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

Reviewer 2

The study presents an updated ocean-sea ice BGC model with updates to the sea ice component as well as to both the biogeochemical and physical ocean model. The model runs include spinup and pre-industrial and historical-present atmospheric forcing, respectively. The model output is compared to observations and reanalysis products, first for ocean physics (AMOC, SST, SSS) and then for ocean biogeochemistry. The authors do a good job of introducing the model and describing updates. The comparisons to other data products covers a wide range, but I would like to see a bit more on the side of quantitative rather than qualitative comparisons. With that and the corrections or responses to the comments below, I would be happy to recommend for publication.

We thank the reviewer for the overall positive assessment of our manuscript and answer the detailed points of criticism below, point by point.

General Comments:

Throughout the paper, there are comparisons between two datasets without any real quantification. They can be found at lines 254, 289, 449, 453. One exception is line 435. Please quantify the other comparisons similarly or in an appropriate way.

We agree with the reviewer that we did not provide statistical metrics for all modeled fields. This is now fixed and we systematically provide at least correlation coefficients and root mean squared errors for all variables.

There are also comparisons between model output for 2012-2021 with an observed/reanalysis product that doesn't have a time period: PHC (line 254), WOA DIN and DSi (line 286), GLODAP (line 435), OC-CCI (338, actually noted in Fig caption, but should be included in text),

Reviewer #1 asked the same question about the time span of climatologies and the reviewer is referred to the answer of its specific comment #1.

Figure 3, A map of just SST and SSS (in addition to the difference plots) would be helpful, either for Clim or run A.

We agree with the reviewer. The figure is updated.

262-263: I don't think runs A minus B should be compared to A minus climatology. The former can be considered de-drifted output or anthropogenic signal, while the latter is how well your model matches to observations (or reanalysis product).

We took out the A-B from the figure and rephrased the corresponding sentence in the text.

310-330: discussion of simulated iron. In this discussion, I don't see any reference to the Aeolian iron deposition (other than intro at line 221-222) as an external (atmospheric) forcing. I see that the modelled value is about twice that of the Huang product. Could you please discuss the forcing here briefly?

The only reference of dust input is given in L221-222, as the referee also mentioned. For all our experiments the same field of monthly dust input was used. Albani et al. (2014)

provides an improved representation of dust using the Community Atmosphere Model and their products for present-day and the Last Glacial Maximum are widely used as aeolian source of iron in global biogeochemical models (e.g. Kurahashi-Nakamura et al. 2022, Du et al. 2022) or as reference to evaluate presentation of dust fluxes by the CMIP6 (Coupled Model Intercomparison Project) models (Zhao et al. 2022). Recent products of dust deposition from other atmosphere models are available as well (e.g. Myriokefalitakis et al., 2018) and Albani et al. (2014) with a total iron input of 16 Tg Fe more or less represents the average of the total range of those models between 10-30 Tg Fe. Therefore we do agree that there are uncertainties in applied dust input data and added some sentences in the revised version. However, this might play a minor role compared with other reasons mentioned in the submitted version.

Figure 10, and text 349-365. Fig 9 seems sufficient for your comparison. The high coastal values in OC-CCI of Fig 9, especiallly in the coastal argument, already supports the discussion of the issue with ocean color products and turbid but abiotic signal in coastal waters. I suggest removing Fig 10. I would still keep most of the turbid water discussion, but link it to Fig 9. I also suggest removing the lines 357 (from “Therefore”) to 362 up to “(Lee and Marr, 2022).”

We totally agree with the reviewer that the turbid coastal waters represent a challenge for the assessment of both remotely sensed chlorophyll and primary production. Therefore, following reviewer’s recommendations, we decided to merge the results and discussions together to avoid repetition, shorten significantly the text and improve readability. However, we preferred to keep Figure 10 showing NPP fields as we believe this is an essential variable.

Fig. 11, Sometimes the PP aren’t very limitted at all. Instead of just which is the most limitting factor, maybe use white for regions where all limitting factors are above 0.5 maybe? Also, should MLD be considered in concert with light limitation (i.e., a very deep MLD, despite high light at the surface, could limit PP because the plankton could sink deeper in the ML).

The reviewer is right that this analysis should be seen as a “likely limitation”, most often at the beginning or the end of the productive season. But a limitation whatsoever must at some point intervene so that the PP remains finite as in the “real world”, either by light or nutrients. In this sense, there is always a factor limiting the PP. In response to the reviewer’s comment, we have added shading to the nutrient limitation figure to illustrate where limitation is weak (>0.05).

About the MLD, the classic Sverdrup Theory refers to the MLD as a light limiting factor. Indeed, if plankton is transported too long outside the euphotic layer, then, it will not be exposed sufficiently to sunlight to maintain growth over loss terms. In other words, the MLD effect can be seen as a light limiting effect.

Fig 14. Note that GLODAP is biased towards later in the sampling period when more measurements have been taken. For DIC in particular, that means GLODAP may have a higher value (DIC in surface ocean is increasing with time), but your model output is weighted equally over the time period. This would contribute to model-obs bias, but would not indicate a fault in the model. (This argument does not apply to Alk.)

We thank the reviewer for raising this important point. This also led us to realize that the time-period for DIC comparison was not chosen wisely. As DIC is changing over time, it

needs to be compared to the same period as the GLODAP data (normalized to the year 2002).

454: “largely agree”: please quantify; “although the magnitude differs”: I’m not sure what this means, please clarify

We reformulated the sentence to now read: “The spatial patterns of reconstructed pCO2 and CO2 flux from different pCO2-products largely agree with each other in contrast to the amplitude and timing of variability of regionally or globally integrated fluxes (Fay et al., 2021; Fay and McKinley, 2021).”

499-505: note that the magnitude of drift is reduced, but also the sign. Both have drift that leads annual air-sea CO2 flux to increase in magnitude (FESOM1.4 is positive with a positive drift, and FESOm2.1 is neg with a neg drift), therefore, I don’t think it is self evident from that information that the flux will tend to zero with a longer spin-up.

The reviewer is right. We think that the flux will not tend to be zero before the deep ocean reaches a steady state. We removed “ and could be further reduced towards zero with a longer spin-up” from the sentence.

Specific/technical comments

21: “preindustrial” should be “industrial”, or “end of preindustrial”

Amended as suggested. “industrial” is used instead of preindustrial.

29: suggest changing “of the seasonal cycle” to “in accurately representing the seasonal cycle”

Amended as suggested.

40: reference needed for “large interbasin gradient between the Pacific and Atlantic”

A reference is added.

110: add “off” after parameterization

Amended as suggested.

274: suggest changing “of both AMOC” to “, for both runs, of AMOC”

Amended as suggested.

275: suggest adding to “a nearly constant” with “nearly constant, but with a small increase,” (I assume this is from continued model spinup drift?)

Amended as suggested.

280: replace “pursues” with “is in the same area as” - This similarity is not necessarily surprising to me. In general, MLD differences are largest where MLD is largest and where horizontal gradients are largest (i.e. Labrador Sea and Southern Ocean). I assume the same for SST and SSS (I recommended to include those with Fig 3)

We agree with the reviewer and we do this.

280: “Fig1” should be “Fig 3”

Amended as suggested.

320 “quite a bit smaller” quantify please (e.g., “smaller by about a quarter magnitude”)

There is no uniform scaling factor for the amplitudes of the patterns between the Huang reconstruction and the model-generated field. The largest discrepancy is in the tropical Atlantic, where the model overestimates the increase in dFe under the Saharan dust plume by more than a factor of two. We added some quantification
Here is the main added text: “The largest discrepancy in amplitude is found under the Saharan dust plume in the tropical Atlantic, where the model produces maximal dissolved iron values that are almost three times as high as the reconstruction from \citet{Huang2022}. Direct observations in the tropical Atlantic also show dissolved iron concentrations that reach 1.2 nmol L-1 (e.g. Hatta et al., 2015), while modeled maxima are > 3 nmol L-1.”

375 ”annal” should be “annual”

Amended as suggested.

440: “(too low” please add “(which was too low in”

Amended as suggested.

Fig 16, please add that positive indicates into the ocean in the caption

It now reads: ”Negative numbers indicate a flux into the ocean.”

Fig. 17, use pCO2 instead of fCO2 to be consistent with rest of paper

Good point, done.

488: suggest replacing “misfits” with “biases” (two times)

Amended as suggested.

490 suggest changing “northern high latitudes” with “northern high latitudes negative bias”

Amended as suggested.

491: suggest “Pacific” change to “Pacific positive bias”

Amended as suggested.

Fig. 19, I can’t really tell the difference between the lines for pCO20products and for Models. Use a more distinct pair of line-colors please

We use colors that have better contrast.

508 (from 512-513), please move “with a constant atm...forcing (simulation B)” up to line 508, just after “control simulations”

Good idea, done.

524: change “is with 27.7 PgC” to “is 27.7 PgC, ”

Amended as suggested.

525: I suggest removing “best”

Amended as suggested.

533: remove one “also”

Amended as suggested.

535: suggest change “too weak” to “too weak of uptake or too strong release”

Amended as suggested.

545: Replace “Compared ... compared” with “Compared to a model intercomparison study of”, and also, what was the value of Cocco et al?

We agree with the reviewer and we do this.

553: Change “sensitivities of” to “sensitivities to”

Amended as suggested.