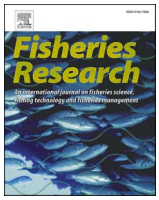
Fisheries Research 255 (2022) 106452

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres



Focusing on the front end: A framework for incorporating uncertainty in biological parameters in model ensembles of integrated stock assessments

Nicholas D. Ducharme-Barth [a,b,*,1], Matthew T. Vincent [c,1]

[a] Pacific Community, 95 Promenade Roger Laroque, B.P. D5, 98848 Noumea, New Caledonia
[b] NOAA National Marine Fisheries Service, Pacific Islands Fisheries Science Center, 1845 Wasp Boulevard, Building 176, Honolulu, HI 96818, USA
[c] NOAA National Marine Fisheries Service, Southeast Fisheries Science Center, Beaufort Lab, 101 Pivers Island Rd, Beaufort, NC, USA

ARTICLE INFO

Handled by: A.E. Punt

Keywords:
Stock assessment
Model ensemble
Uncertainty
MULTIFAN-CL
Swordfish

ABSTRACT

Uncertainty in population status estimates from stock assessments is important for providing a comprehensive picture of current knowledge of a stock. The use of model ensembles to encapsulate model uncertainty has become increasingly prevalent. The uncertainty of biological parameters that are often fixed in stock assessment models can be quantified for presentation of management advice through model ensembles. An ensemble can be created by randomly drawing values from the likely parameter space using a Monte-Carlo/bootstrap (MCB ensemble) or fixed at either a high, medium, or low value that encapsulates the variability in the parameter and applied in a full factorial grid across the fixed parameters (factorial ensemble). We calculated the management advice from MCB ensembles of various sizes and a 243 model factorial ensemble for Southwest Pacific swordfish (*Xiphias gladius*) and compared reference points which included model uncertainty only, model and estimation uncertainty, or both uncertainties weighted by sampling importance resampling. Median reference points were significantly different between the two ensemble types with the factorial ensemble having a significantly larger estimate of model uncertainty than the MCB ensemble. Stock assessments with fixed biological parameters can characterize uncertainty in these parameters more efficiently using a MCB ensemble approach. A factorial ensemble approach is appropriate for comparing different model structure assumptions and functional forms of relationships and can be used in combination with a MCB ensemble approach. Incorporation of both model and estimation uncertainty in estimates of reference points is important when providing management advice because including only model uncertainty can lead to biased estimates of the precision of reference points. Further work is needed regarding appropriate weighting of ensembles which incorporate different data sources or have different likelihood weightings.

1. Introduction

Modern management of exploited fisheries relies on estimates of historical trends in population biomass and fishing mortality or reference points of these quantities. This stock status information is then used by managers to set appropriate limits and targets that are used to determine regulations on harvest. The most frequently used stock assessment approach to estimate stock status is the integrated, statistical catch-at-age model (Fournier and Archibald, 1982; Deriso and Quinn, 1985; Methot and Wetzel, 2013; Fournier et al., 1998). The complexity of these models has evolved and generally increased over time (Hilborn, 2003); recent catch at age models include sex-specific dynamics and/or spatially discrete areas with multiple stocks (Berger et al., 2017; Maunder and Piner, 2017). Hundreds to thousands of model parameters are necessary in order to meet the parametric structure of these complex integrated stock assessment models. In many instances data are insufficient to internally estimate all parameters simultaneously, so a subset are held fixed during the analysis. Fixing parameters in an integrated assessment model makes a strong assumption about the uncertainty (zero) associated with that particular parameter. However, small changes in fixed biological parameters can result in large differences in estimates of stock status (Minte-Vera et al., 2017). Characterization and

* Corresponding author at: NOAA National Marine Fisheries Service, Pacific Islands Fisheries Science Center, 1845 Wasp Boulevard, Building 176, Honolulu, HI 96818, USA.
E-mail address: nicholas.ducharme-barth@noaa.gov (N.D. Ducharme-Barth).
[1] Note: Both authors contributed equally to this study and share first authorship.

https://doi.org/10.1016/j.fishres.2022.106452
Received 29 November 2021; Received in revised form 26 July 2022; Accepted 27 July 2022
Available online 16 August 2022

quantification of uncertainty for presentation of management advice is becoming more widespread as awareness of the magnitude of this uncertainty in stock assessments increases (Privitera-Johnson and Punt, 2020).

There are two main types of uncertainty that afflict fisheries management: scientific uncertainty and management uncertainty (Privitera-Johnson and Punt, 2020). This study focuses on the former, while the latter can be addressed through management strategy evaluation (MSE; Punt et al., 2016). Scientific uncertainty due to imprecision and bias in the stock assessment process can be further subdivided into four categories. First, observation uncertainty is the measurement error in the observed quantities such as catch, length, weight, age estimates, and catch-per-unit effort (CPUE). Second, process uncertainty is variability in underlying stock dynamics such as stochasticity in recruitment or growth of fish. To a certain extent both observation and process uncertainty influence the third type of scientific uncertainty, estimation uncertainty. Estimation uncertainty arises due to the imprecision or bias in parameters estimated within the model. Some refer to estimation uncertainty as parameter uncertainty but this creates ambiguity between model and estimation uncertainties (e.g., multiple models that assume different fixed constant values of the natural mortality parameter is model uncertainty). Therefore, we prefer the use of estimation uncertainty and advocate for not using the term parameter uncertainty. Fourth, model uncertainty is the uncertainty or misspecification of fixed model parameters or functional forms of assumed dynamics. Examples of model uncertainty include biological assumptions such as the form of the spawner-recruit relationship or somatic growth curves, fisheries assumptions such as functional forms of selectivity or number of fisheries, and modeling assumptions such as different spatial structures or sex-specific dynamics.

Reference points of stock status provided for management advice are model predictions, which are directly affected by model and estimation uncertainty. The quantification of the uncertainty in reference points is necessary for understanding the risk of various management actions. Historically, point estimates of stock status from a single model were used to provide management advice and did not quantify uncertainty in the estimates (Haddon, 2001; Quinn and Deriso, 1999). Calculations of estimation uncertainty for presentation of management advice first occurred as a result of greater computational abilities to estimate variances of model quantities using the covariance matrix and the delta method (Fournier et al., 2012; Magnusson et al., 2012; Maunder and Piner, 2015). Monte-Carlo/bootstrap simulations have also been used to estimate uncertainty in estimated quantities for use in management (Restrepo et al., 1992; Legault et al., 2002). However, estimation uncertainty from a single model is now generally thought to be modest compared to model uncertainty representing different states of nature (Scott et al., 2016). Use of ensemble model methods and superensembles (Anderson et al., 2017) has led to the expansion of quantifying uncertainty from multiple models in recent years to present to managers (Brodziak and Piner, 2010; Stewart and Martell, 2014; Scott et al., 2016; Jardim et al., 2021). These ensemble methods can more truthfully capture the broader uncertainty from numerous models representing different states of nature and lead to greater stability in estimates (Stewart and Martell, 2015; Stewart and Hicks, 2018). This study will demonstrate how model uncertainty in fixed parameters within stock assessments can be integrated into a model ensemble to provide a better estimate of uncertainty in reference points.

It is important to consider how model ensembles are combined to provide management advice because the chosen approach can influence both the point estimate (e.g., median) and the estimated uncertainty in stock status. The simplest approach is to assume that all models are equally likely and thus all alternative states of nature have the same probability of being true. The other alternative is to combine models according to a weighting scheme where the derivation of model weights can come from a subjective (based on expert opinion), objective (based on model convergence or other diagnostics), or hybrid approach (Maunder et al., 2020). Equally important to how management advice is presented and combined is the choice of models included within an ensemble (Scott et al., 2016; Maunder et al., 2020; Brodziak and Piner, 2010). However, a research gap exists as guidance regarding which models to include or exclude has generally been left up to individual analysts to decide. Additionally, explicit examples of including inappropriate models within an ensemble and the resulting impact on management advice have not been evaluated. Previous applications of model ensembles used to provide management advice have generally used one of two different approaches to encapsulate uncertainty in biological parameters (e.g., natural mortality or growth) that are fixed within the assessment models.

The first approach to incorporate biological uncertainty into an ensemble is to take a Monte-Carlo/bootstrap (MCB) approach where randomly drawn values of biological parameters, taken from distributions obtained by external analyses, are used to parameterize each model in the ensemble (Restrepo et al., 1992; Legault et al., 2002; Scott et al., 2016; Nadon, 2017; SEDAR, 2021). This approach is similar to creating prior distributions for parameters within a Bayesian framework. However, values drawn from distributions using an MCB approach are then fixed within the stock assessment. Despite recent advances in algorithms for mapping the posterior distributions (STAN Development Team, 2021; Monnahan et al., 2019, 2016; Monnahan and Kristensen, 2018), Bayesian analyses remain computationally impractical for use in complex age-structured stock assessments due to extremely long run times. However, this should not prevent the creation of "prior-like" joint parameter distributions (univariate or multivariate) that can incorporate biological model uncertainty into model ensembles that are fit to data using maximum likelihood approaches. These parameter distributions can be formulated using a range of approaches; the simplest approach would use the estimate of uncertainty from a single study (e.g., growth curve estimate), whereas a more complex approach would be a meta-analytic approach of numerous studies on similar species (Horswill et al., 2019) such as the one implemented in the R package FishLife (Thorson et al., 2017; Thorson, 2020). The second approach to incorporate model uncertainty into an ensemble involves bounding the uncertainty of a fixed parameter, typically using the associated 95% confidence (credible) interval. The high and low estimates of a parameter would be combined with the point estimate of the analysis to represent the uncertainty in the fixed parameter. Uncertainty in biological parameters is quantified and presented in management advice by combining with other model uncertainties in a full factorial combination of "axes of uncertainty".

Application of model ensembles has commonly been used by the Pacific Community (SPC) for assessments conducted for the Western Central Pacific Fisheries Commission (WCPFC) as well as by the International Pacific Halibut Commission (IPHC) to formulate management advice (Takeuchi et al., 2017; WCPFC-SC, 2017; Stewart and Hicks, 2022). Specifically, as it relates to assessments produced by SPC for the WCPFC, assessments implementing the full factorial ensemble approach have been produced under a constrained timeline. Input data are finalized in mid-May, though sometimes as late as early July, with the assessment reports due to the WCPFC scientific committee by the end of July. Model run times for *MULTIFAN-CL* (Fournier et al., 1998) vary by spatial model complexity and species data from 30 min to several hours, with average run times of about 8–12 h for the spatially structured tropical tuna assessment models. One to three assessments are conducted at the same time each year on a limited number of computational cores. Though recent advancements have been made to MULTIFAN-CL to reduce model run times, there are computational limits to the size of model ensembles that can be completed within the timeline for assessments. As a result, model ensembles typically contain 3–5 axes of uncertainty with 27–243 models assuming 3 levels (high, median, and low) per "axis". Model results are presented to managers as the probability of current stock status exceeding limit reference points using Kobe and Majuro plots, typically calculated as a weighted proportion of

models in the ensemble. Historically, these stock status estimates were calculated using model uncertainty only and omitted estimation uncertainty. Weights for each axis have previously been determined by expert opinion of the scientific committee and applied to each model. Projections of future dynamics using current management policies are conducted for all models within the ensemble and reported to managers.

A comparison of these two approaches for ensemble construction (full factorial and MCB) has not been conducted on a set of biological parameters. Theoretically, the MCB distribution approach is superior in many aspects. First, the full factorial approach can result in combinations of parameters that would be considered biologically implausible according to life history theory. For example, a high level of natural mortality is unlikely with a lower level of growth capacity (k in the von Bertalanffy growth curve). Conversely, the MCB approach can be constructed in a way that preserves the inherent correlation between parameters and self-censors the ensemble to more likely parameter combinations. The implicit behavior of the MCB parameter distribution will give more weight to the most plausible parameter values, whereas the full factorial approach will result in more weight in the tails compared to a distribution. Finally, the full factorial approach can quickly become computationally impractical to conduct beyond a few axes of uncertainty when models have lengthy run times, such as those run with MULTIFAN-CL. This computational restriction compels the analyst to triage the potential sources of uncertainty, effectively ignoring the impact of those sources of uncertainty deemed less important. In theory, the range of uncertainty from the MCB approach could be characterized in a more computationally efficient manner using a smaller model ensemble depending on the departure from multivariate normality.

In the present study, we attempt to address the apparent research gap by providing guidance on the construction of model ensembles for integrated stock assessment. We provide an explicit example of how model ensemble construction and model ensemble combination relating to biological parameters fixed in an assessment model can impact the resulting management advice and associated scientific uncertainty, specifically model and estimation uncertainty. In this case, model uncertainty is defined by alternative assumptions for fixed parameter values. This is accomplished by addressing the following six objectives using the 2017 southwest Pacific Ocean (SWPO) swordfish (*Xiphias gladius*) stock assessment model as a case study: (i) we demonstrate the difference in management advice arising from creating a model ensemble using the full factorial and MCB distribution approaches; (ii) using the MCB approach, we evaluate the number of models needed to characterize the model uncertainty; (iii) we show how the MCB approach can be used to identify which fixed parameters are most influential in the reference point estimates; (iv) we illustrate the difference in management advice from ensembles that just characterize model uncertainty versus ensembles that characterize both model and estimation uncertainty; (v) we describe how total uncertainty across an ensemble can be partitioned between model and estimation uncertainty; and (vi) lastly we display how model ensemble construction can be combined with an ensemble combination approach (equal weighting vs. sampling importance resampling (SIR) weighting) in calculating management reference points. Results from this study do not constitute management advice. The use of the 2017 SWPO swordfish case study is to illustrate potential differences between methods for constructing a model ensemble and characterizing uncertainty in reference points.

2. Methods

2.1. Case study description

Details of the 2017 SWPO swordfish stock assessment are presented in Takeuchi et al. (2017) and we refer readers to this report for a complete description as it formed the foundation for all models used in this study. For context, the 2017 SWPO swordfish stock assessment was conducted using the integrated assessment platform MULTIFAN-CL based on data from 1952 to 2015. The model is spatially stratified into two regions in the SWPO delineated at 165° E and uses 13 longline fisheries based on sub-area boundaries, nationality, and time period. The assessment employs a size-based (length and weight) statistical catch-at-age with a catch-errors method. Data used in the swordfish assessment for the SWPO consisted of fishery-specific catch (in numbers) and standardized effort data for the Japanese, Chinese Taipei, Australian, and European Union fleets (which provided indices of relative abundance), length-frequency data, and weight-frequency data. The models used in this analysis were identical to the 2017 SWPO diagnostic model in terms of the model structure and input data except for the treatment of the following biological assumptions. We investigated model uncertainty by modifying five different biological assumptions that were fixed within the 2017 stock assessment: growth, natural mortality (M), length-weight relationship, maturity-(or spawning-potential)-at-length relationship, and steepness. Input data and code for replicating the analysis can be found at the following repository: dx.doi.org/10.6084/m9.figshare.16775860.

2.2. Ensemble construction

2.2.1. MCB approach

The methods and data used to create the joint parameter distribution are of limited importance to the conclusions drawn in this study and could be created through a variety of approaches depending on the species. Briefly, we describe the methods and data used in the current analysis to create the joint parameter distribution; however, we urge readers to consult the supporting information S1 Appendix for further information and details. Four independent Bayesian analyses using the STAN probabilistic language, implemented in R (*v*4.0.3) using the *rstan* package (*v*2.21.2) (Core Team, 2021; STAN Development Team, 2021) were used to create posterior distributions for the parameters needed to parametrize the growth, spawning potential, and length-weight relationships. Growth was modeled as a von Bertalanffy growth relationship, spawning potential was modeled as the product of the logistic relationship of maturity-at-length (lower jaw fork length; LJFL) and the logistic relationship of sex-ratio at lower jaw fork length (LJFL), and length-weight was modeled using an exponential relationship. The length-at-age and maturity-at-length data used to estimate these relationships were initially collected from longline sampled swordfish captured in the Coral Sea (Young and Drake, 2002, 2004; Young et al., 2003), though the aging and histological data came from a subsequent re-analysis (Farley et al., 2016). Additionally, length and weight data by sex of longline captured swordfish, taken as a part of the Pacific Islands Regional Observer Program (PIRFO) were also used in the current analysis.

A joint posterior of these 3 relationships was created by randomly drawing 1255 samples without replacement from each of the independent posteriors. These samples were then used to calculate a MCB parameter distribution for the natural mortality at age, based on the empirical relationship with the von Bertalanffy L_∞ and k, using a combination of the method described in Then et al. (2015), Lopez-Quintero et al. (2017) and Lorenzen (2000). Variability in the parameters in the $\text{Pauly}_{\text{nls-T}}$ relationship (Then et al., 2015) was included when calculating the natural mortality by drawing from their associated covariance matrix following the approach described by Lopez-Quintero et al. (2017). Combining the uncertainty in von Bertalanffy growth parameters and the uncertainty in the $\text{Pauly}_{\text{nls-T}}$ encapsulates the uncertainty in all of the modeled processes and also preserves the parameter correlation from each external analysis. Steepness was assumed to be independent of the other biological processes and was drawn from a censored-beta distribution with a median of 0.88 (Myers et al., 1999) and a variance which matched the range from the previous assessment (0.65–0.95) (Takeuchi et al., 2017). Note that the median steepness value of 0.88 used in the current study differs from the steepness of 0.8 assumed in the diagnostic

case of the previous assessment (Takeuchi et al., 2017). The steepness value of 0.88 reflects the available scientific information for this species (Myers et al., 1999).

The resulting distributions of the biological relationships for the MCB ensemble are presented in Fig. 1. The MCB ensemble approach created 1255 models, each with a different set of biological parameters that was fixed within the assessment model. All models were then fit to the same data used in the 2017 stock assessment using the program MULTIFAN-CL (*v*2.0.8.0).

Multiple MCB ensembles were created in order to investigate how uncertainty in management reference points changed with ensemble size. Thus, the 1255 models from the MCB ensemble were separated into independent subsets to create new ensembles with sample sizes of 500, 300, 200, 100, 75, 50, and 30. An ensemble with a sample size of 243 was randomly drawn from the 1255 models without replacement. The 243 model MCB ensemble was compared to the factorial ensemble because that was the size of the factorial ensemble approach (see below).

Three reference points commonly used to assess stock status, two based on maximum sustainable yield (MSY) and one based on depletion from the unfished condition, were calculated for each model in the MCB ensemble. The two MSY-based reference points, where MSY is based on the average fishing mortality at age in the last 5 years of the model excluding the last year, SB/SB_{MSY} and F/F_{MSY}, show terminal spawning biomass (SB) and fishing mortality (F) relative to the SB or F that produces MSY. The depletion-based reference point, $SB/SB_{F=0}$, is a derived quantity from the model that is calculated as the terminal SB relative to the unfished SB in the terminal year. Unfished SB is calculated using the estimated stock recruitment relationship and the time series of recruitment deviates to determine the biomass that would be present without fishing mortality. A classification and regression tree (CART) analysis for the MCB ensemble with 243 models was conducted to determine which variables in the MCB joint parameter distribution were most influential in explaining variance in the reference point. Separate CART models were fit for each of the three reference points. The CART models included all fixed biological parameters from the joint parameter distribution as covariates and one of the three stock status variables as the response. CART models were fit using the *rpart* package (*v*4.1–15, Therneau and Atkinson, 2019) in R. Default settings were assumed except that trees were pruned using a complexity parameter (*cp*) of 0.02 rather than the default 0.01. This larger *cp* value resulted in a slightly less complex tree which facilitated graphical visualization.

2.2.2. Factorial approach

The factorial approach typically assigns a high, medium, and low

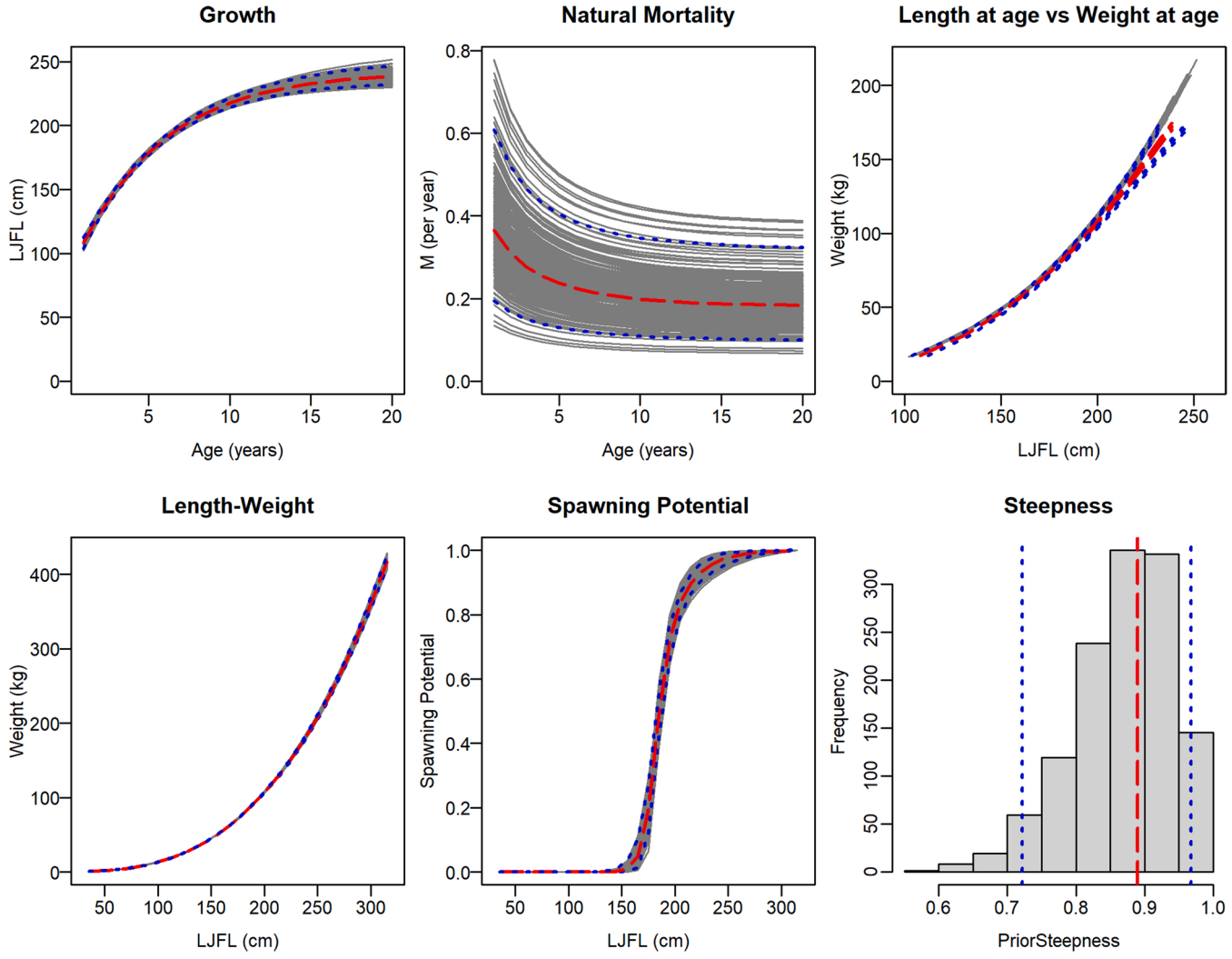


Fig. 1. Plot of biological relationships assumed within the ensembles where the solid grey lines are from the MCB ensemble, the red dashed line is the median used in the factorial ensemble and the two blue dotted lines are the 95% confidence interval. Top left: growth relationship (LJFL; lower jaw fork length); top center: natural mortality at age; top right: length at age from the von-Bertalanffy against the weight-at-age; bottom left: length- weight relationship; bottom center: spawning potential at length; bottom right: steepness of stock recruitment function.

value to be used for each axis of uncertainty in the model ensemble. To this end, the 2.5, 50, and 97.5 percentiles from the MCB joint parameter distribution were calculated for the growth, length-weight, spawning potential, natural mortality, and steepness. For biological relationships that were input into the stock assessment model as vectors (e.g., natural mortality at age or spawning potential at length), the percentiles were calculated across ages or lengths. The values of the biological relationships used in the factorial ensembles are shown as the dotted and dashed lines in Fig. 1. This created five axes of uncertainty (growth, natural mortality, length-weight, spawning potential, and steepness) with three options for the fixed parameters defining these relationships in the assessment. A full factorial combination of these axes of uncertainty was conducted to create a total of 243 models in the factorial ensemble. These models were fit to the data using the program MULTIFAN-CL in the same manner as the MCB ensemble.

2.2.3. Ensemble comparisons

Distributions of reference points were compared among the MCB ensembles and the factorial ensemble by boxplots of the converged and non-converged models. Given that the 2017 SWPO swordfish stock assessment was used as a base model, convergence was determined using the same criteria: (i) presence of a positive definite Hessian solution and (ii) a maximum gradient less than 10^{-3} (Takeuchi et al., 2017). Non-parametric tests were used to determine if the medians and variances of the estimated reference points between ensembles were similar. Wilcox rank sum tests (i.e., Mann–Whitney U test) for each ensemble were conducted to compare median estimates of reference points. Fligner tests were conducted between the ensembles to determine differences in variance estimates of reference points. Significant differences between ensemble medians and variances were based on $P \leq 0.05$.

Estimates of uncertainty for the two ensembles were calculated through four methods, and density plots of each reference point are shown for converged models in both ensembles. The first method incorporated only model uncertainty and the density distribution is from the maximum likelihood point estimates from converged models in each ensemble. The second method incorporated both the model and estimation uncertainty (Stewart and Martell, 2014; Stewart and Hicks, 2021). For each model retained in the ensemble, the estimation uncertainty for the three reference points was approximated by drawing 10,000 samples from a multivariate lognormal (MVLN) distribution. This distribution was created from the estimates of log-transformed MSY--based reference points and $SB/SB_{F=0}$ where the variance-covariance matrix of the log-transformed reference points were approximated with the delta method (Fournier et al., 2012) using a second order Taylor approximation. The model specific parameter distributions were combined across the m models in the ensemble such that the final combined parameter distribution had $10,000 \times m$ samples. Measures of central tendency (e.g., median) and variance were calculated based on this final combined parameter distribution. This is similar to the approach used by Winker et al. (2019) and Walter and Winker (2019) for combining uncertainty across *Stock Synthesis* (Methot and Wetzel, 2013) models. The third method used the model uncertainty but weighted each model through sampling importance resampling (SIR; McAllister and Ianelli, 1997). To conduct SIR, 8000 models were sampled with replacement from the ensemble with a probability of each model drawn as the log-likelihood of the model divided by the sum of all log-likelihoods in the ensemble. The sample size of 8000 was chosen to ensure that the maximum importance ratio was less than 0.04 and the maximum single density was less than 0.01 (McAllister and Ianelli, 1997). The fourth method was similar to the previous but incorporated both measures of uncertainty and weighted the models through SIR described above. From each sampled model in the SIR, 10,000 values of each reference point were drawn from the multivariate normal distribution of reference points from the approximation based on the estimated covariance matrix. Further details on how the reference points were transformed to log-space and how the variance-covariance matrix was constructed can be found in the S1 Appendix.

As SWPO swordfish does not have formerly agreed upon limit reference points for management within the WCPFC, we chose reference points that were representative of agreed limit reference points for other species under WCPFC management. The term *limit reference point* in this instance refers to the reference point threshold at which a stock is designated overfished or undergoing overfishing. These limit reference points are for illustrative purposes only and do not constitute management advice for this species. The three limit reference points chosen were $F/F_{MSY} > 1$, $SB/SB_{MSY} < 1$, and $SB/SB_{F=0} < 0.2$. The probability of exceeding these limit reference points was calculated for each model ensemble and uncertainty combination.

2.2.4. Variance partitioning

The total uncertainty in reference point X from an ensemble of models Y can be derived from the Law of Total Variances and is similar to the two-stage cluster sampling variance calculation described in Cochran (1977; Equations 10.15–10.16):

$$Var(X) = \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}(X|Y)) \tag{1}$$

where the first portion of Equation 1 corresponds to the average variation in X given any model Y_i in the ensemble and can be thought of as the variance within primary sampling units from cluster sampling. This "within-model" variance or estimation uncertainty can be generalized to account for the probability of X being drawn from any model Y_i:

$$\mathbb{E}[Var(X|Y)] = \sum_{i=1}^{m} \breve{w}_i \sigma_i^2 \tag{2}$$

where σ_i^2 is the variance of X from model Y_i, $\breve{w}_i$ is the normalized weighting of model Y_i in the ensemble ($\breve{w}_i = w_i / \sum_i w_i$). The second component of Equation 1 corresponds to how mean estimates of X may differ between ensemble models Y_i and can be thought of as the variance between sampling units from cluster sampling. This "between-model" variance in the mean of X or the model uncertainty across models Y_i can be generalized as:

$$Var(\mathbb{E}(X|Y)) = \sum_{i=1}^{m} \breve{w}_i (\mu_i - \overline{\mu}) \tag{3}$$

where μ_i is the mean of X from model Y_i and $\overline{\mu}$ is the grand mean $\sum_i \breve{w}_i \mu_i$. Combining Equations 2 and 3 results in the total variance of X across the Y_i models in the ensemble as the sum of estimation and model uncertainty:

$$Var(X) = \sum_{i=1}^{m} \breve{w}_i \sigma_i^2 + \sum_{i=1}^{m} \breve{w}_i (\mu_i - \overline{\mu}). \tag{4}$$

Partitioning the total variance from an ensemble in this way can allow for the calculation of the proportion of total uncertainty in reference point X attributable to either estimation uncertainty

$$\frac{\sum_{i=1}^{m} \breve{w}_i \sigma_i^2}{Var(X)}; \tag{5}$$

model uncertainty

$$\frac{\sum_{i=1}^{m} \breve{w}_i (\mu_i - \overline{\mu})}{Var(X)}; \tag{6}$$

or a particular model k

$$\frac{\breve{w}_k \sigma_k^2 + \breve{w}_k (\mu_k - \overline{\mu})}{Var(X)}. \tag{7}$$

Total variance was calculated and partitioned on a lognormal scale because samples drawn from MVLN distributions were used to combine the estimation uncertainty across models.

3. Results

Overall model convergence was good with 1433 of 1498 (95.7%) having both a positive definite Hessian solution and maximum gradient component less than 10^{-3}. The convergence rate of the full factorial ensemble (222 of 243 models; 91.4%) was marginally lower than the rate for the MCB ensemble (236 of 243 models; 97.1%). The full factorial ensemble had some problems with convergence for models with the combination of low natural mortality, high growth, and high or low length-weight. There were no obvious parameter combinations that resulted in poor convergence for the MCB ensemble models. The median SB/SB_{MSY} from the converged factorial ensemble was 2.768 with an interquartile range of 1.676, while the median F/F_{MSY} was 0.567 with an interquartile range of 0.358. The estimates of $SB/SB_{F=0}$ from models in the full factorial ensemble with a positive definite hessian had a median estimate of 0.422 and an interquartile range of 0.151.

Sample size of the MCB ensemble over the range investigated did not have a large influence on the reference point estimates either in terms of the median or the variability, though there were fewer outliers with smaller sample sizes (Fig. 2). The medians and interquartile ranges of reference points were similar for models that obtained a positive definite Hessian. For all sample sizes of the converged MCB ensembles, median SB/SB_{MSY} ranged between 2.037 and 2.204, median $SB/SB_{F=0}$ ranged between 0.327 and 0.341, and median F/F_{MSY} ranged between 0.686 and 0.719.

Wilcox rank sum tests on the median of F/F_{MSY} between the ensembles were significantly different (p-value $\leq$0.001) for the factorial ensemble and all MCB ensembles except those with sample sizes of 75 (p-value = 0.052) or 30 models (p-value = 0.720). Fligner tests between the factorial ensemble and all sample sizes of the MCB ensemble showed that the variance of F/F_{MSY} from the factorial ensemble was significantly different and larger than the variance from all MCB ensembles (p-value $\leq$0.038 for all MCB ensembles). Comparison of F/F_{MSY} among the other MCB ensembles was not significantly different from each another. Comparison of the median SB/SB_{MSY} by Wilcox rank sum tests showed that the factorial ensemble was significantly different from all MCB ensembles (p-value $\leq$0.013) except those with sample sizes of 30 (p-value = 1) and 75 (p-value = 0.887). The variance of SB/SB_{MSY} from the factorial ensemble was significantly different and larger than all MCB ensembles (Fligner test p-value <0.001). Additionally, Fligner tests on the variance of SB/SB_{MSY} showed that the variances from the MCB 50 and MCB 300 ensembles were significantly different (p-value = 0.045). Comparison of median $SB/SB_{F=0}$ by the Wilcox rank sum tests showed that the factorial ensemble was significantly different (p-value $\leq$0.031 for all MCB ensembles, but the MCB ensembles were not significantly different from one another. Fligner tests of the variance of $SB/SB_{F=0}$ were not significantly different among the MCB ensembles, but the factorial ensemble was significantly different from all MCB ensembles (p-value <0.008). Visualizations of all non-parametric comparisons of reference point medians and variances between ensembles can be found in S2 Appendix (Figures 1–3), and the corresponding p-values can be found in S1 Table.

Steepness and natural mortality were the primary variables selected by the CART models for all three reference points as best explaining the variance Fig. 3, 4, 5. The CART model for SB/SB_{MSY} indicated that higher values of steepness and natural mortality resulted in higher values of this reference point. Larger values of steepness were predicted to result in a lower value of F/F_{MSY}. F/F_{MSY} was predicted to be lowest

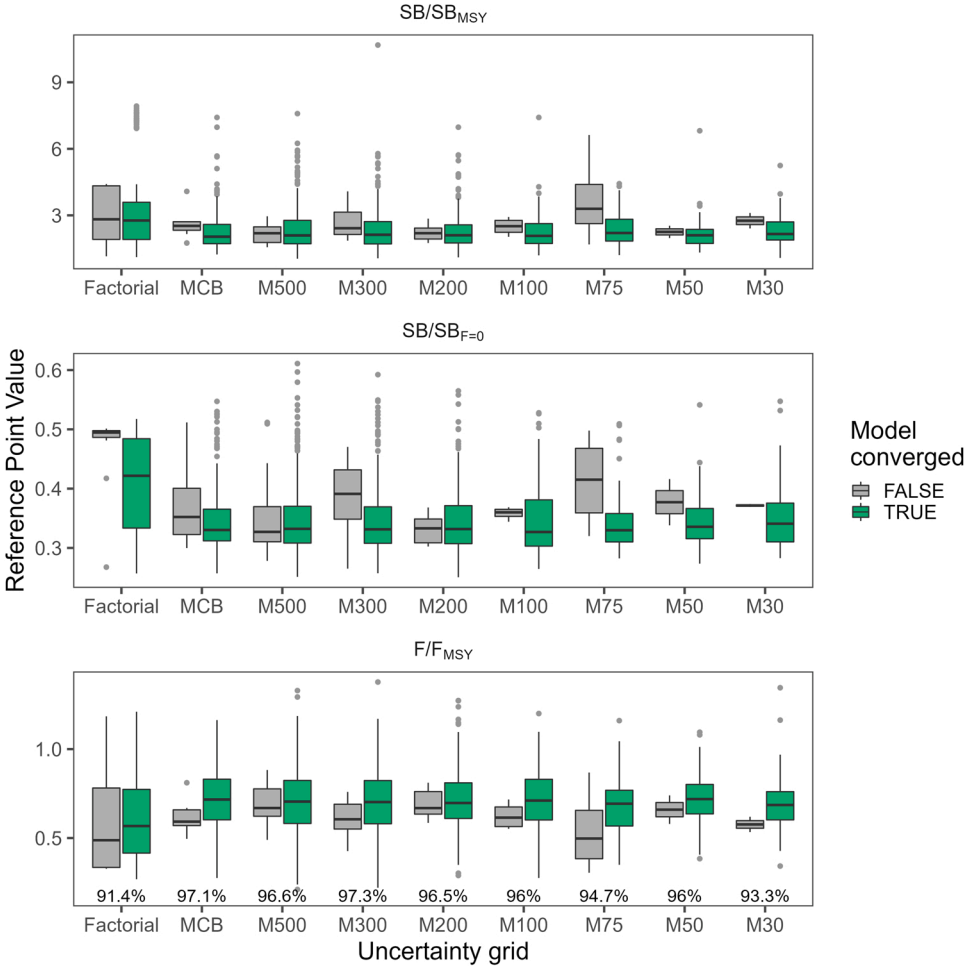


Fig. 2. Boxplots of reference points from a full factorial ensemble, MCB ensembles with different sample sizes (denoted by M and the number of the sample size), and an MCB ensemble with the same number of models as the factorial ensemble (243). The boxes indicate the 25th and 75th percentiles, the whiskers extend to two times the interquartile range, the thick black line is the median, and outliers are plotted as points. The percentages listed at the bottom of the bottom panel indicate the percentage of models that converged.

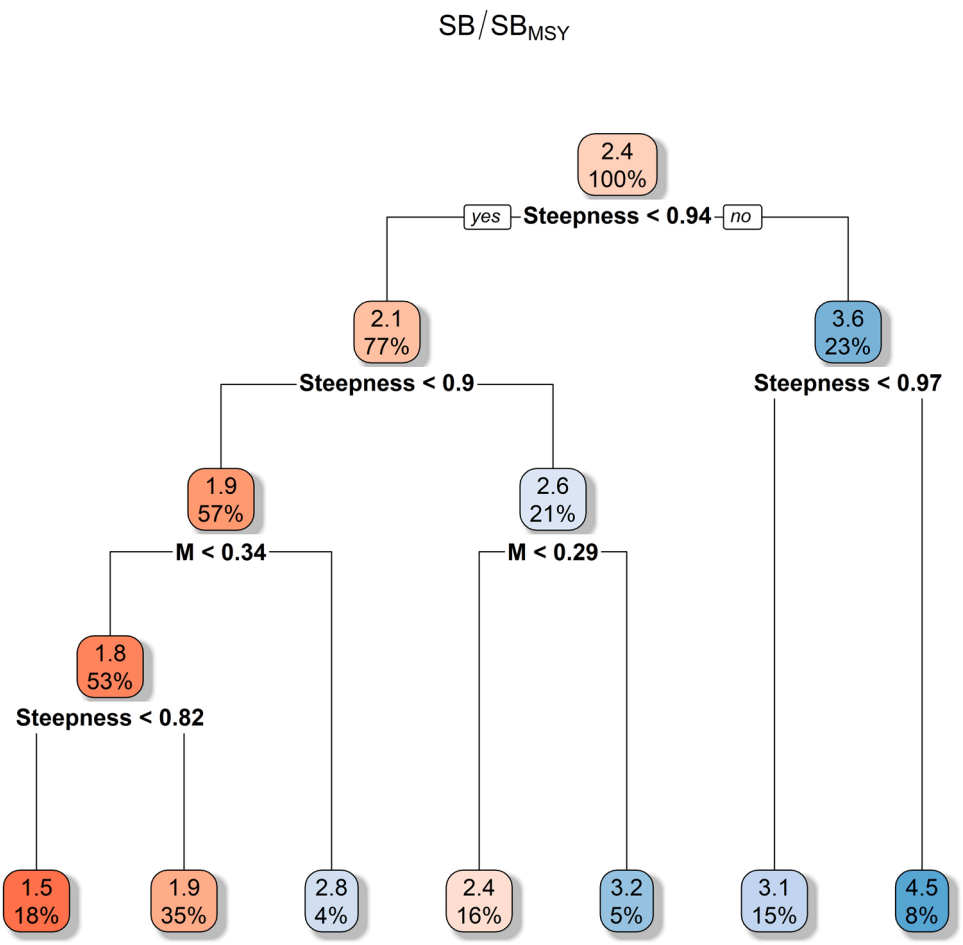


Fig. 3. Classification and regression tree (CART) analysis of biological parameters to explain SB/SB_{MSY} for the MCB ensemble with 243 models.

when steepness was greater than or equal to 0.9 and natural mortality was less than 0.16; with all other values of steepness F/F_{MSY} was predicted to be larger with smaller values of natural mortality. $SB/SB_{F=0}$ appeared to have a nonlinear relationship with natural mortality, where the largest and smallest values resulted in large values of $SB/SB_{F=0}$. Larger values of t_0 were predicted to result in smaller values of $SB/SB_{F=0}$ for intermediate values of natural mortality (0.19–0.28).

Uncertainty in estimates of the reference points for both the factorial and MCB ensembles was influenced depending on whether it included model uncertainty, model and estimation uncertainty, model uncertainty with SIR, or both uncertainties with SIR (Fig. 6). The incorporation of estimation uncertainty with the model uncertainty predictably resulted in a larger range in estimates for all reference points and both ensembles (Fig. 6). The marginal distributions with SIR were similar to the corresponding error type, though SIR for model uncertainty only resulted in a more jagged distribution due to resampling with replacement of individual models. The factorial marginal distributions for SB/SB_{MSY} had a thicker tail to the right compared to the MCB ensemble models for all error types. The marginal distribution of $SB/SB_{F=0}$ from the factorial ensemble had a larger median value and a larger variance compared to the MCB ensemble. F/F_{MSY} from the factorial ensemble had a more right skewed distribution compared to the MCB distribution, where the former had a mode less than 0.5 but the latter had a mode greater than 0.5.

The probability of reference points exceeding their respective limit reference points (for illustrative purposes) was influenced mostly by the error type and to a lesser degree the ensemble (Table 1). In the current case study, incorporation of estimation error into the reference points always resulted in an increase in the probability of exceeding the limit reference point, albeit a small increase. Reference points calculated only using model uncertainty showed zero probability of exceeding the limits for SB/SB_{MSY} and $SB/SB_{F=0}$, but when estimation uncertainty was included, the probability increased to ~ 2%. The incorporation of sampling importance resampling had very similar probability of exceeding reference point limits as the equal weighting of models for both ensembles. The factorial ensemble was more likely to exceed the limits for F/F_{MSY} than the MCB ensemble but was similar for the other two reference points.

Examination of the bivariate distributions of the reference points in terms of Kobe (Fig. 7) and Majuro (Fig. 8) plots shows similar patterns in how uncertainty changed across ensembles. The joint probability of being overfished and undergoing overfishing based on the Kobe plot (Fig. 7; quadrant D) increased when estimation uncertainty was incorporated from 0% to 2.2% for the MCB ensemble and from 0% to 2.7% for the factorial ensemble. In terms of the Majuro plot, the joint probability of being overfished and undergoing overfishing (Fig. 8; quadrant D) increased when estimation uncertainty was incorporated from 0% to 1.6% for the MCB ensemble and from 0% to 1.2% for the factorial ensemble. The bivariate distributions for the MCB ensemble in both the Kobe and Majuro plots were also less dispersed than the full factorial ensemble.

For MCB ensembles, risk (based on model and estimation uncertainty) of exceeding the reference point as a function of ensemble size appeared to be fairly consistent across MCB ensemble size for the depletion based reference point $SB/SB_{F=0}$ (Fig. 9). Risk levels were also generally consistent across different sizes of MCB ensembles for the two MSY based reference points. However, it should be noted that the MCB ensemble with 75 models indicated noticeably lower levels of risk.

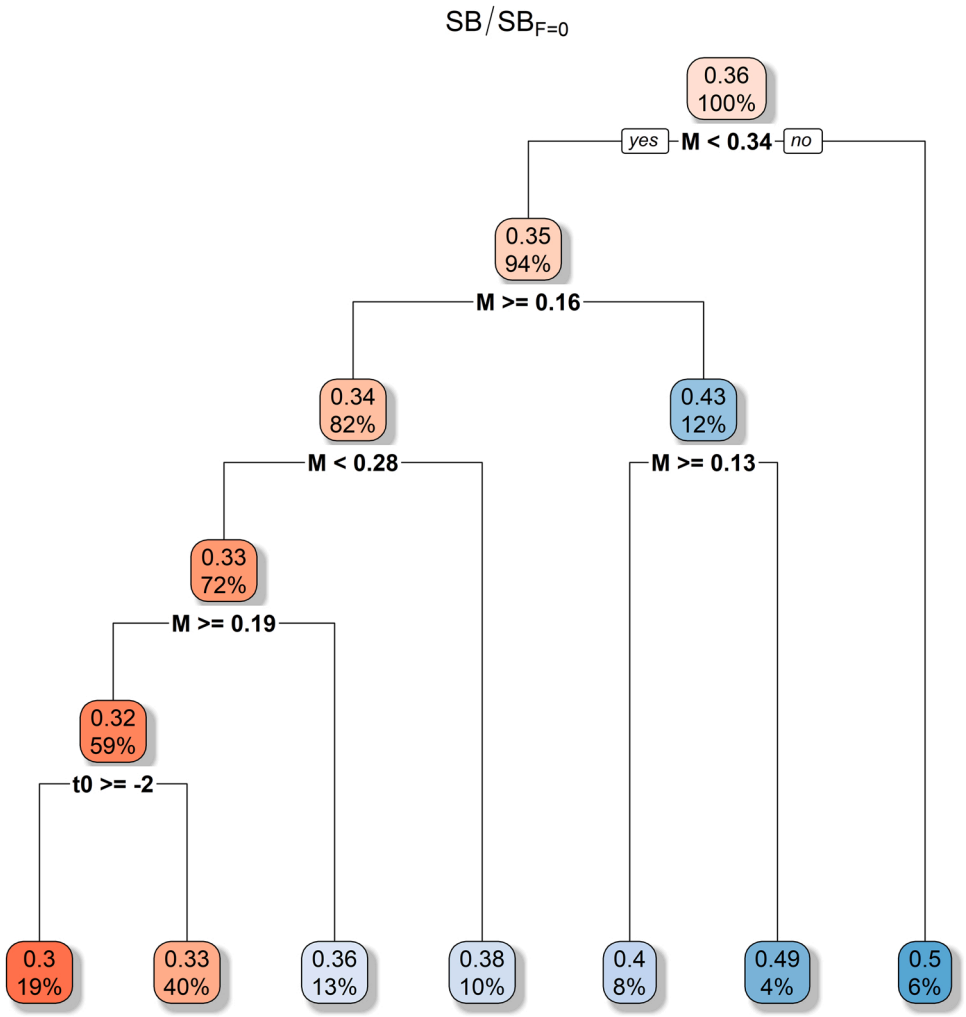


Fig. 4. Classification and regression tree (CART) analysis of biological parameters to explain SB/SB$_{F=0}$ for the MCB ensemble with 243 models.

Applying Equation 5 and Equation 6 to the 243 model MCB and factorial ensembles, with and without SIR, allowed us to partition the total variance in each of the three reference points (Table 2). Across all three reference points, the percentage of total variance attributed to model uncertainty was larger for the factorial ensemble than the MCB ensemble. For the MSY based reference points, the majority of total variance in the reference points came from model uncertainty while the opposite was true for the depletion based reference point. Variance partitioning was consistent between models with equal model ensemble weights or ensemble weights informed by SIR.

4. Discussion

In this study we investigated differences in management advice that would be provided from two model ensembles that used model uncertainty, model and estimation uncertainty, and both uncertainties with sampling importance resampling. The median reference points were statistically different between the two ensembles and the ensemble with the full factorial design showed more uncertainty. The higher variance and difference in the median value seen in the factorial ensemble could lead to different management advice depending on the probabilities of exceeding limits which are used in decision making. This is likely due to the factorial ensemble including biologically unreasonable parameter combinations and the choice of using the upper and lower bounds of the parameter 95% confidence interval to define the factorial levels. Using a smaller confidence interval to define the parameter range (e.g., 50% confidence interval) would not over-represent the tails of the distribution in the factorial approach. However, it would under-represent the uncertainty associated with that particular parameter. Additionally, our analysis also showed the MCB approach was computationally more efficient as reference point estimates, associated model uncertainty, and risk of exceeding reference points were consistent across MCB ensembles of varying sizes. Therefore, in this case a MCB ensemble could be created with as few as 50–100 models to capture uncertainty in fixed biological parameters used within the assessment. The sample size of 50–100 models may only be applicable for this scenario and the number of models required will be dependent on the relative model uncertainty, estimation uncertainty, and the covariance among the models. Further case studies with other species and models are required to define a generalized minimum model ensemble size. We recommend creating an ensemble of models that draws biological parameters that are fixed in the assessment model from a MCB distribution due to the increased model efficiency, better representation of input parameters and less spurious combinations of fixed parameters compared to the factorial approach. In cases where it is not computationally feasible to construct a 50 + model ensemble, ensemble "stacking" (Ting and Witten, 1999) of models parameterized with biological parameters sampled from a Gaussian-Hermite quadrature approximation of the MCB joint parameter distribution can appropriately capture uncertainty in data-limited assessments with as few as 4–30 models (Rudd et al., 2019) and could be extended to the recommended approach for integrated assessments.

Though the CART analysis partitioning variance between reference

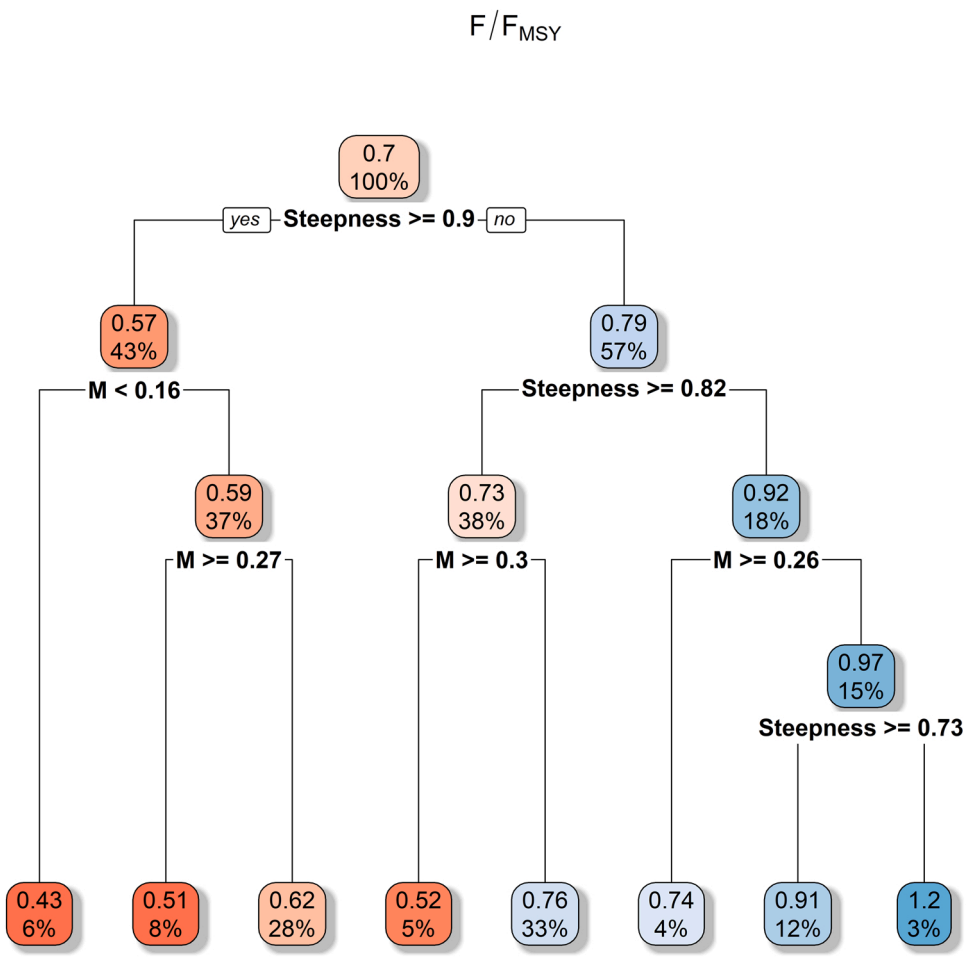


Fig. 5. Classification and regression tree (CART) analysis of biological parameters to explain F/F_{MSY} for the MCB ensemble with 243 models.

points based on fixed biological parameters can be applied to either the full factorial or MCB ensembles, an additional advantage of the MCB ensemble is that this analysis can be done with greater resolution. This can be useful to identify which parameters are influential on the model results particularly if there are non-linear interactions between parameters. By identifying which parameters are most influential on a model, future research on the biology of a species can be prioritized to reduce uncertainty in management advice. For example, in the current SWPO swordfish case study, a better understanding of the natural mortality rate could reduce uncertainty in all reference points. Similarly, reducing the uncertainty in the steepness of the Beverton-Holt stock recruitment relationship could reduce the variability in SB/SB_{MSY} and F/F_{MSY}. More precise estimates of t_0 could reduce some uncertainty in $SB/SB_{F=0}$. These uncertainties could be used to direct and prioritize future research on SWPO swordfish.

The MCB ensemble approach could be applied to the specification of operating models within an MSE framework (Punt et al., 2016). However, an MSE framework does not need to be in place for an ensemble approach to be used since an MSE is used to develop robust management procedures while an ensemble is used to characterize key sources of uncertainty. Additionally, even if current management decisions do not incorporate uncertainty in reference point estimates, these should still be presented to managers. This will provide a realistic picture of the current understanding of the stock and could lead to management practices that incorporate the uncertainty explicitly, consistent with the precautionary approach. Uncertainty in biological parameters and input data has been incorporated into management advice for numerous species under federal jurisdiction in the southeast United States and Hawaiian Islands using MCB ensembles (Legault et al., 2002; Nadon, 2017; SEDAR, 2021). Management of these species does not currently entail an MSE, but the uncertainty in reference points resulting from the uncertainty in biological parameters and data is incorporated into setting the catch limits through the P-star approach. Other approaches for quantifying uncertainty for management advice are decision tables produced for Pacific hake and halibut assessments (Johnson et al., 2021; Stewart and Hicks, 2022). Therefore, the ensemble approach applied in this paper can be used as a part of the framework for quantifying uncertainty in stock status, which then feeds into setting management measures.

Despite our recommendation to use ensembles through an MCB approach, a full factorial ensemble is a valid and warranted approach for creating ensembles in some scenarios. We do note that our recommendation of the MCB ensemble is conditional on the sensible construction of the joint parameter distribution and that it contains the "truth". In cases where the joint parameter distribution can not be sensibly constructed to reflect the best understanding of the stock or where the potential for bias in the joint parameter distribution is high, a more conservative precautionary approach would be to consider the full factorial ensemble given that it produces over-dispersed uncertainty relative to the MCB ensemble. A factorial approach can be created in the absence of a joint parameter distribution by choosing the highest and lowest values (along with an intermediate value) that are deemed plausible for a specific parameter. Additionally, a full factorial ensemble design should be used when there are discrete choices between model structures that cannot be characterized as distributions. A good example of a full factorial axis of uncertainty would be models with differing

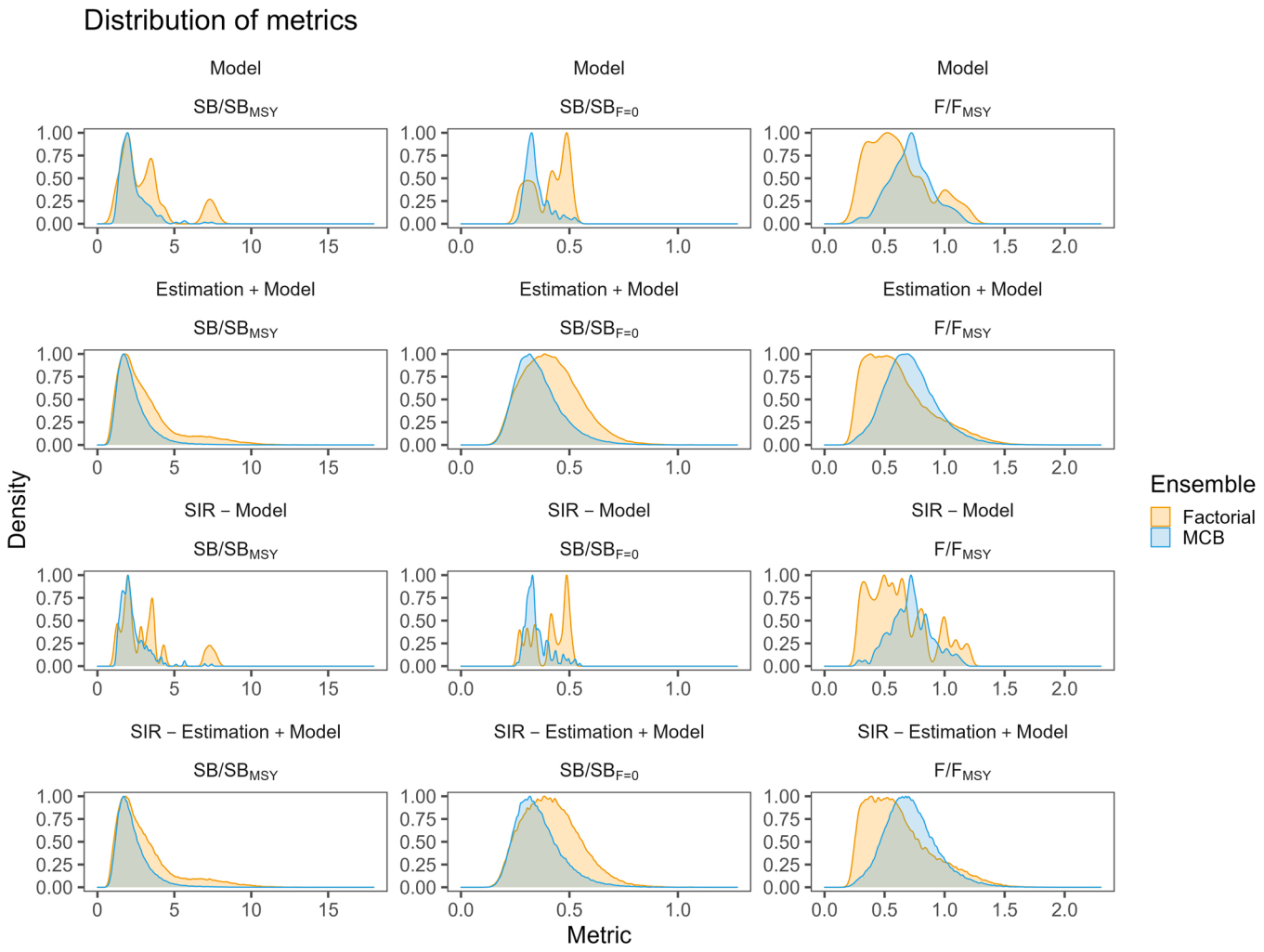


Fig. 6. Estimated marginal distributions of reference points for two ensembles where the left column is SB/SB_{MSY}, the center column is $SB/SB_{F=0}$, and the right column is F/F_{MSY}. The top row is for model uncertainty only, the second row is the estimation and model uncertainty, the third row is the sampling importance resampling for the model uncertainty only, and the bottom row is the sampling importance resampling with both estimation and model uncertainty.

Table 1
The percent of samples for each reference point (SB/SB_{MSY}, $SB/SB_{F=0}$, and F/F_{MSY}) that exceeded their respective limit reference points for the factorial and MCB ensembles under the error distributions of model only, model and estimation, and both weighted by sampling importance resampling (SIR).

	$SB/SB_{MSY} < 1$		$SB/SB_{F=0} < 0.2$		$F/F_{MSY} > 1$	
Error Type	Factorial	MCB	Factorial	MCB	Factorial	MCB
Model	0	0	0	0	10.4	7.2
Estimation + Model	2.6	2.3	1.9	2.4	10.7	9.8
SIR - Model	0	0	0	0	10.7	6.9
SIR - Estimation + Model	2.7	2.3	1.9	2.4	11.1	9.8

hypotheses regarding the functional form of the stock recruitment relationship (e.g., Ricker relationship, a Beverton-Holt relationship, or constant recruitment). Other examples where a factorial ensemble approach could be applied include: the functional forms of selectivity, assessment spatial structure, alternative catch reconstruction time series, and different standardization approaches of CPUE indices. Model ensembles that are a hybrid between the factorial and MCB approaches could easily be created to incorporate the uncertainty in fixed biological parameters and competing hypotheses of states of nature. For example, a hybrid ensemble could use parameter sets drawn from the joint parameter distribution for each of the axes or models in the factorial design.

In the current analysis, the steepness distribution was developed independently from the other biological parameters which is consistent with how uncertainty in steepness is treated in most WCPFC assessments (Tremblay-Boyer et al., 2017; Takeuchi et al., 2017; Vincent et al., 2019, 2020; Ducharme-Barth et al., 2020). There is a lack of scientific consensus (summarized by Munyandorero, 2020; Zhou et al., 2020) on how steepness correlates with life history characteristics (e.g., longevity or natural mortality) with some studies indicating that steepness could be positively correlated, negatively correlated (Forrest et al., 2010), or show no correlation with longevity (Thorson, 2020). Sensitivity to how steepness may correlate with the other biological parameters was not considered, though it is unlikely to qualitatively alter the main conclusions of this study. In the future, should uncertainty in steepness be considered as a part of a model ensemble approach, analysts are encouraged to develop a steepness distribution with an appropriate correlation with other life history parameters.

Though the focus of this study was on how ensemble construction and the characterization of uncertainty impact the reference points used to provide management advice, we did not explicitly consider how projections would be combined with an ensemble approach. Projections themselves are typically ensembles in their own respect as they often consider multiple future fishing mortality and recruitment scenarios. Further work on appropriate methods to integrate projections into the ensemble framework is needed. In the interim our recommendation, if

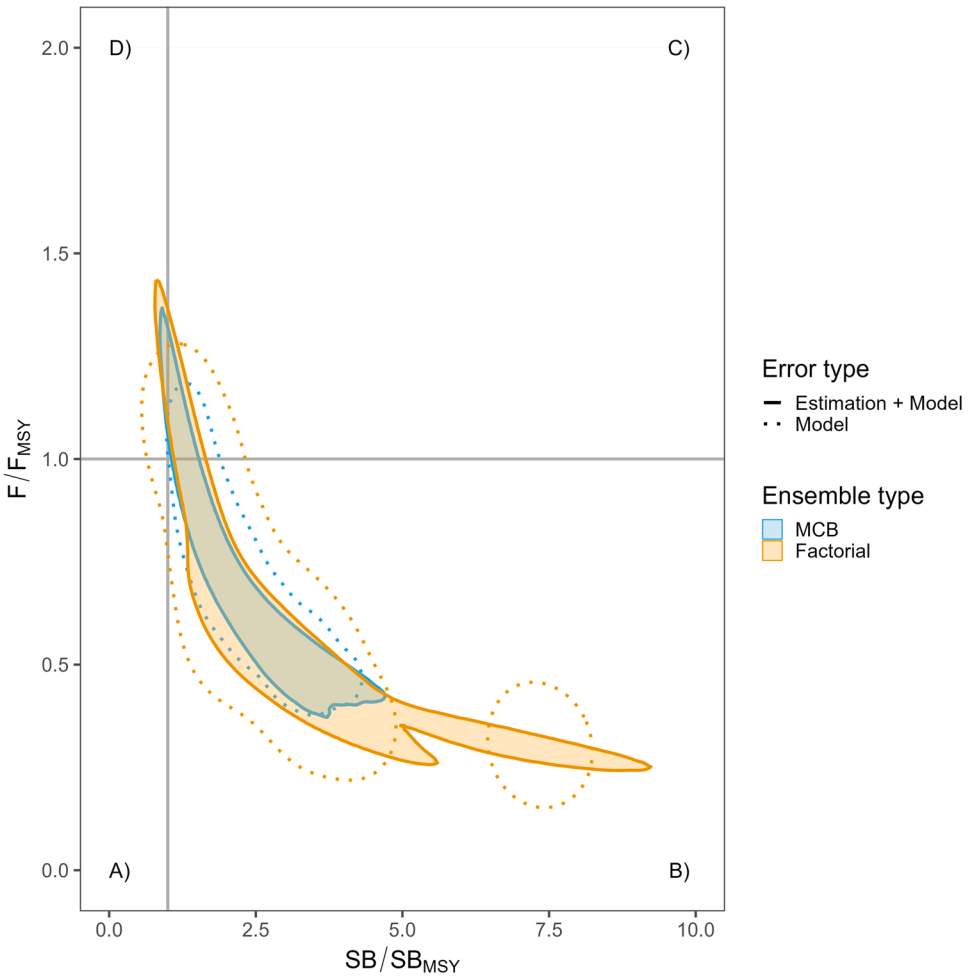


Fig. 7. A Kobe plot or the bivariate distribution of SB/SB_{MSY} and F/F_{MSY} for the 243 model MCB (blue) and factorial (orange) ensembles. The 90th percentile of the kernel density is shown where the line type denotes the model error type: Estimation + Model (solid) and Model only (dashed). The different quadrants indicate stock status with quadrant A) indicating the stock is overfished but not undergoing overfishing, quadrant B) indicating the stock is not overfished and not undergoing overfishing, quadrant C) indicating the stock is not overfished but undergoing overfishing, and quadrant D) indicating that the stock is overfished and undergoing overfishing.

computationally feasible, would be to evaluate all projection scenarios for each model in the ensemble resulting in an "ensemble of ensembles". Uncertainty could then be combined across projection scenarios *within* each model in the ensemble before combining uncertainty *across* models in the ensemble.

Quantification of uncertainty through the total variance equation (Equation (4)) allowed for determining the contribution of model and estimation uncertainty to the overall uncertainty for each reference point. Uncertainty in management advice is generally thought to be greater from structural model uncertainty than it is from estimation uncertainty (Maunder et al., 2020; Scott et al., 2016). However, this was not the case for all reference points that are presented in this study. For the MSY-based reference points, the majority of uncertainty was from the model uncertainty, whereas the estimation uncertainty accounted for about a quarter of the uncertainty. Uncertainty in $SB/SB_{F=0}$ was equally attributed to model and estimation uncertainty for the factorial models, but estimation uncertainty accounted for about three quarters of the uncertainty for the MCB ensemble. We note that the $SB/SB_{F=0}$ is influenced much less by the assumed stock-recruitment steepness compared to MSY reference points, which is an inherent property of the former confirmed by the CART analysis. Thus in the current case, if only model uncertainty from the ensemble was used in the creation of management advice based on $SB/SB_{F=0}$, then the uncertainty would be underrepresented and could lead to risk prone management.

The reference point estimates from an ensemble can be combined through a multitude of techniques. These methods include simple averaging, likelihood weighting (e.g., AIC), and cross validation (Scott et al., 2016). Simple averaging of reference points can easily be conducted for a large number of models and can incorporate both estimation and model uncertainty (Ianelli et al., 2016). Simply averaging across models in a factorial ensemble is implicitly assuming equal probability of the states of nature represented by all models (Maunder et al., 2020). The MCB approach implicitly puts additional weight on combinations of parameters that are the most representative of our current understanding of the biology of the species given data external to the assessment model. Thus, averaging across models may be a reasonable assumption to make for the MCB ensembles. Conversely, the full factorial approach may present different assumptions about modeled relationships that have differing levels of plausibility. For example, the combination of the high growth and high length-weight relationship from the factorial ensemble resulted in a length at age and weight at age relationship which was well outside the range seen from the MCB approach (top right Fig. 1). This value outside the expected range is not observed when looking at the growth or length-weight relationships individually. However, this resulting interaction could potentially explain the difficulty in convergence for certain combinations in the full factorial ensemble.

Assigning weights to various hypotheses in a full factorial ensemble is difficult and often resolved through 'expert opinion' (Maunder et al., 2020). These expert opinions (i.e., subjective weightings) regarding the multiple hypotheses present in a full factorial ensemble should be assigned before the results of the assessment are revealed. This reduces the possibility that the weighting of the hypotheses is driven by the resulting stock status of the models. However, this does not always prevent such bias from occurring because some modeling assumptions can have predictable results (e.g., higher steepness will have a higher

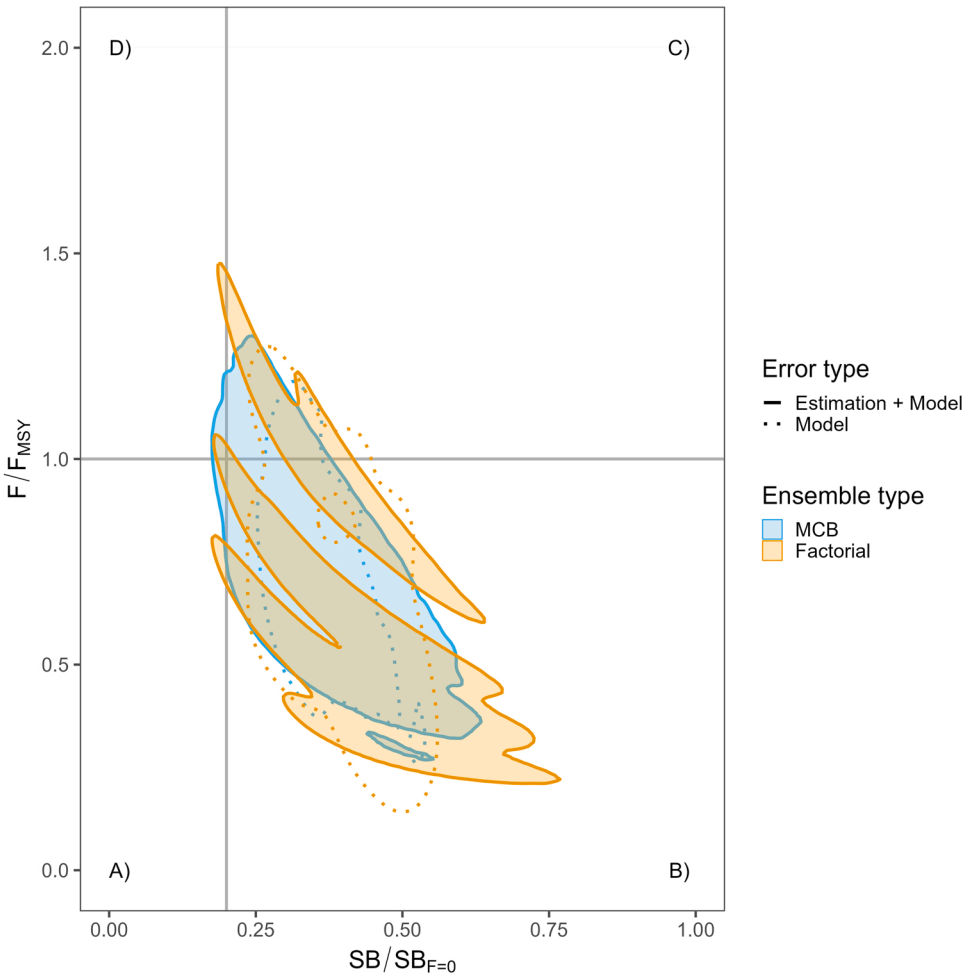


Fig. 8. A Majuro plot or the bivariate distribution of $SB/SB_{F=0}$ and F/F_{MSY} for the 243 model MCB (blue) and factorial (orange) ensembles. The 90th percentile of the kernel density is shown where the line type denotes the model error type: Estimation + Model (solid) and Model only (dashed). The different quadrants indicate stock status with quadrant A) indicating the stock is overfished but not undergoing overfishing, quadrant B) indicating the stock is not overfished and not undergoing overfishing, quadrant C) indicating the stock is not overfished but undergoing overfishing, and quadrant D) indicating that the stock is overfished and undergoing overfishing.

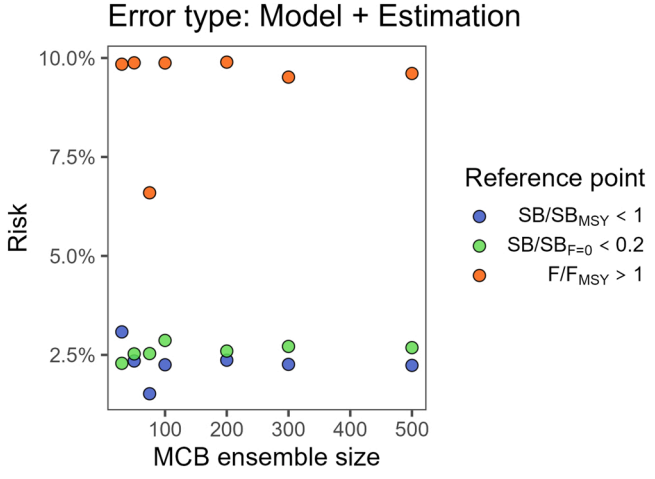


Fig. 9. The risk of exceeding limit reference points based on MCB ensembles with both model and estimation uncertainty. The risks of exceeding the limit reference points SB/SB_{MSY}, $SB/SB_{F=0}$, and F/F_{MSY} are shown as a function of ensemble size.

F_{MSY}). Thus, difficult discussions regarding the incorporation and weighting of uncertainty in stock assessments of managed species should occur on the front end of the assessment process. This prevents political motivations from driving the advice that is presented for management of a species. The advice will instead be influenced by an understanding of the biology of the species. Averaging of results based on expert opinion (even with multiple experts) is less than ideal because the results would not be reproducible with a different analyst or group of experts.

Alternatives to expert based weighting schemes for model ensembles exist but may not be feasible for all situations. Likelihood weighting methods have been proposed as a more objective way of model averaging. However, these do not always select the 'correct' model from the ensemble and could potentially lead to providing biased management advice. Additionally, these methods only work when the same data and likelihoods are used in the models (Jardim et al., 2021). Therefore, these methods cannot be used when different data weightings are assumed in the ensemble or when different datasets are used in an ensemble. Thus, the applicability of these likelihood methods is limited for most assessment ensemble contexts. Cross validation methods can be applied in cases where data and likelihoods differ. However, these can be computationally intensive and thus may not be practical for models that take a long time to converge or for large ensembles (Maunder and Harley, 2011).

Model diagnostics (e.g., those described in Carvalho et al., 2021) could be used to develop a more objective, data/likelihood invariant model weighting scheme. However, there is a lack of consensus on which diagnostics to use and how they can be combined to create objective ensemble weights. Recent work using hindcast predictions of CPUE indices (or composition data) has proposed the use of mean absolute scaled error (MASE) of the hindcast predictions as a potential diagnostic based approach for model ensemble weighting (Kell et al., 2021). This approach is promising in that hindcast MASE scores are comparable across models with differences in the input data and/or

Table 2
The percentage of total variance in each reference point attributed to model or estimation uncertainty from the 243 model factorial and MCB ensembles with model and estimation uncertainty, and either equal or SIR model ensemble weighting.

	SB/SB_{MSY}		$SB/SB_{F=0}$		F/F_{MSY}	
Ensemble	Model	Estimation	Model	Estimation	Model	Estimation
Factorial	82.3	17.7	46.4	53.6	83.9	16.1
SIR - Factorial	82.4	17.6	46.2	53.8	83.9	16.1
MCB	64	36	27.4	72.6	68.8	31.2
SIR - MCB	63.1	36.9	27.2	72.8	68	32

likelihoods. However, the details of performing such weighting need additional evaluation for determining which data source should be removed for hindcasting, whether all data sources need to be individually hindcast, and which prediction interval is most appropriate (e.g., 1, 3, or 5 years hindcast predictions). Additionally, investigation is needed in the correct way to combine the metric across multiple CPUE indices or metrics from removing different data sources.

If differences in likelihoods between models are small relative to total likelihood values, then re-weighting approaches based on the likelihood will generally produce management advice similar to equal weighting. This was demonstrated in the study when using sampling importance resampling. Differences in total likelihood may be larger for models with structural differences among models, but these likelihoods may not always be comparable, which would preclude the use of this method. Management advice after model weighting is only likely to differ significantly if model weighting removes models from the tails of the distribution. However, it is always possible that choosing different weighting methods could allow/prevent management criteria based on probability of exceeding a reference point from being activated. Thus, further research on the best approach for ensemble averaging is required.

This analysis attempted to quantify the total variance of reference points in the estimation and model uncertainty ensemble by combining k samples drawn from unique MVLN distributions for each m model in the ensemble (Winker et al., 2019; Walter and Winker, 2019). This approach is commonly used to derive a proxy for total variance of reference points because it is relatively straightforward to implement the sampling procedure (i.e. *SSdeltaMVLN* function of the *ss3diags* R package for Stock Synthesis users; Carvalho et al., 2021). This sampling scheme can also readily account for different ensemble weighting schemes by manipulating the number of samples taken from any given model, and calculate measures of central tendency and variance from the combined parameter distribution of $m \times k$ samples. While combining MVLN samples in this way preserves *within* model parameter correlation, it does not require knowing the correlation *between* models to calculate the total variance of the ensemble.

Model ensembles are commonly used in other scientific arenas to reduce prediction variance among models (Dormann et al., 2018). Ensemble outputs are combined by taking a weighted average across models for each k sample from the parameter distributions resulting in a variance for the averaged quantity ($\overline{X}$) given by:

$$Var(\overline{X}) = \sum_{i=1}^{m} \breve{w}_i^2 \sigma_i^2 + \sum_{i=1}^{m} \sum_{j \neq i} \breve{w}_i \breve{w}_j \rho_{ij} \sigma_i \sigma_j, \quad (8)$$

where ρ_{ij} is the correlation in predictions *between* models. This is fundamentally different to approximating the ensemble variance by concatenating all draws from the MVLN distributions across models into a single distribution. Calculating ensemble variance through model averages has the benefit of reducing prediction error through bias and variance reduction (Dormann et al., 2018). This formulation (Equation 8) crucially relies on accounting for the *between* model correlation to accurately characterize the variance in the model averaged quantities.

Reference points from ensembles derived by combining (Winker et al., 2019; Walter and Winker, 2019) or averaging (Dormann et al., 2018) samples across models can yield two very different outcomes in terms of the uncertainty portrayed. Indeed, numerical simulations (S3 Appendix) show that the variance reducing property of the Dormann et al. (2018) approach consistently results in smaller estimates of variance relative to approximating ensemble variance by combining samples from multiple MVLN distributions. This is predictable since the two definitions of ensemble uncertainty are not strictly comparable. The Dormann et al. (2018) approach characterizes the weighted variance in ensemble model means while the Walter and Winker (2019) approach characterizes the weighted variance of the combined ensemble estimates. Moving forward, the Dormann et al. (2018) approach could easily be incorporated as an option into functions such as *SSdeltaMVLN* by allowing users the option to average across rather than combine samples. However, the Dormann et al. (2018) approach depends on knowing the between model correlation ρ_{ij} which is challenging to derive. Ignoring the correlation component of Equation 8 (i.e., assuming $\rho_{ij} = 0$) could lead to a biased approximation of variance when averaging across samples in the Dormann et al. (2018) approach. It may be reasonable to expect some level of positive correlation between models given the similarity in model structures used in most ensembles for stock assessment. Assuming $\rho_{ij} = 0$ in Equation 8 when it is positive and non-negligible would result in an underestimate of the model-averaged variance (J. Brodziak, personal communication). Additionally, the Dormann et al. (2018) approach effectively removes the tails of the distribution by averaging across samples. This reduction in variance may not be desirable from a risk standpoint since model ensembles are often used to characterize the uncertainty among alternative model formulations. Currently, an appropriate choice between these two methodologies is unclear. Further investigation to formally evaluate (ideally using simulation) the stock assessment implications of quantifying ensemble variance using the two approaches is needed. Future research into quantifying ensemble variance using the Dormann et al. (2018) approach should also seek to develop a robust, computationally feasible framework for calculating between model correlation.

In conclusion, both model and estimation uncertainty should be included in reference point calculations for management advice. This will allow the most appropriate representation of the current knowledge from the assessment models. Ensembles could be created using a hybrid approach where fixed parameters are drawn from joint parameter distributions and competing hypotheses of functional forms of the states of nature should be included in a full factorial fashion. Further research on objective model averaging that can be used in situations with differing likelihoods and an appropriate way to quantify total ensemble variance is required.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Input data and code for replicating the analysis can be found at the following repository: dx.doi.org/10.6084/m9.figshare.16775860.

Acknowledgements

The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author(s) and do not necessarily reflect those of NOAA, the Department of Commerce, or Pacific Community (SPC). We thank P. Hamer, J. Hampton, and members of the Stock Assessment and Modeling Team in the SPC-Oceanic Fisheries Programme for reviewing drafts of the manuscript and serving as a sounding board for ideas. We thank N. Davies for updating the MULTIFAN-CL software in order to address reviewer comments. We thank F. Bouyé for managing the SPC HTCondor network used for conducting model calculations. Thank you to J. Brodziak and R. Ahrens for productive discussions on model and estimation uncertainty. We also thank M. Fujiwara, A. Hicks, A. Punt, and H. Winker for formally reviewing previous drafts of this manuscript. Exterior funding was not used in this research and the authors declare no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2022.106452.

References

Anderson, S.C., Cooper, A.B., Jensen, O.P., Minto, C., Thorson, J.T., Walsh, J.C., Afflerbach, J., Dickey-Collas, M., Kleisner, K.M., Longo, C., Osio, G.C., Ovando, D., Mosqueira, I., Rosenberg, A.A., Selig, E.R., 2017. Improving estimates of population status and trend with superensemble models. Fish Fish. 18, 732–741. https://doi.org/10.1111/faf.12200.

AnonSTAN Development Team (2021). STAN Modeling Language Users Guide and Reference Manual.

AnonSEDAR (2021). SEDAR 73 South Atlantic Red Snapper Stock Assessment Report. Technical report, SEDAR, North Charleston SC.

Berger, A.M., Geothel, D.R., Lynch, P.D., Quinn II, T.J., Mormede, S., McKenzie, J., Dunn, A., 2017. Space oddity: the mission for spatial integration. Can. J. Fish. Aquat. Sci. 74 (11), 1698–1716.

Brodziak, J., Piner, K., 2010. Model averaging and probable status of North Pacific striped marlin, Tetrapturus audax. Can. J. Fish. Aquat. Sci. 67 (5), 793–805.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R., Maunder, M.N., Taylor, I., Wetzel, C.R., Doering, K., Johnson, K.F., Methot, R.D., 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. 240, 105959.

Cochran, W., 1977. Sampling Techniques. John Wiley & Sons third ed.

Deriso, R.B., Quinn II, T.J., 1985. Catch-Age analysis with auxiliary information. Can. J. Fish. Aquat. Sci. 42 (4), 815–824.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C.M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J.J., Pollock, L.J., Reineking, B., Roberts, D.R., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Wood, S.N., Wüest, R.O., Hartig, F., 2018. Model averaging in ecology: a review of Bayesian, information-?theoretic, and tactical approaches for predictive inference. Ecol. Monogr. 88 (4), 485–504.

Ducharme-Barth, N., Vincent, M., Hampton, J., Hamer, P., Williams, P., and Pilling, G. (2020). Stock assessment of bigeye tuna in the western and central Pacific Ocean (30 July) - Rev.03. Technical Report SC16-SA-WP-03.

Farley, J., Clear, N., Kolody, D., Krusic-Golub, K., Eveson, P., and Young, J. (2016). Determination of swordfish growth and maturity relevant to the southwest Pacific stock. Technical Report WCPFC-SC12–2016/SA-WP-11, Bali, Indonesia, 3–11 August 2016.

Forrest, R.E., McAllister, M.K., Dorn, M.W., Martell, S.J., Stanley, R.D., 2010. Hierarchical Bayesian estimation of recruitment parameters and reference points for Pacific rockfish (Sebastes spp.) under alternative assumptions about the stock-recruit function. Can. J. Fish. Aquat. Sci. 67 (10), 1611–1634.

Fournier, D., Archibald, C., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39 (8), 1195–1207.

Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, Thunnus alalunga. Can. J. Fish. Aquat. Sci. 55, 2105–2116.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., Sibert, J., 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. 27 (2), 233–249.

Haddon, M., 2001. Quantitative Methods in Fisheries. Chapman and Hall.

Hilborn, R., 2003. The state of the art in stock assessment: where we are and where we are going. Sci. Mar. 67 (S1), 15–20.

Horswill, C., Kindsvater, H.K., Juan-Jorda, M.J., Dulvy, N.K., Mangle, M., Matthiopoulos, J., 2019. Global reconstruction of life-history strategies: A case study using tunas. J. Appl. Ecol. 56 (4), 855–865.

Ianelli, J., Holsman, K.K., Punt, A.E., Aydin, K., 2016. Multi-model inference for incorporating trophic and climate uncertainty into stock assessment. Deep Sea Res. Part II Top. Stud. Oceanogr.

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E.N., Johnson, K.F., Klibansky, N., Millar, C.P., Minto, C., Mosqueira, I., Nash, R.D.M., Vasilakopoulos, P., Wells, B.K., 2021. Operationalizing ensemble models for scientific advice to fisheries management. ICES J. Mar. Sci. 78 (4), 1209–1216.

Johnson, K.F., Edwards, A.M., Berger, A.M., and Grandin, C.J. (2021). Status of the Pacific Hake (whiting) stock in U.S. and Canadian waters in 2021. Technical report, Joint Technical Committee of the Pacifc Hake/Whiting Agreement Between the Governments of the United States and Canada.

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., Fu, D., 2021. Validation of stock assessment methods: is it me or my model talking? ICES J. Mar. Sci.

Legault, C.M., Powers, J.E., Restrepo, V.R., 2002. Incorporating uncertainty into fishery models. volume Symposium 27, page 208. American Fisheries Society (Section: Mixed Monte Carlo/Bootstrap Approach to Assessing King and Spanish Mackerel in the Atlantic and Gulf of Mexico: Its Evolution and Impact).

Lopez-Quintero, F.O., Contreras-Reyes, J.E., Wiff, R., 2017. Incorporating uncertainty into a length-based estimator of natural mortality in fish populations. Fish. Bull.

Lorenzen, K., 2000. Allometry of natural mortality as a basis for assessing optimal release size in fish-stocking programs. Can. J. Fish. Aquat. Sci. 57, 2374–2381.

Magnusson, A., Punt, A.E., Hilborn, R., 2012. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish Fish. 14 (3), 325–342.

Maunder, M.N., Harley, S.J., 2011. Using cross validation model selection to determine the shape of nonparameteric selectivity curves in fisheries stock assessment models. Fish. Res. 110 (2), 283–288.

Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72 (1), 7–18.

Maunder, M.N., Piner, K.R., 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fish. Res. 192, 16–27.

Maunder, M.N., Xu, H., Lennert-Cody, C.E., Valero, J.L., Aires-daSilva, A., and Minte-Vera, C. (2020). Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. Technical Report Document SAC-1 INF-F REV 3, Inter-American Tropical Tuna Commission: Scientific Advisory Committee, San Diego, California, USA.

McAllister, M.K., Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54 (2), 284–300.

Methot Jr., R.D., Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.

Minte-Vera, C.V., Maunder, M.N., Aires-daSilva, A.M., Satoh, K., Uosaki, K., 2017. Get the biology right, or use size-composition data at your own risk. Fish. Res. 192, 114–125.

Monnahan, C.C., Kristensen, K., 2018. No-U-turn sampling for fast Bayesian inference in ADMB and TMB: introducing the adnuts and tmbstan R packages. PLoS One 13 (5), e0197954.

Monnahan, C.C., Thorson, J.T., Branch, T.A., 2016. Faster estimation of bayesian models in ecology using hamiltonian monte carlo. Methods Ecol. Evol. 8 (3), 339–348.

Monnahan, C.C., Branch, T.A., Thorson, J.T., Stewart, I.J., Szuwalski, C.S., 2019. Overcoming long Bayesian run times in integrated fisheries stock assessments. ICES J. Mar. Sci. 76 (6), 1477–1488.

Munyandorero, J., 2020. Inferring prior distributions of recruitment compensation metrics from life-history parameters and allometries. Can. J. Fish. Aquat. Sci. 77 (2), 295–313.

Myers, R.A., Bowen, K.G., Barrowman, N.J., 1999. Maximum reproductive rate of fish at low population sizes. Can. J. Fish. Aquat. Sci. 56 (12), 2404–2419. https://doi.org/10.1139/f99-201.

Nadon, M. O. (2017). Stock assessment of the coral reef fishes of Hawaii, 2016.

Privitera-Johnson, K.M., Punt, A.E., 2020. A review of approached to quantify uncertainty in fisheries stock assessment. Fish. Res. 226.

Punt, A.E., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A.A., Haddon, M., 2016. Management strategy evaluation: best practices. Fish Fish. 303–334. ⟨https://onlinelibrary.wiley.com/doi/pdf/10.1111/faf.12104⟩.

Quinn II, T.J., Deriso, R.B., 1999. Quantitative Fish Dynamics. Oxford University Press.

R Core Team (2021). R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria.

Restrepo, V.R., Hoenig, J.M., Powers, J.E., Baird, J.W., Turner, S.C., 1992. A simple simulation approach to risk and cost analysis, with applications to swordfish and cod fisheries. Fish. Bull. 90 (4), 736–748.

Rudd, M.B., Thorson, J.T., Sagarese, S.R., 2019. Ensemble models for data-poor assessment: accounting for uncertainty in life-history information. ICES J. Mar. Sci. 76 (4), 870–883.

Scott, F., Jardim, E., Millar, C.P., Cervino, S., 2016. An applied framework for incorporating multiple sources of uncertainty in Fisheries Stock Assessments. PLoS One 11 (5), e0154922.

Stewart, I.J., Hicks, A.C., 2018. Interannual stability from ensemble modelling. Can. J. Fish. Aquat. Sci. 75 (12), 2109–2113.

Stewart, I.J. and Hicks, A.C. (2021). Assessment of the Pacific halibut (Hippoglossus stenolepis) stock at the end of 2020. Technical Report IPHC-2021-SA-01.

Stewart, I.J. and Hicks, A.C. (2022). Assessment of the Pacific halibut (Hippoglossus stenolepis) stock at the end of 2021. Technical Report IPHC-2022-SA-01, International Pacific Halibut commission.

Stewart, I.J. and Martell, S. (2014). Assessment of the Pacific Halibut stock at the end of 2013. Technical Report IPHC Report of Assessment and Research Activities 2013.

Stewart, I.J., Martell, S.J.D., 2015. Reconciling stock assessment paradigms to better inform fisheries management. ICES J. Mar. Sci. 72 (8), 2187–2196.

Takeuchi, Y., Pilling, G., and Hampton, J. (2017). Stock assessment of sowrdfish(Xiphias gladius) in the southwest Pacific Ocean. Technical Report WCPFC-SC13–2017/SA-WP-13, Western and Central Pacific Fisheries Commission: Scientific Committee, Rarotonga, Cook Islands.

Then, A.Y., Hoenig, J.M., Hall, N.G., Hewitt, D.A., 2015. Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. ICES J. Mar. Sci. 72 (1), 82–92.

Therneau, T. and Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees.

Thorson, J.T., 2020. Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. Fish Fish 21 (2), 237–251.

Thorson, J.T., Munch, S.B., Cope, J.M., Gao, J., 2017. Predicting life history parameters for all fishes worldwide. Ecol. Appl. 27 (8), 2262–2276.

Ting, K.M., Witten, I.H., 1999. Issues in Stacked Generalization. J. Artif. Intell. Res. 10, 271–289.

Tremblay-Boyer, L., Hampton, J., McKechnie, S., and Pilling, G. (2017). Stock assessment of South Pacific albacore tuna Rev 2 (29 July 2017). Technical Report WCPFC-SA14-SA-WP-05.

Vincent, M.T., Pilling, G., and Hampton, J. (2019). Stock assessment of skipjack tuna in the western and central Pacific Ocean (25July) - Rev.02. Technical Report WCPFC-SC15-SA-WP-05.

Vincent, M.T., Ducharme-Barth, N.D., Hamer, P., Hampton, J., Williams, P., and Pilling, G. (2020). Stock assessment of yellowfin tuna in the western and central Pacific Ocean (31July) - Rev.03. Technical Report WCPFC-SA16-SA-WP-04.

Walter, J. and Winker, H. (2019). Projections to create Kobe 2 Strategy Matrices using the multivariate log-normal approximation for Atlantic yellowfin tuna.In Collective Volume of Scientific Papers, volume 76 of Report of the 2019 ICCAT yellowfin tune stock assessment meeting, 725–739.International Commission for the Conservation of Atlantic tunas.

WCPFC-SC (2017). Thirteenth Regular Session of the Scientific Commitee: Summary Report. Technical report, The commision for the conservation and management of highly migratory fish stocks in the Wester and Central Pacific Ocean, Rarotonga, Cook Islands.

Winker, H., Kell, L., Fu, D., Sharma, R., Courtney, D., Carvalho, F., Schirripa, M., and Walter, J. (2019). A rapid approach to approximate Kobe posteriors from Stock Synthesis assessment models with applications to north Atlantic shortfin mako. Technical report, SCRS/2019/093.

Young, J. and Drake, A. (2002). Reproductive dynamics of broadbill swordfish (Xiphias gladius) in the domestic longline fishery off eastern Australia. Technical Report Project FRDC 1999/108, CSIRO.

Young, J. and Drake, A. (2004). Age and growth of broadbill swordfish (Xiphias gladius) from Australian waters. Technical Report FRDC Project 2001/014, CSIRO.

Young, J., Drake, A., Brickhill, M., Farley, J., Carter, T., 2003. Reproductive dynamics of broadbill swordfish, Xiphias gladius, in the domestic longline fishery off eastern Australia. Mar. Freshw. Res. 54 (4), 1–18.

Zhou, S., Punt, A.E., Lei, Y., Deng, R.A., Hoyle, S.D., 2020. Identifying spawner biomass per-?recruit reference points from life-?history parameters. Fish Fish. 21 (4), 760–773.