**Reply to Reviewer #1**

We thank the reviewer for the helpful comments. Our responses are listed below in black, while the reviewer’s comments are marked in blue. The line numbers are based on the tracked-change version of the manuscript word document.

I was looking forward to reading the manuscript as the topic very much interests me, but unfortunately I struggled. The writing was not as strong as it could be and that made it difficult to not only understand the methods but also evaluate them. For instance, there were a lot of compound sentences, which are hard to read as there is a lot of information being thrown at the reader, quickly. So it took several reads to make sure I understood everything, and sometimes even then I was not sure. Also, the organisation seemed off at times. For instance, sometimes I felt like the information was given to me backwards. That is, when I would get later on in a section, I would read a sentence that would make me go: "Oh! I wish I was able to read this sentence earlier." This extends to how some of the sections were broken up. Specifically, I do not understand why section 3.0 was used. I think it would be much easier to read through with all of the information in Sec. 3.0 moved to Sec. 2.2. I really encourage the authors to go over this with very critical eyes and ears to improve the writing. I tried to provide some specific suggestions in the comments below but it got harder to be very through as I continued, so re-writing should extend beyond these specific comments.

Re-organization of sections and subsections, particularly methods and results.

By the time I finished my first read, I had the impression that the manuscript was rushed into review. I especially got this after reading the acknowledgements and seeing the project numbers identified with "xxxxxxx". I was a bit surprised and frustrated seeing this. I was left with the impression that the effort to ensure the document was a completed draft upon submission was not made.

Will add.

I have a bit of a disagreement with the interpretation of the results. I understand the metric of "Total Accuracy" was used to select the best model, but there are some risks with that. Just using one metric that is. There is a growing literature behind the need to consider multiple metrics in decision making. Maybe not in machine learning, but in other aspects of ocean research. Looking at all three of the BCV results, I would argue that SVM-RVM - not SVM - preformed the best. I go into more detail about this below in specific comments. I am not saying that the authors need to change their results because of course this is all up for debate. But there should be something about in the discussion.

To evaluate the prediction of a machine learning classifier, one either gets right or wrong. The accuracy measures the percentage when the classifier predicted correctly. Maybe add a table for all classifiers and include accuracy for both HABs and non-HABs events.

Speaking of the discussion, I do not think it gave enough. Some parts seemed more appropriate for the introduction. For example, after the introduction I was under the impression this type of work has not been pursued before for the study area, so I was surprised when similar examples were brought up in the discussion.

Move the text on machine learning studies of HABs to introduction.

So, I encourage the authors to spend more time with the discussion to determine the information that is important to bring up to the reader earlier (introduction), and build up the discussion a bit more based of what was learned from this study.

Good suggestion worthy of considering.

Also, I do not recall reading anything about problems / challenges with the methodology of the work. For example, using one metric to determine a best model. Or, weaknesses of the modelling approaches. It is really important for models to be open and transparent about modelling weaknesses / issues / problems / etc., or else it appears like we have something to hide. So, it is important - necessary even - to incorporate some honest criticisms of the methodology. This of course opens the door to discuss improvements for the future.   
  
Also good suggestions. Add discussions on the limitations of your approach and results.

Some Specific Comments:  
  
The title reads as if machine learning can confidently predict blooms. Although the highest total accuracy being 62% is not bad, I do not consider this to be confident. The title should be re-worked to focus on what was done. For example: Machine Learning Classification Algorithms For Predicting Karenia brevis Blooms on the West Florida Shelf.

Good suggestion. Follow it.

A graphical abstract should interpretable upon the first glance, and I am not sure what I should be taking in from it. Evan after reading the paper, the graphical abstract does not tell me what was done with machine learning that the focus was red tides. The graphical abstract needs to be re-worked or removed if it is not required.

Think about it and ask Pat.

Highlights: In the first bullet, it would be better if you just stick to the facts and state the rate of accuracy and remove the text "high accuracy". As it is currently written it gives me the impression that the results are being spinned up. In the discussion you can then argue how the prediction strength is relatively high (i.e., compare to other studies).

Are there any other studies for comparison?  
  
Abstract; Lacking a bit, should be more references to the methodology. Where did the data come from? How were models evaluated? Etc.

Take a look.

pg. 1, line 21: I believe you need to add a comma after "magnitude".

Done. Please see line 21.

pg. 1, lines 25-28; \*\* A lot is happening in this sentence. break it down.

Rewrote the sentence. Please see lines 25-29.

pg. 1, line 30; the SVM model was used, just the SVM model?

Should do the sensitivity runs using RVM and check if the results are similar. Add the results to the text (maybe supplemental figures) and mention it in the abstract.

pg. 1, line 30; "to show that" makes it sound like this was already previously theorized. Maybe just used "showed".

Corrected. Please see line 31.

pg. 1, abstract; \*\* the last 2/3 of the abstract have a lot of compound sentences. Break things down. It would be much easier to the reader the follow and reduce the chance of confusion.

Rewrote the sentences as suggested. Please see lines 31-35.

pg. 2, line 43; \*\* Break down this first sentence - too much information crammed into one sentence.

Dr. Glibert, please help.

pg. 2, line 48; you could use the acronym WFS for West Florida Shelf throughout the rest of the manuscript to reduce a little text.

Done as suggested. Please see line 63.

pg. 2, lines 52-53; delete "a region about the length of the state of New Jersey" - it is not really necessary as you say the actual value in the following text, and also this sentence will not mean anything to someone who is not familiar with the relative sizes of the US states.

Delete the words. Please see lines 66-67.

pg. 2, line 62; comma after "pathways"

Added. Please see line 76.

pg. 3, line 66; replace "grazing, suggesting" with "grazing - suggesting"

Edited as suggested. Please see line 80.

pg. 3, lines 69-70; \*\* re-work the sentence

Edited the sentence. Please lines 83-85.

pg. 3, line 74; replace "recent" with "2017-2019". It helps the reader to be very specific even if it seems repetitive.

Replaced as suggested. Please line 88.

pg. 3, lines 74-77; \*\* rework sentence.

Rewrote the sentence. Please lines 89-91.

pg. 3, line 86; why is it necessary to model long term trends?

Need to add a motivation.

pg. 3, line 86; Start a new paragraph when discussing what is being done for this research. It helps the reader.

Done as suggested.

pg. 4, line 96; "the database" is confusing - what database? Does it have a name? Or should it just be "the Florida Fish and Wildlife Conservation Commission database"?

Edited the sentence as suggested. Please see line 112.

pg. 4, Section 2.1.1; Is there a reason why this data is not mapped on Fig. 1? Are there restrictions on how to present the data, like no sharing the coordinate data? One way I got around that is by presenting the density distributions.

Dr. Glibert, can we show sampling stations on WFS and data density distribution?

pg. 4, line 95; when you present data you should immediately state the units. What are the density units in the dataset you used?

Added the unit. Please see lines 111.

pg. 4, line 97; insert "water" in front of samples. The more little details just help the reader.

Added “water”. Please see line 113.

pg 4, lines 99-100; In my experience it is common and often recommended to round to the hundreds place. I suggest rounding the values to 25.85 and 29.14.

Rounded the digits. Please see lines 115-116.

pg. 4; line 100; You should make it clear you made restrictions to the data. It may be obvious to you, but it just helps the reader. Why did you make latitude restrictions? Why cut off at off at 9 km?

Added a sentence for justification. Please see lines 117-118.

pg. 4, line 101; I do not think this should be a new paragraph. I think the whole section is one paragraph.

Combined into one paragraph. Please see line 118.  
  
pg. 4, line 106; insert "the" in between "in" and "major"

Added the word. Please see line 124.

Section 2.1.2; Fig. 1 is not referenced in the first paragraph and it should be as USGS data are indicated on Fig. 1.

Added the reference. Please see line 129.

pg. 4, line 101; insert a comma after "(USGS 2298830)" and "and".

Added the comma. Please see line 129.

pg. 5, line 111; you should say exactly what "nutrient data" you got.

Added the nutrient info. Please see line 129.

pg. 5, line 113; The sentence "and were combined with USGS streamflow data to estimate ..." is going to be really confusing to readers unfamiliar with this methodology. You estimated something? How? What method? Is this common practice?

Added a sentence to explain the calculations of the total TN and TP from river flows and nutrient concentration. Please see lines 131-132.

pg. 5, line 115; replace "over" with "across the"

Changed as suggested. Please see line 135.

pg. 5, line 116; Did you use the website to calculate the weekly averages? What is the method used? References (other than a website)?

No, we obtained the original hourly wind data from NDBC. We then used these wind data to calculate the weekly averaged wind speeds using a vector average. Please see lines 135-137.

Fig. 1; The figure description can be improved. It should be readable independent from the text, so I would redefine the acronyms in the figure at least. And you only highlight the region where the intense bloom was but you have bloom data from other years. Did that span the same region? Different? It is easy for a non-familiar reader to get confused here.   
  
pg. 5, line 121; The three algorithms should be specified right away in this section. For instance: "To hindcast K. brevis cell density and test the strength of various explanatory variables, three machine learning algorithms were used: i) Support Vector Machine (SVM), ii) Naive Bayes (NB), and iii) Artificial Neural Network (ANN)."

Rewrote the sentence as suggested. Please see lines 142-146.

pg. 5, line 127; I do not think this should be a new paragraph.

The two paragraphs were combined as suggested. Please see line 151.

pg. 5, line 133; is there a reason you used k = 10? This is not a criticism - just curious if you had justification or just needed to select a number. In either case, you should specify just in case the reader has a similar question.

Add a sentence to explain why you chose k=10.

pg. 5, Section 2.2; Why is there a section 3.0? It would make more sense to move everything in section 3.0 to 2.2.   
  
pg. 6, line 143; I was anticipating a reference for block cross-validation after reading "the data herein were further validated by block cross-validation".

Reference added. Please see line 165.

Section 2.3; Why do both methods of cross-validation when time series data violates the k-fold cross-validation method?

The k-fold cross-validation method assumes independence among data, and this assumption may be invalid in certain time series which show strong autocorrelation. We gave an example of such time series. This is why we also used the blocked cross-validation method to further test the algorithms. Nevertheless, the k-fold method has been widely used to test machine learning algorithms. We would like to keep it. We have reworded the sentences. Please see lines 162-168.

pg. 6, line 156; The acronym SVM has yet to be defined in-text. It has been defined in the abstract but it needs to be defined again in-text.

Moved the descriptions of machine learning algorithms to section 2.2. SVM is defined by now.

pg. 6, line 156; Why SVM? At this point the reader does not know that SVM gave some of the better results so it is confusing. Be general.

Replaced SVM with the machine learning algorithm. Please see line 181.

pg. 6, beginning of section 2.4; I do not know anything about Platt Scaling Analysis so I am missing some sort of brief yet organised description about what it is and what it is doing, and it should be one of the first things done in the section - before the description of the calculation.

We have reorganized and rewrote the sentences. Please see lines 181-188.

pg. 7, line 162; what about sea surface height? Isnt that one you are using also? (pg. 5, line 118).

Have you done this? We found that the model results are not sensitive to sea surface height difference. We have added a note in the discussion section. Please see lines 352-353.  
  
Equations; equation #s should be right justified - easier for the reader to find equations

Done. Thank you for the suggestions.

Equations; Some equations lacking proper formula descriptions. After a formula, all the parameters and variables need to be properly described. For example, there is nothing following equations (2) and (3).

A lot of problems with equations.

pg. 8, line 195; so isnt there technically 4, not 3, models? Arnt SVM and SVM-RVM two different models?

Yes. We now describe 4 models.

pg. 9, lines 201 and 202; Two different spellings. You need to be consistent throughout the document.

Corrected the spelling. Please see line 247.

pg. 10, line 236; error in equation references.

pg. 11, line 240; Again, four different models results are reported - not three.

Corrected. Please see line 289.

pg. 11, section 4.1; I disagree with the interpretation of the results in that it appears SVM-RVM gives the more robust results - not SVM. Yes, SVM has the best non-HAB accuracy but its HAB accuracy is only 0.38. Thus, nearly 60% of the time SVM will fail to predict a HAB. Isn't predicting a HAB the point? Do we not want to predict HAB events in order to be proactive? On that note, one could argue that NB gave the better results in predicting HABs, but its non-HAB prediction is 0.47. So nearly 50% of the time it incorrectly predicts a HAB event. Meanwhile, SVM-RVM is pretty evenly matched between predicting HABs and non-HABs. I understand that the "Total Accuracy" metric was used to determine the better model, but SVM, SVM-RVM, and ANN all gave really close Total Accuracy metrics. I think this all highlights the risks in making decisions with a single metric. More thought is required and decisions need to be based on multiple metrics/indicators - not just one.

Good points. Perhaps select RVM as the model for the sensitivity analysis. Calculate recall and precision too, compare the four models and then select.

pg. 11, line 243; Fig. 3a shows results from SVM-RVM - not SVM - according to the figure.

Need to check.

Section 4.1; you only discuss BCV results so, again, why present k-fold CV results?

Need to decide whether to keep k-fold in Table 1. If so, mention the results in text.

Section 4.1; I have the urge to look at the results for all the models, not just what is presented in Fig. 3. Is it possible to have an appendix with additional results so the reader can view those as well?

Worth considering this suggestion and add more results from the other models. Depends on time. Do Figure 3 for all four algorithms.  
  
pg. 12, line 270; "discharge was varied by 1-2 ...", shouldnt this be in methods?

Good point. We have added a sentence in the methods section. Please see line 192-193. We have deleted similar sentences in the results section. Please see lines 307-309 and 320-321.

Fig. 5; I do not remember reading why Suwanee would not be included. Did I miss it or was the text missing?

Suwanee River flow data were available but nutrient concentrations were not available. ?  
  
pg. 13, line 284-286; I think this would be a useful sentence in the methods - not results- to provide clarity to the reader.

Do not understand the reviewer’s point.

Fig. 6; why are there only these rivers shown? It would be more compelling if more river-to-river comparisons are made. With just the two I almost get the impression that something is being masked. I am not saying that is what the authors are doing - just being honest about the impression I am being given.

Can add more figures for different river combinations as supplementals. Maybe argue these are most sensitive and Figure 5 already show individual effects.

pg. 14, line 316; These studies should have been mentioned in the introduction. I was under the impression that this has not been done before.

We have now moved the descriptions of these previous studies in the introduction section. Please see lines 102-115.

pg. 14, line 326; "the model" - which one, be specific. The reader needs help keeping the information straight.   
  
Acknowledgements; Acronyms need to be defined and grant numbers need to be added.   
  
Grant numbers are listed in the acknowledgements. The contribution numbers will be requested from UMCES and NOAA ECOHAB program once the manuscript is accepted for publication.

**Reply to Reviewer #2**

We thank the reviewer for the helpful comments. Our responses are listed below in black, while the reviewer’s comments are marked in blue. The line numbers are based on the tracked-change version of the manuscript word document.

Summary and relevance of the study  
  
This manuscript optimizes and compares four classification models (Support Vector Machine, a Relevance Vector Machine extension of the SVM, Naïve Bayes classifier, and a feed-forward Artificial Neural Network) for the prediction of toxic blooms of the dinoflagellate Karenia brevis in the West Florida Shelf. The classification models use temperature, wind velocity and direction, and river nutrient discharge as predictor variables, and in-situ K. brevis cell density measurements from the period between 1998 and 2018 as the target variable. The classification model with the highest accuracy for bloom prediction (Support Vector Machine) was used to assess the role of wind velocity and direction on the recurrence frequency of K. brevis blooms, and the role of river nutrient discharge on the maintenance of bloom conditions in the West Florida Shelf using a sensitivity analysis. Strong northerly and westerly winds were found to increase bloom occurrence probability. The mechanism by which wind direction affects bloom occurrence probability is different for northerly and westerly winds. Strong northerly winds result in upwelling in the West Florida Shelf, which transports K. brevis inshore from lower layers of the offshore water column, and westerly winds hold the bloom inshore, once the bloom has reached the shelf. The river nutrient discharge from various rivers is able to sustain K. brevis bloom conditions inshore.   
  
The occurrence frequency of harmful algal blooms has increased both on a global and local scale due to higher nutrient availability (Brand and Compton, 2007; James et al., 2010), which in the context of climate change the increasing trend in occurrence frequency of extreme precipitation around the Gulf of Mexico (Risser and Wehner, 2017; Emanuel, 2017; Oldenborgh et al., 2017) will likely lead to a further increased nutrient availability and phytoplankton activity as demonstrated for other regions (e.g. the Great Barrier Reef; Parker et al., 2017). Thus, the development of an accurate and locally optimized prediction model for toxic blooms in the West Florida Shelf can be seen as a case study or proof of concept for the future implementation of prediction models in other regions and possibly on a global scale. In addition, a study that determines the drivers of toxic K. brevis blooms is worthwhile, as their identification is relevant for informing local policy makers for the implementation of future mitigation and prevention strategies.  
  
General assessment  
  
Based on my review, I recommend the manuscript to go through major revisions. This is because of issues in the motivation and novelty of the study. Additionally, I found issues in the manuscript structure, study design and chosen methodology, which impact the readability, results, and the interpretation of results. The main issue I found throughout the manuscript is the missing motivation behind the development and implementation of a new prediction model. The development of a new prediction model should emerge from the limitations of current existing models. However, the study fails to quantify or assess the (dis)advantages of developing a new prediction model using machine learning algorithms compared to current models, and does not compare the prediction accuracy to current models. Without the comparison to existing models and motivation for development, this study represents only an application of four well established classification models to in-situ measurements instead of a novel, quantifiable, and statistically robust approach. Thus, I encourage the authors to make the reasoning behind the development of the new model a focal point of the study and to highlight the novelty of this approach with its advantages and disadvantages. The issues concerning the structure of the manuscript, study design and methodology are addressed in detail below.

NOAA operational model does not predict K. brevis blooms but instead use observed remote sensing images of chl a and in-situ cell density to predict respiratory impacts on beaches. Wiesburg’s operational model was based on hydrodynamic model only. Check other coupled hydrodynamic-biogeochemical models available for K. brevis.

Comments on the structure of the manuscript  
  
The structure and ow of the current manuscript is overall confusing, and it needs to be revised. In the following sections, issues encountered in each section are presented.  
  
Introduction:  
For the introduction, there is no transition or connection between paragraphs, and following the story line was difficult to follow. I propose the authors to restructure the introduction and guide the reader from the general aspects of harmful algal blooms towards the need for a new improved model. I propose the following structure:

-       General, global description of the term harmful algal blooms.  
-       Introduction of global changes of HABs in recent history.  
-       Description of study region and its relevance.  
-       Description of drivers of K. brevis bloom life-cycle and seasonality in the West Florida Shelf.  
-       Description of recent bloom events in the West Florida Shelf in relation to climate change and population growth.

Leave those to Pat.

-       Description or review of current modeling approaches and their (dis)advantages.

Add a paragraph.

-       Summarized reasoning behind the study and aim of the study.

Add a paragraph.

Methods:  
The description of the methods is split across two separate sections, whose purpose is missing or stated in vague terms. The study aims to compare the predictive power of different models, and thus a link or comparison between the models should be shown. The reasoning behind the choices in the different algorithms, thresholds, predictor variables, data transformation and aggregation should be motivated or discussed in a quantitative way. Additionally, section 3.0 (Calculation) does not seem appropriate, since it only contains the description of established classification models and does not represent a practical development from a theoretical basis" (see Guide for Authors). As for the introduction, I propose the authors to restructure the methods section and provide the reader with crucial information on the motivations behind the chosen algorithms and methods. I propose the following structure:  
1.      Description of biological data,  
2.      Description of physical/biogeochemical data,  
3.      Preparation of target and predictor variables,  
4.      General description of classification algorithms, and specific description for:  
(a)     Support Vector Machine (SVM),  
(b)     Relevance Vector Machine (RVM),  
(c)     Naïve Bayes classifier (NB),  
(d)     Artificial Neural Network (ANN).  
5.      K-fold cross validation scheme,  
6.      Evaluation criteria with contingency tables (accuracy, recall, precision),  
7.      Sensitivity analysis.

Good suggestions and can follow them.

Concerning the reference style, the source of each data product should be indicated using a regular citation and not just an URL. The R packages and version used should be mentioned and referenced. The references for models, equations, and other methods presented are missing. The explanation and equations of the models should be consistent with those presented in the references provided (e.g. equation (4) and its description in the manuscript does not match the equation in Cortes and Vapnik, 1995).  
  
Results:  
The results section does not provide results for all models used. The study was presented as an inter-comparison of different prediction algorithms, and thus the reader expects for all models to be inter-compared with respect to well-defined and quantitative evaluation criteria. Otherwise, the title, abstract, introduction, and methods should be changed and a section detailing the criteria chosen to identify the optimal model should be in the methods section. Additionally, some sentences are misplaced, since they describe methods (e.g. l.257-259, l.270-271) or discuss the results (e.g. l.261- 263, l.265, l.282, l.290-297).

Add figures from other models in the supplemental.

Discussion:  
The current discussion appears to be an extended introduction rather than a discussion of results. There are results mentioned in the discussion that were not presented previously (e.g. l.325-327). The current discussion presents new concepts (e.g. l.316-328) that should have been mentioned in the introduction. In the discussion section, the accuracy, sensitivity, and applicability of the new prediction models should be compared to current existing ones, and the limitations of the new model should be explored. Without such a comparison, the merits of the study cannot be recognized and the reader is unable to perceive the advantages of the new prediction models over current ones. The main motivating factors or assumption for the study depicted in the introduction should re-emerge and either be confirmed or rejected.  
  
Conclusion:  
The conclusion section is missing. This section should present the main take-home message of the study and provide an outlook for future research, which would likely improve the current limitations of this and other studies.

Good point.

Comments on methodological flaws  
  
Overall, I found major flaws in the methods presented in this study. These flaws represent a source of bias in the results and their interpretation. The main issue I identified in the study design is that the reasoning for testing the role of wind speed, wind direction, river discharge, and river nutrient supply on the occurrence frequency of K. brevis appears circular in nature. Since 32 out of 34 predictor variables of the models consisted of the aforementioned variables, any change in these variables will by design result in an effect of the predicted occurrence frequency of blooms. Thus, if a model was fitted to only a couple of predictor variables, a change in the prediction due to changes in the predictor variables cannot be interpreted to mean that they are a driver.

Mis-understanding on the reviewer part. The model is still fit over all predictor variables, like a complex function dependent on many input variables. The analysis is to hold other variables fixed and vary one variable at a time. This sensitivity is done frequently. Maybe add references.

I suggest the authors to provide a robust reasoning for their choice in predictor variables, to select predictor variables that cover different aspects that may influence bloom conditions, and to test the effect of single or combinations of single predictor variables during the training of the model.

Weren’t those were exactly done in the manuscript. Clarifications.

Presentation and preparation of the target variable:

The presentation and preparation of the target variable appears to be awed. In section 2.1.1 the authors aggregate the data into weekly means for the entire study region, based on the top five cell counts. However, as stated in the manuscript, the study region is characterized by a high spatial and temporal heterogeneity, and thus such an aggregation can potentially lead to bloom conditions despite the fact that only five measurements in the study region are above the selected threshold. I suggest the use of a robust measure of aggregation, such as the median and interquartile range, or to use the top 5, 25, 50, 75, and 95% of the cell counts to calculate the weekly means/medians and perform the training and validation with each of these values. Such an approach can provide a confidence interval for the model predictions, which is currently missing and needs to be included.

Provide justifications and cite others doing weekly averages. Blooms are high cell density events. The traditional statistics do not apply here.   
  
Presentation and preparation of predictor variables:  
The spatial and temporal resolution of the predictor variables are not compatible with the aggregated resolution of the target variable. In section 2.1.2 the physical and biogeochemical data is mentioned, but information on the spatio-temporal resolution and units of the data is missing. Such information is crucial, since the study aims to assess the role of different physical/biogeochemical variables on phytoplankton, which are known to exhibit rapid response times to perturbations (Kavanaugh et al., 2016) and are short-lived (Padisák, 1994). Given that the resolution of predictor variables do not match the resolution of the target variable, those predictors with the highest variability will likely dominate the results of the prediction, i.e. one single process on relatively local scale dominates the outcome of the prediction for the entire study region. The spatial resolution of the predictor variables should match the spatial resolution of the target variable. Since the spatial variability of the biological data in the study regions is smoothed into a weekly mean, I suggest to change the predictor variables accordingly to match the smoothed target variable, e.g. use the weekly mean total nutrient input as a single predictor variable instead of the nutrient input of individual rivers, or the resulting vector average wind speed and direction over the study area instead of the individual measurements in different stations divided into the U and V components.

Weekly averages were used as explanatory variables. Clarified in Methods section.

Optimization of hyperparameters:  
The optimization of the hyperparameters is awed and might have resulted in an overfitted model. In section 2.3 the authors test the classification models and optimize the corresponding hyperparameters simultaneously on the validation set in each iteration of the k-fold cross validation. However, this approach has been shown to result in overfitting to the training data (Cawley and Talbot, 2010), and thus diminishes the applicability of the model for subsequent predictions.

But the prediction is based on an independent testing data. More clarifications.

To avoid overfitting, I suggest to use a nested k-fold cross validation approach (Schratz et al., 2019), which optimizes the hyperparameters on the training set in an inner loop, and tests the model on the validation set in an outer loop. After the optimization, the hyperparameters should be reported.

Blocked validation was used to add weights to the k-fold cross validation.

Overcoming class imbalance:  
The implementation of the oversampling approach is awed. For the oversampling approach of the minority class (bloom), the study by (Fernandez et al., 2018) was referenced, which suggests as a best practice for constructing minority class observations to select a random point in the minority class and its k-nearest neighbours to calculate a synthetic minority observation. However, the implementation (see code cross validation.R line 94) is a simple resampling of observations of the minority class, which has been shown to result in overfitting **(Galar et al., 2012).** As was the case for the hyperparameter optimization, this diminishes the applicability of the model for subsequent predictions.

Seems to be a fair point. Worth checking different resampling method.

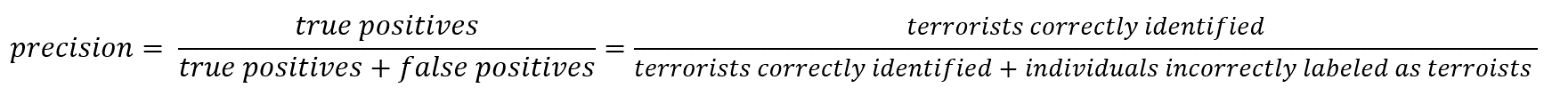
Chosen evaluation criteria:

The chosen evaluation criteria (accuracy) is not an adequate metric to evaluate imbalanced classification problems (**Sun et al., 2009**). As an example using the provided dataset, there are 318 instances of the positive class (bloom) and 755 instances of the negative class (no bloom). A dummy model that predicts the negative class for all possible combination of predictor variables reaches an accuracy (see formula for accuracy in Sun et al., 2009) of roughly 70% since there are zero true positives (TP), zero false positives (FP), 318 false negatives (FN), and 755 true negatives (TN). I recommend to use **precision and recall** as evaluation criteria instead, as these focus on the positive class. The evaluation metric should be calculated for the entire dataset and not separated into weeks with blooms and weeks without blooms. For each optimized model the resulting contingency tables with true/false positives/negatives should be reported together with the number of observations in the majority/minority class.

Worth looking into this. Did resampling already solve this unbalanced problem?

The metric our intuition tells us we should maximize is known in statistics as [**recall**](https://en.wikipedia.org/wiki/Precision_and_recall), or the ability of a model to find all the relevant cases within a dataset. The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives. True positives are data point classified as positive by the model that actually are positive (meaning they are correct), and false negatives are data points the model identifies as negative that actually are positive (incorrect). In the terrorism case, true positives are correctly identified terrorists, and false negatives would be individuals the model labels as *not* terrorists that *actually were* terrorists. Recall can be thought as of a model’s ability to find all the data points of interest in a dataset.

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as positive that are actually negative, or in our example, individuals the model classifies as terrorists that are not. While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.



Implementation of k-fold cross validation:  
The implementation of the k-fold cross validation does not appear to be correct. As implemented in cross validation.R on line 13, ten missing values in the dataset were transformed to the majority class (no bloom). Missing values should always be treated as missing values. During the training and tuning of the four different models in cross validation.R between line 25-29 the data is prepared and normalized for the subsequent training of the models, however, on line 31 this data is overwritten by data stored in alldata.csv, which is not normalized, and does not contain the target variable. Thus, features with the highest magnitude will likely dominate over all other features in the training phase. Additionally, on line 114-121 during the training of the ANN, the predictor variables are specified differently from those for the SVM, RVM, and NB. Since the study compares different models for the prediction of algal blooms, the same predictors should be used in all models, or the differences in predictor variables in different models should be reported and explained.

Overall, reviewer examined the wrong file in the github repository. The cross validation results in this paper can be found in the R files containing the abbreviation “cv”. Mis-understanding by reviewer.

Corrected. Missing data values removed.

Reviewer looked at incorrect R Program. Standardization was applied to “alldata.csv” in other R programs. Final version of programs can now be found here (“final\_version … .R”). Mis-understanding by reviewer.

At the time the research was conducted, the neuralnet package for R did not allow the implicit description of explanatory variables, i.e. cannot write “y~.”, has to write “y~x”. The explanatory variables had to be explicitly provided for the ANN but can be implicitly provided for the SVM, RVM, and NB. New version provides for implicit description. Corrected.

Description of results:  
The description of results lacks quantitative meaning. Statistically significant differences were mentioned throughout the manuscript (l.150, l.287, l.324). However, there was no quantification for the statistical significance of the results, and the methodology used to quantify statistically significantly differences was not provided. The results were reported without standard deviation or confidence interval. Reporting a confidence interval is crucial, as it informs the reader on the robustness of the prediction. A possible way to calculate a confidence interval of the prediction would be to implement a prediction ensemble model (e.g. **Araújo and New, 2007; Bouska et al., 2014; Righetti et al., 2019**) instead of individual models.  
  
Graphical abstract, highlights and minor comments  
  
The graphical abstract does not summarize the findings. It should rather show the different effects of wind, starting from the inshore transport of the cells and the maintenance by westerly winds and nutrient input, which would then summarize how blooms initiated offshore are transported inshore and maintained. The second highlight is misleading, since the study mainly considered wind, river discharge, and nutrient input as predictor variables are only two out of 34 variables were not in this category. Overall, with respect to the language used in the manuscript, I suggest to refrain from using non-standard units of measurement (e.g. "bloom covered a region about the length of the state of New Jersey" instead of quantitative measure in SI-units, l.52-53), avoid using vague (e.g. "in earlier years" instead of giving the exact year/period, l.55; "prolonged duration" instead of providing the duration in days/months, l.56), subjective descriptions (e.g. "clearly" instead of providing evidence or a reference, l.70, l.252), and emotive expressions (e.g. "unrelenting wet weather", l.74-75). These are only a few examples and I encourage the authors to have the language fully examined prior to publication.  
  
References  
  
Araúujo, Miguel B. and Mark New (2007). "Ensemble forecasting of species distributions".  
In: Trends in Ecology Evolution 22.1, pp. 42 -47. ISSN: 0169-5347. DOI: <https://doi.org/10.1016/j.tree.2006.09.010>. URL: <http://www.sciencedirect.com/science/>  
article/pii/S016953470600303X.  
  
Bouska, Kristen L., Gregory W. Whitledge, and Christopher Lant (2014). "Development and evaluation of species distribution models for fourteen native central U.S. fish species". In: Hydrobiologia 747.1, pp. 159-176. DOI: 10.1007/s10750-014-2134-8. URL: https:  
//[doi.org/10.1007%2Fs10750-014-2134-8](http://doi.org/10.1007%2Fs10750-014-2134-8).  
  
Brand, Larry E. and Angela Compton (2007). "Long-term increase in Karenia brevis abundance  
along the Southwest Florida Coast". In: Harmful Algae 6.2, pp. 232-252. DOI: 10.1016/j.hal.2006.08.005. URL: <https://doi.org/10.1016%2Fj.hal.2006.08.005>.  
  
Cawley, Gavin C. and Nicola L.C. Talbot (2010). "On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation". In: J. Mach. Learn. Res. 11, 2079-2107. ISSN: 1532-4435.  
  
Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: Machine Learning 20.3, pp. 273-297. DOI: 10.1007/bf00994018. URL: <https://doi.org/10.1007%2Fbf00994018>.  
  
Emanuel, Kerry (2017). "Assessing the present and future probability of Hurricane Harvey's  
rainfall". In: Proceedings of the National Academy of Sciences 114.48, pp. 12681-12684. DOI: 10.1073/pnas.1716222114. URL: <https://doi.org/10.1073%2Fpnas.1716222114>.  
  
Fernandez, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla (2018).  
"SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the  
15-year Anniversary". In: Journal of Artificial Intelligence Research 61, pp. 863-905.  
DOI: 10.1613/jair.1.11192. URL: <https://doi.org/10.1613%2Fjair.1.11192>.  
  
Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera (2012). "A Review  
on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based  
Approaches". In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Ap-  
plications and Reviews) 42.4, pp. 463-484. DOI: 10.1109/tsmcc.2011.2161285. URL:  
<https://doi.org/10.1109%2Ftsmcc.2011.2161285>.  
  
  
James, K. J., B. Carey, J. O'Halloran, F. N. A. M. van Pelt, and Z. Skrabáková (2010).  
"Shellfish toxicity: human health implications of marine algal toxins". In: Epidemiology  
and Infection 138.7, pp. 927\_940. DOI: 10.1017 / s0950268810000853. URL:  <https://doi.org/10.1017%2Fs0950268810000853>.  
  
Kavanaugh, Maria T, Matthew J Oliver, Francisco P Chavez, Ricardo M Letelier, Frank E Muller-Karger, and Scott C Doney (2016). \Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation". In: ICES Journal of Marine Science: Journal du Conseil 73.7, pp. 1839-1850.  
  
Oldenborgh, Geert Jan van, Karin van der Wiel, Antonia Sebastian, Roop Singh, Julie Arrighi,  
Friederike Otto, Karsten Haustein, Sihan Li, Gabriel Vecchi, and Heidi Cullen (2017). "Attribution of extreme rainfall from Hurricane Harvey, August 2017". In: Environmental Research Letters 12.12, p. 124009. DOI: 10.1088/1748-9326/aa9ef2. URL: <https://doi.org/10.1088%2F1748-9326%2Faa9ef2>.  
  
Padisák, Judit (1994). "Identification of relevant time-scales in non-equilibrium community dynamics: conclusions from phytoplankton surveys". In: New Zealand Journal of Ecology, pp. 169-176.   
  
Parker, Chelsea L., Amanda H. Lynch, Stephanie A. Spera, and Keith R. Spangler (2017).  
"The Relationship between Tropical Cyclone Activity, Nutrient Loading, and Algal Blooms over the Great Barrier Reef". In: Biogeosciences Discussions, pp. 1-35. doi: 10.5194/bg-2017-23. url: <https://doi.org/10.5194%2Fbg-2017-23>.  
  
Righetti, Damiano, Meike Vogt, Nicolas Gruber, Achilleas Psomas, and Niklaus E. Zimmermann (2019). "Global pattern of phytoplankton diversity driven by temperature and environmental variability". In: Science Advances 5.5, eaau6253. DOI: 10.1126/sciadv.aau6253. URL: <https://doi.org/10.1126%2Fsciadv.aau6253>.  
  
Risser, Mark D. and Michael F. Wehner (2017). "Attributable Human-Induced Changes in  
the Likelihood and Magnitude of the Observed Extreme Precipitation during Hurricane  
Harvey". In: Geophysical Research Letters 44.24, pp. 12,457-12,464. DOI: 10.1002 /  
2017gl075888. URL: <https://doi.org/10.1002%2F2017gl075888>.  
  
Schratz, Patrick, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning  
(2019). "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data". In: Ecological Modelling 406, pp. 109-120. DOI: 10.1016/j.ecolmodel.2019.06.002. URL: <https://doi.org/10.1016%2Fj.ecolmodel.2019.06.002>.  
  
Sun, Yanmin, Andrew K. C. Wong, and Mohamed S. Kamel (2009). "Classification of imbalanced data: a review". In: International Journal of Pattern Recognition and Artificial Intelligence 23.04, pp. 687-719. DOI: 10.1142/s0218001409007326. URL: https: //[doi.org/10.1142%2Fs0218001409007326](http://doi.org/10.1142%2Fs0218001409007326).