ICES Journal of Marine Science, 2022, **00**, 1–15
DOI: 10.1093/icesjms/fsac043
Original Article

ICES CIEM International Council for the Exploration of the Sea Conseil International pour l'Exploration de la Mer

Evaluation of the skill of length-based indicators to identify stock status and trends

Laurence T. Kell [1,*], Cóilín Minto [2] and Hans D. Gerritsen [3]

[1]Centre for Environmental Policy, Imperial College London, Weeks Building, 16-18 Princes Gardens, London SW7 1NE, UK
[2]Marine and Freshwater Research Centre, Galway-Mayo Institute of Technology, Dublin Road, Galway H91 T8NW, Ireland
[3]Marine Institute, Rinville, Oranmore Co., Galway H91 R673, Ireland

*Corresponding author: tel: +44 7517071190; e-mail: l.kell@imperial.ac.uk.

In data-poor situations, length-based indicators (LBIs) and reference points based on life history parameters have been proposed to classify stocks according to conservation status and yield optimization. Given the variety of potential LBIs, life history traits, and fisheries, it is necessary to evaluate the robustness of length-based advice to ensure that despite uncertainty that management objectives will still be met. Therefore, a simulation procedure was employed where an Operating Model conditioned on life history parameters was used to generate pseudo data. Receiver operator characteristics and the true skill score were then used to screen LBIs based on their ability to identify overfishing and recovery. It was found that LBIs performed better for long-lived species with low individual growth rates, those aimed at ensuring the conservation of mature fish performed better than those aimed at the conservation of immature fish, are better at indicating trends than at quantifying exploitation level, and in general were robust to uncertainty about dynamic processes.

Keywords: data-poor, evaluation, length-based indicators, life history, receiver operator characteristic, screening, simulation, stock assessment, true skill score.

Introduction

The adoption of the voluntary Code of Conduct on Responsible Fishing and the United Nations Fish Stocks Agreement (PA, Garcia, 1996) requires that reference points and management plans are developed for all stocks—not just targeted commercial stocks, but also by-caught, threatened, endangered, and protected species (Sainsbury and Sumaila, 2003). Reference points are used in management plans as targets to maximize surplus production and as limits to minimize the risk of depleting a resource to a level where productivity is compromised. Reference points must integrate dynamic processes such as growth, fecundity, recruitment, mortality, and connectivity into indices for exploitation level and spawning reproductive potential. An example of a target reference point is the fishing mortality (F) that will produce the maximum sustainable yield (F_{MSY}), commonly defined as the fishing mortality with a given fishing pattern and current environmental conditions that gives the long-term maximum yield. To ensure sustainability requires preventing a stock from becoming overfished, so that there is a low probability of compromising productivity. Therefore, many fishery management bodies also define a limit reference point, e.g. B_{lim}, at a biomass at which recruitment or productivity is impaired (Restrepo and Powers, 1999). When assessing stocks, it is also important to consider trends as well as state since a stock at a target biomass may be declining due to overfishing, while, a depleted stock may be recovering due to management action (Hilborn, 2020).

However, half of the fisheries worldwide exploit resources without formal stock assessments (Hilborn *et al.*, 2020). These are termed data-limited, data-poor, information-poor, or capacity-limited (Dowling *et al.*, 2015). For example: although the United Nations' Food and Agriculture Organization (FAO) landings database includes over 20 000 individual catch histories by FAO region, country, and taxon, the RAM Legacy Stock Assessment Database (www.ramlegacy.org), which includes most of the publicly available stock assessments contains only 1200 assessments (Ovando *et al.*, 2021). Therefore, status, productivity, and exploitation levels of many stocks and species are largely unknown (Thorson *et al.*, 2015). In addition to the risk of overexploitation, the lack of formal assessments may hamper progress towards the Ecosystem Approach to Fisheries Management (EAFM), which requires as a first step the assessment of the impacts on non-target species, trophic structure, and habitat (Hilborn, 2011).

In data-poor situations, life history parameters such as maximum size and size at first maturity have been used as proxies for productivity (Roff, 1984; Jensen, 1996; Caddy, 1998; Reynolds *et al.*, 2001; Denney *et al.*, 2002). For example, ICES has implemented a framework for data-poor stocks that uses length-based indicators (LBIs) and life history parameters to classify stocks according to conservation and sustainability status, and yield optimization. Table 1 summarizes the LBIs, reference points, and reference levels, proposed in ICES (2015).

Some indicators aim to prevent growth overfishing, for example a high proportion of fish should be allowed to spawn at least once before they are caught. To ensure this, the ratio between the 25th percentile of the length distribution ($L_{25\%}$) and the length at 50% maturity (L_{50}) should be greater than 1. To ensure the conservation of large individuals, the mean length of the largest 5% of the length distribution ($L_{max5\%}$) should be greater than 0.8 L_{∞}. Miethe *et al.* (2019) further developed this approach by deriving reference points consistent

Received: September 2, 2021. **Revised:** February 23, 2022. **Accepted:** March 2, 2022

Table 1. LBIs and reference levels.

Indicator	Calculation	Reference point	Indicator ratio	Expected value	Property
L_c	Length at 50% of modal abundance	L_{50}	L_c/L_{50}	> 1	Conservation (immatures)
$L_{25\%}$	25th percentile of length distribution	L_{50}	$L_{25\%}/L_{50}$	> 1	Conservation (immatures)
L_{maxY}	Length class with maximum biomass in catch	$L_{opt} = 2/3\ L_\infty$	L_{maxY}/L_{opt}	$= 1$	Optimal yield
L_{bar}	Mean length	L_{50}	L_{bar}/L_{50}	> 1	Optimal yield
L_{mean}	Mean length of individuals	$L_{F=M} = 0.75L_c + 0.25L_\infty$	$L_{mean}/L_{F=M}$	$= 1$	Optimal yield
$L_{max5\%}$	Mean length of largest 5%	L_∞	$L_{max5\%}/L_\infty$	> 0.8	Conservation (large individuals)
$L_{95\%}$	95th percentile	L_∞	$L_{95\%}/L_\infty$	> 0.8	Conservation (large individuals)
P_{mega}	Proportion of individuals above L_{opt} + 10%, L_{opt} is estimated from L_∞.		P_{mega}	> 0.3	Conservation (large individuals)

with a spawning potential ratio of 40%, which, if estimates of natural mortality, maturity, and growth (M, L_{50}, L_∞, k, and CV_{L_∞}) are available, can be tailored by stock.

Indicators can also be used to provide advice as part of empirical rules, for example Fischer *et al.* (2020a) incorporated the LBI L_c as a proxy for F: F_{MSY} into a harvest control rule. Where L_c is the first length class having at least 50% of the mode in the observed catch–length frequency. The reference point is the length at maximum sustainable yield ($L_{F=M}$); assuming $F_{MSY} = M$, as proposed by Beverton and Holt (1993), using the simplification that $M/k = 1.5$, then $L_{F=M} = 0.75L_c + 0.25L_\infty$.

To be an effective management tool, a LBI should be robust so that it still functions despite uncertainty (Radatz *et al.*, 1990; Zhou *et al.*, 1996). Indicators should also be reliable and stable. An indicator is reliable if it provides an accurate result despite uncertainty, and is stable if, despite random error, similar results are produced across multiple trials. Therefore, to evaluate the robustness of LBIs, a simulation procedure was employed where an Operating Model conditioned on life history parameters was used to generate pseudo data using Monte Carlo simulation and an Observation Error Model. LBIs, proxy reference points, and reference levels were then compared to the actual (simulated) state of the resource and screened for their ability to classify stock status relative to reference points.

Material and methods

Case study stocks considered represent a range of fisheries and life history types to allow comparison across taxa, and do not represent any specific stocks. Species selected were sprat (*Sprattus sprattus sprattus*), brill (*Scophthalmus rhombus*), turbot (*Psetta maxima*), pollack (*Pollachius pollachius*), and thornback ray (*Raja clavata*).

An age-structured simulation model was conditioned using life history theory to provide a theoretical basis for developing hypotheses about population dynamics. The parameters were: growth model parameters (k, L_∞, and t_0; Von Bertalanffy, 1957); a and b of the length–weight relationship; and the length at which 50% were mature (L_{50}). Natural mortality-at-age was modelled as a function of length (Gislason *et al.*, 2010) and spawning stock biomass (SSB) was used as a proxy for stock reproductive potential (SRP; Trippel, 1999). It was assumed that fecundity is proportional to the weight-at-age of the sexually mature portion of the population irrespective of the demographic composition of adults (Murawski *et al.*, 2001), and that processes such as sexual maturity are simple functions of age and independent of sex (Matsuda *et al.*, 1996).

Life history parameters were extracted from FishBase (www.fishbase.org), and Figure 1 summarizes L_∞, k, L_{50}, and b; with species ordered by k. There are only a few observations for pollack and brill, which implies high uncertainty, since data in FishBase are often based on small sample sizes, have limited coverage, and life history parameters (e.g. maturity and growth) and generally come from different studies. There are clear relationships, both between and within species, as L_∞ is inversely correlated with k, and L_{50} is correlated with L_∞. Brill has a value of k, similar to sprat, but its value of L_∞ is closer to that of turbot, while sprat has a large variation in the relationship between length and weight (i.e. b in the length–weight relationship $W = aL^b$).

To create an Operating Model, the FLR (Kell *et al.*, 2007) packages mydas and FLife were used, (see Supplementary Materials). First, an equilibrium per-recruit model was parameterized for growth, maturity, and natural mortality-at-age; where the means of the available values for each parameter by species were used. The per-recruit model then was combined with a stock–recruit relationship (Beverton and Holt, 1993). To model uncertainty about parameters and relationships, a number of scenarios were considered for each species (see Table 2). Virgin biomass was set at a constant value across all stocks and scenarios, as results are presented in terms of exploitation level and relative stock size.

Historical exploitation was simulated for stocks that were initially lightly exploited before fishing mortality (F) gradually increased until the stock became over-fished, after which a recovery plan was implemented to bring fishing down to 70% of F_{MSY} (Figure 2). This exploitation history provides contrasting periods of under-, over-, and maximally sustainable exploitation. Inter-annual variability in yield and SSB depends largely on k, e.g. sprat shows the largest and ray the lowest variations in yield, and SSB.

Scenarios

Even for data-rich stock assessments there is often large uncertainty about the dynamics (i.e. model uncertainty; Punt, 2008), for example estimates of stock status are highly sensitive to assumptions about natural mortality-at-age (Jiao *et al.*, 2012), vulnerability of age classes to the fisheries (Brooks *et al.*, 2009), and the relationship between stock and

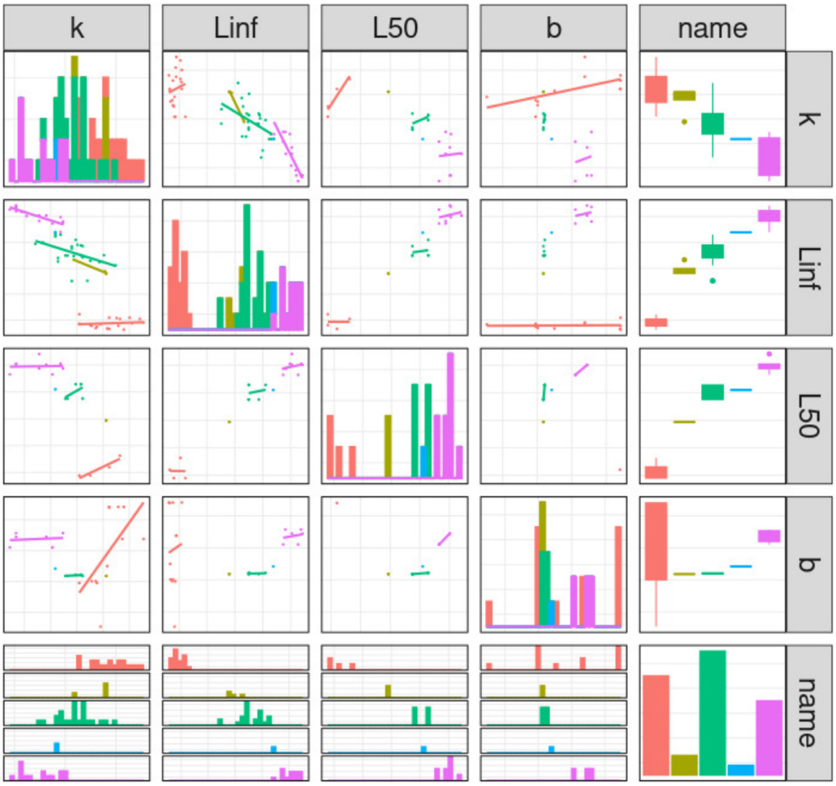

Figure 1. Life history parameters extracted from FishBase for the case study species; the bottom right-hand panel shows the number of observations available. Species are ordered from left to right in order of high to low *k*, i.e. sprat, brill, turbot, and pollack ray.

Table 2. Operating Model scenarios; base case values in bold; *N* is the number of levels per factor and Π is the cumulative number of scenarios.

Factor	*N*	Π	Levels
Steepness	2	2	**0.9**, 0.7
Recruitment deviates	3	6	**0.3**, 0.5, 0.3+AR
Natural mortality (M)	2	12	**Gislason**, Constant
Length sample size	2	24	**500**, 250
Selectivity	2	48	**Maturity**, Dome

recruitment (e.g. Cury *et al.*, 2014). Therefore, a base case was defined and scenarios developed representing the main sources of uncertainty (Table 2); namely the steepness of the stock–recruitment relationship, recruitment variability, natural mortality, selection pattern, and sample size (Boorman and Sefton,1997; Ono *et al.*, 2015).

Selectivity depends on the vulnerability of individuals to fishing, and is typically either asymptotic or dome-shaped. The former indicates an initial increase with age or size followed by a levelling off, while in the later case selectivity-at-age declines. Selectivity will differ between fisheries depending on gear characteristics, when and where the fishery operates, and the biology of the species. For example, estimates for species captured with gill nets or hooks will be biased if the model incorrectly assumes logistic selectivity because the missing large fish will be assumed to have been caught, whereas they could just be missed by the gear. Fisheries selectivity was, therefore, modelled as a double normal, as this allowed both asymptotic and dome-shaped selectivity to be simulated. Logistic selectivity in the base case was based on the maturity ogive, so that *MSY* reference points are comparable across case studies. The consequences of shifting the selection patterns is well-known, since if you fish below or above L_{opt} (the length at which a cohort attains its maximum biomass) you reduce $_{MSY}$, while if you take fish before L_{50}, F_{MSY} is reduced and B_{MSY} is increased as older fish need to be conserved. Therefore, scenarios were not case-specific

An additional scenario was modelled where the ages greater than 0 were sampled equal to their abundance to provide a benchmark against which the impact of the assumed selection pattern could be assessed, this is referred to as the fishery independent survey. Although, even surveys have biases, due to their design.

Observation error model

LBIs may be biased and have poor precision due to uncertainty about life history parameters, lags between exploitation levels and changes in fishery selection pattern, variability in year–class strength, and biased sampling. Therefore, length distributions were derived from the Operating Model catch and stock-at-age by applying an inverse age–length key. The inverse age–length key was based on the von Bertalanffy growth curve for each stock, and variation in length-at-age was included by applying a normal distribution to the expected length at age with a *C.V.* of 10%. Sampling was

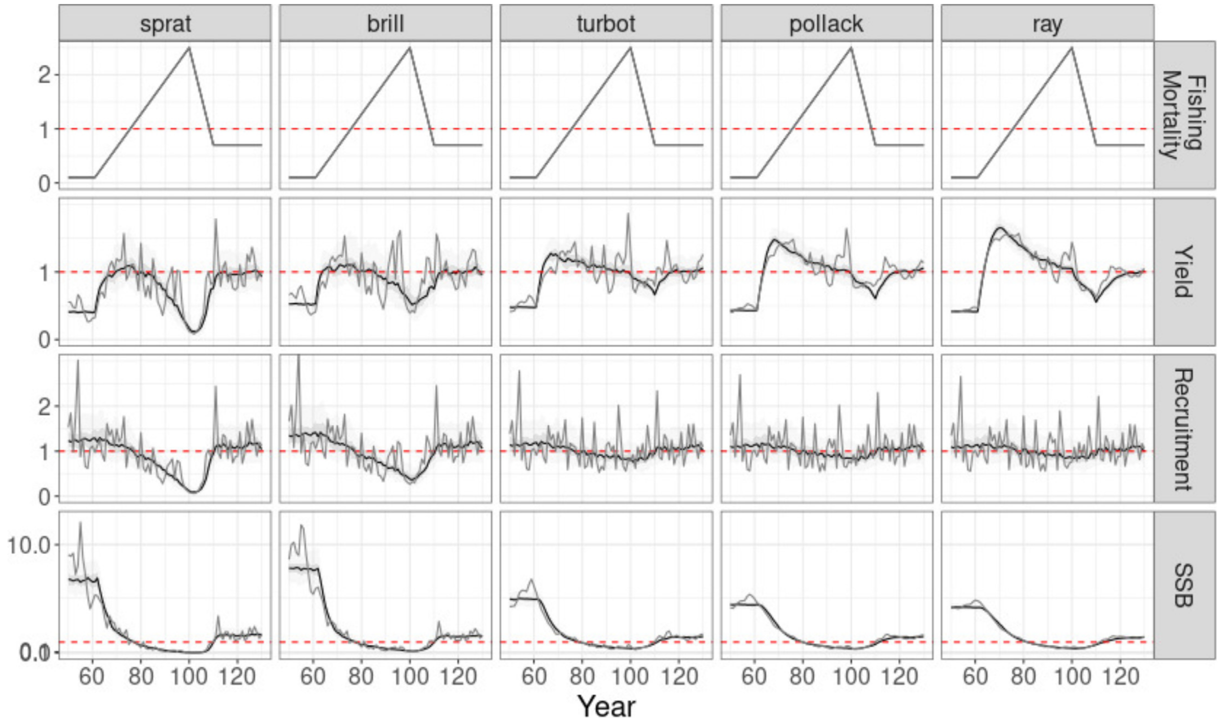

Figure 2. Base case Operating Models, showing fishing mortality, yield, recruitment, and SSB relative to *MSY* reference points (dashed line); fishing was initially low then increased to 2.5 times F_{MSY}, following which a recovery plan was implemented to reduce *F* to 70% of F_{MSY}. The stocks are ordered by *k*. The grey line is an individual Monte Carlo realization.

performed randomly across the catch and the stock proportional to the frequency of an age class for a given sample size. The inverse age–length key was then used to generate a random length for each individual, which were then combined into a length frequency distribution.

Indicators

Empirical indicators were calculated from the length frequency distributions (Table 1). There are three elements in making an indicator operational, the indicator itself, a reference point, and a reference level. For example, for an indicator based on $L_{max5\%}$ the reference level is L_∞ and if the ratio is less than 0.8 the stock is considered to be overfished.

For the conservation of immature fish; LBIs are based on the left-hand limb, i.e. lower percentiles of the length distribution and include: $L_{25\%}$ (the 25th percentile of the length distribution); and L_c (the length at 50% of modal abundance). Indicators based on central tendencies are proxies for F_{MSY} and include: L_{mean} (the mean length of individuals $> L_c$); L_{maxY} (the length class with maximum biomass in catch); and L_{bar} (the mean length). Those based on the right-hand limb and the upper percentiles are for the conservation of larger individuals and include: $L_{max5\%}$ (the mean length of largest 5%); $L_{95\%}$ (the 95th percentile); and P_{mega} (the proportion of individuals above $L_{opt} + 10\%$) where L_{opt} is estimated as $2/3L_\infty$ or $L_\infty \frac{3}{3+M/k}$ as in this study, when the life history parameters are known. A proxy for F_{MSY} is the length at MSY ($L_{F=M}$) proposed by Beverton and Holt (1993), under the assumption that $F = M$, and using the simplification that $M/k = 1.5$ calculated as: $L_{F=M} = 0.75L_c + 0.25L_\infty$.

Indicators are generally based on commercial catches (fishery-dependent) as they represent how fishing mortality is exerted. However, they can also be derived from surveys (e.g. Karnauskas *et al.*, 2011) to help monitor trends as part of the EAFM. We, therefore, also simulated perfect survey data by sampling the stock from ages 1 onwards with full selectivity in the middle of the year. However, even survey data are likely to have a selection pattern, as samples are generally collected with fishing gear at a particular time of year and place. Therefore, the assumptions made in generating the fishery-independent samples likely to be violated to some extent.

Receiver operating characteristics

There are two main questions to be asked when choosing LBIs, namely can a combination of indicator, reference point, and reference level correctly classify a stock, e.g. as being overfished; or can an indicator be used to rank stocks or identify trends in stock status, i.e. should some stocks be assigned higher priority for management intervention in a risk assessment or are things getting better or worse?

For a particular stock and LBI, the best discriminate threshold, i.e. the ratio of the indicator to the reference level, for classifying overfishing, is unlikely to be the one listed in Table 1. This is due to variations in stock and fishery characteristics and uncertainty about the assumptions made. For example, in the case of $L_{max5\%}$ and L_∞, the ratio with the best classification skill may not be 0.8. High random, e.g. measurement, error may also lead to poor classification skill. We, therefore, calculate the true positive rate (TPR, i.e. sensitivity), and the true negative rate (TNR, i.e. specificity). Sensitivity ($\frac{TP}{TP+FN}$) measures the ability of a test to identify positive cases, i.e. the proportion of positives that are correctly identified, while specificity ($\frac{TN}{TN+FP}$) measures the proportion of negatives that are correctly identified. This allows the true skill score (TSS) to be calculated, i.e. TSS = TPR + TNR - 1. A perfect prediction would receive a score of 1, random predictions receive a

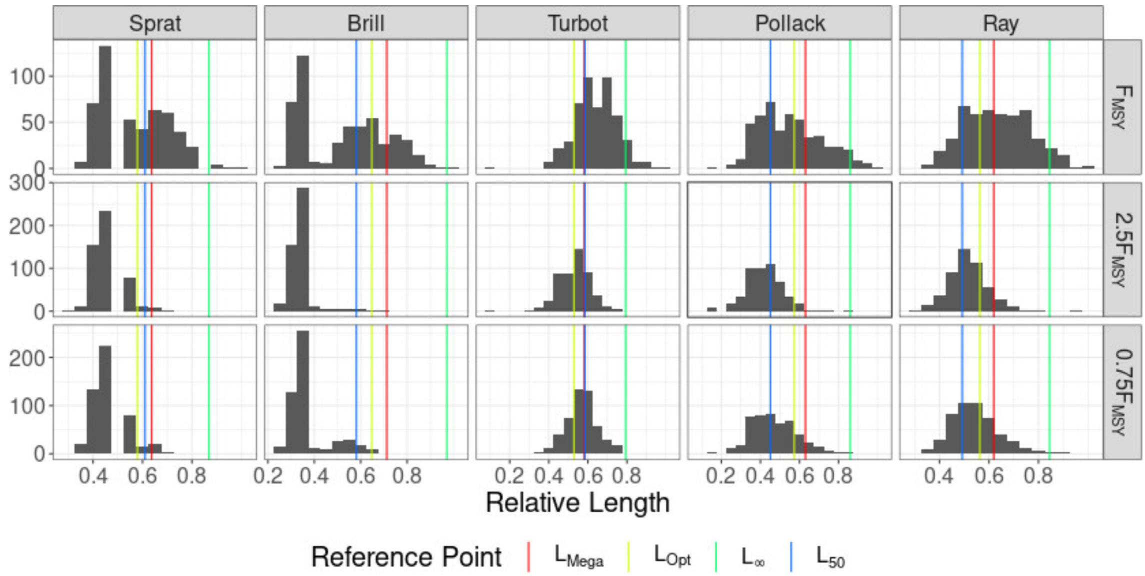

Figure 3. Simulated length samples for three periods where fishing mortality first reached F_{MSY} (year 80), was high at $2.5F_{MSY}$ (year 100), and at 70% of F_{MSY} (year 120), the vertical lines show selected reference points; *x*-axis is length relative to L_{∞}.

score of 0, and predictions inferior to random ones receive a negative score.

Receiver operating characteristic (ROC) curves (Green *et al.*, 1966) can be used to estimate the ability of LBIs to assess status. ROC curves were constructed by sorting the values of F/F_{MSY}, with the highest values first, from the Operating Model and then comparing these to each LBI. The cumulative TPR and TNR are then calculated for the ordered observed outcomes, and the TPR is plotted against the false positive rate (FPR = 1 − TNR) for the different observed indicator and reference ratios, i.e. the potential threshold settings. ROC curves can be thought of as a plot of the power as a function of the Type I Error of the decision rule, and so provides a tool to select the best candidate indicators. This also allows tuning, i.e. calibration, by choosing a reference level that has the best classification skill, and allows the bias in the standard reference points and levels to be evaluated.

The ROC curve is a probability curve, and the area under the curve (AUC) is an important metric for measuring performance. For example, a coin toss would produce a curve that fell along the $y = x$ line and the AUC would be equal to 0.5. The closer the AUC is to 1 the better an indicator is at ranking. The ROC curve can also be used to graphically identify the performance of a choice of indicator ratio (i.e. discriminant threshold): since the best reference points have the shortest euclidean distance between the top left-hand corner (TPR = 1, FPR = 0) and the corresponding point on the curve.

Risks are also asymmetric, i.e. the risk of indicating overfishing is occurring when the stock is sustainably exploited is not the same as the risk of failing to identify overfishing. It may be desirable, therefore, to adjust the threshold to increase or decrease the sensitivity to false positives or false negatives. While some indicators may perform better at identifying the start of overfishing than recovery, and *vice versa*.

Results

Examples of length samples generated by the Observation Error Model are shown for the base case in Figure 3; these are for three time periods corresponding to the initial period when fishing mortality first reached the F_{MSY} level, the overfishing ($2.5F_{MSY}$), and then the recovery ($0.7F_{MSY}$) period. The accumulation of length classes in the right-hand limb is affected by growth and natural mortality, it is, therefore, expected that the performance of LBIs, will vary on a case-specific basis. For sprat and brill, both fast-growing species with high k, there is less overlap between lengths at the early ages, and so bimodal distributions are seen. Exploitation history influences the age and length structure, as there are less large length classes in a recovering stock than a declining stock, although fishing mortality is less. As F increases, the modes of the distributions shift to the left and the decline in the right-hand limb of the length distributions becomes steeper. When F is reduced, the opposite occurs, although the relative abundance of larger individuals does not immediately recover to earlier levels. More contrast is seen in the tails of the distribution than the mode.

The LBIs for the base case (defined in Table 2) are summarized in Figures 4 and 5 for fishery dependent (catch samples) and independent (survey samples) length data. The indicators, summarized by their medians and interquartile ranges, vary with exploitation level. The trend in an indicator should be the inverse of the trends in F/F_{MSY}. Stocks are ordered by the von Bertalanffy growth parameter k and the indicators by the percentile from which they are derived. L_c and $L_{25\%}$ are based on the lower percentiles, L_{maxY}, L_{bar}, and L_{mean} on the central tendency; and $L_{max5\%}$, $L_{95\%}$, and P_{mega} on the upper percentiles. Patterns are seen by stock and indicator. There are also differences between the fishey-dependent and independent indicators, and between overfishing and recovery.

The indicators, L_c and $L_{25\%}$ for the conservation of immature, are primarily a check for the selection pattern and whether growth overfishing is occurring. However, for turbot, pollack, and ray they do show a relationship with F, suggesting that they can indicate recruitment overfishing. If these indicators are below L_{50}, this shows that immature individuals are being caught, which is not a problem if fishing mortality is low.

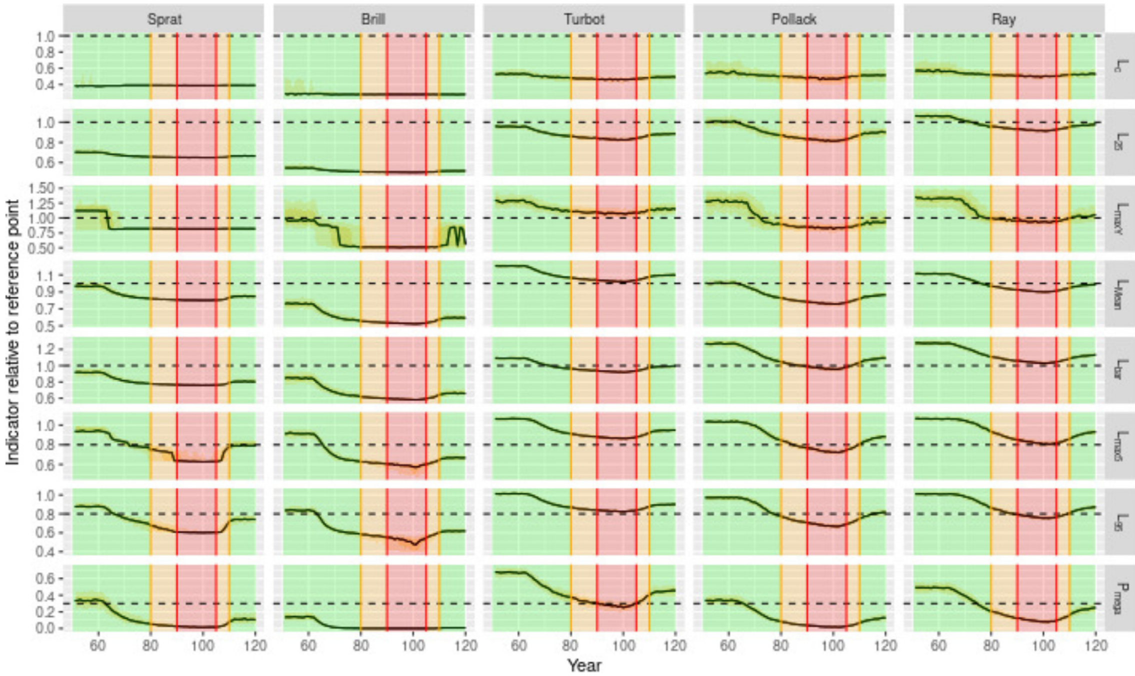

Figure 4. LBIs derived from the fishery dependent length samples for the base case; the coloured regions indicate exploitation levels relative to F_{MSY}, $F < F_{MSY}$ (green), $F_{MSY} < F < 1.5F_{MSY}$ (yellow), and $F \geq 1.5F_{MSY}$ (red). The dashed lines indicate the reference levels in Table 1.

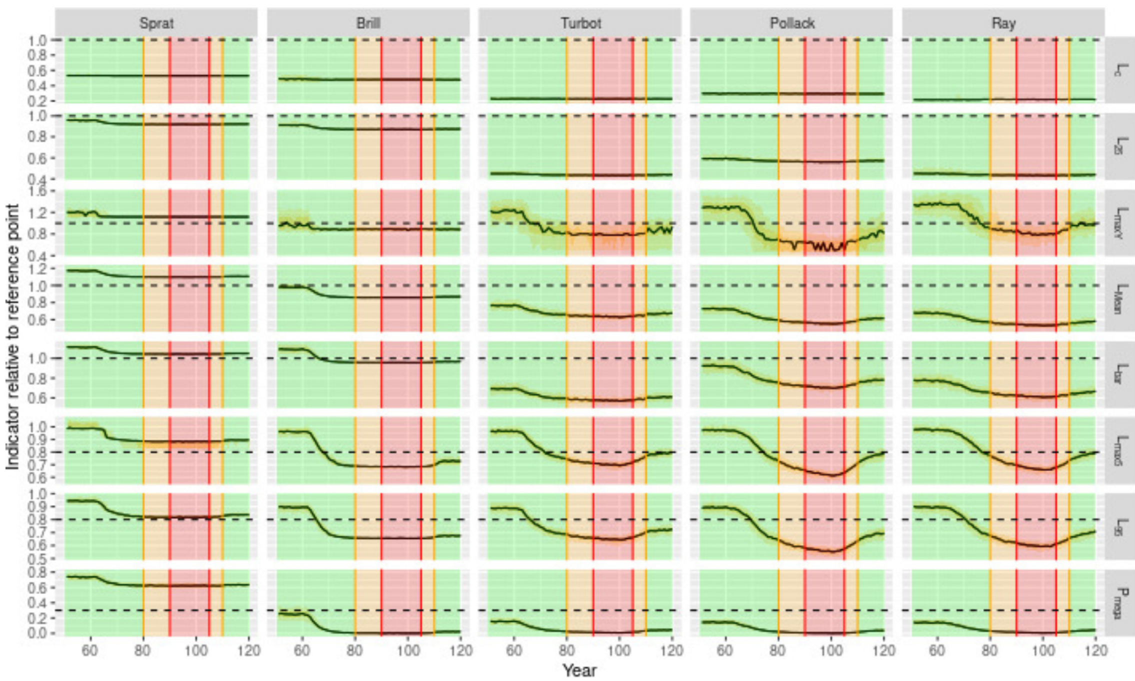

Figure 5. LBIs derived from the fishery independent length samples for the base case; the coloured regions indicate exploitation levels relative to F_{MSY}, $F < F_{MSY}$ (green), $F_{MSY} < F < 1.5F_{MSY}$ (yellow), and $F \geq 1.5F_{MSY}$ (red); The dashed lines indicate the reference levels in Table 1.

More contrast was seen for L_{maxY}, L_{mean}, and L_{95}, particular for pollack and ray, with low k and high L_∞ The discrimination threshold also depends on k, and L_{maxY} showed large inter-annual variation. $L_{max5\%}$ and $L_{95\%}$, based on the upper percentiles, showed good classification skill particularly for stocks with low k, and hence low natural mortality and a range of year–classes in the population. For LBIs based on catch, more contrast was seen, and the reference levels performed better. This is because fishing occurs mainly on mature age classes, and so smaller individuals make less contribution to the length indicators derived from the catch.

Indicators are summarized as ROC curves in Figures 6 and 7 for the overfishing and recovery periods, respectively. The points on the ROC curves indicate the reference levels. The

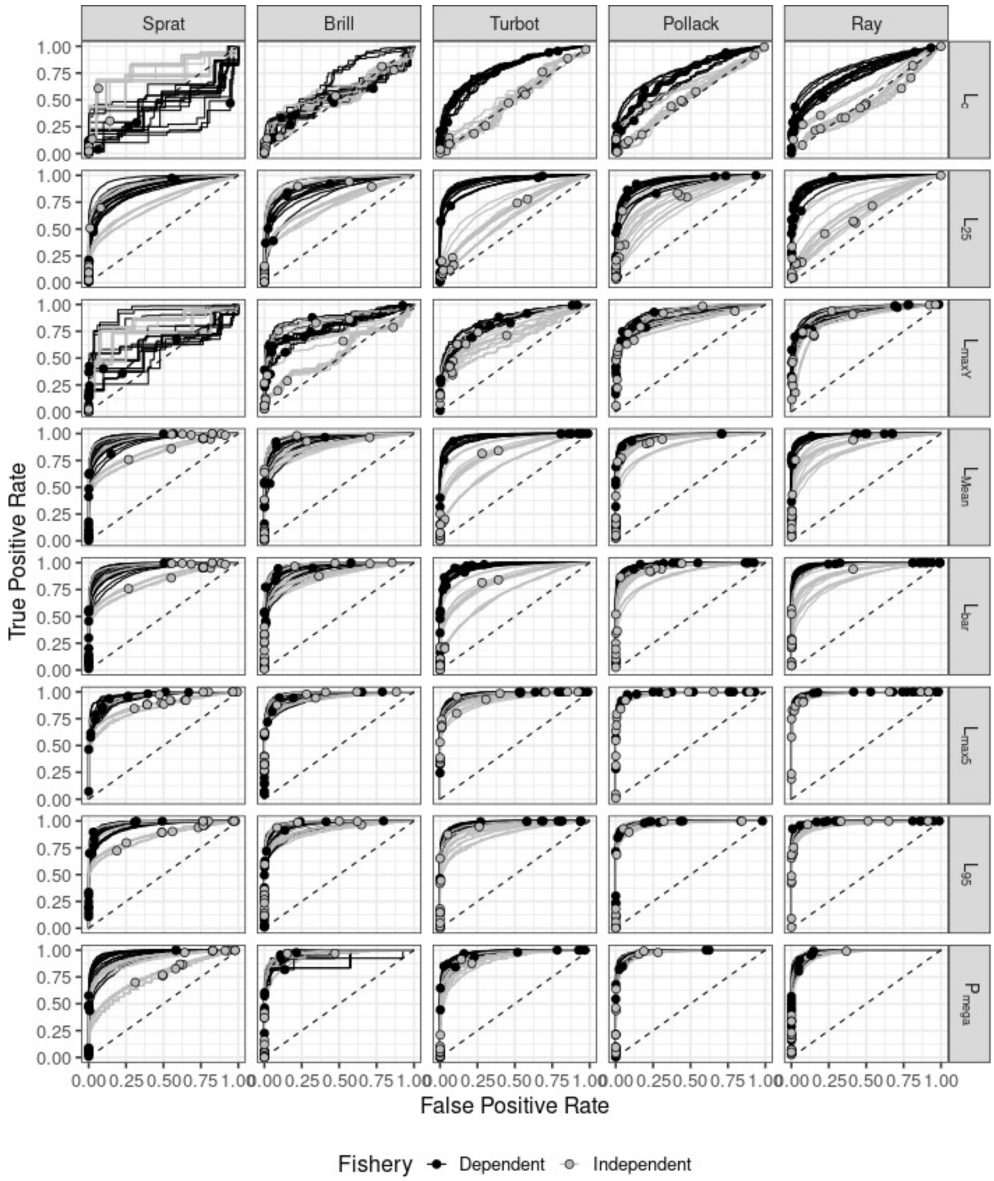

Figure 6. ROC curves for overfishing period, each line represents a scenario that was simulated using a Monte Carlo procedure to obtain the line. The points correspond to the reference level, and the dashed line is the $y = x$ line.

plots confirm the summary for the base case, i.e. the indicators based on the upper percentiles perform well, as the AUC is high, but classification skill is low as points tend to fall at the ends of the curve, rather than being close to TPR = 1 and FPR = 0. Fishery-dependent indicators perform better than those based on fishery-independent data. Additional features are that some indicators are more robust to uncertainty than others, as the curves (corresponding to the scenarios) show less variation. It is also easier to detect overfishing than recovery, and performance can be improved if the reference levels, the discrimination threshold, is varied on a stock by stock basis, i.e. tuned. Patterns in bias of a reference level (point on the curve) and precision (AUC) are seen across the indicators and stocks related to k, which decreases from sprat to ray. For example, as k decreases performance improves, particularly for the LBIs based on upper percentiles, e.g. those where the length distributions have many age classes.

The true skill statistic is used to summarize the ability to classify status in Figure 8. Results are presented for the reference levels (recommended by ICES), and the best obtainable by calibration, where the reference level or discriminant threshold is optimized on a case-specific basis. The fishery-dependent indicators perform better than the fishery-independent ones, and performance varied depending on stock and whether a stock was being overfished or recovering. For example, for sprat L_{mean} performed best for detecting recovery but $L_{max5\%}$ for detecting overfishing, while for turbot L_{bar} performed the best. Performance, therefore, can be improved by developing case-specific indicators. Also, calibration greatly improved skill, however, since calibration was done by scenario this requires there to be no uncertainty about the true value of natural mortality, steepness, and selection pattern.

The areas under the curve are summarized in Figure 9, by boxplots that combine the scenarios. For perfect ranking, skill

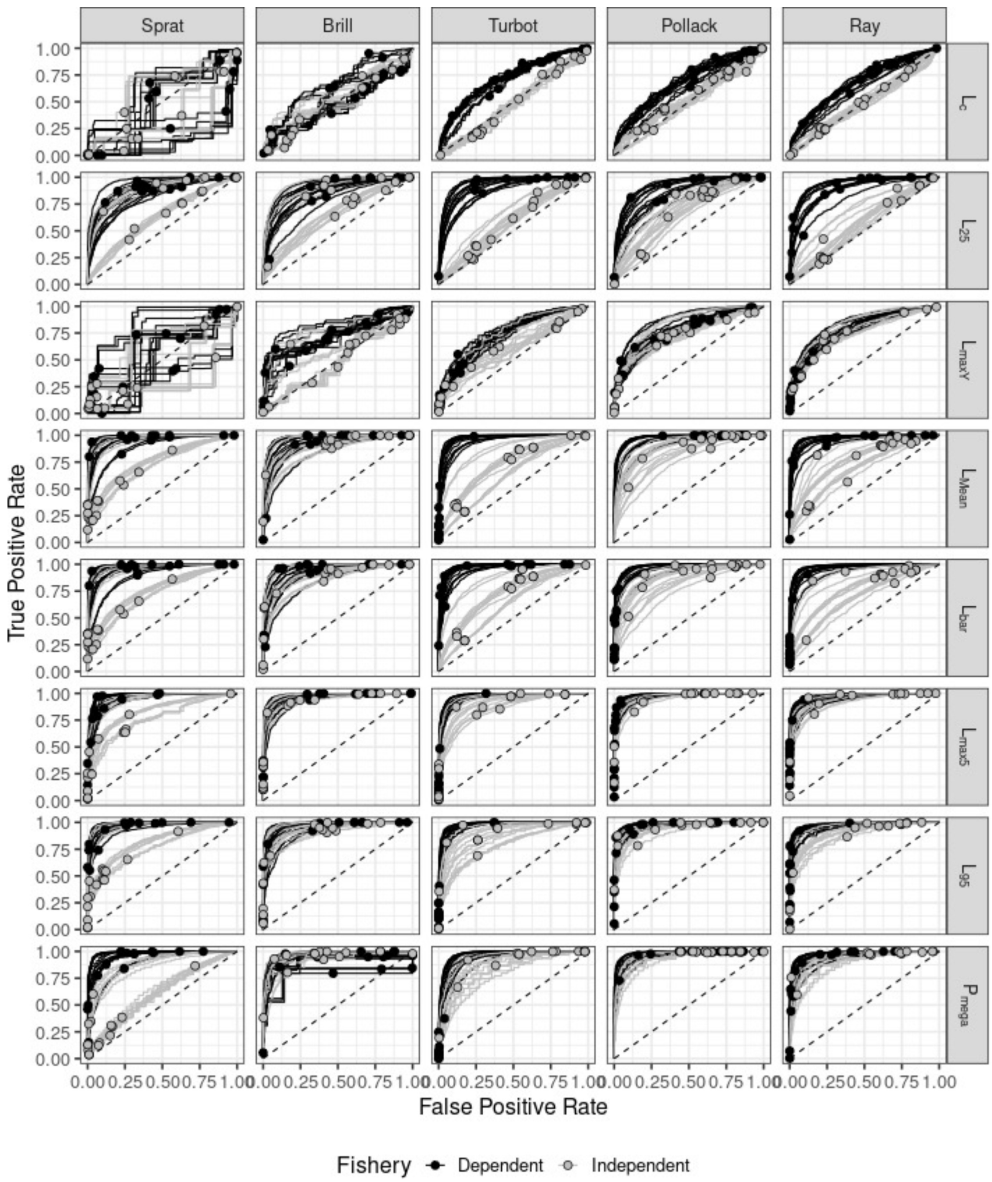

Figure 7. ROC curves for recovery period, each line represents a scenario that was simulated using a Monte Carlo procedure to obtain the line. The points correspond to the reference level, and the dashed line is the $y = x$ line.

values should be close to 1, values close to 0.5 show the indicator is no better than a coin toss. The robustness of a LBI to uncertainty can be inferred by the location and width of a boxplot, since regardless of the assumed values of steepness, natural mortality, or selection pattern if the AUC will be close to 1. Indicators perform better for low-k stocks and for LBIs based on upper quantiles.

TSS (TPR − FPR) for the reference levels and the areas under the curve are compared for stocks and indicators by scenario in Figures 10 and 11 for over-fishing and recovery, respectively. A robust indicator should have high TSS for all scenarios, for example, L_{Mean} and P_{mega} vary little by scenario for sprat under recovery. If the TSS depends on the scenario, indicator, and stock, and whether a stock is undergoing recovery or being overfished, then it will be difficult to specify the reference level required to classify status. However, it may still be able to indicate trends if the AUC is high across scenarios. For example, although the TSS varies for L_{mean}, L_{bar}, $L_{max5\%}$, and L_{95} all have high areas under the curve. P_{mega} also has a high AUC, apart from Brill.

It is relatively easier to increase sample size or to estimate selection pattern than to estimate steepness or M. Therefore, a robust indicator will be little affected by the assumed value of steepness or M. For example, increasing sample size increases skill for L_{95}, while for L_{bar} if M is constant at age then skill is increased for sprat, brill, and turbot (species with high k), while skill is increased for pollack and ray if selectivity is dome-shaped. This shows that skill can be improved by developing case specific indicators.

Discussion

The objective of this work was to evaluate the ability of LBIs to identify overfishing and their robustness to uncertainty of LBIs, rather than to develop stock-specific advice. Therefore, scenarios were selected that provide contrasting hypotheses.

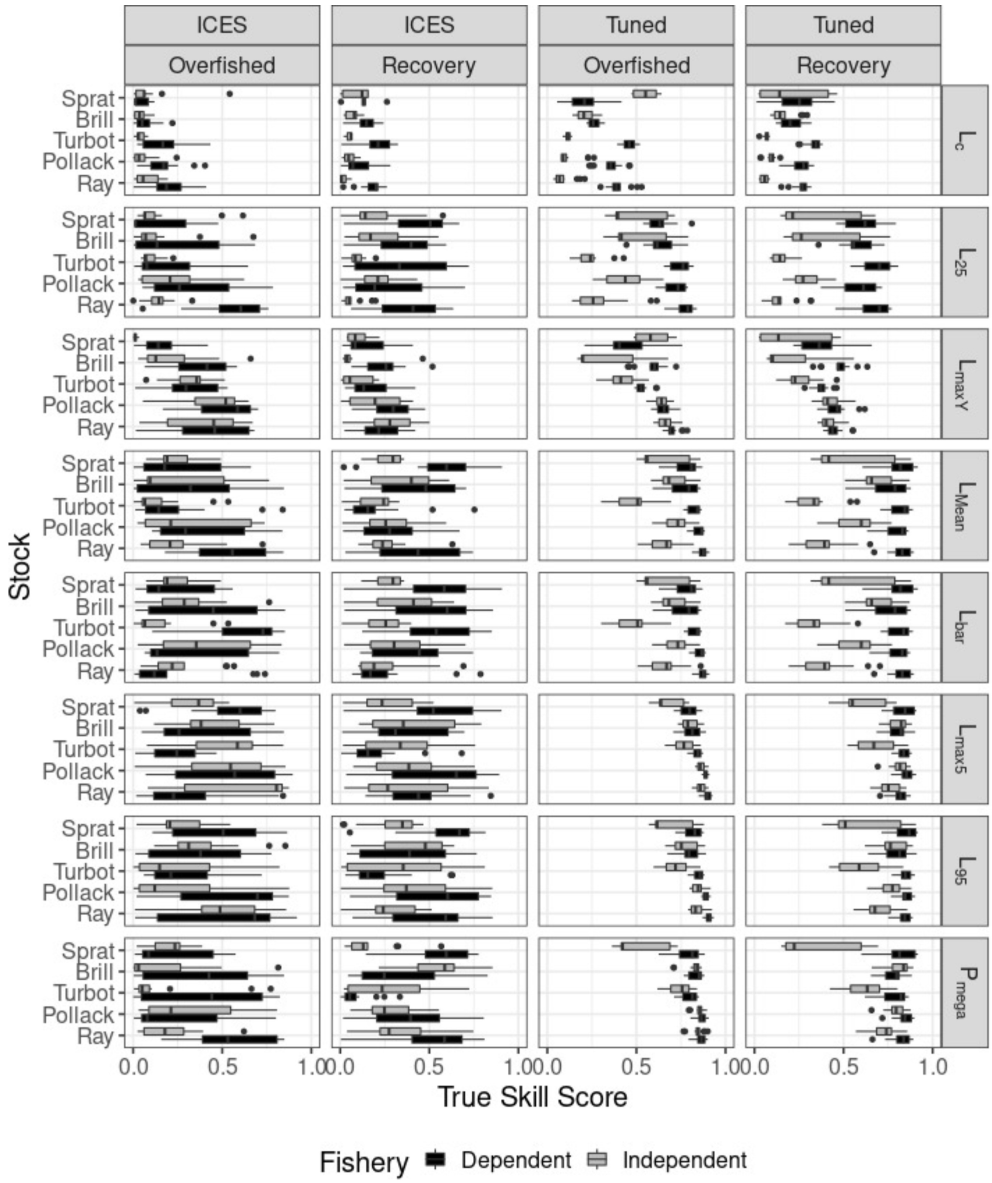

Figure 8. TSS (TPR − FPR) for the reference levels in Table 2 (ICES) and the best reference level (tuned); the boxplot hinges correspond to the interquartile range, while the whiskers extends from the hinges to the values that are within 1.5 times the interquartile.

Although this limited the number of scenarios considered, the number (90) was comparable to other data-limited studies. (e.g. Carruthers *et al.*, 2016; Pons *et al.*, 2019, 2020; Fischer *et al.*, 2020a; Mildenberger *et al.*, 2021). Provision of case-specific advice requires agreement on the plausibility of alternative hypotheses. A highly plausible scenario is one that fits prior knowledge, with many sources of corroboration, without the complexity of explanation, and with minimal conjecture (Connell and Keane, 2006). Plausibility may be determined formally, based on a statistical approach to determine whether a system equivalent to the model generated the data or specified based on expert judgement. This was beyond the scope of this study, but our approach could be developed to allow case-specific applications to be developed.

LBIs based on the right-hand limb of the length frequency distribution, i.e. L_{bar}, $L_{max5\%}$, L_{95}, and P_{mega}, were able to identify trends in fishing mortality and were also robust to uncertainty about the dynamics, e.g. about natural mortality or the steepness of the stock relationship. This is because the main assumption of LBIs is that individuals accumulate into the larger length classes, depending on the mortality (natural and fishing) and the growth rate (k). As long as natural mortality and k do not vary over time, then changes in these indicators will depend on fishing mortality, vulnerability to fishing, and variability in recruitment. If variability in vulnerability and natural mortality is random, then indicators will have skill to identify exploitation rate, while those with a long generation time will show better ranking skill than short-lived species. If changes in survival and recruitment have occurred independently of fishing, then a change in the indicators would be expected. For indicators based on the left-hand limb or central tendency of the length–frequency distribution (L_c, $L_{25\%}$, L_{maxY}, and L_{mean}), variability in recruitment has a large effect.

Skill to classify stocks as being over- or under-fished was poor and depends on the chosen reference point and the corresponding reference level, and hence the assumed biology. In

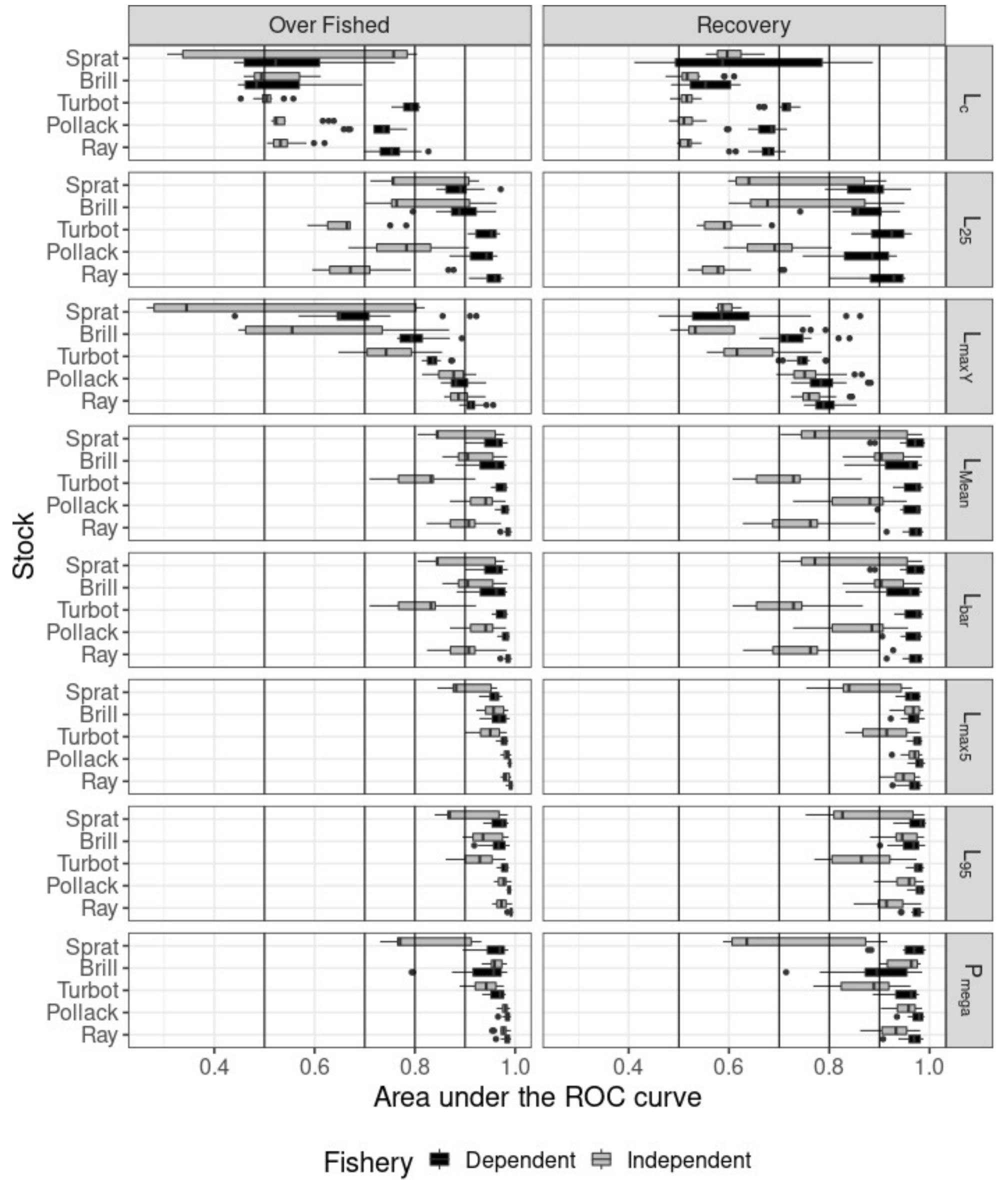

Figure 9. Area under ROC curve, for the reference levels in Table 2 (ICES) and the best reference level (tuned). A LBI with a perfect skill will have a value of 1. The boxplot hinges correspond to the inter-quartile range, while the whiskers extends from the hinges to the values that are within 1.5 times the inter-quartile.

a data-limited situation, there is likely to be uncertainty about the true natural mortality-at-age, the stock–recruitment relationship, and vulnerability, all of which determine target and limit reference points. Therefore, managing stocks on trends will be more robust than managing them relative to reference points.

A possibility is to choose a reference period, when a stock was agreed to be in good health, and compare current indicator levels to this. However, indicators vary depending on whether the stock is declining or recovering. In the former, there will be more year–classes and larger individuals in the population than in the latter. Therefore, choosing a historical period before a stock was overfished may mean that apparent recovery, especially for a stock with low k and hence low natural mortality, will take longer than if a biomass reference point was used that ignored the population structure. This is not necessarily a weakness of LBIs, as the Marine Strategy Framework Directive (MSFD), which has the overarching objective of achieving and maintaining Good Environmental Status (GES), includes a legal requirement to consider the impact of fishing on population demography (Kell *et al.*, 2015) and a healthy stock should have a range of year–classes contributing to spawning reproductive potential (Kell *et al.*, 2015).

Risk is asymmetric, as allowing a stock to be overexploited will eventually require a long-term recovery plan and loss of yield in the medium-term, while underexploitation can be corrected as soon as it is identified. Therefore, different LBIs and reference points may be appropriate for triggering management action and for use as targets. Identifying the best reference level for an indicator reference point combination can be done by tuning to find the reference levels that can best

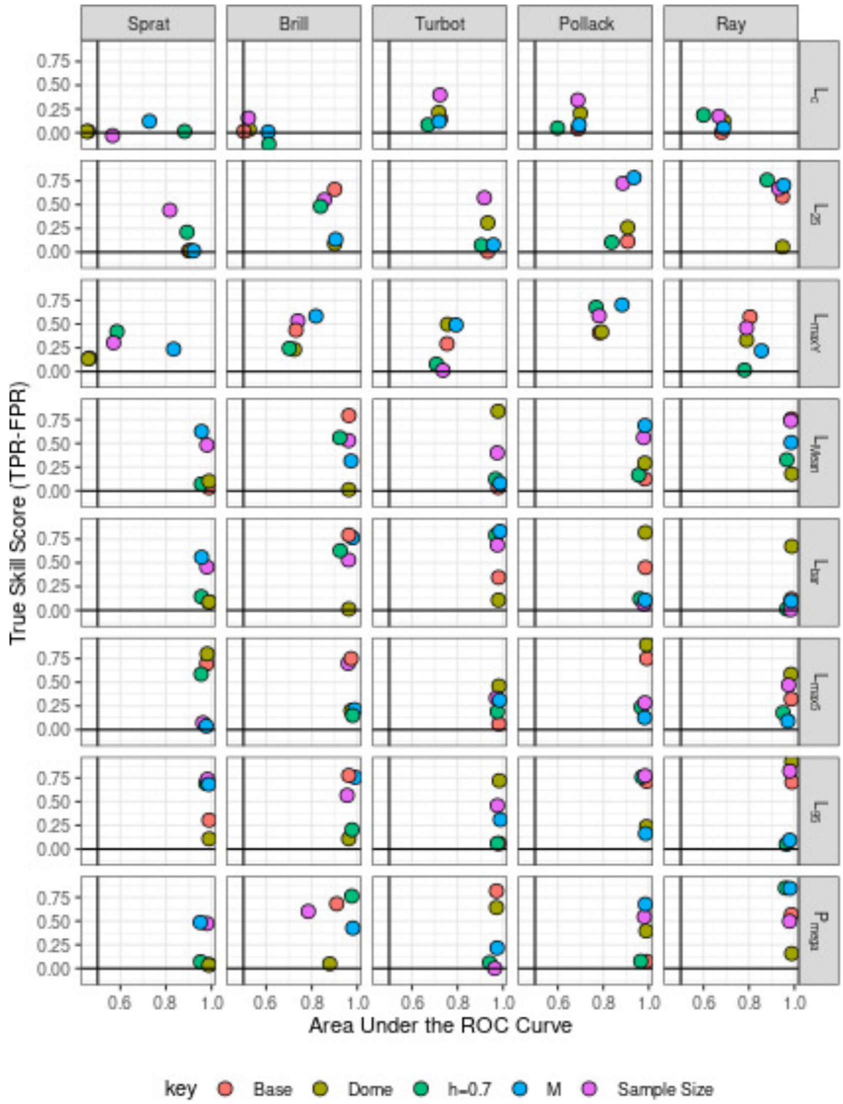

Figure 10. Overfishing: TSS compared to area under the ROC curve, for stock and indicator.

meet management objectives (e.g. Shephard *et al.*, 2018; Fischer *et al.*, 2020b). While receiver operator characteristics can be used to develop generic reference levels, it is thought more appropriate to develop stock-specific reference points (e.g. Miethe *et al.*, 2019).

The area under the ROC curve is a performance measure for machine learning algorithms, and exhibits a number of desirable properties when compared to overall accuracy: increased sensitivity in Analysis of Variance (ANOVA) tests; a standard error that decreased as both AUC and the number of test samples increased; decision threshold independent; and it is invariant to a priori class probabilities (Bradley, 1997). The use of ROC can, therefore, help in tuning, i.e. calibration, and also for unsupervized learning applications to find patterns or general rules.

Length-based assessment methods have also been used in data and capacity limited situations (e.g. Pons *et al.*, 2019, 2020). Length-based methods allow goodness of fit diagnostics to be evaluated and estimates of uncertainty to be derived, and reference points such as F_{MSY} and those based on spawner per recruit (SPR) to be derived. They can also potentially be validated using observations (Kell *et al.*, 2021).

There are several methods that use life history information and length composition from the catch to estimate fishing intensity and derive values of SPR that can be used as a proxy for stock status. For example, length-based spawning potential ratio (LBSPR; Hordyk *et al.*, 2014). LBSPR uses the Beverton–Holt life history ratios in an equilibrium-based population model applying the shape of the length composition data compared to the expected unfished length structure to estimate the ratio of fishing mortality and natural mortality (F/M) and derive SPR. Another method is the length-based integrated mixed effects model (LIME; Rudd and Thorson, 2018), which also requires biological information and length composition data to derive SPR, but relaxes the equilibrium conditions by treating recruitment as a random effect over time and estimating annual F as fixed effects. The inputs are: M, k, L_∞, the CV for L_∞, t_0, selectivity parameters (L_{50} and L_{95}), the steepness of the stock–recruitment relationship, and the parameters of the length–weight relationship a and b. However, if these

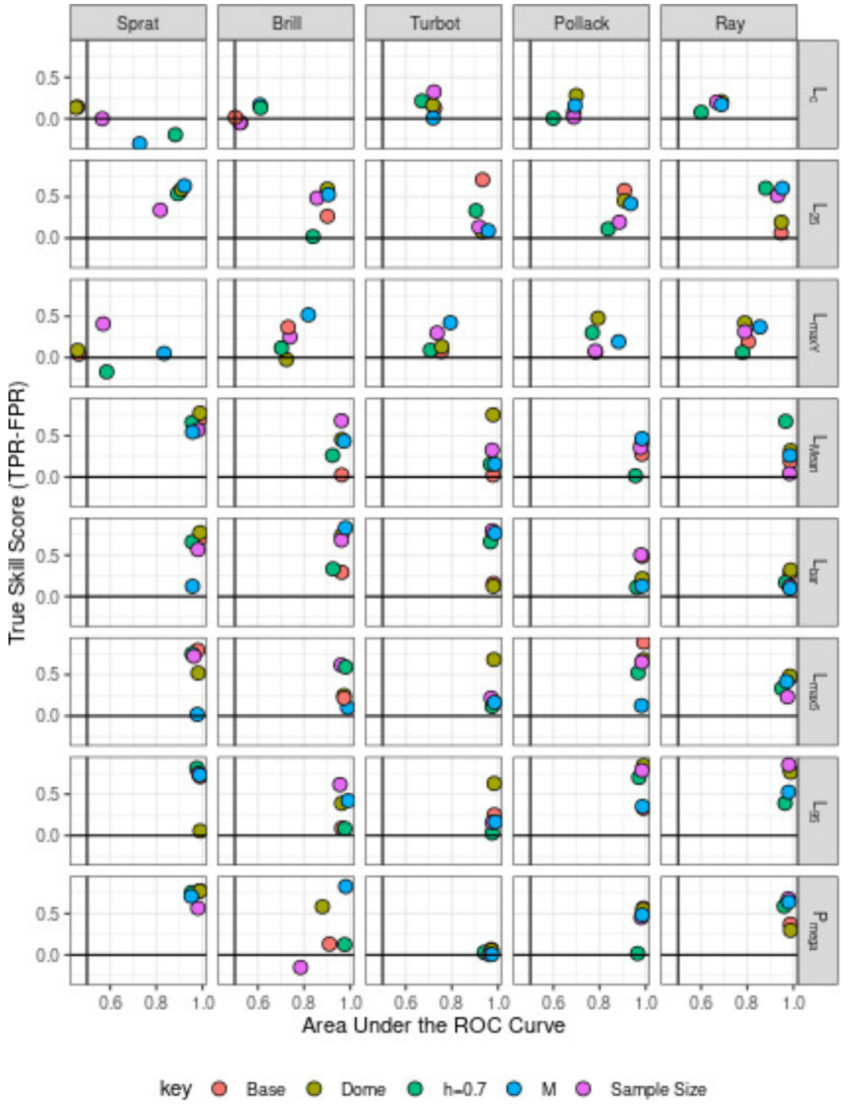

Figure 11. Recovery: TSS compared to area under ROC curve, for stock and indicator.

parameters are uncertain, as seen in this study, the ability to classify stocks will be poor.

The procedure used in this study could be used to compare LBIs and estimates from length-based methods. To screen indicators used in management procedures and compare them to length-based methods before developing case-specific management strategy evaluation for verification that the procedures work as expected despite uncertainty. This can be conducted by conditioning the Operating Models on a stock-by-stock basis (e.g. Geromont and Butterworth, 2014; Fischer *et al.*, 2021).

LBIs can be used to set management advice, as part of an empirical control rule (e.g. Fischer *et al.*, 2020a) for a relatively low cost of data collection, compared to data-rich stock assessments, which typically rely on time series of survey and catch data by age class. Also, in data-rich cases, there is often considerable uncertainty about model structure and the values of parameters such as natural mortality and the steepness of the stock–recruitment relationship. For example, although, integrated models bring together all relevant information into a single framework, problems remain due to missing data, inadequate theory, latent state variables, and unpredictable future elements (Gass, 1983). While, processes that impact estimates of quantities of management interest may be misspecified (e.g. Lee *et al.*, 2011, 2012; Jiao *et al.*, 2012; Simon *et al.*, 2012). Therefore, different parameters values or models with quite different structures may provide equally good fits (Kell *et al.*, 2021). Both kinds of lack of identifiability are common where the observations are incomplete, or latent (Bartholomew *et al.*, 2011). Therefore, LBIs when used as part of empirical control rules may out-perform data-rich stock assessment methods, if they can be selected to be robust to the assumptions that may bias data-rich assessments.

The procedure can also be used to quantify the value of information, for example, what will be the increase in yield (i.e. value) for a reduction in uncertainty, i.e. resolving a hypothesis or for an improvement in data quality for control rules based LBIs designed to achieve *MSY*. For example, if the performance of an indicator used as part of a harvest control rule depends on a hypothesis, and there has to be a high probability of avoiding stock collapse, then this may result in lower yields than if the correct hypothesis could be identified. The cost of the research required to identify the correct hypothesis can be compared to the foregone yield.

LBIs as well as providing single species advice, are potentially useful as part of EBFM (e.g. Blanchard *et al.*, 2005; Babcock *et al.*, 2013). In this study, LBIs were shown to be sensitive to fishing impacts and to respond to management action, they, therefore, have potential to help managers to assess whether changes in the fish community are a desirable or undesirable response to management (Nicholson and Jennings, 2004), and whether environmental effects are impacting the fish community.

Conclusions

Various LBIs have been proposed, and there are many potential proxy reference points against which to assess them, (e.g. M, L_{50}, L_∞, k, and CV_{L_∞}). To be effective and cost-efficient, indicators should be minimized, complementary, and non-redundant (Shin *et al.*, 2010; Kershner *et al.*, 2011). Therefore, indicators should be screened to ensure that they are robust proxies for system attributes and pressures (Fulton *et al.*, 2005). This is true whether the purpose is to develop single species management advice or to develop ecosystem indicators based on groups of species by life history types or guilds.

We, therefore, developed a procedure based on receiver operator characteristics to screen LBIs and to tune reference levels to indicate overfishing. The procedure used an Operating Model based on life history theory and an Observation Error Model to simulate length samples. It was found that LBIs performed better for identifying trends than state. Reference levels were also stock specific, and so LBIs are better at indicating trends than at quantifying exploitation level. Therefore, LBIs should be calibrated, i.e. tuned to meet agreed objectives, if they are to be used for management.

It was shown by Fischer *et al.* (2020b) for data-limited stocks that to reduce the risk of overexploitation and foregoing yield management should be linked to life history traits. Therefore, given the variety of LBIs, life history traits, fishery types, and the associated uncertainty (Shephard *et al.*, 2018), it is necessary to ensure that advice is robust to uncertainty. This requires calibration, verification, and validation. Calibration is the establishment of a relation between an observed quantity that will trigger management action, such as mean size and fishing mortality. Verification is the provision of objective evidence that a given procedure meets the specified requirements. While validation is ensuring that management objectives are met. In this study, we perform calibration to compare LBIs, and reference points to exploitation level. Verification is best performed using management strategy evaluation to identify the best performing rules, and validation by reviewing the performance of rules after implementation.

There are limitations to the approach, however, since parameterization of the Operating Model and Observation Error Model did not account for all the complexities of real stocks. Therefore, the true uncertainties are likely to be underestimated. However, the analysis did show clear patterns in the performance of the different LBIs depending on life history characteristics, whether data are based on surveys or catch, and for stock undergoing over exploitation or recovery. The procedure can also be used to compare LBIs to length-based methods based on their ability to meet management objectives. Future work should include case-specific management plans where, for a particular stock, the most robust LBI and the most appropriate reference point are identified. Additionally, the approach can be used to explore the most-cost effective ways to collect length data (surveys or commercial catch) and information about the life history of the stock, while ensuring that these can be used to trigger and monitor management action despite uncertainty.

Supplementary material

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Data availability

The data underlying this article are available in the github repository flrpapers and can be found at https://github.com/flrpapers.

Acknowledgements

Laurence Kell's and Cóilín Minto's involvement was funded through the MyDas project under the Marine Biodiversity Scheme, which is financed by the Irish government and the European Maritime and Fisheries Fund (EMFF, grant no. ITT17-015) as part of the EMFF Operational Programme for 2014-2020.

References

Babcock, E. A., Coleman, R., Karnauskas, M., and Gibson, J. 2013. Length-based indicators of fishery and ecosystem status: Glover's reef marine reserve, Belize. Fisheries Research, 147: 434–445.

Bartholomew, D. J., Knott, M., and Moustaki, I. 2011. Latent Variable Models and Factor Analysis: A Unified Approach, 904, John Wiley & Sons, New York NY.

Beverton, R., and Holt, S. 1993. On the Dynamics of Exploited Fish Populations, 11, Springer, Berlin, Heidelberg.

Blanchard, J. L., Dulvy, N. K., Jennings, S., Ellis, J. R., Pinnegar, J. K., Tidd, A., and Kell, L. T. 2005. Do climate and fishing influence size-based indicators of celtic sea fish community structure? ICES Journal of Marine Science, 62: 405–411.

Boorman, D. B., and Sefton, C. 1997. Recognising the uncertainty in the quantification of the effects of climate change on hydrological response. Climatic Change, 35: 415–434.

Bradley, A. P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30: 1145–1159.

Brooks, E. N., Powers, J. E., and Cortés, E. 2009. Analytical reference points for age-structured models: application to data-poor fisheries. ICES Journal of Marine Science, 67:165–175.

Caddy, J. 1998. A Short Review of Precautionary Reference Points and Some Proposals for their Use in Data-Poor Situations, 379, Food and Agriculture Organisation, Rome.

Carruthers, T. R., Kell, L. T., Butterworth, D. D., Maunder, M. N., Geromont, H. F., Walters, C., McAllister, M. K. *et al.* 2016. Performance review of simple management procedures. ICES Journal of Marine Science, 73: 464–482.

Connell, L., and Keane, M. T. 2006. A model of plausibility. Cognitive Science, 30: 95–120.

Cury, P. M., Fromentin, J.-M., Figuet, S., and Bonhommeau, S. 2014. Resolving Hjort's dilemma how is recruitment related to spawning stock biomass in marine fish? Oceanography, 27: 42–47.

Denney, N. H., Jennings, S., and Reynolds, J. D. 2002. Life–history correlates of maximum population growth rates in marine fishes. Proceedings of the Royal Society of London B: Biological Sciences, 269: 2229–2237.

Dowling, N., Dichmont, C., Haddon, M., Smith, D., Smith, A., and Sainsbury, K. 2015. Empirical harvest strategies for data-poor fisheries: a review of the literature. Fisheries Research, 171: 141–153.

Fischer, S., José, D., and Kell, T. L. 2020a. Linking the performance of a data-limited empirical catch rule to life-history traits. ICES Journal of Marine Science, 77:1914–1926.

Fischer, S. H., De Oliveira, J. A., and Kell, L. T. 2020b. Linking the performance of a data-limited empirical catch rule to life-history traits. ICES Journal of Marine Science, 77: 1914–1926.

Fischer, S. H., De Oliveira, J. A., Mumford, J. D., and Kell, L. T. 2021. Using a genetic algorithm to optimize a data-limited catch rule. ICES Journal of Marine Science, 78:1311–1323.

Fulton, E. A., Smith, A. D., and Punt, A. E. 2005. Which ecological indicators can robustly detect effects of fishing? ICES Journal of Marine Science, 62: 540–551.

Garcia, S. 1996. The precautionary approach to fisheries and its implications for fishery research, technology and management: an updated review. FAO Fisheries Technical Paper, pp. 1–76. FAO, Rome.

Gass, S. I. 1983. Decision-aiding models: validation, assessment, and related issues for policy analysis. Operations Research, 31: 603–631.

Geromont, H. F., and Butterworth, D. S. 2014. Complex assessments or simple management procedures for efficient fisheries management: a comparative study. ICES Journal of Marine Science, 72: 262–274.

Gislason, H., Daan, N., Rice, J. C., and Pope, J. G. 2010. Size, growth, temperature and the natural mortality of marine fish. Fish and Fisheries, 11: 149–158.

Green, D. M., and Swets, J. A. 1966. Signal Detection Theory and Psychophysics, 1, Wiley, New York, NY.

Hilborn, R. 2011. Future directions in ecosystem based fisheries management: a personal perspective. Fisheries Research, 108: 235–239.

Hilborn, R. 2020. Measuring fisheries management performance. ICES Journal of Marine Science, 77: 2432–2438.

Hilborn, R., Amoroso, R. O., Anderson, C. M., Baum, J. K., Branch, T. A., Costello, C., de Moor, C. L. *et al.* 2020. Effective fisheries management instrumental in improving fish stock status. Proceedings of the National Academy of Sciences, 117: 2218–2224.

Hordyk, A., Ono, K., Valencia, S., Loneragan, N., and Prince, J. 2014. A novel length-based empirical estimation method of spawning potential ratio (SPR), and tests of its performance, for small-scale, data-poor fisheries. ICES Journal of Marine Science, 72: 217–231.

ICES 2015. Report of the fifth workshop on the development of quantitative assessment methodologies based on life-history traits, exploitation characteristics and other relevant parameters for data-limited stocks (WKLIFE v). 153pp.

Jensen, A. 1996. Beverton and holt life history invariants result from optimal trade-off of reproduction and survival. Canadian Journal of Fisheries and Aquatic Sciences, 53: 820–822.

Jiao, Y., Smith, E. P., O'Reilly, R., and Orth, D. J. 2012. Modelling non-stationary natural mortality in catch-at-age models. ICES Journal of Marine Science, 69: 105–118.

Karnauskas, M., McClellan, D. B., Wiener, J. W., Miller, M. W., and Babcock, E. A. 2011. Inferring trends in a small-scale, data-limited tropical fishery based on fishery-independent data. Fisheries Research, 111: 40–52.

Kell, L., Mosqueira, I., Grosjean, P., Fromentin, J., Garcia, D., Hillary, R., Jardim, E. *et al.* 2007. FLR: an open-source framework for the evaluation and development of management strategies. ICES Journal of Marine Science, 64: 640.

Kell, L. T., Nash, R. D., Dickey-Collas, M., Mosqueira, I., and Szuwalski, C. 2015. Is spawning stock biomass a robust proxy for reproductive potential?. Fish and Fisheries, 17: 596–616.

Kell, L. T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., and Fu, D. 2021. Validation of stock assessment methods: is it me or my model talking?. ICES Journal of Marine Science, 78: 2244–2255.

Kershner, J., Samhouri, J. F., James, C. A., and Levin, P. S. 2011. Selecting indicator portfolios for marine species and food webs: a Puget sound case study. PLoS ONE, 6: e25248.

Lee, H.-H., Maunder, M. N., Piner, K. R., and Methot, R. D. 2011. Estimating natural mortality within a fisheries stock assessment model: an evaluation using simulation analysis based on twelve stock assessments. Fisheries Research, 109: 89–94.

Lee, H.-H., Maunder, M. N., Piner, K. R., and Methot, R. D. 2012. Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models?. Fisheries Research, 125: 254–261.

Matsuda, H., Fukase, K., Kotaki, K., and Asano, K. 1996. Inconsistency between the per capita fecundity and estimate of the yearly egg production of the chub mackerel *Scomber japonicus* in Japan. Fisheries Science, 62: 178–183.

Miethe, T., Reecht, Y., and Dobby, H. 2019. Reference points for the length-based indicator l_{max5} for use in the assessment of data-limited stocks. ICES Journal of Marine Science, 76: 2125–2139.

Mildenberger, T. K., Berg, C. W., Kokkalis, A., Hordyk, A. R., Wetzel, C., Jacobsen, N. S., Punt, A. E. *et al.* 2021. Implementing the precautionary approach into fisheries management: biomass reference points and uncertainty buffers. Fish and Fisheries, 23:73–92.

Murawski, S., Rago, P., and Trippel, E. 2001. Impacts of demographic variation in spawning characteristics on reference points for fishery management. ICES Journal of Marine Science, 58: 1002–1014.

Nicholson, M. D., and Jennings, S. 2004. Testing candidate indicators to support ecosystem-based management: the power of monitoring surveys to detect temporal trends in fish community metrics. ICES Journal of Marine Science, 61: 35–42.

Ono, K., Licandeo, R., Muradian, M. L., Cunningham, C. J., Anderson, S. C., Hurtado-Ferro, F., Johnson, K. F. *et al.* 2015. The importance of length and age composition data in statistical age-structured models for marine species. ICES Journal of Marine Science, 72: 31–43.

Ovando, D., Hilborn, R., Monnahan, C., Rudd, M., Sharma, R., Thorson, J. T., Rousseau, Y. *et al.* 2021. Improving estimates of the state of global fisheries depends on better data. Fish and Fisheries, 22: 1377–1391.

Pons, M., Cope, J. M., and Kell, L. T. 2020. Comparing performance of catch-based and length-based stock assessment methods in data-limited fisheries. Canadian Journal of Fisheries and Aquatic Sciences, 77:6.

Pons, M., Kell, L., Rudd, M. B., Cope, J. M., and Lucena Frédou, F. 2019. Performance of length-based data-limited methods in a multifleet context: application to small tunas, mackerels, and bonitos in the Atlantic Ocean. ICES Journal of Marine Science, 76: 960–973.

Punt, A. 2008. Refocusing stock assessment in support of policy evaluation. *In* Fisheries for Global Welfare and Environment, pp. 139–152. TERRAPUB, Tokyo.

Radatz, J., Geraci, A., and Katki, F. 1990. IEEE Standard Glossary of Software Engineering Terminology, 610121990, pp. 121990.IEEE Standards Board, New York, NY.

Restrepo, V. R., and Powers, J. 1999. Precautionary control rules in us fisheries management: specification and performance. ICES Journal of Marine Science, 56: 846–852.

Reynolds, J. D., Jennings, S., and Dulvy, N. K. 2001. Life histories of fishes and population responses to exploitation. *In* Conservation Biology Series, pp. 147–168. Cambridge University Press, Cambridge.

Roff, D. 1984. The evolution of life history parameters in teleosts. Canadian Journal of Fisheries and Aquatic Sciences, 41: 989–1000.

Rudd, M. B., and Thorson, J. T. 2018. Accounting for variable recruitment and fishing mortality in length-based stock assessments for data-limited fisheries. Canadian Journal of Fisheries and Aquatic Sciences, 75: 1019–1035.

Sainsbury, K., and Sumaila, U. R. 2003. 20 incorporating ecosystem objectives into management of sustainable marine fisheries, Including'Best Practice'Reference points and use of marine protected areas. *In* Responsible Fisheries in the Marine Ecosystem, pp. 343. CABI Publishing, Wallingford.

Shephard, S., Davidson, I. C., Walker, A. M., and Gargan, P. G. 2018. Length-based indicators and reference points for assessing data-poor stocks of diadromous trout *Salmo trutta*. Fisheries research, 199: 36–43.

Shin, Y.-J., Bundy, A., Shannon, L. J., Simier, M., Coll, M., Fulton, E. A., Link, J. S. *et al*. 2010. Can simple be useful and reliable? Using ecological indicators to represent and compare the states of marine ecosystems. ICES Journal of Marine Science, 67: 717–731.

Simon, M., Fromentin, J.-M., Bonhommeau, S., Gaertner, D., Brodziak, J., and Etienne, M.-P. 2012. Effects of stochasticity in early life history on steepness and population growth rate estimates: an illustration on Atlantic bluefin tuna. PLoS ONE, 7: e48583.

Thorson, J. T., Kell, L. T., De Oliveira, J. A., Sampson, D. B., and Punt, A. E. 2015. Introduction. Fisheries Research, 171: 1–3.

Trippel, E. 1999. Estimation of stock reproductive potential: history and challenges for Canadian Atlantic gadoid stock assessments. Journal of Northwest Atlantic Fishery Science, 25: 61–82.

Von Bertalanffy, L. 1957. Quantitative laws in metabolism and growth. Quarterly Review of Biology, 32:217–231.

Zhou, K., Doyle, J. C., and Glover, K. 1996. Robust and Optimal Control, 40. Prentice Hall, Hoboken, NJ.

Handling Editor: Ernesto Jardim