SCIENTIFIC COMMITTEE
NINETEENTH REGULAR SESSION

Koror, Palau

16–24 August 2023

Addressing uncertainty in WCPFC stock assessments: Review and recommendations from WCPFC Project 113

WCPFC-SC19-2023/SA-WP-12
28 July 2023

Philipp Neubauer[1], **Kyuhan Kim**[1], **Teresa A'mar**[1] **and Kath Large**[1]

[1] **Dragonfly Data Science, Wellington, Aotearoa New Zealand**

Addressing uncertainty in WCPFC stock assessments

Review and recommendations from WCPFC Project 113

Authors:
Philipp Neubauer
Kyuhan Kim
Teresa A'mar
Kath Large

DRAGONFLY
Data Science

Cover Notes

To be cited as:

Neubauer, Philipp; Kim, Kyuhan; A'mar, Teresa; Large, Kath (2023). Addressing uncertainty in WCPFC stock assessments, 67 pages. WCPFC-SC19-2023/SA-WP-12. Report to the WCPFC Scientific Committee. Nineteenth Regular Session, 16–24 August 2023.

CONTENTS

EXECUTIVE SUMMARY

Model weighting is a central challenge in stock assessments, because the retention or rejection of models, and relative weights given to models and their respective uncertainties can markedly affect quantities measuring risk of available management options. In the Western and Central Pacific Fisheries Commission (WCPFC), there are currently no explicit terms of reference that guide the development and subsequent weighting of model ensembles. For this reason, a range of approaches for developing stock assessment models and model ensembles have been employed, ranging from single base-case stock assessments with relatively few key sensitivities to grid-based ensembles with a large number of models; other assessments employed intermediate approaches that considered a limited number of models.

This research for WCPFC project 113 aimed to provide both general and specific review components to develop recommendations for the presentation of stock assessment and management advice uncertainty by the WCPFC scientific committee. The terms of reference for the general review were to:

1. Review and summarise the different approaches used for characterising uncertainty in WCPFC stock assessments for tuna, billfish and sharks over the last five years.
2. Describe how uncertainty was communicated in the context of management risks and its influence on decision-making processes used by the WCPFC.
3. Comment on the suitability of the recent approaches to characterising uncertainty for the management systems, including the harvest strategy approach.

The specific review aimed to:

1. Critically review the ensemble approach that was applied for the assessment of southwest Pacific Ocean swordfish assessment in 2021 (SC17-SA-WP-04, Ducharme-Barth et al. 2021) to capture both "structural" and "estimation" uncertainty.
2. Conduct a similar review of the approaches used in the analysis pertaining to the stock assessment of southwest Pacific Ocean blue shark (SC18-SA-WP-03, Neubauer et al. 2022a).
3. Based on these reviews, provide recommendations for model ensemble construction, model retention, and weighting of models included within ensembles in the context of the WCPFC tuna, billfish, and shark assessments.

The expected outcomes of the project were to provide:

1. a basis for stock assessment teams to consider and apply alternative approaches for characterising stock assessment uncertainty (including model selection and weighting) across the WCPFC tuna, billfish, and shark assessments;

2. guidance for the Scientific Committee (SC) about the approaches for capturing assessment uncertainty in the provision of management advice; and
3. improved understanding for managers and stakeholders of the implications of alternative approaches to characterising uncertainty for their perception of risk.

Based on these term of reference, the two most recent assessments for all stocks considered by the WCPFC SC were reviewed using a structured approach. The review focused on the stock status and management advice as provided by the SC. It considered how uncertainty was addressed through the use of ensembles and sensitivities, and whether this uncertainty was used in management advice. In addition to the review, the present study included discussions with working group members of the Pacific Community (SPC) and International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean (ISC) to gain a thorough understanding of the rationale for the approaches used for addressing uncertainty in stock assessments.

The findings from this study revealed clear and long-standing differences in approaches between the ISC and SPC working groups for addressing uncertainty in stock assessments; however, there was a recent rapprochement of approaches. Although the ISC working groups traditionally preferred to present a single model with different levels of uncertainty for management advice, their recent advice included a more explicit consideration of alternative models and estimation uncertainty. In contrast, some SPC assessments previously used a considerable number of models in "structural uncertainty grids"; these assessments were without explicit consideration for a single "best" model, often considering all models in the grid equally plausible. Nevertheless, more recent assessments by SPC have attempted to constrain these model grids to sets of plausible models, and have explored metrics to weight models.

There is currently no best practice identified to weight models, or to address uncertainty in stock assessments more broadly, but there may be a valuable "middle ground" for both aspects of stock assessments. This approach would explicitly acknowledge and explore uncertainty, with standardised reporting to allow consistent management advice of the uncertainties considered for each assessment.

To illustrate some of the observations and recommendations from the current analysis, a set of simulations was set up. The simulations were designed to highlight differences in approaches to the characterisation of uncertainty, and were not intended to be a a realistic representation of typical stock assessments.

Based on the current review and simulations, we developed a set of recommendations for consideration by the SC19 relating to the use of model ensembles for management advice, and for the communication of assessment uncertainty.

Model ensembles and weighting

1. Develop joint priors and explicit rationales for grid axes and their values.
2. Either draw from, or weight axes over parameters according to the joint prior
3. Consider observation error, structural, parameter, and estimation uncertainty in management advice.

4. Where possible, express priors for model outcome space to avoid post-hoc selection/weighting.
5. Where post-hoc weighting is necessary (unexpected outcomes), this weighting should be proposed by analysts.
6. Clarity about uncertainties addressed by the grids address, including clear and consistent terminology around uncertainty.

Communicating uncertainty and risk

1. Develop a template for reporting management advice and uncertainties; ideally this template is a standardised table format to help managers and stakeholders locate key quantities easily.
2. Agree on terminology and the set of required measures (ideally probabilities relative to reference points).
3. Clear communication about quality of information determining stock status and management advice:
    - Qualification and quantification of uncertainties.
        (a) Data quality.
        (b) Model/population: structural uncertainty. (Note the use of "structural" here refers to models with different likelihoods, rather than different parameter values.)
        (c) Key parameters (parameter and estimation uncertainty).
    - Key uncertainties and potential impacts.
4. With respect to item 3, develop a set of research recommendations to address key uncertainties.
5. A review of timelines and capacity for tuna stock assessments may be necessary to allow sufficient time and capacity to adequately address uncertainty. Sufficient time is also needed to enable the provision of management advice that is consistent with the application of the precautionary approach as outlined in the WCPFC convention text.

Further development and future research

In addition, we suggest that the SC19 consider recommending:

- the provision of a project to develop a standardised reporting template for the reporting of uncertainty and risk that incorporates recommendations made in the present review; and
- the further development of methodology and idealised simulations to develop principled model ensemble approaches, in particular to consider the ability of alternative model diagnostics to identify model plausibility and weights.

1. INTRODUCTION

Ecological models, including fisheries stock assessments, are necessarily incomplete representations of the natural world, as they are unable to fully capture the complexity of open natural systems with all external and internal influences. For this reason, these models include a non-negligible degree of uncertainty, which leads to associated risk that management according to management advice based on these models leads to non-desirable outcomes (i.e., a loss relative to otherwise achievable outcomes for a quantity of interest, such as long-term yield (Rosenberg & Restrepo 1994, Francis & Shotton 1997).

The Western and Central Pacific Fisheries Commission (WCPFC) convention (Article 6) prescribes the application of a precautionary approach to fisheries management by the commission. In particular, "[...the Commission shall:] take into account, inter alia, uncertainties relating to the size and productivity of the stocks, reference points, stock condition in relation to such reference points, levels and distributions of fishing mortality and the impact of fishing activities on non-target and associated or dependent species, as well as existing and predicted oceanic, environmental and socio-economic conditions"[1] and "Members of the Commission shall be more cautious when information is uncertain, unreliable or inadequate..."[2]. These articles follow the precautionary principle as outlined by the Food and Agricultural Organization of the United Nations, which requires that "...where the likely impact of resource use is uncertain, priority should be given to conserving the productive capacity of the resource" and "...harvesting and processing capacity should be commensurate with estimated sustainable levels of resource, and that increases in capacity should be further contained when resource productivity is highly uncertain" (Food and Agriculture Organization of the United Nations 1996, de Bruyn et al. 2013). Uncertainty in stock assessment models is, therefore, a fundamental aspect of fisheries management advice that allows a precautionary approach to resource management (Rosenberg & Restrepo 1994, Francis & Shotton 1997, Cadrin et al. 2015, Privitera-Johnson & Punt 2020, Merino et al. 2020).

Despite efforts over the past few decades to develop frameworks for communicating uncertainty in the scientific advice provided for fisheries management, there is currently no standard framework available. The lack of a standard framework may be due to the diversity of approaches that are applied in fisheries science to characterise uncertainty (Privitera-Johnson & Punt 2020). Although there is some consistency in reporting of stock status and trends with respect to targets and limits in tuna Regional Fisheries Management Organisations (RFMOs) through the application of the Kobe framework (i.e., Kobe and Majuro plots), the quantification of uncertainties is currently inconsistent among tuna RFMOs (Merino et al. 2020).

While uncertainty may be categorised in a number of ways, Rosenberg & Restrepo 1994, amongst others, distinguished four types of uncertainty relating to scientific advice:

1. measurement error (e.g., uncertain total catch);
2. natural variability relates to temporally or spatially variable production parameters;

[1] Article 6, paragraph 1b of the Convention on the Conservation and Management of Highly Migratory Fish Stocks in the Western and Central Pacific Ocean.

[2] Article 6, paragraph 2.

3. model error and parameter uncertainty relates to imperfect specification or knowledge of relevant model parameters; and,
4. estimation error stems from imprecise estimates of model parameters, often due to the above types of uncertainty.

While some of these uncertainties, such as the precision of total catch, may be improved over time due to more accurate and informative data (epistemic uncertainties), other uncertainties, such as natural variability in population processes, are not reducible (termed "aleatory uncertainties" or simply "process error"). The latter may nevertheless be quantified in stock assessments (Merino et al. 2020).

On the basis of the above classification of uncertainties, it has been suggested that a broad approach of:

1. Analysis of estimation error,
2. sensitivity to model error,
3. stochastic projections, and
4. quantification of risk,

provides an overarching framework for risk analysis in fisheries science (Rosenberg & Restrepo 1994).

Each of the steps in the risk management approach has its unqiue challenges. For estimation error analysis, data, complexity, and computational constraints mean that no method for quantifying estimation error is universally applicable in practice, and the performance of different methods may vary with the assessment setup (Magnusson et al. 2013). Nevertheless, it is generally acknowledged that estimation error is an important part of characterising overall uncertainty (e.g., Privitera-Johnson & Punt 2020, Ducharme-Barth & Vincent 2022).

For model error and subsequent steps in a fisheries risk analysis, there are two long-standing schools of thought: either, management advice is given on the basis of a single "base-case" model — representing the model deemed most plausible and appropriate by the analysts (or assessment team; see Rosenberg & Restrepo 1994, for arguments for this approach) — or, a range of models are explored (Hilborn et al. 1993, e.g.,) as an "ensemble" to integrate over model error in steps 2–4 of the risk analysis. While in the base-case approach sensitivities to model assumptions are often explored, the key difference between approaches is that the sensitivities are not usually used to formulate the scientific advice itself (Merino et al. 2020).

Within tuna RFMOs, the ensemble model approach (called a "grid" when factorial designs over all uncertainties are considered) has been the most prevalent approach to quantifying uncertainties in stock assessments (Merino et al. 2020). Although the ensemble approach has been called the "state of the art" for characterising uncertainty (Punt et al. 2023), significant challenges remain with respect to the application of model ensembles in fisheries stock assessments.

This study for WCPFC project 113 aimed to provide both general and specific review components to develop recommendations about the presentation of stock assessment and management advice uncertainty by the WCPFC scientific committee. The terms of reference for the general review were to:

1. Review and summarise the different approaches used for characterising uncertainty in WCPFC stock assessments for tuna, billfish, and sharks over the last five years;
2. describe how uncertainty was communicated in the context of management risks and its influence on decision-making processes used by the WCPFC; and,
3. comment on the suitability of the recent approaches to characterising uncertainty for the management systems, including the harvest strategy approach.

The specific aims of the review were to:

1. Critically review the ensemble approach that was applied for the assessment of southwest Pacific Ocean swordfish assessment in 2021 (SC17-SA-WP-04, Ducharme-Barth et al. 2021) to capture both "structural" and "estimation" uncertainty.
2. Conduct a similar review of the approaches used in the analysis pertaining to the stock assessment of southwest Pacific Ocean blue shark (SC18-SA-WP-03, Neubauer et al. 2022a).
3. Based on these reviews, provide recommendations for model ensemble construction, model retention, and weighting of models included within ensembles in the context of the WCPFC tuna, billfish, and shark assessments.

The expected outcomes of the project were to provide:

1. a basis for stock assessment teams to consider and apply alternative approaches for characterising stock assessment uncertainty (including model selection and weighting) across the WCPFC tuna, billfish, and shark assessments;
2. guidance for the Scientific Committee (SC) about the approaches for capturing assessment uncertainty in the provision of management advice; and
3. improved understanding for managers and stakeholders of the implications of alternative approaches to characterising uncertainty for their perception of risk.

2. METHODS

The current project included engagement with members of the stock assessment working groups of of the Pacific Community (SPC) and International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean (ISC), and a structured review of existing stock assessments.

2.1 Discussions with stock assessment groups

Discussions with the SPC and ISC stock assessment teams were conducted during the early phase of the present project to gain an understanding of the rationale for different approaches to representing stock assessment uncertainty. A particular focus of this engagement was the use of model ensembles in the provision of management advice. All meetings were minuted, and the minutes were shared with individual groups to ensure that there was misunderstanding or misrepresentation of the contributions from the stock assessment teams.

2.2 Review of uncertainty characterisation in stock assessments and management advice

Due to the considerable number of stock assessments to be reviewed for the present project, reviews were undertaken in a structured way to allow consistent appraisal of stock assessments across the project team. For this approach, the overall review structure was agreed on at the beginning of review process, including the development of a review template. The template was updated throughout the review process to address shortcomings in its initial structure. Reviews were moderated across the review team by discussing review outcomes and comparing notes on the interpretations to ensure that a consistent methodology was applied across stock assessments. The review spreadsheet was subsequently shared with assessment teams to ensure accuracy of individual fields, and any errors were corrected.

The review template consisted of three main sections. In the first section, metadata was collected for each assessment, describing the assessment authors (or teams), assessment year, and software and modelling approach. The second section described the base case and whether it was derived from a previous model or represented a new approach, whether sensitivities where used, and whether an ensemble approach was applied. The latter was defined in the context of the stock structure and management advice; i.e., if more than one model was used by the SC to provide advice, then we considered that an ensemble was used. For all assessments, we attempted to record the diagnostics that were used, and whether they were applied to a base case (or diagnostic) model only, or if they were applied more broadly. The third section pertained to the use of the assessments in the stock status and management advice as recorded by the SC and provided to the WCPFC commission. The focus here was exclusively on the use of assessment outputs in the provision of advice, not on the assessment itself. The latter may be more detailed and evaluate alternative options that are not presented in advice papers. Nevertheless, we considered that the advice papers contained the agreed level of management advice and, therefore, provide the most appropriate and consistent reference.

Using this template, the two most recent assessments for all stocks listed on WCPFC website under "Current Stock Status and Advice"[3] were included in the review (Table 1).

[3] see https://www.wcpfc.int/current-stock-status-and-advice, accessed April 2023

Table 1: Stock assessments included in the present review for WCPFC project 113. SC, Scientific Committee; ISC, International Scientific Committee for Tuna and Tuna - like Species in the North Pacific Ocean; MIST, maximum impact sustainable threshold; BDM, biomass dynamics model.

Species group	Species	Assessment	Team	Approach	Reference
WCPO Tuna	Bigeye tuna	2020 (SC16)	SPC	Multifan-CL	Ducharme-Barth et al. 2020
WCPO Tuna	Bigeye tuna	2017 (SC13/SC14)	SPC	Multifan-CL	McKechnie et al. 2017
WCPO Tuna	Yellowfin tuna	2020 (SC16)	SPC	Multifan-CL	Vincent et al. 2020
WCPO Tuna	Yellowfin tuna	2017 (SC13)	SPC	Multifan-CL	Tremblay-Boyer et al. 2017
WCPO Tuna	Skipjack tuna	2022 (SC18)	SPC	Multifan-CL	Jordan et al. 2022
WCPO Tuna	Skipjack tuna	2019 (SC15)	SPC	Multifan-CL	Vincent et al. 2019
WCPO Tuna	SP albacore tuna	2021 (SC17)	SPC	Multifan-CL	Jordan et al. 2021
WCPO Tuna	SP albacore tuna	2018 (SC14)	SPC	Multifan-CL	Tremblay-Boyer et al. 2018
WCPO Tuna	NP albacore	2020 (SC16)	ISC	Stock Synthesis	ISC 2020a
WCPO Tuna	NP albacore	2017 (SC13)	ISC	Stock Synthesis	ISC 2017b
WCPO Tuna	Pacific bluefin tuna	2022 (SC18)	ISC	Stock Synthesis	ISC 2022b
WCPO Tuna	Pacific bluefin tuna	2020 (SC16)	ISC	Stock Synthesis	ISC 2020b
WCPO Billfish	NP Swordfish	2018 (SC14)	ISC	Stock Synthesis	ISC Billfish Working Group 2018
WCPO Billfish	SWPO swordfish	2021 (SC17)	SPC	Multifan-CL	Ducharme-Barth et al. 2021
WCPO Billfish	SWPO swordfish	2017 (SC13)	SPC	Multifan-CL	Takeuchi et al. 2017
WCPO Billfish	SWPO striped marlin	2019 (SC15)	SPC	Multifan-CL	Ducharme-Barth et al. 2019
WCPO Billfish	SWPO striped marlin	2012 (SC08)	SPC	Multifan-CL	Davies et al. 2012
WCPO Billfish	NP striped marlin	2019 (SC15)	ISC	Stock Synthesis	ISC 2019
WCPO Billfish	NP striped marlin	2015 (SC11)	ISC	Stock Synthesis	ISC 2015b
WCPO Billfish	Pacific blue marlin	2021 (SC17)	ISC	Stock Synthesis	ISC 2021
WCPO Billfish	Pacific blue marlin	2016 (SC12)	ISC	Stock Synthesis	ISC 2016
WCPO Sharks	Oceanic Whitetip Shark	2019 (SC15)	SPC	Stock Synthesis	Tremblay-Boyer et al. 2019
WCPO Sharks	Oceanic Whitetip Shark	2012 (SC08)	SPC	Stock Synthesis	Rice & Harley 2012a
WCPO Sharks	Silky shark	2018 PO (SC14)	SPC	Stock Synthesis	Clarke et al. 2018
WCPO Sharks	Silky shark	2018 WPO (SC14)	SPC	Stock Synthesis	Clarke et al. 2018
WCPO Sharks	Silky shark	2013 WPO (SC09)	SPC	Stock Synthesis	Rice & Harley 2013
WCPO Sharks	Silky shark	2012 WPO (SC08)	SPC	Stock Synthesis	Rice & Harley 2012b
WCPO Sharks	SP blue shark	2022 (SC18)	SPC	Stock Synthesis	Neubauer et al. 2022b
WCPO Sharks	SP blue shark	2021 (SC17)	SPC	Stock Synthesis	Neubauer et al. 2021
WCPO Sharks	SP blue shark	2016 (SC12)	SPC	Multifan-CL	Takeuchi et al. 2016
WCPO Sharks	NP blue shark	2022 (SC18)	ISC	Stock Synthesis	ISC 2022a
WCPO Sharks	NP blue shark	2017 (SC13)	ISC	Stock Synthesis	ISC 2017a
WCPO Sharks	NP shortfin mako	2018 (SC14)	ISC	Stock Synthesis	ISC 2018
WCPO Sharks	NP shortfin mako	2015 (SC11)	ISC	Indicator analysis	ISC 2015a
WCPO Sharks	SWPO shortfin mako	2022 (SC18)	SPC	Stock Synthesis	Large et al. 2022
WCPO Sharks	Pacific bigeye thresher shark	2017 (SC13)	NIWA	MIST	Fu et al. 2017
WCPO Sharks	Porbeagle shark	2017 (SC13)	NIWA	MIST + BDM	Hoyle et al. 2017
WCPO Sharks	Whale Shark	2018 (SC14)	Dragonfly	Spatial Risk Assessment	Neubauer et al. 2018

2.3 Illustrative simulations

To illustrate some of the observations and recommendations from the review, a set of simulations was set up. These simulations were more technical than the review aspect of this project, and are provided as an optional detail.

The simulation setup was designed to illustrate and highlight differences in approaches to uncertainty characterisation, and were not aimed to be a realistic representation of typical stock assessments. In view of their purpose, the simulations were set up using idealised scenarios, where most of the model parameters were known exactly, and a small set of key parameters were treated as unknown (see detailed description of the model and the simulation assumptions in Appendix A).

Specifically, simulations were set up using albacore-like parameters from the stock assessment and risk analysis by Punt et al. (1995) (see Appendix A, Table A-3). Using a known catch time series for albacore tuna from the International Commission for the Conservation of Atlantic Tunas (ICCAT) for the period from 1967 to 2019, we simulated the dynamics and associated data from a simple age-structured model (Figure 1). The resulting time-series were based on drawing random recruitment deviations each year according to an assumed σ_R of 0.4, and drawing observations from the resulting population time series.

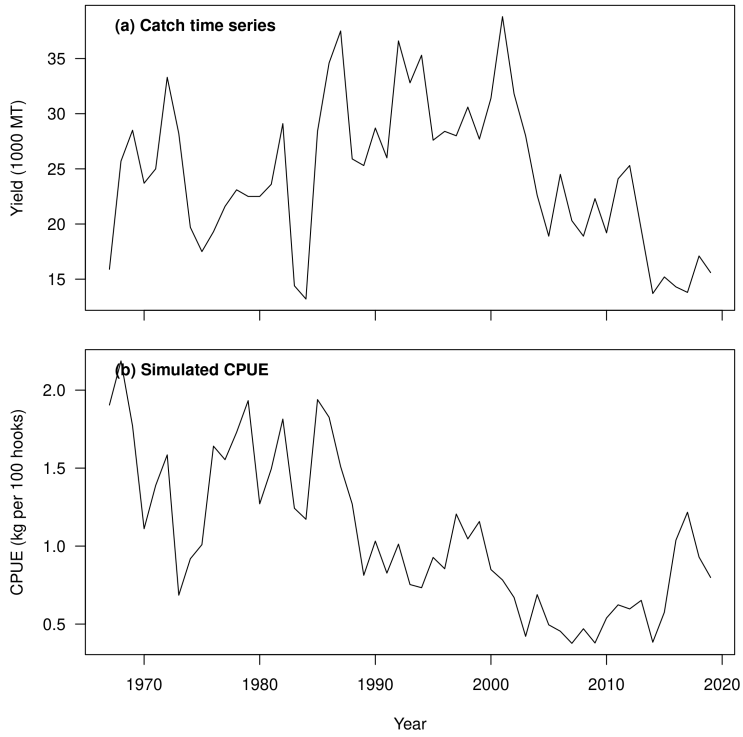


Figure 1: An illustrative catch time series (top) based on North Atlantic Ocean albacore tuna, and the corresponding simulated catch-per-unit-effort (CPUE) from the age-structured population model (bottom).

We assumed that the key unknowns for this simulated population were productivity parameters relating to natural mortality (M) and steepness (h). For this simulation,

all other parameters were held constant at their true value, and only the unfished equilibrium recruitment (R_0) and the recruitment deviations were free parameters in all models.

We compared four approaches to characterising uncertainty in assessments:

1. An idealised scenario was used as a reference. In this scenario, all productivity parameters can be estimated and uncertainty can be fully estimated in a Bayesian assessment model. The simulated time series was chosen (via selection of a random seed) to be weakly informative about stock productivity. Both natural mortality (M) and steepness (h) were estimated using an informative joint prior (e.g., derived from life history parameters), which also encoded information about the covariance of M and h. Population time series and management quantities were estimated from the full posterior distribution of the model parameters given the data and model structure. A variation of this model was also run with productivity parameters fixed at the posterior median, and estimation uncertainty applied to (R_0) and the recruitment deviations only.

2. In this approach, a grid was constructed which integrated over the joint prior from the idealised scenario (the previous approach) instead of fitting natural mortality (M) and steepness (h) within the model. This approach was applied by evaluating the assessment model at all combinations of values defined by a symmetric set of quantiles for the joint prior (e.g., 0.2, 0.5, 0.8). All models were given the same weight in the grid, and time series and management quantities were estimated across this fixed-weight grid.

3. The third approach was similar to the second approach, but the time series and management quantities were estimated across the grid with grid runs weighted by the joint prior.

4. The fourth approach applied a bootstrap Monte Carlo approach (Ducharme-Barth & Vincent 2022) by drawing an equivalent number of fixed parameter combinations from the joint prior used under the idealised scenario (the first approach). These combinations were then integrated over these model runs to estimate the time series and management quantities.

By making comparisons with a single estimated model, with either fixed or estimated productivity parameters, it was possible to highlight how different approaches to the inclusion of uncertainty in the model can affect estimates of stock status and management quantities. These simulated comparisons were used to develop an understanding of the ability of model ensembles to approximate an "ideal" model where all uncertainties are included in the estimation process.

3. RESULTS

3.1 Review of current stock assessment practice

Across the WCPFC, stock assessments have been mainly conducted by SPC and ISC assessment teams (see Table 1). Although some shark stock assessments have been conducted as risk assessments by other assessment teams under contract to WCPFC with

direct FAO funding (e.g., silky shark, porbeagle shark), recent shark stock assessments have followed ISC and earlier SPC approaches of using integrated stock assessment models (Table 3). All recent models, including shark stock assessments since 2018, have been fully integrated stock assessments.

Recent assessments were in two broad categories, consistent with the "base case" and "model ensemble" paradigms. The SPC assessments followed practice in other tuna RMFOs, with (often factorial) "structural uncertainty" grids applied for all assessments (see also Merino et al. 2020). For accepted recent assessments, these grids integrated over between 18 and 648 models to derive stock status and assessment advice. Most often (5 of 8 assessments), these ensembles were not weighted, while 3 of 8 assessments used *a priori* plausibility to remove models from the grid that provided implausible results. There was no consistent definition of implausibility applied across assessments.

Although a range of sensitivities and diagnostics were usually applied to a diagnostic model, they were not applied to the uncertainty grids (Table 2). An exception was the 2022 assessment of blue shark in the Southwest Pacific Ocean (Neubauer et al. 2022b), which employed a range of diagnostics, including jittering and retrospective analysis, to a full factorial uncertainty grid. Similarly, only the assessment of South Pacific Ocean swordfish (Ducharme-Barth et al. 2021) included full estimation uncertainty in the calculation of stock status and management advice.

Table 2: Stock, use of model ensemble, and diagnostics applied to various levels of the analysis (diagnostic / base-case only vs sensitivities or ensembles / uncertainty grids). MASE: mean absolute square error; ASPM: Age-structured production model

Stock	Assessment	Ensemble	Residuals	Likelihood prof.	Retrospectives	Hindcast/MASE	ASPM	Jitter	Hessian	Ensemble weighting
Bigeye tuna	2020 (SC16)	Yes	Base	Base	Base	No	No	Base	Base	None
Bigeye tuna	2017 (SC13/SC14)	Yes	No	Base	Base	No	No	No	Base	None
Yellowfin tuna	2020 (SC16)	Yes	Base	Base	Base	No	No	Base	No	None
Yellowfin tuna	2017 (SC13)	Yes	Base	Base	Base	No	No	No	No	None
Skipjack tuna	2022 (SC18)	Yes	Base	Base	Base	No	No	Base	Ensemble, attempted	None
Skipjack tuna	2019 (SC15)	Yes	No	Base + sensitivities	Base	No	No	No	Base	None
SP albacore tuna	2021 (SC17)	Yes	Base	Base	Base	No	No	No	Base	None
SP albacore tuna	2018 (SC14)	Yes	No	Base	Base	No	No	No	No	None
NP albacore	2020 (SC16)	No	Base	Base	Base	No	Base	Base	Base + sensitivities	None
NP albacore	2017 (SC13)	No	Base	Base	Base	No	Base	Base	No	None
Pacific bluefin tuna	2022 (SC18)	No	Base	Base	Base	Base	Base	No	Base	None
Pacific bluefin tuna	2020 (SC16)	No	Base	Base	Base	No	Base	Base	No	None
NP Swordfish	2018 (SC14)	No	Base	Base	Base	No	Base	Base	Base	None
SWPO swordfish	2021 (SC17)	Yes	Base	Base	Base	Base	Base	No	Ensemble	a-priori plausibility
SWPO swordfish	2017 (SC13)	Yes	Base	Base	Base	No	No	No	Base	None
SWPO striped marlin	2019 (SC15)	Yes	Base	Base	Base	No	No	No	Base	None
SWPO striped marlin	2012 (SC08)	Yes	Base	Base + a few others	No	No	No	No	Base	None
NP striped marlin	2019 (SC15)	No	Base	Base	Base	No	Base	Base	Base	None
NP striped marlin	2015 (SC11)	No	Base	Base	Base	No	No	No	No	None
Pacific blue marlin	2021 (SC17)	Yes	Ensemble	Ensemble	Ensemble	Ensemble	Ensemble	Ensemble	Ensemble	None
Pacific blue marlin	2016 (SC12)	No	Base	Base	Base	No	No	Base	Base	None
Oceanic Whitetip Shark	2019 (SC15)	Yes	No	Base	Base	No	No	No		a-priori plausibility
Oceanic Whitetip Shark	2012 (SC08)	Yes	No	No	No	No	No	No	Base + a few others	a-priori plausibility
Silky shark	2018 PO (SC14)	No	No	No	No	No	No	No		None
Silky shark	2018 WPO (SC14)	No	No	No	No	No	No	No		None
Silky shark	2013 WPO (SC09)	Yes	No	No	No	No	No	No	Base + a few others	a-priori plausibility
Silky shark	2012 WPO (SC08)	Yes	No	No	No	No	No	No	Base + a few others	a-priori plausibility
SP blue shark	2022 (SC18)	Yes	Base	Base	Ensemble	Ensemble	No	Ensemble	Ensemble	a-priori plausibility
SP blue shark	2021 (SC17)	Yes	Base	Base	Base	Base	No	No	Ensemble	None
SP blue shark	2016 (SC12)	Yes	No	Base	No	No	No	No		None
NP blue shark	2022 (SC18)	Yes	Ensemble	Ensemble	Ensemble	Ensemble	Ensemble	Ensemble	Ensemble	hypothesis tree
NP blue shark	2017 (SC13)	No	Base	Base	Base	No	Base	No	Base	None
NP shortfin mako	2018 (SC14)	No	Base	Base	Base	No	Base	Base		None
NP shortfin mako	2015 (SC11)	No								
SWPO shortfin mako	2022 (SC18)	No	Base	Base	Base	No	No	No		None
Pacific bigeye thresher shark	2017 (SC13)	No								
Porbeagle shark	2017 (SC13)	No								
Whale Shark	2018 (SC14)	No								

Stock status and management advice from SPC-led assessments were usually delivered from the full ensembles (uncertainty grids; Table 3), and projections were conducted from the full grid to understand medium-term risk. The quantification of stock status in SC advice was consistently provided in the form of a table of mean values and quantiles of management quantities (reference points and status relative to reference points) from the grid results. No equivalent format was applied for projections or management advice from these assessments to date, but probability statements in management advice text generally encapsulated model uncertainty and provided a quantification of risk with respect to the considered uncertainties.

Table 3: Stock, use of model ensemble, source of estimates for stock status, structural uncertainty, estimation uncertainty and projections, and an indicator if the assessments were used for management advice (MA). Note, this classification is based on management advice provided, and individual assessment documents often provided additional detail about sensitivities and their relevant results, which were not necessarily reported for management advice.

Stock	Assessment	Ensemble	Est. Stock Status	Structural Uncertainty	Estimation uncertainty	Projections	Use for MA
Bigeye tuna	2020 (SC16)	Yes	Ensemble: entire	Ensemble: entire	None	Ensemble: entire	Yes
Bigeye tuna	2017 (SC13/SC14)	Yes	Ensemble: entire	Ensemble: entire	None	Ensemble: entire	Yes
Yellowfin tuna	2020 (SC16)	Yes	Ensemble: entire	Ensemble: entire	None	Ensemble: entire	Yes
Yellowfin tuna	2017 (SC13)	Yes	Ensemble: entire	Ensemble: entire	None		Yes
Skipjack tuna	2022 (SC18)	Yes	Ensemble: entire	Ensemble: entire	None	Ensemble: entire	No
Skipjack tuna	2019 (SC15)	Yes	Ensemble: entire	None	None	Ensemble: entire	Yes
SP albacore tuna	2021 (SC17)	Yes	Ensemble: entire	Ensemble: entire	None	Ensemble: entire	Yes
SP albacore tuna	2018 (SC14)	Yes	Ensemble: entire	Ensemble: entire	None		Yes
NP albacore	2020 (SC16)	No	Single model	Sensitivities	Base case	Base case	Yes
NP albacore	2017 (SC13)	No	Single model	Sensitivities	None	Base case	Yes
Pacific bluefin tuna	2022 (SC18)	No	Single model	None	None	Sensitivities	Yes
Pacific bluefin tuna	2020 (SC16)	No	Single model	None	None	Base case	Yes
NP Swordfish	2018 (SC14)	No	Single model	None	None	Base case	Yes
SWPO swordfish	2021 (SC17)	Yes	Ensemble: partial	Ensemble: partial	Ensemble: partial	Ensemble: partial	Yes
SWPO swordfish	2017 (SC13)	Yes	Ensemble: entire	Ensemble: entire	None		Yes
SWPO striped marlin	2019 (SC15)	Yes	Ensemble: entire	Ensemble: entire	None	None	Yes
SWPO striped marlin	2012 (SC08)	Yes	Ensemble: entire	Ensemble: entire	None	None	Yes
NP striped marlin	2019 (SC15)	No	Single model	None	None	Base case	Yes
NP striped marlin	2015 (SC11)	No	Single model	None	None	Base case	Yes
Pacific blue marlin	2021 (SC17)	Yes	Ensemble: entire	Ensemble: entire	Ensemble: entire	Ensemble: entire	Yes
Pacific blue marlin	2016 (SC12)	No	Single model	None	None	None	Yes
Oceanic Whitetip Shark	2019 (SC15)	Yes	Ensemble: entire	Ensemble: entire	None	Ensemble: entire	Yes
Oceanic Whitetip Shark	2012 (SC08)	Yes	Ensemble: entire	Ensemble: entire	Ensemble: entire	None	Yes
Silky shark	2018 PO (SC14)	No	None	None	None	None	No
Silky shark	2018 WPO (SC14)	No	Single model	Single model	Base case	None	Yes
Silky shark	2013 WPO (SC09)	Yes	Ensemble: partial	Ensemble: partial	Ensemble: partial	None	Yes
Silky shark	2012 WPO (SC08)	Yes	Ensemble: partial	Ensemble: partial	Ensemble: partial	None	No
SP blue shark	2022 (SC18)	Yes	Ensemble: partial	Ensemble: partial	None	None	Yes
SP blue shark	2021 (SC17)	Yes	Ensemble: entire	Ensemble: entire	None	None	No
SP blue shark	2016 (SC12)	Yes	None	None	None	None	No
NP blue shark	2022 (SC18)	Yes	Ensemble: entire	Ensemble: entire	Ensemble: entire	Ensemble: entire	Yes
NP blue shark	2017 (SC13)	No	Single model	Sensitivities	Sensitivities	Base case	Yes
NP shortfin mako	2018 (SC14)	No	Single model	Single model	None	Base case	Yes
NP shortfin mako	2015 (SC11)	No			None	None	No
SWPO shortfin mako	2022 (SC18)	No	Single model	Single model	None	None	No
Pacific bigeye thresher shark	2017 (SC13)	No	Single model	Single model	None	None	Yes
Porbeagle shark	2017 (SC13)	No	Single model	None	None	None	Yes
Whale Shark	2018 (SC14)	No	Single model	Single model	None	None	Yes

In contrast to SPC stock assessments, the ISC stock assessments used three models at most to provide management advice. The stock assessment teams expressed a clear preference for presenting only a single "best" and thoroughly vetted model to deliver stock assessment advice where possible. To date, alternative models and non-binary model weights were, therefore, only applied when a clear best model could not be identified due to conflicting data sources (e.g., growth curves, length composition data).

Although most ISC assessments used some form of sampling (e.g., delta-lognormal) to estimated stock status, and provide stochastic projections, the use of uncertainty in stock status statements and projections varied among assessments according to the assessment teams. For example, the SC reported stock status and advice from assessments of Pacific Ocean bluefin tuna by providing probabilities of achieving rebuilding targets from the base-case model, thereby accounting for the considered sources of uncertainty in the model developed in ISC (2022b) , but there was no uncertainty referenced in the provision of stock status estimates or status with respect to potential reference points (there have been no agreed reference points for Pacific Ocean bluefin tuna). In contrast, the SC reported uncertainty about status and projection biomass levels for North Pacific Ocean albacore tuna (ISC 2020a, based on) in the form of confidence intervals, but these uncertainties were not used to provide quantitative statements about risk in the form of probabilities; however, qualitative statements ("low probability of falling below limit reference points" and "risk increases [with alternative growth assumptions]") were provided.

When more than a single base-case model was used in ISC assessments, it sometimes led to qualitative statements about robustness of the base case and consequence for risk (as for North Pacific Ocean albacore, ISC (2020a)).In other assessments (e.g., Pacific Ocean blue marlin, North Pacific Ocean blue shark), estimates from ensembles were combined to provide management advice. The latter was not consistently reported; for example, stock status was reported in terms of probabilities for Pacific Ocean blue marlin (ISC 2019, based on), but medium-term risk, as quantified by projections, was only described qualitatively. The most recent advice for North Pacific Ocean blue shark contained probability statements for both the current status and projections based on model and estimation uncertainty (ISC 2022a, derived in), but the SC noted that these probabilities did not integrate over uncertainty in production parameters.

3.2 Case studies: recent ensembles and model weighting

Two recent SPC assessments — for southwest Pacific Ocean swordfish (SWPO-SWO; Ducharme-Barth et al. 2021) and southwest Pacific Ocean blue shark (SWPO-BSH; Neubauer et al. 2022a), and one ISC assessment (north Pacific Ocean blue shark, NP-BSH ISC 2022a) explicitly considered approaches for developing and weighting model ensembles. The former two projects were included in the terms of reference for the present project. In comparison, the latter assessment presented a somewhat different approach and was included here for completeness.

The two recent SPC assessments had a similar starting point to many of the SPC-led tuna assessments: an ensemble was constructed as an uncertainty grid over what were considered to be the main characteristics that could affect the estimation of uncertainty— albeit with a different approach for deriving the grid. Both approaches then developed methods for "filtering" these grids to retain a subset of models, and investigated possible

measures that could be used to weight models in the ensemble according to their "plausibility" as defined by the tested measures.

In contrast, the recent ISC NP-BSH ensemble started from the other end of the spectrum, using a single working model, and moving to three models when different data sources (CPUE indices) were found to be inconsistent, but none could be definitively considered "better" than another. This approach can be described as a hypothesis tree, with individual branches emanating from a set of central assumptions.

These recent examples of model ensembles encapsulate three key steps in any stock assessment:

1. Ensemble construction: Developing (a set of) models to use for the assessment.
2. Model weighting: determining the plausibility of alternative models (hypotheses).
3. Characterising uncertainty: Developing assessment advice and determining risk based on the chosen models.

These aspects are important components of implementing a risk analysis that supports a precautionary approach as mandated by the WCPFC convention. In the following, we elaborate on each of these aspects by contrasting the approaches in recent ensembles used for management advice in the WCPFC.

3.2.1 Ensemble construction: *a priori* definition of models to use for assessments

All stock assessments, throughout the assessment process, trial a large number of models, many of which will be deemed unsatisfactory by the analyst or assessment team. At minimum, stock assessment models are usually expected to fit to key biomass indices, while providing overall satisfactory fits to composition data (Francis 2011, 2017). From this perspective, all WCPFC assessments tend to start with a single model (diagnostic or base-case), with the aim of finding a model that can adequately fulfil these requirements. To this point, many models developed during the process are often discarded, and only key steps in the development are usually captured in "stepwise" updates from previous assessment models. Once the base or diagnostic case has been determined, a number of sensitivities are typically run to understand the robustness of the model to alternative assumptions and formulations. These steps are common to all assessments [4] in the WCPFC, and form the basis for the development of recent ensembles.

The key difference in recent ensembles, and between most ISC and SPC stock assessments, is whether (and how) assumptions were chosen to provide the basis for stock status statements and management advice. The recent ISC NP-BSH assessment included alternative catch-per-unit-effort (CPUE) indices and data weighting options because no model could be conclusively identified as performing better on the basis of diagnostics. This approach has been used across ISC assessments where more than one model was used for management advice. For SPC assessments, including SWPO-SWO and SWPO-BSH, many factors are considered uncertain. In addition, uncertainty for the factors that are important for the calculation of reference points or biomass levels

[4] Risk assessments typically used few sensitivities, but to date have also not had a prior model to build from.

are often included in an ensemble for management advice by developing a factorial grid around the diagnostic model, usually without reference to model performance or plausibility. The latter is a common criticism of the grid approach, and was a key focus of developments in both the SWPO-SWO and SWPO-BSH assessments.

Within this context, the main contribution from the SWPO-SWO assessment was to replace full-factorial grids, which may include implausible combinations of parameters (for example, high natural mortality and low steepness). These grids were replaced with an ensemble based on a joint prior that forces explicit consideration of the *a priori* plausibility of parameter combinations, and assigns probabilities to any given combination of parameters. Combinations of parameters were then drawn randomly using Monte Carlo sampling from the joint prior to derive an ensemble that reflects a clear set of hypotheses about parameters/data and their plausibility, to replace the factorial grid design.

Although the SWPO-BSH employed a more traditional approach by constructing a factorial model grid, it assigned prior probabilities to each axis according to the perceived quality of information and likelihood of scenarios. The grid was constructed over uncertainties in the data inputs and model components (e.g., CPUE representing low or high latitude fleets). Across the factorial gird, weights were multiplied leading to low probabilities for scenarios combining settings that were deemed less plausible *a priori*. As a result, some model assumptions were down-weighted, leading to markedly lower influence of extreme outcomes in the final stock advice.

Although the mechanism for constructing and weighting model axes was different between the SWPO-SWO and SWPO-BSH assessments, the key steps and outcomes were largely the same, consisting of the construction of an explicit prior distribution over uncertainty axes, and weighting according to the prior in the resulting ensemble.

3.2.2 Model weighting: determining the plausibility of alternative models (hypotheses) *a posteriori*

Although a joint prior distribution over parameters may serve to develop a more parsimonious ensemble in terms of prior model weights, not all models may be able to fit the data equally well or provide equally plausible outputs. Differential fits to data among models in an ensemble imply that there is information about a certain parameter or setting in a fixed set of data (it is not possible to formally compare fits across different data sets that imply different likelihoods). However, if data are informative about the plausibility of models in an ensemble, models may be formulated to estimate parameters rather than to include them as an axis in an ensemble. More often, however, ensembles explore fundamental uncertainties about data inputs or parameters about which the data carry limited information. In this instance, or when different datasets are used, model fit alone is unlikely to provide a means to distinguish among sets of models with different settings, and the prior weight alone will determine the weight in an ensemble.

Both the SWPO-SWO and SWPO-BSH assessments attempted a two-step procedure to weight models based on plausibility. As a first step, filtering was carried out that assigned a weight of zero to all models that were outside of expected outcomes. Expected outcomes were largely defined in terms of estimated and derived quantities for the SWPO-SWO assessment. For example, the total estimated biomass, was defined

in relation to the biomass of all tuna species in the area. Alternatively, the estimated biomass distribution in space was used to constrain the model ensemble to a set that was considered plausible. In the second step, models that were deemed technically inadequate were removed (e.g., no positive definite Hessian matrix, or parameter estimates on bounds). The resulting model ensemble was, thereby, reduced from 384 to 25 models, documenting that a joint prior alone can still lead to models that are considered implausible *a posteriori* for technical or biological reasons.

For the SWPO-BSH and ISC NP-BSH assessments, a comprehensive suite of diagnostics was applied across the model ensemble. This suite of model diagnostics was not able to distinguish among models for ISC NP-BSH, leading to the inclusion of multiple models in an ensemble used for management advice. For SWPO-BSH, models fell into sets that could be distinguished on the basis of the presence or absence of i) strong retrospective patterns and, ii) high M estimates of > 0.3. In both cases, models grouped naturally along the growth data axis, suggesting that one of the growth datasets was not compatible with other model assumptions and led to implausible assessment outcomes. These models were removed from the model grid, reducing the size of the grid by two-thirds.

The SWPO-SWO and SWPO-BSH assessments both attempted to further weight models according to model predictive capacity with respect to CPUE, using the mean absolute square error (Kell et al. 2021), or model stacking (Yao et al. 2018). In both cases, these criteria led to only minor differences in the distribution of outcomes when applied after the filtering step. In addition, they still required some subjective decisions about the diagnostics to use for weighting, and their settings (i.e., how to translate diagnostics into weights).

3.2.3 Characterising uncertainty: Developing management advice and determining risk based on chosen models

Although reporting of assessment outcomes and management advice for all recent ensembles largely followed patterns described above for SPC and ISC assessments, a key difference in both the SWPO-SWO and the NP-BSH assessments was that they integrated over structural and estimation uncertainty. Estimation uncertainty was not included in the SWPO-BSH assessment, and was generally not calculated for SPC model ensembles due to computational demand and the lack of consistent positive definite Hessians.

Despite reporting differences, and notwithstanding the inclusion or omission of estimation uncertainty in the provision of management advice, recent ensembles used by both SPC and ISC have integrated across models in the ensemble to provide management advice that includes specific consideration of uncertainty.

3.3 Illustration: pros and cons of ensemble approaches

The simple simulation study highlighted some of the advantages and potential shortcomings of different approaches to constructing and weighting a model ensemble.

With a precise and accurate prior, all approaches encompassed the "true" simulated value, but the grid approaches provided inflated uncertainty. For this reason, they implied higher risk than a single model that estimates productivity, or a model where

productivity is fixed at the prior median (i.e., the true value; Figures 2 and 3, Table B-1). The inflated tail risk was particularly evident when priors were chosen to be wide, there was no weighting, or no random Monte Carlo draws were performed to weight the models that make up the ensemble. In this simulation scenario, the models often estimated overfishing risk when overfishing was clearly not occurring in the simulation (Figures 4 and 5, Table B-1).

The pattern was reversed when a slightly-biased prior was applied, which implied that the prior expectation about productivity was inflated. In this scenario, fixing the productivity parameters at the prior median (i.e., the overestimated productivity) biased the analysis and underestimated risk with respect to fishing mortality reference points (Figures 6 and 7, Table B-1). Only the fully-estimated model provided sufficient coverage, while all ensembles also underestimated relative harvest levels. Only an approach that combined estimation and parameter uncertainty in the ensembles led to the inclusion of the true simulated value in the confidence bounds. A wider prior would have mitigated this limited coverage.

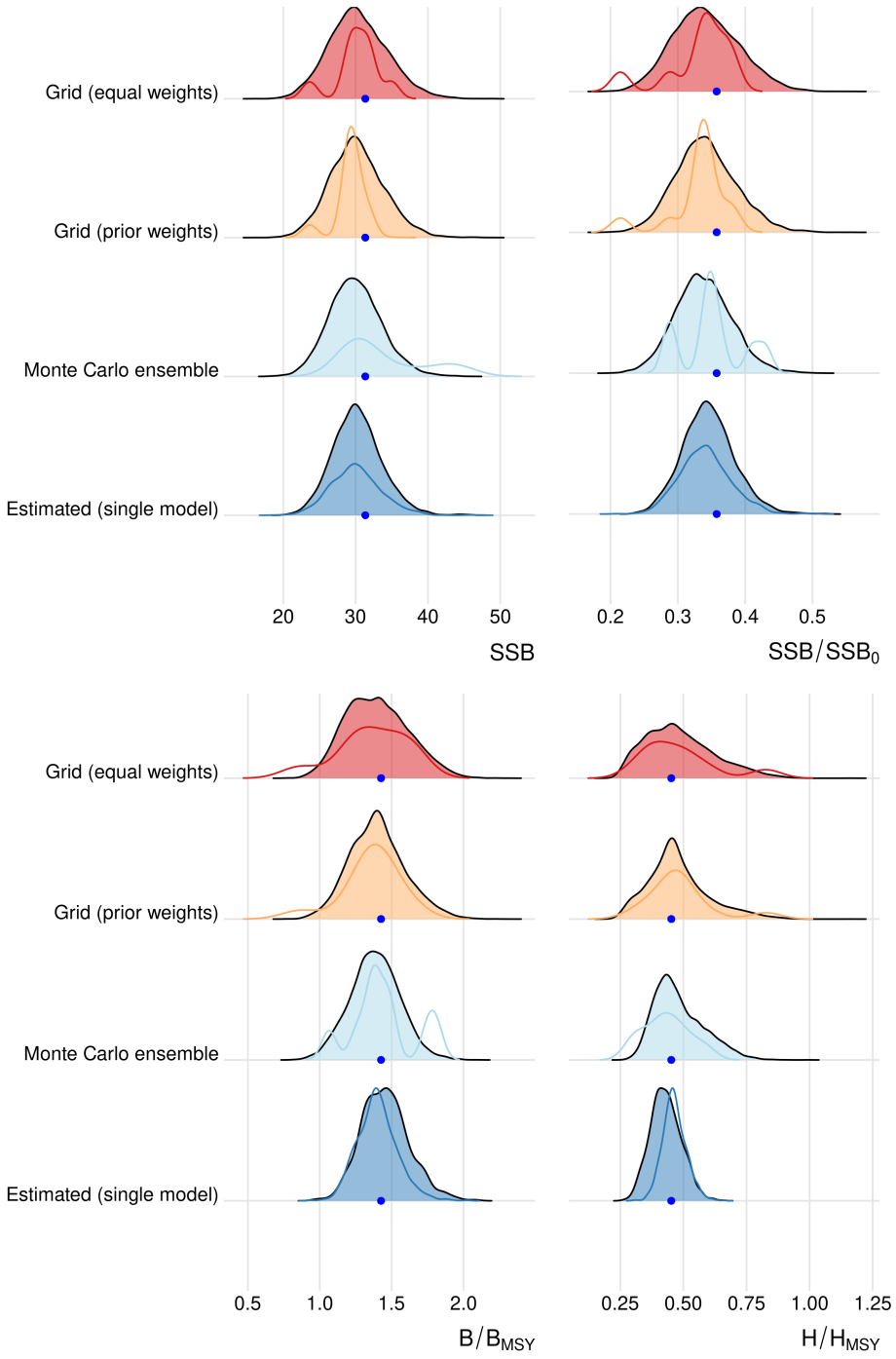


Figure 2: Distribution of key management quantities (SSB: spawning stock biomass, SSB_0: unfished spawning biomass, MSY: maximum sustainable yield, H: harvest rate), for the four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods use the same productivity prior, which was centered on the true value with a low coefficient of variation (10%). Solid lines show densities across maximum likelihood estimates for the grid approaches, whereas filled densities incorporate estimation uncertainty using Markov chain Monte Carlo (MCMC). The dashed dark blue line shows a model with estimation error only where productivity parameters were fixed at the prior median. The true simulated value is shown by the dark blue point.

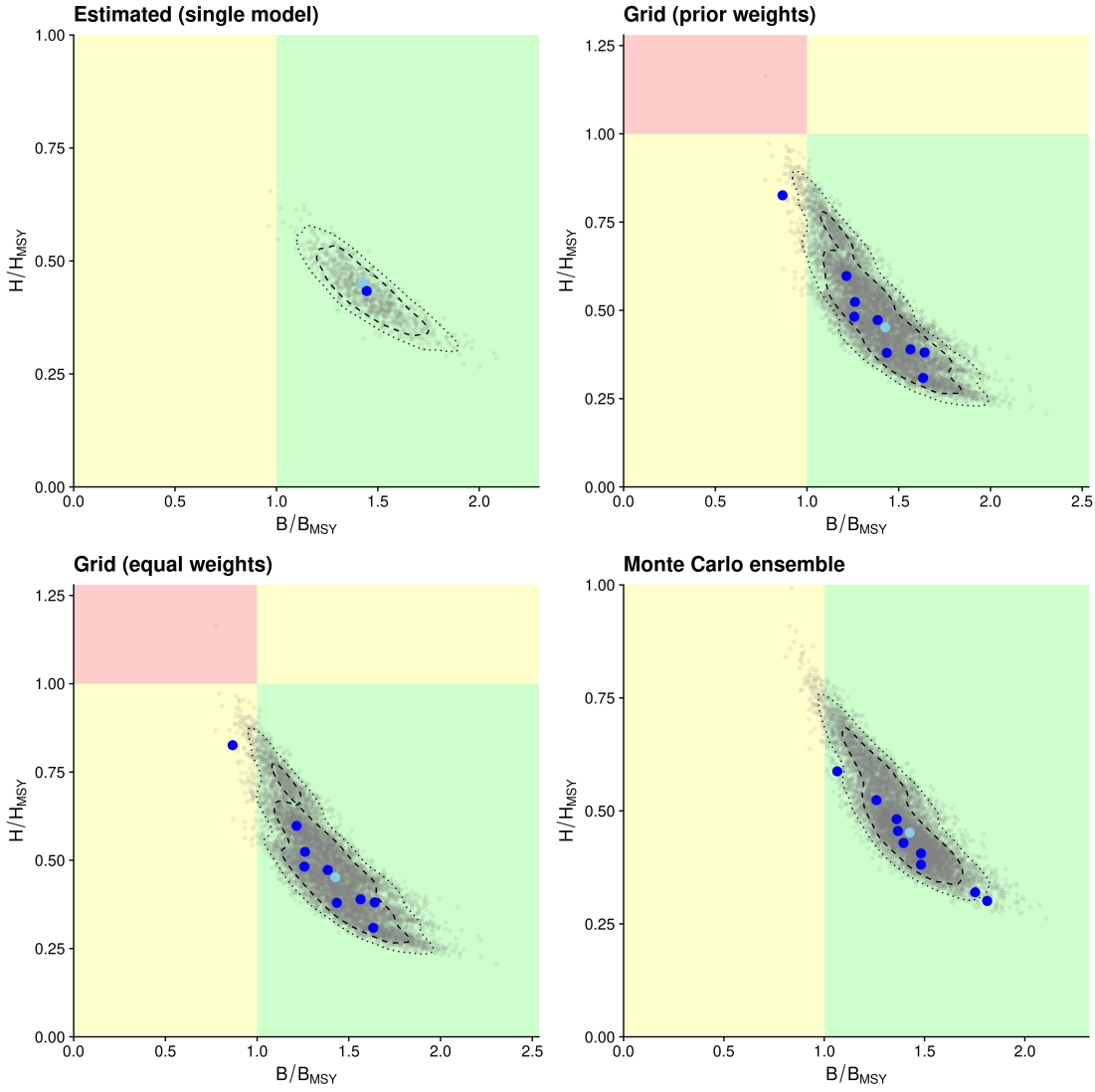


Figure 3: Kobe plot for the four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods use the same productivity prior, which was centered on the true value with a low coefficient of variation (10%). Dashed and dotted lines show 80% and 95% credible intervals from Markov chain MC draws, dark blue points show maximum likelihood estimates across model runs for grid approaches, and estimated status (prior mean) for a model with fixed productivity at the prior median. The true simulated value is shown by the light blue point.

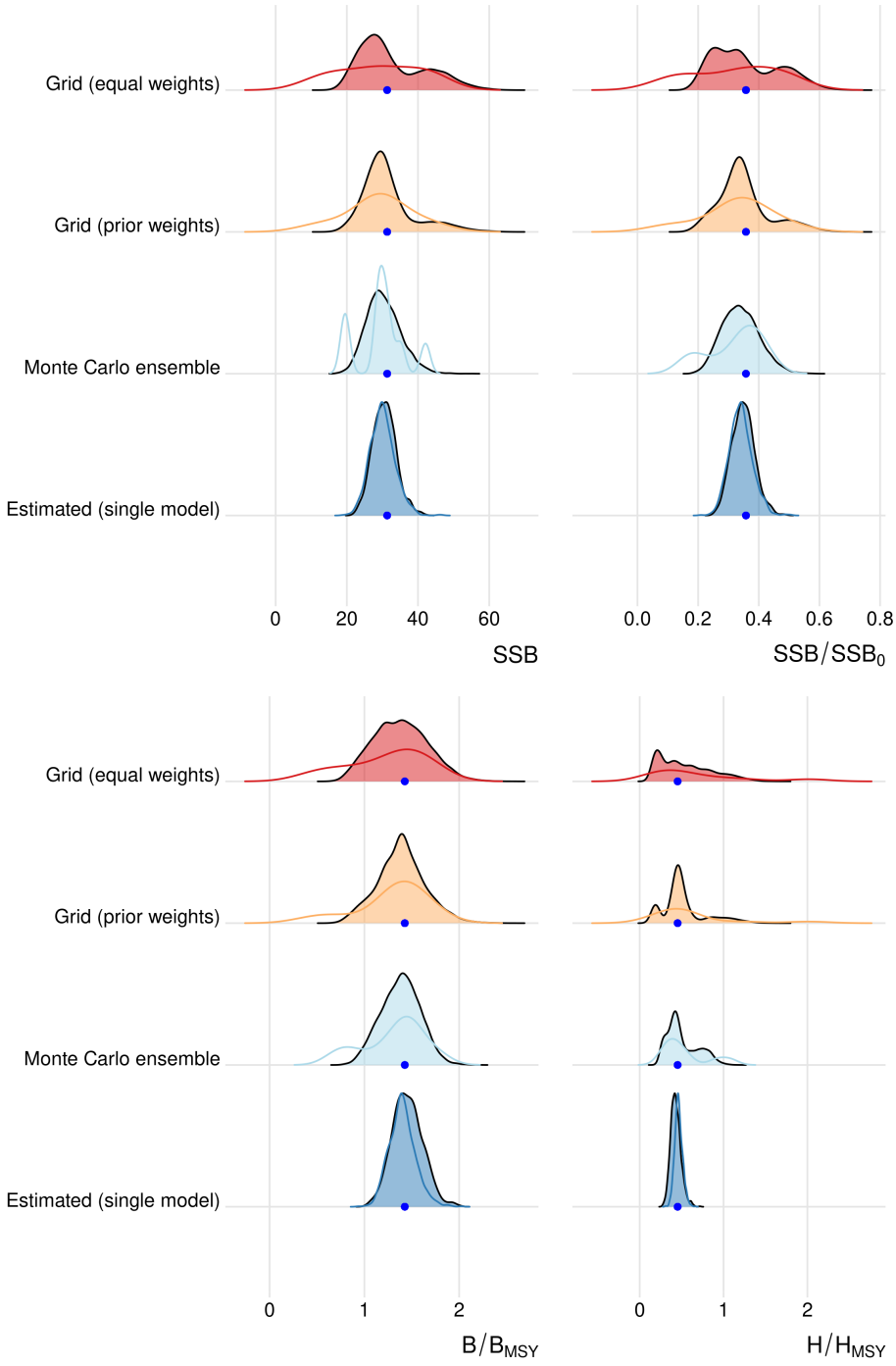


Figure 4: Distribution of key management quantities (SSB: spawning stock biomass, SSB_0: unfished spawning biomass, MSY: maximum sustainable yield, H: harvest rate), for the four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods use the same productivity prior, which is centered on the true value with a high coefficient of variation (30%). Solid lines show densities across maximum likelihood estimates for the grid approaches, whereas filled densities incorporate estimation uncertainty using Markov chain Monte Carlo. The dashed dark blue line shows a model with estimation error only where productivity parameters were fixed at the prior median. The true simulated value is shown by the dark blue point.

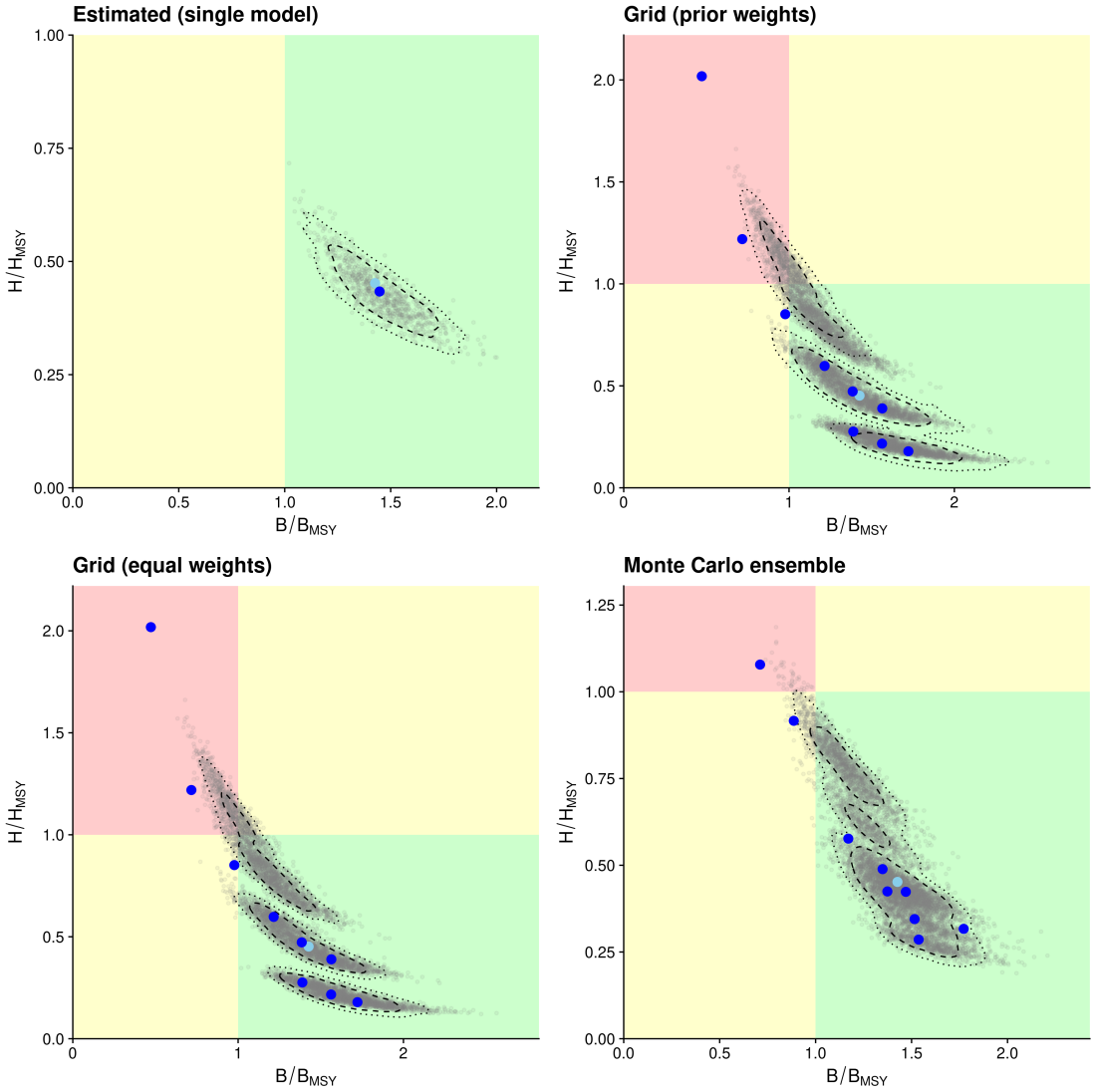


Figure 5: Kobe plot for the four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centered on the true value with a high coefficient of variation (30%). Dashed and dotted lines show 80% and 95% credible intervals from Markov chain MC draws, dark blue points show maximum likelihood estimates across model runs for grid approaches, and estimated status (prior mean) for a model with fixed productivity at the prior median. The true simulated value is shown by the light blue point.

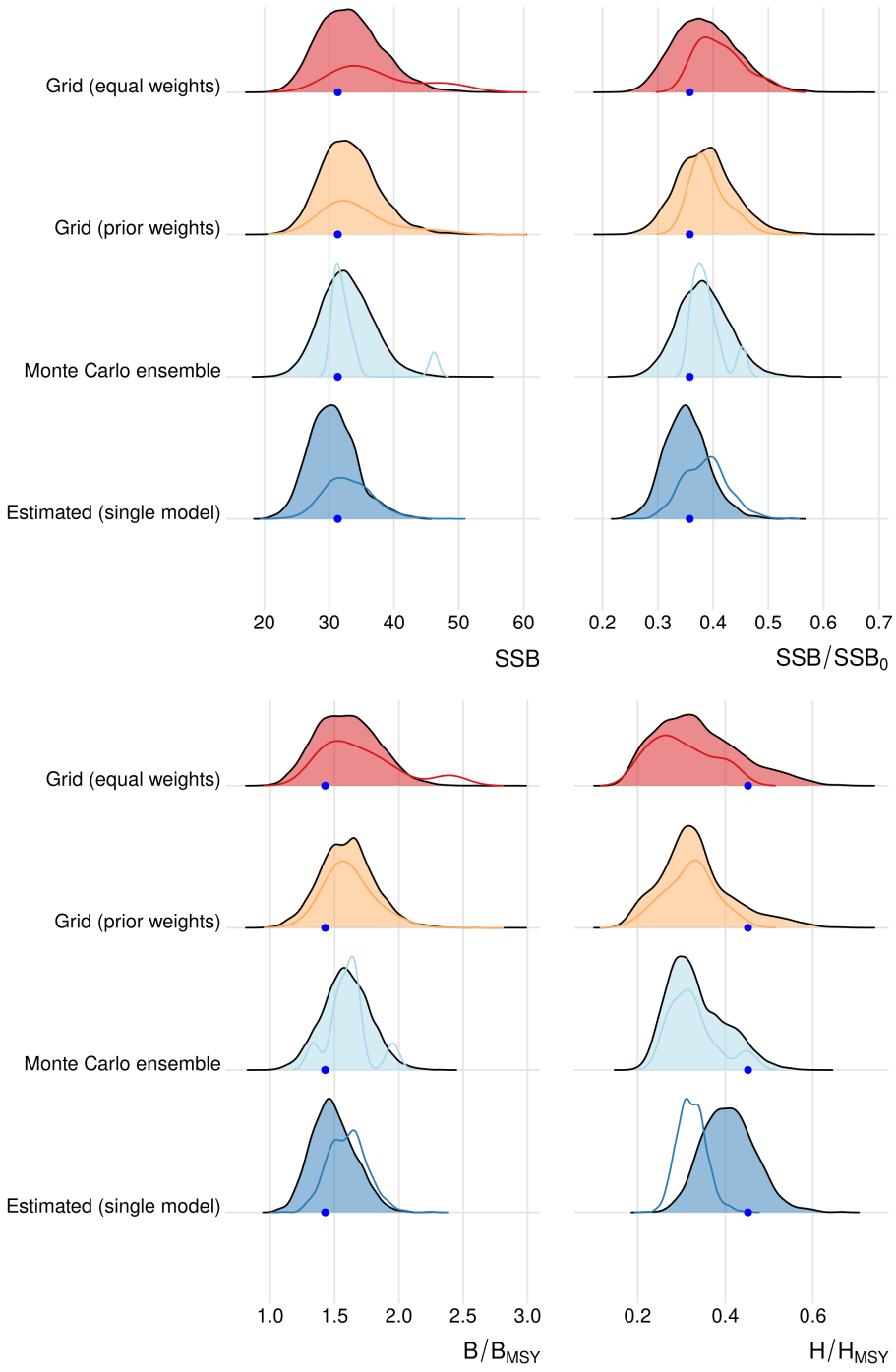


Figure 6: Distribution of key management quantities (SSB: spawning stock biomass, SSB_0: unfished spawning biomass, MSY: maximum sustainable yield, H: harvest rate), for the four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods use the same productivity prior, which is biased high by 0.05 for both parameters relative to the true value, with a tight prior (coefficient of variation 10%) representing seemingly good understanding of productivity. Solid lines show densities across maximum likelihood estimates for the grid approaches, whereas filled densities incorporate estimation uncertainty using Markov chain Monte Carlo. The dashed dark blue line shows a model with estimation error only where productivity parameters were fixed at the prior median. The true simulated value is shown by the dark blue point.

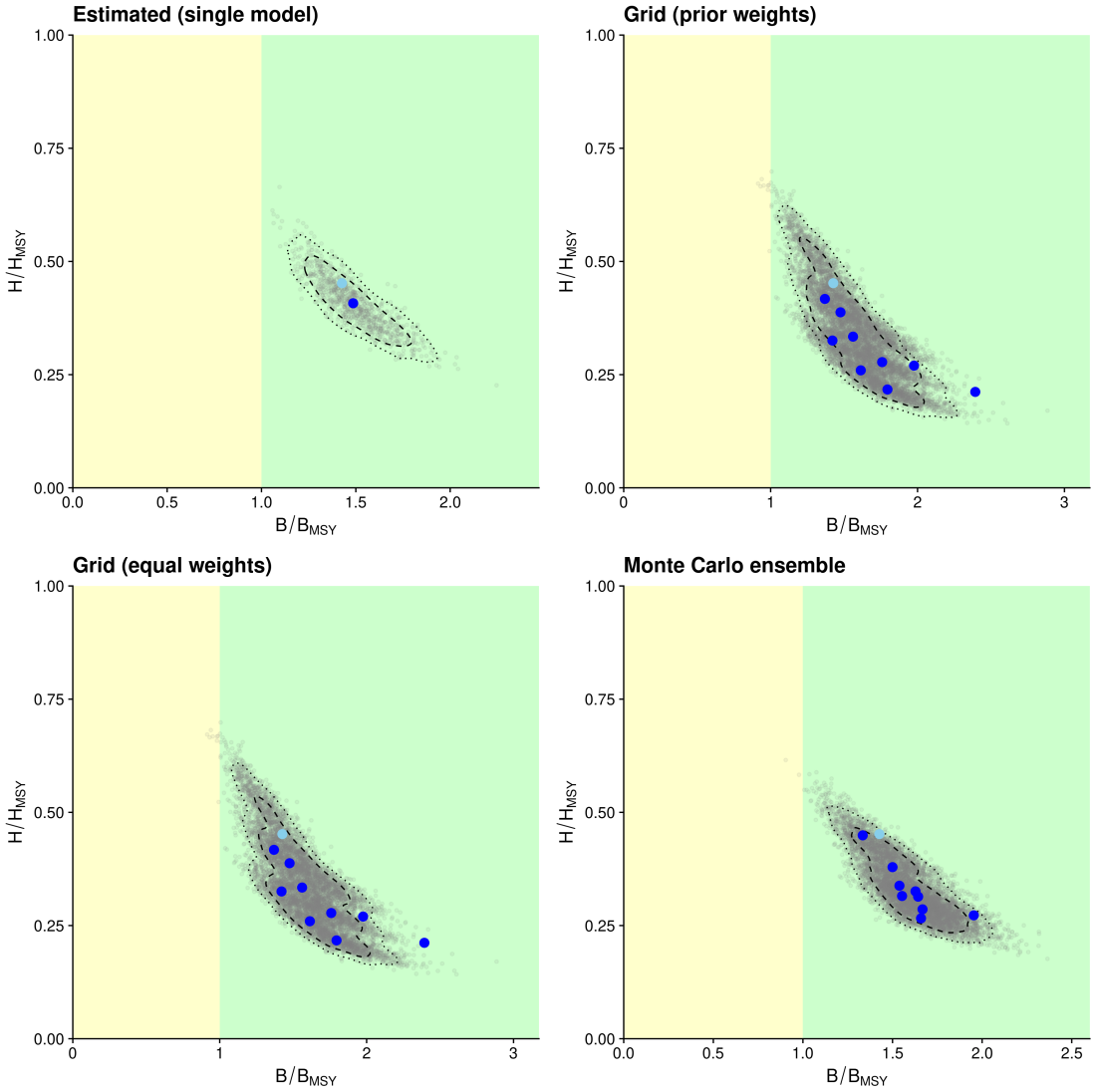


Figure 7: Kobe plot for the four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is biased high by 0.05 for both parameters relative to the true value, with a tight prior (coefficient of variation 10%) representing seemingly good understanding of productivity. Dashed and dotted lines show 80% and 95% credible intervals from Markov chain MC draws, dark blue points show maximum likelihood estimates across model runs for grid approaches, and estimated status (prior mean) for a model with fixed productivity at the prior median. The true simulated value is shown by the light blue point.

4. DISCUSSION

The present review focused broadly on the characterisation of stock assessment uncertainty in management advice, and also specifically on the potential for recent developments in model ensembles applied in the WCPFC to improve uncertainty characterisation.

Accounting for uncertainty stock assessments

In WCPFC assessments, a "structural uncertainty grid" over a large number of models across a fixed parameter grid has often been used to describe the most prevalent approach to SPC stock assessments. This approach was in contrast to the ISC base-case approach, which accounts for estimation uncertainty in a single or small number of models. To reconcile these approaches, Ducharme-Barth and Vincent (2022) suggested a Monte Carlo approach to defining the grid axes based on a joint prior, combined with calculation or simulation of estimation error.

For the set of uncertainties outlined in this review, structural uncertainty grids often integrate over a range of uncertainties. For this reason, the term "structural" may be too imprecise to adequately capture which uncertainties are being addressed. In effect, alternative data sources or treatments, model formulations (e.g., likelihoods or spatial structures) may lead to models with different likelihoods that may be considered "structurally" different. Nevertheless, data weights within these models, or the value of productivity parameters may be considered uncertain parameters—they do not change the structure of the model itself.

The distinction between structural and parameter uncertainty is largely pragmatic (structurally different models may be subsets of a more general model). Nevertheless, from a practical perspective, the distinction identifies parameters that could in theory be estimated. When parameter uncertainly is addressed by estimating parameters, there is an implicit expectation that the data have information that will ultimately reduce the total uncertainty in model outcomes (i.e., the *a posteriori* uncertainty due to estimation is lower than the *a priori* parameter uncertainty). When there is limited information in the data about parameter values, or if estimation is biased (e.g., by unrepresentative data or particular model formulation), it may be necessary to integrate over prior parameter uncertainty only. In this instance, the idealised simulations from this study, and investigations outlined by Ducharme-Barth and Vincent 2022 suggest that the formulation of an explicit prior to either draw from, or weight axes according to a prior over parameters, may be a useful way to avoid overemphasising unrealistic parameter combinations. The decision to integrate over parameter uncertainty within the model, by estimating parameters, or outside of the model, by establishing an ensemble over parameter values, therefore, depends largely on the information content in the data. The latter is often dependent on having sufficient contrast (in biomass levels for estimating steepness, and fishing mortality for estimating natural mortality Magnusson 2016).

Depending on the model specification and the number of parameters estimated, estimation error may be small (e.g., if few parameters are estimated) or large (when many production parameters are estimated, or data are uninformative about estimated parameters). Although the importance of estimation error may difficult to ascertain in the absence of its explicit evaluation, the simulations here and research by Ducharme-

Barth and Vincent (2022) highlight that the inclusion of estimation error may lead to different estimates of management quantities and risk across a model ensemble. Based on these findings, we suggest additional emphasis on addressing this uncertainty in addition to any parameter uncertainty that is addressed via uncertainty grid axes.

Structural uncertainties (e.g., spatial population structure and movement) and uncertainties about data (e.g., alternative CPUE indices) are more difficult to address, especially within the limited time available for any stock assessment. Both types of uncertainty can markedly contribute to overall uncertainty. At a minimum, these uncertainties need to be considered by specifically addressing data quality and structural uncertainties, and their contribution to (often non-quantifiable) overall uncertainty.

Irreducible errors in population parameters, such as recruitment variability and non-stationarity in demographic parameters (i.e., process error), are most appropriately addressed in projections and harvest strategy evaluations. Although models usually incorporate temporally (and spatially) variable recruitment estimates, introducing more than one type of time-varying process in estimation models can reduce retrospective bias, it can also lead to bias in estimated management quantities if mis-specified (Johnson et al. 2015, Stawitz et al. 2019, Szuwalski et al. 2018).

Operationalising the precautionary principle in the WCPFC

Operationalising the precautionary principle requires consistent accounting of uncertainties that contribute to risk (Food and Agriculture Organization of the United Nations 1996). There are legitimate arguments that uncertainty has been used to delay action (Rosenberg 2007), and an undue emphasis on overfishing risk alone may lead to socio-economically detrimental decisions (Hilborn et al. 2001).

The FAO guidelines suggest that harvest strategies be developed that incorporate, and are robust to uncertainties (Punt 2006) to allow efficient decision making in the face of uncertainty. These harvest strategies, when developed in the context of socio-economic objectives can be used to manage risk explicitly while addressing broader objectives of sustainable development and intergenerational equity (Food and Agriculture Organization of the United Nations 1996, Hilborn et al. 2001).

Across tuna RFMOs, management strategies have been or are being developed to give effect to the precautionary approach (Merino et al. 2020). In the WCPFC, Conservation and Management Measure (CMM) 2014-06 and its successor CMM 2022-03 detail the agreed WCPFC process for "[e]stablishing a Harvest Strategy for key fisheries and stocks in the Western and Central Pacific Ocean". In particular, Annex 1 of CMM2022-03 suggest that "[a]s part of this process, the Scientific Committee and other relevant subsidiary bodies, as appropriate shall estimate or describe key uncertainties including with respect to stock assessments and available data". These documents highlight that under WCPFC agreements, uncertainties must be addressed when developing harvest strategies. Nevertheless, the detail about this aspect is currently not specified in CMM 2022-03.

Consistency in addressing uncertainty in WCPFC management advice

This review highlighted two recent key developments for stock assessments: first, although practice has varied markedly between ISC and SPC assessments, there has been a lessening of this difference recently. This rapprochement was evident in more explicit acknowledgement of uncertainty in ISC assessments and also in greater effort to constrain model ensembles towards more plausible configurations for non-tuna SPC assessments.

The complexity of tropical tuna stock assessments conducted by the SPC has provided a barrier to adopting some of the recent approaches; however, assessments currently underway have attempted to incorporate a number of recent suggestions. The long model run time (often up to 24 h) and the complexity of models make it difficult to determine some types of diagnostics that require multiple model runs. In addition, the difficulty to consistently obtain positive definite Hessians for these models makes it difficult to characterise estimation uncertainty. The latter will be increasingly important if productivity parameters such as natural mortality and growth are estimated within assessments. Although "structural" uncertainty grids in SPC assessments integrate across data, structural and parameter uncertainties, they do not consistently include estimation error in the calculation of risk. For this reason, risk estimates across approaches are not directly comparable.

Similar inconsistencies exist for ISC stock assessments, where reported uncertainties range from qualitative statements about the impact of parameter and structural uncertainties to explicit integration over structural uncertainties and estimation error.

Within recent assessments, the current review identified only one assessment that explicitly addressed all potential sources of uncertainty in the provision of management advice. This assessment was the Southwest Pacific swordfish assessment (Ducharme-Barth et al. 2021), which developed a model ensemble that accounted for uncertainty in key input data (by manipulating CPUE CVs), model setup (manipulating weights), production parameters (M, growth, steepness—formulating explicit priors), and estimation error. Because productivity parameters were not estimated, the resulting uncertainty was considerable, including regions where the stock was overfished and undergoing overfishing.

The second development was a move towards more consistent use of risk metrics in the form of probability statements in stock status and management advice. These statements derive from an explicit acknowledgement and calculation of uncertainty. This development has been slow and, currently, advice statements are neither worded consistently, nor consistent in content. To allow for a consistent application of the precautionary principle, as mandated by the WCPFC convention, more consistent reporting should be developed.

International and best practice

Recent projects to develop and weight model ensembles for WCPFC assessments were conducted in the context of global research in the domain both within tuna RFMOs (Maunder & Minte-Vera 2022) and other management bodies, such as ICES (Jardim et al. 2021) and the International Pacific Halibut Commission (Stewart & Hicks 2022). The Center for the Advancement of Population Assessment Methodology (CAPAM)

workshops (i.e., online on model ensembles and weighting in November 2022 (Maunder & Minte-Vera 2022), and on "best practices in tuna stock assessments" in New Zealand in March 2023) explored the topic in some depth. Nevertheless, there has been no clear best practice identified for establishing and weighting model ensembles. At the same time, there are important unknowns concerning the ability of certain diagnostics to reliably identify poorly-performing models. In addition, weighting of data components in assessment models may interact with weighting methods (e.g., models with lower weight for CPUE will have relatively poor performance for diagnostics based on CPUE). By focusing, in a prescriptive way, on certain diagnostics to retain or weight models, biases may be introduced that are currently poorly understood. We, therefore, suggest that assessments be viewed from a broader perspective of key principles; e.g., fitting of CPUE needs to be prioritised and adequate, while leading to acceptable fits elsewhere. Remaining uncertainties need to be adequately described and cataloged to prioritise research to reduce important uncertainties.

A range of methods have been trialed to restrict the number of models or weight models in ensembles for stock assessments in recent years. Fractional factorial designs have been successfully used to restrict the number of models in a grid to low-order interactions only (Hoyle et al. 2008, Kolody et al. 2020), while giving similar outcomes to a full factorial grid. Nevertheless, these approaches are difficult to implement with more than two options per axis (Berger et al. 2013). The bootstrap Monte Carlo approach (Ducharme-Barth & Vincent 2022) can be viewed as a continuous approximation of this notion, because *a priori* unlikely high-order combinations of parameters or structural values are highly unlikely to be drawn at random under a joint prior.

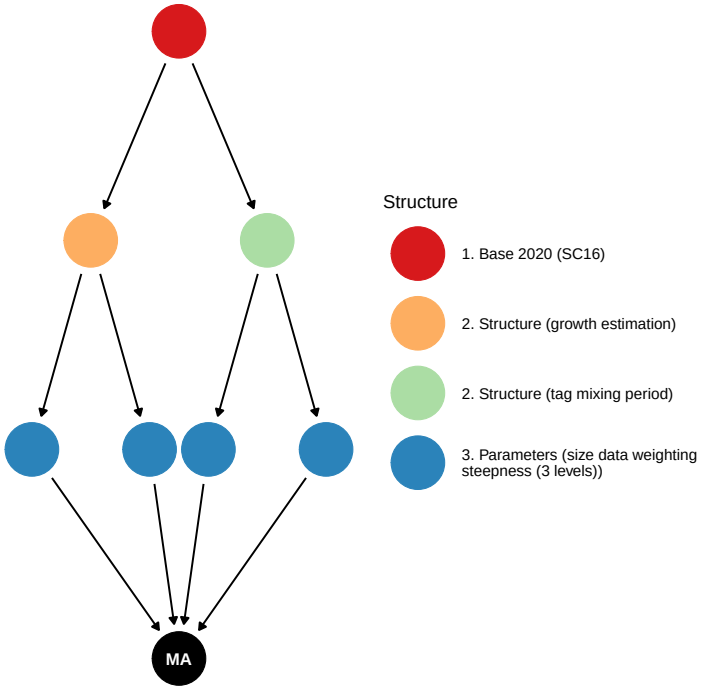
Another related approach that has recently been tried in the WCPFC is the hypothesis-tree approach, where a (or a set of) structural hypothesis (hypotheses) are placed at the base of the tree; within each hypothesis, a number of uncertainties are explored Maunder et al. 2020. This approach has two advantages: first, it limits the number of higher-order interactions (but requires to fix one base level for each structural assumption). For example, the recent yellowfin tuna assessment (Vincent et al. 2020) would have required 35 model runs (assuming no factorial combinations of parameters; 60 models with factorial combinations) instead of 72 model runs (Figure 8). Similarly, the recent SP-BSH (Neubauer et al. 2022b) assessment would have required 80 model runs instead of 648 even with full factorial combinations of parameter uncertainties. A joint prior over parameters at the parameter stage can provide an alternative to a full factorial exploration at this stage. Computational gains may then allow more comprehensive exploration, or inclusion of parameter uncertainty, for a smaller number of models overall. The second advantage of the hypothesis-tree approach is from a conceptual and communication perspective, because this approach clearly distinguishes uncertainties in a series of steps. This aspect provides a clear overview that can be linked to management advice, and also to a loose justification of theoretical Bayesian model averaging (Appendix C).

Although there is no unique best practice about the specifics of formulating and accounting for stock assessment uncertainty, reporting of uncertainty has been researched. In this context, many national (e.g., in Aotearoa New Zealand) and international (e.g., International Council for the Exploration of the Sea, ICES) management agencies use templates for reporting of management advice. Two aspects of advice templates are worth highlighting: first, consistent reporting means that it

is easy for stakeholders to identify key messages (e.g., likelihood of rebuilding or of approaching reference points) from a consistent format. It also ensures that key quantities of interest are consistently reported across all stock assessments. This aspect means there is no reliance on specific stock assessment authors and session conveners to ensure consistency in the format and content of management advice.

A more consistent reporting format can also ensure that results are easily comparable among stocks and years, even if assessment approaches are not the same. Explicitly stating the certainty about key outcomes and qualifying the advice by key uncertainties support a consistent application of the precautionary principle. An example for this reporting are the New Zealand "status of the stocks" reporting tables (Fisheries New Zealand 2022); for example, these tables explicitly reference data quality, key uncertainties, and qualify assessment by type and input data quality. Management quantities (stock status and projections) are reported in standardised terminology (probabilities of exceeding thresholds and reaching targets). At the same time, the management quantities explicitly address calculated uncertainty, and also make provision for unaccounted uncertainties by moderating statements from calculated distributions.

(a) 2021 yellowfin tuna



(b) 2022 South Pacific blue shark

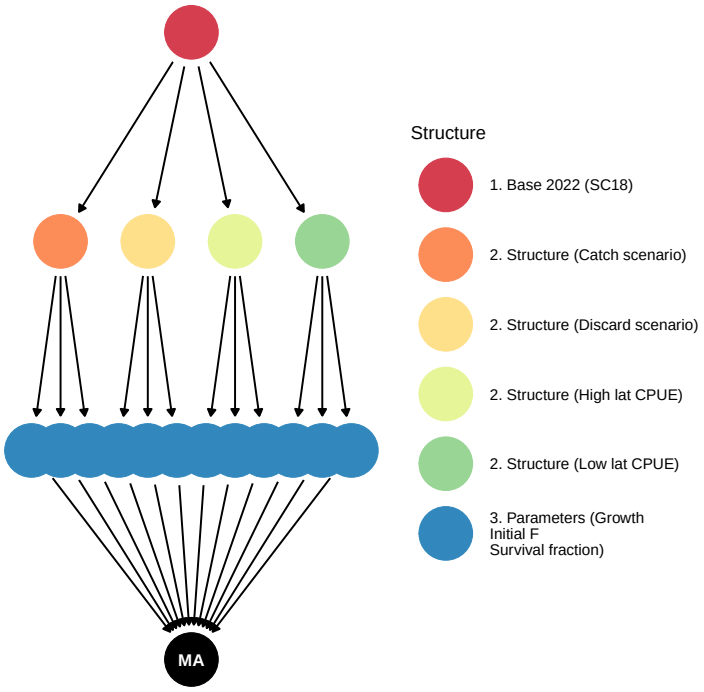


Figure 8: Examples of the stock assessment grid as hypothesis tree, separating structural and parameter uncertainty, and indicating which were used for management advice (MA). Hypothesis trees are for the 2021 yellowfin tuna (a) and the 2022 South Pacific Ocean blue shark (b) stock assessments.

5. RECOMMENDATIONS

Based on the current review of WCPFC stock assessments, simulations, and international best practice, we developed a set of recommendations relating to the use of model ensembles for management advice and the communication of assessment uncertainty. We invite the SC19 to consider the following recommendations, outlined below.

5.1 Model ensembles and weighting

1. **Develop joint priors and explicit rationales for grid axes and their values**

   Where possible, a joint prior should be developed based on life-history or meta-analysis for important but uncertain parameters—especially productivity-related parameters that may not be informed by data (where information is available in the data, fitting parameters may be preferable, (Punt et al. 2021)).

   For non-demographic uncertainties, a consistent rationale could be developed to weight axes; for example, if a default weight for composition data can be identified that leads to acceptable fits for abundance indices while providing fits to composition data (Francis 2011, 2017), then alternative weighting choices (lower or higher weights) may be considered less appropriate, and down-weighted appropriately.

   Similar rationale could be developed for the perceived suitability of alternative datasets, acknowledging that at times, axes may have a marginally uniform prior when no decisions about data suitability can be reached.

2. **Either draw from, or weight axes over parameters according to the joint prior**

   This step mitigates the influence of extreme combinations of parameters that are *a priori* unlikely and that would likely result in more extreme estimates of stock status and management advice.

3. **Consider observation error, structural, parameter, and estimation uncertainty in management advice**

   Current practice is inconsistent in the types of uncertainty that are considered, and how the different types are reported. Although some types of uncertainty can be considered minor, a clear rationale should be developed for prioritising or omitting some types of uncertainty from analyses and management advice. It needs to be clear which uncertainties models and ensembles address, and which ones they considered minor or could not address for technical reasons. For this reason, we suggest a clear terminology around uncertainty.

4. **Where possible, express priors for model outcome space to avoid post-hoc selection/weighting**

   The CAPAM workshops considered it best practice to *a priori* agree on measures to include and/or weight models in an ensemble—with the intent to select models based on objective measures, and avoid selecting models based on results.

   Although some diagnostics can be used in weighting model axes, parameters or model inputs and settings are often uncertain because the data do not contain information to weight or eliminate models along these axes, or data themselves are uncertain.

A joint prior may be formulated for expected outcomes in terms of diagnostics (e.g., models are expected to achieve a certain level of fit to abundance indices) and derived quantities (e.g., biomass distribution or total biomass levels). Similar to input priors, these priors can then be used to weight models or eliminate models that have little or no weight under the outcome prior. Ideally, such an outcome prior should not be based on management-relevant quantities such as biomass or fishing mortality relative to reference points.

5. **Where post-hoc weighting is necessary (unexpected outcomes), it should be proposed by analysts**

   A joint prior over uncertainty axes and outcomes only provides a prior, and does not necessarily constrain models to a sensible outcome space. In particular, with complex models, it may be almost impossible to exhaustively list possible outcomes and their likelihood in an outcome prior. For these models, it may be necessary to further constrain the model grid based on diagnostics.

   We suggest that these diagnostics and decisions to subset model ensembles should be undertaken by the assessment team, and presented to the SC. Although the decision to accept ensembles lies with the SC, we recommend that alternative ensembles are not constructed as part of the SC meeting. This recommendation would avoid decision-making by interested parties based on management-relevant quantities.

5.2 Communicating uncertainty and risk

1. **Develop a template for reporting management advice and uncertainties**

   A straightforward way to improve and standardise reporting of uncertainty and risk is to develop a template for reporting uncertainties alongside the provision of stock status and management advice. The international expert group on "Addressing Uncertainty in Fisheries Science and Management" convened by the National Aquarium in the United States in 2015 suggested that an "[i]nnovative approach" would be to "create a table or checklist indicating the major sources of uncertainty for that fishery, how they are addressed and by whom, and at what point in the process they are considered...This tool would promote understanding among all participants and would also highlight to all how the system already accounts for certain types of uncertainty and where effort needs to be focused to address concerns" (Cadrin et al. 2015). In addition, clear addressing of uncertainties is important in the context of adequately representing risk when developing harvest strategies, to ensure that these strategies are robust to key uncertainties.

   Templates for reporting assessment advice are in use in a number of jurisdictions and councils. For example, ICES provides a standardised structure to provide management advice, and New Zealand's plenary reports have a standardised tabulated reporting format for both status and uncertainty (see "Guidelines for Status of the Stocks Summary Tables" in Fisheries New Zealand 2022). Standardised table formats help managers to focus on key quantities. The following points detail aspects of this type of framework. Development of the latter needs to consider international best practice in relation to WCPFC convention requirements. For this reason, we propose the following recommendations for

the process of developing a framework.

2. **Agreed terminology and set of required measures**

   Consistent terminology is an important aspect of efficiently communicating advice and risk. Categories with associated labels may be used to communicate uncertainty categories and associated risk. For example, the standardised status of the stock tables in New Zealand distinguishes stock status and advice based on the certainty of the statements, following the classification of the Intergovernmental Panel on Climate Change. The classification rates statements as certain (>99% likelihood), highly likely (>90%), likely (>60%), about as likely as not (40–60%), unlikely (<40%), very unlikely (<10%), or exceptionally unlikely (<1%).

   Similarly, types of uncertainties need to be clearly identified. Although the distinction between types of uncertainty is not always clearly-defined (e.g., structural versus parameter uncertainty), a practical and useful classification should be established to allow consistent reporting of uncertainty.

3. **Clear communication about quality of information determining stock status and management advice**

   Where possible, reports of stock assessments should include sections about key uncertainties, how they were addressed, and which uncertainties remain. These uncertainties include:

   - **Qualification and quantification of uncertainties.**
     (a) Data quality.
     (b) Model/population: structural uncertainty (note the use of "structural" here refers to models with different likelihoods, rather than different parameter values).
     (c) Key parameters (parameter and estimation uncertainty).
   - Key uncertainties and potential impacts: qualification of how these uncertainties (listed above) influence management advice.

4. **Based on identified uncertainties, develop a set of research recommendations to address key uncertainties.**

   To improve assessments and reduce uncertainty and risk over time, uncertainties need to be clearly identified and ranked in their importance for determining risk in management advice.

5. **A review of timelines and capacity for tuna stock assessment**

   The review of timelines may be required to allow sufficient time and capacity to adequately address uncertainty. Sufficient time is also needed to enable the provision of management advice that is consistent with the application of the precautionary approach, as outlined in the WCPFC convention text. A similar review was recently conducted for shark stock assessments, which may provide a precedent to further develop the tuna stock assessment process.

5.3 Further development and future research

We further suggest that the SC consider recommending the following suggestions:

- The provision a project to develop a standardised reporting template for uncertainty and risk reporting that accounts for recommendations made in the present review, and
- further development of methodology and idealised simulations to develop principled model ensemble approaches, in particular to consider the ability of alternative model diagnostics to identify model plausibility and weights.

ACKNOWLEDGMENTS

The authors thank the ISC and SPC stock assessment teams for helpful discussions about their respective stock assessment approaches, and Paul Hamer at SPC for organising the pre-assessment workshop, which provided stimulating discussions for the present project.

6. REFERENCES

Berger, A. M.; Pilling, G. M.; Harley, S. J., & Kirchner, C. (2013). Approaches to describe uncertainty in current and future stock status. WCPFC-SC9-2013/MI-WP-04. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Nineth Regular Session Pohnpei, Federated States of Micronesia, 6–14 August 2013.

Cadrin, S.; Henderschedt, J.; Mace, P.; Mursalski, S.; Powers, J.; Punt, A., & Restrepo, V. (2015). *Addressing uncertainty in fisheries science and management*. National Aquarium. Retrieved from https://www.fao.org/3/bf336e/bf336e.pdf. 40 p.

Carpenter, B.; Gelman, A.; Hoffman, M.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M. A.; Guo, J.; Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Clarke, S.; Langley, A.; Lennert-Cody, C.; Aires-da-Silva, A., & Maunder, M. (2018). *Pacific-wide silky shark (*Carcharhinus falciformis*) stock status assessment*. WCPFC-SC14-2018/SA-WP-08. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourteenth Regular Session Busan, Korea, 8–16 August 2018.

Davies, N.; Hoyle, S., & Hampton, J. (2012). Stock assessment of striped marlin (*Kajikia audax*) in the southwest Pacific Ocean. WCPFC-SC8-2012/SA-WP-05. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eighth Regular Session. Busan, Republic of Korea, 7–15 August 2012.

de Bruyn, P.; Murua, H., & Aranda, M. (2013). The precautionary approach to fisheries management: How this is taken into account by tuna regional fisheries management organisations (RFMOs). *Marine Policy*, *38*, 397–406. doi:10.1016/j.marpol.2012.06.019.

Draper, D. (1995, January). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45–70. doi:10.1111/j.2517-6161.1995.tb02015.x.

Ducharme-Barth, N.; Castillo-Jordán, C.; Hampton, J.; Williams, P.; Pilling, G., & Hamer, P. (2021). *Stock assessment of southwest pacific swordfish*. WCPFC-SC17-2021/SA-WP-04. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Seventeenth Regular Session. Online, 11–19 August 2021

Ducharme-Barth, N.; Pilling, G., & Hampton, J. (2019). *Stock assessment of SW Pacific striped marlin in the WCPO*. WCPFC-SC15-2019/SA-WP-07. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fifteenth Regular Session. Pohnpei, Federated States of Micronesia, 12–20 August 2019.

Ducharme-Barth, N.; Vincent, M.; Hampton, J.; Hamer, P., & Williams, P. (2020). *Stock assessment of bigeye tuna in the western and central Pacific Ocean*. WCPFC-SC16-2020/SA-WP-03-Rev3. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Sixteenth Regular Session. Online, 11–20 August 2020.

Ducharme-Barth, N. D. & Vincent, M. T. (2022). Focusing on the front end: A framework for incorporating uncertainty in biological parameters in model ensembles of integrated stock assessments. *Fisheries Research*, *255*, 106452. doi:10.1016/j.fishres.2022.106452

Fisheries New Zealand (2022). *Fisheries Assessment Plenary, May 2022: stock assessments and stock status*. Fisheries New Zealand, Wellington, New Zealand.

Food and Agriculture Organization of the United Nations (1996). Precautionary approach to capture fisheries and species introductions. Retrieved from https://www.fao.org/3/w3592e/w3592e.pdf.

Fournier, D.; 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*. *Canadian Journal of Fisheries and Aquatic Sciences*, *55*, 2105–2116.

Fournier, D.; Sibert, J. R.; Majkowski, J., & Hampton, J. (1990). MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). *Canadian Journal of Fisheries and Aquatic Sciences*, *47*(2), 301–317.

Francis, R. I. C. C. & Shotton, R. (1997). Risk in fisheries management: A review. *Canadian Journal of Fisheries and Aquatic Sciences*, *54*, 1699–1715.

Francis, R. I. C. C. (2011). Data weighting in statistical fisheries stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*(6), 1124–1138. doi:10.1139/f2011-025

Francis, R. I. C. C. (2017). Revisiting data weighting in fisheries stock assessment models. *Fisheries Research*, *192*, 5–15. doi:10.1016/j.fishres.2016.06.006

Fu, D.; Clarke, M.-J. R. S.; Francis, M.; Dunn, A.; Hoyle, S., & Edwards, C. (2017). *Pacific-wide sustainability risk assessment of bigeye thresher shark (*Alopias superciliosus*)*. WCPFC-SC13-2017/SA-WP-11-Rev3. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

Hilborn, R.; Maguire, J.-J.; Parma, A. M., & Rosenberg, A. A. (2001). The precautionary approach and risk management: Can they increase the probability of successes in fishery management? *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(1), 99–107. doi:10.1139/f00-225.

Hilborn, R.; Pikitch, E. K., & Francis, R. C. (1993). Current trends in including risk and uncertainty in stock assessment and harvest decisions. *Canadian Journal of Fisheries and Aquatic Sciences*, *50*(4), 874–880. doi:10.1139/f93-100.

Hoeting, J. A.; Madigan, D.; Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci*, *14*, 382–417.

Hoyle, S. D.; Edwards, C. T. T.; Roux, M.-J.; Clarke, S. C., & Francis, M. P. (2017). *Southern Hemisphere porbeagle shark (*Lamna nasus*) stock status assessment*. WCPFC-

SC13-2017/SA-WP-12-Rev2. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

Hoyle, S.; Bouyé, F.; Langley, A., & Hampton, J. (2008). Sensitivity of the bigeye stock assessment to alternative structural assumptions. WCPFC-SC4-2008/SA-WP-3. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourth Regular Session Port Moresby, Papua New Guinea, 11–22 August 2008.

International Commission for the Conservation of Atlantic tunas (2022). *Report of the Standing Committee on Research and Statistics (SCRS)*. Retrieved from https://www.iccat.int/Documents/Meetings/Docs/2022/REPORTS/2022_SCRS_ENG.pdf. Madrid (Spain)/Hybrid, 26–30 September 2022 (Revision, 6 October 2022)

ISC (2015a). *Indicator-based analysis of the status of shortfin mako shark in the North Pacific Ocean*. WCPFC-SC11-2015/SA-WP-08. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eleventh Regular Session. Pohnpei, Federated States of Micronesia, 5–13 August 2015.

ISC (2015b). *Stock assessment update for striped marlin (*Kajikia audax*) in the western and central North Pacific Ocean through 2013*. WCPFC-SC11-2015/SA-WP-10. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eleventh Regular Session. Pohnpei, Federated States of Micronesia, 5–13 August 2015.

ISC (2016). *Stock assessment update for blue marlin (*Makaira nigricans*) in the Pacific Ocean through 2014*. WCPFC-SC12-2016/SA WP-12. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Twelfth Regular Session. Bali, Indonesia, 3–11 August 2016.

ISC (2017a). *Stock assessment and future projections of blue shark in the North Pacific Ocean through 2015*. WCPFC-SC13-2017/SA-WP-10. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

ISC (2017b). *Stock assessment of albacore in the North Pacific Ocean in 2017*. WCPFC-SC13-2017/SA-WP-09-Rev2. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

ISC (2018). *Stock assessment of shortfin mako shark in the North Pacific through 2016*. WCPFC-SC14-2018/SA-WP-11. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourteenth Regular Session, 8–16 August, Busan, Korea.

ISC (2019). Stock assessment report for striped marlin (*Kajikia audax*) in the western and central North Pacific Ocean through 2017. WCPFC-SC15-2019/SA-WP-09. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fifteenth Regular Session. Pohnpei, Federated States of Micronesia, 12–20 August 2019.

ISC (2020a). *Stock assessment of albacore tuna in the North Pacific Ocean in 2020*. WCPFC-SC16-2020/SA-WP-05. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Sixteenth Regular Session. Online, 11–20 August 2020.

ISC (2020b). *Stock assessment of Pacific bluefin tuna in the Pacific Ocean in 2020*. WCPFC-SC16-2020/SA-WP-06. Report to the Western and Central Pacific Fisheries

Commission Scientific Committee. Sixteenth Regular Session. Online, 11–20 August 2020.

ISC (2021). *Stock assessment report for Pacific blue marlin (*Makaira nigricans*) through 2019*. WCPFC-SC17-2021/SA-WP-08. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Seventeenth Regular Session. Electronic Meeting, 11–19 August 2021.

ISC (2022a). *Stock assessment and future projections of blue sharks in the North Pacific Ocean through 2020*. WCPFC-SC18-2022/SA-WP-06. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eightteenth Regular Session. Electronic Meeting, 10–18 August 2022.

ISC (2022b). *Stock assessment of Pacific bluefin tuna in the Pacific Ocean in 2022*. WCPFC-SC18-2022/SA-WP-05. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eightteenth Regular Session. Electronic Meeting, 10–18 August 2022.

ISC Billfish Working Group (2018). *Stock assessment for swordfish (*Xiphias gladius*) in the western and central North Pacific Ocean through 2016*. WCPFC-SC14-2018/SA-WP-07-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourteenth Regular Session Busan, Korea, 8–16 August 2018.

Jardim, E.; Azevedo, M.; Brodziak, J.; Brooks, E. N.; Johnson, K. F.; Klibansky, N.; Millar, C. P.; Minto, C.; Mosqueira, I.; Nash, R. D., et al. (2021). Operationalizing ensemble models for scientific advice to fisheries management. *ICES Journal of Marine Science*, *78*(4), 1209–1216.

Johnson, K. F.; Monnahan, C. C.; McGilliard, C. R.; Vert-pre, K. A.; Anderson, S. C.; Cunningham, C. J.; Hurtado-Ferro, F.; Licandeo, R. R.; Muradian, M. L.; Ono, K., et al. (2015). Time-varying natural mortality in fisheries stock assessment models: Identifying a default approach. *ICES Journal of Marine Science*, *72*(1), 137–150.

Jordan, C. C.; Hampton, J.; Ducharme-Barth, N.; Xu, H.; Vidal, T.; Williams, P.; Scott, F.; Pilling, G., & Hamer, P. (2021). *Stock assessment of South Pacific albacore tuna*. WCPFC-SC17-2021/SA-WP-02-Rev2. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Seventeenth Regular Session. Electronic Meeting, 11–19 August 2021.

Jordan, C. C.; Teears, T.; Hampton, J.; Davies, N.; Phillips, J. S.; McKechnie, S.; Peatman, T.; Macdonald, J.; Day, J.; Magnusson, A.; Scott, R.; Scott, F.; Pilling, G., & Hamer, P. (2022). *Stock assessment of skipjack tuna in the western and central Pacific Ocean*. WCPFC-SC18-2022/SA-WP-01-Rev5. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eightteenth Regular Session. Electronic Meeting, 10–18 August 2022.

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 Journal of Marine Science*, *78*(6), 2244–2255.

Kolody, D.; Jumppanen, P., & Day, J. (2020). Indian Ocean bigeye tuna Management Procedure Evaluation Update March 2020. IOTC-2020-WPM11-11. Prepared for the IOTC MSE Task Force, originally scheduled for March 2020.

Large, K.; Neubauer, P., & Brouwer, S. (2022). *Stock assessment of southwest pacific shortfin mako shark*. WCPFC-SC18-2022/SA-WP-02-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eightteenth Regular Session. Electronic Meeting, 10–18 August 2022.

Mace, P. M. & Doonan, I. (1988). *A generalised bioeconomic simulation model for fish population dynamics*. MAFFish, NZ Ministry of Agriculture and Fisheries.

Magnusson, A. (2016). *Informative data and uncertainty in fisheries stock assessment* (Doctoral dissertation, University of Washington, United States).

Magnusson, A.; Punt, A. E., & Hilborn, R. (2013). Measuring uncertainty in fisheries stock assessment: The delta method, bootstrap, and MCMC. *Fish and Fisheries*, *14*(3), 325–342. doi:10.1111/j.1467-2979.2012.00473.x.

Maunder, M. N. & Minte-Vera, C. (2022). *2nd workshop on improving the risk analysis for tropical tunas in the eastern pacific ocean: Model weighting in integrated stock assessments: Chair's reports*.

Maunder, M. N.; Xu, H.; Lennert-Cody, C. E.; Valero, J. L.; Aires-da-Silva, A., & Minte-Vera, C. (2020). Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses (no. sac-11-inf-f). *Scientific Advisory Commitee, Inter-American Tropical Tuna Commission, San Diego*.

McKechnie, S.; Pilling, G., & Hampton, J. (2017). *Stock assessment of bigeye tuna in the western and central Pacific Ocean*. WCPFC-SC13-2017/SA-WP-05-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

Merino, G.; Murua, H.; Santiago, J.; Arrizabalaga, H., & Restrepo, V. (2020, October 7). Characterization, communication, and management of uncertainty in tuna fisheries. *Sustainability*, *12*(19), 8245. doi:10.3390/su12198245

Neubauer, P.; Carvalho, F.; Ducharme-Barth, N.; Large, K.; Brouwer, S.; Day, J., & Hamer, P. (2022a). *Report on WCPFC project 107b: Improved stock assessment and structural uncertainty grid for Southwest Pacific blue shark*. WCPFC-SC18-2022/SA-WP-03 (Rev.01). Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eighteenth Regular Session, 10–18 August 2022. Electronic meeting.

Neubauer, P.; Carvalho, F.; Ducharme-Barth, N.; Large, K.; Brouwer, S.; Day, J., & Hamer, P. (2022b). *Report on WCPFC project 107b: Improved stock assessment and structural uncertainty grid for Southwest Pacific blue shark*. WCPFC-SC18-2022/SA-WP-03-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eightteenth Regular Session. Electronic Meeting, 10–18 August 2022.

Neubauer, P.; Large, K., & Brouwer, S. (2021). *Stock assessment of southwest pacific blue shark*. WCPFC-SC17-2021/SA-WP-03-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Seventeenth Regular Session. Electronic Meeting, 11–19 August 2021.

Neubauer, P.; Richard, Y., & Clarke, S. (2018). *Risk to the Indo-Pacific Ocean whale shark population from interactions with Pacific Ocean purse-seine fisheries*. WCPFC-SC14-2018/SA-WP-12-Rev2. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourteenth Regular Session Busan, Korea, 8–16 August 2018.

Polacheck, T.; Hilborn, R., & Punt, A. E. (1993). Fitting surplus production models: Comparing methods and measuring uncertainty. *Canadian Journal of Fisheries and Aquatic Sciences*, *50*(12), 2597–2607.

Privitera-Johnson, K. M. & Punt, A. E. (2020). A review of approaches to quantifying uncertainty in fisheries stock assessments. *Fisheries Research*, *226*, 105503. doi:10.1016/j.fishres.2020.105503.

Punt, A. E. (2006). The FAO precautionary approach after almost 10 years: Have we progressed towards implementing simulation-tested feedback-control

management systems for fisheries management? *Natural Resource Modeling*, *19*(4), 441–464. doi:10.1111/j.1939-7445.2006.tb00189.x.

Punt, A. E.; Maunder, M. N., & Ianelli, J. N. (2023). *Independent review of recent WCPO yellowfin tuna assessment*. WCPFC-SC19-2023/SA-WP-01. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Nineteenth Regular Session. Koror, Palau, 16–24 August 2023.

Punt, A.; Butterworth, D., & Penney, A. (1995). Stock assessment and risk analysis for the South Atlantic population of albacore *Thunnus alalunga* using an age-structured production model. *South African Journal of Marine Science*, *16*(1), 287–310.

Punt, A. E.; Castillo-Jordán, C.; Hamel, O. S.; Cope, J. M.; Maunder, M. N., & Ianelli, J. N. (2021). Consequences of error in natural mortality and its estimation in stock assessment models. *Fisheries Research*, *233*, 105759.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rice, J. & Harley, S. (2012a). Stock assessment of oceanic whitetip sharks in the western and central Pacific Ocean. WCPFC-SC8-2012/SA-WP-06-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eighth Regular Session. Busan, Republic of Korea, 7–15 August 2012.

Rice, J. & Harley, S. (2012b). Stock assessment of silky sharks in the western and central Pacific Ocean. WCPFC-SC8-2012/SA-WP-07-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Eighth Regular Session. Busan, Republic of Korea, 7–15 August 2012.

Rice, J. & Harley, S. (2013). *Updated stock assessment of silky sharks in the western and central pacific ocean*. WCPFC-SC9-2013/SA-WP-03. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Ninth Regular Session. Pohnpei, Federated States of Micronesia, 6–14 August 2013.

Rosenberg, A. A. (2007). Fishing for certainty. *Nature*, *449*(7165), 989–989. doi:10.1038/449989a.

Rosenberg, A. A. & Restrepo, V. R. (1994). Uncertainty and risk evaluation in stock assessment advice for U.S. marine fisheries. *Canadian Journal of Fisheries and Aquatic Sciences*, *51*(12), 2715–2720. doi:10.1139/f94-271.

Stawitz, C. C.; Haltuch, M. A., & Johnson, K. F. (2019). How does growth misspecification affect management advice derived from an integrated fisheries stock assessment model? *Fisheries Research*, *213*, 12–21.

Stewart, I. & Hicks, A. (2022). *Assessment of the Pacific halibut (Hippoglossus stenolepis) stock at the end of 2022*. IPHC-2023-SA-01 Prepared by: IPHC Secretariat.

Szuwalski, C. S.; Ianelli, J. N., & Punt, A. E. (2018). Reducing retrospective patterns in stock assessment and impacts on management performance. *ICES Journal of Marine Science*, *75*(2), 596–609.

Takeuchi, Y.; Pilling, G., & Hampton, J. (2017). *Stock assessment of swordfish (*Xiphias gladius*) in the southwest Pacific Ocean*. WCPFC-SC13-2017/SA-WP-13. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

Takeuchi, Y.; Tremblay-Boyer, L.; Pilling, G. M., & Hampton, J. (2016). Assessment of blue shark in the southwestern Pacific. WCPFC-SC12-2016/SA-WP-08-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Twelfth Regular Session. Bali, Indonesia, 3–11 August 2016.

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

Tremblay-Boyer, L.; Carvalho, F.; Neubauer, P., & Pilling, G. (2019). *Stock assessment for oceanic whitetip shark in the Western and Central Pacific Ocean*. WCPFC-SC15-2019/SA-WP-06. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fifteenth Regular Session. Pohnpei, Federated States of Micronesia, 12–20 August 2019.

Tremblay-Boyer, L.; Hampton, J.; McKechnie, S., & Pilling, G. (2018). *Stock assessment of south pacific albacore tuna*. WCPFC-SC14-2018/SA-WP-05-Rev2. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourteenth Regular Session Busan, Korea, 8–16 August 2018.

Tremblay-Boyer, L.; McKechnie, S.; Pilling, G., & Hampton, J. (2017). *Stock assessment of yellowfin tuna in the western and central pacific ocean*. WCPFC-SC13-2017/SA-WP-06-Rev1. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Thirteenth Regular Session, Rarotonga, Cook Islands, 9–17 August 2017.

Vincent, M. T.; Pilling, G. M., & Hampton, J. (2019). *Stock assessment of skipjack tuna in the western and central Pacific Ocean*. WCPFC-SC15-2019/SA-WP-05-Rev2. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fifteenth Regular Session. Pohnpei, Federated States of Micronesia, 12–20 August 2019.

Vincent, M.; Ducharme-Barth, N.; Hamer, P.; Hampton, J.; Williams, P., & Pilling, G. (2020). *Stock assessment of yellowfin tuna in the western and central Pacific Ocean*. WCPFC-SC16-2020/SA-WP-04-Rev3. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Sixteenth Regular Session. Online, 11–20 August 2020.

Yao, Y.; Vehtari, A.; Simpson, D.; Gelman, A., et al. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007.

APPENDIX A SIMULATION EXPERIMENTS: MODEL DETAIL

For the simulation-estimation experiments, we developed a length-based, age-structured model, which followed the MULTIFAN-CL model (Fournier et al. 1990, Fournier et al. 1998) in a simplified form. The development of the model was motivated by the need to evaluate the performance of ensemble approaches in the context of the length-based, age-structured model (Fournier et al. 1998), which was used for a number of tuna stock assessments. The operating model to simulate data was developed in *R* (R Core Team 2021), and the estimation model was developed in *Stan* (Carpenter et al. 2017) to allow efficient estimation.

The simulation study was conducted in four steps. First, we simulated an age-structured population dynamics of an albacore-like stock using the age-structured model. Second, we generated abundance indices and length-frequency data from the simulated population dynamics. Third, we applied the same age-structured model to the simulated data, where we assumed two different sets of joint prior probability distributions (e.g., prior medians matched with true values versus prior medians not matching true values) for the two focal parameters, steepness and natural mortality, and estimated management quantities using different ensemble weightings. In the fourth step, we compared the estimated management quantities with the true values to evaluate the performance of the different ensemble weightings.

This section provides a description of the age-structured model, developed for this simulation study. To mimic albacore tuna population dynamics in the South Pacific Ocean, we used the parameter values from a previous study (Punt et al. 1995). The catch data for the albacore stock were used to derive annual exploitation rates, and to generate pseudo data (Q: true values with random error applied?) on abundance index and length composition. The annual catch data for 1967 to 2019 were obtained from published literature (Polacheck et al. 1993, International Commission for the Conservation of Atlantic tunas 2022).

A.1 Model structure

The age-structured model was divided into two parts: process models and observation models (see Tables A-1 and A-2 for definitions of parameters used in these models). The description here first provides information of process models, where the top hierarchy describes age-structured transitions of the population, and the bottom hierarchies describe the life history and mortality processes of the population. Following this description, the observation models are presented, which were used to link unobserved quantities (e.g., stock size) in the process models to (simulated) observed quantities (e.g., abundance index and length composition data).

Table A-1: Definitions for the terms used in the age-structured model. DM, Dirichlet-multinomial.

Notation	Description
a	Index for ages.
A	Maximum age.
j	Index for length classes (bins).
r	Bin width.
J	Index for the last bin.
t	Index for years.
T	Index for the last year.
$N_{a,t}$	Abundance of fish of age a at year t.
H_t	Exploitation rate at year t.
Y_t	Yield (i.e., catch in weight) at year t.
$\tilde{N}_a$	Abundance of fish of age a at unexploited equilibrium.
R_0	Recruitment at unexploited equilibrium.
R_t	Recruitment at year t.
σ_R^2	Variance of recruitment deviations.
M	Instantaneous rate of natural mortality.
h	Steepness parameter.
$G_{j\|a}$	Probability of fish of age a being in length bin j.
l_a	Mean length at age a .
L_∞, κ, a_0	von Bertalanffy parameters.
σ_a	Standard deviation of the length-at-age distribution.
λ_1, λ_2	Parameters that determine σ_a.
ζ	Brody coefficient (i.e., $\zeta = e^{-\kappa}$).
SSB_t	Spawning stock biomass at year t.
SSB_0	Spawning stock biomass at unexploited equilibrium.
W_a	Mean weight at age a.
ω_1, ω_2	Length-weight relationship parameters.
S_a	Selectivity at age a.
a_{50}, ν	Selectivity parameters.
Mat_a	Maturity at age a.
a_{Mat}	Age at 100% maturity.
VB_t	Vulnerable biomass at year t.
VB_0	Vulnerable biomass at unexploited equilibrium.
φ	Proportion of females.
I_t	Abundance index at year t.
q	Catchability coefficient.
τ^2	Observation error variance of the abundance index I_t.
$\boldsymbol{n}_t$	Vector of number of fish in length bin j at year t in the length composition data.
E_t	Total sample size of the length-composition data at year t.
E_t^{eff}	Effective sample size of the length-composition data at year t.
$\boldsymbol{\delta}_t$	Vector of the concentration parameters at year t in the DM model.
$\delta_{j,t}$	Concentration parameter of length bin j at year t in the DM model.
θ	Parameter that determines the value of $\delta_{j,t}$ in the DM model and E_t^{eff}.
$\hat{\boldsymbol{P}}_t$	Vector of the model-estimated length-composition proportions at year t.
$\hat{P}_{j\|t}$	Model-estimated length-composition proportion of length bin j at year t.
$\hat{P}_{a\|t}$	Model-estimated age-composition proportion of age a at year t.

Table A-2: Prior distributions for the parameters used in this simulation study (not including steepness h and natural mortality M).

Description	Prior distributions
Recruitment of unexploited equilibrium	$\log(R_0) \sim \text{uniform}(13, 17)$
Catchability coefficient	$\log(q) \sim \text{uniform}(-6.908, 0)$
Observation error variance	$\tau^2 \sim \text{inverse-gamma}(0.001, 0.001)$
Dirichlet Multinomial parameter	$\theta \sim \text{exponential}(1)$

A.1.1 Process models

Age-structured dynamics The age-structured dynamics of the population are defined as:

$$N_{a,t} = \begin{cases} R_t, & \text{for } a = 1 \\ N_{a-1,t-1} \cdot (1 - S_{a-1} \cdot H_{t-1}) \cdot e^{-M}, & \text{for } 1 < a < A, \\ N_{a-1,t-1} \cdot (1 - S_{a-1} \cdot H_{t-1}) \cdot e^{-M} + N_{a,t-1} \cdot (1 - S_a \cdot H_{t-1}) \cdot e^{-M}, & \text{for } a = A \end{cases}$$

where $N_{a,t}$ is the abundance of fish of age a at the beginning of year t, R_t is the recruitment at age 1 in year t, S_a is the time-invariant age-dependent selectivity, H_t is the exploitation rate ($0 \leq H_t \leq 1$) in year t, A is the maximum age which is the plus group, and M is the instantaneous rate of natural mortality.

The initial population (i.e., $N_{a,1}$) was assumed to be at unexploited equilibrium:

$$\tilde{N}_a = \begin{cases} R_0, & \text{for } a = 1 \\ \tilde{N}_{a-1} \cdot e^{-M}, & \text{for } 1 < a < A \\ \dfrac{\tilde{N}_{a-1} \cdot e^{-M}}{1 - e^{-M}}, & \text{for } a = A \end{cases},$$

where $\tilde{N}_a$ is the abundance of fish of age a at unexploited equilibrium, R_0 is the recruitment at age 1 at unexploited equilibrium, M is the instantaneous rate of natural mortality, and A is the maximum age which is the plus group.

Stock-recruitment relationship The Beverton-Holt stock-recruitment function, reparameterised in terms of the steepness parameter h (Mace & Doonan 1988), was used to model the annual recruitment at age 1 R_t:

$$R_{t+1} = \frac{4 \cdot h \cdot R_0 \cdot SSB_t}{(1 - h) \cdot SSB_0 + (5 \cdot h - 1) \cdot SSB_t} \cdot e^{\varepsilon_{t+1} - 0.5 \cdot \sigma_R^2},$$

where SSB_t is the spawning stock biomass in year t, SSB_0 is the spawning stock biomass at unexploited equilibrium, ε_t are the annual recruitment deviations, which are normally distributed with mean 0 and variance σ_R^2, R_0 is the recruitment at age 1 at unexploited equilibrium, and the subtracted term $-0.5 \cdot \sigma_R^2$ is the bias correction term.

Length-at-age distribution of the catch For simplicity, the length-at-age distribution of the catch $G_{j|a}$ was assumed to be the same as the length-at-age distribution of the population (i.e., there is no length-dependent selectivity):

$$G_{j|a} = \begin{cases} \int_0^{\bar{L}_j + r/2} f(L|l_a, \sigma_a^2) dL, & \text{for } j = 1 \\ \int_{\bar{L}_i - r/2}^{\bar{L}_j + r/2} f(L|l_a, \sigma_a^2) dL, & \text{for } 1 < j < J \\ 1 - \int_0^{\bar{L}_j - r/2} f(L|l_a, \sigma_a^2) dL, & \text{for } j = J \end{cases}$$

where r is the length bin width, $\bar{L}_j$ is the midpoint of the length bin j, l_a is the mean length-at-age, σ_a is the standard deviation of the length-at-age distribution, $f(L|l_a, \sigma_a^2)$ is the normal density of the random variable L with mean l_a and variance σ_a^2, and J is the last length bin.

The mean length-at-age, l_a, was modelled with the von Bertalanffy function:

$$l_a = L_\infty \cdot [1 - e^{-\kappa \cdot (a - a0)}],$$

where L_∞ is the asymptotic length, κ is the growth rate, and a_0 is the theoretical age at length 0.

The standard deviation of the length-at-age distribution σ_a was modelled with the function, which was adopted from the MULTIFAN-CL model (Fournier et al. 1990, Fournier et al. 1998):

$$\sigma_a = \lambda_1 \cdot e^{\lambda_2 \cdot \left[-1 + 2 \cdot \left(\frac{1 - \zeta^{a-1}}{1 - \zeta^{A-1}} \right) \right]},$$

where λ_1 determines the scale of the standard deviations, λ_2 determines the length-dependent increase in the standard deviations, and ζ is the Brody growth coefficient (i.e., $\zeta = e^{-\kappa}$).

Length-weight relationship The mean weight-at-age, W_a, was modelled with the length-weight relationship function:

$$W_a = \omega_1 \cdot \mu_a^{\omega_2},$$

where ω_1 and ω_2 are the two parameters which determine the allometric curve.

Selectivity The age-dependent selectivity, S_a, was modelled with a two-parameter logistic curve:

$$S_a = \frac{1}{1 + e^{-(a - a_{50})/\nu}},$$

where a_{50} is the age at 50% selectivity, and ν is the slope of the curve.

Maturity The sexual maturity-at-age Mat_a was assumed to follow a knife-edged function (Punt et al. 1995):

$$Mat_a = \begin{cases} 0 & \text{if } a < a_{Mat} \\ 1 & \text{if } a \geq a_{Mat} \end{cases},$$

where a_{Mat} is the age at 100% maturity.

Biomass quantities The spawning stock biomass SSB_t and its unexploited value at equilibrium SSB_0 are:

$$SSB_t = \varphi \cdot \sum_{a=1}^{A} N_{a,t} \cdot W_a \cdot Mat_a; \quad SSB_0 = \varphi \cdot \sum_{a=1}^{A} \tilde{N}_a \cdot W_a \cdot Mat_a,$$

where φ is the proportion of females in the stock.

Similarly, the vulnerable (or exploitable) biomass VB_t and its unexploited equilibrium value VB_0 are:

$$VB_t = \sum_{a=1}^{A} N_{a,t} \cdot W_a \cdot S_a; \quad VB_0 = \sum_{a=1}^{A} \tilde{N}_a \cdot W_a \cdot S_a.$$

Exploitation rate We assumed the catch data Y_t had no error, thus treating the exploitation rates H_t as derived quantities:

$$H_t = Y_t / VB_t.$$

Catch-at-age Based on the exploitation rate H_t, population abundance $N_{a,t}$, and age-dependent selectivity S_a, the model estimated catch-at-age $\hat{C}_{a,t}$ as:

$$\hat{C}_{a,t} = N_{a,t} \cdot S_a \cdot H_t.$$

A.1.2 Observation models

Abundance index We assumed that the abundance index I_t had a log-normal error:

$$I_t = q \cdot VB_t \cdot e^{\eta_t - 0.5 \cdot \tau^2}, \quad \text{where } \eta_t \overset{\text{i.i.d}}{\sim} N(0, \tau^2),$$

where q is the catchability coefficient, η_t is the normal observation error with mean 0 and variance τ^2, and $-0.5 \cdot \tau^2$ is the bias correction term.

Length composition In the simulation study, it was assumed that the length composition data were collected every five years (i.e., $t^* \in \{1967, 1972, 1977, ..., 2017\}$), which were modelled using a Dirichlet-multinomial (DM) distribution (Thorson et al. 2017):

$$\boldsymbol{n}_{t^*} \sim \text{DM}(E_{t^*}, \boldsymbol{\delta}_{t^*}, \hat{\boldsymbol{P}}_{t^*}),$$

where $\boldsymbol{n}_{t^*}$ is the vector of the number of fish in each length bin j at year t^* in the length composition data (i.e., $\boldsymbol{n}_{t^*} = [n_{1,t^*}, n_{2,t^*}, ..., n_{2,t^*}]'$), E_t is the sample size at year t, $\boldsymbol{\delta}_{t^*}$ is the vector of concentration parameters at year t^* (i.e., $\boldsymbol{\delta}_{t^*} = [\delta_{1,t^*}, \delta_{2,t^*}, ..., \delta_{2,t^*}]'$), and $\hat{\boldsymbol{P}}_{t^*}$ is the vector of model-estimated length-composition proportions at year t (i.e., $\hat{\boldsymbol{P}}_{t^*} = [\hat{P}_{1|t^*}, \hat{P}_{2|t^*}, ..., \hat{P}_{J|t^*}]'$). Then, the model-estimated length-composition proportion of length bin j at year t (i.e., $P_{j|t^*}$) was calculated as:

$$\hat{P}_{j|t^*} = \sum_a \hat{P}_{a|t^*} \cdot G_{i|a},$$

where the model-estimated age-composition proportions $\hat{P}_{a|t^*}$ were derived by normalising the model estimated catch-at-age $\hat{C}_{a,t^*}$:

$$\hat{P}_{a|t^*} = \frac{\hat{C}_{a,t^*}}{\hat{C}_{t^*}}; \quad \hat{C}_{t^*} = \sum_{a=1}^{A} \hat{C}_{a,t^*}.$$

To reduce the number of parameters to be estimated, the concentration parameters δ_{i,t^*} were assumed to be proportional to the sample size E_{t^*} and the model-estimated length-composition proportions $\hat{P}_{i|t^*}$ (Thorson et al. 2017):

$$\delta_{j,t^*} = \theta \cdot E_{t^*} \cdot \hat{P}_{j|t^*}.$$

The effective sample size $E_{t^*}^{\text{eff}}$ is then defined as

$$E_{t^*}^{\text{eff}} = \frac{1 + \theta \cdot E_{t^*}}{1 + \theta} = \frac{1}{1 + \theta} + E_{t^*} \cdot \frac{\theta}{1 + \theta}.$$

A.1.3 Prior distribution

Correlation was incorporated between the two parameters h and M by assuming their transformed forms followed a bivariate normal distribution. Because the steepness h is bounded by 0.2 and 1, h was transformed to x using the logit transformation:

$$x = \text{logit}\left(\frac{h - 0.2}{0.8}\right).$$

Because a log-normal marginal prior distribution was assumed for M, M was log-transformed to $\log(M)$. Thus, the joint prior distribution of x and $\log(M)$ was given by:

$$\begin{bmatrix} x \\ \log(M) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_{\log(M)} \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho \cdot \sigma_x \cdot \sigma_{\log(M)} \\ \rho \cdot \sigma_x \cdot \sigma_{\log(M)} & \sigma_{\log(M)}^2 \end{bmatrix}\right),$$

where μ_x and $\mu_{\log(M)}$ are the mean of the joint prior distribution of x and $\log(M)$, respectively, σ_x^2 and $\sigma_{\log(M)}^2$ are the variance of the joint prior distribution of x and $\log(M)$,

respectively, and ρ is the correlation between x and $\log(M)$ (see Table A-2 for prior distributions imposed on the other parameters (i.e., R_0, τ^2, q, θ).

A.2 Estimation

We estimated the model in *Stan* (Carpenter et al. 2017). The Bayesian model requires the specification of the full likelihood and the prior probability distributions of the parameters. The complete joint likelihood (vectors indicated in bold font) $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{R}|\boldsymbol{I}, \boldsymbol{n}; \boldsymbol{Y})$, including the parameters (i.e., $\boldsymbol{\Theta} = [\tau^2, h, M, R_0, q, \theta]'$) and recruitment (i.e., $\boldsymbol{R} = [R_1, R_2, ..., R_T]'$) as latent variables, is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{R}|\boldsymbol{I}, \boldsymbol{n}; \boldsymbol{Y}) =& \pi(\tau^2, h, M, R_0, q, \theta) \\ &\times \prod_{t=1967}^{2019} f_R(R_t|h, R_0, M; Y_t) \\ &\times \prod_{t=1967}^{2019} f_I(I_t|M, q, R_0, h, \tau^2; Y_t) \\ &\times \prod_{t^*=1967}^{2017} f_{LF}(\boldsymbol{n}_{t^*}|\theta, M, R_0, h; Y_{t^*}),\end{aligned}$$

where $\boldsymbol{I}$, $\boldsymbol{n}$, and $\boldsymbol{Y}$ denote the abundance index, length-frequency data, and the catch, respectively, $\pi(\tau^2, h, M, R_0, q, \theta)$ is the full joint prior distribution for the model parameters, which was derived by the product of the joint prior distribution of x and $\log(M)$ and the priors of the other parameters, $f_R(R_t|h, R_0, M; Y_t)$ is the likelihood of the recruitment, $f_I(I_t|M, q, R_0, h, \tau^2; Y_t)$ is the likelihood of the abundance index data, and $f_{LF}(\boldsymbol{n}_{t^*}|\theta, M, R_0, h; Y_{t^*})$ is the likelihood of the length-frequency data.

A.3 Simulation-estimation experiments

To simulate an albacore-like stock, we obtained most of input parameter values from a previous study (Punt et al. 1995) (see Table A-3). Some of the input values, such as θ, E_t, λ_1, and λ_2, were chosen arbitrarily for simulation purposes.

The main parameters of interest were steepness h and natural mortality M. For this reason, the other parameters R_0, q, τ, and θ were estimated with the same prior distributions across all of simulation-estimation experiments. For the two focal parameters of interest, we considered two scenarios: (1) the medians of the priors were set to the true values, and (2) the medians of the priors were set to the slightly biased values.

Table A-3: Input parameter values used in the age-structured model for simulation. Values were either from a previous study by Punt et al. (1995), otherwise were chosen arbitrarily.

Parameter	Quantity (unit)	Reference
A	12 (years)	Punt et al. 1995
a_{50}	3.5 (years)	Punt et al. 1995
ν	0.2	Punt et al. 1995
a_{Mat}	5 (years)	Punt et al. 1995
L_∞	124.74 (cm)	Punt et al. 1995
κ	0.228 (year^{-1})	Punt et al. 1995
a_0	-0.989 (years)	Punt et al. 1995
ω_1	$1.3718 \cdot 10^{-5}$ (kg)	Punt et al. 1995
ω_2	3.0973	Punt et al. 1995
σ_R	0.385	Punt et al. 1995
φ	0.5	
θ	0.01	
E_t	$10^4 \quad \forall t$	
λ_1	2	
λ_2	0.3	

APPENDIX B SIMULATION EXPERIMENTS: ADDITIONAL TABLES AND FIGURES

Table B-1: Probability that the estimated harvest rate H exceeds the simulated value H_{Sim}, for three scenarios on the prior probability distribution (lower and upper range: 0.025 and 0.975 quantiles of harvest rate estimates, and P). Scenarios were: an accurate and precise prior centred on the true value with a low coefficient of variation (CV) (10%), an accurate but imprecise prior centred on the true value with a high CV (30%), and an inaccurate but precise prior that was biased high by 0.05 for both production parameters relative to the true value, with a tight prior (CV 10%).

		High acc. & high prec.			High acc. & low prec.			Low acc. & high prec.		
Approach	Est. err.	Low	High	P($H > H_{Sim}$)	Low	High	P($H > H_{Sim}$)	Low	High	P($H > H_{Sim}$)
Estimated (single model)	Yes	0.32	0.56	0.34	0.33	0.57	0.34	0.30	0.53	0.22
Estimated (single model)	No	0.38	0.56	0.59	0.38	0.56	0.59	0.27	0.39	0.00
Grid (prior weights)	Yes	0.28	0.77	0.55	0.16	1.19	0.56	0.19	0.54	0.10
Grid (prior weights)	No	0.38	0.78	0.75	0.25	1.86	0.75	0.26	0.41	0.00
Grid (equal weights)	Yes	0.28	0.79	0.55	0.16	1.23	0.53	0.19	0.55	0.13
Grid (equal weights)	No	0.32	0.78	0.56	0.19	1.86	0.56	0.21	0.41	0.00
Monte Carlo ensemble	Yes	0.34	0.71	0.55	0.25	0.91	0.50	0.23	0.48	0.06
Monte Carlo ensemble	No	0.30	0.57	0.44	0.29	1.05	0.44	0.27	0.44	0.00

B.1 Accurate and precise prior

Plots from the simulations with an accurate and precise prior probability distribution centred on the true value with a low coefficient of variation (10%) (Figures B-1 to B-4).

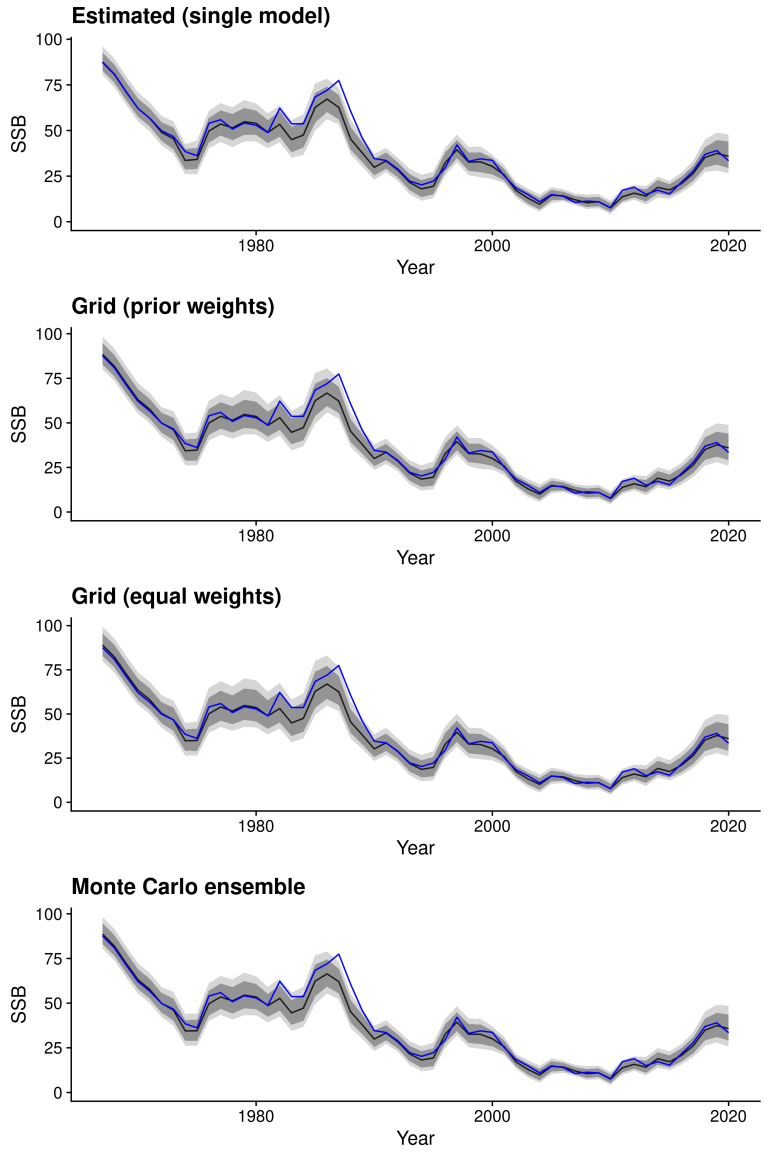


Figure B-1: Time series of spawning stock biomass (SSB) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which was centred on the true value with a low coefficient of variation (10%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

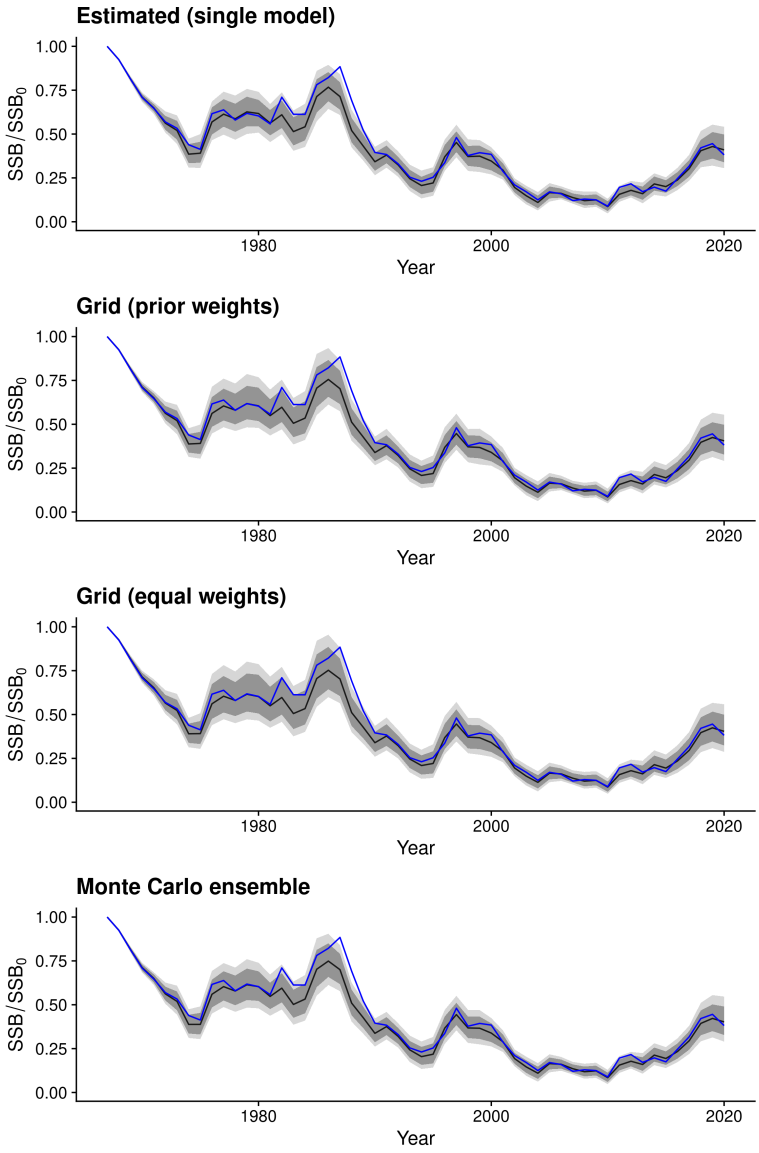


Figure B-2: Time series of stock status (spawning stock biomass SSB relative to unfished spawning biomass SSB_0) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a low coefficient of variation (10%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

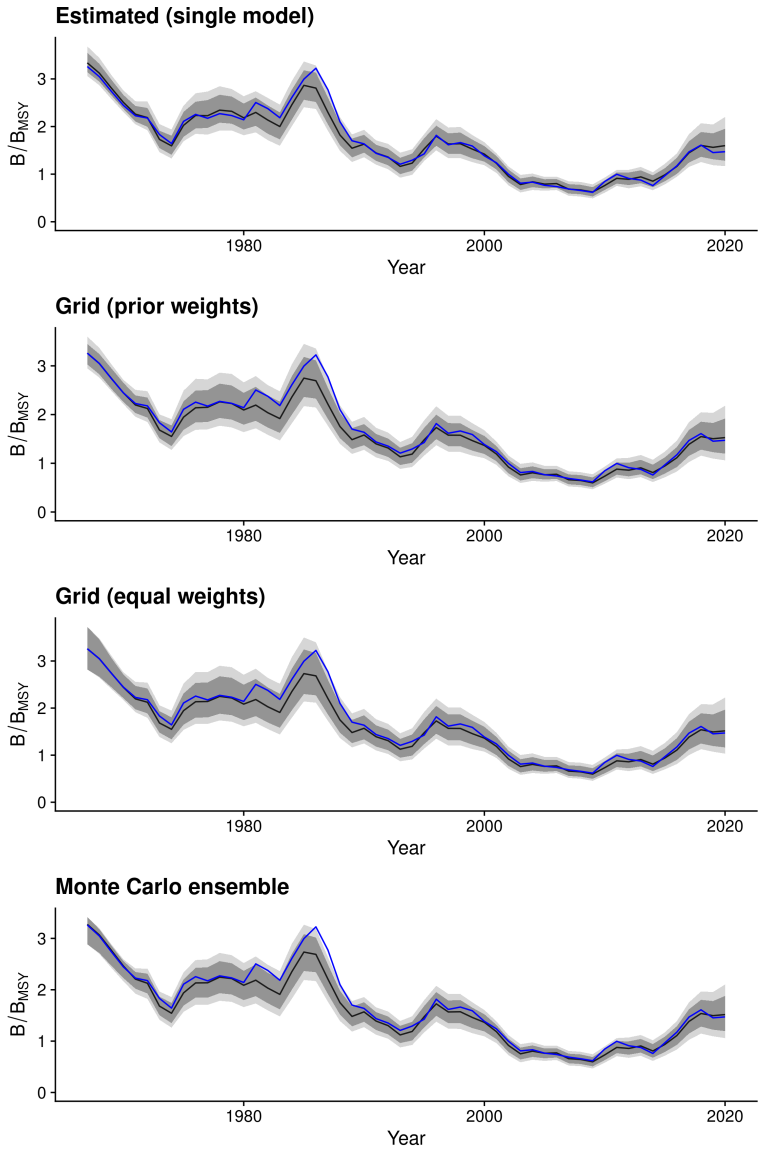


Figure B-3: Time series of stock status (biomass B relative to B that produces maximum sustainable yield (MSY)) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a low coefficient of variation (10%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

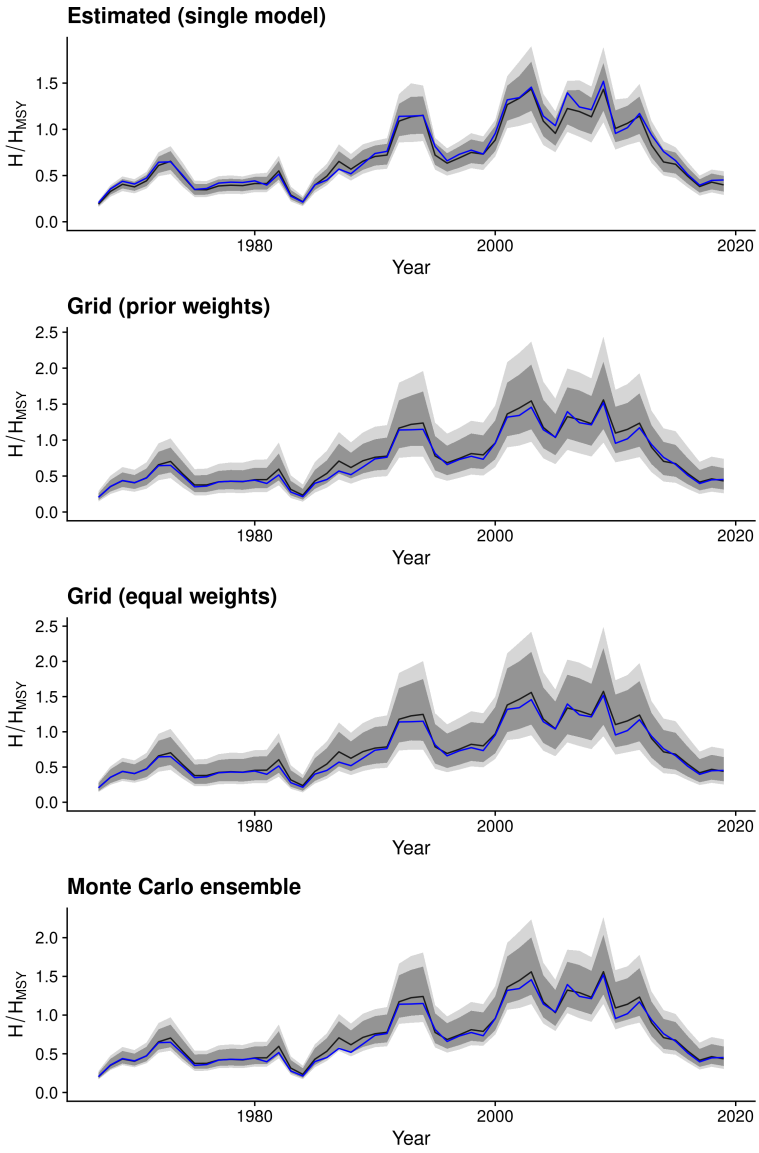


Figure B-4: Time series of harvest rate (H) (relative to H that produces maximum sustainable yield (MSY)) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a low coefficient of variation (10%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

B.2 Accurate but imprecise prior

Plots from the simulations with an accurate but imprecise prior probability distribution centred on the true value with a high coefficient of variation (30%) (Figures B-5 to B-8).

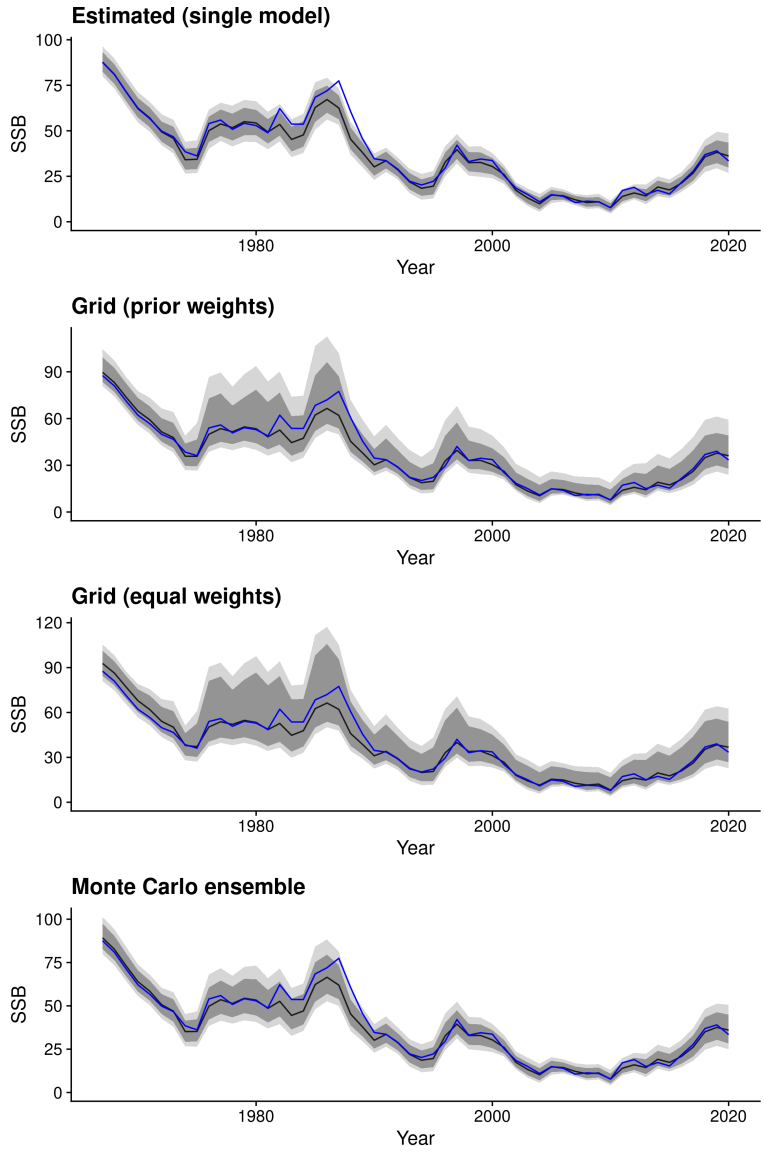


Figure B-5: Time series of spawning stock biomass (SSB) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a high coefficient of variation (30%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). T he true simulated time series is shown by the dark blue line.

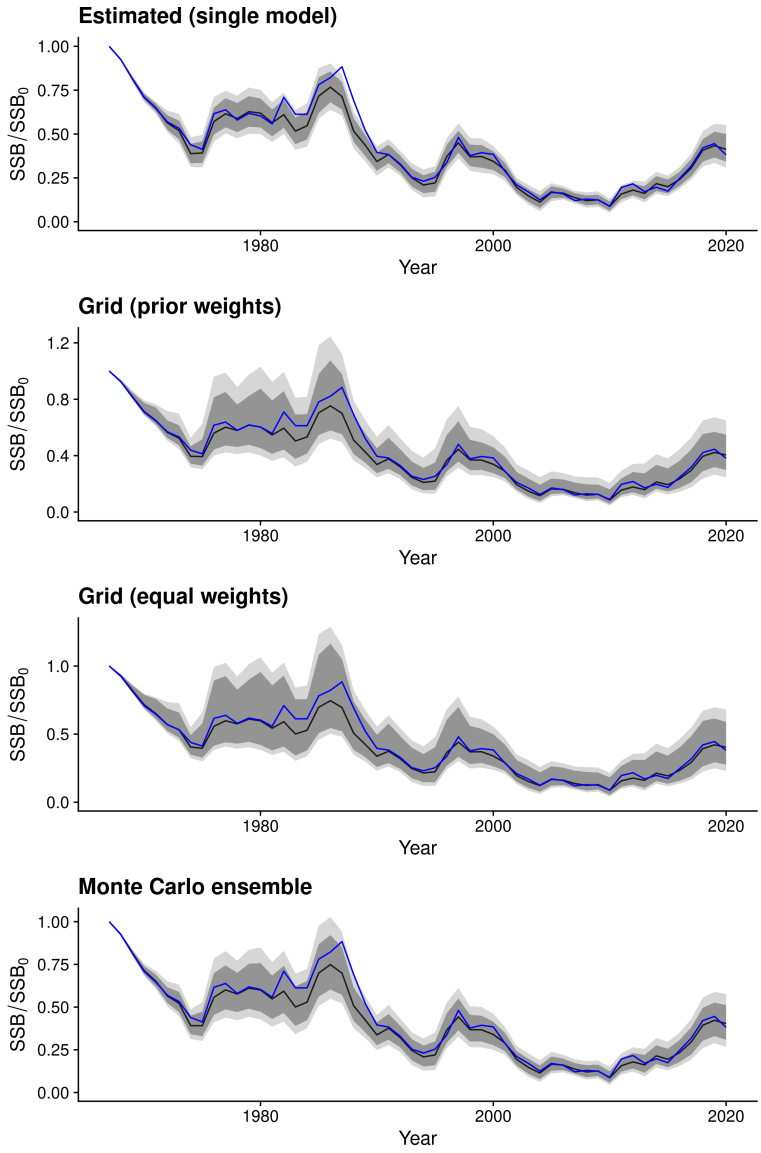


Figure B-6: Time series of stock status (spawning stock biomass SSB relative to unfished spawning biomass SSB_0) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a high coefficient of variation (30%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

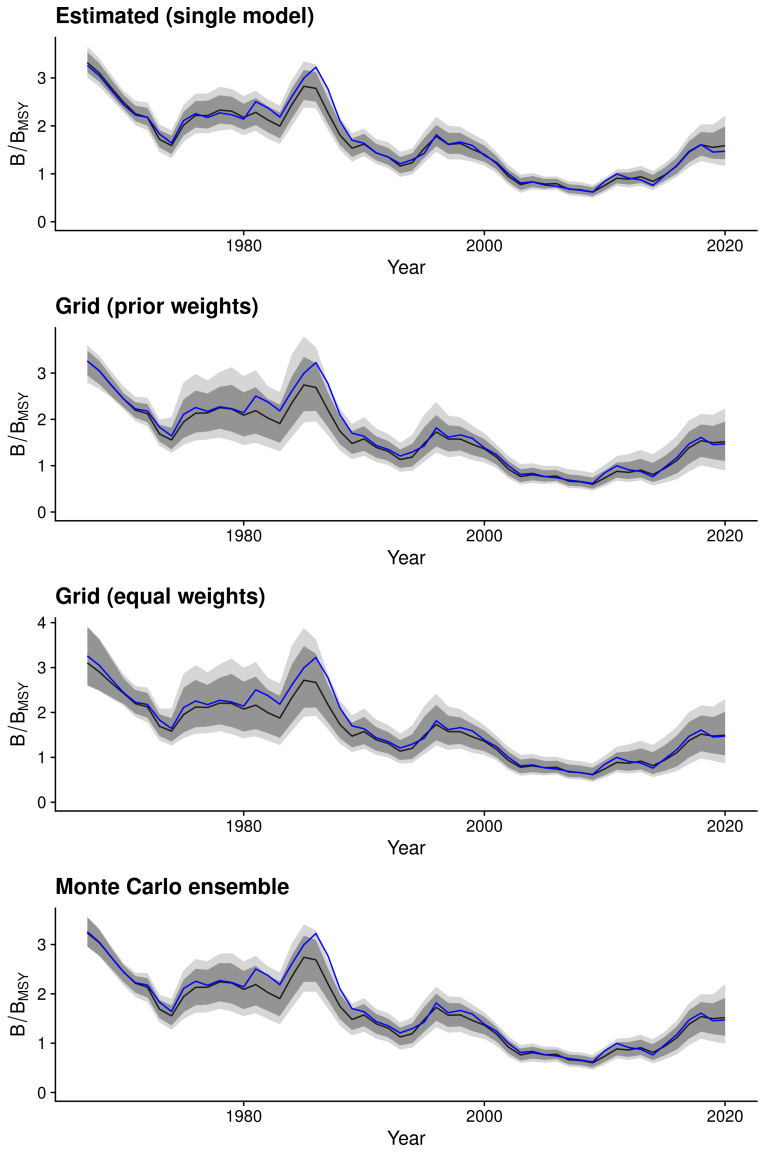


Figure B-7: Time series of stock status (biomass B relative to B that produces maximum sustainable yield (MSY)) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a high coefficient of variation (30%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

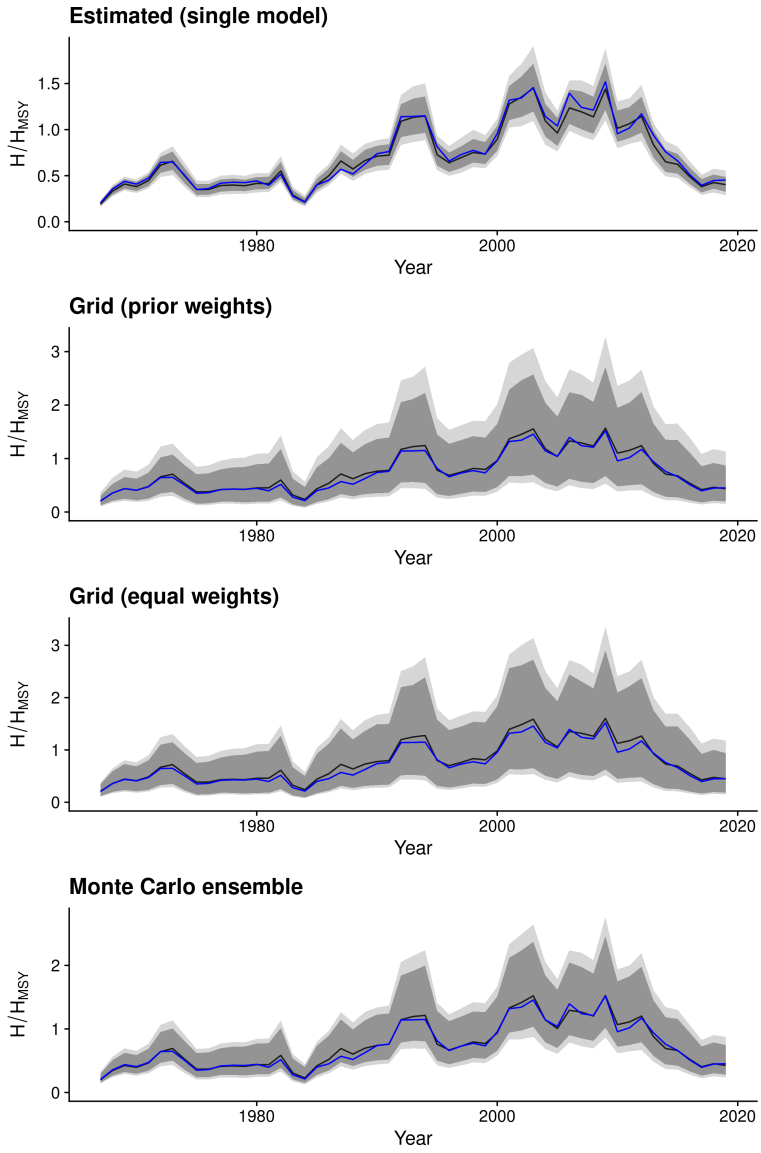


Figure B-8: Time series of harvest rate (H) (relative to H that produces maximum sustainable yield (MSY)) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a high coefficient of variation (30%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

B.3 Inaccurate but precise prior

Plots from the simulations with an inaccurate but precise prior probability distribution biased high by 0.05 for both production parameters relative to the true value, with a tight prior (Coefficient of variation 10%) (Figures B-9 to B-12).

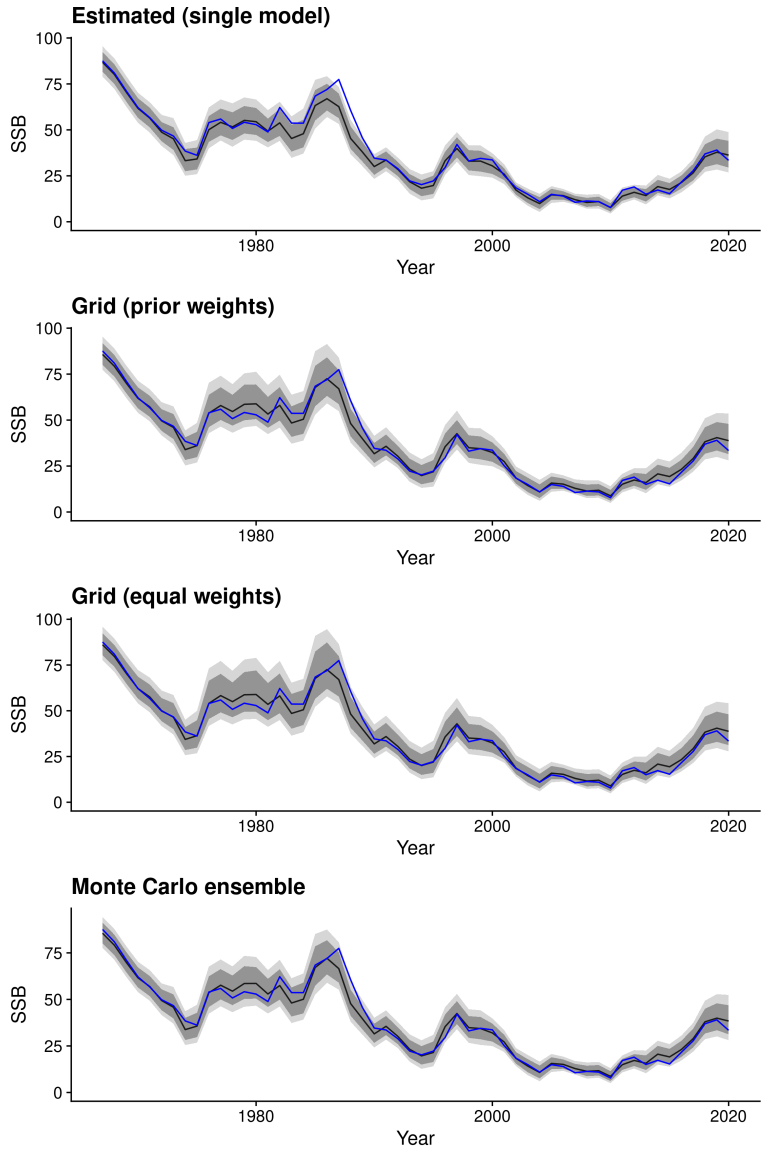


Figure B-9: Time series of spawning stock biomass (SSB) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is centred on the true value with a low coefficient of variation (10%). Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

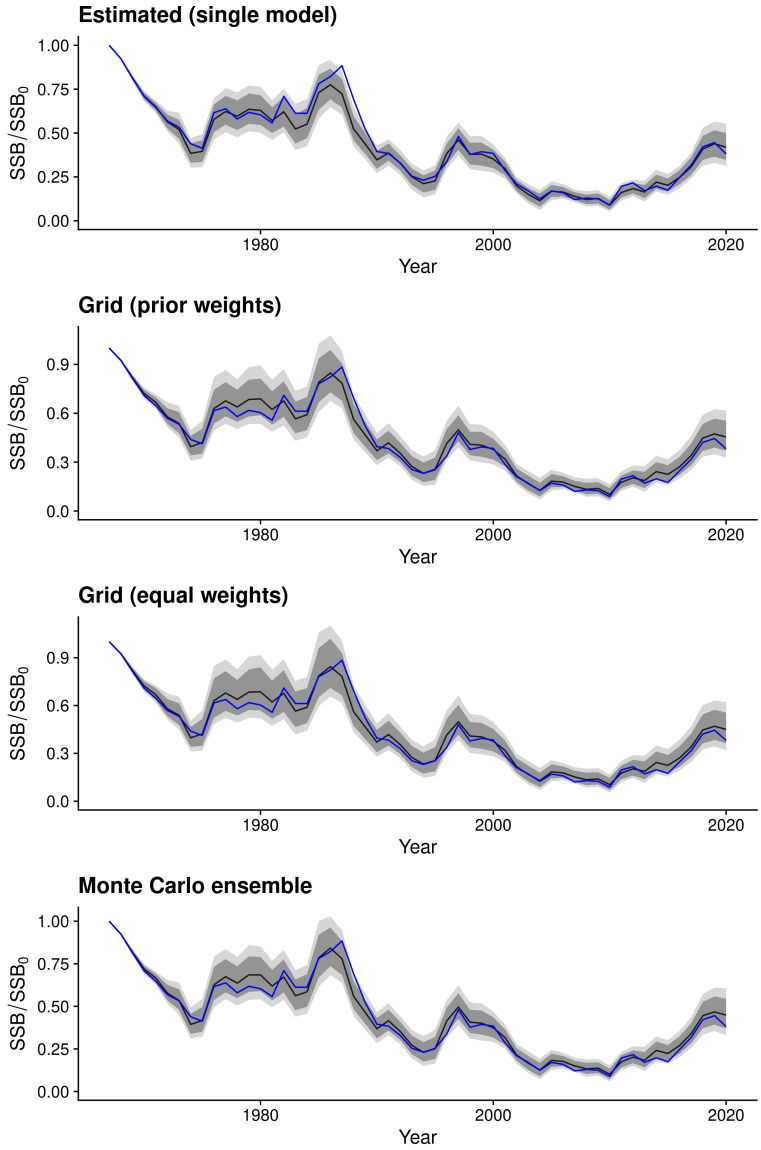


Figure B-10: Time series of stock status (spawning stock biomass SSB relative to unfished spawning biomass SSB_0) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which was biased high by 0.05 for both parameters relative to the true value, with a tight prior (coefficient of variation 10%) representing seemingly good understanding of productivity. Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

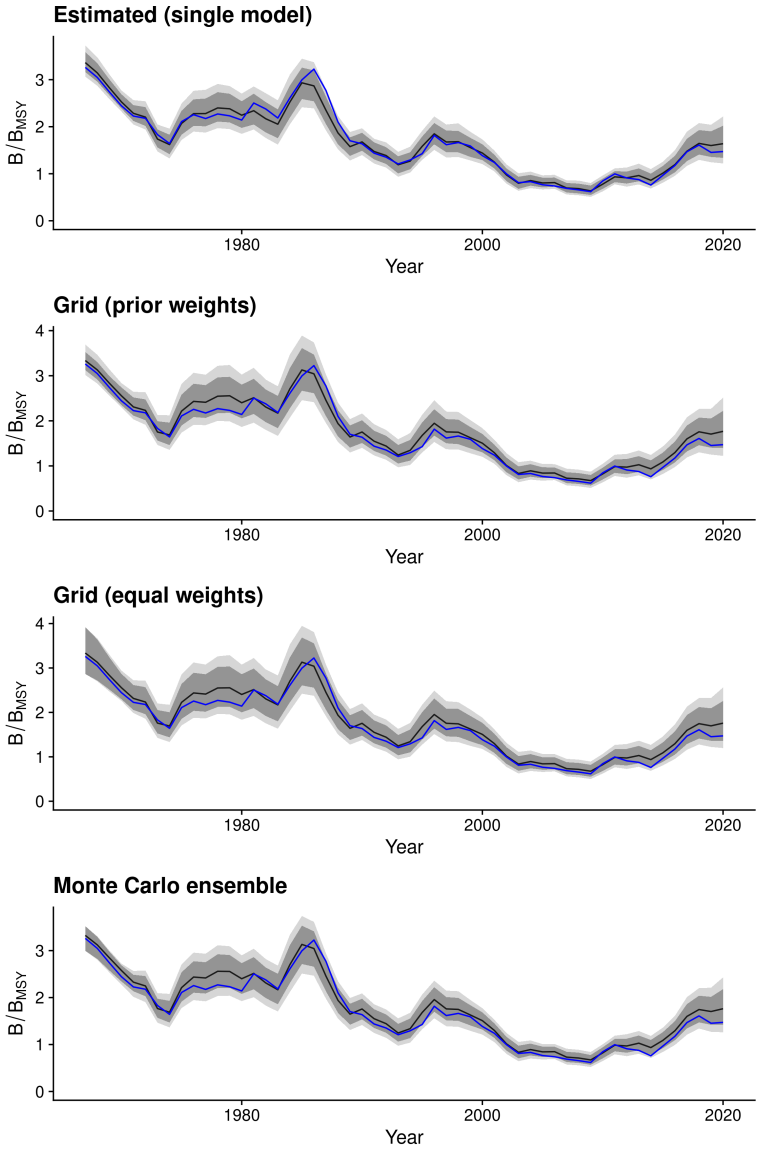


Figure B-11: Time series of stock status (biomass B relative to B that produces maximum sustainable yield (MSY)) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is biased high by 0.05 for both parameters relative to the true value, with a tight prior (coefficient of variation 10%) representing seemingly good understanding of productivity. Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

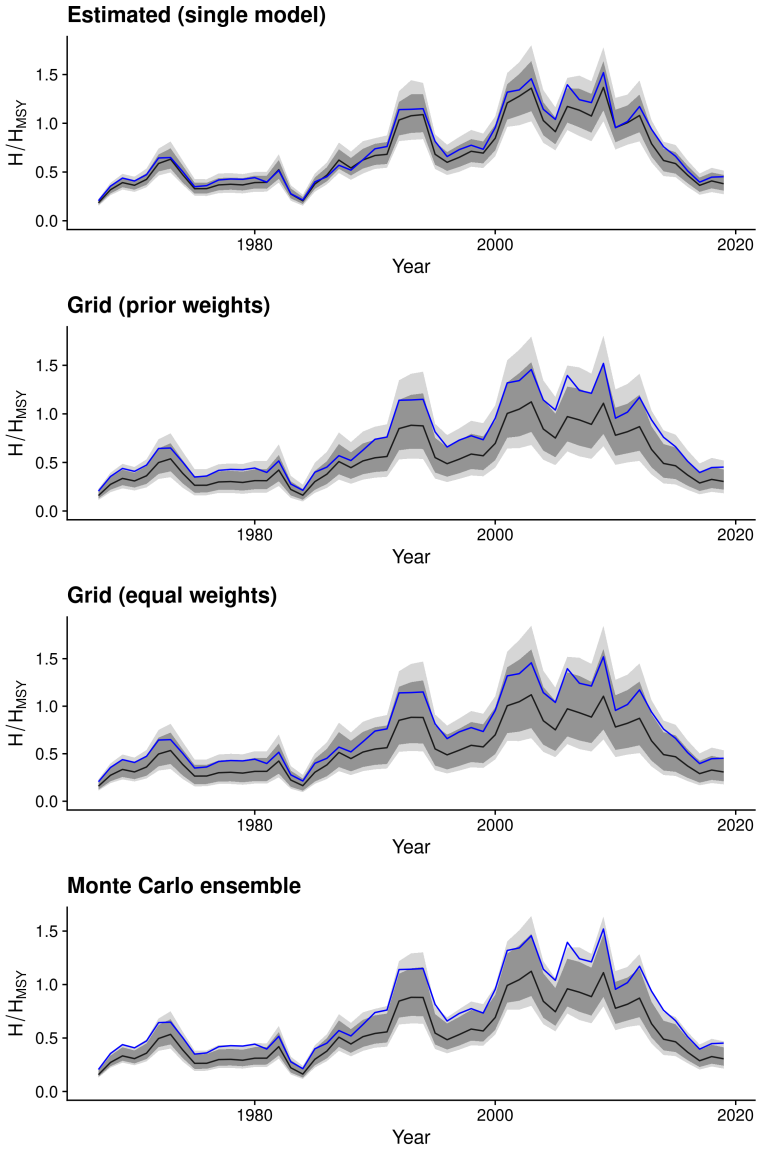


Figure B-12: Time series of harvest rate (H) (relative to H that produces maximum sustainable yield (MSY)) for four assessment approaches using estimated productivity (filled dark blue), bootstrap Monte Carlo (MC) draws from a productivity prior (light blue), and prior-weighted and equal-weight factorial uncertainty grids over the productivity prior. All methods used the same productivity prior, which is biased high by 0.05 for both parameters relative to the true value, with a tight prior (coefficient of variation 10%) representing seemingly good understanding of productivity. Solid lines indicate the median across ensembles (or the posterior median for the estimated case). The true simulated time series is shown by the dark blue line.

APPENDIX C A BAYESIAN VIEW ON MODEL ENSEMBLES AND HYPOTHESIS TREES

A loose theoretical justification for a hypothesis tree approach to ensemble construction can be provided from a Bayesian hierarchical breakdown of the challenge of model averaging (Draper 1995, Hoeting et al. 1999, 4).

For any prediction y, derived over a set of models $M = \{S, \theta\}$ from data X, the posterior predictive probability distribution for y can be split into nested integrals over model structures S and model parameters θ, with predictions y at any θ and S ($p(y|\theta, S, X)$), weighted by the joint posterior probability of the model structure and corresponding parameters ($p(S, theta|X)$). In full the distribution is:

$$p(y|X, M) = \int_M p(y|M, X)p(M|X)dM = \int_S \int_\theta p(y|\theta, S, X)p(S, theta|X)d\theta dS.$$

The right-hand side of the distribution can be further split to provide a hierarchical structure (Draper 1995), which provides an approach for model ensembles as an approximation to the challenge:

$$\int_S \int_\theta p(y|\theta, S, X)p(S, theta|X)d\theta dS = \int_S \int_\theta p(y|\theta, S, X)p(theta|S, X)p(S|X)d\theta dS \tag{C-1}$$

$$= c \int_S \int_\theta p(y|\theta, S, X)p(X|theta, S)p(\theta|S)p(S)d\theta dS \tag{C-2}$$

This composition can be written in hierarchical (or generative) notation, such that:

$$S \sim p(S), \tag{C-3}$$

$$\theta|S \sim p(\theta|S), \tag{C-4}$$

$$X|\theta, S \sim p(X|\theta, S), \tag{C-5}$$

$$y|X, \theta, S \sim p(y|X, \theta, S). \tag{C-6}$$

This notation suggests a generative split into:

1. A prior (decisions) about plausible model structures; i.e., a (set of) hypothesis (hypotheses) about plausible model structures and their associated weights. In practice, many plausible structures have weight zero because a limited number of structures can be feasibly explored. Nevertheless, this limitation can be explicitly acknowledged.
2. Given the model structure, a prior probability distribution over the associated parameters is developed. This development may be the same across all structures, or it may differ depending on the likelihood formulation in the different structures.

3. The data are viewed as generated from a distribution (likelihood) given the structure and parameters.

4. Predictions are made from a distribution given the structure, parameters, and data.

The first step will, in practice, usually under-value uncertainty because not all plausible model structures are evaluated, and most structures are given a weight of zero *a priori*. For a discrete set of models, the probability (or weight) for model i can be written as:

$$p(S_i|X_i) \propto p(X_i|S_i)p(S_i) = p(S_i) \int_{\theta_i} p(X|\theta_i, S_i)p(\theta_i|S_i)d\theta_i,$$

which shows that, in this context, the likelihood $p(X|\theta_i, S_i)$ is the appropriate model weighting tool. Nevertheless, in practice, the likelihood surface may be flat with respect to θ_i and S_i. In this situation, a prior-weighted ensemble (i.e., assuming a flat distribution for $p(X|\theta, S)$ in the hierarchical decomposition) would result in the same outcome as a full model. This aspect can only be tested by attempting to fit certain parameters within the model.

The full likelihood weighting also does not apply where likelihoods are not comparable across structures (i.e., when the data are specific to a given structure). The Bayesian breakdown, however, provides a useful justification for a hypothesis tree approach that attempts to, in principle, cover the steps required for a full integration over model and parameter uncertainty, even if model weighting is often a difficulty that cannot currently be solved in a theoretically justifiable manner.