On automating assessment model diagnostics and the need for simulation testing

Henning Winker[1*], Massimiliano Cardinale[2], Francesco Masnadi[3,4], Laurence Kell[5], Iago Mosqueira[6], Felipe Carvalho[7]

*Henning.Winker@ec.europa.eu

CAPAM, January 2022

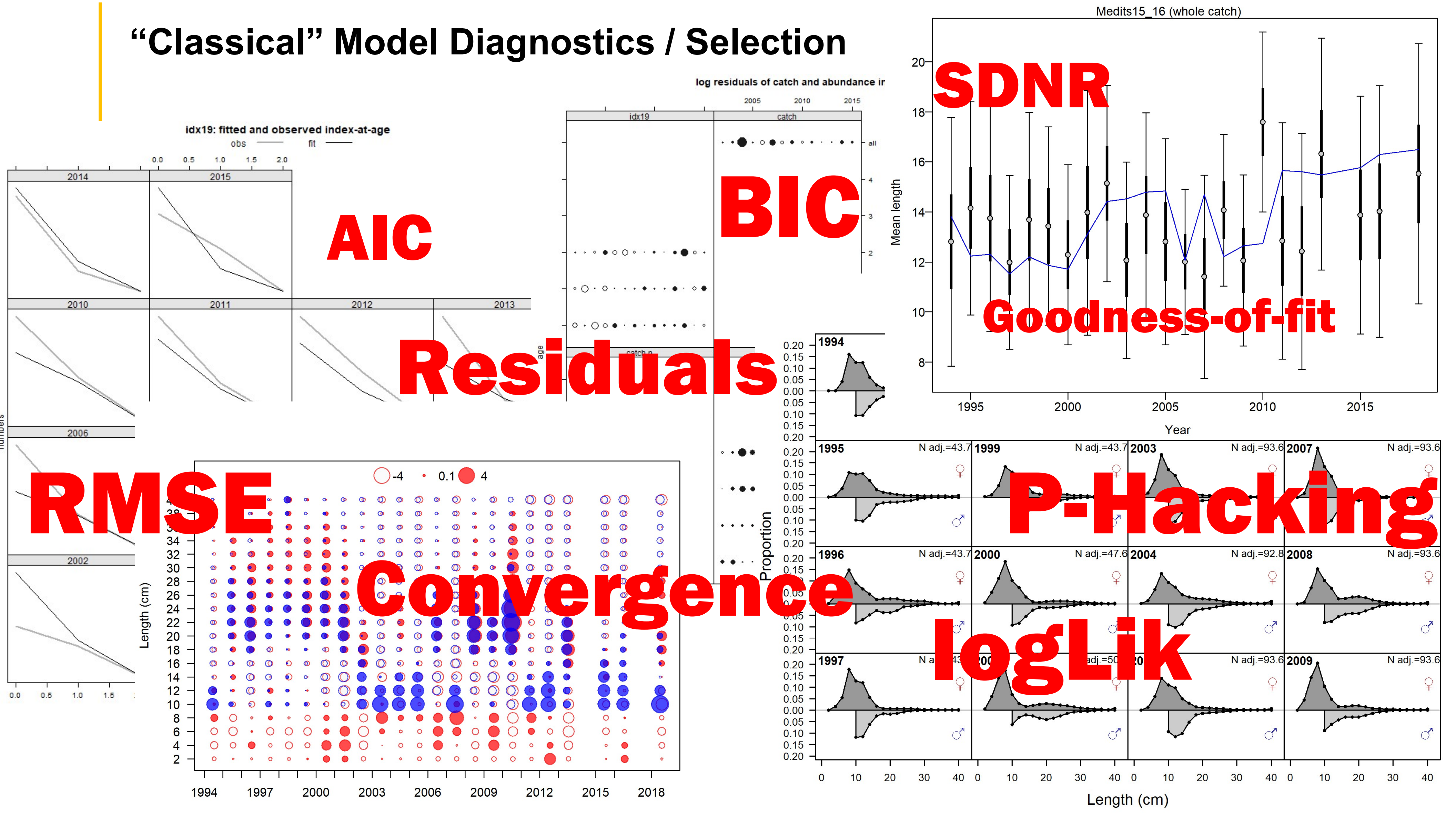
[1]Joint Research Centre (JRC), European Commission, TP 051, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy
[2]Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Lysekil, Sweden
[3]National Research Council, Institute for Marine Biological Resources and Biotechnology (CNR IRBIM), Ancona, Italy
[4]Department of Biological, Geological, and Environmental Sciences (BiGeA), Alma Mater Studiorum—University of Bologna, Bologna, Italy
[5]Centre for Environmental Policy, Imperial College London, Weeks Building, 16-18 Princes Gardens, London SW7 1NE, UK
[6]Wageningen Marine Research, Haringkade 1, 1976CP IJmuiden, the Netherlands
[7]NOAA Fisheries, Pacific Islands Fisheries Science Center, Honolulu, HI, United States

European Commission

Outline

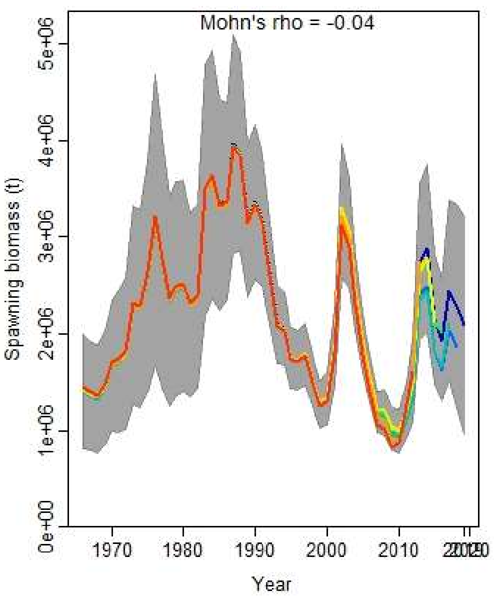
- A very brief evolution of model diagnostic approaches
- Some real world examples
- Towards automation of model diagnostic in benchmarks
- The need for simulation testing
- A simplified trial how a simulation could look like
- Remarks on model plausibility

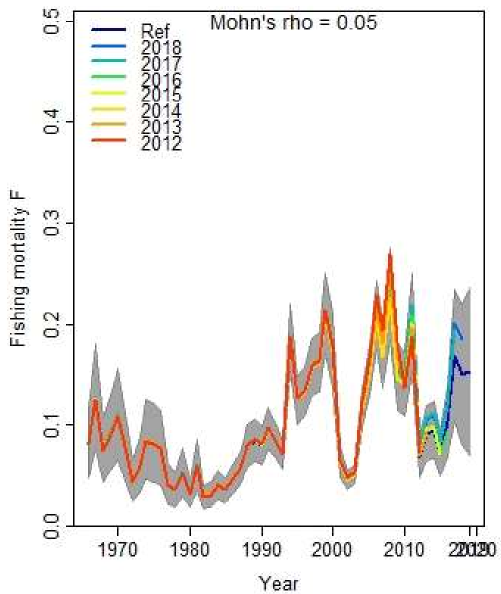
“Classical” Model Diagnostics / Selection

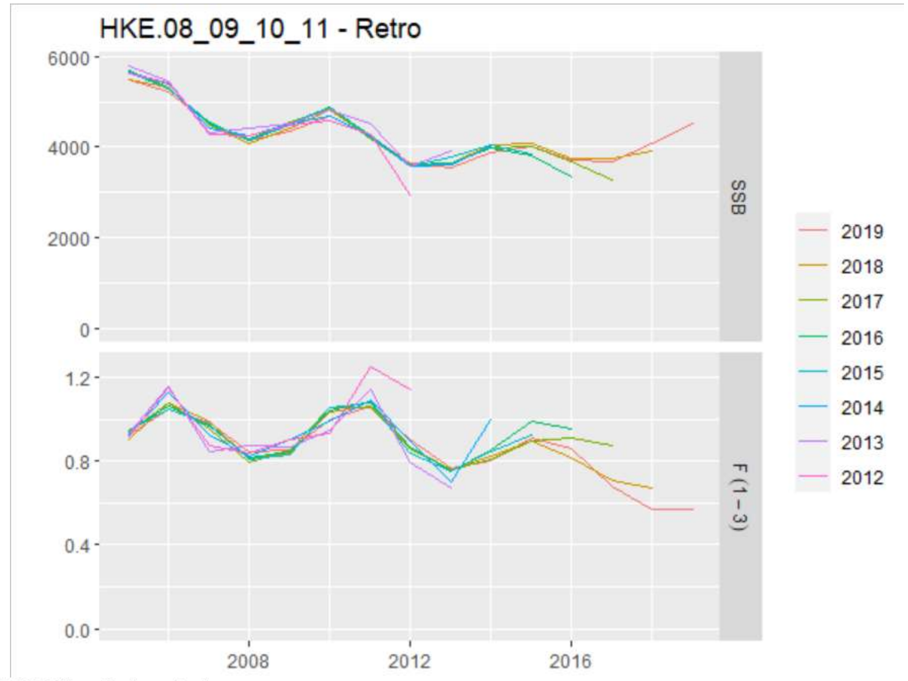


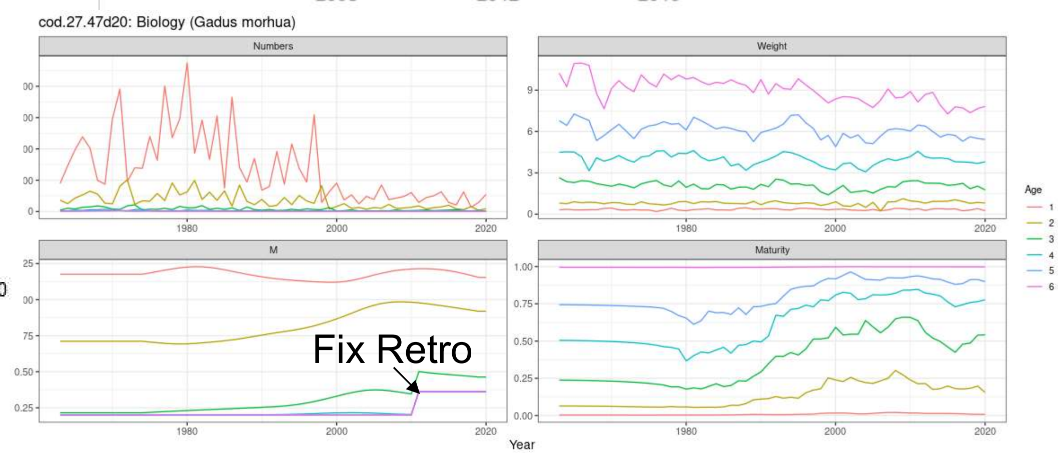
“Neo-Classical”: Model diagnostics

+ Mohn’s ρ









“Modern”: Model diagnostics

+ Data conflicts

M.N. Maunder, K.R. Piner / Fisheries Research 192 (2017) 16–27

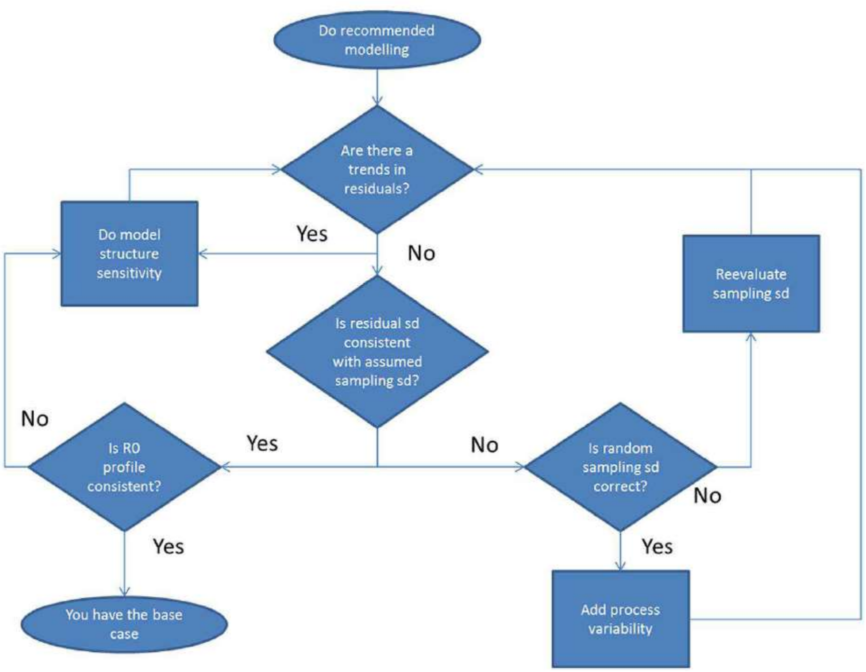
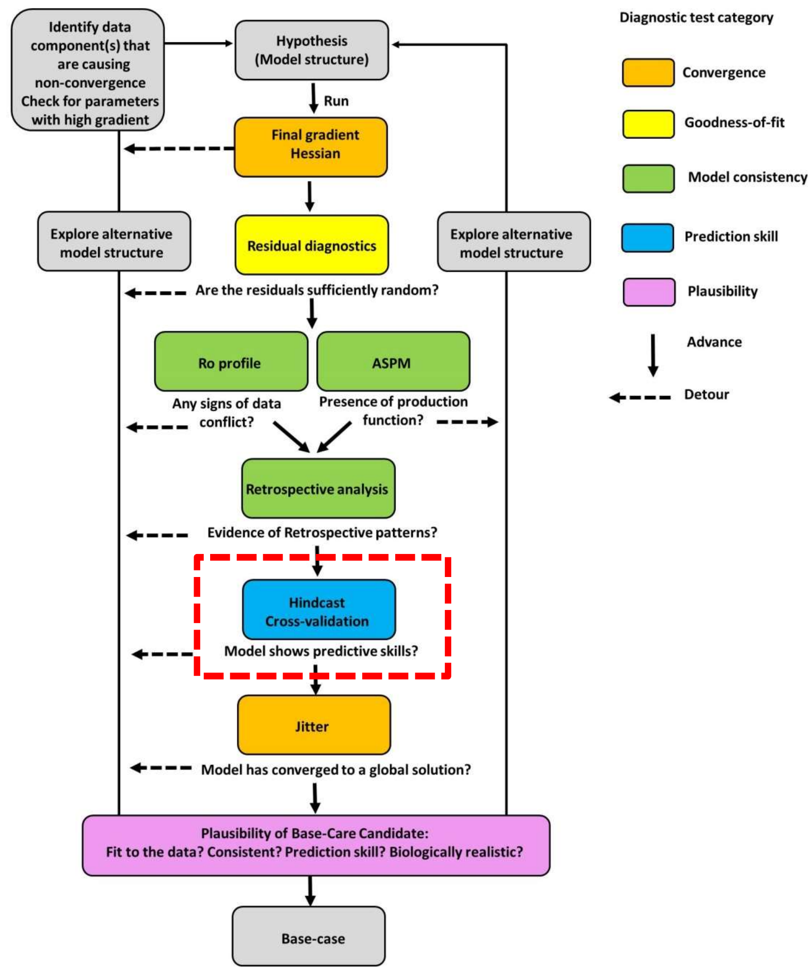


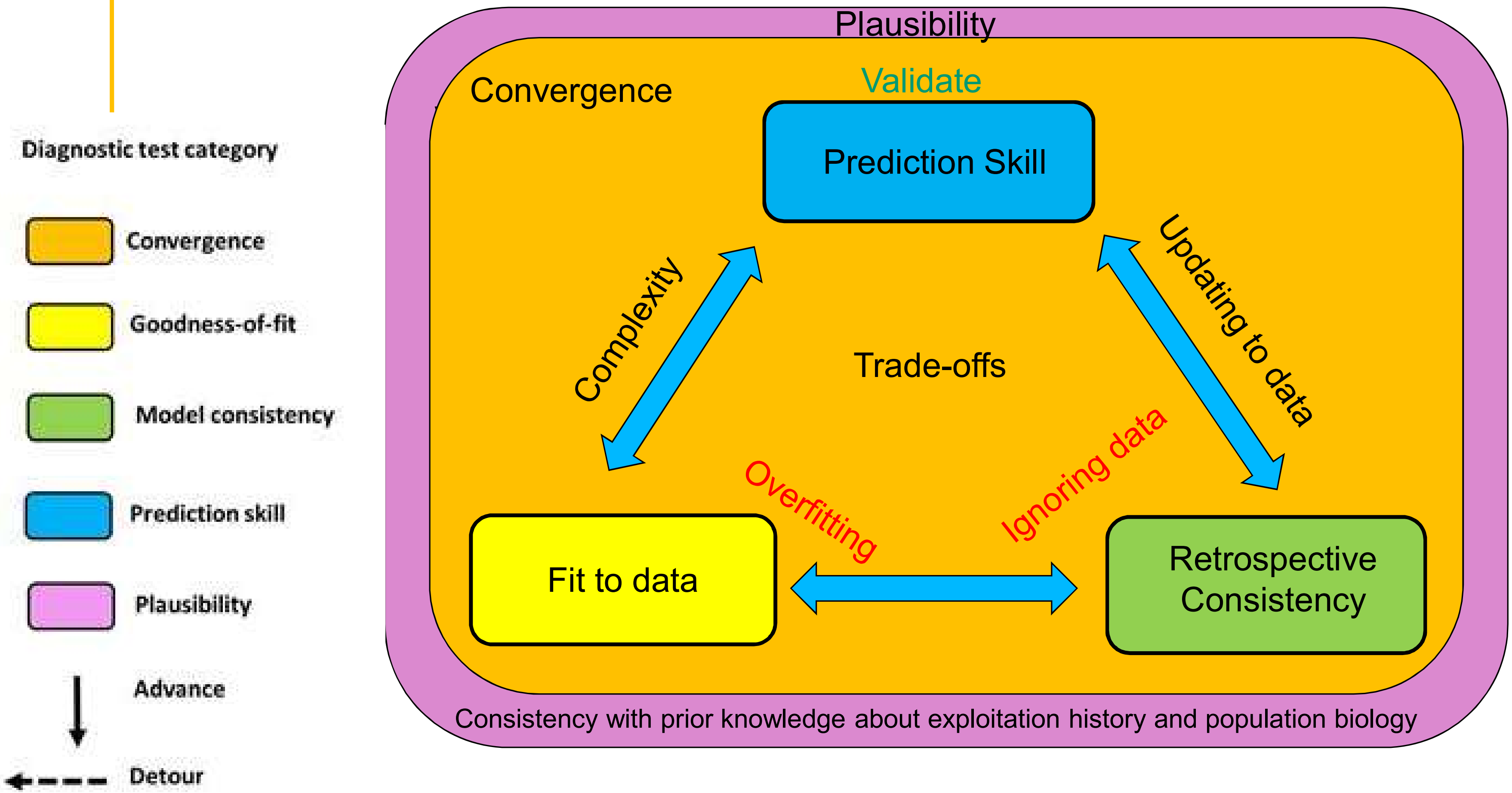
Fig. 5. Flow chart depiction of a simplification of the recommended modelling approach.

+ Prediction Skill



n
sion

“Modern” Model diagnostics



European Commission

'ss3diags': Stock Synthesis

R package hosted on Github

https://github.com/JABBAmodel/ss3diags

Winker H, Carvalho F, Cardinale M, Kell LT

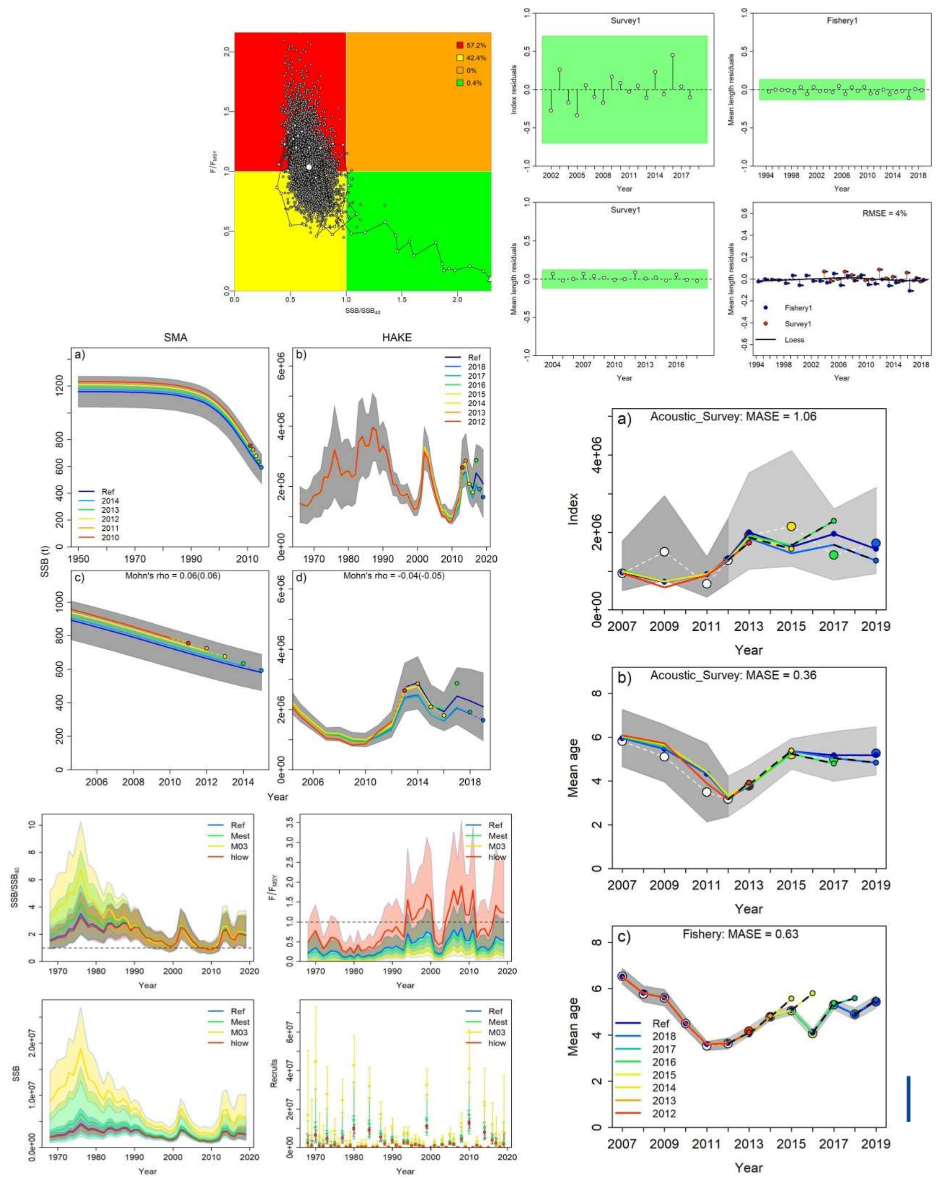
The R package `ss3diags` automates a set of model diagnostics tools plots and ensemble weighting functions for Stock Synthesis

User guidelines for Advanced Model Diagnostics with ss3diags

Hennig Winker (JRC-EC) Felipe Carvalho (NOAA)
Massimiliano Cardinale (SLU) Laurence Kell (Sea++)

20 August, 2021

Contents

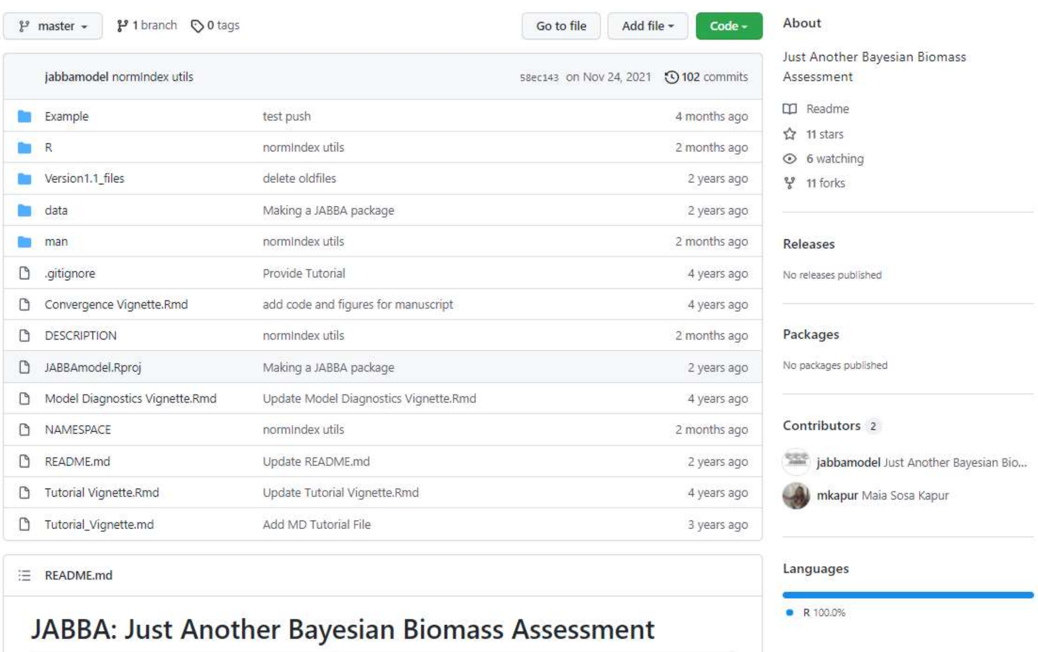


'JABBA': First was JABBA

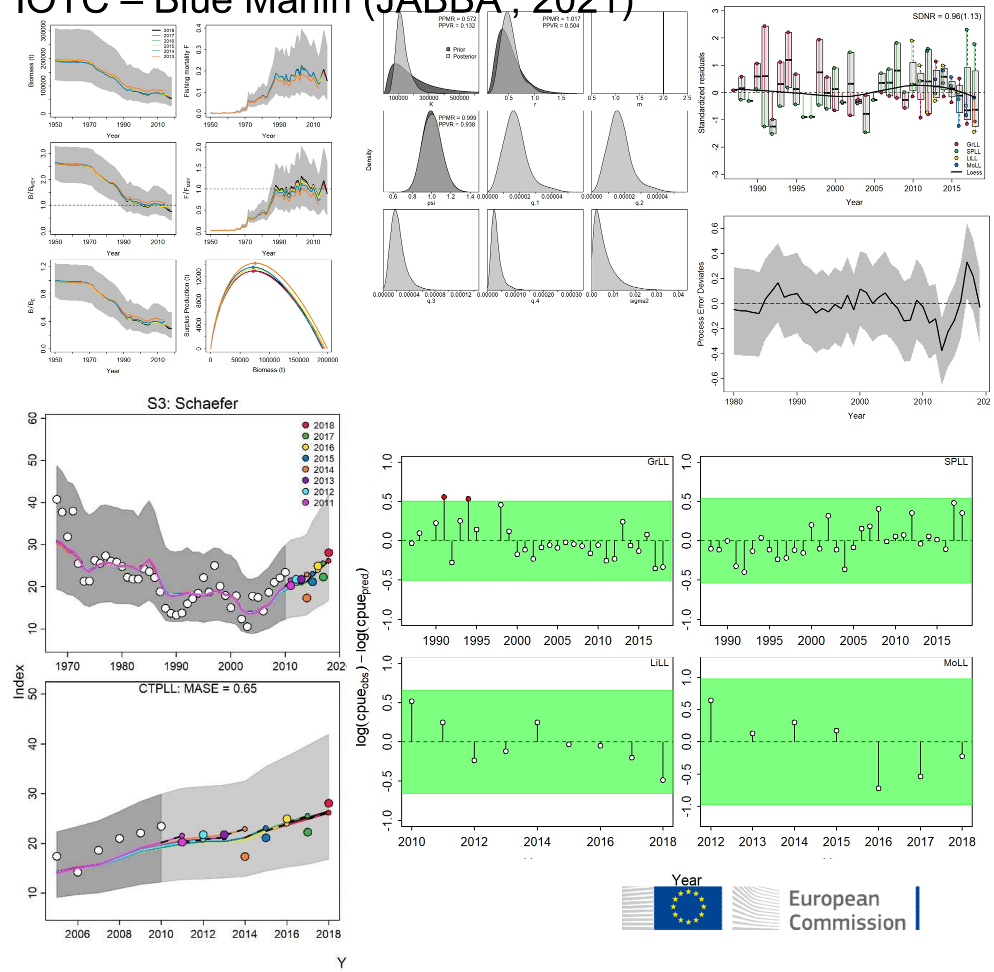
R package hosted on Github

https://github.com/JABBAmodel/JABBA

Winker H, Carvalho F, Kapur M



ICCAT - S.Atl. Albacore (JABBA, 2020)
ICCAT – Med. Swordfish (JABBA , 2020)
ICCAT – Med. Albacore (JABBA , 2021)
IOTC – Blue Marlin (JABBA , 2021)

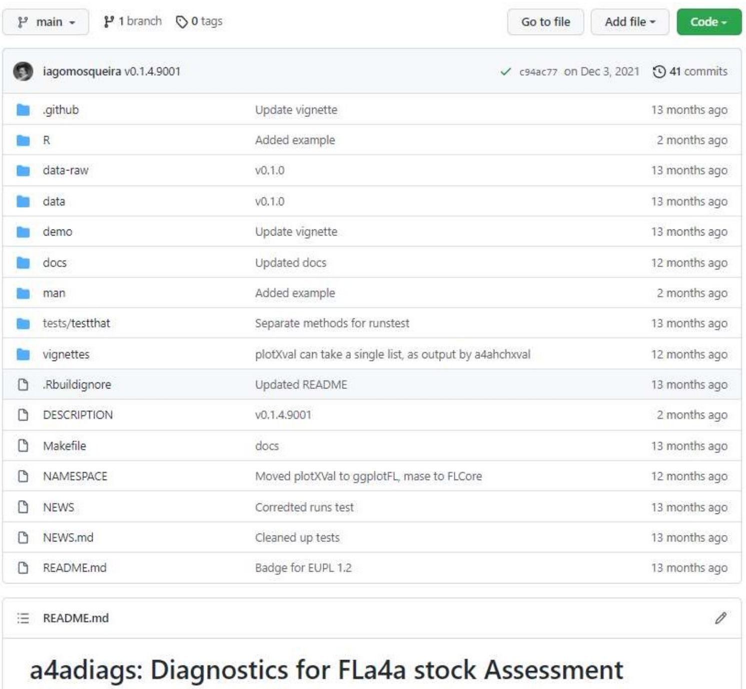


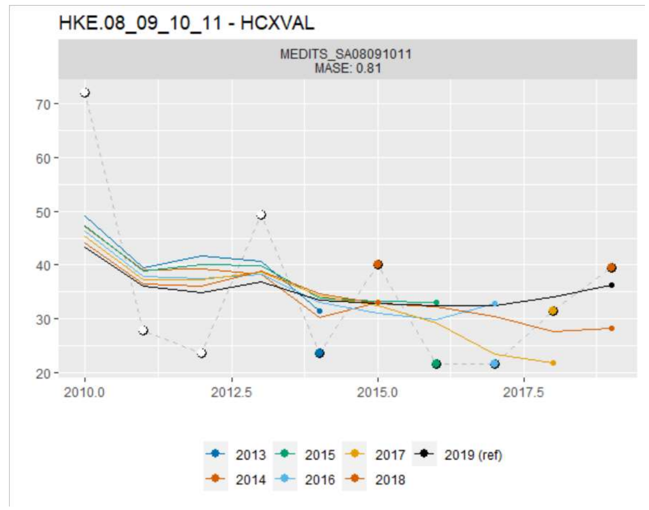
‘a4adiags’: Statistical-Catch-Age

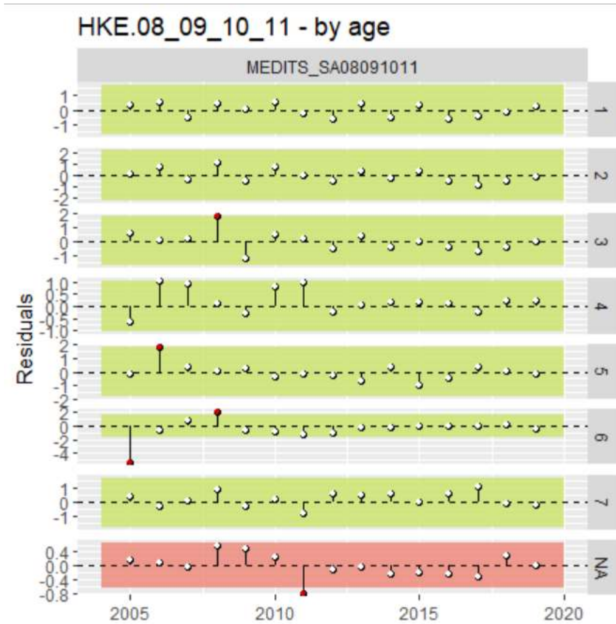
R package hosted on Github

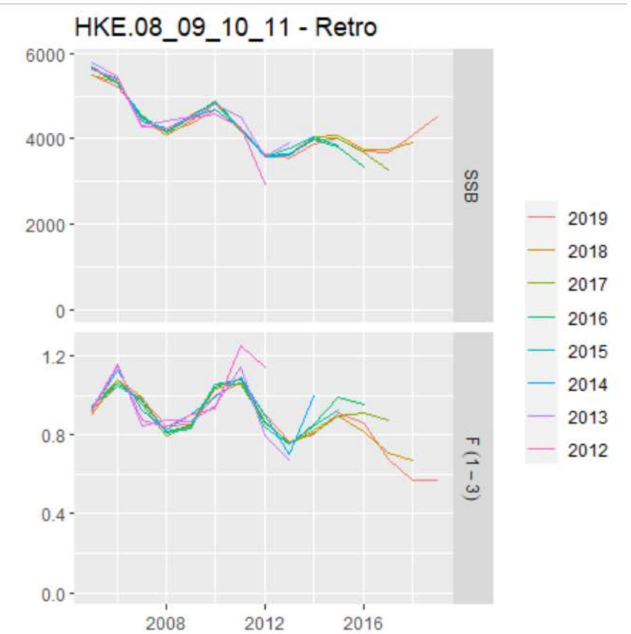
https://github.com/flr/a4adiags

Mosqueira I & Winker H







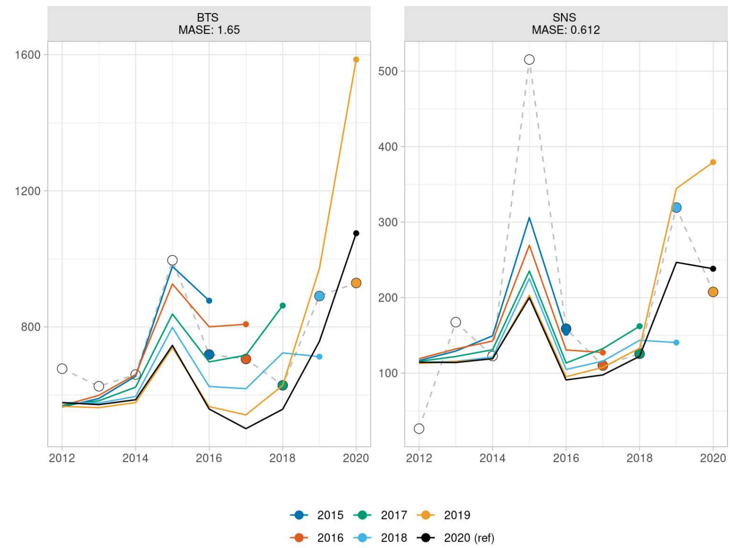


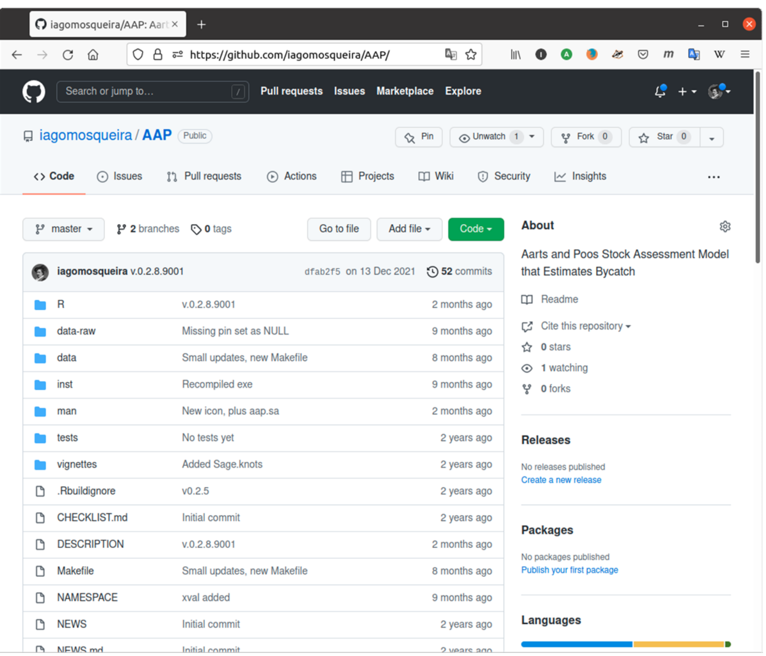
‘AAP’: Statistical-Catch-Age (Aaarts & Poos)

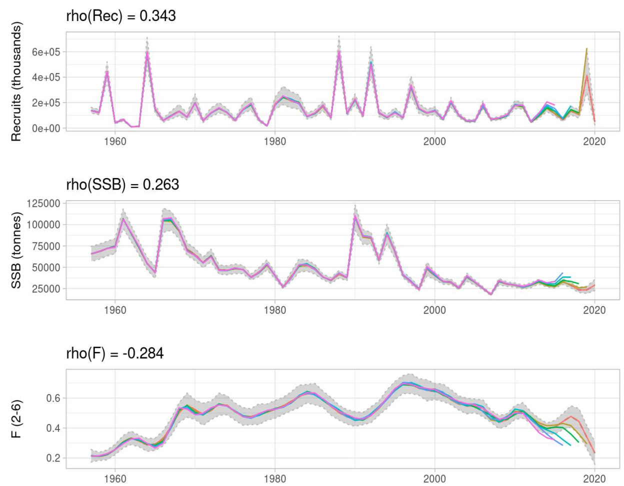
R package hosted on Github

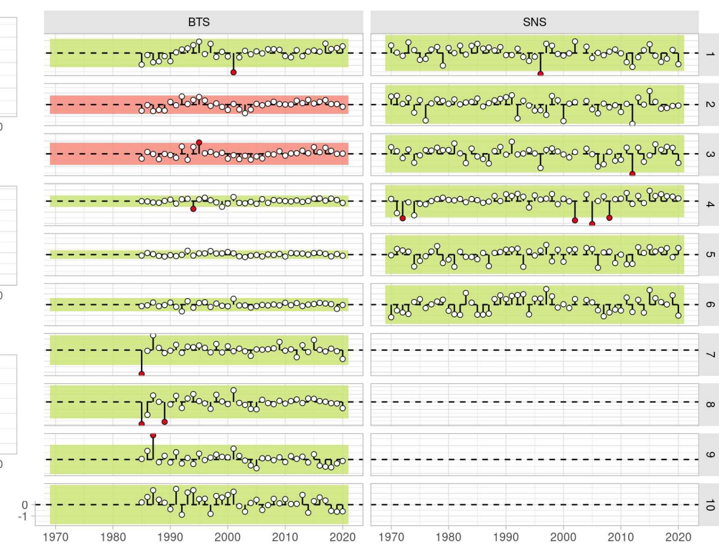
https://github.com/iagomosqueira/AAP

Mosqueira I & Poos JJ



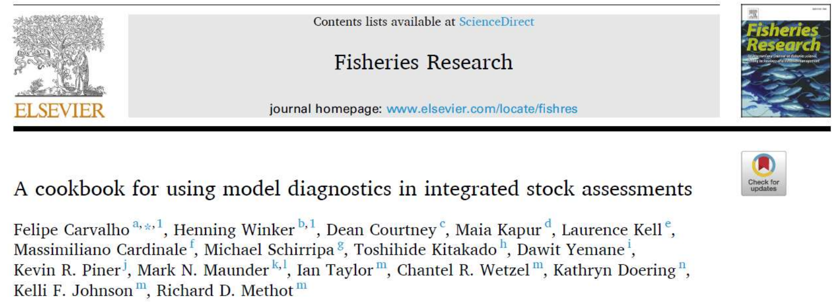






Cookbook in action

1. ICCAT Shortfin mako **2019** (SS3: Benchmark evaluation)
2. GFCM Sicilian Hake **2019** (SS3: Benchmark)
3. ICES - Eastern Baltic Cod **2019** (SS3: Benchmark)
4. ICES - Bothnian Herring **2019** (SS3: Benchmark)
5. IOTC Skipjack tuna **2020** (SS3: Benchmark Ensemble)
6. GFCM - Adriatic Sole **2020** (SS3: Benchmark Ensemble)
7. ICES - Iberian sardine **2020** (SS3: Base-Case development)
8. ICCAT - Med Swordfish **2020** (JABBA, Benchmark)
9. ICCAT - S.Atl. Albacore **2020** (JABBA, Benchmark)
10. IOTC Yellowfin tuna **2021** (SS3: Benchmark Ensemble)
11. ICCAT Bigeye tuna **2021** (**SS3** JABBA, mpb: Benchmark Ensemble)
12. WCPFC - Blue marlin **2021** (SS3, Benchmark Ensemble)
13. IOTC Blue shark **2021** (**SS3**,JABBA: Benchmark)
14. IOTC Albacore **2021** (SS3, MSE grid)
15. IOTC Swordfish **2021** (SS3, MSE grid)
16. SEDAR **2021** Gulf of Mexico Scamp Grouper (SS3, Benchmark)
17. ICCAT - Mediterranean Albacore **2021** (JABBA, Benchmark)
18. Sweden - Vendace **2022** (SS3: Benchmark Ensemble)
19. ICES - Pandalus shrimp **2022** (SS3, Benchmark Ensemble)
20. GFCM - Adriatic Hake **2022** (SS3: Updated Assessment)
21. GFCM – Mantis Shrimp **2022** (SS3: Advice Assessment)


Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

ELSEVIER

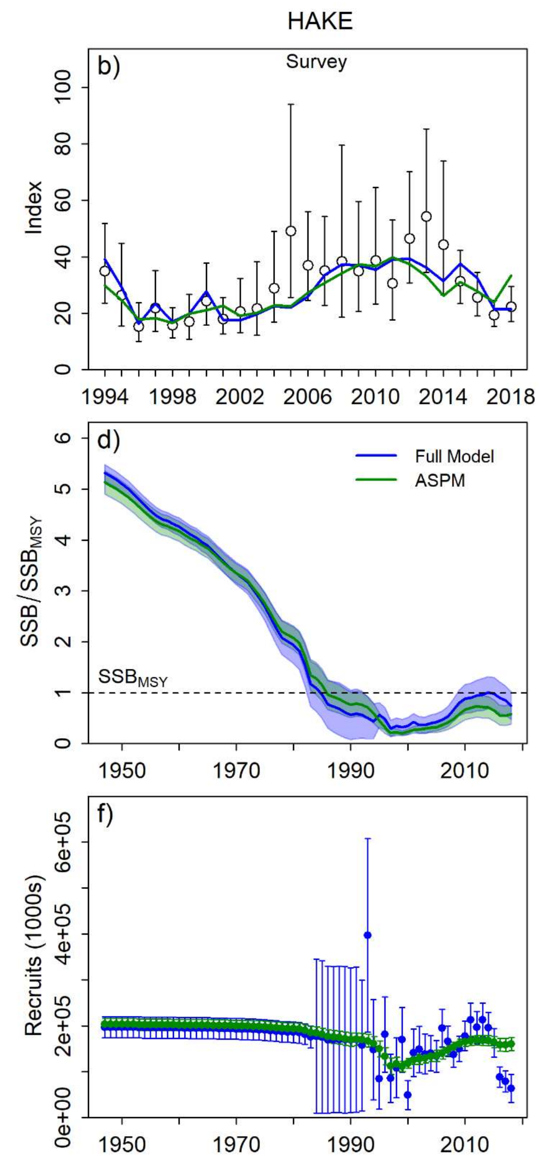
A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho[a,*,1], Henning Winker[b,1], Dean Courtney[c], Maia Kapur[d], Laurence Kell[e], Massimiliano Cardinale[f], Michael Schirripa[g], Toshihide Kitakado[h], Dawit Yemane[i], Kevin R. Piner[j], Mark N. Maunder[k,l], Ian Taylor[m], Chantel R. Wetzel[m], Kathryn Doering[n], Kelli F. Johnson[m], Richard D. Methot[m]

Cookbook in action (2019)

General Fisheries Commission of the Mediterranean (GFCM)
Stock: Sicilian Hake
Model: Stock Synthesis

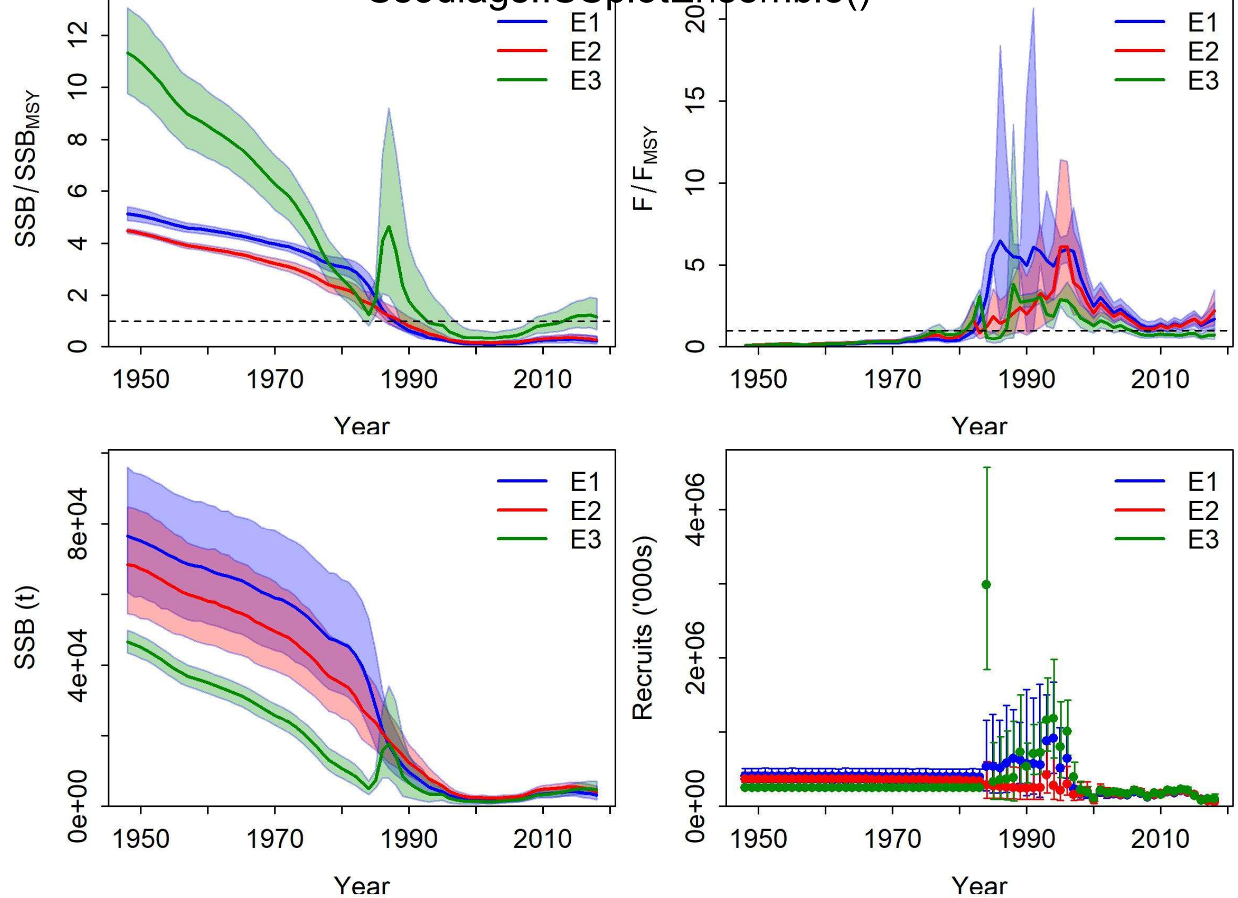
Diagnostics:
- Runs Tests
- Joint-Residual Plots (RMSE)
- ASPM
- R0 Profiling
- Retrospective Analysis
- Hindcast cross-validations



Cookbook in action (2019)
GFCM
Stock: Sicilian Hake
Model: Stock Synthesis

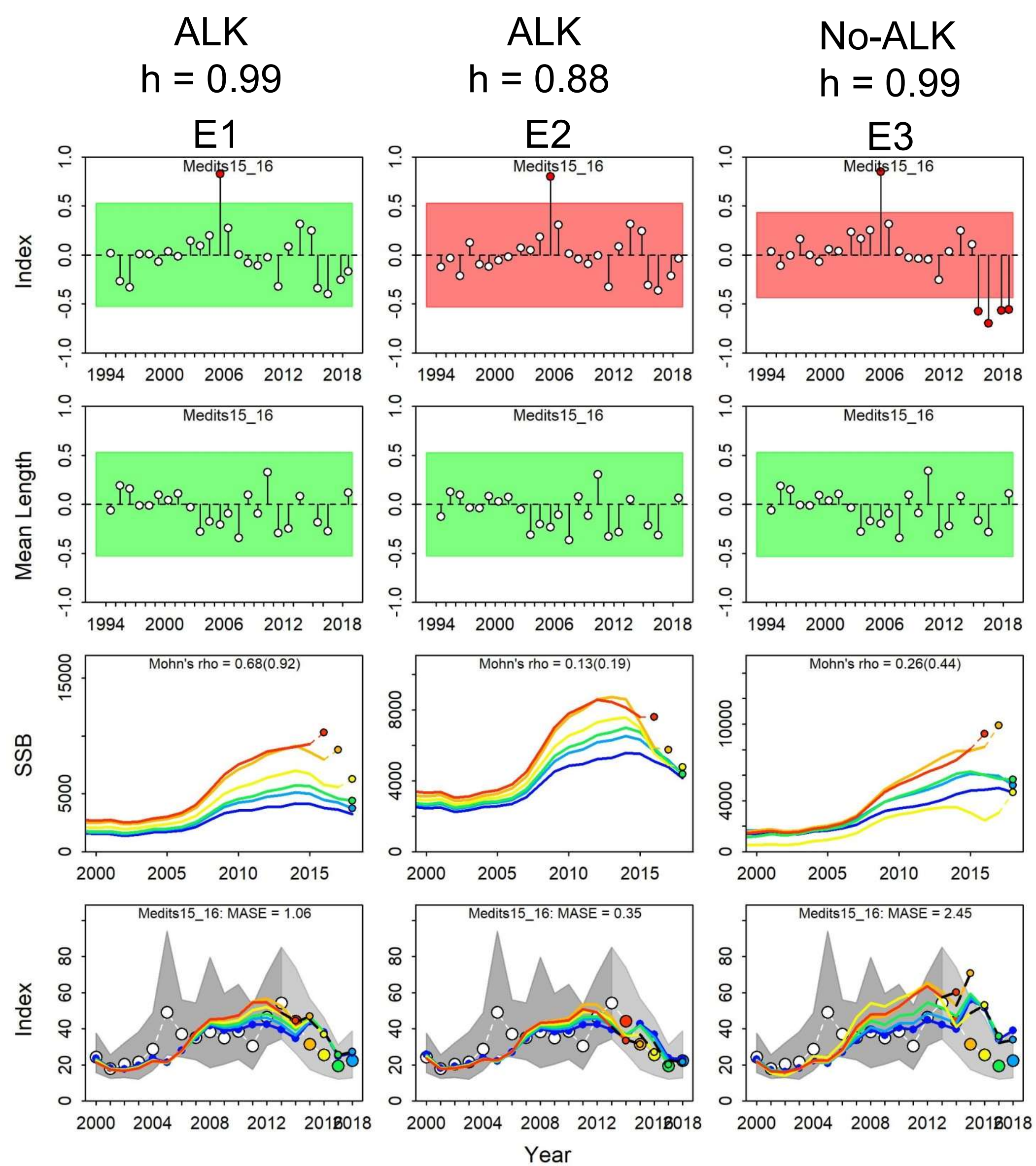
	ALK h = 0.99	ALK h = 0.88	No-ALK h = 0.99
Model	E1	E2	E3
Convergence	**9.74E-05**	**2.02E-04**	9.41E-05
Likelihood	**1289**	1304	**1761**
AIC	**2653**	**2737**	3001

Ss3diags::SSplotEnsemble()



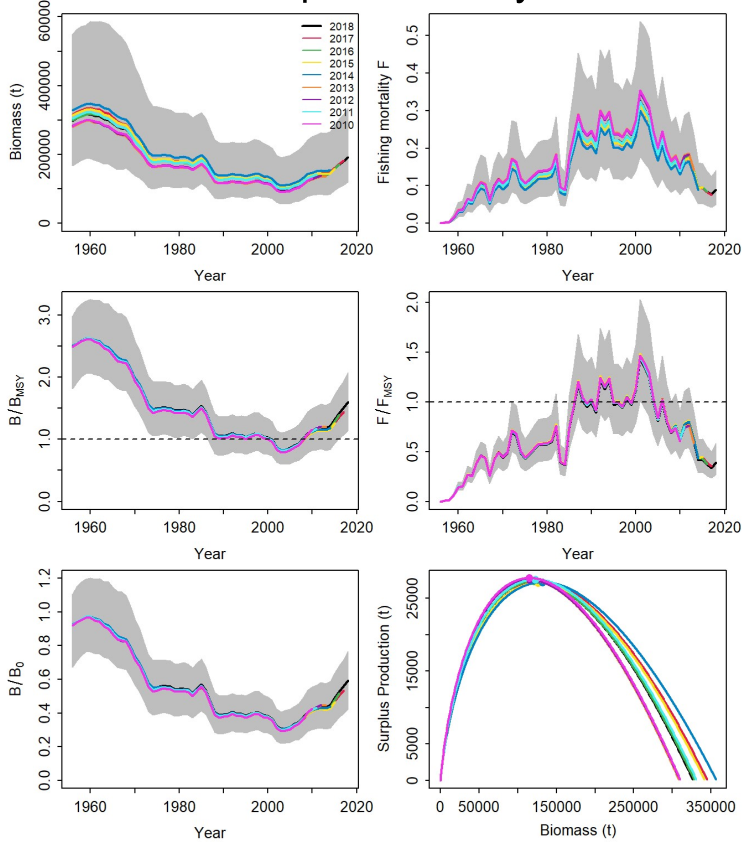
Cookbook in action (2019)
GFCM
Stock: Sicilian Hake
Model: Stock Synthesis

Model	E1	E2	E3
Convergence	**9.74E-05**	**2.02E-04**	9.41E-05
Likelihood	**1289**	1304	**1761**
AIC	**2653**	**2737**	3001
RunsTest Index	**P**	**F**	**F**
Runs Test Length	**P**	**P**	**P**
Restrospective bias	**F**	**P**	**F**
Forecast bias	**F**	**P**	**F**
MASE	**F**	**P**	**F**
Weighting Score	0.4	**0.8**	**0.2**

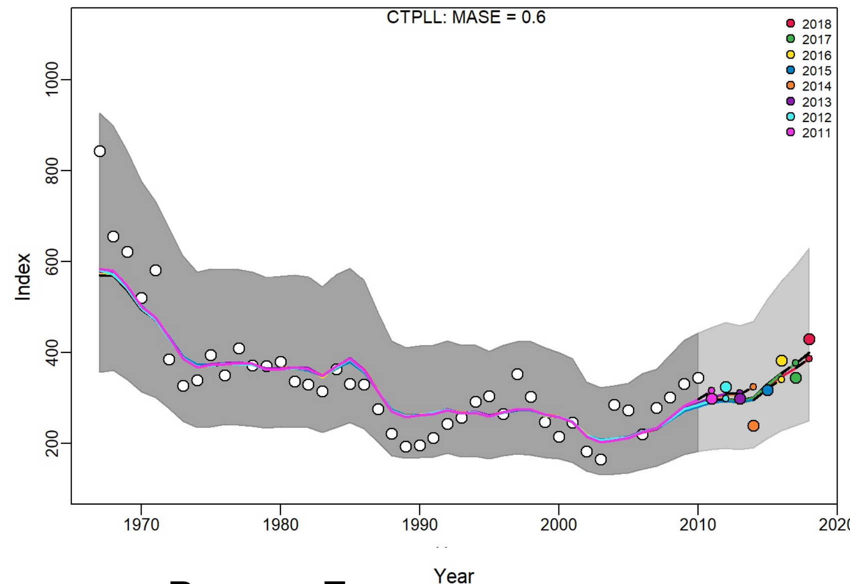


Cookbook in action (2020)
ICCAT
Stock: South Atlantic Albacore
Model: JABBA

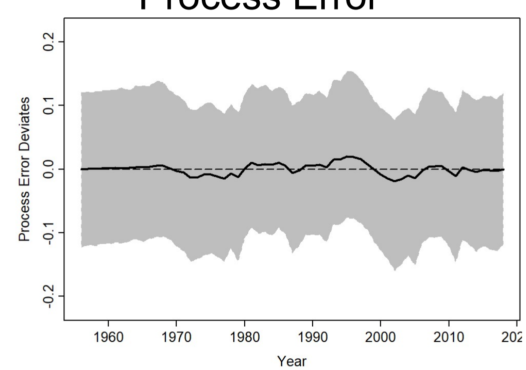
Retrospective Analysis



Hindcast Cross-Validation



Process Error



European Commission

Towards automation of model diagnostics

FAO-GFCM WGSAD

24-28/01/2022

Stock assessment of *Solea solea* in GSA 17 Benchmark update

Masnadi F., Cardinale M., Carbonara P., Donato F., Sabatini L., Pellini G., Scanu M., Dragičević B., Armelloni E. N., Martinelli M., Domenichetti F., Santojanni A., Colella S., Giovanardi O., Raicevich S., Fabi G., Marceta B., Vrgoč N., Isajlovic I., Milone N., Arneri E., Scarcella G.

CNR IRBIM ISTITUTO PER LE RISORSE BIOLOGICHE E LE BIOTECNOLOGIE MARINE

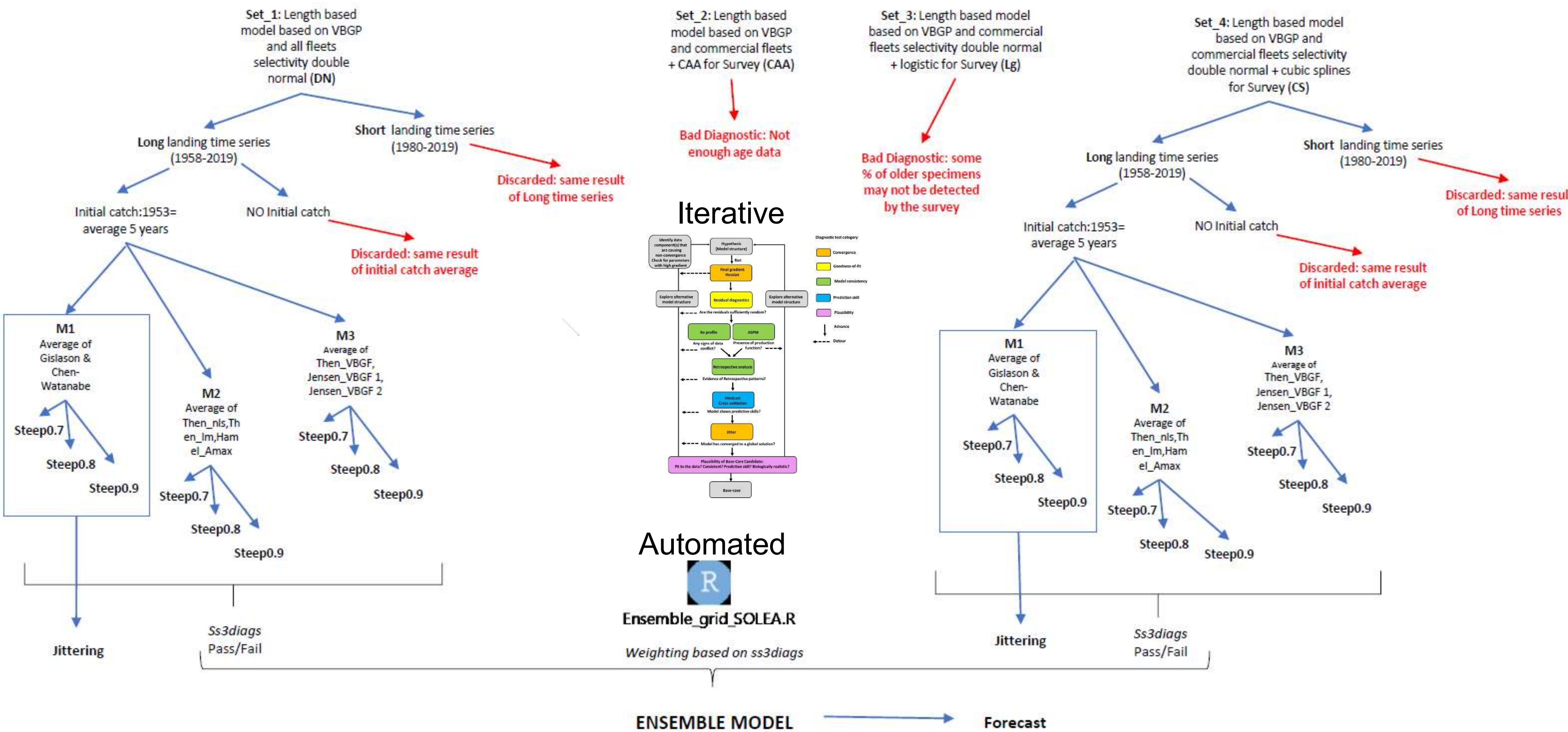
FishMed-PhD

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

COISPA Tecnologia & Ricerca Stazione sperimentale per lo studio delle risorse del mare

SLU Swedish University of Agricultural Sciences

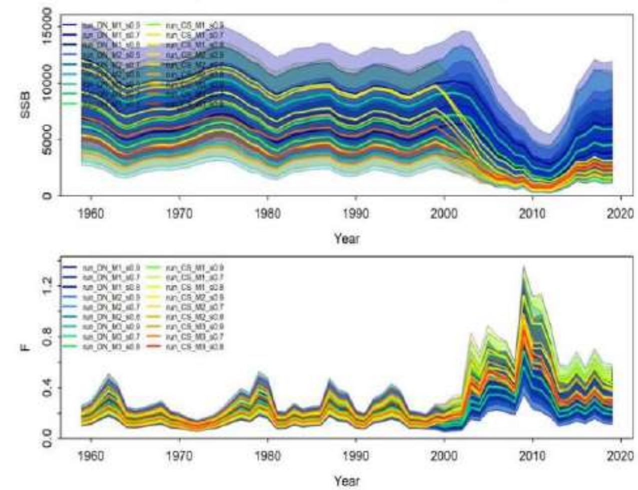
Stock Assessment workflow for Benchmark of *Solea solea* in GSA 17



- ***Stock Synthesis 3 + Ensemble approach (delta-MVLN)***

To directly integrate uncertainty in several key parameters, a range of alternative models to be stitched together was selected through diagnostics

Grid of alternative runs



Diagnostic test category

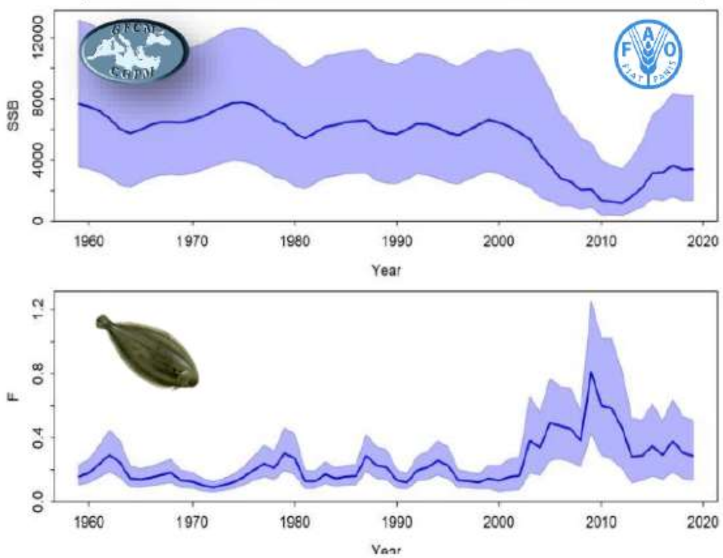
- Goodness-of-fit
- Information sources and structure
- Prediction skill
- Convergence
- Plausibility

ss3diags

A multivariate lognormal Monte-Carlo approach for estimating structural uncertainty about the stock status and future projections for Indian Ocean Yellowfin tuna

Henning Winker[1,], John Walter[2], Massimiliano Cardinale[1], and Dan Fu[3]*

Scientific Advisory Committee On Fisheries (SAC)
Working Group on Stock Assessment of Demersal Species (WGSAD)
Benchmark session for the assessment of common sole in GSA 17



Diagnostic - Convergence & stability (ref run 1)

A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho[a,*,1], Henning Winker[b,1], Dean Courtney[c], Maia Kapur[d], Laurence Kell[e], Massimiliano Cardinale[f], Michael Schirripa[g], Toshihide Kitakado[h], Dawit Yemane[i], Kevin R. Piner[j], Mark N. Maunder[k,l], Ian Taylor[m], Chantel R. Wetzel[m], Kathryn Doering[n], Kelli F. Johnson[m], Richard D. Methot[m]

jabbamodel / ss3diags Public

Stability of the model: No runs under reference one (161-162 LK)

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

- Joint-residuals
- Runs tests

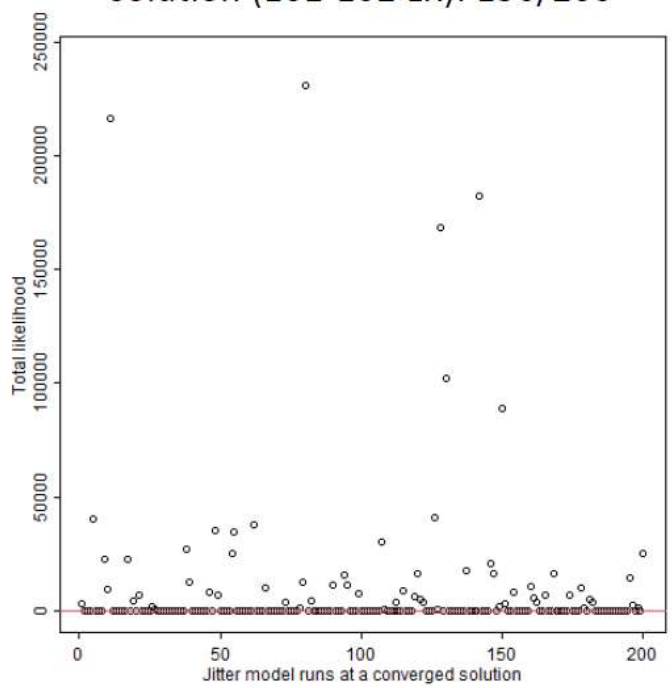
3. Consistency

- Retrospective analysis

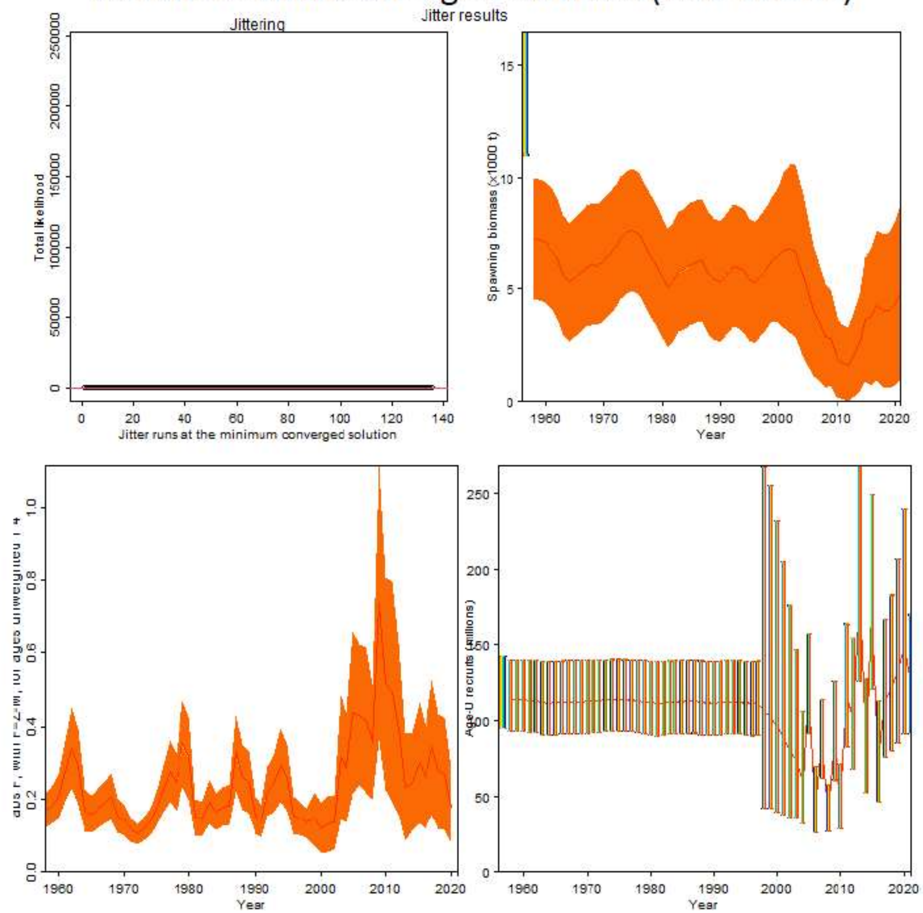
4. Prediction skills

- Hindcasting (CrossValidation)

Num runs at min converged solution (161-162 LK): 136/200



Results at min converged solution (161-162 LK)



Diagnostic - Goodness of the fit (ref run 1)

A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho[a,*,1], Henning Winker[b,1], Dean Courtney[c], Maia Kapur[d], Laurence Kell[e], Massimiliano Cardinale[f], Michael Schirripa[g], Toshihide Kitakado[h], Dawit Yemane[i], Kevin R. Piner[j], Mark N. Maunder[k,l], Ian Taylor[m], Chantel R. Wetzel[m], Kathryn Doering[n], Kelli F. Johnson[m], Richard D. Methot[m]

jabbamodel / ss3diags Public

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

- Joint-residuals
- Runs tests

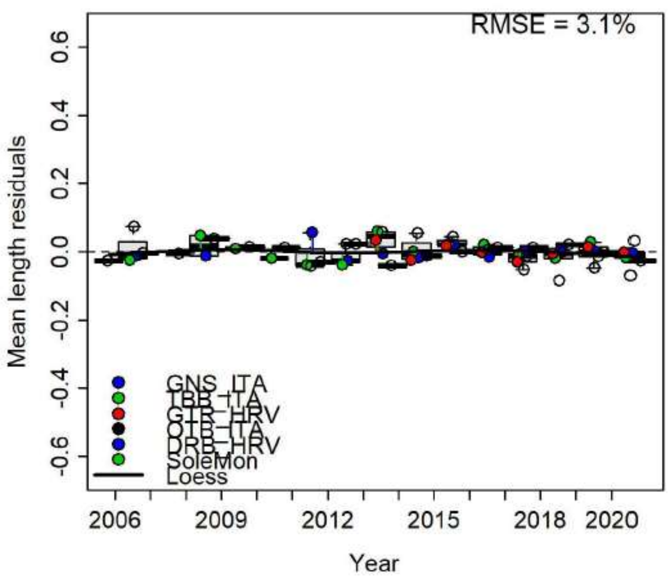
3. Consistency

- Retrospective analysis

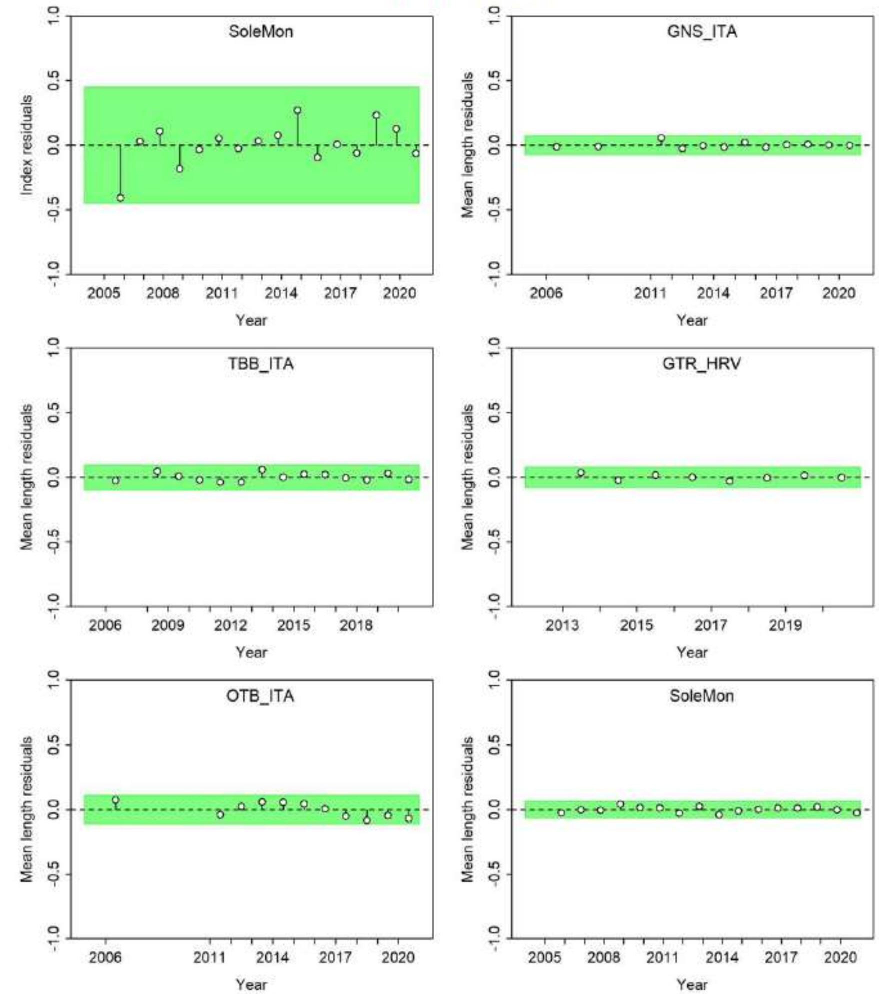
4. Prediction skills

- Hindcasting (CrossValidation)

Joint-residuals



Runs tests



Diagnostic - Consistency (ref run 1)

A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho[a,*,1], Henning Winker[b,1], Dean Courtney[c], Maia Kapur[d], Laurence Kell[e], Massimiliano Cardinale[f], Michael Schirripa[g], Toshihide Kitakado[h], Dawit Yemane[i], Kevin R. Piner[j], Mark N. Maunder[k,l], Ian Taylor[m], Chantel R. Wetzel[m], Kathryn Doering[n], Kelli F. Johnson[m], Richard D. Methot[m]

jabbamodel / ss3diags Public

Mohn´s indices (both ρM and ρF) limit value:

- Long lived species: 0.20 ; -0.15
- Short lived species: 0.30 ; -0.22

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

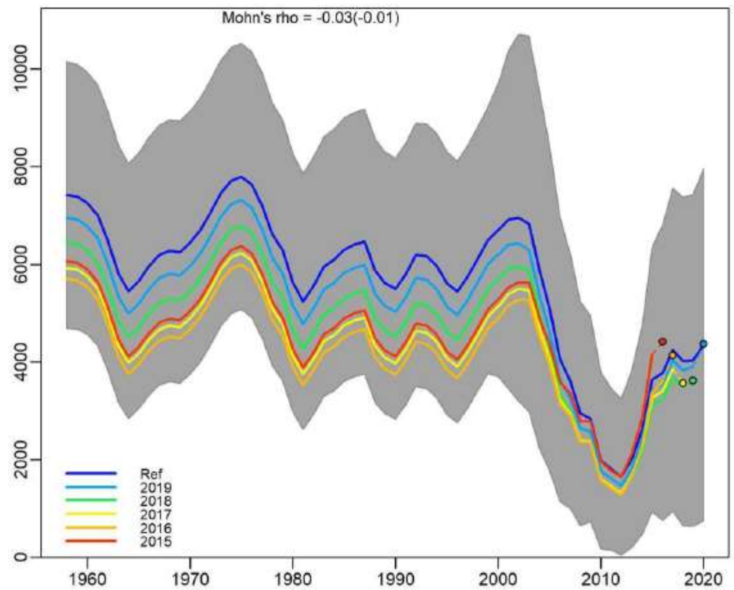
- Joint-residuals
- Runs tests

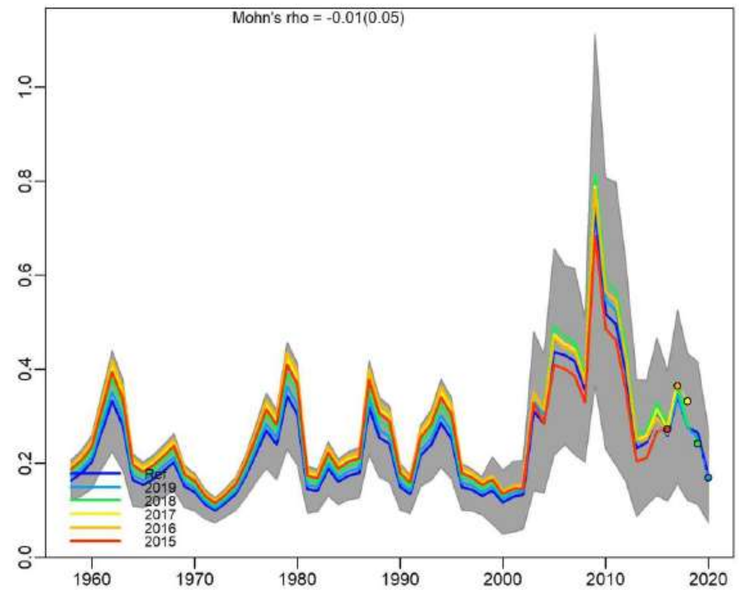
3. Consistency

- Retrospective analysis

4. Prediction skills

- Hindcasting (CrossValidation)





Diagnostic - Prediction skills (ref run 1)

A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho[a,*,1], Henning Winker[b,1], Dean Courtney[c], Maia Kapur[d], Laurence Kell[e], Massimiliano Cardinale[f], Michael Schirripa[g], Toshihide Kitakado[h], Dawit Yemane[i], Kevin R. Piner[j], Mark N. Maunder[k,l], Ian Taylor[m], Chantel R. Wetzel[m], Kathryn Doering[n], Kelli F. Johnson[m], Richard D. Methot[m]

jabbamodel / ss3diags Public

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

- Joint-residuals
- Runs tests

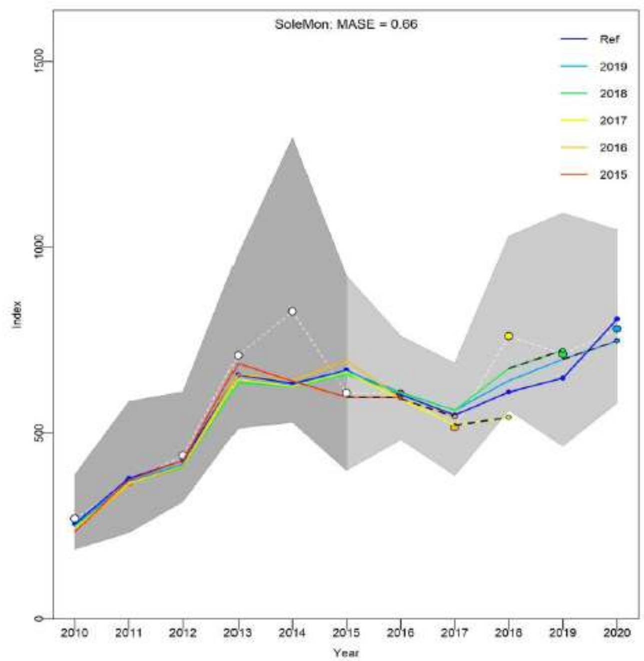
3. Consistency

- Retrospective analysis

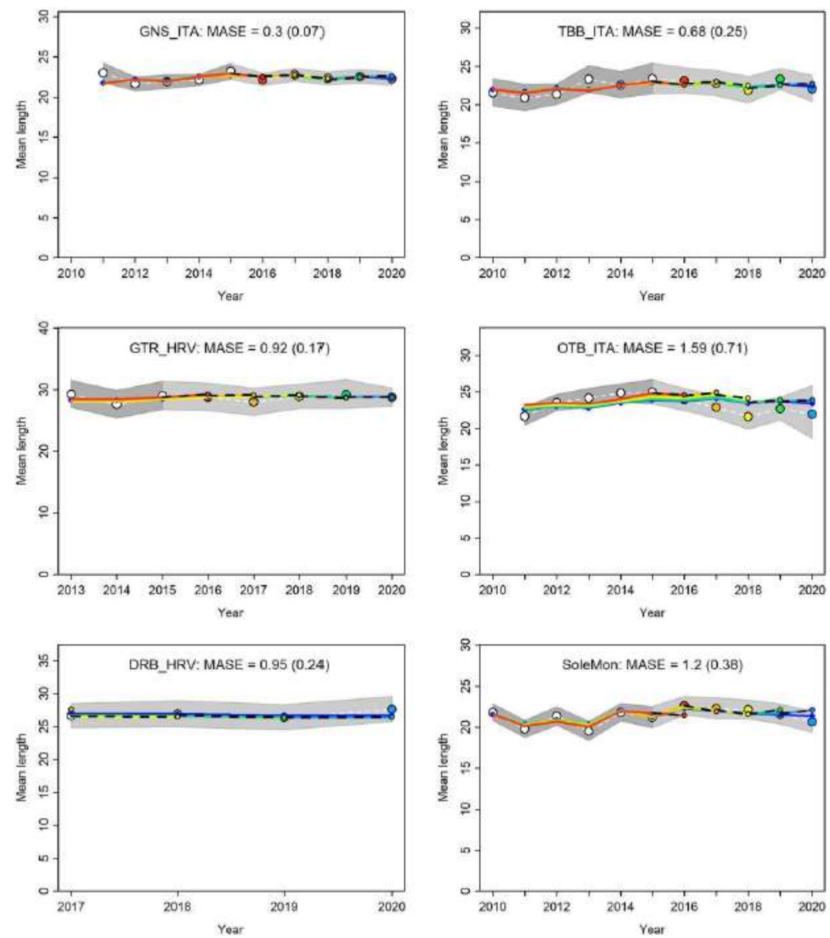
4. Prediction skills

- Hindcasting (CrossValidation)

MASE CPUE Index



MASE Length comp



Model weighting (diagnostic scores)

A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho[a,*,1], Henning Winker[b,1], Dean Courtney[c], Maia Kapur[d], Laurence Kell[e], Massimiliano Cardinale[f], Michael Schirripa[g], Toshihide Kitakado[h], Dawit Yemane[i], Kevin R. Piner[j], Mark N. Maunder[k,l], Ian Taylor[m], Chantel R. Wetzel[m], Kathryn Doering[n], Kelli F. Johnson[m], Richard D. Methot[m]

jabbamodel / ss3diags Public

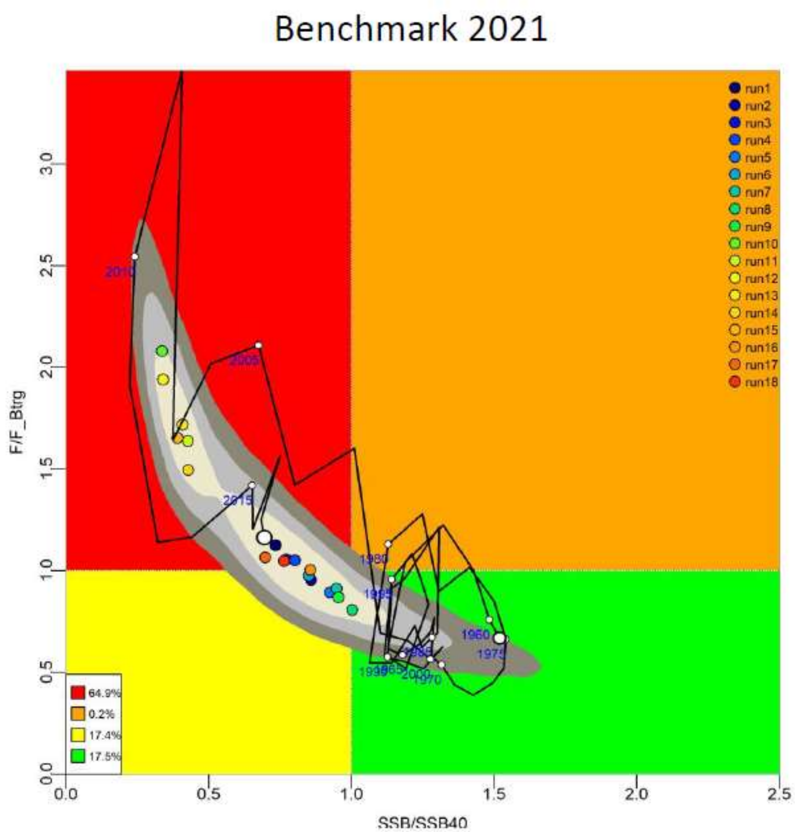
$$W(Diagnostics): \frac{\boldsymbol{W}(\textbf{Diags 1}) + \boldsymbol{W}(\textbf{Diags 2}) + \boldsymbol{W}(\textbf{Diags 3}) \ldots + \boldsymbol{W}(\textbf{Diags N})}{\textbf{Num of } \boldsymbol{W}(\textbf{Diags})}$$

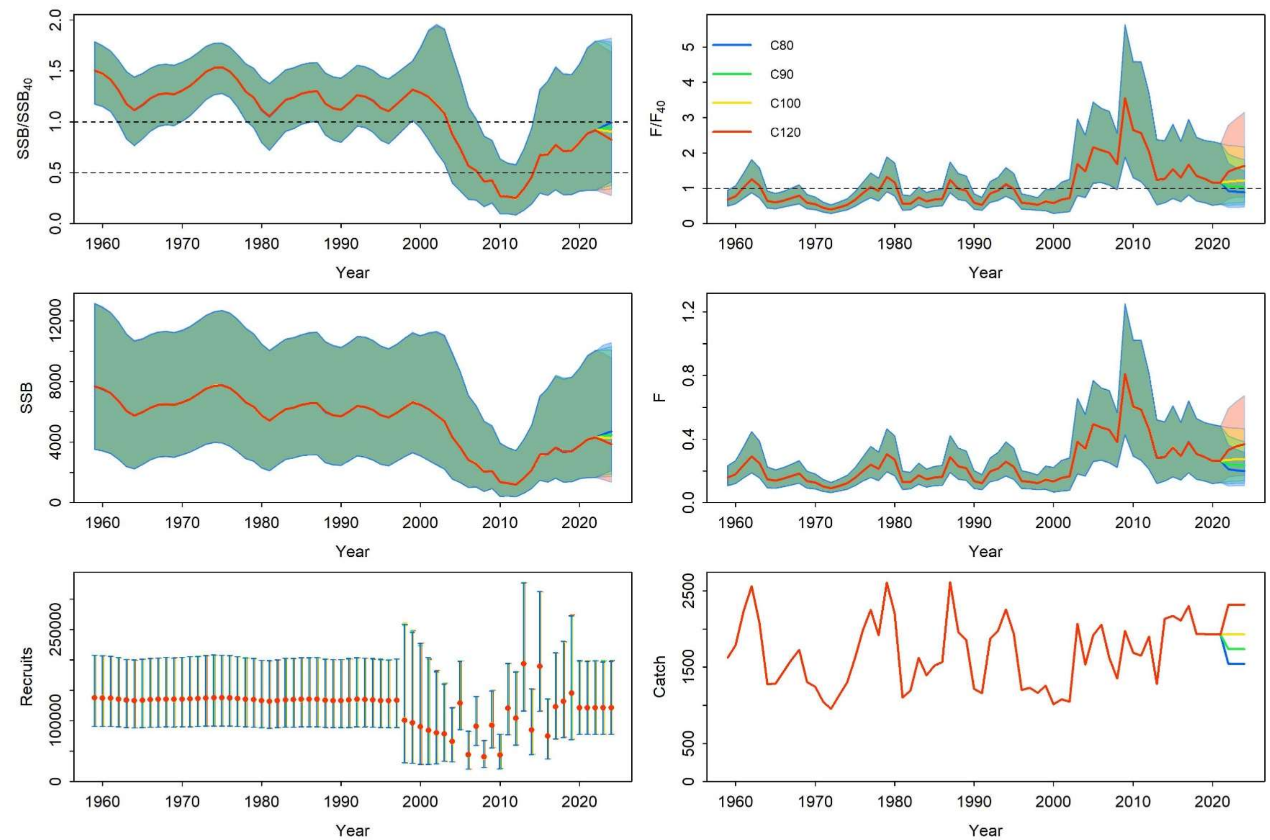
1. Convergence & stability
   - Positive Hessian
   - Jittering
2. Goodness of the fit
   - Joint-residuals
   - Runs tests
3. Consistency
   - Retrospective analysis
4. Prediction skills
   - Hindcasting (CrossValidation)

	Convergence and stability		Goodness of the fit								Consistency				Prediction skills			
	Positive Hessian	Jittering	Run test						Joint-residuals		Retrospective analysis				Hindcasting (MASE)			
Run name			Index	lenGNS_ITA	lenTBB_ITA	lenGTR_HRV	lenOTB_ITA	lenSoleMon	Index	Length	Retro_SSB	Forecast_SSB	Retro_F	Forecast_	Index	SurveyLen	COMfleet	*W(Diagnostics)*
Run1	Passed		Passed	Passed	Passed	Passed	Passed	Passed	15.2	3.1	-0.083	-0.070	0.021	0.035	0.726	0.399	0.320	**1.00**
Run2	Passed		Passed	Passed	Passed	Passed	Passed	Passed	14.7	3.1	-0.058	-0.054	0.026	0.052	0.863	0.363	0.312	**1.00**
Run3	Passed		Passed	Passed	Passed	Passed	Passed	Passed	14.9	3.1	-0.061	-0.053	0.016	0.036	0.766	0.382	0.316	**1.00**
Run4	Passed		Passed	Passed	Passed	Passed	Passed	Passed	15.4	3.1	-0.074	-0.059	0.018	0.029	0.714	0.407	0.319	**1.00**
Run5	Passed		Passed	Passed	Passed	Passed	Passed	Passed	14.7	3.1	-0.040	-0.036	0.014	0.040	0.842	0.370	0.312	**1.00**
Run6	Passed		Passed	Passed	Passed	Passed	Passed	Passed	14.9	3.1	-0.036	-0.030	0.008	0.026	0.743	0.334	0.316	**1.00**
Run7	Passed		Passed	Passed	Passed	Passed	Passed	Passed	15	3.1	-0.078	-0.064	0.034	0.047	0.744	0.410	0.317	**1.00**
Run8	Passed		Passed	Passed	Passed	Passed	Passed	Passed	14.4	3.1	-0.037	-0.033	0.017	0.042	0.825	0.377	0.312	**1.00**
Run9	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	14.7	3.1	-0.054	-0.044	0.021	0.040	0.750	0.396	0.315	**1.00**
Run10	Passed		Passed	Passed	Passed	Passed	Passed	Passed	21.2	3.3	0.126	0.157	-0.106	-0.072	0.967	0.455	0.375	**1.00**
Run11	Passed		Passed	Passed	Passed	Passed	Passed	Passed	20.1	3.6	0.013	0.003	-0.009	0.041	1.362	0.450	0.351	**0.93**
Run12	Passed		Passed	Passed	Passed	Passed	Passed	Passed	21.1	3.4	0.083	0.092	-0.067	-0.037	1.166	0.388	0.367	**0.93**
Run13	Passed		Passed	Passed	Passed	Passed	Passed	Passed	20.2	3.2	0.123	0.162	-0.113	-0.087	0.796	0.472	0.362	**1.00**
Run14	Passed		Passed	Passed	Passed	Passed	Passed	Passed	19.4	3.4	0.042	0.043	-0.040	-0.001	1.098	0.464	0.344	**0.93**
Run15	Passed		Passed	Passed	Passed	Passed	Passed	Passed	20.1	3.2	0.086	0.102	-0.078	-0.044	0.957	0.463	0.354	**1.00**
Run16	Passed		Passed	Passed	Passed	Passed	Passed	Passed	16.7	3.1	0.070	0.081	-0.067	-0.024	0.777	0.423	0.346	**1.00**
Run17	Passed		Passed	Passed	Passed	Passed	Passed	Passed	16.6	3.1	0.049	0.051	-0.045	0.001	0.887	0.421	0.340	**1.00**
Run18	Passed		Passed	Passed	Passed	Passed	Passed	Passed	16.7	3.1	0.062	0.070	-0.058	-0.014	0.810	0.423	0.344	**1.00**

13.33 x 7.50 in

Ensemble model Results

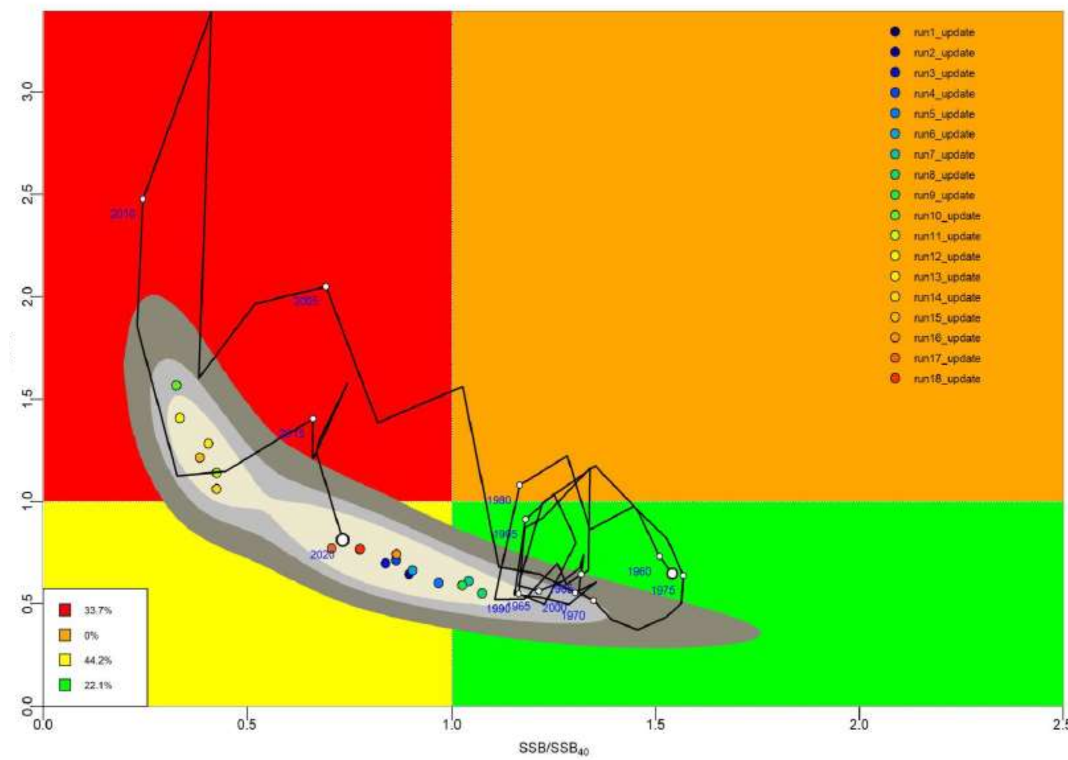




Ensemble model Results

Name	Selectivity	M	h	Weighting
run1	DN	M1	0.9	**1.00**
run2	DN	M1	0.7	**1.00**
run3	DN	M1	0.8	**1.00**
run4	DN	M2	0.9	**1.00**
run5	DN	M2	0.7	**1.00**
run6	DN	M2	0.8	**1.00**
run7	DN	M3	0.9	**1.00**
run8	DN	M3	0.7	**1.00**
run9	DN	M3	0.8	**1.00**
run10	CS	M1	0.9	**1.00**
run11	CS	M1	0.7	**0.93**
run12	CS	M1	0.8	**0.93**
run13	CS	M2	0.9	**1.00**
run14	CS	M2	0.7	**0.93**
run15	CS	M2	0.8	**1.00**
run16	CS	M3	0.9	**1.00**
run17	CS	M3	0.7	**1.00**
run18	CS	M3	0.8	**1.00**

Update 2022



Survey Index increasing &
-25% landings compare to 2019 !!!

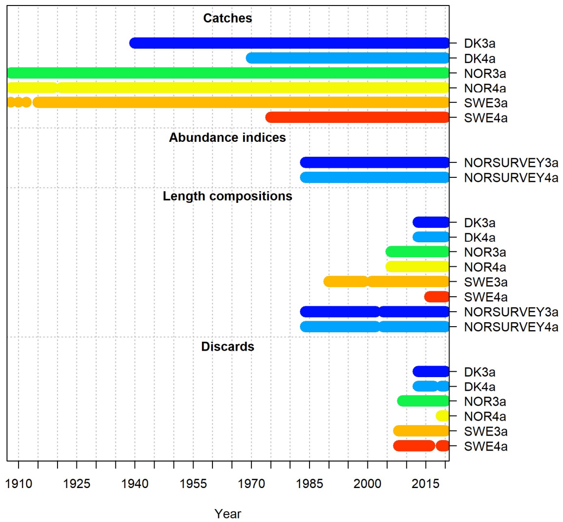
Cookbook in action (2022)

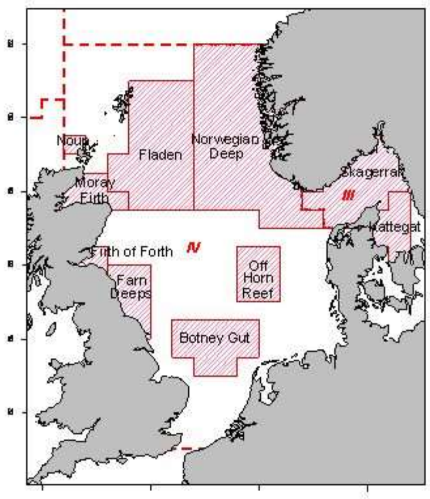
ICES

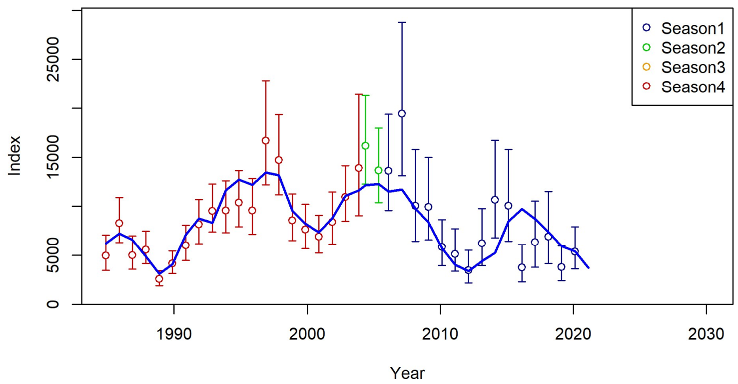
Stock: Northern shrimp (*Pandalus borealis*)

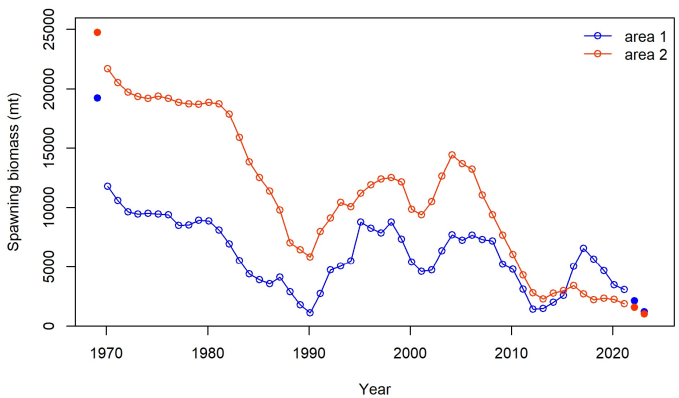
Model: Stock Synthesis – Length-based 2 Area Model

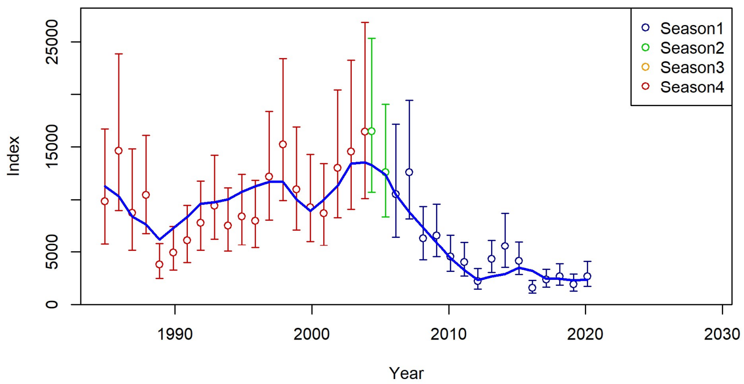
Max Cardinale, Francesco Masnadi, Alessandro Orio, Mikaela Bergenius, Katja Nören, Christopher Griffiths











ICES Benchmark of *Northern shrimp in 3a and 4a east*

Set_1: Length based model based on VBGP and all fleets selectivity logistic discard retention fixed

Evolved model: Two sex, two areas, hermaphroditic model with discards separated from landings

Ancestor model: Single sec, single area without separated discards

Area specific sex ratio

Discarded: Worse fit of the LFD

Long landing time series (1908-2020)

Kept: same results, same fit

Hermaphroditism fixed

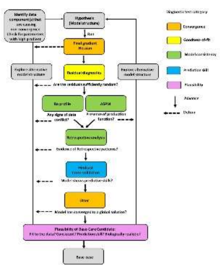
Discarded: Worse fit of the LFD

Short landing time series (1970-2020)

Discards retention time varying: improved fit, improved diagnostic

Survey index 2016 excluded

Iterative



Automate

Run name	Convergence and stability: Positive Hessian	Convergence and stability: Jittering	Goodness of the fit: Run test: CPUE1	CPUE2	Len1	Len2	Len3	Len4	Goodness of the fit: Run test: Len5	Len6	Len7	Len8	Index	Length	Consistency: Retrospective analysis: Retro_SSB	Forecast_SSB	Retro_F	Forecast_F
Run1	Passed	Passed	Passed	Passed	Failed	Passed	Failed	Passed	Failed	Passed	Passed	Failed	Passed	Passed	Passed	Passed	Passed	Passed
Run2	Passed		Passed	Passed	Failed	Passed	Failed	Passed	Failed	Passed	Passed	Failed	Passed	Passed	Passed	Passed	Passed	Failed
Run3	Passed		Passed	Passed	Failed	Passed	Failed	Passed	Failed	Failed	Passed	Failed	Passed	Passed	Failed	Failed	Passed	Passed
Run4	Passed		Passed	Passed	Failed	Passed	Failed	Passed	Failed	Failed	Passed	Failed	Passed	Passed	Passed	Passed	Passed	Failed

Prediction skills: Hindcasting (MASE): Survey3a	Surve4a	Joint	Len3a	Len4a	lenA1S1	lenA1S2	lenA1S3	lenA1S4	lenA2S1	lenA2S2	lenA2S3	lenA2S4	Weight
Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	0.86
Passed	Failed	Failed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Failed	Failed	0.69
Passed	Passed	Passed	Failed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	0.72
Passed	Failed	Passed	Failed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	0.72

Jittering and MCMC

M1 (base, h=0.7)

Predation M (h=0.91)

M2 (low, h=0.89)

M3 (high, h=0.52)

Ss3diags
Pass/Fail

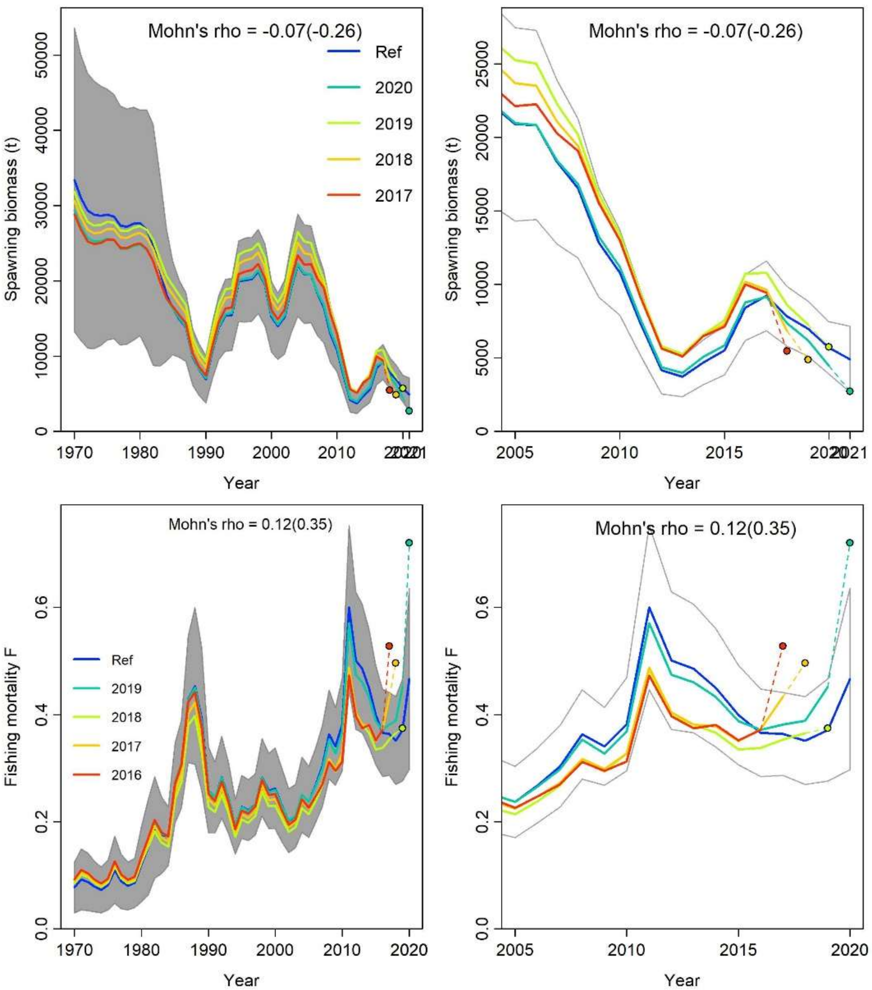
Weighting based on ss3diags

ENSEMBLE MODEL → **Forecast**

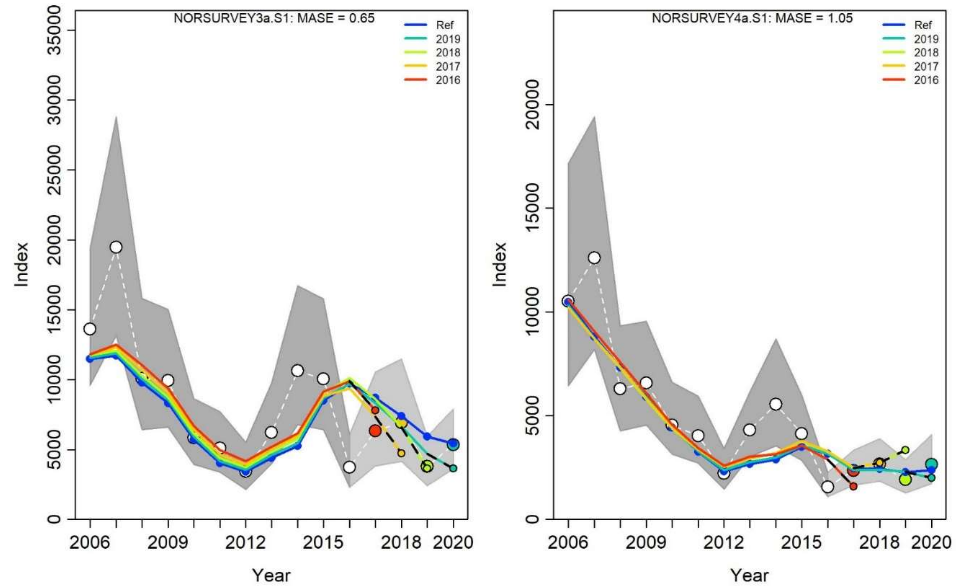
European Commission

Stock: Northern shrimp (*Pandalus borealis*)

Retrospective Analysis



Hindcast Cross-Validation



Joint MASE < 1

The need for simulation testing

The need for more simulation testing

1. Identify Causes of Misspecifications
    - Observation process or system dynamics?

2. Specificity/Sensitivity of Diagnostic tests?
    - False Negative vs False Positive

3. Implications on status/advice?
    - Not all misspecifications may be equally consequential
    - Is there asymmetric risk among diagnostic?

Original Article

Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models

Felipe Hurtado-Ferro[1*], Cody S. Szuwalski[1,2], Juan L. Valero[3], Sean C. Anderson[4], Curry J. Cunningham[1], Kelli F. Johnson[1], Roberto Licandeo[5], Carey R. McGilliard[6†], Cole C. Monnahan[7], Melissa L. Muradian[7], Kotaro Ono[1], Katyana A. Vert-Pre[8], Athol R. Whitten[1], and André E. Punt[1]

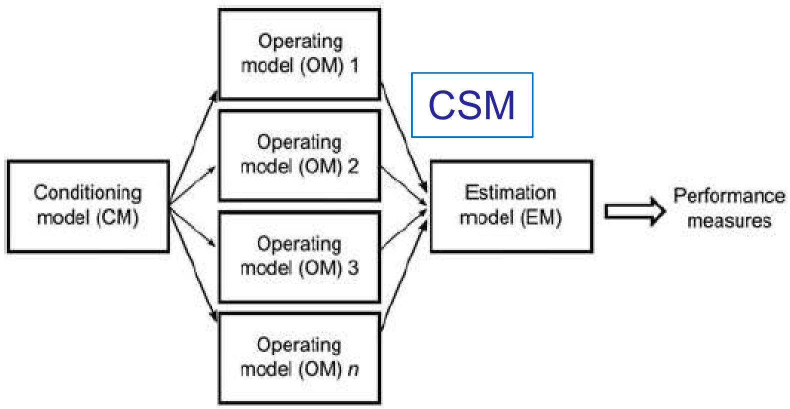


Figure 2. General design of the simulation study.

Time-varying

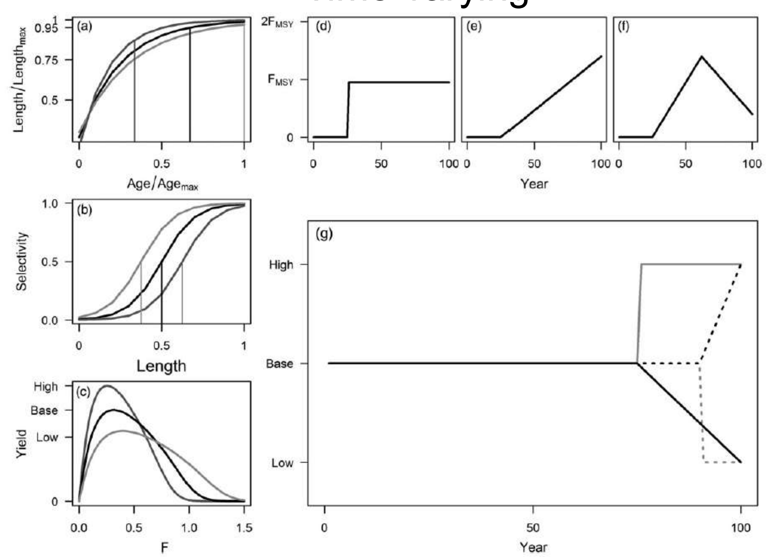


Figure 4. Experimental design showing the time-varying processes (a – c), fishing patterns (d – f), and time-varying patterns for the processes (g). For the patterns of time variance, only some cases are shown to reduce clutter (solid and dotted lines show old and recent timing, respectively; black and grey lines show gradual and sudden patterns, respectively). See text for further explanation.

Causes for retrospective patterns with focus on unaccounted time-varying growth, M and selectivity

Rule-of-thumb Mohn's ρ ranges:

Low-medium productivity: **-0.15 – 0.2**

High productivity: **-0.22 - 0.30**

European Commission

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Full length article

Can diagnostic tests help identify model misspecification in integrated stock assessments?

Felipe Carvalho[a,b,*], André E. Punt[c], Yi-Jay Chang[d], Mark N. Maunder[e,f], Kevin R. Piner[g]

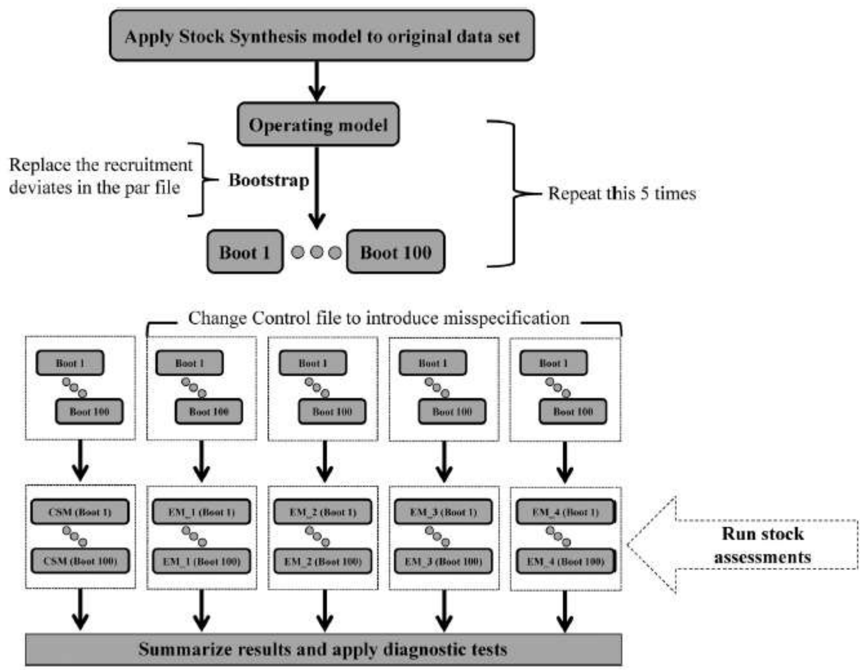
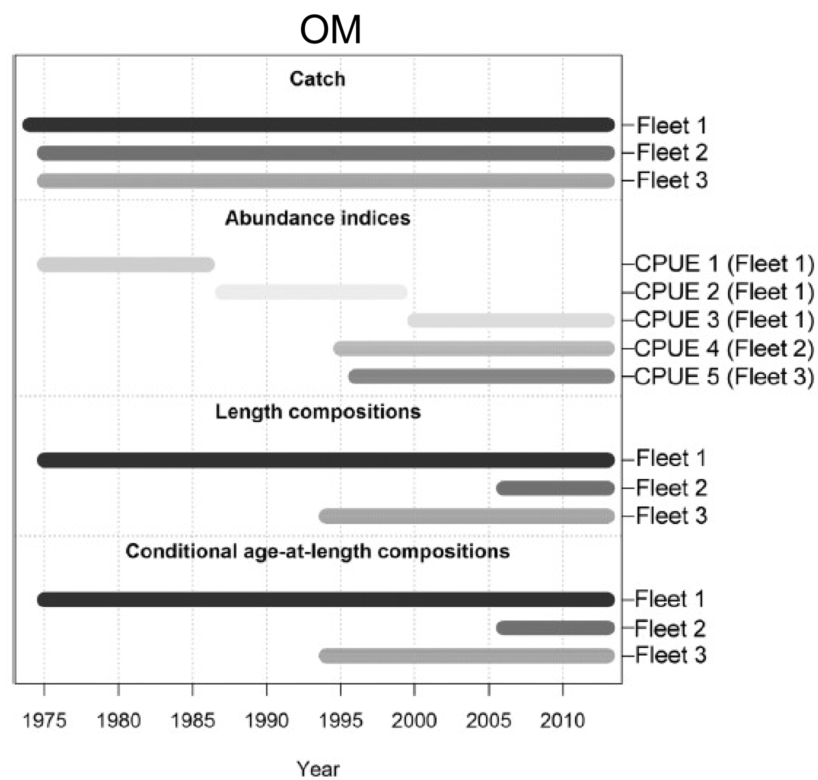


Fig. 3. General design of the simulation study.



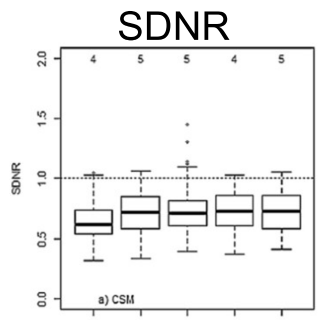
European Commission


Contents lists available at ScienceDirect
Fisheries Research
journal homepage: www.elsevier.com/locate/fishres

Full length article

Can diagnostic tests help identify model misspecification in integrated stock assessments?

Felipe Carvalho[a,b,*], André E. Punt[c], Yi-Jay Chang[d], Mark N. Maunder[e,f], Kevin R. Piner[g]



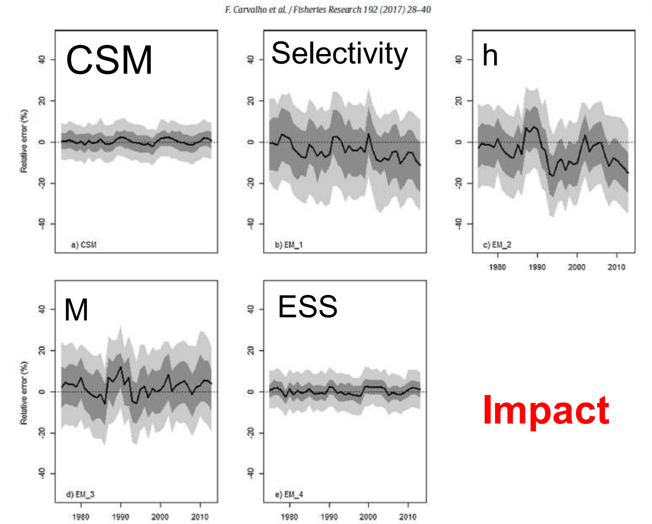


Fig. 6. Relative error distributions (median relative errors, and 50% and 90% intervals) for the time-trajectory of spawning stock biomass of WCNP striped marlin.

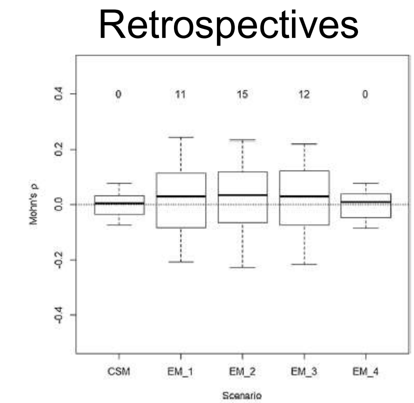


Fig. 8. Box-plots of Mohn's ρ values for stock spawning biomass for each of the five scenarios for WCNP striped marlin. Numbers in the top are the proportion of models where Mohn's ρ values fell outside the range −0.15 to 0.2.

Diagnostic tests:

(1) Runs Test + SDNR

(2) Retrospective analysis

(3) R0 likelihood profile **0%**

(4) ASPM

(5) Catch-Curve Analysis 91% False Negatives

Misspecifications:

(1) Selectivity: all logistic vs 2 of 3 logistic (OM)

(2) Low steepness h vs high h (OM)

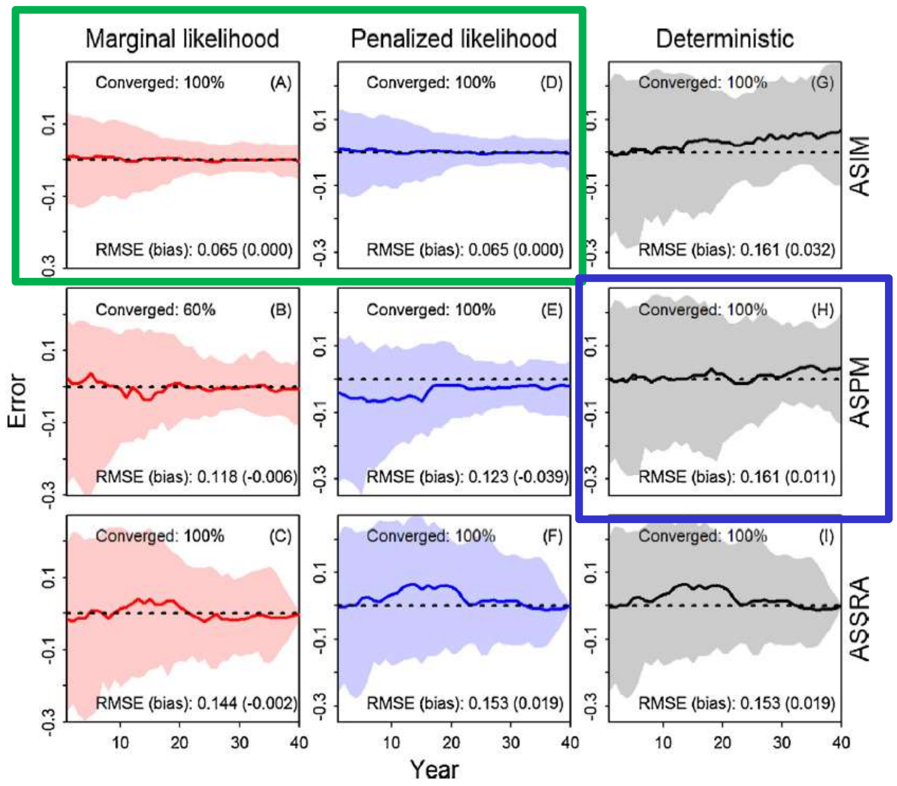
(3) Constant M vs M-at-age (OM)

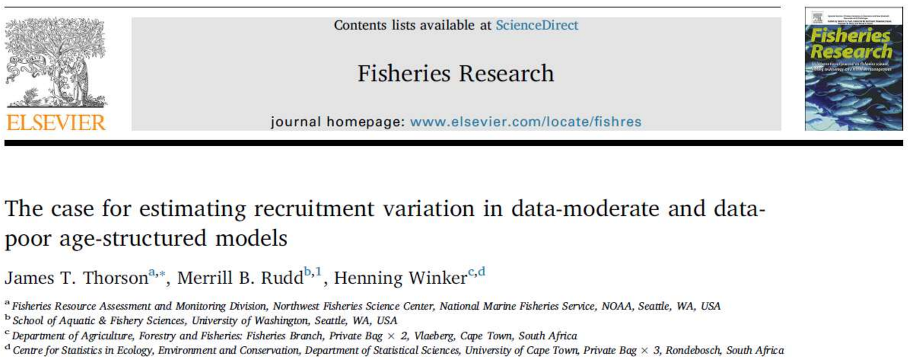
(4) Over-weighting of size-/age composition

European Commission

ASPMs

- Promising candidate to identify misspecifications in the system dynamics (Carvalho et al. 2016)
- Apart from overprecision, ASPMs appear to perform well in simulations to capture the dynamics




Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

The case for estimating recruitment variation in data-moderate and data-poor age-structured models

James T. Thorson[a,*], Merrill B. Rudd[b,1], Henning Winker[c,d]

[a] Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA
[b] School of Aquatic & Fishery Sciences, University of Washington, Seattle, WA, USA
[c] Department of Agriculture, Forestry and Fisheries: Fisheries Branch, Private Bag × 2, Vlaeberg, Cape Town, South Africa
[d] Centre for Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town, Private Bag × 3, Rondebosch, South Africa

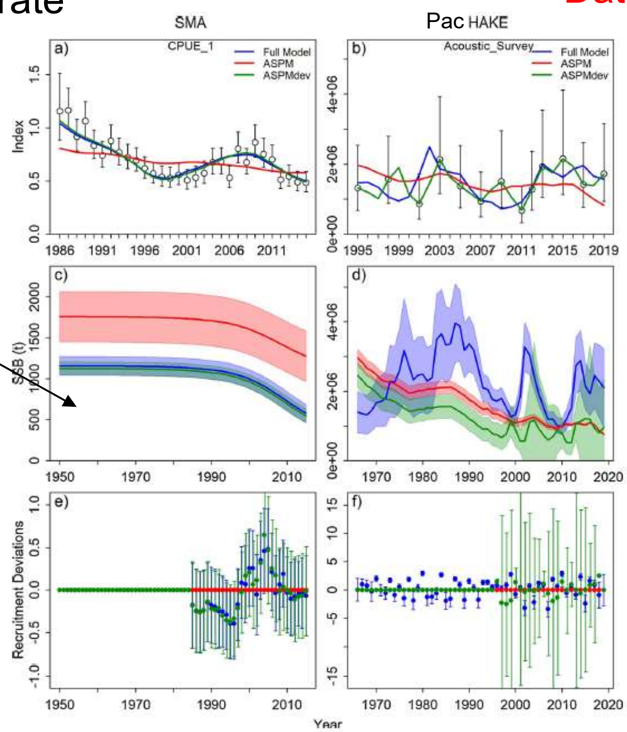
Emerging ASPM questions:

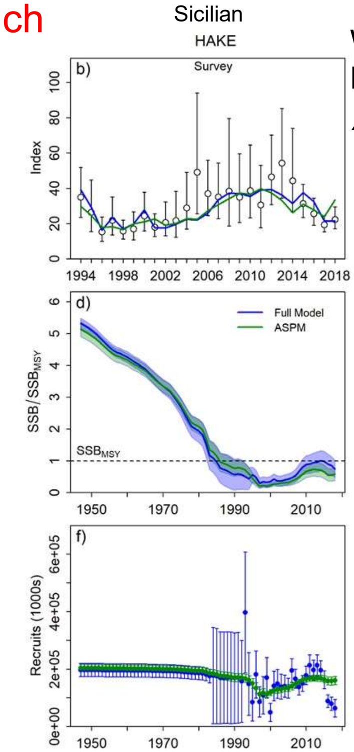
- When do correctly specified ASPMs fail in simulations? - e.g. combinations of σ_R, ρ and generation time
- How to treat time-varying empirical input data of weight-at-age, maturity-at-age and M-at-age?
- Are there any data-rich age-based examples where the ASPM was consistent with the full model?

Data-moderate

Data-Rich

Pass SB_{end}/SSB_{init}?

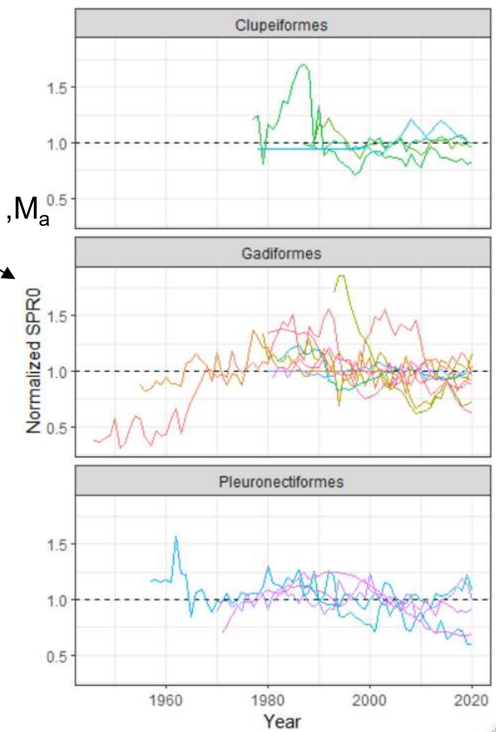




Works well in a few Data-moderate cases

ICES stocks with Empirical W_a, Mat_a ,M_a

Time-varying response of SPR0



Focus Question: Hind casting

Is there a management strategy that relates closely to what we observer (and can then test predictions for)?

Perhaps we should go one step back first and ask:

Are latent management quantities more reliable from models that can be cross-validated using observations that are unknown to the model (i.e. have prediction skill)?

We can test this with simulations!

ICES Journal of Marine Science

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

ICES Journal of Marine Science (2021), 78(6), 2244–2255. https://doi.org/10.1093/icesjms/fsab104

Validation of stock assessment methods: is it me or my model talking?

Laurence T. Kell[1,*], Rishi Sharma[2], Toshihide Kitakado[3], Henning Winker[4], Iago Mosqueira[5], Massimiliano Cardinale[6], and Dan Fu[7]

Prediction Skill with Extension: The Vantage Point Approach

935

NRC Research Press

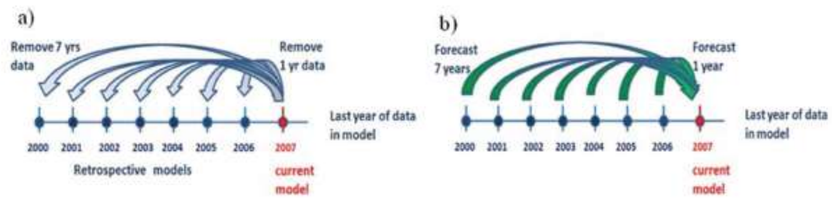
ARTICLE

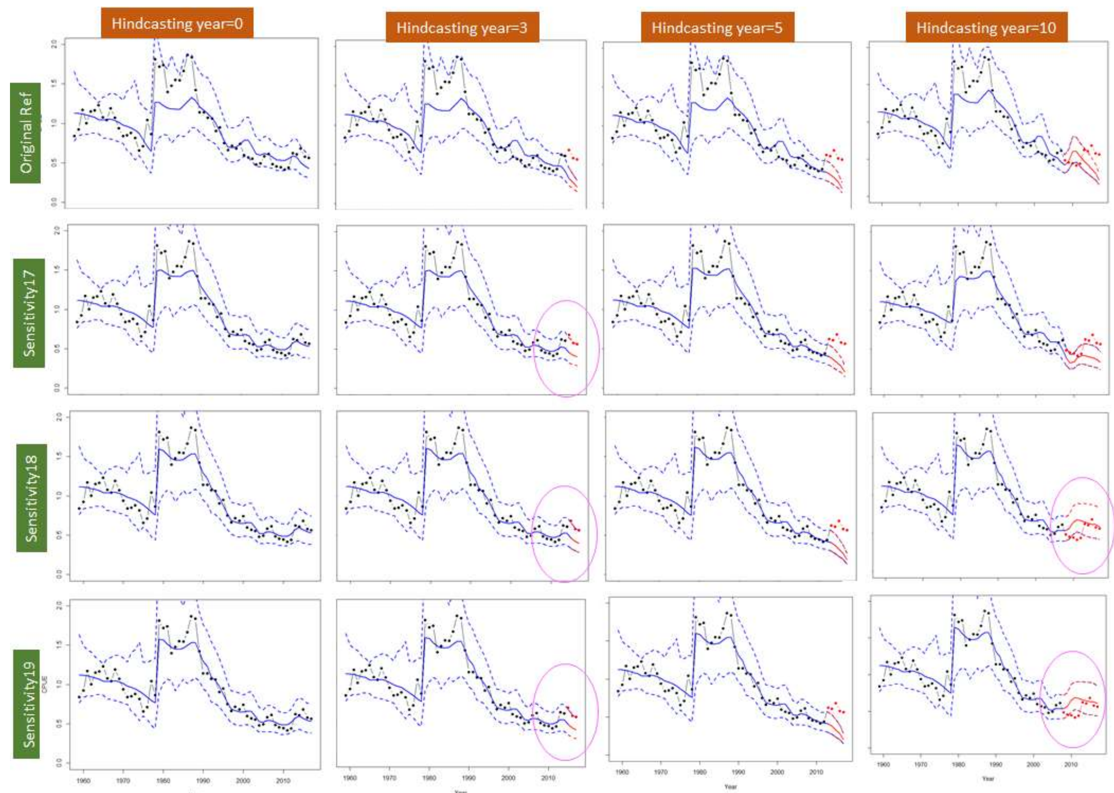
Retrospective forecasting — evaluating performance of stock projections for New England groundfish stocks

Elizabeth N. Brooks and Christopher M. Legault

Toshihide Kitakado
ICCAT Bigeye 2018

Fig. 1. Illustration of "retrospective forecasting". (a) Retrospective models are created by removing 1, 2, …, 7 years of data from the full time series (current model). (b) Forecasts are then made from each retrospective model to the end of year 2007.





- Increased prediction residuals (1+2+3,…)
- More contrast?
- How far can we forecast?

Another look at measures of forecast accuracy

Rob J. Hyndman[a,*], Anne B. Koehler[b,1]

[a] Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia
[b] Department of Decision Sciences and Management Information Systems, Miami University, Oxford, Ohio 45056, USA

SSmase(re)

Index	Season	MASE	MAE.PR	MAE.base	MASE.adj	n.eval
Survey	1	1.065	0.302	0.284	1.065	4

Prediction Skill with Extension: The Vantage Point Approach

935

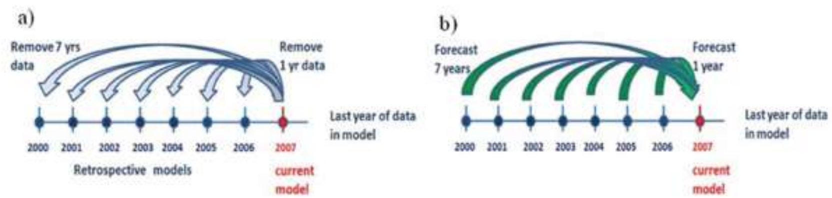
NRC Research Press

ARTICLE

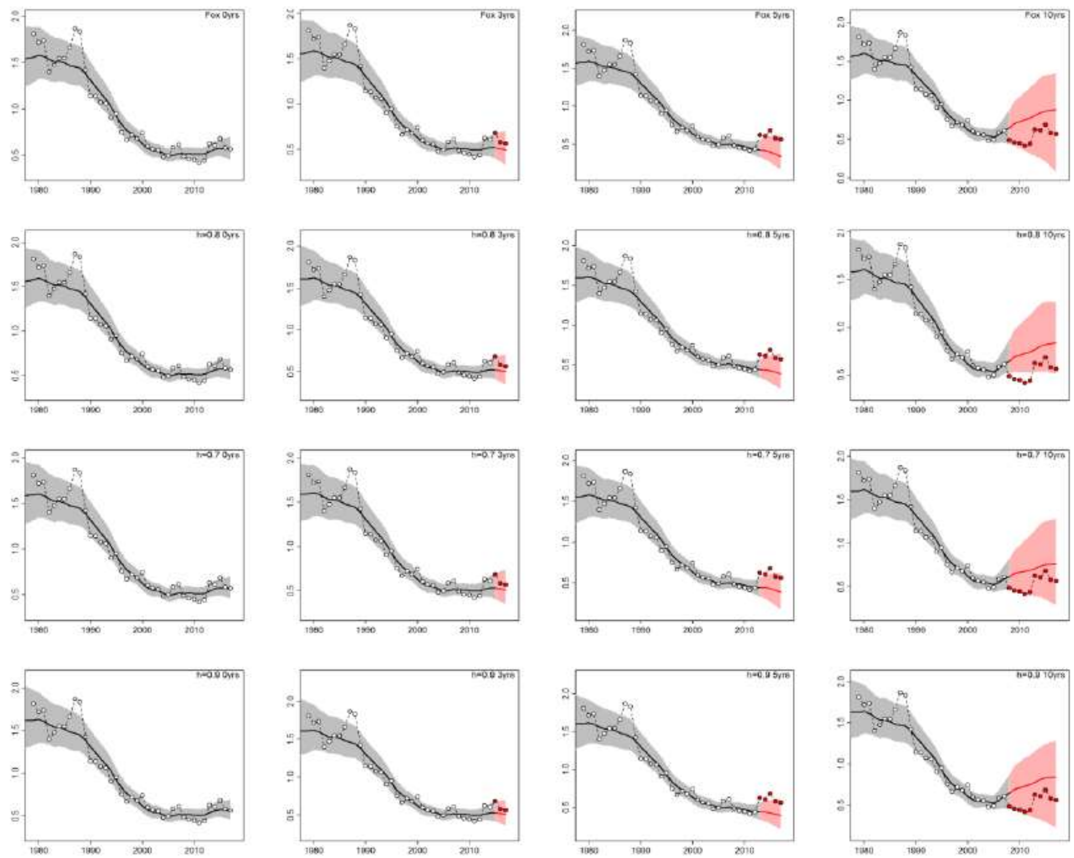
Retrospective forecasting — evaluating performance of stock projections for New England groundfish stocks

Elizabeth N. Brooks and Christopher M. Legault

Fig. 1. Illustration of "retrospective forecasting". (*a*) Retrospective models are created by removing 1, 2, …, 7 years of data from the full time series (current model). (*b*) Forecasts are then made from each retrospective model to the end of year 2007.



Toshihide Kitakado
ICCAT Bigeye 2018



- Increased prediction residuals (1+2+3,…)
- More contrast?
- How far can we forecast?

Another look at measures of forecast accuracy

Rob J. Hyndman [a,*], Anne B. Koehler [b,1]

[a] Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia
[b] Department of Decision Sciences and Management Information Systems, Miami University, Oxford, Ohio 45056, USA

SSmase(re)

Index	Season	MASE	MAE.PR	MAE.base	MASE.adj	n.eval
Survey	1	1.065	0.302	0.284	1.065	4

The options for Simulation testing are infinite

OM generation?

1. **Condition + Simplify**
2. **Off the shelf OMs**
3. **Coverage + variation in F-trajectory**
4. **N Stock examples**
5. **Time-varying processes (e.g. Selectivity)**

Misspecifications EM?

1. **Natural Mortality M**
2. **Steepness h**
3. **Somatic Growth**
4. **Maturity**
5. **Selectivity**
6. **Catchability**
7. **Weighting: CV, ESS**
8. **Catch history (underreporting)**
9. Recruitment function
10. Spatial

Influences?

1. **Life history (productivity, generation time)**
2. **Process error (sigmaR, rho)**
3. **Time series length (observation horizon)**
4. **Sampling (e.g. ESS, CV, AR1)**
5. **Stock depletion (left or right of Bmsy?)**
6. **Contrast in exploitation history**
7. **Data (survey vs CPUE; age vs length)**

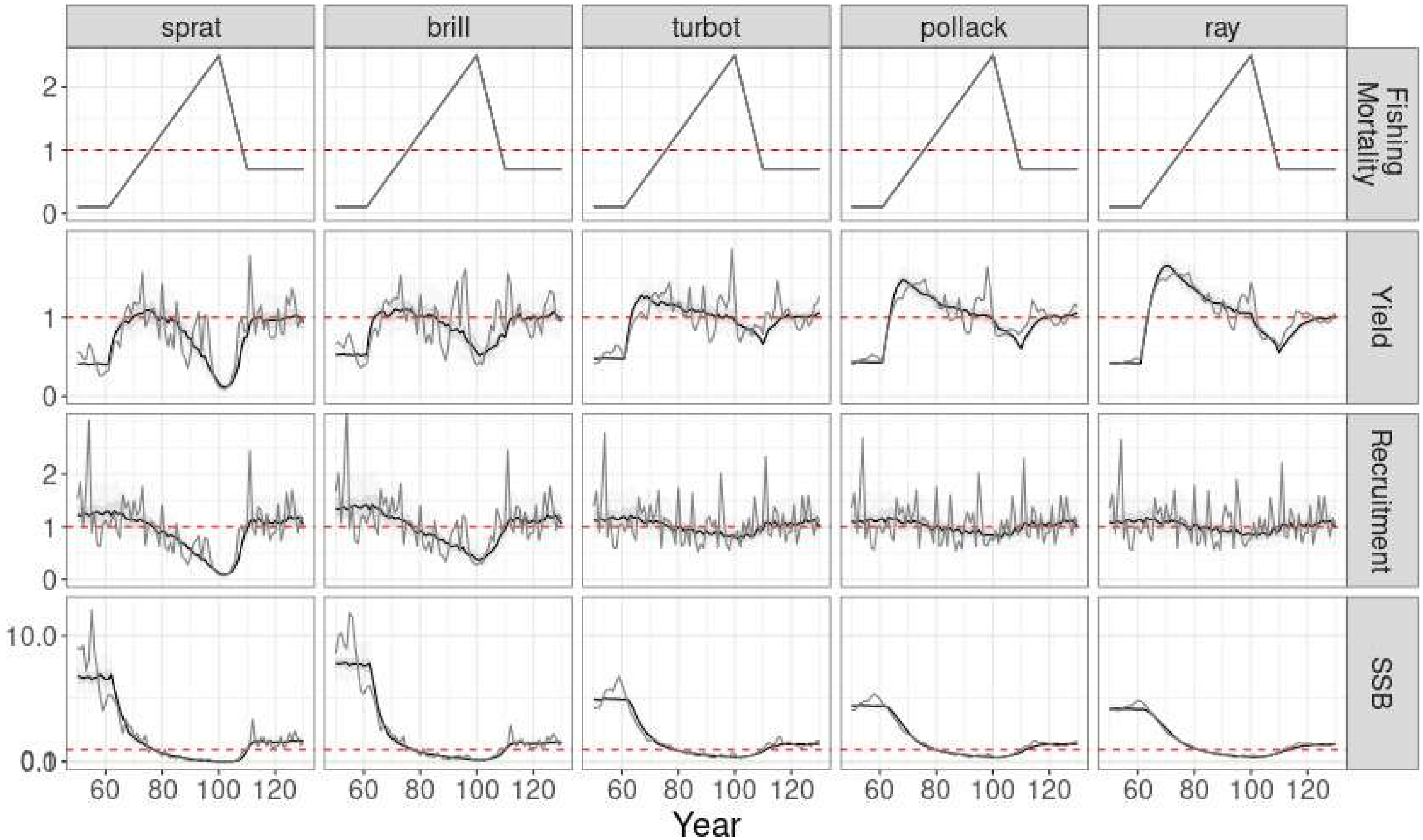
Model complexity?

1. **N Indices**
2. **N Size/Age comps setup**
3. **Historical vs Recent time series**
4. **N Fleets**
5. Sex-structured?
6. Seasonal?
7. Spatial?

European Commission

Proposal: Start with simplified “off the shelf” OMs for the system dynamics (1 Fleet)

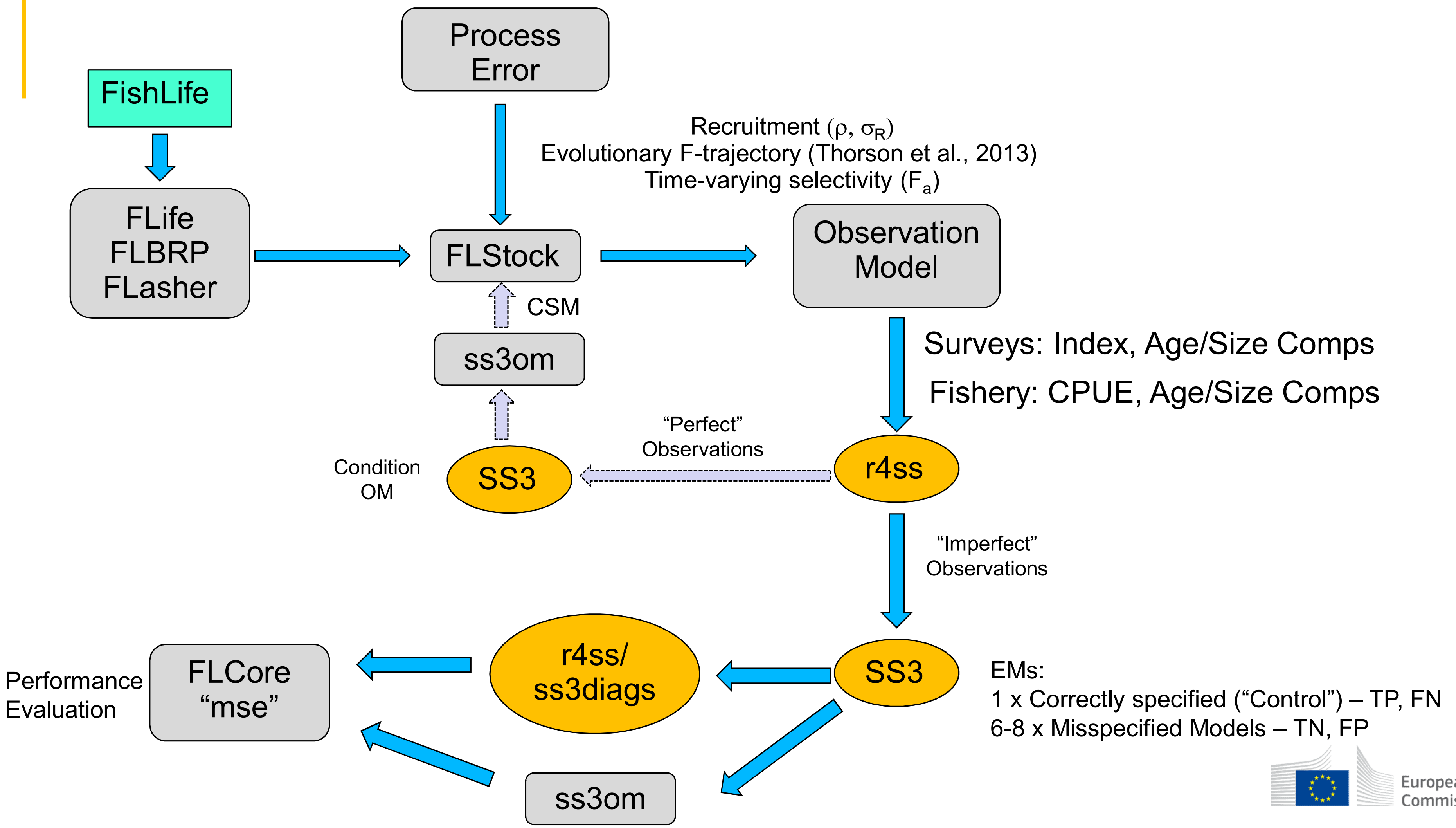
- Cover more life histories, exploitation patterns and process error dynamics, data-rich vs data moderate
- Minimize run time to 1-3 min (1 OM = ~ 8 EMs x 5 hindcasts x 100 iterations = 1 hour on 75 cores)
- Easier interpretation of main effects



Evolutionary F-trajectory (Thorson et al., 2013)

σ_R, ρ

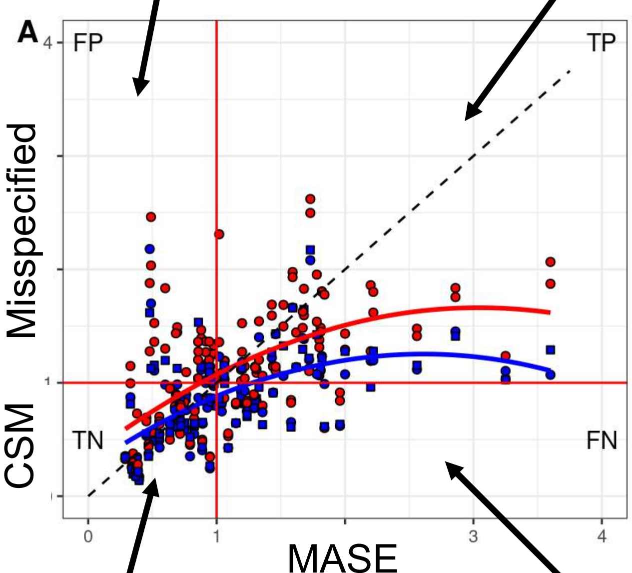
Strawman: Simulation design



Sensitivity: the ability of a test to correctly identify a misspecification
Specificity: the ability of a test to identify a correctly specified model (Model)

Misspecification not detected

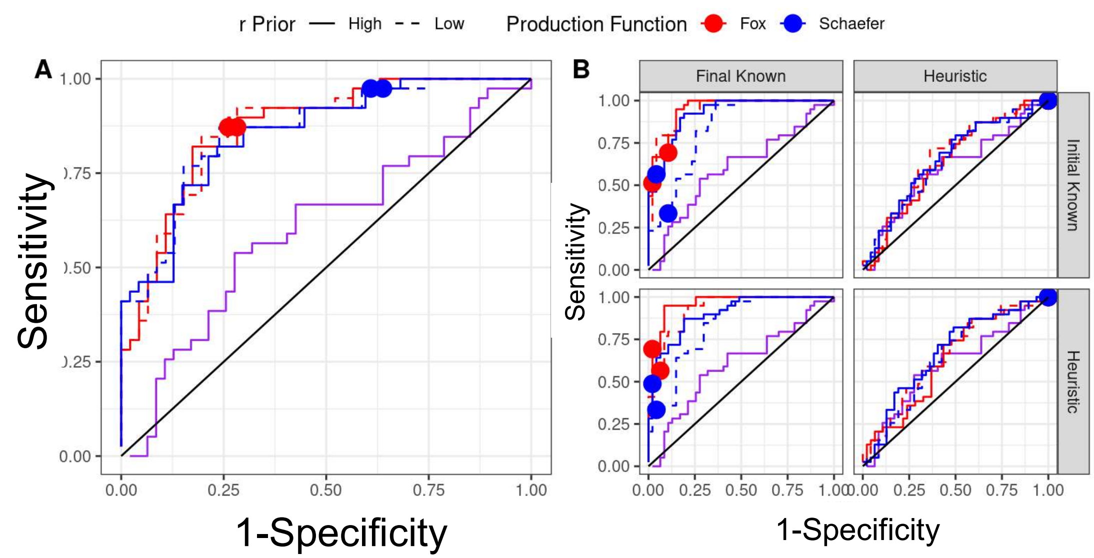
Misspecification correctly diagnosed



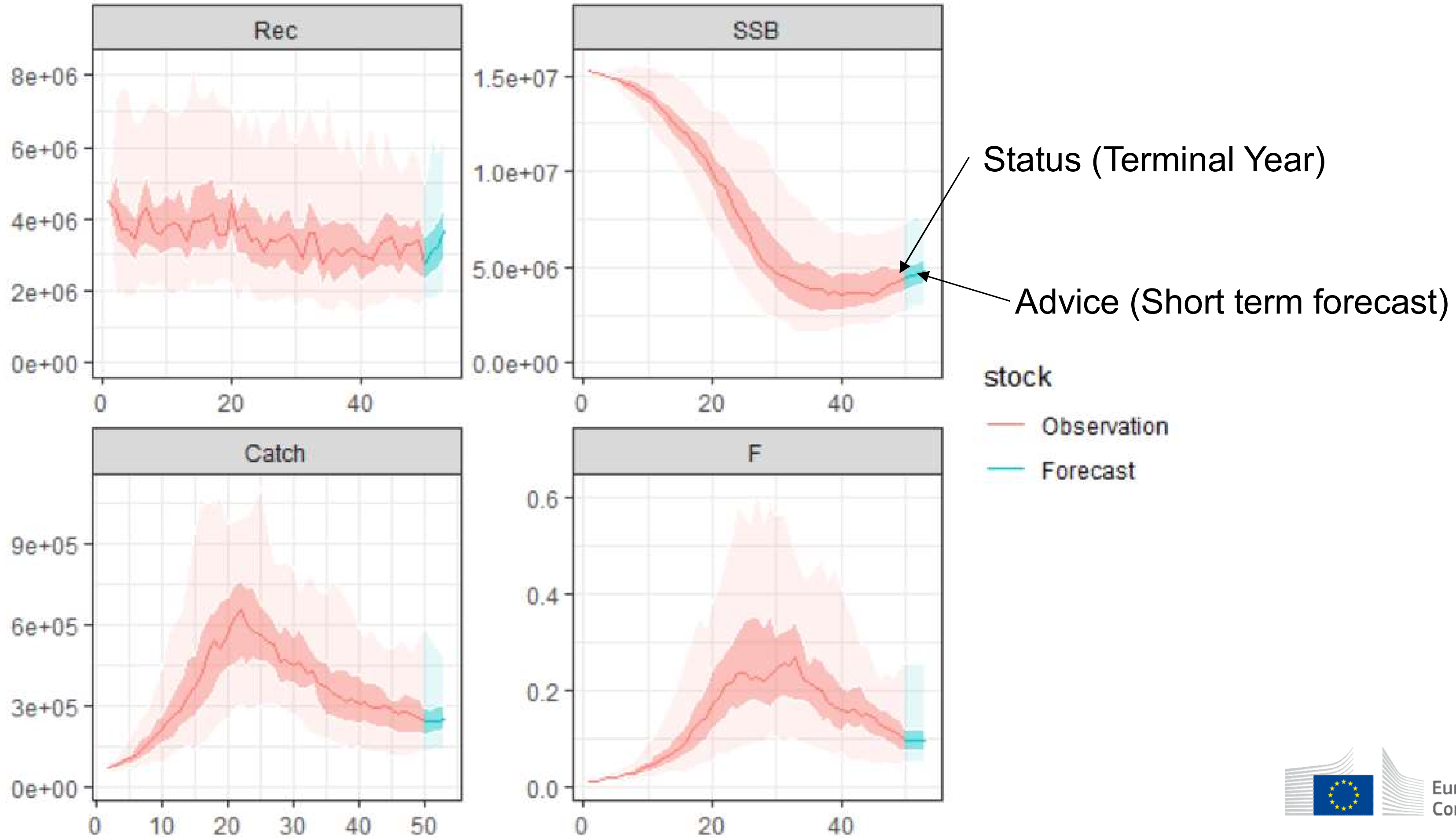
CSM correctly accepted

CSM falsely rejected

Receiver Operating Characteristics

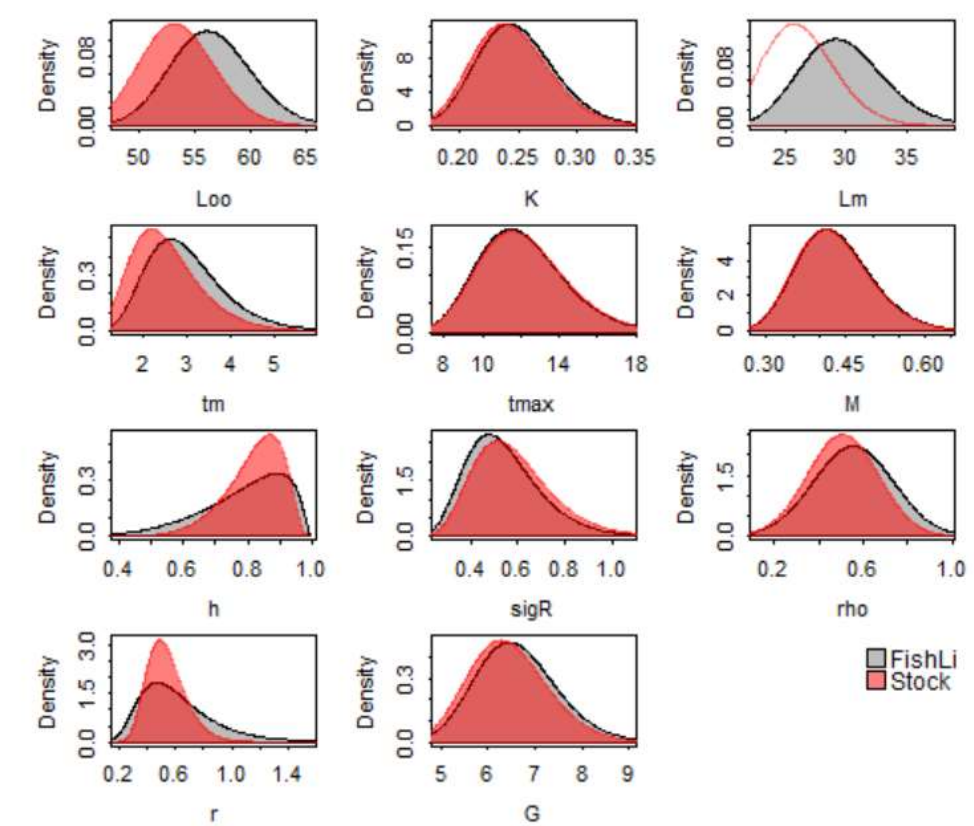


Evaluating impacts on status/advice?



A simple simulation trial

Simulation Run with LIME: Sparus aurata



Received: 17 October 2018 | Revised: 21 October 2019 | Accepted: 31 October 2019
DOI: 10.1111/faf.12427

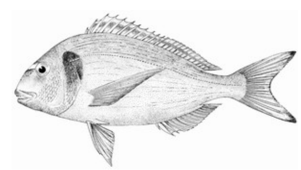
ORIGINAL ARTICLE

FISH and FISHERIES WILEY

Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model

James T. Thorson[1,2]

- SSR BevHold with h = 0.87
- M = 0.43



https://github.com/henning-winker/spmpriors

stk = flmvn_traits(Genus="Sparus",Species="aurata",M=c(0.43,0.1),h=c(0.6,0.9),Plot=T)

European Commission

Simulation Run with LIME: Sparus aurata

NRC Research Press 1829 ARTICLE

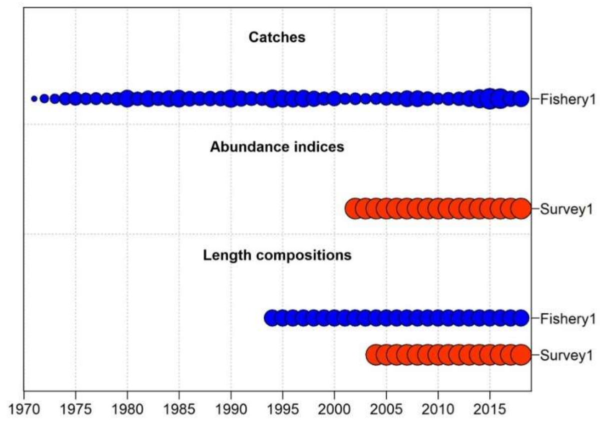
A new role for effort dynamics in the theory of harvested populations and data-poor stock assessment

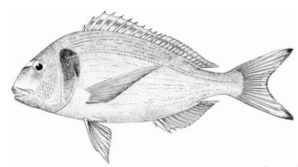
James T. Thorson, Cóilín Minto, Carolina V. Minte-Vera, Kristin M. Kleisner, and Catherine Longo

NRC Research Press 1 ARTICLE

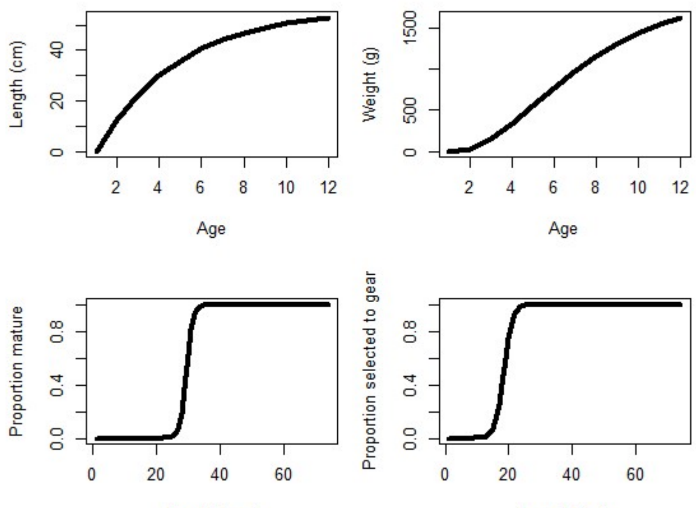
Accounting for variable recruitment and fishing mortality in length-based stock assessments for data-limited fisheries

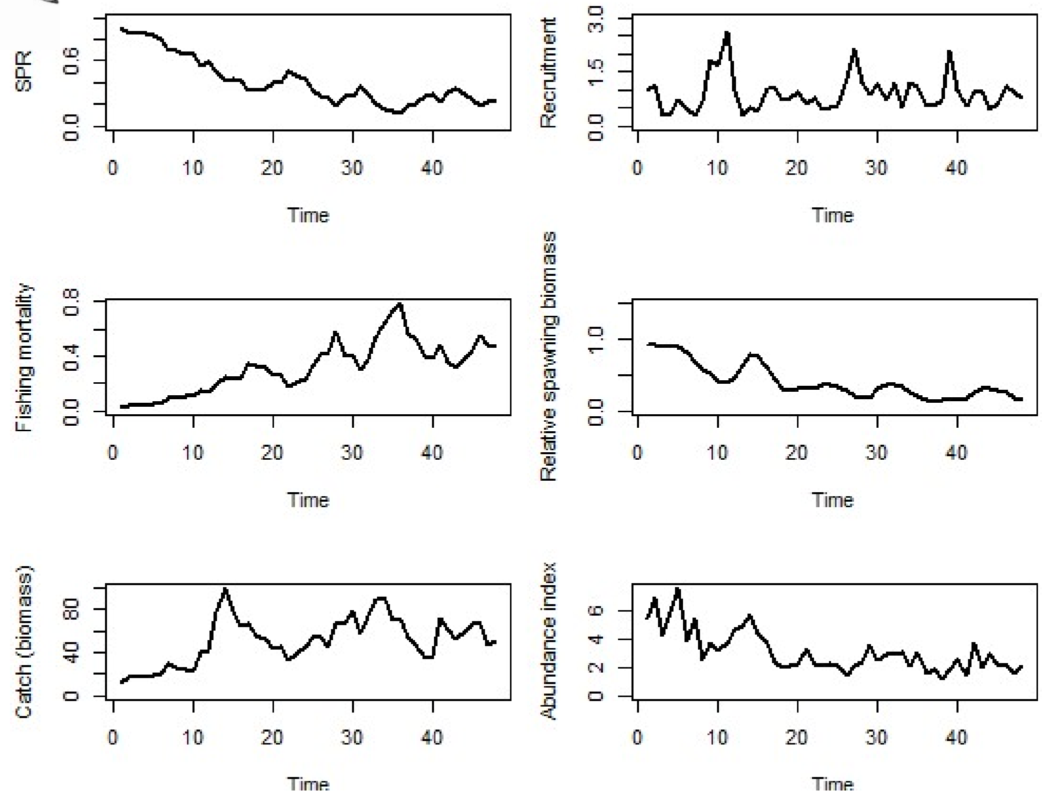
Merrill B. Rudd and James T. Thorson



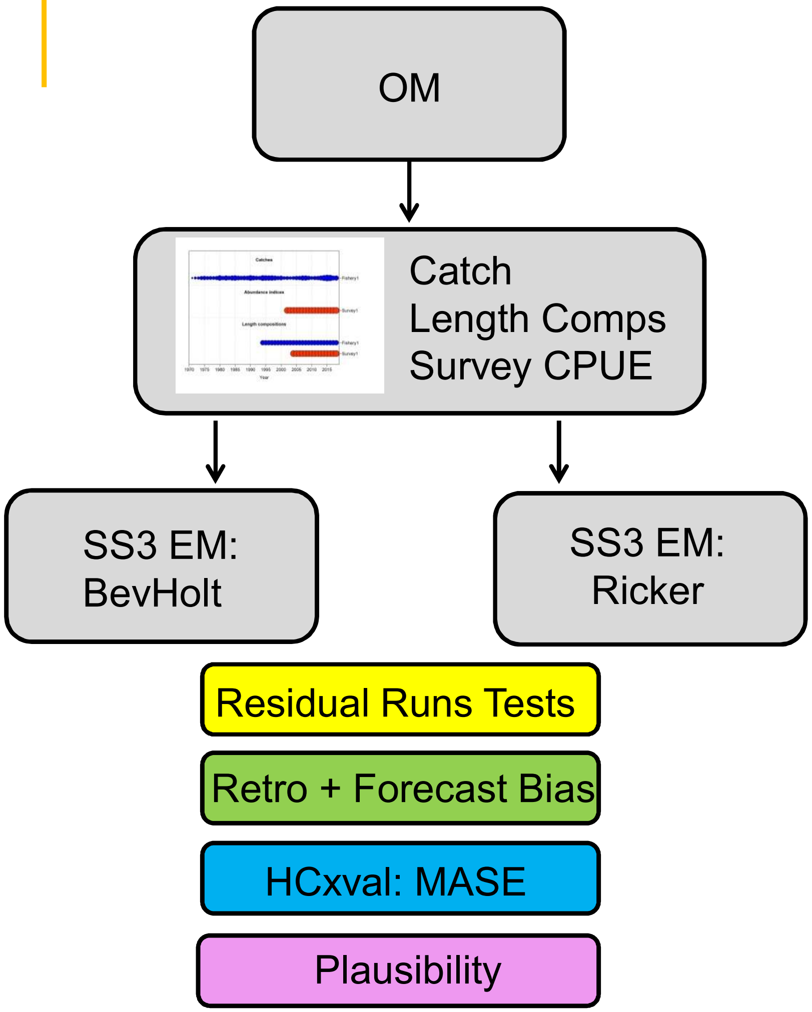


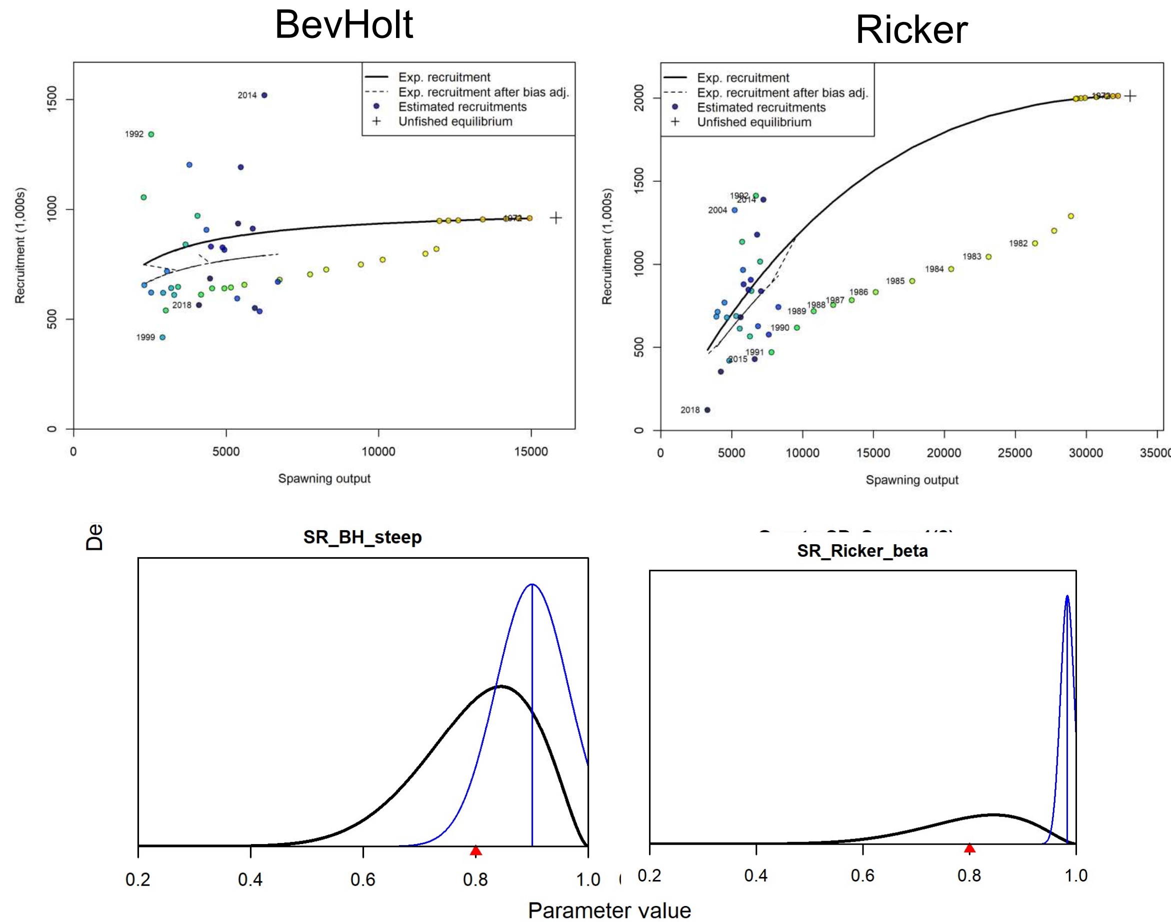
- SSR BevHold with h = 0.87
- M = 0.43





Simulation Example with LIME: Sparus aurata

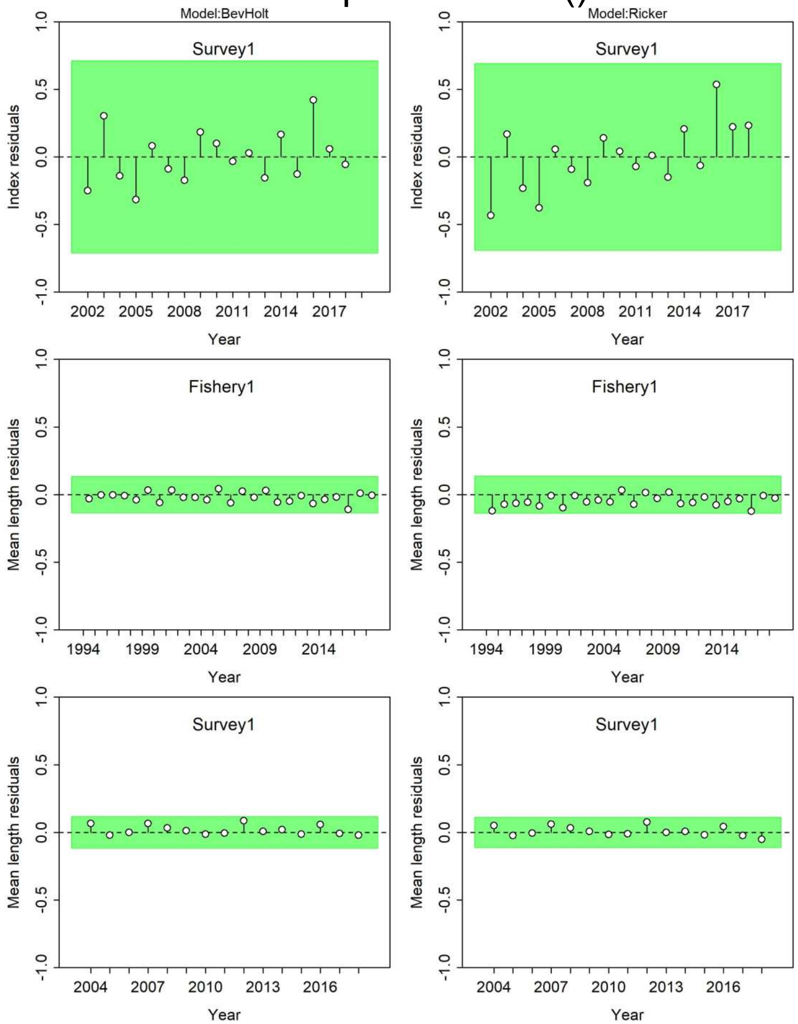


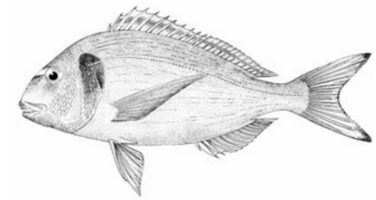


European Commission

Simulation Example with LIME: Sparus aurata

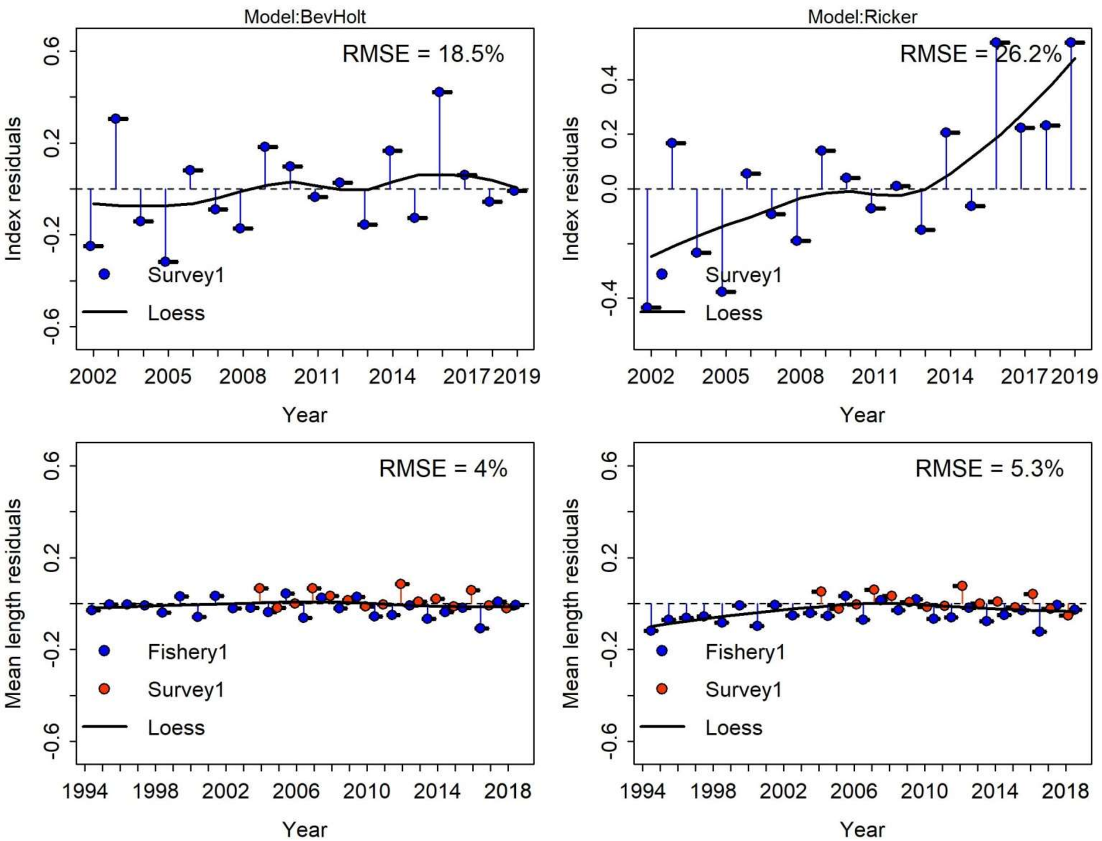
SSplotRunstest()

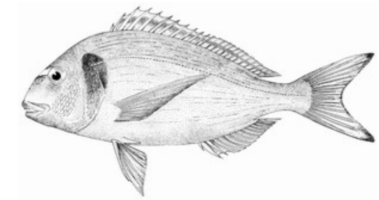




Simulation Example with LIME: Sparus aurata

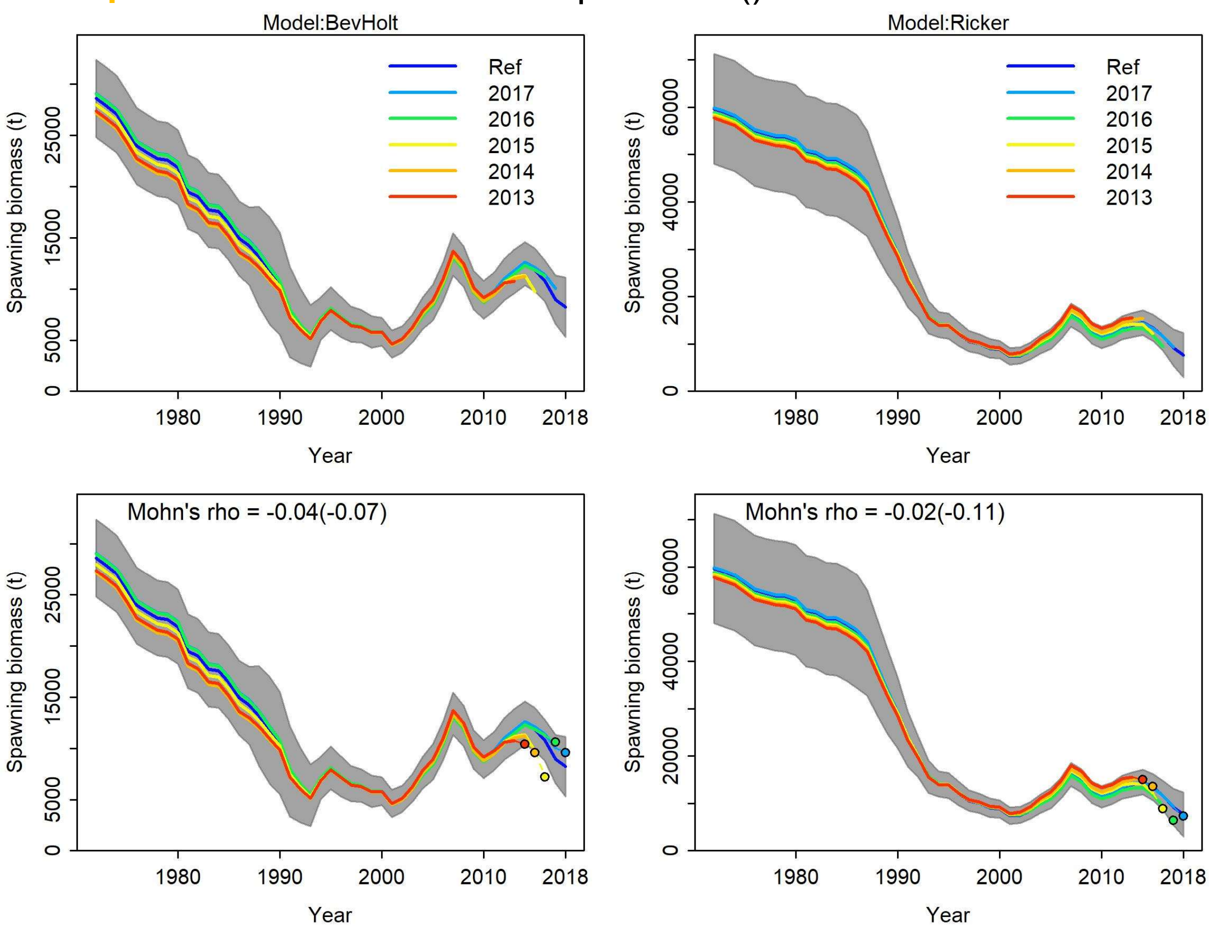
SSplotJABBAres()

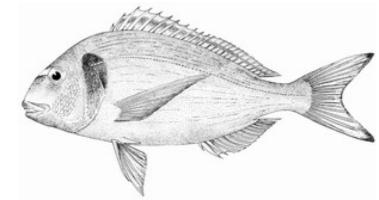




Simulation Example with LIME: Sparus aurata

SSplotRetro()

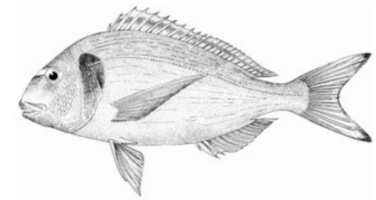


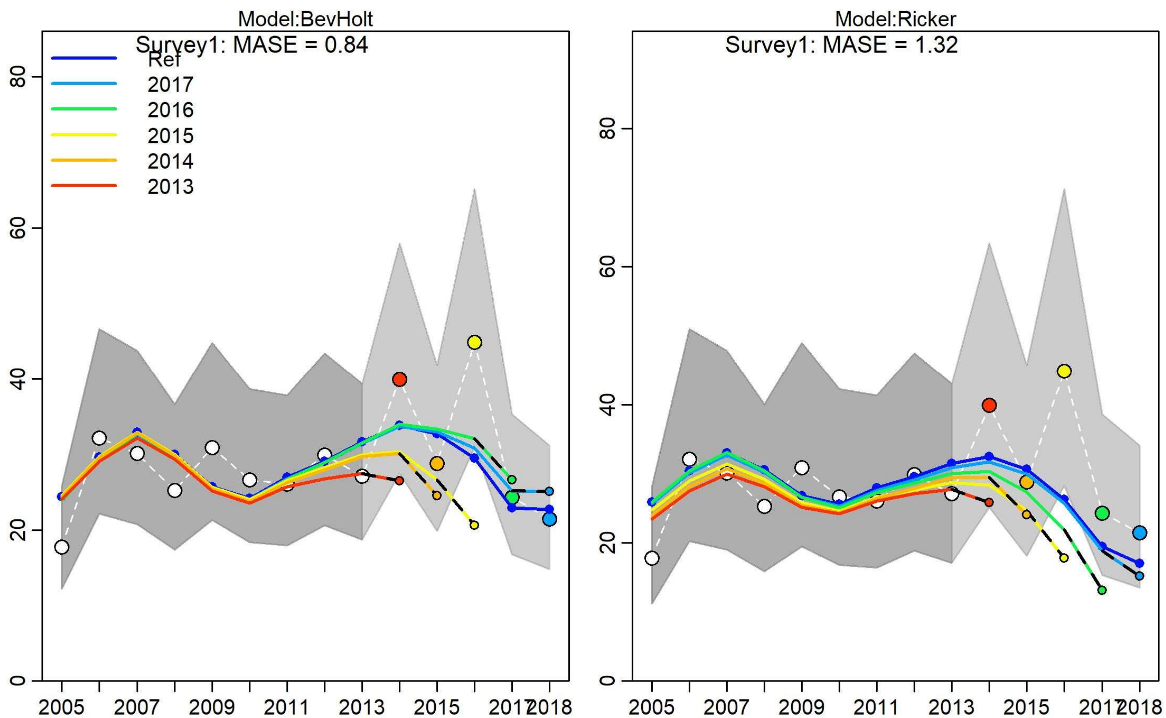


European Commission

Simulation Example with LIME: Sparus aurata

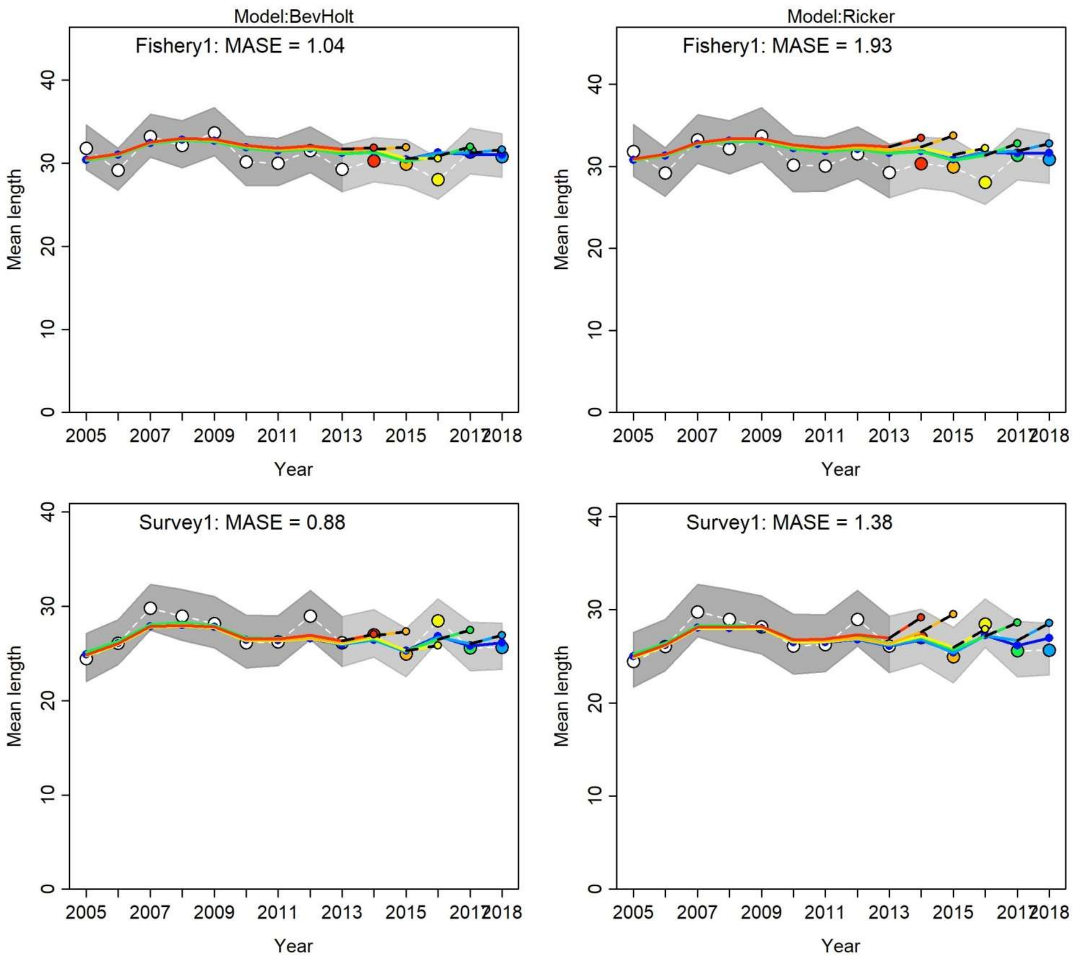
SSplotHCxval()

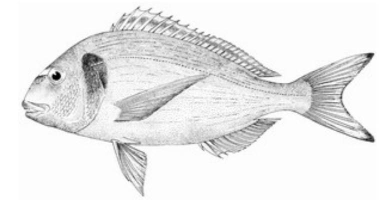




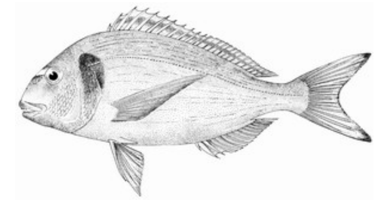
Simulation Example with LIME: Sparus aurata

SSplotHCxval()





Simulation Example with LIME: Sparus aurata



Plausibility

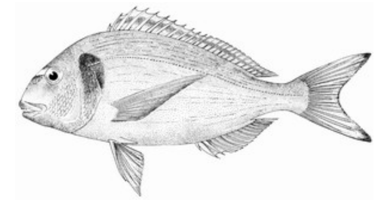
Diagnostics	Data	Model: BevHolt	Model: Ricker
Joint RMSE	Survey Index	**18.5**	26.2
	Length Comps	**4**	5.3
Runs Test	Survey Index	Passed	Passed
	Survey Mean Length	Passed	Passed
	Fishery Mean Length	Passed	Passed
Mohn's Rho	SSB	-0.04	**-0.02**
Forecast Bias	SSB	**-0.07**	-0.11
HCxval MASE	Survey Index	**0.84**	1.32
	Survey Mean Length	**0.88**	1.38
	Fishery Mean Length	**1.04**	1.93

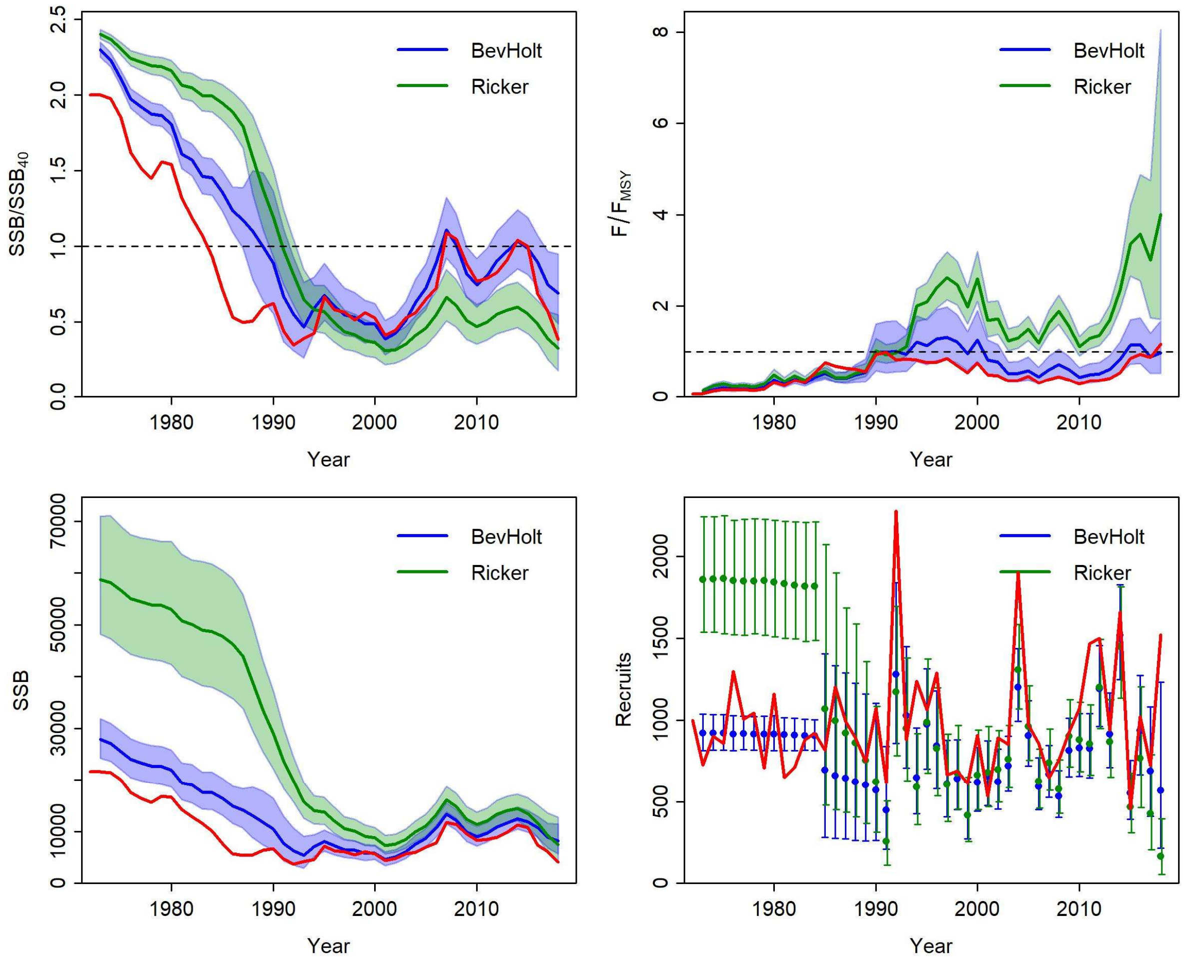
Diagnostics to add: ASPM class

Statistics to note: LL, AIC, convergence

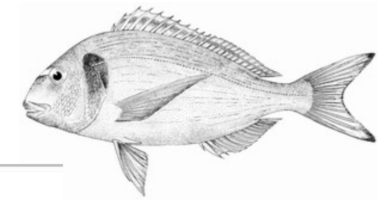
European Commission

Simulation Example with LIME: Sparus aurata

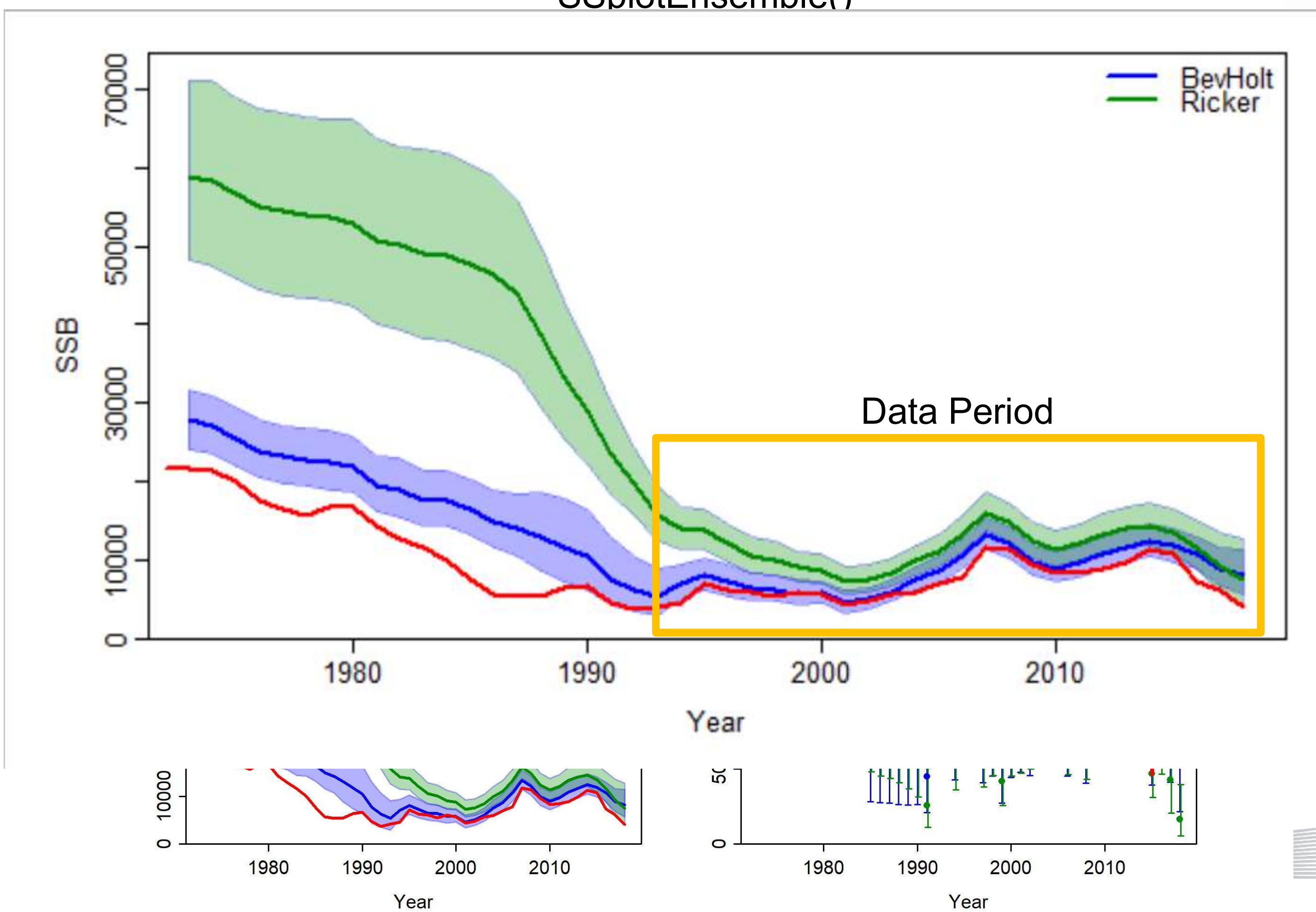




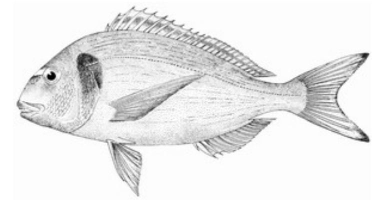
Simulation Example with LIME: Sparus aurata



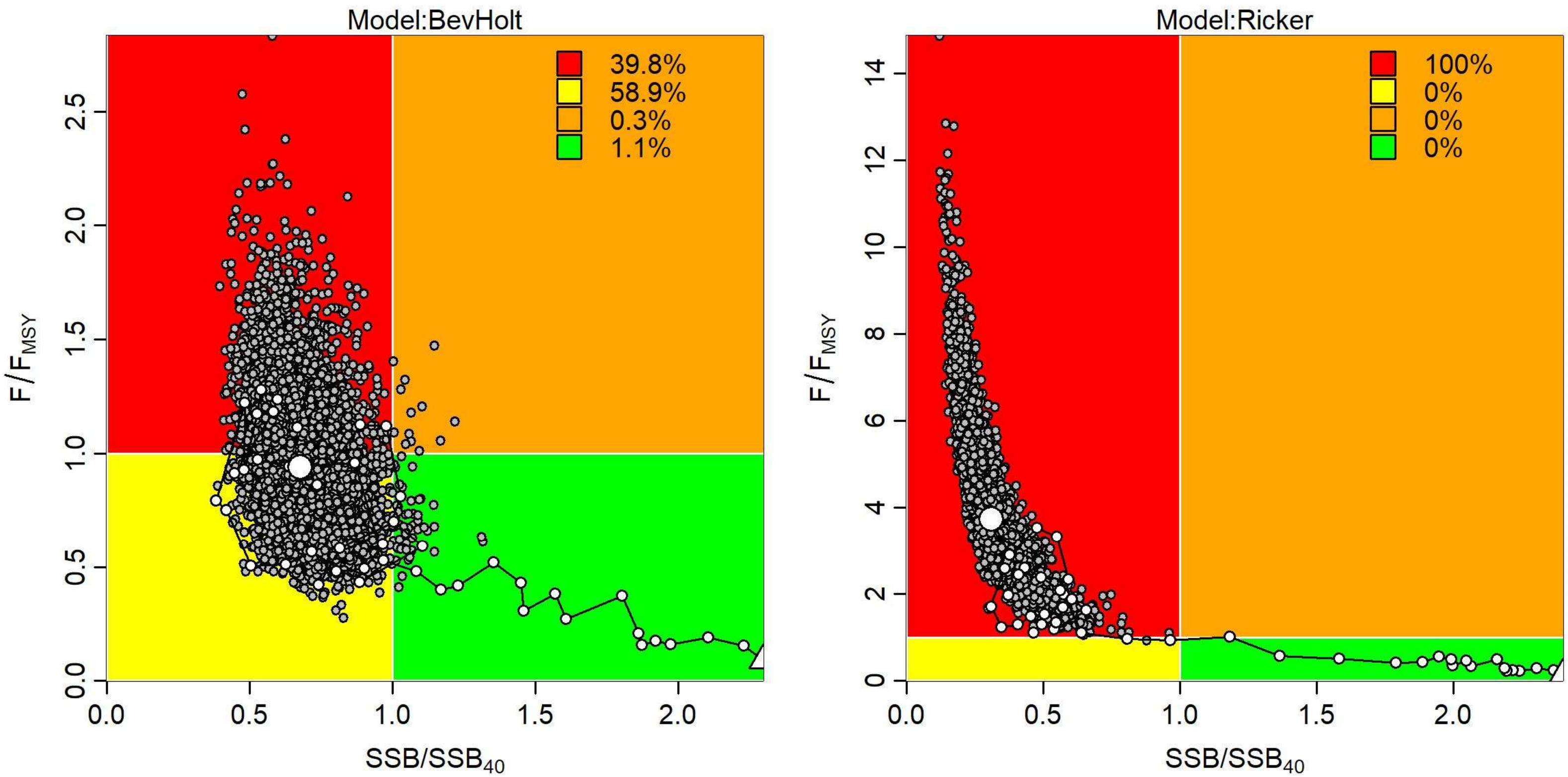
SSplotEnsemble()



Simulation Example with LIME: Sparus aurata



Implications for status advice



Remarks on model plausibility

"The likelihood of a scenario considered in simulation trials representing reality, relative to other scenarios also under consideration. Plausibility may be estimated formally based on **some statistical approach**, or specified based on expert judgement, and can be used to weight performance statistics when integrating over results for different scenarios. […] The aim of conditioning is to select those OMs consistent with the data and reject OMs that do not fit these data satisfactorily and, as such, are considered implausible."

RFMO MSE Glossary 2018

“The **SC AGREED** that it is useful to develop a set of generic criteria for model plausibility, utilising best practice in evaluating model **convergence** and **data fits**, **retrospective pattern** and **forecast bias**, and **prediction skill**, as well as other potential aspects of model diagnostics. The **SC NOTED** that establishing such guidance and criteria can help ensure that the stock assessments are transparent and comprehensive and allow stakeholders to have a good grasp of the scientific process. Stock specific plausibility criteria can also be considered to evaluate if assessment results are consistent with prior knowledge about the exploitation history and population biology.

Report of IOTC Scientific Committee 2020

"A highly **plausible** scenario is one that fits prior knowledge, with many sources of corroboration, without the complexity of explanation, and with minimal conjecture (Connell, 2006). Plausibility may be determined formally, based on a statistical approach to determine whether a **system equivalent to the model generated the data** or based on expert judgement."

Laurence T. Kell, Coilin Minto, and Hans D. Gerritsen. IJMS. Evaluation of the skill of length-based indicators to identify stock status and trends, accepted 2022.