Response to reviewers RC1 and RC2

Manuscript number: egusphere-2025-3290
Manuscript title: Prediction of present and future spatial occurrence of cyanobacteria and the toxin nodularin in the Baltic Sea

Authors: Mohanad Abdelgadir, Bengt Karlson, Elin Dahlgren, and Malin Olofsson

Principal Investigator: Mohanad Abdelgadir
Preprint link: https://doi.org/10.5194/egusphere-2025-3290

Report #1

Second Review of the manuscript "Prediction of present and future spatial occurrence of cyanobacteria and the toxin nodularin in the Baltic Sea" by Mohanad Abdelgadir, Bengt Karlson, Elin Dahlgren, Malin Olofsson)

Major comments:

The authors did quite some effort to work on the text and to justify their approach (mainly by providing new, generic references). Unfortunately, however, they did not provide new supporting evidence for their results and did not rule out my major concerns. The following points remain:

Re: We greatly appreciate the effort made by the reviewer (RC1) on our manuscript.

(1) I find it still almost impossible to keep overview in the many details (e.g. please avoid many software-package names; these could be acknowledged in the appendix). In this respect the new manuscript hardly improved.

Re: We believe that it's important that all software and packages used to produce the analysis should be appreciated, acknowledged and cited in a proper way given that this study provides a methodological modeling approach. However, we will leave it to the editor to decide if we should move some of the details to Supplementary text.

(2) The authors could not convincingly rule out overfitting. Few results on validation data appear somewhat hidden, are not well explained and are, above all, not convincing.

Re: The issue of overfitting is detailed described in the text and can be read as: *"Algorithms overcome the overfitting by generating what are called "pseudo-absence data points" (Hysen et al., 2022), which represent those unsampled sites across the study area and in a way compare the observed environment (represented by the presence) against what is available. ." see* ***lines 113-115***

The method and procedure for handling overfitting are described in the revised version and can be read as follows:

"Algorithms also avoid overfitting by dividing the data into training and testing sets, which allows the model to learn on one subset of the data (training set) and evaluate its performance on another subset (testing set). All these procedures together ensure that the model generalizes well to new data, making it more robust and reliable." see ***lines 115-118***

Here, the dataset is split into two segments: 70% training and 30% testing data, which allows the model to learn on one subset of the data (training set) and evaluate its performance on another subset (testing set). This ensures that the model generalizes well to new data, making it more robust and reliable. For more information, see below how the model was generated in R code:

```
#Current prediction
#Here I add the pseudo-absence data to the background. I add 10000 pseudo-points as background using "bg" argument
dc <- sdmData(species~., train=newssp, predictors= env.c, bg=list(n=10000))
#let's fit the model by adding different algorithm methods, replication techniques and K-fold cross-validation
mc <- sdm(species ~ . , dc, methods=c('GBM','RF','GLM','MAXENT','MARS','CART'), replication=c("boot"),k=10,test.percent=30)

#Future prediction
#Here, I add the pseudo-absence data to the background. I add 10000 pseudo-points as background using "bg" argument
df <- sdmData(species~., train=newssp, predictors= env.f, bg=list(n=10000))
#let's fit the model by adding different algorithm methods, replication techniques and K-fold cross-validation
mf <- sdm(species ~ . , df, methods=c('GBM','RF','GLM','MAXENT','MARS','CART'), replication=c("boot"),k=10,test.percent=30)
```

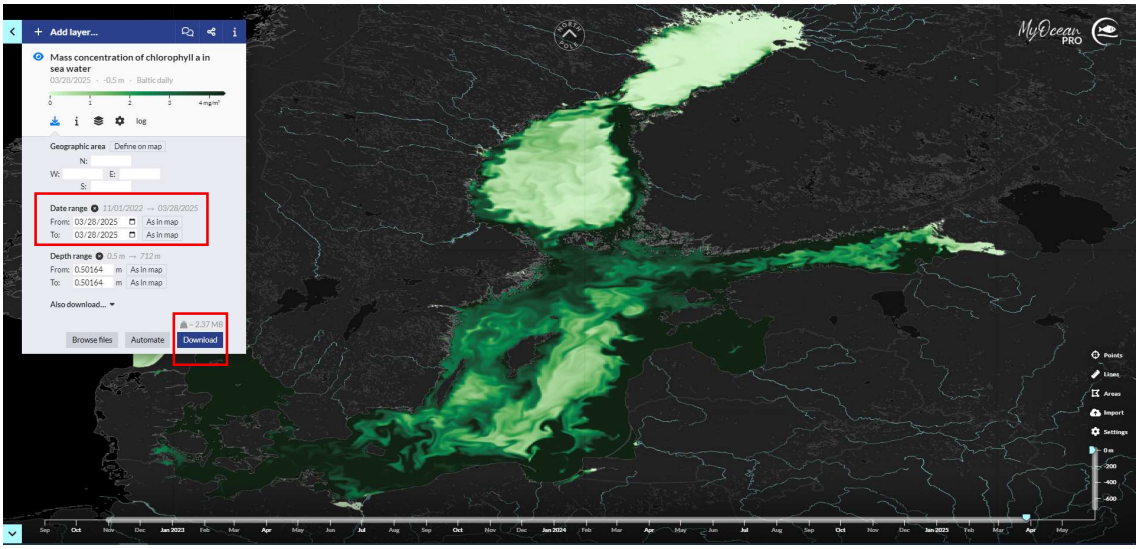
Abbreviations: dc/df= data object (current c/future f), bg=background data object consists of 10000 randomly pseudoabsence points, newssp= data object contains occurrences, species= column of the data object contain the species (Nodularia sp.), methods= set of algorithms used, replication= method technique which is "bootstrapping", test.percent= percentage of data set used for testing which is 30%, k= 10-fold bootstrapping, mc/mf= model setting (current c/future f).

However, in this revised version, we added the overfitting as a cavity in this study. See ***lines 537-538***. The statement can be read as follows: *"The risk of overfitting due to data limitations is still very high, regardless of using ensemble and model accuracy observed in this study."*

(3) The study still combines many different data sources from global to regional models which are almost for certain inconsistent and might make it very difficult to draw robust conclusions. Important information is still lacking (e.g. the time sampling of the predictors remains unclear; how were different salinity products combined?).

Re: The time of sampling of the predictors was already mentioned and clarified in 1st revision under section 2.2 (Data collection and environmental parameters), **lines 173-175.** It can be read as *"....modelled data with a horizontal resolution of 1 nm (nautical mile) and vertical depth layers varying between 1-24 meters were limited to the period of sampling from June to September 2023 at a depth range of 0.5 to 10 meters....."*.

To clarify the method on how the predictors were downloaded, see below screenshot as an example from Copernicus Marine service on how the date of sampling for predictors can be specified from the "Date range" field (marked in red rectangle) and then downloaded into NetCDF-4 format.



There is still limited quality assessment on the predictor data and not much information on the nodularin observations either (e.g. are these data balanced? are the test data independent?).

Re: In this revised version we added two additional analyses testing the independency of the predictors from Copernicus and SMHI. The test of multicollinearity using variance inflation factor (**VIF**: the measure of how much the variance of an estimated regression coefficient increases due to collinearity with other predictors) and correlation matrix between predictors were performed in way to examine the independency of each predictor. All VIFs were below the common threshold of 5 and all pairwise correlations are typically low ($|r| < 0.3$), both showing that there is no indication of severe multicollinearity and most variable-pairs vary independently.

Both analyses were described in this revised version under section **2.3 and 3.1** in ***lines 211-216 and lines 260-268*** *respectively and illustrated in supplementary figures* (see ***Fig S7a&b***).

"2.3. Test of independency and multicollinearity between predictors

To assess the independency of predictors downloaded from Copernicus and SMHI, a variance inflation factor (VIF) for testing multicollinearity and a correlation matrix analysis based on Pearson correlation coefficient (r) were performed. VIF was assessed under the assumption that if VIF < 5: low multicollinearity, 5 ≤ VIF ≤ 10: moderate multicollinearity, and if VIF > 10: high multicollinearity. Furthermore, tests for independency were also performed in two-dimensional principal component analysis (PCA)." ***lines 211-216***

"3.1. Test of independency and multicollinearity between predictors

Tests of multicollinearity show that all VIFs were below the common threshold of 5 (and certainly below 10), showing that there is no indication of severe multicollinearity in Copernicus and SMHI predictors (Fig S7a&b). Test of pairwise correlation show that salinity, PO4 and temperature had highest positive correlation (r ≈ 0.6–0.7). Salinity correlation with temperature reflecting seasonal temperature–salinity structure which could be driven by atmospheric conditions, runoff, and water density properties, while PO4 correlation with temperature typically reflects that higher temperatures accelerate internal phosphorus loading from sediments (Mallissery et al., 2025; Stockmayer and Lehmann, 2023; Walve et al., 2018). On the other hand, all pairwise correlations are typically low (|r| < 0.3), indicating that most variable-pairs vary independently (Fig S7a&b)." ***lines 260-268***

(4) I did still not find convincing evidence for the reliability of the predictions, uncertainties are not convincingly quantified and the plausibility of the obtained results is not assessed. On the contrary, Fig,7a shows nodularin production in response to "current environmental variables" (whatever this means), where to my knowledge nodularia sp. blooms were never observed while the authors indicate "excellent model performance" in the caption. Please note in this context, that predicting nodularia sp. blooms from the environmental factors remains challenging until today - despite many efforts (e.g. Karlson et al., 2008; Kahru et al., 2020).

Re: We acknowledge the reviewer concern. The use of the word "current" in the context is intended to differentiate between the time of sampling of predictors (**current predictors**: downloaded in June-September at the time of toxin sampling) and same set of **future predictors** under SSP5-8.5 scenarios. The rationale of using the words *current vs future* so that the readers can compare.

The model assessment based on AUC values (Area Under Curve) is generally classified into different ranks: AUC >0.9 is outstanding, 0.8–0.9 is excellent, 0.7–0.8 is acceptable, and <0.6 is poor. Since our current and future model predictions showed an AUC between 0.80 and 0.85 which falls under excellent performance.

The challenge of predicting *Nodularia sp.* is now added as cavities in the study; see ***lines 538-539*** and can be read as: *"It is also worth mentioning that, despite many efforts made, predicting Nodularia sp. blooms from the environmental factors remains challenging (Kahru et al., 2020; Karlson et al., 2010)."*

I am afraid that the authors need to address these concerns more carefully until I can recommend publication and I strongly suggest to lower their ambitions which conclusions can be drawn from just a few observational samples. Please don't get me wrong – I really appreciate the nodularin data set and the attempts to draw conclusions from it. However, after all the power of advanced statistics and machine learning depends on a sufficient amount of high-quality training data and expectations what to gain from it, should be realistic. Also, the rationale behind data choices and methods should be clearly explained (including prerequisites).

Re: We acknowledge the reviewer concern. The rationale behind the methods is described in the introduction part ***(lines 60-73)*** , which can be read as:

"There are several factors that may limit the proper spatial estimation of nodularin occurrence across the Baltic Sea. Such factors include sampling difficulty due to spatial variability of nodularin concentrations and the large spatial scale of the area. In addition, seasonal variability of nodularin concentrations poses a sampling challenge due to the short period of time when the toxin nodularin is produced, mainly in the late summer. Moreover, limited, infrequent, and sparse monitoring, especially in the open sea, can lead to underreporting or missing peak toxin periods. The fact that nodularin cannot be directly measured by satellite and remote sensing along with lack of standardized modeling approaches poses great challenges to make predictions about current and future occurrence of nodularin in the Baltic Sea. Additionally, it has been shown that filamentous cyanobacteria over longer time scales not so common in Bothnian Bay (Olofsson et al., 2021) nor in Kattegat (Olofsson et al., 2020b). Furthermore, future prediction of nodularin occurrence under climate change is a modeling challenge, especially in the long-term time frame when the worst-case climate scenario is projected to occur in the Baltic Sea. For instance, climate scenario SSP5-8.5 for 2100 is critical for the Baltic Sea with projected surface temperature increases up to 3.2°C (Meier et al., 2022). This scenario is combined with increased runoff, accelerated oxygen depletion, decreased salinity, and worsened algal blooms (Friedland et al., 2012; Meier, 2006). All these factors, together with lack of proper modeling approach, make the current and future prediction of nodularin a methodological challenge." ***lines 60-73.***

Also Read the paragraphs in ***lines 74-90*** and ***lines 91-111***

Specific comments

Ln 224: AOGCMs typically contain severe biases on regional scales and these need to be addressed for the Baltic Sea region. Apparently, these were not even assessed. To address these issues by a short, general statement is not sufficient.

Re: We are very aware of the AOGCMs biases, however, the (AOGCMs) used in Bio-ORACLE dataset are **downscaled** to provide high-resolution marine data layers suitable for species distribution modeling. The AOGCMs are downscaled using a "change-factor" (or delta) approach to downscale data. This involves applying the projected magnitude of climate change (from AOGCMs) onto high-resolution, present-day baseline data. Please read (Assis et al., 2017 and 2024) at: https://doi.org/10.1111/geb.12693 and https://onlinelibrary.wiley.com/doi/10.1111/geb.13813

To point this in the text we added the words **"downscaled dataset"** in ***line 187***. The sentence can now be read as : *"Raster layers were derived from Atmosphere-Ocean General Circulation Model (AOGCM) downscaled dataset as raster grids at 2.5 arc-minute spatial resolution ..."*

References

Kahru, M., Elmgren, R., Kaiser, J., Wasmund, N., & Savchuk, O. (2020). Cyanobacterial blooms in the Baltic Sea: Correlations with environmental factors. Harmful Algae, 92, 101739.

Karlson, B., Eilola, K., & Hansson, M. (2008). Cyanobacterial blooms in the Baltic Sea–correlating bloom observations with environmental conditions. In Proc 13th Int Conf on Harmful Algae (pp. 247-252).

##########

Report #2

Most of my initial concerns were concerned. I really appreciate the efforts that the authors put into revising this manuscript. I think with some minor revisions the paper is ready to be published. Here are my 3 concerns with the revised version.

Re: We greatly appreciate the effort made by the reviewer (RC2) on our manuscript.

Concern 1. The geostatistical part of the model is not explored in the results. I didnt see the variogram being used either the theoretical or the experimental. The authors talk about the nugget effect in the introduction (line 82). But do not discuss it any further other than mentioning kriging was done in a software (lines 215-218). I would like to see the results of the geostatistical model. The nugget, Anisotropy, sill and the variogram being used.

Re: The model variogram and nugget are now provided in supplementary figures **Fig S2.** and shortly described in ***lines 284-285***. The statement can be read as follows: *"These findings can also be described by exploring the variogram and nugget analysis provided in Fig S2."*

Concern 2. Line 112 "Algorithms in ensemble learning can resolve the issues concerning autocorrelation, clustered samples, small datasets, and unsampled sites across the study area." This is not completely true. There are a number of concerns when using ensemble modelling with limited data. The risk of over fitting still exists. Maybe discuss this being one of the caveats to be considered with the dataset being so small. There is some discussion of this in the final paragraph of the discussion section but the risk of overfitting in this model is

particularly very high given the accuracy statistics we are seeing with the crossvalidation. This doesn't in anyway make light of the model framework which I still believe is novel and particularly useful and shows a lot of promise for modelling environmental phenomenon which are constantly plagued by limited observation data.

Re: In this revised version, we added the overfitting as a cavity in this study. See ***lines 537-538***. The statement can be read as follows: *"The risk of overfitting due to data limitations is still very high, regardless of using ensemble and model accuracy observed in this study."*

Concern 3. This could be more of a knowledge gap for me. But most cyanobacterial bloom and hypoxia modelling I have seen use lagged variables for environmental variables because effect of these variables is not immediate on the toxins or blooms. Is there a particular reason that lagged variables weren't considered in this study. If so I would like to see a couple of points discussing them.

Re: We very much appreciate this point of view. The use of lagged environmental variables in this study is not applicable due to limited data and irregular sampling between June and September. The reason for this is that using lagged variables requires continuous data and given that our sampling was sparse and has significant period gaps, it makes it challenging to define meaningful time lags. It is shown that using lags for multiple predictors increases the dimensionality of the model and consequently leads to overfitting (read also ***lines 63-68*** *in the introduction ("...seasonal variability of nodularin concentrations poses a sampling challenge ..... limited, infrequent, and sparse monitoring, especially in the open sea, can lead to underreporting or missing peak toxin periods ...."*). These points of lagged environmental variables were not specifically discussed in our manuscript because of the nature of our data, however caveats and future outlooks in general are discussed in ***lines 534-541***.