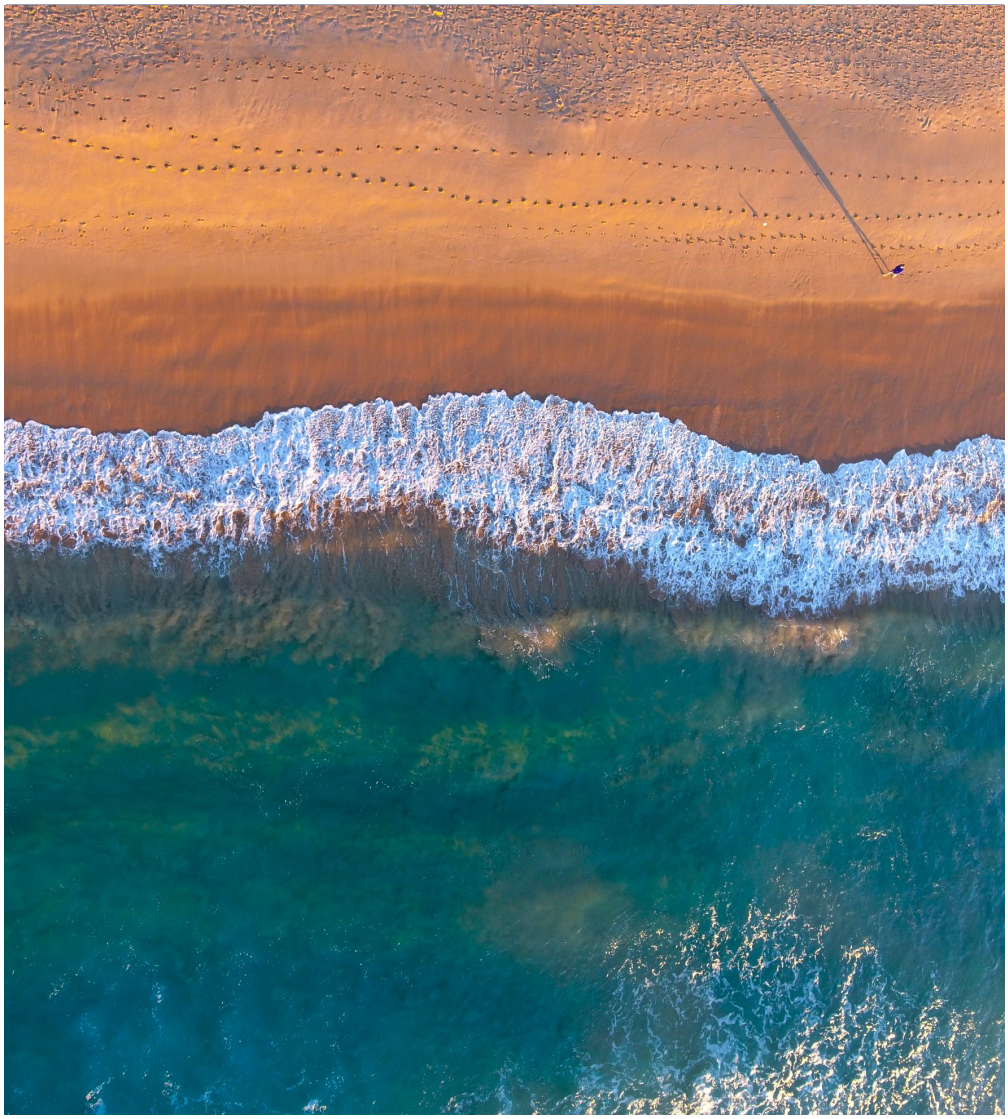
Advice Technical Guidelines

ICES Guidelines for Benchmarks

Version 1 I March 2023

ICES GUIDELINES AND POLICIES



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

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

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

This document is approved by the ICES Advisory Committee and produced under the auspices of the International Council for the Exploration of the Sea.

ICES Guidelines and Policies - Advice Technical Guidelines

ICES Guidelines for Benchmarks

Version 1 I March 2023

Recommended format for purpose of citation:

ICES. 2023. ICES Guidelines for Benchmarks. Version 1. ICES Guidelines and Policies - Advice Technical Guidelines. 26 pp. https://doi.org/10.17895/ices.pub.22316743

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

Contents

i Background i
ii Key points ii
1 Principles and elements of a benchmark system 3
1.1 Elements of a benchmark system 4
2 Identification of the need for a benchmark 5
3 Types of benchmark processes 6
3.1 Expert Group level 6
3.2 Review 6
3.3 Full benchmark 7
3.4 Prioritization Process 11
3.4.1 Oversight and Ownership 11
3.5 Communication 12
3.5.1 Internal communication 12
3.5.1.1 Prior to the process: 12
3.5.1.2 During the process: 13
3.5.1.3 At conclusion of the process: 13
3.5.2 External Communication 13
3.5.2.1 Prior to the process: 13
3.5.2.2 During the process: 13
3.5.2.3 At conclusion of the process: 13
3.6 When a problem with a benchmark is encountered 14
4 References 15
Annex 1: Flow diagrams for different benchmark processes 17
Annex 2: Decision matrix for ICES benchmark prioritization 20
Annex 3: ICES Benchmark toolbox for model diagnostics 22
Annex 4: Detailed version history 26

i Background

Guideline scope

The guidelines describe the principles, elements, purpose, types, and prioritization of ICES benchmarks. The information in this document is relevant for the ICES expert groups delivering scientific evidence for benchmark processes and reviews but also for interested policy makers, stakeholders and the wider scientific community.

Changes since the last version

Location	Change description
NA	*This is the first version of these guidelines*

Other relevant information

These Guidelines were written by the Benchmark Oversight Group (BOG) 2022/2023. The Guidelines for Benchmarks are curated by the Benchmark Oversight Group, and reviewed and approved by ACOM.

ii Key points

- ICES uses a 'benchmark process' as a means to peer-review and incorporate new science for use in provision of all types of recurrent advice. The benchmark process is a critical element in ICES advice to ensure a sound scientific basis.
- There are three types of benchmark process – **Expert Group level**, **Review**, and **Full benchmark**. The work flow for each type is illustrated in Annex 1.
  - **Expert group level**: Adequate for small changes to the method that are mainly technical in nature. For ecosystem services and effects advice, the Expert Group review will require coordination among all Expert Groups involved in generating the evidence base for advice.
  - **Review:** Adequate when addressing one or two larger issues such as changing/correcting an entire data series, or for more substantive revisions to model setting such as changes to age ranges or natural mortality assumptions. This process will, in general, require one year to complete; Expert Groups should carefully consider their workload and their ability to work within the time frame. Peer review will be done by a Review Group composed of members external to the Expert Group(s).
  - **Full benchmark:** A full benchmark process is a full review of methods, underlying conceptual assumptions, and data; it can also provide the technical basis for the provision of new advice. This review must include an evaluation of the appropriateness of the chosen method. Annex 3 provides guidance for category 1 stock assessments, but the content may be useful for other benchmarks as well. As the full benchmark is a multi-stage process and includes members from outside a specific Expert Group, it is not associated with an Expert Group meeting. It is vital that the process be finished far enough in advance of the Expert Group meeting to allow for full documentation to be ready for the Expert Group, and to allow time for review by the Benchmark Oversight Group (BOG) and ACOM.
- To plan ahead and utilize network resources in the best way possible, a prioritization process is in place. It considers aspects such as benchmark preparedness, perceived risks, and time since last benchmark, among others. The same factors should be considered by the Expert Group(s) as well as their own resource capacity when embarking on an Expert Group level or review level process. A detailed prioritization scheme is presented in Annex 2.
- The prioritization and approval of full benchmark processes is done by ACOM following recommendations from BOG. A full benchmark process will, in general, require more than one year from proposal to completion and approval.
- As with all aspects of ICES advice, ACOM is ultimately responsible for the outcome of benchmarks and as such has the responsibility to decide which benchmarks are conducted and to review and approve benchmark outcomes. There is an important ownership role for Expert Groups, however, as they will be responsible for implementing the outcome of benchmarks and should be important contributors to benchmarks from start to finish.

1 Principles and elements of a benchmark system

Two key elements of the ICES approach to providing advice are:

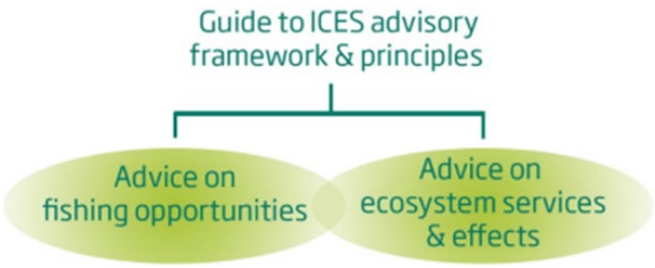
i. incorporating new knowledge into the advisory process to contribute effectively to the creation of advice on meeting conservation, management, and sustainability goals, and
ii. assuring that quality encompasses the entire process from data collection to the publication of objective and independent advice.

ICES uses a 'benchmark process' as a means to peer-review and incorporate new science (new knowledge, data, analyses, and assessment methods) for use in provision of recurrent advice in response to regular annual requests. Methods and data series used in recurrent advice are expected to be valid for several years (often between five and ten). ICES must, however, remain responsive to changes in data, ecosystem, fishery, and model performance; this could result in a more frequent benchmark timing.

The benchmark process is a critical element in ICES advice to assure a sound scientific basis. During the process ICES depends on the willingness of independent experts to serve as peer-reviewers. The peer-review and benchmark process is vital to ensure that ICES advice continues to be based on best available science.

The ICES advisory framework and principles are set out for both fishing opportunities and ecosystem services and effects (ICES, 2023). The ICES Advisory Plan establishes the ecosystem approach as the central tenet governing how ICES provides independent advice on the management of human activities in our seas and oceans. Principle 7 sets out the need for peer-review and the benchmark process.

Principle 7. *To ensure that the best available, credible science has been used and to confirm that the analysis provides a sound basis for advice, all analyses and methods are peer reviewed by at least two independent reviewers. For recurrent advice, the review is conducted through a benchmark process; for special requests through one-off reviews.*



In addition to the Advice principles, ICES recognizes the following Principles for Benchmarks:

1. Adequate peer review is a cornerstone of ICES advice.
2. ICES is using a flexible benchmark system tailored to the current advice requests; it is based on the strong elements of the former 2016 benchmark proposal as well as the current benchmark process.

3. The scope of ICES benchmarks is to review data and methods in support of the production of all types of ICES recurrent advice where the approaches can be expected to be valid for some period of time.
4. Benchmarks are to be conducted at the scale that is most appropriate given the issues to address. Benchmarks conducted at the regional scale are desirable, but not always possible.
5. Appropriate oversight and a sense of ownership is applied to integrate benchmarks in the ICES advisory system.
6. Benchmarks are prioritized to ensure peer-review efforts are best placed.

1.1 Elements of a benchmark system

1. Identification, prioritization, and approval of benchmarks as per the approach approved by ACOM in 2020 (ICES, 2020).
2. A benchmark process that includes scoping (with identification of resources and deadlines), data evaluation workshops, progress meetings, and a final benchmark workshop tailored to the issues to be addressed by the process.
3. Increased oversight and support for benchmarks achieved through a Benchmark Oversight Group (BOG).
4. Review of benchmark processes.
5. Formal endorsement or identification of remedial measures by ACOM for completed benchmark processes.

The benchmark process applies to all recurrent advice, and it has been most regularly applied to fishing opportunities advice. There are clear management objectives for fishing opportunities advice (management plans or the ICES MSY and the precautionary approach framework). For ecosystem services and effects advice the definition of management objectives is often less prescribed and clear. The benchmark process in this context, used in dialogue with managers, can guide the identification of incrementally-achievable objectives based on best available and developing knowledge. A clear distinction between presently-achievable objectives and aspirational objectives will assist with clarifying the narrative in ecosystem-informed advice. ICES is committed to providing the evidence base to inform management decisions across the suite of pressures resulting from a range of human activities.

There is a very broad portfolio of scientific disciplines available to the ICES network, and this expertise is used to construct the best available knowledge for advice. It ensures that this advice is developed considering services and effects across appropriate spatial and temporal scales. Method development often involves dialogue with requesters and stakeholders. The roles and responsibilities of researchers and stakeholders in that dialogue are described in the stakeholder engagement strategy.

Recurrent advice for both ecosystem services and effects and for fishing opportunities should follow a similar benchmark process, but with some important differences. One aspect of ecosystem services and effects advice is that multiple Expert Groups will often be involved in delivering the scientific basis for advice. While new frameworks for ecosystem-informed advice are being developed, there is often a requirement for an iterative, review-type benchmark approach involving multiple Expert Groups. The guidelines here are meant to apply to all forms of recurrent advice, but will require modification in some cases.

2 Identification of the need for a benchmark

Most benchmark proposals will come from the groups responsible for the analysis used to provide the advice, usually Expert Groups. Meetings of Expert Groups are tasked with reporting on data and analyses quality, and with maintaining issue lists. This information should be used by the Expert Group when proposing which advice requires a benchmark process and which process should be pursued. Expert Groups should consider their resource capacity when proposing a benchmark process.

ACOM may also recommend a benchmark be conducted and Advice Drafting Groups (ADG) may recommend benchmarks to ACOM. For ecosystem services and effects advice, ACOM may be the main source in the identification of the need for a benchmark. If an Expert Group delivering part of the scientific advice determines that there is a need for a benchmark process, ACOM or the relevant ADG should be informed.

3 Types of benchmark processes

There are three main types of benchmark processes which differ in their complexity and type of review (Table 1; flow diagrams in Annex 1):

- Expert Group level;
- Review; and
- Full Benchmark.

The Expert Group will conduct the initial determination of the type of process that is needed, and this should be considered carefully. The work will be reviewed at the ADG and ACOM during the approval of advice. If the work exceeds the scope of the particular process it will be rejected.

3.1 Expert Group level

In advice production, Expert Groups review and, where considered appropriate, may implement small changes to the method that are mainly technical in nature. At this level, the review is conducted completely within the Expert Group, with the Expert Group members providing the peer review. Final review is the exception, and this as with all advice occurs within the ADG and ACOM. This process addresses issues such as small adjustments to model settings, or revisions of one or two years of data. The process will generally be completed within the Expert Group meeting and applied to the advice for that current year.

Issues identified by an ADG or ACOM during the drafting/approval of advice can be addressed in the following year's Expert Group. The change to the method or data needs to be fully documented in the Expert Group report and the technical documentation (e.g. benchmark method, stock annex) which gives the details of the method/data must be updated. The Expert Group chair and the person leading the development of the advice (advice leads, e.g. stock coordinator in fishing opportunities, currently ACOM leadership for ecosystem services and effects advice) must ensure that these changes to the methods are fully documented in advance of the ADG and that the ADG is informed of the changes.

For ecosystem services and effects advice, the Expert Group review will require coordination among all Expert Groups involved in generating the evidence base for advice. This will be necessary to understand and document the potential consequences of technical changes made by one Expert Group/in one science input, for other Expert Groups/science inputs or outputs.

3.2 Review

A review level process addresses either one or two larger issues such as changing/correcting an entire data series, or more substantive revisions to model setting such as changes to age ranges or natural mortality assumptions. The issue is identified by an Expert Group (or by multiple Groups), by ADG, or ACOM. This process will generally require one year to complete. A review process is meant to address the method/data for a single advice product and the changes applied in subsequent years. Because this process addresses larger issues than the Expert Group Level process, it will not be possible to complete the process before the current year ADG. Rather, members of the Expert Group(s) will work intersessionally via correspondence to address the issue before their next meeting. A Review Group composed of members external to the Expert Group(s) will provide peer review and report back to the Expert Group(s). The Review Group

will have a chair that will coordinate their work and ensure that the report of the group is available at least one month prior to the Expert Group meeting. This will allow time for the Expert Group(s) to evaluate it prior to the meeting at which the change is meant to be implemented. The BOG and ACOM should be informed of the issue and planned work for the review group before the process is initiated. The Expert Group chair and advice lead will work with ICES Secretariat to identify reviewers and to ensure that summaries, overviews, and details of the proposed assessment changes are available to the Review Group. Stakeholders will be notified of review processes and invited to attend if they have relevant information. Working documents should be available to the Review Group at least two weeks prior to their meeting. Any changes to methods or data recommended by the Review Group need to be fully documented in the Expert Group report and the technical documentation (e.g. stock annex) which gives the details of the method/data must be updated. The Expert Group Chair(s) and advice leads must ensure that these updates to documentation are complete in advance of the ADG and that the ADG is informed of the change. Expert Groups should consider their workload and their ability to complete the required work within the time frame.

3.3 Full benchmark

A full benchmark process is a full review of methods, underlying conceptual assumptions, and data; it can also provide the technical basis for the provision of new advice. This review must include evaluation of the appropriateness of the chosen method. Annex 3 provides guidance for category 1 stock assessments but the content may be useful for other benchmarks as well. Convergence and residual analyses, for example, should be standard for most model evaluations. The process may require more than one year to complete. The timing is independent of the timing of Expert Group meetings, with the main consideration being the completion and review of all work. As is the case for the other types of benchmark processes, the need for a full benchmark may be identified by the Expert Group(s), ADG, or ACOM. The Expert Groups must scope the problem (issues list) and possible solutions, as well as identify potential reviewers. The prioritization table (Annex 2) must be filled out, including enough detail to allow BOG to evaluate. BOG will evaluate all proposed benchmarks for a given year at the same time. When making a proposal, Expert Groups should keep in mind that benchmarks that have investigated the solutions prior to the initiation of the process are more likely to be successful and that preparedness will be considered in the prioritization (see section below on Prioritization Process).

Multiple advice products may be considered (e.g. grouping of stocks that have similar life histories or that have similar methods proposed, or models that could be used in different advisory products) in a single benchmark process. A single advice product can be put forward by an Expert Group and may then be grouped as appropriate by BOG, in conjunction with the relevant Expert Groups and ICES Secretariat. Each year, BOG will recommend benchmarks for ACOM approval. Once approved, advice leads will work with the Secretariat and BOG to draft Terms of Reference (ToRs) for approval. Many benchmarks will involve the evaluation of new/revised data, therefore data calls should be made very early in the process. Data evaluation workshops (or other workshops as required) should be held to review the available knowledge, data, and any associated methods. If stock identification is an issue, for example, a stock ID workshop will be scheduled appropriately. The benchmark workshop itself, as much as possible, should be used to review developed methods and fine tune if necessary, rather than first development/application.

As the full benchmark is a multi-stage process and includes members from outside a specific Expert Group, it is not associated with an Expert Group meeting. It is vital that the process be finished far enough in advance of the Expert Group meeting to allow for full documentation to

be ready for the Expert Group and to allow time for review by BOG and ACOM. Specific workshops (e.g. data evaluation or stock ID workshops) should be held long enough in advance (one or two months at least) of the methods part of the benchmark process to allow the proposed methods to be applied prior to the final workshop (benchmark workshop). The final benchmark workshop, including the full benchmark report, must be completed far enough in advance of the Expert Group at which the results are meant to be applied to allow for BOG recommendation and ACOM approval. This should be 2 months prior to the Expert Group meeting. Full documentation is necessary to give BOG and ACOM the required background. Although the final schedule for a full benchmark should be built around the timing of the Expert Group(s), the proposals need to be made at the same time during the year so that prioritization can occur (see Section 3.4 below, on the *Prioritization Process*).

Full benchmarks will generally take place in a hybrid format. There will be two chairs, one considered an ICES chair (someone who is more familiar with ICES processes) and the other an external chair, who may be external to ICES but must at least be external to the relevant Expert Group. Members of the benchmark should be comprised of relevant Expert Group members, including the advice lead(s), and other experts outside the relevant Expert Group. Reviewers will generally also attend the methods meeting. A professional officer (PO) from the Secretariat will also attend. Stakeholders will be invited to attend and are particularly valuable in the data evaluation part of the benchmark.

Chairs will guide the process, ensuring that progress is evaluated and that any problems are reported to ACOM, who will decide on the need for any possible postponement or for cancellation – see Section 3.6, entitled *When a problem with a benchmark is encountered*. As part of this process it is advisable for the chairs to arrange a series of shorter, issue specific, meetings to review work as it is being done to ensure that any problems are detected early. Chairs will aid in the preparation of the benchmark report and help ensure that it is completed on time (at least two months prior to the Expert Group meeting) and that the work of the meeting is fully documented. Although the chairs work as a team, there are two roles that are specific to each chair. The ICES chair helps ensure that the meeting follows ICES guidelines, while the external chair leads the review group.

Benchmark members (including advice leads) are responsible for completing the work required to address the ToRs and for fully documenting that work. This includes completing the relevant benchmark report sections, entering stock assessments into the ICES Transparent Assessment Framework (TAF), and updating any existing technical documentation once the new methods are approved by ACOM. For advice on fishing opportunities, this includes updating the stock annex. For ecosystems services and effects advice the benchmark report will serve as the technical documentation. The benchmark report is a vital document. It must detail all the work considered, including work that is not accepted by the benchmark. The final accepted method(s) and data must be given in enough detail to allow the reader to know what final model and data have been accepted by the benchmark, along with the rationale for that. This is best achieved by having a final conclusion section for each advice product (e.g. each stock, or type of ecosystem advice) that details the accepted method(s) and data.

Reviewers, led by the external chair, are responsible for providing a written review (within one week of the completion of the final meeting) detailing their evaluation of the process. They should ask questions as the process progresses in order to allow any concerns to be addressed or clarified, and document both the questions and the resulting response in their report. The Review group must complete a written review within one week of the conclusion of the process.

The PO helps draft the ToRs, helps to arrange chairs and reviewers (the benchmark proponents should provide the names of possible candidates for both), and helps address questions from participants about process (the format of reports and timelines, for example). The PO also produces a summary of the benchmark for BOG which outlines the conclusions and highlights any issues. This report should be available within one week of the completion of the process.

Table 1 **Scope and nature of the different review processes.**

Type of process	Types of issues	Nature of process
Expert Group Level	Peer-review small issues that are mostly technical in nature such as: Fixing issues with input data (generally one or two data points rather than entire time series unless issue is minor) Small adjustment to model settings Software updates Year range for which future recruitment are averaged/sampled from Changes to forecast assumptions related to biology (e.g. weight, maturity), selectivity Corrections to reference point calculations if there are errors found How the information is presented or mapped Addition of new scenarios	Issues identified by Expert Group(s), ADG or ACOM Occurs during Expert Group(s) Change needs to be fully documented in EG report Change should be reported to ADG – if the issue is deemed by ADG to be not appropriate for Expert Group level process this should be identified to ACOM Technical documentation (e.g. stock annex) must be updated Each process will review a single advice method only
Review Level	One or two larger issues: Correcting an entire data series or revision to procedures/criteria used to estimate one index Revision to model settings/assumptions that is more substantive (e.g. changing the age range used in the assessment for a well-defined and documented survey series or letting the last age group in a tuning series be a plus group and not a true age group) Incorporating new M (note if M is being derived from multispecies models the new multispecies runs should align with benchmarks if possible) Review of reference points, including change to one or two assumptions in the calculation of reference points	Issues identified by Expert Group(s), ADG or ACOM BOG and ACOM informed of issues and planned work Expert group(s) works with Secretariat to identify reviewers external to Expert Group(s) Work done intersessionally Working documents prepared in advance of next Expert Group meeting(s) Review group provides review to Expert Group(s) in advance of their meeting(s) Review report evaluated by Expert Group(s) prior to their meeting(s) By correspondence Changes fully reported in Expert Group report(s) with working documents and reviewer reports included as annex Technical documentation (e.g. stock annex) must be updated Each process will review a single advice method only
Full Benchmark	Full review of method, underlying conceptual assumptions and data or provision of new advice: New data series may be incorporated Current indices reviewed New analytical method may be introduced Challenge to major structural assumptions of the model Stock id Often multiple advice (for example several stocks of the same species, area or advice using similar methods) grouped Incorporation of ecosystem effects on population dynamics	Multistage process not associated with Expert Group meeting(s) but aligned to finish not less than 2 months prior to Expert Group meeting(s) ICES and external chair Reviewers Often hybrid For fishing opportunities stock should be entered into TAF Benchmark report reviewed by BOG to recommend to ACOM Technical documentation (e.g. stock annex) must be updated if changes adopted

3.4 Prioritization Process

While the impetus for benchmarks should mostly come from the Expert Groups, there needs to be a process to look across all Expert Groups in order to prioritize. The prioritization scheme needs to consider the following six points.

1. Need to improve the quality of the analyses used to provide advice
2. Opportunity to improve the assessment (e.g. commitment of resources, availability of new or corrected data, new methods for the advice)
3. Benchmark preparedness
4. Perceived risks
5. Changing ecosystem and ability to include impact in advice
6. Time since last benchmark

The Expert Group(s) should consider these factors. as well as their own resource capacity, when embarking on an Expert Group level or review level process.

A detailed prioritization scheme is presented in Annex 2.

The prioritization and approval of full benchmark processes is done by ACOM following recommendations from BOG. The scores in the prioritization scheme used to prioritize full benchmark proposals should be checked for consistency and to ensure that the priority is given to aspects essential to provision of advice. In general, a schedule for the highest priority benchmarks should be completed first and according to the schedule proposed by the Expert Group, with lower priority processes added in where possible. Final benchmark workshops should, as much as possible, be scheduled so that the results are available for the upcoming assessment (i.e. no long gap between the benchmark and the next assessment which would prompt stakeholders to request an updated assessment) but with sufficient time for approval and implementation.

A full benchmark process will generally require more than one year from proposal to completion and approval. ACOM must, therefore, plan ahead when prioritizing. The prioritization scheme aids in completing a tentative list of benchmarks. In each year, proposals from the Expert Groups that include the scoring for prioritization and initial scoping (encompassing the list of issues and solutions) will be reviewed by BOG. BOG will make recommendations to ACOM, and a list of benchmarks will be established.

Proposals for benchmarks should be made directly to BOG. Because of the need to prioritize full benchmarks, the proposals must all be made to BOG according to a specified timeline. Currently, this is no later than November 30. Time is required for BOG to work with the benchmark proponents and Secretariat in combining proposals into a manageable number of processes.

If there are more full benchmarks proposed than the network can accommodate, then a pool system will be established. The prioritized list is divided into high, medium, and low priority proposals, with the lowest priority eliminated or rescheduled for another time.

3.4.1 Oversight and Ownership

As with all aspects of ICES advice, ACOM is ultimately responsible for the outcome of benchmarks and as such has the responsibility to decide which benchmarks are conducted; ACOM also reviews and approves benchmark outcomes. To aid in this task ACOM formed a Benchmark Oversight Group (BOG) to focus on benchmark issues with the goal of providing

background information, analyses, and recommendations for the consideration of ACOM on benchmarks. Specifically, the purpose of the BOG is to:

i. Develop solutions to address generic issues with benchmarks.
ii. Prepare proposal of the list of benchmarks to be conducted for approval by ACOM.
iii. Review completed benchmarks and recommend remedial actions as necessary.

The BOG should be composed of the Fisheries Resources Steering Group (FRSG) chair, a member of the ACOM Leadership, at least one ACOM member (preferably two), a SCICOM member, a Professional Officer from the ICES Secretariat, and other members BOG requires to ensure adequate review of the benchmark process. BOG should report at the annual ACOM meeting, at the consultations held during the Annual Science Conference, and as necessary through the ACOM Forum. BOG proposes actions for approval by ACOM, and relays benchmark issues that need further exploration and consideration.

There is an important ownership role for Expert Groups. Members will be responsible for implementing the outcome of benchmarks and should be important contributors to benchmarks from start to finish. The Expert Group level process is run entirely by the Expert Group and the review level process is run by the Expert Group in cooperation with the Secretariat.

3.5 Communication

Regular internal- and external communication is an important contributor to benchmark success. The key points of this are the timely completion of reports, an awareness of the full process by Expert Group members, engagement of all parts of the network, as well as outreach to stakeholders and requesters of advice.

3.5.1 Internal communication

3.5.1.1 Prior to the process:

- Ensure the description of the benchmark process is given to all Expert Group chairs and include this as a link on all of the Expert Group SharePoint sites – *Secretariat.*
- Review the benchmark process at WGCHAIRS – *ACOM Leadership.*
- Develop, and update as necessary, a presentation to be given at the beginning of relevant Expert Group meetings – *BOG and Secretariat.*
- Inform members of Expert Groups about benchmark processes – *Expert Group Chairs.*
- Communicate with, and through, the relevant steering groups (e.g. FRSG, HAPISG, IEASG) and other Expert Group chairs on data issues, new knowledge and information, and appropriate methods that could help with a proposed benchmark – *Expert Group chairs, SG chairs.*
- Communicate proposal for benchmark to BOG before the November 30 deadline – *Expert Group Chairs.*
- Communicate with chairs of Expert Groups and responsible experts (e.g. stock assessors) regarding missing issues lists and prioritization scores – *Secretariat.*
- Communication with benchmark chairs, reviewers, Expert Group chairs (both advisory and science), stock assessors, SG chairs, and ACOM Leadership relating to the organization, issues and outcomes of benchmark processes including agreed timelines – *Secretariat.*
- Communicate to ACOM the rationale for decisions on which benchmark processes should proceed – *BOG.*

- Communicate clear decisions to Expert Groups on which benchmark processes are and are not approved – *ACOM Leadership.*

3.5.1.2 During the process:

- Communicate early with the participants of the scoping, data evaluation, and benchmark workshops about the process and general organization issues – *Benchmark chairs.*
- Facilitate timely completion of benchmark report – *Benchmark chairs, POs.*
- Communicate with the PO assigned to the benchmark and Expert Group chairs about any issues or difficulties that might impact on the benchmark success – *Advice Leads/Responsible experts.*

3.5.1.3 At conclusion of the process:

- Document any work and complete the relevant report sections on time (at least two months prior to Expert Group meeting) – *Advice Leads.*
- Update any technical documentation (e.g. stock annex, TAF) – *Advice Leads.*
- Complete PO report of the benchmark process within one week of the end of that process – *POs.*
- Complete written review of benchmark within one week of the end of the process – *Review group.*
- Communicate to ACOM the rationale for recommending either approval or rejection of the benchmark results – *BOG.*
- Communicate clear decisions to Expert Groups on approval and rejection of benchmark results – *ACOM Leadership.*

3.5.2 External Communication

3.5.2.1 Prior to the process:

- Communicate the list of benchmarks (including the stocks to be reviewed) to advice requesters and stakeholders. This should include the list of issues to be considered – *Secretariat, ACOM Leadership.*
- Details on the location, dates, times, chairs, and ICES contact information for the benchmark should be provided when available; data evaluation workshops should be specifically highlighted – *Secretariat.*
- Solicit feedback from advice requesters on any particular concerns about specific benchmarks on the list – *Secretariat.*
- Issuing of invitation to attend and provide information/observations considered relevant to the review level process or benchmark – *Secretariat.*

3.5.2.2 During the process:

- Communicate any changes to timelines and/or cancellation of benchmarks to advice requesters and stakeholders – *Secretariat.*

3.5.2.3 At conclusion of the process:

- Communicate the development of new methods from all of the processes to advice requesters and stakeholders when the new advice is released – *Secretariat, ACOM Leadership.*

3.6 When a problem with a benchmark is encountered

A benchmark could encounter difficulties for a number of reasons. The anticipated resources may not materialize so that the process is not complete, there might not be required consensus reached among participants, or the proposed approach may not be successful in solving the issues. Any difficulties should be reported to BOG by the benchmark chairs; a suggested way forward should be included, if possible. These will be relayed to ACOM for evaluation and could result in the postponement or cancellation of a benchmark process.

Unless the proposed way forward can be easily achieved, the item (assessment/model/analysis) should be removed from the benchmark priority list until the Expert Group or proposer of the benchmark makes another full proposal for the benchmark. In such a case the Expert Group/proposer should identify how the new proposal differs from the original, particularly with respect to what has changed that improves the probability of a successful process.

If possible, the existing basis for advice is maintained until the issues can be re-examined and resolved. If it is not possible to keep the existing basis, advice may be provided using simpler approaches (e.g. empirical methods in the case of advice for fishing opportunities (ICES, 2022a)) or the advice may be postponed.

In all cases of benchmark failure ACOM is to approve the next steps, including either the extension of the benchmark process or its termination.

4 References

Carvalho, F., Punt, A.E., Chang, Y.-J., Maunder, M.N., and Piner, K.R., 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fisheries Research 192, 28–40. https://doi.org/10.1016/j.fishres.2016.09.018

Carvalho, F., Winker, H., Courtney, D., Kell, L., Kapur, M., Cardinale, M., Schirripa, M., *et al.* 2021. A Cookbook for Using Model Diagnostics in Integrated Stock Assessments. Fisheries Research, 240: 105959. https://doi.org/10.1016/j.fishres.2021.105959

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

Hyndman, R.J. and Koehler, A.B. 2006. Another look at measures of forecast accuracy. International Journal of Forecasting, 22: 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

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

Hilborn, R. 2016. Correlation and causation in fisheries and watershed management. Fisheries, 41: 18–25. https://doi.org/10.1080/03632415.2016.1119600

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R., *et al.* 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES Journal of Marine Science, 72: 99–110. https://doi.org/10.1093/icesjms/fsu198

ICES. 2020. Minutes of the meeting of the ICES Advisory Committee (ACOM), Copenhagen, Denmark, 10–12 March 2020. 58 pp. http://doi.org/10.17895/ices.pub.7452

ICES. 2022a. ICES technical guidance for harvest control rules and stock assessments for stocks in categories 2 and 3. *In* Report of ICES Advisory Committee, 2022. ICES Advice 2022, Section 16.4.11. https://doi.org/10.17895/ices.advice.19801564

ICES. 2022b. Benchmark workshop on *Pandalus* stocks (WKPRAWN). ICES Scientific Reports. 4:20. 249 pp. http://doi.org/10.17895/ices.pub.19714204

ICES. 2023. Guide to ICES advisory framework and principles. *In* Report of the ICES Advisory Committee, 2023. ICES Advice 2023, section 1.1. https://doi.org/10.17895/ices.advice.22116890

Kell, L.T., Kimoto, A., and Kitakado, T. 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries Research, 183: 119–127. https://doi.org/10.1016/j.fishres.2016.05.017

Kell, K., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., and Fu, D. 2021. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science, 78: 2244–2255. https://doi.org/10.1093/icesjms/fsab104

Maunder, M.N., and Piner, K.R. 2015. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science, 72: 7-18. https://doi.org/10.1093/icesjms/fsu015

Merino, G., Urtizberea, A., Fu, D., Winker, H., Cardinale, M., Lauretta, M., Murua, H., *et al.* 2022. Investigating trends in process error as a diagnostic for integrated fisheries' stock assessments. Fisheries Research, 256: 106478. https://doi.org/10.1016/j.fishres.2022.106478

Minte-Vera, C.V., Maunder, M.N., Aires-da-Silva, A.M., Satoh, K., and Uosaki, K. 2017. Get the biology right, or use size-composition data at your own risk. Fisheries Research, 192, 114–125. https://doi.org/10.1016/j.fishres.2017.01.014

Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B. 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 102(6). https://doi:org/e03336. 10.1002/ecy.3336

Trijoulet, V., Albertsen, C.M., Kristensen, K., Legault, C.M., Miller, T.J., and Nielsen, A. 2023. Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. Fisheries Research, 257: 106487. https://doi.org/10.1016/j.fishres.2022.106487

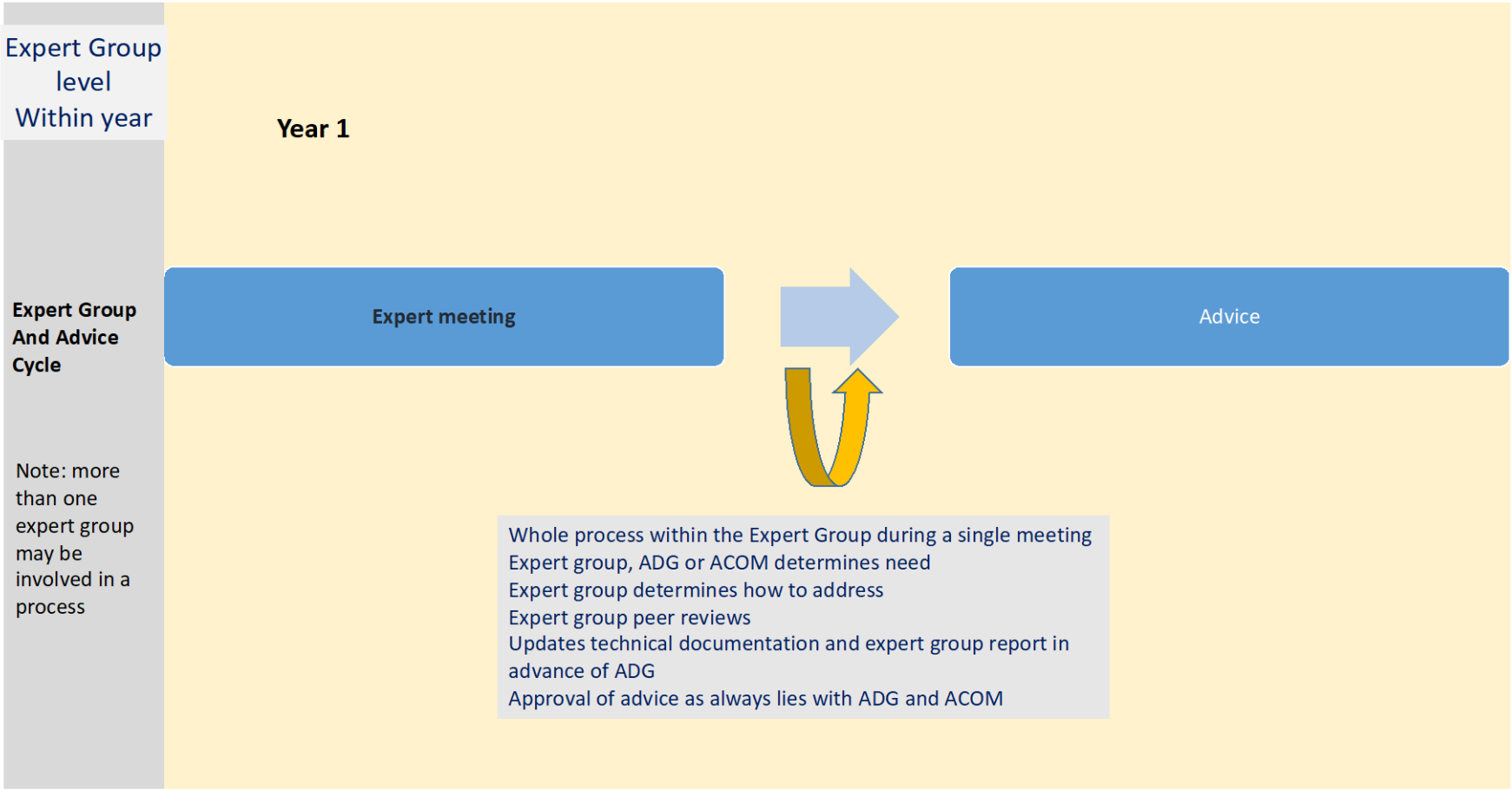
Siekmann, I., Sneyd, J., and Crampin, E. J. 2012. MCMC can detect nonidentifiable models. Biophysical Journal, 103: 2275–2286. https://doi.org/10.1016/j.bpj.2012.10.024

Subbey, S. 2018. Parameter estimation in stock assessment modelling: caveats with gradient-based algorithms. ICES Journal of Marine Science, 75(5): 1553–1559. https://doi.org/10.1093/icesjms/fsy044

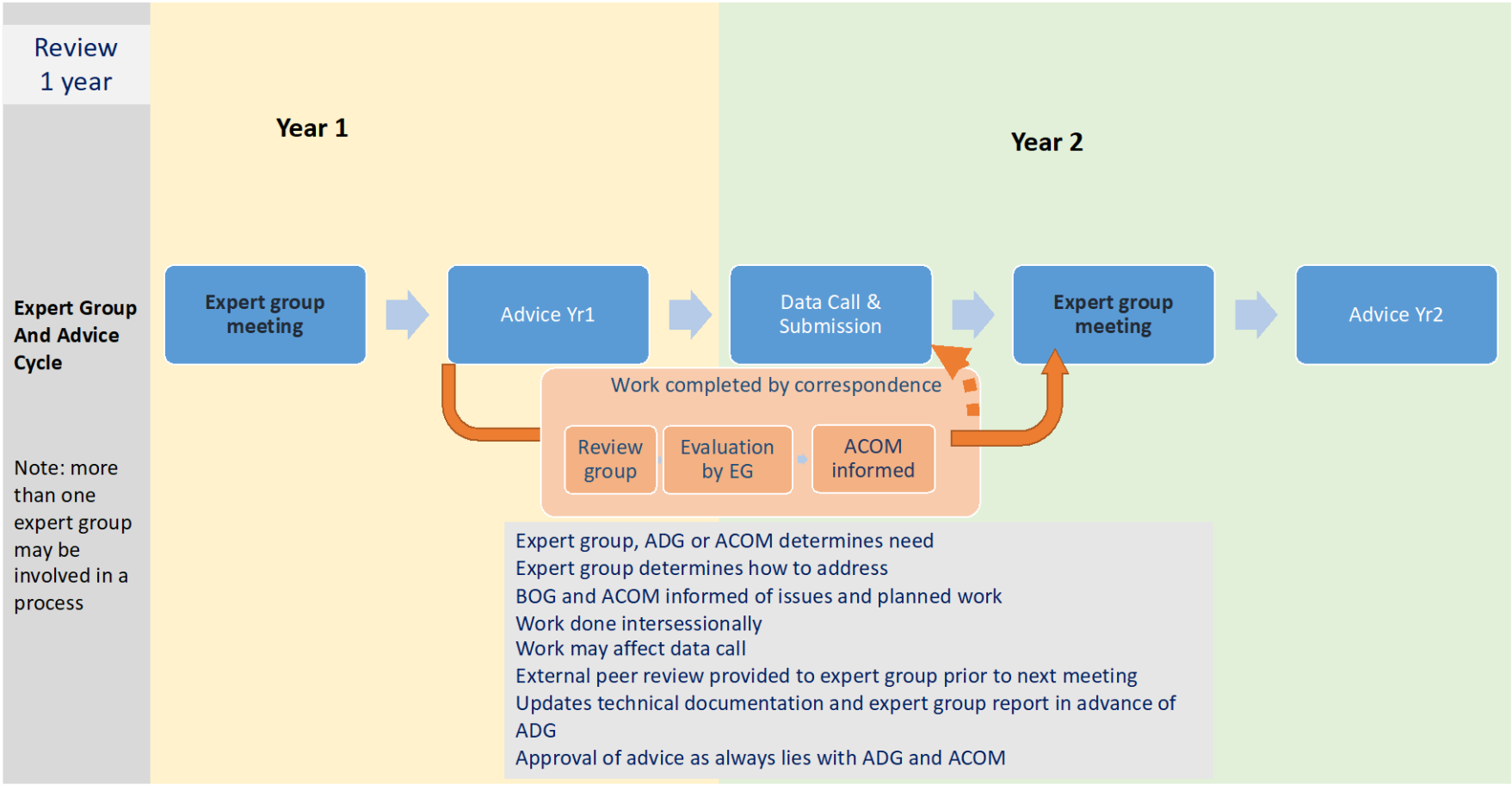
Walters, C. J., Hilborn, R., and Christensen, V. 2008. Surplus production dynamics in declining and recovering fish populations. Canadian Journal of Fisheries and Aquatic Sciences, 65(11): 2536–2551. https://doi.org/10.1139/F08-170

Annex 1: Flow diagrams for different benchmark processes

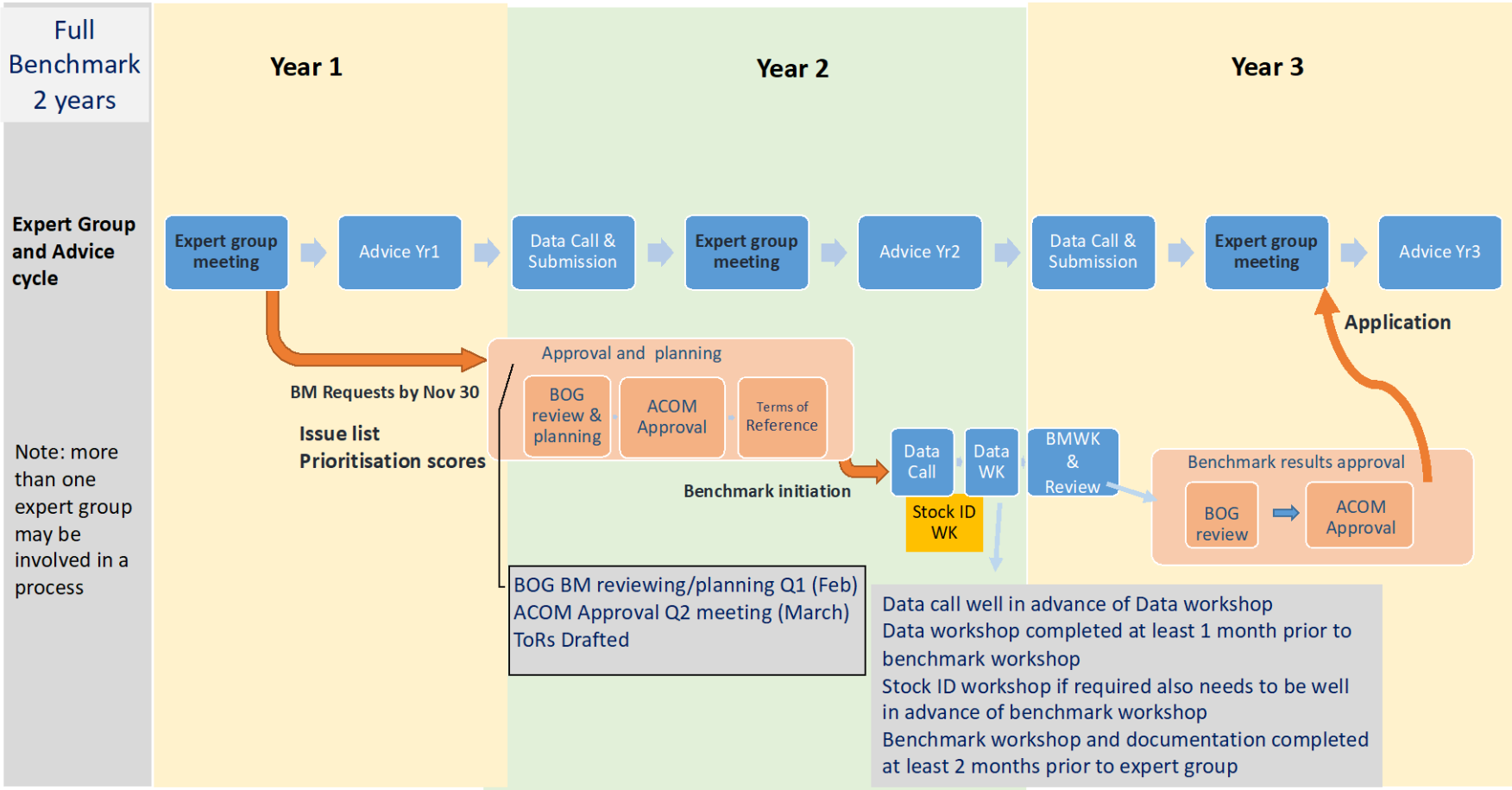
Expert Group Level



Review


Review
1 year
Year 1
Year 2
Expert Group And Advice Cycle
Note: more than one expert group may be involved in a process
Expert group meeting
Advice Yr1
Data Call & Submission
Expert group meeting
Advice Yr2
Work completed by correspondence
Review group
Evaluation by EG
ACOM informed
Expert group, ADG or ACOM determines need
Expert group determines how to address
BOG and ACOM informed of issues and planned work
Work done intersessionally
Work may affect data call
External peer review provided to expert group prior to next meeting
Updates technical documentation and expert group report in advance of ADG
Approval of advice as always lies with ADG and ACOM

Full Benchmark



Annex 2: Decision matrix for ICES benchmark prioritization

SCORE	Criteria 1 – Need to improve the quality of the analyses to provide advice or new recurrent advice	Criteria 2 – Opportunity to improve the analyses	Criteria 3 – Benchmark preparedness	Criteria 4 – Perceived risks	Criteria 5 – Changing ecosystem and ability to include impact in advice	Criteria 6 – Time since last benchmark
Weight	**0.3**	**0.3**	**0.3**	**0.2**	**0.2**	**0.1**
5	Analyses to provide advice does not currently exist or analyses judged to be insufficiently robust to form basis of advice	Significant new data sources or critical corrections in data, and significant new methods for the advice will be available	Data and analyses have been tested and timelines are proposed, potential reviewers and chairs identified/suggested	Status unknown or below precautionary thresholds and under imminent threat from human activities (e.g. fish stock below B_{lim}, increased bycatch of PET species)	Strong evidence of major directional change in ecosystem and good potential to include impact in advice (e.g. ability to calculate F_{ECO}, major change in VME boundary due to change in ocean temperature)	More than 10 years or first benchmark
4	Analyses have high potential to be upgraded from empirical to more analytically based	Significant new data sources or critical corrections in data, or significant new methods for the advice will be available	Data and analyses have been tested but remainder of planning not complete	Status less than optimal and under imminent threat (e.g. fish stock between B_{lim} and MSY $B_{trigger}$)	Evidence of some directional change in ecosystem with ability to include impact in advice (e.g. change in population productivity that can be incorporated in model)	More than 5 but less than 10 years
3	Analyses judged to have some significant deficiencies (models and/or data) but considered acceptable	Some improvement in data or methods will be available	Data have been identified but not incorporated into new analyses – new analyses tested on previous data	Status close to optimal and some imminent threat (e.g. fish stock close to MSY $B_{trigger}$)	Evidence of some directional change in the ecosystem with some ability to include impact in advice	3–5 years

SCORE	Criteria 1 – Need to improve the quality of the analyses to provide advice or new recurrent advice	Criteria 2 – Opportunity to improve the analyses	Criteria 3 – Benchmark preparedness	Criteria 4 – Perceived risks	Criteria 5 – Changing ecosystem and ability to include impact in advice	Criteria 6 – Time since last benchmark
Weight	**0.3**	**0.3**	**0.3**	**0.2**	**0.2**	**0.1**
2	Analyses have no significant or only minor issues	Minor improvement in data or methods will be available	Data and potential analyses have been identified but not tested	Status close to optimal and no imminent threat (e.g. fish stock above MSY $B_{trigger}$)	Evidence of small directional change in ecosystem with some ability to include impact in advice	2 years
0	Analyses has no obvious issues	Improvement in data or methods unlikely	No preparation	Status optimal and no threat (e.g. fish stock at or above BMSY)	No evidence of ecosystem change	Less than 2 years

Annex 3: ICES Benchmark toolbox for model diagnostics

General text

Stock assessment models are deeply scrutinised for model misspecification during development within benchmark workshops. Traditionally in ICES, diagnostics have been based on retrospective and residuals analyses. However, recent papers by Carvalho *et al.*, (2021) showed that when several diagnostic tests are considered together, the power to detect model misspecification improves without a substantial increase in the probability of incorrectly rejecting a correctly specified model (Carvalho *et al.*, 2017, 2021). Consequently, all applicable diagnostics should be used routinely during benchmarks. When the criterion for rejecting a model is a failure of at least one of the diagnostic tests, nearly 90% of misspecifications are detected with no real increase in the probability of a false detection (Carvalho *et al.*, 2017, 2021). For example, residual analyses were easily the best detector of misspecification in the observation model, while the retrospective analysis had low rates of detection of misspecified models (Carvalho *et al.*, 2017, 2021), although retrospective analysis is effective in detecting un-modelled temporal variation (Hurtado-Ferro *et al.*, 2015). Finally, and opposed to the widely used maximum-likelihood estimator, MCMC gives clear warning signs when a non-identifiable model is used for fitting (Siekmann *et al.*, 2012).

An example of the application of many of these model diagnostics can be found in ICES (2022b).

Convergence

The first step for checking model convergence is to verify if parameters are estimated at a bound, which can suggest problems with data or the assumed model structure. The second is checking that the final gradient of the model is relatively small (e.g., ≤ 1.00E-04 or smaller). The third is to determine whether the Hessian (i.e., the matrix of second derivatives of the log-likelihood concerning the parameters, from which the asymptotic standard error of the parameter estimates is derived) is positive definite (Carvalho *et al.*, 2021). Other convergence diagnostics include (i) examining the correlation matrix for highly correlated (e.g., > 0.95) parameter pairs; and (ii) examining parameters for excessively high variance as an indication that they do not influence the fit to the data (Carvalho *et al.*, 2021).

Residual analysis

A non-random pattern of residuals in integrated assessment models may indicate that some heteroscedasticity is present, or there is some leftover serial correlation as for example serial correlation in sampling/observation error and/or that the model is mis-specified. Several well-known nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test (Gibbons and Chakraborti, 1992). Standardized residuals are commonly used, although recent analysis showed that one-step-ahead (OSA) should be used instead in stock assessment model diagnostic (Trijoulet *et al.*, 2023).

Runs test have been recently proposed to be used to evaluate whether residuals were normally distributed and/or displayed time trends. The runs test was chosen as this test has recently been

used to diagnose fits to indices and other data components in ICES assessment models (e.g., ICES, 2022b).

The RMSE runs test (see Carvalho *et al.*, 2021 for details) indicates that the fit is satisfactory if no residuals are larger than 1 and the RMSE is below 30%, indicating the presence of a random pattern in the length frequency distributions and in the survey indices. The RMSE plot is frequently used as a tool for identifying trends in residuals, and if the standard deviation is small on a given year this means the fleets included in the model agree, even if not fitting well, which is a useful diagnostic. Its purpose is to visualize multiple residuals at once, pick up on periods of substantial data conflicts (width of boxes) and systematic departures in median residuals (LOESS smoothers).

Merino *et al.*, (2022) has described and applied a novel model diagnostic to identify trends in process error in recruitment deviation estimates within integrated assessment models. Significant trends in recruitment deviates can be caused by misspecification of the biological parameters used as fixed values in integrated assessment models. The process error diagnostic described there can provide a statistical criterion in support for hypotheses and assumptions when using best case or ensembles of models to develop fisheries management advice.

Jittering

The jittering procedure allows users to verify the stability of a model and its parameter estimates by examining the effect that small changes in its starting values have on model results. An accurate model should converge on a global solution (i.e., not be stuck in local minima of the likelihood surface) across a reasonable range of starting values for all input parameters. For example, a 10% jitter of all initial parameters means that a small random jitter is added to the initial parameter values and the model is rerun. It is, however, important to stress that the absence of a local minima when running jittering is not a guarantee that the model is not indeed stuck in a local minimum, although its absence does reduce the risk that this occurs (Subbey, 2018).

Retrospective analyses

Retrospective analysis is a diagnostic approach to evaluate the reliability of parameter and reference point estimates and to reveal systematic bias in the model estimation. It involves fitting a stock assessment model to the full dataset. The same model is then fit to truncated datasets where the data for the most recent years are sequentially removed and the Mohn´s rho statistics is calculated. Given that the variability of Mohn's rho index depends on life history, and that the statistic appears insensitive to F, Hurtado-Ferro *et al.* (2015) proposed the following rule of thumb when determining whether a retrospective pattern should be addressed. Values of Mohn's rho index higher than 0.20 or lower than −0.15 for long-lived species (upper and lower bounds of the 90% simulation intervals for the flatfish base case), or higher than 0.30 or lower than −0.22 for short-lived species (upper and lower bounds of the 90% simulation intervals for the sardine base case) should be cause for concern and taken as indicators of retrospective patterns. However, Mohn's rho index values smaller than those proposed should not be taken as confirmation that a given assessment does not present a retrospective pattern, and the choice of 90% means that a "false positive" will arise 10% of the time. In both cases, model misspecification would be correctly detected more than half the time.

Hindcasting

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

Kell *et al.* (2016, 2021) and Carvalho *et al.* (2021) showed that hindcasting can be used to evaluate model prediction skill of CPUE time series. When conducting hindcasting, a model is fit to the first part of a time series and then projected over the period omitted in the original fit. Prediction skill can then be evaluated by comparing the predictions from the projection with the observations using, for example, the MASE indicator (Hyndman and Athanasopoulos, 2013). If a model is used for prediction, the specific tool used for model selection is less important than the approach used to validate predictions. Quantifying predictive skill using independent data in ecology is therefore essential (Tredennick *et al.*, 2021).

MASE can be calculated for single components as length, age distributions and surveys index, but also combining all available components in a single joint MASE statistics as for example when several surveys indices are used in the assessment model.

MCMC

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

For Northern shrimp in divisions 3.a and 4.a East, an MCMC run was performed as a diagnostic (thus not for inference, as that would necessitate a much larger number of iterations) using the NUTS algorithm with 3 chains of 50 000 iterations each. We discounted the first 50% of the iterations as burn-in period and used no thinning. The results showed that the MCMC is almost identical to the MLE estimated, which is an indication of the robustness of the model.

Likelihood profile

Likelihood profiling allows to evaluate model performance across a range of values of an input parameter (generally R_0, σR, and steepness but any other parameter can be profiled) and existence of conflict between sources of information (Carvalho *et al.*, 2021). This diagnostic reports the likelihood over each data component across a particular parameter profile. A profile is conducted by sequentially fixing a given parameter to a range of values and then examining the change in the total and data-component likelihoods. A relatively large change in negative log-likelihood units along the profile suggests a relatively informative data source for that model. Also, a difference in the location of the minimum negative log-likelihood along the profile between data sources might suggest either conflict in the data or model misspecification (or both).

Analysis of surplus production trend

Estimates of Surplus Production (Walters *et al.*, 2008) can provide a check of whether predictions of changes in biomass can be made reliably based on catch and current biomass (clockwise or linear behaviour) or whether there has been non-stationarity in production processes, i.e., are dynamics driven by climate and oceanic conditions (counter clockwise). This is important, for example, for the development of management procedures (MPs) in the MSE process.

ASPM

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

The results from the ASPM should be similar to those from the fully integrated model if the size- and age-composition data are not informing absolute abundance or the trend in abundance and there is no strong pattern in recruitment. The ASPM test (Maunder and Piner, 2015) appears to have promise in detecting systems dynamic misspecification (h and M), where the runs test showed lower power, and ASPM showed good power.

The ASPMdev is a variation of the ASPM diagnostic and designed to evaluate if composition data is needed to estimate the variability in recruitment (Minte-Vera *et al.*, 2017). It involves fitting to indices of abundance while simultaneously estimating recruitment deviates in the absence of the composition data. Suppose the ASPMdev produces results substantially different from the fully integrated model and the ASPM. This would indicate that the composition data provide the primary source of information for estimating recruitment deviations.

Annex 4: Detailed version history

Version	Date	Major changes
1	March 2023	Guidelines established