Practical applications of diagnostic weighing in ensemble models: three case studies, Northern shrimp, Adriatic sole and Gulf of Bothnian vendace

Max Cardinale, Henning Winker, Francesco Masnadi, David Gilljam & Chris Griffiths

FishMed-PhD
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

SLU Swedish University of Agricultural Sciences

CNR IRBIM ISTITUTO PER LE RISORSE BIOLOGICHE E LE BIOTECNOLOGIE MARINE

Why Ensemble Modeling

Current practices in stock assessment consists in comparing different models and choosing the **“best case scenario”**

In this context, discarded models have no effect (weight = 0) on management even if they represent a plausible realization of the truth

The main objective when using an **ensemble of models** in stock assessments is to quantify the **total uncertainty** across all plausible models, where the structural uncertainty is likely to be much greater than the within model uncertainty

Moreover, ensemble forecasting has been shown to generally being able to **improve forecast accuracy** in many fields, particularly in weather forecasting where the method originated and transport science

Why weighting models in an ensemble

The need to **weigh models** based on information in the available data is well recognised in several sciences but it is often difficult to do so within the context of fisheries stock assessment models

Model weighting is a necessary step because assigning the same weight (reliability) to all hypotheses could introduce biases into the management advice if some hypotheses are, in fact, more unlikely than others

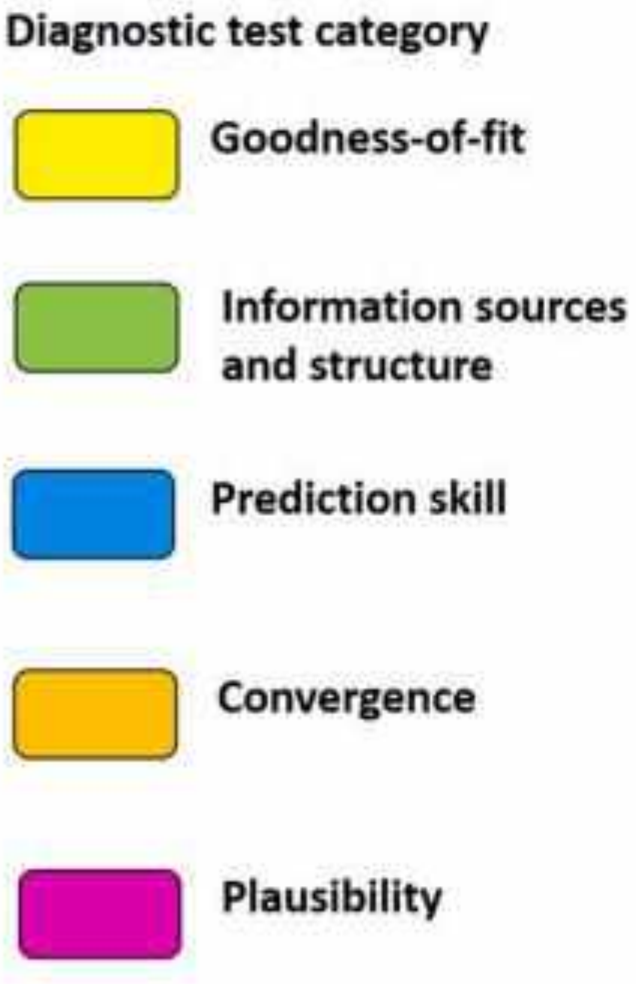
Model diagnostics have the potential to be used to weight models

Model diagnostics

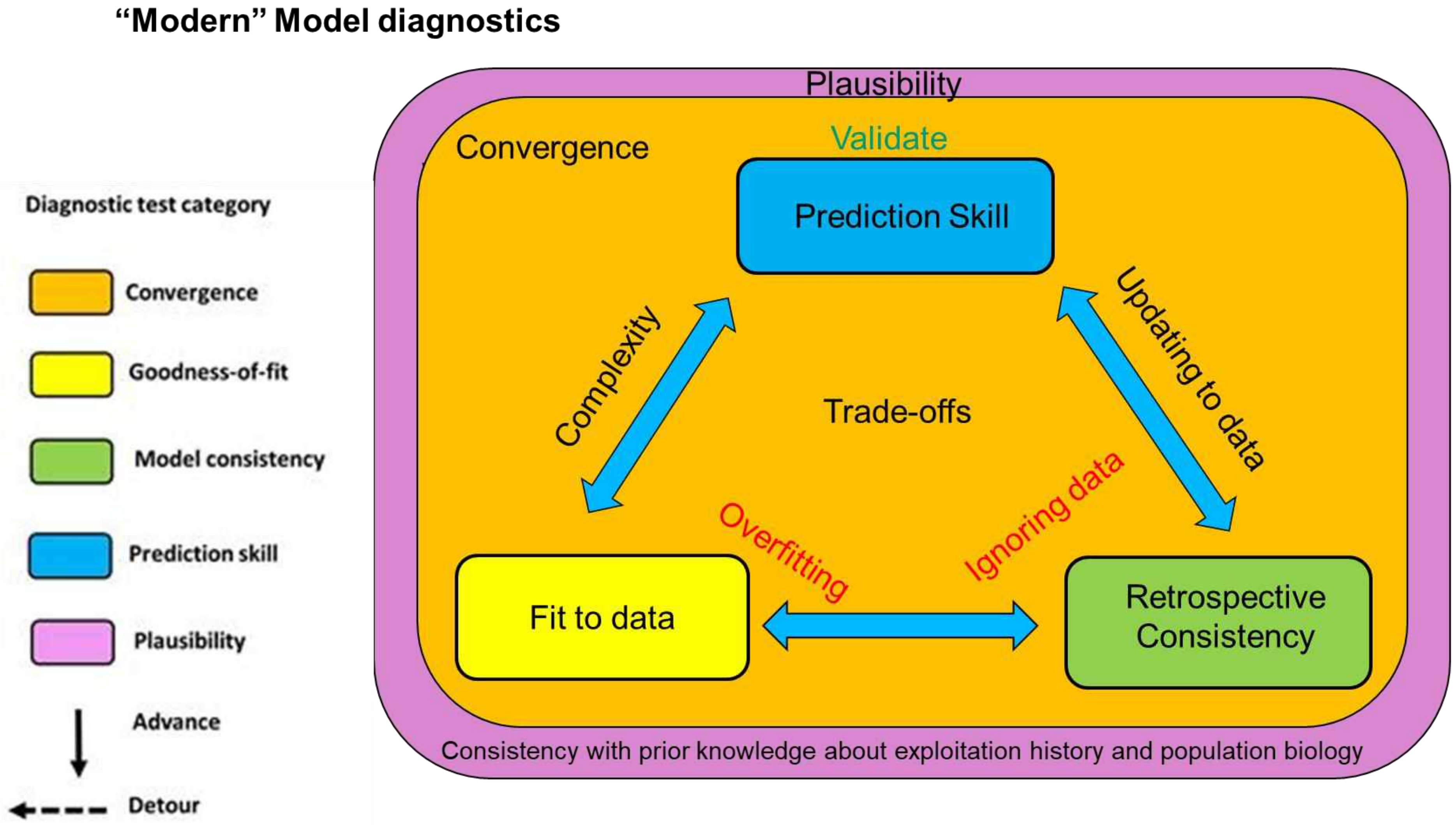
1. Felipe Carvalho, Henning Winker, Dean Courtney, Laurence Kell, Maia Kapur, Massimiliano Cardinale, Michael Schirripa, Toshihide Kitakado, Dawit Y. Ghebrehiwet, Kevin R. Piner, Mark N. Maunder, Rick Methot, 2021. A Cookbook for Using Model Diagnostics in Integrated Stock Assessments. Fisheries Research, https://doi.org/10.1016/j.fishres.2021.105959.

2. Laurie Kell, Rishi Sharma, Toshihide Kitakado, Henning Winker, Iago Mosqueira, Massimiliano Cardinale, Dan Fu, 2021. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science, https://doi.org/10.1093/icesjms/fsab104.

3. Gorka Merino, Agurtzane Urtizberea, Dan Fu, Henning Winker, Massimiliano Cardinale, Mathew Lauretta, Hilario Murua, Toshihide Kitakado, Haritz Arrizabalaga, Robert Scott, Graham Pilling, Ane Laborda, Maite Erauskin-Extraminiana, Josu Santiago 2022. Investigating trends in process error as a diagnostic for integrated fisheries' stock assessments. Fisheries Research, https://doi.org/10.1016/j.fishres.2022.106478.

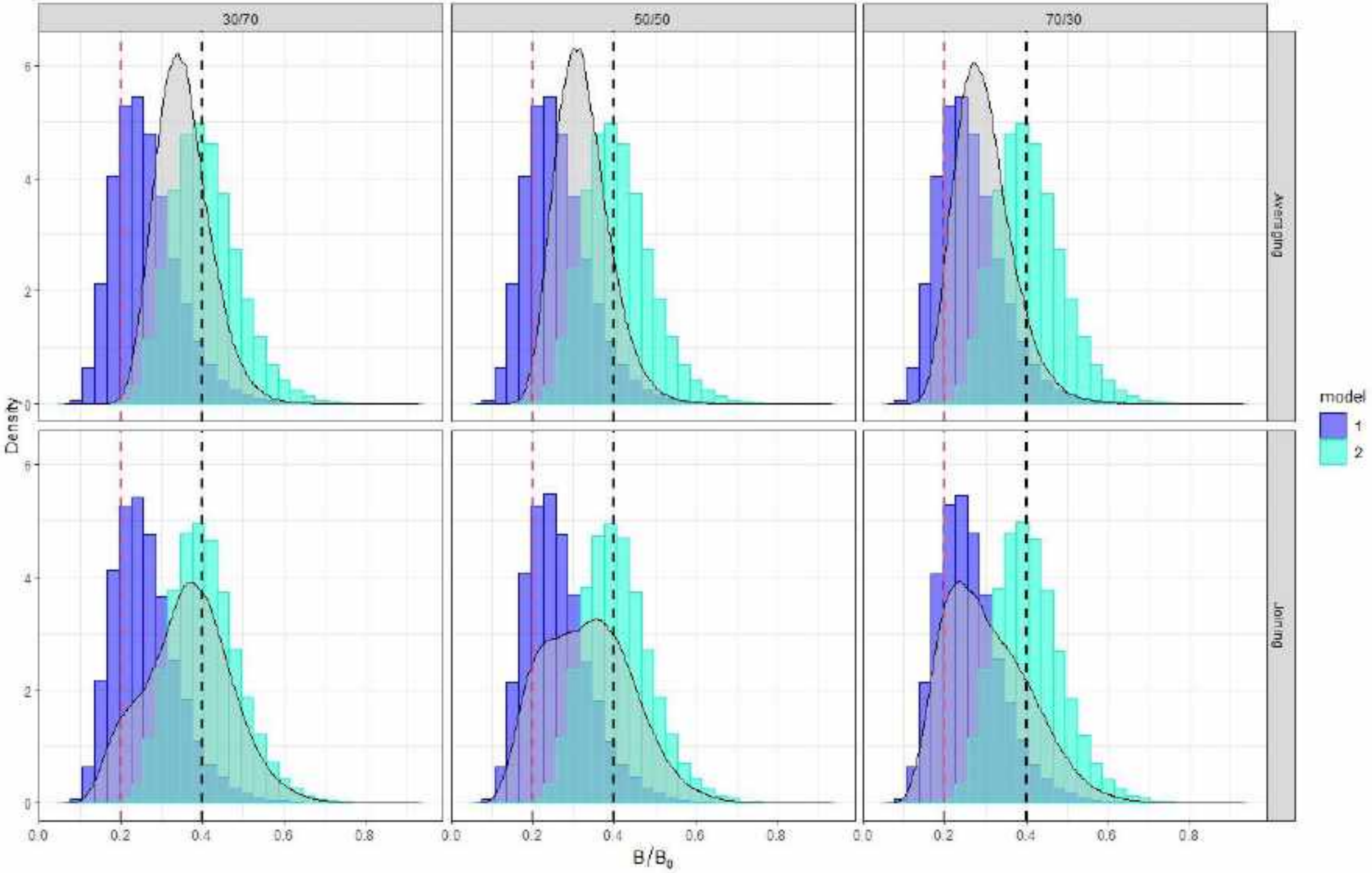


Model diagnostics



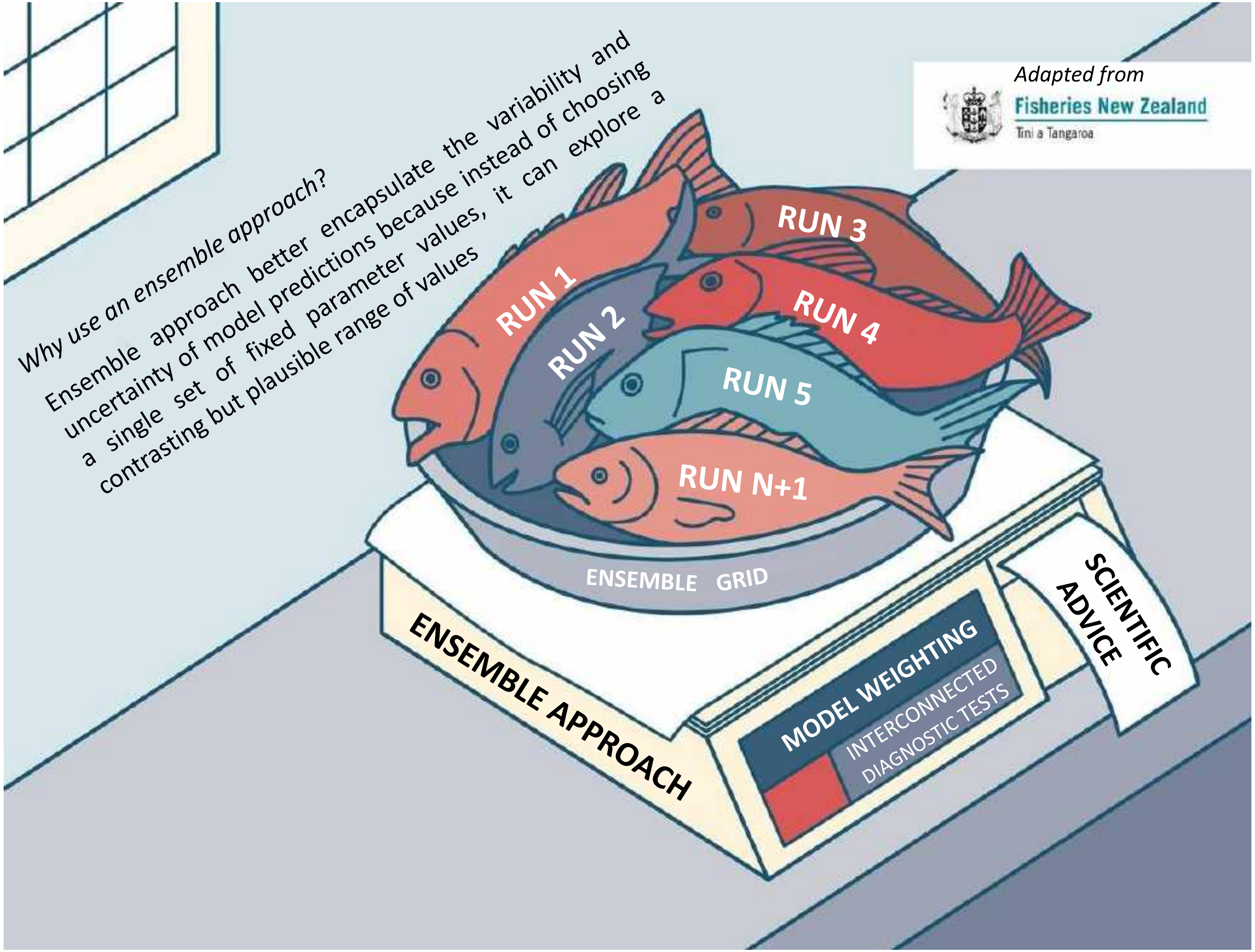
When all **model diagnostic** tests are considered together, the power to detect model misspecification improves without a substantial increase in the probability of incorrectly rejecting a correctly specified model

Model averaging and model stitching



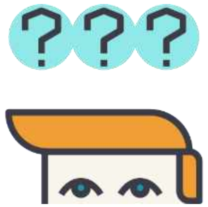
Thus, even if bias divergence is large (even if bidirectional) and correlation between models is also large, then we might loose the benefit of improving accuracy, but we still preserve the tails for a better quantification of risk.

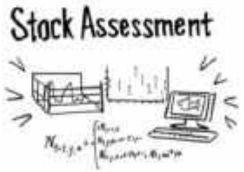
Assessment model framework: the ensemble approach



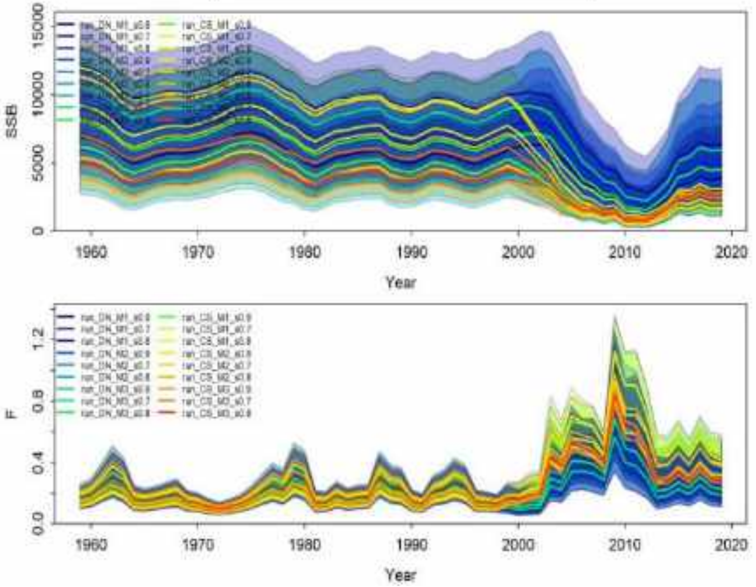
- ***Stock Synthesis 3 + Ensemble approach (delta-MVLN)***

to address structural uncertainties, a range of alternative models was selected through diagnostics, to be stitched together





Grid of alternative runs



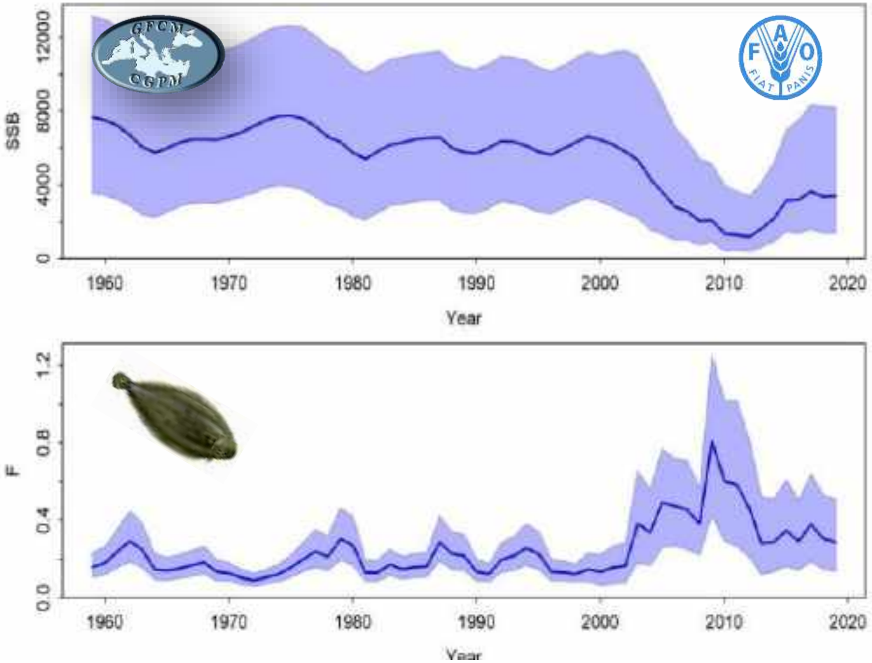
Diagnostic test category

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

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

Henning Winker, John Walter, Massimiliano Cardinale, and Dan Fu

Ensemble model



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

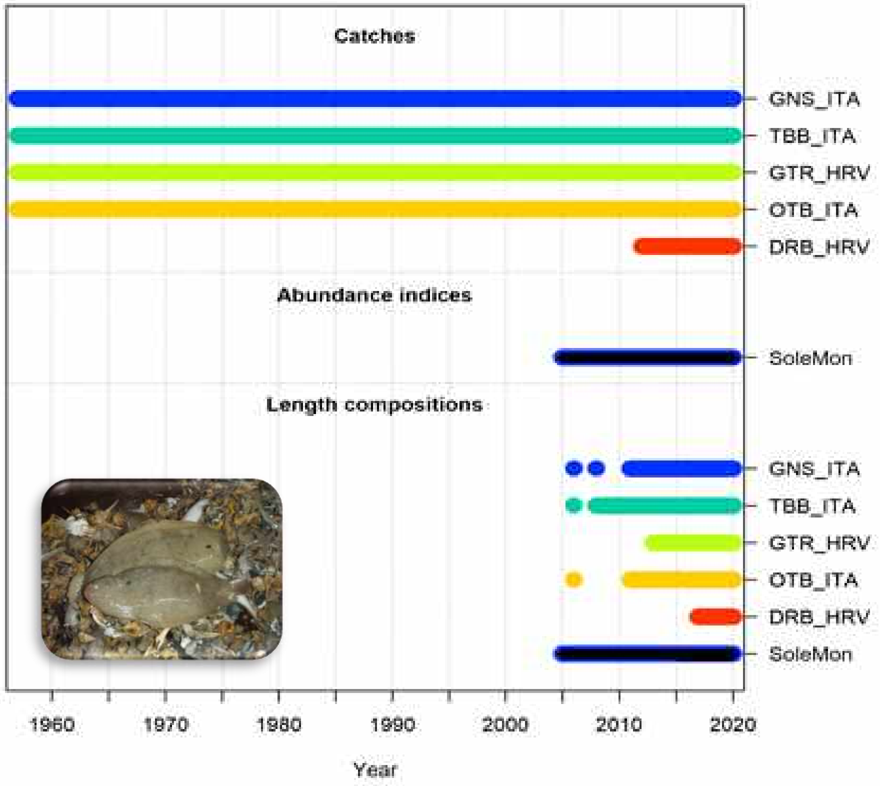
The 1st application of ensemble modelling in GFCM stock assessment and advice

Adriatic sole

- ***Stock Synthesis 3 -*** Statistical age structured population modelling using a wide range of minimally processed fishery and survey data (Maunder and Punt 2013)

https://vlab.ncep.noaa.gov/web/stock-synthesis

Model structure: one-area yearly model where the population is comprised of 15+ age-classes with sexes combined



GNS ITA

TBB ITA

GTR HRV

OTB ITA

DRB HRV

SoleMon

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

Set_1: Length based model based on VBGP and all fleets selectivity double normal **(DN)**

- **Long** landing time series (1958-2019)
  - Initial catch:1953= average 5 years
    - M1 — Average of Gislason & Chen-Watanabe: Steep0.7, Steep0.8, Steep0.9 → Jittering
    - M2 — Average of Then_nls,Then_lm,Hamel_Amax: Steep0.7, Steep0.8, Steep0.9
    - M3 — Average of Then_VBGF, Jensen_VBGF 1, Jensen_VBGF 2: Steep0.7, Steep0.8, Steep0.9
    - → *Ss3diags* Pass/Fail
  - NO Initial catch → Discarded: same result of initial catch average
- **Short** landing time series (1980-2019) → Discarded: same result of Long time series

Set_2: Length based model based on VBGP and commercial fleets + CAA for Survey (CAA)

→ Bad Diagnostic: Not enough age data

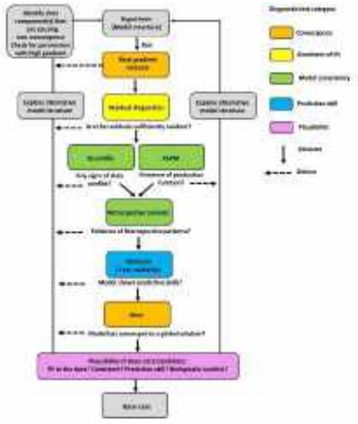
Set_3: Length based model based on VBGP and commercial fleets selectivity double normal + logistic for Survey (Lg)

→ Bad Diagnostic: some % of older specimens may not be detected by the survey

Set_4: Length based model based on VBGP and commercial fleets selectivity double normal + cubic splines for Survey (CS)

- **Long** landing time series (1958-2019)
  - Initial catch:1953= average 5 years
    - M1 — Average of Gislason & Chen-Watanabe: Steep0.7, Steep0.8, Steep0.9 → Jittering
    - M2 — Average of Then_nls,Then_lm,Hamel_Amax: Steep0.7, Steep0.8, Steep0.9
    - M3 — Average of Then_VBGF, Jensen_VBGF 1, Jensen_VBGF 2: Steep0.7, Steep0.8, Steep0.9
    - → *Ss3diags* Pass/Fail
  - NO Initial catch → Discarded: same result of initial catch average
- **Short** landing time series (1980-2019) → Discarded: same result of Long time series

Iterative



Automated

Ensemble_grid_SOLEA.R

Weighting based on ss3diags

ENSEMBLE MODEL → **Forecast**

Diagnostic - Convergence & stability

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

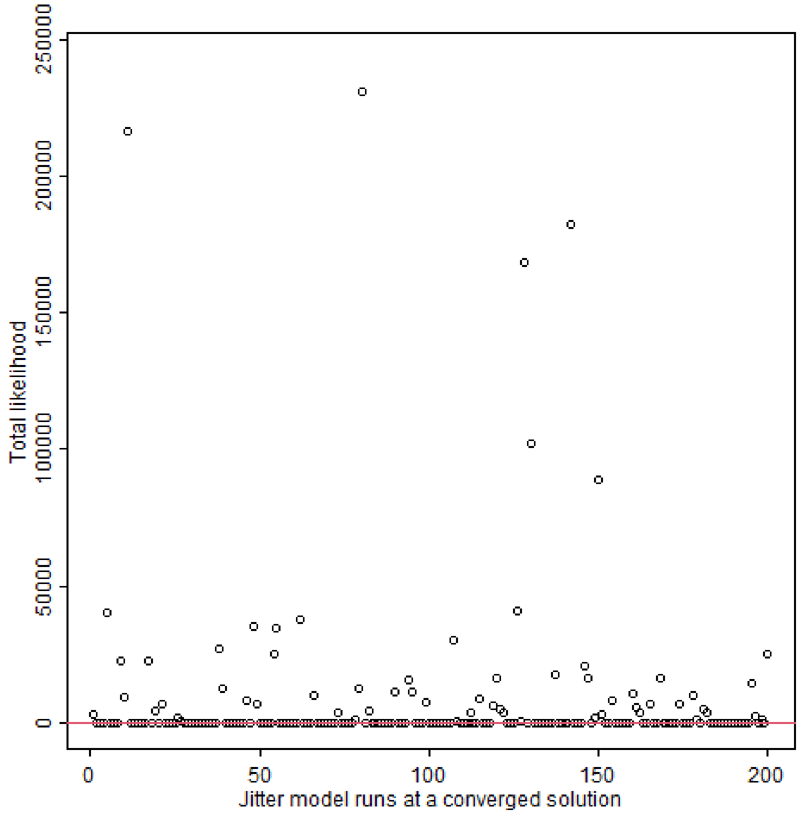
- Joint-residuals
- Runs tests

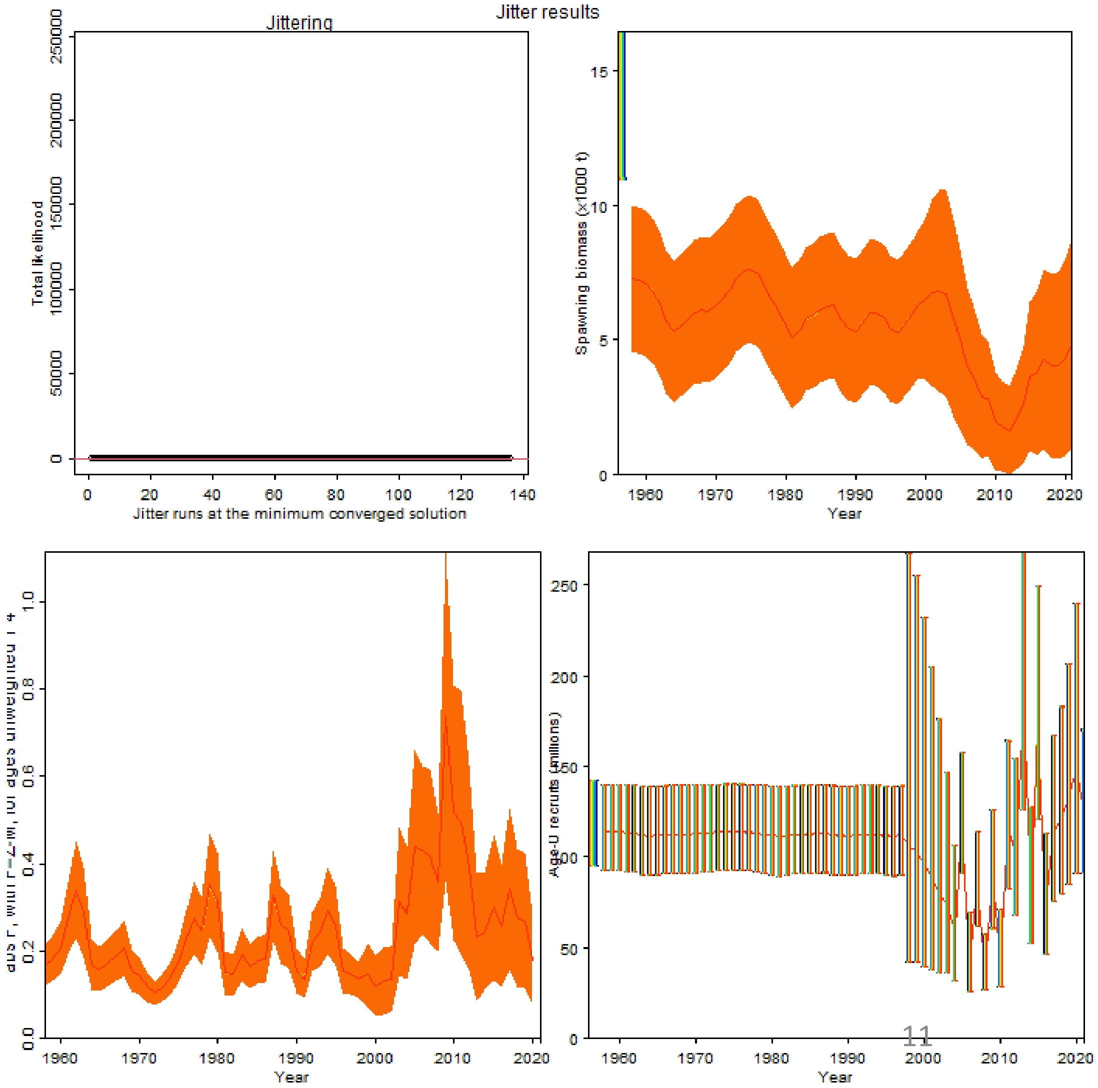
3. Consistency

- Retrospective analysis

4. Prediction skills

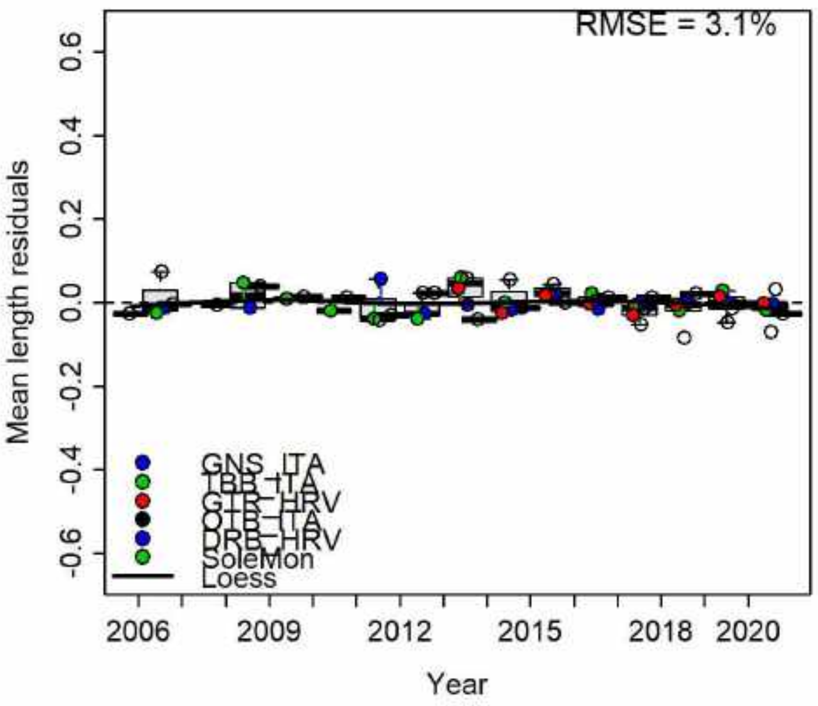
- Hindcasting (CrossValidation)



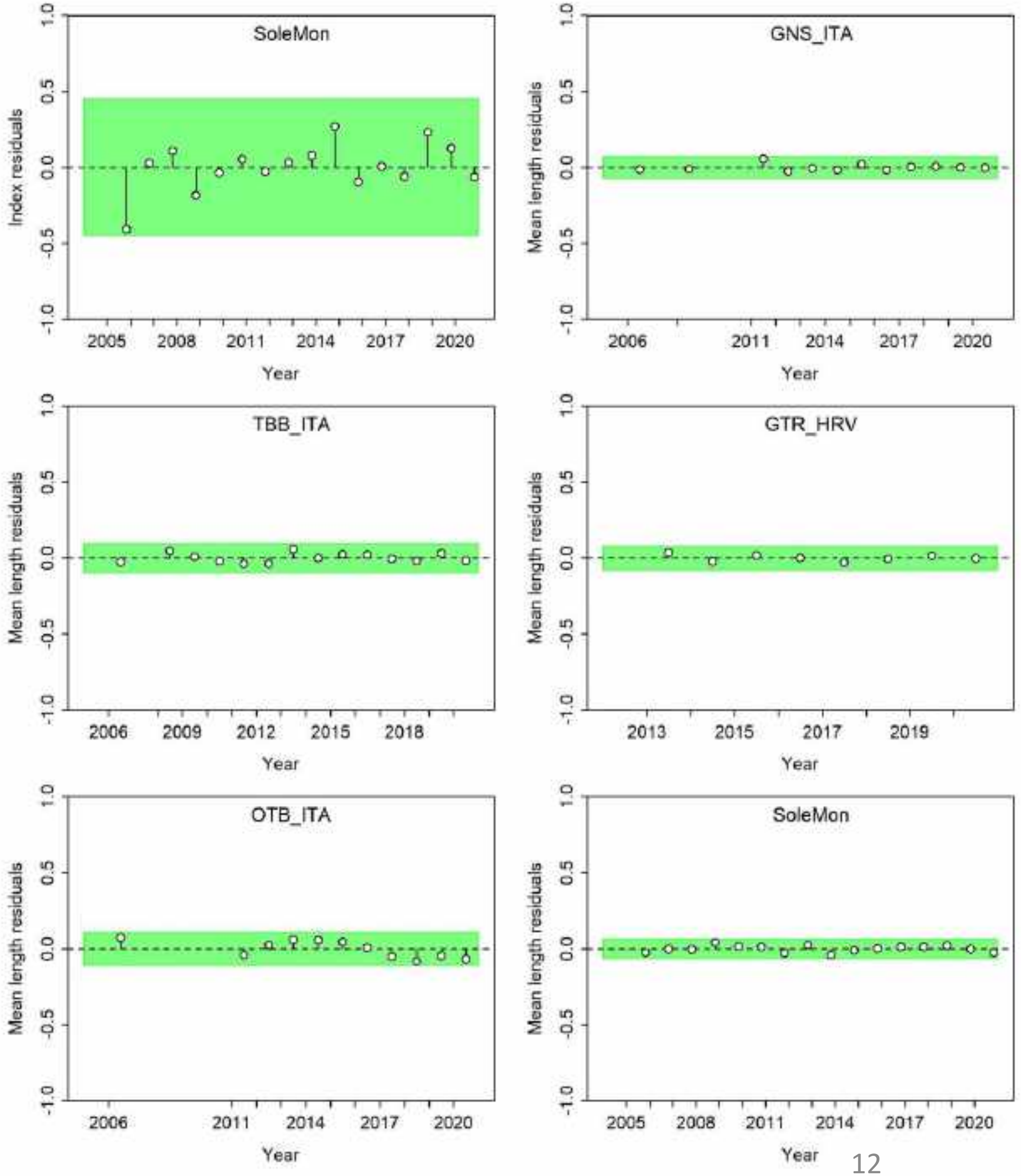


Diagnostic - Goodness of the fit

Joint-residuals



Runs tests



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 - Consistency

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

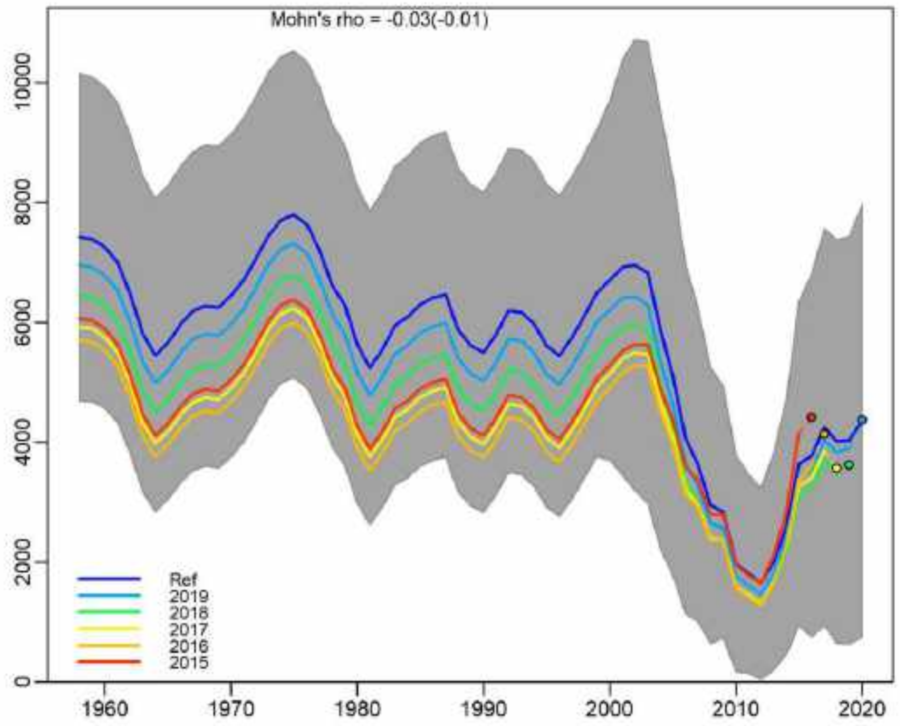
- Joint-residuals
- Runs tests

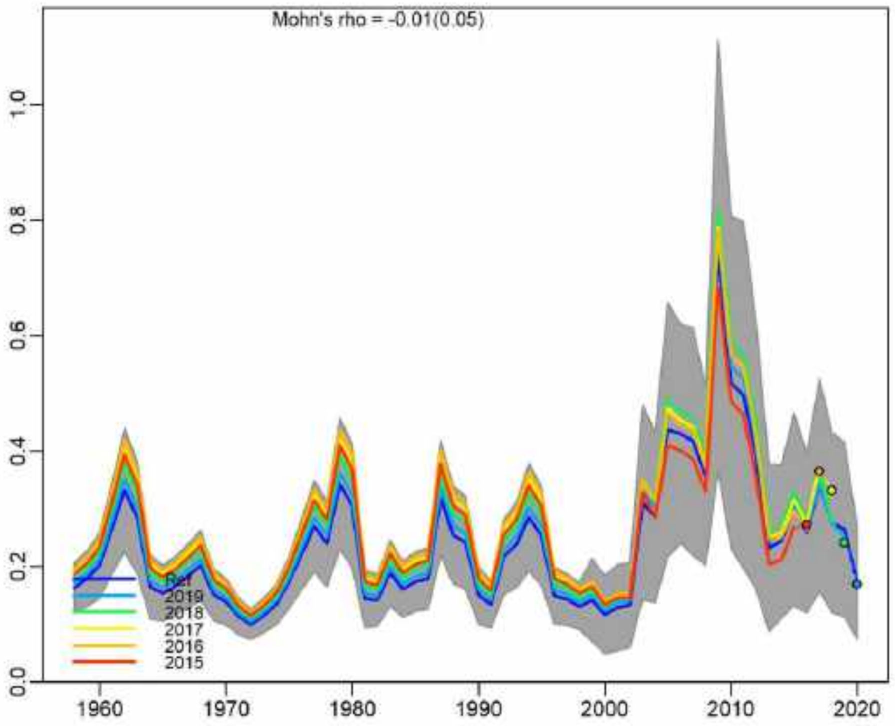
3. Consistency

- Retrospective analysis

4. Prediction skills

- Hindcasting (CrossValidation)





Diagnostic - Prediction skills

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

- Joint-residuals
- Runs tests

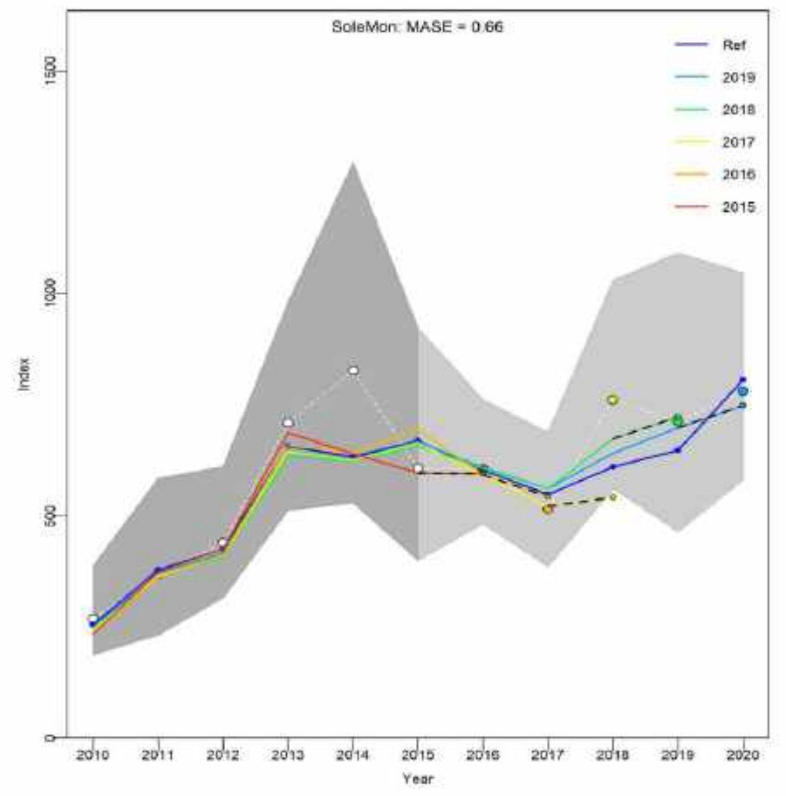
3. Consistency

- Retrospective analysis

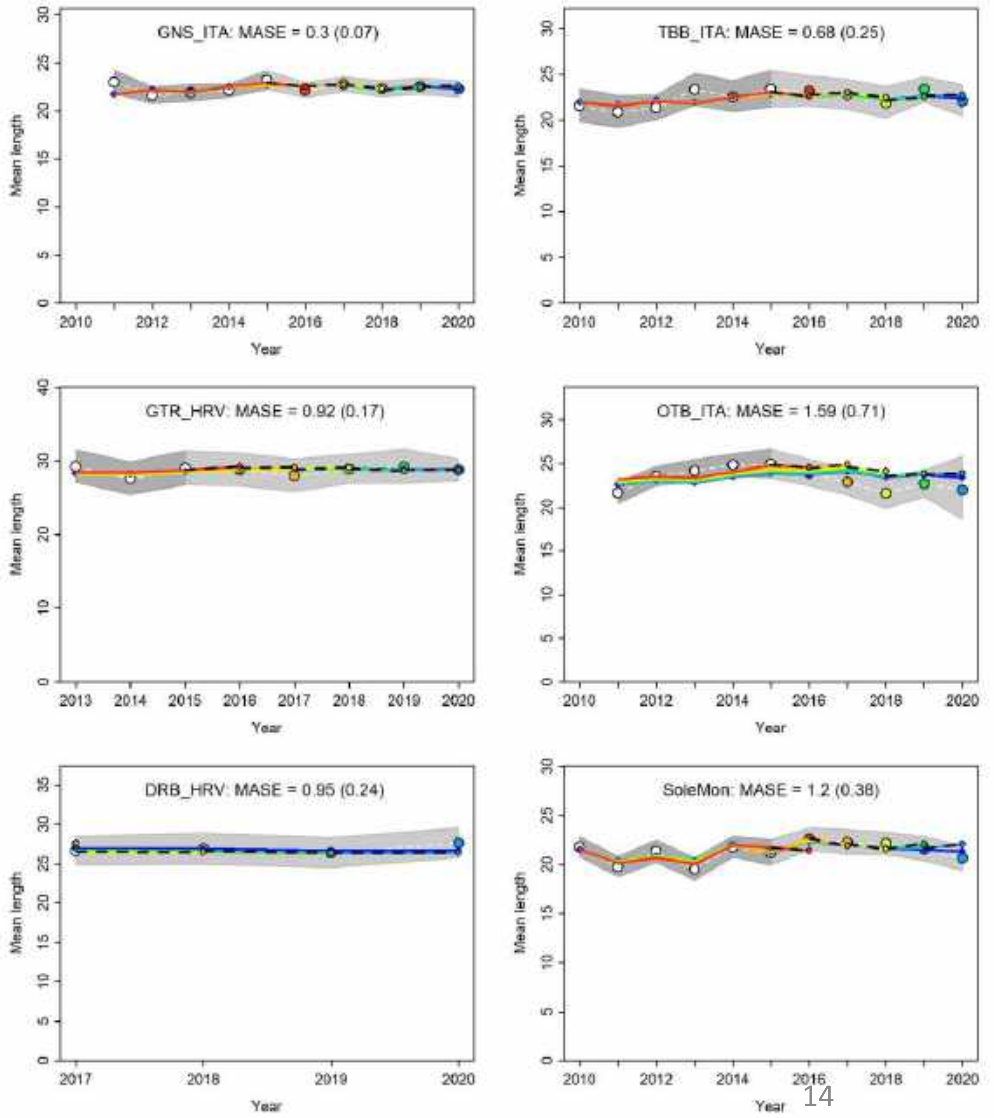
4. Prediction skills

- Hindcasting (CrossValidation)

MASE CPUE Index



MASE Length comp



Model weighting (diagnostic scores)

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

Where Runs test and MASE were aggregated in a single weight (balanced) with a 70% threshold

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

- Joint-residuals
- Runs tests

3. Consistency

- Retrospective analysis

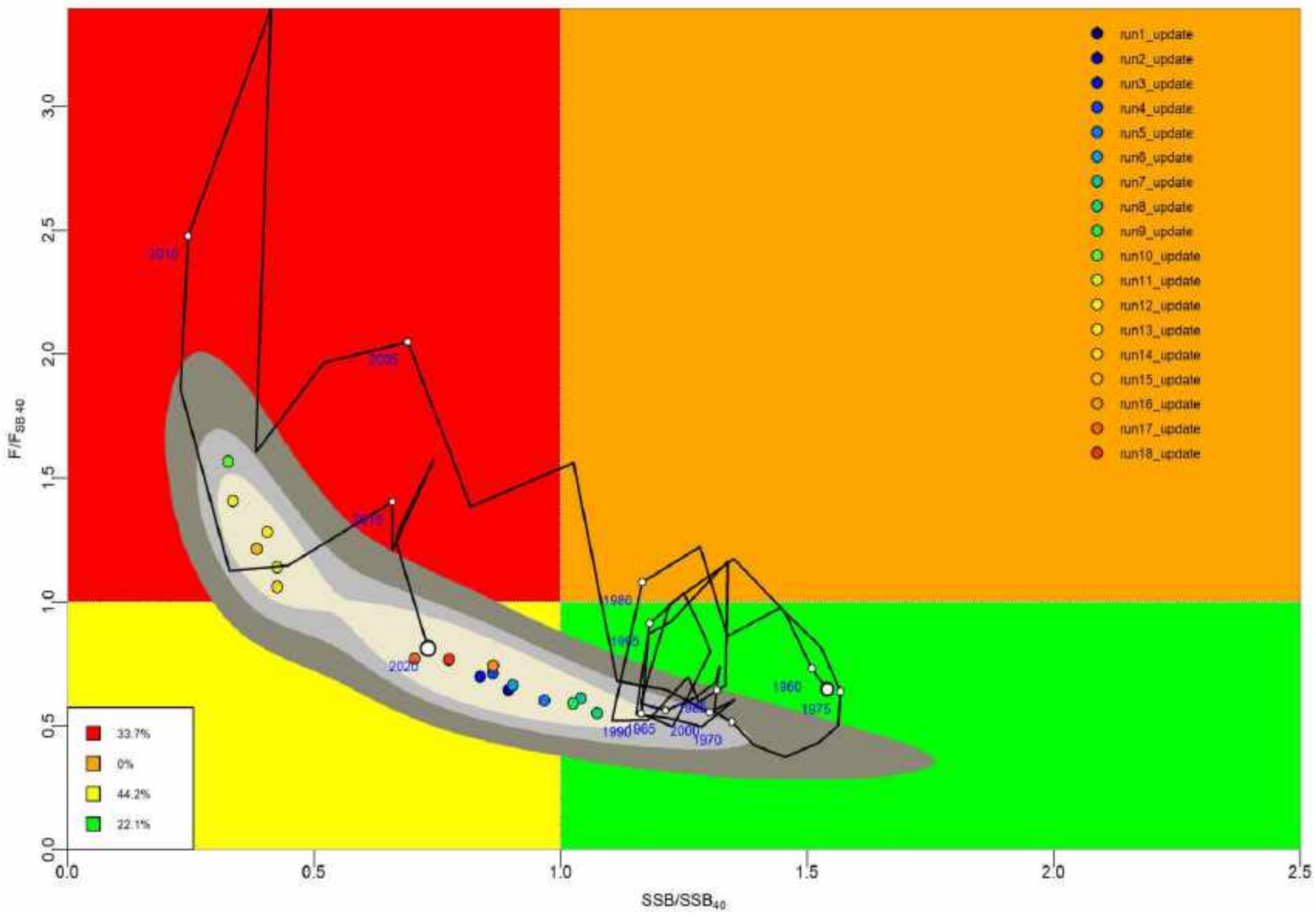
4. Prediction skills

- 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_F	Index	SurveyLen	COMfleet	*W(Diagnostics)*
Run1	Passed	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	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**

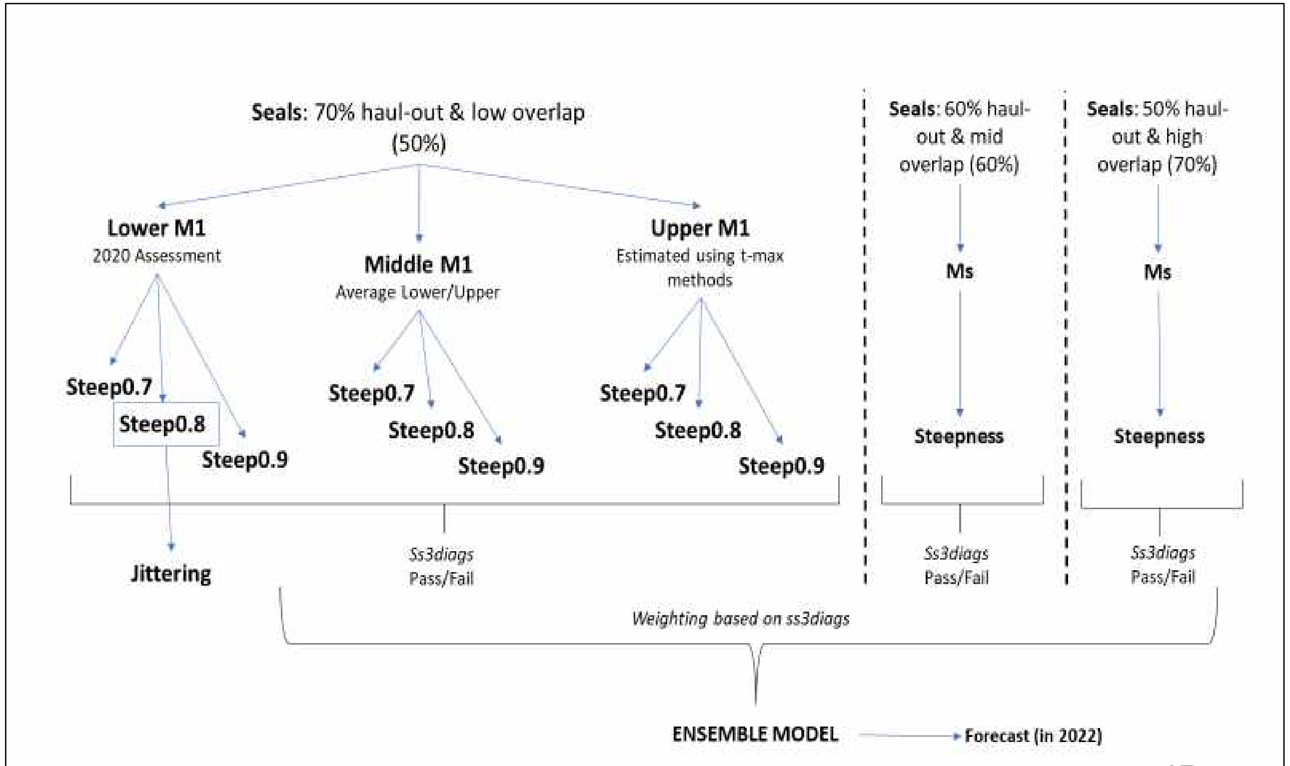
Ensemble model Results

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**



Bothnian Sea vendace

- The model uses a weighted median of plausible scenarios
  - In total 27 scenarios, exploring 3 x 3 x 3 levels of
    - Seal predation
    - Basal Natural mortality
    - Steepness



Bothnian Sea vendace

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

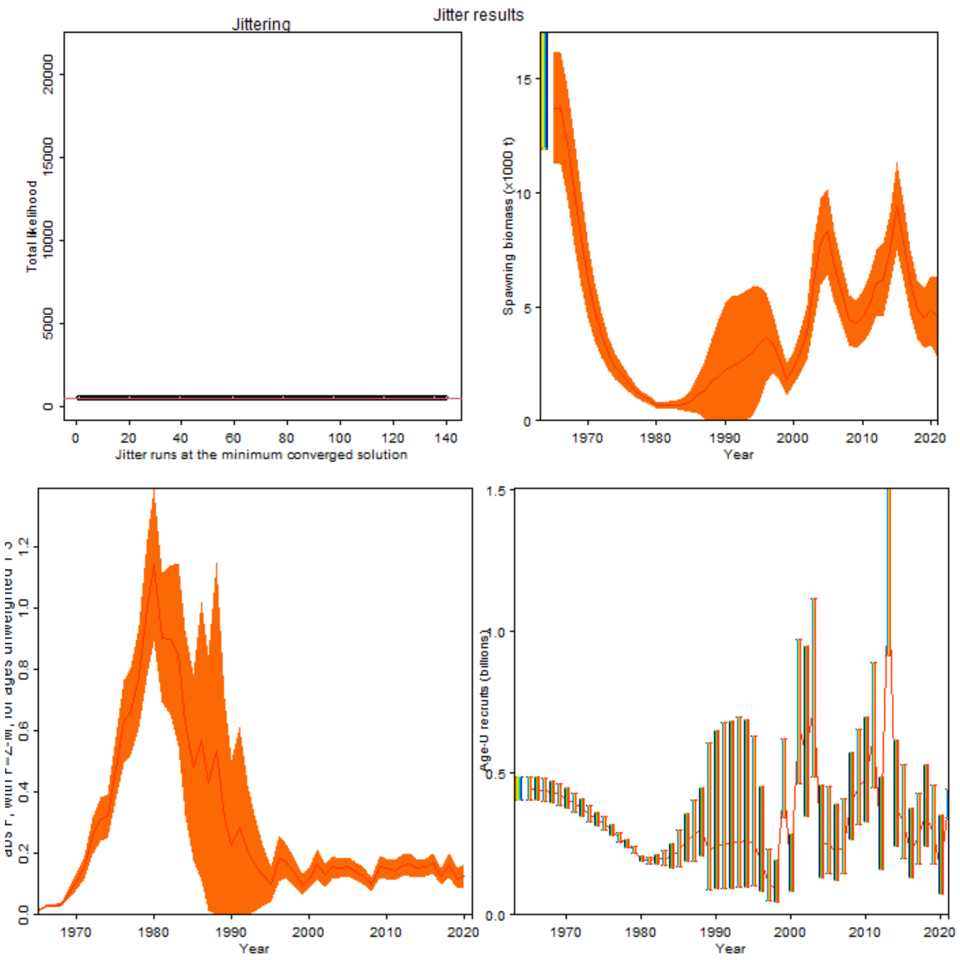
- Joint-residuals
- Runs tests

3. Consistency

- Retrospective analysis

4. Prediction skills

- Hindcasting (CrossValidation)



Model weighting

Where Runs test and MASE were aggregated in compartmental weight but no threshold

1. Convergence & stability

- Positive Hessian
- Jittering

2. Goodness of the fit

- Joint-residuals
- Runs tests

3. Consistency

- Retrospective analysis

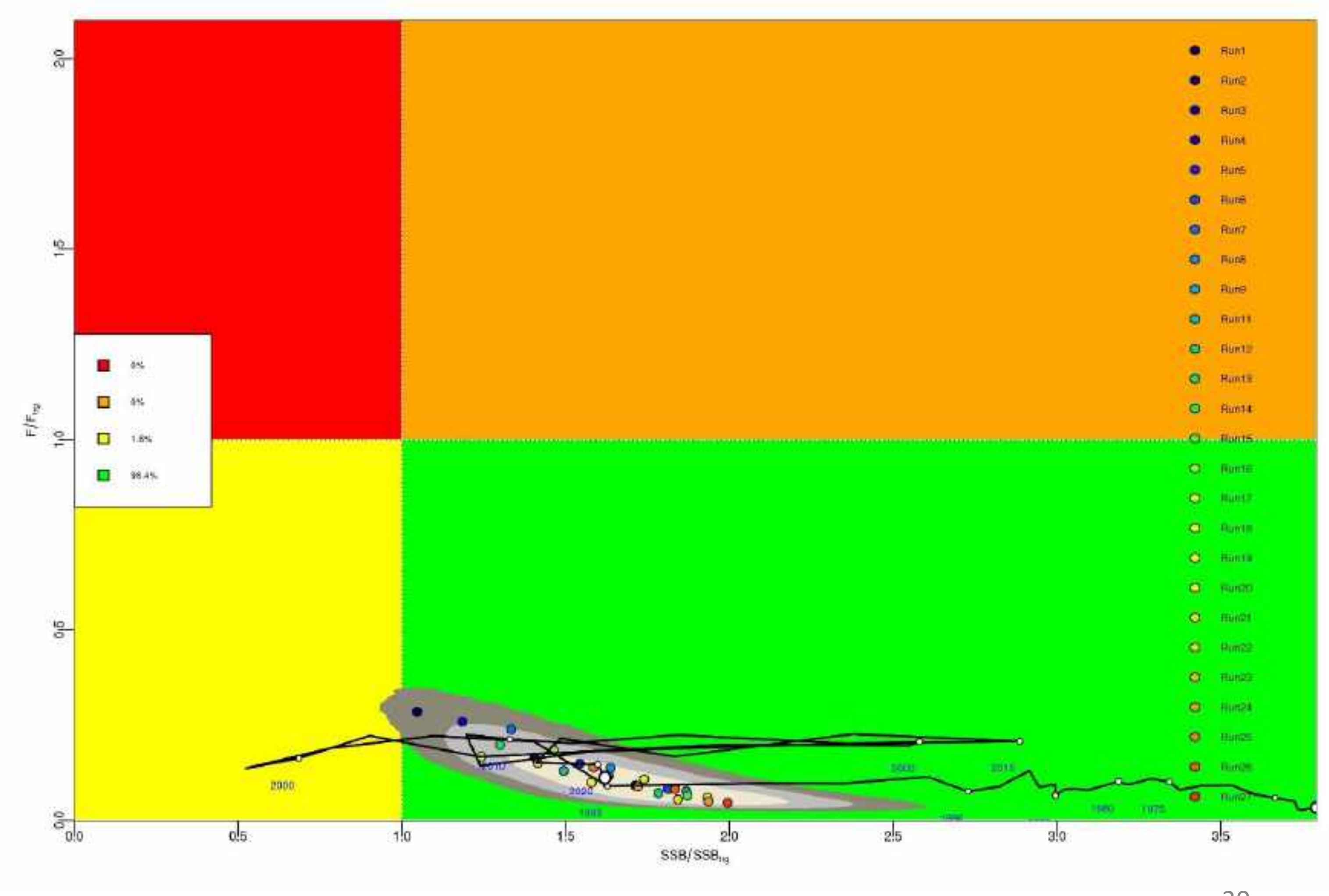
4. Prediction skills

- Hindcasting (CrossValidation)

		Goodness of fit							Consistency				Prediction Skill							
		Runs test					Joint residuals		Retrospective analysis				Hindcasting (MASE)							
Ensemble scenario	Convergence & stability	CPUE	Survey	Length catch	Length Seal	Length survey	Index	Length	SSB retro	SSB forecast	F retro	F forecast	CPUE	Survey	Combined	Length catch	Length Seal	Lengty Survey	Length combined	Weight
Run1	Passed	Passed	Passed	Passed	Passed	Passed	38.2	3.5	0.005	-0.133	0.122	0.274	0.689	1.605	1.062	0.497	0.400	0.243	0.469	0.83
Run2	Passed	Passed	Passed	Passed	Passed	Passed	36	3.4	0.138	0.096	-0.102	-0.088	0.977	0.704	0.866	0.506	0.218	0.224	0.423	0.94
Run3	Passed	Passed	Passed	Passed	Passed	Passed	34.8	3.4	0.135	0.096	-0.110	-0.087	0.982	0.692	0.864	0.500	0.233	0.252	0.424	0.94
Run4	Passed	Passed	Passed	Passed	Passed	Passed	37.6	3.5	0.092	-0.012	0.009	0.107	1.230	1.213	1.223	0.509	0.215	0.174	0.425	0.78
Run5	Passed	Passed	Passed	Passed	Passed	Passed	35.9	3.4	0.125	0.088	-0.097	-0.086	0.983	0.702	0.869	0.511	0.236	0.209	0.432	0.94
Run6	Passed	Passed	Passed	Passed	Passed	Passed	34.6	3.4	0.117	0.080	-0.101	-0.083	0.962	0.697	0.854	0.520	0.243	0.236	0.441	0.94
Run7	Passed	Passed	Passed	Passed	Passed	Passed	37.3	3.4	0.124	0.072	-0.078	-0.069	0.963	0.826	0.907	0.502	0.224	0.196	0.423	0.94
Run8	Passed	Passed	Passed	Passed	Passed	Passed	35.8	3.4	0.126	0.093	-0.100	-0.090	0.988	0.704	0.872	0.513	0.242	0.203	0.435	0.94
Run9	Passed	Passed	Passed	Passed	Passed	Passed	34.7	3.4	0.126	0.089	-0.106	-0.084	0.976	0.700	0.864	0.503	0.239	0.244	0.428	0.94
Run11	Passed	Passed	Passed	Passed	Passed	Passed	36.3	3.4	0.153	0.111	-0.110	-0.099	0.976	0.706	0.867	0.505	0.217	0.217	0.422	0.94
Run12	Passed	Passed	Passed	Passed	Passed	Passed	34.9	3.4	0.146	0.107	-0.116	-0.095	0.980	0.679	0.857	0.499	0.232	0.248	0.422	0.94
Run13	Passed	Passed	Passed	Passed	Passed	Passed	38	3.5	0.115	0.047	-0.064	-0.047	0.923	0.834	0.887	0.495	0.212	0.175	0.414	0.94
Run14	Passed	Passed	Passed	Passed	Passed	Passed	36.2	3.4	0.135	0.098	-0.102	-0.093	0.981	0.705	0.869	0.510	0.235	0.203	0.431	0.94
Run15	Passed	Passed	Passed	Passed	Passed	Passed	34.8	3.4	0.139	0.100	-0.113	-0.096	0.957	0.682	0.845	0.518	0.236	0.244	0.437	0.94
Run16	Passed	Passed	Passed	Passed	Passed	Passed	37.7	3.5	0.126	0.073	-0.080	-0.076	0.966	0.757	0.881	0.500	0.222	0.170	0.421	0.94
Run17	Passed	Passed	Passed	Passed	Passed	Passed	36.1	3.4	0.150	0.114	-0.110	-0.102	0.985	0.698	0.868	0.507	0.228	0.224	0.428	0.94
Run18	Passed	Passed	Passed	Passed	Passed	Passed	34.9	3.4	0.152	0.113	-0.120	-0.099	0.974	0.684	0.856	0.500	0.231	0.252	0.423	0.94
Run19	Passed	Passed	Passed	Passed	Passed	Passed	39	3.6	0.135	0.053	-0.061	-0.038	0.875	0.872	0.874	0.483	0.198	0.177	0.401	0.94
Run20	Passed	Passed	Passed	Passed	Passed	Passed	36.5	3.4	0.158	0.118	-0.113	-0.105	0.967	0.710	0.862	0.503	0.222	0.214	0.423	0.94
Run21	Passed	Passed	Passed	Passed	Passed	Passed	35	3.4	0.167	0.126	-0.128	-0.107	0.974	0.677	0.853	0.496	0.224	0.258	0.419	0.94
Run22	Passed	Passed	Passed	Passed	Passed	Passed	38.4	3.5	0.135	0.070	-0.073	-0.063	0.926	0.818	0.882	0.496	0.213	0.198	0.415	0.94
Run23	Passed	Passed	Passed	Passed	Passed	Passed	36.4	3.4	0.151	0.113	-0.109	-0.102	0.977	0.702	0.865	0.507	0.230	0.212	0.428	0.94
Run24	Passed	Passed	Passed	Passed	Passed	Passed	34.9	3.4	0.174	0.137	-0.129	-0.113	0.953	0.692	0.847	0.518	0.234	0.261	0.437	0.94
Run25	Passed	Passed	Passed	Passed	Passed	Passed	38.1	3.5	0.191	0.145	-0.093	-0.093	0.951	0.747	0.868	0.491	0.212	0.219	0.411	0.94
Run26	Passed	Passed	Passed	Passed	Passed	Passed	36.3	3.4	0.141	0.106	-0.105	-0.099	0.981	0.705	0.869	0.511	0.241	0.195	0.434	0.94
Run27	Passed	Passed	Passed	Passed	Passed	Passed	35	3.4	0.133	0.096	-0.110	-0.089	0.972	0.677	0.852	0.502	0.243	0.225	0.428	0.94

Ensemble model results

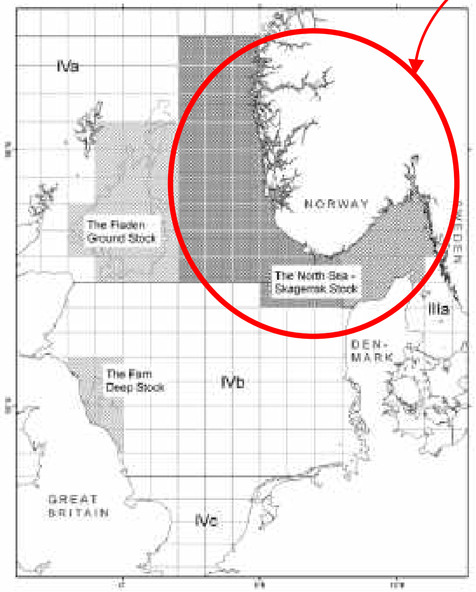
Name	Seals	Natural Mortality	Steepness	Weight
run1	low	M1 low	0.7	0.83
run2	low	M1 middle	0.7	0.94
run3	low	M1 high	0.7	0.94
run4	low	M1 low	0.8	0.78
run5	low	M1 middle	0.8	0.94
run6	low	M1 high	0.8	0.94
run7	low	M1 low	0.9	0.94
run8	low	M1 middle	0.9	0.94
run9	low	M1 high	0.9	0.94
run10	middle	M1 low	0.7	0.83
run11	middle	M1 middle	0.7	0.94
run12	middle	M1 high	0.7	0.94
run13	middle	M1 low	0.8	0.94
run14	middle	M1 middle	0.8	0.94
run15	middle	M1 high	0.8	0.94
run16	middle	M1 low	0.9	0.94
run17	middle	M1 middle	0.9	0.94
run18	middle	M1 high	0.9	0.94
run19	high	M1 low	0.7	0.94
run20	high	M1 middle	0.7	0.94
run21	high	M1 high	0.7	0.94
run22	high	M1 low	0.8	0.94
run23	high	M1 middle	0.8	0.94
run24	high	M1 high	0.8	0.94
run25	high	M1 low	0.9	0.94
run26	high	M1 middle	0.9	0.94
run27	high	M1 high	0.9	0.94

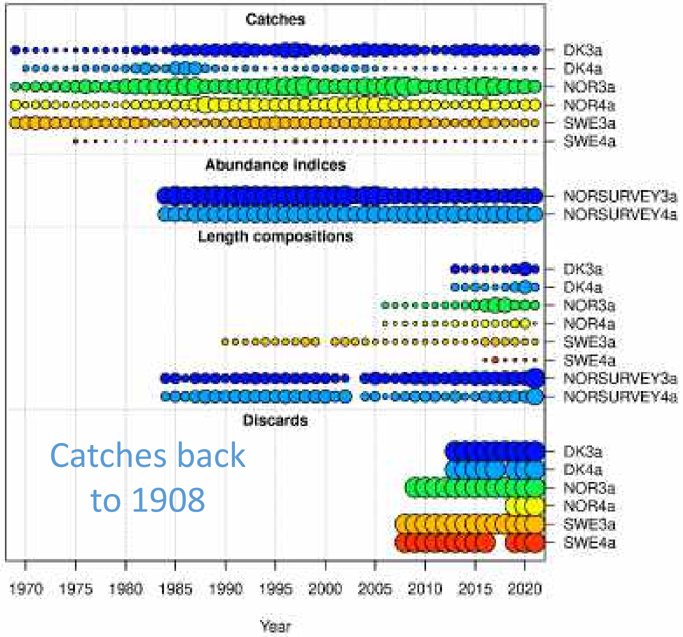


Northern shrimp

The 1st application of ensemble modelling in ICES stock assessment and advice

Two area/age-based using SS3





Catches back to 1908

ICES Advice on fishing opportunities, catch, and effort
Greater North Sea ecoregion
Published 09 May 2022

ICES CIEM

Northern shrimp (*Pandalus borealis*) in divisions 3.a and 4.a East (Skagerrak and Kattegat and northern North Sea in the Norwegian Deep)

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

Ancestor model: Single sec, single area without separated discards

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

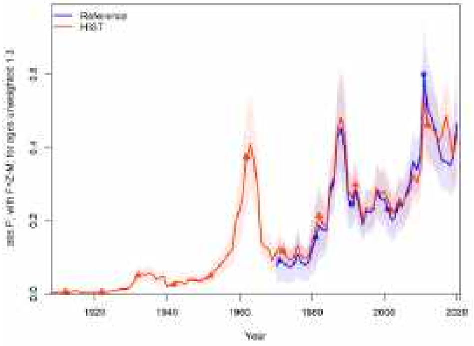
Set_1: Length based model based on VBGP and all fleets selectivity logistic discard retention fixed

Area specific sex ratio

Discarded: Worse fit of the LFD

Long landing time series (1908-2020)

Kept: same results, same fit



Short landing time series (1970-2020)

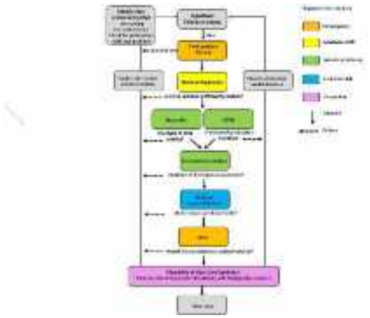
Discards retention time varying: improved fit, improved diagnostic

Survey index 2016 excluded

Hermaphroditism fixed

Discarded: Worse fit of the LFD

Iterative



Jittering and MCMC

Ss3diags
Pass/Fail

M1 (base, h=0.7)

M2 (low, h=0.89)

M3 (high, h=0.52)

Predation M (h=0.91)

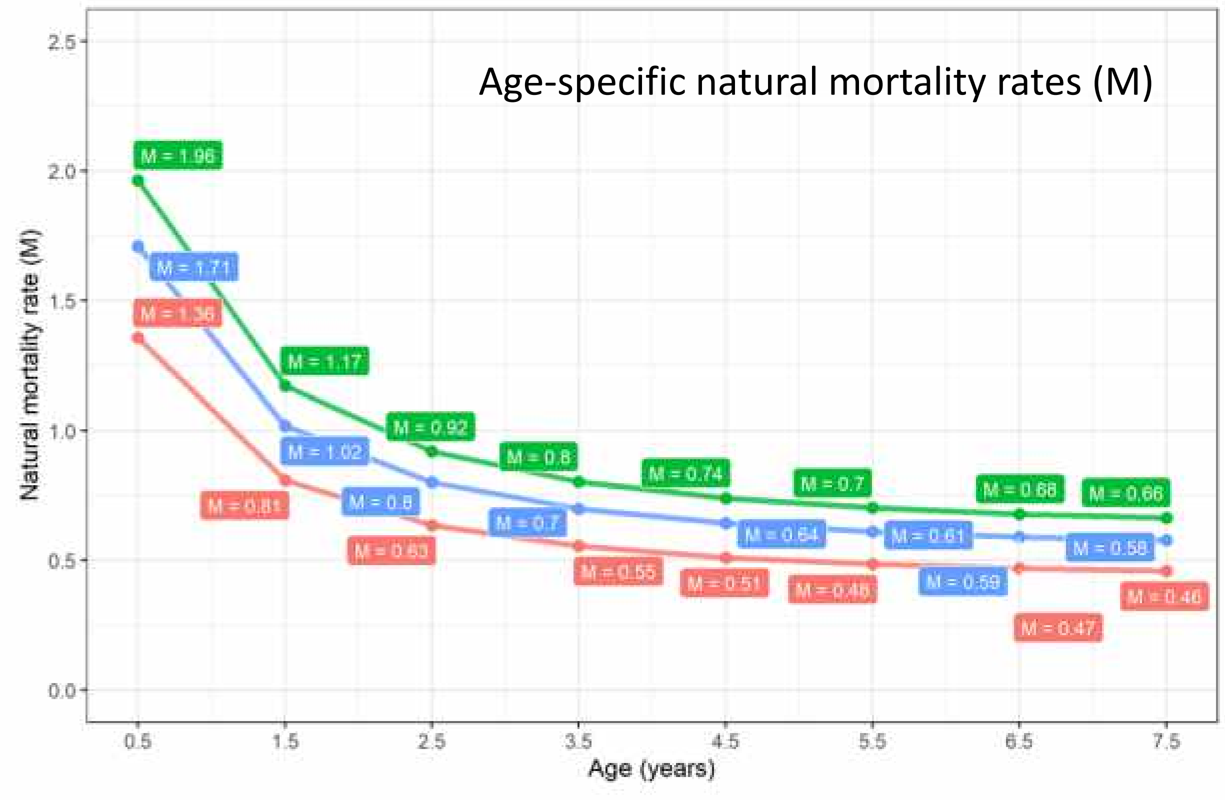
Weighting based on ss3diags

ENSEMBLE MODEL → **Forecast**

Why use an ensemble of M?

- Pandalus is an important prey species in the North Sea
- Eaten by a range of predators including cod and saithe (Jørgensen et al., 2014; Skorda, 2018)

- Despite this, we remain uncertain about the levels of natural mortality (M)
- Ensemble modelling allows us overcome this and incorporate 3 different but equally plausible scenarios for M (median, high and low) in our assessment of the stock



Weights

- As with vendace and sole, the model uses a weighted median of the 3 plausible scenarios based on ss3diags (Carvalho et al. 2021)

jabbamodel / ss3diags Public

A cookbook for using model diagnostics in integrated stock assessments

Felipe Carvalho, Henning Winker, Dean Courtney, Maia Kapur, Laurence Kell, Massimiliano Cardinale, Michael Schirripa, Toshihide Kitakado, Dawit Yemane, Kevin R. Piner, Mark N. Maunder, Ian Taylor, Chantel R. Wetzel, Kathryn Doering, Kelli F. Johnson, Richard D. Methot

- The weights are assigned based on 4 criteria:

1. Convergence and Stability
2. Goodness of fit
3. Consistency
4. Predictive skill

Positive Hessian/MCMC Jittering

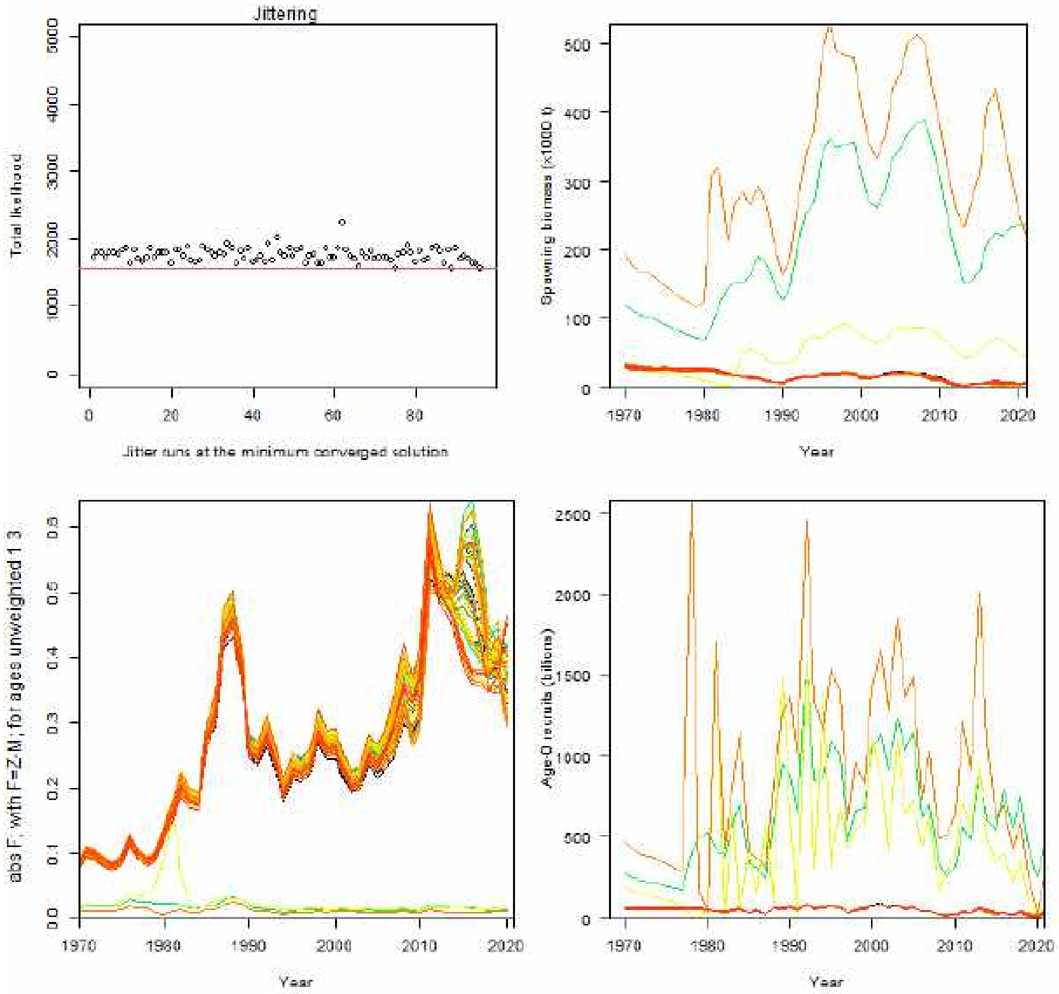
Joint residuals Runs test

Retrospective analysis

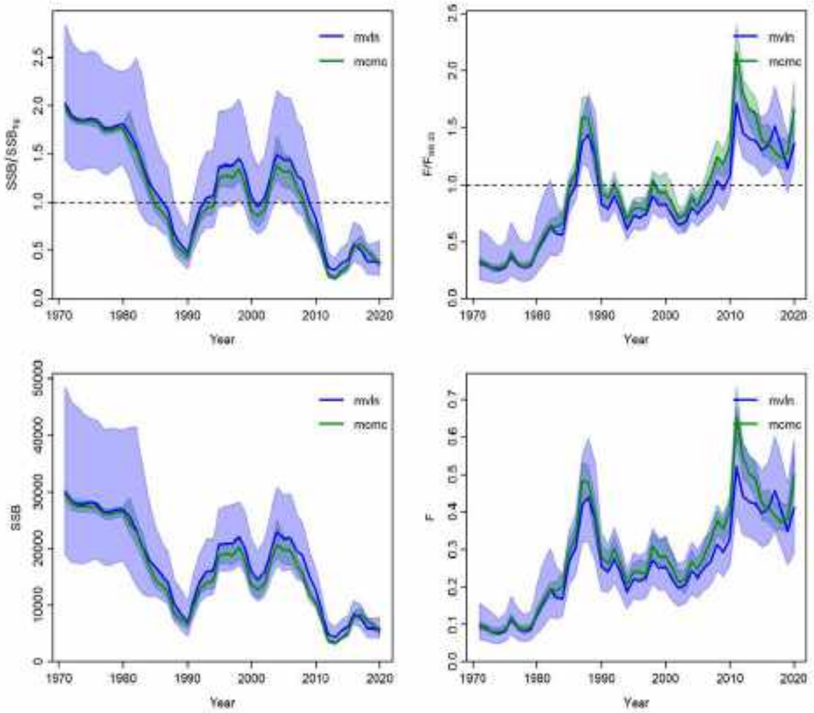
Hindcasting (Cross Validation)

Convergence and Stability

Jittering



MCMC



Positive Hessian/MCMC
Jittering

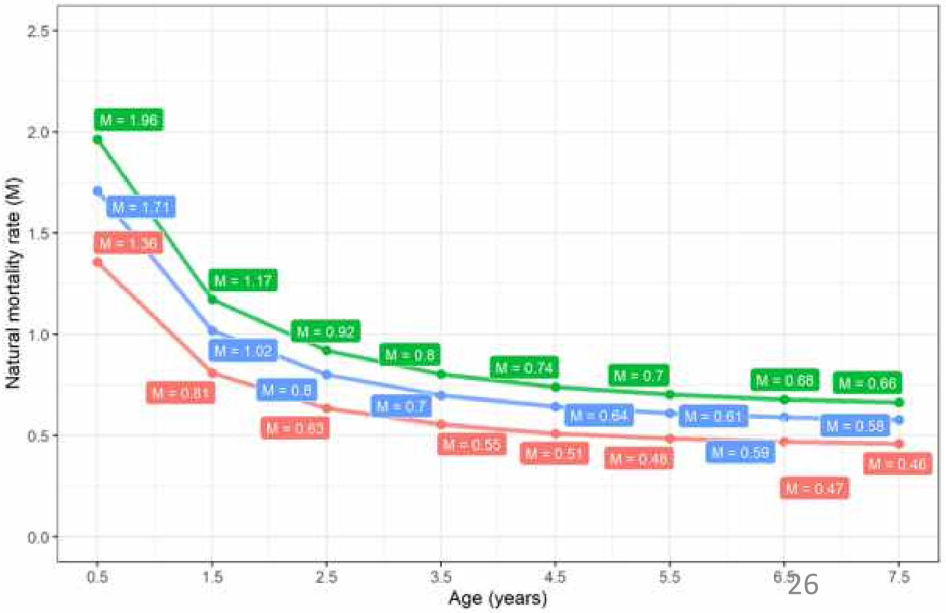
Model weights

Where Runs test and MASE were not aggregated in a single weight (plain) (no threshold)

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

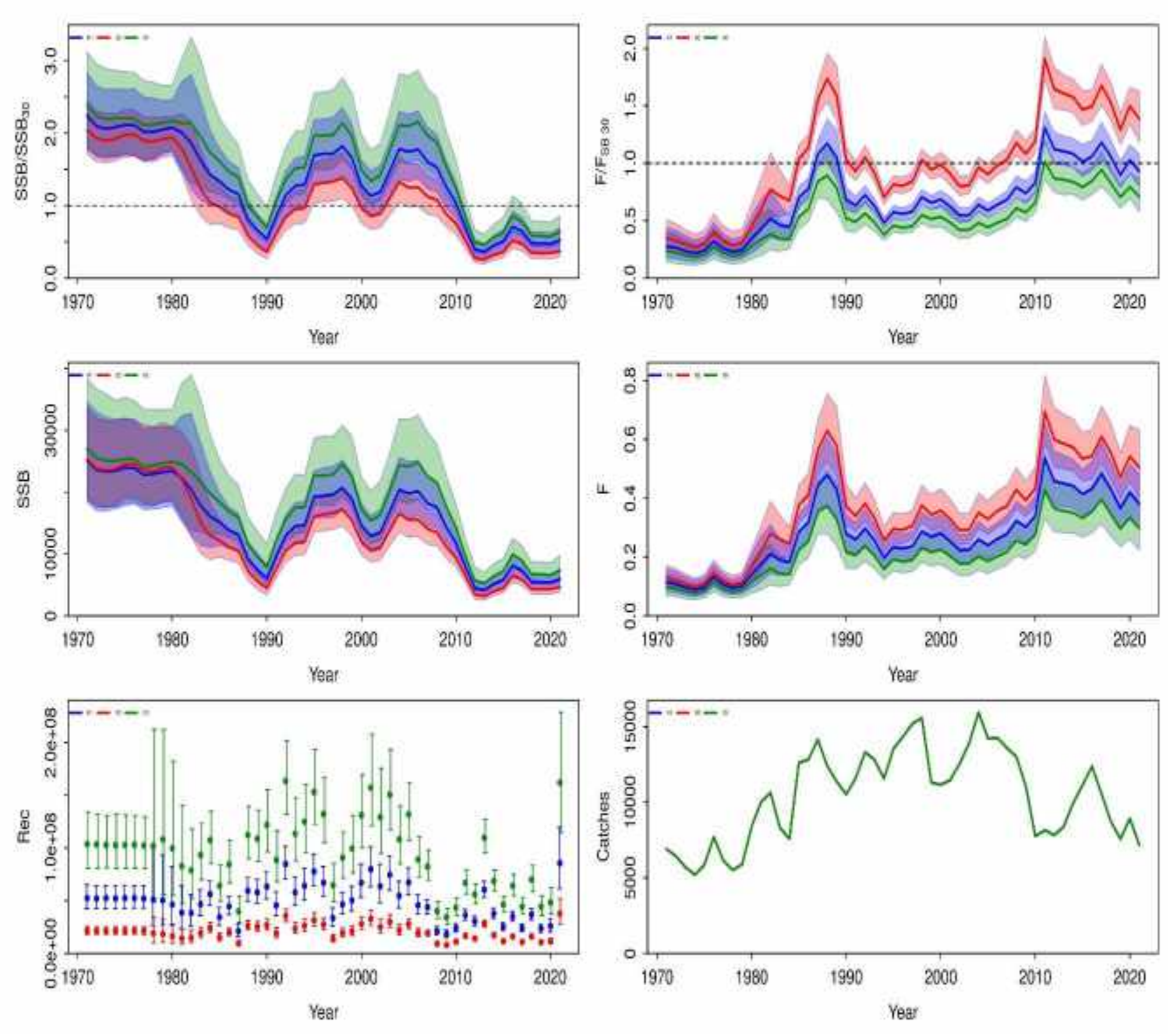
Prediction skills												
Hindcasting (MASE)												
Survey3a	Surve4a	Joint	Len3a	Len4a	lenA1S1	lenA1S2	lenA1S3	lenA1S4	lenA2S1	lenA2S2	lenA2S3	lenA2S4
Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed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

Model	Ensemble weight
R1 (blue)	0.78
R2 (red)	0.66
R3 (green)	0.72

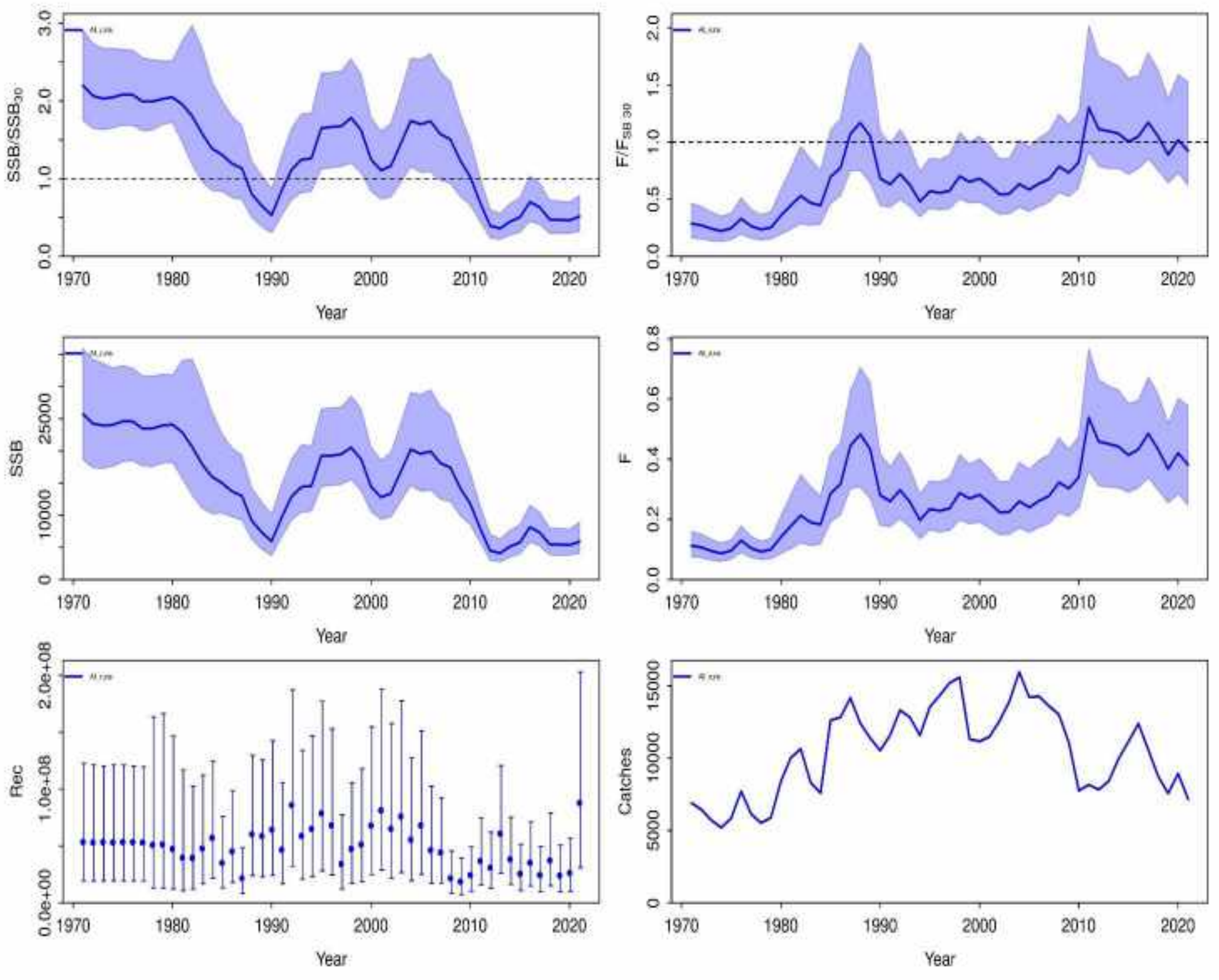


Results

Multi-model estimates

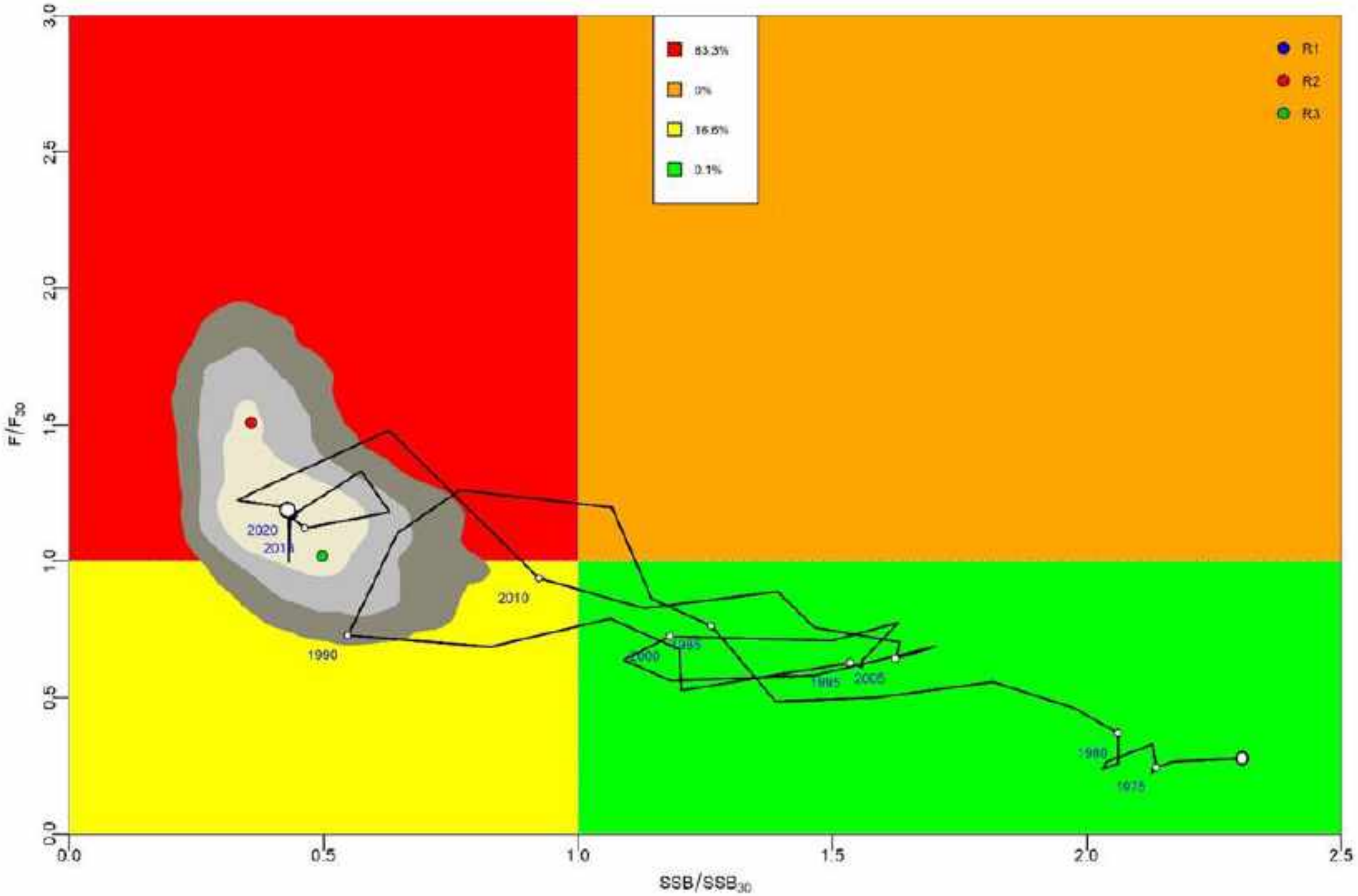


Ensemble model estimates

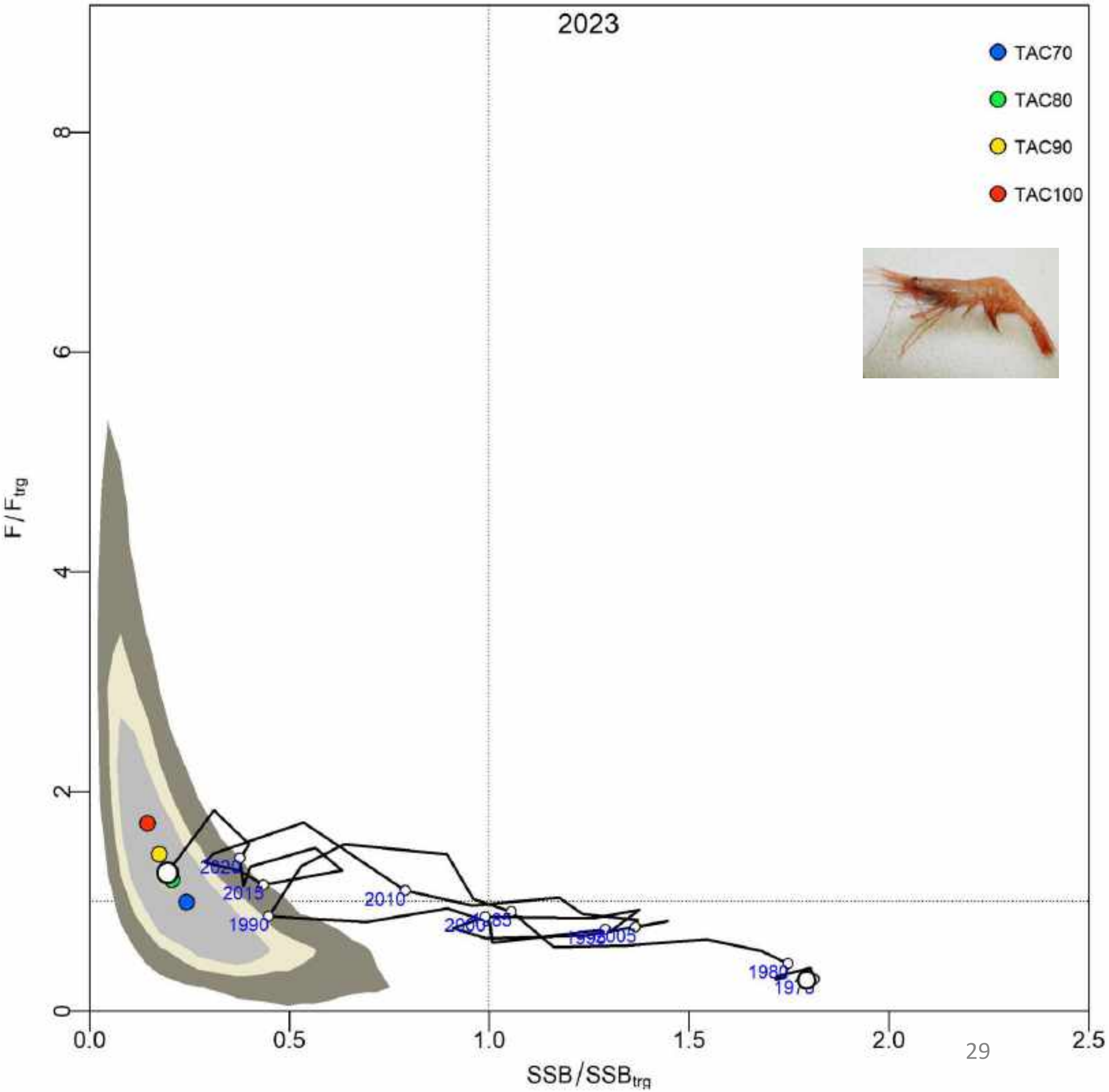


Results

- Healthy stock size being depleted by over fishing
- Ongoing overfishing, stock too small to produce MSY
- Biomass is small and still recovering, a reduction in F is needed
- Target area for management, sustainable F and healthy stock size



Predictions uncertainty



Why weighting

The entire process of stock assessment is pervaded by weighting

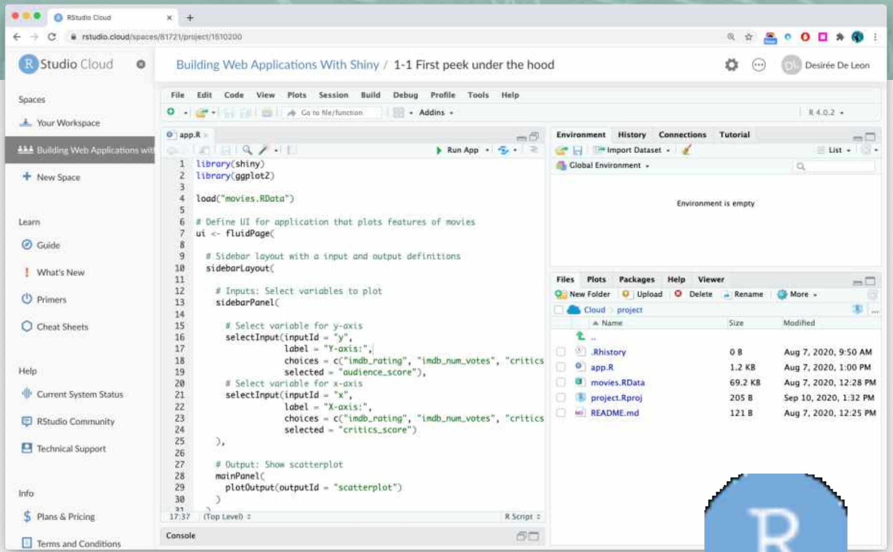
When weighting, the simple questions we asked to ourselves was: would we prefer a model that can predict the CPUE trend or a model the persistently overestimates the trend? Or a model that is retrospectively stable instead of one that is not?

From a “tactical” perspective, model weights are parameters to be chosen in such away as to achieve best predictive performance. No specific interpretation of the model is attached to the weights; they must only perform

As rarely models pass all diagnostics and model performances might change with time, we preferred to weight than to exclude or equally weight

Lessons learned

1. The development of a reference model is the key aspect of the ensemble process
2. A weighting scheme must be agreed beforehand to avoid cherry picking
3. A pass/fail system works well for stock assessment models for which diagnostics is used to weigh the single models in the ensemble
4. Differences in diagnostics performances between models is often small
5. Model stitching preserve the tails for a better quantification of risk
6. Combine diagnostics as for example MASE and run tests into a single value to create more balanced weighting scheme
7. Needs for simulations testing



app.R

https://framasnadi.shinyapps.io/AppSOL/

https://maxcardinale.shinyapps.io/Ensemble/

https://maxcardinale.shinyapps.io/Ensemble_vendace/