Alternative Assessment for the Indian Ocean Yellowfin Tuna Stock; with Generic Goodness of Fit Diagnostics.

Laurence Kell and Rishi Sharma

Summary

- The objective of this work is to assist the IOTC Scientific Committee in providing robust management advice for yellowfin tuna by evaluating alternative assessment methods and scenarios that reflect uncertainty about the assumptions and datasets used in the assessment.
- Specific tasks are to agree datasets (CPUE series and catch series) based on alternative assumptions about mis-reporting, run biomass dynamic based stock assessments. and compare the results to the base case SS3 assessment using a common set of diagnostics.
- The report summarises the analysis and provides a set of diagnostics that can be used for comparison across different modelling platforms
- All scenarios other than 2 (high productivity and the reference case fitted to the estimate of biomass from stock synthesis) indicate that the stock is overfished and experiencing overfishing.
- The work is based on data available on July 8th, 2019.

Preliminary Report

Contents

Executive Summary

Seven alternative scenarios were developed for yellowfin based on hypotheses that examined the impact of high grading on the Catch Per Unit Effort (CPUE) series that are used as indices of abundance in the assessment. Five of these were based on the LL CPUE indices with different assumptions about high grading and the length of the time series used. The initial model set up (i.e. the base case) included the PS index as well as the LL indoices. In addition scenario that included only the PS CPUE series was run. The scenarios with the PS index were discounted, however, as the diagnostics indicated poor fit due to little informationon on stock abudance in the PS series. In the case of the fit to the PS series alone R_0 parameter that influences B_0 (the scale parameter) was unrealistically large, and so the stock was unaffected by catches.

Alternative models therfore focussed on the LL series with different assumptions. These series were then fitted to a surplus production model (JABBA, Winker, et al., 2018). Diagnostics indicated that the indices were highly correlated with each other and were also positively correlated with Stock Synthesis (SS3, Methot and Wetzel, 2013) estimates of stock abundance. This supports the assumption that the biomass dynamic model and the integrated assessment inputs are consistent. Furthermore the absence of lags between the indices supports the use of a biomass dynamic without regional structure, since there does not appear to be any spatial hetrogeneity of year classes.

Diagnostics used to evaluate goodness of fit based on model residuals, indicate that there is a systematic over or underprediction of of relative abudance by the indices in the different areas, other than the case of region 3 with vessel effects. The results from hindcasting suggested that the assessment has poor prediction skill in recent years as it under predicts abundance by a large proportion. This could be due to changes in catchability or reporting. An examination of process error, i.e. the difference between the expected and realised productivity indicate that there are potentially regime shifts due to changes in the distribution of the stock in recent years, particularly in Regions 2 and 4.

Other than under the high productivity scenario, where the population growth rate (r) is greater than 1.2, and the scenario fitted to the overall biomass trajectory from the last integrated assessment, all models indicate that the stock is overfished and experiencing overfishing.

The tools presented here (along with the R scripts developed) provide an generic set of diagnostics that allows models to be validated using prediction residuals and prediction skill. Importantly the framework allows model performance to be compared across different platforms with differeent model assumptions and datasets In the biomass dynamic model case examined here model fits are primarily driven by the indices of abundance, but the same diagnistics can be used for model with different structures, e.g. integrated models, allowing them to be compared to the biomass dynamic model scenarios.

Introduction

To assess the status of the Indian Ocean stock of yellowfin tuna the Scientific Committee has relied mainly on Multifan-CL and and Stock Synthesis III. These are both statistical, length-based, age-structured models, that integrate fishery data including total catch, effort, and length-frequency data, from all Indian Ocean fleets catching yellowfin tuna. The IOTC Scientific Committee, however, has identified several issues that may hamper its capacity to provide sound advice for yellowfin tuna. In particular there are poor estimates of catch for many Indian Ocean fleets; significant changes in longline time-area coverage and non-reporting of discards (high-grading), that may affect the quality of the indices of abundance derived from longline logbook data, especially over the last decade (IOTC, 2019a). Very little length-frequency data are available from the longline fleets and their is a lack of small lengths in samples due to the alleged non-reporting of discards (at least since 2004 due to high-grading), and a lack of catch-and-effort data with which to derive indices of abundance.

There are also various issues related to the configuration of the models and interpretation of model results. Therefore different stock assessment modelling approaches were examined in the 2018 Indian Ocean yellowfin assessment. Although Stock Synthesis 3 was applied there were major problems due to conflicting data inputs, choice of data weighting, changes in catchability, uncertainty in size data and changes in selectivity over time particularly since the development of the industrial purse seine. The IOTC Scientific Committee therefore recommended that alternative, less complex, stock assessment methods should also be tried to assess the stock of yellowfin tuna.

The objective of this work therefore is to assist the IOTC Scientific Committee in providing robust management advice for yellowfin tuna in 2019, by evaluating alternative assessment methods and scenarios that reflect uncertainty about the assumptions and datasets. Specific tasks are to agree datasets based on alternative assumptions about mis-reporting, run biomass dynamic based stock assessments, and compare the results to the base case SS3 assessment using a common set of diagnostics. Diagnostics include an examination of model and prediction residuals, estimation of prediction skill, and a comparison of production functions process error and stock status relative to reference points.

Material and Methods

Material/Data

It was assumed that there is a single yellowfin stock with no spatial structure and an annual time step was assumed.

Assessments specifications

The assessment was conducted using a biomass dynamic model with a Pella Tomlinson production function so that alternative assumptions about productivity, stock status and reference points can be evaluated. The assessment model used was JABBA (Winker et al. 2018) which also allows process error to be fitted.

The original specifications for the indices of abundance were reviewed since they need to provide plausible and consistent hypothesis based on the available data.

Assessment assumptions were

- Stock structure: single Indian Ocean YFT stock
- Spatial stratification: Whole ocean
- Temporal stratification: Year
- Fisheries: All fisheries combined
- Time series for CPUEs: Early 1950's-2018 (based on aggregated data without vesselleffects); 1972-2018 (2 values of q i) for data from 1972-2018 with no vessel effects, and ii) from 1979-2018 with vessel effects);
- Catch data: time-series of (best-estimate) catch published by the IOTC Secretariat
- CPUE/Indices of abundance:
    - joint LL index, as provided by IOTC, (2019b);
    - joint-CPUE LL index corrected for high grading (IOTC, 2019b);
    - PS indices, as provided by EU scientists
- Considering the findings from Geehan and Hoyle (2013), analyse the time-series of length frequency data from Taiwan, Japan, Seychelles, and the Republic of Korea, as available at the IOTC, in order to quantify the number of yellowfin tuna discarded which has been potentially lost from length samples, due to high-grading. Based on this, develop scenarios for adjusting the joint-Asian longline index for high-grading and examine in the assessments. Note this will affect both the abundance index and the catch series used in these assessments. However, since these are small fish (discards) the effects on biomass are marginal (numbers change by 5-10% so the biomass changes are smaller, i.e. <5%). We therefore did not change the estimates of biomass that were used in the model fitting exercise.
- Age and Length data: Not included
- Tagging data: Not included
- Environmental indices: Not included
- Model structure and assumptions: equivalent life history assumptions to those used in SS3 assessment in 2018
- Any other alternative indices of abundance, where available and appropriate.

CPUEs/Indices of abundance: joint LL index, as provided by IOTC (2019b); joint-CPUE LL index corrected for high grading (one or more scenarios, as required); PS index(es), as provided by EU scientists (the fishery dependent index may be subject to bias as well as catch estimates are likely doto be wrong for the YFT).

CPUE indices were computed by IOTC (2019b). accounting for vessel, area and high grading effects. 5 scenarios were examined for the 4 regions used in the assessment. The following were developed:

1. Scenario 1: Reference case model from updated data for Synthesis, accounting for vessel effects from 1979 onwards and non-vessel effects from 1972-1979.
2. Scenario 2: Vessel effects from IOTC (2019b) from 1979-2018 for all areas using Area 2 (Hoyle analysis) representing Area 1 in the assessment.
3. Scenario 3: Vessel effects from IOTC (2019b) from 1979-2018 for all areas using Area 7 (Hoyle analysis) representing Area 1 in the assessment.

4. Scenario 4: Longer term series with no vessel effects from early 1950's to current period using aggregated data. Area 2 (IOTC, 2019b) representing Area 1 in the assessment.
5. Scenario 5: Longer term series with no vessel effects from early 1950's to current period using aggregated data. Area 7 (IOTC 2019b) representing Area 1 in the assessment.

Scenarios for high grading are mostly occurring in Area 2 and Area 7 (IOTC, 2019b). For further details, refer to the paper there as there are many reasons for these changes.

It was however notedthat the LL indices incorporate a correction for high grading but that correction is based on logbook reports and it is very likely that those reports do not contain all fish discarded as logbooks usually miss quite a lot of fish. This is confirmed in the publication that uses observer data from previous years that mentions that discards may be as much as 40% in some areas. In other oceans (for eg. Western central Pacific), the Electronic Monitoring Systems (EMS) confirmed that most discards are not recorded on logbooks. Thus, the real magnitude is probably unknown, and relying solely on LL indices may be subject to large biases.

These hypothesized changes are examined for diagnostics and performance in how the model does, and what is most representative of the stock.

Model structure and assumptions: As a biomass dynamic assessment model was used priors for r, K and the shape of the production function were derived for assumptions about natural mortality, and stock recruitment. Details of all models will be provided at the meeting on the working group server.

Scenarios

The biomass dynamic model was used to evaluate the impact on the assessment of alternative assumptions about the indices of abundance used for calibration. A base case, which included the long line and purse seine indices, was run first to compare the impact on the assessment of indices based on the different gears.

Considering the findings from Geehan and Hoyle (2013) the time-series of length frequency data from Taiwan, Japan, Seychelles, and the Republic of Korea, as available at the IOTC, were analysed in order to quantify the number of yellowfin tuna discarded and which have potentially been lost from the length samples, due to high-grading. Based on this, scenarios were developed to adjust the joint-Asian longline index for high-grading for examination in the assessment.

The CPUE were adjusted based on the discarding rates estimated during the Joint CPUE workshop. The estimates were considered preliminary and were what was available to the analysts on July 8th, 2019. Quarter based CPUE's developed by IOTC (2019b) were averaged for a yearly index that was used in the model fitting. Discard rates were 5-10% in recent years, but as this is in numbers are mostly small fish, we used the reported biomass estimates in model fitting.

Methods

When fitting stock assessment models, it is assumed that CPUE time series are indices of relative abundance. If indices are in conflict, however, then model estimates may be uncertain or biased. Therefore the CPUE series used were first summarised and compared with each other. If indices are not correlated this may suggest alternative data weightings and/or runs. A reason for poor correlation may be because indices represent different age classes or due to spatial hetrogenity. In which case there may be lags between indices, which can be evaluated by plotting cross-correlations.

For the biomass dynamic model two reference case scenarios were developed for comparison with the SS base case. The reference cases were fitted using the SS estimates of SSB and biomass, i.e. used as perfect indices of abundance. This provides a benchmark against which the scenarios with the biomass dynamic assessment scenarios based on alternative datasets and assessment model assumptions can be compared. Following this biomass based assessments scenarios were developed based on alternative CPUE series and assumptions about productivity.

A common set of diagnostics, applicable to all stock assessment methods, both biomass and age based, are used to compare model fits. These include an examination of model and prediction residuals using runs tests, estimation of prediction skill, and a comparison of production functions process error and stock status relative to reference points.

Diagnostic Procedure

Hindcasting (Kell et. al. 2016) was used to estimate prediction residuals and prediction skill. Hindcasting was used for model validation as it allows comparisons to be made across model structures and datasets. Validation examines if a model family should be modified or extended and is complementary to model selection and hypothesis testing. In comparison model selection searches for the most suitable model within a family, whilst hypothesis testing examines if the model structure can be reduced. This will be important as it will allow alternative model frameworks to be compared and valiadated in a working group setting

A hindcast is a multi-step prediction where a model is fitted using a tail-cutting procedure, where data are deleted from year *t−n* to *t* and then using the data from year *1* to *t-n-1* to make predictions of what will happen in years *t-n* to *t*. Hindcasting is a special form of cross validation and evaluates the predictive error of a model by testing it on a set of data not used in fitting. There is often insufficient data, however, in stock assessment datasets to allow some of it to be kept back for testing. A more sophisticated way to create test datasets is, like the jackkife, to leave out one (or more) observation at a time. Hindcasting also allows prediction residuals to be calculated, i.e. the difference between fitted and predicted values where the later is calculated from the out-of-sample predictions.

The expected (the form of the production function) and the stochastic dynamics (time series) will be compared across models, the latter will include an examination of process error. The uncertainty in stock status due to parameter and model uncertainty will be compared in the form of the Kobe phase plot.

Two other diagnostics are produced based on the fits to the index of abundance:

1 Runs test: This evaluates whether the data is randomly distributed around a central tendency without any systematic patterns.

2. Regime shifts: determines if there are systematic patterns do we observe this by regimes. We can then hypothesize if these patterns are driven by environmental changes or by changes in q (fishery driven).

Residual Analysis

Analysis of residuals is a common way to determine a model's goodness-of-fit (Cox and Snell, 1968). For example non-random patterns in the residuals may indicate model misspecification, serial correlation in sampling/observation error, or that heteroscedasticity is present. Various statistics exist to evaluate the residuals for desirable properties and nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test.

If the process of interest shows only random variation, the data points will be randomly distributed around the median. Random meaning that we cannot know if the next data point will fall above or below the median, but that the probability of each event is 50%, and that the data points are independent. Independence means that the position of one data point does not influence the position of the next data point, that is, data are not auto-correlated. If the process shifts, these conditions are no longer true and patterns of non-random variation may be detected by statistical tests.

Non-random variation may present itself in several ways. If the process centre is shifting due to improvement or degradation we may observe unusually long runs of consecutive data points on the same side of the median or that the graph crosses the median unusually few times. The length of the longest run and the number of crossings in a random process are predictable within limits and depend on the total number of data points in the run chart (Anhoej, 2014).

A shift signal is present if any run of consecutive data points on the same side of the median is longer than the prediction limit, round(log2(n) + 3). Data points that fall on the median do not count, they do neither break nor contribute to the run (Schilling, 2012). A crossings signal is present if the number of times the graph crosses the median is smaller than the prediction limit, qbinom(0.05, n - 1, 0.5) (Chen, 2010). n is the number of useful data points, that is, data points that do not fall on the median. The shift and the crossings signals are based on a false positive signal rate around 5% and have proven useful in practice.

Prediction Skill

The provision of fisheries management advice requires the assessment of stock status relative to reference points, the prediction of the response of a stock to management, and checking that predictions are consistent with reality. To evaluate uncertainty often a number of scenarios are considered corresponding to alternative model structures and dataset choices (Hilborn, 2016). It is difficult, however, to empirically validate the various stock assessment models as it is seldom possible to observe fish populations directly. Stock assessments, however, sometimes proven to be wrong in retrospect, due to poor model assumptions or to data that do not reflect the key processes (Schnute and Hilborn, 1993). Therefore techniques such as retrospective analysis, where a model is fitted to increasing periods of data to identify systematic inconsistencies (Mohn, 1999)

A key concept to understand therefore in the evaluation of prediction is the concept of *skill*. A prediction is said to have skill if it improves upon a naive baseline. A naive baseline is the predictive performance that you could achieve without really having any expertise in the subject. For instance, in weather forecasting a naive forecast could be the weather tomorrow will be the same as today. Prediction skill is a measure of the accuracy of a forecast compared to an observation or estimate of the actual value of what is being predicted (Huschke, 1959), and can be used to compare alternative models or observations to a reference set of estimates or data (e.g., Jin et al., 2008; Weigel et al., 2008; Balmaseda et al., 1995).

If data are regarded as being representative of the dynamics of the stock then they can be used as a model-free validation measure (Hjorth, 1993), and the best performing scenarios (e.g., choice of models and data) can be identified by comparing predictions with observations. In contrast quantities such as stock biomass and fishing mortality are model estimates not data and so cannot actually be observed so if estimates from a stock assessment model are compared this is model-based validation.

To evaluate prediction skill a one-step-ahead projection was performed and projetion residuals estimated for the indices of abundance. The first step is similar to a retrospective analysis where the final year(s) of data are deleted and the model refitted, The fitted model is then projected forward over the omitted years and predictions of the missing observations made. The difference between the observation and the predictions are the prediction residuals. If a model is overfitted the variance of the prediction residuals will be greater than the model residuals. While if non-stationarity is present, i.e. a regime shift has occurred, then the recent prediction residuals may show a pattern that is not evident in the model residuals. In this case the predictions were made for a single year, hence one-step-ahead.

Results

Indices of abundance Exploratory Data Analysis

The indices of abundance used in the assessment are shown in **Figure 2** by fleet and scenario; **Figure 3** summaries the abundance from 1979 onwards for the same indices. The trends in the indices are compared in **Figure 4** where a lowess smoother is fitted to year with series as a factor to show the average trend. The residuals from the lowess fit are summarised in **Figure 5**. Pattern are evident, e.g. in jr_r4 where the recent residuals are all positive which could be due to an increase in catchability or a change in the distribution of the stock. **Figures 6** and **7** repeat the analysis for each scenarios. aSimilar patterns are seen across the long line indices across the scenarios, however **Figures 3 & 4** show that there is no particular trend in the purse seine index. While **Figure 5** shows strong patterns and auto-correlation in the residuals from the loess fit to the indices.

Correlations between the stocks are shown in **Figure 8**, in the form of pairwise scatter plots. The estimates of SSB and biomass from the 2018 SS base case are included for comparison. The purse seine index shows poor correlation with the long line indices and the SS3 estimates (<0.24), while the correlations between the long line indices are greater than 0.5 and most are greater than 0.8, There are therefore no obvious data conflicts between the long line indices that require alternative data weightings to be explored for these indiecs.That the long line indices are positively correlated with the SS estimates of stock abundance, suggests that the biomass dynamic model inputs are consistent with the integrated assessment, and that there are no age or spatial effects that would require a more complicated structure than that assumed in the biomass dynamic model implementation.

The cross correlations between the base Case indices are plotted in F**igure 9** to identify potential lags due to year-class effects and differences in spatial distributions at age. The absence of lags for the long line indices confirms the conclusions with respect to the appropriate use of a biomass dynamic without regional structure. The lags in the case of the purse seine indices indicates that the index does not represent the same components of the stock as the long line index. Due to the conflict between the purse seine and long line indices two scenarios were developed namely; i) the base case, where all indices were included; and ii) a scenario where only the purse seine index was fitted.

Discards (numbers) are summarised in F**igure 10** by region from the Taiwanese commercial logbook based on scenario (2) which excluded vessels that did not report any discards (IOTC, 2019b).

Reference Case

The two reference cases (where the biomass model is fitted to the SS estimates of SSB and biomass) are summarised in **Figures 11, 12 and 13**. **Figure 11** summarises the trends of catch, SSB and harvest rate; **Figure 12** presents phase plots with production functions; and **Figure 13** presents the time series of process error where blocks correspond to regimes, estimated by the STARS (Sequential T-test Analysis of Regime Shifts) method, known also as the Rodionov test (Rodionov 2004).

Indicators suggest that the stock is overfished and experiencing overfishing if fitted to SSB but not if fitted to biomass estimated by SS3 (**Figure 12**). The variability in process error is greatest when fitted to biomass (**Figure 13**).

Model Fitting Diagnostics

First the model fits are reviewed by examination of the model residuals using run plots (**Figures 14** and **15**) and regime plots (**Figures 16** and **17**) for model and prediction residuals.

The runs tests (**Figure 14**) indicate that the model fits are poor in most cases other than region 3 vessel effects as the residual patterns are systematically under or overpredicting. **Figure 15** for the prediction residuals shows similar patterns. When the purse seine index is included in the assessment scenarios adds no information, therefore and that there are conflicts with the long line index, therefore the hindcast was not run for the scenarios that include the purse seine index. The prediction residuals indicate that regime shifts are probably occurring around 1990, corresponding to large scale global changes e.g. the strong ENSO event of 1990-1991 that effected the Indian Ocean (Francis Marsac, IFREMER, Sete, Pers. Com).

Both model and prediction residuals using the STARS algorithm determines that there are 3 states/regimes that can be indicated by both the model residuals (**Figure 16**) and prediction residuals (**Figure 17**). Prior to 1990's, through the 1990's and early 2000's to recent years. The patterns vary by region, but in most cases are shown to be stable in regions 1 and 3, increasing positive pattern in region 2, and have a stronger negative pattern in region 4, perhaps suggesting a shift in distribution in those 2 regions from one to the other (see Figure on Regional distribution).

Note that assessment areas are only 4, where CPUE's from Area 2 represent Area 1 (left panel) or Area 7 (Area 2) represent Area 1. Area 3 in the CPUE standardization is Area 2 in the Assessment, Area 4 in CPUE standardization is Area 3 in the Assessment and Area 5 in CPUE standardization is Area 4 in the Assessment.

The implications of this is a probable change in distribution or q getting lower in region 2, and higher in region 4 for the assessments, implying that catchability is changing either due to fishing changes or environmental changes in 2 and 4.

Assessment Results

The phase plots with the production functions are presented in **Figure 18,** (for purse seine estimates of stock size and carrying capacity are large and so the production function is truncated in the panel) the trends in biomass in **figure 19** and time series of current biomass and fishing mortality with respect to reference point is shown in **Figure 20** and **21.** Time series of process error with regimes in **Figure 22**.

As there is little signal in the purse seine index the assessment assumes that the stock is large and is being mined, **Figures 18 and 19.**

A retrospective with projection for SSB is presented in **Figure 22**. A plot showing all results examined on the phase is shown in **Figure 24**, with temporal trajectories with different regimes as a function of Kobe or Majuro based reference points are shown in **Figure 25** and **26**.

Stock status and reference points are summarised in **Table 1.**

We focus on **Figures 19, 20 and 21** where temporal trends indicate a sever decline in overall abundance, and increase in fishing mortality. In all cases, F is greater than F_{MSY} reference point (**Figure 21**) and in only 2 cases B is greater than the reference point B_{MSY} when r, i.e. intrinsic productivity is greater than 1, and when we follow the Biomass trajectory from the last SS3 assessment to force the model to behave similar to the synthesis trajectory. Regardless, of scenario all models show a decline in abundance in recent years with respect to initial estimates. The nature of the time series is similar for scenario 2 and 4, but differs quite substantially from scenario 3 and 5. Scenario 1 is also quite different from the other scenarios.

Regime shifts appear to have occurred in all regions similar to those described above in the diagnostics. Howver, the splits are more apparent to be before the late 90's, 90's to 2008/2009 and from 2009 onwards with regard to the scenarios examined. In the diagnostics, above the regions where the model and prediction residuals varied by area where lagged in different blocks (prior to 90's, till mid 2000's and after that). The retrospective analysis indicates all Regions are fairly stable other than Region 4 in 2017 (**Figure 23**). Reference model is systematically overestimating results from the hindcast.

Discussion

Based on this examination using a Bayesian surplus production model with process error, the here we summarise the main points:

With regard to the CPUE data inputs:

- Plots of correlation and cross-correlation between indices showed no apparent structure due to age or spatial effects that could not be taken into account by the biomass dynamic assessment model used.
- There is no systematic pattern across different regions and essentially all areas indicate a similar declining trend with or with size sorting issues.
- Diagnostics with respect to CPUE series indicated that all indices were highly correlated, and that the indices were also positively correlated with the SS estimates of stock abundance, which suggests that the biomass dynamic model inputs are consistent with the integrated assessment. The absence of lags supports the use of a biomass dynamic without regional structure.

With regard to the Assessment the following could be concluded:

- Model diagnostics indicate regime shifts for a majority of the regions and scenarios examined suggesting non-stationarity of the stock recruit dynamics.
- For indices with more than seven observations the runs test was only passed for index jr_r3ves for scenarios cpueS3, cpueS4 and cpueS5. The other long indices showed systematic patterns, i.e. jr2 showed positive residuals in the recent period and jr_r4 negative residuals in the last two decades when all the series was included. This could suggest that an increase in catchability or highgrading.
- Prediction residuals for jr_r4 and jr_r4ves were always negative and for jr_jr2 and jr_jr2ves positive.
- Prediction skill was poor over three years for all scenarios other than cpueS3, scenario cpueS4 had fitting problems, and prediction skill was poor over 10 years coinciding with a change in process error.
- The only runs that are above BMSY are high sp and SS bio. all other runs F>FMSY and B<BMSY.

Conclusions

In all likelihood, the stock is currently being overfished and is likely experiencing overfishing. The methods developed here provide an objective methodology to assess model performance. Our results indicate that there are non-stationary dynamics that probably effect the distribution, catchability and stock recruit function. Regardless of data series examined, the weight of evidence suggests that the stock is experiencing overfishing and is likely overfished.

Acknowledgements

This work was supported by The Sustainable Indian Ocean Tuna Initiative (SIOTI) a large-scale Fisheries Improvement Project (FIP) comprising the major purse seine fleets and tuna processors in the region. The FIP is supported by Seychelles and WWF,

References

Anhoej, J. (2015).. PLoS ONE 10(3): e0121349.

Balmaseda, M.A., Davey, M.K., Anderson, D.L., 1995. Decadal and seasonal dependence of ENSO prediction skill. J. Clim. 8 (11), 2705–2715.

Chen, Z. (2010). A note on the runs test. Model Assisted Statistics and Applications 5, 73-77

Cox, D.R. and Snell, E.J., 1968. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(2), pp.248-265.

Geehan, J. and Hoyle, S., 2013. Review of length frequency data of the Taiwanese distant water longline fleet. *IOTC Proceedings. IOTC, San Sebastian, Spain*, pp.23-28.

Hjorth, J.U., 1993. Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap. CRC Press.

Huschke, R.E., 1959. Glossary of meteorology. American Meteorological Society

IOTC–WPTmT07(DP) 2019. Report of the Seventh Session of the IOTC Working Party on Temperate Tunas. Kuala Lumpur, Malaysia, 14–17 January 2019. IOTC–2019–WPTmT07(DP)–R[E]: 43 pp.

IOTC, 2019b. Report of the Sixth IOTC CPUE Workshop on Longline Fisheries. San Sebastian, April 28th – May 3rd, 2019.

Jin, E.K., Kinter III, J.L., Wang, B., Park, C.-K., Kang, I.-S., Kirtman, B., Kug, J.-S.,Kumar, A., Luo, J.-J., Schemm, J., et al., 2008. Current status of ENSO prediction skill in coupled ocean-atmosphere models. Clim. Dyn. 31 (6), 647–664.

Kell, L.T., Kimoto, A. and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. *Fisheries research*, *183*, pp.119-127.

Methot Jr, R.D. and Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, *142*, pp.86-99.

Mohn, R., 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56 (4), 473–488.

Rodionov, S.N., 2004. A sequential algorithm for testing climate regime shifts. Geophysical Research Letters, 31:L09204, doi:10.1029/2004GL019448.

Schilling, M. F., 2012. The Surprising Predictability of Long Runs. Math. Mag. 85, 141-149

Schnute, J.T., Hilborn, R., 1993. Analysis of contradictory data sources in fish stock assessment. Can. J. Fish. Aquat. Sci. 50 (9), 1916–1923.

Weigel, A., Liniger, M., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134 (630), 241–260.

Winker, H., Carvalho, F. and Kapur, M., 2018. JABBA: just another Bayesian biomass assessment. *Fisheries Research*, *204*, pp.275-288.

Tables

Table 1. Stock status in 2018 and reference points (we should add the 2 cases of PS just for illustration).

A

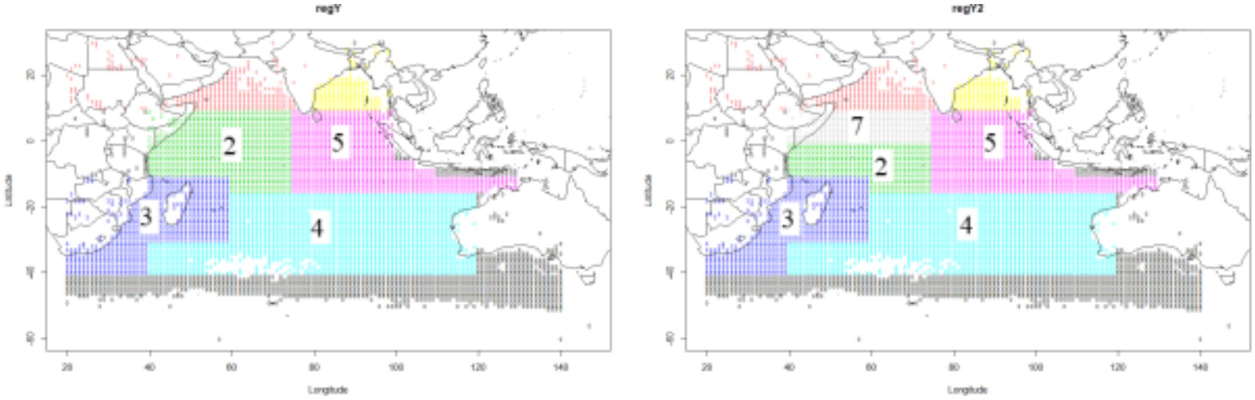

Figure 1 Regional structure

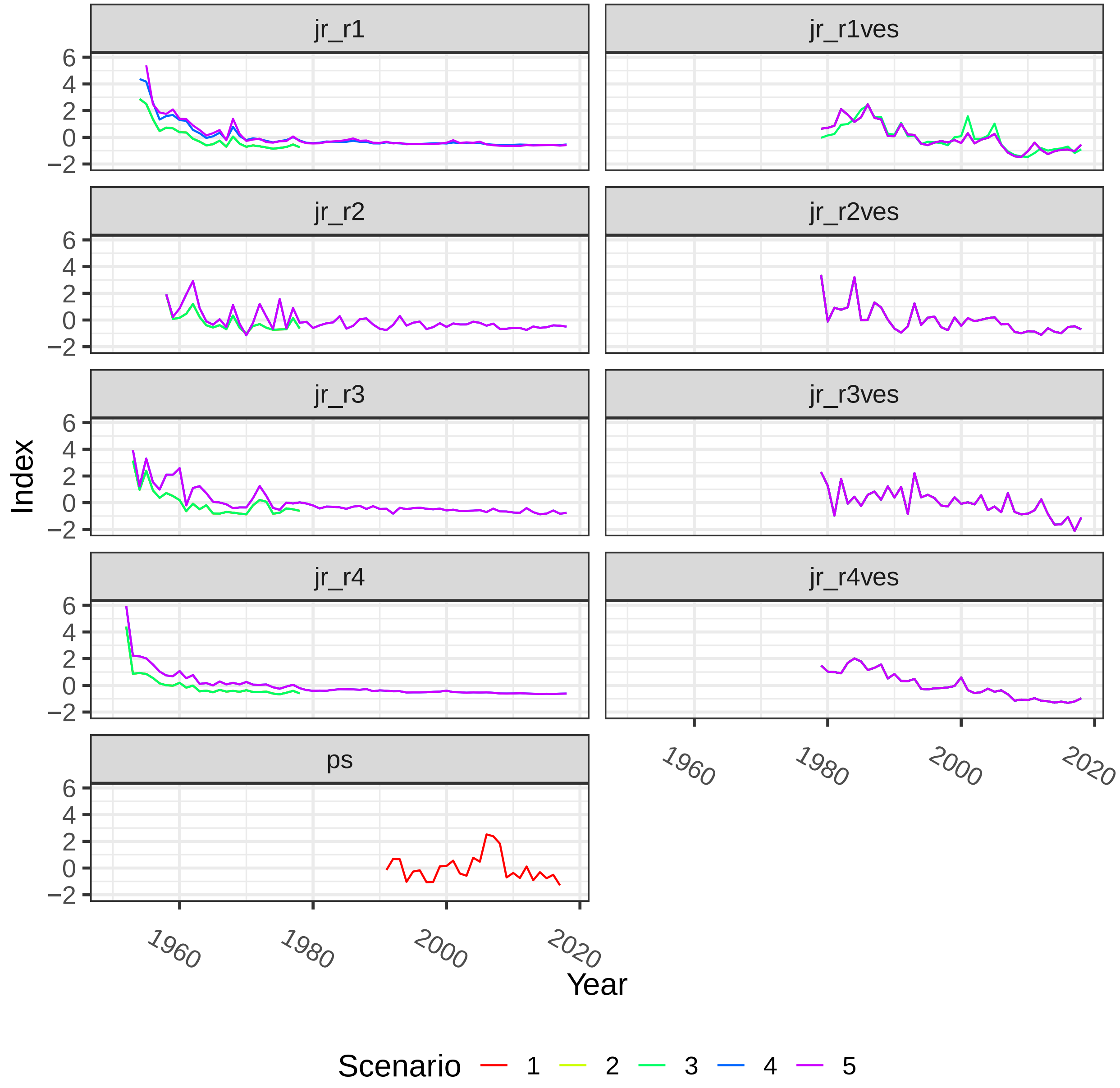

Figure 2 Catch per unit effort indices of abundance by fleet and scenario.

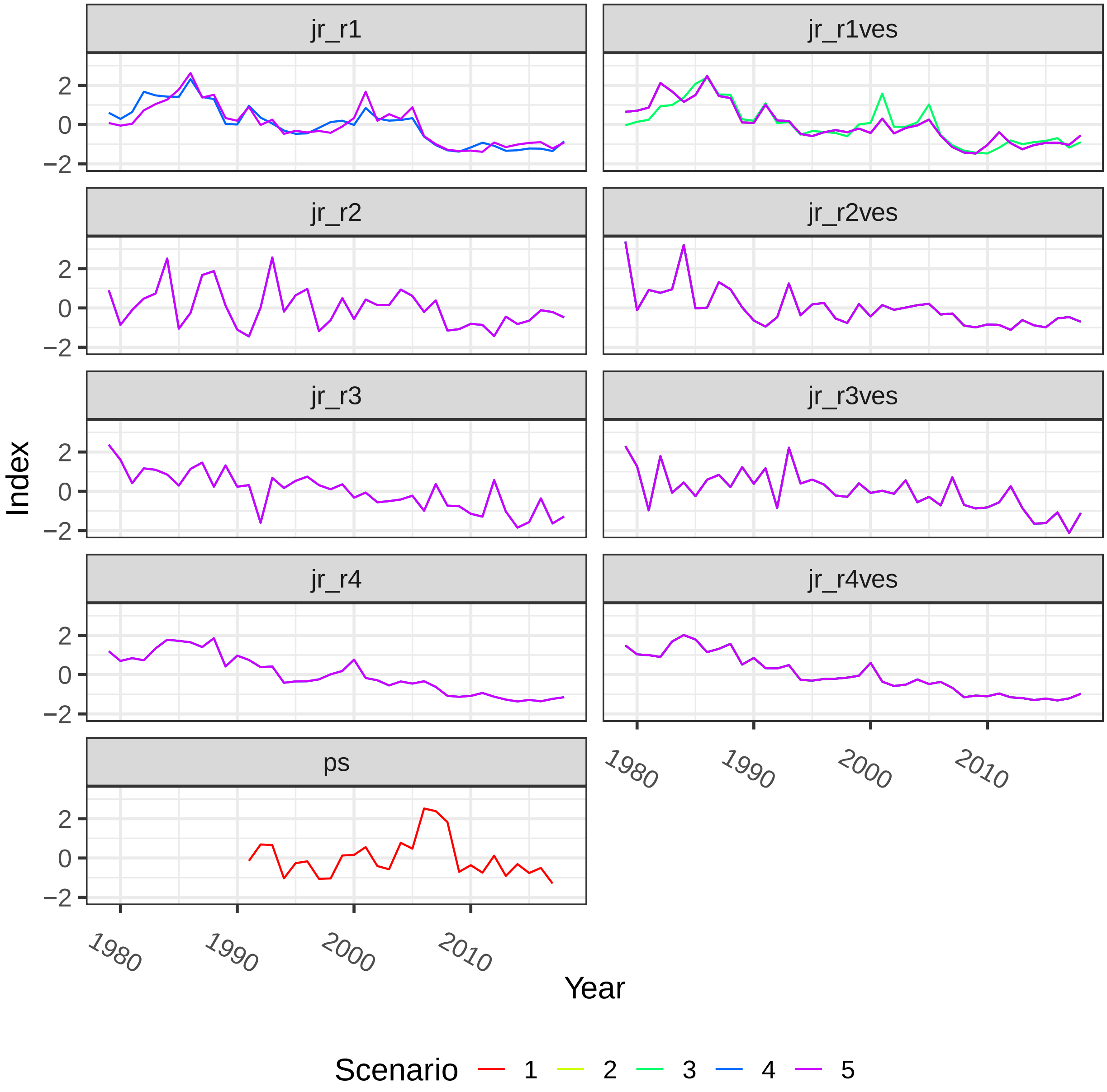

Figure 3 Catch per unit effort indices of abundance from 1979 onwards by fleet and scenario.

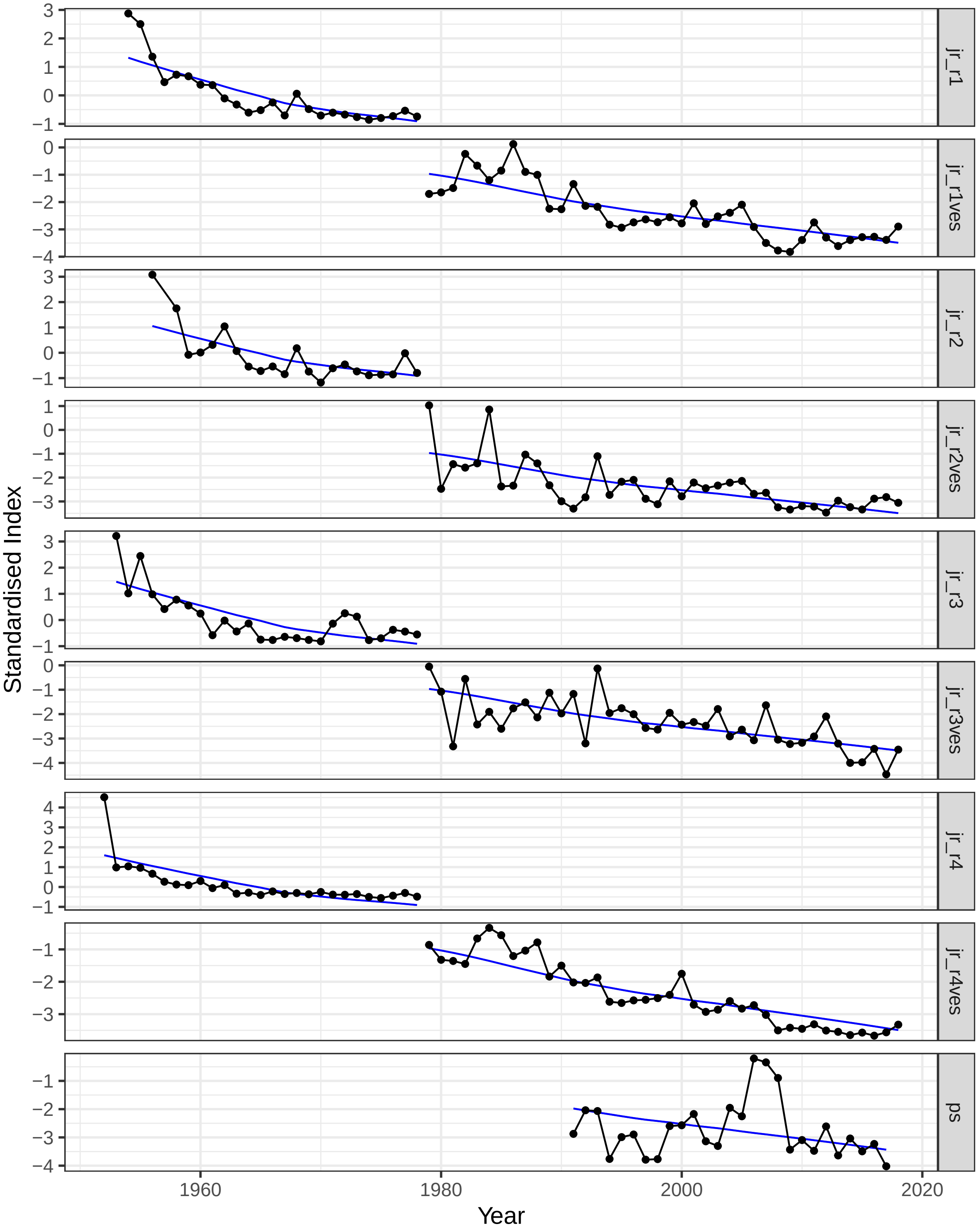

Figure 4 Time series of Base Case CPUE indices, continuous blue line is a lowess smother showing the average trend by area (i.e. fitted to year with series as a factor).

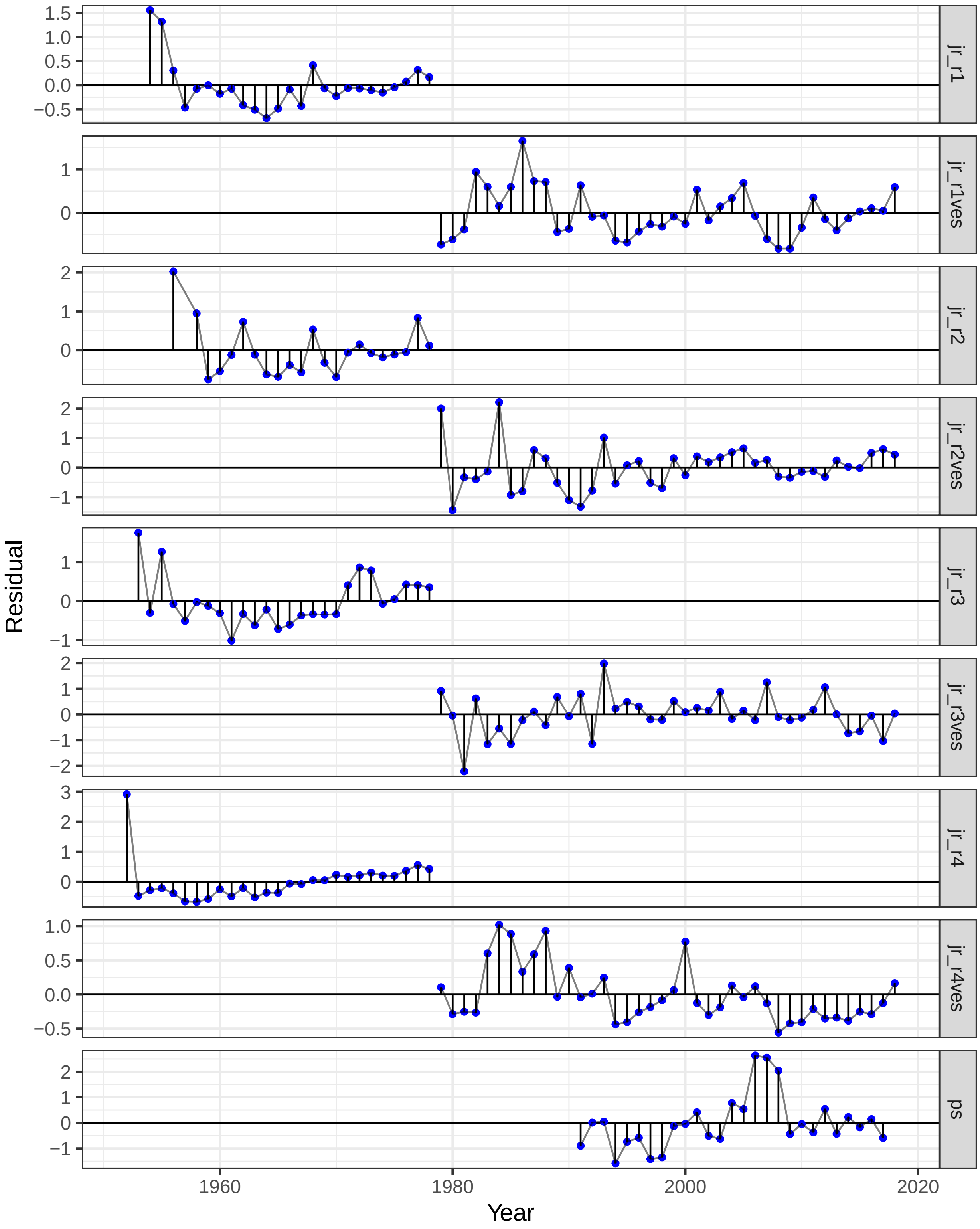

Figure 5 Residuals from the lowess fit to the Base Case.

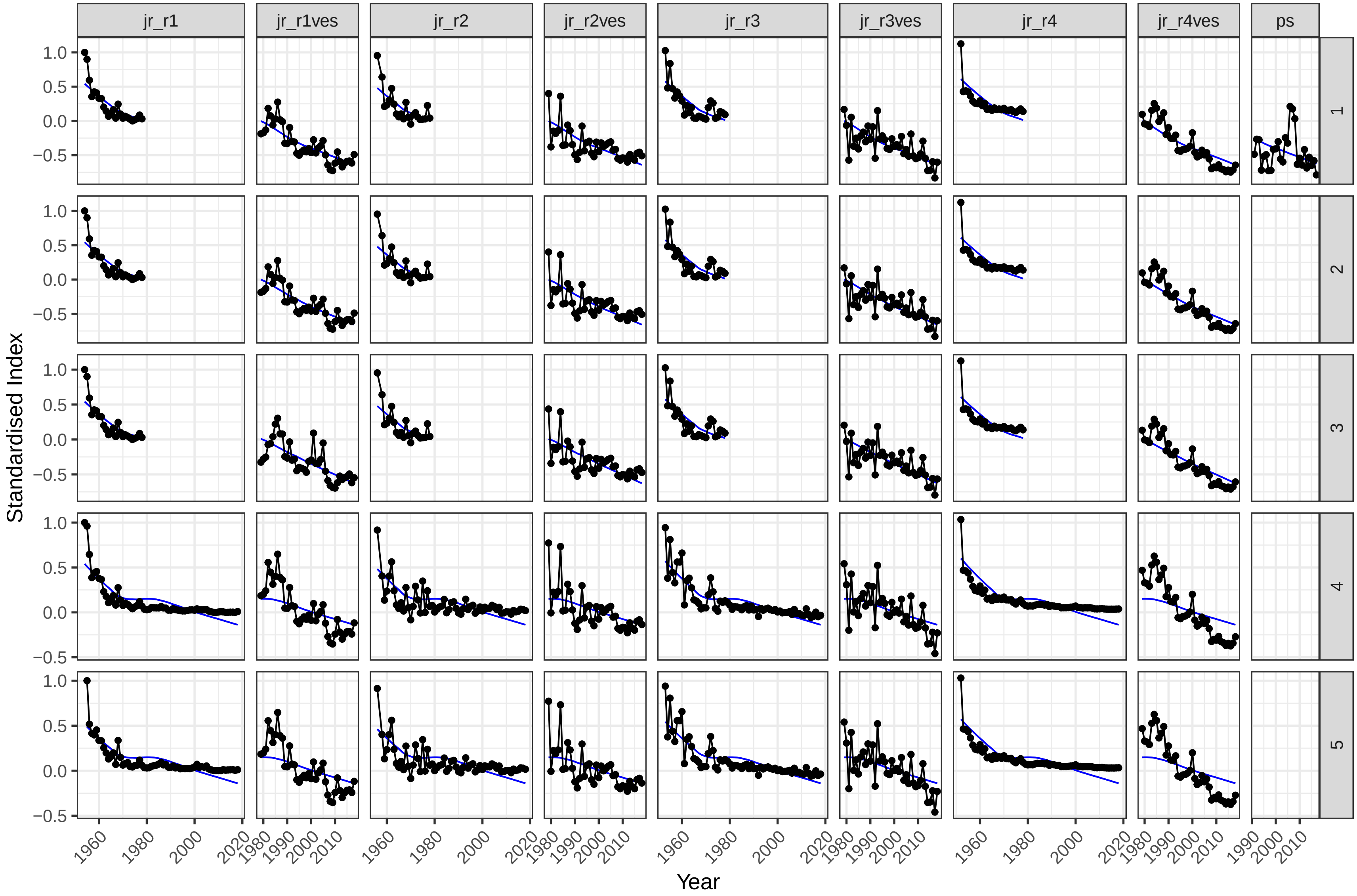

Figure 6 Time series of Base Case CPUE indices, continuous blue line is a lowess smother showing the average trend by area (i.e. fitted to year with series as a factor).

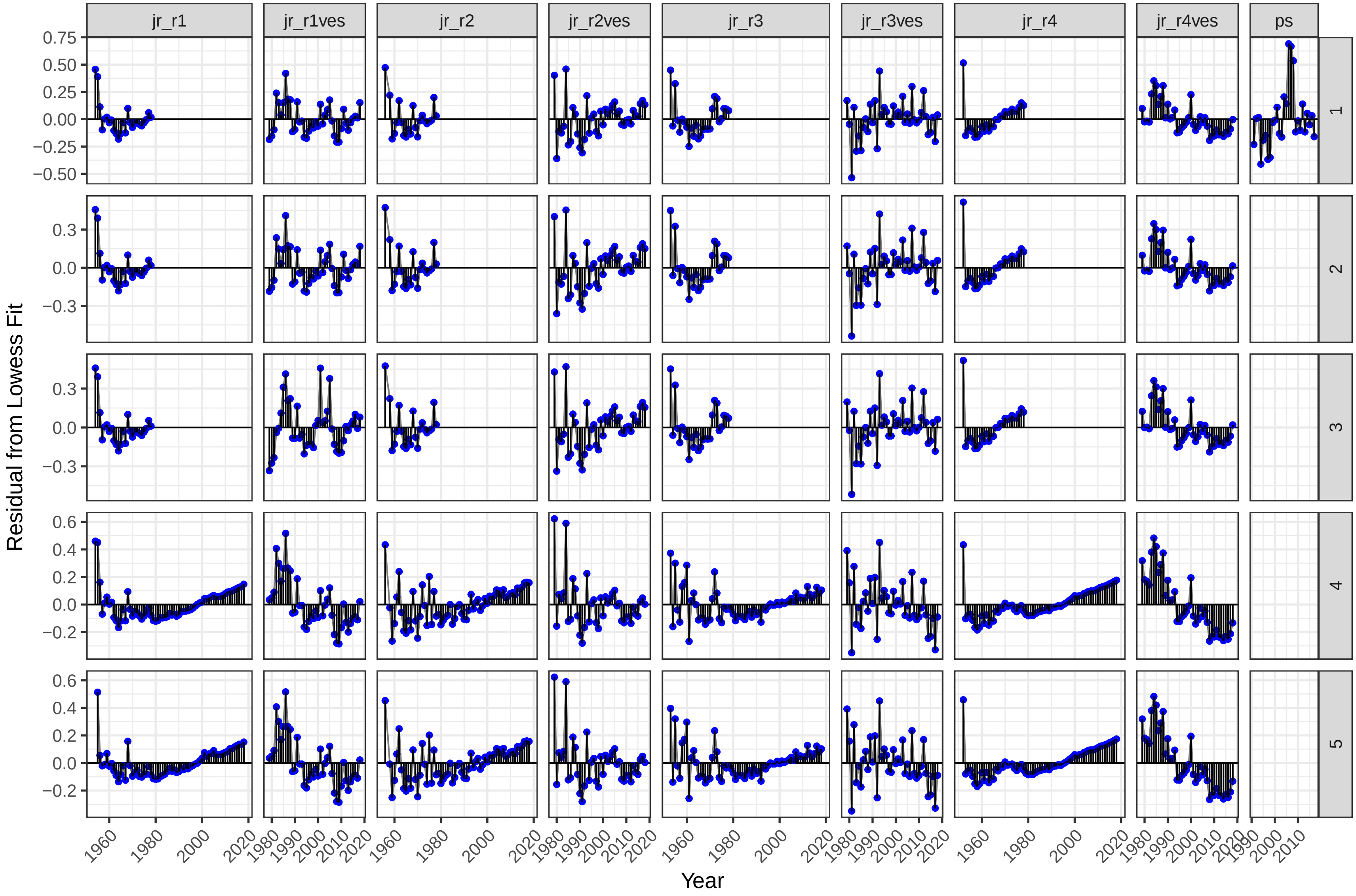

Figure 7 Residuals from the lowess fit to the Base Case.

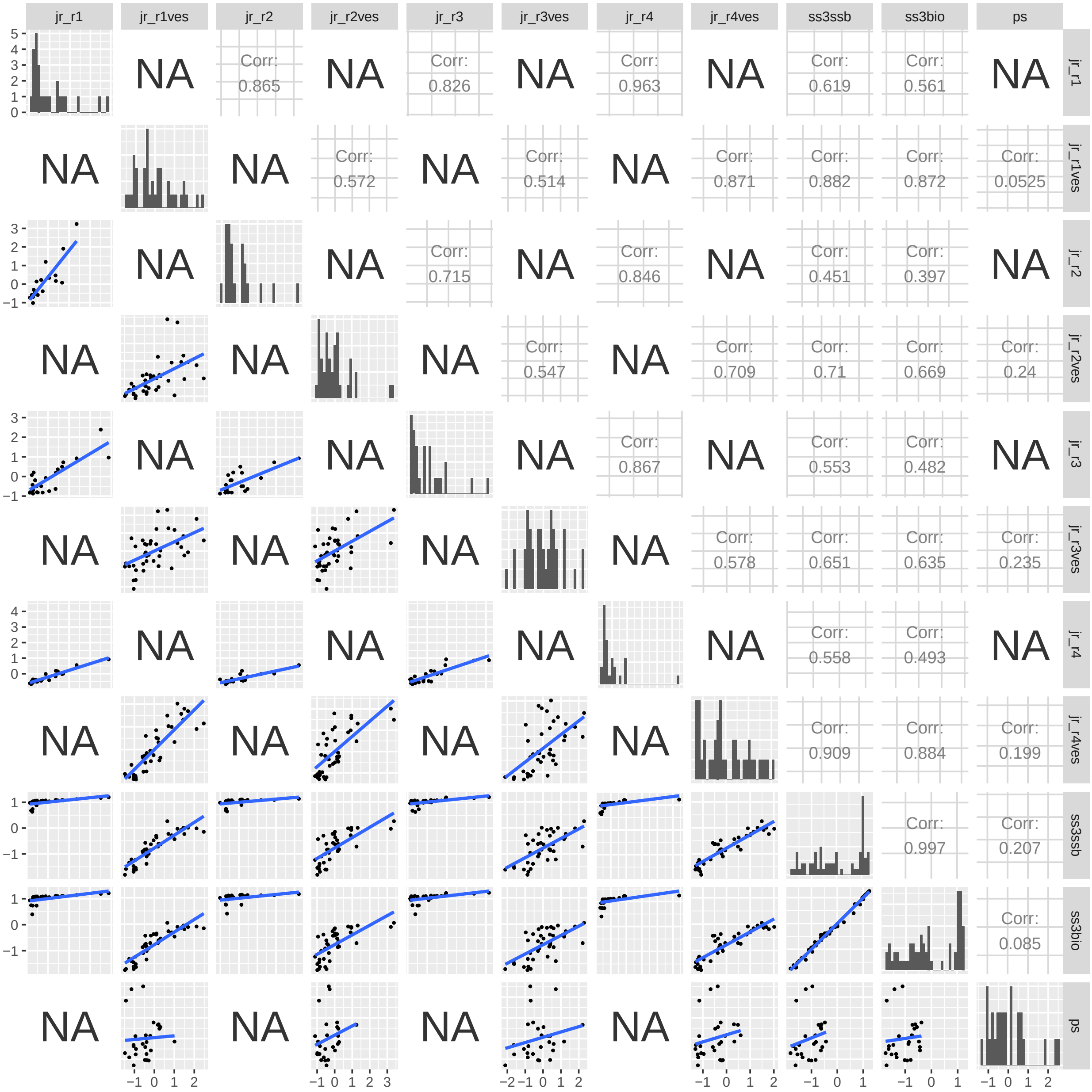

Figure 8 Pairwise scatter plots to look at correlations between Base Case Indices.

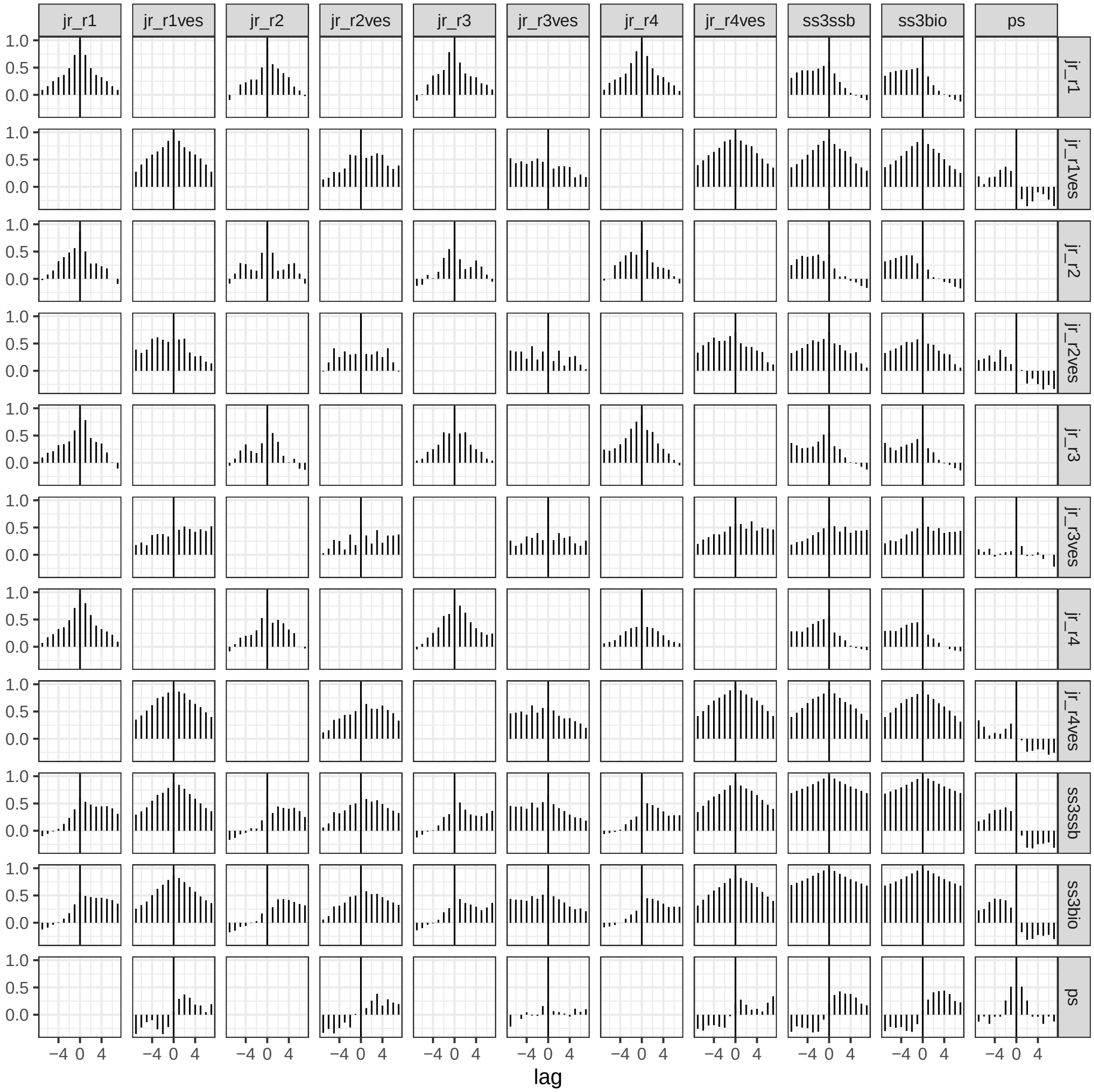

Figure 9 Cross correlations between Base Case indices, to identify potential lags due to year-class effects.

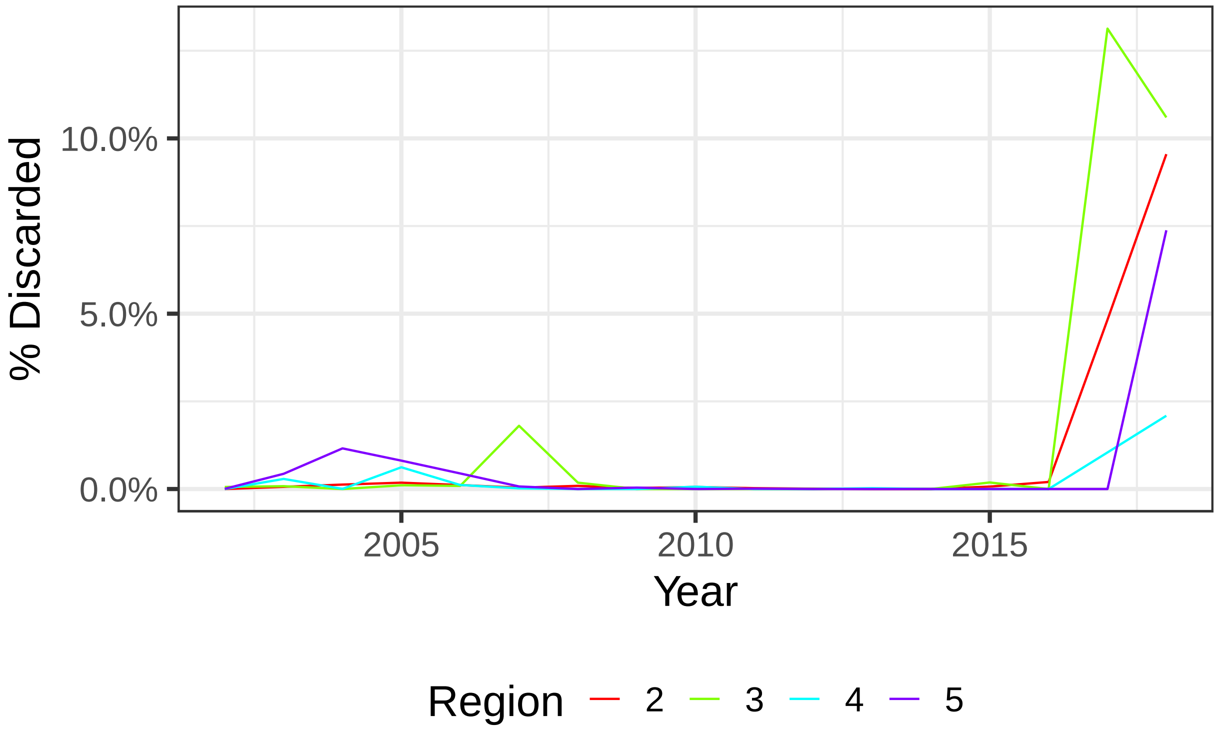

Figure 10 Summary of discards (numbers) by region from the Taiwanese commercial logbook based on scenario (2) which excluded vessels that did not report any discards (Report of the Sixth IOTC CPUE Workshop on Longline Fisheries San Sebastian, April 28 th – May 3 rd , 2019).

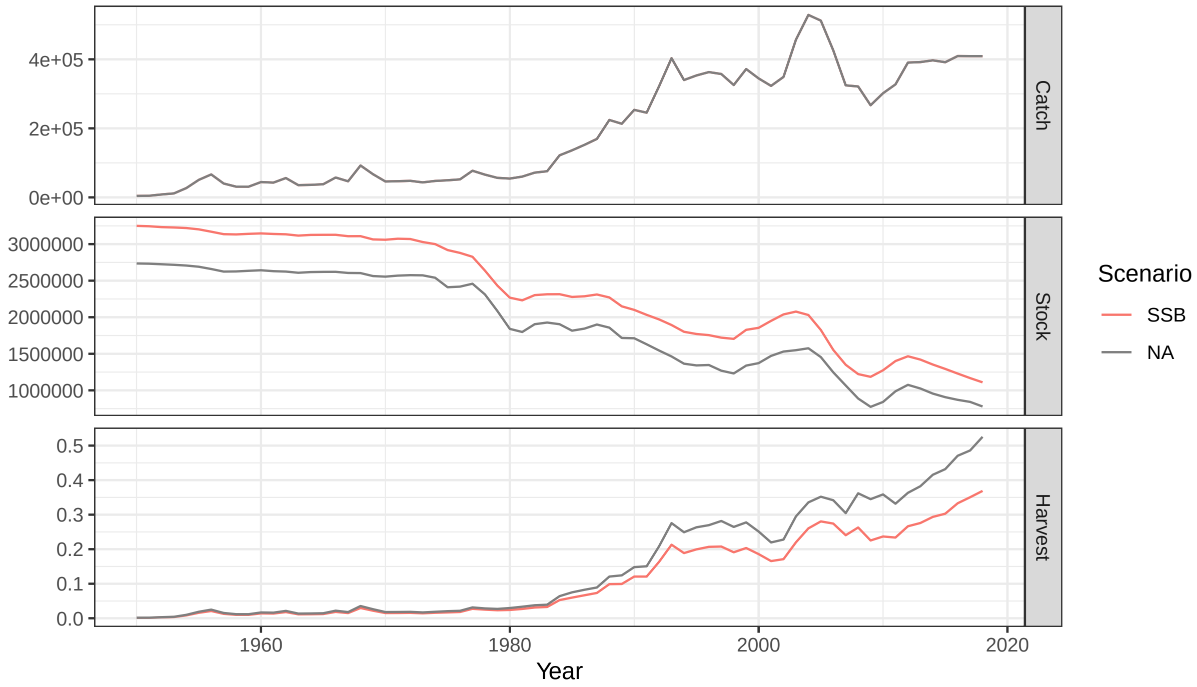

Figure 11 Trends of catch, SSB and harvest rate in the reference set and Base Case Jabba assessments. Trends for the reference set Jabba assessments.

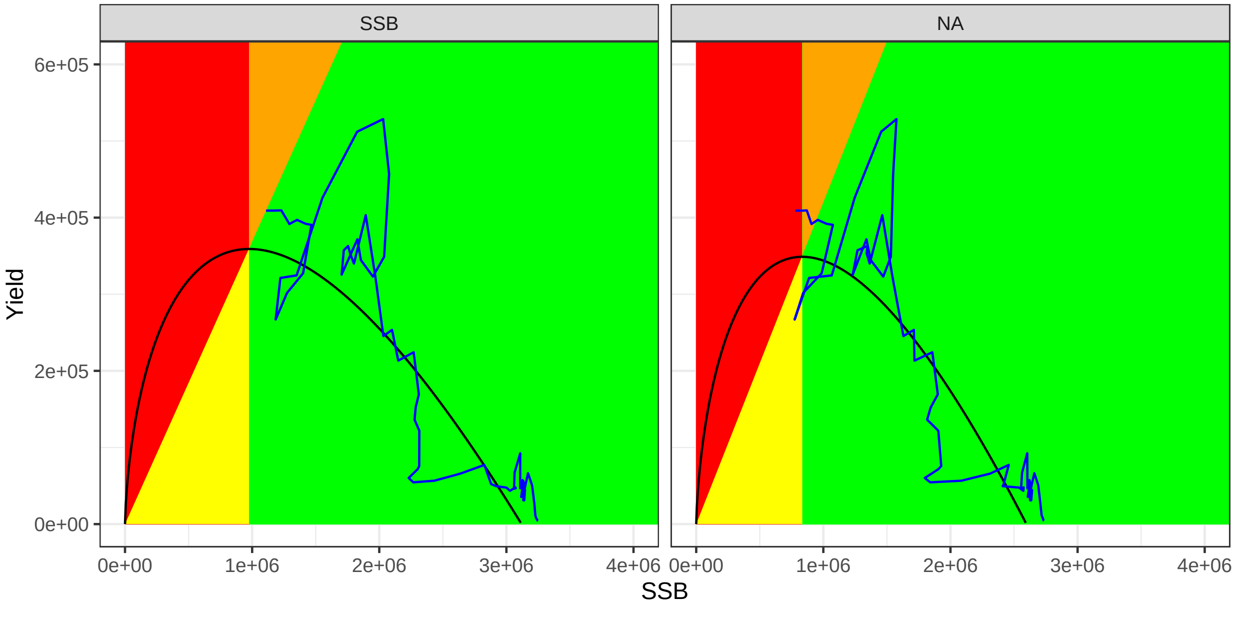

Figure 12 Phase plots with production functions for Jabba reference sets and Base Case assessment. Phase plots for reference sets.

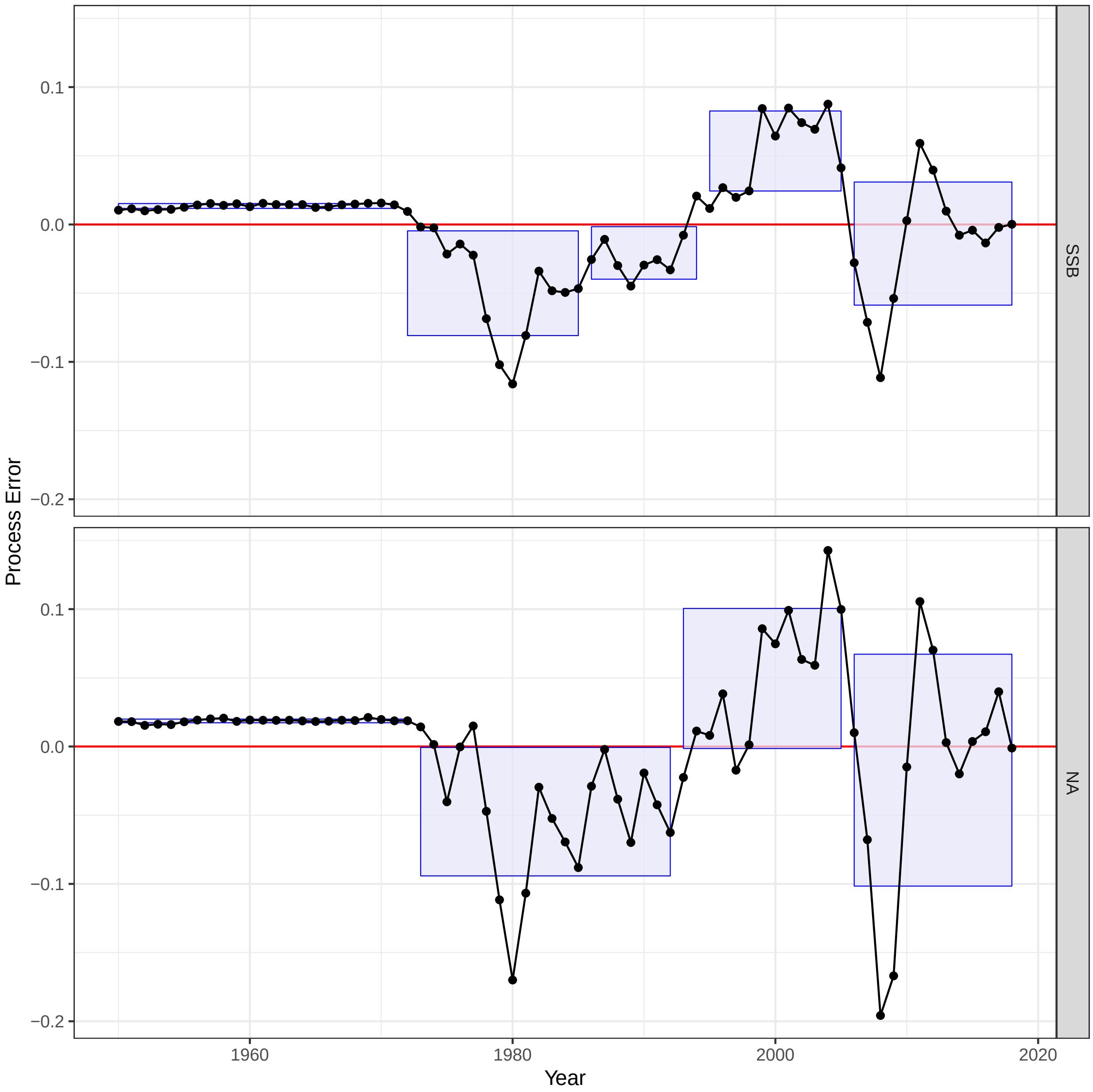

Figure 13 Time series of process error for Jabba reference sets, blocks correspond to regimes identified by the STARS algorithm for Jabba reference sets and Base Case assessment.

Process error with regimes

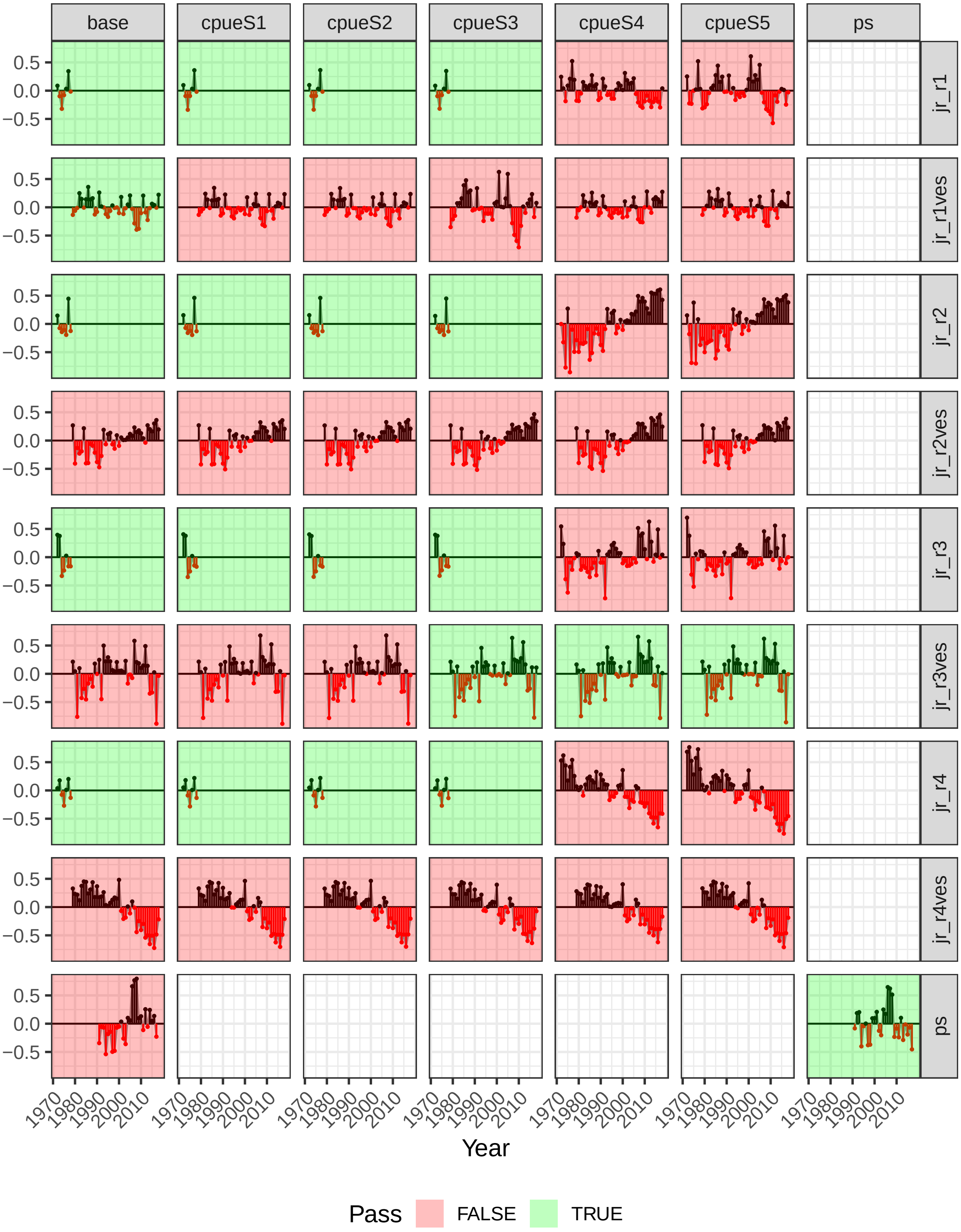

Figure 14 Model Residuals, red background indicates unusually long runs or unusually few crossings.

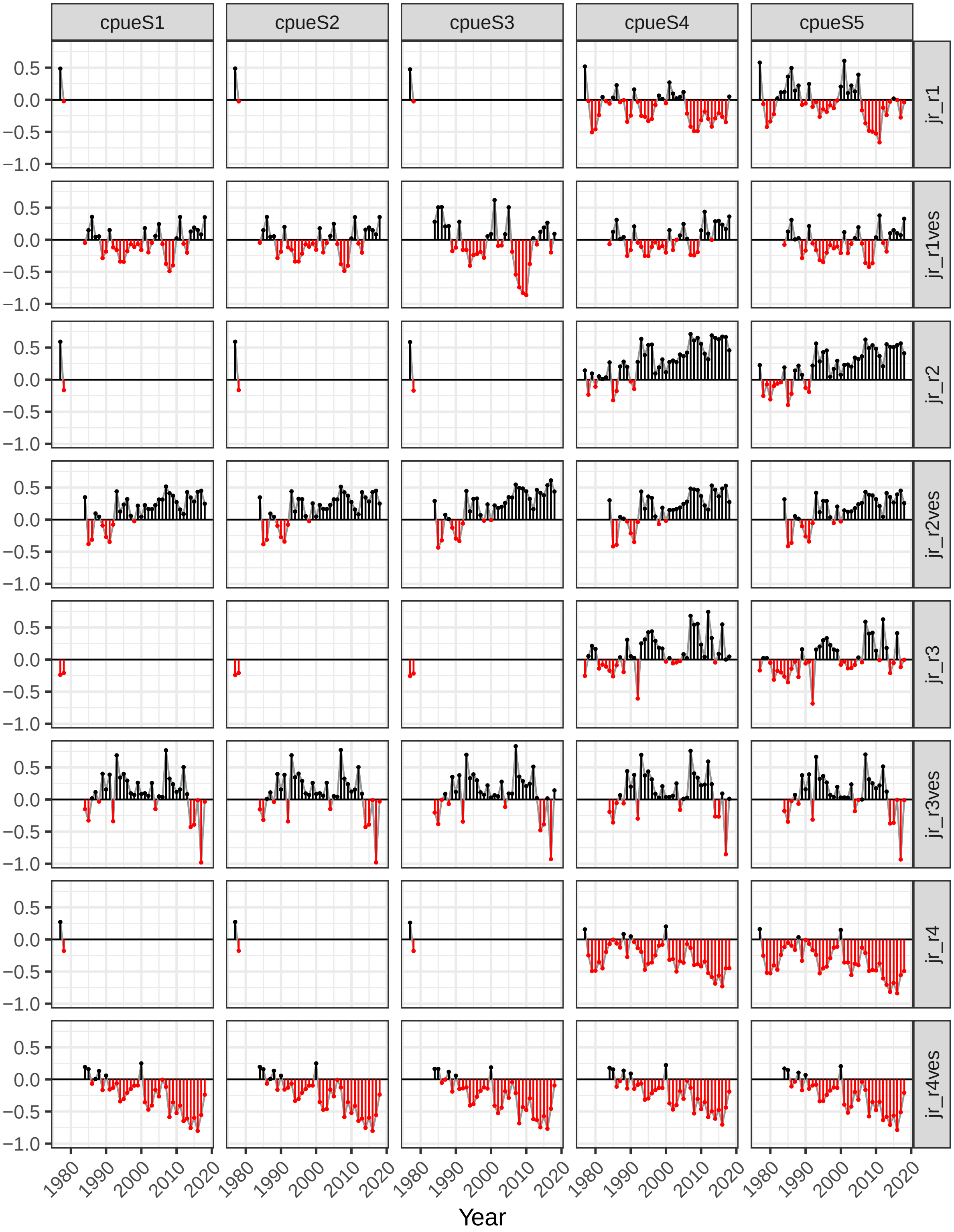

Figure 15 Prediction residuals.

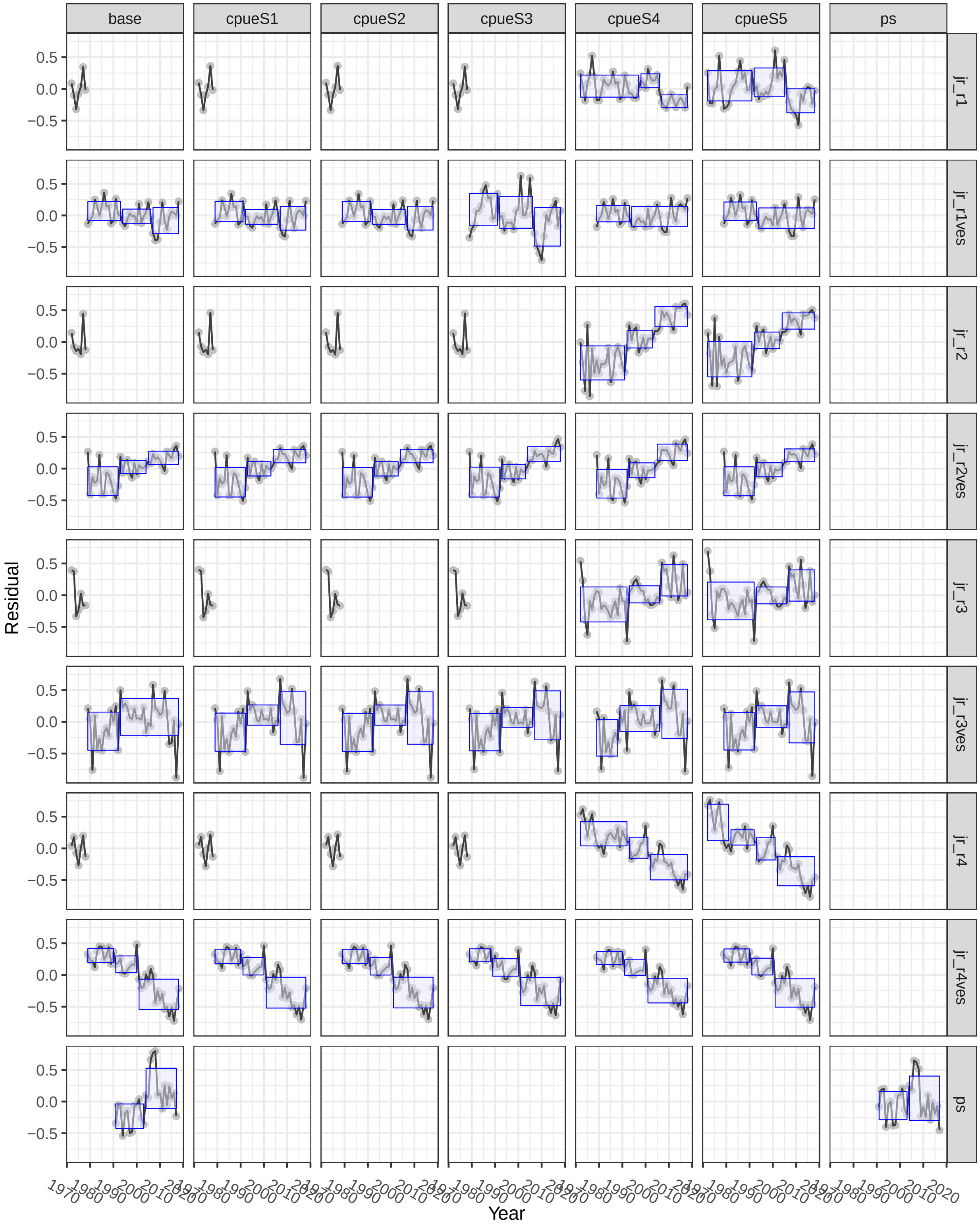

Figure 16 Regime shifts using STARs algorithm for model residuals.

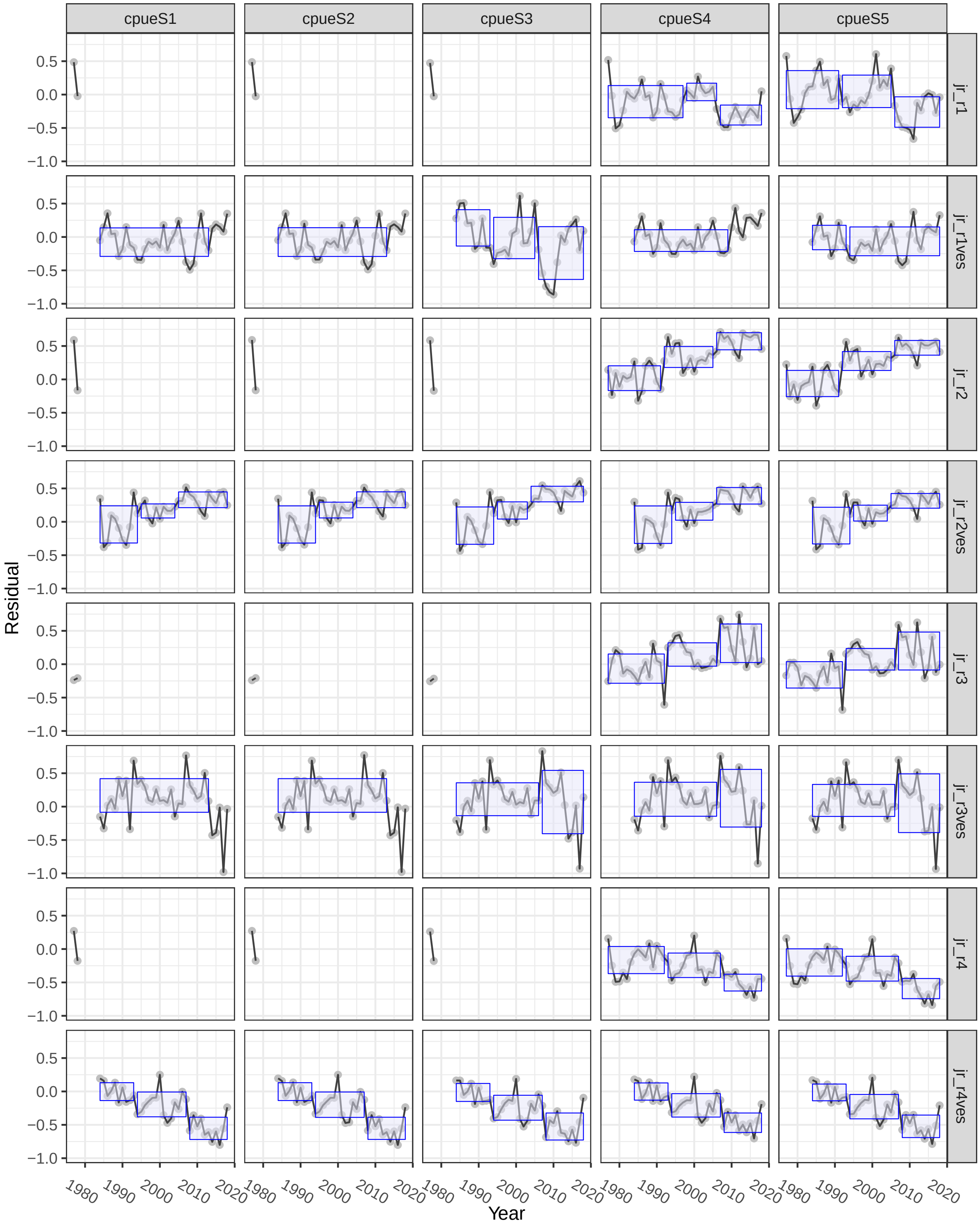

Figure 17 Regime shifts using STARs algorithm for prediction residuals.

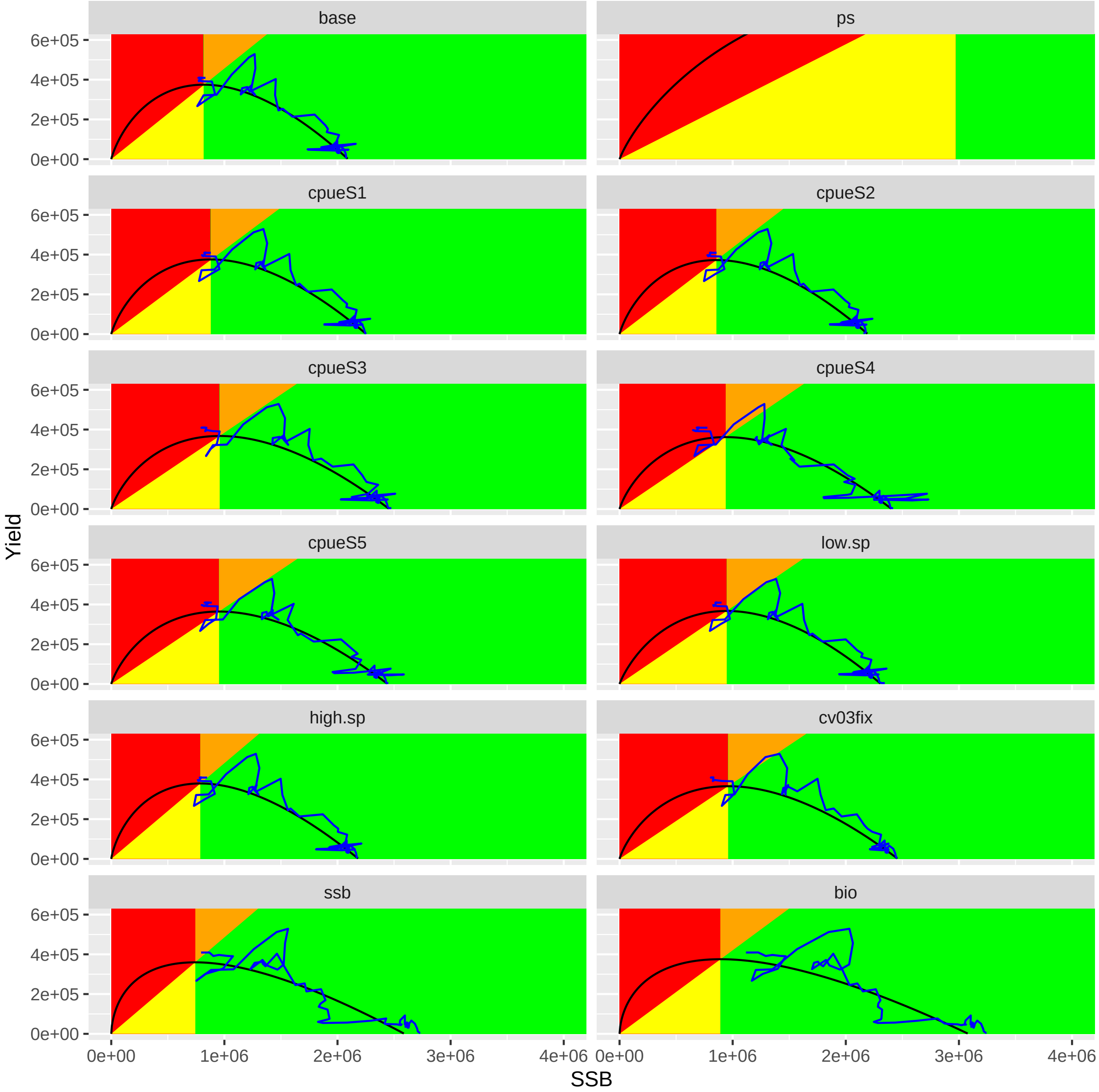

Figure 18 Phase plots.

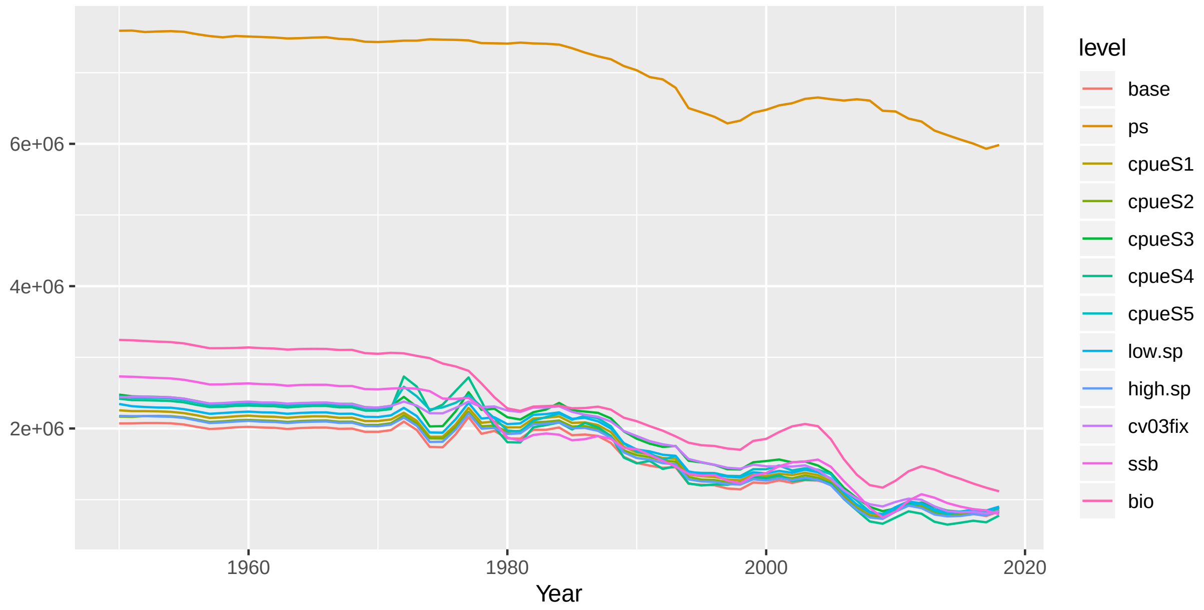

Figure 19 Trends in SSB.

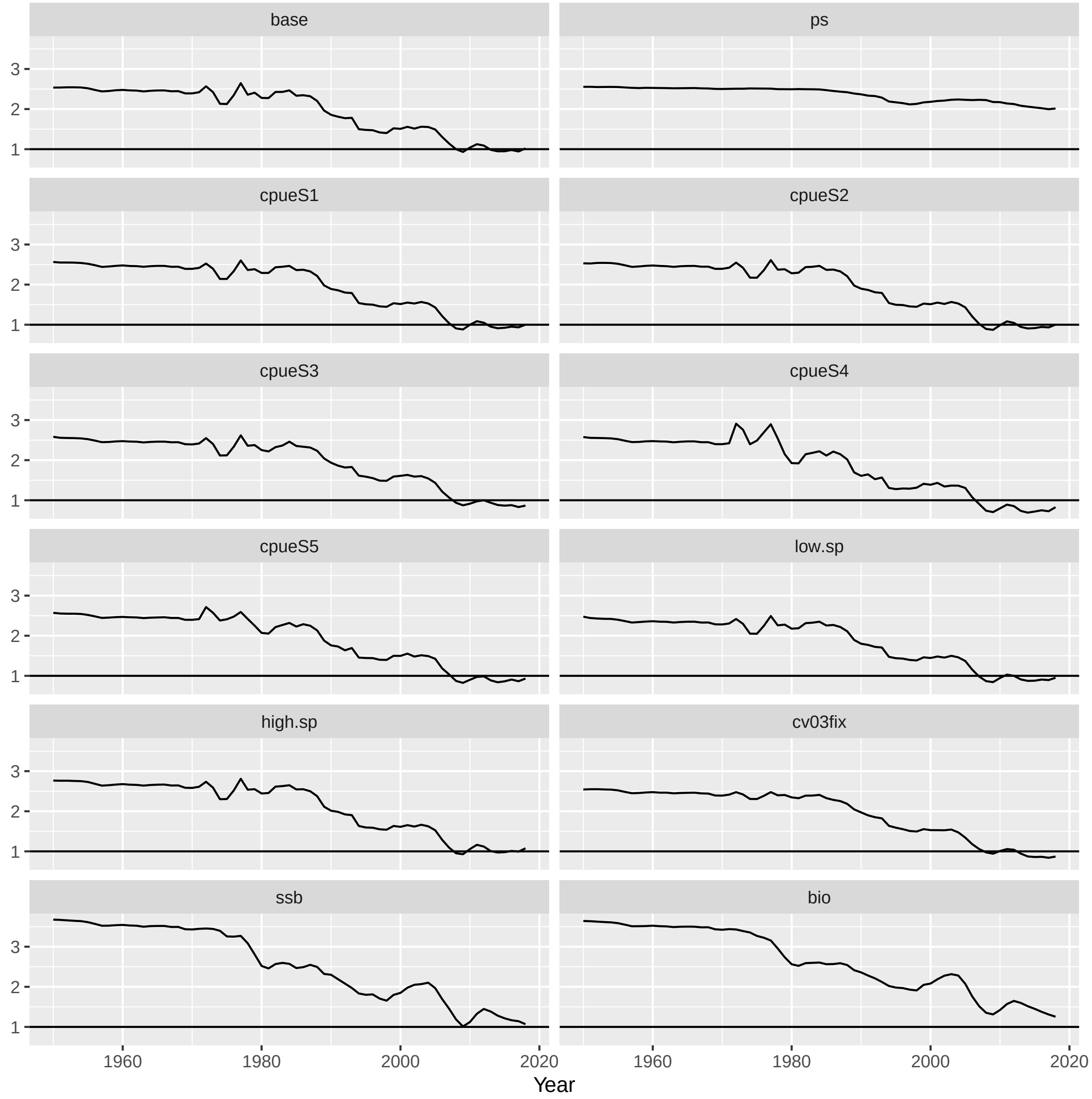

Figure 20 Trends in SSB/Bmsy.

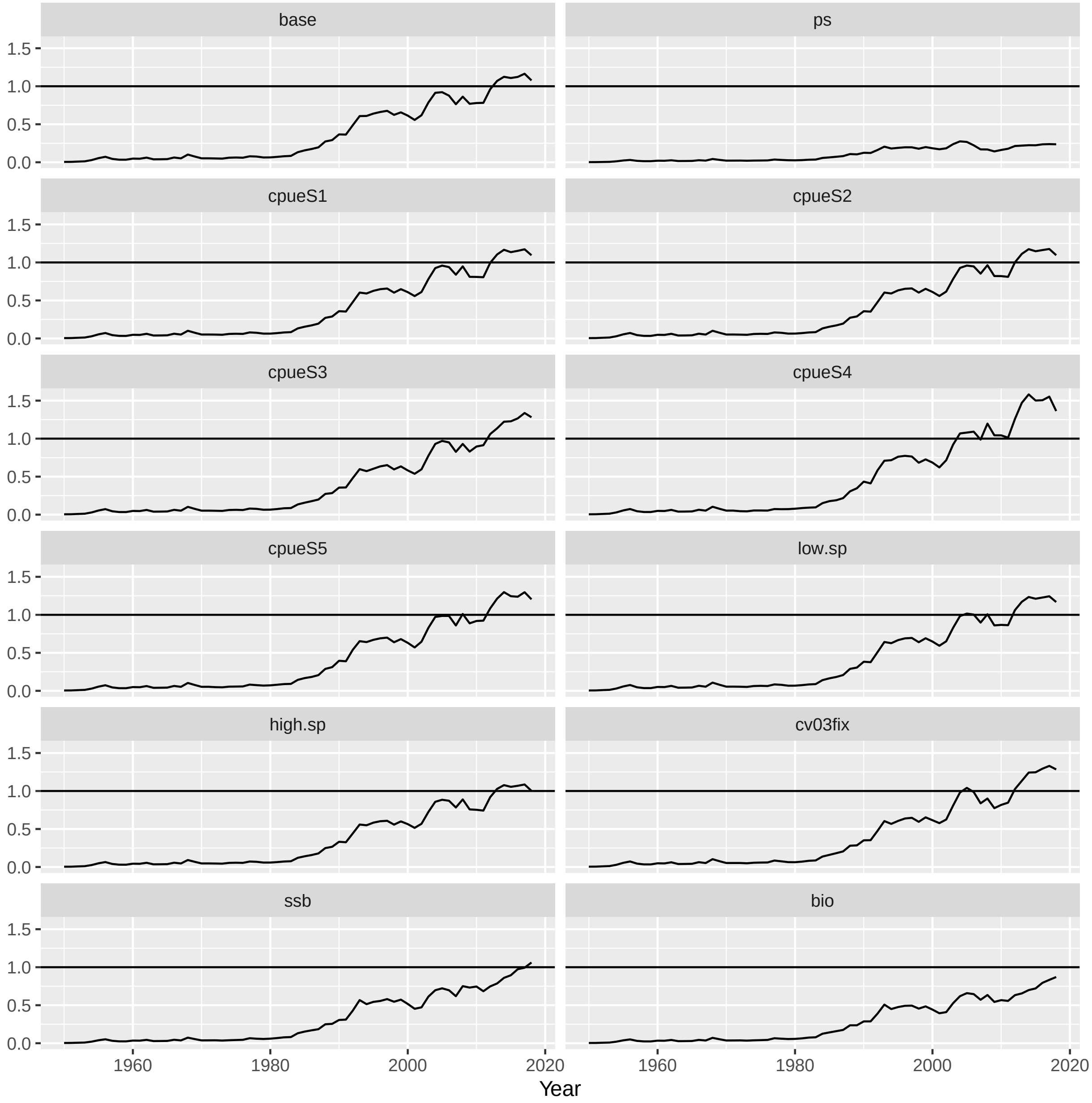

Figure 21 Trends in F/Fmsy$.

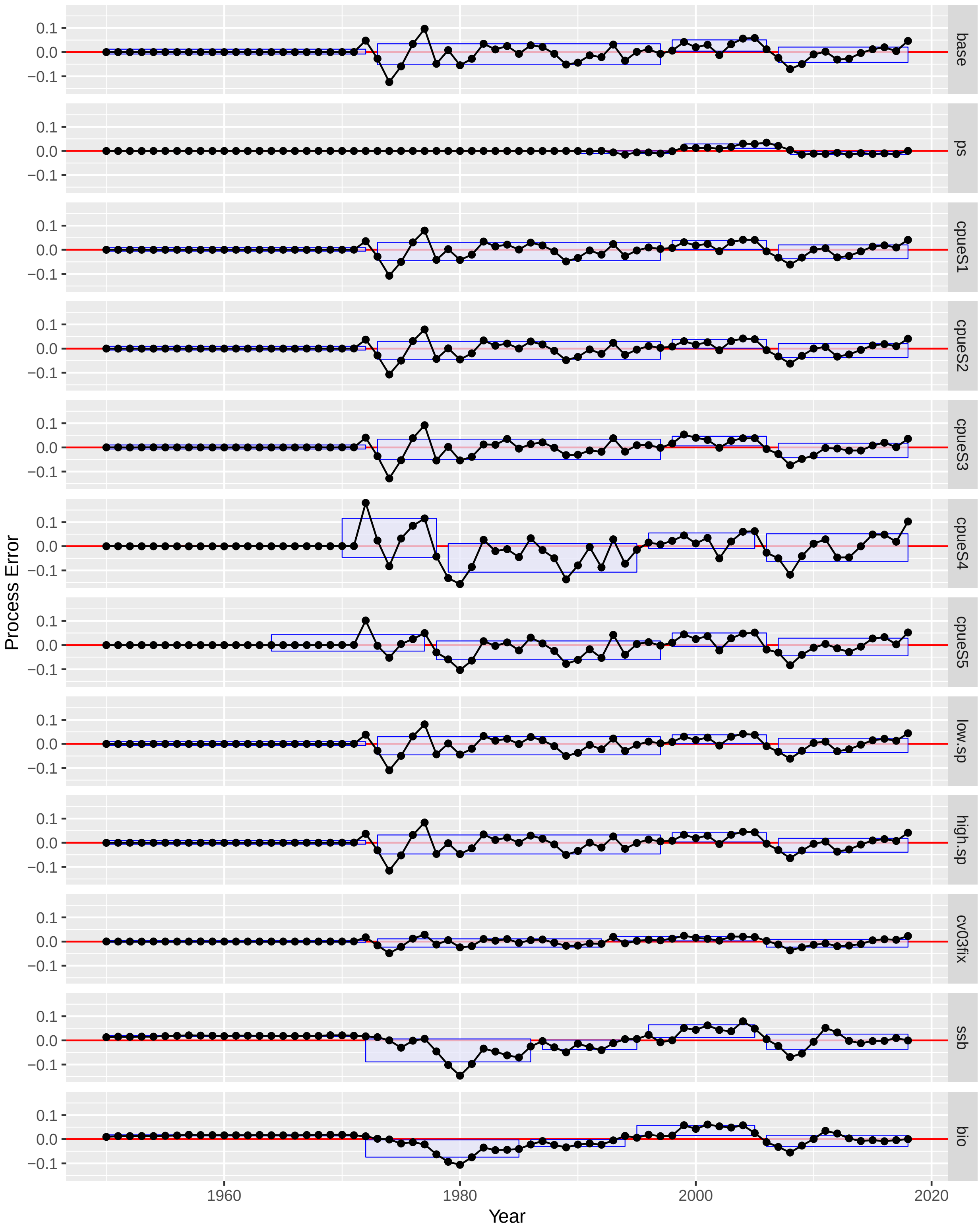

Figure 22 Time series of process error with regimes.

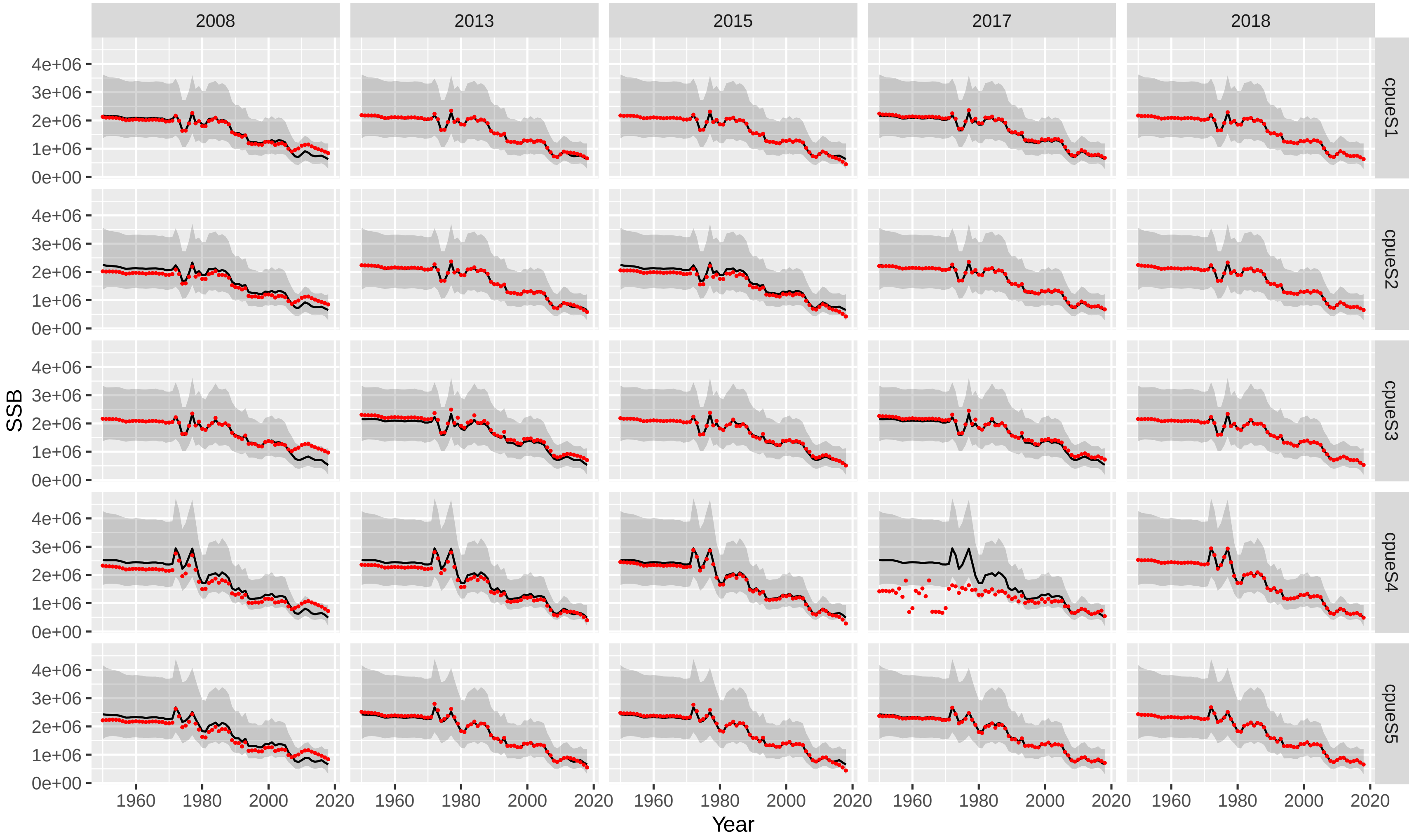

Figure 23 SSB estimates form retrospective analysis.

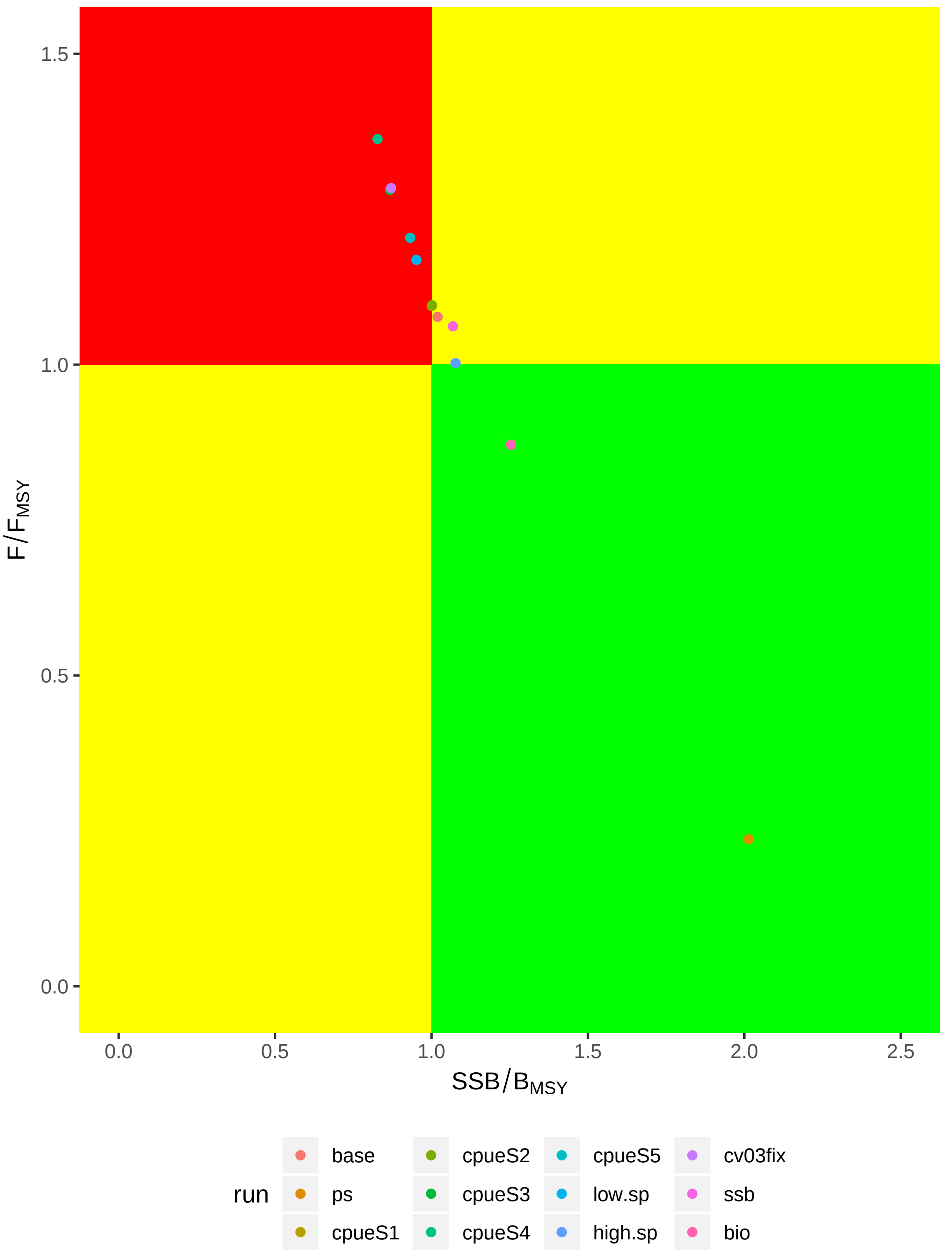

Figure 24 Kobe Phase Plots.

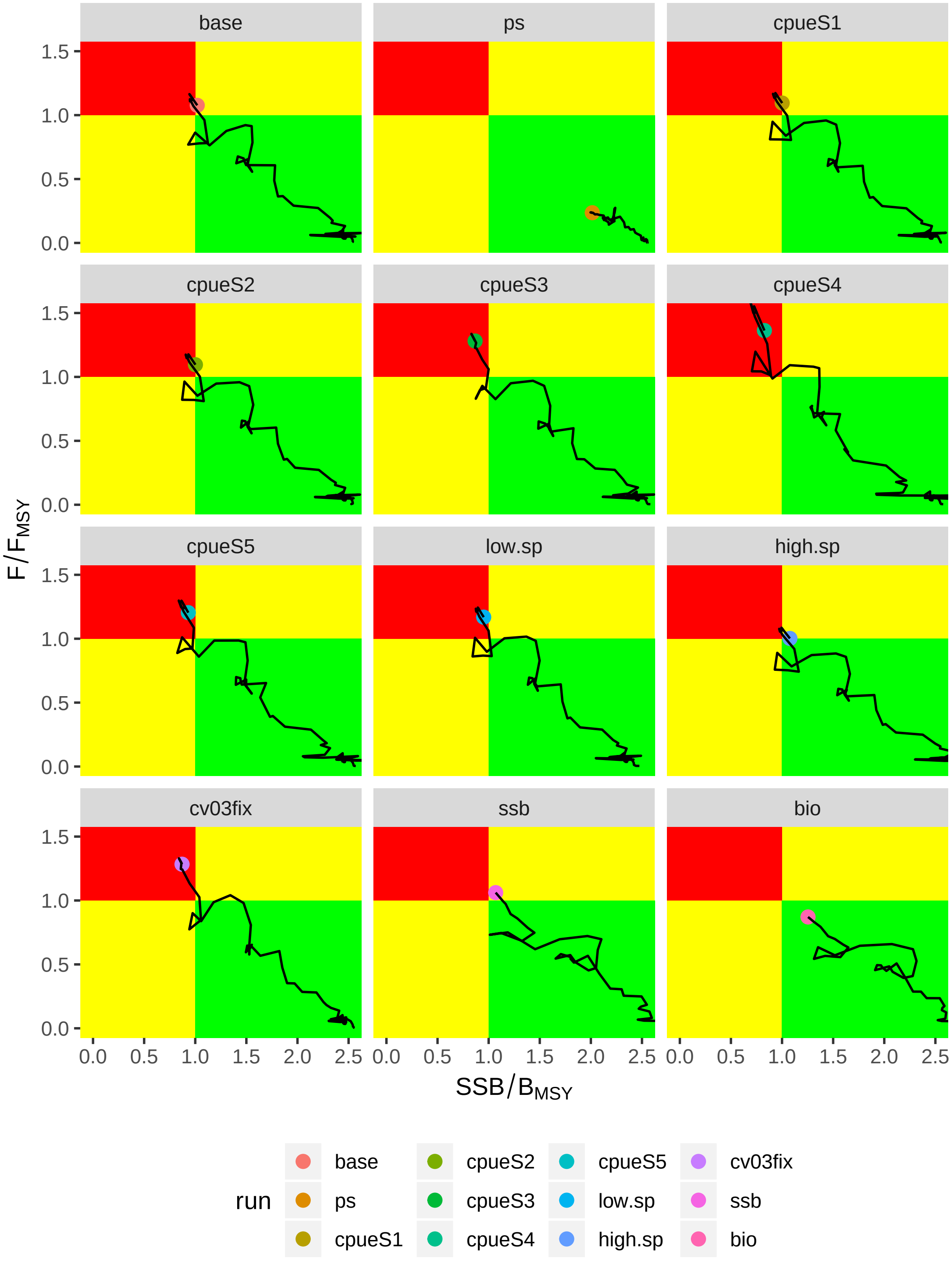

Figure 25 Kobe Phase Plots.

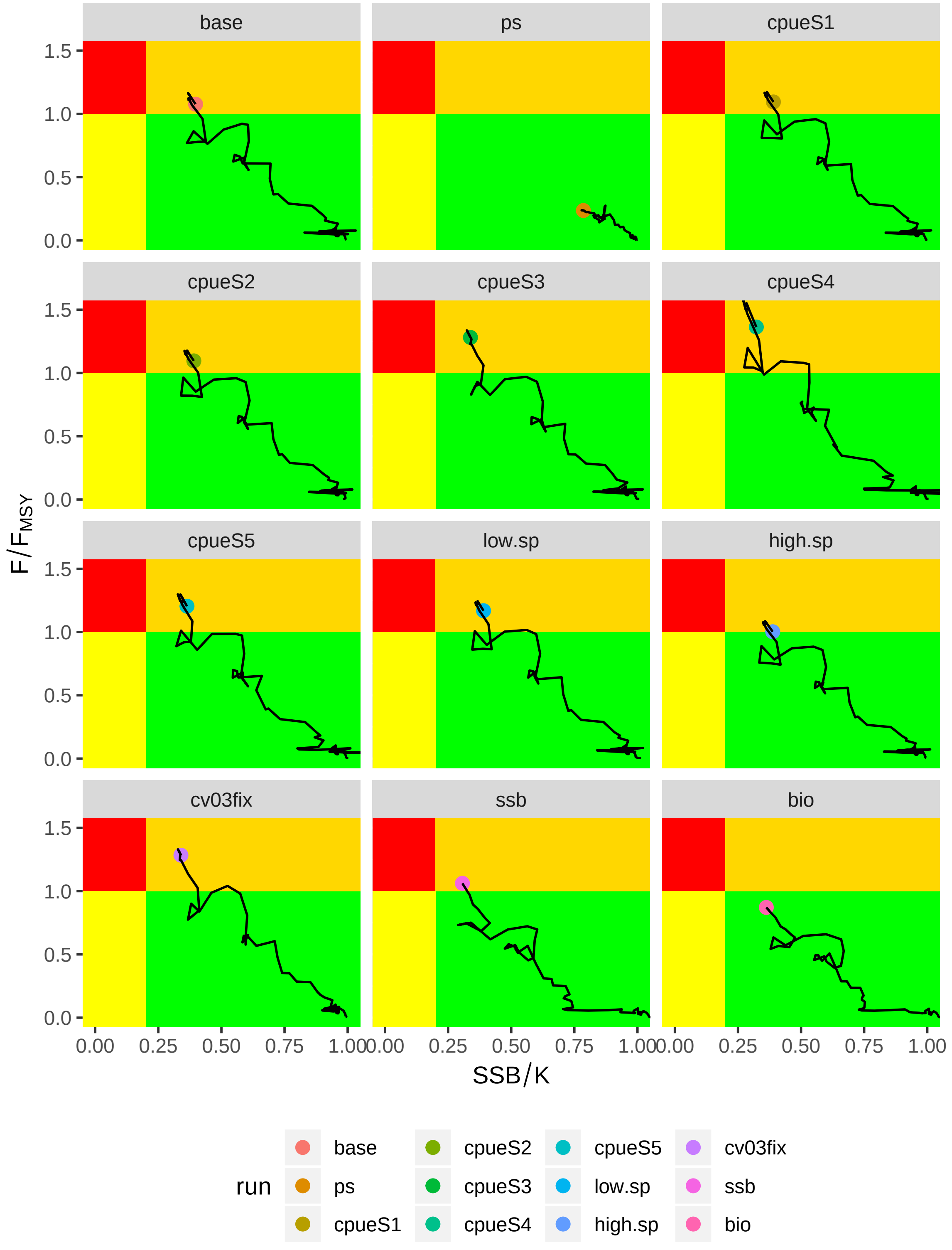

Figure 26 Majuro Phase Plots.