Dear Editor of *Geoscientific Model Development,*

We further modified the manuscript according to the comments and queries from the reviewers where applicable and justify where we don't follow them.

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

---

**Referee #1**

I thank the authors for their careful consideration of the comments and I am happy with the manuscript as it is, except one final question or comment:

In addition to the previous manuscript version in lines 880-883 it is now stated, "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."

It is not clear to me, why particularly this method (MFNK) should complicate determination of inter-model spreads (on equilibration time scales?). As any new numerical scheme, of course it has to be evaluated and tested carefully with respect to its applicability and suitability for a global coupled biogeochemical ocean model. On the other hand, as an efficient numerical treatment it may rather help to disentangle biogeochemical from physical effects on final model state, because it can offer a chance to spin up different biogeochemical models coupled to different circulations quite quickly. I would suggest to explain in more detail, why this scheme, among the other, probably highly non-linear model processes such as the complex biogeochemical dynamics mentioned above, might give rise to further complications.

The referee is right that the MFNK can be applied to isolate physical and biogeochemical influences on the final individual model's state. However, the method only provides approximation of the steady state and ideally still needs to be followed up with additional forward integration. If this method is not applied in a same way, it will only increase the number of spin-up protocols. This is why we state this kind of methods could further complicate model intercomparison (since the array of spin-up protocols is already quite diverse).

We have modified the corresponding paragraph to avoid further confusion.

Submitted:
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.

Revised:
A novel numerical algorithm to accelerate the spin-up integration time for computationally expensive ocean biogeochemical models has emerged (Khatiwala, 2008), which could help to disentangle physical from biogeochemical contribution to the inter-model spreads, but at the same time, could also potentially complicate the determination of inter-model spreads by increasing the diversity of spin-up protocols.

Referee #2

Unfortunately after a further reading of the manuscript I still think there is much work to be done. In general, the language is often rather impenetrable or imprecise. I have provided a few examples below, but I feel that the manuscript as a whole should be carefully re-read by the team of co-authors (I suspect this hasn’t happened yet). Also, unless I’m missing something I find many aspects of the analysis to be inappropriate.

General comment on figures. I don’t know if this is the policy of the journal but its extremely frustrating not to have figure captions next to the figures, The reader has to be constantly jumping around the document. Also all figures should have panel labels added.

We thank the referee for his careful reading. We hope that the revised version of the manuscript satisfied now his/her queries.

Specific Comments:
In the comments below I refer to line numbers from the track changed pdf: gmd-2015-194-author_response-version2.pdf

94-97. Fig 1 should be referred to. You should also acknowledge other initialisation protocols (e.g. use of previous model runs)

Done and acknowledged.

100: Recent work suggests that these models will not reach a steady state as there are energy leaks in the system: http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-15-0477.1
I’m not sure if similar issues exist for BGC components, but if the physical system has a perpetual drift this is also going to affect the BGC. This puts into question the idea of an exponentially decreasing drift.

Done and acknowledged
Submitted:
“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 (Phillips et al., 2004) to several centuries for ocean and sea ice components (Stouffer et al., 2004). The equilibration of ocean biogeochemical tracers across the entire water column amounts to several thousands of years (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 (Aumont et al., 1998; Bryan, 1984; Gnanadesikan, 2004; Marinov et al., 2008). “

Revised:
“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 (Phillips et al., 2004) to several centuries for ocean and sea ice components (Stouffer et al., 2004). The equilibration of ocean biogeochemical tracers across the entire water column amounts to several thousands of years (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 (Aumont et al., 1998; Bryan, 1984; Gnanadesikan, 2004; Marinov et al., 2008). The equilibration time can be different in coupled model configuration (i.e., ocean-atmosphere general circulation models or ESMs) compared to stand-alone climate components due to leaks in the energy budget (Hobbs et al., 2015).”

108-110: I disagree, there are plenty of studies that do drift correction (including the analysis of ocean variables done by the IPCC) - so an equilibrium is not assumed it is corrected for.

There are indeed several studies that employ a drift correction approach. Yet the method to remove this drift still differs between studies (Gleckler et al., 2016; 2012; Palmer and McNeall, 2014).

We do not omit this point from the present manuscript since we discuss it in section 4-2 of the manuscript:
“So far, the most frequently used approach relies on long preindustrial control simulations to ‘remove’ the drift 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.”

Again, we agree with the referee but this assumption is often used for biogeochemical tracers since processes like nutrient limitation, biological growth rate or interactive carbon cycle coupling (fully-resolved atmospheric CO2) require a quasi-equilibrium of
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).

So far, 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 in space and time in spite of considerable efforts such as the GEOTRACES program for trace elements, or MAREDAT for biomasses of plankton functional types.

Depending on ocean circulation, it ranges from 1500 years for subsurface water masses to 10,000 years for the deep water masses (Wunsch and Heimbach, 2008).

For the deep water masses, this time is about 1500 years in the Atlantic Ocean and reaches up to 10,000 years in the North Pacific Ocean (Wunsch and Heimbach, 2008).

Not clear what is meant here. If there is a continuous flux of carbon into the ocean then how can there be an equilibrium? Do you mean: … ~10,000 years to reach a state where a global sea-air carbon flux is less than 0.01 Pg C y^-1.

Exactly. We have corrected the sentence accordingly.

It depends on how carbon-related reservoirs have been initialized.
For example, in CMCC-ESM, the DIC reservoir has been initialized from modern DIC climatology (Key et al., 2004). This climatology includes not only the preindustrial stock of inorganic carbon but also the anthropogenic perturbation due to the ocean carbon uptake. An acceleration method is then applied to remove the imprint of the anthropogenic carbon uptake by enhancing the outgasing of the model by a factor of 20 (adapted from Alessandri 2006 and Alessandri et al. 2011). For others, including more including interactive components like atmospheric CO2 or sediments, the equilibration of the preindustrial DIC depends on a subtle balance of sink minus source processes occurring at different time scales.

372: do you mean 'computed at different depths'?

Yes, we have corrected the text accordingly.

380-382: Your choice of what constitutes drift appears to be purely subjective. Why do you think that variability below 100yrs is internally driven while variability above 200yrs is associated with drift?

We have explained the reason of this trade-off in our previous rebuttal. To support our statements we have included an assessment of our approach with the supplemental Figure S1.

A real investigation of what constitutes the drift and what constitutes the long-term multicentennial variability would have required a multi-millennial-long simulation as used in (Delworth and Zeng, 2012). Such kind of analysis, although interesting, is not the scope of the present study.

425: total difference (RMSE)
Why not just call it RMSE, why introduce an ambiguous term like ‘total difference’

Please see our response below.

425-426: an RMSE error gives the typical error at a location. Not the error in the global average

In this study, we use the definition of the RMSE which includes the global average bias between the two streams of data. Therefore our metric is a measure of the total difference between the two datasets including not only the global average difference in mean but also the difference in the pattern of errors.

447: … drifts reported for CMIP5 ESMs.
This requires a reference

Done and acknowledged.
To assess the global sea-to-air carbon flux …
Do you mean: ‘To estimate the observed pre-industrial global carbon flux …’?

This sentence introduces the evaluation of the global carbon flux from the IPSL model.
It has been modified as follows:

Submitted:
*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)).*

Revised:
*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)) to evaluate the global sea-to-air carbon flux simulated by IPSL-CM5A-LR.*

Do you mean that this metric has been used to assess if models are equilibrated?

Correct. We have changed the wording of this sentence as follows.

Submitted:
*The temporal evolution of sea-to-air CO₂ 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.*

Revised:
*To assess if ocean carbon cycle reservoirs are equilibrated, we track the temporal evolution of sea-to-air CO₂ fluxes during the spin-up simulation. This metrics was used in phase 2 of the Ocean Carbon Model Intercomparison Project (OCMIP-2, (Orr, 2002)) and has still widely been used during CMIP5 as an equilibration metric for the marine biogeochemistry.*

It is weaker…
I suspect that this is wishful thinking. Is there a significant difference between the two trends?

We have compared the difference between the fitted values obtained with the two regressions (1) from years 250 to 500 and (2) from years 400 to 500. The difference in the mean indicated that the fitted values (and hence the slopes) differ from each other with a p-value < 2 \times 10^{-16}. This result was somehow expected since the slope of the two linear fits differs by one order of magnitude. This information is now included in the manuscript.
I feel that this section needs to be much more clearly explained. I don’t understand what you are doing here. A number of issues:

- I suspect that the statement that the linear trend over the last 100 years is smaller than the 250-500 trend is wishful thinking. Is there a significant difference between the two trends?
- I’m not clear how you are constructing the exponential model. Do you calculate a sequence of 100yr trends from all the data and then fit your exponential as you do later in the paper?
- I don’t really understand what the 0.4-0.56 number represents. Are you saying that there was a pre-industrial outgassing of this amount? Given model biases why would we expect model equilibrium to be the same as the observed equilibrium?
- If I understand correctly, you use a linear trend to calculate an intersection between the trend and the 0.4-0.56 range. So you are assuming that the drift magnitude now stays the same over the next 1100-1300 years. You then use an exponential model to estimate the rate of drift after 1100-1300 years (that was obtained using a linear drift model). Either I have misunderstood this section, or your method doesn’t seem to make sense.

We apologize for our lack of clarity. We address the referee’s comments below:

(1) As mentioned above, we have tested the difference in the two fits using a t-test. With this approach we have assessed that the difference between the two fits is different from zero. The t-test gives a p-value of about $2 \times 10^{-16}$, indicating that the slope computed over years 250-500 and years 400-500, respectively, differ between each other.

(2) Since our model uses external riverine carbon input the expected outgassing of carbon is about $0.45 \text{ Pg C y}^{-1}$ under equilibrated condition. This flux of carbon is not included in the inversion-based data product (Mikaloff Fletcher et al., 2007) which complicates the comparison between data and model. Therefore, we have added a riverine-induced carbon outgassing of about $0.45 \text{ Pg C y}^{-1}$ to the MF2007 inversion-based preindustrial carbon flux in agreement with previous inversion estimates (Jacobson et al., 2007). The derived data MF2007+riverine-induced carbon flux is considered as a target range of values.

(3) Following the referee's suggestion, we have modified this section as follows.

Submitted:

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 $0.001 \text{ Pg C y}^{-2}$ from year 250 to 500, it is weaker over the last century of the simulation ($7 \times 10^{-4} \text{ Pg C y}^{-2}$). Using a linear fit over the last century of the simulation with a drift of $7 \times 10^{-4} \text{ Pg C y}^{-2}$, we estimate that the simulated sea-to-air carbon flux would reach the range of $0.4-0.56 \text{ Pg C y}^{-1}$ 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 $2 \times 10^{-7}$ and $7 \times 10^{-7} \text{ Pg C y}^{-2}$ after this 1100 to 1300 supplemental years of spin-up simulation.
Revised:

Figure 2b shows that the global sea-to-air carbon flux is still lower than the range of values estimated from preindustrial natural ocean carbon flux inversions (0.4-0.56 Pg C y\(^{-1}\)). Besides, Figure 2b shows that the drift in the global sea-to-air carbon flux becomes smaller more slowly after a strong decline during the first 50 years of the spin-up simulation. From year 250-500 this drift is about 0.001 Pg C y\(^{-2}\), and still weaker over the last century of the simulation (7x10\(^{-4}\) Pg C y\(^{-2}\)). A one-sided t-test indicates that the two drifts differ from each other with a p-value < 2x10\(^{-16}\). When fitted with drifts computed from overlapping time segments of 100 years, our simple drift model (Equation 1) gives a relaxation time of around 160 years. We use this relaxation time and the drift of 7x10\(^{-4}\) Pg C y\(^{-2}\) to estimate the additional spin-up time required for the model to reach an outgassing of 0.4-0.56 Pg C y\(^{-1}\) as 1100 to 1300 years. However, even after this integration time, the drift in ocean-atmosphere carbon flux estimated with our simple drift model still ranges from 2x10\(^{-7}\) to 7x10\(^{-7}\) Pg C y\(^{-2}\).

482: couldnt this equally lead to an overestimate of the time?

We rather consider that our estimate underestimates the time to reach equilibrium given the non-linearity of the ocean carbon cycle. This point had already been introduced in the submitted manuscript to discuss our estimates and present the limitation of this approach. It is reported below:

“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.”

Figure 3 caption: what are the [X]’s

In the caption of Figure 3, it is indicated:

“Time series of globally averaged concentration ([X] in solid lines)”. Therefore “[X]” represents the global average concentration for O\(_2\), NO\(_3\) and Alk-DIC tracers. For sake of clarity, we have removed this symbol from the Figure’s caption.

495-496: But according to your figure Alk-DIC lies within the observational range of uncertainty.

Done and acknowledged.

522: The panel labels in Figure 4 are not shown. Please check this for all figures. Also where appropriate panel and figure should be provided in the text (not just figure number)

Panel labels are given at the bottom left side of each panel in submitted Figures 4 to 6.
The small paragraph starting the subsection 3-4 from lines 521 to 525 just introduces the Figures. Each Figure and its corresponding panels are then used in the current version of the manuscript, for example at line 526.

538: …pattern of errors are well correlated. well correlated with what? Do you mean there is a high level of spatial autocorrelation? If so, what is the implication of this?

We apologize for the unclear information. We have amended the text as follows.

Submitted:
*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\textsubscript{2} stochiometry of PISCES).*

Revised:
*Figure 5 illustrates that patterns of error for O\textsubscript{2}, NO\textsubscript{3} and Alk-DIC fields are well correlated with each other (R>0.6). This reflects that in PISCES carbon, nitrogen and oxygen concentrations are linked by the elemental C:N:-O\textsubscript{2} stochiometry fixed in space and time.*

538-539: It directly translates the assumptions employed in the biogeochemical model what does this mean?

Please see revised text above.

The descriptions of the figures 4 5 & 6 are largely the same. What insight is gained from the changes at different depths. Unless something interesting can be concluded I suggest deleting 1 or 2 of these figures

We agree that to some extent patterns of error are correlated with each other for a given depth level, but they present different structures at different depth levels. We prefer to keep these figures in the main text rather than moving one of them to the supplementary material.

545: total mismatch

It would be very helpful if precise terminology was used (this applies at many places in the manuscript). What does ‘total mismatch’ mean? I presume its the evolution of globally averaged rmse at different depth levels. The figure caption isnt clear either.

We have corrected the text accordingly and replaced all occurrences of ‘total mismatch’ by RMSE.
551-552: after few decades within the upper hundred meters …
But in figure 3 you show that it takes O[250yrs] to reach a quasi steady state.

We did omit a part of the information indeed; the correct statement is “a few decades after 250 years of spin-up”.

555: respectively in relation with the structure of the large-scale ocean circulation.
I don’t understand what this means.

Please see the response below.

556-557: RMSE evolves much slower because this depth corresponds to the depth of the very old radiocarbon age
this statement of causality makes no sense. I presume you are trying to say that ventilation is slow in these regions as evidenced by large radiocarbon age.

Please see the response below.

559-560: Why?

We apologize for this lack of clarity. We have modified the paragraph text to improve readability as follows.

Submitted:
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 (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.

Revised:
Patterns of errors within the thermocline and upper 1000m water masses evolve relatively fast (within a few centuries) due to the relatively short mixing time in the upper ocean. Mid-depth (~1500-2500m) RMSE evolves much slower because of the slow ocean circulation at these depth levels. Characteristic time scales here are thousands of years as evidenced by radiocarbon age (Wunsch and Heimbach, 2007; 2008). This explains why at the end of the spin-up simulation, two maxima of comparable amplitude are found for RMSE at 150 m and 3750 m for O$_2$ and at 50 m and 3800 m for Alk-DIC (Figure 7).

564-565: equilibration after a longer of …
There’s something wrong with this sentence

We have rephrased this sentence as follows.
Submitted:
With the evolution of the RMSE established, we can use the simple drift model (Equation 1) to determine the relaxation time, $\tau$, required to reach equilibration after a longer of spin-up simulation.

Revised:
With the evolution of the RMSE established, we can use the simple drift model (Equation 1) to determine the relaxation time, $\tau$, which characterizes the e-folding time scale of the RMSE.

568: is fitted level to the 80 drift values for
what does 'is fitted level' mean?

We have corrected this typo error ('level' removed).

571: The simple drift model fits well the evolution of the drift …
Based on your definition of drift i.e. 100 yr linear trend, the exponential model looks very different to the data you are trying to fit to. You are not getting any of the centennial variability (that by your definition constitutes variability in the drift).
In general, I am not clear why you have taken this approach. If you think that drift follows an exponential model why not just fit an exponential to the raw RMSE data without first calculating linear trends.
Also in figure 8 caption you say that the magenta line is: 'The best-fit linear regressions'.
This cannot be correct as the lines aren’t linear.

Our simple drift model is a conceptual way to analyze the temporal change of the drift. It does not intend to capture or replicate the centennial variability of the drift simulated by a model. The quality of the fit is assessed with correlation coefficient which provides an objective measure of the fit's accuracy.

In terms of methodology, what the referee proposes is different from our approach: working on the raw RMSE implies that one tries to capture the evolution of the RMSE which is quite different from analyzing the rate at which the drift in RMSE changes in course of time. In other words, we are working on the time derivative of the RMSE (dX/dt) which is an accurate way to find the time at which dX/dt=0 (i.e., equilibrium period).
Regarding the point on the curves —that are not linear—, we justify our nomenclature as follows. An exponential model belongs to the family of the Poisson generalized linear models meaning that there is a transfer function that links the parameters of the exponential model to those of a linear model. In our case, the transfer function is a natural logarithm. Therefore, the parameters of the model have been estimated using least square regression similar to any linear regression.
With that said, we have chose to remove the word ‘linear’ to avoid further confusion.

573: Unless I'm misunderstanding what you are doing this comparison in term of correlation is not meaningful. What are you trying to test? that the slopes are same? (for this you wouldn't look at correlation) or that the variability is coherent? (you this you
would use correlation but the model doesn't have any variability). Moreover the data certainly does not have 80 degrees of freedom - there is massive temporal auto correlations. The fitted model would have far less.

The referee is right. We have assessed the effective degrees of freedom using the formulation proposed by (Bretherton et al., 1999), which accounts for the autocorrelation of the time series in the computation of the degrees of freedom. This reduces drastically the degrees of freedom from 80 to 15 as the referee expected. Yet, this does not change the accuracy of the fit (determined with correlation coefficients) except for Alk-DIC at 2000 m. We have amended the text accordingly.

Figure 9 caption. As in figure 8, the green line is not a linear fit. This is presumably the fit to the exponential model.

As mention above, we have chosen to remove the word ‘linear’ to avoid further confusion.

608: this relationship suggests a general decrease of the drift as a function of spin-up duration.
For 150m it looks like the CI would suggest that the fit is not different from a slope of zero i.e. no relationship. You also contradict yourself below when you say: ‘at 150 m depth and hence indicating that there is no link’
Similar to my comments for figure 8, I don’t believe this correlation analysis is meaningful. 1. Correlation is a test for a linear relationship 2. Its not used to assess the relationship between discrete data and a line of best fit. A fit line would have very few degrees of freedom, so you couldn’t assume 15 independent samples.

As mentioned above, the exponential model belongs to the family of generalized linear models. Therefore, an exponential relationship can be fitted using a transfer function (here a natural log). Thanks to this approach, a least square regression can be used to fit the data and a correlation (or squared correlation) can be used to determine the quality of the fit.

615: This low significance level do you mean low correlation? your terminology here and elsewhere needs checking

As justified above, in a case of a generalized linear model our terminology is adequate.

620-623:
Is this an extremely complicated way of saying that in IPSL the drift also reduces with time?

We have improved the wording of this sentence accordingly.

625-627: If the spin up is too short to determine robust drifts, doesn’t this invalidate your whole section regarding this model?
This point has been mentioned to clearly state the limitation of our approach. We use a dedicated spin-up simulation with IPSL-CM5A-LR, starting at rest and using observed initial conditions to assess our simple drift model. This simulation differs not only in length but also in protocol compared to the IPSL-CM5A-LR spin-up simulation that has served to produce all of the CMIP5 dedicated simulation. Therefore, a comparison between our spin-up simulation and the CMIP5 results is not straightforward. A naïve but reasonable assumption is that our spin-up simulation would have converged toward the CMIP5 one if the spin-up simulation had been continued.

639: why do you call it distance. I presume you mean the 'ensemble mean globally averaged RMSE'. As mentioned above it would be really helpful if you stick to precise terminology

We use this terminology to stick to that of (Knutti et al., 2013). Besides, a distance has a clear statistical meaning: https://en.wikipedia.org/wiki/Statistical_distance
Therefore, we prefer to keep this terminology in the revised manuscript.

Figure 10 x-axis should have units
Done and acknowledged.

697: Just because circulation reaches a steady state doesn’t mean that other important physical variables like temperature aren’t still drifting

We agree, but for this simulation climate reaches a quasi-equilibrium after 250 years of spin-up, not only the large-scale ocean circulation. We have corrected this sentence accordingly.

701: I don’t understand what you mean by error propagation. You haven’t mentioned this in the results.
We have changed the wording of this sentence. ‘Error propagation’ implies that the RMSE generally grows with time.

734-756. This point is illustrated …
I don’t see how this illustrates the above point. I'm not clear what the above point means.

This point justifies why some biogeochemical fields have to reach a quasi-equilibrium before simulating any climate change scenarios. Indeed, some tracer concentrations are used to characterize specific oceanic domains (high nutrient low chlorophyll regions, oxygen minimum zones) and are also essential to define critical thresholds for some biological processes like the nutrient-to-light limitation or the nitrification. Although climate change impacts can be detected with any drift-correction approach a better understanding of the drivers of these changes is required to simulate accurately
tracer concentrations. This implies that it is necessary to better define if a model is drifting or not, and, of course, to improve model components.

745-748. Sentence doesn’t make sense to me.

Please see the response below.

753-756. Again this doesn’t make sense to me. Spread in what? In fact Im struggling to understand what this whole paragraph is trying to say.

We have modified the sentence as follows.

Submitted: 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.

Revised: Although (Vancoppenolle et al., 2013) and (Laufkötter et al., 2015) explain a part of the difference in simulated nutrient concentration by the differences in the spatial resolution and the complexity of the models, the authors of both studies qualitatively invoked differences in spin-up duration to explain the remaining differences in simulated concentrations.

826-829. Im not sure that an exponential model is most appropriate here in many global cases (let alone regional cases) (see your fig 8)

We disagree with the referee. We have assessed the accuracy of our model with global average RMSE. Most of the time the quality of the fit is significant at 90% confidence level (even at 95% for some depth levels). Note that we introduced this section based on previous referee comments considering that fitting an exponential decay on regional data is not straightforward. In some cases, the hypothesis of a negative decay of the drift in course of time does not hold.

831: The above-mentioned remark
Which remark?

We apologize for this misleading information. The text is now amended as follows.

Submitted: The above-mentioned remark can explain the relatively low confidence level of the fit to drift across the multi-model CMIP5 ensemble (Figure 9).
Besides, differences in simulated processes and resolution can explain the relatively low significance level of the fit to drift across the multi-model CMIP5 ensemble (Figure 9).

831: the relatively low confidence level of the fit This terminology is very awkward. Do you mean the poor fit of the exponential model to the drift data? In one sentence you talk about confidence in the next you talk about significance.

Our statement is correct. Confidence level and significance level are synonymous. In wikipedia: “Select a desired level of confidence (significance level, p-value or alpha level) for the result of the test.”
https://en.wikipedia.org/wiki/Pearson's_chi-squared_test

840-841: it is unlikely that model fields drift at the same rate along the spin-up simulation Isn't this why you use an exponential model which changes over time?

The referee is right. This statement is linked to the complexity of the models. Therefore, we have arranged the subsection as follows.

Submitted:
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 $\tau$ estimates need to be used with caution when computed from short spin-up simulation because they can be subject to large uncertainties.

Revised:
Besides, difference in simulated processes and resolution 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 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. Indeed, as shown in (Kriest and Oschlies, 2015), various parameterizations of the particle sinking speed 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 $\tau$ estimates need to be used with caution when computed from short spin-up simulations because they can be subject to large uncertainties. 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.

859: 3-dimensional growth rate

What is this?

This refers to the time derivative of the RMSE or any other skill-score metrics. To avoid any further confusion, we have removed this word from the main text.


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., Reich, 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.


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


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

Roland Séférian1, Marion Gehlen2, Laurent Bopp3, Laure Resplandy3,2, James C. Orr2, Olivier Marti1, John P. Dunne4, James R. Christian5, Scott C. Doney6, Tatiana Ilyina7, Keith Lindsay8, Paul R. Halloran9, Christoph Heinze10,11, Joachim Segschneider11, Jerry Tjiputra11, Olivier Aumont13, Anastasia Romanou14,15

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
Abstract

During the fifth phase of the Coupled Model Intercomparison Project (CMIP5) substantial efforts were made to systematically assess the skill of Earth system models. One goal was to check how realistically representative marine biogeochemical tracer distributions could be reproduced by models. In routine assessments model historical hindcasts were compared with available modern biogeochemical observations. However, these assessments considered neither how close modeled biogeochemical reservoirs were to equilibrium 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 (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 holds when confronted with a larger ensemble of CMIP5 models. This shows that drift has implications for 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 quantitatively more correct ESM results on marine biogeochemistry and carbon cycle feedbacks.

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. The ESMs that contributed to CMIP5 substantially differed from each other in terms of their simulations of physical and biogeochemical components of the Earth System. These differences in design...
translate into a significant variability between the skill with which the different models reproduce the observed biogeochemistry and carbon cycle, which in turn may impact projected climate change responses (IPCC, 2013).

In the typical objective evaluation and intercomparison of these models, a suite of standardized statistical metrics (e.g., correlation, root-mean-squared errors) are 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 (e.g., the length of time for which the model has been run since initialization, and therefore the degree to which it has approached its own equilibrium) has the potential to exert a significant control over the statistical metrics calculated for each model, using those observations.

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 from either climatologies (derived from the observed fields...
or previous model simulations) or spatially 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). The equilibration time can be different in
coupled model configuration (i.e., ocean-atmosphere general circulation models or
ESMs) compared to stand-alone climate components due to leaks in the energy budget
(Hobbs et al., 2016). In practice, these simulations, called “spin-ups”, often span in
general only several hundred years, at the end of which a quasi-equilibrium state is
assumed for the interior ocean tracers. The present degree of complexity and spatial as well as temporal resolution of marine
biogeochemical ESM components (as well as other physical and chemical
components), however, often precludes a spin-up to reach adequate equilibration of
biogeochemical tracers. This is a consequence of the large 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 spatial as well as temporal resolution. This
number of state variables 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 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). So far, 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 in space and time in spite of considerable efforts
such as the GEOTRACES program for trace elements, or MAREDAT for biomasses
of plankton functional types. The latter set of variables is 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 CO₂ or the excess of nutrients inputs and pollutants
(e.g., Doney, 2010). Although methods exist to remove the anthropogenic
perturbation from some 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). For the deep water masses, this time is about 1500 years in the Atlantic Ocean and reaches up to 10000 years in the North Pacific Ocean (Wunsch and Heimbach, 2008).

In a context of model-to-model intercomparison, this time range contributes to the model uncertainty. Lessons from the previous Ocean Carbon Model Intercomparison...
Project phase 2 (OCMIP-2) exercise have demonstrated that some models required ~10,000 years to reach a state where the global sea-air carbon flux is about 0.01 Pg C yr$^{-1}$.

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 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 in 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.

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, individual modeling groups
have discussed the impacts of their particular spin-up procedure on model
performance individually (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 controlling the ‘final’
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.

2- Methods

2-1- Model simulations

This study exploits in particular results from one simulation performed with IPSL-CM5A-LR (Dufresne et al., 2013), considered here to be representative of the likely behavior of other CMIP5 Earth system models. Like other 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° x 1.87° 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
the 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-ΔO₂ ratios after
Takahashi et al. (1985) are 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.

The model’s 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 burial 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
Riverine carbon inputs induce a natural outgassing of carbon of 0.6 Pg C y\(^{-1}\) which has been shown to be essential to model the inter-hemispheric gradient of atmospheric CO\(_2\) under preindustrial state (Aumont et al., 2001).

The core simulation used within 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 given 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 including data more recently collected than that made us of in WOA94. 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 ensemble of industrial-revolution to present-day simulation (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 at different depth levels, 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:

\[ \text{drift}(t) = \text{drift}(t = 0) \times \exp\left(-\frac{t}{\tau}\right) \]  

where \( \tau \) 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 (\( \text{NO}_3 \)), dissolved oxygen (\( \text{O}_2 \)) 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, $[\text{CO}_3^{2-}]$, 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 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 $[\text{CO}_3^{2-}]$, e.g., is not explicitly transported as a tracer with the velocity fields, 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 \( \sim 10^{-4} \, ^\circ C \, y^{-1} \) which falls into the range of the
drifting reported for CMIP5 ESMs (Sen Gupta et al., 2013). 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).

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

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)) to evaluate
the global sea-to-air carbon flux simulated by IPSL-CM5A-LR. 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 \( \pm 0.08 \, Pg \, C \, y^{-1} \) (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 \( Pg \, C \)
Figure 2b shows that the global sea-to-air carbon flux is still lower than the range of values estimated from preindustrial natural ocean carbon flux inversions (~0.4–0.56 Pg C y\(^{-1}\)). Besides, Figure 2b shows that the drift in the global sea-to-air carbon flux becomes smaller more slowly after a strong decline during the first 50 years of the spin-up simulation. From year 250-500 this drift is about 0.001 Pg C y\(^{-2}\) and still weaker over the last century of the simulation (7\times 10^{-4} Pg C y\(^{-2}\)). A one-sided t-test indicates that the two drifts differ from each other with a p-value < 2\times 10^{-16}. When fitted with drifts computed from overlapping time segments of 100 years, 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 2\times 10^{-7} and 7\times 10^{-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 $NO_3$ presents persistent deviations from WOA2013. At the surface, the convergence of the simulated oxygen to observed value is expected since the 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). 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 little 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, substantial 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 patterns of errors for O$_2$, NO$_3$, and Alk-DIC fields are well correlated with each other (R>0.6). This reflects that in PISCES carbon, nitrogen and oxygen concentrations are linked by the elemental C:N::O$_2$ stoichiometry fixed in space and...
Figure 6 shows that model-data deviations at 2000 m have substantially increased at a regional level 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 RMSE 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 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 a few decades after 250 years of spin-up 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 upper 1000 m water masses evolve relatively fast (within a few centuries) due to the relatively short mixing time in the upper ocean. Mid-depth (~1500-2500 m) RMSE evolves much slower because of the slow ocean circulation at these depth levels. Characteristics time scales here are thousand of years as evidenced by the observed radiocarbon age of seawater (e.g., Wunsch and Heimbach, 2007; 2008). This explains why, at the end of the spin-up simulation, two maxima of comparable amplitude are found for RMSE at 150 m and 3750 m for O$_2$ and at 50 m and 3800 m for Alk-DIC (Figure 7).

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, $\tau$, which characterizes the e-folding time scale of the RMSE. 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 to the 80 drift values for each field and each depth level (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 ($r^*=0.3$ determined with a student distribution with significance level of 90% and $\sim 15$ effective degrees of freedom estimated with the formulation of Bretherton et al., (1999)), except for $NO_3$ at surface and Alk-DIC at 150 m and 2000 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 $\tau$ for $O_2$ RMSE are $\approx 90, 564$ and 1149 y at the surface 150 m and 2000 m, respectively. These values match reasonably well $\tau$ 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, $\tau$ estimates for $O_2$ RMSE are $\approx 114\pm 67, 375\pm 140$ and $1116\pm 527$ y at the surface 150 m and 2000 m depth. These large uncertainties associated with $\tau$
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 O2 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 O2 RMSE versus the spin-up duration for each CMIP5 ESM. The analysis shows that the drift in O2 RMSE differs substantially between models. For a given model, drifts in other biogeochemical tracers (NO3 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 O2 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). Indeed, the drifts in RMSE decreases in course of time at the various depth levels for the IPSL-CM5A-LR model, 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
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, \( \Delta \text{RMSE} \), to the above-
mentioned RMSE. This incremental error due to the drift is computed using the
relaxation time $\tau$ 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 ($m$):

$$\Delta \text{RMSE}_m(z) = \int_0^T \text{drift}_m(z, t=0) \times \exp\left(-\frac{1}{\tau(z)} t\right) dt$$

(2)

where $\Delta \text{RMSE}$ 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 $\Delta \text{RMSE}$ 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.

The 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 3th 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 $\Delta \text{RMSE}$
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 main physical climate fields as well as the large-scale ocean circulation reach quasi-equilibrium after 250 years of spin-up, but our analyses indicate that biogeochemical tracers do not (Figure 3).

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).

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)).
mechanisms can be nonetheless invoked to explain differences or similarities in behavior between biogeochemical fields. For example, the evolution of surface concentrations for O\textsubscript{2} and Alk-DIC is controlled in part by the solubility of O\textsubscript{2} and CO\textsubscript{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\textsubscript{2} and Alk-DIC surface fields once the physical equilibrium is to a large degree 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 model (Figure 4, panels b and f). By contrast, the evolution of NO\textsubscript{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\textsubscript{2}, NO\textsubscript{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 have important implications 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 Vancoppenolle et al. (2013) and Laufkötter et al. (2015) explain a part of the difference in simulated nutrient concentration by the differences in the spatial resolution and the complexity of the models, the authors of both studies qualitatively invoked differences in spin-up duration to explain the remaining differences in simulated concentrations. 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 strategies used to spin-up models in CMIP5 have resulted in
significant source of uncertainty to the multi-model spread. 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 on 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 frequently used approach relies on long preindustrial control simulations running parallel to transient simulations, allowing the ‘removal’ of the drift 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 one 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 perturbation 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.
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 standard 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., (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 $\tau$ at regional scale. Moreover, the combination of regional ocean physics and biogeochemical processes in each individual model may drive an evolution of a 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.

Besides, difference in the simulated processes and resolution in the different models 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 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
Indeed, as shown in Kriest and Oschlies (2015), various parameterizations of the particle sinking speed 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 behavior within a time-scale of $O(10^3)$ years. As such, drift and $\tau$ estimates need to be used with caution when computed from short spin-up simulations because they can be subject to large uncertainties. 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.

5- Conclusions and recommendation for future intercomparison exercises

Skill-score metrics are expected to be widely used in the framework of the upcoming 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 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 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 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 help to disentangle physical from biogeochemical contribution to the inter-model spreads, but at the same time, could also potentially complicate the determination of inter-model spreads by increasing the diversity of spin-up protocols. 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 intercomparison 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.

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 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 makes 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.

References:


Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung,


Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M. and Bladé, I.: The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field, J. Climate,


The importance of realistic physics, Global Biogeochem. Cycles, 18(3),

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.,
upper ocean ecosystem-biogeochemistry models against field and remote sensing
data, Journal of Marine Systems, 76(1-2), 95–112,

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

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,

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., Etche, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P.,
Grandpeix, J. Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourtin, F., Idelkadi, A.,
Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetteulle, S., Lallement, 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 Vuchard, N.: Climate change projections using the
IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim Dyn, 40(9-10), 2123–

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.,
Malshevi, S. L., Milly, P. C. D., Philipps, P. J., Sentman, L. A., Samuels, B. L.,
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.

fast did the ocean—atmosphere system run during the last deglaciation? Earth

Eyring, V., Righi, M., Evaldsson, M., Lauer, A., Wenzel, S., Jones, C., Anav, A.,
Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehlbrecht, 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.

Folloc, M. J., Ductkieviez, S., Grant, S. and Chisholm, S. W.: Emergent
Biogeography of Microbial Communities in a Model Ocean, Science, 315(5820),

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), 33

Ciotti, A. M., Doney, S. C., Dowell, M., Dunne, J. P., Gentili, B., Gregg, W.,
Hoepfner, N., Ishizaka, J., Kameda, T., Lima, I., Marra, J., Mélín, F., Moore, J. K.,
Morel, A., O'Malley, R. T., O'Reilly, J., Saba, V. S., Schmeltz, M., Smyth, T. J.,
uncertainties of model estimates of primary productivity in the tropical Pacific Ocean,

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

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.,


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.
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.,
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
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
Mikaloff Fletcher, S. E., Gruber, N., Jacobson, A. R., Gloo, 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
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), −,
Moore, J., Doney, S., Kleypas, J., Glover, D. and Fung, I.: An intermediate
complexity marine ecosystem model for the global domain, Deep Sea Research Part
2002.
General Circulation Models: Climate Simulation Meets Weather Prediction, Bull.
Resplandy, L., Bopp, L., Orr, J. C. and Dunne, J. P.: Role of mode and intermediate
Resplandy, L., Séférian, R. and Bopp, L.: Natural variability of CO 2and O 2fluxes:
What can we learn from centuries-long climate models simulations? Journal of
Rodgers, K. B., Lin, J. and Fröhlicher, T. L.: Emergence of multiple ocean ecosystem


Séférian, R., Ribes, A. and Bopp, L.: Detecting the anthropogenic influences on recent


Wu, T., Li, W., Ji, J., Xin, X., Li, L., Wang, Z., Zhang, Y., Li, J., Zhang, F., Wei, M.,
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,
Wunsch, C. and Heimbach, P.: Practical global oceanic state estimation, Physica D:
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,
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-
Zeebe, R. E. and Wolf-Gladrow, D. A.: CO2 in seawater: equilibrium, kinetics,
<table>
<thead>
<tr>
<th>Models</th>
<th>spin-up procedure</th>
<th>initial conditions</th>
<th>offline time</th>
<th>online time</th>
<th>total spin-up duration</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCC-CSM1-1</td>
<td>sequential</td>
<td>WOA2001, GLODAP</td>
<td>200</td>
<td>100</td>
<td>300</td>
<td>(Wu et al., 2013)</td>
</tr>
<tr>
<td>BCC-CSM1-1-m</td>
<td>sequential</td>
<td>WOA2001, GLODAP</td>
<td>200</td>
<td>100</td>
<td>300</td>
<td>(Wu et al., 2013)</td>
</tr>
<tr>
<td>CanESM2</td>
<td>sequential (forced w/ obs.)</td>
<td>OCMIP profiles, CanESM1</td>
<td>6000</td>
<td>600</td>
<td>6600</td>
<td>(Arora et al., 2011)</td>
</tr>
<tr>
<td>CESM1-BGC</td>
<td>direct</td>
<td>CCSM4</td>
<td>0</td>
<td>1000</td>
<td>1000</td>
<td>(Lindsay et al., 2014)</td>
</tr>
<tr>
<td>CMCC-CESM</td>
<td>sequential (w/ acc.)</td>
<td>WOA2001, GLODAP</td>
<td>100</td>
<td>100</td>
<td>200</td>
<td>(Vichi et al., 2011)</td>
</tr>
<tr>
<td>CNRM-CM5</td>
<td>sequential</td>
<td>WOA1994, GLODAP, IPSL</td>
<td>3000</td>
<td>100</td>
<td>3100</td>
<td>(Séférian et al., 2013)</td>
</tr>
<tr>
<td>CNRM-CM5-2</td>
<td>sequential</td>
<td>WOA1994, GLODAP, CNRM</td>
<td>3000</td>
<td>100</td>
<td>3100</td>
<td>(Schwinger et al., 2014)</td>
</tr>
<tr>
<td>CNRM-ESM1</td>
<td>sequential</td>
<td>CNRM-CM5</td>
<td>0</td>
<td>1300</td>
<td>1300</td>
<td>(Séférian et al., 2015)</td>
</tr>
<tr>
<td>GFDL-ESM2G</td>
<td>direct</td>
<td>WOA2005, GLODAP</td>
<td>0</td>
<td>1000</td>
<td>1000</td>
<td>(Dunne et al., 2013)</td>
</tr>
<tr>
<td>GFDL-ESM2M</td>
<td>direct</td>
<td>WOA2005, GLODAP</td>
<td>0</td>
<td>1000</td>
<td>1000</td>
<td>(Dunne et al., 2013)</td>
</tr>
<tr>
<td>GISS-E2-H-CC</td>
<td>direct</td>
<td>WOA2005, GLODAP DIC*</td>
<td>0</td>
<td>3300</td>
<td>3300</td>
<td>(Romanou et al., 2013)</td>
</tr>
<tr>
<td>GISS-E2-R-CC</td>
<td>direct</td>
<td>WOA2005, GLODAP DIC*</td>
<td>0</td>
<td>3300</td>
<td>3300</td>
<td>(Romanou et al., 2013)</td>
</tr>
<tr>
<td>HadGEM2-CC</td>
<td>sequential</td>
<td>HadCM3LC, WOA2011</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>(Collins et al., 2011; Wassmann et al., 2010)</td>
</tr>
<tr>
<td>HadGEM2-ES</td>
<td>sequential</td>
<td>HadCM3LC, WOA2010</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>(Collins et al., 2011)</td>
</tr>
<tr>
<td>INMCM4</td>
<td>sequential</td>
<td>Uniform DIC</td>
<td>3000</td>
<td>200</td>
<td>3200</td>
<td>(Volodin et al., 2010)</td>
</tr>
<tr>
<td>IPSL-CM5A-LR</td>
<td>sequential</td>
<td>WOA1994, GLODAP</td>
<td>3000</td>
<td>600</td>
<td>3600</td>
<td>(Séférian et al., 2013)</td>
</tr>
</tbody>
</table>
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 $\Delta C^*$ method. w/ acc. and forced w/ obs. indicates the strategy using 'acceleration' and observed atmospheric forcings during the spin-up, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Strategy</th>
<th>Source</th>
<th>Duration (yr)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPSL-CM5A-MR</td>
<td>sequential</td>
<td>WOA1994, GLODAP, IPSL</td>
<td>3000 300 3300</td>
<td>(Dufresne et al., 2013)</td>
</tr>
<tr>
<td>IPSL-CM5B-LR</td>
<td>sequential</td>
<td>IPSL-CM5A-LR</td>
<td>0 300 300</td>
<td>(Dufresne et al., 2013)</td>
</tr>
<tr>
<td>MIROC-ESM</td>
<td>sequential</td>
<td>GLODAP/constant values</td>
<td>1245 480 1725</td>
<td>(Watanabe et al., 2011)</td>
</tr>
<tr>
<td>MIROC-ESM-CHEM</td>
<td>sequential</td>
<td>GLODAP/constant values</td>
<td>1245 484 1729</td>
<td>(Watanabe et al., 2011)</td>
</tr>
<tr>
<td>MPI-ESM-LR</td>
<td>sequential</td>
<td>HAMOCC/constant values</td>
<td>10000 1900 11900</td>
<td>(Ilyina et al., 2013)</td>
</tr>
<tr>
<td>MPI-ESM-MR</td>
<td>sequential</td>
<td>HAMOCC/constant values</td>
<td>10000 1500 11500</td>
<td>(Ilyina et al., 2013)</td>
</tr>
<tr>
<td>MRI-ESM1</td>
<td>sequential</td>
<td>(forced w/ obs.)</td>
<td>GLODAP</td>
<td>550 395 945</td>
</tr>
<tr>
<td>NorESM</td>
<td>direct</td>
<td>WOA2010, GLODAP</td>
<td>0 900 900</td>
<td>(Tjiputra et al., 2013)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>O₂</th>
<th>NO₃</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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).

<table>
<thead>
<tr>
<th>Depth</th>
<th>surface</th>
<th>150 m</th>
<th>2000 m</th>
<th>surface</th>
<th>150 m</th>
<th>2000 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>7.19</td>
<td>8.75</td>
<td>5.50</td>
<td>2.07</td>
<td>2.90</td>
<td>2.08</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
<td>0.92</td>
<td>0.94</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>O$_2$</th>
<th>NO$_3$</th>
<th>Alk-DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>metrics</td>
<td>mean</td>
<td>RMSE</td>
<td>RMSE$_{max}$</td>
</tr>
<tr>
<td>Surf</td>
<td>-0.2</td>
<td>2.6</td>
<td>55.8</td>
</tr>
<tr>
<td>150 m</td>
<td>3.4</td>
<td>39.0</td>
<td>31.5</td>
</tr>
<tr>
<td>2000 m</td>
<td>30.4</td>
<td>144.3</td>
<td>-40.1</td>
</tr>
</tbody>
</table>
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).

Figure 3: Time series of globally averaged concentration (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. Globally averaged concentration 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 globally averaged concentration.

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 beginning 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 regressions between drifts in 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. Least square correlation coefficients are tested against a one-tailed t-test with significance level of 90% and ~15 effective degrees of freedom estimated with the formulation of Bretherton et al. (1999); * indicates if a given least square correlation coefficient passes the test.

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 beginning every 5 years from the beginning until the end of the simulation.
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 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 ($\Delta$RMSE) 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. $\Delta$RMSE 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.