This manuscript describes the incremental development of the marine biogeochemical component of the Canadian Earth System Model(s) v.5 and the contribution of these models to the 6th phase of the climate model intercomparison project (CMIP6).

My overall judgment is that this is a good technical publication, whose main aim is to describe the features of the new marine biochemistry component (CanOE) and provide useful insights on the historical simulations performed with both versions of CanESM5.

In commending the authors for their achievement, I would however point out that the discussion of the results obtained with CanESM5 and CanESM5-CanOE is not as detailed as one would expect and it should revised to better exploit the material presented in the results section (see detailed comment below).

I would also recommend the authors to revise the ending section of the manuscript with clearer perspective on future research directions and foreseen model development.

Specific comments

Section 3: In the results section the authors widely discuss about differences and biases in the comparison of the model outcomes with observations and CMIP6 multi model ensemble. So, it would be more effective to swap the Figures 2, 3, 6, 7 with the corresponding ones of the Supplementary Material (which directly show the differences against observations).

L46: Is the NEMO model implemented on a T63 horizontal grid? I think this statement is not correct and should be modified to correctly address the ocean model configuration (likely ORCA1 grid). I suggest to add more details on the configuration and resolution of the different CanESM5 components in Section 2.

L123: This aspect could be improved by adopting the SolveSAPHE solver (Munhoven, 2021 https://doi.org/10.5194/gmd-14-4225-2021) in the future development of the model.

L127: I think it should be stated in here that carbon chemistry variables are computed
offline instead of discovering it at L360

L360: GLODAP is a versioned dataset. It would be clearer to refer to it as GLODAPv2.

L360-363: How different is the carbon chemistry obtained with the online computation? If applicable, I suggest authors to detail this aspect to support the offline approach.

L382: A more substantiated explanation should be provided to explain the use of such a coarse horizontal sampling (2°x2°) of CanESM2 and CMIP6 datasets.

L401: The CMIP6 multimodel ensemble data are treated here (and in the following paragraphs) as a “single model” results, but I think that authors are missing the opportunity to exploit this information to better characterize CanESM5 performance in the broad CMIP context

L410: A description of the Oxygen Minimum Zones spatial patterns would a good complement to this paragraph.

Figure 5: Axis labels should be increased in size to make them easily readable

L460-464: The observation-based Aragonite saturation state is here recomputed using GLODAPv2 and WOA2013 data instead of using the original field made available within the GLODAPv2 dataset. The rationale for this choice should be specifically addressed.

L484-485: I don’t think that the conclusion made by Lambert and Boer (2001) in the analysis of atmospheric fields (air temperature, precipitation, sea level pressure) from CMIP1 exercise can be extended in such a way to the DIC, and more generally, to any ocean biogeochemistry.

L515: Figure 11b could be moved to Supplementary material.

L526: The comparison of dFe observations with different models outcome in Figure S4d could be improved by reporting also the tendency lines of each model along with the ideal fit (1:1) black line. Besides, these results could be further discussed in the light of the findings from Seferian et al. (2020, https://doi.org/10.1007/s40641-020-00160-0)

Figure 14: revise caption by adding “CanESM5 is not included because it does not have prognostic iron”

L568: here it would be interesting to specify which are the two models that do not fall along the spectrum.

L574 and L577: Use 260°E instead of 100°W, coherently with the Longitude units used in Fig. 15.

L577: It would be useful to have this “not shown” figure in the Supplementary material

Section 3.4: Differently from the previous ones, this part is largely intertwined with comments on results that better fits the discussion section.

L587: Add reference to Tesdal et al. (2016)

Figure 15: Authors should consider to add a shaded area showing, e.g., the min-max range obtained from the CMIP6 model ensemble and include some considerations with respect to CanESM5.
Figure 16: Axis labels should be increased in size to make them easily readable.

L601-603: The expected behavior of phytoplankton size distribution is clearly visible only in subarctic regions, while in the North Atlantic the monthly variability is rather similar between small and large classes.

Figure 17: The supplementary Table S4 could be easily replaced with a map illustrating in a more straightforward way the location and extent of selected marine regions.

L655: I guess it should read as “... with the range of other CMIP6 models.”

L671-673: It could be useful to address in a dedicated table the residual drifts of the piControl simulation for the CO₂ uptake and also the other biogeochemical variables presented in the previous sections.

Section 4: There are several parts of the discussion section that are not fitting the real purpose. For example, L704-716 describes the evolution of the model which in my option should pertain to the introduction. L729-741 focuses on the differences between CanESM2 and CanESM5 formulations which was already stated previously in the manuscript. Lastly, in many points the authors refer to not shown figures that is not helpful to the discussion. I suggest to revise the entire section by tackling the various outcomes of the result sections, which is very rich in content and material.

L771: Typo in the model name “CanESM5”

L774-776: This sentence is not clear.

L792-794: The impact of the different ocean circulation between CanESM2 and CanESM5 is supported only by the comparison of DIC (Fig.8) and the one “not shown” at Line 750. I think that this part should be better supported (maybe with some additional analysis on other variables) to robustly claim that only ocean circulation is responsible for the observed differences between the two model versions.