Dear Editor and Reviewers,

We sincerely appreciate your time and thoughtful feedback on our manuscript, "Refining marine net primary production estimates: Advanced uncertainty quantification through probability prediction models" (MS No.: egusphere-2024-3221). Your critiques have significantly strengthened our work, and we are pleased to submit this revised version, which addresses all remaining concerns. Below, we provide point-by-point responses to your latest comments (in plain text), with revisions highlighted in blue.

This revision represents a collaborative effort by all co-authors, and we believe the manuscript now offers enhanced methodological clarity and scientific rigor. We are grateful for your guidance throughout this process.

Best regards,

Mengyu Xie (on behalf of all co-authors)

The revised version of the manuscript is a much better read, and the authors have spent considerable effort in addressing my comments. However, I still have some reservations about the methodology and framing in the revised version.

general comments

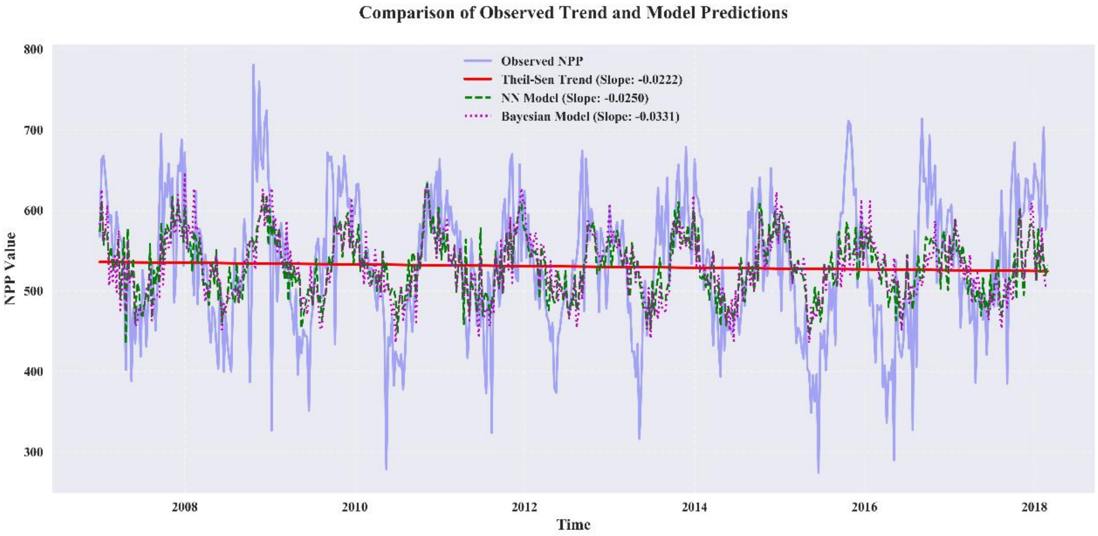
The authors have incorporated feedback from my previous comments in the manuscript, and importantly, they acknowledge that the uncertainty estimates do not reflect the full model uncertainty. However, the first such acknowledgment appears late in the manuscript, in line 476 in the results and discussion section. Later, the authors still claim that "Our objective extends beyond merely reproducing satellite NPP products. We aim to improve the overall accuracy and uncertainty quantification of NPP estimates by incorporating a robust probabilistic framework." (l. 697). But the uncertainty is not fully qualified, in particular, this approach does not capture structural uncertainty, i.e. model bias or inadequacy. The estimates of CAFE may be heavily biased, but we do not know, and the uncertainty analysis conducted here would not show it. A more careful language is needed.

We appreciate the reviewer’s insightful comments regarding uncertainty quantification. In response, we have (1) revised the abstract (lines 43 - 46) to explicitly state the structural uncertainty limitations, ensuring early transparency; (2)

refined the uncertainty quantification phrasing (lines 704-708) to more cautiously articulate the scope of our analysis and acknowledge unresolved challenges. These edits adopt measured language throughout, balancing our contributions with a clear discussion of limitations. We are grateful for the reviewer's emphasis on rigor in communicating uncertainty, which has strengthened the manuscript's scientific integrity.

The authors claim that "The results reveal that both models are competent in quantifying CAFE uncertainty." (l. 726). Beyond the problem mentioned in my comment above, it remains unclear if the two methods actually capture main parts of the CAFE signal. Based on Fig. 7 and 10, the NN and Bayes model can capture the seasonal dynamics of the CAFE output. But is there a trend in the CAFE data, and do the two models capture that trend?

We appreciate the reviewer's insightful comment regarding characterization of the "CAFE signal" and the potential presence of a trend. Our analysis demonstrates that both the neural network (NN) and Bayesian models clearly preserve the seasonal dynamics present in the original Net Primary Productivity (NPP) data (Figures 7 and 10). These cyclical patterns, evident in the low-pass filtered NPP datasets (Figure 3), are retained in the model outputs, indicating that the models preserve key cyclical features of the system. Regarding long-term trends, the Mann-Kendall test shows no statistically significant trend ($p = 0.852$) in the CAFE data, though a weak negative slope exists (−8.11 units/year, Theil-Sen estimator). The corresponding figure illustrating this trend analysis is appended at the end of this response for your reference. Both models correctly reproduce this behavior. The NN slope (−0.0250) closely matches the observed slope (−0.0222), while the Bayesian slope (−0.0331) stays within expected interannual variability ranges (Figures 7 and 10). This consistency confirms neither model artificially alters the data's inherent trends, supporting their reliability for uncertainty quantification. We thank the reviewer for raising this important point and hope this explanation addresses the concerns regarding trend preservation in our analysis.



Furthermore, what evidence is there that the NN and Bayes model perform better than climatology? My concern is that one could build a simple climatological NPP model for Weizhou Island with uncertainty that would produce very similar output to the NN or Bayes model. For example, one could use a + b * sin((c + time)/d) + epsilon where epsilon ~ Normal(0, sigma) is a random variable. After estimating the model parameters (a, b, c, d, sigma) from CAFE data, it would require only time input and produce NPP estimates with uncertainty. Of course, this a very simple model and every year is the same, there is no trend, and the uncertainty does not vary with time. But then the NN and Bayes model seem to produce nearly identical output for each year as well, and the uncertainty envelope in Fig. 7 and 10 are very similar from year to year. Thus, it is important to show that NN and Bayes model perform better than a simple climatology model.

We thank the reviewer for the thoughtful suggestion and agree that comparing our models to a simple climatological baseline is an important benchmark. While a sinusoidal climatology model of the form proposed (e.g., $a + b * \sin((c + t) / d) + \epsilon$) could indeed replicate seasonal dynamics, it does not incorporate external drivers or respond to changes in environmental conditions. In contrast, both the NN and Bayesian models in our study utilize real-time environmental inputs (e.g., temperature, precipitation, radiation), enabling them to adapt to interannual variability and capture ecosystem responses under non-stationary conditions. Additionally, the probabilistic nature of both models allows for dynamic uncertainty quantification, which varies over time based on input conditions—unlike the fixed uncertainty envelope in the proposed climatological model. We acknowledge that during the relatively stable period analyzed (2007–2018), interannual differences in environmental inputs were limited, which resulted in similar model outputs across years. Rather than indicating a lack of model sensitivity, this consistency reflects the stable behavior of the ecosystem under non-extreme conditions. In response to the reviewer’s valuable point, we have expanded the discussion (lines 745–748) to include a more explicit comparison with a climatological baseline and highlight the added value of our models.

An aspect that is important but not described well in the manuscript is the required model input compared to that of VGPM, CbPM, and CAFE. In one statement, the authors write: "These inputs overlap substantially with those used in VGPM, CbPM, and CAFE, demonstrating that the NN and Bayesian models do not require additional or more complex inputs." (l. 315). Later the manuscript states: "These probabilistic models do not require additional input variables beyond those used by VGPM, CbPM, and CAFE." (l. 720) Are really all 11 inputs listed in Table 1 used in VGPM, CbPM, and CAFE? Did the authors perform any experiments limiting the inputs to the NN and Bayes model further to examine which inputs are actually required to produce the output?

We appreciate the reviewer's meticulous examination of the input variable descriptions. We acknowledge that the original manuscript contained imprecise statements regarding input variables. Specifically, not all 11 variables listed in Table 1 were used by VGPM, CbPM, and CAFE; for instance, variables such as height of tide (m) and 1/10th significant wave height (m) are novel to our modeling framework. We have revised the relevant text (lines 320-321, 721) to clarify that our NN and Bayesian models extend beyond the input requirements of VGPM, CbPM, and CAFE, rather than matching them exactly. To assess input necessity, we conducted ablation experiments systematically removing individual input variables. In all cases, performance declined, confirming that each variable contributes meaningfully to model accuracy. Therefore, the full set of 11 variables was retained to ensure robust predictions. We hope this addresses the reviewer's concerns and provides sufficient context for the methodological choices made.

When the data used for training a NN or model is very limited, a common thing to do is bootstrapping, i.e. dividing the data into different training and testing datasets repeatedly. Did the authors try different testing and training data configurations? It may shed more light on the differences in the CDF curves that are discussed in Section 3.2.2.

When dividing the training set and the dataset, different ratios have been tried to explore the model effectiveness in different cases, and the final chosen ratio is 8:2. Not only because the model evaluation metrics are better in this case, but also previous studies have indicated that the 8:2 ratio is a widely adopted practice in the field of machine learning and deep learning, which strikes a balance between providing a sufficiently large training set to efficiently learn features and patterns, and providing a smaller test set to robustly evaluate the generalization ability of the model. It strikes a balance between providing a sufficiently large training set to effectively learn features and patterns, and providing a smaller test set to robustly evaluate the generalization ability of the model.

specific comments

L 54: "Conventional methods of NPP measurement, such as ship-based sampling and bottle incubations, are beset with challenges like human errors and inadequacies in capturing spatial and temporal dynamics. This underscores the necessity for more sophisticated and comprehensive methods (Yang et al., 2021; Li et al., 2020)." True, but this study relies very much on monitoring data from a station and thus does not capture spatial dynamics -- it further relies on continuous measurements to capture the temporal dynamics. The authors mention this later: "Due to factors such as equipment malfunctions and adverse weather conditions, some data for the eleven variables were incomplete." (l. 198).

We agree that spatial variability remains a limitation of our current setup and have revised the manuscript (lines 55-60) to more clearly distinguish between spatial and temporal dynamics, and to acknowledge that our approach primarily addresses the latter. We also clarified the limitations posed by data gaps due to equipment malfunctions and environmental constraints (lines 201-203). We thank the reviewer for pointing out this important distinction.

L 79: "Currently, the most widely utilized models for estimating NPP include the

Vertically Generalized Production Model (VGPM), [...], have been proposed.": This sentence needs to be rephrased.

We have rephrased this sentence as “Currently, the estimation of NPP primarily relies on three mainstream models: the Vertically Generalized Production Model (VGPM), the Carbon-based Productivity Model (CbPM), and the Carbon, Absorption, and Fluorescence Euphotic-resolving Model (CAFE). These models were successively proposed by Behrenfeld et al. (1997), Westberry et al. (2008), and Silsbe et al. (2016), respectively, and have become benchmark methods in this research field.” in line 81-86.

L 156: "The proportion of excellent water quality in Guangxi's near-shore waters reaches more than 90% all year round": It is not clear what this means. What is this measure of water quality, and is this based on a study or survey that could be cited? Similarly, what does "the quality of the marine ecological environment has remained at the forefront of the country" imply? More specific language and references would be useful here.

The description of the study area has been modified to “The island extends in a NE-SW direction and has an elliptical shape. It is approximately 6 km long from north to south, 5 km wide from east to west, and has an area of approximately 25 km2, making it the largest and youngest volcanic island in China (Li and Wang, 2004). Weizhou Island is an inhabited volcanic island, the annual average water surface temperature is about 24°C, and ranges from 19°C to 30°C. The annual average seawater salinity is 32‰, seawater pH ranges from 8.0 to 8.23, and seawater transparency ranges from 3 m to 10 m (Yu et al., 2019). In addition, Weizhou Island is the northernmost island in the Gulf of Tonkin, where coral reefs have developed. These coral reefs are mainly found in shallow waters along the southwest, northwest, and northeast coasts, with widths ranging from 0.86 to 2.56 km (He and Huang, 2019). ” in line 159-169.

L 163: "Weizhou Island, located in the southern subtropical monsoon zone, experiences a pleasant climate with abundant heat and precipitation throughout the year." Phrases like "pleasant climate" or "abundant heat and precipitation" are not specific or quantitative. The next sentence already specifies average (air?) temperatures, so the "pleasant climate" is not necessary here.

We have removed the adjectives like “pleasant” and “abundant” in the revised manuscript.

Eq. 1: Mention right away what theta and D represent in the equation.

An explanation of theta and D has been added to the text “where θ denotes the model parameters, and *D* represents the training dataset” in line 277.

L 367: "In probabilistic forecasting, the focus extends beyond mere point estimates to encompass the shape and dispersion of the probability distribution.": This sentence and the next could go to the beginning of the section to give a better motivation for the use of CRPS.

We have repositioned these two sentences to the beginning of section 2.3.2.

L 382: "y the predicted value, x the observed value". This works, but is not conventional. Typically, x are the predicted values and y denotes observations.

We have modified the formula accordingly.

L 393: The CDF is introduced here, but it has already been used above in the definition of CRPS. I would suggest switching the section order.

The order of presentation of CDF and CRPS has been adjusted.

L 483: "On using CAFE as a prediction target, both models show more consistent performance.": The term model has now been used to describe VGPM, CbPM, and CAFE, but also the NN and Bayesian model. Please ensure that the reader always knows what models are referenced in the text. Furthermore, this statement about consistent performance for both models seems to contradict a later one: "In addition, for NN model's MAPD index value for CAFE is lower than that for Bayes model" (l

487).

A clearer representation of the individual models has been made in the text to avoid ambiguity. What this section is trying to convey is that both NN and Bayesian models have better performance when CAFE is used as the prediction target than when the other two NPPs are used as the prediction target, and the presentation in section 3.1 of the article has been adjusted.

L 490: "Overall evaluation indicates that under both models' assessment criteria, CAFE demonstrates superior accuracy in predicting effects compared to VGPM and CbPM.": This paragraph is not very helpful. What are the two assessment criteria used here? (Fig. 5 uses three metrics, not two.) What does "predicting effects" mean? It is not helpful to the reader that the remaining paragraph discuss VGPM and CbPM results and not CAFE.

Section 3.1 has been restructured to use three evaluation criteria, CRPS, RMSD and MAPD, which are presented in L454-456, and the lower the three metrics, the better the model performance. L460-487 analyze these three metrics for NN and Bayesian models when different NPPs are used as prediction targets in order to evaluate the performance, which reveals that CAFE is used as the prediction target. NN and Bayesian performance was more favorable when the target was CAFE, and thus CAFE was chosen as the main prediction target for subsequent analysis.

L 499: "(1) prior research indicating that CAFE provides relatively accurate estimates of NPP in marine ecosystems with characteristics similar to the Weizhou Island area, due to its advanced parameterization of phytoplankton dynamics". Please cite this prior research or provide some evidence for this statement.

We have added the reference as below.

"(1) Previous studies have shown that for other NPP models analyzed for the same dataset, the CAFE model explains the most variance and has the lowest model bias, and also reproduces the magnitude and seasonality of field-measured NPP better than other satellite remote sensing models (Silsbe et al., 2016)."

L 520: Is this analysis based on the testing data or the full CAFE-based dataset?

This analysis was based on the full CAFE-based dataset.

L 523: Are these confidence intervals credible intervals for the Bayesian model?

This is the confidence interval, which has been described in line 526-528.

L 590: "Fig. 8 demonstrates the CDF curves of the predicted mean values after the normalization process and the CDF curves of the CAFE." This sentence and the next are difficult to understand. Are they meant to emphasize the advantages of normalizing the values? Why make this point right after stating that divergence between these two CDFs should be minimal? Please rephrase.

We have rephrased this paragraph as below.

"Fig. 8 demonstrates the CDF curves of the predicted mean values after the normalization process and the CDF curves of the CAFE. The CDF plots of the normalized data can reflect the statistical distribution of the datasets, especially when the different datasets have different magnitudes or scales, and the normalization can eliminate these differences, which makes the comparisons and analyses between the different datasets more accurate and intuitive. Fig. 9 specifically quantifies the difference between the two CDF curves in Fig. 8 at each point, which is accomplished by calculating the difference between the y-values of the two CDF curves at the same x-value. Optimally, the divergence between these two CDFs should be minimal, manifested as extensive overlap between the yellow and blue curves in Fig. 8, and the blue curve in Fig. 9 approaching zero."

L 671: Is the only difference between the estimates in this section and previous ones the daily resolution?

We thank the reviewer for the question. Yes, the primary difference in this section is the temporal resolution. While earlier sections focused on 8-day estimates aligned with remote sensing data, this section presents daily NPP estimates derived from our models. The objective is to bridge gaps between remote sensing observations and enable finer-resolution analysis of NPP dynamics. This higher temporal resolution also facilitates time series analysis to identify periodic patterns that may be obscured at coarser scales. We have clarified this point in the revised manuscript (line 677-683).

L 722: "By prioritizing variables such as SST and AP, the models can be optimized to reduce reliance on less influential inputs, improving efficiency without compromising accuracy." Was this actually shown? Did the authors try to run the NN or Bayes model with fewer input variables?

To assess input necessity, we conducted ablation experiments systematically removing individual input variables. In all cases, performance declined, confirming that each variable contributes meaningfully to model accuracy. Therefore, the full set of 11 variables was retained to ensure robust predictions.