SPRFMO
South Pacific Regional Fisheries Management Organisation

14th MEETING OF THE SCIENTIFIC COMMITTEE

7 to 12 September 2026, Tórshavn, Faroe Islands

SC14 - JM 04
Independent desk review of the SPRFMO Jack Mackerel Management Strategy Evaluation

JMWG - Quantifish Limited

QUANTIFISH
Quantitative Fisheries Science

Independent desk review of the SPRFMO Jack Mackerel Management Strategy Evaluation

AUTHOR
Darcy Webber

AFFILIATION
Quantifish Limited

PUBLISHED
August 8, 2026

1 What is the SPRFMO Jack Mackerel MSE?

A **management strategy evaluation (MSE)** uses simulation to test proposed catch-advice rules *before* they are used in real decisions. It asks a practical question: if SPRFMO used the same rule for many years, how would the stock and fishery be likely to perform across a range of plausible futures?

The main parts are:

- **Management objectives** — what managers want the procedure to achieve, such as stock safety, useful catches, and reasonably stable advice.
- **Operating models (OMs)** — simulated versions of the stock and fishery. A reference OM is used for the main comparison. Other OMs test harder but plausible situations, such as recruitment shocks, two stocks, or movement between areas.
- **Candidate management procedures (CMPs)** — the complete rules being tested. A CMP states which data to use, how to combine them, how the **harvest control rule (HCR)** converts the stock indicator into advice, and how much the **total allowable catch (TAC)** may change from year to year. The main families here are hockey-stick (HS) and power-ramp (PR) rules.
- **Closed-loop simulation** — in each simulated year, the OM produces imperfect data, the CMP turns those data into catch advice, and the catch then affects the simulated stock. This is repeated across many possible futures.
- **Performance measures** — results used to compare CMPs, including stock safety, catch, year-to-year stability, and whether the rule is practical to apply.
- **SPRFMO decision pathway** — the MSE Task Team and workshop evaluate CMPs and propose a shortlist. The Scientific Committee reviews the evidence and advises the Commission, which decides whether to adopt a procedure. If one is adopted, it is applied each year. An **exceptional-circumstances protocol (ECP)** checks whether conditions have moved outside the range tested in the MSE (SPRFMO 2026b).

The annual stock assessment, benchmark work, and survey and CPUE analyses still matter. They help build the OMs and supply the indices used by CMPs. Until an MSE-tested procedure is adopted, adjusted Annex K remains the source of catch advice.

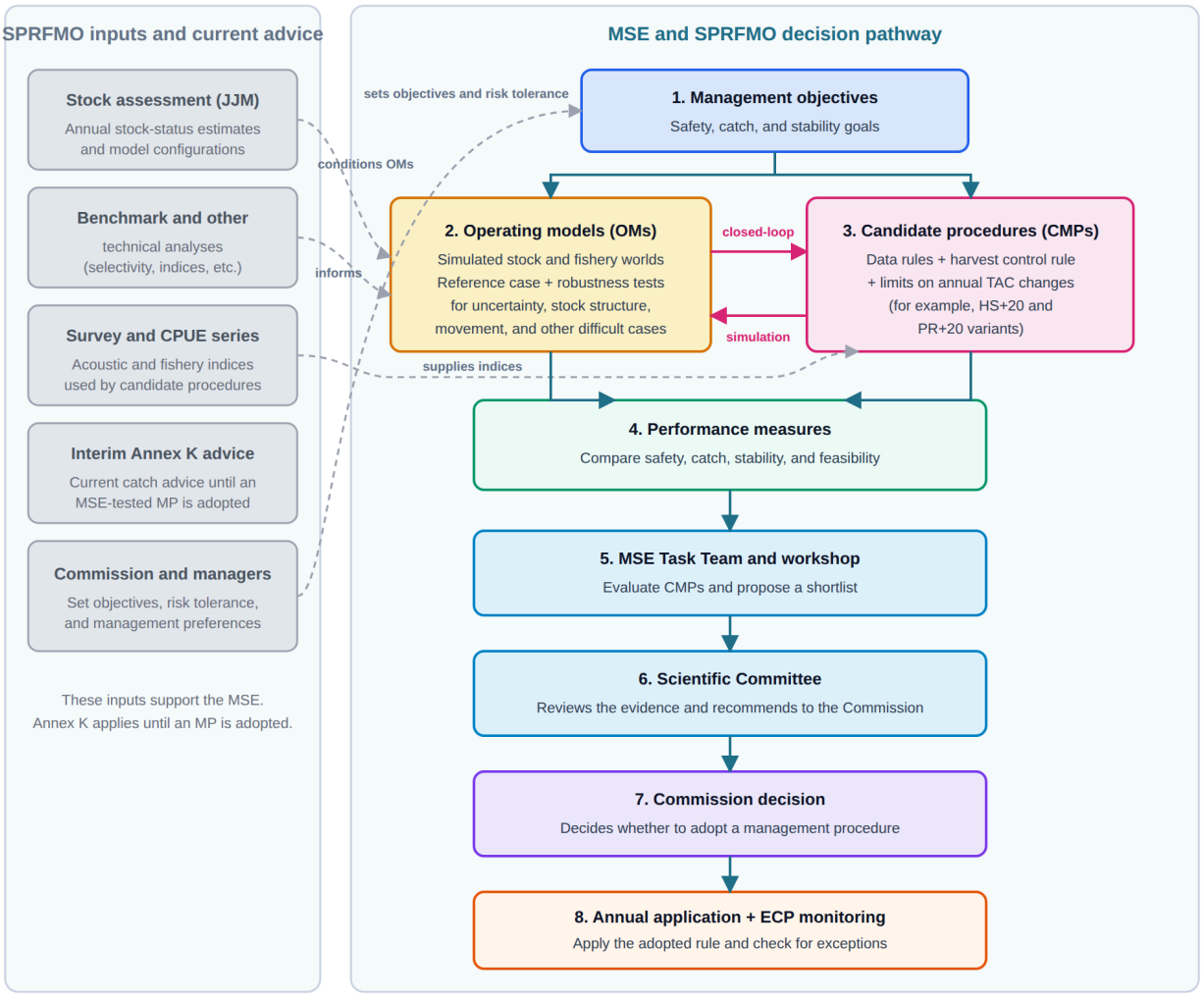

Figure 1: How the SPRFMO Jack Mackerel MSE connects to the formal SPRFMO decision pathway.

2 What I reviewed

This review follows the Terms of Reference. Its main subject is the draft *Jack mackerel management strategy evaluation for SC14* dated 29 July 2026 (SPRFMO 2026a). The materials considered are summarized below.

Main SPRFMO documents

- Draft *Jack mackerel management strategy evaluation for SC14* (29 July 2026), in source and rendered form (SPRFMO 2026a).
- *SCW17 MSE Workshop Report* and Papers 01, 02, 03, 06, 07, 08, and 09 (SPRFMO 2026d).
- *JM MSE Task Team meeting report* (24 July 2026) (SPRFMO 2026c).
- 2026 JMWG Benchmark meeting report.
- SC13 adopted report (2025), including the interim Annex K advice (SPRFMO 2025).
- COMM8 Annex 8b, *Jack Mackerel MSE Management Objectives* (SPRFMO 2020).

- COMM14 Annex 4b, *Jack Mackerel MSE Implementation Pathway* (SPRFMO 2026b).
- Relevant SPRFMO MSE workshop reports from 2024 and 2025.

Files and results checked 30 July 2026

- `jmMSE26` commit `e3cafc5`: report source, controls, registry, CMP and application pages, traceability, scorecard and figure scripts, validation script, and source and deployed candidate CSVs.
- Verified `output/jm_candidates.slick`, inspected with Slick 1.0.2.
- The public `mse` package definition of the Kobe-green statistic.

Guidance and research

- MSE design, robustness, implementation, and communication (Punt et al. 2016; Rademeyer et al. 2007; ICES 2019).
- Testing exceptional-circumstances protocols (Carruthers and Hordyk 2019).
- Sensitivity to dynamic reference points (Berger 2019).
- Multi-index harvest strategies and stakeholder design (Harford et al. 2021).
- Multi-objective performance measures and preference weighting (Dowling et al. 2020).

The **Slick file** is a saved set of MSE results used for interactive comparisons. I downloaded it and confirmed its SHA-256 file fingerprint — a check that it is exactly the same file — as `0e18d665c2a70f953d6727b805cbc47ad22063021ef306bc5a250d977a748a9c`. It matches the fingerprint on the traceability page. I inspected the file and recalculated selected results. I did **not** rerun the underlying simulations or audit all of the code.

COMM14 asks for a workable interim package for SC14, with the final report and source code due in November 2026. Even as an interim product, the package would benefit from presenting one consistent account if it is used to support the SC14 report. SC13 kept adjusted Annex K in place because the MSE was not yet complete.

The Terms of Reference ask for a one- to two-page review based on about four hours of work, without rerunning the analysis or auditing all of the code (Appendix A). Sections 3 and 4 give that short review. Section 5 gives the supporting detail. I included some extra numerical checks because they materially change the view of readiness.

3 Overall view

The draft is a strong scientific summary of a credible MSE framework. It models errors across several indices together and openly discusses the risk of there being two stocks. These are real strengths. I support using the report to continue SC14 discussion and identify a provisional shortlist.

However, the draft SC14 report may benefit from some further work before submission. The main concern is consistency. The report, CMP specifications, source data, downloadable result files, public CMP catalogue, and saved Slick results do not appear to describe the same rules or results. This may make it difficult for a reader to tell which CMP settings, data-timing rules, registry entries, and

performance results are intended to be official. The criteria for deciding whether a candidate procedure is sufficiently safe also do not yet appear to have been agreed; questions about catch allocation, geographic scope, movement, and indices remain open; the HS-versus-PR comparison combines rule shape with other differences; the agreed performance set is still being developed; banking-and-borrowing work is pending; and several rules for applying a CMP each year would benefit from clarification.

These issues appear readily addressable.

4 Answers to the Terms of Reference

The original Terms of Reference are reproduced in Appendix A. My answers to the six review questions follow.

1. Does the report reflect the workshop design, later analyses, and current results?

Mostly for the design, although the reported results are less consistent. The description of the OMs and CMPs broadly follows SCW17. However, the public files appear to come from different analysis versions. The CMP registry — the public catalogue of candidate rules and their status — describes an older set of CMPs in the Slick file; one figure script does not assign labels to four CMPs and combines them into one group; and HS+20 and HSsym give identical results under most saved OMs (Issue 1).

2. Are the OMs, CMPs, tuning, robustness tests, and performance measures appropriate?

Partly. The overall MSE structure is appropriate, but some key results were not readily reproducible from the available files, and several specifications and decision criteria appear to remain under development. These include the tuning and two-stock results; the catch-allocation rule; a comparison that separates HCR type from other design differences; evidence that the number of simulations is sufficient; the complete set of performance measures; and criteria for judging CMP performance across the harder robustness tests (Issues 1, 2, 3, 4, 5, and 6).

3. Are uncertainties, limitations, and trade-offs clear?

They are explained well in words, but are not yet fully supported by readily checkable numbers. The report is open about the two-stock risk, but I was unable to reproduce the statements in the text from the saved results, and the very low northern performance under om23 does not appear to be discussed. A single checkable table would also help bring together the main safety results, effort-limit checks, reference-point tests, assumptions about index errors and actual catch, and uncertainty due to the number of simulations (Issues 3, 5, and 7).

4. Are the conclusions and requests to the Scientific Committee supported?

A provisional shortlist may be supportable once the files are aligned. The evidence for selection or adoption would benefit from further development. Some rules for the annual calculation would benefit from clarification. The report also says banking and borrowing appear feasible even though those runs were still pending, and its wording about having met Commission requests may be stronger than the available evidence supports (Issues 2 and 3).

5. Are there important inconsistencies, omissions, or misleading statements?

Several potentially important inconsistencies and omissions are present. Files differ in their analysis versions, HCR trigger values, data timing, which CMPs are in the registry and Slick file, numerical results, and how some figure summaries were generated. Two candidates that are defined differently are identical under most saved OMs, and the two headline CMPs lack published achieved tuning probabilities. These differences may affect the ranking, interpretation, or annual catch advice (Issues 1, 2, and 4).

6. Does the report separate scientific findings from management choices?

Not always clearly. The 60% target, acceptable safety levels, scorecard weights, judgments about which OMs are plausible, geographic scope, and whether to accept a known weakness may involve management judgment as well as scientific input. The report would be clearer if it distinguished the scientific evidence from choices that may ultimately need to be made by managers (Issues 3 and 6).

5 Main issues to address

The issues below are ordered by priority. **[High]** issues appear important to address before submission. **[Medium]** issues could be addressed or more clearly acknowledged.

5.1 [High] Align the published files

The published package does not yet give one clear answer about which CMPs were tested, which parameter values define them, or what results they produced.

1. **The HCR trigger values disagree.** A trigger is the relative stock-index value that controls where an HCR changes slope or reaches its target catch. It is part of the rule: if the trigger changes, the same observed index can produce different catch advice.

   The SC14 report, `tuns_ctrl.csv`, and the source `cmp-registry.csv` give triggers of **2.3125** for MP29, **1.703125** for MP32, **2.3125** for MP43, and **1.65625** for MP44. The individual CMP pages instead give **2.12** for MP29 and **1.42** for MP32. The annual-application page gives **2.125**, **1.421875**, **3.0625**, and **1.9140625** for MP29, MP32, MP43, and MP44, respectively. The deployed registry table uses the same older values as those public pages, and its accompanying text repeats **3.0625** and **1.9140625** for MP43 and MP44. Although 2.12/2.125 and 1.42/1.421875 appear to be rounded versions of one another, neither pair matches the completed 500-posterior controls. These sources therefore specify materially different catch rules.

2. **The source files and public downloads give different results.** The repository contains one set of candidate-results CSVs for the report and another for the public website. Of the 12 files present in both sets, 11 contain different results. All **48 inputs** used to build the scorecard differ between the source and public versions. The report's source data rank PRsym first (equal-weight score 71.36; balanced score 67.00), while the live CSV linked from the report ranks HS-20 first (81.65; 73.75).

The deployed *reference* `catch_vb2025_*` and `catch_vbmsy_*` summaries use $n = 100$, while the matching source summaries use $n = 500$. The deployed *robustness* catch summaries use $n = 400$ or 500, while the source robustness summaries use $n = 2000$ or 2500. The Kobe-period summaries already say $n = 500$ in both places. Taken together, the public downloads still reflect the exploratory runs based on 100 simulations per OM that were recorded on 24 July, while the report source uses the final runs based on 500. A reader following the report's data link therefore gets a different ranking and sample size from those used in the report.

3. **The reported tuning probabilities were not reproducible from the published materials using the stated definition.** I applied the definition printed in the report — $SB/SB_{\mathrm{MSY}} \geq 1$ and $F/F_{\mathrm{MSY}} \leq 1$ — directly to the 500 saved trajectories for 2041–2050 under the reference OM. This gives 0.570 for each HS variant and 0.522–0.545 for the PR variants. The report instead gives 0.609 for HS-20/HSsym/HS-30 and 0.6054, 0.6000, and 0.5912 for PR-20/PRsym/PR-30. It **does not give the achieved probability as a number for either headline candidate**: HS+20 is said only to have "met the 0.60 objective within the specified 0.01 tolerance", and PR+20 has no achieved value at all.

   This discrepancy does **not** necessarily indicate an error in the report. It may use a different way of calculating reference points that change over time, or a different saved set of results. It does suggest that an independent reader may be unable to reproduce the values from the published definition and files. The same calculation also gives different results from the report's statements about two stocks: under the printed definition, both components of om21 (no movement) are well above 0.60 for all eight CMPs, and under om21_1 (with movement) the PR variants exceed 0.60 on both components while the HS variants fall short on the southern component. The public package also lacks the reported set of tests using different catch splits. The tuning and two-stock results would therefore be better described as **not yet reconciled with the published files**, rather than independently verified (Issue 3).

4. **HS+20 and HSsym give the same results under most OMs.** The registry defines them differently — both share trigger 2.3125 and target 2,000 kt, but HS+20 has an annual increase limit of 1.20 and HSsym of 1.15. In the verified Slick file, their trajectories match to machine precision — apart from negligible computer rounding — for the reference OM and all five single-stock OMs, across all six saved measures. They differ only in the two-stock OMs. The source scorecard ties them at equal rank and score; the deployed scorecard places them five ranks apart (second and seventh). These results suggest either that the increase-limit difference has no effect in those runs or that one candidate duplicates the other there.

5. **One vulnerable-biomass figure script appears to combine four CMPs.** The script `build_candidate_evidence_figures.R` recognizes MP29, MP43, MP32, and MP44 for the vulnerable-biomass summaries, but not MP45–MP48. Those four CMPs therefore receive a missing label and are pooled into one group in `vb_reference_summary.csv` and `vb_robustness_summary.csv`. The same script does label MP45–MP48 correctly in the catch–

VB summaries. The deployed VB summaries use a different, older ten-CMP set that includes MP18/MP23/MP24/MP31/MP35/MP36. The plots and summaries therefore do not show one consistent eight-CMP comparison.

6. **The public CMP registry appears to list an outdated candidate set for the Slick file.** Here, **registry** means the public catalogue of CMP names, settings, status, and which saved results include each CMP. It says the current Slick file contains “ten of the fourteen” registered CMPs, that MP45–MP48 still need a rebuild, and that an older set including MP18/MP23/MP24/MP31/MP35/MP36 is present. The SC14 report, traceability page, and verified Slick file instead contain exactly eight CMPs: MP29, MP32, and MP43–MP48, including MP45–MP48. The “10 of 14” statement appears to be an older inventory and would benefit from correction.

A fingerprint confirms the identity of one saved file. It does not show that all parts of the analysis agree or allow someone else to rerun it. The traceability page cites an untagged `jmMSE` branch based on commit `278c309…`, while the annual-application page cites the same commit on branch `devel`. No public link to that branch was available on the review date, and the complete run results were not archived.

Suggested actions:

- Choose the official candidate and control files, tag the exact code version, and archive the inputs and complete run results.
- Rebuild the report, registry, CMP and application pages, public `docs/` files, and Slick file from that release, then publish their fingerprints.
- Publish the achieved tuning probabilities for HS+20 and PR+20 as numbers.
- Explain why HS+20 and HSsym duplicate each other under the reference and single-stock OMs, or remove one from those comparisons.
- Extend `validate_sc14_parity.R` so it checks the actual numbers, not just the size of each result table or array. It should cover triggers, performance results, period summaries, scorecards, public files, which CMPs are listed, tuning, and the two-stock results.
- Correct the CMP labels and rebuild all vulnerable-biomass summaries and figures.

5.2 [High] Limit the current request to shortlist consideration

The analysis may support a shortlist, but the published material would benefit from further specification before two people could independently repeat the annual calculation and obtain the same answer. Items requiring clarification include:

- catch units and rounding;
- the official data providers and annual cut-off dates;
- what to do with missing or revised indices;
- which previous advice value to use when applying annual change limits;
- how total catch is allocated among fisheries and which data update that split;
- fallback calculations and checks that the calculation can be repeated; and

- ECP triggers, review steps, and possible responses.

These are substantive parts of the procedure. For example, the current code averages whichever index series are available. If one is missing, its weight disappears and the others automatically receive more weight. That would ideally happen only if it is the agreed rule. An operational MP would generally be detailed enough that no new, untested judgement is needed when it is applied (Rademeyer et al. 2007; Punt et al. 2016).

Catch allocation would benefit from further clarification. The registry and CMP pages use `split.is`, which divides the total TAC among fisheries using their average catch shares over the previous five years. The report's two-stock conclusions instead use exploratory **fixed** catch splits. These are different procedures. It may be premature to use the spatial-risk results for a decision until the report states which split would actually be used, matches it to the two-stock design, and tests it again.

The 24 July record also says that Peru intends to continue separate northern monitoring, assessment, and management. If so, SPRFMO's main question is whether a one-stock CMP protects the southern component. The SC14 draft does not explain the results in those terms. At the same time, the tested CMPs use `split.is` to allocate a total TAC across all fleets. It may be preferable not to describe a shortfall for the northern component as a failure of a jointly managed northern TAC unless that joint management system is actually included in the OM and CMP.

The data-timing descriptions also appear inconsistent. The report says most indices arrive with a one-year lag and offshore CPUE with a two-year lag. The annual-application page says all three-year CMPs use a one-year delay and a 2023–2025 window. These rules produce different indicators. The 2019–2023 reference period also differs from the earlier 2015–2020 trials. The draft does not explain or test that change, even though the reference-period mean contains no 2022 acoustic value.

It is not clear from the package whether the simulations include the gap between the CMP's advice, the TAC adopted by managers, and the catch actually taken. If this gap was left out because it was considered negligible, it would be helpful for the report to explain why. Otherwise, it could be considered for inclusion in the simulations (Punt et al. 2016).

The report mentions fishing-effort feasibility and the 270-kt minimum catch but does not show the results. SCW17 treated about three times current effort as a plausible upper bound and asked for checks of low catches, catch rates, and how often the effort cap is reached. The SC14 report gives neither the cap nor those results. All eight CMPs also impose the 270-kt minimum, but the report does not show how often it applies, $P(C < 270)$, or what happens if the CMPs are retuned without it.

Finally, the report says banking-and-borrowing and implementation "appear feasible". Table 6 and the 24 July record say the 10% banking-and-borrowing runs were still pending. That conclusion appears to be ahead of the evidence currently available.

Suggested actions:

- Consider asking for **endorsement of a shortlist for further consideration** rather than “adoption of a subset”, unless all specifications and analyses are completed before submission.
- Either omit the banking-and-borrowing feasibility claim or state clearly that the work is pending.
- Document and retest `split.is` against the fixed catch-split tests, and explain how Peru’s northern management fits within the CMP’s scope.
- Clarify and, if needed, retest the lag, missing-data, and reference-period rules.
- Report the effort cap, how often the 270-kt minimum and effort cap apply, and the low-catch and catch-rate results.
- Simulate the gap between advice and actual catch, or explain why it was left out.

5.3 [High] Clarify safety criteria and review difficult test cases

The mean 60% Kobe-green probability is a useful tuning target, but it may not by itself constitute a complete safety standard. It averages across simulations and years. A CMP can meet that average while still having a potentially concerning chance of crossing a biomass limit in a particular year.

SC13 records discussion of the 60% Kobe-green target in science-manager sessions. COMM8 Annex 8b, however, records Commission approval of a **draft** objective of at least 50% probability above B_{MSY} in 2030 and 95% probability above B_{lim} throughout 2025–2040 (SPRFMO 2020). The SC14 report does not show these tests or identify a later decision that replaced them. It may therefore be premature to say the results “satisfy the Commission requests and recorded management preferences”.

The OM set is broad and scientifically useful, but the report does not say which harder cases should be considered decisive, how well a CMP should perform, or how plausible each case is. The reported two-stock results matter if they are confirmed: without movement, one component misses the 60% benchmark; with movement, no tested fixed catch split meets it for both components. I was unable to reproduce these statements from the results in the verified Slick file using the printed Kobe-green definition (Issue 1). A difficult result need not automatically exclude a CMP, but choosing to accept the associated risk may involve management judgment and would benefit from being stated clearly.

The report also presents only some of the robustness results. It discusses om21 and om21_1 in detail, but under om23 (two stocks, model 1.14 configuration) the northern component has a 2041–2050 Kobe-green probability of only about 0.002–0.021 under every one of the eight candidates. This very low result does not appear to be discussed. A complete safety framework would ideally bring such an outcome forward explicitly.

The movement result may not be fully addressed through ECP monitoring alone. In the report’s movement test, the combined index can rise while the southern component falls. An ECP can flag an unexpected or unlikely situation, but monitoring does not itself resolve a vulnerability already identified by the MSE. ECP triggers would also benefit from testing under other OMs to measure false alarms and missed detections before use (Carruthers and Hordyk 2019).

Before selecting a CMP, the analysis could add a regional safeguard or regional indicators; define and test its geographic scope and catch-allocation rule, including any reliance on separate northern management; or explain why the movement and model-1.14 cases are considered too implausible to affect the choice.

SCW17 also asked for a comparison based on $F_{35\%}$ because a Kobe classification can change with fishing selectivity and with the way reference points are updated. The SC14 report does not show that comparison. Nor does it say whether F_{MSY} changes each year or is held fixed within each simulation, or how B_0 is recalculated each year for SB/SB_{MSY}. The code that builds the published Slick results divides by the OM `refpts` ratios. Until the equations, code, and description are better aligned, the tuning remains difficult to check fully. Reference points that change over time may be appropriate when stock productivity varies, but they can also change the apparent stock status during a trend or regime shift. It would be helpful for the report to show whether another reasonable calculation changes the conclusions (Berger 2019).

Suggested actions:

- Consider safety limits for each stock component, including the chance of crossing a limit in any year and the highest risk in a single year.
- Explain how the COMM8 objective relates to the later 60% target, and remove or qualify the statement that Commission requests have been met.
- Publish a CMP-by-OM table showing at least Kobe-green probability, low-biomass or collapse risk, overfishing risk, catch, interannual catch change (IACC), and whether the rule can be applied in practice, including om23.
- Reconcile the two-stock statements with the saved Slick results, or publish the results file that supports those statements.
- Describe judgments about the plausibility of each OM and the standards a CMP would need to meet to remain under consideration.
- Clarify the movement, index, allocation, and geographic-scope questions.
- Publish the reference-point equations and the requested $F_{35\%}$ or another suitable comparison.
- Complete and test the ECP before it is used.

5.4 [High] Describe the HS-versus-PR comparison as more than shape alone

The report calls HS+20 versus PR+20 its main “rule-shape comparison”. The two CMPs differ in more than shape. They also use different target catches (2,000 versus 1,500 kt), different tuned triggers, and different ways of turning an index value into catch advice. Paragraph 47 of the SCW17 report (official PDF, p. 12) notes that the effect of catch level should be separated from the effect of HCR slope.

The higher catch under HS or the faster response under PR therefore cannot be attributed with confidence to rule shape alone. The result is more accurately described as a comparison of two complete CMP configurations.

Suggested action: call this a “comparison of two complete CMP configurations” and avoid strong claims about the effect of shape alone. If a shape-only conclusion is needed, compare rules with the same catch target and constraints and retune each one to the same objective.

5.5 [Medium] Assess whether 500 simulations provide sufficient precision

The report gives a tuning tolerance of 0.01, but that describes only how close the numerical search must get to its target. It does not show the random variation caused by running a limited number of simulations (Monte Carlo uncertainty). Nor does it explain whether the 500 “posterior iterations” (model draws) also contain 500 independent draws of future stock and index error. These are different sources of uncertainty. ICES guidance recommends increasing the number of runs until the results that matter for the decision stop changing (ICES 2019).

As a rough guide, a probability near 0.60 based on 500 independent yes/no simulation outcomes has a standard error of about 0.022. Averaging over ten years may improve precision, but the improvement depends on how strongly the years within each trajectory are related. Values such as 0.591, 0.600, and 0.609 may therefore not be meaningfully different unless the results are shown to be stable.

Suggested action: explain how the 500 runs were constructed, including how stock, index, and actual-catch uncertainty was added and whether candidates used the same random draws. Report uncertainty intervals caused by the finite number of runs, and show whether the tuning and key safety results change with more runs or new random seeds. Cite the decision that set the 60% target — SC13 and the COMM14 pathway refer to K60 — rather than presenting it as a scientific constant.

5.6 [Medium] Treat scorecard rankings as preliminary

The report appropriately notes that the scorecard is illustrative. Even after the version differences are resolved, it would be premature to use it as a definitive ranking of CMPs because it currently:

- uses only long-term median results from the reference OM;
- leaves out the spread of safety results and the harder test cases;
- counts some closely related stock measures more than once;
- can change when candidates are added or removed because of its scaling; and
- uses weights that decision-makers have not agreed.

SCW17 agreed that the full set of safety, yield, stability, low-catch, and feasibility statistics would be calculated and reported, and it highlighted six headline measures for Scientific Committee review, including Kobe-green probability and the risk of a shutdown or low catch. The shared SC14 Slick data instead contain a different set of six annual measures — SB/SB_{MSY}, F/F_{MSY}, Catch, IACC, VB/VB_{2025}, and VB/VB_{MSY} — so P(Green), low-catch, and shutdown probability are absent from the saved results. The report itself says the plots omit measures needed for a decision, and Table 6 says the definitions and time periods are still being finalized. The rankings are therefore best treated as preliminary and may be better kept out of the executive summary at this stage.

Safety measures may be clearest when they are used first as pass/fail checks. Candidates meeting agreed safeguards could then move on to a weighted comparison of trade-offs. Any overall weighted ranking would also benefit from clear objectives and stakeholder weights that have been agreed and documented (Dowling et al. 2020).

Suggested action: consider removing rankings from the executive summary until the files agree and the full set of results is available. Apply the safety checks first. Then show the spread of results and trade-offs, and test whether the ranking changes with the measures, periods, weights, OMs, or random variation from the simulations.

5.7 [Medium] Document and further test the model for index errors

Simulating the five indices together is a real strength because it allows their errors to be correlated. The report says the closed-loop test exposes the CMP to "realistic combinations of agreement, redundancy, and divergence among indicators". That claim may be stronger than the current documentation supports. The report would benefit from explaining:

- how the correlations were estimated;
- which years overlap among the indices;
- which transformations were used;
- whether the correlation matrix had to be adjusted before it could be used;
- how uncertainty in the estimated correlations was treated; and
- whether the simulations reproduce the important features of the observed data.

Historical correlation may reflect shared stock abundance, but it can also come from changes in catchability, selectivity, fleet behaviour, or the way an index was standardized. Correlated random errors do not cover gradual drift or sudden changes in an index.

Equal weights for the five series also do not mean equal regional weights. Two series are Peruvian and three are associated with the south. The reported sensitivity compares all five indices with one case that excludes the acoustic index. It does not test leaving out each series in turn, balancing the regions, or weighting indices by their reliability. In their comments, external reviewers David Miller and Rosana Ourens said that "exploring different combinations of indices for use in the MP will be useful" (Appendix D of the SCW17 report; official PDF, p. 23).

Suggested action: publish the final equations and values used to simulate index errors. Then test gradual changes in catchability, sudden changes, missing survey years, revised data, region-balanced combinations, and weights based on index reliability. State clearly which uncertainties the correlation model covers and which important index issues remain outside the OM (Punt et al. 2016; ICES 2019; Harford et al. 2021).

5.8 [Medium] Show how the candidates changed and retain the reviewers' caveats

The report would benefit from a simple table linking the SCW17 run18/run20/run22 set to MP29/MP32 and then to MP43–MP48. Such a table could show what changed after the workshop: the five-index combination, the power-ramp correction, the 2019–2023 reference period, the new annual limits, and the retuning with 500 model draws.

The 24 July record was a draft. It did not select a final MP and listed the final runs based on 500 simulations and several tests of how the procedure would be applied as future work. The supplied material contains no later Task Team decision showing that it reviewed the 29 July results and the expanded eight-CMP set. Unless a later record is supplied, it may be clearest to describe these as comparisons added by analysts after the meeting and still awaiting Task Team and Scientific Committee review.

The report fairly summarizes the external reviewers' positive overall views, but presents two caveats more gently than the source: CMP design would have benefited from more stakeholder input, and earlier agreement on the assessment and reference points would have kept the discussion more focused. It would be helpful to retain those caveats clearly and keep the authors' response separate. Continued stakeholder input matters most for candidate design and the weighting of objectives (Punt et al. 2016; Harford et al. 2021).

6 References

Berger, Aaron M. 2019. “Character of Temporal Variability in Stock Productivity Influences the Utility of Dynamic Reference Points.” *Fisheries Research* 217: 185–97. https://doi.org/10.1016/j.fishres.2018.11.028.

Carruthers, Thomas R., and Adrian R. Hordyk. 2019. “Using Management Strategy Evaluation to Establish Indicators of Changing Fisheries.” *Canadian Journal of Fisheries and Aquatic Sciences* 76 (9): 1653–68. https://doi.org/10.1139/cjfas-2018-0223.

Dowling, Natalie A., Catherine M. Dichmont, George M. Leigh, et al. 2020. “Optimising Harvest Strategies over Multiple Objectives and Stakeholder Preferences.” *Ecological Modelling* 435: 109243. https://doi.org/10.1016/j.ecolmodel.2020.109243.

Harford, William J., Ricardo Amoroso, Richard J. Bell, et al. 2021. “Multi-Indicator Harvest Strategies for Data-Limited Fisheries: A Practitioner Guide to Learning and Design.” *Frontiers in Marine Science* 8: 757877. https://doi.org/10.3389/fmars.2021.757877.

ICES. 2019. *Workshop on Guidelines for Management Strategy Evaluations (WKGMSE2)*. No. 33. Vol. 1. ICES Scientific Reports. International Council for the Exploration of the Sea. https://doi.org/10.17895/ices.pub.5331.

Punt, André E., Doug S. Butterworth, Carryn L. de Moor, José A. A. De Oliveira, and Malcolm Haddon. 2016. “Management Strategy Evaluation: Best Practices.” *Fish and Fisheries* 17 (2): 303–34. https://doi.org/10.1111/faf.12104.

Rademeyer, Rebecca A., Éva E. Plagányi, and Doug S. Butterworth. 2007. “Tips and Tricks in Designing Management Procedures.” *ICES Journal of Marine Science* 64 (4): 618–25. https://doi.org/10.1093/icesjms/fsm050.

SPRFMO. 2020. *Jack Mackerel MSE Management Objectives*. COMM 8 Report Annex 8b. South Pacific Regional Fisheries Management Organisation. https://www.sprfmo.int/assets/Meetings/01-COMM/8th-Commission-2020-COMM8/reports/annex/Annex-8b-JM-MSE-Management-Objectives.pdf.

SPRFMO. 2025. *Report of the 13th Scientific Committee Meeting*. SC13 Report. South Pacific Regional Fisheries Management Organisation. https://sprfmo.int/assets/Meetings/02-SC/13th-SC-2025/SC13-REPORT-ADOPTED-13SEPT2025_posted-16Oct2025.pdf.

SPRFMO. 2026a. *Jack Mackerel Management Strategy Evaluation for SC14*. Draft SC14 working paper. South Pacific Regional Fisheries Management Organisation. https://sprfmo.github.io/jmMSE26/evidence/sc14-mse-report.html.

SPRFMO. 2026b. *Jack Mackerel MSE Implementation Pathway*. COMM 14 Report Annex 4b. South Pacific Regional Fisheries Management Organisation. https://www.sprfmo.int/assets/Meetings/01-COMM/14th-Commission-2026/Reports/Annex-4b-SC-JM-MSE-Implementation-Pathway.pdf.

SPRFMO. 2026c. *JM MSE Task Team Meeting Report*. Task Team meeting report. South Pacific Regional Fisheries Management Organisation. https://sprfmo.github.io/jmMSE26/.

SPRFMO. 2026d. *SCW17 MSE Workshop Report*. SCW17 Report. South Pacific Regional Fisheries Management Organisation. https://www.sprfmo.int/assets/Meetings/SC_WS/SCW17-JM-MSE-Workshop/SCW17-MSE-Workshop-report-v2.pdf.

Appendix A. Original Terms of Reference

The text below is reproduced unchanged from the independent desk-review consultancy materials supplied for this assignment.

A.1 Purpose

Undertake an independent, half-day desk review of the Jack Mackerel Management Strategy Evaluation (MSE), focusing primarily on the draft **SC14 MSE Report** and considering whether it provides a scientifically sound, clear, and balanced synthesis for Scientific Committee consideration.

A.2 Background material

The review should consider:

- The draft SC14 MSE Report as the principal document;
- The SCW17 MSE Workshop Report, which records the agreed MSE design and workshop conclusions; and
- Relevant subsequent analyses, simulation updates, supporting outputs, and documentation available from the Jack Mackerel MSE 2026 project site.

A.3 Scope of review

The expert is requested to assess:

1. Whether the SC14 MSE Report accurately and coherently reflects the workshop design, subsequent analyses, and current results.
2. Whether the operating models, candidate management procedures, tuning approach, robustness tests, and performance measures are described and interpreted appropriately.
3. Whether important uncertainties, limitations, trade-offs, and preliminary results are sufficiently transparent.
4. Whether the conclusions and matters presented for Scientific Committee consideration are supported by the analyses.
5. Whether any material inconsistencies, omissions, technical concerns, or potentially misleading statements should be addressed before submission.
6. Whether the report clearly distinguishes scientific findings from management choices, including the selection and weighting of performance measures and candidate procedures.

The review is not expected to reproduce the analyses, audit the complete codebase, or independently rerun the MSE.

A.4 Deliverable

A concise written review of approximately 1–2 pages containing:

- An overall assessment of the report's scientific adequacy and readiness;
- A short list of material issues requiring attention;
- Suggested corrections or clarifications, prioritized where possible; and
- Any recommendations for subsequent MSE development or documentation.

Comments may alternatively be provided directly against the report, accompanied by a brief overall assessment.

A.5 Level of effort

Up to half a working day, approximately four hours.

A.6 Timing

The review should be completed by August 8, 2026, to allow comments to be considered before finalization of the SC14 submission.