Is It You or Your Model Talking

Laurence Kell, Toshihide Kitakado, Rishi Sharma, Henning Winker,
Iago Mosqueira, Dan Fu, Max Cardinale, ...

May 28, 2020

Outline

- 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 (Holt pers com.)
- Often when conducting a stock assessment multiple models with different structures and datasets, are used to explore uncertainty. This means that it is difficult to compare models using conventional metrics such as AIC. The use of metrics based on prediction skill allows different data components and model to be compared in order to explore data conflicts and potential model misspecification.
- Retrospective analysis is commonly used to evaluate the stability of stock assessment estimates, however, stability can be at the expense of prediction skill, i.e. by using shrinkage. We therefore predict forward the retrospective analyses and then compare model predictions with historical estimates.The absence of retropective patterns, however, while reassuring is not sufficient alone as it is not possible to validate models based on model outputs. We therefore conduct model free hindcasts to compare observations with model estimates.
- We compare SS, SS-ASPM, and Jabba assessments for Indian Ocean yellowfin tuna using multiple metrics, make recommendations for benchmarking of stock assessments and discuss the consequences for MSE, i.e. weighting of OMs and developing OEMs.

Contents

Abstract

Evaluating how well the model fits data has been receiving much attention in fisheries science, both in terms of goodness-of-fit and retrospectively. This however merely tells us how well we can describe the past, yet little how well we can predict the future under alternative management actions. In this paper, we revisit the concepts behind hindcasting cross-validation (hcxval) as an important model-free validation tool for predictive modelling. Together with conventional residual diagnostics and retrospective analysis, we apply hcxval to three examples of alternative candidate models using the recent Indian Ocean yellowfin tuna assessment as a case study. These models comprise the 2019 spatially structured reference model implemented in Stock Synthesis (ss-ref), a deterministic age-structured production model (ss-aspm) of ss-ref and a simplied spatially aggregated stochastic surplus production model implemented in the 'JABBA' package. To assess prediction skill, we computed the Mean-Absolute-Scaled-Error (MASE), which, unlike e.g. Aikaike's Information Criterion, enables to compare across different models fitted different data. The best MASE values (MASE < 1) were determined for ss-asem, which indicates that recruitment deviations in ss-ref were poorly estimated due to no or limited information in the 'noisy' length composition data. By contrast, the area effects retained in ss-aspm best explained its superior prediction skill compared to the spatially aggregated jabba model. We suggest that one-step ahead predictions are efficient for detecting overfitting and for model validation in general, but for future quota advice the forecast horizon should preferably at least match the assessment interval to ultimately increase confidence in the model-based scientific advice by stake holder and managers and policy makers.

1 Introduction

In stock assessment most goodness of fit diagnostic are based on residuals obtained from fits to historical observations. To provide fisheries management advice, however, requires predicting the response of a stock to management and checking that the predictions are consistent with reality (pers. com. Sidney Holt). The accuracy and precision of predictions depend on the validity of the model, the information in the data, and how far ahead we wish to predict. Validation examines if a model should be modified or extended and is complementary to model selection and hypothesis testing. Model selection searches for the most suitable model within a family, whilst hypothesis testing examines if the model structure can be reduced.

Model validation is important in many fields, e.g. in energy and climate models, as it increases confidence in the outputs of a model and leads to an increase in trust amongst the public, stake and asset-holders and policy makers. For models to be valid they must satisfy four prerequisites Hodges et al. (1992), the situation being modelled must be observable and measurable, it must be possible to collect sufficient data, exhibit constancy of structure in time, and exhibit constancy across variations in conditions not specified in the model. The first two prerequisites should be straight forward, but many stock assessments depend on fisheries dependent data rather than scientific observation. For example highly migratory stocks fished in areas beyond national jurisdiction (ABNJ). Prerequisite 3 ensures that the model has predictive skill for the same conditions under which the validation tests were conducted. Prerequisite 4 ensures that the model will still be valid for conditions that differ from those in the validation tests, i.e. can be used to set robust management advice.

To explore the robustness of advice to uncertainty requires different model structures to be condition on alternative and potentially conflicting datasets. In such cases model selection criteria such as AIC, however, cannot be applied. The first prerequisite means it is not possible to validate a model, using derived quantities, such as SSB and F. The key concept in this case is prediction skill, defined as any measure of accuracy of a forecasted value to the actual (i.e. observed) value that is not known by the model (Glickman and Zenk, 2000). Therefore An alternative is to use model-free hindcasting, a form of crossvalidation where observations are compared to their predicted values.

To illustrate the utility of hindcasting we develop a case study based on bigeye and yellowfin tuna stocks in the Indian, Atlantic and Eastern Pacific Oceans, and four assessment methods, SS, SS-ASPM, Jabba-Select and Jabba.

2 Material and Methods

Kell et al. (2016) proposed a model-free hindcasting using crossvalidation where observations (e.g. CPUE) are compared to their predicted future values. The hindcasting algorithm is similar to that used in retrospective analysis (Hurtado-Ferro et al., 2014), which involves sequentially removing observations from the terminal year (peels), fitting the model to the truncated series, and then comparing the difference between model estimates from the truncated time-series to those estimated using the full time series. In a model-free hindcast an additional step is included, i.e. projecting over the missing years and then cross-validating these forecasts against observations to assess the model's prediction skill.

2.1 Assessment Methods

Case study based on Indian Ocean yellowfin tuna stocks and four assessment methods, SS, SS-ASPM, Jabba-Select and Jabba.

```
\input{assess.tex}
```

2.2 Procedure

2.2.1 Residuals

```
\input{residuals.tex}
```

2.2.2 Retrospective

```
\input{retro.tex}
```

2.2.3 Retrospective with Projection

When conducting projections to provide managers with advice,such as a total allowable catch (TAC), the short term is of primary importance as usually the immediate consequences of management advice is a major concern of stakeholders (Fricker et al., 2013). Therefore we also conduct projections for 3 years as part of the retrospective analysis.

2.3 Hindcast

```
\input{hindcast.tex}
```

2.3.1 Summary Metrics

```
\input{metrics.tex}
```

3 Results

Figure 1 Spatial stratification of the Indian Ocean for the four region assessment model (R1a and R1b were treated as one model region but were retained for the fleet definition). The black arrows represent the configuration of the movement parameterization. Density contours represent of the dispersal of tag releases (red) and subsequent recaptures from Indian Ocean Regional tuna tagging program. Green circles represent the distribution of catches from the longline fishery aggregated by 5 longitude * 5 latitude for 1980 – 2017 (max. = 133 770 t)

Figure 2 Regional longline CPUE indices included in the 2018 stock assessment. The difference in scales represents the relative distribution of longline vulnerable biomass amongst regions.

The first step was to conduct a retrospective analysis and the estimates of stick biomass (SSB for ASPM and SS, and biomass for JABBA) and F (instantaneous for ASPM and SS, and rate for JABBA) are shown in Figure 3. The terminal estimates were then project forward for three years assuming the reported catches and the estimated recruitment from the model with all years included (Figure 4). ASPM "predicted" values are close to those of the assessment that includes all years. For SS, however, there is a large overestimation of future biomass, and a underestimate of F . For JABBA the strongest retrospective pattern is seen in F, which is underestimated in the predictions.

The residuals from the model fits are shown in Figure 5, the backgroud indicates whether they passed (green) or failed (red) the runs tests.

The results from the model-free Hindcast with one year ahead predictions are shown in Figure 6 and from the three year ahead predictions in Figure 7.

Figure 8 shoes the predictions residuals, and the fits are summarised in Figures 9 and Figure 10 in the form of Taylor diagrams

Table 1 and 2 show Mohn's ρ for the retrospective analysis and retrospective with projection, Table shows the values of RMSE and MASE for the model free hindcast.

Retrospective analysis for F/FMSY and B/BMSY

- Figure 3 and Table 1 summarise the retrospective analysis, taking 0.2 and -0.15 as the cut off points for accepting an assessment all assessments pass.
- The strongest retrospective patterns are seen for SS, where F is negatively biased, and although the value of Mohn's ρ is low for SSB a strong pattern is seen with underestimates followed by overestimates.
- Although recent Jabba estimates are unbiased historical estimates of F and Biomass are negatively biased.
- Jabba estimates are problematic as it appears that even if F<FMSY the stock will decline below BMSY
- ASPM estimates appears to have little bias

Retrospective analysis and projections for F/FMSY and B/BMSY

- Figure 4 and Table 2 summarise the retrospective analysis combinded with a three prediction, again taking the cut off as 0.2 to -0.15, both SS and Jabba fail.
- ASPM appears not to show patterns in the projections
- ASPM performs best
- Survey 2 performs poorly across all models
- It appears that the length compositions add noise (SS), and that area effects (JABBA) are important.
- However these is no objective way to choose an assessment based on retrospective analysis as the best model would always be B/MSY = 1

Residuals

- Figure 5 summarises the residuals, in the form of runs test; green backgrounds denote a pass. Only two indices pass for all models Survey 1 quarter 1 and Survey 3 quarter 4
- Over half of the indices fail for Jabba, half for SS while for ASPM the majority pass. This indicates strong data conflicts while failure of the runs test may explain the retrospective patterns.
- Compare MASE to other measures and use Taylor Diagrams to explain differences

Model-free cross-validation confirms the relative performance of the models

- Figures 6 and 7 shows that SS performs poorly, possibly because length compositions can only be explained by variations in year-class strengths, and projections predict large increase in biomass.
- RMSE is difficult to interpret (Table 3), MASE easier (Table 4).
- Taylor diagram, shows that there is a big difference between model residuals and prediction residuals. Figure 9 summarises the historical model fits survey 4 performs the best, i.e. has high correlation for all models, while survey 3 performs poorly.
- However, when three step ahead projections are considered (Figure 10) survey 4 perform the best for ASPM, although the correlation is reduced. SS has high RMSE, poor correlation and high variance.
- Figure 8, need to say something or delete it.

MSE

- The results have consequences for MSE, both for selection of OMs which are used to simulate future states under feedback control, and the choice of indices to simulate in the OEM.
- If the SS base case has poor prediction skill how can it be used as an OM?
- The OEM generates data for use in the MP, the TDs characterise the process and measurement error, i.e. only indices with high correlation should be used as input to the MP.
- What if more than 1 OM scenarios has been conditioned with alternative data weightings and in each case there is a different preferred index?

4 Discussion

`\input{discussion.tex}`

5 Conclusions

Primary objectives were:

1. In fisheries unlike other fields we try to account for the past but not for the future. Here we propose a way to assess model predictive performance and to account for alternative models within a common diagnostic framework.

2. Unifying platform for evaluating across models
3. Advantage of MASE : What are the new properties of this stat
4. Consequences for stock assessment benchmarking
5. Consequences for uncertainty, risk and MSE.
6. Next steps

References

Fricker, T. E., Ferro, C. A., and Stephenson, D. B. (2013). Three recommendations for evaluating climate predictions. *Meteorological Applications*, 20(2):246–255.

Glickman, T. S. and Zenk, W. (2000). *Glossary of meteorology*. American Meteorological Society.

Hodges, J. S., Dewar, J. A., and Center, A. (1992). *Is it you or your model talking?: A framework for model validation*. Santa Monica, CA: Rand.

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., et al. (2014). Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. page fsu198.

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

6 Tables

Table 1: Mohn's ρ for retrospective analysis.

	run	variable	V1
1	aspm	stock	-0.03
2	aspm	harvest	0.03
3	base	stock	0.04
4	base	harvest	-0.15
5	jabba	stock	-0.09
6	jabba	harvest	-0.09

Table 2: Mohn's ρ for retrospective analysis with three year projection.

	run	variable	V1
1	aspm	stock	-0.09
2	aspm	harvest	0.08
3	base	stock	0.32
4	base	harvest	-0.24
5	jabba	stock	-0.21
6	jabba	harvest	-0.28

Table 3: RMSE.

	name	quarter	SS	ASPM	JABBA
1	SURVEY1	1.00	0.37	0.24	0.25
2	SURVEY1	2.00	0.28	0.24	0.26
3	SURVEY1	3.00	0.47	0.45	0.29
4	SURVEY1	4.00	0.29	0.14	0.24
5	SURVEY2	1.00	0.36	0.34	0.43
6	SURVEY2	2.00	0.64	0.62	0.50
7	SURVEY2	3.00	0.30	0.29	0.23
8	SURVEY2	4.00	0.42	0.42	0.38
9	SURVEY3	1.00	0.28	0.19	0.21
10	SURVEY3	2.00	0.34	0.31	0.38
11	SURVEY3	3.00	0.22	0.22	0.30
12	SURVEY3	4.00	0.37	0.36	0.40
13	SURVEY4	1.00	0.37	0.15	0.41
14	SURVEY4	2.00	0.45	0.30	0.53
15	SURVEY4	3.00	0.49	0.25	0.43
16	SURVEY4	4.00	0.47	0.18	0.48

Table 4: MASE.

	name	quarter	SS	ASPM	JABBA
1	SURVEY1	1.00	0.94	0.53	0.77
2	SURVEY1	2.00	0.63	0.67	0.96
3	SURVEY1	3.00	1.23	1.17	0.85
4	SURVEY1	4.00	0.82	0.45	0.74
5	SURVEY2	1.00	1.52	1.73	2.41
6	SURVEY2	2.00	2.11	2.10	1.56
7	SURVEY2	3.00	0.95	0.82	0.83
8	SURVEY2	4.00	1.33	1.65	1.26
9	SURVEY3	1.00	0.86	0.57	0.88
10	SURVEY3	2.00	0.92	0.81	0.92
11	SURVEY3	3.00	0.93	0.76	1.22
12	SURVEY3	4.00	0.85	0.91	0.99
13	SURVEY4	1.00	2.10	0.76	3.00
14	SURVEY4	2.00	0.91	0.63	1.14
15	SURVEY4	3.00	2.07	0.93	1.92
16	SURVEY4	4.00	3.29	1.09	3.90

7 Figures

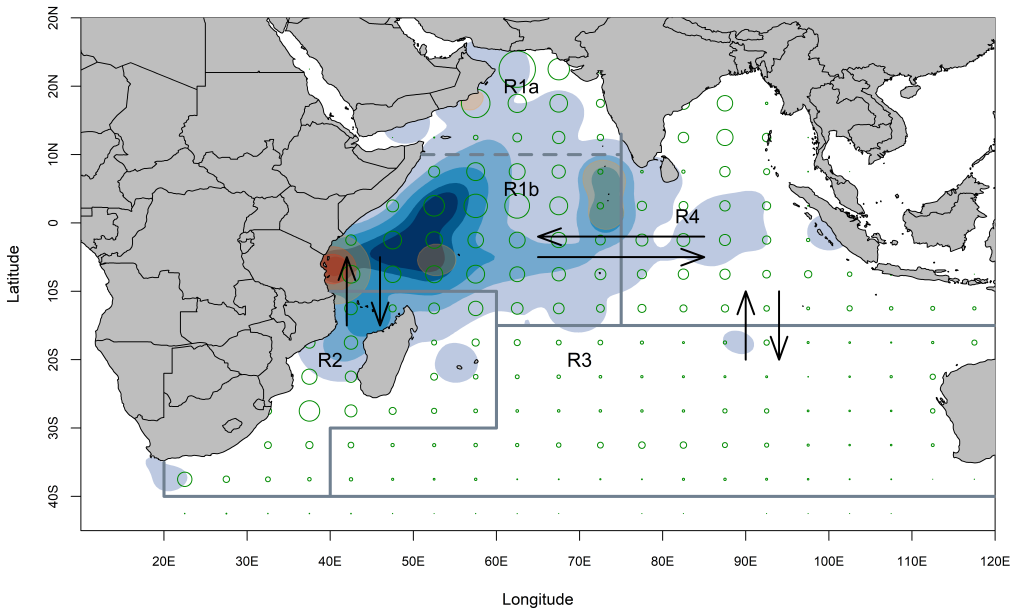

Figure 1: Spatial stratification of the Indian Ocean for the four region assessment model (R1a and R1b were treated as one model region but were retained for the fleet definition). The black arrows represent the configuration of the movement parameterization. Density contours represent of the dispersal of tag releases (red) and subsequent recaptures from Indian Ocean Regional tuna tagging programme. Green circles represent the distribution of catches from the longline fishery aggregated by 5 longitude * 5 latitude for 1980 – 2017 (max. = 133 770 t).

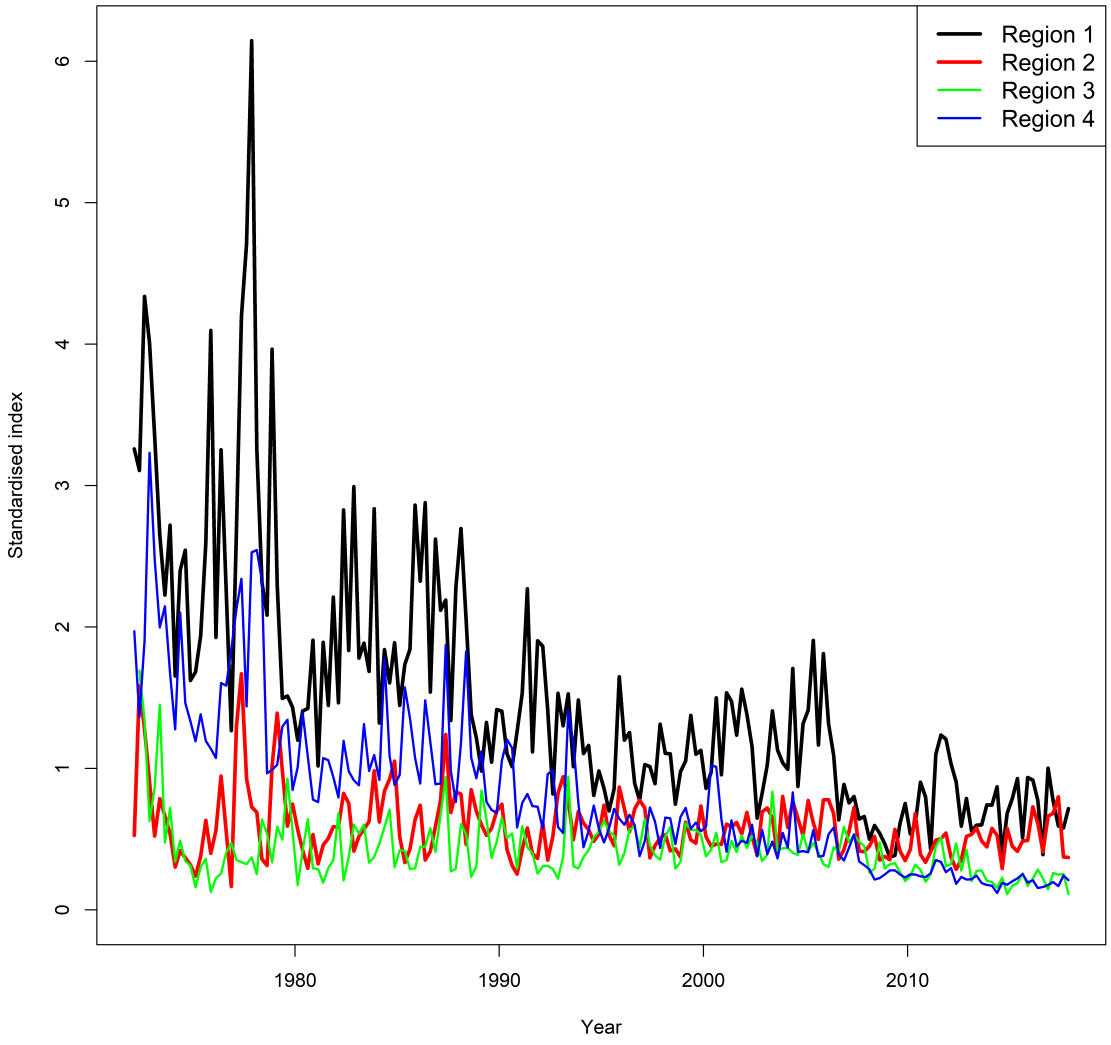

Figure 2: Regional longline CPUE indices included in the 2018 stock assessment. The difference in scales represents the relative distribtuion of longline vulnearable biomass amonst regions.

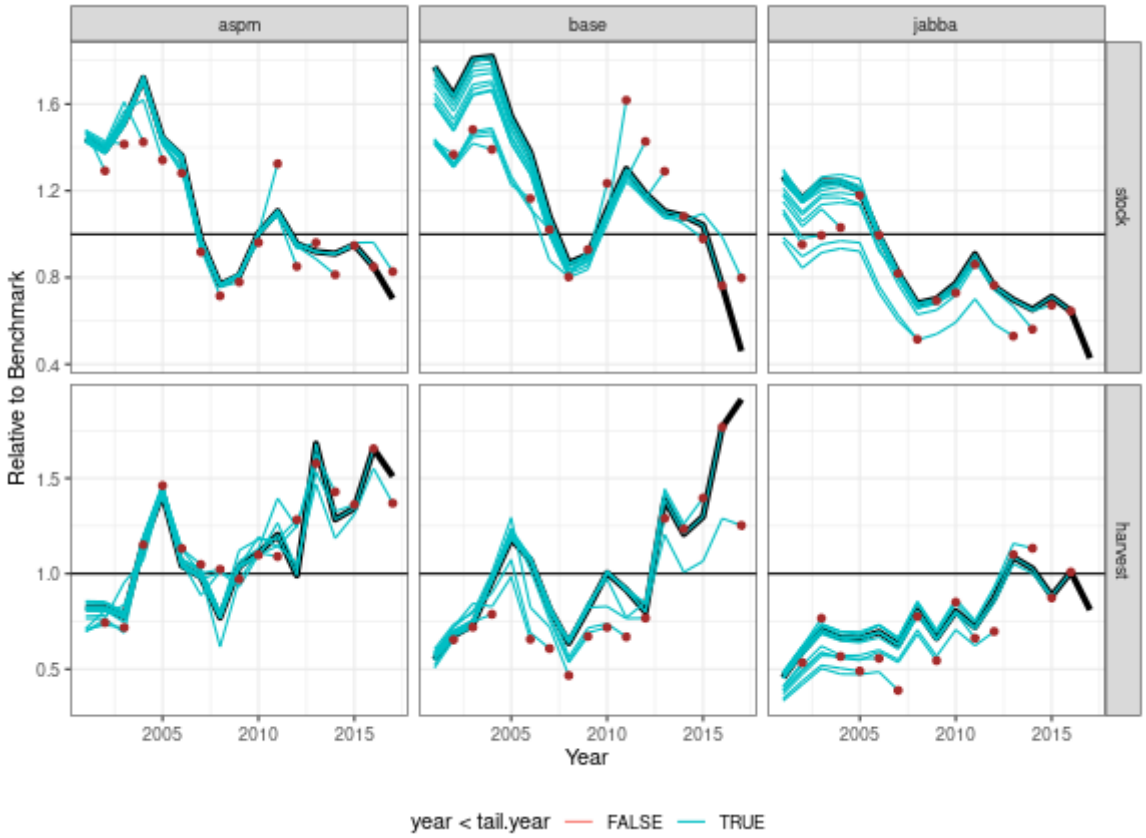

Figure 3: Retrospective analysis for the three models, points indicate the terminal years, and the think line the assessment using all the data.

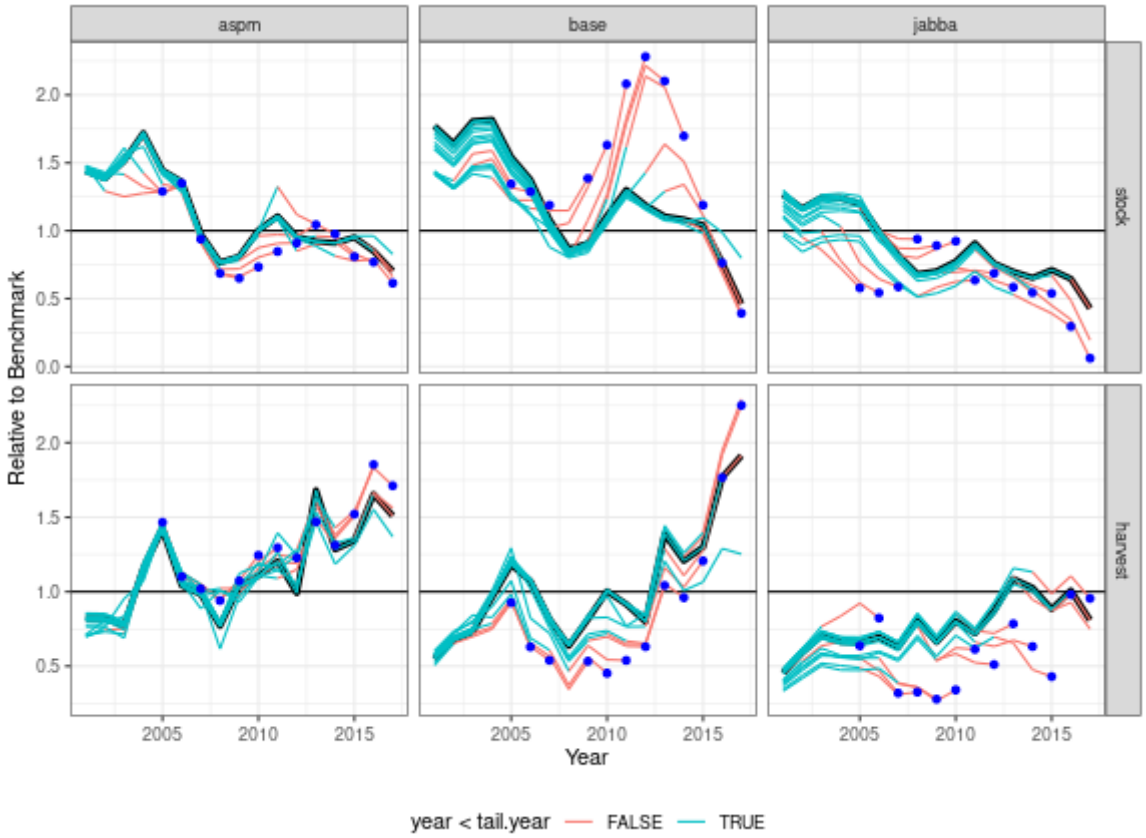

Figure 4: Retrospective analysis with three year predictions for the three models, points indicate the terminal years, and the think line the assessment using all the data.

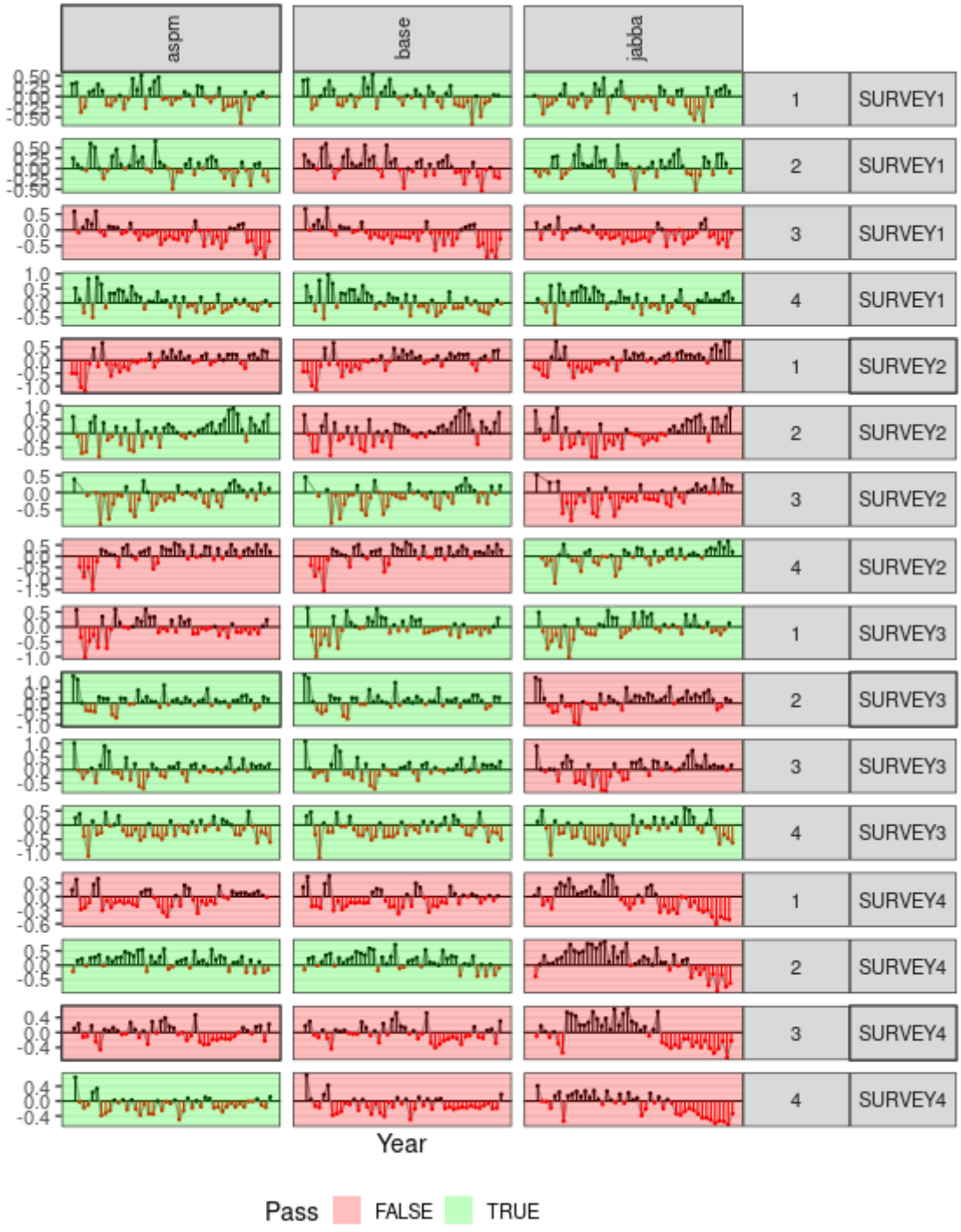

Figure 5: Residual runs tests for fits to the three models; green background indicates series where runs tests are passed.

Figure 6: Hindcast with one year ahead predictions, red dots are the observed CPUE values and lines are the fits with terminal hincast year indicated by a point.

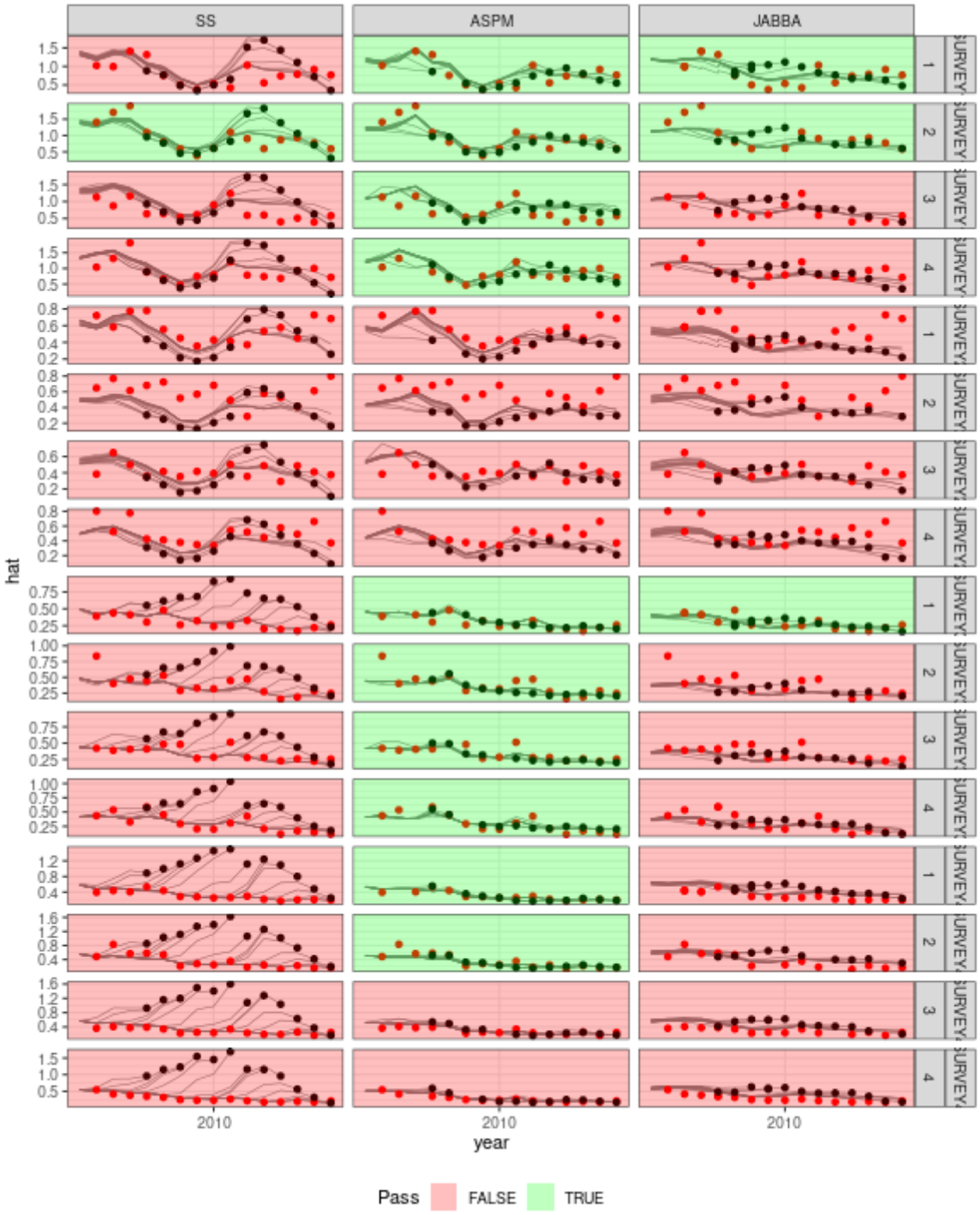

Figure 7: Hindcast with three year ahead predictions, red dots are the observed CPUE values and lines are the fits with terminal hincast year indicated by a point.

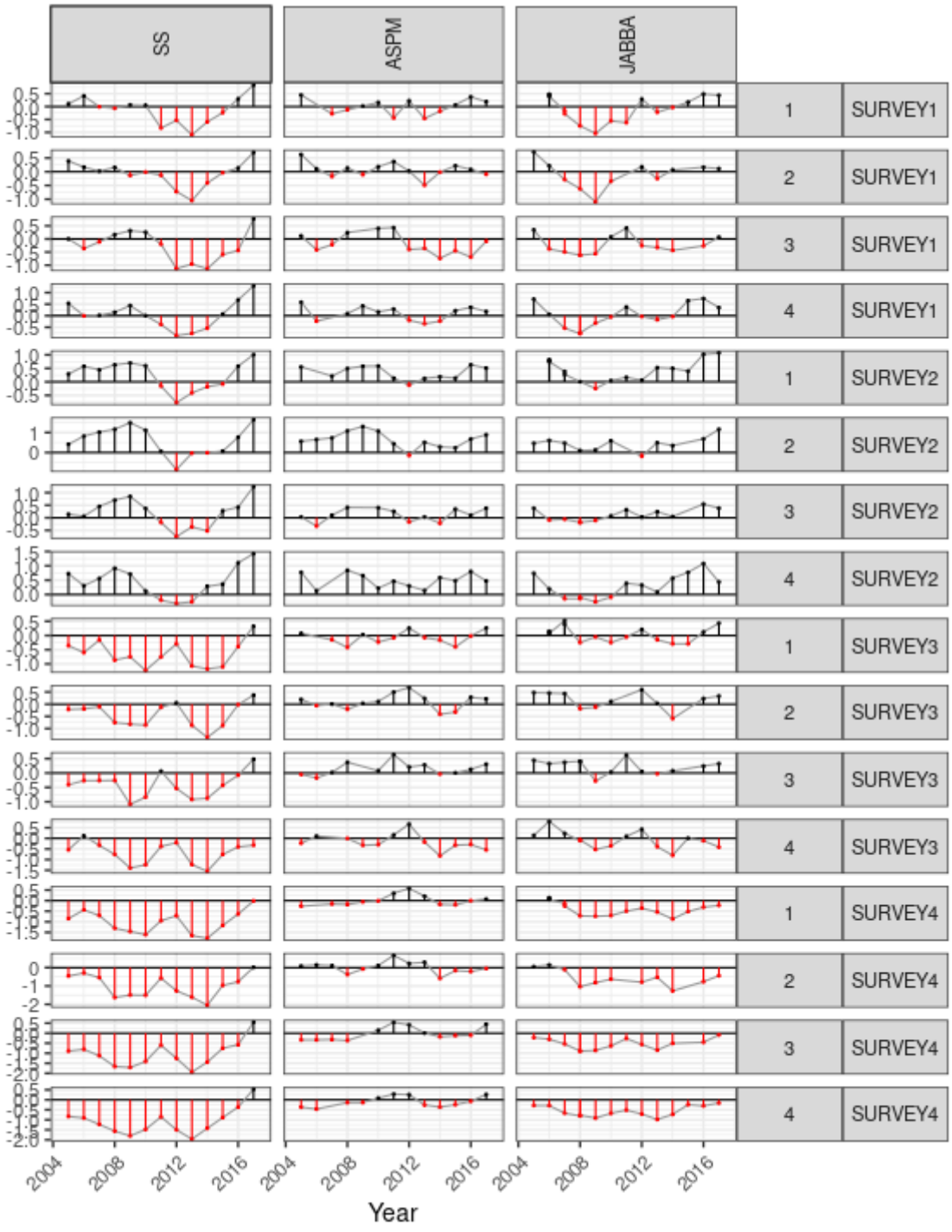

Figure 8: Runs tests for one step ahead residuals.

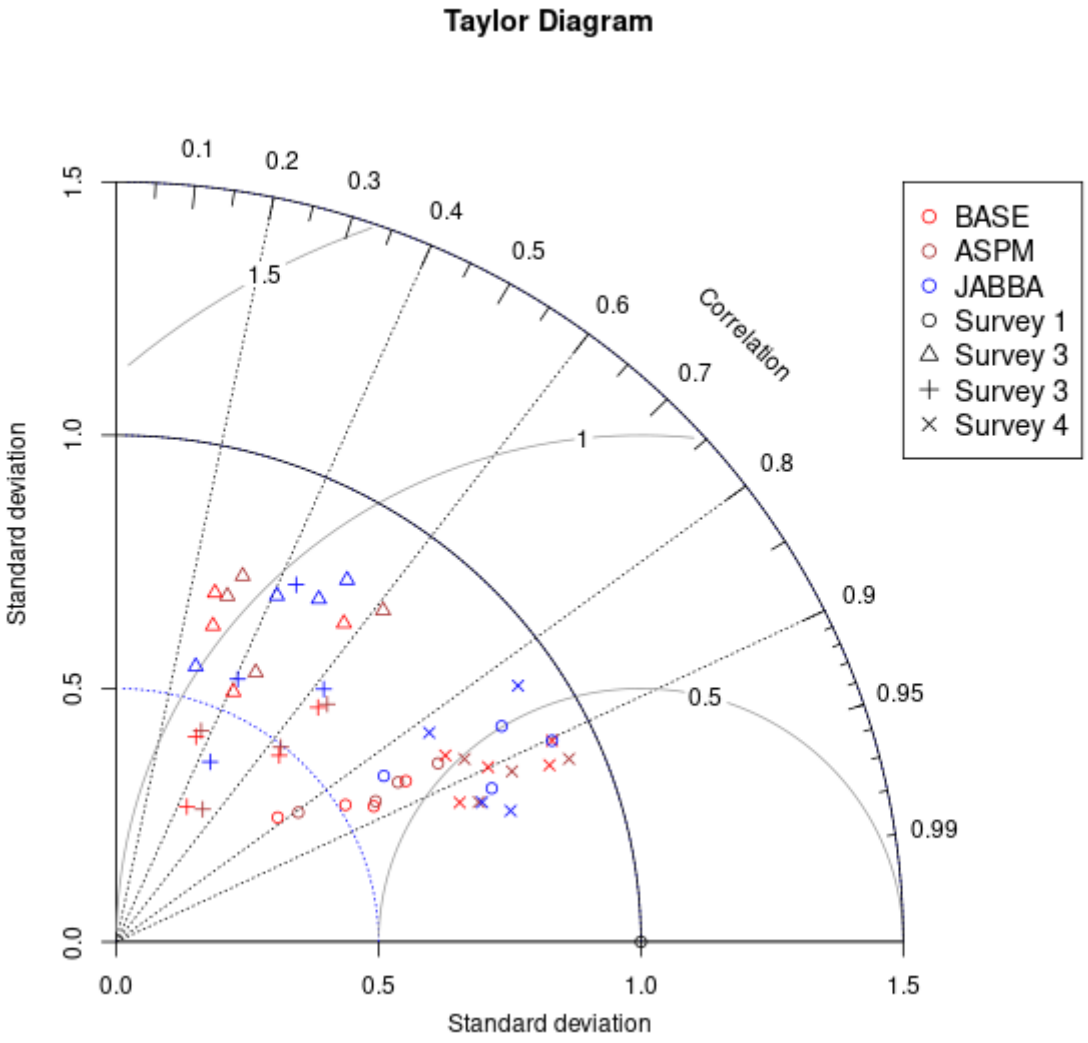

Figure 9: Taylor diagram for fits to CPUE summarising the similarity between the observed time series of CPUEs and the predicted relative stock abundance. Each point quantifies how closely predictions match observations, the angle indicates the correlation, the centred root-mean-square error difference between the predicted and observed patterns is proportional to the distance to the point on the x and the contours around this point indicate the RMSE values; the standard deviations of the predictions are proportional to the radial distance from the origin, scaled so the observed pattern has a value of 1. The open circle corresponds to a series which is identical to the reference series. The colours correspond to the model and shape to the survey.)

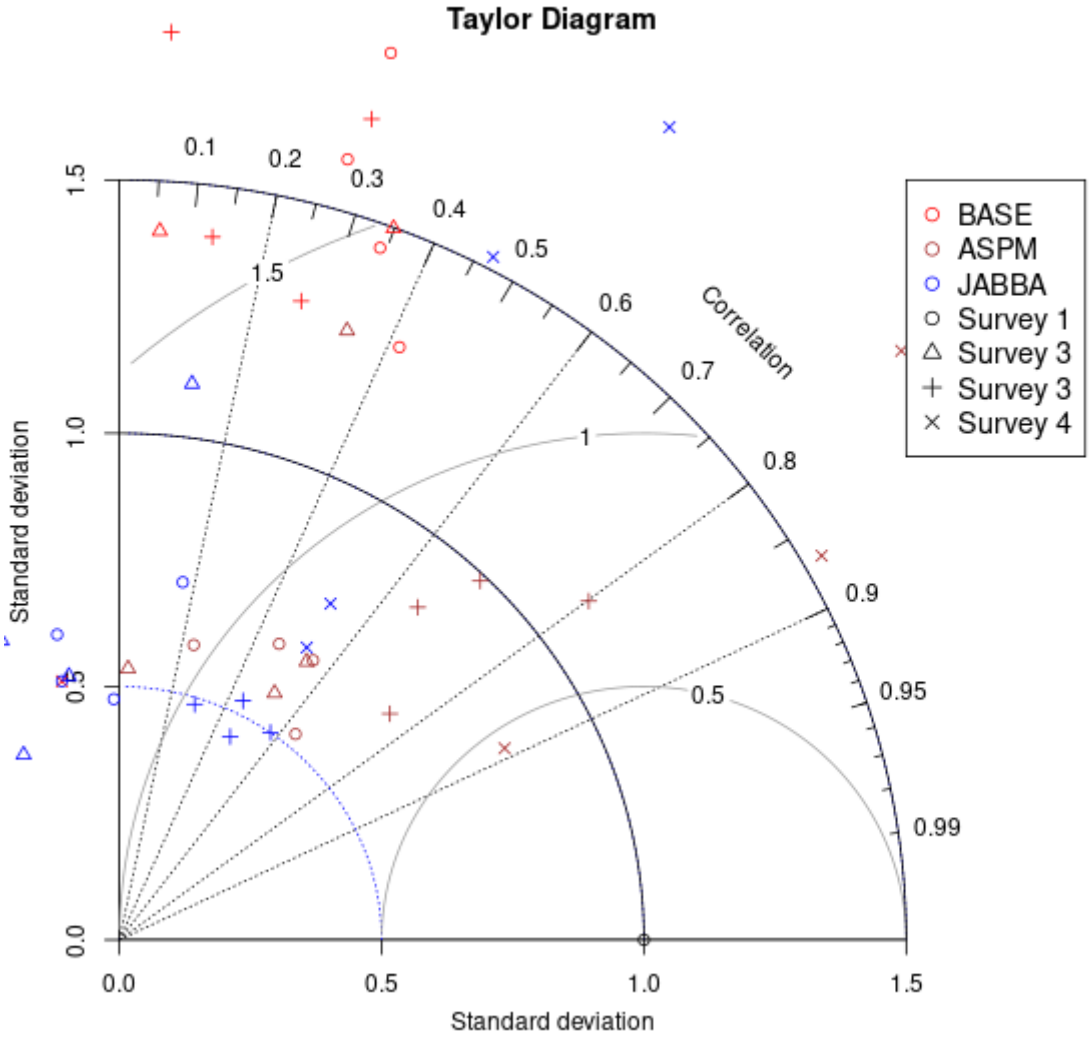

Figure 10: Taylor diagram for 3 year ahead predictions, summarising the similarity between the observed time series of CPUEs and the predicted relative stock abundance. Each point quantifies how closely predictions match observations, the angle indicates the correlation, the centred root-mean-square error difference between the predicted and observed patterns is proportional to the distance to the point on the x and the contours around this point indicate the RMSE values; the standard deviations of the predictions are proportional to the radial distance from the origin, scaled so the observed pattern has a value of 1. The open circle corresponds to a series which is identical to the reference series. The colours correspond to the model and shape to the survey.