The manuscript "Ocean biogeochemistry in the Canadian Earth System Model version 5.0.3: CanESM5 and CanESM5-CanOE" provides critical documentation for the CanESM5 and CanESM5-CanOE contributions to the 6th phase of the Coupled Model Intercomparison Project used in the IPCC 6th Assessment and as a general community research tool for studies in coupled carbon-climate change. The manuscript compares these models to the group's previous generation CMIP5, CanESM2, as well as the CMIP6 ensemble mean to demonstrate marked increase in skill with CanESM5-CanOE improving both biogeochemical skill while also increasing in comprehensiveness for interactive elemental cycles as a 'state of the art' model. As such, it is an important contribution to the scientific literature and should be published with minor revision to address technical points below.

We thank the reviewer for a thorough and constructive review. As we understand it, this is a new reviewer who did not review the original submission. So in a few places we have provided context in terms of explaining why some things were changed or added in response to previous reviews.

38-41 – The phasing of this sentence is confusing. I would shorten it to, "Cumulative ocean uptake of anthropogenic carbon dioxide through 2014 is lower in both CanESM5-CanOE (122 PgC) and CanESM5 (132 PgC) than in observation-based estimates (145 PgC) or the model ensemble mean (144 PgC)."

done

46 – It is not clear why the two Arora model application papers are being cited here in a statement about model development… a better reason to cite these papers would be in justifying an assertion that the Canadian models "have been contributing to coupled carbon-climate research" for over a decade and leave the citation of the necessary model development in support of these contributions to the Christian, et al., 2010 citation.

Arora et al 2009 is expendable here, but Arora et al 2011 is not, as it contains key aspects of the CanESM2 model description. Possibly it is not really necessary to cite any of them at the end of this lead-in sentence, but it directs the reader to the most relevant publications for historical background. We deleted Arora et al 2009 and kept the other two.

48 – CANESM5 includes not only a new ocean but updated atmosphere and land, as well… the atmospheric change could easily be more important than the ocean change.

This is addressed in subsequent paragraphs (see below L94, 96). As this is an ocean-focused paper we chose to emphasize the ocean in the first paragraph of the Introduction. But actually the atmosphere did not change nearly as much as the ocean, and the differences in ocean circulation between CanESM2 and CanESM5 are mostly due to the adoption of the new ocean model.

61 – need a comma after "7"

done

62 – The statement "2-3 times greater" seems very vague, computationally speaking. Is the variance between model computational costs really up to 50%? I would have expected a single number, e.g. "2.5 times greater, or much narrower range, e.g. "2.5-2.8 times greater"

Yes this was a simplistic characterization based on the assumption that computation time scales linearly with the number of tracers, with some additional cost associated with the ocean model itself that is the same in each case. In practice it is ~2X. Wording changed to "CanOE is substantially more expensive computationally (19 tracers vs 7, so the total computation time to integrate the ocean model with biogeochemistry is approximately double)".

87 – "CMOC N2 fixation" should be "that in CMOC" to avoid saying "Dinitrogen fixation"/"N2 fixation" twice.

The small amount of extra text here is justified to avoid ambiguity; "that in" here could refer to N2 fixation rate, or to the model N2 fixation parameterization.

88 – "there is no" should be "CMOC does not include"

done

94 – I would remove "ocean" as the authors are not attempting to distinguish which physical changes are due to the updated atmosphere model versus the updated ocean model, and the changes to the atmosphere could easily be more important than the changes to the ocean model.

changed "ocean" to "climate"

96 – Again, I would remove "ocean" as the Swart paper covers the full coupled model, not just the ocean.

Changed to "An overall evaluation of the CanESM5 climate including the physical ocean is given in ..."

144 – What is the evidence for this assertion? Is there a citation that could be provided and quantification, e.g. Orr et al., 2015 https://bg.copernicus.org/articles/12/1483/2015/

146-147 – Again, "This will affect the total ocean inventory of DIC but not the spatial distribution if the model is well equilibrated." Is an unsupported assertion… Is there evidence that the model is well equilibrated?

We agree that both of the assertions are made without any reference to data, but these are fairly trivial points. If global total alkalinity is conserved and atmosphere CO2 is fixed, the model will converge on a total ocean inventory of DIC that is determined by the inventory of alkalinity. If the initial condition is a constant value or a fixed depth profile, it will take many years for the model to converge on its own internal equilibrium distribution of DIC, but this should not ultimately depend on the initial condition. That our models are relatively well equilibrated is demonstrated later in the paper (Table S6). To cite Orr et al 2015 here could be construed as misleading because the group of methods that they tested does not include the exact method employed in CanESM2. So the difference in pCO2 or Omega between the CanESM2 and

CanESM5 carbon chemistry could be a bit larger than it is in Orr et al, but it is still very small compared to the differences in the distribution of DIC between the two models.

149 – While it is true that CanOE is a substantial advance over CMOS, the current wording of "The CanOE biology model is a substantially new model based on the cellular regulation model of Geider et al. (1998)." implies that CanOE is the first of its kind when in reality the equivalent IPSL, UK, NCAR, GFDL and other models have included this scope of ecosystem complexity intrudued by Moore et al., 2004 and proliferated in the CMIP5 era… Rather, this statement should acknowledge that CanOE is not "substantially new", but rather an effort to meet the complexity that has become standard at other modeling centers. Suggest citing Seferain et al, (2020; https://link.springer.com/article/10.1007/s40641-020-00160-0) which documents the scope of BGC complexity in the CMIP6 class of models. The reader should understand that CanOE contains a level of BGC complexity akin to other models in CMIP6.

"a substantially new model" was deleted (see also below 823-824)

Table 1 – Many of the units are missing the element in which they are applied (e.g. kdnf is listed as "mmol m-3 d-1", but I think this should be "mmol N m-3 d-1", though there is some chance it might be "mmol C m-3 d-1").

We reviewed the Table and the only parameters where there is potential ambiguity are K_DNF and K_Fe. The elemental symbols were added to the units of these.

191 –The definition of "E" should have units… W m-2?

units added

287-289 – Given that CanOE represents N2 Fixation in a novel way that does not include PO4 limitation, it would be helpful to expand this discussion of N2 fixation with a couple of sentences and citations. In particular, "Dinitrogen fixation is parameterized as an external input of ammonium dependent on light, temperature and Fe availability, and inhibited by high ambient concentrations of inorganic N" should be augmented to point out how the maximum rate translates to a maximum vertically integrated rate given light availability much lower than typical recycled productivity in N-limited areas (I'm guessing something like ~0.5 mmol N m-2 d-1 for N2-fixation compared to ~3 mmol N m-2 d-1 for NH4 production) and the inhibition term serving to prevent runaway accumulation of fixed nitrogen. How were these scaling factors derived? Observations?, theory?, other models? An attempt to match denitrification and other terms to have a stable N-budget?, other?

The parameterization derives partly from PISCES and partly from CMOC. This is made explicit in the revised text: "The temperature, iron and light limitation terms are based on PISCES (Aumont et al,, 2015); the N-inhibition term is from CMOC (Zahariev et al., 2008) (CMOC implicitly combines nitrate and ammonium into a single inorganic N pool)." We did not make much attempt to fine-tune the parameters because the global total N2 fixation falls within what we considered to be a reasonable range. The global integral is 68 TgN/y in the piControl, increasing to 82 Tg in 1995-2014.

0.5 mmol N m^-2 d^-1 is in the ballpark, although arguably on the high side. If we take HOT (22.75N, 158W) as a relatively well-characterized example, annual DNF was estimated as 40 mmolN m^-2 by Karl et al 1997 (Nature 388: 533). For a 40 m mixed layer with k_PAR=0.045 m^-1 and surface PAR of 150 W m^-2 (approximate summer value for HOT), the light-limitation function (Elim) will be ~0.75. If we assume T=25C, dFe=0.1 nM, and negligible DIN, the T and Fe limitation terms will each be around 0.5 and the N term ~1. So this gives a realized rate of ~0.15 mmol m^-2 d^-1 or an annual total of 54 mmolN m^-2. Of course, we also have to consider the seasonal and diel variation of surface PAR (time mean Elim not equal to Elim calculated for time-averaged irradiance) and its covariance with temperature. But this simple back-of-the envelope shows that the model produces rates in the expected range. The actual mean rate at this location in CanESM5-CanOE is 27 mmolN m^-2 y^-1 (mean of last 20 years of historical run).

316-317 – The statement, "Burial in the sediments is represented as a simple 'on/off' switch dependent on the calcite saturation state (zero when ΩC<1 and 1 when ΩC≥1)." It is unclear how the saturation state calculated given the statement on lines 139-140, "CanESM5 does not solve the carbon chemistry equations in the subsurface layers."… is carbon chemistry solved at the bottom? If so, lines 139-140 should be changed. If bottom saturation state is calculated some other way, this should be specified here.

The reviewer is correct that in CanESM5 carbon chemistry is not solved in the subsurface layers, but in CanESM5-CanOE it is, as stated on 138-139. We added "In CanESM5-CanOE" at the beginning of this sentence to make sure there is no ambiguity.

323 – comma needed after "production"

330 – comma needed before "but"

both added

339 – Is sediment burial of organic nitrogen considered?

No. There is no burial of organic matter: all C and N in detritus reaching the sea floor are returned to the water column as DIC and NO3. We seem to have neglected to state this in the previous drafts, so we added "There is no burial of organic matter; organic matter reaching the seafloor is instantaneously remineralized." to section 2.5

349 – "length scale of about 200 m"… shouldn't this length scale be precisely known rather than "about 200 m"

No, because the parameterization is a bit more complex than a simple exponential decay (see Aumont et al 2015 eqs 85a-c); it approximates an exponential decay but the length scale is not exactly constant across the depth levels. The approximate e-folding length scale is actually closer to 600 m; this has been corrected.

358-360 – Unless there are further constraints to be cited in the calibration, "particles available for it to precipitate onto, and assumes that POC is strongly positively correlated with total particulate matter." should just be "sinking particles" as equation 24 is not based

on “POC” (which includes phytoplankton, bacteria, and some zooplankton) but rather only sinking detritus concentration, making any “assumption” between POC and total particulate matter only relevant insofar as there was some calibration performed which has not been referenced.

Added "detrital" before "POC". The parameterization is based on the concentration of suspended detritus (i.e., the nonliving fraction of POC). It is assumed that total particulate matter (including living biomass and mineral phases like CaCO3 and biogenic silica) scales approximately linearly with Ds+Dl.

373-374 – I interpret the statement “When the total fixed N adjustment is applied, one mole of alkalinity is removed per mole of N added or removed” that any positive and negative adjustments to the N budget are applied as a negative adjustment to alkalinity such that alkalinity will eventually be exhausted as the model runs to equilibrium… is that what was intended?

This was incorrect; thanks for pointing this out. It has been changed to "one mole of alkalinity is added (removed) per mole of N removed (added)". This adjustment is necessary to maintain global conservation of alkalinity because nonphysically adding or removing nitrate short-circuits the overall pathway of N atoms through the 'fixed' pool (N2 fixation -> nitrification -> denitrification) which has a net alkalinity source/sink of 0 (Table S2).

381-384 – Use of “inherently imperfect” in the statement “However, the OPA free surface formulation is inherently imperfect with regard to tracer conservation.” Is vague… is this formulation globally conservative for mass, and/or Salinity, but not locally, or globally non-conservative? Is global non-conservation in mass or Salinity the source of drift for Alk and N, or is Salinity conserved? The statement “losses due to the free surface are generally larger for tracers with less homogeneous distributions” seems inconsistent with the drift being 3 times greater for surface Alk (range of about 25%; 2.2-2.7 eq m-3) versus NO3 (range of orders of magnitude; 0.00001-0.035 mol m-3)… my guess is that the non-conservation for salinity will be similar to that for Alk, and that the smaller effect for N is because much of the ocean has very low values relative to the global mean rather than a “homogeneous distribution”.

Yes the free surface is globally nonconservative, and this drift occurs in the salinity as well. The numbers were expressed as a fraction of the total tracer pool: 0.01%/ky for alkalinity and 0.03% for N. So it is ~3X larger for N. But it is true that the explanation for this discrepancy (the relative inhomogeneity of the surface distribution) is somewhat speculative and not something we have rigorously tested. So this parenthesis has been deleted.

414/Figure 18 – “excluded”… were areas with Sat Chl >1 just set to 1, or really excluded from the comparison? The justification for excluding these observations as “mostly associated with coastal regions not resolved by coarse resolution global ocean models” is weak. A much better justification would be that these areas represent only a tiny fraction of the global ocean and would unduly weight any error calculation. It looks from Figure 18 that the model is able to represent Chl over 1 since these are provided, which makes this

artificial cutoff for the observations quite strange… at least one would expect the model and obs to both be capped at 1.

Figure has been redrawn so that only data <1 mg m^-3 are shown for both model and observations.

423 – Does "0.0011" have units? Is this the average anomaly between the 1x1 and 2x2 maps in mmol O2 m-3?

No there are no units. It is the difference of pattern correlation coefficients calculated for a 1° or a 2° grid on the same depth level. The text has been revised slightly to make sure this is clear.

427-428 – "as 20 year averages are used, internal variability is assumed to have little effect" would be more effectively put as "20 year averages are used to minimize expression of internal variability"

done

431 – need comma before "and"

done

446 – remove "here"

done

461 – "differences of" should be "and differences from that observational product for"

done

Figure 2, 3, and 6 – this figure should include statistics such as bias, r2 and std error or variance so that the reader can quantitatively compare the results. The r2 and relative variance are supplied graphically in the Taylor diagrams, so only the bias need be calculated anew.

Average bias was added to each panel in Figures 2, 3, 6, and 7 (except OBS). Correlation is not really appropriate for anomalies. Standard deviation is arguably misleading to compare to the MEM, i.e., it is higher, but the same might be true for any individual model (as the reviewer notes, the pattern correlation and relative standard deviation are already shown for the depth levels in Figures 4+9). This argument could apply to bias as well (MEM cancels out opposing biases in various models), but we think the reviewer's suggestion to show average bias here is a good one. As expected, O2 and Omega are biased high except for Omega at 3500 m.

531-532 – Looking at Figure 6 and Figure 7, the statement "CanESM5 and CanESM5-CanOE generally compare well with other models and observations." Appears true for 3500 m, but not at all true for 400 and 900 m in Figure 6 or the most of the water column in Figure 7 due to the strong high bias in saturation state that does not appear in the MEM. Of course comparing any single model to the ensemble mean is not a fair comparison. The statement should be revised to leverage quantitative information on model bias, r2, std error metrics. It is unfortunate that there is not more description of DIC and Alk biases to

help the reader understand why the saturation is biased so high. Is it just that the O2 is high and AOU is low? The patterns don't really look similar to me, so maybe not. Is the North Pacific ventilated too much? How does the model compare with natural 14C which was mentioned at the beginning of the manuscript?

This is one of the pitfalls of showing only anomalies. The original submission showed the model values alongside the observed, with the anomalies (model minus observations) as Supplemental, but this was reversed in the first revision. The text states that the patterns "generally compare well" because when you look at the full modelled value the overall spatial pattern is similar to the observational data product or the MEM (Figure S2). The reviewer is correct that the positive bias is considerably larger in CanESM models than in the MEM, and that this is largely due to having a strongly ventilated mid-depth ocean (not only in the North Pacific, there is a substantial bias in the North Atlantic (Figure 6)). The text has been revised to include more information about the nature of the biases.

We have not yet published any dissi14c data. Looking at the "unofficial" data, the distribution is as expected: high in the mid-depth North Atlantic and low in the mid-depth North Pacific (e.g., 1000 m). One could make an argument that recently ventilated water penetrates deeper into the water column in CanESM5 than in the real world e.g. in the western North Pacific subtropical gyre (see e.g. Figure 2). But for this analysis to contribute significantly to understanding the underlying processes we would have to compare with a suite of other models, and there does not seem to be any DI14C in the GLODAP gridded data product.

Caption to Figure 14 – Should "CMIP5" be "CMIP6"?

Yes. Good catch. Thanks

647 – Is "living" necessary here? Aren't all the phytoplankton groups "living"

deleted

720 – "CanESM models are biased low relative to observation based estimates"… is this true even after accounting for the missing mechanisms in an CMIP models as described in Bronselaer et al., 2017 (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017GL074435) of lack of spinup at 1750 pCO2, 1750-1850 uptake, and legacy lack of uptake from the deltapCO2 between 1750 and 1850 leading to an additional 10-30 PgC(depending on integration period) not represented due to the experimental design? I would think the CanESMs would be pretty close to obs after accounting for these limitations of the CMIP experimental design. Also, is the saturation state high bias in these models also associated with a Revelle Factor bias? I would think it could lead to overly high CO2 uptake. It might be helpful to compare surface excess Alkalinity (Alk-DIC) to answer this.

We thank the reviewer for pointing out this reference. We agree that some of the discrepancy between models and observations is due to it being in part an "apples and oranges" comparison. However, this affects other CMIP6 models as well, and both CanESM models are biased low not only relative to the observations but to other CMIP6 models that would share the same

corrections. The following sentence was added to the text: "Some of the difference may be attributable to differences in the way cumulative uptake is calculated in models vs observations (Bronselaer et al., 2017), although this should apply to other CMIP6 models as well."

We added a new Supplementary Figure to show zonal mean DIC-TA and added to the text "CanESM5 and CanESM5-CanOE show a high bias in near surface DIC relative to alkalinity (a measure of the ocean's capacity to absorb CO_2) in the mid-latitudes of both hemispheres (Supplementary Figure S8), which may in part explain the weak uptake of CO_2." If we draw global maps of DIC-TA at the surface they look very similar to the observations or the MEM, but there are slight biases, and the sign of the bias is consistent with the reviewer's hypothesis. The subsurface saturation state is a different question, and maybe not readily explicable in these terms. Possibly the cumulative CO2 uptake over 1850-2014 will depend on the bias in different depth strata and not just at the surface, but it is reasonable to assume that the biases in surface values calculated for e.g., 1986-2005 are fairly stable over time.

785 – "the preindustrial control simulations had different degrees of equilibration when the historical experiment was launched (cf. Séférian et al., 2016, Supplementary Table S6)." The reader should not have to go to the supplemental of a model comparison paper to find the spin up length… however, when I went to Seferain Table 3, I see CanESM5 and CanESM5-CanOE as having been both spun up for 1000 years… so I am confused by this statement.

Table S6 was added at the request of a reviewer of the previous draft, to put our comments about the drift fraction of DIC change in the historical run in context. It shows that the net ocean DIC change in the piControl is approximately 2X larger in CanESM5-CanOE than in CanESM5, but still small compared to anthDIC uptake. As discussed in Séférian et al. (2016), the spin-up was a bit ad hoc: possibly not as bad as in CMIP5, but not really ideal Best Practice either. We spun both models up ocean-only for as long as we could before launching the 1000 y coupled spinup, but the ocean-only spinup was longer for CanESM5 (partly because it runs faster and partly because it was ready sooner).

815 – "It is also possible that the lower export production in CanESM5-CanOE is due to low iron supply to the surface waters of the Southern Ocean, but comparison with available observations do not suggest that this is the case." I don't see the logic for looking at iron biases here… Wouldn't it rather be a high bias in surface N and/or a low bias in the vertical N gradient that would suggest that the biological pup is too weak? The role of iron is really just to scale the degree of HNLC.

Here we are trying to explain the rather large differences in global and especially Southern Ocean export production between CanESM2, CanESM5, and CanESM5-CanOE. Since only the latter has a prognostic iron cycle, it seems remiss not to discuss this. Since CanESM5-CanOE uses both an in-house parameterization of iron scavenging and a CanESM-derived aeolian Fe deposition field, it is possible that biases in one or both of these could explain the low EP; we think it is of interest to readers that this does not appear to be the case.

823-824 – The statement "The development of CanOE was undertaken in response to some of the most severe limitations of CanESM2, and in light of our collective experience." Should indicate that these "limitations" were specifically with respect to biogeochemical and ecological comprehensiveness in other CMIP5 models such that CanOE would have "state-of-the-art" BGC for CMIP6 suitable for multi-hypothesis testing of interactions between elemental cycles in the climate change context.

Added "Many of the additional features that CanOE introduces were already in the models published by other centres even in CMIP5." ("and in light of our collective experience" was deleted because with the new text it disrupts the flow.)

877 – need comma after "achieved"

done

880 – "As to whether the gains in skill with CanESM5-CanOE justify the extra computational cost" – This simple expression of trade-off between tracer cost versus fidelity really undersells the achievement in CanOE as it ignores the difficulty of adding comprehensiveness and degrees of freedom without reducing fidelity (complex models are typically more difficult to optimize than simple models), the gain in robustness when mechanisms are more fully resolved (whether the model performance should be believed as "getting the right answer for the right reason"), and the gains in representing changing elemental interactions allowing the testing of new hypotheses. These are all achievements that should also be highlighted as justifying the extra computational cost.

The first two points are actually addressed in the latter half of this paragraph: less parameterized models usually perform better in novel environments in part because they are more mechanistically based. We added a bit of additional text to emphasize this point more. We also appended to the end of the paragraph: "Inclusion of a prognostic iron cycle and C/N/Fe stoichiometry also open up additional applications and scientific investigations that are not possible with CMOC."

899 – The phrase "and it is likely that the simplification of having a single particle sinking speed is not well suited to a domain with complex topography and prominent continental shelf and slope." Is very specific to a "domain" that has not been specified. As such the criticism, unless it is part of an established literature that is not currently cited, comes across as a non-sequitur pronouncing sentence before the crime has occurred and should be removed.

Good point. This is the final paragraph and we were trying to communicate what we envision ourselves doing going forward and how it guides our choices about model structure. We changed it to: "We are also developing CanOE for regional downscaling applications (Hayashida, 2018; Holdsworth et al., 2021). The regional domains have complex topography and prominent continental shelf and slope, and the single remineralization length scale in CMOC may not be well suited to such an environment."