In the following, referee comments are in black, while our responses are in green and added material is indicated in blue.

Referee 2

In this manuscript, Yool et al. present in detail the performance of the ocean and marine biogeochemical component of UKESM1, a novel Earth system models contributing to CMIP6. The manuscript is clearly written and provides a large array of standard analysis to understand the performance of UKESM1 at replicating observed features of the ocean and marine biogeochemical dynamics. This work presents an important basis for the users and other climate research groups in the context of multi-model analysis. I only have three majors and a set of minor comments that aims to clarify some point of the paper.

Again, we would like to first thank the referee for their careful reading of what is a very long manuscript, made much longer by its supplementary material. Thank you!

Major comments:

1- Although the authors did a great job assessing model modern climatology (2000-2009) against available modern observations, they didn't provide any key statistical metrics that might be useful to support their assessment (correlation, total root-mean squared errors, etc.). This would represent an added value in the current manuscript In addition, The abstract states that this paper investigate the driving mechanisms behind model-data errors. This is misleading. The manuscript remains largely speculative on what causes model biases and what mechanisms are at play to explain errors propagation or amplification. Within digging to much in those mechanisms (I reckon the paper is already long), I would like to see more properties-to-properties diagrams. This kind of analysis would strengthen the manuscript and support the conclusions.

While we appreciate that our analysis is incomplete, we are uncertain as to the sort of additional analysis that the referee would like to see (e.g. specific property-property plots). We are happy to act on further guidance here. Given the already long nature of this paper, we likely prefer to defer a detailed analysis of the root cause of different biases to subsequent papers.

Those limited range of analysis mirrors the paper structure. Indeed there is no Methodology section. The reader remains without information on how model data are compared to observation (regridding for instance); how the mixed-layer depth are calculated; what is the working hypothesis to compute anthropogenic carbon and so on. This section, even short, would be useful.

We have addressed these omissions by adding the following paragraphs to the "Datasets and Evaluation" subsection of the manuscript's Methods section.

"In addition, several derived variables are calculated from observational and model fields.

– Mixed layer depth (MLD) is calculated in the same way from both observed and modelled 3D fields of potential temperature. MLD is determined to be the depth at which the vertical profile of potential temperature is 0.5C lower than that at the depth of 5 m. Alternative MLD schemes using similar thresholds in potential density (either fixed or variable with temperature) were also examined, but global coverage was less complete with these (especially in sea-ice regions), so the potential temperature criterion was favoured.

– Modelled integrated AMOC and Drake Passage transports are calculated here using the BGC-val toolkit (de Mora et al., 2018). In the case of AMOC, the calculations are based on those of Kuhlbrodt et al. (2007) and McCarthy et al. (2015) and use the cross-sectional area at the 26N transect to calculate the maximum depth-integrated current. Drake Passage transport is calculated following Donohoe et al. (2016) as the total, depth-integrated current along a north-south transect between the South American continent and the Antarctic Peninsula. The methods for both transports are described in de Mora et al. (2018).

– Model anthropogenic CO2 is estimated by differencing DIC fields from the Historical simulation of each ensemble member with the corresponding DIC field from the piControl at the same relative timepoint. For example, we estimate anthropogenic CO2 in 1990 from a given Historical ensemble member as the difference between this member's DIC field at this particular time and the DIC field from the piControl simulation from the same timepoint, i.e. the time that corresponds to 140 years (i.e. 1990 - 1850 = 140) after the Historical ensemble member branched from the piControl. This approach aims to account for drift in the simulations, although it omits changes driven by divergence in circulation and biogeochemistry between the Historical and piControl simulations. These are assumed to be small in this method."

"Throughout, fields of observational and model properties are plotted on their original horizontal and vertical grids. Where these properties are directly intercompared, for instance in difference plots, observational fields are first regridded to the model grid (using the scatteredInterpolant function of Matlab v2020a). In Section 4.3, horizontal fields of UKESM1 output are compared with those from fellow CMIP6 models, and here all models are regridded to a common, uniform 1° grid."

2- Following major comment #1, I find it difficult to understand the choices of the authors team for the analysis. For instance, why DJF or JJA are taken for ocean analysis instead of JAS more widely accepted to studied the Southern Ocean winter dynamics. Does it has something to do with biases in the atmospheric model? Further explanation would be useful.

The choice of DJF and JJA is simply following meteorological conventions for "winter" and "summer". We have added the following text into the Model Evaluation section:

"A number of figures illustrate observed and modelled properties (and the biases of the latter) for the June-July-August (JJA) and December-January-February (DJF) meteorological seasons that correspond respectively to northern hemisphere summer and winter (and southern hemisphere winter and summer)."

3- Manuscript structure: Although the manuscript is well written, some parts could be improved. For instance, I find it unexpected to find the description of the ocean and marine biogeochemical model in the appendix (while central in this work) but not in the main text to guide analyzes. Besides, the opening to the future scenarios seems out of the scope in the manuscript. This latter could be removed to give more space to develop key aspects (see point 1).

Both the ocean physical and biogeochemical submodels are outlined in the main body. While the ocean physical model is documented more completely elsewhere (and this is noted via cited papers), the precise version of the biogeochemical model used here is not, hence the appendix sections. However, we have retained it as supplementary so as not to "break the flow" of the manuscript by adding a large block of text descriptive of a single submodel.

Minor comments:

L6 observational properties = observed fiels ?

Yes, this is more accurate. Amended.

L15 compares favourably = you mean “outperforms” ? or just compares to other models?

This is a fair point. The model's actual performance is mixed, so outperforms sometimes (e.g. surface DIC, ALK), underperforms in others (e.g. surface DIN). Overall, "favourably" tried to capture this, but not well. We have rewritten this to:

"performs well alongside its fellow members of the CMIP6 ensemble"

L29 in response to the release = driven by the release

More accurate. Amended as suggested.

L30 chemical composition = CO2 airborne fraction

This is better. Amended as suggested.

L57 to identify avenues for future: : : addressing model limitation and weakness

Yes, this was rather clunky. We have reworded to:

"Third, to identify avenues for addressing model limitations and weaknesses in future versions"

L70 built to “simulate”

This is better. Amended.

L80 do you simply mean that land-surface model is a submodule of the atmosphere model ?

The suggested phrasing downplays the actual separation of the land and atmosphere submodules (they can be run independently). To address the confusion, we have reworded to:

"In outline, UKESM1 is comprised of closely-coupled atmosphere and land submodules that are linked through an explicit coupler module, OASIS3-MCT_3.0, to coupled ocean and sea-ice submodules."

L92 Mulcahy et al. (subm.) I don’t know if it match GMD standard to refer to submitted papers (I counted two papers with this status)

One of the three submitted manuscripts is now accepted, and its entry in the bibliography has been updated to reflect this. The two others are in late stages of pre-publication and we anticipate publication ahead of this manuscript. should this not occur, we will revise to replace with appropriate alternative publications.

L102-123 I would further develop this paragraph because it describes the model description central to this paper. In addition I suggest the authors to include the MEDUSA description here. Further details on how couplings may influence MEDUSA results (for instance, dust deposition).

Per our previous remarks, we have decided to retain the description of MEDUSA within the appendix. However, the referee's suggestion that we identify the specific couplings between MEDUSA and the rest of the model is good, and we have added the following text:

"Within UKESM1, MEDUSA interacts with other model components via the following feedback connections: atmosphere-ocean exchange of CO2; ocean-to-atmosphere fluxes of DMS and PMOA; deposition of terrestrial iron to the ocean via atmospheric dust transport."

L145 physics=hydrodynamics

We have amended "physics" to "physical" - T and S (the variables in question) are more state than hydrodynamics.

L150 it is unclear what is meant with "ocean circulation", please provide further detail

ECCO essentially estimates the patterns and magnitudes of ocean circulation. For clarity, we have altered this to: "hydrodynamic circulation state".

L162: Please include Khatiwala et al. (2009) dataset

We will add this dataset to the final version of the Zenodo archive that accompanies this manuscript.

L175 "% citep: : :" I guess there is a typo here. Please check the following sentence

Thank you. Amended.

L215 "thermohaline transects" or "thermohaline circulation" Figures: this naming is misleading. What about "basin-averaged section" ?

We have retained the description, but have added the following to make it clear how the figures have been created

"These transects are created from basin zonal means of the plotted properties."

L213 Section 3.2 please include detail on how to determine NADW and AABW properties in the model

We have not determined NADW and AABW properties directly during our analysis. Rather, the model's meridional overturning circulation (MOC) is calculated by integrating its velocity field in the same manner as the ECCO product. As the watermass structure is very similar between UKESM1 and ECCO, we use the same terminology, although our analysis does not go as far as formally identifying different watermasses. We do not judge that this level of in-depth analysis is necessary at this point.

L254: RAPID-MOCHA time coverage poorly matches with the time period chosen for computing the model climatology. Does this difference in time period influences model assessment?

The ensemble mean ranges 16-17 Sv in the period of RAPID-MOCHA (with variability 15-18 Sv). This lies within the range (14.6-19.3 Sv) observed at the array. To make the comparison clearer, we have added the RAPID-MOCHA time-series to the appropriate panel of Figure 9, and altered the text appropriately. Because the RAPID-MOCHA era is short compared to the full Historical period, we have also added Supplementary Figure S8 that focuses in on this period.

L295 "63 vs." ? please check

The values for Southern Ocean silicic acid quoted in the manuscript are correct. The model estimates both higher maximum values, and a much larger area of high concentrations within the Southern Ocean. A compounding factor in the numbers appearing questionable is a mistake in the description of the regions. This has been amended to more clearly describe the Southern Ocean.

"In the southern hemisphere, the subpolar region falls primarily within the Southern Ocean, although as its northern margin is delineated at -50N rather than -40N, the southern margins of the southern subtropical Atlantic and Pacific extend to -50N."

L317 biological community = marine biology ?

Yes, this is clearer. Amended.

L334 three models = three algorithms ?

Yes, they are much simpler models than the dynamic models within ESMs so "algorithm" is more accurate. Amended.

L348-362: All the Figures referred in these paragraphs should be included in the manuscript. Otherwise please consider moving this paragraph and associated text in the suppl. Mat.)

We agree. While other supplementary figures augment the description of properties which already feature in main body figures, these two figures do not. We have moved them into the main body accordingly and amended the text to reflect this.

L596: Section 4.2: While interesting, I would suggest to stratify this section by comparing first NEMO-based model (IPSL, CNRM, CanESM) and then the other models. This would help to discuss the

performance of the marine biogeochemical in a more constrained modelling framework. Please note that the Danabasoglu et al. and Voldoire et al. are inaccurate.

The aim of this section is simply to place the performance of UKESM1 within the context of fellow CMIP6 models. Identifying and attributing common patterns of bias is beyond the scope of this manuscript (but see Seferian et al. 2020; cited here). In any case, the behaviour of these three models is not completely convergent in the plots shown, and we judge that it would not be straightforward (or brief) to attempt to identify parallels. However, the referee's point around shared components is worth emphasising, and we have added the following text to this section:

"While the full configurations of these models are diverse, the CNRM-ESM2-1, CanESM5, and IPSL-CM6A-LR models share a common NEMO physical ocean with UKESM1, though they diverge on other components, including marine biogeochemistry."

We have also tried to amend the references to the correct ones.

L681-695: Section 5: some key points would require further detail: resolution dependent surface biases : is it assessed in this paper or based on published literature ?; same does for the aerosol-driven strengthening of the AMOC.

Both points are made elsewhere in the manuscript, and supported by published analyses. See 187-189 for resolution dependence, and 299-303 on aerosol forcing. They have not been separately evaluated here.

Figure 1: please detail somewhere in the ms why JJA and DJF has been used for seasonal analyses

See earlier comments.

Figure 2: 0.15 isolines may help to compare model and observation

Following a suggestion by another referee, the figure has been modified so that model values below 0.15 are omitted (the colour bar is altered to reflect this). This permits easier comparison of the model and observations. (A similar modification is made for the sea-ice thickness supplementary figure.)

Figure 3: please considering remove indiv ens members and show 1 standard deviation. Please explain what is behind seasonal minima/maxima.

We would agree that the figure gets rather dense from 1980 onwards when the observational data is also plotted. We have followed the referee's suggestion to amend this.

Regarding the seasonal minima and maxima, these are simply the months in which sea-ice usually records its minimum and maximum extents. We do not follow the question.

Figure 6: please see the comments above

See earlier comments.

Figure 17: further discussion would be required to explain why model and obs-derived estimates differ over the preindustrial period (1800-1850) and why models fail at capturing carbon uptake over the recent years (2000-2014). Please include the database in the reference list

On the first point, CMIP6 starts its Historical era in 1850, approximately a century after the industrial revolution began. The Khatiwala et al. (2009) product estimates CO2 uptake over the more complete period during which atmospheric pCO2 has been elevated by anthropogenic emissions.

On the second point, UKESM is typically in the lower portion of the uncertainty range of the Khatiwala et al. (2009) dataset throughout the Historical era. The seeming greater divergence at the very end of the simulated period may be exaggerated by a sudden uptake spike in the dataset, which itself is an estimate. However, to resolve this would require a detailed analysis of factors in the model (e.g. SST, SSS, wind speed) that we judge is beyond the scope of this manuscript.

Nonetheless, the description of UKESM1's comparison with Khatiwala et al. (2009) does not note that it falls within the lower portion of the uncertainty range throughout the Historical period. The text has been amended so that this is clear to readers.

"With some variability, particularly in the early decades, the ensemble tracks the observationally-estimated uptake, reproducing the same pace and features, but with the ensemble estimating a slightly lower flux than estimated (88.5%; integrated 1850-2013)."

"The plot also shows UKESM1's piContol simulation to illustrate the magnitude and period of variability with constant background atmospheric CO2. This shows CMIP6's Historical period beginning (and the piControl period ending) in 1850, approximately a century after the industrial revolution and significant fossil fuel CO2 emissions began. This differs from the Khatiwala et al. (2009) product, which estimates ocean CO2 uptake over the more complete period of anthropogenic emissions."

Finally, we will include the Khatiwala et al. (2009) dataset within the Zenodo archive that accompanies this manuscript.

Figure 18: please explain what is your working hypothesis to compute anthropogenic carbon

Historical ensemble members branch off from the piControl at different time points in its evolution. In a Historical simulation, the 1990s occur 140-149 years after this branch point. So, to estimate anthropogenic DIC concentration, the 1990s decade of each Historical ensemble member are differenced from the piControl decade 140-149 years after the branching point of this ensemble member. Other alternative approaches are possible, but we chose this approach here to account for drift in the piControl.

While lines 399-400 already describe the approach used, the description is very short and may be misunderstood. To avoid this, we have added the following text to the Methods section:

"Model anthropogenic CO2 is estimated by differencing DIC fields from the Historical simulation of each ensemble member with the corresponding DIC field from the piControl at the same relative timepoint. For example, to estimate anthropogenic CO2 in 1990 from a particular Historical ensemble member, we subtract the piControl state from the timepoint that corresponds to 140 years (i.e. 1990 - 1850 = 140) after the Historical ensemble member branched from it."

Figure 20: typo on the Figure “CO 2”

Amended.

Figure 21 and 22: These two Figures merits further explanations. They show major model biases in the subsurface Atlantic waters whereas the modelled meridional transport matches well with the observed one.

Modelled transport in UKESM1 is reasonable, although it is clear from a several properties (salinity, DIN, oxygen) that AABW is more pronounced in UKESM1. It is also clear that the AAIW limb of fresher water from the Southern Ocean is less pronounced than observed, transporting less nutrient to the North Atlantic.

The balance of watermasses - and their constituents - in UKESM1 suggests that nutrients are more efficiently "trapped" in AABW, decreasing the concentrations in upper watermasses and contributing to the pattern observed.

in the case of DIC, the biases match those of DIN, but with an additional negative bias imparted by reduced surface DIC concentrations. This is already described in the manuscript as a consequence of a negative surface alkalinity bias.

The manuscript has been modified around the distributions of DIN as follows:

"Meanwhile, in deeper waters the bias is reversed to strong positive, as the more sluggish AABW circulation shown in Figure 8 is ventilated less efficiently, accumulating excess DIN while accruing an oxygen deficit (Supplementary Figure S16)."

Technical suggestion:

please consider to adjust color scales for some figures and use red-green color-blind color palette (where relevant)

This is a fair point. To address this, we have adopted the perceptually-neutral Google "Turbo" palette where appropriate, and have retained the blue-white-red palette for "delta" plots. The "Turbo" palette is described, and its perceptual qualities evaluated, here ...

https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html

A notice about the palette is included in the acknowledgements to assist others wishing to use it.