PRELIMINARY STOCK ASSESSMENT OF MEDITERRANEAN ALBACORE (*THUNNUS ALALUNGA*) USING THE BAYESIAN STATE-SPACE SURPLUS PRODUCTION MODEL *JABBA*

H. Winker[1], C. Pinto[2], A. Kimoto[3]

SUMMARY

Bayesian State-Space Surplus Production Models were fitted to Mediterranean albacore (Thunnus alalunga) catch and relative abundance indices using the 'JABBA' R package. This document presents details on the model diagnostics and stock status estimates for three preliminary scenarios, S1, S2 and S4. S1 was fitted to the three indices used in 2017, while S2 also included fits to four more historical indices and S4 included in addition a Spanish Tournament index. The prior assumptions and a Fox production function were kept consistent with the last assessment in 2017. We evaluated model plausibility using four objective model diagnostics: (1) model convergence, (2) fits to the data, (3) retrospective consistency and (4) prediction skill. Our results suggest that S2 and S4 represent the most plausible candidate models. These models reduced uncertainty about the absolute biomass estimates, and additional sensitivity runs indicated that these models were robust to alternative productivity and variance assumptions, while a Jackknife analysis revealed that either removing Balearic larval index or the Italian long-line index had the strong effects on the stock status estimates.

RÉSUMÉ

Les modèles de production excédentaire état-espace de type bayésien ont été ajustés aux données de capture et aux indices d'abondance relative du germon de la Méditerranée (Thunnus alalunga) en utilisant le progiciel « JABBA » R. Ce document présente des détails sur les diagnostics du modèle et les estimations de l'état des stocks pour trois scénarios préliminaires, S1, S2 et S4. S1 a été ajusté aux trois indices utilisés en 2017, tandis que S2 incluait également des ajustements à quatre indices historiques supplémentaires et S4 incluait, en outre, un indice des tournois espagnols. Les postulats a priori et une fonction de production de Fox correspondaient à la dernière évaluation de 2017. Nous avons évalué la plausibilité des modèles en utilisant quatre diagnostics de modèle objectifs : (1) la convergence des modèles, (2) les ajustements aux données, (3) la cohérence rétrospective et (4) la capacité de prédiction. Nos résultats suggèrent que S2 et S4 constituent les modèles potentiels les plus plausibles. Ces modèles réduisaient l'incertitude quant aux estimations de la biomasse absolue et des analyses de sensibilité supplémentaires indiquaient que ces modèles étaient robustes face aux autres postulats de productivité et de variance, alors qu'une analyse du Jackknife révélait que la suppression de l'indice larvaire des Baléares ou de l'indice de la palangre italienne avait de forts effets sur les estimations de l'état du stock.

RESUMEN

*Los modelos de producción excedente bayesianos de estado espacio se ajustaron a los índices de captura y CPUE del atún blanco (*Thunnus alalunga*) del Mediterráneo utilizando el paquete R de «JABBA». Este documento presenta detalles sobre los diagnósticos del modelo y las estimaciones del estado del stock para tres escenarios preliminares, S1, S2 y S4. S1 se ajustó a los tres índices utilizados en 2017, mientras que S2 incluía también ajustes a cuatro índices históricos más y S4 incluía además un índice de torneos español. Los supuestos previos y una función de producción Fox se mantuvieron coherentes con la última evaluación, en 2017. La plausibilidad de estos modelos se evaluó mediante cuatro diagnósticos objetivos del modelo: (1) la convergencia del modelo, (2) el ajuste a los datos, (3) la coherencia retrospectiva y (4) la*

[1] JRC, Joint Research Centre, Ispra, Italy. Email: henning.winker@ec.europa.eu
[2] University of Genova, Genova, Italy
[3] ICCAT Secretariat. Calle Corazón de Maria 8, Madrid Spain 28002.

capacidad de predicción. Nuestros resultados sugieren que S2 y S4 representan los modelos candidatos más plausibles. Estos modelos redujeron la incertidumbre acerca de las estimaciones de biomasa absoluta y ensayos de sensibilidad adicionales indicaron que estos modelos eran robustos ante supuestos de productividad y varianza alternativos, mientras que un análisis Jackknife reveló que, eliminar bien el índice larval balear o bien el índice de palangre italiano, tenía grandes efectos sobre el estado del stock.

KEYWORDS

Mediterranean Albacore, stock status, biomass dynamic model, Model diagnostics, hindcasting

1. Introduction

The albacore (*Thunnus alalunga*) is widely distributed in temperate and tropical waters of all oceans, including the Mediterranean Sea (Collette and Nauen, 1983). This species has habitat preferences for an epipelagic and mesopelagic realm and prefers cooler sea temperatures in comparison to other tropical tuna species. In the Atlantic Ocean, due to its wide distribution, albacore has been intensively exploited by a variety of fisheries. For management purposes, the International Commission for the Conservation of Atlantic Tunas (ICCAT) considers three stocks, a North Atlantic, a South Atlantic and a Mediterranean stock.

The most recent stock assessment for Mediterranean albacore was carried out by ICCAT in 2017 (ICCAT, 2017a). This assessment presented the stock status estimates for the assessment horizon 1980-2015, using an early version (BSPSP_ICCATv2.R) of the Bayesian state-space surplus production model JABBA (Just Another Bayesian Biomass Assessment; Winker *et al*., 2018a). The model was fitted to relatively short, joint CPUE index that started in 2001 and which was derived by applying state-space approach (Winker *et al*. 2018) to obtain an average estimate of the following three relative abundance indices: (i) standardized Spanish longline CPUE (ESP-LL), (ii) standardized Italian longline CPUE (ITA-LL) and (iii) the fisheries-independent Western Mediterranean larval density index based on ichthyoplankton surveys around the Balearic Islands (Larval-Index) that was considered as a proxy for spawning stock biomass.

The models were fitted using alternative parameterization of the generalized Pella-Tomlinson production function, of which the Fox model, which produced more optimistic than the Schaefer model, was selected for the base-case scenario to estimate the stock status. The stock status estimates were deemed highly uncertain as judged by the performance of model diagnostics, including jackknife of CPUE time, retrospective analysis and hindcast cross-validation (Kell *et al*. 2016) applied to the averaged CPUE index. In particular, the last CPUE data point was found to be highly influential and deemed suspect. As a result, the 2017 stock status advice was based on a model that was fitted to abundance indices through 2014, excluding the 2015 data points. This model indicated that the 2015 biomass was approximately at B_{MSY} level and that fishing mortality was below F_{MSY}. It was noted, however, that the alternative fits including the 2015 data points are not biologically implausible, highlighting the risk that the stock status could be more pessimistic with current catches substantially exceeding sustainable fishing level. Due to the high uncertainty about the stock status characterization, no future projections were conducted.

In this paper, we present the stock assessment results for Mediterranean albacore stock based on the Bayesian State-Space Surplus Production Model software, JABBA (Winker 2018a), using updated catch and standardized longline CPUE time series through 2019. The main improvement compared to the initial 2017 assessment include that the updated indices enabled fitting JABBA to the indices directly, without the need for an averaging approach for pre-processing indices. In general, index averaging introduces artificial smoothing, which can lead to overfitting, a lack of random residual pattern and model instability, and is therefore not recommended as a good practice in stock assessments. In addition, a number of standardized historical CPUE indices were made available, which were explored as part of this assessment update for Mediterranean albacore.

2. Material and Methods

This preliminary stock assessment is implemented using the Bayesian state-space surplus production model framework JABBA (Winker *et al*., 2018a). JABBA's inbuilt options include: automatic fitting of multiple CPUE time series and associated standard errors; estimating or fixing the process variance, optional estimation of additional observation variance for individual or grouped CPUE time series, and specifying a Fox, Schaefer or

Pella-Tomlinson production function by setting the inflection point B_{MSY}/K and converting this ratio into shape a parameter *m*. JABBA also provides a comprehensive toolbox to conduct model diagnostics to objectively evaluate the four model plausible criteria recommended in Carvalho et al. 2021: (1) model convergence (2) fit to the data, (3) model consistency (retrospective bias) and (4) prediction skill through hindcast cross-validation (Kell *et al*. 2017; 2021). A full JABBA model description, including formulation and state-space implementation, prior specification options and diagnostic tools is available in Winker *et al* (2018a). Following its first application to Mediterranean albacore in 2017 based on an early development version in 2017 (ICCAT, 2017a), JABBA has evolved into a fully documented, open-source R package (https://github.com/JABBAmodel/JABBA), which has been included in the ICCAT stock catalogue (https://github.com/ICCAT/software/wiki/2.8-JABBA). JABBA has subsequently been applied in a number of recent ICCAT stock assessments, including South Atlantic swordfish (ICCAT, 2017b; Winker et al., 2018), Atlantic shortfin mako shark stocks (south and north) (ICCAT, 2017c; Winker et al., 2020, 2017), Atlantic blue marlin (Mourato *et al.*, 2019), Atlantic bigeye tuna (Winker *et al*., 2019), Atlantic White marlin (Mourato *et al*., 2020), Atlantic yellowfin tuna (Sant'Ana *et al.*, 2020), Mediterranean swordfish (Winker *et al*. 2020a) and South Atlantic albacore (Winker *et al.* 2020b).

2.1 Fishery data

Fishery catch data for Mediterranean albacore were made available by ICCAT Secretariat for the period 1950-2019 (**Figure 1**). For this assessment, the time series was subset to 1980-2019, consistent with the 2017 assessment. In 2021, the standardized CPUE time series used in the 2017 final model were updated. For Scenario 1 (S1), the fisheries dependent CPUEs for the Spanish longline (ESP-LL) and for the Italian longline (ITA-LL), and the fisheries-independent 'larval index' from Balearic region (Balearic-Larval) were used.

In addition, we considered an alternative scenario (Scenario 2), where JABBA was fitted to additional four historical indices. These included a historical nominal CPUE from Italian drifting longlines fishery in the southern Adriatic Sea from FAO-MiPAF document (Marano *et al*., 2005), which covered early years 1984 to 2000, but with no data for 1988-89 and 1996-97, and three new indices based on data collected in the frames of past European and national projects. The data were obtained from the MEDPEL project dataset, which was based both on landings and observation on board information, from available historical time series. The three relative abundance indices comprised a standardized Italian long-line CPUE index for the Ionian Sea from 1995 to 2003 (LL-Ionian) collected by the University of Bari, a standardized (non-target) longline index for Ligurian Sea from 1991 to 2009 (Ligurian_Bycatch) collected by the University of Genoa, a standardized Southern Mediterranean longline index (MedSouth_LL) based on data provided by the Aquastudio Research Institute for the period 2004 to 2009, covering five southern Mediterranean areas: Libya, North and South Ionian Sea, South of Sicily, South Thyrrenian Sea (SCRS/2021/115). The indices used in this assessment were provided in mass per unit effort and assumed to be proportional to biomass.

Furthermore, a new Spanish Tournament index was provided that covered a relatively long period from 2005-2019, but only represented limited spatial and seasonal coverages (SCRS/2021/103). This index was fitted together will all indices in Scenario 2, in an alternative scenario (Scenario 4).

2.2 Model specifications and sensitivity runs

Initially, three candidate model scenarios were considered:

- S1: a reference scenario, fitted to the updated three indices (ESP-LL, Larval-Index and ITA-LL) used in the 2017 assessment

- S2: an "extended" model, fitted to the seven available candidate indices (Spanish, Italian, Ionian, Ligurian, Med-South, and historical Italian longline indices, and western Mediterranean larval index (providing information on the trends of the spawning biomass)).

- S4: a "full" model, fitted to all eight available candidate indices (S2 + Spanish Tournament index).

An additional, intermediate Scenario S3 represented a variation of S2 by excluding the fisheries-independent larval index. The results of this scenario are presented in the Supplementary Material.

For the unfished equilibrium biomass *K*, we used default settings of the JABBA R package in the form of vaguely informative lognormal prior with a large CV of 100% and a central value that corresponds to eight times

the maximum total catch and is consistent with other platforms, such as Catch-MSY (Martell and Froese, 2013) or the initial value for K in SpiCt (Pederson and Berg 2017). We assumed a Fox production function (B_{MSY}/K = 0.37), a lognormal prior distribution for r with mean of log(0.153) and a standard deviation for log(r) of 0.457 and initial beta prior for the relative biomass ($\varphi = B_{1980}/K$) with mean = 0.85 and CV of 10%, which is consistent with the 2017 base-case model (ICCAT, 2017). All catchability parameters were formulated as uninformative uniform priors, while the process error of log(B_y) in year y was estimated "freely" by the model using an uninformative inverse-gamma distribution with both scaling parameters set at 0.001. Initial trials indicated that it was challenging to reliably estimate observation errors using an additional variance approach for model internal weighting (e.g. *Winker et al.* 2020), because several of the indices covered only a few years and were subject to missing values and irregular spacing. To address this, we considered a fixed observation error approach by assuming a standard error for log(CPUE) of 0.25 for the three most recent indices, while slightly down-weighting the historical indices and the Spanish Tournament CPUE by setting the standard error to 0.35. To explore sensitivity, additional robust tests were conducted for alternative observation and process error variance settings (**Table 1**), as is common practice in many age-structured tuna assessments with Stock Synthesis. The robustness tests also included an alternative assumption of a 50% higher prior mean for r (**Table 1**). To examine the sensitivity of the assessment results to the inclusion of individual CPUE indices, we iteratively re-fitted the models while excluding one index at the time and refitting the model (i.e. Jackknife index analysis).

2.3 Model diagnostics

The evaluation model diagnostics follows the principles in Carvalho *et al.* (2021), who recommended to objectively evaluate the base-case candidate model based on the following four model plausible criteria: (1) model convergence (2) fit to the data, (3) model consistency (retrospective pattern) and (4) prediction skill through hindcast cross-validation (Kell *et al.* 2016; 2021).

JABBA is implemented in R (R Development Core Team, https://www.r-project.org/) with JAGS interface (Plummer, 2003) to estimate the Bayesian posterior distributions of all quantities of interest by means of a Markov Chains Monte Carlo (MCMC) simulation. In this study, three MCMC chains were used. Each model was run for 30,000 iterations, sampled with a burn-in period of 5,000 for each chain and thinning rate of five iterations. Basic diagnostics of model convergence included visualization of the MCMC chains using MCMC trace-plots as well as Heidelberger and Welch (Heidelberger and Welch, 1992) and Geweke (1992) and Gelman and Rubin (1992) diagnostics as implemented in the coda package (Plummer *et al*., 2006).

To evaluate the JABBA fit to the abundance index data, the model predicted values were compared to the observed indices. JABBA-residual plots were used to examine (1) color-coded lognormal residuals of observed versus predicted CPUE indices by fleet together with (2) boxplots indicating the median and quantiles of all residuals available for any given year; the area of each box indicates the strength of the discrepancy between CPUE series (larger box means higher degree of conflicting information) and (3) a loess smoother through all residuals which highlights systematically auto-correlated residual patterns to evaluate the randomness of model residuals. In addition it depicts the root-mean-squared-error (RMSE) as a goodness-of-fit statistic. We conducted run tests to evaluate the randomness of residuals (Carvalho *et al.*, 2017). The runs test diagnostic was applied to residuals of the CPUE fit on log-scale using the function runs.test in the R package tseries, considering the 1-sided p-value of the Wald-Wolfowitz runs test (Carvalho *et al*. 2021). The runs test results can be visualized within JABBA using a specifically designed plot function that illustrates which time series passed or failed the runs test and highlights individual time-series data points that fall outside the three-sigma limits (e.g. Anhøj and Olesen, 2014).

To check for model consistency with respect to the stock status estimates, we also performed a retrospective analysis by removing one year of data at a time sequentially (n = 5), refitting the model and comparing quantities of interest (i.e. biomass, fishing mortality, B/B_{MSY}, F/F_{MSY}, B/B_0 and MSY) to the reference model that is fitted to full time series. To compare the bias between the models, we computed Mohn's (Mohn, 1999) rho (ρ) statistic and specifically the commonly used formulation Hurtado-Ferro *et al*. (2015).

Although the above model diagnostics are important to evaluate model convergence, the fit to the data and retrospective consistency, providing scientific advice should also involve checking that the model has prediction skill of future states under alternative management scenarios (Carvalho *et al.* 2021). To validate a model's prediction skill requires that the system be observable and measurable (Kell *et al.* 2021). Therefore, we applied a hindcasting cross-validation (HCXval) technique (Kell *et al.* 2016), where observations are compared to their predicted future values. HCxval is a form of cross-validation where, like retrospective analysis, recent data are removed, and the model refitted with the remaining data, but HCXval involves the additional steps of projecting

ahead over the missing years and then cross-validating these forecasts against observations to assess the model's prediction skill. A robust statistic for evaluating prediction skill is the Mean Absolute Scaled Error (MASE), which scales the mean absolute error of prediction residuals to a naïve baseline prediction, where a 'prediction' is said to have 'skill' if it improves the model forecast when compared to the naïve baseline (Kell *et al*. 2021). A widely used baseline forecast for time series is the 'persistence algorithm' that takes the value at the previous time step to predict the expected outcome at the next time step as a naïve in-sample prediction, e.g., tomorrow's weather will be the same as today's. The MASE score scales the mean absolute error of the prediction residuals to the mean absolute error of a naïve in-sample prediction. A MASE score higher than one can then be interpreted such that the average model forecasts are no better than a random walk. Conversely, a MASE score of 0.5 indicates that the model forecasts twice as accurately as a naïve baseline prediction; thus, the model has prediction skill.

3. Results and Discussion

The MCMC convergence tests by Heidelberger and Welch (Heidelberger and Welch, 1992) and Geweke (1992) and Gelman and Rubin (1992) were passed by all estimable key parameters for all scenarios S1, S2 and S4. Adequate convergence of the MCMC chains was also corroborated by visual inspection of trace plots (results available on request), which showed good mixing in general (i.e., moving around the parameter space).

The model fits to each of the relative abundance indices, comprising the standardized CPUE indices ESP-LL and ITA-LL and the fisheries-independent Larval-Index, are shown for the scenarios S1, S2 and S4 in **Figures 3 - 5**, respectively. The fits to the four historical indices, included in S2 and S4 are presented in **Figures 6 - 7**. All three models appeared to fit the abundance trends reasonably well, and run tests conducted on the log-residuals provided no evidence to reject the hypothesis of randomly distributed residual patterns for the three indices used in S1, S2 and S4 (**Figures 3 - 5**), as well as, the four historical indices used in S2 and S4 (**Figures 6 - 7**).

The overall goodness-of-fit indicated a fairly low precision of the fits from the reference run S1 (RMSE = 50.1%) and S4 (RMSE = 50.1%), which were slightly better for S2 (RMSE = 44.1%**)**, which incorporated the historical indices but not the Spanish Tournament CPUE index (**Figure 8**). The residual patterns of the last years 2018-2019 indicated some conflicts between positive residuals from the ITA-LL index and negative 2019 residuals for the ESP-LL and Balearic-Larval indices (**Figure 8**). This still resulted in an, on average, positive trend in the residual pattern for most recent years, which is probably due to the sharp increase in CPUE in the LL-ITA in 2018-2019. Notably, the LL-ITA index attains additional influence because of the lack of estimates for 2018 in both the ESP-LL and Balearic-Larval (**Figure 8**). The estimated process error deviations had a similar trend for all scenarios, showing particularly strong variations in most recent years. The process deviations for the terminal year are close to zero and therefore to average expectation (**Figure 8**).

The medians of marginal posteriors for r were estimated to be lower for S1 at 0.134 than for S2 at 0.174 and S4 at 0.186 (**Table 2**). The scale of absolute biomass estimates for K and B_{MSY} was the lowest for S4 (**Table 2**), which was also associated with an improved, smaller posterior to prior ratio of variance for K (PPRV = 0.094) compared to S1 (PPRV = 0.161) and S2 (PPRV = 0.101) as shown in **Figures 9-11.** This indicates that the historical indices may hold information about the total biomass to effectively update the posterior of K given its relatively vague prior. Estimates of the median MSY were very similar among all scenarios, and so were the posterior medians of B_{2019}/B_{MSY} and F_{2019}/F_{MSY} (**Table 2**). The most notable difference among the three scenarios was therefore the reduced uncertainty about the total biomass scale for S2 and S4, as result of the inclusion of the historical indices in the model (**Figure 12**). The addition of Spanish Tournament index to S2 reduced uncertainty about the estimates in the 2000's, however both scenarios S2 and S4 showed similar results for the recent years.

The sensitivity runs indicated that the 'extended' model S2 was largely robust to alternative assumptions about r and the observation and process errors (**Figure 13**). Assuming a higher prior mean for r showed surprisingly little effect on the current stock status in terms of B_{2019}/B_{MSY} and F_{2019}/F_{MSY}, and only resulted in a slight increase on the MSY estimate as indicated by the height of the production function in **Figure 13**. The only effects on stock status estimates were observed in the form of higher F_{2019}/F_{MSY} estimates for the runs with a fixed process error (fixed.pe) and larger observation error (high.oe), which also resulted in lower MSY estimates, while no effect on B_{2019}/B_{MSY} was observed (**Figure 13**). Both settings resulted in similar properties, because fixing the process error at 0.1 and thus lower than the estimate (**Table 3**), results in larger observation errors to fit to the data, whereas admitting higher observation error *a priori*, results in lower process error.

The Jackknife index analysis was applied to S4 by removing one index at the time and showed that removing the 'Larval-Index' was most influential with regards to stock status trajectories and resulted in current estimate of B_{2019} being approximately at B_{MSY} and F_{2019} below F_{MSY} (**Figure 14**). Thus excluding the Larval-Index would change the perception of the stock to a more optimistic status that is also associated with a higher MSY (**Figure 14**). The second most influential effect was the removal of the ITA-LL, resulting in a more pessimistic estimate of B_{2019}/B_{MSY} and a much higher F_{2019} that corresponded to two times that of F_{MSY} (**Figure 14)**. The relatively strong influence of the ITA-LL index is probably related to the strong increase in CPUE for the 2018-2019, which is contrary to the recent trends in the ESP-LL CPUE and the Larval-Index.

The retrospective analysis applied over a horizon of five years to S1, S2 and S4 (**Figures 15-17**) revealed a much improved retrospective pattern for S2 and S4 (**Figures 16** and **17**), which included the additional fits to the historical indices. For S1, the Mohn's ρ values fell outside the acceptable thresholds of -0.15 and 0.2 for longer lived species (Huerto-Ferro *et al.*, 2015) for the three quantities *B*, *F* and F/F_{MSY} (**Table 3**; **Figure 15**). In particular, *F* and F/F_{MSY} were associated with large retrospective bias estimates Mohn's ρ = 0.32 and 0.41, respectively (**Table 3**). For S2 and S4 models, the retrospective bias was reduced for quantities, with Mohn's ρ values falling within acceptable thresholds for all evaluated quantities (*B*, *F*, B/B_{MSY}, *B/K* and MSY), with F/F_{MSY} for S2 being the only exception that a slightly higher Mohn's ρ of 0.24 (**Table 3, Figures 16** and **17**).

Hindcasting cross-validation results suggested that the ESP-LL and ITA-LL have prediction skill for all S1, S2 and S4 scenarios as judged by the MASE scores < 1 (**Figure 18**), which provides a means to validate that short-term forecast are consistent with the 'future' observations that were unknown to the model (Kell *et al.* 2021). The MASE scores for the Balearic-Larval index were close to one and the prediction residuals appeared relatively small. This suggests that forecasts are also relatively precise for the Balearic-Larval index. Generally, MASE scores were marginally better in S2.

Due to the strong influence of the Balearic-Larval index on the stock status estimates (**Figure 14, Figure S1)**, we also evaluated an additional scenario S3, which was based on S2 but excluded the Balearic-Larval index (see Supplement **Figures S1-S5**). Although the S3 model appeared to fit the data well, with no evidence for an undesirable residual pattern (**Figure S3**), it resulted in a concerning increase of retrospective bias for all evaluated key quantities, but MSY (**Figure S4)**, while prediction skill for two retained indices ESP-LL and ITA-LL could be improved and had even slightly decreased (**Figure S5)**.

The surplus production phase plots were similar for the S1, S2 and S4 scenarios, suggesting that the stock has been in an overfished state since the late 2000s, with current fishing pressure remaining unsustainable (**Figure 15**). Catches exceeded MSY for several years while biomass remained above B_{MSY} and from 2006 onwards these high catches were longer sustainable. The probability that the stock is overfished and that overfishing is currently occurring is estimated to be 65.6% for S1 and 74.3% for S2 (**Figure 15**). The current 2019 catch is slightly below the expected average surplus production, so, while biomass may slowly increase under current catch levels, the current fishing mortality is high for rebuilding the stock to biomass levels that are capable of producing MSY (**Figure 15**).

In line with the recommendations by the 2021 Working Group of Stock Assessment Methods (WGSAM), we evaluated the plausibility of two alternative JABBA model scenarios for the Mediterranean albacore based on best practice in using model diagnostics (Carvalho *et al*. 2021). These criteria are: (1) model convergence, (2) fits to the data, (3) model consistency (e.g. retrospective bias) and (4) prediction skill. Our results suggest that S2 and S4 represent the most plausible candidate model for the Mediterranean albacore stock status. Specifically, S2 and S4 converged adequately, provided a robust fit to the data, is largely consistent retrospectively and indicated that two of three abundance indices that covered the most recent years have desirable prediction skill. The most notable improvement compared to the alternative candidate models is a substantially reduced retrospective bias and reduced uncertainty about the absolute biomass estimates.

References

Anhøj, J., Olesen, A.V., 2014. Run charts revisited: A simulation study of run chart rules for detection of non-random variation in health care processes. PLoS One 9, 1–13. https://doi.org/10.1371/journal.pone.0113825

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? Fish. Res. 192, 28–40. https://doi.org/10.1016/j.fishres.2016.09.018

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., and Methot R. D.2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research Volume 204. https://doi.org/10.1016/j.fishres.2021.105959

Collette, B. B. and C. E. Nauen. 1983. FAO species catalogue, vol. 2, Scombrids of the world. FAO. Fisheries synopsis 125 (2): 137 p.

Gelman, A., Rubin, D.B., 1992. lnference from Iterative Simulation Using Multiple Sequences. Stat. Sci. 7, 457–472. https://doi.org/10.2307/2246093

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments., in: Berger, J.O., Bernardo, J.M., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting. Clarendon Press, Oxford, pp. 169–193.

Heidelberger, P., Welch, P.D., 1992. Simulation run length control in the presence of an initial transient. Oper. Res. 31, 1109–1144. https://doi.org/10.1287/opre.31.6.1109

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, in: ICES Journal of Marine Science. pp. 99–110. https://doi.org/10.1093/icesjms/fsu198

ICCAT, 2017a. Report of the 2017 ICCAT albacore species group intersessional meeting (including assessment of Mediterranean albacore). Collect. Vol. Sci. Pap. ICCAT 74, 45.

ICCAT, 2017b. Report of the 2017 ICCAT Atlantic swordfish stock assessment session. Collect. Vol. Sci. Pap. ICCAT 74, 841–967.

ICCAT, 2017c. Report of the 2017 ICCAT shortfin mako assessment meeting. Collect. Vol. Sci. Pap. ICCAT 74, 1465–1561.

Kell, L. T., Kimoto, A., and Kitakado, T. 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries Research, 183: 119–127. http://www.sciencedirect.com/science/article/pii/S0165783616301540.

Kell, L. T., Sharma, R., Kitakado, T., Mosqueira, I., Winker, H., Cardinale, M., and Fu, D. 2021. Validation of stock assessment models using prediction skill: Is it me or my model talking? ICES Journal of Marine Science: https://doi.org/10.1093/icesjms/fsab104.

Martell, S., Froese, R., 2013. A simple method for estimating MSY from catch and resilience 504–514. https://doi.org/10.1111/j.1467-2979.2012.00485.x

Mohn, R., 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56, 473–488. https://doi.org/10.1006/jmsc.1999.0481

Mourato, B.L., Winker, H., Carvalho, F., Kimoto, A., Ortiz, M., 2020. Developing of Bayesian State-Space Surplus Production JABBA for Assessing Atlantic white marlin (*Kajikia albida*) stock. Col. Vol. Sci. Pap. ICCAT 76, 235–254.

Mourato, B.L., Winker, H., Carvalho, F., Ortiz, M., 2019. Stock Assessment of blue marlin (*Makaira nigricans*) using a Bayesian State-Space Surplus Production Model JABBA. Collect. Vol. Sci. Pap. ICCAT 75, 1003–1025.

Plummer, M., 2003. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling, 3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria.

Plummer, M., Nicky Best, Cowles, K., Vines, K., 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News 6, 7–11.

Sant'Ana, R., Mourato, B., Kimoto, A., Walter, J., Winker, H., 2020. Atlantic Yellowfin tuna stock assessment: An Implementation of a Bayesian State-Space Surplus Production Model using JABBA. Col. Vol. Sci. Pap. ICCAT 76, 699–724.

Winker, H., Carvalho, F., Kapur, M., 2018. JABBA: Just Another Bayesian Biomass Assessment. Fish. Res. 204, 275–288. https://doi.org/http://doi.org/10.1016/j.fishres.2018.03.01

Winker, H., Carvalho, F., Kerwath, S., 2020. Age-structured biomass dynamics of north Atlantic shortfin mako with implications for the interpretation of surplus production models. Col. Vol. Sci. Pap. ICCAT 76, 316–336.

Winker, H., Carvalho, F., Sharma, R., Parker, D., Kerwath, S., 2017. Initial results for North and South Atlantic shortfin mako (*Isurus oxyrinchus*) stock assessments using the Bayesian surplus production model JABBA and the catch-resilience method CMSY 74, 1836–1866.

Winker, H., Kerwath, S.E., Merino, G., Ortiz, M., 2019. Bayesian State-Space Surplus Production Model JABBA of Atlantic bigeye tuna (*Thunnus obesus*) stock. Col. Vol. Sci. Pap. ICCAT 75, 2129–2168.

Winker, H., Mourato, B., Parker, D., R, S. A., Kimoto, A., and Ortiz, M. 2020a. Preliminary stock assessment of South Atlantic albacore tuna (*Thunnus alalunga*) using the Bayesian state-space surplus production model JABBA. Collective Volume of Scientific Papers ICCAT, 77: 352–376.

Winker, H., Kimoto, A., Mourato, B. L., Tserpes, G., and Ortiz, M. 2020b. Development of Bayesian state-space surplus production model JABBA for assessing the Mediterranean swordfish (*Xiphias gladius*) stock. ICCAT Coll.Vol.Sci.Pap., 77: 508–536.

Table 1. Model specifications of alternative productivity and variance parameters used in the sensitivity analysis for Scenario1. *r*-prior: mean value of the prior for *r*, σ_{obs_cur}: observation error assumption for the three current indices, σ_{obs_hist}: observation error assumption for the four historical indices.

Run	*r*-prior	σ_{obs_cur}	σ_{obs_hist}	σ_{proc}	Description
high-r	0.230	0.25	0.35	est	Increased *r* prior mean by a factor of 1.5
fixe.pe	0.153	0.25	0.35	0.1	Fixed process error
low.oe	0.153	0.20	0.30	est	lower fixed observation error
high.oe	0.153	0.30	0.40	est	higher fixed observation error
equal.oe	0.153	0.25	0.25	est	equal observation for current and historical indices

Table 2. Summary of posterior quantiles presented in the form of marginal posterior medians and associated the 95% credibility intervals of parameters for the Bayesian state-space surplus production models for Mediterranean albacore Scenarios 1,2, and 4 model runs.

	S1			**S2**			**S4**		
Estimates	Median	2.50%	97.50%	Median	2.50%	97.50%	Median	2.50%	97.50%
K	75670.0	41077.9	169938.2	56453.4	33409.6	107606.6	53240.7	31551.1	99528.3
r	0.134	0.063	0.255	0.173	0.085	0.319	0.186	0.093	0.339
ψ *(psi)*	0.858	0.638	0.970	0.851	0.629	0.969	0.852	0.623	0.970
σ_{proc}	0.190	0.141	0.211	0.189	0.139	0.211	0.032	0.013	0.044
F_{MSY}	0.132	0.063	0.252	0.171	0.084	0.315	0.184	0.091	0.335
B_{MSY}	28003.6	15201.9	62890.0	20892.0	12364.1	39822.6	19703.1	11676.3	36833.0
MSY	3742.1	2231.3	6384.8	3641.4	2319.5	5228.6	3653.9	2445.9	5090.1
B_{1980}/K	0.825	0.534	1.183	0.8	0.508	1.168	0.805	0.527	1.153
B_{2019}/K	0.214	0.106	0.39	0.207	0.113	0.368	0.211	0.119	0.372
$B_{2019}/B_{\mathrm{MSY}}$	0.580	0.286	1.054	0.560	0.307	0.994	0.57	0.322	1.004
$F_{2019}/F_{\mathrm{MSY}}$	1.182	0.49	2.488	1.246	0.621	2.386	1.213	0.618	2.175

Table 3. Summary Mohn's rho statistic for the Scenarios S1, S2, and S4 models, computed for a retrospective evaluation period of five years. The larger the threshold the stronger is the retrospective bias.

	Stock Quantity					
Scenario	B	F	B/B_{MSY}	F/F_{MSY}	B/K	MSY
S1	-0.168	0.233	-0.106	0.374	-0.106	0.162
S2	-0.009	0.110	0.162	0.259	-0.062	-0.127
S4	-0.034	0.044	-0.027	0.172	-0.027	-0.095

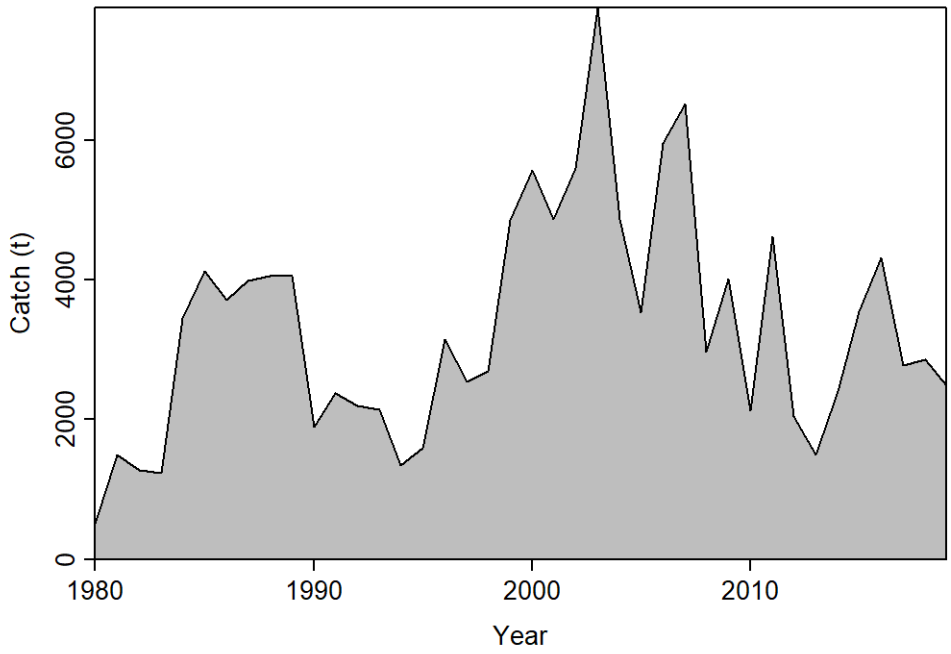

Figure 1 - Catch times series 1980-2019 in metric tons (t) for Mediterranean albacore.

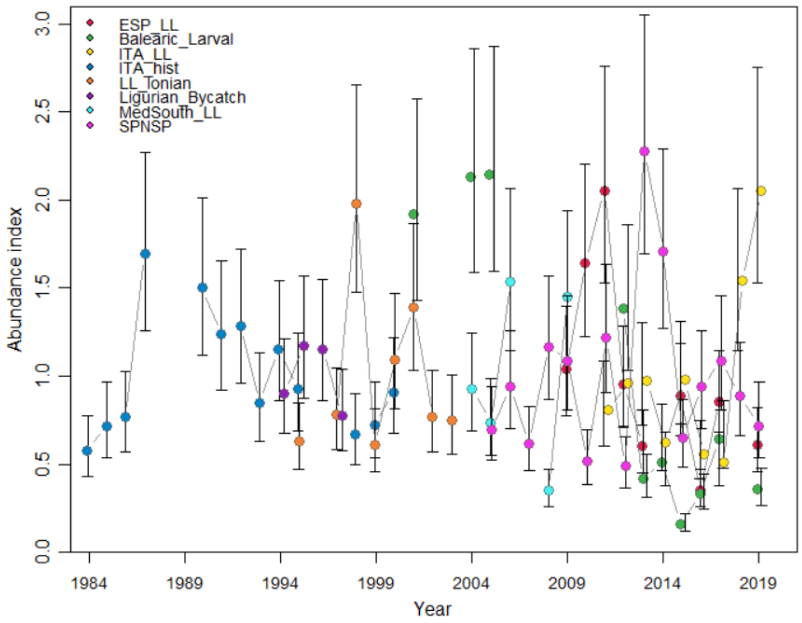

Figure 2 - Time-series of four recent and four historical relative abundance indices considered in the preliminary JABBA stock assessment for Mediterranean albacore. Error bars represent the 95% Confidence Interval corresponding to the assumed standard errors on logarithm of the annual index estimates. SPNSP: Spanish Tournament Index

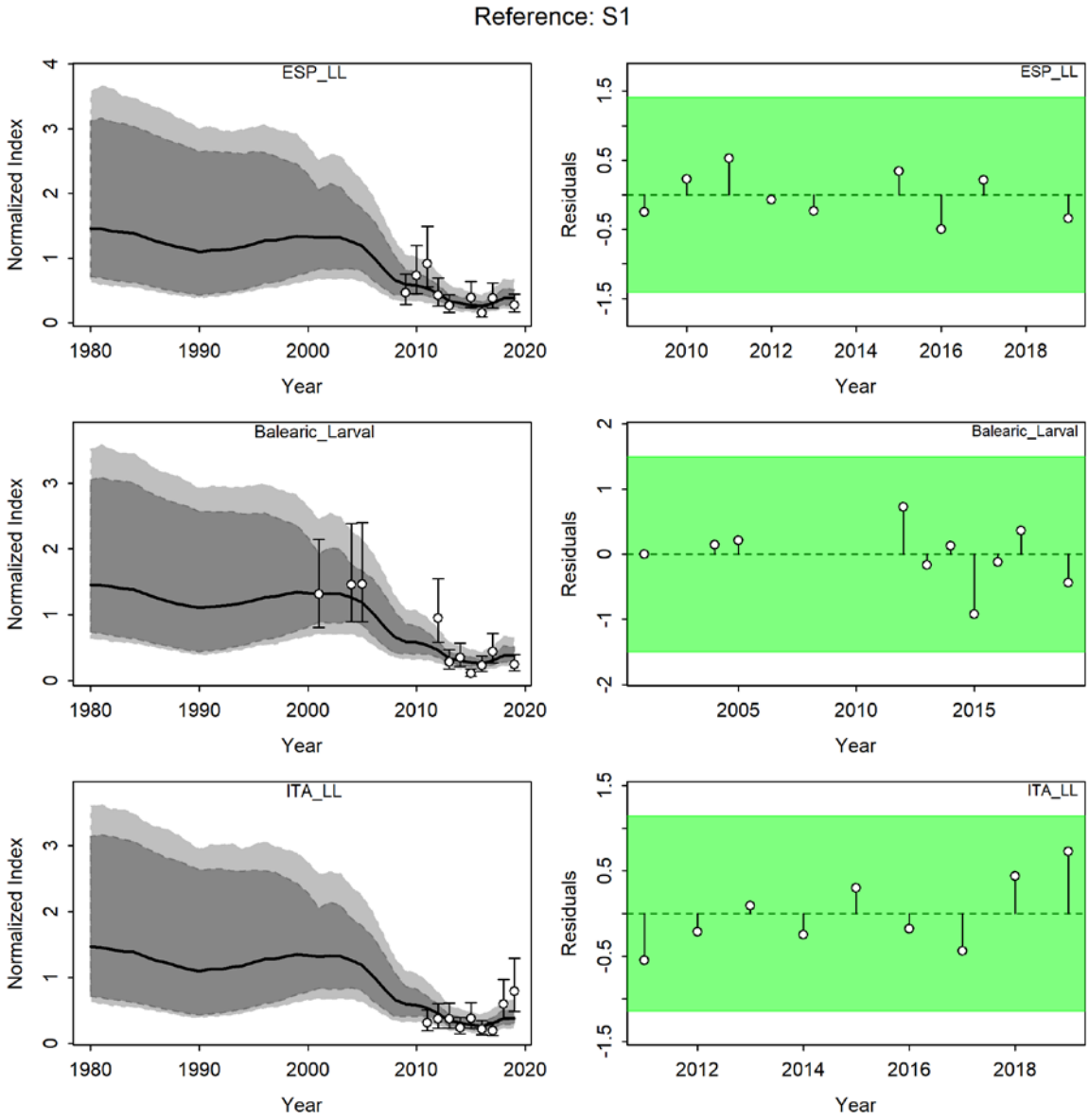

Figure 3. Time-series of observed (circle) with error 95% CIs (error bars) and predicted (solid line) CPUE (left) and Runs tests to quantitatively evaluate the randomness of the time series of CPUE residuals (right) for the Mediterranean albacore reference scenario (S1). On the left panel, the Dark shaded grey areas show 95% credibility intervals of the expected mean CPUE and light shaded grey areas denote the 95% posterior predictive distribution intervals. On the right panel, green areas indicate no evidence of lack of randomness of time-series residuals (p>0.05) while red panels (not shown here) indicate the opposite. The inner shaded area shows three standard errors from the overall mean and red circles identify a specific year with residuals greater than this threshold value (3- sigma rule).

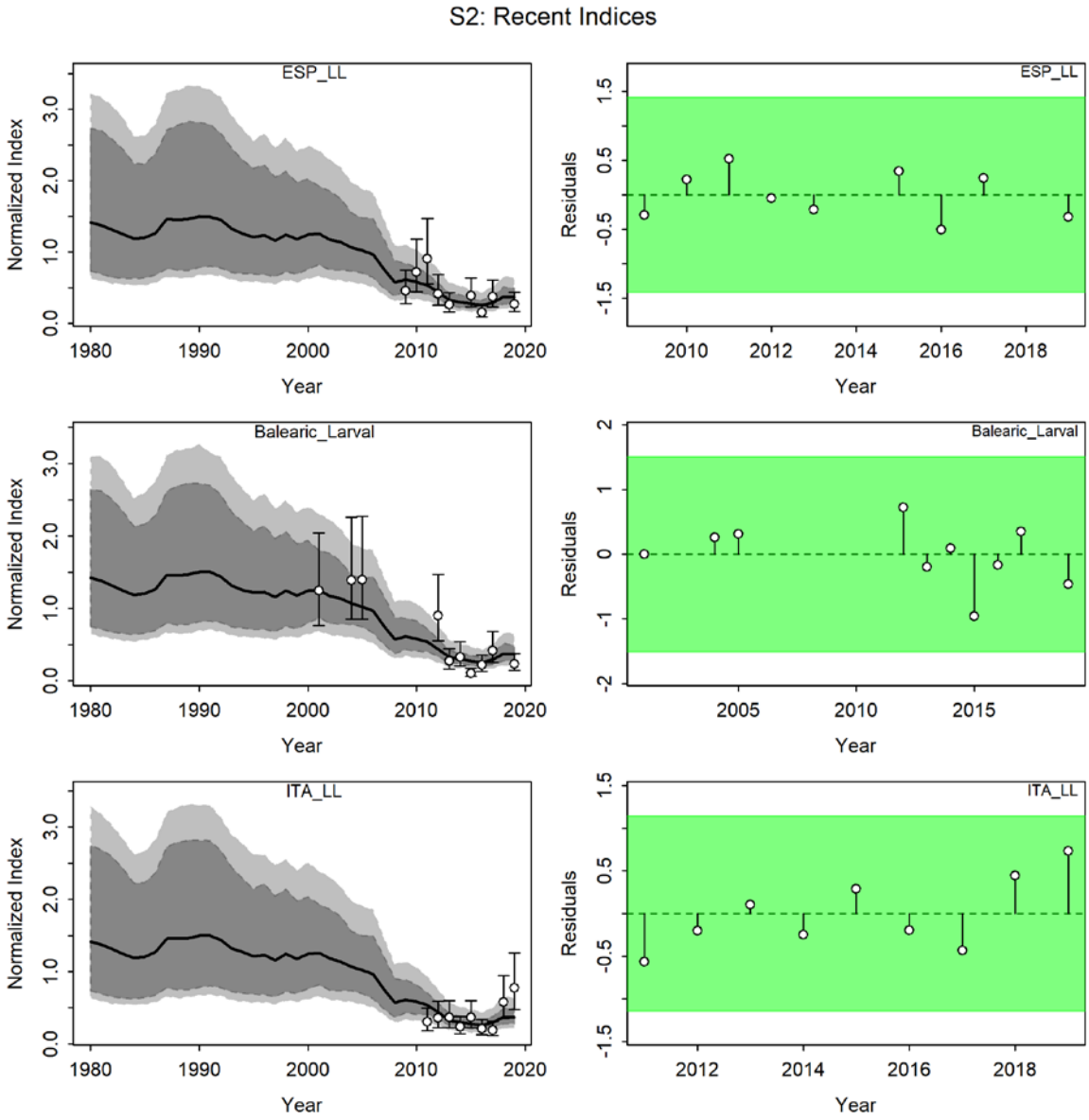

Figure 4. Time-series of observed (circle) with error 95% CIs (error bars) and predicted (solid line) CPUE (left) and Runs tests to quantitatively evaluate the randomness of the time series of CPUE residuals (right) for Mediterranean albacore "extended" model (S2), shown for the most recent indices that were also used in the reference run S1. On the left panel, the Dark shaded grey areas show 95% credibility intervals of the expected mean CPUE and light shaded grey areas denote the 95% posterior predictive distribution intervals. On the right panel, green areas indicate no evidence of lack of randomness of time-series residuals (p>0.05) while red panels (not shown here) indicate the opposite. The inner shaded area shows three standard errors from the overall mean and red circles identify a specific year with residuals greater than this threshold value (3- sigma rule).

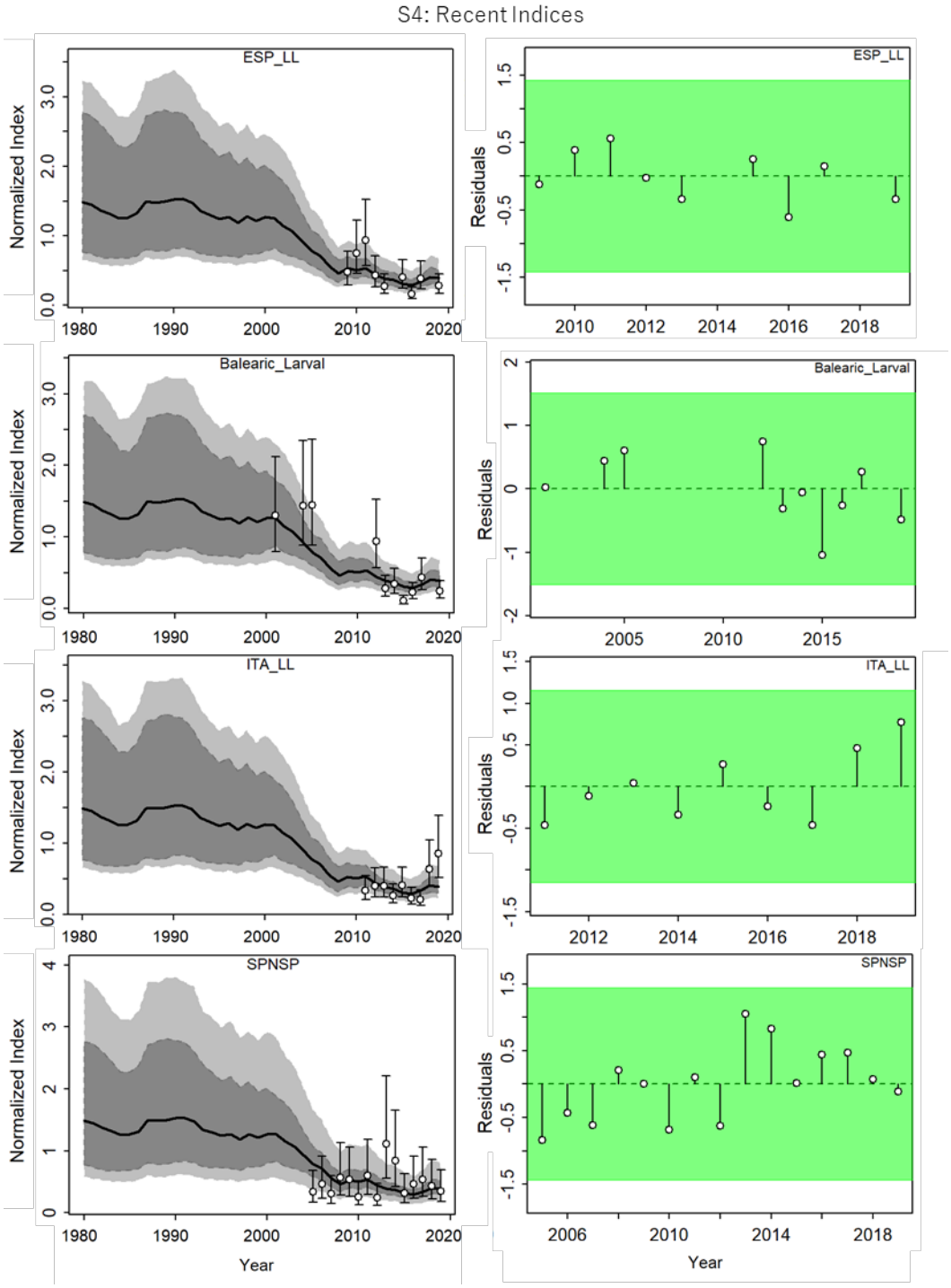

Figure 5. Time-series of observed (circle) with error 95% CIs (error bars) and predicted (solid line) CPUE (left) and Runs tests to quantitatively evaluate the randomness of the time series of CPUE residuals (right) for Mediterranean albacore "full" model (S4), shown for the most recent indices. On the left panel, the Dark shaded grey areas show 95% credibility intervals of the expected mean CPUE and light shaded grey areas denote the 95% posterior predictive distribution intervals. On the right panel, green areas indicate no evidence of lack of randomness of time-series residuals (p>0.05) while red panels (not shown here) indicate the opposite. The inner shaded area shows three standard errors from the overall mean and red circles identify a specific year with residuals greater than this threshold value (3- sigma rule).

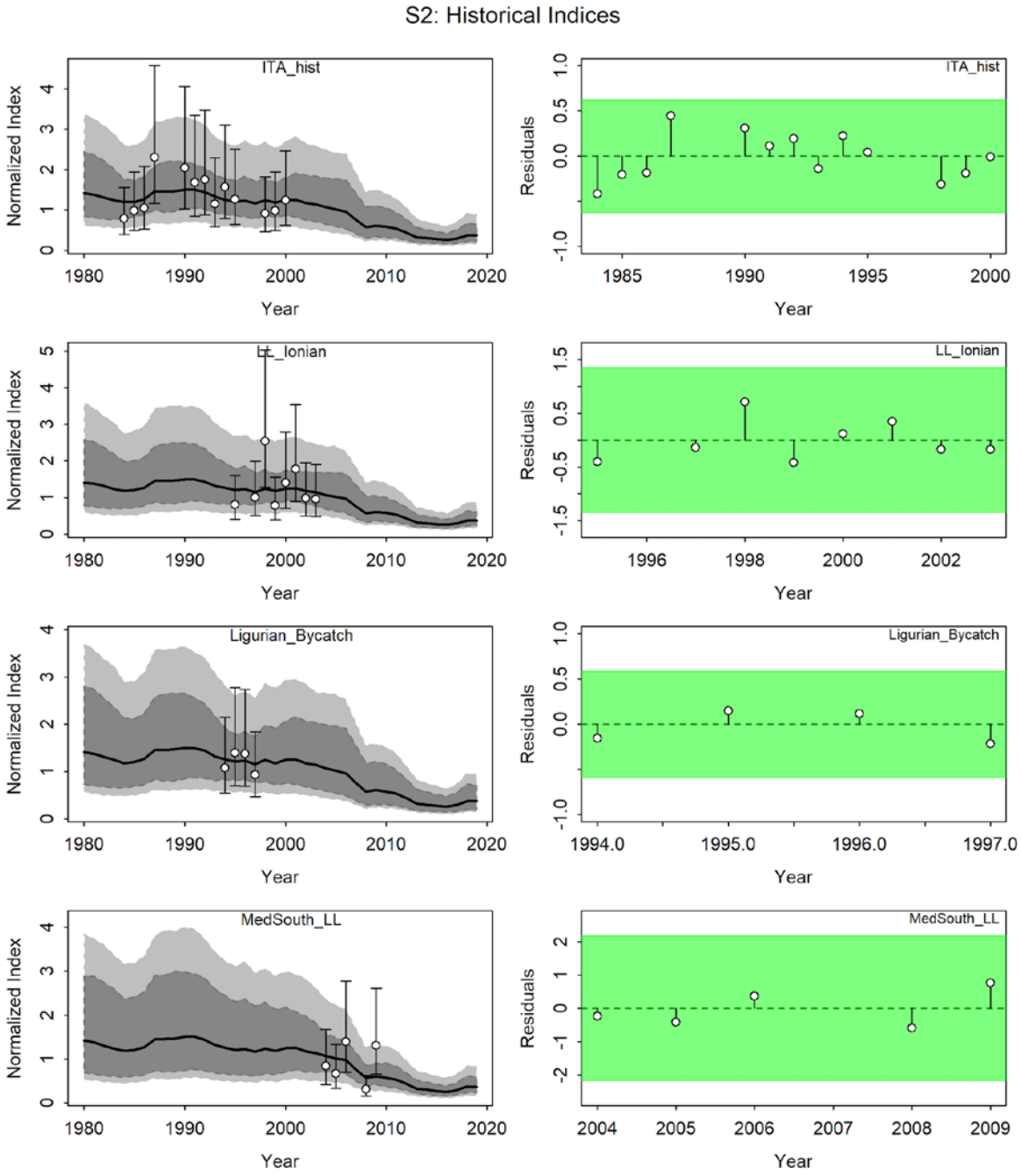

Figure 6. Time-series of observed (circle) with error 95% CIs (error bars) and predicted (solid line) CPUE (left) and Runs tests to quantitatively evaluate the randomness of the time series of CPUE residuals (right) for Mediterranean albacore "extended" model (S2), shown for the four historical indices that were included as additional indices in S2. On the left panel, the Dark shaded grey areas show 95% credibility intervals of the expected mean CPUE and light shaded grey areas denote the 95% posterior predictive distribution intervals. On the right panel, green areas indicate no evidence of lack of randomness of time-series residuals (p>0.05) while red panels (not shown here) indicate the opposite. The inner shaded area shows three standard errors from the overall mean and red circles identify a specific year with residuals greater than this threshold value (3- sigma rule).

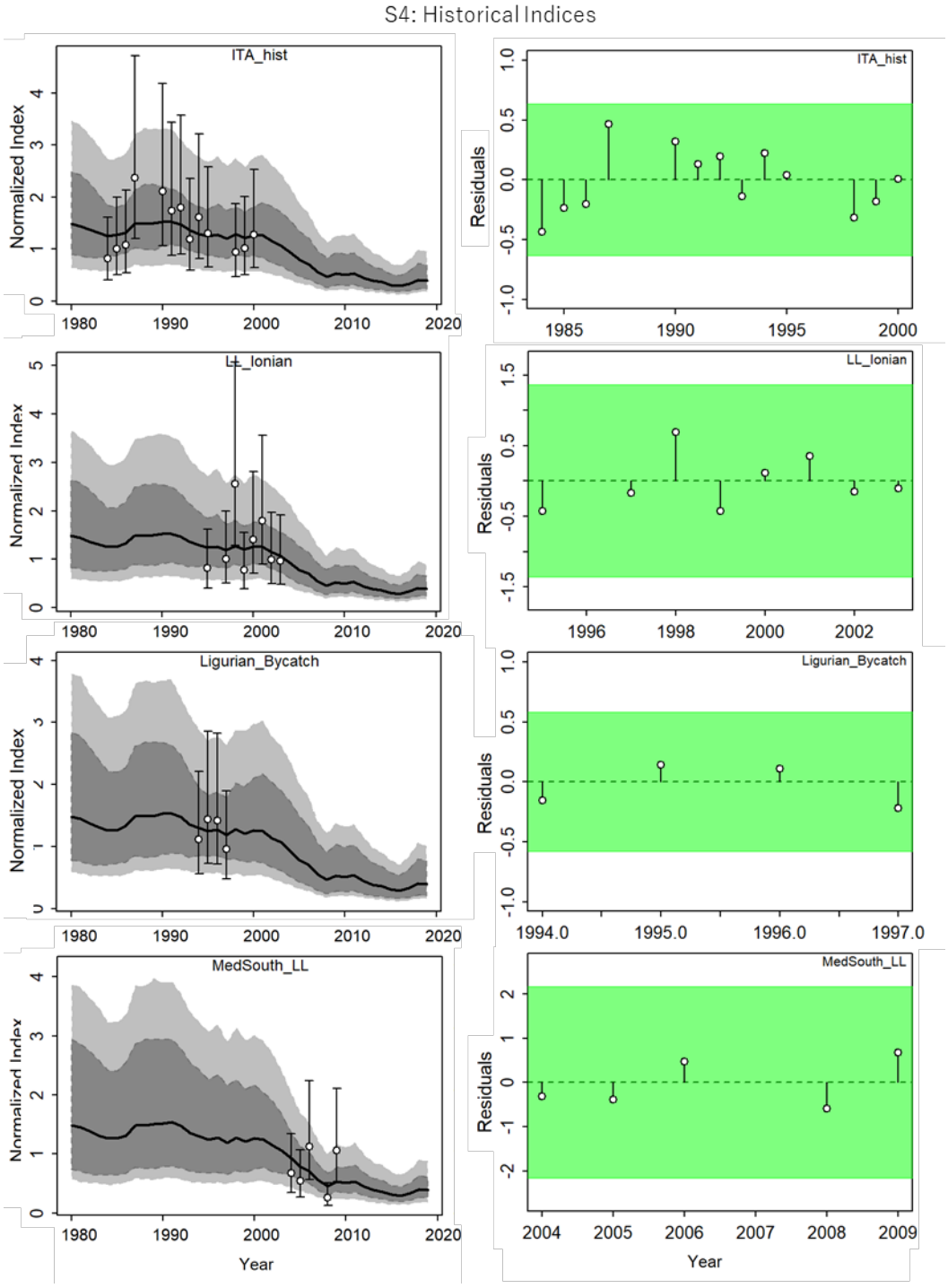

Figure 7. Time-series of observed (circle) with error 95% CIs (error bars) and predicted (solid line) CPUE (left) and Runs tests to quantitatively evaluate the randomness of the time series of CPUE residuals (right) for Mediterranean albacore "full" model (S4), shown for the four historical indices that were included as additional indices in S4. On the left panel, the Dark shaded grey areas show 95% credibility intervals of the expected mean CPUE and light shaded grey areas denote the 95% posterior predictive distribution intervals. On the right panel, green areas indicate no evidence of lack of randomness of time-series residuals (p>0.05) while red panels (not shown here) indicate the opposite. The inner shaded area shows three standard errors from the overall mean and red circles identify a specific year with residuals greater than this threshold value (3- sigma rule).

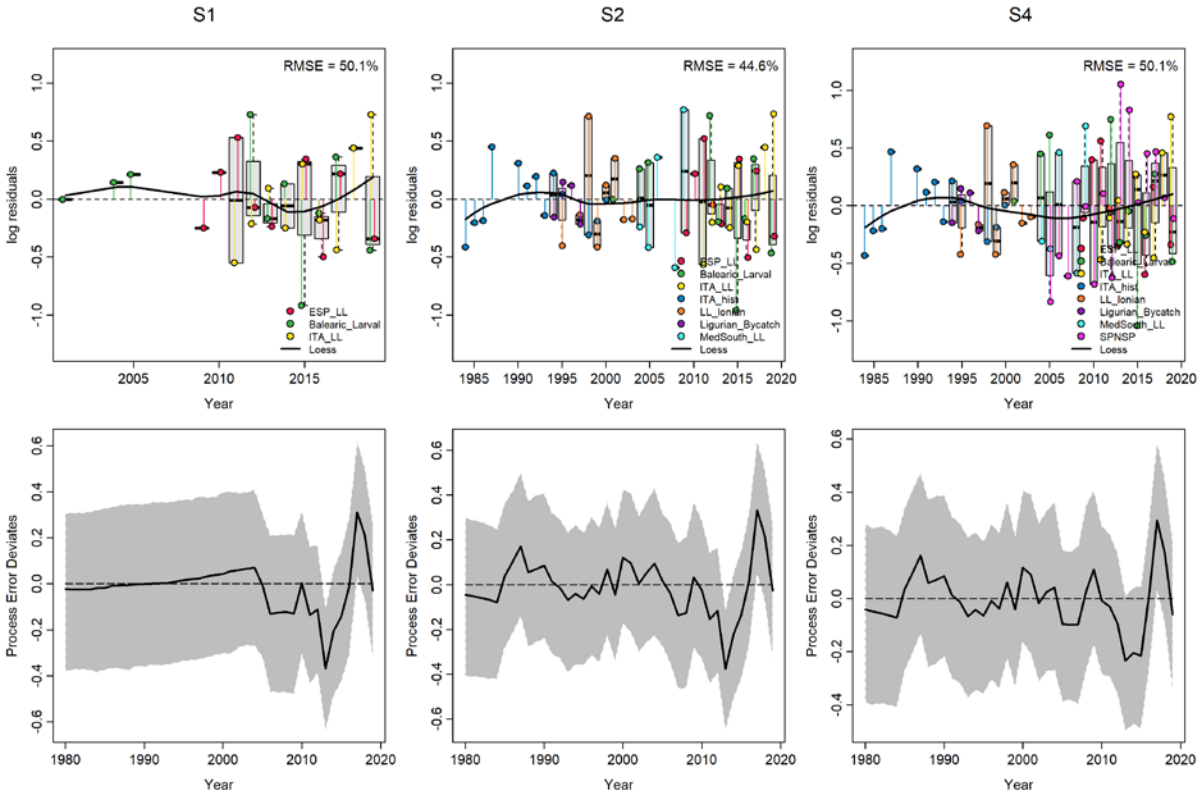

Figure 8. JABBA residual diagnostic plots for alternative sets of relative abundance indices examined for each scenario (Left: reference S1; Middle: "extended" S2; Right: "full" S4) for Mediterranean albacore. Top panels: Boxplots indicating the median and quantiles of all residuals available for any given year, and solid black lines indicate a loess smoother through all residuals. Bottom panels: Process error deviates (median: solid line) with shaded grey area indicating 95% credibility intervals.

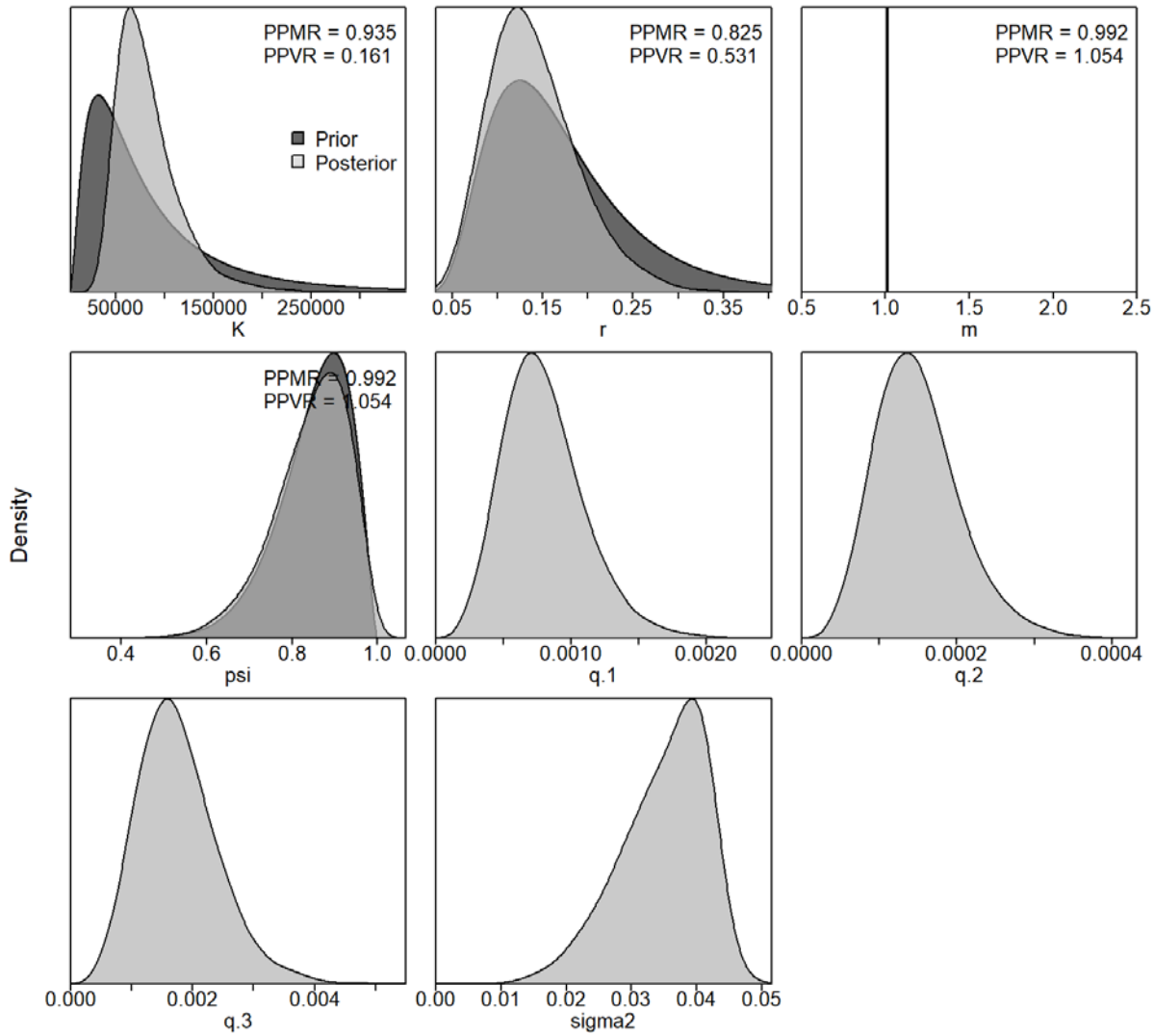

Figure 9. Prior and posterior distributions of various model and management parameters for reference run S1 for Mediterranean albacore. PPRM: Posterior to Prior Ratio of Means; PPRV: Posterior to Prior Ratio of Variances.

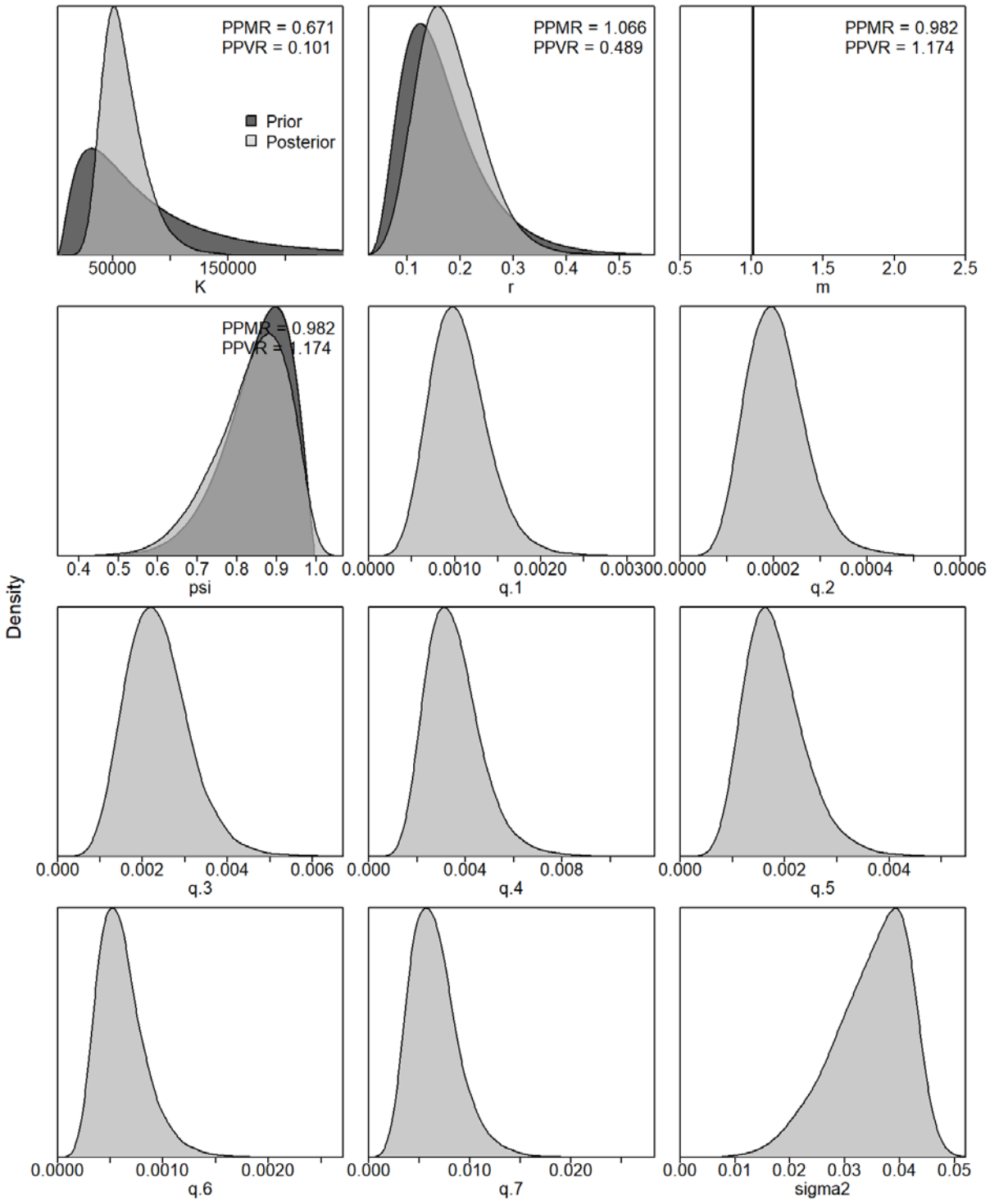

Figure 10. Prior and posterior distributions of various model and management parameters for the "extended" S2 model for Mediterranean albacore. PPRM: Posterior to Prior Ratio of Means; PPRV: Posterior to Prior Ratio of Variances.

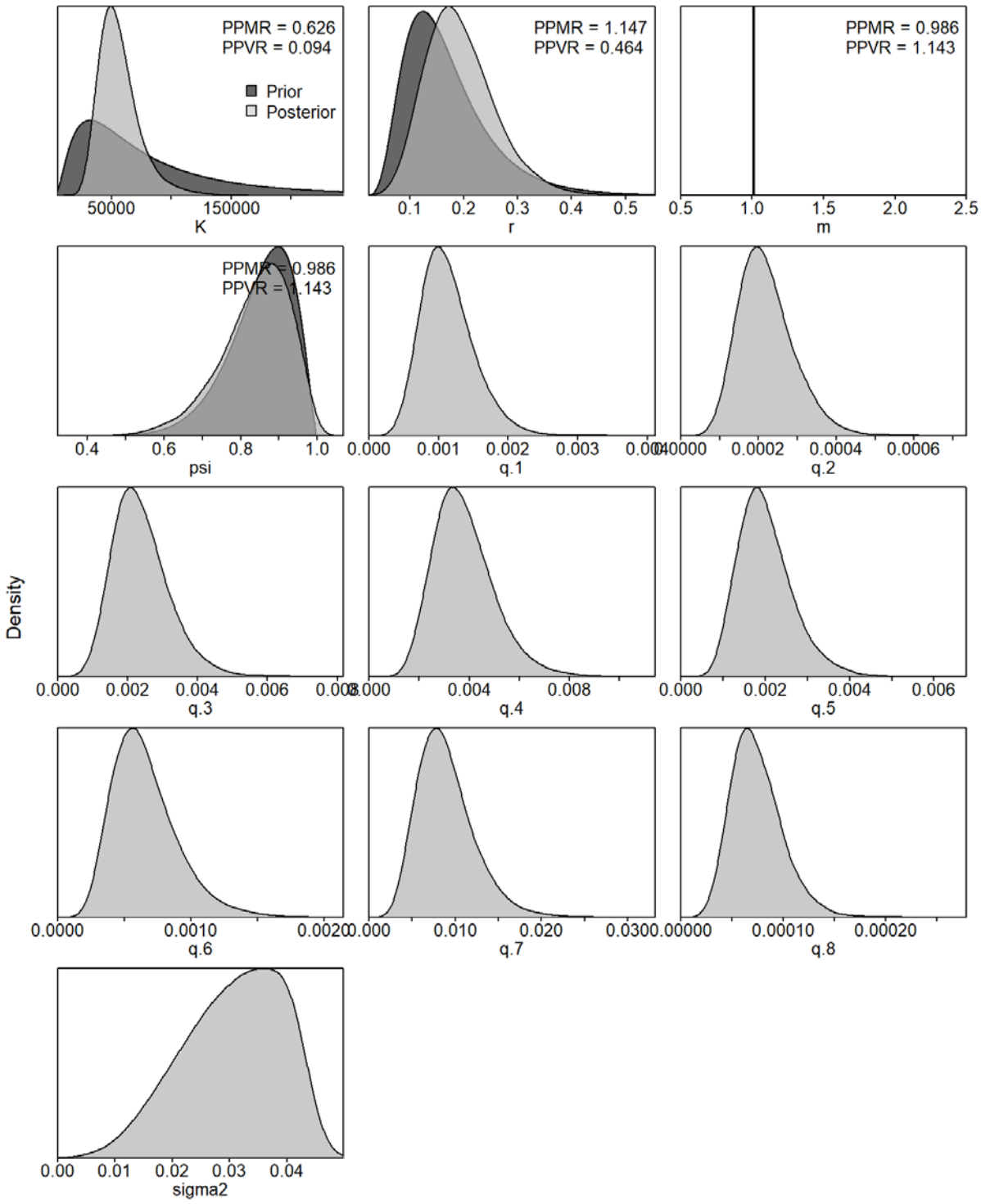

Figure 11. Prior and posterior distributions of various model and management parameters for the "full" S4 model for Mediterranean albacore. PPRM: Posterior to Prior Ratio of Means; PPRV: Posterior to Prior Ratio of Variances.

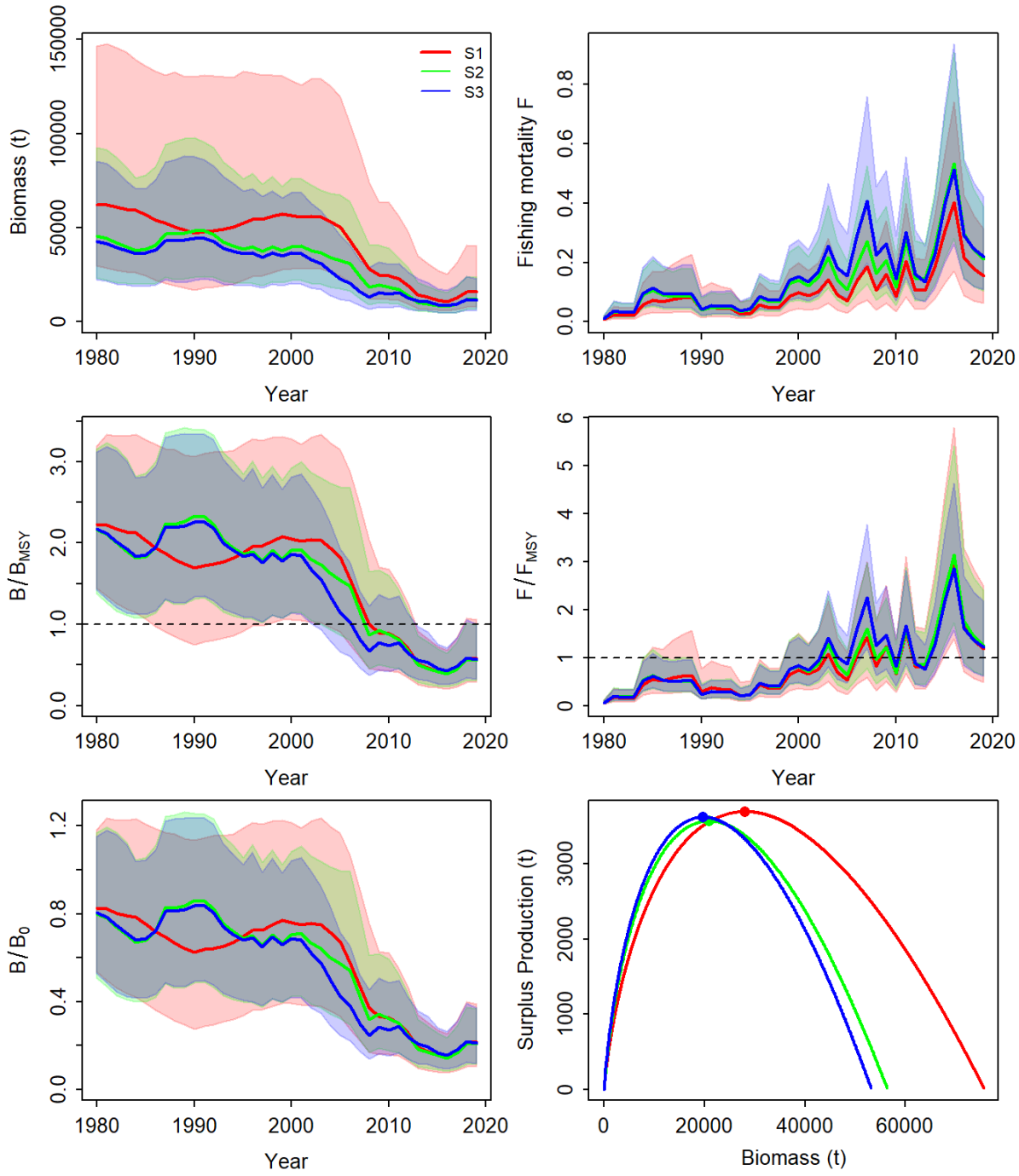

Figure 12. Comparison stock trajectory estimates for the Mediterranean albacore S1, S2 and S4 scenarios, showing trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels)

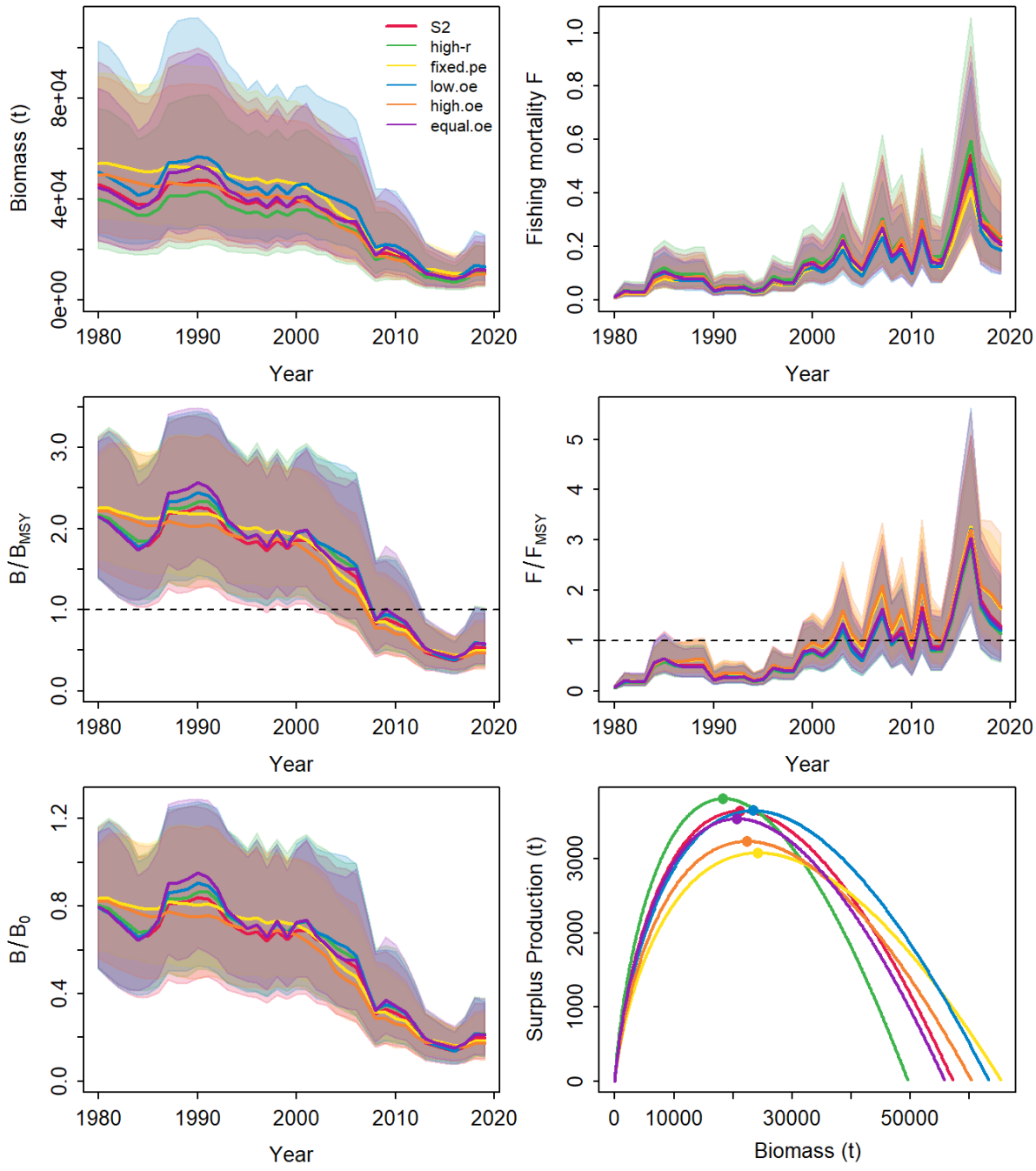

Figure 13. Sensitivity analysis performed on the "extended" model S2 for alternative productivity (*r*) and variance parameter assumptions described in Table 2. high-r: Increased r prior mean by a factor of 1.5, fixed.pe: process error fixed to 0.1, low.oe: lower fixed observation error, high.oe: higher observation error, equal.oe: equal observation for current and historical indices

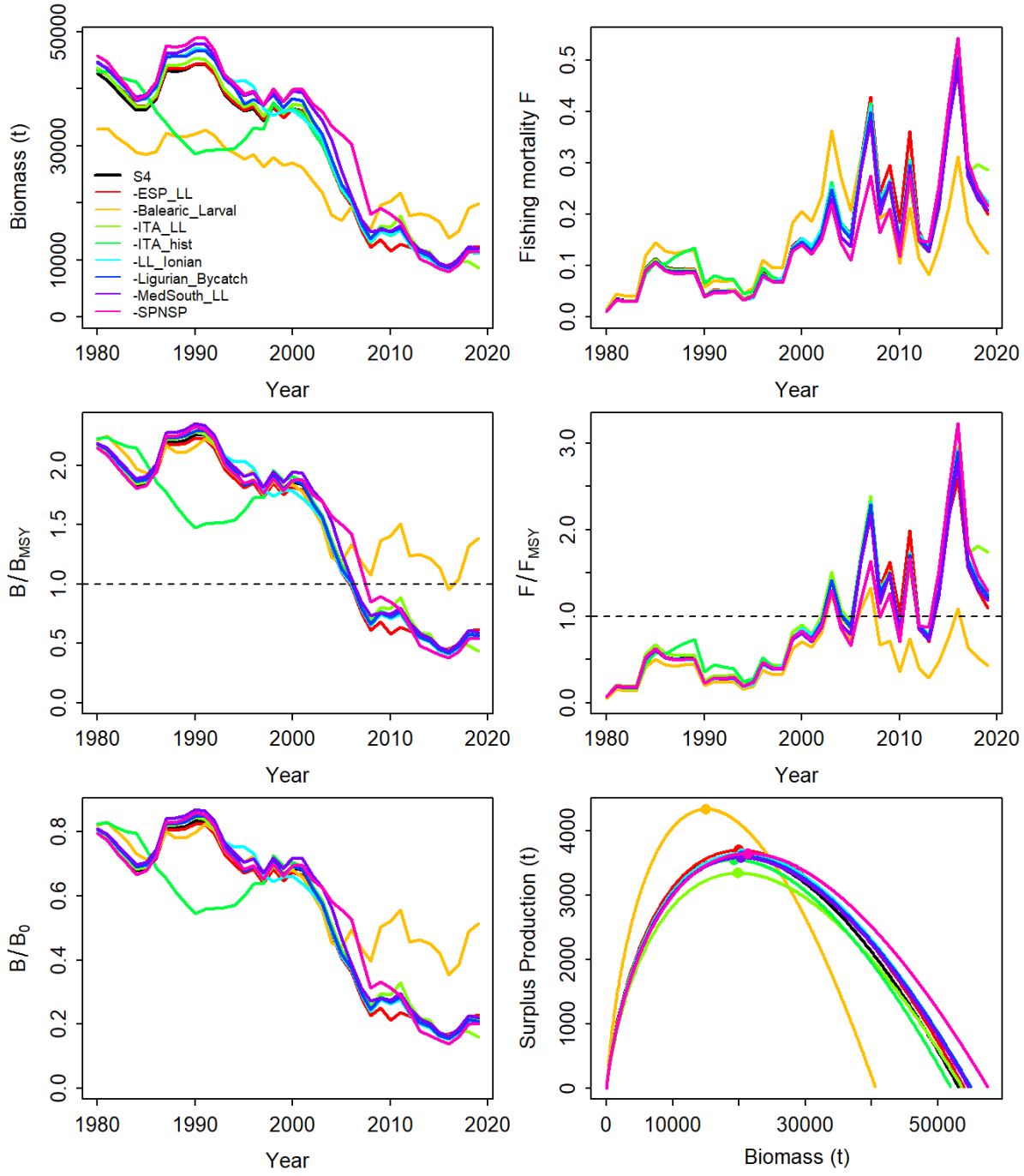

Figure 14. Jackknife index analysis performed on the "full" model S4, by removing one CPUE fleet at a time and predicting the trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels) for each scenario from the Bayesian state-space surplus production model fits to Mediterranean albacore.

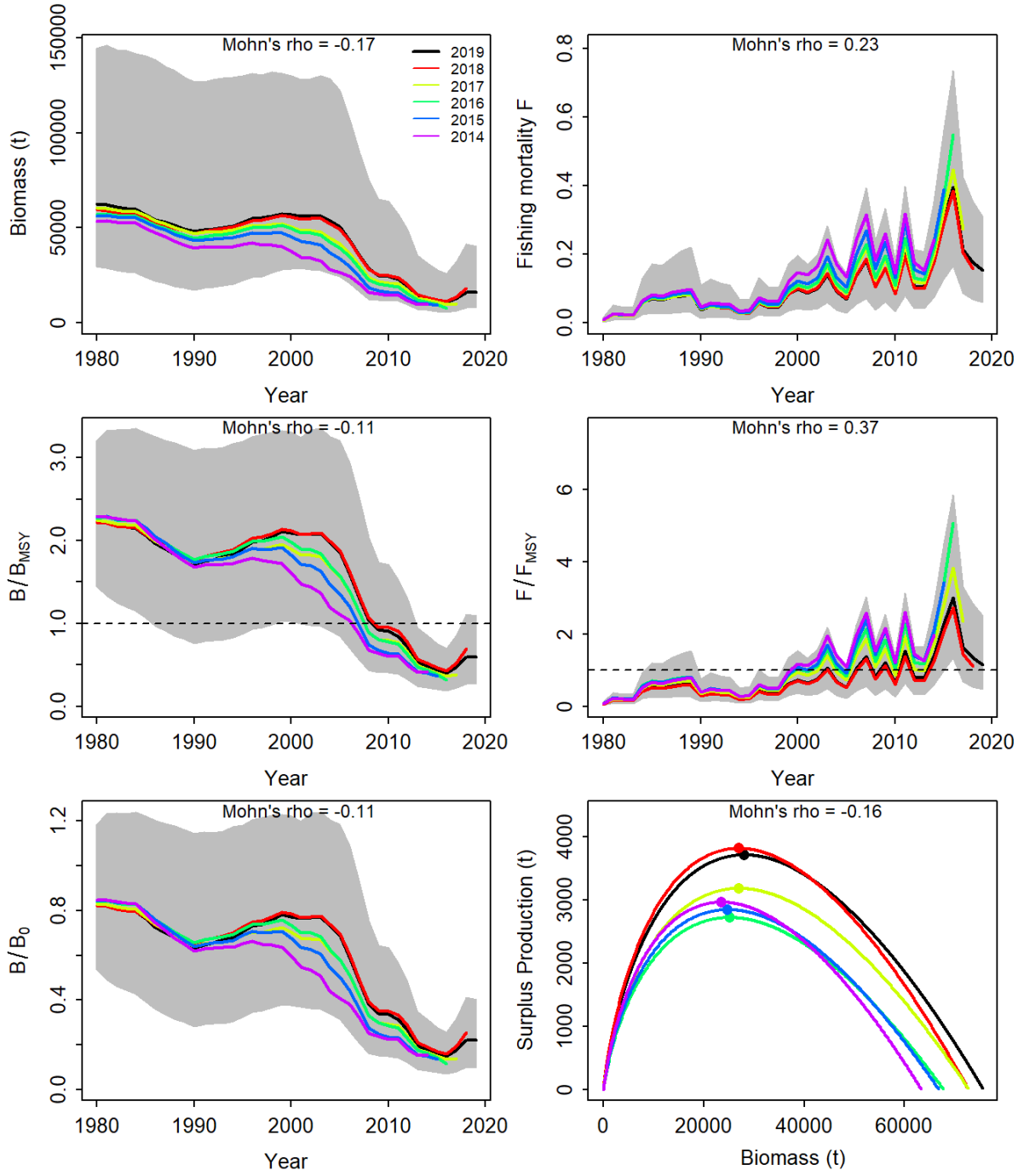

Figure 15. Retrospective analysis performed for the reference model run S1, by removing one year at a time sequentially (n=5) and predicting the trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels) for each scenario from the Bayesian state-space surplus production model fits to Mediterranean albacore.

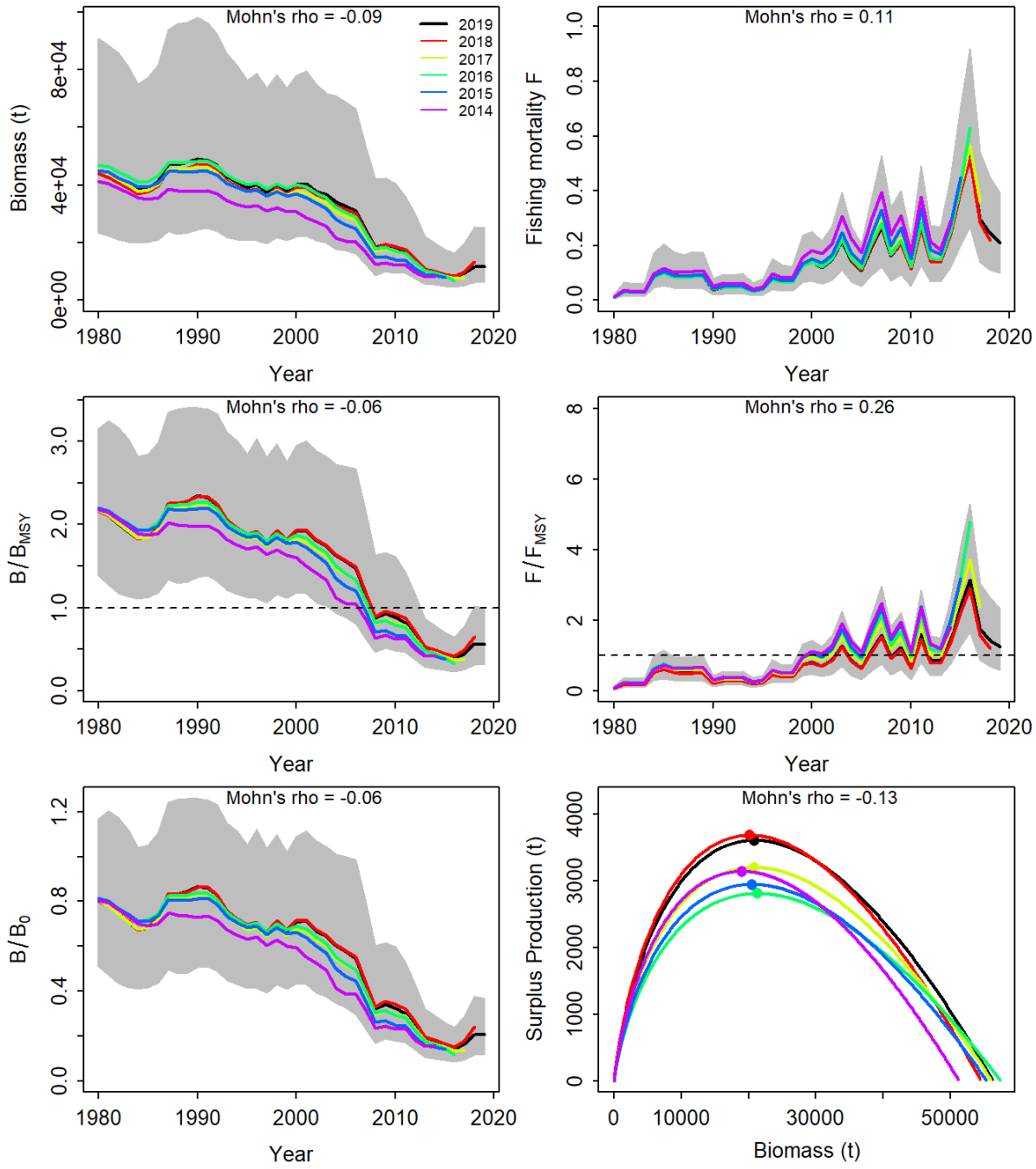

Figure 16. Retrospective analysis performed for the "extended" model run S2, by removing one year at a time sequentially (n=5) and predicting the trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels) for each scenario from the Bayesian state-space surplus production model fits to Mediterranean albacore.

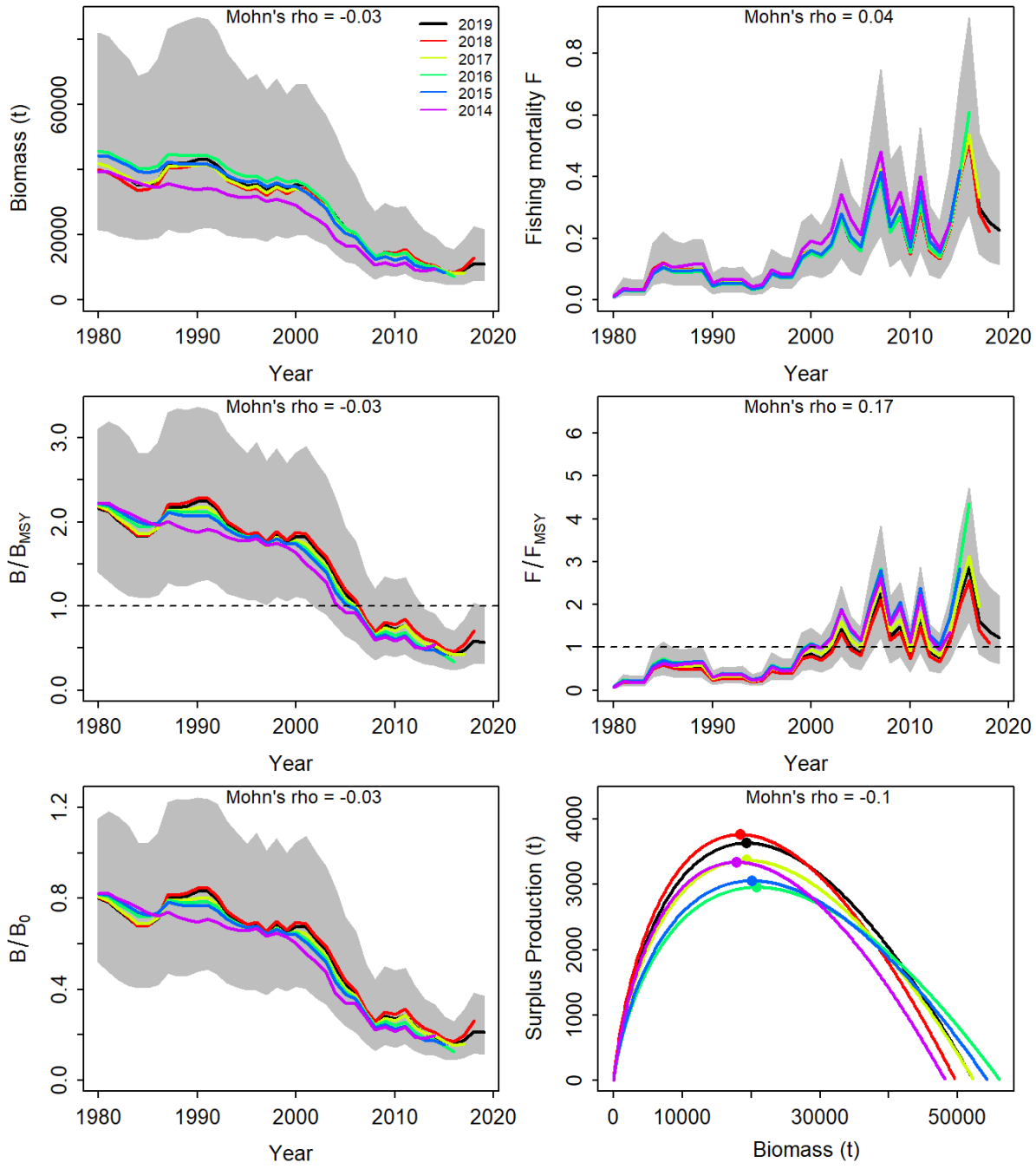

Figure 17. Retrospective analysis performed for the "full" model run S4, by removing one year at a time sequentially (n=5) and predicting the trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels) for each scenario from the Bayesian state-space surplus production model fits to Mediterranean albacore.

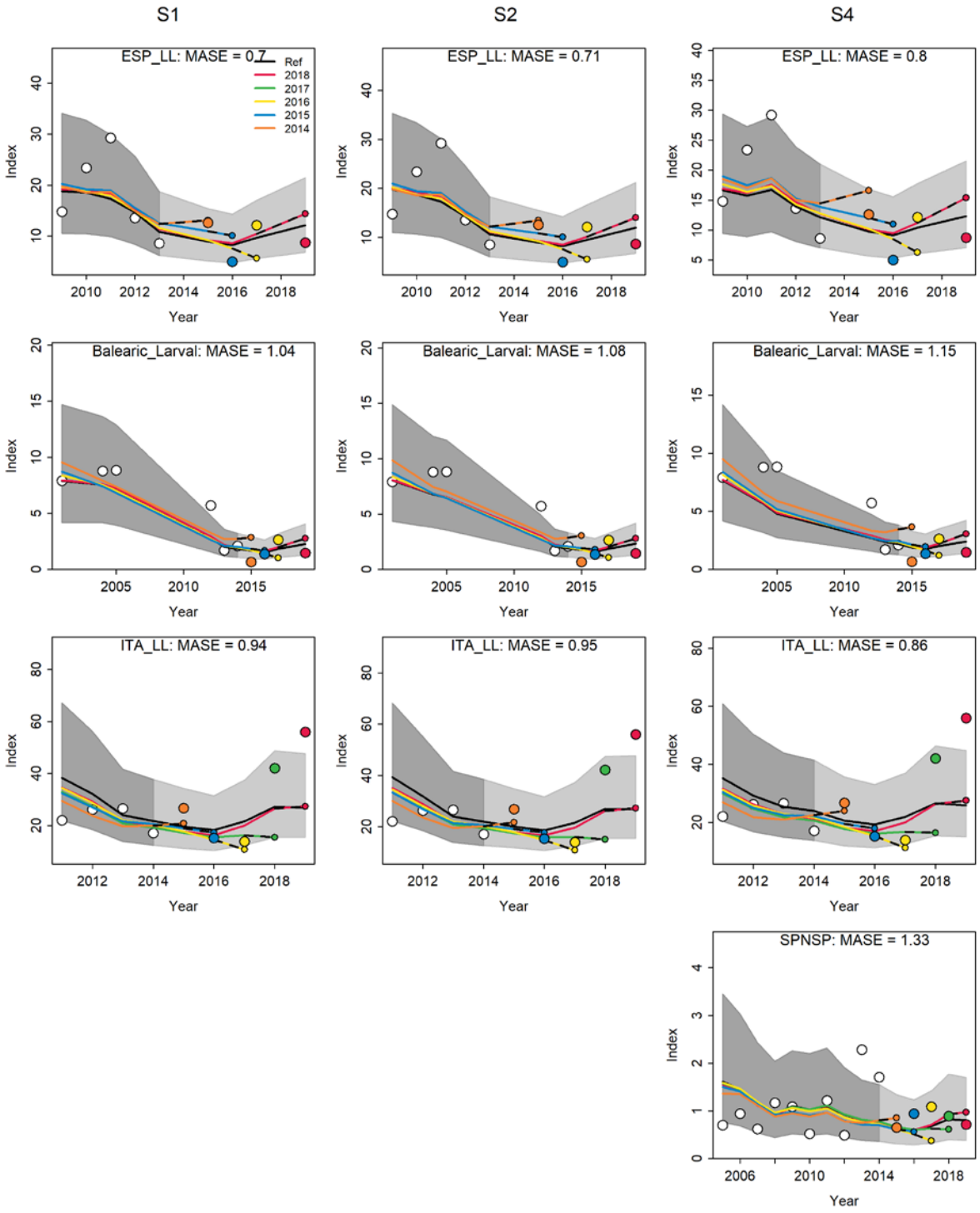

Figure 18. Hindcasting cross-validation results (HCxval) for the scenarios S1, S2, and S4 for Mediterranean albacore, showing one-year-ahead forecasts of CPUE values (2015-2019), performed with five model hindcast runs. The CPUE observations, used for cross-validation as prediction residuals, are highlighted as color-coded solid circles with associated light-grey shaded 95% confidence interval. The model reference year refers to the end points of each one-year-ahead forecast and the corresponding observation (i.e. year of peel + 1).

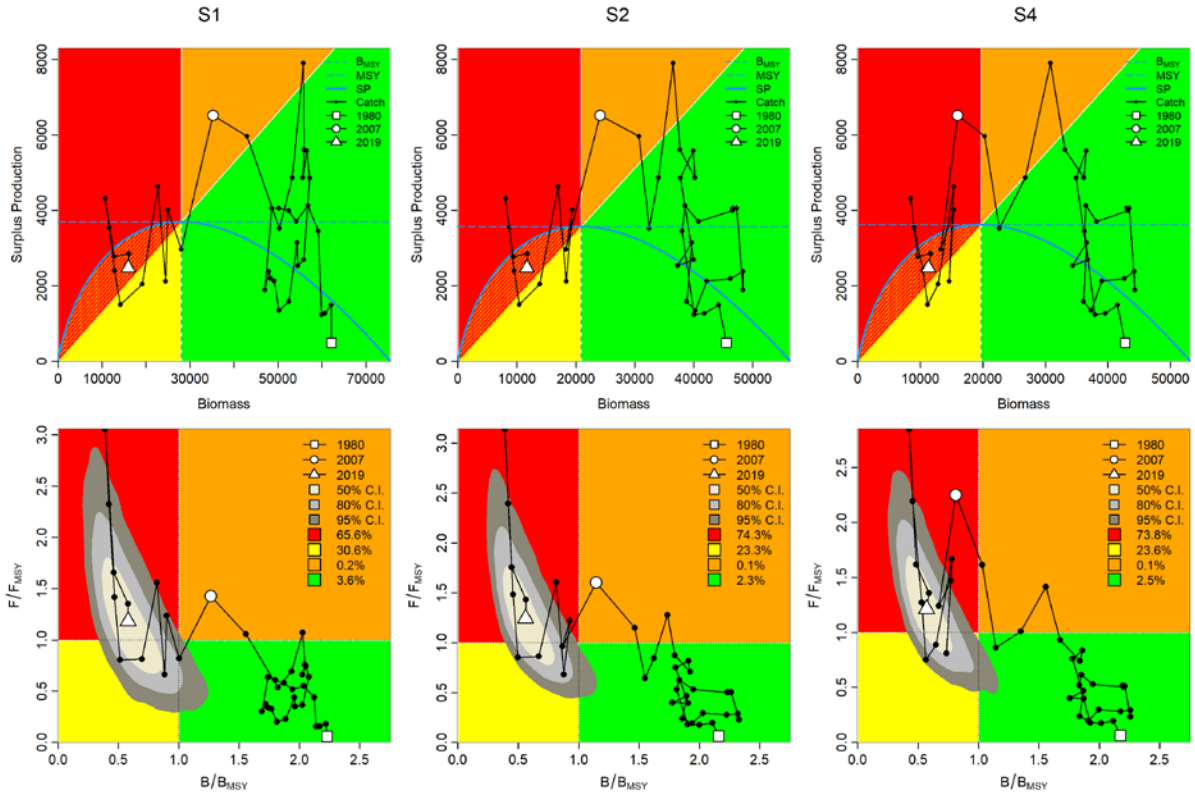

Figure 19. JABBA surplus production phase plot for the reference run S1 (left panels), "full" model S2 (middle panels) and "full" model S4 (right panels) showing trajectories of the catches in relation to B_{MSY} and *MSY* (top panels) and Kobe phase plot showing estimated trajectories (1980-2019) of B/B_{MSY} and F/F_{MSY} for the Bayesian state-space surplus production model for the Mediterranean albacore (bottom panels). Different grey shaded areas denote the 50%, 80%, and 95% credibility interval for the terminal assessment year. The probability of terminal year points falling within each quadrant is indicated in the figure legend.

Supplement: Additional Model Diagnostics for a third scenario that excluded the Balearic Larval Index

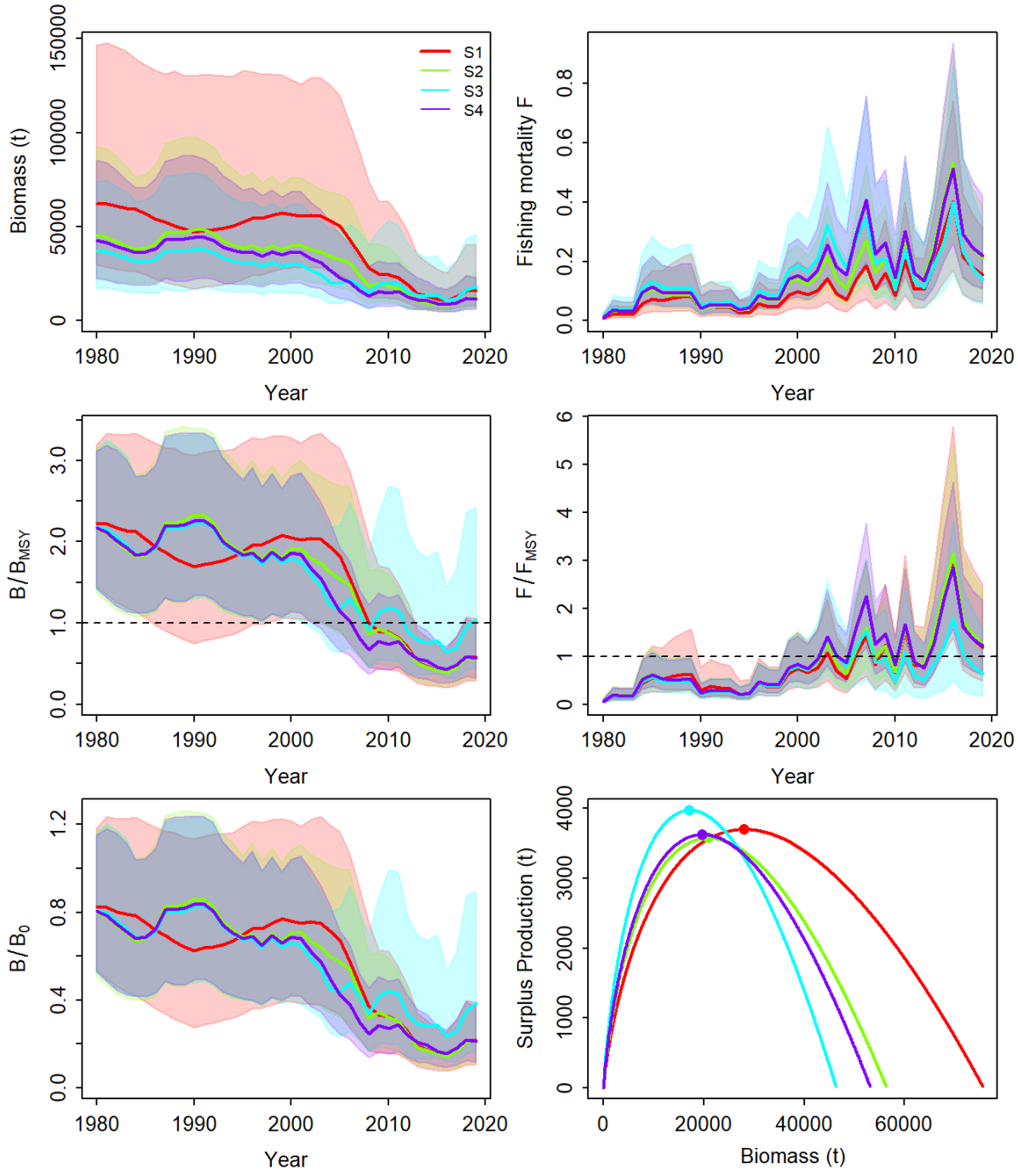

Figure S1. Comparison stock trajectory estimates for the Mediterranean albacore reference runs S1, the "full" S2 model and third scenarios S3, based on S3 but excluding the Larval-Index, showing trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels)

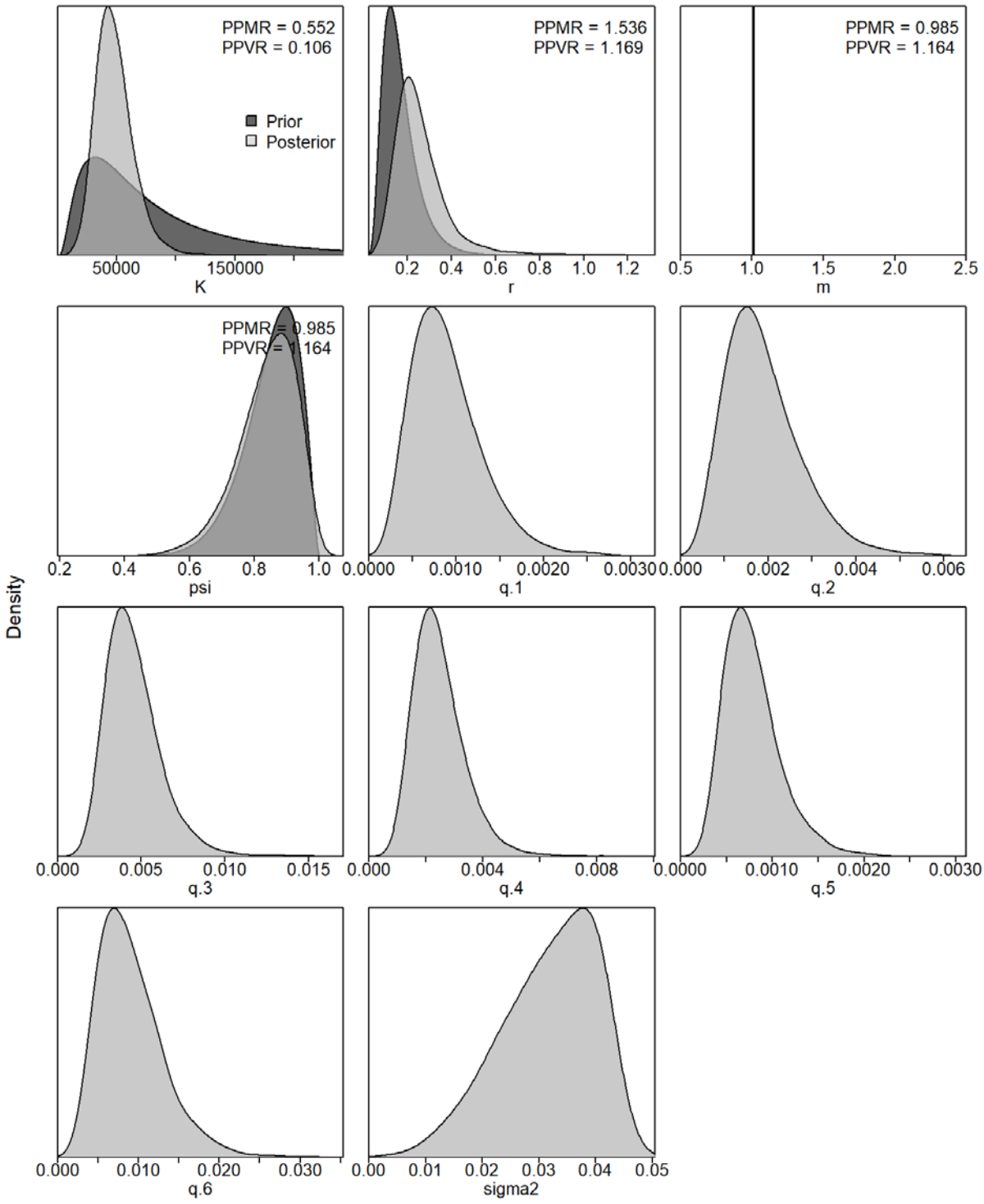

Figure S2. Prior and posterior distributions of various model and management parameters for the additional S3 scenario, excluding the 'Larval-Index', for Mediterranean albacore. PPRM: Posterior to Prior Ratio of Means; PPRV: Posterior to Prior Ratio of Variances.

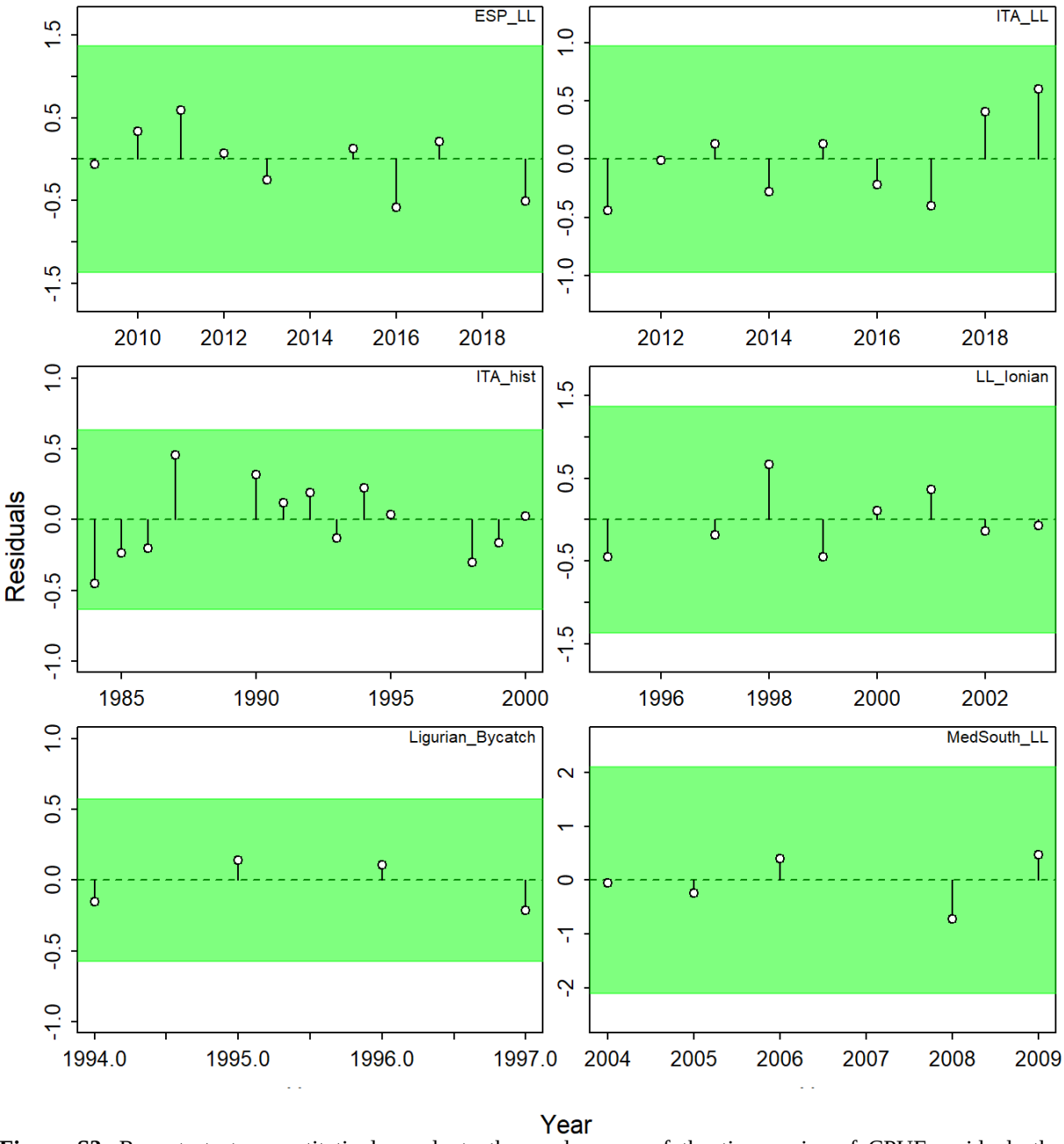

Figure S3. Runs tests to quantitatively evaluate the randomness of the time series of CPUE residuals the additional S3 scenario, excluding the 'Larval-Index'. On the left panel, the Dark shaded grey areas show 95% credibility intervals of the expected mean CPUE and light shaded grey areas denote the 95% posterior predictive distribution intervals. On the right panel, green areas indicate no evidence of lack of randomness of time-series residuals ($p>0.05$) while red panels (not shown here) indicate the opposite. The inner shaded area shows three standard errors from the overall mean and red circles identify a specific year with residuals greater than this threshold value (3- sigma rule).

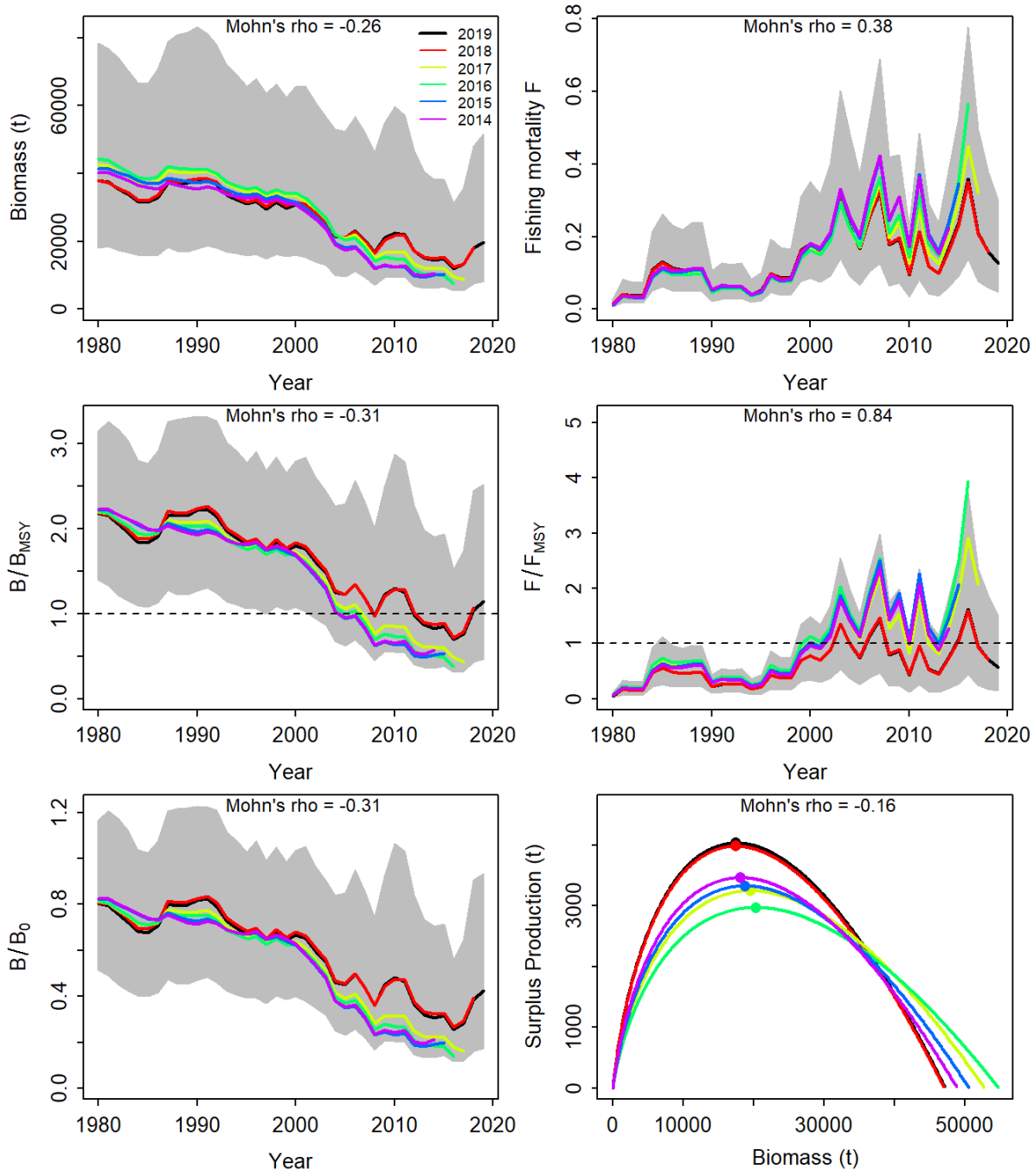

Figure S4. Retrospective analysis performed for S3, which is based on S2 but excluding the Larval-Index, by removing one year at a time sequentially (n=5) and predicting the trends in biomass and fishing mortality (upper panels), biomass relative to B_{MSY} (B/B_{MSY}) and fishing mortality relative to F_{MSY} (F/F_{MSY}) (middle panels) and biomass relative to K (B/K) and surplus production curve (bottom panels) for each scenario from the Bayesian state-space surplus production model fits to Mediterranean albacore.

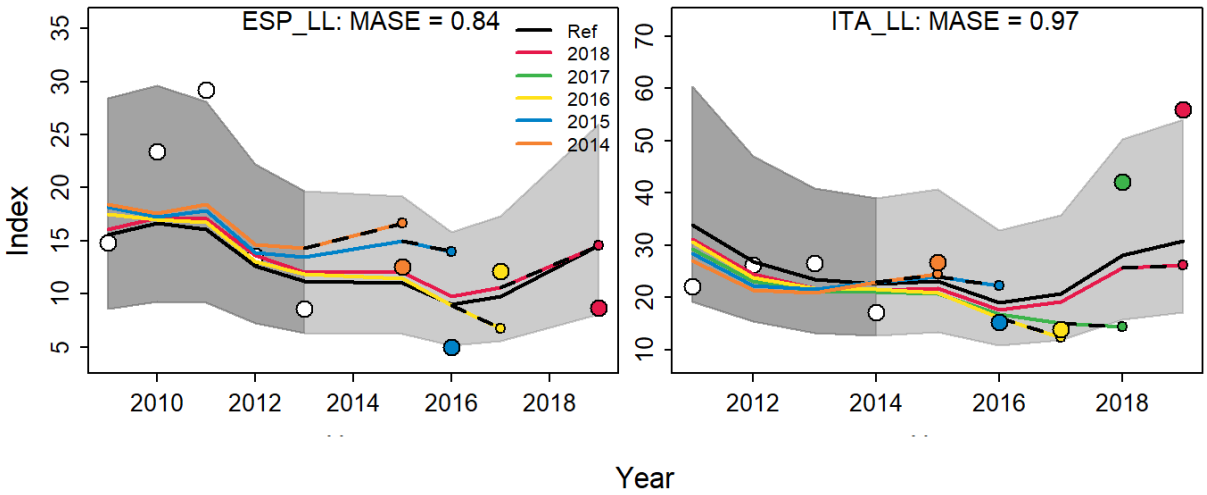

Figure S5. Hindcasting cross-validation results (HCxval) for the scenario S3 for Mediterranean albacore, showing one-year-ahead forecasts of CPUE values (2015-2019), performed with five model hindcast runs. The CPUE observations, used for cross-validation as prediction residuals, are highlighted as color-coded solid circles with associated light-grey shaded 95% confidence interval. The model reference year refers to the end points of each one-year-ahead forecast and the corresponding observation (i.e. year of peel + 1).