Reply to CEC:

The Erken data we used in this study is stored under same DOI:10.5281/zenodo.7149563 as code, in the 'training data' folder. Here we also provide the model forcing data in the format used in the machine learning models. Data collected by the Erken laboratory, in the archived format used by the Swedish Infrastructure for Ecosystem Science (SITES) is available from the SITES data archive https://data.fieldsites.se/portal/

We have included the GPLv3 License within the latest Zenodo repository and fixed the hyperlinks in the manuscript.

Reply to RC1:

1. The authors mentioned some machine learning models in the Introduction. These are important in the development of algal bloom prediction. The authors should analyze the disadvantages of these ML models and the improvements of their own model.

   I have added some sentences in Line 37-46 that summarizes the disadvantage of the present ML models, and their limitations in water quality and plankton dynamic prediction. The following paragraph illustrates the improvements we made in our models.

2. The literature review of ML models is too simple, which makes it difficult to find the development of models.

   As I replied to the last comment, I have added the information about the development and application of ML models in water quality and algal bloom prediction (Line 37-46).

3. Page3, Line 73. The significance of designing three workflows needs to be further clarified.

   The significance of designing three workflows has been clarified in the Introduction (Lines 47-53), and the details about the workflow were illustrated in the 2.4 section.

4. Page 3, Line 80. Why do the authors use GBR and LSTMï¼Ÿ

   The characteristics and benefits of these two ML has been added (2.3.2; Line 89-105)

5. The advantage of two-step method is accurate prediction when observations are insufficient. However, workflow 1 performs better than workflow 2 or 3 (Table 1). From this comparison, the two-step method is not an important step that affects the accuracy.

   Yes. But workflow 1 can only predict Chl concentration when lake nutrients observations are available, which could be infrequent in most of lakes, and it is also hard to apply this workflow in the algal bloom forecast due to lack of water quality forecast. The advantage of workflows 2 and 3 is therefore a wider potential range of application (e.g., interpolation, reconstruct historical data, algal bloom forecast) at only a minor decrease in performance for this particular lake. The advantages of workflows 2 and 3 were illustrated in the discussion (4.1 Performance of ML models)

6. From Fig. 3 (e.g., Kappa scores), the PB model also works well. What is the advantage of ML models?

The advantages of ML models were revealed by the higher TPR. Although PB works well in terms of Kappa scores, it means PB model can correctly predict chlorophyll concentrations during most of the no-algal bloom period. However, our goal is to capture the algal bloom onset as skillful as we can. ML models have a clear advantage in terms of predicting algal blooms event, which only happened in relatively small proportion of time (Fig. 3a, b).

Reply to RC2

Major remarks

1. Possible overfitting

The authors discuss the potential overfitting issues a few times. L259, "there was overfitting issues in all three workflows, in both GBR and LSTM models, indicated by higher MAE and RMSE in the testing dataset compared to the training dataset especially for GBR". This statement is not completely accurate: higher error in the testing dataset does not immediately imply overfitting. In particular, the authors discuss the peculiarities of the algal bloom in July-August 2019, which is not properly predicted by any model; as this occurrence is in the testing phase, it means that the errors are expected to be large anyway.

The overfitting issue means the model is too closely aligned to the training dataset so that it can not predict the peculiarity in the testing dataset. We think this is the issue existing in the three workflows we tested here. The algal dynamics varied from year to year (Fig. S3). If the models show relatively lower RMSE and MAE consistently in the training dataset which include the data from 2004-2016, than testing dataset, overfitting is a likely explanation.

However, we also agree with your point that peculiarities of the algal bloom in 2019 could also contribute to consistent high errors in testing dataset since the most of observed data points were way higher than usual values. Thus, we adjusted our explanations about the higher RMSE and MAE in testing datasets (Lines 172-175, 189-190, 304-307).

On the other hand, overfitting issues may effectively exist in the model application. In fact, Text S2 reports on the hyperparameters of the LSTM model: by adopting 3 layers and 100 neurons, it approximately implies 300 degrees of freedom (parameters). No information is provided for GBR (please add it).

We added some information about GBR model in Line 93-97, "The hyperparameters in GBR are optimized via *RandomizedSearchCV* function within Scikit-Learn library. The loss function of model is chosen as '*huber*', which is a combination of the squared error and absolute error of regression. Since the target variable in our research Chl concentration has peak values during algal blooms which could be regarded as outliers, the '*huber*' loss function is more robust and gives greater weight to peak values than the mean squared error function."

How does the high number of parameters to be calibrated in the ML model compare with the number of available data? If the number of data is not large enough, the model is intrinsically prone to overparameterization. Did the authors test different hyperparameters for LSTM, e.g. smaller number of neurons and layers?

We have tried different combination of numbers of layers and neurons (1-3 layers, 20-200 neurons), but larger numbers of layers and neurons did not obviously improve the results but increased the computational time a lot, and worse results were achieved when the number of layers and neurons were decreased. We added these details into Supporting Information Text S2.

2. Intrinsic variability in the model's results

The authors analyze the variation of the results obtained in the testing period (2019-2020) when shuffling the training years (section 3.4), and of other possible modifications of the dataset, e.g., by artificially reducing the frequency of the data. As I already mentioned, this is very important, and the analysis is well conceived. Nevertheless, single realizations of ML models may provide non-optimal results. For this reason, it is a common practice to repeat ML runs several times and then average the results (e.g., Piotrowski et al., 2021; Yousefi and Toffolon, 2022). Did the authors account for this?

Yes. That's the reason we conducted these two shuffling year tests. As we mentioned in Line 335-338, "We suggest that testing strategies similar to the shuffle methods used in this study are needed to accurately evaluate the expected accuracy of ML models when applied to any given site. The estimated uncertainty in shuffling training year tests (Fig. 4) and shuffling training/testing year tests (Fig. 5) can be used to better represent the uncertainty of ML derived forecasts."

Minor remarks and typos

1. The Supporting Information contains some data and plots that would fit well in the main text. For instance, Table S1 is useful to understand the procedure used in the analysis.

   We have moved the table from SI to the main text (Table 1).

2. "Even the LSTM algorithms could not account for previous condition so far back in time". How long is the expected memory of the model?

   L270: The expected memory to consider the formation and deposition of cyanobacteria akinetes may require couples of months extending to the previous ice-free season (Lines 304-305).

3. L56 "beings": begins

4. L136-137: "modified Kappa" is not a common metric. Please give a short description of what it represents.

   L136-137: modified accuracy (Kappa) which considers the possibility of the agreement occurring by chance (Table S2; McHugh, 2012)

5. L228 "even though the GBR model usually performs better in Fig. 5c the testing period chosen for use in Fig. 3, showed the opposite result." Cumbersome sentence, please rephrase it. Moreover, the whole section contains a weird use of commas.

   L228: Consequently, even though the GBR model usually performs better in most of chosen 4-year testing periods (Fig. 5), Fig. 3, which shows the results of 2017-2020 testing period, presented the opposite result.

6. Figure 4. The subplots (b) and (c) are not described in the caption.

   The caption has been modified.

7. Text S3, reference to Wilson et al. (2020): it is not in the bibliography.

   The bibliography has been updated.

8. Figure S3, specify that boxplots refer to the period 2004-2020. Maybe it would be more interesting to compute the boxplots excluding 2019-2020?

   This was a typo. The boxplots refer to the period 2004-2018.

9. Figure "Penal": Panel:

   Panel (c)