Table S1: Examples of applying the framework of guiding questions to a range of advice questions and models with different complexity. The models range from a single species assessment model (SAM, Nielsen et al. 2021) for Western Baltic Spring-Spawning Herring (WBSS) over models of intermediate complexity (Georges Bank Hydra (Gaichas et al. 2017), Baltic Sea SMS (ICES WGSAM 2019)) to ecosystem/food web models (Georges Bank Rpath (Lucey et al. 2020), Irish Sea Ecopath with Ecosim (EwE, Bentley et al. 2021)). More details on analyses carried out and diagnostics can be found in ICES WGSAM 2019 (Baltic Sea SMS and Irish Sea EwE) and 2023 (Georges Bank Hydra and Rpath) as well as ICES HAWG 2023 (SAM WBSS).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Step/question in the framework** | | **Model example** | | | | |
| **Overall question** | **Detailed question** | **SAM model for Western Baltic Spring-Spawning Herring** | **Baltic Sea SMS** | **Georges Bank Hydra** | **Georges Bank Rpath** | **Irish Sea EwE** |
| **General questions** | | | | | | |
| **What is needed for a successful delivery of the advice product?** | **What is the advice question?** | What is the catch advice for next year? | Which are the time variable natural mortalities of central Baltic herring and Baltic sprat for use in their single species stock assessments? | Can this length based multispecies model be used as an assessment tool? | How does management of a few species affect the rest (other fish and unmanaged or protected apex predators)? | Can some of the outputs of the Irish Sea EwE be used to influence quota setting? EwE output would be used as a synthesized ecosystem indicator to help inform the choice of Ftarget (called Feco)within predefined (from single species benchmarks) FMSY ranges (Rindorf et al. 2016, Bentley et al. 2021). |
| **Has the model the right complexity?** | Yes, the model is a multi-fleet state-space stock assessment for herring in ICES Subdivisions 20-24. The quotas for this stock are given at the fleet level (human consumption or industrial fishing fleet) in Subdivisions 20-21 and 22-24. The multi-fleet assessment and forecast capture therefore the correct level of complexity. However, the stock mixes with North Sea herring in Subdivisions 20-21 and the eastern parts of the North Sea. Part of the data is split between the two stocks. Recent genetic findings show that more stocks are mixing in these areas. The model currently does not account for the correct level of mixing and multi-stock assessments might be better to predict the split within the model. | Yes, SMS is a model of intermediate complexity. It represents the age structure of the species of interest and it fits data typically used in stock assessment models. In addition, it fits stomach data to inform about predator-prey interactions. | Yes, the previous length-based multispecies simulation model did not have estimation capability. An objective function to fit the model to survey and fishery index, length, and diet composition data was developed. | Yes, Rpath is a food web model that can evaluate impacts of changes to a few modeled species on the rest of the modeled species through changing predator-prey interactions and feedback within the system. | Yes, EwE is a food web model that allows to identify major energy pathways within a system. The Irish Sea EwE has 41 functional groups covering all trophic levels and treating commercially important stocks (e.g., cod, sole herring, *Nephrops*, monkfish) as individual groups, some with multi-stanza representation (e.g., cod). EwE is also able to incorporate environmental covariates as forcing functions. Therefore, the model is suitable to provide a summary information on current environmental impacts on commercially important fish stocks. |
| **Can the model deliver the output needed at the right spatial and temporal scales?** | Yes, the model runs for 1991-present and the multi-fleet short-term forecast allows estimating catch advice in the next year given an intermediate year catch constraint per fleet. | Yes, the model estimates natural mortality of the two clupeid stocks by age for the period 1974-present as required by the stock assessment models. Model outputs are quarterly and require annual aggregation for use in the single species stock assessments. | Yes, the model is implemented for Georges Bank, the spatial scale of interest, and can track annual changes matching management time scales. However, the spatial scales of modeled stocks ranging beyond Georges Bank may require different treatment within the model than stocks occupying Georges Bank. | Yes, the model is implemented for Georges Bank, the spatial scale of interest, and can track annual changes matching management time scales. | Yes, next to the simulation of food web dynamics, EwE allows for the incorporation of environmental covariates as forcing functions. Therefore, the model is appropriate to address the need for synthesis of ecosystem indicators for particular stocks. Since the model covers the years 1974 to 2016 future updates are needed to track changes in ecosystem indicators over years. |
| **Which are the most important outputs and metrics relevant for the advice?** | Spawning stock biomass, recruitment and fishing mortality estimates. The MSY reference points estimated during the last benchmark are used to set up the ICES harvest control rule and determine stock and fishing status. | Clupeids biomass and fishing mortality need to be consistent with the single species stock assessment. Main output is time variable estimates of natural mortality at age for herring and sprat. | Model estimation performance (framework review): comparison of model estimates with both observed and simulated data. | Biomass and production for a range of species across trophic levels that are outside the scope of groundfish management. | Ecosystem indicators describing environmental impacts on commercially important fish stocks (e.g., predation, food availability, environmental drivers impacting productivity). |
| **What type of model is available and which skill assessment methods are appropriate?** |  | This is a single species multi-fleet assessment commonly used with different diagnostics available to check estimation and prediction skills (residual analysis, retrospective analysis, leave one survey out analysis, convergence indicators e.g. gradient of the objective function, standard errors, parameter correlation, etc.). | Stochastic multi species assessment model. The skill assessment can be carried out with commonly used diagnostics to check estimation and prediction skills (residual analysis, retrospective analysis). | The length based multispecies model framework has been implemented for research but untested for operational assessment; (in the future) hindcast and forecast diagnostics comparable to those used for operational assessment models (residual analysis, retrospective analysis, likelihood component profiles) will be applied to test estimation and prediction skill. | A food web model is needed to address full food web questions; food web thermodynamic diagnostics and (in the future) fits to biomass observations for both hindcasts and forecasts will be used to assess skill. | For EwE, a set of standard diagnostics has been developed. This includes e.g., pre-balance (PREBAL) diagnostics (Link et al. 2010) for a sanity check on key features (e.g., biomass slope across trophic levels, ratio between predator and prey biomass) of the static Ecopath base model and fit to time series data (e.g., biomass and catch; sum of squared deviations) for the time-dynamic Ecosim part. |
| **What real world observations are available for skill assessments?** | **Are sufficient observations available?** | Yes, the model is fit to the fleets catch-at-age in numbers as well as time and age varying survey indices. Inputs include stock and catch mean weight-at-age, natural mortality, maturity ogive, proportion of F and M before spawning. | Yes, the model uses catch-at-age matrices and survey indices of herring and sprat like in the stock assessment but at quarterly resolution, the stock number-at-age of cod from the assessment (cod is not a dynamic species in the present key-run) and a time series of cod stomach data. | Yes, a 50-year time series of Fall and Spring survey biomass, survey length compositions, and diet composition data are available for Georges Bank. Fishery landings and estimated discards were used from 1978 onwards. Age and growth information from Georges Bank for model parameterization is available for all but 3 modeled species. | Yes, a 50-year time series of Fall and Spring survey biomass, and diet composition data are available for Georges Bank. The current static model was initialized in the 1980s to evaluate skill running forward to the present day in dynamic mode. Landings and discard data for a comparable period can be included in the dynamic model.  Observations for unmanaged or protected apex predators are intermittent in space and time. | Yes, the Irish Sea EwE model uses data gathered and quality checked by ICES WKIRISH and updated stock assessments to parameterize 41 functional groups. The model uses information from literature, an open access stomach database and local ecological knowledge from stakeholders. Catch time series are available as well as survey time series for certain functional groups. Single species assessment results are also used as “observations” where available. |
|  | **Are the observations at the right spatial and temporal scale?** | Yes, some data is split into stocks prior to entering the assessment model. | Yes, this is a single area quarterly model operating at the stock level. All data are available at this resolution. | Yes, surveys providing index and diet observations are designed for estimation at the Georges Bank scale, annually for Fall and Spring. Fishery observations require spatial scaling due to mismatch of larger catch reporting areas with Georges Bank survey areas. Fishery length observations have some gaps in spatial and temporal coverage by gear type for some species. | Yes, surveys providing index and diet observations are designed for estimation at the Georges Bank scale, annually for Fall and Spring. Fishery observations require spatial scaling due to mismatch of larger catch reporting areas with Georges Bank survey areas.  Observations for higher trophic level apex predators are designed for coastwide assessment at a much larger scale than Georges Bank. | Yes, surveys, single species assessments and catch time series are available at the right spatial and temporal resolution (i.e. for the Irish Sea and by year). Diet data are not available by year, but used aggregated over the years 1960 to 2016. |
|  | **How certain and/or biased are the observations?** | As any assessment the data can be prone to observation errors due to sampling and estimation. These are estimated in the model in addition to process errors. The uncertainty in mixing is not currently accounted for in the model. | The survey indices and catch data are the same as those used for the stock assessment, hence with comparable levels of quality from sampling to estimation.  Diet data are from a long time series of cod stomachs which are known to be geographically and seasonally unbalanced over the entire time period. | Survey indices and catch data are the same as those used for single species stock assessments in the region, hence comparable levels of certainty or bias for the modeled subset of fish.  Diet data are mainly from fish stomachs collected aboard bottom trawl surveys so best represent Fall and Spring diet with high variation between seasons. | Survey indices and catch data are the same as those used for single species stock assessments in the region, hence comparable levels of certainty or bias for fish.  Observations for higher trophic level apex predators are those used for protected species stock assessment which is at the coastwide scale, so are likely to be less certain at the Georges Bank scale.  Diet data are mainly from fish stomachs collected aboard bottom trawl surveys so best represent Fall and Spring diet with high variation between seasons. Diet data for apex predators is assembled from multiple sources including some outside the region and much more uncertain. | Time series of catch and survey data are as far as possible used in line with input to ICES stock assessments in the region, hence with comparable levels of quality as accepted during ICES benchmark workshops. Time series for scarce species are more uncertain than time series for abundant species. The use of biomass time series from single species assessments as “truth” transfers the associated uncertainties and/or bias into the EwE model. Ecopath parameters are not for all functional groups based on literature specific for the Irish Sea. |
| **Hindcast** | | | | | | |
| **Which are the most sensitive parameters and is there room for improvement?** |  | Difficulties in estimating the parameter for the correlation in fishing mortality of the second fleet were encountered at the benchmark and the parameter was therefore fixed. Some convergence problems have occurred since the benchmark due to the difficulty in estimating the variance in the fishing mortality process error. | Predation mortality estimates are particularly sensitive to the choice of predator-prey size selection (i.e., uniform vs. log-normal) and to consumption rates of the predator. Assumption on the “other food” component (constant vs based on an index of overlap of benthic food with hypoxic bottoms) and assumption on residual mortality are also influential on the estimated M. | The reviewed model was not final. However, model structural review suggested evaluating sensitivity to the current fixed other food parameter, different size bin configurations, decoupling model size bins and timestep, estimating vulnerabilities of prey relative to other food for each predator, estimating predator size selection of prey, and reconfiguring fishery selectivity and efficiency by species. | The reviewed model was not final and had no hindcast. However, initial model balance sensitivity to different time series of diet input was low. For dynamic modeling, review recommended the use of Ecosense procedure to generate alternative parameter values for the static model and evaluate fit to historic data performance across a range of initial conditions. | Monte Carlo simulations were run wherein cod production over biomass (PB ratio) was permeated between 0.820 (Ecopath value) and 0.1, with and without the Atlantic Multidecadal Oscillation (AMO) driving the recruitment rate of juveniles to the adult stage. Results suggest that a reduced PB would indeed reduce the recovery rate of cod, however, without the inclusion of the AMO as a driver of cod recruitment rate all simulations overestimated the observed recovery response to reduced fishing effort. |
| **What is the performance of the (final) model hindcasts?** | **How good is the agreement between model output and real-world observations?** | The fit to all observations is generally good. The fit to the catch of the first fleet could be improved. Diagnostics currently include One Step Ahead (OSA) residuals for both observations and processes. Some large residuals are present but overall residuals are acceptable. | The fitting is good and estimates of the main stock metrics (i.e., SSB, F, Recruitment) are comparable with those from the single species assessments. Diagnostics include traditional and One Step Ahead residuals of all fitted data components, sensitivity analyses on a number of model assumptions (i.e., multi-annual aggregation of stomach data, functional form of the predator-prey size selection, residual mortality). | The reviewed model was not final and therefore not expected to have good agreement. However, given the known oversimplified processes in the model, initial model fits were responsive to the input data. Diagnostics currently include One Step Ahead (OSA) residuals and more traditional residual visualizations evaluating fits to indices and compositions. Retrospective analysis and likelihood profiles are planned when the model structure is closer to final. | Rpath static model reviewed with PREBAL, diagnostics were satisfactory. Dynamic model not yet developed. Residuals from fits to indices will be examined for the dynamic model. | Patterns of temporal variability are generally captured (emergent or forced with e.g. recruitment time-series). The Irish Sea model identified six key species: Atlantic cod, haddock, Atlantic herring, whiting, plaice, and *Nephrops*. Using a visual analysis and sum of squared deviation calculation, the trends for Atlantic cod, herring, and *Nephrops* were sufficient. The modeled biomass for whiting was low in the early part of the hindcast but matched well with more recent data. Several ecosystem indicators were examined and model representation is sufficient. Fitting all bumps and drops within a time-series is not necessary as long as the general global assessment of the ecosystem is represented. |
| **How large are the estimated parameter uncertainties?** | The standard deviation of the fixed effect parameters is considered acceptable (varying between 0.07 and 0.53 for the 42 parameters). | CVs of the catch are in the range 0.36-0.59 for herring and 0.38-0.76 for sprat. CVs of the survey range 0.33-0.44 for herring and 0.35-0.51 for sprat.  Number-at-age of the predator cod stock are assumed to be without errors from the stock assessment while they are known to have considerable uncertainty associated. For this reason, uncertainty estimates of the SMS model are considered of limited information. | The reviewed model was not final so parameter uncertainties were not evaluated. | Not relevant until the model is dynamic. | Uncertainty was addressed through a Monte Carlo Markov Chain (MCMC) simulation. Ecopath parameters (e.g., Biomass, PB, diet) have an associated pedigree that allows the user to specify their level of confidence in an input parameter. The Irish Sea model was run 1,000 times with a uniform distribution around input parameters based on data pedigree. Results show substantial uncertainties for biomass of some functional groups, but overall uncertainties were within expected boundaries. |
| **Are there indications for major structural uncertainties (e.g, identified by sensitivity analyses)?** | No, there are no major year or age effects visible in the residuals. | Yes, the value to be assumed as residual mortality is unknown.  Residual patterns in the fitting of stomach data  suggest possible (non-constant) trends in processes such as prey vulnerabilities or availability of “other food”, which are so far not captured by the model. | The reviewed model was not final. However, several structural uncertainties were identified during review including the length bin structure, coupling of length bins with timesteps, predator feeding configuration, and fishery selectivity. These will be evaluated with sensitivity analysis as well as skill assessment with simulated datasets. | The reviewed model was not final. However, review recommended the use of Ecosense procedure to generate alternative parameter values for the static model and evaluate fit to historic data performance across a range of initial conditions. In particular, sensitivity to different diet inputs (1980s vs full time series), could be evaluated using Ecosense. | Yes, WGSAM suggests that additional sensitivity analysis be performed with respect to the PB ratios for groups other than cod in the future. These values could have a fundamental impact on the interpretation of the model. |
| **Are the parameterization and emerging properties from the (final) model sound according to scientific knowledge?** |  | The leave-one-survey-out analysis shows that the stock perception is mainly driven by one survey that brings the stock down (increase in SSB when leaved out). The survey indices are considered accurate since this survey sampled most of the stock distribution. | Yes, natural mortality in the multispecies model includes two components: 1) a constant residual mortality (M1) and 2) a time variable cod predation (M2). Estimates of natural mortality at age emerge from the abundance and age composition of both cod and clupeids. | The reviewed model was not final. However, planned skill assessment with simulated datasets is designed to address this. | Yes, Rpath static model reviewed with PREBAL and found to be sound for general thermodynamic properties of food webs. | Yes, all of the PREBAL diagnostics (e.g., biomass slope around 5 - 10% decline with increasing trophic level, biomasses span 5-7 orders of magnitude, ratios between predator and prey between 0.01 and <1, production to biomass ratio and consumption to biomass ratio declines with trophic level, production to consumption ratio below 1 for all functional groups and between 0.1 and 0.3 for finfish) were within reason.  Ecosim stability tests show expected model behavior at extreme fishing scenarios to perturbate the ecosystem. |
| **Are there retrospective patterns?** |  | Yes, but the retrospective patterns are acceptable but could be improved with Mohn’s rho in SSB of 16% in 2023. | Analytical retrospective analysis shows only minor patterns in the SSB, R, F and very consistent estimates on natural mortality. | Not yet tested as the model is still being developed, but this diagnostic will be used in the future. | Not applicable for static model; could be tested with dynamic model. | No retrospective analysis was performed. However, when the procedure for using model outputs in the management process is more clearly specified, the model’s ability to provide stable estimates of ecosystem state over a range of data endpoints could be evaluated. |
| **Forecast** | | | | | | |
| **What is the predictive skill of the (final) model?** | **Do predictions show expected model behavior?** | Yes, the forecast is built within SAM and has the same structural assumptions as the hindcast (e.g. multi-fleet). It behaves as expected depending on the assumptions taken regarding fleet selectivity, biological input, recruitment, process errors, etc. | Natural mortality estimates are only used as input for the historic part and no forecast is needed. | Not yet tested; short term forecast skill similar to single species stock assessment would be the relevant test. | Not yet tested. | No forecast is needed to answer the advice question in place. |
| **How does the model perform in cross validations?** | Not yet tested. | N.A. | Not yet tested. | Not yet tested. | N.A. |
| **How large are the estimated uncertainties?** | The forecast is deterministic so does not account for any uncertainty. | N.A. | Not yet tested. | Not yet tested, but review recommended the use of Ecosense procedure to generate alternative parameter values for the static model and evaluate forecast performance across a range of initial conditions. | N.A. |
| **Are there indications for major structural uncertainties (e.g, identified by sensitivity analyses)?** | Yes, the forecast is built within SAM and has the same structural assumptions as the hindcast (e.g. multifleet) except that the process errors are turned off. | N.A. | Not yet tested. | Not yet tested, but review recommended testing sensitivity to different diet inputs (1980s vs full time series), which could be projected using Ecosense. | N.A. |
| **How is the performance of short-, medium- and long-term forecasts?** | The forecast is run assuming a catch constraint in the intermediate year and the ICES advice rule in the following year. Some assumptions have to be made regarding stock split fleet behavior in the forecast, which makes it particularly uncertain. | N.A. | Not yet tested; short term forecast skill similar to single species stock assessment would be the relevant test. | Not yet tested, but all will be important for characterizing wider ecosystem responses to fishery management measures. | N.A. |
| **Judgment by stakeholders and use of model output** | | | | | | |
|  |  | The current multi-fleet SAM configuration was reviewed and validated in 2018 during an ICES benchmark (ICES WKPELA 2018).  The model is updated every year with new observations and the short-term forecast provides future fishing opportunities under certain scenarios. After the assessment and short-term forecast have been audited, an advice drafting group produces the advice for Western Baltic herring. The advice has to be finally agreed by the ICES Advisory Committee and is used afterwards by managers as a basis for negotiations.  The diagnostics discussed here concern the 2023 assessment model and catch advice for 2024. | The Baltic Sea SMS key-run has been internally reviewed by ICES WGSAM (2023). The model has been resilient to several updates (ICES WGSAM 2012, 2019, 2023). In 2019 comparison of predation mortality estimates between SMS and a Gadget model showed overall good consistency which contributed to build further confidence in the key-run estimates (ICES WGSAM 2019).  In 2022/2023, the key-run results were presented and discussed at the single-species stock assessment benchmark of Baltic herring and sprat (ICES WKBBALTPEL 2023). Here, alternative scaling of the residual mortality was applied which resulted in multiple time series of natural mortalities that were used to serve an ensemble of stock assessment models. | Not final yet. | Not final yet. | Feco is provided as a catch option for cod in the Irish Sea:<https://ices-library.figshare.com/articles/report/Cod_Gadus_morhua_in_Division_7_a_Irish_Sea_/19447895>.  *Regarding Feco in general, fishers involved in the North Western Advisory Council were supportive (NWWAC, 2022). Several decision makers expressed encouragement and were looking forward to further work with ICES on EBFM (ICES MIRIA 2023).*  *ICES WKREF2 (2022) validated the use of Feco for advice under certain conditions and recommended that ICES guidelines include the possibility to use an Feco approach to adjust the target F based on ecosystem model information.* |

**References:**

Bentley, J. W., Lundy, M. G., Howell, D., Beggs, S. E., Bundy, A., De Castro, F., Fox, C. J., Heymans, J. J., Lynam, C. P., Pedreschi, D., Schuchert, P., Serpetti, N., Woodlock, J., & Reid, D. G. (2021). Refining fisheries advice with stock-specific ecosystem information. Frontiers in Marine Science, 8. <https://doi.org/10.3389/fmars.2021.602072>

Gaichas, S. K., Fogarty, M., Fay, G., Gamble, R., Lucey, S., & Smith, L. (2017). Combining stock, multispecies, and ecosystem level fishery objectives within an operational management procedure: Simulations to start the conversation. ICES Journal of Marine Science, 74(2), 552–565. <https://doi.org/10.1093/icesjms/fsw119>

ICES HAWG. (2023). Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG). ICES Scientific Reports, 5:23 <https://doi.org/10.17895/ices.pub.22182034.v1>

ICES WGSAM. (2012). Report of the working group on multispecies assessment methods (WGSAM), 22-26 October 2012, Venice, Italy. ICES CM 2012/SSGSUE:10.: 145 pp.

ICES WGSAM. (2019). Report of the Working Group on Multispecies Assessment Methods (WGSAM). ICES Scientific Reports, 1:91, 320 pp. <http://doi.org/10.17895/ices.pub.5758>

ICES WGSAM. (2023). Working Group on Multispecies Assessment Methods (WGSAM; outputs from 2022 meeting). ICES Scientific Reports. 5:12. 233 pp. <https://doi.org/10.17895/ices.pub.22087292>

ICES WKBBALTPEL. (2023). Benchmark Workshop on Baltic Pelagic stocks (WKBBALTPEL). ICES Scientific Reports. 5:47. 350 pp. <https://doi.org/10.17895/ices.pub.23216492>.

ICES WKPELA. (2018). Report of the Benchmark Workshop on Pelagic Stocks (WKPELA 2018), 12–16 February 2018, ICES HQ, Copenhagen, Denmark. ICES CM 2018/ACOM:32. 313 pp. https://doi.org/10.17895/ices.pub.5432

Link, J. S. (2010). Adding rigor to ecological network models by evaluating a set of pre-balance diagnostics: A plea for PREBAL. *Ecological Modelling*, *221*(12), 1580–1591. <https://doi.org/10.1016/j.ecolmodel.2010.03.012>

Lucey, S. M., Gaichas, S. K., & Aydin, K. Y. (2020). Conducting reproducible ecosystem modeling using the open source mass balance model Rpath. Ecological Modelling, 427, 109057. <https://doi.org/10.1016/j.ecolmodel.2020.109057>

Nielsen, A., Hintzen, N. T., Mosegaard, H., Trijoulet, V., & Berg, C. W. (2021). Multi-fleet state-space assessment model strengthens confidence in single-fleet SAM and provides fleet-specific forecast options. ICES Journal of Marine Science, 78(6), 2043-2052. <https://doi.org/10.1093/icesjms/fsab078>

Rindorf, A., Cardinale, M., Shephard, S., De Oliveira, J. A., Hjorleifsson, E., Kempf, A., Luzenczyk, A., Millar, C., Miller, D. C., Needle, C. L., Simmonds, J., & Vinther, M. (2016). Fishing for MSY: Using “pretty good yield” ranges without impairing recruitment. ICES Journal of Marine Science, 74(2), 525-534. <https://doi.org/10.1093/icesjms/fsw111>

ICES WKREF (2022). Workshop on ICES reference points (WKREF2). ICES Scientific Reports. Report. <https://doi.org/10.17895/ices.pub.20557008.v1>

ICES Miria. (2023). Report of the Annual Meeting between ICES and Requesters of ICES Advice (MIRIA), 2023. ICES Business Reports. Report. <https://doi.org/10.17895/ices.pub.22154555.v1>

NWWAC. 2022. Minutes of the March meeting. <https://www.nwwac.org/_fileupload/Minutes%20and%20Reports/2022/March%202022/Minutes_WG1_March_2022.pdf>