SCIENTIFIC COMMITTEE
EIGHTEENTH REGULAR SESSION

ELECTRONIC MEETING
10 – 18 August 2022

Report on WCPFC project 107b: Improved stock assessment and structural uncertainty grid for Southwest Pacific blue shark

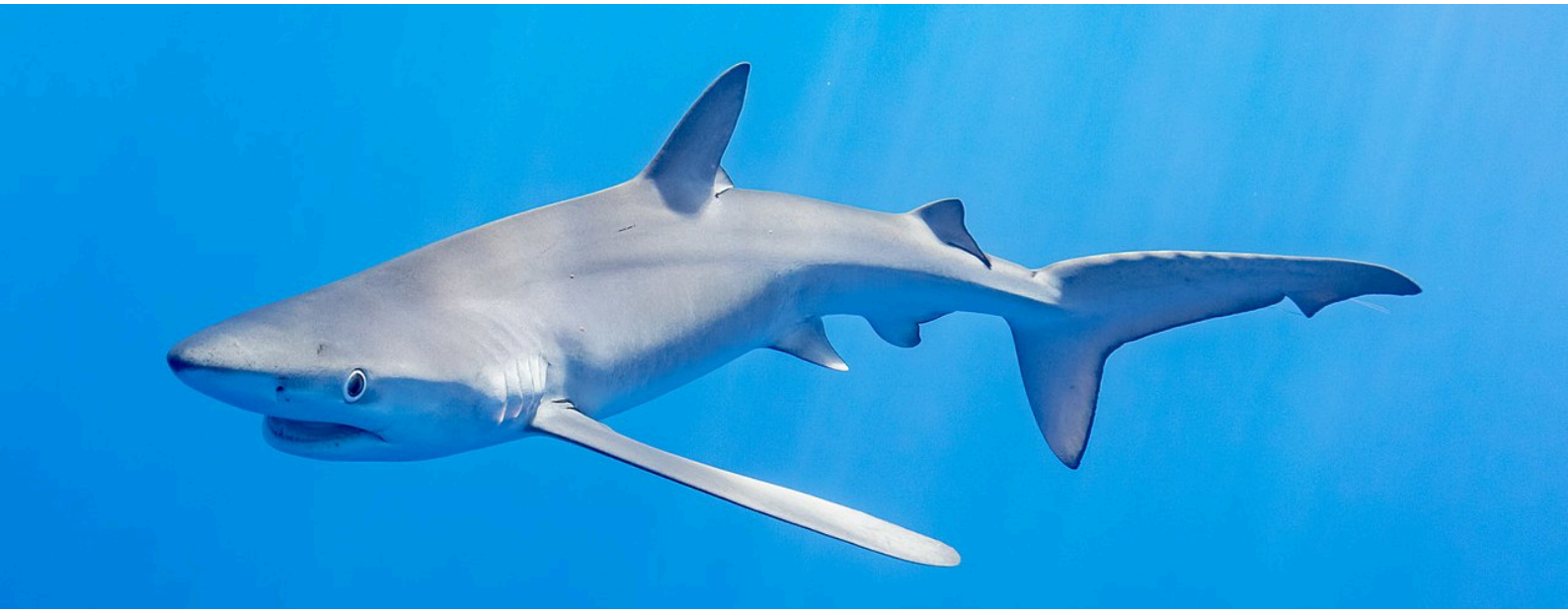
WCPFC-SC18-2022/SA-WP-03 (Rev.01)
8 August 2022

Philipp Neubauer[1], Felipe Carvalho[2], Nicholas Ducharme-Barth[2], Kath Large[1], Stephen Brouwer[3], Jemery Day[4] and Paul Hamer[4]

[1] Dragonfly Data Science
[2] NOAA Pacific Islands Fisheries Science Center
[3] Saggitus Consulting
[4] SPC Oceanic Fisheries Programme

Revision 1:

Corrected typo in the third bullet point in the assessment conclusions for the % of grid runs with recent spawning biomass above SB_{MSY}, corrected from 97% to 87%.



Improved stock assessment and structural uncertainty grid for Southwest Pacific blue shark

Authors:
Philipp Neubauer
Felipe Carvalho
Nicholas Ducharme-Bath
Kath Large
Stephen Brouwer
Jemery Day
Paul Hamer

DRAGONFLY
Data Science

SAGGITUS

Cover Notes

To be cited as:

Neubauer, Philipp; Carvalho, Felipe; Ducharme-Bath, Nicholas; Large, Kath; Brouwer, Stephen; Day, Jemery; Hamer, Paul (2022). Improved stock assessment and structural uncertainty grid for Southwest Pacific blue shark, 70 pages. WCPFC-SC18-2022/SA-WP-03. Report to the WCPFC Scientific Committee. Eighteenth Regular Session, 10–18 August 2022.

CONTENTS

EXECUTIVE SUMMARY

This analysis presents additional work in order to constrain the model grid employed for the 2021 south Pacific blue shark (BSH) stock assessment in the Western and Central Pacific Ocean (WCPO). The 2021 stock assessment for BSH was accepted by the Scientific Committee at SC17. However, due to a number of uncertainties about the relative merit (model fit, plausibility) of individual models within the large (3888) model grid, SC17 was hesitant about using such a large grid to provide management advice. The SC17 recommended improving the manner in which the grid was selected before approving the results for providing management advice.

The present analysis attempted to address these concerns by running a number of standard diagnostics across all grid model runs to ensure that:

1. models had sufficiently converged and results were robust to jittering of starting values;
2. models were consistent and did not show large retrospective patterns (as evidenced by Mohn's ρ); and
3. models had reasonable predictive skill.

Acknowledging that the stock was not unfished at the start of the assessment, and that references to unfished biomass may be misleading, as B_0 is likely poorly estimated; we also explored the results using an alternative reference point, namely $SB/SB_{F=0}$.

Our initial investigation of models in the 2021 assessment grid found that all models appeared to have converged to global solutions with small gradients for all estimated parameters across all models, all models had positive definite Hessian solutions, and jittering did not lead any models to find alternative optima, likely due to the low number of estimated parameters. Retrospective analyses of the 2021 grid showed that only a small number of models had large retrospective patterns, but these models were not consistently associated with a particular uncertainty axis. The majority of models had Mohn's ρ values near zero. Filtering by these diagnostics did not significantly reduce the spread of outcomes from the initial 2021 model grid.

Given the lack of reduction of over-all uncertainty in the model grid, we further addressed model assumptions and inputs that were found to drive the spread in uncertainties, namely CPUE and natural mortality (M). Assuming low M, for example, accounted for most high estimates of $SB/SB_{F=0}$. Alternative CPUE assumptions had high impact, largely driven by inconsistent trends in early CPUE, and differences in recovery rates in recent CPUE among alternative indices. The 2021 stock assessment grid ignored process error, thereby placing high weight on CPUE indices (i.e., assuming high signal and low uncertainty). As a result, differences in indices were accentuated in grid runs that re-weighted or used alternative CPUE indices.

Two important decisions lead to a strong reduction in both the number of assessment models in the grid, as well as the spread of uncertainty in the outcomes. First, estimating M with an informative prior meant that one structural uncertainty axis could be dropped from the analysis. An additional two axes that contributed little to over-all outcome uncertainty were also dropped, resulting in a substantial reduction in the size of the initial (i.e., pre-diagnostic) grid. In addition, we included allowance for process error in CPUE, which may be large given unknown reporting trends for sharks. Acknowledging this process error in the models leads to less extreme trends, for both the diagnostic assessment scenario as well as the new model grid.

Estimating M also allowed for a closer inspection of the relationship between growth and M. In the previous formulation of the grid, having both growth and M as fixed values allowed for biologically inappropriate combinations of fast growth and low M. Estimating M alleviated this to a certain extent however it identified that in order for the fast growth hypothesis to fit the existing data, M needed to be implausibly large. As a result, the grid was further reduced by excluding the fast growth scenario.

Lastly, we followed recent analyses that have attempted to use various metrics to weight models in the uncertainty grid. We propose an iterative procedure that first excludes models that fail diagnostic criteria. We then weighted input axes for remaining models according to prior probabilities derived from either input analyses or analyst assessments of the relative utility of different inputs (e.g., CPUE time series). This *a priori* weighting can then be supplemented with *a posteriori* weighting for model fit or predictive skill.

We investigated a range of possible *a posteriori* weighting measures for the model grid, namely inverse variance weighting, MASE weights and stacking weights. Using predictive skill in the form of the MASE criterion did not reduce the outcome space significantly. We suggest that MASE is largely a measure of the degree to which a stock is production driven relative to being recruitment ("regime") driven. The MASE criterion will likely select for production-driven, over recruitment driven models, which may or may not be desirable. We show that stacking weights, weighting the model ensemble directly to maximise model predictive skill, does not appear to share this property. Over-all none of these model-weighting approaches appeared to lead to substantial changes in the range of outcomes from the reduced uncertainty grid. We suggest that more research is required on the topic of model ensemble weighting, and we therefore formulate our recommendations on the basis of prior (input axis) weighting only.

Taken together, these analyses restrict the number of candidate models from 3888 in the 2021 uncertainty grid, to 228 models in the revised uncertainty grid, and lead to lower uncertainty compared with the 2021 model grid. Nevertheless, the over-all model conclusions and recommendations from the 2021 blue shark assessment remain valid. Substantial uncertainties about inputs and biological parameters remain. Our analyses underscore that for low- to medium information stocks, such as most sharks, uncertainties in model outcomes are not necessarily reducible in the short-term. Only improved biological data collection and recording of interactions with bycatch species will lead to improved precision in stock assessment. Nevertheless, we suggest that consistency in estimated recent recovery trends, as well as robustness of these trends to alternative model assumptions provide evidence for effectiveness of recent non-retention measures for sharks, and BSH in particular.

Although the sensitivity analysis highlighted a number of uncertainties, we found a number of consistent patterns in the outcomes. Based on these consistent trends, and using a restricted, weighted set of 228 uncertainty grid runs, we **conclude** that:

- The most influential axis within the reduced uncertainty grid was the initial F assumption.
- The stock biomass was low throughout the region through the early 2000s following the expansion of longline fishing effort in the region. But the estimates across the uncertainty grid of 228 models largely indicated that the stock has been recovering since then.
- All 228 model runs indicate that fishing mortality at the end of the assessment period was below F_{MSY} and 87% of (weighted) model runs show that the biomass is above SB_{MSY} (median $SB_{recent}/SB_{\text{MSY}}$ = 1.64 (90^{th} percentiles 0.88 and 1.87; Table 6), with the median estimated depletion $SB_{recent}/SB_{F=0}$ = 0.71 (90^{th} percentiles 0.37 and 0.82), and

SB_{recent}/SB_0 = 0.80 (90^{th} percentiles 0.43 and 0.90).

- Fishing mortality has declined over the last decade and is currently relatively low with the median $F_{recent}/F_{\text{MSY}}$ = 0.65 (90^{th} percentiles 0.43 and 0.86; Table 6). This may be a result of most sharks being released upon capture from by most longline fleets.
- Finally, considered against all conventional reference points the stock on average does not appear to be overfished and overfishing is not occurring.

Given some of the uncertainties highlighted above, we recommend that SC18 consider:

- Providing more time, either as inter-session projects, or by extending time-frames for shark analyses. This will allow more thorough investigation of input data quality and trends, which shape assessment choices. In addition, it would allow input analyses to be completed in time to be presented to the pre-assessment workshop prior to the stock assessment. In addition, allowing more time for the assessments themselves will allow a more thorough investigation of alternative model structures, which may include comparisons with low-information methods such as spatial risk assessments.
- Increased effort to re-construct catch histories for sharks (and other bycatch species) from a range of sources. Our catch reconstruction models showed that model assumptions and formulation can have important implications for reconstructed catches. Additional data sources, such as log-sheet reported captures from reliably reporting vessels, may be incorporated into integrated catch-reconstruction models to fill gaps in observer coverage.
- Additional tagging be carried out using satellite tags in a range of locations, especially known nursery grounds in South-East Australia and New Zealand, as well as high seas areas to the north and east of New Zealand, where catch-rates are high. Such tagging may help to resolve questions about the degree of natal homing and mixing of the stock.
- Tagging may also help to obtain better estimates of natural mortality, if carried out in sufficient numbers. This could be taken up as part of the WCPFC Shark Research Plan to assess the feasibility and scale of such an analysis.
- Additional growth studies from a range of locations could help build a better understanding of typical growth, as well as regional growth differences. Current growth data are conflicting, despite evidence that populations at locations of current tagging studies are likely connected or represent individuals from the same population.
- Genetic/genomic studies could be undertaken to augment the tagging work to help resolve these stock/sub-stock structure patterns. To support this work, a strategic tissue sampling program for sharks is recommended with samples to be stored and curated in the Pacific Marine Specimen Bank.

1. INTRODUCTION

Southwest Pacific Blue shark (*Prionace glauca*) in the Western and Central Pacific Fisheries Commission Convention Area (WCPFC-CA) were assessed in 2021 (Neubauer et al. 2021a). That analysis was based on CPUE, reconstructed catch and length frequency analyses presented in Neubauer et al. (2021b). The assessment, as well as preceeding input analyses, followed from suggestions in Brouwer and Hamer (2020), stating that conditional on an appropriate catch reconstruction, an integrated assessment could be attempted.

The input data anlyses and stock assessment showed a number of uncertainties that were captured in the structural uncertainty grid of the assessment. Key uncertainties resulted from uncertain early fishing mortality (see Neubauer et al. (2021a) for discussion), poor logsheet reporting and therefore possibly non-representative CPUE, and biolgical assumptions such as natural mortality and stock structure. As a result of these uncertainties, the stock assessment produced a large grid of models (3888), which explored these uncertainties in a factorial design.

The large number of uncertainty axes in the 2021 Southwest Pacific blue shark stock assessment led to a large spread in stock status estimates prompting SC17 to request follow-up work to scrutinise models within the model grid, as well as to reconsider input assumptions and reference points, in an effort to constrain the number of models and the spread of outcomes of the model grid. Specifically, the present project included three objectives: i) a re-examination of the input data, ii) development of an objective criteria for evaluating the performance of the proposed models to be included in the final grid used for the management advice, and iii) evaluation of dynamic reference point for southwest Pacific blue shark.

A preliminary analysis of the 2021 model grid based on a range of diagnostics was presented to the Pacific Community (SPC) pre-assessment workshop (Hamer 2022). That analysis suggested that much of the uncertainty in the initial grid was irreducible based on diagnostics alone, suggesting that the inputs and uncertainty axes themsleves would need to be reviewed in order to provide a more parsimonious uncertainty grid than that presented in 2021. Following from the initial analysis of the 2021 grid, we therefore revised key model assumptions that, together, drove much of the spread in outcomes. These changes lead to a considerable reduction in the over-all structural uncertainty relative to 2021, and we recommend that stock status estimates (including uncertainty) and trends from the updated model ensemble be used for management of Southwest Pacific blue shark.

2. METHODS

2.1 Summary of key stock assessment assumptions

Based on inferences from southwest Pacific and global blue shark tagging data, the 2021 stock assessment for southwest Pacific blue shark was structured into three fleets with respect to trends in observed length frequencies by fleet, and corresponding trends in CPUE indices:

1. High latitude fleets catching juvenile and mature blue shark south of 35°South, mainly in New Zealand and the South Tasman Sea;
2. The EU-Spain fleet fishing at intermediate latitudes to the northeast and northwest of the New Zealand EEZ, capturing a broad size range from just mature to large individuals (>250 cm); and
3. Low latitude fleets, capturing largely mature fish, but with a notable absence of large individuals.

Catches were reconstructed between 1990 and 2020 using a spatial GLMM model (Neubauer et al. 2021a). Catch estimates were combined with a model for annual discard rates per flag, which was used to produce scenarios of total fishing-induced mortalities.

Diagnostic model runs were established on the basis of the CPUE series (Figure 1) that were found to be the most robust and representative:

1. The New Zealand logsheet CPUE was used to represent relative biomass trends in high latitudes;
2. EU-Spain CPUE based on reported catches in weight; and,
3. Japanese logsheet CPUE to represent distant water and low-latitude fisheries.

We note here that the diagnostic model in 2021 used estimation error from CPUE as error on the CPUE index, and indices were therefore initially weighted according to their relative error (Figure 2). The setup ignored likely process error, a consequence of attempts early in the assessment process to improve fits to CPUE. This assumption is revisited in section 2.3.

2.1.1 2021 structural uncertainty grid

To adequately represent major uncertainties in assessment inputs, the 2021 stock assessment incorporated nine axes of uncertainty and 3888 models. The grid considered:

1. **Catch scenarios:** (Table 1) - posterior mean catch (base) and 90^{th} percentile of the posterior distribution of predicted catches.
2. **Discard scenario:** (Table 1) - low (25^{th} percentile), mean (base) and high (75^{th} percentile) estimated discard rates.
3. **Initial F:** initial fishing mortality associated with equilibrium catch - assuming baseline F or high (50% higher) initial exploitation.
4. **High latitude CPUE:** using the New Zealand CPUE series with (base) or without pre-2004 years (i.e., removing years when logsheet and observer CPUE differ - *RM early New Zealand*), or down-weighting both New Zealand and EU-Spain index to 25% of their original weight in favour of low latitude/high seas indices.
5. **Low latitude CPUE:** replacing the Japanese index (base) with the Australian low-latitude index, removing the EU-Spain index.
6. **Recruitment deviation:** low ($\sigma_R = 0.2$; base), forcing smaller recruitment deviations in the model (i.e., the model acts more like an age-structured production model; ISC 2018), or allowing greater variation in recruitment ($\sigma_R = 0.4$).
7. **Natural mortality:** base (0.2) or low (0.16) M.
8. **Survival fraction/density dependent recruitment:** $S_{frac} = 0.391, \beta = 2$ vs. scenarios described in ISC 2018: $S_{frac} = 0.378, \beta = 1$ (low) or $S_{frac} = 0.467, \beta = 3$ (high). Higher β indicates increased over-compensation.
9. **Growth:** replacing Manning and Francis (2005) (base) with Joung et al. (2018) growth equations.

2.1.2 Reference points

Clarke and Hoyle (2014) and Zhou et al. (2018) evaluated methods to derive reference points for elasmobranchs in the Western and Central Pacific Ocean (WCPO). However, to date, there are no formally agreed reference points for sharks in the WCPO. Recent assessments of oceanic whitetip shark, for example, compared fishing mortality to F_{lim} as a tentative limit reference point for sharks, and to F_{crash}, the fishing mortality that would lead to extinction in the long-term. If one assumes a simple Schaefer surplus production model, then $F_{crash} = R_{max}$, the maximum population growth rate (intuitively, a population cannot be sustained if fishing removes more individuals than the population can maximally produce), and $F_{lim} = 0.75R_{max}$. Because the versions of these reference points as used in the present assessment were approximated from integrated stock assessment runs, we use a subscript AS to show that these are not derived from R_{max}, but from $F_{crash,AS}$, as calculated from the stock synthesis models.

For blue shark, which have higher productivity than many other shark species, we also applied alternative reference points used for target fisheries. These include MSY based reference points (i.e., F_{MSY} and SB_{MSY}), as well as spawning biomass relative to spawning biomass at average initial recruitment levels (SB_0) and under current recruitment and $F = 0$ ($SB_{F=0}$). We note that the last reference point is used in tuna and billfish assessments, acknowledging that long-term average (or unfished) recruitment levels are not necessarily relevant in a dynamic pelagic environment. In the following we focus on comparisons between SB_0 and $SB_{F=0}$ based reference points.

2.2 Diagnosing the 2021 model grid

Following from comments made by SC17 and subsequent terms of reference for Project 107b: "Additional work to provide scientific advice for southwest Pacific blue shark (*Prionace glauca*) based on the 2021 stock assessment", we revisited the 2021 structural uncertainty grid in light of the following criteria:

- **Model convergence and stability:** the analysis should assess the final gradient (it should be relatively small; <1e-4), and check that the Hessian matrix is definite. The jitter procedure is applied to verify the stability of the model to evaluate whether the model has likely converged to a global solution rather than a local minimum.
- **Model consistency:** Retrospective analysis can be used to check the consistency of model estimates, for example, the invariance in SB and F as the model is updated with new data in retrospect.
- **Prediction skill:** Hindcasting analysis could be done to evaluate the model prediction skill of the CPUE. When conducting hindcasting, a model is fitted to the first part of a time series and then projected over the period omitted in the original fit. Prediction skill can then be evaluated by comparing the predictions with the observations.

We note that additional information and context was provided by the 2022 CAPAM workshop "Model Diagnostics in Integrated Stock Assessments", held virtually by the Center for the Advancement of Stock Assessment Methodology between Jan 31^{th}–Feb 3^{rd} 2022.

Goodness-of-fit tests to evaluate whether patterns in the residuals of the CPUE and length-frequency distributions were normally distributed and/or had temporal trends were not used in a first instance to scrutinise the 2021 model grid. This is due to the poor temporal resolution

of length frequencies and conflicts in CPUE for some parts of the time-series, which lead standard diagnostics to provide limited information to scrutinise the models in relative terms. Nevertheless, the apparent conflicting signals in CPUE make a case for higher process error than previously assumed in the 2021 BSH models. We revisit this issue in the next section and the Discussion.

2.3 Revisiting the diagnostic model input assumptions

Initial scrutiny of the 2021 BSH structural uncertainty grid, as described in the previous section, did not substantially reduce the model grid based on the proposed diagnostics. However, the lack of contrast in the model diagnostics with respect to the model grid axes is likely explained by the stock assessment setup:

- **Model convergence and stability:** Previous blue shark models employed an iterative procedure to fix most model parameters, leaving only unfished recruitment and recruitment deviations to be estimated. It is perhaps not surprising that convergence and stability are achieved by models with few freely estimated parameters.
- **Model consistency:** Retrospective patterns relate to changes in estimated productivity as new data is added to the model. Since the 2021 BSH model grid fixed most production relevant parameters (M, stock recruit parameters, growth), the degree to which the shape of the estimated production function could change was also relatively constrained. Observing only small retrospective patterns is therefore not necessarily unexpected.
- **Prediction skill:** Conflicts in the degree of recent stock rebuilding as well as early CPUE among the indices included in 2021 BSH models, coupled with low assumed error for these indices, led to poor predictive performance of models for most indices in all models. However, this is related to fitting all indices in the same model, highlighting unaccounted for process error in the assessment model setup.

The last point here is particularly relevant: we previously argued that this acknowledgement of process error for various input datasets can be made post-hoc (i.e., in interpreting model fits) — if we acknowledge process and sampling error, then we may be willing to allow relatively non-satisfactory fits to some datasets, as long as there was sufficient consistency in the outcomes. However, this argument ignores that the assumed error in the indices will affect the relative weighting in the stock assessment. As a consequence, the New Zealand CPUE index, for example, had very high weight relative to other indices, due to low estimation error in the CPUE standardisation. When applying the procedure advocated by Francis (2011) to assume total error (estimation plus observation error) – fitting a LOESS smoother through the index and calculating the resulting CV in residuals – the New Zealand index has the highest total error by a factor of 1.5 relative to high latitude (JP) and EU time-series, which are more temporally consistent. Consequently, the previous 2021 diagnostic case and model grid may have over-weighted the New Zealand index, which accounts for a fleet that only accounts for a relatively small proportion of total fishing mortality.

Based on the above considerations, we applied two important changes to the diagnostic case model, and re-evaluated the structural uncertainty grid with respect to those changes. First, we made explicit allowance for process error in indices, as per Francis (2011). We estimated expected total error by fitting a LOESS smoother with a span of 0.5 to all indices, and calculated the resulting standard error (in log space). The assumed error for each index was then set to the maximum of either observation error or total error for each year (i.e., if estimation error

was higher than the assumed total error for any one year, than that estimation error was used instead of the total error estimate).

Second, we allowed M to be estimated with an informative prior based on previous assumptions in the structural uncertainty grid: we previously assumed a base-case value of 0.2 with a sensitivity of 0.16 as a low end of plausible values. To reflect this spread in our prior, we used a (truncated at zero) normal prior with mean 0.2 and SD 0.025. Letting this parameter be estimated had two important consequences:

1. It allowed us to remove a consequential parameter from the structural uncertainty grid, and
2. it allows more flexibility in estimated production, as well as allowing trade-offs with assumed growth.

Estimating M is often seen as a difficult exercise, but may be associated with smaller assessment error (Punt et al. 2021). In addition, despite some inconsistencies in CPUE, the stock appears to react to reductions in catch with increases in all CPUE indices, suggesting that there may be sufficient contrast in the time-series to estimate natural mortality.

2.4 Structural uncertainty grid for revised diagnostic model

We sought to revise the 2021 structural uncertainty grid to reduce the number of grid axes and therefore the number of models. Relative to the previous structural uncertainty grid we note that:

- There was very limited differences between models run with different priors for Σ_R: most model setups appeared to produce relatively strong recruitment trends, with realised Σ_R usually higher than its prior. Given that this axis did not drive outcomes, it was omitted in the revised (2022) structural uncertainty grid.
- The natural mortality axis was dropped as M was estimated for all models.
- Only high initial F was included as an alternative to assumed initial F values. The previous assumption of low initial F was regarded as unlikely, and less relevant to quantify risk of recent harvest levels or future management.

Together, these steps lead to a significant reduction in the total number of initial models from 3888 in 2021 to 648 in 2022. The grid was then further reduced using diagnostics and weighting steps in the next section.

2.5 Weighting the structural uncertainty grid

The 2021 presentation of the structural uncertainty grid did not explicitly weight any of the axes. However, such weighting may be desirable to eliminate implausible models, down-weight less plausible inputs and to provided a more coherent picture of plausible stock trajectories. We distinguish three aspects of weighting, the first two weighting axes *a priori*, before observing outcomes, and the last weighting *a posteriori* based on model outcomes.

1. Weighting data inputs or biological axes by their *a priori* likelihood. Based on our input analyses, for instance, we can weight axes for catch and discards: low and high estimates

for these axes correspond to 90% confidence intervals from catch re-construction and fate (discard) models, respectively. These probabilities can be used directly to weight outcomes from these uncertainty axes. Weights for these axes are given in table 2.

2. Alternative datasets, such as alternative CPUE indices, may be weighted by analysts assigning weights to alternative indices. In our case, we may replace the diagnostic-case index for high-seas/low latitude fisheries based on Japanese high-seas data, with one based on Australian low latitude (<35 °S) fisheries. However, the latter index comes from a more restricted area, and may not reflect the over-all population as well. We may therefore (arbitrarily) assign a lower weight to models based on this alternative index. In our case, we initially assigned arbitrary weights of 0.5 to alternative scenarios in the uncertainty grid, on the basis that the diagnostic case assumptions reflect what we considered the best available data. It turns out that this weighting has relatively little influence in the case of the 2022 uncertainty grid, and we therefore did not explore this weighting any further.

3. Weighting models according to performance criteria is an open research question (e.g., Punt 2022). In order for this weighting to be meaningful, a set of objectives needs to be agreed on, and performance against these objectives measured. Some recent analyses suggest that predictive ability may be used as an objective measure to judge model adequacy or to weight models (Kell et al. 2021, Ducharme-Barth et al. 2021). The Mean Absolute Square Error (MASE) criterion, which measures model predictive skill relative to a random walk, may be used to this end, although a range of other options are available to weight models based on predictive performance (Dormann et al. 2018). Alternatively to assessing or weighting individual models based on predictive performance, which does not guarantee optimal predictive capacity in the resulting ensemble, models may be weighted such that the resulting ensemble minimises predictive loss functions. The latter approach is known as stacking (Dormann et al. 2018, Yao et al. 2018).

We caution here that the validity of predictive measures as tools for model selection depends on the context within which these models are employed - in a management setting, where long term responses to harvest policies are of interest, such model selection based on predictive performance may lead to the selection of sub-optimal harvest or conservation policies (Boettiger 2022). We hypothesize that this phenomenon is due to selection of production-driven over environmentally (e.g., recruitment regime) driven models. Dynamics of the latter are more difficult to predict from stock status and productivity alone – they require knowledge of future environmental states – and may therefore be penalised in such predictive model selection even in situations where they more truthfully describe the mechanistic process.

Herein we only show how predictive model selection would impact on the *a priori* weighted ensemble as an illustrative example. For this purpose, we employed MASE, stacking and inverse-variance (a model-free weighting procedure, where we used the variance of stock status estimates to weight models, following Dormann et al. 2018). Stacking was performed using leave-future-out predictive density, such that weights were determined according to

$$\arg\max_K w_k f(\hat{ix}_t | ix_t, \sigma_{ix}),$$

where $\hat{ix}_t$ is a prediction for index ix_t based on a model fitted to data from 1 to $t-1$. It therefore selects a set of weights that will optimise predictive performance (weighted by index uncertainty σ_{ix}) across all CPUE indices employed within the structural uncertainty grid (within and across models). This formulation is similar to stacking of posterior distributions in Bayesian models (Yao et al. 2018), but uses index weighting rather than prediction uncertainty used in Bayesian stacking.

Given theoretical and practical concerns with predictive model selection, all inferences about stock status, however, are based on the *a priori* weighted ensemble only. Weights for these axes are given in table 2.

3. RESULTS

3.1 Diagnosing the 2021 model grid

Viewing the 2021 structural uncertainty grid in light of $SB_{latest}/SB_{F=0}$ rather than SB_{latest}/SB_0 showed a more consistent grouping of stock status estimates around $SB_{latest}/SB_{F=0}$ =1 (Figure 3). Nevertheless, the model grid also produced a long tail in status estimates, with $SB_{latest}/SB_{F=0}$ of up to 3 and as low as 0.25. The high $SB_{latest}/SB_{F=0}$ estimates are possible due to over-compensatory stock-recruit relationships assumed here, and high status estimates were largely associated with model runs that fixed natural mortality (M) at low values (Figure 4). Using this alternative reference point therefore did not offer a consistent reduction in the range of outcomes from the 2021 structural uncertainty grid. We nevertheless maintained the reference point throughout as it may be regarded as a more appropriate reference point for a stock that was not unfished at the start of our stock assessment time series, and which is likely subject to long-term fluctuations of the pelagic environment.

Standard convergence and stability diagnostics, such as jittering (using a jitter fraction of 0.2), inspecting gradients (Figure 5) and checking for the presence of a positive definite Hessian suggested that all models converged towards a stable solution.

Retrospective analysis using 6 peels suggested mainly relatively low retrospective trends in spawning biomass (Figure 6) and fishing mortality (Figure 7), with ρ estimates for both quantities centered on zero. Very few runs fell outside of thresholds suggested by Hurtado-Ferro et al. (2015), meaning retrospective analyses do not constrain the 2021 grid runs. Forecast ability, on the other hand, was found to be relatively poor across all models (all MASE >1; Figure 8). This result does not come as a surprise given that the model aims to fit three concurrent CPUE indices, which show different levels of decline and rebuilding in recent years. As the model fit represents a compromise, no single index it fitted particularly well, thereby compromising the forecast ability of the model for any individual CPUE index.

3.2 Revised diagnostic case model

The inclusion of additional process error markedly changed the fit to CPUE for the diagnostic case (Figures 9 vs. Figures 2): relative to the 2021 diagnostic case, which assumed very low errors for the NZ CPUE index, and which was correspondingly largely driven by that index. The present analysis provides an improved fit to low-latitude (JP) and EU-Spain CPUE, at the expense of fit to the New Zealand CPUE index (NZ). While fits to aggregate length-frequencies did not change (Figures 10), the model is now driven by the CPUE for the fleet which accounts for the majority of catch and fishing mortality (Figures 11,12).

The updated diagnostic case produced less extreme outcomes than the previous assessment (Figure 13), largely driven by the more subtle trends in JP CPUE relative to NZ CPUE. The model estimated a slightly lower initial status, and a substantially lower recent status in terms of SB_0 than the 2021 diagnostic case. As a consequence, the model did not require large recruitment deviations to explain stock trajectories, with the updated recruitment trajectory effectively smoothing through previous, much larger fluctuations (Figure 13). This suggests that the model went from a model which was largely recruitment driven for the 2021 diagnostic

case, to a model that was largely production driven for the 2022 diagnostic case.

Likelihood profiles revealed that the over-all stock size estimate was still driven by the relative weighting between the New Zealand and remaining indices (Figure 14), but is now not dominated by the former. The conflict in CPUE indices therefore supports the continued use of uncertainty axes that explore relative CPUE weighting. Composition data minima largely fall in the same area, and slightly lower than those of the aggregate CPUE indices.

Estimated natural mortality was 0.19 (Figure 15), with information mainly derived from length-frequency data. Similarly to the profile for R_0, the CPUE indices were in some conflict about natural mortality, with the New Zealand index suggesting higher M (near 0.3) while other CPUE indices favoured lower CPUE near 0.15. Length-frequencies suggested an intermediate M near 0.2.

MASE suggested limited predictive ability for all indices (Figure 17). Similar to the 2021 outcomes, this is likely a reflection of including multiple, not entirely compatible CPUE indices in the model. Model forecasts are close to naive forecasts, reflecting slow stock variation, but also reflect the conflict between New Zealand and high latitude series in terms of the degree of increase in recent years.

3.3 Structural uncertainty grid outcomes

3.3.1 Model weighting: constraining the uncertainty grid

Applying the same criteria as for the 2021 grid showed similar stability of estimates in terms of residual gradients, stability to jittering of input values, and positive definite Hessian solutions.

The grid results for the updated diagnostic case had a much lower spread of status estimates, but also showed more diversity in retrospective patterns (Figures 18, 19). Many of the most strongly positive runs had large $|\rho|$ values relative to 2021. Excluding runs with $|\rho|$>0.2 led to a reduction of the grid that excluded many models with high stock status. Trends in $|\rho|$ were not uniquely explained by any one variable, but by combinations of variables (Figure 20). Retrospective bias estimates for SB and F fell either within a cluster relatively close to zero, or were dispersed away from the main cluster of models at values $|\rho| > 0.2$. We therefore used 0.2 as a cutoff value for retrospective patterns that were deemed acceptable for inclusion within the final grid (see also Hurtado-Ferro et al. 2015).

We further found some models had unrealistically high estimates of M (Figures 22), and these patterns were largely associated with assuming fast growing alternative growth curve (i.e., Joung et al. 2018). Biological plausibility of M was informed based on reasonable M values [0.15; 0.27] determined by empirical relationships with growth parameters and maximum age for large, long lived individuals (Then et al. 2015; see http://barefootecologist.com.au/shiny_m.html). Models using this growth also had the most extreme recent increases in biomass. Based on the plausibility of these outcomes, we decided to drop this growth model from the analysis.

Restricting the grid to models with relatively small retrospective patterns and base growth lead to a reduction from 648 models to 228 models across the structural uncertainty grid. Notably, this filtering also removed models with very high estimates of M (Figures 22), leading to a slightly bi-modal distribution of M estimates for the remaining model runs (Figure 23). While one mode was centered on the prior (and near the estimate of the diagnostic case of 0.19), a second mode of M estimates corresponded to slightly higher estimates of M.

Over-all, the reduction in the model grid based on retrospective patterns and outcomes for M and population trajectories, led to a substantial reduction in the range of grid outcomes

(Figure 25). The procedure did, however, lead to a more strongly apparent bimodal structure in status estimates for SB_{latest}/SB_0 (top row of Figure 26). Applying weights to the individual grid axes shows that the peak at lower stock status for SB_{latest}/SB_0 is due to sensitivities that carry less weight - namely high catch/high initial F estimates. For $SB_{latest}/SB_{F=0}$, the opposite pattern emerges. For example, while the initial subset by retrospective patterns leads to a more pessimistic picture of status, re-weighting with respect to prior input weights down-weights lower status estimates. By down-weighting more extreme catch scenarios, the outcome-density becomes markedly more uni-modal for both reference points, giving a more consistent picture of likely current depletion levels (middle row in Figure 26). Applying weights for CPUE axes had relatively little effect on status estimates.

A posteriori weights were not applied in the grid used for management advice, but were explored as a sensitivity. Weights based on MASE scores for individual models did little to change the distribution of grid outcomes (Figure 26). Weighting model runs based on the inverse of the variance for biomass estimates gave a slight emphasis to more pessimistic model runs, but did not substantially change the over-all picture. Stacking weights had the comparatively largest impact, effectively further up-weighting outcomes with low stock status, though again results were qualitatively similar.

Over-all, our filtering and weighting removed models with the most extreme estimated recruitment deviations, which also showed the most extreme retrospective patterns (Figure 27). MASE weights favoured models with small average recruitment deviations, whereas stacking weights did not appear to favour models with low recruitment deviations.

3.3.2 Analysis of the weighted uncertainty grid

Models in the retained grid showed consistent increases for $SB_{F=0}$ (Figure 28). Models with larger recruitment deviations were associated with higher recent $SB_{F=0}$, resulting in lower estimates of recent status.

Grid models showed high consistency in overall population trajectories (Figure 29, Table 6), with a median current (latest) stock status at $0.90SB_0$ ($0.79SB_{F=0}$), with an 90 percentile range from 0.49–1.01 SB_0 (0.43–$0.93SB_{F=0}$). Model runs with low current SB were associated with high initial F estimates, whereas models using the base initial F assumption showed higher recent status (Figures 30, 32, 33, 31). Relative stock status was only strongly influenced by these initial F assumptions, with alternative CPUE series axes only accounting for secondary effects (Figures 31, 32, 33, A-1). Seventy percent of the grid models had a final year biomass above SB_{MSY}; when accounting for model weights, the probability that 2020 biomass was above MSY was 90%.

MSY varied between a mean of 8 400 mt for scenarios with base catch and initial F assumptions, to up to 25 000 mt for scenarios including high catch and slow growth (Figure 35). The distribution of MSY had a long tail depending on discard and productivity assumptions, with 90% of weighted grid runs showing a MSY between 7 500 and 16 000 mt.

Fishing mortality reference points were determined by a combination of factors in the grid models. F_{MSY} was largely driven by the applied stock-recruit assumption, but varied only slightly from 0.134 to 0.181. Current F relative to F_{MSY} was largely determined by discard levels and initial F (Figure 35). While the high discard scenario lead to estimates of low F relative to F_{MSY} (mean 0.33 for those runs), estimates of $F_{latest}/F_{\text{MSY}}$ were as high as 0.66 with low discard assumptions (Figures 35, 36, 37, 39, 38): Crucially, given relatively low estimated fishing mortality rates compared to F_{MSY}, no models were estimated to exceed other potential reference points ($F_{lim,\text{AS}}$, $F_{crash,\text{AS}}$; Figures A-2, A-3, A-4, A-5)

4. DISCUSSION

In this analysis, we attempted to constrain a large range of previous models to give a more consistent picture of Southwest Pacific blue shark stock status based on a subset of plausible models. To this end, we went beyond simply constraining the initial model grid, and instead re-defined the diagnostic case on which the model grid was formulated. The rationale for this decision came from the realisation that; i) model diagnostics provided little contrast to support the reduction of the initial 2021 structural uncertainty grid, and ii) the lack of contrast was likely driven by overly restrictive assumptions in the initial diagnostic case.

These restrictive assumptions for productivity and error associated with CPUE time series may have led to extreme outcomes in the initial model grid due to mismatches of parameters among uncertainty axes. For example, most very high status estimates for $SB_{latest}/SB_{F=0}$ in the 2021 model grid were from model runs that had low M and fast growth, a combination that counters conventional wisdom that growth and M ought to be correlated (i.e., the M/k invariant). Such correlations are realised when we estimate M in the revised 2022 grid, highlighting the implausibility of the high growth model. Consequently, freeing up M in the diagnostic case served a dual purpose: it not only added additional flexibility to the model, but it also provided a diagnostic tool and an additional lens through which assessment grid outcomes could be further scrutinized for plausibility. In addition, recent advice on estimating M suggests that it is often better to estimate M than to apply fixed values (Punt et al. 2021), and in our case, likelihood profiles support the estimation of M.

The most notable difficulty in assessing the merit of shark stock assessments is the difficulty in judging the usefulness of input data. Many diagnostics (e.g., residuals, predictive model checks), treat stock assessments input as the ground truth. However, it is often difficult to know the degree of uncertainty in inputs, especially for bycatch species where inputs are often derived in a series of more-or-less complex analyses of often poorly representative data. The degree to which input analyses can overcome fundamental deficiencies in the data, such as reporting trends, is often difficult to ascertain. As a result, applying standard diagnostics for model fit to "data" may provide insights into systematic deviations from inputs (e.g., systematically poor fits to CPUE), however, the interpretation of such deviations is more difficult than their statistical formulation suggests. Inconsistencies in CPUE trends for blue shark, for example, suggest process error affecting one or all series. It is near impossible to partition the process error robustly between the different CPUE series, yet the amount of assumed process error will dictate the numerical value or passing of statistical tests (e.g., runs tests of CPUE residuals).

Here, we reversed previous decisions to not explicitly assign process error, and to account for process error in the interpretation of model fits. We assigned process error on the basis of the method proposed in (Francis 2011), but we note that this procedure corresponds to only one assumption about process error, and it ignores CPUE inconsistencies. Nevertheless, from a technical and practical perspective, including additional process error in CPUE led to less extreme trends in model runs, and provided a more consistent set of outcomes.

Selecting the suite of models for advice based on an ensemble of models is an ongoing area of research. One that is unlikely to be definitively resolved as thresholds for particular diagnostics, such as retrospective patterns, likely depend on life history as well as individual stock trajectories (Hurtado-Ferro et al. 2015). In addition, in terms of managing risk of management policies, it is often not clear that the "right" model can be determined from diagnostics alone, and may only be found from applying management and observing stock responses over time (Walters 1981, Boettiger 2022). In that context it may be more important to consider alternative productivity assumptions, such as those considered in the structural

uncertainty grid (Figure 40).

We employed weighting on the basis of prior biological and technical aspects to constrain the model grid from an initial set of 648 models in the revised 2022 uncertainty grid (reduced from 3888 in 2021) to a set of 228 weighted models. While model consistency from retrospective analysis is difficult to interpret, it does provide a way to eliminate highly inconsistent models. Biological plausibility is more easily determined. For example, models with growth modeled by parameters in Joung et al. (2018) gave estimates of M that were inconsistent with the prior, suggesting a mismatch between that data and other biological assumptions. Weighting on the basis of elicited prior weights, for example from the catch reconstruction, is a reasonably interpretable and straightforward process given posterior distributions for predicted catch and discards from the catch reconstruction and fate model, respectively.

Ducharme-Barth et al. (2021) highlighted some of the difficulties with further weighting of grid axes based on model performance, such as predictive performance, noting it is often unclear *a priori* how much weight such a measure should carry, especially if more than one measure is applied. In addition, we do not understand well enough what features measures such as MASE select for and how such selection interacts with estimates of risk associated with management decisions (Boettiger 2022). We suggest that MASE is largely a measure of the degree to which a stock is production driven relative to being recruitment ("regime") driven. The MASE criterion will likely select for production-driven, over recruitment driven models, which may or may not be desirable. We suggest that future research into the application of measures of model predictive capacity to weight models needs to assess the degree to which these measures may affect risk estimates for different types of stocks (e.g., production vs. recruitment regime driven stocks). For models remaining in the blue shark grid after initial filtering, the three metrics we explored had little effect.

Over-all, the weighted model grid removed some of the more optimistic models leaving a set of models with SB_{latest}/SB_0 similar to the initial 2021 grid, but with lower estimates of $SB_{latest}/SB_{F=0}$. Most notably, extreme values of stock status well beyond values of $SB_{latest}/SB_{F=0}$ =1 are not a part of the present grid, and the distribution of stock status is largely uni-modal with relatively constrained estimates. However, we also note that the use of dynamic reference points in the context of shark life-histories is poorly explored, and interpretation of $SB/SB_{F=0}$ is made more difficult by the assumed stock recruitment assumption, which is over-compensatory at parameter values used here. This may lead to non-intuitive outcomes of $SB > SB_{F=0}$ because recruitment at lower SB is higher than in the unfished state, leading to non-trivial interactions between fished stock levels and recruitment dynamics. Reference points based on $SB_{F=0}$ are conditioned on estimates of the latter, the determination of which places considerable faith in the functional form recruitment assumptions and deviations. These may be poorly estimated or represent process error that does not reflect productivity shifts, especially in models without age or informative length data to constrain recruitment estimates. Future assessments could consider switching to more straightforward stock-recruitment assumptions such as Beverton-Holt (ISC 2022) to facilitate interpretation of status with respect to dynamic reference points, and explore how the estimation of dynamic reference points interacts with stock-recruit assumptions.

Uncertainty from the 2022 uncertainty grid did not account for estimation uncertainty and as a result is an underestimate of total uncertainty from the ensemble. While mean/median estimates will be largely unaffected, this could impact the level of risk with overfishing/overfished status (Ducharme-Barth and Vincent in press). We found that, in practice, estimation uncertainty was small relative to structural uncertainties (Figure 41). Nevertheless, future assessments should aim to include both types of uncertainty in reporting from the structural uncertainty grid.

We further highlight that the restriction of the grid outcomes was largely determined by the redevelopment of the diagnostic case and associated changes in the spread of model outcomes, as well as by the application of relatively straightforward filters. Diagnostics and filtering tools need to be interpreted in the context of the input data and modelling decisions. For sharks, there will always remain unanswerable questions about input data, and likely a paucity of biological data on which to structure assessments. We therefore suggest that assessments in this context need to strike a balance between exploring uncertainty due to input and biological unknowns, and finding a set of models that is consistent with best available data.

We suggest that our updated model grid and associated weights provide such a set of models, which acknowledge unknown process error (by varying CPUE weights and dropping conflicting time series, e.g., early NZ CPUE), uncertainty in production (by estimating M) and input data (via uncertainty on initial F and catch). The models nevertheless show a consistent picture of recent rebuilding of the stock due to reductions in catch and associated fishing mortality. We therefore retain most of our conclusions and recommendations from the 2021 blue shark assessment model presented to SC17.

4.1 Main Assessment Conclusions

- The most influential axis within the reduced uncertainty grid was the initial F assumption.
- The stock biomass was low throughout the region through the early 2000s following the expansion of longline fishing effort in the region. But the estimates across the uncertainty grid of 228 models largely indicated that the stock has been recovering since then.
- All 228 model runs indicate that fishing mortality at the end of the assessment period was below F_{MSY} and 87% of (weighted) model runs show that the biomass is above SB_{MSY} (median $SB_{recent}/SB_{\text{MSY}}$ = 1.64 (90^{th} percentiles 0.88 and 1.87; Table 6), with the median estimated depletion $SB_{recent}/SB_{F=0}$ = 0.71 (90^{th} percentiles 0.37 and 0.82), and SB_{recent}/SB_0 = 0.80 (90^{th} percentiles 0.43 and 0.90).
- Fishing mortality has declined over the last decade and is currently relatively low with the median $F_{recent}/F_{\text{MSY}}$ = 0.65 (90^{th} percentiles 0.43 and 0.86; Table 6). This may be a result of most sharks being released upon capture from by most longline fleets.
- Finally, considered against all conventional reference points the stock on average does not appear to be overfished and overfishing is not occurring.

Given some of the uncertainties highlighted above, we recommend that SC18 consider:

- Providing more time, either as inter-session projects, or by extending time-frames for shark analyses. This will allow more thorough investigation of input data quality and trends, which shape assessment choices. In addition, it would allow input analyses to be completed in time to be presented to the pre-assessment workshop prior to the stock assessment. In addition, allowing more time for the assessments themselves will allow a more thorough investigation of alternative model structures, which may include comparisons with low-information methods such as spatial risk assessments.
- Increased effort to re-construct catch histories for sharks (and other bycatch species) from a range of sources. Our catch reconstruction models showed that model assumptions and formulation can have important implications for reconstructed catches. Additional

data sources, such as log-sheet reported captures from reliably reporting vessels, may be incorporated into integrated catch-reconstruction models to fill gaps in observer coverage.

- Additional tagging be carried out using satellite tags in a range of locations, especially known nursery grounds in South-East Australia and New Zealand, as well as high seas areas to the north and east of New Zealand, where catch-rates are high. Such tagging may help to resolve questions about the degree of natal homing and mixing of the stock.
- Tagging may also help to obtain better estimates of natural mortality, if carried out in sufficient numbers. This could be taken up as part of the WCPFC Shark Research Plan to assess the feasibility and scale of such an analysis.
- Additional growth studies from a range of locations could help build a better understanding of typical growth, as well as regional growth differences. Current growth data are conflicting, despite evidence that populations at locations of current tagging studies are likely connected or represent individuals from the same population.
- Genetic/genomic studies could be undertaken to augment the tagging work to help resolve these stock/sub-stock structure patterns. To support this work, a strategic tissue sampling program for sharks is recommended with samples to be stored and curated in the Pacific Marine Specimen Bank.

5. ACKNOWLEDGEMENTS

The authors would like to thank WCPFC for funding this project, which greatly improved the models for blue shark in the southwest pacific ocean.

6. REFERENCES

Boettiger, C. (2022). The forecast trap. *Ecology Letters*, *25*(7), 1655–1664.

Brouwer, S. & Hamer, P. (2020). *2021-2025 Shark Research Plan* (tech. rep. No. EB-IP-01 Rev1). WCPFC.

Clarke, S. & Hoyle, S. (2014). *Development of limit reference points for elasmobranchs* (tech. rep. No. SC10-MI-WP-07). WCPFC.

Dormann, C. F.; Calabrese, J. M.; Guillera-Arroita, G.; Matechou, E.; Bahn, V.; Bartoń, K.; Beale, C. M.; Ciuti, S.; Elith, J.; Gerstner, K., et al. (2018). Model averaging in ecology: A review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504.

Ducharme-Barth, N. & Vincent, M. (in press). Focusing on the front end: A framework for incorporating uncertainty in biological parameters in model ensembles of integrated stock assessments. *Fisheries Research*.

Ducharme-Barth, N.; Castillo-Jordan, C.; Hampton, J.; P, W.; G, P., & P, H. (2021). Stock assessment of southwest pacific swordfish. (WCPFC-SC17-2021/SA-WP-04).

Francis, R. C. (2011). Data weighting in statistical fisheries stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*(6), 1124–1138.

Hamer, P. (2022). *Report from the SPC Pre-assessment Workshop - March 2022* (tech. rep. No. WCPFC-SC18-2022/SA-IP-xx). WCPFC.

Hurtado-Ferro, F.; Szuwalski, C. S.; Valero, J. L.; Anderson, S. C.; Cunningham, C. J.; Johnson, K. F.; Licandeo, R.; McGilliard, C. R.; Monnahan, C. C.; Muradian, M. L., et al. (2015). Looking in the rear-view mirror: Bias and retrospective patterns in integrated, age-structured stock assessment models. *ICES Journal of Marine Science*, *72*(1), 99–110.

ISC (2018). *Stock assessment and future projections of blue shark in the North Pacific Ocean through 2015* (tech. rep. No. SC14-SA-IP-13). WCPFC.
ISC (2022). *Stock assessment and future projections of blue sharks in the North Pacific Ocean through 2020* (tech. rep. No. WCPFC-SC18-2022/SA-WP-06SC14-SA-IP-13). WCPFC.
Joung, S. J.; Lyu, T., G; Hsu, H. H.; Liu, K. M., & Wang, S. B. (2018). Age and growth estimates of the blue shark (*Prionace glauca*) in the central South Pacific Ocean. *Marine and Freshwater Research*. doi:10.1071/MF17098
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.
Manning, M. J. & Francis, M. P. (2005). Age and growth of blue shark (*Prionace glauca*) from the New Zealand Economic Exclusive Zone. *New Zealand Fisheries Assessment Report 2005/26*. 52 p.
Monnahan, C. C.; Branch, T. A.; Thorson, J. T.; Stewart, I. J., & Szuwalski, C. S. (2019). Overcoming long bayesian run times in integrated fisheries stock assessments. *ICES Journal of Marine Science*, *76*(6), 1477–1488.
Neubauer, P.; Large, K.; Brouwer, M., S.and Kai; Tsai, W.-P., & Liu, K.-M. (2021a). *Input data for the 2021 South Pacific Blue Shark stock assessment* (tech. rep. No. WCPFC-SC17-2021/SA-IP-18). WCPFC.
Neubauer, P.; Large, K., & Brouwer, S. (2021b). *Stock assessment for south Pacific blue shark in the Western and Central Pacific Ocean* (tech. rep. No. WCPFC-SC17-2021/SA-WP-03). WCPFC.
Punt, A. E. (2022). *Diagnostics: Yesterday, today and tomorrow*. Retrieved from www . capamresearch.org/content/diagnostics-workshop-presentations
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.
Then, A. Y.; Hoenig, J. M.; Hall, N. G.; Hewitt, D. A., & editor: Ernesto Jardim, H. (2015). Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. *ICES Journal of Marine Science*, *72*(1), 82–92.
Walters, C. J. (1981). Optimum escapements in the face of alternative recruitment hypotheses. *Canadian Journal of Fisheries and Aquatic Sciences*, *38*(6), 678–689.
Yao, Y.; Vehtari, A.; Simpson, D., & Gelman, A. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007.
Zhou, S.; Deng, R.; Hoyle, S., & Dunn, M. (2018). *Identifying appropriate reference points for elasmobranchs within the WCPFC*. WCPFC-SC14-2018/MI-WP-07. Report to the Western and Central Pacific Fisheries Commission Scientific Committee. Fourteenth Regular Session, 8–16 August 2018, Busan, Korea.

7. TABLES

Table 1: Description of the 8 catch scenarios used in the stock assessment. The scenario used for the diagnostic case is highlighted in bold. The total mortality is the cumulative mortality assumed for individuals from the time they are hooked to after they are released back to the water. Further information, see Neubauer et al. (2021a).

Catch scenario	Catch levels	Discard and post-release-mortality
Catch (Post exp)	Mean	100% mortality on all catches, independently of discard status
Low Disc. (Post exp)	Mean	25% quantile of posterior distribution of annual discard rates; 17% post-release mortality
Mean Disc. (Post exp)	**Mean**	**Posterior mean of annual discard rates; 17% post-release mortality**
High Disc. (Post exp)	Mean	75% quantile of posterior distribution of annual discard rates; 17% post-release mortality
High Catch (90%)	90^{th} quantile	100% mortality on all catches, independently of discard status
Low Disc (90%)	90^{th} quantile	25% quantile of posterior distribution of annual discard rates; 17% post-release mortality
Mean Disc (90%)	90^{th} quantile	Posterior mean of annual discard rates; 17% post-release mortality
High Disc (90%)	90^{th} quantile	75% quantile of posterior distribution of annual discard rates; 17% post-release mortality

Table 2: Description of the seven axes for the updated 2022 structural uncertainty grid. Base settings used under the diagnostic case are highlighted in bold. Weights used for alternative values in the weighting of the grid axes are given in parentheses .

Axis	Description
Catch scenario	**Base (0.9)**, high (0.1)
Discard scenario	Low (0.25), **base (0.5)**, high (0.25)
Initial F	**base (0.9)**, high (0.1)
High latitude CPUE	**Base (1)**, low weight (0.5), remove (RM) early New Zealand (0.5)
Low latitude CPUE	**Japan (1)**, Australia (0.5), remove EU CPUE (0.5)
Survival fraction	**Base**, low, high
Growth	**Manning and Francis (2005)**, Joung et al. (2018)

Table 3: Description of the symbols used in the yield and stock status analyses. In this assessment, 'recent' is the average of the metric over the period 2017–2020, and 'latest' is 2020.

Symbol	Description
C_{latest}	Catch in the last year of the assessment (2020)
C_{recent}	Catch in a recent period of the assessment (2017–2020)
MSY	Equilibrium yield at MSY
SB_0	Equilibrium unfished spawning biomass under average recruitment
$SB_{F=0}$	Average spawning biomass predicted in the absence of fishing and using estimated recruitment deviations for the period 2010–2019.
SB_{MSY}	Spawning biomass that will produce MSY
SB_{latest}	Spawning biomass in the last year of the assessment (2020)
SB_{recent}	Spawning biomass in a recent period of the assessment (2017–2020)
SB_{latest}/SB_0	Spawning biomass in the latest time period (2020) relative to the equilibrium spawning biomass under $F = 0$ and average recruitment
SB_{recent}/SB_0	Spawning biomass in the recent time period (2017–2020) relative to the equilibrium spawning biomass under $F = 0$ and average recruitment
$SB_{latest}/SB_{F=0}$	Spawning biomass in the latest time period (2020) relative to the average spawning biomass predicted in the absence of fishing and using estimated recruitment deviations for the period 2010–2019.
$SB_{recent}/SB_{F=0}$	Spawning biomass in the recent time period (2017–2020) relative to the average spawning biomass predicted in the absence of fishing and using estimated recruitment deviations for the period 2010–2019.
SB_{latest}/SB_{MSY}	Spawning biomass in the latest time period (2020) relative to that which will produce the maximum sustainable yield (MSY)
SB_{recent}/SB_{MSY}	Spawning biomass in the recent time period (2017–2020) relative to that which will produce the maximum sustainable yield (MSY)
F_{MSY}	Fishing mortality producing the maximum sustainable yield (MSY)
F_{limAS},	Fishing mortality resulting in 0.5 of SB_{MSY}
$F_{crashAS}$	Fishing mortality resulting in population extinction when sustained on the long-term
F_{latest}/F_{MSY}	Average fishing mortality-at-age for the last year of the assessment (2020)
F_{recent}/F_{MSY}	Average fishing mortality-at-age for a recent period (2017–2020)
F_{latest}	Latest fishing mortality (2020) compared to that producing maximum sustainable yield (MSY)
F_{recent}	Recent fishing mortality (2017–2020) compared to that producing maximum sustainable yield (MSY)
F_{latest}/F_{limAS}	Latest fishing mortality (2020) compared to that resulting in 0.5 of SB_{MSY}
F_{recent}/F_{limAS}	Recent fishing mortality (2017–2020) compared to that resulting in 0.5 of SB_{MSY}
$F_{latest}/F_{crashAS}$	Latest fishing mortality (2020) compared to that resulting in population extinction
$F_{recent}/F_{crashAS}$	Recent fishing mortality (2017–2020) compared to that resulting in population extinction

Table 4: Summary of reference points for the subset of 648 grid models in the structural uncertainty grid, before sub-setting and re-weighting of grid axes.

	Mean	Median	Min	10%	90%	Max
C_{latest}	6176	6224	3505	3840	8992	9601
C_{recent}	7085	7429	4133	4508	9301	9864
MSY	27082	12500	8968	9734	114242	133588
SB_0	51779	21340	12776	15452	198755	250137
$SB_{F=0}$	53092	24356	13490	16697	197250	233167
SB_{MSY}	25965	10446	6098	7557	100582	127141
SB_{latest}	48322	17101	10836	12878	237144	278838
SB_{recent}	47828	15677	11007	12299	248124	291087
SB_{latest}/SB_0	0.85	0.88	0.42	0.49	1.12	1.23
SB_{recent}/SB_0	0.80	0.82	0.37	0.43	1.07	1.28
$SB_{latest}/SB_{F=0}$	0.79	0.79	0.32	0.43	1.18	1.29
$SB_{recent}/SB_{F=0}$	0.74	0.74	0.29	0.37	1.14	1.29
$SB_{latest}/SB_{\text{MSY}}$	1.72	1.79	0.85	0.99	2.32	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.62	1.69	0.76	0.87	2.12	2.53
F_{MSY}	0.174	0.173	0.134	0.141	0.210	0.231
$F_{lim,\text{AS}}$	0.277	0.274	0.211	0.224	0.336	0.374
$F_{crash,\text{AS}}$	0.396	0.393	0.299	0.318	0.479	0.538
F_{latest}	0.069	0.074	0.003	0.007	0.114	0.153
F_{recent}	0.084	0.087	0.004	0.008	0.141	0.176
$F_{latest}/F_{\text{MSY}}$	0.41	0.42	0.01	0.03	0.70	0.78
$F_{recent}/F_{\text{MSY}}$	0.50	0.52	0.02	0.04	0.88	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.26	0.26	0.01	0.02	0.44	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.32	0.33	0.01	0.02	0.55	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.18	0.19	0.01	0.02	0.31	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.22	0.23	0.01	0.02	0.39	0.48

Table 5: Summary of reference points and stock status for the subset of 228 grid models in the structural uncertainty grid, after sub-setting the grid for model runs that showed acceptable retrospective patterns and estimates for natural mortality, but before weighting of grid axes.

	Mean	Median	Min	10%	90%	Max
C_{latest}	6183	6421	3707	3876	8919	9601
C_{recent}	7042	7514	4322	4513	9247	9577
MSY	13976	13765	8968	9564	18737	25629
SB_0	27533	22828	15686	19042	37955	53503
$SB_{F=0}$	31556	27905	17559	20335	44692	66434
SB_{MSY}	13432	11162	7564	9095	18703	26684
SB_{latest}	18099	17101	12973	13832	24631	38004
SB_{recent}	16017	15257	11320	12139	21686	33654
SB_{latest}/SB_0	0.71	0.69	0.42	0.45	0.96	1.19
SB_{recent}/SB_0	0.63	0.60	0.37	0.39	0.84	1.05
$SB_{latest}/SB_{F=0}$	0.63	0.60	0.32	0.38	0.89	1.29
$SB_{recent}/SB_{F=0}$	0.56	0.53	0.29	0.33	0.79	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.45	1.44	0.85	0.91	1.94	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.29	1.25	0.76	0.81	1.71	2.19
F_{MSY}	0.148	0.147	0.134	0.136	0.162	0.181
$F_{lim,\text{AS}}$	0.235	0.232	0.211	0.215	0.255	0.291
$F_{crash,\text{AS}}$	0.335	0.331	0.299	0.306	0.367	0.419
F_{latest}	0.078	0.078	0.039	0.051	0.104	0.120
F_{recent}	0.098	0.095	0.048	0.065	0.133	0.160
$F_{latest}/F_{\text{MSY}}$	0.53	0.54	0.24	0.34	0.72	0.78
$F_{recent}/F_{\text{MSY}}$	0.67	0.65	0.30	0.43	0.94	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.33	0.34	0.15	0.21	0.46	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.42	0.41	0.19	0.27	0.60	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.23	0.24	0.11	0.15	0.32	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.30	0.29	0.13	0.19	0.42	0.48

Table 6: Summary of reference points and stock status for the subset of 228 grid models in the structural uncertainty grid, after sub-setting the grid for model runs that showed acceptable retrospective patterns and estimates for natural mortality. Grid axes are weighted by prior input weights.

	Mean	Median	Min	10%	90%	Max
C_{latest}	5965	5671	3707	3978	7593	9601
C_{recent}	6912	6744	4322	4596	8926	9577
MSY	11413	9993	8968	9313	16333	25629
SB_0	22772	20603	15686	18524	32263	53503
$SB_{F=0}$	25894	22658	17559	20161	38033	66434
SB_{MSY}	11104	9985	7564	9008	15854	26684
SB_{latest}	18420	17904	12973	15902	20424	38004
SB_{recent}	16344	15907	11320	14000	17670	33654
SB_{latest}/SB_0	0.85	0.90	0.42	0.49	1.01	1.19
SB_{recent}/SB_0	0.76	0.80	0.37	0.43	0.90	1.05
$SB_{latest}/SB_{F=0}$	0.76	0.79	0.32	0.43	0.93	1.29
$SB_{recent}/SB_{F=0}$	0.67	0.71	0.29	0.37	0.82	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.75	1.84	0.85	1.00	2.10	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.55	1.64	0.76	0.88	1.87	2.19
F_{MSY}	0.144	0.142	0.134	0.136	0.158	0.181
$F_{lim,\text{AS}}$	0.228	0.225	0.211	0.214	0.248	0.291
$F_{crash,\text{AS}}$	0.325	0.320	0.299	0.304	0.351	0.419
F_{latest}	0.073	0.072	0.039	0.051	0.093	0.120
F_{recent}	0.094	0.094	0.048	0.065	0.117	0.160
$F_{latest}/F_{\text{MSY}}$	0.51	0.52	0.24	0.35	0.67	0.78
$F_{recent}/F_{\text{MSY}}$	0.65	0.65	0.30	0.43	0.86	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.32	0.33	0.15	0.22	0.43	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.41	0.41	0.19	0.27	0.55	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.23	0.23	0.11	0.15	0.30	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.29	0.29	0.13	0.19	0.39	0.48

8. FIGURES

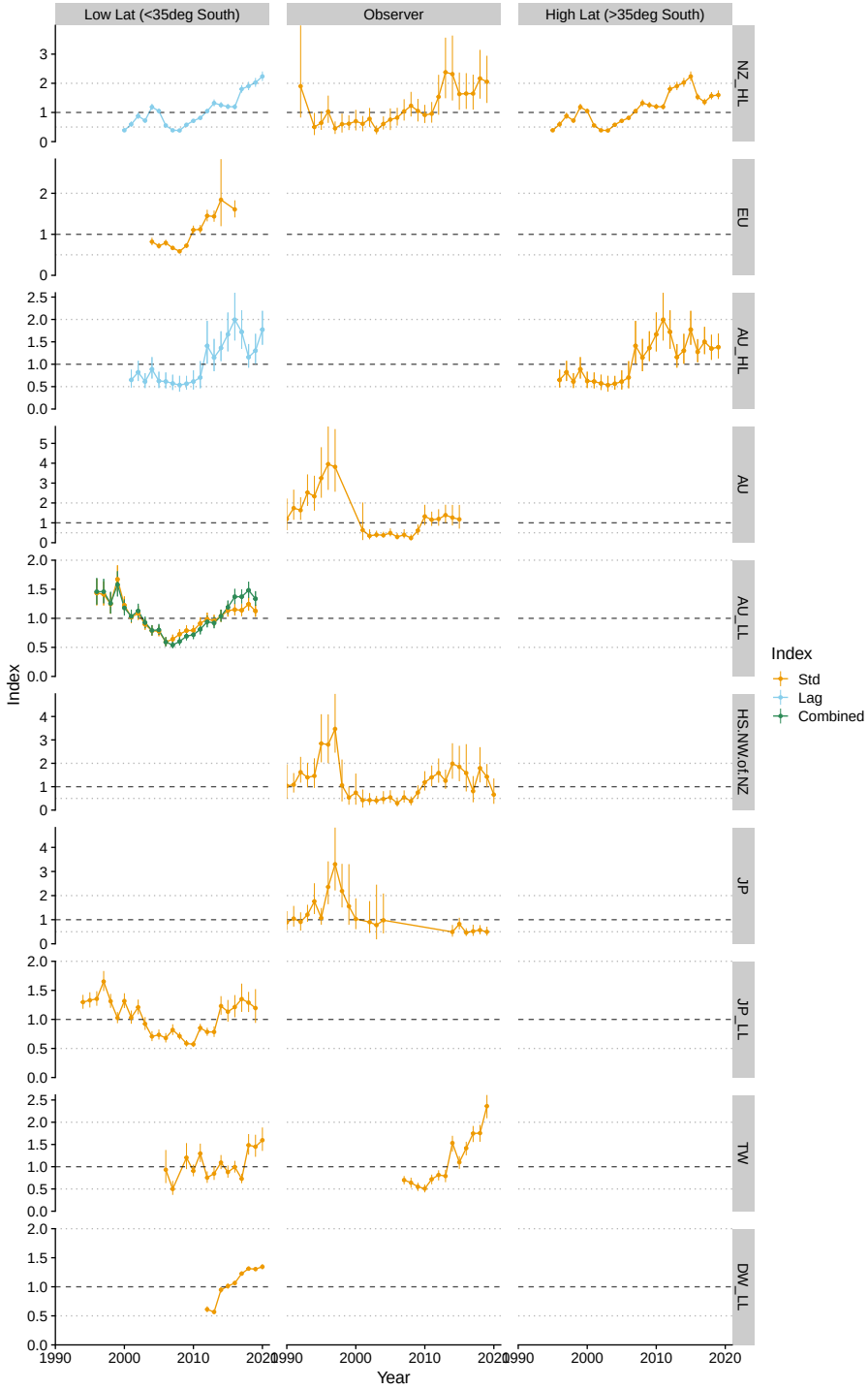


Figure 1: Standardised (circles with standard error) CPUE indices for CCMs included in the logsheet CPUE analyses. The Chinese–Taipei observer CPUE is included for comparison. To aid comparison between high- and low- latitude CPUE series, the high-latitude indices were lagged by 5 years and re-plotted (blue CPUE) with the low-latitude indices; 4-5 years is the apparent lag given length frequencies observed in the high latitude fisheries.

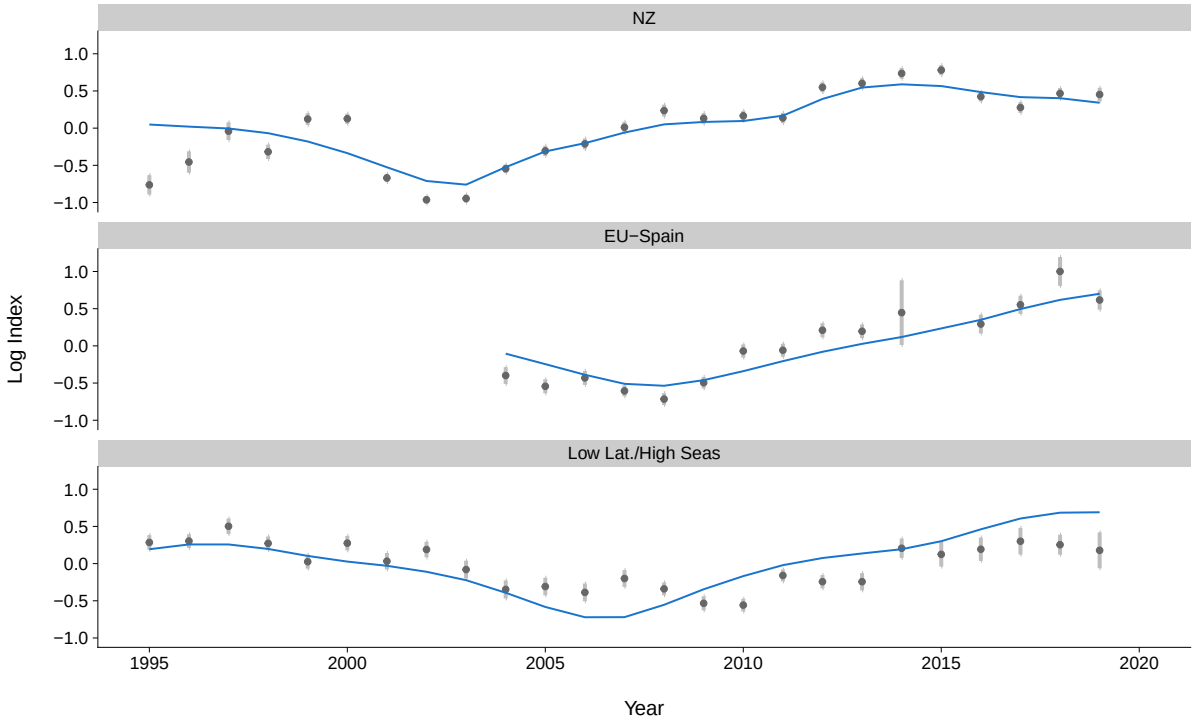


Figure 2: Observed (grey dots) vs. predicted (blue line) CPUE on the log-scale for index longline fleets under the 2021 diagnostic case, with vertical light grey bands showing the 95% confidence interval for each year's index.

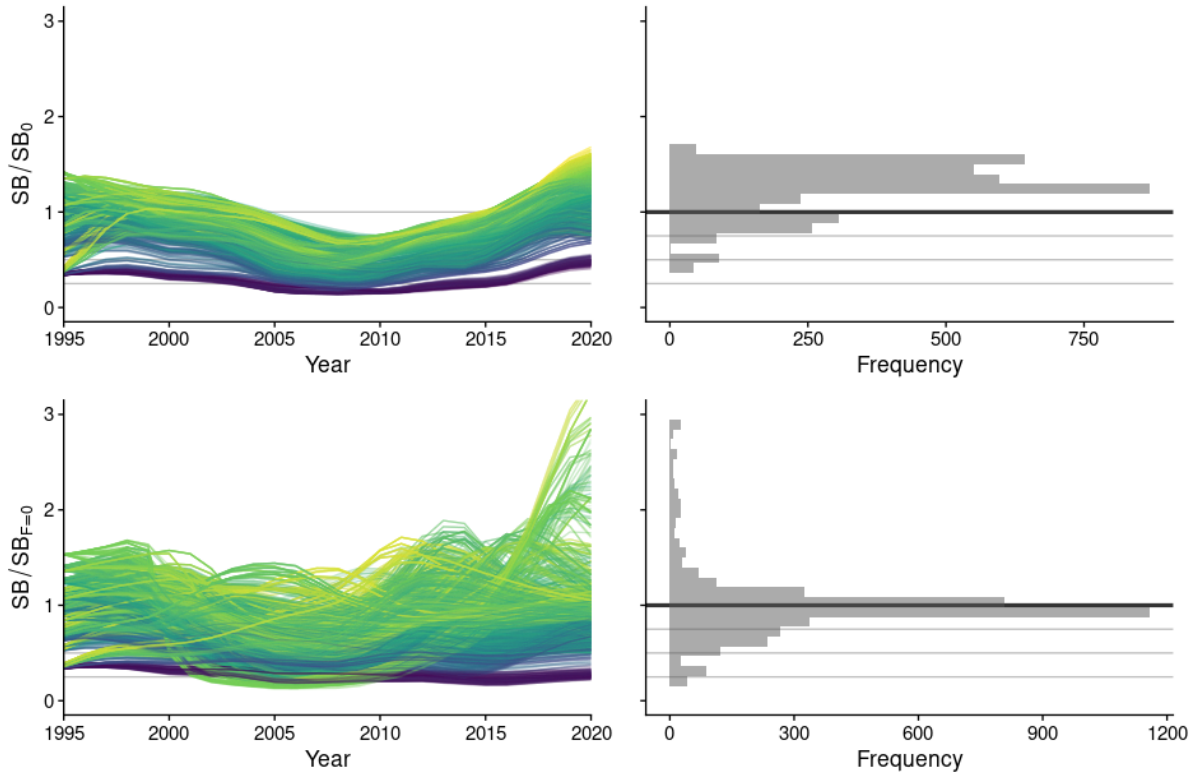


Figure 3: Stock trajectories relative to reference points (SB_0 or $SB_{F=0}$) for model runs in the 2021 structural uncertainty grid (left plots). Latest stock status (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) is graphed on right panels, and trajectories are coloured by latest stock status with regards to SB_0.

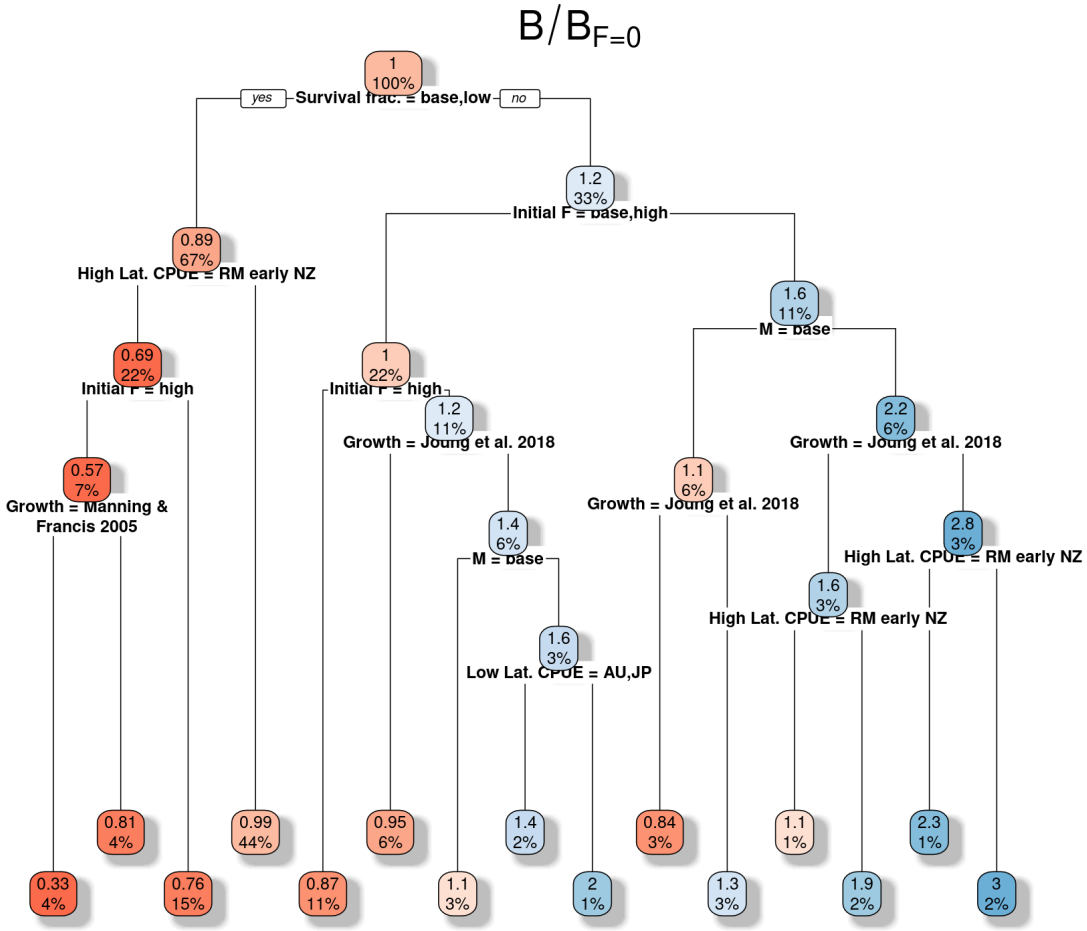


Figure 4: Decision tree for $SB_{latest}/SB_{F=0}$ across the 2021 structural uncertainty grid for blue shark: positive ('yes') values for each split are on the left, leafs on the decision tree show the mean value of $SB_{latest}/SB_{F=0}$ by leaf, as well as the percentage of records on that leaf.

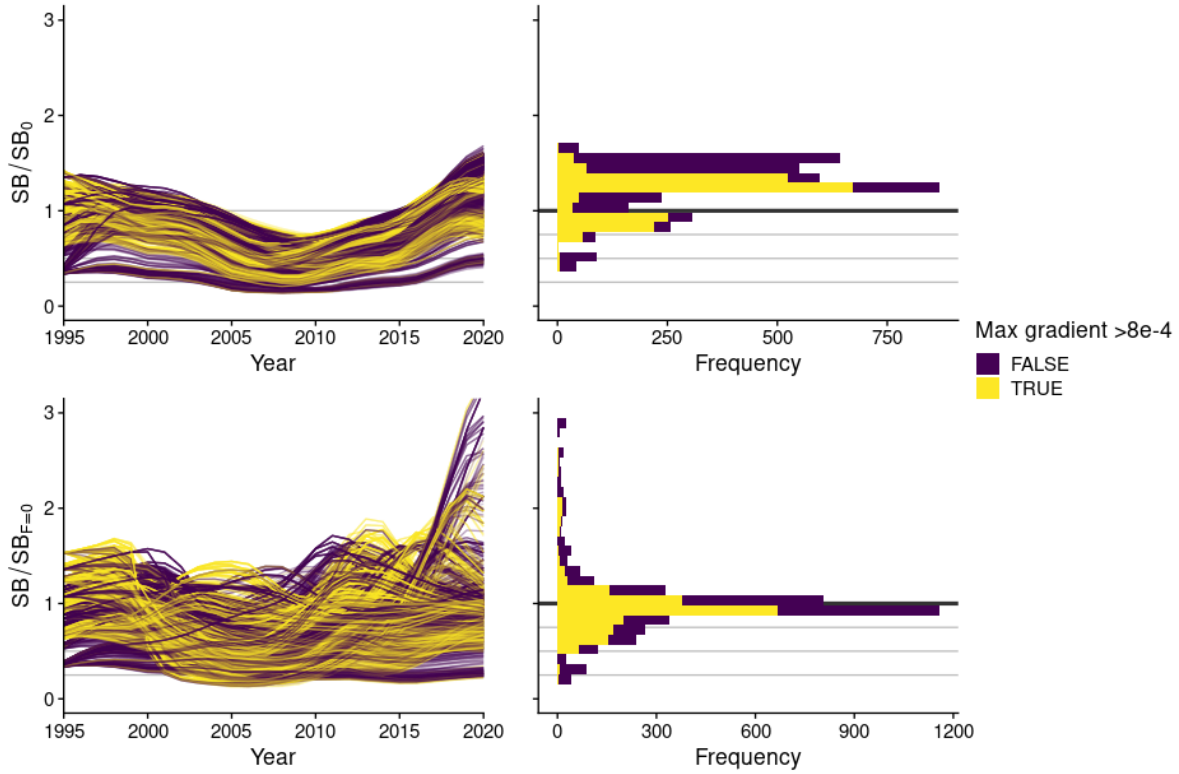


Figure 5: Stock trajectories relative to reference points (SB_0 or $SB_{F=0}$) for model runs in the 2021 structural uncertainty grid (left plots). Latest stock status (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) is graphed on right panels. Trajectories and status are coloured by gradients exceeding an arbitrary threshold value that appeared to divide model runs along growth assumptions.

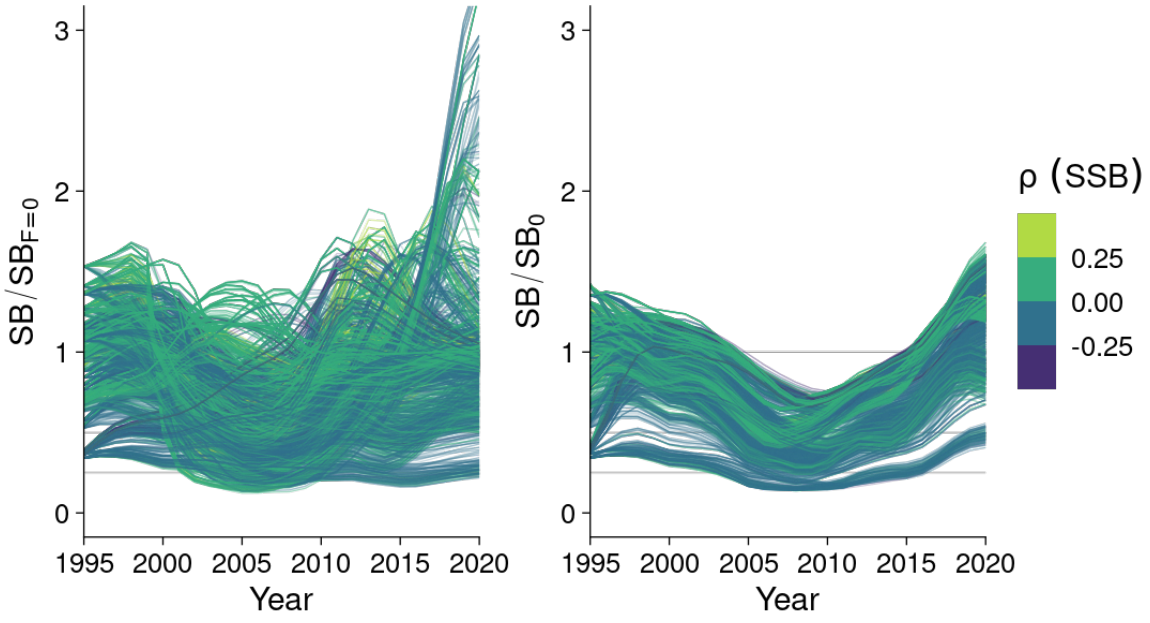


Figure 6: Stock trajectories relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) for model runs in the 2021 structural uncertainty grid for blue shark, coloured by Mohn's ρ for spawning biomass estimates.

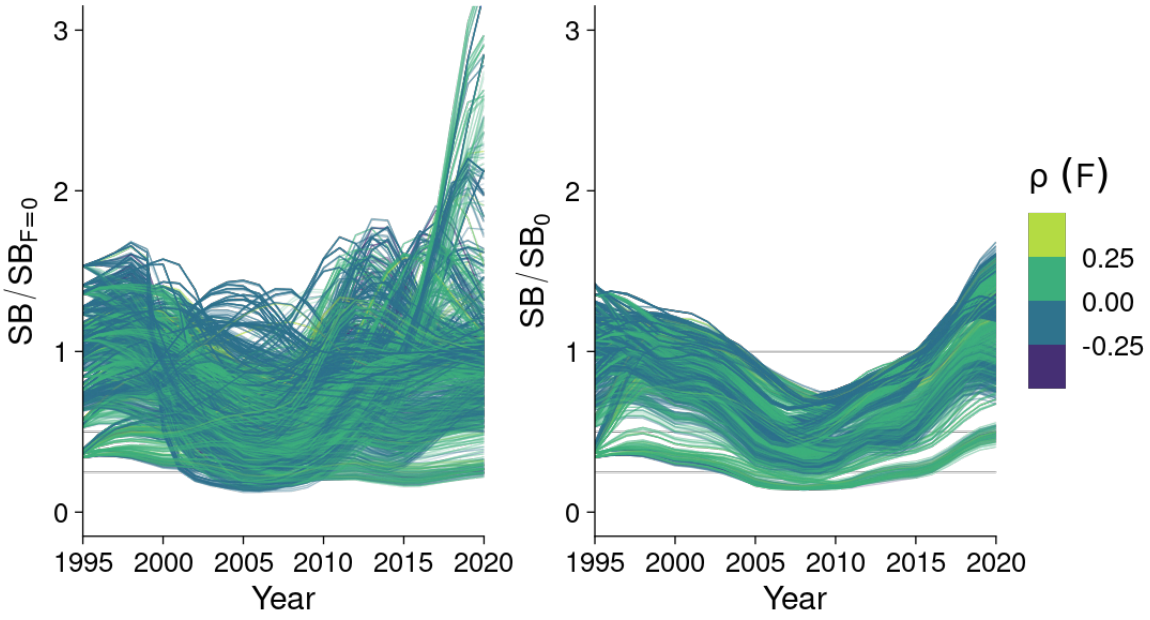


Figure 7: Stock trajectories relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) for model runs in the 2021 structural uncertainty grid for blue shark, coloured by Mohn's ρ for fishing mortality (F) estimates.

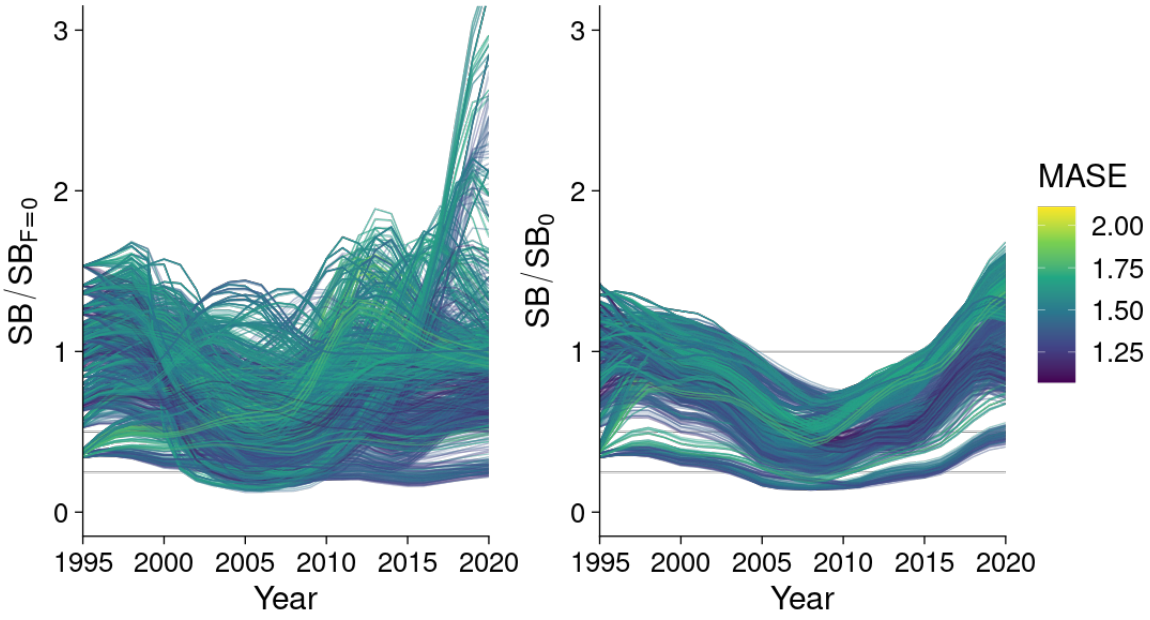


Figure 8: Stock trajectories relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) for model runs in the 2021 structural uncertainty grid for blue shark, coloured by average mean absolute square error (MASE) across 6 peels used for cross-validation

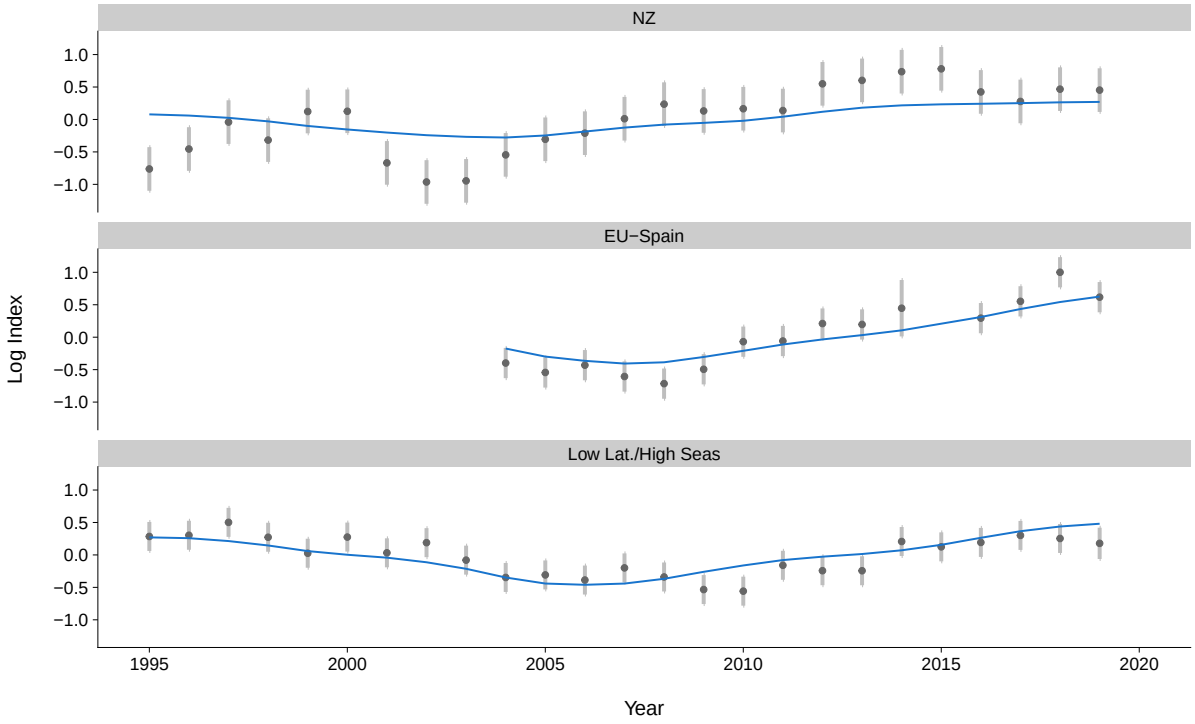


Figure 9: Observed (grey dots) vs. predicted (blue line) CPUE on the log-scale for index long-line fleets under the 2022 diagnostic case, with vertical light grey bands showing the 95% confidence interval for each year index.

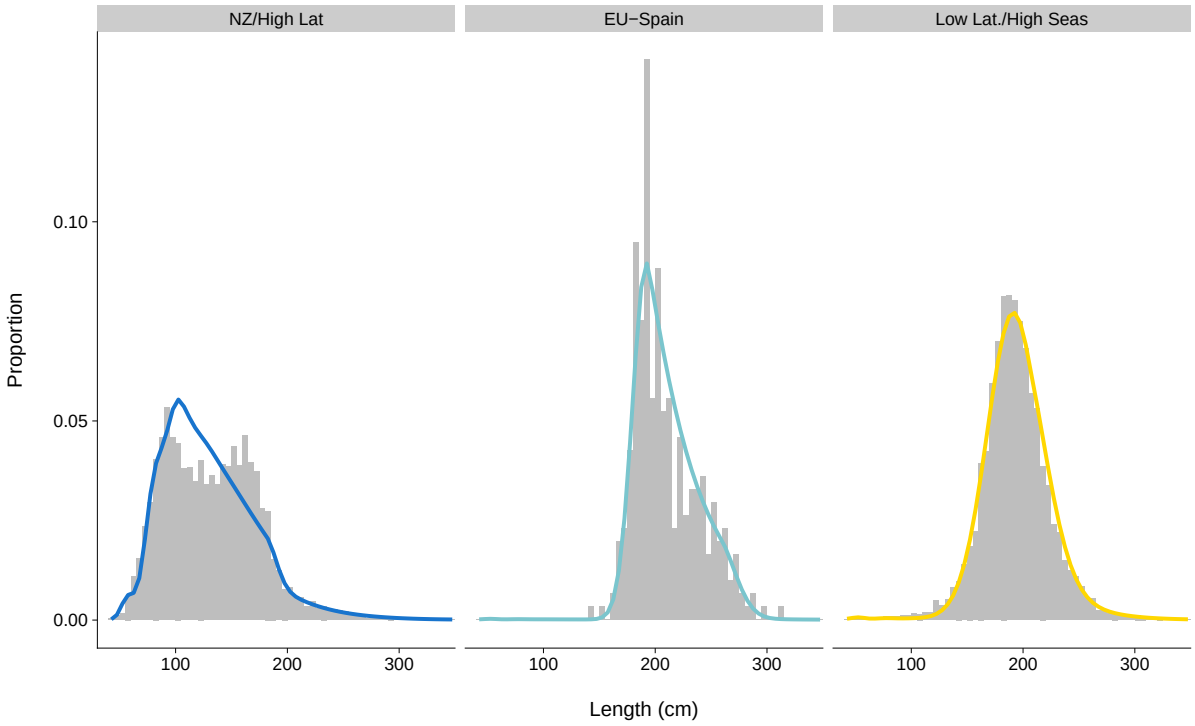


Figure 10: Observed (grey bars) vs. predicted (coloured line) catch-at-length for each fleet aggregated over all years for the 2022 diagnostic case.

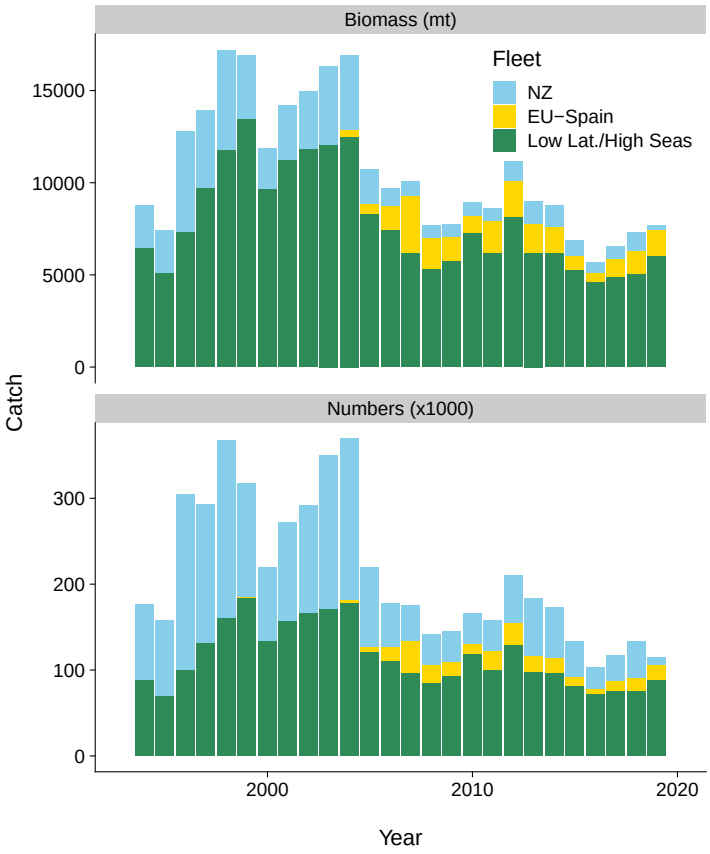


Figure 11: Catch by fleet in biomass and numbers for the 2022 diagnostic case.

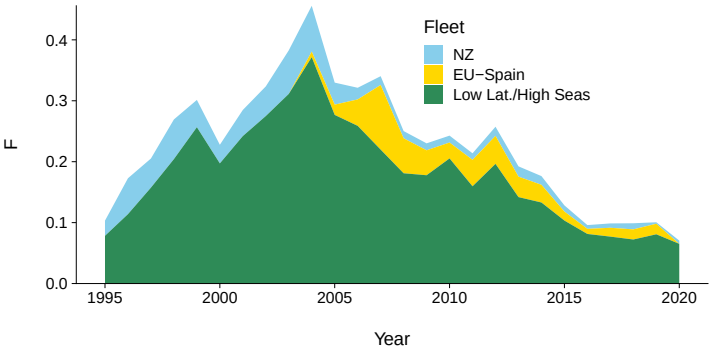


Figure 12: Fishing mortality by fleet estimated for the 2022 diagnostic case over the time - span of the assessment.

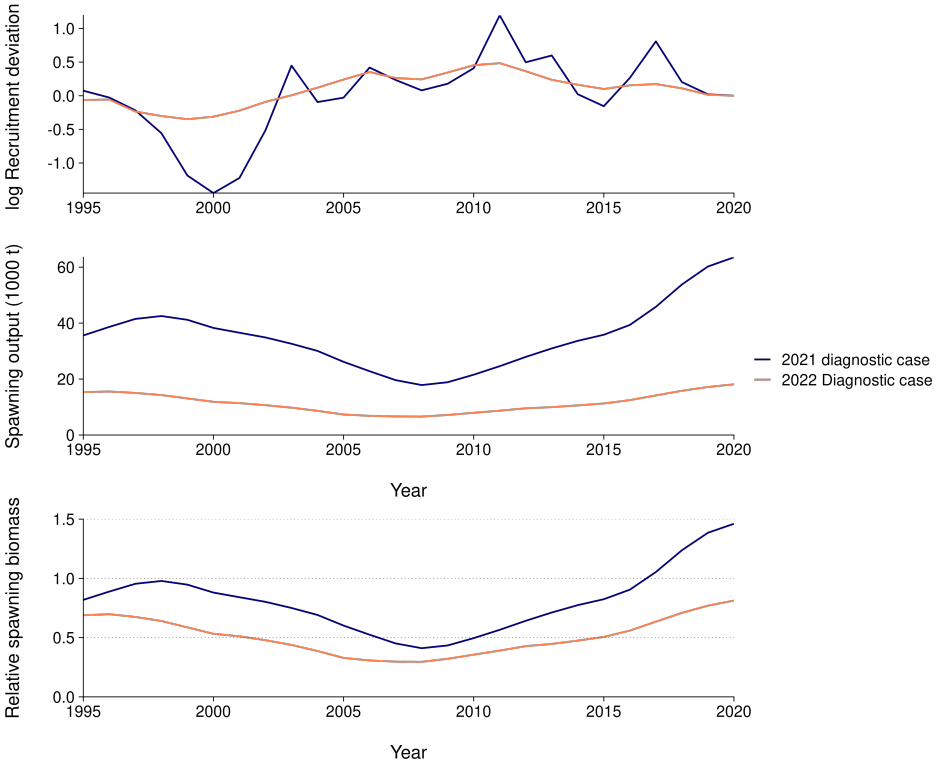


Figure 13: Total biomass, recruitment and spawning biomass for the 2022 diagnostic case estimated between 1995–2020.

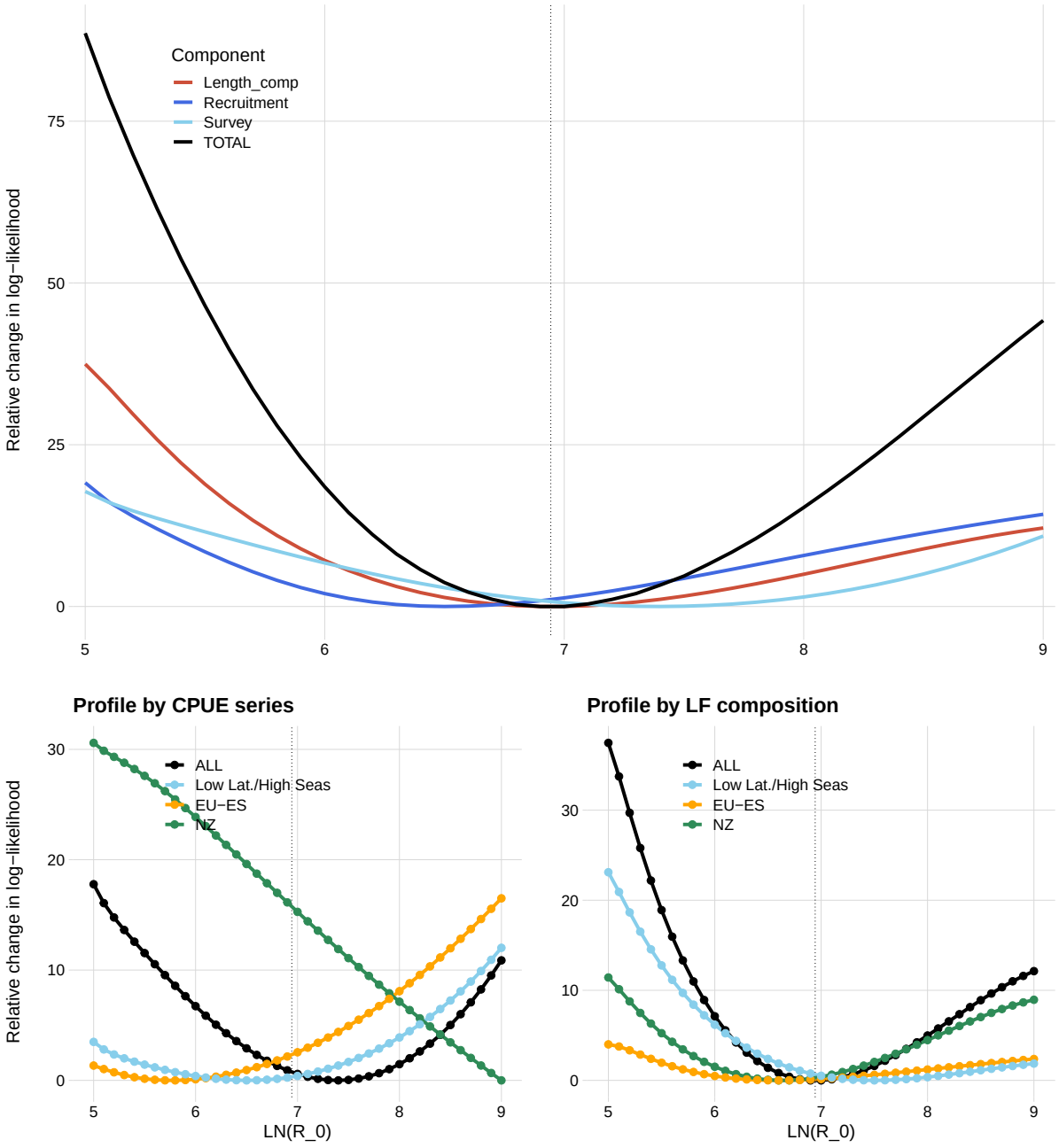


Figure 14: Relative change in log-likelihood for different values of $LN(R_0)$ for the 2022 diagnostic case. The top panel shows the total likelihood and contribution by each component. The bottom panels show individual components by fleet for the CPUE (left) and catch-at-length data (right). The dotted line shows the value for $LN(R_0)$ estimated under the diagnostic case.

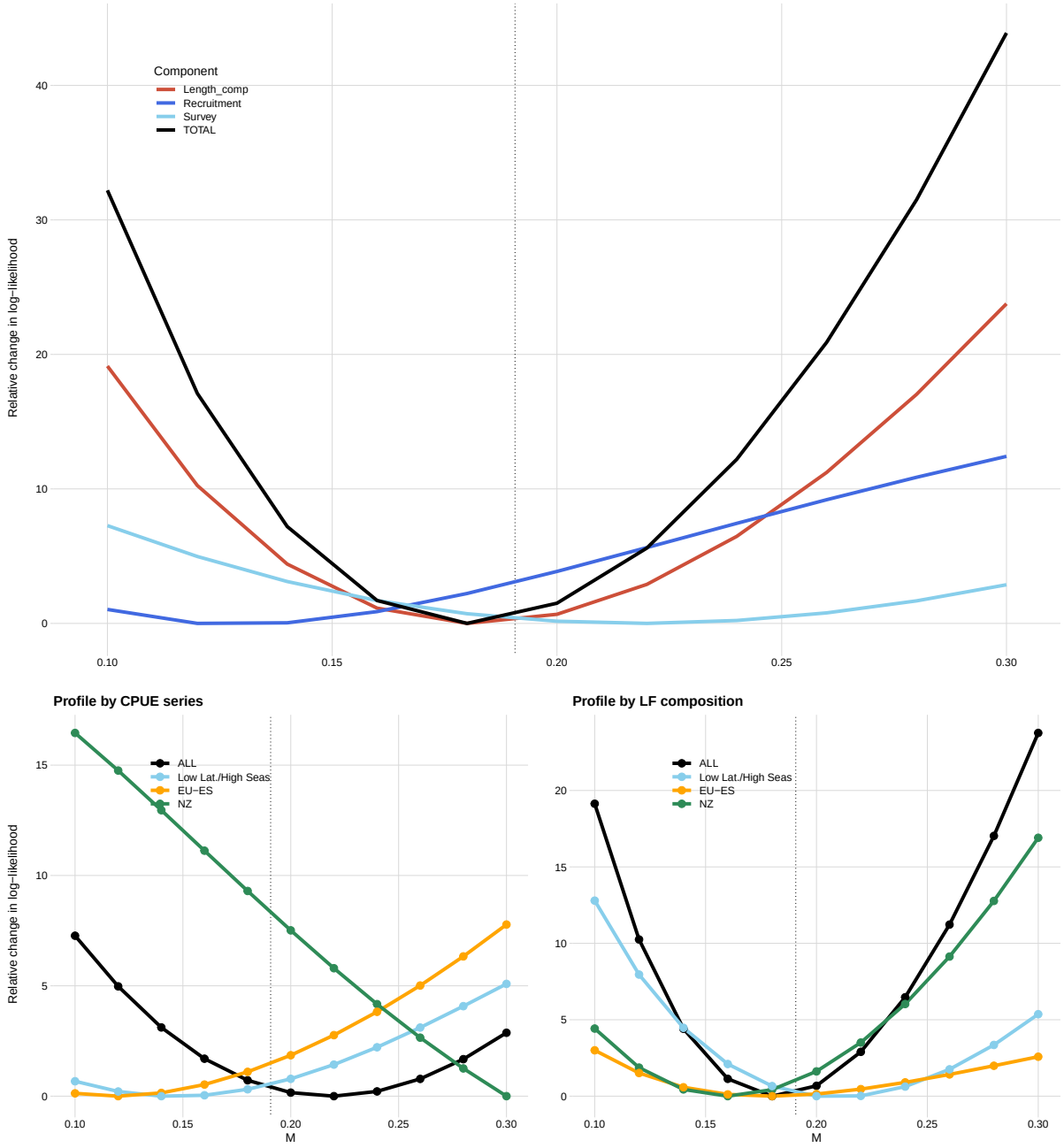


Figure 15: Relative change in log-likelihood for different values of M for the 2022 diagnostic case. The top panel shows the total likelihood and contribution by each component. The bottom panels show individual components by fleet for the CPUE (left) and catch-at-length data (right). The dotted line shows the value for M estimated under the diagnostic case.

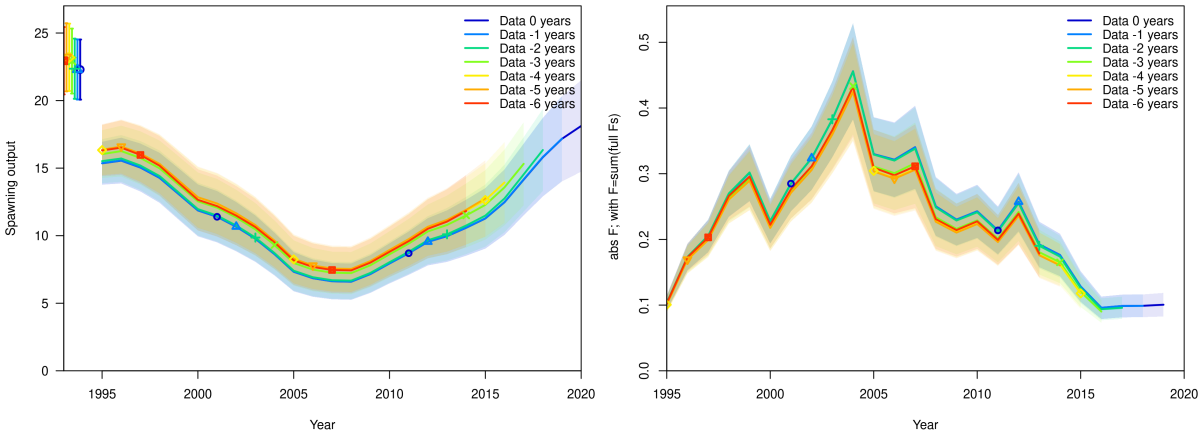


Figure 16: Retrospective patterns of spawning biomass and fishing mortality for the 2022 diagnostic case, compared with estimated uncertainty intervals.

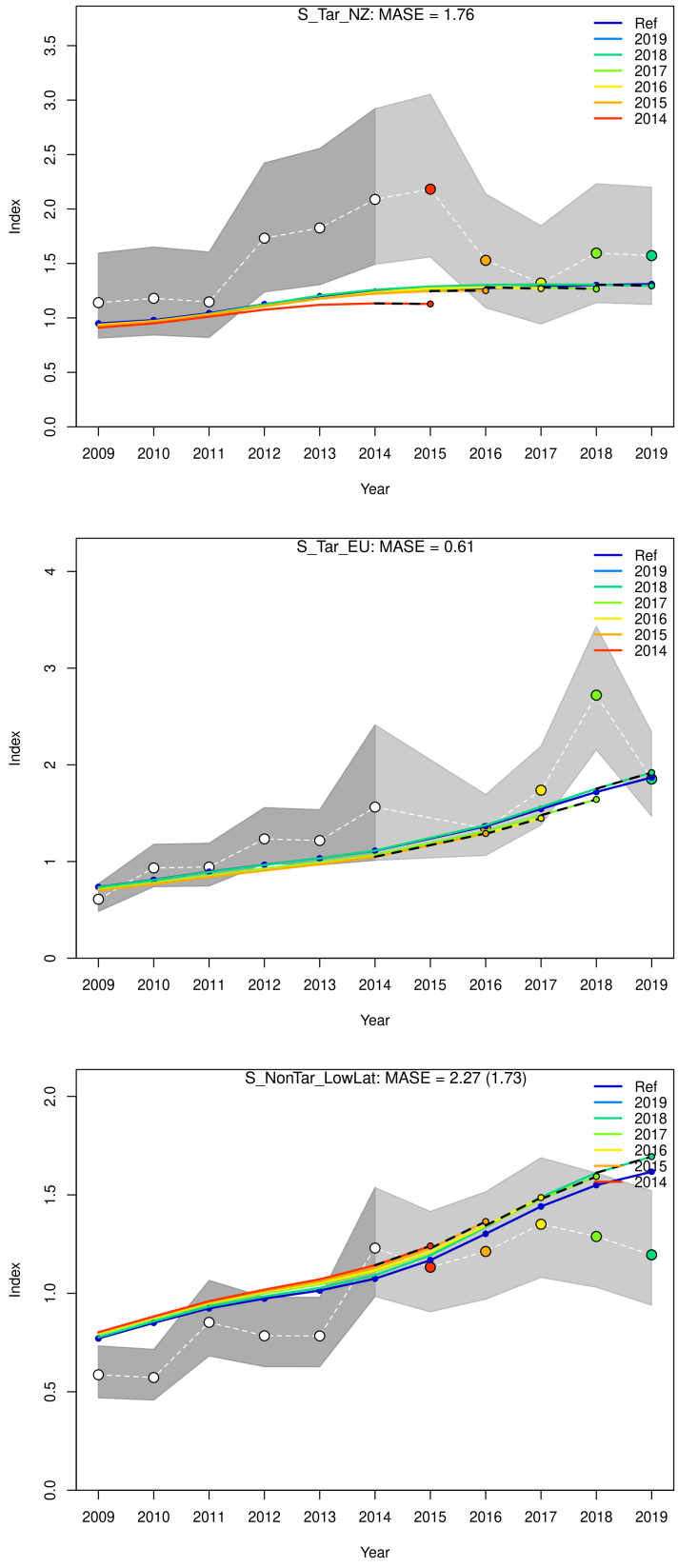


Figure 17: MASE for predictions from the model (coloured points) relative to a naive prediction (blue) for the 2022 diagnostic case.

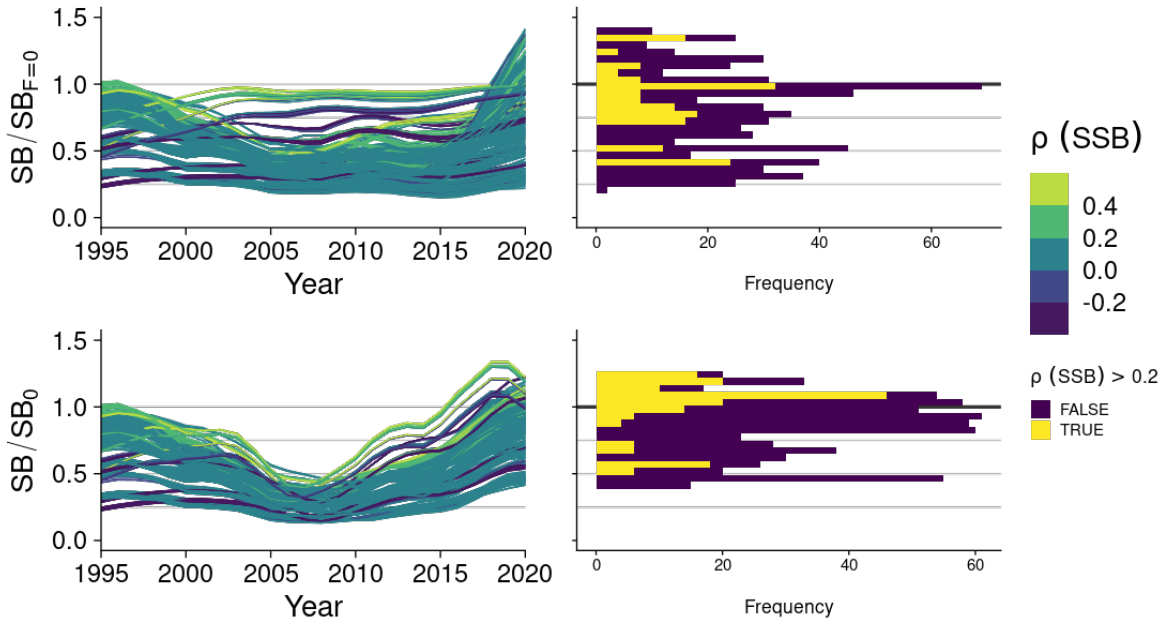


Figure 18: Stock trajectories relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) for model runs in the 2022 structural uncertainty grid for blue shark, coloured by Mohn's ρ for spawning biomass estimates. Models with $|\rho| > 0.2$ were excluded from subsequent analyses.

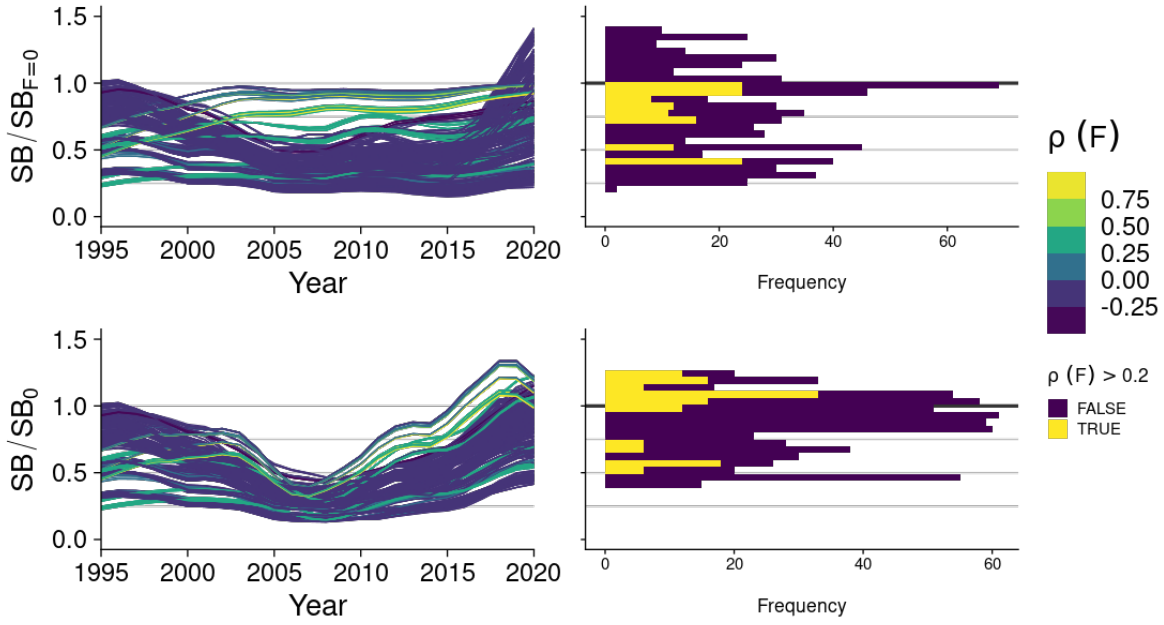


Figure 19: Stock trajectories relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) for 648 model runs in the 2022 structural uncertainty grid for blue shark, coloured by Mohn's ρ for fishing mortality (F) estimates. Models with $|\rho| > 0.2$ were excluded from subsequent analyses.

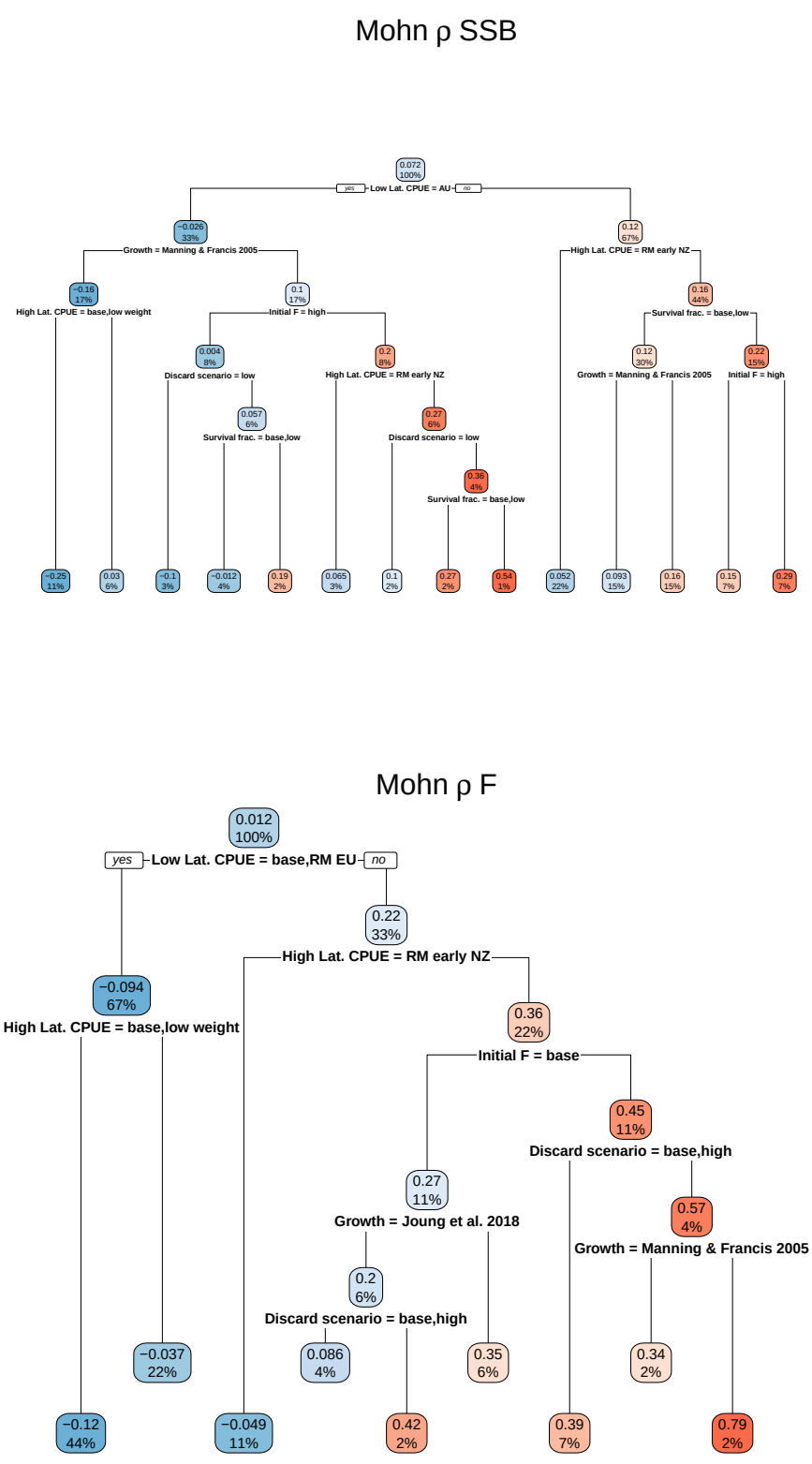


Figure 20: Decision tree for Mohn's ρ for spawning biomass (top) and fishing mortality (F; bottom) across 648 models in the 2022 structural uncertainty grid for blue shark: positive ('yes') values for each split are on the left, leafs on the decision tree show the mean value of ρ by leaf, as well as the percentage of records on that leaf.

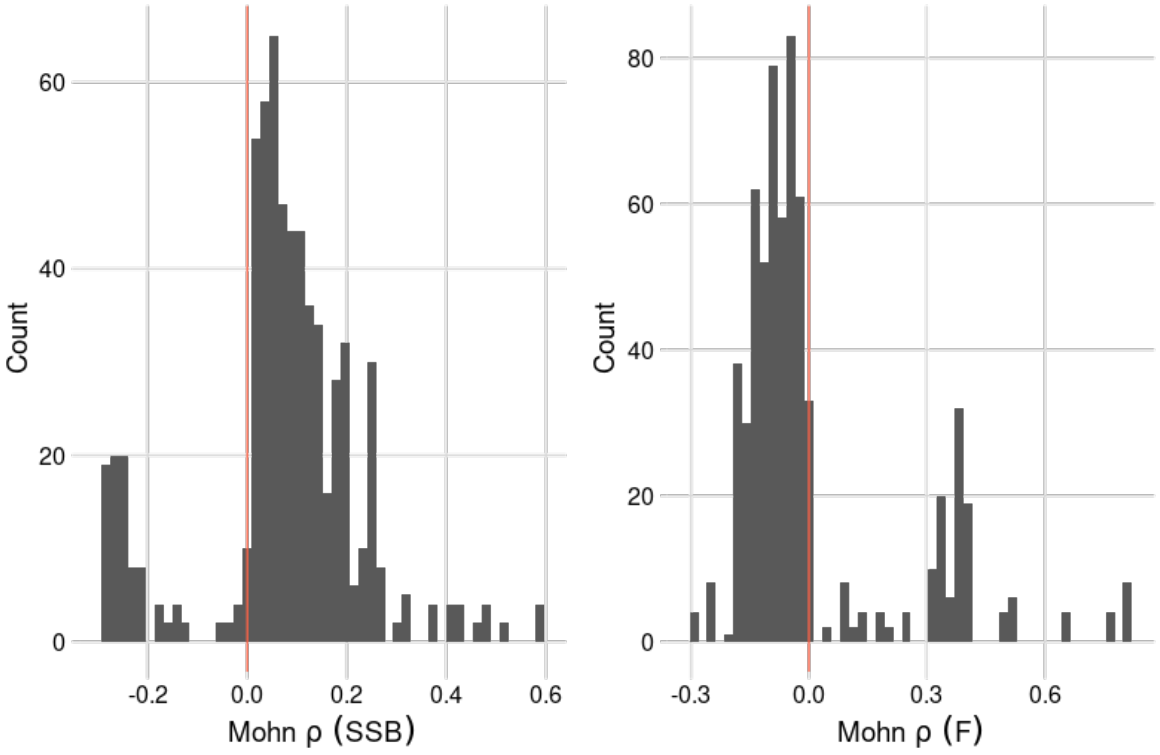


Figure 21: Mohn's ρ for spawning biomass and fishing mortality across the 648 models in the 2022 structural uncertainty grid for blue shark.

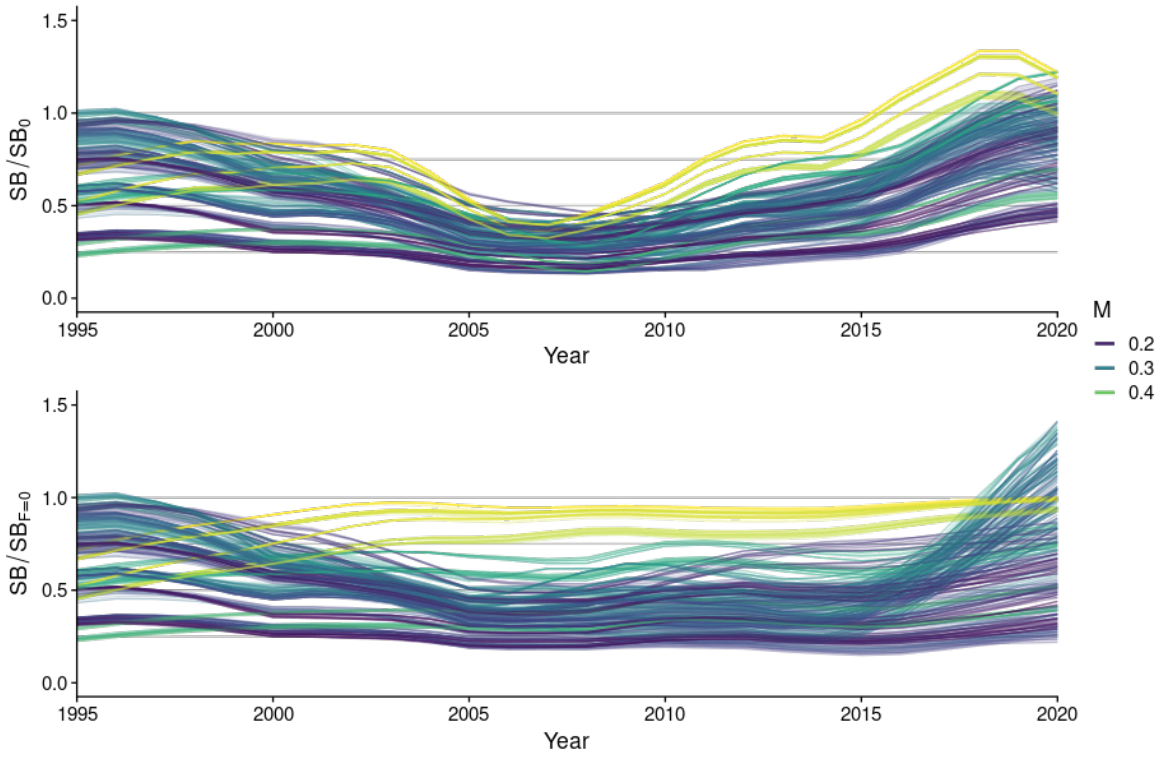


Figure 22: Stock trajectories relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) for 648 model runs in the 2022 structural uncertainty grid for blue shark, coloured by estimated natural mortality (M) estimates.

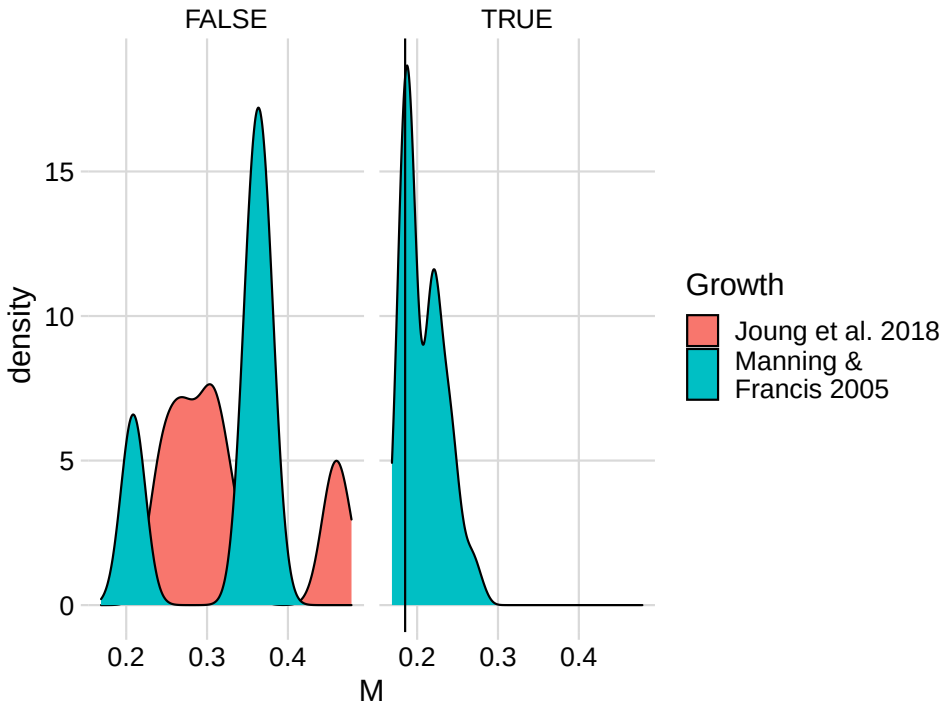


Figure 23: Estimated natural mortality (M) across the 648 models in the 2022 structural uncertainty grid for blue shark, showing bimodal outcomes for retained models. Panel labels indicate if models were retained (TRUE) or removed (FALSE) based on retrospective patterns and consideration for M. 228 models remained after discarding implausible models.

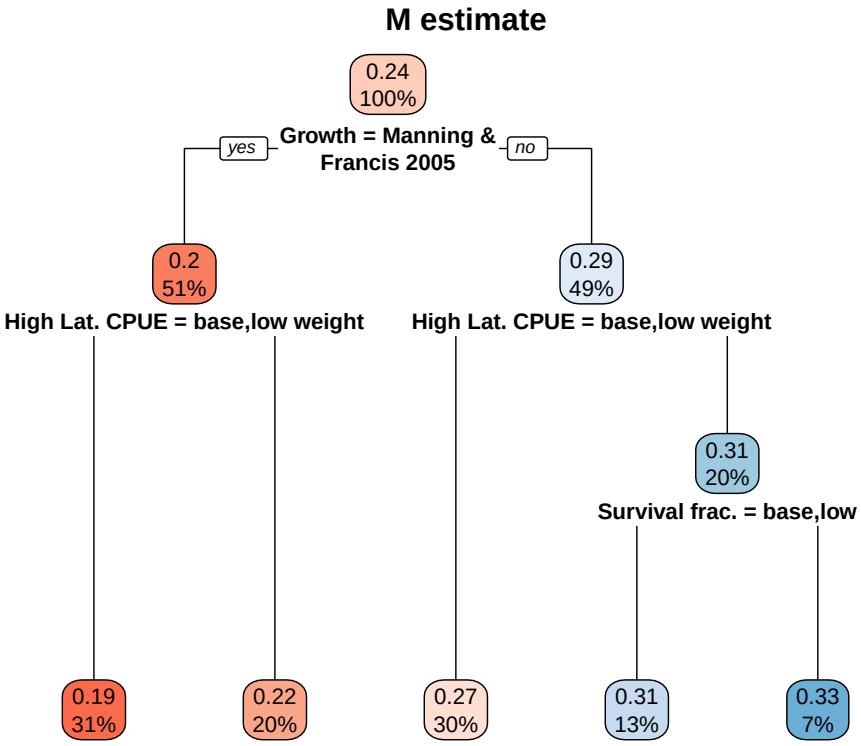


Figure 24: Decision tree for natural mortality estimates across 648 models in the 2022 structural uncertainty grid for blue shark: positive ('yes') values for each split are on the left, leafs on the decision tree show the mean value of M by leaf, as well as the percentage of records on that leaf.

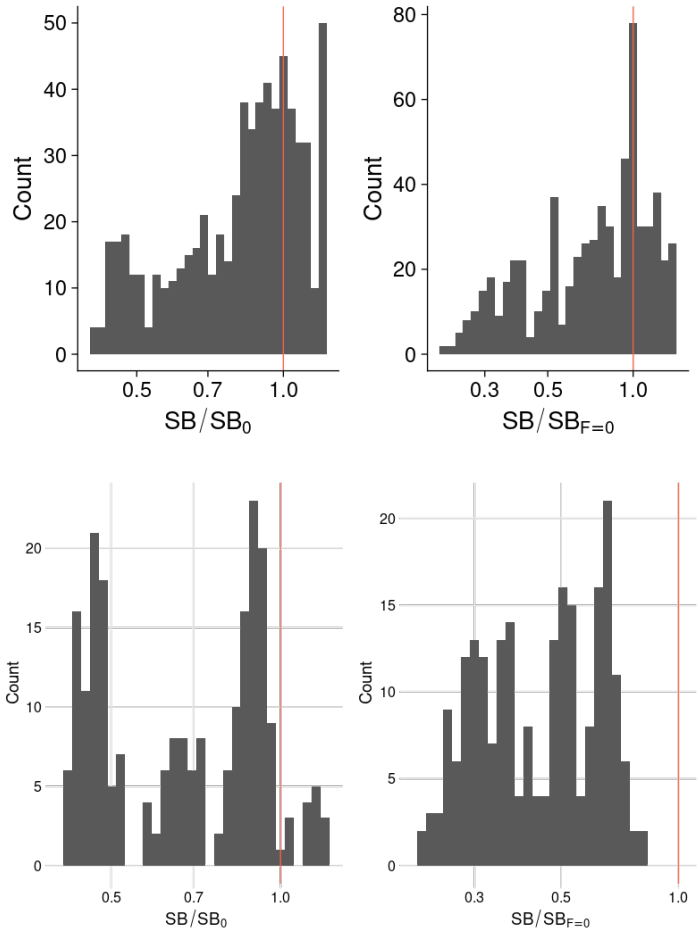


Figure 25: Estimated depletion level as SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$ for 648 models in the 2022 structural uncertainty grid for blue shark before (top), and after discarding model runs based on retrospective patterns and consideration for M (but before weighting model axes). 228 models remained after discarding implausible models.

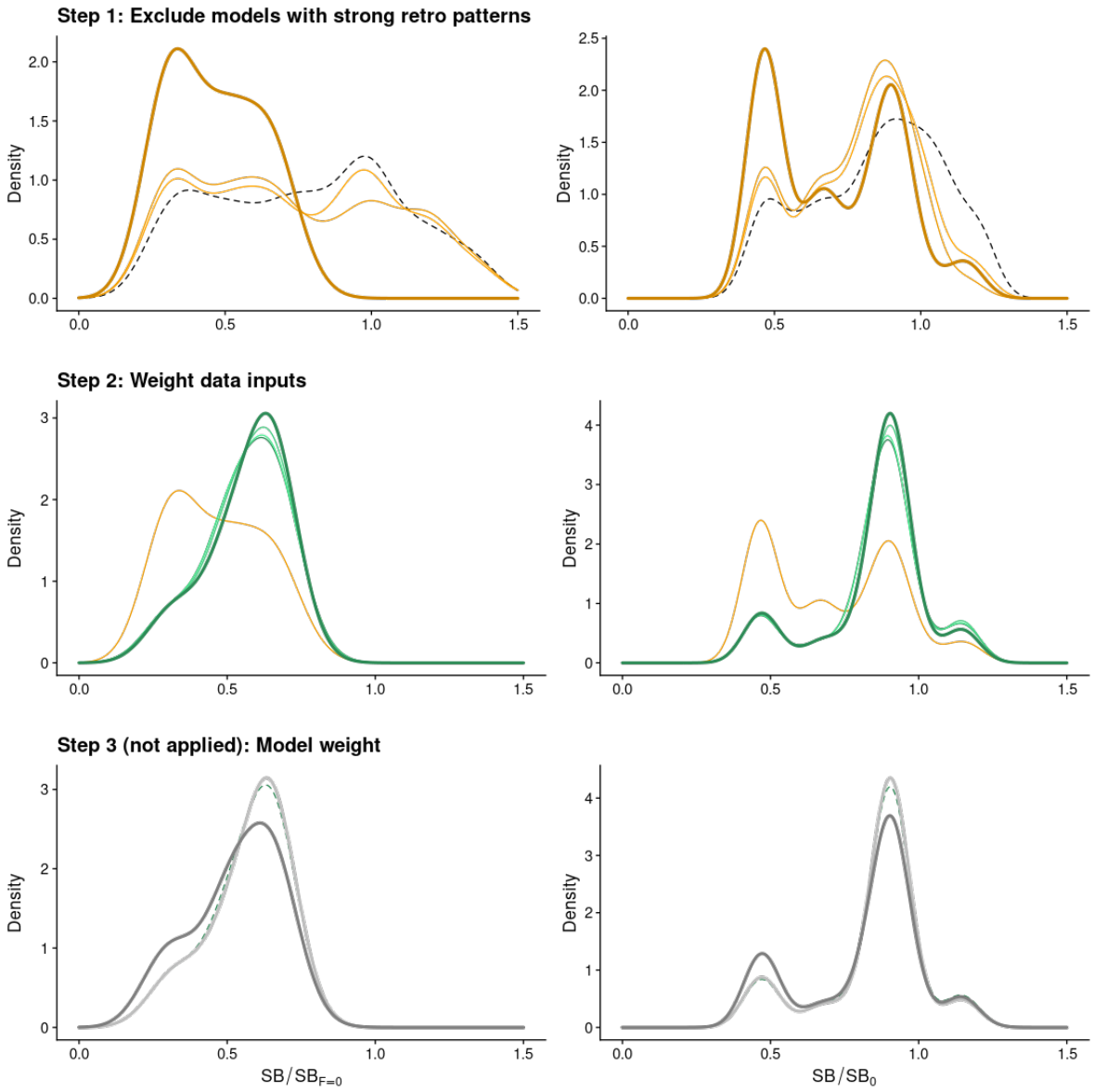


Figure 26: Three step procedure to constrain and weight the 2022 structural uncertainty grid for blue shark. In Step 1 (top panels), the initial set of 648 models shown as dashed density of spawning biomass (SB) relative to reference points (SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) is subset by excluding models with strong retrospective patterns in spawning biomass (thin orange line) and fishing mortality (thick orange line). This subset of (440) models is then weighted (Step 2; second row) according to *a priori* weights for grid axes for catch/discard assumptions (thin green line), which down-weights assumptions of high catch, leading to more constrained distribution of outcomes. CPUE index assumptions only adjust this ensemble in a minor way (thick green line). Model weights from predictive model checks were not applied for further reporting but are illustrated in step 3 (bottom row; MASE - light grey, inverse variance weighting - mid-grey, stacking weights - dark grey) relative to the *a priori* weighted ensemble (green dashed line).

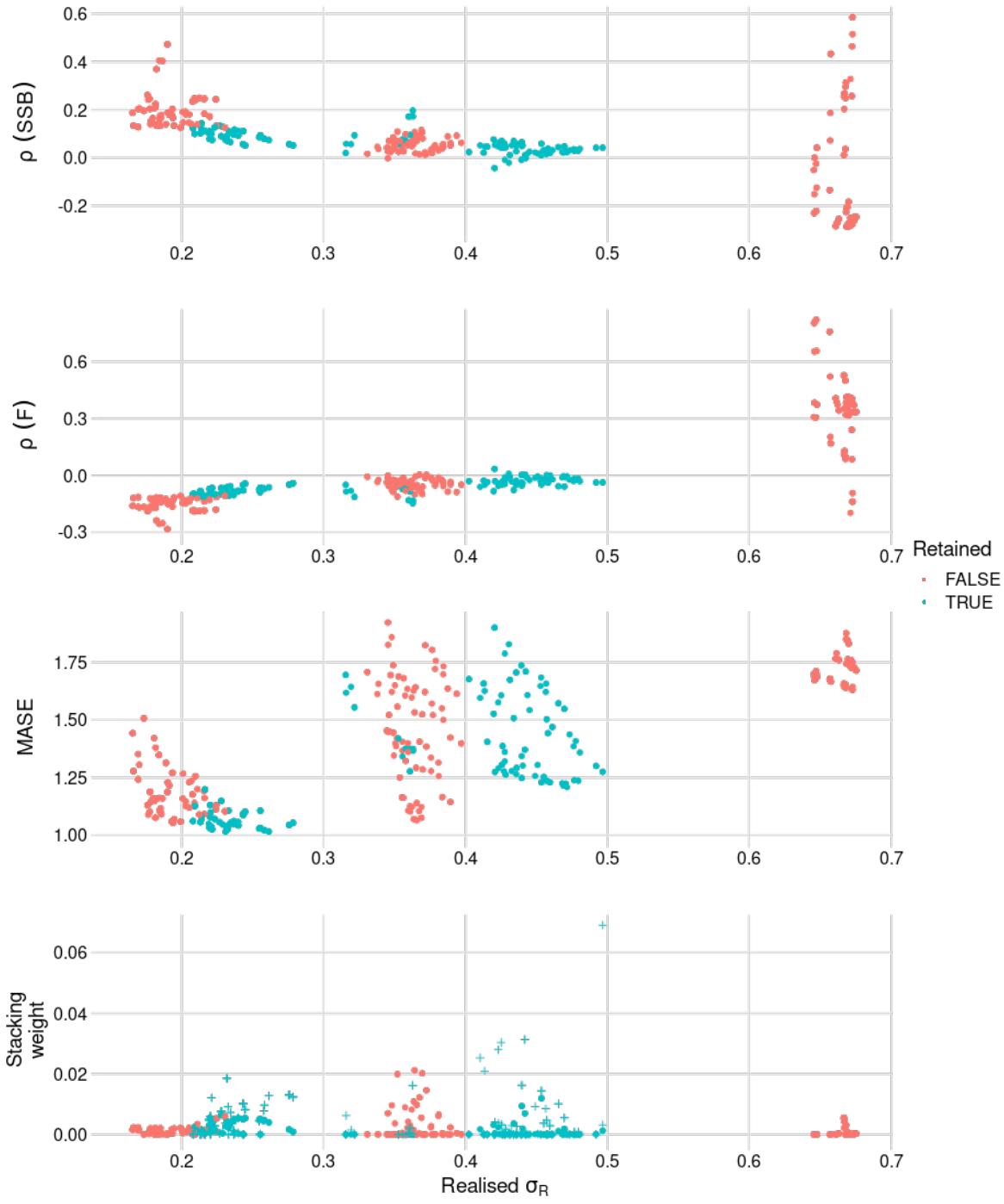


Figure 27: Mohn's ρ for spawning biomass and fishing mortality, MASE and stacking weights relative to the standard deviation of estimated recruitment deviations (realised $sigma_R$). For stacking weights, points show stacking weights before applying filtering with respect to ρ and M; crosses show re-calculated stacking weights post-filtering. Colours indicate if models were retained (TRUE) or removed (FALSE) based on retrospective patterns and consideration for M.

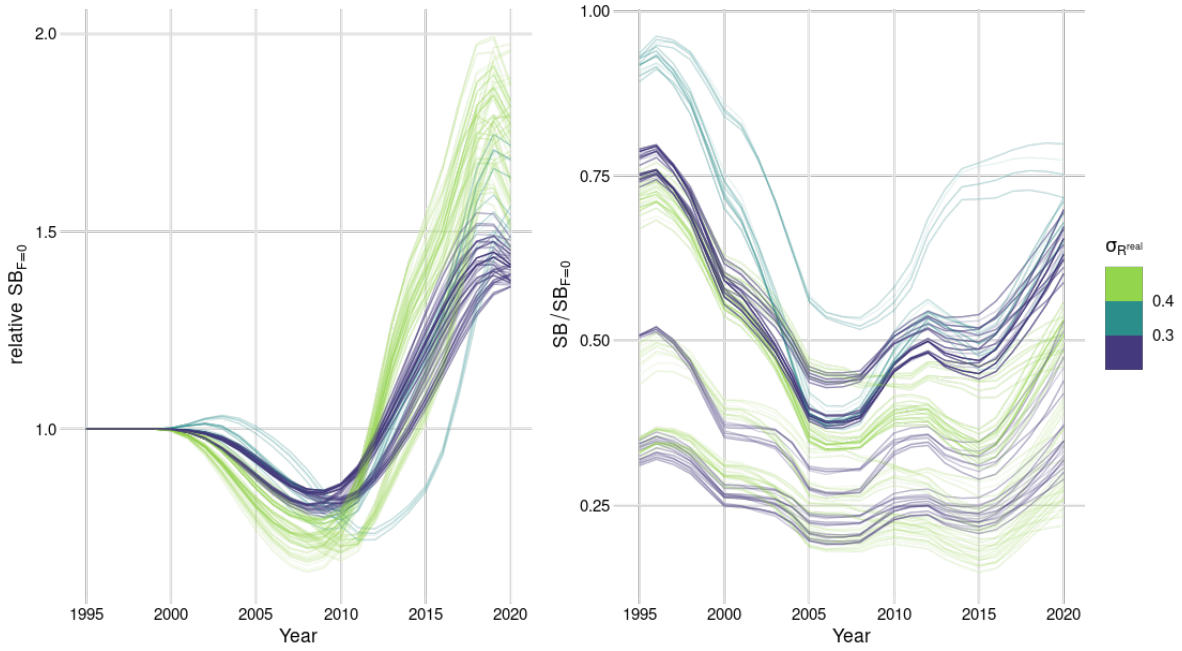


Figure 28: Evolution of dynamic B_0 ($SB_{F=0}$) as a function of realised $sigma_R$ for 228 models retained in the 2022 structural uncertainty grid for blue shark, with shading indicating model weight. Corresponding trends in $SB_{latest}/SB_{F=0}$ are given for comparison.

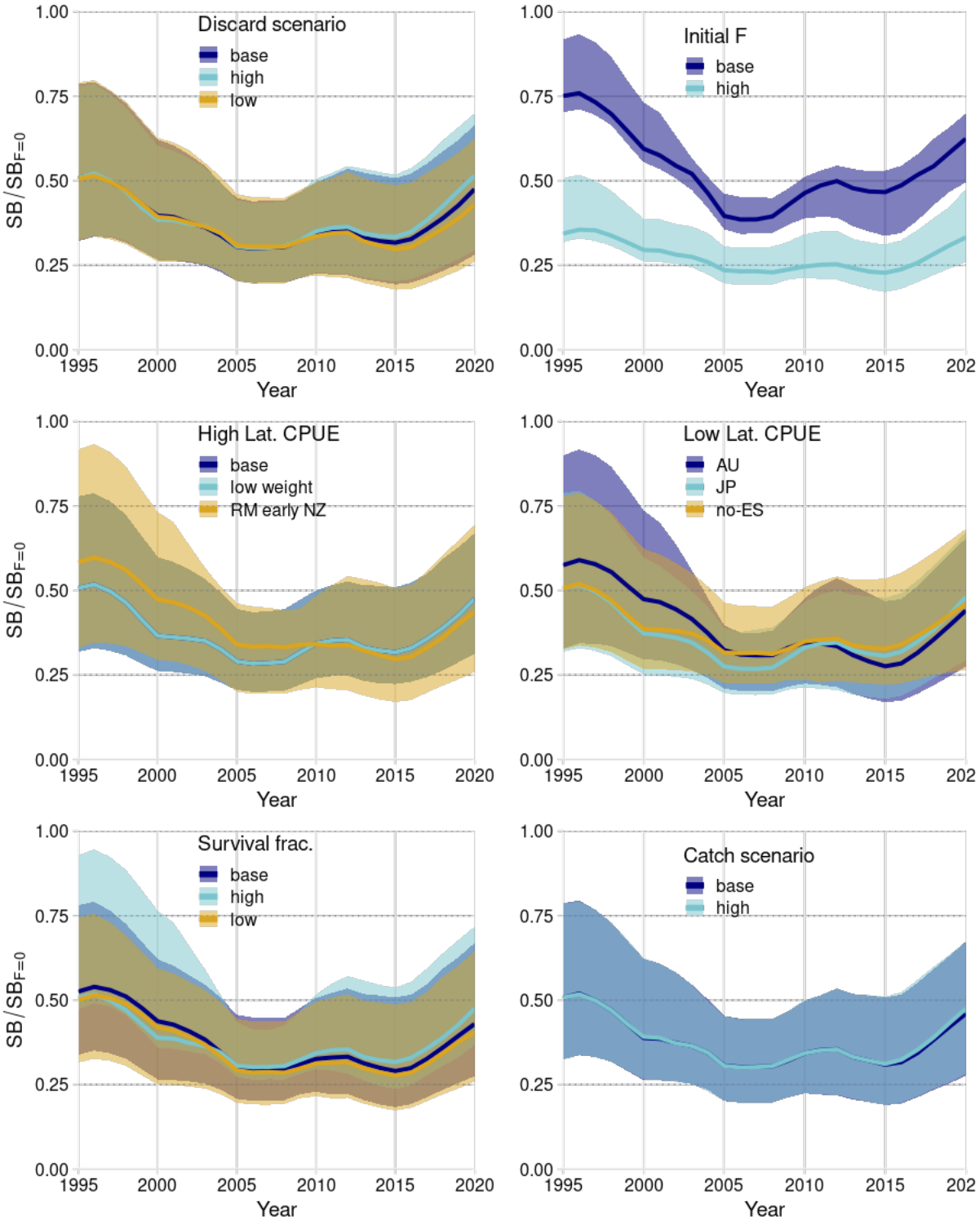


Figure 29: Median and inter-quartile bounds for depletion in spawning biomass for each structural uncertainty axis, colour-code by the level used for each axis and weighted by input model weights across 228 models. The horizontal grey lines are placed at intervals of 25% in the lower part of the graph to aid visualization.

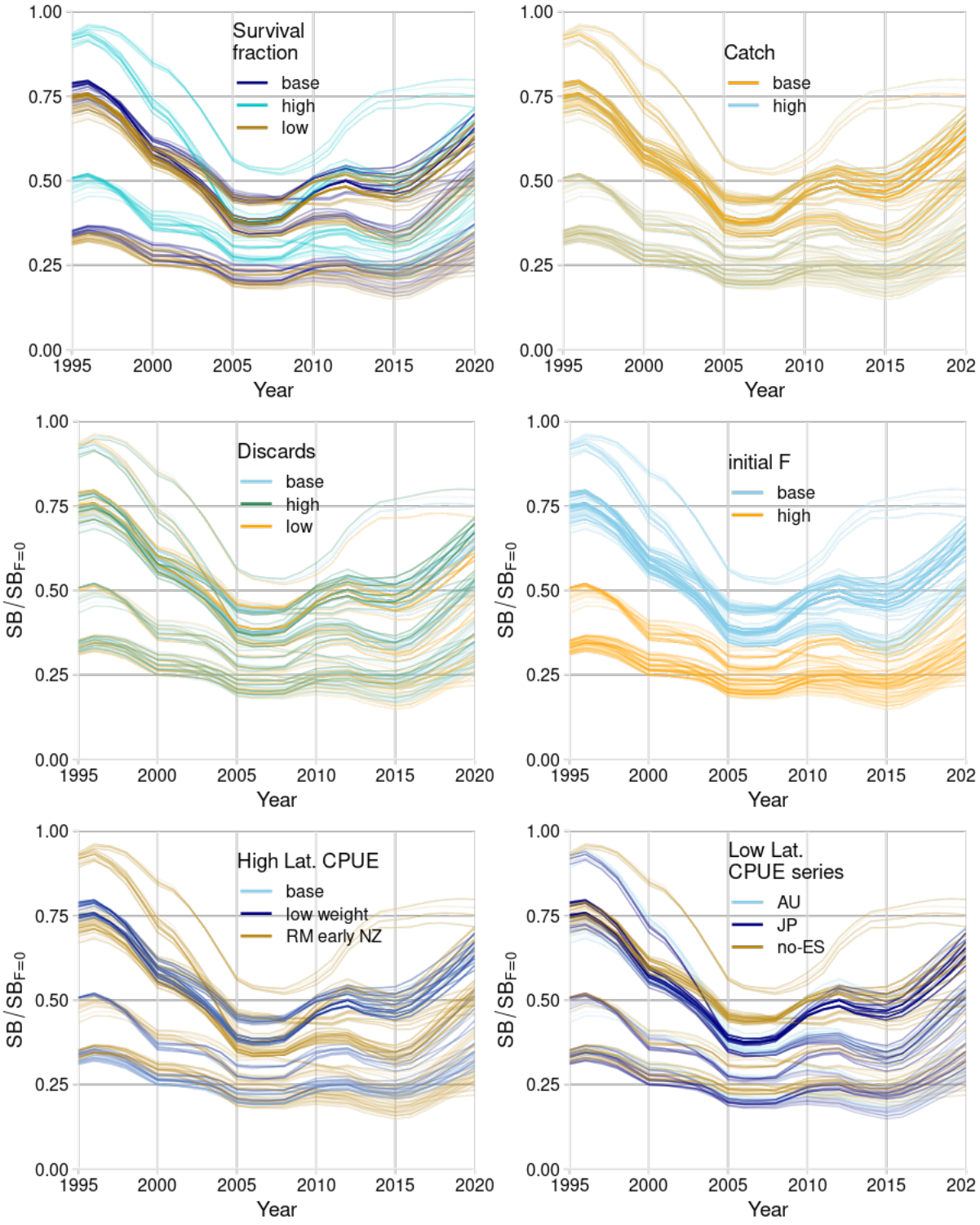


Figure 30: Prediction of depletion in spawning biomass for each structural uncertainty grid run, with one panel for each grid axis highlighting the different levels within. Transparency reflects model weights across 228 models. The horizontal grey lines are placed at intervals of 25% in the lower part of the graph to aid visualization.

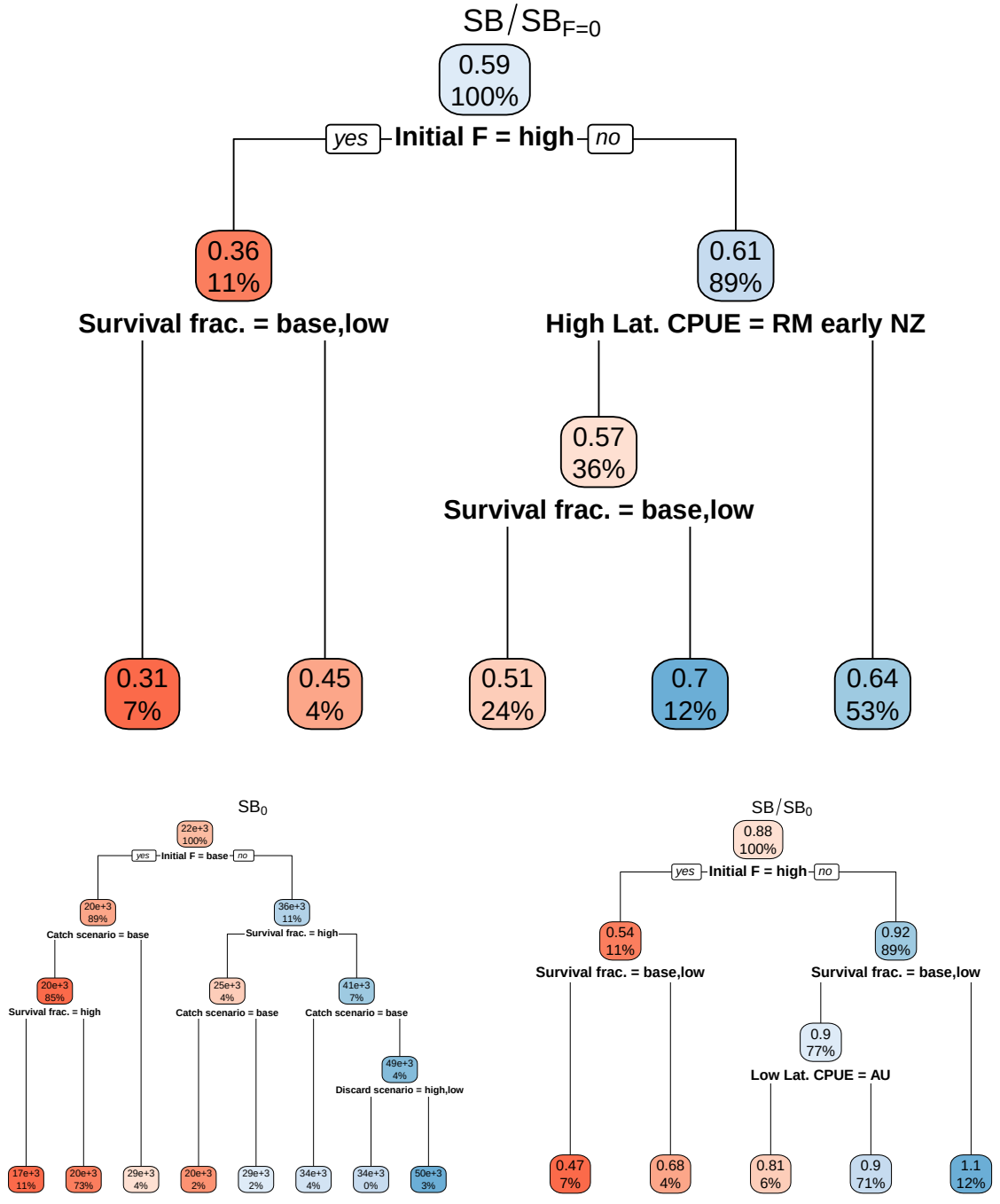


Figure 31: Decision trees weighted by input model weights across 228 models in the structural uncertainty ensemble: positive ('yes') values for each split are on the left, leafs on the decision tree show the mean value (SB_0 or SB_{latest}/SB_0 or $SB_{latest}/SB_{F=0}$) by leaf, as well as the percentage of records on that leaf. (RM=remove).

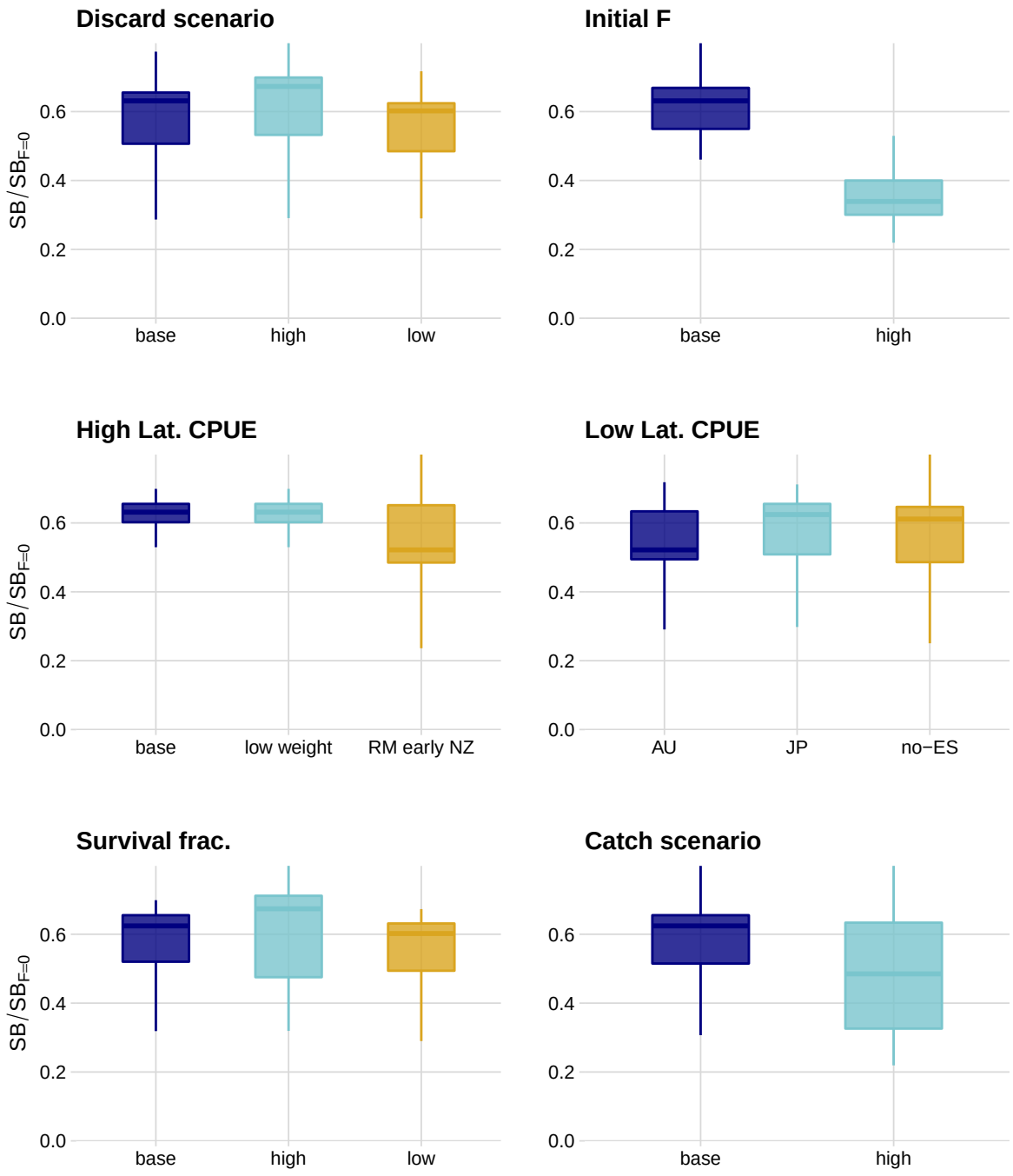


Figure 32: Median (white bar) and inter-quartile bounds (box) for $SB_{latest}/SB_{F=0}$ in the final year of the assessment for each structural uncertainty axis weighted by input model weights across 228 models in the structural uncertainty ensemble. The whiskers extend to $1.5\times$ the interquartile range.

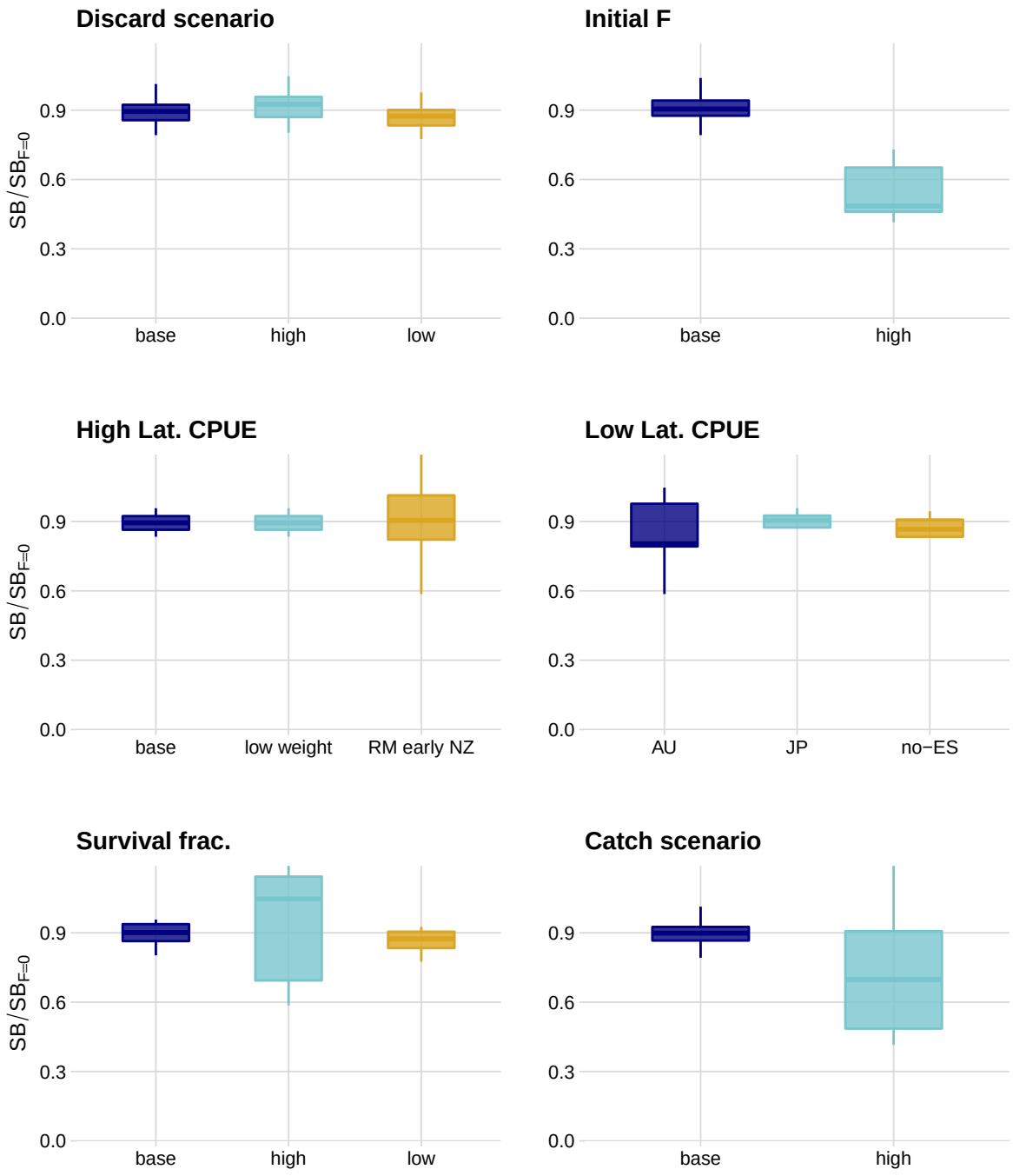


Figure 33: Median (white bar) and inter-quartile bounds (box) for SB_{latest}/SB_0 in the final year of the assessment for each structural uncertainty axis weighted by input model weights across 228 models in the structural uncertainty ensemble. The whiskers extend to 1.5× the interquartile range.

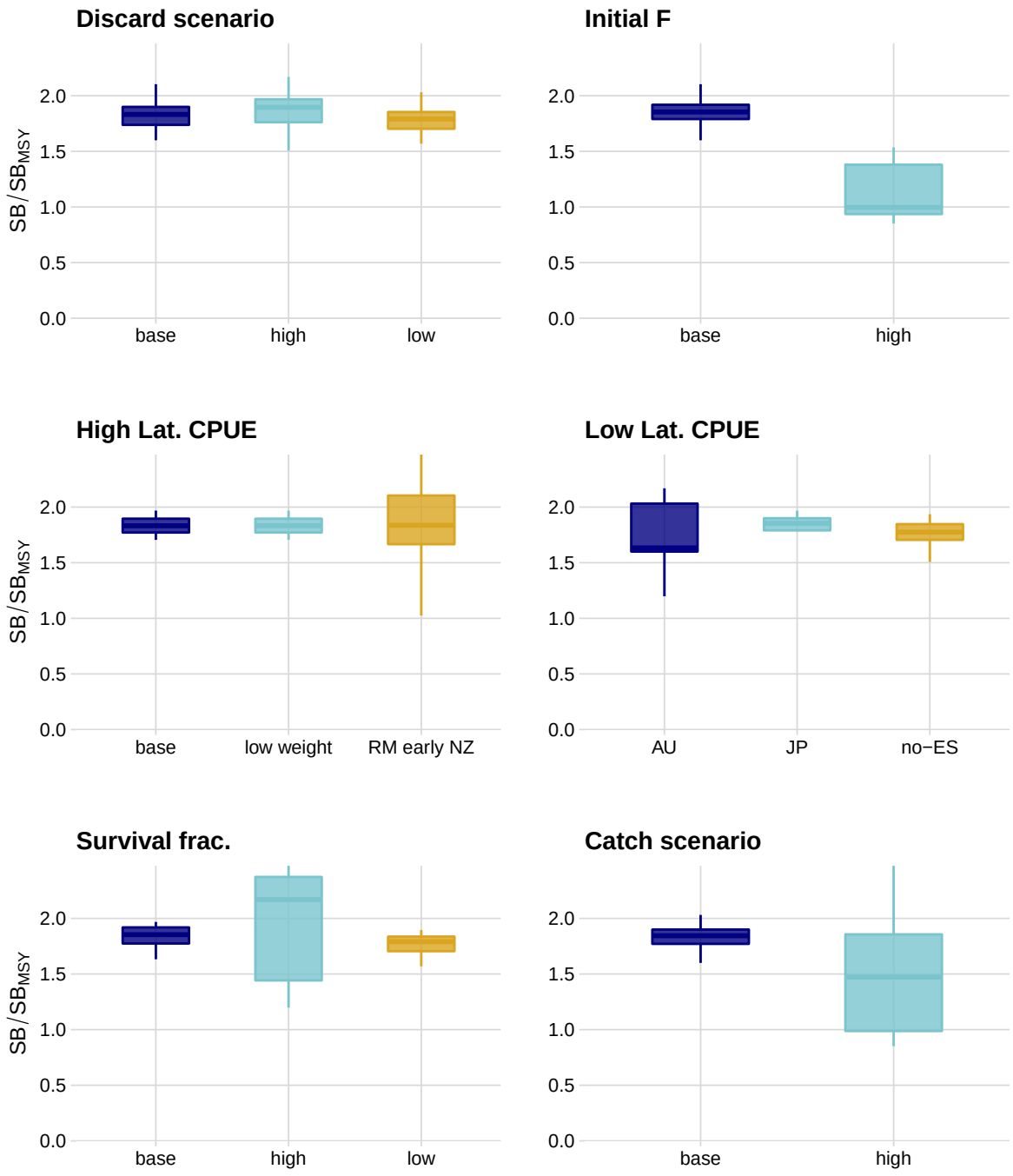
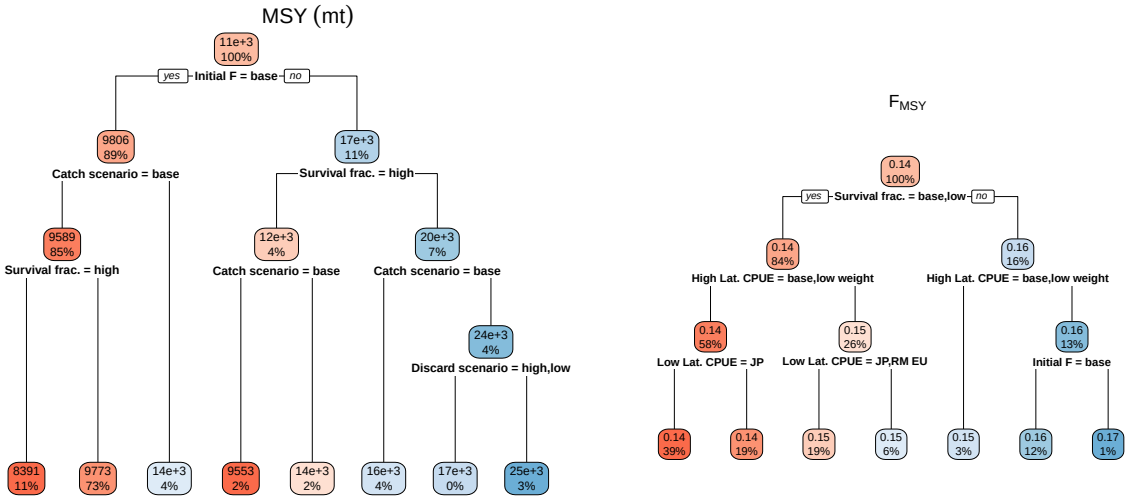


Figure 34: Median (white bar) and inter-quartile bounds (box) for $SB_{latest}/SB_{\text{MSY}}$ in the final year of the assessment for each structural uncertainty axis weighted by input model weights across 228 models in the structural uncertainty ensemble. The whiskers extend to $1.5\times$ the interquartile range.



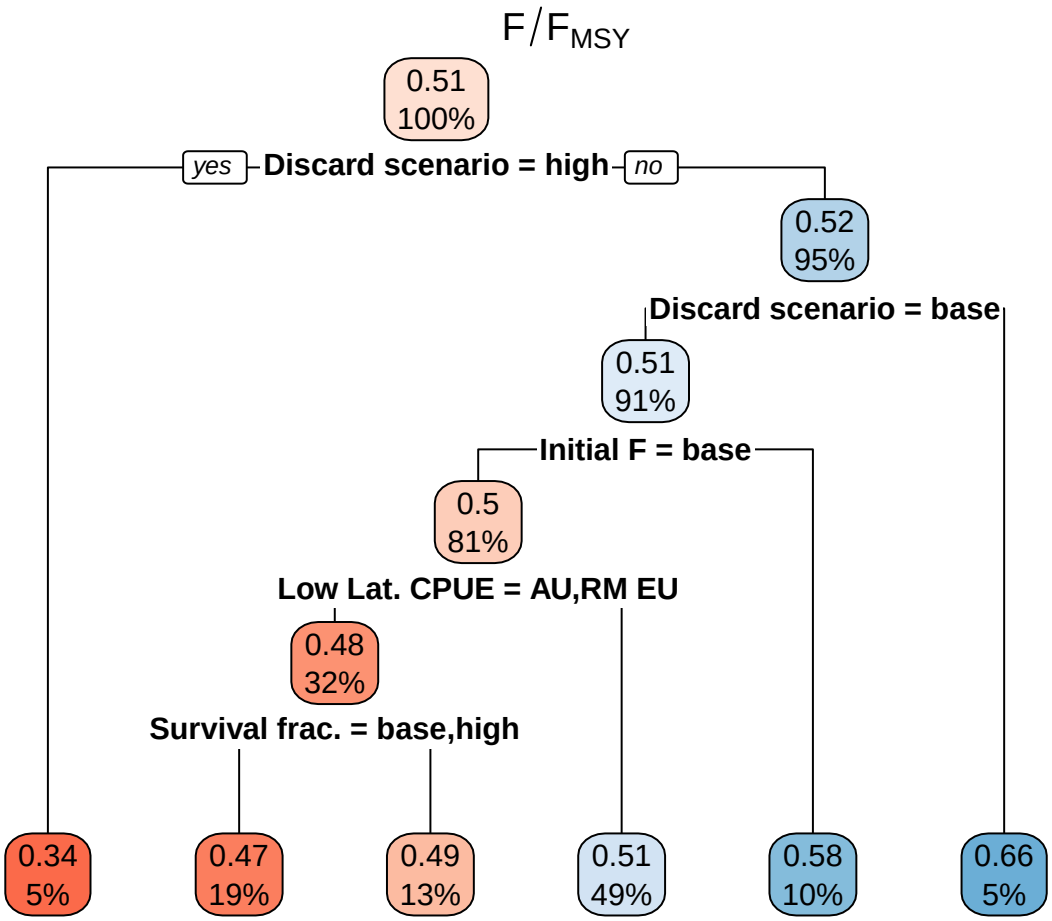


Figure 35: Decision trees weighted by input model weights across 228 models in the structural uncertainty ensemble: positive ('yes') values for each split are on the left, leafs on the decision tree show the mean value (MSY, F_{MSY} and $F_{latest}/F_{\text{MSY}}$) by leaf, as well as the percentage of records on that leaf. Note that for $F_{latest}/F_{\text{MSY}}$ and MSY, 'yes' for Australia or Japan low latitude CPUE also includes EU - Spain; similarly, a 'base' or 'low weight' scenario for high latitude CPUE (New Zealand / EU - Spain) also includes early CPUE from New Zealand (RM=remove).

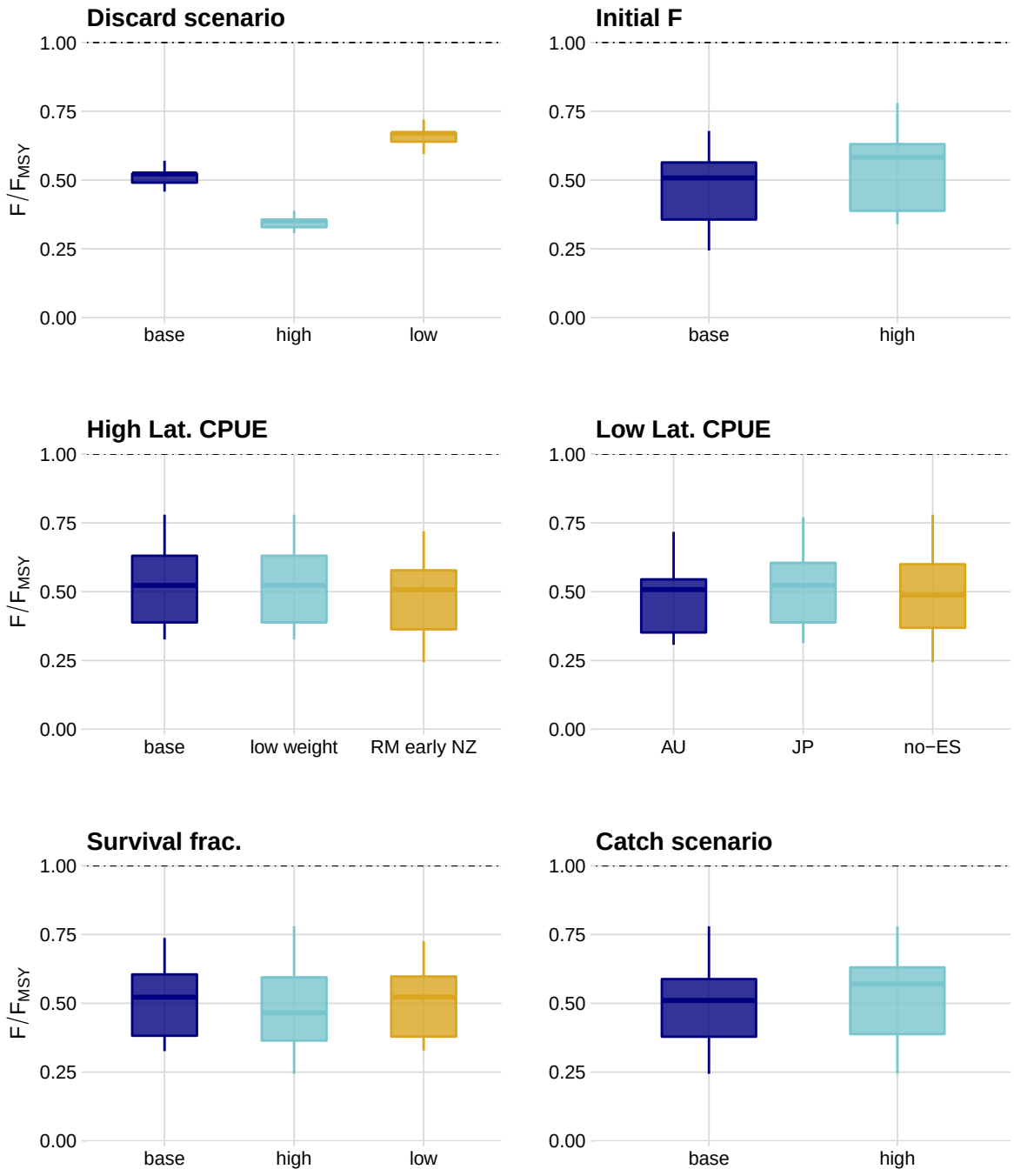


Figure 36: Median (white bar) and inter-quartile bounds (box) for $F_{latest}/F_{\text{MSY}}$ in the final year of the assessment for each structural uncertainty axis weighted by input model weights across 228 models in the structural uncertainty ensemble. The whiskers extend to 1.5× the interquartile range. The dashed line shows the level where $F = F_{\text{MSY}}$.

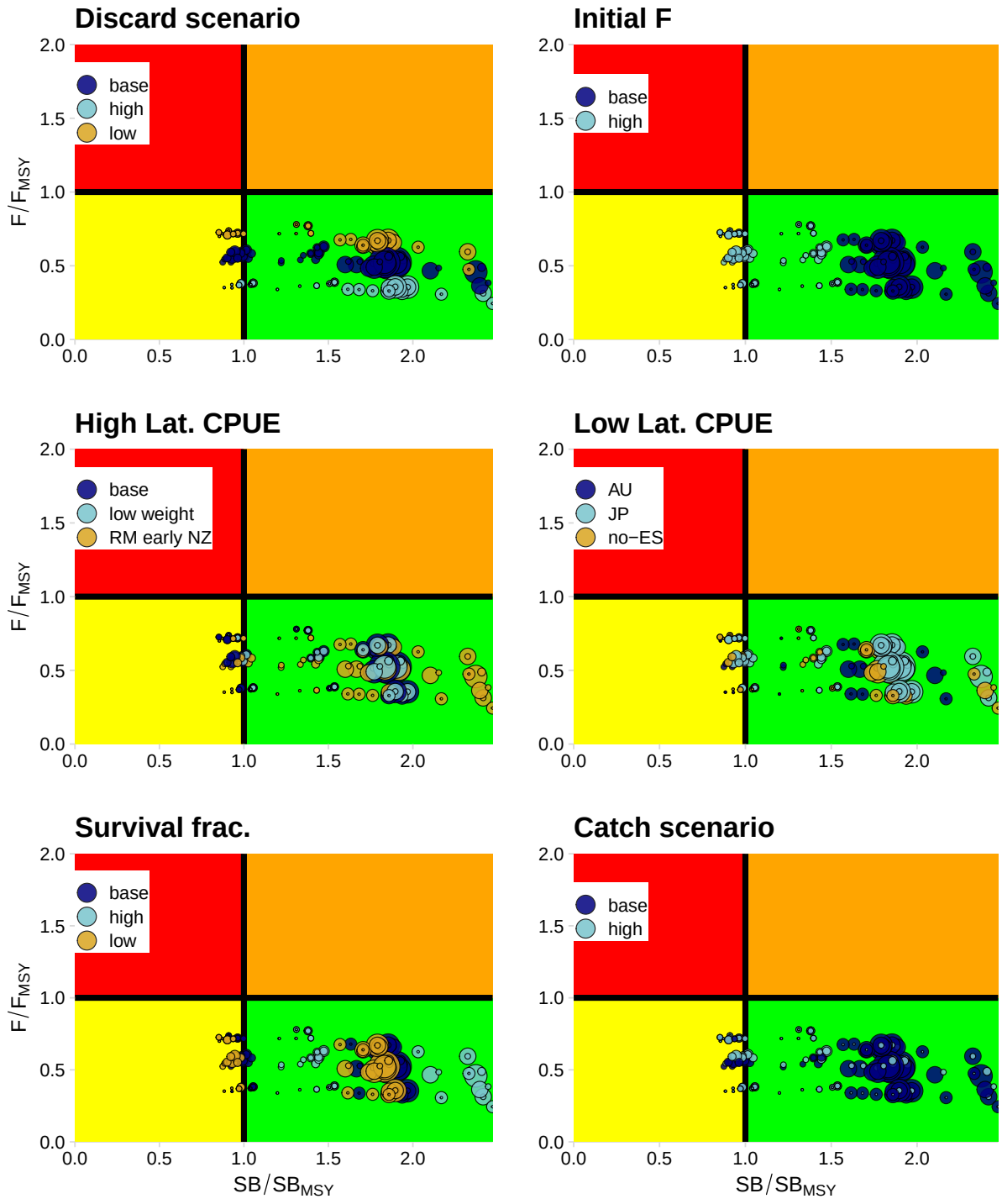


Figure 37: Kobe plots summarising status in the final year for each of the models weighted by input model weights (point size) across 228 models in the structural uncertainty ensemble, based on $SB_{latest}/SB_{\text{MSY}}$ and $F_{latest}/F_{\text{MSY}}$. The stock is considered to be overfished when $SB_{latest}/SB_{\text{MSY}} < 1$ and undergoing overfishing when $F_{latest}/F_{\text{MSY}} > 1$.

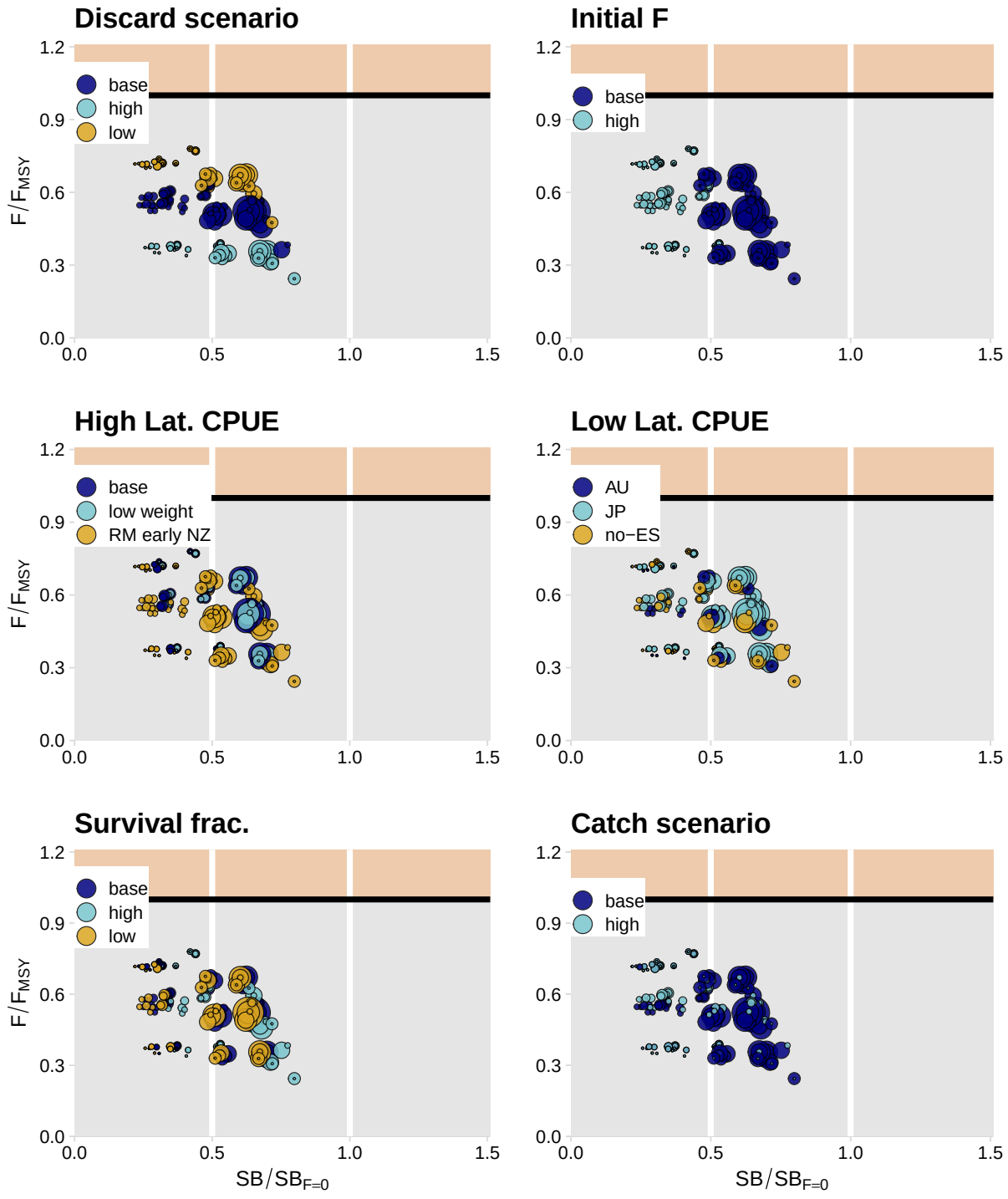


Figure 38: Panel plot summarising stock status in the final year for each of the models weighted by input model weights (point size) across 228 models in the structural uncertainty ensemble for $SB_{latest}/SB_{F=0}$ and $F_{latest}/F_{\text{MSY}}$. The stock is considered to be undergoing overfishing when $F_{latest}/F_{\text{MSY}} > 1$ (beige zone). Guidelines were added in white at $SB_{latest}/SB_{F=0} = 0.5$ and $SB_{latest} = SB_{F=0}$.

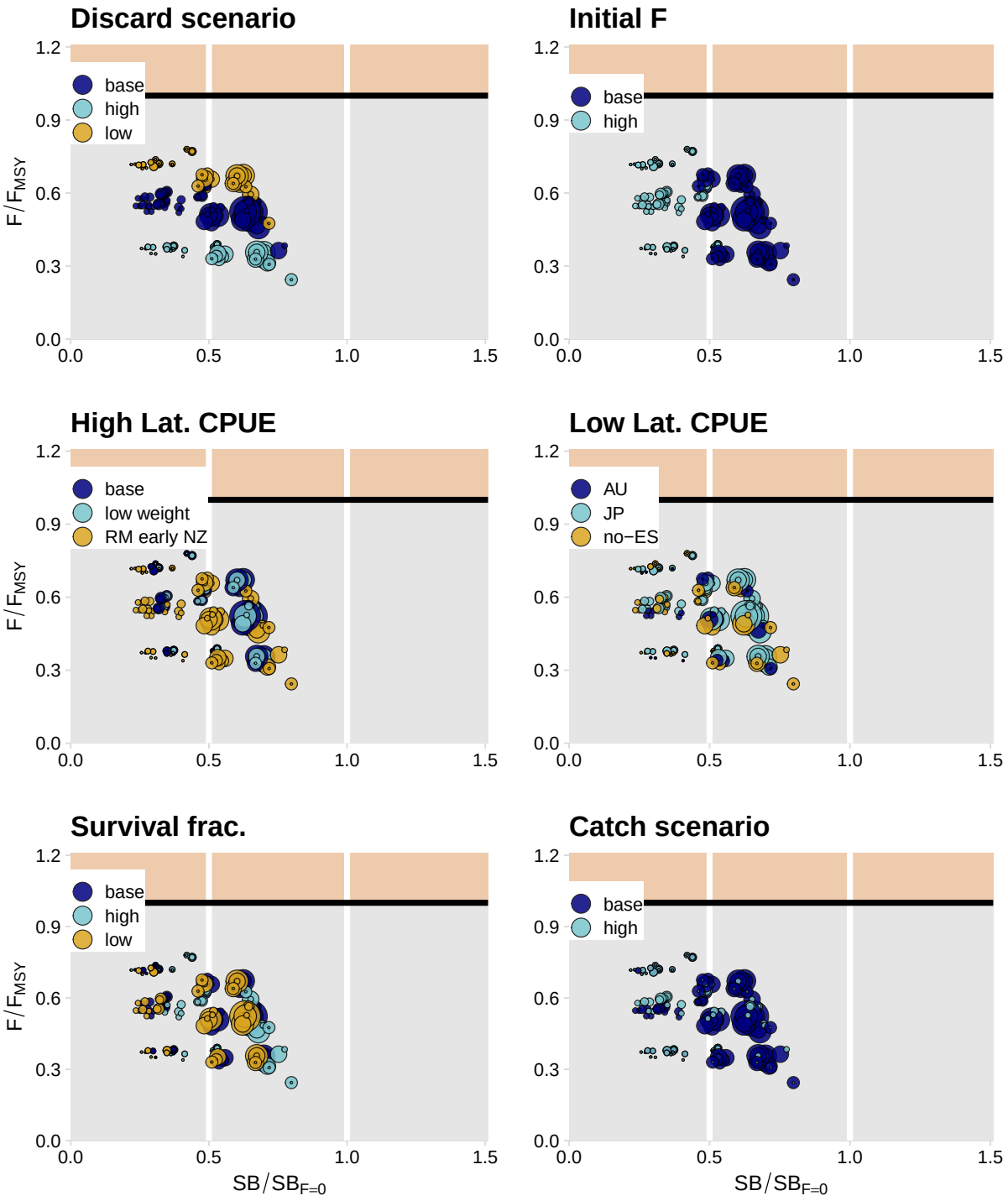


Figure 39: Panel plot summarising stock status in the final year for each of the models weighted by input model weights (point size) across 228 models in the structural uncertainty ensemble for SB_{latest}/SB_0 and $F_{latest}/F_{\text{MSY}}$. The stock is considered to be undergoing overfishing when $F_{latest}/F_{\text{MSY}} > 1$ (beige zone). Guidelines were added in white at $SB_{latest}/SB_{F=0} = 0.5$ and $SB_{latest} = SB_{F=0}$.

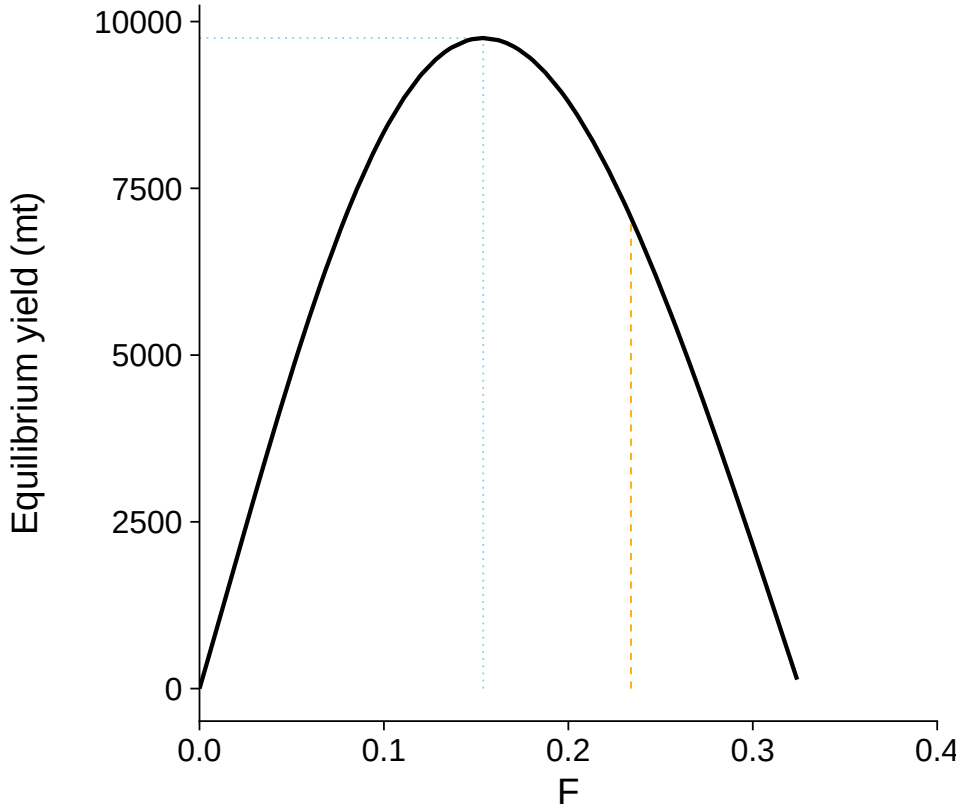


Figure 40: Yield profiles for Southwest Pacific blue shark for different model assumptions, with F_{MSY} indicated by dotted vertical lines, and $F_{lim,\text{AS}}$ shown as dashed lines.

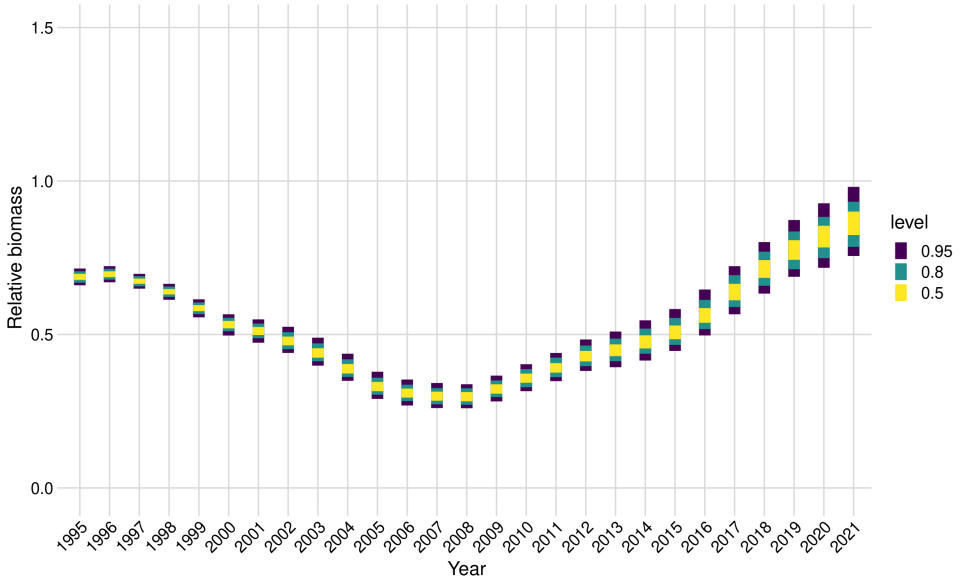


Figure 41: Estimation uncertainty for the updated 2022 diagnostic case, derived using 1000 samples from the posterior distribution of SB/SB_0 using No-U-Turn sampling implemented in the ADNUTS R package (Monnahan et al. 2019).

APPENDIX A: Supplementary figures

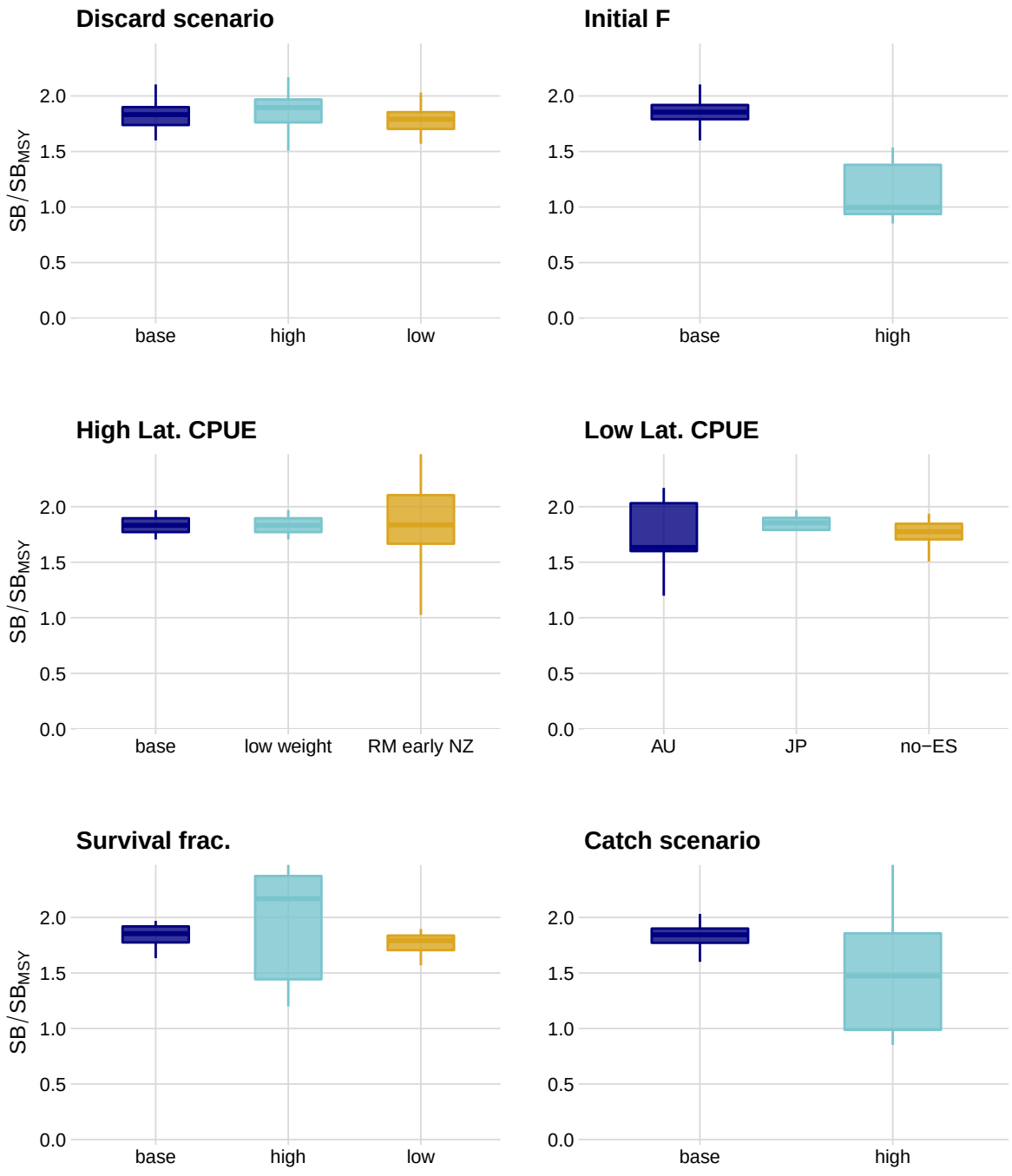


Figure A-1: Median (white bar) and inter-quartile bounds (box) for $SB_{latest}/SB_{\text{MSY}}$ in the final year of the assessment for each structural uncertainty axis. The whiskers extend to 1.5× the interquartile range.

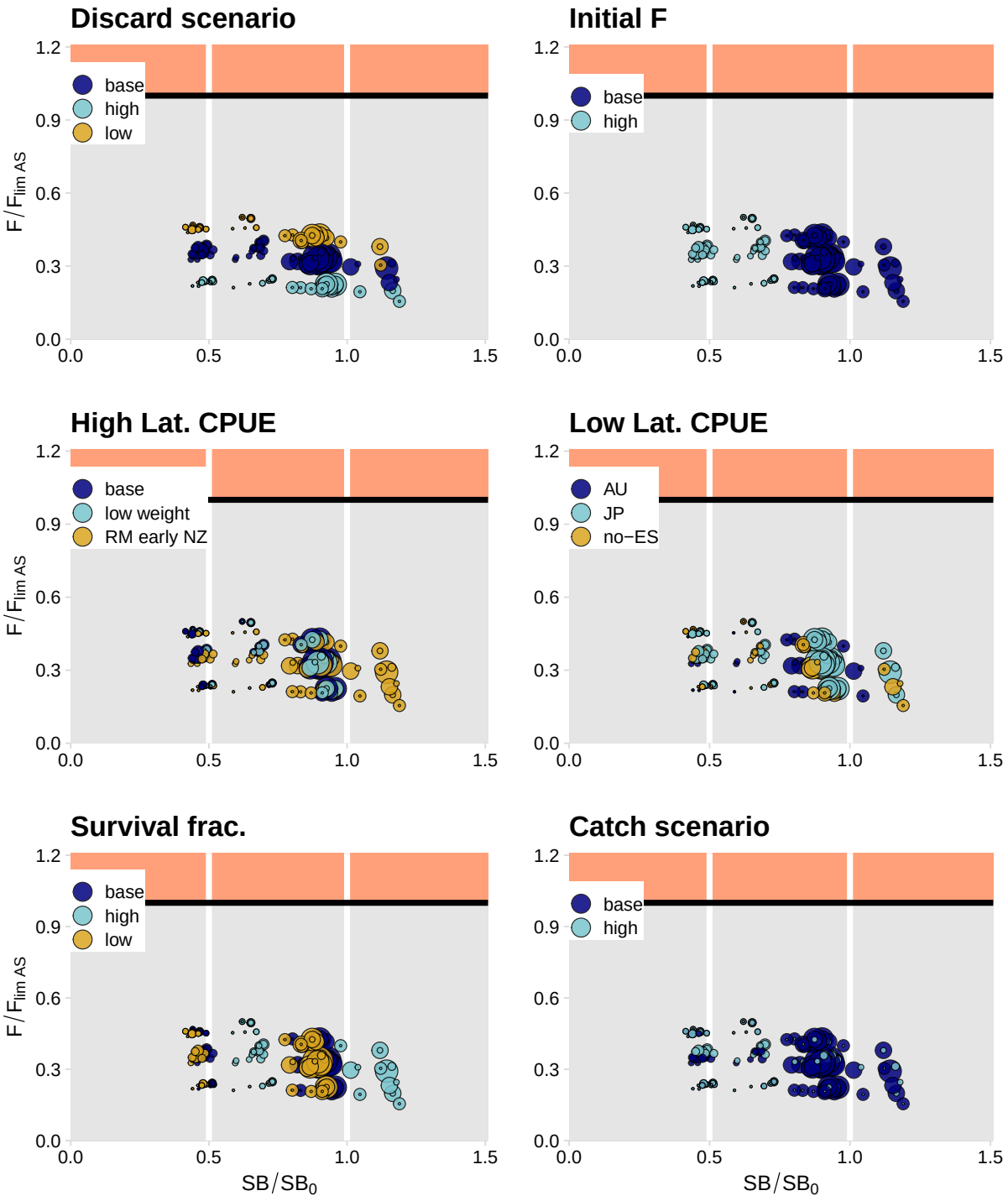


Figure A-2: Panel plot summarising stock status in the final year for each of the models in the structural uncertainty grid for SB/SB_0 and $F/F_{lim,\text{AS}}$. When $F/F_{lim,\text{AS}} > 1$ (orange zone), the spawning biomass has declined below $0.5SB_{MSY}$. Guidelines were added in white at $SB_{latest}/SB_{F=0} = 0.5$ and $SB_{latest} = SB_{F=0}$.

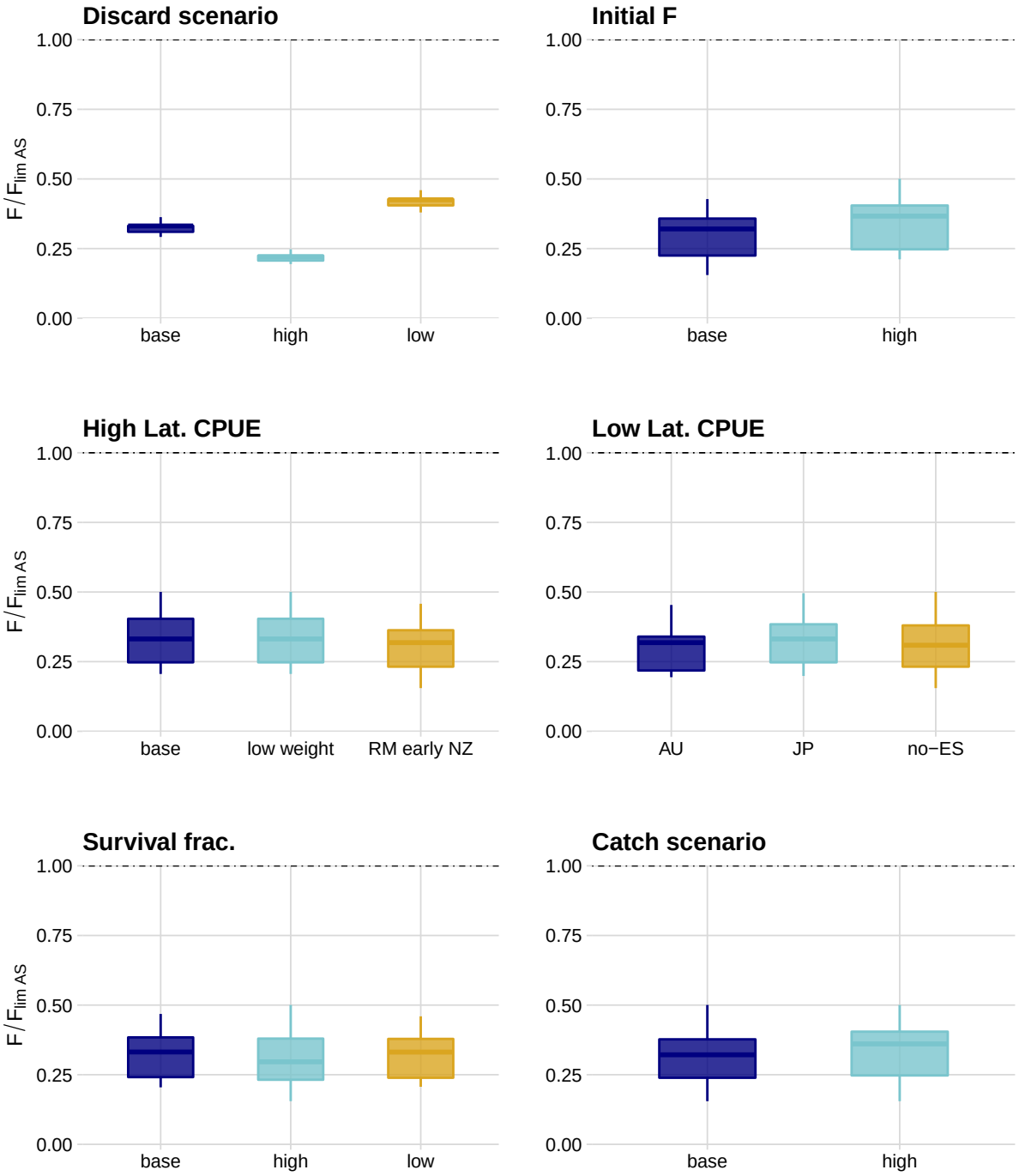


Figure A-3: Median (white bar) and inter-quartile bounds (box) for F/F_{lim} in the final year of the assessment for each structural uncertainty axis. The whiskers extend to 1.5× the interquartile range. The dashed line shows the level where $F = F_{lim}$.

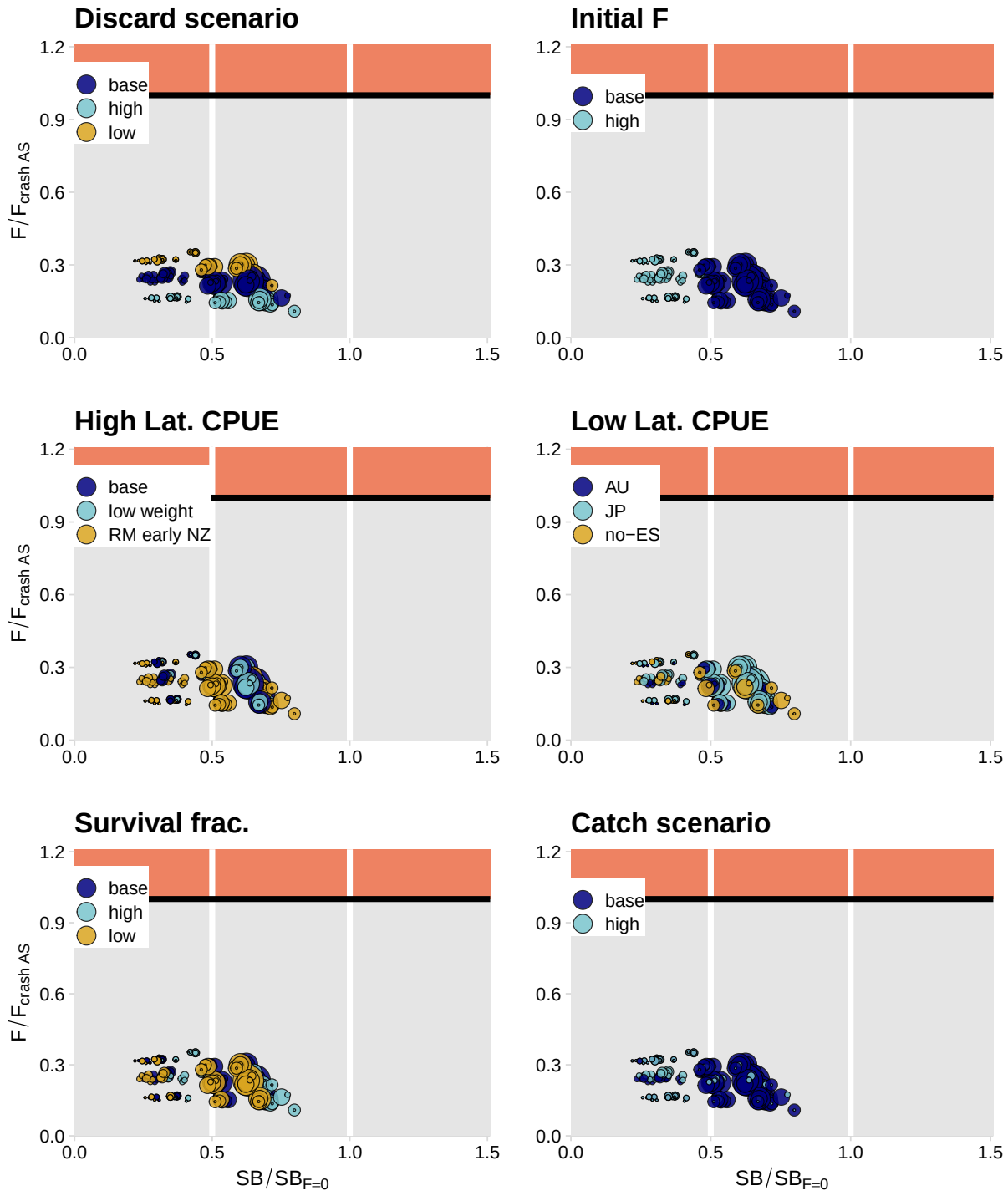


Figure A-4: Panel plot summarising stock status in the final year for each of the models in the structural uncertainty grid for SB/SB_0 and $F/F_{crash,\text{AS}}$. The population is expected to become extinct when levels of F in excess of $F_{crash,\text{AS}}$ (i.e. $F/F_{crash,\text{AS}} > 1$; pink zone) are maintained on the long-term. Guidelines were added in white at $SB_{latest}/SB_{F=0} = 0.5$ and $SB_{latest} = SB_{F=0}$.

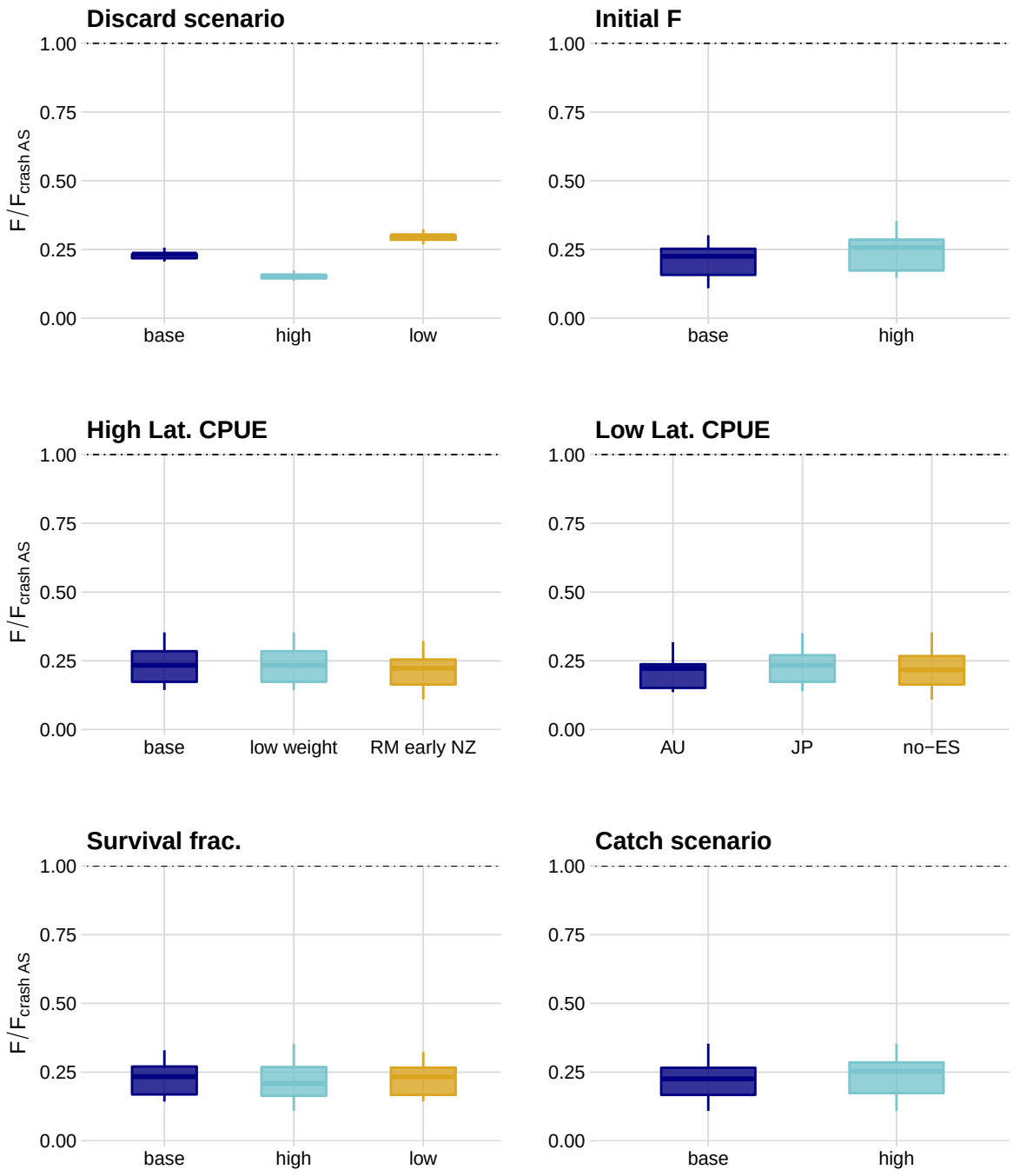


Figure A-5: Median (white bar) and inter-quartile bounds (box) for F/F_{crash} in the final year of the assessment for each structural uncertainty axis. The whiskers extend to 1.5× the interquartile range. The dashed line shows the level where $F = F_{crash}$.

APPENDIX B: Supplementary tables

Table B-1: Summary of reference points for the subset of 114 grid models in the structural uncertainty grid using the median catch scenario and median discards estimates.

	Mean	Median	Min	10%	90%	Max
C_{latest}	5685	5657	3707	4000	7588	7776
C_{recent}	6698	6744	4322	4600	8850	8926
MSY	10465	9868	8968	9313	12779	18737
SB_0	20967	20114	15686	18468	22828	38957
$SB_{F=0}$	23769	22658	17559	19873	27743	48618
SB_{MSY}	10227	9847	7564	9008	11162	18974
SB_{latest}	17919	17800	12973	16193	18903	24706
SB_{recent}	15917	15907	11320	14244	17026	21915
SB_{latest}/SB_0	0.88	0.90	0.42	0.68	1.01	1.19
SB_{recent}/SB_0	0.78	0.80	0.37	0.59	0.90	1.05
$SB_{latest}/SB_{F=0}$	0.78	0.81	0.32	0.59	0.93	1.29
$SB_{recent}/SB_{F=0}$	0.70	0.71	0.29	0.51	0.82	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.81	1.84	0.85	1.44	2.10	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.60	1.64	0.76	1.25	1.87	2.19
F_{MSY}	0.144	0.142	0.134	0.136	0.158	0.181
$F_{lim,\text{AS}}$	0.228	0.225	0.211	0.215	0.248	0.291
$F_{crash,\text{AS}}$	0.324	0.320	0.299	0.306	0.351	0.419
F_{latest}	0.072	0.072	0.039	0.051	0.091	0.120
F_{recent}	0.093	0.094	0.048	0.065	0.117	0.160
$F_{latest}/F_{\text{MSY}}$	0.50	0.51	0.24	0.34	0.67	0.78
$F_{recent}/F_{\text{MSY}}$	0.65	0.65	0.30	0.43	0.86	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.32	0.32	0.15	0.21	0.43	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.41	0.41	0.19	0.27	0.55	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.22	0.23	0.11	0.15	0.30	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.29	0.29	0.13	0.19	0.39	0.48

Table B-2: Summary of reference points for the subset of 114 grid models in the structural uncertainty grid using the high catch scenario and median discards estimates.

	Mean	Median	Min	10%	90%	Max
C_{latest}	7345	8134	3707	3899	9285	9601
C_{recent}	7961	8923	4322	4574	9475	9577
MSY	16067	14970	8968	9868	24337	25629
SB_0	31643	30135	15686	19346	49454	53503
$SB_{F=0}$	36333	33467	17559	21166	57049	66434
SB_{MSY}	15414	14723	7564	9445	24456	26684
SB_{latest}	20881	20104	13209	15153	27435	38004
SB_{recent}	18440	17453	11626	13396	24316	33654
SB_{latest}/SB_0	0.71	0.70	0.42	0.46	0.96	1.19
SB_{recent}/SB_0	0.63	0.60	0.37	0.41	0.84	1.05
$SB_{latest}/SB_{F=0}$	0.63	0.61	0.32	0.39	0.88	1.29
$SB_{recent}/SB_{F=0}$	0.55	0.53	0.29	0.34	0.78	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.46	1.47	0.85	0.93	1.94	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.29	1.27	0.76	0.82	1.71	2.19
F_{MSY}	0.146	0.144	0.134	0.135	0.159	0.181
$F_{lim,\text{AS}}$	0.230	0.228	0.211	0.213	0.252	0.291
$F_{crash,\text{AS}}$	0.327	0.324	0.299	0.301	0.361	0.419
F_{latest}	0.079	0.081	0.039	0.053	0.099	0.120
F_{recent}	0.096	0.094	0.048	0.066	0.131	0.160
$F_{latest}/F_{\text{MSY}}$	0.55	0.57	0.24	0.36	0.72	0.78
$F_{recent}/F_{\text{MSY}}$	0.67	0.64	0.30	0.46	0.91	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.35	0.36	0.15	0.22	0.45	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.42	0.41	0.19	0.29	0.58	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.24	0.25	0.11	0.16	0.32	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.30	0.29	0.13	0.20	0.41	0.48

Table B-3: Summary of reference points for the subset of 76 grid models in the structural uncertainty grid using the median catch scenario and low discards estimates.

	Mean	Median	Min	10%	90%	Max
C_{latest}	7489	7533	7005	7284	7593	7776
C_{recent}	8762	8850	8234	8517	8926	8926
MSY	11528	10222	9659	9683	16582	18737
SB_0	23078	21624	16920	20114	35247	38957
$SB_{F=0}$	26815	25076	19873	20923	41854	48618
SB_{MSY}	11240	10553	8145	9859	17134	18974
SB_{latest}	18238	18650	13209	15912	19035	24706
SB_{recent}	16277	16740	11626	14055	17127	21915
SB_{latest}/SB_0	0.83	0.87	0.42	0.47	0.98	1.12
SB_{recent}/SB_0	0.74	0.78	0.37	0.42	0.87	0.99
$SB_{latest}/SB_{F=0}$	0.72	0.75	0.32	0.39	0.84	1.18
$SB_{recent}/SB_{F=0}$	0.64	0.68	0.29	0.35	0.75	1.05
$SB_{latest}/SB_{\text{MSY}}$	1.70	1.79	0.85	0.96	2.03	2.33
$SB_{recent}/SB_{\text{MSY}}$	1.51	1.59	0.76	0.86	1.81	2.07
F_{MSY}	0.141	0.138	0.134	0.134	0.155	0.167
$F_{lim,\text{AS}}$	0.222	0.218	0.211	0.211	0.243	0.265
$F_{crash,\text{AS}}$	0.315	0.309	0.299	0.299	0.343	0.378
F_{latest}	0.093	0.091	0.073	0.087	0.104	0.120
F_{recent}	0.119	0.115	0.089	0.114	0.132	0.160
$F_{latest}/F_{\text{MSY}}$	0.66	0.67	0.48	0.62	0.72	0.78
$F_{recent}/F_{\text{MSY}}$	0.85	0.86	0.58	0.77	0.96	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.42	0.43	0.30	0.39	0.46	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.54	0.55	0.37	0.49	0.61	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.30	0.30	0.22	0.28	0.32	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.38	0.39	0.26	0.34	0.43	0.48

Table B-4: Summary of reference points for the subset of 76 grid models in the structural uncertainty grid using the median catch scenario and high discards estimates.

	Mean	Median	Min	10%	90%	Max
C_{latest}	3977	4004	3707	3877	4019	4159
C_{recent}	4626	4673	4322	4495	4706	4715
MSY	10552	9618	8968	8985	15238	16518
SB_0	20906	19346	15686	18468	30718	34974
$SB_{F=0}$	23170	21356	17559	19088	34680	42183
SB_{MSY}	10205	9445	7564	8933	14904	17450
SB_{latest}	17641	17731	13704	15634	18752	24682
SB_{recent}	15576	15672	11840	13720	16643	21870
SB_{latest}/SB_0	0.88	0.93	0.44	0.51	1.05	1.19
SB_{recent}/SB_0	0.78	0.82	0.39	0.45	0.93	1.05
$SB_{latest}/SB_{F=0}$	0.80	0.83	0.37	0.45	0.96	1.29
$SB_{recent}/SB_{F=0}$	0.71	0.74	0.33	0.40	0.85	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.80	1.90	0.88	1.06	2.17	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.59	1.68	0.79	0.92	1.93	2.19
F_{MSY}	0.150	0.147	0.142	0.142	0.164	0.181
$F_{lim,\text{AS}}$	0.238	0.233	0.225	0.225	0.259	0.291
$F_{crash,\text{AS}}$	0.340	0.332	0.320	0.320	0.369	0.419
F_{latest}	0.052	0.051	0.039	0.048	0.057	0.064
F_{recent}	0.066	0.066	0.048	0.062	0.074	0.084
$F_{latest}/F_{\text{MSY}}$	0.35	0.35	0.24	0.31	0.38	0.39
$F_{recent}/F_{\text{MSY}}$	0.44	0.44	0.30	0.39	0.51	0.53
$F_{latest}/F_{lim,\text{AS}}$	0.22	0.22	0.15	0.20	0.24	0.25
$F_{recent}/F_{lim,\text{AS}}$	0.28	0.28	0.19	0.24	0.32	0.34
$F_{latest}/F_{crash,\text{AS}}$	0.15	0.16	0.11	0.14	0.17	0.17
$F_{recent}/F_{crash,\text{AS}}$	0.20	0.19	0.13	0.17	0.22	0.24

Table B-5: Summary of reference points for the subset of 96 grid models in the structural uncertainty grid dropping the EU series from the model.

	Mean	Median	Min	10%	90%	Max
C_{latest}	5963	5606	3806	3974	7776	9601
C_{recent}	6926	6744	4395	4673	8881	9577
MSY	11984	10188	9857	9868	16468	25496
SB_0	23942	21132	19099	20210	33493	53440
$SB_{F=0}$	25044	22015	19088	20161	35721	58714
SB_{MSY}	11677	10328	9095	9847	16536	26162
SB_{latest}	19207	18752	12973	15634	23926	38004
SB_{recent}	17064	16615	11320	13834	21207	33654
SB_{latest}/SB_0	0.84	0.87	0.42	0.48	0.94	1.19
SB_{recent}/SB_0	0.75	0.78	0.37	0.42	0.84	1.05
$SB_{latest}/SB_{F=0}$	0.81	0.83	0.38	0.45	0.93	1.29
$SB_{recent}/SB_{F=0}$	0.72	0.74	0.33	0.40	0.82	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.72	1.77	0.85	0.98	1.94	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.53	1.59	0.76	0.87	1.71	2.19
F_{MSY}	0.145	0.142	0.136	0.138	0.154	0.176
$F_{lim,\text{AS}}$	0.228	0.224	0.215	0.218	0.244	0.280
$F_{crash,\text{AS}}$	0.325	0.318	0.306	0.309	0.347	0.403
F_{latest}	0.070	0.069	0.039	0.048	0.088	0.117
F_{recent}	0.091	0.091	0.048	0.062	0.115	0.160
$F_{latest}/F_{\text{MSY}}$	0.49	0.49	0.24	0.33	0.64	0.78
$F_{recent}/F_{\text{MSY}}$	0.63	0.65	0.30	0.43	0.83	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.31	0.31	0.15	0.21	0.41	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.40	0.41	0.19	0.27	0.53	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.22	0.22	0.11	0.14	0.29	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.28	0.29	0.13	0.19	0.37	0.48

Table B-6: Summary of reference points for the subset of 108 grid models in the structural uncertainty grid removing early years (<2005) from the NZ CPUE

	Mean	Median	Min	10%	90%	Max
C_{latest}	5932	5645	3707	4000	7604	9601
C_{recent}	6780	6606	4322	4562	8745	9577
MSY	11080	10023	8968	9046	15155	24091
SB_0	21913	20683	15686	16920	31829	53503
$SB_{F=0}$	26027	23854	17559	18902	37973	66434
SB_{MSY}	10733	9985	7564	8145	15738	26684
SB_{latest}	18643	18070	13286	15912	23926	38004
SB_{recent}	16471	15938	11848	14000	21207	33654
SB_{latest}/SB_0	0.89	0.90	0.42	0.52	1.14	1.19
SB_{recent}/SB_0	0.79	0.80	0.38	0.45	1.00	1.05
$SB_{latest}/SB_{F=0}$	0.77	0.74	0.32	0.45	1.04	1.29
$SB_{recent}/SB_{F=0}$	0.68	0.65	0.29	0.40	0.91	1.15
$SB_{latest}/SB_{\text{MSY}}$	1.83	1.84	0.85	1.05	2.37	2.47
$SB_{recent}/SB_{\text{MSY}}$	1.62	1.62	0.76	0.92	2.08	2.19
F_{MSY}	0.152	0.151	0.140	0.142	0.162	0.181
$F_{lim,\text{AS}}$	0.241	0.241	0.222	0.225	0.259	0.291
$F_{crash,\text{AS}}$	0.343	0.343	0.316	0.320	0.369	0.419
F_{latest}	0.074	0.074	0.039	0.051	0.094	0.120
F_{recent}	0.092	0.093	0.048	0.064	0.116	0.152
$F_{latest}/F_{\text{MSY}}$	0.49	0.51	0.24	0.33	0.66	0.72
$F_{recent}/F_{\text{MSY}}$	0.61	0.61	0.30	0.40	0.82	0.91
$F_{latest}/F_{lim,\text{AS}}$	0.31	0.32	0.15	0.21	0.42	0.46
$F_{recent}/F_{lim,\text{AS}}$	0.38	0.38	0.19	0.25	0.52	0.58
$F_{latest}/F_{crash,\text{AS}}$	0.22	0.22	0.11	0.14	0.29	0.32
$F_{recent}/F_{crash,\text{AS}}$	0.27	0.27	0.13	0.18	0.36	0.41

Table B-7: Summary of reference points for the subset of 126 grid models in the structural uncertainty grid with high initial fishing mortality.

	Mean	Median	Min	10%	90%	Max
C_{latest}	6361	6262	3707	3868	8919	9022
C_{recent}	7216	7461	4322	4542	9277	9384
MSY	17119	16402	11924	12672	24190	25629
SB_0	33051	32480	18814	19587	48510	53503
$SB_{F=0}$	38499	38129	20270	22482	57049	66434
SB_{MSY}	16068	15946	8933	9294	23590	26684
SB_{latest}	17058	15726	12973	13704	23846	24631
SB_{recent}	15009	13870	11320	11840	20932	21686
SB_{latest}/SB_0	0.54	0.49	0.42	0.44	0.70	0.73
SB_{recent}/SB_0	0.47	0.43	0.37	0.39	0.60	0.63
$SB_{latest}/SB_{F=0}$	0.46	0.43	0.32	0.37	0.61	0.68
$SB_{recent}/SB_{F=0}$	0.41	0.37	0.29	0.33	0.53	0.59
$SB_{latest}/SB_{\text{MSY}}$	1.11	1.00	0.85	0.90	1.44	1.54
$SB_{recent}/SB_{\text{MSY}}$	0.98	0.88	0.76	0.80	1.26	1.33
F_{MSY}	0.148	0.147	0.135	0.136	0.162	0.181
$F_{lim,\text{AS}}$	0.234	0.233	0.212	0.215	0.255	0.291
$F_{crash,\text{AS}}$	0.333	0.331	0.301	0.304	0.367	0.419
F_{latest}	0.083	0.083	0.054	0.055	0.106	0.120
F_{recent}	0.106	0.105	0.071	0.074	0.137	0.160
$F_{latest}/F_{\text{MSY}}$	0.57	0.58	0.34	0.38	0.72	0.78
$F_{recent}/F_{\text{MSY}}$	0.72	0.71	0.42	0.50	0.96	1.06
$F_{latest}/F_{lim,\text{AS}}$	0.36	0.37	0.21	0.24	0.46	0.50
$F_{recent}/F_{lim,\text{AS}}$	0.46	0.45	0.26	0.31	0.61	0.68
$F_{latest}/F_{crash,\text{AS}}$	0.25	0.26	0.15	0.16	0.32	0.35
$F_{recent}/F_{crash,\text{AS}}$	0.32	0.32	0.18	0.22	0.43	0.48