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 #1:

Detailed Review for '*Refining Marine Net Primary Production Estimates: Advanced Uncertainty Quantification through Probability Prediction Models*', Jie Niu et al.

1. Line 26: In the abstract, the source of the NPP estimate (i.e., model output or observation) used in the paper should be mentioned.

We have revised the abstract to explicitly mention the source of the NPP estimate and have provided details about the research location and data used in the study (Lines 28-34).

"This study focuses on the aquatic environs of Weizhou Island, located off the coast of Guangxi, China, and introduces an advanced probability prediction model aimed at improving NPP estimation accuracy while addressing its associated uncertainties. The dataset comprises eleven distinct sets of monitoring data and satellite data spanning from January 2007 to February 2018. NPP values were derived using three widely recognized estimation methods — VGPM, CAFE, and CbPM — serving as model outputs for further analysis."

2. Line 30: The author should explain the nature and the sources of uncertainty in NPP estimates. And why it is important.

We have revised the abstract to clarify the sources and nature of uncertainty in NPP estimates and to emphasize their significance. Specifically, we have included information about the challenges arising from measurement difficulties, errors in satellite-based inversion, and the need for reliable uncertainty quantification to improve ecosystem management and global carbon cycle modeling (Lines 23 - 28).

"In marine ecosystems, Net Primary Production (NPP) is important, not merely as a critical indicator of ecosystem health, but also as an essential component in the global carbon cycling process. Despite its significance, the accurate estimation of NPP is plagued by uncertainty stemming from multiple sources, including measurement challenges in the field, errors in satellite-based inversion methods, and inherent variability in ecosystem dynamics."

3. Line 61-61: It is important to mention the recent study in Satyendranath et al. 2020 (Reconciling models of primary production and photoacclimation, Applied Optics)

We thank the reviewer for highlighting this relevant study. In response, we have incorporated a reference to Satyendranath et al. (2020) into the revised manuscript. Specifically, we have added a sentence to emphasize their contribution to improving primary production models by addressing parameter assignment and its impact on reducing uncertainties (Lines 77 - 80).

"Satyendranath et al. (2020) emphasize the critical role of accurately assigning parameters in primary production models as a key strategy for reducing model uncertainties and enhancing the reliability of satellite-based primary production estimates, particularly in the context of climate research."

4. Line 122: Again, it's important to mention why estimating uncertainty is important?

In the revised manuscript (Lines 129 - 135), we have included sentences to emphasize the significance of uncertainty estimation.

“The estimated values of NPP derived from the above three classical models exhibit significant discrepancies, reflecting substantial uncertainties in these methods. These inaccuracies can impede a comprehensive understanding of the role of oceans in the global climate system, particularly in their capacity to act as carbon sinks and regulators of atmospheric CO2 levels. Consequently, quantifying and addressing these uncertainties is primary to improving the reliability of NPP estimates and ensuring their applicability in climate research and marine ecosystem management.”

5. Line 137-138: Authors should rephrase “discloses the results” to “discusses the results”.

Corrected (Line 153).

“ Section 4 discusses the results; and Section 5 presents the conclusions.”

6. Line 167-167: Why are these variables (input features) important in terms of estimating NPP?

We have added detailed explanations in the revised manuscript to clarify the relevance and importance of the input variables for estimating NPP (Lines 182 - 191).

“These data were aggregated to constitute a comprehensive dataset encompassing eleven variables, serving as the input features for the models. Phytoplankton, the primary source of NPP, is directly influenced by variables such as SST, Par, and SH, which are critical to its photosynthetic processes. Additionally, other variables have significant indirect effects on phytoplankton growth. Sal, for example, influences the community structure of phytoplankton (Braarud et al., 1951). Variables such as TH, H/10, and WS indirectly affect phytoplankton dynamics by modulating water column mixing and the vertical distribution of nutrients. AP, RH and SV also indirectly impacts phytoplankton photosynthetic activity by altering environmental conditions.”

7. Line 164: For PAR, SSP, SH and NPP data, authors should mention direct links for the data they used for experiments.

We have provided direct links to the datasets used in this study (Lines 179-182, Lines 195 - 197).

“For the analysis of three NPP algorithms—namely, VGPM, CbPM, and CAFE—we utilized their output datasets, which were obtained at an eight-day temporal resolution from the Ocean Productivity website (http://orca.science.oregonstate.edu/1080.by.2160.monthly.hdf.vgpm.m.chl.m.sst.php, http://orca.science.oregonstate.edu/1080.by.2160.monthly.hdf.cbpm2.m.php, http://orca.science.oregonstate.edu/1080.by.2160.monthly.hdf.cafe.m.php).”

8. Line 783: Table-1: No need to mention the links here, acronyms are sufficient.

We have corrected Table 1 by removing the dataset links and retaining only the acronyms, as suggested.

Table 1. Summary of Variables and Data Sources.

Variable name	Variable description	Data source
SST	Sea surface temperature (°C)	Weizhou Marine environment monitoring station
Sal	Salinity (‰)	
TH	Height of tide(m)	
AP	Air pressure (hPa)	
RH	Relative humidity (%)	
SV	Sea visibility (km)	
WS	Wind speed ($m \cdot s^{-1}$)	

H/10	1/10th significant wave height (m)	
PAR	Photosynthetically active radiation ($W \cdot m^{-2}$)	Oceancolor
SSP	Sea surface precipitation (mm)	Earthdata
SH	Sunshine hours ($h \cdot d^{-1}$)	China Meteorological Administration
VGPM	NPP from the VGPM model ($mgC\ m^{-2} \cdot d^{-1}$)	
CbPM	NPP from the CbPM model ($mgC\ m^{-2} \cdot d^{-1}$)	Ocean Productivity
CAFE	NPP from the CAFE model ($mgC\ m^{-2} \cdot d^{-1}$)	

9. Line 785: Table-2: Authors should be more clear about the “missing quantity” units i.e., days.

Thank you for your reminder. We have updated Table 2 to include the unit "days" for the “missing quantity” column.

Table 2. Summary of Missing Variables.

Variable	SV (km)	H/10 (m)	PAR ($W \cdot m^{-2}$)	SSP (mm)	SH ($h \cdot d^{-1}$)
Missing quantity (days)	31	51	828	378	18

10. Line 187: What specific algorithm was applied to make the time series interpolation.

In our research, we used the ‘interpolate’ function from the Python Pandas library, configured with the 'time' method, to perform the time series interpolation. This approach, while classified as linear interpolation, incorporates the time factor, ensuring that the intervals between timestamps are explicitly considered. This feature enhances its suitability for time series data, particularly datasets with periodic variations like those in our study, enabling more accurate estimation of missing values. Although it is computationally simpler than periodic interpolation methods (e.g., Fourier transform or time series models with seasonal decomposition), the ‘time’ method sufficiently captures the periodicity and variations inherent in our dataset, making it both efficient and effective for this application (Line 224-228).

“ In this study, interpolation was used to address missing variables, and we ensure that the statistical properties of the original data were preserved to the greatest extent possible. This approach allows us to maintain the integrity of our analyses while recognizing the inherent limitations of using interpolated data.”

11. Line 198-216: Authors can drop using “NPP” repeatedly, just the algorithm name is sufficient.

Corrected.

12. Line 208-211: It is not clear why CbPM is negatively correlated with AP. Authors should give an explanation.

Thank you for raising this insightful question. In response, we have elaborated on the relationship between AP and CbPM in the revised manuscript, providing an explanation for the observed negative correlation (Lines 260 - 264).

“Changes in AP affect atmospheric stability, cloudiness, and precipitation, indirectly altering light conditions in the ocean and subsequently affecting phytoplankton photosynthesis. Lower AP often corresponds to unstable atmospheric conditions and increased cloud cover, which may inhibit photosynthesis activity by reducing light penetration.”

13. Line 223: Typo in equation number.

Corrected. (Line 275)

14. Line 282: It is not clear whether the author had normalised the input features since they are in different scales.

At the beginning of Section 2.3 of the article, it has been clarified that the input data of different scales have been normalized (Line 341).

“Prior to model evaluation, we normalized the NPP satellite data.”

15. Line 378: Do the authors have any explanation behind finding the lowest CPRS value than the other models?

In the revised manuscript (Section 3.2.2, Lines 459 – 470, and 476 – 485), we have elaborated on potential factors contributing to the lower CRPS value for the CAFE model, in terms of both variance and cumulative distribution function. Also, Figs. S1 to S6 have been added in the SI to better explain the differences among the training and testing datasets of three NPPs.

“The lower CRPS value for the CAFE NPP, compared to VGPM and CbPM, may stem from the fact that its probability distribution aligns more closely with the prediction of models in terms of both shape and central tendency, since CRPS evaluates the full probability distribution, incorporating factors such as skewness and kurtosis in addition to variance. In the case of CAFE, the probabilistic structure of its predictions may exhibit better congruence with the observed cumulative distribution function (CDF) (Section 3.2.2, Figs. 8 and 9, Figs. S2, S3, S5, and S6), particularly in regions with higher data density. This enhanced alignment could compensate for its slightly larger variance compared to CbPM, thereby resulting in a lower CRPS value. Additionally, the design and parameterization of the CAFE model may inherently emphasize features that lead to improved probabilistic predictions, which warrants further investigation.”

“The neural network and Bayesian models developed in this study were trained using outputs from the VGPM, CbPM, and CAFE models. While this approach allowed us to evaluate the uncertainty in emulating these base models, it also means that our models inherit their underlying biases and errors. As such, the uncertainty estimates reported here reflect the uncertainty in emulating these specific outputs and do not represent the true uncertainty of NPP estimation. Furthermore, as Fig. 3 demonstrates, the outputs of VGPM, CbPM, and CAFE differ significantly, underscoring the need for ground truth data to validate these models. Among these, CAFE NPP is often considered more accurate based on prior studies, but further validation with observational data is necessary to confirm this assumption."

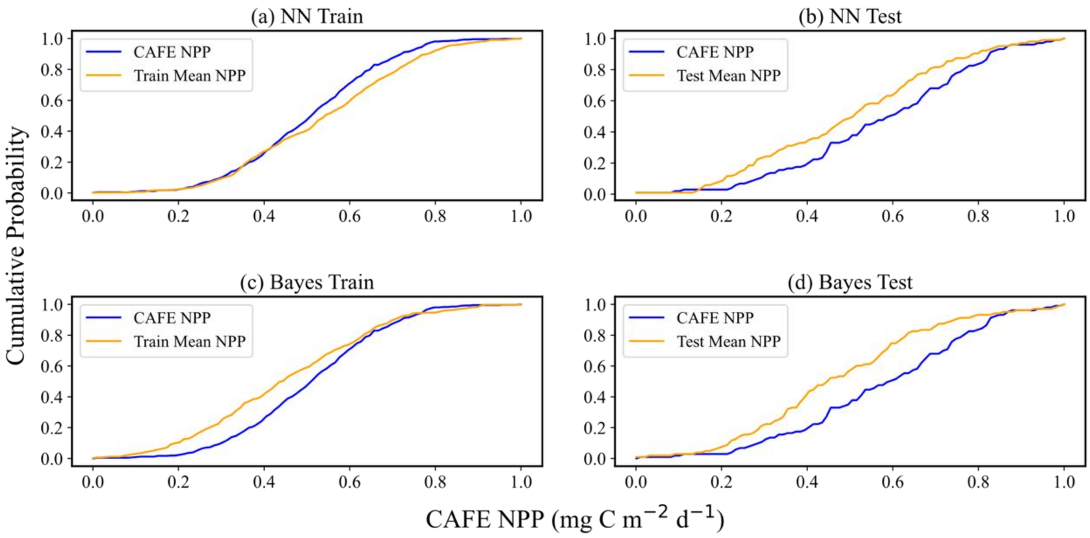

Fig. 8. Comparison of CAFE and predicted mean CDF. Panels (a) and (b) display the performance of the training and test sets, respectively, in the neural network-based probabilistic prediction model. Panels (c) and (d) illustrate the performance of the training and test sets, respectively, in the empirical distribution-based Bayesian probabilistic prediction model. The data has been normalized to a scale of 0–1 to ensure consistency across metrics and facilitate direct comparison between the two models. In each panel, the blue curves represent the CDFs of the CAFE values, while the yellow curves depict the CDFs of the model's predicted mean values.

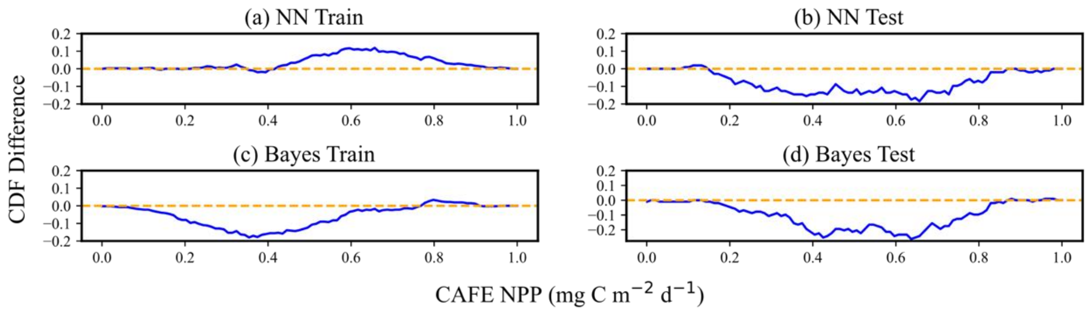

Fig. 9. Difference between the CAFE CDF and predicted mean CDF of model predictions. Panels (a) and (b) represent the performance of the training set and test sets, respectively, in the neural network-based probabilistic prediction model. Panels (c) and (d) showcase the performances of the training set and test sets, respectively, in the empirical distribution-based Bayesian

probabilistic prediction model. The residuals are expressed in normalized units (0–1), enabling consistent assessment of model performance across different NPP ranges. The blue curves in each panel indicate the differential magnitude of the CDFs. Instances where the blue curves align with the yellow lines denote zero discrepancy between the input data CDF and the model's predicted mean CDF.

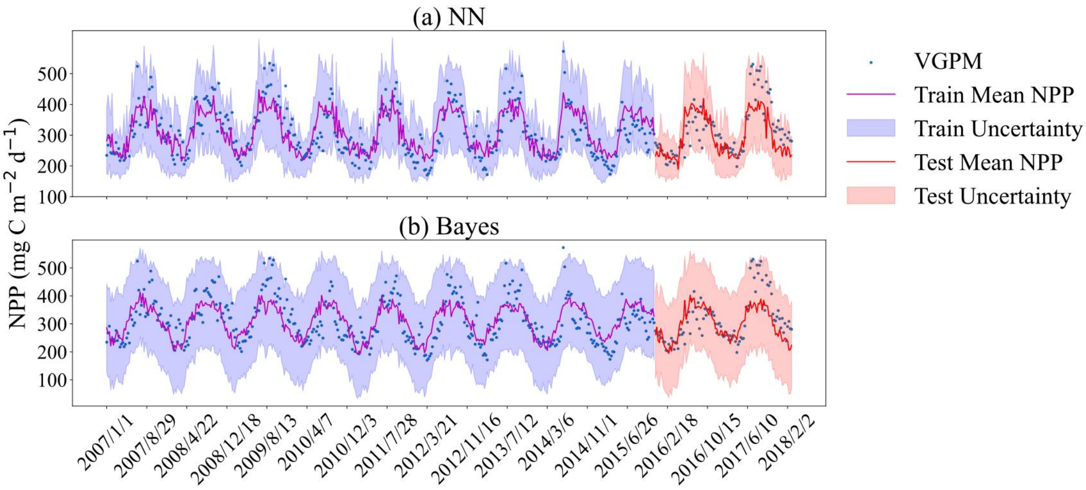

Fig. S1. Comparison of VGPM and predicted mean values at an 8-day temporal resolution within a 95% confidence interval. (a) Probabilistic prediction results are based on neural networks; (b) Bayesian probabilistic prediction results are based on empirical distributions. The dashed lines represent the mean values of the probabilistic predictions. The purple and red shaded areas illustrate the uncertainty ranges for the training and the test sets, respectively. Blue dots signify observed data points. All predictions and observations are presented in chronological sequence.

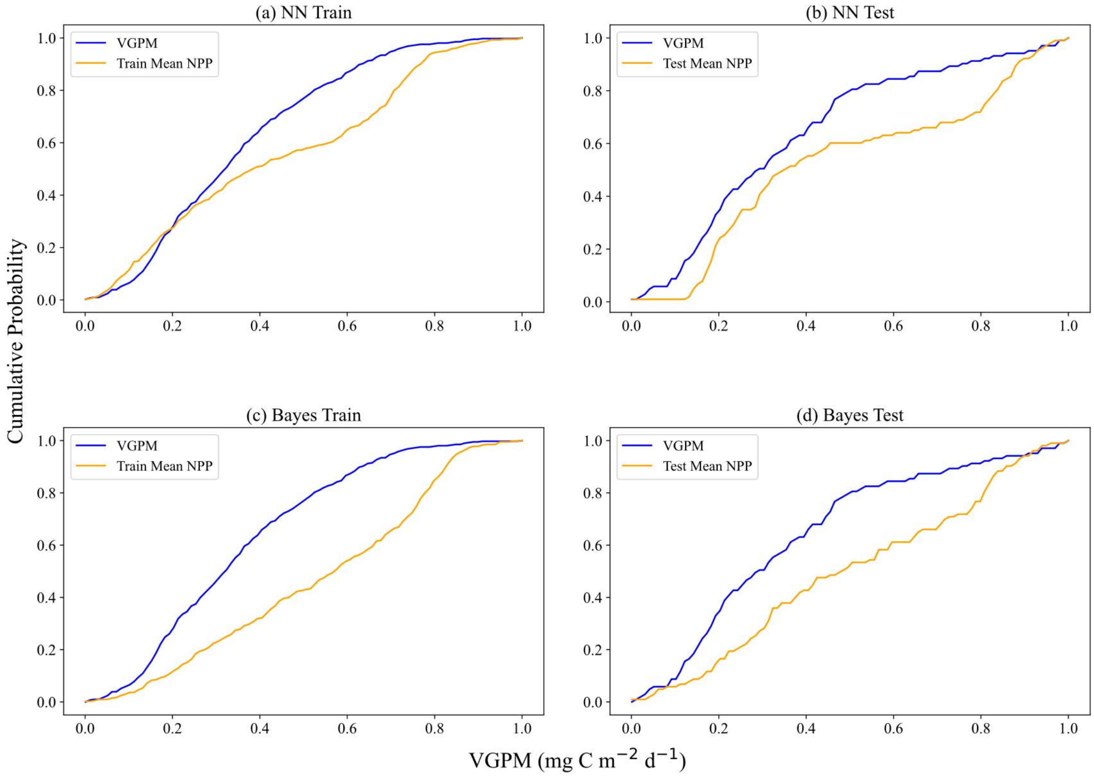

Fig. S2. Comparison of VGPM and predicted mean CDF. Panels (a) and (b) display the performance of the training and test sets, respectively, in the neural network-based probabilistic prediction model. Panels (c) and (d) illustrate the performance of the training and test sets, respectively, in the empirical distribution-based Bayesian probabilistic prediction model. The data has been normalized to a scale of 0–1 to ensure consistency across metrics and facilitate direct comparison between the two models. In each panel, the blue curves represent the CDFs of the CAFE values, while the yellow curves depict the CDFs of the model's predicted mean values.

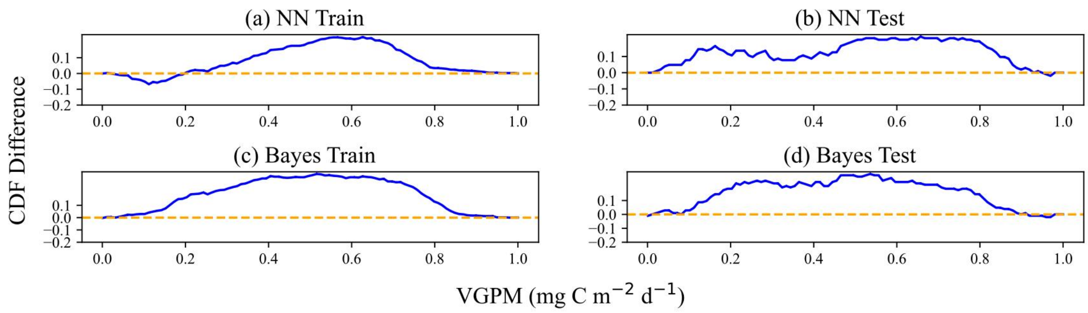

Fig. S3. Difference between the VGPM CDF and predicted mean CDF of model predictions. Panels (a) and (b) represent the performance of the training set and test sets, respectively, in the neural network-based probabilistic prediction model. Panels (c) and (d) showcase the performances of the training set and test sets, respectively, in the empirical distribution-based Bayesian probabilistic prediction model. The residuals are expressed in normalized units (0–1), enabling consistent assessment of model performance across different NPP ranges. The blue curves in each panel indicate the differential magnitude of the CDFs. Instances, where the blue curves align with the yellow lines, denote zero discrepancy between the input data CDF and the model's predicted mean CDF.

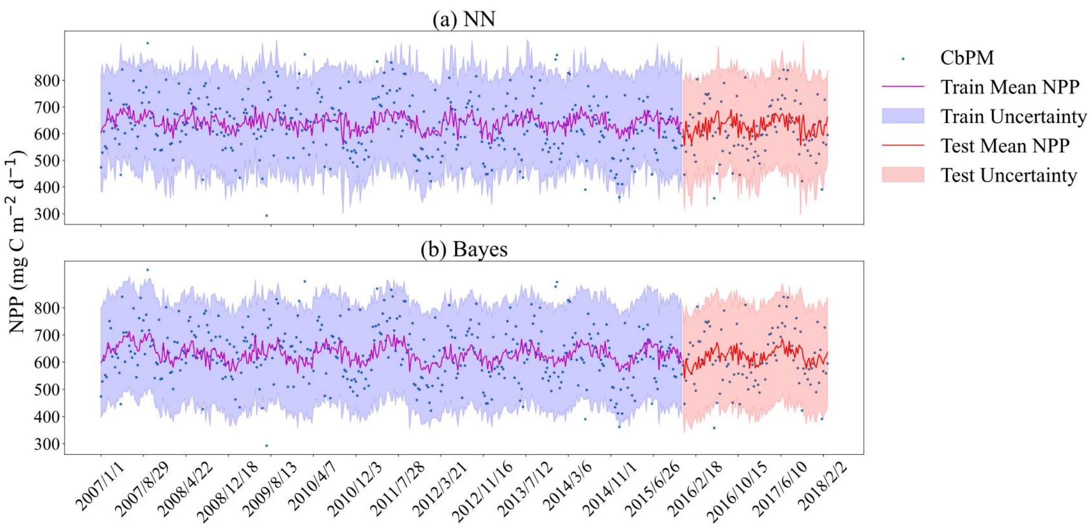

Fig. S4. Comparison of CbPM and predicted mean values shown at an 8-day temporal resolution within a 95% confidence interval. (a) Probabilistic prediction results are based on neural networks; (b) Bayesian probabilistic prediction results are based on empirical distributions. The dashed lines represent the mean values of the probabilistic predictions. The purple and red shaded areas illustrate the uncertainty ranges for the training and the test sets, respectively. Blue dots signify observed data points. All predictions and observations are presented in chronological sequence.

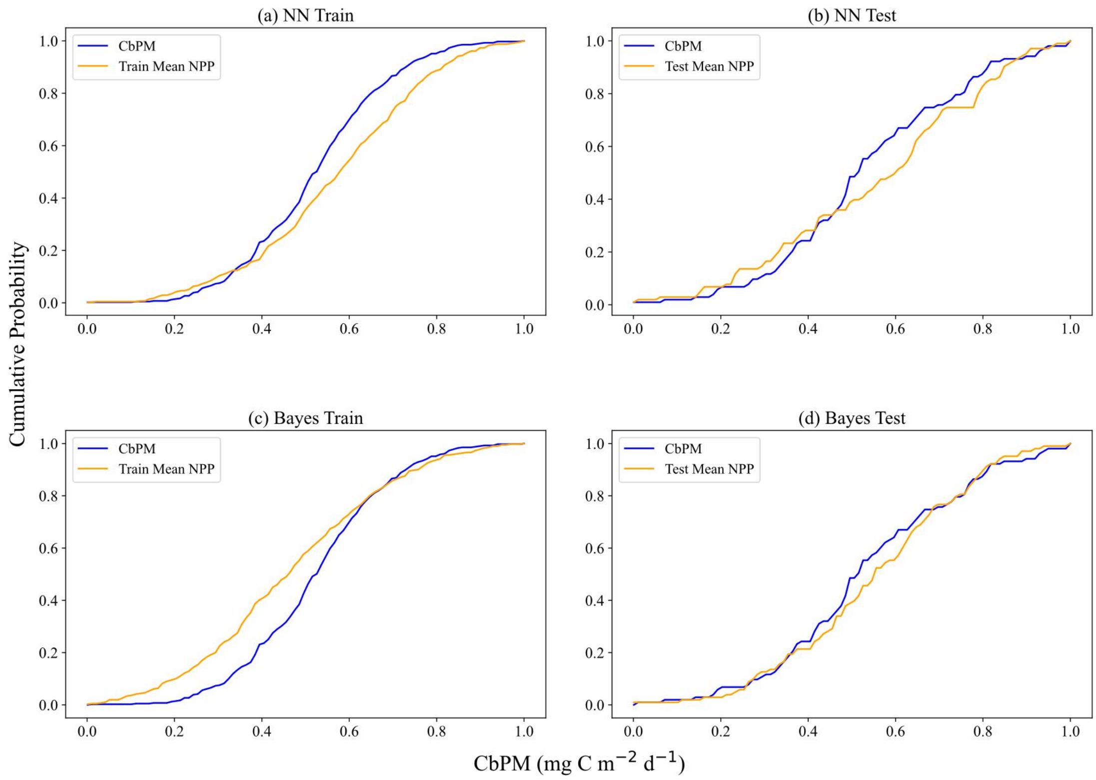

Fig. S5. Comparison of CbPM and predicted mean CDF. Panels (a) and (b) display the performance of the training and test sets, respectively, in the neural network-based probabilistic prediction model. Panels (c) and (d) illustrate the performance of the training and test sets, respectively, in the empirical distribution-based Bayesian probabilistic prediction model. The data has been normalized to a scale of 0–1 to ensure consistency across metrics and facilitate direct comparison between the two models. In each panel, the blue curves represent the CDFs of the

CAFE values, while the yellow curves depict the CDFs of the model's predicted mean values.

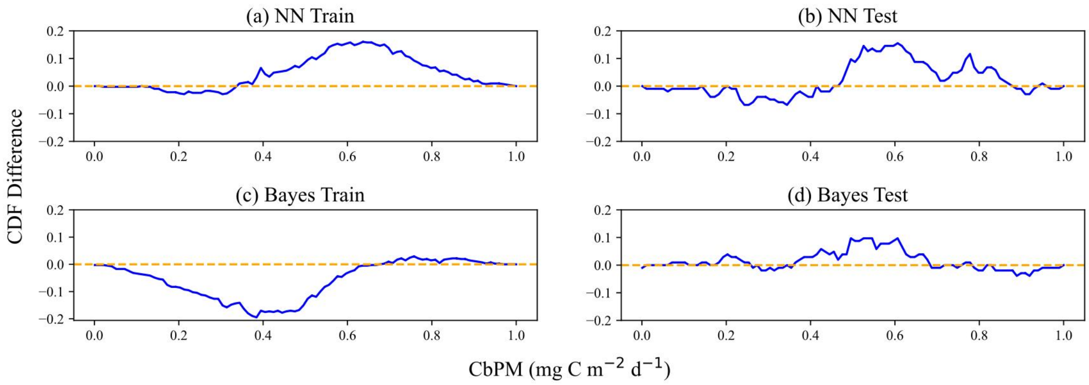

Fig. S6. Difference between the CbPM CDF and predicted mean CDF of model predictions. Panels (a) and (b) represent the performance of the training set and test sets, respectively, in the neural network-based probabilistic prediction model. Panels (c) and (d) showcase the performances of the training set and test sets, respectively, in the empirical distribution-based Bayesian probabilistic prediction model. The residuals are expressed in normalized units (0–1), enabling consistent assessment of model performance across different NPP ranges. The blue curves in each panel indicate the differential magnitude of the CDFs. Instances, where the blue curves align with the yellow lines, denote zero discrepancy between the input data CDF and the model's predicted mean CDF.

16. Line 466-467: Applying a low pass filter on the time series is recommended before reaching this conclusion about long-term trend.

We have applied a low-pass filter to the time series data for the three NPPs to isolate the long-term trends. The filtered results have been included in the revised Figure 3 to visually represent the smoothed trends, ensuring the analysis and conclusions are supported by appropriately processed data (Lines 237 – 243).

"To evaluate the long-term trends in Net Primary Production (NPP), we applied a low-pass filter to the three NPP products (VGPM, CbPM, and CAFE) (Fig. 3). This filtering process removes high-frequency variations, such as noise and short-term fluctuations, while retaining the underlying long-term patterns. It became evident that each exhibits a distinct seasonal periodicity, with the fluctuation ranges remaining stable over time yet the magnitude and timing of them varing significantly among the three NPPs."

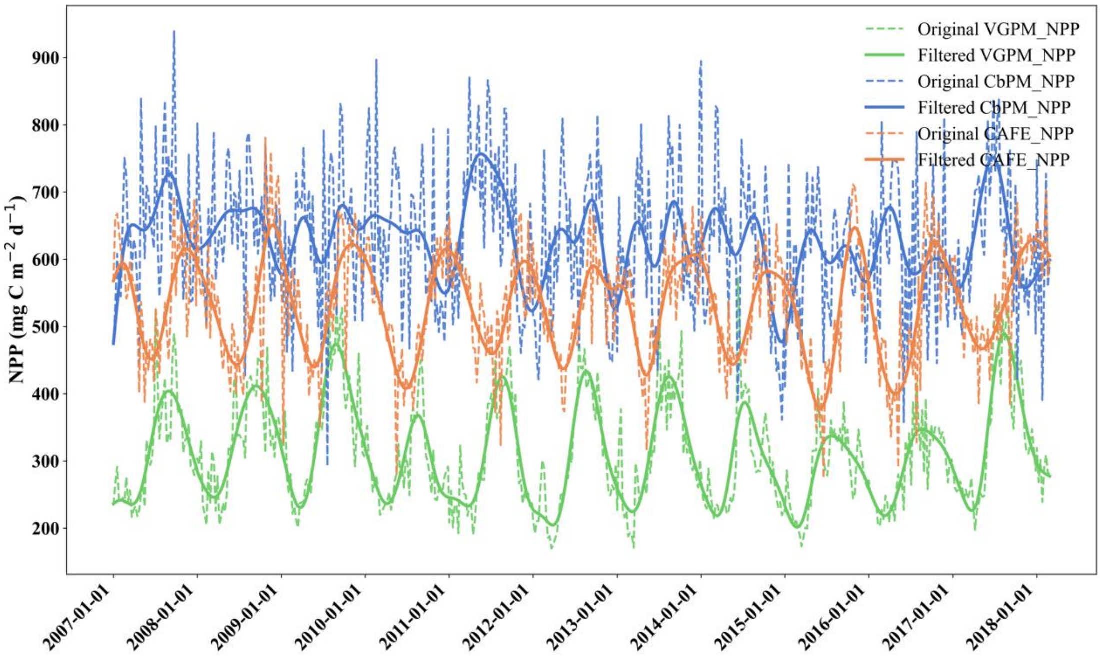

Fig. 3. Time series of VGPM, CbPM, and CAFE from January 2007 to February 2018, where the green line represents VGPM, the blue line represents CbPM, and the orange line represents CAFE . The dashed lines are the original data and the solid ones are the low-pass filtered, which show the seasonal variations more clearly. Abbreviations and data sources can be referenced in Table 1.

17. Line 478-481: Any previous studies (reference papers) that can support the statement about Bayesian model performing better in estimating uncertainty?

Thank you for the comment. In the section introducing the Bayesian method (Lines 270 – 275), we have added citations to relevant literature to support the statement about the Bayesian model's superior performance in estimating uncertainty (lines 274– 275).

"Bayesian models can adeptly quantify the uncertainty in the distribution of predicted outcomes. The Bayesian approach is particularly advantageous in scenarios with limited training data or when potential invisibility in training data cannot be discounted in practical applications (Perfors et al, 2011; Kaplan D, 2021; Zou et al, 2024)."

18. Line 483: What formula did the authors use to estimate the CDFs?

In the revised manuscript, we have added a detailed explanation of the formula used to estimate CDFs (Lines 396 - 407).

"The Cumulative Distribution Function (CDF), also known as the distribution function, is the integral of probability density function (PDF). It provides a complete description of the probability distribution of a real-valued random variable X. The CDF is defined as the probability P that a random variable X is less than or equal to a given value x, expressed as:

$$F(x) = P(X \leq x)$$

To evaluate the predictive performance of the model, we computed the empirical CDF of the input data and compared it with the average predictive CDF generated by the model. This comparison provides a graphical representation of the model's predictive accuracy. A higher degree of overlap between the empirical and predictive CDF curves indicates a greater similarity between the two distributions, thereby reflecting superior model predictions."

19. Line 486-487: As mentioned in the previous comment, the estimation of Train mean NPP and CAFE NPP curves are not clearly mentioned.

Thank you for highlighting this point. In the revised manuscript, we have clarified that since our models generate probabilistic predictions, the curves presented in some figures represent the mean of these predictions. This clarification has been added in Section 2.3 to ensure transparency regarding the methodology and interpretation of the results (Lines 355 - 360).

"In this study, our models provide probabilistic predictions, generating a probability distribution for each time point rather than a single point estimate. To facilitate visualization and interpretation, the curves presented in some figures represent the mean values derived from these predictive distributions. These mean curves summarize the central tendency of the model outputs while inherently accounting for the uncertainty associated with the predictions.”

20. Line 505 “Small” should be replaced by “lower values” for more clarity.

Corrected. (Line 625)

21. Line 509-515: Test mean NPP lying below at lower values and the alteration at higher values is not appearing very significantly. Also, test mean NPP seems to over-estimate at mid-range but this is not the same as seen in the scatter plot (Fig. 6) where it is almost evenly distributed across either side of the 1:1 line.

We appreciate the reviewer’s detailed observation. In response, we have revised the text to provide a clearer explanation of the observed patterns in the CDF curves and their relationship to the scatter plot (Fig. 6). Additionally, we have clarified the interpretation of the differences between the predicted and true value CDFs and provided insights into potential reasons for these discrepancies (Lines 623 - 635).

“As CAFE increases, the difference between the predicted and true CDF curves grows larger, with the predicted mean CDF on the training set generally lying below the CAFE CDF. The difference between the two ranges from 0 - 0.2. For the test set, the predicted mean CDF initially slightly lies below the true CDF curve at lower values, becomes steeper and overestimates at mid-range, and alternates again at higher values. While these trends suggest some instability in the model’s predictions for higher values, the absolute difference between the two CDFs remains within 0.1, indicating limited deviation. It is worth noting that the scatter plot in Fig. 6 shows the test mean NPP predictions distributed more evenly around the 1:1 line. This apparent discrepancy arises from the differing perspectives of the two plots: the CDF curve highlights cumulative differences across the distribution, whereas the scatter plot reflects point-wise deviations. Together, these visualizations suggest that while the model captures the overall distribution trends well, some localized errors in predicting mid-range and higher values may contribute to these patterns."

22. Fig 10: The curves are difficult to distinguish. Different choice of colours recommended.

We appreciate the reviewer's detailed observation. We have revised the colors in Fig 10 to better present the detailed information clearly. However, the contrast is not significant due to the fact that the predicted means of the two models are closer and the folds in the graph overlap more.

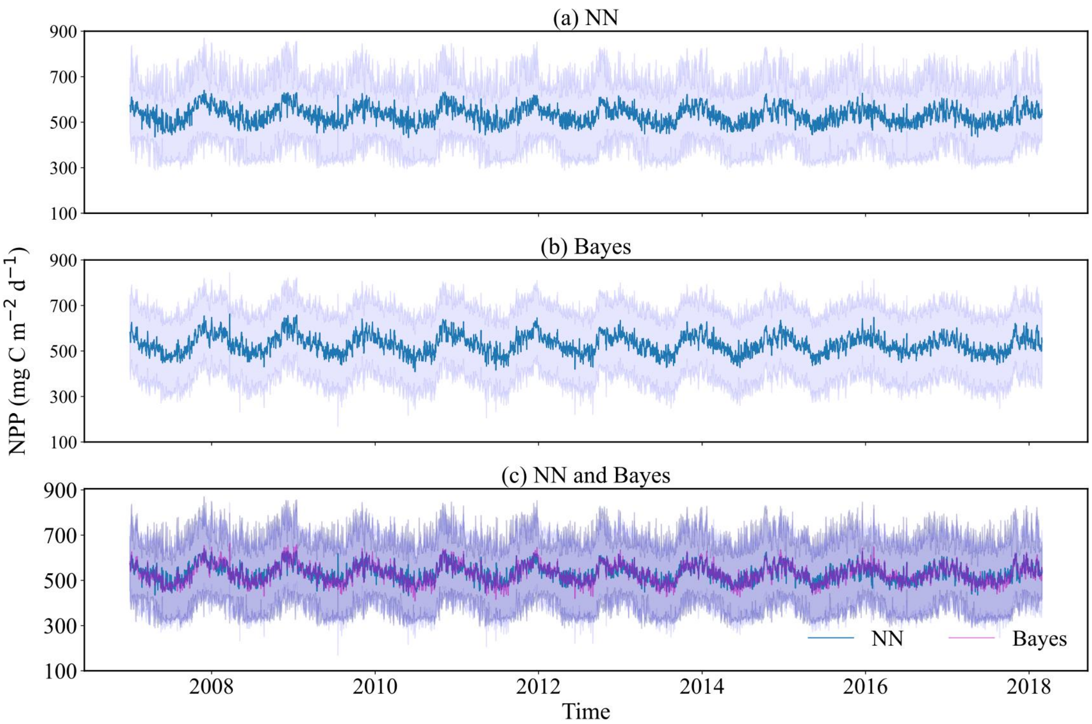

Fig. 10. Time series plots of daily probabilistic NPP predictions in Weizhou Island (2007 – March 2018). (a) Probability prediction results of the neural network model; (b) Bayesian probability prediction results based on empirical distribution; (c) Comparison of the two models' predictions, with the green lines representing the mean predictions from the neural network model and the gray lines depicting the mean predictions from the Bayesian model.

23. Fig 10: Capturing the seasonal cycle is fairly easy as most of the input features contain the same signal. To have a better understanding about how good the models are in reproducing the extreme values, authors should plot the anomaly time series by removing seasonal signals overlayed with observation treated in the same way.

Thank you for highlighting this point. We have drawn anomaly time series plot with seasonal signals removed (Figs. S7 and S8), and compared the ability of two probability prediction models to reproduce extreme values.

"To better understand the model's ability to reproduce extreme values, this article removed the seasonal signals from the original CAFE values and the predicted means of the two probabilistic prediction models and plotted the abnormal time series graphs (Figs. S7 and S8). From Fig. S7, it can be seen that the NN predicted mean values overlap more with the original values, better reflecting the fluctuation size of the original CAFE values, and is superior to Bayes in reproducing extreme values. Fig. S8 compares the prediction means of NN and Bayes when removing seasonal signals. As can be seen from the figure, when the models are applied to the NPP forecast from 2007 to March 2018, the average predictions of the two models are mostly close, but the NN output results fluctuate more significantly, better reflecting the complexity of the actual data."

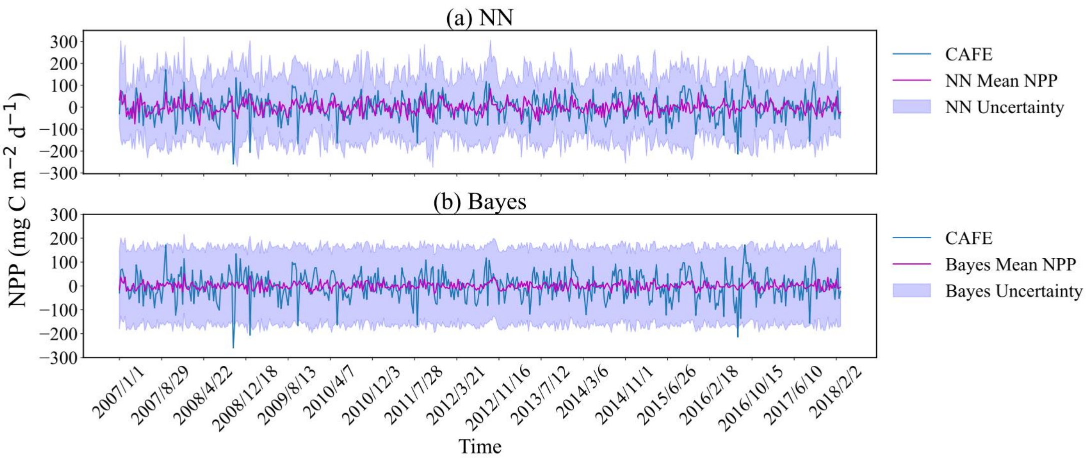

Fig. S7. Comparison of CAFE and predicted mean values shown at an 8-day temporal resolution within a 95% confidence interval. In this case, the seasonal signals have been removed from the original data and the predicted mean values to form an anomalous time series.

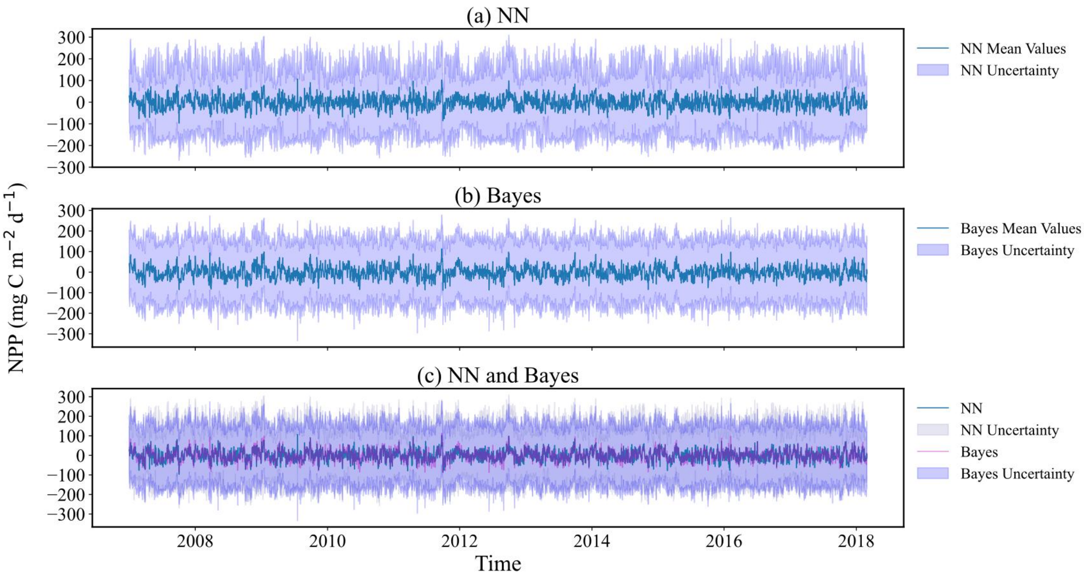

Fig. S8. Time series plots of daily probabilistic NPP predictions in Weizhou Island (2007 – March 2018). In this case, the seasonal signals have been removed from the predicted mean values to form an anomalous time series.