Skill and value in recruitment forecasting

Skill & value in recruitment forecasting

Christian Kiaer*, Stefan Neuenfeldt and Mark R. Payne

** Corresponding Author (E-mail: cmrki@aqua.dtu.dk)*

Address: Section for Oceans and Arctic, National Institute of Aquatic Resources (DTU Aqua), Technical University of Denmark, Kemitorvet B201, 2800 Kongens Lyngby

Keywords: recruitment forecasting, ecological forecasting, predictive skill, forecast value

---

Abstract

Forecasting variation in the recruitment to fish stocks is one of the most challenging problems in fisheries science and one that essentially remains unsolved today. Traditionally recruitment forecasting involves assumptions of stationarity, is evaluated on explanatory power and rarely includes environmental processes.. Here we apply the lessons from fisheries science and other fields to propose a new generic approach to recruitment forecasting that allows the skill and value of recruitment forecasts to be assessed appropriately. We employ empirical multi-model ensembles to account for model structural uncertainty and consider forecasts in terms of both continuous and categorical forecasts. Forecast skill is assessed based on predictive power, using retrospective forecasts, while value is quantified using an economic cost-loss decision model. We demonstrate this framework using four stocks of lesser sandeel (*Ammodytes marinus*) in the North Sea. We find the skill and value of the forecasts for each stock is strongly influenced by the quality of the stock assessment: skilful and valuable forecasts are shown to be feasible in some management areas. This result shows the ability to produce valuable short-term recruitment forecasts, evaluated in a realistic setting, and paves the way for revisiting our ability to make skilful and valuable recruitment forecasts.

Introduction

Recent developments in ocean observations and modelling today make it possible to forecast many of the physical variables in the ocean (Doblas-Reyes *et al.*, 2013; Meehl *et al.*, 2014). Building on top of this data about the ocean environment, forecasts of marine ecological responses have been developed (Payne et al 2017) and provide managers and stakeholders the foresight needed to sustainably manage marine living resources (Tommasi *et al.*, 2017b; Hobday *et al.*, 2018). Examples of operational forecasts already in use include southern Bluefin tuna habitat forecasts (Eveson *et al.*, 2015), the dynamic fisheries bycatch management tool EcoCast (Hazen *et al.*, 2018), and blue whale habitat preference forecast (Hazen *et al.*, 2017). However, these operational fisheries forecast products are currently limited to predictions of distribution and phenology and there are currently no known operational marine fish recruitment forecasts (Payne *et al.*, 2017).

Understanding and forecasting changes in fish stock productivity has, however, been a key aspiration in fisheries science for the last century (Leggett and Deblois, 1994; Subbey *et al.*, 2014; Tommasi *et al.*, 2017a; Haltuch *et al.*, 2019). Recruitment, the number of young individuals produced each year, has a key role in shaping fish population dynamics (Hilborn and Walters, 1992), especially in determining total allowable catches for short-lived species, where the recruiting year-classes that are highly susceptible to environmental forcing, contribute a significant share of the landings. Environmental drivers play an important role in the productivity for such stocks (e.g. via temperature (MacKenzie *et al.*, 2008; Mantzouni and Mackenzie, 2010), salinity (Köster *et al.*, 2005) or phenology (Platt *et al.*, 2003)) and including climate information in stock-assessments can reduce uncertainties in stock status and the risk of over- or under harvesting (Hare *et al.*, 2010; Haltuch and Punt, 2011; Tommasi *et al.*, 2017a, 2017b). The ability to foresee changes in productivity on a short time-scale can therefore enable adaptive and pre-emptive decision-making strategies, benefiting both stakeholders and managers (Hobday et al., 2016; Payne et al., 2017; Welch et al., 2019).

Standard statistical methodology has shown that recruitment forecasts have had limited potential for tactical, i.e. short-term, management. Combined with spurious correlations and the large variety of underlying environmental, physical and ecosystem processes affecting recruitment simultaneously, recruitment forecasts are prone to errors and large uncertainty (Leggett and Deblois, 1994; Browman *et al.*, 1995; Myers, 1998; Tommasi *et al.*, 2017b). Fish population time series are often relatively short in length (Ricard *et al.*, 2012) and hampered by high observation noise, limiting the ability to develop and test predictive models (Clark and Bjørnstad, 2004; Ward *et al.*, 2014). Furthermore, previously published environment-recruitment correlations generally breakdown when confronted with new data, diminishing the uses for management (Myers, 1998; Tommasi *et al.*, 2017b). The relative importance of drivers of recruitment can also change from year to year ("non-stationarity") (Subbey *et al.*, 2014; Haltuch *et al.*, 2019). As a consequence of all of these processes, and a long history of failures, recruitment forecasts are widely viewed with scepticism in the community today.

Nevertheless, the potential of such forecasts to benefit all those that depend on living marine resources is clear. So how can this potential be realised? And even more importantly, how would we know when we have produced a useful forecast? To answer this question, here we take inspiration from other forecasting fields, and in particular from meteorology, a discipline that has also been attempting to predict chaotic and difficult to observe systems for nearly a century (albeit with considerably more success!). In particular, the question of "what makes a good forecast?" is addressed in a seminal 1993 paper in the field by Alan Murphy (Murphy, 1993) that presents two key relevant concepts, skill, and value, which we use as the basis for this work as follows.

Murphy presents forecast "skill" as the quantitative ability of the forecast: is it numerically correct? In the marine setting, and while some recruitment forecasts includes out-of-sample model verification (e.g. Fernandes *et al.*, 2015; Van Beveren *et al.*, 2021), model performance is often measured as in-sample accuracy (i.e. some goodness-of-fit measure) e.g (Lindegren *et al.*, 2018). There is however, a

fundamental difference between explanatory and predictive power: while explanatory models can be used to investigate causal hypotheses, models with high explanatory power cannot be expected to predict well (Levins, 1966; Shmueli, 2009; Dickey-Collas et al., 2014). Skill measures are plentiful and different metrics can result in different model performance rankings, highlighting the need for analysing forecasts using multiple assessment methods (Murphy and Winkler, 1987; Stow *et al.*, 2009; Brun *et al.*, 2016). In the atmospheric and climate sciences, skill is often assessed based on retrospective forecast analysis (Wilks, 2011) i.e. predicting beyond the period over which the model was developed or tuned. Furthermore, meteorology always places its forecasts in the context of a baseline or reference forecast (Jolliffe and Stephenson, 2012). Common baseline forecasts includes random selection of categories or using the average over a given reference time period, often referred to as climatology in atmospheric sciences (Jolliffe and Stephenson, 2012).

Secondly, Murphy discusses the usefulness of a forecast in terms of its "value" in aiding decision-making. A good forecast needs to be of value for an end-user by assisting in decision-making, providing economic value or otherwise benefiting the user (Murphy, 1993). While value in recruitment forecasts has been discussed (e.g. Walters, 1989; Field *et al.*, 2010), a quantitative approach to value is rarely seen in marine science. Simple economic decision models can analyse forecasts under simplified assumptions, helping end-users decide if it is economically wise to follow the forecast (Murphy, 1976a). Quantitatively providing a value assessment can help integrate forecast products directly into a user's framework, allowing users to assess the benefits of a given forecast system and can give a clear insight into how, and when, a forecast should be used (Murphy, 1976a).

Here we apply the concepts of skill and value to the development of recruitment forecasts. We present a generic framework for developing and assessing short-term recruitment forecasts for fish stocks from the perspective of Murphy (1993). Forecast skill is assessed based on predictive performance, using validation techniques currently used in atmospheric and meteorological sciences. Value is assessed quantitively,

using an economic cost-loss decision model, providing insight into the actual monetary value of the forecast product. Furthermore, we also take advantage of the state-of-the-art in marine science and statistics to produce the best forecasts possible, accounting for non-stationarity and model structural uncertainty. We illustrate the framework using multiple stocks of the ecologically and economically important lesser sandeel (*Ammodytes marinus*) in the North Sea, where explanatory studies of recruitment have already highlighted several recruitment drivers (Arnott and Ruxton, 2002; van Deurs *et al.*, 2009; Lindegren *et al.*, 2018).

Methods

Recruitment forecast framework

This work presents a generic framework (Figure 1a) allowing for assessing recruitment forecasts for fish stocks. Borrowing from the atmospheric sciences, we propose an iterative ensemble-modelling approach for productivity predictions. The model ensemble consists of generalized additive models (GAMs) (Hastie and Tibshirani, 1986), weighted using multi-model inference (Burnham and Anderson, 2004). Models are fitted using a combination of stock assessment data and satellite sea surface temperature (SST) measurements, as well as a time varying cohort year effect, allowing for changes in productivity due to other unquantified effects. Models are updated when new data becomes available, repeating the fitting process. Predicted categories indicates the probability of a certain event occurring, using a tercile division of the historical observed recruitment time series (i.e. the probability of a low, medium or high

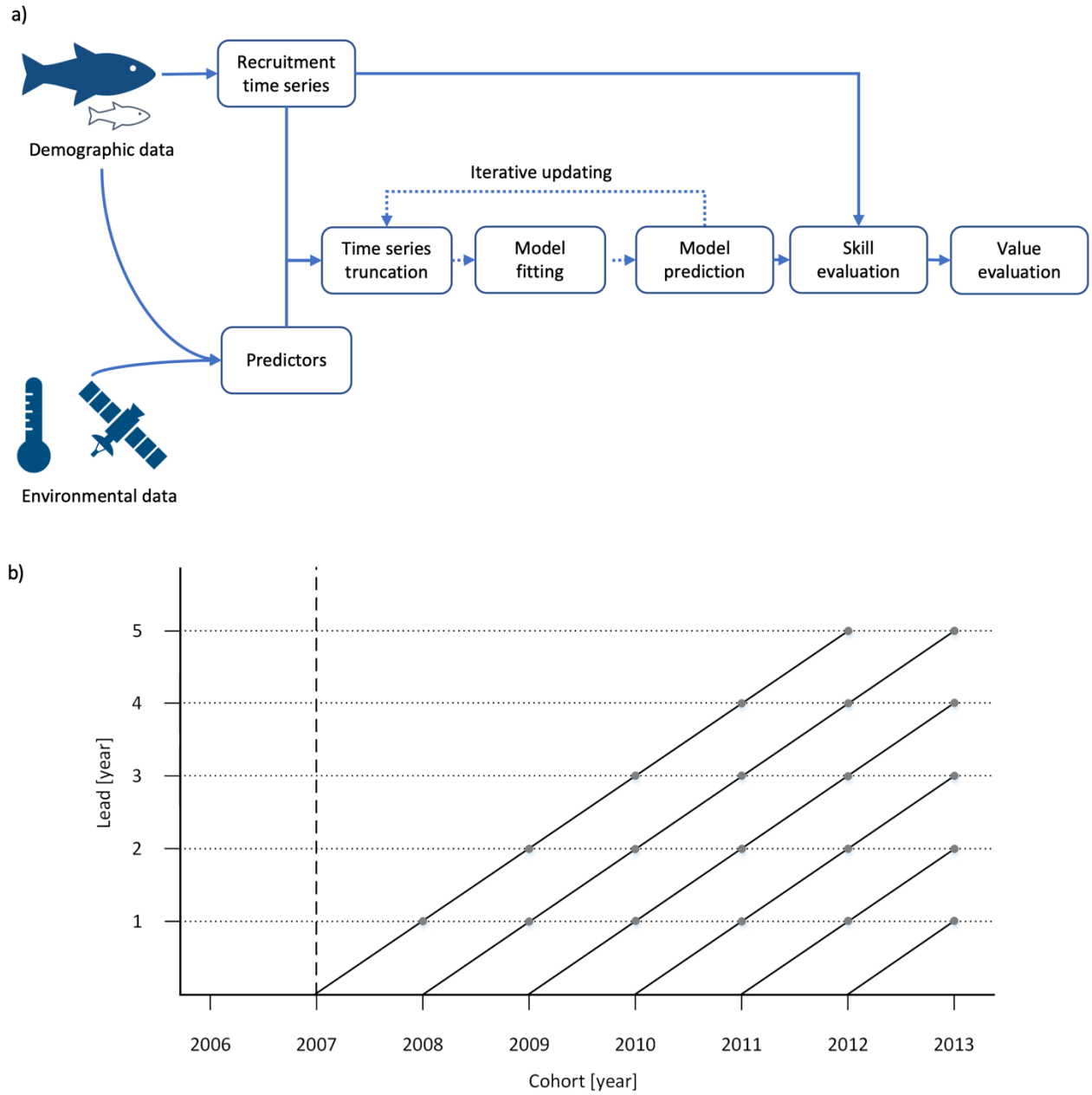


Figure 1 a) Overview of the process in the forecasting framework. Here, data are extracted and combined into the appropriate modelling data. Afterwards, an iterative process of data truncation and model fitting are the basis of all model objects and predictions. These predictions contains both the current and the retrospective predictions, which can be used for skill and value evaluation b) Schematic of retrospective forecast system used to generate a retrospective forecast time series. One time series is generated at each lead time. Dashed line indicates first data cut-off and the start of the retrospective forecasting period. Dotted lines indicates the forecast time series at a given lead. After the first cut-off each subsequent retrospective forecast will include the previous year's observations increasing the size of the model training data set. For each generated retrospective forecast time series skill, value and accuracy will be evaluated. Depending on species, stocks, data availability and period of interest, the evaluated cohort period and the start of the retrospective analysis can vary.

recruitment year based on historical observations). Forecast performance is assessed on predictive power, instead of explanatory power, evaluating out-of-sample retrospective forecasts. The evaluation ranges from continuous accuracy metrics, such as root-mean-squared error (RMSE), to categorical hit rates and true skill score (TSS) (Wilks, 2011). Furthermore, the framework allows for forecast aggregation

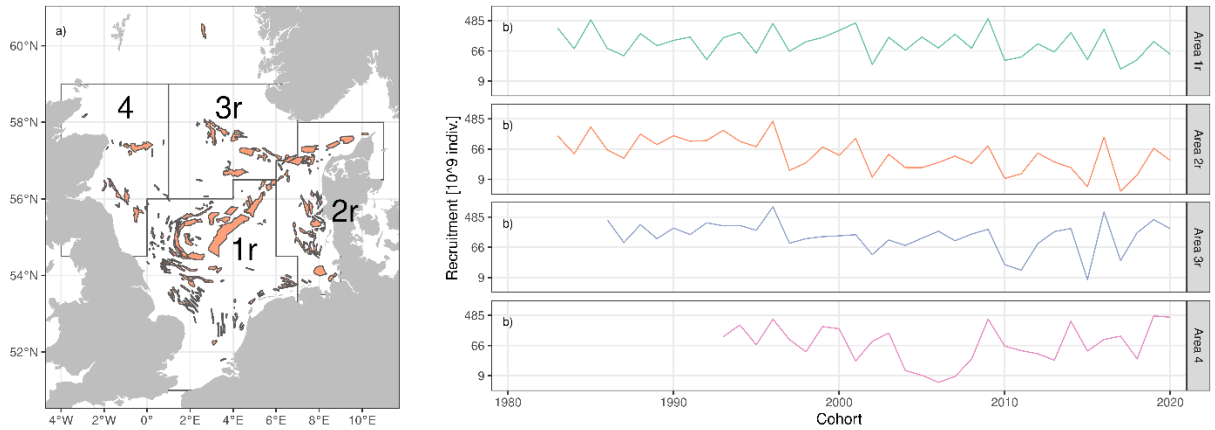


Figure 2 : a) Map of the North Sea showing the four management areas of sandeel assessed analytically. Sandy habitat banks, the predominant sandeel habitat, are shown in orange. b) Recruitment time series for the four sandeel stocks from the official ICES stock assessment. Dashed-horizontal lines mark the delineation of the upper and lower terciles for each stock.

of multiple stocks and species into a single portfolio forecast. This combines information from multiple stocks into a single portfolio forecast, where the overall productivity of the selected system can be assessed. A portfolio effect, as seen in finances can be expected (Schindler *et al.*, 2010), where an overall reduction in the error is seen.

Sandeel Case Study

The lesser sandeel (*Ammodytes marinus*), is a pelagic species of the Ammodytidea family and is one of the most common sandeels found in the North Sea. Adult lesser sandeel habitats are found in most of the North Sea, generally distributed across the shallow sandy banks (van Deurs et al., 2009, Figure 1a). Particle-tracking studies and the sedentary state of post-recruitment sandeel (Christensen et al., 2008; Pedersen et al., 2019) resulted in a division into 7 different individually managed North Sea sandeel stocks. Analytical stock assessments are done in management areas 1r, 2r, 3r and 4 (see Figure 2a), while the remaining three stocks are considered data poor. Sandeel is seen as one of the main links between primary production and the higher trophic levels in the North Sea, observed for both larger piscivorous fish (e.g. cod and haddock) and seabirds (Eliasen *et al.*, 2011). The lesser sandeel has historically supported a large fishery, which has seen a large decline in recent years (Dickey-Collas et al., 2014) . Due to the importance of the species, recruitment to these stocks is well studied (Arnott and Ruxton, 2002; van Deurs *et al.*, 2009; Eigaard *et al.*, 2014; Lindegren *et al.*, 2018). In the south western

part of the North Sea (i.e. management area 1r), the sandeel shows signs of being influenced negatively by temperature, while the abundance of the main prey, *Calanus finmarchus*, has a positive influence (Arnott and Ruxton, 2002; Lindegren *et al.*, 2018). Density dependence is also been found to be an important driver, where competition with young adults and juveniles has a negative effect on recruitment (van Deurs *et al.*, 2009). Currently, stock assessment uses a geometric mean for recruitment predictions (ICES, 2018). These geometric means will be used as continuous reference models during skill evaluation.

Data

Operational forecasts require data to be available at the time of the forecast, potentially excluding some data which otherwise may be of relevance to recruitment prediction. For example, estimates of zooplankton prey, *Calanus finmarchicus* and *Temora longicornis* (Arnott and Ruxton, 2002; van Deurs *et al.*, 2009; Lindegren *et al.*, 2018) used in other studies are only available with 1-2 years delay, and are therefore of limited value in forecasting recruitment in this stock. We focus our analyses on data that is available with a maximum of a few months delay. An overview of the data employed is provided in Table 1 and complete time series are found in Figure S1.

Assessment data

Assessment data used for sandeel modelling is obtained from official ICES advice, based on the stochastic multi-species assessment model, SMS (Pedersen et al., 1999). The SMS model is run in a single-stock mode for sandeel assessments, and integrates data on catches, catch effort, maturity, weight, fishing mortality and natural mortality at a given age (ICES, 2018). All stock assessment data are the current (2021) assessments provided by ICES for area 1r, 2r, 3r and 4 (Figure 2b), where recruits are treated at age 0. From the assessment data, 4 demographic variables are extracted, consisting of spawning stock biomass (SSB), total stock biomass (TSB), number of individuals (N) and number of one-year olds (N1). This

allows for different types of interactions between the demography, including density dependence and SSB impact on recruitment. All demographic data is log-transformed before use in modelling and converted to log-anomalies (relative to the average log-value over the full time series for each stock).

Environmental data

High resolution spatial sea surface temperature data is gathered from the Optimum Interpolation Sea Surface Temperature (OISST) product (Banzon *et al.*, 2016). The product is a 0.25° x 0.25° global daily sea surface temperature (SST) data set on a regular grid. Noting that adult sandeel are bound to specific banks (Christensen *et al.*, 2008), we produced daily average temperatures over the banks in each stock area, and then averaged temporally over quarters as follows: P3 and P4 represents the temperature anomalies experienced by the adult sandeel from July to December before and during spawning (i.e. the temperatures experienced by the spawners just before spawning). Q1, Q2, Q3 and Q4 are SST anomalies experienced during the egg, larval and juvenile stages from January to December for a given cohort. All extracted temperatures were converted to anomalies from the average (climatology) over the complete SST time series period (1983 to 2020) prior to use in modelling.

Model ensembles

Table 1 List of all variables considered and the rationale behind. The parameterisation of each variable in the model is also shown, with s() indicates the use of a spline-smoother and other terms indicating the incorporation of that term as a linear response term.

Variable	***Description***	***Rationale***	***Parameterisation***
Demographic explanatory variables			
SSB	Spawning stock biomass	Adult biomass that determines amount of eggs spawned	s(log(SSB))
N1	Number of 1-year olds	Number of individuals at age 1 inducing a density dependence	log(N1)
SumN	Number of individuals	Entire sandeel population, inducing density dependence	log(SumN)
TSB	Total stock biomass	Combination of all of the above	log(TSB)
Enviromental explanatory variables			
P3	Jul-Sep temperatures	Temperatures experienced by the adults prior to spawning	P3

P4	Oct-Dec temperatures	Temperatures experienced by the adults prior to / during spawning	P4
Q1	Jan-Mar temperatures	Temperature experienced during egg development	Q1
Q2	Apr-Jun temperature	Temperature experienced by larvae during pelagic drift phase	Q2
Q3	Jul-Sep temperature	Temperature experienced by post-settlement juveniles	Q3
Q4	Oct-Dec temperature	Temperature experienced by post-settlement juveniles	Q4
Other explanatory variables			
Cohort	Cohort year	Included to allow time-variation in the mean productivity of the stock due to systematic shifts in other unquantified variables	s(Cohort)

A multi-model ensemble was generated based on the 2048 possible combinations of the 11 variables available (Table 1). However, not all possible considerations were considered: to minimize risk of overfitting due to both collinearity between model parameters and the short time-series, predictors are split into three groups (as shown in Table 1) based on an exploratory analysis of collinearity (i.e. environmental, demographic and other predictors). Models in the ensemble that incorporated more than one variable in a given group were excluded, giving a total of 819 candidate model structures to be considered.

The cohort-based non-parametric time-varying smoother allows for changes in the underlying productivity to be incorporated (e.g. due to unquantified variables). Here, additional effects not captured by the selected variables are implicitly caught and incorporated in the model structure. This allows non-stationarity and systematic shifts in recruitment patterns, which would otherwise not be accounted for. For in-depth model descriptions, see supplementary methods.

Forecast evaluation and skill metrics

The forecast system was evaluated on retrospective forecasts, mimicking operational usage, using the system in a state where only a limited subset of the data is available (Figure 1). Models are first trained on all data up to a cut-off point, selected by the data availability for a given stock. The model ensemble predicts the complete recruitment time series, allowing analysis of both in-sample and out-of-sample performance. This process was repeated moving the cut-off point forwards by one year, creating in-sample and out-of-sample time series for both positive and negative lead times (Figure 1b). Values from the retrospective forecasts are then compared to observed values and skill measures calculated.

The fitted modelsare weighted according to AICc, which allows for smaller sample sizes in relation to model parameters (Anderson, 2008). An ensemble predictive distribution is generated by drawing from the posterior distribution of each individual model weighted by the AICc, in total generating 4000 draws containing the ensemble prediction. Probabilistic categories (i.e. high, medium and low recruitment) are generated from tercile thresholds, based on historical observation, and the ensemble predictive distribution. Continuous forecasts are evaluated as the mean of the predictive forecasts distribution.

Creating time series for all lead times results in a flexible skill evaluation, focusing on the lead period important for a given species. As 1-year olds are selected as the first year class in sandeel assessment (ICES, 2010), this is chosen as the main lead time in this work.

Skill metrics

Multiple performance metrics are used to assess the retrospective forecasts (Table 2), including both continuous and categorical skill evaluations (Stow *et al.*, 2009; Jolliffe and Stephenson, 2012; Brun *et al.*, 2016). Continuous skill uses the mean prediction for a root-mean-square error (RMSE) analysis, giving indications of the accuracy of the forecast. Continuous forecasts can use the mean-squared-error skill score (MSESS) to directly compare the forecast with a reference forecast. The categorical forecasts (high,

medium and low) are analysed using the hit rate, false alarm rate and true skill score (TSS). Using a combination will quantify both the accuracy of the forecast and forecast performance in each terciles (Murphy, 1969).

Table 2 Performancel metrics used to evaluate the skill of the forecast system. These values are calculated over all retrospective forecasts at a given lead time. Continuous skill evaluation is also performed for reference forecasts, while the categorical and binary forecast evaluation are only calculated for probabilistic forecasts (Murphy, 1969). MSE contains the mean of the difference between the forecasted (F) and observed (O). The hit rate (H) consists of the proportion of correct forecasts (i.e. true positives (TP) and true negatives (TN)), while false alarm rate (F) is the proportion of incorrect forecasts (i.e. false positives (FP) and false negatives (FN)).

Name of Forecast Quality Measure	**Definition**	**Range**	**Application**
Mean square error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(F_i - O_i)$	[0,inf]	Cont.
Mean square error skill score (MSESS)	$1 - \frac{MSE}{MSE_{reference}}$	[-inf,1]	Cont.
Root-mean-square error (RMSE)	$\sqrt{MSE}$	[0,inf]	Cont.
Proportion correct / Hit rate	$H = \frac{TP + TN}{(TP + TN) + (FP + FN)}$	[0,1]	Cat.
False alarm rate	$F = \frac{FP + FN}{(TP + TN) + (FP + FN)}$	[0,1]	Cat.
True Skill Score (TSS)	$TSS = H - F$	[-1,1]	Cat.

Reference forecasts were selected according to current stock assessment practices: in this way, it was immediately apparent if the forecast outperforms existing procedures. For the sandeel, the official ICES sandeel advice uses either the 10-years moving geometric mean (Area 2r and 4) or the geometric mean of the full time series (Area 1r and 3r) (ICES, 2018). These models are selected as reference forecasts in the MSESS. The skill score ranges from negative infinity to 1, effectively comparing the performance gains from using a given forecast compared to the reference. For categorical forecasts, the reference forecast is selected to be random guessing (33% correct) baseline, both for True Skill Score (TSS) and Ranked Probability Skill Score (RPSS).

Table 3 a) Confusion matrix generated from the retrospective forecasts at a given lead. Constructed from the sum of observed positives (event occurred) and observed negatives (event didn't occur) with corresponding predicted positives (predict event occurred) and predictive negatives (predicted event not to occur). This results in a matrix of true positive (TP), false positives (FP), false negatives (FN) and true negatives (TN). For recruitment predictions, a TP is when the forecasting system correctly predicts the observed tercile, while a FP is when the system predicts a given tercile, which is not observed. For negative events this is reversed, i.e. FN the tercile is observed while the forecast system doesn't predict it and TN is when the tercile is not observed and the system doesn't predict it. b) Cost matrix used to calculate the value of the forecast system. Here a cost (C) is associated with a precaution and a loss is associated with not taking the precaution and the event occurring.

a) Contingency matrix

	Observed P	Observed N
Predict P	TP	FP
Predict N	FN	TN

b) Cost matrix

	Event occurs	Event does not occur
Precaution taken	C	C
Precaution not taken	L	0

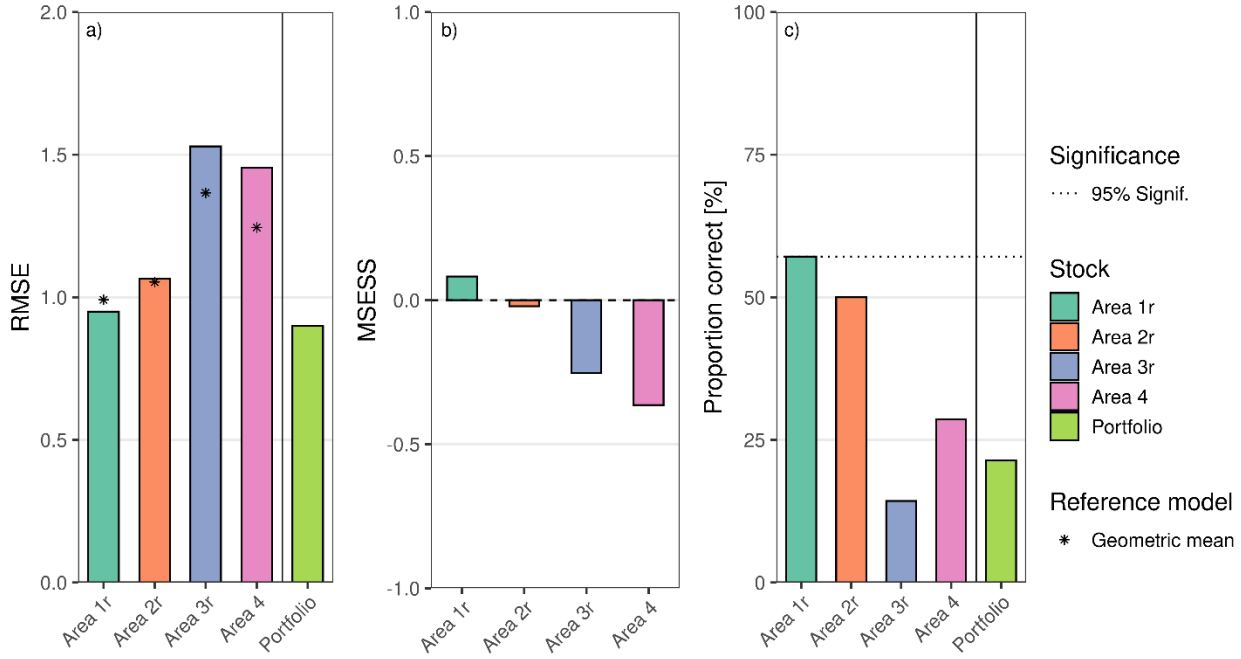
Forecast value

We assess the value of the forecasts using a Richardson cost-loss decision model (Richardson, 2000). Simple economic models, as used here, are widely used in the climate services sector (Pope *et al.*, 2019) to quantify value of e.g. seasonal forecast systems, and provide an intuitive metric for users (Murphy, 1976b). Briefly, the model considers the economic impacts of a particular event that is being forecast (e.g. poor recruitment), and the loss (L) that the user could potentially incur. However, the user also has the ability to avert these losses by implementing precautionary mitigation actions (e.g. based on a forecast), but doing so also incurs a cost (C) (e.g. mothballing processing plants). These two dimensions (i.e. whether the event occurs, and whether the user takes a precaution) each have two outcomes, and therefore form a 2x2 cost matrix (Jolliffe & Stephenson, 2012, see Table 3b). Combining this set of costs with the properties of the forecast system characterised by the contingency matrix (Table 3a) allows the expected expense over the long-term (E) to be calculated when the forecast is always ($E_{forecast}$) or never ($E_{reference}$) followed. The value (V) of the realised forecast system can then be calculated relative to a perfect forecast system as :

$$V = \frac{E_{reference} - E_{forecast}}{E_{reference} - E_{perfect}} \quad (1)$$

The value of the forecast system, V, is expressed as a non-dimensional number less than 1 and varies as a function of the cost-loss ratio (C/L) of a given user (Richardson, 2000, see eq. 1).

We necessarily extend this analysis to account for the (relatively) small sample size associated with our set of retrospective forecasts and therefore estimate the uncertainties in the forecast value. We model the retrospective contingency table (Table 3a) using a Bayesian multinomial model implemented in Stan (Stan Development Team, 2020) to estimate the vector of true probabilities $\boldsymbol{p} = \{p_{TP}, p_{FP}, p_{FN}, p_{TN}\}$ of



each quadrant of the contingency table. The posterior predictive distribution of $\boldsymbol{p}$ was then sampled 4000 times and used to construct a corresponding large set of contingency tables and therefore the statistical distribution of the forecast system value, V.

Results

Assessment of the predictions is presented at a forecast lead of one cohort beyond the final year of the assessment, mimicking potential operational usage in these stocks. Our results parallel those seen in other

Figure 3 a) Root-mean-squared error of the different management areas and the portfolio forecast at lead 1. Area 1r and 2r show highest accuracy of individual forecasts, while the portfolio is the *overall most accurate, indicating the presence of the portfolio effect.Stars show the reference geometric mean RMSE for the individual management areas. b) Mean-squared error skill score of the individual forecast products for lead 1. Official recruitment prediction model is used as a reference model. Here area 1r and 2r shows better or equal performance to the reference models, while area 3 and 4 has a negative skill score. c) Hit rate of the different management areas indicating the percentage of correct retrospective forecasts at lead 1. Dashed line indicates the 95th percentile level of the random guessing reference forecast. Area 1r are significantly better than random guessing at 57% hit rate, while area 2r are borderline significant with 50% hitrate. Area 3r, 4 and the portfolio shows large drop-offs in hit rate with hit rates below 30%.*

sandeel recruitment studies (Arnott and Ruxton, 2002; van Deurs *et al.*, 2009), where relative variable importance shows SSB is not a primary recruitment driver (Figure S2). Furthermore, analysis of the retrospective models shows a clear shift in recruitment dynamics over the evaluated period, indicating that the system is non-stationary (Figure S2). Information-rich stocks (i.e. area 1r and 2r) have the highest continuous forecast accuracy, while the two areas with the shorter time series show a higher RMSE (Figure 3a). Here the portfolio forecast has the highest overall accuracy, showing that the aggregation of predictions can lower the RMSE, indicating a portfolio effect. In the information-rich management area 1r, the continuous forecast accuracy is also better than existing models used in official sandeel stock assessment, as indicated by the positive MSESS at lead 1 (Figure 3b). The performance of Area 2r is on a par with the reference model, with a comparable RMSE to official models. Area 3r and 4 both show a negative MSESS at lead 1, indicating that the selected model ensemble would not be an improvement over the reference model when used as a continuous forecast.

Hit rate metrics (how often the system correctly forecasts high, medium or low recruitment) also show~~s~~ best results in management area 1r, outperforming random guessing (p=0.02, one-tailed test) with 57% correct (Figure 3c). Here, area 2r sees a hit rate of 50% correct (p=0.06, one-tailed test), significant at the

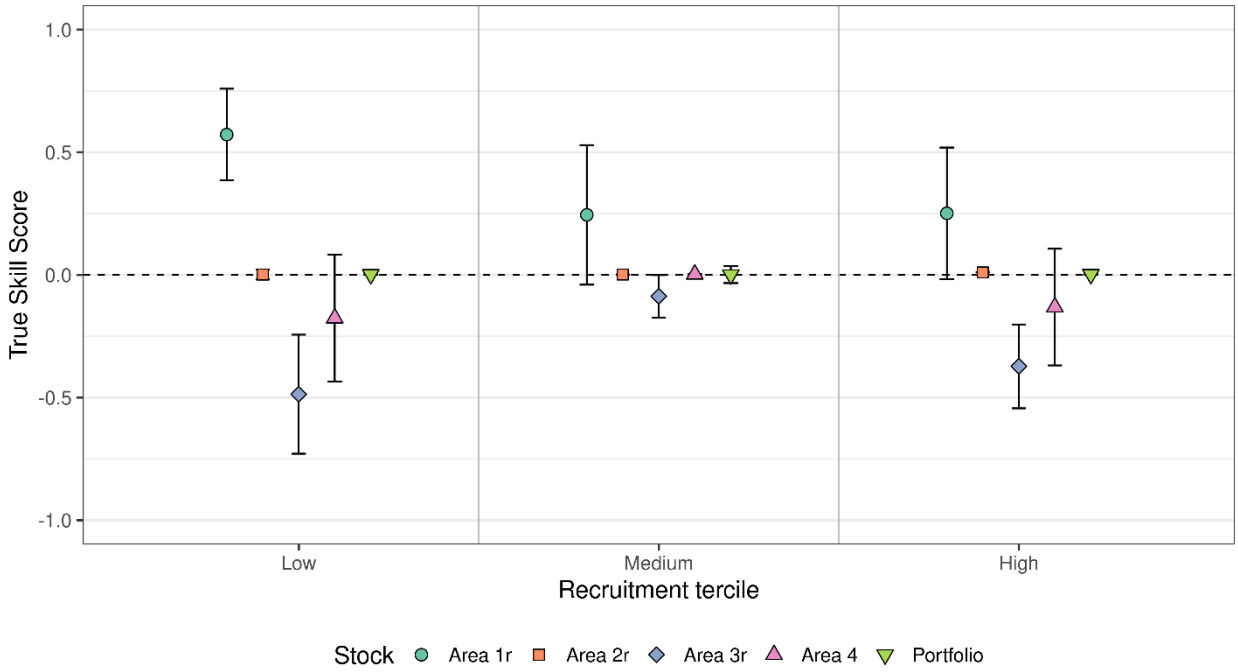


Figure *4* Model skill at lead 1 is represented as the True (Peirce) Skill Score (TSS), which ranges between +1 and -1, and has a value of 1 for perfect skill, and 0 for random guessing (black dashed line). Negative values indicates perverse forecast. The 95% confidence interval for the estimated skill score are shown as error bars on each of the points. Recruitment terciles are shown, with shapes indicating the corresp*onding stock. A positive TSS is seen for all recruitment terciles in area 1r, while all other models shows utility close to or worse than random guessing.*

90% level. A large drop off in hit rate is seen in area 3r and 4 (respectively at 14% and 28%), where performance is not significantly better than random guessing (p=0.90, and p=0.52, one-tailed tests). The portfolio categorical forecast is not significantly better than random guessing at a 21% proportion correct (p=0.74, one-tailed test).

Further insight into the models can be gained by examining the skill of predicting individual terciles. The true-skill score (TSS) metric combines the specificity (true-positive rate) with the sensitivity (true-negative rate) for a categorical forecast and is applied here to each tercile in turn. The TSS indicates area 1r being the only management area where the model can reliably differentiate all three categories (Figure 4), outperforming random guessing (i.e. where TSS=0) consistently in all categories. Most other areas do not show a significant ability to differentiate, either due to the small sample size or poor model skill. For example, area 2r's TSS is not significantly different from zero for all terciles, in part due to the relatively low recruitment seen in the stock in recent years, affecting the ability of the TSS metric to quantify the forecast skill. Areas 3r and 4 have negative or zero skill scores in all categories, likely due to the large

uncertainties propagating into these forecasts and resulting in a wide prediction distribution. The portfolio forecast shows similar TSS values to area 2r, with no categroies reaching levels where the system is able to correctly distinguishing between terciles.

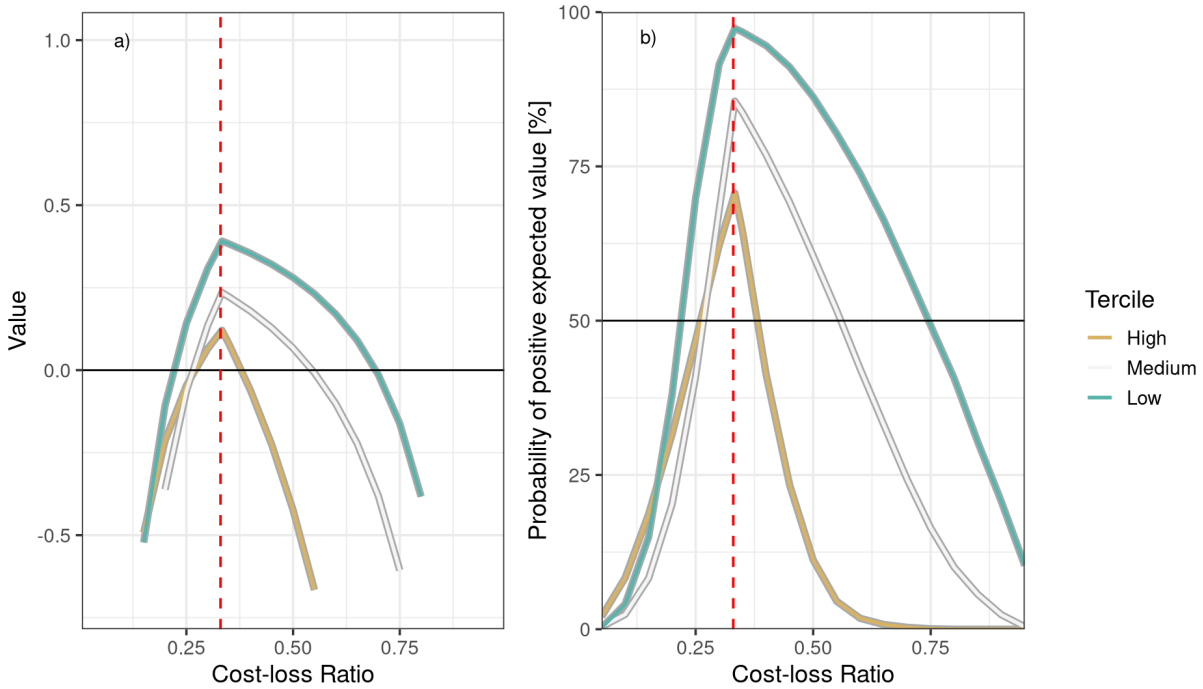


Figure 5 Long-term value of a cost-loss decision model in area 1r, simulated from a multinomial confusion matrix model. a) Tercile divided value given cost-loss ratios. Solid line indicates zero value. Positive value is seen in all terciles, peaking at a cost-loss ratio of 0.33. Most value can be gained by following the low tercile forecast, which corresponds with the highest observed TSS. b) Tercile divided probability of a positive expected value. Calculated from a Bayesian posterior distribution, indicating the probability of drawing a positive value at a given cost-loss ratio. Peak probability is seen at cost-loss ratio of 0.33, where all terciles shows above 65% probability of a positive expected value.

We assessed the value of the forecast for all areas, where the most value can be gained in area 1r (Figure S3 for Area 2r, 3r and 4 value). The cost-loss decision model for Area 1r shows value in all forecast categories, with especially the low recruitment prediction showing the highest value over a range of cost/loss ratios (Figure 5a). All categories peak at a cost-loss ratio of 0.33, as is expected from theoretical analyses of this model (Jolliffe and Stephenson, 2012). We account for the small sample size and propagate the uncertainty that it creates into the forecast value by estimating the probability of a positive expected value for a given cost/loss ratio (Figure 5b): this metric provides decision makers with

302 an indicator when using the forecast will lead to a positive economic return. Here the peak is still seen at
303 a cost-loss ratio of 0.33, where all categories have above 65% probability of a positive expected long-
304 term value. Following the low recruitment forecast for this cost-loss ratio (i.e. 0.33) will result in a 96%
305 probability of positive value from the forecast, but probabilities above 50% are also seen across a wide
306 range of cost-loss ratios. While area 2r, 3r and 4 generally can't provide the same levels of value, area 2r
307 could prove valuable when following the high forecast (Figure S3).

308 **Discussion**

309 Here we present a framework for robustly assessing the skill and value of recruitment predictions. The
310 case study that we have examined, for four sandeel stocks in the North Sea, illustrates several important
311 conceptual points that deserve particular attention.

312 Firstly, we show the importance to assess a forecast system with multiple metrics. While in-sample
313 performance and explanatory metrics are good for finding correlations (and thereby highlighting possible
314 causality), the assessment of predictive skill is quite different and should primarily be shaped by the needs
315 of the forecast user. For example, we identify an overall high forecast accuracy in area 2 (RMSE in Figure
316 3), but the ability to distinguish between the two lower terciles is poor (Figure 4). Stock assessors may
317 focus on the MSESS as a criteria for uptake, while use in industry might be more interested in performance
318 in a specific category (e.g. ability to forecast poor year classes) or long-term economic value. Furthermore,
319 the value of a forecast to users within the same sector (e.g. two different fish processing plants) may differ
320 due to differences in their underlying risk profile (i.e. cost-loss ratio) such that while the forecast system
321 may be advantageous for one user, it may not be of user to another. Understanding the forecast needs of
322 the user is therefore essential to the production of a good forecast (Murphy, 1993; Payne *et al.*, 2017).

323 While the application of the cost-loss model to estimate forecast value has a clear interpretation in a
324 commercial context, it is less clear how relevant this approach is to fisheries management. Here, cost-loss

decision models encapsulate both the costs and losses associated with correctly and incorrectly forecasting recruitment. These ideas can be relevant to fisheries management, as the managers can use this knowledge as the basis of the forecast evaluation, assigning value on e.g. true positives versus false positives. This allows managers and users to understand how the forecast can be incorporated, and how forecasts can and should be used in the management of a given stock. While not an economic gain, the value metric of a forecast can be used to assess and manage stocks sustainably, providing the managers with the tools to properly assess how to incorporate forecasts into decision making.

The framework presented here is generally applicable, allowing recruitment predictability to be assessed on a larger scale. We note that all sandeel stocks considered show no reliance on SSB (Figure S2), indicating environmental and demographic recruitment predictors might be more reliable predictors. This pattern has also been observed in other species (e.g. Baltic Sprat (MacKenzie and Köster, 2004; MacKenzie *et al.*, 2008)), where if the SSB is above a certain limit, the recruitment is not driven by spawning stock, but rather other influences. This study focused on temperature, which has both direct and indirect impacts on recruitment, but identifying other drivers might further improve predictability. For the lesser sandeel additional drivers has been shown to influence recruitment, such as prey concentrations (van Deurs *et al.*, 2009) and circulation patterns (Henriksen *et al.*, 2018). The current framework is able to include such types of drivers, allowing for specialised forecasts for a given species or stock and basing it on already known drivers. However, operational usage does require data availability at the time of forecasting, limiting the options in selection of potential predictors.

Currently, the recruitment modelling framework is based on stock assessement output, which can lead to an increase in uncertainties (Brooks and Deroba, 2015). Here, all models are fitted and predict using the output directly from the stock assessment, and treating it as the primary data source. Using stock assessment output estimated recruitment is a model output, whose values that can change depending on the model structure. Furthermore, retrospective the retrospective forecast analysis is performed with the

available stock assessment output, increasing the risks of retrospective bias in the model assessment. However, the stock assessment output is used for management decisions, and basing the models on the available data is required. Other biomass indicators, such as a recruitment index, is not readily available and cannot easily replace the usage of the assessment data in the case of sandeel recruitment modelling.

The predictability found in the North Sea sandeel show that our understanding of recruitment predictability needs to be re-assessed The potential performance of a forecast system should be evaluated based on predictions instead of explanatory power to properly understand performance when moving out-of-sample. Models need to handle non-stationarity, as ecosystems are not stationary and the recruitment drivers change from year to year. Providing probabilistic predictions provides a more nuanced product, potentially of great use for the industry to incorporate into risk assessments. Recognising and understanding the decision-making need's of the forecast-user is critical to ensuring that the product is ultimately useful, and also shapes the way that both skill and value are assessed. These results therefore open the way for a new paradigm in addressing this long-running, but fundamental question in fisheries management.

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727852 (Blue-Action).

References

Arnott, S. A., and Ruxton, G. D. 2002. Sandeel recruitment in the North Sea: Demographic, climatic and trophic effects. Marine Ecology Progress Series, 238: 199–210.

Banzon, V., Smith, T. M., Chin, T. M., Liu, C., and Hankins, W. 2016. A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. Earth System Science Data, 8: 165–176. https://www.earth-syst-sci-data.net/8/165/2016/.

Brooks, E. N., and 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: 634–641.

Browman, H., Cushing, D., DeBlois, E., Ellertsen, B., Fossum, P., Leggett, W., Myers, R., *et al.* 1995.

Commentaries on current research trends in recruitment studies. Marine Ecology Progress Series, 128: 305–310.

Brun, P., Kiørboe, T., Licandro, P., and Payne, M. R. 2016. The predictive skill of species distribution models for plankton in a changing climate. Global change biology, 22: 3170–3181.

Burnham, K. P., and Anderson, D. R. 2004. Model Selection and Multimodel Inference. Springer New York, New York, NY. http://link.springer.com/10.1007/b97636.

Christensen, A., Jensen, H., Mosegaard, H., St. John, M., and Schrum, C. 2008. Sandeel (Ammodytes marinus) larval transport patterns in the North Sea from an individual-based hydrodynamic egg and larval model. Canadian Journal of Fisheries and Aquatic Sciences, 65: 1498–1511. http://www.nrcresearchpress.com/doi/abs/10.1139/F08-073.

Clark, J. S., and Bjørnstad, O. N. 2004. POPULATION TIME SERIES: PROCESS VARIABILITY, OBSERVATION ERRORS, MISSING VALUES, LAGS, AND HIDDEN STATES. Ecology, 85: 3140–3150. http://doi.wiley.com/10.1890/03-0520.

Dickey-Collas, M., Payne, M. R., Trenkel, V. M., and Nash, R. D. M. 2014a. Food for Thought Hazard warning: model misuse ahead. ICES Journal of Marine Science, 71: 2300–2306. http://icesjms.oxfordjournals.org/content/early/2014/01/09/icesjms.fst215.short.

Dickey-Collas, M., Engelhard, G. H., Rindorf, A., Raab, K., Smout, S., Aarts, G., van Deurs, M., *et al.* 2014b. Ecosystem-based management objectives for the North Sea: riding the forage fish rollercoaster. ICES Journal of Marine Science, 71: 128–142. https://academic.oup.com/icesjms/article/71/1/128/642315.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L. 2013. Seasonal climate predictability and forecasting: Status and prospects. Wiley Interdisciplinary Reviews: Climate Change, 4: 245–268.

Eigaard, O. R., van Deurs, M., Behrens, J. W., Bekkevold, D., Brander, K., Plambech, M., Plet-Hansen, K. S., *et al.* 2014. Prey or predator - Expanding the food web role of sandeel Ammodytes marinus. Marine Ecology Progress Series, 516: 267–273.

Eliasen, K., Reinert, J., Gaard, E., Hansen, B., Jacobsen, J. A., Grønkjær, P., and Christensen, J. T. 2011. Sandeel as a link between primary production and higher trophic levels on the Faroe shelf. Marine Ecology Progress Series, 438: 185–194.

Eveson, J. P., Hobday, A. J., Hartog, J. R., Spillman, C. M., and Rough, K. M. 2015. Seasonal forecasting of tuna habitat in the Great Australian Bight. Fisheries Research, 170: 39–49. Elsevier B.V. http://dx.doi.org/10.1016/j.fishres.2015.05.008.

Fernandes, J. A., Irigoien, X., Lozano, J. A., Inza, I., Goikoetxea, N., and Pérez, A. 2015. Evaluating machine-learning techniques for recruitment forecasting of seven North East Atlantic fish species. Ecological Informatics, 25: 35–42. Elsevier B.V. http://dx.doi.org/10.1016/j.ecoinf.2014.11.004.

Field, J. C., MacCall, A. D., Ralston, S., Love, M. S., and Miller, E. F. 2010. Bocaccionomics: The effectiveness of pre-recruit indices for assessment and management of bocaccio. California Cooperative Oceanic Fisheries Investigations Reports, 51: 77–90.

Haltuch, M. A., and Punt, A. E. 2011. The promises and pitfalls of including decadalscale climate forcing of recruitment in groundfish stock assessment. Canadian Journal of Fisheries and Aquatic Sciences,

68: 912–926.

Haltuch, M. A., Brooks, E. N., Brodziak, J., Devine, J. A., Johnson, K. F., Klibansky, N., Nash, R. D. M., *et al.* 2019. Unraveling the recruitment problem: A review of environmentally-informed forecasting and management strategy evaluation. Fisheries Research, 217: 198–216. Elsevier. https://doi.org/10.1016/j.fishres.2018.12.016.

Hare, J. A., Alexander, M. A., Fooarty, M. J., Williams, E. H., and Scott, J. D. 2010. Forecasting the dynamics of a coastal fishery species using a coupled climate - Population model. Ecological Applications, 20: 452–464.

Hastie, T., and Tibshirani, R. 1986. Generalized Additive Models. Statistical Science, 1: 297–310. http://projecteuclid.org/euclid.ss/1177013604.

Hazen, E. L., Palacios, D. M., Forney, K. A., Howell, E. A., Becker, E., Hoover, A. L., Irvine, L., *et al.* 2017. WhaleWatch : a dynamic management tool for predicting blue whale density in the California Current: 1415–1428.

Hazen, E. L., Scales, K. L., Maxwell, S. M., Briscoe, D. K., Welch, H., Bograd, S. J., Bailey, H., *et al.* 2018. A dynamic ocean management tool to reduce bycatch and support sustainable fisheries. Science Advances, 4: 1–8.

Henriksen, O., Christensen, A., Jonasdottir, S., MacKenzie, B., Nielsen, K., Mosegård, H., and van Deurs, M. 2018. Oceanographic flow regime and fish recruitment: reversed circulation in the North Sea coincides with unusually strong sandeel recruitment. Marine Ecology Progress Series, 607: 187–205. https://www.int-res.com/abstracts/meps/v607/p187-205/.

Hilborn, R., and Walters, C. J. 1992. Quantitative Fisheries Stock Assessment. Springer US, Boston, MA. http://link.springer.com/10.1007/978-1-4615-3598-0.

Hobday, A. J., Spillman, C. M., Eveson, J. P., Hartog, J. R., Zhang, X., and Brodie, S. 2018. A Framework for Combining Seasonal Forecasts and Climate Projections to Aid Risk Management for Fisheries and Aquaculture. Frontiers in Marine Science, 5: 1–9. http://journal.frontiersin.org/article/10.3389/fmars.2018.00137/full.

ICES. 2010. Report of the Benchmark Workshop on Sandeel (WKSAN), 6–10 September 2010, Copenhagen, Denmark. 201 pp.

ICES. 2018. Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG). 29-31 January 2018 and 12-20 March 2018. ICES HQ, Copenhagen, Denmark. ICES CM 2018/ACOM:07.

Jolliffe, I. T., and Stephenson, D. B. 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley & Sons, Ltd, Chichester, UK. http://doi.wiley.com/10.1002/9781119960003.

Köster, F. W., Möllmann, C., Hinrichsen, H. H., Wieland, K., Tomkiewicz, J., Kraus, G., Voss, R., *et al.* 2005. Baltic cod recruitment - The impact of climate variability on key processes. ICES Journal of Marine Science, 62: 1408–1425.

Leggett, W. C., and Deblois, E. 1994. Recruitment in marine fishes: Is it regulated by starvation and predation in the egg and larval stages? Netherlands Journal of Sea Research, 32: 119–134.

Levins, R. 1966. Linked references are available on JSTOR for this article : THE STRATEGY OF MODEL BUILDING IN population biology arises. American Scientist, 54: 421–431.

Lindegren, M., van Deurs, M., MacKenzie, B. R., Worsoe Clausen, L., Christensen, A., and Rindorf, A. 2018. Productivity and recovery of forage fish under climate change and fishing: North Sea sandeel as a case study. Fisheries Oceanography, 27: 212–221.

MacKenzie, B. R., and Köster, F. W. 2004. FISH PRODUCTION AND CLIMATE: SPRAT IN THE BALTIC SEA. Ecology, 85: 784–794. http://doi.wiley.com/10.1890/02-0780.

MacKenzie, B. R., Horbowy, J., and 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: 1334–1341. http://article.pubs.nrc-cnrc.gc.ca/ppv/RPViewDoc?issn=1205-7533&volume=65&issue=7&startPage=1334&ab=y.

Mantzouni, I., and Mackenzie, B. R. 2010. Productivity responses of a widespread marine piscivore, Gadus morhua, to oceanic thermal extremes and trends. Proceedings. Biological sciences / The Royal Society, 277: 1867–74. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2871868&tool=pmcentrez&rendertype=abstract.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., *et al.* 2014. Decadal climate prediction an update from the trenches. Bulletin of the American Meteorological Society, 95: 243–267.

Murphy, A. H. 1969. On the “Ranked Probability Score”. Journal of Applied Meteorology, 8: 988–989. http://journals.ametsoc.org/doi/abs/10.1175/1520-0450%281969%29008%3C0988%3AOTPS%3E2.0.CO%3B2.

Murphy, A. H. 1976a. Decision-Making Models in the Cost-Loss Ratio Situation and Measures of the Value of Probability Forecasts. Monthly Weather Review, 104: 1058–1065. http://journals.ametsoc.org/doi/10.1175/1520-0493(1976)104%3C1058:DMMITC%3E2.0.CO;2.

Murphy, A. H. 1976b. Decision-Making Models in the Cost-Loss Ratio Situation and Measures of the Value of Probability Forecasts. Monthly Weather Review, 104: 1058–1065.

Murphy, A. H., and Winkler, R. L. 1987. A General Framework for Forecast Verification. Monthly Weather Review, 115: 1330–1338. http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%281987%29115%3C1330%3AAGFFFV%3E2.0.CO%3B2.

Murphy, A. H. 1993. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Weather and Forecasting, 8: 281–293.

Myers, R. A. 1998. When Do Environment–recruitment Correlations Work? Reviews in Fish Biology and Fisheries, 8: 285–305. http://www.springerlink.com/index/P76366G6716KN272.pdf (Accessed 29 March 2011).

Payne, M. R., Hobday, A. J., MacKenzie, B. R., Tommasi, D., Dempsey, D. P., Fässler, S. M. M., Haynie, A. C., *et al.* 2017. Lessons from the First Generation of Marine Ecological Forecast Products. Frontiers in Marine Science, 4. http://journal.frontiersin.org/article/10.3389/fmars.2017.00289/full.

Pedersen, E. J., Miller, D. L., Simpson, G. L., and Ross, N. 2019. Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ, 7: e6876.

Pedersen, S. A., Lewy, P., and Wright, P. 1999. Assessments of the lesser sandeel (Ammodytes marinus)

in the North Sea based on revised stock divisions. Fisheries Research, 41: 221–241. https://linkinghub.elsevier.com/retrieve/pii/S0165783699000260.

Platt, T., Fuentes-Yaco, C., and Frank, K. T. 2003. Spring algal bloom and larval fish survival off Nova Scotia. Nature, 423: 398–399.

Pope, E. C. D., Buontempo, C., and Economou, T. 2019. Exploring constraints on the realised value of a forecast-based climate service. Climate Services, 15: 100102. Elsevier. https://doi.org/10.1016/j.cliser.2019.100102.

Ricard, D., Minto, C., Jensen, O. P., and Baum, J. K. 2012. Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. Fish and Fisheries, 13: 380–398.

Richardson, D. S. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society, 126: 649–667. http://doi.wiley.com/10.1002/qj.49712656313.

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., and Webster, M. S. 2010. Population diversity and the portfolio effect in an exploited species. Nature, 465: 609–612. Nature Publishing Group. http://dx.doi.org/10.1038/nature09060.

Shmueli, G. 2009. To Explain or To Predict? Ssrn, 25: 289–310.

Stan Development Team. 2020. Stan Modeling Language Users Guide and Reference Manual, 2.19.0.

Stow, C. A., Jolliff, J., McGillicuddy, D. J., Doney, S. C., Allen, J. I., Friedrichs, M. A. M., Rose, K. A., *et al.* 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems, 76: 4–15. Elsevier B.V. http://dx.doi.org/10.1016/j.jmarsys.2008.03.011.

Subbey, S., Devine, J. A., Schaarschmidt, U., and Nash, R. D. M. 2014. Modelling and forecasting stock–recruitment: current and future perspectives. ICES Journal of Marine Science, 71: 2307–2322. https://academic.oup.com/icesjms/article/71/8/2307/2804451.

Tommasi, D., Stock, C. A., Pegion, K., Vecchi, G. A., Methot, R. D., Alexander, M. A., and Checkley, D. M. 2017a. Improved management of small pelagic fisheries through seasonal climate prediction: Ecological Applications, 27: 378–388.

Tommasi, D., Stock, C. A., Hobday, A. J., Methot, R., Kaplan, I. C., Eveson, J. P., Holsman, K., *et al.* 2017b. Managing living marine resources in a dynamic environment: The role of seasonal to decadal climate forecasts. Progress in Oceanography, 152: 15–49. http://linkinghub.elsevier.com/retrieve/pii/S0079661116301586.

Van Beveren, E., Benoît, H. P., and Duplisea, D. E. 2021. Forecasting fish recruitment in age-structured population models. Fish and Fisheries: 1–14.

van Deurs, M., Van Hal, R., Tomczak, M. T., and Jónasdóttir, S. H. 2009. Recruitment of lesser sandeel Ammodytes marinus in relation to density dependence and zooplankton composition. Marine Ecology Progress Series, 381: 249–258.

Walters, C. J. 1989. Value of Short-Term Forecasts of Recruitment Variation for Harvest Management.

Ward, E. J., Holmes, E. E., Thorson, J. T., and Collen, B. 2014. Complexity is costly: A meta-analysis of parametric and non-parametric methods for short-term population forecasting. Oikos, 123: 652–

661.

Wilks, D. 2011. Statistical Methods in the Atmospheric Sciences. 704 pp. https://linkinghub.elsevier.com/retrieve/pii/B9780123850225000014.