

Investigating trends in process error as a diagnostic for integrated fisheries' stock assessments

Gorka Merino[a,*], Agurtzane Urtizberea[a], Dan Fu[b], Henning Winker[c], Massimiliano Cardinale[d], Matthew V. Lauretta[e], Hilario Murua[f], Toshihide Kitakado[g], Haritz Arrizabalaga[a], Robert Scott[h], Graham Pilling[h], Carolina Minte-Vera[i], Haikun Xu[i], Ane Laborda[a], Maite Erauskin-Extraminiana[a], Josu Santiago[a]

[a] AZTI Fundazioa, Herrera kaia pasealekua z/g, 20100, Pasaia, Gipuzkoa, Spain.

[b] Indian Ocean Tuna Commission, PO Box 1011, Victoria, Seychelles.

[c] Joint Research Centre (JRC), European Commission, Ispra, Varese, Italy.

[d] Department of Aquatic Resources, Institute of Marine Research, Swedish University of Agricultural Sciences, Lysekil, Sweden.

[e] National Oceanic and Atmospheric Administration, Southeast Fisheries Science Center, 75 Virginia Beach Drive, Miami, Florida, 33149 USA.

[f] International Seafood Sustainability Foundation, Washington, DC 20005, USA.

[g] Tokyo University of Marine Science and Technology 5-7, Konan 4, Minato-ku, Tokyo 108-8477, Japan.

[h] Oceanic Fisheries Programme, The Pacific Community (SPC), B.P. D5, 98848 Noumea Cedex, New Caledonia.

[i] Inter-American Tropical Tuna Commission, 8901 La Jolla Shores Drive, La Jolla, CA, 92037-1508, USA.

(*) Corresponding author: gmerino@azti.es.

Abstract

Integrated stock assessments consist of fitting several sources of catch, abundance and auxiliary biological information to estimate parameters of equations that describe the population dynamics of fish stocks. Stock assessments are subject to uncertainty, and it is a common practice to characterize uncertainty using alternative hypotheses and assumptions within an ensemble of models to develop scientific advice for fisheries management. In this context, there is the need to assign levels of plausibility to each of the combinations of factors that ultimately reflect the uncertainty on different biological and fishery processes. In this study, we describe and apply a model diagnostic to identify trends in process error in recruitment deviation estimates within ensembles of integrated assessment models of tropical tunas. We demonstrate that assessment model ensembles for tropical tunas contain distinct scenarios with significant trends in process error that are overlooked, with the associated implications for fisheries management. Using the Indian Ocean yellowfin as a case study, we found that trends in recruitment deviates are linked to extreme productivity scenarios which strongly diverged in scale from deterministic models fitted without recruitment deviates. This indicates that when recruitment deviates show an increasing trend, these can compensate for the loss of biomass in periods of high catch beyond the surplus production. In these cases, variation in recruitment is not a random process, but rather takes the function of a compensatory, systematic driver in productivity. Significant trends in recruitment were positively correlated with increased standard deviations and auto-correlation coefficient, non-random residual pattern in fits to abundance indices, and particularly poor performance of the Age-Structured Production Model (ASPM) diagnostic. We suggest that trends in recruitment deviates can be caused by misspecification of the biological parameters used as fixed values in integrated assessment models. The process error diagnostic described here can provide a statistical criterion in support for hypotheses and assumptions when using ensembles of models to develop fisheries management advice.

Keywords

Stock assessment, process error, recruitment, uncertainty, fisheries management

1. Introduction

In ecology, resilience is defined as the ability of an ecosystem or species to resist and recover from a disturbance and return to equilibrium (Holling 1973; O'Leary et al., 2017; Pimm 1984). In fishery science, the productivity of fisheries reflects the capacity of fish stocks to respond to fishing pressure and overfishing thresholds are determined by fish life-history traits (Froese et al., 2021; Murua et al., 2017; Wang et al., 2020; Zhou et al., 2012) and fishing selectivity (Froese et al., 2016; Sampson and Scott, 2011). The management of fisheries is generally guided by the output of fisheries stock assessments, which estimates the stock's current and historical exploitation levels and maximum productivity and, predicts the levels of catch and fishing mortality that can be sustained by fish stocks. Integrated fishery stock assessment consists of fitting catch, abundance and auxiliary biological information into fish population dynamics equations using specifically tailored models and computer software. The biological information used in stock assessments include growth, reproduction and natural mortality that constrain the estimated productivity and thus resilience of fish stocks. In general, the knowledge of the underlying biological processes and life-history traits (*e.g.,* fecundity, longevity, maturation and somatic growth) is limited (Meador and Brown 2015) and the forms and values of these processes must be assumed. In particular, highly influential, yet difficult to estimate parameters such as natural mortality (M) and the steepness of the stock recruitment relationship (h) are commonly assumed and fixed in age-structured assessments, thereby making strong assumptions about stock's resilience, productivity and associated biological reference points (Mangel et al., 2013; Winker et al., 2020), with associated management implications.

Three major sources of error can cause structural and statistical uncertainty in fisheries stock assessment (Francis and Shotton 1997; Fromentin et al., 2014; Rosenberg and Restrepo 1994): (i) observation errors, directly linked to the measurement accuracy in the data, (ii) model errors, due to the limited ability of models to reproduce population dynamic patterns and, (iii) process errors, due to the inherent variability of the processes underlying fish stock dynamics or fisheries. Process errors usually refers to the excess of variation that cannot be represented by deterministic models; they are used in stochastic models to represent the variability in the data caused by natural population variation (e.g. recruitment strength, life history traits) or unaccounted variations (e.g changes in fisheries operations, time-varying catchability) beyond the deterministic expectation. Structural uncertainty relates to alternative assumptions about functional relationships (e.g. growth, selectivity and recruitment functions), fixed parameter values (e.g. M and h), data weighting, model structure (e.g. spatial and fleet structures). To characterize the structural uncertainty in fisheries, model ensembles are frequently considered for providing advice based on combining the outcomes of multiple model scenarios (Jardim et al. 2021).

Tunas sustain some of the world's most valuable fisheries and dominate global marine ecosystems (Juan-Jordá et al., 2011). Over the recent decades, tuna fisheries have intensified and expanded worldwide, and global catch has steadily grown (Figure 1) with the development of industrial purse seine fisheries, which has placed tuna fisheries management under pressure for timely and effective management (Allen et al., 2010; Merino et al., 2020). The major commercial tropical tuna species are bigeye, skipjack, and yellowfin tuna, which are among the most productive species of tunas, and their assessments are carried out using integrated age-structured fisheries stock assessment models such as Stock Synthesis (Methot Jr and Wetzel 2013) and Multifan-CL (Kleiber et al., 2012).

In the past, the stock assessments of tropical tunas used the best available information to fix key population parameters in a base case configuration. However, recent stock assessments tend to integrate results across alternative hypotheses of influential parameters to capture the full structural uncertainty in the estimates and in the management advice. In the assessments of tropical tuna stocks of 2019, 2020 and 2021, the structural uncertainty has been characterized using ensembles of models with factors such as the steepness of the stock-recruitment relationship, variability in recruitment, natural mortality, growth, longevity, fishing gear selectivity and weighting of different data sources (Merino et al., 2021). Different options for model assumptions are combined in a model ensemble and each models' result is averaged using statistical techniques (Walter et al., 2019; Winker and Walter 2019) to obtain probabilistic estimates of stock status and productivity to develop management advice.

The use of the ensemble or grid approach has raised discussions on the associated plausibility of each factorial combination of hypotheses, factors, and scenarios (Maunder et al., 2020). Recently, specific diagnostics have been compiled to evaluate the convergence, consistency, and prediction skill of integrated stock assessments and to help model development and selection (Carvalho et al., 2021). Specifically, these diagnostics evaluate (i) model convergence, (ii) goodness of fit to the data by analysing differences between estimated and observed quantities (residuals) (Wald and Wolfowitz 1940), (iii) model consistency by identifying the influence of the different sources of information in the likelihood component (Ichinokawa et al., 2014) and retrospective analyses (Brooks and Legault 2016) and, (iv) prediction skill by checking that predictions are consistent with future reality conducting hindcasting by adding steps of projection to retrospective fits (Kell et al., 2021). These diagnostics have been used to develop tuna stock assessments under the grid approach (Urtizberea et al., 2019) but it is computationally intensive and time consuming to run all diagnostics (in particular retrospectives and hindcasting) for all model configurations in a large ensemble. Therefore, it is common to evaluate diagnostics for a reference case or diagnostic case configuration to help model development (Fu et al., 2021), or in a subset of models (Minte-Vera et al., 2020; Xu et al., 2020) and is only seldom the case where diagnostics are used to select or weigh all models of the reference grid used to develop management advice (Castillo et al., 2021; Maunder et al., 2020).

Alternative model assumptions lead to various extents of model misspecifications, where model specification is the difference between the model and reality. It follows that all model are somewhat misspecfied, but some are more parsimonuos and useful for advice than others (Carvalho et al. 2021). Examining residuals pattern of the fitted observation is commonly considered one of the first step for identifying model misspecification. For example, poor model fits can be detected by either the magnitude of the residuals being larger than expected or trends in residuals. However, in stochastic models, model misspecification is likely to cause additional process error and systematic trend in the process deviations, which provides the model with additional flexibility to compensate for misspecification in system dynamics in the fits to the observations. As such, process error deviations may also serve as "sink" of unaccounted time-varying processes and latent model misspecifications. Diagnosing the statistical properties of the recruitment deviates appears to have been somewhat overlooked as a potentially critical aspect in the diagnostic toolbox model for integrated assessment models (Carvalho et al. 2021). However, a diagnostic approach that builds on a similar principle is the use of a deterministic age-structured production model (ASPM) for evaluations against a full stochastic model implementation with respect to scale and the production function (Maunder and Piner 2015; Minte-Vera et al. 2017). The ASPM approach has also shown promising performance in simulation-testing for detecting misspecification the population dynamics (Carvalho et al. 2017).

Fish populations have been shown to exhibit large variation in recruitment about the assumed relationships between spawning stock biomass (Mertz and Myers, 1996; Rose et al., 2001; Thorson et al., 2014). Integrated models are therefore commonly configured in a way that recruitment variation is main (or only) source of process variation. It is common to model recruitment as a random deviation from a stationary functional relationship between the spawners and subsequent recruitment (Sharma et al., 2019). Recruitment deviates are usually considered to originate from a random lognormal process with a mean zero constraint around a log-bias adjusted stock-recruitment curve (Methot Jr and Wetzel 2013). The assumption of a lognormal distribution has been supported by empirical evidence (Allen 1973), as well as biological realism (Hilborn and Walters 1992; Quinn and Deriso 1999). A theoretical justification for the use of this error distribution is that survival from spawning to recruitment can be considered as the combined effect of a series of independent environmental factors that affect mortality during early life stages (Walters and Ludwig 1981). This interpretation of the lognormal error as arising from a combination of multiple environmental effects implies that the recruitment can be occasionally very large when most environmental conditions are favourable, and that the variance of recruitment will increase as the expected stock size and recruitment increase (Hightower and Grossman 1985). The most common approach to estimate recruitment variations remains probably maximizing a penalized likelihood by fixing an assumed standard deviation in recruitment (but see Thorson 2019 for alternative methods), which penalizes the likelihood if the average the recruitment deviates exceed the assumed variation about the stock-recruitment relationship. A bias-adjustment approach is often implemented to ensure that the expected recruitment in each year is equal to the stock-recruit relationship (Cordue, 2001). The implicit model assumptions of this are therefore that recruitment variation is stationary and is less likely to exceed an upper process error threshold given by fixed marginal recruitment standard deviation (sigmaR), for which plausible values may also be informed from meta-analyses (Thorson et al. 2014; Thorson 2019).

This study specifically focuses on potential model process error diagnostics of recruitment deviation estimates in integrated assessment models. We explore the trends in recruitment deviates of alternative model configurations within ensembles of models and illustrate their patterns in response to different hypotheses based on life-history assumptions. We developed a diagnostic tool to objectively evaluate different model scenarios and provide statistical criteria for model selection by identifying the least plausible models from an ensemble. For this, we run numerical experiments using the most recent Stock Synthesis model of Indian Ocean yellowfin tuna (Fu et al., 2021). The analyses include (i) assessing the hypothesis of no-trend in recruitment deviates, (ii) comparing with equivalent scenarios without recruitment deviates, (iii) comparing the probability of no-trend hypothesis with diagnostics developed for integrated stock assessment models (Carvalho et al., 2021) and, (iv) simulating bias in natural mortality and growth parameters within a stock assessment carried out using simulated data. We then evaluate evidence of process error trends in the assessments of tropical tunas across ocean basins.

2. Material and methods

2.1 Data

The data used for our analyses includes files of the Indian Ocean yellowfin and other tropical tuna stock assessments. Yellowfin tuna supports the most valuable tuna fisheries in the Indian Ocean, with catches currently exceeding 400,000 t annually. The stock is harvested by a diverse range of gears, from small-scale artisanal fisheries to large gill netters, industrial longliners, and purse seiners, with the western tropical region being the core area of the fisheries' distribution. The stock is currently determined to be overfished and subject to a building plan (IOTC, 2021).

The yellowfin tuna stock is assessed using an age and spatially structured Stock Synthesis model that incorporates spatial recruitment and movement dynamics and accounts for the different regional exploitation pattern (Fu et al. 2021). The data available for assessing the stock include time series of the total catch, standardised CPUE indices, observations of length compositions, and tagging recaptures data. CPUE are the primary source of information on abundance and are based on a regionally stratified index for adult fish from the main distant water longline fleets, and a region-specific juvenile index from the European Union purse seine fleets. The length composition data are considered sufficient to provide reasonable estimates of fishery selectivity and recruitment trends but not stock abundance trends. Tag-release and recovery data collected from the main phase of the Indian Ocean large-scale tuna tagging programme inform estimates of mortality, abundance, and movement. The Indian Ocean yellowfin assessment has established a model ensemble of 96 models to capture a range of uncertainties arising from assumptions on biological parameters, data weighting, and model configurations: 1) three levels of steepness (0.7 (*h70*), 0.8 (*h80*) and 0.9 (*h90*)); 2) two growth curves (*Gbase* (Fonteneau 2008) and *GDortel (Dortel et al., 2014)*); 3) two natural mortality options (*Mbase* and *Mlow*), 4) two spatial configurations (*io* and *sp*), 5) two assumptions about the effect of piracy in longline catchability (*q0* and *q1*) and, 6) two weighing options for tagging data (low weight (*tagLambda01*) and full weight (*tagLambda1*)).

As for the Indian Ocean yellowfin, the assessments of the other tropical tunas are also carried out using integrated statistical assessment models (Methot Jr and Wetzel 2013; Kleiber et al., 2012). For all cases, an ensemble of models is used to develop scientific advice for management and characterize structural uncertainty. The files of these assessments have been compiled to estimate the trends of the recruitment deviates. Our analysis on the Indian Ocean yellowfin is shown throughout the main manuscript and we also provide an overview of trends in recruitment deviates from all tropical tuna stocks as Supplementary material.

2.2 Analysis of trends in recruitment deviates

Process error in Stock Synthesis is typically implemented as a multiplicative lognormal error component applied to the stock recruitment relationship (equation 1). Recruitment (R) is defined as the expected number of recruits from a Beverton-Holt spawner-recruitment curve multiplied with a bias-adjusted log-normally distributed random recruitment deviation.

$$R_t = \frac{4hR_0SSB_t}{SSB_0(1-h)SSB_t(5h-1)} e^{(\varepsilon_y - 0.5\,\sigma_R^2)} \quad ; \varepsilon_t \sim N(0, \sigma_R^2) \qquad \text{[equation 1]}$$

where R_t is the number of recruits at time t, SSB_t is the spawning stock biomass, h is the steepness of the spawner-recruitment and R_0 is the estimable parameter for the expected recruitment of the unfished stock biomass SSB_0. The process error term ε_t represents the recruitment variability after accounting for the stock recruit relationship given the marginal variance of recruitment deviations σ_R^2 (Johnson et al. 2016).

The recruitment deviates of the main data period (years of the assessment with abundance indices and/or size compositions that are assumed to be informative) have been extracted from Stock Synthesis files using *r4ss* (Taylor et al., 2021) a package that contains a collection of R functions (R_Core_Team 2021) for interacting with Stock Synthesis. The statistical analysis has consisted in validating the hypothesis that there is no temporal trend in recruitment deviates. For this, we applied the Student's t-test using the R package *funtimes* (Lyubchich et al., 2022). The *notrend_test* function includes a combination of time series trend tests to verify the null hypothesis of no trend, versus the alternative hypothesis of a linear trend (Student's test).

2.3 Comparison to deterministic model runs without recruitment deviates

The aim of these runs is to evaluate whether the population dynamics are driven by the underlying production function estimated by the model, or by trends in process error (i.e., recruitment deviates). We compare the models from the stock assessment ensemble of Indian Ocean yellowfin with and without recruitment deviates. The production function represents the changes of yield over the range SSB from 0 to SSB_0 and its maximum corresponds to the Maximum Sustainable Yield (MSY). A key scaling parameter for the biomass is the estimable parameter of the unfished recruitment R_0. The relative productivity of the stock with respect to MSY is governed by the spawner-recruitment function, somatic growth, fecundity, natural mortality and fishery selectivity and can therefore be to a large extent predetermined by the choices about functional relationships and fixing population parameters (Winker et al. 2020).

If the assumed production function is supported by the data, it can be hypothesised that the estimated maximum sustainable productivity (MSY) and scale (R_0) are similar between the model fits with and without recruitment deviates. The hypotheses of this analysis are comparable to the Age Structured Population Model (ASPM) diagnostic (Maunder and Piner, 2015; Minte-Vera et al. 2017; Carvalho et al., 2021), with the difference that all available data sources are used to fit the model, including abundance indices, tagging and size frequency data. To implement the deterministic models, we re-run all models within the ensemble for Indian Ocean yellowfin without recruitment deviates, by deactivating the recruitment deviates' option in the Stock Synthesis control file (i.e. fixing the recruitment deviates to zero). The difference between the full stochastic models and their deterministic implementations was done by computing the percentage differences for MSY and R_0.

2.4 Comparison of process error trends with standard model diagnostics

A flowchart for model development and selection has been used to evaluate model plausibility using diagnostic criteria for model convergence, goodness of fit, consistency and prediction skill (Carvalho et al., 2021). To conduct a comparative analysis with the trends process deviations, we applied selected key diagnostic tests to the model ensemble of Indian Ocean yellowfin (Table 1) that can be can relatively straight forward for automated large ensembles using the R packages *r4ss* and *ss3diags* (https://github.com/JABBAmodel/ss3diags). These included runs tests to evaluate the randomness in the fits to the CPUE indices as a goodness of fit criterion, the ASPM diagnostic to evaluate consistency between the CPUE trends and the production function with respect to productivity (MSY) and scale (R_0) (Maunder and Piner, 2015; Minte-Vera et al. 2017), retrospective bias (Mohn's ρ; Mohn's 1999; Hurtado-Ferro et al., 2015) and the Mean Absolute Scaled Error (MASE) as a measure of prediction skill using hindcast cross-validation of the observations form the four common joint CPUE indices (Kell al. 2021), following the procedures described in Carvalho et al. (2021). In addition, we also evaluated two additional process error measures in the form of the realized marginal standard deviation of recruitment deviates and the first order auto-regressive (AR1) autocorrelation coefficient of recruitment deviates (Johnson et al. 2016) at an annual time step interval across scenarios.

The p-values for the residual runs tests were computed for each of the four joint CPUE indices that are common in all models. The p-values were then combined into a single test statistic using Fisher's method (equation 2):

$$\chi^2_{2k} = -2\sum \log(p_i) \qquad \text{(equation 2)}$$

where p_i is the p-value for CPUE index *i* and *k* are the degree of freedom of the four p-values from joint CPUE indices.

Retrospective analysis and hindcast cross validations were based on sequentially removing five years with data, whereas the hindcast then used one-year ahead predictions to compute the MASE (equation 3). The MASE was computed as a combined across all four joint CPUE indices and four seasons, such that:

$$MASE = \frac{\frac{1}{h}\sum |\tilde{y}_{i,s,t} - y_{i,s,t}|}{\frac{1}{h}\sum |y_{i,s,t} - y_{i,s,t-1}|} \qquad \text{(equation 3)}$$

where $\tilde{y}_{i,s,y}$ is the one-year-ahead forecast of the expected value for the of the log(CPUE) observation of index *i*, in season *s*, and year *t*, $y_{i,s,t}$ is the corresponding observed value, $y_{i,s,t-1}$ is the log(CPUE) observation from the previous year and *h* denotes the number of hindcasting annual retrospective hindcasts steps for which forecasts $\tilde{y}_{i,s,y}$ were made to compare with the observations $y_{i,s,t}$ (c.f. Carvaho et al. 2021). The numerator therefore represents the mean absolute error (MAE) of a total of 80 prediction residuals (4 indices × 4 seasons × 5 hindcasts) and the corresponding denominator the MAE of 80 naïve prediction residuals.

[Insert Table 1]

2.5 Experiment with yellowfin operating model

The aim of this experiment is to reproduce trends in recruitment deviates by intentionally producing bias in natural mortality and growth parameters. For this, we use data generated from an operating model (OM) developed for Indian Ocean yellowfin (Dunn et al., 2020) and develop a grid of Stock Synthesis models defining a range of natural mortality and growth parameters relative to the true values from the OM. The hypothesis is that under the assumption of data with random error, the use of scenarios with biological parameters that deviate from the true value will produce recruitment deviate trends comparable to the trends observed in the stock assessment.

The spatially explicit OM of the tropical tuna population was implemented by the Indian Ocean Tuna Commission as a proof of concept for evaluating potential stock assessments performance (Dunn et al., 2020). The OM development focused on the yellowfin as a case study based on data availability and management priorities. The operating model was conditioned on a range of spatially explicit observations (usually at 5 x 5 latitude and longitude grid), including commercial catch, catch rates, length frequency, and tagging data using maximum likelihood estimation, and incorporated population processes such as recruitment, growth, maturity, spawning, movement, and fishing at relevant spatial and temporal scale in accordance with biological and fishery characteristics of the yellowfin tuna stock. In particular, the OM implements and estimates movement dynamics using preference functions based on spatially discrete environmental layers. Subsequently fine scale randomised observational data for size, catch per unit of effort (CPUE) and tag recoveries generated from the OM were reformatted and fitted by a Stock Synthesis model equivalent to the 2021 IOTC yellowfin stock assessment.

3. Results

3.1 Catch of tropical tunas

Tropical tuna fisheries developed after the 1950s, and in the early years, mainly consisted of longline fleets targeting bigeye and yellowfin tuna. In the 1980's, the purse seine fisheries rapidly developed and increased the catch of tropical tunas worldwide, reaching their maximum total catches between 1990 and 2010. The catch of Indian Ocean yellowfin is currently near its historic maximum levels, likewise the Atlantic and Western Pacific stocks (Figure 1). The four skipjack stocks are currently at their historical maximum levels of catch whilst the catch of the four bigeye stocks has decreased in the recent years.

[Insert Figure 1]

3.2 Analysis of trends in recruitment deviates

The recruitment deviates and trend analysis of the 96 models included in the reference grid of the 2021 assessment of Indian Ocean yellowfin tuna are shown in Figure 2 (p-values for the no-trend hypothesis in Table 2). Black dots and lines represent scenarios where the hypothesis of no trend is verified and no trend in process error is detected (p-value > 0.1) and pink and blue lines and dots represent scenarios where a trend in recruitment deviates is detected (p-value < 0.1). Pink dots and lines represent scenarios with an increasing trend in recruitment deviates and blue dots and lines represent scenarios with a decreasing trend.

[Insert Figure 2]

We detected trends in recruitment deviates in 41 of the 96 models (43%). From these, 5 show a decreasing trend (the average recruitment deviates of the first period are larger than in the second period) and 36 display a positive trend (larger recruitment deviates in the second part of the data series, where the catch of yellowfin is higher). 23 of the 24 models (96%) that use the low natural mortality option and *GDortel* growth combined display an increasing trend. 32 of the 48 scenarios with low natural mortality option show an increasing trend (67%). 27 of the 48 scenarios with the *GDortel* option also show an increasing trend (56%). The scenarios with a decreasing trend are all using the base growth and base mortality options combined with the tagging-data downweighed option. The models that obtain a p-value of more than 0.8 (9 models, 9.4%), include at least once all the values of the factors included in the uncertainty grid (two growth options, two natural mortalities, three steepness values, two spatial configurations, two assumptions on tagging data and two hypotheses on the impact of piracy).

Figure 3 shows the relation between the p-value of the no-trend hypothesis and the range of the productivity of the stock (MSY) as estimated by the assessment models. P-values lower than 0.1 correspond to values of MSY lower than 350,000 tons (pink, increasing trends in recruitment deviates) and larger than 400,000 tons (blue, decreasing trend in recruitment deviates). The scenarios with particularly high probability for the lack of trend in recruitment deviates (p-values>0.8) also correspond to the range of MSY between 350,000 and 400,000 tons. The scenarios in the lowest left side of the figure (20 out of 96 models, 21%) display a very low probability for the no-trend hypothesis p-value<0.01 and MSY values estimated at 310,000 tons or less (average MSY for these 20 models is 286,974 tons). All models with a p-value>0.1 estimate MSY values larger than 314,507 tons. The Indian Ocean yellowfin catch reached 323,688 tons for the first time in 1992 and has remained above thereafter (except for 1999, with 277,771 tons) (Figure 1). The average catch since 1992 has been 382,064 tons. In the last 20 years (2000-2020), the average catch of yellowfin tuna has been 401,999 tons (40% larger than the average MSY estimated by the 20 models with p<0.01). The highest estimated MSY value is 468,488 tons with a model that displays a p-value of 0.012. The second largest estimated MSY is 463,968 tons and its model displays a p-value of 0.199.

[Insert Figure 3]

3.3 Comparison to deterministic model runs without recruitment deviates

Figures 4, 5, 6 and 7 show the differences between the estimated quantities (MSY, R_0 and B/B_{MSY}) between the stock assessment scenarios (SA) and the equivalent runs with the recruitment deviates option deactivated, i.e. without process error (RecDev0). The models identified with the lowest p-values and lowest estimated MSY (Figure 3) are also the models that display the largest differences in the estimated MSY with their equivalent models without recruitment deviates (lower MSY in the stock assessment than without recruitment deviates), reaching a -30% difference or more for 7 models (7%), -20% or more for 17 models (18%) and -10% or more for 49 models (51%) (Figure 4). The models that estimate larger MSYs than their equivalents without process error are also associated with p-values<0.1 (blue points). Two models (2%) from the stock assessment grid estimate MSY 10% larger or more than their equivalents without recruitment deviates. The models with the highest p-value for the no-trend hypothesis show differences of less than 10% with their equivalent model runs without recruitment deviates.

[Insert Figure 4]

Figure 5 shows that the models identified with the lowest p-values and lowest and highest estimates of MSY are the models with largest differences on R_0 compared to their equivalents without process error. The inverse relation between p-value and differences between models with and without recruitment deviates is even more compelling for R_0 than for MSY. The models with the largest p-values obtain very similar estimates of R_0 with and without recruitment deviates. 11 models from the stock assessment reference grid (11%) estimate R_0 30% or more lower than their equivalents without process error, 18 estimate R_0 20% or lower (19% of models) and 40 models 10% or lower (42%). One model from the stock assessment grid estimates R_0 20% larger or more than its equivalents without deviates (1%) and 12 models (12.5%) estimate R_0 larger than 10% or more.

[Insert Figure 5]

Figures 6 and 7 show the differences in relative biomass (B/B_{MSY}) estimated with and without recruitment deviates. Figure 6 shows the two trajectories for each single scenarios of the reference grid. The scenarios with large p-values for the no-trend hypothesis show similar overall trends between the models with and without rec devs (e.g. *GDortel_Mbase_h80_IO_q1_TagLambda01* [8th column, 1st row]) and the models with very low p-value (e.g. *GDortel_Mlow_h70_Sp_q2_TagLambda01* [11th column, 7th row] and *Gbase_Mbase_h90_Sp_q1_TagLambda01K* [3rd column, 5th row]) show a marked difference between the two model trajectories. In general, the trajectories from the models without recruitment deviates (dashed lines) elapse above the trajectory from the stock assessment models with low p-values (blue and pink). Figure 7 shows that as with the differences between MSY and R_0, the models with lowest p-values display large differences between estimated relative biomass. The differences in relative biomass between models with and without recruitment deviates reach 30% or more for 12 models (12.5%), 20% or more for 22 models (23%) and 10% or more for 43 models (45%). The models with the highest p-values estimate relative biomass differences of less than 10%.

[Insert Figure 6]

[Insert Figure 7]

3.4 Comparison of process error trends with standard model diagnostics

Table 2 shows the results of the diagnostics used to evaluate plausibility of the different models and Figure 8 shows the values of the different diagnostics for models identified or not with a trend in recruitment deviates. Figure 9 shows the correlation of diagnostics with the probability of no-trend in recruitment deviates hypothesis. Overall, the models identified with trends in recruitment deviates are linked with autocorrelated deviates, with higher variance, with larger differences between MSY and R_0 estimates relative to their ASPM models' and, poorer scores in runs test of residuals of fit. Trends in recruitment deviates appear independent of MASE and retrospective performance (Mohn's ρ). Consequently, the p-value is negatively correlated with differences on the MSY (-0.67) and R_0 (-0.63) estimates between the stock assessment and the ASPM models. This means that the largest p-values of the no-trend hypothesis are linked with lower differences between stock assessment models and equivalents using catch and effort data only and without recruitment deviates. The p-value is also negatively correlated with the standard deviation (-0.39) and autocorrelation of recruitment deviates (-0.38). Additionally, the p-value is positively correlated to the runs test (0.21).

[Insert Figure 8]

[Insert Figure 9]

[Insert Table 2]

3.5 Experiment with yellowfin operating model

Figure 10 shows the estimated recruitment deviates for a model that uses data generated from a simulated OM (Dunn et al., 2020). This experiment suggests that natural mortality needs to be reduced or increased by 90% (M010 and M190) to produce a trend in recruitment deviates. As with the stock assessment, models with recruitment deviate trends are associated with lower (pink) and higher (blue) MSY than their equivalent without deviates (Figure 11). The differences in MSY between models with and without recruitment deviates ranges between -19% and +24%. The models with higher p-value for the no-trend hypothesis are also the models with the lowest differences between models with and without process error.

[Insert Figure 10]

[Insert Figure 11]

Discussion

Our results demonstrate that the assessments of tropical tunas contain trends in process error that are overlooked, and we highlight that not accounting for this uncertainty can have important implications for stock management. We show that evaluating trends in recruitment deviates from integrated assessment models can contribute to reducing the uncertainty in fisheries' stock assessment and to improve the assessment of stock status. Trends in recruitment deviates were correlated with extreme (lowest and highest) productivity scenarios and, with differences (up to 30% for Indian Ocean yellowfin stock assessment models) in the estimates of model runs with and without recruitment deviates. This indicates that when recruitment deviates show an increasing trend, these can compensate for the loss of biomass in periods of high catch beyond the surplus production. When this happens, the process error is not a random component that describes the variability in the population trends as driven by fish productivity and fishery dynamics but, it is identified to be one of the processes that drive the general population trend in the form of the underlying stocks' response to fishing pressure. Trends in recruitment deviates can be caused by misspecification of biological and other parameters and suggest incompatibility of model assumptions with the data.

The misspecification of key parameters or assumptions in integrated stock assessment models can strongly impact the scientific advice for fisheries management (Carvalho et al., 2021; Mangel et al., 2013). When using integrated models, numerous decisions are required such as whether the models appropriately fit the data, if the optimization has been successful, if estimates are consistent retrospectively and if the model is suitable to predict a stock's future response to fishing (Carvalho et al., 2021). During the development of integrated models, analysts evaluate performance from likelihood profiles, the residuals between estimated and observed quantities, retrospective analyses, and other methods. This process allows for deciding between modelling options, parameters and selecting or discarding specific model assumptions. However, this evaluation of diagnostics is time-consuming, especially when large ensembles of models are the preferred option to characterize structural uncertainty. In cases where the factors of uncertainty, assumptions and the configuration of models are decided during time-limited stock assessment meetings, developing a full set of diagnostics becomes inviable. Producing near-term advice when time pressure is severe and uncertainty looms can lead to decisions guided by priors without statistical support (Schuch and Richter 2021). Evaluating trends in recruitment deviates from stock assessment output files is a relatively straightforward and quick task that can help identify model assumptions that are incompatible with the available observations and thus providing an additional statistical method for model selection in a timely manner.

Figures 2 and 3 demonstrate how positive recruitment deviates are accumulated in the recent period of the assessment models that estimate productivity levels significantly lower than the recent catch history. This suggests that recruitment deviates are an intrinsic factor that is part of the response to the high catch in the recent years and that the recent catch history would not be possible without them. It would be expected that fish stocks with a maximum productivity of 40% below the average catch of the last 30 years would have collapsed but instead, the positive trend in recruitment deviates prevents it. However, when running deterministic projections forward using the stock recruitment relationship without recruitment deviates, these models collapse in a short period of time unless catch is drastically reduced. When models with a decreasing trend in recruitment deviates are projected without deviates, the fish stock increases rapidly because the recruitments from the stock recruitment relationship are larger than the recruitments estimated for the recent period. This causes large uncertainties in the management advice derived from forward projections that omit recruitment deviates (Figure 12), as observed in the advice provided using the 2021 stock assessment of Indian Ocean yellowfin (Urtizberea et al., 2021). Regardless of the identification of trends in recruitment using threshold p-values for the no-trend hypothesis, we recommend that projections carried out to provide management advice based on stock assessment models be developed using recent recruitment deviates for models showing appropriate diagnostic values.

[Insert Figure 12]

The case of Indian Ocean yellowfin is a compelling example because the grid used for advice in 2021 covers a wide range of options for biological parameters and assumptions. However, trends in recruitment deviates are also identified in other tropical tunas' assessments (Supplementary Information). Indian Ocean skipjack assessment (Figures SI1A and SI1B) displays decreasing trends in 4 of 24 scenarios (17%), all associated with the largest productivity levels estimated in the grid of models used for management advice. The Indian Ocean bigeye assessment (Figures SI2A and SI2B) doesn't have any model with a p-value of less than 0.1 for the no-trend hypothesis but neither model with a p-value larger than 0.68. In the Atlantic, there are two stock assessments carried out with integrated models. The Atlantic bigeye assessment (Figures SI3A and SI3B) includes 17 cases from the reference grid of 27 models with increasing trends (63%) associated with the lower range of MSY estimates. There is no recruitment deviate trend identified in the four models of the Atlantic yellowfin assessment reference grid (Figures SI4A and SI4B). In the Eastern Pacific, there are two tropical tunas assessed using integrated models. The lowest p-values for the no-trend hypothesis (13 of 44 models, 29%) correspond to the lower and higher tails of the MSY estimated in the reference grid for Eastern Pacific bigeye (Figures SI5A and SI5B), and for most models the null hypothesis was not rejected. For Eastern Pacific yellowfin (Figures SI6A and SI6B), a significant number of models display a recruitment deviate trend, and, in all cases, this is negative (26 of 48 models, 54%) and, 12 of them are associated with MSY estimates well above the largest historical catch of this stock (443,458 tons in 2002) and also the recent catch (average 2000-2020 is 261,165 tons). When interpreting results for the WCPO stocks, consideration should be taken of the specific approach used to estimate recruitment in a spatially structured assessment model. Within MULTIFAN-CL the spatial distribution of recruitment can be allowed to vary in time such that recruitment by time period in each region is estimated somewhat independently. Subsequently, and by design in the terminal assessment phase, an overall stock recruitment relationship is fitted with a weak penalty term so as not to overly influence the recruitment estimates, with the express purpose of estimating equilibrium management quantities such as MSY. This equilibrium calculation is based upon a single region approximation, with overall recruitment, no movement, and averaged fishing mortality over a specified period. The assessments of Western Central Pacific bigeye (Figures SI7A and 7B) and skipjack (Figures SI9A and 9B) display increasing trends in the majority of their models (16 of 24, 67% and 41 of 54, 76% respectively), linked to the lowest MSY estimates of each of the ensembles. For Western Central Pacific yellowfin (Figures SI8A and 8B), 22 of 72 models (31%) display a decreasing trend and these models are not linked to the highest estimated MSYs seen in other stocks. Overall, except for Western Central yellowfin and some models of East Pacific bigeye, increasing trends (pink) are associated with the lower tail of MSY estimates and decreasing trends (blue) are associated with the higher tail.

The relative roles of intrinsic and extrinsic factors in population dynamics have been investigated in ecology, and ecologists have aimed at quantifying the real drivers of population dynamics (Ahrestani et al., 2013). In fisheries, it is assumed that fish stocks' population dynamics are driven by natural mortality, growth and reproduction as intrinsic biological factors and, fishing as the main extrinsic factor. In this context, the influence of variables that are not understood or that are ignored in the models are assumed to be random. To elucidate if recruitment deviates represent a source of variability, we compared models with and without recruitment deviates. In the Indian Ocean yellowfin assessment, with fixed growth, natural mortality and steepness, the model can only modulate the R_0 to estimate different levels of productivity of the stock and fit the available data. Figures 4-7 show that when recruitment deviates are randomly distributed, the data and model assumptions are used to estimate the general trend of the population and its productivity because model estimates are similar with and without recruitment deviates. Instead, when there is a trend in recruitment deviates, there are large differences between the estimates of models with and without recruitment deviates, which supports the idea that process error is a factor that is driving the dynamics of the population and not a random variable. If trends were detected in all model configurations it might indicate a lack of identification of the main drivers of the population. When this happens only in certain configurations of models it suggests implausible combinations of parameters. The accumulation of positive recruitment deviates in periods of high catch could be due to underestimation of the mean productivity of the stock (*e.g.*, unfished equilibrium recruitment (R_0)) and alternatives, such as allowing higher penalties on recruitment deviates or estimating recruitment deviates variability, may need to be investigated.

Process error and recruitment deviates may also potentially represent the variation in the true population due to factors not included in the equations of the stock assessments such as environmental regime shifts or productivity changes. For example, there is evidence that environmental drivers such as climate change can produce variability and alterations in the underlying productivity of fish stocks that can have important impacts on fisheries and their management (Alheit et al., 2009; Allison et al., 2009; Arnason 2006; Barange et al., 2014; Brander 2007; Chavez et al., 2003; Cheung et al., 2009; Erauskin-Extramiana et al., 2019; Merino et al., 2012). However, we develop our method in the context of large uncertainty ensembles of models where only certain model configurations display trends in recruitment. Should evidence of the impact of factors not considered in stock assessments be available, these factors would need to be included in the stock assessment, which is possible in integrated models. However, systematic examination is necessary to assume stock productivity shifts in stock assessments (Klaer et al., 2015). Also, such evidence of regime shifts and changes in productivity should be used to calculate fish stocks productivity (in this case R_0) in the different years of the stock assessment period. With this, the reference points used to provide management advice would also be adapted to the inferred changes in the productivity of the stocks.

We used data generated from a simulated population to see if our method is able to identify problems with model configurations that are intentionally incorporated as bias in natural mortality and growth. Natural mortality is one of the least well-understood population processes included in stock assessments and the trends observed in the Atlantic bigeye assessment (Figure SI3A) suggest that changes in this parameter would be sufficient to provide trends in recruitment deviates. Figures 8 and 9 show that the expected trends are only observed for very large bias from the true value of the simulated model (±90%). This was somewhat unexpected because the assumptions on M developed for Atlantic bigeye include natural mortality reductions of 23% and 37% respectively for the scenarios M20 and M25 relative to the M17 models, and these changes do produce trends in recruitment deviates. However, the trends observed in the Indian Ocean yellowfin assessment were reproduced and they displayed the expected slope, increasing for low natural mortality and decreasing for large natural mortality. The reasons for the absence of recruitment deviates' trends except for large bias in *M* for the operating model needs to be explored further. However, there is good consistency between the CPUE and catch data in the simulated model. The mis-specified natural mortality produces changes in the overall productivity estimate (e.g., R_0) but doesn't affect the trend as they do in the stock assessment, where inconsistencies between abundance indices, catch, size frequency and tagging data have been identified (Fu et al., 2021). The Indian Ocean yellowfin stock assessment and the operating model are spatially disaggregated and, in the past, trends in the regional recruitment distribution have also been encountered (IOTC 2021). These are shown to have been mostly associated with trends in catch distribution (i.e., large increase of the regional recruitment often coincided with the high catch), and may also reflect model-misspecification (*e.g.,* the prior assumption imposed on the regional abundance distribution).

In the Indian Ocean yellowfin assessment, we observed trends in recruitment deviates in specific model configurations but not in single factors. For example, 23 of 24 models with the low natural mortality and Dortel growth (Dortel et al., 2014) combination display a trend in recruitment deviates. However, there are models with the Dortel growth combined with base natural mortality or models with the low mortality option combined with the base growth curve that show a high probability for the no-trend hypothesis. This suggests that this method should not be used to discard or select entire factors from a reference grid but to identify problems with specific model configurations (combinations of factors) and eventually, discard or select individual models. This also suggests that the cause of the trends is probably not a single parameter but the result of the combination of factors and possible inconsistencies and conflict between observations.

Diagnostics can be used to evaluate model plausibility when using integrated stock assessment models (Carvalho et al., 2021). The p-value of the no-tend hypothesis adds to the statistical tests currently applied to evaluate model performance and help model development and selection when using ensembles of models to develop management advice. The p-value can identify problematic models that are not identified using retrospective analyses, hindcasting and partially with runs tests (Carvalho et al., 2021). Our results suggest that overall, models with trends in recruitment deviates are linked with models with poorer performance in runs test and therefore, with models with residuals to abundance indices that are not random (Carvalho et al., 2017). The ASPM diagnostic has previously shown good power to detect misspecification of system dynamics (steepness of the stock recruitment relationship and natural mortality) and confirmation that stock dynamics are driven by stock's production function (Carvalho et al., 2017; Minte-Vera et al., 2017). Our results indicate that differences between ASPM and the stock assessment models are linked to autocorrelated recruitment deviates which supports the idea that in these models, recruitment deviates are not random and represent part of the stock's response to fishing. Finally, our results show that trends in recruitment deviates are also coincident with larger variability and autocorrelation of recruitment deviates. This also supports the idea that process error is not a random process in many models and furthermore, it suggests that in practice, recruitment deviates do not only explain natural variation but also act as a sink to allow fits to observations in mis-specified models.

The results of the diagnostic analyses show that no single diagnostic can be used in isolation, and it is difficult to assign a single criterion for discarding or selecting models. The p-value of the no-trend hypothesis is fast and easy to calculate which makes it powerful when running stock assessments in dedicated meetings with decisions needed in a short time. Diagnostics as developed by Carvalho et al., (2021) have recently been used for model weighting used to develop management advice (GFCM, 2021). The preliminary model weighting work done by Maunder et al. 2020 shows that problems remain when assigning weights to model according to model diagnostics. Although it is straightforward that the models which perform better in diagnostics should be given higher weights, how to quantify weights given various diagnostics performance can be subjective and controversial. Also, model weighting could also include expert's opinions (e.g., regarding the assumptions of steepness in Maunder et al 2020). The process of translating expert's opinions into quantitative weighting is inherently subjective and can be problematic. Our results indicate that trends in recruitment deviates can provide statistical evidence to help model discard/selection or quantitative weighting when using large ensembles of models. We have used the p–value of 0.1 as the threshold to link the falsehood of the no-trend hypothesis for recruitment deviates with the productivity of stocks and, to elucidate the role of process error as a random variable or as part of the intrinsic factors that drive fish stocks' response to fishing. However, this value is arbitrary. The aim is to identify models that are problematic or mis-specified but we acknowledge that other values could have been used. In other words, the low probability for the no-trend hypothesis helps identify models with potential problems of misspecification of parameters and incompatibility between assumptions and data that need to be investigated. We recommend that models with a p-value below a threshold are analysed carefully before selection for the ensemble of models used to develop management advice.

In conclusion, this study highlights problems in the configuration of tropical tuna stock assessment models and identifies a method to discard assumptions and model configurations that are incompatible with the available information. The investigation of recruitment deviation trends provides opportunities to reduce the uncertainty in stock assessments and to contribute to the improvement of the management of fish and fishery resources. We have based our analysis in the Indian Ocean yellowfin assessment and other tropical tunas, but the methodology can be extrapolated to other fisheries.

Acknowledgments

This paper is contribution nº xxxx from AZTI, Marine Research, Basque Research and Technology Alliance (BRTA).

Funding:

This research has been supported by the Economic Development, Sustainability and Environment directorate from the Basque Government through the program "Acuerdo Marco Pesca (2020–2023)".

References

Ahrestani, F.S., Hebblewhite, M., Post, E., 2013. The importance of observation versus process error in analyses of global ungulate populations. Scientific Reports, 3(1): 3125. 10.1038/srep03125.

Alheit, J., Roy, C., Kifani, S., 2009. Decada-scale variability in populations. Climate Change and Small Pelagic Fish, 10.1017/CBO9780511596681.007.

Allen, K.R., 1973. The influence of random fluctuations in the stock recruitment relationship on the economic return from salmon fisheries. Rapports et procès-verbaux des réunions Conseil permanent international pour l'exploration de la mer, 164: 350–359.

Allen, R., Joseph, J., Squires, D., 2010. Conservation and Management of Transnational Tuna Fisheries. Iowa, USA. Wiley-Blackwell. ISBN: 978-0-813-80567-2.

Allison, E.H., Perry, A.L., Badjeck, M.C., Adger, W.N., Brown, K., Conway, D., Halls, A.S., Pilling, G., Reynolds, J.D., Andrew, L.N., Dulvy, N., 2009. Vulnerability of national economies to the impacts of climate change on fisheries. Fish and Fisheries, 10: 173-196.

Arnason, R., 2006. Global warming, small pelagic fisheries and risk. in: Hanneson R., Barange M., Herrick S.J., eds. Climate change and the economics of the world's fisheries. Cheltenham (UK). Edwar Elgar Publishing Limited.

Barange, M., Merino, G., Blanchard, J., Scholtens, J., Harle, J., Allison, E.H., Allen, I., Holt, J., Jennings, S., 2014. Impacts of climate change on marine ecosystem production in societies dependent on fisheries. Nature Climate Change, 4: 211-216.

Brander, K., 2007. Global fish production and climate change. Proceedings of the National Academy of Sciences of the United States of America, 104: 19709-19714.

Brooks, E.N., Legault, C.M., 2016. Retrospective forecasting-evaluating performance of stock projections for New England groundfish stocks. Canadian Journal of Fisheries and Aquatic Sciences, 73: 935–950.

Carvalho, F., Punt, A.E., Chang, Y.-J., Maunder, M.N., Piner, K.R., 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fisheries Research, 192: 28–40.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R., Maunder, M.N., Taylor, I., Wetzel, C.R., Doering, K., Johnson, K.F., Methot, R.D., 2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research, 240: 105959.

Castillo, C., Hampton, J., Ducharme-Barth, N., Xu, H., Vidal, T., Williams, P., Scott, F., Pilling, G., Hamer, P., 2021. Stock assessment of South Pacic albacore tuna. WCPFC-SC17-2021/SA-WP-02.

Chavez, F.P., Ryan, J., Lluch-Cota, S.E., Niquen, M., 2003. From anchovies to sardines and back: multidecadal change in the Pacific ocean. Science, 299: 217-221.

Cheung, W.W.L., Lam, V.W.Y., Sarmiento, J.L., Kearney, K., Watson, R., Zeller, D., Pauly, D., 2009. Large-scale redistribution of maximum fisheries catch potential in the global ocean under climate change. Global Change Biology, 16:24-35.

Cordue, P.L. 2001. A note on incorporating stochastic recruitment into deterministic age structured 518 population models. ICES Journal of Marine Science, 58: 794–798. https://doi.org/10.1006/jmsc.2001.1069

Dortel, E., Sardenne, F., Bousquet, N., Rivot, E., Million, J., Croizier, G.L., Chassot, E., 2014. An integrated Bayesian modeling approach for the growth of Indian Ocean yellowfin tuna. Fisheries Research, 163: 69-84.

Dunn, A., Hoyle, D.S., Datta, S., 2020. Development of spatially explicit operating models for yellowfin tuna populations in the Indian Ocean. IOTC Working Party on Tropical Tunas 22(AS), IOTC-2020-WPT22(AS)-19.

Erauskin-Extramiana, M., Arrizabalaga, H., Hobday, A.J., Cabré, A., Ibaibarriaga, L., Arregui, I., Murua, H., Chust, G., 2019. Large-scale distribution of tuna species in a warming ocean. Global Change Biology, 25: 2043-2060.

Fonteneau, A. 2008. A working proposal for a Yellowfin growth curve to be used during the 2008 yellowfin stock assessment. IOTC-2008-WPTT-4.

Francis, R.I.C.C., Shotton, R., 1997. “Risk” in fisheries management: a review. Canadian Journal of Fisheries and Aquatic Sciences, 54: 1699-1715.

Froese, R., Winker, H., Gascuel, D., Sumaila, R.U., Pauly, D. 2016. Minimizing the impact of fishing. Fish and Fisheries, 17(3): 785-802. https://doi.org/10.1111/faf.12146.

Froese, R., Zeller, D., Kleisner, K.M., Pauly, D., 2021. What catch data can tell us about the status of global fisheries. Marine Biology, 159:1283-1292.

Fromentin, J.M., Bonhommeau, S., Arrizabalaga, H., Kell, L.T., 2014. The spectre of uncertainty in management of exploited fish stocks: The illustrative case of Atlanticbluefin tuna. Marine Policy, 47: 8-14.

Fu, D., Urtizberea, A., Cardinale, M., Methot Jr, R.D., Hoyle, D.S., Merino, G., 2021. Preliminary Indian Ocean yellowfin tuna stock assessment 1950-2020 (Stock Synthesis). IOTC–2021–WPTT23–12.

GFCM, 2021. Report of the Working Group on Stock Assessment of Demersal Species (WGSAD) – Benchmark session for the assessment of common sole in GSA 17, Scientific Advisory Committee on Fisheries (SAC). Online via Microsoft Teams, 12–16 April 2021.

Hightower, J.E., Grossman, G.D., 1985. Comparison of constant effort harvest policies for fish stocks with variable recruitment. Canadian Journal of Fisheries and Aquatic Sciences. 42: 982–988.

Holling, C.S., 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics, 4: 1-23.

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R., McGilliard, C.R., Monnahan, C.C., Muradian, M.L., Ono, K., Vert-Pre, K.A., Whitten, A.R., Punt, A.E., 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES Journal of Marine Science, 72: 99–110. https://doi.org/10.1093/icesjms/fsu198.

ICES, 2022. Benchmark workshop on Pandalus stocks (WKPRAWN). ICES Scientific Reports. 4:20, 249 pp. http://doi.org/10.17895/ices.pub.19714204.

Ichinokawa, M., Okamura, H., Takeuchi, Y., 2014. Data conflict caused by model misspecification of selectivity in an integrated stock assessment model and its potential effects on stock status estimation. Fisheries Research, 158: 147-157.

IOTC., 2021. Report of the 24th Session of the IOTC Scientific Committee. IOTC–2021–SC24–R[E].

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E.N., Johnson, K.F., Klibansky, N., Millar, C.P., Minto, C., Mosqueira, I., Nash, R.D.M., Vasilakopoulos, P., Wells, B.K. 2021. Operationalizing ensemble models for scientific advice to fisheries management. ICES Journal of Marine Science, 78(4) 1209–1216. https://doi.org/10.1093/icesjms/fsab010.

Johnson, k.f., Councill, E., Thorson, J.T., Brooks, E., Methot, R.D., Punt, A.E. 2016. Can autocorrelated recruitment be estimated using integrated assessment models and how does it affect population forecasts? Fisheries Research, 183: 222-232. 10.1016/j.fishres.2016.06.004

Juan-Jordá, M.J., Mosqueira, I., Cooper, A.B., Freire, J., Dulvy, N.K., 2011. Global population trajectories of tunas and their relatives. Proceedings of the National Academy of Sciences of the United States of America, 108: 20650-20655.

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., Fu, D., 2021. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science, 78: 2244-2255.

Klaer, N.L., O'Boyle, R.N; Deroba, J.J., Wayte, S.E., Little, L.R., Alade, L.A., Rago, P.J., 2015. How much evidence is required for acceptance of productivity regime shifts in fish stock assessments: Are we letting managers off the hook? Fisheries Research, 168, 49-55.
Kleiber, P., Hampton, J., Davies, N., Hoyle, S., Fournier, D.A., 2012. MULTIFAN-CL User's Guide.
Lyubchich, V., Gel, Y.R., Brenning, A., Chu, C., Huang, X., Islambekov, U., Niamkova, P., Ofori-Boateng, D., Schaeffer, E.D., Vishwakarma, S., Wang, X., 2022. Package 'funtimes' https://cran.r-project.org/web/packages/funtimes/funtimes.pdf.
Mangel, M., MacCall, A.D., Brodziak, J., Dick, E.J., Forrest, R.E., Pourzard, R., Ralston, S. A., 2013. Perspective on steepness, reference points, and stock assessment. Canadian Journal of Fisheries and Aquatic Sciences, 70: 930-940.
Maunder, M.N., Hamel, O., Ianelli, J., 2021. Natural Mortality: Theory, estimation and application in fishery stock assessment models. Special Issue for Fisheries Research.
Maunder, M.N., Piner, K.R. 2015. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science, 72(1): 7–18. https://doi.org/10.1093/icesjms/fsu015.
Maunder, M.N., Xu, H., Lennert-Cody, C.E., Valero, J.L., Aires-da-Silva, A., Minte-Vera, C., 2020. Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. Inter-Americal Tropical Tuna Commission, SAC-11 INF-F.
Meador, M.R., Brown, L.M., 2015. Life history strategies of fish species and biodiversity in eastern USA streams. Environmental Biology of Fishes, 98: 663-677.
Merino, G., Barange, M., Blanchard, J., Harle, J., Holmes, R., Allen, I., Allison, E.H., Badjeck, M.C., Dulvy, N., Holt, J., Jennings, S., Mullon, C., Rodwell, L., 2012. Can marine fisheries and aquaculture meet fish demand from a growing human population in a changing climate? Global Environmental Change, 22: 795-806.
Merino, G., Die, D.J., A., U., Laborda, A., 2021. Characterization of structural uncertainty in tropical tuna stocks' dynamics. SCRS/2021/016.
Merino, G., Murua, H., Santiago, J., Arrizabalaga, H., Restrepo, V., 2020. Characterization, communication, and management of uncertainty in tuna fisheries. Sustainability, 12: 8245.
Mertz, G., Myers, R.A. 1996. Influence of fecundity on recruitment variability of marine fish. Canadian Journal of Fisheries and Aquatic Sciences, 53: 1618-1625.
Methot Jr, R.D., Wetzel, C.R., 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research, 142: 86-99.
Methot, Jr, R.D., Taylor, I.G., 2011. Canadian Journal of Fisheries and Aquatic Sciences, 68: 1744-1760.
Minte-Vera, C.V., Maunder, M.N., Xu, H., Valero, J.L. Lennert-Cody, C.E., Aires-da-Silva, A. 2020. Yellowfin tuna in the eastern Pacific Ocean, 2019: benchmark assessment. 11th meeting of the Scientific Advisory Committee. Document SAC-11-07. Available from: https://www.iattc.org/Meetings/Meetings2020/SAC-11/Docs/_English/SAC-11-07-MTG_Yellowfin%20tuna%20benchmark%20assessment%202019.pdf.
Minte-Vera, C.V., Maunder, M.N., Aires-da-Silva, A.M., Satoh, K., Uosaki, K., 2017. Get the biology right, or use size-composition data at your own risk. Fisheries Research, 192: 114–125.
Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES Journal of Marine Science. 56: 473–488. https://doi.org/10.1006/jmsc.1999.0481.
Murua, H., Rodriguez-Marin, E., Neilson, J., Farley, J.H., Juan-Jordá, M.J., 2017. Fast versus slow growing tuna species–age, growth, and implications for population dynamics and fisheries management. Fish Biology and Fisheries, 27: 733-773.

O'Leary, J.K., Micheli, F., Airoldi, L., Boch, C., De Leo, G., Elahi, R., Ferretti, F., Graham, N.A.J., Litvin, S.Y., Low, N.H., Lummis, S., Nickols, K.J., Wong, J., 2017. The resilience of marine ecosystems to climatic disturbances. BioScience, 67: 208-220.

Pimm, S.L., 1984. The complexity and stability of ecosystems. Nature, 307: 321-326.

Quinn, T.J.I., Deriso, R.B., 1999. Quantitative fish dynamics. Oxford University Press, New York.

R Core Team., 2021. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ricker, W.E. 1954. Stock and recruitment. Journal of the Fisheries Research Board of Canada, 11: 559–623.

Rose, K.A., Cowan Jr., J.H., Winemiller, K.O., Myers, R.A., Hilborn, R., 2001. Compensatory density dependence in fish populations: importance, controversy, understanding and prognosis. Fish and Fisheries, 2: 293–327.

Rosenberg, A.A., Restrepo, V.R., 1994. Uncertainty and risk evaluation in stock assessment advice. Canadian Journal of Fisheries and Aquatic Sciences, 51: 2715-2720.

Sampson, D.B., Scott, R.D. 2011. An exploration of the shapes and stability of population-selection curves. Fish and Fisheries, 13(1): 89-104. https://doi.org/10.1111/j.1467-2979.2011.00417.x.

Schuch, E., Richter, A., 2022. Tracing intuitive judgement of experts in fish stock assessment data. Fish and Fisheries, 23(3): 758-767.

Sharma, R., Porch, C., Babcock, E., Maunder, M., Punt, A., 2019. Recruitment: Theory, estimation, and application in fishery stock assessment models. Fisheries Research, 217: 1-4.

Taylor, I.G., Doering, K.L., Johnson, K.F., Wetzel, C.R., Stewart, I.J., 2021. Beyond visualizing catch-at-age models: Lessons learned from the r4ss package about software to support stock assessments. Fisheries Research, 239: 105924.

Thorson, J.T. 2019. Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. Fish and Fisheries, 21(2): 237-251.

Thorson, J.T., Jensen, O.P., Zipkin, E.F. 2014. How variable is recruitment for exploited marine fishes? A hierarchical model for testing life history theory. Canadian Joural of Fisheries and Aquatic Sciences, 71: 973-983. 10.1139/cjfas-2013-0645.

Urtizberea, A., Fu, D., Merino, G., Methot Jr, R.D., Cardinale, M., Winker, H., Walter, J., Murua, H., 2019. Preliminary assessment of Indian Ocean yellowfin tuna 1950-2018 (Stock Synthesis, V3.30). IOTC–2019–WPTT21–50.

Urtizberea, A., Fu, D., Schirripa, M., Methot Jr, R.D., Cardinale, M., Hoyle, D.S., Merino, G., 2021. Indian Ocean yellowfin tuna SS3 model projections. IOTC–2021–SC24-INF08.

Wald, A., Wolfowitz, J., 1940. On a test whether two samples are from the same population. The Annals of Mathematical Statistics, 11: 147-162.

Walter, J., Hiroki, Y., Satoh, K., Matsumoto, T., Winker, H., Urtizberea, A., Schirripa, M., 2019. Atlantic bigeye tuna Stock Synthesis projections and Kobe 2 matrices. ICCAT Collect. Vol. Sci. Pap, 75, 2283–2300.

Walters, C.J., Ludwig, D., 1981. Effects of measurement errors on the assessment of Stock–Recruitment relationships. Canadian Journal of Fisheries and Aquatic Sciences, 38: 704–710.

Wang, H.-Y., Shen, S.-F., Chen, Y.-S., Kiang, Y.-K., Heino, M., 2020. Life histories determine divergent population trends for fishes under climate warming. Nature communications, 11: 4088.

Winker, H., Carvalho, F., Thorson, J.T., Kell, L.T., Parker, D., Kapur, M., Sharma, R., Booth, A.J., Kerwath, S.E. 2020. JABBA-Select: Incorporating life history and fisheries' selectivity into surplus production models. Fisheries Research, 222: 105355.

Winker, H., Walter, J. 2019. Application of a multivariate lognormal approach to estimate uncertainty about the stock status and future projections for Indian Ocean yellowfin tuna. IOTC–2019–WPTT21–51.

Xu, H., Maunder,M.N., Minte-Vera, C., Valero, J.L., Lennert-Cody, C., Aires-da-Silva, A. 2020. Bigeye tuna in the eastern Pacific Ocean, 2019: benchmark assessment. Inter-American Tropical Tuna Commission. 11th meeting of the Scientific Advisory Committee. Document SAC-11-06. Available from: https://www.iattc.org/Meetings/Meetings2020/SAC-11/Docs/_English/SAC-11-06-MTG_Bigeye%20tuna%20benchmark%20assessment%202019.pdf.

Zhou, Shijie., Yin, Shaowu., Thorson, J., Smith, A.D.M., Fuller, M. 2012. Linking fishing mortality reference points to life history traits: An empirical study. Canadian Journal of Fisheries and Aquatic Sciences. 69, 1292-1301. 10.1139/f2012-060.

Diagnostic	Description
sd(rec-devs)	*Standard deviation of the recruitment deviates across the fitting period.*
Autocorrelation	*First order auto-regressive (AR1) autocorrelation coefficient recruitment deviates at an annual time step interval.*
RunsTest	*Runs test for residual analysis (Carvalho and others 2017) applied to the abundance indices available for the stock assessment. It represents the probability of residuals to be random. If p-values larger than 0.05 are considered representative of models with random residuals.*
Mohn's ρ *(B)*	*Mean relative error of the biomass estimate using the full dataset and the estimate of sequentially removing years with data (Carvalho and others 2017; Carvalho and others 2021; Hurtado-Ferro and others 2015; Mohn 1999). The closer the value to zero, the smaller the retrospective bias.*
Mohn's ρ *(F)*	*Mean relative error of the fishing mortality estimate using the full dataset and the estimate of sequentially removing years with data (Carvalho and others 2017; Carvalho and others 2021; Hurtado-Ferro and others 2015; Mohn 1999). The lower the value, the more robust the model.*
MASE	*Mean absolute scaled error (Hyndman and Koehler 2006). Evaluates the prediction skill of a model relative to a naïve baseline prediction by scores of the mean absolute error of forecasts (prediction residuals) (Carvalho and others 2021). The lower the value, the prediction skills of the model are assume better. If the MASE is smaller than one, the model is considered to have prediction skill.*
MSY (ASPM-SA)	*Difference in the estimated Maximum Sustainable Yield (MSY) as a measure of productivity between the stock assessment (SA) models and equivalent runs without recruitment deviates and using only catch and effort data (ASPM).*
R0 (ASPM-SA)	*Difference in the estimated unfished equilibrium recruitment (R0) as a measure of scale between the stock assessment (SA) models and equivalent runs without recruitment deviates and using only catch and effort data (ASPM).*

Table. 1. Diagnostics used for comparison with the no-trend hypothesis for recruitment deviates.

Model name	*p-value NoTrend*	*sd (rec-devs)*	*Autocorrelation*	*RunsTest*	*Mohn's rho (B)*	*Mohn's rho (F)*	*MASE*	*MSY (ASPM-SA)*	*R0 (ASPM-SA)*
io_h70_q1_Gbase_Mbase_tlambda01	0.165	0.377	0.263	0.082	0.123	0.360	1.080	0.059	0.056
io_h70_q1_Gbase_Mbase_tlambda1	0.825	0.395	0.253	0.156	0.084	0.220	1.030	0.030	0.006
io_h70_q1_Gbase_Mlow_tlambda01	0.655	0.392	0.189	0.049	0.137	0.391	1.085	0.031	0.009
io_h70_q1_Gbase_Mlow_tlambda1	0.007	0.425	0.271	0.053	0.114	0.259	1.067	0.119	0.067
io_h70_q1_GDortel_Mbase_tlambda01	0.337	0.458	0.428	0.074	NA	NA	NA	0.129	0.056
io_h70_q1_GDortel_Mbase_tlambda1	0.108	0.473	0.405	0.099	0.073	0.251	1.011	0.120	0.078
io_h70_q1_GDortel_Mlow_tlambda01	0.001	0.536	0.546	0.030	0.016	0.252	1.093	0.167	0.150
io_h70_q1_GDortel_Mlow_tlambda1	0.001	0.568	0.574	0.045	0.112	0.172	1.031	0.183	0.183
io_h70_q2_Gbase_Mbase_tlambda01	0.610	0.367	0.241	0.118	0.087	0.319	1.054	0.016	0.026
io_h70_q2_Gbase_Mbase_tlambda1	0.272	0.393	0.249	0.134	0.102	0.328	1.037	0.071	0.019
io_h70_q2_Gbase_Mlow_tlambda01	0.228	0.394	0.221	0.100	0.128	0.343	1.084	0.084	0.017
io_h70_q2_Gbase_Mlow_tlambda1	0.004	0.434	0.299	0.039	0.143	0.351	1.068	0.158	0.084
io_h70_q2_GDortel_Mbase_tlambda01	0.049	0.462	0.421	0.143	0.084	0.343	1.035	0.145	0.074
io_h70_q2_GDortel_Mbase_tlambda1	0.078	0.485	0.442	0.098	0.050	0.315	1.023	0.182	0.109
io_h70_q2_GDortel_Mlow_tlambda01	0.003	0.557	0.582	0.044	0.085	0.367	1.030	0.235	0.189
io_h70_q2_GDortel_Mlow_tlambda1	0.000	0.598	0.592	0.045	0.083	0.312	1.050	0.264	0.224
io_h80_q1_Gbase_Mbase_tlambda01	0.052	0.383	0.302	0.062	0.100	0.339	1.045	0.050	0.052
io_h80_q1_Gbase_Mbase_tlambda1	0.757	0.387	0.242	0.156	0.080	0.213	1.029	0.003	0.017
io_h80_q1_Gbase_Mlow_tlambda01	0.956	0.389	0.191	0.049	0.127	0.375	1.075	0.014	0.017
io_h80_q1_Gbase_Mlow_tlambda1	0.054	0.411	0.231	0.048	0.123	0.202	1.082	0.104	0.046
io_h80_q1_GDortel_Mbase_tlambda01	0.931	0.453	0.386	0.073	0.030	0.246	1.035	0.069	0.022
io_h80_q1_GDortel_Mbase_tlambda1	0.316	0.464	0.396	0.099	0.067	0.220	0.997	0.106	0.063
io_h80_q1_GDortel_Mlow_tlambda01	0.009	0.527	0.536	0.031	0.042	0.249	1.048	0.226	0.130
io_h80_q1_GDortel_Mlow_tlambda1	0.005	0.547	0.547	0.046	0.003	0.286	1.093	0.188	0.169
io_h80_q2_Gbase_Mbase_tlambda01	0.300	0.366	0.244	0.147	0.067	0.350	1.053	0.020	0.026
io_h80_q2_Gbase_Mbase_tlambda1	0.545	0.390	0.263	0.170	0.119	0.038	1.052	0.062	0.015
io_h80_q2_Gbase_Mlow_tlambda01	0.235	0.385	0.210	0.073	0.126	0.375	1.086	0.062	0.010
io_h80_q2_Gbase_Mlow_tlambda1	0.005	0.423	0.277	0.039	0.088	0.347	1.070	0.151	0.073
io_h80_q2_GDortel_Mbase_tlambda01	0.115	0.454	0.412	0.143	0.034	0.314	1.048	0.134	0.064

io_h80_q2_GDortel_Mbase_tlambda1	0.140	0.478	0.429	0.102	0.089	0.407	1.036	0.168	0.094
io_h80_q2_GDortel_Mlow_tlambda01	0.000	0.544	0.570	0.043	0.188	0.564	1.030	0.243	0.179
io_h80_q2_GDortel_Mlow_tlambda1	0.000	0.579	0.586	0.072	0.043	0.293	1.036	0.277	0.217
io_h90_q1_Gbase_Mbase_tlambda01	0.012	0.387	0.284	0.082	0.080	0.331	1.049	0.072	0.051
io_h90_q1_Gbase_Mbase_tlambda1	0.471	0.389	0.230	0.157	0.083	0.197	1.017	0.020	0.020
io_h90_q1_Gbase_Mlow_tlambda01	0.570	0.389	0.186	0.048	0.143	0.416	1.078	0.000	0.022
io_h90_q1_Gbase_Mlow_tlambda1	0.082	0.408	0.233	0.061	0.114	0.247	1.081	0.113	0.050
io_h90_q1_GDortel_Mbase_tlambda01	0.597	0.454	0.389	0.073	0.079	0.243	1.055	0.042	0.010
io_h90_q1_GDortel_Mbase_tlambda1	0.694	0.463	0.408	0.085	0.074	0.233	1.010	0.109	0.062
io_h90_q1_GDortel_Mlow_tlambda01	0.002	0.520	0.546	0.031	0.099	0.399	1.079	0.212	0.169
io_h90_q1_GDortel_Mlow_tlambda1	0.004	0.531	0.528	0.070	0.070	0.280	1.041	0.193	0.152
io_h90_q2_Gbase_Mbase_tlambda01	0.240	0.367	0.243	0.120	0.093	0.321	1.030	0.023	0.026
io_h90_q2_Gbase_Mbase_tlambda1	0.847	0.397	0.235	0.151	0.033	0.323	1.008	0.049	0.028
io_h90_q2_Gbase_Mlow_tlambda01	0.288	0.383	0.196	0.073	0.286	0.677	1.072	0.046	0.005
io_h90_q2_Gbase_Mlow_tlambda1	0.013	0.413	0.257	0.038	0.098	0.305	1.055	0.138	0.061
io_h90_q2_GDortel_Mbase_tlambda01	0.187	0.449	0.403	0.159	0.049	0.356	1.059	0.120	0.056
io_h90_q2_GDortel_Mbase_tlambda1	0.166	0.485	0.419	0.106	0.060	0.337	1.028	0.162	0.094
io_h90_q2_GDortel_Mlow_tlambda01	0.001	0.535	0.561	0.062	0.024	0.280	1.011	0.279	0.197
io_h90_q2_GDortel_Mlow_tlambda1	0.001	0.559	0.570	0.037	0.045	0.218	1.009	0.281	0.202
sp_h70_q1_Gbase_Mbase_tlambda01	0.073	0.371	0.188	0.016	0.132	0.078	1.092	0.084	0.040
sp_h70_q1_Gbase_Mbase_tlambda1	0.420	0.400	0.227	0.115	0.169	0.307	1.093	0.035	0.035
sp_h70_q1_Gbase_Mlow_tlambda01	0.774	0.386	0.172	0.035	NA	NA	NA	0.030	0.013
sp_h70_q1_Gbase_Mlow_tlambda1	0.153	0.423	0.254	0.049	0.176	0.380	1.090	0.161	0.044
sp_h70_q1_GDortel_Mbase_tlambda01	0.953	0.452	0.403	0.072	0.148	0.344	1.081	0.089	0.002
sp_h70_q1_GDortel_Mbase_tlambda1	0.997	0.481	0.432	0.032	0.140	0.369	1.084	0.097	0.028
sp_h70_q1_GDortel_Mlow_tlambda01	0.012	0.514	0.535	0.017	0.162	0.355	1.070	0.201	0.132
sp_h70_q1_GDortel_Mlow_tlambda1	0.050	0.537	0.551	0.010	0.109	0.319	1.147	0.246	0.163
sp_h70_q2_Gbase_Mbase_tlambda1	0.866	0.408	0.244	0.022	0.113	0.236	1.076	0.008	0.021
sp_h70_q2_Gbase_Mlow_tlambda01	0.096	0.393	0.198	0.021	NA	NA	NA	0.092	0.028
sp_h70_q2_Gbase_Mlow_tlambda1	0.075	0.434	0.269	0.018	NA	NA	NA	0.168	0.078
sp_h70_q2_GDortel_Mbase_tlambda01	0.131	0.448	0.405	0.037	0.116	0.312	1.096	0.143	0.066

sp_h70_q2_GDortel_Mbase_tlambda1	0.558	0.476	0.410	0.019	0.081	0.300	1.058	0.100	0.050
sp_h70_q2_GDortel_Mlow_tlambda01	0.000	0.550	0.581	0.013	0.172	0.454	1.086	0.292	0.189
sp_h70_q2_GDortel_Mlow_tlambda1	0.000	0.594	0.585	0.006	0.081	0.443	1.056	0.323	0.245
sp_h80_q1_Gbase_Mbase_tlambda01	0.039	0.375	0.186	0.016	0.131	0.063	1.132	0.091	0.037
sp_h80_q1_Gbase_Mbase_tlambda1	0.399	0.386	0.222	0.067	0.152	0.256	1.072	0.042	0.031
sp_h80_q1_Gbase_Mlow_tlambda01	0.875	0.383	0.161	0.035	0.203	0.450	1.086	0.002	0.019
sp_h80_q1_Gbase_Mlow_tlambda1	0.321	0.415	0.229	0.068	NA	NA	NA	0.123	0.023
sp_h80_q1_GDortel_Mbase_tlambda01	0.797	0.450	0.392	0.026	0.107	0.254	1.192	0.017	0.021
sp_h80_q1_GDortel_Mbase_tlambda1	0.737	0.473	0.409	0.027	0.110	0.342	1.060	0.062	0.003
sp_h80_q1_GDortel_Mlow_tlambda01	0.013	0.500	0.521	0.012	0.159	0.294	1.244	0.233	0.133
sp_h80_q1_GDortel_Mlow_tlambda1	0.086	0.521	0.526	0.010	0.107	0.293	1.089	0.244	0.151
sp_h80_q2_Gbase_Mbase_tlambda01	0.393	0.370	0.167	0.019	0.114	0.180	1.069	0.038	0.022
sp_h80_q2_Gbase_Mbase_tlambda1	0.491	0.403	0.244	0.018	0.113	0.216	1.126	0.021	0.023
sp_h80_q2_Gbase_Mlow_tlambda01	0.119	0.390	0.169	0.016	0.169	0.410	1.143	0.072	0.018
sp_h80_q2_Gbase_Mlow_tlambda1	0.077	0.425	0.253	0.022	0.136	0.399	1.140	0.136	0.054
sp_h80_q2_GDortel_Mbase_tlambda01	0.072	0.451	0.415	0.033	0.137	0.476	1.063	0.135	0.067
sp_h80_q2_GDortel_Mbase_tlambda1	0.991	0.474	0.406	0.019	0.097	0.316	1.080	0.056	0.029
sp_h80_q2_GDortel_Mlow_tlambda01	0.001	0.539	0.569	0.015	NA	NA	NA	0.293	0.179
sp_h80_q2_GDortel_Mlow_tlambda1	0.002	0.577	0.573	0.006	0.070	0.397	1.057	0.335	0.234
sp_h90_q1_Gbase_Mbase_tlambda01	0.010	0.379	0.197	0.016	0.060	-0.165	1.117	0.088	0.033
sp_h90_q1_Gbase_Mbase_tlambda1	0.141	0.400	0.232	0.085	0.071	0.131	1.070	0.069	0.034
sp_h90_q1_Gbase_Mlow_tlambda01	0.670	0.379	0.148	0.035	NA	NA	NA	0.013	0.021
sp_h90_q1_Gbase_Mlow_tlambda1	0.487	0.405	0.207	0.065	0.177	0.358	1.097	0.087	0.015
sp_h90_q1_GDortel_Mbase_tlambda01	0.516	0.437	0.371	0.019	0.128	0.201	1.035	0.022	0.002
sp_h90_q1_GDortel_Mbase_tlambda1	0.553	0.469	0.411	0.032	0.112	0.297	1.053	0.048	0.007
sp_h90_q1_GDortel_Mlow_tlambda01	0.055	0.492	0.510	0.014	0.082	0.223	1.249	0.229	0.121
sp_h90_q1_GDortel_Mlow_tlambda1	0.166	0.512	0.523	0.010	0.100	0.308	1.114	0.233	0.143
sp_h90_q2_Gbase_Mbase_tlambda01	0.195	0.370	0.159	0.015	0.140	0.182	1.054	0.039	0.023
sp_h90_q2_Gbase_Mbase_tlambda1	0.287	0.403	0.244	0.018	0.110	0.300	1.106	0.028	0.022
sp_h90_q2_Gbase_Mlow_tlambda01	0.209	0.386	0.171	0.021	0.163	0.350	1.145	0.054	0.011
sp_h90_q2_Gbase_Mlow_tlambda1	0.229	0.422	0.243	0.014	0.151	0.394	1.124	0.131	0.050

sp_h90_q2_GDortel_Mbase_tlambda01	0.079	0.448	0.404	0.024	0.152	0.391	1.073	0.120	0.058
sp_h90_q2_GDortel_Mbase_tlambda1	0.770	0.467	0.399	0.025	0.024	0.271	1.063	0.093	0.022
sp_h90_q2_GDortel_Mlow_tlambda01	0.001	0.530	0.551	0.010	0.142	0.487	1.122	0.274	0.163
sp_h90_q2_GDortel_Mlow_tlambda1	0.005	0.566	0.552	0.005	0.121	0.401	1.016	0.328	0.214

Table. 2. Performance of the 2021 stock assessment models estimated through diagnostics.

Carvalho, F.; Punt, A.E.; Chang, Y.-J.; Maunder, M.N.; Piner, K.R. Can diagnostic tests help identify model misspecification in integrated stock assessments? . Fisheries Research. 192:28–40; 2017

Carvalho, F.; Winker, H.; Courtney, D.; Kapur, M.; Kell, L.; Cardinale, M.; Schirripa, M.; Kitakado, T.; Yemane, D.; Piner, K.R.; Maunder, M.N.; Taylor, I.; Wetzel, C.R.; Doering, K.; Johnson, K.F.; Methot, R.D. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research. 240:105959; 2021

Hurtado-Ferro, F.; Szuwalski, C.S.; Valero, J.L.; Anderson, S.C.; Cunningham, C.J.; Johnson, K.F.; Licandeo, R.; McGilliard, C.R.; Monnahan, C.C.; Muradian, M.L.; Ono, K.; Vert-Pre, K.A.; Whitten, A.R.; Punt, A.E. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. Ices J Mar Sci 72:99-110; 2015

Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. . Int J Forecast 22:679–688; 2006

Mohn, R. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES J Mar Sci. 56:473-488; 1999

Figure 1

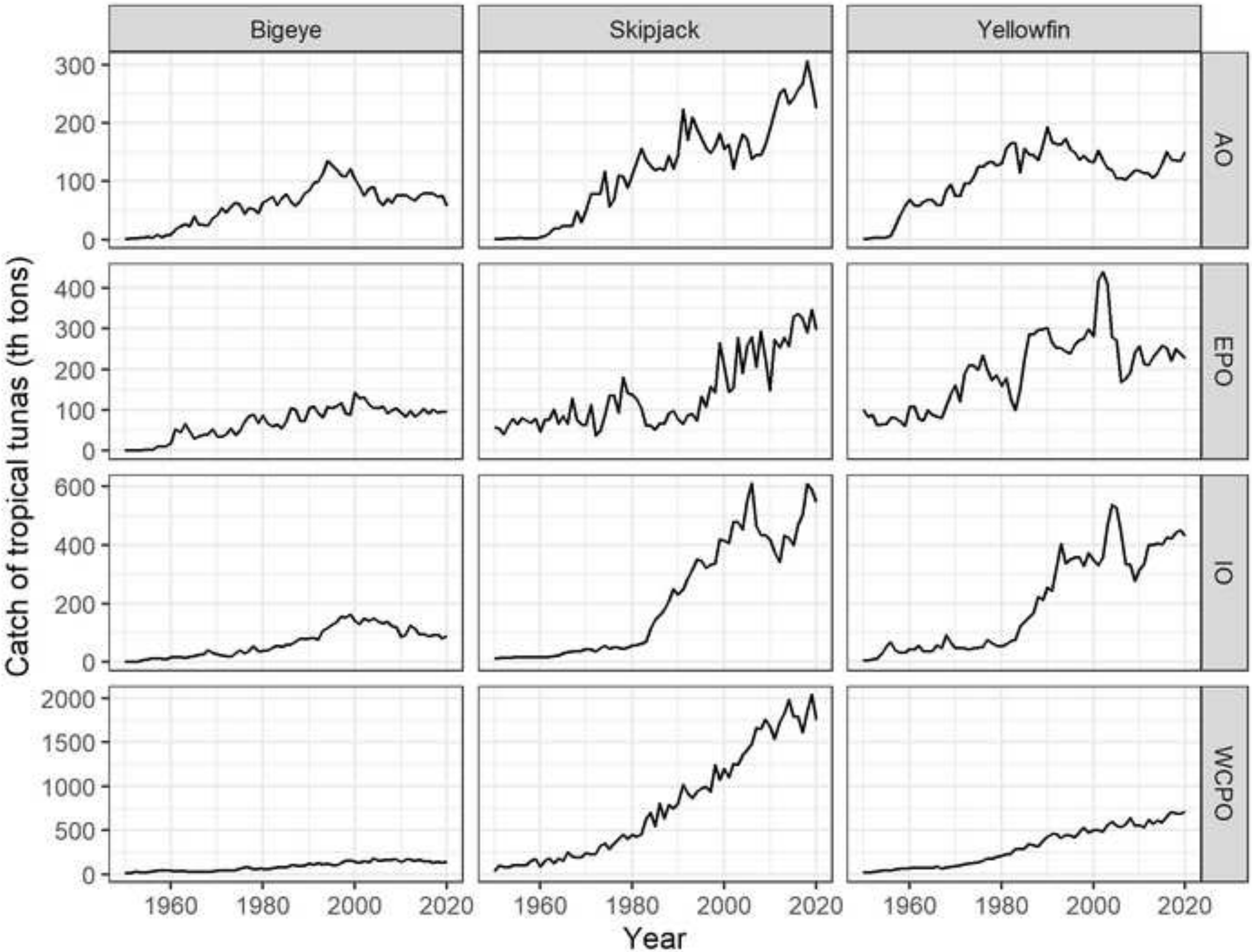

Bigeye
Skipjack
Yellowfin
AO
EPO
IO
WCPO
Catch of tropical tunas (th tons)
Year
0
100
200
300
400
500
600
1000
1500
2000
1960
1980
2000
2020

Figure 2

Click here to access/download;Figure;Figure2.tiff

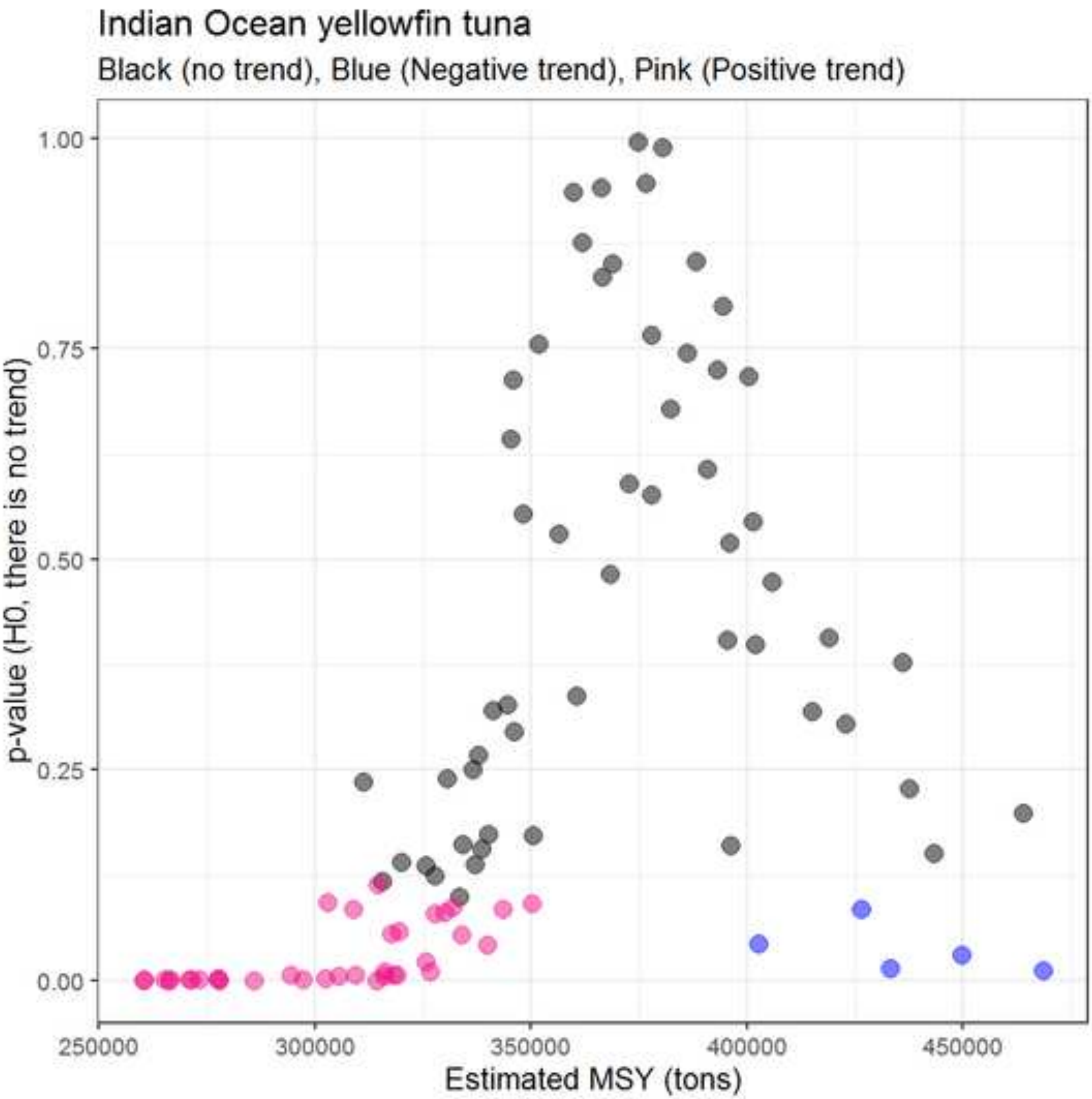

Indian Ocean yellowfin tuna
Black (no trend), Blue (Negative trend), Pink (Positive trend)
p-value (H0, there is no trend)
1.00
0.75
0.50
0.25
0.00
250000
300000
350000
400000
450000
Estimated MSY (tons)

Figure 4

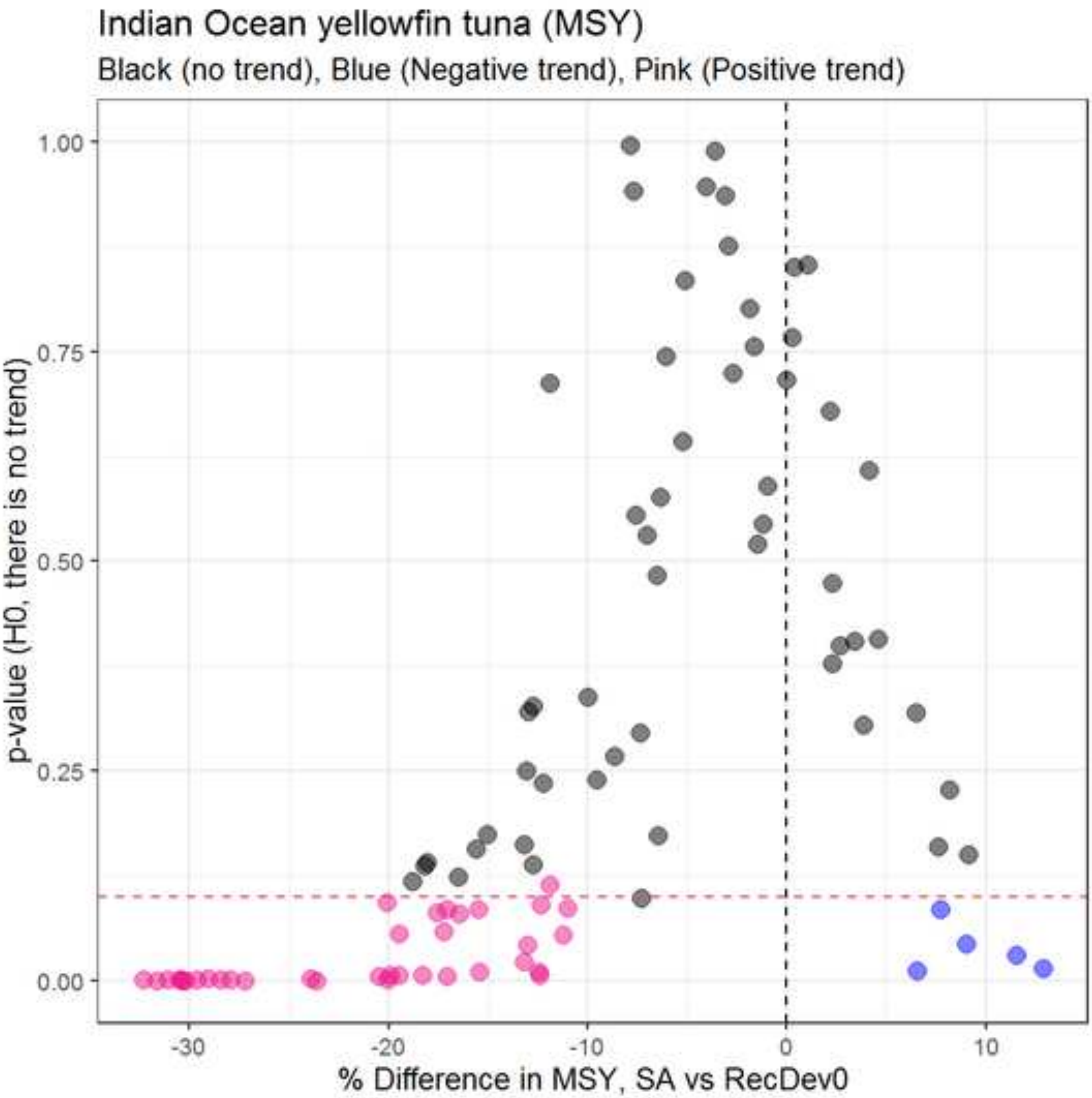

Indian Ocean yellowfin tuna (MSY)
Black (no trend), Blue (Negative trend), Pink (Positive trend)
p-value (H0, there is no trend)
1.00
0.75
0.50
0.25
0.00
-30
-20
-10
0
10
% Difference in MSY, SA vs RecDev0

Figure 5

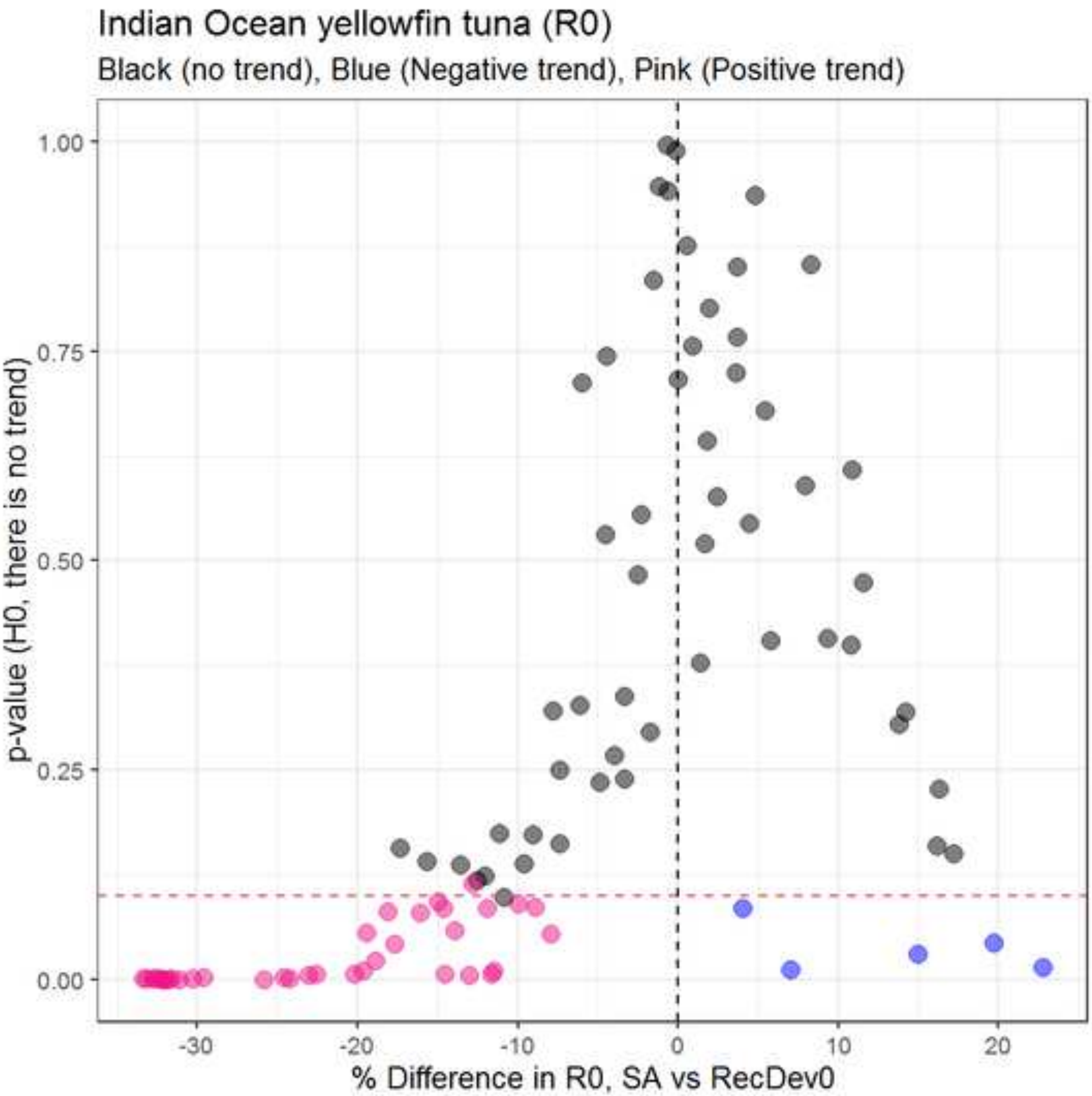

Indian Ocean yellowfin tuna (R0)
Black (no trend), Blue (Negative trend), Pink (Positive trend)
p-value (H0, there is no trend)
1.00
0.75
0.50
0.25
0.00
-30
-20
-10
0
10
20
% Difference in R0, SA vs RecDev0

Figure 6

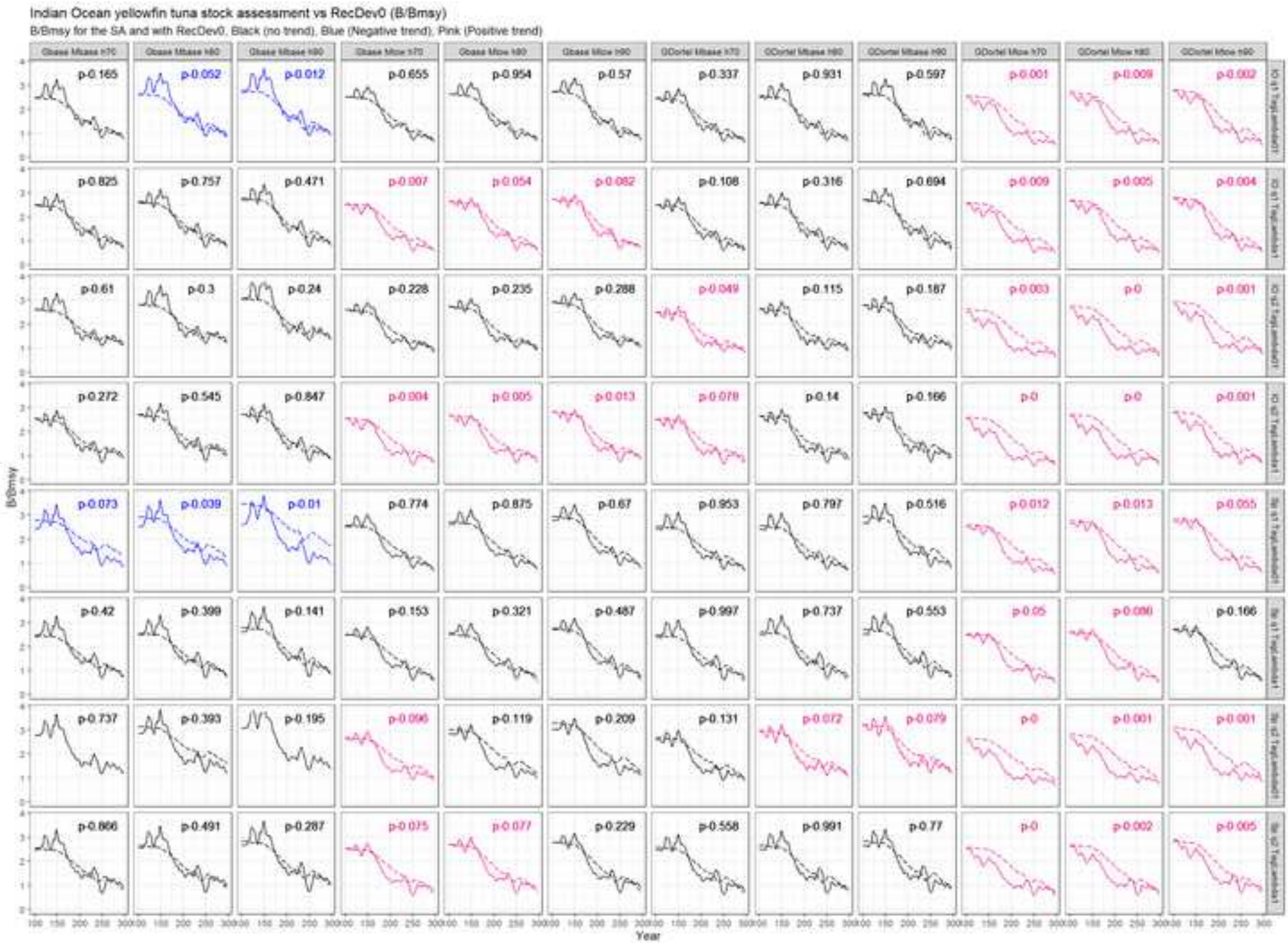

Indian Ocean yellowfin tuna stock assessment vs RecDev0 (B/Bmsy)
B/Bmsy for the SA and with RecDev0. Black (no trend), Blue (Negative trend), Pink (Positive trend)
B/Bmsy
Year

Figure 7

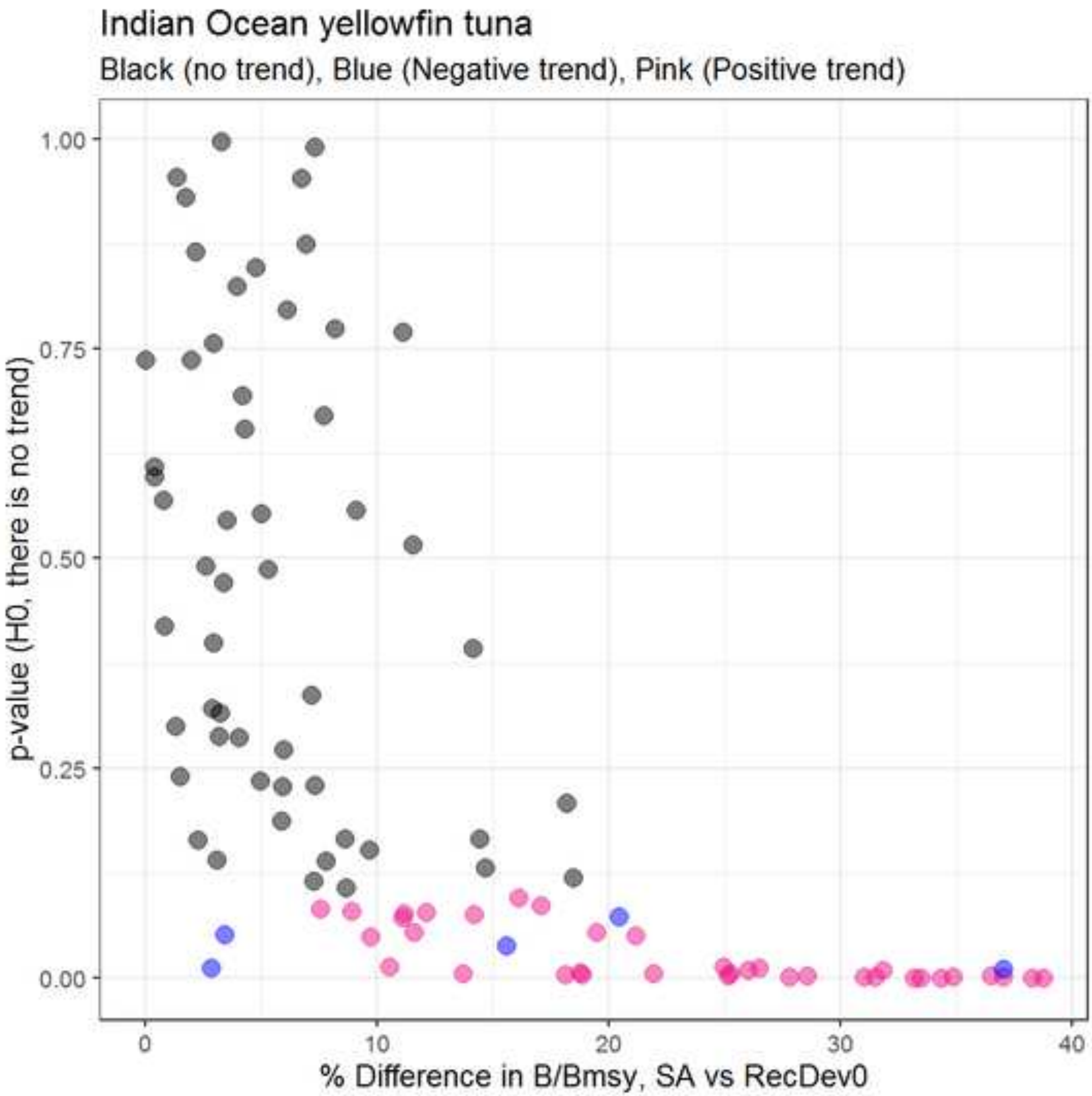

Indian Ocean yellowfin tuna
Black (no trend), Blue (Negative trend), Pink (Positive trend)
p-value (H0, there is no trend)
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
% Difference in B/Bmsy, SA vs RecDev0

Figure 8

Click here to access/download;Figure;Figure8.tiff

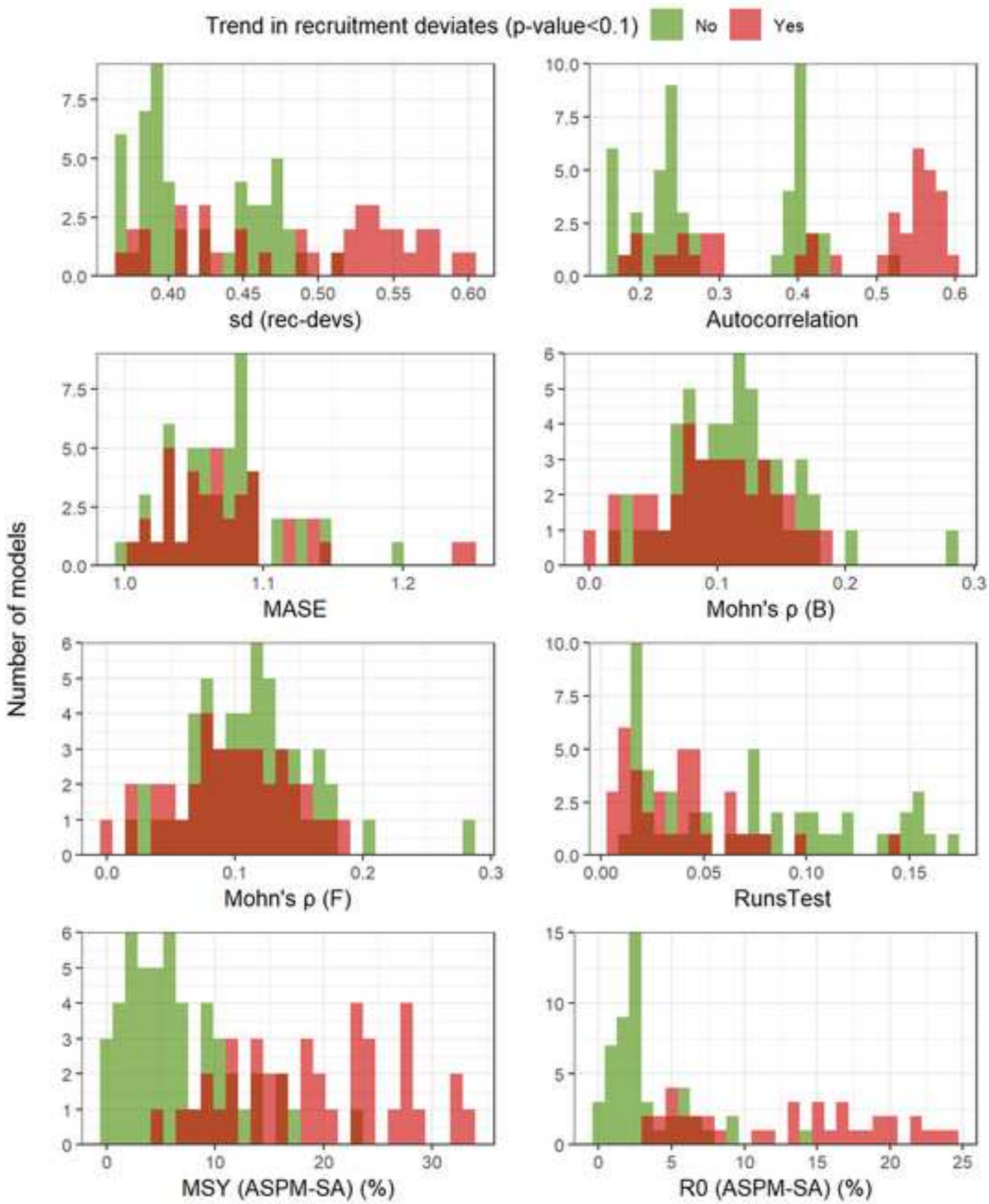


Figure 9

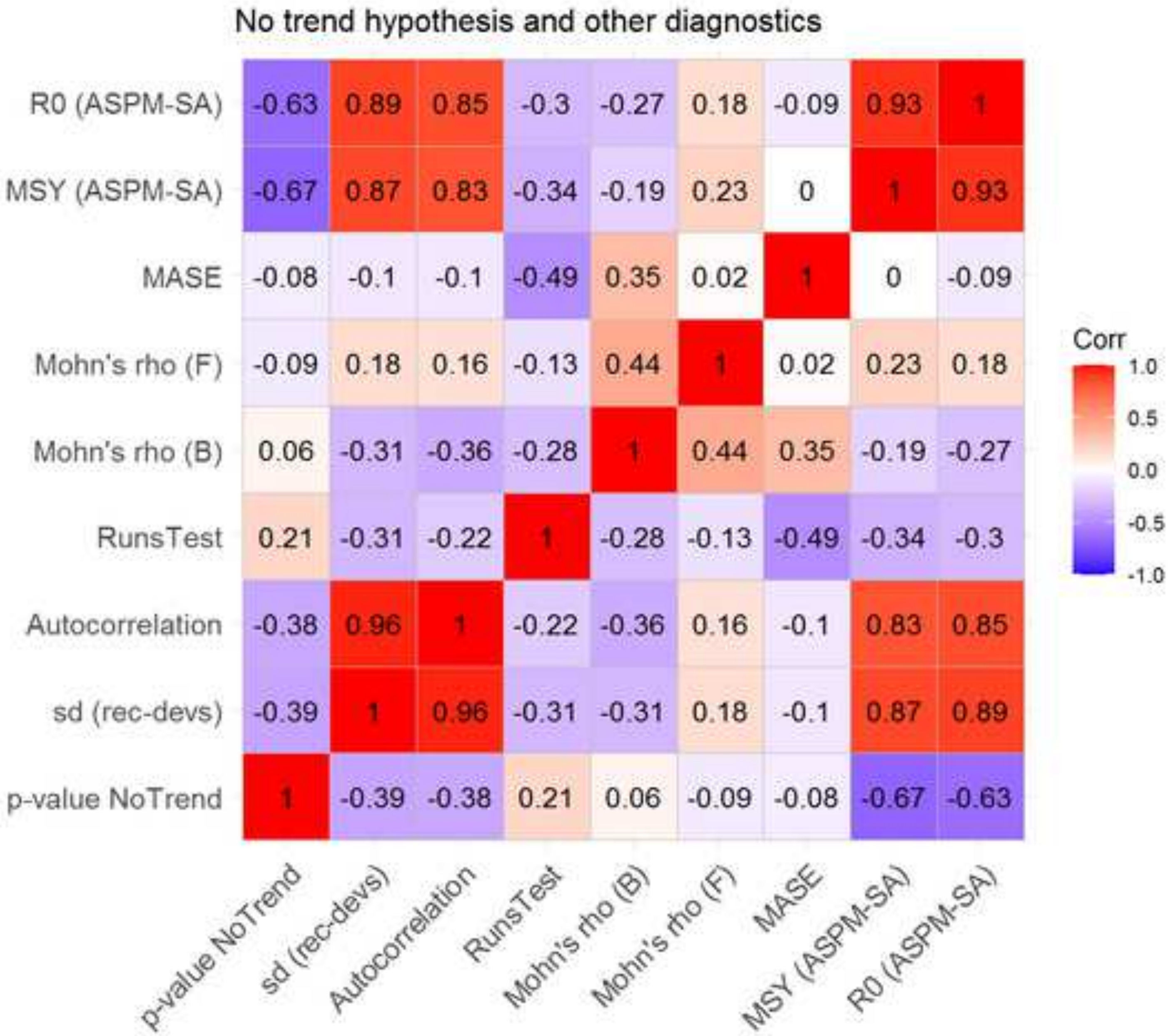
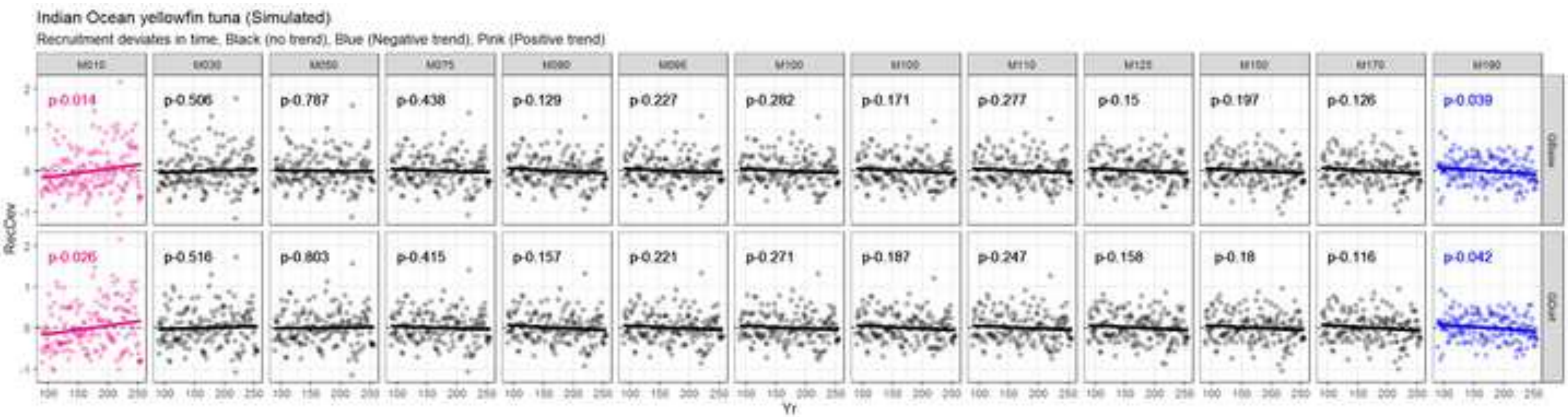

No trend hypothesis and other diagnostics
R0 (ASPM-SA)
MSY (ASPM-SA)
MASE
Mohn's rho (F)
Mohn's rho (B)
RunsTest
Autocorrelation
sd (rec-devs)
p-value NoTrend
-0.63 0.89 0.85 -0.3 -0.27 0.18 -0.09 0.93 1
-0.67 0.87 0.83 -0.34 -0.19 0.23 0 1 0.93
-0.08 -0.1 -0.1 -0.49 0.35 0.02 1 0 -0.09
-0.09 0.18 0.16 -0.13 0.44 1 0.02 0.23 0.18
0.06 -0.31 -0.36 -0.28 1 0.44 0.35 -0.19 -0.27
0.21 -0.31 -0.22 1 -0.28 -0.13 -0.49 -0.34 -0.3
-0.38 0.96 1 -0.22 -0.36 0.16 -0.1 0.83 0.85
-0.39 1 0.96 -0.31 -0.31 0.18 -0.1 0.87 0.89
1 -0.39 -0.38 0.21 0.06 -0.09 -0.08 -0.67 -0.63
p-value NoTrend
sd (rec-devs)
Autocorrelation
RunsTest
Mohn's rho (B)
Mohn's rho (F)
MASE
MSY (ASPM-SA)
R0 (ASPM-SA)
Corr
1.0
0.5
0.0
-0.5
-1.0

Figure 10


Indian Ocean yellowfin tuna (Simulated)
Recruitment deviates in time. Black (no trend), Blue (Negative trend), Pink (Positive trend)
p-0.014
p-0.506
p-0.787
p-0.438
p-0.129
p-0.227
p-0.282
p-0.171
p-0.277
p-0.15
p-0.197
p-0.126
p-0.039
p-0.026
p-0.516
p-0.803
p-0.415
p-0.157
p-0.221
p-0.271
p-0.187
p-0.247
p-0.158
p-0.18
p-0.116
p-0.042
RecDev
Yr

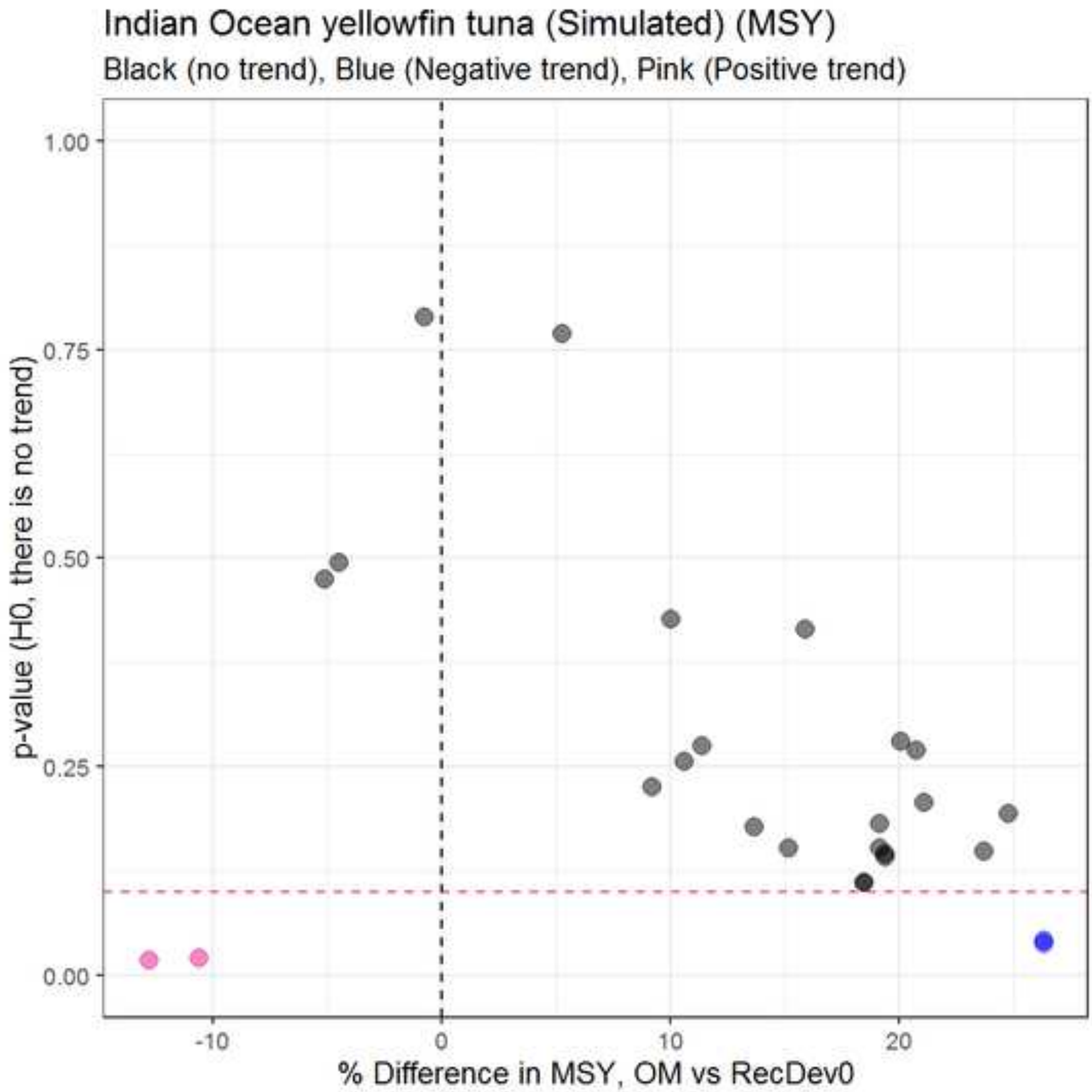

Indian Ocean yellowfin tuna (Simulated) (MSY)
Black (no trend), Blue (Negative trend), Pink (Positive trend)
p-value (H0, there is no trend)
1.00
0.75
0.50
0.25
0.00
-10
0
10
20
% Difference in MSY, OM vs RecDev0

Figure 12

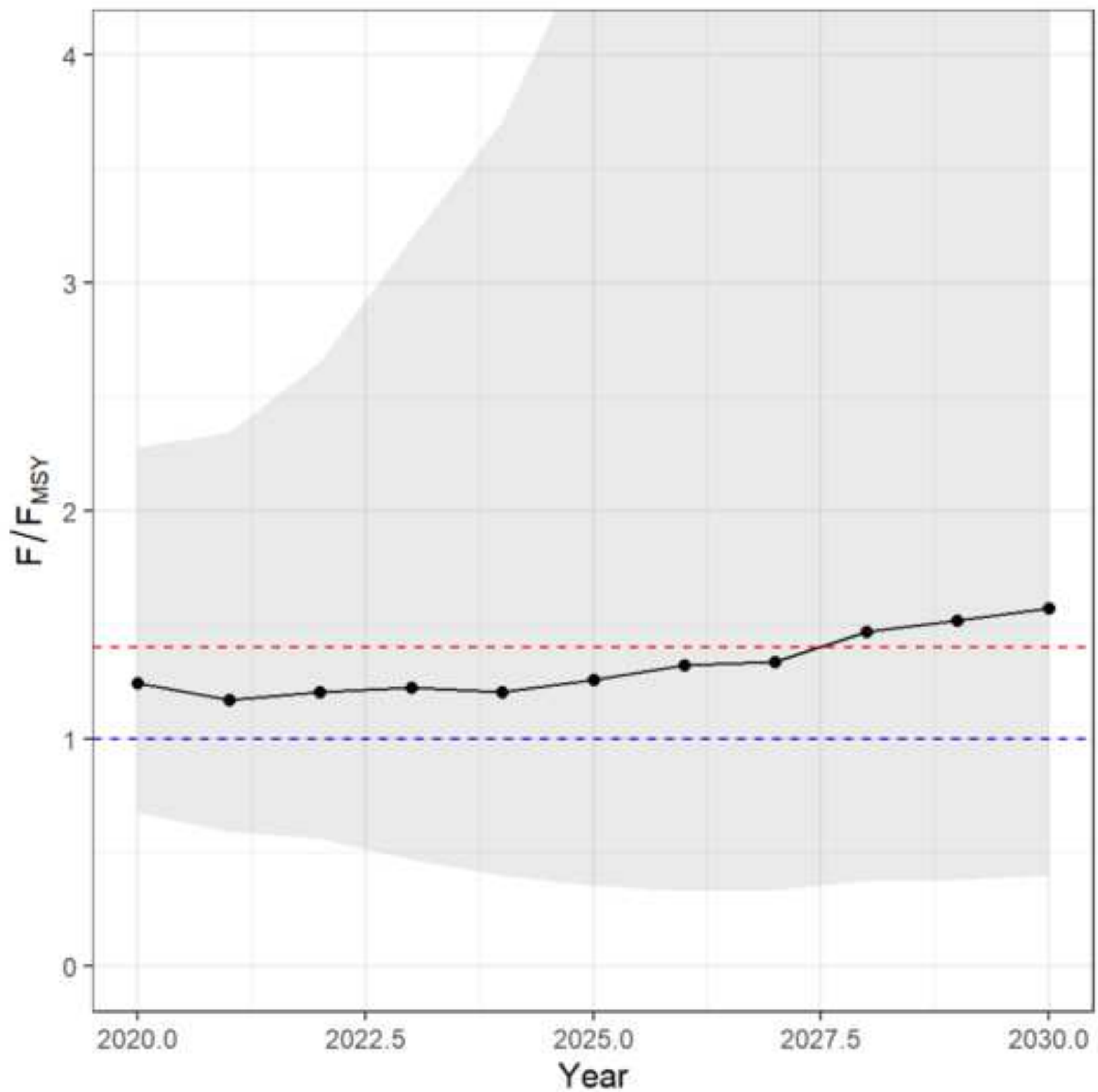


Figure captions

Figure 1. Catch history of tropical tunas (bigeye, yellowfin and skipjack) in the Atlantic Ocean (AO), Eastern Pacific Ocean (EPO), Indian Ocean (IO) and Western Central Pacific Ocean (WCPO).

Figure 2. Recruitment deviates for the 96 models of the Indian Ocean yellowfin stock assessment of 2021 (Fu et al., 2021). Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure 3. Estimated Maximum Sustainable Yield (MSY) for the 96 models of the Indian Ocean yellowfin stock assessment (Fu et al., 2021) and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure 4. Differences in % of MSY between the models of the Indian Ocean yellowfin stock assessment of 2021 (Fu et al., 2021) (SA) and their equivalent models with the recruitment deviates option deactivated (RecDev0) and, p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure 5. Differences in virgin recruitment (% of R0) between the models of the Indian Ocean yellowfin stock assessment of 2021 (Fu et al., 2021) (SA) and their equivalent models with the recruitment deviates option deactivated (RecDev0) and, and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure 6. Differences in the estimated relative biomass trajectory (B/Bmsy) between the models of the Indian Ocean yellowfin stock assessment of 2021 (Fu et al., 2021) (continuous line) and their equivalent models with the recruitment deviates option deactivated (RecDev0, dashed line). Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure 7. Differences in the estimated relative biomass (%B/Bmsy) between the models of the Indian Ocean yellowfin stock assessment of 2021 (Fu et al., 2021) (SA) and their equivalent models with the recruitment deviates option deactivated (RecDev0) and, and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure 8. Comparison of process error trends with standard model diagnostics. Red: Models with trends in recruitment deviates (p-value<0.1); green: Models without trends in recruitment deviates (p-value>0.1).

Figure 9. Correlation between the diagnostics developed in Carvalho et al (2021) and the p-value of the no-trend hypothesis for recruitment deviates. The diagnostics include convergence, likelihood, RMSE (Root mean square error), MASE (Mean average square error) and differences between the stock assessment estimates of MSY and R0 with their corresponding Age Structured Production Models (ASPM).

Figure 10. Recruitment deviates for the 26 models of the simulated Indian Ocean yellowfin operating model (Dunn and others 2020). Columns reflect % changes in the fixed natural mortality (e.g. M010 describes M as 10% of the M in the base case (M100)). Scenarios with a p-

value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure 11. Differences in % of MSY between the simulated Indian Ocean yellowfin operating model (Dunn and others 2020) (OM) and their equivalent models with the recruitment deviates option deactivated (RecDev0) and, p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure 12. Projection of fishing mortality from the 96 models of the assessment of Indian Ocean yellowfin (Fu and others 2021; Urtizberea and others 2021). Dotted black line represents the median trajectory, dashed blue line indicates F_{MSY} and dashed red line indicates the limit fishing mortality (F_{lim}=1.4xF_{MSY}).

Figure SI1A. Recruitment deviates for the 26 models of the Indian Ocean skipjack stock assessment of 2020. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI1B. Estimated Maximum Sustainable Yield (MSY) for the 26 models of the Indian Ocean skipjack stock assessment of 2020 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI2A. Recruitment deviates for the 18 models of the Indian Ocean bigeye stock assessment of 2019. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI2B. Estimated Maximum Sustainable Yield (MSY) for the 18 models of the Indian Ocean bigeye stock assessment of 2019 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI3A. Recruitment deviates for the 27 models of the Atlantic Ocean bigeye stock assessment of 2021. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI3B. Estimated Maximum Sustainable Yield (MSY) for the 27 models of the Atlantic Ocean bigeye stock assessment of 2021 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI4A. Recruitment deviates for the 4 models of the Atlantic Ocean yellowfin stock assessment of 2019. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI4B. Estimated Maximum Sustainable Yield (MSY) for the 4 models of the Atlantic Ocean yellowfin stock assessment of 2019 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI5A. Recruitment deviates for the 44 models of the East Pacific Ocean bigeye stock assessment of 2021. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI5B. Estimated Maximum Sustainable Yield (MSY) for the 44 models of the East Pacific Ocean bigeye stock assessment of 2021 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI6A. Recruitment deviates for the 48 models of the East Pacific Ocean yellowfin stock assessment of 2020. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI6B. Estimated Maximum Sustainable Yield (MSY) for the 48 models of the East Pacific Ocean yellowfin stock assessment of 2020 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI7A. Recruitment deviates for the 24 models of the West Central Pacific Ocean bigeye stock assessment of 2021. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI7B. Estimated Maximum Sustainable Yield (MSY) for the 24 models of the West Central Pacific Ocean bigeye stock assessment of 2021 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI8A. Recruitment deviates for the 72 models of the West Central Pacific Ocean yellowfin stock assessment of 2021. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI8B. Estimated Maximum Sustainable Yield (MSY) for the 72 models of the West Central Pacific Ocean yellowfin stock assessment of 2021 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Figure SI9A. Recruitment deviates for the 63 models of the West Central Pacific Ocean skipjack stock assessment of 2019. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend). Lines represent a linear regression to the recruitment deviates.

Figure SI9B. Estimated Maximum Sustainable Yield (MSY) for the 63 models of the West Central Pacific Ocean skipjack stock assessment of 2019 and p-value of the no-trend hypothesis. Scenarios with a p-value of the no-trend test lower than 0.1 are identified in purple (increasing trend) and blue (decreasing trend).

Dunn, A.; Hoyle, D.S.; Datta, S. Development of spatially explicit operating models for yellowfin tuna populations in the Indian Ocean. IOTC Working Party on Tropical Tunas 22(AS). IOTC-2020-WPT22(AS)-19.; 2020

Fu, D.; Urtizberea, A.; Cardinale, M.; Methot Jr, R.D.; Hoyle, D.S.; Merino, G. Preliminary Indian Ocean yellowfin tuna stock assessment 1950-2020 (Stock Synthesis). IOTC–2021–WPTT23–12. 2021

Urtizberea, A.; Fu, D.; Schirripa, M.; Methot Jr, R.D.; Cardinale, M.; Hoyle, D.S.; Merino, G. Indian Ocean yellowfin tuna SS3 model projections. IOTC–2021–SC24-INF08.; 2021