BENCHMARK WORKSHOP ON HERRING (*CLUPEA HARENGUS*) IN THE GULF OF BOTHNIA (WKCLUB 2021)

VOLUME 3 | ISSUE 9

ICES SCIENTIFIC REPORTS

RAPPORTS
SCIENTIFIQUES DU CIEM

ICES INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CIEM CONSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

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

H.C. Andersens Boulevard 44–46
DK-1553 Copenhagen V
Denmark
Telephone (+45) 33 38 67 00
Telefax (+45) 33 93 42 15
www.ices.dk
info@ices.dk

ISSN number: 2618-1371

This document has been produced under the auspices of an ICES Expert Group or Committee. The contents therein do not necessarily represent the view of the Council.

ICES Scientific Reports

Volume 3 | Issue 9

BENCHMARK WORKSHOP ON HERRING (*CLUPEA HARENGUS*) IN THE GULF OF BOTHNIA (WKCLUB 2021)

Recommended format for purpose of citation:

ICES. 2021. Benchmark Workshop on herring (*Clupea harengus*) in the Gulf of Bothnia (WKCLuB 2021). ICES Scientific Reports. 3:9. 110 pp. http://doi.org/10.17895/ices.pub.5989

Editors

Johan Lövgren• Tapani Pakarinen • Noél Holmgren

Authors

Max Cardinale • Santiago Cerviño • David Gilljam • Zeynep Pekcan Hekim • Noél Holmgren • Jim Ianelli • Pekka Jounela • Olavi Kaljuste • Juha Lilja • Johan Lövgren • Alessandro Orio • Jukka Pönni • Tapani Pakarinen • Jari Raitaniemi

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

Contents

i Executive summary

The assessment for the Gulf of Bothnia herring (SD 3031) in 2019 was not accepted by the Advice Drafting Group and was changed from category 1 to 3. The assessment was not accepted based on the poor retrospective diagnostics where the Mohn's rho values were above 20% for SSB, F and recruitment. The aim for the Benchmark was to evaluate a new model, Stock Synthesis (SS3) as a candidate for the assessment of Gulf of Bothnia Herring SD30–31 in order to minimize the retrospective pattern previously observed.

Following the path of the Benchmark data related issues were revealed. This led to a situation that the benchmark was prolonged one year in order to correct the data related issues. Apart from a misspecification in the model about how the trapnet index was used (abundance index vs. biomass index), the data problem had been related to the acoustic survey. The acoustic survey index used in the assessment was thoroughly examined by the Baltic international fish survey working group (WGBIFS) in a meeting in December 2020. A number of model runs (six prior to meeting, and an additional 16 during the benchmark) were conducted for evaluation at this benchmark. The analysis presented extensive diagnostic tests including the standard ICES criterion related to retrospective patterns. This was considered an enhancement over using one method for accepting or rejecting an assessment. It was noted that the final retrospective pattern had low and acceptable values of Mohn's rho.

In general, the benchmark using the stock synthesis platform with the settings specified during the benchmark are considered acceptable for assessment and advice and have features that should ensure stability as new data are added (e.g. selectivity is assumed to be constant over time).

ii Expert group information

Expert group name	Benchmark Workshop on herring (*Clupea harengus*) in the Gulf of Bothnia (WKCluB)
Expert group cycle	2
Year cycle started	2020
Reporting year in cycle	2/2
Chairs	Johan Lövgren, Sweden
	Tapani Pakarinen, Finland
	Noél Holmgren, Sweden
Meeting venue and dates	4–6 February 2020, ICES HQ, Copenhagen, Denmark, eleven participants 25–27 January 2021, Online meeting, thirteen participants

1 Introduction

The assessment for the Gulf of Bothnia herring (SD 3031) in 2019 was not accepted by the Advice Drafting Group and was downgraded from category 1 to 3. The results from the SAM run for the 2019 assessment can be found in Figure 1. The assessment was not accepted based on the poor retrospective diagnostics where the Mohn's rho values were above 20% for SSB (37%), F (27%) and recruitment (68%) (Figure 1).

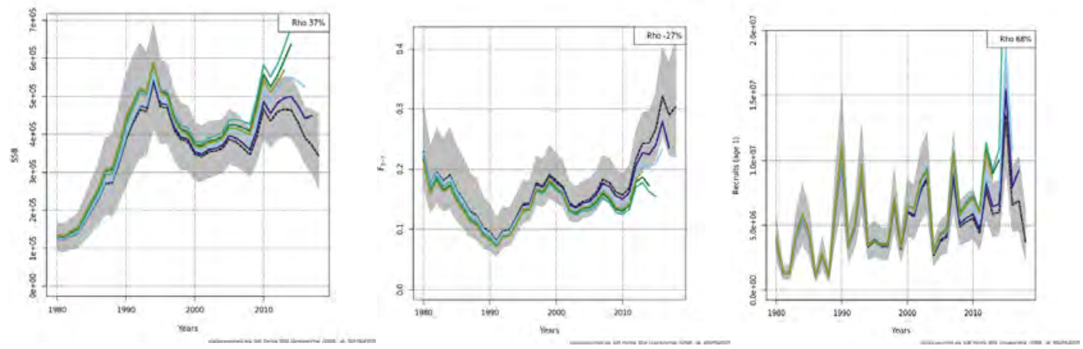


Figure 1. The retrospectives for the stock assessment run in SAM in 2019, which was not accepted. SSB, F and recruitment.

This benchmark was setup in order to investigate the reasons behind the bad retrospective diagnostics. In addition, a stock synthesis model was setup in parallel that showed good performance. Therefore, it was concluded that the benchmark would also investigate the potential use of a new model (SS3) for the herring SD 3031 assessment.

1.1 Terms of Reference

A Benchmark Workshop on herring (*Clupea harengus*) in the Gulf of Bothnia (WKCluB), chaired by Johan Lövgren and Noél Holmgren, Sweden and attended by two invited external experts Jim Ianelli, US and Santiago Cerviño, Spain was established. WKCLuB for the data meeting by correspondence on 19 November 2019 and then for a three-day Benchmark meeting in Copenhagen, Denmark, on 4–6 February 2020:

a) Evaluate the appropriateness of data and methods to determine stock status and investigate methods for short-term outlook taking agreed or proposed management plans into account for the stocks listed in the text table below. The evaluation shall include consideration of:
   1. Examine SS3 as an alternative assessment model to SAM;
   2. Explore impact of all tuning fleets on assessment estimates;

b) Agree and document the preferred method for evaluating stock status and (where applicable) short-term forecast, and update the stock annex as appropriate. Knowledge about environmental drivers, including multispecies interactions, and ecosystem impacts should be integrated in the methodology. If no analytical assessment method can be agreed, then an alternative method (the former method, or following the ICES data-limited stock approach) should be put forward;

c) Update the stock annex as appropriate;

d) Re-examine and update MSY and PA reference points according to ICES guidelines (see Technical document on reference points);

e) Prioritize recommendations for future improving of the assessment methodology and data collection.

f) Produce working documents to be reviewed during the Benchmark meeting at least seven days prior to the meeting.

1.2 Description of the Benchmark process

The data meeting was held by correspondence on 19 November 2019. An error in the acoustic data was found and corrected at the data meeting. Two weeks prior to the benchmark meeting, three working documents were finalised and presented to the group. One working document covered the SS3 model with initial runs, the second one the data input and the third working document consisted of comparison runs with SAM and SS3. The benchmark meeting took place in Copenhagen on 4–6 February 2020, where the assessment with SS3 was presented and accepted, after which the group could proceed with the calculation of the reference points.

However, after the benchmark report was done and before the assessment working group (WGBFAS) in April 2020 a model misspecification was found in the reference run. The trap-net index of abundance had been used in the original reference run as an index of biomass, when it should have been entered as an index of abundance. Therefore, after the benchmark meeting, the reference run for the assessment was run again with the correct specification of the trapnet survey index, reference points were re-calculated and the updated work was included in the reviewers' report.

Unfortunately, just prior to the assessment working group in April 2020, it was detected that the acoustic index input data used in the assessment was not area-corrected (i.e. multiplied) for the years 2013–2018 with a factor taking into account the shallow coastal areas that had not been surveyed. This led to that the benchmark results were considered invalid, and that the assessment of the stock was downgraded from a category 3 to a 5. The WGBIFS was asked to review the acoustic data in a meeting set to December 2020. After the WGBIFS meeting the Benchmark was continued in January 2021 where the reference run with full-reviewed acoustic survey data and new reference points were calculated.

2 Gulf of Bothnia Herring (SD 3031)

2.1 Examine SS3 as an alternative assessment model to SAM

Assessment of herring in SDs 30–31 was conducted using the Stock Synthesis (SS) model (Methot and Wetzel, 2013). Stock Synthesis is programmed in the ADMB C++ software and searches for the set of parameter values that maximize the goodness-of-fit, then calculates the variance of these parameters using inverse Hessian and MCMC methods. The assessment was conducted using the 3.30 version of the Stock Synthesis software under the windows platform (WD 2).

The assessment model of herring in SDs 30–31 is a one area, annual, age-based model where the population is comprised of 20+ age-classes (with age 20 representing a plus group) with sexes combined (male and females are modelled together).

The model starts in 1963 and the initial population age structure was assumed to be in an exploited state, so that the initial catches was assumed to be the average of last three years (1963–1965) in the time-series. Fishing mortality was modelled using hybrid F method (Methot and Wetzel, 2013). Option 5 was selected for the F report basis; this option represents a recent addition to Stock Synthesis and corresponds to the fishing mortality requested by the ICES framework (i.e. simple unweighted average of the F of the age classes chosen to represent the F_{bar} (age 3–7)). Overview of the data included in the final Stock Synthesis model is shown in Figure 3 and described in WD1 (input data).

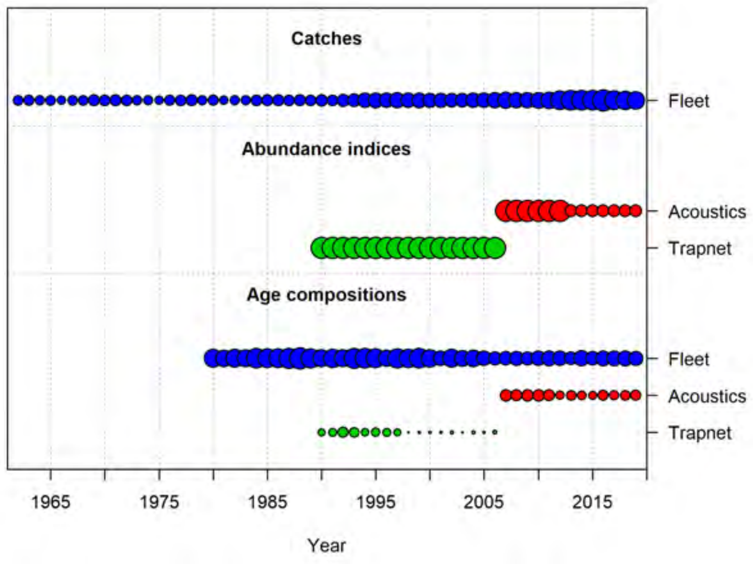


Figure 3. Herring SDs 30–31. Summary of the input time-series included in the model.

The parameter and the configuration for the final assessment run and alternative runs that was tested in the Benchmark meeting is described thoroughly in WD2. After a series of statistics to test the robustness of the final assessment model the model was accepted by the group and external reviewers (WD2).

The 3-years retrospectives of the final model were stable (Figure 4). The estimated Hurtado-Ferro *et al*. (2014) variant of the Mohn´s rho indices were inside the bounds of recommended values for both SSB (-0.11) and F (0.18).

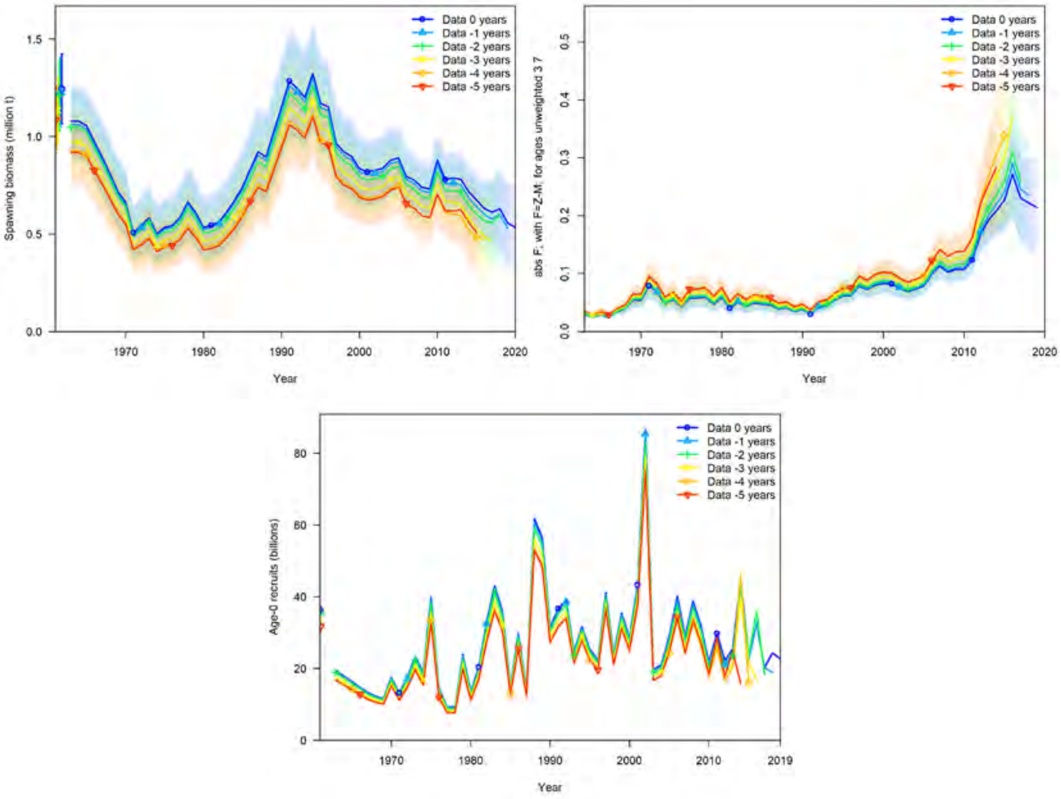


Figure 4. Herring SDs 30–31. Retrospective analyses of the reference model.

2.2 Comparison with SAM (Exploratory run)

A SAM run was performed with the final data used in SS3 assessment model. The comparison of these runs (SSB, F and Recruitment) are presented in Figure 5. The SAM and the SS3 model show the same dynamics of the stock (Figure 5 and Annex 3).

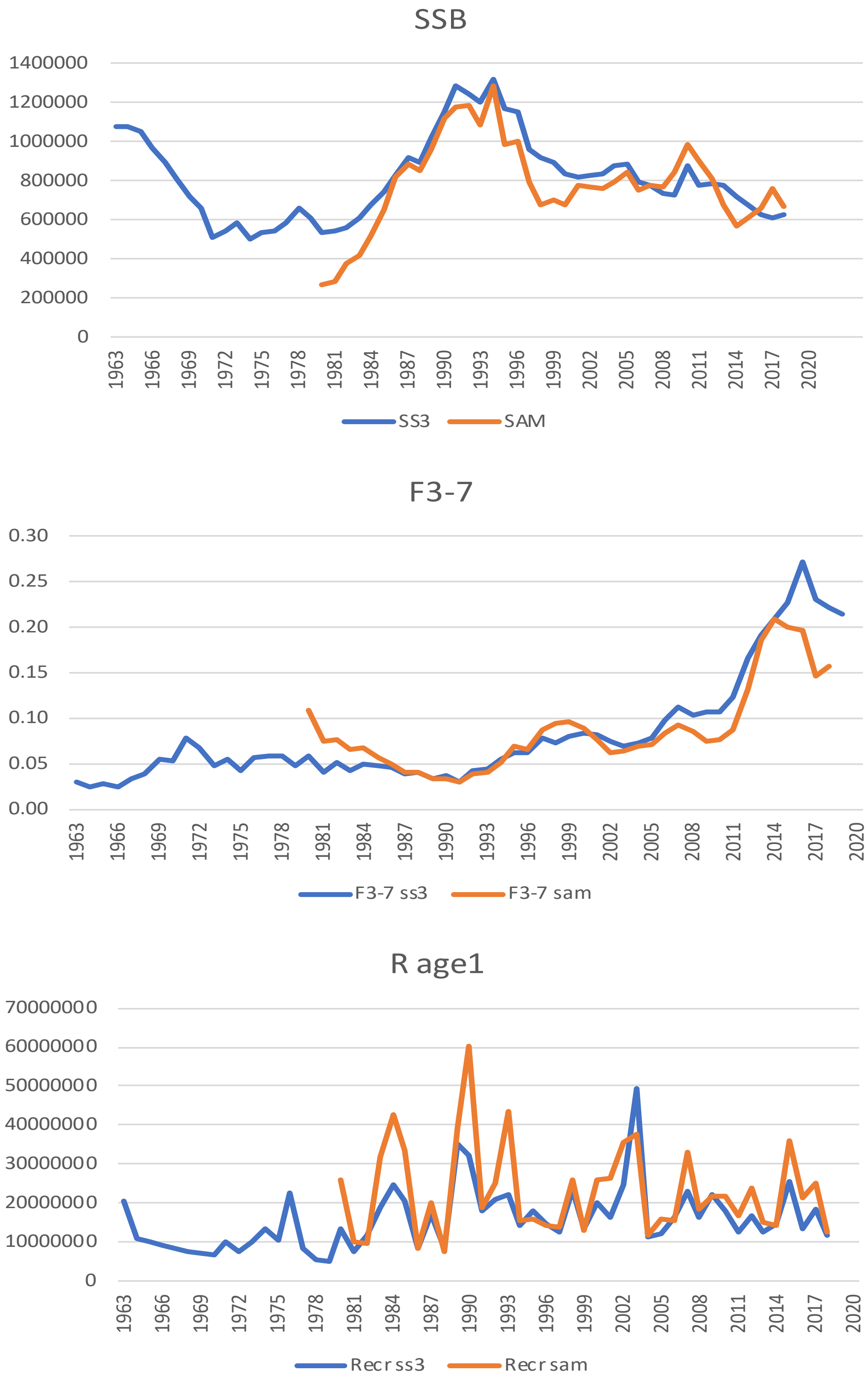


Figure 5. Comparison of runs with same input data for SAM (light blue) and SS3 (dark blue) for SSB, F and Recruitment.

2.3 Short-term projection

The short-term projections were performed in the meeting in 2020, and hence remain with the same settings also for the new runs carried out in 2021. The short-term projections are made with Stock Synthesis using MCMC or the delta-Multivariate log-Normal' (delta-MVLN) estimator

(Walter and Winker, 2019; Winker *et al.*, 2019). MVLN infers within-model uncertainty from maximum likelihood estimates (MLEs), standard errors (SEs) and the correlation of the untransformed quantities and it has demonstrated to be able to mimic the MCMC fairly closely.

Recruitment in the forecast period is set to the average of the last ten years for which recruitment deviations are estimated in the Stock Synthesis model. For maturity and weight-at-age an average of the last three years is used. Constant selectivity is used. Probabilistic forecasts were used. In this approach, catch and SSB levels corresponding to different catch options are calculated as in typical deterministic short-term forecast but using MCMC to make it possible to also include the most correct associated probability of the SSB to be below biomass reference points, for each year of forecast. Therefore, an MCMC with 1 100 000 iterations, 100 000 burn-in and 1000 thinning was run for the different levels of assumed F in the assessment year and assessment year+1, assuming F constraint in the intermediate year. It is important to note that the inputted F values for the forecast will sometime be different from the model realized F in the MCMC (but also in the MLE if this is used for the forecast). This is because the F used is an average across ages and those ages have different F, because they are affected by selectivity. Each draw of the MCMC has different selectivity so the F produced for each draw will be slightly different due to the different selectivity. We have tested running three different MCMC with 110 000 iterations and compared the difference in F inputted and model realized F.

2.4 Explore impact of all tuning fleets on assessment estimates

2.4.1 Data issues detected

At the data meeting (on 19th November 2019) a mistake in the input data used in the April 2019 assessment was detected. The acoustic survey index used was calculated wrongly from the years 2013 to 2015 from the survey. The 2015 year class in the acoustic survey was calculated much higher than it actually was. However, there were more issues detected that were corrected by the WGBIFS (ICES, 2021) and then used in the additional benchmark meeting in January 2021.

2.4.2 Expanding the age groups in the tuning fleets and catch

In the data meeting it was decided to expand the age groups (both in catch and tuning fleets) from 1–15+. We also expanded the time-series of the catch data from 1963 to the assessment year.

Type	Year range	Age
Commercial fleet	1963–assessment year	1–15+
Acoustic survey	2007–assessment year	1–15+
Trapnet survey	1990–2006	1–15+

3 Recommendation for future benchmarks

We recommend that the standardization procedure of the acoustic survey abundance and biomass should be done using spatio-temporal models like for example VAST. The standardization should always include a spatial-time interaction factor and a vessel effect. Also, standardization should include observations (e.g. hauls or trips) with zeroes. The year should also be modelled as a factor and not as a smoother when the CPUE time-series is used in the assessment. (WGBFAS and WGBIFS). Preliminary analysis on the reference point showed a small change on F_{MSY} and B_{MSY} year to year, similar to those observed in the retrospective analysis for F and SSB. This is not critical in the short-term forecast. However, it is worth to follow the progress in reference points, at least every three years, particularly if retrospective patterns increase (WGBFAS).

Two different age-reading methods have been used for ageing. During 1980–2001 whole otoliths were used while from 2002 and onwards cut otoliths were used for ageing. We recommend that a recalibration of the age readings from the period when whole otoliths were used (1980–2001) should be performed. The major concern is that older ages of herring are underestimated when ageing whole otoliths. (WGBIOP)

Adaptation of a more balanced sampling covering all quarters, fishing métiers and the two subdivisions (SD 30 and 31) is recommended in order to estimate fish biological parameters (age, weight, etc.). For historical catch data it is recommended to split the data by subdivision, fishery and quarter. (RCG)

A recommendation to calculate a CV for the acoustic survey and if possible also for the trapnet survey. (WGBFAS and WGBIFS)

3.1 Considerations for future benchmarks

Calculation of a smoothed maturity ogive and a smoothed weight-at-age with an aim to reduce annual unexplained variation in retrospective analysis for fishing mortality, stock biomass and recruitment parameters. (WGBFAS).

Annex 1: Final assessment model reference run of herring in the Gulf of Bothnia (ICES SD 30–31)

By Massimliano Cardinale and Alessandro Orio

A model misspecification was found in the reference run. The trapnet index of abundance was used in the original reference run as an index of biomass while instead it was supposed to be used as an abundance index. The acoustic index has also been reviewed by WGBIFS. The abundance calculations have been made in the StoX software based on the BIAS survey data available in the ICES database for acoustic and trawl surveys. The acoustic survey index time-series in numbers shows a pronounced jump from 2013, and also in 2017 which coincides to changes in survey vessels (RV DANA to RV ARANDA). The survey vessels used by years were RV ARGOS in 2007–2010, RV DANA in 2011–2012 and 2017, RV ARANDA in 2013–2016 and 2018–2020. This pattern is common also to the average CPUE in biomass of herring in the acoustic survey trawl hauls and when excluding age 1 from the index and indicates that a model standardization is necessary to account for a likely vessel effect and possibly for other effects in the survey. In the meantime, to account for this pattern, the CV assigned for years after 2013 (0.2) was 0.3 compared to 0.1 for other years. This improved the retrospective and could be a temporary solution before a full standardization of the acoustic survey index is conducted. This standardisation should also provide annual CV estimates. Therefore, the assessment was run again with the correct specification of the trapnet survey index, with the updated acoustic survey index, with the different CV for the later part of the time-series (2013–2019), and with data up to 2019.

All data inputs are summarized in Table 1 while in Table 2 the configuration of the updated reference model is reported. Note that the selectivity of the fleet has been reduced by one parameter in the updated run to stabilize estimate of selectivity parameters. All the updated diagnostics have been run using the ss3diags package for R (Winker *et al*., 2020).

Table 1. Herring SDs 30–31. Input data used in the Stock Synthesis models.

Type	Description	Year range	Range
Catches	Catches in tonnes for each year	1963–2019	
Age compositions	Catch in numbers (thousand) per age group	Commercial fleet: 1980–2019 Acoustic survey: 2007–2019 Trapnet survey: 1990–2006	0–15+
Weight at age	Weight in kg per age group	Commercial fleet: 1970–2019 Acoustic survey: 2007–2019 Trapnet survey: 1990–2006	0–17+
Maturity ogives	Empirical maturity-at-age estimated from commercial data		
Natural mortality	Natural mortality by age class costant for the entire time-series derived from Then *et al.*, 2015		0–20+
Surveys indices	Abundance index from acoustic survey and abundance index from trapnet survey	Acoustic survey: 2007–2019 Trapnet survey: 1990–2006	
SSB index	SSB proportional to fecundity		

Table 2. Herring SDs 30–31. Settings of the Stock Synthesis assessment reference model. The table columns show: number of estimated parameters, the initial values (from which the numerical optimization is started), the intervals allowed for the parameters, the priors used, the value estimated by the model and its standard deviation. Parameters in bold are set and not estimated by the model.

Parameter	Number estimated	Initial value	Bounds (low,high)	Prior	Value (MLE)	Standard deviation
Natural mortality (age classes 0.5, 1, 3, 5, 8, 15)		**0.563, 0.472, 0.332, 0.290, 0.267, 0.257**				
Stock and recruitment						
Ln(R₀)	1	17.43	(16, 25)	No_prior	17.41	0.07
Steepness (h)	1	0.775	(0.1, 1)	0.74	0.77	0.11
Recruitment variability (σ_R)		**0.60**				
Ln (Recruitment deviation): 1963–2019	56					
Recruitment autocorrelation		**0**				
Initial catches		Average of 1963–1965				
Initial F Commercial fleet	1	0.2	(0.001, 1)	No_prior	0.03	0.005
Selectivity (random walk)						
Commercial fleet						
Change from age1 to age2	1	1.45	(-5, 9)	No_prior	1.29	0.07
Change from age2 to age3	1	0.4	(-5, 9)	No_prior	0.35	0.06
Change from age3 to age4	1	0.15	(-5, 9)	No_prior	0.12	0.06
Change from age4 to age5	1	0.14	(-5, 9)	No_prior	0.11	0.07
Change from age5 to age6	1	0.03	(-5, 9)	No_prior	-0.003	0.07
Acoustic Survey						
Change from age1 to age2	1	0.6	(-5, 9)	No_prior	0.49	0.17
Change from age2 to age3	1	0.2	(-5, 9)	No_prior	0.23	0.18
Change from age3 to age4	1	0.02	(-5, 9)	No_prior	0.02	0.22
Change from age4 to age5	1	0.11	(-5, 9)	No_prior	0.02	0.25
Change from age5 to age6	1	0.14	(-5, 9)	No_prior	0.12	0.22
Trapnet Survey						
Change from age3 to age4	1	0.30	(-5, 9)	No_prior	0.10	0.15

Parameter	Number estimated	Initial value	Bounds (low,high)	Prior	Value (MLE)	Standard deviation
Catchability **(Using float option in Stock Synthesis)**						
Acoustic survey						
Ln(Q) – catchability		**-2.47811**				
Extra variability added to input standard deviation		**0**				
Trapnet survey						
Ln(Q) – catchability		**3.86604**				
Extra variability added to input standard deviation		**0**				

Final model run

Overview of the datasets included in the final Stock Synthesis model is shown in Figure 1.

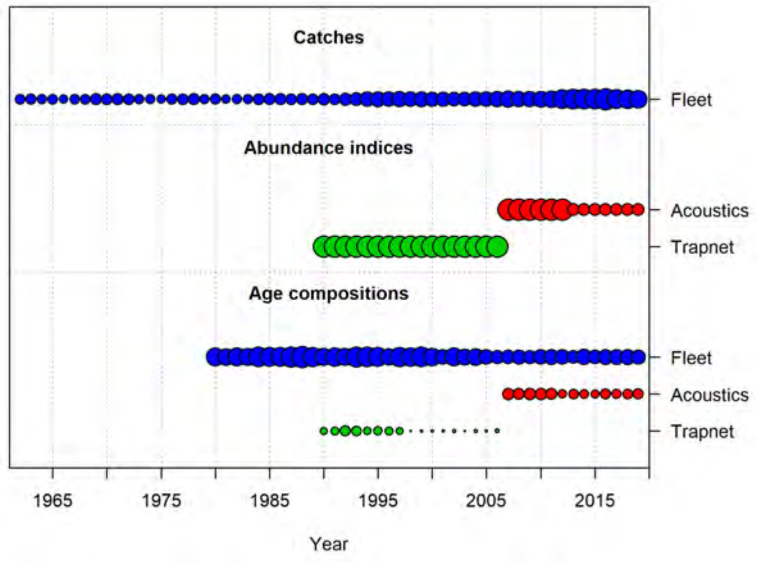


Figure 1. Herring SDs 30–31. Summary of the input time-series included in the model. Circles are proportional to total catch for catches, to precision for indices, and to total sample size for age compositions.

The selectivity of commercial fleet, acoustic and trapnet surveys is well estimated (Figure 2).

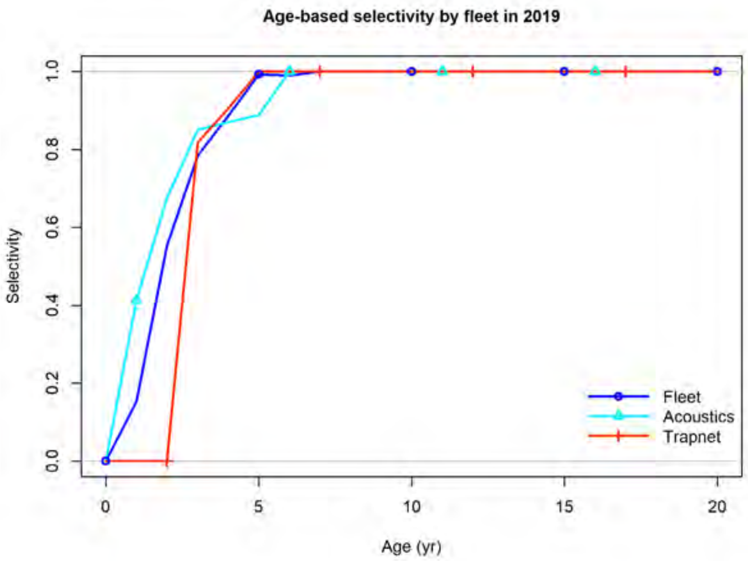


Figure 2. Herring SDs 30–31. Age based selectivity of the commercial fleet, acoustic and trapnet surveys.

The fitting of the model was good, with the age compositions well reconstructed. The residuals are quite low, never below -2.2 and above 2.2, and without particular worrying patterns (Figure 3 and 4). However, there is a positive residual pattern by cohort for acoustics, and a residual pattern with negative residuals in the historical part followed by positive residuals in recent years for older ages changing from negative to positive around year 2000. Figure 3 also shows an overestimation at age 13 for Fleet and Trapnet, which has been associated to a change in the procedure for reading the otoliths.

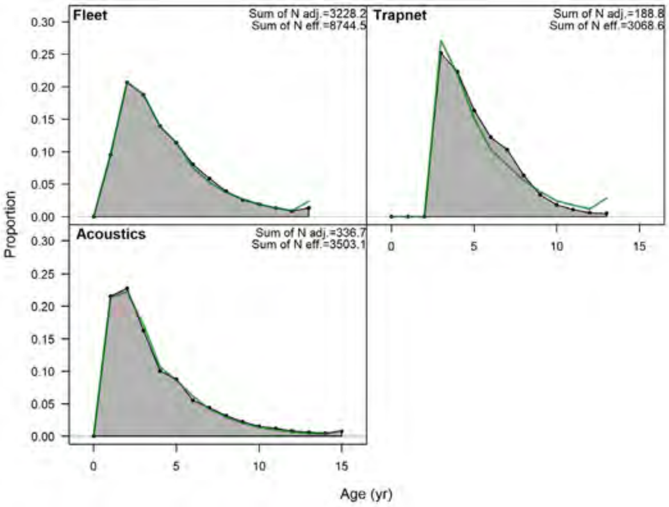


Figure 3. Herring SDs 30–31. Model fits to age composition data for commercial fleet, acoustic and trapnet surveys.

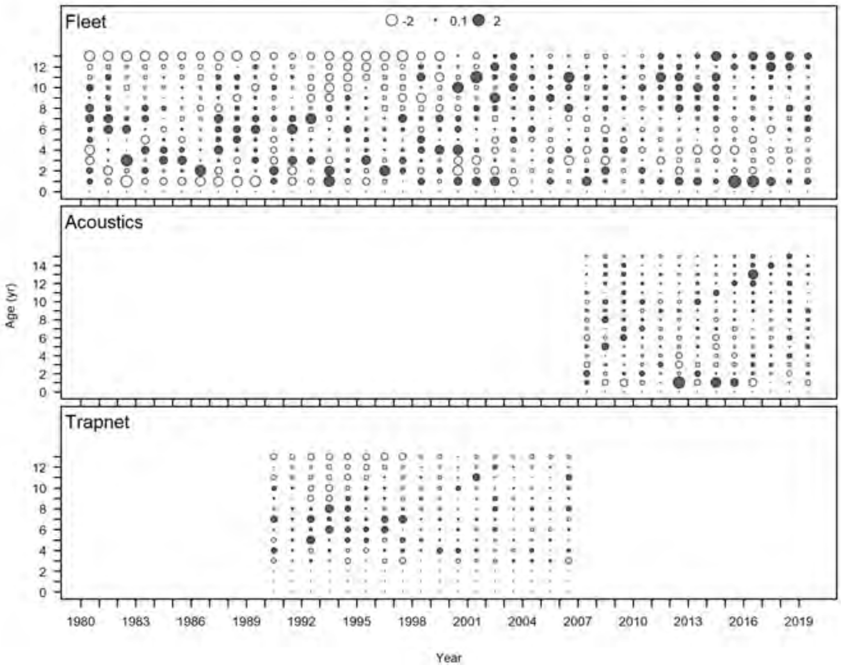


Figure 4. Herring SDs 30–31. Residuals of fits to age composition data for the commercial fleet and acoustic and trapnet surveys.

Overall, the model doesn't provide a very good fit to the trend of the acoustic survey (Figure 5) while the trapnet survey shows a good fit to the data (Figure 6).

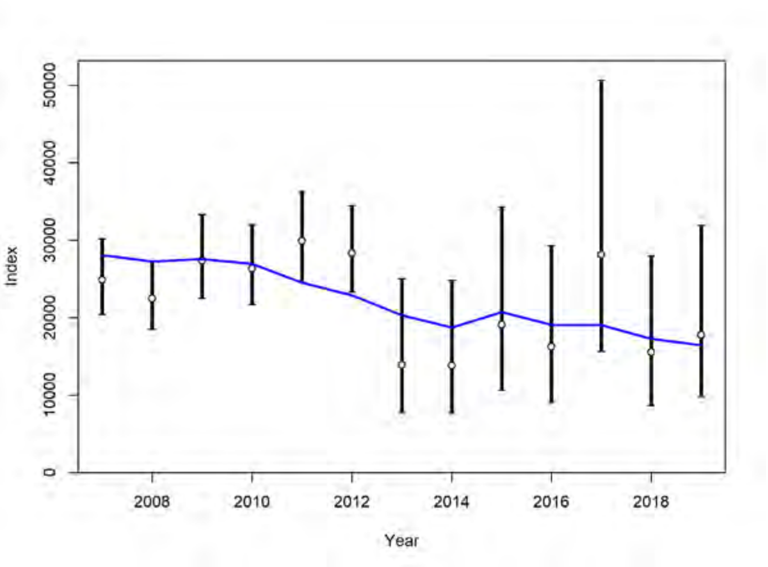


Figure 5. Herring SDs 30–31. Model fits to the acoustic survey abundance index.

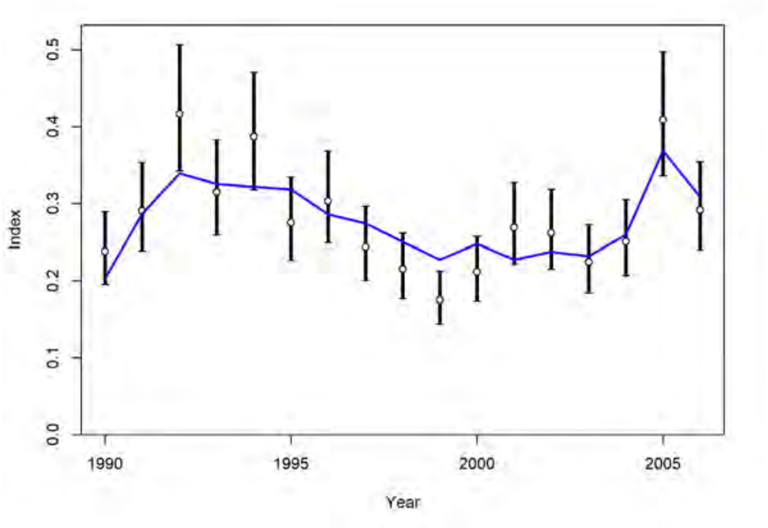


Figure 6. Herring SDs 30–31. Model fits to the trapnet survey abundance index.

A non-random pattern of residuals may indicate that some heteroscedasticity is present, or there is some leftover serial correlation in sampling/observation error or model misspecification. Several well-known nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test (Gibbons and Chakraborti, 1992). Here we used the runs test (RMSE and ordinary runs test) to evaluate the residuals of surveys and age frequency distributions (e.g. SEDAR 40, 2015; Winker *et al.*, 2018). The results of the runs test are presented in Figures 7 and 8. The RMSE runs test indicated that the fit of the CPUE index was good because no residuals were larger than 1 and the root-mean-square error (RMSE) was less than 30%, indicating a random pattern of the surveys residuals and the age frequency distributions. The RMSE plot is considered as a tool for identifying trends in residuals. If the standard deviation is small on a given year this means the fleets are in agreement, even if not fitting well. Its purpose is to visualize multiple residuals at once, pick up on periods of substantial data conflicts (width of boxes) and systematic departures in median residuals (loess). In this case, as we have two non-overlapping in time surveys, the RMSE is the only useful metric. The ordinary runs test was passed for both acoustic and trapnet surveys residuals and also for all age frequency distributions with the exception of the trapnet (Figure 7).

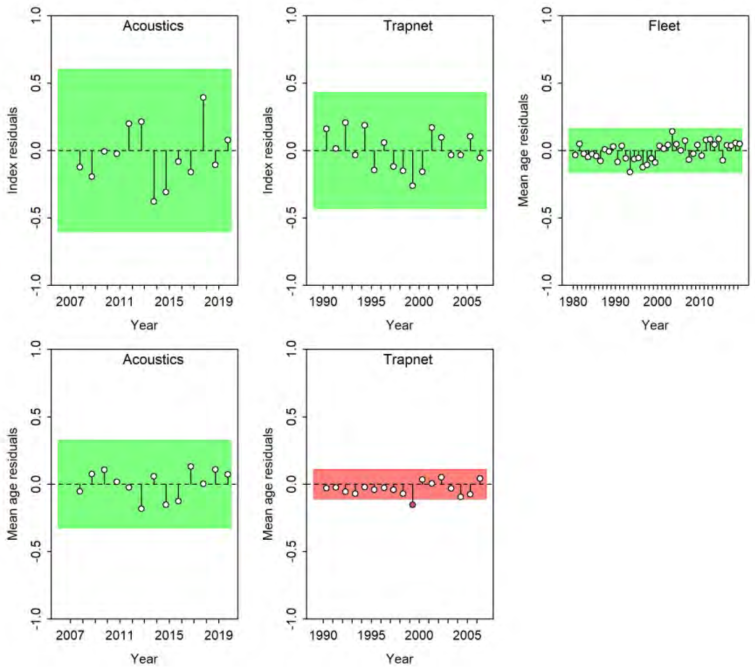


Figure 7. Herring SDs 30–31. Residuals from runs test analyses for the age distributions, and the fit to the acoustic and trapnet survey indices.

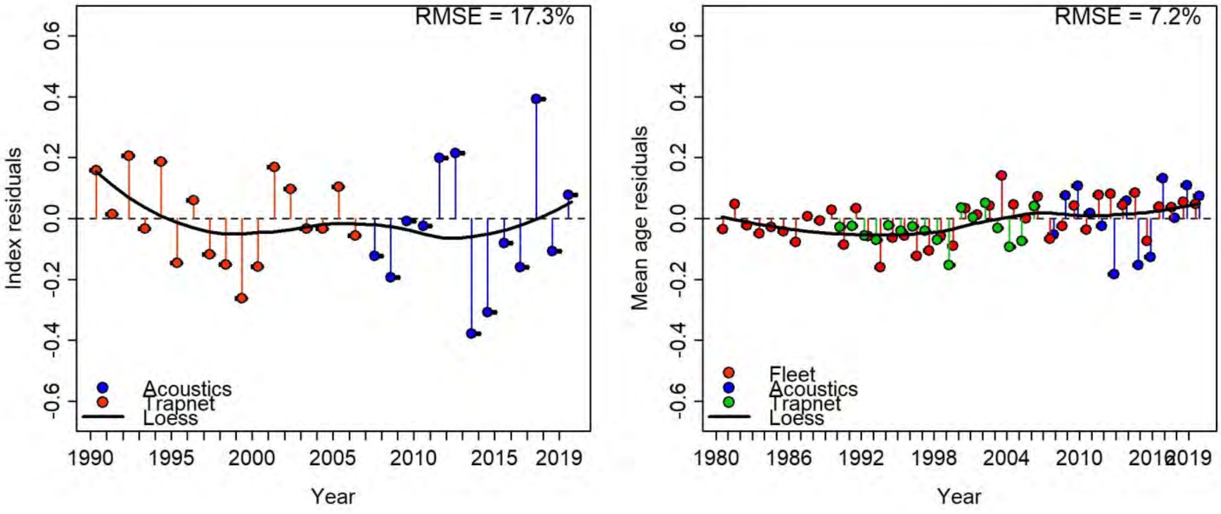


Figure 8. Herring SDs 30–31. Residuals from the JABBA runs test analyses for the age distributions and the fit to the acoustic and trapnet survey indices.

Retrospective analyses

Retrospective analysis is a diagnostic approach to evaluate the reliability of parameter and reference point estimates and to reveal systematic bias in the model estimation. It involves fitting a stock assessment model to the full dataset. The same model is then fitted to truncated datasets where the data for the most recent years are sequentially removed. The retrospective analysis was conducted for the updated reference model for the last five years of the assessment time horizon to evaluate whether there were any strong changes in model results. Given that the variability of Mohn's rho index depends on life history, and that the statistic appears insensitive to F, Hurtado-Ferro *et al*. (2014) proposed the following rule of thumb when determining whether a retrospective pattern should be addressed explicitly. Values of Mohn's rho index higher than 0.20 or lower than -0.15 for long-lived species (upper and lower bounds of the 90% simulation intervals for the flatfish base case), or higher than 0.30 or lower than -0.22 for short-lived species (upper and lower bounds of the 90% simulation intervals for the sardine base case) should be cause for concern and taken as indicators of retrospective patterns. However, Mohn's rho index values smaller than those proposed should not be taken as confirmation that a given assessment does not present a retrospective pattern, and the choice of 90% means that a "false positive" will arise 10% of the time. In both cases, model misspecification would be correctly detected more than half the time. The retrospectives of the updated reference model were rather stable (Figure 9). The estimated Hurtado-Ferro *et al*. (2014) variant of the Mohn's rho indices were inside the bounds of recommended values for SSB (-0.17) but outside the bounds for F (0.28). Also, the forecast Mohn's rho which is a measure of the predictive power of the model for SSB and F is outside the bounds for both SSB (-0.21) and F (0.3).

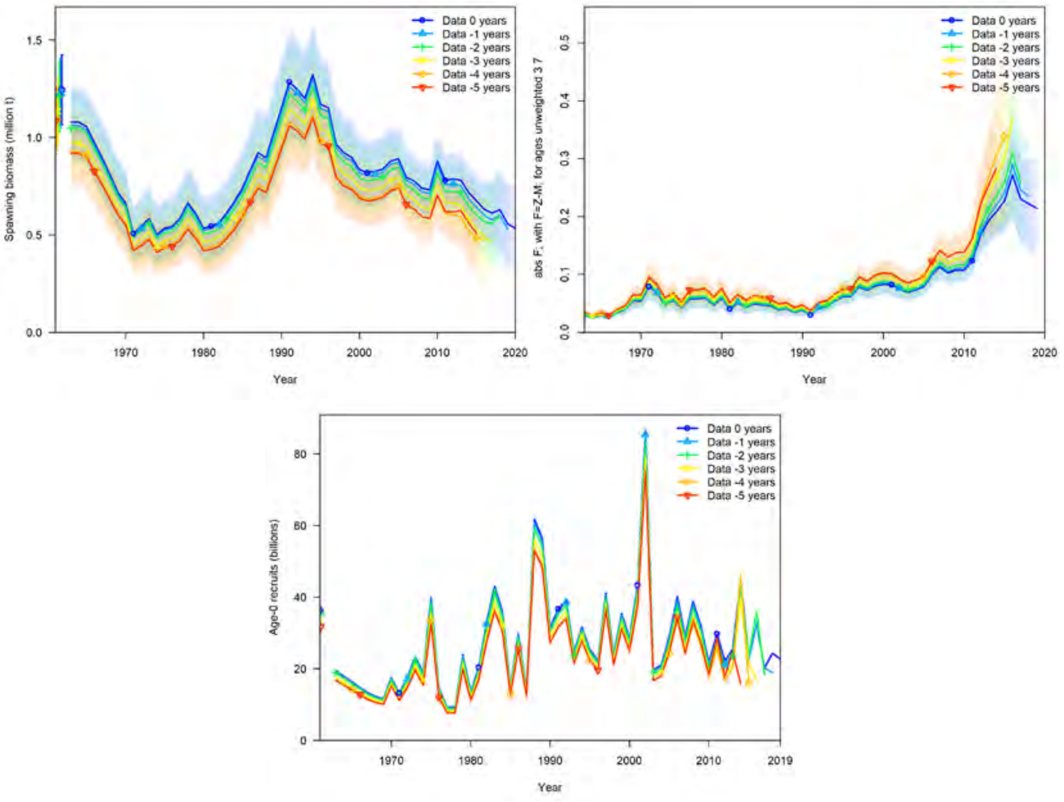


Figure 9. Herring SDs 30–31. Retrospective analyses of the updated reference model.

Following the most recent ACOM guidelines, when the five years retrospective results in a Mohn's rho outside the recommended bounds, results for the three-years retrospective should be calculated and compared. If the Mohn's rho values of the three years retrospective are within the bounds, the model is considered to be robust enough for providing advice.

When using a 3-years retrospective both the estimated Hurtado-Ferro *et al*. (2014) variant of the Mohn's rho indices and the forecast Mohn's rho were inside the bounds of recommended values for SSB (-0.11 and -0.13) and for F (0.18 and 0.19).

There is little or no information in the data to estimate the sizes of the 2019 and 2018 year class. Retrospective analyses of year-class strength for young fish shown the estimates of recent recruitment to be unreliable prior to at least between age 1 and 2 (Figure 10), which implies that you need to observe an year class at least twice to estimate it with a good precision.

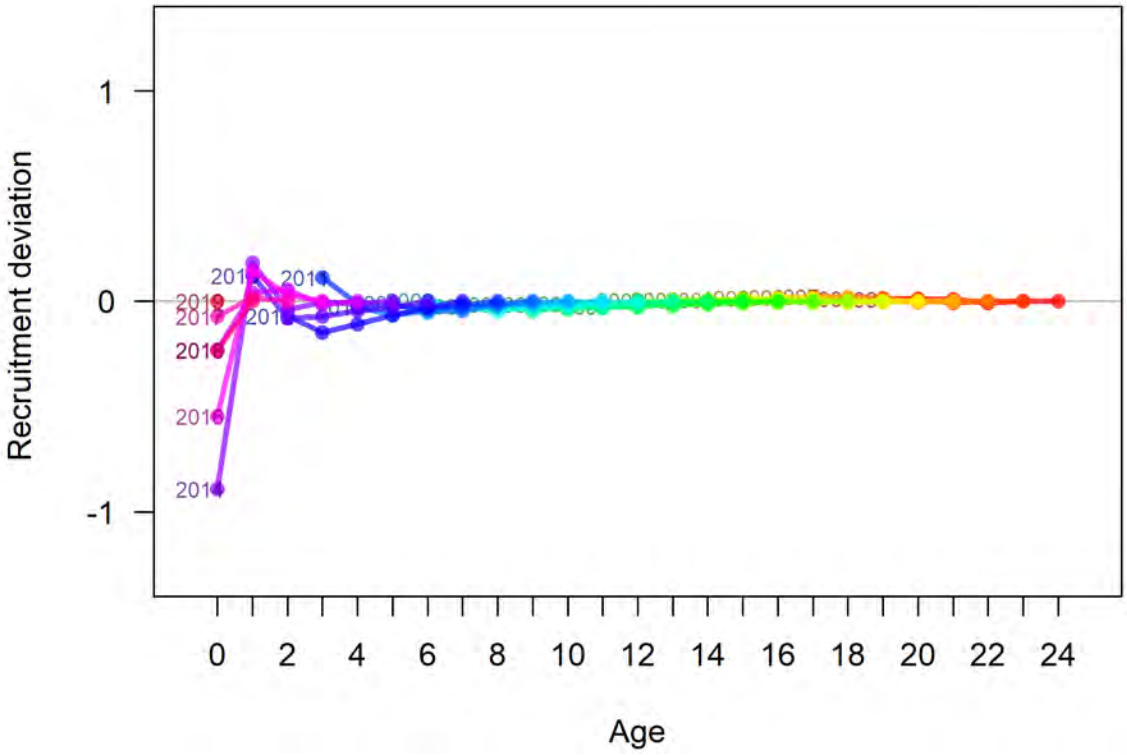


Figure 10. Herring SDs 30–31. Retrospective recruitment estimates scaled relative to the most recent estimate of the strength of each cohort.

Hindcasting

The provision of fisheries management advice requires the assessment of stock status relative to reference points, the prediction of the response of a stock to management, and checking that predictions are consistent with reality. A major uncertainty in stock assessment models is the difference between model estimates and observed quantities as CPUE or age distribution. To evaluate uncertainty often a number of scenarios are considered corresponding to alternative model structures and dataset choices (Hilborn, 2016). It is difficult, however, to empirically validate model prediction, as fish stocks can rarely be observed and counted. Various criteria are available for estimating prediction skill (see Hyndman and Koehler, 2006). One commonly used measure is root-mean-square error (RMSE). RMSE, however, is an inappropriate and misinterpreted measure of average error (Willmott and Matsuura, 2005). On the other hand, mean absolute error (MAE) is a more natural measure of average error, and unlike RMSE is unambiguous. Scaling the average errors using the Mean Absolute Scaled Error (MASE) allows forecast accuracy to be compared across series on different scales. MASE values greater than one indicates that in-sample one-step forecasts from the naïve method perform better than the forecast values under consideration. MASE also penalizes positive and negative errors and errors in large forecasts and small forecasts equally.

Kell *et al*. (2016) showed how hindcasting can be used to evaluate model prediction skill of the CPUE. When conducting hindcasting, a model is fitted to the first part of a time-series and then projected over the period omitted in the original fit. Prediction skill can then be evaluated by comparing the predictions from the projection with the observations using for example the MASE indicator (Hyndman and Athanasopoulos, 2013).

Hindcasting was conducted for the reference model (Figure 11). The results showed that the acoustic survey performs well in hindcasting given that the MASE value is lower than the 1.0 threshold when predicting the index one year ahead.

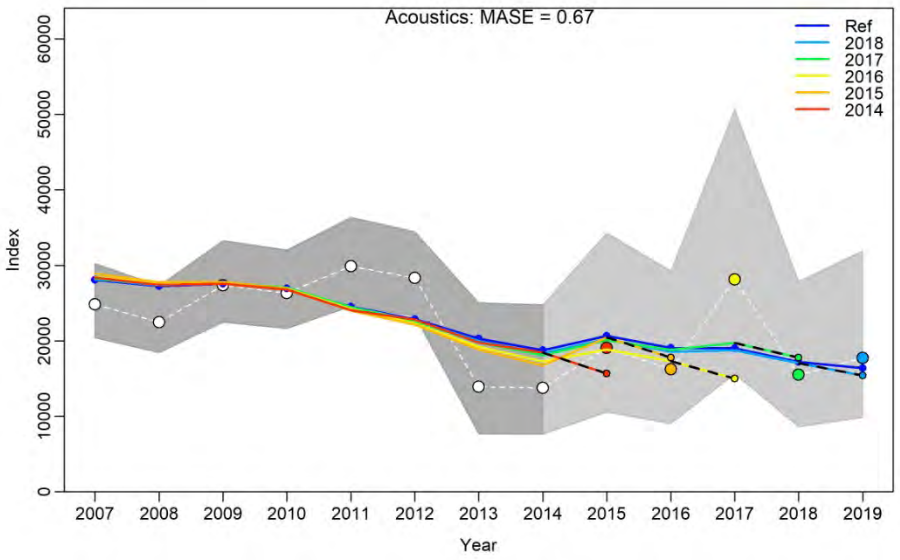


Figure 11. Herring SDs 30–31. Results of hindcasting for the acoustic survey. Black dashed lines are the forecasts while colour coded observations are the corresponding observations that were dropped when making the prediction residual for that specific year.

Trends in SSB, F and R of the updated reference model

The stock status and the trends in SSB, R and F are based on the MLE model. The spawning–stock biomass (SSB) has been declining from the beginning of the time-series up to the 1970s, then it increased during the 1980s reaching levels comparable to the 1960s. During the mid-1990s the SSB decreased and has remained stable at that level since then. Fishing mortality (F) has increased markedly since 1990s, with a peak in 2016. Recruitment (R) has been fluctuating throughout the time-series. In 2002 a very strong year class appeared (Figure 12 and Table 3).

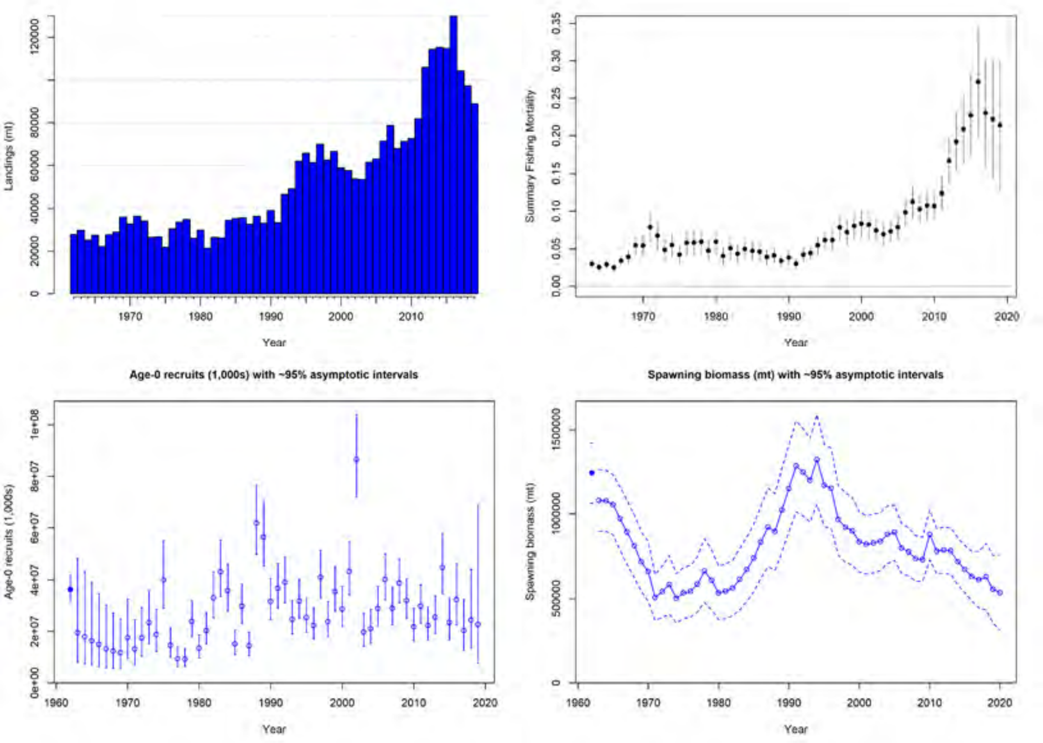


Figure 12. Herring SDs 30–31. Summary of the stock assessment. SSB, F and R with 95% confidence intervals. Catches by fleet and SSB are in tonnes R is in thousands of individuals.

Table 3. Herring SDs 30–31. Summary of the stock assessment. Catches and SSB are in tonnes R is in thousands of individuals.

Year	SSB	F3-7	Recruitment	Catch
1963	1079580	0.03	19482800	29739
1964	1077700	0.03	17851100	25204
1965	1054580	0.03	16323500	27541
1966	972143	0.02	14835300	22164
1967	892272	0.03	13247100	27772
1968	809109	0.04	12215600	28966
1969	719077	0.05	11631400	35996
1970	661513	0.05	17537800	32790
1971	506985	0.08	13103000	36347
1972	541414	0.07	17454900	34092
1973	582867	0.05	23371400	26507
1974	501267	0.05	18695300	26776
1975	533827	0.04	39893000	21811
1976	543697	0.06	14572100	30520
1977	586582	0.06	9348120	33634
1978	664493	0.06	9176120	34873
1979	608636	0.05	23837800	26109
1980	533515	0.06	13401700	29809
1981	544390	0.04	20260200	21526
1982	562869	0.05	33025300	26499
1983	614942	0.04	43064900	26208
1984	674200	0.05	35806900	34545
1985	740536	0.05	15061700	35432
1986	831970	0.05	29816500	35579
1987	921518	0.04	14346400	32628
1988	894602	0.04	61763800	36418
1989	1022130	0.03	56377900	33086
1990	1150840	0.04	31563200	39180
1991	1283730	0.03	36649400	33419
1992	1247380	0.04	39059800	46610
1993	1200110	0.04	24634300	49314
1994	1323160	0.05	31716400	61986
1995	1169380	0.06	25375700	65547
1996	1152110	0.06	22315800	61303
1997	964787	0.08	41019000	69808
1998	920915	0.07	23742000	62474
1999	897984	0.08	35395200	66502
2000	834620	0.08	28576400	58852
2001	818147	0.08	43208900	57806
2002	826030	0.07	86561900	53969
2003	837165	0.07	19652600	53644
2004	878085	0.07	20981200	61423
2005	890207	0.08	28815200	62911
2006	795963	0.10	40137900	71318
2007	776043	0.11	28968200	78678
2008	738291	0.10	38734100	67914
2009	731941	0.11	31944200	71248
2010	877844	0.11	21711500	72590
2011	778787	0.12	29734900	81850
2012	785765	0.17	22215800	106007
2013	781211	0.19	25476200	114396
2014	718138	0.21	44768400	115366
2015	673905	0.23	23387500	114942
2016	631475	0.27	32281900	130029
2017	610912	0.23	20236000	104358
2018	628807	0.22	24319300	97366
2019	555343	0.21	22709300	88907
2020	535314			

Comparison with the unweighted model

In Figure 13 it is possible to see the comparison between the reference run weighted using the Francis method and the unweighted one. The two runs are almost identical. Moreover, a model which uses Dirichlet multinomial distribution for weighing the different components, shows that the Dirichlet parameters are estimated to be 1 for all components. Therefore, we propose to use the unweighted model for advice.

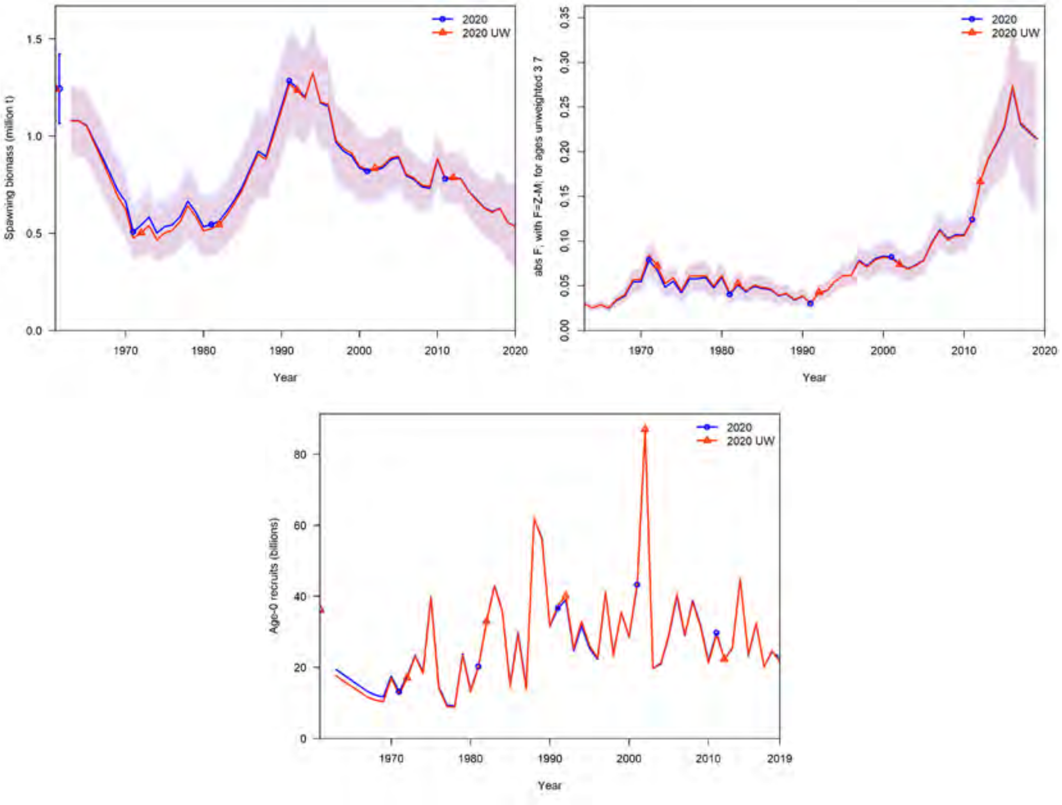


Figure 13. Herring SDs 30–31. Comparison between the reference run weighted using the Francis method (in blue) and the unweighted one (in red). SSB, F and R with 95% confidence intervals. SSB is in tonnes and R is in thousands of individuals.

Ensemble of alternative model runs

An ensemble of four alternative model configurations and the reference run was also run. The alternative four model configurations were created based on hypothesis testing and feedback from the WK and the reviewers. The results of the ensemble were not used to provide advice but as a demonstration of the procedures and the workflow used to build an ensemble model. The models used in the ensemble are summarised in Table 4.

Table 4. Characterisation of the models used in the ensemble.

Model	Main change from reference model
E1	Reference
E2	Age2
E3	Survey CV
E4	Survey CV Biomass
E5	T-distribution

In particular, model E2 refers to the model without age 2 individuals in the acoustic time-series, E3 is the model with CV as estimated by WGBIAS for the acoustic time-series, while E4 is also using estimated CV but with an index of biomass instead of abundance as in the reference model. Finally, E5 uses a t-distribution with four degrees of freedom instead of a log-normal distribution for the acoustic surveys. Figure 14 shows the comparison between the five models used in the ensemble for the main derived quantities (i.e. SSB, F and R). The models were then weighted by diagnostics which are summarised in Table 5 using the threshold method (i.e. assigning 1 to a diagnostic test if passed and 0 if failed). Based on the diagnostics, the average score for each model was calculated and used to weight the models in the final ensemble. Reference models scored best of all models with 0.643, while all other models scored 0.500. In the last step, the ensemble model was build stitching the model results of all models weighted by diagnostic. The delta-Multivariate log-Normal' (delta-MVLN) estimator (Walter and Winker, 2019; Winker *et al.*, 2019) was used to generate bootstrapped observations and estimate model uncertainty, with the number of bootstrapped observation draws for each model proportional to the diagnostic weights. It infers within-model uncertainty from maximum likelihood estimates (MLEs), standard errors (SEs) and the correlation of the untransformed quantities F/F_{MSY} and SSB/SSB_{MSY} and it has demonstrated to be able to mimic the MCMC fairly closely. These quantities are derived with Stock Synthesis using the delta-method to calculate the asymptotic variance estimates from the inverted Hessian. To generate Kobe posteriors from a delta-MVLN distribution requires the means and the variance-covariance matrix (VCM) of $\log(SSB/SSB_{MSY})$ and $\log(F/F_{MSY})$. Figure 15 shows the ensemble un-weighted (i.e. Equal) and weighted (i.e. Threshold) by the threshold method as described above including a 3-year forecast with F set at F_{MSY} and R set at the average of the last three years. Finally, the kobe plot of the ensemble model and the proportion of the bootstrapped observation in each of the quadrant as estimated in 2021 is show in Figure 16.

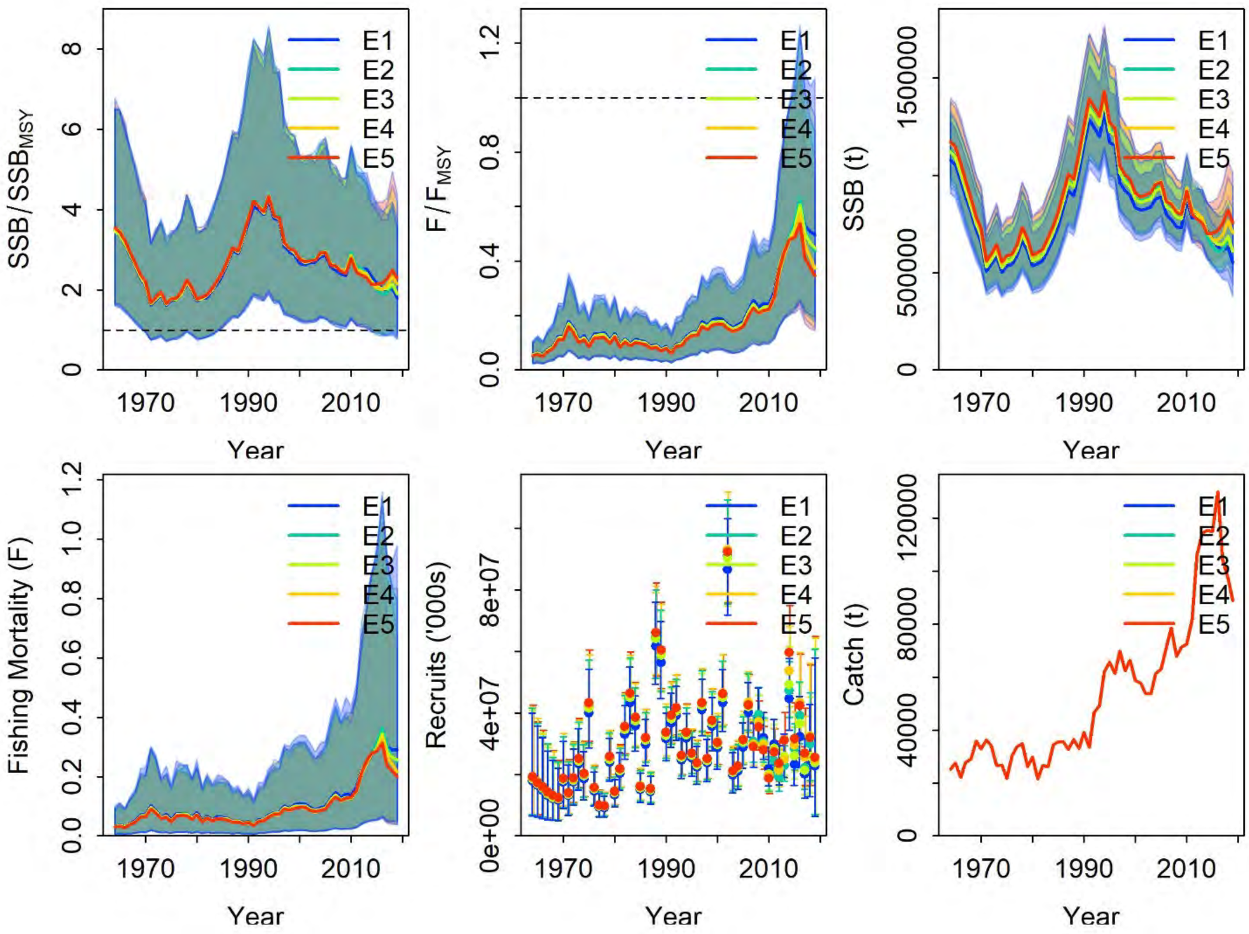


Figure 14. Herring SDs 30–31. Comparison between the five models used in the ensemble.

Table 5. Results of the diagnostics for the 5 models used in the ensemble.

Run	Runs_test_1	Runs_test_2	Runs_test_3	RMSE_Perc	RMSE_Perc_1	MASE_1
E1	Passed	Passed	Failed	21.1	6.6	0.67
E2	Passed	Passed	Passed	19.2	6.8	1.52
E3	Passed	Passed	Passed	20.2	6.5	1.01
E4	Passed	Passed	Passed	27.4	6.8	1.26
E5	Passed	Passed	Passed	29.9	6.9	1.50
	Retro_Rho_1	Forecast_Rho_1	Retro_Rho_2	Forecast_Rho_2	MASE_3	MASE_4
E1	-0.17	-0.22	-0.17	-0.22	0.92	0.90
E2	-0.32	-0.40	-0.32	-0.40	1.49	1.25
E3	-0.28	-0.34	-0.28	-0.34	1.12	1.02
E4	-0.42	-0.48	-0.42	-0.48	1.73	1.55
E5	-0.41	-0.47	-0.41	-0.47	1.69	1.51

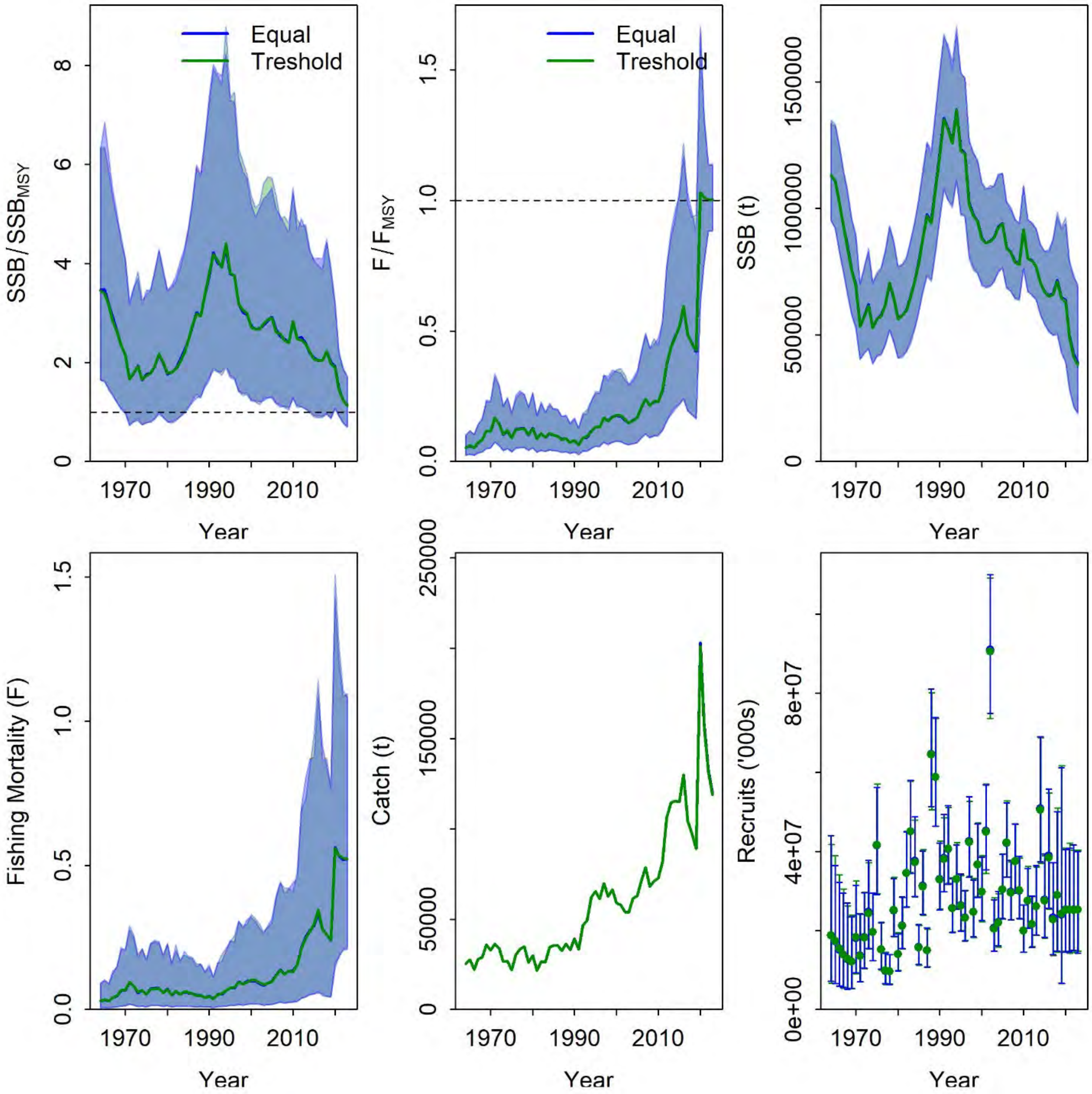


Figure 15. Herring SDs 30–31. Ensemble of the 5 models using equal weight or weighted by threshold method.

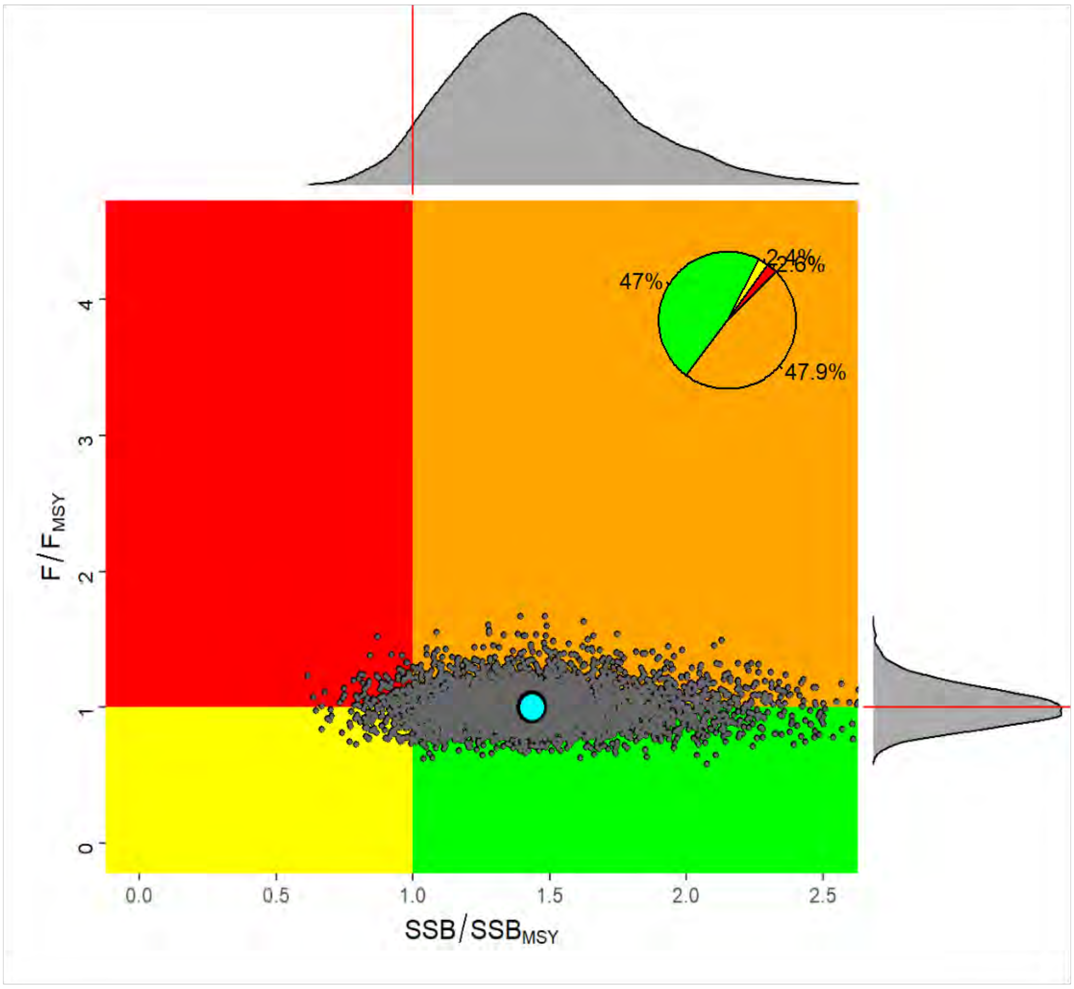


Figure 16. Herring SDs 30–31. Kobe plot of the ensemble model.

Annex 2: Final reference points analysis

Herring in Subdivision 30 and 31 (Gulf of Bothnia)

Current reference points

Summary table of current stock reference points following ICES (2017, 2017a) protocols:

Reference point	Value	Technical basis
Current F_{MSY}	0.21	Maximizes median long-term yield, based on stochastic simulations
Current B_{lim}	202272	B_{pa}/1.4 (as it is not possible to estimate B_{lim} from stock–recruitment data)
Current B_{pa}	283180	MSY $B_{trigger}$ (as it is not possible to estimate B_{lim} from stock–recruitment data)
Current $B_{trigger}$	283180	5th percentile of the distribution of SSB when fishing at F_{MSY}, based on stochastic simulations

Source of EqSim inputs

The results from the Stock Synthesis assessment (1980–2019) were processed for application in EqSim.

EqSim settings and configuration

Data and parameters	Setting	Comments
SSB-recruitment data	1980–2019	
Exclusion of extreme values (option extreme.trim)	Selected	
Mean weights, proportion mature and F-at-age pattern	2017–2019	
Exploitation pattern	2017–2019	
Assessment error in the advisory year. CV of F	0.212	ICES default value
Autocorrelation in assessment error in the advisory year	0.423	ICES default value
Autocorrelation in recruitment	Selected	

Results

Stock–recruitment estimates

The stock–recruitment fit using the three models (Ricker, B&H and segmented regression) weighted by the default "Buckland" method available in EqSim gave 84% of the points as derived from a segmented regression curve. However, the estimated breakpoint for this curve is at 718 138 t, which is around 54% of the maximum observed values. This was considered to be high because it is estimated to be much larger than B_{MSY} and around 54% of the max observed SSB. The influence of the 2002 year class on the estimate of the breakpoint was examined and found to be minimal; when excluding the 2002 year class, the breakpoint is estimated at 700 761 t. In this case the group decided to follow the "type 6" stock–recruitment relationship as described in the ICES guidelines (ICES, 2017b). Therefore, B_{pa} was set at the lowest observed SSB between 1980–2019 (i.e. the period where the recruitment is estimated in the assessment model; $B_{pa} = B_{loss} = 533\ 515$ t). B_{loss} was observed in the 1980s during a period of very low F and high M (due to cod predation). B_{lim} was then calculated as $B_{pa}/(\exp(1.645*SSB_{var})$ which results in a value of 376 571 t. The quantity "SSB_{var}" was taken as the uncertainty associated to the SSB in last assessment year (i.e. 2020; $\sigma = 0.212$). Applying this variance to B_{MSY} and computing the lower 5th percentile of 305 544 t. This is lower than B_{pa} (533515 t) and thus eliminated as a candidate and $B_{trigger}$ was thus set equal to B_{pa}.

Given the B_{lim}, B_{pa} and $B_{trigger}$ estimates, the ICES procedure was implemented to compute the remaining reference points. The S–R relationship selected was a hockey-stick with the breakpoint set at B_{pa} (equal to B_{loss}). The number of samples used to fit the S–R relationship and the number of runs used in all EqSim simulations were 1000 and 200, respectively. Autocorrelation of recruitment was used in all EqSim simulations. According to the ICES guidelines F_{pa} is set equal to $F_{P.05}$ with $B_{trigger}$.

Proposed reference points

Applying the ICES (2017) procedure as detailed above resulted in the following proposed stock reference points:

Stock	
Reference point	Value
$F_{P.05}$ (5% risk to B_{lim}) with $B_{trigger}$	0.384
$F_{P.05}$ (5% risk to B_{lim}) without $B_{trigger}$	0.312
F_{MSY}	0.384
F_{MSY} lower	0.308
F_{MSY} upper	0.384
F_{pa}	0.384
F_{lim}	0.497
F_{MSY} upper precautionary	0.384
F_{MSY} range with $B_{trigger}$	0.384
F_{MSY} range without $B_{trigger}$	0.312
MSY $B_{trigger}$	533 515 t
B_{pa}	533 515 t
B_{lim}	376 571 t

Discussion / Sensitivity

According to the EqSim estimations, $F_{P0.05}$ (0.384) is lower than F_{MSY} (0.407) estimated without $B_{trigger}$ and thus the F_{MSY} and F_{MSY} range are considered precautionary. As an integrated model, Stock Synthesis provides estimate of B_0. The calculated B_{lim} using ICES guidelines, corresponds to around 27% of B_0. It was noted that in other settings and parts of the world, lower limits of spawning biomass are typically a smaller fraction of B_0 (e.g. the minimum stock size threshold in the USA is considered to be half of B_{MSY}, and a standard proxy for B_{MSY} is 40% of unfished). Consequently, the value of B_{lim} (given the model settings and assumptions) would be more precautionary.

The S–R relationship shows no clear recruitment trend. However, the historical fishing mortality has always been well below conventional F_{MSY} (=0.407) and historically SSB has been always well above B_{MSY} (333 000 t) estimated by the Stock Synthesis model. This means that the dynamic range of SSB is little because since this stock has not experienced high Fs. As a consequence, B_{loss}, is well above B_{MSY}, and therefore is not a right proxy for B_{lim} as suggested for Category 5 of the S–R relationship. Category 6, designed for stocks with little dynamic range and low F, suit best in this case, which implies setting B_{pa} as B_{loss}.

References

ICES. 2017a. Report of the Benchmark Workshop on Baltic Stocks (WKBALT2017), 7–10 February 2017. Copenhagen, Denmark. ICES CM 2017/ACOM:30. 42 pp.

ICES. 2017b. ICES fisheries management reference points for category 1 and 2 stocks. In ICES Advice Technical Guidelines. ICES Advice 2017, pp. 1–19.

ICES. 2021. ICES Working Group on Baltic International Fish Survey (WGBIFS; outputs from 2020 meeting). ICES Scientific Reports. 3:02. 539pp. https://doi.org/10.17895/ices.pub.7679.

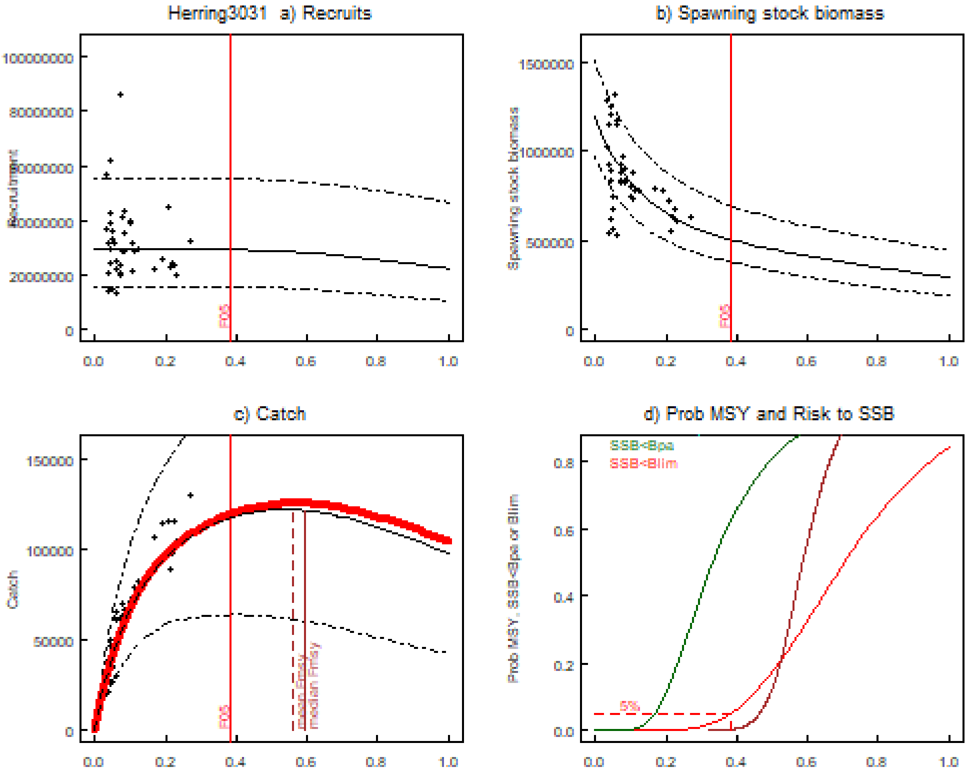


Figure 1. EqSim results for Herring in subdivisions 30 and 31 with $B_{trigger}$.

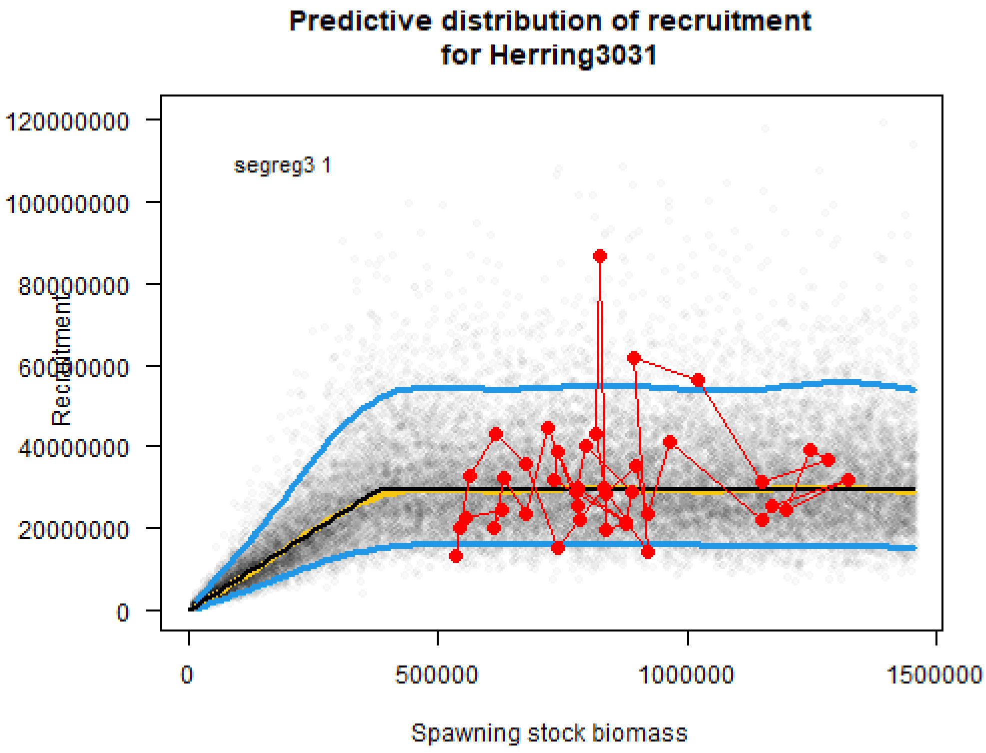


Figure 2. Stock–recruitment relationship (i.e. segmented regression with breakpoint at B_{pa}) for Herring in subdivisions 30 and 31 used in the EqSim simulations for the estimation of the F_{MSY} reference points.

Annex 3: Comparison of SS3 reference run and SAM

Model comparison (SS3 vs SAM) and general performance of the SAM model

The SAM model was rejected in February 2020 benchmark meeting because the five years average Mohn's rho bias was higher than the upper limit defined in the ICES rule/guideline (Mohn's rho ±20%). Thereafter, ACOM set the new (alternative optional) three years retrofit rule/guideline for the Mohn's rho bias. In the 2021 meeting, the same SAM model was applied with the revised rule/guideline and new revised acoustic survey data, which showed much smaller retrospective bias. Hence, the SAM model was used as a contrast (evaluation) model for the SS3 model (Figure 1, Table 1).

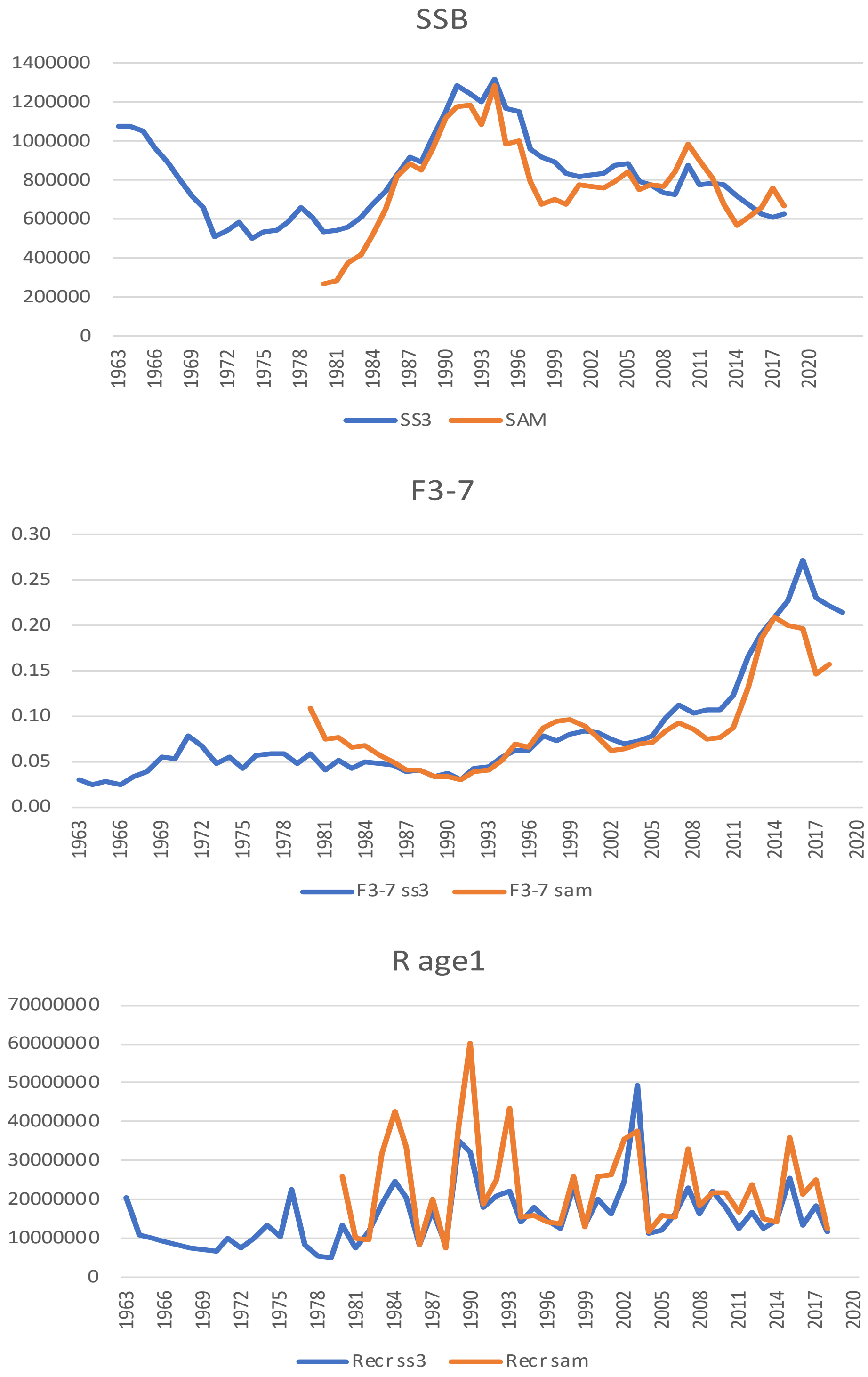


Figure 1. Comparison of the main SS3 and SAM results based on new revised acoustic survey data (after December 2020 WGBIFS meeting).

Table 1. Estimates of retrospective pattern (Mohn's Rho bias in %) with five and three years retrofit, based on the most recent ACOM guidelines and model fit (log(L), AIC) of the SAM model using the old (last year 2018) and new revised (last year 2019) acoustic survey data.

Mohn's Rho	**Old survey data2018**	**New survey data 2018**	**New survey data 2019**	**New survey data 2019**
SSB	-17 (5 yrs)	-18 (5 yrs)	-10 (5 yrs)	4 (3 yrs)
F	34 (5 yrs)	34 (5 yrs)	21 (5 yrs)	0 (3 yrs)
R	-8 (5 yrs)	-8 (5 yrs)	-5 (5 yrs)	6 (3 yrs)
Model fit				
log(L)	-526.9	-520.62	-528.84	-528.84
AIC	1083.79	1071.24	1087.67	1087.67

The changes made in acoustic survey data in December 2020 WGBIFS meeting did not much affect the SAM based recruitment, SSB and fishing mortality estimates (Figure 2).

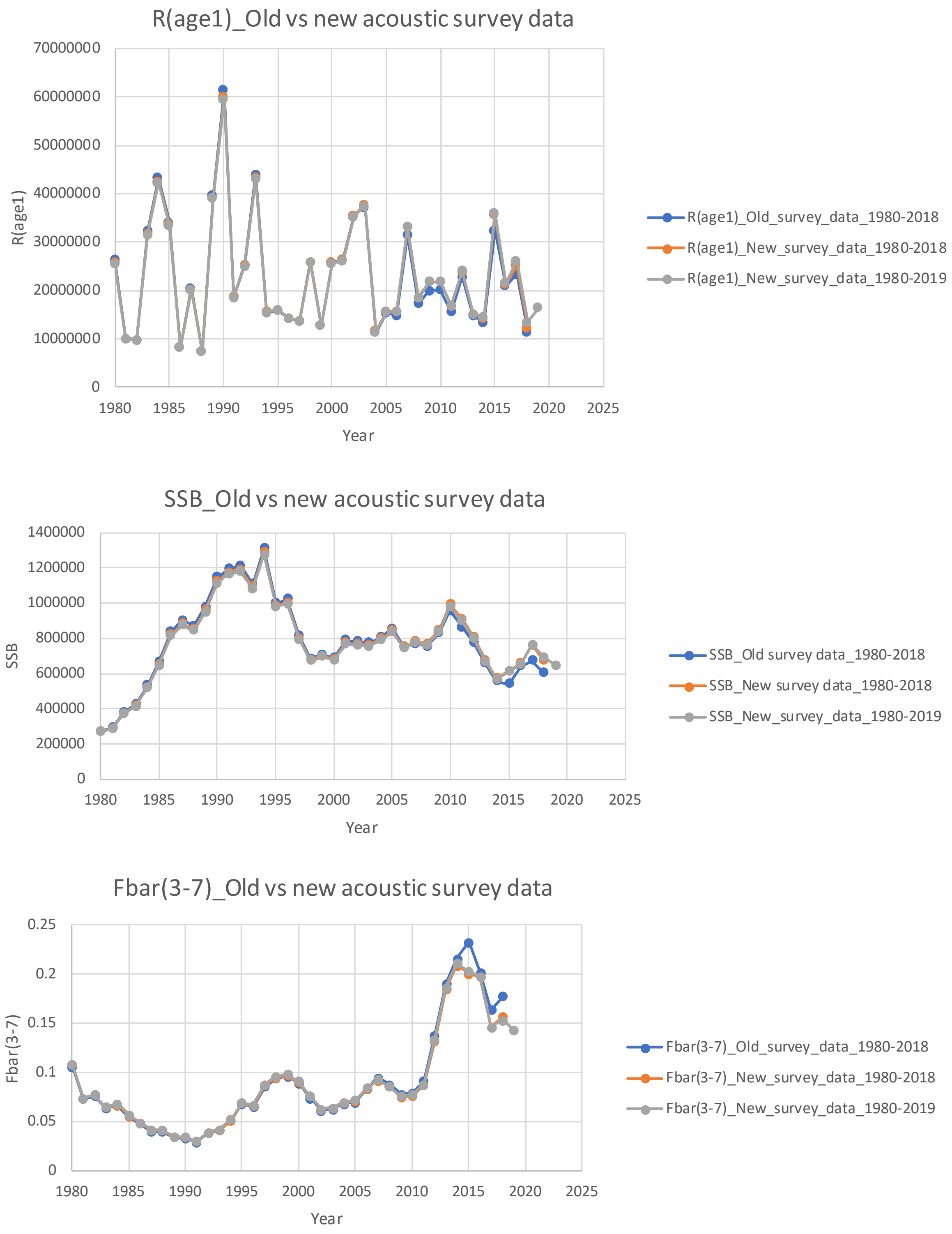


Figure 2. SAM model based estimates of R(age1), SSB and F$_{bar}$(3–7) using the old and new acoustic survey data.

Annex 4: Report of External reviewers for Gulf of Bothnian herring (SD 30 and 31)

Comments from Michael O'Malley

Invited expert to WGBIFS to review the acoustic survey index

As co-chair of WGIPS (Working Group of International Pelagic Surveys) I was asked to participate in this meeting and acted as an additional reviewer, at the request of ACOM. The meeting was held online from 1–3 December 2020 and the focus was to investigate the herring abundance indices generated for SD30 using the StoX software and to compare these with the previous old BIAS calculation methods. The amount of work completed before this meeting was impressive, and the reviewer would like to commend the effort that went into getting the entire survey calculation method for SD30 copied from the old method and converted into a new StoX project framework. This included getting all the data into the ICES DB in the correct format beforehand in order to generate the correct StoX input file format. It was inevitable that there would be errors is such a wholesale change in working up the data from the surveys to an appropriate index. The differences in total number of herring per rectangle between the two methods were generally low, but in some years differences appeared to be greater. It was difficult to get to the bottom of where the differences were in the short time we had, but great effort was put in to investigating this was done by the people involved. This included revisiting the input data to investigate possible errors there. There were some small errors found in the input data, which made proper investigation difficult during the meeting. For example, an error was found in the original BIAS calculations where wrong mean weight for herring length samples were used. Also there were errors found in some rectangles where the allocation of hauls was not given equal weight for all hauls. Comparison of NASC values showed that with some few exceptions the differences in the data were generally small and could usually be explained after investigation of the input data.

Conclusion

Generally, differences in overall estimates appeared to occur from small methodological differences between the new StoX project and the old BIAS calculation method. This is somewhat expected, and although the new StoX project for the survey was developed specifically for WGBIFS, it is not exactly the same as the old method used by WGBIFS previously. It is recommended that the herring abundance time-series of SD30 should be calculated with StoX after all investigations into errors are exhausted, and that this should be used for assessment purposes. The benefits of StoX include better transparency, harmony of calculation methods and replication across surveys. Similar comparisons should be done with all abundance index series in the group before the transition to StoX can be done for the entire area. The results from the SD30 changeover to StoX has given members valuable lessons for the other areas and this should make the process easier.

Further recommendations

Survey design issues were discussed during the meeting and the variability in survey design in the past was acknowledged. The group agreed that it would be better to cover the Bothnian Sea using parallel transects east to west if possible in the future. Survey design should consider the degree of area coverage for all survey rectangles and try to harmonise biological sampling procedures across surveys as much as possible. A lot of the discrepancies appeared to come from differences in methods across surveys (e.g. length–weight relationship procedures, stickleback protocols) and lack of familiarity between methods across surveys. The group also agreed that hauls should be distributed more evenly throughout the survey area.

The reviewer finds that the StoX survey index calculation method is appropriate to be used as input data in the stock assessment for the survey in SD30, once the group agrees that all errors in the input data have been eliminated.

Conclusions from Jim Ianelli and Santiago Cerviño

In general, the benchmark using the stock synthesis platform with the settings specified during the benchmark are considered acceptable for assessment and advice, and have features that should ensure stability as new data are added (e.g. selectivity is assumed to be constant over time).

The base model's retrospective pattern is acceptably small and is primarily due to the characteristics of the acoustic survey data. Notably, the acoustic index is relatively flat during a period when catches increased substantially in the past decade. This suggests that there must be an improved recruitment event. If the age composition data are inconsistent with such an event then this could be the cause of the observed retrospective pattern.

Future work could continue to evaluate changes in selectivity over time. For the SAM runs, the partial Fs should be considered as standard output for comparisons in future assessments. This was noted above in the section on an alternative assessment model done by the review team. Alternatively, splitting the fishery components into separate fleets may reflect spatial patterns that have varied over time. Time varying natural mortality might be explored given the high cod predation in the early period of the model.

Both assessment approaches presented have capabilities for reflecting the uncertainty in the advice. ICES should develop advice that more formally considers uncertainty. This would provide a path for including structural and estimation uncertainty from stock assessment models. For example, given that age-specific schedules of maturity and weight-at-age have changed over the assessment period, it seems appropriate to reflect this variability/uncertainty in providing advice based on future projections.

During follow-up meetings, it was reiterated that fishers have indicated their experiences in relative abundance contrasted from what the assessment indicated. Specifically, that fishing conditions had deteriorated rather than in a stable to improving state suggested by the assessment. To the extent practical, retaining input from fishing conditions during future assessments should play a role. Where such contradictions occur, assessment scientists should strive to understand and articulate (as they did in this case) why perceptions might differ.

Annex 5: List of participants

Name	Institute	Country	E-mail
Max Cardinale	Department of Aquatic Resources-SLU Aqua	Sweden	massimiliano.cardinale@slu.se
Santiago Cerviño Invited Expert	Spanish Institute of Oceanography	Spain	Santiago.cervino@ieo.es
David Gilljam	Department of Aquatic Resources-SLU Aqua	Sweden	david.gilljam@slu.se
Ruth Fernandez	International Council for the Exploration of the Sea-ICES	Denmark	Ruth.fernandez@ices.dk
Zeynep Pekcan Hekim	Department of Aquatic Resources-SLU Aqua	Sweden	zeynep.pekcan.hekim@slu.se
Noél Holmgren Chair by correspondence	Department of Aquatic Resources-SLU Aqua	Sweden	Noel.holmgren@slu.se
Jim Ianelli Invited Expert	National Oceanic and Atmospheric Administration Alaska Fisheries Science	US	jim.ianelli@noaa.gov
Pekka Jounela	Natural Resources Institute Finland	Finland	Pekka.jounela@luke.fi
Olavi Kaljuste	Department of Aquatic Resources-SLU Aqua	Sweden	olavi.kaljuste@slu.se
Juha Lilja	Natural Resources Institute Finland	Finland	Juha.lilja@luke.fi
Johan Lövgren Chair	Department of Aquatic Resources-SLU Aqua	Sweden	Johan.lovgren@slu.se
Alessandro Orio	Department of Aquatic Resources-SLU Aqua	Sweden	alessandro.orio@slu.se
Tapani Pakarinen Chair	Natural Resources Institute Finland	Finland	tapani.pakarinen@luke.fi
Jukka Pönni	Natural Resources and Bioproduction Natural Resources Institute Finland	Finland	jukka.ponni@luke.fi
Jari Raitaniemi	Natural Resources Institute Finland	Finland	jari.raitaniemi@luke.fi

Annex 6: Stock Annex for Herring (*Clupea harengus*)in subdivisions 30 and 31 (Gulf of Bothnia)

The table below provides a link of the stock annex for herring in subdivisions 30 and 31 (Gulf of Bothnia). Stock Annexes for other stocks are available on the ICES website library under the publication type "Stock Annexes". Use the search facility to find a particular Stock Annex, refining your search in the left-hand column to include the *year*, *ecoregion*, *species*, and *acronym* of the relevant ICES expert group.

Stock ID	Stock name	Last updated	Link
her.27.3031	Herring (*Clupea harengus*) in subdivisions 3031 (Gulf of Bothnia)	February 2021	Herring in SD 3031

Annex 7: Working Documents

The following working documents were presented at WKCLuB (February 2020 meeting) and are inserted in full in the following pages:

WD1: Input data for Stock assessment of herring in the Gulf of Bothnia (ICES SD 30–31), February 2020. Jukka Pönni, Zeynep Hekim, Jari Ratianiemi and Pekka Jounela.

WD2: Stock assessment of herring in the Gulf of Bothnia (ICES SD 30–31), February 2020. Massimiliano Cardinale and Alessandro Orio

WD3: Updated exploratory run using SAM, February 2020. Zeynep Hekim.

WD4: Additional reviewers' comments related to WD2 and WD3. February 2020.

Data for Stock assessment of herring in the Gulf of Bothnia (ICES SD 30-31)

By Jukka Pönni, Zeynep Hekim, Jari Raitaniemi and Pekka Jounela

1. Commercial catch

Finnish commercial herring catch statistics is based on catch notifications submitted by fishermen at set intervals. The application of the Act (1139/94) on implementing the Common Fisheries Policy of the European Union obliges all commercial fishermen to submit catch notifications.

The discards are negligible in both countries' commercial fisheries (i.e. Finland and Sweden) and therefore not sampled either. Also, the information of discards from Finnish fishermen's reports is not used in assessment, but the Swedish reported discards are added to the total catches.

The fishing data of vessels ≥ 10 metres long are entered in the EU fishing logbook. The data entered are the dates of fishing by fishing trip, the size of the catch by species, the fishing (statistical) rectangle, the gear and number of gears used in fishing, and the trawling time in hours. A fisherman is obliged to keep an up to date logbook onboard his vessel. The logbook must be returned to the regional authorities within 48 hours of the catch being landed.

With the exception of salmon catches, the Finnish fishing data of vessels ≤ 10 metres long are entered monthly in a coastal fishery form. The data entered are the size of the catch by species by the statistical rectangle, the type and number of gears used in fishing, and the number of fishing days. The forms must be returned to the regional authorities by the fifth day of the following month. All logbooks and most of the other catch notification forms are checked by national authorities.

The proportion of the Baltic herring catch, landed in Finland for the food and processing industry in relation to the total catch of that species, is estimated with the aid of the fish purchasing register that is maintained by the Ministry of Agriculture and Forestry.

Because all the main fisheries (pelagic trawling, deep mid-water trawling and trap-nets) have different exploitation patterns, their catches are also sampled separately. The sampling in the Gulf of Bothnia herring fishery is performed according to EU DCF requirements, covering 12 strata (3 fleets and 4 year-quarters).

Since the study projects funded by DG XIV (International Baltic Sea Sampling Programs I & II) in 1998–2001, a length stratified sub-sampling scheme has been applied to estimate age compositions of the Finnish catches of Baltic herring. This sampling scheme is designed to be compatible with international databases and uses standardized methodologies in data processing. Baltic herring samples are collected mainly in fishing harbours and, if necessary, also on board commercial fishing vessels. In the sampling scheme the annual life cycle of Baltic herring and the presence of the ice coverage during the winter in the Gulf of Bothnia have been taken into account. Because of icing conditions, the three fishing gears are not in use year-round (e.g. trap net fishery is usually conducted only in spawning time during quarters 2 and 3). The sampling effort is roughly based on the proportions of catches in different fisheries. Moreover, the sampling intensity in general is locally adjusted during the year according to temporal and regional changes in fisheries. The seasonal herring fishing intensity is predominantly dependent on the TAC, which may cause fishing restrictions in certain fisheries and/or seasons and may therefore change the sampling intensity from the original plan. A minimum coverage target is at least one sample by fishery per month (or three samples by fishery per year-quarter). The sampling strategy is to have age-length samples from all major gears in each quarter.

The Finnish and Swedish input files are uploaded to ICES InterCatch database. The data can also be found in the national laboratories and with the stock co-ordinator. The national data have been aggregated to international data in InterCatch.

Table 1. Description of the types of data available per country.

	Kind of data				
Country	Caton (catch in weight)	Canum (catch-at-age in numbers)	Weca (weight-at-age in the catch)	Matprop (proportion mature by age)	Length composition in catch
Finland	x	x	x	x	x
Sweden	x	x	x		x

2. Biological sampling

The age and the individual weight data is obtained from both Finnish and Swedish landings from all year-quarters as well as from the catch samples in acoustic surveys in 3rd and 4th quarter. The annual weights at age are weighted by the year-quarterly catch-numbers. The maturity ogives are based on the proportions of mature individuals of each age group before spawning time in the Finnish sampling of commercial data, and are updated every year.

2.1 Calculation of catch at age

In Finland the calculation of catch at age is based on year-quarterly performed length-stratified random sampling of individual fish (at minimum 10 aged individual fish from all prevailing 0.5 cm length-classes) and length-samples of at least 300 specimens per sample from different commercial fisheries per quarter. The average number of individual-samples is 1100 from commercial fisheries and 2500 from surveys in SD 30 and 600 from commercial fisheries in SD 31 annually, and the average number of length measurements is 18500 and 6000 respectively.

The quarterly collected length distributions (from gear-specific length sampling) are converted into age distributions with year quarterly prepared age-length keys, ALKs, which are derived from the sampling of individuals from all gears combined.

The quarterly catches from the main herring fisheries (OTM [Midwater Otter Trawl, single trawling] + PTM [Midwater Otter Trawl, pair trawling] carried out in midwater and deep midwater trawling and trapnets, FPN [pound nets]+ FYK [fyke nets]) are divided by the mean weight of the herring from length samples of respective fisheries in order to get the total catch number of fish for all strata (for all above mentioned national fisheries, 4 quarters). The total catch numbers from each fishery and quarter are then multiplied by the proportions of the age-classes in the age distributions and summed up to get the annual catch at age.

In Sweden, the length-samples of at least 300 specimens per sample from two (main) commercial fisheries (bottom trawls (XTB) and gillnets (GNS)) in SD 30 and only from gillnets in SD 31 are collected quarterly each year. The catches of pelagic trawls (OTM and PTM) fisheries are not sampled, although they can constitute more than 70% of the total Swedish catches (i.e. 2018). Length-stratified random sampling of individual fish (around 20 aged individual fish from all prevailing 0.5 cm length-classes per quarter) is performed only for gillnet fisheries. In SD 30, the average total number of annual length measurements is 5600 from bottom trawls and 2300 from gillnet fisheries, and the average total number of sampled fish individuals is 490, and in SD 31 the average total number of annual length measurements is 1700, and the average number of sampled fish individuals is 450.

The length distributions (from length sampling) by gear are converted into age distributions with quarterly prepared age-length keys (ALKs) using all gears together. For that purpose, additionally Finnish ALK and mean weight at length data from trawl fisheries are borrowed. Finnish deep midwater-trawl ALK and weights are used for Swedish Bottom (demersal) trawls.

The calculation of total annual catch-at-age follows the same procedure as in Finland.

2.2 Calculation of mean weight

The mean weights at age are derived from the individual data collected from commercial catches all year round as well as from the individual data of acoustic survey trawl samples during September-October (2600 individuals annually), and averaged over year and quarters. The annual mean weights at age for assessment are derived by weighting the year-quarterly mean weights by the year-quarterly catch numbers.

2.3 Maturity

The maturities are defined from the individual data that is collected all the year round from commercial catches with other so called “stock related variables” as length, weight and age, and from the trawl samples of the acoustic survey. The data for the maturity ogive used in assessments is collected from commercial samples before spawning (i.e. January to March in SD 30 and March to May in SD 31), because the idea is to get the proportion of spawners by age from the whole population, before the spawning part separates itself from non-spawners by approaching the coastline spawning areas.

The share of mature fish in each age-group is calculated from annual data and the annual number of the individual samples for maturity definitions that are used for the maturity ogives has been on average (2010–2015) 283 in SD 30 and 212 in SD 31.

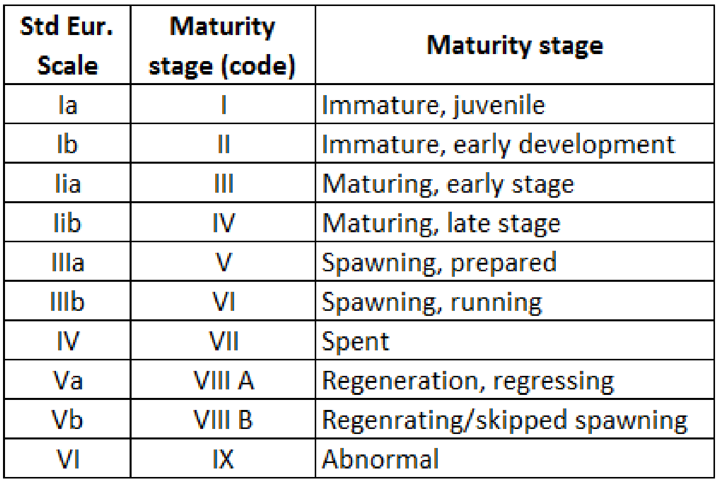
The maturity scale (Table 2) in use is the modified European standard 9-stage scale and the same scale is used both in Finland and Sweden. The stages II–VIII (VIII–A and VIII–B) are considered mature while stage I and IX are counted as “non-mature” although stage nine (abnormal) is usually mature, but not accounted to take part to spawning.

The maturities collected during a Swedish acoustic survey in 4th quarter and the maturities derived from Finnish 1st quarter sampling of commercial catches have showed very small differences.

In the WKPELA 2012 benchmark (ICES, 2012), the sensitivity of the annually changing proportions of spawners in age-groups was tested (by several types of averages over time[1]) and even though there are clearly visible annual changes in mostly 2-year-olds, there was only negligible impact to e.g. the estimates of SSB. It was concluded then that it was still better to have the latest real information on maturity at age.

The reason for the "instability" was found to be the high inter-annual variation in the maturation of 2-year olds in the whole time-series and especially in 2010. The maturity calculations from raw data were examined carefully, and no mistakes were revealed.

Table 2. Maturity scale in use in Finland and Sweden.



Std Eur. Scale	Maturity stage (code)	Maturity stage
Ia	I	Immature, juvenile
Ib	II	Immature, early development
Iia	III	Maturing, early stage
Iib	IV	Maturing, late stage
IIIa	V	Spawning, prepared
IIIb	VI	Spawning, running
IV	VII	Spent
Va	VIII A	Regeneration, regressing
Vb	VIII B	Regenrating/skipped spawning
VI	IX	Abnormal

From 2002 to 2006, Finnish samples were age determined with two different methods in parallel, using traditionally whole otoliths and as a new method, neutral red stained slices of cut otoliths. The effects of the age determination method were presented at the WGBFAS meeting in 2006 (Raitaniemi and Pönni, 2006, working document). The method affects the age distribution as well as the proportion of mature fish at age. Especially in old age groups (from age 5 or 6 on), determination from cut otoliths generally results in an older assessed age compared to whole otoliths. In the comparison, the numbers at age in the total catch differed about 2% on average, but ranged from 0.4% to 52% depending on year and age. On average the proportion of 4 to 8 years old individuals in the catch was 11% lower, and the proportion of ages 9+ was 32% higher, when using neutral red stained slices compared to whole otoliths.

According to Peltonen et al. (2002), the agreement between the determinations of different age readers was better with the cut otoliths technique than with whole otoliths. A combination of age data from Finnish cut otoliths (representing 98% of the catch) and

[1] *Four new combinations of maturity ogives were introduced to XSA (maturity ogive with 3- and 5 years running averages for the whole time series, constant maturity ogive for the whole time series as an average of the whole time series and two different averages over the time series according to periods before and after the alleged regime shift (1973—1988 and 1989—2010)). Resulting estimates of SSB were compared to the annually updated maturity ogive in SPALY run, and the differences were found to be negligible with the exception of year 2010 only.*

Swedish whole otoliths (representing 2% of the catch) was used between 2002–2006. The slicing method was calibrated between Finland and Sweden in 2007, and it has been applied also to Swedish catches as well as Bothnian Sea surveys since 2007. Since age determination using cut otoliths is considered to be more accurate (Raitaniemi and Pönni, 2006), this method is used as the standard method for ageing all the samples, and the time series including ages from whole otoliths from 1980–2001 and cut otoliths from 2002 onwards is used in the assessments of this stock.

2.4 Natural mortality

An age-varying natural mortality is assumed to be constant for the entire time series (Figure 1, Table 1). M was estimated based on the methods described in Then et al. (2015) and Lorenzen (1996). Then et al. (2015) estimation of M is based on maximum observed age (*tmax* = 25) and parameters of the von Bertalanffy growth curve as derived from www.fishbase.org for the same area. The Lorenzen type (Lorenzen, 1996) of M-at-age function assumes a declining relationship between M and the mean weight of fish in successively older age classes. The growth and the length-weight parameters used for the M estimation are reported in Table 2. In all model configurations tested, M gradually decrease from 0.563 to 0.257 for ages 0 to 20. In order to reduce the number of parameters to be used in the model, natural mortality was set using 6 breaks: age 0.5, 1.5, 3.5, 5.5, 8.5 and 15.5, where M for the adjacent ages is simply linearly interpolated using the values estimated for the age breaks.

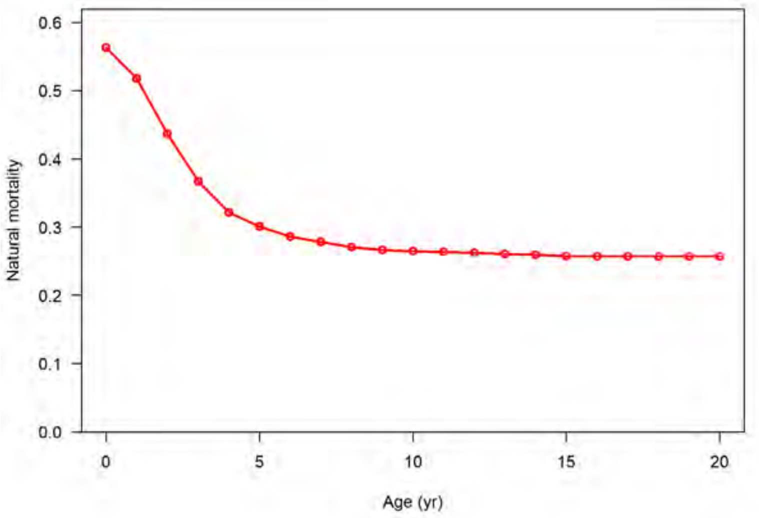


Figure 1. Herring SDs 30-31. The age-specific natural mortality used in the model.

Table 1. Herring SDs 30-31. Natural mortality vector by breaks used in the model.

Age 0.5	Age 1.5	Age 3.5	Age 5.5	Age 8.5	Age 15.5
0.563	0.472	0.332	0.290	0.267	0.257

Table 2. Herring SDs 30-31. Parameters used to estimate the natural mortality vector by age.

Life history parameters		
k (combined sex)	0.34	From Fishbase
L_{inf} (combined sex)	21	From Fishbase
t_0 (combined sex)	-1.1	From Fishbase
a	3.99E-03	Estimated from LW BIAS survey data
b	3.145258	Estimated from LW BIAS survey data

2.5 Surveys

Annual hydroacoustic surveys have been conducted in SD 30 in October from 2007 until 2010 with Swedish R/V Argos. In 2011 and in 2012, the survey was performed with the Danish R/V Dana, 2013–2016 with Finnish R/V Aranda, and in 2017 with R/V Dana again. This survey is co-ordinated by ICES within the Baltic International Acoustic Surveys (BIAS). The annual survey-indices are collected and calculated with standardised methods within the international coordination of ICES WGBIFS and stored in international databases. The actual calculations have been performed in years 2007–2012 in the Swedish marine research institute (Havsfiskelaboratoriet) by Niklas Larson and from 2013 onwards in the Natural Resources Institute Finland by Juha Lilja.

Based on the implementation progress to development StoX software for calculations of WGBIFS acoustic stock indexes it was decided that members of WGBIFS StoX task sub-group should analyze their national survey data with StoX software and compare the results with their official results. In SD 30, the comparison between StoX and official results showed significant difference between 2013 and 2015 but no differences was found between 2016-2018. Consequently, all official results (based on Excel spreadsheets) were recalculated and an error was discovered in calculations for 2013-2015 indexes. The error was corrected and updated values were presented.

The acoustic survey has been considered a reliable tuning fleet and has been included into the SD 30 assessment in 2013 after an independent review process (ICES, 2015).

The SD 30 acoustic estimates are used as abundance indices (tuning fleet) for the assessment of Gulf of Bothnia herring stock (SDs 30 & 31) (see the text table in section C, Assessment: data and method). In the acoustic tuning fleet, age-groups 0–15 (true ages) are applied (Figure 1).

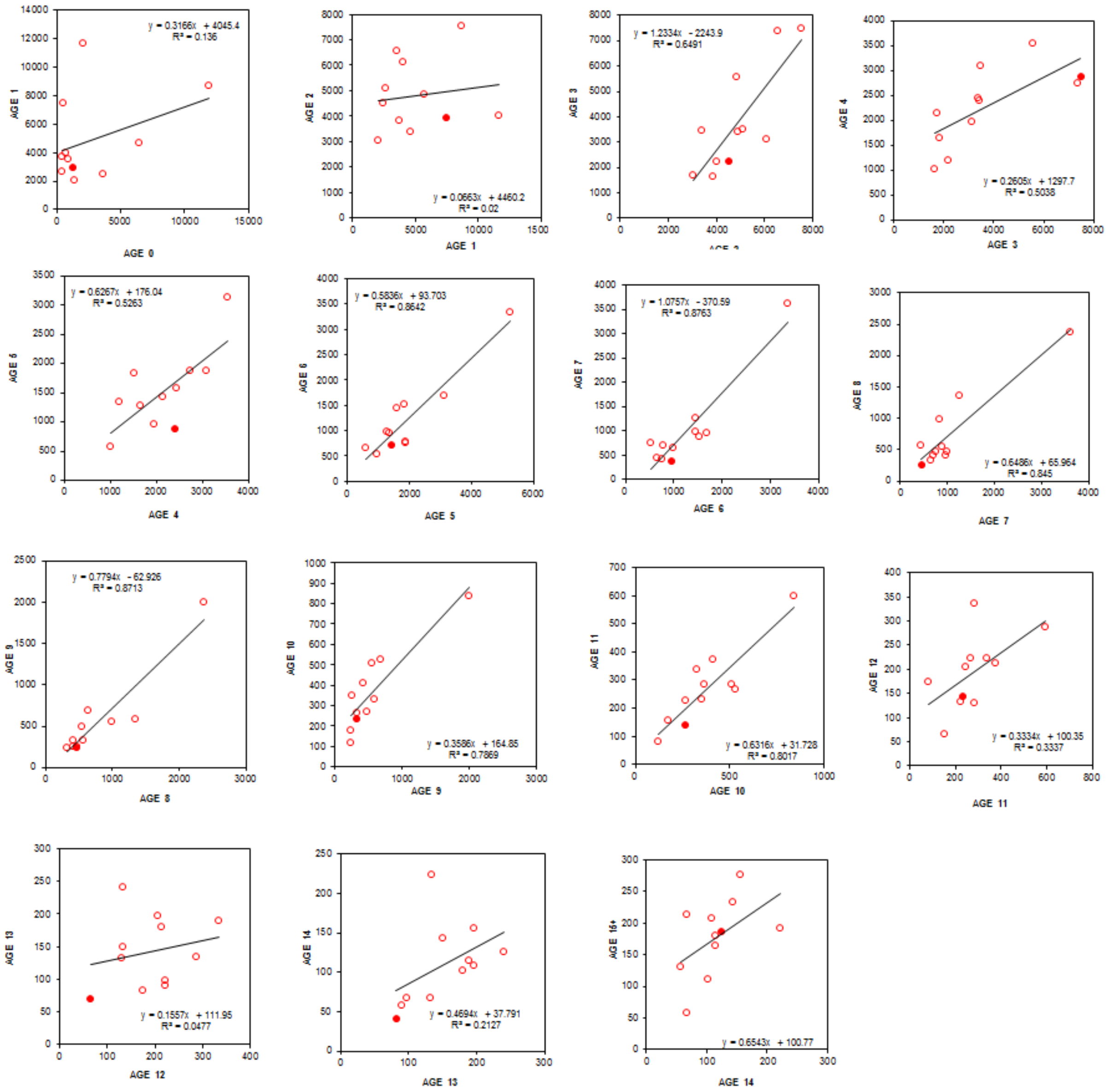


Figure 1. Consistency between consecutive age-classes in acoustic tuning fleet.

The coverage of the acoustic transects and trawl samples has mostly been good. In 2012 the coverage was only half of the “normal” because of a sudden 50% reduction in funding. In 2014 there were problems with the fishing gear, which reduced the trawl hauls, but the spatial acoustic coverage was not affected significantly. In 2015 a storm damaged the ship so that the most northern part of the area had to be skipped due to lack of time after fixing the damage in harbour.

The 2012, 50% reduction in the survey effort, as well as the 2014 and 2015 results were, however, considered acceptable for the index by the survey expert working group, WGBIFS (ICES, 2013; 2015; 2016).

The survey is based on Baltic International Acoustic Surveys (BIAS) manual (ICES, 2016, in annex) with the aim of 60 Nm of acoustic transect and 2 trawl hauls per statistical rectangle. In the catch sampling, at least 300 fish are measured in 0.5 cm length-classes for length distributions, and 10 individuals from all prevailing length-classes are aged per rectangle, comprising normally about 20 000 length-measurements and 2600 age-readings annually.

References

ICES. 2012. Report of the Benchmark Workshop on Pelagic Stocks (WKPELA 2012), 13–17 February 2012, Copenhagen, Denmark. ICES CM 2012/ACOM:47. 525 pp.

ICES. 2013. Report of the Baltic International Fish Survey Working Group (WGBIFS), 21–25 March 2013, Tartu, Estonia. ICES CM 2013/SSGESST:

ICES. 2015. Report of the Baltic International Fish Survey Working Group (WGBIFS), 23–27 March 2015, Öregrund, Sweden. ICES CM 2015/SSGESST:

ICES. 2016 Report of the Baltic International Fish Survey Working Group (WGBIFS), 30 March-4 April 2016, Rostock, Germany. ICES CM 2016/SSGIEOM:07. 8 pp.

Lorenzen K. (1996) – The relationship between body weight and natural mortality in juvenile and adult fish: a comparison of natural ecosystems and aquaculture. Journal of Fish Biology, 49:627-647.

Peltonen, H., Raitaniemi, J., Parmanne, R., Eklund, J., Nyberg, K., and Halling, F. 2002. Age determination of Baltic herring from whole otoliths and from neutral red stained otolith cross sections. – ICES Journal of Marine Science, 59: 323–332.

Then, A. Y., Hoenig, J. M., Hall, N. G., Hewitt, D. A. (2014). Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. ICES Journal of Marine Science, 72(1), 82-92.

Stock assessment of herring in the Gulf of Bothnia (ICES SD 30-31)

By Massimliano Cardinale and Alessandro Orio

Assessment method and settings

Assessment of herring in SDs 30-31 was conducted using the Stock Synthesis (SS) model (Methot & Wetzel 2013). Stock Synthesis is programmed in the ADMB C++ software and searches for the set of parameter values that maximize the goodness-of-fit, then calculates the variance of these parameters using inverse Hessian and MCMC methods. The assessment was conducted using the 3.30 version of the Stock Synthesis software under the windows platform.

Uncertainty measures and likelihood

The total likelihood of the model is composed of a number of components, including the fit to the survey and CPUE indices, tag recovery data (when tagging data are used), fishery length frequency data, age compositions and catch data. There are also contributions to the total likelihood from the recruitment deviates and priors on the individual model parameters (if any). The model is configured to fit the catch almost exactly so the catch component of the likelihood is generally small (although catch penalties might be created and catches are entered with uncertainty). Details of the formulation of the individual components of the likelihood are provided in Methot & Wetzel (2013).

Samples sizes, CVs, data weighting

For the commercial fleet the CV of the catches was set to 0.05. The CV of the initial catches of the commercial fleet was set to 0.1 to add extra variability. The annual sample size associated with the age distribution data for commercial catches is reported as number of trips sampled.

The CV of both the acoustic and trapnet survey indices are not available. Therefore, a value of 0.1 is assumed for all years.

The relative weighting of the age compositions of the reference model were estimated using Francis method as implemented in r4ss package. The Hessian matrix computed at the mode of the posterior distribution was used to obtain estimates of the covariance matrix, which was used in combination with the Delta method to compute approximate confidence intervals for parameters of interest.

Assessment model runs: Reference model

The assessment model of herring in SDs 30-31 is a one area, annual, age-based model where the population is comprised of 20+ age-classes (with age 20 representing a plus group) with sexes combined (male and females are modelled together).

The model starts in 1963 and the initial population age structure was assumed to be in an exploited state, so that the initial catches was assumed to be the average of last three years (1963-1965) in the time series. Fishing mortality was modelled using hybrid F method (Methot & Wetzel 2013). Option 5 was selected for the F report basis; this option represents a recent addition to Stock Synthesis and corresponds to the fishing mortality requested by the ICES framework (i.e. simple unweighted average of the F of the age classes chosen to represent the F_{bar} (age 3-7)).

Spawning stock biomass and recruitment

Spawning biomass was estimated at the beginning of the year and it was considered proportional to fecundity. In the model, the recruitment was assumed to be only a single event occurring at the beginning of the year. Recruitment was derived from a Beverton and Holt (BH) stock recruitment relationship (SRR) and variation in recruitment was estimated as deviations from the SRR. Recruitment deviates were estimated for 1963 to 2018 (55 annual deviations). Recruitment deviates were assumed to have a standard deviation (σ_R) of 0.6, which was derived using the likelihood profile function in r4ss (See section below). The reference model estimates steepness (*h*) for the SRR within the model using with a full Beta prior of 0.74 with a standard deviation of 0.113 as derived for herring in Myers et al. (1999).

Growth, weights and maturity

Empirical weight at age matrices for both commercial fleet and survey indices are provided as input for the model and are estimated using commercial and survey data. Maturity at age matrix is also provided as input and derived from commercial data. Details on how weight and maturity at age were derived are included in the stock annex.

Natural mortality

An age-varying natural mortality is assumed to be constant for the entire time series (Figure 1, Table 1). M was estimated based on the methods described in Then et al. (2015) and Lorenzen (1996). Then et al. (2015) estimation of M is based on maximum age (t_{max} = 25) and parameters of the von Bertalanffy growth curve as derived from www.fishbase.org for the same area. The Lorenzen type (Lorenzen, 1996) of M-at-age function assumes a declining relationship between M and the mean weight of fish in successively older age classes. The growth and the length-weight parameters used for the M estimation are reported in Table 2. In all model configurations tested, M gradually decrease from 0.563 to 0.257 for ages 0 to 20. In order to reduce the number of parameters to be used in the model, natural mortality was set using 6 breaks: age 0.5, 1, 3, 5, 8 and 15, where M for the adjacent ages is simply linearly interpolated using the values estimated for the age breaks.

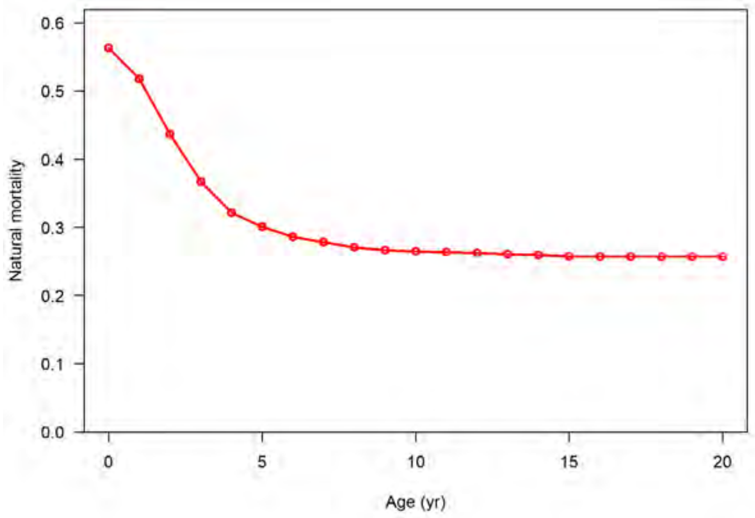


Figure 1. Herring SDs 30-31. The age-specific natural mortality used in the model.

Table 1. Herring SDs 30-31. Natural mortality vector by breaks used in the model.

Age 0.5	Age 1	Age 3	Age 5	Age 8	Age 15
0.563	0.472	0.332	0.290	0.267	0.257

Table 2. Herring SDs 30-31. Parameters used to estimate the natural mortality vector by age.

Life history parameters		
k (combined sex)	0.34	From Fishbase
L_{inf} (combined sex)	21	From Fishbase
t_0 (combined sex)	-1.1	From Fishbase

a	3.99E-03	Estimated from LW BIAS survey data
b	3.145258	Estimated from LW BIAS survey data

Fishery dynamics

Fishery selectivity of the reference model is assumed to be age-specific and time-invariant. For both commercial fleet and surveys, a random walk selectivity was used. This selectivity pattern provides for a random walk in ln(selectivity). For each age $a \geq A_{\mathrm{min}}$, where A_{min} is the minimum age for which selectivity is allowed to be non-zero, there is a selectivity parameter, p_a, controlling the changing selectivity from age $a - 1$ to age a. All data inputs are summarized in Table 3 while in Table 4 the configuration of the reference model is reported.

Table 3. Herring SDs 30-31. Input data used in the Stock Synthesis models.

TYPE	NAME	YEAR RANGE	RANGE
Catches	Catches in tonnes for each year	1963- 2018	
Age compositions	Catch in numbers (thousand) per age class	Commercial fleet: 1980-2018 Acoustic survey: 2007-2018 Trapnet survey: 1990-2006	0 – 15+
Maturity ogives	Empirical maturity at age estimated from commercial data		
Natural mortality	Natural mortality by age class costant for the entire time series derived from Then et al., 2015		0 - 20+
Surveys indices	Density index from acoustic survey and biomass index from trapnet survey	Acoustic survey: 2007-2018 Trapnet survey: 1990-2006	
SSB index	SSB proportional to fecundity		

Table 4. Herring SDs 30-31. Settings of the Stock Synthesis assessment reference model. The table columns show: number of estimated parameters, the initial values (from which the numerical optimization is started), the intervals allowed for the parameters, the priors used, the value estimated by the model and its standard deviation. Parameters in bold are set and not estimated by the model.

Parameter	Number estimated	Initial value	Bounds (low,high)	Prior	Value (MLE)	Standard deviation
Natural mortality (age classes 0.5, 1, 3, 5, 8, 15)		**0.563, 0.472, 0.332, 0.290, 0.267, 0.257**				
Stock and recruitment						
Ln(R_0)	1	18.03	(16, 25)	No_prior	17.36	0.07
Steepness (h)	1	0.66	(0.1, 1)	0.74	0.77	0.10
Recruitment variability (σ_R)		**0.60**				
Ln (Recruitment deviation): 1963 - 2018	55					
Recruitment autocorrelation		**0**				
Initial catches		Average of 1963-1965				
Commercial fleet	1	0.2	(0.001, 1)	No_prior	0.034	0.006
Selectivity (random walk)						
Commercial fleet						
Change from age1 to age2	1	1.45	(-5, 9)	No_prior	1.31	0.07
Change from age2 to age3	1	0.4	(-5, 9)	No_prior	0.37	0.06
Change from age3 to age4	1	0.15	(-5, 9)	No_prior	0.14	0.06
Change from age4 to age5	1	0.14	(-5, 9)	No_prior	0.14	0.07
Change from age5 to age6	1	0.03	(-5, 9)	No_prior	0.07	0.08

Change from age6 to age7	1	0.01	(-5, 9)	No_prior	0.04	0.08
Acoustic Survey						
Change from age1 to age2	1	0.6	(-5, 9)	No_prior	0.50	0.17
Change from age2 to age3	1	0.2	(-5, 9)	No_prior	0.35	0.18
Change from age3 to age4	1	0.02	(-5, 9)	No_prior	-0.05	0.22
Change from age4 to age5	1	0.11	(-5, 9)	No_prior	0.07	0.25
Change from age5 to age6	1	0.14	(-5, 9)	No_prior	0.18	0.22
Trapnet Survey						
Change from age3 to age4	1	0.30	(-5, 9)	No_prior	0.13	0.15
Catchability						
Acoustic survey						
Ln(Q) – catchability		**-2.47811**				
Extra variability added to input standard deviation		**0.001**				
Trapnet survey						
Ln(Q) – catchability		**3.86604**				
Extra variability added to input standard deviation		**0.001**				

Exploratory runs

The following alternative configurations were explored:

Table 5. Herring SDs 30-31. Alternative model runs with different configurations.

id	Number of age 0 individuals from the BIAS acoustic survey	Selectivity of the fleet	Weighting
Reference_run_age0	Included	Time invariant	Francis
Reference_run_DIRICH	Excluded	Time invariant	Dirichelet

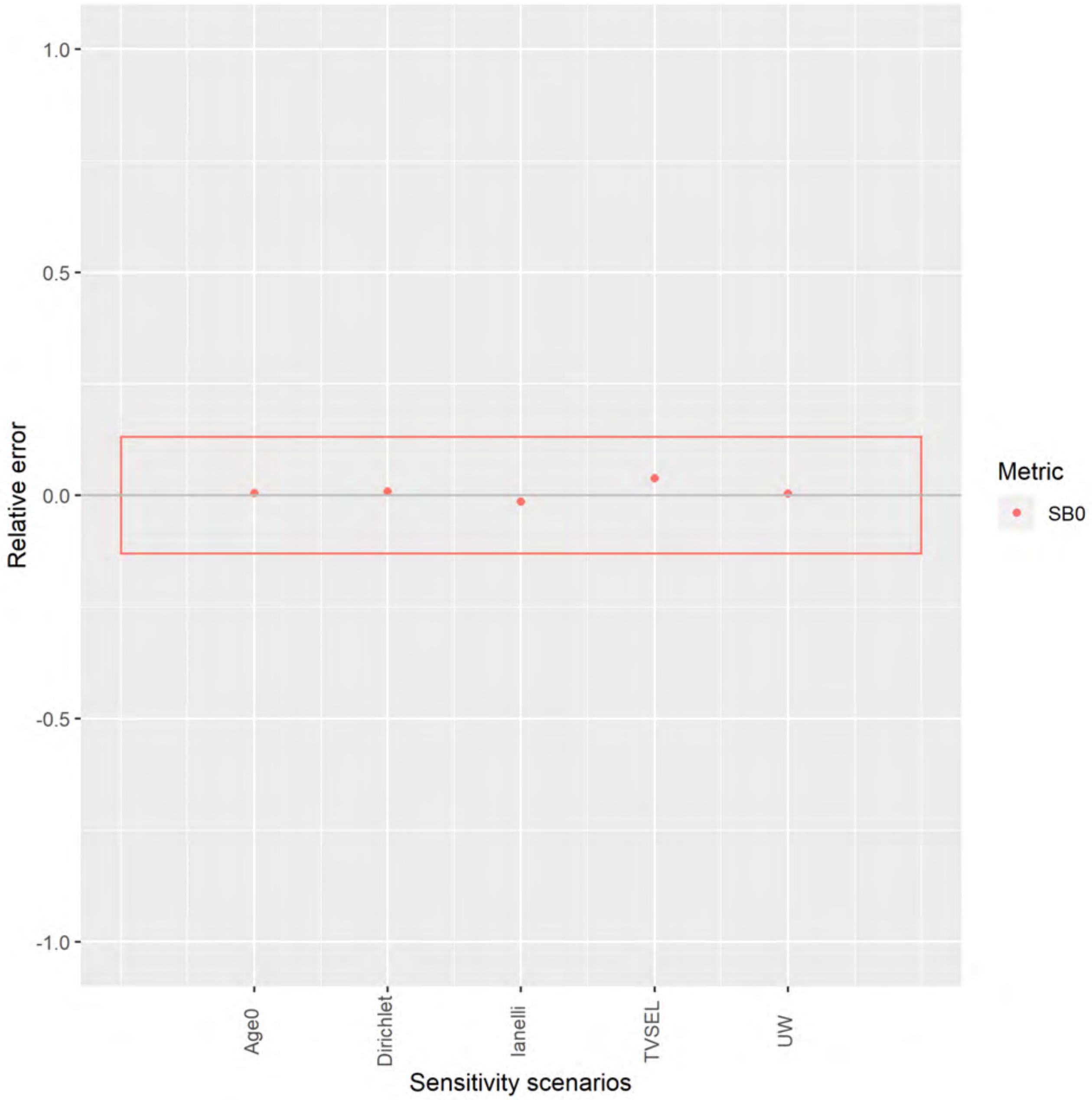
Reference_run_Ianelli	Excluded	Time invariant	Ianelli
Reference_run_TVSEL	Excluded	Time variant	Francis
Reference_run_UW	Excluded	Time invariant	None

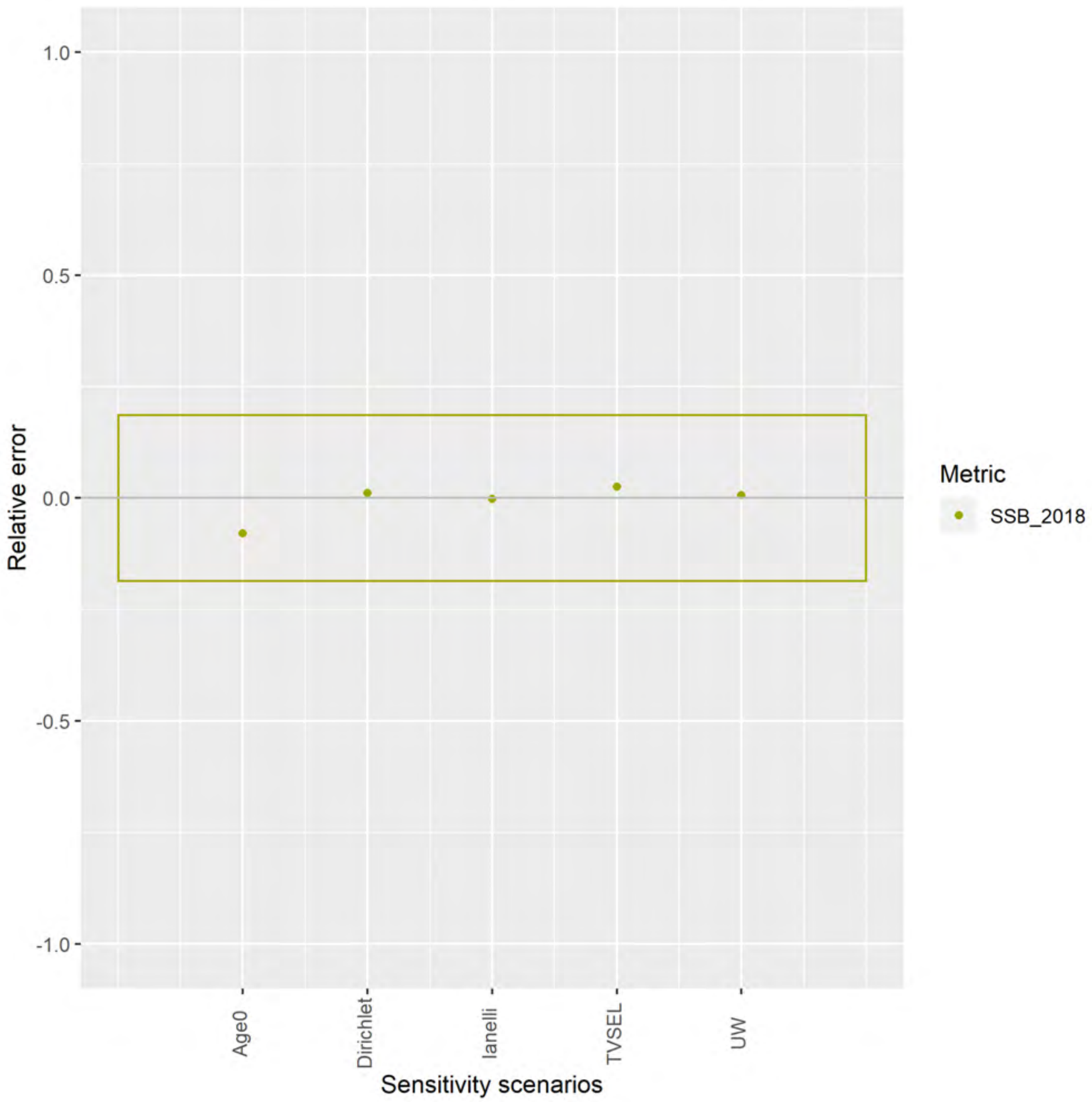
The five alternative models reported in Table 5 generally did not improve the likelihood or convergence of the model, or the fits to different data sources, compared to the reference run (Table 6). The reference run achieved the lowest value of AIC (although the comparison with Age0 model is not statistically correct as the data used in the two models are different) similar values of final convergence.

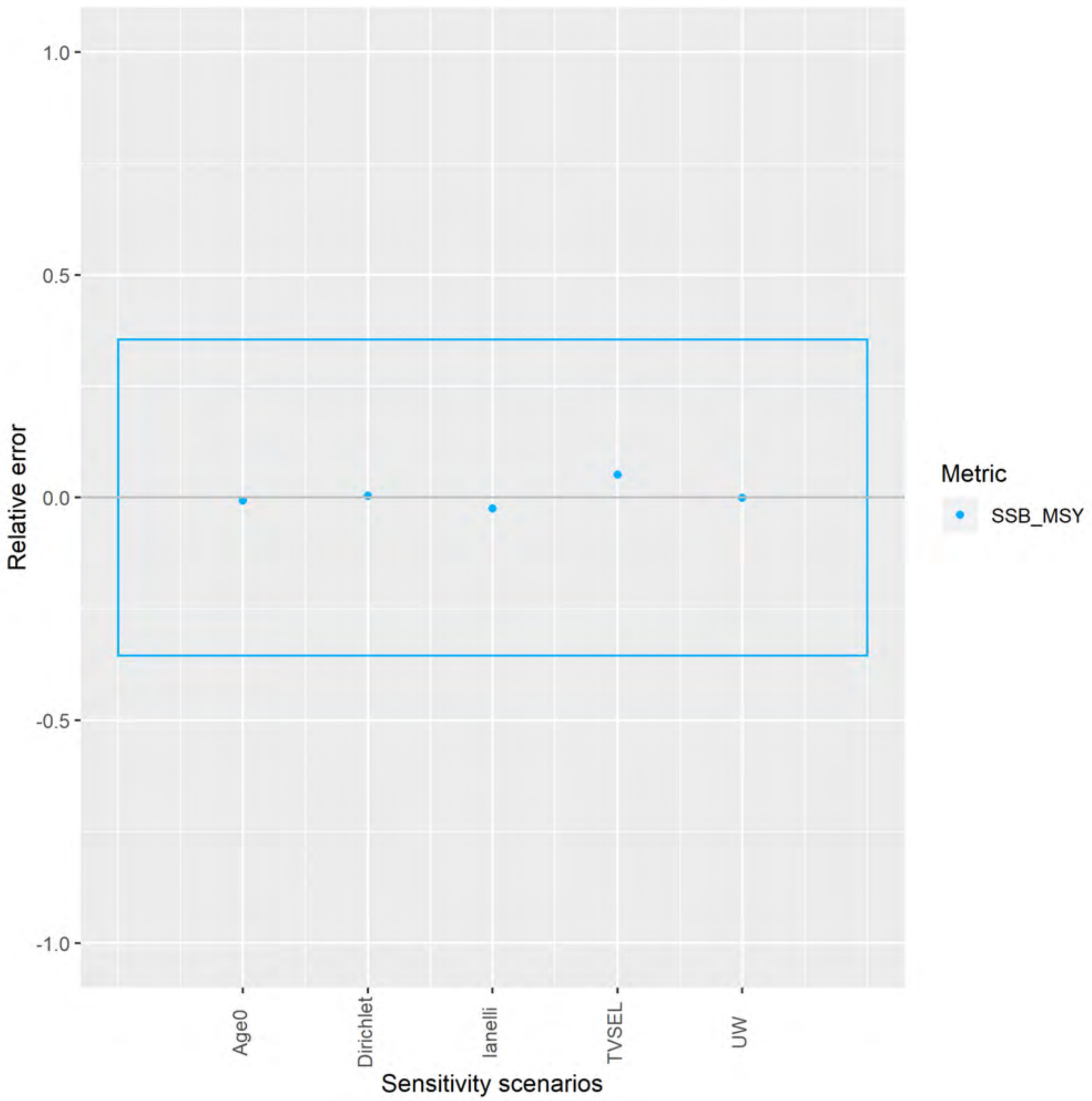
Table 6. Herring SDs 30-31. Likelihood component, parameter values and derived model quantities for the alternative model configurations. The values in the likelihood component of each model indicate changes in likelihood units compared to the reference model. Values +/- 2 likelihood units are considered significantly different.

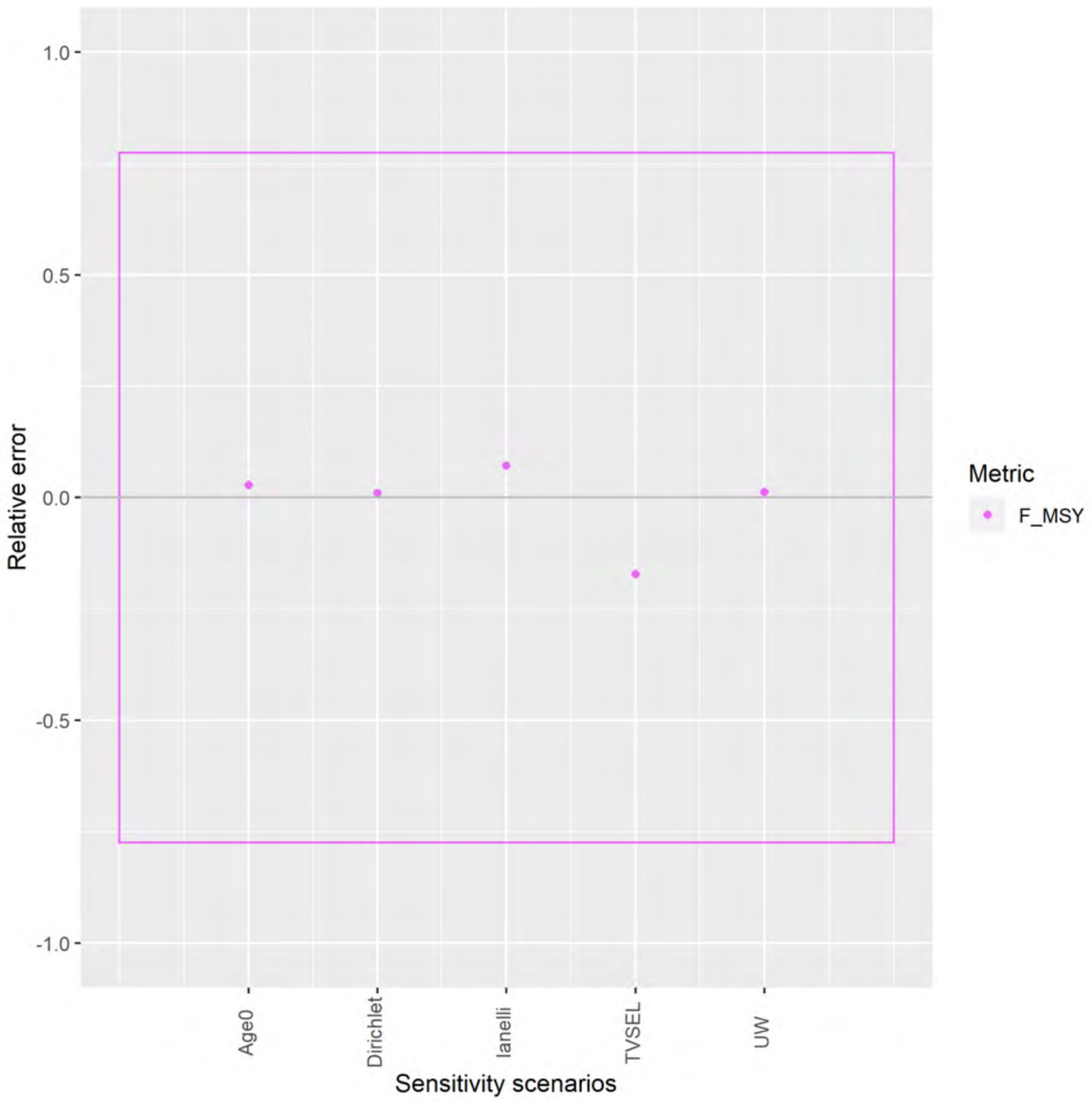
Type	Reference	Age0	Dirichlet	Ianelli	TVSEL	UW
Convergence	5.3E-06	2.8E-05	2.6E-05	1.9E-05	1.1E-05	2.1E-06
TOTAL_likelihood	128	144	149	303	102	147
AIC	404	437	447	754	545	442
deltaAIC	0.0	33.7	43.1	349.9	141.4	38.3
Survey_likelihood						
ALL		1.9	1.1	7.6	-1.8	1.0
Acoustics		17.2	15.8	18.1	14.4	15.8
Trapnet		11.4	12.0	16.2	10.5	12.0
Age_likelihood						
ALL		14.3	19.2	159.1	-24.4	16.7
Fleet		0.0	21.7	85.3	-25.0	20.4
Acoustics		14.3	0.5	15.7	0.4	-0.1
Trapnet		0.0	-2.9	58.1	0.2	-3.5
Derived quantities						
SB0	517425	519650	521995	509920	537000	519300
SSB_2018	407361	374826	411624	406500	417677	409588
F_2018	0.18	0.20	0.18	0.19	0.17	0.18
SSB_MSY	170574	169334	171119	166364	179325	170316

The alternative model configuration with the inclusion of BIAS survey estimates of age 0 individuals estimates a smaller SSB and a higher F in 2018 compared to the reference model and also to all other model configurations. Otherwise, the results from the different model configurations are rather similar (Figure 2).




1.0
0.5
0.0
-0.5
-1.0
Relative error
Age0
Dirichlet
Ianelli
TVSEL
UW
Sensitivity scenarios
Metric
● SSB_2018


1.0
0.5
0.0
-0.5
-1.0
Relative error
Age0
Dirichlet
Ianelli
TVSEL
UW
Sensitivity scenarios
Metric
● SSB_MSY


1.0
0.5
0.0
-0.5
-1.0
Relative error
Age0
Dirichlet
Ianelli
TVSEL
UW
Sensitivity scenarios
Metric
F_MSY

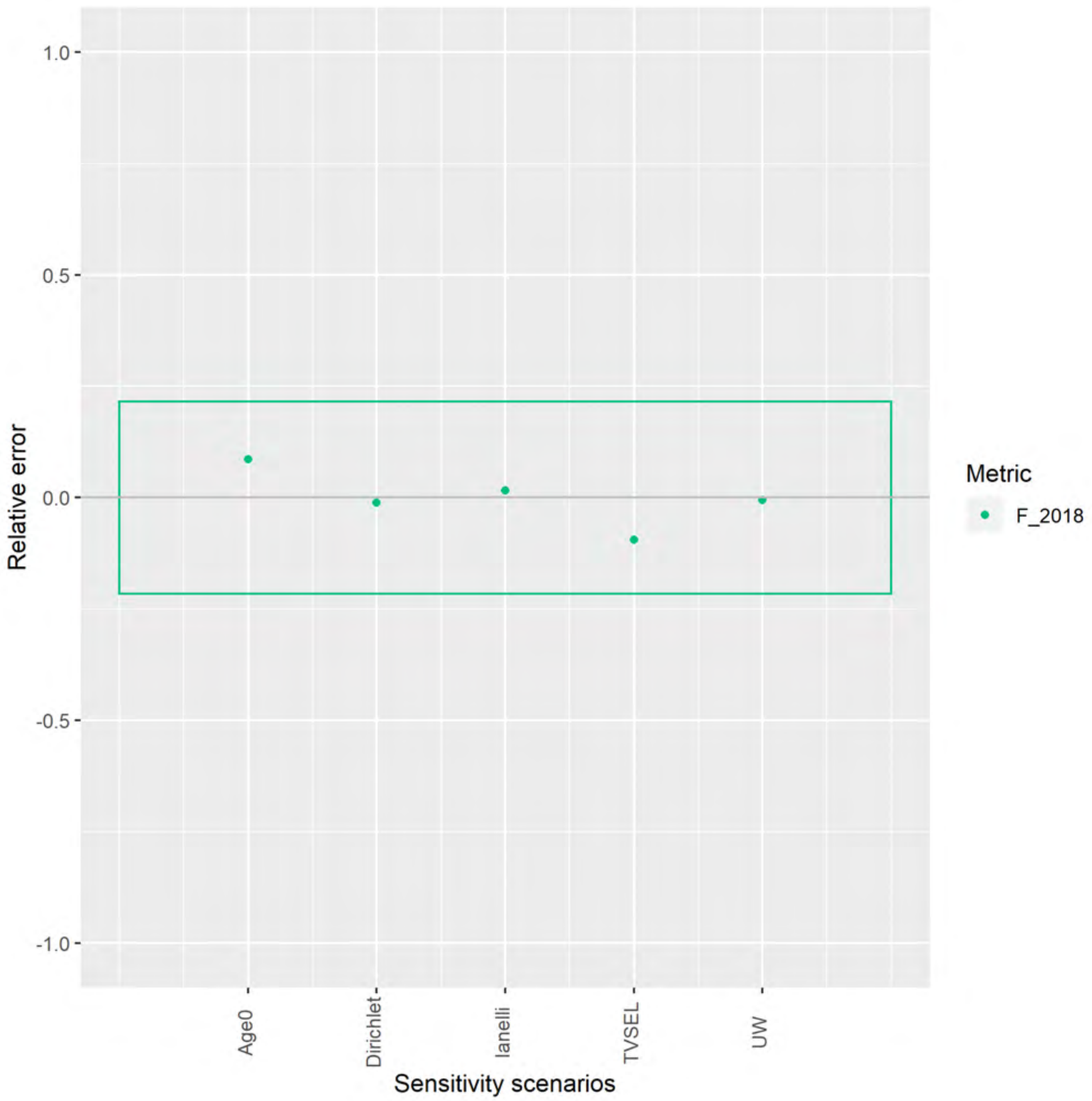


Figure 2. Herring SDs 30-31. Sensitivity to alternative model configurations as described in the text. Boxes correspond to the 95% confidence interval of a derived quantity (indicated by color) in the reference model. Values outside the box would indicate significant difference from the uncertainty provided in the reference model. The metric used were: SB_0 (a), SSB (b) in 2018, B_{MSY} (c), F_{MSY} (d) and F in 2018 (F).

Final model run and diagnostics

Overview of the datasets included in the final Stock Synthesis model is shown in Figure 3. The diagnostic figures included in the following chapters are related to the reference model developed before the benchmark. Those were used to compare between model configurations and are considered valid for this purpose. The final model as agreed at the benchmark with its retrospective diagnostic is presented in the next sections.

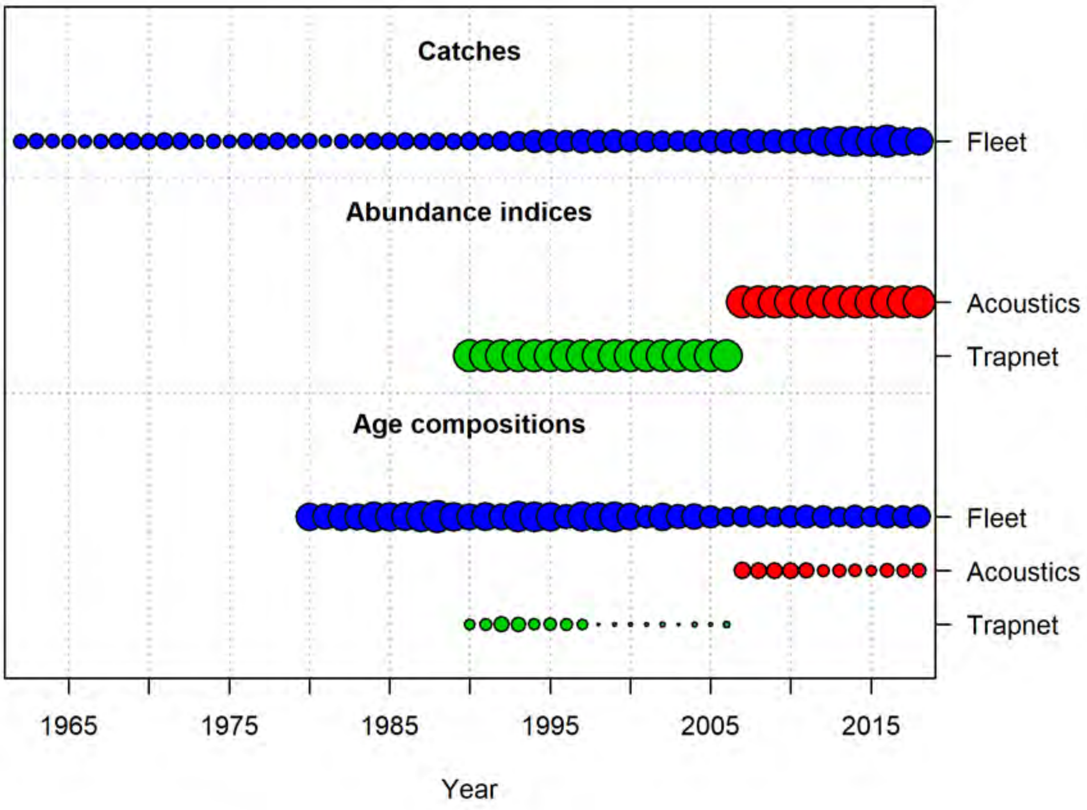


Figure 3. Herring SDs 30-31. Summary of the input time series included in the model.

The selectivity of all fleets is well estimated (Figure 4).

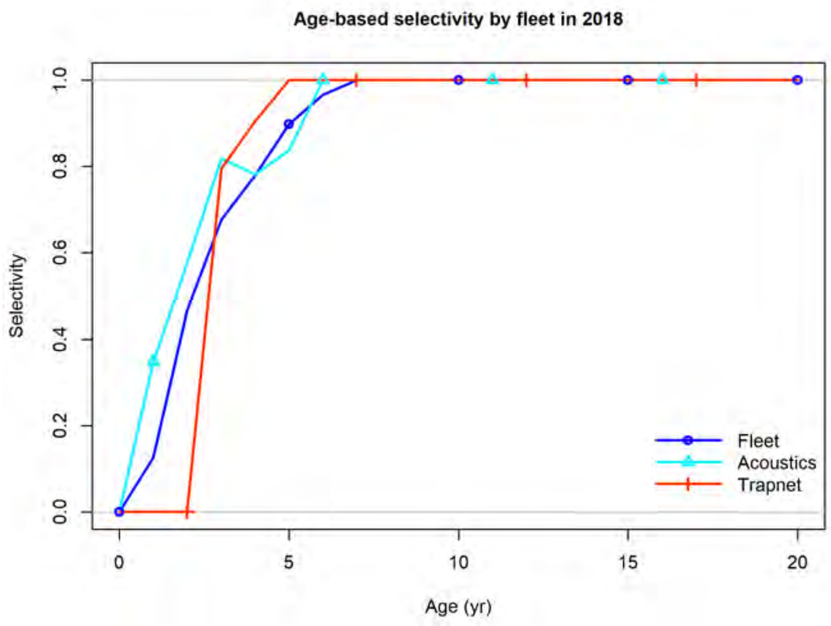


Figure 4. Herring SDs 30-31. Age based selectivity by fleet.

The fitting of the model was good, with the age compositions well reconstructed. The residuals are quite low, never below -2.6 and above 2.6, and without particular patterns (Figure 5 and 6).

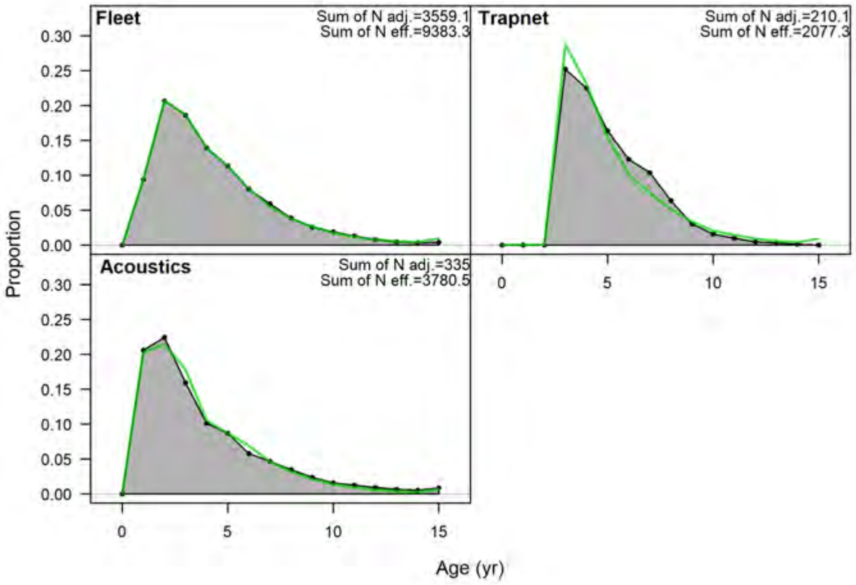


Figure 5. Herring SDs 30-31. Model fits to age composition data.

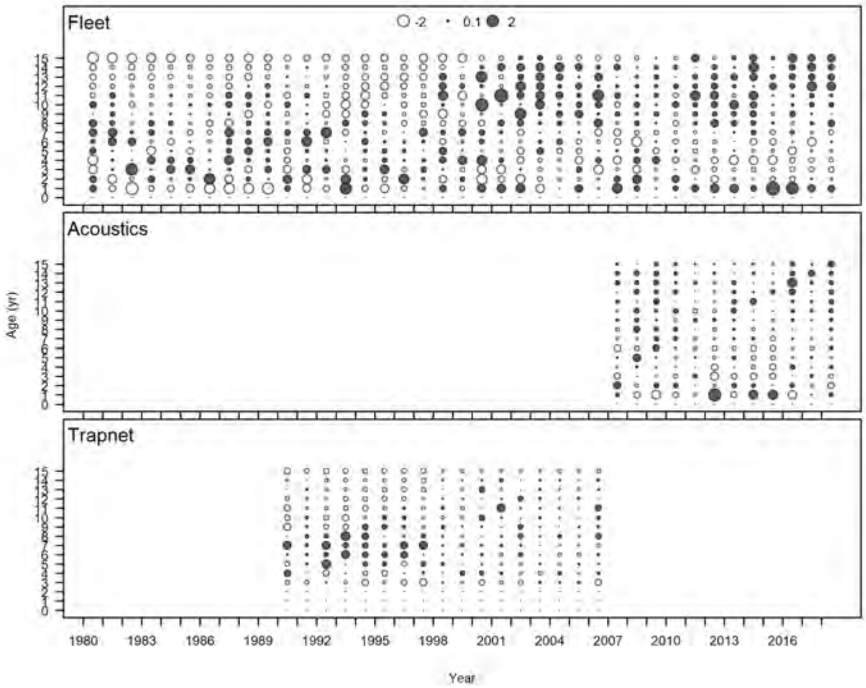


Figure 6. Herring SDs 30-31. Residuals of fits to age composition data for the different fleets.

Overall, the model doesn’t provide a very good fit to the trends in both the acoustic and trapnet surveys (Figures 7 and 8).

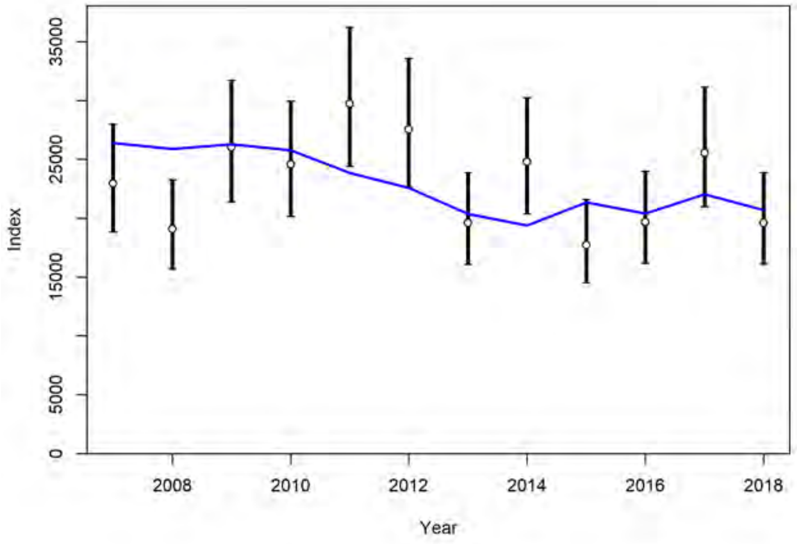


Figure 7. Herring SDs 30-31. Model fits to the acoustic survey density index.

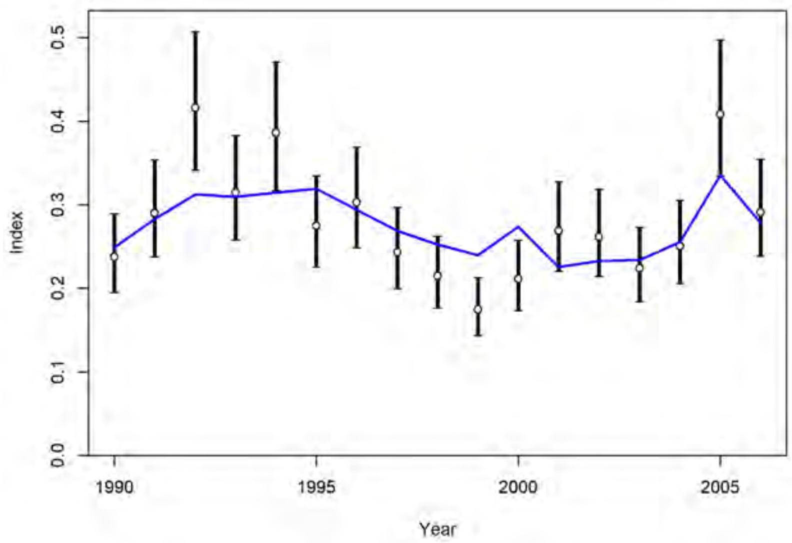


Figure 8. Herring SDs 30-31. Model fits to the trapnet survey biomass index.

A non-random pattern of residuals may indicate that some heteroscedasticity is present, or there is some leftover serial correlation (serial correlation in sampling/observation error or model misspecification). Several well-known nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test (Gibbons and Chakraborti, 1992). Here we used the runs test (JABBA and ordinary runs test) to evaluate whether residuals of the surveys and of the age frequency distributions were normally distributed or/and had time trends because this test has been used recently to diagnose fits to indices and other data components in other assessment models (e.g. SEDAR 40, 2015; Winker et al., 2018). The results of the runs test are presented in Figures 9 and 10. The JABBA runs test indicated that the fit of the CPUE index was good because no residuals were larger than 1 and the RMSE was less than 30%, indicating a random pattern of the surveys residuals and the age frequency distributions. The JABBA plot is considered as a tool for identifying trends in residuals and if the standard deviation is tight on a given year this means the fleets are in agreement, even if not fitting well, which is a useful diagnostic. Its purpose is to visualize multiple residuals at once, pick up on periods of substantial data conflicts (width of boxes) and systematic departures in median residuals (loess). In this case, as we have two surveys not overlapping in time, the RMSE is the only useful metric. The ordinary runs test was passed for both acoustic and trapnet surveys residuals and also for all length frequency distributions.

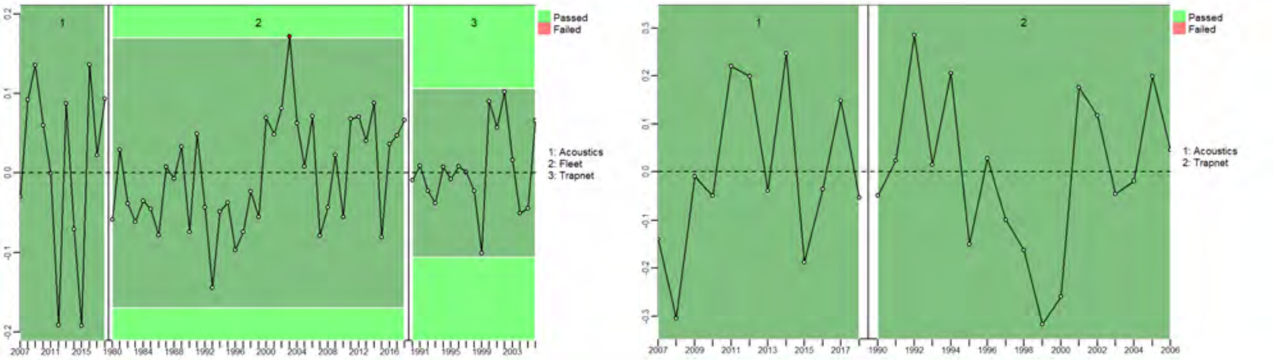


Figure 9. Herring SDs 30-31. Residuals from runs test analyses for the age distributions and the fit to the acoustic and trapnet survey indices.

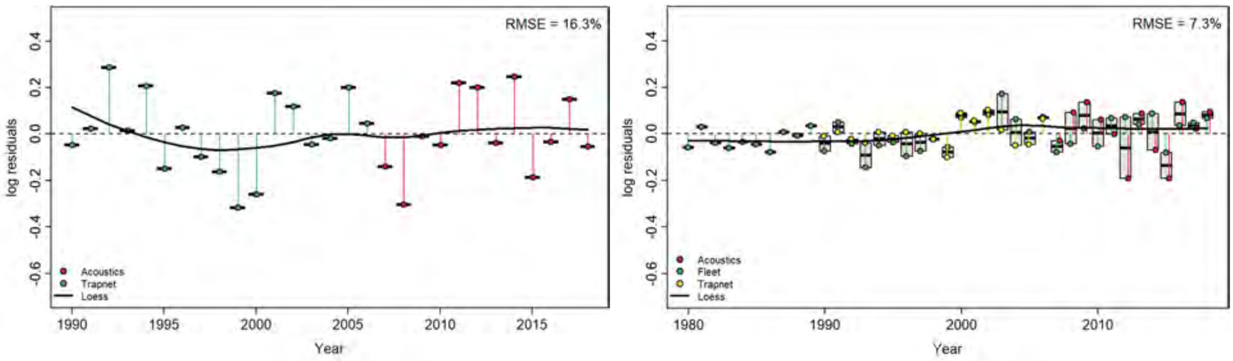


Figure 10. Herring SDs 30-31. Residuals from the JABBA runs test analyses for the age distributions and the fit to the acoustic and trapnet survey indices.

Jittering

The jitter procedure allows to verify the stability of the model examining the effect of varying the starting values of the model input estimated parameters on the model results. An accurate model should converge on a global solution (i.e. not being stuck in local minima of likelihood surface) across a reasonable range of starting values input parameters. In this case, 100 runs were performed considering a 10% of jitter of the initial parameters, which means that a small random jitter is added to the initial parameter values. Starting values are jittered based on a normal distribution based on the pr(PMIN) = 0.1% and the pr(PMAX) = 99.9%.

The 100 iterations of the jitter test for global convergence resulted in the same results as the reference run (Figure 11), so no local minima are observed as no runs have a likelihood lower than the reference run. It is however important to stress that the absence of a local minima when running jittering is not a guarantee that the model is not indeed stuck in a local minimum, although its absence reduced the risks that this occurs (Subbey 2018).

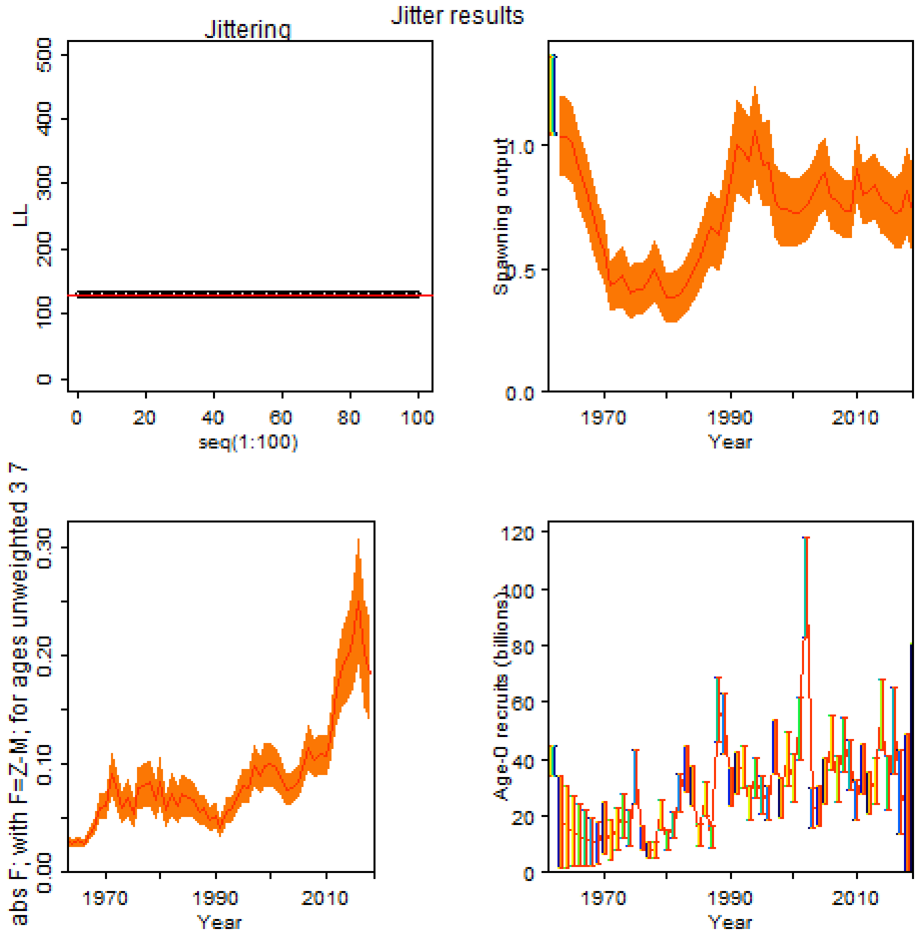


Figure 11. Herring SDs 30-31. Results from jitter using 100 iterations and an average jitter of 10%.

Retrospective analyses

Retrospective analysis is a diagnostic approach to evaluate the reliability of parameter and reference point estimates and to reveal systematic bias in the model estimation. It involves fitting a stock assessment model to the full dataset. The same model is then fitted to truncated datasets where the data for the most recent years are sequentially removed. The retrospective analysis was conducted to the reference model for the last 5 years of the assessment time horizon to evaluate whether there were any strong changes in model results. Given that the variability of Mohn's rho index depends on life history, and that the statistic appears insensitive to F, Hurtado-Ferro et al. (2014) proposed the following rule of thumb when determining whether a retrospective pattern should be addressed explicitly. Values of Mohn's rho index higher than 0.20 or lower than -0.15 for long-lived species (upper and lower bounds of the 90% simulation intervals for the flatfish base case), or higher than 0.30 or lower than -0.22 for short-lived species (upper and lower bounds of the 90% simulation intervals for the sardine base case) should be cause for concern and taken as indicators of retrospective patterns. However, Mohn's rho index values smaller than those proposed should not be taken as confirmation that a given assessment does not present a retrospective pattern, and the choice of 90% means that a "false positive" will arise 10% of the time. In both cases, model misspecification would be correctly detected more than half the time. The retrospectives of the reference model were rather stable (Figure 12). The estimated Hurtado-Ferro et al. (2014) variant of the Mohn´s rho indices were inside the bounds of recommended values for both SSB (-0.07) and F (0.13), but outside the bounds for recruitment (-0.56).

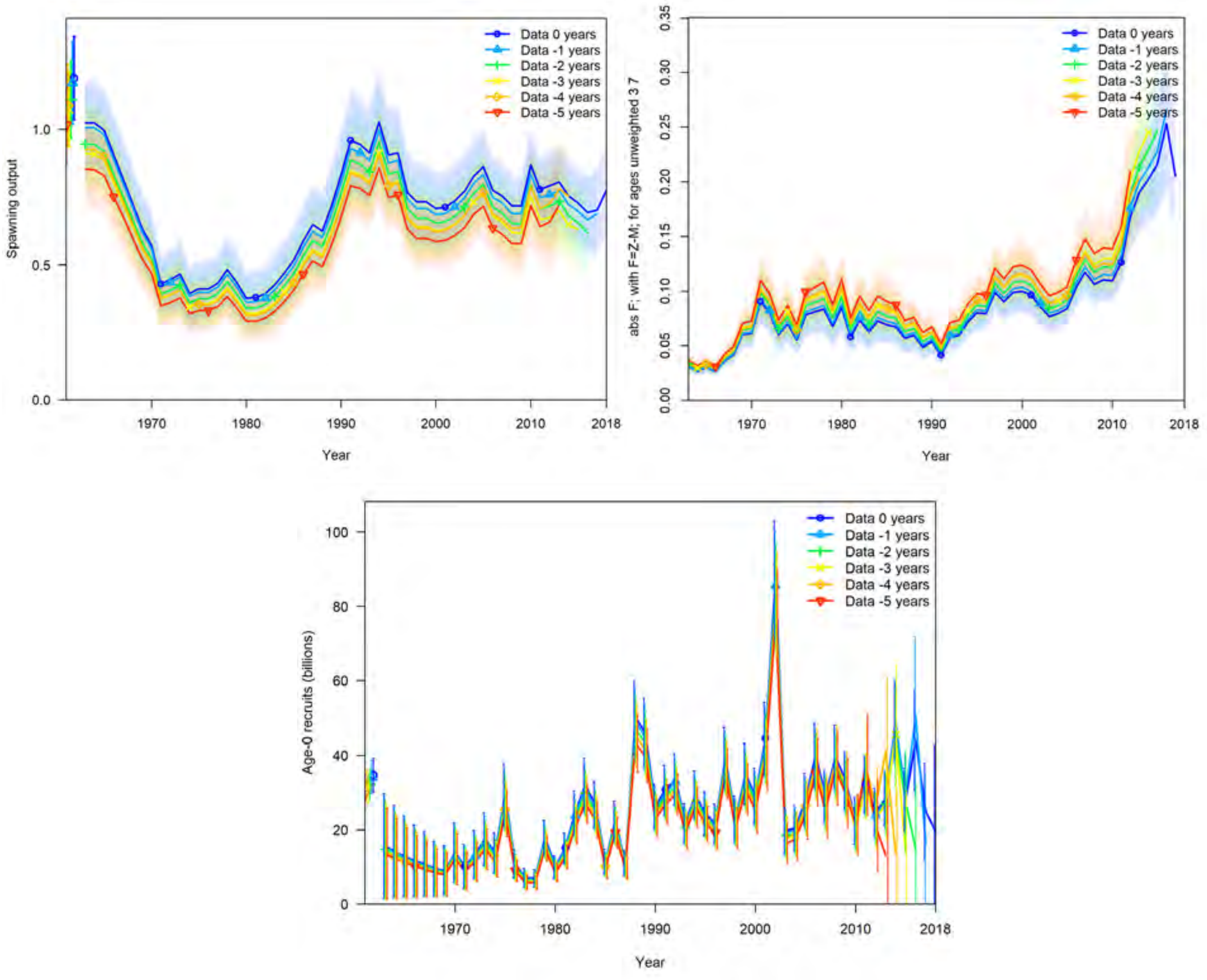


Figure 12. Herring SDs 30-31. Retrospective analyses of the reference model.

There is little or no information in the data to estimate the sizes of the 2018 and 2017 year-class. Retrospective analyses of year class strength for young fish shown the estimates of recent recruitment to be unreliable prior to at least age 2 (Figure 13), which likely explain the retrospective pattern in recruitment.

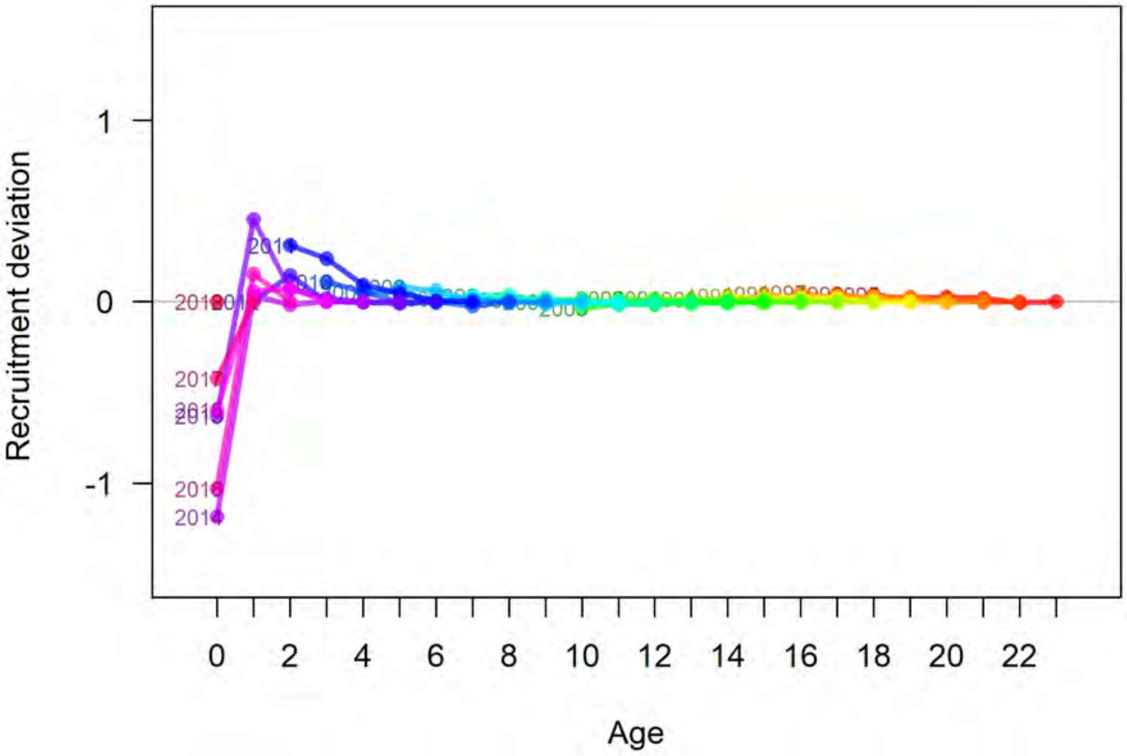


Figure 13. Herring SDs 30-31. Retrospective recruitment estimates scaled relative to the most recent estimate of the strength of each cohort.

Likelihood profiles

Likelihood profiling is an automated routine in Stock Synthesis, which allows to evaluate model performance across a range of values of an input parameter (generally R_0, σ_R and steepness). Here we performed the likelihood profile of R_0 and σ_R for the reference model. The likelihood profile of R_0 shows a minimum at the model estimate minimum, which corresponds to 17.5. There is an apparent conflict between the index and age versus the recruitment likelihood components in the estimate of R_0 (Figure 14).

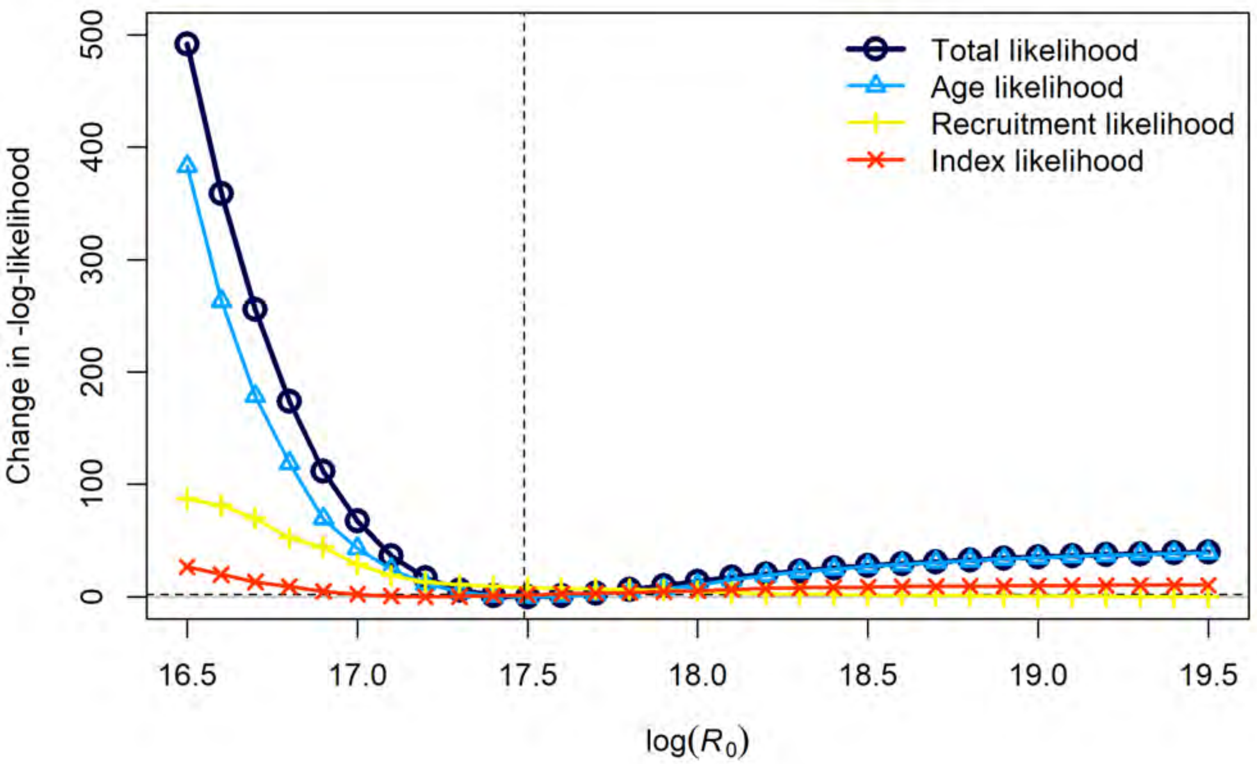


Figure 14. Herring SDs 30-31. Likelihood profile of R_0. The vertical dashed line represent the value estimated in the reference model. The horizontal dashed line represents the threshold of the chi-square test for 95% significance value.

For sigmaR (σ_R), the values ranging from 0.5 to 1.2 did not affect significantly (i.e. they are under the dashed line, which represent the threshold of the chi-square test) the log-likelihood of the surveys (Figure 15). Differently, the values of σ_R tested affect significantly the log likelihood of Recruitment and Age composition, which represent an apparent conflict between the different information. However, those changes in log likelihood resulted low considering that the max change obtained values of around 40 (Figure 15). For the total likelihood, values between 0.45 and 0.7 are not significantly different. Thus, a σ_R of 0.6 was used in all models.

σ_R is the stochastic recruitment process error and the estimation of this parameter within integrated models is generally recognised to be problematic (Kolody et al., 2019) so that σ_R individual recruitment estimates is fixed at a values that is large enough to prevent the SSR from constraining individual recruitment estimates (e.g. analogous to traditional VPA) (Kolody et al., 2019). A meta-analysis of the estimation of σ_R done outside the operative model (ISSF, 2011) yielded a median estimate between 0.2 and 0.5, which suggested that σ_R is often inflated in assessment models. However, models with σ_R down to 0.3 do not change substantially the results, which is reassuring that the management advice is not affected by the choice of the σ_R value (Figure 16).

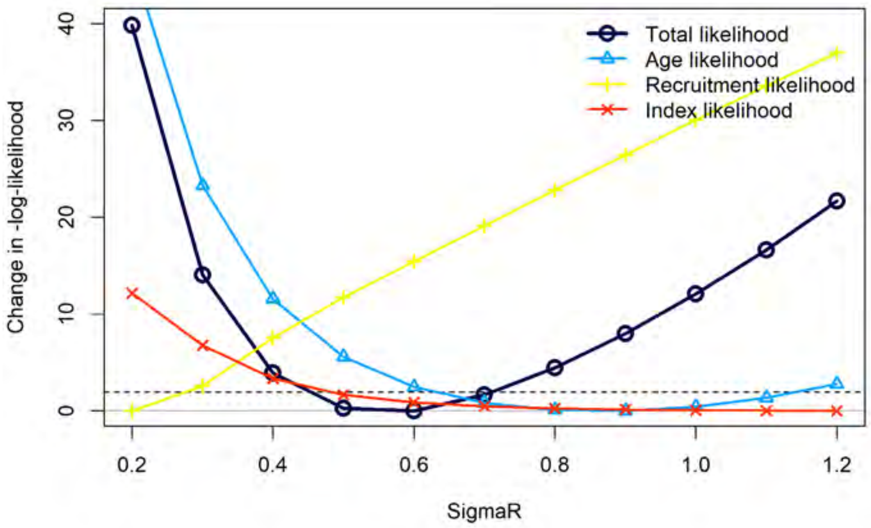


Figure 15. Herring SDs 30-31. Likelihood profile for σR. The horizontal dashed line represents the threshold of the chi-square test for 95% significance value.

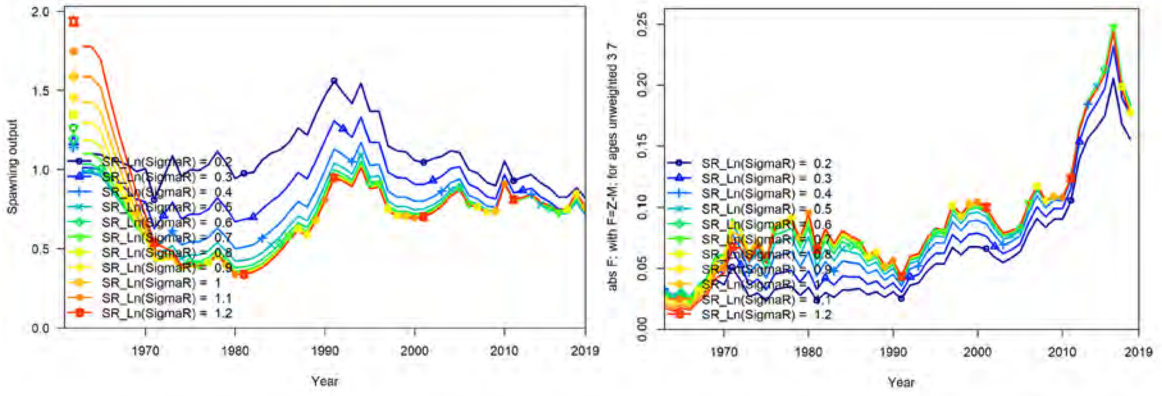


Figure 16. Herring SDs 30-31. Trajectories of SSB (left panel) and F (rigth panel) at different values of σR.

Hindcasting

The provision of fisheries management advice requires the assessment of stock status relative to reference points, the prediction of the response of a stock to management, and checking that predictions are consistent with reality. A major uncertainty in stock assessment models is the difference between model estimates and reality. To evaluate uncertainty often a number of scenarios are considered corresponding to alternative model structures and dataset choices (Hilborn, 2016). It is difficult, however, to empirically validate model prediction, as fish stocks can rarely be observed and counted. Various criteria are available for estimating prediction skill (see Hyndman and Koehler, 2006). One commonly used measure is root-mean-square error (RMSE). RMSE, however, is an inappropriate and misinterpreted measure of average error (Willmott and Matsuura, 2005). On the other hand, mean absolute error (MAE) is a more natural measure of average error, and unlike RMSE is unambiguous. Scaling the average errors using the Mean Absolute Scaled Error (MASE) allows forecast accuracy to be compared across series on different scales. MASE values greater than one indicates that in-sample one-step forecasts from the naïve method perform better than the forecast values under consideration. MASE also penalizes positive and negative errors and errors in large forecasts and small forecasts equally.

Kell et al. (2016) showed how hindcasting can be used to evaluate model prediction skill of the CPUE. When conducting hindcasting, a model is fitted to the first part of a time series and then projected over the period omitted in the original fit. Prediction skill can then be evaluated by comparing the predictions from the projection with the observations using for example the MASE indicator (Hyndman and Athanasopoulos, 2013).

Hindcasting was conducted for the reference model (Fig. 17). The results showed that the acoustic survey performs well in hindcasting given that the MASE value is lower than the 1.0 threshold when predicting the index one year ahead.

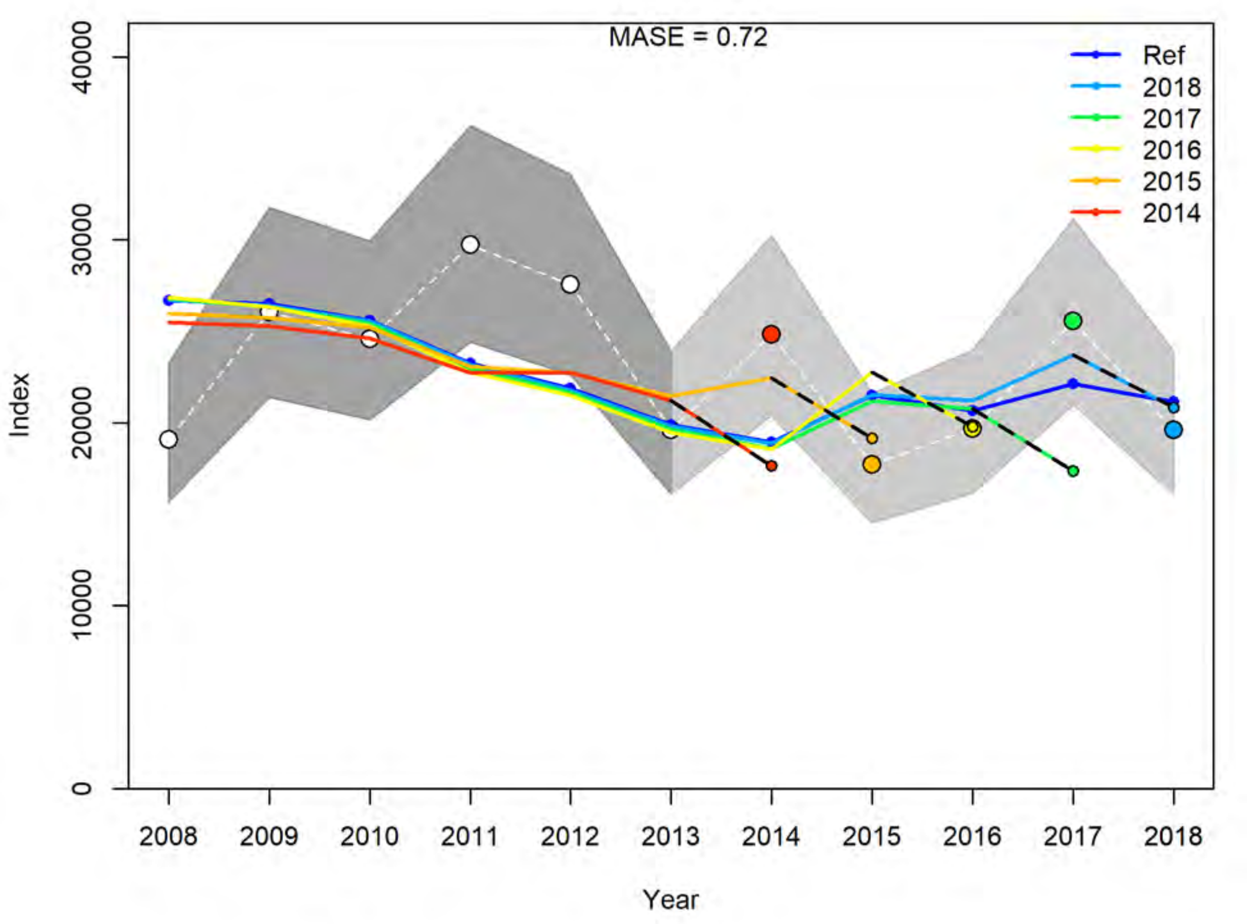


Figure 17. Herring SDs 30-31. Results of hindcasting for the acoustic survey. Black dashed lines are the forecasts while colour coded observations are the corresponding observations that were dropped when making the prediction residual for that specific year. .

MCMC

Markov chain Monte Carlo (MCMC) methods comprise a class of algorithms for sampling from a probability distribution. It is used in integrated models for detecting misspecification in key fixed parameters or issues with estimation of the parameters. By constructing a Markov chain it is possible to obtain a sample of the desired distribution by observing the chain after a number of steps. The more steps there are, the more closely the distribution of the sample matches the actual desired distribution. MCMC methods create samples from a possibly multi-dimensional continuous random variable, with probability density proportional to a known function. These samples can be used to evaluate an integral over that variable, as its expected value or variance. Practically, an ensemble of chains is generally developed, starting from a set of points arbitrarily chosen and sufficiently distant from each other. Those are then used to estimate the posterior distribution of the parameters of interest within the model.

For herring in SDs 30-31, we performed an MCMC run using the Random Walk Metropolis method with 1100000 iterations, with no burn-in period and thinning each 1000 iterations. The results showed that the MCMC is almost identical to the MLE estimated, which is an indication of the robustness of the model (Figure 18).

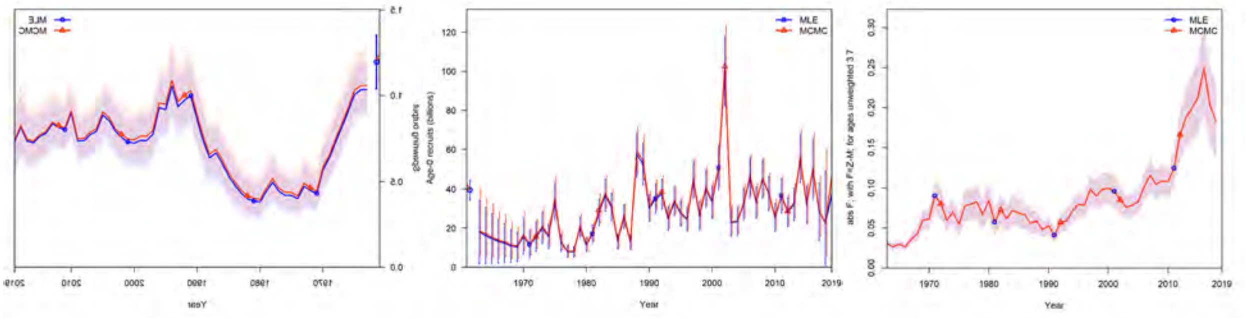


Figure 18. Herring SDs 30-31. Results of the MCMC analysis in terms of SSB, R and F compared to the MLE model.

First, we used also NUTS algorithm in MCMC (Monnahan et al., 2019) to regularize the model, i.e. to check that all parameters are identifiable. MCMC with 10000 iterations, 9 chains, run with NUTS algorithm, with 25000 iterations as burn in and thinning every 100 confirmed that all parameters of the model are identifiable.

Successively, we analysed the plot of the five slowest mixing parameters in the MCMC run with NUTS algorithm, 100000 iterations, 9 chains, with 25000 iterations as burn in and thinning every 100. Almost all estimates are within the 95% confidence interval and the central tendency of the five slowest mixing parameters shows that they are centred around the median (Figure 19), which indicate that the model is not ill-configured. Stationarity of the posterior distribution for model parameters was re-assessed via a suite of standard single-chain and multi-chain diagnostic tests. Diagnostic of the MCMC does not reveal any issue with the key parameters (i.e. SSB in 2018, F in 2018, R in 2018 and R_0) (Figure 20). The objective function, as well as all estimated parameters and derived quantities, showed good mixing during the chain, no evidence for lack of convergence, and low autocorrelation (Figure 20). Finally, analysis with Shinystan library (Monnahan et al., 2019) showed divergence only in less than 1% of the iterations, which confirms that the model is not ill-configured.

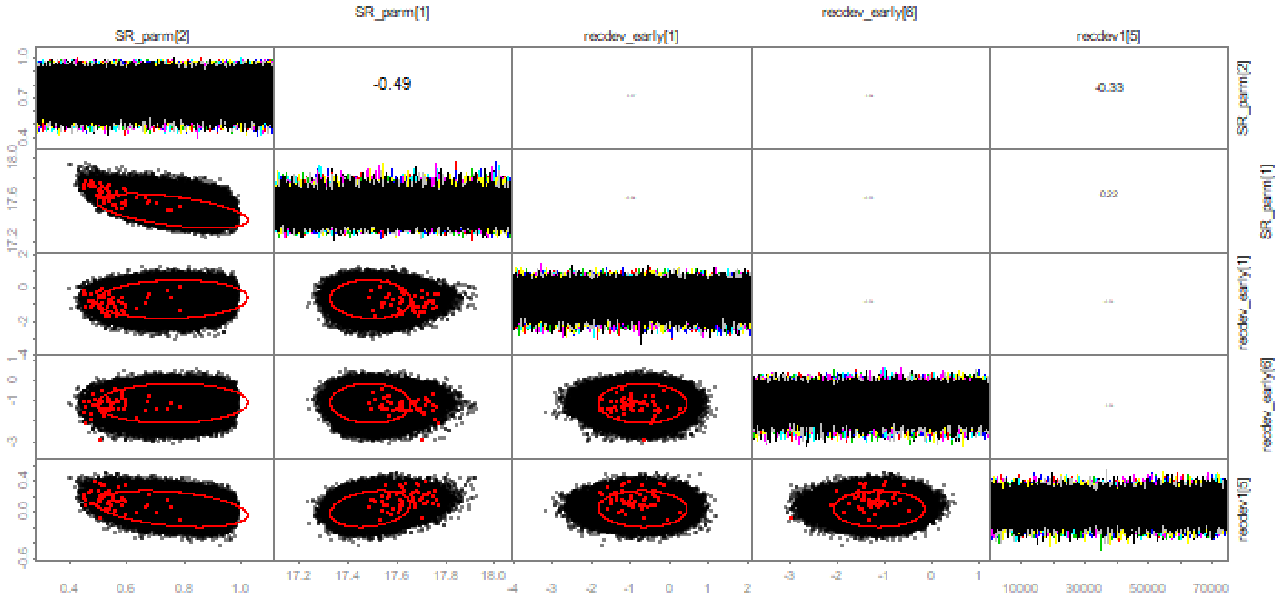
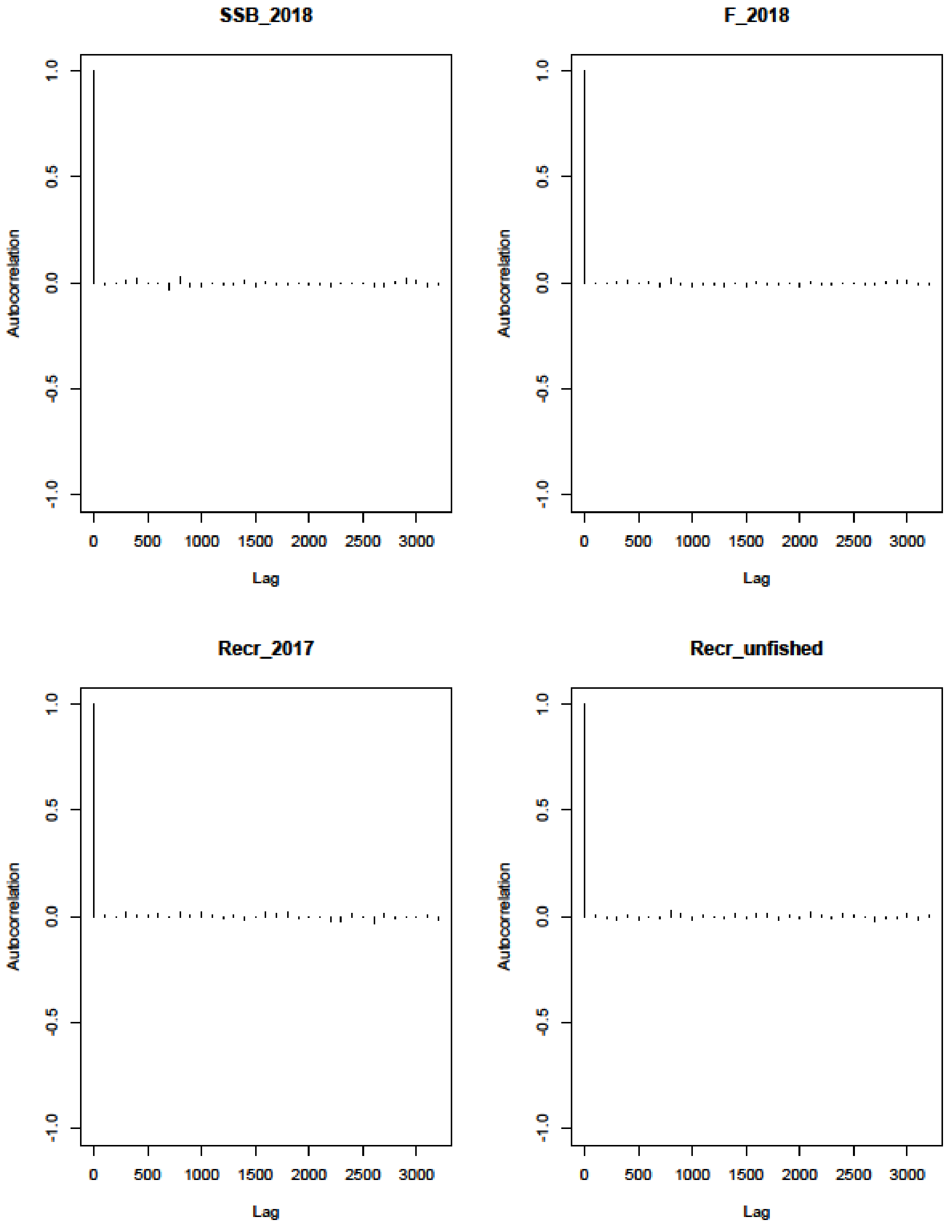


Figure 19. Herring SDs 30-31. Comparison between MLE (red points) against posteriors of the reference model obtained by an MCMC with 10000 iterations, 9 chains, run with NUTS algorithm, with 25000 iterations as burn in and thinning every 100 for the 5 slowest mixing parameters. Red ellipse is 95% confidence interval, points are posteriors draw and lines shows chain traces.


SSB_2018
F_2018
Recr_2017
Recr_unfished
Autocorrelation
1.0
0.5
0.0
-0.5
-1.0
0
500
1000
1500
2000
2500
3000
Lag

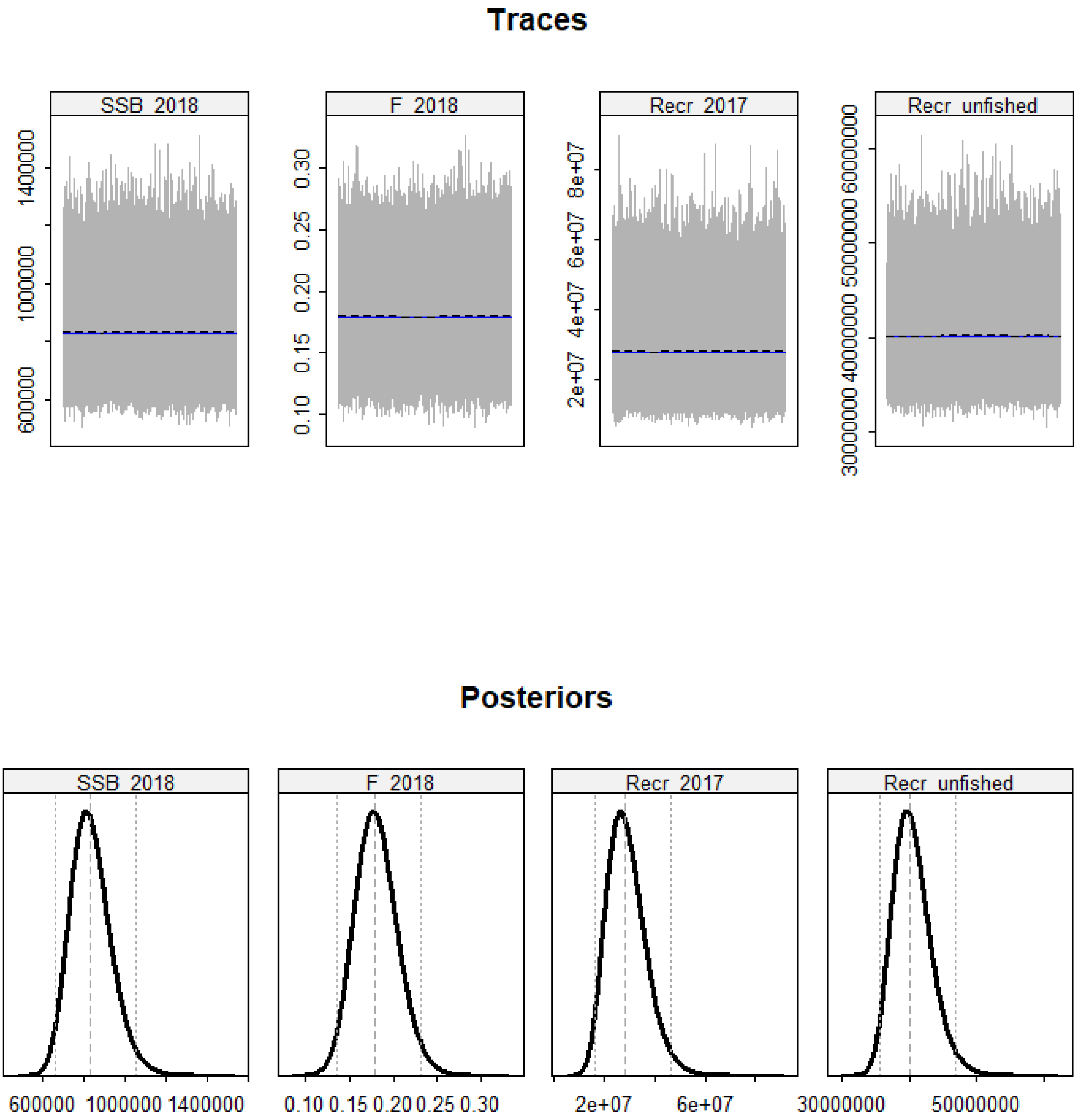


Figure 20. Herring SDs 30-31. Summary of MCMC diagnostics (MCMC with 10000 iterations, 9 chains, run with NUTS algorithm, with 25000 iterations as burn in and thinning every 100) for SSB in 2018, F in 2018, R in 2017 and R_0 in the reference model. The first sub-panels show the autocorrelation present in the chain at different lag times (i.e., distance between samples in the chain), the second sub-panels show the trace of the sampled values across iterations (absolute values, top left; cumulative running mean with 5th and 95th percentiles, top right), and the third sub-panel shows the distribution of the values in the chain (i.e., the marginal density from a smoothed histogram of values in the trace plot).

ASPM

In some integrated stock assessments, the index of abundance provides almost no information on population scale. Consequently, the estimates of the model outputs rely almost completely on the size- and age-composition data and model structure. Maunder and Piner (2015) proposed a diagnostic tool that can be used to evaluate the information content of data about absolute abundance and assess whether the model is correctly specified. This diagnostic consists of comparing the results of an age-structured production model (ASPM) to those from a model estimating all of the model parameters and fitting to all the data (e.g., an integrated analysis). It is inferred that a production function is apparent in the data when the catch data explain indices with good contrast (e.g., declining and increasing trends), therefore providing evidence that the index is a reasonable proxy of stock trend. If the ASPM cannot mimic the index, then either the stock is recruitment-driven, catch levels have not been high enough to have a detectable impact on the population, the model is incorrect, or the index of relative abundance is uncertain or not proportional to abundance. Thus, ASPM is able to evaluate if variations in predicted population dynamics are mainly informed by the relative abundance indices and catches, and governed by the underlying surplus production function and process error or instead is driven by changes in recruitment or other biological characteristics of the stock.

To perform the ASPM diagnostic test, we had to change the original model parameterization. SS can behave like an ASPM (Methot and Wetzel, 2013) when the parameters of the selectivity curve are fixed at those estimated from the fully integrated model, the annual recruitment deviates are not estimated (fixed at zero so that recruitment follows the stock-recruitment relationship), and the age- and size-composition data are not used for parameter estimation. The results from the ASPM should be similar to those from the fully integrated model if the size- and age-composition data are not informing absolute abundance or the trend in abundance and there is no strong pattern in recruitment. The ASPM test (Maunder and Piner, 2015) appears to have promise in detecting systems dynamic misspecification (*h* and M), where the runs test showed lower power, and ASPM showed good power. For herring in SDs 30-31, a production relationship is not evident in the assessment model, with ASPM results leading to very dissimilar estimates of SSB and SSB$_{2018}$ compared to the correspondent fully integrated model (Figure 21). In this case, a clear pattern in recruitment deviations as estimated by the model is evident (Figure 22), which might be related to changes in salinity, ice cover and other key parts of the ecosystem in the area (Pekcan-Hekim et al., 2016). Thus, the stock seems to be mainly environmentally driven in the last two decades, with the appearance of larger than expected year classes, which has determined an increase of the SSB despite increasing fishing mortality. In other words, the stock is recruitment-driven (*sensu* Carvalho et al., 2017).

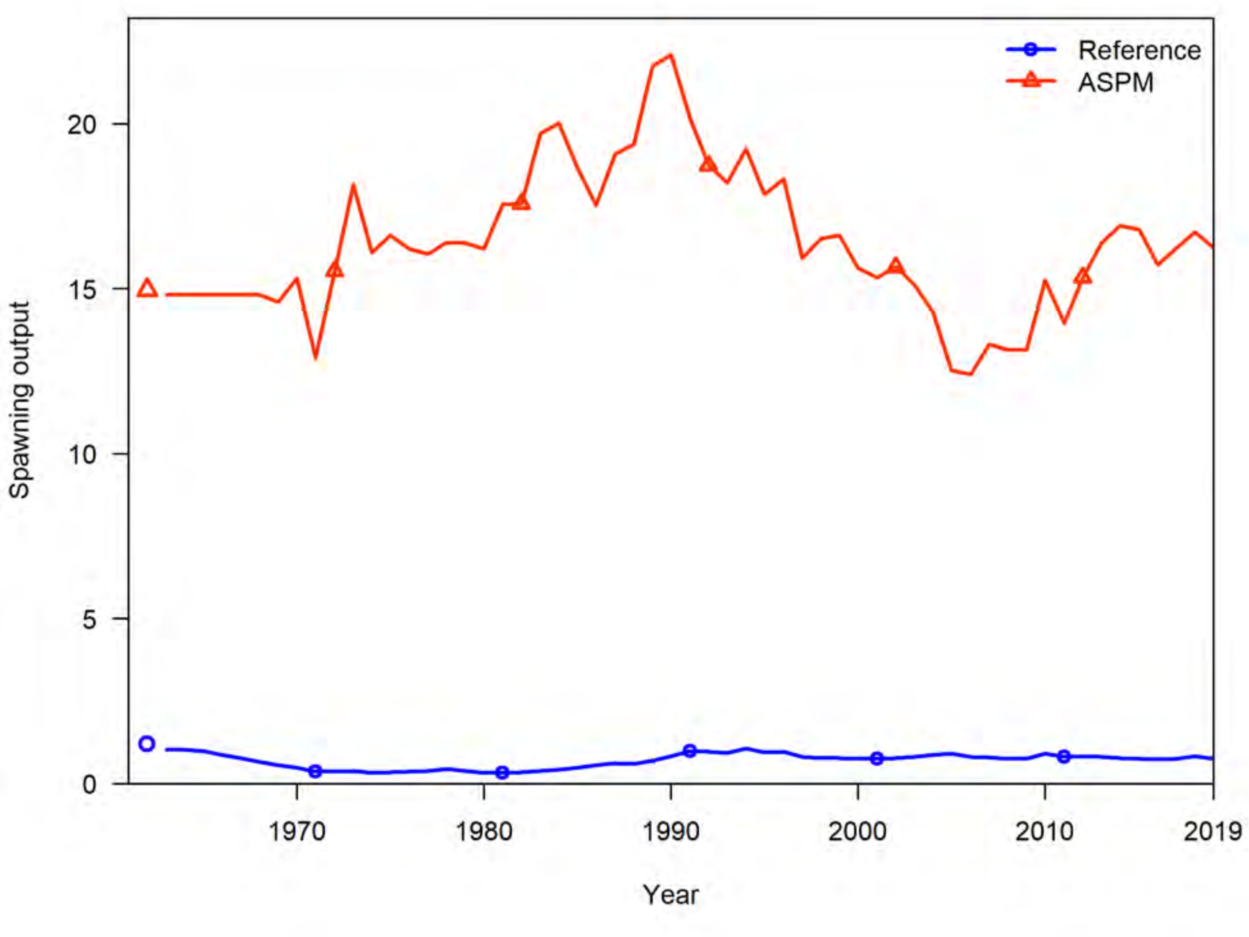


Figure 21. Herring SDs 30-31. Results of the ASMP analysis in terms of SSB compared to the MLE model.

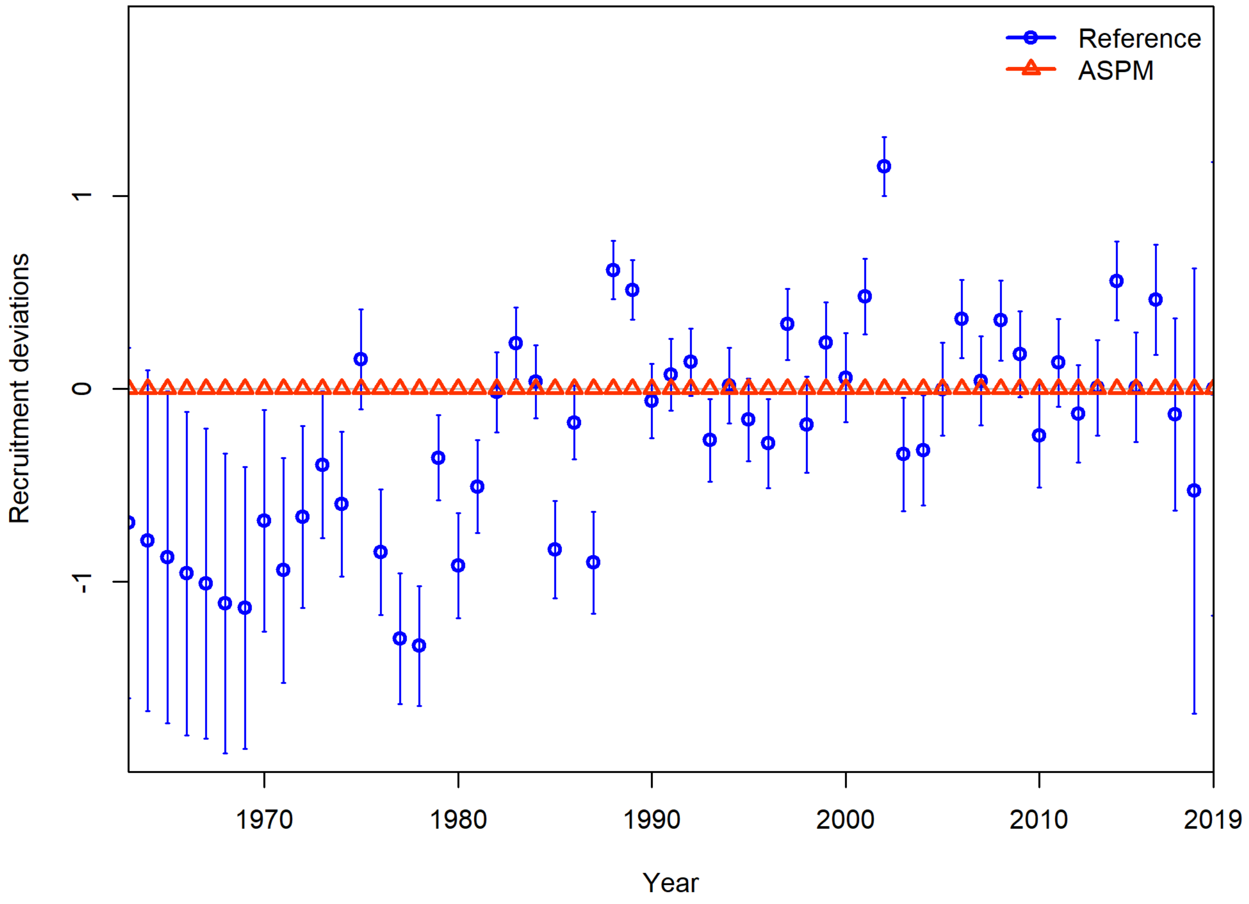


Figure 22. Herring SDs 30-31. Results of the ASMP analysis in terms of recruitment deviations compared to the MLE model.

Analysis of surplus production trend

Estimates of Surplus Production (Walters, et al., 2008) can provide a check of whether predictions of changes in biomass can be made reliably based on catch and current biomass (clockwise or linear behaviour) or whether there has been non-stationarity in production processes, i.e. are dynamics driven by climate and oceanic conditions (counter clockwise). This is important for example for the development of MPs in the MSE process. In the case of herring in SDs 30-31, the figure shows a mixed pattern of surplus production against total biomass predicted by age-structured models with a clockwise or linear behaviour in the beginning of the time series followed by a counter clockwise in the more recent years. This is likely related to the increased recruitment in the latest decades associated with a warming environment (Figure 23).

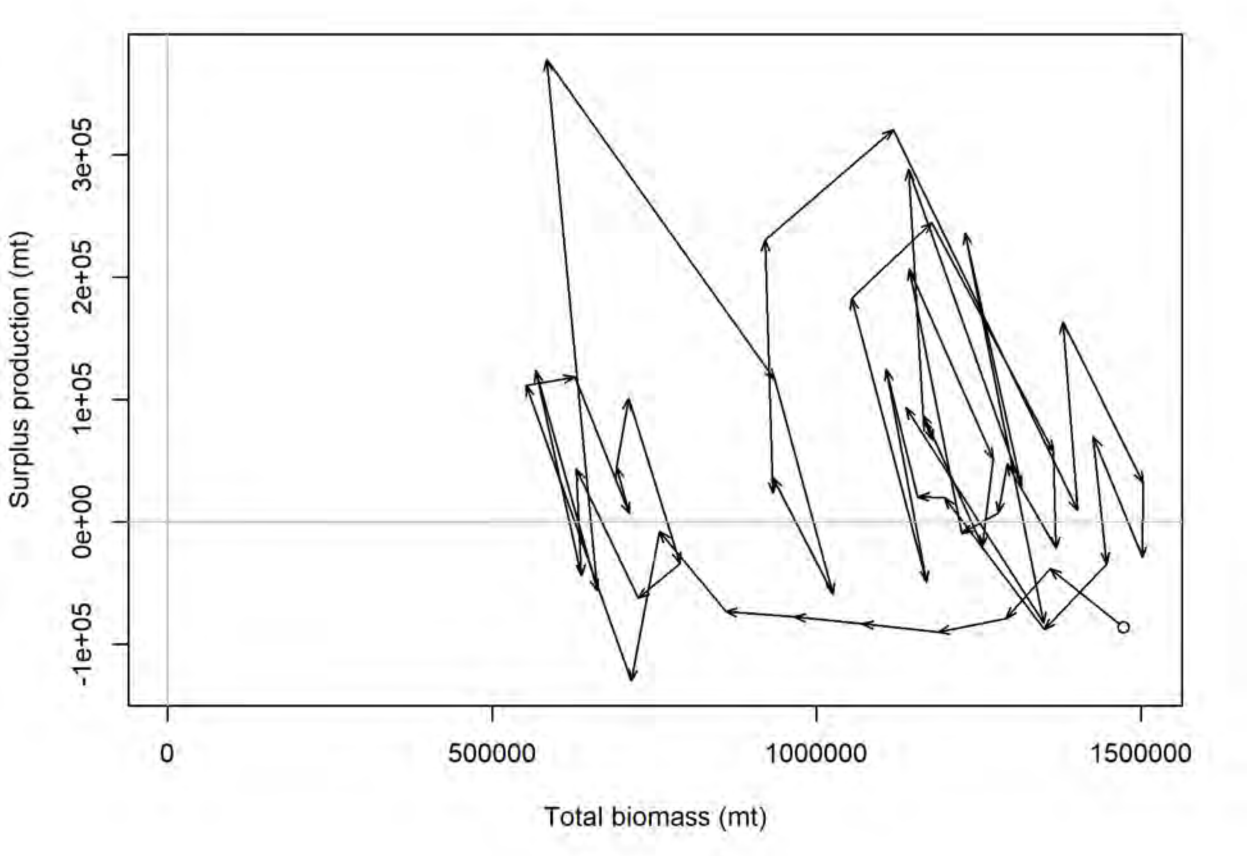


Figure 23. Herring SDs 30-31. Surplus production against biomass plot. The round circle represents the first year of the time series (1963).

When all diagnostic tests are considered together, the power to detect model misspecification improves without a substantial increase in the probability of incorrectly rejecting a correctly specified model (Carvalho et al., 2017) and therefore these diagnostics should be all applied routinely. When the criteria for rejecting a model as correctly specified is a failure of at least one of the diagnostic tests, nearly 90% of most mis-specified are detected with no real increase in the probability of a false detection (Carvalho et al., 2017). Residual analyses were easily the best detector of misspecification of the observation model, while the ASPM is the only good diagnostic for misspecification of system dynamics model (Carvalho et al., 2017). The retrospective analysis and R_0 likelihood component profile had low rates of detection of mis-specified models (Carvalho et al., 2017), although retrospective analysis is effective in detecting un-modeled temporal variation (Hurtado-Ferro et al., 2014). Finally, opposed to the widely used maximum-likelihood estimator, MCMC gives clear warning signs when a non-identifiable model is used for fitting (Siekmann et al., 2012). In this context, we created a table that summarize all diagnostics for the three main candidate models (Table

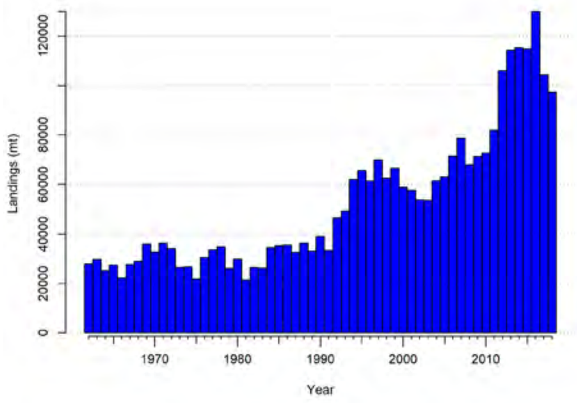
7). The table is an attempt to sum up a multidimensional space and thus it needs to be seen as a guidance more than as a definitive result. However, it is evident from Table 7 that the reference model is the “best” model according to all scoring criteria. Also, it passes or it perform best in two of the key tests (MCMC and jitter test) and in the retrospective. Thus, the reference model was proposed as the final model to be used for advice.

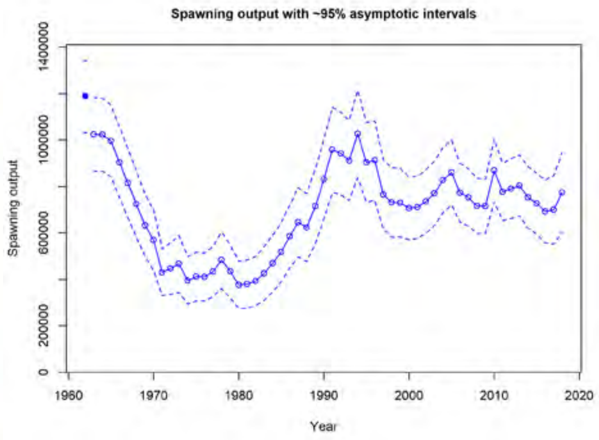
Table 7. Herring SDs 30-31. Summary table of the diagnostics of the three main candidate models. Reference= Age 0 excluded; Age0= Age 0 individuals from BIAS survey included; TVSEL = Time varying selectivity of the fleet. “Passed tests” score refers to the average test passes in % when multiple tests have been conducted. “Best model” score is the sum of the number of times each model is the best model in each of the tests and “weighted ranking” is as the “Best model score” but weighted for the importance of each test to detect model misspecification as described (but not quantified) by Carvalho et al., 2017.

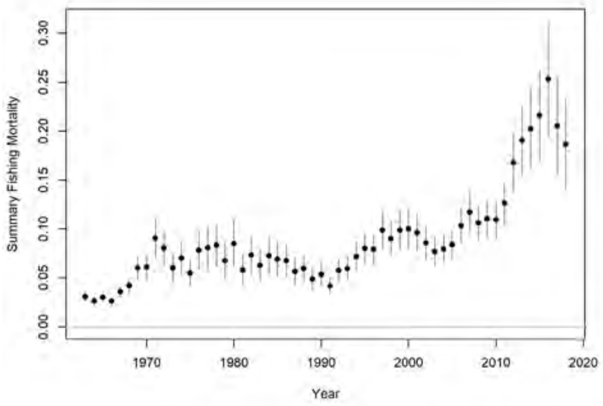
			Model		
Diagnostic	**Indicator**	**Component**	**Reference**	**Age0**	**TVSEL**
Convergence		Model	5.33E-06	2.80E-05	1.13E-05
N. of parameters		Model	71	76	170
Hessian		Model	Yes	Yes	Yes
AIC		Model	404	437	545
Jittering (10%)	% of runs above reference LL	Local minima	100%	100%	100%
	N of runs not different from reference run	Model	100%	99%	93%
Retrospective (5 years)	Mohn´s rho	SSB	-0.06	0.09	-0.13
		F	0.11	0.17	0.34
Hindcasting	MASE	Survey	100%	100%	0%
MCMC	Confidence of intervals of SSB_{2018}	Model	Yes	Yes	Yes
ASPM	Confidence of intervals of SSB_{2018}	Model	No	No	No
Run´s test		Survey	100%	100%	100%
		Recruitment deviations	100%	100%	100%
		Age compositions	100%	100%	66%
		JABBA survey	100%	100%	100%
		JABBA Age compositions	100%	100%	100%
		Passed tests	100.0%	99.9%	82.4%
		Best model	15	8	7
		Weighted ranking	13	5.5	4

Trends in SSB, F and R of the reference model

The stock status and the trends in SSB, R and F are based on the MLE model. The spawning stock biomass (SSB) has been declining from the beginning of the time series up to the 1970s, then it increased during the 1980s reaching levels comparable to the 1960s. At the end of the 1990s the SSB decreased slightly and has remained stable at that level since then. Fishing mortality (F) has increased markedly since 1990s, with a peak in 2016. Recruitment (R) has been fluctuating throughout the time series. In 2002 a very strong year class appeared (Figure 24 and Table 8).


Landings (mt)
0
20000
40000
60000
80000
120000
1970
1980
1990
2000
2010
Year


Spawning output with ~95% asymptotic intervals
Spawning output
0
200000
600000
1000000
1400000
1960
1970
1980
1990
2000
2010
2020
Year


Summary Fishing Mortality
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1970
1980
1990
2000
2010
2020
Year

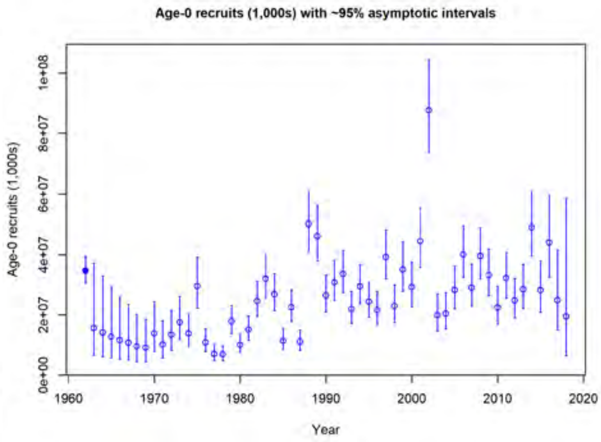

Age-0 recruits (1,000s) with ~95% asymptotic intervals
Age-0 recruits (1,000s)
0e+00
2e+07
4e+07
6e+07
8e+07
1e+08
1960
1970
1980
1990
2000
2010
2020
Year

Figure 24. Herring SDs 30-31. Summary of the stock assessment. SSB, F and R with 95% confidence intervals. Catches by fleet and SSB are in tonnes R is in thousands of individuals.

Table 8. Herring SDs 30-31. Summary of the stock assessment. Catches and SSB are in tonnes R is in thousands of individuals.

Year	SSB	F3-7	Recruitment	Catch
1963	1024620	0.03	15595600	29739
1964	1022950	0.03	14115300	25204
1965	996471	0.03	12777300	27541
1966	903440	0.03	11550700	22164
1967	814425	0.04	10753600	27772
1968	724404	0.04	9486100	28966
1969	632085	0.06	9048650	35996
1970	569045	0.06	13826600	32790
1971	429691	0.09	10109000	36347
1972	446072	0.08	13321000	34092
1973	466322	0.06	17503400	26507
1974	394351	0.07	13771800	26776
1975	411303	0.06	29493500	21811
1976	410046	0.08	10848900	30520
1977	433001	0.08	7002900	33634
1978	482474	0.08	6890340	34873
1979	433988	0.07	17871700	26109
1980	375610	0.08	9982980	29809
1981	378733	0.06	15059900	21526
1982	392367	0.07	24550600	26499
1983	425281	0.06	32042400	26208
1984	468843	0.07	26768000	34545
1985	517475	0.07	11409500	35432
1986	584781	0.07	22438800	35579
1987	645940	0.06	11020600	32628
1988	622741	0.06	50005700	36418
1989	715301	0.05	46083900	33086
1990	830688	0.05	26394600	39180
1991	957853	0.04	30729100	33419
1992	942510	0.06	33576900	46610
1993	911558	0.06	21813600	49314
1994	1026980	0.07	29401100	61986
1995	903463	0.08	24332100	65547
1996	912020	0.08	21558400	61303
1997	764747	0.10	39238500	69808
1998	730476	0.09	22851800	62474
1999	729923	0.10	35017000	66502
2000	706038	0.10	29116400	58852
2001	710749	0.10	44430400	57806
2002	734572	0.09	87681400	53969
2003	769418	0.08	19814600	53644
2004	826281	0.08	20416600	61423
2005	859549	0.08	28173300	62911
2006	772245	0.10	40070100	71318
2007	752860	0.12	29000000	78678
2008	715670	0.11	39495300	67914
2009	714931	0.11	33187200	71248
2010	868905	0.11	22333300	72590
2011	776183	0.13	32193800	81850
2012	790391	0.17	24736500	106007
2013	804053	0.19	28421200	114396
2014	752971	0.20	48992400	115366
2015	726187	0.22	28082600	114942
2016	691646	0.25	43928600	130029
2017	699175	0.21	24893100	104358
2018	774039	0.19	19477000	97366
2019	698669			

Model uncertainty

The reference assessment model run using maximum likelihood estimates (MLE) integrates over the substantial uncertainty associated with several important model parameters including: the magnitude of the stock (via the R_0 parameter for equilibrium recruitment), growth, annual selectivity for key ages and recruitment deviations. The uncertainty portrayed by the posterior distribution when the model is run using MCMC is a better representation of the uncertainty when compared to asymptotic approximations about the MLE because it allows for asymmetry (Figure A1). Note that we use the term MLE even though the priors are involved in the likelihood calculation and so the more accurate term would be the mode of the posterior density. Figure A1 shows the posterior distribution of the key parameters. It can be noted that the distribution of all parameters is generally very similar between the MLE and the MCMC model. Also, MLE and Bayesian (from MCMC) estimates of the trend in spawning biomass, F and recruitment are very similar, overlapping for the entire time series (see Figure 18).

Uncertainty measures in the reference model run underestimate the total uncertainty in the current stock status because they do not account for alternative structural models for herring population dynamics and fishery processes (e.g., natural mortality, steepness, selectivity, the effects of alternative data-weighting choices, and many others). To address structural uncertainties, we investigated a range of alternative models, and we present the key sensitivity analyses along with a suite of other informative sensitivity analyses. However, a model grid has not been used in short-term projections, which are based on the reference model run with MCMC.

For completeness, the results in terms of stock status using the grid of 6 alternative model configurations as described above is shown in Figures 25-27. There are a number of options to generate the joint posterior distributions of plausible outcomes of $SSB_{\mathrm{y}}/SSB_{MSY}$ and F/F_{MSY} as a basis for estimating the probabilities of the stock falling into the respective quadrant of the Kobe phase plot (i.e. stock status). Commonly used approaches to do so include: (i) bootstrap or Markov Chain Monte-Carlo (MCMC) methods to estimate the within model (i.e. reference model) uncertainty (e.g. Walter et al., 2019), (ii) developing a large grid of models to derive the Kobe posterior distribution from a sufficiently large number of point estimates (e.g. n > 500) of $\mathrm{SSB/SSB_{MSY}}$ and $\mathrm{F/F_{MSY}}$ and (iii) a hybrid approach of joining MCMC or bootstrap derived posteriors from alternative model runs to capture both across- and within-model uncertainty (Walter and Winker, 2019). However, in integrated age-structured stock assessment models, such as Stock Synthesis (Methot and Wetzel, 2013), these methods are computationally intense and time consuming as they require to first invert Hessian matrix and then refitting the model (bootstrap) or running sufficiently long MCMC chains (Maunder et al., 2006; Magnusson et al., 2013). This renders them as challenging tasks to complete during typically time-constrained stock assessment meetings. Therefore, the delta-Multivariate log-Normal' (delta-MVLN) estimator (Walter and Winker, 2019; Winker et al., 2019) was used here. It infers within-model uncertainty from maximum likelihood estimates (MLEs), standard errors (SEs) and the correlation of the untransformed quantities F/F_{MSY} and SSB/SSB_{MSY} and it has demonstrated to be able to mimic the MCMC fairly closely. These quantities are derived with Stock Synthesis using the delta-method to calculate the asymptotic variance estimates from the inverted Hessian. To generate Kobe posteriors from a delta-MVLN distribution requires the means and the variance-covariance matrix (VCM) of $\log(SSB/SSB_{MSY})$ and $\log(F/F_{MSY})$. Figure 25 shows that there is about 99% probability that the stock is in the green quadrant of the kobe plot in 2018, i.e. SSB is above SSB_{MSY} and F is below F_{MSY} (F_{MSY} used here was the one estimate by ICES in 2018; $F_{MSY} = 0.21$; ICES 2018). Also, all different model configurations median estimate for 2018 are in the green quadrant of the kobe plot (Figure 26). Finally, the median trend with the 95% confidence intervals of SSB/SSB_{MSY} and F/F_{MSY} for the aggregated 10000 estimates of 2018 stock status of the 11 model options is shown in Figure 27.

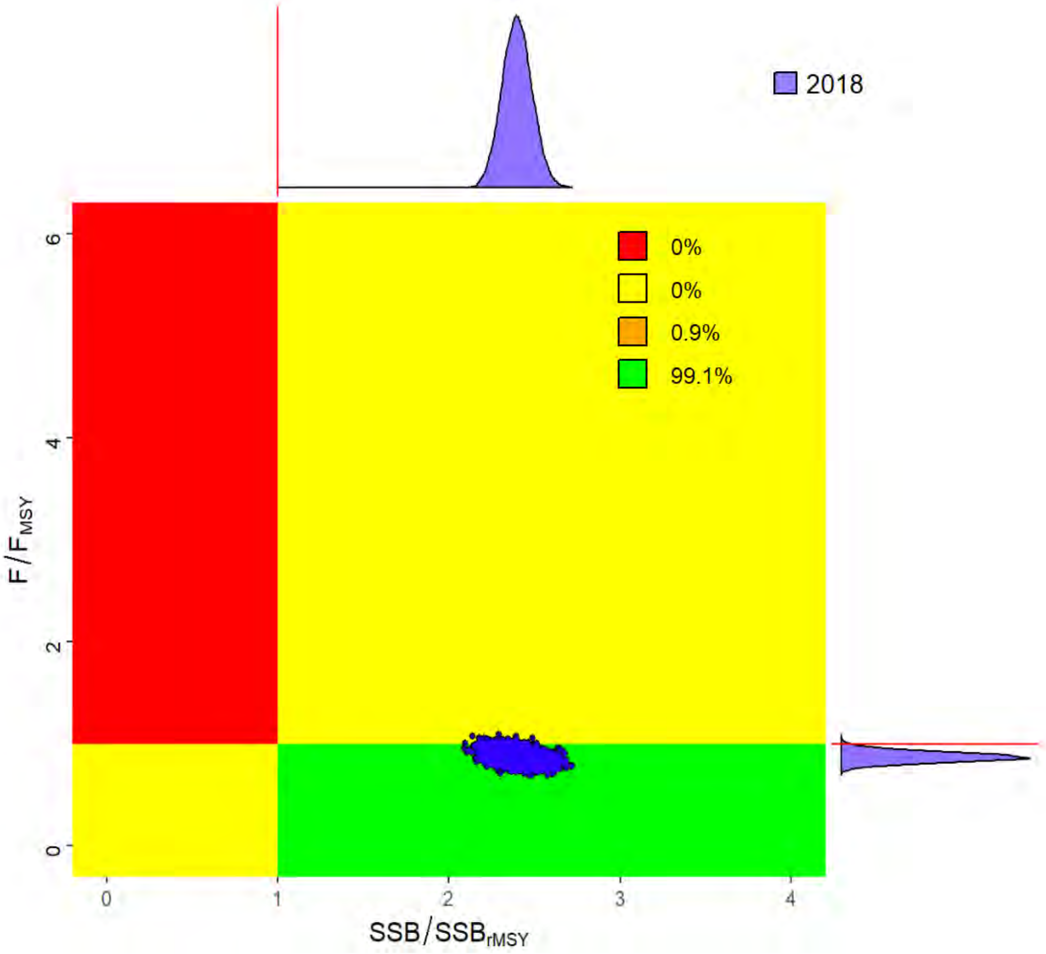

2018
0%
0%
0.9%
99.1%
F / F_MSY
0
2
4
6
0
1
2
3
4
SSB / SSB_rMSY

Figure 25. Herring SDs 30-31. Aggregated assessment Kobe plot. The blue points represent the aggregated 10000 estimates of 2018 stock status from the multinomial approximation from the mean and variance-covariance of the 6 model options. The legend indicates the estimated probability of the stock status being in each of the Kobe quadrant.

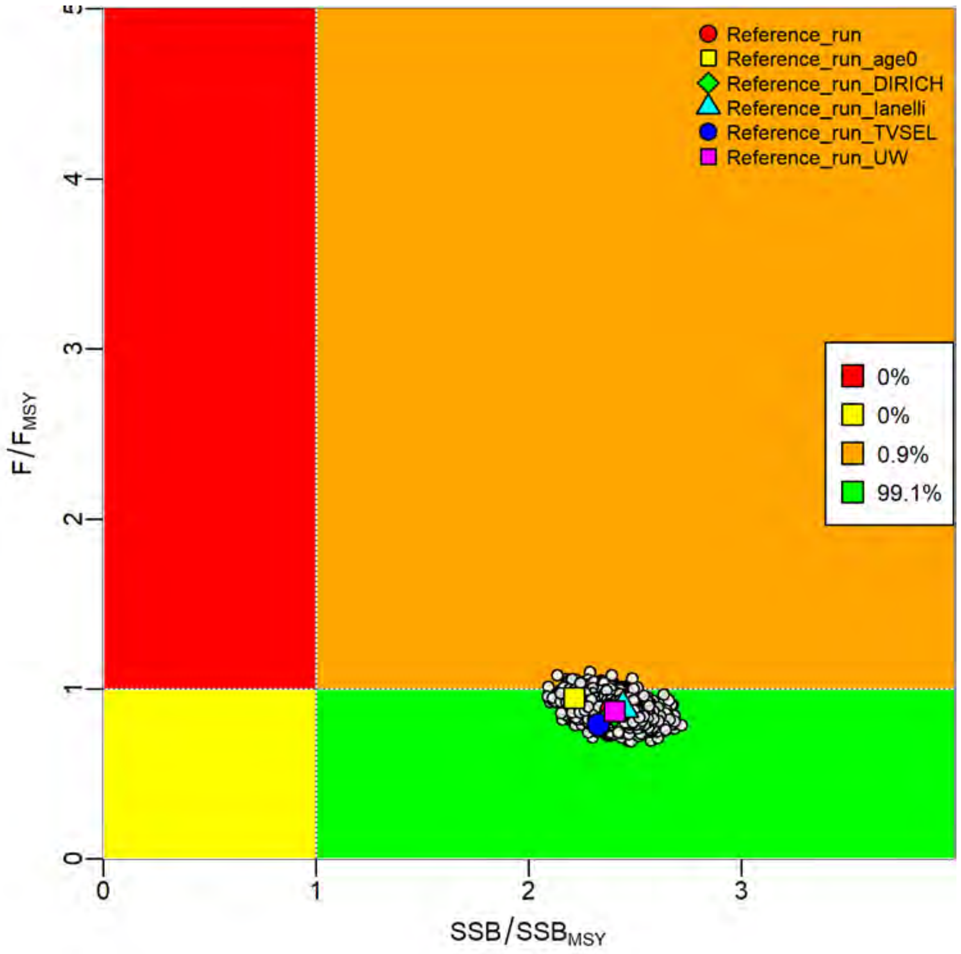


Figure 26. Herring SDs 30-31. Aggregated assessment Kobe plot. The coloured points represent stock status estimates from the 6 model options. The white points represent the aggregated 10000 estimates of 2018 stock status from the multinomial approximation from the mean and variance-covariance of the 11 model options. The legend indicates the estimated probability of the stock status being in each of the Kobe quadrant.

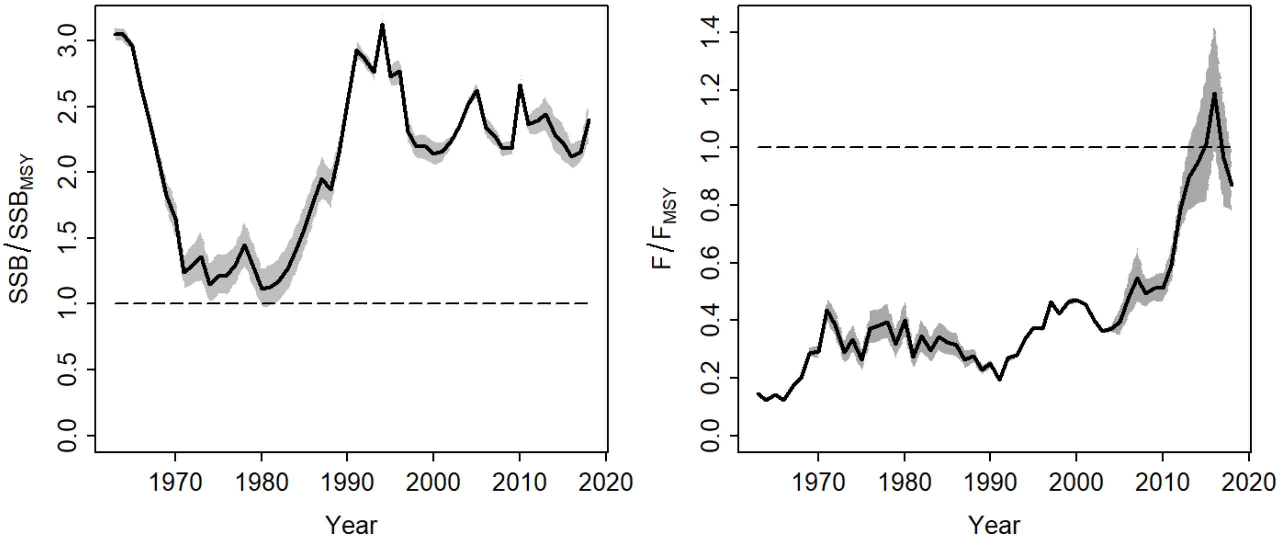


Figure 27**.** Herring SDs 30-31. The median trend with the 95% confidence intervals of SSB/SSB$_{MSY}$ and F/F$_{MSY}$ /for the aggregated 10000 estimates of 2018 stock status from the multinomial approximation from the mean and variance-covariance of the different 6 model options.

Short-term projection

The short-term projections are made with Stock Synthesis using the Reference model using MCMC. Recruitment in the forecast period was set to the average of the last 10 years for which recruitment deviations are estimated in the Stock Synthesis model. Probabilistic forecasts were used. In this approach, catch and SSB levels corresponding to different catch options are calculated as in typical deterministic short term forecast but using MCMC to make it possible to also include the most correct associated probability of the SSB to be below biomass reference points, for each year of forecast. Therefore, an MCMC with 1100000 iterations, 100000 burn-in and 1000 thinning was run for the different levels of assumed F in 2020 and 2021, assuming F in 2019 (i.e. intermediate year) equal to F in 2018. Figure 28 shows the kobe plot of the stochastic forecast for 2021 conducted applying different fishing F levels (F at 80, 90, 100, 110, 115 (i.e. F$_{MSY}$) and 120% of the F in 2018) in 2020. The results show that the stock will remain in the green quadrant of the kobe plot up to F = 0.208. The probability of the SSB to be above B$_{trigger}$ and F below F$_{MSY}$ in the forecast period is 100 (Tables 9 and 10). B$_{trigger}$ used here was the same as reported by ICES in 2018 (ICES 2018).

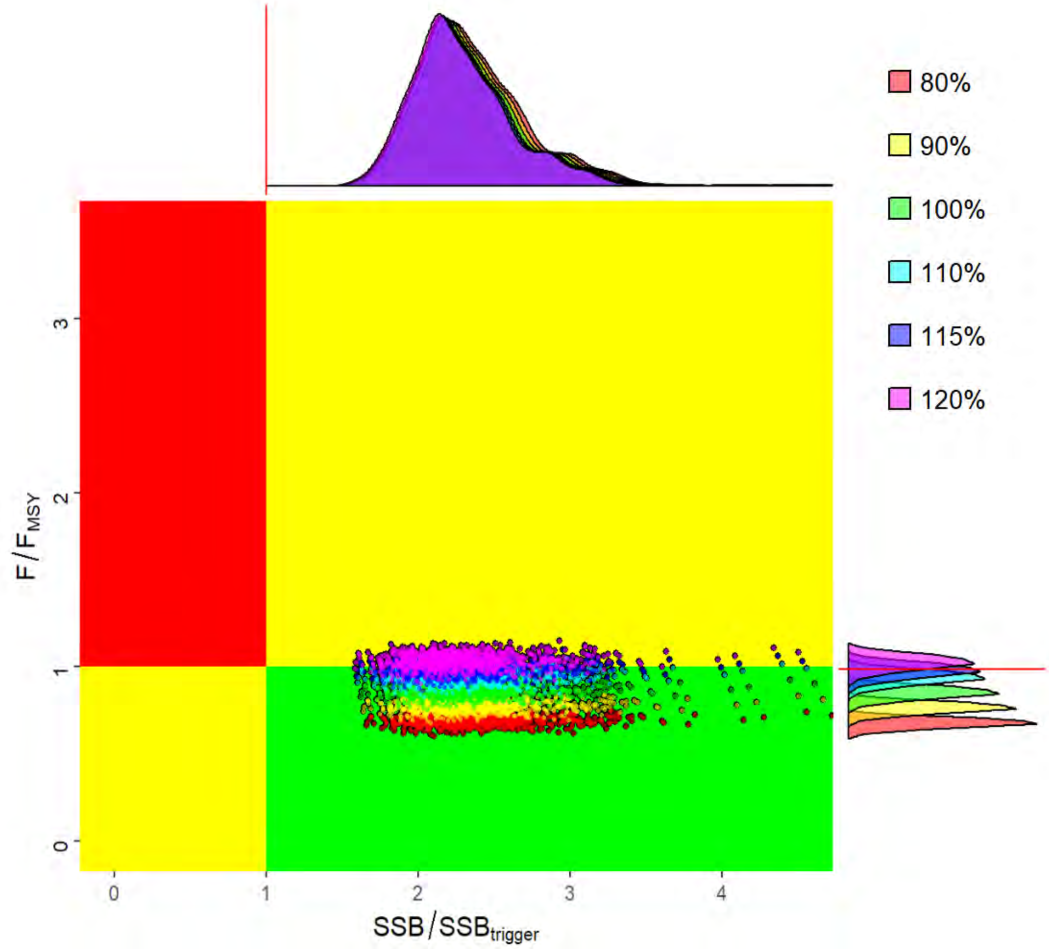

80%
90%
100%
110%
115%
120%
F/F_{MSY}
0
1
2
3
0
1
2
3
4
$SSB/SSB_{trigger}$

Figure 1. Herring SDs 30-31. Kobe plot of the stochastic forecast for 2021 conducted applying different F options in 2020 (i.e. F at 80, 90, 100, 110, 115 (i.e. F_{MSY}) and 120% of the F in 2018), shown in y-axis. The x-axis shows the corresponding spawning stock biomass (SSB) relative to $B_{MSY,}$ while the y-axis shows corresponding F relative to F_{MSY}.

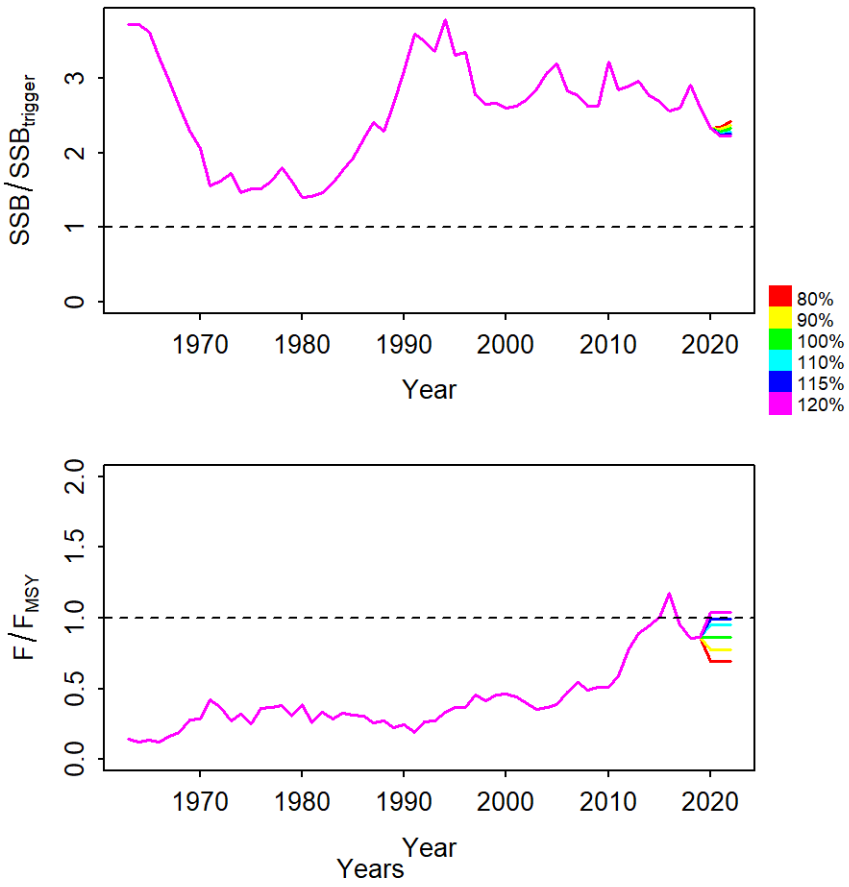
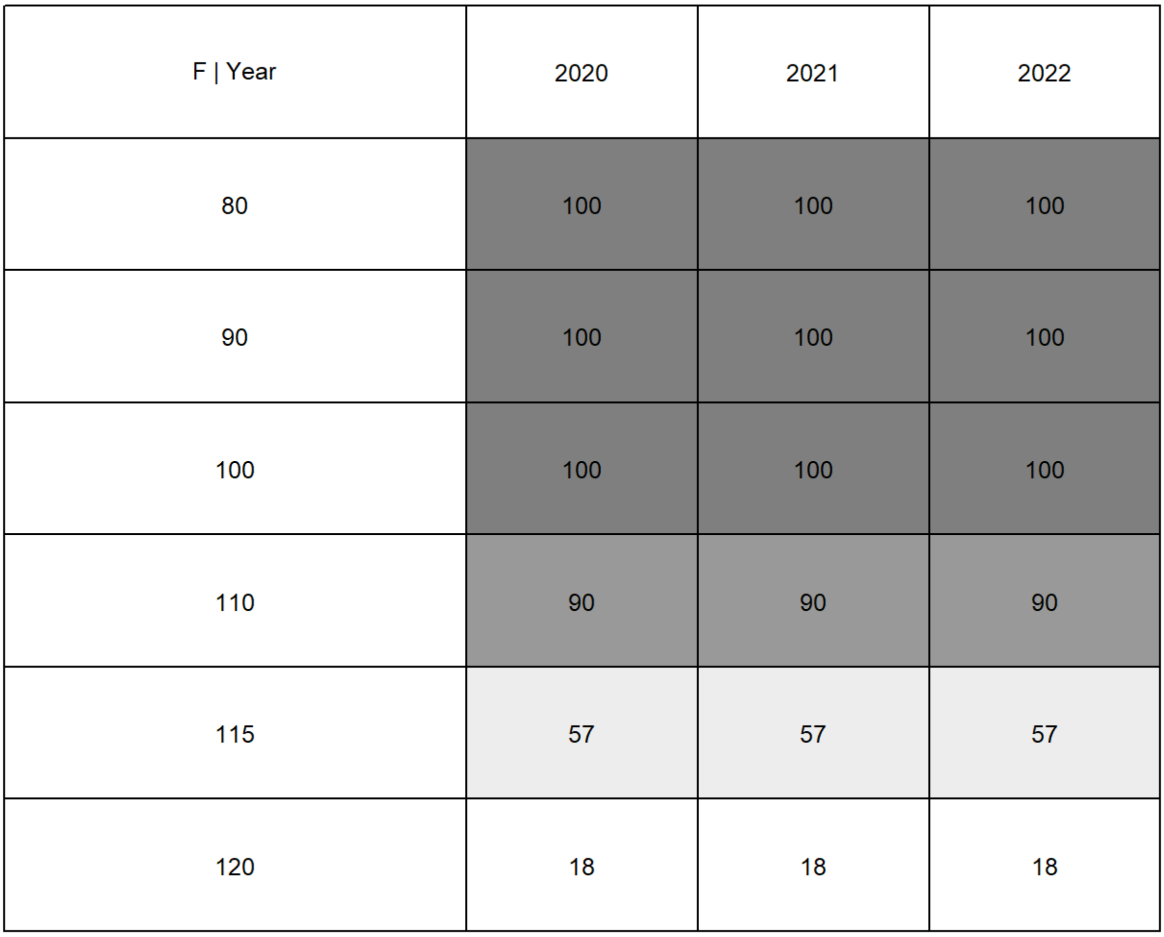


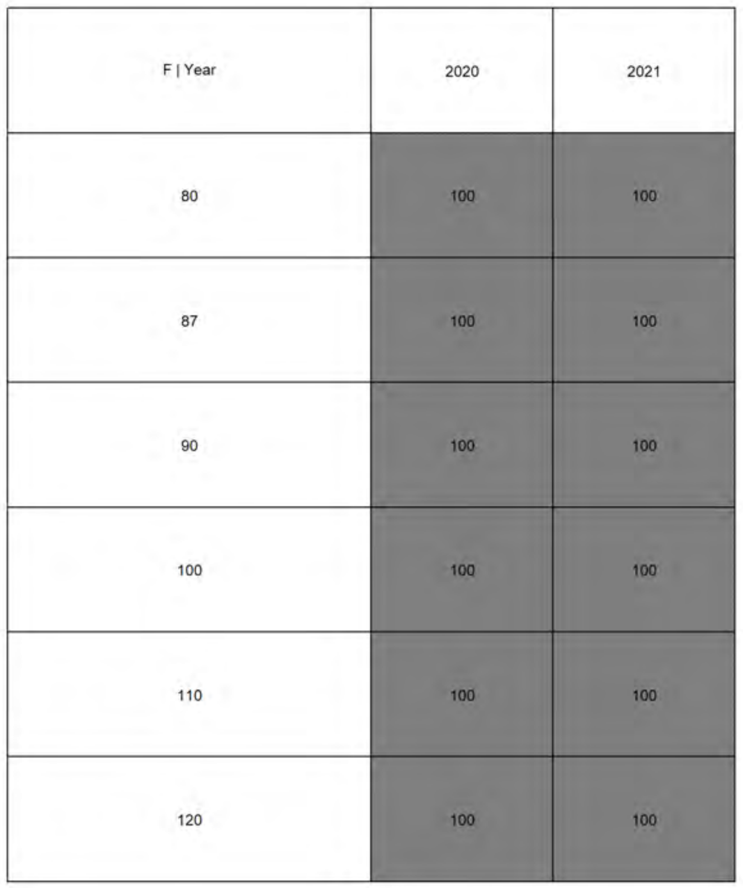
Figure 29. Herring SDs 30-31. Deterministic trend (median) of SSB and F of the stochastic forecast for 2021 conducted applying different F options in 2020 (i.e. F at 80, 90, 100, 110, 115 (i.e. F_{MSY}) and 120% of the F in 2018).

Table9. Herring SDs 30-31. Probability of F to fall below F_{MSY} between 2020 and 2022 at different level of catch options in 2020 and 2021.



F | Year	2020	2021	2022
80	100	100	100
90	100	100	100
100	100	100	100
110	90	90	90
115	57	57	57
120	18	18	18

Table 10. Herring SDs 30-31. Probability of SSB to be above B_{MSY} between 2020 and 2022 at different level of catches in 2020 and 2021.



F | Year	2020	2021
80	100	100
87	100	100
90	100	100
100	100	100
110	100	100
120	100	100

Table 11. Herring SDs 30-31. Short term forecast table. F level is the F expressed as % of the F estimated in 2018.

F levels	Year	SSB	Catch	F
80	2020	658728	72993	0.145
80	2021	663288	72613	0.145
80	2022	685240	74017	0.145
90	2020	658728	81485	0.164
90	2021	654937	79949	0.164
90	2022	671049	80747	0.164
100	2020	658728	89835	0.182
100	2021	646744	87070	0.182
100	2022	658137	87032	0.182
110	2020	658728	98084	0.200
110	2021	639064	93823	0.200
110	2022	644838	92807	0.200
115	2020	658728	101817	0.208
115	2021	635731	96823	0.208
115	2022	638890	95359	0.208
120	2020	658728	106208	0.218
120	2021	631814	100270	0.218
120	2022	631922	98250	0.218

Medium-term projections

Not relevant.

Long-term projections

Not relevant.

Appropriate Reference Points (MSY)

For the time being, we have used F_{MSY} (0.21), B_{lim} (202272 t) and $B_{trigger}$ (283180) as derived by ICES in 2018 (ICES 2018) and B_{MSY} (341148 t) as derived by the Stock Synthesis reference model.

Additional runs at the benchmark

Additional runs were requested by the reviewers during the benchmark. One run was conducted excluding age 1 from the acoustic survey since the poor internal consistency between age 1 and 2 in this survey;, a run with the acoustic index in units of biomass instead of abundance (BIO); a run with all fleet selectivities modelled as logistic (LOGSEL); one run with a lower survey CV (i.e. 0.05) to attempt improving the fit of the survey indices; a run with a higher SigmaR and higher steepness

(SigStep), and four runs with different natural mortality assumptions, fixed 0.15, Lorenzen scaled to 0.15, 0.20 and 0.23 (M015, M015Lor, M020Lor, M023Lor). The nine alternative models generally did not improve the likelihood or convergence of the model, or the fits to different data sources, compared to the reference run (Table 12).

Table 12. Herring SDs 30-31. Likelihood component, parameter values and derived model quantities for the alternative model configurations. The values in the likelihood component of each model indicate changes in likelihood units compared to the reference model. Values +/- 2 likelihood units are considered significantly different.

Type	Reference	Age1	BIO	LOGSEL	M015	M015Lor	M020Lor	SigStep	SurveyCV	M023Lor
TOTAL_likelihood	128	181	131	196	188	168	153	153	189	140
AIC	398	503	404	522	517	478	449	446	521	423
deltaAIC	0	105	7	125	120	80	51	49	123	25
Survey_likelihood										
ALL		-2.9	5.7	5.3	3.3	1.7	-0.4	-0.9	19.1	-0.2
Acoustics		12.3	21.0	17.5	12.2	13.2	14.0	14.9	21.1	14.4
Trapnet		11.5	11.4	14.6	17.8	15.2	12.4	10.9	24.7	12.2
Age_likelihood										
ALL		55.8	-0.9	69.8	-5.7	-9.1	9.0	-0.6	39.8	3.7
Fleet		0.5	-1.5	66.3	-5.8	-9.0	8.7	-0.6	25.5	3.7
Acoustics		55.3	0.5	-0.2	-1.2	0.1	0.5	0.5	10.6	0.1
Trapnet		0.1	0.1	3.7	1.2	-0.2	-0.2	-0.5	3.7	-0.1
Derived quantities										
SB0	517425	533810	498102	727115	50059	48718	409031	844800	476115	439308
SSB_2018	407361	451487	335248	548795	210809	206293	266457	426617	417086	299132
F_2018	0.18	0.16	0.23	0.12	0.29	0.34	0.27	0.18	0.18	0.24
SSB_MSY	151283.5	157126.5	146291.5	200696	258048.5	104768	101553	211511	139480	116090

The BIO run performed in a similar way compared to the reference run. Therefore, more runs were requested by the reviewers to inspect more thoroughly the differences between the use of a biomass or abundance acoustic survey index. One run with the unweighted reference run (UW), one with the unweighted biomass run (BIOUW), one with the reweighted biomass run (BIO), and one with the reference run with time-varying selectivity for the last 6 years in order to try to take into account the drop in the mean age in the catches in the last four years (TVSEL). The new BIO run slightly improved the likelihood of the model, and the fits to different data sources, compared to the reference run (Table 13).

Table 13. Herring SDs 30-31. Likelihood component, parameter values and derived model quantities for the alternative model configurations. The values in the likelihood component of each model indicate changes in likelihood units compared to the reference model. Values +/- 2 likelihood units are considered significantly different.

Type	Reference	BIO	BIOUW	UW	TVSEL
TOTAL_likelihood	124	119	144	143	122
AIC	389	379	429	437	458
deltaAIC	0	-10	40	48	68
Survey_likelihood					
ALL		2.2	3.2	1.0	-0.4
Acoustics		2.5	2.8	0.5	-0.3
Trapnet		-0.3	0.4	0.5	-0.1
Age_likelihood					
ALL		-6.0	16.1	17.4	-3.1
Fleet		-6.0	16.1	17.4	-3.1
Acoustics		-0.1	0.3	0.0	0.0
Trapnet		0.1	-3.4	-3.5	0.0
Derived quantities					
SB0	512310	501425	503470	513560	514885
SSB_2018	387020	345018	347128	388678	389512
F_2018	0.19	0.21	0.21	0.19	0.18
SSB_MSY	151627.5	148853	148527	151071	153265

The results of the different runs in terms of SSB and F are shown in Figure 30. The runs using an acoustic biomass index show a slightly lower SSB and higher F compared to the runs with the acoustic abundance index.

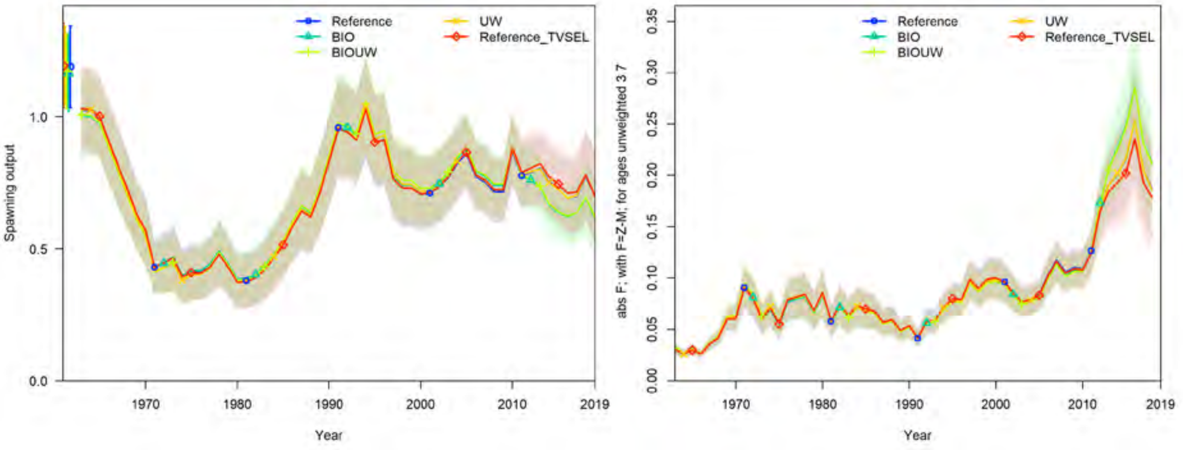


Figure 30. Herring SDs 30-31. Summary of the alternative stock assessment runs. SSB and F with 95% confidence intervals. SSB is in tonnes.

The runs with the acoustic abundance index were preferred compared to the ones with the biomass index since the retrospective analysis failed for the biomass index runs. Thus, the working group in the end decided to accept as final run the reference model (i.e. original model configuration with the acoustic survey index expressed in terms of abundance).

References

Carvalho, F., Punt, A. E., Chang, Y. J., Maunder, M. N., & Piner, K. R. (2017). Can diagnostic tests help identify model misspecification in integrated stock assessments? Fisheries Research, 192, 28-40.

Gibbons, J.D. & Chakraborti, S. (1992). Nonparametric Statistical Inference. New York: Marcel Dekker.

Hilborn, R. (2016). Correlation and causation in fisheries and watershed management. Fisheries, 41(1), 18-25.

Hurtado-Ferro, F., Szuwalski, C. S., Valero, J. L., Anderson, S. C., Cunningham, C. J., Johnson, K. F., & Ono, K. (2014). Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES Journal of Marine Science, 72(1), 99-110.

Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. International journal of forecasting, 22(4), pp.679-688.

Hyndman, R.J. and Athanasopoulos, G. 2013. Forecasting: principles and practice, an online text book. Retrieved September 16, 2012, from http://otexts.com/fpp/.

ICES. 2018. Baltic Fisheries Assessment Working Group (WGBFAS), 6–13 April 2018, ICES HQ, Copenhagen, Denmark. 748 pp.

Kolody, D. S., Eveson, J. P., Preece, A. L., Davies, C. R., & Hillary, R. M. (2019). Recruitment in tuna RFMO stock assessment and management: A review of current approaches and challenges. Fisheries Research, 217, 217-234.

Kell, L.T., Kimoto, A. and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries research, 183, pp.119-127.

Lorenzen K. (1996) – The relationship between body weight and natural mortality in juvenile and adult fish: a comparison of natural ecosystems and aquaculture. Journal of Fish Biology, 49:627-647.

Magnusson, A., Punt, A.E., Hilborn, R., 2013. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish Fish. 14, no-no. doi:10.1111/j.1467-2979.2012.00473.x.

Maunder, M.N., Harley, S.J., Hampton, J., 2006. Including parameter uncertainty in forward projections of computationally intensive statistical population dynamic models. ICES J. Mar. Sci. 63, 969–979. doi:10.1016/j.icesjms.2006.03.016.

M.N. Maunder, K.R. Piner, 2015. Contemporary fisheries stock assessment: many issues still remain

ICES J. Mar. Sci., 72 (1) (2015), pp. 7-18

Methot, R.D., Wetzel, C.R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research 142 (2013) 86–99.

Monnahan, C.C., Branch, T. A., Thorson, J. T., Stewart, I. J., Szuwalski, C. S., 2019. Overcoming long Bayesian run times in integrated fisheries stock assessments. ICES Journal of Marine Science, fsz059, https://doi.org/10.1093/icesjms/fsz059.

Myers, R. A., Bowen, K. G., & Barrowman, N. J. (1999). Maximum reproductive rate of fish at low population sizes. Canadian Journal of Fisheries and Aquatic Sciences, 56(12), 2404-2419.

Subbey, S. (2018). Parameter estimation in stock assessment modelling: caveats with gradient-based algorithms. ICES Journal of Marine Science, 75(5), 1553-1559.

Siekmann, I., Sneyd, J., & Crampin, E. J. (2012). MCMC can detect non identifiable models. Biophysical journal, 103(11), 2275-2286.

Then, A. Y., Hoenig, J. M., Hall, N. G., Hewitt, D. A. (2014). Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. ICES Journal of Marine Science, 72(1), 82-92.

Walters, C. J., Hilborn, R., & Christensen, V. (2008). Surplus production dynamics in declining and recovering fish populations. Canadian Journal of Fisheries and Aquatic Sciences, 65(11), 2536-2551.

Walter, J., Winker, H., 2019. Projections to create Kobe 2 Strategy Matrices using the multivariate log-normal approximation for Atlantic yellowfin tuna. ICCAT-SCRS/2019/145 1–12.

Willmott, C.J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), pp.79-82.

Winker, H., Walter, J., Cardinale, M., Fu, D., 2019. A multivariate lognormal Monte-Carlo approach for estimating structural uncertainty about the stock status and future projections for Indian Ocean Yellowfin tuna. IOTC-2019-WPM10-XX.

Winker, H., Carvalho, F., and Kapur, M., 2018. JABBA: Just Another Bayesian Biomass Assessment. Fisheries Research, Volume 204, August 2018, Pages 275-288.

Zeynep Pekcan-Hekim, Anna Gårdmark, Agnes M. L. Karlson, Pirkko Kauppila, Mikaela Bergenius, Lena Bergström 2016. The role of climate and fisheries on the temporal changes in the Bothnian Bay foodweb. ICES Journal of Marine Science, Volume 73, Issue 7, July 2016, Pages 1739–1749, https://doi.org/10.1093/icesjms/fsw032.

Appendices

Appendix I

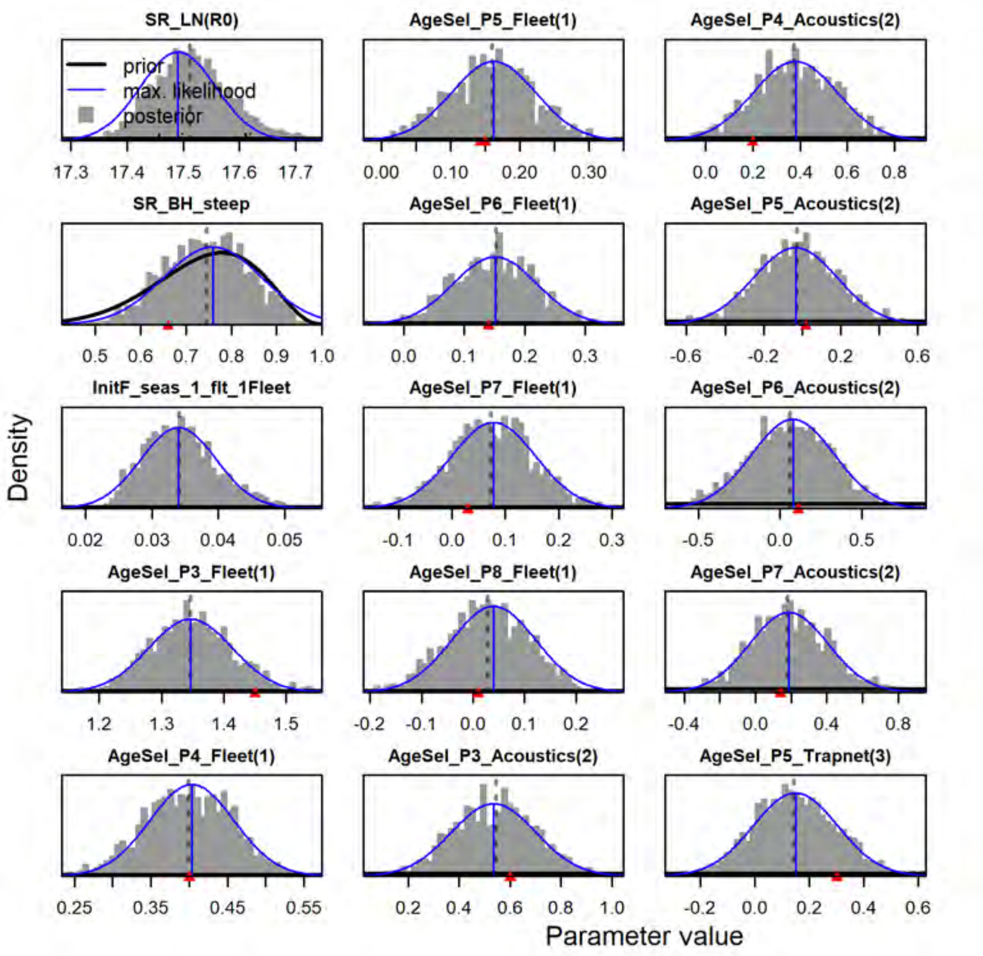


Figure A1. Herring SDs 30-31. Parameter distribution of the reference model. Prior (black lines) and posterior (grey histograms) distributions for key parameters in the reference model obtained by the MCMC run. The MLE and associated symmetric uncertainty intervals are also shown (blue lines).

Appendix II.

GLOSSARY OF TERMS AND ACRONYMS USED IN THIS DOCUMENT

B_0: The unfished equilibrium female spawning biomass.

B_{MSY}: The estimated female spawning biomass which theoretically would produce the maximum sustainable yield (MSY) under equilibrium fishing conditions (constant fishing and average recruitment in every year).

B_{lim}: Spawning biomass below which recruitment is considered to be impaired.

Catchability (q): The parameter defining the proportionality between a relative index of stock abundance (often a fishery-independent survey) and the estimated stock abundance available to that survey (as modified by selectivity) in the assessment model.

Catch-per-unit-effort (CPUE): A raw or (frequently) standardized and model-based metric of fishing success based on the catch and relative effort expended to generate that catch from commercial or survey estimates. Catch per-unit-effort is often used as an index of stock abundance.

Cohort: A group of fish born in the same year. Also see recruitment and year-class.

CV: Coefficient of variation. A measure of uncertainty defined as the standard deviation (SD) divided by the mean.

Fishing mortality rate, or instantaneous rate of fishing mortality (F): A metric of fishing intensity that is usually reported in relation to the most highly selected ages(s) or length(s), or occasionally as an average over an age range that is vulnerable to the fishery.

F_{MSY}: The rate of fishing mortality estimated to produce the maximum sustainable yield (MSY) from the stock.

Markov-Chain Monte-Carlo (MCMC): A numerical method used to sample from the posterior distribution (see below) of parameters and derived quantities in a Bayesian analysis. It is more computationally intensive than the maximum likelihood estimate (see below), but provides a more accurate depiction of parameter uncertainty.

Maximum likelihood estimate (MLE): A method used to estimate a single value for each of the parameters and derived quantities. It is less computationally intensive than MCMC methods (see below), but parameter uncertainty is less well determined.

Maximum sustainable yield (MSY): An estimate of the largest sustainable annual catch that can be continuously taken over a long period of time from a stock under equilibrium ecological and environmental conditions.

NUTS: Hamiltonian Monte Carlo (HMC) is a Markov chain Monte Carlo (MCMC) algorithm that avoids the random walk behaviour and sensitivity to correlated parameters that plague many MCMC methods by taking a series of steps informed by first-order gradient information. These features allow it to converge to high-dimensional target distributions much more quickly than simpler methods such as random walk Metropolis or Gibbs sampling. No-U-Turn Sampler (NUTS), an extension to HMC that eliminates the need to set a number of steps. NUTS uses a recursive algorithm to build a set of likely candidate points that spans a wide swath of the target distribution, stopping automatically when it starts to double back and retrace its steps.

Posterior distribution: The probability distribution for parameters or derived quantities from a Bayesian model representing the result of the prior probability distributions being updated by the observed data via the likelihood equation. For stock assessments, posterior distributions are approximated via numerical methods; one frequently employed method is MCMC.

Prior distribution: Probability distribution for a parameter in a Bayesian analysis that represents the information available before evaluating the observed data via the likelihood equation. For some parameters, non-informative priors can be constructed which allow the data to dominate the posterior distribution (see above). For other parameters, informative priors can be constructed based on auxiliary information and/or expert knowledge or opinions.

R0: Estimated annual recruitment at unfished equilibrium.

Recruits/recruitment: the estimated number of new members in a fish population born in the same age. In this assessment, recruitment is reported at age 0. See also cohort and year class.

Recruitment deviation: The offset of the recruitment in a given year relative to the stock-recruit function; values occur on a logarithmic scale and are relative to the expected recruitment at a given spawning biomass (see below).

Random Walk Metropolis: In statistics and statistical physics, the Metropolis–Hastings algorithm is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution from which direct sampling is difficult. This sequence can be used to approximate the distribution (e.g. to generate a histogram) or to compute an integral (e.g. an expected value). Metropolis–Hastings and other MCMC algorithms are generally used for sampling from multi-dimensional distributions, especially when the number of dimensions is high.

SD: Standard deviation. A measure of variability within a sample.

Steepness (h): A stock-recruit relationship parameter representing the proportion of R0 expected (on average) when the female spawning biomass is reduced to 20% of B0 (i.e., when

Stock Synthesis (SS): The age-structured stock assessment model applied in this stock assessment.

Year-class: A group of fish born in the same year. See also ‘cohort’ and ‘recruitment’.

Background

The assessment for the Gulf of Bothnia herring (SD 3031) in 2019 was not accepted by the Advice Drafting Group and was downgraded from category 1 to 3. The run results from the stockassessment.org for the 2019 assessment can be found under "*GoB_Herring_2019_clonedversfinal*" and in Figure 1 and 2 below. The assessment was not accepted based on the poor retrospective diagnostics where the Mohn's rho values were above 20% for SSB (37%), F (27%) and recruitment (68%) (Figure 2).

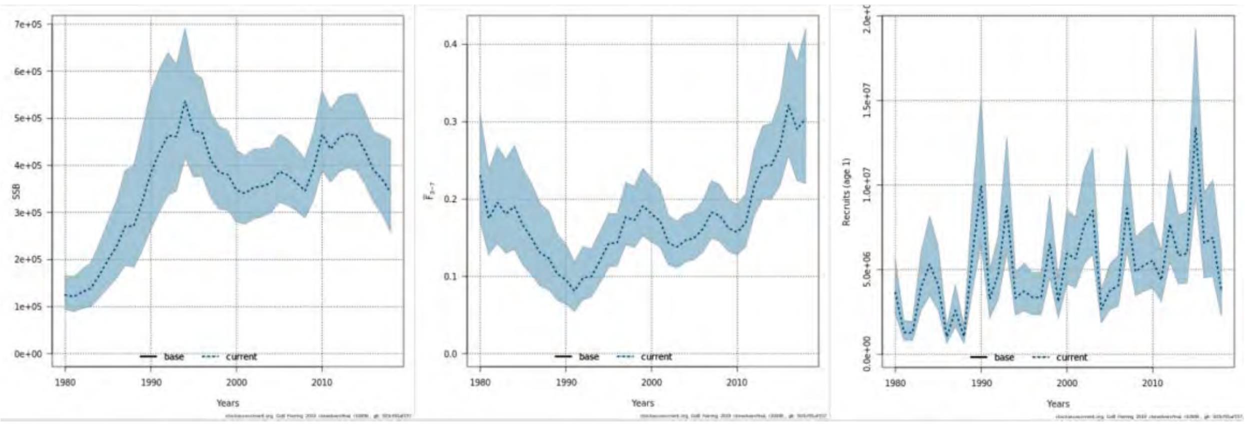


Figure 1. The stock assessment run in SAM in 2019 (*GoB_Herring_2019_clonedversfinal*), which was not accepted. SSB (left), F (middle) and recruitment (right).

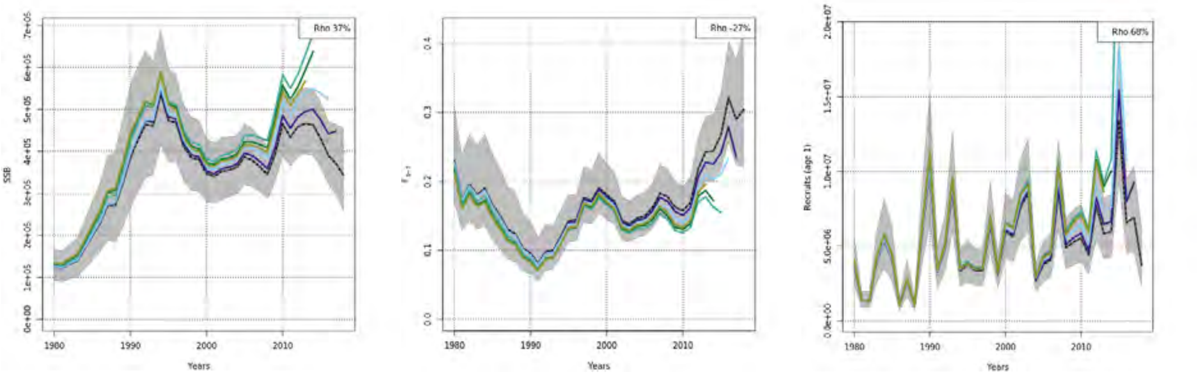


Figure 2. The retrospectives for the stock assessment run in SAM in 2019 (*GoB_Herring_2019_clonedversfinal*), which was not accepted. SSB, F and recruitment.

Due to the downgrading of the stock a benchmark was setup in order to investigate the reasons behind the bad retrospective diagnostics. In addition, a stock synthesis model was setup in parallel that showed good performance. Therefore, it was concluded that the benchmark would also investigate the potential use of a new model (SS3) for the SD 3031 assessment.

Data issues detected

Before the data meeting (19th November) a mistake in the input data in the 2019 assessment was detected. The acoustic survey indices were calculated wrongly from years 2013 to 2015. Years 2013, 2014 and 2015 in the acoustic survey was calculated higher than it actually was. The difference between the correct data and the wrong data used in the 2019 assessment can be found in Figure 3.

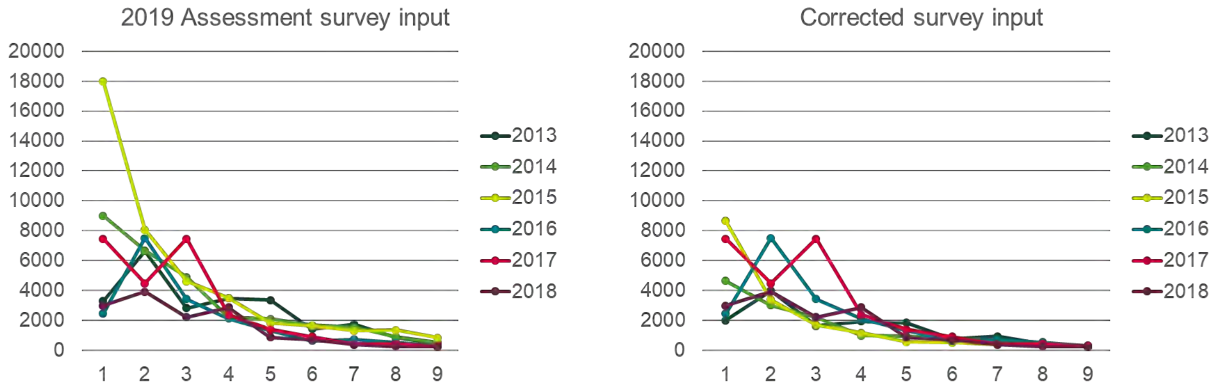


Figure 3. Abundance of herring (Millions) in SD 30 from the Acoustic Survey by age for years 2013-2018. The 2019 assessment survey input (left) and the corrected survey indices (right).

This section of the report contains a run with SAM with corrected acoustic indices compared to the 2019 assessment run. In addition, a SAM run with new data, which is used in the SS3 model was run in order to compare the outputs from the two different models.

SAM run with corrected acoustic index compared to the 2019 assessment run

The SAM run "*GoBHer2020"* includes the corrected data for the acoustic survey index for years 2013-2018 while keeping all other input data the same without changing the configuration from 2019 assessment run. Comparison of the survey input data used in the 2019 assessment and the corrected survey indices used in the *GoBHer2020* can be found in Figure 3.

The SSB, F and recruitment from the SAM run with corrected survey index compared to the assessment in 2019 are shown in Figure 4. There is a decline in SSB in the corrected run starting 2010 instead of an increase and a decline later on in year later in 2015, which showed on the 2019 assessment with the incorrect data. The retrospectives for the run *GoBHer2020* improved from the 2019 assessment run substantially for SSB (Mohn's rho 37% to 19%), and for recruitment (Mohn's rho 68% to 5%). However, the retrospectives for F worsened (Mohn's rho 27% to 34%) (Figure 5).

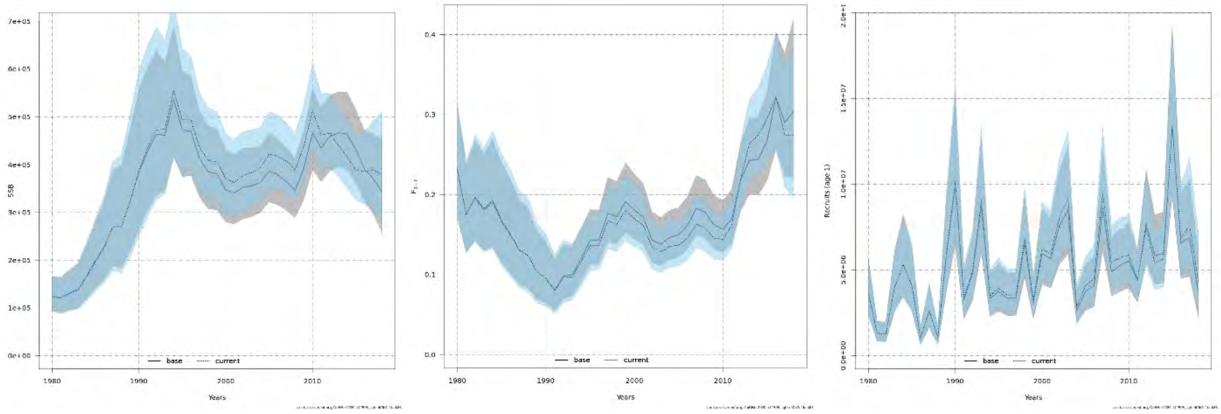


Figure 4. The stock assessment run in SAM "*GoBHer2020"* comparing the Spawning stock biomass (left), F (middle) and recruitment (right) for data with corrected acoustic index values (current, blue) to base run (assessment run in 2019, grey).

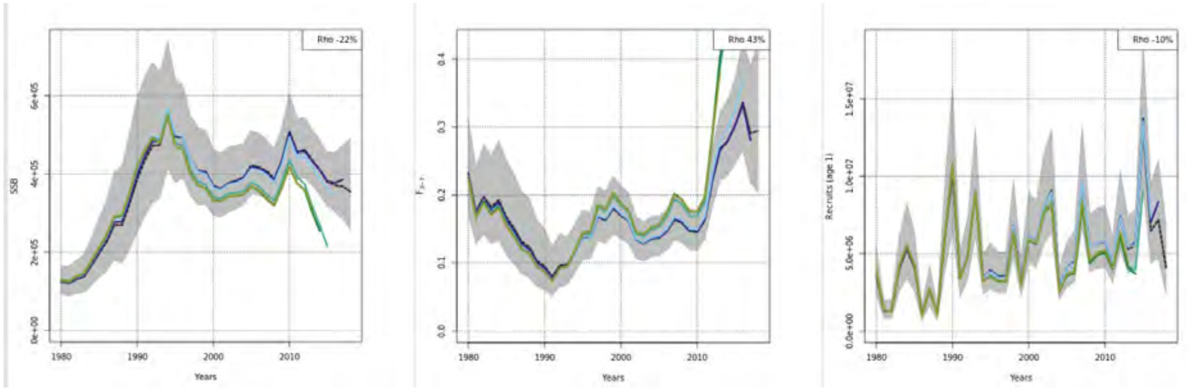


Figure 5. The retrospectives for the stock assessment run in SAM “*GoBHer2020*” for SSB (left), F (middle) and recruitment (right). The Mohn’s rho values for SSB, F and recruitment are -22%, 43% and 10%, respectively.

SAM run with new data (same data as in the SS3)

The SAM run including the same input data as in the SS3 reference run (with ages expanded to 15+, with age varying yearly constant natural mortality and updated survey indices) can be seen for SSB, F and recruitment in figure 6. The retrospectives in Figure 7 show Mohn’s rho values of 17%, 34% and 8% for SSB, F and recruitment, respectively. These values have improved for SSB however worsened for F and recruitment when comparing to the run with corrected survey index (*GoBHer2020*). Note that the configuration is kept the same as in the assessment run in 2019.

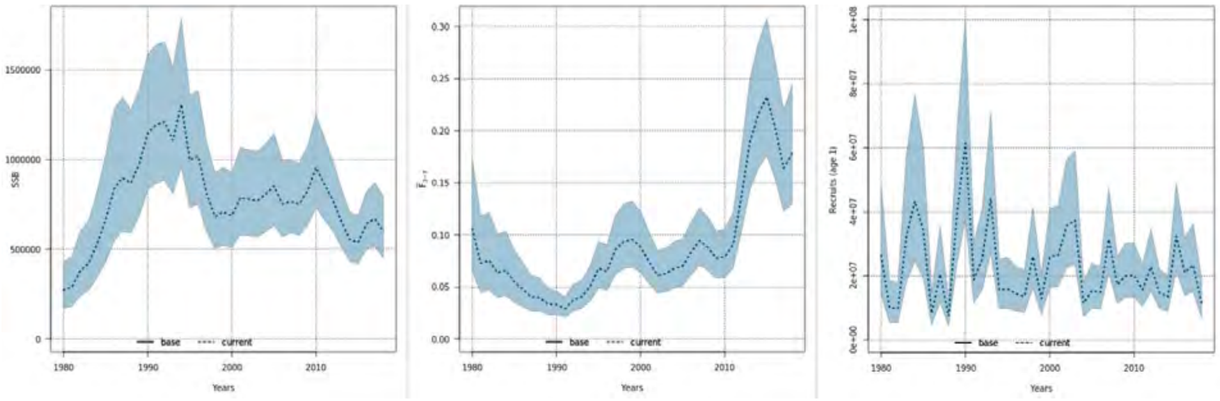


Figure 6. The stock assessment run in SAM “*GoB_Herring_2020_SS_3”* with the new data for Spawning stock biomass (left), F (middle) and recruitment (right).

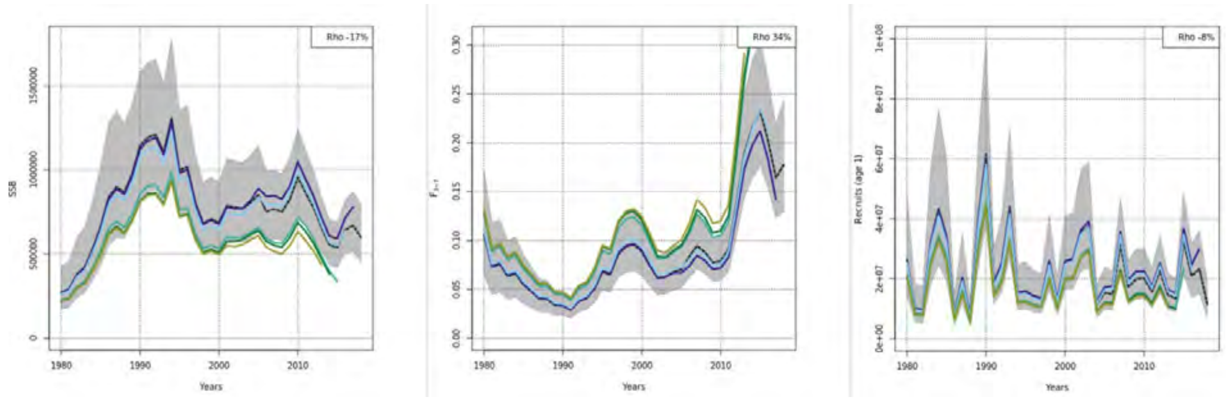


Figure 7. The retrospectives for the stock assessment run in SAM “*GoB_Herring_2020_SS_3*” for SSB (left), F (middle) and recruitment (right). The Mohn’s rho values for SSB, F and recruitment are -17%, 34% and 8%, respectively.

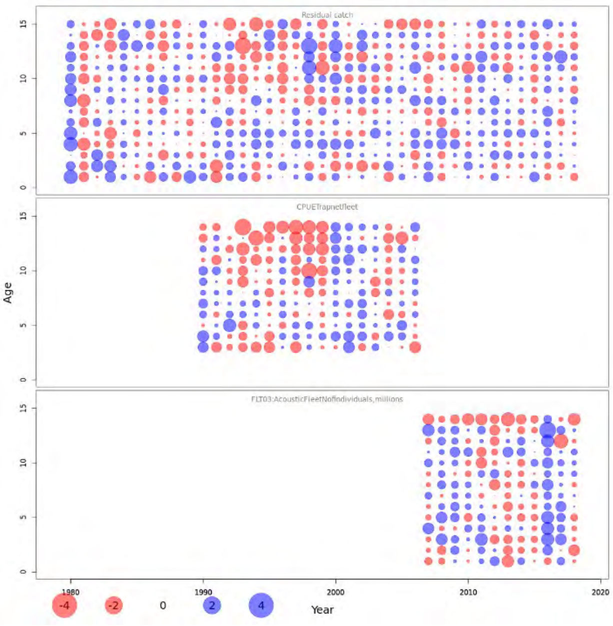


Figure 8. Residuals for the stock assessment run in SAM “*GoB_Herring_2020_SS_3*” for catch (top), trapnet CPUE (middle) and acoustic survey (bottom).

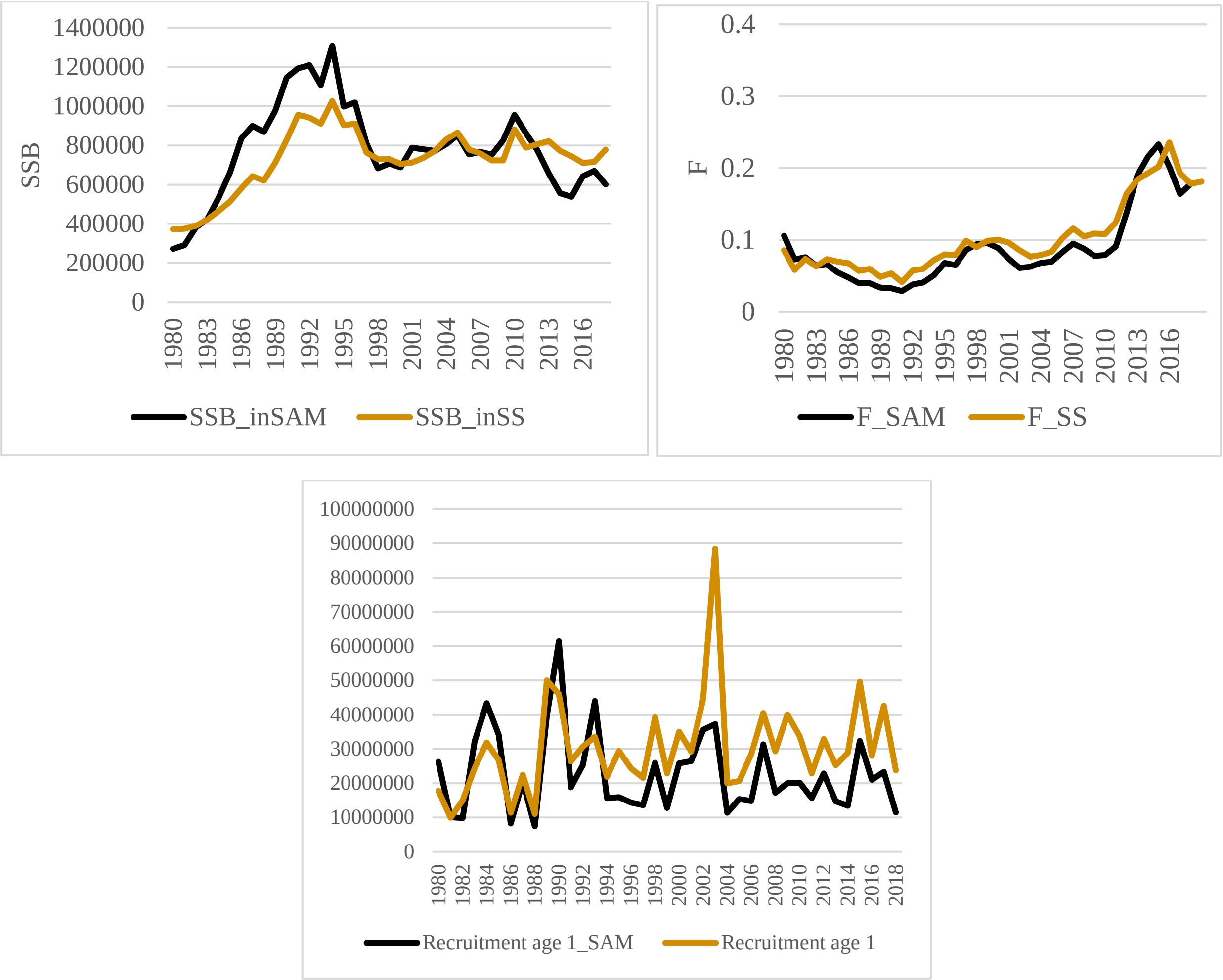


Figure 9. Comparison for SSB, F and Recruitment age 1 for the SS and SAM models with same input data.

WD4. Additional reviewers' comments related to WD2 and WD3. February 2020.

Introduction

A Benchmark Workshop on herring (*Clupea harengus)* in the Gulf of Bothnia (WKCluB) met in Copenhagen, Denmark, in 4–6 February 2020 for a 3 day Benchmark meeting:

a) Evaluate the appropriateness of data and methods to determine stock status and investigate methods for short-term outlook taking agreed or proposed management plans into account for the stocks listed in the text table below. The evaluation shall include consideration of:
   1. Examine SS3 as an alternative assessment model to SAM;
   2. Explore impact of all tuning fleets on assessment estimates;

b) Agree and document the preferred method for evaluating stock status and (where applicable) short-term forecast and update the stock annex as appropriate. Knowledge about environmental drivers, including multispecies interactions, and ecosystem impacts should be integrated in the methodology. If no analytical assessment method can be agreed, then an alternative method (the former method, or following the ICES data-limited stock approach) should be put forward;
   1. Update the stock annex as appropriate;
   2. Re-examine and update MSY and PA reference points according to ICES guidelines (see Technical document on reference points);
   3. Prioritize recommendations for future improvements of the assessment methodology and data collection;
   4. Produce working documents to be reviewed during the Benchmark meeting at least 7 days prior to the meeting.

This assessment follows on from a 2018 benchmark that was rejected for advice given retrospective patterns. As such, data were revisited and alternative models were evaluated.

The reviewers supported and critically reviewed the work performed in the data meeting in January and also, in the final presentation meeting in February in Copenhagen.

Data

Biological data from each gear type were presented and reviewed. In the course of the workshop, some edits were noted in the data files created (e.g. an "outlier" weight-at-age was nearly five times greater than the mean, and this caused a model prediction being abnormally high in one year). Additionally, the recent main information indexing the stock (the acoustic time-series) was "corrected" for the assessment and this was one of the key differences (apparently) from the 2018 benchmark. The source and process of the correction was not presented during the review period other than a verbal explanation having to do with the proportion of herring in the catches.

Key assumptions and inputs were discussed and reviewed, including the extent of variability in some time-varying age-specific schedules. It was also noted that there was a fair amount of variability over time in the body mass-at-age specified for use in the assessment. It was unclear the extent that this variability was due to sampling or actual variability. A similar degree of variability was noted for the time-varying maturity-at-age data. Since there is some indication of density-dependent growth, approaches to smooth or regularize the data might be considered in future assessments.

For example, the figure below shows that given a constant equilibrium age structure (**N**) and computing the time-series as the sum of annual maturity-at-age (or weight-at-age) times N-at-age (and normalizing to have mean 1.0 over time) shows that these biological age-specific schedules introduce variability on their own. This variability, in conjunction with fishery selectivity estimates, will affect reference point calculations. This also illustrates that the stock weight is lower in the recent period; assuming the recent period for future projections and reference points should acknowledge that some variability is being ignored. This also implies that the input spawning biomass-per-recruit for different fixed fishing mortality rates (and selectivity) will vary given these biological schedules alone.

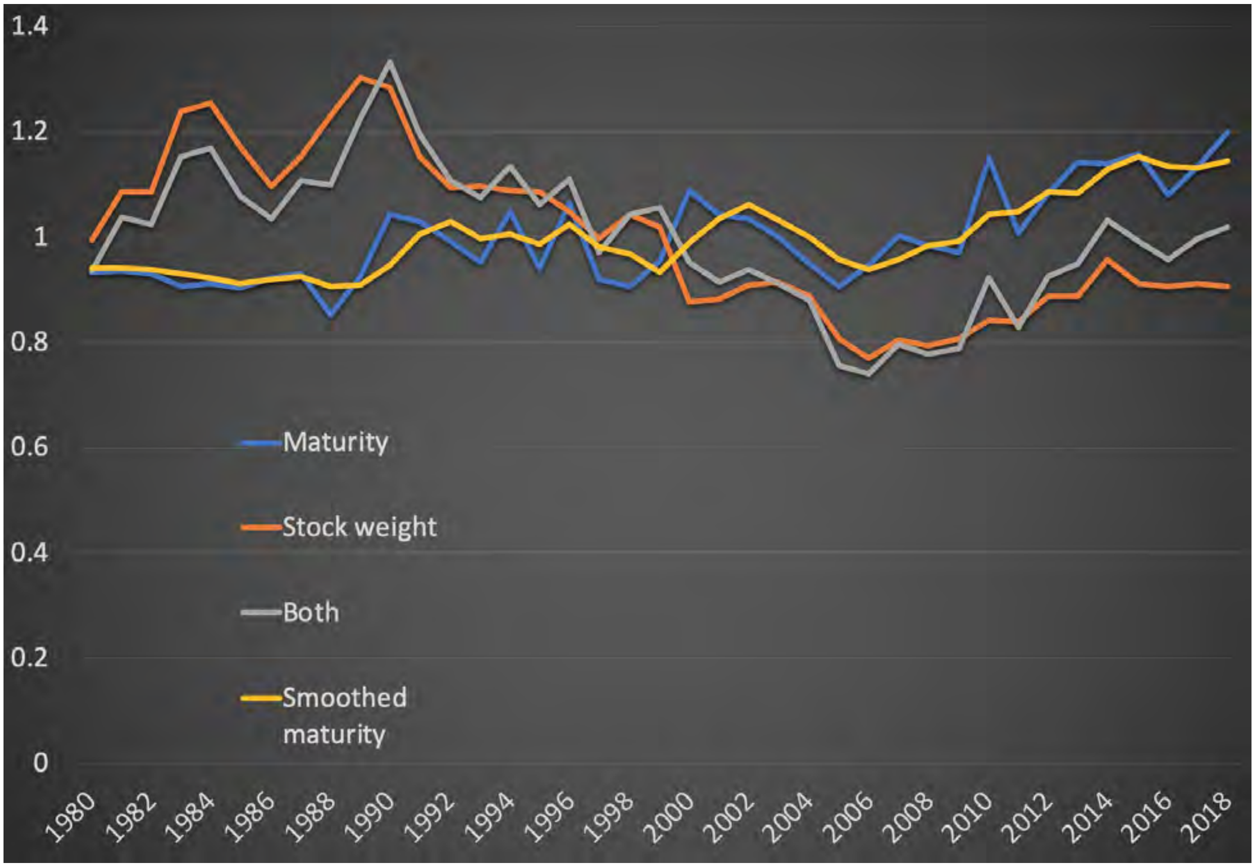


Figure 1. Historical "biomass" or "Mature population" (normalized to have mean 1) given a fixed equilibrium numbers-at-age. "Both" means application of maturity and stock weight combined.

Age-determination precision and accuracy is considered quite high for this stock. However, data are available that could provide some insight on errors (particularly for older ages) and possible biases and this could be explored in future benchmark assessments.

The analysis presented the history and sampling patterns for the data used in the assessment. It was noted that for the 2018 data, the catch by quarter and area were sampled in an irregular way (i.e. some lower catch strata had lots of samples and in some strata with high catches, sample sizes were relatively small). Without further understanding, such sampling discrepancies should be cross-checked so that the relative efficiency/precision can be appropriately accounted for within the assessment.

A novel approach to specifying age-specific natural mortality was proposed and during the benchmark, some alternatives were considered. The alternatives, based on the reference model configurations failed to show improved data fits so the fixed values were considered acceptable (noting that they differ substantially from the previous constant value-at-age of 0.15).

Presentation alternatives of catch-at-age and survey abundance-at-age were used to help clarify how data are affecting model results (examples shown in Figure 2 below). Such depictions help with understanding how recent years fishery, and to a lesser degree survey, data indicate a younger population.

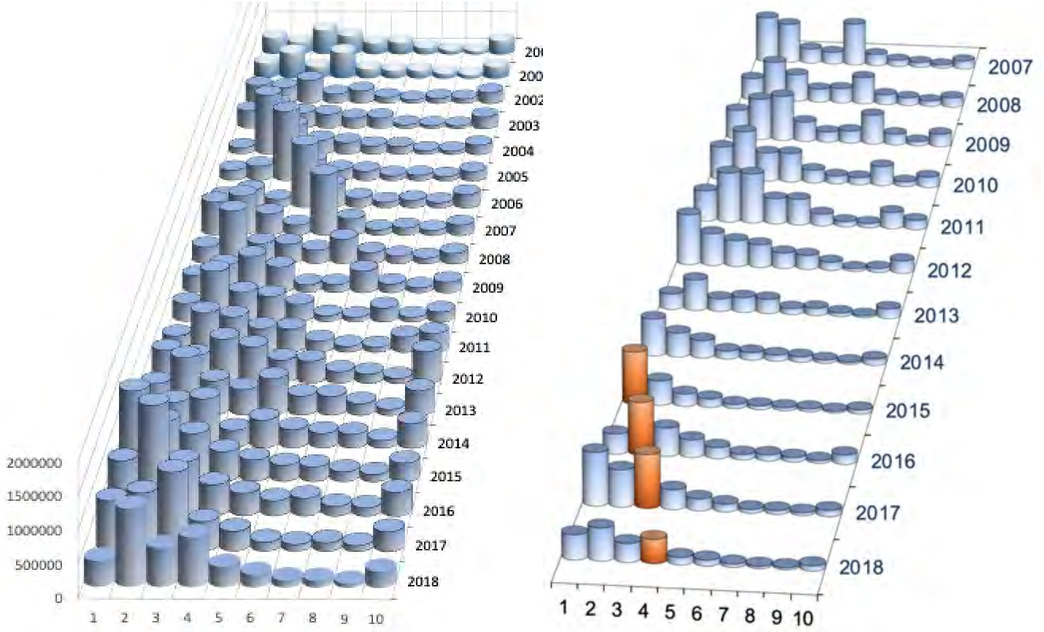


Figure 2. Alternative presentations of recent fishery catch-at-age (left) and acoustic estimates of numbers-at-age (right) for Gulf of Bothnian herring.

Models

This benchmark was primarily focussed on understanding what caused the poor retrospective patterns observed from the 2018 review. Within the SAM model framework, this task was more difficult to diagnose. Consequently, the same data were evaluated within the stock synthesis (SS) framework for comparison and to provide an alternative set of diagnostics. These are summarized below.

Stock Synthesis framework

While fundamentally the data were the same between SAM and SS, the SS configuration decomposed the data components into index and composition data (whereas SAM used indices-at-age). The indices were the trapnet cpue trends (TNT) and acoustic survey abundance trends (ACT). The age composition data were from trapnet (TNAD), acoustic survey (ACAD) and catch age distribution (CAD). The SS fit to the age distributions was generally good, whereas the fit to the trend data was considered acceptable.

A number of model runs (six prior to meeting, and an additional 16 during the benchmark) were conducted for evaluation at this benchmark. The analysts presented extensive diagnostic tests including the standard ICES criterion related to retrospective patterns. This was considered an enhancement over using one method for accepting or rejecting an assessment. It was noted that the final retrospective pattern had low and acceptable values of Mohn's rho. Nonetheless, alternative natural mortality schedules, selectivity patterns, and calibration indices were explored during the benchmark to see if further improvements in the retrospective pattern could be obtained. It was clear that aspects of introducing the ACAD and CAD data were the main cause of this pattern.

SAM

The SAM assessment model approach was refined during the benchmark and in the end, provided similar results to the stock synthesis model runs. In particular, the poor retrospective pattern was largely resolved although remains outside the thumb rule (+-20%) and furthermore, the pattern was consistent with the pattern seen from the stock synthesis model runs. This was encouraging and should be developed further for considerations at the next benchmark (provided more diagnostic evaluations of process errors and partial Fs-at-age (e.g. for comparing selectivities).

Alternative model

During the review, and to aid in understanding the data and assessment approaches, a simple model with many of the same features as stock synthesis was constructed by one of the external reviewers (Jim). This model confirmed many of the results with respect to patterns in spawning biomass and retrospective runs. One investigation was the option to include a fully time-varying selectivity (with some modest constraint). Results shows some differences in the partial Fs that might be worth considering in future benchmarks, as these are qualitatively different than both SAM and SS configurations (Figures 3 and 4).

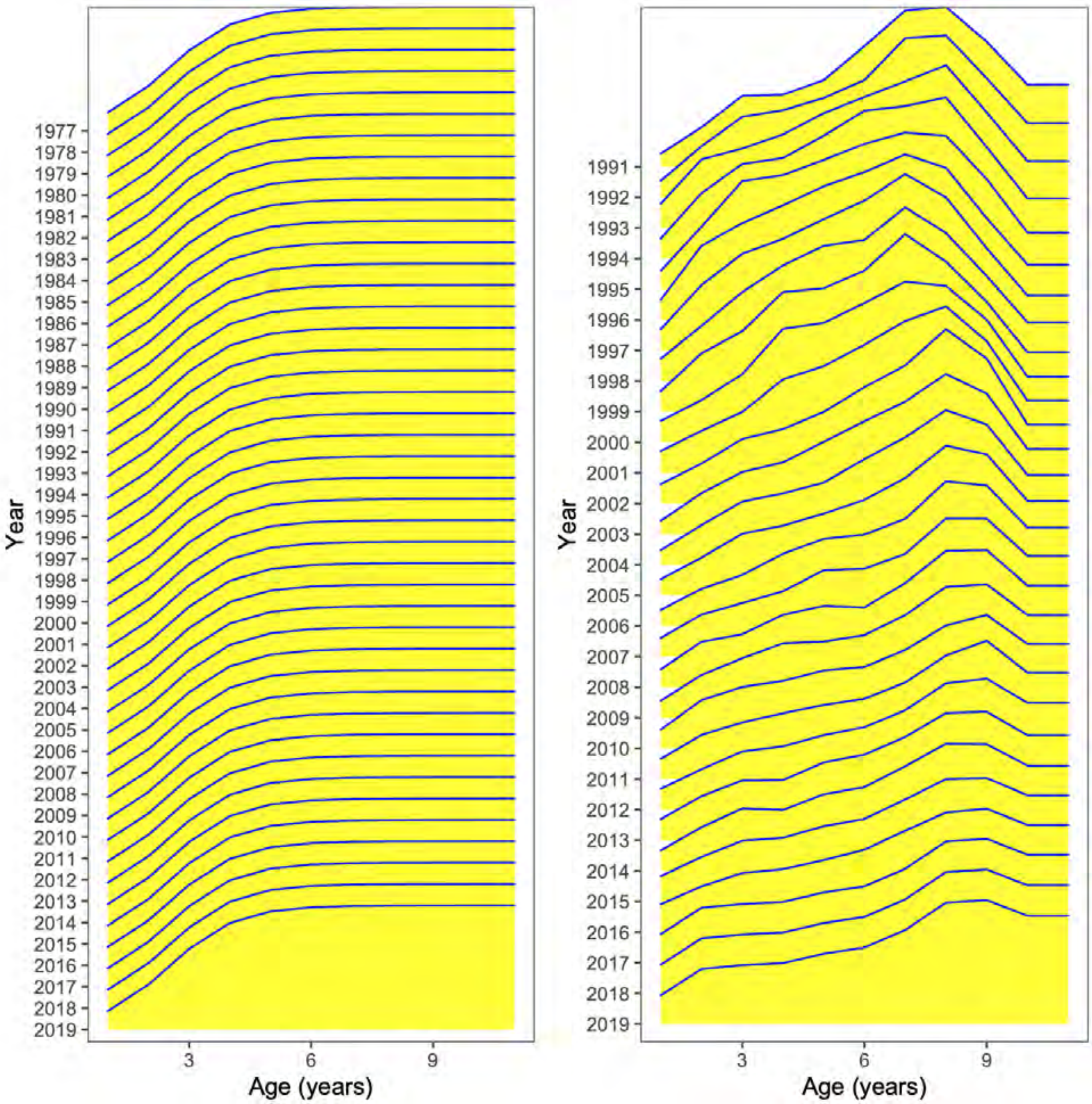


Figure 3. Alternative model runs with constant fishery selectivity (left) and variable (right).

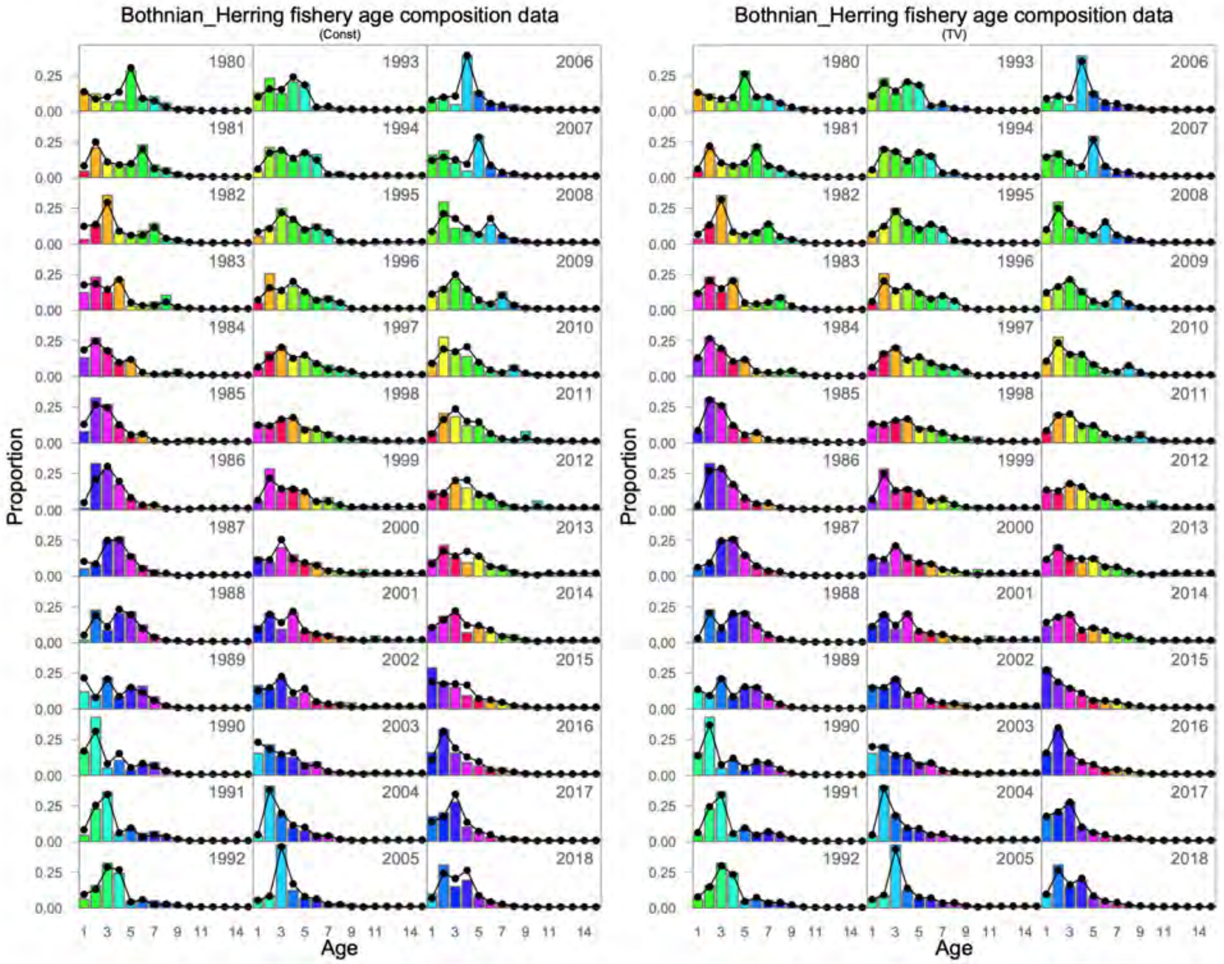


Figure 4. Alternative model fits to fishery age composition data with constant fishery selectivity (left) and variable selectivity (right).

Reference points and forecasts settings

Reference points were estimated for the accepted SS run. The process accomplished ICES guides using EqSim software. Biological parameters (maturity-at-age and weight-at-age) were set as the mean of the last three years. Natural mortality and selection at-age are constant in time in the model, and the same for the reference points. Stock–recruitment data were cut before 1980 because of a lack of age–length distribution before; that makes the S–R data less accurate. Stock–recruitment relationship for 1980–2018 combines Beverton–Holt, Ricker and segmented regression following the EqSim software.

Setting B_{lim}=B_{loss} was considered a reasonably defensible approach given the contrast and variability evident in the stock–recruit relationship. The value of B_{pa} was calculated with sigma derived from the estimated assessment error (estimated CV of SSB in 2019).

The EqSim software was run first without $B_{trigger}$ and then with $B_{trigger}$ resulting in F_{MSY} capped by F_{p05} = 0.189. The reviewers agreed with the process and these decisions.

The WK also discussed the settings for short-term projections for catch advice (time ranges for biological and selection parameters, recruitment ranges for the mean in projected years and F for intermediate year). There were no conflicting issues to highlight, and we agree with the decisions taken.