

# Is It You or Your Model Talking

Laurie Kell, Toshihide Kitakado, Rishi Sharma, ...

May 20, 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 retrospective 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

1 Introduction 3

2 Material and Methods 3

2.1 Assessment Methods . . . . . 4

2.2 Procedure . . . . . 4

2.2.1 Retrospective . . . . . 4

2.2.2 Hindcast . . . . . 4

2.3 Summary Metrics . . . . . 4

3 Results 4

4 Discussion 5

5 Conclusions 5

6 Tables 5

7 Figures 5

## 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 simplified 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. Aikake's Information Criterion, enables to compare across different models fitted different data. The best MASE values ( $\text{MASE} < 1$ ) were determined for ss-aspm, 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 Retrospective

`%include{retro.tex}`

### 2.2.2 Hindcast

`%include{hindcast.tex}`

## 2.3 Summary Metrics

`%include{metrics.tex}`

# 3 Results

Figure 1 shows the stock distribution shows the stock structure assumed in the SS assessments.

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 2. 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 3). 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 4, the background 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 5 and from the three year ahead predictions in Figure 6.

Figure 7 shoos the predictions residuals, and the fits are summarised in Figures 8 and Figure 9 in the form of Taylor diagrams

## Summary

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

- SS projections and hence advice is based upon SS is unreliable.
- Jabba advice is problematic as it appears that even if  $F < F_{MSY}$  the stock will decline below BMSY
- ASPM appears consistent
- It therefore appears that the length compositions just add noise, and that area effects are important
- However these is no objective way to choose an assessment based on retrospective analysis as the best model would always be  $B/MSY = 1$

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

- ASPM performs best
- Survey 2 performs poorly across all models

Make case for why we need MASE and why this new methodology is important

- Tabulate Metrics
- Compare MASE to other measures and use Taylor Diagrams to explain differences

## 4 Discussion

`%include{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

## References

- 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

## 7 Figures

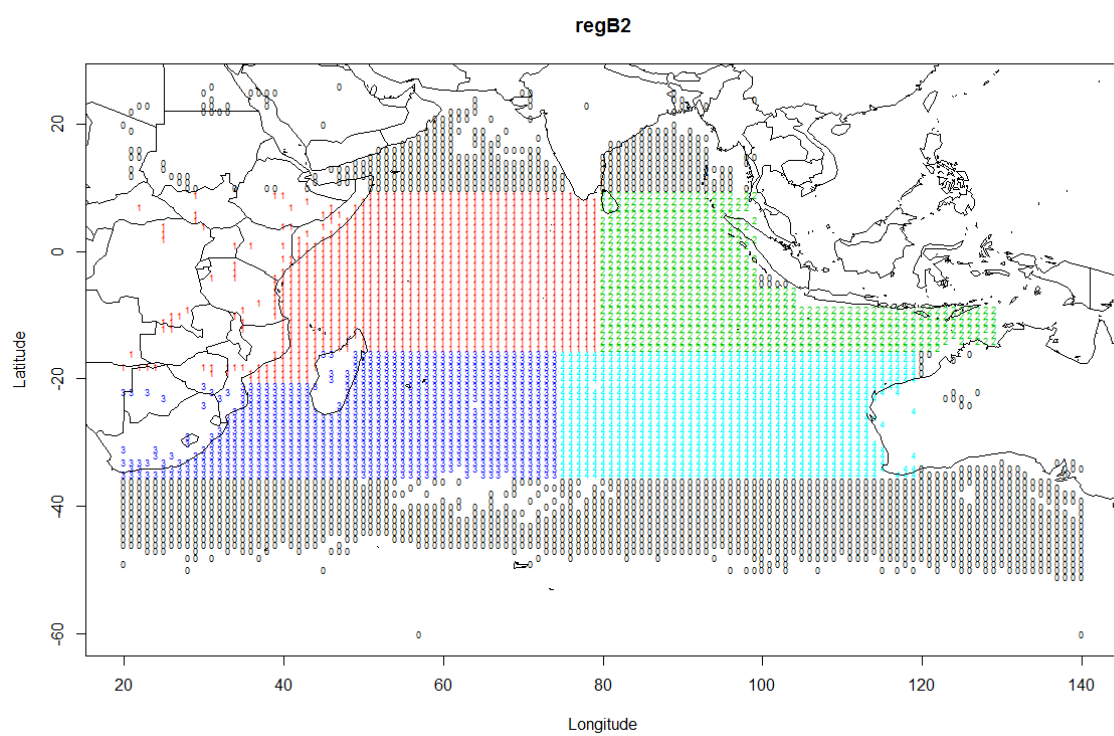


Figure 1: Stock distribution.

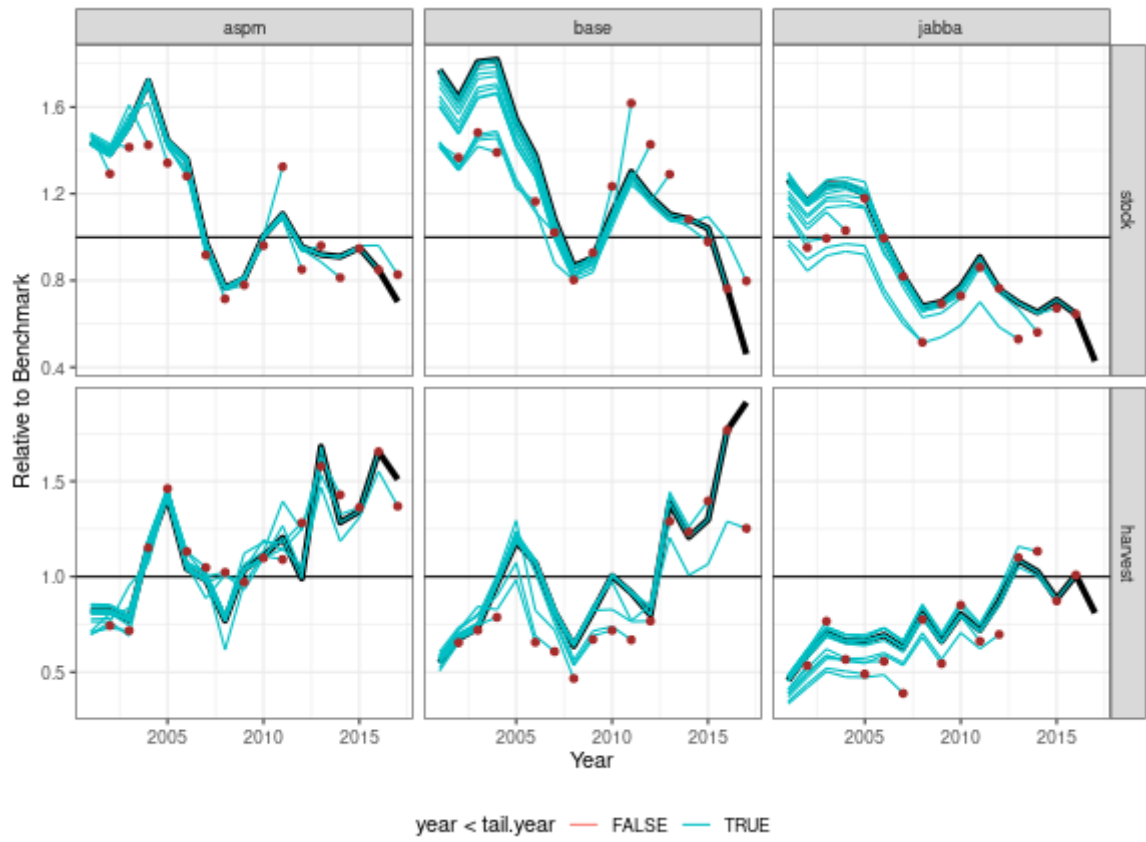


Figure 2: Retrospective analysis for the three models, points indicate the terminal years, and the thick line the assessment using all the data.



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



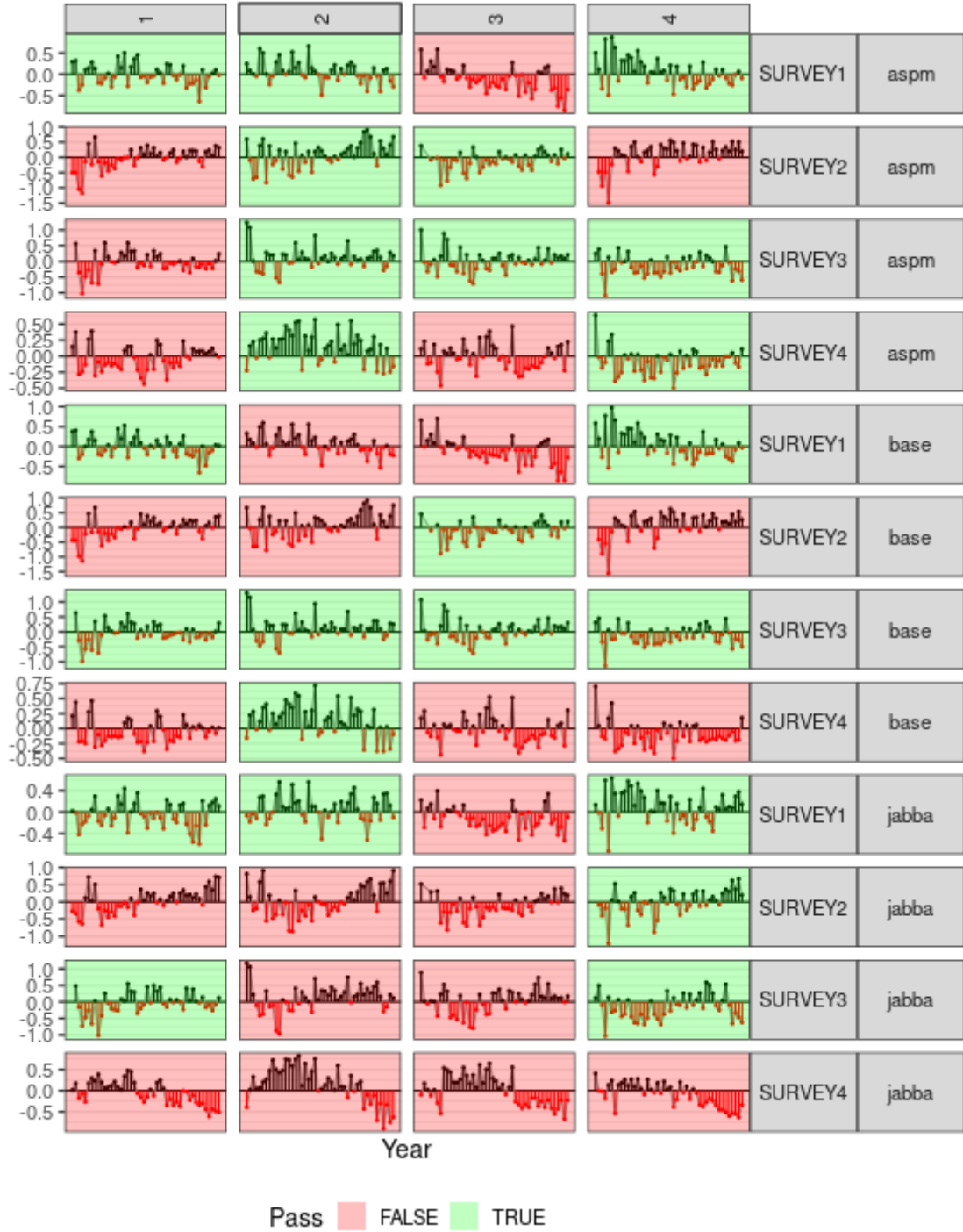


Figure 4: Residual runs tests for fits to the three models; green background indicates series where runs tests are passed.

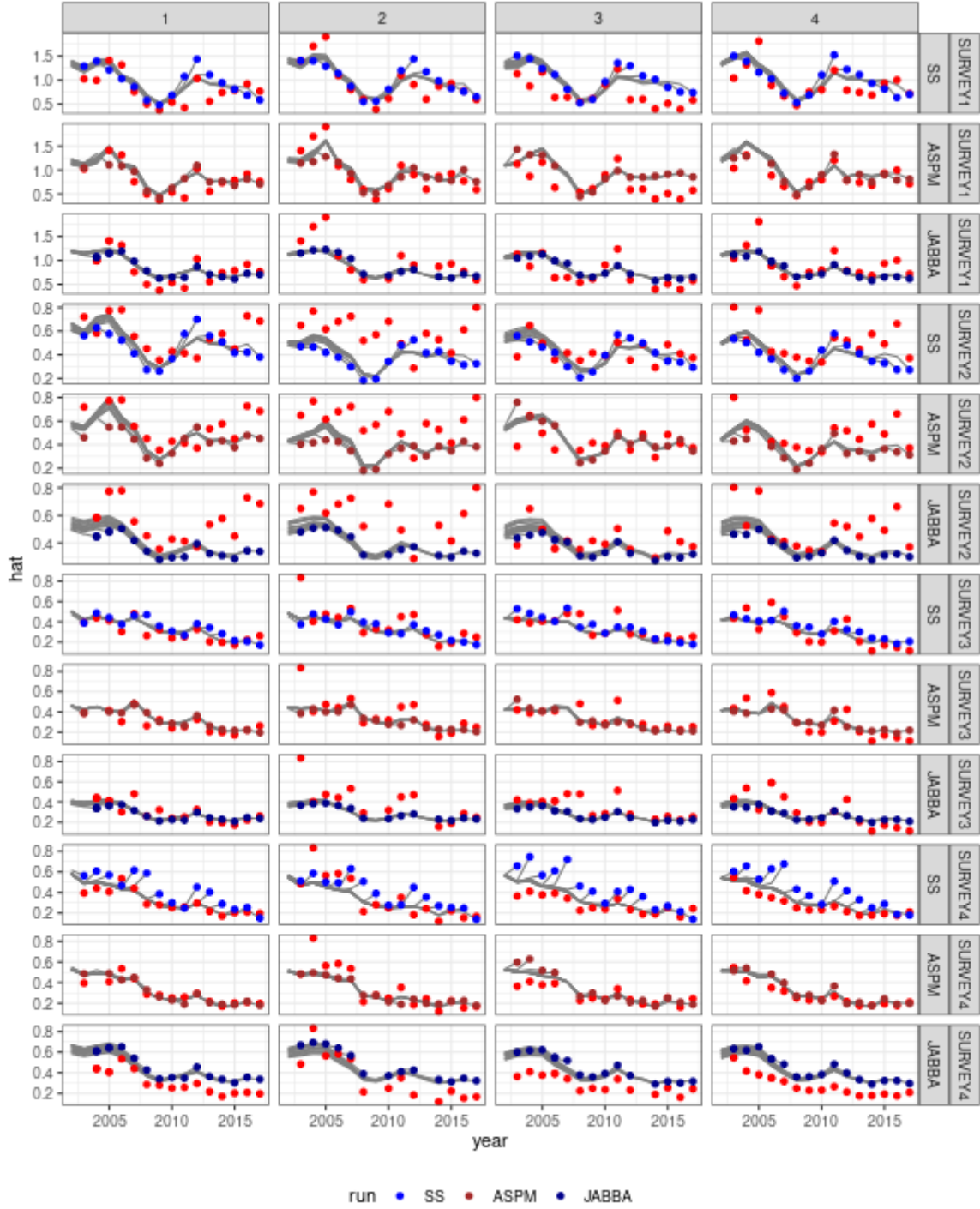


Figure 5: Hindcast with one year ahead predictions, red dots are the observed CPUE values and lines are the fits with terminal hincast year indicated by a point.

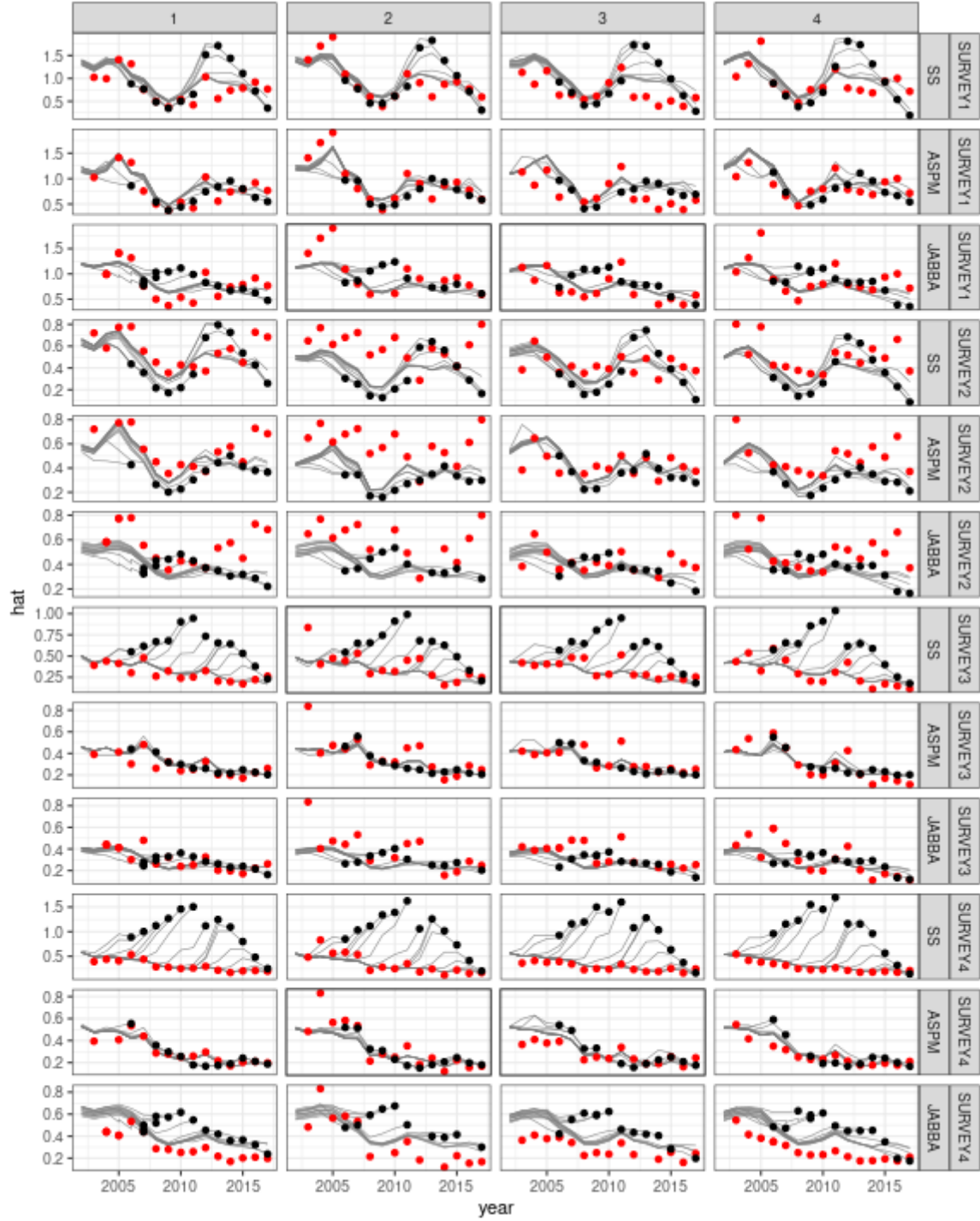


Figure 6: Hindcast with three year ahead predictions, red dots are the observed CPUE values and lines are the fits with terminal hincast year indicated by a point.



Figure 7: Runs tests for one step ahead residuals.

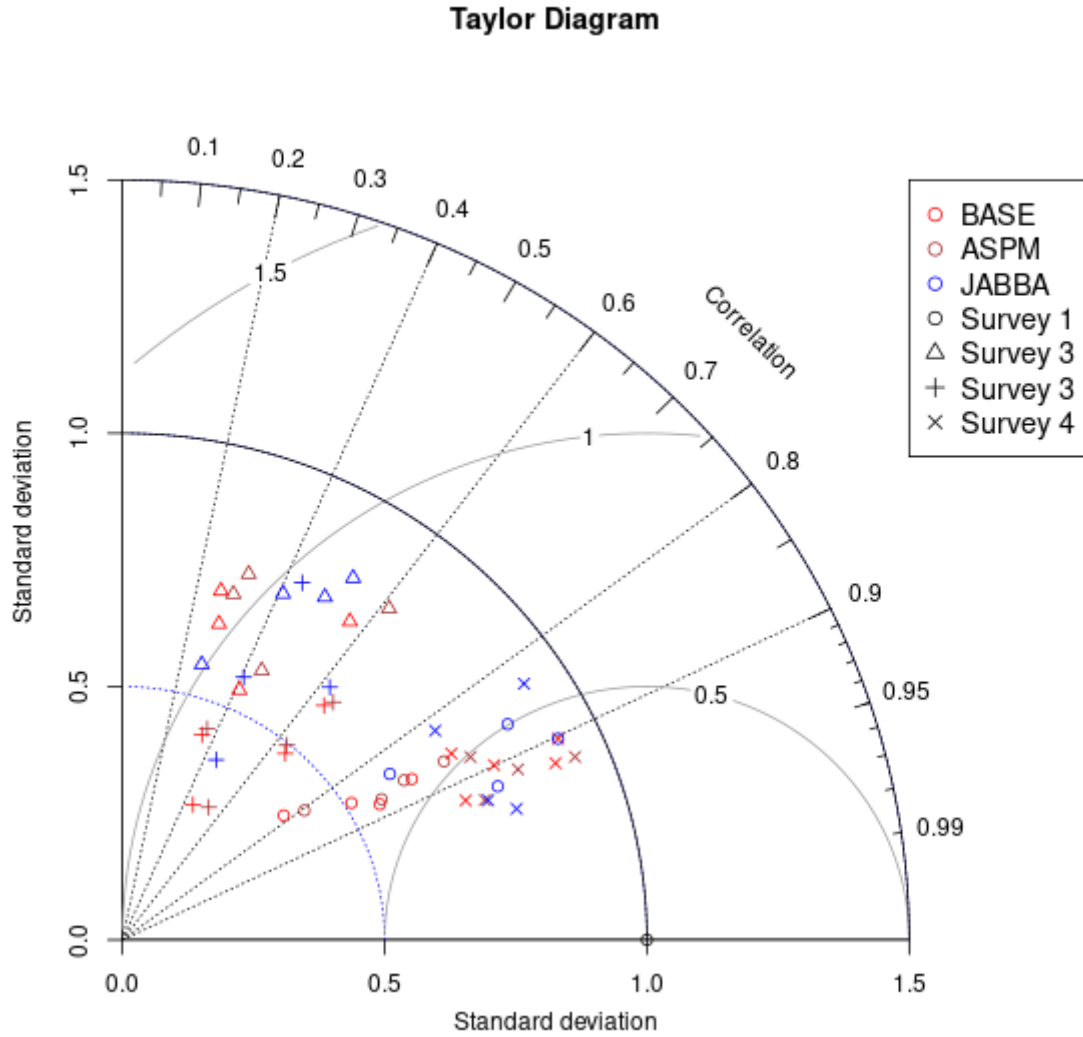


Figure 8: Taylor diagram for fits to CPUE summarising the similarity between the observed time series of CPUEs and the predicted relative stock abundance. Each point quantifies how closely predictions match observations, the angle indicates the correlation, the centred root-mean-square error difference 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.)

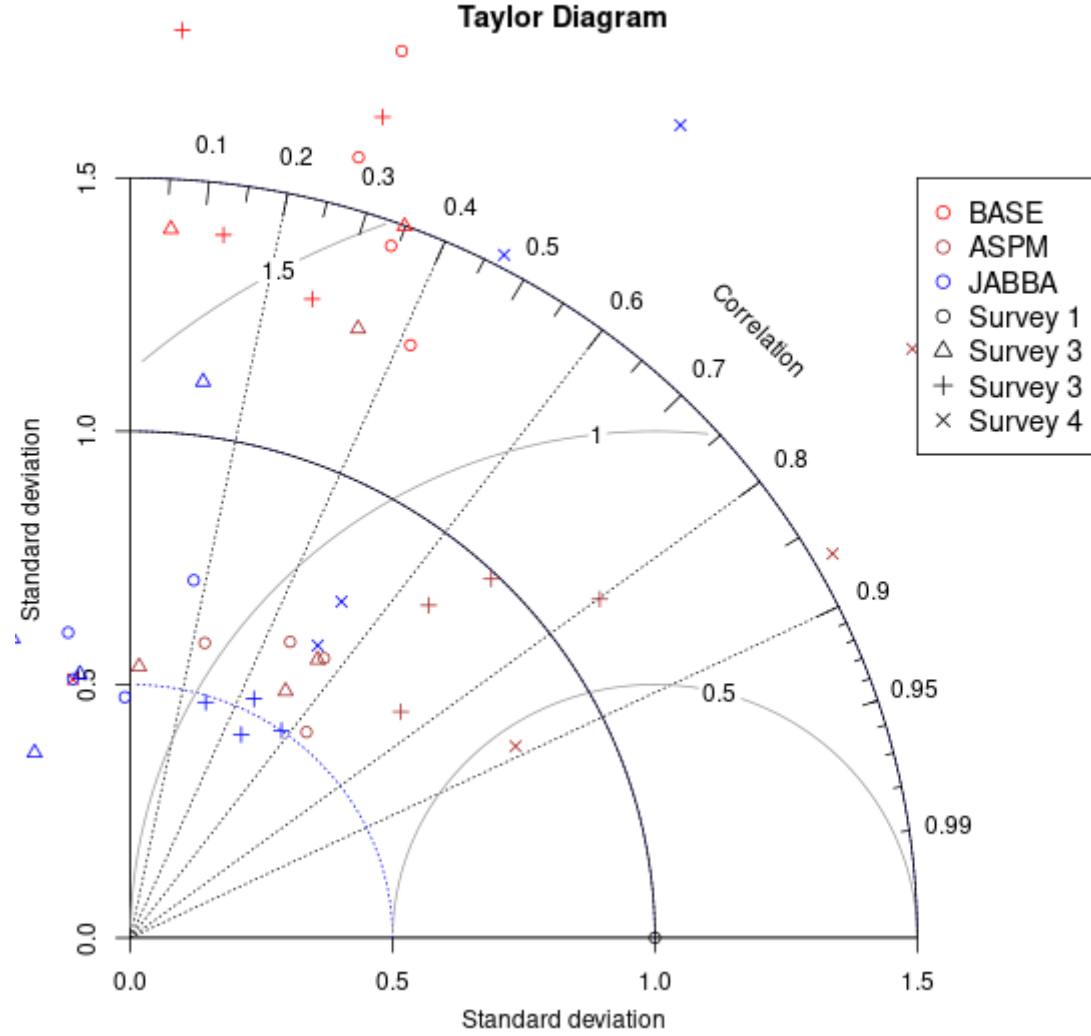


Figure 9: Taylor diagram for 3 year ahead predictions, summarising the similarity between the observed time series of CPUEs and the predicted relative stock abundance. Each point quantifies how closely predictions match observations, the angle indicates the correlation, the centred root-mean-square error 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.