch-at-age models when fitting age composition through the use of an ageing error matrix (Punt et al. 2008, Candy et al. 2012). The essence of this approach is that the numbers-at-age estimated by the assessment model are ‘corrected’ through multiplication with an ageing error matrix, which assigns estimated numbers-at-age to adjacent age-classes depending on the magnitude of the ageing error within the specific age-class. Since the development and implementation of ageing error matrices a number of studies have been devoted to quantifying the effects of ageing error on assessment model estimates Liao et al. (2013). Within each of these studies, and in each application of an ageing error matrix within a stock assessment model, the age composition data fit will be weighted by an input sample size. As described previously, the input sample size selected would reflect the variability in the sampling process, which would also include the variability in the age readings themselves.

In the process of compositional expansion, it is often the case that an age-length key (ALK) is employed to expand population numbers-at-length to population numbers-at-age (Ailloud and Hoenig 2019). It is through the age-length key, and the subsequent age expansion, that observations of age composition are derived from fishery-independent and fishery-dependent sources. Conditional age-at-length (CAAL), in which paired age-length data are used as in indication of the age distribution for a specific length, is used to inform length-at-age and it’s related uncertainty (Taylor and Methot 2013). CAAL data can be used directly within statistical catch-at-age models to inform estimates of growth as well as composition data (Lee et al. 2019) and has been implemented in a number of operational stock assessments (e.g., McGilliard et al. 2019, Hulson et al. 2022). An intrinsic component to both the ALK and CAAL is the variability in length for a given age. Further, when using CAAL data as an additional likelihood component to a statistical catch-at-age model one must determine the input sample size to be used to weight this information.

Despite the acceptance of requiring an input sample size to weight age composition data in statistical catch-at-age models that reflects the added uncertainty caused by overdispersion common to age sampling, and the recognition of the inherent variability in the ageing process when reading otoliths and in the growth process upon which age-length keys are based, these sources of uncertainty have not been previously integrated in an objective estimation method for input sample size. In this study, we use the methods of Stewart and Hamel (2014) to estimate age composition input sample size that includes both ageing error and growth variability in the estimation process. We show, in a step-wise process, the added variability in age composition sample size from including ageing error and growth variability across a number of species that reflect differing life histories and levels of ageing difficulty.

# Methods

## Data: age collections and reader-tester agreement

In this study we used historical data collected from bottom trawl surveys conducted by the Alaska Fisheries Science Center (AFSC) in the Eastern Bering Sea (EBS: Lauth et al. 2019), Aleutian Islands (AI: von Szalay et al. 2017), and Gulf of Alaska (GOA: von Szalay and Raring 2018). Within the AFSC bottom trawl surveys both length frequency data and age specimen data are collected (in addition to other data sources including catch, effort, location, etc.). To facilitate age estimation, individual fish are processed at sea to record sex, length and weight and to remove sagittal otoliths that are sent to the AFSC Age and Growth laboratory for age determination. Periodically, a subset of these otoliths are selected for reader-tester agreement tests, which evaluates the reproducability of an age reading when two different age readers age the same fish, without knowledge of the other reader’s determined age of the otolith (Kimura and Lyons 1991). The average annual bottom trawl survey age sample sizes for the stocks selected for this analysis by region, along with the total number of otoliths that have been selected for reader-tester agreement tests across these regions is shown in Table ??). The stocks selected to use as examples for this analysis were all stocks that had greater than 5,000 reader-tester paired otolith readings. These stocks are also all assessed with statistical catch-at-age models that require input sample sizes to ‘weight’ the age composition data fit in the stock assessment models. Details of how the length frequency and age collections are expanded to population abundance-at-length and age at AFSC to then subsequently used as compositional data in stock assessment models at AFSC are provided in Hulson et al. (in review).

## Simulation-Bootstrap framework

To evaluate the effect of the inclusion of ageing error and growth variability on uncertainty in age compostion datasets we developed a bootstrap-simulation framework that allowed for the addition of these sources of error. In simple terms, the simulation framework that we developed resamples the hauls, then lengths and ages collected within the resampled hauls following from the method introduced by Stewart and Hamel (2014). To implement ageing error, for a given resampled age, the set of tester ages for that reader age were pooled and a new age was resampled from that set. To implement growth variability, for a given resampled age the lengths observed for that age were pooled by sex, and then a new length was resampled from that set of lengths.

The bootstrap-simulation framework is composed of a suite of nested resampling protocols. 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 1 *ben - we need to add where we’re doing the ageing error and growth resampling to this figure*) has the following schedule, with steps 1-2 and 4-6 being optional switches:

1. Resample hauls 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
3. With the resampled length frequency data from step 2, calculate population abundance-at-length
4. Within the resampled hauls from step 1, resample the observed ages from the specimen data
5. For the resampled ages in step 4, resample a length from the set of lengths observed for the given age
6. For the resampled ages in step 4, resample an age from the set of tester ages for the given age
7. With the resampled age data in steps 4-6 and the population abundance-at-length in step 3, calculate the population abundance-at-age

The core of the bootstrap-simulation function (steps 5 and 6 above) were designed to explore inclusion of ageing error and growth variability. The bootstrap-simulation then repeated steps 1-7 iteratively providing iterated population abundance-at-age that was then compared to the historical (the full sample without any resampling of data) population abundance-at-age determined by the bottom trawl surveys.

We applied the bootstrap-simulation in a step-wise manner to evaluate the consequences of adding each source of additional error to the age composition estimates across what we term ‘uncertainty scenarios’. First, we ran the bootstrap-simulation with the historical data without adding any extra error, thus, omitting steps 5 and 6 above (termed the ‘Base’ scenario). Next, we added ageing error (termed the ‘AE’ scenario) and growth variability (termed the ‘GV’ scenario) separately, thus, omitting either step 5 or 6 depending on the source of uncertainty that we wanted to include. Finally, we added both ageing error and growth variability (termed the ‘AE & GV’ scenario) to the bootstrap-simulation framework. To allow for the largest number of samples from which the consequences of these sources of uncertainty could be evaluated we included reader-tester data that was pooled across these three regions. To evaluate the inclusion of growth variability we pooled the sex-specific age-length pairs across all the years of the survey in order to provide the maximum influence of growth variability on the replicated sex-specific age composition estimates. We ran the bootstrap-simulation for 500 iterations, which was a level for which the variability in population abundance-at-age 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/surveyISS>).

## Computing effective and input sample size

A useful statistic that can quantify the variability in age composition is effective sample size, introduced by McAllister and Ianelli (1997). This statistic evaluates the amount of uncertainty in an estimated composition compared to an observed composition and is given by:

where is the estimated proportion for category-*c* (which can be age or any other arbitrary category across which proportions are computed) and is the observed proportion. Here, the underlying age composition derived from the historical bottom trawl surveys with the full and unsampled data was treated as the observed proportions in equation (1). For each iteration of the bootstrap-simulation we computed an estimated proportion () that was then compared to the underlying historical age composition () to determine the effective sample size of the resampled age composition. Thus, across each iteration of the bootstrap-simulation we computed an effective sample size that indicated the amount of uncertainty in the resampled age composition.

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, and 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’ (ISS) we are referring to the harmonic mean of the iterated effective sample sizes. While we present the results of the annual ISS for each stock evaluated when incorporating ageing error, growth variability, or both, we also compute the proportion of ‘base’ ISS in order to present the relative decrease in ISS when incorporating these sources of additional uncertainty. The proportion of base ISS is computed by dividing the ISS as determined after incorporating ageing error, growth variability, or both, by the base ISS without these sources of uncertainty.

## Evaluating life-history relationships to consequences of added uncertainty

In order to quantify trends across stocks we evaluated these stocks based on their species type. Results for three species types are shown: for flatfish (arrowtooth flounder, flathead sole, northern rock sole, and yellowfin sole), for gadids (walleye pollock and Pacific cod), and for rockfish (Pacific ocean perch and northern rockfish). We also investigated the relationship between two statistics and the cases that added ageing error and growth variability. The first statistic we used to evaluate the influence of adding ageing error was the average coefficient of variation (CV) in age agreement. We computed this by computing the CV in the tester ages for a given reader age, then averaged these CVs across the reader ages. The second statistic we used to evaluate the influence of adding growth variability was the average CV in age-length data. We computed the average age-length CV by computing the CV in length for a given age, then averaging the CVs in length-at-age across age.

# Results

While the magnitude of age composition ISS was stock and region specific, the general result that was consistent across the stocks evaluated was a reduction in age composition ISS as additional sources of uncertainty were introduced in the bootstrap procedure (Figure 2). This reduction in age composition ISS resulted for both sex-specific and total (combined sex) age composition ISS. The relative magnitude of adding ageing error compared to growth variability was also stock and region specific. For example, adding ageing error to arrowtooth flounder age data resulted in smaller ISS than adding growth variability in the GOA, but larger ISS in the EBS. Variability in the age composition ISS across uncertainty scenarios evaluated was also stock and region specific, where Pacific cod and walleye pollock in the EBS resulted in the largest variability in ISS and AI Atka mackerel resulted in the smallest.

The proportional reduction in age composition ISS compared to the base case across uncertainty scenarios revealed patterns across species types within each region evaluated (Figure 3). In general, including ageing error in rockfish age data had a larger proportional reduction in age composition ISS than including growth variability. Alternatively, adding growth variability to flatfish and gadid data had a larger proportional reduction in age composition ISS than adding ageing error. Across the regions and uncertainty scenarios investigated, the largest variability in the proportional reduction in age composition ISS in any given region resulted for flatfish. Across the regions and uncertainty scenarios investigated, the largest range among the reduction in region-specific age composition ISS when adding additional sources of error resulted for gadids. For example, after adding both ageing error and growth variability the proportional reduction in age composition ISS was 85-95% of base ISS for gadids in the AI, 40-50% for gadids in the EBS, and 70-80% for gadids in the GOA. Overall, when adding both ageing error and growth variability across the regions evaluated the age composition ISS resulted in 50-90% of the base age composition ISS for flatfish, 40-90% for gadids, and 80-95% for rockfish.

For each of the uncertainty scenarios investigated the relationship between age composition ISS and the number of sampled hauls was not one-to-one for any of the species type evaluated (Figure 4). It was also the case that across the uncertainty scenarios investigated there was large variability in the age composition ISS compared to the number of sampled hauls for each species type evaluated, and any trend was difficult to distinguish. While an increasing trend resulted for rockfish and flatfish between age composition ISS and the number of sample hauls when fitting a linear model to the data, it was generally the case that the age composition ISS was larger than the number of sampled hauls for hauls less than 100-200, and smaller than the number of hauls for hauls larger than 100-200. This increasing trend for rockfish and flatfish degraded as additional sources of uncertainty were added, but for each uncertainty case the trend was not one-to-one. In each of the uncertainty scenarios investigated the trend between age composition ISS and the number of hauls for gadids was different than that for rockfish and flatfish, and while there was a slight increase in age composition ISS as the number of hauls increased, it was not as large of an increase as that which resulted for rockfish and flatfish.

For each species type there was a generally increasing trend that resulted between the number of age samples taken per haul and the age composition ISS per sampled haul (top panel of Figure 5 *i haven’t yet figured out how to match the font in the shared y-axis title with the rest of the figure*). However, as sources of uncertainty were added to age data this increasing trend was dampened for each species type; this was particularly true for gadids as compared to flatfish and rockfish. Across species types the age composition ISS per haul was around half of the number of sampled ages per haul, less for rockfish and gadids. The age composition ISS per haul was, on average, the largest for flatfish, followed by rockfish, and was the smallest for gadids (bottom panel of Figure 5). The variability in the age composition ISS per haul across years, regions, and sex categories was also the largest for flatfish compared to gadids and rockfish.

Comparing between statistics for ageing error and growth variability resulted in different trends in the proportional reduction of age composition ISS upon comparison across species types (top panels of Figure 6). The proportional reduction in age composition ISS when adding ageing error was similar when compared to the average reader-tester CV between gadids and rockfish, where the variability was smaller for rockfish (top left panel of Figure 6). While the range in average reader-tester CV was smaller for rockfish and gadids than for flatfish, the resulting range in proportional reduction in age composition ISS for gadids was larger than rockfish and flatfish. The proportional reductions in age composition ISS when adding growth variability resulted in the smallest range in both ISS reduction and average age-length CV for rockfish, but had a large range and was quite similar between flatfish and gadids (top right panel of Figure 6). On average, the proportional reduction in age composition ISS when adding both ageing error and growth variability was smallest for rockfish, intermediate for flatfish, and largest for gadids (bottom panel of Figure 6). The same trend resulted when evaluating the range in the proportional reduction in age composition ISS when both ageing error and growth variability were added.

# Discussion

In this study we found that adding sources of uncertainty that account for ageing error and growth variability in bootstrap procedures decreased age composition ISS for all stocks included in this analysis. The impact of the sources of uncertainty on resulting ISS was species type specific, where ageing error was more influential for rockfish than growth variability, and growth variability was more infleuntial than ageing error for flatfish and gadids. We propose that these results are due to larger growth variability found in gadids and flatfish compared to rockfish. Further, the effects of ageing error are not surprising for rockfish, since they are so long-lived, but the average CV for reader-tester agreement was much larger for some flatfish stocks as compared to the rockfish stocks evaluated, thus, there was some interaction between the number of age bins and ISS in this analysis. Upon adding both sources of uncertainty, generally the largest effect on reduction in ISS magnitude resulted for gadids, was intermediate for flatfish, and had the least effect on rockfish, but was stock and region specific.

Understanding effects of survey reduction effort has recently been the focus of a number of studies (ICES 2020, 2023), as survey reductions may be inevitable in many regions. A recent study investigated the reductions in length frequency and age collection effort, using AFSC bottom trawl survey as example (Hulson et al. in review) and found that reduction in age collections had larger effect on flatfish and rockfish as compared to gadids. Here we find that including additional sources of uncertainty has largest effect on gadids, and smallest on rockfish, and it is potentially the case that the effect of decreases in sampling effort for gadids and flatfish would be smaller given the magnitude of effect by these sources of uncertainty as compared to rockfish. However, we acknowledge that this should be evaluated in future studies to understand the specific effects on stocks and when comparing among species types.

It is well known that misspecification of ISS when fitting compositional data can lead to bias results in assessment model predictions (e.g., Stewart and Monnahan 2017, Xu et al. 2020). Here we show for gadids that there is upwards of a 50% decrease in ISS for some stocks when adding additional uncertainty as compared to the base case, an upwards of 40% decrease for flatfish, and an upwards of 20% decrease for rockfish. Without these additional sources of uncertainty taken into account, using the bootstrap procedure would result in ISS that are larger than what they should be. While we did not investigate implications to specific assessment model outcomes, it can be inferred that reductions of ISS on this scale would have downstream effects on model predictions and the associated uncertainty. We note that these sources of uncertainty would not be contained only to fishery-independent sources, like evaluated here, but would also be inherent to age collections for fishery-dependent sources as well. Thus, future investigations into the impacts of adding ageing error and growth variability into ISS estimation on assessment model results should also include fishery-dependent ISS implications as well.

A number of operational assessment models use hauls as a proxy in some form when setting ISS for age and length composition data Hulson et al. (2021). This is derived from a result found in Pennington et al. (2002) who investigated length frequency sampling and, based on the level of intra-haul correlation, determined that for the species that were investigated, on average, the ISS was one fish per haul However, we note that the conclusion made in Pennington et al. (2002) was not that the number of hauls should be used as a proxy for ISS in the assessment model data fitting procedure, but rather, that in order to potentially increase the ISS and have a better estimate of the level of intra-haul correlation that samples should be taken from an increased number of hauls. This result was further substantiated in Siskey et al. (2023), who also suggested that increasing the number of hauls while decreasing the sampling effort on any given haul was preferable to increasing the sampling effort within any given haul at the expense of sampling from a greater number of hauls.

In this study, we find that the relationship between age composition ISS and the number of hauls to be extremely weak, if non-existent. While the magnitude of ISS and the number of sampled hauls is on a similar relative magnitude scale, there was not a one-to-one relationship that resulted for any of the stocks or species types evaluated. Further, we found that there was extreme variability when comparing between the number of sampled hauls and the resulting bootstrapped ISS for all of the uncertainty scenarios investigated here, and only became worse as the additional uncertainty sources were added to the bootstrap. Compared to the bootstrap results, it was found that for all the species types evaluated, that if hauls were used as a proxy for ISS in an assessment model the ISS would be an underestimate for sampled hauls less than around 100, and an overestimate for hauls greater than around 100. It is the case that, based on the results of Pennington et al. (2002), that sampling effort at AFSC has transitioned to sampling an increased number of hauls rather than increasing sampling in any given haul, thus, in recent time-series using hauls as a proxy for ISS is likely an over-estimate of ISS as compared to the bootstrap results.

There are several advantages to using the bootstrap procedure introduced here, that follows from Stewart and Hamel (2014), to determine ISS for composition data used in stock assessment models. The primary advantage is that this procedure accounts for the sampling design employed to collect data, and thus, implicitly includes the inter annual heterogeneity in sampling effort. One could argue that the number of sampled hauls can be used as a proxy for this heterogeneity in sampling effort, and we would agree, but also add that the number of hauls itself does not account for the inter annual heterogeneity in intra-haul correlation, which can be large and based on the changes in population demographics (i.e., large recruitment events or changes in composition structure from year to year). Thus, we recommend that the use of hauls as a proxy for ISS be discontinued when using stock assessment models to fit age and length composition data, and rather, assessment scientists consider implementing a bootstrap procedure as shown here to use for setting ISS. As suggested in Stewart and Hamel (2014), we agree that the bootstrap ISS can be used in a stock assessment model as the ‘maximum’ sample size.

Previous work has investigated both the inclusion of ageing error (Punt et al. 2008, Liao et al. 2013) and growth (Taylor and Methot 2013) within stock assessment models, however, none have applied these additional sources of uncertainty in the context of estimating ISS. As it pertains to ageing error, in many current assessments an ageing error matrix is implemented (e.g., Williams et al. 2022) in order to account for ageing error that is inherent to the age composition data. Use of an ageing error matrix effectively ‘smudges’ assessment model estimates of population-at-age into adjacent age classes prior to fitting the ‘observed’ age composition input data from either fishery-independent or fishery-dependent sources. Then, in the model fitting step, an ISS is used to fit the models ‘smudged’ estimates of age composition to the ‘observed’ age composition. Here we argue that unless ageing error is accounted for in the age composition ISS used to fit these data then we are likely ‘over-fitting’ the model estimates to the ‘observed’ age composition and that the use of only an ageing error matrix half-way accounts for this source of uncertainty in the assessment modeling process. The results from this study suggest that in some cases the addition of ageing error to the bootstrap method decreases the age composition ISS by 20-40%, thus, an ISS that does not take into account this source of uncertainty can be 125-165% too large.

It is commonly the case that age composition is produced through a two-stage expansion process (Quinn and Deriso 1999), in which length frequency data is expanded to population-at-length in the first stage, and an ALK is used to expand population-at-length to population-at-age in the second stage (Ailloud and Hoenig 2019). In the second-stage of this expansion process, the ALK is produced through the use of age-length paired data that are obtained in the age sampling collection, and within the ALK the variability in length-at-age is implicitly accounted for. However, there have been no previous attempts to include this source of variability when considering the ISS that is used to fit the expanded age composition data. We find that the magnitude of effect on age composition ISS is species type dependent, in which a larger effect results for species that exhibit larger variability in growth, for example, gadids and flatfish compared to rockfish. When implementing growth variability there was a 10-50% decrease in the magnitude of bootstrapped age composition ISS, thus, if not taking this source of uncertainty into account the ISS can be up to 200% too large.

Overall, we find that including ageing error and growth variability into estimation of age composition ISS can have large effect in reducing the magnitude of ISS. We provide three primary recommendations from this work. First, we recommend that the use of hauls as a proxy for age or length composition ISS in stock assessment models be discontinued. While we found that the scale of age composition ISS and hauls were similar, we did not find any relationship between the bootstrap ISS and hauls for any species that would suggest using hauls as a proxy, rather, we found that using hauls will likely lead to either an over- or under-estimation of ISS, with the former potentially having greater implications to model results (Francis 2011). Second, we recommend that stock assessment scientists consider the use of bootstrap methods like this one (which was developed from Stewart and Monnahan 2017) to set and and length composition ISS. With modern computing power, for a single species using the package we built (<https://github.com/BenWilliams-NOAA/surveyISS>) it takes on the matter of a couple of hours to obtain both age and length composition bootstrap ISS for a historical survey time series (longer than 40 years in some cases); for a single year it takes a matter of minutes. Third, for all estimates of age composition ISS we recommend implementing ageing error and growth variability to more explicitly and thoroughly take these sources of uncertainty into account in stock assessment models. We note that while we used fishery-independent data here as an example, that these sources of uncertainty would also be inherent to fishery-dependent data.

# Acknowledgments

We thank *reviewer1* and *reviewer2* for their helpful reviews of this manuscript. We also thank all the AFSC survey staff who collected the data over the last 40 years used in this analysis.

# 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.

Barbeaux, S.J., Barnett, L., Connor, J., Nielson, J., Shotwell, S.K., Siddon, E., and Spies, I. 2022. Assessment of the Pacific cod stock in the Eastern Bering Sea. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea and Aleutian Islands. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

Candy, S.G., Nowara, G.B., Welsford, D., and McKinlay, J.P. 2012. Estimating an ageing error matrix for patagonian toothfish (*dissostichus eleginoides*) otoliths using between-reader integer errors, readability scores, and continuation ratio models. Fisheries Research 115: 14–23.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences 68: 1124–1138.

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

Henriquez, V., Licandeo, R., Cubillos, L.A., and Cox, S.P. 2016. Interactions between ageing error and selectivity in statistical catch-at-age models: Simulations and implications for assessment of the chilean patagonian toothfish fishery. ICES Journal of Marine Science 73(4): 1074–1090.

Hulson, P.-J.F., Barbeaux, S.J., Ferriss, B., McDermott, S., and Spies, I. 2022. Assessment of the Pacific cod stock in the Gulf of Alaska. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

Hulson, P.-J.F., Williams, B., Bryan, M., Conner, J., and Siskey, M. in review. Reductions in sampling effort for fishery-independent age and length composition: balancing stock assessment input data uncertainty and workforce health and efficiency.

Hulson, P.-J.F., Williams, B.C., Fissel, B.E., Ferriss, B.E., Hall, M., Yasumiishi, E.M., and Jones, D.T. 2021. Assessment of the Pacific ocean perch stock in the Gulf of Alaska. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

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.

ICES. 2020. Workshop on unavoidable survey effort reduction (WKUSER). ICES Scientific Reports.

ICES. 2023. Workshop on unavoidable survey effort reduction 2 (WKUSER). ICES Scientific Reports.

Kimura, D.K., and Lyons, J.J. 1991. Between-reader bias and variability in the age-determination process. Fishery Bulletin 89: 53–60.

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.

Lee, H., Piner, K.R., Taylor, I.G., and Kitakado, T. 2019. On the use of conditional age at length data as a likelihood component in integrated population dynamics models. Fisheries Research 216: 204–211.

Liao, H., Sharov, A.F., Jones, C.M., and Nelson, G.A. 2013. Quantifying the effects of aging bias in atlantic striped bass stock assessment. Transactions of the American Fisheries Society 142(1): 193–207.

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.

McGilliard, C.R., Palsson, W., Havron, A., and Zador, S. 2019. Assessment of the Deepwater Flatfish stock complex in the Gulf of Alaska. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

Monnahan, C.C., Dorn, M.W., Deary, A.L., Ferriss, B.E., Fissel, B.E., Honkalehto, T., Jones, D.T., Levine, M., Rogers, L., Shotwell, S.K., Tyrell, A., and Zador, S. 2021. Assessment of the walleye pollock stock in the Gulf of Alaska. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

Morison, A., Burnett, J., McCurdy, W., and Moksness, E. 2005. Quality issues in the use of otoliths for fish age estimation. Marine and Freshwater Research 56.

Nesslage, G., Schueller, A.M., Rezek, A.R., and Mroch III, R.M. 2022. Influence of sample size and number of age classes on characterization of ageing error in paired-age comparisons. Fisheries Research 249: 106255.

Pennington, M., Burmeister, L.M., and Hjellvik, V. 2002. Assessing the precision of frequency distributions estimated from trawl-survey samples. Fishery Bulletin 100(1): 74–80.

Punt, A.E., Smith, D.C., KrusicGolub, K., and Robertson, S. 2008. Quantifying age-reading error for use in fisheries stock assessments, with application to species in australia’s southern and eastern scalefish and shark fishery. Canadian Journal of Fisheries and Aquatic Sciences 65(9): 1991–2005.

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.

Spencer, P.D., and Ianelli, J.I. 2022. Assessment of the Pacific ocean perch stock in the Bering Sea/Aleutian Islands. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea and Aleutian Islands. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

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.

Stewart, I.J., and Monnahan, C.C. 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. Fisheries Research 192: 126–134.

Taylor, I.G., and Methot, R.D. 2013. Hiding or dead? A computationally efficient model of selective fisheries mortality. Fisheries Research 142: 75–85.

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.

Williams, B.C., Hulson, P.-J.F., Lunsford, C.R., and Ferriss, B. 2022. Assessment of the Northern rockfish stock in the Gulf of Alaska. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. North Pacific Fishery Management Council, 1007 West 3rd Ave., Suite 400, L92 Building, 4th floor, Anchorage, AK 99501.

Xu, H., Thorson, J.T., and Methot, R.D. 2020. Comparing the performance of three data-weighting methods when allowing for time-varying selectivity. Canadian Journal of Fisheries and Aquatic Sciences 77(2): 247–263.

# Tables

Table 1: Average age samples from the AFSC bottom trawl surveys by region (rounded to the nearest 10), and total reader-tester age pairs (rounded to the nearest 100) for the species evaluated in the bootstrap-simulation.

| Stock | Scientific name | AI | EBS | GOA | R-T |
| --- | --- | --- | --- | --- | --- |
| arrowtooth flounder | Atheresthes stomias | 450 | 480 | 850 | 6,100 |
| Atka mackerel | Pleurogrammus monopterygius | 600 | – | – | 12,900 |
| flathead sole | Hippoglossoides elassodon | – | 560 | 520 | 9,400 |
| northern rock sole | Lepidopsetta polyxystra | – | 460 | 450 | 8,900 |
| northern rockfish | Sebastes polyspinis | 570 | – | 450 | 6,400 |
| Pacific cod | Gadus macrocephalus | 800 | 1070 | 650 | 21,200 |
| Pacific ocean perch | Sebastes alutus | 940 | – | 1030 | 13,500 |
| walleye pollock | Gadus chalcogrammus | 790 | 1500 | 1300 | 84,400 |
| yellowfin sole | Limanda aspera | – | 750 | – | 10,300 |

# Figures



Figure 1: Bootstrap-simulation flow chart, the steps refer to the order of operations.

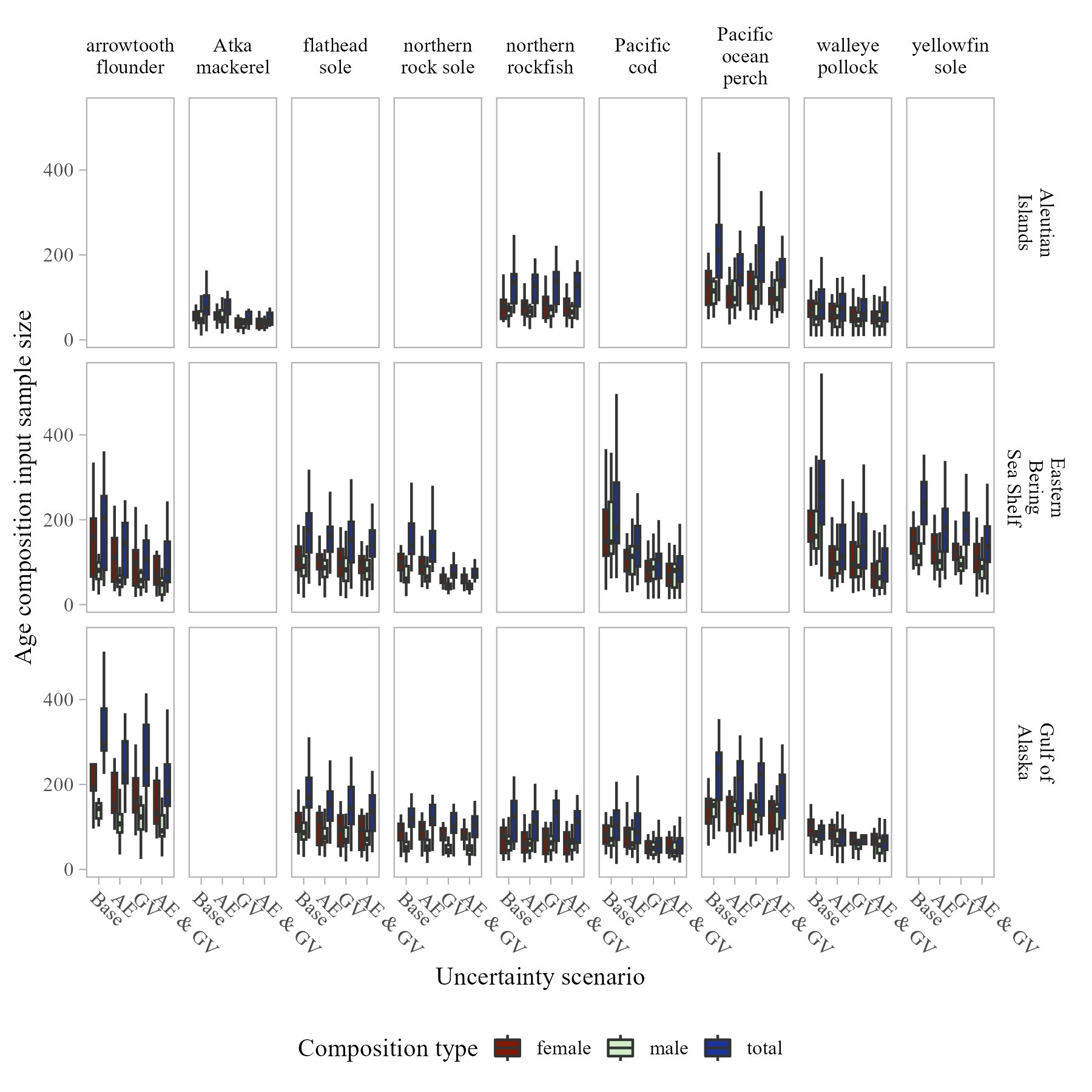


Figure 2: Age composition input sample size for the stocks and regions evaluated across uncertainty scenarios.

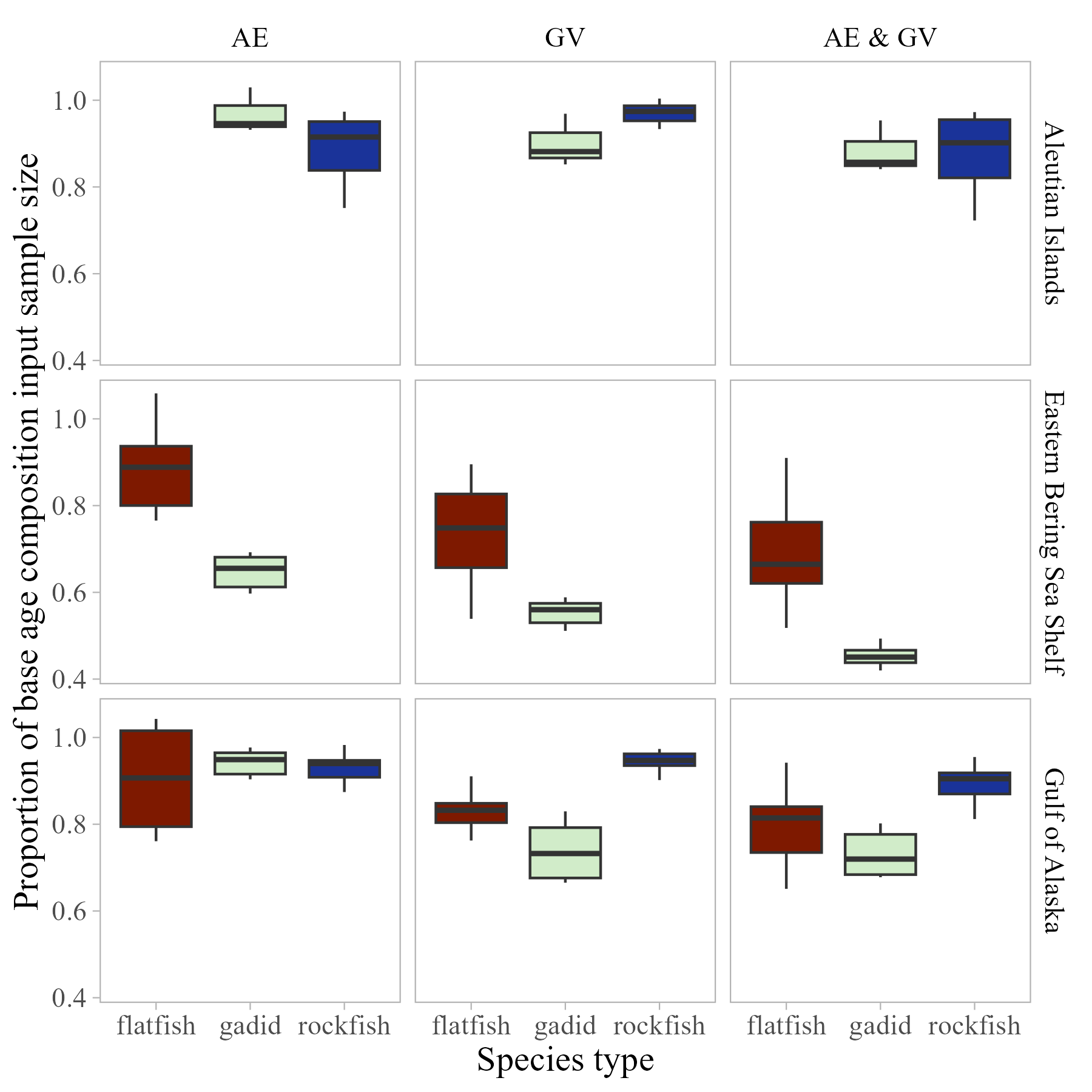


Figure 3: Proportion of base case age composition input sample size by species type across sex categories and uncertainty scenarios.

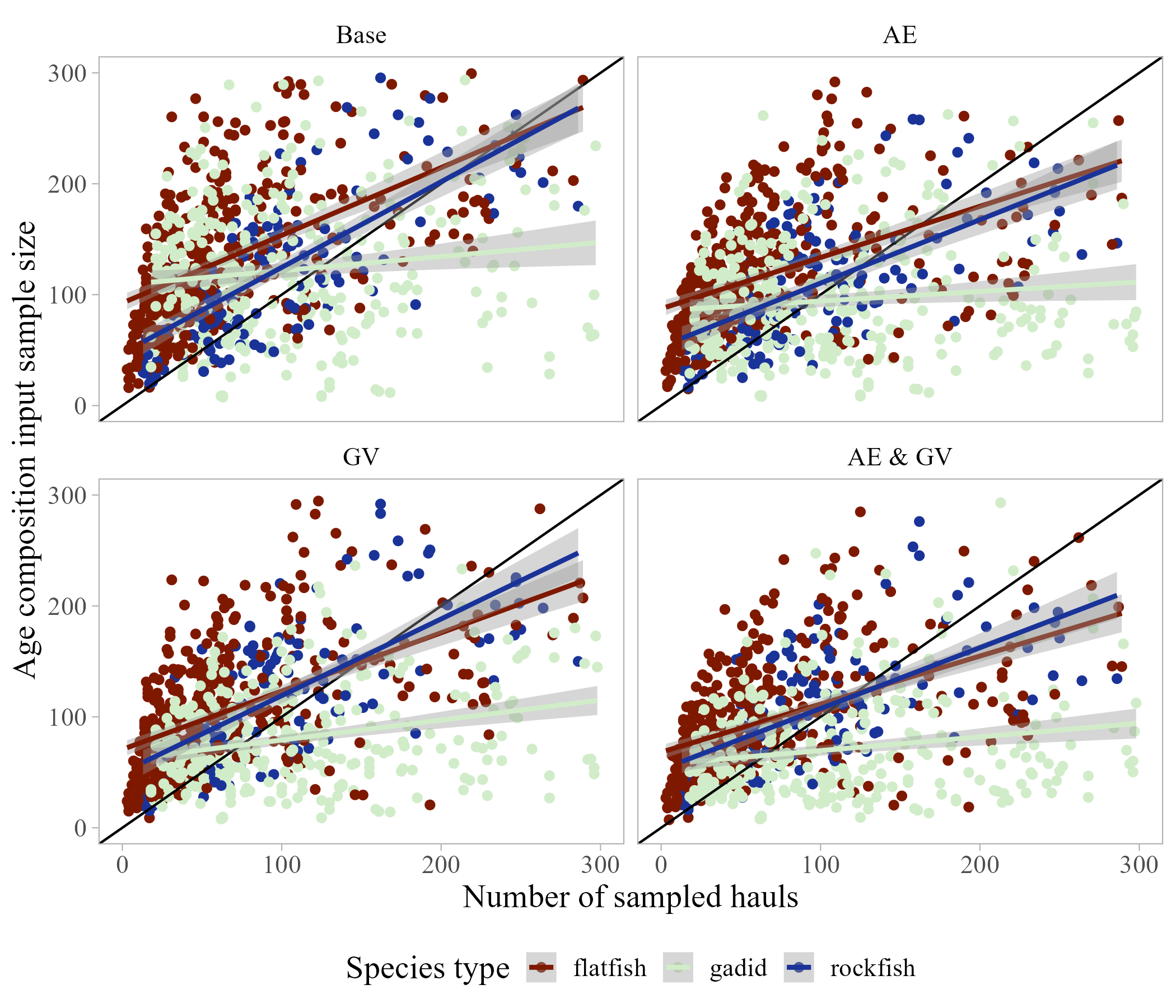


Figure 4: Age composition input sample size compared to number of sampled hauls across uncertainty scenarios and species types (linear fit by species type shown for illustration).

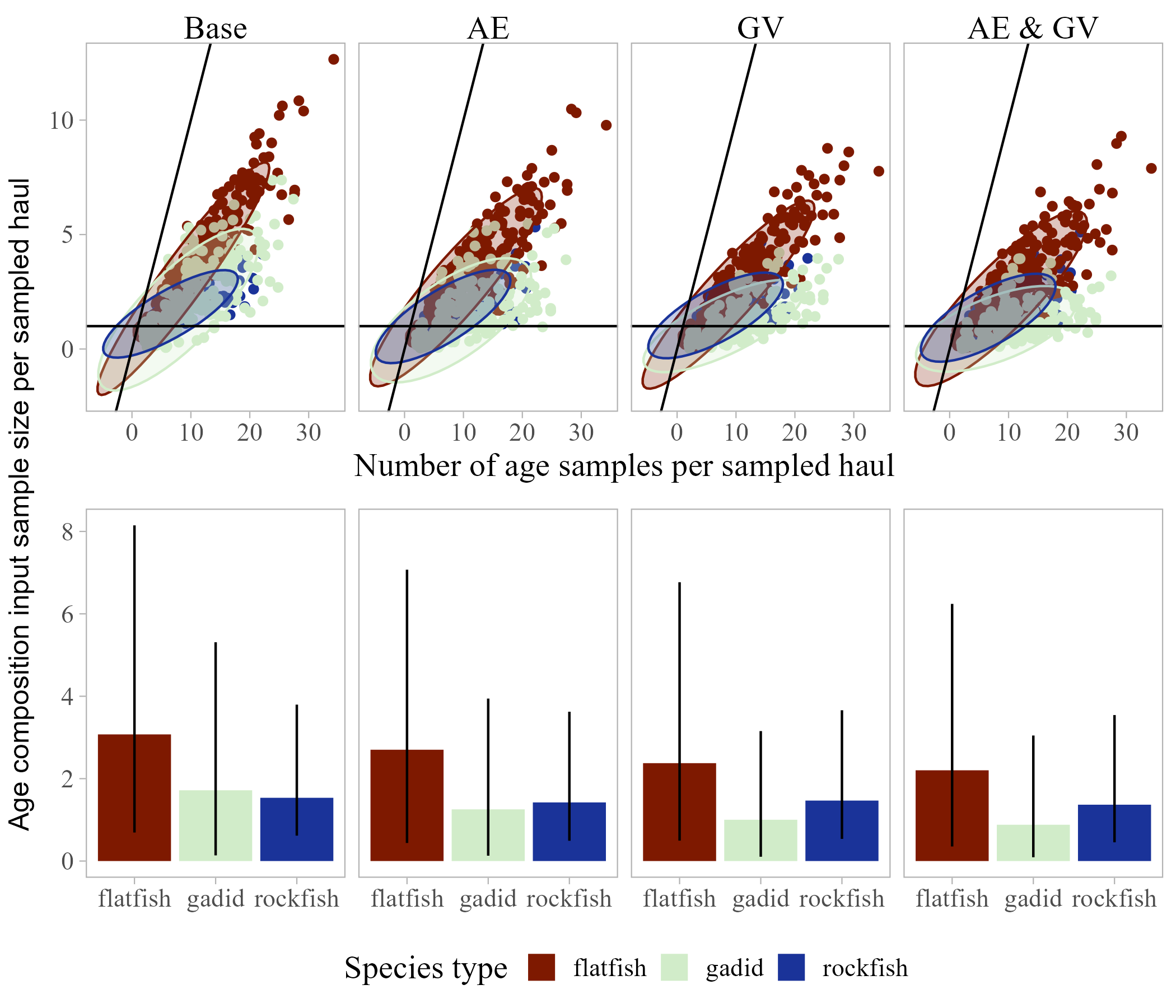


Figure 5: Age composition input sample size per sampled haul compared to number ages sampled within a haul (top panel) across uncertainty scenarios and species types (bottom panel).

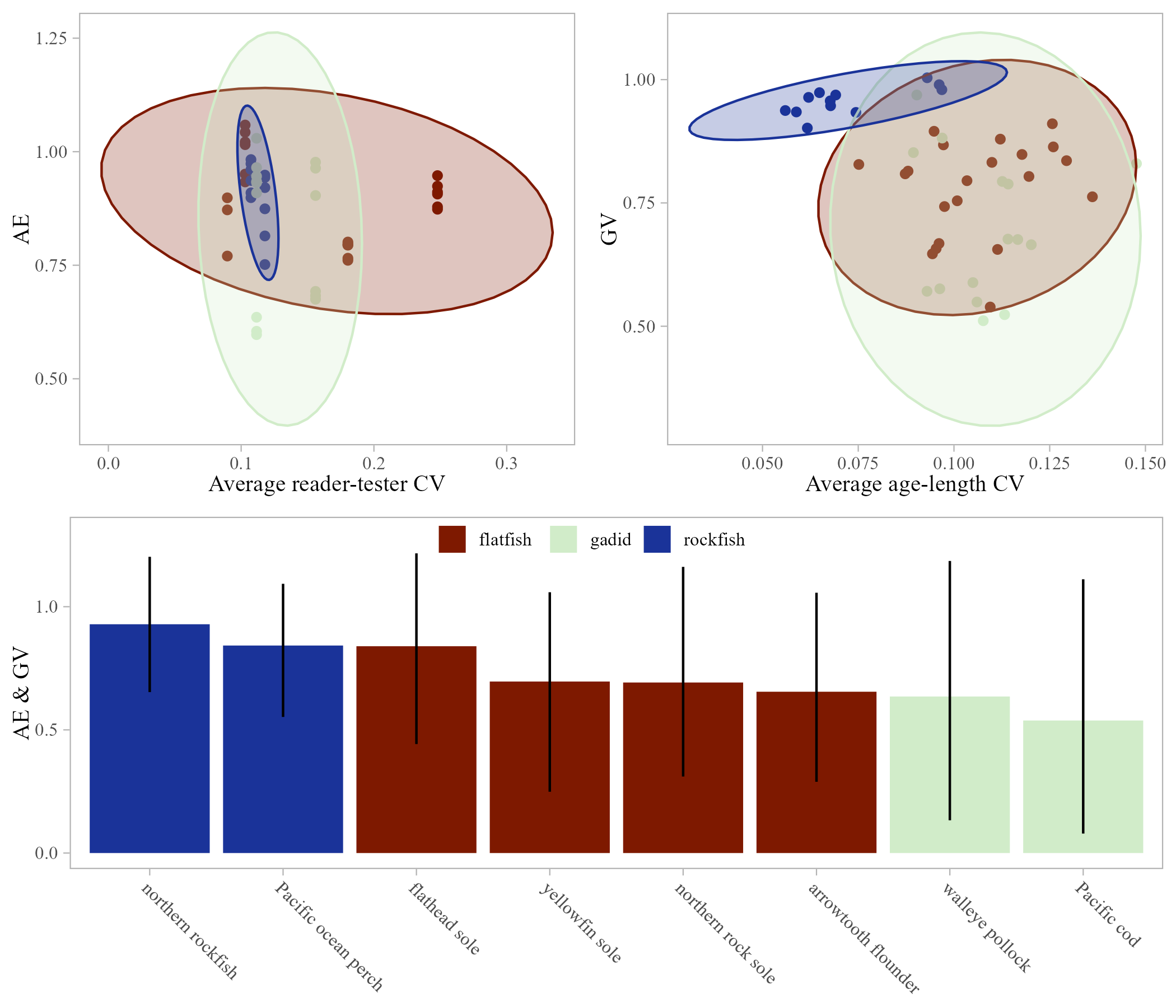


Figure 6: Proportional reduction in age composition input sample size compared to age and growth statistics (top panel) and across stocks evaluated (bottom panel).