Received: 11 January 2021 | Revised: 15 March 2021 | Accepted: 30 March 2021

DOI: 10.1111/faf.12562

ORIGINAL ARTICLE

FISH and FISHERIES WILEY

Forecasting fish recruitment in age-structured population models

Elisabeth Van Beveren | Hugues P. Benoît | Daniel E. Duplisea

Fisheries and Oceans Canada, Institut Maurice-Lamontagne, Mont-Joli, QC, Canada

Correspondence
Elisabeth Van Beveren, Fisheries and Oceans Canada, Institut Maurice-Lamontagne, 850 Route de la Mer, Mont-Joli, G5H 3Z4 QC, Canada.
Email: elisabeth.vanbeveren@dfo-mpo.gc.ca

Abstract

Recruitment in age-structured stock assessment models can be forecasted using a variety of algorithms to provide advice on the anticipated consequences of different possible management actions. Selecting one method over another usually involves some subjectivity, yet can be consequential to the provision of advice. Extensive case-specific testing is not always feasible. We evaluated the forecast skill in 3-, 5- and 10-year forecasts of 16 recruitment forecasting methods under various circumstances to provide a broad evaluation and general guidelines on the reliability of forecasts. We used 31 operating models based on existing stock assessment models applied to a diversity of stocks with empirical data, which we show to be generally representative of assessed stocks worldwide. Although no single best-performing method could be identified, we found that time-series methods were most likely to perform poorly. Both forecast skill across all methods and forecast sensitivity to the selected method were linked to the properties of the stock or assessment: age at maturity and recruitment autocorrelation in 3-year forecasts and previous long-term recruitment variability in 10-year forecasts. In some situations, all forecasting methods resulted in systematic over- or underestimation of spawning stock biomass. The simulation approach employed here to assess forecast performance, rooted directly in the predictions of existing stock assessment models, can be a complementary tool to existing simulation approaches which generate alternative sets of population dynamics or observations and we discussed the advantages and limitations.

KEYWORDS

fish stock, population forecast, recruitment dynamics, stock assessment, stock–recruitment, uncertainty

1 | INTRODUCTION

Fisheries advice often includes forecasts of how a stock will respond to different future harvesting scenarios. This is true in tactical stock assessment that aims to advise on the risks associated with different catch options in the short term and more strategic advice such as that produced by management strategy evaluation, which advises of the risks associated with different management procedures. Likewise, longer-term forecasts are the cornerstone of recovery planning for depleted or threatened species. Forecasting reasonable, plausible outcomes on which managers can rely is not trivial, as it depends on an incomplete, uncertain and sometimes incorrect understanding of past population dynamics and current population state as well as assumptions on the degree to which historical dynamics will persist into the future. Forecasts must balance the admission of uncertainty with the tangible forecasting of future states, as unnecessarily vague forecasts

are impractical and incorrect ones are potentially harmful to stocks if they lead to overly optimistic catch advice or to resource users otherwise. Despite the critical importance of forecasts to fisheries and population management advice and their sensitivity to the underlying assumptions (Punt et al., 2016), considerably less research has been devoted to them compared to research on the construction and fitting of models to describe past and current population states (Kell et al., 2016).

Since at least the seminal work of Hjort (1914), fisheries science has invested a great deal of time and energy into producing forecasts of recruitment. Of all the demographic rates affecting the dynamics of fish stocks, including age- or size-dependent rates of growth, mortality and maturation, recruitment is typically the most temporally variable and one of the most influential (e.g. Rothschild, 1986). Although there must be some dependence on spawning biomass, the relationship is often imperceptible or explains little of the recruitment variance (Cury et al., 2014; Iles, 1994; Szuwalski et al., 2019). Well-defined stock–recruitment relationships can be spurious or change over time (Gilbert, 1997; Kurota et al., 2020; Szuwalski et al., 2019) and are dependent on the data and model structure that were used to estimate them (Brooks & Deroba, 2015; Walters & Ludwig, 1981). Numerous studies have therefore aimed to improve forecasts through the inclusion of external drivers (e.g. Haltuch et al., 2019). Although promising for some stocks, for many others, this approach appears to be of limited value, as variables with enough explanatory power need to be available (De Oliveira & Butterworth, 2005) and they often need to be accurately projectable over a sufficiently long timeframe (Basson, 1999; Planque et al., 2003; Walters & Collie, 1988). One also needs to be confident in current and future effect size and stability (Zwolinski & Demer, 2019). Even in favourable circumstances, forecasting recruitment using external drivers is statistically challenging because of the ubiquity of non-linear and state-dependent drivers (Chen & Irvine, 2001; Sugihara et al., 2012).

Despite the many challenges in forecasting recruitment, stock assessment requires practical approaches to do so. These can generally be classified as parametric, semiparametric and non-parametric (Subbey et al., 2014). Parametric methods model recruitment as a functional expression of one or more variables, including stock biomass (stock–recruitment models), the time-lagged recruitment series itself (time-series models) and environmental indices, in a classical or state-space framework. Non-parametric methods do not require specification of a direct relationship or distribution and have thus fewer assumptions (e.g. sampling algorithms; Kimoto et al., 2007; Paz & Larrañeta, 1992). Semiparametric methods include a parametric component (e.g. a functional relationship) as well as a non-parametric element (e.g. random deviations from the parametric component). Despite the multiplicity of forecasting approaches (e.g. Brodziak, 2018) and the availability of reviews providing general guidelines (Maunder & Thorson, 2019; Needle, 2001; Sharma et al., 2019; Subbey et al., 2014), we are not aware of a broad-scale systematic evaluation of the reliability of these different forecasting methods (for examples of more restricted comparisons, see Evans & Rice, 1988; Kimoto et al., 2007; Paz & Larrañeta, 1992). Time constraints, a limitation of options available in stock assessment software or a lack of awareness thereof can predispose a tendency towards institutional preference, despite evidence that realistic alternatives might result in meaningfully different outcomes, even for short-term forecasts (e.g. MacKenzie et al., 2008). This is particularly concerning because the reliability and robustness of a chosen forecasting method is often not extensively tested either by cross-validation or by simulation.

To begin addressing these issues, we applied 16 commonly employed or recent recruitment forecasting methods on an empirical-based testing set of population dynamics for 31 finfish stocks representing a range of stock characteristics that exist worldwide. We based our study on model output for various stocks because this ensured that our findings represented the properties of a diversity of situations. We began by evaluating and comparing the forecast error and bias in spawning stock biomass (SSB) at the end of different forecast periods for each recruitment forecasting method and stock. We then undertook analyses to identify properties of the different stocks or assessments that are associated with differing levels of forecast skill of the various approaches. The intention was to guide stock assessors in their choice of recruitment forecasting method by providing general measures of forecast skill that can be used to qualify the reliability of forecasts when extensive simulations are not feasible.

2 | MATERIALS AND METHODS

We developed a four-step simulation procedure to evaluate the forecast error and bias of various recruitment forecasting methods (Figure 1).

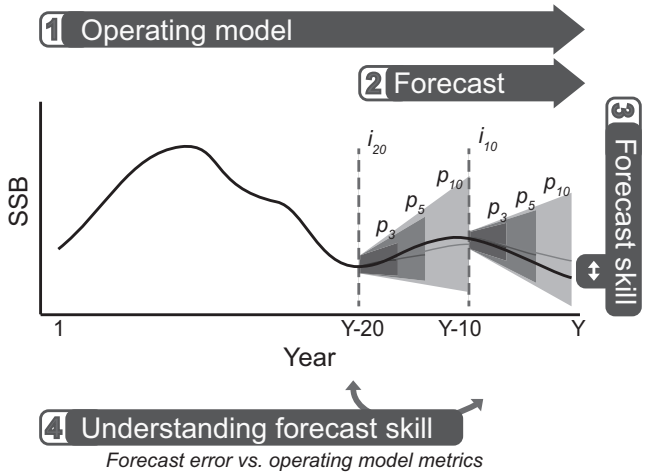

FIGURE 1 Schematic representation of the analyses (SSB = spawning stock biomass, *Y* = terminal year, *p* = forecast period, *i* = initiation time). The black line represents the "true" population state (step 1: operating model), and forecasts using a given method (step 2) are shaded. Forecast skill is measured by comparing forecasted values with a "baseline" or true model and is then linked to metrics of the operating model (step 4)

1. We gathered a diverse set of population dynamic models for different and contrasting fish stocks, each of which was presumed to reflect the "true world" (31 operating models; OMs).
2. Forecasts under various recruitment assumptions but assuming perfect knowledge of other processes (e.g. growth and maturation) were then initiated (initiation time; *i*) at 10 and 20 years prior to the terminal year *Y* for each OM (i_{10} = *Y*-10; i_{20} = *Y*-20). These initiation times were chosen to provide the potential to discriminate forecast skill at contrasting demographic or productivity states within OMs (e.g. low and high abundance periods), while ensuring some independence between the two sets of predictions and that there was sufficient information on which to base the forecasts. Three sets of forecasts were made spanning 3, 5 or 10 years, respectively (forecast period; *p*).
3. The skill of each forecast method was assessed through a comparison of forecasted SSB and values obtained if true recruitment during forecasting was known. The sensitivity of forecast skill with respect to the forecast method and OMs was analysed by contrasting the results obtained by applying the various methods.
4. The skill of each method as well as the dissimilarity in skill over all methods was then linked to potential causes related to the characteristics of the OM.

2.1 | Operating models

To evaluate forecasting approaches over a range of situations, a pertinent set of OMs needs to be established as a test bed. This test bed requires realism and has to be large enough to allow general conclusions yet small enough to be manageable in terms of computation time. The proposed simulation approach is therefore grounded in the output of empirical assessment models, which we use as OMs for the simulations. Specifically, maximum likelihood estimates were considered as the "true" population. This is a complementary alternative to simulating pseudodata or dynamics based on various fits, which is commonly done in other simulation studies (e.g. Anderson et al., 2014), but which is challenging to do for a wide range of dynamics and assessment situations. The proposed approach reduces the flexibility of the simulation framework but results in a set of OMs of workable size (each OM involving 102 forecasts) that approaches most closely real stock and assessment scenarios. Furthermore, by conducting our simulations within the respective stock assessment frameworks, we test recruitment projection methods under the conditions in which they are employed to provide scientific advice (see Discussion).

The integrated age-based state-space stock assessment model (SAM; Berg & Nielsen, 2016; Nielsen & Berg, 2014), as implemented using the R stockassessment package (Nielsen et al., 2019), was used to perform the analyses. This model was chosen because of its popularity and flexibility and because it allowed consistency between the model and forecasts, and among simulated cases. Model inputs and outputs (input data matrices, parameter estimates, etc.) were obtained from two sources (Annex S1); 25 were downloaded from stockassessment.org, which is a repository and online application for stock assessments undertaken using SAM, and an additional 6 were generated by fitting SAM to data available for Canadian stock assessments with which the authors had some familiarity (research documents published by the Canadian Science Advisory Secretariat). Not all fits were peer-reviewed. This is rational for our intent, to generate a test bed suitable for reaching general conclusions on the performance of recruitment forecast methods. As in many simulation studies, individual OMs will also not be discussed because doing so would be inappropriate.

Parameter estimates, related to the overall dynamics or the stock and fishery state (process variance and fishing mortality and numbers at age), were taken as the presumed assessment outputs at either i_{10} or i_{20} and served as a basis of projection. That is, the typical model output to start forecasts was extracted, albeit for pre-defined initiation years rather than the commonly used terminal year. Forecasts proceeded using the parameter estimates and their covariance matrix for the initiation year, without the need to refit the model at each iteration.

2.2 | Forecasts

2.2.1 | Forecasting approach

Forecasting future stock state in age-based assessments requires making assumptions concerning all age-dependent processes (e.g. natural and fishing mortality, maturation and growth), as well as forecasting annual recruitment. Forecasts were based on the forecast function within the stockassessment package, supplemented with

recruitment forecasting options. In short, 1,000 simulations (validated to provide stable results) of abundance and fishing mortality at age (i.e. the states) were initiated from the initiation year so that the uncertainty envelope of that year was carried forward. A given stochastic recruitment forecasting algorithm and the traditional cohort equations were then used to forecast the subsequent years. Process noise included in the annual transition of abundance and fishing mortality at age and which is proper to SAM's state-space framework followed a multivariate normal distribution where the specified mean (μ) and variance (Σ) were based on model estimates. Fishing mortality was solved given catch, which was known without error in the forecasts. Values of natural mortality, proportion mature and weight at age were also assumed known without error. Error in the forecasts was therefore primarily induced by the recruitment forecast method (Table 1) and sources of additional stochasticity (i.e. initial state uncertainty and process noise). While we acknowledge that there are other important factors that can contribute to forecast errors (e.g. growth), it would not have been tractable in one paper to parse of the contribution of multiple factors.

2.2.2 | Recruitment forecasting methods

The 16 recruitment forecasting methods fall into the following four general classes: (a) sampling methods, (b) empirical dynamic modelling, (c) time-series analysis and (d) classical methods.

Sampling methods

Drawing from past values estimated in the assessment is a flexible non-parametric approach to forecast recruitment. Sampling from the past relies on the assumption that these values are good predictors of future recruitment. Because this assumption is less likely to hold for longer-term forecasts, sampling methods are generally intended for short- to medium-term forecasts only. It is especially important to determine an appropriate recruitment reference period or a pool of realistic recruitment values from which to sample. When sampling is independent of SSB, this reference period typically comprises only more recent years to avoid forecasting historical recruitments that are unlikely to reoccur in the near future. Without SSB feedback, a pool of recruitment values was defined in relation to the length of the forecast period ($R_{i\text{-}p}$, …, R_i, where p equals 3, 5 or 10 years). This choice was made to accommodate the evaluation of sampling methods that require larger pools (see assumptions below) and to provide a consistent comparison among all. On the other hand, when sampling included SSB dependence, sufficient information on this dependency is necessary and hence the full historical period was used as a reference period.

Five sampling-based forecasting methods independent of SSB were evaluated.

Random sampling involved equal probability sampling with replacement of values from a fixed set of recruitment values estimated for the past (e.g. Nielsen et al., 2019).

TABLE 1 Recruitment (R) forecasting methods by class, with indication of spawning stock biomass dependence (SSB dep.)

Class	Name	Method	SSB dep.
Sampling	random	random sampling	No
Sampling	tapered	time tapered sampling	No
Sampling	window	expanding window sampling	No
Sampling	block	block sampling	No
Sampling	ecdfR	ecdf sampling of R	No
Sampling	ecdfS	ecdf sampling of R/SSB	Yes
Sampling	ecdfR3	ecdf sampling of R, conditional on SSB	Yes
Sampling	ecdfS3	ecdf sampling of R/SSB, conditional on SSB	Yes
Sampling	markov	markov matrix sampling	Yes
EDM	simplex	simplex (empirical dynamic modelling, EDM)	No
Time series	rw	random walk	No
Time series	arima	autoregressive integrated moving average	No
Time series	ets	exponential smoothing algorithm	No
Classical	bh	Beverton–Holt	Yes
Classical	ri	Ricker	Yes
Classical	mean	mean past value	No

Time tapered sampling involved sampling with selection probabilities that vary inversely with time span between the year of their occurrence and the forecast initiation year (Figure S1).

Expanding window sampling involved equal probability sampling with replacement from a recruitment pool that expands over time (Figure S1, e.g. DFO, 2010, DFO, 2011). It can be perceived as a special case of time tapered sampling, in which selection probabilities are either zero or not, with the string of zeros shortening over time.

Block sampling involved first selecting a continuous block of values from the pool, from which individual values are then drawn. Here, we employed 3-year blocks from which three successive recruitments were drawn with replacement. Depending on the duration of the forecast and size of the blocks, a single forecast may involve selecting more than one block. This method has been implemented to retain some temporal autocorrelation structure of forecasted recruitment (e.g. Licandeo et al., 2020), which is otherwise eventually lost with the other methods above. Note that for stocks with spasmodic recruitment, for which this method has been used, the recruitment pool is not limited to more recent values.

ecdfR involved randomly selecting the recruitment value from the empirical cumulative distribution function (ecdf) constructed using the recruitment pool and smoothed using linear interpolation (as in Brodziak, 2018). In contrast to the above methods, ecdfR is not limited to the exact values from the past.

All of the above non-parametric methods can easily be adapted to incorporate SSB dependence through the use of the recruitment rate ($S_y = \frac{R_y}{SSB_{y-a_r}}$, where a_r is the age at recruitment) and/or by conditioning on SSB (e.g. Brodziak, 2018; Kimoto et al., 2007). The recruitment rate method, which assumes a linear relationship between recruitment and SSB, consists of sampling S_y values, that are then multiplied by the estimated or forecasted SSB to obtain a new recruitment prediction. Conditioning on SSB was achieved by defining three classes (low, medium and high SSB) for each stock, based on k-means clustering (MacQueen, 1967) applied to the complete historical series. Pools of recruitment or recruitment rates are subsequently sampled as a function of the estimated or forecasted SSB class. We included sampling methods with and without a transition matrix between the recruitment (rate) and SSB classes (see Annex S2).

Four non-parametric forecasting methods for recruitment sampling dependent on SSB were evaluated, all making use of the ecdf, although any of the previously outlined methods could have been employed.

ecdfS involved ecdf sampling of S_y values from the complete historical time period.

ecdfR3 involved ecdf sampling of SSB class-specific recruitment values. That is, by sampling the pool of recruitment values associated with the respective SSB class (see Annex S2 for details).

ecdfS3 involved ecdf sampling of SSB class-specific S_y values. That is, by sampling the pool of S_y values associated with the respective SSB class (see Annex S2 for details).

Markov matrix sampling involved ecdf sampling from a pool of recruitment values determined through a transition matrix, which describes the probabilities that an SSB class is associated with the production of given recruitment pools. These pools can be established in various ways, so that there is flexibility in their numbers and sizes. We defined three recruitment ranges based again on k-means clustering. A pool was selected using a multinomial distribution based on the probability of transitioning from the estimated or forecasted SSB to any of the recruitment pools (see Annex S2 for details).

Empirical dynamic modelling

Simplex forecasting (Sugihara, 1994; Sugihara & May, 1990) is a non-parametric and non-linear forecasting method, which is the foundation of the empirical dynamic modelling framework (e.g. Deyle et al., 2018; Sguotti et al., 2019; Ye et al., 2015). Essentially, it is based on the principle that comparable time fragments will evolve similarly. Therefore, the complete historical recruitment time series of an OM was split or embedded in fragments of length E, and the Euclidean distance between the last time fragment and all previous ones was calculated to find the E + 1 past states most comparable to the current situation (see Deyle et al., 2018). The forecast was then taken as the average of the recruitments ensuing these states, weighted by their distance. A self-test of this method on the historical values with different values of E (max = 5) allowed automatic selection of the optimal value (maximal Pearson correlation coefficient) for forecasts. Because this procedure results in a deterministic estimate, we added process noise with autocorrelation (technically making it semiparametric) and a correction for the mean bias when transforming from the log scale (details below; Beverton–Holt).

Time-series analysis

Time-series analyses can be used to forecast recruitment based on temporal patterns in the past recruitment stream and are typically parametric. Time-series methods are only appropriate for series of sufficient length, and therefore, the complete historical recruitment series was used as a reference period. For simplicity and reproducibility, we worked with standard and correspondingly named functions from the R forecast package (Hyndman et al., 2019; default settings), which can all return stochastic forecasts. We refer readers to the online documentation for further details on these methods (including equations and algorithms for automatic modelling; Hyndman & Athanasopoulos, 2018). Forecasts were always done on a log scale, with bias correction for exponentiation to the linear scale.

Random walk (without drift) involved forecasting recruitment presuming that the past does not hold information to predict the future. We applied this method as a baseline for comparing more complex methods, as ecological time series such as recruitment are generally not as erratic.

ARIMA (autoregressive integrated moving average) models are a form of Box–Jenkins models, of which the random walk is a special case and which can incorporate dependence on prior observations and patterns (e.g. Gröger & Fogarty, 2011; Stoker and Noakes, 1988).

ETS (exponential smoothing state space) models are based on the concept that larger weights should be given to more recent observations (similar to time tapered sampling). These forecasts involve taking past observations and down weighing them exponentially over time (Hyndman & Hyndman, 2008).

Classical methods

Three classic parametric methods were used, each assuming autocorrelated error.

Beverton–Holt involved the forecasting of a recruitment value (R_μ) from a standard two-parameter stock–recruitment relationship $\left(R_\mu = \frac{\alpha SSB_{y-a_r}}{(1+\beta SSB_{y-a_r})}\right)$. Most OMs did not explicitly presume such a relationship, and hence, parameters were deterministically estimated on a log scale using the full historical recruitment and SSB series. A modification was made ($R_y = R_\mu e^{\rho\varepsilon_{y-1} + \delta_y\sqrt{1-\rho^2} - \sigma_R^2/2}$) to incorporate temporal autocorrelation and to correct for bias when exponentiating from the log scale (Butterworth et al., 2003; Johnson et al., 2016). Specifically, the random recruitment deviation ($\delta_y \sim N(0, \sigma_R^2)$) was adjusted by the lag-1 autocorrelation coefficient (ρ), both of which were separately estimated outside the model based on the complete historical time

TABLE 2 Stock metrics used in the interstock comparison of forecast skill

Metric	Definition
PV	Proportional variability of recruitment
AC	Lag-1 autocorrelation of recruitment
D	Distance from geometric mean of forecast year recruitment
N	Series duration
A50	Age at which 50% of fish are mature

period (for ρ, the acf function in R was used; R Core Team, 2020). Parameter estimates (α, β) could only be obtained for 30 of the 62 cases based on the observations available at the forecast times, and thus, the method could only be tested on that subset.

Ricker involved the forecasting of a recruitment value based on the Ricker stock–recruitment relationship ($R_\mu = \alpha \mathrm{SSB}_{y-a_r} e^{-\beta \mathrm{SSB}_{y-a_r}}$). Values of R_y were calculated from R_μ in the same manner as for the Beverton–Holt method. Sensible parameters for the relationship could only be obtained for 32 of the 62 cases.

Mean involved taking the average of past recruitment values ($R_\mu = \Sigma(R_{i-p}, \ldots, R_i)/(p+1)$). Because there is no link to SSB, the reference period was set to the recent past ($R_{i-p}, \ldots, R_i$). We added process noise ($\delta_y \sim N(0, \sigma_R^2)$) estimated based on the reference period, autocorrelation (ρ) estimated based on the full past and a correction for the mean bias when transforming from the log scale (details above).

2.3 | Forecast skill and sensitivity

Different stock assessment population forecasting frameworks account for different distinct sources of stochasticity (initial state uncertainty, process noise, uncertainty associated with model parameters, etc.) which might influence overall forecast skill and sensitivity. To ensure that our results are transferable to other modelling frameworks, including non-state-space models, we ran a baseline known-recruitment scenario in parallel to the various recruitment forecasting scenarios for each OM, allowing only forecasted recruitment to differ. This was achieved by using random number seeds for each stochastic parameter unrelated to recruitment that were common across simulated scenarios.

2.3.1 | Forecast skill

The skill of each forecast was measured in terms of forecast error and bias. We focus on predictability of SSB rather than recruitment itself because this variable is generally of higher interest when undertaking forecasts as part of stock assessments and other types of population assessments.

Forecast error was evaluated using the absolute percentage error between the baseline and forecasted median SSB at the final year of the forecast $\left(\text{error} = 100\,\% \frac{\mathrm{SSB}_{i+p} - \mathrm{SSB}_{i+p}^{\text{baseline}}}{\mathrm{SSB}_{i+p}^{\text{baseline}}}\right)$. This metric was chosen among the variety of scale-invariant measures of error (see Hyndman & Koehler, 2006) because of its intuitiveness and popularity. Furthermore, it yielded similar results as other methods evaluated in preliminary analyses.

Bias was determined as the final-year SSB percentage error $\left(\text{bias} = 100\,\% \frac{\mathrm{SSB}_{i+p} - \mathrm{SSB}_{i+p}^{\text{baseline}}}{\mathrm{SSB}_{i+p}^{\text{baseline}}}\right)$.

2.3.2 | Forecast sensitivity

The analyses aim to inform on the forecast skill of different recruitment forecasting methods under various circumstances. In practical situations, the method(s) with optimal performance will however remain unidentifiable with absolute certitude. When there is subjectivity in a potentially impactful choice, this almost inevitably leads to questions concerning the consequences (e.g. "What level of sensitivity can be expected?" and "What is the effect of selecting one particular method over another?") and subsequently what should be done about it (e.g. "What alternatives are consideration worthy?"). To help tackle these concerns, we determined the overall sensitivity to the choice of a recruitment forecast method and included a comparison of SSBs forecasted by the various methods. Overall sensitivity was quantified as the interquartile range (IQR) of biases across algorithms for each OM. The IQR was preferred over other common measures of spread because it is always positive, useful for skewed distributions and not sensitive to a few extremes.

2.4 | Understanding forecast skill and sensitivity

For each OM, we identified five metrics that summarize patterns in the recruitment time series or a relevant aspect of stock biology and assessment that might help explain variation in forecast skill among OMs and forecast periods (Table 2). Each can be calculated before starting a forecast, thereby allowing readers to use our findings in selecting recruitment forecasting methods for a particular application. Note that the selected metrics (Table 2) reflect our perception of stock dynamics and biology but are nonetheless usually dependent on the quality of data and assumptions entering the assessment model.

We identified three metrics calculated from the recruitment series for the second to the forecast year ($R_2, \ldots R_i$).

Relative global recruitment variability was calculated as the **proportional variability (*PV*)** of each recruitment vector:

$$\mathrm{PV} = \frac{2\sum_{R=1}^{n(n-1)/2}\left(\frac{|R_k - R_l|}{max\,(R_k,R_l)}\right)}{n(n-1)},$$

where R_k and R_l are values within the recruitment vector ($k \neq l$) and n is the vector length. The PV measures the average per cent difference between all possible pairwise combinations of the recruitment values in the time series and is unbiased in the presence of rare events and non-Gaussian dynamics (see Heath, 2006; Heath & Borowski, 2013). It is therefore more appropriate for measuring and comparing variability between recruitment time series than more common approaches such as the coefficient of variation or standard deviation (e.g. Fogarty, 2001; Myers & Pepin, 1994).

Lag-1 autocorrelation (*AC*) was calculated as a metric of serial correlation between recruitment values in the time series.

The **relative distance from the geometric mean (D)** measures the extent to which recruitment in the year the forecast was initiated differs from the geometric mean recruitment value ($\overline{R}$), $D = |\overline{R} - R_i|/\overline{R}$.

The metric was included as more extreme recruitment states at the initiation time might evolve towards values outside or at the margins of the historical scope and therefore be harder to forecast.

Two additional metrics were identified:

Series duration (N), or the number of years on which the forecasting was based, was included as it is reasonable to expect that longer series allow forecasts to be better "trained." Series duration could hence improve forecast skill for methods that are conditioned on the full historical period (e.g. time-series analyses and SSB-dependent methods).

Age at fifty per cent maturity (A50) is associated with longevity and with the intrinsic rate of population increase (Denney et al., 2002; Hutchings et al., 2012). It is therefore associated with population turn over and variability, and associations with forecasting skill might be expected. A50 was estimated at the initiation year using a logistic regression of the proportion mature fish as a function of their age, and it was preferred over other correlated characteristics that essentially reflect life history (e.g. the number of modelled age classes or age at recruitment).

Multiple regression was used to relate the log of the forecast error to the above normalized metrics for each recruitment forecasting method and period. All possible variable combinations were fitted, excluding interactions, resulting in 51 regression models (Calcagno, 2020). Model averaging based on Akaike information criterion, corrected for small sample sizes (AICc), was used to quantify the relative importance of each variable using the sum of the relative evidence weights of the averaged models.

The above metrics might also explain the dissimilarity in forecasted SSB between different recruitment forecasting methods for a given OM and forecast period. Put another way; are there characteristics of the OM that cause different recruitment forecasting methods to forecast similar (or dissimilar) values for recruitment? To explore this, we used the log of the interquartile range of the bias of the various recruitment forecasting methods for a given OM as a measure of dissimilarity in forecasts. Among-OM dissimilarities were related to the above metrics using regression. An averaged multiple regression was fitted by forecast period, using the same approach as presented above (resulting in three regression models with $n = 62$).

To validate the generality of the results we obtained from the 31 OMs used in our study, we compared the distribution of the metrics above with distributions estimated from a much broader suite of stocks using recruitment estimates from the RAM Legacy Stock Assessment Database (2020). Based on similar ranges and distributions for these metrics for our chosen OMs and the broader sample available in the RAM database, we conclude that the results obtained here should be broadly applicable (Figure 2).

3 | RESULTS

3.1 | Forecast skill

When recruitment was forecasted over 3 years, the error of the final-year SSB was slightly higher for time series-based methods compared to the other methods (median: 5%–9%, 95th percentile: 34%–60%; Figure 3; see Figures S2 and S3 for results under different initiation times). The discrepancy in forecast error between these two sets of forecasting methods increased as the forecast period was increased. In 10-year forecasts, the time series-based methods had a potential to generate large errors in forecasted SSB (error > 100%), with the random walk performing poorest most often across OMs. The median forecast error across OMs remained similar for all other methods (36%–70%) in 10-year forecasts. The two methods based on a stock–recruitment relationship (Ricker and Beverton–Holt) produced median errors in 10-year forecasts that were not noticeably lower (71%–79%) than those for most other methods (47%–84%, excluding RW, ETS and ARIMA) when comparing for OMs for which all methods could be applied. Overall, the simplex method consistently provided forecasts with among the lowest errors across OMs and forecast periods.

Most methods were about equally likely (33%-66% chance) to produce positively or negatively biased 3-year forecasts (Figure 4). Sampling-based methods using recent recruitment had a minor but consistent tendency towards overestimation as the duration of the forecast period increased. This was also the case for the time series-based methods. The stock–recruitment as well as the simplex methods, in contrast, had a propensity to underestimate SSB.

An important driver of 3-year forecast error was the autocorrelation in the recruitment time series (AC), although the result was statistically significant only for methods using recent recruitment and the simplex forecasts (Figure 5). Thus, in these cases greater AC was associated with lower forecast error. In addition, for a number of forecasting methods, there was an indication that OMs for early maturing stocks (low A50) were associated with more error prone forecasting, although the effect was smaller than that for AC and only occasionally

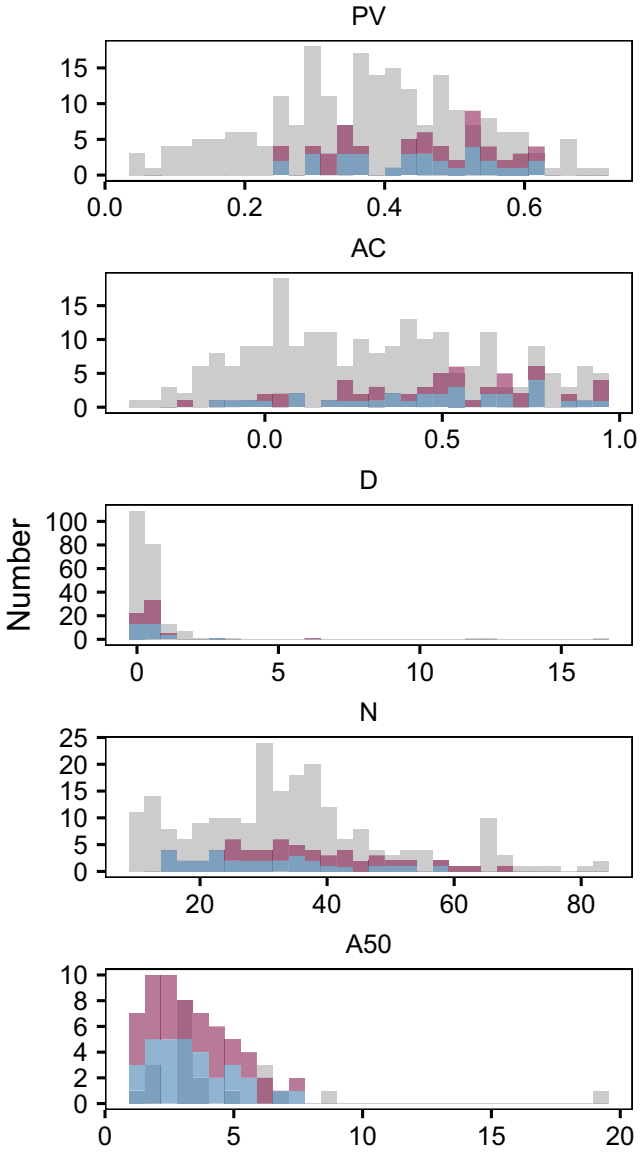

FIGURE 2 Distribution of the metrics for all operating models considered at both initiation times (blue = Y-20 and purple = Y-10). Metrics were compared to those calculated on model output (from virtual population analyses and statistical catch-at-age models) available through the RAM Legacy Stock Assessment Database (grey histograms; PV = proportional variability of recruitment, AC = lag-1 autocorrelation of recruitment, D= distance from the geometric mean of initiation-year recruitment, N = series duration, A50 = age at which 50% of fish are mature). Stocks with globally low recruitment variability (PV) were not covered in this study (typical for sharks and seashells), but these are also stocks for which reliably forecasting recruitment is easier (figure appears in colour in the online version only)

significant. As the forecast period was extended to five and then 10 years, the estimated effect sizes for AC and A50 decreased, and in 10-year forecasts, neither was associated with relative error. In contrast, as the forecast period increased, the importance of interannual recruitment variability (PV) increased. For most forecasting methods, high PV was associated with larger forecast error in 10-year forecasts.

3.2 | Forecast sensitivity

Three-year forecasts were generally not sensitive (difference in bias <5%) to the choice of a forecasting method from within a class of methods (Figure 3). However, between classes there were noticeable divergences. For instance, although the differences in bias between basic sampling and the application of a Beverton–Holt curve in 3-year forecasts were mostly (75% of forecasts) below 14%, it could reach 100%. The largest impact was produced by selecting a time-series method (random walk, ARIMA, etc.) over a stock–recruitment method (Beverton–Holt, Ricker; Figure 4), the former providing generally more optimistic estimations. These patterns of sensitivity to the selection of a forecasting method expectedly amplified as the forecast period increased. For 10-year forecasts, the difference in error between contrasting methods was frequently (>25% of forecasts) above 90%. This was, for example, the case for SSB-dependent sampling methods that aim to simulate a stock–recruitment relationship, which commonly forecasted considerably higher biomasses than if a parametric curve was fitted.

Biases in 3-year forecasts varied the least among forecasting methods for OMs for late maturing stocks and to a lesser extent for OMs associated with high recruitment autocorrelation (Figure 6). The former result is a direct consequence of the prolonged time required for forecasted recruits to reach the SSB in late maturing stocks, resulting in very similar forecasts among methods. This benefit rapidly diminished as the forecast period increased. Autocorrelation in the recruitment series contributed to overall sensitivity to the forecast method for 5-year forecasts in particular, but essentially not at all for 10-year forecasts. Meanwhile, high variability in recruitment time series contributed increasingly to forecast sensitivity as the forecast period lengthened.

4 | DISCUSSION

Despite recognition that the choice of a recruitment forecasting method can be consequential (Punt et al., 2016; Subbey et al., 2014), a review of available methods and their overall skill had to date been surprisingly lacking. We thus first confirm that the choice of a recruitment forecasting method can be important, even for short-term forecasts and increasingly so as the forecast interval increases. We then showed that there is no single best recruitment forecasting method, although some methods like time-series methods are noticeably more likely to perform poorly as the forecast interval lengthens.

Forecasting recruitment as a function of SSB has intuitive appeal. Including a feedback system between the population biomass and recruitment is rational both from a biological and from a managerial point of view (e.g. because of compensatory dynamics). The common parametric forms such as the Ricker and Beverton–Holt have an intensively studied mechanistic foundation, and much of the theory of fishing is routed in these relationships. By consequence, the use of such relationships is widespread and the literature on it

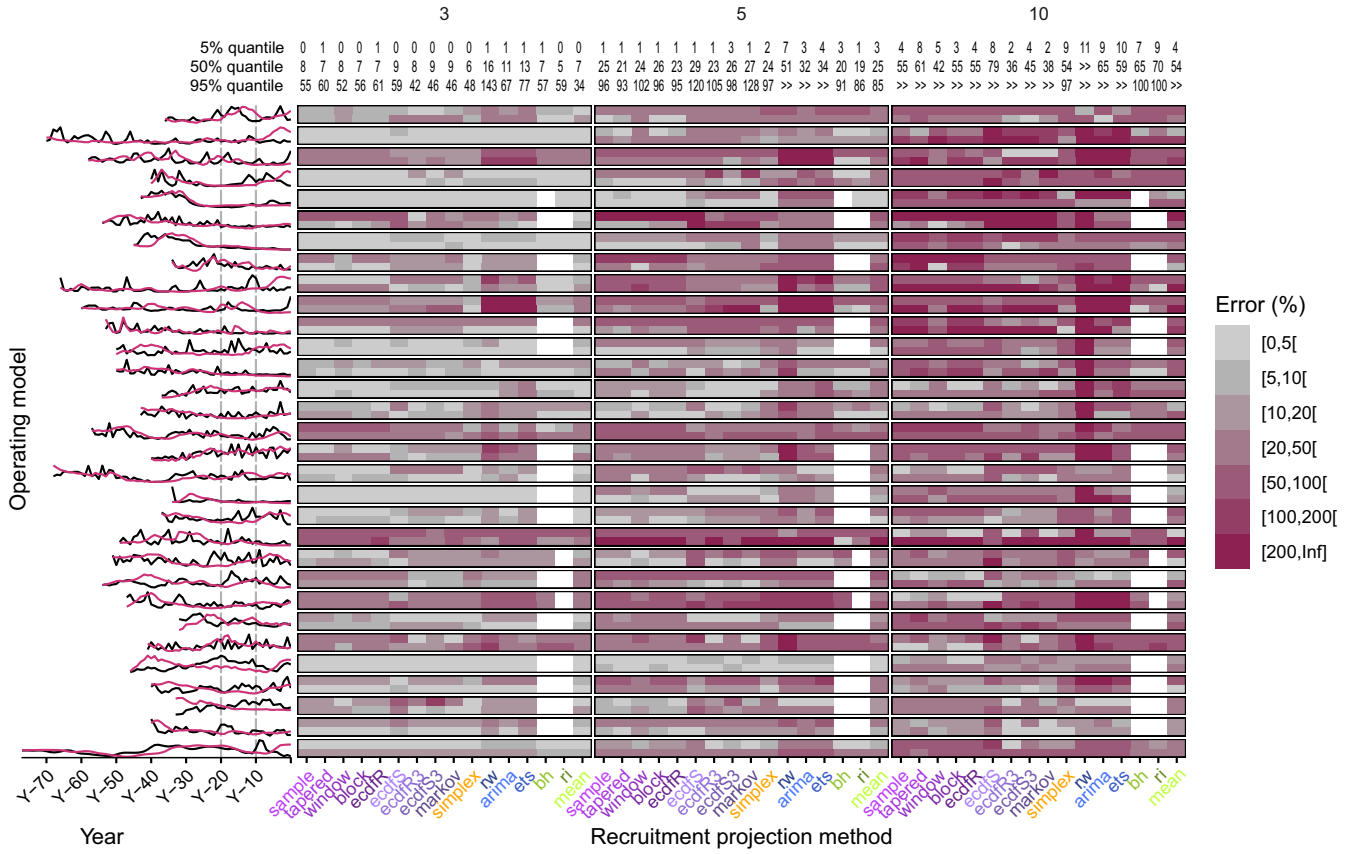

FIGURE 3 Error (%) in forecast spawning stock biomass (SSB) at the end of the forecast period (3, 5 and 10 years; coloured blocks) for every operating model (OM; *y*-axis) at both initiation times (upper row = Y-20, lower row = Y-10 for each OM) for the different recruitment forecasting methods (*x*-axis). The time series (normalized) for recruitment (black) and SSB (red) are shown for each OM on the left-hand side, with initiation times indicated using vertical grey lines. The table provides the 5%, 50% and 95% quantiles of forecast error over the ensemble of OMs and initiation times (values above 200% are indicated as ≫). The labels for the recruitment forecast methods are coloured according to their class (Table 1). Stock–recruitment functions could not always be fitted, preventing testing of these forecasting methods in some cases (white boxes) (figure appears in colour in the online version only)

is vast (Maunder & Thorson, 2019; Sharma et al., 2019). However, the use of these methods to model and predict recruitment appears feasible in a restricted number of applications (e.g. Iles, 1994), as we were only able to meaningfully fit parametric relationships in about 50% of simulated cases (similar to Szuwalski et al., 2019). Regardless, these methods did not noticeably outperform simpler methods in terms of overall error in short- or longer-term forecasts, and they had an overall tendency towards negative bias after 5- and 10-year forecasts. The latter might be a feature of how we parametrized the method or be specific to the OM suite, for which the estimated relationships might have been inexact, which is a common issue (e.g. He & Field, 2019). Simplex forecasts, modelled here assuming the same variance structure as the stock–recruitment methods, had similar patterns of error and bias but could be applied to all forecasts.

Certain methods assuming non-parametric stock–recruitment relationships (e.g. ecdfS3, ecdfR3 and markov), which could be applied to all cases, were overall unbiased and similar in overall error, but could occasionally result in high error. For sampling-based methods that are independent of SSB and use the near past recruitment as a reference period, the likelihood of bias (here positive) increased in longer-term forecasts. This is not surprising as these methods assume that future recruitment will resemble recent recruitment, an assumption that should be reasonable for the short term (e.g. Ward et al., 2014) but that will deteriorate as the interval between the two increases. The demonstrated positive bias appears to be a consequence of the suite of OMs included in our analysis, for which there was a tendency for recruitment to decrease after the point in time at which forecasts were initiated. All non-parametric methods with a similar reference period (e.g. sampling as random or weighted) usually contrasted relatively little, and therefore, there is little basis for selecting one over the other for 3- to 5-year forecasts.

The simple procedure of forecasting recent mean recruitment with autocorrelated deviates resulted in no apparent bias, even after 10 years, and forecast error was comparable to most other methods. The absence of bias contrasts with sampling methods using the same reference period. The main difference between the two types of methods is that sampling-based methods do not or only partially account for or replicate temporal autocorrelation within a fixed reference period. Autocorrelation ensures that the next step in the forecast is similar to the last one, which for ecological time series in general can improve skill (Ward et al., 2014). Forecasts of recruitment time series in specific have also been shown to improve by

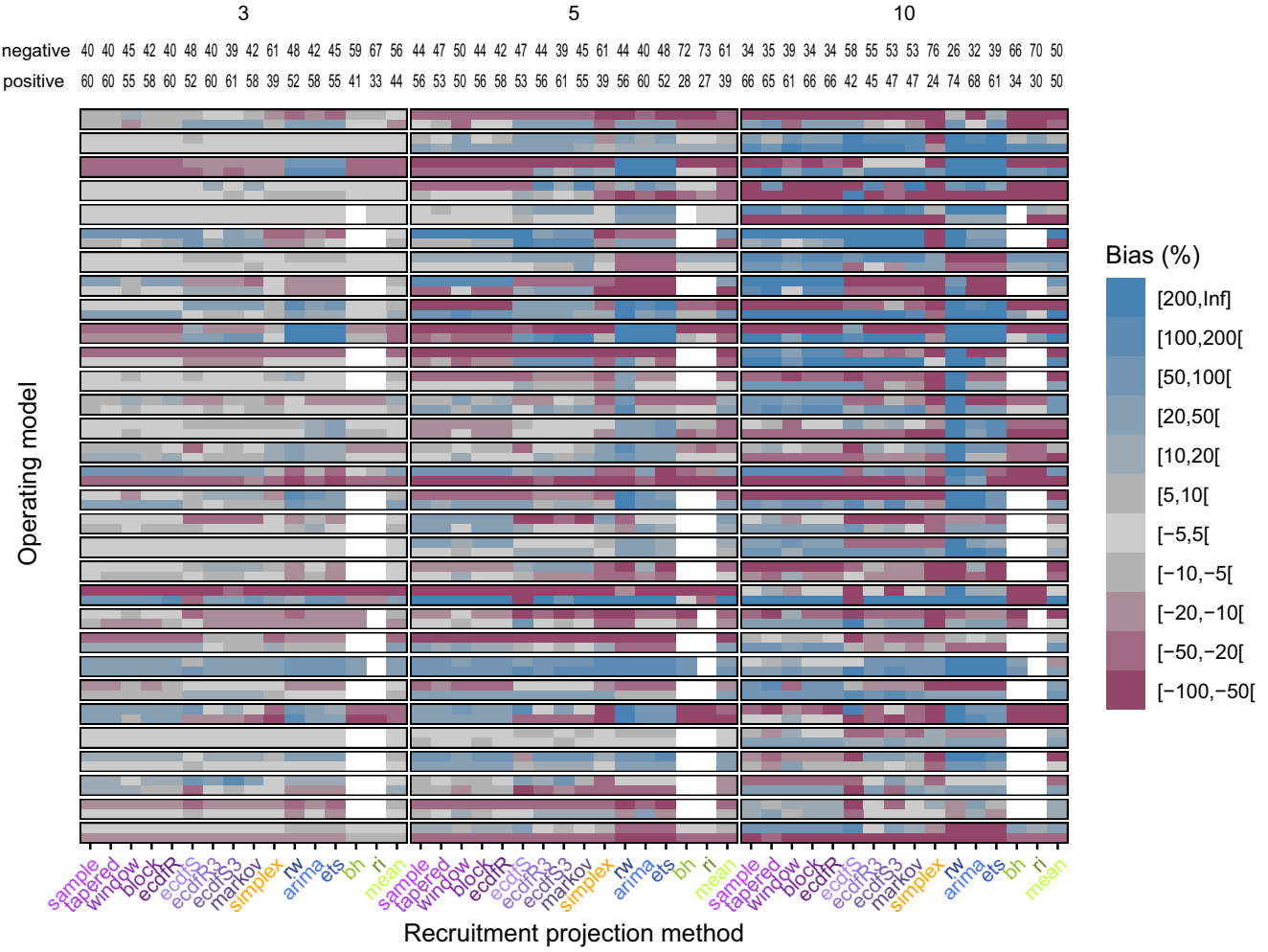

FIGURE 4 Bias (%) in forecast spawning stock biomass (SSB) at the end of the forecast period (3, 5 and 10 years; coloured blocks) for every operating model (OM; y-axis) at both initiation times (upper row = Y-20, lower row = Y-10 for each OM) for the different recruitment forecasting methods (x-axis). Results are ordered vertically in the same manner as in Figure 3. The table provides the percentage of negative and positive bias over the ensemble of OMs and initiation times. The labels for the recruitment forecast methods are coloured according to their class (Table 1). Stock–recruitment functions could not always be fitted, preventing testing of these forecasting methods in some cases (white boxes) (figure appears in colour in the online version only)

correctly specifying temporal autocorrelation (Johnson et al., 2016). However, this is not a general rule, as time-series methods, that also account for temporal autocorrelation but which are more complex and based on a longer-term reference period, proved to be the least reliable. This result corroborates Ward et al., (2014), who indicated that the increase in complexity comes at a price as parameters need to be estimated reliably and the observed long-term dynamics are expected to have some stability into the near future.

Knowledge of the probable performance of a forecast could help define their proper application, in terms of length (e.g. "How many years ahead can we reasonably forecast?"), delivery (e.g. "Should alternative hypotheses be presented?" or "Should estimates of risk be provided as broader classes rather than exact percentages?") and goal (e.g. "What management decisions can they inform?"). For example, the choice of a recruitment forecasting method will be largely inconsequential for 3-year forecast skill for later maturing stocks with low interannual recruitment variability (i.e. high AC). These forecasts—in the absence of other strong uncertainties—might be relatively trustworthy, which should affect how one delivers the results and reduce the necessity for extra analyses (e.g. sensitivity tests). In all situations where the stock matured earlier and recruitment was less autocorrelated, several potential methods could lead to different outcomes, even putting aside time-series methods with overall poor skill. For example, SSB forecasted after 3 years by either the Beverton–Holt or random sampling method could contrast on the order of 14%–100% (25% of all situations). The extreme case is early maturing stocks with important long-term variability in recruitment (high PV; e.g. small pelagic fish) for which the dynamics are forecasted 5–10 years, where the choice of a particular recruitment forecast method over another could be consequential for the evaluation of risks associated with management decisions. Erratic time series have traditionally been harder to forecast, so this is not surprising.

Our inability to identify single best recruitment forecasting methods argues for the use of multiple methods (Maunder &

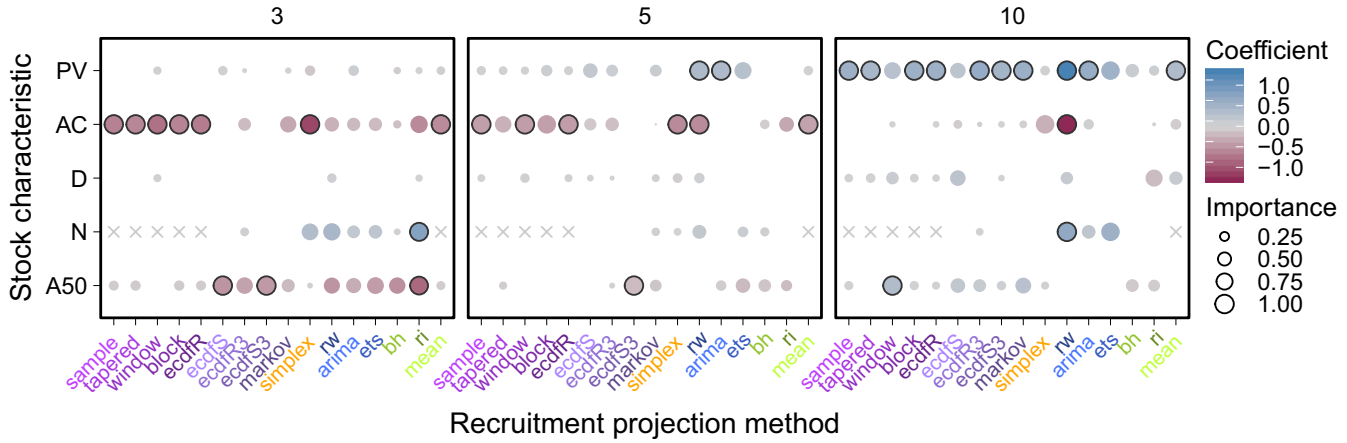

FIGURE 5 Effect size (coefficient) and relative importance of stock metrics on the logged error of each recruitment forecasting method (PV = proportional variability of recruitment, AC = lag-1 autocorrelation of recruitment, D= distance from the geometric mean of initiation-year recruitment, A50 = age at 50% maturity). Each column represents the output of one averaged multiple regression. Significant variables (p <.05) have a black border. For recruitment forecasting methods with a near-past reference period, no effect of series duration (N) was expected (grey crosses). Panels represent different forecast periods (3, 5 and 10 years), and label colours indicate classes of recruitment forecasting methods (Table 1) (figure appears in colour in the online version only)

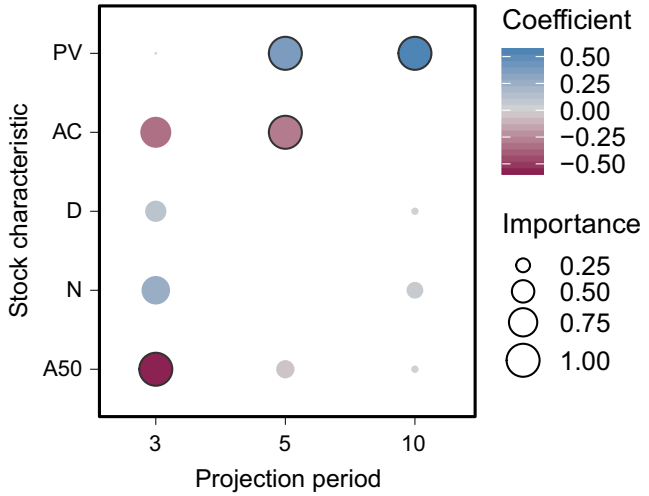

FIGURE 6 Effect size (coefficient) and relative importance of different operating model (OM) metrics on the dissimilarity in forecast bias among recruitment forecast method across OM and initiation periods, as a function of the forecast period (*x*-axis; PV = proportional variability of recruitment, AC = lag-1 autocorrelation of recruitment, D = distance from the geometric mean of initiation-year recruitment, N = series duration, A50 = age at 50% maturity) (figure appears in colour in the online version only)

Thorson, 2019; Punt et al., 2016), accomplished by exploring the consequences of different scenarios for forecasts (i.e. presented as different hypotheses) or by employing an ensemble or model averaging approach. The latter has the potential to augment forecast skill, but is however not guaranteed to do so (Yang, 2004) and might generate additional questions (e.g. "What methods to include?" or "Should they be given equal weight?"). Indeed, in this study there were several situations in which all or nearly all forecast methods consistently over- or underestimated SSB, against which multimodel inference cannot buffer. Such cases can arise because of the dependence of forecasts on the correct characterization of the (recent) past recruitment process and its application forwards. Achieving the former can be difficult because the stock and recruitment processes are estimated through the imperfect lens of a model applied to uncertain observations. Forecasting then relies on assuming stationarity in the assumed recruitment generating process, which will not be true if that process changes, such as due to a regime shift. Although the consideration of multiple alternative recruitment scenarios should thus generally be perceived as good practice, doing so will not eradicate the need to keep an open mind about the appropriate forecast period, presentation and goal.

There exists a much broader range of forecasting methods than those evaluated here and the literature on the forecasting of ecological time series is vast (Ward et al., 2014). We deliberately excluded more complex methods (e.g. machine learning techniques, Sun et al., 2009; cusp model, Sguotti et al., 2019; S-map, Deyle et al., 2018; Ye et al., 2015) as some are inappropriate for the shorter time series and simplicity is often preferable (easier to understand, apply and expand upon). There are also variations to the broad forecasting methods explored here. For example, the stock–recruitment relationships could be specified in other parametric forms (e.g. Needle, 2001), with parameters modelled stochastically and estimated inside the model (rather than outside using model output as in our study), and presuming different residual structures in terms of their distribution (e.g. t-distribution) and temporal autocorrelation (e.g. ARIMA instead of lag-1 autocorrelation). In some instances, error in the stock–recruitment relationships is simulated by sampling from residuals rather than a statistical distribution. Time-series models might need to be, such as all models with a certain complexity, fine-tuned to the stock in question (e.g. after inspection of residuals and parameters), and sampling methods are so flexible that there

are a multitude of possible spin-offs (Kimoto et al., 2007; Paz & Larrañeta, 1992).

Forecasts in stock assessment involve forecasting several demographic characteristics (e.g. selectivity, Kraak et al., 2019; weight at age, Jaworski, 2011). Decisions for all of these are required, and this can have important impacts on the determination of risk associated with different forecasted fishing strategies. Past studies have evaluated the conditioning of forecasts by measuring its skill either through simulation (e.g. Johnson et al., 2016) or through hindcasting (e.g. Kell et al., 2016; also referred to as retrospective forecasting, Brooks & Legault, 2016). The latter is accomplished by refitting an assessment model to truncated time series and comparing forecasted states (e.g. SSB) or forecasted observations (e.g. an SSB index) to the information from the full time series. Hindcasting evaluates "overall" performance because all sources of error are integrated, which includes forecasting but also estimation error through retrospective patterns (Brooks & Legault, 2016). Simulations on the other hand allow analysing elements under controlled circumstances, but are normally conditioned on a particular stock and data set so that most evaluations are constrained to relatively specific settings. They typically ignore the messiness proper to practical situations, which facilitates the analyses and their interpretation, but risks providing overly optimistic results. We have taken a hybrid approach; stock assessment output was directly taken as a "perfect world" test bed for forecasting algorithms. The advantages of this approach are that the suite of OMs as well as the individual models better reflect actual stock assessment conditions and that methods are systematically evaluated within the same framework (stock assessment) in which the forecasts are performed. However, the limited numbers of fixed OMs make this approach less flexible and potentially less general than a typical simulation study. We believe the two approaches can be complementary. The present approach is grounded in the constraints of reality and implicitly accounts for the influence of interacting components that influence projections, while a fully simulated study can validate the generalities of specific conclusions drawn (e.g. the performance of the stock–recruitment relationships under certain conditions or with various configurations) and can be sufficiently replicated to test things such as the coverage of confidence intervals.

While analysts should ideally evaluate the appropriateness and skill of forecasting methods in the specific context of their stock, this is often not feasible or may not be reliable because 1) there are many components to forecast and evaluating the role of each can be laborious, 2) time series can be short, precluding proper evaluation and 3) situation-specific evaluation risks drawing spurious conclusions. The benefits of a baseline comparison such as ours include a more robust evaluation across a range of contexts and a roadmap for selecting forecast methods and identifying the potential impact of this choice on forecasted biomass.

ACKNOWLEDGEMENTS

We would like to greatly thank everyone involved in the assessments of the stocks for which we extracted output and who made this information available. We would also like to express our gratitude to Kelli F. Johnson and the two anonymous reviewers, all of whom provided very thoughtful suggestions that significantly improved this manuscript.

DATA AVAILABILITY STATEMENT

Data will be made available upon request.

ORCID


Elisabeth Van Beveren https://orcid.org/0000-0002-9378-0215

REFERENCES

Anderson, S. C., Monnahan, C. C., Johnson, K. F., Ono, K., & Valero, J. L. (2014). ss3sim: An R package for stock assessment simulation with stock synthesis. *PLoS One*, *9*(4), e92725. https://doi.org/10.1371/journal.pone.0092725

Basson, M. (1999). The importance of environmental factors in the design of management procedures. *ICES Journal of Marine Science*, *56*(6), 933–942. https://doi.org/10.1006/jmsc.1999.0541

Berg, C. W., & Nielsen, A. (2016). Accounting for correlated observations in an age-based state-space stock assessment model. *ICES Journal of Marine Science*, *73*(7), 1788–1797. https://doi.org/10.1093/icesjms/fsw046

Brodziak, J. (2018). *AGEPRO Reference Manual. Version 4.2*. NOAA Fisheries.

Brooks, E., & Deroba, J. J. (2015). When "data" are not data: The pitfalls of post hoc analyses that use stock assessment model output. *Canadian Journal of Fisheries and Aquatic Sciences*, *72*(4), 634–641. https://doi.org/10.1139/cjfas-2014-0231

Brooks, E. N., & Legault, C. M. (2016). Retrospective forecasting — evaluating performance of stock projections for New England groundfish stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, *73*(6), 935–950. https://doi.org/10.1139/cjfas-2015-0163

Butterworth, D. S., Ianelli, J. N., & Hilborn, R. (2003). A statistical model for stock assessment of southern bluefin tuna with temporal changes in selectivity. *African Journal of Marine Science*, *25*(1), 331–361. https://doi.org/10.2989/18142320309504021

Calcagno, V. (2020). *glmulti: Model selection and multimodel inference made easy*. (1.0.7.1) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/glmulti/index.html

Chen, D. G., & Irvine, J. R. (2001). A semiparametric model to examine stock-recruitment relationships incorporating environmental data. *Canadian Journal of Fisheries and Aquatic Sciences*, *58*, 1178–1186. https://doi.org/10.1139/f01-037

Cury, P. M., Fromentin, J.-M., Figuet, S., & Bonhommeau, S. (2014). Resolving Hjort's dilemma: How Is recruitment related to spawning stock biomass in marine fish? *Oceanography*, *27*(4), 42–47. https://doi.org/10.5670/oceanog.2014.85

De Oliveira, J., & Butterworth, D. (2005). Limits to the use of environmental indices to reduce risk and/or increase yield in the South African anchovy fishery. *African Journal of Marine Science*, *27*(1), 191–203. https://doi.org/10.2989/18142320509504078

Denney, N. H., Jennings, S., & Reynolds, J. D. (2002). Life–history correlates of maximum population growth rates in marine fishes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *269*(1506), 2229–2237. https://doi.org/10.1098/rspb.2002.2138

Deyle, E., Schueller, A. M., Ye, H., Pao, G. M., & Sugihara, G. (2018). Ecosystem-based forecasts of recruitment in two menhaden species. *Fish and Fisheries*, *19*(5), 769–781. https://doi.org/10.1111/faf.12287

DFO. (2010). *Proceedings of the Newfoundland and Labrador regional Atlantic cod framework meeting: reference points and projection methods for Newfoundland cod stocks* (Can. Sci. Advis. Sec. Proceed. Ser. No. 2010/053; p. 61).

DFO. (2011). *Recovery potential assessment of the maritime designatable unit of American plaice (*Hippoglossoides platessoides*)*. (Can. Sci. Advis. Sec. Sci. Advis. Rep. No. 2011/043; p. 30).

Evans, G. T., & Rice, J. C. (1988). Predicting recruitment from stock size without the mediation of a functional relation. *ICES Journal of Marine Science*, *44*(2), 111–122. https://doi.org/10.1093/icesjms/44.2.111

Fogarty, M. (2001). Recruitment of cod and haddock in the North Atlantic: A comparative analysis. *ICES Journal of Marine Science*, *58*(5), 952–961. https://doi.org/10.1006/jmsc.2001.1108

Gilbert, D. J. (1997). Towards a new recruitment paradigm for fish stocks. *Canadian Journal of Fisheries and Aquatic Sciences*, *54*(4), 969–977. https://doi.org/10.1139/f96-272

Gröger, J. P., & Fogarty, M. J. (2011). Broad-scale climate influences on cod (*Gadus morhua*) recruitment on Georges Bank. *ICES Journal of Marine Science*, *68*(3), 592–602. https://doi.org/10.1093/icesjms/fsq196

Haltuch, M. A., Brooks, E. N., Brodziak, J., Devine, J. A., Johnson, K. F., Klibansky, N., Nash, R. D. M., Payne, M. R., Shertzer, K. W., Subbey, S., & Wells, B. K. (2019). Unraveling the recruitment problem: A review of environmentally-informed forecasting and management strategy evaluation. *Fisheries Research*, *217*, 198–216. https://doi.org/10.1016/j.fishres.2018.12.016

He, X., & Field, J. C. (2019). Effects of recruitment variability and fishing history on estimation of stock-recruitment relationships: Two case studies from U.S. West Coast fisheries. *Fisheries Research*, *217*, 21–34. https://doi.org/10.1016/j.fishres.2018.06.001

Heath, J. P. (2006). Quantifying temporal variability in population abundances. *Oikos*, *115*(3), 573–581. https://doi.org/10.1111/j.2006.0030-1299.15067.x

Heath, J. P., & Borowski, P. (2013). Quantifying proportional variability. *PLoS One*, *8*(12), e84074. https://doi.org/10.1371/journal.pone.0084074

Hjort, J. (1914). *Fluctuations in the great fisheries of northern Europe, viewed in the light of biological research* (Vol. *20*). Andr. Fred. Høst & Fils.

Hutchings, J. A., Myers, R. A., García, V. B., Lucifora, L. O., & Kuparinen, A. (2012). Life-history correlates of extinction risk and recovery potential. *Ecological Applications*, *22*(4), 1061–1067. https://doi.org/10.1890/11-1313.1

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.), OTexts. OTexts.com/fpp2. Accessed on 10/2020

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y., & Zhou, Z. (2019). *forecast* (8.9) [Computer software]. Retrieved from http://pkg.robjhyndman.com/forecast

Hyndman, R. J., & Hyndman, R. (Eds.) (2008). *Forecasting with exponential smoothing: The state space approach*. Springer.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Iles, T. C. (1994). A review of stock-recruitment relationships with reference to flatfish populations. *Netherlands Journal of Sea Research*, *32*(3–4), 399–420. https://doi.org/10.1016/0077-7579(94)90017-5

Jaworski, A. (2011). Evaluation of methods for predicting mean weight-at-age: An application in forecasting yield of four haddock (*Melanogrammus aeglefinus*) stocks in the Northeast Atlantic. *Fisheries Research*, *109*(1), 61–73. https://doi.org/10.1016/j.fishres.2011.01.017

Johnson, K. F., Councill, E., Thorson, J. T., Brooks, E., Methot, R. D., & Punt, A. E. (2016). Can autocorrelated recruitment be estimated using integrated assessment models and how does it affect population forecasts? *Fisheries Research*, *183*, 222–232. https://doi.org/10.1016/j.fishres.2016.06.004

Kell, L. T., Kimoto, A., & Kitakado, T. (2016). Evaluation of the prediction skill of stock assessment using hindcasting. *Fisheries Research*, *183*, 119–127. https://doi.org/10.1016/j.fishres.2016.05.017

Kimoto, A., Mouri, T., & Matsuishi, T. (2007). Modelling stock–recruitment relationships to examine stock management policies. *ICES Journal of Marine Science*, *64*(5), 870–877. https://doi.org/10.1093/icesjms/fsm054

Kraak, S. B. M., Haase, S., Minto, C., & Santos, J. (2019). The Rosa Lee phenomenon and its consequences for fisheries advice on changes in fishing mortality or gear selectivity. *ICES Journal of Marine Science*, *76*(7), 2179–2192. https://doi.org/10.1093/icesjms/fsz107

Kurota, H., Szuwalski, C. S., & Ichinokawa, M. (2020). Drivers of recruitment dynamics in Japanese major fisheries resources: Effects of environmental conditions and spawner abundance. *Fisheries Research*, *221*, 105353. https://doi.org/10.1016/j.fishres.2019.105353

Licandeo, R., Duplisea, D. E., Senay, C., Marentette, J. R., & McAllister, M. K. (2020). Management strategies for spasmodic stocks: A Canadian Atlantic redfish fishery case study. *Canadian Journal of Fisheries and Aquatic Sciences*, *77*(4), 684–702. https://doi.org/10.1139/cjfas-2019-0210

MacKenzie, B. R., Horbowy, J., & Köster, F. W. (2008). Incorporating environmental variability in stock assessment: Predicting recruitment, spawner biomass, and landings of sprat (*Sprattus sprattus*) in the Baltic Sea. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(7), 1334–1341. https://doi.org/10.1139/F08-051

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam, & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–297). University of California Press.

Maunder, M. N., & Thorson, J. T. (2019). Modeling temporal variation in recruitment in fisheries stock assessment: A review of theory and practice. *Fisheries Research*, *217*, 71–86. https://doi.org/10.1016/j.fishres.2018.12.014

Myers, R. A., & Pepin, P. (1994). Recruitment variability and oceanographic stability. *Fisheries Oceanography*, *3*(4), 246–255. https://doi.org/10.1111/j.1365-2419.1994.tb00102.x

Needle, C. L. (2001). Recruitment models: Diagnosis and prognosis. *Reviews in Fish Biology and Fisheries*, *11*, 95–111. https://doi.org/10.1023/A:1015208017674

Nielsen, A., & Berg, C. W. (2014). Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, *158*, 96–101. https://doi.org/10.1016/j.fishres.2014.01.014

Nielsen, A., Berg, C. W., Kristensen, K., Brooks, M., & Albertsen, C. M. (2019). *stockassessment (0.8.1)* [Computer software]. Retrieved from https://github.com/fishfollower/SAM

Paz, J., & Larrañeta, M. G. (1992). Testing non-parametric methods to estimate cod (*Gadus morhua*) recruitment in NAFO divisions 3NO. *Scientific Council Studies*, *18*, 27–31.

Planque, B., Fox, C. J., Saunders, M. A., & Rockett, P. (2003). On the prediction of short term changes in the recruitment of North Sea cod (*Gadus morhua*) using statistical temperature forecasts. *Scientia Marina*, *67*(S1), 211–218. https://doi.org/10.3989/scimar.2003.67s1211

Punt, A. E., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A. A., & Haddon, M. (2016). Management strategy evaluation: Best practices. *Fish and Fisheries*, *17*(2), 303–334. https://doi.org/10.1111/faf.12104

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

RAM Legacy Stock Assessment Database. (2020). *RAM Legacy Stock Assessment Database (v4.491)*. Zenodo. https://doi.org/10.5281/ZENODO.3676088

Rothschild, B. J. (1986). *Dynamics of marine fish populations*. Harvard University Press.

Sguotti, C., Otto, S. A., Cormon, X., Werner, K. M., Deyle, E., Sugihara, G., & Möllmann, C. (2019). Non-linearity in stock–recruitment relationships of Atlantic cod: Insights from a multi-model approach. *ICES Journal of Marine Science*, *77*(4), 1492–1502. https://doi.org/10.1093/icesjms/fsz113

Sharma, R., Porch, C. E., Babcock, E. A., Maunder, M. N., & Punt, A. E. (2019). Recruitment: Theory, estimation, and application in fishery stock assessment models. *Fisheries Research*, *217*, 1–4. https://doi.org/10.1016/j.fishres.2019.03.015

Stoker, M. & Noakes, D. J. (1988). Evaluating forecasting procedures for predicting Pacific herring (*Clupea harengus pallasi*) recruitment in British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, *45*, 928–935.

Subbey, S., Devine, J. A., Schaarschmidt, U., & Nash, R. D. M. (2014). Modelling and forecasting stock–recruitment: Current and future perspectives. *ICES Journal of Marine Science*, *71*(8), 2307–2322. https://doi.org/10.1093/icesjms/fsu148

Sugihara, G. (1994). Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, *348*(1688), 477–495. https://doi.org/10.1098/rsta.1994.0106

Sugihara, G., & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, *344*(6268), 734–741. https://doi.org/10.1038/344734a0

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496–500. https://doi.org/10.1126/science.1227079

Sun, L., Xiao, H., Li, S., & Yang, D. (2009). Forecasting fish stock recruitment and planning optimal harvesting strategies by using neural network. *Journal of Computers*, *4*(11), 1075–1082. https://doi.org/10.4304/jcp.4.11.1075-1082

Szuwalski, C. S., Britten, G. L., Licandeo, R., Amoroso, R. O., Hilborn, R., & Walters, C. (2019). Global forage fish recruitment dynamics: A comparison of methods, time-variation, and reverse causality. *Fisheries Research*, *214*, 56–64. https://doi.org/10.1016/j.fishres.2019.01.007

Walters, C. J., & Collie, J. S. (1988). Is research on environmental factors useful to fisheries management? *Canadian Journal of Fisheries and Aquatic Sciences*, *45*(10), 1848–1854. https://doi.org/10.1139/f88-217

Walters, C. J., & Ludwig, D. (1981). Effects of measurement errors on the assessment of stock–recruitment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, *38*(6), 704–710. https://doi.org/10.1139/f81-093

Ward, E. J., Holmes, E. E., Thorson, J. T., & Collen, B. (2014). Complexity is costly: A meta-analysis of parametric and non-parametric methods for short-term population forecasting. *Oikos*, *123*(6), 652–661. https://doi.org/10.1111/j.1600-0706.2014.00916.x

Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory*, *20*(01), https://doi.org/10.1017/S0266466604201086

Ye, H., Beamish, R. J., Glaser, S. M., Grant, S. C. H., Hsieh, C., Richards, L. J., Schnute, J. T., & Sugihara, G. (2015). Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, *112*(13), E1569–E1576. https://doi.org/10.1073/pnas.1417063112

Zwolinski, J. P., & Demer, D. A. (2019). Re-evaluation of the environmental dependence of Pacific sardine recruitment. *Fisheries Research*, *216*, 120–125. https://doi.org/10.1016/j.fishres.2019.03.022

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Van Beveren E, Benoît HP, Duplisea DE. Forecasting fish recruitment in age-structured population models. *Fish Fish*. 2021;00:1–14. https://doi.org/10.1111/faf.12562