DRAFT:

Consistency and Robustness testing of candidate reference point systems for North East Atlantic stocks

Henning Winker[1], Massimiliano Cardinale[2], Iago Mosqueira[3], Laurance Kell[4], Christoph Konrad[1], Michael Gras[1], Rishi Sharma[5] & Colm Lordan[6]*

[1]*Joint Research Centre (JRC), European Commission, TP 051, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy*

[2]*Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Lysekil, Sweden*

[3]*Wageningen Marine Research, Haringkade 1, 1976CP IJmuiden, the Netherlands*

[1]*Centre for Environmental Policy, Imperial College London, Weeks Building, 16-18 Princes Gardens, London SW7 1NE, UK*

[5]*Food and Agricultural Organization, Fishery and Aquaculture Policy and Resources Division, Rome, Lazio 00153, Italy*

[6]*Marine Institute, Oranmore, Galway,Ireland*

Abstract

Recently, the ICES Workshop on ICES reference points (WKREF1, 2021) was tasked to provide a thorough review of the ICES reference points system as a basis to re-evaluate the process for estimating, updating and communicating reference points in the context of the ICES advice. The key recommendations of WKREF1 were to: i) revise and simplify how B_{lim} is derived, including the possibility to determine B_{lim} as a fraction of B_0 based on biological principles and international best practice; ii) $F_{P.05}$ should be calculated without $B_{trigger}$; iii) to use biological proxies for deriving F_{MSY}, and the F_{MSY} proxy must not exceed $F_{P.05}$ consistent with ICES Precautionary Approach (PA) ; iv) to report a biomass target (B_{trg}) that corresponds to the F_{MSY} proxy; and v) to set $B_{trigger}$ as either a fraction of B_{trg} or multiplier of B_{lim}. In this paper, we conduct a large-scale simulation testing experiment with feedback control for 64 ICES Category 1 stocks, with the aim to evaluate the consistency and robustness of candidate reference point systems. In accordance with the objectives of ICES advice framework, the evaluation criteria for testing consistency are based on the following objects: (1) to not exceed a 5% probability of SSB falling below B_{lim} , (2) to achieve high long-term yields that correspond to at least 95% of the median yield at constant F_{MSY} (MSY), (3) to attain a high probability that SSB is above the FAO threshold of 80% of the B_{trg} proxy for B_{MSY}. By considering stock-specific productivity and taxonomic grouping, we then put forward the best performing candidate reference point systems for further robustness testing under alternative misspecifications of the stock recruitment relationship . Based

on our simulation results, we present straightforward and transparent guidelines for setting optimal reference points depending on the stock's productivity characteristics. We align this new reference point system with a status classification system that is intended to facilitate clear and unambiguous interpretation of the stock status.

Keywords: North East Atlantic stocks; Reference points; simulation testing; harvest control rule; shortcut MSE

1. Introduction

Central to fisheries advice worldwide are reference points, which are used to classify and communicate current resource status relative to sustainability limits and to provide targets for determining future fishing opportunities, e.g. to set the total allowable catch (TAC) in quota managed fisheries. Stock assessment models are generally considered the basis for scientific advice. In practice, however, the process starts with the processing of fishery dependent and independent observations used by the assessment model, which are typically associated with large and systematic sampling errors (Carruthers et al. 2017). The assessment model itself relies on many assumptions about model structure, in the form of the underlying deterministic relationships (e.g. between stock and recruitment, SR) and population parameters (e.g. natural mortality, M). These contribute to structural and estimation uncertainties associated with stock assessment results (Patterson et al., 2001), where uncertainty is the difference between the model and reality. Accounting for these uncertainties is one of the key challenges for operationalising reference point systems able to provide consistent and robust scientific advice on fishing opportunities (Ralston et al. 2011). However, despite common commitments to maintain or restore stocks at levels capable of producing maximum sustainable yield (MSY) and the Precautionary Approach (PA) to fisheries (UN 1995; FAO 1995), international advice standards vary widely in how this challenge is addressed in particular regarding specifying and estimating the corresponding target- (TRPs) and limit reference points (LRPs), as well as setting the operational trigger points used in harvest control rules.

A main objective of reference points is to prevent overfishing, e.g. growth, recruitment, economic and target overfishing. Growth and recruitment overfishing are generally associated with limit reference points, while economic overfishing may be expressed in terms of either targets or limits. The difference between targets and limits is that indicators may fluctuate around targets, but in general limits should not be crossed. Target overfishing occurs when a target is overshot, although variations around a target are not necessarily considered of serious concern unless a consistent bias becomes apparent. In contrast, even a single violation of the LRP may indicate the need for immediate action in order to be consistent with the PA. On the other hand, triggers are intended to implement action before limits are reached.

In age-structured assessments, MSY based reference points can be either estimated in the model, i.e. when the SR is fitted internally in the assessment model, or derived post-hoc from the model results, using yield and spawner per recruit assumption combined with a SR relationship. These reference points typically assume equilibrium, or an alternative approach is to run long-term stochastic projections. Benefits of the latter approach are that reference points can account for structural uncertainties and estimation errors (e.g. required for ensembles). A problem, however, is that as reference points estimation procedures become more complicated and computationally demanding, they become less transparent and difficult to verify and validate; where verification is the provision of objective evidence that a given procedure meets the specified requirements, and validation is ensuring that management objectives are actually met. This is complicated by the fact that the quantities used to compute reference points are model-based estimated latent quantities, such as numbers-at-age, spawning stock biomass (SSB) and fishing selectivity, which can therefore not be validated by observations (Kell et al. 2021). Thus, verification and validation of reference point systems need to be based on simulation-testing.

Simulation-testing allows verifying consistency of a reference point system in meeting the quantifiable management objectives (e.g. thresholds of TRPs and LRPs) and validating the system's robustness of achieving the underlying goals (e.g. biomass levels at MSY). The consistency of a reference point system relies on the setting TRPs, LRPs and trigger points so that target thresholds are exceeded and the limit thresholds are not breached. By contrast, a reference point system would be internally inconsistent if, for example, the rules for setting the target fishing mortality (F_{trg}) would fail systematically to exceed the corresponding target biomass threshold. Evaluating consistency does not need knowledge of the "true" quantities and can therefore be simulation-tested using "self-tests". The term self-test is used because the assumptions for simulating the stock dynamics are the same as the assumptions for computing biological reference point proxies. Thus, the reference point estimator is correctly specified with respect to the operating model (OM) simulator (Deroba *et al.*, 2015). In contrast to consistency, the term robustness refers in statistics to a model that provides correct inference despite its assumptions being violated; whereas robustness in engineering means that a system functions correctly in presence of uncertainty (Kell *et al.*, 2016). In the context of fisheries advice both meanings are interrelated and highly relevant. Evaluating the robustness of a reference point system therefore requires testing if it can also produce desired outcomes in situations where the reality (OM) differs in assumptions from the reference point estimator (Deroba *et al.*, 2015). Using simulations for robustness testing provides an additional scope beyond a self-test because it can be used to validate that if by meeting management objectives, the desired yet latent state of the stock (e.g. biomass at or above the "true" B_{MSY}) is achieved with high probability despite imperfect knowledge of the true population dynamics.

In the International Council for the Exploration of the Sea (ICES), the PA to fishing (UN 1995; FAO 1995) was introduced first in 1998 without consideration of MSY, and the ICES MSY approach was subsequently integrated into the PA framework in 2009 (ICES, 2012). The ICES reference point system has since evolved and undergone several revisions (Lassen *et al.*, 2014; Silvar-Viladomiu *et al.*, 2021). Part of this evolution is driven by the scientific advances in accounting for risk and uncertainty, but also by policy requirements to implement the ecosystem-based approach to fisheries management, among which multi-species interactions, impacts on bycatch species, adaptation to environmental change and socio-economic considerations, are important drivers. Fisheries advice is therefore becoming increasingly more sophisticated and also more complex. However, sequentially adding more elements, rules and exceptions can also result in ambiguities, inconsistencies and conflicts in achieving multiple objectives. Various ambiguities and potential inconsistencies related to reference points have recently been identified by ICES workshops WKREBUILD (ICES, 2021a), WKGMSE3 (ICES, 2020) and WKRPCHANGE (ICES, 2021b) and concerns have been raised that current reference point estimation procedures are complex and difficult to communicate both internally, among the scientific community, and externally to stakeholders and clients.

For age-structured data rich Category 1 assessments in ICES, F_{MSY} is mostly derived through stochastic forward simulations that are externally implemented in the EQSIM software (Simmonds *et al.*, 2010). The fishing mortality (F) at MSY (F_{MSY}) is in the first instance determined as the F that achieves maximum median long-term yield (F_{MMY}). These projections are commonly run with an HCR, in which the ICES MSY $B_{trigger}$ point instantly reduces fishing mortality linearly if biomass falls below it. Therefore, F_{MMY} in conjunction with MSY $B_{trigger}$ can lead to higher F_{MSY} estimates in comparison to values from projections run at constant fishing mortality (WKREF1, 2021). Both lower and upper ranges for F_{MSY} are provided, but these are bound on the condition to not reduce the long-term yield corresponding to F_{MMY} by more than 5%. However, to be consistent with ICES PA the final estimate of F_{MSY} must not exceed the fishing mortality $F_{P.05,}$ where $F_{P.05}$ is associated with a 5% probability for biomass to fall below B_{lim} (i.e. F_{MSY} is the minimum of $F_{P0.5}$ and F_{MMY}). A, perhaps unique, feature is that ICES MSY approach does not have a formal biomass TRP, with the B_{MSY} estimate corresponding to F_{MSY} being neither used nor reported. Strictly speaking there is one exception, however, in that MSY $B_{trigger}$ can be specified as the lower 5th percentile of the B_{MSY} estimate. This has the seemingly risk-prone property that the higher the uncertainty, the lower biomass has to fall to reduce fishing. In practice, however, this MSY $B_{trigger}$ rule is used rarely, and instead the precautionary biomass (B_{pa}) is normally set equal to MSY $B_{trigger}$, which is approximated by a multiplier of B_{lim} (typically ~ 1.4). Without a biomass TRP, the MSY $B_{trigger}$ not only serves as an operationalized trigger in the ICES Advice Rule, but also as a threshold to classify the stock status to be within 'safe biological' limits if biomass is above it. Without a biomass TRP, the ICES MSY approach is strongly "bottom-up" dependent on B_{lim}. For age-structured models, B_{lim} is a derived deterministic value of absolute spawning stock biomass (SSB) that is independent from any other biomass reference point (e.g. B_{MSY} or B_0). In ICES, there are currently 6 typologies of SR data patterns for determining B_{lim}. Of these, one approach is based on fitting a segmented regression to the SR data to quantify its breakpoint as B_{lim}, but this was only used for 14% of 77 stocks analysed (WKREF1, 2021). Of the other five rule-based approaches, setting B_{lim} to the lowest observed SSB (B_{loss}) was the most common (41%). A meaningful comparison of B_{lim} ranges across stocks or advisory bodies is challenging because the common reference values of B_{MSY} and B_0 are not reported. A recent analysis on 69 ICES stocks, which used a default segmented regression with B_{lim} benchmark estimates as its break-point, indicated a wide variation of B_{lim} value relative to B_0, as estimated by the EQSIM procedure, ranging from 1.3 to 38% of B_0 with a median of just under 10% (WKREF1 2021).

Recently, the ICES Workshop on ICES reference points (WKREF1, 2021) was tasked to provide a thorough review of the ICES reference points system as a basis to re-evaluate the process for estimating, updating and communicating reference points in the context of the ICES advice. The key recommendations of WKREF1 were to: i) revise and simplify how B_{lim} is derived, including the possibility to determine B_{lim} as a fraction of B_0 based on biological principles and international best practice; ii) $F_{P.05}$ should be calculated without $B_{trigger}$; iii) to use biological proxies for deriving F_{MSY}, and the F_{MSY} proxy must not exceed $F_{P.05}$ consistent with ICES PA; iv) to report a biomass target (B_{trg}) that corresponds to the F_{MSY} proxy; and v) to set $B_{trigger}$ as either a fraction of B_{trg} or multiplier of B_{lim}.

In this paper, we first present an overview of international reference point systems. This is to provide the conceptual basis for conducting a large-scale simulation testing experiment with feedback control for 64 ICES Category 1 stocks, with the aim to evaluate the consistency and robustness of candidate reference point systems in accordance with the recommendations made by WKREF1 (ICES, 2021). The evaluation criteria for testing consistency are based on the following three main objectives: (1) to not exceed a 5% probability of SSB falling below B_{lim} (ICES, 2021c), (2) to achieve high long-term yields that correspond to at least 95%

of the median yield at constant F_{MSY} (MSY) (ICES, 2021c), (3) to attain a high probability that SSB is above the FAO threshold of 80% of the B_{trg} proxy for B_{MSY} (DFO, 2009; FAO, 2020; Sharma *et al.*, 2021). By considering stock-specific productivity and taxonomic grouping, we then select the best performing candidate reference point systems for robustness testing. Robustness testing based on simulations enables comparison against 'true' quantities of B_{lim}, MSY and B_{MSY} as derived from the OM, which can differ to various extent from the reference estimator. For example, it allows us to evaluate if a median SSB close to the "true" B_{MSY} is indeed achieved in cases where SSB is above the lower threshold set for B_{trg}. Based on this simulation testing framework, we provide best practice guidelines on the estimation of reference points that are simplified, yet robust, data driven and consistent with the criteria of ICES advice framework.

2. Overview of international reference point systems

Direct estimates for fishing mortality (F_{MSY}) and biomass (B_{MSY}) that correspond to the maximum surplus production MSY are the default TRPs in tuna Regional Management Organizations (RMFOs), such as the International Commission for the Conservation of Atlantic Tunas (ICCAT) and the Indian Ocean Tuna Commission (IOTC). However, there can be exceptions of using a ratio relative to the unfished biomass (B_0) for the biomass and the corresponding TRP if there is high uncertainty about the stock recruitment relationship (e.g. SKJ). For several tuna and billfish stocks, Management Strategy Evaluations (MSEs) and harvest control rules are under development by ICCAT and IOTC but implementation of interim Management Procedures (NAtl. Albio, ICCAT Laurie) or harvest control rules (IO SKJ) are limited. In the absence of a harvest control rule, catch advice is typically based on the Kobe-2-Strategy Matrix, which depicts the probabilities of biomass exceeding B_{MSY} and F remaining below F_{MSY} as derived from medium to long-term projections (7-15 years) over a range of constant catch scenarios. In tune RFMOs, the total allowable catch (TAC) advice has generally to fulfil the minimum requirement that $B > B_{MSY}$ and $F < F_{MSY}$ with 50% probability at the end of the projection horizon. Like harvest control rules, formal implementation LRPs are pending for most stocks, but interim LRPs are increasingly put forward (Refs, Rishi). For example, in the IOTC interim LRPs were specified as a biomass limit at $B_{lim} = 0.4B_{MSY}$ and F_{lim} =1.4-1.5 F_{MSY} for tunas and swordfish, pending further updates as part of the ongoing MSE development process. By contrast, MSE has already been successfully implemented since 2012 by the Commission for the Conservation of Southern

Bluefin Tuna (CCSBT) to provide rigid TAC advice for Southern bluefin tuna (Hillary *et al.*, 2015). Here, the management procedure specifies the interim rebuilding objective to achieve spawning stock biomass (SSB) levels above a LRP of 20% B_0 with a least 70% probability and a TRP of 30% B0 to be achieved with at least 50% by 2035.

In Canada, the maximum acceptable harvest removal reference point is determined analytically as the best estimate of F_{MSY} from the stock assessment model (DFO, 2009). However, the advised fishing mortality (F_{trg}) can be at or below F_{MSY}, but must not exceed it, i.e. $F_{trg} \leq F_{MSY}$. The value for F_{trg} can be set smaller than F_{MSY} by factoring in the impact on other stocks ecosystem considerations and precaution in light of uncertainty. The stock status zones are defined as the Limit Reference Point (LRP) at the *Critical-Cautious* zone boundary, and an Upper Stock Reference Point (USR) at the *Cautious-Healthy* zone boundary. In absence of a pre-agreed harvest rule developed in the context of the PA, DFO (2019; Appendix 1b) provides provisional guidance for specifying the LRP and USR. The stock is considered to be in the *Critical Zone*, if the mature biomass, or its index, is less than or equal to 40% of the B_{MSY} estimate (i.e. $B_{lim} = 0.4\ B_{MSY}$), where B_{MSY} is the expected biomass corresponding to F_{MSY}. The stock is considered to be in the *Cautious Zone* if the biomass, or its index, is higher than 40% of B_{MSY} but lower than 80% of B_{MSY} ($0.4\ B_{MSY} < B < 0.8\ B_{MSY}$). F_{trg} is linearly reduced between the URP and the LRP. The stock is considered to be "healthy" if the biomass, or its index, is higher than 80% of B_{MSY} ($B > 0.8\ B_{MSY}$), with $F_{trg} \leq F_{MSY}$. In this case, the URP therefore serves the purpose of both reference point for stock status classification and an operationalised $B_{trigger}$ point that is bound to B_{trg} (i.e. $B_{trigger} = 0.8 B_{trg}$).

In New Zealand, Australia and the USA, biological proxies for F_{MSY} and B_{MSY} are predominantly used (Punt *et al.*, 2013). For the New Zealand Harvest Strategy Standard (New Zealand Ministry of Fisheries, 2008), detailed guidelines (New Zealand Ministry of Fisheries, 2011) on selecting proxies, so called "MSY-compatible reference points", are specified for B_{MSY} as ratios to B_0 ($B_{\%}$) and F_{MSY} based on the per-recruit spawning potential ratio ($F_{SPR\%}$). The ratios are specified according to biological classifications into very low, low, medium and high productivity species (Musick, 1999; FAO, 2001), where lower productivity is associated with more conservative ratios (e.g. F_{SPR45} and SB_{40}). The default target is to achieve B_{MSY} (or its proxy) with at least a 50% probability. LRPs comprise a "soft-limit" at 0.5 B_{MSY} or 0.2 B_0, whichever is higher, and "hard-limit" at 0.25 B_{MSY} or 0.1 B_0, whichever is higher. The soft-limit is considered breached and the stock classified as depleted if there is a more than 50% probability that the biomass falls below the soft limit, whereas the hard-limit is considered breached and stock classified as collapsed if there is more than 50% that the biomass is below the hard-limit. Catch advice is implemented via a HCR. If biomass falls below the biomass trigger point ($B_{trigger}$) located between the biomass target (B_{trg}) and the soft-limit, fishing mortality is reduced linearly to keep the stock close to the target and away from the soft-limit, where $B_{trigger}$ is typically set relative to B_{trg} (Restrepo *et al.*, 1998). Harvest strategies based on MSE are advocated and tuning criteria are designed to be fully compatible with the minimum requirements of the Harvest Strategy Standard. The default performance criteria for MSEs are therefore specified to ensure that: (1) the probability of achieving the biomass target is at least 50%, (2) the probability of breaching the soft limit does not exceed 10%, (3) and the probability of breaching the hard limit does not exceed 2%.

In the USA, the choices of proxies for F_{MSY} and B_{MSY} vary widely, but those based on $F_{SPR\%}$ (typically $F_{SPR30\%}$ to $F_{SPR45\%}$) and its corresponding $B_{SPR\%}$ or $B_{\%}$ (e.g. B_{40}) are most frequently used. F_{MSY} or its proxy determines the Maximum Fishing Mortality Threshold (MFMT), where F > MFMT invokes a condition of overfishing and associated management interventions. The target fishing mortality F_{trg} is set lower than F_{MSY} so that the probability of overfishing is reduced below 50% according to the degree of scientific uncertainty, which is referred to as P* approach for data-rich assessments (Shertzer *et al.*, 2008). The LRP is referred to as Minimum Stock Size Threshold (MSST) below which the stock is considered to be overfished and invokes requirement for a rebuilding plan. The MSST is explicitly linked to the B_{MSY} or its proxy that is often specified to be larger or equal to $0.5B_{MSY}$.

Horbowy and Luzenzzyk (2012) interpreted the use of more conservative biological proxies for F_{MSY} to be consistent with the guidelines for applying a PA within an MSY framework in Annex 2 of the UN Fish Stocks Agreement (1995), which states that fishing mortality that produces the MSY should be considered as a fishing mortality limit rather than a management target. The basis for this is also well founded in the scientific literature, which frequently found that more conservative biological proxies for F_{MSY} are more robust to asymmetric risk associated with fishing below or above the 'true' unknown F_{MSY} (Mace, 2001; Horbowy and Luzeńczyk, 2012; Hordyk *et al.*, 2019), where asymmetric risk describes the phenomenon that one direction of bias for an estimate leads to disproportionately higher risk than if the bias would occur in the other direction (Hordyk et al. 2019).

The consequence of fishing above F_{MSY} is that the biomass will decrease relative to B_{MSY}, so that yield levels close to MSY cannot be maintained. Subsequent rebuilding requires fishing mortalities lower than F_{MSY} which may come at high costs of reduced catches and long recovery time. Fishing below F_{MSY} can result in short-term yield loss but in contrast to overshooting F_{MSY} the catch opportunity still exists at higher biomass levels. In comparison to the substantial biomass increase at $F < F_{MSY}$, the long-term loss in yield is relatively small (Hordyk *et al.*, 2019). For example, Beverton (1998) noted that instead of striving for F_{max} "a simple management system based on careful monitoring of fishing effort, biological targets such as F_{95} (i.e. a lower fishing mortality the results in 95% of the maximum yield), and exploitation of a diversity of fish resources may suffice to avert further disaster and hedge against uncertainty." Restrepo et al. (1998) showed that fishing at just 75% F_{MSY} would still yield an average 0.949 - 0.989 of MSY based on deterministic age-structured models that was parameterized with 600 combination of variations of life history parameters (Mace, 1994). Hilborn's (2010) concept of 'Pretty Good Yield' is also founded on the principle that fishing near but not at the maximum yield will reduce risk of overfishing and increase robustness to uncertainties with little long-term yield loss. Horbowy and Luzenzzyk (2012) and Punt et al. (2013) showed that fishing mortality corresponding to a biomass at 40% B_0 as a proxy for B_{MSY} leads to high yield and safe biomass levels irrespective of the steepness value of the stock recruitment function. Even fishing under a harvest control rule at F_{MSY} can still be associated with high risk of a stochastic collapse below $0.5B_{MSY}$ as a result of recruitment variability, while this risk can be significantly reduced by fishing somewhat below F_{MSY} (Thorson *et al.*, 2015). Recently, Hordyk et al. (2019) demonstrated by way of simulations with stock assessment feedback-loop that there is much higher risk to long-term yields and sustainable stock biomass levels when positively biased stock parameter (*e.g. M,* steepness and historical catches) lead to an overoptimistic F_{MSY} than with the equivalent negative bias.

3. *Proposed candidate reference point system*

To make a reference point system operational requires general guidelines on how to specify the reference points in practice. A guiding principle for developing these guidelines is that reference points, such as the F_{MSY} proxies, should be set stock-specific by considering its biology, productivity and ecology, and the nature of the fisheries, following international best practice. The reference point system should be based on tangible and transparent rules and should provide a clear and unambiguous interpretation of the stock status. The proposed candidate reference point system builds on the key recommendations by ICES WKREF1 (2021), which interpret and define as follows (Figure 1):

- $\mathbf{B_{lim}}$ is the deterministic biomass limit below which a stock is considered to have reduced reproductive capacity, or productivity. For stocks where quantitative information is available, a reference point B_{lim} may be identified as the stock size below which there is a high risk of reduced recruitment. In this study, we consider Type 1 and Type 2 of the three newly proposed typologies to derive B_{lim} made by WKREF1 (2021):

- **Type 1**: Consider an empirical Hockey-Stick for deriving B_{lim} only if there is a clear relationship between stock and recruitment, the data show contrast and a breakpoint is clearly defined
- **Type 2**: Determine a plausible B_{lim}/B_0 ratio based on biological principles and life history of the stock (e.g. 10% to 25% of B_0 depending on the type of stocks)
- **Type 3:** It meant for those stocks where recruitment is dominated by occasional good year-classes (i.e. spasmodic recruitment), e.g. dynamics are process error driven, the lowest observed SSB(s) that gave rise to a good year class can be used as basis for B_{lim}

- $\mathbf{F_{P.05}}$ is the fishing mortality that is associated with a 5% risk that SSB falls below B_{lim} as derived using stochastic long-term projections.
- $\mathbf{F_{brp}}$ is the biological reference point proxy for F_{MSY} which can be computed at equilibrium or derived from long-term projections to incorporate additional structural uncertainties and estimation errors (e.g. required for ensembles). Here, we consider two type of F_{brp} estimators:
  - $\mathbf{F_{SPR\%}}$: The fishing mortality at which the spawner-biomass-per-recruit (SPR), e.g. 40%, of its unexploited level of SPR_0 at F = 0. $F_{SPR\%}$ requires no assumption about SR. We consider a range is F_{spr35} -F_{spr50} for evaluations using simulation.
  - $\mathbf{F_{B\%}}$**:** The fishing mortality at which the spawning stock biomass (SSB) is e.g. 40% of its unexploited level at B_0, i.e. F_{B40}. Computation of $F_{B\%}$ relies on a SSR assumption. For a Beverton-Holt SRR $F_{B\%}$ is smaller to the equivalent $F_{spr\%}$ (i.e. $F_{B40} < F_{spr40}$). A specific property of the segmented regression SSR (Hockey-Stick) is that $F_{B\%}$ is equal to the equivalent $F_{spr\%}$ if the corresponding biomass is larger than B_{lim}. Here, we therefore consider a lower range of F_{B30} -F_{B45} for simulation-testing.
- $\mathbf{B_{trg}}$ is the **biomass target** ($\mathbf{B_{trg}}$), i.e. the expected average biomass that corresponds to F_{brp}, which can be computed at equilibrium quantity or derived from long-term projections.
- $\mathbf{F_{trg}}$ is the fishing mortality used in the advice rule. In accordance with the ICES PA, F_{trg} must not exceed $F_{P.05}$, such that $F_{trg} = min(F_{brp}, F_{P0.5})$. The definition of F_{trg} is used here as the equivalent of F_{MSY} as defined in the current ICES advice framework to ensure a clear distinction between F_{trg} and "true" F_{MSY} of the simulated stock. Note that if $F_{P0.5} < F_{brp}$, $\mathbf{B_{trg}}$ is not adjusted upward to correspond to F_{trg}. The reasoning is that $F_{P.05}$ is thought to act as precautionary safeguard to prevent biomass to fall below B_{lim} which has no obvious implications for the need of changing B_{trg}.

- **B_0** is not directly used in the advice rule, but included here because it can be considered for specifying **F_{brp}** and **B_{trg}** values based on **$F_{B\%}$** as well as B_{lim} (Type 2). In age-structured models, B_0 is the unfished spawning biomass that is given by the product of recruitment R_0 of an unfished stock (implicit to the SR relationship) and the unfished spawning biomass-per-recruit (SPR_0) being a function of weight-at-age, maturity-at-age and natural mortality under current conditions (e.g. average of the last 3 years). If the biology is time-varying, B_0 will therefore differ from the virgin biomass that is assumed to be representative of historical conditions prior to fishing.
- Like B_{MSY}, it is therefore an implicit property of any age-structured model for which a stock recruitment relationship is estimated or assumed.
- **$B_{trigger}$** is the operationalised biomass trigger point for tuning of the harvest control rule (not a reference point). If biomass falls below $B_{trigger}$, F_{trg} is decreased linearly toward minimum biomass (default is zero) at which the fishery may be closed. The $B_{trigger}$ is a generalization of the MSY $B_{trigger}$. Two options are considered for specifying the $B_{trigger}$ value:
    - (1) as a fraction of B_{trg} (here: 0.6 - 1.0 of B_{trg})
    - (2) as multiplier of B_{lim} (here: 2 x B_{lim})

A new element to the ICES reference point system is the introduction of an explicit B_{trg} reference point that corresponds to the F_{brp} proxy for F_{MSY}. Therefore, guidelines are needed on how to quantify stock status relative to B_{trg}, for example, by specifying the level probability being close or above B_{trg} at biomass levels capable of producing MSY. In real-world stochastic systems, the biomass will fluctuate around B_{trg} when fishing at the corresponding F_{brp}. The extent of biomass fluctuation depends on the variability and autocorrelation of recruitment, as well as time-varying biological processes (e.g. somatic growth, maturation and survival). An important property of stochastic stock dynamics to consider is that the probability of biomass being below the B_{trg} tends to be larger than being above it due to the lognormal nature of the system (Thorson et al. 2015). This is aggravated for species that exhibit high recruitment variability and short generation turn-over, such as many small pelagic foraging species (Thorson et al. 2015; Mildenberger et al. 2021). One option to attain probabilities above 50% of being above B_{trg} is to reduce F_{trg} relative to F_{MSY} or its proxy (Mildenberger et al. 2021). However, considering that F_{brp} proxies tend to be more conservative than F_{MSY}, this could result in increased risk of reduced fishing opportunities by reducing a conservative a F_{brp} to an even more conservative F_{trg}. As an alternative we therefore adopted a target threshold (**B_{thresh}**) at 80% B_{trg}

to be achieved with probability of at least 50%, as used by FAO (e.g. Sharma et al., 2021) and Canada (DFO, 2019) for classifying stock status as "sustainably fished" and within "Healthy Zone (Green)", respectively.

We seek to align the reference point system with a status classification system that facilitates clear and unambiguous interpretation. A clear definition of sustainability is important to make the reference point system operational and useful, so that the achievement of sustainability can be assessed against quantitative objectives and effectively communicated to stakeholders (Quinn and Collie, 2005). Currently, ICES uses pictograms (i.e. green, yellow, red) to represent the status of the stocks and their exploitation, relative to management objectives as defined by separate categories for the ICES MSY and PA reference points. Stocks are classified by "green" and "red" symbols with respect to the reference points for fishing pressure and stock size. Separating the PA and the ICES MSY approach into different categories of reference points results in the current classification system being complex and difficult to illustrate. With the aim to unify the MSY and Precautionary approach within a single reference point system, we integrate the four colour classification system of the Kobe MSY framework used in tuna RFMOs (de Bruyn et al., 2013) with key elements for the PA frameworks drawn from ICES (ICES, 2020), the New Zealand Harvest Standard (New Zealand Ministry of Fisheries, 2008) and the Canadian Harvest Strategy (DFO, 2009).

The harvest control rule is embedded in the stock classification system and is shown together with governing reference points in **Figure 1**. The reference point system includes five stock status zones delineated for stock size by B_{lim} and $\mathbf{B_{thresh}}$, in this example set to 80% B_{trg}, and for fishing pressure by F_{trg}. Stock size for age-structured assessment is usually represented by stock spawning biomass. The $B_{trigger}$ location may vary relative to B_{trg} or B_{lim}, depending on stock's biology, and is therefore explicitly not considered for stock status classification.

The stock status zone below F_{trg} and above $\mathbf{B_{thresh}}$ is the "Sustainable" zone illustrated in green (B > $\mathbf{B_{thresh}}$ and F < F_{trg}). The orange "Overfishing" zone demarcates sustainable biomass levels above $\mathbf{B_{thresh}}$, but unsustainable fishing pressure (B > $\mathbf{B_{thresh}}$ and F > F_{trg}). The stock is classified to be in the yellow rebuilding zone if biomass is below $\mathbf{B_{thresh}}$ but fishing pressure is below F_{trg} so that biomass is predicted to increase (B > B_{tresh} and F > F_{trg}). The stock falls in the red "overfished" zone if fishing mortality is above F_{trg} and biomass falls below $\mathbf{B_{thresh}}$. However, to be consistent with the principles of the PA, a fifth "Critical" zone is introduced, if spawning stock biomass is below B_{lim}. The classification of this zone is conceptually independent of fishing pressure relative to F_{trg}, but the dark red shading that overlays the yellow "Rebuilding" zone still allows depicting if a status stock is "Critical" and under "Rebuilding" (Figure 1).

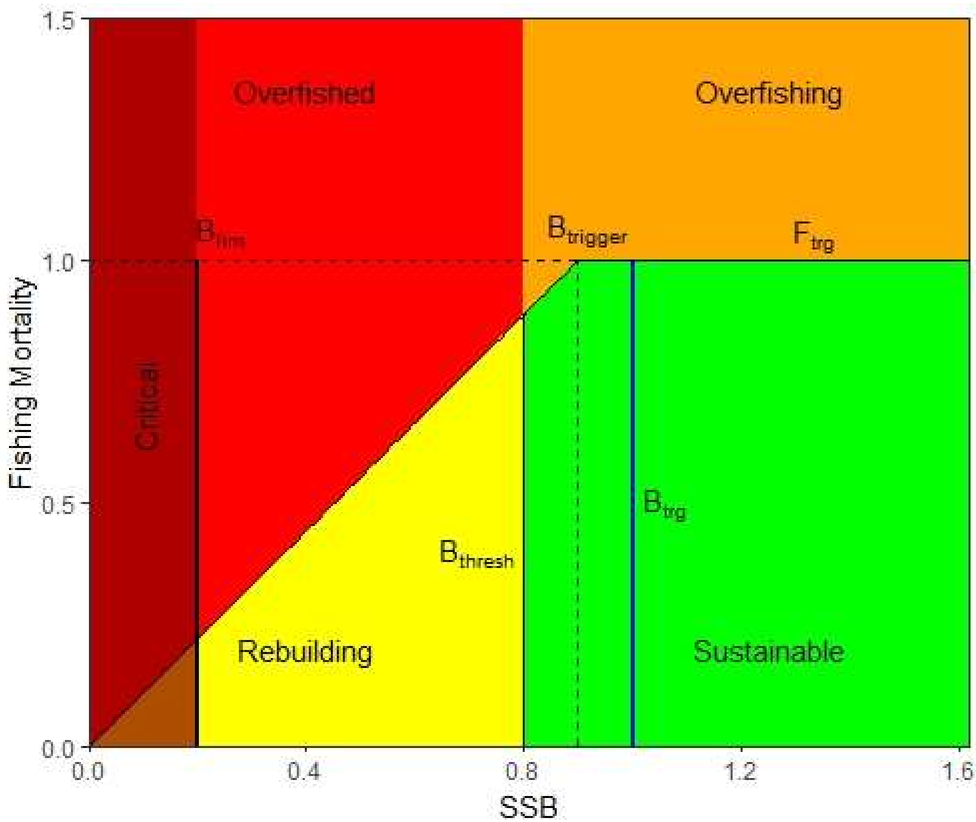

Figure 1: Proposed ICES Reference points system with integrated Harvest Control Rule. (source `FLRef` function plotWKREF(); https://github.com/henning-winker/FLRef)

4. Simulation-test framework

We develop our simulation-testing framework using the tools available in the Fisheries Library for R (FLR; Kell et al., 2007; https://flr-project.org/). The simulation framework was implemented in the FLR library `mse` (https://github.com/flr/mse) with `FLasher` (https://github.com/flr/FLasher) being used to carry out the forward projections. All Stock Recruitment relationships were conditioned using the FLR package `FLSRTMB` (https://github.com/flr/FLSRTMB). Reference points at equilibrium were calculated with `FLBRP` (https://github.com/flr/FLSRTMB). To facilitate customised reference point estimation and visualisation of F_{brp} ($F_{S\%}$ and $F_{B\%}$), B_{lim}, $F_{p0.5}$, B_{trg}, F_{trg}, we developed the FLR package `FLRef` (https://github.com/henning-winker/FLRef). `FLRef` makes use of the new fast forward projection 'ffwd()` in `FLasher` together with the bisection function `bisect()` in 'mse' to efficiently derive precise values of $F_{P0.5}$ based stochastic simulations. All data and R code used in this analysis will be made available in the github repository of `FLRef`.

4.1 *Stock assessment data*

For simulation-testing, we use a unique dataset of detailed stock assessment outputs for 64 stocks that cover the entire ICES region across the North-East Atlantic, which were collated in the form of objects of the `FLStock` class as defined in FLR. The 64 stocks were sourced from a database that include 77 stocks which are assessed as Category 1 by ICES in 2020 and 2021. In the following, stocks are referred to by ICES stock IDs, with details on the assessment outputs of all stocks in the form of a Shiny Application (https://michaelgras.shinyapps.io/WKREF1), which also comes with a range plots visualising various aspects of each stock's population dynamic and productivity characteristics .

Of the 77 stocks, eight stocks were excluded as MSY reference points are undefined (i.e. cod.27.1-2coastN, cod.27.24-32, san.sa.1r, san.sa.2r, san.sa.3r, san.sa.4, spr.27.3a4 and reb.27.1-2). Further five stocks (sol.27.7e, sol.27.8ab, cod.27.7e–k, her.27.20–24, whg.27.47d) were excluded due to challenges of estimating realistic B_0 or F_{MSY} values (i.e. $F_{MSY} < 1$) within plausible limits during Operating Model conditioning (see below) for the given assessment assumptions, such as natural mortality and selectivity. The final set of 64 `FLStock` objects represent the unified assessment outputs of 12 different age-structured assessments platforms, of which SAM (n = 27; Nielsen and Berg, 2014), Stock Synthesis (n = 9; Methot and Wetzel, 2013) and XSA (n = 8) were the most common. The 64 stocks comprised 23 bony fish species representative of nine taxonomic orders as well as one crustacean, *Pandalus borealis* (pra.27.3a4a). The majority of stocks belonged to the following three taxonomic orders Gadiformes (*n* = 27), Pleuronectiformes (*n* = 14) and Clupeiformes (*n* = 11). Note that there is only one chondrichthyes species (North East Atlantic spurdog) assessed as category 1 by ICES, but the assessment is not included in our database.

We characterised the stocks into low, medium and high productivity categories in accordance with the classification scheme proposed by FAO (2001), using the intrinsic rate of population increase *r* and mean generation time *G* (Table 1). In cases where *r* and *G* resulted in different categories, the lower productivity class was chosen. Productivity is a function of somatic growth, reproduction, survival and longevity. More productive species tend to have high somatic growth, early maturation and short generation times. These life history traits are typically associated with high resilience to growth overfishing and fast rebuilding potential if conditions are favourable. High productivity is therefore often perceived as highly resilient to fishing pressure based on their "ability to rebound after perturbation" (Holling 1973). On the other hand, these traits are often associated with high variability in recruit and fewer mature fish to buffer against sequential recruitment failure, which can make them more vulnerable to recruitment overfishing and risk stochastic depletion, even under light fishing pressure (Thorson *et al.*, 2015).

A direct indicator for productivity is *r*, which summarizes several key life history traits into a single metric (Musick, 1999) (Musick, 1999). FAO (2001) suggested the mean generation time *G* as an additional indicator, which quantifies the turnover time of generations and is widely considered for setting targets for rebuilding plans. Both *r* and *G* can be directly derived from a Leslie Matrix (McAllister *et al.*, 2001; Thorson, 2020), which requires weight, maturity, and *M*-at-age from the `FLStock` objects as well as an estimate of recruitment compensation in the form of the steepness *s* of the stock recruitment relationship. We implemented the Leslie matrix generic tool in the R package FLSRTMB (https://github.com/flr/FLSRTMB) and provide details on methods in Appendix B. Stock specific steepness *s* were derived as the expected means from the hierarchical taxonomic FishLife model (Thorson, 2020; https://github.com/James-Thorson/FishLife), which are summarised in **Table B1**. Most stocks fell into the medium productivity category (n = 37), followed by low productivity stocks (n = 17), and high productivity stocks (n = 10).

Table 1: Guidelines used for categorising productivity levels for exploited fish species. Criteria for intrinsic rate population increase *r* are from Musick (1999) and value of Generation Time *G* are adopted from FAO (2001). In cases where *r* and *G* resulted in different categories, the lower productivity class is chosen.

Parameter	Productivity		
	Low	Medium	High
Intrinsic population Growth *r*	< 0.15	0.15 - 0.5	> 0.5
Generation Time *G*	> 10	5 -10	< 5

4.2 Conditioning of Operating models

Operating Models were implemented as single sex and single fleet models with an annual time step. Future projections were run over 60 years (i.e. 2021-2080) with 250 iterations and based on the 3-years average of the most recent data years for weight-at-age (w_a), maturity-at-age (mat_a), natural mortality-at-age (M_a) and the F_a pattern determining the selectivity-at-age (s_a). This choice was made to account for non-stationary processes in these quantities. The performance evaluations were based on the last 10 years of the 60-year projection horizon (i.e. 2071-2080).

For the simulation testing, a generic Beverton-Holt model (BH-SRR) was assumed for all stocks. The recruitment deviation is assumed to be associated with a first-order autocorrelation (AR1) process and a function of recruitment standard deviation σ_r and the AR1 coefficient ρ (Johnson et al. 2016). To ensure an objective and unified approach representative over the wide range of life histories across the 64 stocks, species-specific predictive distributions for steepness *s* were and expected means for σ_r and ρ were sourced from the hierarchical taxonomic FishLife model to fit a Beverton-Holt (BH) to the SR data and generate the recruitment deviations, respectively (Thorson, 2020; https://github.com/James-Thorson/FishLife).

The parameters of stock-recruit curves are notoriously difficult to estimate, and often little inference can be made from a single stock-recruit fit, but meta-analysis and the use of distributions as a Bayesian prior can provide a useful starting point from which meaningful updates could occur. This approach of using prior information to condition the SR to the data, is consistent with discussions and suggestions for future work in WKMSYREF2 (ICES, 2014). Instead of assuming that nothing is known, other than the information that is contained in the stock data alone, this approach assumes that at least within taxonomic groupings (family, species) information from one stock can provide some useful prior information about the recruitment compensation for another (Myers *et al.*, 1999; Thorson, 2020). For stocks with few years of SR data, or where the observations appear uninformative, priors can assist in making less spurious inference about the SR, whereas if the SR data are informative, so that the priors are effectively updated by the data.

The Beverton-Holt SSRs were fitted to S-R data using the R package FLSRTMB (https://github.com/flr/FLSRTMB), which implements a re-parameterised of the BH SR as a function of steepness s and annual unfished spawning biomass per-recruit SPR_0 to accommodate the integration of priors for s (Thorson, 2020). A notable difference to the conventional parameterization is that SPR_{0_y} is treated as non-stationary, being a function of annual quantities of $W_{a,y}$, $Mat_{a,y}$ and $M_{a,y}$. By way of using time-varying $SPR_{0,y}$, it also takes into consideration the recent criticism by Miller and Brooks (2021) that specifying a set biological parameters to define a single time-invariant SPR_0 can be highly sensitive to reference estimation when using steepness values from meta-analysis (See Appendix I for details).

4.3 Implementation system

To facilitate comparability of the tested reference point systems, all considered harvest control rules (HCRs) are kept generic and in the same form of the conventional ICES Advice Rule (ICES, 2021d), where the F

advice decreases from F_{trg} to zero and $B_{trigger}$ and zero SSB (Figure 1). Variations of the tested HCRs are therefore determined by the parameters F_{trg} and $B_{trigger}$.

The HCRs were implemented using a simulated feedback control loop between the implementation system and the operating model, where the implementation system translates the assessment outcome via the HRC into the Total Allowable Catch (TAC) advice (Figure 2). The key difference to a simple stochastic risk simulation, such as EQSIM, is the simulated feedback control loop between the implementation system and the operating model allows accounting for the lag between the last of year data used in the assessment and the implementation year of TAC advice. In ICES, the implementation system of the harvest control rule is based on the assumption that advice is given for year y+1 based on an assessment completed in year y, which is typically fitted to data up until last data year y-1 (ICES, 2020b). Therefore, implementation of the TAC derived through HCR requires projection of the stock dynamics by way of a short-term forecast (Mildenberger *et al.*, 2021). To do this, numbers-at-age were projected through the year of assessment. Status quo recruitment, M_a, w_a and mat_a were set as the mean of the last 3 years. A projection based on a fixed fishing mortality-at-age to the last year (y-1) in the assessment is then made through to the implementation year (y+1).

In contrast to a full Management Strategy Evaluation (MSE) simulation design (Punt et al. 2017), this MSE ‘shortcut’ approach (e.g. ICES, 2020 WKGMSE3), omits the step of the annual updating of the estimation model (assessment) in the feedback control. Instead, it passes the 'true' age-structured dynamics from the OM (or with assumed some error) to the HCR implementation. For testing the robustness of reference point systems across a large number of stocks the merits of a short-cut MSE approach include: (1) the straight-forward implementation using the tools available in ‘FLR’ (Kell et al., 2007), i.e. ‘mse’ and ‘FLasher’, (2) reduced computation time, (3) data requirements are limited the available assessment outputs (FLStock class object) without the need of sourcing auxiliary data to recondition the assessment models, and (4) the incorporation of the lag effect between data, assessment and management implementation.

The limitations of the MSE short-cut approach are that it cannot fully account for uncertainties resulting from imperfect sampling of the full age-structure (e.g. poorly sampled recruits), observation error, model estimation error, misspecified assumptions about the biology (M_a, w_a or mat_a) and selectivity. Therefore, robustness testing is limited here to the structural uncertainty about the externally fitted SR, which determines the stock’s recruitment compensation and the absolute scale of R_0, with direct impacts on reference points such as F_{MSY}, B_{MSY}, MSY, B_0, B_{trg} or $B_{trigger}$.

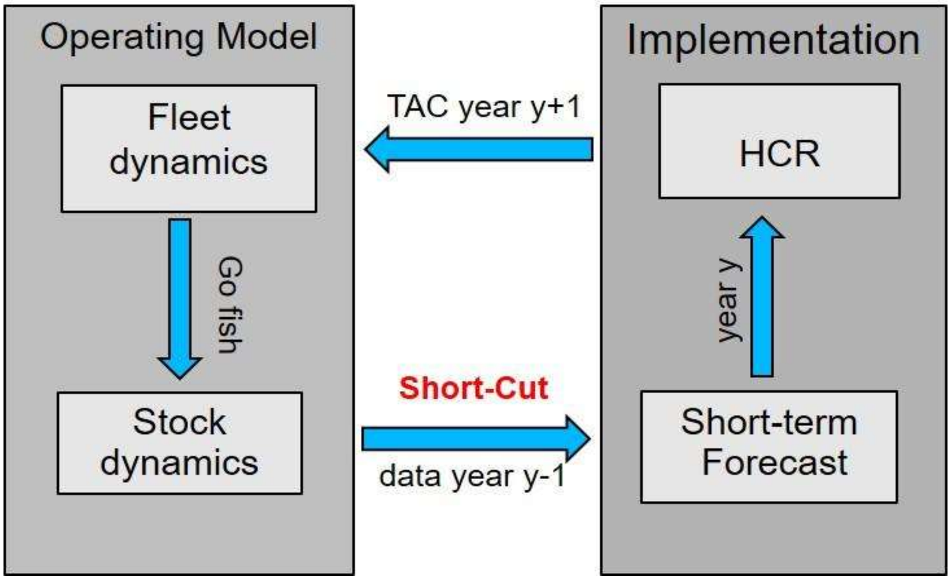

Figure 2: Schematic illustrating the key processes of the short-cut approach to MSE, showing the Operating model that simulates the fishery and stock dynamics on the left and Implementation System including the short-term forecast on the right. The short-cut denotes the omission of the estimation (stock assessment) model that updates to new observations (with estimation error) in conventional MSE implementations with full feedback control loop.

4.3 Performance Evaluation Criteria

The consistency tests were designed to identify the generic rules for specifying F_{brp}, B_{trg} and $B_{trigger}$ according to stock-specific productivity that provide the optimal trade-offs among the following three main objectives: (1) to not exceed a 5% probability of SSB falling below B_{lim} in any single year (2) to achieve high long-term yields that correspond to at least 95% of the median long-term yield attained by fishing at F_{MSY} (MSY), (3) to attain at least 50% probability that SSB is above B_{thresh} set at 80% of B_{trg}. Consistent with the objectives of ICES advice framework (ICES, 2020d), the three objectives are interpreted hierarchically in that objective (1) is the overriding criteria of maintaining stock size above B_{lim} with at least 95% probability to be compliant with the ICES PA. Conditional on objective (1), objective (2) is based on the ICES definition for using plausible values around F_{MSY} in the advice rule, which are derived so that they lead to no more than a 5% reduction of MSY obtained by fishing at F_{MSY} in the long term. The B_{tresh} in objective (3) replaces the current MSY $B_{trigger}$ threshold (which is normally set to 1.4B_{lim}) for classifying stock size to be at biomass levels that can produce MSY (green).

For this performance evaluation, F_{trg} was set to F_{brp}, but we also analysed how often F_{P05} would be invoked based on the specifications of the OM (see Section 6.3). To set B_{lim}, we considered both estimators for Type

1 and Type 2. To derive Type 1 B_{lim}, a generic continuous Hockey-Stick (HS) model was fitted to the SR data (Appendix B). In absence of contrast in a large proportion of S-R dataset, the HS was constrained to assure that B_{lim} falls within a range of $0.1B_0 < B_{lim} < 0.3B_0$ to ensure that Type 1 B_{lim} was estimated within plausible biological limits. Within these constraints B_{lim} is estimated by the breakpoint $b = B_{lim}$, while R_0 is given by the product of the slope *a* and *b* (see details in Appendix B). Type 2 B_{lim} was derived as the 10% of B_0, where B_0 is the equilibrium estimate under F = 0 based on the "true" SR of the OM and average stock biology over the most recent three years. Regressing the so derived Type 1 and Type 2 B_{lim} values against each other on log-scale showed notable variation (CV = 40%) among the 64 stocks but indicated no systematic divergence from a 1:1 relationship. Type "P3" probability was applied to compute the risk for the biomass limits as the maximum of annual probabilities. The performance statistic for MSY was quantified as long term median yield obtained when fixing F_{trg} of HCR to the "true" F_{MSY} of the OM.

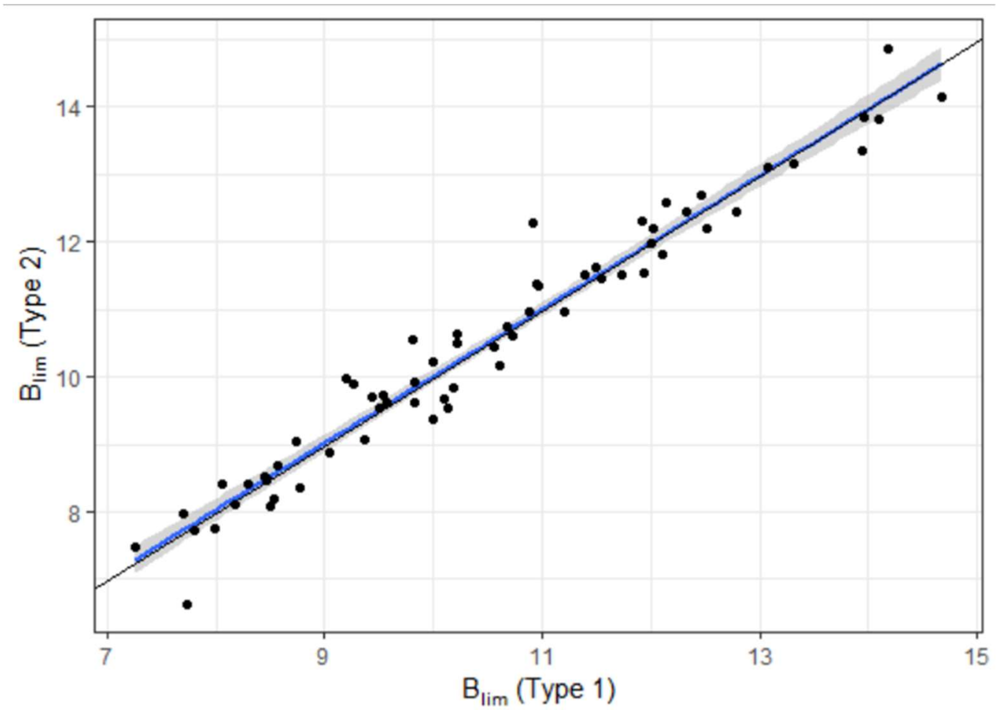

Figure 3. Relationship between Type 1 and Type 2 on a log-scale (CV=40%)

For the robustness evaluation, we retain objectives (1) and (2) as performance criteria, but instead of B_{tresh} from objective 3, we used the "true" B_{MSY} from the OM as the third performance criteria. This allowed us to evaluate if the underlying goal to restore and maintain stocks above average levels that can produce MSY is achieved by the selected candidates reference point systems that were most consistent in meeting the objectives. To test the robustness of the selected "WKREF" candidate reference point systems, we considered two scenarios for violating the assumptions about the SSR with respect to the "true" functional form of the OM. These were: (1) a Beverton Holt SRR, but fitted without informative priors about *s* and (2) the continuous Hockey-Stick SSR (Appendix B). This effectively achieved various extents of misspecification of the SRR and the associated production function across the 64 stocks (Supplement 1). For reference, we also compare the performance of "WKREF" candidate reference systems to: (1) the current ICES advice rule, by setting the official 2021 ICES benchmarks of F_{MSY} as F_{trg} and MSY $B_{trigger}$ as $B_{trigger}$, (2) the New Zealand Standard, and using directly the estimates of F_{MSY} and B_{MSY} to specify F_{trg} and B_{trg}, respectively (See Table 3).

5. *Results*

5.1 Results of Self-test consistency

A total of 32 scenarios in two 4 x 4 grids were tested. The first grid comprises F_{SPR} that ranges from 35 to 50% and the second grid ranges of F_B ranged from 30 to 45%. These F_{brp} ranges were tested in both grids with alternative $B_{trigger}$ set equal to 0.6, 0.8 and 1 × B_{trg} and 2 × $B_{lim,}$ where type 1 B_{lim} was in this case used to estimate $B_{trigger}$.

For low productivity stocks, all tested F_{brp} proxies for F_{MSY} are precautionary with a less than 5% probability of SSB falling below B_{lim} (Figure 4). This is irrespective of how B_{lim} is set (i.e. Type 1 or Type 2) or which fraction of B_{trg} is used to determine $B_{trigger}$. Medium productive species showed higher risk of falling below Type 1 than Type 2 B_{lim}. In accordance with the PA, F_{trg} needs to be set at $F_{B35\%}$ or F_{SPR40} (or larger) in combination with a trigger of at least 80% of B_{trg} or 2 × B_{lim}. In contrast to $F_{B\%}$ proxies, the 5% risk threshold for Type 1 B_{lim} was exceeded for some medium productivity stocks at $F_{SPR40\%}$. In contrast to Type 1 B_{lim}, all F_{brp} proxies consistently met the precautionary objective for low and medium productive species, with the exception of $F_{SPR35\%}$ in combination with the $B_{trigger}$ set at 0.6 B_{trg}. High productivity species were associated with substantially higher risk to fall below B_{lim}, with comparable levels of $F_{SRP\%}$ being substantially more risk prone than $F_{B\%}$. Like for medium productivity, Type 1 B_{lim} was associated with a higher risk than Type 2 B_{lim}. Consistency with the precautionary objective, was only achieved for $F_{B\%40}$ or combinations F_{SRP50} and $B_{trigger}$ set to B_{trg} or higher.

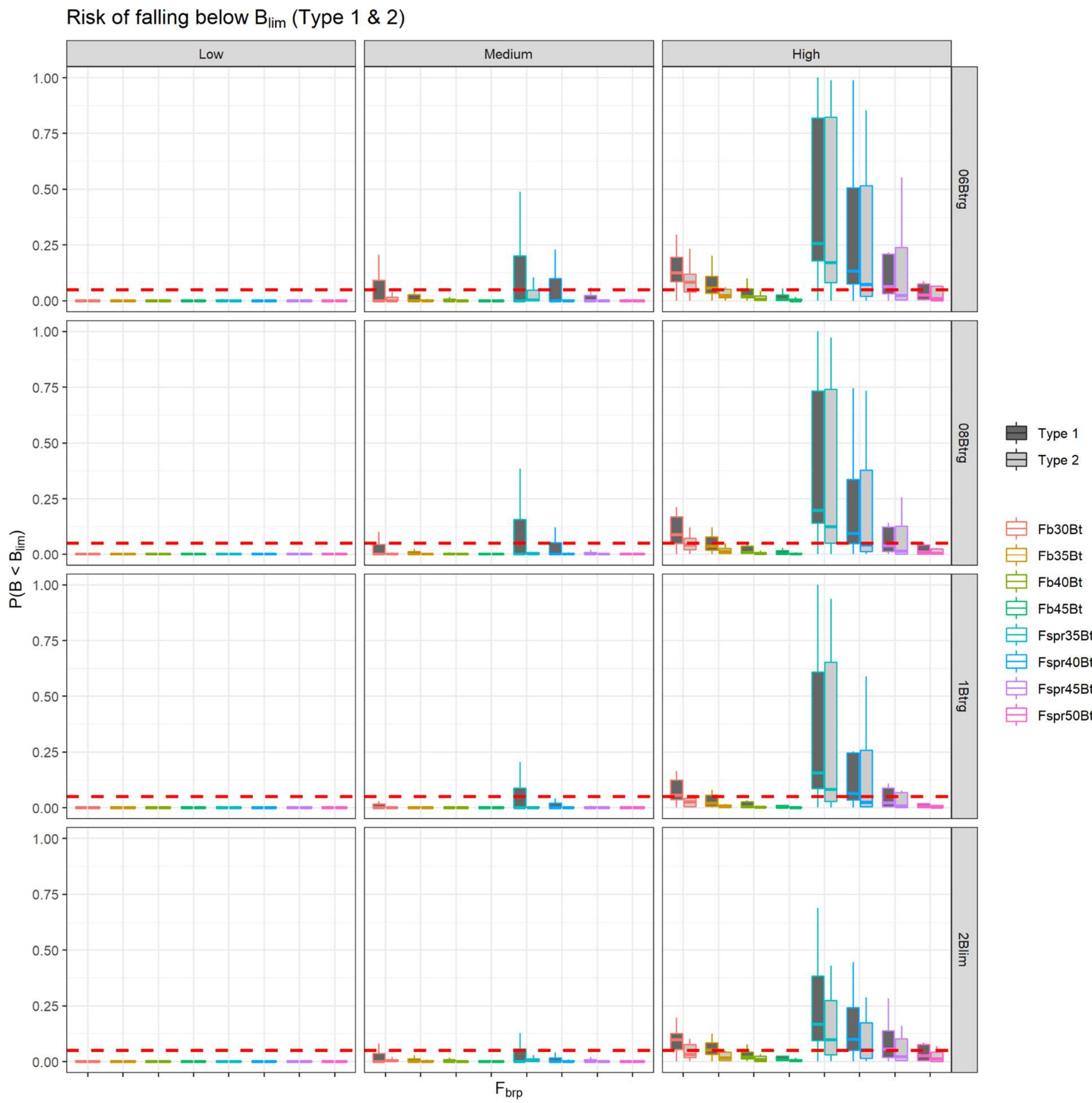

Figure 4: Self-test consistency evaluations of the type 3 risk probability (P3) that SSB falls below B_{lim} shown for low, medium and high productivity stocks (columns) across colour-coded ranges for $F_{SPR\%}$ of 35-50% and $F_{B\%}$ of 30-45 in combinations with alternative $B_{trigger}$ values of fractions of 0.6, 0.8, 1 B_{trg} and a multiplier of 2 x B_{lim} (rows). The red dashed line denotes the limit threshold of a 5% probability in accordance with ICES Precautionary Approach.

For low and medium productivity stocks, highest long term catches in excess of 95% MSY are obtained with F_{brp} proxies specified at levels of 30 - 35% for $F_{B\%}$ and 40 - 45% for $F_{SPR\%}$ in combination with $B_{trigger}$ between 0.8 and 1.0 B_{trg} or at 2 × B_{lim} (Figure 5). The situation is very different for high productivity stocks (e.g. sardine, sprat), for which more conservative proxies of $F_{B\%}$ and $F_{SPR\%}$ lead to increased yield. Here, highest long term catches in excess of 95% MSY are obtained with F_{brp} equal to F_B 40 to 45% and F_{SPR} 45 to 50% in combination with $B_{trigger}$ between 0.8 and 1.0 B_{trg} or equal to 2 × B_{lim}.

The results of the self-test showed that the probability of exceeding B_{tresh} (at 80% B_{trg}) increases by setting $B_{trigger}$ closer to B_{trg}. However, for low and medium productivity stock high $B_{trigger}$ values indicate yield loss and thus creates a conflict with the objective to optimise long yield. High productivity stocks, by contrast, indicated no conflicts among the objects of optimising yield, exceeding B_{tresh} and minimizing the risk of falling below B_{lim}, with optimal trade-off being achievable with more conservative combinations F_{brp} and $B_{trigger}$. Setting $B_{trigger}$ equal to 2 × B_{lim} performs generally similar in terms of the yield and risk objectives when compared to optimal setting of $B_{trigger}$ to 0.8 B_{trg} for low and medium productivity stocks and equal to B_{trg} for high productivity stocks. However, in particular for medium and high productivity species, the probability to exceed B_{tresh} is notably lower when $B_{trigger}$ is set 2 × B_{lim} associated with large variations among species. Setting $B_{trigger}$ to 2 × B_{lim} is therefore associated with increased risk of inconsistent stock status classification, which can be minimised by setting $B_{trigger}$ relative B_{trg}.

Based on these results, we chose to specify the candidate reference points for further robustness testing using $B_{trigger}$ equal to 0.8B_{trg} for low and medium productivity stocks and $B_{trigger}$ equal to B_{trg} for high productivity stocks (Table 3)

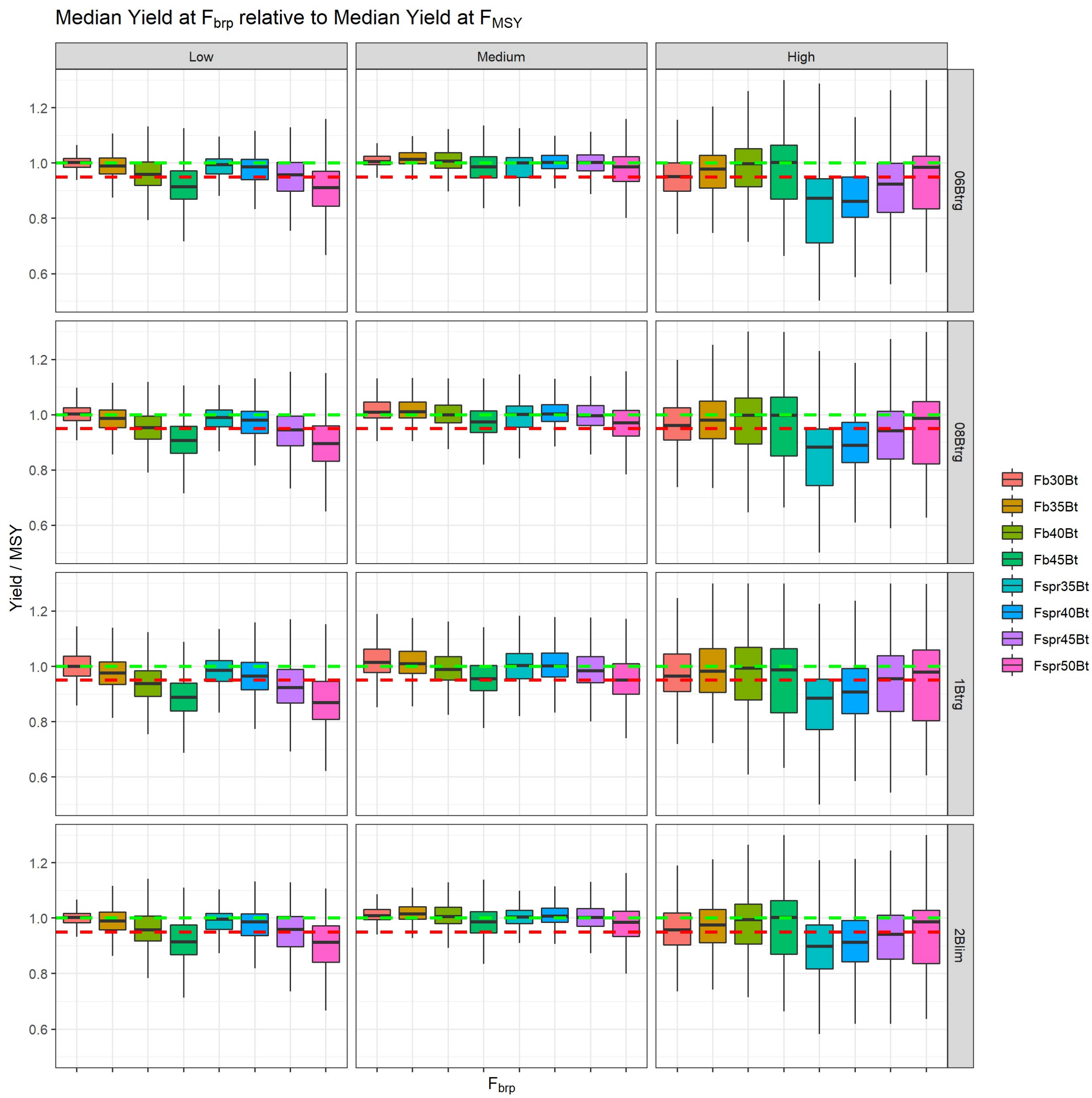

Figure 5: Self-test consistency evaluations of the median long term yield relative the median long-term obtained at fixed "true" F_{MSY} (MSY) shown for low, medium and high productivity stocks (columns) across colour coded ranges for $F_{SPR\%}$ of 35-50% and $F_{B\%}$ of 30-45 in combinations with alternative $B_{trigger}$ values of fractions of 0.6, 0.8, 1 B_{trg} and a multiplier of 2 x B_{lim} (rows). The green dashed line denotes a 1:1 ratio of long term Yield/MSY and the red dashed line denotes the yield threshold at 95% MSY.

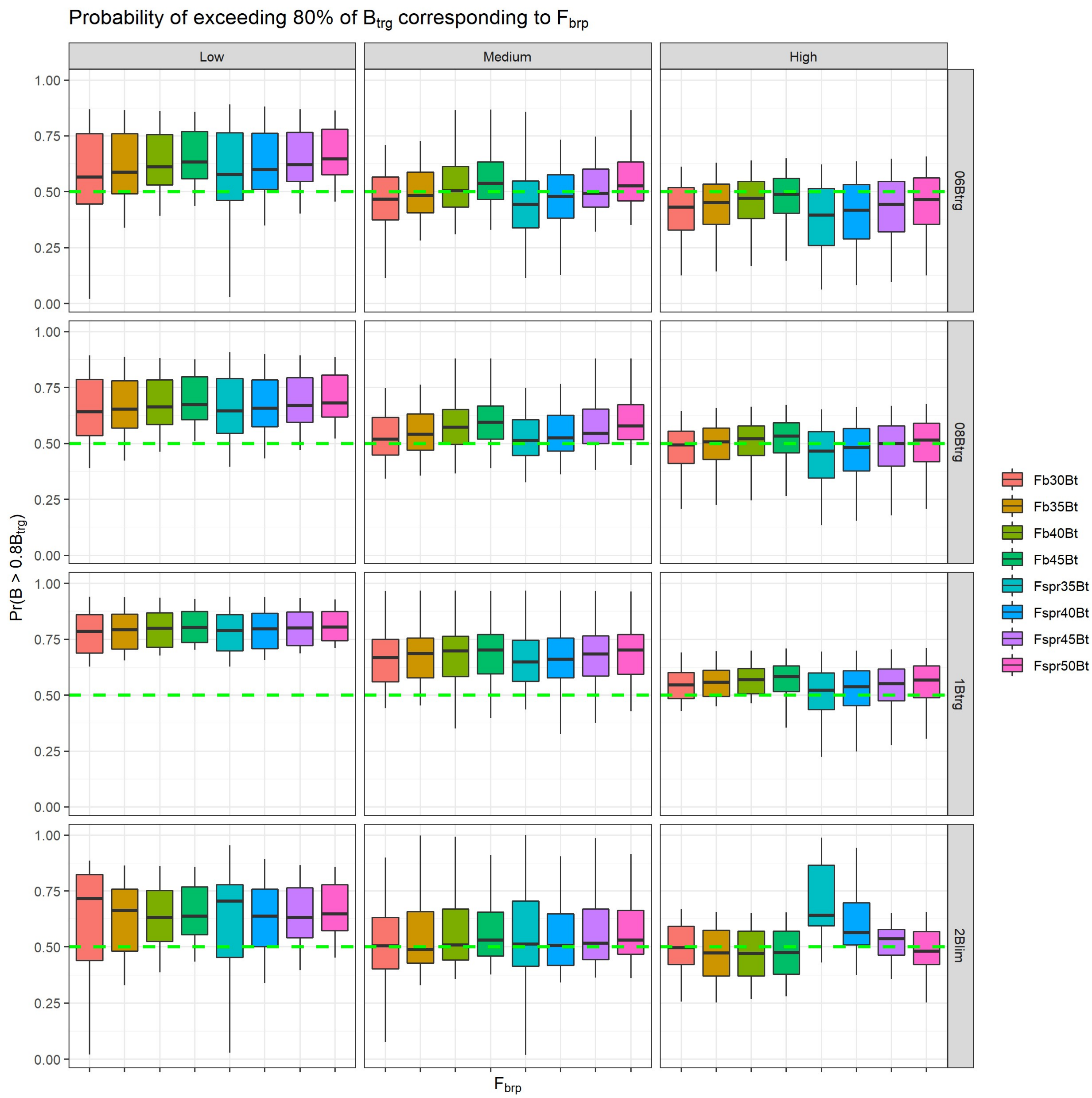

Figure 6: Self-test consistency evaluations of the probabilities that SSB exceeds B_{tresh} at 80% B_{trg} shown for low, medium and high productivity stocks (columns) across colour coded ranges for $F_{SPR\%}$ of 35-50% and $F_{B\%}$ of 30-45 in combinations with alternative $B_{trigger}$ values of fractions of 0.6, 0.8, 1 B_{trg} and a multiplier of 2 x B_{lim} (rows). The green dashed line denotes a 50% probability threshold of exceeding B_{tresh}.

5.2 Results from robustness tests

Details on the specifications of the reference point systems considered for robustness tested are presented in Table 3, together acronyms used hereafter. The candidate reference point systems that showed the best performance in the self-tests are referred to as "WKREF".

Table 3: Specifications of alternative reference point system evaluated by robustness testing. SRR: Stock-recruitment recruitment relationship; BH: Beverton and Holt; HS: Hockey-Stick.

Advice Rule	Productivity	F_{trg}	B_{trg}	$B_{trigger}$	SRR	Acronyms
ICES	-	Advice	-	Advice	N/A	ices.ar
F_{MSY}	All	F_{MSY}	B_{MSY}	0.8 *x* B_{trg}	BH	fmsy.bh
New Zealand	Low	F_{SPR45}	B_{40}	min(1-*M*, 0.5)	BH	
	Medium	F_{SPR40}	B_{35}	min(*1*-*M*, 0.5)	BH	nz.bh
	High	F_{SPR35}	B_{30}	min(*1*-*M*, 0.5)	BH	
WKREF1 SPR%	Low	F_{SPR40}	B_{SPR40}	0.8 *x* B_{trg}	BH / HS	fspr.bh / fspr.hs
	Medium	F_{SPR40}	B_{SPR40}	0.8 *x* B_{trg}	BH / HS	
	High	F_{SPR50}	B_{SPR50}	1 *x* B_{trg}	BH / HS	
B%	Low	F_{B35}	B_{35}	0.8 *x* B_{trg}	BH / HS	fb.bh / fb.hs
	Medium	F_{B35}	B_{35}	0.8 *x* B_{trg}	BH / HS	
	High	F_{B40}	B_{40}	1 *x* B_{trg}	BH / HS	

The ices.ar was found to be the least robust compared to any other tested reference point systems (Figure 7). For low and medium productivity stocks, the risk of falling below either of the two B_{lim} types was substantially higher, yield and SSB were on average lower. By contrast, the ices.ar was among the more precautionary reference point systems for high productivity stock. Similarly, the use of direct estimates F_{MSY} as F_{trg} in fmsy.bh performed generally poorer than the F_{brp} proxies in the WKREF candidates for low and medium productivity stocks, but also improved notably for high productivity stocks. Direct estimates fail in particular to achieve SSB levels at or below B_{MSY} for low and medium productivity species and the risk of

SSB falling below B_{lim} is above the 5% threshold is relatively high, in particular for medium productivity stocks. The fmsy.bh system performs comparably better for high productivity stocks.

Except for the ices.ar, all tested systems were robust to risk of SSB falling below B_{lim} for low productivity species (Figure 7). For medium productivity species, the WKREF fsb.bh was the only candidate system that was fully compliant with the PA for Type 1 B_{lim}, whereas for Type 2 B_{lim}, this also included the nz.bh and WKREF systems. For high productivity stocks, the best performing systems in terms of risk are fspr.bh, fspr.hs and fb.bh, while the nz.bh performed poorly with respected to the PA. This can be explained in that $F_{B\%}$ tends to be notable smaller than its equivalent $F_{spr\%}$ when the production function is based on Beverton-Holt SRR but equal for a Hockey-Stock. Therefore, the specifications fb.hs led to consistently higher F_{brp} (i.e. proxies (i.e. $F_{B30} = F_{SPR30}$), which then led poorer performance in the robustness tests.

Among the WKREF candidates, differences in long term yields are small for low and medium productivity species, with all medians exceeding the 95% MSY threshold and generally low yield variation among stocks. For high productive species largest median yields are attained with fspr.hs. The results indicate that fspr.bh, fsb.bh and fspr.hs lead to median SBB levels at or above $B_{MSY.}$ The exception is fspr.bh, which was generally the least robust of the tested WKREF candidates (Figure 7).

With respect to taxonomic orders, the WKREF candidates performed particular well for pleuronectiformes (flatfishes), which fells within medium productivity group (Figure 8). Pleuronectiformes showed negligible risk of SSB falling below B_{lim}, long-term yields at or above MSY and median SSB at B_{MSY}. With respect to yield and B_{MSY}, similar good performance was achieved for gadoids although some stocks are associated with higher risk to fall B_{lim}. Stocks of the order Clupeiformes showed a similar high risk profile as the high productivity stocks, but generally better performance in terms yield. Maintaining stock levels close to B_{MSY} was only achieved with fb.bh, fmsy.bh and current the ices.ar, albeit the latter with larger variation.

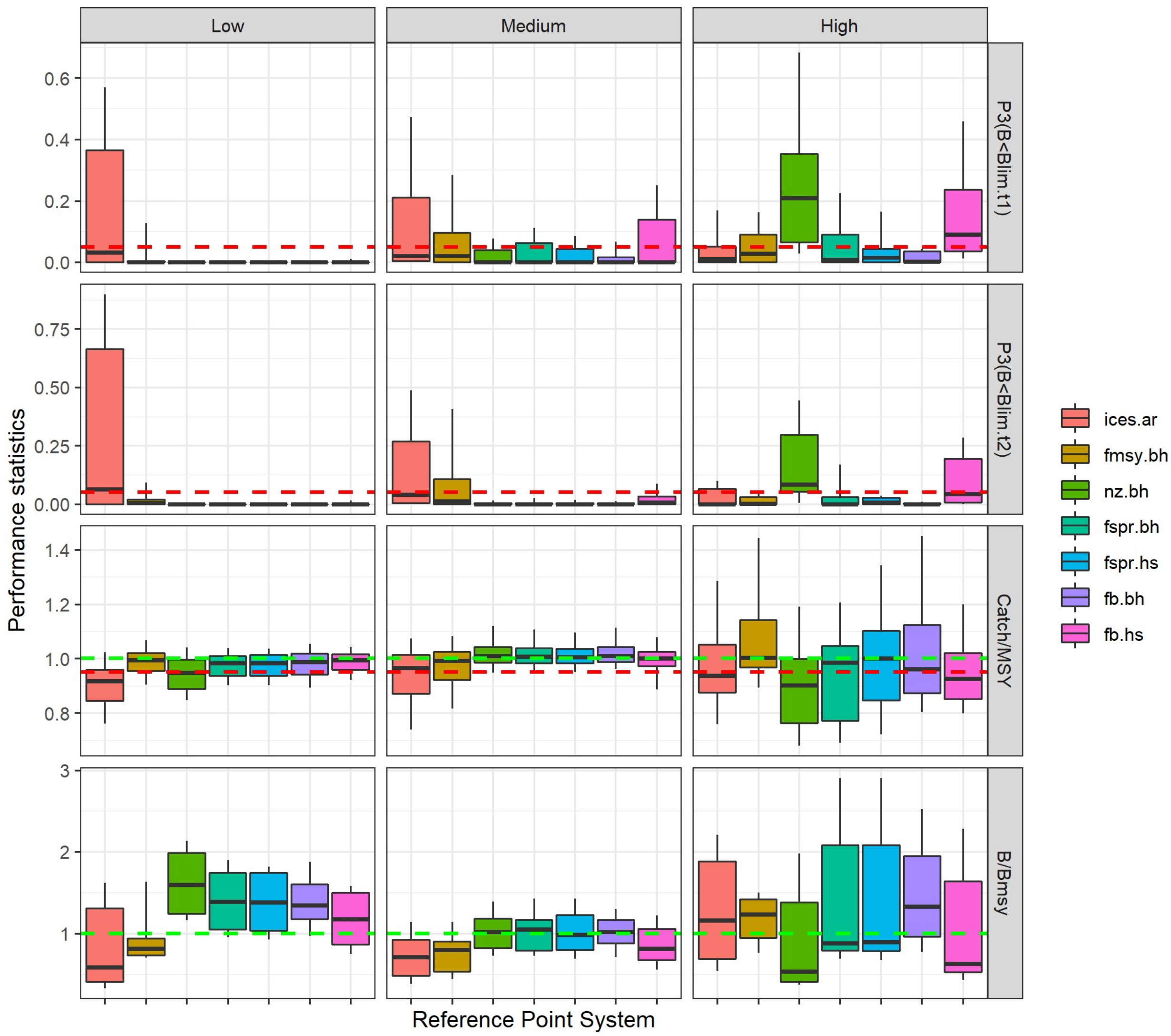

Figure 7: Results of robustness tests of evaluate reference point systems, showing the type 3 risk probabilities (P3) of SSB falling below B_{lim} of Type 1 (top row) and Type 2 (2nd row), the median long term yield relative the median long-term obtained at fixed "true" F_{MSY} (MSY) (3rd row) and the probabilities of SSB exceeding B_{tresh} at 80% B_{trg} (bottom row) for low, medium and high productivity stocks (columns). Green and red dashed lines denoting the target and limit thresholds, respectively. *ices.ar*: ICES Advice Rule; *fmsy.bh*: HCR with $F_{trg} = F_{MSY}$ and $B_{trigger} = 0.8\ B_{MSY}$; *nz.bh*: New Zealand Harvest Standard; *fspr.bh*/*.hs*: WKREF1 candidate based on $F_{SPR\%}$ and *fspr.bh*/*.hs:* WKREF1 candidate based on $F_{B\%}$, where *.bh* and *.hs* denotes if a fitted Beverton-Holt or Hockey-Stick was used, respectively (See Table 3 for details).

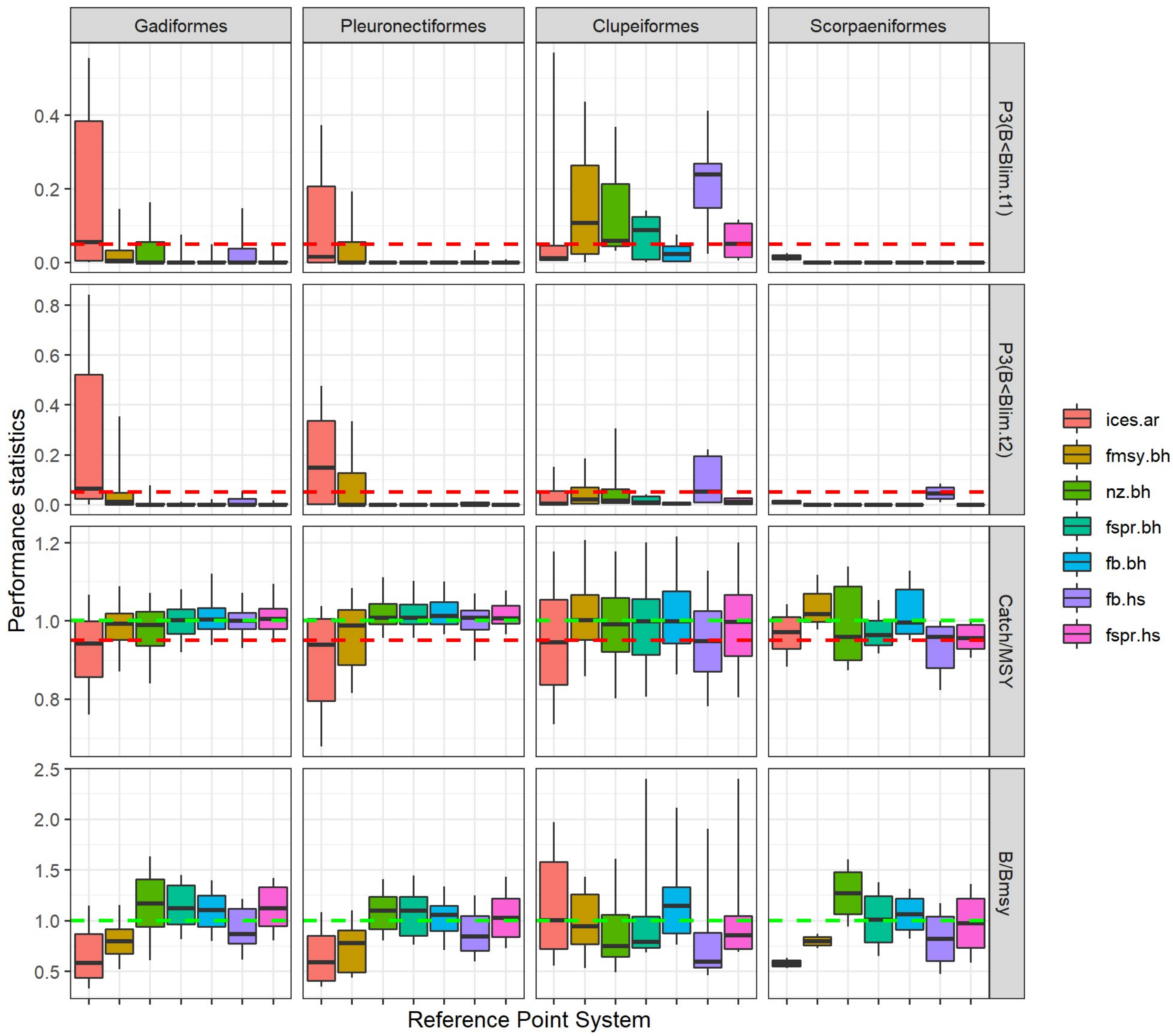

Figure 8: Results of robustness tests of evaluated reference point systems, showing the type 3 risk probabilities (P3) of SSB falling below B_{lim} of Type 1 (top row) and Type 2 (2nd row), the median long term yield relative the median long-term obtained at fixed "true" F_{MSY} (MSY) (3rd row) and the probabilities of SSB exceeding B_{tresh} at 80% B_{trg} (bottom row) for stock of four selected taxonomic orders, (columns). Green and red dashed lines denoting the target and limit thresholds, respectively. *ices.ar*: ICES Advice Rule; *fmsy.bh*: HCR with $F_{trg} = F_{MSY}$ and $B_{trigger} = 0.8\ B_{MSY}$; *nz.bh*: New Zealand Harvest Standard; *Fspr.bh*/*.hs*: WKREF1 candidate based on $F_{SPR\%}$ and *Fspr.bh*/*.hs:* WKREF1 candidate based on $F_{B\%}$, where *.bh* and *.hs* denotes if a fitted Beverton-Holt or Hockey-Stick was used, respectively (See Table 3 for details)

5.3. Invoking the precautionary fishing mortality target $F_{P0.5}$

Based on the SRR of OM we estimated B_{lim} for Type 1 and 2 and used the bisection function in FLFlasher to determine $\mathbf{F_{P.05}}$. As shown in Figure 7, the $\mathbf{F_{P.05}} < F_{SPR\%}$ was invoked for 16% for Type 1 B_{lim} and 6% for Type 2 B_{lim}. $\mathbf{F_{P.05}} < F_{B\%}$ was invoked for 8% of the stocks when using Type 1 B_{lim} but it was never invoked for Type

2 B_{lim}. High and medium productivity stocks were similarly likely to invoke $\mathbf{F_{P.05}}$ rule for Type 1 B_{lim}, whereas this was reduced for medium productivity stocks when Type 2 B_{lim} was used. In total only 10 stocks invoked $\mathbf{F_{P.05}}$ for any of the B_{lim} and F_{brp} combinations. These included six of the herring stocks.

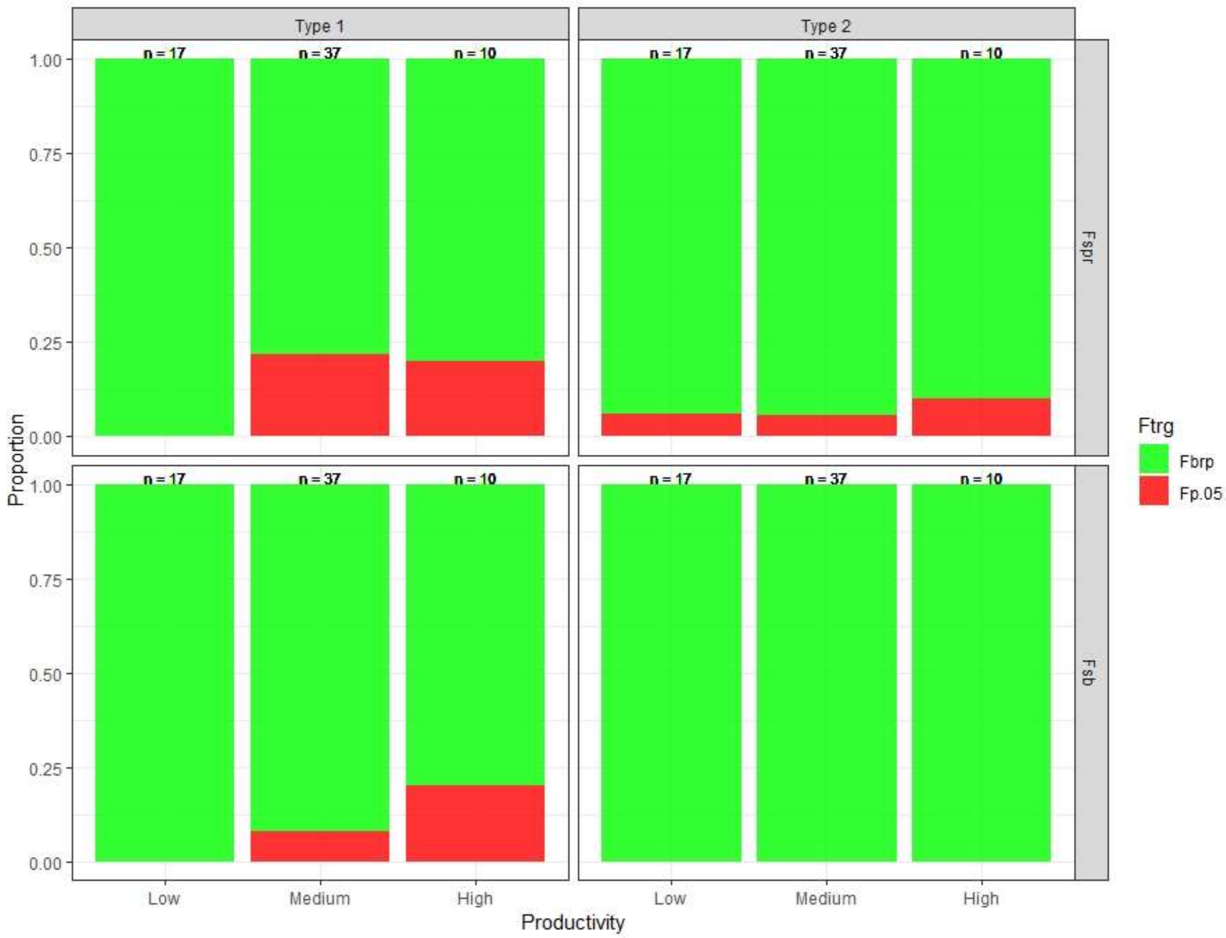

Figure 9: Proportion of stocks triggering $F_{P.05}$ for the different category of productivity when using $F_{SPR\%}$ or $F_{B\%}$

6. Recommendations

The results of both the self-test and robustness test clearly highlights the need to consider the stock's biological and productivity for setting reference points. Based on these results the following guidelines for setting reference points for category 1 stocks assessed by ICES are proposed according to productivity category (Table 4):

Table 4: Guidelines for deriving target and trigger reference points in the newly proposed ICES system. The Type 1 and 2 approaches can be used for all stocks to derive B_{lim}. SRR: Stock-recruitment recruitment relationship; BH: Beverton and Holt; HS: Hockey-Stick.

	Productivity	F_{trg}	B_{trg}	$B_{trigger}$	SRR
SPR%	Low	F_{spr40}	B_{spr40}	0.8 *x* B_{trg}	BH/HS
	Medium	F_{spr40}	B_{spr40}	0.8 *x* B_{trg}	BH/HS
	High	F_{spr50}	B_{spr50}	1*x* B_{trg}	BH/HS
B%	Low	F_{B35}	B_{35}	0.8*x* B_{trg}	BH
	Medium	F_{B35}	B_{35}	0.8*x* B_{trg}	BH
	High	F_{B40}	B_{40}	1*x* B_{trg}	BH

Low productive species $F_{SPR40\%}$ with stock and recruitment modelled as BH or HS fulfils both the PA and the MSY criteria and is proposed as candidate for the future ICES system to derive TRP. $F_{B35\%}$ with stock and recruitment modelled as BH fulfils both the PA and the MSY criteria and is proposed as candidates for the future ICES system to derive TRP. B_{lim} can be derived as the newly proposed Type 1 or Type 2, B_{trg} is the SSB that corresponds to $F_{SPR40\%}$ or $F_{B35\%}$ and $B_{trigger}$ is set at 0.8 B_{trg}.

Medium productive species $F_{SPR40\%}$ with stock and recruitment modelled as Beverton-Holt or Hockey-Stick SRR fulfils both the PA and the MSY criteria and is proposed as candidates for the future ICES system to derive TRP. $F_{B35\%}$ in combination with a Beverton-Holt SRR fulfils both the PA and the MSY criteria and is proposed as candidates for the future ICES system to derive TRP. B_{lim} can be derived as the newly proposed Type 1 or Type 2, B_{trg} is the SSB that corresponds to $F_{SPR40\%}$ or $F_{B35\%}$ and $B_{trigger}$ is set at 0.8 B_{trg}.

High productive species $F_{SPR50\%}$ with stock and recruitment modelled as BH or HS fulfils both the PA and the MSY criteria and is proposed as candidates for the future ICES system to derive TRP. $F_{B40\%}$ with stock and recruitment modelled as BH fulfils both the PA and the MSY criteria and is proposed as candidates for the future ICES system to derive TRP. B_{lim} can be derived as the newly proposed Type 1 or Type 2, B_{trg} is the SSB that corresponds to $F_{SPR50\%}$ or $F_{B40\%}$ and $B_{trigger}$ is set equal to B_{trg} or higher.

The Type 1 and 2 approaches can be used for all stocks to derive $\mathbf{B_{lim}}$ where Type 1 relies on the existence of a discernible relationship between stock and recruitment in that the data show contrast and a breakpoint is clearly defined. The $F_{B\%}$ guidelines should not be used in combination with Hockey-Stick SRR. In all cases it is recommended to estimate $\mathbf{F_{P.05}}$ although with the exception of herring, the newly proposed set of reference points should very rarely trigger $\mathbf{F_{P.05}}$.

References

Beverton, R. 1998. Fish, Fact and Fantasy: a Long View. Reviews in Fish Biology and Fisheries, 8: 229–249. https://doi.org/10.1023/A:1008888411100.

Carruthers, T., Kell, L., and Palma, C. 2017. Accounting for uncertainty due to data processing in virtual population analysis using Bayesian multiple imputation. Canadian Journal of Fisheries and Aquatic Sciences, 75: 883–896. NRC Research Press. https://doi.org/10.1139/cjfas-2017-0165.

de Bruyn, P., Murua, H., Aranda, M., 2013. The precautionary approach to fisheries management: how this is taken into account by Tuna regional fisheries management organisations (RFMOs). Mar. Policy 38, 397–406. http://dx.doi.org/10.1016/j.marpol.2012.06.019.

Deroba, J. J., Butterworth, D. S., Methot Jr, R. D., De Oliveira, J. A. A., Fernandez, C., Nielsen, A., Cadrin, S. X., *et al.* 2015. Simulation testing the robustness of stock assessment models to error: some results from the ICES strategic initiative on stock assessment methods. ICES Journal of Marine Science, 72: 19–30. https://doi.org/10.1093/icesjms/fst237.

DFO. 2009. A fishery decision-making framework incorporating the precautionary approach. https://www.dfo-mpo.gc.ca/reports-rapports/regs/sff-cpd/precaution-eng.htm.

FAO. 1995. Code of Conduct for Responsible Fisheries. FAO Technical Guidelines for Responsible Fisheries, 4 (Suppl.: 1–112. Rome.

FAO. 2001. Second technical consultation on the suitability of the CITES criteria for listing commercially exploited aquatic species. Windhoek, Namibia. https://www.fao.org/3/y1455e/y1455e.htm.

FAO. 2020. The State of World Fisheries and Aquaculture (SOFIA). Rome. https://www.fao.org/documents/card/en/c/ca9229en/.

Hilborn, R. 2010. Pretty good yield and exploited fishes. Marine Policy, 34: 193–196. http://www.sciencedirect.com/science/article/B6VCD-4W99NT8-

1/2/06cc5f8a2afa7621460cea7fd3a5a898.

Hillary, R. M., Preece, A. L., Davies, C. R., Kurota, H., Sakai, O., Itoh, T., Parma, A. M., *et al.* 2015. A scientific alternative to moratoria for rebuilding depleted international tuna stocks. Fish and Fisheries, 17: 469–482. Wiley/Blackwell (10.1111). https://doi.org/10.1111/faf.12121.

Horbowy, J., and Luzeńczyk, A. 2012. The estimation and robustness of F MSY and alternative fishing mortality reference points associated with high long-term yield. Canadian Journal of Fisheries and Aquatic Sciences, 69: 1468–1480.

Hordyk, A. R., Huynh, Q. C., and Carruthers, T. R. 2019. Misspecification in stock assessments: Common uncertainties and asymmetric risks. Fish and Fisheries, 20: 888–902.

ICES. 2014. Report of the Workshop to consider reference points for all stocks (WKMSYREF2). ICESCM 2014/ACOM, 47: 99.

ICES. 2020. The third Workshop on Guidelines for Management Strategy Evaluations (WKGMSE3). ICES Scientific Reports, 2: 1–112.

ICES. 2021a. Workshop on guidelines and methods for the evaluation of rebuilding plans (WKREBUILD). ICES Scientific Reports, 2: 1–79.

ICES. 2021b. Workshop of Fisheries Management Reference Points in a Changing Environment (WKRPChange). ICES Scientific Reports, 3: 39.

ICES. 2021c. ICES fisheries management reference points for category 1 and 2 stocks; Technical Guidelines.

ICES. 2021d. Benchmark Workshop on the development of MSY advice for category 3 stocks using Surplus Production Model in Continuous Time; SPiCT (WKMSYSPiCT). ICES Scientific Reports, 3: 1–317.

Kell, L. T., Nash, R. D. M., Dickey-Collas, M., Mosqueira, I., and Szuwalski, C. 2016. Is spawning stock biomass a robust proxy for reproductive potential? Fish and Fisheries, 17: 596–616.

Lassen, H., Kelly, C., and Sissenwine, M. 2014. ICES advisory framework 1977–2012: from Fmax to precautionary approach and beyond. ICES Journal of Marine Science, 71: 166–172. https://doi.org/10.1093/icesjms/fst146.

Mace, P. M. 1994. Relationships between common biological reference points used as thresholds and targets of fisheries management strategies. Canadian Journal of Fisheries and Aquatic Sciences, 51: 110–122.

Mace, P. M. 2001. A new role for MSY in single-species and ecosystem approaches to fisheries stock assessment and management. Fish and Fisheries, 2: 2–32.

McAllister, M. K., Pikitch, E. K., and Babcock, E. A. 2001. Using demographic methods to construct Bayesian priors for the intrinsic rate of increase in the Schaefer model and implications for stock rebuilding. Canadian Journal of Fisheries and Aquatic Sciences, 58: 1871–1890. http://www.nrcresearchpress.com/doi/abs/10.1139/f01-114.

Methot, R. D., and Wetzel, C. R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research, 142: 86–99. http://www.sciencedirect.com/science/article/pii/S0165783612003293.

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: 1–20.

Miller, T. J., and Brooks, E. N. 2021. Steepness is a slippery slope. Fish and Fisheries, 22: 634–645.

Musick, J. A. 1999. Criteria to Define Extinction Risk in Marine Fishes. Fisheries, 24: 6–14.

Myers, R. A., Bowen, K. G., and Barrowman, N. J. 1999. Maximum reproductive rate of fish at low population sizes. Canadian Journal of Fisheries and Aquatic Sciences, 56: 2404–2419.

New Zealand Ministry of Fisheries. 2008. Harvest Strategy Standard for New Zealand Fisheries. https://fs.fish.govt.nz/Page.aspx?pk=113&dk=16543.

New Zealand Ministry of Fisheries. 2011. Operational Guidelines for New Zealand's Harvest Strategy Standard. New Zealand Government. 80 pp. https://fs.fish.govt.nz/Doc/22847/Operational_Guidelines_for_HSS_rev_1_Jun_2011.pdf.ashx.

Nielsen, A., and Berg, C. W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research, 158: 96–101. http://www.sciencedirect.com/science/article/pii/S0165783614000228.

Patterson, K., Cook, R., Darby, C., Gavaris, S., Kell, L., Lewy, P., Mesnil, B., *et al.* 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish and Fisheries, 2: 125–157.

Punt, A. E., Smith, A. D. M., Smith, D. C., Tuck, G. N., and Klaer, N. L. 2013. Selecting relative abundance proxies for BMSY and BMEY. ICES Journal of Marine Science, 71: 469–483. http://icesjms.oxfordjournals.org/content/71/3/469.short (Accessed 30 May 2014).

Ralston, S., Punt, A. E., Hamel, O. S., Devore, J. D., and Conser, R. J. 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. Fishery Bulletin, 109: 217–231.

Restrepo, V. R., Thompson, G. G., Mace, P. M., Gabriel, W. L., Low, L. L., MacCall, A. D., Methot, R. D., *et al.* 1998. Technical Guidance On the Use of Precautionary Approaches to Implementing National Standard 1 of the Magnuson-Stevens Fishery Conservation and Management Act. 56 pp. http://www.st.nmfs.noaa.gov/StockAssessment/documents/Tech-Guidelines.pdf.

Sharma, R., Winker, H., Levontin, P., Kell, L., Ovando, D., Palomares, M. L. D., Pinto, C., *et al.* 2021. Assessing the potential of catch-only models to inform on the state of global fisheries and the UN's SDGs. Sustainability (Switzerland), 13.

Shertzer, K. W., Prager, M. H., and Williams, E. H. 2008. A probability-based approach to setting annual catch levels. Fishery Bulletin, 106: 225–232.

Silvar-Viladomiu, P., Minto, C., Halouani, G., Batts, L., Brophy, D., Lordan, C., and Reid, D. G. 2021. Moving reference point goalposts and implications for fisheries sustainability. Fish and Fisheries, 22: 1345–1358. John Wiley & Sons, Ltd. https://doi.org/10.1111/faf.12591.

Simmonds, E. J., Portilla, E., Skagen, D., Beare, D., and Reid, D. G. 2010. Investigating agreement between different data sources using Bayesian state-space models: an application to estimating NE Atlantic mackerel catch and stock abundance. ICES Journal of Marine Science, 67: 1138–1153. http://icesjms.oxfordjournals.org/cgi/doi/10.1093/icesjms/fsq013.

Thorson, J. T., Jensen, O. P., and Hilborn, R. 2015. Probability of stochastic depletion: an easily interpreted diagnostic for assessment modelling and fisheries management. ICES Journal of Marine Science, 72: 428–435.

Thorson, J. T. 2020. Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. Fish and Fisheries, 21: 237–251. John Wiley & Sons, Ltd. https://doi.org/10.1111/faf.12427.

UN. 1995. United Nations Conference on Straddling Fish Stocks and Highly Migratory Fish Stocks. New York.

Appendix A:

FLSRTMB: Characterising stock productivity in FLR

Demographic information from FLStock objects can be used to construct an age-structured Leslie matrix **A** of the form:

$$\mathbf{A} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_A \\ \theta_1 & 0 & 0 & 0 & 0 \\ 0 & \theta_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \theta_{T-1} & 0 \end{pmatrix} \quad \text{(B1)}$$

where ϕ_a is the average number of recruits expected to be produced by an adult female at age a and θ_a is the fraction of survivors at age, with T denoting the maximum age (plus group). The value of *r* is obtained from $\lambda = \exp(r)$, where λ is the dominant eigenvalue of **A** and *G*

Age-dependent survival calculated as $\theta_a = \exp(-M_a)$, where M_a is age-dependent natural mortality. The average number of recruits expected to be produced by an adult female at age t is expressed as:

$$\phi_t = \alpha w_a mat_a \quad \text{(B2)}$$

where α denotes the slope of the origin of the spawner-recruitment relationship (i.e. the ratio of recruits to spawner biomass at very low abundance) , w_a is the weight at age a, $mat_a$is the fraction of females that are mature at age a. For the calculation of the annual reproductive rate a first consider the BH-SSR of the form:

$$R = \frac{\alpha SB}{1+\beta SB} \quad \text{(B3)}$$

spr_0

where R is the number of recruits, S is the spawner biomass and β is the scaling parameter (Hilborn and Walters, 1992) . In contrast to alternative formulations of the BH-SSR, the parameter α can be directly interpreted as the slope in the origin of the S-R curve. We re-parameterized α as function of unfished spawner-biomass per recruit SPR_0 and the steepness parameter h of the spawner-recruitment relationship (Myers et al., 1999), such that:

$$\alpha = \frac{4h}{(1-h)} SPR_0^{-1} \quad \text{(B4)}$$

In cases where the quantities $W_{a,y}$, $Mat_{a,y}$ and $M_{a,y}$ varied annually, the averages of *r* and *G* across all years.

Appendix B:

FLSRTMB: Fitting conditioned Stock Recruitment Relationships (SRR) in FLR

Beverton-Holt SSR conditioning with prior information for steepness

The stock-recruitment relationship (SRR) was assumed to follow a Beverton and Holt model (BH-SRR) of the form

$$R_y = \frac{aSB_{y-a_{min}}}{b + SBB_{y-a_{min}}} e^{\epsilon_y - 0.5\sigma_r^2}$$

where R_y is the number of recruits in year y, SSB_{y-a_r} is the spawning biomass in year y minus minimum age a_{min} defined for the stock (typically age-0 or age-1). The recruitment deviation ϵ_t is assumed to be associated with a a first-order autocorrelation (AR1) process (Johnson et al. 2016; Simmonds et al. 2019), such that

$$\epsilon_y = \rho\epsilon_{y-1} + \delta_y\sqrt{1-\rho^2}$$

where ρ is the AR1 coeffient and $\delta_y \sim N(0, \sigma_r)$ determines variation in recruitment as a function of the recruitment standard deviation $sigma_r$.

The BH-SRR was fitted the recruitment R and SSB from FLStock objects using the FLR library FLSRTMB (Winker and Mosquiera; https://github.com/flr/FLSRTMB), which enables straight-forward integration of available prior information on the steepness s of the SSR from a recent meta-analysis (Thorson 2020).

For this purpose, the Beverton-Holt equation in FLSRTMB is re-parameterised as function of steepness s and annual unfished spawning biomass per-recruit SPR_0 (Mace and Doonan, 1988),

$$R_y = \frac{4sSB_{y-a_{min}}R0}{R_0 SPR_{0_y}(1-s) + SBB_{y-a_{min}}(5s-1)}$$

where steepness s is defined as the ratio of recruitment when SSB equals 20% of the unfished SSB_0 to the virgin recruitment R_0 at $SSB0$. A notable difference to the conventional parameterization is that SPR_{0_y} is treated as non-stationary, being function of annual quantities of $W_{a,y}$, $Mat_{a,y}$ and $M_{a,y}$. By way of using time-varying $SPR_{0,y}$, also takes into consideration the recent criticism by Miller and Brooks (2021) that specifying a set biological parameters to define a single time-invariant SPR_0 can be highly sensitive to reference estimation when using steepness values from meta-analysis.

The prior distribution for s is generated from truncated logit distributions ($TrunkLogit$) of the form

$$s = 0.2001 + 0.7999/\left(1 + exp(-s_{logit})\right)$$

$$s \sim TrunkLogit(s_{logit}, \sigma_{logit})$$

where s_{logit} and $sigma_{logit}$ correspond to the input of species-specific predictions for the distribution of s from the hierarchical taxonomic FishLife model (Thorson, 2020, https://github.com/James-Thorson-NOAA/FishLife), summarized in Table A1. The default prior is assuming an approximately uniform prior between 0.3 – 0.9, with a decreasing density (soft bounds) to the limits 0.2 and 1.0 (Figure. A1)

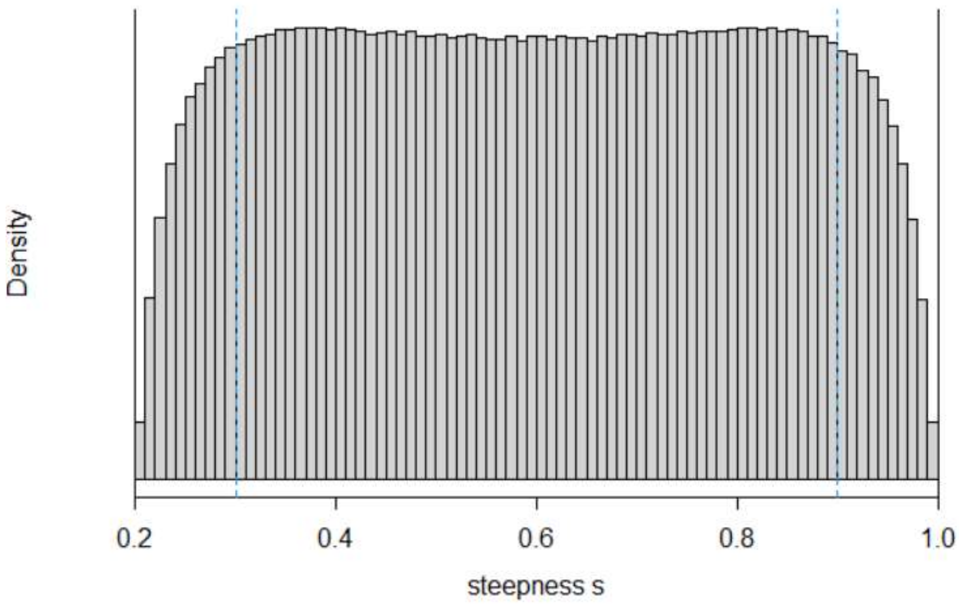

Figure A.1.1 Graphical illustration of default prior for estimating steepness s, with a mean of 0.6 and logit.sd = 1.5

The FLSRTMB estimates of $R0$ and s are then converted into the parameters a and b of the Beverton-Holt formulation in FLR, such that

$$a = \frac{4sR_0SPR_0}{5sSPR_0 - 1} \qquad \text{and} \qquad b = \frac{R_0SPR_0(1-s)}{5s-1}$$

where the reference for SPR_0 to predict a and b was taken the average SPR_{0_y} across all years in the case of the OM.

A conditioned, continuous hockey-stick SSR

A new conditional Hockey-Stick formulation was developed and implemented in 'FLSRTMB`. The new Hockey-Stick is based on a continuous, quadratic hockey-stick (c.f. Barrowman and Myers), which is re-parameterised as a function of SPR_{0_y} and a "re-purposed" steepness parameter s^* given by

$$R_y = \frac{s^*}{2P_{lim}SPR0_y}\left(SSB_y + P_{lim}R_0SPR_{0_y}/s^* - \sqrt{\left(SSB_y - P_{lim}R_0SPR_{0,y}/s^*\right)^2}\right)$$

In addition, the parameter P_{lim} is introduced, which then determines the lower of the ratio B_{lim}/SSB_{0_y}, where $Blim$ corresponds to break point b of the segmented regression and SSB_{0_y} is allowed to be treated as non-stationary being a function of $SSB_{0_y} = R_0SPR_{0_y}$.

The break point b (B_{lim}) and slope a are given by

$b = P_{lim} * R_0SPR_{0,y}/s$ and $a = R_0/b$

In the chosen setting for FLSRTMB, the parameter s^* was bounded by a mostly uniform distribution between $0.2 > s^* \leq 1$, with soft bounds towards the limits to invoke a conditioned Blim range of $0.1B_0 < B_{lim} < 0.3B_0$.

Table B1. List Species arranged by taxonomic order with FishLife (Thorson 2020) predictions for the recruitment standard deviation (σ_R), the auto-correlation coefficient (ρ), steepness (s) and the associated standard error (σ_S) on logit scale.

Species	Order	σ_R	ρ	s	σ_s
Argentina silus	Argentiniformes	0.69	0.38	0.52	1.14
Clupea harengus	Clupeiformes	0.67	0.32	0.58	0.26
Sardina pilchardus	Clupeiformes	0.49	0.50	0.77	0.60
Sprattus sprattus	Clupeiformes	0.70	0.31	0.80	0.67
Brosme brosme	Gadiformes	0.42	0.56	0.57	1.30
Gadus morhua	Gadiformes	0.53	0.39	0.79	0.22
Melanogrammus aeglefinus	Gadiformes	0.80	0.24	0.66	0.34
Merlangius merlangus	Gadiformes	0.64	0.31	0.71	0.43
Merluccius merluccius	Gadiformes	0.23	0.67	0.56	1.20
Micromesistius poutassou	Gadiformes	0.60	0.34	0.55	0.73
Molva molva	Gadiformes	0.38	0.56	0.53	1.33
Pollachius virens	Gadiformes	0.46	0.57	0.79	0.40
Pandalus borealis	Crustacian	0.28	0.27	0.84	0.30
Lophius piscatorius	Lophiiformes	0.30	0.88	0.92	1.28
Dicentrarchus labrax	Perciformes	0.34	0.75	0.90	1.93
Trachurus trachurus	Perciformes	0.53	0.47	0.75	0.87
Glyptocephalus cynoglossus	Pleuronectiformes	0.53	0.47	0.63	1.04
Lepidorhombus boscii	Pleuronectiformes	0.37	0.68	0.87	1.23
Lepidorhombus whiffiagonis	Pleuronectiformes	0.38	0.66	0.84	1.29
Pleuronectes platessa	Pleuronectiformes	0.48	0.58	0.82	0.40
Scophthalmus maximus	Pleuronectiformes	0.60	0.48	0.86	1.15
Solea solea	Pleuronectiformes	0.54	0.34	0.61	0.42
Scomber scombrus	Scombriformes	0.78	0.28	0.64	0.58
Sebastes norvegicus	Scorpaeniformes	0.56	0.61	0.58	0.96