Inconsistent strategies to spin up models in CMIP5: implications for ocean biogeochemical model performance assessment
Séférian et al

This study examine spin-up and drift in ocean biogeochemical properties using a spin-up run from a single model and archived model output from CMIP5. In particular the study demonstrates the need to take drift into account when assessing model skill. I think that this is a useful study that highlights an important issue that is probably not given enough attention.
I do have some issues with the analysis undertaken however that need to be addressed or clarified before I think this manuscript is ready for publication.

8754:
However, since these models are typically initialized from observations, initialization and equilibration of climate variables are the most model-dependent protocols that could introduce errors or drifts in modeled fields with consequences on skill score metrics.
By 'equilibration' do you mean spin-up procedure. Sentence isn't very clear

8755:
First paragraph
There is an assumption here that the model will reach an equilibrium. This is not clear. Sen Gupta et al 2013 show little evidence for equilibration in many physical variables. Work by Will Hobbs and collaborators (soon to be published) shows that a large component of drift in physical variables is associated with spurious energy leaks in the models that are independent of model state. As such the models just keep drifting. Indeed in your Fig 2b I don't really see clear evidence for equilibration.

'quasi-equilibrium state is assumed for the interior ocean tracers.'
I don't think its assumed its either corrected for or neglected.

8757:
It ranges from 1500 to 4000 years depending on the ocean circulation and can reach up to 10 000 years in the deeper domains of the ocean
Doesnt really make sense to give a range of 1500 to 4000 and then say some regions are 10,000. That means the range is 1500 to 10,000.

8759:
Gupta et al. (2012, 2013).
Should be 'Sen Gupta et al'

8763 last paragraph:
The metrics (2–4) are not very well defined can you be more precise? Does 2 mean you calculate the difference between model and obs at each grid point and then average? Is 3 just the spatial correlation between model and observations. 4 I dont really understand. Is this the difference between the spatial standard deviation for model and obs?

Figure 1:
In Fig 1 I think that the direction of the cross hatching for initial conditions

are the opposite way round for ‘model’ and ‘mixed’ in the figure and the legend.

8766:

except some recommendations for the decadal prediction exercise …

I presume however that there was no simulation of BGC in the decadal prediction simulations

8767:

Figure 2b also shows that the drift in the global sea-to-air carbon flux reduces slowly after the first 50 years of the spin-up simulation. While this drift is about 0.001 PgCyr-2 from year 250 to 500, it is much weaker over the last century of the simulation (5_10-4 PgCy-2)

The drift looks pretty linear after about year 50. Are the differences you discuss really significant? For example, if you shifted your analysis 50 years earlier i.e using 150 to 450 do you get robust results?

the simulated sea-to-air carbon flux would reach a steady state after ~500 supplemental years of spin-up.

Im a bit confused. By steady state do you mean when the air-sea flux is zero? But this isnt steady state. Steady state is when dF/dt=0, which will never happen under an exponential model, which is why you have a decay timescale.

Also your time estimates seem strange. If the decay timescale was only 73 years we would expect to see a large slowdown in drift over the course of the experiment, whereas it looks pretty linear. Also, if the trend at the end of the control is 5e-4, and the carbon flux is just less than -0.5PgCy-1 it would take almost 1000years to reach 0 and a further 950 years to reach 0.45. This is without any further reduction in the rate of the drift. Am I missing something?

8770:

… over the last century of spin-up …

Is 100 years really sufficient to get a good estimate? While you need to remove the period of initial coupling shock, this seems to only affect the first 100yrs or so in Fig 2.

These decay timescales seem very short. The tracers dont look like they would reach equilibrium on O[50yr] timescales. Indeed given that there is still substantial drift at the end of the 500yr control, when you exclude the initial coupling shock the timescale for reaching steady conditions look to be much longer.

I would like to see more detail on how you are fitting your drift model as it seems something is going wrong.

…across depth over the first century of simulation for each ESM …

Given that the minimum control is 250yrs I dont see why you would only consider 100ys to obtain your drift estimate. The shorter the time period the more likely it is that you are aliasing low frequency natural variability. Indeed you are assuming that the drift follows an exponential model so why wouldn’t you use the full control run to estimate the decay timescale?

At the very least I would like to see error bars on the drift estimates based on the rest of the control runs (the full period should be subject to the same drift timescale, if your model is appropriate)

8771:
… between the drift in RMSE and the spin-up duration.
The relationship is with the log of the spin up time

fall outside the 90% …
Do you mean 'below' not outside

This low significance level must be put into perspective given the large diversity of spin-up protocols and initial conditions (Fig. 1 and Table 1) that can deteriorate the drift-spin up duration relationship in this ensemble of models.
In addition you are unlikely to find the same drift rates in different models anyway

extrapolated over the 250–1190 spin-up duration range
This is a massive extrapolation. I would like to see the raw data this is based on displayed on the graph as I suspect the drift estimates from the 100yr chunks are very noisy

You might also consider doing this analysis for all depths (and plotting R vs depth) to see how robust the relationship is, although I appreciate that this might be a big task given all the data required

8773:
We employ ÉRMSE to penalize the normalized distance …
Im not really clear what has been done here. Is the following correct?
1. You have taken the RMSE for the mean 1985–2005 historical period relative to available observations
2. You then calculate the drift timescale for each model based on the first 100yrs of picontrol
3. You then calculate the additional RMSE you would expect for a further 3000 years worth of integration and add it to the original RMSE.

If so, some problems I see with this:
1. It assumes that 100yrs from the picontrol is sufficient to get an accurate estimate of the drift.
2. It assumes that the drift at the start of the control is representative of the 1985–2000 period. This depends on when the historical simulation was branched off the control.

(i.e., CMCC-CESM, IPSL-CM5B-LR, NorESM1-ME, CNRM-CM5)
what about the GFDL ESM2M?

8774:
… errors in ocean biogeochemical fields amplify and propagate…
not sure what you mean by propogate in this context

Mignot et al. (2013) with the same model simulation showed that the large-scale ocean circulation reaches quasi-equilibrium after 250 years of spin-up, but our analyzes

indicate that biogeochemical tracers do not …
But all the characteristic timescales you have calculated are <150yrs. This does not match with your assertions of long equilibrium times

8777: that have drifted further away from their initial states …
This doesnt seem to be true always. Examination of Fig 3 shows that in many cases the initial coupling shock is in the opposite direction to the long term drift. Eg in 3e, NO3 is almost back to its initial state after the spin up period

Swart and Fyfe (2011)
Im not sure about the relevance of this study here – please explain

8778:
One issue is that the penalization relates to what the model state will look like around the time of full equilibration. However the transient (historical/RCP) runs are potentially done when the model state is closer to the initial observed state than the final equilibrium state. As such the transient response to greenhouse forcing may be more correct (even if the model is going to keep drifting). In the end the scores are there to help identify the models that produce the most realistic projections

The low confidence level of the fit to drift …
Where in your analysis do you demonstrate this low confidence?

The impact of this penalization approach on model ranking calls for the consideration of spin-up and initialization strategies in the determination of skill assessment metrics…
I dont follow this. Your penalisation process doesnt involve the spin up. It just requires an estimate of the drift which is estimated by looking at the control simulation.
However I agree that it would be very useful to have more spin up information (including the spin up run output) as part of the available archive

8779:
CMIP7 …
What happened to CMIP6?

agree on a set of recommendations for initialization, spin-up protocols and duration
Im not sure that it makes sense to have a common duration as different models drift at different rates