Single species status sub-group: meeting report

JRC, Ispra 3-5th June 2013

Attendees: 8

- Nicolas Gutierrez
- Kristin Kleisner
- Katie Longo
- Carolina Minte Vera
- Cóilín Minto
- Iago Mosqueira-Sanchez
- Giacomo Chato Osio
- James Thorson

Observers: 1

- Alessandro Orio (JRC)

Executive summary:

The purpose of this meeting was to bring together the stock simulation and method developer teams to discuss the performance of the methods on simulated data. Any remaining issues with the simulations and methods were to be finalized and models ranked prior to the Boston meeting. Issues were found with the implementation of Catch-MSY, Costello mimic and the state space catch only model (SSCOM). The issues with Catch-MSY and Costello mimic were resolved and diagnostics produced for SSCOM that will allow for its convergence issues to be resolved post-meeting. Conditional upon the preliminary results, model performance diagnostics and methods to rank the models were discussed and finalized. The structure of the presentations to be delivered in Boston was finalized with a set of tasks assigned to individuals with completion dates.

Discussion overview

The goal for was to bring the simulation and developer teams together to discuss:

- Simulation framework
- Model performance on simulated stocks:
    - Convergence
    - Bugs in the fits
    - Discussion of trends and visual patterns
- Performance diagnostics
- Plotting of diagnostics
- Analyses of diagnostics
- Stochastic implementation

Deterministic simulation framework:

An overview of the simulation framework was provided. Below is a synopsis of discussion on the simulation framework.

Fisheries Libraries in R (FLR) is used to simulate stocks using Gislason life-history trait relationships. A limited number of life histories (3) were implemented. These don't cover all possible life histories but are thought representative of many stock. Earlier discussions ruled out a reef fish species as the generating variables were thought close to the small pelagic category and to keep the number of combinations practical.

Flat-topped selectivity only is implemented but there is the option to do a sensitivity test on the effect of choice of the selectivity keeping in mind the dimensions of the full factorial.

1. Effort dynamics factor levels include: fixed F, F using Thor II effort dynamics with x=0.6 and two options implemented with a model other than ThorII, namely: one way trip (getting to 80% of Fcrash), and Roller Coaster:
    - The roller-coaster is designed to have: 5% increase rate of F up to Fcrash, a fixed 5yr of F=Fcrash level and then a 4% decline rate back to Fmsy
    - Chato will plot the F/Biomass/Catch timeseries for simulation presentation, to aid in interpretation
    - Jim brings up that the roller coaster potentially confounds life-history traits and effort dynamics, it may be worth implementing a second version of rollercoaster, RC2, where the 5 years are still fixed, but we fix the time it takes to ramp up to Fcrash and instead solve for the rate (as this may be a more informative option when we're testing the effect of truncated timeseries). This was implemented subsequently and replaced the earlier roller coaster implementation.
2. For stochastic runs, auto-correlation on recruitment residuals:

- Paper by Pyper and Peterman discusses AR(1) coefficient for salmon, 0.8 currently implemented seems an extreme case. Subsequently changed to 0.6.
- Currently the two autocorrelation factors implemented: no AR, with sigmaR = 0, the second option has autoregression 0.8 on residuals, but it also includes some white noise - this may confound noise and autoregression. Jim suggests a sigmaR=0.2 for the base case, so we're including white noise but not autoregression and thus can disentangle the two effects.
- Jim mentions catch process stochasticity in the case of fixed effort, eg due to a stock being bycatch and thus the result of some other stock's dynamics, it would be nice to have that too, but we must be clear on lack of time for changes.

The only change made to the full factorial design was to the internal implementation of the roller coaster. Carolina updated the factorial design table accordingly (Table 1).

Table 1: Updated settings for full factorial deterministic simulations.

Factor	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Initial depletion (**ID**)	0%	30%	60%	
Effort dynamics (**ED**)	Effort dynamics ED:0 – flat effort	Jim's effort dyn. model with a=1 (Bmsy) and x=0.6	One way trip. F increases to 80% Fcrash at the end of the series	RC Roller Coaster Going up in effort by 5% a year, stays 5 years at 80% Fcrash and subsequently decreases at 4%. RC2 has a specified amount of time going up and coming down with the rate of increase solved for. RC2 is to be used preferably.

Time-series length (**TS**)	20	60	
Life-history (**LH**)	Sardine	Haddock	Tuna
(Gislanson et al 2008 for LH invariants	Small pelagic	Demeral	Large pelagic

Deterministic model implementation

Convergence:

a. Costello et al is always fully converged (by design): should we classify it as 100% convergence rate or NA? Subsequently decided NA.

b. Thor II/SSCOM MCMC chains had some issue. Need to look at traceplots of MCMC chains. Coilin will prepare them by stock. Subsequently prepared and sent to Jim and amendments made to Thor II that improved convergence but still getting non-convergence in many stocks. Further discussion below.

c. COM strong convergence, COM-SIR converged 68 times out of 72, Catch-MSY 68 strong, 2 weak convergence.

d. Jim suggests to focus post-hoc attention on non-convergence cases.

e. For the SS-COM it may be possible to do extra runs, but ding the method in a "ease of use" category when constructing our limitation/applicability appraisal

Following convergence discussion, for each model we looked at the panel time series plot of true and predicted B/Bmsy with uncertainty bounds. From these we qualitatively discussed if there were any obvious bugs in the implementation and also what were the most pertinent trends were by eye. These discussions are summarized below.

Costello mimic

- The second version of Costello et al included a propensity score as an additional covariate (random forest of RAM vs FAO) with main effects and an interaction allowed (using all the covariates that could be used in Costello et al. mimic)
    - is there a bug? ED:OW,TS:60,ID:1,LH:LP has an unexplainable peak. Subsequently tracked down to an erroneous step in catch within that series. Now fixed in the simulations.

    - What drives discrepancies in ID0.4,LH=LP,ED:0.6,TS:20.?

- How important is the role of absolute catch size in the Costello method? Because the initial unfished stock size (B0) was established arbitrarily as 1000tons, i.e. small for commercial catches, for all stocks so this may drive the fact that they appear to be "un-Rammy", but also and more importantly, may drive the base model's performance

- (Gislason models r0, hence B0, as a negative correlation with Lmax – this may be simply something to mention)

- we can simulate a few stocks with high initial biomass to test whether this has a significant effect

- regardless of initial depletion, Costello appears to give the same answer – and tends to perform best in the case of depleted stocks, with particularly good performance for large pelagics -> based on lifehistory it assumes a fixed initial depletion

Discussion section in our write-up:

- do the limitations of the Costello method as implemented here change in a "real world" case, where the set of covariates would be bigger?

- whether the options to set priors in the "real world" would be different, it needs to be explicitly explained how robust are the simulation outputs

Catch-only method (COM)

- No obvious signs of bugs

- increasing confidence bounds, not a good indicator of initial uncertainty

- too narrow confidence bounds in presence of effort dynamics (appropriate increased confidence because this is the situation, by design, where it should best perform based on its assumptions) – too narrow, but predictions pretty good

State space catch only method (SSCOM)

- Convergence was poor with only 4 stocks having strong convergence

- Initial check of whether the convergence rate has influenced the performance:

    - We spot-checked the models with high convergence, 3 out of the 4 we checked happen to have very narrow confidence bounds (2 of these had the effort dynamics, so that may be for the same reason as the COM, but the third was a one way trip) – this would suggest its' not the fault of low convergence rate

- The chains were designed so that as a ballpark there was a minimum effective sample size of 30 per parameter (but actually, the number of parameters varies, so this ballpark estimate may not be accurate in this case?)

- Need to look at trace plots to draw conclusions (Coilin needs some time to produce those). Subsequently produced and provided to Jim.

- Jim suggests to drop from the methods comparison, the cases where effective sample size is too small, even if the model is performing well

- Hypotheses on the priors: the parameters aren't mixing, the mean draw from the prior looks correct

- Discussion is shelved for now, until we have the trace-plots

Subsequent changes implemented for SSCOM include:

- an upper limit of 1 for the harvest ratio, as previous solutions had greater than 1 solution with very low/no biomass.

- Importance re-sampling for initialization

Catch-MSY

- two of the curves appear to have no confidence intervals. Also some of the B/Bmsy values seem way too high – there appears to be a bug

- Subsequently this bug was tracked down to the omission of initial depletion. Code was amended and Catch-MSY run overnight, new runs are much more in line with other models and seem more realistic.

Deterministic model performance

Performance statistics/diagnostics

We had extensive discussions on model performance, which will subsequently be used for model ranking. A set of performance statistics were decided upon:

1. Proportional error ((true-predicted)/true). If a manager is looking for indicative measures, but is prepared to not adhere strictly to the model's own predictions of confidence intervals.

2. Posterior predictive score. This integrates the model's confidence interval and mean and is essentially the density of the model prediction at the true value (large is good; small is bad). Code was worked up to extract this.

3. Coverage – the number of times the true value was within the model 95% CI.

4. Number of times model is negatively or positively biased.

These performance statistics are calculated over:

- Entire time series
- Last five years
- Last 10 years
- Previous 10 to 5 years

The last two year sets are designed to capture periodicity in performance that could otherwise be missed.

Code to take the output from the fits and calculate these diagnostics was developed.

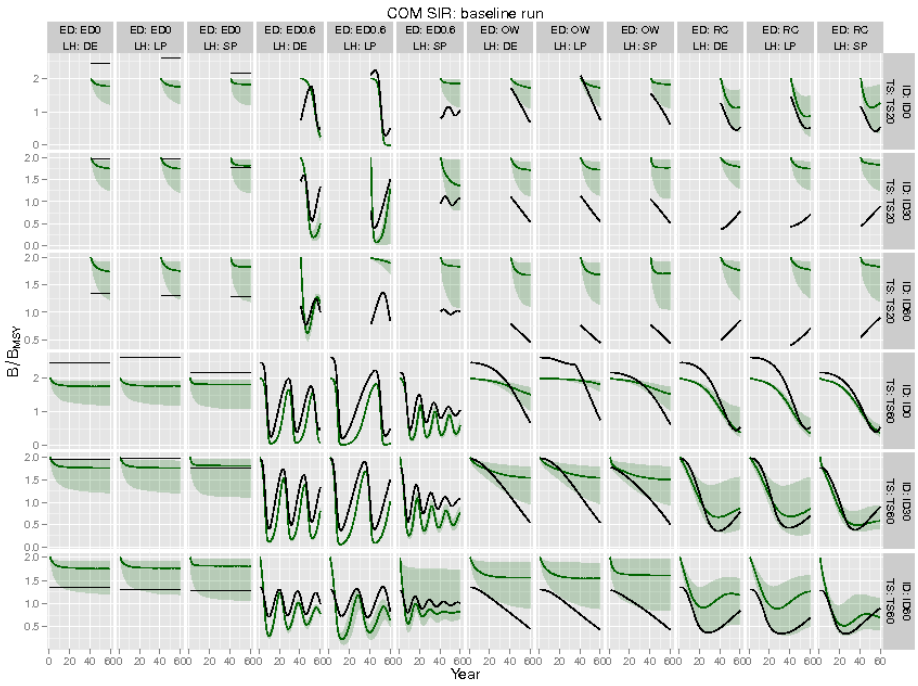
Using all these performance criterion over the different time-spans provides a potentially bewildering set of results. The team will choose a single metric and time-scale to report – likely proportional error over last 5 years – for the main results to ease interpretation.

[Each developer will also insert considerations on robustness of conclusions in real world case, based on simulated stocks limitations.]

Plotting/presenting results

The set of plots (examples below, note not final), which were coded up before and at the meeting include:

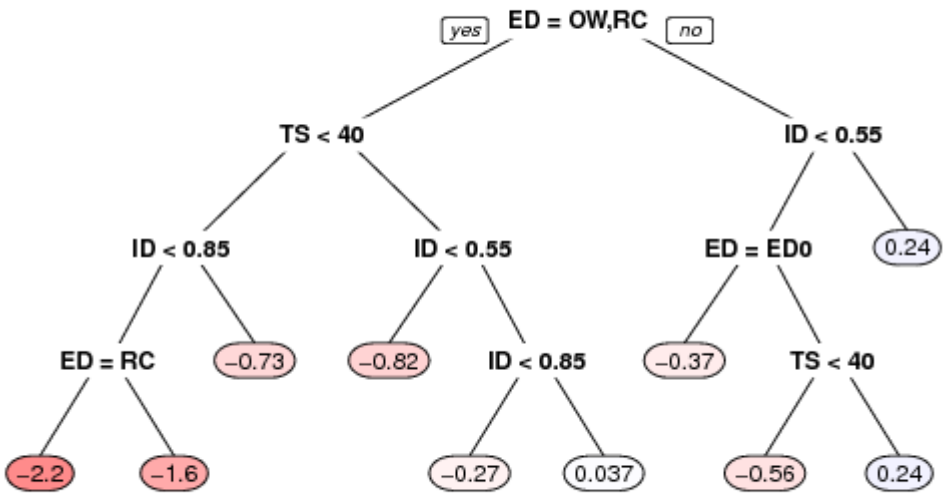
1) Time series performance plot (by method)



Discussed what is the most useful ordering of factors in the plot:

- order the columns by effort dynamics level, with timeseries matched side by side for the same ED factor level
- order the rows so that the first row has no depletion (ID=1) and progressively higher depletion, with LH fixed, then the 30% depletion level (ID=0.7)
- order the LH by response rate from “slow” to “fast” (where small pel is “fast” and dem is “intermediate”)

2) Regression trees (by method). Example is for proportional error in last 5 years. (Note results not final)



3) Matrix plot of best performing model per simulation condition (Note results not final)

A mosaic plot alternative was also coded up to include the second best, third best etc. With the area proportional to a measure of performance. This plot looked quite busy but the concept is returned to below.

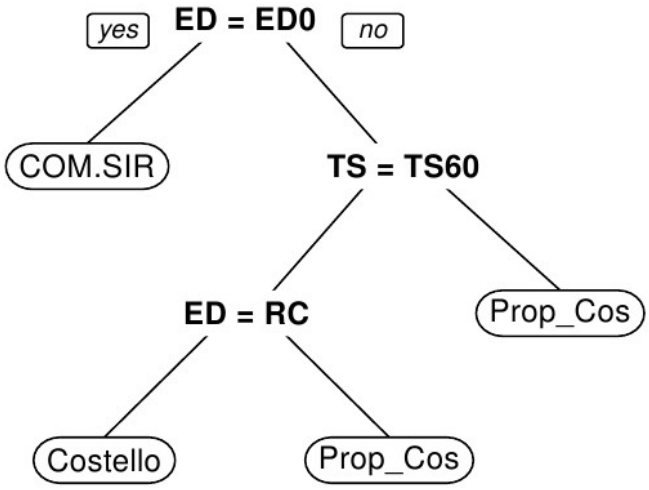
4) Tabulation of frequency of best performance under the various performance statistics, e.g. (Note results illustrative only – need to be calculated based on final runs)

Model	Proportional error	Posterior predictive score	Coverage
Costello mimic	43	10	.
Catch-MSY	20	2	.
COM	1	30	.
SSCOM	8	30	.

This provides a method for ranking performance and a comparison of performance ranking across criterion.

5) Classification/decision trees

The best performing model in each case is used as the response and the simulation variables as the explanatory variables in a classification/decision tree, e.g. (Note results not final)



Boxplots of the performance criterion (e.g., proportional error) could be placed at the terminal nodes, which would indicate how the other models performed, i.e. not just the best model.

Stochastic simulations

Extensive discussions and implementation coding was undertaken for the stochastic simulations during the Ispra meeting. At the Santa Barbara meeting, it was agreed that the stochastic runs could not be implemented for the full factorial because of time and computing limitations and that conditional on the results of the deterministic runs a select base case should be selected for stochastic implementations. This approach is adhered to at the Ispra meeting with a base case selected. Given that we also have a significant amount of additional computing resources on Hexagon, we have also included additional batches that can be run subsequent to the baseline case. The rational for these additional simulations are:

- There were no clear variables that should be excluded from the deterministic runs. Life history did not appear very important but it was felt that the addition of stochasticity could interact with life history to produce an effect.
- To allow for a proper comparison of deterministic versus stochastic performance.
- For the method appraisal to be taken seriously it needs to be applied to realistic looking data, which necessitates different levels of stochasticity.

Stochasticity enters on:

- Recruitment variability (two levels: sigmaR=0.2, sigmaR=0.6)
- Catch error (two levels: sigmaC=0; sigmaC=0.3)

Autoregressive process error on recruitment variability enters at two levels (AR(1)=0; AR(1)=0.6).

The final set up for the stochastic runs is shown in Table 2. The design of this setup is that batches will be run on hexagon with primacy given to the first batch, which is what was agreed in Santa Barbara.

The timeline for completing the stochastic runs is provided below.

Limitations/caveats/applicability

We discussed the limitations of the methods, which will focus on:

- Appraising the performance relative to the truth – we can rank the models but how well are they doing relative to the truth, e.g. COM SIR might perform the best but do we still expect the proportional error or CV on B/Bmsy to be very large?
- Caveats associated with best performers – used for monitoring versus management. Pre-cautionary – what side do they come down on? If used for management should be management strategy tested.
- Applicability: how much know-how needed to implement – compare with next best performing method.

Outreach

- Code repository
- Simulation design paper – led by Chato and Iago
- Method performance paper – led by Coilin with rest of team

Task completion: ownership and timeline

Table 3 contains outstanding tasks and a timeline for completion in preparation for presentation at the Boston meeting.

Please note that we request 2 presentation slots at the Boston meeting:

1) Simulation design

2) Method performance (presentation outlined below)

- Slide 1: Goal and questions

- Slides 2-5: Methods overview
- Slide 6: Diagnostic/performance criterion/statistics defined
- Slide 7: Deterministic results title
- Slides 8-11: Time series of true to estimated by method
- Slides 12-15: Regression trees by method
- Slide 16: Matrix plot of best performer per category
- Slide 17: Tabulation of best performer by criterion
- Slide 18-20: Decision trees by criterion
- Slide 21: Stochastic results title
- Slides 21-31: Stochastic results (similar to above only using average across iterations)
- Slides 32-38: Limitations/caveats/applicability
- Slide 39: Outreach
- Slide 40: Acknowledgements

Table 2: Stochastic run setup. Batch 1 is the baseline setup.

Batch	FACTOR	LH	ID	ED	TS	AR	SigmaC	SigmaR	Total treatments	replicas	models on hexagon	total runs	total hours
1	N. of levels	1	1	4	1	2	2	2	32	10	3	960	9600
	Levels	SP	0.7	all	60								
2	N. of levels	1	2	4	1	2	2	2	64	10	3	1920	19200
	Levels	SP	0.4; 1	all	60								
3	N. of levels	1	1	4	1	2	2	2	32	10	3	960	9600
	Levels	LP	0.7	all	20								
4	N. of levels	1	2	4	1	2	2	2	64	10	3	1920	19200
	Levels	LP	0.4; 1	all	20								
5	N. of levels	1	1	4	1	2	2	2	32	10	3	960	9600
5	Levels	DE	0.7	all	60								
6	N. of levels	1	2	4	1	2	2	2	64	10	3	1920	19200
	Levels	DE	0.4; 1	all	60								
7	N. of levels	1	1	4	1	2	2	2	32	10	3	960	9600
	Levels	DE	0.7	all	20								
8	N. of levels	1	2	4	1	2	2	2	64	10	3	1920	19200
	Levels	DE	0.4; 1	all	20								

9	N. of levels	1	1	4	1	2	2	2	32	10	3	960	9600
	Levels	LP	0.7	all	60								
10	N. of levels	1	2	4	1	2	2	2	64	10	3	1920	19200
	Levels	LP	0.4; 1	all	60								
11	N. of levels	1	1	4	1	2	2	2	32	10	3	960	9600
	Levels	SP	0.7	all	20								
12	N. of levels	1	2	4	1	2	2	2	64	10	3	1920	19200
	Levels	SP	0.4; 1	all	20								
									576	120	36	17280	172800

Task Area	**Specific Taks**	**Person**	**Timeline**
Model Running	Run final models on deterministic full factorial	Coilin-Chato	Friday 7 June
	Run Costello on full factorial stochastic 10 replicates per stock	Chato	Friday 7 June
	Run batch 1 of stochastic on Hexagon	Coilin	Thurs 13 June
	Run batch 2-12 of stochastic on Hexagon	Coilin	Mon 17 June
	Export/email results full factorial runs	Chato	Friday 7 June
	Computation posterior predictive scores	Carolina Katie	Tues 12 June
	Run regression tree on results from Costello stochastic runs (code from Coilin)	Carolina Jim	Friday 14
Plotting	Produce simulation Design Plots	Chato	Monday 10
	Performance plots deterministic full factorial	Chato	Monday 10
	Regression Tree deterministic full factorial (Tigli plots)	Coilin-Andy C	Tuesday 11
	Mosaic Plot Full Factorial	Kristin + Alessandro	Friday 14
	Decision Trees Plots-Terminal Node Boxplot (aka gnocchi plot)	Jim Carolina Coilin	Friday 14

Other	Table of frequencies of best performance Summary of the meeting	Katie Carolina All (Coilin Katie)	Friday 14 Mon 10 June
Stock Simulation	Finalize the simulations with latest modification	Iago	Friday 7 June
Presentation	Outline with Plots Final Presentation	Chato & Coilin All	Monday 17 Wednesday 19