We would like to thank the reviewer for their very positive evaluation and their valuable comments that could further improve the quality of the paper.

Please find below each reviewer comment in blue and italic and our response in black, standard font.

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.*

Thank you, we are happy to include an extra comment on the relevance of the coastal and estuarine systems at the bottom of the page 5 (around line 180), where we already briefly mentioned the impact of coastal areas on global PP. Model parameterization as a source of (whether regional, or global) model uncertainty is the central topic of the paper, but we will add an extra comment highlighting also the other sources of model uncertainty (e.g. input forcing and underlying physical environmental changes).

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.*

Thank you, one issue is that this is slightly difficult to do on the level of the temperature/-nutrient/light modulating functions, as the final PP model depends on the non-trivial combination of all the modulating functions (and also on whether parameters are treated as constants, or are spatio-temporally variable). In this regard we would like to add some text to the paragraph on the lines 277-282, explaining how the differences in modulating functions relate to the differences in the final PP model. As part of this, we suggest explicitly mentioning the VGPM as an example of a model that accounts for nutrient limitation in oligotrophic gyres by using a single temperature-modulating function and constant parameters. In this context, we also suggest discussing the differences between the Eppley function and the VGPM, emphasizing that these differences depend on the broader PP model structure. For instance, the Eppley function, when multiplied by an appropriate nutrient-limitation term, may yield PP behavior in oligotrophic gyres similar to that produced by the VGPM.

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 re-tuning) 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.*

From our understanding, the reviewer suggests here to discuss AI model emulators, which we agree is appropriate. We suggest to add some short text on this in the "A way forward" section, where AI is mentioned (around the lines 665-667). We also suggest to mention explainable AI approaches.

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, yes, we suggest to add a short comment on this in the Conclusions. We'd suggest that considering some of the observational uncertainty, model convergence is indeed one of the several possible indicators of model reliability, but we acknowledge that as every other metric it has potential issues (as highlighted by the reviewer). As often, it might be optimal to use several metrics in the same time, to give a more complex insight into model performance.