Fisheries Research 183 (2016) 119–127

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

# Evaluation of the prediction skill of stock assessment using hindcasting

Laurence T. Kell[a,*], Ai Kimoto[b], Toshihide Kitakado[c]

[a] ICCAT Secretariat, C/Corazón de María, 8, 28002 Madrid, Spain
[b] Bluefin Tuna Resources Division, National Research Institute of Far Seas Fisheries, Fisheries Research Agency, 5-7-1 Orido, Shimizu, Shizuoka 424-8633, Japan
[c] Faculty of Marine Science, Tokyo University of Marine Science and Technology, Department of Marine Biosciences, 5-7, Konan 4, Minato-ku, Tokyo 108-8477, Japan

ARTICLE INFO

*Article history:*
Received 9 February 2016
Received in revised form 18 May 2016
Accepted 19 May 2016
Handled by A.E. Punt
Available online 7 June 2016

*Keywords:*
Abundance indices
Cross-validation
Projection
Retrospective analysis
Stock assessment
Taylor diagrams

ABSTRACT

A major uncertainty in stock assessment is the difference between models and reality. The validation of model prediction is difficult, however, as fish stocks can rarely be observed and counted. We therefore show how hindcasting and model-free validation can be used to evaluate multiple measures of prediction skill. In a hindcast a model is fitted to the first part of a time series and then projected over the period omitted in the original fit. Prediction skill can then be evaluated by comparing the predictions from the projection with the observations. We show that uncertainty increased when different datasets and hypotheses were considered, especially as time-series of model-derived parameters were sensitive to model assumptions. Using hindcasting and model-free validation to evaluate prediction skill is an objective way to evaluate risk, i.e., to identify the uncertainties that matter. A hindcast is also a pragmatic alternative to hindsight, without the associated risks. While the use of multiple measures helps in evaluating prediction skill and to focus research onto the data and the processes that generated them.


## 1. Introduction

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. In most fishery management frameworks a stock is defined on operational rather than an ecological or evolutionary basis (Waples and Gaggiotti, 2006). In this paper a stock is defined as a population or subpopulation of a species for which parameters such as growth, recruitment, mortality, and fishing mortality are regarded as being homogeneous, and which have the main effect on determining the dynamics; extrinsic factors such as immigration and emigration are traditionally ignored.

Stock assessments 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). 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 stock assessment models as it is seldom possible to observe fish populations directly. Therefore techniques such as retrospective analysis, where a model is fitted to increasing periods of data to identify systematic inconsistencies (Mohn, 1999), or simulation are used. Deroba et al. (2015) summarised an extensive state-of-the-art simulation exercise to compare stock assessment models. This was limited to the evaluation of historical and current estimates of stock status based on self- and cross-tests. Both approaches evaluate consistency rather reliability, where a reliable model provides accurate results despite uncertainty.

One approach to address uncertainty in historical estimates of stock status is to integrate multiple diverse datasets to try and extract as much information as possible about modelled processes (Fournier et al., 1998). An implicit assumption is that integrated models can compensate for lack of good data. Models are by definition, however, simplifications of reality and model misspecification can lead to degradation of results when there are multiple potentially conflicting data sets. For example Payne et al. (2009) showed that including all available data in stock assessments may lead to high noise levels and poor-quality assessments, and recommended that the choice of data should be based on rational and justifiable selection criteria. It is therefore critical to determine what drives an assessment (Francis and Hilborn, 2011).

* Corresponding author.
*E-mail address:* Laurie.Kell@iccat.int (L.T. Kell).

http://dx.doi.org/10.1016/j.fishres.2016.05.017
0165-7836/

To check that predictions are consistent with reality it is necessary to evaluate prediction skill (e.g., Walters and Punt, 1994; Patterson et al., 2001; Ralston et al., 2011); a statistical evaluation of the accuracy of a prediction relative to a reference model or dataset. Prediction skill can be used to compare alternative models or observations used for prediction 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. Stock biomass cannot actually be observed so if estimates of population abundance were compared in the hindcast this would be model-based validation.

Errors and uncertainty in historical parameter estimates, particularly in the most recent years, propagate into predictions. Different stock assessment packages often use different methods for estimating and propagating those errors, and the choices made will affect the robustness of management advice (Patterson et al., 2001; Magnusson et al., 2012). Validation of predictions is therefore as least as important as examining diagnostics for fits to historical data. More effort, however, appears to be going into the latter than the former, unlike in other fields such as meteorology and oceanography (e.g., Murphy and Winkler, 1987; Doswell III et al., 1990; Schaefer, 1990; Roebber, 2009), where the ability to predict is more important than the description of past states.

Hindcasting is widely used, in oceanography and meteorology where the state of a system is observable, to evaluate prediction skill (Huijnen et al., 2012). Hindcasting is a conceptually simple form of cross-validation, which has no parametric or theoretic assumptions allowing it to be used for comparisons across different models and datasets. In a hindcast, a model is first fitted using a truncated time series, dynamics are projected forward using the model and predictions compared to recent observations not used in fitting (e.g., Christoffersen and Pelletier, 2004; Pastoors et al., 2007; Heath et al., 2004). Although hindcasting is not commonly used in stock assessment it combines two individual procedures routinely used, namely retrospective analysis and projection.

An objective of the paper is to show how hindcasting and using multiple measures for prediction skill can help in the development of robust stock assessment advice frameworks. We use hindcasting to evaluate the prediction skill of series of catch per unit effort (CPUE) used as indices of stock abundance, across a range of stock assessment scenarios. Since time series of CPUE are often the most influential inputs to stock assessment models (Francis and Hilborn, 2011) it is important to be aware of the limitations of these data when fitting models with them.

## 2. Materials and methods

We chose Atlantic and Mediterranean bluefin (*Thunnus thynnus*) as a case study to show how hindcasting can provide insight into stock assessment uncertainty, to illustrate the benefits of the approach and help identify ways forward. A reason for the choice is because it is a stock of high value with well documented uncertainty about current stock status and response of the stock to management (Fromentin et al., 2014; Leach et al., 2014).

### 2.1. Stock assessment

Atlantic bluefin is assessed by the International Commission for the Conservation of Atlantic Tuna (ICCAT) using Virtual Population Analysis (VPA). VPA sums catch numbers-at-age backwards down a cohort, adjusted by losses due to natural mortality (Pope, 1972). Indices of relative stock abundance allow the numbers in the oldest age class of a cohort (terminal numbers-at-age) to be derived using maximum likelihood.

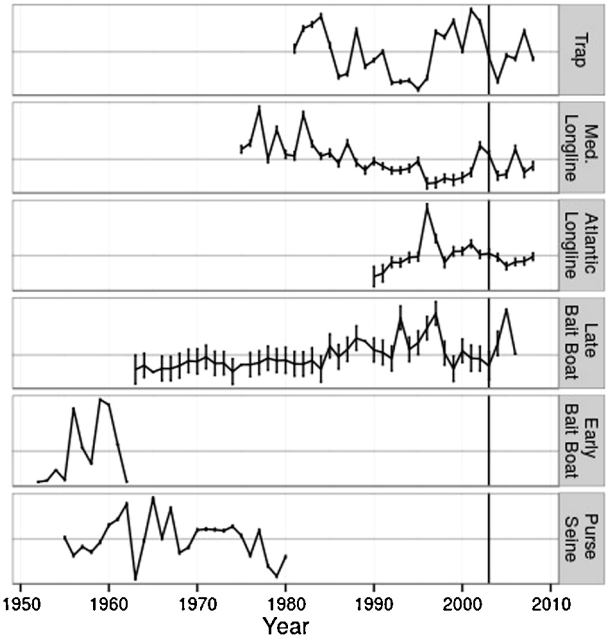


**Fig. 1.** Time series of CPUE, the error bars are the CVs derived from the standard errors of the GLM predictions; the vertical line corresponds to 2003 the start of the hindcast (standardized to have a mean of 0 and variance of 1).

Two main datasets are used in the Atlantic bluefin assessment (ICCAT, 2015), i.e., catch-at-age and CPUE. Catch-at-age data are derived from total reported catches and samples of size data using age slicing. The raw catch and effort data are standardised using generalised linear models (GLMs) to remove the effect of factors that bias CPUE when used as an index of abundance (Maunder et al., 2006). Although the majority of the catch comes from the Mediterranean purse seine fishery, catches taken with this gear do not provide a reliable estimate of stock abundance and the CPUE from this gear are not used in the assessment.

Time series of catch numbers-at-age start in 1950 and the assessment included six CPUE series from fleets using four fishing gears (Table 1; Fig. 1). Data after 2008 are not used in this study since the implementation of the bluefin recovery plan affected catch rates and the selection of age classes by the fisheries. Fig. 1 shows the time series of CPUE, the error bars are the CVs derived from the standard errors of the GLM predictions. Only four series covered the recent period, namely trap, Mediterranean and Atlantic long line and late period bait boat. All the gears target large bluefin tuna except the bait boats that target juveniles. As the purse seine and early period bait boats series do not cover the recent period they were not included in the hindcast analysis.

The time series of CPUE are in biomass and represent all ages in the catch. Therefore when fitting the VPA these have to be transformed into numbers-at-age based on the vulnerability of age-classes to the fishery and the mass-at-age, i.e.

$$\hat{I}_{iy} = q_i \delta_i \Sigma_a v_{ia} w_{iay} \tilde{N}_{ay} \quad (1)$$

where $q_i$ is the catchability coefficient, $\delta_i$ an adjustment for time of fishing, $v_{ia}$ the relative vulnerability-at-age, $w_{iay}$ mass-at-age, and $\tilde{N}_{ay}$ the estimated of numbers-at-age. The subscripts are $i$ for the CPUE series, $a$ for age and $y$ for year.

$v$ is given by

$$v_{ia} = \frac{\Sigma_y C_{iay} F_{ay} / C_{ay}}{\max_a \{C_{iay} F_{ay} / C_{ay}\}} \quad (2)$$

**Table 1**
Catch per unit effort indices: including a summary of fishing gear, period, target ages, main operating area, season, and proportion of the catch taken.

| Series | Gear | Period | Ages | Area | Season | Catch |
|---|---|---|---|---|---|---|
| Trap | Trap | 1981–2008 | 6+ | Gibraltar Straits | Spring | 8% |
| Med. long line | Long line | 1975–2008 | 6+ | East Atlantic and Med. | Spring | 4% |
| Atlantic long line | Long line | 1990–2008 | 4+ | Northeast Atlantic | Autumn–winter | 4% |
| Late bait boat | Bait boat | 1963–2006 | 2–3 | Bay of Biscay | Summer | 10% |
| Early bait boat | Bait boat | 1952–1962 | 5–6 | Bay of Biscay | Summer | 6% |
| Purse seine | Purse seine | 1955–1979 | 10+ | North Sea | Autumn | 10% |

**Table 2**
Virtual population analysis scenarios; there are also three recruitment scenarios (low, medium and high) giving 21 scenarios in total.

| Scenario | Indices | Catch | F-Ratio | M |
|---|---|---|---|---|
| Base Case | All | Reported | Fixed | Varies by age |
| Ex-LL | Exclude Atlantic LL | Reported | Fixed | Varies by age |
| Inflated | All | Inflated | Fixed | Varies by age |
| Revised | All | Revised | Fixed | Varies by age |
| West M | All | Reported | Fixed | 0.14 all ages |
| F-Ratio | All | Reported | Annual | Varies by age |
| Catch-Curve | All | Reported | External | Varies by age |

where $C$ is the partial catch-at-age by year and fleet and $F$ fishing mortality-at-age estimates from the VPA.

### 2.2. Scenarios

When conducting a stock assessment it is common practice to define a Base Case, which comprises a set of model assumptions and specifications thought to be the most likely. In this study the Base Case used the official catches reported to ICCAT and the CPUE indices in Table 1. Natural mortality was assumed to be time-invariant but to vary by age (i.e., 0.490, 0.240, 0.240, 0.240, 0.240, 0.200, 0.175, 0.150, 0.125, 0.100 per year for ages 1–10, respectively). A value for fishing mortality in the plus group (10+) is required to estimate the terminal numbers-at-age in each cohorts. The plus group F is derived from the F-Ratio, the ratio between the plus group and the last true age in a cohort (9) which was fixed at 0.7, 1.0, 0.6, 1.2, 1.0 for the periods 1950–1969, 1970–1984, 1985–1994, 1995–2007 and 2008 onwards, respectively.

Uncertainty about the actual dynamics (i.e., model uncertainty) can have a large impact on achieving management objectives (Punt, 2008). For example estimates of stock status are highly sensitive to the assumptions about natural mortality-at-age (Jiao et al., 2012) and the vulnerability of age classes to the fisheries (Brooks et al., 2010). While the relationship between stock and recruitment, required to perform projections and calculate reference points (Sissenwine and Shepherd, 1987) is difficult to estimate in practice due to lack of contrast in stock assessment datasets (Pepin and Marshall, 2015). Therefore a number of scenarios, corresponding to alternative datasets and model assumptions, were considered when conditioning the stock assessment model.

The main uncertainties in the last assessment were represented by two types of scenarios (Table 2); those that changed the input datasets (Ex LL, Inflated and Revised catch) and those where fixed parameters were altered or estimated (West M, F-Ratio and Catch Curve). Based on these scenarios hindcasting can be used to identify where assumptions are violated, models that have prediction skill and datasets with the best signal to noise ratio.

The impact of the choice of CPUE series was explored by removing the Atlantic long line series (Ex-LL scenario). Uncertainty about the actual level of catches was evaluated by using an inflated catch series (i.e., raised to 50,000 t from 1998 to 2006 and to 61,000 t in 2007; inflated scenario) and a recent revision of the catch-at-age that incorporated new historical information (revised scenario).

The impact of the assumed values for natural mortality was evaluated by changing the values to those assumed in the western stock (i.e., 0.14 per year for all ages, West M scenario). The F-Ratio is difficult to estimate when conducting a VPA and in the Base Case was fixed based on expert judgement. Two alternative scenarios were used to estimate the F-Ratio namely either it was estimated each year as a random walk (i.e., $X_t = X_{t-1} + \eta_t$; $\eta_t \sim N(0, \sigma^2)$, F-Ratio scenario) or fixed based on a Catch Curve analysis of length data independent of the stock assessment model (Catch Curve scenario).

The relationship between stock and recruitment, required to calculate reference points and perform projections, is difficult to estimate thus three future recruitment scenarios were considered, corresponding to high, medium, and low recruitment levels (geometric mean recruitment estimates from 1995 to 2006, 1990 to 2000, and 1970 to 1980, respectively). The projections were performed using the reported total catch and the mean selection pattern from the last three years in the historical assessment to create a catch-at-age matrix based on the reported catches.

### 2.3. Hindcasting

To conduct a hindcast involves fitting a model 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 $t$.

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 in our example $n$ is 5 and $t$ is the last year in the time series, i.e., time series were cut at 2003 and projected from 2004 to 2008.

When conducting a hindcast, as when conditioning stock assessment models, it is assumed that modelled variables are observable, processes exhibit constancy of structure in time, including those not specified in the model, and that collection of accurate and sufficient data is possible (Hodges et al., 1992).

### 2.4. Prediction skill

A well-fitting model results in predicted values close to the reference values, in this case the CPUE observations from 2004 to 2008. The best statistical measure to use depends on the objectives of the analysis, and using more than one measure can be helpful in providing insight into the nature of observation and process error structures.

Root mean square error (RMSE, $E'$) is the square root of the variance of the residuals and indicates how close the observed data points are to the predicted values i.e.

$$E' = \sqrt{\frac{\sum_{y=t-n}^{t}(I_y - \hat{I}_y)^2}{n+1}} \quad (3)$$

Computing the relative errors allows comparisons across the data series to be made. As the square root of a variance it can also be interpreted as the standard deviation of the unexplained variance,

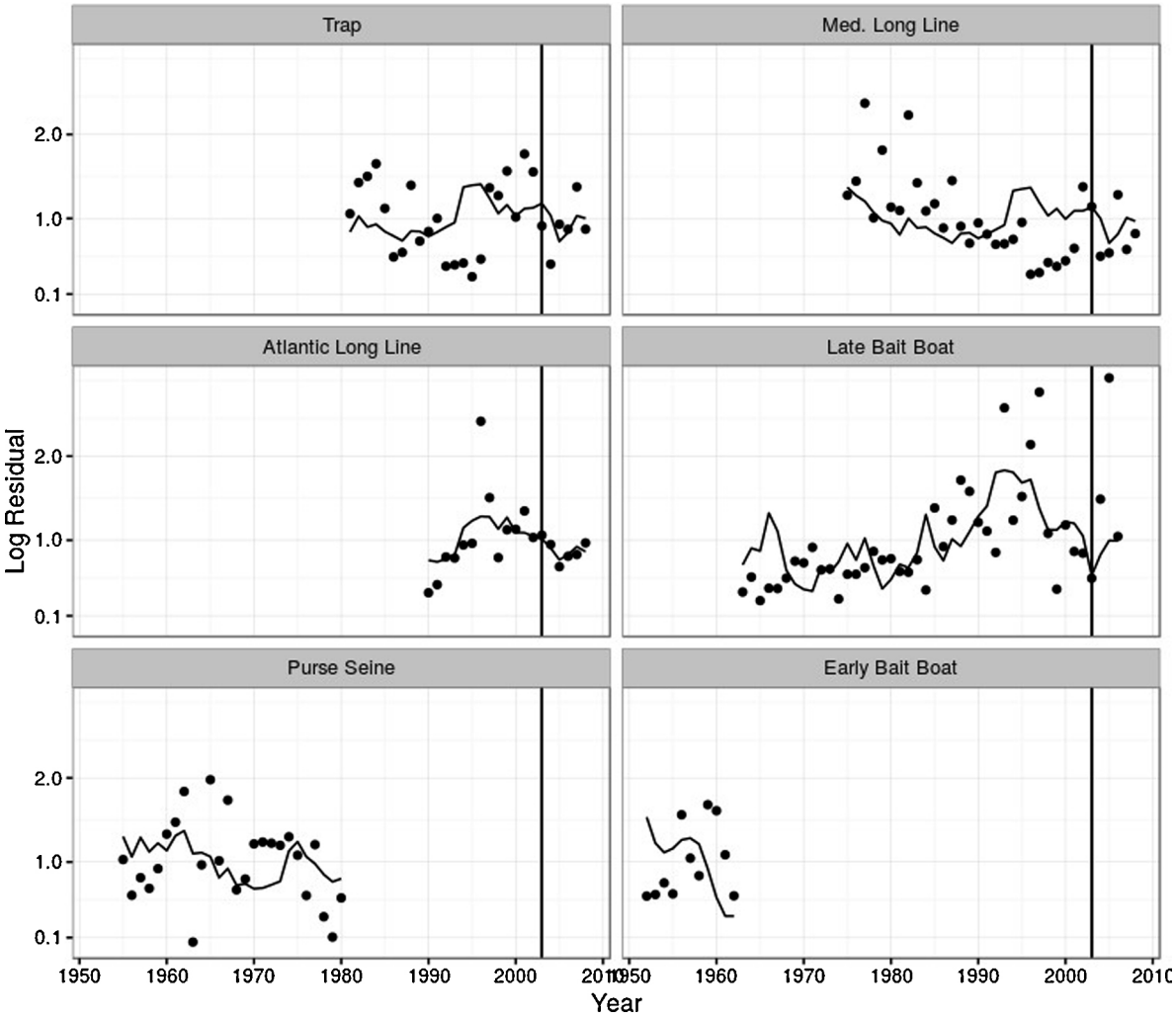


**Fig. 2.** Time series of CPUE (points) with estimates of relative abundance for the Base Case stock assessment (lines); the vertical line corresponds to 2003, the start of the hindcast.

lower values indicate better fits. $E'$ is sensitive to outliers, however, and favours forecasts that avoid large deviations from the mean and cannot be used to compare across series. In comparison, the correlation ($\rho$) between $I_y$ and $\hat{I}_y$ is not affected by the amplitude of the variations, is insensitive to biases and errors in variance, and can be used to compare across series.

$E'$, $\rho$ and the variance of the predictions ($\sigma_f^2$) and observations ($\sigma_o^2$) are related by the cosine rule

$$E'^2 = \sigma_o^2 + \sigma_f^2 - 2\sigma_o\sigma_f\rho \quad (4)$$

The reference set ($o$) are the observed CPUE not included in the retrospective assessment and the values ($f$) are the projected CPUE series up to and including 2008.

This means that $E'$, $\rho$ and $\sigma_f$ can be summarised simultaneously in a single diagram (Taylor, 2001). Taylor diagrams therefore provide a concise statistical summary of how well patterns match each other and are therefore especially useful for evaluating multiple aspects or in gauging the relative skill of different models (Griggs and Noguer, 2002).

Modelling was carried out using FLR (Kell et al., 2007) designed to build simulation models representing alternative hypotheses about stock and fishery dynamics.

## 3. Results

First the CPUE series and their fits are summarised, then Taylor diagrams are used to evaluate the prediction skill of the various scenarios using the individual CPUE series as model-free validation measures. The observed CPUE time series (points) and the VPA fits using all the data are shown in Fig. 2. The indices are in biomass, summed over ages and split into ages within the likelihood estimation using Eq. (2) above.

The difference between the fits and the observations (the residuals) corresponds to the unexplained variance. To summarise the variance the residuals for the CPUE fits are plotted against the estimates from the Base Case and medium recruitment scenario in Fig. 3; only four series are shown as the early bait boat and purse seine data do not cover the hindcast period. The points from the historical stock assessment are shown in grey and from the hindcast projection in black. The variance of the projections ($\sigma_f^2$) appears to be less than that of the historical assessment. The CPUE are assumed to be relative indices of stock abundance, therefore the observations are plotted against the fits for each series as a check that the regression falls along the $y = x$ line as expected (Fig. 4). In the case of the Atlantic long line and trap, the outliers appear to have a strong effect on the regression and some indices appear to be poor fits.

$E'$, $\rho$ and $\sigma_f$ are plotted simultaneously using Taylor diagrams; values are tabulated in the appendix provided as Online

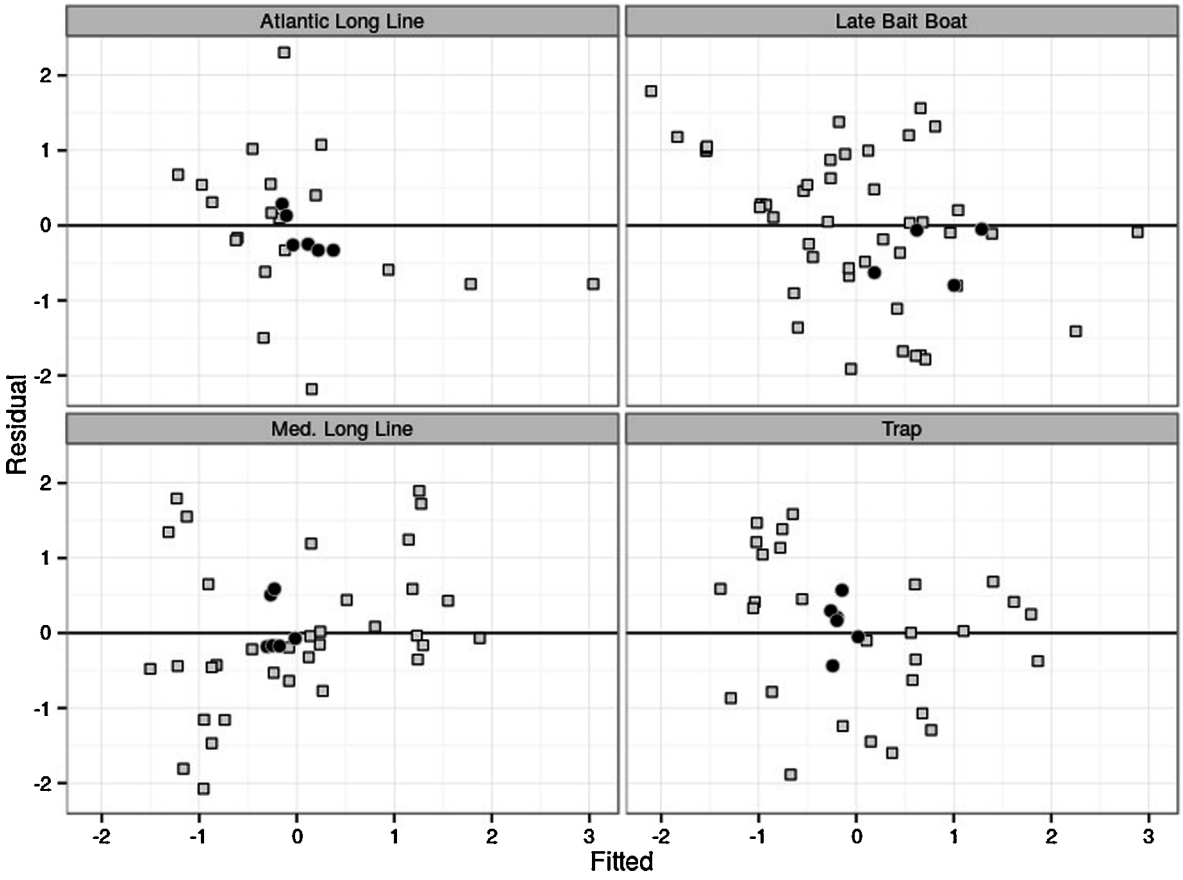


**Fig. 3.** Plots of residuals verses fitted CPUE by series for the Base Case stock assessment fitted up to 2003 and projected with medium recruitment up to 2008; filled circles are data points from 2003 to 2008.

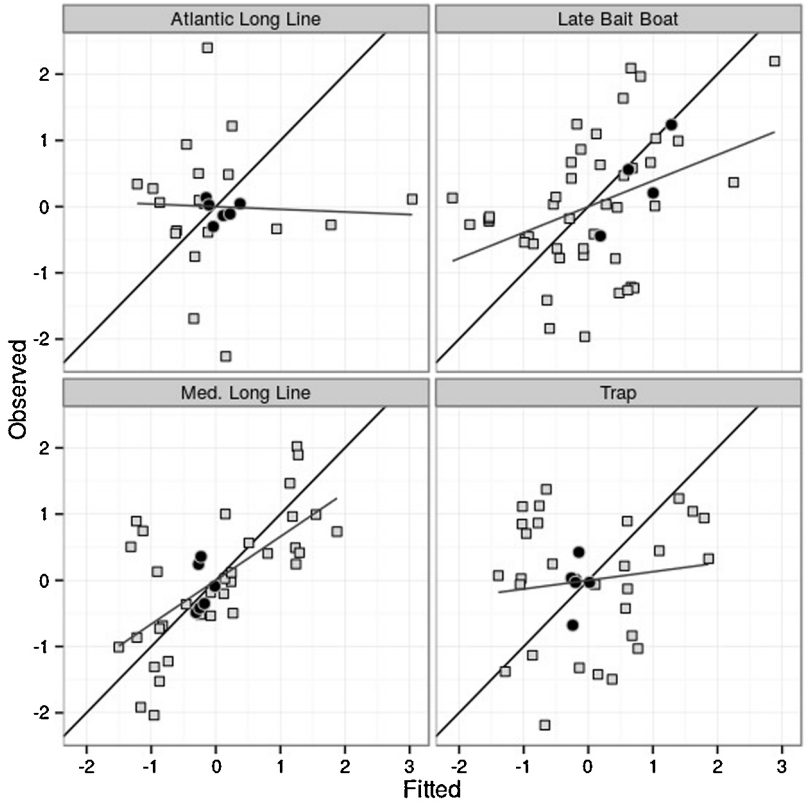


**Fig. 4.** Plots of observed verses fitted CPUE by series for the Base Case stock assessment up to 2002 and projected with medium recruitment up to 2008; also included are a linear regression (using all data) fitted to the points and the $y = x$ line; filled circles are data points from 2003 to 2008.

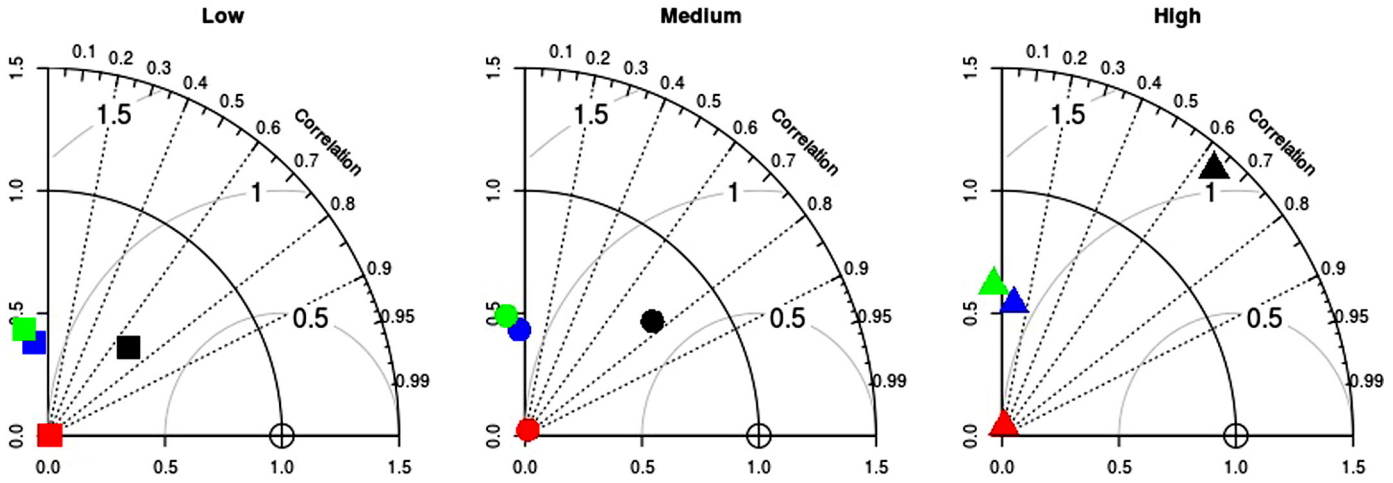


**Fig. 5.** Taylor diagram summarising the similarity between the observed time series of CPUEs and the predicted relative stock abundance, for the Base Case by recruitment scenario. Each point quantifies how closely predictions match observations, the angle indicates the correlation, the centred root-mean-square 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 CPUE series, Atlantic long line (black), Mediterranean long line (green) trap (blue) and bait boat (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

Supplementary Material. The results for the Base Case are shown in Fig. 5 with recruitment scenarios by panel. The CPUE series are Atlantic long line (black), Mediterranean long line (green), trap (blue) and bait boat (red); recruitment levels are low (square), medium (dot) and high (triangle).

The closer a point is to the open circle (the cross hairs) at 1.0 on the *x*-axis the better is the prediction skill. The centred RMSE is proportional to the distance to the cross hairs on the *x*-axis (i.e., the contours around the cross hairs indicate $E'$); correlation between predictions and observations is indicated by the angle from the vertical of the radii on which a point falls; and the standard deviations of the predictions is proportional to the radial distance from the origin (scaled so $\sigma_o = 1$). $E'$ can be used to compare within a series but is sensitive to noise; $\rho$ is less sensitive to outliers and can be used to compare across CPUE; while $\sigma_f$ provides a measure of how smooth are the forecasts and is useful for understanding differences between $E'$ and $\rho$.

Only Atlantic long line shows a correlation greater than 0.6, the other CPUE series all lie on or close to the *y*-axis and so show low or negative correlations. For Atlantic long line, as the recruitment level increased from low to medium, $\rho$ increased from 0.69 to 0.76, $E'$ from 0.91 to 0.97 and $\sigma_f$ from 0.50 to 0.72. Increasing recruitment level from medium to high resulted in a slight decrease in $\rho$ (0.64) but large increases in $E'$ and $\sigma_f$ (2.10 and 1.42). Since $E'$ is sensitive to changes in variance.

All scenarios are compared to the Base Case in Fig. 6, panels are for each assessment scenario; as before colours correspond to CPUE series and shapes to recruitment level. ICCAT explored two types of historical assessment scenarios, i.e., those that changed the input datasets (second row; i.e., Ex LL, Inflated and Revised catch) and those where fixed parameters were altered or estimated (third row; i.e., West M, F-Ratio, and Catch Curve). Values of $E' > 2.0$ and negative correlations less than −0.3 are not shown in the plots; e.g., trap values for Inflated, F-Ratio and Catch Curve scenarios. The bait boat has low prediction skill across all scenarios i.e., $\rho$ is close to 0 or even negative. There is little to choose between scenarios as differences in $E'$ are small and observations are much more variable than the predictions as $\sigma_f$ is small. This is because there is large uncertainty in numbers-at-age in recent incomplete cohorts as there is little information in either the CPUE series or the catch-at-age on incoming year-class strength. Atlantic long line has the best prediction skill across all the scenarios.

Excluding the Atlantic long line CPUE series (Ex LL) improves the prediction skill of the Trap and Mediterranean long line. However, little improvement is seen for bait boat. Inflating the total catch without also revising the catch-at-size accordingly (Inflated) results in poor prediction skill. Revising the catch-at-size (Revised) has little effect compared to the Base Case.

When the parameters, which had been fixed in the Base Case, are altered or estimated (third row) the prediction skill of Trap, Mediterranean long line and bait boat is not noticeably improved. Changes are seen, however, for the Atlantic long line CPUE series. If *M* (West M) is assumed to be constant at age, natural mortality of younger ages is reduced, and variability and $E'$ increases relative to the Base Case but $\rho$ is relatively insensitive. When the assumptions about fishing mortality in the plus group are changed results for the two scenarios (F-Ratio and Catch Curve) are similar; namely in the low recruitment scenario $E'$ increased and $\rho$ decreased (i.e., prediction skill declined), while $\sigma_f$ is less effected. Greater variability was seen in the Catch Curve scenario. This contrasts with the Base Case where increasing recruitment had little effect on $\rho$, and high recruitment increased $E'$ and $\sigma_f$, i.e., added more variability.

## 4. Discussion

The objectives of the study were to show how hindcasting, model-free validation, and the evaluation of prediction skill using tools such as Taylor diagrams to visualise multiple measure can help in the development of robust advice frameworks (e.g., Murphy and Winkler, 1987; Doswell III et al., 1990; Schaefer, 1990; Roebber, 2009).

An insidious problem in stock assessment is systematic changes in model-derived quantities that occur as additional years of data are added or removed. The main cause of such retrospective patterns are processes that do not exhibit constancy of structure in time, and can lead to severe errors in management advice (Hurtado-Ferro et al., 2015). Therefore often a first task in stock assessment is to establish the credibility of predictions by updating the last assessment and comparing the estimates with forecasts made previously. Hindcasting can provide a formal and objective framework for doing this, allowing different model structures and stock assessment packages to be compared.

Model-free validation allows the impact of different input datasets, their information content, and signal to noise ratio to

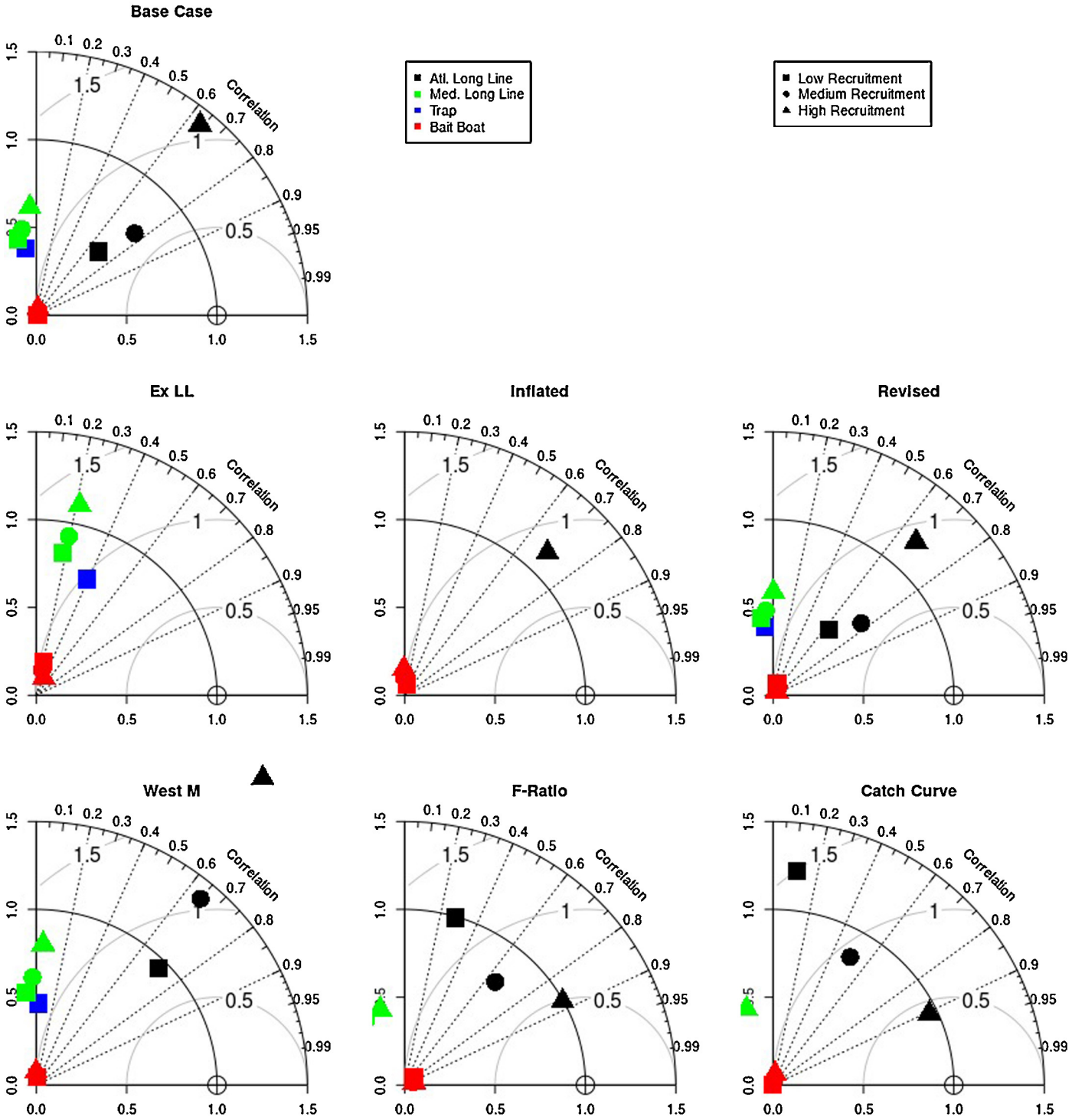


**Fig. 6.** Taylor diagram comparing all scenarios and CPUE series. Values of RMSE that are off the scale (i.e., >2.5) and negative correlations are not shown, e.g., trap values for Inflated, F-Ratio and Catch Curve scenarios. (For interpretation of the references to colour in text, the reader is referred to the web version of the article.)

be evaluated. This is difficult to do by comparing model outputs using methods such as AIC. In the former case if stock trajectories differ between assessment models it can be difficult to make a rational choice between models. While in the latter if different datasets are used then AIC cannot be used to compare models. Using hindcasting also helps to change the focus from historical estimates of stock status to prediction skill and hence management.

In fisheries science and management, uncertainties and the risks they create are pervasive owing to natural variability in aquatic ecosystems, imperfect information, and lack of perfect control over fisheries (Peterman, 2004). For these reasons, Hilborn (2016) observed that fisheries is decision making under uncertainty, and recommended that when actions are proposed, their consequences should be evaluated across as many possible hypotheses as possible. Otherwise there is a danger of assuming that the strongest

correlation identifies a causal relationship and subsequently ignoring other equally plausible hypotheses (Plowright et al., 2008). Using Taylor diagrams to summarise prediction skill allows models conditioned on a wide range of scenarios based on different datasets and alternative model structures to be evaluated simultaneously. While using multiple measures to represent more than one attribute of prediction skill helps provide insight on the reliability of the measures and the error structure of the data.

The study showed that uncertainty increases when multiple datasets are available, as different datasets and model assumptions may result in contradictory outcomes. For example the recruitment time-series were sensitive to the model assumptions. Dickey-Collas et al. (2015) reported a similar finding, in a study based on North Sea herring, where the characteristics of time-series were determined by the model used to generate them rather than underlying ecological phenomena alone. This is especially true when information about catch-at-age and hence cohort abundance is noisy or biased.

In the case of the bluefin assessment scenarios were originally chosen as sensitivity tests and then combined to provide advice on current stock status and time to recover the stock to an underfished state. There was no attempt to weight (either a priori or a posteriori) or to reject scenarios (i.e., hypotheses). An alternative approach could be to take a risk-based approach, where risk is an uncertainty that matters and what matters are management objectives. For example the scenarios where the catch was revised gave very similar results to the Base Case, while the choice of F-Ratio scenario (i.e., the assumed plus group dynamics) had a large effect. Weighting the Base Case, Revised Catch and the F-Ratio scenarios equally may result in unacceptable risks of forgoing yield or overfishing. For example if the management objective is to recover the stock with a 60% probability and the first two lead to high probabilities of stock recovery and the F-Ratio scenario a low probabilities combining the scenarios could give a false perception of the consequence of adopting particular management measure. To provide robust advice you either have to resolve which is the "correct" scenario or make sure that advice achieves management objectives for all scenarios.

A main assumption in most stock assessment models is that CPUE is proportional to stock abundance over the exploitation history and geographic range of a stock (Maunder et al., 2006). However, CPUE series are often conflicting (Schnute and Hilborn, 1993) and the appropriate method to deal with such data conflicts depends on the error structure, which depends on whether they are caused by random sampling error, process variation, observation model misspecification, or system dynamics model misspecification, and the ratio of signal to noise. Using multiple measures helps in understanding the error structure and how assumptions and data affect the robustness of advice (Le Cren and Holdgate, 1962). For example if the objective is to rank models, results using $E'$ may not be stable if random errors are large (Freyer, 2014). Model-free validation and hindcasting is an objective way to compare models based on predictions and observations. It also has the advantage of focusing research on the data themselves and the processes which generated them.

If a model is erroneous, however, in that it does not describe the dynamics, then a hindcast alone can not validate the model as it is a test of consistency rather than robustness. For example, in a fishery where both catch and effort had been constant over time, a stock assessment model conditioned on the catch and CPUE would interpret this as the stock and fishing mortality being stable. Catchability, however, may increase as the stock declines due to hyperstability (Erisman et al., 2011). Although projecting the model predicts the data (i.e., no trend in CPUE) as the assessment on which they are based is biased the stock is actually be declining (e.g., Dickey-Collas et al., 2010).

## 5. Conclusions

Making sense of outputs from stock assessment is challenging, especially when the data and model outputs are uncertain. Dealing with uncertainty is a multi-faceted process that includes the quantification, modelling, propagation and visualisation of the uncertainties and errors intrinsic in the data and arising from data transformations (Correa and Lindstrom, 2013). Often in stock assessment this is done by presenting multiple plots based on the residuals or pairwise comparisons, and a full comparison among a large number of models quickly becomes impractical.

Fisheries, as in many industries requires forecasts to make strategic decisions. Hollander (2015) summarised some common failings when making model-based forecasts including presenting a large number of model runs or scenarios with limited interpretation, avoiding clear statements on how sure we are about projections, and avoiding discussion about the history of the model, which may go back many years and may not be fully understood by current users. Taylor diagrams help tackle these failing by allowing comparisons to be made across a range of scenarios and by focusing on the prediction skill of different datasets and models.

Although a comprehensive evaluation of robustness and the value-of-information can be conducted using MSE and multiple hypotheses, this is a complex and time consuming process and unlikely to be routinely applied by stock assessment working groups. In contrast hindcasting is conceptually simple as it has no parametric or theoretic assumptions and can be used for any stock and for comparison of methods. A hindcast can help to identify how data and assumptions carry through into future projections, i.e., does the overall character of a projection change when different models and datasets are considered. It can also be used to help identify potential time series to use in empirical or model based harvest control rules (e.g., Hillary et al., 2016; Carruthers et al., 2016).

## Acknowledgements

This study does not necessarily reflect the views of ICCAT and in no way anticipates the Commission's future policy in this area. Ai Kimoto was supported by the overseas research program of the Fisheries Research Agency, Japan. The authors would also like to thank the reviewers and the editor, Andre Punt, who made many suggestions which greatly improved the manuscript.

## Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fishres.2016.05.017.

## References

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

Brooks, E.N., Powers, J.E., Cortés, E., 2010. Analytical reference points for age-structured models: application to data-poor fisheries. ICES J. Mar. Sci. 67, 165–175.

Carruthers, T.R., Kell, L.T., Butterworth, D.D., Maunder, M.N., Geromont, H.F., Walters, C., McAllister, M.K., Hillary, R., Levontin, P., Kitakado, T., et al., 2016. Performance review of simple management procedures. ICES J. Mar. Sci. 73 (2), 464–482.

Christoffersen, P., Pelletier, D., 2004. Backtesting value-at-risk: a duration-based approach. J. Financ. Econ. 2 (1), 84–108.

Correa, C.D., Lindstrom, P., 2013. The mutual information diagram for uncertainty visualization. Int. J. Uncertain. Quantif. 3 (3).

Deroba, J.J., Butterworth, D.S., Methot Jr., R.D., De Oliveira, J.A.A., Fernandez, C., Nielsen, A., Cadrin, S.X., Dickey-Collas, M., Legault, C.M., Ianelli, J., Valero, J.L., Needle, C.L., O'Malley, J.M., Chang, Y.-J., Thompson, G.G., Canales, C., Swain, D.P., Miller, D.C.M., Hintzen, N.T., Bertignac, M., Ibaibarriaga, L., Silva, A., Murta, A., Kell, L.T., de Moor, C.L., Parma, A.M., Dichmont, C.M., Restrepo, V.R., Ye, Y., Jardim, E., Spencer, P.D., Hanselman, D.H., Blaylock, J., Mood, M., Hulson, P.-J.F.,

2015. Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. ICES J. Mar. Sci. 72 (1), 19–30.
Dickey-Collas, M., Hintzen, N., Nash, R., Schön, P., Payne, M., 2015. Quirky patterns in time-series of estimates of recruitment could be artifacts. ICES J. Mar. Sci. 72 (1), 111–116.
Dickey-Collas, M., Nash, R., Brunel, T., Van Damme, C., Marshall, C., Payne, M., Corten, A., Geffen, A., Peck, M., Hatfield, E., et al., 2010. Lessons learned from stock collapse and recovery of north sea herring: a review. ICES J. Mar. Sci. 67 (9), 1875–1886.
Doswell III, C.A., Davies-Jones, R., Keller, D.L., 1990. On summary measures of skill in rare event forecasting based on contingency tables. Weather Forecast. 5 (4), 576–585.
Erisman, B.E., Allen, L.G., Claisse, J.T., Pondella, D.J., Miller, E.F., Murray, J.H., Walters, C., 2011. The illusion of plenty: hyperstability masks collapses in two recreational fisheries that target fish spawning aggregations. Can. J. Fish. Aquat. Sci. 68 (10), 1705–1716.
Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to south pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. 55 (9), 2105–2116.
Francis, R.C., Hilborn, R., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68 (6), 1124–1138.
Freyer, L., 2014. Robust rankings. Scientometrics 100 (2), 391–406.
Fricker, T.E., Ferro, C.A., Stephenson, D.B., 2013. Three recommendations for evaluating climate predictions. Meteorol. Appl. 20 (2), 246–255.
Fromentin, J.-M., Bonhommeau, S., Arrizabalaga, H., Kell, L.L., 2014. The spectre of uncertainty in management of exploited fish stocks: the illustrative case of Atlantic bluefin tuna. Mar. Policy 47, 8–14.
Griggs, D.J., Noguer, M., 2002. Climate change 2001: the scientific basis. contribution of working group I to the third assessment report of the intergovernmental panel on climate change. Weather 57 (8), 267–269.
Heath, M., Werner, F., Chai, F., Megrey, B., Monfray, P., et al., 2004. Challenges of modeling ocean basin ecosystems. Science 304 (5676), 1463–1466.
Hilborn, R., 2016. Correlation and causation in fisheries and watershed management. Fisheries 41 (1), 18–25.
Hillary, R.M., Preece, A.L., Davies, C.R., Kurota, H., Sakai, O., Itoh, T., Parma, A.M., Butterworth, D.S., Ianelli, J., Branch, T.A., 2016. A scientific alternative to moratoria for rebuilding depleted international tuna stocks. Fish Fish. 17 (2), 469–482.
Hjorth, J.U., 1993. Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap. CRC Press.
Hodges, J.S., Dewar, J.A., Center, A., 1992. Is it You or Your Model Talking?: A Framework for Model Validation. Rand, Santa Monica, CA.
Hollander, Y., 2015. How Transparent is the Decision on Transport Investment in London? http://www.ctthink.com/publications.html (accessed 25.12.15).
Huijnen, V., Flemming, J., Kaiser, J., Inness, A., Leit ao, J., Heil, A., Eskes, H., Schultz, M., Benedetti, A., Hadji-Lazaro, J., et al., 2012. Hindcast experiments of tropospheric composition during the summer 2010 fires over western Russia. Atmos. Chem. Phys. 12 (9), 4341–4364.
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., 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. 72 (1), 99–110.
ICCAT, 2015. Report of the 2014 Atlantic bluefin tuna stock assessment session. ICCAT Collect. Vol. Sci. Pap. 71 (2), 692–945.
Jiao, Y., Smith, E.P., O'Reilly, R., Orth, D.J., 2012. Modelling non-stationary natural mortality in catch-at-age models. ICES J. Mar. Sci. 69 (1), 105–118.
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., Mosqueira, I., Grosjean, P., Fromentin, J., Garcia, D., Hillary, R., Jardim, E., Mardle, S., Pastoors, M., Poos, J., et al., 2007. FLR: an open-source framework for the evaluation and development of management strategies. ICES J. Mar. Sci. 64 (4), 640–646.
Le Cren, E., Holdgate, M.W., 1962. The Exploitation of Natural Animal Populations. Blackwell Scientific Publications.
Leach, A., Levontin, P., Holt, J., Kell, L., Mumford, J., 2014. Identification and prioritization of uncertainties for management of eastern Atlantic bluefin tuna (*Thunnus thynnus*). Mar. Policy 48, 84–92.
Magnusson, A., Punt, A.E., Hilborn, R., 2012. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish Fish. 14 (3), 325–342.
Maunder, M.N., Sibert, J.R., Fonteneau, A., Hampton, J., Kleiber, P., Harley, S.J., 2006. Interpreting catch per unit effort data to assess the status of individual stocks and communities. ICES J. Mar. Sci. 63 (8), 1373–1385.
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.
Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Weather Rev. 115 (7), 1330–1338.
Pastoors, M.A., Poos, J.J., Kraak, S.B., Machiels, M.A., 2007. Validating management simulation models and implications for communicating results to stakeholders. ICES J. Mar. Sci. 64 (4), 818–824.
Patterson, K., Cook, R., Darby, C., Gavaris, S., Kell, L., Lewy, P., Mesnil, B., Punt, A., Restrepo, V., Skagen, D.W., et al., 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish Fish. 2 (2), 125–157.
Payne, M.R., Clausen, L.W., Mosegaard, H., 2009. Finding the signal in the noise: objective data-selection criteria improve the assessment of western Baltic spring-spawning herring. ICES J. Mar. Sci. 66, 1673–1680.
Pepin, P., Marshall, C.T., 2015. Reconsidering the impossible linking environmental drivers to growth, mortality, and recruitment of fish. Can. J. Fisher. Aquat. Sci. 72 (999), 1–11.
Peterman, R.M., 2004. Possible solutions to some challenges facing fisheries scientists and managers. ICES J. Mar. Sci. 61 (8), 1331–1343.
Plowright, R.K., Sokolow, S.H., Gorman, M.E., Daszak, P., Foley, J.E., 2008. Causal inference in disease ecology: investigating ecological drivers of disease emergence. Front. Ecol. Environ. 6 (8), 420–429.
Pope, J., 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. ICNAF Res. Bull. 9 (10), 65–74.
Punt, A., 2008. Refocusing stock assessment in support of policy evaluation. Fish. Glob. Welf. Environ., 139–152.
Ralston, S., Punt, A.E., Hamel, O.S., DeVore, J.D., Conser, R.J., 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. Fish. Bull. 109 (2), 217–232.
Roebber, P.J., 2009. Visualizing multiple measures of forecast quality. Weather Forecast. 24 (2), 601–608.
Schaefer, J.T., 1990. The critical success index as an indicator of warning skill. Weather Forecast. 5 (4), 570–575.
Schnute, J.T., Hilborn, R., 1993. Analysis of contradictory data sources in fish stock assessment. Can. J. Fish. Aquat. Sci. 50 (9), 1916–1923.
Sissenwine, M., Shepherd, J., 1987. An alternative perspective on recruitment over fishing and biological reference points. Can. J. Fish. Aquat. Sci. 44 (4), 913–918.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmos. 106 (D7), 7183–7192.
Walters, C., Punt, A., 1994. Placing odds on sustainable catch using virtual population analysis and survey data. Can. J. Fish. Aquat. Sci. 51 (4), 946–958.
Waples, R.S., Gaggiotti, O., 2006. Invited review: what is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15 (6), 1419–1439.
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.