Extended abstract of presentations

11.1 Biological Parameters (Anders Nielsen)

State-space models are becoming common in assessment and forecast of fish stocks (Stock and Miller 2021, Cadigan 2015, and Nielsen and Berg 2014). In development of such models the focus has naturally been on correct modelling the observed catches from commercial and scientific fleets, which is important, but for management it is equally important to be able to correctly forecast the biological parameters: stock-weights, catch-weights, natural mortalities, and proportions mature in each age group. Age-based assessment models estimate stock-sizes in numbers at age, but management is based on spawning stock biomass (SSB) and on total catch in weight, so information about weights and maturities are needed to produce those required quantities. The standard approach is that within the data period direct observations of the biological parameters are used, and even if these should strictly be considered as observations subject to observation noise, they are treated as known constants. In the forecast period, which is directly used in management, simple ad-hoc rules are applied (e.g.~average of last 5 years), and the resulting weights and maturities are treated as known quantities. This approach was necessary in deterministic and fully parametric models, which do not have the build-in ability to project beyond the observed period, but state-space models should be able to improve this practice.

However, state-space models have largely inherited the practice of using e.g.~simple averages in the forecast period from deterministic and fully parametric models. The practice has possibly been copied because assessment scientists have been accustomed to consider the biological parameters as known covariates and not as quantities observed with observation noise. In state-space models there are potential benefits to consider the biological parameters as observations subject observation noise. The first benefit is that the observation noise would not simply be ignored, but would be allowed to propagate to the estimated quantities of interest (e.g.~SSB). The second benefit is that the same model could be used to predict the needed biological parameters in the forecast period and quantify the uncertainty of the predictions. Hopefully, the predictions from the state-space approach would also be more accurate than the predictions from simple averages.

A study was designed to evaluate and select the best model-based approach for predicting the biological model parameters focused on stock-weights, but could potentially be copied for other biological parameters of interest. 14 stocks were selected to be used as validation data sets. The stocks were selected based on their availability of fairly long data series of stock-weights. The stocks selected were: Northeast Atlantic Blue Whiting and Mackerel; Faroe Haddock and Saithe; North East Arctic Haddock, Saithe, and Cod; North Sea Cod, Haddock, Herring, Plaice, Saithe, Sole, and Whiting. These 14 stocks were selected before the different models were suggested or developed.

The suggested model structures developed were all designed for the purpose of predicting 1-3 years forward. A large number of different suggestions were implemented and subjected to the same validation procedure on all 14 data sets. The suggested model types can broadly be partitioned into : 0) Current practice (5-year average), 1) Gaussian Markov Random Field with optional correlation in age, year, and/or cohort direction. 2) Separable age and year AR(1) structure with or without added cohort effect. These models were further branched out by transformation of observations (none, logarithmic, or Box-Cox), by observation

error, (none, independent, or correlated), and by their use of covariates (none or N). Further some models were restricted to be increasing within each cohort. A total of 34 model structures were implemented, validated, and compared.

Each of the models predict stock-weights in each age group, but it would be difficult to reach a coherent conclusion if different models predicted different age classes better, so to reach a joint conclusion, and to keep the evaluation focused on the real application of this modelling exercise, it was decided to use the ability to predict SSB as the summarizing criteria. For each predicted year the age-specific predicted stock-weights were used to compute SSB (using fixed maturities and N's from the official assessments). This prediction was compared to the SSB calculated by using same maturities, N's, and the observed stock-weights. The comparison was done on logarithmic scale. Based on e.g.~the 10 predictions calculated per stocks the root-mean-square-error (RMSE) was calculated and summarized by simple averages across stocks. 1, 2, and 3, year ahead predictions were compared. Furthermore, it is important that the model structure is robust, so across all runs it was summarized how often the model converged (should be very close to 100%).

To guard against over-fitting to specific observations while developing and testing the following procedure was applied: In the development phase the last 10 years of data were not used in any way. The preceding ten years were used to validate the predictions in the development phase (so e.g.~SSB(Y-19|Y-20),...,SSB(Y-10|Y-11) were used to test one year ahead predictions. Then only after the all the models had been developed the last 10 years of the data were used in a similar way to evaluate and compare the performance of the models.

One model is the GMRF model with cohort and year correlation. The model is a random effects model, where the unobserved true weights by age and year are described by a so-called Gaussian Markov Random Field (GMRF). A GMRF is a stochastic process, where the correlation structure is expresses via the inverse covariance matrix, which is setup via the neighborhood structure. The inverse specification allows for fast computations, because the most time consuming part of evaluating a multivariate Gaussian is the part where the covariance matrix is inverted. The structure here is the structure where weights from neighboring years are correlated within each age group and where weights from neighboring years are correlated within cohorts. Furthermore, mean weights are estimated for each age group (or combination of age groups). The process model for the true log-stock-weights can be written as:

Here the random effects Lambda represent the true log-stock-weights, mu(a) are the age-specific means, sigma is the standard deviation, and phi(1) and phi(2) define the correlation between age groups within the same year and the correlation within each cohort respectively. n(y) and n(d) are the number of neighboring cells a given cell has in the year and diagonal direction respectively (see Fig. 11.1.1). The observed log-stock-weights are assumed to be independent normally distributed:

$$\log SW_{ay} \sim \mathcal{N}(\Lambda_{ay}^{(SW)}, \sigma_{SW}^2)$$

where sigma is the standard deviation for the log-stock-weight observations.

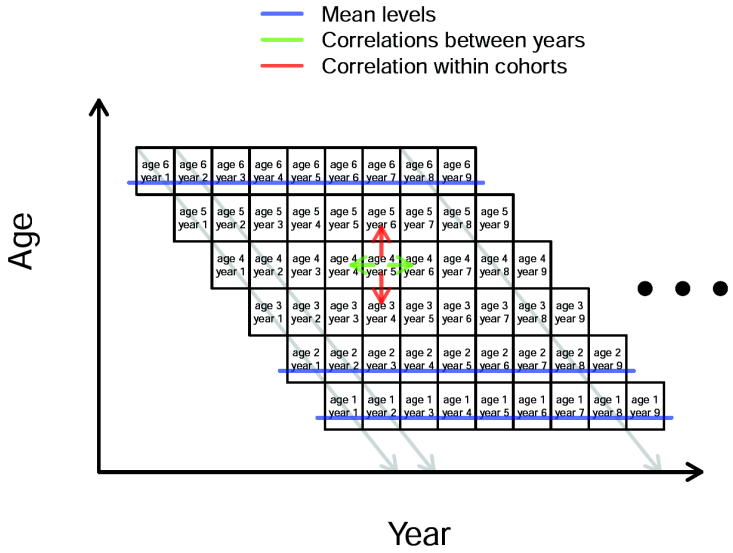

Figure 11.1.1. The structure of the process model, where a mean level is assigned to each age group, and the correlation is across cohort and across years within each age class.

This model can be illustrated as a spatial model where neighbouring cells are correlated (with separate correlations in the directions of the two axis) if the yearly observation vectors are shifted such that the cohorts line up vertically (Fig. 11.1.1.)

The model structure selected for stock-weights was the GMRF with correlations across ages within year and correlation within cohorts. This model is described in details above. The log-scale one year ahead root mean square error (RMSE) of the selected model structure is 0.068 compared to 0.086 for the standard approach of using average over the last 5 years. This is a reduction of more than 20% averaged over 14 stocks (Table 11.1.1).

Model	RMSE(1)	RMSE(2)	RMSE(3)	Jit	Convergence
5-year average	0.086	0.101	0.118	0.00	100%
GMRF age, cohort	0.068	0.088	0.102	0.01	100%
GMRF age, cohort, no obs noise	0.068	0.087	0.102	0.00	100%
ARxAR, cohort, increasing	0.087	0.115	0.140	0.81	100%
GMRF increasing	0.089	0.115	0.139	0.58	98%
ARxAR, no cohort, increasing	0.069	0.093	0.111	0.20	100%

Table 11.1.1: Evaluation results for some of the best performing models.

The aim was to find a solution which could be used as an alternative to the current practice (e.g.~average of most recent five years data), so it was also important to validate that the proposed models were robust w.r.t. convergence. To validate the convergence it was recorded what fraction of the cross-validation runs converged and the maximum difference in the estimated parameter values when starting from random initial values (``jit'' in table 11.1.1). A lot of the suggested models did not pass this investigation.

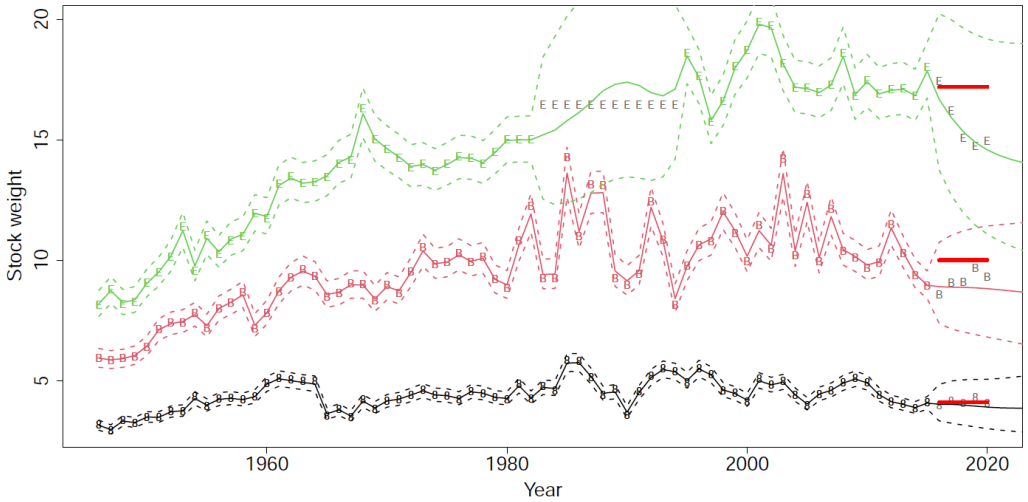

Figure 11.1.2: Observed mean weights in stock for North East Arctic Cod for selected ages 8 (symbol '8'), 11 (symbol 'B'), and 14 (symbol 'E'). The model predictions and 95% confidence intervals (thin solid and dashed lines) and the predictions using the recent 5 year average (thick red lines). The observations used to estimate the model (full color symbols) and the observations not used to estimate the model (gray symbols).

The selected model is a clear improvement for North East Arctic Cod stock weights against simply using the average of the most recent 5 years (Fig. 11.1.2). The model was run using all age groups (3-15+), but without using any observations from the last five years (2016-2020). It was also noticed that for the older age groups (12-15+) in the years 1983-1994 constant values had been used. The constant values were not considered real observations, so they were also omitted. The model provide a prediction where the observations were omitted and in the last five years the prediction is much closer to the omited observations than the average of the five most recent years. Notice also that the model provides predictions (which appear reasonable) for age 14 in the 1983-1994 period. In addition to the predictions the model also supply uncertainty estimates (here illustrated by 95% confidence intervals), and those increase as the model predicts further away from the observed period.

Cadigan, N., 2015. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Canadian Journal of Fisheries and Aquatic Sciences 73.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research 158.

Stock, B.C., Miller, T.J., 2021. The Woods Hole assessment model (wham): A general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental co-variates. Fisheries Research 240.

11.2 One Step Ahead Residuals (Anders Nielsen)

Stock assessment models are often used to inform fisheries management and need therefore to be thoroughly validated. Different diagnostics exist to validate models including the analysis of standardized residuals. Standardized residuals are commonly calculated by subtracting prediction from the observation and dividing the result with the estimated standard deviation (i.e., Pearson residuals). Many

currently applied stock assessment models fit to compositional observations (e.g.,~age, length or stock compositions) using multivariate distributions. These distributions create correlation between observations, which are propagated in the residuals if estimated as Pearson. This study shows that using Pearson residuals to analyze goodness of the fit, when data are fitted using a multivariate distribution, is incorrect and one-step-ahead (OSA) or forecast quantile residuals should be used instead. For such distributions, OSA residuals are independent and standard normally distributed for correctly specified models. This study describes the calculation of OSA residuals specifically to de-correlate compositional observations for the multivariate distributions most commonly used in assessment models (multinomial, Dirichlet, and Dirichlet-multinomial). This allows composition observations to be evaluated with the same statistical rigor as residuals from uncorrelated observations. This also prevents the possible wrong interpretation of Pearson residuals and the rejection of a correct model. We have developed an R-package that estimates OSA residuals externally to the model for models that do not include random processes. For models that use random processes, the distributions are now developed in Template Model Builder and explained in detail here for internal use.

The implementation for use with TMB is available at: **https://github.com/vtrijoulet/OSA_multivariate_dists/** and the R package for external use in fixed effects models is available at **https://github.com/fishfollower/compResidual.**

Method:

In univariate statistical models where the observation noise is assumed independent and normally distributed the standardized residual r_i is often calculated by subtracting the models estimated expectation from the observation and dividing the result with the estimated standard deviation. r=(x-pred)/sd. If the model is correct this will approximately lead to residuals which follow a standard normal distribution. Hereby are defined residuals with the property that they will follow a standard normal distribution if the model is correct, but deviate systematically from a standard normal distribution if the model is incorrect. Plotting such residuals are useful for validating if a model is describing the observed data.

If observations $x_1,...,x_n$ are continuous, univariate, and independent, but originating from a distribution of X which is not a normal distribution, but has cumulative distribution function (cdf) F_x, then residuals with the same properties as above can be obtained via transformation. First notice that transforming the observations via the cdf will lead to quantities which follow a uniform distribution $u_i = F_x(x_i)$ (This can be see by noting that u_i in]0,1[and that the cdf of u is $F_u(u) = P(F_x(X)<u) = P(X<F^{-1}_x(u)) = F_x(F^{-1}_x(u)) = u$, which is the distribution function for the uniform distribution). Next notice that transforming these uniformly distributed quantities $u_1,...,u_n$ by the inverse cdf of the standard normal distribution Phi^{-1} will lead to residuals which follow a standard normal distribution if the model is correct (This can be seen by calculating the cdf for the transformed $P(Phi^{-1}(U)<r) = P(U<Phi(r)) = Phi(r)$, so the wanted cdf). Collectively these quantile residuals are defined simply as: $r_i = Phi^{-1}(F_x(x_i))$. Notice that the model is defining the cdf of the observations, so if the model is incorrect, then the residuals will deviate systematically from a standard normal distribution.

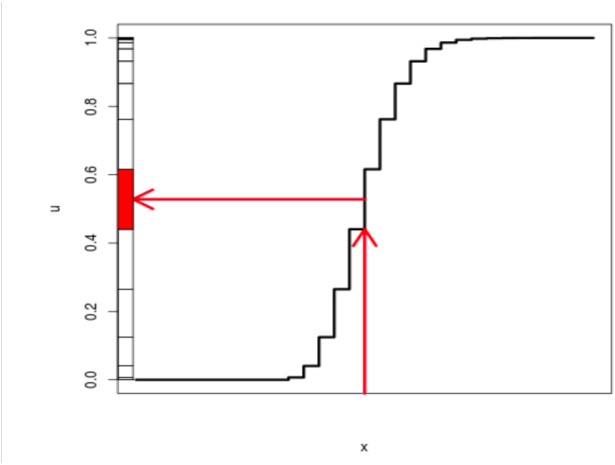

Figure 11.2.1: The distribution function (cdf) of a discrete distribution requites a uniform sampling step in the interval marked red on the y-axis.

If the observations $x_1,...,x_n$ are discrete, but still univariate and independent, then the distribution function F_x is a step function (Figure 11.2.1). In this discrete case the transformation by the distribution function needs an additional step. The probability mass at a given value x_i (e.g. at the value of the red vertical arrow) need to be transformed onto the interval from $F(x_i\text{-epsilon})$ to $F(x_i)$. The transformation into uniform(0,1) distributed quantities can be achieved by sampling from the uniform distribution for that interval, so $u_i \sim U(F(x_i\text{-epsilon}),F(x\_i))$. The final step to get standard normal residuals is the same transformation by the inverse cdf of the standard normal distribution $r_i = Phi^{-1}(u_i)$. These residuals will again have the desired properties.

If observations are not independent, then independent residuals can be obtained by the one-observation-ahead residuals. The residual of the i'th observation is computed using the techniques outlined above, but instead of using the cdf of the observation in isolation, the cdf of the predicted distribution if the i'th conditioned on all previous observations. This allows the resulting residuals to become independent standard normal if the model is correct.

As an example a vector of observations $(x_1,...,x_m)$ follows a multinomial distribution $(x_1,...,x_m) \sim Multi(N,(p_1,...,p_m))$ it is equivalent to express the distribution as successive binomials, as in: $x_1 \sim Bin(N,p_1)$, $x_2 \sim Bin(N\text{-}x_1,p_2/(1\text{-}p_1))$, ..., $x_{m-1} \sim Bin(N\text{-}(x_1+...+x_{m-2}),p_2/(1\text{-}(p_1+...+p_{m-2})))$. So the likelihood of the multinomial can be evaluated as a product of the successive binomials. Expressing the distribution via the successive binomials allows easy access to the predictive distributions needed.

The other distributions commonly used for composition observations can be resolved in similar ways to make the one-step-ahead calculation possible.

Further details of this work can be found in:

Trijoulet, V, Albertsen, CM, Kristensen, K, Legault, CM, Miller, TJ, Nielsen, A (2023). Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. Fisheries Research 257.

11.3 Fishery Integrated Modeling System (FIMS) (Andrea Havron)

The U.S. fisheries stock assessment community has relied on a suite of regionally-developed, individually-maintained software tools to assess 100+ fish stocks annually. These methods are accurate (Li et al.[1]), but difficult to maintain and do not facilitate connection to ecosystem, climate and socioeconomic models. There has been a lag in adoption of new best practices (e.g., random effects, spatial models, ensemble modeling, multi-species, hierarchical Bayes estimation) due to the rigidity of current models and capacity for adding new features. Additionally, separate regional approaches create the unnecessary duplication of effort, and potential for "hidden" differences between similar models (Li et al.). A number of parallel developments make now the ideal time to invest in a Fisheries Integrated Modeling System (FIMS). These developments include advancements in computing techniques, such as cloud computing, version control, parallel processing, and open-source development. Within NMFS, a centralized suite of tools (the Fisheries Integrated Toolbox; FIT[2]) and associated steering committees are increasing interdisciplinary collaboration in tool development. There is growing regional and international interest and effort in coordinating, advancing, and reducing duplication in modeling approaches, as evidenced by the 2019 Center for the Advancement of Population Assessment Methodology (CAPAM) Workshop on Next Generation Models (Punt et al, 2020[3]). Longer-term, FIMS is a stepping stone to facilitate increased interfacing between fish assessment community and ecosystem, climate-fisheries, and socioeconomic communities, and associated models.

Through FIMS, NMFS is leading the development of a modular software system that presents a unified approach to fisheries modeling, meets regional needs for conducting stock assessments and providing scientific advice across a range of data input, provides a focus for greater international collaboration, and builds a bridge between single-species, ecosystem, socioeconomic, and climate-fisheries models. FIMS will scale from data-limited to data-rich stock assessments, and provide population dynamics modules for use in other models and management strategy evaluations (MSEs). This FIMS effort will incorporate models currently in FIT with regionally specific models being developed, and then identify new modeling requirements that would benefit from national coordination. Use of a modular architecture and best practices for software development facilitates maintainability, extensibility, and continuity across a virtual, nationwide development team. These engineering approaches also allow FIMS to leverage cloud-computing infrastructure.

By managing FIMS centrally within the NMFS Office of Science & Technology (OST) and incorporating modular components developed regionally, we provide a well-designed system that meets the variety of needs across regional science centers, is extensible enough to assimilate new research quickly, and that can provide functionality to multiple regions without dependence on individual developers and scientists. A modular, flexible system allows us to incorporate emerging best practices in ensemble modeling, new model types, MSEs, and to adapt to future best practices. The FIMS initiative reduces duplication of effort, enhances stability, provides resources to develop sustainable architecture, supports regional stakeholders, and provides a scalable and accessible tool that meets a variety of needs and can familiarize staff with model capabilities and good practices. Overall, FIMS provides a deliberate national research to operations pathway.

Please see the Governance section of the FIMS Developer Handbook to read the full FIMS TOR[4].

[1] Bai Li, Kyle W. Shertzer, Patrick D. Lynch, James N. Ianelli, Christopher M. Legault, Erik H. Williams, Richard D. Methot Jr, Elizabeth N. Brooks, Jonathan J. Deroba, Aaron M. Berger, Skyler R. Sagarese, Jon K.T. Brodziak, Ian G. Taylor, Melissa A. Karp, Chantel R. Wetzel, and Matthew Supernaw. A comparison of four primary age-structured stock assessment models used in the United States. In review. Fishery Bulletin.

[2] https://noaa-fisheries-integrated-toolbox.github.io/

[3] André E. Punt, Alistair Dunn, Bjarki Þór Elvarsson, John Hampton, Simon D. Hoyle, Mark N. Maunder, Richard D. Methot, Anders Nielsen. 2020. Essential features of the next-generation integrated fisheries stock assessment package: A perspective. Fisheries Research 229. https://doi.org/10.1016/j.fishres.2020.105617.

[4]FIMS Implementation Team. 2022. "3.1 FIMS Governance." FIMS Developers Handbook. https://noaa-fims.github.io/collaborative_workflow/fims-project-overview.html#fims-governance

11.4 Multi-stock extension of SAM and reference points (Christoffer Albertsen)

Marine stocks are predominantly assessed using single-stock models assuming closed populations. However, many marine stocks do not live in isolation. Stocks from the same species often live in the same management areas and are found together in commercial catch and surveys. For example, North Sea autumn spawning (NSAS) and western Baltic spring spawning (WBSS) herring are both found in are 3.a and part of the North Sea; three cod stocks are mixing in the waters west of Greenland; Eastern and Western Baltic cod co-occupy sub division 24; and the North Sea is inhabited by a southern stock, a northwestern stock as well as the Viking stock. In the North Sea, data to split commercial catches and the quarter 3 IBTS survey is not available.

Recently, Albertsen et al. (2018) developed a multi-stock extension of the SAM model that could combine assessments through correlation in the abundance processes. Such correlation can for example occur when stocks have a predator-prey relation, compete for resources, or share environmental conditions. Building on that approach, this presentation introduces a multi-stock multi-fleet version of the SAM model that allow for observations with mixtures of stocks.

In the SAM model, the fishing mortality rate process, F, is modelled by a multivariate random walk process. This provides good reconstructions of historical fishing mortality rates with flexibility to have time-varying selectivity and average fishing mortality rates. In cases where observations are only available as a sum over stocks, there may not be sufficient information to estimate full random walk processes for all stocks. To account for this, three options were implemented in the multi-stock model.

In all three options, one stock retains a full random walk process for the fishing mortality rate. In the first option, the F process of the first stock is scaled by individual univariate random walk processes for each of the other stocks. Thereby, each stock is assumed to have the same selectivity but different average fishing mortality rates. In the second option, the F process of the first stock is scaled by a parametric function of, for example, age and year. The scaling is similar to a time-varying proportional hazards model from survival analysis. This model allows

time-varying selectivity and average fishing mortality rates with the smoothness constraints imposed by using a parametric function. The third option is to include stock composition data in the model.

When including stock composition data (e.g., genotype, phenotype, otolith shape, otolith microchemistry, vertebrae counts), both baseline data af known stock origin as well as samples from catches (either commercial or survey) must be included. Baseline data would typically be samples of spawning individuals that define the population.

Depending on the type of data, X_i, a model can be specified for the chance of a given observation for each stock, $P(X_i) \mid S_i = s)$. For example, this could be a product of multinomials for genotype data, a multivariate normal for microchemistry, or a log-normal for weight-at-length data. Baseline data are used directly in the model to estimate the parameters of $P(X_i) \mid S_i = s)$. For mixed samples, the true stock origin is unknown. Therefore, the chance of a given observation, $(P(X_i)$ is a weighted sum of the stock-wise chances,

$$P(X_i) = \sum_s P\ (X_i \mid S_i = s)P(S_i = s).$$

The sum is weighted by the probability of originating from each stock, which can be calculated from the predicted catch in the assessment model,

$$P(S_i = s) = \frac{C_s}{C_{total}}$$

For most species, fish from the same haul, trip, or sampling, are more similar than fish from different hauls. Therefore, the model can include random effects in $P(S_i = s)$ to account for this correlation. Further, random effects can be included to account for spatio-temporal correlation. Finally, the model can include a conversion matrix for cases where baseline stocks does not map one-to.one to the assessed stocks. For example, a multi-stock model for NSAS and WBSS herring could include a genetic baseline that also included Norwegian spring spawners, Downs, Baltic autumn spawners, and central Baltic herring. In contrast, Southern and Northwestern North Sea cod are genetically similar. Therefore, a multi-stock model for Southern, Northwestern, and Viking cod could include a genetic baseline with Dogger and Viking cod.

Further, a fourth option is being explored. The fourth option aims to make use of spatial quarterly landings in the model. Assuming the landings in a specific quarter can be spatially linked to a stock (i.e., the mixing is limited), the fraction of total landings in that quarter by area can be compared to the fraction of total landings in that quarter by stock in the model.

Recently, Albertsen and Trijoulet (2020) outlined a method for estimating deterministic reference points, assuming no process noise, with confidence intervals in assessment models. The method has been implemented in both the single- and multi-stock versions of the SAM model. The approach was tested and compared to eqsim by Trijoulet et al. (2022).

In this presentation, a new functionality for estimating simulation based stochastic reference points, assuming process noise in the system, in the SAM model is presented. To calculate stochastic reference points, a number of average fishing mortality values are simulated. In turn, each F value is used to do a simulation forecast of the fitted model to equilibrium and a reference point criterion is calculated. For estimating Fmsy, the criterion would be yield. Finally, a curve is

fitted to the simulated values for which the optimum can be found. For curve fitting, two options are available. To optimize mean values, maximum likelihood is used. To optimize median values, quantile regression is used. Likewise, any quantile of interest can be optimized. Confidence intervals are calculated by parametric bootstrap simulations. An implementation for the multi-stock version is in progress.

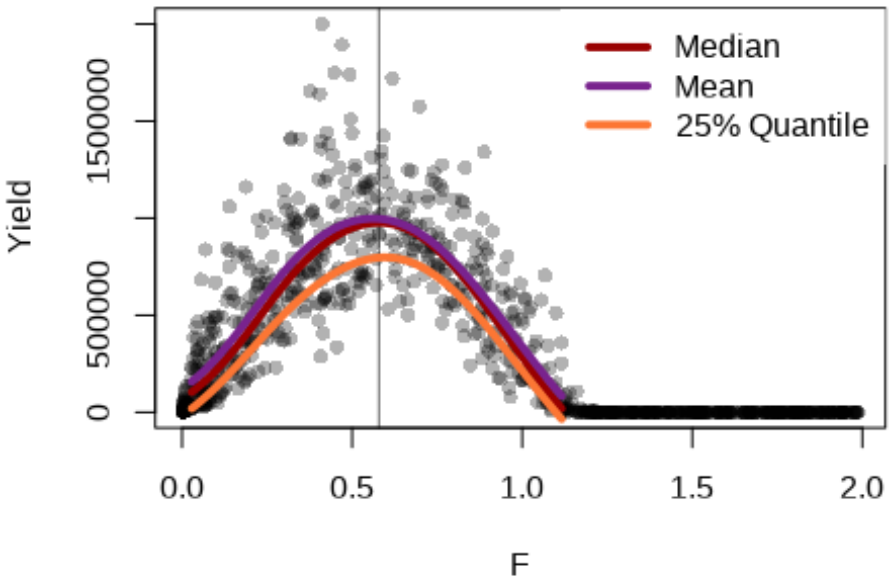

Figure 11.4.1. Example of the curve fitting for estimating stochastic reference points in SAM. Each grey point is a simulated equilibrium yield for a given F. Lines are fitted median (red), mean (purple), and 25% quantile (orange) yield as a function of F.

To validate the approaches for mixed observations, a simulation study with two stocks was conducted. One stock was based on the last North Sea cod benchmark (the stock nscod_bench21_02 on stockassessment.org) but was modified to have Beverton-Holt recruitment. The slope at origin was set to 20 while the maximum recruitment was set to 1.3000.000. The second stock was similar to the first, but was set to have 30% higher natural mortality and 20% lower weight-at-age. Further, both the recruitment curve slope at origin and maximum recruitment was set to 30% of the first stock.

A total of four scenarios were used for the simulations. In all four scenarios, the F process of the first stock was fixed to the fitted value. In the first scenario, the F process of the second stock was scaled by a univariate ARMA(1,1) process. In the second scenario, the scaling was a log-linear function of age and year. In the third scenario, the F processes were identical. In the fourth scenario the F process was scaled by a multivariate ARMA(1,1) process.

For each scenario, 300 data sets were simulated for the two stocks with fixed F processes. For each data set, four model approaches were used as assessments. In the first model approach, the full, non-mixed, data was used to fit two single stock

assessment models. While this is not possible in a situation with mixed observations, it was used as an optimal base case to compare against. In the second model approach, all observations and biological input were combined for the two stocks and a single assessment model was fitted. This corresponds to assessing the two stocks as one. In the third and fourth model approaches, commercial catches and quarter three surveys were combined for the two stocks while quarter one and recruitment surveys were retained for each stock. This corresponds to a situation where stocks are separated at spawning time in quarter one and mixing in the rest of the year. The third model approach used the scaled F by univariate random walk approach while the fourth used proportional hazards.

In all four scenarios, the full information assessments performed best, closely followed by the scaled F and proportional hazard approaches. Both the scaled F and proportional hazard models could adequately reconstruct the stock-wise F processes and SSB. However, for some simulations the level of average fishing mortality and SSB was wrong, while trends were followed. In contrast, the combined assessment model estimated wrong trends in F, especially for the smallest stock. While the combined assessment model could not output stock-wise SSB, the total SSB was estimated well in all scenarios.

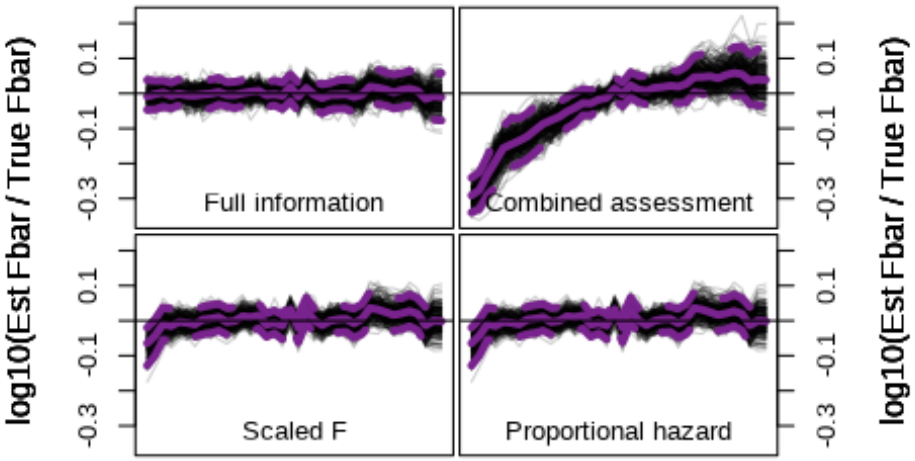

Figure 11.4.2. Log10 difference in estimated and true simulated average fishing mortalities for the second scenario when estimating using the full (not mixed) information (top left panel), using a combined single stock assessment (top right panel), using the scaled F approach (bottom left panel), and using the proportional hazard scaling (bottom right panel).

Besides F and SSB, deterministic Fmsy was calculated for each simulation and model approach. In all four scenarios and for both stocks, the distributions of Fmsy were similar for the full information, scaled F, and proportional hazard approaches. Further, the estimated Fmsy values were close to the true value with a tendency of slightly underestimating the value. In contrast, the combined single-

stock assessment consistently overestimated Fmsy. For the smallest stock, with the lowest true Fmsy, the average estimated Fmsy was up to twice the true value.

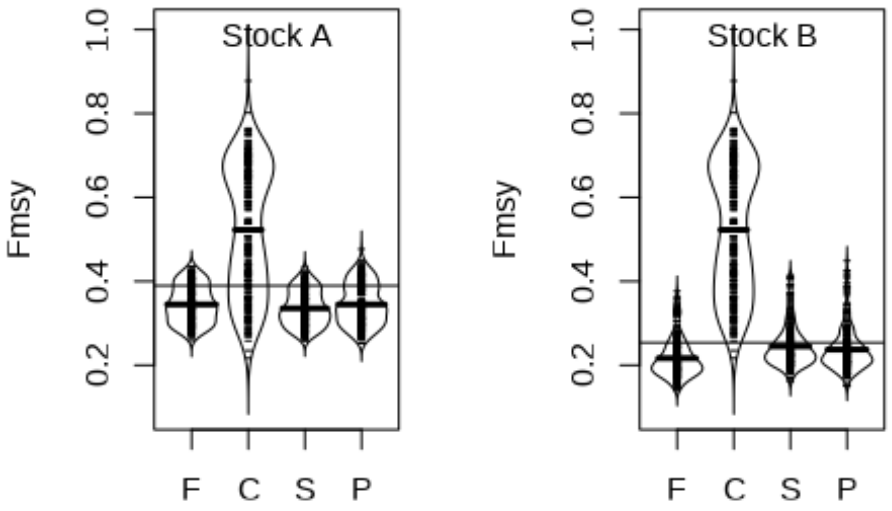

Figure 11.4.3. Bean plot of estimated Fmsy in scenario two for the full information (F), combined single-stock assessment (C), scaled F (S), and proportional hazard (P) model approaches. Horizontal black lines show the true value.

Albertsen, C. M., Nielsen, A. and Thygesen, U. H. (2018) Connecting single-stock assessment models through correlated survival. ICES Journal of Marine Science, 75(1), 235-244. doi: 10.1093/icesjms/fsx114

Albertsen, C. M. and Trijoulet, V. (2020) Model-based estimates of reference points in an age-based state-space stock assessment model. Fisheries Research, 230, 105618. doi: 10.1016/j.fishres.2020.105618

Trijoulet, V., Berg, C. W., Miller, D. C. M., Nielsen, A., Rindorf, A. and Albertsen, C. M. (2022) Turning reference points inside out: comparing MSY reference points estimated inside and outside the assessment model. ICES Journal of Marine Science, 79(4), 1232-1244. doi: 10.1093/icesjms/fsac047

11.5 Catalog stock assessment settings (Colin Millar)

The design of the ICES TAF system has been to store stock assessment source data, code for preparing data, running assessments and producing results, and ultimately to store the results from stock assessments and make them easily available. Recently developments have been made to store stock assessment results in a simple database to allow upload and download of results and allow filtering to be done to extract results of interest. Previous attempts at this had been overly complicated, but a simplifying decision was to contstuct a database which serves as a read only source of stock assessment model runs which are submitted

by the stock assessor, eventually via TAF, but in the first instance via an upload API. The structure of the DB is given in figure 11.5.1

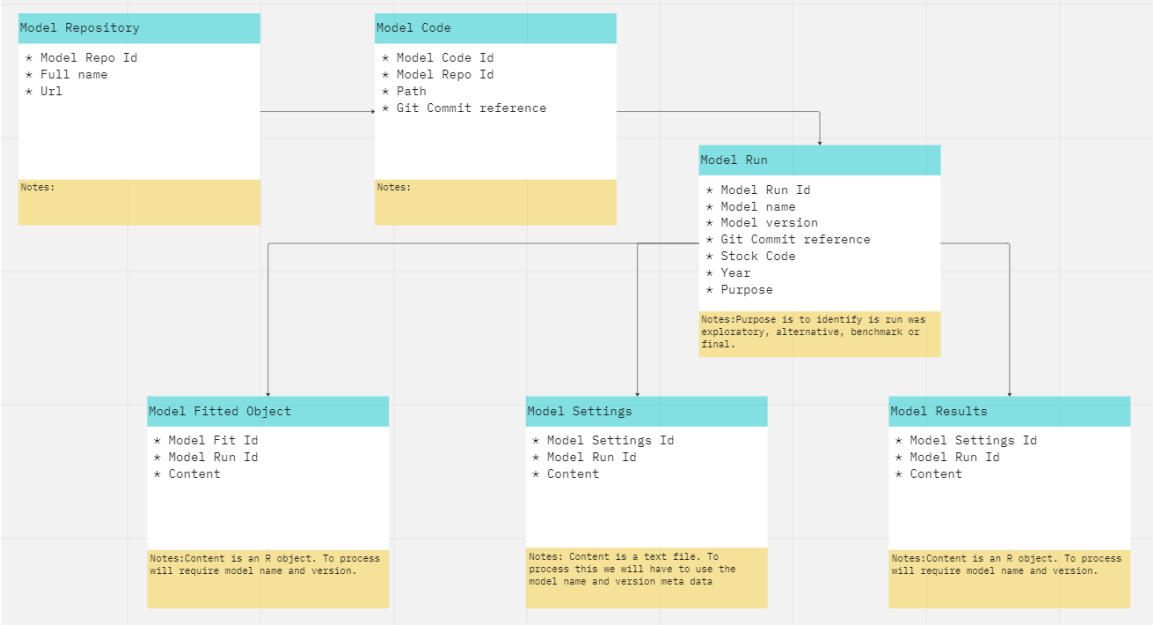

Figure 11.5.1 Design of the wider database to store assessment settings in addition to fitted model objects and model results.

Some technical details are that the content fields are stored as text as base 64 encoded binary serialisation of an R object. A test database has been created and some draft web services produced to allow upload and download of selected aspects of the database (Figure 11.5.2). These services have a single upload point where a user can upload a model run, optionally including a fitted model object, results and model settings. Links to the code and if present a github repository are not implemented yet. A core piece of information is the model name and model version. This is because the content that is stored is model specific, and so after extraction of the configuration files (for example) a model specific R script will need to be developed to produce an output that can be compared across models. The decision to store model specific content was made in order to be as flexible as possible, and to allow the model specific scripts to be tailored to specific questions if needed, for example, to produce a summary of recruitment modelling settings.

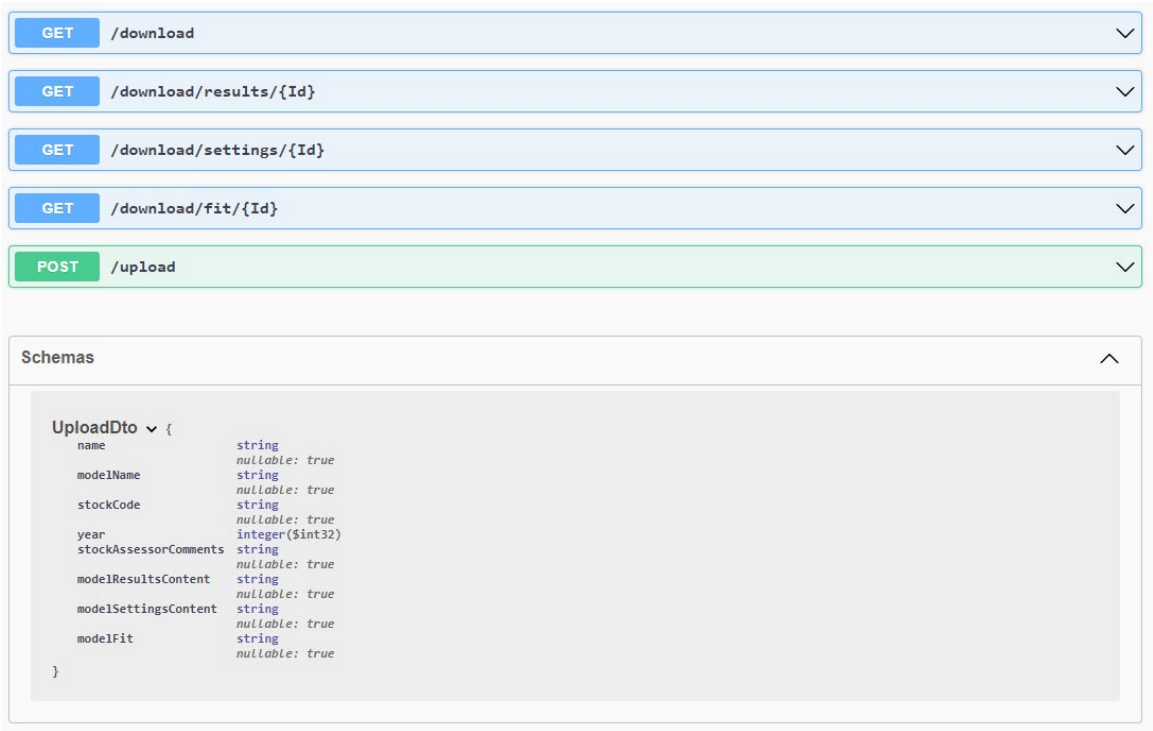

Figure 11.5.2 Web interface to the draft web services to access the catalogue of assessment settings.

As a test of the structure and services, 8 stock assessment model runs from 2020 were extracted from stockassessment.org and uploaded to the database, and then extracted and processed in an R script to produce the table below.

Table 11.5.1 Table of selected outputs taken from SAM configuration files and fitted object.

stock_code	minimum_age	sr_model	recruitment_sd
cod.27.47d20	1	plain random walk	0.781
cod.27.6a	1	plain random walk	0.883
cod.27.7e-k	1	plain random walk	0.977
had.27.7b-k	0	plain random walk	1.154
ple.27.21-23	1	plain random walk	0.356
ple.27.24-32	1	plain random walk	0.676
whg.27.47d	0	plain random walk	0.309
whg.27.7b-ce-k	0	plain random walk	0.436

11.6 Stocks not managed by EU or a country (Laurie Kell)

An example of a Productivity Susceptibility Analysis (PSA) was presented that is being conducted for Areas Beyond National Jurisdiction (e.g. by the Sargasso Sea Commission) following a DPSIR exercise to identify pressures on different ecosystem components by multiple sectors (e.g. fishing, shipping, climate change, pollution). The PSA also includes uncertainty so as data and knowledge improve the benefits can be identified.

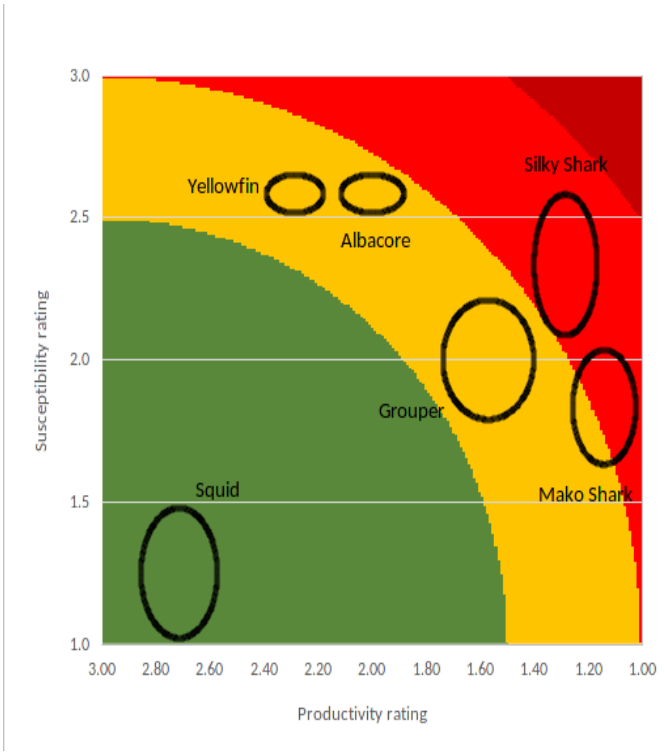

Figure 11.6.1. Example of a productivity susceptibility analysis showing uncertainty in both ratings for a number of species.

11.7 A comparison of length based estimators and indicators (Laurie Kell)

Lmean the mean size of individuals greater than Lc (e.g. the length at first capture), is used by ICES as an indicator of mortality. Figure 11.7.1 plots 1/Lmean (since the smaller the mean size the higher is mortality) against F for cod and whiting by 1st and 3rd quarter for different values of Lc. While Figure 11.7.2 plots the corresponding ROC curves and the AUC to allow the best indicator to be chosen.

Another example is shown in Figure 11.7.3 based on an Operating Model for sprat for three potential indicators a) biomass index, b) length-based indicator, and c) relative harvest rate (catch/index). The lines correspond to years after the change in F. Relative harvest rate has the best classification skill, while the length based indicator takes several years before a change in the length distribution is seen and F can be estimated. This lag will make it difficult to use the length-based indicator in a feedback control rule, e.g. as part of the rfb rule (Fischer et al. 2021). This shows how a ROC curve can be used to select and weight the components of empirical HCRs reducing the need for tuning when conducting an MSE. It can also be used

to evaluate the ability of the 2 over 3 and other rules before conducting full feedback.

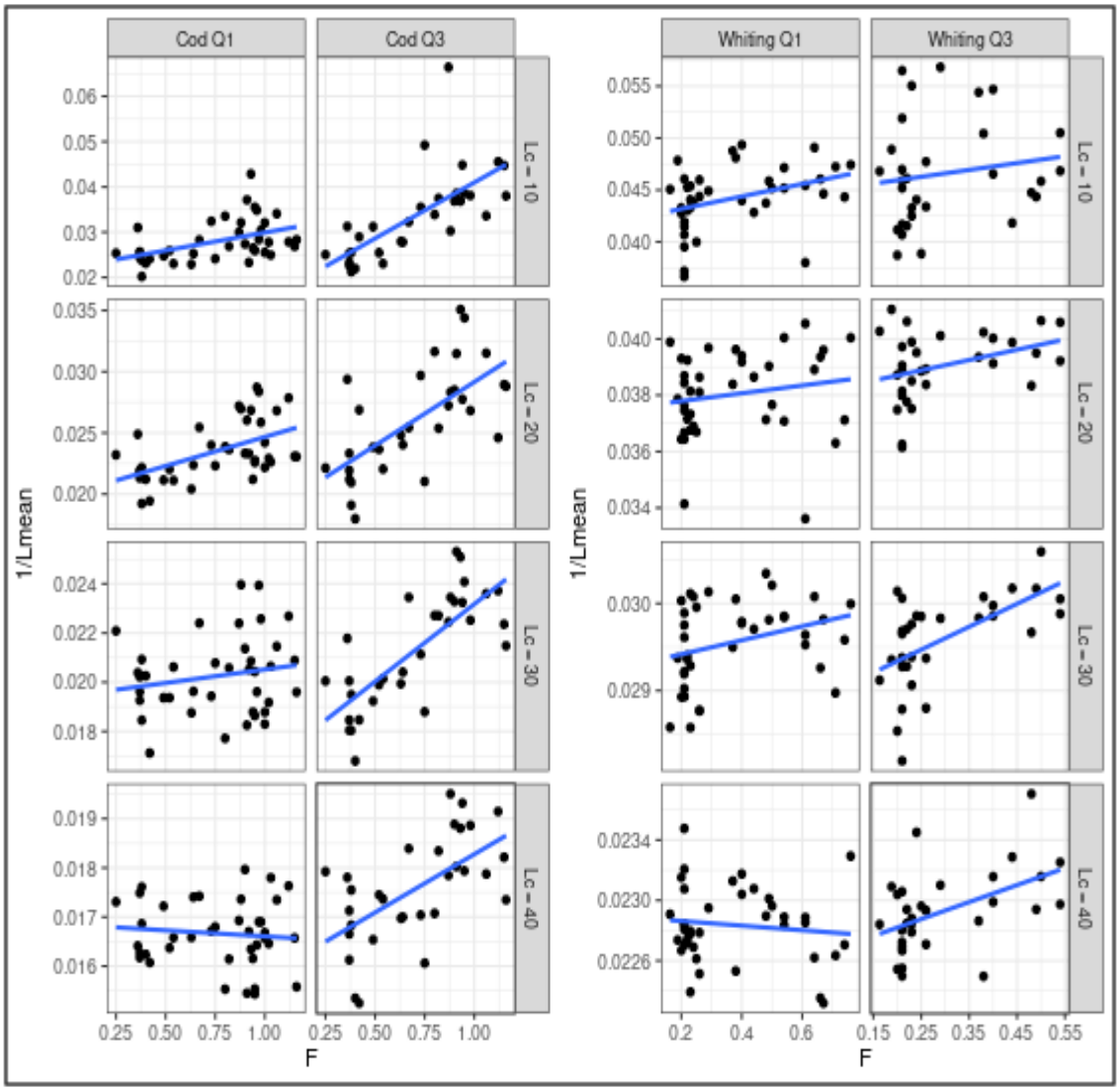

Figure 11.7.1. Plots of the length based indicator 1/Lmean against F for cod and whiting by 1st and 3rd quarter for different values of length at first capture (Lc).

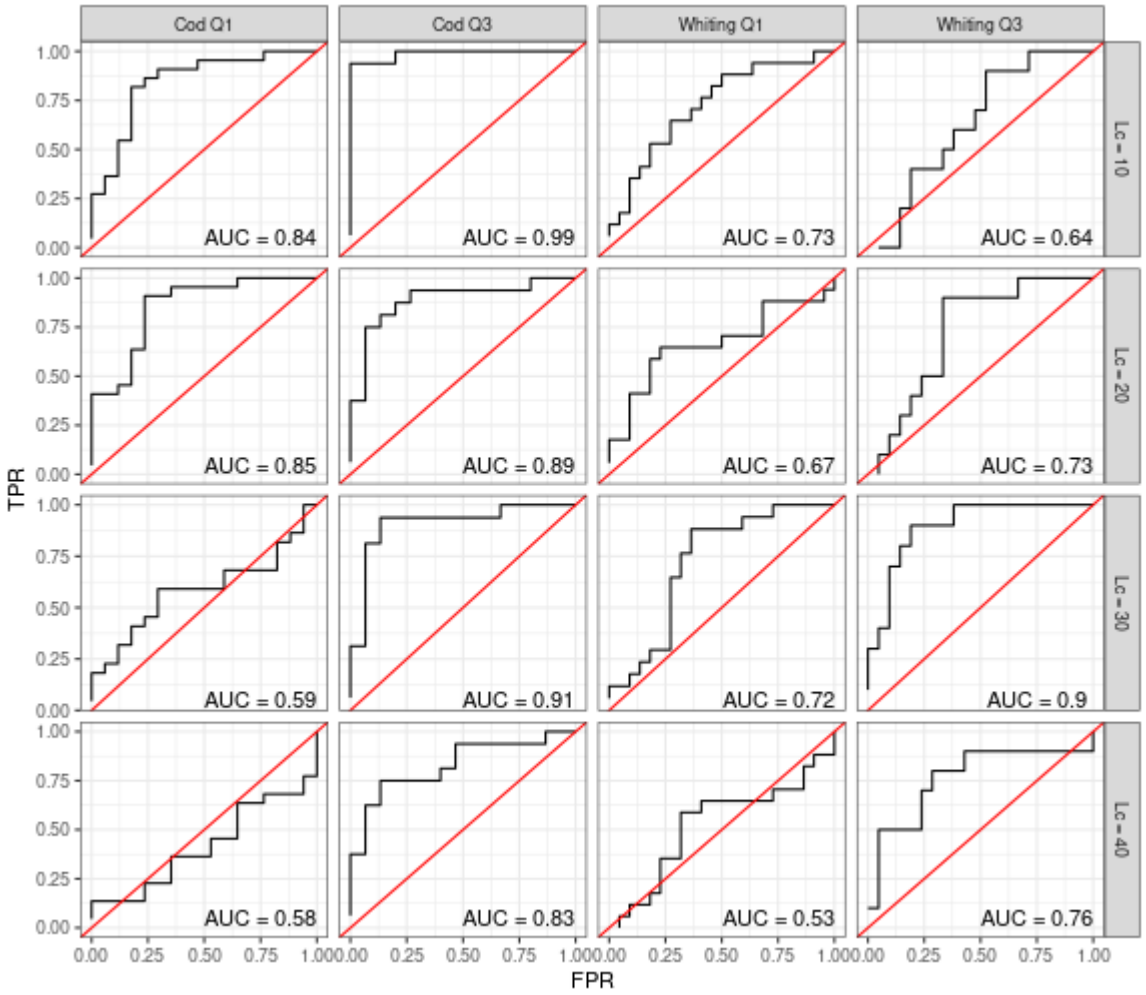

Figure 11.7.2. Receiver Operator Characteristic curves for the 1/Lmean.

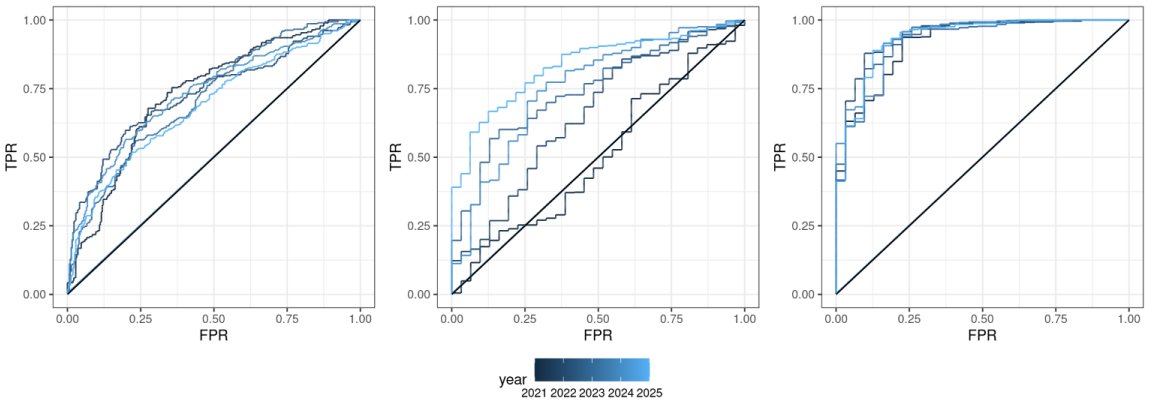

Figure 11.7.3. Receiver Operator Characteristics curves for a) biomass index, b) length-based indicator, and c) harvest rate. The lines correspond to years after a change in F.

11.8 Celtic Sea sprat (Laurie Kell)

Sprat is a target species in the Celtic Sea; however, current harvest advice is based on landings, and there is insufficient information to estimate stock status, trends, or target and limit reference points. As well as being a valuable commercial species, sprat are a major predator on zooplankton, an abundant prey for piscivorous fish and a competitor for herring. Ensuring the sustainable exploitation of sprat is therefore important for both the health of the Celtic Sea's marine ecosystem and for the wider fisheries sector in Ireland. To develop robust advice, we conduct Management Strategy Evaluation using a multi-stock single species sprat operating model conditioned on life-history theory and linked to an Ecosystem Model of Intermediate Complexity. The study shows how ecosystem understanding can be incorporated within the existing precautionary and maximum sustainable yield frameworks. This approach also allows environmental drivers, such as sea surface temperature, to more complex emergent food web indicators to be simulated and the benefits of alternative harvest control rules to be evaluated.

Some preliminary results are shown in Figure 11.8.1. As well as performance metrics for state and yield, forage shows the food availability for predators for different fishing mortalities. A objective is to stress test the current ICES PA and MSY advice to see if EBFM objectives can met "unaided".

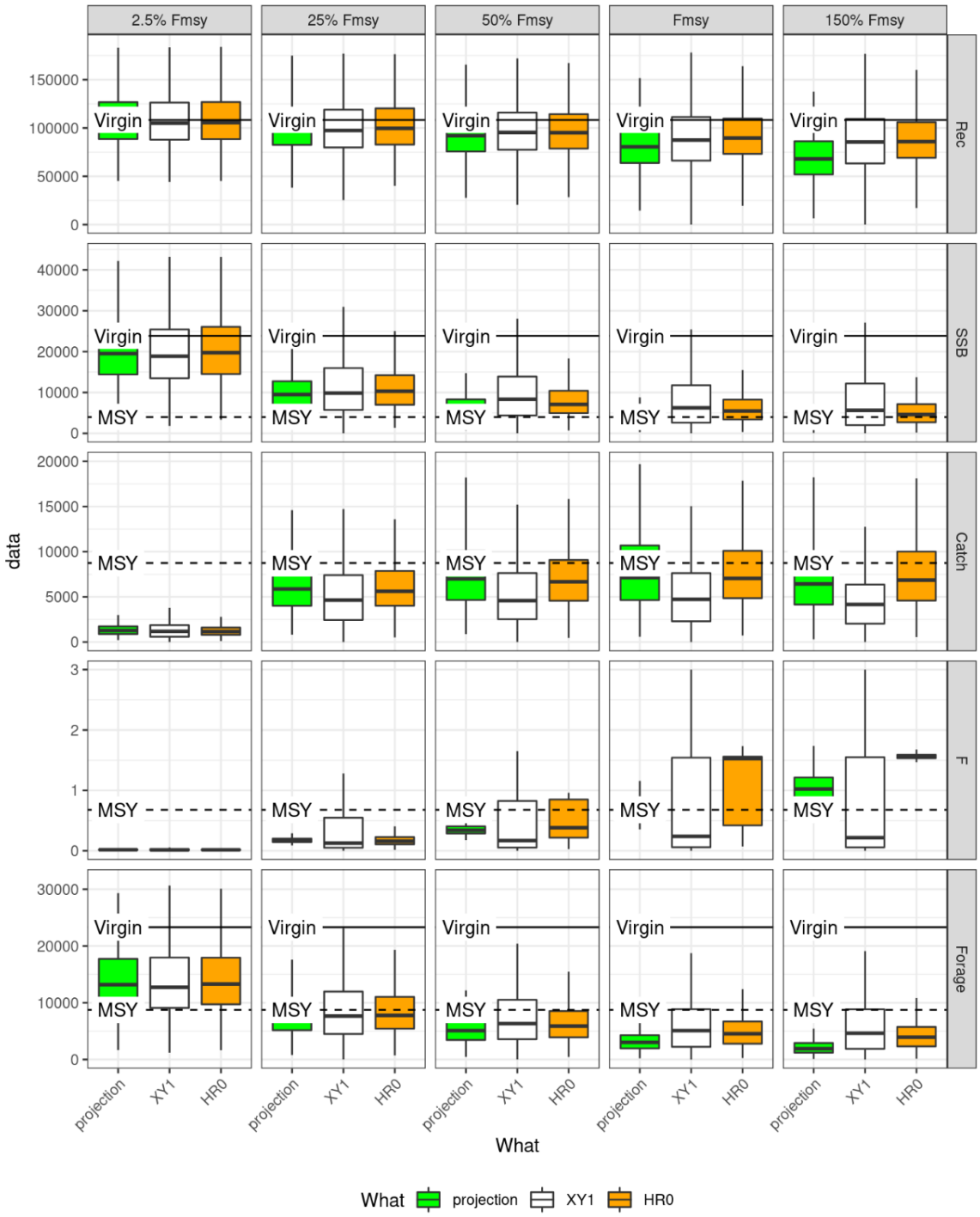

Figure 11.8.1. Results from an MSE for sprat comparing, a simple projection with the 2-over-3 rule and a harvest rate rule.

The MSE also explores empirical control rules which could be implemented for stock assessments based on methods such as Daily Egg Production Method (DEPM, Parker, 1980), Close Kin Mark Recapture (Bravington et al. 2016), Tagging (Bauerlien et al., 2021) acoustic surveys (O'Malley et al., 2022).

Bauerlien, C.J., Crane, D.P., Smith, S., Palmer, G., Young, T. and Goetz, D.B., 2021. Estimates of Abundance and Extreme Catch-and-Release Exploitation in a Southern Riverine Muskellunge Fishery. North American Journal of Fisheries Management, 41(5), pp.1602-1615.

Bravington, M. V., Skaug, H. J., and Anderson, E. C. 2016. Close-kin mark–recapture. Statistical Science, 31: 259–274.

O'Malley, M., Mullins, E. and Nolan, C., 2022. Atlantic Herring in 6aS/7b, Industry Acoustic Survey Cruise Report, December 2021 and January 2022. Marine Institute.

Parker, K. 1980. A direct method for estimating northern anchovy, Engraulis mordax, spawning biomass. Fisheries Bulletin, 78: 541–544.

11.9 Stock assessments: old fish (Tanja Miethe)

Age-based stock assessments, such as SAM, make use of commercial catch data, age-based survey indices and biological data (maturity-at-age, weights-at-age, natural mortality at age). While the estimation of fishing mortality and stock abundance is based on numbers, weights-at-age and maturities-at-age are often used to scale the outputs to units of biomass (SSB, Yield).

Weights-at-age depend on lengths-at-age. Length distributions of a stock are influenced by multiple factors including fishing mortality. Changes in length distributions can be linked to the level of fishing mortality (Hordyk et al. 2015; Jardim et al. 2015; Shin et al. 2015). High fishing mortality is expected to truncate the length distribution and thereby decrease length-at-age, particularly at older ages, as well as weights-at-age. WKLIFE X considered length data (length-based indicators, LBI) to be included in harvest control rules for category 3 stocks (ICES, 2020).

Similarly, for category 1 stocks high fishing mortality is expected to affect lengths distributions. On the example of North Sea whiting (whg.27.47d) assessment, changes in catch-at-age and weights-at-age were presented. NS whiting assessment results (ICES, 2022) show high fishing mortality until about year 2000 followed by a decline in spawning stock biomass with lowest observed SSB in 2007. Looking at catch weights at age for this stock, a decreasing trend can be observed for ages 6+ from the mid 1980ies until 2000 and a subsequent increase in following years. Similarly in maturity at age, estimated as time varying values from survey data, show an decreasing trend in probability mature for ages 2+ with lowest values in the early 2000s and a subsequent increase in the following years. The increase in both maturity and catch weights-at-age could be explained by the reduction in fishing mortality allowing larger individuals to survive. The probability to be mature at age 1 increased since the late 1990ies consistent with a genetic/phenotypic response to selection pressure towards individuals reproducing earlier in life at smaller size (Law, 2000). Number of individuals aged 7+ are still low in recent years due to the low recruitment observed around 2012. This caused retrospective issues in the recent assessment (ICES, 2022) and required a change in the plusgroup from 8+ to 6+ and a recalculation of reference points. Due to the time lag for subsequent growth, an increase in numbers at older ages is expected in the coming years due to stronger recent cohorts entering the stock.

Other factors influence the stock and catch weights-at-age such as growth changes due to warming, food limitation and density dependence (Baudron et al. 2014). For

NS whiting in particular, a decrease in asymptotic length alongside warming temperatures was observed. However, the study suggests an increase in the asymptotic length at the end of the time period, together with a recent increase in catch weights, suggests that the decrease in fishing mortality may be an important factor for this recovery. A lag between the response of LBIs and catch weights-at-age to the release from fishing pressure could be related to generation time and recruitment variability. It could be suggested to making better use of stock or catch weights at older ages or available LBIs to estimate fishing mortality in age-based models.

Baudron, A. R., Needle, C. L., Rijnsdorp, A. D., and Tara Marshall, C. 2014. Warming temperatures and smaller body sizes: synchronous changes in growth of North Sea fishes. Global Change Biology, 20:4, 1023-1031.

Hordyk, A., Ono, K., Valencia, S., Loneragan, N., and Prince, J. 2015. A novel length-based empirical estimation method of spawning potential ratio (SPR), and tests of its performance, for small-scale, data-poor fisheries. ICES Journal of Marine Science, 72: 217–231.

ICES. 2020. Tenth Workshop on the Development of Quantitative Assessment Methodologies based on LIFE-history traits, exploitation characteristics, and other relevant parameters for data-limited stocks (WKLIFE X). ICES Scientific Reports. 2:98. 72 pp. http://doi.org/10.17895/ices.pub.5985

ICES. 2022. Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak (WGNSSK). ICES Scientific Reports. 4:43. http://doi.org/10.17895/ices.pub.19786285

Law, R. 2000. Fishing, selection, and phenotypic evolution. ICES Journal of Marine Science, 57:3, 659–668.

Shin, Y. J., Rochet, M. J., Jennings, S., Field, J. G., and Gislason, H. 2005. Using size-based indicators to evaluate the ecosystem effects of fishing. ICES Journal of Marine Science, 62: 384–396.

11.10 Biopar shorten time series length (Anders Nielsen)

A question was posed at MGWG about the length required to estimate the parameters of the biological parameter model used within SAM. A stand-alone version of the code was provided to the MGWG and the experiments of shortening the time series of stock-weights for north-east arctic cod revealed that with less 8 years of data the confidence intervals tended to collapse (because the variance parameters were not estimable), but with about only 10 years of observations it was able to estimate the parameters needed.

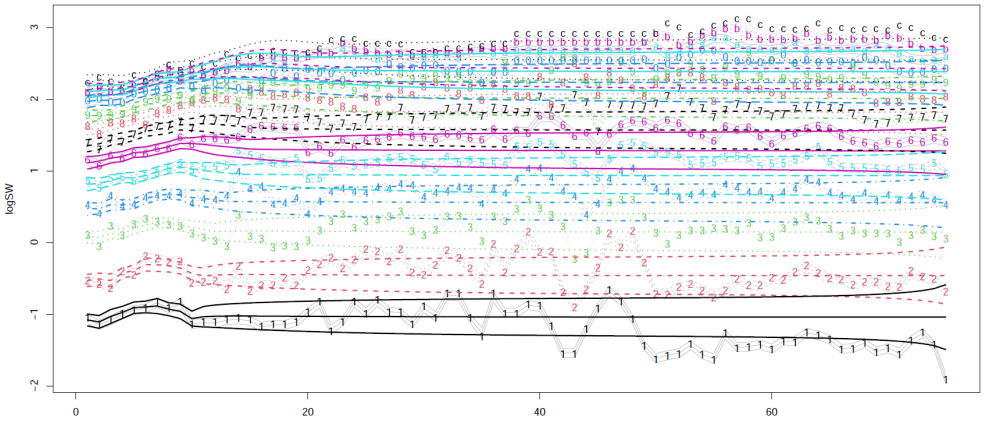

Figure 11.10.1. Observed log-stock-weights (numbers) and predictions from using only the first 10 years of observations (lines).

11.11 Biopar reduce number of ages (Timothy Earl)

Following the example of modelling stock weights presented in Sections 11.1 and 11.10, a sensitivity test was run investigating the influence of the number of ages on the estimated parameters and estimated values. Starting with the youngest three ages (3–5) older ages were successively added to the model until the full range of ages (3–13+) was included. Parameters generally showed a trend with increasing number of ages leading to a higher values of φ_1, φ_2, observation error and process error (Figure 11.11.1). Estimated values were similar within the range of years observed, but showed some differences in the forecast values, with higher maximum age leading to lower forecast stock weights (Figure 11.11.2).

Figure 11.11.1: Estimated parameters in the stock weights model depending on the number of years included.

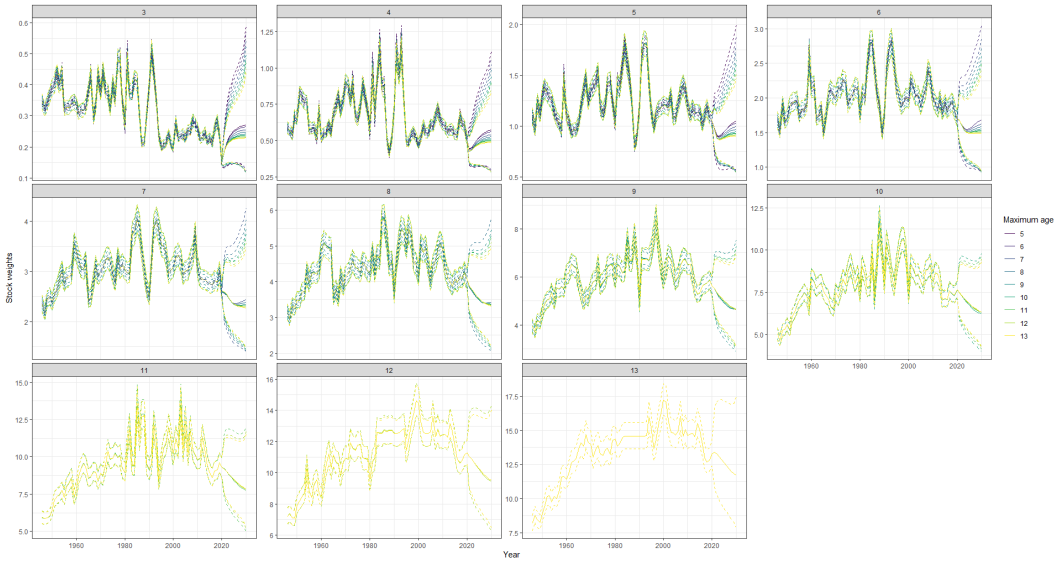

Figure 11.11.2: Estimated stock weights in the stock weights model depending on the number of years included.

11.12 Plus group test for stock weight at age state-space estimation (Chris Legault)

A simple simulation was used to examine the new feature of allowing an additional variance for the plus group in the biological parameters state-space code provided during the meeting. The simulated population had 40 ages and 40 years, constant weight at age over time, variable recruitment (lognormal), and dramatically varying fishing mortality to change the age structure within the plus group and thus the mean stock weight in the plus group. A small amount of lognormal error (sd=0.05) was added to create observed weight at age for all 40 ages and years. Four different plus groups (10, 15, 20, 30) were created and the true and observed weight at age for each plus group was computed using a weighted average based on the population size at age in that year. The biopar code was used to estimate weights at age for the four different plus groups with and without the added parameter to treat plus group random effect differently. The true, observed, and estimated plus group weights at age were compared as were the population stock weight summed across all ages. Including the additional parameter for the plus group weights at age caused a slight reduction in the confidence intervals associated with the plus group weights at age. The difference between including or not this additional parameter would be expected to be larger for forecasted weight of the plus group, but this was not examined in this study. When the plus group was set to young ages and exhibited large changes over time, the estimated values tracked the true estimates quite well. When the plus group was set to old ages there was some bias in the estimates, which may be due to the strong signal of no change over time in weight at age from the many younger ages in this simple study. The summed population stock weights were essentially identical across all plus groups and treatment of the additional parameter. Code to conduct this simulation test is available on the meeting SharePoint site in Section 07 Software, plusgrouptest.R.

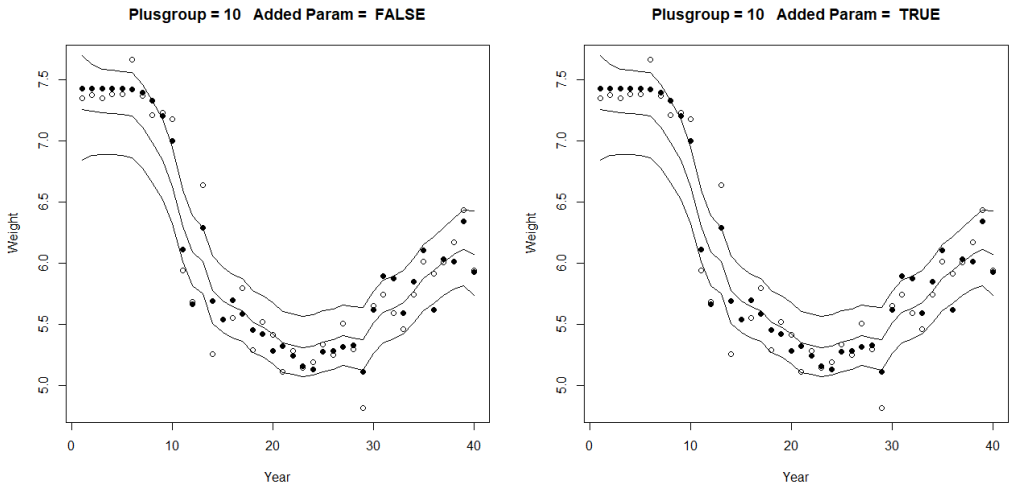

Figure 11.12.1. True (filled circles), observed (open circles), and estimated (lines showing point estimate and 95% confidence interval) for weight at age in the plus group when the plus group was set to age 10 and the added parameter for the plus group was either not included (left panel) or included (right panel).

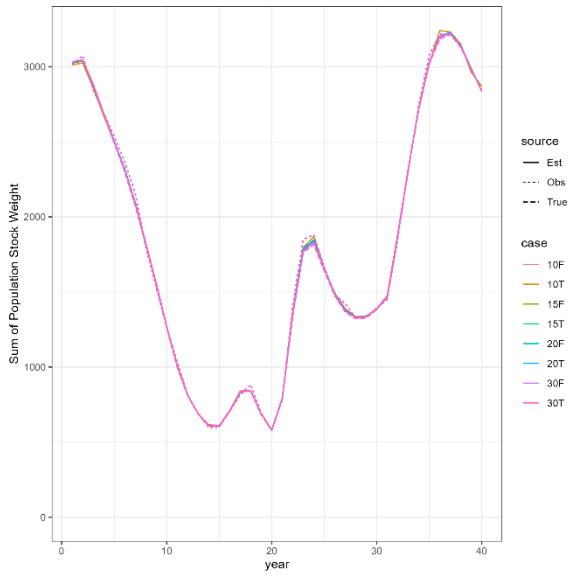

Figure 11.12.2. Sum of population stock weight by year for 24 combinations of four plus group ages (10, 15, 20, and 30) and whether the added parameter was included (T) or not (F) crossed with the true, observed, or estimated weights at age.

11.13 Nodes, connections, and basis functions for GMRF (Colin Millar)

A short presentation was given showing that a random effect with a variance given by a GMRF, such as random walks, can be recast in terms of spline basis functions with the distribution of the random effect now being independent normal. This is achieved by taking the eigen decomposition of the precision matrix of the GMRF, and extracting the eigen vectors where the eigen values are greater than zero. This formulation allows for a reduction in the dimension of the model space by removing the basis functions with the highest variability, resulting in what is

known as reduced-rank spline approximation to the GMRF. It was suggested that further investigation be done on this topic with regard to using the (reduced rank) spline formulations for predictions, comparing them to the equivalent models fitted as full random effects.

11.14 MSE for DLMs when retro present (Chris Legault)

This work was conducted by the Index-Based Methods Working Group at the Northeast Fisheries Science Center and recently published (dx.doi.org/10.1139/cjfas-2022-0045). In this region, the local practice is to immediately adopt a data-limited method (DLM) when an age-based stock assessment is rejected due to a strong retrospective pattern. This is because catch advice is needed for the next fishing year. The main question posed to the group was whether this practice is justified. The group conducted a large simulation study loosely based on local groundfish biology, data collection, and a range of DLMs. The DLMs, along with a retrospective-adjusted statistical catch-at-age (SCAA) model, were applied every other year in a forty year feedback period that occurred after a 50 year simulated starting condition. There were 16 conditions examined, importantly the source of the retrospective pattern was either under-reported catch or an increase in natural mortality that was not known in the assessments. The 12 DLMs, an ensemble of some of the DLMs, and the SCAA were applied and a large number of response metrics were collected. There were a number of DLMs that performed poorly in these simulations across all scenarios. This does not mean these DLMs are always a poor choice, just that they should not be expected to perform well when catch under-reporting or a large (and unknown) increase in natural mortality has occurred. The SCAA performed at least as well as the remaining DLMs, indicating that the current practice is not expected to produce better catch advice than sticking with the retrospective-adjusted SCAA.

11.15 Diagnostics for SAM – a speculative idea (Timothy Earl)

Experience from stock assessment Working Groups suggests that excessive retrospective bias (ϱ) is the most common reason for analytical stock assessments to be rejected, in part because ICES has adopted quantitative guidelines for this diagnostic, unlike other assessment diagnostics. Once a large value of ϱ has been identified (e.g. outside the range -0.15–0.2 for long-lived stocks), it can be difficult to know which aspect of the data, biological assumptions or model mis-specification is the cause of the issue.

The presentation proposed investigating a modification to the source code of models such as SAM that would allow ϱ to be calculated within the model, therefore allowing the impact of catch at age, survey data and biological parameters on ϱ to be investigated through examining the partial derivatives of ϱ with respect to each data point. Such an approach could lead to identifying data particularly associated with the retrospective, providing a starting point for identifying the source of the retrospective bias.

11.16 Individual approach to movement estimation (Anders Nielsen and Tobias Mildenberger)

The movement model describes an advection-diffusion process that utilizes environmental fields and smooth functions to inform advection and diffusion rates (Thorson et al. 2021). Two approaches are considered for the estimation of parameters: the matrix exponential and the extended Kalman filter.

The matrix exponential approach can be summarized by setting up the movement intensity matrix, which describes the intensity of moving from one cell i to another cell j. Note that this definition of M* includes the fishing and natural mortality rates, which addressed in more detail in the next section.

$$M^{\star}_{i\to j} = \begin{cases} D_i/\Delta^2 + 0.5\alpha_i^{(x)}/\Delta, & \text{if cell } j \text{ right of cell } i \\ D_i/\Delta^2 - 0.5\alpha_i^{(x)}/\Delta, & \text{if cell } j \text{ left of cell } i \\ D_i/\Delta^2 + 0.5\alpha_i^{(y)}/\Delta, & \text{if cell } j \text{ above cell } i \\ D_i/\Delta^2 - 0.5\alpha_i^{(y)}/\Delta, & \text{if cell } j \text{ below cell } i \\ F_i, & \text{if cell } j \text{ is 'caught'} \\ M_i, & \text{if cell } j \text{ is 'dead'} \\ -\sum_{j\neq i} M^{\star}_{i,j}, & \text{if } i = j \quad \text{(obviously calculated last in row)} \\ 0, & \text{otherwise} \end{cases}$$

The matrix is constructed from a diffusion field D and an advection vector-field α. The fields are defined by splines applied to the environmental fields. The vector-field for advection is defined as the gradient of a habitat field h, which is defined as:

$$h(i) = S_1^{(h)}(I_1(i)) + \cdots + S_m^{(h)}(I_m(i)) \text{ used as } \alpha(i) = \nabla h(i)$$

Where S denotes spline functions and I denote environmental input fields. As such the advection field is defined in each grid cell and the gradient is approximated by finite differences. The diffusion field is similarly but more directly defined as:

$$\log D(i) = S_1^{(D)}(I_1(i)) + \cdots + S_m^{(D)}(I_m(i))$$

As the environmental input fields are defined only at a discrete grid, then the advection and diffusion fields are also restricted to a discrete grid. This matches well with the conventional tags, which are observed to move from one specific cell (i) to another (j) and possibly with a hidden Markov model approach to the archival tags (Pedersen, 2010). The discrete definition of the advection and diffusion fields does not match well a more continuous approach to modelling the tags. For such approaches to work we need the advection and diffusion fields to be defined – and be differentiable – in all points in our study area. If we imagine that we could find a differentiable representation of each field (e.g. replace I 1with

*I*1˜), then the fields for advection and diffusion would also be continuous and differentiable defined everywhere.

Two approaches were used to represent the environmental field in a differentiable way. First a neural network with 3 inputs (lon, lat, and '1' (representing an intercept)), 15 hidden nodes, and one output (the corresponding environmental value) was set up. Such a neural network has 60 model parameters and after they have been estimated, then the environmental field can be approximated by evaluation the network at any (lon, lat)-point. The network representation is continuous and differentiable, so it can be used as basis for more continuously defined movement models.

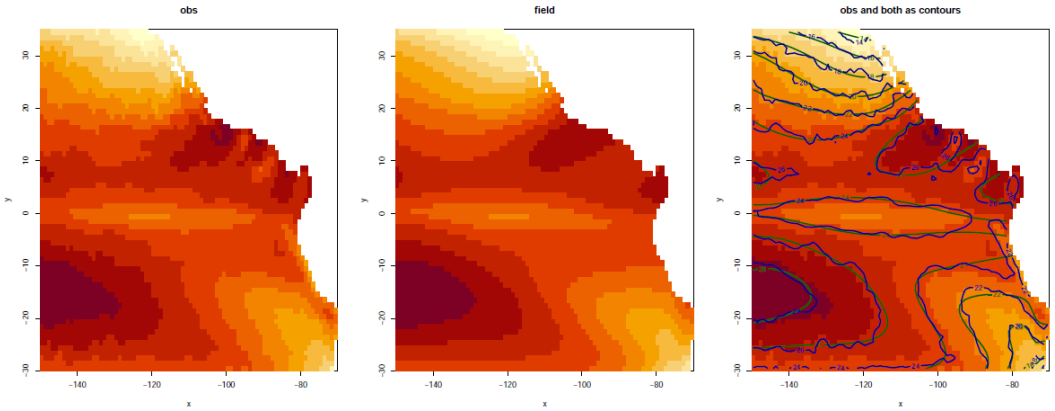

Figure 11.16.1. Raw temperature field (left), smooth neural network representation (middle), and compared contours (right).

While the neural network approach does work and gives a parametric approximation of the field it uses a relatively high number of parameters and could possibly be poorly defined outside the data area. The second approach to represent the environmental field in a differentiable way uses local interpolation. The locally interpolated field is defined in any position (lon, lat) by a weighted average of the input field values from a radius 'R' around the position. The distance-weighting of the local points is defined by an iterated cosine function to ensure differentiability (to a high enough order) when observation points are smoothly added or excluded from the average, as the position changes. If the radius is defined to be exactly equal to the distance between neighboring positions, then the value of the differentiable representation evaluated at an actual observation position will be exactly equal to the observed value (because exactly one observation is included), and at any other position the value will be a smooth weighted average of neighboring points.

The advection field is defined as the gradient to the habitat field, and it is possible (easy) to evaluate the gradient of the local interpolation calculations. That gradient is, however, not what we want. If the radius is defined to be equal to the distance between neighboring points, then the gradient of the smooth representation will be zero if evaluated at the position of an observation (because exactly at such a point the calculation only involves taking the average of one point). To get useful gradient fields (one in the longitude direction and one in the latitude direction) the input fields for delta-longitude and delta-latitude were computed from each of the environmental fields and then the local interpolation method applied to get smooth versions. Hence each discretely given environmental field I is converted into 3 smooth fields (*I* ˜, ˜, and ˜). Once these are defined, the gradient of the

habitat field itself can be computed and that is what is needed to define advection and diffusion everywhere in a differentiable way.

With the advection and diffusion fields defined it is possible to formulate (and estimate) movement models, which are not defined on a grid. For the archival tag observations, the model becomes:

$$o_t \sim N(\psi_t, \Sigma_o)$$

$$\psi_{t+\Delta t} \sim N(\psi_t + \alpha(\psi_t, t)\Delta t, 2D(\psi_t, t)\Delta t I_{2\times 2})$$

Here psi represents the true unobserved position, o is the observed position with observation noise Σ_o. Starting from the known release position the model likelihood is computed via a classic Kalman filter (Harvey, 1990).

For the conventional tags only the release and recapture positions are known. In order to better capture the non-linearities in the spatial fields and the temporally changing environmental fields a number of intermediate time points are inserted between release and recapture time. The likelihood contribution of each conventional tag is ten computed exactly at for the archival tags. Starting from the release location a Kalman filter is used to step from each timepoint and to the next (updating the distribution of the true unobserved position) and finally evaluating the likelihood of the recapture position. The only difference is that no observations are available at all the intermediate timesteps.

As multiple recovered tags are linked to the same release locations (but various recovery times), the computations are optimized for computation speed and memory allocation by estimating the movement matrices dependent on the unique release locations.

Estimation of mortality rates

The recovery of a tag at a given location and time does not only depend on the movement from the release to recapture location, but also the probability of the survival of the fish until being recaptured and the probability of capture at the recapture location by fleet f. Defining the instantaneous fishing mortality Fg,t,f in grid cell g at time t of fleet f proportional to the effort of that fleet and the instantaneous natural mortality rate M as a constant rate in space and time, allows to calculate the likelihood of recapture of tag h at time m as:

$$\boldsymbol{L}_h(M, \lambda|E) = \frac{F_{g_m,t_m,f}}{M + \sum_{f=1}^{F} F_{g_m,t_m,f}} (1 - e^{-(M+\sum_{f=1}^{F} F_{g_m,t_m,f})\Delta t_m}) \prod_{s=1}^{m-1} e^{-(M+\sum_{f=1}^{F} F_{g_s,t_s,f})\Delta t_s}$$

On the other hand, a conventional tag that was not recovered, is either lost with the death of the fish due to natural causes or still attached to the living fish. Note, that this assumes that the tag would have been reported if it was caught by any fleet (non-reporting=0). Thus, the likelihood of a non-recaptured tag h with assumed maximum age A:

$$\boldsymbol{L}_h(M,\lambda|E) = \frac{M}{M+\sum_{f=1}^{F} F_{g_1,t_1,f}}(1-e^{-(M+\sum_{f=1}^{F} F_{g_1,t_1,f})\Delta t_1}) + \sum_{a=1}^{A} \frac{M}{M+\sum_{f=1}^{F} F_{g_a,t_a,f}}(1-e^{-(M+\sum_{f=1}^{F} F_{g_a,t_a,f})\Delta t_a}) \prod_{s=1}^{a-1} e^{-(M+\sum_{f=1}^{F} F_{g_s,t_s,f})\Delta t_s} + \prod_{a=1}^{A} e^{-(M+\sum_{f=1}^{F} F_{g_a,t_a,f})\Delta t_a}$$

Similar to the environmental fields, the effort must be defined in a continuous differentiable space for the Kalman filter approach. Therefore, we used the same local interpolation approach described above for the interpolation of the effort.

Population model

A rough estimate of the population biomass in space and time and based on information from the different fleets ($B^r_{g,t,f}$) can be calculated from observed catches and estimated fishing and natural mortality rates corresponding to fleet f by means of the Baranov catch equation.

11.17 Summary of CAPAM-related issues relative to ICES MGWG (All)

A number of meeting participants also attended the recent CAPAM pretty good practice meeting in Rome. The group discussed ideas from that meeting that could be relevant to this group, especially as topics for future meetings. Close kin mark recapture, acoustic estimates of total stock biomass, and tagging models similar to those presented at this meeting are all examples of approaches that can be used to provide catch advice without the use of traditional stock assessment. This area is of interest to the MGWG as it could provide new avenues for research and development in fisheries science and management. Mackerel, elasmobranchs, and anglerfish were all suggested as possible case studies for application of close kin mark recapture. A simulation study based on mackerel could explore the probability of finding siblings or half-siblings in different size samples given the large population of mackerel under different assumptions of mixing.

The CAPAM session on spatial modeling was of interest to this group, especially after hearing about the use of multispecies SAM applied to a single species with multiple areas at this meeting. The group feels the latter approach is worth pursuing in future meetings either through application to ICES stocks or in simulations. The consequences of modeling the stocks incorrectly was of particular interest. A speaker at the CAPAM meeting showed results from inland tagging studies with a much narrower spatial range than typically used in marine fisheries assessment. It was suggested that using tagging information from lakes might be useful for addressing both spatial and tagging issues in future MGWG meetings due to the reduced scale and easier ability to identify boundaries.

Conflicts in assessment data were discussed based on the CAPAM session of data weighting. Many ICES stock assessments do not use data weighting, but do experience conflicts in the data. Identifying the incorrect data and removing it from the model is the obvious, but most difficult, solution. How to deal with conflicting

data within a stock assessment model remains a challenge and would be a good topic for future meetings of the MGWG. There may be approaches used in other fields for dealing with conflicting data that could be applied to fisheries stock assessment.

11.18 Using a Markov model to describe movement cuttlefish of in the English Channel (Michael A. Spence)

Cuttlefish in the English Channel have a peculiar lifestyle, where they die shortly after reproduction. They spawn in coastal regions, and move through inshore regions, where they are susceptible to otter trawlers, to offshore regions where they are caught by beam trawlers. Information on effort, catches and some observer programs that can record the length of the catches for the different fleets, as well as a survey. To account for the movement between the different gear types, a continuous time Markov model was added. The movement model is described by rate functions that describe the time of each movement, the intensity, and the how long it happens.

11.19 Estimate sd of rho in SAM (Christoffer Albertsen)

Based on the ideas presented by Timothy Earl (section 11.15), the multi-stock version of SAM was modified to calculate Mohn's rho. For a single-stock SAM fit, n clones are made where 1 to n years of data are removed. Using the multi-stock version of SAM, all n+1 models are fitted simultaneously and Mohn's rho is calculated for SSB, Fbar, and recruitment. Using the ADREPORT functionality of TMB, both estimates and their standard deviation can be obtained. The approach was tested and gave the same Mohn's rho values as the corresponding SAM functions with the added benefit of confidence intervals. Further, the coverage of the confidence intervals was tested in a simulation study. For correctly specified models, the coverage of nominal 95% confidence intervals were found to be good (95-98%). The idea of using partial derivatives of Mohn's rho has not been explored yet, but the building blocks needed are now available.

Further, there is a need to look into the consequences of fitting the retro runs as independent assessments. In reality, the data before the "peeling" period is the same which would lead to correlation in the estimates. Such a correlation was not accounted for in the first implementation, but could improve the confidence intervals. Finally, Mohn's is calculated as

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\frac{v_{T-i}^{(i)} - v_{T-i}}{v_{T-i}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{v_{T-i}^{(i)}}{v_{T-i}} - 1\right),$$

where $v_{T-i}^{(i)} > 0$ is the estimated value of interest in year $T-i$ when removing i years of data from the assessment and v_{T-i} is the corresponding value from the full data fit. Therefore, the value is bounded from below at -1. Further, for a SAM model that is a good approximation of reality, the estimators of $v_{T-i}^{(i)}$ and v_{T-i} are random variables with presumably the same mean. However, in general

$$E\left(\frac{v_{T-i}^{(i)}}{v_{T-i}}\right) \neq \frac{E(v_{T-i}^{(i)})}{E(v_{T-i})} = 1$$

Therefore, the expected value of Mohn's rho is not guaranteed to be zero for a good SAM model. The expectation will depend on the variance of v_{T-i} as well as the correlation between $v_{T-i}^{(i)}$ and v_{T-i}. Therefore, it may be an advantage to calculate confidence intervals on a different scale.

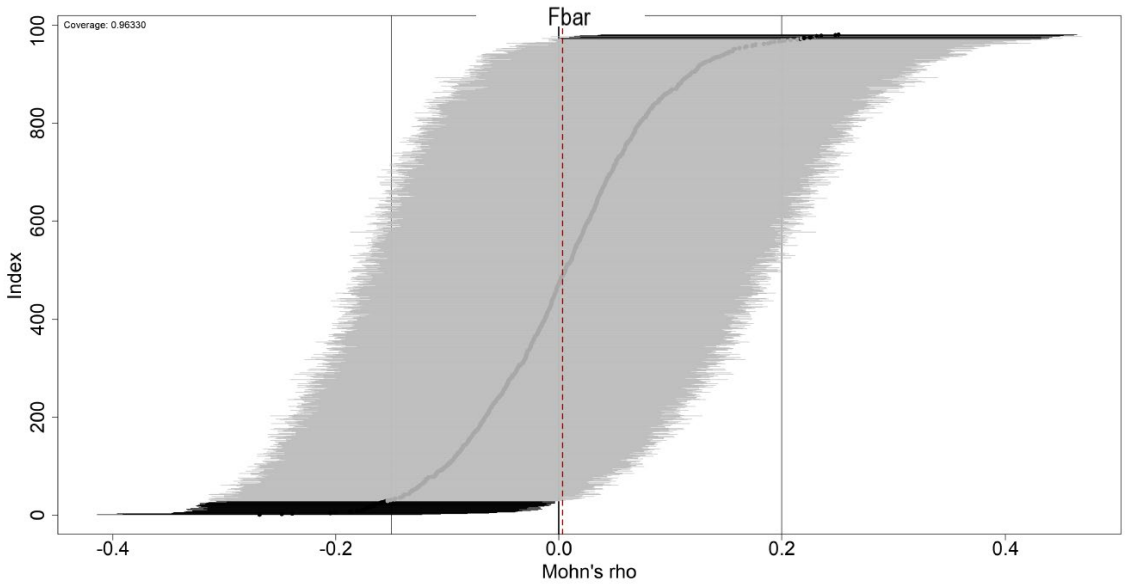

Figure 11.19.1. Point estimates and nominal 95% confidence intervals of Mohn's rho for Fbar under a correct model specification.

11.20 Add plus group option to SAM biological parameters (Anders Nielsen)

A fairly recent addition to the state-space assessment model SAM is to consider the biological parameters (stock-weight (SW), catch-weight (CW), fraction mature (MO), and natural mortality (M)) as observations subject to observation noise and estimate process describing the development of each quantity. This has several advantages, which are: separating observation and process noise, natural prediction framework, and quantifying prediction uncertainty. The model suggested used the same correlation model for all age groups, ant MGWG suggested to test the option of allowing additional correlation connecting the neighbouring plus-groups. This makes sense, because if a plus group is consisting of many age groups, then it should be expected to be more similar from year to year.

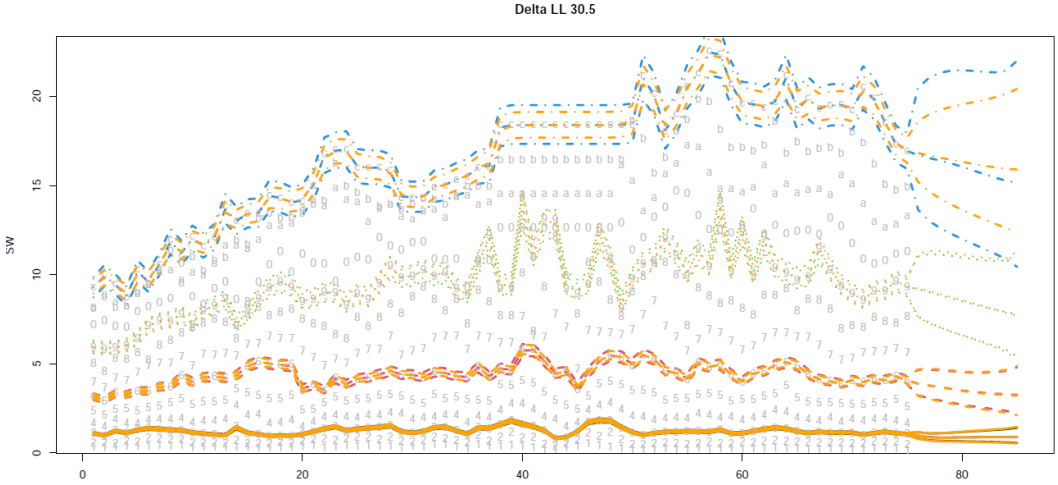

Figure 11.20.1: Observed stock mean weights. Fitted and predicted stock weights from model without plus-group (blue, green) and from model with plus-group (orange).

The option was implemented in SAM and it did indeed perform better in the test-example provided (Fig. 11.20.1). The prediction interval became more narrow and the likelihood improved by 30.5 log-likelihood units.

This option is now available in SAM for all 4 biological parameters. It can be configured e.g. for SW in the configuration file by setting:

```
# Integer code describing the treatment of stock weights in the model (0 use as known, 1 use as observat
ions to inform stock weight process (GMRF with cohort and within year correlations)), 2 to add extra cor
relation to plusgroup
 2

$keyStockWeightMean
# Coupling of stock-weight process mean parameters (not used if stockWeightModel==0)
 0 1 2 3 4 5

$keyStockWeightObsVar
# Coupling of stock-weight observation variance parameters (not used if stockWeightModel==0)
 0 0 0 0 0 0
```

11.21 Added flag to TMB in OSA for mixed distributions (Andrea Havron)

The oneStepPredict function in TMB was modified to allow for OSA residual calculations for data that are a mixture of discrete and continuous observations (mixed distributions). The current implementation of the oneStepPredict function results in errors when run on mixed distributions. Three elements of the code were modified. First, flags were added to catch errors and allow unique calculations of OSA residuals when distributions were specified to be mixed. Second, code was written to correctly add random noise to discrete observations from a mixed distribution. The method was based on the qres.tweedie function from the statmod R package1. Third, oneStepGeneric code was modified within a mixed flag to attempt to correctly calculate the cdf of the mixed distribution.

A case study was developed using a simple Tweedie model, modified from the TMB example suite, to include a keep vector needed to implement the OSA oneStepGeneric method. Data were simulated in R and OSA residuals were calculated using the modified oneStepPredict function. OSA residuals were compared to quantile residuals calculated in R empirically. The OSA residuals do not appear to be implemented correctly yet as the results suggested model misspecification (Fig. 11.21.1). In the comparison, the OSA residuals matched the

empirical ones exactly when observations were zeros, however, accuracy declined as observations approached zero (Fig 11.21.2., Fig 11.21.3).

Experts at the ICES MGWG suggested an additional comparison of the oneStepGeneric method using the unmodified oneStepPredict function and a fully discrete distribution. A comparison was done with the Poisson distribution and it was found that the OSA residuals matched the empirical quantile residuals (Fig. 11.21.4, Fig 11.21.5). The discreteSupport option was required for low lambda values to prevent the code from using spline-based methods to calculate residuals. This was likely due to the difficulty of defining a spline for a small number of discrete values. The next steps will be to contact the TMB development team and discuss this issue further.

1Dunn, P. K., and Smyth, G. K. (1996). Randomized quantile residuals. J. Comput. Graph. Statist., 5, 236-244.

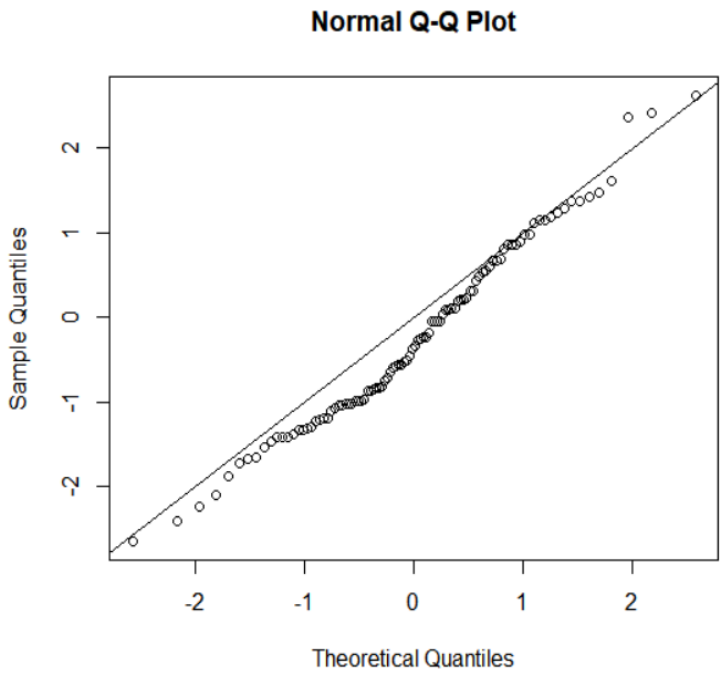

Figure 11.21.1. QQnorm plot of OSA residuals generated using the modified oneStepGeneric method for Tweedie data fit using the correct model.

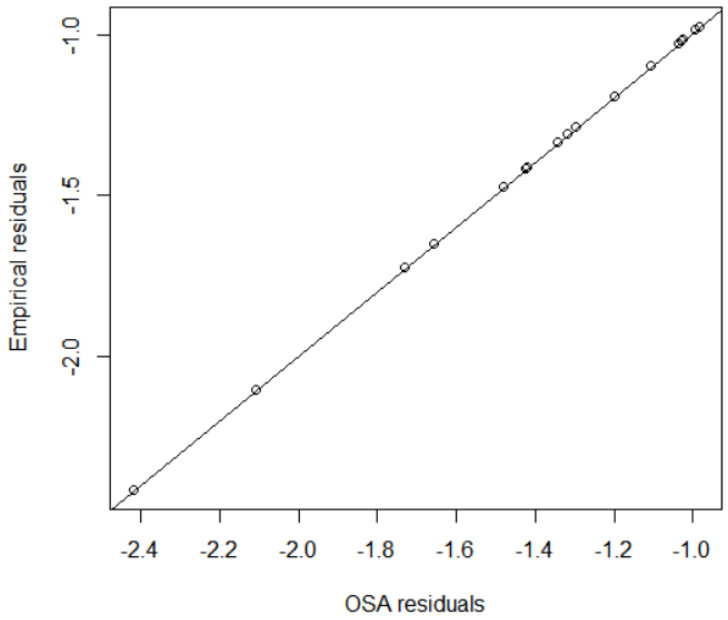

Figure 11.21.2. A comparison of OSA and empirical residuals when observations of a mixed distribution were zero.

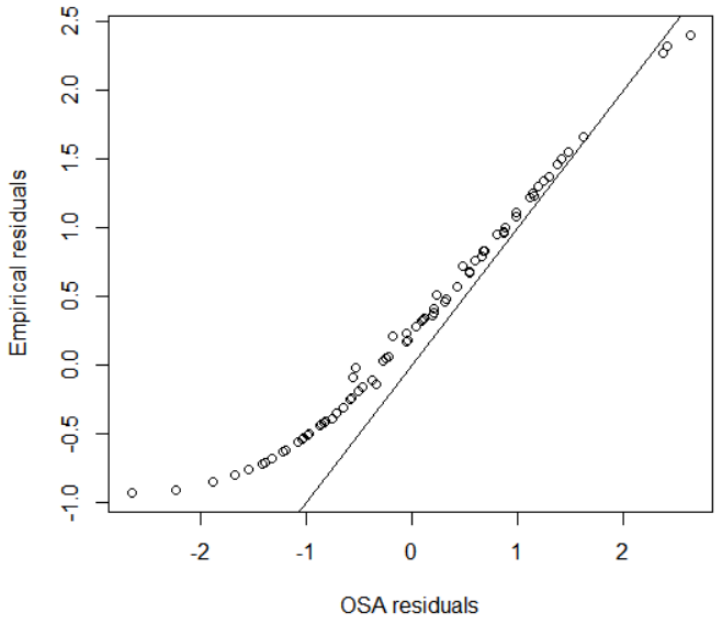

Figure 11.21.3. A comparison of OSA and empirical quantile residuals when observations of a mixed distribution were continuous and positive.

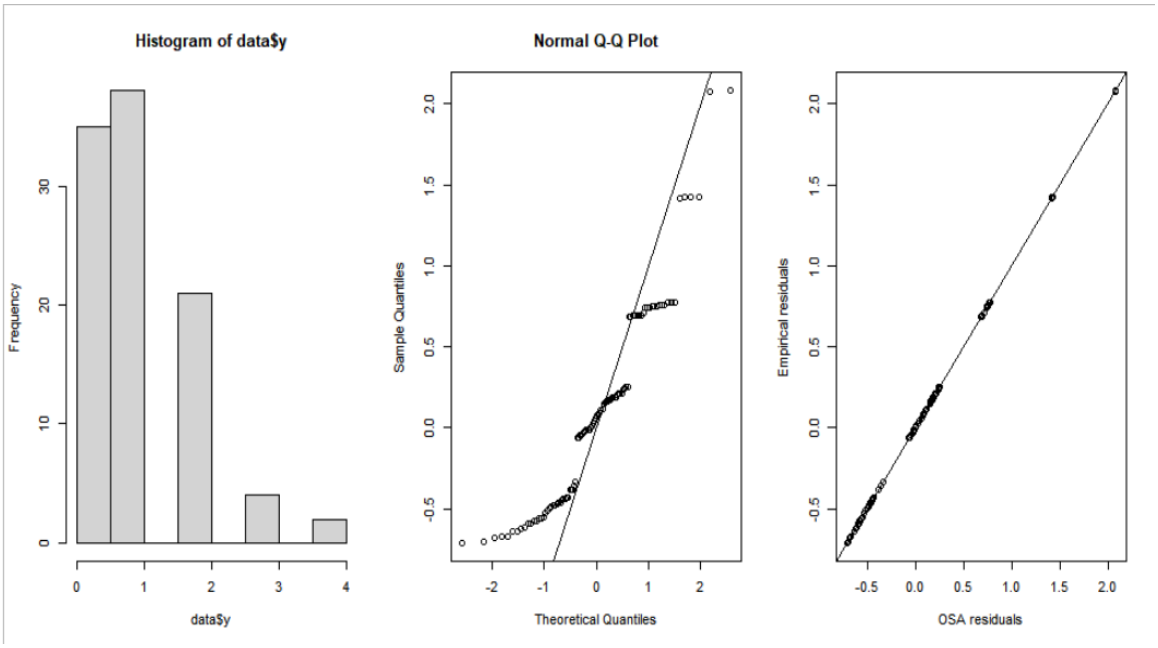

Figure 11.21.4. From left to right: Histogram of Poisson data for lambda = 1, OSA residual qqnorm plot, Comparison of empirical quantile and OSA residuals with discreteSupport set to 0:10.

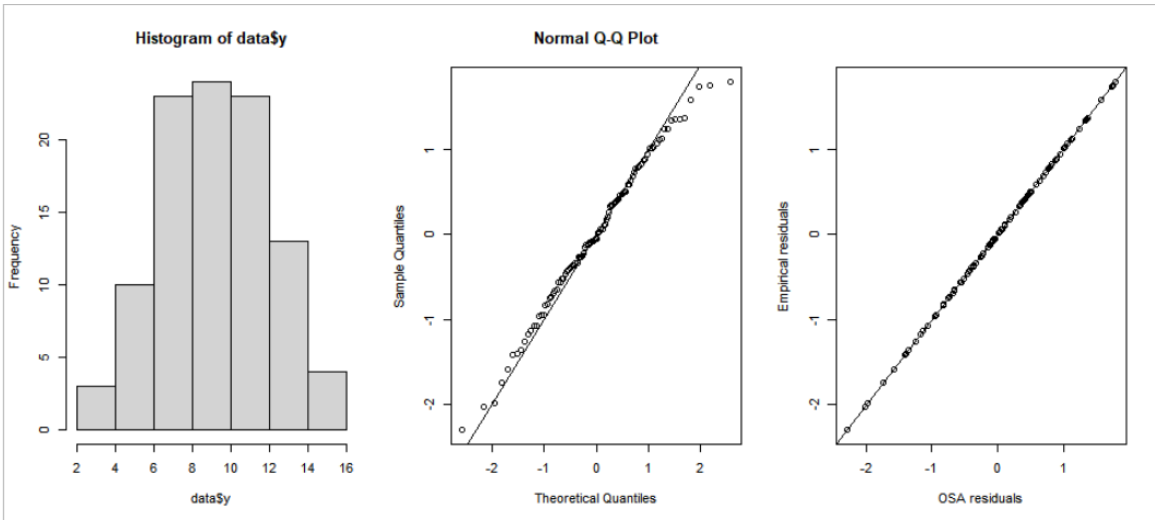

Figure 11.21.5. From left to right: Histogram of Poisson data for lambda = 10, OSA residual qqnorm plot, Comparison of empirical quantile and OSA residuals.

11.22 SS3diags package (Henning Winker presenter, abstract kindly provided by Laurie Kell)

Figure 11.22.1 shows a procedure for selecting a best assessment. A first steps is to agree on a best fit and then to run models to determine the sensitivity of outputs to elements of the model that are subject to uncertainty. This helps better understand the relationships between model inputs and outputs and can help identify which input factors have a small or large influence. It is important that stakeholders should be allowed to propose scenarios and that there is be an objective way to reject and develop new scenarios.

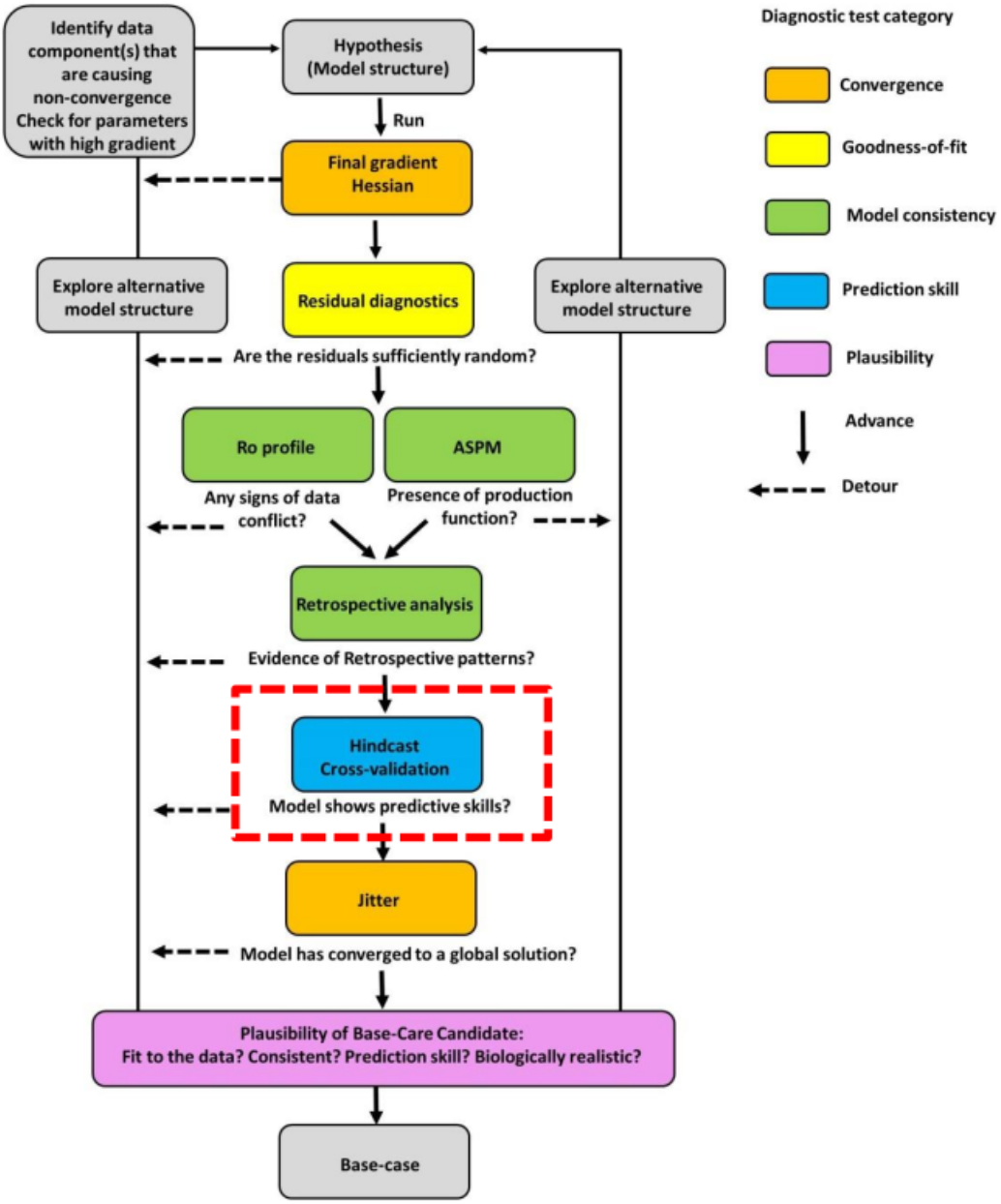

Figure 11.22.1. Procedure for selecting a best model using a range of diagnostics.

11.23 Tuning SAM to F from mean length (Anders Nielsen)

A request was made at the MGWG to extend the state space assessment model SAM to allow the inclusion of an index assumed to be proportional to Fbar. SAM is already capable of including indices of biomass, catch, landings, and more, so making this extension was not a major undertaking. The code was amended and verified via an artificial example where an index of Fbar was constructed by

adding noise to the estimates of Fbar. The model extension performed as expected (fitted the data well and reduced the uncertainty of the Fbar estimates).

The new addition can be used in sam by adding an additional survey fleet with one data column to the data and artificially assigning the ages in the datafile to -1 and -1. The negative ones are indicating that the model need to consider this not as an age–based index. Then in the configuration of the model it need to be specified that "keyBiomassTreat" is 10 for the corresponding fleet, as in:

$keyBiomassTreat

 # To be defined only if a biomass survey is used (0 SSB index, 1 catch index, 2 FSB index, 3 total catch, 4 total landings, 5 TSB index, 6 TSN index, and 10 Fbar idx).

-1 -1 -1 10

11.24 ROC curves for F and mean length (Laurie Kell)

Receiver Operator Characteristics (ROCs) are used to estimate classification skill, and several examples were shown during the meeting. A ROC curve is a graph showing the performance of an classifier at all discriminant thresholds. The ROC curve plots two parameters the True Positive Rate (TPR=TP/(TP+FN)) and the False Positive Rate (FPR=FP/FP+TN), where TP are the number of true positives, TN the true negatives, FP the false positives, and FN the false negatives. The TPR, or sensitivity, measures the ability of a test to identify positive cases, while the true negative rate, TNR = TN/(TN+FP), measures the proportion of negatives that are correctly identified.

The true skill score (TSS, i.e. TSS = TPR + TNR – 1) can can also be calculated. A perfect prediction would receive a score of 1, random predictions receive a score of 0 and predictions inferior to random ones receive a negative score.

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

An example was developed during the meeting comparing length-based indicators and F. See section 11.7 for this example.

Recommendations

Based on work conducted and discussed at the 2022 meeting, the MGWG makes the following recommendations.

Stock Assessment Working Groups and Benchmark Workshops

- Account for covariances/correlations in observations when presenting model residuals (e.g., the one step ahead method).
- Account for uncertainties in data inputs to the extent possible in stock assessment, reference point calculation, and provision of management advice. Similarly, evaluate the consequences of alternative fixed parameter values.
- Use statistical properties of estimates to determine when diagnostic tests fail. Consider both prediction skill as well as model fit and other diagnostics when characterizing the performance of a stock assessment.
- Evaluate data conflicts to determine if they can be removed before modeling and if not address through uncertainty estimates.
- Biological parameters (weights, maturity, and mortality) are important in the forecast and reference point calculations, so model based alternatives should be considered in place of the ad-hoc moving averages currently being applied.
- The newly developed model-consistent confidence limits for Mohn's rho appears logical and promising and should be considered as an alternative to the currently applied "one size fits all" limits.

Methods Developers and Benchmark Workshops

- Consider using the new approach for evaluating tagging data presented at this meeting to improve speed and stability of analysis.
- Explore open science features, such as buddy coding, automated testing, and version control, to develop and advance stock assessment models in order to facilitate collaboration and innovation.
- Consider the use of parameter covariances when confronted with challenging stock boundary issues and calculation of biological reference points (e.g., through the use of multi-SAM or other state-space approaches).
- Develop systems to catalog stock assessment model configurations to improve understanding of current practices in the field, increase transparency, repeatability, and validation.
- Explore approaches that allow use of data as close as possible to its original collection to improve the use of prediction skill measures. Data should be made available in as raw a basis as possible (including spatially).
- Create a catalog of stock assessment diagnostics with quantitative definitions to determine when fail keeping in mind the issue of multiple comparison significance levels. Develop a general software approach and workflow to providing diagnostic tests.