Dear editor and reviewers,

we would like to thank again the two reviewers for their positive evaluation and their valuable comments that will further improve the quality of the paper. We have now supplied the revised version of the manuscript. In the track-changes version we highlight the biggest changes in the blue font. In our response below, we always highlight the reviewer comment in blue, with our response in the standard font and the quoted text from the revised paper in the green font.

Please find the specific answers below:

Reviewer 1

We would like to thank the reviewer for their very positive evaluation. Please see our responses below:

I have a few minor suggestions that I think could help strengthen the arguments further and/or provide clarification. These should not be considered conditional for publication, take or leave as you all see fit.

1. While the focus of this synthesis is understandably marine, a brief acknowledgment of the contribution of coastal and estuarine systems to global and annual NPP could provide useful context, especially in light of choosing to include Figure 3. These regions are ecologically and socioeconomically important, disproportionately impacting fisheries and food security under climate change. Resolution issues are mentioned briefly, as well as sparse coverage, but perhaps a sentence or two would add helpful perspective, noting that different environmental forcing, model parameterization, input uncertainties, validation biases, etc. are all compounded in these complex, dynamic systems.

To address this, we have included new text on the lines 182-184:

"Furthermore, global models, with their coarse horizontal resolution, struggle to capture coastal and estuarine processes that enhance PP (coastal regions account for 14-33% of global PP; Gattuso *et al.*, 1998), which makes them also prone to underestimate global PP."

We have also better emphasized the different sources of model uncertainty mentioned by the reviewer, by slightly expanding the paragraph on the lines 669-680. Please see the lines 676-680:

"For example, differences in how models treat external forcing, such as micro- and macronutrient depositions from the atmosphere, could still contribute to the growth of uncertainties as models become more complex. Furthermore, the spread in the underlying environmental changes such as warming, stratification, changes in irradiation, and ocean circulation among others, contributes significantly to uncertainties in the PP projection."

2. Section 3.1 provides a robust technical comparison of model structures. To further engage readers beyond the specialist community, it may be useful to briefly explain why models differ in specific regimes a bit more explicitly. For instance, in oligotrophic gyres, the Eppley function captures potential higher photosynthetic efficiency at elevated temperatures, whereas the standard VGPM adjusts efficiency based on nutrient limitation in high-SST waters. Other mechanistic considerations, like estimating NPP below the mixed layer or incorporating fluorescence, might also be briefly mentioned to help readers interpret the practical

consequences of different parameterizations, particularly among satellite-based models. Because these models often rely on a common, highly correlated set of input variables, differences in assumptions and internal formulations tend to play an outsized role in shaping their divergent behavior, especially across distinct oceanographic regimes.

We have taken a rather cautious approach to replying to this thoughtful comment by the reviewer. While we agree that multiplying an Eppley-type temperature response function with a nutrient response function could yield a function that had similarities with the VGPM type of empirical temperature-dependence function, we would prefer not to speculate on the underlying causes of PBopt-temperature relationship, in the absence of corroborating evidence on nutrient conditions in the work of Behrenfeld and Falkowski (1997). Furthermore, there might be other reasons for the empirical relationship: e.g., Blackford et al. (2004) invoke enzyme inhibition for the decrease in relative growth rates at high temperatures. We have therefore only added the following statement on the lines 293-296: "Thus, models (the combined functions F) differ depending on (i) how many environmental factors are included in the model, (ii) the explicit functional forms selected for each modulating function; and (iii) the parameter values A_i, D_i assigned to those modulating functions, and whether they are allowed to vary with region and time."

3. The brief mention of AI could be expanded slightly to address common skepticism. Beyond predictive applications, AI can be extremely useful in the creation of "numerical laboratories" for assessing (or retuning) parameter sensitivity under varying environmental forcings, enabling process-based insights rather than treating models as black boxes. This framing aligns well with the broader comparative focus of the manuscript.

Thank you, we have now expanded the text on AI, mentioning the use of emulators (numerical laboratories) and several other topics, such as explainable AI. Please see the new text on the lines 711-716:

"AI can be used in a variety of different ways, either as a direct prediction approach to optimise model parameterisation and also to emulate models, allowing it to explore a range of model behaviours at reduced computational cost for parameter sensitivity analyses and model calibration (e.g., Mattern *et al.,* 2012; Schartau *et al.,* 2017). Furthermore, recent statistical approaches unique to the ML field enable insights into what the ML model has learned, for example, using Explainable AI, or physically constrained machine learning."

4. In the conclusion, it would be helpful to clarify (even philosophically) if "success" is defined by model convergence (all models better agreeing on a NPP number), by increased mechanistic realism, by reduced uncertainty in common input parameters, or something else altogether. If different models converge on the same result for the wrong reasons, we haven't actually improved our predictive foundation. How do we deem something "fit-for-purpose" in the end? This paper is a good opportunity to assert some of those absolute metrics of improvement with confidence.

Thank you, we have now included several statements on the model convergence and inter-model consistency comparisons, but also explicitly warning that any increase in model consistency must be supported by an independent validation (so the models don't converge for the wrong reason). Please see as an example of this the lines 657-668, or 247-253. We think that when it comes to complex problems such as modelling PP, the model evaluation depends on the precise questions asked and the exact context in which the metrics are applied. This means, if possible, we would like to avoid singling out a specific metric defining the success and making overly generalizing comments in this paper.

Reviewer 2:

We would like to thank the reviewer for their overall positive evaluation. Please see our responses below.

In their major comments the reviewer raises two main points:

Main point 1: The uncertainty in the primary production measurement methods, leading to high uncertainties in (in situ) observations of primary production, e.g.

"First and foremost, in my opinion, the single largest challenge in improving estimates of primary production and its changes is the difficulty of reliably measuring primary production. Any attempt to use data assimilation or machine learning tools to understand productivity patterns requires reliable observations, or at least observations with a reliable error model to extract those insights from."

and

"Different NPP data sets, and even the same data set from different times, support different models. It is difficult to know which data to trust, or whether all models were simply wrong in substantial ways in different places and times."..

and

"The authors come back to the issue of NPP measurement uncertainty just before the conclusions (Fig. 10), but I think it is important to address it earlier, present it in a more cohesive way, and propose a more specific productive path forward to support the parameter-estimation and variation aspirations described elsewhere in the text. Is the development of a comprehensive NPP error model across data types and times, for example, possible? Would this enable the data-driven extraction of the insights that you are aiming for?"

Thank you for your comment. We fully recognize the challenges associated with measuring PP *in situ*, and we are happy to highlight these issues early in the manuscript. We acknowledge that the potentially large uncertainties in PP field measurements can importantly influence the calibration of the models discussed in our paper. At the same time, model calibration can also rely on observations of many other variables beyond PP, which may help mitigate this concern (observability of PP-relevant parameters from observations of chlorophyll, nutrients, oxygen etc. was analysed e.g. in Ciavatta et al. 2025, referenced in our manuscript). We indeed agree that having a reliable PP error model could enhance the ability to learn from observational data, and this is now explicitly mentioned. While we agree that the difficulties of PP measurement should be clearly stated upfront, we would still like to keep the primary focus of the paper on the modelling aspects of PP.

To address these points, we have now included a paragraph early in the Introduction section on the lines 98-104:

"When it comes to *in situ* observations, experimental methodologies and carefully assembled protocols exist for measurement of each of these components (IOCCG, 2022); however, practical constraints may limit the extent to which the components may be differentiated from each other. Furthermore, various observational methods of the same component could yield differing values. For example, multiple methods for measuring the same component could have different intrinsic timescales that are applicable to them, making direct comparisons difficult. Improving observational tools for PP, including

developing reliable PP error models, is a priority for the scientific community, in addition to the modelling issues that are the primary focus of this paper."

Furthermore, there are several other places in the manuscript where we now highlight the implications of the current status quo in *in situ* PP observations and emphasize the need for continued improvement in PP measurement methodologies. These include:

(i) Stating right in the Abstract on the lines 39-41: "Model-observation comparisons are also complex, because paucity of data, differences in measurement techniques, and evolving methodologies could all lead to difficulties with the interpretation of results."

(ii) In the section 3.1 on satellite data, see lines 339-340: "A contributing factor to this outcome could be the uncertainty in field measurements of PP and also to issues related to mismatches between the temporal and spatial resolutions of models and observations."

(iii) See also the lines 673-674.

The reviewer's main point 2:

"My second observation relates to the potential synergies between satellite NPP models and ecosystem models that estimate NPP. The paper outlines many differences and commonalities between these two types of models. I think it would benefit, however, from a more concise distillation of the basic differences and pathways for productive interaction that leverage their relative strengths."

and

"I see a tangible path forward here that could start with ecosystem models simpler than Darwin. Even a biogeochemical model with a basic representation of allometric tradeoffs for nutrient uptake, grazing and light harvesting, for example, offers a potential step forward relative to most current size-agnostic satellite-based algorithms. There are published examples of approaches to relate the more complex patterns emerging from ecosystems models to the emergent growth functions employed by satellite-based NPP models (Stock et al., 2019). The utility of the ecosystem- and satellite-model synergy, of course, would be contingent on the capacity of the ecosystem model to skillfully capture observed chlorophyll, Chl:C, nutrient, zooplankton and other patterns, but fortunately skill is improving on this front (thanks in part to satellite-based products!).

In short, this specific approach seems to provide a concrete strategy for how ecosystem- and satellite-based models could be integrated in a way that helps overcome NPP data challenges. Systematic addition and removal of physiological and ecological tradeoffs, for example, could provide "fingerprints" that help explain and justify data-driven findings. I would consider giving this potential interaction a more central place and reducing some background/descriptive elements that are less actionable. A more specific call-to-action would make the paper more impactful."

We agree with the reviewer that this point can be made more central. However, we also emphasize the important caveat that inter-model consistency must always be supported by validation against independent data, to avoid the risk of models reinforcing shared errors. We have thus addressed this reviewer comment by:

(i) Making a statement right in the Abstract on the lines 53-56: "An important aspect of this unification could be the ability to infer the spatio-temporal variability of parameters in the less complex models from the emergent behaviour of the more complex ones. This could

include ecosystem model simulations of nutrients, temperature, phytoplankton classes, or vertical distributions informing satellite-based models."

(ii) In the Introduction on the lines 247-253: "Furthermore, these highly complex models could provide valuable information for estimating the spatio-temporal variability of parameters used in less complex ecosystem and satellite-based models. This would also help increase consistency among different models, making them more comparable. At the same time, caution must be applied to ensure that increased consistency and convergence is not confused with increase accuracy. For this, we would need to continue independent assessments of accuracy, for example by comparisons with *in situ* observations, with full recognition of the caveats that such comparisons entail, as discussed above."

(iii) In section 5, lines 657-668: "For example, high-complexity ecosystem models (such as the DARWIN ecosystem model) could be used in some cases to deduce the degree of parameter variability of simpler ecosystem models or help inform spatio-temporally varying parameter calibration of those models (always keeping in mind that inter-model consistency does not automatically imply model quality). Comparison studies across models of different complexity would be desirable in this case (for some examples see Friedrichs *et al*., 2007; Xiao and Friedrichs, 2014). Similarly, emergent properties of ecosystem models can be leveraged to provide specific information for satellite-based models, such as vertical and class-distribution of phytoplankton (e.g., Stock, 2019), or information about nutrient distributions. Such inter-calibrations of models against each other could potentially improve satellite PP products and conversely make the satellite PP data more useful for ecosystem model development. However, one has to be cautious here: model-model intercomparisons and tuning would help models look more like each other, but independent information would be needed to ensure that the simulations are also getting closer to key features in the real world that the models are designed to reproduce."

(iv) And Conclusions, lines 752-757: "It would also create opportunities for improved intercomparison across models of different complexity, including the ability to understand more about unresolved variability in simpler models by comparing them with the higher-complexity models. One could argue that the spatio-temporally varying parametrisation could help reduce the existing high uncertainty both in historical estimates and future projections of marine PP, provided that independent information is used to avoid all models converging towards a systematically biased outcome."

The reviewer had also several minor comments that are addressed below:

1. Line 39-40; 45-47: Consider this statement relative to my first primary comment. It is important to acknowledge that the in-situ data is also highly uncertain.

Thank you, this has been addressed, please see the lines 39-41 (also see the response to reviewer major point 1).

2. Line 48-52: The concrete and achievable synergy between ecosystem and satellite-based models (lines 626-631) highlighted in my second primary comment seems central to these points. I would consider a clearer and more prominent articulation of this synergistic path forward.

Thank you, we have addressed this, please see the Abstract on the lines 53-56 and also our response to reviewer major point 2.

3. Line 71-73: Is the purpose here only to review, or do you intend to suggest a path forward?

Thank you, this has been now rephrased.

4. Line 90: Perhaps reframing this paragraph to address my first general comment is: “A central challenge in understanding primary production and predicting primary production is that it is difficult to measure”, then consolidating these issues here? You would then just need to articulate how these challenges could be overcome.

Thank you, we have now added the whole new paragraph in the Introduction section, please see the lines 98-104 and the response to reviewer major point 1.

5. Line 97-98: Most ecosystem models predict and report NPP, not GPP. Respiration is often included in the growth function (e.g., Geider), basal respiration is either omitted or accounted for in reported NPP.

Thank you, we rephrased this by adding “in many ecosystem models”.

6. Line 132-133: I would argue that not including T-dependent parameters for a global algorithm, or even a large regional one across seasons, is not defensible given the data. Can’t we at least eliminate this?

We agree with the reviewer that such approaches are problematic, but we would like to try to refrain from commenting on specific approaches that are being compared. If possible, we prefer to remain sufficiently general in both, reviewing the state-of-the-art PP models, and suggesting the way forward.

7. Line 156: I would call the 17-83 range the range, not the uncertainty. The lower and upper bounds can clearly be rejected with existing constraints. There seems to be an unwillingness to do so directly, but maybe the effort you are calling for could help us do so in an objective and constructive fashion.

Thank you, we addressed the first comment by rephrasing it to (see line 159) “with values reported in the 17-83 Pg C y^{-1} range”. As we mentioned previously, whilst we agree with the reviewer that the estimates are unlikely accurate, we would avoid outright rejecting some of these estimates in this paper.

8. Line 207-213: See general comment 1. The difficulty of observing NPP deserves a more central, consolidated place in your narrative and needs to be more directly addressed in your call to action.

Thank you, we made this now more central, please see our response to reviewer’s major point 1.

9. Line 232-234; 238-241: Is there enough data to do this, or would an ecosystem-/satellite-based model synergy be the primary driver of this process? Would intermediate-/high-

complexity ecosystem models be an intended target of such activity, or would it be a tool helping to understand and introduce this added structure into other models?

We believe that both have value: calibration against the same observational data sets and, wherever possible, inter-calibration among models (by keeping in mind that the latter is no substitute for the former). Please note that when we refer to the availability of large observational data sets for model calibration, we mean data covering a broad range of biogeochemical variables (chlorophyll-a, oxygen, pH, nutrients, etc.). These variables are generally better constrained than the PP measurements mentioned by the reviewer. Therefore, stating that sufficient data may exist for a highly constrained model calibration is, in our view, fully consistent with the acknowledged need to improve the accuracy of in situ PP measurements.

10. Line 322-327: See general comment 1.

Thank you, we have addressed these points, please see our response to reviewer's major point 1.

11. Line 347-349: Not always ignored; Sometimes treated implicitly.

Thank you, we said here "often ignored" which we hope was accurate enough?

12. Line 404-405: Do you think it is defensible to not have a temperature dependence?

Please see our response to the reviewer comment 6.

13. Line 434-440: This seems key to the synergy between these two model types, and comes up again on lines 626-631. It would be nice to have a more cohesive presentation of these points (see general comments). Emergent Patterns: The satellite-based models that do not have explicit nutrient and temperature dependencies, inherently contain those dependencies in the model parameter values (line 434-435); BGC models have these dependences and allow you to probe emergent outcomes.

Thank you, please see our response to reviewer's major point 2.

14. Line 460: See general comments.

Thank you, please see our response to reviewer's major point 2.

15. Fig. 6A: See Stock (2019) for an example of how to derive the equivalent photosynthesis-irradiance parameter from an ecosystem model.

Thank you, the reference was now included.

16. Line 535-537: This flexibility allows models to account for the diversity of plankton and processes that are not explicitly represented in current models (BGC models capture quite a

bit, but I could see a datadriven approach capture more. The only problem is that we don't have reliable measures of NPP to find emergent patterns).

Whenever we discuss calibration of parameters, we consider also other measurements than PP, please see our response to the reviewer's comment 9 and also our response to their major point 1. But we agree that the uncertainty of PP measurements needs to be better discussed in the paper and we did this as listed in our response to reviewer's major point 1.

17. Line 585: "Together, these considerations suggest that investigating parameter assignment and parameter variability may be an important route to understand and potentially reduce many of the apparent differences between marine PP models, and hence in the estimated magnitudes of production." I feel like the recommendations that follow could be more concrete, and worry a bit that this statement comes nearly 600 lines into the paper.

These points are discussed prominently, e.g. in Abstract (lines 53-56), and Background section paragraph on the lines 233-253. We hope that the discussion in those sections is concrete enough to be actionable, whilst being still general enough to be of wide interest.

18. Line 628-631: This seems like a concrete path forward yet the treatment is very limited (see general comments).

Thank you, this was addressed as described in our response to reviewer major point 2.

19. Line 634-636: An considerable uncertainty in in-situ NPP observations?

Thank you, these were addressed (please see our response to reviewer's major point 1).

20. Line 655: New paragraph?

Yes, thank you, this was addressed.

21. Line 656-675, Fig. 10: Yes, data volume has increased rapidly, but has it yielded reliable, consistent NPP estimates across global ocean biomes? If this hasn't happened yet, is the time really right? Is there a way to address the fact that it hasn't in a rigorous way using new techniques to provide a self-consistent data source for much of the analysis you've proposed. I feel like this section should have come earlier in the text as one of the central challenges to executing what you are proposing, rather than just before the conclusion.

The uncertainty of in situ observations is discussed now much more thoroughly (see our response to reviewer's major point 1). Regarding the data-sets and the calibration techniques (including the data assimilation and machine learning) we would like to point out also our response to reviewer question 9: the PP-relevant parameters can be estimated from non-PP observations as well (such as chlorophyll-a, oxygen, nutrients...), as long as the parameters are observable from those data. We would also like to emphasize that we are referring to a highly constrained estimation (e.g. seasonal climatology across Longhurst provinces), so we are being very cautious in our statements - please see the discussion in the two paragraphs on the lines 620-629 and 630-642.

22. Line 690: I'm not sure what you mean by "assignment is quite rare".

We mean assignment of spatio-temporally variable parameters. We have made this now more explicit, please see the lines 739-740: “where assignment of variable parameters is still quite rare”

23. Line 693-695: Would the natural next sentence be: A combination of data-driven approaches and natural synergies with intermediate- and high-complexity ecosystem models could provide the means to fruitfully uncover these variations.”

Thank you, we have now tried to emphasize better the value of model synergy (please see our answer to reviewer major point 2).

Thanks again to the reviewers for their valuable input to the paper, we hope that after these revisions the manuscript is now ready to be accepted for publication in Ocean Science.

Best wishes,

Jozef Skakala and the co-authors