SCIENTIFIC COMMITTEE

SEVENTEENTH REGULAR SESSION

Online meeting

11-19 August 2021

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

WCPFC-SC17-2021/SA-WP-05

24 July 2021

Nicholas Ducharme-Barth[1] and Matthew Vincent[2]

[1] OFP (Oceanic Fisheries Programme)
Pacific Community (SPC), Noumea, New Caledonia
[2] NOAA National Marine Fisheries Service, Southeast Fisheries Science Center

Executive Summary

In the Western & Central Pacific Fisheries Commission (WCPFC), uncertainty in management reference points are derived from one of two stock assessment modeling approaches: 1) one that solely incorporates the statistical (estimation) uncertainty from a single "best" model, or 2) one that characterizes the model uncertainty across a model ensemble or structural (model) uncertainty grid. Either approach, when considered independently, is likely to under-represent the uncertainty in management reference points. However, these approaches are not mutually exclusive and can be combined to characterize uncertainty in a more holistic and transparent manner. We encourage the SC to recommend that combining both the statistical and structural uncertainty across an ensemble of models be the standard approach for characterizing uncertainty for all assessed stocks under the management of the WCPFC. We also point the SC to the 2021 southwest Pacific Ocean swordfish assessment which applies this approach to a WCPFC assessment for the first time. This work also responds to concerns expressed at SC15 that approaches are needed to reduce the requirement to make subjective decisions on model weighting in structural uncertainty grids and provides a strong basis for addressing this issue (below).

Adopting the approach of combining statistical and structural uncertainty across an ensemble necessitates having a sound framework for developing the ensemble. The principal criticisms levied against the structural uncertainty grid approach was that the choice of axis levels could be subjective, and that a clear approach for objectively weighting different models in the grid was lacking.

In the current paper, we describe a framework for creating an ensemble that addresses both of these criticisms, and demonstrate it using the 2017 southwest Pacific Ocean swordfish stock assessment as a case-study. This approach centers on developing a joint prior distribution for parameters that are fixed within an assessment model.

We invite the SC to:

- Recommend that the WCPFC considers adopting a standard approach for presenting uncertainty in management reference points and that the standard approach combines the statistical and structural uncertainty across an ensemble of models.
- Consider the merits of the framework outlined in this paper as a suitable approach for combining statistical and structural uncertainty across an ensemble of models for WCPFC assessments.
- Note the application of this framework in the 2021 southwest Pacific Ocean swordfish assessment.
- Support additional research into ensemble modeling and model weighting for the provision of management advice.

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

Nicholas D. Ducharme-Barth[1*], Matthew T. Vincent[2],

1 Pacific Community, Oceanic Fisheries Programme, BP D5, Noumea, New Caledonia 98848
2 NOAA National Marine Fisheries Service, Southeast Fisheries Science Center, 101 Pivers Island Road, Beaufort, North Carolina, USA 28516

* nicholasd@spc.int

Abstract

Uncertainty in population status estimates from stock assessments are important for providing a more comprehensive picture of current knowledge of a stock. The use of ensemble models to encapsulate model uncertainty has become increasingly prevalent. The incorporation of uncertainty of biological parameters that are often fixed in stock assessment models can be incorporated through model ensembles. An ensemble can be created by randomly drawing values from the likely parameter space (prior ensemble) or fixed at either a high, medium, or low value that encapsulates the variability in the parameter and applied in a fully factorial grid across the fixed parameters (factorial ensemble). We calculated the management advice from a prior ensemble and a factorial ensemble for Southwest Pacific Swordfish (*Xiphias gladius*) and compared reference points which incorporated model uncertainty only, model and estimation uncertainty, or both uncertainties weighted by sampling importance resampling. Median reference points were not significantly different for the two ensembles, but the factorial ensemble had a significantly larger estimate of model uncertainty than the prior ensemble. Stock assessments with fixed biological parameters can characterize uncertainty in these parameters more effectively using a prior ensemble approach. A factorial ensemble approach is appropriate for comparing different model structure assumptions and functional forms of relationships, but can be used in combination with a prior 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 overestimation of the precision of estimates. Further work is needed regarding appropriate weighting of ensembles which incorporate different data sources or have different likelihood weightings.

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 [1–4]. The complexity of these models has evolved and generally become greater

through time [5]; recent catch at age models include sex-specific dynamics, and/or spatially discrete areas with multiple stocks [6]. 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 there is not sufficient data 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 [7]. Characterization and incorporation of uncertainty into management advice is becoming more widespread as awareness of the magnitude of this uncertainty in stock assessments increases [8].

There are two main types of uncertainty that afflict fisheries management: scientific uncertainty and management uncertainty [8]. This study focuses on the former, while the latter can be addressed through Management Strategy Evaluation (MSE) [9]. Scientific uncertainty arises due to imprecision and bias in the stock assessment process, which 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. Third, 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. Fourth, estimation uncertainty is 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.

Historically, point estimates of stock status from a single model were used to provide management advice and did not quantify the uncertainty in the estimates [10, 11]. Incorporation of estimation uncertainty into management advice first occurred as a result of greater computational abilities to estimate variance of model quantities using the covariance matrix and the delta method [12–14]. Monte Carlo and bootstrap simulations have also been used to estimate uncertainty in estimated quantities for use in management [15, 16]. However, estimation uncertainty from a single model is now generally thought to be modest compared to model uncertainty representing different states of nature [17]. Use of ensemble model methods and superensembles [18] have led to the expansion of incorporating uncertainty from multiple models into management advice in recent years [17, 19–21]. 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 [22, 23].

It is important to consider how model ensembles are combined to provide management advice, because the chosen methodology can influence 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 weighted average, 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 [24]. Equally important to how management advice is presented and combined, is the choice of models included within an ensemble [17, 19, 24]. 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 incorporating 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 methodologies to encapsulate uncertainty in biological parameters (e.g., natural mortality or growth) that are fixed within the assessment models.

The first methodology to incorporate uncertainty into an ensemble is to randomly draw values of biological parameters from distributions obtained by external analyses [15–17,25]. This methodology is similar to creating prior distributions for parameters within a Bayesian framework. Despite recent advances in algorithms for mapping the posterior distributions [26–29], Bayesian analyses remain computationally infeasible for use in complex age-structured stock assessments. However, this should not prevent the creation of prior distributions in order to incorporate biological model uncertainty within ensembles fit using maximum likelihood approaches. These prior distributions can be formulated using a range of methodologies; the simplest approach would use the estimate of uncertainty from a single study (e.g., growth curve estimate), whereas a more complex method would be a meta-analytic approach of numerous studies on similar species such as the one implemented in the R package FishLife [30,31]. The second methodology to incorporate model uncertainty into an ensemble involves bounding the uncertainty, typically using the associated 95% confidence (credible) interval, of a fixed parameter. 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. This application of model ensembles has been commonly used by the Western Central Pacific Fisheries Commission (WCPFC) to formulate management advice [32,33]. Uncertainty in biological parameters is incorporated into management advice by combining with other model (structural) uncertainties in a full factorial combination of "axes of uncertainty".

A comparison of these two methodologies has not been conducted on a set of biological parameters, but theoretically the prior distribution method is superior in many aspects. First, the full factorial method 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 prior 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 prior 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 method can quickly become computationally impractical to conduct beyond 3-5 axes of uncertainty[1]. This computational restriction compels the analyst to triage the potential sources of uncertainty, effectively ignoring the impact of those sources of uncertainty deemed to be less important. In theory, the range of uncertainty from the prior approach could be characterized in a more computationally efficient manner using a smaller model ensemble (25-40 models) 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 can impact the resulting management advice. This is accomplished by addressing the following five objectives using the 2017 southwest Pacific Ocean (SWPO)

[1]This corresponds to an ensemble size of 27 - 243 models assuming 3 levels (high, median, and low) per "axis"

swordfish (*Xiphias gladius*) stock assessment as a case study: (i) we demonstrate the difference in management advice arising from creating a model ensemble using the full factorial and prior distribution approaches; (ii) using the prior distribution approach we evaluate the number of models needed to characterize the model uncertainty; (iii) we show how the prior distribution 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; and (v) 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.

Methods

Case study description

Details of the SWPO swordfish stock assessment are presented in [32] and we refer readers to this for a complete description of the model. For context, the 2017 SWPO swordfish stock assessment was conducted using the integrated assessment platform Multifan-CL [4], using data from 1952-2015. 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 uses 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. Using this model as a baseline, we investigated model uncertainty in five different biological assumptions that were fixed within the 2017 stock assessment. For four of these assumptions, there was sufficient data to conduct external analysis to estimate relationships to be used in our ensembles: growth, natural mortality (M), length-weight relationship, and maturity (or spawning-potential) at length relationship. There was not sufficient data to conduct an external analysis for the stock-recruitment steepness parameter, which was the fifth biological uncertainty in the ensemble.

Ensemble construction

Prior approach

The methods and data used to create the joint prior are of limited importance to the conclusions drawn in this study and could be created through a variety of methods depending on the species. For example, estimates of uncertainty in the biological parameters could be created using the FishLife package [30], or any other multivariate or meta-analytic approach [34]. Briefly, we describe the methods and data used in the current analysis to create the joint prior, however we urge readers to consult the 2021 southwest Pacific Ocean swordfish stock assessment data inputs paper [35] for further information and detail. Four independent Bayesian analyses, implemented in R using the STAN package [26, 36] 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 [37–39], though the aging and histological data come from a subsequent re-analysis [40]. 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 500 samples without replacement from each of the independent posteriors. These samples were then used to calculate a prior 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 [41–43]. Variability in the parameters in the $\text{Pauly}_{\text{nls-T}}$ relationship [41] was included when calculating the natural mortality by drawing from their associated covariance matrix. This methodology 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 uniform distribution between 0.65 and 0.95 which matched the range from the previous assessment [32]. The resulting distributions of the biological relationships for the joint-prior ensemble are presented in Fig 1. The prior ensemble approach created 500 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.

Multiple prior ensembles were created in order to investigate how uncertainty in management reference points changed with ensemble size. Thus, the 500 models from the prior ensemble were sampled without replacement to create new ensembles with sample sizes of 350, 243, 100, 75, 50, and 30. The sample size of 243 was chosen because this was the size of the factorial ensemble approach (see below). The 243 model prior ensemble was used in comparison with the factorial ensemble.

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 prior ensemble. The two MSY based reference points [2], $\text{SB}/\text{SB}_{\text{MSY}}$ and $\text{F}/\text{F}_{\text{MSY}}$, show terminal spawning biomass (SB) and fishing mortality (F) relative to the SB or F that produces MSY. The depletion based reference point, $\text{SB}/\text{SB}_{\text{F=0}}$, shows terminal SB relative to the unfished SB in the terminal year. A generalized linear model of each reference point was created for the prior ensemble with the 243 models. The model included all fixed biological parameters as covariates, scaled to a mean of 0 and standard deviation of 1, against the selected reference point. The effect and p-value of the regression were displayed graphically to determine which biological relationship was most influential in stock status reference point estimates.

Factorial approach

The factorial approach typically assigns a high, medium and low value to be used for each axis of uncertainty in the model ensemble. To this end, the 2.5, 50, and 97.5% percentile from the joint prior were calculated for the growth, length-weight, spawning potential, and natural mortality. The values of the biological relationship used in the factorial ensembles are shown as the dotted and dashed lines in Fig 1. The value for steepness was assumed to be either 0.65, 0.8, or 0.95. 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 prior ensemble.

[2] MSY is based on the average fishing mortality at age in the last 5 years of the model, excluding the last year

Ensemble Comparisons

Distributions of reference points were compared among the prior ensembles and the factorial ensemble by boxplots of the converged and non-converged models. Convergence was determined based on the presence of a positive definite Hessian solution. Pairwise Wilcox tests for each ensemble were conducted to compare mean estimates of reference points. Flinger tests were conducted between the ensembles to determine differences in variance estimates of reference points.

Estimates of uncertainty for the two ensembles were calculated through three 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 point estimates from converged models in each ensemble. The second method incorporates both the model and estimation uncertainty, and follows the approach taken by the International Pacific Halibut Commission (IPHC) [20, 44]. For each model retained in the ensemble, the estimation uncertainty for the three reference points was generated by drawing 100 samples in a parametric bootstrap from a distribution, where the mean was the estimate and the standard deviation the estimated standard error computed using the delta method applied to the variance-covariance matrix of the model parameters [12]. The MSY based reference points were drawn from a normal distribution. However, the $SB/SB_{F=0}$ reference point was estimated on a natural logarithm-scale in MULTIFAN-CL, therefore a lognormal distribution was used for this parameter. The third method uses the model uncertainty but weights each model through sampling importance resampling (SIR) [45]. To conduct SIR, 8000 models were drawn 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 1% [45]. The fourth methodology is similar to the previous but incorporates both measures of uncertainty and weights the models through SIR described above. From each sampled model in the SIR 100 values were drawn for each reference point from the approximation based on the estimated standard error.

Results

All models achieved the specified maximum gradient component stopping criteria of 10^{-3}. The convergence rate of the full factorial ensemble (138 of 243 models; 57%) was marginally lower than the rate for the prior ensemble (147 of 243 models; 60%). The full factorial ensemble had problems with convergence for models with the combination of low natural mortality, low growth, and high or medium length-weight. There were not any specific parameters or combinations that resulted in poor convergence for the ensemble models. The estimates of $SB/SB_{F=0}$ from models in the full factorial ensemble with a positive definite hessian had a median estimate of 0.374 and an interquartile range of 0.030, whereas models without a positive definite hessian had a lower median of 0.347 and a larger IQR of 0.035. For the fully factorial ensemble, the F/F_{MSY} was marginally higher for the models without a positive definite Hessian. Conversely, SB/SB_{MSY} was marginally lower for models without a positive definite Hessian compared to models that had converged (Fig 2).

Sample size of the prior ensemble over the range investigated did not have a large influence on the reference point estimates both in terms of the median and the variability (Fig 2). The median and interquartile range of reference points were similar for models that obtained a positive definite Hessian and those that did not for all prior ensembles. For the converged prior ensembles for all sample sizes, median F/F_{MSY}

ranged between 0.767 and 0.818, median SB/SB_{MSY} ranged between 1.633 and 1.757, and median $SB/SB_{F=0}$ ranged between 0.358 and 0.362.

Pairwise Wilcox tests on the mean of the three reference points across all ensembles were not significantly different (P-value >0.1). A Flinger test between the factorial ensemble and all sample sizes of the prior ensemble showed that the variance was significantly larger for the factorial ensemble (P-value <0.001). Flinger tests among the sample sizes of the prior ensemble were not significantly different.

Reference point estimates from models with a positive definite Hessian from the prior ensemble with 243 models were regressed against the input biological parameters for the model (Fig 3). SB/SB_{MSY} was strongly positively correlated with steepness (P-value <0.001). The hypothetical age where length is zero (t0), the Brody growth coefficient (k), both length-weight relationship parameters (LW-a and LW-b), and the length at 50% maturity (L_{50}) were strongly correlated with $SB/SB_{F=0}$ (P-value <0.001). The k parameter had the largest positive coefficient and t0 was a similar magnitude but negative for the $SB/SB_{F=0}$ regression. The length weight parameters had the second largest positive magnitude whereas L_{50} had a negative coefficient but was not very large. $SB/SB_{F=0}$ was also correlated with steepness and the slope of the maturity function (P-value <0.01) and weakly correlated with L_{inf} and the asymptotic value of the sex ratio (P-value <0.1). F/F_{MSY} was strongly negatively correlated with steepness, the length weight parameters (LW-a and LW-b), and k, but was strongly positively correlated with t0 and L_{50} (P-value <0.001). The reference point F/F_{MSY} was also positively correlated with the slope of the maturity function (P-value <0.01).

Uncertainty in estimates of the reference points for both the factorial and prior ensembles were influenced depending on whether they included model uncertainty, model and estimation uncertainty, or both uncertainties with SIR. The incorporation of estimation uncertainty with the model uncertainty predictably resulted in less precise estimates for all reference points and both ensembles (Fig 4).

The factorial ensemble distribution for SB/SB_{MSY} was distinctly bimodal with peaks around 1.75 and 4 when only model uncertainty was used (Fig 4). When estimation uncertainty was incorporated into the density, the mode around 4 became much less influential and the overall distribution became more similar to the density for the prior ensemble. The distribution of SB/SB_{MSY} for the SIR for model uncertainty only had three models for the factorial ensemble and did not have a smooth distribution for the prior ensemble. The distribution from the SIR with both estimation and model error was nearly identical to the estimation and model error for both ensembles.

The distribution of $SB/SB_{F=0}$ for the prior ensemble with only model uncertainty was less variable than the factorial distribution and had a mode at a slightly lower value (Fig 4). When estimation and model uncertainty were incorporated into $SB/SB_{F=0}$, the estimates of this reference point became much more uncertain, but more similar between the two ensembles. The distributions for the SIR showed very similar distributions to the corresponding errors without resampling.

The density of F/F_{MSY} for the factorial ensemble contained three modes when only model uncertainty was accounted for and was more variable than the prior ensemble (Fig 4). The density of $F/F_{MSY}$with model and estimation uncertainty for the factorial ensemble was bimodal but remained more variable than the prior ensemble. The density curve for the sampling importance resampling for model uncertainty displayed a jagged density curves and the number of distinct modes increased for both ensembles. The distribution of F/F_{MSY} from SIR with estimation and model uncertainty showed some smoothing of the mode for both ensembles compared to the estimation and model error distributions but overall were very similar.

The probability of reference points exceeding their respective chosen values [3] was

[3]These limits have not been agreed to for management purposes and are for illustrative purposes

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. 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 5%. 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 SB/SB_{MSY} and F/F_{MSY} than the prior ensemble.

Discussion

In this study we investigated the difference 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 not statistically different between the two ensembles (or for prior ensembles of different sizes), but the ensemble with the full factorial design showed more uncertainty. The higher variance and bimodality seen in the factorial ensemble could lead to different management advice depending of 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 prior approach was computationally more efficient as reference point estimates (and associated model uncertainty) were consistent across prior ensembles of varying sizes. Therefore,in this case a prior ensemble could be created with as few as 30 models to capture uncertainty in fixed biological parameters used within the assessment. Therefore, we recommend creating an ensemble of models that draws fixed biological parameters from a prior distribution.

An additional advantage of using the prior ensemble over the fully factorial approach is the ability to regress the fixed model covariates against the reference points. This can be useful to identify which parameters are influential on the model results. 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 SWPO swordfish case study, a better understanding of the steepness of the Beverton-Holt stock recruitment would reduce the variability in MSY based reference points. Conversely, better understanding of the growth and length-weight parameters would reduce variability in the $SB/SB_{F=0}$ reference point the most.

The prior ensemble approach could be applied to the specification of operating models within a MSE framework [9]. However, a MSE framework does not need to be in place for an ensemble approach to be used. Additionally, even if current management decisions do not incorporate uncertainty in reference point estimates, the presentation of uncertainty in the reference points to managers should nevertheless occur. 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 have been incorporated into management advice for numerous species under federal jurisdiction in the southeast United States using a Monte Carlo bootstrap ensemble (MCBE) [16, 25]. 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. Therefore, the methodology applied in this paper can be used to provide robust advice to fisheries managers without a full MSE framework.

Despite our recommendation to use ensembles from prior distributions, a fully factorial ensemble is a valid and warranted approach for creating ensembles in some scenarios. A fully factorial ensemble design should be used when there are discrete choices between model structures that cannot be characterized as a distribution. A good example of a fully factorial axis of uncertainty would be models with differing 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 would be applied could include the functional forms of selectivity, spatial structure assumptions of the assessment, alternative catch reconstruction time series and different standardization approaches of CPUE indices. Model ensembles that are a hybrid between the factorial and prior approach 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 prior for each of the axes or models in the factorial design.

Uncertainty in management advice is generally thought to be greater from model uncertainty than it is from estimation uncertainty [17, 24]. However, this was not the case for all reference points that are presented in this study. For the MSY based reference points, the CV from model uncertainty was on the same scare or slightly larger than the CV from estimation uncertainty (Table 2). However, for some individual models the standard error for these reference points was larger than the standard deviation of the model estimates. Conversely, the CV of $SB/SB_{F=0}$ from the model estimates was an order of magnitude smaller than the CV from the estimation error. Thus in the current case, if only model uncertainty from the ensemble were used in the creation of management advice for this reference point, then the uncertainty would be underrepresented and could lead to risk prone management. The influence of including estimation error will depend on the precision of the reference point estimates within the assessment model, but will generally result in an increase in uncertainty. Therefore, the incorporation of both model and estimation uncertainty into management advice is necessary to accurately capture the current knowledge of stock status.

The reference point estimates from an ensemble can be combined through a multitude of techniques. These methods can range from simple averaging, likelihood weighting (e.g., AIC), or cross validation [17]. Simple averaging of reference points can easily be conducted for a large number of models and can incorporate both estimation and model uncertainty [46]. Simply averaging across models in a factorial ensemble is implicitly assuming equal probability of the states of nature represented by all models [24]. The prior method 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 prior ensembles. Conversely, the full factorial method 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 prior method (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 fully factorial ensemble. Assigning weights

to various hypotheses in a fully factorial ensemble is difficult and typically is resolved through 'expert opinion'. 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 are 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 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 and will be driven more by the 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. Likelihood weighting methods have been proposed as an alternative 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 [21]. 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 computationally intensive and thus may not be practical for models that take a long time to converge or for large ensembles [47]. Recent work using hindcasting has proposed the use of mean absolute scaled error (MASE) as potential method for model ensemble weighting [48]. However, the details of performing such weighting need additional evaluation for determining which data source should be removed for hindcasting or whether all data sources need to individually hindcast; additionally, investigation is needed in the correct way to combine the metric across multiple CPUE indices or metrics from removing different data sources. If estimation uncertainty and model uncertainty are at a similar scale, then the different weighting methodologies will generally produce similar results if both model uncertainty and estimation uncertainty are incorporated into management advice. This was seen in this study when using the sampling importance resampling. Management advice is only likely to be significantly different if model uncertainty is much larger than the estimation uncertainty and model weighting removes models from the tails of the distribution. However, it is always possible that different weighting methods chosen could allow/prevent management criteria based on probability of exceeding a reference point from being activated. Thus, further research on the best method for ensemble averaging is required.

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

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 J. Farley (Commonwealth Scientific and Industrial Research Organization, CSIRO) for aiding us

in assembling the data required to create the biological prior for the swordfish case study. We thank C. Monnahan for assistance with the STAN modeling and discussions on priors, posteriors, and combining model uncertainty. We also thank Paul Hamer, John Hampton, and members of the Stock Assessment and Modeling (SAM) Team in the SPC-Oceanic Fisheries Programme for reviewing drafts of the manuscript and serving as a sounding board for ideas.

References

1. Fournier D, Archibald C. A general theory for analyzing catch at age data. Canadian Journal of Fisheries and Aquatic Sciences. 1982;39(8):1195 – 1207. doi:10.1139/f82-157.

2. Deriso RB, Quinn T. Catch-Age analysis with auxiliary information. Canadian Journal of Fisheries and Aquatic Sciences. 1985;42(4):815 – 824. doi:10.1139/f85-104.

3. Richard D Methot J, Wetzel CR. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research. 2013;142:86 – 991. doi:10.1016/j.fishres.2012.10.012.

4. Fournier DA, Hampton J, Sibert JR. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, Thunnus alalunga. Canadian Journal of Fisheries and Aquatic Sciences. 1998;55:2105 – 2116. doi:10.1139/F98-100.

5. Hilborn R. The state of the art in stock assessment: Where we are and where we are going. Scientia Marina. 2003;67(S1):15 – 20. doi:10.3989/scimar.2003.67s115.

6. Maunder MN, Piner KR. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fisheries Research. 2017;192:16 – 27. doi:10.1016/j.fishres.2016.04.022.

7. Minte-Vera CV, Maunder MN, da Silva AMA, Satoh K, Uosaki K. Get the biology right, or use size-composition data at your own risk. Fisheries Research. 2017;192:114 – 125. doi:10.1016/j.fishres.2017.01.014.

8. Privitera-Johnson KM, Punt AE. A review of approached to quantify uncertainty in fisheries stock assessment. Fisheries Research. 2020;226. doi:10.1016/j.fishres.2020.105503.

9. Punt AE, Butterworth DS, de Moor CL, De Oliveira JAA, Haddon M. Management strategy evaluation: best practices. Fish and Fisheries. 2016;17(2):303–334. doi:https://doi.org/10.1111/faf.12104.

10. Haddon M. Quantitative methods in fisheries. Chapman and Hall; 2001.

11. Quinn TJ, Deriso RB. Quantitative Fish Dynamics. Oxford University Press; 1999.

12. Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, et al. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software. 2012;27(2):233 – 249. doi:10.1080/10556788.2011.597854.

13. Magnusson A, Punt AE, Hilborn R. Measuring uncertainty in fisheries stock assessment: The delta method, bootstrap, and MCMC. Fish and Fisheries. 2012;14(3):325 – 342. doi:10.1111/j.1467-2979.2012.00473.x.

14. Maunder MN, Piner JR. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science. 2015;72(1):7 – 18. doi:10.1093/icesjms/fsu015.

15. Restrepo VR, Hoenig JM, Powers JE, Baird JW, Turner SC. A simple simulation approach to risk and cost analysis, with applications to swordfish and cod fisheries. Fishery Bulletin. 1992;90(4):736 – 748.

16. Legault CM, Powers JE, Restrepo VR. Mixed Monte Carlo/Bootstrap Approach to Assessing King and Spanish Mackerel in the Atlantic and Gulf of Mexico: Its Evolution and Impact. In: Berkson JM, Kline LK, Orth DJ, editors. Incorporating uncertainty into fishery models. vol. Symposium 27. American Fisheries Society; 2002. p. 208.

17. Scott F, Jardim E, Millar CP, Cervino S. An applied framework for incorporating multiple sources of uncertainty in Fisheries Stock Assessments. PLoS ONE. 2016;11(5):e0154922. doi:10.1371/journal.pone.0154922.

18. Anderson SC, Cooper AB, Jensen OP, Minto C, Thorson JT, Walsh JC, et al. Improving estimates of population status and trend with superensemble models. Fish and Fisheries. 2017;doi:10.1111/faf.12200.

19. Brodziak J, Piner K. Model averaging and probable status of North Pacific striped marlin, Tetrapturus audax. Canadian Journal of Fisheries and Aquatic Sciences. 2010;67(5):793 – 805. doi:10.1139/F10-029.

20. Stewart I, Martell S. Assessment of the Pacific Halibut stock at the end of 2013; 2014. IPHC Report of Assessment and Research Activities 2013.

21. Jardim E, Azevedo M, Brodziak J, Brooks EN, Johnson KF, Klibansky N, et al. Operationalizing ensemble models for scientific advice to fisheries management. ICES Journal of Marine Science. 2021;doi:10.1093/icesjms/fsab010.

22. Stewart IJ, Martell SJD. Reconciling stock assessment paradigms to better inform fisheries management. ICES Journal of Marince Science. 2015;72(8):2187 – 2196. doi:10.1093/icesjms/fsv061.

23. Stewart IJ, Hicks AC. Interannual stability from ensemble modelling. Canadian Journal of Fisheries and Aquatic Sciences. 2018;75(12):2109–2113. doi:10.1139/cjfas-2018-0238.

24. Maunder MN, Xu H, Lennert-Cody CE, Valero JL, da Silva AA, Minte-Vera C. Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. San Diego, California, USA: Inter-American Tropical Tuna Commission: Scientific Advisory Committee; 2020. Document SAC-1 INF-F REV 3.

25. SEDAR. SEDAR 73 South Atlantic Red Snapper Stock Assessment Report. North Charleston SC: SEDAR; 2021.

26. STAN Development Team. STAN Modeling Language Users Guide and Reference Manual; 2021. `https://mc-stan.org`.

27. Monnahan CC, Branch TA, Thorson JT, Stewart IJ, Szuwalski CS. Overcoming long Bayesian run times in integrated fisheries stock assessments. ICES Journal of Marine Science. 2019;76(6):1477 1488. doi:10.1093/icesjms/fsz059.

28. Monnahan CC, Kristensen K. No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. PLoS ONE. 2018;13(5):e0197954. doi:10.1371/journal.pone.0197954.

29. Monnahan CC, Thorson JT, Branch TA. Faster Estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. Methods in Ecology and Evolution. 2016;8(3):339 – 348. doi:10.1111/2041-210X.12681.

30. Thorson JT, Munch SB, Cope JM, Gao J. Predicting life history parameters for all fishes worldwide. Ecological Applications. 2017;27(8):2262 – 2276. doi:10.1002/eap.1606.

31. Thorson JT. Predicting recruitment density dependence and instrinsic growth rate for all fishes workldwide using a data-integrated life-history model. Fish and Fisheries. 2020;21(2):237 – 251. doi:10.1111/faf.12427.

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

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

34. Horswill C, Kindsvater HK, Juan-Jorda MJ, Dulvy NK, Mangle M, Matthiopoulos J. Global reconstruction of life-history strategies: A case study using tunas. Journal of Applied Ecology. 2019;56(4):855 – 865. doi:10.1111/1365-2664.13327.

35. Ducharme-Barth N, Peatman T, Hamer P. Background analyses for the 2021 stock assessment of Southwest Pacific swordfish; 2021. WCPFC-2021-SC17-SA-IP-07.

36. R Core Team. R: A Language and Environment for Statistical Computing; 2021. Available from: `https://www.R-project.org/`.

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

38. Young J, Drake A, Brickhill M, Farley J, Carter T. Reproductive dynamics of broadbill swordfish, *Xiphias gladius*, in the domestic longline fishery off eastern Australia. Marine and Freshwater Research. 2003;54(4):1–18.

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

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

41. Then AY, Hoenig JM, Hall NG, Hewitt DA. Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. ICES Journal of Marine Science. 2015;72(1):82 – 92. doi:10.1093/icesjms/fsu136.

42. Lopez-Quintero FO, Contreras-Reyes JE, Wiff R. Incorporating uncertainty into a length-based estimator of natural mortality in fish populations. Fishery Bulletin. 2017;doi:10.7755/FB.115.3.6.

43. Lorenzen K. Allometry of natural mortality as a basis for assessing optimal release size in fish-stocking programs. Canadian Journal of Fisheries and Aquatic Sciences. 2000;57:2374 – 2381. doi:10.1139/f00-215.

44. Stewart I, Hicks A. Assessment of the Pacific halibut (*Hippoglossus stenolepis*) stock at the end of 2020; 2021. IPHC-2021-SA-01.

45. McAllister MK, Ianelli JN. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences. 1997;54(2):284 – 300. doi:10.1139/f96-285.

46. Ianelli J, Holsman KK, Punt AE, Aydin K. Multi-model interence for incorporating trophic and climate uncertainty into stock assessment. Deep Sea Research Part II: Topical Studies in Oceanography. 2016;doi:10.1016/j.dsr2.2015.04.002.

47. Maunder MN, Harley SJ. Using cross validation model selection to determine the shape of nonparameteric selectivity curves in fisheries stock assessment models. Fisheries Research. 2011;110(2):283 – 288. doi:10.1016/j.fishres.2011.04.017.

48. Kell LT, Sharma R, Kitakado T, Winker H, Mosqueira I, Cardinale M, et al. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science. 2021;doi:10.1093/icesjms/fsab104.

Tables

Table 1. Probability reference points exceeding limits by error type. The percent of the reference points (SB/SB_{MSY}, $SB/SB_{F=0}$, and F/F_{MSY}) exceeding their respective limits for the factorial and prior ensembles under the error distributions of model only, model and estimation, and both weighted by sampling importance resampling (SIR).

	SB/SB_{MSY} <1		$SB/SB_{F=0}$ <0.3		F/F_{MSY} >1	
Error Type	Factorial	Prior	Factorial	Prior	Factorial	Prior
Model	0	0	0	0	12.3	6.1
Model and Estimation	5.4	4.9	28.5	30.2	16.5	13.7
SIR	5.4	5.1	27.8	30.6	16.4	13.3

Table 2. Model and Estimation Uncertainty for Ensembles. Model uncertainty quantified by the coefficient of variation (CV) of reference points estimated from a factorial and prior ensemble. Estimation uncertainty was quantified as the median (maximum) of the CV estimated from the models in the ensembles for each reference point.

	SB/SB_{MSY}		$SB/SB_{F=0}$		F/F_{MSY}	
Ensemble	Model	Estimation	Model	Estimation	Model	Estimation
Factorial	0.45	0.22 (0.25)	0.069	0.22 (0.26)	0.28	0.14 (0.17)
Prior	0.28	0.23 (0.26)	0.032	0.23 (0.26)	0.18	0.15 (0.17)

Figures

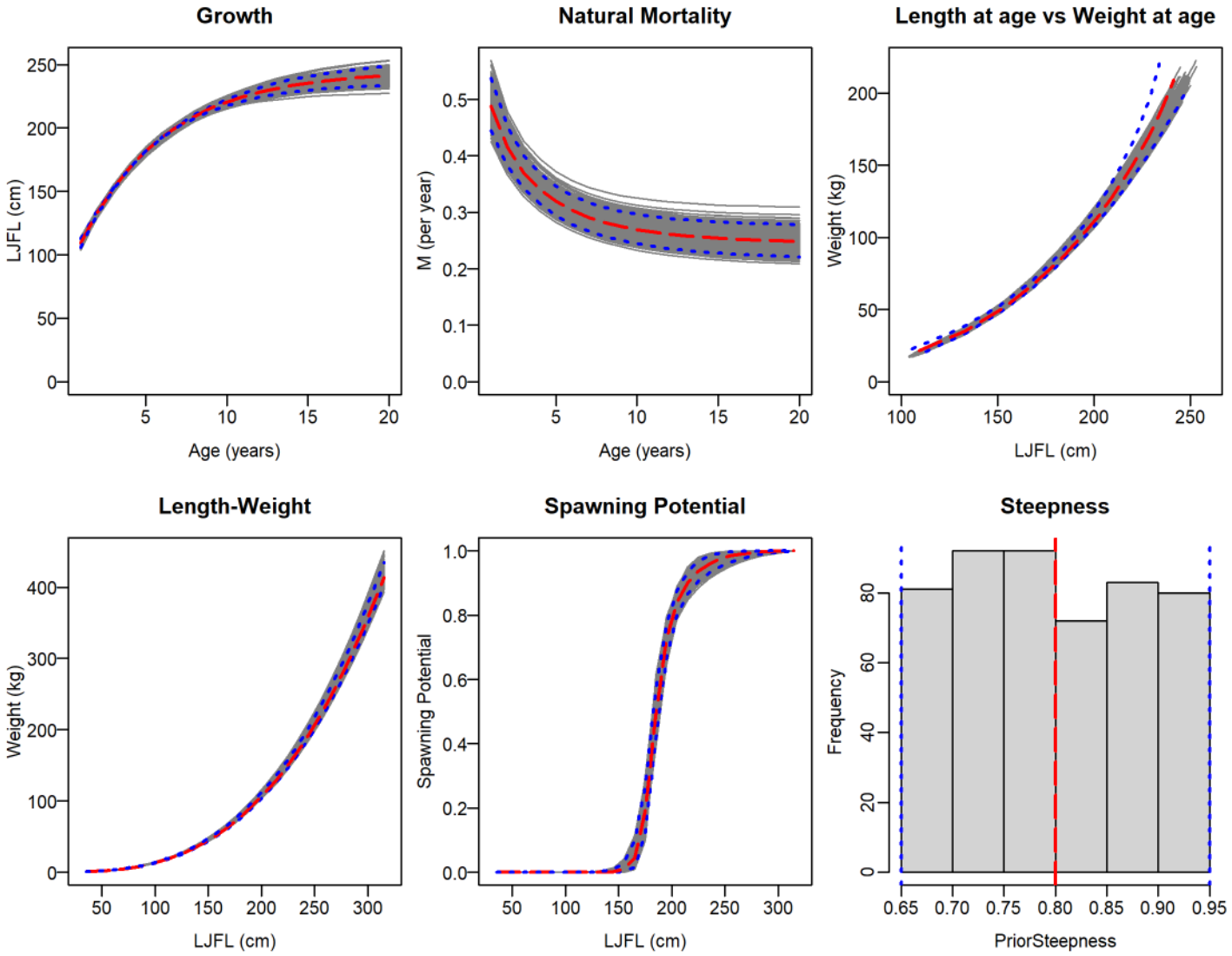

Fig 1. Prior distributions of biological parameters. Plot of biological relationships assumed within the ensembles where the solid grey lines are from the prior ensemble, the red dashed line is the median used in the factorial ensemble and the two blue dotted lines are the 95% confidence interval. Top left: growth relationship, 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, where LJFL is lower jaw fork length.

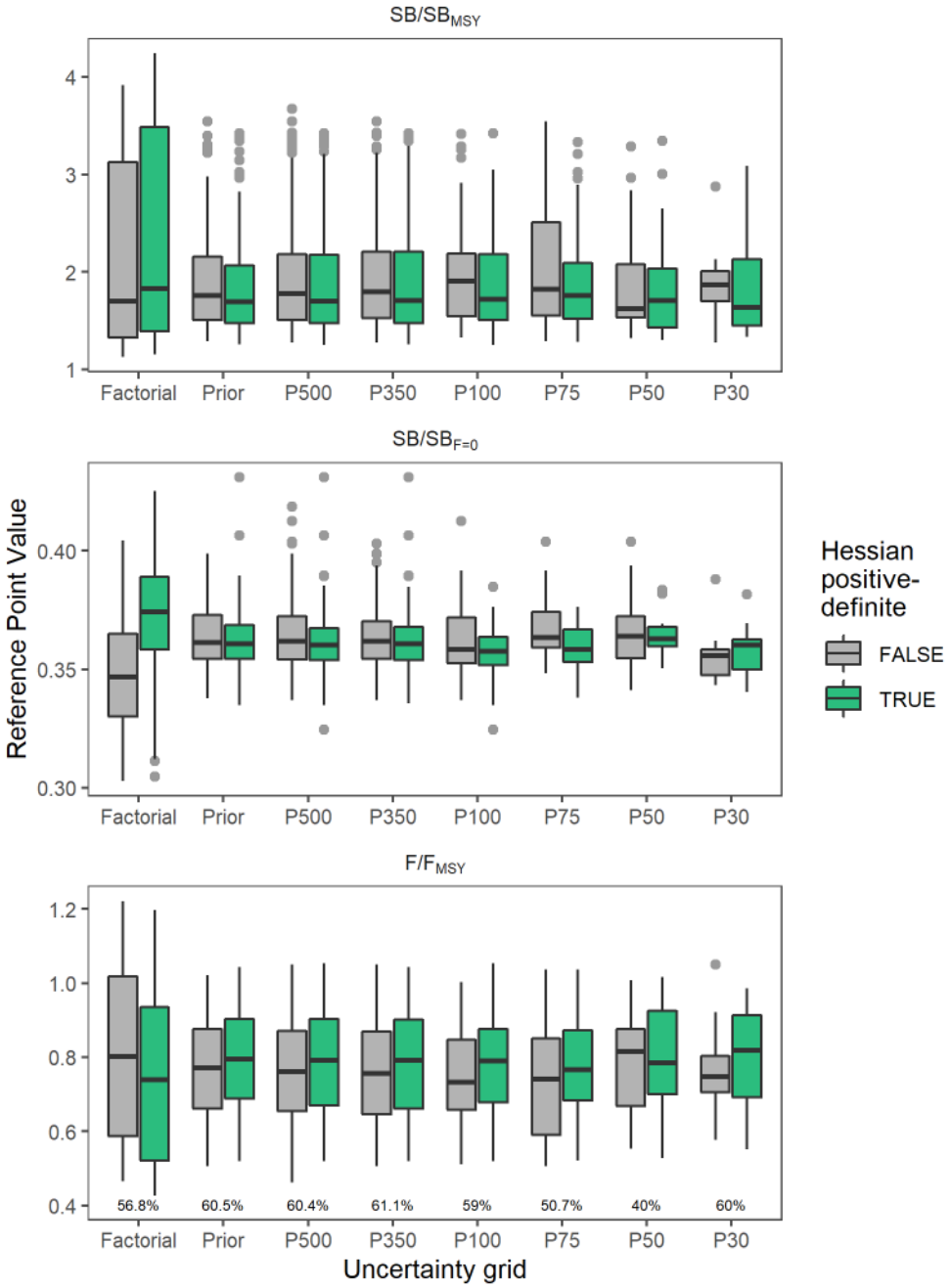

Fig 2. Reference points from ensembles. Boxplots of reference points from a full factorial ensemble and a prior ensemble with different sample sizes given by P and the number of the sample size, where the Prior ensemble has the same number of models as the factorial ensemble (243). The boxes indicate 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.

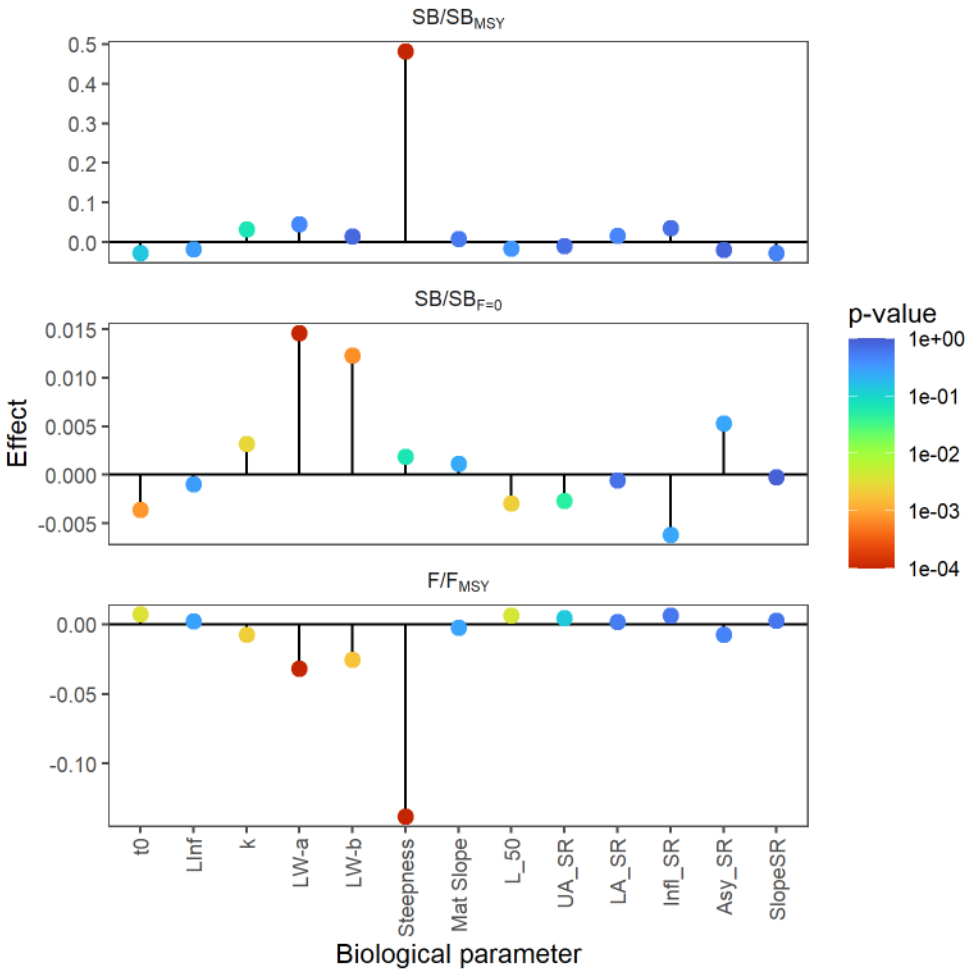

Fig 3. Covariate estimates from generalized linear models. Estimated covariates from a generalized linear model of biological parameters against reference points where the color of the point is the p-value of the effect and warmer colors indicate greater significance.

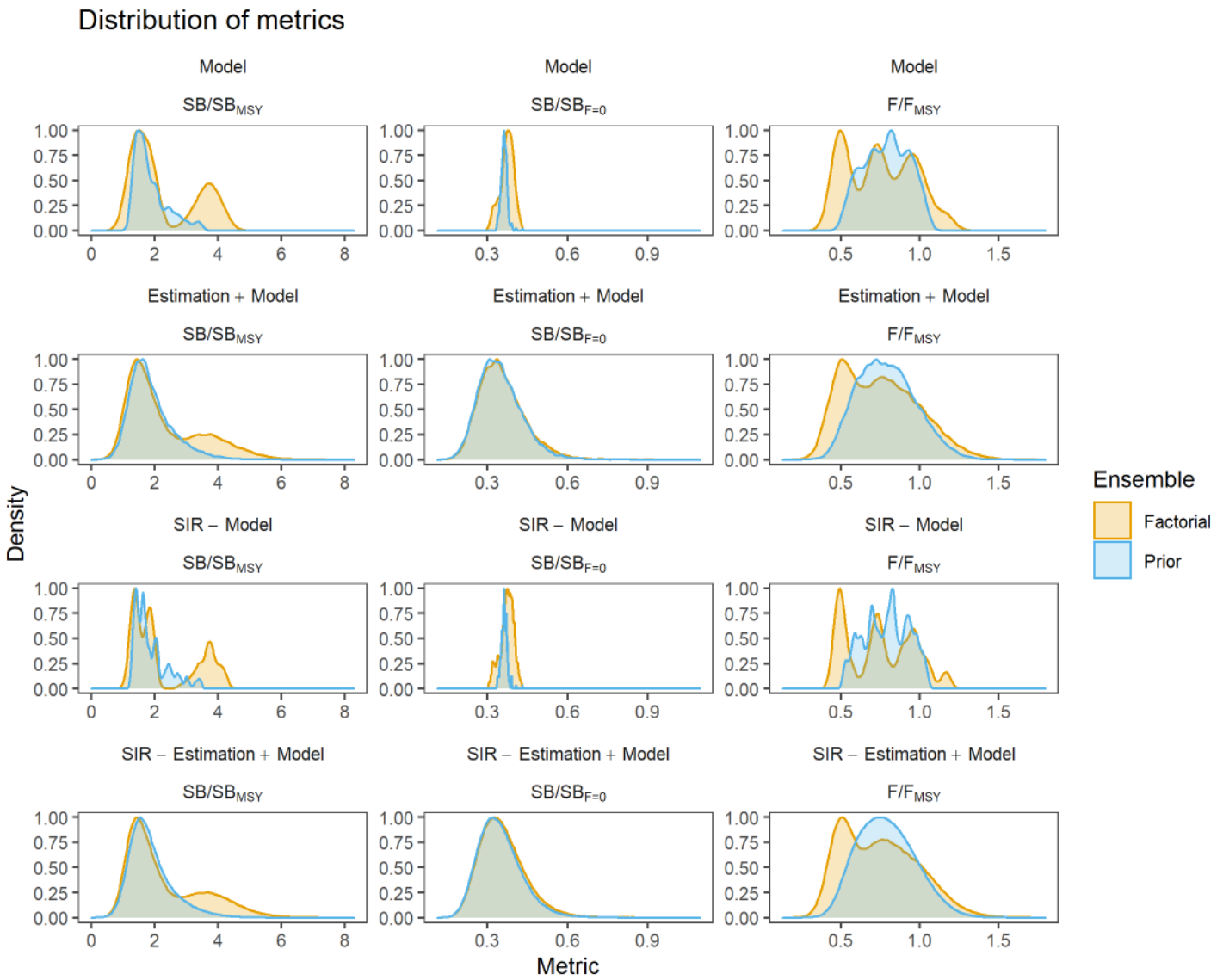

Fig 4. Error distributions of reference points. Estimated distributions of reference points for two ensembles where the left column is SB/SB_{MSY}, the center column is $SB/SB_{F=0}$, and the right column is F/F_{MSY}, the top row is for model uncertainty only, the center row is the model and estimation uncertainty and the bottom row is for the sampling importance resampling.