Dear Editor of *Geoscientific Model Development,*

We modified the manuscript according to the comments and queries from the three reviewers. To fulfill their requests, we

1) Clarified several sections of the main text by focusing our analysis on the root-mean squared error (RMSE) and removing information on the difference between modeled and observed spatial variability ($\Delta\sigma$). We included supplementary figures to several sections in order to support our hypotheses and strengthen our analysis.

2) Included one additional IPCC-class Earth system model (ESM) to our analyses. This extends the current analysis to an ensemble of 15 IPCC-class ESMs. Accordingly, we have revised all computations and updated all figures of the manuscript.

3) Added two new subsections to the revised manuscript to satisfy the request to include further discussion on the limitation of our framework. These are a new subsection dealing with the assessment of the simple drift model with IPSL-CM5A-LR as well as an extended discussion of the limitations of our framework.

Finally, we also introduce two co-authors (O. Aumont and A. Romanou) for their expertise on their model and their help to address the comments and queries from the three reviewers.

Please find a detailed response to each question/comment hereafter in blue (text fragments are *in blue italics*).

 # Editor

The column for model INMCM4 is shown in a colour (red) that does not appear in the key to the figure. Presumably this colour is a simple error as the text does not identify INMCM4 as undertaking a unique spin-up strategy.

We thank the Editor for his comment. It was indeed an error. Figure 1 was updated to improve its readability (bar plot enlarged and cross hatching modified). An additional model (CNRM-ESM1) was added to the analysis. The model is detailed in Séférian et al. (2015).

Referee #1

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

This study examines 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.

We appreciate the reviewer's careful reading and suggestions for corrections. Most of his suggestions are addressed in the revised manuscript.

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

We agree that the use of "equilibration" might be confusing.

We consider the spin-up procedure to encompass: (1) the initial conditions, (2) the spin-up time (duration of the spin-up simulation at the end of which a "quasi-steady-state equilibrium" is declared) and (3) the method to achieve this quasi-steady-state equilibrium (it can consist in an offline mode simulation or a coupled mode simulation — called online in the manuscript).

We modified the sentence as follows:

Revised:

"However, since these models are typically initialized from observations, the spin-up procedure of climate variables are the most model-dependent protocols that could introduce errors or drifts in modeled fields with consequences on skill score metrics."

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.

We agree with the reviewer that "reaching an equilibrium" is a strong assumption. It is, however, mentioned in several published papers:

In (Dunne et al., 2013):

"The fully coupled models were then integrated over 1000 model years with 1860 solar and radiative forcing before declaring ''quasi-steady-state equilibrium'' and beginning the 1860 control and perturbation integrations. In addition to the many qualitative requirements, we define acceptable quasi-steady-state equilibrium with quantitative metrics: net top-of-the-atmosphere radiative fluxes less than 0.5 W m-2, surface temperature drifts less than 0.18C century-1, stable Atlantic meridional overturning circulation (AMOC; Delworth et al. 1993) above 10 Sv (1 Sv [106 m3 s-1), local sea surface tem- perature (SST) biases less than ;98C, global 708S–708N root-mean-square SST errors less than 1.98C, and global net CO2 fluxes between the atmosphere and both land and ocean lower than 20 PgC century-1 (averaged over two centuries)."

In (Mignot et al., 2013):
"First of all, although it is clear that the oceanic adjustment requires several hundreds of years, this figure illustrates that all simulations approaches an equilibrium state after 300 years. The latter is nevertheless not reached and may require thousands of years, as it was necessary in CM5_piCtrl."

In (Stouffer et al., 2004):
"The initialization of coupled atmosphere-ocean climate models (AOGCMs) has been a long-standing problem in models used to study climate change on multi-century time scales (Moore et al. 2001). To perform simulations of the twentieth century and into the future, modellers must start with an initial equilibrium prior to extensive industrialization (typically near year 1850)."

In (Vichi et al., 2011):
"Oceanic DIC and alkalinity pools have been initialized from current climate data reconstructions (Key et al. 2004) and DIC has been spun up to equilibrium with the preindustrial atmospheric CO2 concentration by means of an acceleration method (adapted from Alessandri 2006 and Alessandri et al. 2011) consisting of increasing the air–sea CO2 outgassing flux of a factor 20 and removing the corresponding DIC amount homogeneously from the oceanic pool."

Besides, in most of the reference papers we reviewed for this study it is indicated that a near steady-state equilibrium is declared or assumed before performing any CMIP5 control and 20th century simulation. Therefore, we prefer to keep the term "assumed" in the revised version of the manuscript.

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.

We have amended the sentence as follows.

Revised:

"Depending on ocean circulation, it ranges from 1500 years for subsurface water masses to 10000 years for deep water masses (Wunsch and Heimbach, 2008)."

8759:

Gupta et al. (2012, 2013).

Should be ‘Sen Gupta et al’

Done and acknowledged.

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?

In the revised version of the manuscript, we chose to simplify the methodology and to concentrate the analysis on:

- The error or bias (metric 2 in the submitted manuscript)
- The root-mean-squared error (RMSE, the metric 5 in the submitted manuscript)

Since most of the analyses were performed with the RMSE, we have chosen to remove mention to the metrics 4 ($\Delta\sigma$) in the revised manuscript.

Consequently, we have amended the sentence as follows.

Original:

“The skill score metrics are (1) the global averaged concentrations for overall drift; (2) the average error or bias for mismatches between modeled and observed fields; (3) spatial correlation for mismatches between modeled and observed large-scale structures; (4) the difference between modeled and observed spatial variability ($\Delta\sigma$) and (5) the root-mean squared error (RMSE) to assess the total cumulative errors between modeled and observed fields.”

Revised:

“The skill score metrics are (1) the global averaged concentrations for overall drift; (2) the error or bias between modeled and observed fields at each grid-cell; (3) spatial correlation between model and observations to assess mismatches between modeled and observed large-scale structures; (4) the root-mean squared error (RMSE) to assess the total cumulative errors between modeled and observed fields.”

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.

As stated above in the response to the Editor, we updated Figure 1 in order to improve its readability: we have enlarged the barplot and made cross hatching in only one way direction.

8766:

except some recommendations for the decadal prediction exercise …

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

At least one ESM participating in the CMIP5 decadal prediction exercise included BGC, namely IPSL-CM5A-LR (Séférian et al., 2014)).

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?

We appreciate the reviewer's careful reading. To address this question, we have computed the drift in fgCO2 over several time windows (Table R1):

Time window (model year)	251-350	301-400	401-500
Drift (PgC y-1 y-1)	0.0030	0.0004	0.0007

Table R1: Drift in ocean carbon flux (fgCO2 in PgC y-2) over various time window of 100 years.

Drift estimates differ by one order of magnitude between each other and they decrease with time. Although our computations show drift to fluctuate in course of the spin-up simulation, only the two last time window (301-400 and 401-500) are statistically different from the long-term drift of 0.001 PgC y-2 at 90% confidence level.

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

I'm a bit confused. By steady state do you mean when the air-sea flux is zero?
But this isn't 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?

We apologize for the confusion. The subsection discusses two distinct criteria:
(1) The ocean carbon flux: the model set up includes a prescribed input of riverine carbon which should induce an outgassing of about 0.45-0.6 PgC y-1 at preindustrial state and under quasi-steady-state equilibrium (see (Aumont et al., 2001)).
(2) The drift in ocean carbon flux: the simple drift model is used to track the temporal evolution of drift until it approaches a value close to zero. At this stage, we consider that the variable has reached a quasi-steady-state equilibrium.

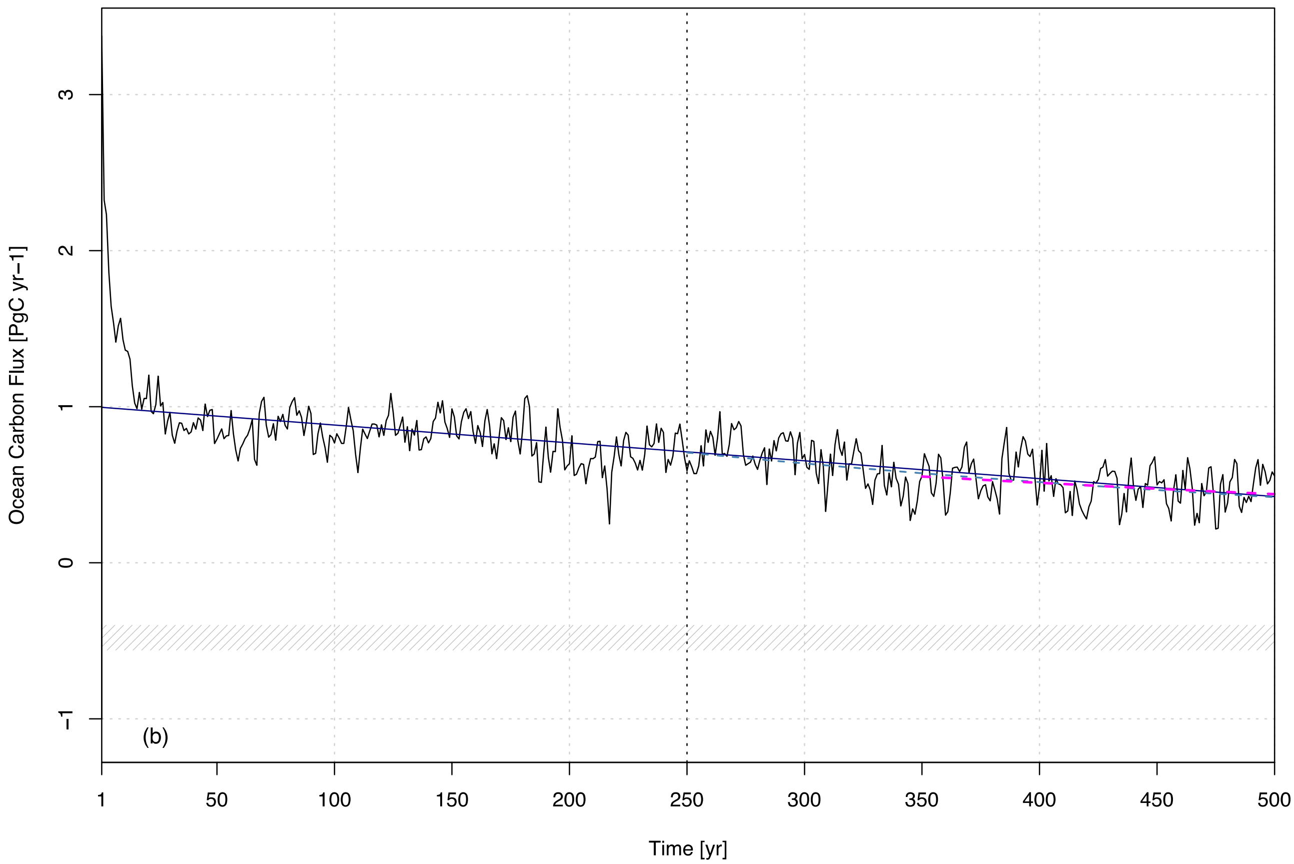


Figure R1: Comparison of linear (solid blue line) and exponential (dashed magenta and green lines) regression in ocean carbon fluxes from years 250 to 500.

To address the first point, we fitted a linear model (in solid blue line in Fig R1) and two exponential models (in green and magenta dashed lines in Fig R1) to the time series of carbon flux. The linear model was fitted over 251-500 years. The exponential models were fitted over years 251-500, respectively 351-500. While the differences between the three models are overall small, the two exponential models show a flattening of the long-term tendency towards the end of the simulation (Fig. R1).

From these fits, we estimated that fgCO2 would reach the range of 0.44-0.56 PgC y-1 after 1627-1838 years of simulation, This range corresponds to the multi-model inverse estimate of preindustrial fluxes of CO2 estimated by (Mikaloff Fletcher et al., 2007) plus a river-induced outgassing of 0.45 PgC y-1.

Since the model has already completed 500 years of spin-up, we substracted this duration and concluded that the model would fit the target flux after an additional of 1127-1338 years of simulation.

The above is consistent with the estimate computed by the referee considering that he/she did not account for the 500 years of spin-up simulation already performed by the model. After correction, his/her estimate of 1450 years is close to ours.

To improve the readability, we modified the following subsection.

<u>Original:</u>

"The temporal evolution of sea-to-air CO_2 fluxes was used in phase 2 of the Ocean Carbon Model Intercomparison Project (OCMIP-2, (Orr, 2002)*) as an equilibration metric for the marine biogeochemistry and was still widely used during CMIP5. Figure*

2b presents its evolution in the 500-year long spin-up simulation. The global ocean sea-to-air CO_2 flux is ~-0.7 Pg C y^{-1} over the last decades of the spin-up simulation (negative values indicate ocean CO_2 uptake). The global sea-to-air carbon flux does not fit the range of values estimated from preindustrial natural ocean carbon flux inversions (e.g. (Gerber and Joos, 2010) *or* (Mikaloff Fletcher et al., 2007)*, referred to as MF2007 on Figure 2), which amounts to 0.03 ± 0.08 Pg C y^{-1} (to which an open-ocean river-induced carbon outgassing of 0.45 Pg C y^{-1} has to be added on Figure 2 accordingly to* (IPCC, 2013; Le Quéré et al., 2015)*. This indicates that the ocean carbon cycle has not reached a steady state in the model system following the 500 years of integration.*

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 Pg C y^{-2} from year 250 to 500, it is much weaker over the last century of the simulation (5x10^{-4} Pg C y^{-2}). Using an approximate relaxation time of 73 years estimated from the simple drift model (Equation 1) over years 250-500 and the drift of 0.001 Pg C y^{-2}, we find that the simulated sea-to-air carbon flux would reach a steady state after ~500 supplemental years of spin-up. After the additional 500 supplemental years of spin-up, sea-air carbon flux would fall into the range of inverse estimates of MF2007 with accounting for outgassing of river carbon of 0.45 Pg C y^{-1}. This estimate does not account for the non-linearity of the ocean carbon cycle and the associated process uncertainties (Schwinger et al., 2014). This estimation potentially underestimates the time required to equilibrate the ocean carbon cycle and sea-to-air carbon fluxes in the range of inversion estimates. The duration of the spin-up simulation would have to be increased by an additional 500 years to account for the estimated river-induced natural CO_2 outgassing of about 0.45 Pg C y^{-1} (IPCC, 2013). The drift of 0.001 Pg C y^{-2} is, however, much smaller than the oceanic sink for anthropogenic carbon. Even if not fully equilibrated in terms of carbon balance, it is likely that this run would have given consistent estimates of anthropogenic carbon uptake in transient historical hindcasts.“

<u>Revised:</u>

“The temporal evolution of sea-to-air CO_2 fluxes was used in phase 2 of the Ocean Carbon Model Intercomparison Project (OCMIP-2, Orr (2002)) as an equilibration metric for the marine biogeochemistry and was still widely used during CMIP5. Figure 2b presents its evolution in the 500-year long spin-up simulation. The global ocean sea-to-air CO_2 flux is ~-0.7 Pg C y^{-1} over the last decades of the spin-up simulation (negative values indicate ocean CO_2 uptake).

To assess the global sea-to-air carbon flux, we use the range of values estimated from preindustrial natural ocean carbon flux inversions (e.g. Gerber and Joos (2010) or Mikaloff Fletcher et al. (2007)). Since, these estimates do not account for the preindustrial carbon outgassing induced by the river input, while our model does, we have added a constant outgassing of 0.45 Pg C y^{-1} to the range of 0.03 ± 0.08 Pg C y^{-1} (Mikaloff Fletcher et al. 2007). This value of 0.45 Pg C y^{-1} corresponds to the global open-ocean river-induced carbon outgassing accordingly to IPCC (2013) or Le Quéré et al. (2015). Consequently, in our modeling framework, the target value of the global sea-to-air carbon flux ranges between 0.4 and 0.56 Pg C y^{-1}.

Figure 2b shows that the global sea-to-air carbon flux does not fit our range of values estimated from preindustrial natural ocean carbon flux inversions. Besides, Figure 2b shows that the drift in the global sea-to-air carbon flux reduces more slowly after a strong decline during the first 50 years of the spin-up simulation. While this drift is about 0.001 Pg C y^{-2} from year 250 to 500, it is weaker over the last century of the simulation (7x10^{-4} Pg C y^{-2}). Using a linear fit over the last century of the simulation with a drift of 7x10^{-4} Pg C y^{-2}, we estimate that the simulated sea-to-air carbon flux would reach the range of 0.4-0.56 Pg C y^{-1} after 1100 to 1300 supplemental years of spin-up simulation. Our simple drift model (Equation 1) gives a relaxation time of around 160 years, which *indicates that drift in ocean carbon flux should range between $2x10^{-7}$ and $7x10^{-7}$ Pg C y^{-2} after this 1100 to 1300 supplemental years of spin-up simulation."*

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.

The reviewer is right. We apologize for errors in reporting results of our computation. Fig R2 presents the fit of the drift model at three depth levels. For this Figure, we computed drift in oxygen RMSE over a time window of 100 years starting from model year 200 to model year 400 every 5 years. The simple drift model was fitted to the resulting drift estimates presented with black circles in Fig R2.

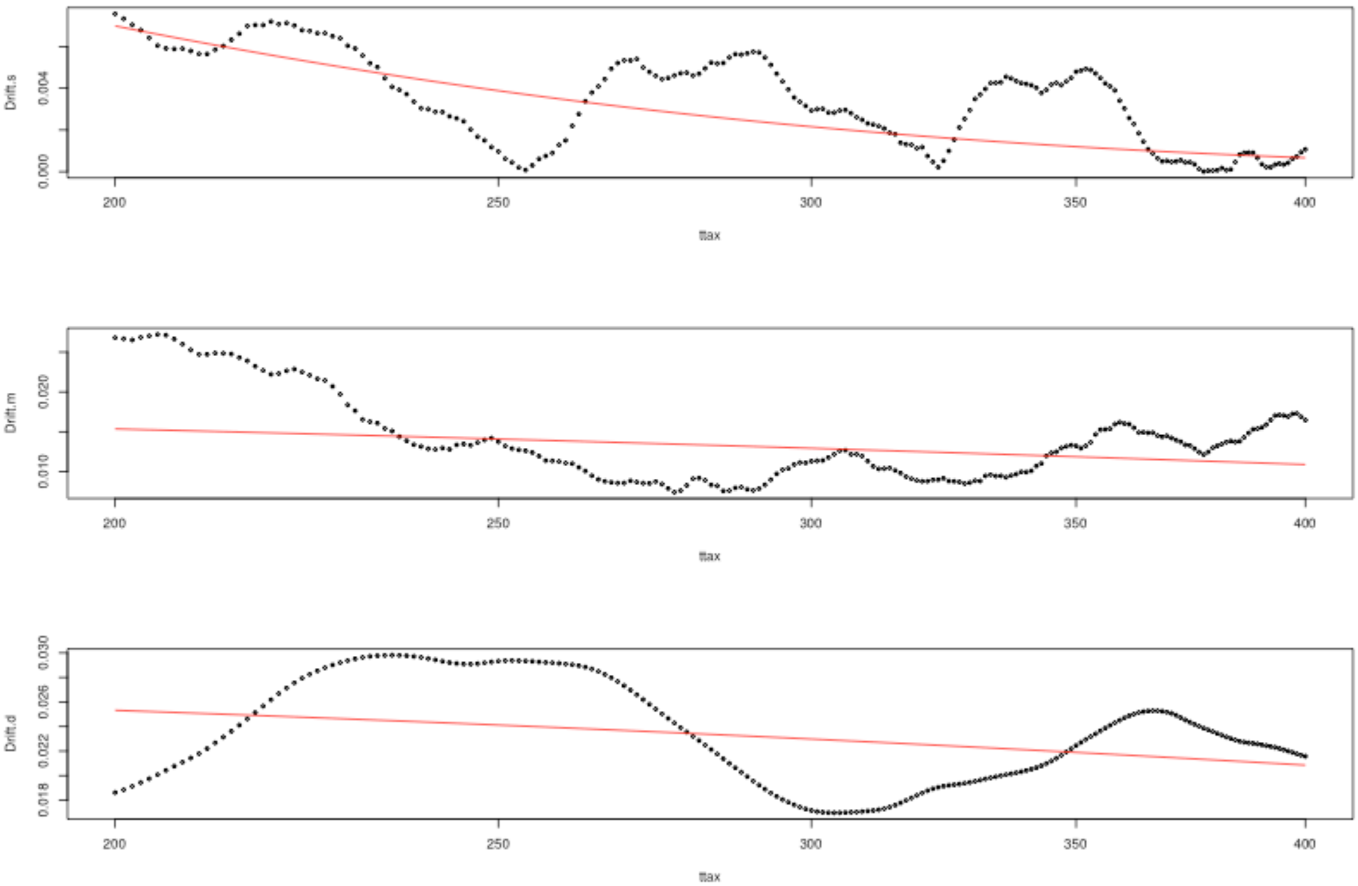


Figure R2: Evaluation of a simple drift model to fit drift in O2 RMSE for time windows of 100 years starting every 5 years from year 200 to 400 as simulated by IPSL-CM5A-LR 500-year-long spin-up simulation. Top: surface, middle 150m, bottom 2000m.

The corrected relaxation times of drift in the oxygen field are 90, 564 and 1149 years at surface, 150 m, respectively 2000 m depth.

We agree with the referee's comment on the noise in the fit (below). Figure R2 clearly shows that there are substantial fluctuations in the drift across the spin-up simulation. To assess uncertainty in relaxation time, we repeated the analysis for time windows of 100, 150, 200 and 250 years. Table R2 presents the relaxation time for oxygen RMSE for these time windows.

Depth levels	100 years	150 years	200 years	250 years	Mean±sd
surface	90	200	126	40	114±67
150 m	564	391	238	306	375±140
2000 m	1149	590	895	1829	1116±527

Table R2: Relaxation time estimated from the 500-year-long spin-up simulation performed with IPSL-CM5A-LR using a simple drift model and different time windows.

We modified the subsection as follows.

Original:

"From these two metrics, the simple drift model (Equation 1) enables us to determine the relaxation time τ required to reach equilibration over the last century of spin-up simulation. The relaxation times for oxygen RMSE are about 4, 13 and 140 y at the surface 150 m and 2000 m, respectively. Different values are derived for oxygen $\Delta\sigma$ with

8, 7 and 46 y at surface, 150 and 2000 m, respectively. Values for other biogeochemical fields are quite similar to those for O_2 except for NO_3 at 150 m. This contrasting result between the two skill score metrics expresses the fact that RMSE accounts for the total distance between modeled and observed oxygen distributions, while Δσ considers solely the difference in spatial structure between model fields and observations. This shows that the time scale for equilibration of spatial structure is not necessarily the same as the drift."

Revised:

*"**3-5 Drifts in IPSL-CM5A-LR spin-up simulation***

With the evolution of the RMSE established, we can use the simple drift model (Equation 1) to determine the relaxation time, τ, required to reach equilibration after a longer of spin-up simulation. To use this simple drift model, we compute the drift in RMSE determined from time segments of 100 years distributed evenly every 5 years from year 250 to 500 for O_2, NO_3 and Alk-DIC tracers. The drift model (magenta lines in Figure 8) is fitted level to the 80 drift values for each field and each depth (colored crosses in Figure 8).

The simple drift model fits well the evolution of the drift in RMSE for the biogeochemical variables along the spin-up simulation of IPSL-CM5A-LR (Figure 8). Correlation coefficients are mostly significant at 90% confidence level ($r^*=0.14$ determined with a student distribution with significance level of 90% and 80 degrees of freedom), except for NO_3 at surface and Alk-DIC at 150 m. Another exception is found for NO_3 at 150 m where the drift does not correspond to an exponential decay of the drift as function of time. The large confidence interval of the fit indicates that the fit would have been considered as non-significant given a longer spin-up simulation or a higher confidence threshold.

When significant, estimates of τ for O_2 RMSE are ≈ 90, 564 and 1149 y at the surface 150 m and 2000 m, respectively. These values match reasonably well τ estimated for NO_3 RMSE at 2000 m (1130 y) and those for Alk-DIC RMSE at surface and 2000 m (137 and 1163 y). However, these estimates are sensitive to the time windows used to compute the drift. For a subset of time windows between 100 and 250 years by step of 50 years, τ estimates for O_2 RMSE are ≈ 114±67, 375±140 and 1116±527 y at the surface 150 m and 2000 m depth. These large uncertainties associated with τ estimates are essentially due to the length of the spin-up simulation. A longer spin-up simulation would improve the quality of the fit (see Figure S1)."

We added Figure S1 to the supplementary materials to show how the fit is sensitive to the time-window. Estimates of the relaxation time are quite similar when using a time-window greater than 80 years. Below this threshold, the quality of the fit is significantly reduced ($R<0.3$)

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

Given that the minimum control is 250yrs I don't 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)

We thank the referee for his/her thoughtful comment. Accordingly, we re-run the analysis for the different CMIP5 models using the full available control simulation and a time window of 100 years. We modified Figure 8 which now presents this new computation and includes error bars for each model drift.
We removed the fit performed with the IPSL spin-up simulation from the Figure 9 (previous Fig. 8), acknowledging that extrapolation IPSL drift up to 11900 years is subject to large uncertainties. We have nonetheless included results of the drift computation performed with this simulation in the Figure to strengthen our conclusions. They are represented in magenta cross over the available period (1 to 500 model year).

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

Please see response below

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

Please refer to text changes presented below.

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

Please see the modification of the text below.

extrapolated over the 250–11900 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

As mentioned above, we have removed the extrapolation from the original version of Figure 8. Besides, we have introduced a new subsection in the revised manuscript with new Figures. Previous section 3.4 and 3.5 are now splitted in 3.4 3.5 and 3.6. Major changes are presented below.

Please note that a new Figure has been introduced as Figure 8 (see modification below). Therefore, the Figure 8 of the submitted manuscript now becomes Figure 9 in the revised manuscript.

"3-5 Drifts in IPSL-CM5A-LR spin-up simulation

With the evolution of the RMSE established, we can use the simple drift model (Equation 1) to determine the relaxation time, τ, required to reach equilibration after a longer of spin-up simulation. To use this simple drift model, we compute the drift in RMSE determined from time segments of 100 years distributed evenly every 5 years from year 250 to 500 for O_2, NO_3 and Alk-DIC tracers. The drift model (magenta lines in Figure 8) is fitted level to the 80 drift values for each field and each depth (colored crosses in Figure 8).

The simple drift model fits well the evolution of the drift in RMSE for the biogeochemical variables along the spin-up simulation of IPSL-CM5A-LR (Figure 8). Correlation coefficients are mostly significant at 90% confidence level (r=0.14 determined with a student distribution with significance level of 90% and 80 degrees of freedom), except for NO_3 at surface and Alk-DIC at 150 m. Another exception is found for NO_3 at 150 m where the drift does not correspond to an exponential decay of the drift as function of time. The large confidence interval of the fit indicates that the fit would have been considered as non-significant given a longer spin-up simulation or a higher confidence threshold.*

When significant, estimates of τ for O_2 RMSE are ≈ 90, 564 and 1149 y at the surface 150 m and 2000 m, respectively. These values match reasonably well τ estimated for NO_3 RMSE at 2000 m (1130 y) and those for Alk-DIC RMSE at surface and 2000 m (137 and 1163 y). However, these estimates are sensitive to the time windows used to compute the drift. For a subset of time windows between 100 and 250 years by step of 50 years, τ estimates for O_2 RMSE are ≈ 114±67, 375±140 and 1116±527 y at the surface 150 m and 2000 m depth. These large uncertainties associated with τ estimates are essentially due to the length of the spin-up simulation. A longer spin-up simulation would improve the quality of the fit (see Figure S1).

3-6 Drifts in CMIP5 ESMs preindustrial simulations

In this subsection, the analysis is extended to the CMIP5 archive. We focus on oxygen fields in the long preindustrial simulation, piControl, for the 15 available CMIP5 ESMs. From these simulations that span from 250 to 1000 years, we compute the drift in O_2 RMSE across depth from several time segments of 100 years distributed evenly every 5 years from the beginning until the end of the piControl simulation. These drifts are used as a surrogate for drift computed from the spin-up of each model since such simulations are not available through the data portal.

Figure 9 represents the drift in O_2 RMSE versus the spin-up duration for each CMIP5 ESM. The analysis shows that the drift in O_2 RMSE differs substantially between models. For a given model, drifts in other biogeochemical tracers (NO_3 and Alk-DIC) display similar features (not shown). The between-model differences in drift are not surprising since there are no reasons for different models to exhibit similar drift for a given field. Yet, Figure 9 shows that a global relationship emerges from this ensemble when using the simple drift model to fit the drift in O_2 RMSE as function of the spin-up duration (solid green lines in Figure 9). With a 90% confidence level, this relationship suggests a

general decrease of the drift as a function of spin-up duration for all depth levels. At the surface and at 2000 m depth, the quality of fits is low with correlation coefficients of about ~0.4. These are however significant at 90% confidence level (r=0.34 determined with a student distribution with significance level of 90% and 15 models as degree of freedom). The weakest correlation coefficient is found for the fit at 150 m depth and hence indicating that there is no link between the drift in O_2 RMSE and the duration of the spin-up simulation. This low significance level must be put into perspective given the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1) that can deteriorate the drift-spin up duration relationship in this ensemble of models.*
The drift versus spin up duration relationship established from the 15 CMIP5 ESMs is nonetheless consistent with the results obtained with IPSL-CM5A-LR (The results in Figure 8 have been reported in Figure 9 with magenta crosses). Consistency is indicated by the sign of the drift versus spin up duration relationship of the IPSL-CM5A-LR model at the various depth levels, although their magnitudes differ. This difference in magnitude is not surprising if one considers that drift is highly model and protocol dependent and that the length of the IPSL-CM5A-LR spin-up simulation is potentially too short to determine accurate estimates of the long-term drift in O_2 RMSE. Despite these differences, our analyses show that a relationship between the drift in O_2 RMSE versus the spin-up duration emerges from an ensemble of models and is broadly consistent with our theoretical framework of a drift model established from the results of the IPSL-CM5A-LR model (Figure 8).”

8773:
We employ ΔRMSE 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.

Correct.

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.

The referee is right.
In the revised version of the manuscript, we updated the different computations taking into account the referee’s comments. In particular, we accounted for uncertainties associated with the long-term drift estimate and those due to the different starting dates of the historical hindcast.

In the revised version, we now use

(1) The average of several drift estimates computed over a time window of 100 years from year 1 to the end of the preindustrial simulation every 5 years.

(2) The ensemble-mean of all historical hindcast members (over 1986-2005).

We preferred this approach rather than computing a single drift estimate from the full control simulation (since this latter is not equal between models).

We updated Figure 9 (now Figure 10 in the revised manuscript) accordingly of the manuscript and we have amended the text as follows.

Original:

"To assess the impact of model drift inherited from the diversity of spin-up strategies (Figure 1 and Table 1) on model performance metrics, the incremental deviation due to drift in biogeochemical fields is estimated from the simple drift model (Equation 1). The incremental deviation, ΔRMSE, is computed using the relaxation time τ *determined from the piControl simulations of each CMIP5 model (Figure 8) and a common duration of T=3000 years for all models:*

$$\Delta RMSE = \int_0^T drift(t=0) \times \exp(-\frac{1}{\tau}t)dt \qquad (2)$$

where ΔRMSE has the same unit as RMSE. The common duration T is used to bring model drift close to zero and hence to make models comparable to each other.

We employ ΔRMSE to penalize the normalized distance from the observations assuming that this drift-induced deviation in tracer fields can be added to RMSE. This means that the effect of the penalty is to increase the normalized distance giving a consistent measure of the equilibration error."

Revised:

"To assess the impact of model drift inherited from the diversity of spin-up strategies (Figure 1 and Table 1) on the performance metrics, we use a simple additive assumption to incorporate an incremental error due to the drift, ΔRMSE, to the above-mentioned RMSE. This incremental error due to the drift is computed using the relaxation time τ *determined from the piControl simulations of each CMIP5 model at each depth level (Equation 1 and Figure 9) and a common duration of T=3000 years for all models (m):*

$$\Delta RMSE_m(z) = \int_0^T drift_m(z, t=0) \times \exp(-\frac{1}{\tau(z)}t)dt \qquad (2)$$

where ΔRMSE has the same unit as RMSE.

The common duration T is used to bring model drift close to zero and hence to make models comparable to each other.

We employ ΔRMSE to penalize the distance from the observations assuming that this drift-induced deviation in tracer fields can be added to RMSE. This means that the effect of the penalty is to increase the distance giving a consistent measure of the equilibration error."

In addition to this modification, we extended the discussion of our approach in a new subsection :

"4-4 Limitations of the framework

In this work, the analyses focus on the globally averaged O_2 *RMSE across a diverse ensemble of CMIP5 models, which differ in terms of represented processes, spatial* *resolution and performance in addition to differences in spin-up protocols. Major limitations of the framework are presented below.*

Due to their specificities in terms of processes and resolution (e.g., Cabré et al., (2015), Laufkötter et al. (2015)), regional drift in CMIP5 models may differ from the drift computed from globally averaged skill-score metrics (see Figure S2 and S3). These differences may lead to different estimates of the relaxation time τ at regional scale. Moreover, the combination of regional ocean physics and biogeochemical processes in each individual model may drive an evolution of regional drift in RMSE that does not fit the hypothesis of an exponential decay of the drift during the course of the spin-up simulation.

The above-mentioned remark can explain the relatively low confidence level of the fit to drift across the multi-model CMIP5 ensemble (Figure 9). The relatively low significance level of the fit directly reflects not only the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1) but also the large diversity of processes and resolution of the CMIP5 models. An improved derivation of the penalization would require access to output from spin-up simulations for each individual model or, at least, a better quantification of model-model differences in terms of initial conditions.

Finally, it is unlikely that model fields drift at the same rate along the spin-up simulation, even under the same spin-up protocols. Indeed, as shown in Kriest and Oschlies (2015), various parameterizations of the particles sinking speeds in a common physical framework may lead to a similar evolution of the globally averaged RMSE in the first century of the spin-up simulation but display very different behaviour within a time-scale of $O(10^3)$ years. As such, drift and τ estimates need to be used with caution when computed from short spin-up simulation because they can be subject to large uncertainties."

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

Our focus is on the identification of main patterns, rather than on the description of individual models. We nevertheless added a sentence specific to GFDL ESM2:
"The ranking of GFDL-ESM2G and GFDL-ESM2M slightly evolves with penalization but both models stay close to the ensemble median and ensemble mean."

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

We removed the word 'propagate' from the revised manuscript.

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

As mentioned above, we have corrected the relaxation time in the revised manuscript. Except at surface, subsurface and deep ocean relaxation times are greater than 150 years.

8777: that have drifted further away from their initial states …
This doesn't 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

In the ideal case of a model perfectly reproducing all the processes occurring in the real world (which is not the case), the model field will fit the observed field some time after the initial coupling shock (years to thousand of years).
Figure 9 abc confirms that none of the CMIP5 model represents an ideal case since none of them displays an RMSE close to zero for oxygen fields.
However, we acknowledge that a 500-year-long spin-up simulation might be too short to accurately determine the long-term drift of the model. The use of output from the spin-up simulations performed for CMIP5 would have provided a solution to the problem, but these have not been archived. We included further discussion on the limitation of our framework in the revised manuscript.

Swart and Fyfe (2011)
I'm not sure about the relevance of this study here - please explain

We removed this sentence from the text and the reference list.

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

This is not always true. Indeed Figure 1 indicates that several CMIP5 modelling groups have used previous simulations to initialize their model, some others have used mixed sources of initialization (both models and observations).
Nonetheless, we agree with the referee that drift in model field are one or two order of magnitude smaller than the climate change trends. This is why we emphasize the fact that our penalization approach does not totally turn upside down model standard ranking (i.e., done with standard RMSE over the historical period).
Besides, we have already mentioned this point in the submitted manuscript:
"The drift of 0.001 Pg C y^{-2} is, however, much smaller than the oceanic sink for anthropogenic carbon. Even if not fully equilibrated in terms of carbon balance, it is

likely that this run would have given consistent estimates of anthropogenic carbon uptake in transient historical hindcasts."

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

Please see the response below.

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 don't follow this. Your penalisation process doesn't 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.

In this section, we have discussed our results.
First, we have highlighted the fact that the fit of our model is quite low. Even if, correlation coefficients are larger than zero with a 90% confidence interval at surface and 2000m, there are substantial uncertainties on the drift estimates (shown in the revised Figure 8 with error bars). These uncertainties influence the confidence we can have on the fit of the exponential model.
Next, we have attributed the large diversity in drift to both the protocols employed for spin-up and the initial condition (observations, models, mixing of both or constant values). These have to be considered to explain part of the model drift. As mentioned above, we introduce a new subsection "**4-4 Limitations of this framework"** where we further discuss the limitations and caveats of our approach.

8779:
CMIP7 …
What happened to CMIP6?
We have corrected this error.
Yet, we acknowledge that CMIP6 has been omitted purposely since we (all of the co-authors) that it is/was too late to agree on a common set of spin-up protocol for CMIP6.

agree on a set of recommendations for initialization, spin-up protocols and duration
I'm not sure that it makes sense to have a common duration as different models drift at different rates
We understand the referee's point of view. Therefore, we have simplified the message with "*the community should agree on a set of simple recommendations for spin-up protocols*". Yet, we could agree that drift is a direct metrics of model performance. Consequently, a common set of recommendations including the duration of the spin-up simulation should contribute to valuable information for model assessment protocols. This suggestion needs further discussion and, of course, to be tested in a forthcoming study

Referee #2 (F. Joos)

This is a nice and timely paper that addresses an important issue – model drift. It reflects the authors' broad knowledge in the field of coupled modelling. The authors show that short spin-up simulations initialized with observations lead to a too optimistic error statistic and biases model ranking. The authors also make proposal how model drift may be accounted for in future model assessments. This is an important and original contribution to the field.
I recommend publication after the following comments have been addressed

We appreciate the thoughtful suggestions from F. Joos. We incorporated most of the suggestions in the revised manuscript.

1) I am concerned about the way the drift model is presented and introduced and that the drift model may be used inappropriately in future work. The authors apply an exponential model with a single relaxation time scale to approximate the evolution of drift. However, the application of a single time scale is most likely not appropriate to determine the drift in whole ocean RMSE or other global error statistics. For example, this is implicitly demonstrated by the results in section 3.5 where the authors apply the drift models for different depth levels individually and show that time scales are different between depth levels.
In my opinion, the following point should be made very clear in this manuscript and in the method, results and discussion/conclusion section: different relaxation time scales apply for different regions (and variables). This requires that the drift in RMSE and other metrics is to be determined for different regions or even for different grid boxes individually before the drift in RMSE for the whole ocean is to be determined. In this way, multiple time scales would be applied to estimate the evolution of whole ocean RMSE and to correct error statistics for drift.

Please see our response below.

2) I am not convinced that selecting depth levels as regions is a good approach. For example, drift at 2000 m in the well-ventilated North Atlantic Deep Water may be quite different from drift in the slowly ventilated North Pacific. It would be illustrative to compute the relaxation time scale, tau, individually for each grid cell and plot tau along sections in the Atlantic, Pacific and Indian (or similar). A grid-cell based approach is generally also applied when removing model drift from projections by using a control simulation. Computing tau for individual grid cells would be comparable with such an approach.

We agree with F. Joos and his comments are addressed in the revised manuscript by adding in a new subsection "4.4 Limitations of the framework" and including corresponding results to the supplementary material. At the scale of individual grid cells, drift displays a large temporal and spatial variability. The larger variability reflects the mismatch between model output and observations, i.e. model fields vary on inter-annual to decadal timescales, while observations are climatological means based on sparse observations. A similar problem arises when analyzing temporal trends and requires to be solved using either longer time series or smoothing procedures. Extending the analysis of drift to basin-scale improves the signal-to-noise ratio and facilitates the determination of drift without smoothing procedure.

The preceding is addressed in a new subsection:

"4-4 Limitations of the framework

In this work, the analyses focus on the globally averaged O_2 RMSE across a diverse ensemble of CMIP5 models, which differ in terms of represented processes, spatial resolution and performance in addition to differences in spin-up protocols. Major limitations of the framework are presented below.

Due to their specificities in terms of processes and resolution (e.g., Cabré et al., (2015), Laufkötter et al. (2015)), regional drift in CMIP5 models may differ from the drift computed from globally averaged skill-score metrics (see Figure S2 and S3). These differences may lead to different estimates of the relaxation time τ at regional scale. Moreover, the combination of regional ocean physics and biogeochemical processes in each individual model may drive an evolution of regional drift in RMSE that does not fit the hypothesis of an exponential decay of the drift during the course of the spin-up simulation.

The above-mentioned remark can explain the relatively low confidence level of the fit to drift across the multi-model CMIP5 ensemble (Figure 9). The relatively low significance level of the fit directly reflects not only the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1) but also the large diversity of processes and resolution of the CMIP5 models. An improved derivation of the penalization would require access to output from spin-up simulations for each individual model or, at least, a better quantification of model-model differences in terms of initial conditions.

Finally, it is unlikely that model fields drift at the same rate along the spin-up simulation, even under the same spin-up protocols. Indeed, as shown in Kriest and Oschlies (2015), various parameterizations of the particles sinking speeds in a common physical framework may lead to a similar evolution of the globally averaged RMSE in the first century of the spin-up simulation but display very different behaviour within a time-scale of $O(10^3)$ years. As such, drift and τ estimates need to be used with caution when computed from short spin-up simulation because they can be subject to large uncertainties."

The discussion of limitations is supported by two new supplementary Figures:

- Figure S2 presents the sensitivity of the drift profile computed either from global-averaged RMSE or form 3D RMSE. The figure suggests that while the approach selected for computing global drift might impact its magnitude, the general form of vertical profiles appears robust.
- Figure S3 presents basin-scale drift in O_2 RMSE and its structure for the ensemble of CMIP5 models. The results are broadly consistent with the outcome of the penalization approach (Figure 10) with models displaying the largest drift having the greatest penalization.

Further comments:

———————

1) A sufficiently long spin up over several hundred years is a prerequisite to estimate drift in error statistics and other variables. (High-resolution) models that are initialized with observed fields and not spun-up over several centuries very likely suffer from serious drift problems. It may be very difficult to estimate the future evolution of the drift from a short spin-up. This should be mentioned explicitly in the manuscript. (May be this could even be quantitatively illustrated by estimating relaxation time scales from an initial period, e.g., first 50 or 100 yr as compared to time scales from the last 100 year of the simulation as already presented for three different depth levels.)

As mentioned above, we have introduced a new subsection in the revised version of the manuscript in which we further discussed the limitation of our approach. One of the limitations is of course the duration of the spin-up simulation employed to determine the drift.
It is worth mentioning that the scope of the study emphasizes the impact of drift on skill-score assessment and not the assessment time required, for each CMIP5 models, to reach a quasi-steady-state equilibrium.

2) The authors may also note that rate of drifts (e.g. in the surface) may increase when the mode of model operation is changed, e.g. from prescribed atmospheric CO2 to freely simulated atmospheric CO2.

We agree with F. Joos. But this point was already mentioned in the submitted version of the manuscript:
"These developments will go along with an increase in the diversity and complexity of spin-up protocols applied to Earth system models, especially those including an interactive atmospheric CO_2 or interactive nitrogen cycle (Dunne et al., 2013; Lindsay et al., 2014). *The additional challenge of spinning-up emission-driven simulations with interactive carbon cycle will also require us to extend the assessment of the impact of spin-up protocols to the terrestrial carbon cycle. Processes such as soil carbon accumulation, peat formation as well as shift in biomes such as tropical and boreal ecosystems for dynamic vegetation models require several long time-scales to equilibrate (Brovkin et al., 2010; Koven et al., 2015)."*

3) The authors do hardly evaluate the validity of their exponential model. It would be nice if this model could be validated, e.g. in the context of a millennium long control simulation or similar?

An important part of the analysis was dedicated to the evaluation of the simple drift model. However, we did not present any material to support its reliability in the submitted version of the manuscript. Consequently, in agreement with F. Joos suggestion, we included the assessment of our simple drift model with a long millennial-scale control simulation of IPSL-CM5A-LR. The result of this assessment is presented in Figure S1.

In response to a suggestion by reviewer 1, Figure S1 shows sensitivity tests on the length of the time-window to compute the drift in O_2 RMSE. It supports the fact that long time series are required to accurately estimate the time of relaxation (R<0.3 with a time-window < 80 years).

Sec 3.6: It is not entirely clear whether the same time scale is applied here across all models considered. Please make this clear. It is also not clear whether different time scales are used for different depth levels. Please clarify.

We apologize for the lack of clarity. Pending on your comments and those of reviewer 1, we have amended the following section.

Original:

"To assess the impact of model drift inherited from the diversity of spin-up strategies (Figure 1 and Table 1) on model performance metrics, the incremental deviation due to drift in biogeochemical fields is estimated from the simple drift model (Equation 1). The incremental deviation, ΔRMSE, is computed using the relaxation time τ determined from the piControl simulations of each CMIP5 model (Figure 8) and a common duration of T=3000 years for all models:

$$\Delta RMSE = \int_0^T drift(t=0) \times \exp(-\frac{1}{\tau}t)dt \qquad (2)$$

where ΔRMSE has the same unit as RMSE. The common duration T is used to bring model drift close to zero and hence to make models comparable to each other.

We employ ΔRMSE to penalize the normalized distance from the observations assuming that this drift-induced deviation in tracer fields can be added to RMSE. This means that the effect of the penalty is to increase the normalized distance giving a consistent measure of the equilibration error."

Revised:

"To assess the impact of model drift inherited from the diversity of spin-up strategies (Figure 1 and Table 1) on the performance metrics, we use a simple additive assumption to incorporate an incremental error due to the drift, ΔRMSE, to the above-mentioned RMSE. This incremental error due to the drift is computed using the relaxation time τ determined from the piControl simulations of each CMIP5 model at each depth level (Equation 1 and Figure 9) and a common duration of T=3000 years for all models (m):

$$\Delta RMSE_m(z) = \int_0^T drift_m(z,t=0) \times \exp(-\frac{1}{\tau(z)}t)dt \qquad (2)$$

where ΔRMSE has the same unit as RMSE.

The common duration T is used to bring model drift close to zero and hence to make models comparable to each other.

We employ ΔRMSE to penalize the distance from the observations assuming that this drift-induced deviation in tracer fields can be added to RMSE. This means that the effect of the penalty is to increase the distance giving a consistent measure of the equilibration error."

Sec 3.2: I am somewhat confused here about the role of river outgassing. The clarity of the text should be increased. It is not readily clear whether the model should actually achieve a flux of 0 GtC/yr or an outgassing of ~0.4 to 0.6 GtC/yr at equilibrium.

Reviewer #1 also critized the lack of clarity of this section. To improve its readability and to clarify our computation, we modified this subsection as follows.

Original:

"The temporal evolution of sea-to-air CO_2 fluxes was used in phase 2 of the Ocean Carbon Model Intercomparison Project (OCMIP-2, (Orr, 2002)*) as an equilibration metric for the marine biogeochemistry and was still widely used during CMIP5. Figure 2b presents its evolution in the 500-year long spin-up simulation. The global ocean sea-to-air CO_2 flux is ~-0.7 Pg C y^{-1} over the last decades of the spin-up simulation (negative values indicate ocean CO_2 uptake). The global sea-to-air carbon flux does not fit the range of values estimated from preindustrial natural ocean carbon flux inversions (e.g.* (Gerber and Joos, 2010) *or* (Mikaloff Fletcher et al., 2007)*, referred to as MF2007 on Figure 2), which amounts to 0.03 ± 0.08 Pg C y^{-1} (to which an open-ocean river-induced carbon outgassing of 0.45 Pg C y^{-1} has to be added on Figure 2 accordingly to* (IPCC, 2013; Le Quéré et al., 2015)*. This indicates that the ocean carbon cycle has not reached a steady state in the model system following the 500 years of integration.*

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 Pg C y^{-2} from year 250 to 500, it is much weaker over the last century of the simulation (5x10^{-4} Pg C y^{-2}). Using an approximate relaxation time of 73 years estimated from the simple drift model (Equation 1) over years 250-500 and the drift of 0.001 Pg C y^{-2}, we find that the simulated sea-to-air carbon flux would reach a steady state after ~500 supplemental years of spin-up. After the additional 500 supplemental years of spin-up, sea-air carbon flux would fall into the range of inverse estimates of MF2007 with accounting for outgassing of river carbon of 0.45 Pg C y^{-1}. This estimate does not account for the non-linearity of the ocean carbon cycle and the associated process uncertainties (Schwinger et al., 2014). This estimation potentially underestimates the time required to equilibrate the ocean carbon cycle and sea-to-air carbon fluxes in the range of inversion estimates. The duration of the spin-up simulation would have to be increased by an additional 500 years to account for the estimated river-induced natural CO_2 outgassing of about 0.45 Pg C y^{-1} (IPCC, 2013). The drift of 0.001 Pg C y^{-2} is, however, much smaller than the oceanic sink for anthropogenic carbon. Even if not fully equilibrated in terms of carbon balance, it is likely that this run would have given consistent estimates of anthropogenic carbon uptake in transient historical hindcasts."

Revised:

"The temporal evolution of sea-to-air CO_2 fluxes was used in phase 2 of the Ocean Carbon Model Intercomparison Project (OCMIP-2, Orr (2002)) as an equilibration metric for the marine biogeochemistry and was still widely used during CMIP5. Figure 2b presents its evolution in the 500-year long spin-up simulation. The global ocean sea-to-air CO_2 flux is ~-0.7 Pg C y^{-1} over the last decades of the spin-up simulation (negative values indicate ocean CO_2 uptake).

To assess the global sea-to-air carbon flux, we use the range of values estimated from preindustrial natural ocean carbon flux inversions (e.g. Gerber and Joos (2010) or Mikaloff Fletcher et al. (2007)). Since, these estimates do not account for the preindustrial carbon outgassing induced by the river input, while our model does, we have added a constant outgassing of 0.45 Pg C y^{-1} to the range of 0.03 ± 0.08 Pg C y^{-1} (Mikaloff Fletcher et al. 2007). This value of 0.45 Pg C y^{-1} corresponds to the global open-ocean river-induced carbon outgassing accordingly to IPCC (2013) or Le Quéré et

al. (2015). Consequently, in our modeling framework, the target value of the global sea-to-air carbon flux ranges between 0.4 and 0.56 Pg C y^{-1}.
Figure 2b shows that the global sea-to-air carbon flux does not fit our range of values estimated from preindustrial natural ocean carbon flux inversions. Besides, Figure 2b shows that the drift in the global sea-to-air carbon flux reduces more slowly after a strong decline during the first 50 years of the spin-up simulation. While this drift is about 0.001 Pg C y^{-2} from year 250 to 500, it is weaker over the last century of the simulation ($7x10^{-4}$ Pg C y^{-2}). Using a linear fit over the last century of the simulation with a drift of $7x10^{-4}$ Pg C y^{-2}, we estimate that the simulated sea-to-air carbon flux would reach the range of 0.4-0.56 Pg C y^{-1} after 1100 to 1300 supplemental years of spin-up simulation. Our simple drift model (Equation 1) gives a relaxation time of around 160 years, which indicates that drift in ocean carbon flux should range between $2x10^{-7}$ and $7x10^{-7}$ Pg C y^{-2} after this 1100 to 1300 supplemental years of spin-up simulation."

8767, line 17: additional compared to what?
Please refer to the modification of the text above.

8778 line 24: conclusion: Is it sufficient to report the drift in global RMSE? Perhaps this clause should be deleted or refined
This section has been amended as follows.
Original:
"Skill-score metrics are expected to be widely used in the framework of the future CMIP6 (Meehl et al., 2014) with the development of international community benchmarking tools like the ESMValTool (http://www.pa.op.dlr.de/ESMValTool (Eyring et al., 2015)). The assessment of model skill to reproduce observations will focus on the modern period. In order to increase the reliability of these traditional metrics, additional metrics that allow us to determine the equilibrium state of the model like the 3-dimensional growth rate or drift of relevant skill score metrics (e.g., RMSE) over the last decades or centuries of the spin-up, should be included in the set of standard assessment tools for CMIP6."
Revised:
"Skill-score metrics are expected to be widely used in the framework of the future CMIP6 (Meehl et al., 2014) with the development of international community benchmarking tools like the ESMValTool (http://www.pa.op.dlr.de/ESMValTool (Eyring et al., 2015)). The assessment of model skill to reproduce observations will focus on the modern period. Complementary to this approach, our results call for the consideration of spin-up and initialization strategies in the determination of skill assessment metrics (Friedrichs et al., 2009; Stow et al., 2009) and, by extension, to model weighting (Steinacher et al., 2010) and model ranking (Anav et al., 2013). Indeed, the use of equilibrium-state metrics of the model like the 3-dimensional growth rate or drift of relevant skill score metrics (e.g. RMSE) could be employed to increase the reliability of these traditional metrics and, as such, should be included in the set of standard assessment tools for CMIP6."

Referee #3 (I. Kriest)
This paper examines the impact of different initialization procedures and spinup times in CMIP5 models, the resulting drift, and its impact on model skill assessement. I am

delighted to see that finally the issue of spinup times and drift is addressed comprehensively for the CMIP5 model suite. However, I have two concerns or comments, that I think should be kept in mind, and a few minor issues.

We appreciate I. Kriest careful reading. We included most of her suggestions and corrections in the revised version of the manuscript.

(1) As far as I understand, the core model experiment, IPSL-CM5A-LR, was spun up from rest for 500 years. I am aware that it is sometimes quite expensive - in terms of computational cost - to simulate global or earth system models over a long time. However, I am not quite sure that a spinup time of 500 years, as used for this experiment, is always sufficient to draw conclusions about the long-term model drift. As has been shown recently (Kriest and Oschlies, 2015; www.geosci-model-dev.net/8/2929/2015/), simulated global average oxygen, nitrate, or total fixed nitrogen can exhibit a non-linear trajectory over time, sometimes with inflection points within the first few centuries of spinup; i.e., the model drift may not only decrease or increase, but change its sign. In practice, it means that, due to the many timescales involved, a model that shows a bad fit and negative trend within the first few hundred years e.g., with respect to global average oxygen, may cease to do so after some more centuries, and finally show a quite good fit to observed oxygen after some millenia.

We agree with the referee. In the revised version of the manuscript, we clearly stated that our 500-year-long spin-up simulation, as used for this study, is maybe too short to draw robust conclusions on the long-term drift. An ideal solution would have been to use output of the spin-up simulation performed in the context of CMIP5 but these latter have not been archived. This will be tested in a forthcoming study in the context of CMIP6 for which we hope some modeling groups will store output from the spin-up simulation.

(2) The above doesn't have to hold for all model types. It can depend on the biogeochemical time scales involved, i.e. on particle sinking speed or remineralization (Kriest and Oschlies, 2015), circulation, and probably other parameters as well. Given that the CMIP5 biogeochemical models involve a huge variety of these parameterizations (e.g., Cabre et al., 2015; www.biogeosciences.net/12/5429/2015/; Fig. 6), together with very different circulations, resolutions, etc., the time scales associated with model equilibration, as well as their transient may be very different, and not always follow linear relationships for the decay term.
Therefore, I would suggest to include some discussion on this in the paper. Overall, nevertheless I think this paper gives a helpful and timely overview about potential limitations of model-model and model-data comparison of this suite of models.

As mentioned above to the first referee and to F. Joos, we have amended the manuscript with the inclusion of this new subsection:
"4-4 Limitations of the framework
In this work, the analyses focus on the globally averaged O_2 RMSE across a diverse ensemble of CMIP5 models, which differ in terms of represented processes, spatial

resolution and performance in addition to differences in spin-up protocols. Major limitations of the framework are presented below.

Due to their specificities in terms of processes and resolution (e.g., Cabré et al., (2015), Laufkötter et al. (2015)), regional drift in CMIP5 models may differ from the drift computed from globally averaged skill-score metrics (see Figure S2 and S3). These differences may lead to different estimates of the relaxation time τ at regional scale. Moreover, the combination of regional ocean physics and biogeochemical processes in each individual model may drive an evolution of regional drift in RMSE that does not fit the hypothesis of an exponential decay of the drift during the course of the spin-up simulation.

The above-mentioned remark can explain the relatively low confidence level of the fit to drift across the multi-model CMIP5 ensemble (Figure 9). The relatively low significance level of the fit directly reflects not only the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1) but also the large diversity of processes and resolution of the CMIP5 models. An improved derivation of the penalization would require access to output from spin-up simulations for each individual model or, at least, a better quantification of model-model differences in terms of initial conditions.

Finally, it is unlikely that model fields drift at the same rate along the spin-up simulation, even under the same spin-up protocols. Indeed, as shown in Kriest and Oschlies (2015), various parameterizations of the particles sinking speeds in a common physical framework may lead to a similar evolution of the globally averaged RMSE in the first century of the spin-up simulation but display very different behaviour within a time-scale of $O(10^3)$ years. As such, drift and τ estimates need to be used with caution when computed from short spin-up simulation because they can be subject to large uncertainties."

Other comments:

p. 8760, line 27ff: "Oxygen is prognostically simulated using two different oxygen-to-carbon ratios, one for the oxic remineralization of matter and one for the sub-oxic pathway (Sarmiento and Gruber, 2006)." - It is not clear to me what is meant with "oxygen-to-carbon ratios": the ratio of organic matter, or of the process itself? If the latter, how can oxygen be used in sub-oxic pathways? If the former: doesn't this imply that either oxygen or carbon is not conserved when switching between these processes? E.g. consider that - implicitly - organic matter built during photosynthesis has a composition according to Anderson (1995, Deep-Sea Res. I, 42(9), 1675-1680), with C:H:O:N:P = 106:175:42:16:1. Of course, one usually does not describe OM in models exactly this way; but the assumption particularly about C:H:O (in some way: the amount of carbohydrates) is reflected in the stoichiometry for O2 release and CO2 consumption. If then the C:O-ratio of OM is different between remineralization and denitritification/anammox (whatever is considered), wouldn't this affect mass conservation of either C or O?

We apologize for this misleading information. We have amended the description of

PISCES accordingly.

<u>Original:</u>

"Oxygen is prognostically simulated using two different oxygen-to-carbon ratios, one for the oxic remineralization of organic matter and one for the sub-oxic pathway (Sarmiento and Gruber, 2006)."

<u>Revised:</u>

"Oxygen is prognostically simulated. The model distinguishes between oxic and suboxic remineralization pathways, the former relying on oxygen as electron acceptor, the latter on nitrate."

Therefore the total amount of C and O is conserved in PISCES.

p. 8763, subsection 2.3: I would suggest to more clearly define drift, to make this term more easily accessible for users outside the modeling or CMIP5 community.

We refined the subsection describing the way we determine the drift. In addition, in the revised subsection, we briefly discuss the sensitivity to the time-window used to compute the drift.

<u>Original:</u>

"The drift is determined for either concentrations in simulated biogeochemical fields or for skill score metrics (e.g., RMSE or Δσ) using a linear regression fit over a time window of 100 years. This time window of 100 years was chosen as a trade off between longer time window (>200 years) that smoothes the drift signal and shorter time window (<80 years) that introduces fluctuations due to internal variability."

<u>Revised:</u>

"The drift is determined for either concentrations in simulated biogeochemical fields or for skill score metrics (e.g., RMSE) using a linear regression fit over a time window of 100 years. This time window of 100 years was chosen as a trade off between a longer time window (>200 years) that smoothes the drift signal and a shorter time window (<100 years) that introduces fluctuations due to internal variability and hence impacting the quality of the fit (see the assessment performed with the millennial-long CMIP5 piControl simulation of IPSL-CM5A-LR in Figure S1)."

p. 8773, lines 10-11: "We employ ΔRMSE to penalize the normalized distance from the observations assuming that this drift-induced deviation in tracer fields can be added to RMSE. " - Why choose an additive model?

We acknowledge that there is no justification to employ a simple additive model rather than a multiplicative model in our case. That said, we aimed at keeping our framework as simple as possible for this study.

The additive approach is coherent with current 'drift-correction' approaches which are based on an additive hypothesis. As indicated in the submitted version of the manuscript:

"So far, the most frequent approach relies on the use of long preindustrial control simulations to 'remove' the drift embedded in the simulated fields over the historical period or future projections (Bopp et al., 2013; Cocco et al., 2013; Friedlingstein et al., 2006; 2013; Frölicher et al., 2014; Gehlen et al., 2014; Keller et al., 2014; Steinacher et

al., 2010; Tjiputra et al., 2014). *Although this approach allows to determine relative changes, it does not allow to investigate the underlying reasons of the spread between models in terms of processes, variability and response to climate change. The "drift-correction" approach, much as the one used for this study, assumes that drift-induced errors in the simulated fields can be isolated from the signal of interest."*

Testing the validity of both hypothesis (additive or multiplicative amplification of the errors) is not easy. We think that it would have required for example a large ensemble of historical simulations starting at various date of the spin-up, with an important computation cost. To our knowledge, this question remains an uncharted territory that would require further analyses to be answered.

References:

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. and Zhu, Z.: Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models, J. Climate, 26(18), 6801–6843, doi:10.1175/JCLI-D-12-00417.1, 2013.

Aumont, O., Orr, J. C., Monfray, P., Ludwig, W., Amiotte-Suchet, P. and Probst, J.-L.: Riverine-driven interhemispheric transport of carbon, Global Biogeochem. Cycles, 15(2), 393–405, doi:10.1029/1999GB001238, 2001.

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Séférian, R., Tjiputra, J. and Vichi, M.: Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models, Biogeosciences, 10(10), 6225–6245, doi:10.5194/bg-10-6225-2013, 2013.

Brovkin, V., LORENZ, S. J., JUNGCLAUS, J., RADDATZ, T., Timmreck, C., Reick, C. H., Segschneider, J. and SIX, K.: Sensitivity of a coupled climate-carbon cycle model to large volcanic eruptions during the last millennium, Tellus B, 62(5), 674–681, doi:10.1111/j.1600-0889.2010.00471.x, 2010.

Cabré, A., Marinov, I., Bernardello, R. and Bianchi, D.: Oxygen minimum zones in the tropical Pacific across CMIP5 models: mean state differences and climate change trends, Biogeosciences, 12(18), 5429–5454, doi:10.5194/bg-12-5429-2015, 2015.

Cocco, V., Joos, F., Steinacher, M., Frölicher, T. L., Bopp, L., Dunne, J., Gehlen, M., Heinze, C., Orr, J., Oschlies, A., Schneider, B., Segschneider, J. and Tjiputra, J.: Oxygen and indicators of stress for marine life in multi-model global warming projections, Biogeosciences, 10(3), 1849–1868, doi:10.5194/bg-10-1849-2013, 2013.

Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., Stouffer, R. J., Cooke, W., Dunne, K. A., Harrison, M. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Phillipps, P. J., Sentman, L. A., Samuels, B. L., Spelman, M. J., Winton, M., Wittenberg, A. T. and Zadeh, N.: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics, J. Climate, 25(19), 6646–6665, doi:doi: 10.1175/JCLI-D-11-00560.1, 2013.

Eyring, V., Righi, M., Evaldsson, M., Lauer, A., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt, K. D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., Kunert, D., Levine, R., Loew, A., Mäkelä, J., Martin, G., Mason, E., Phillips, A., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., van Ulft, L. H., Walton, J., Wang, S. and Williams, K. D.: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP, Geosci. Model Dev. Discuss., 8(9), 7541–7661, 2015.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Bloh, Von, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K. G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C. and Zeng, N.: Climate–Carbon Cycle Feedback Analysis: Results from the C 4MIP Model Intercomparison, J. Climate, 10(14), 3337–3353, doi:10.1175/JCLI3800.1, 2006.

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K. and Knutti, R.: Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks, J. Climate, 27(2), 511–526, doi:10.1175/JCLI-D-12-00579.1, 2013.

Friedrichs, M. A. M., Carr, M.-E., Barber, R. T., Scardi, M., Antoine, D., Armstrong, R. A., Asanuma, I., Behrenfeld, M. J., Buitenhuis, E. T., Chai, F., Christian, J. R., Ciotti, A. M., Doney, S. C., Dowell, M., Dunne, J. P., Gentili, B., Gregg, W., Hoepffner, N., Ishizaka, J., Kameda, T., Lima, I., Marra, J., Mélin, F., Moore, J. K., Morel, A., O'Malley, R. T., O'Reilly, J., Saba, V. S., Schmeltz, M., Smyth, T. J., Tjiputra, J., Waters, K., Westberry, T. K. and Winguth, A.: Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean, Journal of Marine Systems, 76(1-2), 113–133, doi:10.1016/j.jmarsys.2008.05.010, 2009.

Frölicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P. and Winton, M.: Dominance of the Southern Ocean in Anthropogenic Carbon and Heat Uptake in CMIP5 Models, J. Climate, 28(2), 862–886, doi:10.1175/JCLI-D-14-00117.1, 2014.

Gehlen, M., Séférian, R., Jones, D. O. B., Roy, T., Roth, R., Barry, J., Bopp, L., Doney, S. C., Dunne, J. P., Heinze, C., Joos, F., Orr, J. C., Resplandy, L., Segschneider, J. and Tjiputra, J.: Projected pH reductions by 2100 might put deep North Atlantic biodiversity at risk, Biogeosciences, 11(23), 6955–6967, 2014.

Gerber, M. and Joos, F.: Carbon sources and sinks from an Ensemble Kalman Filter ocean data assimilation, Global Biogeochem. Cycles, 24(3), n/a–n/a, doi:10.1029/2009GB003531, 2010.

IPCC: IPCC, 2013: Climate Change 2013: The Physical Science Basis., edited by T. F. Stoker, D. Qin, G. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Cambridge Univ Press, Cambridge, United Kingdom and New

York, NY, USA. 2013.

Keller, K. M., Joos, F. and Raible, C. C.: Time of emergence of trends in ocean biogeochemistry, Biogeosciences, 11(13), 3647–3659, doi:10.5194/bgd-10-18065-2013, 2014.

Koven, C. D., Chambers, J. Q., Georgiou, K., Knox, R., Negron-Juarez, R., Riley, W. J., Arora, V. K., Brovkin, V., Friedlingstein, P. and Jones, C. D.: Controls on terrestrial carbon feedbacks by productivity vs. turnover in the CMIP5 Earth System Models, Biogeosciences Discuss., 12(8), 5757–5801, 2015.

Kriest, I. and Oschlies, A.: MOPS-1.0: towards a model for the regulation of the global oceanic nitrogen budget by marine biogeochemical processes, Geosci. Model Dev, 8(9), 2929–2957, doi:10.5194/gmd-8-2929-2015, 2015.

Laufkötter, C., Vogt, M., Gruber, N., Aita-Noguchi, M., Aumont, O., Bopp, L., Buitenhuis, E., Doney, S. C., Dunne, J., Hashioka, T., Hauck, J., Hirata, T., John, J., Le Quéré, C., Lima, I. D., Nakano, H., Séférian, R., Totterdell, I., Vichi, M. and Völker, C.: Drivers and uncertainties of future global marine primary production in marine ecosystem models, Biogeosciences, 12(23), 6955–6984, 2015.

Le Quéré, C., Moriarty, R., Andrew, R. M., Peters, G. P., Ciais, P., Friedlingstein, P., Jones, S. D., Sitch, S., Tans, P., Arneth, A., Boden, T. A., Bopp, L., Bozec, Y., Canadell, J. G., Chini, L. P., Chevallier, F., Cosca, C. E., Harris, I., Hoppema, M., Houghton, R. A., House, J. I., Jain, A. K., Johannessen, T., Kato, E., Keeling, R. F., Kitidis, V., Klein Goldewijk, K., Koven, C., Landa, C. S., Landschützer, P., Lenton, A., Lima, I. D., Marland, G., Mathis, J. T., Metzl, N., Nojiri, Y., Olsen, A., Ono, T., Peng, S., Peters, W., Pfeil, B., Poulter, B., Raupach, M. R., Regnier, P., Rödenbeck, C., Saito, S., Salisbury, J. E., SCHUSTER, U., Schwinger, J., Séférian, R., Segschneider, J., Steinhoff, T., Stocker, B. D., Sutton, A. J., Takahashi, T., Tilbrook, B., van der Werf, G. R., Viovy, N., Wang, Y. P., Wanninkhof, R., Wiltshire, A. and Zeng, N.: Global carbon budget 2014, Earth Syst. Sci. Data, 7(1), 47–85, doi:10.5194/essd-7-47-2015, 2015.

Lindsay, K., Bonan, G. B., Doney, S. C., Hoffman, F. M., Lawrence, D. M., Long, M. C., Mahowald, N. M., Moore, J. K., Randerson, J. T. and Thornton, P. E.: Preindustrial Control and 20th Century Carbon Cycle Experiments with the Earth System Model CESM1(BGC), J. Climate, 141006111735008, doi:10.1175/JCLI-D-12-00565.1, 2014.

Meehl, G. A., Moss, R., Taylor, K. E., Eyring, V., Stouffer, R. J., Bony, S. and Stevens, B.: Climate Model Intercomparisons: Preparing for the Next Phase, Eos Trans. AGU, 95(9), 77–78, doi:10.1002/2014EO090001, 2014.

Mignot, J., Swingedouw, D., Deshayes, J., Marti, O., Talandier, C., Séférian, R., Lengaigne, M. and Madec, G.: On the evolution of the oceanic component of the IPSL climate models from CMIP3 to CMIP5: A mean state comparison, Ocean Modelling, 72 IS -(0 SP - EP - PY - T2 -), 167–184, 2013.

Mikaloff Fletcher, S. E., Gruber, N., Jacobson, A. R., Gloor, M., Doney, S. C., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Müller, S. A. and Sarmiento, J. L.: Inverse estimates of the oceanic sources and sinks of natural CO 2and the implied oceanic carbon transport, Global Biogeochem. Cycles, 21(1), GB1010, doi:10.1029/2006GB002751, 2007.

Orr, J. C.: Global Ocean Storage of Anthropogenic Carbon, Gif-sur-Yvette, France. 2002.

Sarmiento, J. L. and Gruber, N.: Ocean Biogeochemical Dynamics, Princeton University Press. 2006.

Schwinger, J., Tjiputra, J. F., Heinze, C., Bopp, L., Christian, J. R., Gehlen, M., Ilyina, T., Jones, C. D., Salas-Mélia, D., Segschneider, J., Séférian, R. and Totterdell, I.: Nonlinearity of Ocean Carbon Cycle Feedbacks in CMIP5 Earth System Models, J. Climate, 27(11), 3869–3888, doi:10.1175/JCLI-D-13-00452.1, 2014.

Séférian, R., Bopp, L., Gehlen, M., Swingedouw, D., Mignot, J., Guilyardi, E. and Servonnat, J.: Multiyear predictability of tropical marine productivity, Proceedings of the National Academy of Sciences, 111(32), 11646–11651, 2014.

Séférian, R., Delire, C., Decharme, B., Voldoire, A., Salas y Mélia, D., Chevallier, M., Saint-Martin, D., Aumont, O., Calvet, J.-C., Carrer, D., Douville, H., Franchistéguy, L., Joetzjer, E. and Sénési, S.: Development and evaluation of CNRM Earth-System model – CNRM-ESM1, Geosci. Model Dev. Discuss., 8(7), 5671–5739, 2015.

Steinacher, M., Joos, F., Fr olicher, T. L., Bopp, L., Cadule, P., Cocco, V., Doney, S. C., Gehlen, M., Lindsay, K., Moore, J. K., Schneider, B. and Segschneider, J.: Projected 21st century decrease in marine productivity: a multi-model analysis, Biogeosciences, 7(3), 979–1005, doi:10.5194/bg-7-979-2010, 2010.

Stouffer, R. J., Weaver, A. J. and Eby, M.: A method for obtaining pre-twentieth century initial conditions for use in climate change studies, Clim Dyn, 23(3-4), 327–339, doi:10.1007/s00382-004-0446-5, 2004.

Stow, C. A., Jolliff, J., McGillicuddy, D. J. J., Doney, S. C., Allen, J. I., Friedrichs, M. A. M., Rose, K. A. and Wallheadg, P.: Skill assessment for coupled biological/physical models of marine systems, Journal of Marine Systems, 76, 4–15, doi:10.1016/j.jmarsys.2008.03.011, 2009.

Tjiputra, J. F., Olsen, A., Bopp, L., Lenton, A., Pfeil, B., Roy, T., Segschneider, J., Totterdell, I. and Heinze, C.: Long-term surface pCO 2 trends from observations and models, Tellus B; Vol 66 (2014), 66(2-3), 151–168, doi:10.1007/s00382-007-0342-x, 2014.

Vichi, M., Manzini, E., Fogli, P. G., Alessandri, A., Patara, L., Scoccimarro, E., Masina, S. and Navarra, A.: Global and regional ocean carbon uptake and climate change: sensitivity to a substantial mitigation scenario, Climate Dynamics, 37(9-10), 1929–1947, doi:10.1007/s00382-011-1079-0, 2011.

Wunsch, C. and Heimbach, P.: How long to oceanic tracer and proxy equilibrium? Quaternary Science Reviews, 27(7-8), 637–651, doi:10.1016/j.quascirev.2008.01.006, 2008.

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

Roland Séférian[1,*], Marion Gehlen[2], Laurent Bopp[2], Laure Resplandy[3,2], James C. Orr[2], Olivier Marti[2], John P. Dunne[4], James R. Christian[5], Scott C. Doney[6], Tatiana Ilyina[7], Keith Lindsay[8], Paul Halloran[9], Christoph Heinze[10,11], Joachim Segschneider[12], Jerry Tjiputra[11], Olivier Aumont[13], Anastasia Romanou[14,15]

roland seferian 28/1/16 19:34
Formatted: Superscript

[1] CNRM-GAME, Centre National de Recherches Météorologiques-Groupe d'Etude de l'Atmosphère Météorologique, Météo-France/CNRS, 42 Avenue Gaspard Coriolis, 31057 Toulouse, France

[2] LSCE/IPSL, Laboratoire des Sciences du Climat et de l'Environnement, Orme des Merisiers, CEA/Saclay 91198 Gif-sur-Yvette Cedex, France

[3] Scripps Institution of Oceanography, UCSD, La Jolla, CA, USA

4 Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey, USA

[5] Fisheries and Oceans Canada and Canadian Centre for Climate Modelling and Analysis, Victoria, B.C., Canada

[6] Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole MA, USA

[7] Max Planck Institute for Meteorology, Bundesstraße 53, 20146 Hamburg, Germany

[8] Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado

[9] College of Life and Environmental Sciences, University of Exeter, Exeter, EX4 4RJ, UK

[10] Geophysical Institute, University of Bergen and Bjerknes Centre for Climate Research, Bergen, Norway

[11] Uni Research Climate, Allegt. 55, 5007 Bergen and Bjerknes Centre for Climate Research, Bergen, Norway

[12] University of Kiel, Kiel, Germany

[13] Sorbonne Universités (UPMC, Univ Paris 06)-CNRS-IRD-MNHN, LOCEAN-IPSL Laboratory, 4 Place Jussieu, F-75005 Paris, France

[14] Dept. of Applied Math. and Phys., Columbia University, 2880 Broadway, New York, NY 10025, USA

[15] NASA-GISS, NY, USA

*Corresponding author address: Roland Séférian, CNRM-GAME, 42 av. Gaspard Coriolis 31100 Toulouse. E-mail: roland.seferian@meteo.fr

roland seferian 28/1/16 19:35
Formatted: Superscript

roland seferian 28/1/16 19:35
Formatted: Font:Times New Roman, 12 pt

roland seferian 28/1/16 19:35
Formatted: Font:Times New Roman, 12 pt, Superscript

roland seferian 28/1/16 19:35
Formatted: Font:Times New Roman, 12 pt

roland seferian 28/1/16 19:35
Formatted

Abstract

During the fifth phase of the Coupled Model Intercomparison Project (CMIP5) substantial efforts were made on the systematic assessment of the skill of Earth system models. One goal was to check how realistically representative marine biogeochemical tracer distributions could be reproduced by models. Mean-state assessments routinely compared model hindcasts to available modern biogeochemical observations. However, these assessments considered neither the extent of equilibrium in modeled biogeochemical reservoirs nor the sensitivity of model performance to initial conditions or to the spin-up protocols. Here, we explore how the large diversity in spin-up protocols used for marine biogeochemistry in CMIP5 Earth system models

roland seferian 28/1/16 09:08
Deleted: carried out

(ESM) contribute to model-to-model differences in the simulated fields. We take advantage of a 500-year spin-up simulation of IPSL-CM5A-LR to quantify the influence of the spin-up protocol on model ability to reproduce relevant data fields. Amplification of biases in selected biogeochemical fields (O_2, NO_3, Alk-DIC) is assessed as a function of spin-up duration. We demonstrate that a relationship between spin-up duration and assessment metrics emerges from our model results and is consistent when confronted against a larger ensemble of CMIP5 models. This shows that drift has implications on performance assessment in addition to possibly aliasing estimates of climate change impact. Our study suggests that differences in spin-up protocols could explain a substantial part of model disparities, constituting a source of model-to-model uncertainty. This requires more attention in future model intercomparison exercises in order to provide realistic ESM results on marine biogeochemistry and carbon cycle feedbacks.

roland seferian 28/1/16 09:08
Deleted: their

roland seferian 26/1/16 14:07
Deleted: c

1- Introduction

1-1 Context

Earth system models (ESM) are recognized as the current state-of-the-art global coupled models used for climate research (e.g., Hajima et al., 2014; IPCC, 2013). They expand the numerical representation of the climate system used during the 4th IPCC assessment report (AR4) that was limited to coupled physical general circulation models, to the inclusion of biogeochemical and biophysical interactions between the physical climate system and the biosphere. ESMs that contributed to CMIP5 substantially differ in terms of their simulations of physical and biogeochemical components. These differences in design translate into a significant variability of the models' ability to reproduce the observed biogeochemistry and carbon cycle, which in turn may impact projected climate change responses (IPCC, 2013).

roland seferian 28/1/16 17:45
Deleted: climate

In the typical objective evaluation and intercomparison of these models, a suite of standardized statistical metrics (e.g., correlation, root-mean-squared errors) is applied to quantify differences between modeled and observed variables (e.g., Doney et al., 2009; Rose et al., 2009; Stow et al., 2009; Romanou et al., 2014; 2015). With the goal of constraining future projections, statistical metrics are often used for model ranking (e.g., Anav et al., 2013), weighting of model projections (e.g., Steinacher et al., 2010) or selection of the most skillful models across a wider ensemble (e.g., Cox et al., 2013; Massonnet et al., 2012; Wenzel et al., 2014). Most of these approaches can be considered as "blind" given that they are routinely applied without considering models' specific characteristics and treat models *a priori* as equivalently independent of observations. However, since these models are typically initialized from observations, the spin-up procedure of climate variables are the most model-dependent protocols that could introduce errors or drifts in modeled fields with consequences on skill score metrics.

roland seferian 5/1/16 15:22
Deleted: initialization and equilibration

1-2 Initialization of biogeochemical fields and spin-up protocols in CMIP5

Ocean initialization protocols aim at obtaining stable and equilibrated distributions of model state variables, such as temperature or concentrations of dissolved tracers. Most commonly used initialization protocols consist of initializing both physical and biogeochemical variables with either climatologies of the observed fields or constant values before running the model to equilibrium. In theory, equilibrium corresponds to steady-state and, hence, temporal derivatives of tracer fields close to zero. The time needed to equilibrate tracer distributions or, in other words, the integration time needed by the model to converge towards its own attractor (which is different from the true state of the climate system) varies greatly between components of the climate system. It spans from several weeks for the atmosphere (e.g., Phillips et al., 2004) to several centuries for ocean and sea ice components (e.g., Stouffer et al., 2004). The equilibration of ocean biogeochemical tracers across the entire water column amounts to several thousands of years (e.g., Heinze et al., 1999; Wunsch and Heimbach, 2008) and depends on the state of background ocean circulation as well as the turbulent mixing and eddy stirring parameterizations (e.g., Aumont et al., 1998; Bryan, 1984; Gnanadesikan, 2004; Marinov et al., 2008). In practice, these simulations, called "spin-up", span in general only several hundreds of years at the end of which a quasi-equilibrium state is assumed for the interior ocean tracers.

roland seferian 28/1/16 17:47
Deleted: diapycnal diffusivity

The present degree of complexity and increasing spatial as well as temporal resolution of marine biogeochemical ESM components, however, often precludes a spin-up to reach adequate equilibration of biogeochemical tracers. This is a consequence of the increasing number of state variables present in most of the current generation of biogeochemical models (e.g., for each tracer a separate advection equation has to be solved via a numerical CPU time demanding algorithm), more complex process descriptions (e.g., including more plankton functional types than before), and increasing spatial as well as temporal resolution. This number has continuously increased from simple biogeochemical models (e.g., HAMOCC3, Maier-Reimer and Hasselmann (1987)) to marine biodiversity models (e.g., Follows et al., 2007). Current generation biogeochemical models embedded in CMIP5 ESMs contain roughly two to four times more state variables than the physical models (e.g., atmosphere, ocean, sea-ice), which makes their equilibration computationally costly and difficult. The initialization of biogeochemical state variables is further complicated by the scarcity of biogeochemical observations as compared to observations of physical variables (e.g., temperature, salinity). While three-dimensional observation-based climatologies exist for macro-nutrients, oxygen, dissolved carbon and alkalinity, for other tracers such as dissolved iron, dissolved organic carbon and biomass of the various plankton functional types data are still sparse and represent measurements done over different time periods and climate conditions (in-spite of considerable efforts such as the GEOTRACES program for trace elements, or MAREDAT for biomasses of plankton functional types). The latter are initialized either with constant values (e.g. global average estimates) or with output from a previous model run. An additional difficulty stems from the use of modern climatologies to initialize the ocean state, implicitly assuming a long-term steady state, which does not necessarily represent the preindustrial state of the ocean. These climatologies incorporate the ongoing anthropogenic perturbation of marine biogeochemical fields, be it the uptake of anthropogenic CO_2 or the excess of nutrients inputs and pollutants (e.g., Doney, 2010). Although methods exist to remove the anthropogenic perturbation from observed ocean carbon tracer fields, their use is still debated since they lead to non-unique results (e.g., Tanhua et al., 2007; Yool et al., 2010).

The equilibration of marine biogeochemical tracer distributions is driven not only by the ocean circulation but also by numerous internal biogeochemical processes acting at various time scales. For example, while the transport and degradation of sinking organic matter spans days to perhaps several months, the associated impact on deep water chemistry accumulates over several decades to centuries as zones of differential remineralization are mixed across water masses and follows the ocean circulation (Wunsch and Heimbach, 2008). For models including interactive sediment modules, the sediment equilibration takes even longer (*O*(10^4) years; e.g., Archer et al. (2009) and Heinze et al. (1999)). As a consequence of the interplay between ocean circulation and biogeochemical processes, biogeochemical models require long spin-up times to equilibrate (e.g., Khatiwala et al., 2005; Wunsch and Heimbach, 2008). Modeling studies of paleo-oceanographic passive tracers such as $\delta^{18}O$ or $\Delta^{14}C$ (Duplessy et al., 1991), or global ocean passive tracers (Wunsch and Heimbach, 2008), as well as more recently available modern global scale data compilations (e.g., Key et al., 2004; Sarmiento and Gruber, 2006) and GEOTRACES Intermediate Data product 2014 (Version 2) http://www.bodc.ac.uk/geotraces/data/idp2014/) provide an estimate of the time required for the ocean biogeochemical reservoir to equilibrate with the climate systems (excluding continental weathering and reaction with marine sediments). Depending on ocean circulation, it ranges from 1500 years for subsurface water masses to 10000 years for the deep water masses (Wunsch and Heimbach, 2008).

In a context of model-to-model intercomparison, this time range contributes to the model uncertainty. Lessons from the previous OCMIP-2 exercise have demonstrated that some models required ~10,000 years to equilibrate to a global sea-air carbon flux of 0.01 Pg C y^{-1}.

While it is recognized that long time-scale processes influence the length of spin-up to equilibrium, the spin-up duration is usually defined *ad hoc* based on external

roland seferian 28/1/16 19:39
Formatted: Font:Italic

roland seferian 5/1/16 15:51
Deleted: It ranges from

roland seferian 5/1/16 15:51
Deleted: to 4000

roland seferian 5/1/16 15:52
Deleted: depending on the ocean circulation and can reach up to

roland seferian 23/1/16 12:39
Deleted: in

roland seferian 23/1/16 12:39
Deleted: deeper domains

roland seferian 5/1/16 15:52
Deleted: of the ocean

constraints or internal biogeochemical criteria. The computational cost is commonly invoked as external constraint to shorten and limit the spin-up duration. It is directly related to model complexity (e.g., Tjiputra et al., 2013; Vichi et al., 2011; Yool et al., 2013) and spatial resolution (Ito et al., 2010). The internal biogeochemical criteria applied to derive the duration of the spin-up simulations are generally defined by (i) reaching a steady-state, quasi equilibrium of the long-term global-mean CO_2 fluxes between the ocean and the atmosphere (e.g., Dunne et al., 2013; Ilyina et al., 2013; Lindsay et al., 2014; Romanou et al., 2013; Séférian et al., 2013), (ii) determining the amount of carbon stored into the ocean at preindustrial state (e.g., Dunne et al., 2013; Vichi et al., 2011) or (iii) representing relevant biogeochemical tracer patterns (e.g., oxygen minimum zone in Ito and Deutsch (2013)).

Despite its importance, only limited information on spin-up procedures is available through the CMIP5 metadata portal (http://metaforclimate.eu/trac). Information on spin-up protocols and model initialization is usually not taken into account in model intercomparison studies (e.g., Andrews et al., 2013; Bopp et al., 2013; Cocco et al., 2013; Frölicher et al., 2014; Gehlen et al., 2014; Keller et al., 2014; Resplandy et al., 2013; 2015; Rodgers et al., 2014; Séférian et al., 2014). This information, if available, can only be found separately in the reference papers of individual models (e.g., Adachi et al., 2013; Arora et al., 2011; Collins et al., 2011; Dunne et al., 2013; Ilyina et al., 2013; Lindsay et al., 2014; Romanou et al., 2013; Séférian et al., 2013; Séférian et al., 2015; Tjiputra et al., 2013; Vichi et al., 2011; Volodin et al., 2010; Watanabe et al., 2011; Wu et al., 2013). The duration of spin-up simulations of CMIP5 ocean biogeochemical components spans from one hundred years (e.g., CMCC-CESM) to several thousand years (e.g., MPI-ESM-LR, MPI-ESM-MR) (Figure 1 and Table 1).

Model initialization and spin-up procedures are equally variable across the model ensemble (Figure 1 and Table 1). Four different sources of initialization and four different procedures of model equilibration emerge from the 24 ESMs reviewed for this study.

roland seferian 26/1/16 14:08
Deleted: ive

roland seferian 26/1/16 14:08
Deleted: 3

Biogeochemical state variables were mostly initialized from observations, although from various releases of the same World Ocean Atlas global climatology (WOA1994, WOA2001, WOA2006, WOA2010). A small subset of ESMs relied either on a mix between previous model output and observations or solely on model output from a previous simulation for initialization. Similarly, spin-up procedures fall into two categories. The first one may be called "sequential": it consists in decomposing the spin-up integration into one long offline simulation (~200-10000 years) and one shorter online simulation (~100-1000 years). During the offline simulation, the biogeochemical model is forced by dynamical fields from the climate model or from reanalysis (CanESM2, MRI-ESM, Figure 1 and Table 1). Some modeling groups have adopted a "direct" strategy, which consists in running solely one online or coupled spin-up simulation (e.g., CNRM-ESM1, GFDL-ESM2M, GFDL-ESM2G, GISS-E2-H-CC, GISS-E2-R-CC, NorESM1-ME). Finally, a spin-up "acceleration" procedure is used by CMCC-CESM. This technique consists of enhancing the ocean carbon outgassing to remove anthropogenic carbon from the ocean, a legacy from initialization with modern data (Global Data Analysis Project or GLODAP following Key et al., 2004). None of these spin-up procedures, durations and sources of initialization can be considered as "standard"; each of them is unique and subjectively employed by one modeling group.

Objective arguments and hypotheses justifying the choice of one method of spin-up rather than the others have been the focus of previous studies (e.g., Dunne et al., 2013; Heinze and Ilyina, 2015; Tjiputra et al., 2013). Similarly, modeling groups discussed impacts of their particular spin-up procedure on model performance (e.g., Dunne et al., 2013; Lindsay et al., 2014; Séférian et al., 2013; Vichi et al., 2011). However, no study has addressed the potential for the large diversity of spin-up procedures found across the CMIP5 ensemble to translate into model-to-model differences in terms of comparative model performance assessments or model evaluations in terms of future projections.

1-3 Objectives of this study

This study assesses the role of the spin-up protocol in the representation of biogeochemical fields and subsequent model skill assessment, providing a complementary analysis from the studies of Sen Gupta et al. (2012; 2013). It relies on a 500-year long spin-up simulation from a state-of-the-art Earth system model, IPSL-CM5A-LR to investigate the impacts of spin-up strategy on selected biogeochemical tracers and residual model drift across the various ESMs of the CMIP5 ensemble. We demonstrate that the duration of the spin-up has implications for the determination of robust and meaningful skill-score metrics that should improve future intercomparison studies such as CMIP6 (Meehl et al., 2014).

Section 2 describes the model, the observations, the model experiments, as well as the methods used for assessing the impacts of spin-up protocols on the representation of biogeochemical fields in IPSL-CM5A-LR, as well as across the ensemble of CMIP5 ESMs. Section 3 presents the analysis developed for the assessment of the impact of spin-up duration on the representation of biogeochemical structures. Implications and recommendations are discussed in Sections 4 and 5, respectively.

roland seferian 26/1/16 14:09
Deleted: respectively

2- Methods

2-1- Model simulations

This study exploits in particular results from one simulation performed with IPSL-CM5A-LR (Dufresne et al., 2013) as representative for other CMIP5 Earth system models. As a typical representative of the current generation of ESMs, IPSL-CM5A-LR combines the major components of the climate system (Chap 9, Table 9.1, (IPCC, 2013). The atmosphere is represented by the atmospheric general circulation model LMDZ (Hourdin et al., 2006) with a horizontal resolution of 3.75°x1.87° and 39 levels. The land surface is simulated with ORCHIDEE (Krinner et al., 2005). The oceanic component is NEMOv3.2 in its ORCA2 global configuration (Madec, 2008). It has a horizontal resolution of about 2° with enhanced resolution at the equator (0.5°) and 31 vertical levels. NEMOv3.2 includes the sea-ice model LIM2 (Fichefet and Maqueda, 1997), and the marine biogeochemistry model PISCES (Aumont and Bopp, 2006). PISCES simulates the biogeochemical cycles of oxygen, carbon and the main nutrients with 24 state variables. The model simulates dissolved inorganic carbon and total alkalinity (carbonate alkalinity + borate + water) and the distributions of macronutrients (nitrate and ammonium, phosphate, and silicate) and micronutrient iron. PISCES represents two sizes of phytoplankton (i.e., nanophytoplankton and diatoms) and two zooplankton size-classes: microzooplankton and mesozooplankton. PISCES simulates semi-labile dissolved organic matter, and small and large sinking particles with different sinking speeds (3 m d^{-1} and 50 to 200 m d^{-1}, respectively). While fixed elemental stoichiometric C:N:P-O_2 ratios after Takahashi et al. (1985) are

roland seferian 15/1/16 16:11
Formatted: Justified, Space Before: 0 pt, After: 0 pt, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers, Tabs: 0,99 cm, Left + 1,98 cm, Left + 2,96 cm, Left + 3,95 cm, Left + 4,94 cm, Left + 5,93 cm, Left + 6,91 cm, Left + 7,9 cm, Left + 8,89 cm, Left + 9,88 cm, Left + 10,86 cm, Left + 11,85 cm, Left

imposed for these three compartments the internal concentrations of iron, silica and calcite are simulated prognostically . The carbon system is represented by dissolved inorganic carbon, alkalinity and calcite. Calcite is prognostically simulated following Maier-Reimer (1993) and Moore et al. (2002). Alkalinity in the model system includes the contribution of carbonate, bicarbonate, borate, protons, and hydroxide ions. Oxygen is prognostically simulated. The model distinguishes between oxic and suboxic remineralization pathways, the former relying on oxygen as electron acceptor, the latter on nitrate. For carbon and oxygen pools, air-sea exchange follows the Wanninkhof (1992) formulation.

roland seferian 16/1/16 11:39
Formatted: English (US)

roland seferian 15/1/16 16:11
Deleted: Oxygen is prognostically simulated using two different oxygen-to-carbon ratios, one for the oxic remineralization of organic matter and one for the sub-oxic pathway (Sarmiento and Gruber, 2006). F

roland seferian 15/1/16 16:11
Formatted: Font:(Default) Times, Font color: Blue, French

The boundary conditions account for nutrient supplies from three different sources: atmospheric dust deposition for iron, phosphorus and silica (Jickells and Spokes, 2001; Moore et al., 2004; Tegen and Fung, 1995), rivers for nutrients, alkalinity and carbon (Ludwig et al., 1996) and sediment mobilization for sedimentary iron (de Baar and de Jong, 2001; Johnson et al., 1999). To ensure conservation of nitrogen in the ocean, annual total nitrogen fixation is adjusted to balance losses from denitrification. For the other macronutrients, alkalinity and organic carbon, the conservation is ensured by tuning the sedimental loss to the total external input from rivers and dust. In PISCES, an adequate treatment of external boundary conditions has been demonstrated to be essential for the accurate simulation of nutrient distributions (Aumont and Bopp, 2006; Aumont et al., 2003). Riverine carbon inputs induce a natural outgassing of carbon of 0.6 Pg C y^{-1} which has been shown essential to model the inter-hemispheric gradient of atmospheric CO_2 under preindustrial state (Aumont et al., 2001).

The core simulation of this study is a 500-year long coupled preindustrial run. It uses the same atmospheric, land surface and ocean configurations as IPSL-CM5A-LR (Dufresne et al., 2013) for which the marine biogeochemistry has been extensively evaluated (see e.g., Séférian et al. (2013) for modern-state evaluation). The only difference between the “standard” preindustrial simulation contributed to CMIP5 and the present one is the initial conditions. While the CMIP5 preindustrial simulation starts from an ocean circulation after several thousand years of online physical adjustment, the present simulation starts from an ocean at rest using the January temperature and salinity fields from the World Ocean Atlas (Levitus and Boyer, 1994). Biogeochemical state variables were initialized from data compilations or climatologies as explained in the following section. Atmospheric CO_2 and other greenhouse gases, as well as natural aerosols, were set to their 1850 preindustrial values. The simulation is extensively described in terms of ocean physics by Mignot et al. (2013). Mignot and coworkers show that the strength of the Atlantic meridional overturning circulation and the Antarctic circumpolar current as well as the upper 300 m ocean heat content stabilize after 250 years of simulation.

Although the spin-up protocol used to conduct this 500-year long simulation is not readily comparable to the one used to produce the initial conditions for the CMIP5 preindustrial simulation, its duration is greater than the median length of on-line adjustment computed from the multiple spin-up protocols applied during CMIP5 (~395 years, Figure 1 and Table 1). Besides, the methodology of initializing biogeochemical state variables from data fields is not broadly employed by the various modeling groups that have contributed to CMIP5. Despite the above-mentioned methodological shortcuts, we take this 500-year long preindustrial simulation as a representative example of a spin-up protocol for the diversity of approaches used by CMIP5 models.

2-2- Observations for initialization and evaluation

Two streams of data sets were used in this study. The first stream combines data from the World Ocean Atlas 1994 (WOA94, Levitus and Boyer (1994) and Levitus et al., (1993)) for the initialization of 3-dimensional fields of temperature and salinity, dissolved nitrate, silicate, phosphate and oxygen, and data from GLODAP (Key et al., 2004) for preindustrial dissolved inorganic carbon and total alkalinity. This stream of data was chosen purposely in our experimental setup to be slightly different than the second stream of data, World Ocean Atlas 2013 (WOA2013, Levitus et al. (2013)), the evaluation data set.

A second stream of data was used to compare modeled biogeochemical fields. It includes up-to-date observed climatologies of nitrate and oxygen from the WOA2013. This database is based on samples collected since 1965, and incorporates also data from WOA94 onwards. For the concentrations of preindustrial dissolved inorganic carbon and total alkalinity, we still use GLODAP. The second stream of data was selected to be as close as possible to the "standard" evaluation procedure of skill-assessment protocols found in CMIP5 model reference papers (Adachi et al., 2013; Arora et al., 2011; Collins et al., 2011; Dunne et al., 2013; Ilyina et al., 2013; Lindsay et al., 2014; Romanou et al., 2013; Séférian et al., 2013; Séférian et al., 2015; Tjiputra et al., 2013; Vichi et al., 2011; Volodin et al., 2010; Watanabe et al., 2011; Wu et al., 2013). Differences between these two streams of data are minor and are further detailed below.

2-3- Approach and statistical analysis

To quantify the impacts of a large diversity of spin-up procedures on the representation of biogeochemical fields in CMIP5, we employ a three-fold approach.

(1) The 500-year long spin-up simulation described in Section 2.1 is used to determine the influence of the spin-up procedure on the representation of biogeochemical fields in IPSL-CM5A-LR.

(2) In the next step, relationships between biases in modeled fields, model-data mismatches and the duration of the spin-up simulation are identified across the CMIP5 ensemble. For this step, drifts in biogeochemical fields are determined from the first century of the preindustrial simulation (referred to as *piControl*) of each CMIP5 ESM.

(3) Finally, the various ensemble of modern hindcast (referred to as *historical*) from each available CMIP5 ESM are used to estimate the impact of these drifts in biogeochemical fields on the ability of models to replicate modern observations. For a given model, we use the ensemble average of the available 'historical' members if several realizations are available.

For this purpose, several statistical skill score metrics are computed following Rose et al. (2009) and Stow et al. (2009) from model fields interpolated on a regular 1° grid and to fixed depth levels. The skill score metrics are (1) the global averaged concentrations for overall drift; (2) the error or bias between modeled and observed fields at each grid-cell; (3) spatial correlation between model and observations to assess mismatches between modeled and observed large-scale structures; (4) the root-mean squared error (RMSE) to assess the total cumulative errors between modeled and observed fields. These statistical metrics are computed across the water column,

roland seferian 28/1/16 09:09
Deleted: is

roland seferian 7/1/16 12:59
Deleted: choose to

roland seferian 7/1/16 12:59
Deleted: first member

roland seferian 7/1/16 13:00
Deleted: ensemble

roland seferian 5/1/16 16:03
Formatted: Justified, Right: -0,04 cm, Tabs: 15,24 cm, Left

roland seferian 5/1/16 16:02
Deleted: The skill score metrics are (1) the global averaged concentrations for overall drift; (2) the average error or bias for mismatches between modeled and observed fields; (3) spatial correlation for mismatches between modeled and observed large-scale structures; (4) the difference between modeled and observed spatial variability ($\Delta\sigma$) and (5) the root-mean squared error (RMSE) to assess the total cumulative errors between modeled and observed fields.

but for clarity we focus on surface, 150 m (thermocline) and 2000 m (deep) levels. These statistical metrics were chosen among those described in the literature, because they proved to yield the most indicative scores for tracking model errors or improvement along the various intercomparison exercises (IPCC, 2013).

The drift is determined for either concentrations in simulated biogeochemical fields or for skill score metrics (e.g., RMSE) using a linear regression fit over a time window of 100 years. This time window of 100 years was chosen as a trade off between a longer time window (>200 years) that smoothes the drift signal and a shorter time window (<100 years) that introduces fluctuations due to internal variability and hence impacting the quality of the fit (see the assessment performed with the millennial-long CMIP5 *piControl* simulation of IPSL-CM5A-LR in Figure S1). The drift is assumed to decrease exponentially during the spin-up simulation and is described by a simple drift model:

$$drift(t) = drift(t=0) \times \exp(-\frac{1}{\tau}t) \qquad (1)$$

where τ is the relaxation time of the respective field at a given depth level. It corresponds to the time required to nullify the drift.

Our analyses focus on the global distribution of nitrate (NO_3), dissolved oxygen (O_2) and the difference between total alkalinity and dissolved inorganic carbon (Alk-DIC). The latter serves as an approximation of carbonate ion concentration following Zeebe and Wolf-Gladrow (2001). We use this approximation of the carbonate ion concentration rather than its concentration, $[CO_3^{2-}]$, since the latter was poorly assessed in CMIP5 reference papers and was not provided by a majority of ESMs. These three biogeochemical tracers were chosen because (1) most current

roland seferian 5/1/16 16:02
Formatted: French

roland seferian 11/1/16 19:20
Deleted: or Δσ

roland seferian 14/1/16 10:38
Deleted: 8

roland seferian 14/1/16 10:40
Formatted: Font:Italic

roland seferian 14/1/16 10:38
Deleted: .

biogeochemical models simulate Alk, DIC, NO_3 and O_2 prognostically and (2) they are frequently used in state-of-the-art model performance assessment (e.g., Anav et al., 2013; Bopp et al., 2013; Doney et al., 2009; Friedrichs et al., 2009; 2007; Stow et al., 2009), and (3) DIC and Alk are both used as “master tracers” for the carbonate system in the ocean biogeochemistry models (while $[CO_3^{2-}]$, e.g., is not explicitly advected as a tracer but diagnosed from temperature, salinity, DIC, Alk, $[H^+]$, and pCO_2 when needed) . Modeled distributions of NO_3, O_2 and Alk-DIC reflect the representation of biogeochemical processes related to the biological pump (CO_2, NO_3, O_2), the air-sea gas exchange and ocean ventilation (CO_2 and O_2), as well as carbonate chemistry (Alk-DIC). These biogeochemical processes are of particular relevance for investigating the impact of climate change on marine productivity (e.g., Henson et al., 2010), ocean deoxygenation (e.g., Gruber, 2011; Keeling et al., 2009) and the ocean carbon sink, processes for which future projections with the current generation of ESMs yield large inter-model spreads (e.g., Friedlingstein et al., 2013; Resplandy et al., 2015; Séférian et al., 2014; Tjiputra et al., 2014).

3 Results

3-1 Comparison of observational datasets

Our review of spin-up protocols for CMIP5 ESM shows that several modeling groups have employed different streams of datasets to initialize their biogeochemical models (e.g., WOA1994, WOA2001), while model evaluation relies on the most up-to-date stream of data. Differences between the two data streams used for initializing and assessing, respectively, NO_3 and O_2 concentrations are analyzed. Table 2 summarizes RMSE and correlation between WOA1994 and WOA2013 for these two biogeochemical fields.

Table 2 indicates that differences between the two streams of data are fairly small. The total difference (RMSE) represents a departure between 5 to 10% from the global average concentrations of WOA2013 across depth levels. It is generally lower in regions where the sampling density has not increased markedly between the two releases. These values can be used as a baseline for model-to-model comparison assuming that errors attributed to the various sources of initialization cannot be larger than 10%. Considering that some models have used outputs from previous model simulations or globally averaged concentrations as initial conditions, we acknowledge that this baseline is not a perfect criterion for benchmarking model performance. There is, however, no ideal solution to address this issue since there is no standardized set of initial conditions in CMIP5 except some recommendations for the decadal prediction exercise in which specific attention was paid to initialization (e.g., Keenlyside et al., 2008; Kim et al., 2012; Matei et al., 2012; Meehl et al., 2013; 2009; Servonnat et al., 2014; Smith et al., 2007; Swingedouw et al., 2013).

3-2 Equilibration state metrics in IPSL-CM5A-LR

The global mean sea surface temperature (SST) is a common metric to quantify the energetic equilibrium of the model. This metric has been widely used in various papers referenced in this study to determine the equilibration of ESM physical components. Figure 2a shows the evolution of this metric during the 500-year long spin-up simulation. The global average SST sharply decreases during the first 250 years of the simulation. In the last 250 years of the simulation, the global averaged SST displays a small residual drift of ~-10^{-4} °C y^{-1} which falls into the range of the drifts reported for CMIP5 ESMs. The evolution over the last 250 years is comparable to those of other physical equilibration metrics, such as the ocean heat content or the meridional overturning circulation (Mignot et al., 2013).

The temporal evolution of sea-to-air CO_2 fluxes was used in phase 2 of the Ocean Carbon Model Intercomparison Project (OCMIP-2, Orr (2002)) as an equilibration metric for the marine biogeochemistry and was still widely used during CMIP5. Figure 2b presents its evolution in the 500-year long spin-up simulation. The global ocean sea-to-air CO_2 flux is ~-0.7 Pg C y^{-1} over the last decades of the spin-up simulation (negative values indicate ocean CO_2 uptake). To assess the global sea-to-air carbon flux, we use the range of values estimated from preindustrial natural ocean carbon flux inversions (e.g. Gerber and Joos (2010) or Mikaloff Fletcher et al. (2007)). Since, these estimates do not account for the preindustrial carbon outgassing induced by the river input, while our model does, we have added a constant outgassing of 0.45 Pg C y^{-1} to the range of 0.03 ± 0.08 Pg C y^{-1} (Mikaloff Fletcher et al. 2007). This value of 0.45 Pg C y^{-1} corresponds to the global open-ocean river-induced carbon outgassing accordingly to IPCC (2013) or Le Quéré et al. (2015). Consequently, in our modeling framework, the target value of the global sea-to-air carbon flux ranges between 0.4 and 0.56 Pg C y^{-1}.

Figure 2b shows that the global sea-to-air carbon flux does not fit our range of values estimated from preindustrial natural ocean carbon flux inversions. Besides, Figure 2b shows that the drift in the global sea-to-air carbon flux reduces more slowly after a strong decline during the first 50 years of the spin-up simulation. While this drift is about 0.001 Pg C y^{-2} from year 250 to 500, it is weaker over the last century of the simulation (7x10^{-4} Pg C y^{-2}). Using a linear fit over the last century of the simulation

roland seferian 6/1/16 12:25
Deleted: T

roland seferian 6/1/16 12:26
Deleted: the

roland seferian 6/1/16 12:26
Deleted: (e.g. Gerber and Joos (2010) or Mikaloff Fletcher et al. (2007), referred to as MF2007 on Figure 2), which amounts to 0.03 ± 0.08 Pg C y^{-1} (to which an open-ocean river-induced carbon outgassing of 0.45 Pg C y^{-1} has to be added on Figure 2 accordingly to IPCC (2013) or Le Quéré et al. (2015). This indicates that the ocean carbon cycle has not reached a steady state in the model system following the 500 years of integration

roland seferian 6/1/16 12:27
Deleted:

roland seferian 6/1/16 12:27
Deleted: also

roland seferian 6/1/16 12:28
Deleted: i

roland seferian 6/1/16 12:28
Deleted: s much

roland seferian 6/1/16 12:28
Deleted: 5

roland seferian 6/1/16 12:32
Deleted:

roland seferian 6/1/16 12:33
Deleted: an approximate relaxation time of 73 years estimated from the simple drift model (Equation 1) over years 250-500

with a drift of $7x10^{-4}$ Pg C y^{-2}, we estimate that the simulated sea-to-air carbon flux would reach the range of 0.4-0.56 Pg C y^{-1} after 1100 to 1300 supplemental years of spin-up simulation. Our simple drift model (Equation 1) gives a relaxation time of around 160 years, which indicates that drift in ocean carbon flux should range between $2x10^{-7}$ and $7x10^{-7}$ Pg C y^{-2} after this 1100 to 1300 supplemental years of spin-up simulation.

These estimates do not account for the non-linearity of the ocean carbon cycle and the associated process uncertainties (Schwinger et al., 2014), and hence potentially underestimate the time required to equilibrate the ocean carbon cycle and sea-to-air carbon fluxes in the range of inversion estimates. The drift of 0.001 Pg C y^{-2} is, however, much smaller than the oceanic sink for anthropogenic carbon. Even if not fully equilibrated in terms of carbon balance, it is likely that this run would have given consistent estimates of anthropogenic carbon uptake in transient historical hindcasts.

3-3 Temporal evolution of model errors in IPSL-CM5A-LR

Figure 3 shows the temporal evolution of globally averaged concentrations for O_2, NO_3 and Alk-DIC at the surface (panels a, b and c), 150 m (panels d, e and f) and 2000 m (panels g, h, and i). Globally averaged concentrations of O_2, NO_3 and Alk-DIC (solid lines) reach steady state after 100 to 250 years of spin-up at the surface. While modeled nominal values for O_2 concentration converge toward the observed concentration (i.e., 172.3 μmol L^{-1}), that of NO_3 and to a lesser extent Alk-DIC present persistent deviations from WOA2013 and GLODAP. At the surface, the convergence of the simulated oxygen to observed value is expected since the

roland seferian 6/1/16 12:33
Deleted: and the

roland seferian 6/1/16 12:34
Deleted: 0.001 Pg C y^{-2}

roland seferian 6/1/16 12:34
Deleted: find

roland seferian 6/1/16 12:51
Deleted: a steady state after ~500 supplemental years of spin-up. After the additional 500 supplemental years of spin-up, sea-air carbon flux would fall into the range of inverse estimates of MF2007 with accounting for outgassing of river carbon of 0.45 Pg C y^{-1}.

roland seferian 6/1/16 12:53
Deleted: is

roland seferian 6/1/16 12:53
Deleted: es

roland seferian 6/1/16 12:53
Deleted: . This estimation

roland seferian 28/1/16 09:12
Deleted: s

roland seferian 6/1/16 12:54
Deleted: The duration of the spin-up simulation would have to be increased by an additional 500 years to account for the estimated river-induced natural CO_2 outgassing of about 0.45 Pg C y^{-1} (IPCC, 2013). T

roland seferian 6/1/16 12:55
Deleted: , however,

roland seferian 6/1/16 12:55
Deleted: E

dominant governing process of thermodynamic saturation (through the air-sea gas exchange) is well understood and modeled. The deviation in surface NO_3 highlights uncertainty related to near surface biological processes and upper ocean physics. Below the surface, concentrations of biogeochemical tracers drift away from the globally averaged concentrations computed from WOA2013 or GLODAP (Figure 3, panels d-i). At 150 and 2000 meters, the drift in global averaged concentrations for these fields, computed over the last 250 years, is still significant with $p<10^{-4}$ (Table 3). Dashed lines in Figure 3 indicate the temporal evolution of RMSE, which quantifies the total mismatch between simulated and observed fields. Except for the surface fields, Figure 3 shows that RMSE globally increases with time for all biogeochemical fields. The linear drift in RMSE over the last 250 years of the spin-up simulation falls within the 2-3 % ky^{-1} range at the surface. It is much larger at 2000 m (144-280 % ky^{-1} ; Table 3). This is also the case regionally, because the latitudinal maximum in RMSE ($RMSE_{max}$) is similar to the global RMSE. Table 3 also shows that the magnitude of drift in RMSE for O_2, NO_3 and Alk-DIC differs at a given depth as different processes affect the interior distribution of these biogeochemical fields.

3-4 Evolution of geographical mismatches in IPSL-CM5A-LR

To further explore the evolution of mismatch in biogeochemical distributions, we analyze differences (ε) between simulated and observed fields of O_2, NO_3 from WOA2013 and Alk-DIC from GLODAP after the initialization and at the end of the spin-up, i.e., the first year and the last year of the core spin-up simulation performed with the IPSL-CM5A-LR model (Figures 4, 5 and 6).

Figure 4 (panels a, c, and e) shows that surface concentrations of biogeochemical fields are associated with small biases at initialization. This error represents less than 5% of the observed surface concentrations for O_2, NO_3 and Alk-DIC and reflects the weak difference between the data stream employed for initialization and validation. After 500 years of spin-up, deviations between the modeled and observed fields at the surface have increased locally by up to ~40% (Figure 4, panels b, d, and f). The largest deviations are found in high-latitude oceans for O_2 and NO_3 and also to some extent in the tropics for NO_3 and Alk-DIC.

Below the surface, distributions of modeled biogeochemical fields compare well to the observations at 150 m at initialization with averaged errors close to zero (Figure 5, panels a, c, and e). This result was expected since WOA2013 and WOA1994 differ weakly at these depth levels. Subsurface distributions at initialization strongly contrast with the concentrations that resulted from 500 years of spin-up (Figure 5, panels b, d, and f). After 500 years of spin-up, strong mismatches characterize the distribution of O_2, NO_3 and Alk-DIC fields in the high-latitude oceans and in the tropics. Figure 5 illustrates that pattern of errors are well correlated. It directly translates the assumptions employed in the biogeochemical model (here the elemental C:N:-O_2 stochiometry of PISCES). Figure 6 shows that model-data deviations at 2000 m have substantially increased regionally after 500 years of simulation, showing large errors in the southern hemisphere oceans. This appears clearly in Figure 6, panels d and f for NO_3 and Alk-DIC fields, respectively.

The temporal evolution of the total mismatch between modeled and observed fields of O_2, NO_3 and Alk-DIC over the whole water column is presented in Figure 7 in terms of RMSE (Figure 7, panels a-c). As expected, Figure 7 illustrates that there is a good

roland seferian 6/1/16 14:23
Deleted: ac

roland seferian 6/1/16 14:23
Deleted: bd

roland seferian 26/1/16 14:09
Formatted: Subscript

roland seferian 6/1/16 14:24
Deleted: and the difference in spatial standard deviation between modeled and observed fields, Δσ (Figure 7, panels d-f)

match during the first years of simulation for all biogeochemical fields at all depth levels with low RMSE. After a few centuries, patterns of error evolve differently across depth for O_2, NO_3 and Alk-DIC.

The temporal evolution of RMSE shows that patterns of error have reached a steady state after few decades within the upper hundred meters of the ocean but continue to evolve at greater depths, even after 500 years. Patterns of errors within the thermocline and deep water masses evolve at time scales of few decades and few centuries, respectively in relation with the structure of the large-scale ocean circulation. Mid-depth (~1500-2500m) RMSE evolves much slower because this depth corresponds to the depth of the very old radiocarbon age (e.g., Wunsch and Heimbach, 2007; 2008) whose characteristics time scale spans over thousand of years. At the end of the spin-up simulation, two maxima of comparable amplitude are found for RMSE at 150 and 3750 m for O_2 and at 50 m and 3800 m for Alk-DIC.

3-5 Drifts in IPSL-CM5A-LR spin-up simulation

With the evolution of the RMSE established, we can use the simple drift model (Equation 1) to determine the relaxation time, τ, required to reach equilibration after a longer of spin-up simulation. To use this simple drift model, we compute the drift in RMSE determined from time segments of 100 years distributed evenly every 5 years from year 250 to 500 for O_2, NO_3 and Alk-DIC tracers. The drift model (magenta lines in Figure 8) is fitted level to the 80 drift values for each field and each depth (colored crosses in Figure 8).

The simple drift model fits well the evolution of the drift in RMSE for the biogeochemical variables along the spin-up simulation of IPSL-CM5A-LR (Figure 8).

roland seferian 6/1/16 14:24
Deleted: and Δσ

roland seferian 6/1/16 14:25
Deleted: Patterns of Δσ present different features than those of RMSE and demonstrate that the simulated fields substantially underestimate the spatial variation of observed biogeochemical fields on subsurface depth surfaces except for Alk-DIC.

roland seferian 6/1/16 14:25
Deleted: both metrics

roland seferian 13/1/16 15:36
Deleted:
From these two metrics, the simple drift model

roland seferian 13/1/16 15:36
Deleted: enables us

roland seferian 6/1/16 14:26
Deleted:

roland seferian 6/1/16 14:26
Deleted:

roland seferian 6/1/16 14:26
Deleted: over the last century

roland seferian 13/1/16 15:42
Formatted: Subscript

roland seferian 13/1/16 15:42
Formatted: Subscript

Correlation coefficients are mostly significant at 90% confidence level ($r^*=0.14$ determined with a student distribution with significance level of 90% and 80 degrees of freedom), except for NO_3 at surface and Alk-DIC at 150 m. Another exception is found for NO_3 at 150 m where the drift does not correspond to an exponential decay of the drift as function of time. The large confidence interval of the fit indicates that the fit would have been considered as non-significant given a longer spin-up simulation or a higher confidence threshold.

When significant, estimates of τ for O_2 RMSE are ≈ 90, 564 and 1149 y at the surface 150 m and 2000 m, respectively. These values match reasonably well τ estimated for NO_3 RMSE at 2000 m (1130 y) and those for Alk-DIC RMSE at surface and 2000 m (137 and 1163 y). However, these estimates are sensitive to the time windows used to compute the drift. For a subset of time windows between 100 and 250 years by step of 50 years, τ estimates for O_2 RMSE are ≈ 114±67, 375±140 and 1116±527 y at the surface 150 m and 2000 m depth. These large uncertainties associated with τ estimates are essentially due to the length of the spin-up simulation. A longer spin-up simulation would improve the quality of the fit (see Figure S1).

3-6 Drifts in CMIP5 ESMs preindustrial simulations

In this subsection, the analysis is extended to the CMIP5 archive. We focus on oxygen fields in the long preindustrial simulation, *piControl*, for the 15 available CMIP5 ESMs. From these simulations that span from 250 to 1000 years, we compute the drift in O_2 RMSE across depth from several time segments of 100 years distributed evenly every 5 years from the beginning until the end of the piControl simulation. These drifts are used as a surrogate for drift computed from the spin-up of each model since

- roland seferian 13/1/16 15:52 **Formatted** ... [1]
- roland seferian 6/1/16 14:27 **Deleted:** The relaxation times for oxygen RMSE… … about …4…13…4…0 ... [2]
- roland seferian 13/1/16 16:18 **Formatted:** Subscript
- roland seferian 6/1/16 14:28 **Deleted:** . Different values are derived for oxygen Δσ with 8, 7 and 46 y at surface, 150 and 2000 m, respectively. …V…alues for other biogeochemical fields are quite similar to those for O_2 except for NO_3 at 150 m… ... [3]
- roland seferian 13/1/16 16:19 **Formatted:** Font:(Default) Times New Roman
- roland seferian 6/1/16 14:33 **Deleted:** This contrasting result between the two skill score metrics expresses the fact that RMSE accounts for the total distance between modeled and observed oxygen distributions, while Δσ considers solely the difference in spatial structure between model fields and observations. This shows that the time scale for equilibration of spatial structure is not necessarily the same as the drift.
- roland seferian 13/1/16 15:34 **Deleted:** 5
- roland seferian 23/1/16 13:07 **Deleted:** we extend our…4…While the duration of CMIP5 *piControl* simulations ranges from 250 to 1000 years, we choose to…relative …and Δσ …over… the first century of simulation for each ESM (Figure 8). is… …is ... [4]

such simulations are not available through the data portal.

Figure 9 represents the drift in O_2 RMSE versus the spin-up duration for each CMIP5 ESM. The analysis shows that the drift in O_2 RMSE differs substantially between models. For a given model, drifts in other biogeochemical tracers (NO_3 and Alk-DIC) display similar features (not shown). The between-model differences in drift are not surprising since there are no reasons for different models to exhibit similar drift for a given field. Yet, Figure 9 shows that a global relationship emerges from this ensemble when using the simple drift model to fit the drift in O_2 RMSE as function of the spin-up duration (solid green lines in Figure 9). With a 90% confidence level, this relationship suggests a general decrease of the drift as a function of spin-up duration for all depth levels. At the surface and at 2000 m depth, the quality of fits is low with correlation coefficients of about ~0.4. These are however significant at 90% confidence level (r*=0.34 determined with a student distribution with significance level of 90% and 15 models as degree of freedom). The weakest correlation coefficient is found for the fit at 150 m depth and hence indicating that there is no link between the drift in O_2 RMSE and the duration of the spin-up simulation. This low significance level must be put into perspective given the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1) that can deteriorate the drift-spin up duration relationship in this ensemble of models.

The drift versus spin up duration relationship established from the 15 CMIP5 ESMs is nonetheless consistent with the results obtained with IPSL-CM5A-LR (The results in Figure 8 have been reported in Figure 9 with magenta crosses). Consistency is indicated by the sign of the drift versus spin up duration relationship of the IPSL-

Unknown
Deleted:

roland seferian 13/1/16 16:29
Deleted: 8…relative …(Figure 8, panels a, b, and c) and Δσ (Figure 8, panels d, e, and f) for O_2 concentrations … (log scale)… numerous models are far from equilibrium with a… …concentration …greater than 10% over the first century of the CMIP5 preindustrial simulation. …D…show …The magnitude of the drifts in RMSE and Δσ tends to increase with depth. It spans within 0-35% range at the surface up to a 0-400% range at 2000 m. …8…also …the best-fit linear regression relationship (or exponential in linear scale) between t… (or Δσ)…and…This relationship suggests a general decrease of the drift as a function of spin-up duration for all depth levels. This decrease in drift displays low significance. … the…generally …ranging between 0.3 and 0.7…The lower correlation coefficients fall outside the 90% significance thre…shold…6…4…C…of RMSE at 150 m is even lower than the 66% significance threshold (0.1…)…,…decrease in drift as a function of spin-up duration… ... [5]

roland seferian 23/1/16 13:12
Deleted: is…with…4… the… model…shown as solid red lines… in…8…, for which the relationship was (1) determined from overlapping time slices of 100 years from year 250 to 500 and (2) extrapolated over the 250-1190 spin-up duration range… ... [6]

CM5A-LR model at the various depth levels, although their magnitudes differ. This difference in magnitude is not surprising if one considers that drift is highly model and protocol dependent and that the length of the IPSL-CM5A-LR spin-up simulation is potentially too short to determine accurate estimates of the long-term drift in O_2 RMSE. Despite these differences, our analyses show that a relationship between the drift in O_2 RMSE versus the spin-up duration emerges from an ensemble of models and is broadly consistent with our theoretical framework of a drift model established from the results of the IPSL-CM5A-LR model (Figure 8).

3-7 Impact of the drift on model skill score assessment metrics across CMIP5 ESMs

In the following, we investigate the influence of model drift on skill score assessment metrics that are routinely used to benchmark model performance. For this purpose, we use the ensemble-mean O_2 RMSE as a metrics to assess the distance between the biogeochemical observations and model results. For this purpose, we compute O_2 RMSE from each ensemble member of the CMIP5 models averaged from 1986 to 2005 with respect to WOA2013 observations. The model-data distance is then determined for each CMIP5 model using the mean across the available ensemble members.

The left hand side panels of Figure 10 present the performance of available CMIP5 models in terms of distance to oxygen observations at the surface, 150 m and 2000 m, respectively. In these panels, the various CMIP5 models are ordered as function of their distance to the oxygen observations. Following Knutti et al. (2013), either the ensemble mean or the ensemble median is used to identify groups of models with similar skill within the CMIP5 ensemble. The left hand side panels of Figure 10 show

roland seferian 6/1/16 18:56
Deleted: . The importance of the spin-up is emphasized by the fact that the regression fit (red line on Figure 8 determined from the spin-up simulation) and data point (blue point on Figure 8 determined from the first century of the CMIP5 piControl simulation) derived from IPSL-CM5A-LR results differ at some depth levels.

roland seferian 13/1/16 15:34
Deleted: 6

roland seferian 13/1/16 16:37
Deleted: the

roland seferian 13/1/16 16:36
Deleted: normalized

roland seferian 13/1/16 16:37
Formatted: Subscript

roland seferian 13/1/16 16:37
Deleted:

roland seferian 13/1/16 16:40
Formatted: Subscript

roland seferian 13/1/16 16:41
Deleted: output from *historical* simulations of individual CMIP5 ESMs is computed following the approach widely employed to evaluate CMIP5 models (Adachi et al., 2013; Arora et al., 2011; Collins et al., 2011; Dunne et al., 2013; Ilyina et al., 2013; Lindsay et al., 2014; Romanou et al., 2013; Séférian et al., 2013; Tjiputra et al., 2013; Vichi et al., 2011; Volodin et al., 2010; Watanabe et al., 2011; Wu et al., 2013). The normalized distance consists of the scaled RMSE of model output averaged from 1985 to 2005.

roland seferian 13/1/16 16:35
Deleted: 9

roland seferian 11/1/16 12:09
Deleted: normalized distance to oxygen observations

roland seferian 11/1/16 12:11
Deleted: The normalized distance is used for model ranking

roland seferian 13/1/16 18:19
Deleted: 9

roland seferian 13/1/16 18:22
Deleted: s

that the ability of models to reproduce oxygen observations varies across depth levels.

The RMSE in the simulated O_2 fields in CESM1-BGC, HadGEM2-ES, HadGEM2-CC, GFDL-ESM2M, MPI-ESM-LR and MPI-ESM-MR is generally smaller than the ensemble mean or ensemble median RMSE across the various depth levels (Figure 10 panels a, b and c). On the other side of the ranking, CMCC-CESM, CNRM-CM5, CNRM-CM5-2, IPSL-CM5B-LR and NorESM1-ME exhibit RMSE generally higher than the ensemble mean and median RMSE across the various depth levels. The other models, i.e., CNRM-ESM1, GFDL-ESM2G, IPSL-CM5A-LR and IPSL-CM5A-MR display O_2 RMSE that is generally close to the ensemble mean or the ensemble median.

To assess the impact of model's drift inherited from the diversity of spin-up strategies (Figure 1 and Table 1) on the performance metrics, we use a simple additive assumption to incorporate an incremental error due to the drift, ΔRMSE, to the above-mentioned RMSE. This incremental error due to the drift is computed using the relaxation time τ determined from the *piControl* simulations of each CMIP5 model at each depth level (Equation 1 and Figure 9) and a common duration of T=3000 years for all models (*m*):

$$\Delta RMSE_m(z) = \int_0^T drift_m(z,t=0)\times \exp\left(-\frac{1}{\tau(z)}t\right)dt \qquad (2)$$

where ΔRMSE has the same unit as RMSE.

The common duration T is used to bring model drift close to zero and hence to make models comparable to each other.

We employ ΔRMSE to penalize the distance from the observations assuming that this drift-induced deviation in tracer fields can be added to RMSE. This means that the

roland seferian 11/1/16 13:48
Deleted: CC

roland seferian 11/1/16 13:48
Deleted: ES

roland seferian 11/1/16 13:51
Deleted: and CESM1-BGC

roland seferian 13/1/16 21:25
Deleted: For the HadGEM2 models, this result might arise from the fact that these models (1) are closer to the initial state since they have been spun up for a relatively short integration time or (2) display a better representation of the observed ocean. In either case, this would lead to a departure from the initial state that is smaller than for models that have a longer spin-up duration.

roland seferian 13/1/16 21:22
Deleted: , IPSL models

roland seferian 13/1/16 21:22
Deleted: CMCC-CESM

roland seferian 13/1/16 21:24
Formatted: Subscript

roland seferian 13/1/16 21:25
Deleted: model

roland seferian 7/1/16 12:28
Deleted: the incremental deviation due to drift in biogeochemical fields is estimated from the simple drift model (Equation 1). The incremental deviation, ΔRMSE,

roland seferian 13/1/16 18:20
Deleted: 8

roland seferian 7/1/16 12:41
Formatted: Font:Italic

roland seferian 7/1/16 12:28
Formatted: Font:(Default) Times

roland seferian 7/1/16 12:37
Formatted: Font:Times, Font color: Auto, English (US), Lowered by 13 pt

roland seferian 7/1/16 12:44
Deleted:

roland seferian 11/1/16 13:54
Deleted: normalized

effect of the penalty is to increase the distance giving a consistent measure of the equilibration error.

Right hand side panels of Figure 10 show the influence of this penalization approach on the model ranking at the various depth levels. They show that several models have been upgraded in the ranking while others have not. For example, both MPI-ESM-LR, MPI-ESM-MR have been upgraded at the surface and 2000 m. On the other hand, the rank of HadGEM2-ES and HadGEM2-CC has been downgraded to the 5th and 3rd position due to the large drift in surface oxygen concentrations in comparison to that of the other models. The surface drift might be attributed to drivers in oxygen fluxes (e.g., SST, SSS). The ranking of GFDL-ESM2G and GFDL-ESM2M slightly changes with penalization but both models stay close to the ensemble mean or the ensemble median. At the bottom of the ranking, models with large deviation from the oxygen observations (i.e., CMCC-CESM, IPSL-CM5B-LR, NorESM1-ME, CNRM-CM5) are found. For these models, the computed ΔRMSE and RMSE result in similar ranking, because even a small drift and hence relatively low ΔRMSE cannot compensate for their large RMSE.

4- Discussion

4-1 Implications for biogeochemical processes

Our results show that errors in ocean biogeochemical fields amplify during the spin-up simulation but not at the same rate at all depths. These differences in error evolution are consistent with an increasing contribution of biogeochemical processes in setting the distribution of tracers at depth. Indeed, Mignot et al. (2013) with the same model simulation showed that the large-scale ocean circulation reaches quasi-

roland seferian 11/1/16 13:54
Deleted: normalized

roland seferian 13/1/16 18:20
Deleted: 9

roland seferian 11/1/16 16:43
Deleted: The GFDL-ESM2G and MPI-ESM models that

roland seferian 11/1/16 16:44
Deleted: models

roland seferian 11/1/16 16:44
Deleted: 6

roland seferian 11/1/16 16:44
Deleted: 7

roland seferian 13/1/16 21:28
Deleted: .

roland seferian 11/1/16 18:59
Deleted: Implications

roland seferian 7/1/16 13:44
Deleted: and propagate

equilibrium after 250 years of spin-up, but our analyses indicate that biogeochemical tracers do not (Figure 3).

roland seferian 28/1/16 09:17
Deleted: z

roland seferian 26/1/16 14:11
Deleted: 1

Besides, our analysis demonstrates that error propagation and biogeochemical drift are highly model dependent. For example, despite having the same initialization strategy and comparable spin up duration, the GFDL-ESM2G, GFDL-ESM2M, and NorESM1-ME models display considerable difference in drift (Figures 9 and 10) that mirror large differences in model performance and properties (e.g., resolution, simulated processes).

roland seferian 26/1/16 14:11
Deleted: 8

roland seferian 26/1/16 14:11
Deleted: 9

The identification of the dynamical or biogeochemical processes responsible for these errors is not within the scope of this study and would required additional long simulations with additional tracers targeted for attribution of the various biogeochemical processes and the underlying ocean physics (e.g., Doney et al., 2004) involved (e.g. using abiotic, passive tracers as suggested in Walin et al. (2014)). Some mechanisms can be nonetheless invoked to explain differences or similarities in behavior between biogeochemical fields. For example, the evolution of surface concentrations for O_2 and Alk-DIC is controlled by the solubility of O_2 and CO_2 in seawater and the concentration of these gases in the atmosphere (set to the observed values and kept constant in all experiments performed with IPSL-CM5A-LR discussed here) and the biological soft-tissue and calcium carbonate counter pumps (in relation with the vertical transport of nutrients and alkalinity). Therefore, the equilibration of the O_2 and Alk-DIC surface fields once the physical equilibrium is reached (~250 years of spin-up) is expected (Figure 3, panels a and c and Figure 7). Nevertheless, spatial errors could increase depending on the physical state of the

roland seferian 26/1/16 14:11
Deleted: -

model (Figure 4, panels b and f). By contrast, the evolution of NO_3 concentration is predominantly determined by ocean circulation, biological processes, and to a lesser extent by external supplies from rivers and atmosphere. Below the surface, concentrations of O_2, NO_3, and Alk-DIC evolve in response to the combined effect of ocean circulation and biogeochemical processes. The combination of dynamical and biogeochemical processes on the one hand, and the spin-up strategy on the other hand both shape the modeled distributions of large-scale biogeochemical tracers.

Consequences of the difficulty in achieving the correct equilibration procedure are even larger for biogeochemical features that are defined by regional characteristics in tracer concentrations, such as high nutrient/low chlorophyll regions, oxygen minimum zones and nutrient-to-light colimitation patterns. This point is illustrated by recent studies focusing on future changes in phytoplankton productivity (e.g. Vancoppenolle et al. (2013) and Laufkötter et al. (2015). Vancoppenolle and co-workers report a wide spread of surface mean NO_3 concentrations (1980-1999) in the Arctic with a range from 1.7 to 8.9 μmol L^{-1} across a subset of 11 CMIP5 models. The spread in present day NO_3 concentrations translates into a large model-to-model uncertainty in future net primary production. Laufkötter and colleagues determined limitation terms of phytoplankton production for a subset of CMIP5 and MAREMIP (Marine Ecosystem Model Intercomparison Project) models. The authors demonstrate that nutrient-to-light colimitation patterns differ in strength, location and type between models and arise from large differences in the simulated nutrient concentrations. Although large differences between models were reported by Vancoppenolle et al. (2013) and Laufkötter et al. (2015) such as the spatial resolution and the complexity of biogeochemical models, differences in nutrient concentrations were identified as the largest source of model-to-model spread in addition to simply model error. The authors of both studies qualitatively invoked differences in spin-up duration to explain this spread. Besides, a recent assessment of interannual to decadal variability of ocean CO_2 and O_2 fluxes in CMIP5 models, suggests that decadal variability can range regionally from 10 to 50% of the total natural variability among a subset of 6 ESMs (Resplandy et al., 2015). In that study, the authors demonstrate that, despite the robustness of driving mechanisms (mostly related to vertical transport of water masses) across the model ensemble, model-to-model spread can be related to differences in modeled carbon and oxygen concentrations. In light of present results, it appears likely that differences in spin-up strategy and sources of initialization could also contribute to the amplitude of the natural variability of the ocean CO_2 and O_2 fluxes.

4-2 Implications for future projections

The inconsistent strategy to spin-up models in CMIP5 is a significant source of model uncertainty. It needs to be better constrained in order to draw robust conclusions on the impact of climate change on the carbon cycle as well as its climate feedback (e.g., Arora et al., 2013; Friedlingstein et al., 2013; Roy et al., 2011; Schwinger et al., 2014; Séférian et al., 2012) and on marine ecosystems (e.g., Bopp et al., 2013; Boyd et al., 2015; Cheung et al., 2012; Doney et al., 2012; Gattuso et al., 2015; Lehodey et al., 2006). So far, the most frequent approach relies on the use of long preindustrial control simulations to 'remove' the drift embedded in the simulated fields over the historical period or future projections (e.g., Bopp et al., 2013; Cocco et al., 2013; Friedlingstein et al., 2013; 2006; Frölicher et al., 2014; Gehlen et al., 2014; Keller et al., 2014; Steinacher et al., 2010; Tjiputra et al., 2014). Although this approach allows to determine relative changes, it does not allow to investigate the underlying reasons of the spread between models in terms of processes, variability and response to climate change. The "drift-correction" approach, much as the one used for this study, assumes that drift-induced errors in the simulated fields can be isolated from the signal of interest. Verification of this fundamental hypothesis would require a specific experimental set-up consisting of the perturbation of model fields (e.g., nutrients or carbon-related fields) to assess by how much the model projections would be modified. So far, several modeling groups have generated ensemble simulation in CMIP5 using a similar approach. However, the perturbations were applied either to physical fields only or to both the physical and marine biogeochemical fields. To assess impacts of different spin-up strategies and/or initial conditions on future projections of marine biogeochemical tracer distributions, ensemble simulations in which only biogeochemical fields are perturbed would be needed.

roland seferian 28/1/16 09:17
Deleted: in

4-3 Implications for multi-model skill-score assessments.

While the importance of spin-up protocols is well accepted in the modeling community, the link between spin-up strategy and the ability of a model to reproduce modern observations remains to be addressed.

Most of the recent CMIP5 skill assessment approaches were based on *historical* hindcasts that were started from preindustrial runs of varying duration and from various spin-up strategies. Therefore, in typical intercomparison exercises, Earth system models with a short spin-up, and hence modeled distributions still close to initial fields, are confronted with Earth system models with a longer spin-up duration and modeled distributions that have drifted further away from their initial states. Our study highlights that such inconsistencies in spin-up protocols and initial conditions across CMIP5 Earth system models (Figure 1 and Table 1) could significantly contribute to model-to-model spread in performance metrics. The analysis of the first century of CMIP5 *piControl* simulations demonstrated a significant spread of drift between CMIP5 models (Figure 9). An approximate exponential relationship between the amplitude of drift and the spin up duration emerges from the ensemble of CMIP5 models, which is consistent with results from IPSL-CM5A-LR. For example, while the global average root-mean square error increased up to 70% during a 500-year spin-up simulation with IPSL-CM5A-LR, its rate of increase (or drift) decreased with time to a very small rate (0.001 Pg C y^{-1}). Combining a simple drift model and this relationship, we propose a penalization approach in an effort to assess more objectively the influence of documented model differences on model-data biases. Figure 10 compares the state-of-the-art approach to assess model performance (left hand side panels) to the drift-penalized approach (right hand side panels). This novel approach penalizes models with larger drift without affecting the models with smaller drift. Taking into account drift in modeled fields results in subtle adjustments in ranking, which reflect differences in spin-up and initialization strategies.

4-4 Limitations of the framework

In this work, the analyses focus on the globally averaged O_2 RMSE across a diverse ensemble of CMIP5 models, which differ in terms of represented processes, spatial resolution and performance in addition to differences in spin-up protocols. Major limitations of the framework are presented below.

Due to their specificities in terms of processes and resolution (e.g., Cabré et al.,

roland seferian 11/1/16 17:09
Deleted: These inconsistencies result in systematic model biases, somewhat in line with Swart and Fyfe (2011), and affect model reliability.

roland seferian 26/1/16 14:12
Deleted: 8

roland seferian 11/1/16 18:38
Deleted: , suggesting that results should be reliable in the context of transient historical hindcast or future projections

roland seferian 26/1/16 14:12
Deleted: 9

roland seferian 11/1/16 18:40
Deleted: , while not inducing major changes in model ranking

roland seferian 15/1/16 10:48
Formatted: Subscript

roland seferian 15/1/16 11:17
Formatted: English (US)

roland seferian 15/1/16 12:39
Formatted: Justified, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbers, Tabs: 0,99 cm, Left + 1,98 cm, Left + 2,96 cm, Left + 3,95 cm, Left + 4,94 cm, Left + 5,93 cm, Left + 6,91 cm, Left + 7,9 cm, Left + 8,89 cm, Left + 9,88 cm, Left + 10,86 cm, Left + 11,85 cm, Left

(2015), Laufkötter et al. (2015)), regional drift in CMIP5 models may differ from the drift computed from globally averaged skill-score metrics (see Figure S2 and S3). These differences may lead to different estimates of the relaxation time τ at regional scale. Moreover, the combination of regional ocean physics and biogeochemical processes in each individual model may drive an evolution of regional drift in RMSE that does not fit the hypothesis of an exponential decay of the drift during the course of the spin-up simulation.

The above-mentioned remark can explain the relatively low confidence level of the fit to drift across the multi-model CMIP5 ensemble (Figure 9). The relatively low significance level of the fit directly reflects not only the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1) but also the large diversity of processes and resolution of the CMIP5 models. An improved derivation of the penalization would require access to output from spin-up simulations for each individual model or, at least, a better quantification of model-model differences in terms of initial conditions.

Finally, it is unlikely that model fields drift at the same rate along the spin-up simulation, even under the same spin-up protocols. Indeed, as shown in Kriest and Oschlies (2015), various parameterizations of the particles sinking speeds in a common physical framework may lead to a similar evolution of the globally averaged RMSE in the first century of the spin-up simulation but display very different behaviour within a time-scale of $O(10^3)$ years. As such, drift and τ estimates need to be used with caution when computed from short spin-up simulation because they can be subject to large uncertainties.

roland seferian 15/1/16 11:17
Formatted: English (US)

roland seferian 15/1/16 11:17
Formatted: English (US)

roland seferian 15/1/16 11:17
Formatted: English (US)

roland seferian 15/1/16 11:17
Formatted: English (US)

roland seferian 15/1/16 11:17
Formatted: English (US), Superscript

roland seferian 15/1/16 11:17
Formatted: English (US)

5- Conclusions and recommendation for future intercomparison exercises

Skill-score metrics are expected to be widely used in the framework of the future CMIP6 (Meehl et al., 2014) with the development of international community benchmarking tools like the ESMValTool (http://www.pa.op.dlr.de/ESMValTool , see also Eyring et al. (2015)). The assessment of model skill to reproduce observations will focus on the modern period. Complementary to this approach, our results call for the consideration of spin-up and initialization strategies in the determination of skill assessment metrics (e.g., Friedrichs et al., 2009; Stow et al., 2009) and, by extension, to model weighting (e.g., Steinacher et al., 2010) and model ranking (e.g., Anav et al., 2013). Indeed, the use of equilibrium-state metrics of the model like the 3-dimensional growth rate or drift of relevant skill score metrics (e.g. RMSE) could be employed to increase the reliability of these traditional metrics and, as such, should be included in the set of standard assessment tools for CMIP6.

In an effort to better represent interactions between marine biogeochemistry and climate (Smith et al., 2014), future generations of Earth system models are likely to include more complex ocean biogeochemical models, be it in terms of processes (e.g., Tagliabue and Völker, 2011; Tagliabue et al., 2011) or interactions with other biogeochemical cycles (e.g., Gruber and Galloway, 2008) or increased spatial resolution (e.g., Dufour et al., 2013; Lévy et al., 2012) in order to better represent mesoscale biogeochemical dynamics. These developments will go along with an increase in the diversity and complexity of spin-up protocols applied to Earth system models, especially those including an interactive atmospheric CO_2 or interactive nitrogen cycle (e.g., Dunne et al., 2013; Lindsay et al., 2014). The additional

roland seferian 15/1/16 11:53
Deleted: The low confidence level of the fit to drift across the multi-model CMIP5 ensemble is a limitation of our approach and the robustness of the fit used to penalize models might be criticized. The low significance level of the fit directly reflects the large diversity of spin-up protocols and initial conditions (Figure 1 and Table 1). The improved derivation of the penalization factor would require access to output from spin-up simulations for each individual model or, at least, a better quantification of model-model differences in terms of initial conditions. 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 (e.g., Friedrichs et al., 2009; Stow et al., 2009) and, by extension, to model weighting (e.g., Steinacher et al., 2010) and model ranking (e.g., Anav et al., 2013).

roland seferian 15/1/16 11:45
Deleted: In order to increase the reliability of these traditional metrics, additional metrics that allow us to determine the equilibrium state of the model like the 3-dimensional growth rate or drift of relevant skill score metrics (e.g., RMSE) over the last decades or centuries of the spin-up,

roland seferian 15/1/16 11:50
Deleted:

roland seferian 15/1/16 11:45
Formatted: Font:(Default) Times

challenge of spinning-up emission-driven simulations with interactive carbon cycle will also require us to extend the assessment of the impact of spin-up protocols to the terrestrial carbon cycle. Processes such as soil carbon accumulation, peat formation as well as shift in biomes such as tropical and boreal ecosystems for dynamic vegetation models require several long time-scales to equilibrate (Brovkin et al., 2010; Koven et al., 2015). In addition, the terrestrial carbon cycle has large uncertainties in terms of carbon sink/source behavior (Anav et al., 2013; Dalmonech et al., 2014; Friedlingstein et al., 2013) which might affect ocean CO_2 uptake (Brovkin et al., 2010). A novel numerical algorithm to accelerate the spin-up integration time for computationally expensive ocean biogeochemical models has emerged (Khatiwala, 2008), which could further complicate the determination of inter-model spreads.

To evaluate the contribution of variable spin-up and initialization strategies to model performance, these should be documented extensively and the corresponding model output should be archived. Ideally, for future coupled model intercomparision exercises (i.e., CMIP6, CMIP7, Meehl et al., (2014)), the community should agree on a set of simple recommendations for spin-up protocols, following past projects such as OCMIP-2. In parallel, any trade-off between model equilibration and computationally efficient spin-up procedures has to be linked with efforts to reduce model errors due to the physical and biogeochemical parameterizations.

roland seferian 15/1/16 11:52
Deleted: like

roland seferian 15/1/16 11:53
Deleted: initialization,

roland seferian 15/1/16 11:53
Deleted: and duration

Acknowledgement:

We sincerely thank I. Kriest, F. Joos, the anonymous reviewer and A Yool for their useful comments on this paper. *This work was supported by* *the H2020 project*

roland seferian 19/1/16 09:07
Formatted: Font:Italic

roland seferian 15/1/16 11:59
Formatted: Left, Indent: First line: 0,85 cm

roland seferian 19/1/16 09:07
Formatted: Font:Italic

CRESCENDO "Coordinated Research in Earth Systems and Climate : Experiments, kNowledge, Dissemination and Outreach" which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641816 and by the EU FP7 project CARBOCHANGE "Changes in carbon uptake and missions by oceans in a changing climate" which received funding from the European community's Seventh Framework Programme under grant agreement no. 264879. Supercomputing time was provided by GENCI (Grand Equipement National de Calcul Intensif) at CCRT (Centre de Calcul Recherche et Technologie), allocation 016178. Finally, we are grateful to the ESGF project which make data available for all the community. R.S. is grateful Aurélien Ribes for his kind advices on statistics. JT acknowledges ORGANIC project (239965/F20) funded by the Research Council of Norway. CH and TJ are grateful for support through project EVA - Earth system modelling of climate variations in the Anthropocene (229771/E10) funded by the Research Council of Norway, as well as CPU-time and mass storage provided through NOTUR project NN2345K as well as NorStore project NS2345K. K.L. and S.C.D. acknowledge support from the National Science Foundation.

roland seferian 15/1/16 11:59
Formatted: Font:Italic

roland seferian 15/1/16 11:59
Formatted: Font:Italic

roland seferian 15/1/16 11:59
Formatted: Font:12 pt, Not Bold, Italic

roland seferian 15/1/16 11:59
Formatted: Font:12 pt, Not Bold

roland seferian 15/1/16 11:59
Formatted: Font:12 pt, Not Bold, Italic

roland seferian 15/1/16 11:59
Formatted: Font:Not Bold

References:

Adachi, Y., Yukimoto, S., Deushi, M., Obata, A., Nakano, H., Tanaka, T. Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Shindo, E., Tsujino, H., Mizuta, R., Yabu, S., Koshiro, T., Ose, T. and Kitoh, A.: Basic performance of a new earth system model of the Meteorological Research Institute (MRI-ESM1), Papers in Meteorology and Geophysics, 64, 1–18, doi:10.2467/mripapers.64.1, 2013.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. and Zhu, Z.: Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models, J. Climate, 26(18), 6801–6843, doi:10.1175/JCLI-D-12-00417.1, 2013.

Andrews, O. D., Bindoff, N. L., Halloran, P. R., Ilyina, T. and Le Qu 'er 'e, C.: Detecting an external influence on recent changes in oceanic oxygen using an optimal fingerprinting method, Biogeosciences, 10(3), 1799–1813, doi:10.5194/bg-10-1799-2013, 2013.

Archer, D., Buffett, B. and Brovkin, V.: Ocean methane hydrates as a slow tipping point in the global carbon cycle, Proceedings of the National Academy of Sciences, 106(49), 20596–20601, 2009.

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., Bonan, G., Bopp, L., Brovkin, V., Cadule, P., Hajima, T., Ilyina, T., Lindsay, K., Tjiputra, J. F. and Wu, T.: Carbon–Concentration and Carbon–Climate Feedbacks in CMIP5 Earth System Models, J. Climate, 26(15), 5289–5314, doi:10.1175/JCLI-D-12-00494.1, 2013.

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W. G. and Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases, Geophys.

Res. Lett., 38(5), L05805, doi:10.1029/2010GL046270, 2011.

Aumont, O. and Bopp, L.: Globalizing results from ocean in situ iron fertilization studies, Global Biogeochem. Cycles, 20(2), GB2017, doi:10.1029/2005GB002591, 2006.

Aumont, O., Maier-Reimer, E., Blain, S. and Monfray, P.: An ecosystem model of the global ocean including Fe, Si, P colimitations, Global Biogeochem. Cycles, 17(2), 1060, doi:10.1029/2001GB001745, 2003.

Aumont, O., Orr, J. C., Monfray, P., Ludwig, W., Amiotte-Suchet, P. and Probst, J.-L.: Riverine-driven interhemispheric transport of carbon, Global Biogeochem. Cycles, 15(2), 393–405, doi:10.1029/1999GB001238, 2001.

Aumont, O., Orr, J., Jamous, D., Monfray, P., Marti, O. and Madec, G.: A degradation approach to accelerate simulations to steady-state in a 3-D tracer transport model of the global ocean, Climate Dynamics, 14(2), 101–116, 1998.

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Séférian, R., Tjiputra, J. and Vichi, M.: Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models, Biogeosciences, 10(10), 6225–6245, doi:10.5194/bg-10-6225-2013, 2013.

Boyd, P. W., Lennartz, S. T., Glover, D. M. and Doney, S. C.: Biological ramifications of climate-change-mediated oceanic multi-stressors, Nature Clim. Change, 5(1), 71–79, 2015.

Brovkin, V., Lorenz, S. J., Jungclaus, J., Raddatz, T., Timmreck, C., Reick, C. H., Segschneider, J. and Six, K.: Sensitivity of a coupled climate-carbon cycle model to large volcanic eruptions during the last millennium, Tellus B, 62(5), 674–681, doi:10.1111/j.1600-0889.2010.00471.x, 2010.

roland seferian 28/1/16 09:46
Deleted: Brovkin, V., LORENZ, S. J., JUNGCLAUS, J., RADDATZ, T., Timmreck, C., Reick, C. H., Segschneider, J. and SIX, K.

Bryan, K.: Accelerating the Convergence to Equilibrium of Ocean-Climate Models, J. Phys. Oceanogr., 14(4), 666–673, doi:10.1175/1520-0485(1984)014<0666:ATCTEO>2.0.CO;2, 1984.

Cheung, W. W. L., Sarmiento, J. L., Dunne, J. P., Frölicher, T. L., Lam, V. W. Y., Palomares, M. L. D., Watson, R. and Pauly, D.: Shrinking of fishes exacerbates impacts of global ocean changes on marine ecosystems, Nature Climate change, 2(10), 1–5, doi:10.1038/nclimate1691, 2012.

Cabré, A., Marinov, I., Bernardello, R. and Bianchi, D.: Oxygen minimum zones in the tropical Pacific across CMIP5 models: mean state differences and climate change trends, Biogeosciences, 12(18), 5429–5454, doi:10.5194/bg-12-5429-2015, 2015.

Cocco, V., Joos, F., Steinacher, M., Frölicher, T. L., Bopp, L., Dunne, J., Gehlen, M., Heinze, C., Orr, J., Oschlies, A., Schneider, B., Segschneider, J. and Tjiputra, J.: Oxygen and indicators of stress for marine life in multi-model global warming projections, Biogeosciences, 10(3), 1849–1868, doi:10.5194/bg-10-1849-2013, 2013.

roland seferian 15/1/16 16:18
Deleted:

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A. and Woodward, S.: Development and evaluation of an Earth-System model – HadGEM2, Geosci. Model Dev, 4(4), 1051–1075, doi:10.5194/gmd-4-1051-2011, 2011.

Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D. and Luke, C. M.: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, Nature, 494(7437), 341–344, doi:10.1038/nature11882, 2013.

Dalmonech, D., Foley, A. M., Anav, A., Friedlingstein, P., Friend, A. D., Kidston, M., Willeit, M. and Zaehle, S.: Challenges and opportunities to reduce uncertainty in projections of future atmospheric CO_2: a combined marine and terrestrial biosphere perspective, Biogeosciences Discuss., 11(2), 2083–2153, doi:10.5194/bgd-11-2083-2014, 2014.

de Baar, H. J. W. and de Jong, J. T. M.: The biogeochemistry of iron in seawater, edited by D. R. Turner and K. A. Hunter, John Wiley, Hoboken, N. J., 2001.

Doney, S. C. , Lindsay, K., Caldeira, K., Campin, J.-M., Drange, H., Dutay, J.-C., Follows, M., Gao, Y., Gnanadesikan, A., Gruber, N., Ishida, A., Joos, F., Madec, G., Maier-Reimer, E., Marshall, J. C., Matear, R. J., Monfray, P., Mouchet, A., Najjar, R., Orr, J. C., Plattner, G.-K., Sarmiento, J., Schlitzer, R., Slater, R., Totterdell, I. J., Weirig, M.-F., Yamanaka, Y. and Yool, A: Evaluating global ocean carbon models: The importance of realistic physics, Global Biogeochem. Cycles, 18(3), doi:10.1029/2003GB002150, 2004.

Doney, S. C.: The Growing Human Footprint on Coastal and Open-Ocean Biogeochemistry, Science, 328(5985), 1512–1516, doi:10.1126/science.1185198, 2010.

Doney, S. C., Lima, I., Moore, J. K., Lindsay, K., Behrenfeld, M. J., Westberry, T. K., Mahowald, N., Glover, D. M. and Takahashi, T.: Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data, Journal of Marine Systems, 76(1-2), 95–112, doi:10.1016/j.jmarsys.2008.05.015, 2009.

Doney, S. C., Ruckelshaus, M., Emmett Duffy, J., Barry, J. P., Chan, F., English, C. A., Galindo, H. M., Grebmeier, J. M., Hollowed, A. B., Knowlton, N., Polovina, J., Rabalais, N. N., Sydeman, W. J. and Talley, L. D.: Climate Change Impacts on Marine Ecosystems, Annu. Rev. Marine. Sci., 4(1), 11–37, doi:10.1146/annurev-marine-041911-111611, 2012.

Dufour, C. O., Sommer, J. L., Gehlen, M., Orr, J. C., Molines, J.-M., Simeon, J. and Barnier, B.: Eddy compensation and controls of the enhanced sea-to-air CO2 flux during positive phases of the Southern Annular Mode, Global Biogeochem. Cycles, 27(3), 950–961, doi:10.1002/gbc.20090, 2013.

Dufresne, J.-L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., Noblet, N., Duvel, J. P., Ethe, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P.,

roland seferian 26/1/16 14:14
Deleted: Doney, S. C.: Evaluating global ocean carbon models: The importance of realistic physics, Global Biogeochem. Cycles, 18(3), doi:10.1029/2003GB002150, 2004.

Grandpeix, J. Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M.-P., Lefèvre, F., Lévy, C., Li, Z. X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N. and Vuichard, N.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim Dyn, 40(9-10), 2123–2165, doi:10.1007/s00382-012-1636-1, 2013.

Dunne, J. P., John, J. G., Adcroft, A. J., Griffies, S. M., Hallberg, R. W., Shevliakova, E., Stouffer, R. J., Cooke, W., Dunne, K. A., Harrison, M. J., Krasting, J. P., Malyshev, S. L., Milly, P. C. D., Phillipps, P. J., Sentman, L. A., Samuels, B. L., Spelman, M. J., Winton, M., Wittenberg, A. T. and Zadeh, N.: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics, J. Climate, 25(19), 6646–6665, doi:doi: 10.1175/JCLI-D-11-00560.1, 2013.

Duplessy, J. C., Bard, E., Arnold, M., Shackleton, N. J., Duprat, J. and Labeyrie, L.: How fast did the ocean—atmosphere system run during the last deglaciation? Earth and Planetary Science Letters, 103(1-4), 27–40, doi:10.1016/0012-821X(91)90147-A, 1991.

Eyring, V., Righi, M., Evaldsson, M., Lauer, A., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt, K. D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., Kunert, D., Levine, R., Loew, A., Mäkelä, J., Martin, G., Mason, E., Phillips, A., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., van Ulft, L. H., Walton, J., Wang, S. and Williams, K. D.: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth System Models in CMIP, Geosci. Model Dev. Discuss., 8(9), 7541–7661, 2015.

Fichefet, T. and Maqueda, M. A. M.: Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics, J. Geophys. Res., 102(C6), 12609–12646, 1997.

Follows, M. J., Dutkiewicz, S., Grant, S. and Chisholm, S. W.: Emergent Biogeography of Microbial Communities in a Model Ocean, Science, 315(5820), 1843–1846, doi:10.1126/science.1138544, 2007.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Bloh, Von, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K. G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C. and Zeng, N.: Climate–Carbon Cycle Feedback Analysis: Results from the C 4MIP Model Intercomparison, J. Climate, 10(14), 3337–3353, doi:10.1175/JCLI3800.1, 2006.

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K. and Knutti, R.: Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks, J. Climate, 130917124100006, doi:doi: 10.1175/JCLI-D-12-00579.1, 2013.

Friedrichs, M. A. M., Carr, M.-E., Barber, R. T., Scardi, M., Antoine, D., Armstrong, R. A., Asanuma, I., Behrenfeld, M. J., Buitenhuis, E. T., Chai, F., Christian, J. R., Ciotti, A. M., Doney, S. C., Dowell, M., Dunne, J. P., Gentili, B., Gregg, W., Hoepffner, N., Ishizaka, J., Kameda, T., Lima, I., Marra, J., Mélin, F., Moore, J. K., Morel, A., O'Malley, R. T., O'Reilly, J., Saba, V. S., Schmeltz, M., Smyth, T. J., Tjiputra, J., Waters, K., Westberry, T. K. and Winguth, A.: Assessing the uncertainties of model estimates of primary productivity in the tropical Pacific Ocean, Journal of Marine Systems, 76(1-2), 113–133, doi:10.1016/j.jmarsys.2008.05.010, 2009.

Friedrichs, M. A. M., Dusenberry, J. A., Anderson, L. A., Armstrong, R. A., Chai, F., Christian, J. R., Doney, S. C., Dunne, J. P., Fujii, M., Hood, R., McGillicuddy, D. J., Jr., Moore, J. K., Schartau, M., Spitz, Y. H. and Wiggert, J. D.: Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups, J. Geophys. Res., 112(C8), doi:10.1029/2006JC003852, 2007.

Frölicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P. and Winton, M.: Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models, J. Climate, 141031131835005, doi:10.1175/JCLI-D-14-00117.1, 2014.

Gattuso, J. P., Magnan, A., Bille, R., Cheung, W. W. L., Howes, E. L., Joos, F., Allemand, D., Bopp, L., Cooley, S. R., Eakin, C. M., Hoegh-Guldberg, O., Kelly, R. P., Portner, H. O., Rogers, A. D., Baxter, J. M., Laffoley, D., Osborn, D., Rankovic, A., Rochette, J., Sumaila, U. R., Treyer, S. and Turley, C.: Contrasting futures for ocean and society from different anthropogenic CO2 emissions scenarios, Science, 349(6243), aac4722–aac4722, doi:10.1126/science.aac4722, 2015.

Gehlen, M., Séférian, R., Jones, D. O. B., Roy, T., Roth, R., Barry, J., Bopp, L., Doney, S. C., Dunne, J. P., Heinze, C., Joos, F., Orr, J. C., Resplandy, L., Segschneider, J. and Tjiputra, J.: Projected pH reductions by 2100 might put deep North Atlantic biodiversity at risk, Biogeosciences, 11(23), 6955–6967, 2014.

Gerber, M. and Joos, F.: Carbon sources and sinks from an Ensemble Kalman Filter ocean data assimilation - Gerber - 2010 - Global Biogeochemical Cycles - Wiley Online Library, Global Biogeochem. Cycles, 24, GB3004, doi:10.1029/2009GB003531, 2010.

Gnanadesikan, A.: Oceanic ventilation and biogeochemical cycling: Understanding the physical mechanisms that produce realistic distributions of tracers and productivity, Global Biogeochem. Cycles, 18(4), doi:10.1029/2003GB002097, 2004.

Gruber, N.: Warming up, turning sour, losing breath: ocean biogeochemistry under global change, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 369(1943), 1980–1996, doi:10.1098/rsta.2011.0003, 2011.

Gruber, N. and Galloway, J. N.: An Earth-system perspective of the global nitrogen cycle, Nature, 451(7176), 293–296, doi:doi:10.1038/nature06592, 2008.

Sen Gupta, A. S., Muir, L. C., Brown, J. N., Phipps, S. J., Durack, P. J., Monselesan,

D. and Wijffels, S. E.: Climate Drift in the CMIP3 Models, J. Climate, 25(13), 4621–4640, doi:10.1175/JCLI-D-11-00312.1, 2012.

Sen Gupta, A. S., Jourdain, N. C., Brown, J. N. and Monselesan, D.: Climate Drift in the CMIP5 models, J. Climate, 26(21), 8597–8615. http://doi.org/10.1175/JCLI-D-12-00521.s1.

Hajima, T., Kawamiya, M., Watanabe, M., Kato, E., Tachiiri, K., Sugiyama, M., Watanabe, S., Okajima, H. and Ito, A.: Modeling in Earth system science up to and beyond IPCC AR5, Progress in Earth and Planetary Science, 1(1), 29–25, doi:10.1186/s40645-014-0029-y, 2014.

Heinze, C., Maier-Reimer, E., Winguth, A. and Archer, D.: A global oceanic sediment model for long-term climate studies, Global Biogeochem. Cycles, 13(1), 221–250, 1999.

Heinze, M. and Ilyina, T.: Ocean biogeochemistry in the warm climate of the late Paleocene, Climate of the Past, 11(1), 1933–1975, doi:10.5194/cp-11-63-2015, 2015.

Henson, S. A., Sarmiento, J. L., Dunne, J. P., Bopp, L., Lima, I., Doney, S. C., John, J. and Beaulieu, C.: Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity, Biogeosciences, 7(2), 621–640, doi:10.5194/bg-7-621-2010, 2010.

Hourdin, F., Musat, I., Bony, S., Braconnot, P., Codron, F., Dufresne, J.-L., Fairhead, L., Filiberti, M.-A., Friedlingstein, P., Grandpeix, J.-Y., Krinner, G., LeVan, P., Li, Z.-X. and Lott, F.: The LMDZ4 general circulation model: climate performance and sensitivity to parametrized physics with emphasis on tropical convection, Climate Dynamics, 27, 787–813, doi:10.1007/s00382-006-0158-0, 2006.

Ilyina, T., Six, K. D., Segschneider, J., Maier-Reimer, E., Li, H. and Núñez-Riboni, I.: Global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-Earth system model in different CMIP5 experimental realizations, J. Adv. Model. Earth Syst., 5(2), 287–315, doi:10.1029/2012MS000178, 2013.

IPCC: Climate Change 2013: The Physical Science Basis, edited by: Stoker,T. F., Qin, D., Plat- tner, G., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge Univ. Press, Cambridge, UK, and New York, NY, USA, 2013..

Ito, T. and Deutsch, C.: Variability of the Oxygen Minimum Zone in the Tropical North Pacific during the Late 20th Century, Global Biogeochem. Cycles, n/a–n/a, doi:10.1002/2013GB004567, 2013.

Ito, T., Woloszyn, M. and Mazloff, M.: Anthropogenic carbon dioxide transport in the Southern Ocean driven by Ekman flow, Nature, 463(7277), 80–83, doi:10.1038/nature08687, 2010.

Jickells, T. and Spokes, L.: The biogeochemistry of iron in seawater, edited by D. R. Turner and K. A. Hunter, John Wiley, Hoboken, N. J., 2001.

Johnson, K., Chavez, F. and Friederich, G.: Continental-shelf sediment as a primary source of iron for coastal phytoplankton, Nature, 398(6729), 697–700, 1999.

Keeling, R. F., Körtzinger, A. and Gruber, N.: Ocean Deoxygenation in a Warming World, Annu. Rev. Marine. Sci., 2(1), 199–229, doi:10.1146/annurev.marine.010908.163855, 2009.

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L. and Roeckner, E.: Advancing decadal-scale climate prediction in the North Atlantic sector, Nature, 453(7191), 84–88, doi:10.1038/nature06921, 2008.

Keller, K. M., Joos, F. and Raible, C. C.: Time of emergence of trends in ocean biogeochemistry, Biogeosciences, 11(13), 3647–3659, doi:10.5194/bgd-10-18065-2013, 2014.

Key, R., Kozyr, A., Sabine, C., Lee, K., Wanninkhof, R., Bullister, J., Feely, R., Millero, F., Mordy, C. and Peng, T.: A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP), Global Biogeochem. Cycles, 18(4), doi:10.1029/2004GB002247, 2004.

Khatiwala, S., Visbeck, M. and Cane, M. A.: Accelerated simulation of passive tracers in ocean circulation models, Ocean Modelling, 9(1), 51–69, doi:10.1016/j.ocemod.2004.04.002, 2005.

Khatiwala, S.: Fast spin up of ocean biogeochemical models using matrix-free Newton-Krylov, Ocean Modelling, 23, 121-129, 2008.

Kim, H.-M., Webster, P. J. and Curry, J. A.: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts, Geophys. Res. Lett., 39(10), L10701, doi:10.1029/2012GL051644, 2012.

Knutti, R., Masson, D. and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40(6), 1194–1199, doi:10.1002/grl.50256, 2013.

Koven, C. D., Chambers, J. Q., Georgiou, K., Knox, R., Negron-Juarez, R., Riley, W. J., Arora, V. K., Brovkin, V., Friedlingstein, P. and Jones, C. D.: Controls on terrestrial carbon feedbacks by productivity vs. turnover in the CMIP5 Earth System Models, Biogeosciences Discuss., 12(8), 5757–5801, 2015.

Kriest, I. and Oschlies, A.: MOPS-1.0: towards a model for the regulation of the global oceanic nitrogen budget by marine biogeochemical processes, Geosci. Model Dev., 8(9), 2929–2957, doi:10.5194/gmd-8-2929-2015, 2015.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochem. Cycles, 19(1), 1–33, 2005.

Laufkötter, C., Vogt, M., Gruber, N., Aita-Noguchi, M., Aumont, O., Bopp, L., Buitenhuis, E., Doney, S. C., Dunne, J., Hashioka, T., Hauck, J., Hirata, T., John, J., Le Quéré, C., Lima, I. D., Nakano, H., Séférian, R., Totterdell, I., Vichi, M. and

roland seferian 15/1/16 16:18
Formatted: Don't hyphenate

roland seferian 15/1/16 16:18
Formatted: Font:(Default) Times, Font color: Black, (Asian) Chinese (PRC), (Other) French

Völker, C.: Drivers and uncertainties of future global marine primary production in marine ecosystem models, Biogeosciences, 12(23), 6955–6984, 2015.

Le Quéré, C., Moriarty, R., Andrew, R. M., Peters, G. P., Ciais, P., Friedlingstein, P., Jones, S. D., Sitch, S., Tans, P., Arneth, A., Boden, T. A., Bopp, L., Bozec, Y., Canadell, J. G., Chini, L. P., Chevallier, F., Cosca, C. E., Harris, I., Hoppema, M., Houghton, R. A., House, J. I., Jain, A. K., Johannessen, T., Kato, E., Keeling, R. F., Kitidis, V., Klein Goldewijk, K., Koven, C., Landa, C. S., Landschützer, P., Lenton, A., Lima, I. D., Marland, G., Mathis, J. T., Metzl, N., Nojiri, Y., Olsen, A., Ono, T., Peng, S., Peters, W., Pfeil, B., Poulter, B., Raupach, M. R., Regnier, P., Rödenbeck, C., Saito, S., Salisbury, J. E., Schuster, U., Schwinger, J., Séférian, R., Segschneider, J., Steinhoff, T., Stocker, B. D., Sutton, A. J., Takahashi, T., Tilbrook, B., van der Werf, G. R., Viovy, N., Wang, Y. P., Wanninkhof, R., Wiltshire, A. and Zeng, N.: Global carbon budget 2014, Earth Syst. Sci. Data, 7(1), 47–85, doi:10.5194/essd-7-47-2015, 2015.

Lehodey, P., Alheit, J. and Barange, M.: Climate variability, fish, and fisheries, Journal of Climate, 2006.

Levitus, S. and Boyer, T.: World ocean atlas 1994, volume 4: Temperature, PB--95-270112/XAB, National Environmental Satellite, Data, and Information Service, Washington, DC (United States). 1994.

Levitus, S., S., Antonov, J. I., Baranova, O. K., Boyer, T. P., Coleman, C. L., Garcia, H. E., Grod- sky, A. I., Johnson, D. R., Locarnini, R. A., Mishonov, A. V., Reagan, J. R., Sazama, C. L., Seidov, D., Smolyar, I., Yarosh, E. S., and Zweng, M. M.: The World Ocean Database TI, Data Science Journal, 12, WDS229–WDS234, 2013.

Levitus, S., Conkright, M. E., Reid, J. L., Najjar, R. G. and Mantyla, A.: Distribution of nitrate, phosphate and silicate in the world oceans, Progress in Oceanography, 31(3), 245–273, 1993.

Lévy, M., Lengaigne, M., Bopp, L., Vincent, E. M., Madec, G., Ethe, C., Kumar, D. and Sarma, V. V. S. S.: Contribution of tropical cyclones to the air-sea CO 2flux: A global view, Global Biogeochem. Cycles, 26(2), doi:10.1029/2011GB004145, 2012.

Lindsay, K., Bonan, G. B., Doney, S. C., Hoffman, F. M., Lawrence, D. M., Long, M. C., Mahowald, N. M., Moore, J. K., Randerson, J. T. and Thornton, P. E.: Preindustrial Control and 20th Century Carbon Cycle Experiments with the Earth System Model CESM1(BGC), J. Climate, 141006111735008, doi:10.1175/JCLI-D-12-00565.1, 2014a.

Ludwig, W., Probst, J. and Kempe, S.: Predicting the oceanic input of organic carbon by continental erosion, Global Biogeochem. Cycles, 10(1), 23–41, 1996.

Madec, G.: NEMO ocean engine, Institut Pierre-Simon Laplace (IPSL), France. Institut Pierre-Simon Laplace (IPSL). [online] Available from: http://www.nemo-ocean.eu/About-NEMO/Reference-manuals, (last access: Novem- ber 2013) 2008.

Maier-Reimer, E.: Geochemical cycles in an ocean general circulation model. Preindustrial tracer distributions, Global Biogeochem. Cycles, 7(3), 645,

roland seferian 15/1/16 16:17

Deleted: Laufkötter, C., Vogt, M., Gruber, N., Aita-Noguchi, M., Aumont, O., Bopp, L., Buitenhuis, E., Doney, S. C., Dunne, J., Hashioka, T., Hauck, J., Hirata, T., John, J., Le Quéré, C., Lima, I. D., Nakano, H., Séférian, R., Totterdell, I., Vichi, M. and Völker, C.: Drivers and uncertainties of future global marine primary production in marine ecosystem models, Biogeosciences Discuss., 12(4), 3731–3824, 2015.

roland seferian 28/1/16 09:47

Deleted: SCHUSTER

doi:10.1029/93GB01355, 1993.

Maier-Reimer, E. and Hasselmann, K.: Transport and storage of CO2 in the ocean ——an inorganic ocean-circulation carbon cycle model, Clim Dyn, 2(2), 63–90–90, doi:10.1007/BF01054491, 1987.

Marinov, I., Gnanadesikan, A., Sarmiento, J. L., Toggweiler, J. R., Follows, M. and Mignone, B. K.: Impact of oceanic circulation on biological carbon storage in the ocean and atmospheric pCO 2, Global Biogeochem. Cycles, 22(3), GB3007, doi:10.1029/2007GB002958, 2008.

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M. and Barriat, P. Y.: Constraining projections of summer Arctic sea ice, The Cryosphere, 6(6), 1383–1394, doi:10.5194/tc-6-1383-2012, 2012.

Matei, D., Baehr, J., Jungclaus, J. H., Haak, H., Muller, W. A. and Marotzke, J.: Multiyear Prediction of Monthly Mean Atlantic Meridional Overturning Circulation at 26.5 N, Science, 335(6064), 76–79, doi:10.1126/science.1210299, 2012.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., Karspeck, A., Kimoto, M., Kumar, A., Matei, D., Mignot, J., Msadek, R., Pohlmann, H., Rienecker, M., Rosati, T., Schneider, E., Smith, D., Sutton, R., Teng, H., van Oldenborgh, G. J., Vecchi, G. and Yeager, S.: Decadal Climate Prediction: An Update from the Trenches, Bull. Amer. Meteor. Soc., doi:doi: 10.1175/BAMS-D-12-00241.1, 2013.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D. and Stockdale, T.: Decadal Prediction, Bull. Amer. Meteor. Soc., 90(10), 1467–1485, doi:10.1175/2009BAMS2778.1, 2009.

Meehl, G. A., Moss, R., Taylor, K. E., Eyring, V., Stouffer, R. J., Bony, S. and Stevens, B.: Climate Model Intercomparisons: Preparing for the Next Phase, Eos Trans. AGU, 95(9), 77–78, doi:10.1002/2014EO090001, 2014.

Mignot, J., Swingedouw, D., Deshayes, J., Marti, O., Talandier, C., Séférian, R., Lengaigne, M. and Madec, G.: On the evolution of the oceanic component of the IPSL climate models from CMIP3 to CMIP5: A mean state comparison, Ocean Modelling, 72 IS -(0 SP - EP - PY - T2 -), 167–184, 2013.

Mikaloff Fletcher, S. E., Gruber, N., Jacobson, A. R., Gloor, M., Doney, S. C., Dutkiewicz, S., Gerber, M., Follows, M., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Müller, S. A. and Sarmiento, J. L.: Inverse estimates of the oceanic sources and sinks of natural CO2 and the implied oceanic carbon transport, Global Biogeochem. Cycles, 21(1), GB1010, doi:10.1029/2006GB002751, 2007.

Moore, J., Doney, S. and Lindsay, K.: Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model, Global Biogeochem. Cycles, 18(4), –, doi:10.1029/2004GB002220, 2004.

Moore, J., Doney, S., Kleypas, J., Glover, D. and Fung, I.: An intermediate

roland seferian 28/1/16 09:39
Deleted:

complexity marine ecosystem model for the global domain, Deep Sea Research Part II: Topical Studies in Oceanography, 49, 403–462, 2002.

Orr, J. C.: Global Ocean Storage of Anthropogenic Carbon, Gif-sur-Yvette, France. 2002.

Phillips, T. J., Potter, G. L., Williamson, D. L., Cederwall, R. T., Boyle, J. S., Fiorino, M., Hnilo, J. J., Olson, J. G., Xie, S. and Yio, J. J.: Evaluating Parameterizations in General Circulation Models: Climate Simulation Meets Weather Prediction, Bull. Amer. Meteor. Soc., 85(12), 1903–1915, doi:10.1175/BAMS-85-12-1903, 2004.

Resplandy, L., Bopp, L., Orr, J. C. and Dunne, J. P.: Role of mode and intermediate waters in future ocean acidification: Analysis of CMIP5 models, Geophys. Res. Lett., 40(12), 3091–3095, 2013.

Resplandy, L., Séférian, R. and Bopp, L.: Natural variability of CO_2 and O_2 fluxes: What can we learn from centuries-long climate models simulations? Journal of Geophysical Research-Oceans, 120(1), 384–404, doi:10.1002/2014JC010463, 2015.

Rodgers, K. B., Lin, J. and Frölicher, T. L.: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an earth system model, Biogeosciences Discuss., 11(12), 18189–18227, doi:10.5194/bgd-11-18189-2014, 2014.

Romanou, A., Gregg, W. W., Romanski, J. and Kelley, M.: Natural air–sea flux of CO2 in simulations of the NASA-GISS climate model: Sensitivity to the physical ocean model formulation, Ocean Modelling, 66 IS -, 26–44, doi:10.1016/j.ocemod.2013.01.008, 2013.

Romanou, A., J. Romanski, and W.W. Gregg, 2014: Natural ocean carbon cycle sensitivity to parameterizations of the recycling in a climate model. Biogeosciences, 11, 1137-1154, doi:10.5194/bg-11-1137-2014.

Romanou, A., W.W. Gregg, J. Romanski, M. Kelley, R. Bleck, R. Healy, L. Nazarenko, G. Russell, G.A. Schmidt, S. Sun, andN. Tausnev, 2013: Natural air-sea flux of CO2 in simulations of the NASA-GISS climate model: Sensitivity to the physical ocean model formulation. Ocean Model., 66, 26-44, doi:10.1016/j.ocemod.2013.01.008.

Rose, K. A., Roth, B. M. and Smith, E. P.: Skill assessment of spatial maps for oceanographic modeling, Journal of Marine Systems, 76(1-2), 34–48, doi:10.1016/j.jmarsys.2008.05.013, 2009.

Roy, T., Bopp, L., Gehlen, M., Schneider, B., Cadule, P., Frölicher, T. L., Segschneider, J., Tjiputra, J., Heinze, C. and Joos, F.: Regional Impacts of Climate Change and Atmospheric CO_2 on Future Ocean Carbon Uptake: A Multimodel Linear Feedback Analysis, J. Climate, 24(9), 2300–2318, doi:10.1175/2010JCLI3787.1, 2011.

Sarmiento, J. L. and Gruber, N.: Ocean Biogeochemical Dynamics, Princeton University Press, Princeton, New Jersey, USA, 526 pp., 2006.

Schwinger, J., Tjiputra, J. F., Heinze, C., Bopp, L., Christian, J. R., Gehlen, M.,

roland seferian 28/1/16 17:46
Formatted: Font:Times New Roman

Ilyina, T., Jones, C. D., Salas-Mélia, D., Segschneider, J., Séférian, R. and Totterdell, I.: Nonlinearity of Ocean Carbon Cycle Feedbacks in CMIP5 Earth System Models, J. Climate, 27(11), 3869–3888, doi:10.1175/JCLI-D-13-00452.1, 2014.

Servonnat, J., Mignot, J., Guilyardi, E., Swingedouw, D., Séférian, R. and Labetoulle, S.: Reconstructing the subsurface ocean decadal variability using surface nudging in a perfect model framework, Clim Dyn, 44(1-2), 1–24–24, doi:10.1007/s00382-014-2184-7, 2014.

Séférian, R., Bopp, L., Gehlen, M., Orr, J., Ethé, C., Cadule, P., Aumont, O., Salas y Mélia, D., Voldoire, A. and Madec, G.: Skill assessment of three earth system models with common marine biogeochemistry, Climate Dynamics, 40(9-10), 2549–2573, doi:10.1007/s00382-012-1362-8, 2013.

Séférian, R., Iudicone, D., Bopp, L., Roy, T. and Madec, G.: Water Mass Analysis of Effect of Climate Change on Air–Sea CO2 Fluxes: The Southern Ocean, J. Climate, 25(11), 3894–3908, doi:10.1175/JCLI-D-11-00291.1, 2012.

Séférian, R., Ribes, A. and Bopp, L.: Detecting the anthropogenic influences on recent changes in ocean carbon uptake, Geophys. Res. Lett., 2014GL061223, doi:10.1002/2014GL061223, 2014.

Séférian, R., Delire, C., Decharme, B., Voldoire, A., Salas y Mélia, D., Chevallier, M., Saint-Martin, D., Aumont, O., Calvet, J.-C., Carrer, D., Douville, H., Franchistéguy, L., Joetzjer, E. and Sénési, S.: Development and evaluation of CNRM Earth-System model – CNRM-ESM1, Geosci. Model Dev. Discuss., 8(7), 5671–5739, 2015.

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R. and Murphy, J. M.: Improved Surface Temperature Prediction for the Coming Decade from a Global Climate Model, Science, 317(5839), 796–799, doi:10.1126/science.1139540, 2007.

Smith, M. J., Palmer, P. I., Purves, D. W., Vanderwel, M. C., Lyutsarev, V., Calderhead, B., Joppa, L. N., Bishop, C. M. and Emmott, S.: Changing how Earth System Modelling is done to provide more useful information for decision making, science and society, Bull. Amer. Meteor. Soc., 140224132934008, doi:10.1175/BAMS-D-13-00080.1, 2014.

Steinacher, M., Joos, F., Frölicher, T. L., Bopp, L., Cadule, P., Cocco, V., Doney, S. C., Gehlen, M., Lindsay, K., Moore, J. K., Schneider, B. and Segschneider, J.: Projected 21st century decrease in marine productivity: a multi-model analysis, Biogeosciences, 7(3), 979–1005, doi:10.5194/bg-7-979-2010, 2010.

Stouffer, R. J., Weaver, A. J. and Eby, M.: A method for obtaining pre-twentieth century initial conditions for use in climate change studies, Clim Dyn, 23(3-4), 327–339, doi:10.1007/s00382-004-0446-5, 2004.

Stow, C. A., Jolliff, J., McGillicuddy, D. J. J., Doney, S. C., Allen, J. I., Friedrichs, M. A. M., Rose, K. A. and Wallhead, P.: Skill assessment for coupled biological/physical models of marine systems, Journal of Marine Systems, 76, 4–15,

roland seferian 11/1/16 17:10
Deleted: .

roland seferian 28/1/16 09:40
Deleted: o

roland seferian 28/1/16 09:41
Deleted: g

doi:10.1016/j.jmarsys.2008.03.011, 2009.

Swingedouw, D., Mignot, J., Labetoulle, S., Guilyardi, E. and Madec, G.: Initialisation and predictability of the AMOC over the last 50 years in a climate model, Clim Dyn, 40(9-10), 2381–2399, doi:10.1007/s00382-012-1516-8, 2013.

Tagliabue, A. and Völker, C.: Towards accounting for dissolved iron speciation in global ocean models, Biogeosciences, 8(10), 3025–3039, 2011.

Tagliabue, A., Bopp, L. and Gehlen, M.: The response of marine carbon and nutrient cycles to ocean acidification: Large uncertainties related to phytoplankton physiological assumptions, Global Biogeochem. Cycles, 25(3), GB3017–n/a, doi:10.1029/2010GB003929, 2011.

Takahashi, T., Broecker, W. and Langer, S.: Redfield Ratio Based on Chemical-Data From Isopycnal Surfaces, Journal of Geophysical Research-Oceans, 90, 6907–6924, 1985.

Tanhua, T., Koertzinger, A., Friis, K., Waugh, D. W. and Wallace, D. W. R.: An estimate of anthropogenic CO2 inventory from decadal changes in oceanic carbon content, P Natl Acad Sci Usa, 104(9), 3037–3042, doi:10.1073/pnas.0606574104, 2007.

Tegen, I. and Fung, I.: Contribution to the Atmospheric Mineral Aerosol Load From Land-Surface Modification, J Geophys Res-Atmos, 100, 18707–18726, 1995.

Tjiputra, J. F., Olsen, A., Bopp, L., Lenton, A., Pfeil, B., Roy, T., Segschneider, J., Totterdell, I. and Heinze, C.: Long-term surface pCO 2 trends from observations and models, Tellus B; Vol 66 (2014), 66(2-3), 151–168, doi:10.1007/s00382-007-0342-x, 2014.

Tjiputra, J. F., Roelandt, C., Bentsen, M., Lawrence, D. M., Lorentzen, T., Schwinger, J., Seland, Ø. and Heinze, C.: Evaluation of the carbon cycle components in the Norwegian Earth System Model (NorESM), Geosci. Model Dev, 6(2), 301–325, doi:10.5194/gmd-6-301-2013, 2013.

Vancoppenolle, M., Bopp, L., Madec, G., Dunne, J. P., Ilyina, T., Halloran, P. R. and Steiner, N.: Future Arctic Ocean primary productivity from CMIP5 simulations: Uncertain outcome, but consistent mechanisms, Global Biogeochem. Cycles, 27(3), 605–619, 2013.

Vichi, M., Manzini, E., Fogli, P. G., Alessandri, A., Patara, L., Scoccimarro, E., Masina, S. and Navarra, A.: Global and regional ocean carbon uptake and climate change: sensitivity to a substantial mitigation scenario, Climate Dynamics, 37(9-10), 1929–1947, doi:10.1007/s00382-011-1079-0, 2011.

Volodin, E. M., Dianskii, N. A. and Gusev, A. V.: Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations, Izv. Atmos. Ocean. Phys., 46(4), 414–431, doi:10.1134/S000143381004002X, 2010.

Walin, G., Hieronymus, J. and Nycander, J.: Source-related variables for the

roland seferian 11/1/16 17:09
Deleted: Swart, N. C. and Fyfe, J. C.: Ocean carbon uptake and storage influenced by wind bias in global climate models, Nature Climate change, 2(1), 47–52, doi:10.1038/nclimate1289, 2011.

description of the oceanic carbon system, Geochem. Geophys. Geosyst., 15(9), 3675–3687, doi:10.1002/2014GC005383, 2014.

Wanninkhof, R.: A relationship between wind speed and gas exchange over the ocean, J. Geophys. Res., 97(C5), 7373–7382, 1992.

Wassmann, P., Duarte, C. M., AGUSTÍ, S. and SEJR, M. K.: Footprints of climate change in the Arctic marine ecosystem, Global Change Biol, 17(2), 1235–1249, doi:10.1111/j.1365-2486.2010.02311.x, 2010.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S. and Kawamiya, M.: MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments, Geosci. Model Dev, 4(4), 845–872, doi:10.5194/gmdd-4-1063-2011, 2011.

Wenzel, S., Cox, P. M., Eyring, V. and Friedlingstein, P.: Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models, J. Geophys. Res. Biogeosci., 2013JG002591, doi:10.1002/2013JG002591, 2014.

Wu, T., Li, W., Ji, J., Xin, X., Li, L., Wang, Z., Zhang, Y., Li, J., Zhang, F., Wei, M., Shi, X., Wu, F., Zhang, L., Chu, M., Jie, W., Liu, Y., Wang, F., Liu, X., Li, Q., Dong, M., Liang, X., Gao, Y. and Zhang, J.: Global carbon budgets simulated by the Beijing Climate Center Climate System Model for the last century, J Geophys Res-Atmos, 118(10), 4326–4347, doi:10.1002/jgrd.50320, 2013.

Wunsch, C. and Heimbach, P.: Practical global oceanic state estimation, Physica D: Nonlinear Phenomena, 230(1-2), 197–208, doi:10.1016/j.physd.2006.09.040, 2007.

Wunsch, C. and Heimbach, P.: How long to oceanic tracer and proxy equilibrium? Quaternary Science Reviews, 27(7-8), 637–651, doi:10.1016/j.quascirev.2008.01.006, 2008.

Yool, A., Oschlies, A., Nurser, A. J. G. and Gruber, N.: A model-based assessment of the TrOCA approach for estimating anthropogenic carbon in the ocean, Biogeosciences, 7(2), 723–751, 2010.

Yool, A., Popova, E. E. and Anderson, T. R.: MEDUSA-2.0: an intermediate complexity biogeochemical model of the marine carbon cycle for climate change and ocean acidification studies, Geosci. Model Dev, 6(5), 1767–1811, doi:10.5194/gmd-6-1767-2013-supplement, 2013.

Zeebe, R. E. and Wolf-Gladrow, D. A.: CO_2 in seawater: equilibrium, kinetics, isotopes, Elsevier Science Ltd. 2001.

Models	spin-up procedure	initial conditions	offline time	online time	total spin-up duration	References
BCC-CSM1-1	sequential	WOA2001, GLODAP	200	100	300	(Wu et al., 2013)
BCC-CSM1-1-m	sequential	WOA2001, GLODAP	200	100	300	(Wu et al., 2013)
CanESM2	sequential (forced w/ obs.)	OCMIP profiles, CanESM1	6000	600	6600	(Arora et al., 2011)
CESM1-BGC	direct	CCSM4	0	1000	1000	(Lindsay et al., 2014)
CMCC-CESM	sequential (w/ acc.)	WOA2001, GLODAP	100	100	200	(Vichi et al., 2011)
CNRM-CM5	sequential	WOA1994, GLODAP, IPSL	3000	100	3100	(Séférian et al., 2013)
CNRM-CM5-2	sequential	WOA1994, GLODAP, CNRM	3000	100	3100	(Schwinger et al., 2014)
CNRM-ESM1	sequential	CNRM-CM5	0	1300	1300	(Séférian et al., 2015)
GFDL-ESM2G	direct	WOA2005,	0	1000	1000	(Dunne et al.,

		GLODAP				2013)
GFDL-ESM2M	direct	WOA2005, GLODAP	0	1000	1000	(Dunne et al., 2013)
GISS-E2-H-CC	direct	WOA2005, GLODAP DIC*	0	3300	3300	(Romanou et al., 2013)
GISS-E2-R-CC	direct	WOA2005, GLODAP DIC*	0	3300	3300	(Romanou et al., 2013)
HadGEM2-CC	sequential	HadCM3LC, WOA2011	400	100	500	(Collins et al., 2011; Wassmann et al., 2010)
HadGEM2-ES	sequential	HadCM3LC, WOA2010	400	100	500	(Collins et al., 2011)
INMCM4	sequential	Uniform DIC	3000	200	3200	(Volodin et al., 2010)
IPSL-CM5A-LR	sequential	WOA1994, GLODAP, IPSL	3000	600	3600	(Séférian et al., 2013)
IPSL-CM5A-MR	sequential	WOA1994, GLODAP, IPSL	3000	300	3300	(Dufresne et al., 2013)
IPSL-CM5B-LR	sequential	IPSL-CM5A-LR	0	300	300	(Dufresne et al., 2013)
MIROC-ESM	sequential	GLODAP/constant values	1245	480	1725	(Watanabe et al., 2011)
MIROC-ESM-CHEM	sequential	GLODAP/constant values	1245	484	1729	(Watanabe et al., 2011)
MPI-ESM-LR	sequential	HAMOCC/constant values	10000	1900	11900	(Ilyina et al., 2013)
MPI-ESM-MR	sequential	HAMOCC/constant values	10000	1500	11500	(Ilyina et al., 2013)
MRI-ESM1	sequential (forced w/ obs.)	GLODAP	550	395	945	(Adachi et al., 2013)
NorESM	direct	WOA2010, GLODAP	0	900	900	(Tjiputra et al., 2013)

Table 1: Summary of spin-up strategy, sources of initial conditions, offline/online durations and references used to equilibrate ocean biogeochemistry in CMIP5 ESMs. The so-called direct and sequential strategies inform whether the spin-up of the ocean

biogeochemical model is run directly in online/coupled mode or first in offline (ocean biogeochemistry only) and then in online/coupled mode. DIC* refers to the observation-derived estimates of preindustrial dissolved inorganic carbon concentration using the ΔC* method. w/ acc. and forced w/ obs. indicates the strategy using 'acceleration' and observed atmospheric forcings during the spin-up, respectively.

roland seferian 28/1/16 09:24
Deleted: I

roland seferian 28/1/16 09:24
Deleted: ir

	O_2			NO_3		
Depth	surface	150 m	2000 m	surface	150 m	2000 m
RMSE	7.19	8.75	5.50	2.07	2.90	2.08
R^2	0.98	0.98	0.99	0.96	0.92	0.94

Table 2: Differences between the oxygen (O_2, μmol L^{-1}) and nitrate (NO_3, μmol L^{-1}) datasets used for initializing IPSL-CM5A-LR (WOA1994) and the datasets used for assessing its performances (WOA2013).

	O_2			NO_3			Alk-DIC		
metrics	mean	RMSE	$RMSE_{max}$	mean	RMSE	$RMSE_{max}$	mean	RMSE	$RMSE_{max}$
Surf	-0.2	2.6	55.8	-0.1	-0.1	34.2	1.6	-0.1	-0.1
150 m	3.4	39.0	31.5	-15.9	33.4	55.2	6.1	27.9	24.7

2000 m	-30.4	144.3	-40.1	2	51.8	-34.8	-69.6	281.8	47.5

Table 3: Drift in % ky^{-1} for oxygen (O_2), nitrate (NO_3) and total alkalinity minus DIC (Alk-DIC) at surface, 150 and 2000 meters as simulated by the IPSL-CM5A-LR model. The drift has been computed over the last 250 years of the spin-up simulation using a linear regression fit of the globally averaged concentrations, root-mean squared error (RMSE) and latitudinal maximum root-mean squared error ($RMSE_{max}$) with respect to the values at year 250.

Figure 1: Spin-up protocols of CMIP5 Earth system models. Color shading represents strategies of the various modeling groups. *Online* and *Offline* steps refer to runs performed with coupled climate model and with stand-alone ocean biogeochemistry model, respectively. Sources of initial conditions for biogeochemical component of CMIP5 Earth system models are indicated as hatching below the barplot.

Figure 2: Time series of two climate indices over the 500-year spin-up simulation of IPSL-CM5A-LR. They represent the global averaged sea surface temperature (a) and the global mean sea-air carbon flux (b). For sea-air carbon flux, negative value indicates uptake of carbon. Steady state equilibrium of physical components as described in Mignot et al., (2013) is reached at ~250 years and is indicated with a vertical dashed line. Drifts in sea surface temperature and global carbon flux are indicated with dashed blue lines. They are computed using a linear regression fit over years 250 to 500. Hatching on panel (b) represents the range of inverse modeling estimates for preindustrial global carbon flux as described in Mikaloff Fletcher et al., (2007), i.e., 0.03±0.08 Pg C y^{-1} plus 0.45 Pg C y^{-1} corresponding to the riverine-induced natural CO_2 outgassing outside of near-shore regions consistently with Le Quéré et al. (2015).

roland seferian 28/1/16 09:24
Deleted: in

roland seferian 28/1/16 09:25
Deleted: for

roland seferian 28/1/16 09:26
Formatted: Hyphenate, Don't adjust space between Asian text and numbers, Tabs:Not at 0,99 cm + 1,98 cm + 2,96 cm + 3,95 cm + 4,94 cm + 5,93 cm + 6,91 cm + 7,9 cm + 8,89 cm + 9,88 cm + 10,86 cm + 11,85 cm

roland seferian 28/1/16 09:26
Deleted: to

roland seferian 28/1/16 09:26
Deleted:

Figure 3: Time series of globally averaged concentration (**[X]** in solid lines) and globally averaged root-mean squared error (RMSE in dashed lines) for dissolved oxygen (O_2), nitrate (NO_3) and difference between alkalinity and dissolved inorganic carbon (Alk-DIC) as simulated by IPSL-CM5A-LR. **[X]** and RMSE are given at surface (a,b and c), 150 m (d, e and f), and 2000 m (g, h and i) for these three biogeochemical fields. Their values are indicated on the left-side and right-side y-axis, respectively. Hatching represents the ±σ observational uncertainty due to optimal interpolation of in situ concentrations around the observed **[X]**.

Figure 4: Snap-shots of spatial biases, ε, in surface concentrations (μmol L^{-1}) in biogeochemical fields during the 500-year spin-up simulation of IPSL-CM5A-LR. ε in dissolved oxygen (O_2), nitrate (NO_3) and difference between alkalinity and dissolved inorganic carbon (Alk-DIC) is given for the first year (a, c and e, respectively) and for the last year of spin-up simulation (b, d and f, respectively).

Figure 5: As Figure 4 but for concentrations at 150 m. Note that color shading does not represent the same amplitude in spatial biases as in Figures 4 and 6.

Figure 6: As Figure 4 but for concentrations at 2000 m. Note that color shading does not represent the same amplitude in spatial biases as in Figures 4 and 5.

Figure 7: Temporal-vertical evolution in root-mean squared error (RMSE) for biogeochemical tracers during the 500-year-long spin-up simulation of IPSL-CM5A-LR. RMSE is given for (a) dissolved oxygen O_2, (b) nitrate NO_3 and (c) difference between alkalinity and dissolved inorganic carbon Alk-DIC.

Figure 8: Temporal evolution of drift in root-mean squared error (RMSE) for dissolved oxygen (O_2, blue crosses), nitrate (NO_3, green crosses) and difference between alkalinity and dissolved inorganic carbon (Alk-DIC, orange crosses) during the 500-year-long spin-up simulation of IPSL-CM5A-LR. Drift in RMSE is given at surface (a,b and c), 150 m (d, e and f), and 2000 m (g, h and i) for these three biogeochemical fields. Drift in RMSE is computed from time segments of 100 years begenning every 5 years from the beginning until year 400 of the spin-up simulation for O_2, NO_3 and Alk-DIC tracers. The best-fit linear regressions between drifts in

roland seferian 16/1/16 14:21
Deleted:

roland seferian 14/1/16 07:40
Deleted: and difference in spatial standard deviation between observations and model fields (Δσ)

roland seferian 14/1/16 07:40
Deleted: and Δσ

roland seferian 28/1/16 09:27
Deleted: are

roland seferian 28/1/16 09:27
Deleted: (

roland seferian 28/1/16 09:27
Deleted:)

roland seferian 28/1/16 09:27
Deleted: (

roland seferian 28/1/16 09:27
Deleted:)

roland seferian 28/1/16 09:27
Deleted: (

roland seferian 28/1/16 09:27
Deleted:) in panels (a,d), (b,e) and (c,f), respectively

RMSE and spin-up duration over year 250 to 500 are indicated in solid magenta lines; their 90% confidence intervals are given by thin dashed envelopes.

Figure 9: Scatterplot of drifts in root-mean squared error (RMSE) in O_2 concentration versus the duration of the spin-up simulation for the available CMIP5 Earth system models. Drifts in O_2 RMSE are respectively given for surface (a), 150 m (b) and 2000 m (c) for oxygen concentrations. Drift in O_2 RMSE is computed from several time segments of 100 years begenning every 5 years from the beginning until the end of the piControl simulation for the available CMIP5 models. Coloured symbols indicate the mean drift in O_2 RMSE while vertical lines represent the associated 90% confidence interval. The best-fit linear regressions between models' mean drifts in RMSE and spin-up duration are indicated as solid green lines; their 90% confidence intervals are given by thin dashed envelopes. Fits are assumed robust if correlation coefficients are significant at 90% (i.e., r^*>0.34). For comparison, drift in O_2 RMSE from our spin-up simulation with IPSL-CM5A-LR (Figure 8) are represented by magenta crosses.

Figure 10: Rankings of CMIP5 Earth system models based on standard and penalized version of the distance from oxygen observations. The standard distance metric is calculated as the ensemble-mean root-mean squared error (RMSE) for O_2 concentrations at surface (a), 150 m (b) and 2000 m (c). The penalized distance metric incorporates drift-induced changes in O_2 RMSE (ΔRMSE) to O_2 RMSE at surface (d), 150 m (e) and 2000 m (f). Ensemble-mean RMSE are calculated using available ensemble members of Earth system models oxygen concentrations averaged over the 1986-2005 historical period relative to WOA2013 observations. ΔRMSE is determined using Equation 2 and fits derived from first century of the CMIP5 piControl simulations. Solid red and magenta lines indicate the multi-model mean standard and penalized distance from O_2 observations, respectively. With the same colour pattern, dashed lines are indicative of the multi-model median for the standard and penalized distance from O_2 observations.

roland seferian 14/1/16 07:42
Deleted: 8…relative …and difference in spatial standard deviation (Δσ) … (in logscale)…Relative d… and Δσ are…,d…,e…e,f…They are estimated from the first century of the piControl simulation relative …and Δσ …in…R…6…similar best-fit linear regressions haven derived …; they are indicated in solid red line… They have been obtained from drifts computed over 100-year-long overlapping time slices of O_2 RMSE from years 250 to 500 at each depth levels and (2) extrapolated over the CMIP5 spin-up duration range (250-10190 years). ... [7]

roland seferian 14/1/16 07:55
Deleted: 9…normalized …normalized normalized ... [8]

roland seferian 14/1/16 07:46
Formatted: Subscript

roland seferian 14/1/16 07:47
Deleted: 5…Red s…and dashed …mean and median for …different ensembles ... [9]

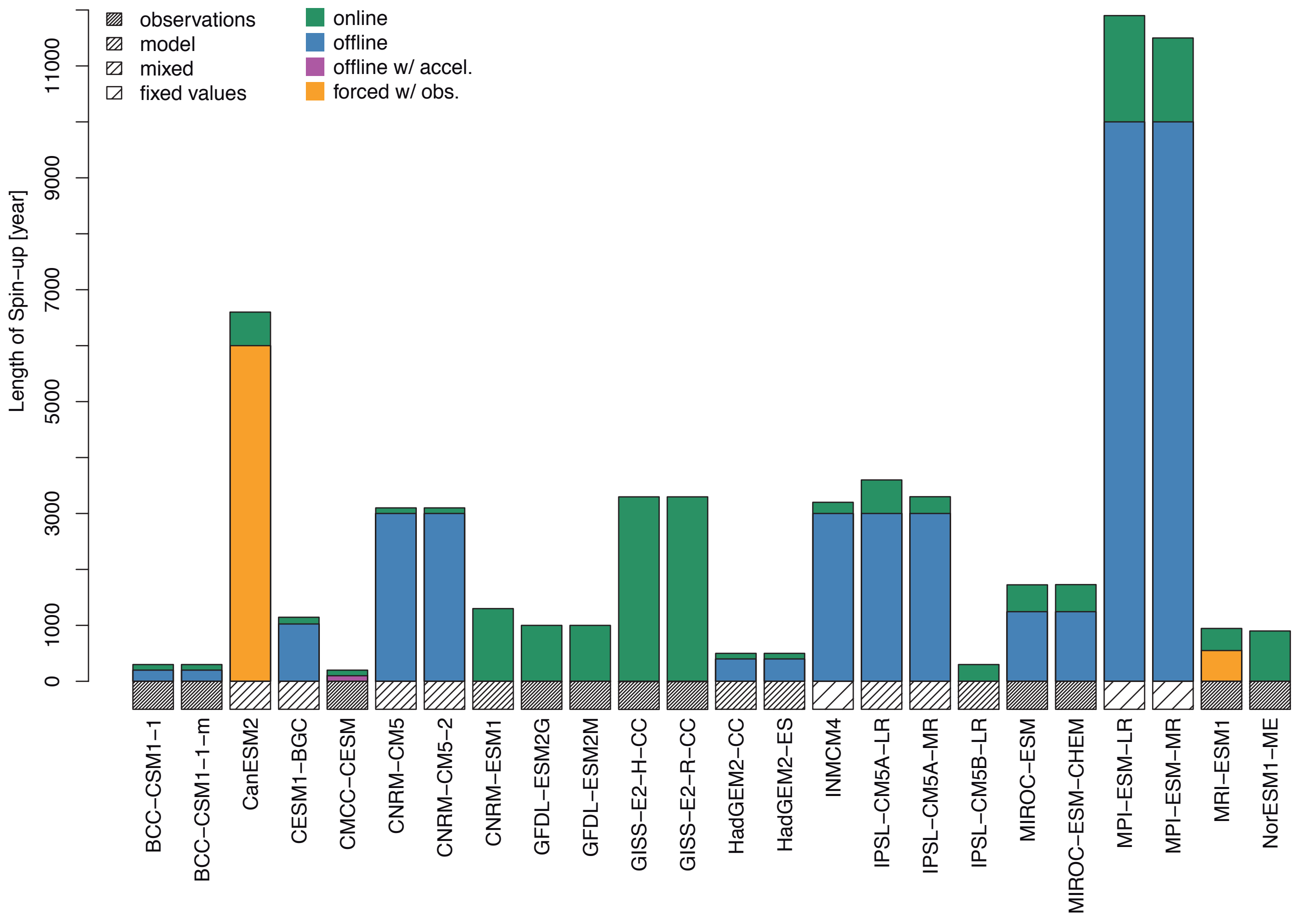
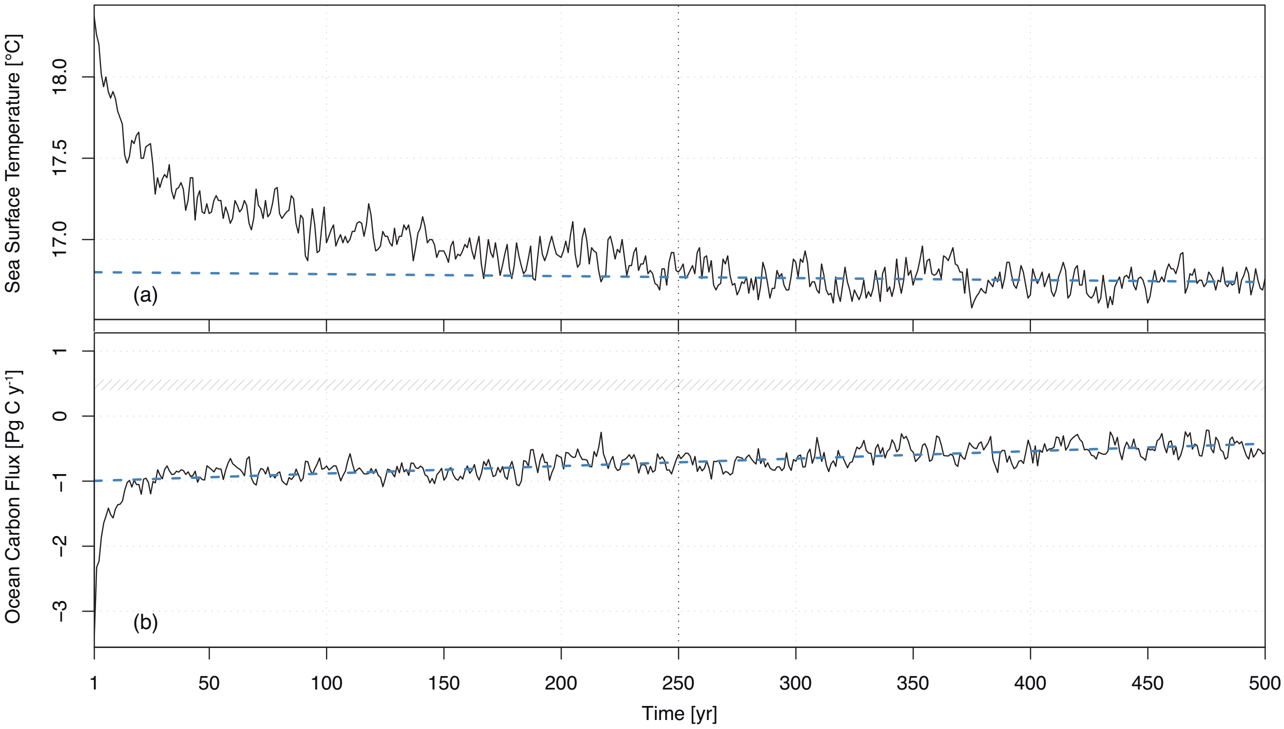
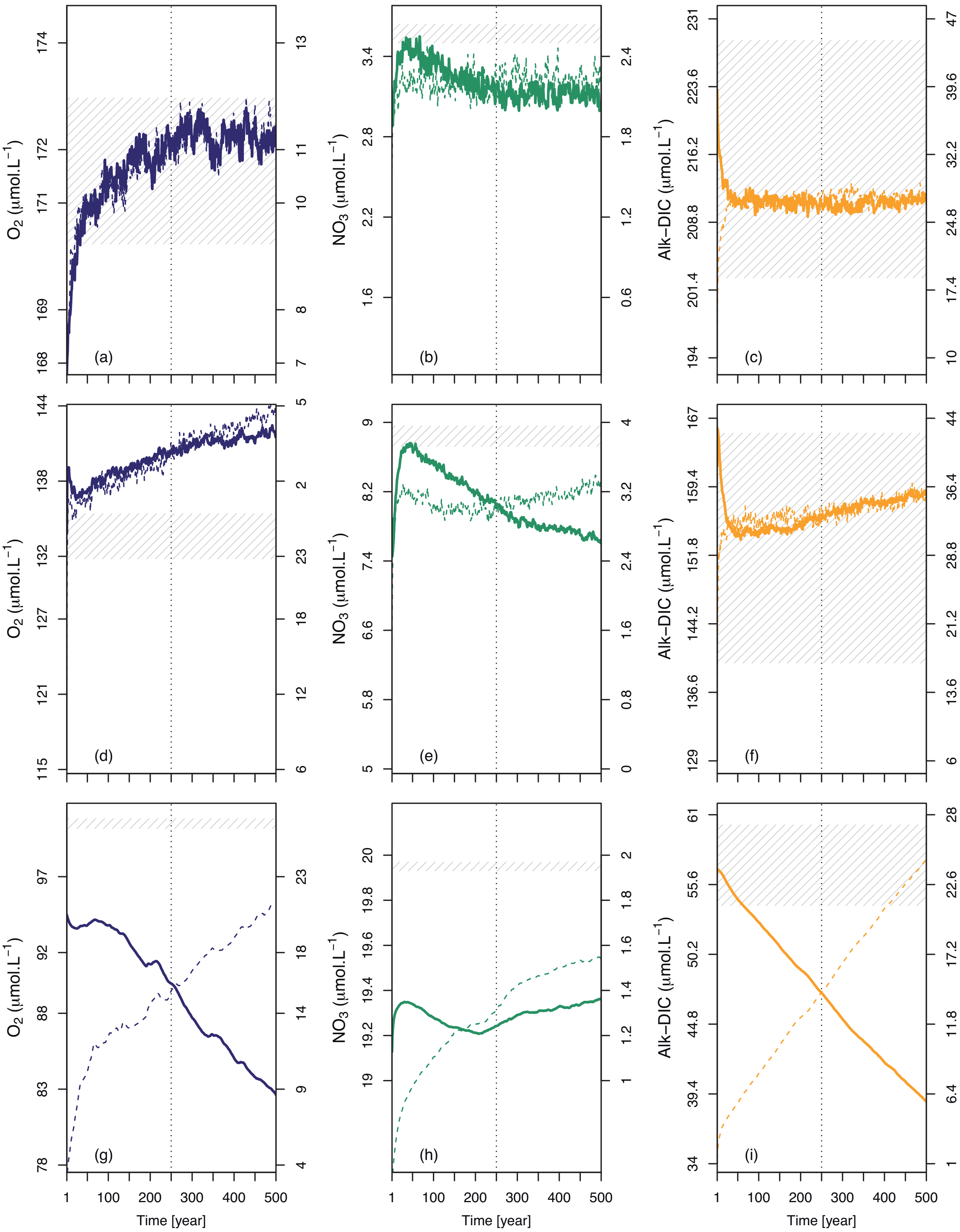

Length of Spin−up [year]
0
1000
3000
5000
7000
9000
11000
observations
model
mixed
fixed values
online
offline
offline w/ accel.
forced w/ obs.
BCC−CSM1−1
BCC−CSM1−1−m
CanESM2
CESM1−BGC
CMCC−CESM
CNRM−CM5
CNRM−CM5−2
CNRM−ESM1
GFDL−ESM2G
GFDL−ESM2M
GISS−E2−H−CC
GISS−E2−R−CC
HadGEM2−CC
HadGEM2−ES
INMCM4
IPSL−CM5A−LR
IPSL−CM5A−MR
IPSL−CM5B−LR
MIROC−ESM
MIROC−ESM−CHEM
MPI−ESM−LR
MPI−ESM−MR
MRI−ESM1
NorESM1−ME

Figure 1:



Sea Surface Temperature [°C]
18.0
17.5
17.0
(a)
Ocean Carbon Flux [Pg C y-1]
1
0
−1
−2
−3
(b)
1
50
100
150
200
250
300
350
400
450
500
Time [yr]

Figure 2:



O2 (µmol.L-1)
NO3 (µmol.L-1)
Alk−DIC (µmol.L-1)
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
Time [year]
Time [year]
Time [year]

Figure 3:

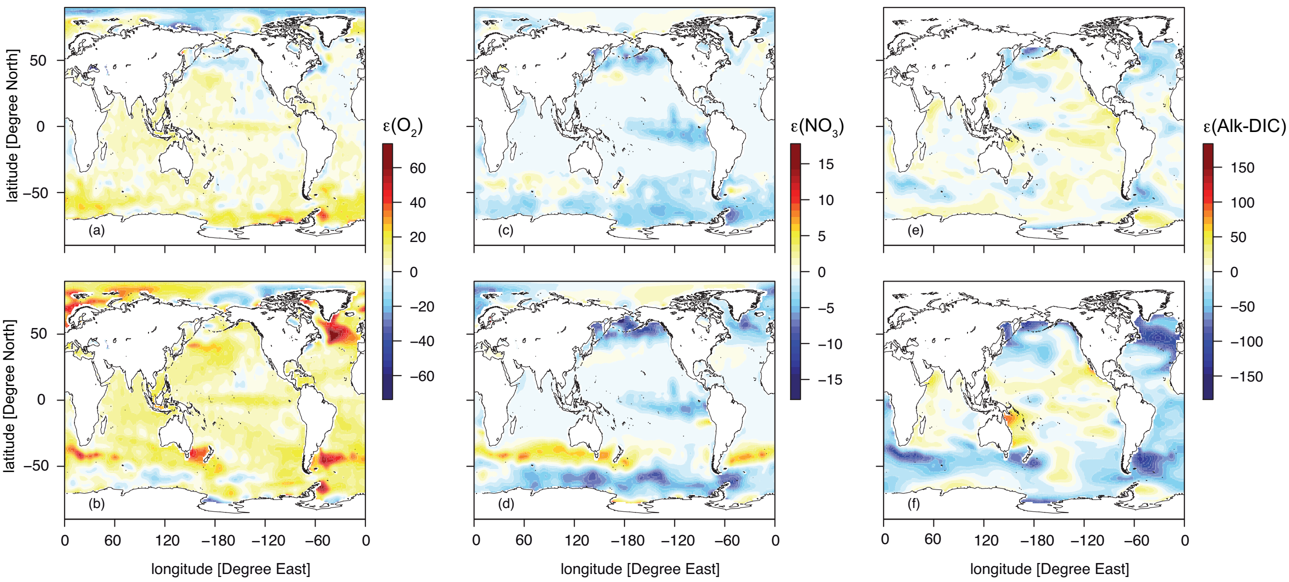


Figure 4:

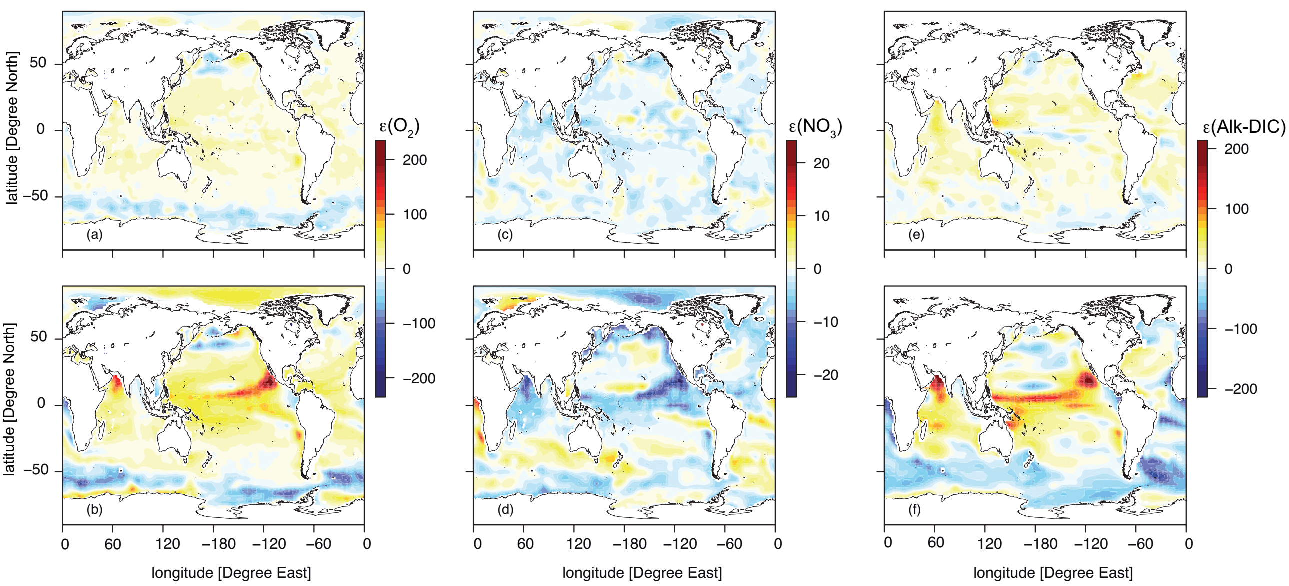


Figure 5:

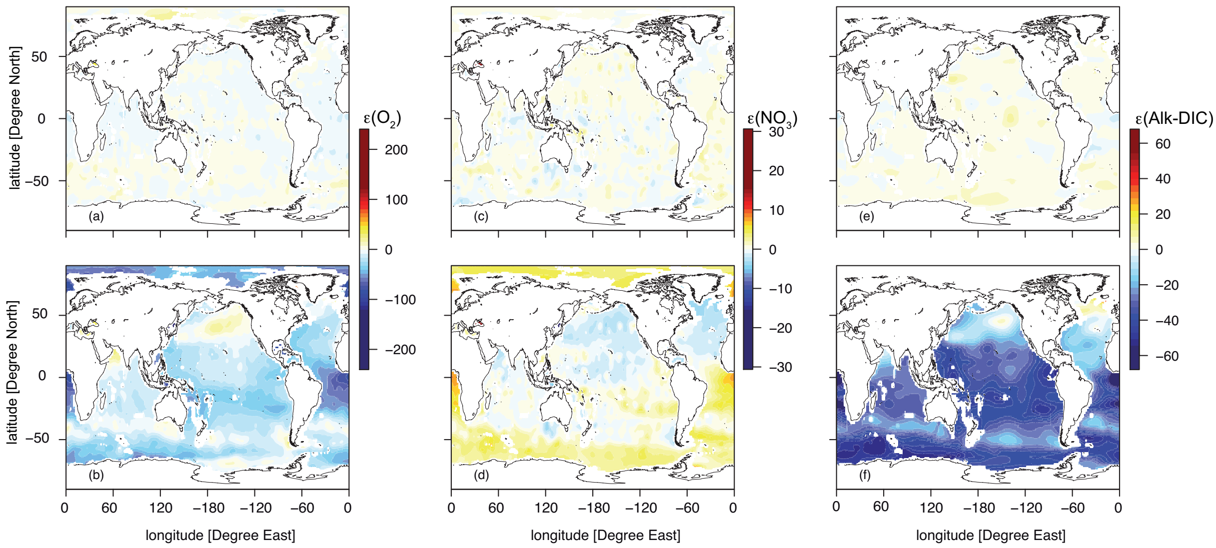


Figure 6:

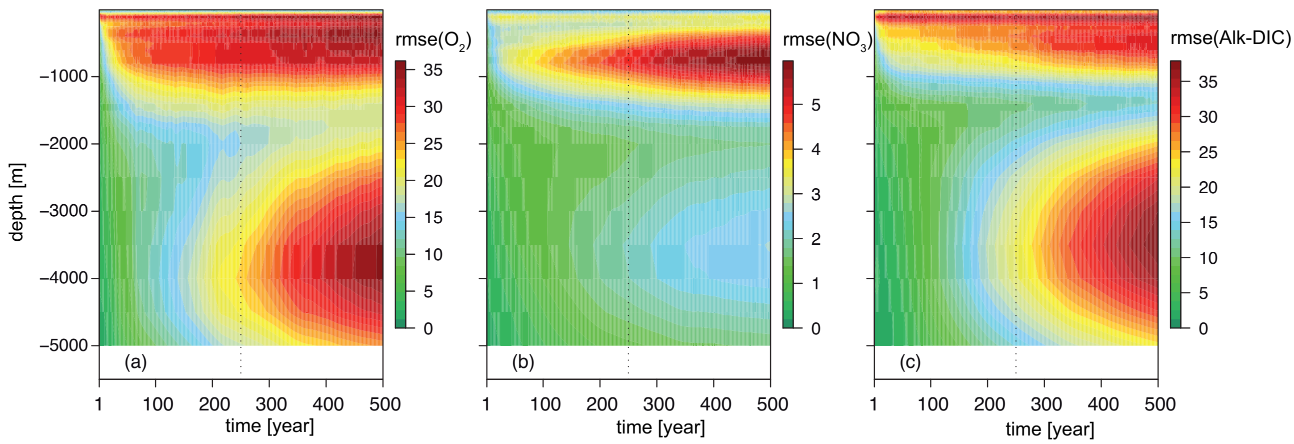
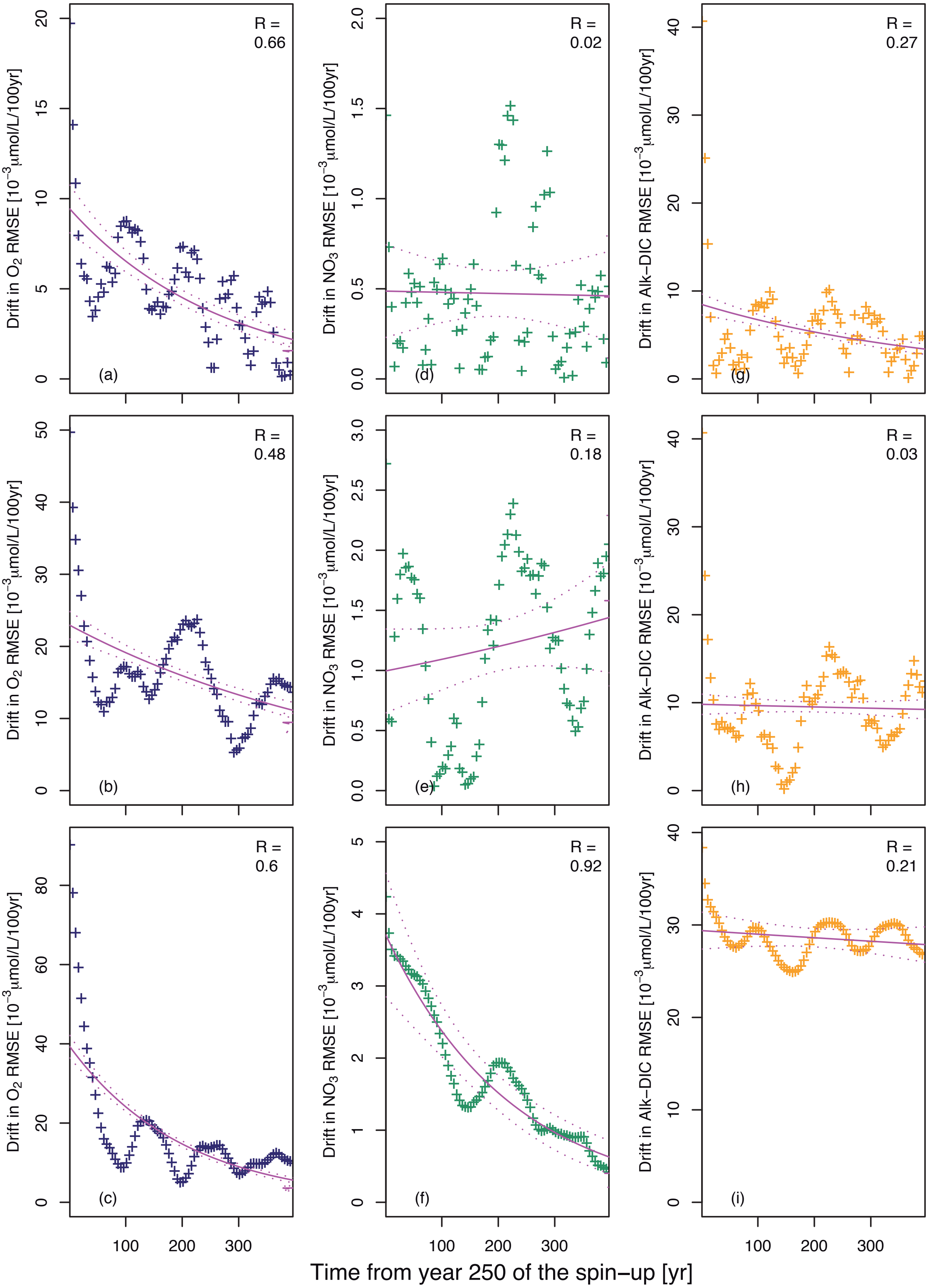
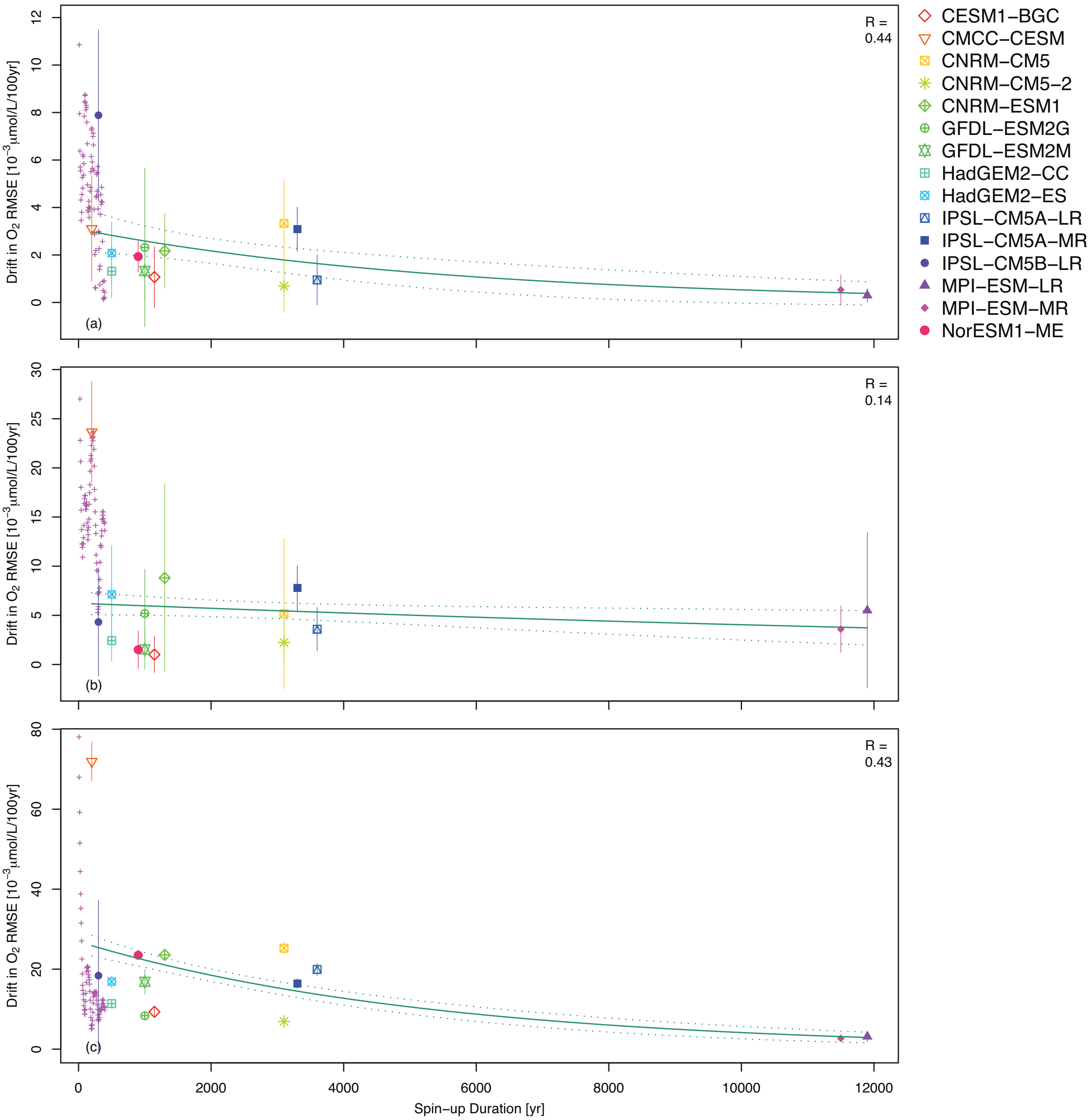


Figure 7:



R = 0.66
(a)
R = 0.48
(b)
R = 0.6
(c)
R = 0.02
(d)
R = 0.18
(e)
R = 0.92
(f)
R = 0.27
(g)
R = 0.03
(h)
R = 0.21
(i)
Drift in O2 RMSE [10^-3 µmol/L/100yr]
Drift in NO3 RMSE [10^-3 µmol/L/100yr]
Drift in Alk−DIC RMSE [10^-3 µmol/L/100yr]
Time from year 250 of the spin−up [yr]

Figure 8:



R = 0.44
(a)
R = 0.14
(b)
R = 0.43
(c)
Drift in O2 RMSE [10^-3 µmol/L/100yr]
Spin−up Duration [yr]
0
2000
4000
6000
8000
10000
12000
CESM1−BGC
CMCC−CESM
CNRM−CM5
CNRM−CM5−2
CNRM−ESM1
GFDL−ESM2G
GFDL−ESM2M
HadGEM2−CC
HadGEM2−ES
IPSL−CM5A−LR
IPSL−CM5A−MR
IPSL−CM5B−LR
MPI−ESM−LR
MPI−ESM−MR
NorESM1−ME

Figure 9:

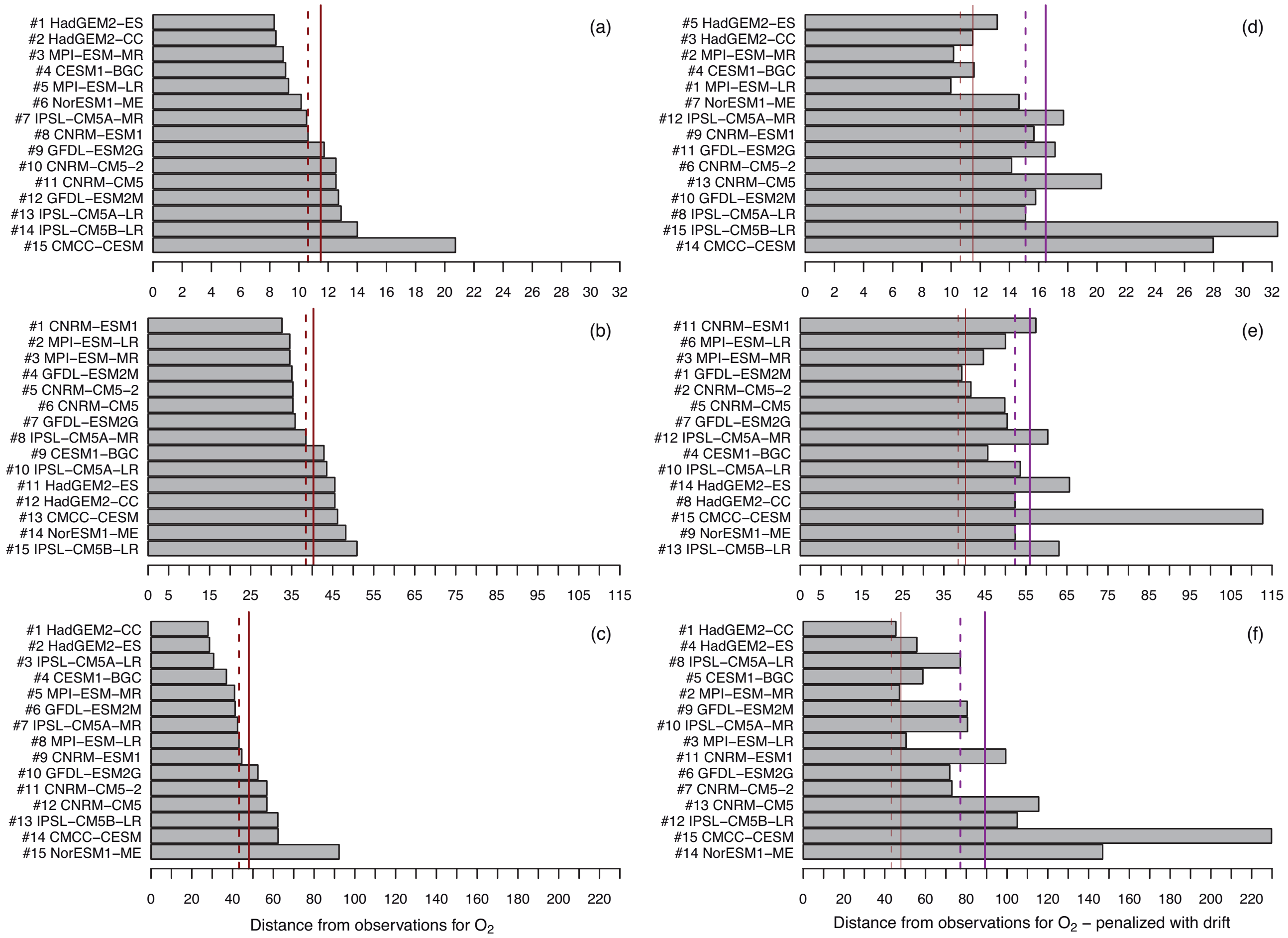

(a)
#1 HadGEM2–ES
#2 HadGEM2–CC
#3 MPI–ESM–MR
#4 CESM1–BGC
#5 MPI–ESM–LR
#6 NorESM1–ME
#7 IPSL–CM5A–MR
#8 CNRM–ESM1
#9 GFDL–ESM2G
#10 CNRM–CM5–2
#11 CNRM–CM5
#12 GFDL–ESM2M
#13 IPSL–CM5A–LR
#14 IPSL–CM5B–LR
#15 CMCC–CESM
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
(b)
#1 CNRM–ESM1
#2 MPI–ESM–LR
#3 MPI–ESM–MR
#4 GFDL–ESM2M
#5 CNRM–CM5–2
#6 CNRM–CM5
#7 GFDL–ESM2G
#8 IPSL–CM5A–MR
#9 CESM1–BGC
#10 IPSL–CM5A–LR
#11 HadGEM2–ES
#12 HadGEM2–CC
#13 CMCC–CESM
#14 NorESM1–ME
#15 IPSL–CM5B–LR
0 5 15 25 35 45 55 65 75 85 95 105 115
(c)
#1 HadGEM2–CC
#2 HadGEM2–ES
#3 IPSL–CM5A–LR
#4 CESM1–BGC
#5 MPI–ESM–MR
#6 GFDL–ESM2M
#7 IPSL–CM5A–MR
#8 MPI–ESM–LR
#9 CNRM–ESM1
#10 GFDL–ESM2G
#11 CNRM–CM5–2
#12 CNRM–CM5
#13 IPSL–CM5B–LR
#14 CMCC–CESM
#15 NorESM1–ME
0 20 40 60 80 100 120 140 160 180 200 220
Distance from observations for O_2
(d)
#5 HadGEM2–ES
#3 HadGEM2–CC
#2 MPI–ESM–MR
#4 CESM1–BGC
#1 MPI–ESM–LR
#7 NorESM1–ME
#12 IPSL–CM5A–MR
#9 CNRM–ESM1
#11 GFDL–ESM2G
#6 CNRM–CM5–2
#13 CNRM–CM5
#10 GFDL–ESM2M
#8 IPSL–CM5A–LR
#15 IPSL–CM5B–LR
#14 CMCC–CESM
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
(e)
#11 CNRM–ESM1
#6 MPI–ESM–LR
#3 MPI–ESM–MR
#1 GFDL–ESM2M
#2 CNRM–CM5–2
#5 CNRM–CM5
#7 GFDL–ESM2G
#12 IPSL–CM5A–MR
#4 CESM1–BGC
#10 IPSL–CM5A–LR
#14 HadGEM2–ES
#8 HadGEM2–CC
#15 CMCC–CESM
#9 NorESM1–ME
#13 IPSL–CM5B–LR
0 5 15 25 35 45 55 65 75 85 95 105 115
(f)
#1 HadGEM2–CC
#4 HadGEM2–ES
#8 IPSL–CM5A–LR
#5 CESM1–BGC
#2 MPI–ESM–MR
#9 GFDL–ESM2M
#10 IPSL–CM5A–MR
#3 MPI–ESM–LR
#11 CNRM–ESM1
#6 GFDL–ESM2G
#7 CNRM–CM5–2
#13 CNRM–CM5
#12 IPSL–CM5B–LR
#15 CMCC–CESM
#14 NorESM1–ME
0 20 40 60 80 100 120 140 160 180 200 220
Distance from observations for O_2 – penalized with drift

Figure 10:

- Supplementary Figures -

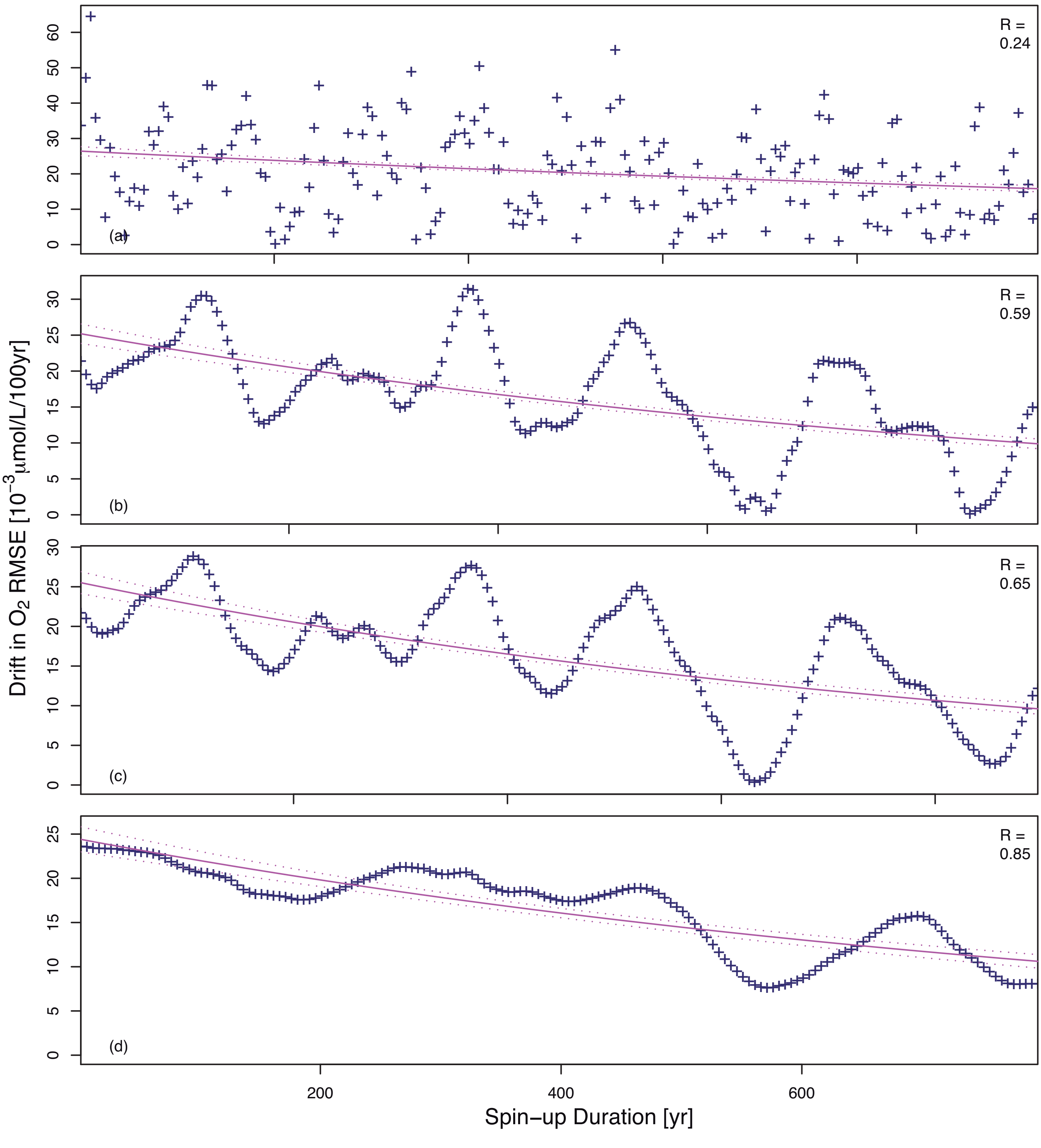


Figure S 1: Temporal evolution of the drift in O_2 root-mean squared error (RMSE) at 2000 m over the 1000-year-long CMIP5 piControl simulation of IPSL-CM5A-LR. Drift in O_2 RMSE is computed from time segments of (a) 20, (b) 50, (c) 80, and (d) 100 years distributed evenly every 5 years from the beginning until the end of the piControl simulation. The best-fit linear regressions between drifts in O_2 RMSE and simulation duration are indicated in solid magenta lines; their 90% confidence intervals are given by thin dashed envelope.

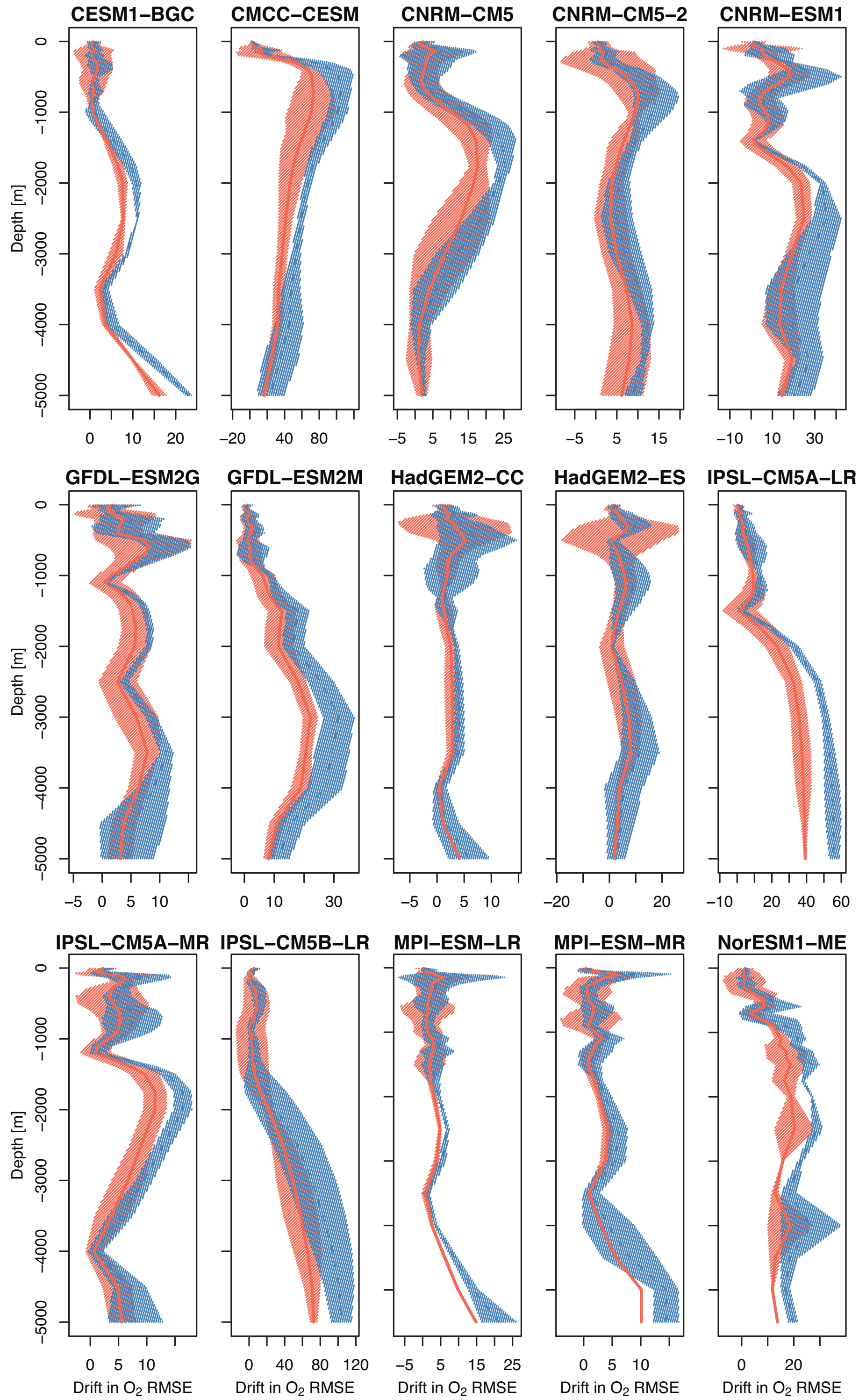


Figure S 2: Vertical profiles of the globally averaged drift in O_2 root-mean squared error (in 10^{-3} μmol L^{-1} kyr^{-1}) from the 15 CMIP5 Earth system models used in this study. Two ways to determine the globally averaged drift are presented in this Figure: vertical profiles determined from global-averaged O_2 RMSE are indicated in blue while those computed from the globally averaged 3-dimensionnal drift (i.e., estimated from 3-dimensionnal O_2 RMSE over domains where the drift in O_2 RMSE fits the simple drift model) are given in red. Solid lines represent the mean vertical profile of the drift in O_2 RMSE; the 90% confidence interval around the mean profile is represented with hatching patterns.

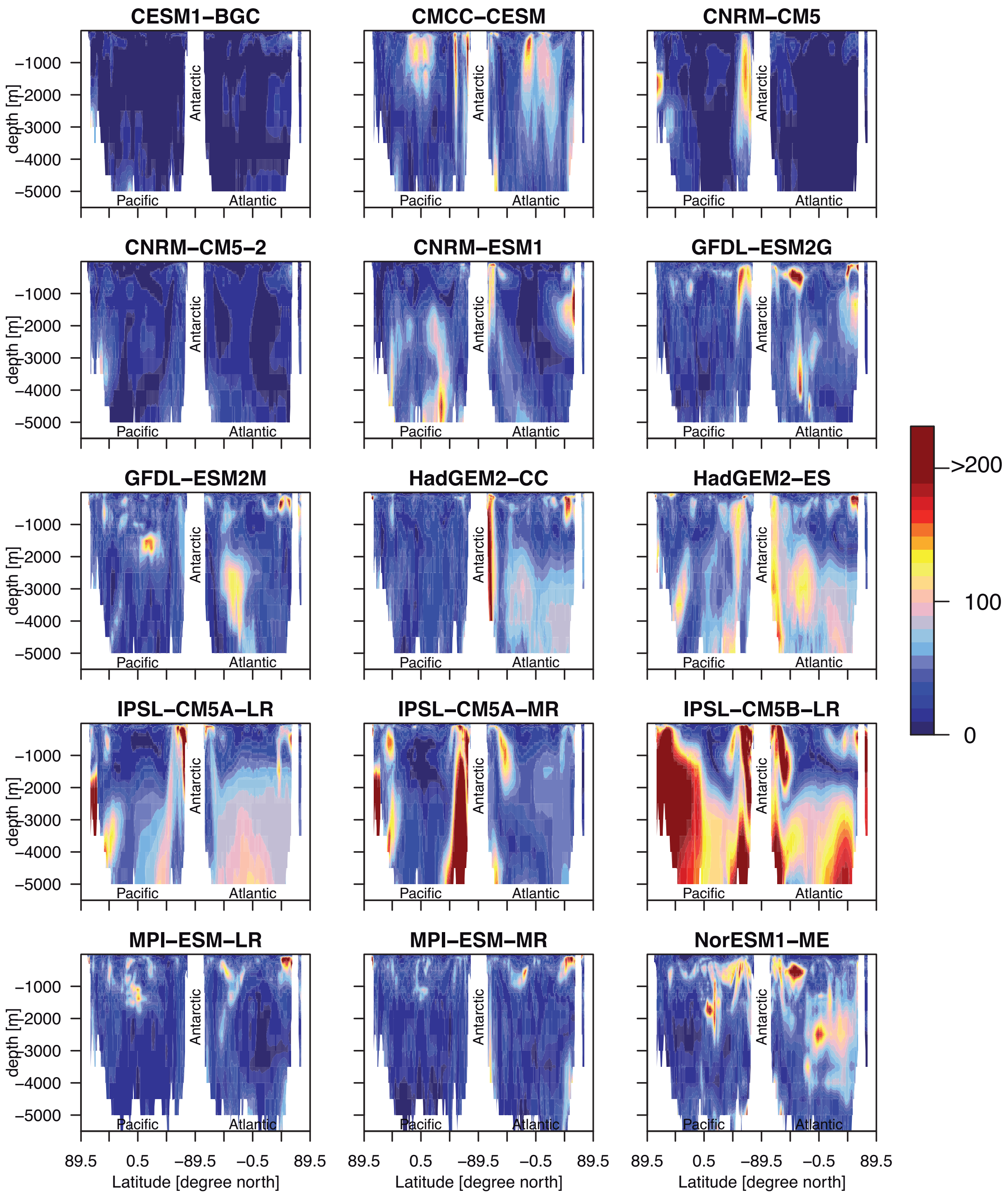


Figure S 3: Vertical structures of the basin-scale drift in O_2 root-mean squared error (in 10^{-3} μmol L^{-1} kyr^{-1}) from the 15 CMIP5 Earth system models used in this study. Basin-scale drift in O_2 RMSE has been computed from 3-dimensionnal drift averaged over Atlantic and Pacific oceans (i.e., estimated from 3-dimensionnal O_2 RMSE over domains where the drift in O_2 RMSE fits the simple drift model).