Remote Sens. **2015**, *7*, 13564-13585; doi:10.3390/rs71013564

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Technical Note

Learning-Based Algal Bloom Event Recognition for Oceanographic Decision Support System Using Remote Sensing Data

Weilong Song [1], John M. Dolan [2], Danelle Cline [3] and Guangming Xiong [1,*]

[1] School of Mechanical Engineering, Beijing Institute of Technology, No. 5 South Zhongguancun Street, Beijing 100081, China; E-Mail: songweilong8896@gmail.com

[2] Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA; E-Mail: jmd@cs.cmu.edu

[3] Monterey Bay Aquarium Research Institute, 7700 Sandholdt Rd., Moss Landing, CA 95039, USA; E-Mail: dcline@mbari.org

***** Author to whom correspondence should be addressed; E-Mail: xiongguangming@bit.edu.cn; Tel.: +86-10-6891-8652; Fax: +86-10-6891-8290.

Academic Editors: Deepak R. Mishra, Eurico J. D'Sa, Sachidananda Mishra, Magaly Koch and Prasad S. Thenkabail

Received: 28 August 2015 / Accepted: 13 October 2015 / Published: 19 October 2015

Abstract: This paper describes the use of machine learning methods to build a decision support system for predicting the distribution of coastal ocean algal blooms based on remote sensing data in Monterey Bay. This system can help scientists obtain prior information in a large ocean region and formulate strategies for deploying robots in the coastal ocean for more detailed *in situ* exploration. The difficulty is that there are insufficient *in situ* data to create a direct statistical machine learning model with satellite data inputs. To solve this problem, we built a Random Forest model using MODIS and MERIS satellite data and applied a threshold filter to balance the training inputs and labels. To build this model, several features of remote sensing satellites were tested to obtain the most suitable features for the system. After building the model, we compared our random forest model with previous trials based on a Support Vector Machine (SVM) using satellite data from 221 days, and our approach performed significantly better. Finally, we used the latest *in situ* data from a September 2014 field experiment to validate our model.

Keywords: remote sensing; machine learning; random forest; Monterey Bay

1. Introduction

The coastal ocean has climate effects and plays a significant role in economies and society. In recent years, anthropogenic inputs have significantly affected organism abundance and community structure. One result of such inputs and their resultant nutrients is algal blooms. Harmful algal blooms (HABs) can not only threaten the diversity of ecosystems (e.g., coral reef communities [1]), but also cause negative economic impacts around the world [2–5]. As a result, detecting and predicting HABs has become a popular research topic.

Significant prior works have used remote sensing images to manually detect algal blooms. Ryan, J.P., *et al.* [6–8] used the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) imagery instrument to explain algal bloom events and discover the impact of upwelling on the plankton ecology in the Monterey Bay coastal domain. Allen *et al.* [9] predicted algal bloom events in Europe using a coupled hydrodynamic ecosystem model based on a comparison of satellite estimates, and Gower *et al.* [10] used the Medium Resolution Imaging Spectrometer (MERIS) satellite sensor to detect peaks in algal bloom response. Li *et al.* [11] provided a framework for monitoring HABs based on satellite remote sensing data.

Correspondingly, several researchers have also applied machine learning (ML) methods to enhance oceanographic experiments. Linear discriminant analysis was used in [12] to detect HABs using multiple remote sensing variables, and Tomlinson *et al.* [13] used ocean color imagery from SeaWiFS to provide an early warning system for HABs in the eastern Gulf of Mexico. These efforts aimed to detect marine events manually and directly from satellite data, processed with a small set of ground truth generated from *in situ* data. However, such efforts do not focus on how to predict such events automatically. SVM regression is used in [14] for predicting chlorophyll from water reflectance data, whereas we are focused on classification. Our own previous work on the current project [15] used chlorophyll-A and fluorescence line height data as inputs to predict ocean events based on an SVM model, but the initial result was less accurate than desired and not validated by field experiments.

In this paper, we use machine learning to predict marine phenomena automatically in the coastal ocean using heterogeneous data sources. The resultant predictions can be used to reposition autonomous underwater vehicles (AUVs). These AUVs are equipped with advanced sensors used to obtain data from physical, biological, chemical, and geological phenomena. In addition, they can gather water samples for offline laboratory analysis, including that of algal blooms. However, the cost of AUV sampling is rather high, especially when searching without guidance over large ocean areas. For example, covering mesoscale (> 50 km^2) regions requires robotic assets to be used for weeks at a time. Prior models used to answer the questions of *where* and *when* to sample may result in significant cost savings via targeted sampling.

Our work is based on the Monterey Bay Aquarium Research Institute (MBARI) Oceanographic Decision Support System (ODSS) [16,17]. The ODSS is an online system designed for MBARI's Controlled Agile and Novel Observation Network (CANON) program that includes situational awareness, planning, data analysis, water management and the STOQS database [18,19]. In addition, the ODSS provides a portal for designing, testing and evaluating algorithms using mobile robots in ocean experiments. The Dorado-class *Gulper* AUV at MBARI has an advanced control system and onboard water sample collection system, which are suitable for targeted sampling. The purpose of the ODSS is to couple human decision-making with probabilistic modeling and learning to build and refine environmental field models. The models in the ODSS can help scientists make better decisions in selecting sampling patterns and schedules for AUVs [20].

In this paper, we focus on how to build an informative model for predicting algal blooms, a typical marine event. Remote sensing images are used as the system inputs to predict algal bloom areas, which can be used in the ODSS for AUVs' adaptive sampling of the ocean and scientists' further surveys of harmful algal bloom events combined with the *in situ* AUV data. This model can additionally be used as an event detection toolbox for CANON during field experiments [15], reducing the number of sampling targets and providing the AUVs' prior knowledge for their adaptive sampling work.

Several challenges exist in the model establishment process. First, there are insufficient *in situ* data to create a direct statistical machine learning (ML) model with satellite data inputs. Second, building a statistical machine learning model requires mass data and developing a preprocessing model to obtain the training inputs automatically is difficult. Third, the training datasets are extremely imbalance (e.g., the positive points are less than 1%), which increases difficulty in classifier selection and evaluation. In face of these challenges, we extend our former efforts [15], test more oceanographic features from satellites and obtain an ML model based on random forests that predicts algal bloom events better than before. Our model is also validated by *in situ* field experimental data. The main contributions of this paper can be summarized as:

(1) Building an informative model using Moderate Resolution Imaging Spectro-radiometer (MODIS) data, Medium Resolution Imaging Spectrometer (MERIS) data and machine learning approaches to predict the distribution of algal blooms that can be used in ODSS for guiding actual field experiments in Monterey Bay.

(2) Developing preprocessing methods to automatically obtain the training inputs of statistical machine learning model using MODIS and MERIS data.

(3) Testing model performance based on remote sensing data, as well as *in situ* data from actual field experiments in Monterey Bay, which proves the effectiveness of our model.

2. The Data

2.1. Study Area

The study area is Monterey Bay, California, and its related ocean region is from 36.33 °N to 38.5 °N and 121.77 °W to 123.5 °W. Monterey Bay is a gulf in the Pacific Ocean that is situated off the coast of California in the U.S. (Figure 1). The bay is located to the south of San Jose and between Monterey and Santa Cruz.

Monterey Bay is situated in the California Current System [21] and is the largest open bay along the western seacoast of the US. In Monterey Bay, harmful algal bloom events have been related to several phenomena, such as wind-driven upwelling [22] and massive mortality of seabirds [23,24]. Therefore, predicting potential algal bloom event areas could be helpful in reducing their harmful effects on the ocean ecology.

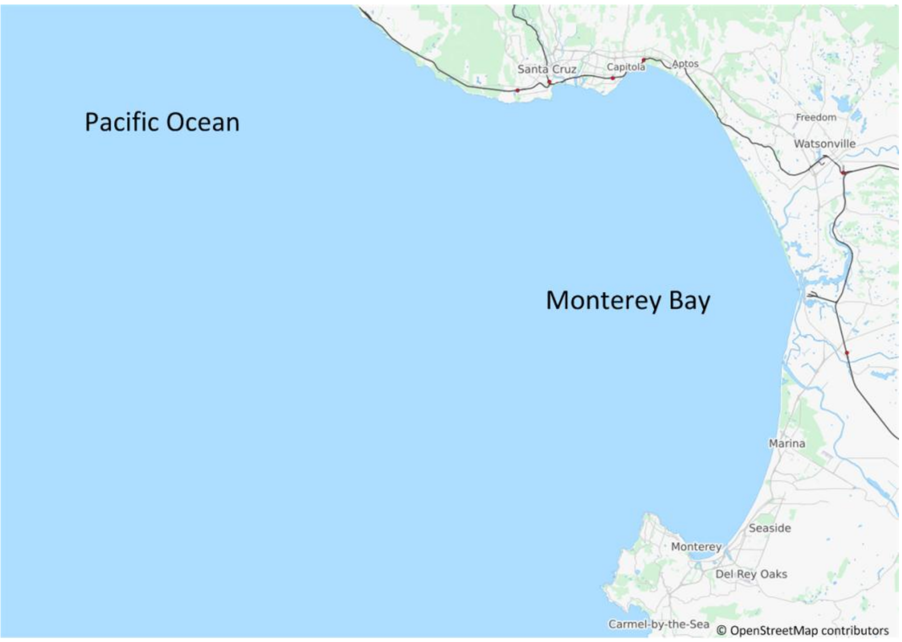


Figure 1. Overview of the study region. Monterey Bay area.

2.2. Satellite Data

Satellite ocean color observations are important for atmospheric correction and the analysis of pigment concentrations [25,26]. MODIS [27], launched in 1999 on the Terra satellite and in 2002 on the Aqua satellite, and MERIS [28], launched in 2002 on the ENVISAT platform, are two popular data sources equipped with phytoplankton fluorescence ocean color bands.

The satellite data sources used in this paper consist of L3 MODIS data and L2 MERIS data. MODIS datasets have different products (e.g., fluorescence line height and chlorophyll_A) and frequencies (e.g., 1 day, 3 day and 8 day), and we use MODIS 1-day composite products, which is the same frequency as MERIS data. MODIS and MERIS data were obtained beginning with October 2010. However, MERIS data after April 2012 are not available due to loss of communication with the satellite, while MODIS data are still available. Therefore, we collected approximately 221 days of imagery between October 2010 and April 2012 to obtain our classification model where data from both MERIS and MODIS satellites are available. MODIS imagery for each day obtained in this work is 321 × 321 pixels, and the resolution of one pixel is 1 km^2. However, the resolution of MERIS data is 0.0754 km^2 with 869 × 693 pixels. To select proper products in MODIS datasets and deal with different resolutions between MODIS and MERIS, a feature selection process is performed to obtain the most reliable model inputs, which will be discussed in detail in Section 3.

The Maximum Chlorophyll Index (MCI) (Figure 2) derived from the MERIS instrument is potentially an excellent indicator of extreme blooms in Monterey Bay [18] because MERIS contains a sensing band at 709 nm, corresponding to a peak in the ocean surface reflectance for red tide blooms. Because ground truth data for blooms are notoriously difficult and/or expensive to acquire, we use the MERIS MCI, which is generally regarded as a good surrogate for bloom presence above a certain threshold, as ground truth in developing our classifier. Because contact with the MERIS satellite has been lost since 8 April 2012, the MERIS MCI can no longer be used as a surrogate for bloom activity, at least for the time being. However, because MODIS data continue to be available, we use historical MERIS and MODIS data to predict the former based on the latter and then use current MODIS data to predict bloom activity in current field trials.

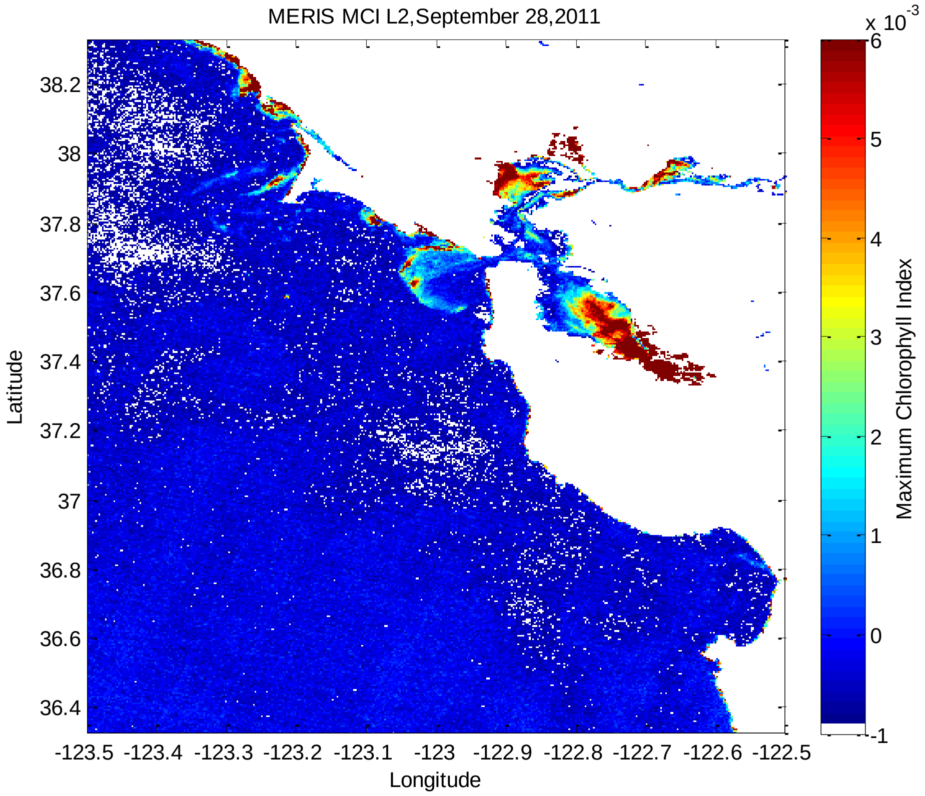
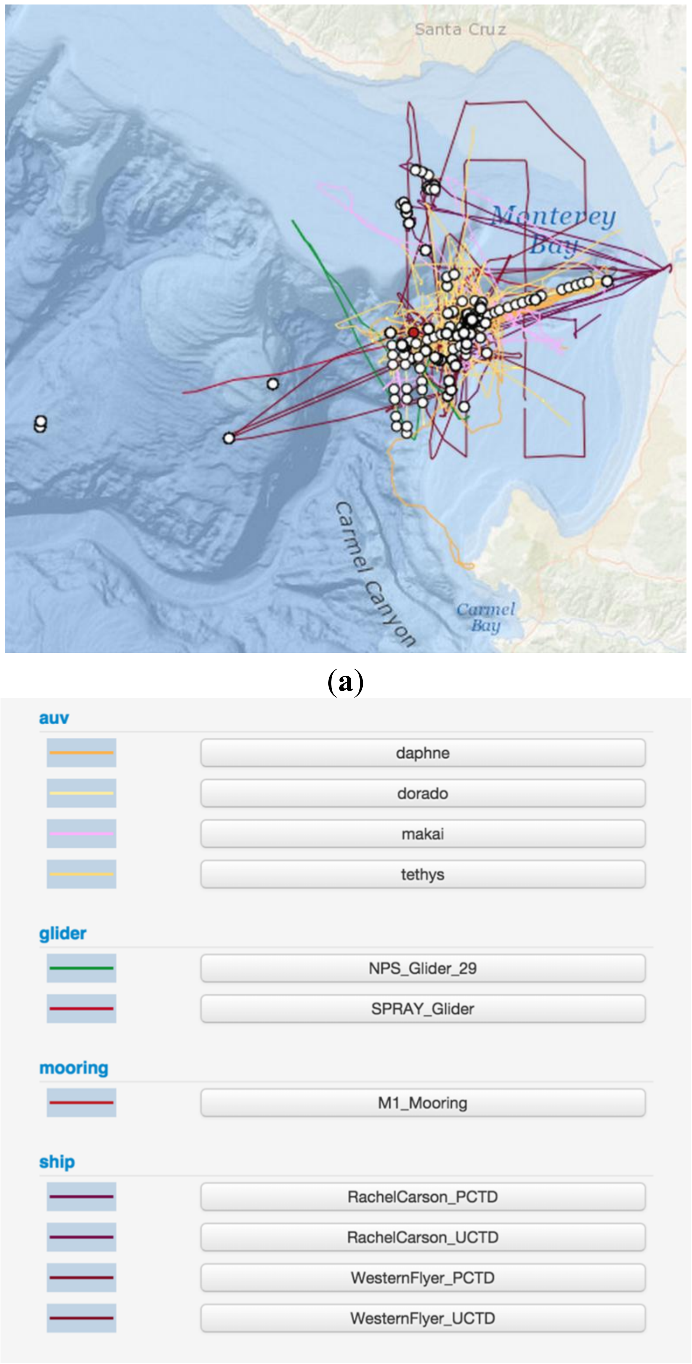


Figure 2. Bloom-like events in the northern part of Monterey Bay and San Francisco Bay through MERIS on 28 September 2011.

2.3. In Situ Data

There have been many field experiments in the Southern California Bight area in the past few years, with water samples collected by various platforms (e.g., AUVs, gliders, and ships) and analyzed by human experts. The resultant experimental data are stored in the Spatial Temporal Oceanographic Query System (STOQS) database and can be accessed to evaluate our model [29].

We used the CANON-ECOHAB-September 2014 data as the *in situ* dataset in this paper, obtained from the Fall 2014 Dye Release Experiment in Monterey Bay. This experiment began on 2 September 2014 21:21:36 GMT and ended on 11 October 2014 23:30:00 GMT. The experimental information is shown in Figure 3. The trajectories of different platforms (e.g., AUVs, gliders, ships) are marked with different colors. The red circle is the M1 mooring observation position, and the white circles denote the sampling positions of various platforms. Samples were taken at different depths at every sampling point, and the corresponding environmental parameters were recorded. Based on *in situ* data, we can obtain the corresponding MODIS data to predict the bloom area through our machine learning model and make further comparisons. Processing and analysis of these *in situ* data will be discussed in Section 5.



(**a**)

(**b**)

Figure 3. CANON-ECOHAB-September 2014 experimental data in STOQS (Copyright 2015 MBARI [30]). (**a**) is the map of the vehicle tracks and (**b**) describes the annotation of lines. The red circle is the M1 mooring observation position, and the white circles denote the sampling positions at different depths.

3. Satellite Data Analysis

To obtain a classification model in machine learning, we should obtain training inputs and their corresponding labels. This section will discuss how to obtain them from MODIS and MERIS data automatically.

3.1. Feature Extraction for Obtaining Training Input Data

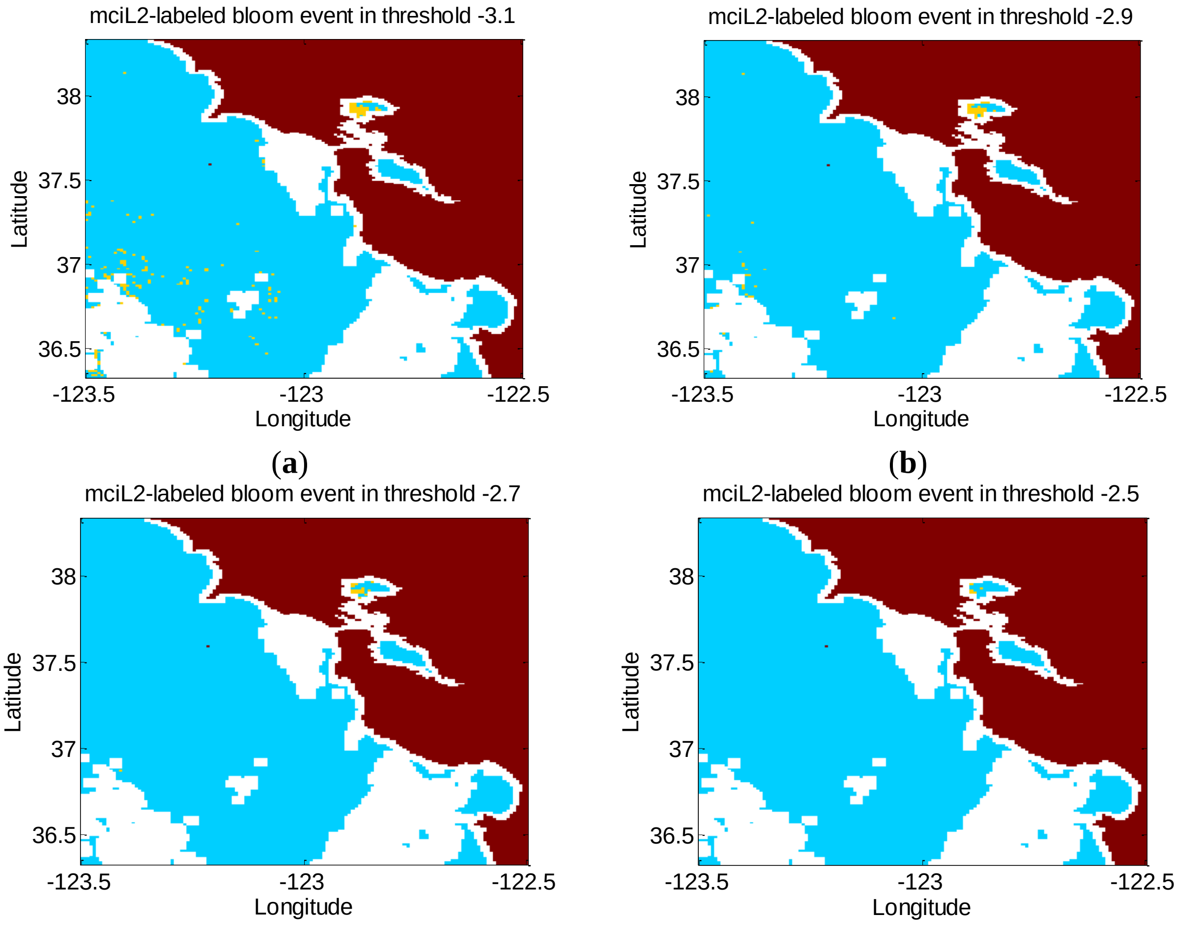
There are multiple features of MODIS data that can be used as inputs. We select suitable features through two steps: (a) finding features which can be not only obtained in MODIS datasets with suitable resolution but also relative to bloom events based on the oceanography researches; (b) testing the performance of different combinations. In this paper, we focus on the following five features: chlorophyll-A (chlA), fluorescence line height (flh), sea surface temperature (sst), the diffuse attenuation coefficient at 490 nm (k490) and cloud cover (cloud). The MODIS data of these five features can be obtained with 1-day period and 1 km^2-per-pixel resolution. ChlA and flh are widely used for bloom identification, and our previous work [15] also used these two features as the training inputs. However, k490 also plays a significant role in many oceanographic processes, including as an indicator of the quantification of radiative heating of the ocean [31], and sst has a positive relationship with the chlA concentration in phytoplankton bloom areas [32]. Finally, we consider that the cloud cover data could potentially affect the satellite sensing process, and therefore also add it to the input datasets too.

We had access to data from both MERIS and MODIS between October 2010 and April 2012. There were no MODIS data and/or no MERIS data on some specific dates, resulting in only 221 days possessing all five MODIS features as well as the corresponding MERIS MCI data during this time period. Because MERIS and MODIS data have different resolutions, they had to be processed to the same resolution. We used nearest-neighbor interpolation to down-sample the 0.0754 km^2-per-pixel MERIS data to the 1 km^2-per-pixel resolution MODIS data. In addition, remote sensing data can be uneven due to atmospheric conditions, leading to an absence of data in some areas. Therefore, we set a filter size to restrict our interpolation process to only search for a small range. We then used spatial filters to process the data and aligned them to a final input vector. After testing a number of spatial filters, we selected a blur filter to obtain the result of averaging over a 5×5 spatial grid. After obtaining the final inputs, the MERIS MCI data were then processed to obtain the corresponding labels, which are discussed in the next section.

3.2. Threshold Filter for Labeling the MERIS and in Situ Data

The MERIS MCI data are used as an indicator for algal blooms. Because the MCI values are quite small, we applied a base-10 logarithm to the data to measure the threshold between bloom and non-bloom areas. If the log10 of an MCI value is higher than this threshold, the corresponding area is considered a bloom area and labeled as 1 in the training process. Previous work [15] has simply used a threshold of −2.5 to divide bloom and non-bloom areas. However, although using a smaller threshold may lead to predicting more non-bloom areas as bloom areas, it also tends to include more real bloom areas in the prediction results. Thus, a trade-off needs to be considered in this process. In this paper, we try to deal with this problem in two ways. On one hand, the performance of our model with different thresholds is tested in the experimental Section 5 to validate the robustness. On the other hand, we analyzed the bloom event process to obtain the proper threshold in the rest of this section.

In a bloom event, the process generally begins close to shore, then moves out into the California current. As a result, we used different MCI thresholds to classify HABs in the MERIS images. Bloom event distributions for different thresholds on two consecutive days are shown in Figures 4 and 5. The bloom distribution is quite different for different thresholds. If the bloom threshold is under −3.0, we find the event in the California current [6], which is more believable than the absence of blooms in the California current given by the previous threshold of −2.5.


mciL2-labeled bloom event in threshold -3.1
mciL2-labeled bloom event in threshold -2.9
mciL2-labeled bloom event in threshold -2.7
mciL2-labeled bloom event in threshold -2.5
Latitude
38
37.5
37
36.5
Longitude
-123.5
-123
-122.5

(**a**) (**b**) (**c**) (**d**)

Figure 4. MCI-based bloom phenomenon appearance test using various thresholds on 31 January 2011. The MCI thresholds are at the top of each subfigure (**a**–**d**). Yellow means HAB, red means land, blue represents ocean without HAB, and white means no data.

Binding *et al.* [25] describe the relationship between MERIS and *in situ* data in Lake Erie. Although their data were evaluated in a different location, we used them as a guide to obtain the MERIS threshold. The relationship between the MERIS MCI and chlorophyll (Chl) is given by Equation (1):

$$MCI(Chl) = 0.0077Ln(Chl) - 0.017 \quad (1)$$

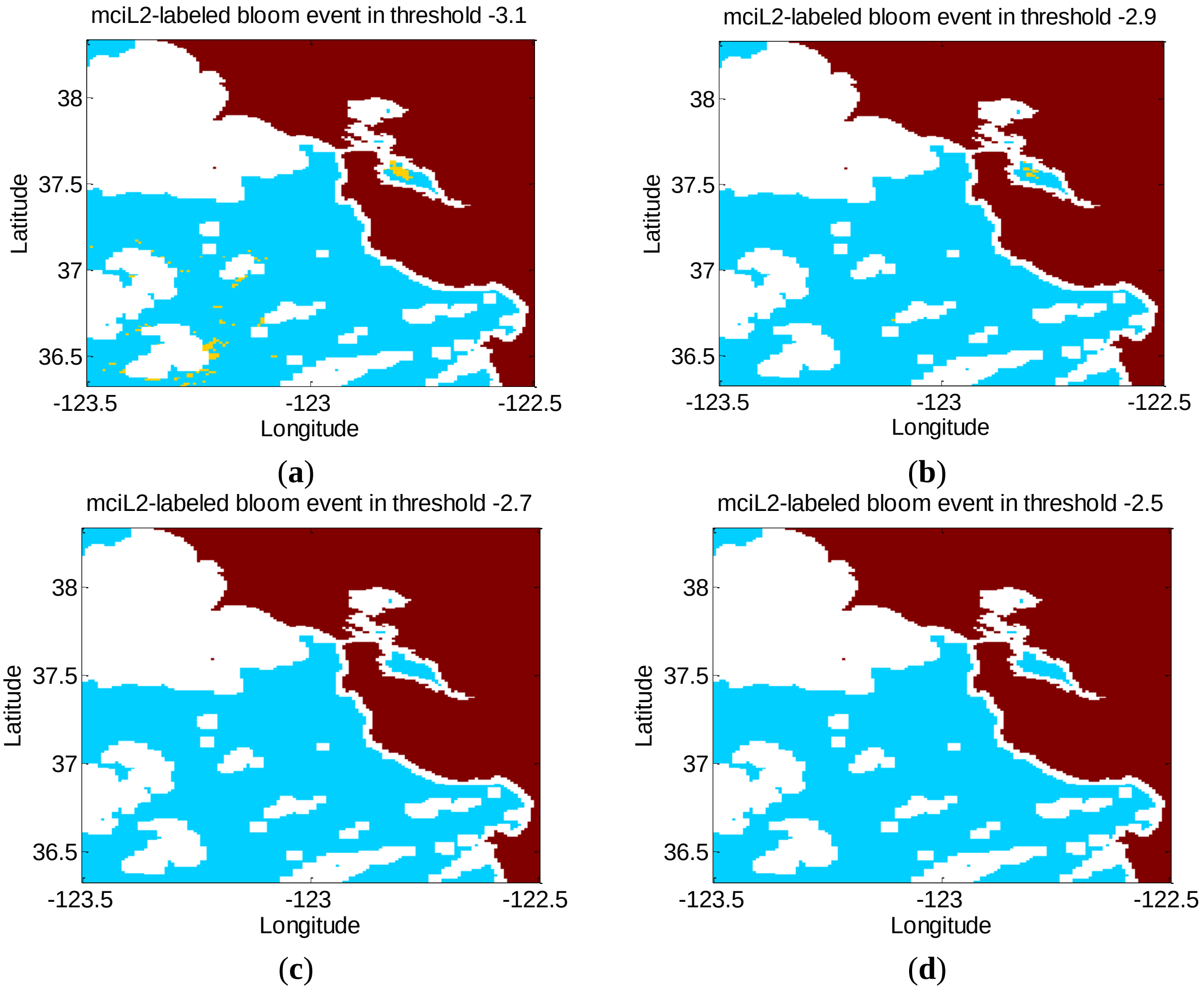
Chlorophyll is an indicator of blooms, and we define β as the chlorophyll threshold for bloom detection; therefore, the MCI threshold T in our model is calculated as

$$T = log_{10}(MCI(\beta)) \quad (2)$$

In addition, we simply use the same chlorophyll threshold β as a threshold detector Equation (3) to obtain the relationship between *in situ* data and bloom events, which is denoted as

$$F(y_x) = \begin{cases} 1 & if\ y_x > \beta \\ 0 & if\ y_x \leq \beta \end{cases} \tag{3}$$

where y_x is the *in situ* chlorophyll measurement. $F(y_x)$ is the prediction result for the *in situ* data: 1 and 0 indicate a bloom or not a bloom area.


mciL2-labeled bloom event in threshold -3.1
mciL2-labeled bloom event in threshold -2.9
mciL2-labeled bloom event in threshold -2.7
mciL2-labeled bloom event in threshold -2.5
Latitude
38
37.5
37
36.5
-123.5
-123
-122.5
Longitude

(**a**) (**b**)

(**c**) (**d**)

Figure 5. MCI-based bloom phenomenon appearance test using various thresholds on 1 February 2011. The MCI thresholds are at the top of each subfigure (**a**–**d**). Yellow means HAB, red means land, blue represents ocean without HAB, and white means no data.

4. Machine Learning for Bloom Event Prediction

4.1. Overview of the Bloom Event Prediction Framework

To build the bloom event prediction model, we need to determine the relationship between satellite data and *in situ* experimental data. However, there are not enough *in situ* experimental data to create a direct machine learning model, for which satellite data are used as model inputs and *in situ* data are processed as labels. For this reason, two satellite data products, MERIS and MODIS, are used in our framework. As shown in Figure 6, the inaccurate MODIS dataset is used as the machine learning training inputs and a threshold filter is designed to obtain the relationship between the accurate MERIS datasets and the *in situ* experimental data. Then, the training labels, experimental labels and corresponding MODIS test data are obtained. Based on the training inputs and training labels, a machine learning model (e.g., SVM, random forests) is built to predict potential bloom event areas. Various features are tested to improve our model, and the best combination is selected to obtain the final model. Finally, we use the experimental labels and corresponding MODIS data to validate our model.

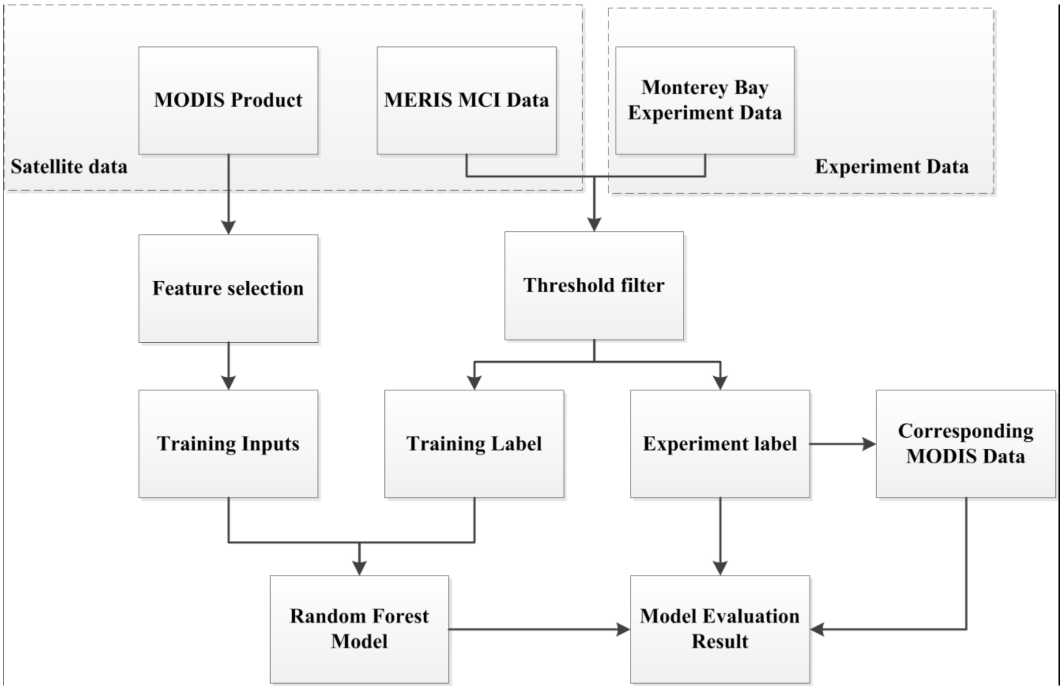


Figure 6. The framework of bloom event prediction.

4.2. Machine Learning for Classification

After we obtain the training inputs and outputs, we need to use the appropriate machine learning methods to address these data and classify the bloom and non-bloom areas. A classification model must be built by a set of observations and labels. Because the training data for this project are extremely imbalanced (e.g., the positive points are less than 1%), the selection of classification methods and the evaluation of our model are challenging problems. In this paper, support vector machine and random forest are used as classifiers. The Matthews Correlation Coefficient is used as an evaluation index, and Receiver operating characteristics (ROC) analysis is also used as an evaluation method in this paper.

4.2.1. Support Vector Machine

A Support Vector Machine (SVM) [33] is a supervised learning method based on margin-maximization. The SVM approach for classification is based on mapping training data and finding the optimal separating hyperplane in a higher-dimensional space.

Given a labeled training dataset $\{\boldsymbol{x}_i, y_i\}$, where $i = 1,2,\dots N$, $\boldsymbol{x}_i \in \mathbb{R}^k$, $y_i \in \{+1,-1\}$, N is the number of training inputs, y_i is the training label, and each vector contains k elements, a test vector $\boldsymbol{x}'$ can be classified into one label by the following equation:

$$F(\boldsymbol{x}') = sign(\sum_{i=1}^{N} y_i a_i k(\boldsymbol{x}_i, \boldsymbol{x}') + b) \tag{4}$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}')$ is the kernel function [34] and b is the bias term. $\boldsymbol{a}$ is determined by the following optimization function:

$$\max_{\boldsymbol{a}} w(\boldsymbol{a}) = -\frac{1}{2}\sum_{i,j=1}^{N} y_i y_j a_i a_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{N} a_i \tag{5}$$

subject to

$$\sum_{i=1}^{N} y_i a_i = 0, a_i \geq 0 \tag{6}$$

It is essential to obtain a proper type of kernel function for an SVM classifier. The Gaussian radial basis function is selected during the experiment evaluation in this paper, and can be defined as follows:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right), \gamma > 0 \tag{7}$$

The parameters in the SVM model are determined by a 6-fold cross-validation [34] in our experiment and the libsvm toolbox is used to train our SVM model [35].

4.2.2. Random Forest

Random forest (RF) [36] is an ensemble learning method for classification, which is performed by constructing multiple decision trees$\{h(\boldsymbol{x}, \Theta_k), k = 1, ...\}$. Each tree $\{\Theta_k\}$ is independent and identically distributed with the other trees in the forest. RF has many good features. It runs efficiently on large datasets. It is computationally much faster than boosting-based ensemble methods and somewhat faster than simple bagging. It can address thousands of training inputs instead of variable deletion and it has the ability to balance training/testing error in classification problems with unbalanced datasets. Finally, RF does not have an overfitting problem when more trees are added [36], which can save much time in the training process.

Training an RF classification model requires three steps. First, we use the bootstrap technique [37] to obtain *n* samples from the original training data. Second, unpruned classification trees are built from each bootstrap sample. In each node, instead of choosing all predictors, *m* predictors are randomly sampled to select the best split through these variables. Finally, new test data can be predicted by the majority voting of *n* trees in the classification problem.

In the RF model, two parameters must be selected: the number of Decision Trees (*k*) in the ensemble process and the minimum number of observations per tree leaf (*b*). For a given classification process, we can generally set the *m* value to one-third of the total input variables. However, to build a robust RF classifier, we should optimize *m* and *k* to minimize the generalization error. Using six-fold cross-validation, we select *b* as 1 and *k* as 300 to construct our final model.

4.2.3. Evaluation Methods

Several methods can be used to judge the performance of a classifier. The basic and most widely used methods are based on the confusion matrix [38]. If we test the classifier with a set of N observations, then the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) can be obtained. The accuracy is defined as $(TP + FP)/N$, the recall is $TP/(TP + FN)$, and the precision is $TP/(TP + FP)$.

However, this method is not effective in the bloom prediction problem because the data are extremely imbalanced and the blooms are very sparse in our datasets, constituting less than 1% of the data. In this situation, the accuracy is determined by the true negatives, which makes the number of true positives (bloom points) unimportant. For instance, a model that classifies all pixels as negatives would have over 99% accuracy. Similarly, the precision and recall could also be highly influenced by false positives and false negatives, respectively. However, these two methods could still be used as indicators for our model. Precision indicates the percentage of true blooms among the areas identified as blooms, and recall indicates the percentage of the true bloom region that we have correctly identified.

However, we wish to use a criterion that can summarize our performance with a single value. As a result, the Matthews Correlation Coefficient [34] is selected because it can measure the quality of a model operating on imbalanced data. True and false positives and negatives are considered in MCC, which is generally considered to be a balanced evaluation of the binary classification even if the classes are imbalanced. Based on the confusion matrix, the MCC can be calculated using the criterion Equation (8):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FP)(TP + FN)(TN + FN)}} \tag{8}$$

where the MCC value of +1 represents a perfectly correct prediction, −1 represents a perfectly wrong prediction and 0 represents a no better than random prediction.

However, a single value may not certainly reflect the actual status of classifiers. To deal with this problem, we use ROC analysis (see [39] for a tutorial) as another evaluation method in this paper. ROC graphs are able to visualize performance of classifiers and are insensitive to skew class distribution. In ROC space, the False Positive Rate (FPR) is plotted on the *x*-axis and the True Positive Rate (TPR) is plotted on the *y*-axis. The area under the ROC curve (AUC) can be used to compare different ROC curves.

5. Experimental Results

This section evaluates our model in two parts. First, we use different feature inputs and different MERIS thresholds to train the model based on a six-fold cross-validation and test its performance compared to our previous work based on SVM [15]. We attempt to determine the best features for predicting blooms. Second, we use *in situ* data from field experiments to validate our RF model.

5.1. Evaluation of the MCI-Based Model

In our former efforts, we chose the chlorophyll-A (chlA) and fluorescence line height (flh) features to train the SVM model [15]. In this paper, we used SVM/RF to train the model and evaluated another three features to obtain the final model.

The pixel-based intersections of the different feature inputs are different because the non-data area is not the same for each feature. To easily process the data, we only consider the final dataset including five filtered features and the corresponding MCI-based labels. In this case, each feature is represented in each pixel. To compare with our earlier results, we select the MCI threshold as −2.5 in this part, and the dataset includes 283,826 pixels, with only 544 bloom points. We divide these data into six folds, five folds for training and one fold for testing. The test results are shown in Tables 1 and 2.

Table 1. Test results based on SVM.

Features	MCC	Accuracy	Recall	Precision	Confusion Matrix (TP, TN, FP, FN)
flh, chlA	0.3194	0.998	0.121	0.846	(11, 47,212, 2, 80)
flh, chlA, cloud	0.3140	0.998	0.132	0.750	(12, 47,210, 4, 79)
flh, chlA, k490	0.3403	0.998	0.143	0.813	(13, 47,211, 3, 78)
flh, chlA, sst	0.3571	0.998	0.176	0.727	(16, 47,208, 6, 75)
flh, chlA, k490, sst	0.3827	0.998	0.209	0.703	(19, 47,206, 8, 72)
flh, chlA, k490, sst, cloud	0.5590	0.998	0.385	0.814	(35, 47,206, 8, 56)

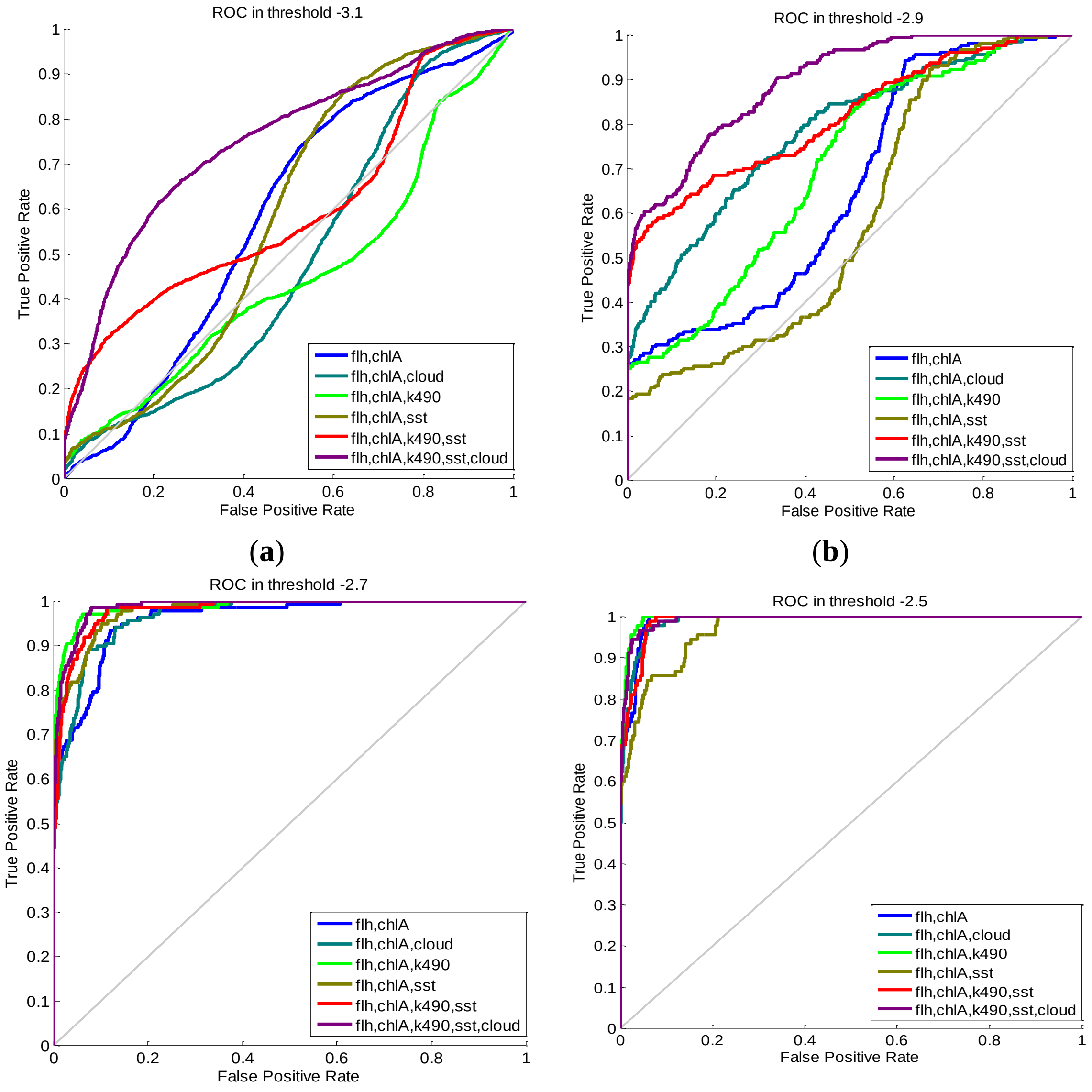
Table 2. Test results based on Random Forest.

Features	MCC	Accuracy	Recall	Precision	Confusion Matrix (TP, TN, FP, FN)
flh, chlA	0.4380	0.998	0.319	0.604	(29, 47,195, 19, 62)
flh, chlA, cloud	0.5637	0.999	0.451	0.707	(41, 47,197, 17, 50)
flh, chlA, k490	0.4396	0.998	0.330	0.588	(30, 47,193, 21, 61)
flh, chlA, sst	0.5775	0.999	0.462	0.724	(42, 47,198, 16, 49)
flh, chlA, k490, sst	0.6761	0.999	0.593	0.771	(54, 47,198, 16, 37)
flh, chlA, k490, sst, cloud	0.7062	0.999	0.615	0.812	(56, 47,201, 13, 35)

The tables clearly show that the Random Forest-based approach achieves more accurate results than the SVM-based method. RF achieves the highest MCC of 0.7062, whereas the maximum obtained using SVM is only 0.5590. In addition, the added k490, sst and cloud cover features play important roles in the bloom prediction process, and the highest MCCs occur when all five features are used.

We then performed ROC analysis through various MCI thresholds of both SVM and RF classifier. As shown in Figures 7 and 8, the ROC curves of four and five features in RF classifier are more northwest than the other three curves at different thresholds, which means that the RF classifier maintains a good performance. However, the performance of SVM is not really stable through its ROC curves, which move further away from the northwest section of the ROC graph as the threshold is decreased. These visualized graphs clearly reflect that the RF model has a better performance than the SVM model with various thresholds. In addition, these results show that if a human expert needs to change the prediction threshold, our RF model can still retain a high prediction quality, which is useful for implementation in the ODSS. We then tested the RF performance for a recommended MCI threshold of −3.1 in Table 3. The RF model still maintains a high MCC of 0.6743 and 0.7321 for four and five features, respectively.

However, the use of more features also results in fewer data inputs, because we must obtain the intersection of all the feature inputs to test our model. The datasets that have valid values for all five features are very small. As shown in the table, using four out of the five available features results in an MCC (0.6761) that is nearly as high as the maximum MCC (0.7062) achieved with five features, and resulted in more valid pixel-based datasets (638,839) than that of five features (283,826). In addition, the ROC curves show the performance achieved with between four and five features. Therefore, we ultimately use flh, chlA, k490, and sst to obtain our final model. The prediction results from 31 January 2011 are described in Figure 9, and the validation with *in situ* data will be discussed in the next subsection.



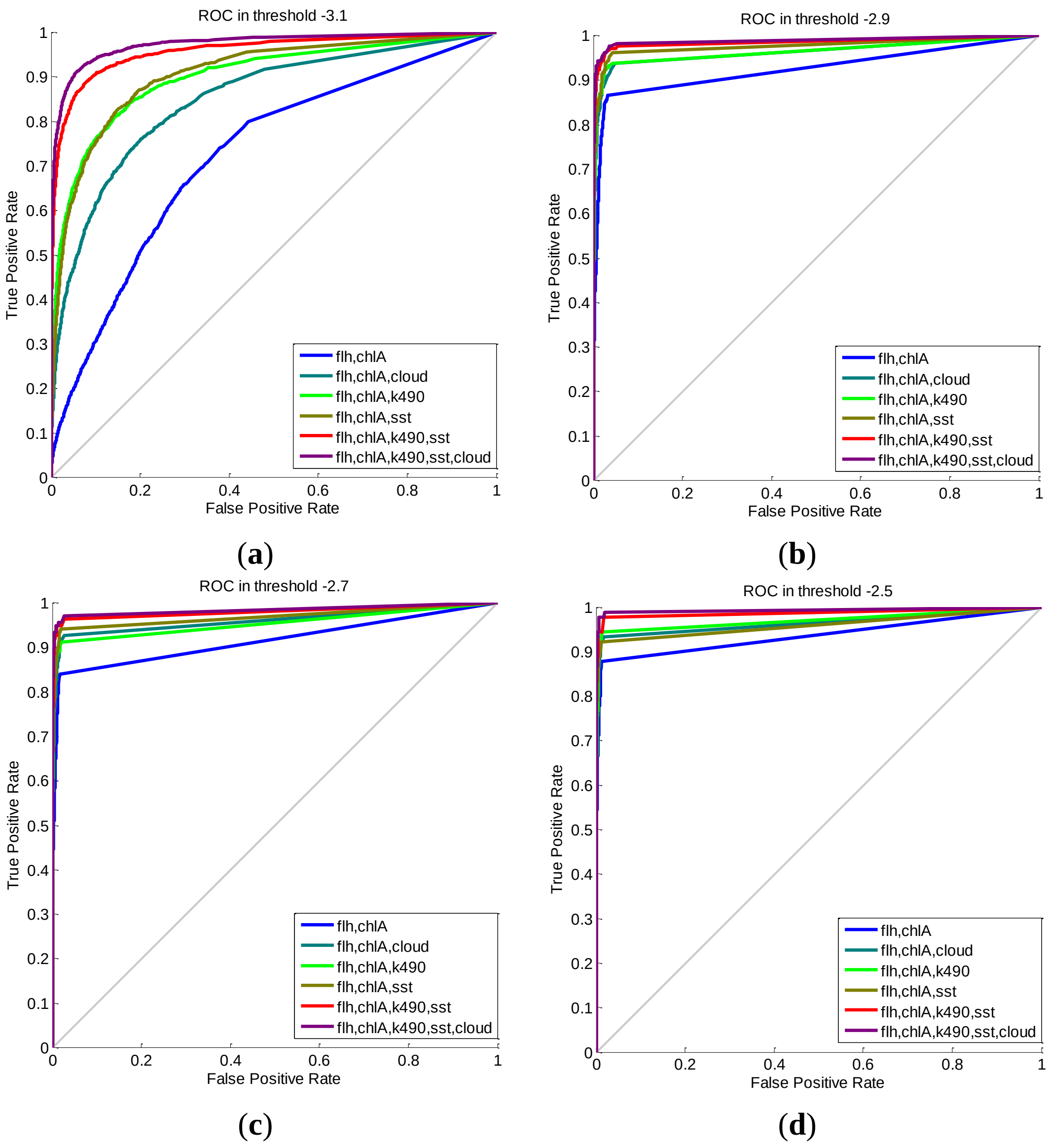
(a) **(b)**

(c) **(d)**

Figure 7. ROC curves based on SVM at different MCI thresholds with various combinations of features. The MCI thresholds are at the top of each subfigure (**a**–**d**).

Table 3. Test results based on Random Forest with MCI threshold −3.1.

Features	MCC	Accuracy	Recall	Precision	Confusion Matrix (TP, TN, FP, FN)
flh, chlA	0.1168	0.962	0.535	0.311	(90, 45,425, 199, 1591)
flh, chlA, cloud	0.2864	0.965	0.164	0.545	(276, 45,394, 230, 1405)
flh, chlA, k490	0.4633	0.971	0.329	0.688	(553, 45,373, 251, 1128)
flh, chlA, sst	0.3966	0.968	0.276	0.610	(464, 45,327, 297, 1217)
flh, chlA, k490, sst	0.6743	0.980	0.553	0.845	(929, 45,453, 171, 752)
flh, chlA, k490, sst, cloud	0.7321	0.983	0.622	0.880	(1045, 45,482, 142, 636)



(**a**) (**b**)

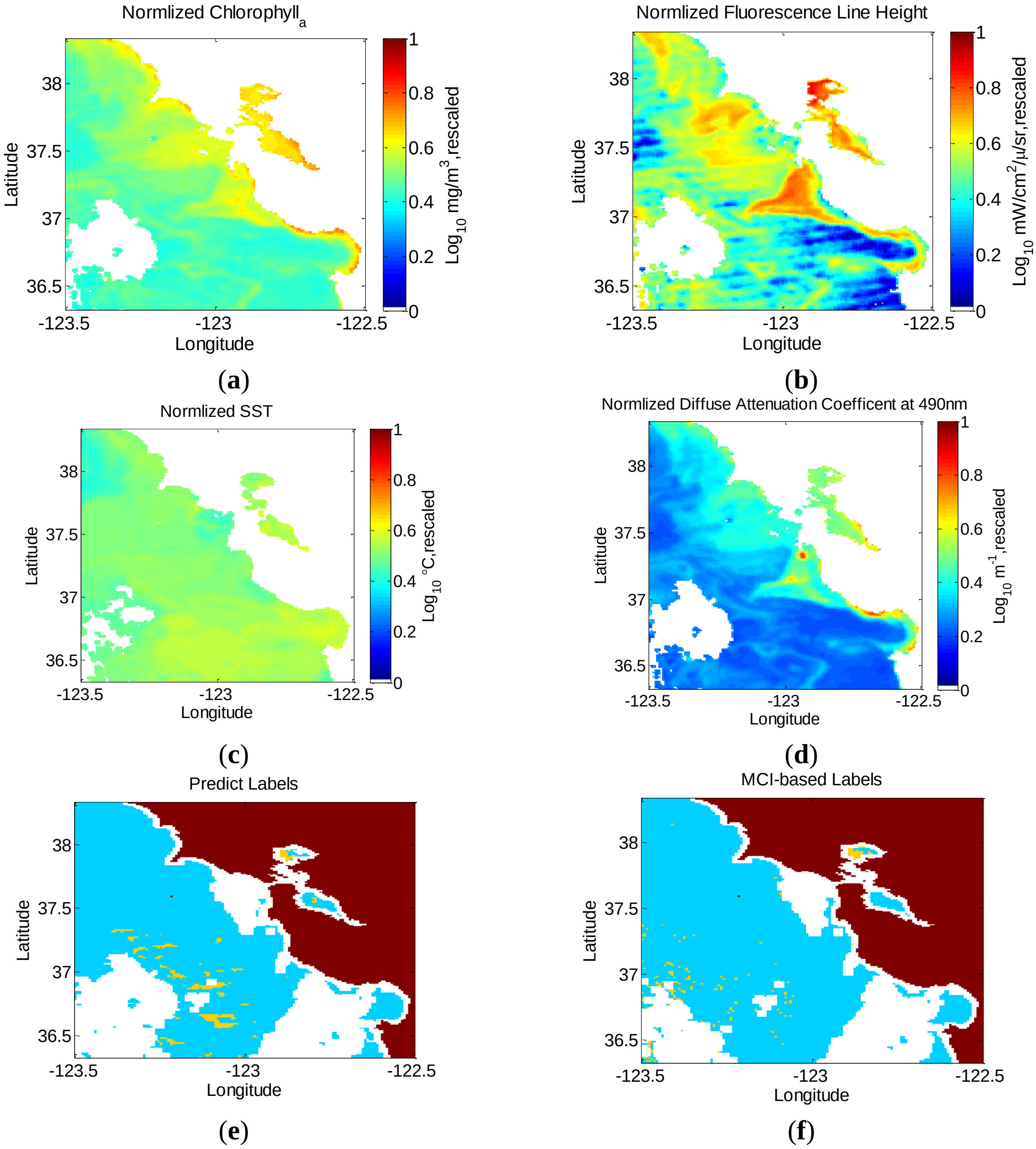
(**c**) (**d**)

Figure 8. ROC curves based on RF in different MCI thresholds with various combinations of features. The MCI thresholds are on top of each subfigure (**a**–**d**).

5.2. Evaluation of the Final Model Using in Situ Data from Field Experiment

In this section, we evaluate our model using CANON-ECOHAB-September 2014 *in situ* data from the STOQS database. We process the sampling data in three steps. (a) We resample the data and use the feature mass concentration of chlorophyll in sea water for our validation. The value of chlorophyll and its corresponding depth and (lat/lon) position are selected and classified into seven groups. Each group corresponds to a specific date because the chlorophyll data were only detected on seven days; (b) In each group, we reformulate the data into different position-value groups and down-sample the data into a low-resolution dataset; (c) At each position, we use linear regression (LR) to determine the relationship between the depth and chlorophyll value. Then, we use the LR model to predict the chlorophyll value at a depth of zero (0) meters (*i.e.*, on the water's surface) at each position, because this is the water depth upon which our model input and output are based.

In step (c), we calculate the determination coefficient of every linear regression model, which is an indicator of how well the data fit a statistical model. We ignore some inaccurately predicted values (e.g., determination coefficient less than 0.9) and obtain the final test datasets. In addition, we obtain the corresponding MODIS data from the website to prepare the remote sensing inputs.


Normlized Chlorophyll$_a$
Latitude
Longitude
38
37.5
37
36.5
-123.5
-123
-122.5
Log_{10} mg/m^3,rescaled
0
0.2
0.4
0.6
0.8
1
Normlized Fluorescence Line Height
Log_{10} mW/cm^2/μ/sr,rescaled
Normlized SST
Log_{10} °C,rescaled
Normlized Diffuse Attenuation Coefficent at 490nm
Log_{10} m^{-1},rescaled
Predict Labels
MCI-based Labels

(**a**) (**b**)

(**c**) (**d**)

(**e**) (**f**)

Figure 9. Test results for 31 January 2011. (**a**–**d**) represent MODIS-measured chlA, flh, sst and k490. (**e**–**f**) show labels predicted from the RF model (left) and MERIS MCI (right), respectively. Bloom pixels are shown in yellow, and non-bloom pixels in blue. White areas represent unknown pixels and land is red in the bottom row. Since the labels are generated only for the subset of known MERIS and MODIS imagery, the labels contain a smaller set of pixels than the MODIS images alone.

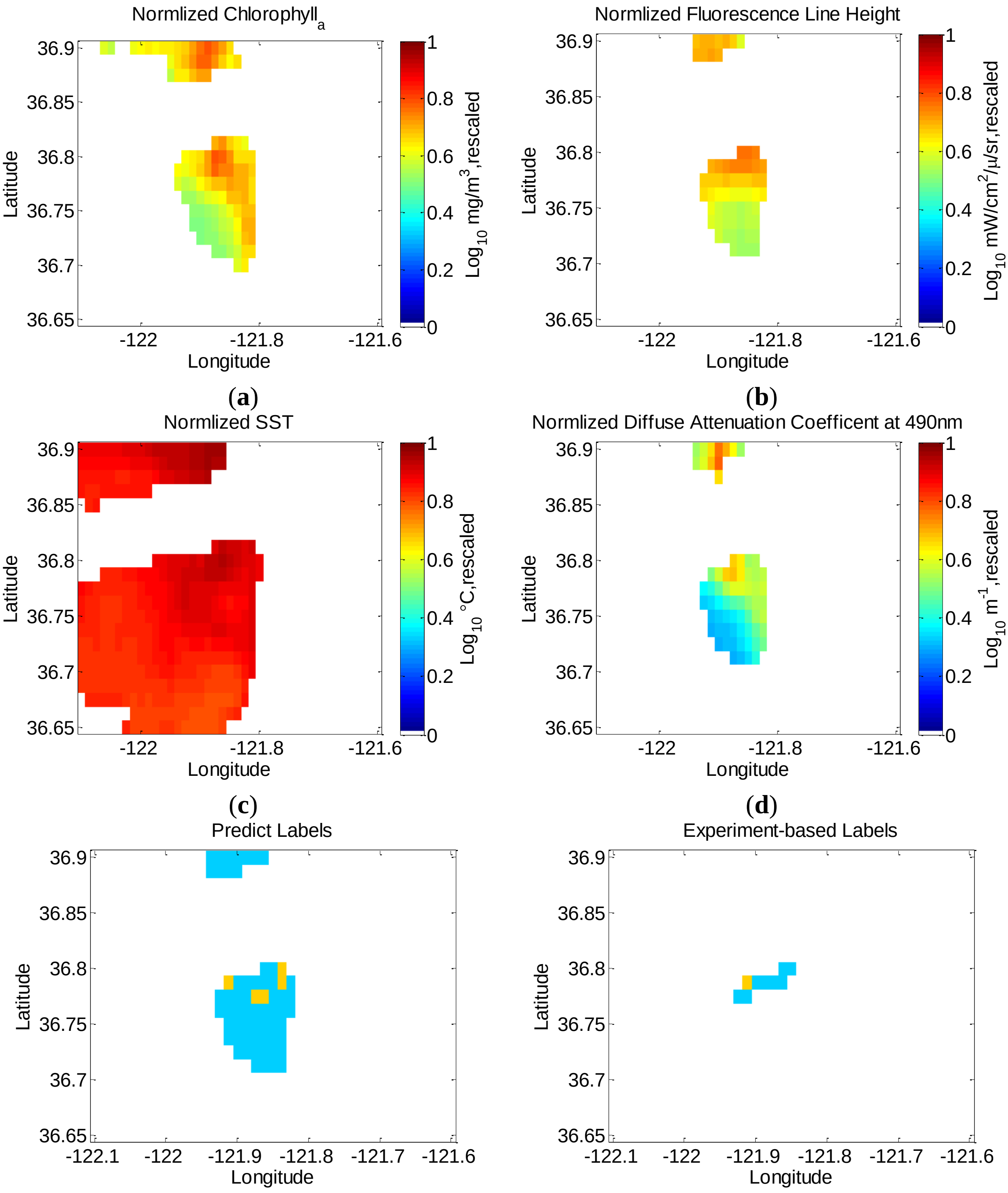
After obtaining all the data, we use our RF model to predict the blooms in this area. When we train the model, the bloom threshold is set to 10 mg/m^3, and the MCI threshold can be calculated as −3.1 using Equations (1) and (2). We then use this threshold to train an RF model that includes four features (chlA, flh, k490, sst).

For several yet to be confirmed reasons, some positions lack one or more MODIS data points, indicating that we can only validate the areas possessing both effective remote sensing data and *in situ* field experiment data. As a result, we select the data dated 19 September to test our model because a bloom event occurred on this date based on the *in situ* data. Because the *in situ* data have higher resolution than the MODIS data, there are several sets of *in situ* data in each MODIS prediction area, and we use the nearest interpolation to obtain the corresponding *in situ* data point to determine what the value should be at a point in the center of each MODIS data area. However, our goal is to locate the bloom area, and we therefore assume that, if one of those *in situ* measurements exceeds the bloom threshold, its corresponding area can be classified as a bloom area, which determines the experiment-based label. Therefore, we only consider the maximum *in situ* value in each prediction area. The test results are shown in Figure 10 and Table 4. The bloom threshold is set to 10 mg/m^3, which is the same as the threshold in the training process. We use RF to train the model, and the feature inputs are chlA, flh, k490 and sst run through a 5 × 5 blur filter, which is the result of averaging through a 5 × 5 grid.

As shown in Figure 10, our RF model predicts five pixels as a potential bloom area but only one pixel can be correlated with the experimental area. Although there are only nine pixels in the subset of both available areas, our model can successfully predict the bloom area. The detailed information about the testing locations is listed in Table 4.

Table 4. Comparison of predicted-label and experiment-based label.

Position	Feature Input	Predicted-Label	Maximum Chlorophyll	Experiment-Based Label
36.775 °N 121.925 °W	**chlA, flh, k490, sst**	0	7.50	0
36.775 °N 121.9125 °W	**chlA, flh, k490, sst**	0	8.72	0
36.7875 °N 121.9125 °W	**chlA, flh, k490, sst**	1	14.90	1
36.7875 °N 121.9 °W	**chlA, flh, k490, sst**	0	6.16	0
36.7875 °N 121.8875 °W	**chlA, flh, k490, sst**	0	4.14	0
36.7875 °N 121.875 °W	**chlA, flh, k490, sst**	0	5.51	0
36.7875 °N 121.8625 °W	**chlA, flh, k490, sst**	0	3.60	0
36.8 °N 121.8625 °W	**chlA, flh, k490, sst**	0	3.97	0
36.8 °N 131.85 °W	**chlA, flh, k490, sst**	0	5.77	0


Normlized Chlorophyll$_a$
Normlized Fluorescence Line Height
Normlized SST
Normlized Diffuse Attenuation Coefficent at 490nm
Predict Labels
Experiment-based Labels
Latitude
Longitude
36.9
36.85
36.8
36.75
36.7
36.65
-122.1
-122
-121.9
-121.8
-121.7
-121.6
Log_{10} mg/m^3,rescaled
Log_{10} mW/cm^2/μ/sr,rescaled
Log_{10} °C,rescaled
Log_{10} m^{-1},rescaled
0
0.2
0.4
0.6
0.8
1

(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 10. Field experiment evaluation results for 19 September 2014. (**a**–**d**) represent MODIS-measured chlA, flh, sst and k490. (**e**–**f**) show labels predicted from the RF model (left) and *in situ* experimental data (right), respectively. Bloom pixels are shown in yellow, and non-bloom pixels in blue. White areas represent unknown pixels. Because the labels are generated only for the subset of known MODIS imagery, the labels contain a smaller set of pixels than the MODIS images alone. The experiment-based labels are derived by *in situ* data so that they only cover a small area.

6. Conclusions

This paper uses remote sensing data to detect and predict bloom events automatically. In contrast to our prior work, we use an RF-based approach to build a model that can be used for ODSS and targeted sampling by AUVs. Several features (sst, cloud cover data, k490) were added to the chlorophyll-a and flh features in our model that effectively increased the bloom event prediction performance. In addition, we use *in situ* field experiment data from STOQS to validate our model.

In the future, we hope to build an oceanographic event response toolbox not only for blooms but also for other events, which could lead to efficient event prediction. More features (e.g., water color types, upwelling index) will be tested for the current model, and some new machine learning methods such as deep learning may be used to perform an automatic feature selection process. We will also evaluate the refinement of our model by adding time information or time-related features. Obtaining more *in situ* field experiment data will help to validate our model and facilitate its use in future field tests. Finally, a replacement for MERIS called the Sentinel-3 is planned to be launched in the middle of 2015, which can serve as another data source.

Acknowledgments

This work was funded by the National Science Foundation (grant No. 1124975) and NOAA (grant No. NA11NOS4780055). The MBARI author is also funded by a block grant from the David and Lucile Packard Foundation.

Author Contributions

Weilong Song developed the algal bloom event prediction framework and analyzed the satellite data and *in situ* data in this article. John M. Dolan provided funds, guidance in system design, data analysis and experiment evaluation. Danelle Cline provided satellite and *in situ* data sources and guidance in the feature selection process. Guangming Xiong contributed to the drafting of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bauman, A.G.; Burt, J.A.; Feary, D.A.; Marquis, E.; Usseglio, P. Tropical harmful algal blooms: An emerging threat to coral reef communities? *Mar. Pollut. Bull.* **2010**, *60*, 2117–2122.
2. Trainer, V.; Pitcher, G.; Reguera, B.; Smayda, T. The distribution and impacts of harmful algal bloom species in eastern boundary upwelling systems. *Prog. Oceanogr.* **2010**, *85*, 33–52.
3. Li, H.M.; Tang, H.J.; Shi, X.Y.; Zhang, C.S.; Wang, X.L. Increased nutrient loads from the Changjiang (Yangtze) River have led to increased Harmful Algal Blooms. *Harmful Algae* **2014**, *39*, 92–101.
4. Hoagland, P.; Anderson, D.; Kaoru, Y.; White, A. The economic effects of harmful algal blooms in the United States: Estimates, assessment issues, and information needs. *Estuaries* **2002**, *25*, 819–837.

5. Hudnell, H.K. The state of US freshwater harmful algal blooms assessments, policy and legislation. *Toxicon* **2010**, *55*, 1024–1034.
6. Ryan, J.; Dierssen, H.; Kudela, R.; Scholin, C.; Johnson, K.; Sullivan, J.; Fischer, A.; Rienecker, E.; McEnaney, P.; Chavez, F. Coastal ocean physics and red tides: An example from Monterey Bay, California. *Oceanography* **2005**, *18*, 246–255.
7. Ryan, J.; Fischer, A.; Kudela, R.; McManus, M.; Myers, J.; Paduan, J.; Ruhsam, C.; Woodson, C.; Zhang, Y. Recurrent frontal slicks of a coastal ocean upwelling shadow. *J. Geophys. Res.* **2010**, *115*, doi:10.1029/2010JC006398.
8. Ryan, J.P.; Gower, J.F.; King, S.A.; Bissett, W.P.; Fischer, A.M.; Kudela, R.M.; Kolber, Z.; Mazzillo, F.; Rienecker, E.V.; Chavez, F.P. A coastal ocean extreme bloom incubator. *Geophys. Res. Lett.* **2008**, *35*, doi:10.1029/2008GL034081.
9. Allen, J.I.; Smyth, T.J.; Siddorn, J.R.; Holt, M. How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? *Harmful Algae* **2008**, *8*, 70–76.
10. Gower, J.; King, S.; Borstad, G.; Brown, L. Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer. *Int. J. Remote Sens.* **2005**, *26*, 2005–2012.
11. Shen, L.; Xu, H.; Guo, X. Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework. *Sensors* **2012**, *12*, 7778–7803.
12. Miller, P.I.; Shutler, J.D.; Moore, G.F.; Groom, S.B. SeaWiFS discrimination of harmful algal bloom evolution. *Int. J. Remote Sens.* **2006**, *27*, 2287–2301.
13. Tomlinson, M.C.; Stumpf, R.P.; Ransibrahmanakul, V.; Truby, E.W.; Kirkpatrick, G.J.; Pederson, B.A.; Vargo, G.A.; Heil, C.A. Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico. *Remote Sens. Environ.* **2004**, *91*, 293–303.
14. Matarrese, R.; Morea, A.; Tijani, K.; de Pasquale, V.; Chiaradia, M.T.; Pasquariello, G. A specialized support vector machine for coastal water Chlorophyll retrieval from water leaving reflectances. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Boston, MA, USA, 7–11 July 2008; pp. 910–913.
15. Bernstein, M.; Graham, R.; Cline, D.; Dolan, J.M.; Rajan, K. Learning-based event response for marine robotics. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013; pp. 3362–3367.
16. Oceanographic Decision Support System. Available online: https://odss.mbari.org/odss/ (accessed on 23 September 2015).
17. Gomes, K.; Cline, D.; Edgington, D.; Godin, M.; Maughan, T.; McCann, M.T.; O'Reilly, T.; Bahr, F.; Chavez, F.; Messi é, M. Odss: A decision support system for ocean exploration. In Proceedings of the IEEE 29th International Conference on Data Engineering Workshops (ICDEW), Brisbane, Australia, 8–12 April 2013; pp. 200–211.
18. Das, J.; Rajan, K.; Frolov, S.; Pyy, F.; Ryan, J.; Caron, D.; Sukhatme, G.S. Towards marine bloom trajectory prediction for AUV mission planning. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Anchorage, AK, USA, 3–7 May 2010; pp. 4784–4790.
19. Das, J.; Maughan, T.; McCann, M.; Godin, M.; Reilly, T.O.; Messi é, M.; Bahr, F.; Gomes, K.; Py, F.; Bellingham, J.G. Towards mixed-initiative, multi-robot field experiments: Design, deployment, and lessons learned. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 3132–3139.

20. Das, J.; Harvey, J.; Py, F.; Vathsangam, H.; Graham, R.; Rajan, K.; Sukhatme, G. Hierarchical probabilistic regression for AUV-based adaptive sampling of marine phenomena. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 6–10 May 2013; pp. 5571–5578.
21. Ryan, J.P.; Davis, C.O.; Tufillaro, N.B.; Kudela, R.M.; Gao, B.C. Application of the hyperspectral imager for the coastal ocean to phytoplankton ecology studies in Monterey Bay, CA, USA. *Remote Sens.* **2014**, *6*, 1007–1025.
22. Ryan, J.; Greenfield, D.; Marin, R., III; Preston, C.; Roman, B.; Jensen, S.; Pargett, D.; Birch, J.; Mikulski, C.; Doucette, G. Harmful phytoplankton ecology studies using an autonomous molecular analytical and ocean observing network. *Limnol. Oceanogr.* **2011**, *56*, 1255–1272.
23. Jessup, D.A.; Miller, M.A.; Ryan, J.P.; Nevins, H.M.; Kerkering, H.A.; Mekebri, A.; Crane, D.B.; Johnson, T.A.; Kudela, R.M. Mass stranding of marine birds caused by a surfactant-producing red tide. *PLoS ONE* **2009**, *4*, e4550.
24. Kudela, R.M.; Lane, J.Q.; Cochlan, W.P. The potential role of anthropogenically derived nitrogen in the growth of harmful algae in California, USA. *Harmful Algae* **2008**, *8*, 103–110.
25. Binding, C.; Greenberg, T.; Bukata, R. The MERIS maximum chlorophyll index; its merits and limitations for inland water algal bloom monitoring. *J. Great Lakes Res.* **2013**, *39*, 100–107.
26. Hu, C.; Cannizzaro, J.; Carder, K.L.; Muller-Karger, F.E.; Hardy, R. Remote detection of Trichodesmium blooms in optically complex coastal waters: Examples with MODIS full-spectral data. *Remote Sens. Environ.* **2010**, *114*, 2048–2058.
27. Justice, C.O.; Vermote, E.; Townshend, J.R.G.; DeFries, R.; Roy, D.P.; Hall, D.K.; Salomonson, V.V.; Privette, J.L.; Riggs, G.; Strahler, A.; *et al*. The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1228–1249.
28. Rast, M.; Bezy, J.; Bruzzi, S. The ESA Medium Resolution Imaging Spectrometer MERIS a review of the instrument and its mission. *Int. J. Remote Sens.* **1999**, *20*, 1681–1702.
29. Campaign List. Available online: http://odss.mbari.org/canon (accessed on 23 September 2015).
30. Stoqs_September 2014. Available online: http://odss.mbari.org/canon/stoqs_september2014/query/ (accessed on 23 September 2015).
31. Stramska, M.; Świrgoń, M. Influence of atmospheric forcing and freshwater discharge on interannual variability of the vertical diffuse attenuation coefficient at 490 nm in the Baltic Sea. *Remote Sens. Environ.* **2014**, *140*, 155–164.
32. Wang, S.; Tang, D. Remote sensing of day/night sea surface temperature difference related to phytoplankton blooms. *Int. J. Remote Sens.* **2010**, *31*, 4569–4578.
33. Bishop, C. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 2nd ed.; 2006; Springer: New York, NY, USA, 2007.
34. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C.A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **2000**, *16*, 412–424.
35. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27, doi:10.1145/1961189.1961199.
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
37. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.

38. Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **1997**, *62*, 77–89.
39. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.