Dear Editor and anonymous Reviewer,

We express our sincere gratitude for the insightful comments and constructive criticisms on our manuscript titled "*Refining marine net primary production estimates: Advanced uncertainty quantification through probability prediction models*" (MS No.: egusphere-2024-3221). In response to your valuable feedback, we have meticulously revised our manuscript to enhance its clarity, coherence, and overall scientific contribution. Specific modifications have been made to address each point raised by the reviewers, and these are detailed in the subsequent pages, where we provide a point-by-point response to your comments. Reviews' comments are in normal text, whereas our responses are in blue.

This revision process has been a collaborative effort among all co-authors, and we believe that the adjustments made significantly improve the manuscript. We are confident that these changes have addressed your concerns and enriched the manuscript.

With kind regards,

Mengyu Xie (on behalf of all co-authors)

Reviewer #2:

The manuscript presents a comparative analysis of Bayesian and neural network-based probability prediction models for estimating Net Primary Production (NPP) at a location near Weizhou Island (though this spatial focus is not clearly stated in the abstract or introduction). While the study demonstrates interesting methodological approaches to uncertainty quantification, it requires major revisions and clarifications.

general comments

The spatial scope and context of the study need to be clearly defined in the abstract and introduction. The location or spatial extent of the study is not mentioned in the title, abstract or introduction, suggesting a global analysis of marine NPP, when in fact the study focuses on a specific (point) location near Weizhou Island off the Chinese coast. Given the large number of inputs required for the Neural Network (NN) and Bayesian technique used in the study, it would not be easy to scale the approach to a larger region.

Response: We thank the reviewer for this constructive comment. In response, we have revised the abstract and introduction to clearly define the spatial scope and context of the study. Specifically, we have included the geographic focus on the aquatic ecosystem near Weizhou Island, located off the Chinese coast, to ensure transparency and accuracy in presenting the study's scope (Lines 28 - 32). Additionally, we have addressed the scalability limitations of the proposed approach in the discussion section (Lines 786 – 799) to provide a balanced view of the study’s applicability.

A critical limitation of the study is the data used for training the NN and the Bayesian model. The models are trained on outputs from existing NPP models (VGPM, CbPM, and CAFE) rather than directly on NPP data. Effectively, the NN and Bayesian model serve as emulators of the NPP models, inheriting their underlying errors and biases. Thus, the uncertainty estimates reported in the manuscript reflect the uncertainty in emulating the output, but not the uncertainty in estimating actual NPP. Furthermore, as shown in Fig. 3, estimates from VGPM, CbPM, and CAFE differ strongly, and it is not clear which output is more accurate. These points need to be explicitly acknowledged in the manuscript, as it means the reported uncertainty estimates do not represent true NPP estimation uncertainty.

Response: Thank you for raising this critical point. We acknowledge that the neural network (NN) and Bayesian models in this study are trained on outputs from the VGPM, CbPM, and CAFE models, rather than directly on observational NPP data. This is indeed a limitation, as it means that our models inherit the inherent biases and errors of these three base models. Consequently, the uncertainty estimates reported in the manuscript reflect the uncertainty in emulating these models' outputs, rather than representing the true uncertainty of NPP estimation. We have revised the manuscript to explicitly acknowledge this limitation in both the discussion and conclusion sections (Lines 476–483 and Lines 763 – 776).

To further clarify, as shown in Fig. 3, the outputs of VGPM, CbPM, and CAFE differ significantly, highlighting the variability in model estimates and the lack of ground truth data to determine which output is more accurate. This variability contributes to the challenges in validating our models' uncertainty estimates against true NPP values. Based on the current knowledge and previous reviews, it is reasonable to consider CAFE NPP estimates as potentially more accurate, but this assumption requires further validation with in situ measurements.

In our previous review process, a similar concern was raised regarding the VGPM model's potential underperformance at the study site, particularly for NPP values exceeding 350 mg C m^{-2} d^{-1}. In response, we normalized the data before calculating CRPS, RMSD, and MAPE to ensure fair comparisons between the models. These revisions improved the performance evaluation of our statistical models and aligned the results more closely with expectations, showing that the models perform best with CAFE NPP as the prediction target. We have now extended this discussion to acknowledge the broader implications of relying on modeled NPP outputs for training (Lines 763 – 776).

The differences between VGPM, CbPM, and CAFE output raise questions about which model provides the best NPP estimates and the most reliable training data. The current version of the manuscript initially does not mention which of the 3 models provided the output used to generate the full time series of NPP estimates near Weizhou Island in Section 3.3. Section 4 finally reveals that CAFE was used to generate the NPP training data, but that choice appears to have been motivated by results showing that the NN and the Bayesian model can emulate CAFE output well and not that CAFE output best represents true NPP.

Response: Thank you for this insightful comment. We agree that the differences between VGPM, CbPM, and CAFE outputs raise critical questions about which model provides the most accurate representation of true NPP. To address this concern, we have revised the manuscript to clarify our rationale for focusing on CAFE as the prediction target, as highlighted in Section 3.1. Specifically, our choice was not solely based on the NN and Bayesian models' ability to emulate CAFE output effectively but also on prior studies suggesting that CAFE tends to provide more accurate estimates of NPP under certain conditions, particularly in the study area near Weizhou Island. These revisions can be found in Section 3.1 (Lines 500 – 509) and Section 4 (Lines 770 – 781).

Additionally, we have explicitly clarified earlier in the manuscript (Section 3.3, Lines 684 – 687) that CAFE outputs were used to generate the full time series of NPP estimates for subsequent analysis. This revision ensures consistency and transparency throughout the manuscript.

In the context of the above comments, it would be interesting for the reader to know what inputs VGPM, CbPM, and CAFE used to generate their results. If the NN or the Bayesian model require more or more difficult to measure input data than VGPM, CbPM, or CAFE, why use them at all? Similarly, it would be interesting to investigate which of the inputs to the NN or the Bayesian model are actually required to obtain good performance.

Response: Thank you for your thoughtful comment. We would like to clarify that the two probabilistic models (NN and Bayesian) used in our study do not necessarily require more or more difficult-to-measure inputs compared to VGPM, CbPM, and CAFE. Instead, the inputs used by these models overlap substantially with the inputs used to generate VGPM, CbPM, and CAFE, such as sea surface temperature (SST), photosynthetically active radiation (PAR), and atmospheric pressure (AP). This ensures that the NN and Bayesian models are comparable to traditional models in terms of data requirements.

To address the reviewer's suggestion about investigating which inputs are most critical, we conducted a Pearson correlation analysis (Figure 4). This analysis helps identify the most relevant variables for predicting NPP and reveals variability in their influence across VGPM, CbPM, and CAFE. By leveraging these correlations, it is possible to filter out less relevant variables, thereby refining the models and reducing their reliance on less critical inputs without sacrificing predictive performance.

We have revised the manuscript to include detailed information on the input variables used by VGPM, CbPM, and CAFE in Section 2.1 (Lines 229 – 236). Furthermore, we have highlighted the role of Pearson correlation in identifying the most influential variables for the NN and Bayesian models Section 2.2 (Lines 327 – 330) and discussed how this process can optimize model prediction in Section 4 (Lines 732 – 735).

The manuscript's writing style suggests the use of AI-assisted writing, which, while not problematic in itself, has led to the use of emphatic language and filler words (such as "pivotal", "integral", "advanced", "comprehensive", "indispensable", "paramount", etc.). The manuscript would benefit from removing these words in places and rewording passages.

Response: We appreciate your observation regarding the use of emphatic language and filler words. As English is not our first language, we recognize that achieving the appropriate academic tone can be challenging. To address your comment, we have revised the manuscript to remove or replace excessive emphatic language and filler words, focusing on more precise and concise expressions. Specific examples include replacing “pivotal” with “important” or “key”, “integral” and “indispensable” as “essential”, and etc. These changes ensure the writing aligns with a formal academic tone and emphasizes the scientific content of the study. We hope these changes enhance the manuscript’s readability and tone while addressing your concerns.

A few passages in the manuscript appear to suggest surprise in discovering periodicity in NPP values: "Upon visualizing the values of the three NPP products (VGPM, CbPM, and CAFE) (Fig. 3), it became evident that each exhibits a distinct periodicity" (l 198). "The analysis of the annual change of NPP shows a clear periodicity, which means that the change of NPP is not random, but follows certain laws and patterns." (l 571). Even at 21 degrees north, one can expect seasonal patterns in marine primary production - this context should be provided in the text.

Response: Thank you for highlighting this point. We agree that periodicity in marine primary production, particularly in regions around 21 degrees north, is to be expected due to seasonal variations in environmental conditions such as light availability, temperature, and nutrient dynamics. In the revised manuscript, we have added contextual information to clarify that the observed periodicity in NPP values aligns with established knowledge of seasonal patterns in marine ecosystems. These revisions provide appropriate context and avoid any unintended implication of surprise. We have updated the relevant sentences to emphasize that our analysis confirms these expected periodic trends and highlights how the three NPP products (VGPM, CbPM, and CAFE) capture this periodicity differently (Lines 237 – 253).

specific comments

L 117: What are "stochastic optimization" and "advanced chance constraints"? They are only used here and nowhere else in the manuscript. It would be useful to describe relevant new concepts to the reader right away, or not mention them when they are not used or described in the manuscript.

Response: Thank you for your comment. We agree that mentioning "stochastic optimization" and "advanced chance constraints" without further elaboration may confuse the reader, especially since these terms are not central to the rest of the manuscript. To address this, we have revised the text to focus on the broader advantages of probabilistic forecasting without introducing concepts that are not further explored in the study (Lines 123 - 128). This ensures that the text remains clear and directly relevant to the manuscript's objectives.

L 149: What does "sea accumulation" mean?

Response: Thank you for pointing out this ambiguity. The term "sea accumulation" was intended to refer to depositional features created by the accumulation of marine sediments, such as beaches, sandbars, and other sedimentary formations resulting from wave action and tidal processes. To improve clarity, we have revised the text to use more precise terminology (Lines 164 - 165).

L 149: "Surrounded by the sea on all sides, Weizhou Island ...": I think this is the definition of an island.

Response: Thank you for your observation. We acknowledge that the phrase "surrounded by the sea on all sides" is redundant and unnecessary. We have revised the text to remove this redundancy and focus on the unique climatic and oceanographic conditions of Weizhou Island, which are relevant to marine primary production. The revised text clarifies the study's focus on the marine environment rather than terrestrial variables (Lines 156 - 171).

L 168: "For the analysis of three NPP algorithms - namely, VGPM, CbPM, and CAFE - we acquired datasets at an eight-day temporal resolution ...": Here it is unclear to the reader if the "acquired datasets" are the input required to run the algorithms or their output. I assume it is the latter, but that should be made more explicit.

Response: Thank you for this helpful observation. You are correct that the "acquired datasets" refer to the outputs of the VGPM, CbPM, and CAFE algorithms rather than the inputs required to run them. To clarify this, we have revised the text to explicitly state that the datasets represent the outputs of the three NPP algorithms (Lines 192 – 194).

L 177/Table 2: Just listing the numbers of missing entries is not very informative. At which frequency were they recorded?

Response: Thank you for pointing out this issue. Most missing data are due to satellite equipment malfunctions or severe weather conditions, which occur randomly and are not tied to any specific frequency. Therefore, while we have quantified the number of missing values, analyzing their frequency is not feasible or meaningful for this study. To clarify this in the manuscript, we have revised the accompanying text for Table 2 to explain the source and nature of the missing data (Lines 202 - 207).

L 198: "Upon visualizing the values of the three NPP products (VGPM, CbPM, and CAFE) (Fig. 3), it became evident that each exhibits a distinct periodicity, with the fluctuation ranges remaining stable yet markedly varied among them." What exactly does this mean? Do the signals not have an underlying annual periodicity?

Response: Thank you for pointing out this ambiguity. The periodicity observed in the NPP products is primarily seasonal rather than annual. To clarify this, we have revised the text to explicitly describe the seasonal periodicity of the NPP signals and to avoid confusion regarding the nature of the observed fluctuations (Lines 240 – 253).

L 311: Samples are mentioned here for the first time and need a better introduction.

Response: Thank you for highlighting this point. To provide a better introduction to the term "samples," we have revised the text to clarify its meaning and context in this study. The updated text ensures that the term is clearly defined before it is used (Lines 376 – 379).

Eq. 3: This looks like a recursive definition of CRPS, I would suggest using different names for the "CRPS" used in Eq. 2 and 3.

Response: Thank you for your suggestion. We agree that using the same name "CRPS" in both equations may cause confusion, as Eq. 2 refers to the CRPS for a single prediction-observation pair, whereas Eq. 3 represents the average CRPS over multiple samples. To address this, we have revised the text and equations to use distinct names for these two forms of CRPS. Specifically, we have renamed the CRPS in Eq. 2 as "CRPS_individual" to denote its use for individual pairs and retained "CRPS" in Eq. 3 to indicate the aggregated metric over all samples.

Eq. 4: The notation is inconsistent: In Eq. 2 and 3, x denotes the observed value and y the predicted value, but in Eq. 4 and 5, y is used for the actual/observed value and y-hat for the predicted value.

Response: Thank you for pointing out the inconsistency in the notation. To ensure clarity and uniformity, we have revised the manuscript to maintain consistent notation throughout. Specifically: For Eq. 4 (now Eq. 5 in the revised manuscript), x is used for observed values and y for predictions, aligning with the notation in earlier equations. For Eq. 5 (now Eq. 6 in the revised manuscript), the same notation is applied, and the equation and accompanying text have been updated to reflect this change.

L 501: The test data distribution for CAFE NPP does not look similar to that of the train data distribution, suggesting that the test data may not be well-represented by the train data.

Response: Thank you for your observation. While we agree that the CDF curves for the test and train datasets appear different in Fig. 8, this discrepancy may not necessarily indicate that the test data are poorly represented by the training data. The difference can be attributed to the smaller size of the test dataset relative to the training dataset, which can lead to visual differences in the CDF curves. Furthermore, Fig. 7 demonstrates that the patterns for simulating the training set and predicting the test set are consistent for both the NN and Bayesian models. This similarity suggests that the models generalize well to the test data despite the apparent differences in the CDF curves. To clarify this point, we have revised the text (Lines 608 - 635) to provide additional context and improve the rigor of the discussion.

L 503: What is the difference between the values shown in Table 5 and Fig. 4?

Why not combine the two?

Response: Thank you for raising this point. We believe the reference to Fig. 4 in the comment is a typo and that the reviewer intended to refer to Fig. 5, as the content in Fig. 5 is more closely related to Table 5. To clarify, the content described in Table 5 and Figure 5 is different and serves distinct purposes in the manuscript. Table 5 presents the evaluation metrics (e.g., CRPS, RMSD, and MAE) specifically for the CAFE, providing detailed numerical results from the NN and Bayesian models when CAFE is the prediction target. Figure 5 visually compares the evaluation metrics (CRPS, RMSD, and MAE) for all three NPP models (VGPM, CbPM, and CAFE) under both NN and Bayesian models, offering a broader view of model performance across all prediction targets. While both Table 5 and Figure 5 address evaluation metrics, they serve complementary roles. Table 5 provides precise numerical data for CAFE-specific metrics, while Figure 5 visually demonstrates trends and comparisons across all three NPP models. Keeping them separate allows readers to access both detailed and visualized information relevant to different aspects of the study. We have clarified these in the revised manuscript to avoid confusion (Lines 447 - 448， Lines 653 - 654).

Fig. 2 and 3: The date label locations 2007/1/1, 2008/3/13, 2009/5/25, ... make it difficult to interpret the plot and detect seasonality.

Response: Thank you for your suggestion. We appreciate your observation regarding the difficulty in interpreting seasonality based on the current time annotations in Figures 2 and 3. However, the date labels in the graph are divided based on equal and reasonable time intervals to ensure consistency in visual representation. Since our research does not specifically focus on seasonal distribution patterns, we did not classify or annotate the data explicitly by seasons. That said, we understand the importance of making the plots easier to interpret. To address your concern, we have revised the date labels in Figures 2 and 3 to align with annual markers (e.g., January 1st of each year) to improve readability and facilitate interpretation of potential seasonal trends. These updates allow readers to better observe patterns over time while maintaining the integrity of the original analysis. The revised figures can be found in the updated manuscript.

Fig. 4: The caption mentions "input variables". Are these inputs to VGPM, CAFE, and CbPM?

Response: Yes, it has been clearly stated in the manuscript that the 11 variables presented in Fig. 4 are used as input variables for the models, while VGPM, CAFE, and CbPM serve as the model outputs (Lines 316 - 320). To further clarify, these input variables represent environmental and oceanographic factors that are used by the probabilistic models (NN and Bayesian) to predict NPP values derived from VGPM, CAFE, and CbPM as outputs. We have reviewed the caption for Fig. 4 to ensure this connection is explicitly stated, improving clarity for readers. The revised caption now specifies that the "input variables" are those used to train the probabilistic models, and their correlations with the VGPM, CAFE, and CbPM outputs are visualized in the figure.

Fig. 5: Why does the y-axis go past 0.8 in panels a, b, d and e, when the values all stay below 0.4? Also, the units are missing.

Response: Thank you for pointing out this issue. We have re-plotted panels (a), (b), (d), and (e) of Fig. 5 to set the maximum y-axis value to 0.5, which better reflects the observed range of values and improves visual clarity. Additionally, we have clarified the units in the figure captions. Specifically, the NPP values in these panels are normalized to the range of 0–1, making the axes dimensionless. For MAPD in panels (c) and (f), the units are expressed as percentiles (%). These have been clarified in the figure caption. The updated figure is included in the revised manuscript.

Fig. 8 and 9: The NPP units here are incorrect. The data appears to have been normalized, but why? Without normalization, it would be easier to interpret for which

NPP ranges the NN and the Bayes model over- or underestimate VGPM NPP.

Response: Thank you for your observation regarding the units in Figures 8 and 9. The data in these figures were normalized to a scale of 0–1 to ensure consistency across different metrics and facilitate direct comparison between the NN and Bayesian models' predictions. Normalization is particularly meaningful in this context, as it mitigates the effects of differences in magnitude between datasets, enabling a fair assessment of the models' performance. While we understand that presenting the data in its original units could provide direct insight into specific NPP ranges, the normalized format helps maintain a unified framework across the study, especially when comparing metrics like CRPS and residuals. To address potential confusion, we have updated the captions for Figures 8 and 9 in the revised manuscript to explicitly state that the data is normalized and explain the rationale for this approach.