Title

The reproducibility crisis meets stock assessment science: Sources of inadvertent bias in the stock assessment prioritization and review process

Author

Satterthwaite, William H

Publication Date

2023-10-01

DOI

10.1016/j.fishres.2023.106763

Accepted manuscript version of Satterthwaite, W. H. 2023. The reproducibility crisis meets stock assessment science: Sources of inadvertent bias in the stock assessment prioritization and review process. Fisheries Research 266:106763. https://doi.org/10.1016/j.fishres.2023.106763

Title: The reproducibility crisis meets stock assessment science: Sources of inadvertent bias in the stock assessment prioritization and review process

Author: William H. Satterthwaite
Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Santa Cruz, CA, USA

Corresponding author. 110 McAllister Way, Santa Cruz, California, USA, 95060.
will.satterthwaite@noaa.gov

Highlights

- Prioritization, review, and adoption of assessments is subject to inadvertent bias
- Many of the factors introducing bias are analogous to "p-hacking" in science broadly
- This compromises the interpretation of risk metrics based on probability statements
- Bias may be comparable to differences between commonly-applied uncertainty buffers
- Solutions/mitigations proposed for p-hacking broadly have analogs for use here

Abstract

The broader scientific community is struggling with a reproducibility crisis brought on by numerous factors, including “p-hacking” or selective reporting that may increase the rate of false positives or generate misleading effect size estimates from meta-analyses. This results when multiple modeling approaches or statistical tests may be brought to bear on the same problem, and there are pressures or rewards for finding “significant” results. Fisheries science is unlikely to be immune to this problem, with numerous opportunities for bias to inadvertently enter into the process through the prioritization of stocks for assessment, decisions about competing model approaches or data treatments within complex assessment models, and decisions about whether to adopt assessments for management after they are reviewed. I present a simple simulation model of a system where many assessments are performed each management cycle for a multi-stock fishery, and show how asymmetric selection of assessments for extra scrutiny or re-assessment within a cycle can turn a process generating unbiased advice on fishing limits into one that is biased high. I show similar results when sequential assessments receive extra scrutiny if they show large proportional decreases in catch limits compared to a prior assessment for the same stock, especially if there are only small changes in true stock size or status over the interval between assessments. The level of bias introduced by a plausible level of asymmetric scrutiny is unlikely to fundamentally undermine scientific advice, but may be sufficient to compromise the nominal “overfishing probabilities” used in a common framework for accommodating uncertainty, and introduce a level of bias comparable to the difference between buffers corresponding to commonly-applied levels of risk tolerance.

Keywords: Reproducibility, bias, p-hacking, unintended consequences

1. Introduction

While reproducibility has long been a cornerstone of science, in the last decade there has been an explosion of references to a reproducibility "crisis" (Fanelli 2018). This has led to alarming suggestions that most published scientific studies are false (Ioannidis 2005), or that meta-analyses synthesizing average effect sizes across studies may yield misleading results (Schooler 2011). Other authors have taken more nuanced views suggesting that the problem, while real, may cause fewer false positives (Jager and Leek 2014) or confound meta-analyses less (Head et al. 2015) than the worst-case predictions. Nevertheless, reliability and reproducibility remain major concerns across all fields of science (Baker 2016), with the potential to affect both the advancement of knowledge and specific policy and management decisions guided by scientific advice. Recent challenges to reproducibility of hot topics in marine science (e.g., Clark et al. 2020, Provencher et al. 2020) suggest that fisheries science, which is often closely linked to economically and culturally impactful policy and management decisions, is unlikely to be immune to these problems.

Although some instances of non-reproducible science may reflect data fabrication or other scientific misconduct (Viglione 2020), this is widely believed to be uncommon relative to much more prevalent issues resulting from selective reporting of statistically "significant" results, pressure to publish, and improper application of statistics or experimental design (Baker 2016). The data and models used in fisheries stock assessments vary by region and stock, but are often subject to multiple layers of oversight and review, which should make outright data fabrication or other forms of deliberate misconduct unlikely in systems with robust review processes. However, I will argue that fisheries science and stock assessments can be, and likely are, affected by forces analogous to other sources of the reproducibility crisis.

Much of the lack of reproducibility in the scientific literature may be attributable to "p-hacking" on the part of researchers (Benjamin et al. 2018), and/or selective reporting on the part of authors and journals that may only, or at least preferentially, publish statistically "significant" results (Schooler 2011). Statistical "significance" is often established in a frequentist setting via a null hypothesis testing framework, wherein results are deemed "significant" if a null model suggests less than a 5% probability (p-value) of generating a pattern at least as strong as the one observed in the data by chance alone (Wasserstein and Lazar 2016). In "p-hacking", multiple statistical tests are performed, but only the "significant" results are retained and reported, meaning that false positives are likely to be reported at a higher rate than the nominal value assigned for a single test. This is a particular problem when analysts have large datasets with multiple potential predictors and/or response variables (leading to decisions on which to include and how to weight them) and a variety of plausible statistical model forms to test. However, even in simpler situations where only a single testing approach is considered, interpretation of p-values is clouded when additional data are accumulated and tests are repeated for larger datasets without clear stopping rules (Wicherts et al. 2016).

P-hacking is not necessarily done with ill intent, rather it may reflect an innocent lack of understanding of statistics (Peng 2015) or the strong ability of post-hoc reasoning to convince scientists that whatever course of decisions led to the “significant” (and publishable, thus rewarding) analysis were the correct ones (Simmons et al. 2011). Even when the scientists performing individual studies do everything right, if journals tend to publish “significant” results while rejecting inconclusive studies, this may lead to larger mean effects in the published literature compared to the means that would be concluded from all valid studies performed, including those that did not produce “significant” results.

2. P-hacking and selective reporting: parallels in fisheries and stock assessments

Stock assessments are fundamental to scientifically informed fisheries management (Hilborn and Walters 1992, Hilborn et al. 2020) for purposes such as setting allowable catch limits. These catch limits are often set using control rules intended to maintain population abundance and spawning output levels near those expected to produce maximum sustainable yield (Melnychuk et al. 2013, Methot et al. 2014). Stock assessments are also the product of often-complex models based on imperfectly-measured data and that require numerous choices on the part of the analysts about data to include versus exclude, how to weight different data sets, parameters to fix versus estimate, and functional forms to assume (Maunder and Piner 2015, 2017). As a result, stock assessment outputs are unavoidably uncertain (Hilborn and Walters 1992, Mildenberger et al. 2022), and different but more or less equally defensible decisions on the part of the analysts (and/or reviewers, who often drive final model form in conjunction with the original analysts) could lead to different results (Ralston et al. 2011).

One approach to dealing with this uncertainty that has been adopted in multiple regions of the United States, and in somewhat similar forms in other countries, is the P*/sigma approach (Shertzer et al. 2008). This approach assumes that overfishing limits (OFLs) are estimated without bias, but with uncertainty expressed by assuming a lognormally distributed ratio between the true OFL and the assessed OFL, where the median is equal to one and the log-scale standard deviation is sigma (see Ralston et al. 2011 and Privitera-Johnson and Punt

2020a for approaches to estimating sigma). Then, an acceptable biological catch (ABC) is determined by multiplying the OFL by the P* quantile of the distribution. If all statistical assumptions were met, this would result in a P* probability that the ABC exceeds the “true” OFL that would have been estimated given perfect knowledge. When applied to multiple stocks simultaneously, this implies that a fraction P* of the ABCs established for a multi-stock fishery would be larger than the OFLs that would have been estimated given perfect knowledge, loosely analogous to the expectation that 5% of scientific results reported as significant at the $p<0.05$ level would be false positives. Just as p-hacking and selective reporting may lead to a higher than nominal false positive rate in the scientific literature, and over-estimate mean effect size, it would seem unavoidable that if analogous pressures operate in the analysis and adoption of assessments of stock status and fishing limits, the interpretation of P* may be similarly clouded.

2.1 Scope of the problem

I suggest that just like academic scientists may face pressure to produce studies with $p<0.05$ that can be published (with resultant career benefits), and journals may be more likely to publish “significant” results, stock assessors (and the review bodies that can influence the final structure of stock assessments) may face pressures to produce “favorable” results, and/or management bodies may be more likely to adopt and use assessments perceived as “less pessimistic” (Seagraves and Collins 2012).

2.1.1 Assessment prioritization

In most regions, there are far more stocks in need of assessment than there are resources to assess them. As a result, management agencies in the United States (primarily the National Marine Fisheries Service that implements many assessments and Fishery Management Councils that lead development of management responses) have adopted a comprehensive prioritization process (Methot 2015), which typically considers many factors (NMFS 2022) including economic and ecological importance, trends in survey data, time since last assessment, and status at most recent assessment. Of note, although this is only one of the many factors considered, stocks which were last assessed to be in high status are given the lowest score for the status component of their overall prioritization score, while stocks recently assessed to be in poor status are given the highest score – higher even that stocks which lack recent (or even any) assessments and with attributes associated with high vulnerability (i.e., stocks with low productivity and high susceptibility to the fishery [Cope et al. 2011]). This strategy runs the risk that a new assessment of a low-status stock means a new error in the inevitably uncertain assessment; thus, a more favorable assessment is not conclusive proof of better status. When considering repeated assessments of multiple stocks overall, this strategy could lead to an asymmetry where there may be more chances to incorrectly reverse an assignment of poor status than there are to incorrectly re-assign a stock from good status to poor. At the same time, it may divert resources from other stocks in need of assessment where an assessment might do more to address important management uncertainties (Cadrin et al. 2015).

Given limited resources for assessments and the large number of species/stocks to assess, stock assessment analysts have also developed “data-moderate” approaches that consider fewer types of input data, have less flexibility in choices among alternative assumptions, and have fewer model structure alternatives (e.g., Rudd et al. 2021), with the hope of increasing throughput. When this approach was first proposed to the Pacific Fishery Management Council (PFMC), it was suggested that the output of a “data-moderate” assessment might be acceptable for use in management if it returned a favorable estimate of stock status but not be used as the basis for determining that a stock was overfished (PFMC 2013), or that there be an option for an “out-of-cycle” assessment to provide a second estimate of status before adopting an overfished status from a data-moderate assessment (NMFS 2013). Such an asymmetric standard of proof, especially when confounded with the lack of priority given to re-assessing stocks with favorable assessment outputs, seems likely to introduce bias at the level of the suite of stocks subject to the same management process, such as all the stocks in a Fishery Management Plan (FMP).

2.1.2 Conduct of assessments and reviews

Once a stock has been chosen for assessment, stock assessment analysts still face numerous decisions about the specific datasets to include as well as the treatment of putative “outliers” within accepted datasets, the weightings applied to different data sources, potential use of priors, parameters to fix versus estimate, and various functional forms. Assessments are typically subjected to review panels where data treatments and other modeling choices are scrutinized and alternatives are explored. In the vast majority of cases, the model endorsed at the end of the review process has some differences from the initially proposed base model, reflecting the combined deliberations of the assessors and reviewers. Ideally, the reviewers would be guided solely by scientific considerations, but just as with academic scientists (Simmons 2011), there is the potential for conscious or subconscious considerations leading to post-hoc reasoning to support the outcome perceived as likely to be most palatable to managers at the next step in the review and adoption process. There may also be incentives to be more critical of proposed model changes that reduce status, or to be less likely to recommend ultimate acceptance of an assessment yielding low status. These pressures may be most acute when the reviewers are drawn from bodies whose members are dependent on managers or politicians for their appointment or renewal (Crosson 2013).

2.1.3 Adoption of assessments for management

Following the initial peer review, adoption of stock assessments by Fishery Management Councils in the United States comes after review by a Scientific and Statistical Committee (SSC). Ideally, the SSC would apply equal standards of proof for acceptance of any assessment, but consciously or subconsciously they may apply extra scrutiny to assessments outputting poor status, although assessments yielding unexpectedly positive outcomes have also faced extra scrutiny. Once the SSC has endorsed an assessment, although Councils "may not exceed" the fishing level recommendations of the SSC, the Council must act to formally adopt the assessments recommended by the SSC, and this does not always happen (Crosson 2013, Nies 2022). Councils are intended to represent the public interest, but with an emphasis on the fishing industry. For example, as of 2022 the 72 appointed seats on U.S. Fishery Management Councils consisted of 29 representatives of commercial fishing interests, 25 representatives of recreational fishing interests, and 18 representatives of "other" interests (NMFS 2021), which can include tribal fishery representatives and individuals formerly more closely associated with a fishing interest. This in itself is of course not a definitive basis for concluding that Councils are more likely to reject a "pessimistic" assessment and accept an "optimistic one", but suggests that it could be a reasonable expectation. Indeed, in response to concerns expressed about increased and repeated scrutiny of poor-status assessments by the PFMC in 2021 (SSC 2021), a voting Council member stated that from the perspective of a manager and/or Council member, it is logical to expect that more attention will be given to more pessimistic assessments, consistent with the need to instill confidence for managers and stakeholders that the results are robust (SSC 2022, p. 13), suggesting that managers do indeed apply different standards of proof depending on management implications and may not appreciate the parallels with p-hacking or the potential biases introduced by shifting standards of proof.

2.2 Quantifying the problem

To quantify the likely magnitude of bias that might be introduced by various approaches to selecting assessments remanded for further scrutiny, revised, and/or re-assessed on a very short timeline (i.e., before another year is added to the assessment), I developed a simple simulation model. As with any model, it requires numerous simplifying assumptions, and does not incorporate all of the potential qualitative sources of bias described above. Nevertheless, I hope it is useful in demonstrating the potential scope of the problems introduced by asymmetric scrutiny during the review and adoption steps of the stock assessment process.

I simulated a system in which a large number of stocks are assessed in each assessment cycle, with the assumption that (prior to any additional selective scrutiny at the review and adoption stage) assessments are uncertain but median-unbiased. I then examined the distributional properties of the outputs of the full suite of assessments after a selected subset of assessments had been re-done within the same cycle, with the assumption that the redone assessments were also median unbiased, exploring various scenarios for the degree of independence between the initial and redone assessment. I also explored a scenario where sequential assessments are performed for a stock, and the degree of scrutiny applied to the assessment done at the later timestep depends on the proportional difference in OFL compared to the outcome of the assessment from the first timestep (Section 2.2.3).

2.2.1 Model structure – assessments redone in same year

Following the assumptions at the heart of the P*/sigma approach, I assumed that for the initial version of each assessment:

1) $$\log\left(\frac{OFL_{true}}{OFL_{assessed}}\right) = \epsilon_s + \epsilon_a$$

where

2) $$\epsilon_s \sim Normal(0, \sigma_s)$$

represents persistent errors for a given stock (arising from underlying issues in the primary data, assumptions shared across all candidate models [e.g., assumptions about steepness (Thorson et al. 2019) or natural mortality (Hamel 2014)], and if applicable persistent assessor and/or reviewer effects – and how all of these interact with the biology of the stock in question), and

3) $$\epsilon_a \sim Normal(0, \sigma_a)$$

represents assessment-specific errors arising from choices about datasets used, data weightings, and selection of model assumptions from within the general scope of acceptable assumptions.

To preserve the assumption that if all assessments are performed only once,

4) $$\log\left(\frac{OFL_{true}}{OFL_{assessed}}\right) \sim Normal(0, \sigma)$$

I used the equation for variance of a sum of random variables (assuming independence, which may be reasonable given how the respective errors were defined, but potentially problematic if certain features of the data lead to a tendency to select certain modeling options) to obtain

5) $$\sigma_a = \sqrt{\sigma^2 - \sigma_s^2}.$$

To simulate an assessment cycle that assesses *N* total stocks, I simulated the distribution of true versus assessed OFLs by drawing vectors of length *N* for ϵ_s and ϵ_a across various values of σ_s (chosen such that anywhere from 0% to 100% of the variance of assessment outputs was driven by assessment-specific factors) while holding σ constant at 0.5 (and thus determining σ_a via equation 5), then exponentiated the sum of these vectors. 0.5 is the default value of sigma applied by the PFMC for "category 1" assessments, the most data-rich and complex models used (PFMC 2022). To simulate a distribution of true versus assessed OFLs after some stocks were re-assessed, I retained all draws of ϵ_s, and all draws of ϵ_a for stocks that were not re-assessed, while drawing new values of ϵ_a for re-assessed stocks. I then exponentiated the summed vectors as before. For realistic values of *N*, stochastic variation from run-to-run is expected to predominate, thus I chose *N*=2,000,000 to approximate the asymptotic expectation. I focused on P* values of 0.45 and 0.40, as they are the most commonly used by the PFMC.

I explored various scenarios for the selection of assessments to be redone within a single assessment cycle: 1) selecting *x*% of assessments at random (exploring varying levels of *x*), 2) assuming skillful selection of problematic assessments by including the *x*/2% highest and *x*/2% lowest ratios of true versus assessed OFLs, or 3) assuming skillful selection of problematic assessments aimed only at the pessimistic ones by including the *x*% highest ratios of true:assessed OFLs.

2.2.2 Model outputs for within-cycle scrutiny

As expected, redoing assessments at random does not change the FMP-wide distribution of ratios between true and assessed OFLs (compare Figure 1a versus Figure 1b). If inaccurate assessments are skillfully selected regardless of the direction of error, the median ratio remains fixed at 1.0 (i.e., there is still no bias at the FMP-wide level) while the distribution narrows (but remains symmetric while no longer lognormal) and sigma becomes smaller (Figure 1c). If the least accurate assessments with errors in the direction of poor status are redone, the median ratio drops below 1.0 (indicating FMP-wide bias) and the distribution becomes non-symmetric (Figure 1d). The degree of bias introduced increases with the fraction of assessments redone and with the fraction of variance in assessment outputs attributable to assessment-specific factors (Figure 2). Note that this FMP-wide bias occurs even though the individual assessments and re-assessments are assumed to provide unbiased estimates.

Figure 1. Example outputs showing the ratio between true and assessed overfishing limits (OFLs) for the initial set of unbiased assessments (a), after redoing some assessments within-cycle at random (b), redoing the least accurate assessments within-cycle regardless of the direction of error (c), or redoing (within-cycle) the assessments with the largest errors in the direction of low status (d). Q45 and Q40 denote the 45th and 40th quantiles, respectively.

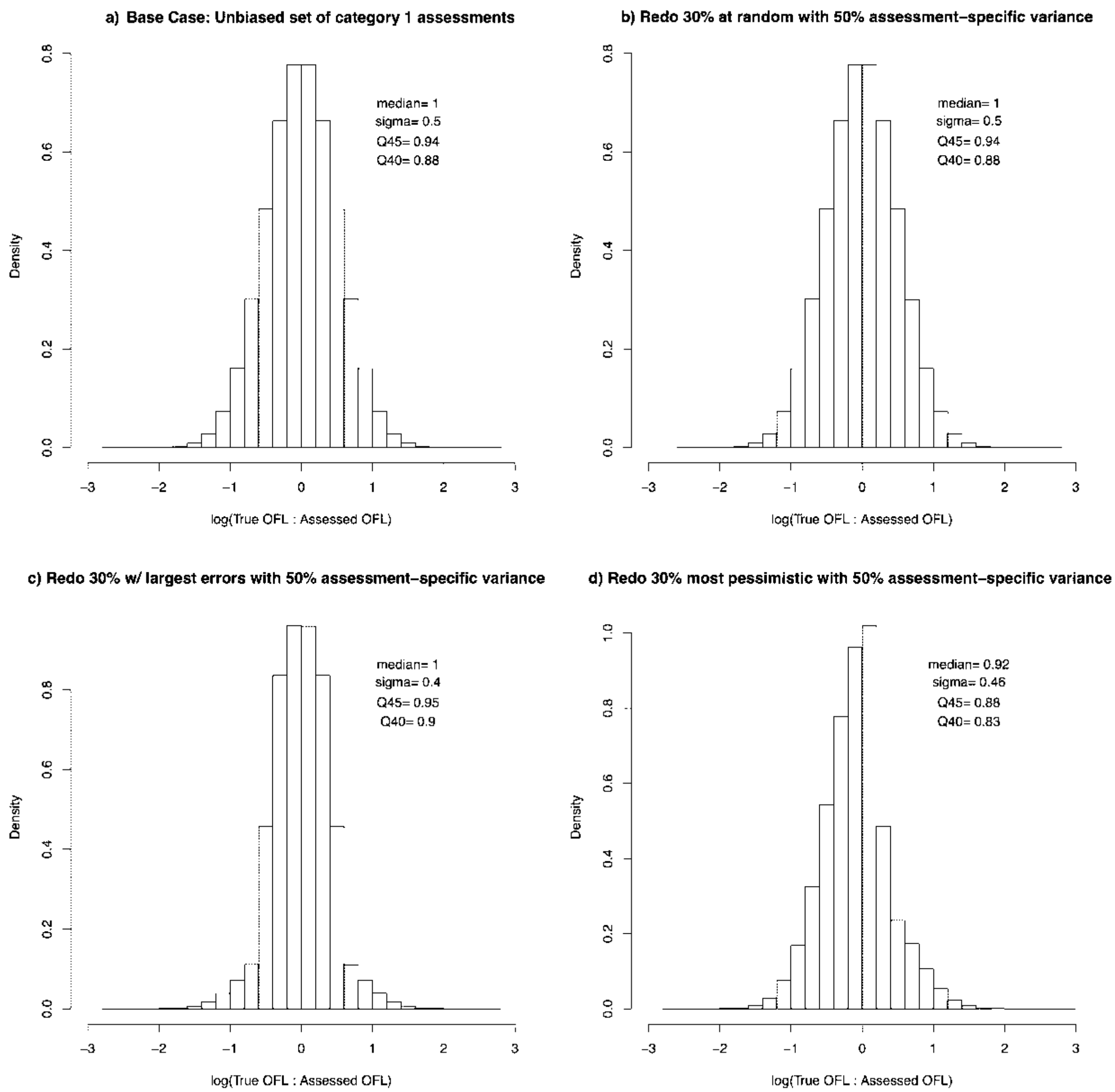
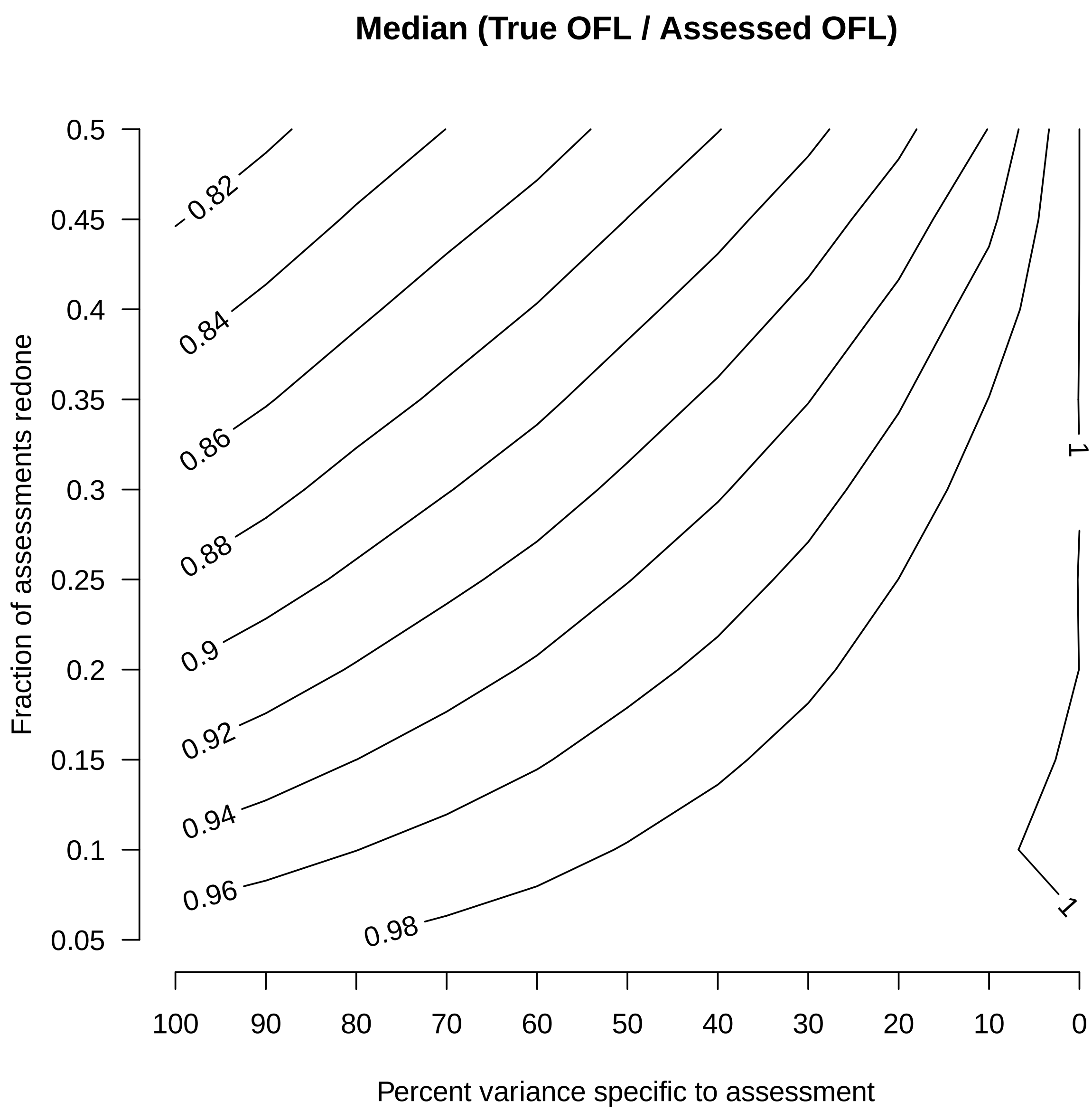
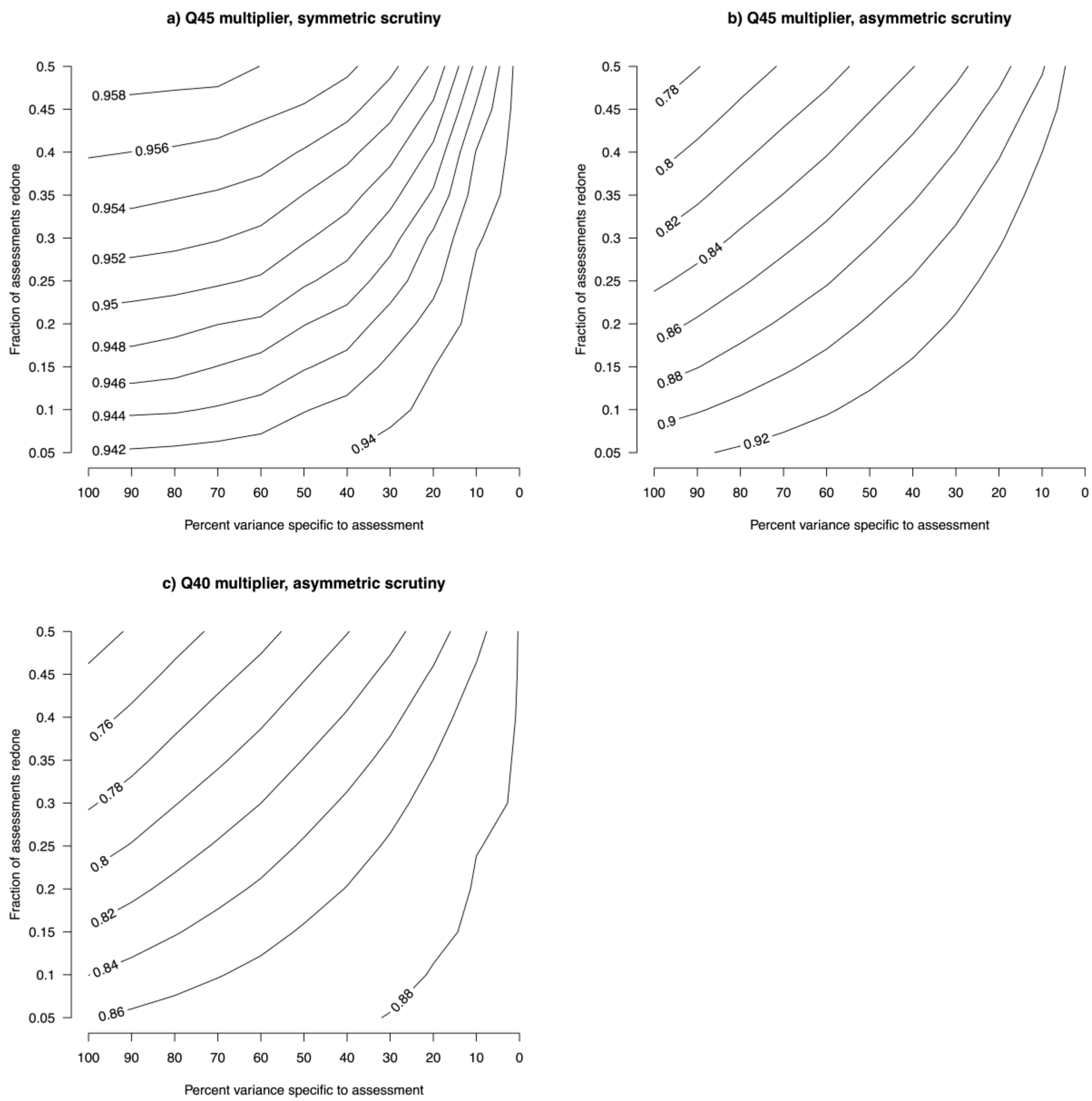


Figure 2. Median ratio between the true and assessed OFL following the skillful selection of assessments with the largest proportional errors in the direction of low status to be redone within a single assessment cycle.



These changes in the distribution of true versus assessed OFLs in the skillful selection scenarios for redoing assessments lead to changes in the multiplier needed to result in a specified probability that the ABC calculated by applying the multiplier to the assessed OFL will be higher than the true OFL. In the base case of lognormal distribution with sigma=0.5, a buffer Q45=0.945 results in the expectation that the ABC will be greater than the true OFL 45% of the time. If assessments are skillfully selected to redo, but without attention to the direction of error, a slightly larger multiplier (i.e., reduced buffer between ABC and OFL) can be used to achieve the same expectation (Figure 3a), although the change in multiplier is small. In contrast, skillfully selecting the most pessimistic assessments to be redone requires larger changes in the multiplier to preserve the nominal probability of ABCs exceeding the true OFL, and the multiplier needs to be decreased (Figure 3b). Even larger changes in the multiplier are required to achieve a nominal 40% probability of ABCs exceeding the true OFL (Figure 3c), which requires a multiplier of 0.881 in the base case where no assessments are redone.

Figure 3. Multiplier required to achieve a 45% (Q45, panels a and b) or 40% (Q40, panel c) probability that the ABC obtained by applying the multiplier to the assessed OFL is larger than the true OFL, under skillful and symmetric selection of the least accurate assessments to be redone within an assessment cycle (a) or under skillful selection of the most pessimistic assumptions to be redone within an assessment cycle (b and c). Note that the contour spacings are different in panel a versus b and c.



2.2.3 Responses to changes between consecutive assessments

So far, my quantitative analysis focused on actions taken within a single assessment cycle, such that both the original and redone assessment are estimating status for the same terminal year and setting catch limits for the same management year(s). It may also be the case that managers and reviewers would give extra scrutiny to an assessment of a stock that differed substantially in its status estimate or OFL determination compared to the previous assessment of that stock, although such changes could be unsurprising given long intervals between assessments or significant changes in the environment or management affecting the stock.

To simulate a system where consecutive assessments of the same stock are compared and unexpected changes may trigger further scrutiny of the more recent assessment, I assumed a set of OFLs was determined for a suite of stocks in both timestep 1 and timestep 2. In timestep 1 I assumed the ratios between true and assessed OFLs were determined as in equation 1, but with ϵ_a subscripted by timestep to reflect its potential to vary between timestep 1 and timestep 2:

6) $$\log\left(\frac{OFL_{true,1}}{OFL_{assessed,1}}\right) = \epsilon_s + \epsilon_{a,1}.$$

I assumed that the dynamics of the stock between timestep 1 and timestep 2 changed the true OFL by a proportion given by a lognormal distribution:

7) $$\log\left(\frac{OFL_{true,2}}{OFL_{true,,1}}\right) = \partial,$$

where

8) $$\partial \sim Normal(0, \sigma_d)$$

and σ_d represents the variability in stock dynamics. For each stock, the proportional difference between assessed OFLs in timestep 2 and timestep 1 is

9) $$\log\left(\frac{OFL_{assessed,2}}{OFL_{assessed,,1}}\right) = \partial + \epsilon_{a,2,initial} - \epsilon_{a,1}$$

where $\epsilon_{a,2,initial}$ is the assessment-specific error associated with the initial iteration of the timestep 2 assessment (with the "initial" subscript reflecting the potential that some timestep 2 assessments will be redone), noting that ϵ_s affects the assessed OFLs in both timesteps equally and so it drops out of the comparison.

For the initial set of timestep 2 assessments,

10) $$\log\left(\frac{OFL_{true,2}}{OFL_{assessed,2,initial}}\right) = \epsilon_s + \epsilon_{a,2,initial}$$

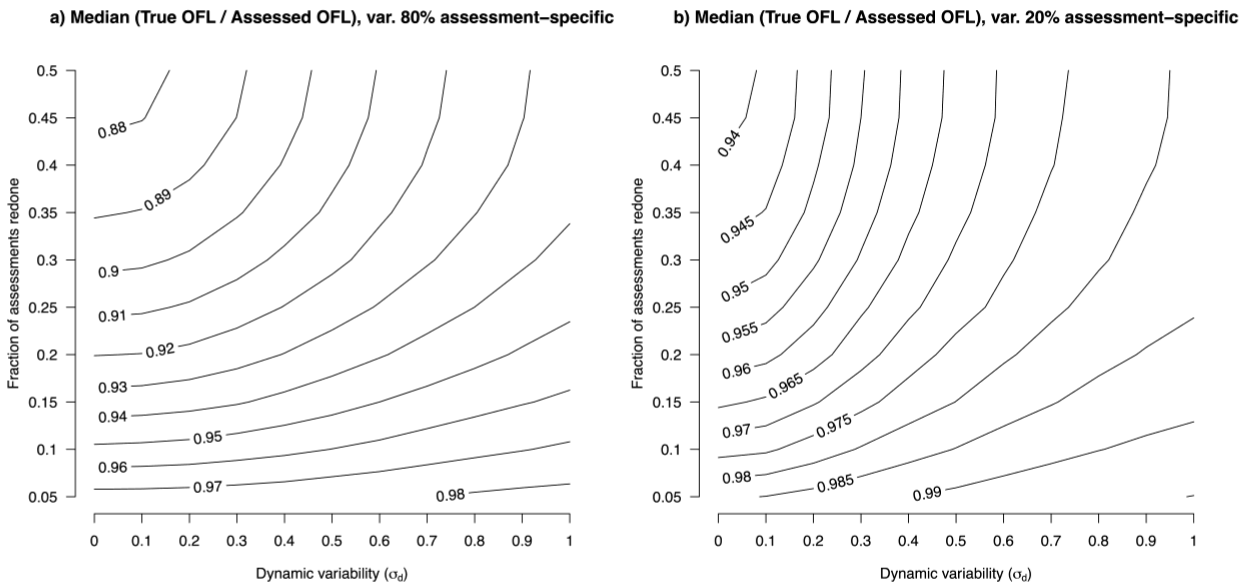
and I simulated scenarios where the timestep 2 assessments with the largest proportional changes in OFL compared to timestep 1 (as determined in equation 9) were redone, resulting in new errors:

10) $$\log\left(\frac{OFL_{true,2}}{OFL_{assessed,2,redone}}\right) = \epsilon_s + \epsilon_{a,2,redone}$$

and explored the distribution of the ratio between true and assessed OFLs given different levels of σ_d and different proportions of assessments being redone within timestep 2. I explored scenarios where the *x*% of largest proportional decreases in OFL or the *x*/2% of largest proportional changes in either direction led to re-assessment. For these simulations, I held σ_a constant at either 0.447 (80% of total variance given σ=0.5) or 0.225 (20% of total variance) to reduce the dimensionality of the simulations. Given the meta-analytic approach to estimating σ based on repeated assessments (Ralston et al. 2011, Privitera-Johnson and Punt 2020b), it is likely that estimates of σ are dominated by assessment-specific factors, with intrinsic factors largely constant across assessments and so not revealed by comparison of sequential assessments. Therefore, the larger value of σ_a may be more appropriate, since σ_a likely drives current estimates of σ.

Applying extra scrutiny to large proportional decreases in OFL from one assessment to the next can yield similar biases to redoing assessments within-cycle if the changes in true stock size/status between assessment periods is small relative to the error in assessments (compare Figure 4 to Figure 2, comparing Figure 4a to the part of Figure 2 where a large proportion of variance is assessment-specific and 4b to the part of Figure 2 where a small proportion of variance is assessment-specific). As the variation in true stock size/status between assessment periods becomes larger, the bias introduced is less (Figure 4) and results become more similar to picking assessments to be redone at random, because when changes in true stock size/status between assessments are large, assessment error has relatively little effect on the probability of observing a large overall change in the estimated OFL. As expected, redoing assessments that show large proportional changes in OFLs regardless of the direction of change does not introduce a bias (results not shown, but are produced by the code available online). Implications for appropriate uncertainty buffers or multipliers are similar to the effects on bias, requiring larger buffers in the case of directional scrutiny and having minimal effects on appropriate buffers in the case of symmetric scrutiny (results not shown, but are produced by the code available online).

Figure 4. Median ratio between the true and assessed OFL after redoing the second in a pair of sequential assessments if there was a large proportional reduction in the OFL for the second assessment compared to the first. In panel a, 80% of the variance associated with assessment error is due to assessment-specific factors, whereas in panel b 80% of the variance is associated with intrinsic factors that do not vary between assessment iterations.



2.2.4 Caveats and potential extensions

The models presented here are admittedly oversimplifications of complicated processes where there may be no bright line between "redoing" an assessment within a cycle and the revisions that normally occur during the process of model development and review. Numerous factors could lead to the expectation that true OFLs would change between assessments performed at different times, with the direction of change depending on the scenario (e.g., OFL likely to increase through time for a rebuilding stock, or decrease for a newly-targeted stock currently assessed to be well above its biomass target). The assumption that first-pass assessments are median unbiased, or that a single distribution can describe the uncertainty associated with each assessment, is also a gross oversimplification. Numerous factors could affect the covariance between estimates from initial and revised assessments, and the expectation that revised assessments would be median-unbiased is questionable. Requests for model changes in revised assessments may have anticipated directional effects, or new research projects may be funded with the anticipation of directional changes in assessment outcomes (Terceiro 2018, Lynn et al. 2022). Nevertheless, I chose to model a scenario in which redone assessments were median unbiased to illustrate the potential for inadvertent bias to be introduced even when requests for assessment revisions were not made with the intent or anticipation of driving the results in a particular direction. It may often be the case that proposed revisions to historical datasets, addition or removal of datasets, changes in data treatments, and/or revised prior specifications would have predictable effects. These sorts of predictable biases are not included in my simulations, and might be guarded against by restricting the opportunities for requesting such changes to early in the assessment process, before results are known.

In addition, stock assessments output numerous other quantities of scientific and management importance beyond the OFL, including estimates of status or depletion and estimates of biological parameters that affect productivity and so influence the projections needed for multi-year catch advice and for identifying sustainable fishing rates. The uncertainty in many of these estimates may also be reasonably described by a lognormal distribution (Bi et al. 2023) amenable to exploration with a similar approach, although there may be less of an established basis for the value of sigma to assume. I focused on OFLs and ABCs given the clear frequentist interpretation of P* and a well-established existing framework for using this approach to characterize uncertainty. However, while incentives associated with p-hacking and publication of novel results is expected to lead to increased Type I error rates (i.e., false positives), it might be argued that a tendency to be suspicious of low-status assessments and favor model variants giving more moderate status could be more akin to increasing the rate of Type II errors (false negatives or incorrectly rejecting an assessment of poor status).

It is important to realize that full ABCs may not be attained, and thus an ABC higher than the true OFL does not necessarily mean that biological overfishing will or is likely to occur (i.e., the fishing mortality rate actually estimated for recent years is often well below the proxy for the rate expected to produce maximum sustainable yield [e.g., Figure 3.4.1 of Harvey et al. (2022)]). Thus, even if the calculated P* is lower than the true probability of an ABC exceeding the true OFL, biological overfishing may still be acceptably unlikely given expected attainment levels. This might reduce concern about the accuracy of P* calculations, but this complication would be better addressed through a framework where P* represents the probability that the expected harvest, as opposed to the ABC, would result in overfishing. Such an approach has not yet been developed.

Empirically quantifying the degree to which assessments are “redone” prematurely, and the extent to which low status predicts assessments receiving extra scrutiny, would be a formidable task requiring extensive review of the grey literature and likely a fair amount of informed speculation about the motivation underlying incompletely documented decisions. Silvar-Viladomiu et al. (2021) and Bi et al. (2023) did not detect evidence of bias when comparing year-specific status estimates between repeated assessments of the same stocks, but each assessment was the product of a similar process that might exert similar effects on each iteration of the assessment. In addition, these analyses may not have sufficient power to detect small biases. Note also that my simulations assumed the re-assessments were themselves unbiased, and the potential bias arises when considering advice on the managed suite of stocks as whole rather than an expectation that the redone assessments are themselves biased. Some reviews of the scientific literature have used techniques like p-curves to test whether distributions of critical test statistics of published papers have discontinuities at critical “significance” levels that could be indicative of p-hacking or publication bias (Simonsohn et al. 2015). A similar approach might be used to examine the frequency of assessments indicating status just above versus below target or limit reference points, although one might expect that successful management would naturally lead to discontinuities around such reference points. Additionally, one could test whether assessments of status just above limits or targets in the terminal year of an earlier assessment tended to be consistent with updated perceptions of status for that year based on future assessments (Bi et al. 2023). Future simulation work could relax the assumption that initial or redone assessments are median unbiased, model the effects of directed requests for assessment revisions, and/or partition the assessment-specific error into additional components such as individual assessor effects, institutional effects, modeling platform effects, and the like. This might be addressed through a hierarchical modeling framework.

2.2.5 Likely magnitude of the problem

In practice, reviewers and decision makers likely have some skill in identifying less accurate assessments, but do not have perfect knowledge, suggesting that the outcome in practice is likely to be somewhere in between the random selection and skillful selection scenarios. Note that the effects of moderate changes in the fraction of assessments redone or the proportion of variance attributable to assessment-specific factors can change the multiplier by amounts comparable to or larger than the changes needed to achieve 45% versus 40% probability of establishing an ABC higher than the true OFL. The bias in the OFL estimate introduced by asymmetric scrutiny is larger than can be countered by the default Q45 when as few as about 10% of assessments are redone if a high proportion of variance in OFL estimates is attributable to assessment-specific factors, or about 20% of assessments if 50% of OFL variance is assessment-specific (Figure 2). If assessments are not redone within-cycle, but large relative decreases in OFLs between sequential assessments prompt extra scrutiny, this can introduce comparable levels of bias if the interval between assessments is short relative to the rate of change in true stock size/status. Thus, these problems may be more acute for long-lived, frequently assessed species and less acute for short-lived, infrequently assessed species.

Overall, these simulations suggest that randomly selecting assessments to be redone within a cycle is a waste of time and resources. Skillfully and symmetrically selecting assessments to be redone will not introduce bias, but is likely an inefficient use of resources, given the small changes in suitable multipliers. Redoing only the most pessimistic assessments within a cycle would introduce a bias that may be comparable in magnitude to the differences between choice of P*=0.45 versus 0.40. Similar biases could result when applying extra scrutiny to reductions in OFL or status between consecutive assessments, although the bias is reduced when changes in true stock size/status over the relevant interval are likely to be large, such as for infrequent assessment of a short-lived and dynamic stock. Differences on the order of a few percent may be acceptable relative to other uncertainties in the process, but the selection of assessments to be redone should be judicious to avoid the introduction of larger biases.

3. Potential solutions

Similar to several broad reviews of the scientific literature, the examples and models explored here suggest that various processes operating at the analysis and “publication” or adoption stage for stock assessment science are introducing a bias into fishery-wide OFL specification, and the rate at which ABCs exceed the OFLs that would have been established given perfect knowledge is likely higher than the nominal value of P*, similar to how publications of “significant” results likely have a higher false positive rate than the nominal p-value. While the magnitude of the bias may not be sufficient to call the scientific or assessment enterprise as a whole into question, it does seem sufficient to warrant caution and efforts to limit known sources of bias as much as possible.

For the broader scientific enterprise, several courses of action have been proposed (Wicherts et al. 2016), most of which have clear analogs in the assessment process. 1) Analyses should be driven by clear *a priori* hypotheses that lead directly to a parsimonious set of candidate explanatory covariates, with objective methods for model and variable selection. 2) There should be transparency in statistical model selection and significance criteria. 3) Preregistration should be considered and employed to the extent possible. At the funding or pre-publication stage, clear statements should be made of the motivation for a study, the hypotheses it will test, the data to be collected and the analyses to be performed (ideally with a power analysis indicating sufficient power to detect meaningful effects if present), and the statistical tests to be performed along with the significance criterion and its justification. If all of these are satisfactory, publication should be assured regardless of the p-value obtained. 4) When multiple statistical tests or model formulations are applied to the same dataset, some adjustments like the Holm-Bonferroni procedure or Šidák correction should be applied. 5) A more stringent "significance" standard than $p<0.05$ should be considered for novel findings (Benjamin et al. 2018).

The stock assessment prioritization and review process along with the application of the P*/sigma system for developing ABCs from OFLs offers analogies to all of these recommendations. 1) At the beginning of each assessment cycle, the species to be assessed, the spatial boundaries in assessment and management units, the data sources to be considered for inclusion, and the standards for review should all be specified in advance. 2) Review criteria should be clearly tied to the strength of scientific evidence, not management implications. 3) Assessments should not be aborted or rejected for use in management based on a politically unfavorable outcome. 4) The buffer between the ABC and OFL should be increased beyond that implied by the nominal choice of P*, based on an approach similar to the models explored in Section 2.3. Similar adjustments to estimates of depletion and status may also be warranted. 5) Strict standards should be adopted for revising, further reviewing, or rapidly revisiting an assessment that has been endorsed by scientific reviewers. For example, the SSC of the Mid-Atlantic Council will only reconsider a recommendation if new data are found or an error is discovered in an assessment (Crosson 2013), and the New England Fishery Management Council has similar limits on when an SSC recommendation can be remanded (Nies 2022). There may also be benefits in determining *a priori* criteria for how large a change would be required to deem a revised assessment sufficiently different from the initial assessment to revisit the adoption of the original assessment (SSC 2022), which might be based on evaluating the magnitude of the difference between two model alternatives relative to the overall level of uncertainty (Cope and Gertseva 2020).

4. Conclusions

Overall, there seem to be numerous pathways by which inadvertent bias may be introduced into the stock assessment prioritization, review, and adoption process; and these pathways have commonalities in concerns raised about “p-hacking” in the scientific enterprise more broadly. Fortunately, the broader scientific literature also poses potential solutions or at least steps to reduce the influence of p-hacking, and many of these steps have direct analogs that can be applied to reduce the chances of introducing inadvertent bias into the fisheries stock assessment process. Simply raising awareness of the issue may go a long way toward fostering more careful work that is less likely to create bias (Peng 2015).

Data Availability

No original data were used in this paper. R code to run the simulations is available in a Mendeley archive at https://data.mendeley.com/datasets/d49ct4fypr/2.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author and do not necessarily reflect the views of NOAA or the Department of Commerce. The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The author is a member of the Pacific Fishery Management Council's Scientific and Statistical Committee at the time of writing, and his experiences there may have affected his perception of issues discussed in this paper, but he does not speak on their behalf.

Acknowledgments

This paper was informed and improved by feedback from members of the PFMC’s SSC and stock assessment teams, attendees at a University of Washington Think Tank seminar, Eric Ward, Corey Ridings, John Field, Michael O’Farrell, Olaf Jensen, and an anonymous reviewer.

References

Baker, M., 2016. Is there a reproducibility crisis? Nature 533, 542-454. https://doi.org/10.1038/533452a.

Benjamin, D.J., Berger, J.O., Johannesson, M.J., Nosek, B.A., Wagenmakrers, E.-J., et al., 2018. Redefine statistical significance. Nature Human Behaviour 2, 6-10. https://doi.org/10.1038/s41562-017-0189-z.

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://dx.doi.org/10.1111/faf.12714.

Cadrin, S., Henderschedt, J., Mace, P., Mursalski, S., Powers, J., Punt, A.E., Restrepo, V., 2015. Addressing Uncertainty in Fisheries Science and Management. National Aquarium. http://www.fao.org/3/a-bf336e.pdf.

Clark, T.D., Raby, G.D., Roche, D.G., Binning, S.A., Speers-Roesch, B., Jutfelt, F., Sundin, J. 2020. Ocean acidification does not impair the behaviour of coral reef fishes. Nature 577, 370-375. https://doi.org/10.1038/s41586-019-1903-y.

Cope, J.M., DeVore, J., Dick, E.J., Ames, K., Budrick, J., Erickson, D.L., et al., 2011. An approach to defining stock complexes for U.S. West Coast groundfishes using vulnerabilities and ecological distributions. N. Am. J. Fish. Manag. 31, 589–604. https://dx.doi.org/10.1080/02755947.2011.591264.

Cope, J.M., Gertseva, V., 2020. A new way to visualize and report structural and data uncertainty in stock assessments. Can. J. Fish. Aquat. Sci. 77, 1275-280. https://doi.org/10.1139/cjfas-2020-0082.

Crosson, S., 2013. The impact of empowering scientific advisory committees to constrain catch limits in US fisheries. Science and Public Policy 40, 261-273. https://doi.org/10.1093/scipol/scs104.

Fanelli, D., 2018. Is science really facing a reproducibility crisis, and do we need it to? Proc. Nat. Acad. Sci. 115, 2628-2631. https://doi.org/10.1073/pnas.1708272114.

Harvey, C., et al. 2022. 2021-2022 California Current Integrated Ecosystem Assessment (CCIEA) California Current ecosystem status report. Report to the Pacific Fishery Management Council. https://www.pcouncil.org/documents/2022/02/h-2-a-cciea-team-report-1-2021-2022-california-current-ecosystem-status-report-and-appendices.pdf/.

Hamel, O.S., 2014. A method for calculating a meta-analytical prior for the natural mortality rate using multiple life history correlates. ICES J. Mar. Sci. 72, 62-69. https://doi.org/10.1093/icesjms/fsu131.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D., 2015. The Extent and Consequences of P-Hacking in Science,. PLoS Biol. 13, e1002106. https://dx.doi.org/doi:10.1371/journal.pbio.1002106.

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., 2020. Effective fisheries management instrumental in improving fish stock status. Proc. Nat. Acad. Sci. 117, 2218-2224.

Hilborn, R., Walters, C., 1992. Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty. Chapman and Hall.

Ioannidis, J.P.A., 2005. Why most published research findings are false. PloS Medicine 2, e124. https://dx.doi.org/10.1371/journal.pmed.0020124.

Jager, L.R., Leek, J.T., 2014. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15, 1-12. https://dx.doi.org/10.1093/biostatistics/kxt007.

Lynn, K., Dorval E., Porzio, D., Nguyen, T., 2022. A collaborative survey of coastal pelagic species in nearshore California waters. Fisheries 47, 500-508. https://doi.org/10.1002/fsh.10840.

Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. M. Sci. 72, 7-18. https://doi.org/10.1093/icesjms/fsu015.

Maunder, M.N., Piner, K.R., 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fish. Res. 192, 16-27. https://doi.org/10.1016/j.fishres.2016.04.022.

Melnychuk, M.C., Banobi, J.A., Hilborn, R., 2013. Effects of management tactics on meeting conservation objectives for western North American groundfish fisheries. PLoS One 8, e56684. https://doi.org/10.1371/journal.pone.0056684.

Methot Jr., R.D. (editor). 2015. Prioritizing fish stock assessments. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-F/SPO-152. https://repository.library.noaa.gov/view/noaa/12874.

Methot Jr, R.D., Tromble, G.R., Lambert, D.M. Greene, K.E., 2014. Implementing a science-based system for preventing overfishing and guiding sustainable fisheries in the United States. ICES J. Mar. Sci. 71, 183-194. https://doi.org/10.1093/icesjms/fst119.

Mildenberger, T., 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 Fish. 23, 73-92. https://doi.org/10.1111/faf.12599.

NMFS (National Marine Fisheries Service), 2013. Use of stock status results from data-moderate assessments. Presentation to Pacific Fishery Management Council. https://www.pcouncil.org/documents/2013/03/h-groundfish-management-march-2013.pdf/#page=40.

NMFS (National Marine Fisheries Service), 2021. 2021 Report to Congress on the Regional Fishery Management Councils and Scientific and Statistical Committee members' financial interest disclosure and recusal requirements and on the Regional Fishery Management Councils membership apportionment developed pursuant to section 302(b)(2)(b) and section 302(j)(9) of the Magnuson-Stevens Fishery Conservation and Management Act. https://media.fisheries.noaa.gov/2022-03/Final-2021%20Report%20to%20Congress%20on%20Councils.pdf.

NMFS (National Marine Fisheries Service), 2022. Prioritizing groundfish stock assessments: Preliminary analysis and ranking of Pacific coast groundfish species for assessment in 2023. Presentation to Pacific Fishery Management Council. https://www.pcouncil.org/documents/2022/03/e-8-a-supplemental-noaa-fsc-presentation-1-preliminary-analysis-and-ranking-of-pacific-coast-groundfish-species-for-assessment-in-2023-dr-hastie-and-dr-wetzel.pdf/

Nies, T. 2022. SSC Roles and Responsibilities. SSC Planning Meeting, presentation of the NEFMC Executive Director. https://s3.us-east-1.amazonaws.com/nefmc.org/SSC-roles-and-responsibilities_for-web.pdf.

Peng, R., 2015. The reproducibility crisis in science: A statistical counterattack. Significance 6, 30-32. https://doi.org/10.1111/j.1740-9713.2015.00827.x.

PFMC (Pacific Fishery Management Council), 2013. Status determination criteria for data-moderate stocks. https://www.pcouncil.org/documents/2013/03/h-groundfish-management-march-2013.pdf/#page=27.

PFMC (Pacific Fishery Management Council), 2022. Terms of reference for the groundfish stock assessment review process for 2023-2024. https://www.pcouncil.org/documents/2022/06/terms-of-reference-for-the-groundfish-stock-assessment-review-process-for-2023-2024-june-2022.pdf/.

Privitera-Johnson, K.M., Punt, A.E., 2020a. A review of approaches to quantifying uncertainty in fisheries stock assessments. Fish. Res. 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 J. Mar. Sci. 77, 515-526. https://doi.org/10.1093/icesjms/fsz237.

Provencher, J.F., Covernton, G.A., Moore, R.C., Horn, D.A., Conkle, J.L., Lusher, A.L. 2020. Proceed with caution: The need to raise the publication bar for microplastics research. Sci. Tot. Env. 748, 141426. https://doi.org/10.1016/j.scitotenv.2020.141426.

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. Fish. Bull. 109, 217–231. https://spo.nmfs.noaa.gov/content/meta-analytic-approach-quantifying-scientific-uncertainty-stock-assessments.

Rudd, M.B., Cope, J.M., Wetzel, C.R., Hastie, J., 2021. Catch and length models in the stock synthesis framework: expanded application to data-moderate stocks. Front. Mar. Sci. 8, 663554. https://doi.org/10.3389/fmars.2021.663554.

Schooler, J., 2011. Unpublished results hide the decline effect. Nature 470, 437. https://doi.org/10.1038/470437a.

Seagraves, R., Collins, K. (eds), 2012. Fourth National Meeting of the Regional Fishery Management Council's Scientific and Statistical Committees. Report of a National SSC Workshop on Scientific Advice on Ecosystem and Social Science Considerations in U.S. Federal Fishery Management. Mid-Atlantic Fishery Management Council, Williamsburg, VA. http://www.fisherycouncils.org/ssc-workshops/fourth-national-ssc-workshop-2011.

Shertzer, K.W., Prager, M.H., Williams, E.H., 2008. A probability-based approach to setting annual catch levels. Fish. Bull. 106, 225–232. https://media.fisheries.noaa.gov/dam-migration/ns1-shertzer-et-al-2008.pdf.

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

Simmons, J.P., Nelson L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psych. Sci. 22, 1359-1366. https://dx.doi.org/10.1177/0956797611417632.

Simonsohn, U., Simmons, J. P., Nelson, L. D., 2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller

(2015). J. Experimental Psychology: General 144, 1146–1152. https://doi.org/10.1037/xge0000104.

SSC (Scientific and Statistical Committee of the Pacific Fishery Management Council), 2021. Scientific and Statistical Committee Report on Harvest Specifications for 2023-2024 Including Final Overfishing Limits and Acceptable Biological Catches. https://www.pcouncil.org/documents/2021/11/e-3-a-supplemental-ssc-report-1-2.pdf/.

SSC (Scientific and Statistical Committee of the Pacific Fishery Management Council), 2022. SSC Groundfish Subcommittee Report on Groundfish Stock Assessment Process Review Webinar held on January 25, 2022. https://www.pcouncil.org/documents/2022/02/e-8-attachment-6-ssc-groundfish-subcommittee-report-on-groundfish-stock-assessment-process-review-webinar-held-on-january-25-2022.pdf/.

Terceiro, M., 2018. The summer flounder chronicles III: struggling with success, 2011-2016. Rev. Fish Biol. Fisheries 28, 381-404. https://doi.org/10.1007/s11160-017-9506-x.

Thorson, J.T., Dorn, M.W., Hamel, O.S., 2019. Steepness for West Coast rockfishes: results from a twelve-year experiment in iterative regional meta-analysis. Fish. Res. 217, 11-20. https://doi.org/10.1016/j.fishres.2018.03.014.

Viglione, G., 2020. 'Avalanche' of spider-paper retractions shakes behavioural-ecology community. Nature 578, 199-200. https://www.nature.com/articles/d41586-020-00287-y.

Wasserstein, R.l., Lazar, N. A., 2016. The ASA statement on p-values: Context, process, and purpose. Am. Stat. 70, 129-133. https://dx.doi.org/10.1080/00031305.2016.1154108.

Wicherts, J.M., Veldkamp, C.L.S., Augusteijn, H.E.M., Bakker, M., van Aert, R.C.M. and van Assen, M.A.L.M., 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-Hacking. Front. Psychol. 7,1832. https://dx.doi.org/10.3389/fpsyg.2016.01832.