Check for updates

Received: 2 April 2022 | Revised: 26 September 2022 | Accepted: 4 October 2022

DOI: 10.1111/faf.12714

ORIGINAL ARTICLE

FISH and FISHERIES WILEY

How consistent is the advice from stock assessments? Empirical estimates of inter-assessment bias and uncertainty for marine fish and invertebrate stocks

Rujia Bi[1] | Chip Collier[2] | Roger Mann[3] | Katherine E. Mills[4] | Vincent Saba[5] | John Wiedenmann[6] | Olaf P. Jensen[1]

[1]Center for Limnology, University of Wisconsin–Madison, Madison, Wisconsin, USA

[2]South Atlantic Fishery Management Council, North Charleston, South Carolina, USA

[3]Virginia Institute of Marine Science, William & Mary, Gloucester Point, Virginia, USA

[4]Gulf of Maine Research Institute, Portland, Maine, USA

[5]National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Northeast Fisheries Science Center, Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, New Jersey, USA

[6]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, New Jersey, USA

Correspondence
Rujia Bi, Center for Limnology, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA.
Email: rbi24@wisc.edu

Funding information
Lenfest Ocean Program

Abstract

Fishery management frequently involves precautionary buffering for scientific uncertainty. For example, a precautionary buffer that scales with scientific uncertainty is used to calculate the acceptable biological catch downward from the overfishing limit in the US federal fishery management system. However, there is little empirical guidance to suggest how large buffers for scientific uncertainty should be. One important component of uncertainty is variation among different assessments of the same stock in estimates of management-relevant quantities. We analysed commercially exploited marine fish and invertebrate stocks around the world and developed Bayesian hierarchical models to quantify inter-assessment variation in terminal year biomass and fishing mortality estimates, reference points, relative biomass and fishing mortality estimates, and overfishing limits. There was little evidence of inter-assessment bias; stock assessment estimates in the terminal year of the assessment were not consistently higher or lower than estimates of the same quantities in future years. However, there was a tendency for extreme values from the terminal year to be pulled closer to the mean in future years. Inter-assessment variation in all estimates differed across regions, and a longer inter-assessment interval generally resulted in greater variation. Inter-assessment uncertainty was greatest for estimates of the overfishing limit, with coefficients of variation ranging from 17% in Europe (non-EU) to 107% for Pacific Ocean pelagic stocks. Because inter-assessment variation is only one component of scientific uncertainty, we suggest that these uncertainty estimates may provide a basis for determining the minimum size of precautionary buffers.

KEYWORDS
annual catch limits, fisheries management, probability of overfishing, scientific uncertainty, stock assessment consistency

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *Fish and Fisheries* published by John Wiley & Sons Ltd.

1 INTRODUCTION	127
2 METHODS	128
2.1 Stock selection	128
2.2 Assessment output comparison	129
2.3 Model framework	129
2.3.1 Common coefficient of variation	130
2.3.2 Varied coefficient of variation across regions	130
2.3.3 Varied coefficient of variation across regions and time periods	130
2.4 Model fitting and comparison	131
3 RESULTS	132
3.1 Inter-assessment variations and potential causes	132
3.2 Model comparisons and results	132
3.3 Inter-assessment uncertainty in stock status	134
4 DISCUSSION	134
4.1 Overall variability in estimates of stock status on overfished and overfishing	134
4.2 Implications for fisheries management	134
4.3 Factors associated with higher variability	137
ACKNOWLEDGEMENTS	138
CONFLICT OF INTEREST	137
DATA AVAILABILITY STATEMENT	138
REFERENCES	139

1 | INTRODUCTION

Stock assessments, mathematical models of the population dynamics of harvested fish and invertebrates, provide key inputs used to guide decision-making within many fishery management systems. Most notably, they provide estimates of biological reference points and stock status with respect to these reference points, as well as an estimate of the target or limit harvest that can be taken in the following year or years. The efficacy of stock assessments in guiding fisheries management is challenged by uncertainties inherent to the ecosystem and management process (Garcia, 2000; Mildenberger et al., 2022). Uncertainties originate from biological, economic, and political factors that influence fisheries and interface with the ability to develop effective management measures. Such uncertainties in fisheries stem primarily from these inescapable facts: unstable and unpredictable states of nature, observational errors, model misspecification, and multiple, sometimes conflicting, scientific and economic goals (Charles, 1998; Hanna, 1997; Hilborn, 1987; Sethi, 2010; Sissenwine, 1984). Errors in the management advice provided by stock assessments have been implicated in overfishing and the failure of some depleted stocks to recover (Brooks & Legault, 2016; Wiedenmann & Jensen, 2018). Identifying and quantifying uncertainty helps to estimate the probability that different harvest levels will prevent overfishing (Edwards, 2016), and many fishery management systems include buffers in harvest levels to explicitly account for scientific uncertainty. However, despite previous reviews and simulation studies (Ralston et al., 2011; Privitera-Johnson & Punt, 2020a), there remains much that is unknown regarding both the magnitude of scientific uncertainty in different stock assessment outputs and the potential for systematic bias (i.e., a non-zero mean difference between a stock assessment output and a future understanding of the value of this same quantity).

The fisheries management process is composed of the following steps: data collection, data analysis, harvest control rule specification, and regulation implementation. Uncertainty may occur in any component of this cycle. For example, process uncertainty underlies changes in population dynamics, such as variation in growth and natural mortality (Edwards, 2016). Observation uncertainty is produced from variation in measurement during data collection (Rosenberg & Restrepo, 1994). Another source, model uncertainty, can come from misspecification of model parameters (e.g., fixed natural mortality rate or catchability) or model structure (e.g., age-aggregated or age-structured), and retrospective biases (i.e., systematic changes in model outputs that may arise when additional periods of data are added to or removed from a stock assessment) (Brooks & Deroba, 2015; Dorn & Zador, 2020; Hurtado-Ferro et al., 2015; Legault, 2009; Mohn, 1999; Privitera-Johnson & Punt, 2020a). Lastly, estimation uncertainty, such as inaccurate and imprecise estimates from model fitting (Francis & Shotton, 1997), can manifest in the data analysis step and undermine the efficacy of fisheries management. Process, observation, model, and estimation uncertainties are collectively called scientific uncertainty (Privitera-Johnson & Punt, 2020b).

Recognition of uncertainties throughout the fisheries management process led to the widespread adoption of a precautionary approach to fisheries management in the 1990s (Hilborn et al., 2001). Precautionary fishery management requires an understanding of the magnitude of scientific uncertainty in estimates of stock status, biological reference points, and harvest levels that will achieve management goals (Dettloff, 2020; Hilborn et al., 2001; Ralston et al., 2011). Application of the precautionary approach is perhaps most formalized in the US federal fishery management system where it follows a multistep process. First, the overfishing limit (OFL), the best estimate of the maximum amount of a stock that can be caught without resulting in overfishing, is estimated in a stock assessment (Shertzer et al., 2010). Next, an acceptable biological catch (ABC) is set at or below the OFL to account for scientific uncertainty (Prager & Shertzer, 2010). One approach used by a number of US Regional Fishery Management Councils ("Councils" hereafter) for setting the ABC is the p^* (pronounced "p-star") method developed by Shertzer et al. (2008). Under this approach, a distribution of the OFL is assumed to be centred on the assessment's OFL point estimate with an assumed level of variation to account for scientific uncertainty. In some regions of the US, the OFL is assumed to follow a lognormal distribution with a coefficient of variation (CV) that is specified by scientific advisory committees which review the stock assessments and provide an

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ABC recommendation to managers (MAFMC, 2011; PFMC, 2010). The target probability of overfishing (p^*), is then selected by fisheries managers (Shertzer et al., 2008) and the resulting ABC is calculated. For example, with a $p^* = .4$, the 40th percentile of the OFL distribution, corresponding with a 40% chance of overfishing, is selected as the ABC. Regulations promulgated for implementing National Standard 1 of the US Magnuson-Stevens Fishery Conservation and Management Act mandate that p^* must never exceed .5 (Federal Register, 2009, 2016). Finally, the annual catch limit (ACL) is set equal to or lower than the ABC to account for conservation objectives, socioeconomic concerns, management goals and implementation uncertainty (the uncertainty associated with achieving a certain target catch). Because uncertainty in the ABC setting step propagates through to final harvest rules and implementation, quantifying scientific uncertainly is important to develop appropriate management limits, and to properly specify the risk of overfishing.

Beyond setting the ABC, there are other reasons for trying to quantify and understand uncertainty and bias in stock assessment estimates. Large changes in estimates between assessments can lead to a lack of trust in the scientific process amongst stakeholders and reduced catch levels. For example, Wiedenmann and Jensen (2018) found that for many species of groundfish in New England, although the fishery typically stayed within annual catch limits, overestimation of abundance led to continued overfishing. Changes to the overfished status could invoke rebuilding plans which typically require large reductions in fishing mortality to enable the population to recover. Large changes in assessment estimates could also result in a high-level bias in stock status indicators (i.e., is the stock overfished or experiencing overfishing?) and negatively impact effectiveness of fish stock rebuilding plans (Parma et al., 2013). For example, some overfished stocks were previously misclassified as not overfished, and the inverse may also have occurred, a consequence of uncertainties in estimating fish stock biomass or fishing mortality (Parma et al., 2013). Besides the catch limits, reference points to safeguard against low biomass or high fishing mortality in the face of high uncertainty have been recommended and implemented in precautionary fishery management (Da-Rocha et al., 2016; Mildenberger et al., 2022). Understanding uncertainty in assessment outputs can also improve the performance of Management Strategy Evaluation (MSE) which often attempt to mimic realistic levels of scientific uncertainty in order to evaluate the performance of different harvest strategies (Mildenberger et al., 2022; Punt et al., 2016).

There are several ways to quantify scientific uncertainty in stock assessments. One way is to calculate uncertainty estimates such as standard errors and confidence/credible intervals from stock assessment models. But such methods would underestimate the true uncertainty as they are conditional on the assumption that the underlying model is an accurate and complete representation of the system (i.e., population; Brodziak & Walsh, 2013; Stewart & Hicks, 2018). Another approach uses simulation methods, in which the dynamics of a population are simulated, observations (with error) are extracted, and assessment models are fit to these observations (Conn et al., 2010; Magnusson & Hilborn, 2007; Yin & Sampson, 2004). However, the extent to which uncertainty estimates derived from simulation studies represent the uncertainty to be expected in real assessments depends on the extent to which the simulation model represents the dynamics of the real population. A third approach is to compare outputs between different assessments of the same stock (Ralston et al., 2011; Wiedenmann & Jensen, 2018; Privitera-Johnson & Punt, 2020b; Silvar-Viladomiu et al., 2021). The variation in historical estimates among multiple stock assessments for the same stock can be used to evaluate bias and uncertainty in assessment outputs. For example, Ralston et al. (2011) quantified the variation in historical time series of spawning biomass estimates among multiple assessments of the same stock based on US West Coast groundfish and coastal pelagic species stocks using this approach. Although estimating scientific uncertainty in stock assessment outputs will always be hindered by the fundamental fact that the true values of these outputs can never be known for real fish populations, this latter approach improves on previous methods and is a useful way to quantify precision, but not accuracy, among stock assessments in model estimates.

In this study, we expanded on previous studies quantifying inter-assessment variation, analysed multiple commercially exploited marine fish and invertebrate stocks around the world and quantified inter-assessment variations in biomass and fishing mortality estimates, reference point estimates, relative biomass and fishing mortality estimates, as well as catch limits. In doing so, we provide a basis for determining the minimum buffer limit for scientific uncertainty in fisheries management.

2 | METHODS

2.1 | Stock selection

We made use of the RAM Legacy Stock Assessment Database (RAMLDB, v4.491, https://www.ramlegacy.org/), an open-access compilation of data-rich stock assessment output for commercially exploited marine fish and invertebrate populations from around the globe (Figure S1 in the Appendix S1; Ricard et al., 2012). Data retrieved from the assessments included time series of catch and model-estimated biomass and fishing mortality rates as well as reference points such as maximum sustainable yield (MSY), the expected equilibrium biomass (B_{MSY}) for a stock harvested at MSY, and the fishing mortality rate that results in B_{MSY} and MSY at equilibrium (F_{MSY}). For many stocks, the RAMLDB contains model estimates from multiple, sequential assessments. We chose stocks with more than one assessment in the database for our analysis (Table S1 in the Appendix S1) and augmented the data available in the RAMLDB through the addition of data from 34 assessments for 14 stocks that were taken directly from stock assessment documents. A total

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of 838 assessments of 277 stocks were included in our analysis (Figure 1, Table S1 in the Appendix S1).

2.2 | Assessment output comparison

For each stock, we conducted pairwise comparisons among model estimates from all available assessments, ranging from 2 to 8 assessments for a given stock. For each comparison, we determined the target year (ty) as the final (terminal) year in older assessments and obtained biomass and fishing mortality estimates in the target year (B_{ty} and F_{ty}) from all available assessments. The total number of pairwise comparisons was 1 for a stock with 2 assessments, and 3, 6, 10, 15, 21 or 28 for a stock with 3, 4, 5, 6, 7, or 8, respectively. For example, Argentine anchoita (*Engraulis anchoita*, Engraulidae) in South America has three assessments with terminal years of 2007 (A1), 2015 (A2), and 2016 (A3). We conducted pairwise comparisons between model estimates in 2007 from A1 and A2, between estimates in 2007 from A1 and A3, and between estimates in 2015 from A2 and A3. We chose to focus on terminal year estimates as these are the most relevant to management decision-making. Pairwise model estimates among assessments for the same stock were fitted with linear regressions to detect if the slopes were different from 1 or the intercepts were different from 0, to assess patterns of the variations and evaluate the presence of systematic bias.

We also compared reference points at MSY (B_{MSY} and F_{MSY}) or their proxies; some were provided in the RAMLDB, and others were found in the assessment documents, relative biomass and fishing mortality rate in the target year (B_{ty}/B_{MSY} and F_{ty}/F_{MSY}), and the OFL in the target year (OFL_{ty}) from all available assessments for each stock. The OFL_{ty} was a function of B_{ty} and F_{MSY}:

$$OFL_{ty} = \begin{cases} F_{MSY}B_{ty}, & \text{when} F_{MSY} \text{ is a discrete rate} \\ \frac{F_{MSY}}{F_{MSY}+M}B_{ty}\left(1-e^{-M-F_{MSY}}\right), & \text{when} F_{MSY} \text{ is an instantaneous rate} \end{cases}$$

The biomasses used in the OFL calculation were sometimes defined differently for different stocks (e.g., spawning stock biomass or total biomass), but were consistent across assessments for the same stock. Thus, the OFL calculated for this analysis was not always equal an OFL calculated for use in management.

We gathered information for each stock on its region, management council (for federally managed US stocks), assessment model structure (e.g., age-aggregated or age-structured), model assumption (e.g., fixed or time-varying parameters like catchability), data input (e.g., changed time-series catch data, changed time-period of data input), and natural mortality (M) for all available assessments. For assessments without information on assessment model, data input, or M in the RAMLDB, we obtained the information from the corresponding stock assessment documents.

We also evaluated inter-assessment consistency in threshold-based categorization of stock status: $F_{ty}/F_{MSY} > 1$ indicated that the stock was experiencing overfishing, $B_{ty}/B_{MSY} < 0.5$ indicated that the stock was overfished. This is a definition typically used in the US and several other countries (Hilborn et al., 2020). For overfished status, we defined a binary outcome (O^B_{ty}) that = 0 if B_{ty}/B_{MSY} from the newer assessment ≥0.5, and =1 if B_{ty}/B_{MSY} from the newer assessment <0.5 (indicating overfished). For overfishing status, we defined a binary outcome (O^F_{ty}) that = 0 if F_{ty}/F_{MSY} from the newer assessment ≤1, =1 if F_{ty}/F_{MSY} from the newer assessment >1 (indicating overfishing). Pairwise comparisons among assessments were conducted for each stock. Logistic regression curves were fitted to B_{ty}/B_{MSY} from the older assessment and O^B_{ty}, and F_{ty}/F_{MSY} from the older assessment and O^F_{ty}, separately, using the glm function in R.

2.3 | Model framework

We assumed that the B_{ty}, F_{ty}, B_{MSY}, F_{MSY}, B_{ty}/B_{MSY}, F_{ty}/F_{MSY}, and OFL_{ty} estimates from all available assessments for each stock followed lognormal distributions. We chose this distribution because the distribution of estimated biomass, fishing mortality and other quantities in assessment models are bounded by zero and often exhibit a long right tail, and the lognormal distribution had been widely adopted for the OFL estimate (MAFMC, 2011; PFMC, 2010). The lognormal distribution was characterized by a mean and a CV. In the following section, we present different scenarios based on the hierarchical structures of the CV. We first developed a model with globally

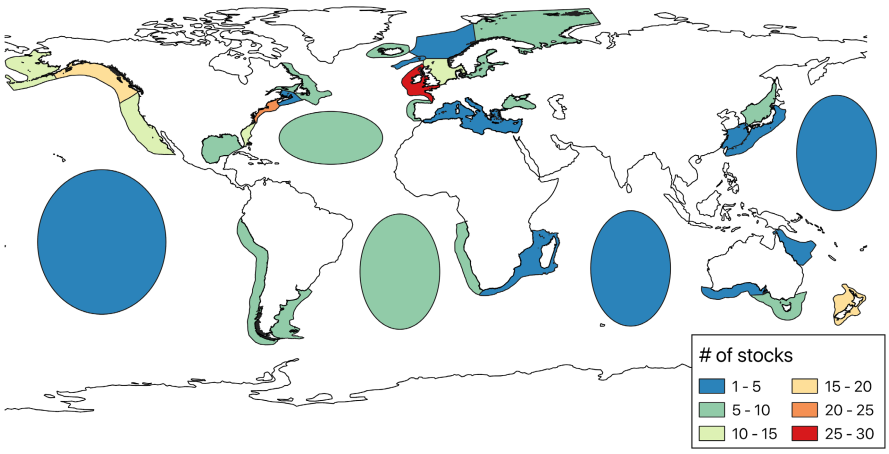


FIGURE 1 A map of stocks incorporated in this analysis. Each stock is assigned to a large marine ecosystem (LME), which encompass the continental shelves of the world's oceans and represent the most productive areas of the oceans. Large highly migratory oceanic species, such as tuna, are assigned to high seas areas (represented by ovals) that are not included in the LME classification. Map is built based on LME classification by Ricard et al. (2012) and using QGIS3.10.

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

constant CV, which assumed that the lognormal distributions for all stocks have the same CV, against which different scenarios of varied CV were compared. And then we developed models with CVs varying across regions and time periods between assessments. Symbols used in model equations are defined in Table 1. Model equations are displayed in Table 2.

2.3.1 | Common coefficient of variation

In the constant CV scenario (termed M1), the log-transformed estimates in year ty from the two assessments under the *j*th comparison for stock *i* ($Y_{i,j,\mathrm{ty}}$) were modelled as normally distributed with a mean $\mu_{i,\mathrm{ty}}$ and a globally constant variance parameter σ_g^2 (Equation 1.1). The CV_g of log-normal distribution was derived from σ_g^2 (Equation 1.2). For B_{MSY} or F_{MSY}, there was no subscript about year (ty), so μ_i would be the same for a certain stock.

2.3.2 | Varied coefficient of variation across regions

Next, because CV might differ across regions, instead of independent uniform priors, we modelled a second scenario (termed M2) where log-transformed $Y_{i,j,\mathrm{ty}}$ were modelled as normally distributed with a mean $\mu_{i,\mathrm{ty}}$ and a region (*r*)-specific variance parameter σ_r^2 (Equation 2.1). The region-specific CV_r was derived from σ_r^2 (Equation 2.2). The logarithms of CV_r were modelled as normally distributed with a global mean $\log(\mathrm{CV}_g)$ and a common variance parameter υ_r (Equation 2.3). For US stocks, region was further classified to management council.

2.3.3 | Varied coefficient of variation across regions and time periods

In the third scenario, we not only included the variations on CV across regions, but also considered the impacts of the time elapsed between two assessments on the magnitude of CV estimates. When the time period between the terminal years of the two assessments was larger, it was hypothesized to have a greater CV estimate (Van Beveren et al., 2021). We defined different CVs (i.e., $\mathrm{CV}_{\mathrm{long}}$ and $\mathrm{CV}_{\mathrm{short}}$) based on the time period (Equation 3.1).

To better understand how to define $\mathrm{CV}_{\mathrm{long}}$ and $\mathrm{CV}_{\mathrm{short}}$, especially for stocks with more than two assessments, we provide an example here. The Argentine anchoita in South America has three assessments with terminal years of 2007 (A1), 2015 (A2), and 2016 (A3). We conducted pairwise comparisons and assumed that model estimates in 2007 from A1 and A2 followed a lognormal distribution with a mean of μ_{2007} and a $\mathrm{CV}_{\mathrm{long}}$, estimates in 2007 from A1 and A3 followed a lognormal distribution with a mean of μ_{2007} and a $\mathrm{CV}_{\mathrm{long}}$, and estimates in 2015 from A2 and A3 followed a lognormal distribution with a mean of μ_{2015} and a $\mathrm{CV}_{\mathrm{short}}$.

TABLE 1 Symbols used in model equations

Symbol	Description
	Indicator variables
i	Stock
j	Comparison for a stock
ty	Target year
r	Region
tp	Time periods (long = more than 5 years, short = less than or equal to 5 years)
	Observed data
$Y_{i,j,\mathrm{ty}}$	One of the 7 model estimates in target year ty (i.e., B_{ty}, B_{MSY}, $B_{\mathrm{ty}}/B_{\mathrm{MSY}}$, F_{ty}, F_{MSY}, $F_{\mathrm{ty}}/F_{\mathrm{MSY}}$ and $\mathrm{OFL}_{\mathrm{ty}}$) from the two assessments in the *j*th comparison for stock *i*
	Estimated parameters
$\mu_{i,\mathrm{ty}}$	Mean of log-transformed model estimates in year ty for stock *i*
σ_g	Globally constant standard deviation of log-transformed model estimates
CV_g	Globally constant CV of model estimates
σ_r	Region-specific standard deviation of log-transformed model estimates
CV_r	Region-specific CV of model estimates
υ_r	Variance of log-transformed CV_r deviation
$\sigma_{r,\mathrm{tp}}$	Region and time period-specific standard deviation of log-transformed model estimates
$\mathrm{CV}_{r,\mathrm{tp}}$	Region and time period-specific CV of model estimates
υ_{tp}	Variance of log-transformed $\mathrm{CV}_{r,\mathrm{tp}}$ deviation
ω	Difference between $\mathrm{CV}_{\mathrm{long}}$ and $\mathrm{CV}_{\mathrm{short}}$ in M3-b

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Model equations

Model	Description	Equation	Eq.
M1	Common coefficient of variation	$\log(Y_{i,j,\mathrm{ty}}) \sim N\left(\mu_{i,\mathrm{ty}}, \sigma_g^2\right)$	1.1
		$\sigma_g^2 = \log\left(\mathrm{CV}_g^2 + 1\right)$	1.2
M2	Varied coefficient of variation across regions	$\log(Y_{i,j,\mathrm{ty}}) \sim N(\mu_{i,\mathrm{ty}}, \sigma_r^2)$	2.1
		$\sigma_r^2 = \log\left(\mathrm{CV}_r^2 + 1\right)$	2.2
		$\log(\mathrm{CV}_r) \sim N(\log(\mathrm{CV}_g), \upsilon_r)$	2.3
M3	Varied coefficient of variation across regions and time periods	$\mathrm{CV}_{\mathrm{tp}} = \begin{cases} \mathrm{CV}_{\mathrm{long}}, \text{if time period} > 5 \\ \mathrm{CV}_{\mathrm{short}}, \text{if time period} \leq 5 \end{cases}$	3.1
		$\log(Y_{i,j,\mathrm{ty}}) \sim N\left(\mu_{i,\mathrm{ty}} \sigma_{r,\mathrm{tp}}^2\right)$	3.2
		$\sigma_{r,\mathrm{tp}}^2 = \log\left(\mathrm{CV}_{r,\mathrm{tp}}^2 + 1\right)$	3.3
	M3-a	$\log(\mathrm{CV}_r) \sim N(\log(\mathrm{CV}_g), \upsilon_r)$	3.4
		$\log(\mathrm{CV}_{r,\mathrm{tp}}) \sim N(\log(\mathrm{CV}_r), \upsilon_{\mathrm{tp}})$	3.5
	M3-b	$\log(\mathrm{CV}_{r,\mathrm{short}}) \sim N(\log(\mathrm{CV}_g), \upsilon_r)$	3.6
		$\mathrm{CV}_{r,\mathrm{long}} = \mathrm{CV}_{r,\mathrm{short}} + \omega$	3.7

The log-transformed $Y_{i,j,\mathrm{ty}}$ were modelled as normally distributed with a mean $\mu_{i,\mathrm{ty}}$ and a region (r) and time period (tp)-specific variance parameter $\sigma_{r,\mathrm{tp}}^2$ (Equation 3.2). The region and time period-specific $\mathrm{CV}_{r,\mathrm{tp}}$ was derived from $\sigma_{r,\mathrm{tp}}^2$ (Equation 3.3). We developed two models to estimate CV in this scenario. In the first model (termed M3-a), the CV parameter was indexed by both region (r) and time period (tp). Region-specific CV (CV_r) were modelled as log-normally distributed with a global mean CV_g and a common variance parameter υ_r (Equation 3.4). Time period variations in CV were nested in regions, and region and time period-specific CV ($\mathrm{CV}_{r,\mathrm{tp}}$) were modelled as log-normally distributed with a region-specific CV_r and a common variance parameter υ_{tp} (Equation 3.5). In the second model (termed M3-b), region-specific $\mathrm{CV}_{\mathrm{short}}$ ($\mathrm{CV}_{r,\mathrm{short}}$) were modelled as log-normally distributed with a global mean CV_g and a common variance parameter υ_r (Equation 3.6), and the difference between $\mathrm{CV}_{\mathrm{long}}$ and $\mathrm{CV}_{\mathrm{short}}$ was assumed to be a positive constant (ω ; Equation 3.7). In this way, we decreased the number of parameters in the model.

2.4 | Model fitting and comparison

We used Bayesian estimation methods because of their convenience for specifying hierarchical models. The Bayesian models incorporated prior probability distributions, modelled dynamics and structured the likelihood and finally used posterior distributions to quantify uncertainty. Uniform prior probability distributions were adopted (for details, please see Table S2 in the Appendix S1).

To simulate Markov Chain Monte Carlo (MCMC) samples from the posterior, we used JAGS 4.0 (Plummer, 2003) with the R packages rjags (Plummer, 2016) and runjags (Denwood, 2016) implemented in R (R Core Team, 2019). For each model, five chains with different initial conditions were simulated, and the convergence of different chains was checked by Gelman-Rubin convergence diagnostics (Gelman & Rubin, 1992) and trace plots (Giudici & Castelo, 2003).

We compared model performance using the deviance information criterion (DIC; Spiegelhalter et al., 2002), Watanabe-Akaike information criterion (WAIC; Watanabe, 2010), and leave-one-out cross-validation (LOO; Vehtari et al., 2017).

The DIC is defined as:

$$\mathrm{DIC} = \overline{D} + P_D$$

where $\overline{D}$ is the posterior mean of the deviance of the model, and p_D is the effective number of parameters in the model.

The WAIC is defined as:

$$\mathrm{WAIC} = -2^*(\mathrm{LPPD} - P_D)$$

where LPPD is the log posterior predictive density.

The LOO is defined as:

$$\mathrm{LOO} = \sum_{i=1}^{n} \log p(y_i \mid y_{-i})$$

where y_{-i} denotes the observations y with the ith component removed. It expresses the posterior probability of observing the value of y_i when the model is fitted to all data except y_i.

The WAIC and LOO were computed with R package loo (Vehtari et al., 2016). The WAIC is known to be more stable than DIC because it is fully Bayesian and uses the entire posterior distribution (Vehtari et al., 2017). The LOO was computed using Pareto smoothed importance sampling that provides a more accurate and reliable estimate by applying a smoothing procedure to the importance weights (Vehtari et al., 2017; Vehtari & Gelman, 2015). A smaller value of DIC, WAIC, or LOO indicates a better model performance. If all

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

three criteria showed the same preference for a model, we had more evidence that the preference was correct. If they showed different preferences, we picked the best model based on WAIC and LOO.

3 | RESULTS

3.1 | Inter-assessment variations and potential causes

Pairwise comparisons on log-transformed model estimates (i.e., B_{ty}, F_{ty}, B_{MSY}, F_{MSY}, B_{ty}/B_{MSY}, F_{ty}/F_{MSY}, and OFL_{ty}) among assessments for the same stock were displayed in Figure 2. Fitted linear regressions were, in general, visually similar to the one-to-one line. However, except for OFL_{ty}, the slopes of the regression lines were significantly lower than 1 with the 95% credible intervals (CIs) below 1. The intercepts of fitted linear regressions for B_{ty} and B_{MSY} were significantly greater than 0 with the 95% CIs above 0; the intercepts of fitted linear regressions for B_{ty}/B_{MSY}, F_{ty}, and F_{MSY} were significantly lower than 0 with the 95% CIs below 0; the intercepts of fitted linear regressions for F_{ty}/F_{MSY} and OFL_{ty} were not significantly different from 0 because the 95% CIs overlapped 0. Although fishing mortality-related estimates (i.e., F_{ty}, F_{MSY}, and F_{ty}/F_{MSY}) appeared unbiased overall, there was some tendency for unusually low values from the older assessment to be adjusted upward in the more recent assessment and high values to be adjusted downward. Some visually obvious outliers were observed in biomass-related estimates, especially, B_{ty}/B_{MSY}. Many of these variations or biases were associated with changes in the underlying assessment model structure, assumed values of natural mortality, definitions of reference points, and input data.

Changes in the input data were sometimes implicated in larger changes in model estimates. For example, for bight redfish (*Centroberyx gerrardi*, Berycidae) in Southeast Australia, a 27% decrease in $\log(B_{MSY})$ was associated with a marked increase in the number of data sources included in the assessment (Figure 2, point 1; Haddon, 2016).

Changes in the underlying model structure were associated with some of the largest inter-assessment differences (e.g., Figure 2, points 2 and 3; ICES, 2013a, 2015a; SEDAR, 2017). For example, for Atlantic herring (*Clupea harengus*, Clupeidae) in the International Council of the Exploration of the Sea (ICES) 5a-7bc region, a 47% increase in $\log(F_{ty})$ was associated with a change in assessment method from a trends-based exploratory assessment to an age-based analytical assessment in 2015 (Figure 2, point 2).

Changes in the assumed value of natural mortality also led to changes in model estimates, especially *F*-related estimates. For example, for Alaska plaice (*Pleuronectes quadrituberculatus*, Pleuronectidae) in the Bering Sea and Aleutian Islands, the natural mortality rate assumed in the assessment decreased from 0.25 y^{-1} in the 2009 assessment to 0.13 y^{-1} in later assessments and led to large changes in F_{MSY}, F_{ty}/F_{MSY}, and OFL_{ty} (Figure 2, point 4; Wilderbuer et al., 2010).

Changes in definitions of reference points led to changes in B_{MSY}, F_{MSY}, and related model estimates (e.g., Figure 2, points 5, 6 and 7; Chute et al., 2013; ICES, 2012a, 2012b, 2013b, 2014, 2015b, 2016). For example, for the US Atlantic stock of ocean quahog (*Arctica islandica*, Arcticidae), more conservative reference points for the biomass threshold, fishing mortality threshold and target fishing mortality were implemented in 2009 (Figure 2, point 5).

Changes in inter-assessment model estimates might also be induced by the estimation procedure. For example, for Atlantic cod (*Gadus morhua*, Gadidae) in the Irish Sea, the estimation procedure was changed from the state-space to conventional likelihood, and new reference points were estimated from the EqSim simulation, a stochastic equilibrium software used to explore MSY reference points, in the 2017 assessment (Figure 2, point 8; ICES, 2017).

3.2 | Model comparisons and results

For all models, the Gelman-Rubin statistics for all the posterior samples were found to be smaller than 1.1, and the trace plots showed that the chains mixed well and moved back and forth over the space, both suggesting that the convergence of the posteriors was validated. The DIC, WAIC, and LOO results for the four models with different CV configurations for the seven model estimates are presented in Table S3 in the Appendix S1. For B_{ty}, F_{ty}, B_{MSY}, and F_{MSY}, model with a varied CV across regions and time periods (M3-a) achieved better performance than other models in terms of the smallest WAIC and LOO values. For B_{ty}/B_{MSY}, F_{ty}/F_{MSY}, and OFL_{ty}, model with a varied CV across regions and time periods but $CV_{long} - CV_{short}$ was a positive constant (M3-b) achieved better performance.

The estimated global mean CVs of F_{ty} and OFL_{ty} were greatest, followed by F_{ty}/F_{MSY} and B_{ty}; the global mean CVs of B_{MSY}, B_{ty}/B_{MSY}, and F_{MSY} were smaller (Table 3). The region-specific CV of B_{ty} derived from model M3-a was greatest for the US Gulf of Mexico Fishery Management Council (GMFMC; Figure 3a). The region-specific CV of B_{MSY} derived from model M3-a was greatest for the Pacific Ocean High Seas (POHS), followed by the US GMFMC (Figure 3a). The region-specific CV of B_{ty}/B_{MSY} derived from model M3-b was greatest for Australia, followed by the US Mid-Atlantic Fishery Management Council (MAFMC; Figure 3a). The region-specific CV of F_{ty} derived from model M3-a was greatest for the US GMFMC (Figure 3b). The region-specific CV of F_{MSY} derived from model M3-a was greatest for the US MAFMC and GMFMC (Figure 3b). The region-specific CV of F_{ty}/F_{MSY} derived from model M3-b was greatest for the US MAFMC (Figure 3b). The region-specific CV of OFL_{ty} derived from model M3-b was greatest for the POHS, followed by the US MAFMC, and lowest in Europe (Figure 3c).

Regression analyses on the average frequency of assessment updates and estimated CV_{long} and CV_{short} in each region suggested that CV (i.e., uncertainty) increased with increased assessment update intervals, except for CV_{short} of B_{ty} (Figure 4). That is, a region with a less frequent assessment update (greater update interval) would have a greater uncertainty.

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

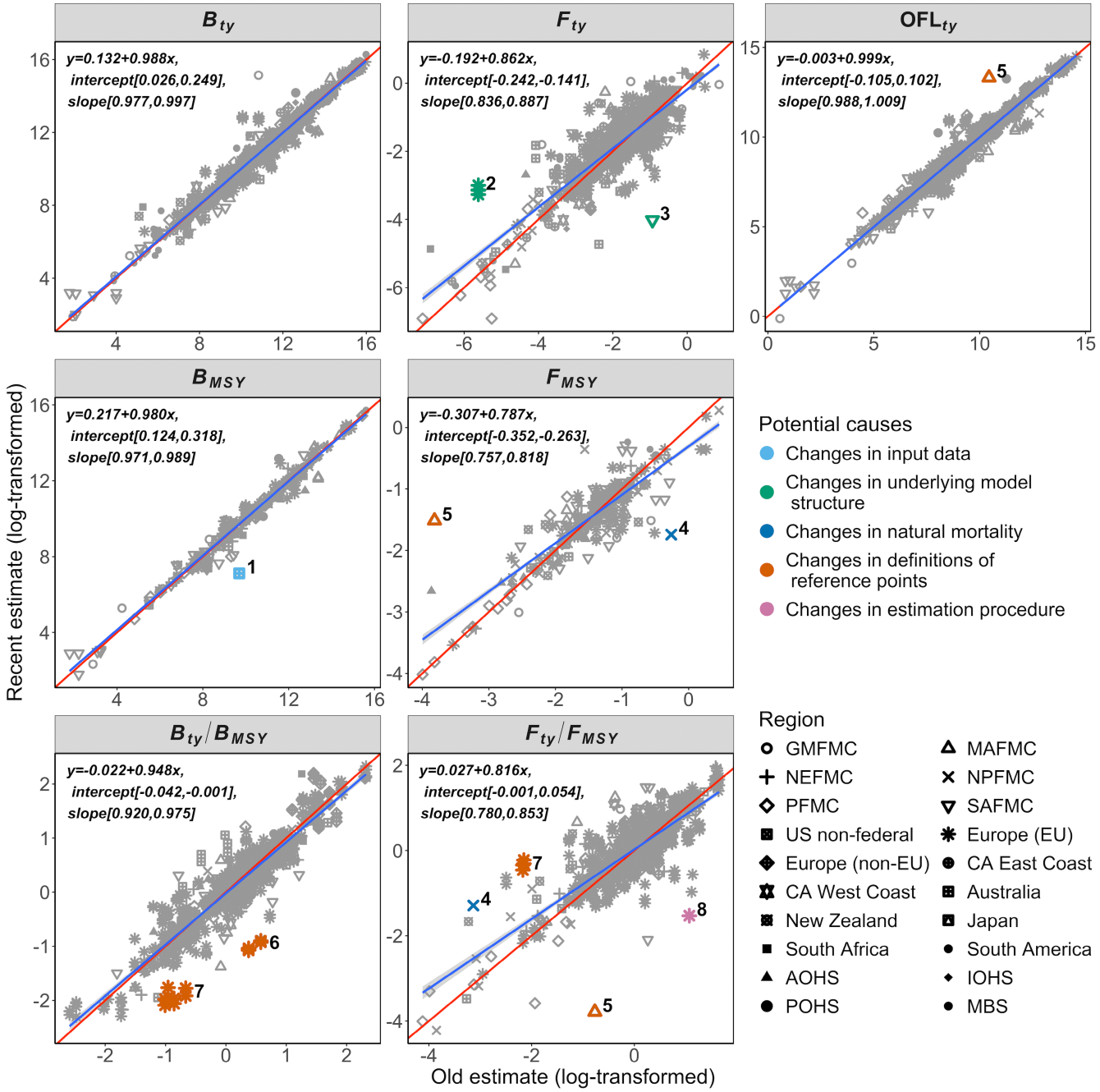


FIGURE 2 Estimates from more recent assessment against estimates from older assessment. Estimates are on log scale. The red line in each panel is the 1:1 line. The blue line in each panel is the fitted linear regression line. The equation is the fitted linear regression with posterior means of intercept and slope. The values in the brackets are the 95% credible intervals for intercept and slope. The plot numbers correspond to stocks with obvious variations in model estimates from different assessments: (1) Bight redfish in the Southeast Australia; (2) Herring in the ICES 5a-7bc, managed by the International Council of the Exploration of the Sea (ICES); (3) Blueline tilefish in the US South Atlantic, managed by the South Atlantic Fishery Management Council (SAFMC); (4) Alaska plaice in the Bering Sea and Aleutian Islands, managed by the North Pacific Fishery Management Council (NPFMC); (5) Ocean quahog in the Atlantic Coast, managed by the Mid-Atlantic Fishery Management Council (MAFMC); (6) Atlantic cod in the Western Baltic, managed by the ICES; (7) Whiting in the West of Scotland, managed by the ICES; (8) Atlantic cod in the Irish Sea, managed by the ICES. Abbreviations for management councils or regions represent the following: GMFMC, Gulf of Mexico Fishery Management Council; MAFMC, Mid-Atlantic Fishery Management Council; NEFMC, New England Fishery Management Council; NPFMC, North Pacific Fishery Management Council; PFMC, Pacific Fishery Management Council; SAFMC, South Atlantic Fishery Management Council; CA, Canada; AOHS, Atlantic Ocean High Seas; IOHS, Indian Ocean High Seas; POHS, Pacific Ocean High Seas; MBS, Mediterranean-Black Sea.

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Global mean CV for each model estimate (results only for the selected model). Median values and 95% credible intervals (in the brackets) are listed

Estimate	Selected model	Global mean CV (%)
B_{ty}	M3-a	39 (29, 52)
F_{ty}	M3-a	46 (36, 59)
B_{MSY}	M3-a	31 (21, 49)
F_{MSY}	M3-a	23 (14, 38)
B_{ty}/B_{MSY}	M3-b	25 (18, 35)
F_{ty}/F_{MSY}	M3-b	43 (29, 64)
OFL_{ty}	M3-b	45 (31, 69)

3.3 | Inter-assessment uncertainty in stock status

Comparison of stock status (i.e., whether a stock is overfished or experiencing overfishing) determined in the recent assessment and the B_{ty}/B_{MSY} and F_{ty}/F_{MSY} estimated in the older assessment revealed considerable uncertainty but was centred on the nominal values. For example, when B_{ty}/B_{MSY} from the older assessment equalled 0.5, that is the critical value used in the US to define if a stock is overfished, there was a 48% probability (95% CI: 42%–54%) that the stock was overfished based on the more recent assessment (Figure 5a). When B_{ty}/B_{MSY} from the older assessment equalled 1, indicating that the stock was not overfished, there was still a 7% probability (95% CI: 5%–10%) that the stock was overfished based on the more recent assessment (Figure 5a). When F_{ty}/F_{MSY} from the older assessment equalled 1, there was a 49% probability (95% CI: 45%–53%) that the stock was experiencing overfishing based on the more recent assessment (Figure 5b). When F_{ty}/F_{MSY} from the older assessment equalled 0.5, indicating the stock was not experiencing overfishing, there was still a 15% probability (95% CI: 11%–18%) that the stock was experiencing overfishing based on the more recent assessment (Figure 5b).

4 | DISCUSSION

This study quantifies inter-assessment uncertainty around management-relevant model outputs from multiple stock assessments in different regions. Uncertainty differed by assessment estimates and regions. The OFL_{ty} was a most uncertain model output because the variation in OFL_{ty} captures both uncertainties in B_{ty} and F_{MSY}. The F_{MSY} estimate was the least uncertain assessment output. The variations in B_{ty}/B_{MSY} and F_{ty}/F_{MSY} reflect uncertainties in B_{ty} and B_{MSY}, and uncertainties in F_{ty} and F_{MSY}, but variations may be counteracted when both B_{ty} and B_{MSY}, or both F_{ty} and F_{MSY}, change in the same direction, which potentially explain why the global mean CV of B_{ty}/B_{MSY} was lower than those of B_{ty} and B_{MSY} (Figure S2 in the Appendix S1).

4.1 | Overall variability in estimates of stock status on overfished and overfishing

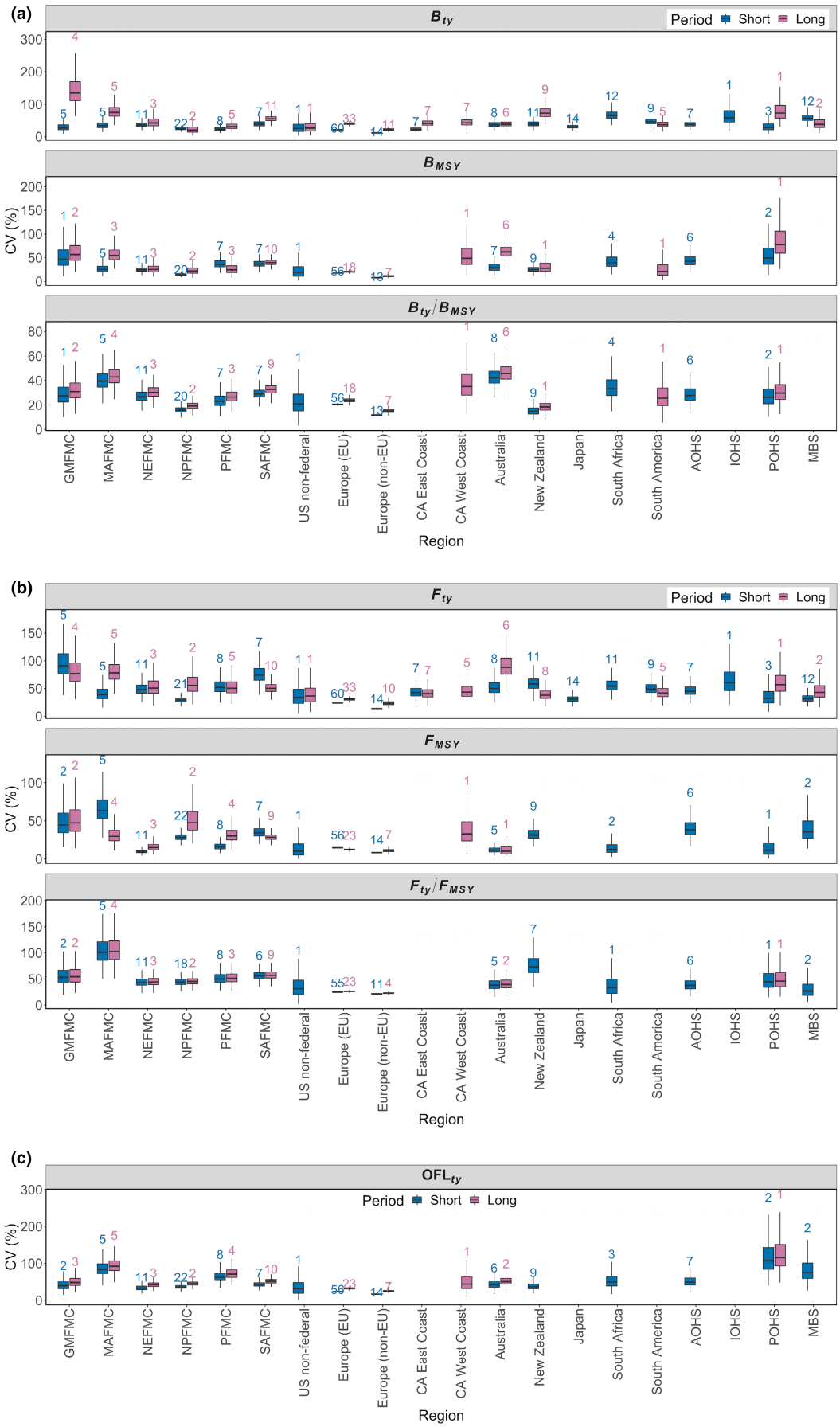
Stock status determinations from two assessments for each stock may be different. The probability for the newer assessment to determine that the stock is overfished or experiencing overfishing is close to 50% when $B_{ty}/B_{MSY} = 0.5$ or $F_{ty}/F_{MSY} = 1$ in the older assessment (Figure 5). However, when the older assessment determines that the stock is not overfished nor experiencing overfishing, there is still a non-zero probability that the newer assessment determines that the stock was actually overfished or was experiencing overfishing (Figure 5). The variability in estimates of stock status with respect to overfished status and overfishing reveals the importance of incorporating scientific uncertainty in fisheries management.

4.2 | Implications for fisheries management

Fishery management bodies currently use a variety of different approaches to set scientific uncertainty buffers. For example, the MAFMC classifies stocks into three uncertainty categories based on scoring of a table of stock attributes. These categories correspond to assumed CVs of the OFL of 60%, 100%, or 150%. The PFMC used Ralston et al. (2011) meta-analysis of historical time-series of spawning biomass estimates to define a lower bound on the uncertainty buffer, but uses uncertainty estimates directly from the stock assessment if they exceed this value. In practice, a minimum CV of 36% is used for data-rich stocks (Category 1), and a higher value of 50% is used for an extra buffer for staleness. Data-limited and data-poor stocks (Categories 2 and 3) are deemed to have increasing levels of uncertainty and therefore higher CVs. The US New England Fishery Management Council (NEFMC) uses 75% of the F_{MSY} proxy to set the ABC, that is $ABC_t = 0.75 \times F_{MSY} \times B_t$, which corresponds to an equivalent OFL CV of approximately 162% with a $p^* = .4$ (Figure S3 in the Appendix S1). If the stock is in a rebuilding plan, the NEFMC

FIGURE 3 CV estimates for the seven assessment estimates. (a) CV estimates for biomass-related assessment estimates (B_{ty}, B_{MSY}, B_{ty}/B_{MSY}) from the best models. (b) CV estimates for fishing mortality-related assessment estimates (F_{ty}, F_{MSY}, F_{ty}/F_{MSY}) from the best models. (c) CV estimates for OFL_{ty} from the best model. Whiskers, boxes, and horizontal middle lines are 95% and 50% interquartile ranges, and medians of posterior distributions. Blue and pink boxplot shading correspond to short and long-time period, respectively. Blue and pink numbers show the number of stocks with estimates available for each time period group in each region, respectively. Groups in which there are zero stocks with the corresponding model estimate are not displayed. Abbreviations for management councils or regions are defined in the caption for Figure 2.

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

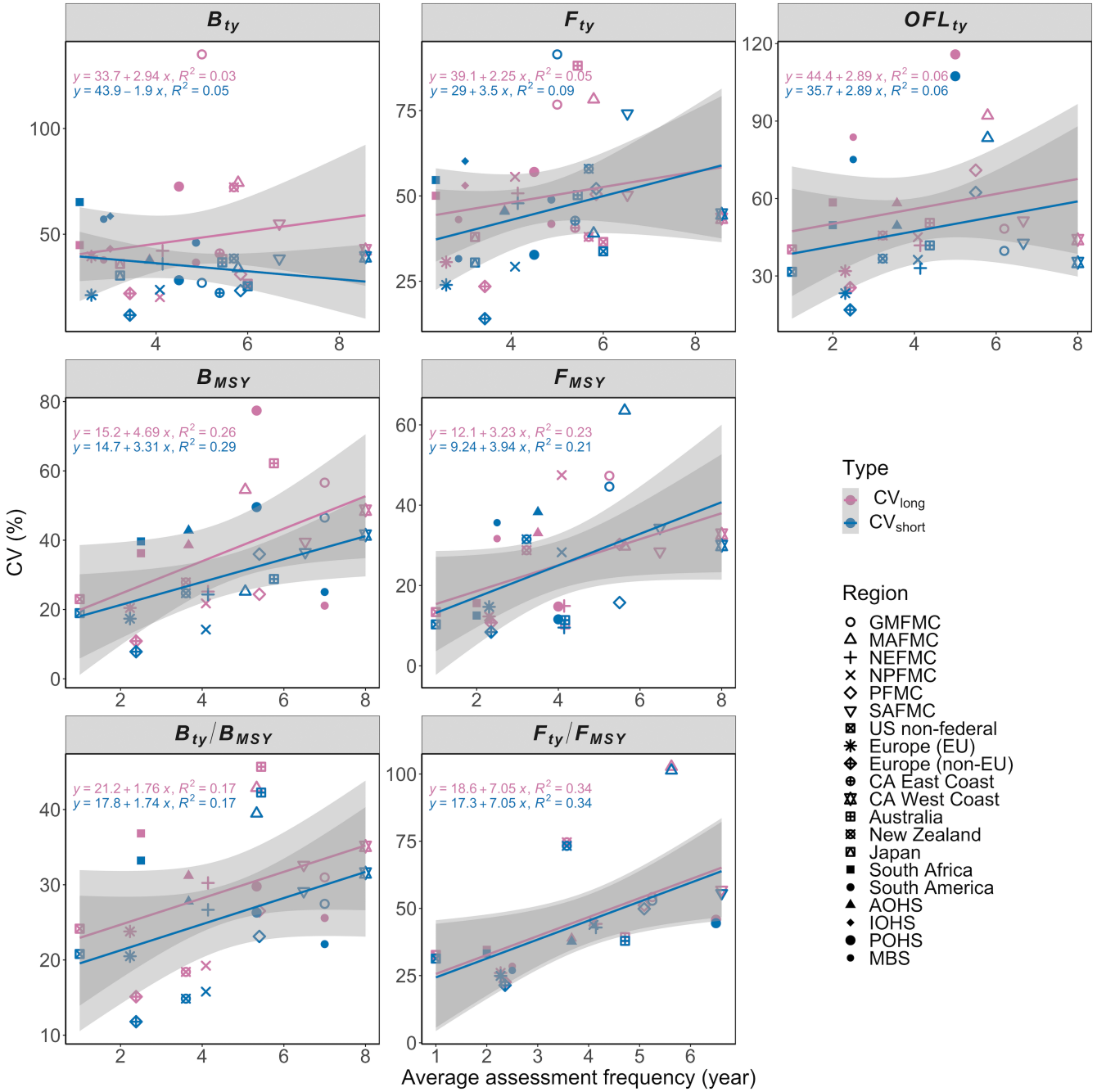


FIGURE 4 Regional CV estimates compared to the average time interval between assessment updates in each region. For each model estimate, a linear regression is fitted to CV_{long} and CV_{short} estimates, separately. Abbreviations for management councils or regions are defined in the caption for Figure 2.

uses the lesser of 75% F_{MSY} or $F_{rebuild}$, a fishing mortality rate associated with a specific rebuilding trajectory. For depleted stocks that are projected to increase, the ABC is fixed across years at the ABC estimated in the first year of the projection interval. The North Pacific Fishery Management Council (NPFMC) once empowered assessment authors to decide the buffer individually based on their expert opinion, and now assessment authors create a risk table, and the most "risky" score determines the buffer automatically. The SAFMC uses a Monte Carlo Bootstrap Ensemble approach to estimate the uncertainty associated with the OFL and the p^* distribution is set based on assessment information, productivity and susceptibility analysis of the stock, stock status, and uncertainty characterization. Our analyses of the data-rick stocks in the RAMLDB reveal that the OFL CVs calculated from interassessment uncertainty are 83% and 62% (CV_{short}) and 92% and 71% (CV_{long}) for the MAFMC and PFMC, respectively. These CVs are greater than the currently

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

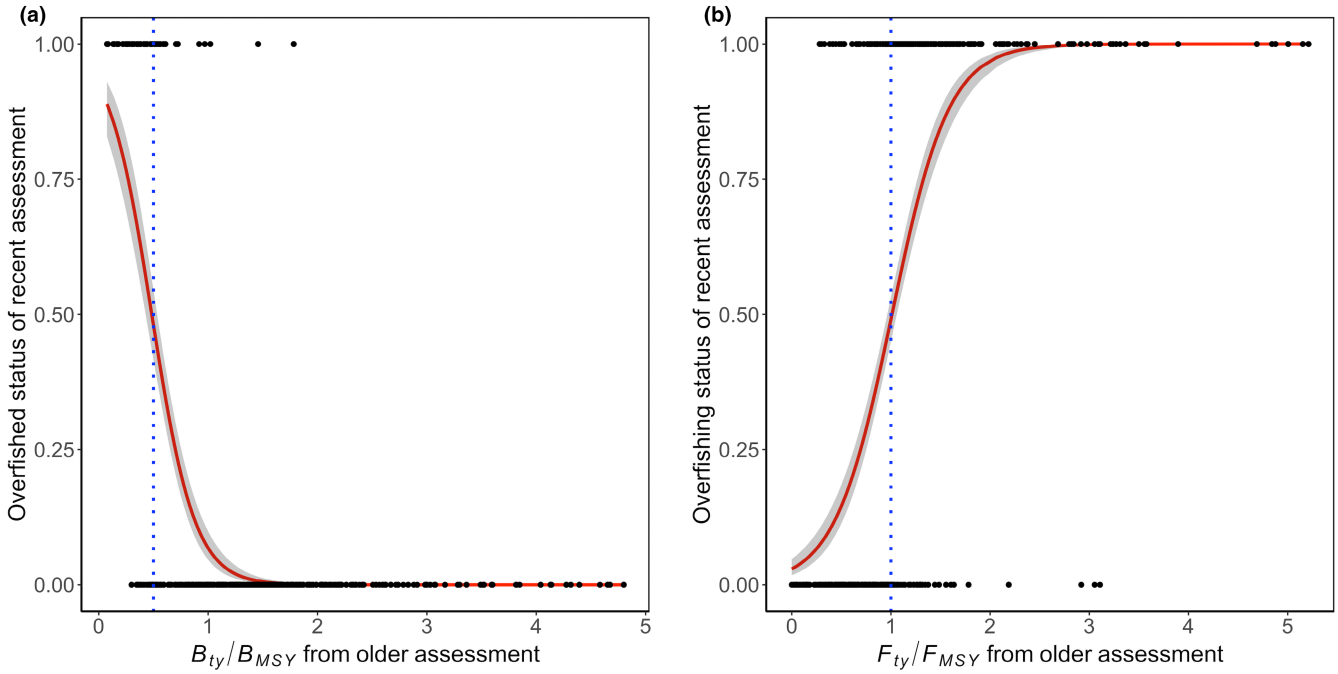


FIGURE 5 Inter-assessment variation in threshold-based categorization of stock status. (a) The probability that a subsequent assessment will consider the stock to have been overfished for different values of B_{ty} / B_{MSY}. (b) The probability that a subsequent assessment will consider the stock to have been experiencing overfishing for different values of F_{ty} / F_{MSY}. Pairwise comparisons on stock status among assessments are conducted for each stock. For panel a, the *x*-axis is B_{ty} / B_{MSY} from the older assessment in a pairwise comparison, *y*-axis is a binary outcome that equals 0 if B_{ty} / B_{MSY} from the newer assessment in the pairwise comparison ≥0.5, equals 1 if B_{ty} / B_{MSY} from the newer assessment in the pairwise comparison <0.5 (indicating overfished). For panel b, the *x*-axis is F_{ty} / F_{MSY} from the older assessment in a pairwise comparison, *y*-axis is a binary outcome that equals 0 if F_{ty} / F_{MSY} from the newer assessment in the pairwise comparison ≤1, equals 1 if F_{ty} / F_{MSY} from the newer assessment in the pairwise comparison >1 (indicating overfishing). To get a better visualization, there are 24 points at B_{ty} / B_{MSY} > 5 that are not shown in the panel a. The red curves are the fitted logistic regression curves, and the grey ribbons are the 95% confidence intervals. The blue dashed vertical lines separate the standard breakpoints for overfished and overfishing status (0.5 in panel a, and 1 on panel b).

used minimum CVs in these regions suggesting a possible need to reconsider the minimum CV values used to determine the buffer for scientific uncertainty in the OFL.

Although we estimated region-specific measures of assessment uncertainty, the CVs reported here should be considered a lower bound in the context of setting management advice for a given stock in a region. There is an additional component of the catch-setting process that can lead to scientific uncertainty in achieving management objectives. We focused on the terminal estimates from an assessment, but projections are typically used to calculate the OFL and ABC for a number of years in the future. Projections are generally even more uncertain than terminal year estimates (Wiedenmann & Jensen, 2018). Uncertainty in the terminal year estimates can have a large impact on the accuracy of the projections, but the projections also rely on assumptions about stock productivity in the future (e.g., recruitment, growth). Deviations in future productivity from what was assumed in the projection (e.g., poorer than average recruitment) can lead to large differences in the projected and realized stock size or OFL (Wiedenmann & Jensen, 2018). Our analyses were based on data-rich stocks from the RAMLDB, and greater uncertainties are expected for data-moderate and data-poor stocks (Ralston et al., 2011). Our models assumed that the truth was essentially the mean estimate and assessment results were random and independent samples from an underlying distribution. Other options, such as that the last assessment is the best, are possible. Models that set the model estimate from the last assessment as the mean value of the underlying distribution resulted in similar but slightly greater CV estimates (Table S4 in the Appendix S1).

4.3 | Factors associated with higher variability

Previous studies have shown that there are numerous potential causes for variations in stock assessment outputs over time (Hurtado-Ferro et al., 2015; Punt et al., 2018; Ralston et al., 2011; Silvar-Viladomiu et al., 2021; Wiedenmann & Jensen, 2018). Potential factors include, but are not limited to, changes in model assumptions or structure (e.g., age-aggregated or age-structured, changes in the shape of the selectivity curve, fixed vs. time-varying parameters like catchability), changes in data inputs (e.g., revised survey data, changed time-series catch data, changed time-period of data input), changes in how a stock is defined spatially (which impacts the specific inputs), changes in life history information (e.g., natural mortality, length- and mass-at-age), or changes in how reference points are defined (Hurtado-Ferro et al., 2015; Magnusson

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

& Hilborn, 2007; Punt et al., 2002; Silvar-Viladomiu et al., 2021; Wiedenmann & Jensen, 2018). These changes between assessments produce variations in assessment outputs and represent multiple forms of scientific uncertainty.

A full exploration of each factor and its impact on inter-assessment variation in model estimates is beyond the scope of this work, but we provide some examples. For herring in the ICES 5a-7bc area, the assessment model changed from a trends-based exploratory assessment to an age-based analytical assessment, resulting in an approximately 75% decrease in F_{ty} from the 2010 assessment to the 2015 assessment. For ocean quahog in the US Atlantic Coast, F_{MSY} was revised from $F_{25\%}$ (i.e., the fishing mortality rate that reduces lifetime egg production to 25% of its potential) to a more conservative reference point ($F_{45\%}$, egg production at 45% of potential) in 2009 (Chute et al., 2013), which led to large changes in the OFL as well as the reference points. For Alaska plaice in the Bering Sea and Aleutian Islands, the natural mortality rate was re-estimated and a fixed $M = 0.13\ y^{-1}$ was used for both sexes in 2010, in comparison with $M = 0.25\ y^{-1}$ in the previous assessments (Wilderbuer et al., 2010). Many stocks in Europe also experienced changes in reference points. For example, for Atlantic cod in the Western Baltic, advised reference points were based on the EU management plan (EC 1098/2007) in the 2012 and 2013 assessments, and changed to be based on the MSY approach in assessments from 2014 (ICES, 2012a, 2013b, 2014); for Whiting (*Merlangius merlangus*, Gadidae) in the West of Scotland, the basis of advised reference points changed from the precautionary approach in assessments from 2012 to 2015 to the MSY approach in assessments of 2016 and 2017 (ICES, 2012b, 2015b, 2016). A previous study found that the reference point definition and the technical basis for estimation were the most important reason for reference point changes (Silvar-Viladomiu et al., 2021).

High inter-assessment uncertainty can also reflect a culture of willingness to revise the assumptions and reanalyse the underlying assessment data within an assessment, or a willingness of managers to adjust the target reference points (see examples above). For example, in the US Northeast, South Atlantic, Gulf of Mexico and Caribbean regions, assessments for a given stock are grouped into two broad categories. In the first category, called management track assessments (previously called updates), the existing model structure remains the same but the models are updated with more recent data. In research track assessments (previously called benchmark assessments), a wide range of changes to the model may be explored. Although large changes can occur between assessments with just updated data (e.g., Wiedenmann & Jensen, 2018), it is more likely that the large-scale changes to the model structure which may happen during research track assessments will result in greater changes in estimates between assessments. We did not characterize the degree of changes across assessment in this work, but larger changes typically require more time between assessments, and our finding that there was greater uncertainty in model estimates for longer periods between assessments (Figure 4), possibly reflecting the degree of change between successive models. As such, inter-assessment CVs cannot be interpreted as an index of assessment quality.

Dramatic ecological changes within an ecosystem can also lead to greater uncertainty among assessment estimates. For example, simulation models have shown that differences between the assumed and the true natural mortality rate can lead to retrospective patterns in sequential model estimates (Hurtado-Ferro et al., 2015; Mohn, 1999). Changing environmental conditions, such as warming, could impact survival, abundance, or productivity of a large number of stocks within a region (e.g., Hare et al., 2016; Pershing et al., 2015), and such changes could result in disconnects between assessment model assumptions and signals in the data used in the model fitting (e.g., Wiedenmann & Legault, 2022). Changes in the abundance of top generalist predators can also have wide-ranging impacts on the survival of many stocks within a region (e.g., Swain & Benoît, 2015). Exploration of the roles of these and other environmental factors on assessment uncertainty is warranted and should be a focus of future work that expands on this analysis.

In summary, we have quantified region-specific uncertainties in estimates of biomass, fishing mortality, reference points, and relative biomass and fishing mortality rate, as well as OFL among assessments for the same stock. This study presents one method of comparing uncertainty among assessments and provides a base for determining the minimum buffer for scientific uncertainty. Which climatic, environmental, ecological, and assessment-related factors best predict assessment performance remains unclear, but should be a focus of future empirical analyses based on these results.

ACKNOWLEDGMENTS

We thank all members of the Jensen Lab at the University of Wisconsin-Madison's Center for Limnology for valuable feedback on this work. We appreciate the insightful discussions of this topic with Paul Rago, Martin Dorn, Michael Wilberg, Chris Anderson, Pat Sullivan, and Ian Stewart. This study is financially supported by the Lenfest Ocean Program.

CONFLICT OF INTEREST

No authors have competing interests.

DATA AVAILABILITY STATEMENT

RAM Legacy Stock Assessment Database is available from https://www.ramlegacy.org/. Additional assessment documents for stocks not included in RAM Legacy are available as follows: in the US Southeast and Gulf: http://sedarweb.org/sedar-projects; US Alaska: https://www.fisheries.noaa.gov/alaska/population-assessments/alaska-stock-assessments; US Northeast: https://apps-nefsc.fisheries.noaa.gov/rcb/publications/assessment-documents.html; US West Coast:https://www.pcouncil.org/stock-assessments-star-reports-stat-reports-rebuilding-analyses-terms-of-reference/groundfish-stock-assessment-documents/. Europe (ICES region): https://www.ices.dk/sites/pub/Publication-Reports.

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ORCID

Rujia Bi https://orcid.org/0000-0002-1954-6797
Katherine E. Mills https://orcid.org/0000-0001-6078-7747

REFERENCES

Brodziak, J., & Walsh, W. A. (2013). Model selection and multimodel inference for standardizing catch rates of bycatch species: A case study of oceanic whitetip shark in the Hawaii-based longline fishery. *Canadian Journal of Fisheries and Aquatic Sciences*, *70*(12), 1723–1740. https://doi.org/10.1139/cjfas-2013-0111

Brooks, E. N., & Deroba, J. J. (2015). When "data" are not data: The pitfalls of post hoc analyses that use stock assessment model output. *Canadian Journal of Fisheries and Aquatic Sciences*, *72*(4), 634–641. https://doi.org/10.1139/cjfas-2014-0231

Brooks, E. N., & Legault, C. M. (2016). Retrospective forecasting — Evaluating performance of stock projections for New England groundfish stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, *73*(6), 935–950. https://doi.org/10.1139/cjfas-2015-0163

Charles, A. T. (1998). Living with uncertainty in fisheries: Analytical methods, management priorities and the Canadian groundfishery experience. *Fisheries Research*, *37*(1–3), 37–50. https://doi.org/10.1016/S0165-7836(98)00125-8

Chute, A., Hennen, D., Russell, R., & Jacobson, L. (2013). *Stock assessment update for ocean quahogs (Arctica islandica) through 2011*. National Marine Fisheries Service Technical Report NEFSC-13-17.

Conn, P. B., Williams, E. H., & Shertzer, K. W. (2010). When can we reliably estimate the productivity of fish stocks? *Canadian Journal of Fisheries and Aquatic Sciences*, *67*(3), 511–523. https://doi.org/10.1139/F09-194

Da-Rocha, J. M., Garcia-Cutrin, J., & Gutierrez, M. J. (2016). *Harvesting control rules that deal with scientific uncertainty*. University Library of Munich.

Denwood, M. J. (2016). Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(1), 1–25. https://doi.org/10.18637/jss.v071.i09

Dettloff, K. (2020). Uncertainty in National Marine Fisheries Service Stock Assessments. https://doi.org/10.25923/ym2v-9802

Dorn, M. W., & Zador, S. G. (2020). A risk table to address concerns external to stock assessments when developing fisheries harvest recommendations. *Ecosystem Health and Sustainability*, *6*(1), 1813634. https://doi.org/10.1080/20964129.2020.1813634

Edwards, C. T. T. (2016). *Feedback control and adaptive Management in Fisheries*. Routledge.

Federal Register. (2009). *Magnuson-Stevens act provisions; annual catch limits; national standard guidelines*. https://www.federalregister.gov/documents/2009/01/16/E9-636/magnuson-stevens-act-provisions-annual-catch-limits-national-standard-guidelines

Federal Register. (2016). *Magnuson-Stevens act National Provisions; national standard guidelines*. https://www.federalregister.gov/documents/2016/10/18/2016-24500/magnuson-stevens-act-provisions-national-standard-guidelines

Francis, R. I. C. C., & Shotton, R. (1997). "Risk" in fisheries management: A review. *Canadian Journal of Fisheries and Aquatic Sciences*, *54*(8), 1699–1715. https://doi.org/10.1139/f97-100

Garcia, S. M. (2000). The precautionary approach to fisheries: Progress review and main issues (1995–2000). In M. N. Nordquist & J. N. Moore (Eds.), *Current fisheries issues and the food and agriculture organization of the united nations* (pp. 479–560). Center for Oceans Law.

Gelman, A., & Rubin, D. B. (1992). A single sequence from the Gibbs sampler gives a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 625–631). Oxford University Press.

Giudici, P., & Castelo, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, *50*(1), 127–158. https://doi.org/10.1023/A:1020202028934

Haddon, M. (2016). Bight redfish (*Centroberyx gerrardi*) stock assessment using data to 2014/2015. In G. N. Tuck (Eds.), *Stock assessment for the southern and eastern Scalefish and shark fishery 2015, part 1. Australian fisheries management authority and CSIRO oceans and atmosphere* (pp. 9–50). Australian Fisheries Management Authority and CSIRO Oceans and Atmosphere.

Hanna, S. S. (1997). The new frontier of American fisheries governance. *Ecological Economics*, *20*(3), 221–233. https://doi.org/10.1016/S0921-8009(96)00082-1

Hare, J. A., Morrison, W. E., Nelson, M. W., Stachura, M. M., Teeters, E. J., Griffis, R. B., Alexander, M. A., Scott, J. D., Alade, L., Bell, R. J., Chute, A. S., Curti, K. L., Curtis, T. H., Kircheis, D., Kocik, J. F., Lucey, S. M., McCandless, C. T., Milke, L. M., Richardson, D. E., … Griswold, C. A. (2016). A vulnerability assessment of fish and invertebrates to climate change on the northeast U.S. continental shelf. *PLoS One*, *11*(2), e0146756. https://doi.org/10.1371/journal.pone.0146756

Hilborn, R. (1987). Living with uncertainty in resource management. *North American Journal of Fisheries Management*, *7*(1), 1–5. https://doi.org/10.1577/1548-8659(1987)7<1:LWUIRM>2.0.CO;2

Hilborn, R., Amoroso, R. O., Anderson, C. M., Baum, J. K., Branch, T. A., Costello, C., de Moor, C. L., Faraj, A., Hively, D., Jensen, O. P., Kurota, H., Little, L. R., Mace, P., McClanahan, T., Melnychuk, M. C., Minto, C., Osio, G. C., Parma, A. M., Pons, M., … Ye, Y. (2020). Effective fisheries management instrumental in improving fish stock status. *Proceedings of the National Academy of Sciences*, *117*(4), 2218–2224. https://doi.org/10.1073/pnas.1909726116

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

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., Ono, K., Vert-Pre, K. A., Whitten, A. R., & Punt, A. E. (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. https://doi.org/10.1093/icesjms/fsu198

ICES. (2012a). *Report of the Baltic Fisheries Assessment Working Group (WGBFAS). ICES Headquarters, 12–19 April 2012. ICES CM 2012/ACOM:10*.

ICES. (2012b). *Report of the Working Group on Celtic Seas Ecosystems (WGCSE), 9–18 May 2012, Copenhagen, Denmark. ICES CM 2012/ACOM:12*.

ICES. (2013a). *Report of the Herring Assessment Working Group for the Area South of 62°N, 13–22 March 2013. ICES CM 2013/ACOM:06*.

ICES. (2013b). *Report of the Baltic Fisheries Assessment Working Group (WGBFAS). ICES Headquarters, 10–17 April 2013. ICES CM 2013/ACOM:10*.

ICES. (2014). *Report of the Baltic Fisheries Assessment Working Group (WGBFAS), ICES Headquarters, 3–10 April 2014. ICES CM 2014/ACOM:10*.

ICES. (2015a). *Report of the Herring Assessment Working Group for the Area South of 62°N 10–19 March 2015. ICES CM 2015/ACOM:06*.

ICES. (2015b). *Report of the working Group for the Celtic Seas Ecoregion (WGCSE), Copenhagen, Denmark. ICES CM 2015/ACOM:12*.

ICES. (2016). *Report of the Working Group for the Celtic Seas Ecoregion (WGCSE), 04–13 May 2016, ICES Headquarters, Copenhagen, Denmark. ICES CM 2016/ACOM:13*.

ICES. (2017). *Report of the Benchmark Workshop on the Irish Sea Ecosystem (WKIrish3), 30 January–3 February 2017, Galway, Ireland. ICES CM 2017/BSG:01*.

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Legault, C.M. (2009). *Report of the Retrospective Working Group, January 14–16, 2008, Woods Hole, Massachusetts. U.S. Department of Commerce, Northeast Fisheries Science Center Ref Doc. 09–01; 30 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543–1026.*

MAFMC. (2011). *Omnibus amendment*. Mid-Atlantic Fisheries Management Council. https://www.greateratlantic.fisheries.noaa.gov/nero/regs/frdoc/11/11OmnibusAmendmentEA&CommentsFinal.pdf

Magnusson, A., & Hilborn, R. (2007). What makes fisheries data informative? *Fish and Fisheries*, *8*(4), 337–358. https://doi.org/10.1111/j.1467-2979.2007.00258.x

Mildenberger, T. K., Berg, C. W., Kokkalis, A., Hordyk, A. R., Wetzel, C., Jacobsen, N. S., Punt, A. E., & Nielsen, J. R. (2022). Implementing the precautionary approach into fisheries management: Biomass reference points and uncertainty buffers. *Fish and Fisheries*, *23*, 73–92. https://doi.org/10.1111/faf.12599

Mohn, R. (1999). The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. *ICES Journal of Marine Science*, *56*(4), 473–488. https://doi.org/10.1006/jmsc.1999.0481

Parma, A. M., Sullivan, P. J., Collie, J., Hartley, T. W., Heyman, W., Johnson, R., Punt, A. E., Rose, K. A., Sanchiro, J., Sissenwine, M. P., & Sugihara, G. (2013). *Evaluating the effectiveness of fish stock rebuilding in the United States*. National Academies Press.

Pershing, A. J., Alexander, M. A., Hernandez, C. M., Kerr, L. A., Le Bris, A., Mills, K. E., Nye, J. A., Record, N. R., Scannell, H. A., Scott, J. D., Sherwood, G. D., & Thomas, A. C. (2015). Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. *Science*, *350*(6262), 809–812. https://doi.org/10.1126/science.aac9819

PFMC. (2010). FMP amendment addresses National Standard 1 – Annual catch limits, accountability measures [online]. Pacific Fishery Management Council. *Pacific Council News*, 34(1), 5. http://www.pcouncil.org/wp-content/uploads/Spring_2010_Newsletter.pdf.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*. Technische Universität Wien. https://www.R-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Plummer, M. (2016). *rjags: Bayesian Graphical Models Using MCMC*. R package version 4–6. https://CRAN.R-project.org/package=rjags

Prager, M. H., & Shertzer, K. W. (2010). Deriving acceptable biological catch from the overfishing limit: Implications for assessment models. *North American Journal of Fisheries Management*, *30*(1), 289–294. https://doi.org/10.1577/M09-105.1

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

Privitera-Johnson, K. M., & Punt, A. E. (2020b). Leveraging scientific uncertainty in fisheries management for estimating among-assessment variation in overfishing limits. *ICES Journal of Marine Science*, *77*(2), 515–526. https://doi.org/10.1093/icesjms/fsz237

Punt, A. E., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A. A., & Haddon, M. (2016). Management strategy evaluation: Best practices. *Fish and Fisheries*, *17*(2), 303–334. https://doi.org/10.1111/faf.12104

Punt, A. E., Day, J., Fay, G., Haddon, M., Klaer, N., Little, L. R., & Wayte, S. (2018). Retrospective investigation of assessment uncertainty for fish stocks off Southeast Australia. *Fisheries Research*, *198*, 117–128. https://doi.org/10.1016/j.fishres.2017.10.007

Punt, A. E., Smith, A. D. M., & Cui, G. (2002). Evaluation of management tools for Australia's south east fishery 2. How well can management quantities be estimated? *Marine and Freshwater Research*, *53*(3), 631–644.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ralston, S., Punt, A. E., Hamel, O. S., DeVore, J. D., & Conser, R. J. (2011). A meta-analytic approach to quantifying scientific uncertainty in stock assessments. *Fishery Bulletin*, *109*(2), 217–231.

Ricard, D., Minto, C., Jensen, O. P., & Baum, J. K. (2012). Examining the knowledge base and status of commercially exploited marine species with the RAM legacy stock assessment database. *Fish and Fisheries*, *13*(4), 380–398. https://doi.org/10.1111/j.1467-2979.2011.00435.x

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

SEDAR. (2017). *SEDAR 50 – Atlantic blueline tilefish assessment report* (p. 542). SEDAR. http://sedarweb.org/sedar-50

Sethi, S. A. (2010). Risk management for fisheries. *Fish and Fisheries*, *11*(4), 341–365. https://doi.org/10.1111/j.1467-2979.2010.00363.x

Shertzer, K. W., Prager, M. H., & Williams, E. H. (2008). A probability-based approach to setting annual catch limits. *Fishery Bulletin*, *106*(3), 225–232.

Shertzer, K. W., Prager, M. H., & Williams, E. H. (2010). Probabilistic approaches to setting acceptable biological catch and annual catch targets for multiple years: Reconciling methodology with National Standards Guidelines. *Marine and Coastal Fisheries*, *2*(1), 451–458. https://doi.org/10.1577/C10-014.1

Silvar-Viladomiu, P., Minto, C., Halouani, G., Batts, L., Brophy, D., Lordan, C., & Reid, D. G. (2021). Moving reference point goalposts and implications for fisheries sustainability. *Fish and Fisheries*, *22*, 1345–1358. https://doi.org/10.1111/faf.12591

Sissenwine, M. (1984). The uncertain environment of fishery scientists and managers. *Marine Resource Economics*, *1*(1), 1–30.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (statistical methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stewart, I. J., & Hicks, A. C. (2018). Interannual stability from ensemble modelling. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(12), 2109–2113. https://doi.org/10.1139/cjfas-2018-0238

Swain, D. P., & Benoît, H. P. (2015). Extreme increases in natural mortality prevent recovery of collapsed fish populations in a Northwest Atlantic ecosystem. *Marine Ecology Progress Series*, *519*, 165–182. https://doi.org/10.3354/meps11012

Van Beveren, E., Benoît, H. P., & Duplisea, D. E. (2021). Forecasting fish recruitment in age-structured population models. *Fish and Fisheries*, *22*(5), 941–954. https://doi.org/10.1111/faf.12562

Vehtari, A., & Gelman, A. (2015). Pareto smoothed importance sampling. *arXiv*. https://doi.org/10.48550/arXiv.1507.02646

Vehtari, A., Gelman, A., & Gabry, J. (2016). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 0.1.6. https://github.com/stan-dev/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.48550/arXiv.1507.04544

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594. https://doi.org/10.48550/arXiv.1004.2316

Wiedenmann, J., & Jensen, O. P. (2018). Uncertainty in stock assessment estimates for New England groundfish and its impact on achieving target harvest rates. *Canadian Journal of Fisheries and*

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Aquatic Sciences, *75*(3), 342–356. https://doi.org/10.1139/cjfas-2016-0484

Wiedenmann, J., & Legault, C. M. (2022). Something strange in the neighborhood: Diverging signals in stock assessment data for Northeast U.S. fish stocks. *Fisheries Management and Ecology*, *29*, 269–285. https://doi.org/10.1111/fme.12532

Wilderbuer, T. K., Nichol, D. G., & Spencer, P. D. (2010). Assessment of the Alaska plaice stock in the Bering Sea and Aleutian Islands. In *Stock assessment and fishery evaluation document for groundfish resources in the Bering Sea/Aleutian Islands region*. North Pacific Fishery Management Council. https://apps-afsc.fisheries.noaa.gov/REFM/docs/2010/BSAIplaice.pdf

Yin, Y., & Sampson, D. B. (2004). Bias and precision of estimates from an age-structured stock assessment program in relation to stock and data characteristics. *North American Journal of Fisheries Management*, *24*(3), 865–879. https://doi.org/10.1577/M03-107.1

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bi, R., Collier, C., Mann, R., Mills, K. E., Saba, V., Wiedenmann, J., & Jensen, O. P. (2023). How consistent is the advice from stock assessments? Empirical estimates of inter-assessment bias and uncertainty for marine fish and invertebrate stocks. *Fish and Fisheries*, *24*, 126–141. https://doi.org/10.1111/faf.12714

14672979, 2023, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/faf.12714 by University Of Wisconsin - Madison, Wiley Online Library on [25/07/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License