Response to Reviews: CJFAS-2025-0093 — Evaluating the impact of log-normal bias-correction on a state-space stock assessment model

Authors: Li et al.

Contents

1 Cover Letter to the Associate Editor

Dear Associate Editor,

We sincerely appreciate the time and effort that you and the reviewers have invested in evaluating our manuscript *"Evaluating the impact of log-normal bias-correction on a state-space stock assessment model" (CJFAS-2025-0093)*. We have carefully considered all comments and suggestions, and in this revised version we have addressed each point in detail.

The major revisions include:
- Adding 50 additional replicates to address concerns about low sample size.
- Calculating and presenting additional performance metrics, all of which support our main conclusions.
- Revising and strengthening the **Methods** and **Discussion** sections based on reviewer feedback.
- Updating figures and tables, including several new supplementary figures to address specific concerns.

We believe these revisions have substantially improved the clarity, rigor, and overall quality of the manuscript. Thank you again for your thoughtful comments and guidance, which have been invaluable in strengthening this work.

Sincerely,
Li et al.

2 Associate Editor Comments to the Authors

Major Comments

> Multiple reviewers mention sloppy notation, question the meaning/use of "«" in Line 105, and have concerns about dropping unconverged runs and simulating again until 50 runs were converged for each scenario. Also, there is general consensus that you should discuss your convergence criteria. Some suggestions for improving Figures and Tables are provided. Two reviewers mention Aldrin et al. (2020), which seems relevant and the authors should discuss relative to differences in finding/guidance. Reviewer 4 raises a good conceptual question about section 2.3, please expand on this point.

Response

We thank the Associate Editor and reviewers for their constructive feedback. A detailed point-by-point response is provided below. The resubmission was delayed due to the extended time required to complete an additional 50 replicates for each scenario under limited computational resources at the center.

Major Comment 1

> Multiple reviewers mention sloppy notation and question the meaning/use of "«" in Line 105.

Response

We have carefully revised the manuscript to ensure notation is used consistently and clearly throughout. To avoid ambiguity, we have replaced this symbol with explicit wording in the revised manuscript (i.e., is always less than).

Major Comment 2

> Concerns were raised about dropping unconverged runs and simulating again until 50 runs were converged for each scenario. Reviewers also requested a discussion of convergence criteria.

Response

We've completed an additional 50 replicates for each scenario.

In the revised manuscript (Section 2.5), we now clearly describe our convergence criteria: (1) successful convergence of the optimizer and (2) an invertible Hessian. Simulations failing to meet these criteria were discarded. To strengthen robustness, we increased the number of replicates from

50 to 100 complete and converged runs for each scenario. Additional iterations with new random seeds were conducted until 100 converged replicates were obtained.

We also highlight in the main text (Lines 179–186) that "convergence failures were often realization-specific rather than model-specific. In most cases, if one EM failed to converge for a given realization, all EMs failed for that same realization, suggesting that convergence issues were driven by factors other than the bias correction setting. Therefore, to better isolate the effect of bias correction, we calculated a conditional convergence rate: the proportion of simulations that converged for the misspecified model, given that the correctly specified self-test had already converged. This provides a clearer evaluation of how the bias correction mismatch impacts model convergence, with results summarized across all stocks (Table S4)."

Major Comment 3

> Some suggestions for improving Figures and Tables are provided.

Response

We thank the reviewers for these helpful suggestions. In response: 1. Figures have been re-generated to reflect the increased number of simulations (100 replicates per scenario instead of 50). 2. Captions have been refined to ensure they are more informative and clear. 3. In the main text, we have retained the current style of summarizing results, as we believe it most effectively conveys the high-level patterns to the reader. However, to provide full transparency, the supplementary materials now include additional figures showing the complete set of simulation results (all OM $\times$ EM combinations for each random-effects structure). 4. We have also added figures using alternative performance measures beyond relative error, including log-relative error (symmetric), median symmetric signed percentage bias (symmetric), and root mean squared relative error (combining bias and variance). These complementary metrics all support the same overall conclusions presented in the manuscript.

Major Comment 4

> Two reviewers mention Aldrin et al. (2020), which seems relevant, and suggest the authors discuss relative to differences in findings/guidance.

Response

We appreciate this recommendation and now include a dedicated discussion of Aldrin et al. (2020) in the revised manuscript (Discussion, Lines 286–299). We compare our results with theirs, highlighting both methodological and interpretive differences. In particular, Aldrin et al. (2020) found that applying bias correction to catch data could improve model performance. A key distinction lies in the treatment of observation error: in their state-space assessment model (SAM), the catch observation error variance was freely estimated, whereas in our study it was fixed at low levels. Because the catch data in our models had relatively small and fixed error (and were thus heavily weighted), the potential benefit of observation-level bias correction was limited. We emphasize this contrast in the Discussion to place our findings in context and to clarify why our conclusions differ from Aldrin et al. (2020).

Major Comment 5

> Reviewer 4 raises a conceptual question about Section 2.3 and requests expansion on this point.

Response

We thank the reviewer for this thoughtful comment. In the revised manuscript (Section 2.3), we expand the motivation for bias correction in log-normal observation models. We clarify that, with bias correction, the mean of the observation distribution equals the true catch (mean-unbiased); without bias correction, the median equals the true catch (median-unbiased), and the observation is mean-unbiased on the log scale. In our study, observation variance was fixed and small, so choosing a mean- vs. median-unbiased representation has negligible practical impact; the two are related by the $\sigma^2/2$. For survey indices, we note that bias correction is less influential because catchability (q) can absorb much of the adjustment (Section 2.3). We have made these points explicit in the revised text, and the broader implications are discussed in the Discussion (Lines 286–299).

Major Comment 6

> Figures S10 and S11 – Are you isolating the effect of observation error or process error by setting the CV so low? Or do you mean that you are minimizing the effect of observation error?

Response

This is an additional case study where we fixed the observation error for both catch and survey at very low levels to minimize its influence and thereby isolate the effect of process variance (recruitment and NAA) on model performance. The results clearly show that model performance was not sensitive to the magnitude of recruitment process variance, but was highly sensitive to the magnitude of NAA process variance. Specifically, higher NAA process variance led to greater bias when models were misspecified with respect to bias correction structure. This finding supports our main conclusion that when the OM excludes bias correction but the EM includes it, increasing NAA process variance produces significantly larger bias in recruitment and SSB estimates.

Major Comment 7

> Bias in model-estimated recruitment and SSB when bias correction is applied is not highlighted in the Abstract, but it should be.

Response

We agree with this suggestion and have revised the Abstract to explicitly highlight that applying bias correction can lead to bias in estimates of SSB and NAA.

Major Comment 8

> Concerns were raised about the stability of random effects when AR(1) approaches 1 (i.e., random walk behavior), and about the source of higher estimation bias for flounder relative to the other two stocks.

Response

We agree that a true random walk (AR1 = 1) would raise concerns; however, when AR1 < 1 the process remains stable. Thus, even when ρ is close to 1, the random effects are still stable, particularly in cases where the EM converges. We believe the greater magnitude of estimation bias observed for flounder is related to the higher process variance for NAA specified in the OM for that stock, compared to haddock and mackerel. Because the process variance estimates directly influence the bias-correction term, this likely explains the stronger estimation bias observed for flounder. Our additional sensitivity analysis (Figures S17 and S18) indirectly supports this interpretation.

Major Comment 9

> The paragraph on Lines 296–307 was not a well-developed line of analysis, not mentioned in the methods, and the reader has no information about what improvement in data was relative to earlier years. If this is an important part of the analysis, I suggest developing it further and providing more details that quantify data improvements. At present, the 'conclusion' that data quality and quantity are less influential is more of a hypothesis, but stating it as such without a well-designed test could work against efforts to maintain high-quality data streams.

Response

We appreciate this comment. Our intention in this paragraph, together with the sensitivity analyses in Figures S19–S20, was to provide a snapshot of model performance during an intermediate period with richer data. This was meant to indirectly illustrate that, in our specific case study, model performance was not strongly linked to data availability or quantity (since observation error was fixed at low levels), but instead more related to internal estimation bias in model parameters. We agree that this analysis is case-specific, and we have clarified in the revised Discussion that these findings apply only under our study conditions. We have also emphasized that future research should more rigorously evaluate how variability in data availability and quality influences state-space models, as noted in our revised manuscript (Discussion, Lines 335–337).

Major Comment 10

> In Lines 310–311, the authors state "In the absence of strong evidence in support of bias correction…" – what would such evidence look like? It would be good to give readers some insight on what they should be looking for.

Response

We appreciate this suggestion. We view this as an open question that warrants further exploration through future simulation–estimation studies, and thus did not attempt to prescribe what "strong evidence" should look like in this manuscript. However, our results suggest that bias correction is most defensible only when variance parameters associated with random effects can be estimated accurately, potentially through approaches such as REML. In the absence of such reliable variance estimation, applying bias correction risks introducing additional bias. Therefore, while we do not provide prescriptive guidance here, we emphasize in the Discussion that further research should focus on identifying conditions under which variance parameters can be estimated without bias, and only then would bias correction be justified.

3 Reviewer 1 Comments to the Authors

Major comment 1

> I'm not the biggest fan of the presented plots. You have 4 cases for each random effects structure regarding whether the bias correction is on or off for the random effects and observation terms. But all of the figures only use BC-ON and BC-OFF but it's not clear how things are divided. Where does an OM with bias correction on in random effects and off in observations? Is that BC-ON or BC-OFF? I think the x-axis should include all 4 cases in all the relevant plots. I think it would be much clearer and hopefully more revealing about what is happening. Some of the text could also be clearer in this regard to which of the 4 cases it's referring to.

Response

We apologize for the confusion and have modified Sections 2.4 and 2.5 of the manuscript to better describe the study design. For each random-effects structure (e.g., recruitment random effects only), we considered four OMs (process BC-ON/obs BC-ON, process BC-ON/obs BC-OFF, process BC-OFF/obs BC-ON, and process BC-OFF/obs BC-OFF). Each OM was then fitted with four EMs under the same four bias-correction combinations, resulting in $4 \times 4 = 16$ OM–EM combinations per structure. With four random-effects structures examined across three stocks, this produced $64 \times 3 = 192$ performance outcomes in total. Presenting all of these in the main text would likely overwhelm the reader and obscure the key takeaways.

To provide a clear high-level summary, we simplified the visualizations in two ways. First, we found that observation-level bias correction had minimal influence in our case study (e.g., process BC-ON/obs BC-ON was nearly indistinguishable from process BC-ON/obs BC-OFF). Therefore, in the main text we focused on the two extreme cases: both process and observation BC-ON vs. both BC-OFF. Second, we observed very similar patterns between the AR(1) and IID autocorrelation structures, so we combined them in the main figures to further simplify the presentation.

With these steps, the results could be summarized in 24 combinations, which we believe captures the essential takeaways without losing clarity. For transparency, the full set of results (all four OM $\times$ four EM cases for each random-effects structure and stock) is included in the supplementary material, and this is now explicitly noted in the revised text and figure captions.

Major comment 2

> This is really more of a question but I'm wondering if you've thought about the fact that SAM and some other state-space stock assessment models fit directly to the log numbers at age rather than aggregate indices and proportions, do you think this would impact the results of bias correction on observations differently from what was seen here on WHAM?

Response

We appreciate this thoughtful question. We agree that fitting directly to log numbers-at-age, as in SAM and some other state-space models, could lead to some differences in the influence of bias correction on observations, particularly depending on the likelihood assumptions used for age composition data. However, we do not anticipate that the general patterns we observed would change. In our view, there is nothing unique about using log numbers-at-age that would fundamentally resolve issues related to imprecise or biased estimates of process variance, or the resulting problems associated with bias correction. Thus, while model-specific details could produce some variation in outcomes, we expect the overall conclusions from our study to hold.

Minor Comments

Comment
> Change N in Equations 2, 3, and 6 to use N as in Equation 7 to better distinguish between the numbers-at-age matrix and the normal distribution.
Response
Fixed.

Comment
> In Equation 8, the numerator should be ˆi,y.
Response
Fixed.

Comment
> Lines 144 and 145 use theta(i, y) and theta_hat(i, y), but they should match the notation in Equations 8 and 10.
Response
Fixed.

Comment
> Line 157: AIC should be defined. You should also define dAIC here and not only in Figure 5.

Response
Both AIC and dAIC are now defined in the main text.

Comment
> Table 1: It is unclear what Dirichlet-miss0, etc., mean for the age composition likelihood. Does it mean there are zero missing values? This should be better explained in the text or in the table caption. Also, consider expanding "Age Comp." at least once. How many parameters are estimated for the age-specific selectivity — all A of them, or fewer?
Response
A detailed explanation has been added to the table caption for clarity, including an expansion of "Age Comp." We chose not to include the specific number of selectivity parameters, as this detail is not directly relevant to the focus of the paper.

Comment
> Equations 9 and 10: Define what the abbreviation MdLQ stands for.
Response
Defined in the main text.

Comment
> Lines 110–111: The notation for the term involving C is not consistent. In one place you use the version with a bar over C, while in Equation 7 and earlier on line 110 you use the version without the bar. Please make this consistent by choosing one form and using it throughout.

Response
Fixed.

Comment
> Figure S1: Would be better as a table listing the proportions of EMs that converged, since it's hard to read values from a bar plot. A grouped table would be more compact and useful.
Response
Converted to a table (see Table S4).

Comment
> Many figure captions have quotation marks that need fixing (should be consistent — either proper directional quotes "like this" or straight quotes).
Response
Fixed.

Comment
> Caption for Figure 5: Should say "proportion of models where AIC …" instead of "probability."
Response
Fixed.

Comment
> Section 3.4: Could use a reference for the dAIC rule of thumb.
Response
We note that there is no universal rule of thumb for dAIC, but a threshold of 2 is commonly used in statistical analyses.

Comment
> Figure S13 caption: Typo, should be "estimates," not "estiamtes."
Response
Fixed.

4 Reviewer 2 Comments to the Authors

Major Comment 1

> The authors conclude that using bias-correction may do more harm than the potential gain of using it. This conclusion is based on the effect of BC of the process error for age > 1, where using an EM with BC when the OM has no BC is worse than using an EM without BC when the OM has BC, but mainly only for one of the three fish stocks. I think one out of three is not enough evidence for a strong conclusion. And I think the authors misinterpret the results of the simulation study presented in Figure S10, where they ignore the asymmetry of the measure of relative error.

Response

We thank the reviewer for this valuable comment. In the revised manuscript, we further evaluated our results using multiple performance measures beyond the asymmetric relative error originally presented. Specifically, we now include: (i) the root mean squared relative error (RMSR), which incorporates both bias and variance (Figures S4–S6); (ii) symmetric metrics such as the Median Log-Quotient (MdLQ, or log-relative bias; Figures S7–S9); and (iii) the Median Symmetric Signed Percentage Bias (SSPB; Figures S11–S13). We also increased the number of simulations from 50 to 100 per scenario, to address concerns about sample size raised by other reviewers and to ensure each distribution is more representative.

Across all these metrics, including the symmetric ones noted by the reviewer, we found consistent evidence that supports our conclusion: using bias correction on process error often increases bias, particularly when estimation models are misspecified relative to the operating model. Thus, the expanded analysis strengthens rather than weakens our conclusion that caution is warranted when applying process-level bias correction.

Major Comment 2

> The measure for relative error takes first the median over years, which ignores variance. Then the variation between simulations is presented by box-plots, which tells us both about bias and variance. I would prefer a measure that takes into account the variance in the first step as well.

Response

We thank the reviewer for this helpful suggestion. As noted in our previous response, we have added the root mean squared relative error (RMSR) as an additional performance metric. RMSR incorporates both bias and variance directly, addressing the concern about variance being ignored in the first step. The results using RMSR (Figures S4–S6) are consistent with our other metrics, and the overall conclusion remains unchanged.

Major Comment 3

> This comment is personal, and maybe out of the scope of the paper: According to the results, it is only BC of the process error for age > 1 that has any effect. But why is the process error needed and what does it represent? Mortality, and therefore survival, is already accounted for by fishing mortality and natural mortality, and WHAM also allows for random variation in the natural mortality.

Response

We appreciate this thoughtful comment. A full exploration of the role of process error in numbers-at-age (NAA) is beyond the scope of this paper, but we have added a short clarification and citation in the revised manuscript. As discussed in Stock and Miller (2021) and Stock et al. (2021), NAA random effects allow additional deviations from expected survival beyond what is captured by fishing and natural mortality alone (e.g. migration, disease). These random effects provide flexibility in cases where mortality components alone do not fully explain the observed dynamics and avoid attributing all variation specifically to natural mortality. Stock et al. (2021) also found that including random effects on both NAA and M improved model fit compared to restricting random effects to one component only, as measured by AIC.

Citations

- Stock, B.C. and Miller, T.J. (2021). The Woods Hole Assessment Model (WHAM): A general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental covariates.

- Stock, B.C., Deroba, J.J., & Miller, T.J. (2021). Implementing two-dimensional autocorrelation in either survival or natural mortality improves a state-space assessment model for Southern New England-Mid Atlantic yellowtail flounder.

Minor comments

Comment
> Line 44: "Bias, or estimation error, …". It seems that the authors use this as synonyms, but estimation error should include all errors, i.e. both bias and variance.
Response
Change to "Bias"

Comment
> Eq. (1): Here, log recruitment is log(f(SSB)) + epsilon, and SSB depends on year. But in line 154, the authors mention mu_Rec, which is a constant. So, is log recruitment log(f(SSB)+epsilon, or mu_Rec + epsilon ?
Response
Fixed.

Comment
> Eq. (1): My question here is probably out of the scope for the paper, but I ask anyway. It

is about the process error for a>1: For a>1, what does epsilon represent? Is it migration, since mortality (Z=F+M) is already accounted for by Z? In line 249, the process error is variation in survival, but when mortality is accounted for by Z, then the survival is also accounted for.

Response

We've changed the equation 1 for clarity. here can referred to the Stock and Miller (2021): "The deviations are akin to survival, which is clearly related to , but they can also be caused by mis-specification of selectivity of the fishery or indices, movement into or out of the population, or misreported catch (Gudmundsson and Gunnlaugsson, 2012, Perretti et al., 2020, Schnute and Richards, 1995, Stock et al., 2021)."

Comment

> Eqs. (3)-(6) and the text around. There is need to use ^ here and talk about estimated values. It is sufficient and better to only use Ry and E(Ry), i.e. the true values in the model itself. And Rbar,y can be mean recruitment, it is not yet estimated. And if you as here want to use R for recruitment, you should write that Ry=N1y in Eq. (1) and Rbar,y=f(SSBy).

Response

Fixed.

Comment

> Line 105: I am used to that "«" means much less than, but here the authors maybe mean "Ry is always less than E(Ry) and the difference becomes larger as the variance sigma^2 increases.

Response

Fixed.

Comment

> Eq. (7): You should write what C and Chat is, because the notation here and in Stock and Miller (2021) The Woods Hole Assessment Model (WHAM) is the opposite of what I would prefer and expect. In statistics, it is usual to write the model itself without hats on (true, but unknown) parameters and other quantities, and introduce hats only on estimates. In your case, Chat is the unobserved, true catch, and C is the observed catch, i.e. a point estimation of the true catch.

Response

Fixed.

Comment

> Line 137: When one model did not converge, was another pseudo-data set generated, so it always ended up with 50 pseudo-datasets?

Response

Yes. This has been rewritten for clarity see section 2.5: "The simulation experiment followed a full factorial design (Table 2). For each OM variant, fixed-effect parameters (including variance parameters for random effects) were used to generate 100 realizations of recruitment and population dynamics. Observation error was applied to produce 100 pseudo-datasets per OM. These pseudo-datasets were then fitted withall four estimation model (EM) variants differing in bias correction treatment, producing all OM–EM bias correction combinations within the same random-effects structure. This design produced both self-tests (EM bias correction treatment matches OM bias correction treatment) and cross-tests (EM bias correction treatment differs from OM bias correction treatment) (Deroba et al., 2015). A model was considered converged if two criteria were met: (1) the optimizer had successfully converged, and (2) the Hessian matrix was invertible. Simulations in which any of the four EMs failed to converge were discarded, and additional iterations were run with new random seeds until 100 complete and converged simulation replicates were obtained for every scenario".

Comment
> Section 2.3: I think both the authors and the readers at this point should think about the motivation for using bias-correction or not here. If the point estimate C is constructed such that it is intended to be mean-unbiased, then one should consider using bias-correction. If it is median-unbiased (not so usual, I think), one don't need bias-correction. On the other hand, the abundance indices are only indices, so bias-correction is probably not important at all for estimation SSB and NAA, unless the variance of epsilon change over years. This is because the catchabilities q_a will cancel the effect of exp(0.5 sigma^2).

Response
We thank the reviewer for this thoughtful comment. In the revised manuscript (Section 2.3), we expand the motivation for bias correction in log-normal observation models. We clarify that, with bias correction, the mean of the observation distribution equals the true catch (mean-unbiased); without bias correction, the median equals the true catch (median-unbiased), and the observation is mean-unbiased on the log scale. In our study, observation variance was fixed and small, so choosing a mean- vs. median-unbiased representation has negligible practical impact; the two are related by the $^2/2$. For survey indices, we note that bias correction is less influential because catchability (q) can absorb much of the adjustment (Section 2.3). We have made these points explicit in the revised text, and the broader implications are discussed in the Discussion (Lines 286–299).

Comment
> Lines 124 and Table 2: Is the first 8 rows in the table without random effect for Nay for a>1, i.e. only for recruitment N1y?

Response
We think our current version is clear and informative.

Comment
> Eq. (8): theta_hat_i in the numerator should include the year y. And the notation must be consistent. Either write theta_i_y with i and y as subscripts or theta(i,y).

Response
Fixed.

Comment
> Eq. (8):
> a) By taking the median over years, one average out errors. Two different methods, one with large errors and one with small errors, can still have similar median (or mean) over years. I think one should compute a measure that contain all the errors, for instance the median of the absolute values. One could perhaps alse keep the present measure, which focus more on bias.
> b) But when one is in year y (or y-1?) and want to estimate SSBy and set fishing quotas, is then the estimates SSBx for previous years x<y of interest? I think it is more relevant to focus only on the last year. For each pseudo-dataset, one can start with 2/3 of the data as training set and predict SSBy for the relevant year y. Then one can increase the training set by one year and predict for year y+1 etcetera.
> c) NAA depends on both age and year, and it is not obvious how this is handled here. Is the median over both ages and years?

Response
a) We evaluated our results using multiple performance measures beyond the asymmetric relative error originally presented. Specifically, we now include: (i) the root mean squared relative error (RMSR), which incorporates both bias and variance (Figures S4–S6); (ii) symmetric metrics such as the Median Log-Quotient (MdLQ, or log-relative bias; Figures S7–S9); and (iii) the Median Sym-

metric Signed Percentage Bias (SSPB; Figures S11–S13). All these results support our conclusion.
b) this paper does not address projections or prediction. This is beyond the scope.
c) That's the median over both ages and years, to provide a big picture about NAA estimates.

Comment
> Lines 147-153: This could maybe be put into the supplementary material, since the results are not shown. And why not simply use log(theta_hat)-log(theta) instead of SSPB?
Response
We now have results using log relative error in the supplementary file.

Comment
> Line 157: What is delta AIC here?
Response
We now define dAIC in the main text. See lines 162-164.

Comment
> Lines 165-166 and Figure S1: As far as can see, there are no self-tests in Figure S1. And if you are going to talk about convergence, shouldn't you say a little bit more in the text?
Response
Fixed. We provide reasoning in lines 179-186: "We found that convergence failures were often realization-specific rather than model-specific. In most cases, if one EM failed to converge for a given realization, all EMs failed for that same realization, suggesting that convergence issues were driven by factors other than the bias correction setting. Therefore, to better isolate the effect of bias correction, we calculated a conditional convergence rate: the proportion of simulations that converged for the misspecified model, given that the correctly specified self-test had already converged. This provides a clearer evaluation of how the bias correction mismatch impacts model convergence, with results summarized across all stocks (Table S4)."

Comment
> Figure S1: It is difficult to read the figure since everything is close to 1. It is better to plot the convergence failure rate instead.
Response
Table 4 with values of convergence rate were produced to replace Fisgure S1

Comment
> Lines 166-168: Does this mean that results for half of the experiments in Table 2 are not shown, i.e. rows 2, 3, 6, 7, 10, 11, 14,15?
Response
Yes, we provide reasoning in lines 168-178.

Comment
> Line 173 and Figure 1. One can expect this result,mu exp(epsilon) = $\text{mu}^{\text{exp(epsilon-0.5sigma}}$2), where mu=mu exp(-0.5sigma^2), this is only a transformation. This could be mentioned already in Section 2.2. This is only important for the interpretation of mu as the mean or the median recruitment.
Response
We think estimation error of mu and sigma is less known in state-space models when only recruitment is rando meffects and need to be included to see wheterh the corresponding mean or median is estimated accurately, along with annual recrtuiemtn estimates.

Comment
> Figure 1: From the figure, it is impossible to understand when BC is on or off for OM and EM.

Bu t by reading the text at lines 178-181 I understand that the text below the figure is for EM. I suggest including one line in the figure with BC on/off for OM as well, and clearly state what line is for OM and what is for EM.

Response

title for x-axis shows the bias correction for EM and facet tile shows self-test or cross-test, which can be translated to OM bias correction structure.

Comment

> Line 176: What do you mean by "scale"? Do you mean the variability, i.e. the width of the boxes?

Response

This section has been revised. See section 3.1

Comment

> Section 3.4: In general, if you have two models with equally many parameters and they are equally simple and you have no prior preference, one would choose the one with best likelihood and the green columns in Figure 5 show that you do that in 60-80% of the cases. But when you use that the difference in AIC should be larger than 2, it seems that you have preference for the wrong model, since you in some cases never choose the correct model. This should be formulated differently.

Response

We revised Figure 5, but the results are still basically the same. The figure shows how model choice depends on the cutoff we use for AIC differences (ΔAIC).

- If we use a relaxed rule ($\Delta\text{AIC} > 0$), then in many cases the correct model is more likely to be chosen.
- If we use a stricter rule ($\Delta\text{AIC} > 2$), we only look at simulations where the AIC clearly favors one model over another. For example, if Model 1's AIC is at least 2 points lower than Model 2's, then Model 1 is treated as the "best-fit" model. We then count how often this best-fit model is also the correct one (in terms of bias correction).

If the AIC difference is smaller than 2, we don't count that case in the figure because it means the models are too close to call. This is why the percentage in the figure doesn't simply mean "the rest are wrong"—some of the missing percentage is made up of these "uncertain" cases where neither model was clearly better.

Comment

> Line 229: Assume there are no bias-correction in either OM or EM: If sigma_Rec is underestimated, then exp(0.5 sigma_Rec^2) is also underestimated, and the mu_Rec must be OVERESTIMATED to compensate.

Response log(Rec) = log(mu_Rec) - 0.5 * sigma_Rec^2 + random_effects. In this formula, mu_Rec is the parameter for the true average (mean) recruitment. If mu_Rec is underestimated, the bias correction term (-0.5 * sigma_Rec^2) becomes a smaller negative number. To compensate and get the right answer for log(Rec), the model must underestimate sigma_Rec.

Comment

> Line 233: The "model" here is maybe the operating model?

Response

Fixed.

Comment
> Lines 261-263: What do you mean with "observations with a log-normal assumptions have limited impact on derived population quantities"? The observations are of course important. Do you mean that the bias correction has little impact? And when you write "have", do you men in general or for the three stocks that you have investigated?

Response
Fixed. See lines 285-291.

Comment
> Lines 263-265: This is probably correct for catch. But still, bias-correction for catch will have a systematic effect on the estimate of SSB. With bias-correction, the estimate of SSB will be systematically higher, but usually not much, see for instance Aldrin et al (2020) https://www.sciencedirect.com/science/article/pii/S0165783620301028?via%3Dihub

Response
We provide the comparison between our study and Aldrin et al. (2020) in discussion lines 292-302.

5 Reviewer 3 Comments to the Authors

Major Comment 1

> I do agree that bias-correction should be applied to observations, as in Section 2.3. Eqn (7) is saying that the observed catches C are biased estimates of the modelled catches hatC. The authors should justify why this bias occurs. Further, from a theoretical statistical standpoint the maximum likelihood estimator is invariant to transformation of the data, i.e. we get the same parameter estimates whether we fit to C or log(C) (https://en.wikipedia.org/wiki/Maximum_likelihood_estimation: "The MLE is also equivariant with respect to certain transformations of the data…."). For normally distributed data (log(C)), nobody would suggest a bias correction. I may be wrong, but if I am right, the authors are bringing confusion into the discussion about bias correction. (My reading of the results and conclusion of the paper, bias correction applied to observations does not have a big effect)

Response

We thank the reviewer for this thoughtful comment. In the revised manuscript (Section 2.3), we expand the motivation for bias correction in log-normal observation models. We clarify that, with bias correction, the mean of the observation distribution equals the true catch (mean-unbiased); without bias correction, the median equals the true catch (median-unbiased), and the observation is mean-unbiased on the log scale. In our study, observation variance was fixed and small, so choosing a mean- vs. median-unbiased representation has negligible practical impact; the two are related by the $\sigma 2/2$. For survey indices, we note that bias correction is less influential because catchability (q) can absorb much of the adjustment (Section 2.3). We have made these points explicit in the revised text, and the broader implications are discussed in the Discussion (Lines 286–299).

Major Comment 2

> My next point is similar, but applies to Section 2.2. I do not agree that there should be an operating model without bias correction. If the expectation of exp(epsilon_1) in eqn (1) is not 1, it affects the interpretation of the recruitment function f(SSB). The paper does not give details about f(SSB), so we do not know if f(SSB) will be correct for the missing bias correction. Hence, by opening up for E[exp(epsilon_1)] neq 1, I feel the authors are making the discussion too general.

Response

Results are invariant to using density-dependent stock–recruit functions: with multiplicative lognormal noise of constant variance, dropping the lognormal bias correction rescales productivity by $\exp\left(\frac{1}{2}\sigma^2\right)$) but leaves the density-dependence unchanged; hence our conclusions about the direction of bias and estimator behavior are unaffected.

Major Comment 3

> The number of simulation replica is low only 50. By increasing by x10 would improve the precision of the results, but if the authors feel they can show their results with only 50 replica, I guess it is OK.

Response

we've increased the sample size from 50 to 100 for this study.

Major Comment 4

> The paper is in general clearly written, but a bit sloppy written at some places:
> a. Line 104: "larger than" followed by "neq"?
> b. In eqn (8), is it median over y? Also, first it says theta_{i,y}, and right after is says theta(i,y)

Response

Fixed.

6 Reviewer 4 Comments to the Authors

Major Comment 1

> Only 50 simulation replicates? This seems quite low. I hope there is more compute power available at the NEFSC to increase this number to at least 100!
> Lines 114-122. I need a few more details here. Also please cite the assessments. My current best assumption: So for each of the 16 scenarios (OM * BC), the model was fit to actual data to estimate parameters. Then new data were simulated (50 times). And then within a given OM structure, a full factorial (basically each of the 4 EMs wit BC options) was to fit to each BC option on the OM. Please re-write this paragraph to make it explicitly clear what was done. Also best to use consistent terminology throughout for the OM structure and the bias correction so as to not confuse the two. I see some places referred to as OM structure, others as random effects structure.

Response

Response: we've increased to 100.
Response: Assessments are cited. The paragraph has been revised. See section 2.4 and Section 2.5.

Major Comment 2

> If one EM of the four did not converge you threw out that simulated iteration in favor of another that resulted in the four EMs converging? Could you not be biasing your results here? There could be some aspect of a simulated dataset where some EMs perform very poorly (and thus aren't converging) that you are ignoring/throwing out? This is always a difficult aspect of simulation studies as in the real world there may be a number of options an analyst may employ to try to get a model to converge. One possible way around this is to crank up the simulations and report the % non-converged as a performance metric in the main results such that you're considering RE of the converged models (perhaps even only of the iterations where all 4 converged) and % non-convergence together. I see you report this perhaps in Figure S1. I would argue bring this into main text. Seems to potentially be slightly more evidence in favor of no bias-correct in EM.

Response

We provide our reasoning in (Lines 179–186) "we found that convergence failures were often realization-specific rather than model-specific. In most cases, if one EM failed to converge for a given realization, all EMs failed for that same realization, suggesting that convergence issues were driven by factors other than the bias correction setting. Therefore, to better isolate the effect of bias correction, we calculated a conditional convergence rate: the proportion of simulations that converged for the misspecified model, given that the correctly specified self-test had already converged. This provides a clearer evaluation of how the bias correction mismatch impacts model convergence, with results summarized across all stocks (Table S4)."

Major Comment 3

> Is the negative bias in the variance terms not simply the bias associated with not using REML? In some previous research (Figure 7, Fisch et al., 2023) we found this bias to be incredibly minimal, which I think your paper shows as well in Figure 3.

Response

We agree that some downward bias in variance components can stem from using ML rather than REML, and the bias in recruitment variance is minimal in our study (Fig. 3). In contrast, the NAA variance shows a non-trivial downward bias (Fig. 4). Because the NAA process is also lognormal, underestimating can propagate and amplify bias in states and derived quantities on the natural scale.

Minor Comments

Comment
> Line 29. Consider using alternative language to "noise" throughout.
Response
We used "observation error" instead

Comment
> Lines 70-76. I see this in the Intro but nowhere in the methods. i.e., describing the four RE structures. Perhaps remind the reader on line 117.
Response
Fixed, see lines 72-74: "We explored scenarios where either recruitment only or both recruitment and NAA were treated as random effects, with different autocorrelation structures (see Section 2.4 for more details).

Comment
> Line 105. Much less than symbology («). Are you not meaning to say the Rbar decreases relative to E[R] as sigma increases (or vice-versa)?
Response
Fixed.

Comment
> Equation 7. What is the meaning of i subscript here? Is it needed for this paper?
Response
Fixed. No need to have i.

Comment
> Lines 120-121. I assume here you mean EMs were fit to the pseudo-data generated from the OM options? If not, I am not entirely following. The OMs were fit to their own generated data?
Response
The section 2.4 and section 2.5 have been revised for clarity.

Comment
> Line 128. Consider replacing the term noise. Perhaps with error here. Or random sampling variability.
Response
Fixed. Used observation error instead.

Comment
> Equation 8. Missing y subscript in numerator
Response
Fixed.

Comment
> Methods. Please list your standards for convergence.
Response
Fixed. See lines 137-141.

Comment
> Line 164. Does this result in 100 simulation replicates for a given RE result or plot? If so please note.
Response
Actually 200 simulation replicates, we combined results of iid and ar1 to simplfy visualization. We clarify this in lines 168-174.

Comment
> Line 165: How are you calculating convergence rate? I only recall mentioning dropping simulations when any model did not converge in favor of an additional simulation where all 4 converged.
Response
We added more text for explaination and justification (see Lines 179–186) "we found that convergence failures were often realization-specific rather than model-specific. In most cases, if one EM failed to converge for a given realization, all EMs failed for that same realization, suggesting that convergence issues were driven by factors other than the bias correction setting. Therefore, to better isolate the effect of bias correction, we calculated a conditional convergence rate: the proportion of simulations that converged for the misspecified model, given that the correctly specified self-test had already converged. This provides a clearer evaluation of how the bias correction mismatch impacts model convergence, with results summarized across all stocks (Table S4)."

Comment
> Figure S1. What are the meaning of the colors here?
Response
This figure has been updated.

Comment
> Lines 166-168. I am not entirely following what is meant here. Are you dropping some results? If so can you place the word "only" on line 168? "Results were only reported… "
Response
See lines 176-178: "Therefore, to simplify the analysis, we restricted both the OMs and EMs to two primary configurations: one where bias correction was applied to both processes and observations (hereafter 'BC-ON'), and another where it was turned off for both (hereafter 'BC-OFF')."

Comment
> Lines 183-184. But more so for biascorrect=OFF in the EM.
Response

Throughout the Results section, BC-ON denotes that the lognormal bias-correction term is applied to both the process and observation models; BC-OFF denotes that it is applied to neither. We did not include operating or estimation models with mixed settings (e.g., BC-ON for the process but OFF for the observation model, or vice versa).

Comment
> Figure 3. This must simply be the bias in not using REML.
Response
Agree.

Comment
> Line 200. Please provide citation for your rule of thumb. My old rule of thumb was 3 units from Burnham and Anderson (2002)
Response
We note that there is no universal rule of thumb for dAIC, but a threshold of 2 is commonly used in statistical analyses.

Comment
> Figures S10-S11. Strange that the difference in rec sd here has no effect at all. Double check.
Response
In our sensitivity analyses, varying the recruitment SD had negligible effect on bias; the patterns were driven primarily by the magnitude of the NAA process SD. We have clarified this point in the text and note the range explored.

Comment
> Lines. 265-268. Is this true for composition data as well? I would assume you're allowing for the estimation of the Dirichlet and logistic normal terms.
Response
Yes, for composition data, we modeled overdispersion with Dirichlet and logistic-normal likelihoods and estimated their 'dispersion' parameters. Observation error variance for fleet catch was fixed at known, low values (CVs <= 0.2), following the configuration used in recent assessments for these stocks. Fixing the variance prevents it from being confounded with the process variance during estimation but may also limit the potential influence of the observation-level bias correction.

Comment
> Lines 268-270. I don't think we examined bias of the sampling variance terms in Fisch et al. (2023) as the simulator was not the same as estimator (it was a fine scale simulator and zoomed out EM). I think we focused on whether it was estimable. In fact, I think the results of Fisch et al. (2023) align more with lines 270-275. We found the opposite to the referenced sentence, that the sampling variance terms of the composition likelihoods were actually accounting for process variance.
Response
Agree and fixed. Please see lines 268-270.

Comment
> Lines 277-286. I personally have found the REML bias to be very minimal, and I think your results show that as well (Figure 3). I also think the iterative procedure that it requires add some additional complications and downsides.
Response
Agree.