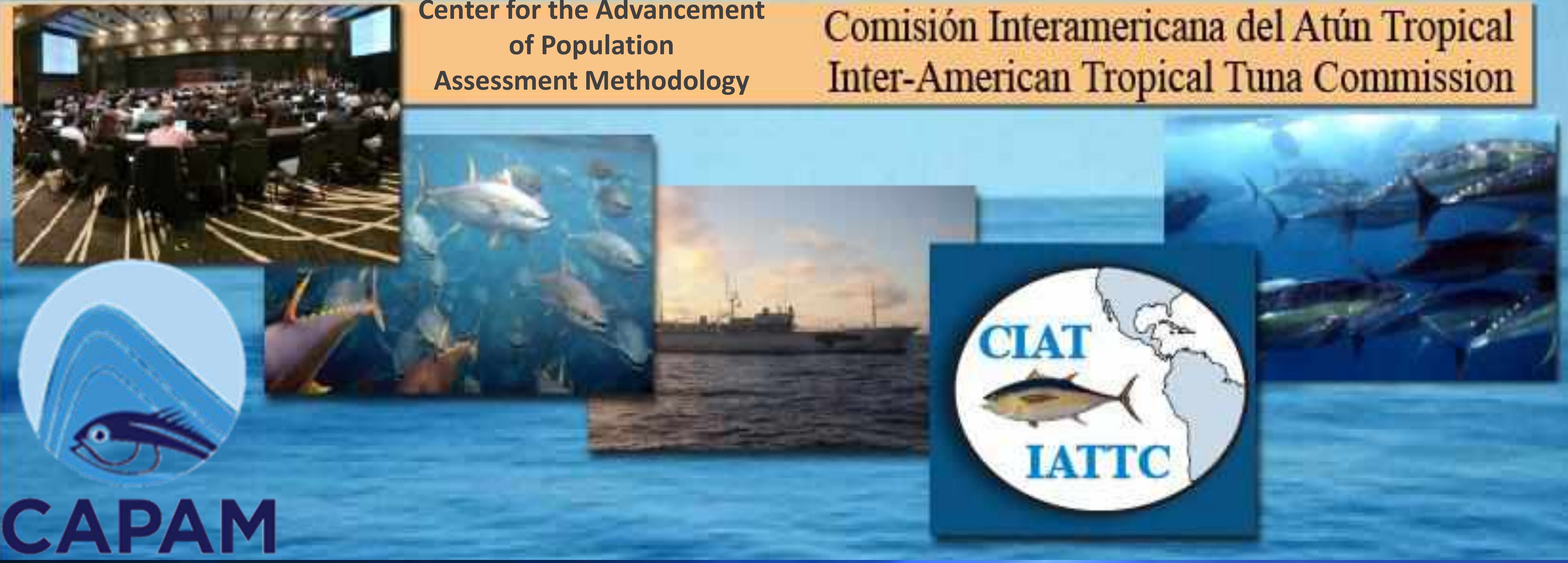


Center for the Advancement of Population Assessment Methodology

Comisión Interamericana del Atún Tropical
Inter-American Tropical Tuna Commission

Using diagnostics to fix and eliminate models when constructing an ensemble

Mark Maunder, Felipe Carvalho, Maia Sosa Kapur, and Andre Punt

Virtual meeting, 28 Nov – 2 Dec (8am to 11am - San Diego)

HOW DO WE INTERPRET & USE DIAGNOSTIC RESULTS?

Model selection

Model weighting,

Characterizing uncertainty

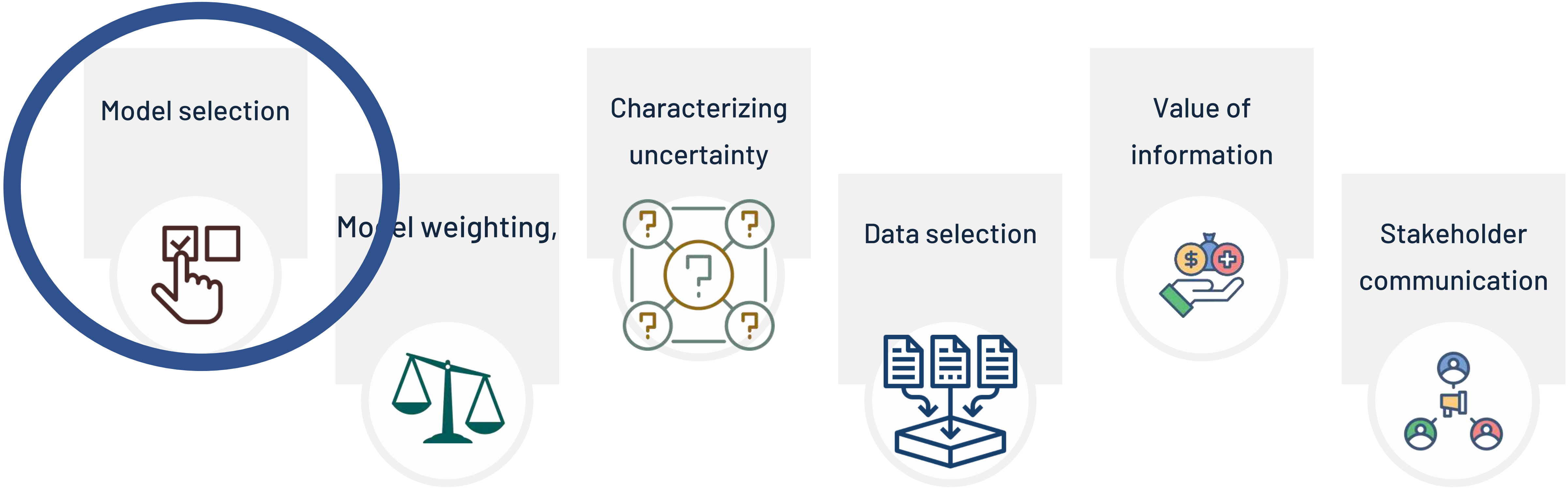
Data selection

Value of information

Stakeholder communication

we make decisions without clear, consensus-based thresholds.

HOW DO WE INTERPRET & USE DIAGNOSTIC RESULTS?



Model selection

Model weighting,

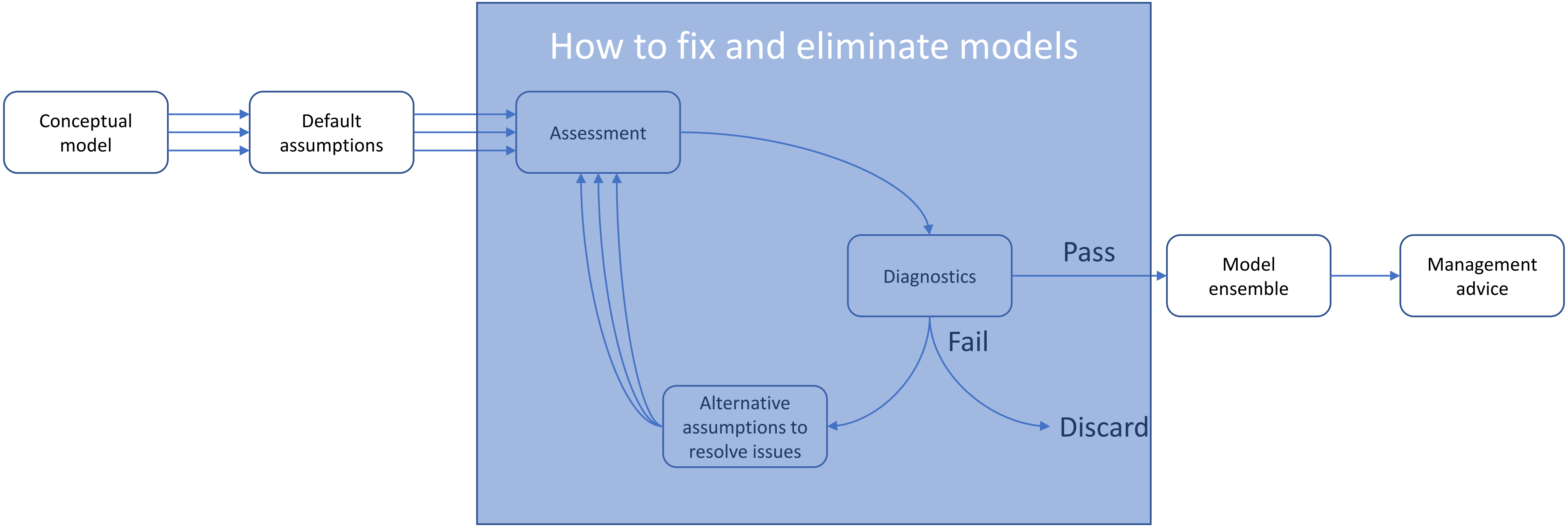
Characterizing uncertainty

Data selection

Value of information

Stakeholder communication

we make decisions without clear, consensus-based thresholds.



How to fix and eliminate models

Conceptual model

Default assumptions

Assessment

Diagnostics

Pass

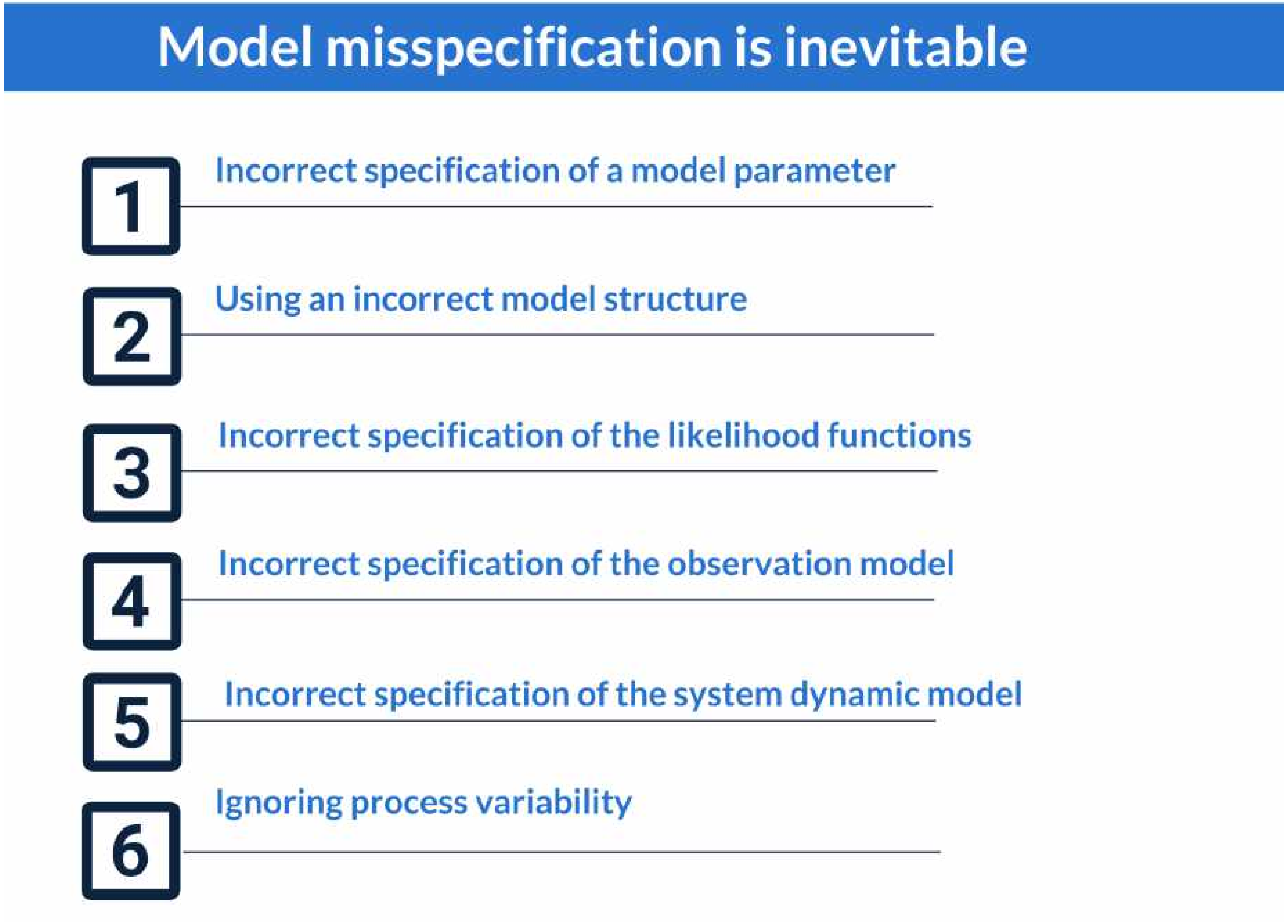
Fail

Alternative assumptions to resolve issues

Discard

Model ensemble

Management advice



Model misspecification is inevitable

1 Incorrect specification of a model parameter

2 Using an incorrect model structure

3 Incorrect specification of the likelihood functions

4 Incorrect specification of the observation model

5 Incorrect specification of the system dynamic model

6 Ignoring process variability

7 Unrepresentative or poorly “standardized” data

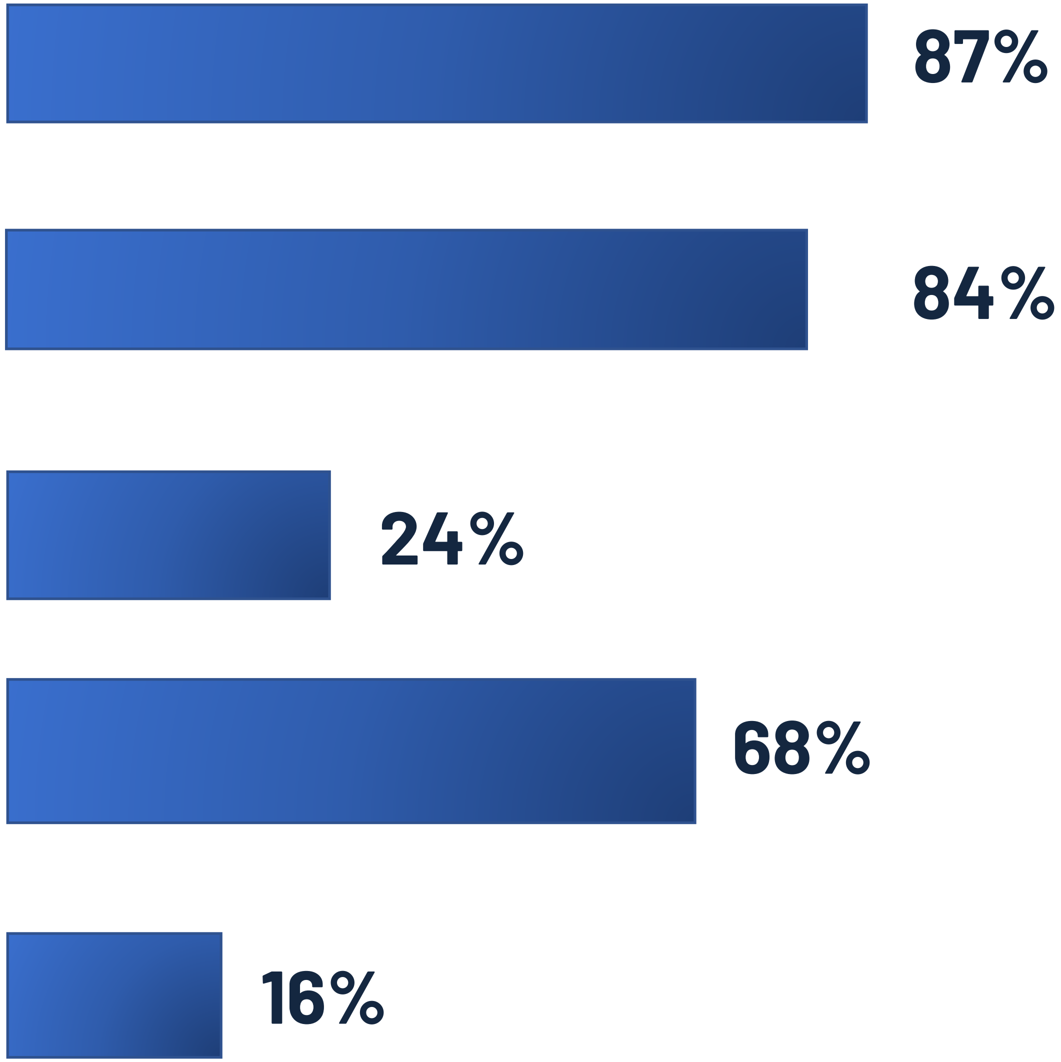
IATTC – CAPAM diagnostics workshop

- Standard diagnostics
  - Evaluation of residuals
  - Effective sample sizes and variances
  - Cross validation and hindcasting
  - Bayesian model checking
- Stock Assessment specific
  - R0 likelihood component profile
  - Age-structured Production Model (ASPM)
  - Catch curve analysis
  - Epriical selectivity
- Plausibility
  - Parameter values
  - Results

Commonly-used diagnostics

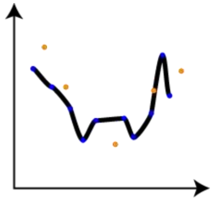
- Goodness-of-Fit
- Retrospective
- Hindcasting, MASE
- Likelihood profiles
- ASPM, catch curve analyses

How often do you use them?



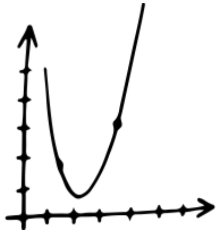
				Participation: 46%
	What diagnostics/statistics…			
	do you routinely perform to assess your integrated models?	should be the **minimum standard** to evaluate the **performance of the “base case model” or “reference set of models”**?	should **a model pass to be acceptable to use for management** advice?	could be used for **weighting models** in an ensemble to produce inference for management advice
None or Diagnostics should not be used	0%	0%	2%	2%
Simple residuals or Pearson residuals	87%	62%	52%	30%
PIT, simulation/ quantile residuals	11%	33%	27%	37%
Addressing variances	57%	52%	57%	37%
R0 Likelihood profile	68%	56%	38%	29%
ASPM	16%	19%	13%	14%
Retrospective analysis	84%	86%	76%	63%
Hindcasting/prediction skill evaluation	24%	57%	52%	65%
“Red-face test” = subjective evaluation of the plausibility of the results	65%	56%	57%	41%
Other	19%	13%	16%	22%

Commonly-used diagnostics

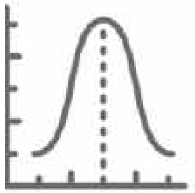


Goodness-of-Fit

Retrospective analysis, hindcasting, MASE



Likelihood profiles



ASPM, catch curve analyses

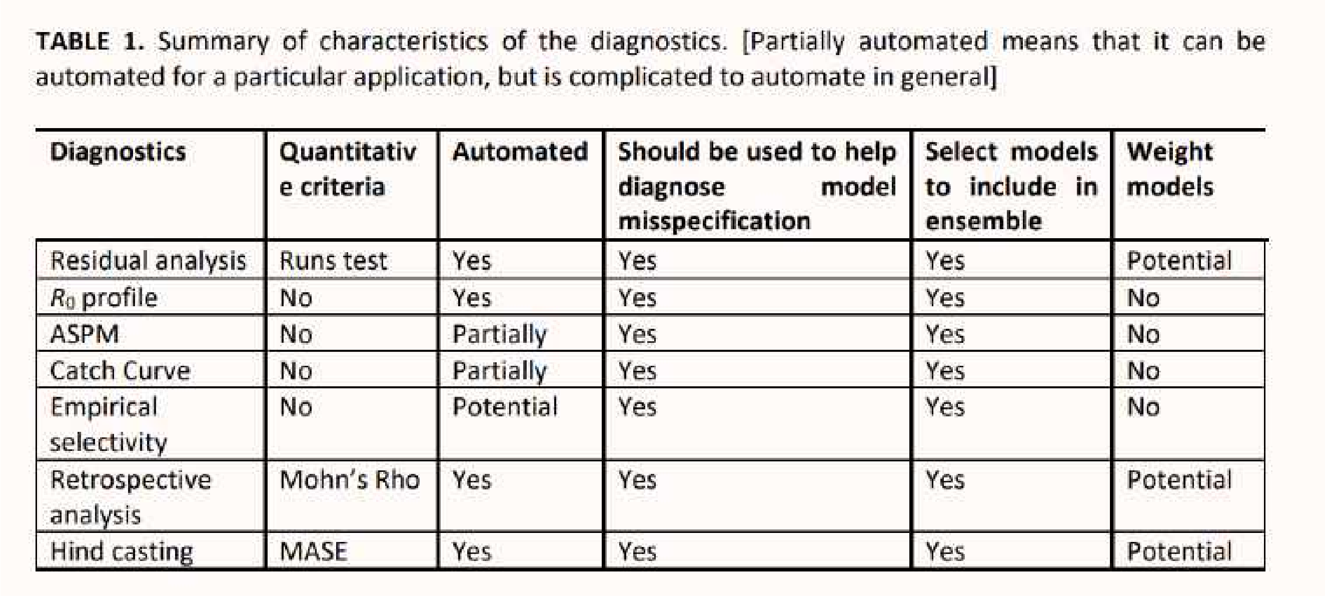


TABLE 1. Summary of characteristics of the diagnostics. [Partially automated means that it can be automated for a particular application, but is complicated to automate in general]

Diagnostics	Quantitativ e criteria	Automated	Should be used to help diagnose model misspecification	Select models to include in ensemble	Weight models
Residual analysis	Runs test	Yes	Yes	Yes	Potential
R_0 profile	No	Yes	Yes	Yes	No
ASPM	No	Partially	Yes	Yes	No
Catch Curve	No	Partially	Yes	Yes	No
Empirical selectivity	No	Potential	Yes	Yes	No
Retrospective analysis	Mohn's Rho	Yes	Yes	Yes	Potential
Hind casting	MASE	Yes	Yes	Yes	Potential

IATTC – CAPAM diagnostics workshop conclusions

- Current model diagnostics are good for model development, but less so for other purposes
- Provide tools to detect if there is a problem with the model
- Can’t identify the exact source of the problem
- Do not guarantee that the model is an adequate representation of the "true" population dynamics nor whether the estimates of management quantities are reliable
- The development and understanding of diagnostics are not at the stage that diagnostics can be used for weighting models.
- Current metrics from the diagnostics (e.g., Mohn’s rho from retrospective analysis and MASE from hind casting) cannot be turned into P(Model) or made consistent with AIC.
- Alternative validation-based metrics should be explored, e.g., a “prediction likelihood” based on prediction errors from hindcast cross-validation (c.f., Dormann et al., 2018)

Diagnostics

- Failure criteria
- Indications of what is misspecified and how to fix it
- Rejection criteria

Diagnostics

- Failure criteria: Limited often they visual and subjective
- Indications of what is misspecified and how to fix it: Mostly vague or unknown
- Rejection criteria: Same as failure criteria after alternative assumptions tried

Convergence

- Failure criteria
  - Hessian matrix is not positive define
  - Gradient is large > 0.1?
  - Parameter on a bound within 0.1%? of bound
  - Large parameter CV > 0.5?
  - Parameter correlation is large > 0.5?
  - Jittering leads to different optima
- Indications of what is misspecified
  - Lack of information about a parameter

Plausibility

- Failure criteria
  - F < 0.05 F > 2.0?
  - M outside range of empirical relationships
  - h < 0.6? for a pelagic spawner (or use meta analysis)
  - Application specific
- Indications of what is misspecified
  - Parameters compensate for other misspecifications
- Data to compare it with should be used in the model or as a prior
-

Residual analysis

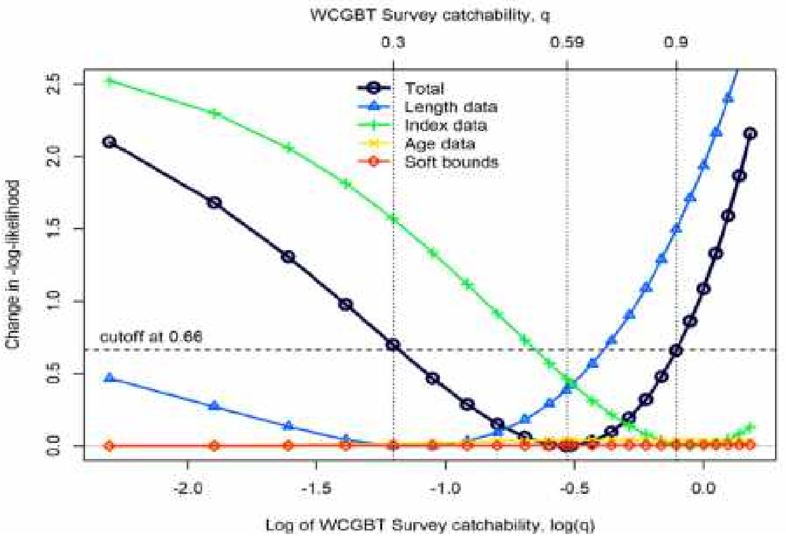
- Failure criteria
  - Examined visually and subjectively
  - Runs test
  - SDNR ≠ 1 (standard deviation of the normalized residual)
- Indications of what is misspecified
  - Conduct runs tests over age/length, time, and cohort.
  - Age/length or consecutive groups of ages/lengths
    - Misspecified selectivity curve, growth model, or other process
  - Year or block of years
    - Changes in selectivity, growth, or other processes
  - Cohort
    - Cohort targeting or cohort-specific growth or other processes.
  - Patterns in residuals may indicate unmodelled temporal variation in system or sampling processes.
  - Allowing variation in one process can eliminate residual patterns caused by time-variation in other parameters
  - SDNR > 1
    - The input sample sizes have not correctly accounted for the way the data were collected
    - the model is too stiff
  - SDNR < 1
    - The sample size was based on the wrong measure (e.g. tows sampled)

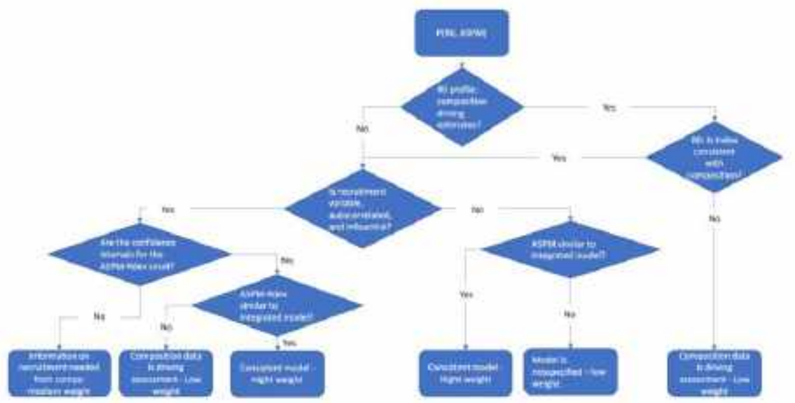
Empirical selectivity

- Failure criteria
  - Visual and subjctive
- Indications of what is misspecified
  - Too inflexible selectivity
  - Temporal trends in selectivity

Likelihood component profile

- Failure criteria
  - Wang and Maunder’s quantitative metric
  - Maunder et al. (2020) flow chart combining the R0 profile and the ASPM
  - Low power to detect model misspecification (Carvalho et al. 2017)
- Indications of what is misspecified
  - Conflict may be with data not directly associated with misspecification




R0, ASPM
R0 profile: composition driving estimates?
Yes
No
R0: is index consistent with composition?
Yes
No
Is recruitment variable, autocorrelated, and influential?
Yes
No
Are the confidence intervals for the ASPM index small?
No
No
ASPM-Rdev similar to integrated model?
No
Yes
ASPM similar to integrated model?
Yes
No
Information on recruitment needed from comps - medium weight
Composition data is driving assessment - Low weight
Consistent model - High weight
Consistent model High weight
Model is misspecified – low weight
Composition data is driving assessment - Low weight

Age structured production model (ASPM)

- Failure criteria
  - Visual and subjective or confidence bounds
  - When ASPM-Rdev differs from the full assessment, conflict between comp and index data
  - When ASPM differs from ASPM-Rdev means that recruitment dev information is needed to interpret the index of abundance (which comes from composition)
- Indications of what is misspecified
  - Stock dynamics are recruitment-driven
  - The stock has not yet declined to where catch is influencing abundance
  - Indices of relative abundance are not proportional to abundance
  - CPUE index may not be sufficiently standardized to detect the impact of the catch
  - The model is incorrectly specified
  - Data are unrepresentative (biased)

Catch-curve diagnostic

- Failure criteria
  - Visual and subjective
  - High type I error, indicates problems when none exist (Carvalho et al. 2017)
- Indications of what is misspecified
  - Changes in selectivity (or M) or growth (length comp)

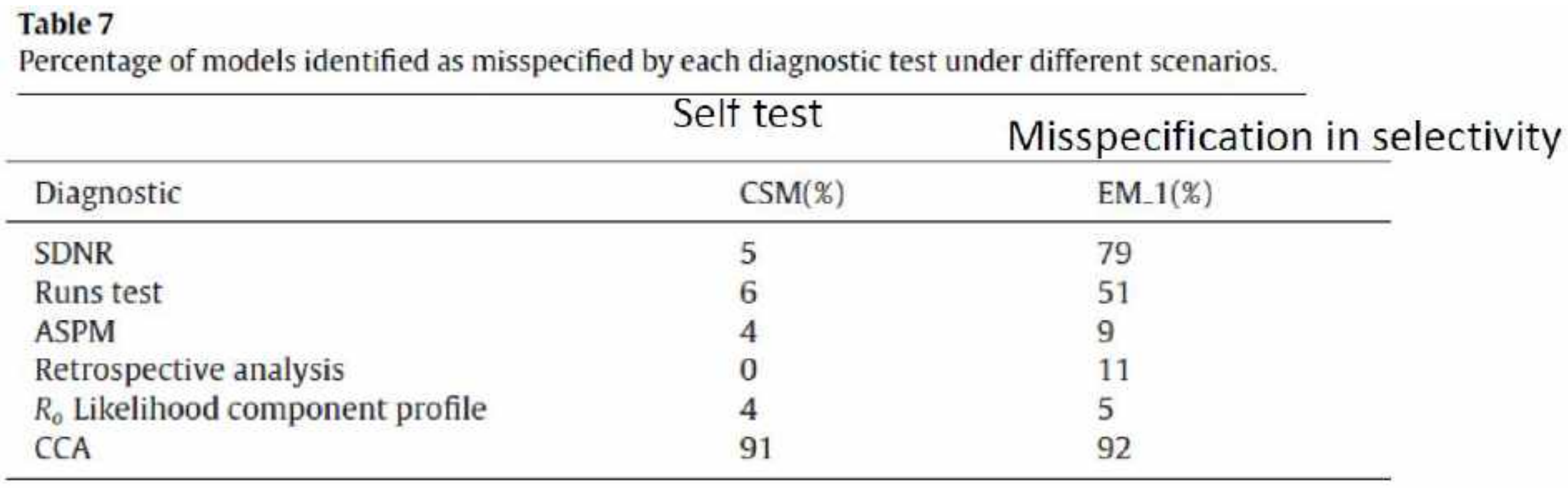


Table 7
Percentage of models identified as misspecified by each diagnostic test under different scenarios.

	Self test	Misspecification in selectivity
Diagnostic	CSM(%)	EM_1(%)
SDNR	5	79
Runs test	6	51
ASPM	4	9
Retrospective analysis	0	11
R_0 Likelihood component profile	4	5
CCA	91	92

Retrospective analysis

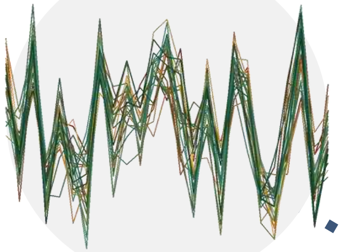
- Failure criteria
  - Mohn’s rho: ICES uses range [-0.15-0.2] as acceptable (ICES, 2019)
  - Rho-adjustement
  - Determine if adjustment factor is outside the uncertainty estimates (Legualt)
  - Evaluate if the Mohn’s rho uncertainty interval from a parametric bootstrap overlaps zero (Legault)
- Indications of what is misspecified
  - Errors in catch time series
  - Processes are time varying but not modelled
  - Single large error: ignore
  - Large but random: uncertainty, so take into consideration in management
  - Moderate to large pattern: need to fix model
  - Adding time varying process may reduce retrospective error but may not improve the management related quantity (Szuwalski et al. 2018)

Cross validation/Hindcasting

- Failure criteria
  - Root mean squared error (RMSE)
  - Mean absolute scaled error (MASE)
  - Others
  - Simple cross validation does not deal with autocorrelation
- Indications of what is misspecified
  - Stock is recruitment driven
  - Production function is not estimable from the data
  - Production function changes over time
  - Model is misspecified
  - Can inform whether there is overfitting or bias

What are model diagnostics good for?

Simulation Approach



Simulate population trajectories

bootstrap CPUE & length composition data sets using simulated recruitment deviations

run many mis-specified assessments

growth, selectivity, steepness, mortality fixed to inaccurate values (high or low), all possible combinations of correct/incorrect

calculate error

MARE in terminal biomass, depletion, and management quantities

run diagnostic tests

tabulate performance on diagnostics from Cookbook

explore trends

How consistently do diagnostics pass/fail mis-specified models?

How does diagnostic performance square with error?

Summary

	Fully-specified	Automated	Threshold	Notes
Convergence	Yes	Generally	Yes	
Residual patterns	Yes	Yes*	Yes	Move to PIT residuals
Variances	Yes	Yes	No	We really don’t what to do fix the problem
Retrospective patterns	Yes	Yes	Yes*	
R_0 profile	Yes	Yes	No	Issues with the recruitment deviations
ASPM	Yes	No	No	Need for recruitment deviations
Catch curve	Yes	No	No	
Hindcasting	Perhaps	No	Yes	Many ways to do this. Also, what does MARE > 1 mean practically
Empirical selectivity	Yes?	Yes	N/A	