Dear editor,

Add as additional change / comment below Discovered that first line of results had an error (simulations with 1 maternal female were excluded, when they actually weren’t – but these usually failed and were excluded due to lack of convergence)

|  |  |
| --- | --- |
| Green | Agree to change |
| Red | Disgreed to change |
| Blue | Comment / no change required |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Comment | Response |
| 1 | Associate editor | It occurs to me that placing a prior/penalty on Pmax could be helpful in practice given the inaccuracy of parameter estimation with many realistic sample sizes. This could be thought of as a statistical way to achieve a middle ground (as discussed by Reviewer 1) between a highly uncertain but estimated parameter and a model with a fixed parameter. I don't think you need to implement that here, but the authors may consider discussing this as well as how one might choose a prior (e.g., from related populations or species or using something like the algorithm used to choose a fixed Pmax). |  |
| 2 |  | Line 224: Would MARE not be better referred to as accuracy (a measure combining bias and precision) as it is on line 261 or with the word ‘accurately’ in Figure 3 and 6 caption? | Yes this is correct, changed |
| 3 |  | I would encourage the authors to publish or share what code and data they can with the manuscript. Even if the empirical data component cannot be shared, the simulation testing, and most importantly, the TMB models with examples of fitting them to data in R would be highly valuable to readers and increase the chances for uptake of the recommendations in this manuscript. | A Github repository will be made public to share the code, including an example using the empirical dataset. |
| 4 | Reviewer 1 | .. a main point illustrated in the ms is that sample sizes must be extremely large in cases where reproductive periodicity is less than annual, and even in the best-case examples of focused studies such as the sandbar dataset used in this paper, the outcome is still a bit uncertain. In a field that is inherently undersampled, this produces some difficulty in finding real utility for this type of model in routine stock assessment, where decades of study might produce sample sizes of 500-1000 individuals of a species, even globally. The sandbar shark example was the product of an entire scientific research fishery effort, a rare feat in elasmobranch science. Even at 1000+ samples, the sandbar data are uncertain, and that uncertainty translates to the maternity function models. The question then becomes: Is it better to know the model is wrong (relative to empirical data) and to be able to quantify how wrong it is, than to empirically estimate parameters with some less-knowable and less quantifiable amount of uncertainty associated? Reproductive biologists would say the latter is most logical, but assessment scientists may disagree. I think the answer lies somewhere in the middle, and is where this type of exercise can be quite useful, but great care and caution is needed in portraying the results for future use. This is where the ms needs some improvement, as I think a careful review of the results captures the above sentiments, but the discussion leads off with more optimism than is warranted. | I think this statement is overly pessimistic. I suspect many fisheries agencies (my own included) are probably sitting on data sets that could utilised by this method but currently aren’t due to a lack of guidance. Yes, there are many data-poor stocks, but there are still A LOT of species and stocks.  The second point is very valid. I have touched on it with the comparison |
| 5 |  | … a key point that is overlooked in the sandbar analysis is the fact that the age data used to inform those models have been proven to be underestimated (Andrews et al. 2011). The lead author knows better than most how common this is in elasmobranchs, and yet it is not even introduced as a discussion point despite the conclusion that life history affects outcomes of different modeling methods (i.e. estimated vs fixed Pmax). How would increased Amax impact the estimated Pmax and other outcomes of the sandbar analysis? It’s logical that increasing Amax would yield higher lifetime productivity no matter the actual Pmax, but perhaps this adjustment could account for the difference in modeled Pmax=0.5 and the 2.5 (range of 2-3) year breeding frequency observed empirically in the Baremore study. A sensitivity analysis to explore this may be beyond the scope of the paper, but at a minimum should be discussed thoroughly. Lines 284-293 may be a good spot, and lines 368-377 in the discussion need an entire revamp. Growth models based on validated age data are rare and unvalidated models are still routinely used in assessments because they are still the “best available science” given no better ageing alternatives. Hence, this point is germane to the conversation around how useful this type of modeling could be in reality. | I disagree with this comment. As far as I can tell, age underestimation shouldn’t have a major impact on the outcomes of this study.  A per-recruit model was used to generate catch at length data used for the simulations. This model is relatively insensitive to changes in maximum age (see Figure 1 below). Growth curves for both these species have also been validated, that is partly why they were chosen. The calculation of R0 (Section 2.2.3) used a continuous time model that doesn’t make assumptions about maximum age (in effect there is no Amax). The sandbar shark empirical example doesn’t incorporate age in any way, and is wholly length based, so age underestimation isn’t a relevant consideration here either. |
| 6 |  | Abstract: A sentence with caveat about the need for large sample sizes to produce “feasible” results is warranted. |  |
| 7 |  | Line 14: The sentence reads more easily to me as “Maturity parameters could also be estimated with greater precision; however, substituting them for maternity parameters overestimated lifetime reproductive output.” | Changed |
| 8 |  | Line 83: a mention here that sampling for well over a year is generally necessary for species with resting periods is warranted | Added |
| 9 |  | Line 116: see also Natanson et al. 2018 regarding porbeagles | Natanson identified that the porbeagle reproductive cycle was biennial, rather than annual, due to lack of sampling in previous studies. It wasn’t an example of plasticity in reproductive cycles, and as such as I haven’t cited here. |
| 10 |  | Line 227-228: From a non-modeller perspective, this seems like a low bar for well-performing prediction. I suggest carrying this idea through to the discussion about interval coverage to help keep things in perspective. | This is to check that the model can adequately capture uncertainty. It isn’t related to the predictive capacity of the model. |
| 11 |  | Line 275-276: This is entirely expected! Also the same for lines 279-281… less datapoints means less certainty | These do make intuitive sense, but important points to note for those that may not be familiar with the analysis. |
| 12 |  | Lines 284-293: I really like this paragraph and I think these data are very interesting. Expanding to include how age at length and Amax uncertainty can affect these outcomes would really add a lot of value. Also, while potential spatial variability in periodicity is kind of implied earlier, it would be good to mention it as another possible biasing factor, if not here then elsewhere. |  |
| 13 |  | Discussion: add life history and sample size caveats where appropriate. |  |
| 14 |  | Lines 314-315: must add caveats to this… perhaps some advantages, but at what sample size? And are they true advantages or just better estimation of uncertainty but not necessarily of the desired parameters? |  |
| 15 |  | Line 321: not desirable, necessary! | Changed |
| 16 |  | Lines 324-325: this sentence and really the entire paragraph should lead off this section, in my opinion. It more accurately reflects the reality of these findings |  |
| 17 |  | 344-345: Fujinama (2017) parenthetical is missing and name misspelled. This section is also interesting re: the age uncertainty discussion, perhaps warrants a sentence there. | Changed |
| 18 |  | 368-377: This needs total revision and incorporation of caveats surrounding age underestimation and the effects on these results. | See Point 5. I don’t believe age underestimation is likely to have a major impact, and it doesn’t factor into the empirical analysis as this was entirely length-based. |
| 19 |  | 396-398: While true, the uncertainty resulting from sample sizes typical of assessments would still be crippling. Please acknowledge that here. |  |
| 20 | Reviewer 2 | …Table S8 provides median number of females in maternal condition, which I believe is the more meaningful number to report (and possibly Table S9, although I can’t tell how important the number of immature females was to the analysis). | Yes, you’re absolutely right, I’m not sure why I hadn’t thought of this. |
| 21 |  | L101-104: Can you demonstrate that using maturity with fecundity divided by gestation period produces a different (and biased) result compared to the 3PLF? | No I can’t. The point, as stated at the end of the paragraph, is that this is an important, and widely-used assumption that has never been tested, as far as I’m aware. |
| 22 |  | L179: why use a cut-off Amax instead of a plus group (using formula for the sum of a geometric series)? There can be non-negligible survival beyond Amax that impacts calculation of lifetime reproduction (which appears to be how it is done in 214-217). | I was under the impression that the formula in Line 178 described a plus group rather than a cut-off Amax.  As noted, I also did use this approach in lines 214-217 for the calculation of lifetime reproduction, where survival beyond Amax is likely to have a bigger impact. |
| 23 |  | L181-182: did you just scale N to achieve the catch in number scenarios? And do the sample sizes refer to catch of all individuals selected by the gear or just mature? This seems germane to the interaction of sample size, reproductive frequency and selectivity (Figure S2, S3) – with a lower sample size, there are fewer individuals in maternal state for quadrennial compared to annual and this discrepancy is augmented by low selectivity. | No, lengths were randomly sampled based on probabilities from the catch equation. I have added more details in the methods. |
| 24 |  | Figure 2: there is barely a difference between 3PLF estimated vs 3PLF fixed; why? What sample size did this plot correspond to? | Based on this comment and comment 20, I have decided to redo this Figure. I have removed the ‘fixed’ panels, as they show a similar trend to the ‘estimated’. I have matched the color palette to the other graphs and also scaled the point size by the number of maternal females. |
| 25 |  | Also, if it is always the case that L’50 >= L50, would it help to impose L50 as a lower bound or would that increase the bias in Pmax? |  |
| 26 |  | L207-209: did you record in your results how often your guess was based on L99 vs L95 vs L50 vs Lall?  Presumably that would affect comparisons for methods that guessed at Pmax. Perhaps affecting coverage (Fig 4) and other performance aspects? | I didn’t record this, but yes it would potentially affect it in various ways. The choice of PMAX when using a fixed value is subjective. This is one of the downsides of using the 3PLF method in general, and also makes it tricky to evaluate its performance relative to other methods. I discuss this in the final paragraph of Section 4.1. |
| 27 |  | Figure 5 – the pattern is a result of the x-axis (selectivity:sample size) referring to different numbers of maternal individuals sampled. I think if you just change the order of categories on the x-axis from high:N, low:N, etc. to having the low N categories first (low:N, high:N, etc.) then the gaps disappear (or at least most of them, and then maybe one can better ask the significance of the remaining gaps). | Yes, this is particularly evident in Fig S8. I have changed throughout |
| 28 |  | L294: for the empirical case study, what was the selectivity of the gear relative to maturity and maternal condition? Were samples skewed to mature and/or maternal individuals? How did 2PLF-maturity compare to the 3PLF estimates? Are samples sufficient to separately estimate Gulf of Mexico and western North Atlantic populations? Were samples:maternal state similar for both populations? It would be a consequential switch to move from triennial to biennial reproductive frequency, so understanding this result is important. Also, Figure 7 does not seem to be cited (it refers to this case study). Can you relate the sandbar case study to any of the scenarios in your simulation to guess at potential bias? |  |
| 29 |  | L310-311: There were also scenarios where 2PLF-maturity performed better, yet this statement emphasizes the cases where 2PLF did not perform as well.  I think a more balanced statement would be “This study also demonstrated that recruitment can be overestimated in some cases…” |  |
| 30 |  | L319: it would be good to give advice on sample sizes of maternal condition fish, rather than raw sample sizes.  It is pretty clear from your results that there is an interaction between reproductive frequency, selectivity, and raw sample size. |  |
| 31 |  | L347-348: since you note that the more common approach is to divide fecundity by gestation period, it would have been very useful to know how that approach compares to weighting maturity curve by frequency of parturition.  As far as I can tell, dividing by fecundity assumes a constant reproductive frequency at all ages (or lengths) while weighting maturity curve by frequency of parturition will derive an average Pmax that would only differ from the common approach if reproductive frequency differs at length.  Since your simulation imposed a probability of or maternity at length (L186), one would expect the common approach to be biased. But, it might depend on the selectivity of the gear (or depletion of the population).  Can you report what the ‘guess’ at Pmax would be for the 2PLF-maturity with your outlined approach for the simulation (described on L205-209)? |  |
| 32 |  | L338-345: regarding the issue of L50 (and A50) underestimating L’50 and A’50 by 1 year or more, would your advice be that if assessment models follow the status quo approach (use maturity and scale fecundity by gestation period) then a sensitivity run should be attempted that shifts the maturity ogive ahead by one year (to account for the A50 vs A’50 difference)? |  |
| 33 |  | L35: …the proportion of “mature” females that will give … | Added |
| 34 |  | L65-66: does assigning maternal condition require more subjectivity than simply mature/immature? (“…if they would have given birth…by the end of a given year).  One would have to be able to judge stage of gestation and/or time to parturition/egg deposition. | Yes, assignment of maternal condition requires a detailed understanding of the timing and duration of the uterine and ovarian cycles. This is stated on lines 83 and 84. |
| 35 |  | L111: “Wider use of maternity functions…” – shouldn’t this be predicated on wider collection of data to be able to estimate maternity functions? | I would argue that in many cases the data are collected, the analysis just isn’t undertaken |
| 36 |  | L116 and L120: these two points seem contradictory (“despite observations of plasticity in this trait within discrete populations” vs “help understand the temporal stability of reproductive cycles”)? | Yes, this is actually the point I was trying to make here. We assume reproductive periodicity is fixed even though there is evidence that it isn’t (Line 114). I have slightly reworded this sentence to make it clearer that it is referring to reproductive periodicity. |
| 37 |  | Figure 1: The difference in selectivity for females in maternal condition is not clear. Perhaps additional text in the figure caption, or a different visualization would help. For example, both the bars in the background and the line plot in the foreground are different. I presume the line is selectivity, but why is the bar chart different? Maybe indicating where lengths equal to 25% and 75% of the population in maternal condition is would help? | Agree. I have added maternal status as a grouping variable in the length structure example in Figure 1. |
| 38 |  | L206: use either “a” or “the”  in the phrase “…condition above a the length…” |  |
| 39 |  | L219-221: what about looking at a threshold for the gradient? | Optimization was undertaken by passing the gradient of the TMB model objective function to nlminb. I used a positive definite hessian matrix to evaluate convergence, but yes an additional check could have been to examine the threshold of the gradient. |
| 40 |  | L246: perhaps “successful convergence” rather than “a success rate” (because it is not clear success refers to convergence here) | Changed |
| 41 |  | L261: precision rather than accuracy of parameter estimates? | This has been addressed as part of comment 2. Accuracy is correct. |
| 42 |  | Fig S10: other than at very low sample sizes with very infrequent reproduction (quadriennial, and to a lesser extent triennial), there is no real improvement using 3PLF-estimated over 2PLF-maturity. | Figure 5 and Figure 6 provides a clearer visualisation of this. The estimated method wasn’t particularly effective at minimising bias in gummy sharks, but worked better in school sharks. It was overall more accurate in most scenarios for both species |
| 43 |  | Figure 6: This is supposed to be summarizing precision, but the legend says something about accuracy. | This has been addressed as part of comment 2. Accuracy is correct. |
| 44 |  | L301: one fewer estimated parameter instead of parameters? | Changed |
| 45 |  | L331: how do you characterize “successfully estimate” for the Colonello et al study? | Success probably isn’t the correct word here. I have changed to plausible instead. |
| 46 |  | L374: knowledge of reproductive frequency is also important for estimating abundance in CKMR (see Swenson et al. 2024 DOI: 10.1002/ece3.10854 ) | Yes, this is an important reference I missed. I have added it, although think it fits better in the introduction. |
| 47 |  | L386: delete one instance of “that it” | Deleted |
| 48 |  | L389: simply estimating 10 plausible scenarios would seem to introduce subjectivity because to do the assessment, wouldn’t just one value need to be chosen? If not, then are all 10 values used in 10 different assessments? | Yes, I agree, I have slightly reworded this sentence |
| 49 |  | L394: “less frequently than” (not less frequently that) | Changed |
| 50 |  | L401: need a space between typus and are | Added |
| 51 |  | L412: odd spacing in population densities | Changed |
| 52 |  | L413: add “in” before “many cases” | Added |
| 53 |  | Various places: et al should be et al. (with period after al) | Added period |
| 54 | Reviewer 3 | I suggest changing the names of the four models to be more descriptive of what is being estimated, fixed, and what's the data source, something along the lines of 3PLF-maternity, 2PLF-maternity, 2PLF-maternity-noPmax (or 2PLF-maternity-annual), and 2PFL-maturity. I also think adding a clearer description of the four models used (as outlined above) would be very beneficial for readers who are unfamiliar with these maturity/maternity growth models. |  |
| 55 |  | Line 19. Suggest using a comma after "100". | Added |
| 56 |  | Line 50. In the second equation line, I think `x` should be `x\_i`. | Correct, changed |
| 57 |  | Line 140. Please provide version numbers for R and TMB. | Added |
| 58 |  | Line 147. The concept of r- and K-selection is considered "r-K-aic" by many as it oversimplifies axes of variation in life histories, particularly differences in reproductive output. It would be more accurate to say that "... were chosen to be representative of fast and slow chondrichthyan life histories, respectively", as both species differ mostly in their time-related LH traits. | Changed |
| 59 |  | Line 183. Clarify that parameters Psi(t) and Psi'(t) are dichotomous data on female maturity and maternal condition, respectively (as described on line 63). | Changed |
| 60 |  | Line 191. Add hyphen to "2PLF maturity". | Added |
| 61 |  | Line 233. Add space between 0.12 and yr-1. | Added |
| 62 |  | Line 401. Add space after comma. | Added |
| 63 |  | Figure 7. Please describe the black dots in the caption. | Added, and have also described what the marginal rug plots are as this was also left out. |

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 1 Example of simulated catch at length data for Galeorhinus galeus comparing two different values for AMAX.54 is the maximum empirical age based on tag-recapture, and 27 is half that value. Facets represent different sample sizes of 100, 1000 and 2500. Code available upon request.