EASTERN ATLANTIC BLUEFIN TUNA ASSESSMENT REVIEW

James Ianelli[1]

SUMMARY

The assessment models applied to various levels of aggregation for the eastern bluefin tuna stock (EBFT) included VPA, ASAP, and stock synthesis (SS). For VPA and ASAP, I found these results useful but limited in that they fail to use all available information. These two models incorporated data in aggregate and consequently, failed to provide a complete picture on how the different fisheries interact with the stock. For SS, the diagnostics have been sufficiently applied and documented. However, the development requires further work in several areas to be fully acceptable. The SS estimates seem most reliable in terms of uncertainty estimations. However, the uncertainty in the data (e.g., catch totals) suggests bias in the point estimates and uncertainty may be underestimated. We conclude that the results from the different models are indicative that the abundance levels have improved from low values and is likely to continue improving given recent patterns of fishing mortality (effort). However, the projection and ability to advice on levels of change in TAC is limited and should be deferred until the MSE is completed.

RÉSUMÉ

Les modèles d'évaluation appliqués à divers niveaux d'agrégation du stock de thon rouge de l'Est (EBFT) étaient VPA, ASAP et Stock Synthesis (SS). Pour VPA et ASAP, les résultats semblaient utiles mais limités dans la mesure où ils n'utilisent pas toutes les informations disponibles. Ces deux modèles ont incorporé des données agrégées et par conséquent, n'ont pas réussi à fournir une image complète de la façon dont les différentes pêcheries interagissent avec le stock. Dans le cas de SS, les diagnostics ont été suffisamment appliqués et documentés. Cependant, son développement nécessite un travail supplémentaire dans plusieurs domaines pour être pleinement acceptable. Les estimations de SS semblent les plus fiables en termes d'estimation de l'incertitude. Cependant, l'incertitude des données (par exemple, les totaux de capture) suggère un biais dans les estimations ponctuelles et l'incertitude peut être sous-estimée. Nous concluons que les résultats des différents modèles indiquent que les niveaux d'abondance se sont améliorés à partir de valeurs faibles et qu'ils continueront probablement à s'améliorer compte tenu des tendances récentes de la mortalité par pêche (effort). Cependant, la projection et la capacité de donner un avis sur les niveaux de changement du TAC sont limitées et devraient être reportées jusqu'à l'achèvement de la MSE.

RESUMEN

Los modelos de evaluación aplicados a varios niveles de agregación para el stock de atún rojo del este (EBFT) incluyeron VPA, ASAP y Stock Synthesis (SS). Para VPA y ASAP, los resultados parecieron útiles, pero limitados en el sentido de que no utilizan toda la información disponible. Estos dos modelos incorporaban datos agregados y, en consecuencia, no ofrecían una imagen completa de cómo interactuaban las distintas pesquerías con el stock. En el caso de la SS, los diagnósticos se han aplicado y documentado suficientemente. Sin embargo, su desarrollo requiere trabajos adicionales en varias áreas para ser plenamente aceptable. Las estimaciones SS parecen las más fiables en términos de estimación de la incertidumbre. Sin embargo, la incertidumbre de los datos (por ejemplo, los totales de capturas) sugiere un sesgo en las estimaciones de valores y la incertidumbre puede estar subestimada. Se concluye que los resultados de los distintos modelos son indicativos de que los niveles de abundancia han mejorado a partir de valores bajos y es probable que sigan mejorando dados los recientes patrones de mortalidad por pesca (esfuerzo). Sin embargo, la proyección y la capacidad de asesoramiento sobre los niveles de cambio en el TAC son limitadas y deben aplazarse hasta que se complete la MSE.

KEYWORDS

Stock assessment

[1]Jim.Ianelli@gmail.com, 10628 Marine View Drive SW, Seattle WA 98146

Executive Summary

The assessment models applied to various levels of aggregation for the eastern bluefin tuna stock (EBFT) included VPA, ASAP, and stock synthesis (SS). For VPA and ASAP, I found these results to be informative but limited in the ability to reasonably take all the information available into account. These two models incorporated data in aggregate and consequently, failed to provide a complete picture on how the different fisheries interact with the stock. This information is needed for management advice. Nonetheless, these models show that the stock status seems to have improved. For SS, the diagnostics have been sufficiently applied and documented. However, the development requires further work in several areas to be fully acceptable. In terms of consistency among the models, results were broadly similar; as such, projections depended mainly on the assumed approach for future recruitment and assumptions about selectivity relating to fishing mortality. Such projections can only be indicative as it seems implausible that predictions on where different age groups will be most vulnerable is highly variable, as observed from historical patterns in the length frequencies and general distributions. Uncertainty is encompassed to some degree by examining the three different modeling approaches. The SS estimates may be the most reliable in terms of uncertainty. However, given uncertainty in data (e.g., catch totals for a significant period of the assessment), there is likely bias point estimates and greater uncertainty than that captured by all the methods. The communication of the results was most clear from the SS application. However, due to the virtual setting of the meeting and the form of interactions within the workgroup, the presentations and in-meeting updated results made following things difficult. We considered that including the extensive work performed in developing an operating model for the MSE work should be a formal part of the "assessment" process since the data and the conditioning should reflect likely distribution and stock dynamics. We conclude that the results from the different models are indicative and sufficient to provide general management advice that the abundance levels have improved from low values and is likely to continue improving given recent patterns of fishing mortality (effort). However, we recommend that science- and risk-based advice on the magnitude of how TACs should change in future years should be done following the results of simulation-tested and data driven management procedures—an expected product of the MSE work that is underway.

1. Introduction

This report is based on the cumulative work from the data-workshop meeting held in April and the follow-up on the assessment during the first week of July 2022. We address the requested (from the Terms of Reference) broad evaluations on:

1. Whether model diagnostic performance criteria have been sufficiently applied, sufficiently documented and sufficiently met to provide a basis for models to be used to provide management advice;

2. The methods used to estimate population benchmarks and stock status (e.g., target fishing mortality proxy for, e.g., F0.1);

3. The adequacy, appropriateness and application of the methods used to conduct projections, given the Commission's objectives;

4. the adequacy, appropriateness and application of the methods used to characterize the uncertainty and to provide probabilistic catch limit advice. Comment on whether the implications of uncertainty in technical conclusions are clearly stated.

Additionally, I was requested to provide:

5. comments on whether the stock assessment results have been presented clearly and accurately in the detailed report of the stock assessment.

Since these items basically hinge upon determining the "best available science" we considered it worthwhile to briefly review how this should be considered. The general guidelines on this topic focused on the main objectives of the science.[2] Among the attributes outlined in the paper cited, we considered the following to be the most relevant to this work:

a) having standardized methods for collecting data,
b) applying sound logic and statistical rigor for interpreting results
c) having clear documentation of methods applied (including results and conclusions), and
d) supporting adequate peer review.

For example, item a) on data collection: the processes appear to be improving. However, given inconsistencies (and apparent lack of knowledge) about some important data collection activities, this is probably the item that is most critically lacking. This shortcoming feeds into the issues that remain for item b) in that the lack of understanding data collection characteristics impacts the ability to apply appropriate statistical procedures and on how to interpret results. For item c), the documentation was reasonably provided but is conditioned on how well the available data collection process is understood and interpreted. As far as "adequate peer review", I suppose this question is partly satisfied by having a single external reviewer involved. The fact that member scientists should also be critically reviewing the activities within the SCS process (and working groups) I think the peer review is adequate, though hampered in this period due to constraints of having virtual and hybrid meetings during the pandemic.

For the models included in the EBFT assessment process, I note that for the ASAP model, the engagement was particularly difficult given the remote/virtual communication and the complexity of the data and the time available for configuring the assessment. For example, the assumptions about the age-specific selectivities (/catchabilities) for the index data could have been refined. Also, they failed to implement a multi-fleet version due to data issues. The interactions among member scientists in configuring the VPA and SS models was better since key assessment leads attended the meeting in person. However, even for these cases, more time was needed to develop these approaches more fully, especially for the SS application since it required fewer data assumptions compared to the VPA.

This report follows on from SCRS/2022/132 where I discussed aspects of the available data relative to assessment approaches. While outside the specific terms of reference requested for this report, I include an additional examination of the indices used for the assessment since I wish to highlight how data may affect the assessment results. I believe these evaluations can help with understanding how best to specify model configurations. In the following subsection I present a short evaluation of the indices and then I present a summery related to the Terms of Reference in the subsequent sections.

1.1 Issues with indices

In SCRS/2022/132 in the discussion of fishery independent indices we noted that the larval index showed a major increase in recent years and is responsible for some, rather incredible, increased stock sizes in projections. We noted that there was a work group recommendation to inflate the recent years' CVs as they are relatively small (and among the smallest in the series). Furthermore, using a larval sampling program (which by definition will have a fair amount of sampling variability) as an index of spawning biomass (akin to egg-production methods) seems to be highly unusual and unique. Egg production methods (Ref) typically perform poorly in indexing spawning stock abundance. This is for a variety of reasons Fisheries stock assessments require data that are reliably collected and compiled. Secondarily, assessment models should be configured to match the assumptions associated with the observed data. To gain some appreciation of the indices and the specified observation errors, we adopted a simple state-space random walk model that has a number of options for how process-errors can be specified and estimated. The observation model applies the observation-error variances ($\sigma^2_{j,t}$) for the j^{th} index in year t $\left(x_{j,t}\right)$. The indices are fit to latent state variables, e.g., the underlying population trend $ln\left(\hat{Z}_{j,t}\right)$ as follows:

$$ln\left(Z_{j,t}\right) = ln\left(\hat{Z}_{j,t}\right) + \epsilon_{j,t}, \text{where } \epsilon_{j,t} \sim N\left(0, \sigma^2_{j,t}\right)$$

and the state equation and associated process error variance σ^2_{PE} is defined as

$$ln\left(\hat{Z}_{j,t+1}\right) = ln\left(\hat{Z}_{j,t}\right) + \eta_{j,t}, \text{where } \eta_{j,t} \sim N\left(0, \tau^2_j\right).$$

[2] Defining and Implementing Best Available Science for Fisheries and Environmental Science, Policy, and Management

The process error variances τ_j^2 (which may or may not vary across indices) are fixed effect parameters and the unobserved population $ln(Z_{j,t})$ is estimated as a series of random effects. The model is fit using maximum likelihood estimation in TMB using the R package "rema" (Sullivan 2022). The index data were extracted from the SS base 8 files and all normalized to have mean value of 1.0. The original CVs were applied for observation error specifications and values for τ_j^2 were estimated over all the indices in one case, and a separate run was made where all τ_j^2 values were equal (and a single value for this process error was estimated). The full period was used (the earliest data point was in 1952) but the focus of this illustration was for the recent period.

1.1.1 Results from evaluation of indices

For the case where each index is treated independently (i.e., they all have their own process error term and state variables), the indices with a flat trend (i.e., S2_MOR_POR_TP[3]), the process error term is small as indicated by the narrow confidence band (**Figure 1**). For the indices which vary substantially the process errors are large as indicated by the large confidence intervals in hindcast and projection periods (i.e., in two out of three aerial survey series and in the larval survey). In terms of what is feasible for the state-variables over time, neither of these configurations seems plausible. As such, the model was configured to have a common process error variance among all the indices. This is meant to imply that the unobserved state variability should be similar for the same species (though arguments could be made on indices that represent young fish should be more variable than those that represent old fish). Among the 11 indices, the S1, S5, S6 series showed little difference in the estimates between the two model configurations. This means that they represent data that are consistent with the "average" amount of state variability (for this given set of indices). For the other indices, when estimating the PE variances individually some were apparently overly constant (e.g., S2, S3, and S9) which simply reflects that the data are either limited (e.g., S9) or relatively invariant over time (S2).

While this is a simple-minded way to evaluate the information content of the indices (and their specified observation/measurement errors) and should be interpreted with caution, it does raise an important point that relates to how such data could or should be used within an assessment model. I found it strange that indices on older fish (specifically the larval survey and to a lesser extent, the aerial indices) had the highest process-error variances. This contrasts with other relatively large fish fisheries (i.e., S1 and S2) which had moderate to low variability. For the latter, the S2 index was completely flat compared to the other indices in the recent period. In most fish stocks, the variability of younger fish tends to be driven variable recruitment and that year-to-year variability of the "spawning stock" is less variable, especially for relatively long-lived fish such as BFT.

1.1.2 Use of larval survey as an index of spawning biomass

As noted, the rapid changes and recent increases in the estimated larval survey seems contrary to other indices of similar sized fish. That is, it seems very unlikely that the adult population could increase to the extent observed. The simple-minded analysis showed that the process error assumption would differ substantially from the other indices should the index be reliable. All of the models fit these data poorly with a serious pattern of negative results for the first several years. This is because the increase suggested by the Larval survey is inconsistent with most all other information on incoming recruitment (that would have contributed to the spawning biomass which this index is intended to apply). Using this index in any assessment as a proxy for adult spawning biomass. This is basically applying these data as if there were reliable egg-production sampling that has been done in other parts of the world for pelagic and other species. The track record for these data being useful within stock assessment settings is poor (e.g., see Armstrong *et al.*, 2001).

2. Considerations relative to each element in the terms of reference

2.1 Adequacy of alternative models

Under this request we provide feedback on if performance criteria have been sufficiently applied, documented and if they are useful to provide management advice.

The three models put forward were identified to have deficiencies in sufficiently integrating diverse and highly variable data characteristics (see SCRS/2022/132). At one end of the spectrum, the model configured using the stock synthesis applied size composition data and many fleets (16) and indices (11). While the effort was admirable

[3] See Anon 2022 Table 6 on page 20 for descriptions of these names.

in dealing with the complexity, it was clear that the way selectivity was specified caused estimation difficulty and rigidity in the separable (age-specific) fishing mortality assumptions. Within each gear type the available (/selected) sizes (ages) was mostly constant yet the size composition data clearly showed that the sizes available to the different fisheries varied broadly. At the other end of the spectrum, the VPA model aggregated all the catch-at-age over fleets *and* allowed fishing mortality to vary broadly by age over time. A somewhat intermediate model (ASAP) attempted to disaggregate the fishery data to the main fleets but during the meeting, the paper submitted had an issue with how the catch data were compiled so they abandoned runs with disaggregated fleets and went back to the aggregated version as presented in April. This model also had relatively rigid assumptions about the partial Fs at age over time. They used different time "blocks" to allow for some variability.

2.2 Methods used to estimate stock status

For ASAP and VPA models, the issue of aggregating over fleets will affect assumptions about the uniqueness of computing a target fishing mortality rate such as F0.1. For ASAP, the partial F's (or selectivity) for the recent period appears to be mainly fish aged 10 and older. However, in earlier periods the pattern was for substantially younger fish. The extent that the F0.1 can be based on an age-selection pattern some 4 years earlier raises concerns over reliability since the selectivity clearly varies considerably. This problem becomes exacerbated with the VPA method where the selection pattern varies tremendously. Within the SS application, this problem also exists, but it should be possible to tease out the relative contributions to the F0.1 estimation based on the relative future catches anticipated by each fleet.

Regarding stock status estimates, this involves reliable time-series estimates of reproductive output compared to what is estimated to exist under no fishing. Interestingly, while they were not overplotted to make an easy comparison, it does appear that the VPA and SS models had near identical spawning biomass levels in 2019 but the VPA indicated a more rapid increase.

2.3 Projection methods

Methods presented to conduct projections from each model were overly complex and difficult to understand and follow. There should be a simple table as a result from the different models showing the year and starting input numbers-at-age, the age-specific Fs, and how recruitment is generated. Then each estimate and projection method could be easily compared. Instead, there was opaque configurations within each model approach and I failed to see a way to make any comparisons among methods.

As for specifics relative to the Commission's objectives, an assessment model is inappropriate to accommodate without details. For example, do the Commission's objectives include catch stability? Effort control? Equity in opportunities among members? For this type of application, the MSE process should be used to compare projections and management procedures more directly against the Commission's objectives. To this end, and as noted above, using the operating model from the MSE as a supplemental assessment tool seems most appropriate for evaluating projections (and the relative uncertainty).

2.4 Uncertainty estimates

As noted in SCRS/2022/132, I obtained the important data files used in different assessment software packages. This provided the ability to understand the extent and issues related to how scattered the data appear through time by fishery. I attempted to evaluate the SS run using the MCMC package *adnuts* (Monnahan 2021). With relatively few parameters, the mixing rates were poor due to a couple of parameters that had divergences. The run time exceeded 50 hrs due to this aspect. I conclude that with more time to evaluate model specifications, this model could have benefitted with the use of more advanced methods to obtain more reliable model uncertainty estimates.

2.5 Assessment presentation and documentation

Assessments were presented clearly and accurately given the time constraints. The meeting process failed to compare results in a timely way so model comparisons could be discussed in more detail (a table was constructed and included in the meeting report).

2.6 Conclusions

The data organization is necessarily complex given the number of fisheries and highly interested members of ICCAT. This poses serious challenges for any assessment approach. Examining the index data in a simple way shows that there are inconsistencies between the specified observation errors and what should be a common process error (for EBFT).

At this meeting, the three draft forms of the models presented were far from being appropriate as an "acceptable" assessment in the conventional sense. Nonetheless, each approach attempts to integrate the available data and on their own can provide insights on stock trends, issues with assumptions, and what type of data might best help in the future. The issues I noted from the April and July meetings were:

- SS3:
    - Size-selectivity may interact with growth in unclear ways
    - The need to handle variable selectivity over time (burdens computation time)
    - High fishing mortality issue and the "hybrid F" option (may need to estimate Fs directly)
    - Constant, fixed length-weight relationship
    - Having so many fleets adds to the complexity of assumptions and on the computation of reference points
    - The model may have difficulty anchoring the scale of the population (e.g., based on examinations of flat likelihoods when profiled over the R_0 parameter
    - The conflicts in the data require more study in how the indices are interacting and whether the implied assumptions are valid (that the stock availability to the indices has been constant with respect to size/age).
- ASAP:
    - Single fleet (apparently to be revised in subsequent developments; but then same issues will arise a in the SS3 case)
    - "selectivity blocks" may require added evaluations/justifications (including adding variability); lacks flexibility in selectivity-at-age. Separable assumptions may be too severe
    - Unclear how body mass-at-age is derived/used and the extent it varies over time
    - The index fleet selectivity assumptions were fixed based on estimates from a separate model. While this may be a good first order approximation, it will affect the uncertainty arising from these data since the age-specific catchabilities are rigidly linked.
    - Unclear how catch-at-age data can be shown as consistent by areas and times/seasons
    - Results of spawning biomass appear to be about 2 times higher than those from the other two models in 2019.
- VPA:
    - Single fleet
    - Cohort sliced data
    - Ad-hoc methods
    - Lack of statistics, data are "known" variables
    - Known to have issues in terminal years

Each of these approaches as drafted can handle (to some degree) the uncertainty in the catch during the 1996-2006 period (historically dubbed as inflated catch) since this has been identified as an issue of importance. I think that the VPA approach, which lacks any formal statistical basis, is hard to defend for providing useful advice to the commission. The authors are striving to at least evaluate its application in traditional and rigorous ways, but at the end, the trade-off in the number of assumptions required may be less useful.

As noted earlier, it may be useful to allow some efficient non-parametric smoothing (see SPRFMO 2021 or Butterworth *et al.,* 2003). These models are efficient (can be fitted to a broad variety of data without much computational overhead) and can provide full Bayesian integration based on recent experience with the adnuts R package (Monnahan and Kristensen 2018, Monnahan 2018). Such models may help with developing simple models for management procedure testing (in the future).

In general, I believe the science standards for the EBFT assessment could be improved. This assessment and process was partly affected by the circumstances brought on by the COVID19 pandemic but mostly due to the complexity of the available data and the commitment required to undertake a comprehensive stock assessment. Resources that might normally have been devoted to this assessment have focused on developing management

procedures through the MSE process. I support this direction as I believe developing a well-tested management procedure is the best way for ICCAT to proceed in managing this fishery. However, I think that the "assessment" could have benefitted by having some focus on the - presumably data-driven and tuned - operating model (OM) used for the MSE. That is, the OM should have the features that are based on available data modeled in a way that can provide both the estimates of stock status (and uncertainty) and methods for projecting future conditions (in addition to "future" data). Including the OM in the mix, at a minimum, should provide an alternative indication of how the stock is responding to available data and how future catch scenarios might impact the status. Finally, the recent paper that shows how the suitable habitat for Atlantic BFT has changed (Payne *et al.*, 2022) should be brought forth and discuss in terms of impacts on different indices and future availabilities to different fisheries.

References

Anon. 2022. Report of the 2022 ICCAT Eastern Atlantic and Mediterranean bluefin tuna stock assessment meeting. (Madrid, Spain, hybrid meeting, 4-9 July 2022).

Armstrong, M. J., Connolly, P., Nash, R. D. M., Pawson, M. G., Alesworth, E.,Coulahan, P. J., Dickey-Collas, M., Milligan, S. P., O'Neill, M. F., Witthames, P. R., and Woolner, L. 2001. An application of the annual egg production method to estimate the spawning biomass of cod (*Gadus morhua* L.), plaice (*Pleuronectes platessa* L.) and sole (*Solea solea* L.) in the Irish Sea. – ICES Journal of Marine Science, 58:183–203.

Butterworth, D.S, J. N. Ianelli & R. Hilborn (2003) A Statistical Model for Stock Assessment of Southern Bluefin Tuna with Temporal Changes in Selectivity, African Journal of Marine Science, 25:1, 331-361, DOI: 10.2989/18142320309504021.

Ortiz, M., and Palma, C. 2011. Summary of comparison and verification of the AgeIT program for age-slicing of bluefin tuna catch-at-size (CAS) information. ICCAT Collective Volume of Scientific Papers, 66: 918–934.

Monnahan CC, Kristensen K (2018). "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages." PLoS ONE, 13(5), e0197954.

Monnahan CC (2018). adnuts: No-U-Turn MCMC Sampling for ADMB Models. R package version 1.1.2.

Payne, M.R., Danabasoglu, G., Keenlyside, N., D. Matei, A.K. Miesner, S. Yang and S.G. Yeager. 2022. Skillful decadal-scale prediction of fish habitat and distribution shifts. Nat Commun 13, 2660. https://doi.org/10.1038/s41467-022-30280-0.

SPRFMO (2021). 9th Scientific Committee meeting report. 79 p. Wellington, New Zealand 2021. See Technical Annex.

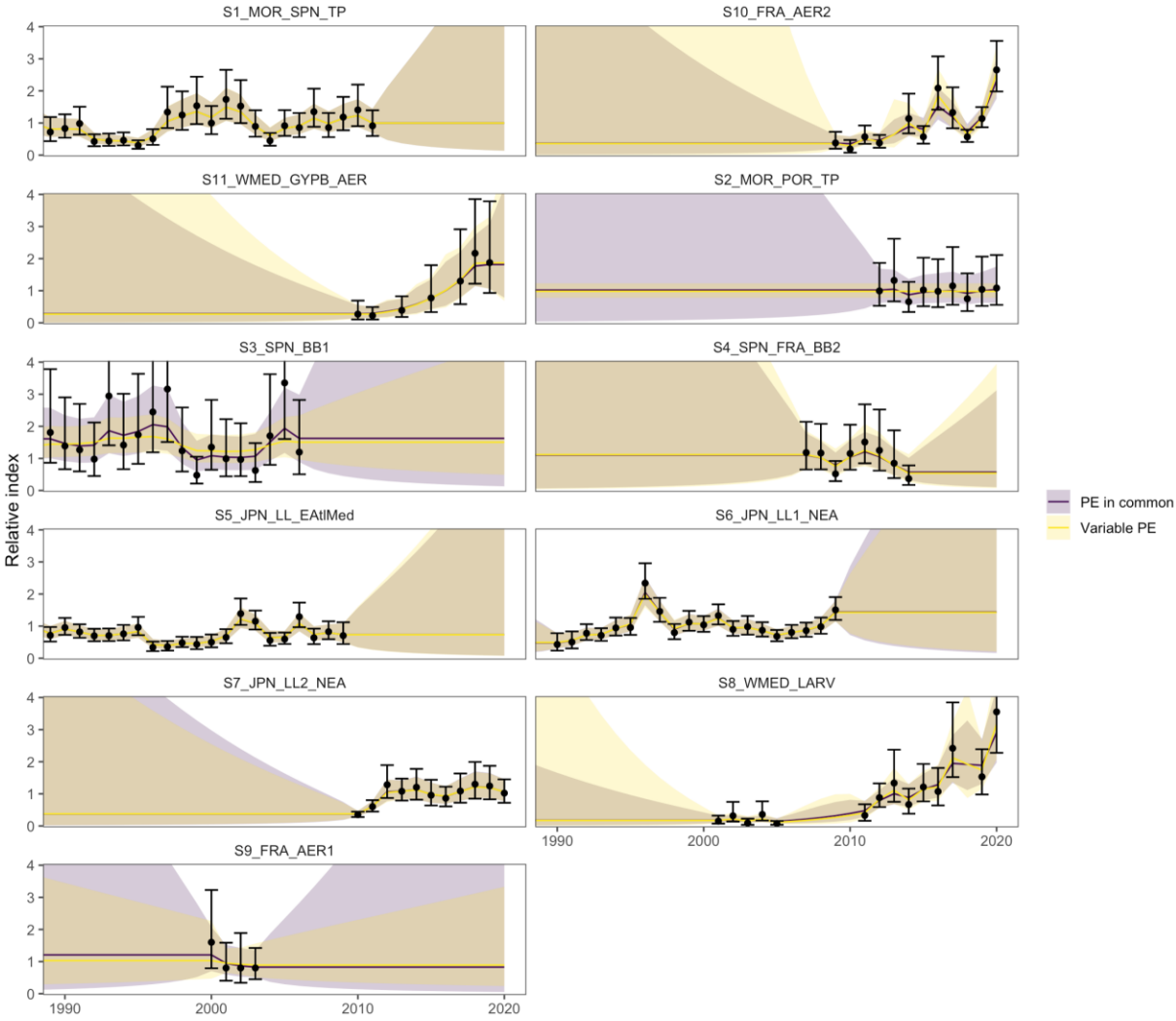

Figure 1. Random-walk model fits to each of the 11 indices (solid bullets with error bars) used in tuning the "base 8" SS model rescaled to each have a mean value of 1.0. The shadings indicate uncertainty bounds based on process error (PE) estimates. The two cases are where the process errors have a common variance (purple shading) and the other is where a separate process error variance term is estimated for each index (gold shading).