



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Campus de São José dos Campos
Instituto de Ciência e Tecnologia



Cemaden
Centro Nacional de Monitoramento
e Alertas de Desastres Naturais

UNIDADE DE PESQUISA DO
MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA,
INOVAÇÕES E COMUNICAÇÕES



PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

PEDRO HENRIQUE MORAES ANANIAS

**DETECÇÃO, MONITORAMENTO E PREVISÃO DE ANOMALIAS
COM APLICAÇÃO EM ESTUDOS DE CASO RELACIONADOS À
FLORAÇÃO DE ALGAS POTENCIALMENTE TÓXICAS UTILIZANDO
DADOS MULTITEMPORAIS DE SENSORIAMENTO REMOTO E
ALGORITMOS DE APRENDIZADO DE MÁQUINA**

PEDRO HENRIQUE MORAES ANANIAS

**DETECÇÃO, MONITORAMENTO E PREVISÃO DE ANOMALIAS COM
APLICAÇÃO EM ESTUDOS DE CASO RELACIONADOS À FLORAÇÃO DE ALGAS
POTENCIALMENTE TÓXICAS UTILIZANDO DADOS MULTITEMPORAIS DE
SENSORIAMENTO REMOTO E ALGORITMOS DE APRENDIZADO DE MÁQUINA**

Dissertação apresentada ao Instituto de Ciência e Tecnologia, Universidade Estadual Paulista (Unesp), Campus de São José dos Campos; Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden), como parte dos requisitos para a obtenção do título de MESTRE pelo Programa de Pós-Graduação em DESASTRES NATURAIS.

Área: Desastres naturais. Linha de pesquisa: Instrumentação e análise de dados.

Orientador: Prof. Dr. Rogério Galante Negri

São José dos Campos

2021

Instituto de Ciência e Tecnologia [internet]. Normalização de tese e dissertação [acesso em 2021]. Disponível em <http://www.ict.unesp.br/biblioteca/normalizacao>

Apresentação gráfica e normalização de acordo com as normas estabelecidas pelo Serviço de Normalização de Documentos da Seção Técnica de Referência e Atendimento ao Usuário e Documentação (STRAUD).

Ananias, Pedro Henrique Moraes

Detecção, monitoramento e previsão de anomalias com aplicação em estudos de caso relacionados à floração de algas potencialmente tóxicas utilizando dados multitemporais de sensoriamento remoto e algoritmos de Aprendizado de Máquina / Pedro Henrique Moraes Ananias. - São José dos Campos : [s.n.], 2021.

61 f. : il.

Dissertação (Mestrado) - Pós-Graduação em Desastres Naturais - Universidade Estadual Paulista (Unesp), Instituto de Ciência e Tecnologia, São José dos Campos, 2021.

Orientador: Rogério Galante Negri.

1. Sensoriamento Remoto Hiperespectral. 2. Floração de Algas. 3. Aprendizagem de Máquina. 4. Análise de Séries Temporais. I. Negri, Rogério Galante, orient. II. Universidade Estadual Paulista (Unesp), Instituto de Ciência e Tecnologia, São José dos Campos. III. Universidade Estadual Paulista 'Júlio de Mesquita Filho' - Unesp. IV. Universidade Estadual Paulista (Unesp). V. Título.

BANCA EXAMINADORA

Prof. Dr. Rogerio Galante Negri (Orientador)

Universidade Estadual Paulista (Unesp)

Instituto de Ciência e Tecnologia

Campus de São José dos Campos

Profa. Dra. Tatiana Sussel Gonçalves Mendes

Universidade Estadual Paulista (Unesp)

Instituto de Ciência e Tecnologia

Campus de São José dos Campos

Prof. Dr. Thales Sehn Körting

Instituto Nacional de Pesquisas Espaciais (INPE)

Divisão de Observação da Terra e GeoInformática

São José dos Campos, 12 de agosto de 2021.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS	4
RESUMO	5
<i>ABSTRACT</i>	6
1 INTRODUÇÃO	7
2 ARTIGOS	10
2.1 Artigo - “Anomalous behaviour detection using one-class support vector machine and remote sensing images: a case study of algal bloom occurrence in inland waters”	10
2.2 Artigo - “Forecasting algal bloom events in inland waters using remote sensing data and machine learning methods”	33
3 CONSIDERAÇÕES FINAIS	55
REFERÊNCIAS.....	57

LISTA DE ABREVIATURAS E SIGLAS

ABD	<i>Anomalous Behaviour Detection</i>
ABF	<i>Anomalous Behaviour Forecasting</i>
AM	Aprendizado de Máquina
AP	Aprendizado Profundo
API	<i>Application Programming Interface</i>
FAI	<i>Floating Algae Index</i>
GEE	<i>Google Earth Engine</i>
HAB	<i>Harmful Algal Blooms</i>
IA	Inteligência Artificial
LSTM	<i>Long-short Term Memory</i>
MNDWI	<i>Modified Normalized Difference Water Index</i>
NDVI	<i>Normalized Difference Vegetation Index</i>
OC-SVM	<i>One-class Support Vector Machines</i>
RF	<i>Random Forest</i>
SR	Sensoriamento Remoto
SABI	<i>Surface Algae Bloom Index</i>
SVM	<i>Support Vector Machines</i>

ANANIAS, M. H. P. Detecção, monitoramento e previsão de anomalias com aplicação em estudos de caso relacionados à floração de algas potencialmente tóxicas utilizando dados multitemporais de sensoriamento remoto e algoritmos de Aprendizado de Máquina. Dissertação. São José dos Campos: Universidade Estadual Paulista (Unesp), Instituto de Ciência e Tecnologia; Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden), 2021.

RESUMO

Globalmente, foram observadas nos últimos anos severas mudanças ambientais e climáticas. Neste cenário, destacam-se o aumento de eventos de floração de algas tóxicas, responsáveis pela degradação da qualidade da água e, principalmente, ameaça à saúde dos seres vivos. Estudos apontam para a necessidade de monitoramento e previsão desse fenômeno, os quais podem ser conduzidos segundo a concentração de clorofila. Perante esta motivação, utilizando-se técnicas de Aprendizado de Máquina e séries multitemporais de dados obtidos por Sensoriamento Remoto, foram desenvolvidos dois novos métodos capazes de proporcionar suporte à detecção e previsão de floração de algas. O primeiro visa a detecção automática do fenômeno e aplica conceitos de classificação de imagens por meio do emprego do algoritmo One-class Support Vector Machine. O segundo método é responsável por prever o seu surgimento. Como forma de evidenciar o potencial e viabilidade das propostas, os algoritmos foram aplicados em estudos de caso em áreas suscetíveis à ocorrência de floração de algas tóxicas.

Palavras-chave: Sensoriamento Remoto. Floração de Algas. Aprendizado de Máquina. Anomalia. Classificação. Séries multitemporais

ANANIAS, M. H. P. *Anomaly detection, monitoring, and forecasting with application in case studies related to the of potentially harmful algal blooms using multitemporal remote sensing data and Machine Learning algorithms*. Research projetct. São José dos Campos: São Paulo State University (Unesp), Institute of Science and Technology; National Center for Monitoring and Early Warning of Natural Disasters (Cemaden), 2021.

ABSTRACT

Globally, severe environmental and climate changes have been observed in recent years. In such a scenario, there was also an increase in the bloom of toxic algae, responsible for the degradation of water quality and threatening living beings' health. Studies point to the need to monitor and predict this phenomenon with basis on the chlorophyll concentration information. In the face of this motivation, using Machine Learning techniques and Remote Sensing multitemporal image series, two novel methods were developed to detect and forecast algae bloom. The first aims at the automatic detection of the phenomenon and applies concepts of image classification through the use of the One-class Support Vector Machine algorithm. The second method is responsible for predicting its emergence. Different case studies were carried in order to prove the potential of the proposed methods.

Keywords: Remote Sensing. Algal Bloom. Machine Learning. Anomaly. Classification, Multitemporal series

1 INTRODUÇÃO

Em um contexto global, mudanças ambientais como desmatamento, efeito estufa, desertificação ou perda da biodiversidade estão frequentemente relacionadas ao aumento da população humana e ao consequente descompasso na disponibilidade de recursos (GRIMMOND, 2007; CHAWLA; KARTHIKEYAN; MISHRA, 2020). Nesse sentido, Nagendra et al. (2013) salientam a importância em monitorá-las, mesmo diante do seu alto custo, a fim de permitir que grupos técnicos e o poder público tenham parâmetros realistas para a tomada de decisões.

No Brasil, observa-se a organização de movimentos a partir dos anos 60 com o objetivo de pressionar autoridades a adotarem políticas e agendas de prevenção de problemas ambientais (FOWLER; AGUIAR, 1993). Em um estudo recente, Nobre et al. (2011) afirmam que os governos federal e estaduais têm trabalhado com foco em ampliar o conhecimento necessário para que o país possa responder aos efeitos das mudanças climáticas nos diversos setores da sociedade.

Essencial para a vida, um dos recursos mais exigidos é a água, seja para consumo doméstico, seja para consumo agrícola ou industrial (MISHRA; COULIBALY, 2009). Segundo Wells et al. (2015), um dos problemas decorrentes desse aumento e responsável por severos danos aos ambientes aquáticos é a proliferação de cianobactérias (i.e. *Harmful Algal Blooms* – HAB). Causado por ações humanas (CARVALHO et al., 2013) ou mudanças climáticas (CASTRO; MOSER, 2012), esse fenômeno, também conhecido como floração, proliferação, afloramento ou *bloom*, é de difícil previsão (WELLS et al., 2015) e deve ser foco de estudos que busquem o seu monitoramento e forneçam suporte ao alerta precoce de sua ocorrência.

Nesse sentido, a clorofila-a (Chl-a), um importante pigmento fotossintético e causadora da coloração esverdeada de plantas, algas e cianobactérias (MILENKOVIĆ et al., 2012), é ponto central em discussões sobre a qualidade da água. Diversos parâmetros são descritos por Chawla, Karthikeyan e Mishra (2020) como responsáveis pela sua qualidade, incluindo sedimentos em suspensão, turbidez, fósforo total, conteúdo orgânico dissolvido, temperatura e disco de Secchi. No caso da Chl-a, o seu desenvolvimento está diretamente relacionado às mudanças repentinas em temperatura da superfície, velocidade do vento, precipitação, estratificação da coluna de água e direção do fluxo de água (WELLS et al., 2015; SHI et al., 2019; ROUSSO et al., 2020). Ao monitorá-la, é possível determinar se um reservatório ou rio, por exemplo, encontra-se próprio para consumo humano (MATTHEWS; BERNARD; ROBERTSON, 2012; CARVALHO et al., 2013). Por sua vez, conforme destacado por Chawla, Karthikeyan e Mishra (2020), observa-se a deficiência de estações de monitoramento, sendo muitas vezes negligenciadas pelo poder público.

Além do registro de sua ocorrência em diversos países do mundo (STUMPF et al., 2012; DUAN et al., 2015; YI et al., 2018; BINDING et al., 2018), florações de algas são frequentemente encontradas em território nacional (CARVALHO et al., 2013) em locais responsáveis pelo abastecimento urbano (BEYRUTH, 2000; OGASHAWARA et al., 2014) e pela geração de energia (MATSUMURA-TUNDISI; TUNDISI, 2005; WATANABE et al., 2018), por exemplo.

Além de afetarem o ecossistema e ameaçarem a saúde da população, apresentam risco para o tratamento de água, pois alteram seu odor e sabor (CARVALHO et al., 2013).

A ampla observação da Terra por meio de Sensoriamento Remoto (SR) permite o desenvolvimento de soluções, especialmente na avaliação quantitativa, qualitativa e temporal da proliferação de algas potencialmente tóxicas. Atualmente, uma variedade de satélites (e.g., *Terra*, *Aqua*, *Landsat-8*, *Sentinel-2*) apresenta-se como base para estes algoritmos.

Embora ferramentas que detectam automaticamente e permitam a geração de alertas precoces sejam escassas, são encontrados na literatura estudos baseados em dados empíricos e aplicação de índices espectrais com foco na detecção de HABs e estimação da concentração de Chl-a em ambientes aquáticos (GOWER et al., 2005; MATTHEWS; BERNARD; ROBERTSON, 2012). Um exemplo é aplicação do índice espectral Normalized Difference Vegetation Index (NDVI) (JR et al., 1973), inicialmente desenvolvido para o mapeamento de vegetação e utilizado em questões relacionadas às algas (ZHAO, 2003). Pode-se citar ainda o Floating Algae Index (FAI) (HU, 2009), onde Oyama et al. (2015) o utiliza no monitoramento dos níveis de cianobactéria em lagos do Japão. Outros exemplos são o Modified Normalized Difference Water Index (MNDWI) (HAN-QIU, 2005), inicialmente desenvolvido para detectar a presença de água limpa em corpos d'água (XU, 2006) e Surface Algae Bloom Index (SABI) (ALAWADI, 2010), outro índice exclusivamente utilizado na detecção de algas. Adicionalmente, são propostos estudos que fazem uso de processos matemáticos como a aplicação dos modelos Empirical Orthogonal Function (EOF) para estimar a concentração de Chl-a no Lago Taihu, China (QI et al., 2014) e do Medium Resolution Continental Shelf (MRCS) (ALLEN et al., 2008) para detecção de eventos de HABs em zonas costeiras europeias.

Diante do aumento na quantidade e qualidade de dados fornecidas por avançados sensores imageadores (YUAN et al., 2020; MARTÍNEZ-ÁLVAREZ; BUI, 2020), torna-se possível o aprimoramento da acurácia dos resultados de monitoramento, estimação e previsão da floração de algas. Esse ganho pode ser obtido com a aplicação de técnicas de Aprendizagem de Máquina (AM) e Aprendizado Profundo (AP), já aplicadas em outras áreas do conhecimento e expandidas na ciência do SR. O uso de Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM) (CORTES; VAPNIK, 1995), por exemplo, é abordado em estudos realizados por Sun, Li e Wang (2009) e Zhang, Huang e Wang (2020). Já os resultados com a aplicação do algoritmo de Florestas Aleatórias (*Random Forests* – RF), inicialmente proposto por Breiman (2001), podem ser observados nas publicações de Song et al. (2015) e Kupssinskü et al. (2020). A implementação de classificação supervisionada é possível quando os dados de treinamento estão disponíveis. Da mesma forma, quando o objetivo é automatizar a detecção de anomalias, o modelo One-class Support Vector Machine (OC-SVM) proposto por Schölkopf et al. (2000) surge como alternativa. Munoz-Mari et al. (2010) destacaram os desafios de usar OC-SVM para classificar dados pouco representativos, o que é um problema recorrente neste tipo de aplicação. No entanto, esses estudos mostram que modificações no processo de modelagem de OC-SVM podem aumentar sua eficácia de classificação e potencial detecção de anomalias.

Na mesma direção, são observadas ferramentas que fazem uso de AP (CHO; CHOI; PARK, 2018; VILÁN et al., 2013). Pode-se citar o estudo desenvolvido por Lee et al. (2019), que implementa redes neurais de múltiplas camadas na detecção de marés vermelhas (FLAGELLATES, 1979) na península coreana. Em outro estudo, Barzegar, Aalami e Adamowski (2020) propõem um modelo híbrido composto por Redes Neurais Convolucionais (*Convolutional Neural Network* – CNN) (LECUN et al., 1989) e Memória Longa de Curto Prazo (*Long Short-term Memory* – LSTM) (HOCHREITER; SCHMIDHUBER, 1997) na previsão da quantidade de Chl-a em um lago localizado na Grécia.

Conforme abordado anteriormente, a maioria dos modelos que se utilizam de imagens de SR se baseiam em modelos físicos (SATHYENDRANATH et al., 2001; FRANKLIN et al., 2020) (i.e. empíricos) ou são direcionados ao ambiente onde as amostras de campo foram coletadas, não sendo facilmente replicáveis em outros locais (YUAN et al., 2020). Desta forma, faz-se necessário o desenvolvimento de ferramentas universais e independentes de dados *in situ*.

Portanto, utilizando-se de técnicas de AM e análise de séries temporais de imagens de SR adquiridas da plataforma *Google Earth Engine Application Programming Interface* (GEE API), foram desenvolvidos dois novos algoritmos que possibilitam, de forma totalmente automatizada, a detecção e previsão de anomalias em ambientes aquáticos. O primeiro, denominado *Anomalous Behaviour Detection* (ABD), é responsável por detectar a ocorrência de proliferação de algas em águas interiores com uso do modelo RF. Sua performance foi avaliada com base em dados *in situ* dos Lagos Erie (USA) e Taihu (China). O segundo, denominado *Anomalous Behaviour Forecasting* (ABF) e modelado a fim de prever a ocorrência do fenômeno em questão, foi avaliado por meio de três estudos de casos que compreendem os lagos Erie, Chilika (Índia) e Taihu.

Este documento encontra-se organizado da seguinte forma: o Capítulo 2 apresenta os artigos “*Anomalous behaviour detection using one-class support vector machine and remote sensing images: a case study of algal bloom occurrence in inland waters*” e “*Forecasting algal bloom events in inland waters using remote sensing data and machine learning methods*”; O Capítulo 3 aborda as considerações finais do presente trabalho.

2 ARTIGOS

2.1 Artigo - “Anomalous behaviour detection using one-class support vector machine and remote sensing images: a case study of algal bloom occurrence in inland waters”

O artigo a seguir foi publicado em 17 de março de 2021 no periódico INTERNATIONAL JOURNAL OF DIGITAL EARTH <<http://dx.doi.org/10.1080/17538947.2021.1907462>>.



Anomalous behaviour detection using one-class support vector machine and remote sensing images: a case study of algal bloom occurrence in inland waters

Pedro Henrique Moraes Ananias ^{a,b} and Rogério Galante Negri ^{a,b}

^aGraduate Program in Natural Disasters, UNESP/CEMADEN, São José dos Campos, São Paulo, Brazil; ^bSciences Technology Institute, São Paulo State University (UNESP), São José dos Campos, São Paulo, Brazil

ABSTRACT

Algal blooms are a frequent subject in scientific discussions and are the focus of many recent studies, mainly due to their adverse effect on society. Given the lack of ground truth data and the need to develop tools for their detection and monitoring, this research proposes a novel method to automate detection. Concepts derived from multi-temporal image series processing, spectral indices and classification with One-class Support Vector Machine (OC-SVM) are used in this proposal. Imagery from multi-spectral sensors on Landsat-8 and MODIS were acquired through the Google Earth Engine API (GEE API). In order to evaluate our method, two bloom detection case studies (Lake Erie (USA) and Lake Taihu (China)) were performed. Comparisons were made with methods based on spectral index thresholds. Also, to demonstrate the performance of the OC-SVM classifier compared to other machine learning methods, the proposal was adapted to be used with a Random Forest (RF) classifier, having its results added to the analysis. In situ measurements show that the proposed method delivers highly accurate results compared to spectral index thresholding approaches. However, a drawback of the proposal refers to its higher computational cost. The application of the new method to a real-world bloom case is demonstrated.

ARTICLE HISTORY

Received 11 September 2020
Accepted 17 March 2021

KEYWORDS

Remote sensing; spectral indices; unsupervised classification; anomalies; algal bloom detection

1. Introduction

Natural disasters such as floods, fires, deforestation, thawing, water pollution, earthquakes, are becoming more widely reported (Marzuoli and Liu 2019). They are commonly related to anthropogenic causes, climate change, population growth, and overuse of land resources (Grimmond 2007; Chawla, Karthikeyan, and Mishra 2020). As described by the United Nations, I.S.f.D.R. (2015), a disaster is characterised by its ability to disorganise a society, causing human, material, economic or environmental losses.

In this sense, and as part of a management system, it is strongly desirable that the authorities know about these events in advance, so that they can build early warning systems to protect society (He et al. 2019).

Regarding aquatic environments, the appearance of phytoplankton is often associated with economic losses and the consequent impact on society (Klemas 2011). In addition to the damage to the local ecosystem, some species are responsible for poisoning humans and animals (Ghatkar, Singh, and Shanmugam 2019). On the other hand, these algae comprise an important share of the aquatic

food base and are accountable for fixing approximately 50% of the CO² present in the atmosphere (Ghatkar, Singh, and Shanmugam 2019).

In the face of environmental degradation and recent climate changes, controlling algal blooms has become a challenge for humankind (Yi et al. 2018). Observation of the Earth through satellites and remote sensing images allows quantitative, qualitative, and temporal assessments of these blooms, making room for the development of algorithms to solve this problem.

An example covered in Binding et al. (2018) highlights the concern about cyanobacterial proliferation in Lake Winnipeg (Canada) in recent decades, as this region provides essential support for multiple ecosystems, recreational and commercial activities, and is a source of hydroelectric power generation. The author also discusses that the impact is a function of the exposure time, location, quantity, and composition of the proliferating species.

Chlorophyll-*a* (Chl-*a*) is a pigment necessary for photosynthesis and is found in all types of organisms that perform it, mainly algae (Wetzel 2001). Chl-*a* has been the central focus of recent discussions on its classification in inland waters (Watanabe et al. 2018). Song et al. (2015) pointed out several challenges in this process, including a lack of ground truth data and unbalanced samples. Another issue is the complications generated by obtaining field data, such as high costs and remote locations (Klemas 2011).

Several studies have concerned the detection of potentially harmful algae using remote sensing. However, there have only been a few efforts to develop automatic detection and alert models for its presence (Song et al. 2015). One approach to this discussion is given by Sun, Li, and Wang (2009), who created a unified model to estimate the concentration of Chl-*a* in Lake Taihu (China) using support vector machines (SVM).

In parallel, Song et al. (2015) discussed the use of artificial intelligence methods to predict algal bloom in Monterey Bay (California, USA). In the study, given a lack of field data, the authors propose the use of the Random Forest model (RF) (Breiman 2001) and satellite images obtained from a Moderate Resolution Imaging Spectroradiometer (MODIS) and Medium Resolution Imaging Spectrometer (MERIS) to detect algal distribution in the bay coast.

The use of spectral indices and its thresholds for algal bloom detection in recent studies are found in the literature. One is the Normalized Difference Vegetation Index (NDVI) (Rouse et al. 1973), which was initially developed for vegetation mapping, but is formally used in studies related to algae. Another is the Floating Algae Index (FAI) (Hu 2009), which focuses on detecting floating algae.

However, the use of these indices based on thresholds is not a permanent solution, as it does not consider the historical behaviour of the algae. The existing algorithms, such as Maximum Chlorophyll Index (MCI) and Maximum Peak Height (MPH), were also considered. As pointed out by Shi et al. (2019), and following the scope of this research, the mentioned indices were not designed to be used with the Landsat-like data to solve the addressed problem, due to its limited spectral information and low signal-to-noise ratios. In this sense, (Shi et al. 2019) also suggested that new studies should follow the direction of elucidating algal bloom behaviour over time, including the development of tools that enable its prediction.

Moreover, the implementation of supervised classification in remote sensing, including Chl-*a* concentration estimation, is possible when training data is available. Similarly, when the objective is to automate anomaly detection, the One-class Support Vector Machine (OC-SVM) model proposed by Schölkopf et al. (2000) emerges as a potential alternative. Muñoz-Marí et al. (2010) highlighted the challenge of using OC-SVM to classify poorly representative data, which is a recurring problem in this type of application. However, these studies show that modifications in the modelling process of OC-SVM may increase its classification effectiveness and potential anomaly detection.

In light of the problems caused by algae presence, as well as the importance of its continuous monitoring, we propose a new algorithm based on fully automated unsupervised anomaly detection approaches, in order to verify the occurrence of such phenomenon in inland waters by using remote

sensing images acquired through the Google Earth Engine Application Programming Interface (GEE API). The formalisation and construction of this new proposal uses concepts related to spectral indices, image processing, and classification based on OC-SVM. In order to validate it, two case studies with *in situ* measurements are carried out on Lake Erie (USA) and a portion of Lake Taihu (China). A comparison of the proposal, based on a Random Forest (RF) classifier, with the other methods mentioned earlier (Zhao 2003; Jia, Zhang, and Dong 2019) is also included in this study. Finally, to demonstrate the application of the proposed method in a real-world algal bloom case, we present an annual mapping (i.e. 2014 to 2018) of this event in Lake Taihu, comparing the results obtained using the study presented by Jia, Zhang, and Dong (2019).

This article is organised as follows: Section 2 briefly reviews data classification and spectral indices; Section 3 presents the proposed algorithm; Section 4 describes the study areas and data as well as the experiment design and the results; Section 4.1 discusses details about the study areas, images used in the algae identification process and selected reference samples for results evaluation; Section 5 presents and discusses the results obtained; Finally, Section 6 concludes this paper.

2. Theory rational

2.1. Preliminary notations

Let $\mathcal{I}$ be the matrix representation of an image obtained by remote sensing. Each position of $\mathcal{I}$ is expressed in terms of s , defined over a regular grid $\mathcal{S} \subset \mathbb{N}^2$. Usually, s is called pixel and is related to a given geographic location and its sensor measurements, expressed by the $\mathbf{x} \in \mathcal{X}$ vector. We call the $\mathcal{X}$ vector space per attribute space. Thus, according to these elements and notations, $\mathcal{I}(s) = \mathbf{x}$ determines which behaviour of $\mathcal{I}$, in the respective s position, is expressed by the components of $\mathbf{x} = (x_1, x_2, \dots, x_\ell)$.

Assuming a defined geographic region over $\mathcal{S}$ support observed at distinct times t_i , $i = 1, 2, \dots, n$, it is convenient to adopt $\mathcal{I}^{(i)}$. Thus, $\mathcal{I}^{(i)}$ and $\mathcal{I}^{(j)}$ express the behaviour of the targets contained in the same region, but at different times.

Due to the need to delimit certain portions of a given image, the use of masks becomes convenient. For a given $\mathcal{I}^{(i)}$ image, the $\mathcal{M}^{(i)}$ image, also defined on the same support $\mathcal{S}$, whose positions are associated with binary values (i.e. 0 or 1). Under these conditions, $\mathcal{M}^{(i)}(s) = 0$ has singular use for occluding values/vectors assigned to the position s in $\mathcal{I}^{(i)}$. On the other hand, $\mathcal{M}^{(i)}(s) = 1$ identifies s positions where there is a particular interest. Consequently, the process of hiding attributes of $\mathcal{I}^{(i)}$ may be achieved through $\mathcal{I}^{(i)} \otimes \mathcal{M}^{(i)}$, where $\otimes$ represents the multiplication between $\mathcal{I}^{(i)}$ and $\mathcal{M}^{(i)}$, with respect to every $s \in \mathcal{S}$. In time, $\overline{\mathcal{M}}^{(i)}$ denotes the complement (i.e. $0 \rightleftharpoons 1$) of the binary values in $\mathcal{M}^{(i)}$.

Among different applications that use of remote sensing images, it is necessary to identify and distinguish several types of targets that compose the observed landscape. For this purpose, image classification techniques are usually employed.

The classification process consists of applying $F: \mathcal{X} \rightarrow \Omega$ to the attribute vectors $\mathbf{x}$ of each $s \in \mathcal{S}$ in order to associate a class $\omega_j \in \Omega$, $j = 1, \dots, c$. The distinct classification methods proposed in the literature establish ways of modelling F .

Usually, supervised and unsupervised approaches are adopted in remote sensing. In supervised learning, a set of labelled attribute vectors is used to learn how F should assign an unlabelled vector $\mathbf{x}$ to a class of Ω . Regarding the unsupervised approach, no labelled information is available, and F is modelled though the structural organisation found in the set where it is applied.

Additionally, according to Congalton and Green (2009), it is essential before the analysis of mappings obtained on remote sensing data that assessments are made regarding the accuracy of these mappings. Among different measures available in the literature, the kappa coefficient and the global accuracy allow to account for omissions and inclusions according to certain classes, as well as to compare different mappings between themselves (Bishop et al. 1977). F1-Score (van Rijsbergen

1979) comprises an alternative way of measuring the ratio between true and false detections (Yang and Liu 1999).

2.2. One-class support vector machines

Support Vector Machine (SVM) is a widespread classification technique, especially in remote sensing applications. A solid mathematical formulation, simple algorithmic architecture and high generalisation ability are some features that highlight such method (Bruzzone and Persello 2009). Furthermore, as reported in Mountrakis, Im, and Ogole (2011), SVM has achieved similar or even higher accuracy results compared to other classification methods.

Based on the original SVM conception, several variations have been proposed, for example, laplacian (Gu and Feng 2013), transductive (Li et al. 2018), context-sensitive (Bruzzone and Persello 2009; Negri, Dutra, and Sant'Anna 2014) and One-class (Schölkopf et al. 2001) SVMs. The last example, One-class SVM (OC-SVM), regards into an unsupervised approach motivated by quantile estimation (Glazer, Lindenbaum, and Markovitch 2013) and may be adopted for change detection (Negri et al. 2020) applications. Conceptually, from a given set of unlabelled observations, the OC-SVM obtains a model able to classify elements as part of such set with a false-positive/negative occurrence rate ν .

Formally, we may write a function $F: \mathcal{D} \subset \mathcal{X} \rightarrow \{+1, -1\}$, where $+1$ output implies that input elements are in $\mathcal{D}$, and -1 otherwise. The classifier F is given by:

$$F(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) - b \right), \quad (1)$$

where $b = \sum_{j=1}^m \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$ for any $\mathbf{x}_i \in \mathcal{D}$, and $K(\cdot, \cdot)$ is a kernel function. The coefficient α_i , $i = 1, \dots, m$, are obtained as solution of the following optimisation:

$$\begin{aligned} \min_{\alpha_1, \dots, \alpha_m} \quad & \sum_{i,j=1}^m \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & \begin{cases} \alpha_i \in \left[0, \frac{1}{\nu m} \right] \\ \sum_{i=1}^m \alpha_i = 1 \end{cases} \end{aligned} \quad (2)$$

The most significant characteristic that distinguishes this method from classic SVM approach lies in the optimisation problem expressed by Equation (2). It is worth noting that the OC-SVM is parameterised by $\nu \in [0, 1]$ and additional parameters related to the adopted kernel function. For example, when the RBF kernel (i.e. $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$) is adopted, $\gamma \in (0, \infty)$ should also be adjusted. More details on kernel functions can be found in Shawe-Taylor and Cristianini (2004).

As mentioned, the method is formalised on unsupervised learning and anomaly detection approaches, thus avoiding issues related to the difficulty of acquiring labelled data, since it only considers 'regular samples' in the learning process. Alternatively, it is essential to highlight that it is sensitive to the presence of noise in the dataset used in its decision rule modelling process; otherwise, it becomes liable to understand anomalies as regular occurrences.

2.3. Random forests

Random Forest (RF) is another classifier employed in the recent remote sensing studies. Introduced by Breiman (2001), its idea lies in the use of a forest of decision trees. Moreover, it exploits the

ensemble learning technique in order to combine the output of multiple decision trees through a major voting process, finally producing a classification decision.

Assuming a training set $\mathcal{D}$, n_{est} sets with the same cardinality are replicated by a bootstrap sampling. Posteriorly, for each replica, it is considered a random attribute subset with a maximum of n_{att} attributes and then used to build a decision tree. Parameters regarding such trees, like the maximum depth (p_{depth}), minimum samples to split (p_{split}), and minimum samples per leaf (p_{leaf}) should be tuned before the training process. More details and discussions regarding those parameters are found in Breiman (2001).

Concerning the classification stage, as previously mentioned, the vector $\mathbf{x}$ is assigned to a class in Ω that produces significant concordance among all individual trees. Accordingly to Belgiu and Drăguț (2016), RF is a computationally efficient algorithm that does not overfit the final decision rule.

2.4. Spectral indices thresholding for algal bloom detection

Spectral indices allow the extraction and analysis of remote sensing data. Despite a feature of interest, a spectral index can assist in its identification. This approach, also called ‘spectral enhancement’, is essential, given the impossibility of modifying orbiting imaging sensors and the difficulty of obtaining field data with the same spatial and temporal resolution (Verstraete and Pinty 1996).

Generally, spectral indices are derived from algebraic operations on the attributes of $\mathcal{X}$ that characterise the behaviour of $\mathbf{x}$ assigned with every pixel s of $\mathcal{I}^{(i)}$.

Examples of spectral indices proposed in the literature are focussed on the characterisation of vegetation (Rouse et al. 1973), components in water bodies (Gao 1996), constructed areas (Huang, Lu, and Zhang 2014), soil moisture (Khanna et al. 2007), and others (Xue and Su 2017). Among the different applications benefited by the use of spectral indices is the detection of algal bloom (Hu 2009; Mishra and Mishra 2012; Zhang et al. 2014; Houborg et al. 2016; Watanabe et al. 2018). Such problem has received attention due to its harmful potential to man and the environment.

Zhao (2003) proposes to distinguish algal concentration and bloom using thresholds in NDVI values. Admitting $\mathcal{I}^{(i)}$, an image whose attributes express the behaviour of imaged targets at wavelength ranges in the electromagnetic spectrum, NDVI is characterised by the following function:

$$f_{NDVI}(\mathbf{x}) = \frac{x_{NIR} - x_{Red}}{x_{NIR} + x_{Red}}, \quad (3)$$

where x_{NIR} and x_{Red} refer to the behaviour of the target at near-infra-red and red wavelengths, respectively. Without loss of generality, f_{NDVI} is applied to every $s \in \mathcal{S}$ since $\mathcal{I}^{(i)}(s) = \mathbf{x}$. Also, since $\mathbf{x} = (x_1, x_2, \dots, x_\ell)$, we have $NIR, Red \in \{1, 2, \dots, \ell\}$. Consequently, the approach proposed in Zhao (2003) can be expressed in terms of the following decision rule:

$$F_{NDVI}(\mathbf{x}) = \begin{cases} \omega_1, & f_{NDVI}(\mathbf{x}) \leq -0.15 \\ \omega_2, & \text{otherwise;} \end{cases} \quad (4)$$

being ω_1 and ω_2 classes referring to non-occurrence and occurrence of algae, respectively.

Additionally, other approaches can be used to detect algal bloom, to mention the study by Jia, Zhang, and Dong (2019). In this case, the following threshold for FAI (Hu 2009) values is set:

$$F_{FAI}(\mathbf{x}) = \begin{cases} \omega_1, & f_{FAI}(\mathbf{x}) \leq -0.004 \\ \omega_2, & \text{otherwise.} \end{cases} \quad (5)$$

such that:

$$f_{FAI}(\mathbf{x}) = x_{NIR} - \left(x_{Red} + (x_{SWIR} - x_{Red}) \times \frac{\lambda_{NIR} - \lambda_{Red}}{\lambda_{SWIR} - \lambda_{Red}} \right), \quad (6)$$

where x_{SWIR} is the behaviour of the target at short-wave infra-red wavelength. Also, λ_{SWIR} ranges from 1608 to 1640 nm.

3. Automatic algal bloom detection

3.1. Conceptual formalisation

Using a multi-temporal series of remote sensing images and based on the concepts of spectral indices and image classification, this work proposes an algorithm based on anomaly detection with case studies applied to algal blooms in aquatic environments. The diagram illustrated in Figure 1 presents an overview of this proposal and its steps, discussed according to its representation.

Initially, the user limits the *region of interest*, which contains the water body, a *spanning period* for characterising a time series of images, a *particular instant* for which anomalies are to be detected

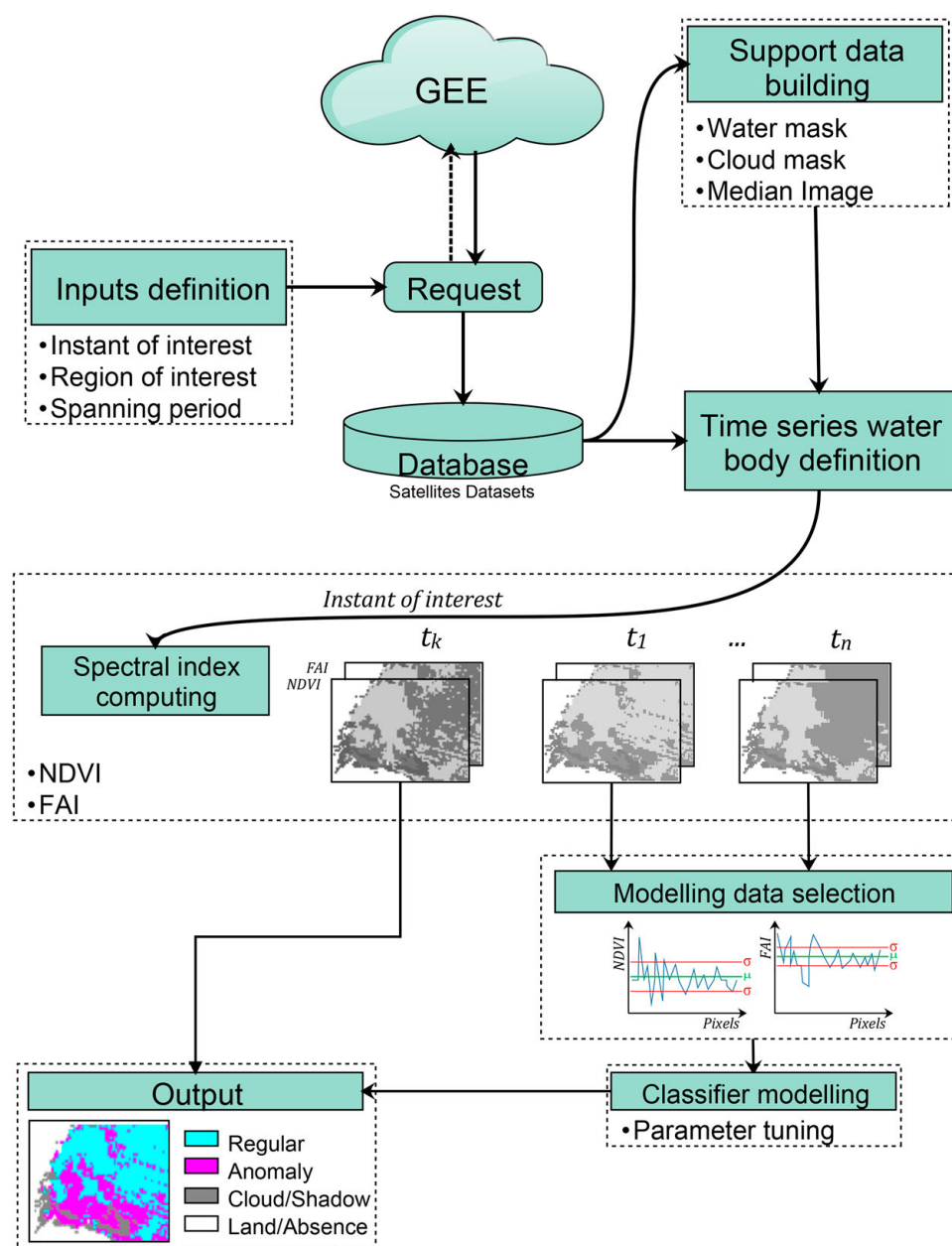


Figure 1. General organisation of the proposed method. The 'Modelling data selection' section refers to the process of creating a reference set by automatically selecting between pixels associated with the possible occurrence or non-occurrence of algal bloom based on indices median variation.

and the *sensor*, limited to Landsat-8 OLI (30 m spatial resolution) or MODIS MOD09GA.006 (500 m spatial resolution). According to the purposes of this research, anomalies are interpreted as an algal bloom occurrence. However, sudden changes in the target's spectral response can also be detected as non-regularities.

Once input parameters are set, a search for remote sensing images is performed. The Google Earth Engine (GEE) (Gorelick et al. 2017) platform is used to acquire data according to established criteria. As a result, images derived from the chosen sensor, with spectral bands ranging from blue to shortwave infra-red, are returned. To reduce the computational cost, those images are stored in a caching area, allowing them to be reused by the algorithm. Also, for each image, an auxiliary product is provided, allowing identification of elements such as water bodies and the presence of clouds. For convenience, further details regarding this subproduct are discussed later in Section 3.2 (see *support data* topic).

Subsequently, the query performed allows the construction of time series images representing $\mathcal{I}^{(1)}, \mathcal{I}^{(2)}, \dots, \mathcal{I}^{(n)}$. Such series have temporal ordering, where $\mathcal{I}^{(n)}$ refers to the most recent instant. Also, we have $\mathcal{I}^{(k)}$, where $k \approx \frac{n}{2}$, expressing the *particular instant*. The difference between $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(k)}$; or the equivalent between $\mathcal{I}^{(k)}$ and $\mathcal{I}^{(n)}$, expresses the input *spanning period*.

Through the mentioned auxiliary product, masks are constructed to allow delimitation of the existing water body in the area of interest as well as mapping locations associated with the occurrence of clouds and cloud shadows. Since its occurrence has a dynamic behaviour over time, it is necessary to define masks $\mathcal{M}_{cl}^{(1)}, \dots, \mathcal{M}_{cl}^{(n)}$ for each image/instant $\mathcal{I}^{(1)}, \dots, \mathcal{I}^{(n)}$. On the other hand, the regions that comprise water are defined by a single mask $\mathcal{M}_{wb}$.

It is a common issue that, in applications involving the observation of the Earth's surface via optical sensors, the occurrence of clouds provides a lack of information about the targets. This work considers images with good visibility, i.e. those where cloud cover is less than or equal to 50%. More information on this procedure is shown in Section 3.2 (*cloud threshold* topic). Under these conditions, the regions affected by this atmospheric phenomenon are ignored. For the sake of mathematical and computational simplicity, it is convenient to determine the median image $\tilde{\mathcal{I}}$, whose attribute vector associated with each position s represents the median of each attribute, also in s , for the considered time series.

From the initial time series and its supporting data (i.e. masks and median image), each image is limited to the water body, followed by correcting the target spectral response with the removal of cloud occurrences and median trend subtraction. Such a process is formally expressed by:

$$\mathcal{I}^{(i)} := \left(\mathcal{I}^{(i)} - \tilde{\mathcal{I}} \right) \otimes \overline{\mathcal{M}_{cl}^{(i)}} \otimes \mathcal{M}_{wb}, \quad i = 1, \dots, n, \quad (7)$$

where ‘ $-$ ’ denotes the usual matrix subtraction and $\overline{\mathcal{M}_{cl}^{(i)}}$ is the reverse of mask $\mathcal{M}_{cl}^{(i)}$. Note that the expressed process provides a redefinition of each $\mathcal{I}^{(i)}$.

When in possession of the multi-temporal series of images, adjusted according to Equation (7), and facing the central objective of this proposal, which is the detection of algal blooms, measures that favour the identification of this phenomenon are computed. Such measures refer to the NDVI and FAI indices discussed in Section 2.4.

With the exception of the image $\mathcal{I}^{(k)}$ (i.e. the *instant of interest*), values from the spectral indices considered are observed in relation to all pixels of the time series. According to these index values, processed separately, it is possible to extract their average trend μ_z and its variation level, given in terms of the standard deviation σ_z , with $z \in \{\text{NDVI}, \text{FAI}\}$. In turn, the values performed by each index, and not limited by the range $[\mu_z - \sigma_z, \mu_z + \sigma_z]$, may represent an anomaly occurrence according to z .

While the OC-SVM model uses the $\mathcal{D}$ reference set represented by non-occurrence of algal bloom (i.e. $[\mu_z - \sigma_z, \mu_z + \sigma_z]$, for $z = \{\text{NDVI}, \text{FAI}\}$), the RF algorithm is trained with all datasets, which also comprehends the anomaly class (i.e. $]-\infty, \mu_z - \sigma_z] \cup [\mu_z + \sigma_z, +\infty[$, for $z = \{\text{NDVI}, \text{FAI}\}$). However, before modelling, the set of observations defined as regular or anomalous is used as auxiliary information in the process of selecting the method's parameters. Thus, it

makes it possible to obtain a decision rule capable of identifying the occurrence of anomalies according to the behaviour of the considered spectral indices. It is important to note that the adjustment of parameters associated with each method (i.e. the ν and kernel function parameters related to OC-SVM or depth (p_{depth}), number of estimators (n_{est}), feature subset size and minimum samples required to split (p_{split}) or to be a leaf node (p_{leaf}), regarding RF method) is conducted automatically. Further details on this procedure are discussed in Section 3.2 (see *model parameter tuning* topic).

Lastly, $\mathcal{I}^{(k)}$ is expressed in terms of the NDVI and FAI indices and then submitted for classification. As a result (i.e. *output*), maps in Tagged Image File Format (TIFF) and GeoJSON (Butler et al. 2016) formats delimit the water body between *anomaly* and *regular* classes (i.e. no anomalous observation).

Along with the purposes, which drove this development, the introduced algorithm is denoted by the acronym ABD-OCSVM, where ABD stands for *Anomalous Behaviour Detection*. In this context, ABD-RF stands for the use of the RF model.

3.2. Implementation details

The previous formalisation was simplified to make it clear. The following information completes the proposal. The code of the proposed algorithm is freely available at <https://github.com/pedroananas/abd>.

Programming language and libraries: The Python 3.6 (van Rossum and Drake 2011) programming language was used to implement the proposed algorithm. Additionally, functions provided in the Numpy (Van Der Walt, Colbert, and Varoquaux 2011) and Pandas (McKinney 2010) libraries were used for data manipulation. The Scikit-Learn (Pedregosa et al. 2011) library was used in OC-SVM and RF modelling and classification processes.

Google Earth Engine API: To access Landsat-8 OLI and MODIS MOD09GA.006 satellites imageries, Google Earth Engine Application Programming Interface (API) (GEE-API 2019), compatible with the Python language, was used. This API allows automation of the image search process for a given period and region of interest. Also, it is worth mentioning that images from Landsat (Landsat 8 Surface Reflectance Code (LaSRC) algorithm (USGS 2017)) and MODIS (USGS 2020) sensors returned by this API are previously subjected to atmospheric correction and all processing and modelling steps after its extraction are performed outside the GEE platform.

Model parameter tuning: As discussed in Section 2.2 and 2.3, OC-SVM and RF methods require the adjustment of parameters inherent in its formalisation (i.e. the ν parameter or numbers of trees in the forest). Given the high degree of freedom associated with the process of selecting appropriate parameters for the method modelling step, a Randomized Grid Search procedure was employed (Rastrigin 1963; Baba 1981; Bergstra and Bengio 2012). This procedure consists of testing a finite set of parameter settings and selecting one that ensures higher accuracy. For OC-SVM, the search space that determines the tested settings is given from the $\nu \in \{10^{-1}, 10^{-2}, \dots, 10^{-7}\}$ value and a RBF kernel with $\gamma \in \{10^{-1}, 10^{-2}, \dots, 10^{-7}\}$. Regarding the RF parameters, with the Gini impurity measure guiding the nodes splitting process, the tested settings are $n_{est} \in \{1, 5 \dots 250\}$, $p_{depth} \in \{1, 2, \dots, 30\}$, $p_{split} \in \{2, 4, \dots, 20\}$, $p_{leaf} \in \{2, 4, \dots, 20\}$ and $n_{att} \in \{\sqrt{\dim(\mathcal{X})}, 100\%, 75\%, 50\%\}$. In addition, this procedure is replicated in the decision rules modelling the data set for each possible configuration, according to a 10-fold cross-validation process.

Decision rule modelling dataset size: As described above, the samples comprising the decision rule modelling dataset are defined as a function of the trend range $[\mu - \sigma, \mu + \sigma]$. Given the extremely high number of examples (i.e. attribute vectors associated with each pixel) in the time series employed, the use of all available data for modelling the OC-SVM method becomes computationally impeding, thus motivating the use of randomly selected subsets. After preliminary testings with 10%, 5%, 2.5% and 1% of available examples, it was found that using 1% allowed

results similar to other higher proportions with a lower computational cost. There is an exponential increase in the computational cost with an increasing decision rule modelling set size, as a time series can reach more than ten million pixels in its processing dataset. Thus, the 1% ratio is employed in determining the modelling set.

Support data: Masks that delimit water bodies and cloud occurrence within the region of interest are used as supporting data in the developed algorithm. In turn, bitwise operations are used to operate binary values directly (Cavanagh 2013). In addition, the AND operator is commonly used to create filters and masks. In our study, the masks are derived from the operation performed on the object *qa*, which represents the quality assessment band of the working product, using function *bitwiseAnd*. More information can be found in Gorelick et al. (2017). For Landsat-8 OLI, the quality band is known as 'pixel_qa', comprising bits 3 and 5 for cloud/shadow. MODIS MOD09GA.006 sensor uses a bitwise operation for cloud occurrence (bit 10) extraction through the 'state_1km' subproduct. The extraction of water bodies must be done with auxiliary products, called 'MOD44W.006 Terra Land Water Mask Derived from MODIS and SRTM Yearly Global 250m' and 'NASA GLCF Landsat Global Inland Water' throughout bands 'water_mask' and 'water', respectively. This is necessary to ensure that images comprising the time series have the same number of pixels.

Cloud threshold: The occurrence of clouds in remote sensing data brings little or no information. The implementation of the proposed algorithm considered occurrence thresholds according to the percentages 25%, 50% and 75% to determine the best value for the selection and use of the images (i.e. good visibility). Preliminary analysis for the NDVI index, based on the data in Section 4.1, showed a mean and variance equal to -0.42 and 0.11 , when considering images with up to 25% of cloud occurrence. This was -0.48 and 0.11 for 50%; -0.57 and 0.11 for 75% occurrence. Since the difference between these values is not statistically significant at 5%, it was decided to use the intermediate percentage equal to 50%. However, it is worth emphasising that the sites affected by the occurrence of this phenomenon are discarded from the composition of the database used in modelling the method.

Spanning period: The *spanning period* is an input parameter responsible for determining the size of the time series used in the anomaly detection process. Note that the general trend that characterises the occurrence of anomalies (i.e. algal bloom) is determined as a function of time series characteristics. Thus, the spanning period may directly influence the anomaly detection process. If considered a *spanning period* of d days, the time series comprehends the period between d days after and before the *particular instant* (i.e. the image $\mathcal{I}^{(k)}$). In addition, the algorithm will automatically seek to allocate this period so that the excess on both ends of the series is reallocated from one to the other, considering the selected value and the image $\mathcal{I}^{(k)}$. Preliminary tests using 90, 180 and 365 as inputs indicated that the 180-value is the most stable, in terms of accuracy.

4. Experiments

In order to evaluate the proposed method, two case studies were conducted on the detection and mapping of algal bloom occurrence in water bodies. The first was Lake Erie, located in Ohio, USA. A portion of Lake Taihu, located in China, was the second study area.

The assessment and validation of the proposed method were conceived by choosing dates and sensors according to the availability of ground truth data provided by the National Centers for Environmental Information of USA (Burtner et al. 2019, 2020) and the National Earth System Science Data Center, National Science and Technology Infrastructure of China (Ma 2016, 2017). Correlating its spatial and temporal occurrences, the definition of anomalous samples was based on the chlorophyll concentration >5.6 g/L, total phosphorus >21.7 g/L and Secchi depth <3.0 m for Lake Erie (Chapra and Dobson 1981; Kasich, Taylor, and Butler 2014) and chlorophyll

>20 g/L for Lake Taihu (Xu et al. 2015; Wang et al. 2019). Regular samples comprised data not covered by the rules in either case.

The results obtained were compared to the methods proposed in Gandhi et al. (2015) and Jia, Zhang, and Dong (2019), previously discussed in Section 2.4, which detect the occurrence of algal blooms based on thresholds for NDVI (Equation (4)) and FAI (Equation (5)) spectral indices, respectively. Overall accuracy, the kappa coefficient of agreement (Bishop et al. 1977) and F1-Score (van Rijsbergen 1979) were computed from reference samples identified on distinct dates. Also, hypothesis tests derived from the kappa measure were performed to evaluate the significance of the results. Furthermore, the analysed methods were compared on the basis of false/true positives/negatives occurrence ratios.

Additionally, the *spanning period* of 180 days, previously discussed in Section 3.2, was considered in the following experiments. The same inputs and analysis were tested through a modified version of the proposal using an RF classifier, comparing its performance against the OC-SVM method.

The experiments were performed on a computer with an AMD Ryzen 9 3900X 12-core processor, and 32 GB of RAM running the Ubuntu Linux version 20.04 operating system.

4.1. Study areas and data

As mentioned at the beginning of Section 4, two case studies were conducted. Regions referring to parts of Lake Erie (Figure 2) and Lake Taihu (Figure 3) determined the respective study areas.

Lake Erie is one of the five Great Lakes in North America and has the eleventh largest surface area in the world; it is located between Ohio, Michigan, Pennsylvania, and New York states and covers an area of 25,657 km². Lake Erie has a total volume of 484 km³ and an average and maximum depth of approximately 19 m and 64 m, respectively (Bolsenga and Herdendorf 1993). Historically, severe algal blooms have been recorded since the '70s, predominantly during the summer, peaking between August and September (Stumpf et al. 2012). Through Figure 2, it is also possible to observe the water quality monitoring stations belonging to the Great Lakes Environmental Research Laboratory (GLERL), a multidisciplinary environmental research laboratory linked to the National Oceanic and Atmospheric Administration (NOAA).

The third-largest lake in China, Lake Taihu is a source of water for 40 million people. Recently, high algal blooms affected more than 4 million residents and industries in the region (Duan et al. 2015). According to Gao et al. (2020), it has an area of 2,338 km² and an average depth of 1.9 m. According to the authors, the period of rain begins in April and extends until July. However, the lake faces its driest period between February and March.

According to the discussions held in Section 3, the algal bloom detection process uses images obtained from the Landsat-8 OLI and MODIS MOD09GA.006 sensors. This detection process, however, only uses one sensor at a time. Such images have 30 and 500 m spatial resolutions, respectively, and spectral bands ranging from blue to short-wave infra-red wavelengths.

After analysis of the good visibility images obtained by matching *in situ* data with the passage of the Landsat and MODIS satellites (i.e. less than 50% of cloud presence), and motivated by the temporal variation in algal bloom occurrence over Lake Erie, the following images were selected: June 3rd, 2019 – moderate; July 1st, 2019 – very high; August 19th, 2019 – very high; September 24th, 2019 – high (MODIS MOD09GA.006); and September 21st, 2015 – moderate (Landsat-8 OLI). Figure 4 illustrates the images by sensor of origin, highlighting the dates and validation samples considered for the occurrence or not of anomalies (that is, algae proliferation). The yellow arrows point to the exact location where the samples were selected, based on the NOAA/GLERL monitoring station locations. Table 1 summarises the number of pixels and polygons selected, as well as the total of training instants.

For the study area on Lake Taihu, two images matching the ground truth dataset and the passage of Landsat satellite were acquired: September 13th, 2016 – moderate; and December 2nd, 2016 – moderate (Figure 5(c,d)). The yellow arrows point to the exact location where the samples were

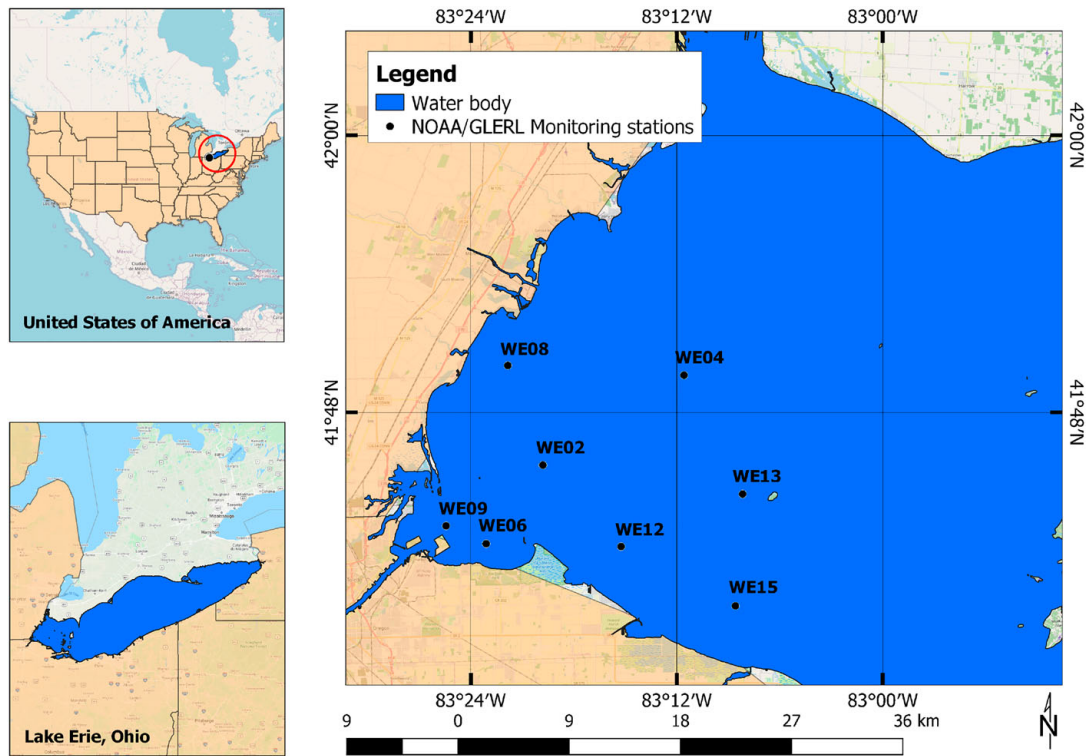


Figure 2. Spatial location of the study area 1 – Lake Erie, Ohio, USA.

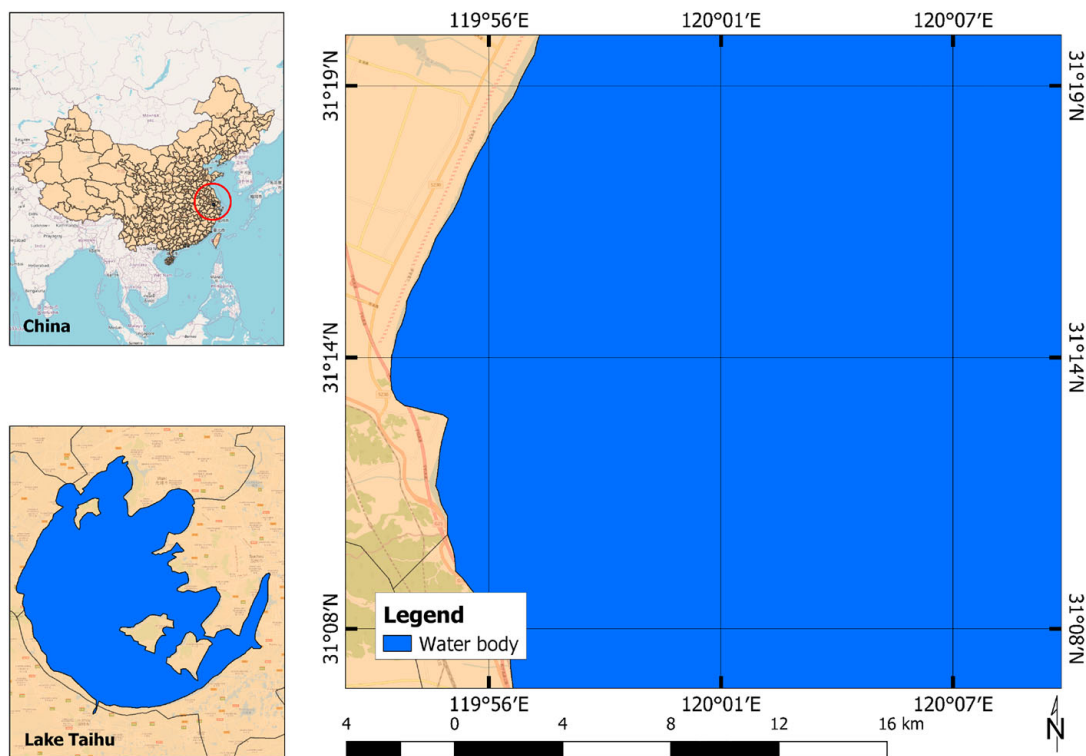


Figure 3. Spatial location of the study area 2 – Lake Taihu, China.

selected, based on the mapping of the chlorophyll concentration for days September 13th and December 2nd (Figure 5(a,b), respectively) provided by the National Earth System Science Data Center, National Science and Technology Infrastructure of China (Ma 2016, 2017). The thresholds used to determine whether blooms occurred are discussed in Section 4. Reference samples related to the occurrence or not of anomalies, as well as the total of training instants, are expressed in Table 2.

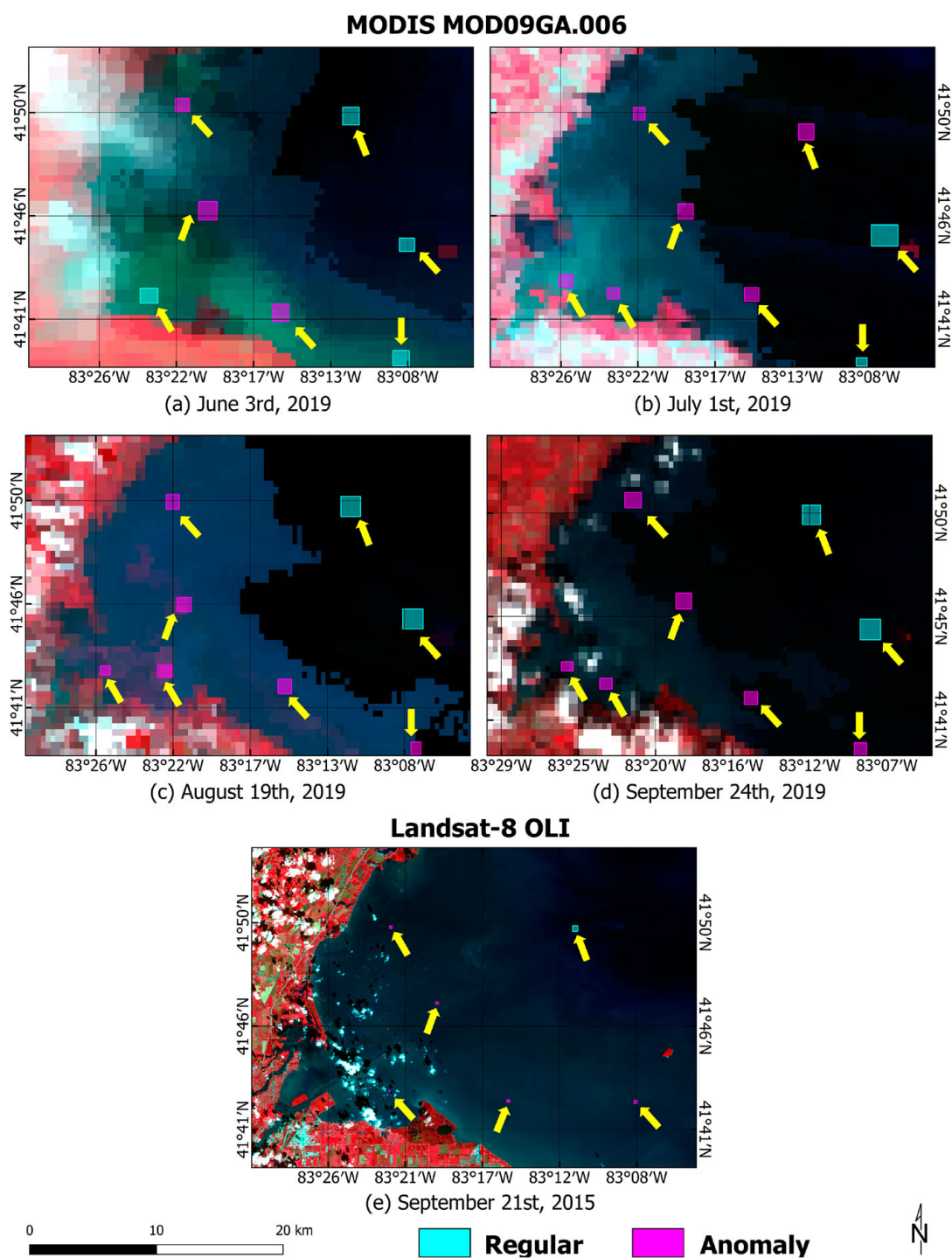


Figure 4. Images in false-color composition (NIR, red and green bands) from the Lake Erie study area for each considered instant. Regular and anomaly samples are identified by cyan and magenta polygons, respectively.

Table 1. Summary of regular and anomaly reference samples related to Lake Erie for the considered instants.

	Pixels/Polygons				
	MODIS MOD09GA.006				Landsat-8 OLI
	2019				
	Jun. 3rd	Jul. 1st	Aug. 19th	Sep. 24th	
Regular	29/4	20/2	31/2	27/2	266/1
Anomaly	27/3	37/6	35/6	34/6	281/5
Training instants	105	104	99	104	24

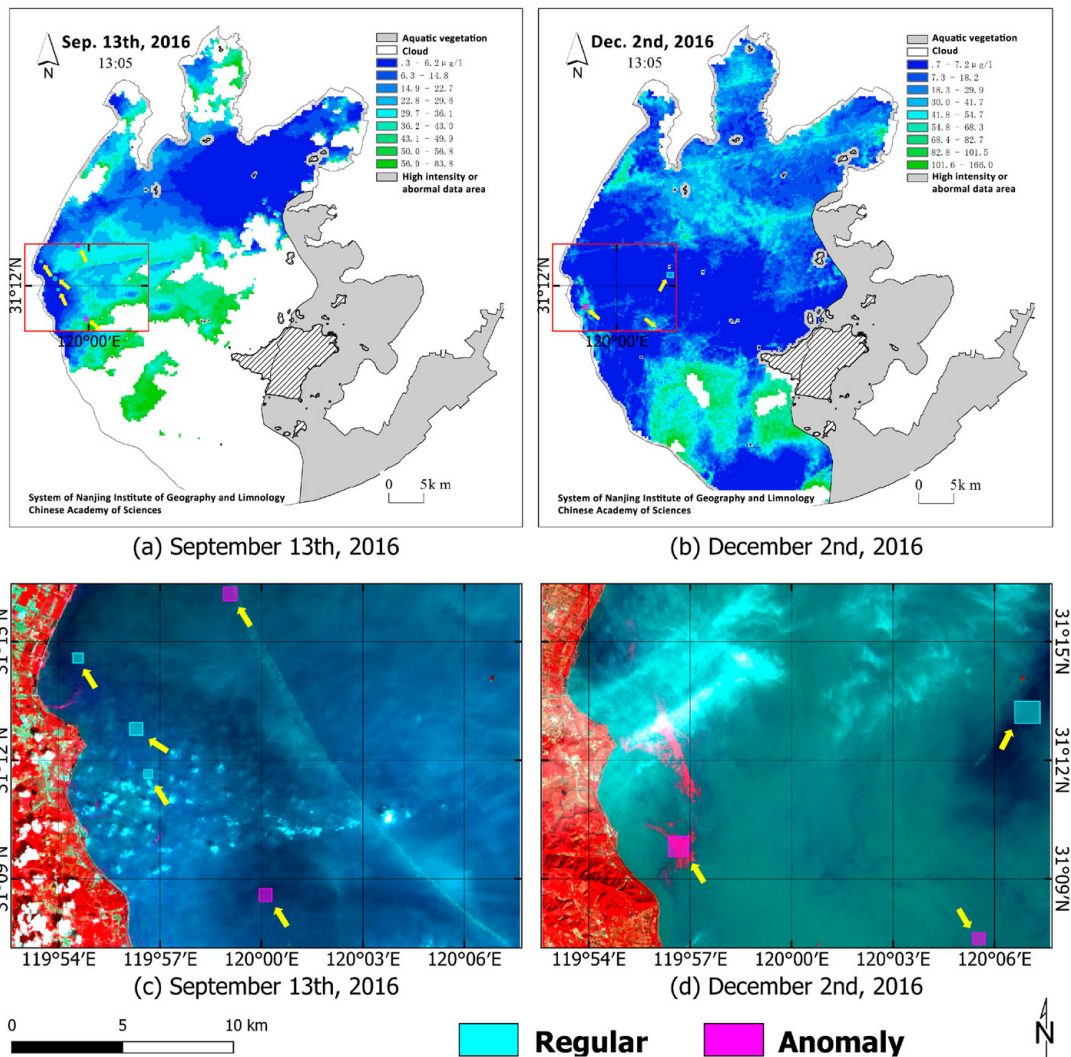


Figure 5. Images in false-color composition (NIR, red and green bands) from the Lake Taihu study area for each considered instant. Regular and anomaly samples are identified by cyan and magenta polygons, respectively. At the top, Chl-*a* concentration maps from the National Earth System Science Data Center, National Science and Technology Infrastructure of China are shown.

5. Results and discussion

Taking the case study of Lake Erie as a starting point, anomaly mapping was obtained for each considered date. Figure 6 shows the accuracy values associated with each tested method, expressed in terms of overall accuracy, kappa coefficient and F1-Score, and computed from the reference samples.

Regarding the first image from June 3rd, 2019, the proposal outperforms its competitors. Concerning the second date (July 1st, 2019), the proposal again demonstrates higher accuracy, delivering the first and second-best results with the application of the OC-SVM and RF algorithms, respectively. Similar to the first, the image from August 19th, 2019 presents the best result of the proposal. However, NDVI shows similar performance compared to the OC-SVM variation,

Table 2. Summary of regular and anomaly reference samples related to Lake Taihu for the considered instants.

	Pixels/Polygons	
	September 13th, 2016	December 2nd, 2016
Regular	915/3	1452/1
Anomaly	922/2	1424/2
Training instants	10	10

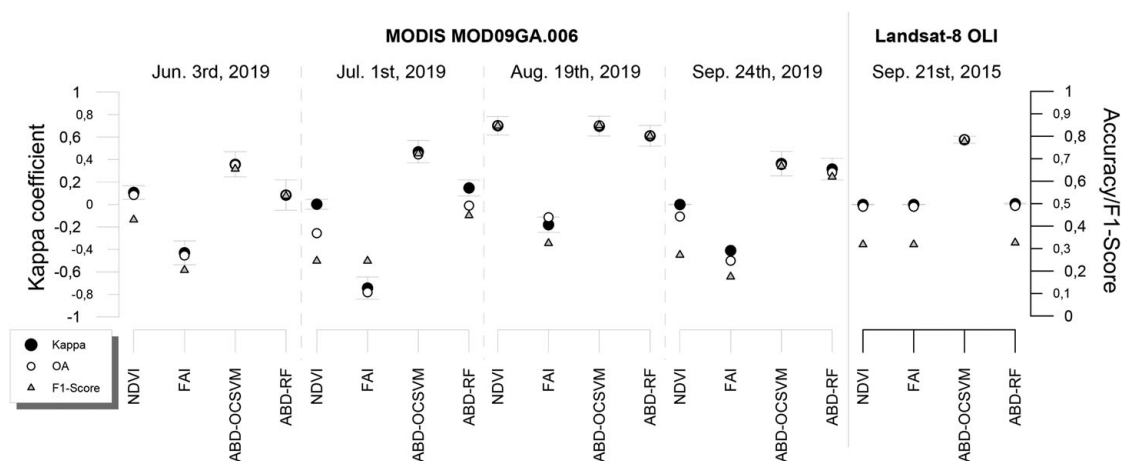


Figure 6. Accuracy of the analysed methods related to the Lake Erie study area at distinct instants. Error bars denotes the respective kappa standard deviation value.

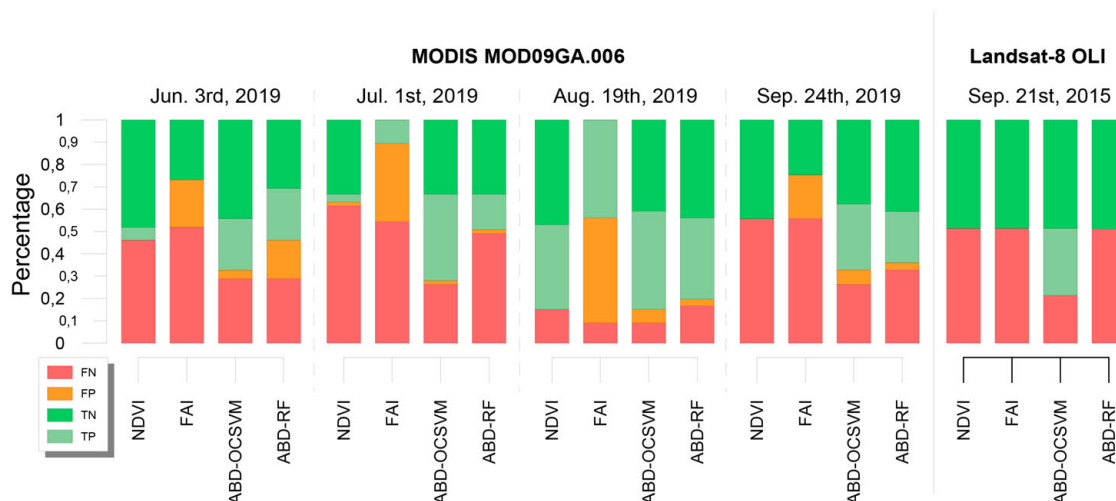


Figure 7. True(T)/False(F)-Positive(P)/Negative(N) ratio graph for the Lake Erie detection results.

producing the best result among the analysed indices. The last image, acquired on September 24th, 2019 with the MODIS sensor MOD09GA.006, also demonstrates the superiority of the proposal's classification, with OC-SVM and RF showing the first and second-best results, respectively. On this date, the performance of competitors was still low, with FAI exhibiting the worst metrics. The Landsat-8 OLI sensor, the single available image for this study area and acquired on September 21st, 2015, presents the best performance with ABD-OCSVM variance.

Given the above results, the proposal using the OC-SVM algorithm showed the most regularity among its competitors. In order to accept or reject the superiority of its performance, hypothesis tests comparing kappa values were performed (Congalton and Green 2009). It was observed that the proposal is statistically superior at 1% significance (i.e. 99% confidence), except for the image from August 19th, where NDVI, ABD-OCSVM and ABD-RF are statistically equivalent. This behaviour is also true for the image from September 24th with ABD-OCSVM and ABD-RF. Figure 7 shows the True/False-Positive/Negative ratio graph for the first study area, where the acronyms FN, FP, TN and TP relate to false-negative, false-positive, true-negative and true-positive, respectively. The term 'positive' refers to the occurrence of anomalies and 'negative' refers to regular pixels/regions. First, it should be noted that among the competitors, the ABD-OCSVM proposal delivers the lowest FN rate for all compared dates. The opposite is observed with the NDVI and FAI indices, which present high rates of FN throughout the evaluated period. Although the proposal using the

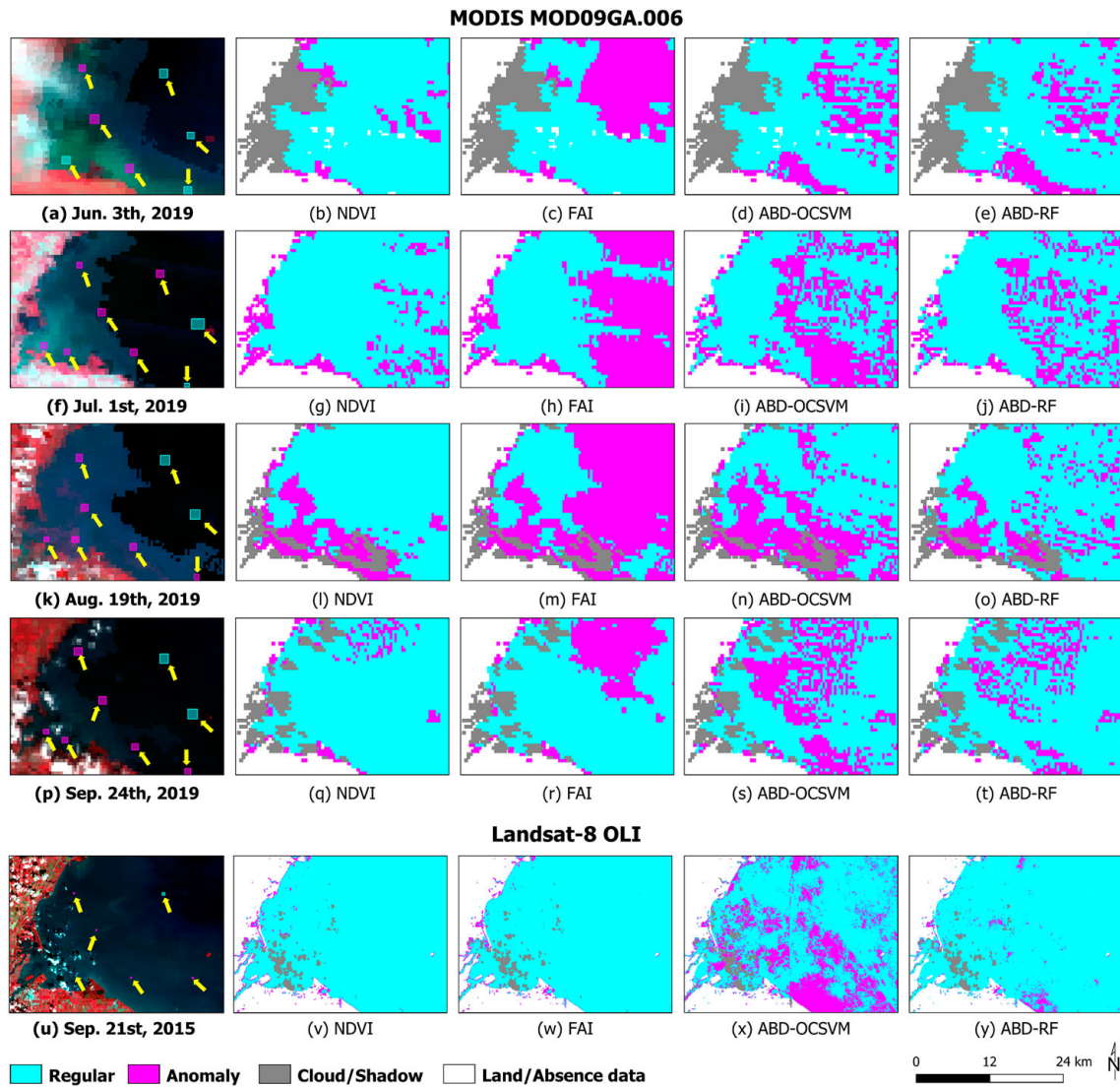


Figure 8. Mapping results obtained by the analysed methods for the Lake Erie study area related to the MODIS MOD09GA.006 (a)–(t) and Landsat-8 OLI (u)–(y) sensors. Regular, anomaly, cloud and land/absence data regions are denoted in cyan, magenta, grey and white, respectively.

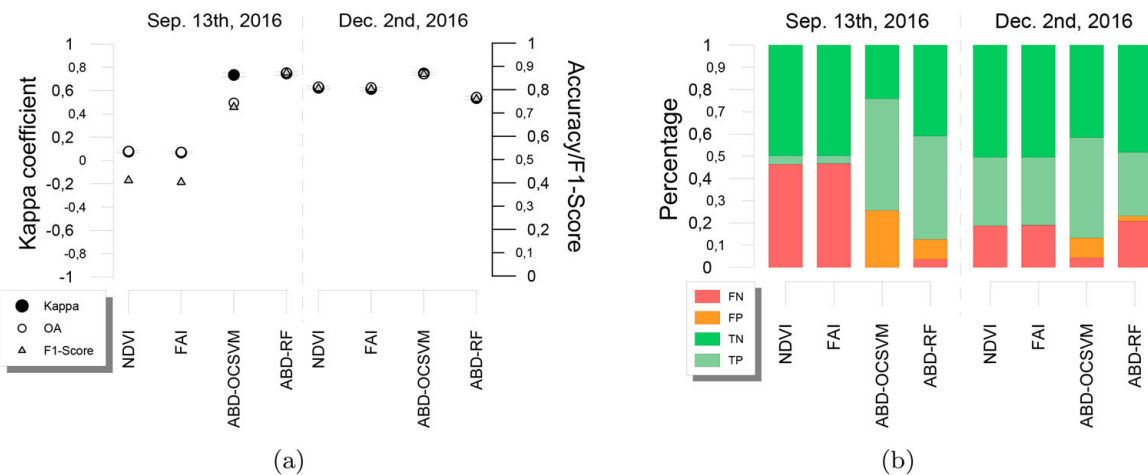


Figure 9. Images considering the performance obtained by the tested methods for the Lake Taihu study area: (a) kappa coefficient/Accuracy/F1-Score of the analysed methods at distinct instants. Error bars denotes the respective kappa standard deviation value. (b) True(T)/False(F)-Positive(P)/Negative(N) ratio graph.

RF algorithm presents similar results of TP compared with the ABD-OCSVM variation for most of the evaluated dates, there is a decrease in the detection of TN and an increase in FP on June 3rd, 2019 for the model in question. These results indicate that competitors overestimate the occurrence of regular pixels in the tested samples, presenting high values of FN and ignoring the occurrence of algae in all images. As shown in the previous analysis, ABD-OCSVM has the most regular detection.

Figure 8(a,f,k,p,u) represent the areas of interest and polygons of *in situ* samples for the evaluated dates (i.e. yellow arrows). Alternatively, Figure 8(b–e,g–j,l–o,q–t,v–y) refer to the detection results delivered by each one of the tested methods.

Regarding Figure 8(b,c) from the MODIS NDVI and FAI first dates, respectively, a comparison with *in situ* samples (Figure 8(a)) shows an overestimation of FN by both approaches, which is corroborated by Figure 7. Concerning the ABD-RF model, the opposite behaviour is true as it overestimates FP. Regarding the image from July 1st, 2019, ABD-OCSVM shows the best performance, followed by ABD-RF. For the August 19th, 2019 image, the FAI indice delivered high FP compared with NDVI, ABD-OCSVM and ABD-RF, the last three being statistically equivalent. Figure 8(p–t) (MODIS last image), show high FN delivered by spectral indices and corroborate the detection equivalence and superiority of the ABD-OCSVM and ABD-RF variations.

Finally, Figure 8(u–y) represent the only acquired image for Lake Erie using the Landsat-8 OLI sensor. It demonstrates the superiority of the proposal, having its variations showing lower FN compared to tested indices.

Regarding the second study area through Figure 9(a), the superior performance of the proposal variants, in relation to NDVI and FAI thresholds, is evident for both dates (September 13th and December 2nd, 2016). This analysis is confirmed by Figure 9(b), where the image acquired in September 13th, 2016 shows low FN and FP performance by the ABD-RF model, in contrast with the high FN proportion detected by spectral indices. In the same direction, the ABD-OCSVM model showed low FN values for December 2nd, 2016. In this image, however, NDVI and FAI obtained results equivalent to the proposal's RF algorithm.

Visual analysis of the second area's results is possible through Figure 10(a–j). In this sense, Figure 10(a,f) represent the areas of interest selected according to ground truth data (i.e. yellow arrows). From another angle, Figure 10(b,c) confirm the non-detection of anomalous regions found in this image by NDVI and FAI indices, validating the results from the True/False-Positive/Negative graph. However, Figure 10(i) stands out, and it is possible to observe the detection of a vast anomalous region by the ABD-OCSVM variant, corroborated by samples from Figure 10(f).

Lastly, hypothesis tests derived from the kappa coefficient confirm the superiority of the ABD-OCSVM and ABD-RF models, being statistically superior to the spectral indices with a significance

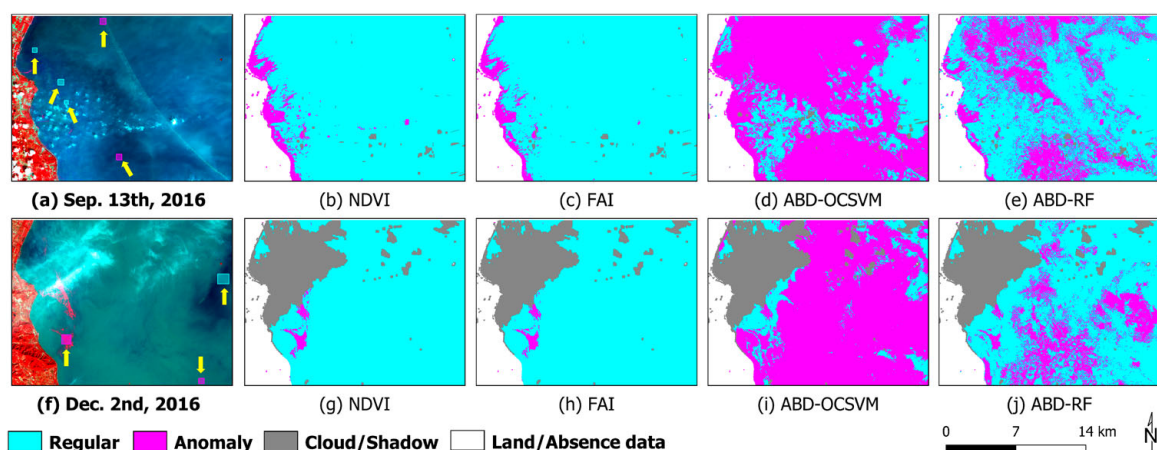


Figure 10. Mapping results obtained by the analysed methods for the Lake Taihu study area. Regular, anomaly, cloud and land/absence data regions are denoted in cyan, magenta, grey and white, respectively.

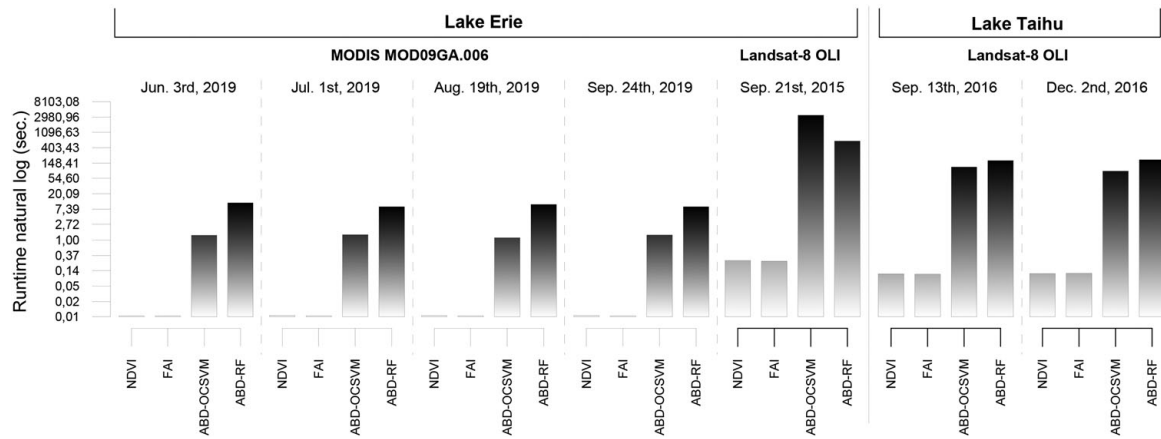


Figure 11. Run-time of analysed methods for both study areas.

of 1%. However, the ABD-RF variant is superior to the ABD-OCSVM model for the first date (September 13th, 2016) and the opposite is true for the second date, where ABD-OCSVM presents a better result than ABD-RF.

Although the proposal has considerably higher computational costs (as shown in Figure 11), it has the advantage of considering anomalies temporal behaviour and being an unsupervised

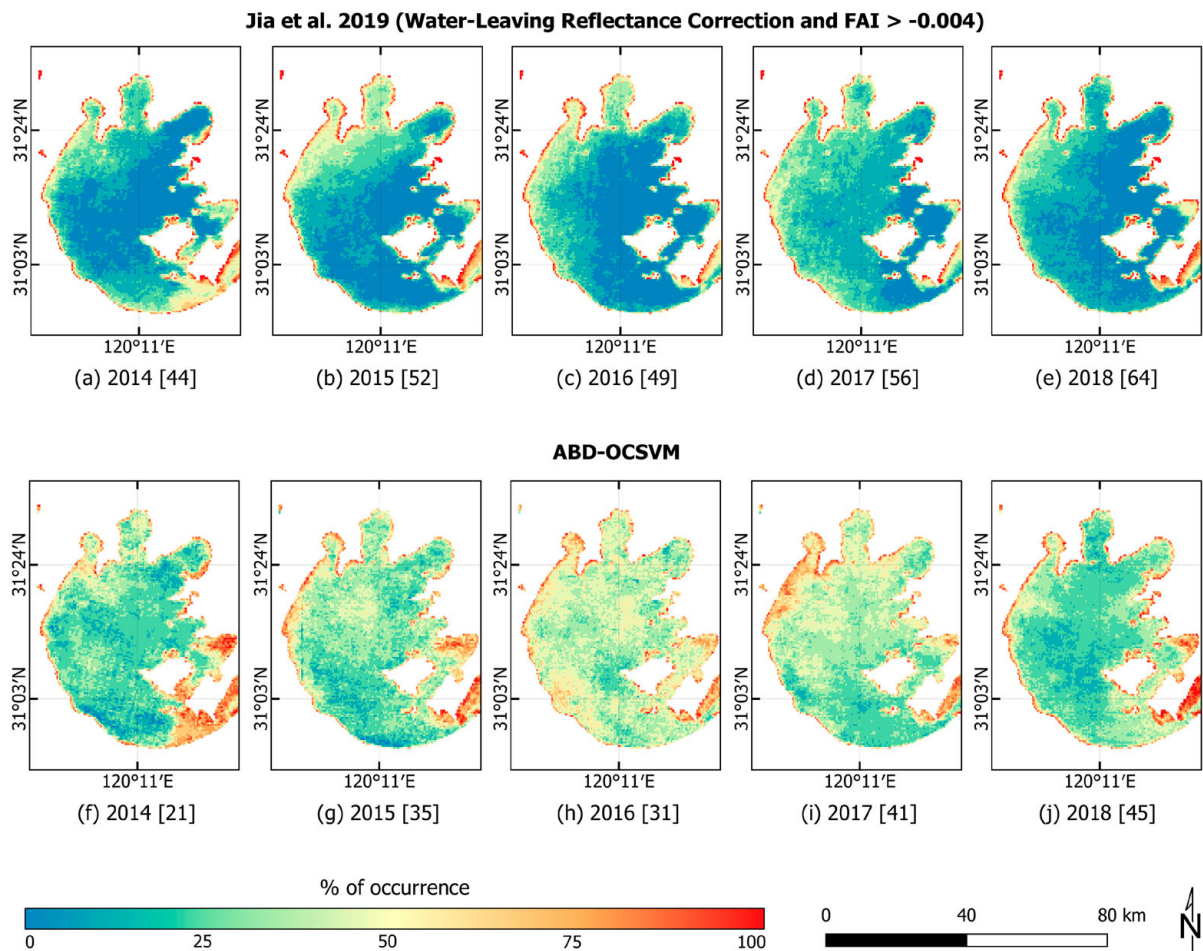


Figure 12. Annual mapping of cyanobacteria blooms occurrence in the Lake Taihu from 2014 to 2018 using multi-temporal series of images extracted from GEE API and MODIS MOD09GA.006 sensor: (a)–(e) algal bloom coverage with water-leaving reflectance correction (Shenglei et al. 2016; Wang et al. 2018) and FAI > -0.004 (threshold proposed by Jia, Zhang, and Dong 2019). (f)–(j) algal bloom coverage using ABD-OCSVM algorithm. Brackets show the total instants mapped by each year.

approach, where no manual selection of training data is needed. The proposal creates datasets larger than millions of pixels, mainly in the parameter optimisation and decision rule modelling. These sets are filtered and automatically obtained, reaching 20 or even 100 instants (MODIS MOD09GA.006 sensor), depending on the spanning period selected. To minimise the impact of network resource and consequently reduce the computational cost, a caching system was adopted, storing the images according to the area of interest locally. During the classifier construction, several images are processed. In this step, acquisition time is therefore reduced, if already in the disk.

5.1. Real-world application: 5-year mapping of the spatial occurrence of cyanobacterial blooms in Lake Taihu

In their study, Jia, Zhang, and Dong (2019) proposed a workflow responsible for annually mapping the occurrence of cyanobacteria in Lake Taihu in addition to the $FAI > -0.004$ threshold (Section 2.4). With results obtained using the algorithm implemented via the GEE platform (Jia 2019), the authors discuss the spatio-temporal patterns of the mentioned phenomenon and its annual characteristics, as well as the effects of environmental factors on the proliferation of these bacteria between the years 2000 and 2018. They also highlight the application of the water-leaving reflectance correction, recommended for Chinese inland waters greater than 25 km^2 (Shenglei et al. 2016; Wang et al. 2018).

To demonstrate the application of the ABD-OCSVM model in a real case study related to algae problems, the algorithm proposed by Jia, Zhang, and Dong (2019) was replicated, resulting in an annual mapping of algae blooming in Lake Taihu between 2014 and 2018 (Figure 12(a–e)). Simultaneously, respecting the aforementioned spatiality and temporality, the results of the proposal are shown in Figure 12(f–j).

In addition to the higher concentration verified by the ABD-OCSVM mapping compared to the algorithm proposed by Jia, Zhang, and Dong (2019), algae blooms occur in the central region of the lake, which is not widely verified in the first model. However, there is a spatial compatibility between both algorithms, especially in the most extreme regions, where the occurrence of the phenomenon is close to 100%.

6. Conclusions

In summary, regarding the problem of algal blooms and their adverse effect, whether related to health (i.e. poisoning, water quality) or economic losses, it is essential to develop studies aimed at monitoring it. Given this motivation, an algorithm based on anomaly detection approaches to detect algal bloom occurrence in inland water, supported by spectral indices and the OC-SVM model, was proposed. In order to validate it, two case studies were conducted with *in situ* measurements. In this direction, comparisons with similar proposals based in NDVI and FAI indices thresholding, were also included. The proposal was modified to be used with an RF classifier, comparing its results against the OC-SVM approach. Finally, to demonstrate the proposed algorithm in a real-world application and compare it with other studies, a 5-year mapping of algal bloom occurrence in Lake Taihu was presented, showing its spatial behaviour over time.

The results showed that the proposed method using the OC-SVM algorithm has higher accuracy levels than the analysed competitors, showing averages for kappa coefficient, overall accuracy and F1-Score of 68%, 87% and 86%, respectively; against 47%, 78% and 75% for ABD-RF; 28%, 67% and 57% for NDVI; and –4%, 53% and 46% for FAI.

Furthermore, it has been observed that the tested methods overestimate the non-occurrence of algal blooms (i.e. regular regions). On the other hand, it should be mentioned that the proposal is highly dependent on network connection and images with good visibility (i.e. less than 50% of clouds) in the training set and requires a higher computational effort. Still, given the lack of ground truth data, it was not possible to validate the behaviour of the proposal in regions with the presence

of snow or ice. As mentioned before, sudden changes in the target's spectral response can be detected as anomalies. The opposite is true when the phenomenon's occurrence is too subtle, being detected as regular. Additionally, a positive or negative performance variation is expected at different instants, both for the proposal and for the evaluated indexes.

Future perspectives for this study include (i) the development of alternative strategies to reduce computational cost; (ii) evaluate the behaviour of other spectral indices in the modelling process in order to extend the application for the proposed algorithm; (iii) implementation of moving average based strategies in the modelling process.

Acknowledgments

The authors would like to thank the National Centers for Environmental Information of USA (<https://www.ncei.noaa.gov>) and the National Earth System Science Data Center, National Science and Technology Infrastructure of China (<http://www.geodata.cn>) for providing data support for Lake Erie and Lake Taihu, respectively. Finally, they would like to thank the anonymous reviewers for their suggestions and comments that highly improved this work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (grants 2018/01033-3) for their financial support of this research.

ORCID

Pedro Henrique Moraes Ananias  <http://orcid.org/0000-0002-0924-1236>

Rogério Galante Negri  <http://orcid.org/0000-0002-4808-2362>

References

- Baba, N. 1981. "Convergence of a Random Optimization Method for Constrained Optimization Problems." *Journal of Optimization Theory and Applications* 33 (4): 451–461.
- Belgiu, M., and L. Drăguț. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31.
- Bergstra, J., and Y. Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *The Journal of Machine Learning Research* 13 (1): 281–305.
- Binding, C., T. Greenberg, G. McCullough, S. Watson, and E. Page. 2018. "An Analysis of Satellite-Derived Chlorophyll and Algal Bloom Indices on Lake Winnipeg." *Journal of Great Lakes Research* 44 (3): 436–446.
- Bishop, Y. M., S. E. Fienberg, P. W. Holland, R. J. Light, and F. Mosteller. 1977. "Book Review: Discrete Multivariate Analysis: Theory and Practice." *Applied Psychological Measurement* 1 (2): 297–306.
- Bolsenga, S. J., and C. E. Herdendorf. 1993. *Lake Erie and Lake St. Clair Handbook*. Detroit, MI: Wayne State University Press.
- Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.
- Bruzzzone, L., and C. Persello. 2009. "A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabelled Training Samples." *IEEE Transactions on Geoscience and Remote Sensing* 47 (7): 2142–2154.
- Burner, A., C. Kitchens, D. Fyffe, C. Godwin, T. Johengen, D. Stuart, R. Errera, D. Palladino, D. Fanslow, and D. Gossiaux. 2020. "Physical, Chemical, and Biological Water Quality Data Collected From a Small Boat in Western Lake Erie, Great Lakes From 2019-04-30 to 2019-10-07 (NCEI Accession 0209116)" (dataset). NOAA National Centers for Environmental Information. Accessed June 13, 2020. <https://accession.nodc.noaa.gov/0209116>.
- Burner, A., D. Palladino, C. Kitchens, D. Fyffe, T. Johengen, D. Stuart, D. Fanslow, and D. Gossiaux. 2019. "Physical, Chemical, and Biological Water Quality Data Collected From a Small Boat in Western Lake Erie, Great Lakes

- From 2012-05-15 to 2018-10-09 (NCEI Accession 0187718)" (dataset). NOAA National Centers for Environmental Information. Accessed June 13, 2020. <https://accession.nodc.noaa.gov/0187718>.
- Butler, H., M. Daly, A. Doyle, S. Gillies, S. Hagen, and T. Schaub. 2016. *The GeoJSON Format*. Internet Engineering Task Force (IETF).
- Cavanagh, J. 2013. *X86 Assembly Language and C Fundamentals*. Boca Raton, FL: CRC Press.
- Chapra, S. C., and H. F. Dobson. 1981. "Quantification of the Lake Trophic Typologies of Naumann (Surface Quality) and Thienemann (Oxygen) with Special Reference to the Great Lakes." *Journal of Great Lakes Research* 7 (2): 182–193.
- Chawla, I., L. Karthikeyan, and A. K. Mishra. 2020. "A Review of Remote Sensing Applications for Water Security: Quantity, Quality, and Extremes." *Journal of Hydrology* 585: Article ID: 124826.
- Congalton, R. G., and K. Green. 2009. *Assessing the Accuracy of Remotely Sensed Data*. Boca Raton, FL: CRC Press.
- Duan, H., S. A. Loiselle, L. Zhu, L. Feng, Y. Zhang, and R. Ma. 2015. "Distribution and Incidence of Algal Blooms in Lake Taihu." *Aquatic Sciences* 77 (1): 9–16.
- Gandhi, G. M., S. Parthiban, N. Thummalu, and A. Christy. 2015. "NDVI: Vegetation Change Detection Using Remote Sensing and GIS-a Case Study of Vellore District." *Procedia Computer Science* 57: 1199–1210.
- Gao, B.-C. 1996. "NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water From Space." *Remote Sensing of Environment* 58 (3): 257–266.
- Gao, Y., G. Zhu, H. W. Paerl, B. Qin, J. Yu, and Y. Song. 2020. "A Study of Bioavailable Phosphorus in the Inflowing Rivers of Lake Taihu, China." *Aquatic Sciences* 82 (1): 1.
- GEE-API. 2019. "Google Earth Engine API." Accessed 29 October 2019. <https://github.com/google/earthengine-api>.
- Ghatkar, J. G., R. K. Singh, and P. Shanmugam. 2019. "Classification of Algal Bloom Species From Remote Sensing Data Using An Extreme Gradient Boosted Decision Tree Model." *International Journal of Remote Sensing* 40 (24): 9412–9438.
- Glazer, A., M. Lindenbaum, and S. Markovitch. 2013. "q-ocsvm: A q-Quantile Estimator for High-Dimensional Distributions." In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Stateline, NV.
- Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. "Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone." *Remote Sensing of Environment* 202: 18–27.
- Grimmond, S. U. 2007. "Urbanization and Global Environmental Change: Local Effects of Urban Warming." *Geographical Journal* 173 (1): 83–88.
- Gu, Y., and K. Feng. 2013. "Optimized Laplacian SVM with Distance Metric Learning for Hyperspectral Image Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6 (3): 1109–1117.
- He, M., S. Pathak, U. Muaz, J. Zhou, S. Saini, S. Malinchik, and S. Sobolevsky. 2019. "Pattern and Anomaly Detection in Urban Temporal Networks." Preprint. arXiv:1912.01960.
- Houborg, R., M. F. McCabe, Y. Angel, and E. M. Middleton. 2016. "Detection of Chlorophyll and Leaf Area Index Dynamics From Sub-Weekly Hyperspectral Imagery." In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XVIII*, 999812. Vol. 9998. International Society for Optics and Photonics.
- Hu, C. 2009. "A Novel Ocean Color Index to Detect Floating Algae in the Global Oceans." *Remote Sensing of Environment* 113 (10): 2118–2129.
- Huang, X., Q. Lu, and L. Zhang. 2014. "A Multi-Index Learning Approach for Classification of High-Resolution Remotely Sensed Images Over Urban Areas." *ISPRS Journal of Photogrammetry and Remote Sensing* 90: 36–48.
- Jia, T. 2019. "Earth Engine Code – Long-Term Spatial and Temporal Monitoring of Cyanobacteria Blooms Using Modis on Google Earth Engine: A Case Study in Taihu Lake." Accessed 5 February 2020. <https://code.earthengine.google.com/e6c4d627ec7f4f0dbc1a4f7fdeb3bb3>.
- Jia, T., X. Zhang, and R. Dong. 2019. "Long-Term Spatial and Temporal Monitoring of Cyanobacteria Blooms Using Modis on Google Earth Engine: A Case Study in Taihu Lake." *Remote Sensing* 11 (19): 2269.
- Kasich, J. R., M. Taylor, and C. W. Butler. 2014. *Public Water System Harmful Algal Bloom Response Strategy*. Ohio Environmental Protection Agency. Columbus, OH.
- Khanna, S., A. Palacios-Orueta, M. L. Whiting, S. L. Ustin, D. Riaño, and J. Litago. 2007. "Development of Angle Indexes for Soil Moisture Estimation, Dry Matter Detection and Land-Cover Discrimination." *Remote Sensing of Environment* 109 (2): 154–165.
- Klemas, V. 2011. "Remote Sensing of Algal Blooms: An Overview with Case Studies." *Journal of Coastal Research* 28 (1A): 34–43.
- Li, Y., Y. Wang, C. Bi, and X. Jiang. 2018. "Revisiting Transductive Support Vector Machines with Margin Distribution Embedding." *Knowledge-Based Systems* 152: 200–214.
- Ma, R. 2016. "Lake Taihu Chlorophyll Inversion Product Data Set (2016)" (dataset). National Earth System Science Data Center, National Science and Technology Infrastructure of China. Accessed June 13, 2020. <http://www.geodata.cn/data/datadetails.html?dataguid=122425543649945>.
- Ma, R. 2017. "Lake Taihu Chlorophyll Inversion Product Data Set (2017)" (dataset). National Earth System Science Data Center, National Science and Technology Infrastructure of China. Accessed June 13, 2020. <http://www.geodata.cn/data/datadetails.html?dataguid=107063659544667>.

- Marzuoli, A., and F. Liu. 2019. "Monitoring of Natural Disasters Through Anomaly Detection on Mobile Phone Data." In *2019 IEEE International Conference on Big Data (Big Data)*, 4089–4098. IEEE.
- McKinney, W. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 51–56. Vol. 445. Austin, TX.
- Mishra, S., and D. R. Mishra. 2012. "Normalized Difference Chlorophyll Index: A Novel Model for Remote Estimation of Chlorophyll-a Concentration in Turbid Productive Waters." *Remote Sensing of Environment* 117: 394–406.
- Mountrakis, G., J. Im, and C. Ogole. 2011. "Support Vector Machines in Remote Sensing: A Review." *ISPRS Journal of Photogrammetry and Remote Sensing Society* 66 (3): 247–259.
- Muñoz-Mari, J., F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camp-Valls. 2010. "Semisupervised One-Class Support Vector Machines for Classification of Remote Sensing Data." *IEEE Transactions on Geoscience and Remote Sensing* 48 (8): 3188–3197.
- Negri, R. G., L. V. Dutra, and S. J. S. Sant'Anna. 2014. "An Innovative Support Vector Machine Based Method for Contextual Image Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 87: 241–248.
- Negri, R. G., A. C. Frery, W. Casaca, S. Azevedo, M. Araújo, E. Silva, and E. Alcântara. 2020. "Spectral-Spatial Aware Unsupervised Change Detection with Stochastic Distances and Support Vector Machines." *IEEE Transactions on Geoscience and Remote Sensing* 59 (4): 2863–2876.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (Oct): 2825–2830.
- Rastrigin, L. 1963. "The Convergence of the Random Search Method in the Extremal Control of a Many Parameter System." *Automaton & Remote Control* 24: 1337–1342.
- Rouse Jr., J. W., R. H. Haas, J. Schell, and D. Deering. 1973. "Monitoring the Vernal Advancement and Retrogradation (Green Wave Effect) of Natural Vegetation." *Third ERTS Symp. 1*: 309–317.
- Schölkopf, B., J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. 2001. "Estimating the Support of a High-Dimensional Distribution." *Neural Computation* 13 (7): 1443–1471.
- Schölkopf, B., R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. 2000. "Support Vector Method for Novelty Detection." In *Advances in Neural Information Processing Systems*: 582–588. Denver, CO.
- Shawe-Taylor, J., and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge University Press.
- Shenglei, W., L. Junsheng, Z. Bing, S. Qian, Z. Fangfang, and L. Zhaoyi. 2016. "A Simple Correction Method for the Modis Surface Reflectance Product Over Typical Inland Waters in China." *International Journal of Remote Sensing* 37 (24): 6076–6096.
- Shi, K., Y. Zhang, B. Qin, and B. Zhou. 2019. "Remote Sensing of Cyanobacterial Blooms in Inland Waters: Present Knowledge and Future Challenges." *Science Bulletin* 64 (20): 1540–1556.
- Song, W., J. Dolan, D. Cline, and G. Xiong. 2015. "Learning-based Algal Bloom Event Recognition for Oceanographic Decision Support System Using Remote Sensing Data." *Remote Sensing* 7 (10): 13564–13585.
- Stumpf, R. P., T. T. Wynne, D. B. Baker, and G. L. Fahnenstiel. 2012. "Interannual Variability of Cyanobacterial Blooms in Lake Erie." *PloS One* 7 (8): e42444.
- Sun, D., Y. Li, and Q. Wang. 2009. "A Unified Model for Remotely Estimating Chlorophyll-a in Lake Taihu, China, Based on SVM and in Situ Hyperspectral Data." *IEEE Transactions on Geoscience and Remote Sensing* 47 (8): 2957–2965.
- United Nations, I.S.f.D.R. 2015. *Global Assessment Report on Disaster Risk Reduction 2015: Making Development Sustainable: The Future of Disaster Risk Management*. UN.
- USGS. 2017. "Product Guide: Landsat 8 Surface Reflectance Code (LASRC) Product" (techreport). United States Geological Service. Accessed 1 August, 2019. <https://www.usgs.gov/media/files/land-surface-reflectance-code-lasrc-product-guide>.
- USGS. 2020. "MODIS/Terra Surface Reflectance Daily L2G Global 1 km and 500 m." Acesso em 5 mar. 2020. <https://lpdaac.usgs.gov/products/mod09gav006/>.
- Van Der Walt, S., S. C. Colbert, and G. Varoquaux. 2011. "The Numpy Array: a Structure for Efficient Numerical Computation." *Computing in Science & Engineering* 13 (2): 22.
- van Rijsbergen, C. J. 1979. *Information Retrieval*. 2nd ed. London: Butterworths. <http://www.dcs.gla.ac.uk/Keith/Preface.html>.
- van Rossum, G., and F. L. Drake. 2011. *The Python Language Reference Manual*. Godalming, UK: Network Theory Ltd.
- Verstraete, M. M., and B. Pinty. 1996. "Designing Optimal Spectral Indexes for Remote Sensing Applications." *IEEE Transactions on Geoscience and Remote Sensing* 34 (5): 1254–1265.
- Wang, S., J. Li, B. Zhang, E. Spyarakos, A. N. Tyler, Q. Shen, F. Zhang, T. Kuster, M. K. Lehmann, and Y. Wu, et al. 2018. "Trophic State Assessment of Global Inland Waters Using a Modis-Derived Forel-Ule Index." *Remote Sensing of Environment* 217: 444–460.

- Wang, M., M. Stokal, P. Burek, C. Kroeze, L. Ma, and A. B. Janssen. 2019. "Excess Nutrient Loads to Lake Taihu: Opportunities for Nutrient Reduction." *Science of the Total Environment* 664: 865–873.
- Watanabe, F., E. Alcantara, T. Rodrigues, L. Rotta, N. Bernardo, and N. Imai. 2018. "Remote Sensing of the Chlorophyll-a Based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita Reservoir, Brazil)." *Anais da Academia Brasileira de Ciências* 90 (2): 1987–2000.
- Wetzel, R. G. 2001. *Limnology: Lake and River Ecosystems*. 3rd ed. San Diego, CL: Academic Press.
- Xu, H., H. Paerl, B. Qin, G. Zhu, N. Hall, and Y. Wu. 2015. "Determining Critical Nutrient Thresholds Needed to Control Harmful Cyanobacterial Blooms in Eutrophic Lake Taihu, China." *Environmental Science & Technology* 49 (2): 1051–1059.
- Xue, J., and B. Su. 2017. "Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications." *Journal of Sensors* 2017: Article ID: 1353691.
- Yang, Y., and X. Liu. 1999. "A Re-Examination of Text Categorization Methods." In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42–49. Berkeley, CL.
- Yi, H.-S., S. Park, K.-G. An, and K.-C. Kwak. 2018. "Algal Bloom Prediction Using Extreme Learning Machine Models at Artificial Weirs in the Nakdong River, Korea." *International Journal of Environmental Research and Public Health* 15 (10): 2078.
- Zhang, Y., R. Ma, H. Duan, S. A. Loiselle, J. Xu, and M. Ma. 2014. "A Novel Algorithm to Estimate Algal Bloom Coverage to Subpixel Resolution in Lake Taihu." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7 (7): 3060–3068.
- Zhao, D. 2003. "Application of NDVI to Detecting Algal Bloom in the Bohai Sea of China from AVHRR." In *Ocean Remote Sensing and Applications*, 241–246. Vol. 4892. International Society for Optics and Photonics.

2.2 Artigo - “Forecasting algal bloom events in inland waters using remote sensing data and machine learning methods”

O artigo a seguir foi submetido para apreciação do corpo editorial do periódico ENVIRONMENTAL MODELLING & SOFTWARE em 23 de julho de 2021.

Forecasting algal bloom events in inland waters using remote sensing data and machine learning methods

ARTICLE INFO

Keywords:

Remote sensing
Algal bloom
Forecasting
Google Earth Engine
Classification

ABSTRACT

The monitoring of water quality and algal blooms has received considerable attention from the scientific community, as these can pose risks to the health of living beings. The use of Remote Sensing can minimize costs in the *in loco* analysis processes, as well as being a generating source of data. Additionally, Machine Learning techniques and concepts favor the development of solutions for environmental analysis and monitoring. In this context, supported by classification models and multitemporal data series, it is proposed an automated method able to predict algal blooms. Such a proposal uses Modis images and meteorological/climatic products acquired from the Google Earth Engine platform. Three case studies involving predicting the phenomenon in lakes Erie (USA), Chilika (India) and Taihu (China) demonstrate a median global accuracy of 95%. The computational cost comprises the main drawback of the proposal.

1. Introduction

Water quality is of vital importance for the Earth, mainly due to the recent increase in population and climate changes [11]. Thus, there is a growth in demand for both domestic and agricultural, increasing pressure on this resource and, consequently, on the environment [40]. Another consequence is the proliferation of cyanobacteria, responsible for severe damage to ecological structures and aquatic ecosystems [68]. This phenomenon, also known as Harmful Algal Blooms (HABs), implies severe risks to the behavior and health of living beings [53].

Several parameters are described by [11] as responsible for water quality, including suspended sediment, turbidity, total phosphorus, dissolved organic content, temperature, and Secchi disk. However, chlorophyll-a (Chl-a) stands out as the most used parameter [23]. Its extension is directly related to sudden changes in components such as surface temperature, wind speed, precipitation, water column stratification, and water flow direction [68, 53, 50].

[40] point to sustainable water resources management. Terrestrial monitoring stations are described as precarious, presenting outdated data and non-existent information. As highlighted by [11], they are not uniformly distributed and operate intermittently, primarily due to lack of investment and government interest.

As a way of getting around this limitation, [11] address the use of remote sensing images in water quality management. An example can be seen in [45], where the authors present an algorithm based on the Empirical Orthogonal Functions model to estimate the concentration of Chl-a in Lake Taihu, China. [2] make use of images from the Moderate Resolution Imaging Spectroradiometer (Modis) sensor to assess the ability of the Medium Resolution Continental Shelf model to predict high algal bloom events in a eutrophic coastal sea.

With the increase in the quantity and quality of data, technologies, and information from imaging sensors and Artificial Intelligence [72, 36], several studies are presented focusing on increasing the accuracy of Chl-a concentration estimation and prediction models of algal blooms. An example is a use of Support Vector Machine (SVM) and Landsat-8/OLI images to estimate Chl-a concentration in [73] lakes. Also, [3] propose a study to detect algal bloom occurrences in inland waters with One-class Support Vector Machine. Using Sentinel-2/MSI images and the Random Forest (RF) model, the study of [54] addresses the prediction of Chl-a in two small water bodies. Additionally, approaches using deep learning models [12, 31], which often accurately approximate nonlinear environmental parameters [72], have become popular. In their study, [5] uses a hybrid model of Convolutional Neural Network and Long Short-term Memory (LSTM) capable of predicting the concentration of Chl-a in a lake in Greece. Applying LSTM nets, [12] demonstrate a 1- and 4-day Chl-a concentration prediction model in the Geum River (South Korea). In a study published by [63], a multilayer perceptron is applied to generate a cyanobacterial prediction model.

However, it should be stressed that physical models, such as the one presented by [19], are strongly dependent on environment variables [72]. Similarly, [72] it also highlights that most of the models based on Remote Sensing data and Machine Learning concepts are trained with ground-truth/reference samples collected from the study area [72]. In

ORCID(s):

both cases, obtaining the needed information may be critical. Thus, efforts are needed in the development of tools that not depends on reference data.

In this context, by the use of multitemporal series of remote sensing images obtained by the Modis sensor, meteorological/climatic attributes and Machine Learning methods, this work proposes an algorithm able to forecast the occurrence of anomalies in aquatic environments. Three case studies related to the prediction of algal blooms in Lakes Erie (USA), Chilika (India) and Taihu (China) are carried to demonstrate the potential of the proposed method.

In summary, the main contributions of this research are:

- (i) A unsupervised method designed to predict anomalous behavior;
- (ii) The proposed method is modular and, therefore, flexible regarding the use of other classification models, in addition to those presented in the following formalization;
- (iii) A conceptual formalization that, after convenient changes and adaptations, can be applied to other environmental issues, in addition to the detection of anomalies in inland waters.

This article is organized as follows: Section 2 briefly analyzes data classification, machine learning models and spectral indices; Section 3 performs the formalization of the proposed algorithm; Section 4 describes the experiment, the study areas, the reference data, presents and discusses the obtained results; Finally, Section 5 presents conclusions about the proposal and studies developed.

2. Theory Rational

2.1. Preliminary notations

A more extensive and comprehensive discussion about the LSTM method and its parameters are found in [26]. Let $\mathcal{I}$ be an image remotely obtained by a sensor whose pixels $s \in \mathcal{S} \subset \mathbb{N}^2$ are associated to a attribute vector $\mathbf{x}_s = [x_{s;1}, \dots, x_{s;n}] \in \mathcal{X} \subset \mathbb{R}^n$. Assuming that a given set of images composes a multitemporal series with t distinct instants take from a same spatial domain $\mathcal{S}$, the notation $\mathcal{I}^{(\ell)}$, with $\ell = 1, \dots, t$, is adopted as a spatio-temporal representation of these data.

The image classification process comprises associating a class $\omega_k \in \Omega = \{\omega_1, \dots, \omega_c\}$ to each $s \in \mathcal{S}$ by applying a function $G : \mathcal{X} \rightarrow \mathcal{Y}$ over $\mathbf{x}_s$, where $\mathcal{Y} = \{1, \dots, c\}$. The different classification methods proposed in the literature differ in how G is modeled. When modelling of this function is performed in a supervised way, it is used information from a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \dots, m\}$, where $(\mathbf{x}_i, y_i)$ indicates that $\mathbf{x}_i$ is assigned to the class ω_a if the scalar $y_i \in \mathbb{N}$ is equal to a . Conveniently, we denote by $\mathcal{C}^{(\ell)}$ the result that arises when applying G on each position s of $\mathcal{I}^{(\ell)}$.

2.2. Machine learning models

Given the importance of image classification in several applications involving Remote Sensing data, the development of more efficient and accurate methods has become a constant challenge [30, 66]. Among various examples, consolidated methods such as Support Vector Machine (SVM) [14], Random Forest (RF) [8] and the recent models based on neural networks have proven potential in the context of Remote Sensing applications.

Introduced by Vladimir Vapnik [62], the SVM method comprises a supervised learning algorithm that aims to separate classes though a surface $g(\mathbf{x}) = K(\mathbf{x}, \mathbf{w}) + b$ whose margin is maximum [14, 41]; $\mathbf{w}$ and b are parameters that determine the separation surface, and $K(\cdot, \cdot)$ is a kernel function [52] conveniently adopted according to the complexity of the classification problem.

As described by [9], starting from a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, +1\} : i = 1, \dots, m\}$, where $y_i = \pm 1$ indicates membership between two classes, the training of the SVM method comprises the calculation of the parameters $\mathbf{w}$ and b after solve the following optimization problem:

$$\begin{aligned} \max_{\alpha} & \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \\ \text{s.t.} & \begin{cases} \sum_{i=1}^m y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C \\ i = 1, \dots, m \end{cases} \end{aligned} \quad (1)$$

where $\alpha_i \in \mathbb{R}$ are Lagrange multipliers and $C \in \mathbb{R}_0^+$ is a penalty factor applied to misclassifications. Regarding the kernel, the radial basis function (RBF) $K(\mathbf{x}_i, \mathbf{w}) = e^{-\gamma \|\mathbf{x}_i - \mathbf{w}\|^2}$, with $\gamma \in \mathbb{R}_0^+$, is highlighted as a convenient option.

The Lagrange multipliers obtained when solving Equation 1 allows then define the classification rule $G(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \right)$. More details are found in [15] and [57].

Proposed by [8], the RF method comprises a classification rule $G : \mathcal{X} \rightarrow \mathcal{Y}$ resulting from an ensemble of decision trees [16]. Formally, from a training set $\mathcal{D}$, it is made n_{est} replications with same cardinality by bootstrap sampling. For each replica, a subset with up to n_{att} attributes are randomly considered and then used to train a single decision tree. Parameters such as the maximum depth (p_{depth}), and the minimum examples for splitting and per leaf (p_{split} and p_{leaf} , respectively) need to be tuned before the training process. A detailed discussions about these parameters are found in [8]. After training each tree $G_k : \mathcal{X} \rightarrow \mathcal{Y}$, with $k = 1, \dots, n_{est}$, a given unlabeled attribute vector $\mathbf{x}$ is classified according to a class $\omega_a \in \Omega$ whose agreement among the n_{est} decision trees is maximum, that is:

$$G(\mathbf{x}) = \arg \max_{a \in \{1, \dots, c\}} \left\{ \sum_{k=1}^{n_{est}} \delta_a(G_k(\mathbf{x})) \right\} \quad (2)$$

where $\delta_a(G_k(\mathbf{x})) = 1$ when $G_k(\mathbf{x}) = a$, otherwise, $\delta_a(G_k(\mathbf{x})) = 0$.

Concurrently with the previous methods, models based on neural networks concepts emerge as an effective alternative for the classification of Remote Sensing images [72]. Such models are characterized by their high generalization ability when dealing with uncorrelated data.

Among dozen neural network proposals, the Long Short-term Memory (LSTM) comprises a model that inserts and exploits the temporal effects over the learning process. The diagram depicted in Figure 1 yields to understand this method. Initially, we assume $\mathbf{x}_i$, with $i = 1, \dots, m$, as a set of patterns (i.e., attribute vectors) ordered by its index i and which are sequentially submitted to concatenation, element-wise multiplication and vector sum ($\odot$, $\otimes$ and $\oplus$) and then applied to sigmoid (σ) and hyperbolic tangent (φ) functions.

Also, worth highlighting some elements in the training process, for example, $\mathbf{c}_{i-1}$ and $\mathbf{c}_i$ as a vector with the “previous state” and “current” of the network; $\mathbf{f}_i$ a vector comprising “forgetting factors” for each component of the input vector; $\mathbf{i}_i$ as a “signal modulation” of input information; and $\mathbf{h}_{i-1}$ and $\mathbf{h}_i$ as the network input and output vectors in the “previous” and “current” states.

After successively insert the patterns $\mathbf{x}_i$, and consider the respective predictions $\mathbf{h}_i$, the training process stops when occurs the network convergence. It is worth noting that the classification is performed by a softmax function [18] adjusted using the outputs $\mathbf{h}_i$, which class indicators y_i are known from $(\mathbf{x}_i, y_i) \in \mathcal{D}$. According to this method, the combination between the described network and the softmax function comprise the classification model G . A more extensive and comprehensive discussion about the LSTM method and its parameters are found in [26].

2.3. Algal bloom detection through spectral index thresholding

Spectral indices can be understood as an attribute extraction process for Remote Sensing images whose purpose is to aid in the identification of specific targets. Over the years, several studies have proposed and used spectral indices, for example, for the characterization of water bodies [37, 20, 69] and vegetation health [49, 27, 33, 67]. In the context of Chl-a estimation, the detection of algal blooms has been performed through the adoption of rigid thresholds. In [74] the *Normalized Difference Vegetation Index* (NDVI) [49] is used to identify algae when such index achieve values greater than -0.15 . Similarly, in [69], the *Modified Normalized Difference Water Index* (MNDWI) [25] is used to characterize algal blooms when such index is less than zero. The *Surface Algae Bloom Index* (SABI) [1] and *Floating Algae Index* (FAI) [42] are proposals exclusively for detecting algae, whose occurrence is verified when such indices surpass the values -0.1 and -0.004 , respectively.

Table 1 summarizes the different above-mentioned spectral indices and the respective thresholds that characterize the occurrence of algae. Regarding the expressions, x_{Red} , x_{Green} , x_{Blue} , x_{NIR} and x_{SWIR} represent the spectral behavior measured in the red, green, near-infrared and short-wave infrared wave spectral bands, respectively; λ_{Red} , λ_{NIR} and λ_{SWIR} are the midpoint of the red, near-infrared, and short-wave infrared wavelength bands.

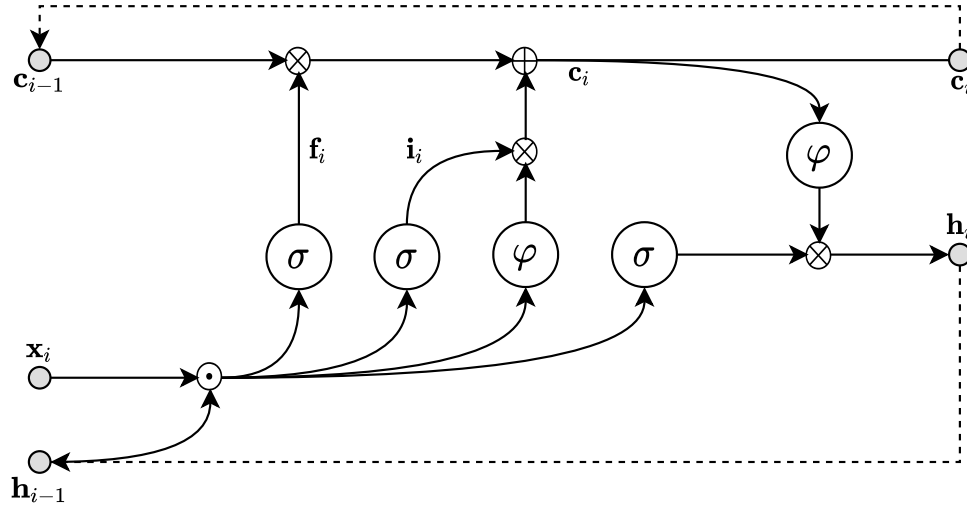


Figure 1: An overview of the LSTM architecture.

Table 1

Summary of spectral indices and thresholds used for algal bloom detection.

Spectral Index	Expression	Algae threshold	Reference
NDVI	$\frac{x_{NIR} - x_{Red}}{x_{NIR} + x_{Red}}$	> -0.15	[74]
MNDWI	$\frac{x_{Green} - x_{SWIR}}{x_{Green} + x_{SWIR}}$	< 0	[70]
SABI	$\frac{x_{NIR} - x_{Red}}{x_{Blue} + x_{Green}}$	> -0.1	[1]
FAI	$\frac{x_{NIR} - [x_{Red} + (x_{SWIR} - x_{Red}) \times \frac{\lambda_{NIR} - \lambda_{Red}}{\lambda_{SWIR} - \lambda_{Red}}]}{x_{Blue} + x_{Green}}$	> -0.004	[42]

3. Anomalous behaviour forecasting

This section introduces a novel proposal for predicting algal blooms in inland waters. In the following discussions, the emergence of this phenomenon will be called “anomaly”; otherwise, we adopt the term “regular”. Section 3.1 presents a conceptual formalization of the proposal. Section 3.2 gives relevant details regarding the proposal’s implementation.

3.1. Conceptual formalization

Based on Machine Learning concepts and multitemporal data obtained by Remote Sensing, a method for predicting anomalous events in aquatic environments is proposed. For this reason, this method will be called *Anomalous Behavior Forecasting* (ABF). Figure 2 illustrates the main components of this proposal, which are structured in four main blocks.

The first block is intended to build a multitemporal database for predicting algal blooms. Five basic parameters are needed at this stage: a *region of interest* where the forecasting is executed; a *period of analysis* for information extraction and modelling; the number of *past step instants* used to composite the information (i.e., the attribute vector) to build the forecast model; the number of instants that comprise the *forecasting period* and the *spatial context size*, used in the final estimates about the local concentration of anomalies.

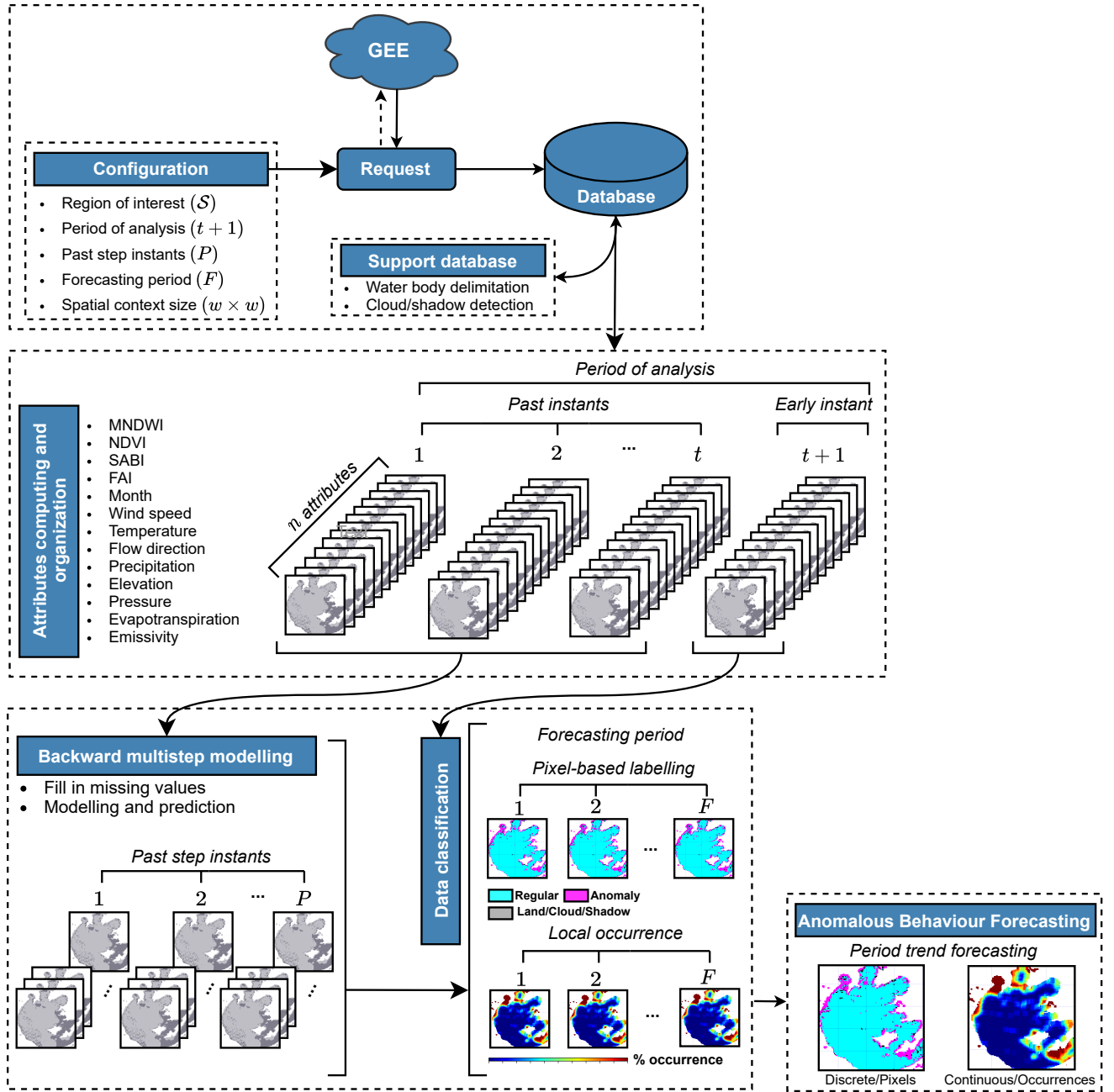


Figure 2: Framework of the proposed method.

After defined the basic parameters, are acquired data from different sensors and products in relation to the *region of interest* and period comprised by the *period of analysis*. Regarding this query, $t + 1$ instants are obtained, where the first t are employed to build the prediction model and the last instant (the *early instant*) is taken as a data reference to perform the predictions. The acquisition of these data is automatically executed using functions available in the Application Programming Interface (API) of the Google Earth Engine (GEE) platform. Specifically, this database is composed of multitemporal image series obtained by the Moderate-Resolution Imaging Spectroradiometer sensor (Modis – MOD09GA.006 – 500 m of spatial resolution) and products provided by the National Aeronautics and Space Administration Global Land Data Assimilation projects System 2.1 (NASA GLDAS 2.1) [48] (wind speed, temperature, precipitation, pressure and evapotranspiration), World Wide Fund for Nature Hydrological Data and Maps Based on Shuttle Elevation Derivatives at Multiple Scales (WWF HydroSHEDS) [32] (drainage direction), Japan Aerospace eXploration Agency Advanced Land Observing Satellite Digital Surface Model World 3D (JAXA

AW3D30) [56] (elevation) and National Aeronautics and Space Administration Modis Terra Land Surface Temperature and Emissivity Daily Global (NASA MOD11A1.006) [64] (emissivity). All data are re-sampled in terms of spatial resolution in order to match with Modis data (i.e. 500 m).

In parallel, using the images from the Modis sensor, it is built a *support database* with the purpose of identifying the boundaries of the water body (see Section 3.2 – Support data) contained in the *region of interest* as well as to identify and register the occurrence of clouds and shadows over the Modis images, which characterize information gaps on the multitemporal series and demands for a treatment conducted later (in the “fill in missing values” step). Given the delimitation of the water body, performed by the auxiliary product, and motivated by the reduction of the computational cost of subsequent procedures, the database is immediately restricted to its interior. In the following discussions we must assume that S is limited to the water body area.

After defining the main and support databases, distinct features are extracted from the remote sensing images and then integrated with climate and environmental information to produce an attribute vector that may express the phenomenon under analysis. It is important to mention that such procedures are locally executed without using cloud-computing facilities offered by the GEE platform. Formally, assuming that the *period of analysis* is composed of $t + 1$ instants, with one representing the actual (current) and other t representing past instants, a time series $\mathcal{I}^{(\ell)}$ is constructed for $\ell = 1, \dots, t + 1$, such that $\mathcal{I}^{(\ell)}(s) = \mathbf{x}_s^{(\ell)} = [x_{s:1}^{(\ell)}, \dots, x_{s:n}^{(\ell)}]$ is an attribute vector composed with information from the ℓ -th instant. The components of this vector refer to: MNDWI, NDVI, SABI and FAI spectral indices; month number referring to the instant; environmental components (i.e. meteorological and climatic) such as wind speed (m s^{-1}), temperature (K), drainage direction, precipitation ($\text{kg/m}^2/\text{s}$), elevation (m), pressure (Pa), evapotranspiration ($\text{kg/m}^2/\text{s}$) and emissivity.

Based on the previous organization, the data are restructured according to a “backward-multistep” approach, where the $P - 1$ previous instants of each observation are concatenated in the form of a single attribute vector. Precisely, for every $s \in S$ taken from $\mathcal{I}^{(\ell)}$, for $\ell = P + 1, \dots, t - F + 1$, is defined the attribute vector $\mathbf{x}_s^{(\ell|P)} = [x_{s:1}^{(\ell-P)}, \dots, x_{s:n}^{(\ell-P)}, \dots, x_{s:1}^{(\ell)}, \dots, x_{s:n}^{(\ell)}] \in \tilde{\mathcal{X}}$. Under situations where some component $x_{s:i}^{(k)}$ is under the influence of cloud/shadow occurrence, the respective values are substituted by an interpolation over the time according to the implementation of [38].

In addition, given a generic position s and instant ℓ , are assigned labels related to the occurrence of anomalies in the next F instants, herein denoted by $\mathbf{y}_s^{(\ell|F)} = [y_s^{(\ell+1)}, \dots, y_s^{(\ell+F)}]$, where $y_s^{(k)} = +1$ or -1 indicate the occurrence or not of an anomaly (i.e. algae bloom) at the k -th instant. The establishment of this label is done by an ensemble scheme of majority voting involving the thresholding rules performed on the NDVI, MNDWI, SABI and FAI spectral indices (discussed in Section 2.3). Under these conditions, $y_s^{(k)} = +1$ when all the four spectral indices point to the occurrence of anomaly with respect to position s and instant k , otherwise, $y_s^{(k)} = 0$.

Before discuss the next steps of the proposed method, it is valid to highlight the reason behind use only the instants from $P + 1$ to $t - F + 1$ to define the vectors $\mathbf{x}_s^{(\ell|P)}$ and $\mathbf{y}_s^{(\ell|F)}$. Assuming $\ell = P - 1$, the attempt of retrieve the $P - 1$ previous instants to determine $\mathbf{x}_s^{(\ell|P)}$ leads to lower-bound breaking of the time series. Analogously, $\ell = t - F + 2$ causes an upper-bound break when trying to define $\mathbf{y}_s^{(\ell|F)}$. Figure 3 elucidates the presented explanation.

Note that the vectors $\mathbf{x}_s^{(\ell|P)}$ and $\mathbf{y}_s^{(\ell|F)}$, for $\ell = P + 1, \dots, t - F + 1$ and $s \in S$, favor the organization of different training sets $\mathcal{D}^{(q)} = \{(\mathbf{x}_s^{(\ell|P)}, y_s^{(\ell+q)}) : s \in S; \ell = P + 1, \dots, t - F + 1; \}$, with $q = 1, \dots, F$. In turn, each $\mathcal{D}^{(q)}$ is used in the modelling of a classifier $G_q : \tilde{\mathcal{X}} \rightarrow \{+1, -1\}$, intended to predict the occurrence of anomalies at the q -th instant after the most recent instant (i.e. $t + 1$), characterized by the attribute vectors associated with each position/pixel of $\mathcal{I}^{(t+1|P)}$, resulting then into a pixel-based classification $\mathcal{C}^{(q)}$, with $q = 1, \dots, F$.

Derived from the classifications $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(F)}$, are estimated the local anomaly concentration by the application of a convolution filter with $w \times w$ pixels wide that computes the local percentage of anomalies. Positions outside the support S or affected by the occurrence of cloud/shadow are not considered in this calculation. The resulting output from this process are denoted by $\mathcal{O}^{(1)}, \dots, \mathcal{O}^{(F)}$.

Finally, based on the F predictions from $\mathcal{I}^{(t+1|P)}$ is computed a *period trend forecasting*. For this purpose, it is considered the median of the estimates made at each position. Under these conditions, the representations $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(F)}$ gives place to a discrete occurrence trend representation. Similarly, it is obtained a representation of occurrence concentration for the forecasting period through $\mathcal{O}^{(1)}, \dots, \mathcal{O}^{(F)}$. The reason that justifies using a general trend for the

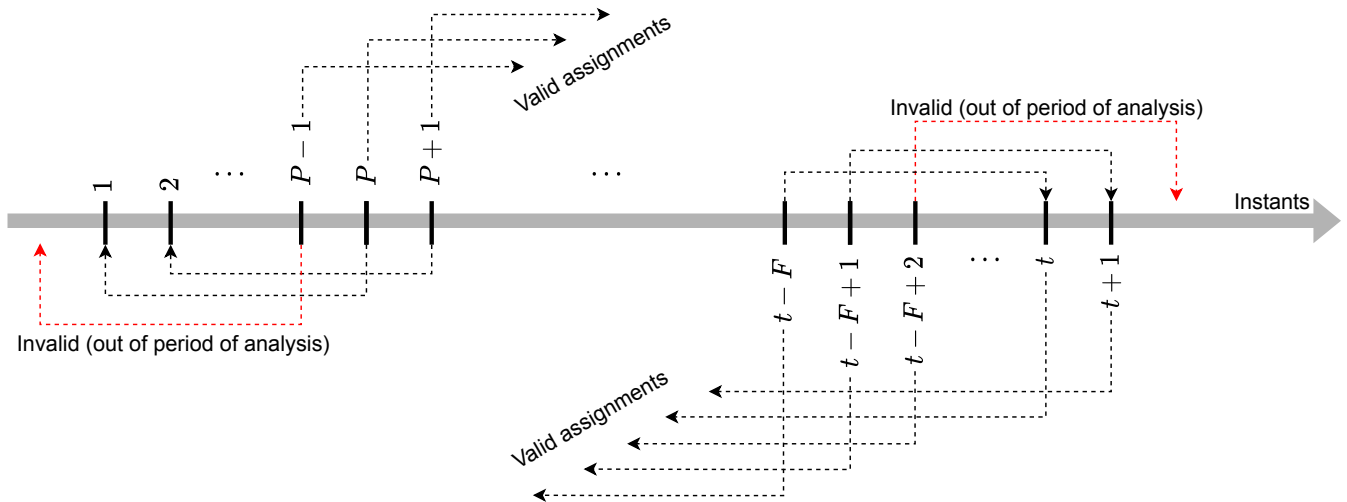


Figure 3: Representation for instants and its past and future relations.

forecasting period is intended to compose a future scenario and, simultaneously, exempt the method from the need to predict the behavior of the phenomenon in a specific instant.

3.2. Implementation details and parametrization

In order to provide an implementation overview about the proposed method, the following discussions are limited to the most relevant details. The complete source code is available at https://drive.google.com/file/d/1e9zvWku3J81F_U95VffkXXt9tVGdtYlJ/view.

Programming language and libraries: Python 3.8 [61] was adopted as programming language. The modules Numpy [59], Pandas [39], Scikit-Learn [44], responsible for the implementation of SVM and RF methods, and Keras/Tensorflow, for the implementation of the LSTM method [29], were also adopted in the coding.

Google Earth Engine API: The Google Earth Engine Application Programming Interface (API) [22] was used to access data from Modis sensor (MOD09GA.006 and MOD11A1.006), and the products GLDAS 2.1, HydroSHEDS and AW3D30. It is important to point out that the product MOD09GA.006, provided by the afore-mentioned API, is atmospheric-corrected [58].

Model parameter tuning: Given the need of tune the parameters of the SVM, RF and LSTM methods (Section 2.2), it is convenient to adopt a technique able to help and automatize this process. At this stage, in addition to a 5-time cross-validation, a Randomized Grid Search procedure [47, 4, 6] was used to obtain the most suitable parameter configurations. The space-search defined to the LSTM has $n_{epochs} \in \{5, 10, 20, \dots, 100\}$, $p_{dropout} \in \{0, 0.1, \dots, 0.5\}$, $n_{layer} \in \{16, 32, \dots, 512\}$. The parameters related to the RF method are space-delimited by $n_{est} \in \{1, 5, \dots, 250\}$, $p_{depth} \in \{1, 2, \dots, 30\}$, $p_{split} \in \{2, 4, \dots, 20\}$, $p_{leaf} \in \{2, 4, \dots, 20\}$ and $n_{att} \in \{\sqrt{\dim(\tilde{\mathcal{X}})}, 100\%, 75\%, 50\%\}$. Regarding the SVM parameters, in addition to the RBF kernel and $\xi = 0.001$, the tested values are $\gamma \in \{10^{-1}, 10^{-2}, \dots, 10^{-7}\}$ and $C \in \{1, 2, 3, \dots, 1000\}$.

Support data: Identifying the water body and the occurrences of cloud/shadow are required to compose a support database. Concerning the water body extraction, it is used the “water_mask” sub-product provided by the Modis – MOD44W.006/“Terra Land Water Mask”, which provides a mapping of the terrestrial water surface with a spatial resolution of 250 m. Alternatively, bitwise operations on the sub-product “state_1km” are performed to detect clouds and shadows. More information can be found at [24].

Cloud threshold: In view of the low or no information coming from clouds and shadows, it was necessary to establish a threshold (i.e. good visibility). When considering the cloud occurrence tolerance percentages of 25%, 50%,

and 75% in the NDVI data from Section 4, preliminary analysis showed all values are statistically equivalent with 5% significance. The mean and variance for 25% were equal to -0.44 and 0.04. For 50% and 75%, it was -0.47/0.04 and -0.49/0.04, respectively. The NDVI was adopted as a comparative reference for being the most well-known remote sensing index.

Period of analysis: The period of analysis used in forecasting model may influence the prediction performance. After preliminary tests considering 30, 60, 90, 180, 365, 730, and 1825 days, it was observed that better results are achieved with 180 days.

Past steps instants and forecasting period: The parameters *past steps instants*, responsible for defining the number of previous instants used in the composition of the attribute vectors; and the *forecasting period*, related to the period of prediction, plays a direct influence on the model performance and its computational cost. Face with the need of select the more suitable values for such parameters, the proposed method was previously executed on distinct study areas and periods (Sections 4.1 and 4.2) varying both values between 1 to 15. Figure 4 depicts a median profile, in terms of R^2 Score regarding the concentration occurrences estimation. The results allows to observe that four *past steps instants* and a *forecasting period* of five instants comprises an appropriated choice.

Spatial context size: Represents the dimension $w \times w$ of the convolution filter applied on the classifications $C^{(1)}, \dots, C^{(F)}$ to result in the occurrence concentration. After performing a battery of tests with w varying in 3, 5, and 7 pixels, the use of $w = 7$ provided a better consistency.

Data projection: Aiming to alleviate the computational cost, the Principal Component Analysis (PCA) [28] technique was applied on the entire set of vectors $\mathbf{x}_s^{(\ell|P)}$, referring to all existing instants and positions, allowing then a projection of the data on an attribute space of reduced dimension but without significant loss in the algorithm performance. For this purpose, projections with an explained variance of 99%, 95% and 90% were compared in terms of algorithm performance and the latter value was selected.

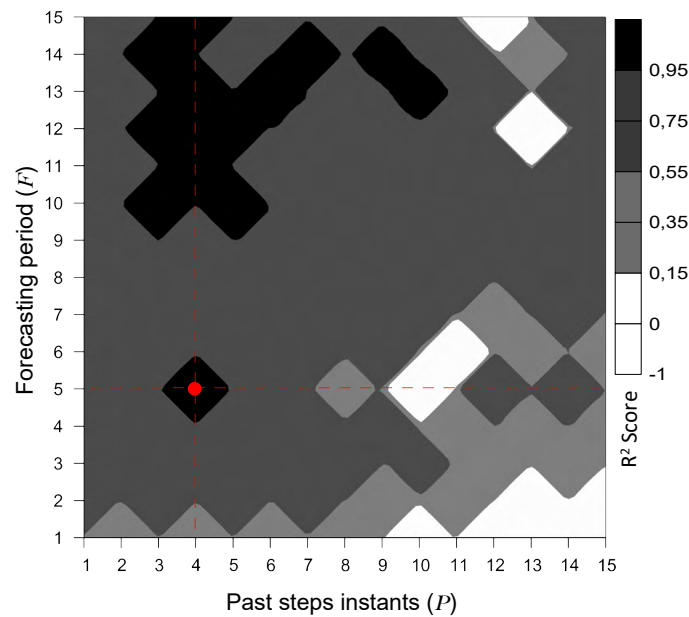


Figure 4: Median performance in terms of R^2 Score for different configurations of *past steps instants* and *forecasting period* parameters. The red dot and dotted lines indicate the best performance values.

4. Experiments

Three case studies were carried in order to validate the proposed method. The regions of Lake Erie (USA), Chilika (India), and Taihu (China) are the respective study areas (Figure 5). Regarding the first study area, two instants (i.e., the early instant) and the five posterior days (i.e. the forecasting period) were selected. Similarly, one early instant and

its five posterior days were considered for the second study area. Regarding the third area, two early instants were selected from a reference dataset provided by the National Earth System Science Data Center (NESSDC) [34, 35]. It is worth mentioning that the selected periods were conditioned to low cloud/shadow occurrence at both early instant and forecasting period.

Using the LSTM, RF and SVM methods as classification models, three algorithm variants were included in the experiments, herein referred as ABF-LSTM, ABF-RF and ABF-SVM, respectively. The performance of each variant is measured in terms of global accuracy, kappa coefficient [13], percentages of true/false-positive/negative (denoted by TP, TN, FP and FN), F1-Score [60], R^2 -Score [17] and Root Mean Square Error (RMSE) [10]. Hypothesis tests derived from the kappa coefficient [13] were performed in order to compare the results with a significance of 5%.

The experiments were run on a computer with an AMD Ryzen 9 3900X 12-core processor, 32 GB of RAM and operating system Ubuntu Linux version 20.04.

4.1. Study areas

Located between the states of Ohio, New York, Pennsylvania and Michigan, Lake Erie (Figure 5 (a)) is one of the Great Lakes of Saint Lawrence. According to [7], this lake has approximately 388 km of length, 92 km wide and an average depth of 19 m. It also presents the total area and volume of 25,657 km² and 484 km³, respectively. Affected by the urbanization and agriculture growing, this lake has been demonstrated algal blooms since the late 1960's [75]. Usually, peaks of algal bloom are verified between August and September, during the summer in the northern hemisphere [55].

Lake Chilika (Figure 5 (b)) is located on the east coast of Orissa, India. Beyond considered an income source for about 150,000 fishers [51], such lake comprise the largest brackish water body in Asia and deeply contributes to the regional economy [43]. Furthermore, according to [43], this lake it presents an area that varies between 906 km² (summer) and 1165 km² (monsoon) and depth between 0.40 and 4.90 m. In addition to problems such as constant changes in salt concentration, freshwater weeds, the decline in fish productivity, the lake faces a constant process of eutrophication [46].

The third study area corresponds to Lake Taihu (Figure 5 (c)), China. Although considered a water source for more than 40 million people, this lake is a scenario of frequent algal blooms. According to [21], the Lake Taihu has an area of 2338 km² and an average depth of 1.9 m. The wet and dry periods usually occur between April to July and February to March, respectively.

4.2. Reference data

As previously mentioned, the proposed method is assessed at different study areas and instants. Regarding the first area (Lake Erie – Figure 5 (a)), October 10th, 2013, and July 26th, 2015 are taken as *early instants*. Consequently, the respective forecasting periods becomes October 11th–15th, 2013 (Figures 6 (a)-(e)) and July 27th–31st, 2015 (Figures 7 (a)-(e)). For the second area (Lake Chilika – Figure 5 (b)), November 24th, 2019 is taken as *early instants*, then the forecasting period is November 25th–29th, 2014 (Figures 8 (a)-(e)). At last, the third area (Lago Taihu – Figure 5 (c)) adopts August 24th, 2016 and July 24th, 2017 as *early instants*, thus providing the *forecasting period* according to the Figures 9 (a)-(e) and Figures 10 (a)-(e). It is worth stress that in the following experiments, in order to simulate a real-world application, the *forecasting period* is unknown for the ABF method. Thus, based on the current behavior observed at the *early instants*, a prediction is made about the occurrence of anomalies (algal bloom) in the *forecasting period*.

In order to define a reference dataset (i.e. ground-truth data) and adopt it to measure the prediction performance of the ABF method in the Lake Erie and Chilika areas, we consider a process similar to those adopted to determine the labels of training patterns (i.e. $\mathbf{y}_s^{(\ell|F)}$), as discussed in Section 3. In summary, the steps of such a process are: (i) each instant into the *forecasting period* is submitted to a thresholding of the spectral indices MNDWI, NDVI, FAI and SABI, calculated from Modis images (Section 2.3); (ii) locations/pixels simultaneously identified as anomalies by the four thresholding processes are admitted as “anomaly”; (iii) locations/pixels simultaneously identified as non-anomalies by the four thresholding processes are admitted as “regular”; (iv) locations/pixels with a partial agreement regarding the occurrence of anomalies or under the influence of cloud/shadow are admitted as “indeterminate” or “land/cloud/shadow” and then discarded as a reference; (v) from the anomalies and non-anomalies identified over the instants into the *forecasting period*, the median value is taken over each location to determine the local trend and generate a discrete reference; finally (vi), to generate the reference which expresses the local anomaly concentration,

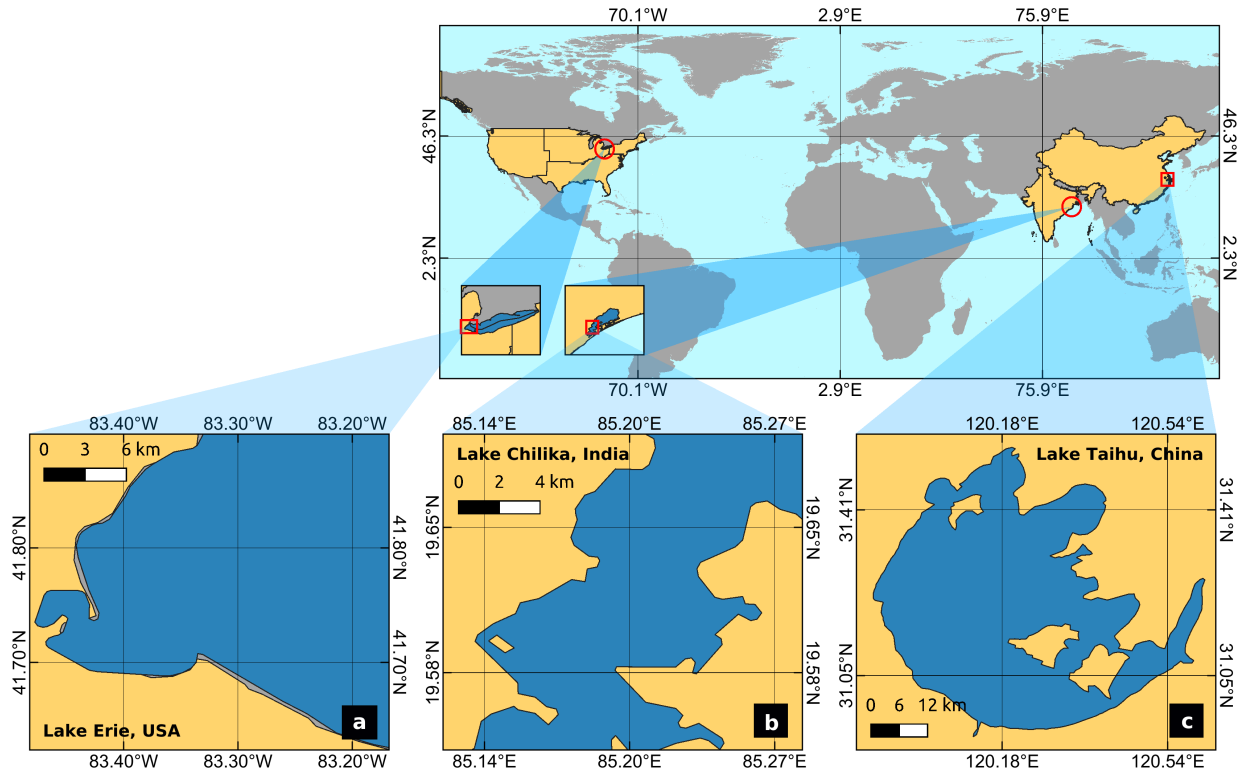


Figure 5: Spatial location of study areas.

firstly, a convolution filter of 7×7 pixels wide is applied to calculate the local percentage of anomaly occurrence at each instant of the *forecasting period*, followed then by the calculation of the local trend through the median.

Concerning the reference dataset to assess the ABF at Lake Taihu study area, as already mentioned, we used information provided by the NESSDC. In this case, the anomaly occurrence is defined by locations/pixels where the Chl-a concentration is above $20 \mu\text{g}/\text{L}$ [71, 65], otherwise, the location is assumed as regular/non-anomaly. After that, the discrete and local anomaly concentration references are determined following the steps (v) and (vi), respectively.

The resulting references datasets are shown by the sub-figures (k) and (q) at Figures 6 to 10. For the sake of presentation, the estimates at each instant are depicted in sub-figures (f)–(j) and (l)–(p) at Figures 6 to 10. Table 2 summarizes the number of anomaly and regular positions/pixels for each discrete reference. The average and standard deviation values express the local anomaly concentration references.

4.3. Results and discussions

According to the established accuracy measures and reference datasets, Figure 11 depicts the forecasting performances achieved by the ABF-SVM, ABF-RF and ABF-LSTM variants in each study area and considered instants.

Regarding the first study area (Lake Erie), although the ABF-RF demonstrated a global accuracy and F1-Score with a slight difference compared to the ABF-LSTM and ABF-SVM, such variant achieved the best kappa coefficient among its competitors. Regarding the True-Positive/False-Negative percentages, all the variants present a similar performance. However, it was observed that in the first period (October 11th–15th, 2019) that the ABF-LSTM model does not detect TP values. In the context of this analysis, “positive” means the occurrence of anomalies (alga bloom). The opposite is verified for the term “negative”, where anomalies are not verified (i.e. regular). Additionally, according to the R^2 -Score and RMSE values, computed through the occurrence concentration reference, it is possible to observe that the ABF-RF tends to deliver a high R^2 -Score, low RMSE and a good balance between the expected (reference) and predicted (forecasted). Also, it is worth observing that all the variants predict the anomaly occurrence with a low difference to the reference data (i.e. the dashed red line at zero).

Concerning the second study area (Lake Chilika), the ABF-SVM provided better results than the ABF-RF variant, as expressed by the global accuracy, F1-Score and kappa coefficient. Moreover, the ABF-SVM variant demonstrated lower FP values when compared to its competitors. Analyzing the forecasting results in terms of concentration

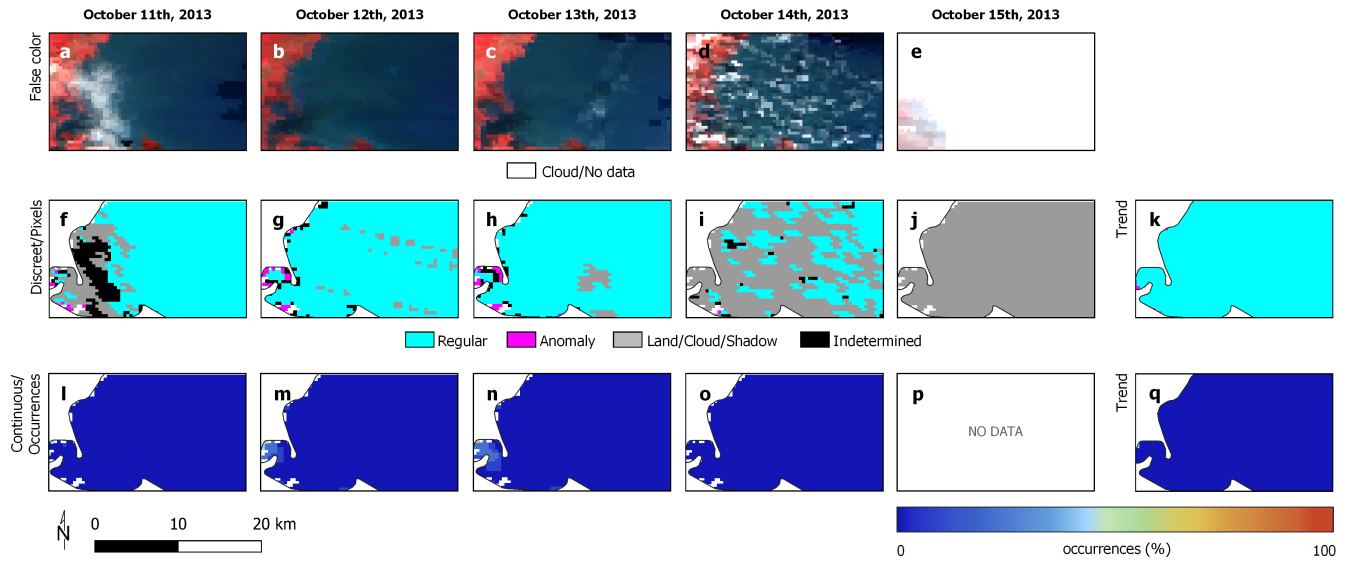


Figure 6: Images (a)-(e) in false-color composition (NIR, red and green bands) for Lake Erie study area from October 11th–15th, 2013. Regular, anomaly, land/cloud/shadow and indetermined samples (f)-(j) are identified by cyan, magenta, gray and black polygons, respectively; (l)-(p) show the local anomaly concentration at each instant; (k) and (q) are the discrete and local anomaly concentration references for the whole forecasting period.

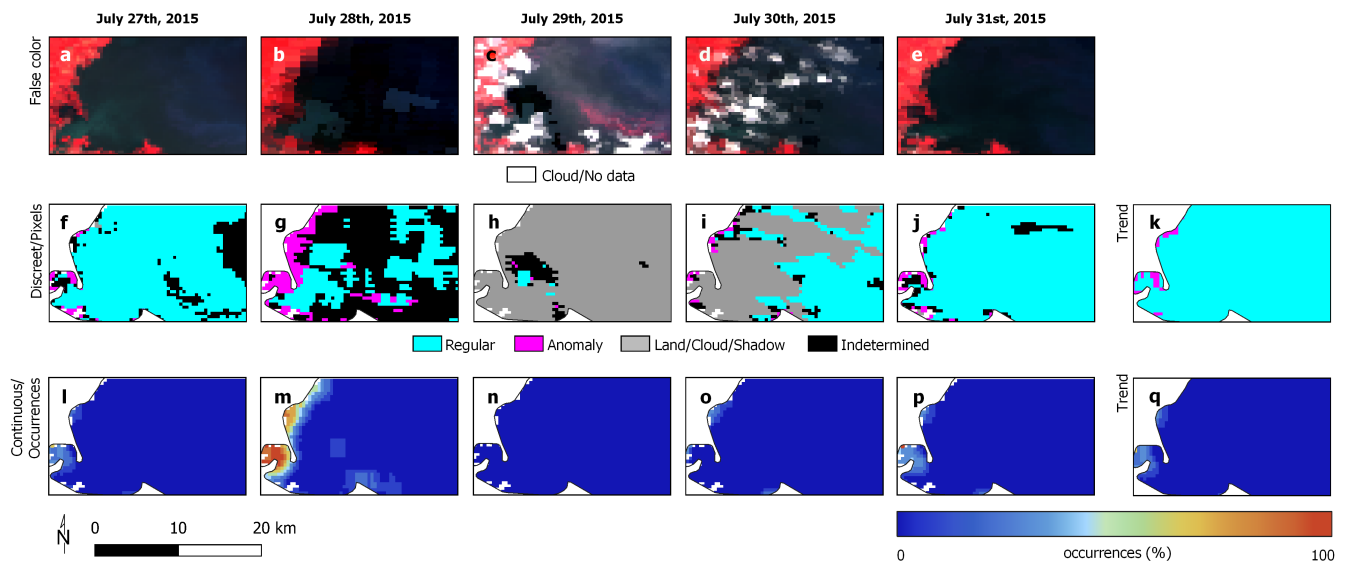


Figure 7: Images (a)-(e) in false-color composition (NIR, red and green bands) for Lake Erie study area from July 27th–31st, 2015. Regular, anomaly, land/cloud/shadow and indetermined samples (f)-(j) are identified by cyan, magenta, gray and black polygons, respectively; (l)-(p) show the local anomaly concentration at each instant; (k) and (q) are the discrete and local anomaly concentration references for the whole forecasting period.

occurrences, although the error assigned to the ABF-SVM variant in this area is bigger than the previous study area, such variant still provides more suitable R^2 -Score and RMSE values than ABF-RF and ABF-LSTM. On the other hand, we may observe that the ABF-LSTM generates a prediction with a higher negative divergence to the expected values (i.e., omission errors) when compared to ABF-RF and ABF-SVM, whose divergences are positive (i.e., inclusion errors).

For the Lake Taihu study area, with respect to the first period (August 25th–29th, 2016), although the ABF-LSTM variant showed high global accuracy and F1-Score, the FN percentage is also high and leads to a lower kappa value in

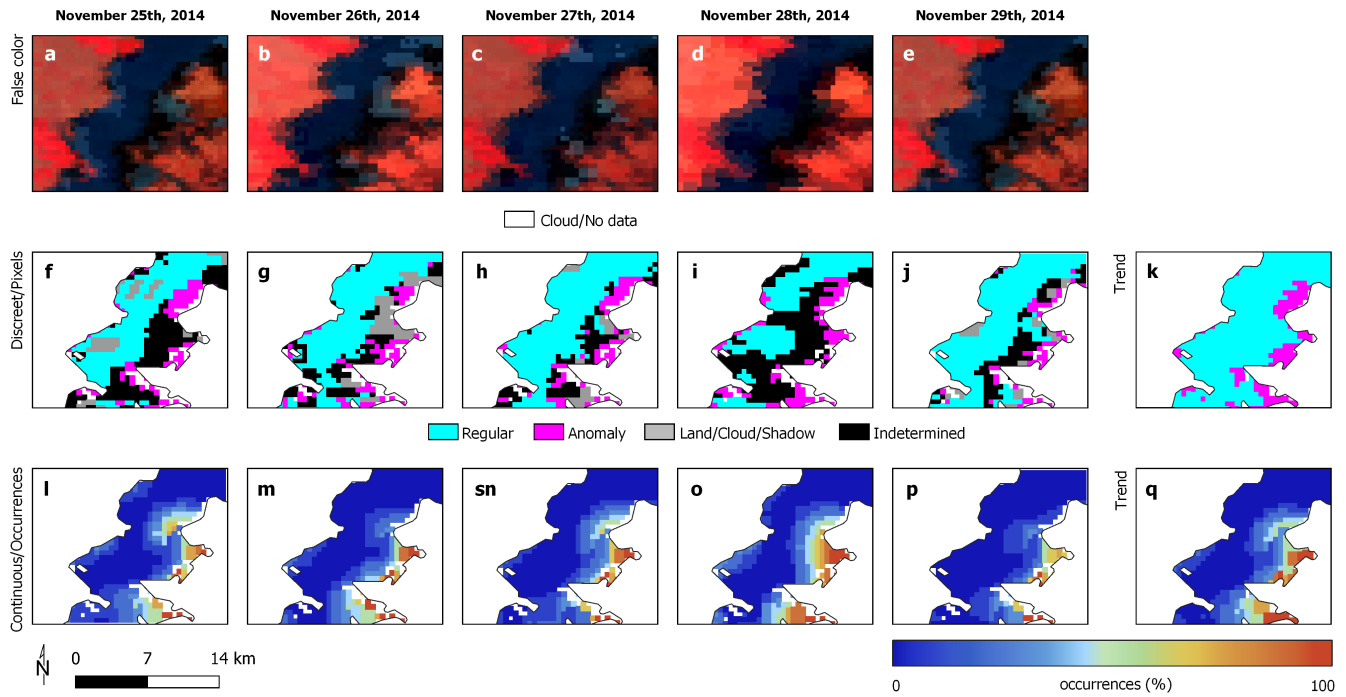


Figure 8: Images (a)–(e) in false-color composition (NIR, red and green bands) for Lake Chilika study area from November 25th–29th, 2014. Regular, anomaly, land/cloud/shadow and indetermined samples (f)–(j) are identified by cyan, magenta, gray and black polygons, respectively; (l)–(p) show the local anomaly concentration at each instant; (k) and (q) are the discrete and local anomaly concentration references for the whole forecasting period.

comparison to the other variants. Although the ABF-RF demonstrates a high FP percentage, the best result is delivered by this variant. A similar result is achieved by the ABF-SVM variant, except for the FP count. However, all models performed poorly in FN, which means that some anomalies occurrences were not predicted. Regarding the performance analysis about occurrence concentration, the negative R^2 -Score and the high RMSE values corroborate the omission behavior demonstrated by all the variants. Nonetheless, it is worth point out that the ABF-RF provided the best result.

Focusing on the second period (July 25th–29th, 2017) at Lake Taihu, both high global accuracy values and F1-Scores are assigned to the three variants. However, when considering the kappa coefficient, the ABF-SVM model was superior to its competitors. Also, we should observe a higher inclusion error frequency (i.e., FP values) generated by the SVM-based variant. In the occurrence concentration analysis, although the ABF-SVM showed a small yet positive R^2 -Score, the related RMSE value is higher than the other variants.

When evaluating the Figure 12 (b) with respect to Lake Erie at October 11th–15th, 2013, it was observed that anomaly pixels were not predicted by the ABF-LSTM, which contradicts the discrete reference (Figure 12 (a)). The ABF-RF and ABF-SVM variants predicted a closer behavior to the expected/reference. With respect to the second period (July 27th–31st), 2015, the Figures 13 (b)–(d) showed similar performance between the three proposals and the reference data (Figure 13 (a)). The occurrence mapping depicted in Figures 13 (e)–(h) also reveal the FP errors (Figure 13 (e)). However, it is valid to affirm that the ABF tends to be correct in an approximate/regional point of view.

Figure 14 presents the forecasting map for Lake Chilika (period November 25th–29th, 2014). While the ABF-LSTM fails in predicting the anomalies (Figure 14 (b)), the RF (Figure 14 (c)) and SVM (Figure 14 (d)) variants exhibited more assertive results in both discrete and occurrence concentration representations (Figures 14 (a) and (e)).

Regarding Lake Taihu in the first period (August 25th–29th, 2016), several divergences were observed between the ABF variants (Figures 15 (b)–(d)) and the reference data (Figure 15 (a)). While the first image showed a marked anomalous behavior in the upper left region that extends to the center, ABF-LSTM and ABF-SVM did not demonstrate this prediction. The same is true for the ABF-RF variant, although it has mapped a larger portion of anomalies in the upper-central region. The same behavior was observed for the concentration occurrence mappings (Figure 15 (f)–(h)).

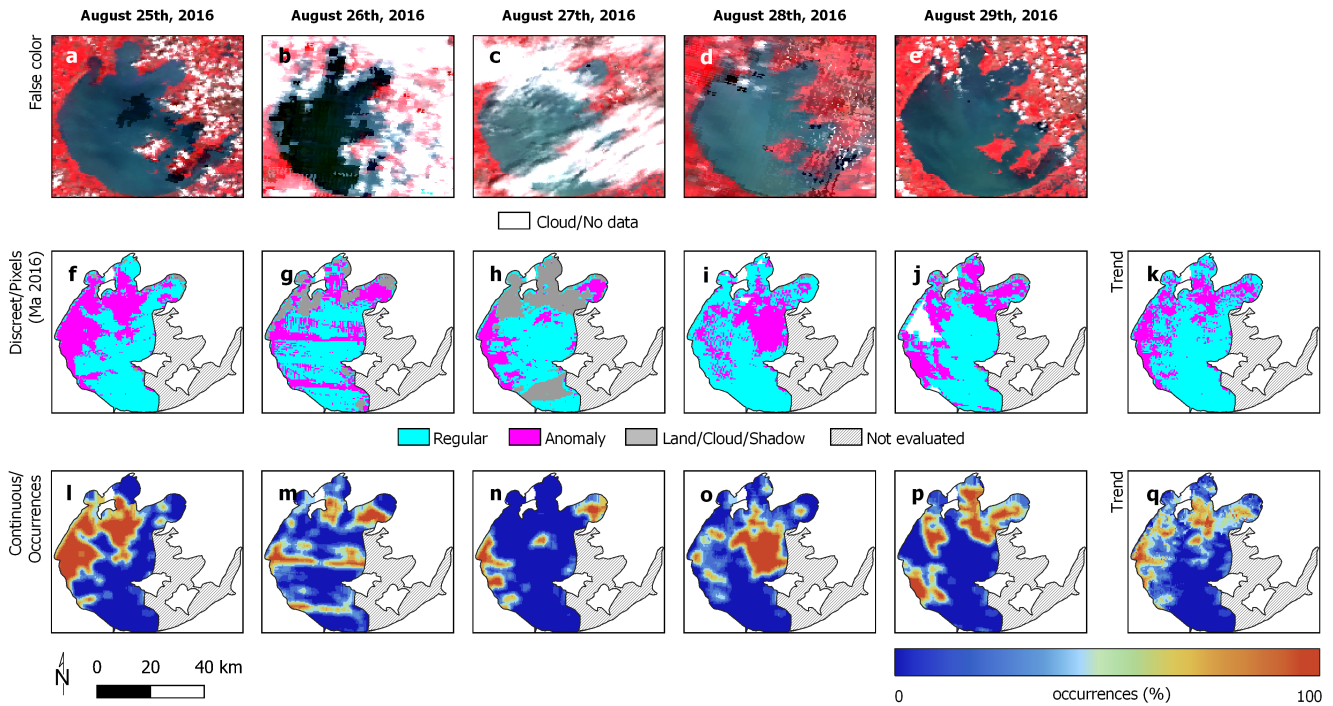


Figure 9: Images (a)-(e) in false-color composition (NIR, red and green bands) for Lake Taihu study area from August 25th–29th, 2016. Regular, anomaly, land/cloud/shadow and indetermined samples (f)-(j) are identified by cyan, magenta, gray and black polygons, respectively; (l)-(p) show the local anomaly concentration at each instant; (k) and (q) are the discrete and local anomaly concentration references for the whole forecasting period.

As shown in Figures 16 (a)-(h), which refers to the second period of the Chinese lake, the performance previously discussed through the Figure 11 was confirmed, where the three evaluated models obtained superior performance in comparison to the first period. A high correlation was observed between the reference data (Figure 16 (a)) and the ABF-RF variant (Figures 16 (c) and (g)). The ABF-LSTM, in turn, did not detect anomaly occurrences over the study area. The opposite is observed for the ABF-SVM variant.

In order to assess the significance of the results, hypothesis tests derived from the kappa coefficient were performed. In the first period evaluated for Lake Erie, it is revealed the equivalence between the ABF-RF and ABF-SVM variants. In the second period, all the variants are statistically equivalent. Regarding the second study area, the ABF-RF and ABF-SVM were again statistically equivalent and significant to ABF-LSTM. The results for Lake Taihu demonstrated the superiority of the ABF-RF, however, such variant is equivalent to the ABF-SVM in the second period.

Finally, Figure 17 shows the median run-time of each variant over the different study areas and periods. A higher run-time is expended by the ABF-LSTM variant due to the low-convergence behavior commonly presented by the neural network models. Reversely, the ABF-RF model had the lowest cost for all datasets. Given that the increase of pixels to be processed is proportional to the study area dimensions, the cost assigned to Lake Taihu tends to be higher in comparison to the other areas.

5. Conclusions

Forecast the proliferation of potentially harmful algae in water resources is essential for all sector in the society. Face with this relevant issue, this study developed an algorithm for Anomalous Behaviour Forecasting (ABF) through the use of Remote Sensing multitemporal data and Machine Learning concepts. Three study cases were performed in order to validate the proposed method. Also, considering the RF, SVM and LSTM classification methods, the proposed method was unfold into the ABF-RF, -SVM and -LSTM variants.

The obtained results shows that the ABF model is able to forecast the analyzed phenomenon. Regarding the specific variants, the ABF-RF showed more regular results with lower computational cost, obtaining mean values of global

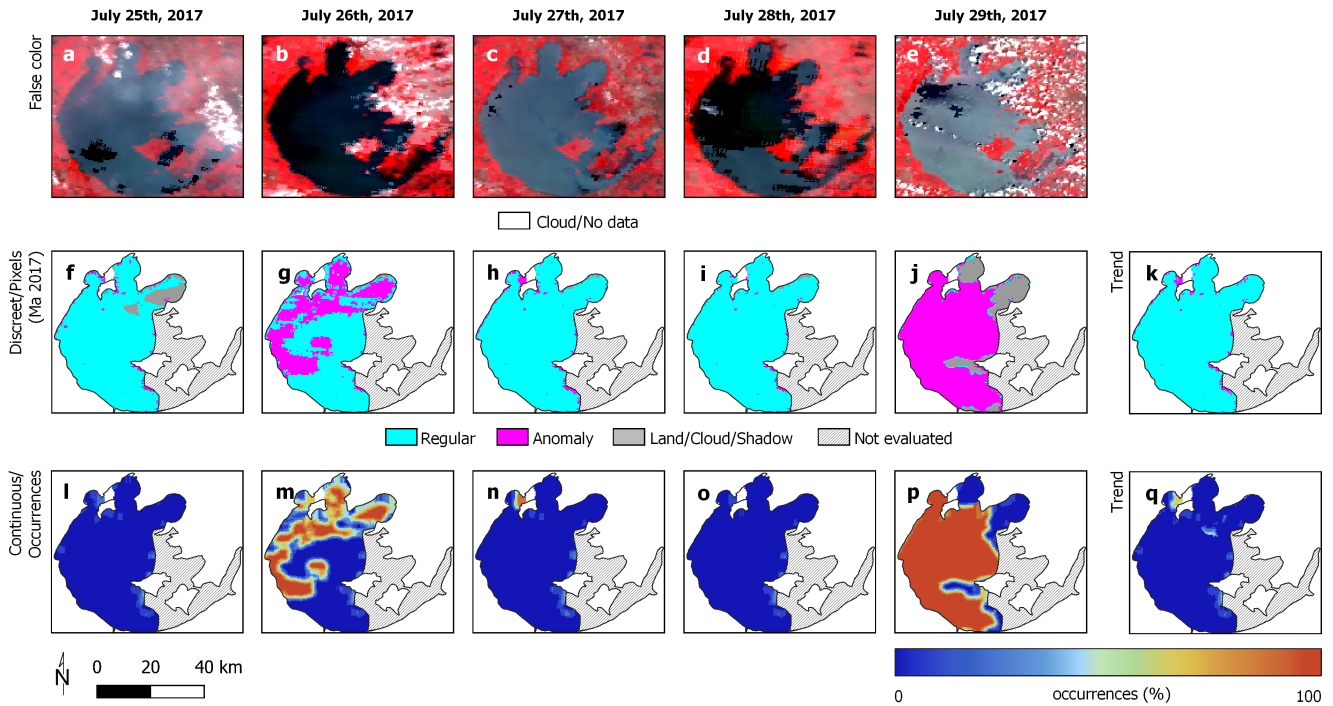


Figure 10: Images (a)–(e) in false-color composition (NIR, red and green bands) for Lake Taihu study area from July 25th–29th, 2017. Regular, anomaly, land/cloud/shadow and indetermined samples (f)–(j) are identified by cyan, magenta, gray and black polygons, respectively; (l)–(p) show the local anomaly concentration at each instant; (k) and (q) are the discrete and local anomaly concentration references for the whole forecasting period.

accuracy, F1-Score, kappa coefficient, R^2 -Score and RMSE between 95%, 94%, 51%, 0.38 and 8.22, respectively, against 90%, 92%, 51%, 0.61 and 20.98 for the ABF-SVM; and 95%, 93%, 0%, 0.0 and 8.05 for the ABF-LSTM.

Nonetheless, it is worth highlight some weaknesses in the proposed method. The first one regards the need for images with good visibility (i.e. less than 50% clouds/shadows). If the multitemporal image series is small due to strong cloud incidence, the temporal profile of the analyzed phenomenon tends to be uninformative. The second weak point stands for the effects of study area dimensions into the computational cost increasing. Lastly, the method is subject to the Internet connection availability and requires updated meteorological/climatic products.

Finally, future perspectives include (i) testing other variants by considering other classification models; (ii) test the proposed method in other study areas; (iii) evaluate the algorithm through new applications domains like forest burning, deforestation and flooding.

Acknowledgement(s)

The authors would like to thank the National Earth System Science Data Center, National Science and Technology Infrastructure of China (<http://www.geodata.cn>) for providing data support for Lake Erie and Lake Taihu, respectively.

References

- [1] F. Alawadi. Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI). In C. R. B. Jr., S. P. Mertikas, X. Neyt, and M. Velez-Reyes, editors, *Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010*, volume 7825, pages 45 – 58. International Society for Optics and Photonics, SPIE, 2010.
- [2] J. I. Allen, T. J. Smyth, J. R. Siddorn, and M. Holt. How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? *Harmful Algae*, 8(1):70–76, 2008. HABs and Eutrophication.
- [3] P. H. M. Ananias and R. G. Negri. Anomalous behaviour detection using one-class support vector machine and remote sensing images: a case study of algal bloom occurrence in inland waters. *International Journal of Digital Earth*, 14(7):921–942, 2021.
- [4] N. Baba. Convergence of a random optimization method for constrained optimization problems. *Journal of Optimization Theory and Applications*, 33(4):451–461, 1981.

Table 2

Summary of regular and anomaly pixels and training images related to ground truth data of each study area and its considered instants.

	Lake Erie	
	October 11th–15th, 2013	July 27th–31st, 2015
Regular	277328	233137
Anomaly	19041	13210
Average occurrence	0.8%	3.2%
Stand. deviation occ.	4.5%	11.3%
	Lake Chilika	
	November 25th–29th, 2014	
Regular	41272	
Anomaly	21692	
Average occurrence	19.4%	
Stand. deviation occ.	26.5%	
	Lake Taihu	
	August 25th–29th, 2016	July 25th–29th, 2017
Regular	631398	1259410
Anomaly	72399	142685
Average occurrence	22.6%	19.1%
Stand. deviation occ.	27.7%	33.8%

- [5] R. Barzegar, M. T. Aalami, and J. Adamowski. Short-term water quality variable prediction using a hybrid cnn-lstm deep learning model. *Stochastic Environmental Research and Risk Assessment*, pages 1–19, 2020.
- [6] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.
- [7] S. J. Bolsenga and C. E. Herdendorf. *Lake Erie and Lake St. Clair Handbook*. Wayne State University Press, 1993.
- [8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [9] L. Bruzzone and C. Persello. A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2142–2154, 2009.
- [10] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.
- [11] I. Chawla, L. Karthikeyan, and A. K. Mishra. A review of remote sensing applications for water security: Quantity, quality, and extremes. *Journal of Hydrology*, 585:124826, 2020.
- [12] H. Cho, U. Choi, and H. Park. Deep learning application to time-series prediction of daily chlorophyll-a concentration. *WIT Trans. Ecol. Environ*, 215:157–63, 2018.
- [13] R. G. Congalton and K. Green. *Assessing the accuracy of remotely sensed data*. CRC Press, Boca Raton, 2009.
- [14] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [15] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, 2000.
- [16] T. G. Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125, 2002.
- [17] N. R. Draper and H. Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.
- [18] L. R. Duncan. Individual choice behavior: A theoretical analysis, 1959.
- [19] J. B. Franklin, T. Sathish, N. V. Vinithkumar, and R. Kirubakaran. A novel approach to predict chlorophyll-a in coastal-marine ecosystems using multiple linear regression and principal component scores. *Marine Pollution Bulletin*, 152:110902, 2020.
- [20] B.-C. Gao. NDWIA normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266, 1996.
- [21] Y. Gao, G. Zhu, H. W. Paerl, B. Qin, J. Yu, and Y. Song. A study of bioavailable phosphorus in the inflowing rivers of lake taihu, china. *Aquatic Sciences*, 82(1):1, 2020.
- [22] GEE-API. Google Earth Engine API, 2019. Accessed: 2019-10-29.
- [23] H. J. Gons. Optical teledetection of chlorophyll a in turbid inland waters. *Environmental Science & Technology*, 33(7):1127–1132, 1999.
- [24] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.
- [25] X. Han-Qiu. A study on information extraction of water body with the modified normalized difference water index (mndwi). *Journal of remote sensing*, 5:589–595, 2005.
- [26] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.
- [27] A. Huete. A soil-adjusted vegetation index (savi). *Remote Sensing of Environment*, 25(3):295–309, 1988.
- [28] I. Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.
- [29] N. Ketkar. Introduction to keras. In *Deep learning with Python*, pages 97–111. Springer, 2017.

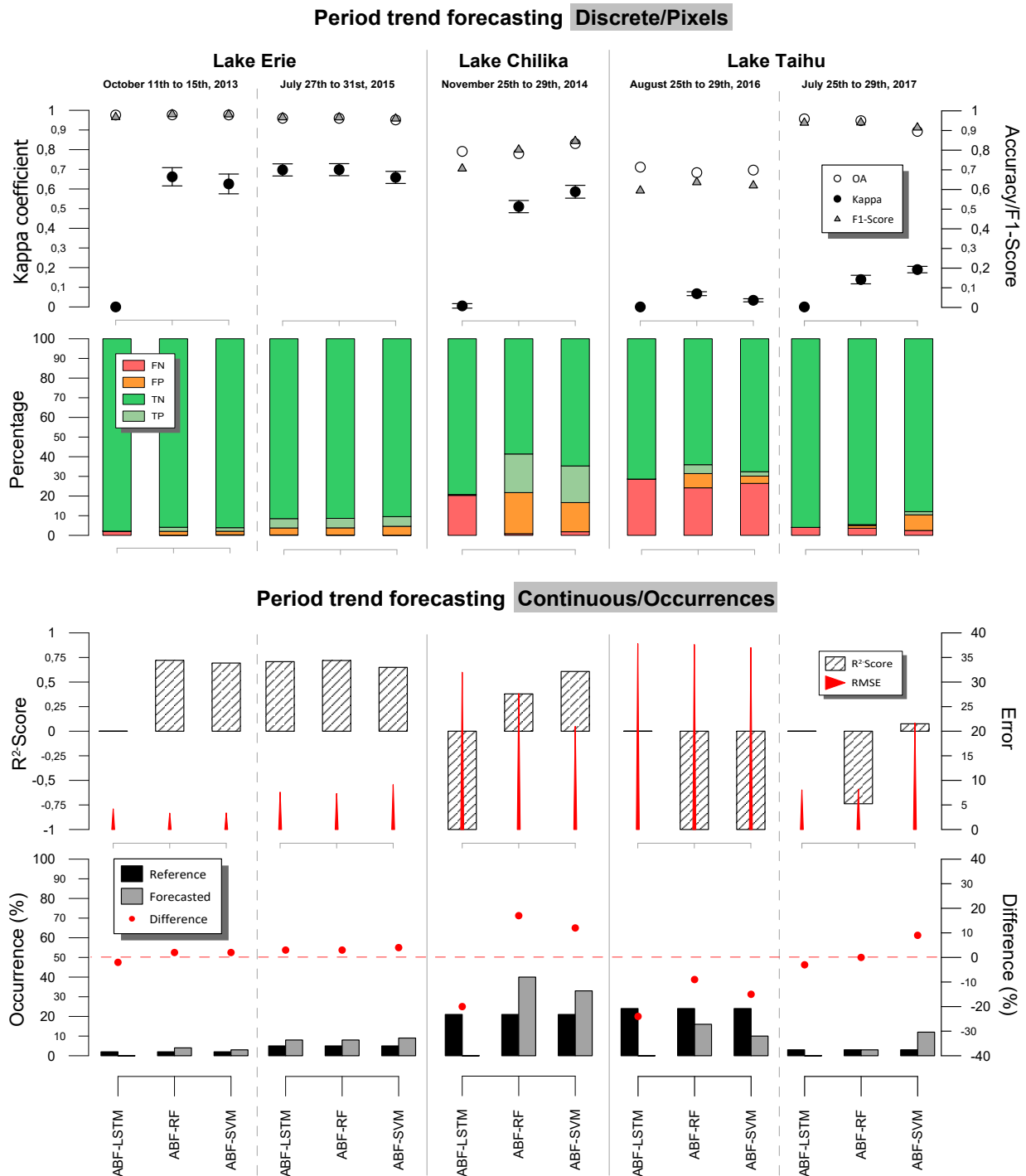


Figure 11: Performance graphs for each study area, period and variants. The first graph represents the global accuracy, F1-Score and kappa coefficient values. The second one demonstrates the performance of each variant based on True(T)-False(F)/Positive(P)-Negative(N) values. The third compares the variants through the R^2 -Score RMSE measures. The latter graph evaluates the scene based on predicted/referenced occurrence percentages. The dashed red line represents the value where the difference is null.

- [30] T. S. Körting, L. M. G. Fonseca, E. F. Castejon, and L. M. Namikawa. Improvements in sample selection methods for image classification. *Remote Sensing*, 6(8):7580–7591, 2014.
- [31] M.-S. Lee, K.-A. Park, J. Chae, J.-E. Park, J.-S. Lee, and J.-H. Lee. Red tide detection using deep learning and high-spatial resolution optical satellite imagery. *International Journal of Remote Sensing*, 41(15):5838–5860, 2020.
- [32] B. Lehner, K. Verdin, and A. Jarvis. New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union*, 89(10):93–94, 2008.

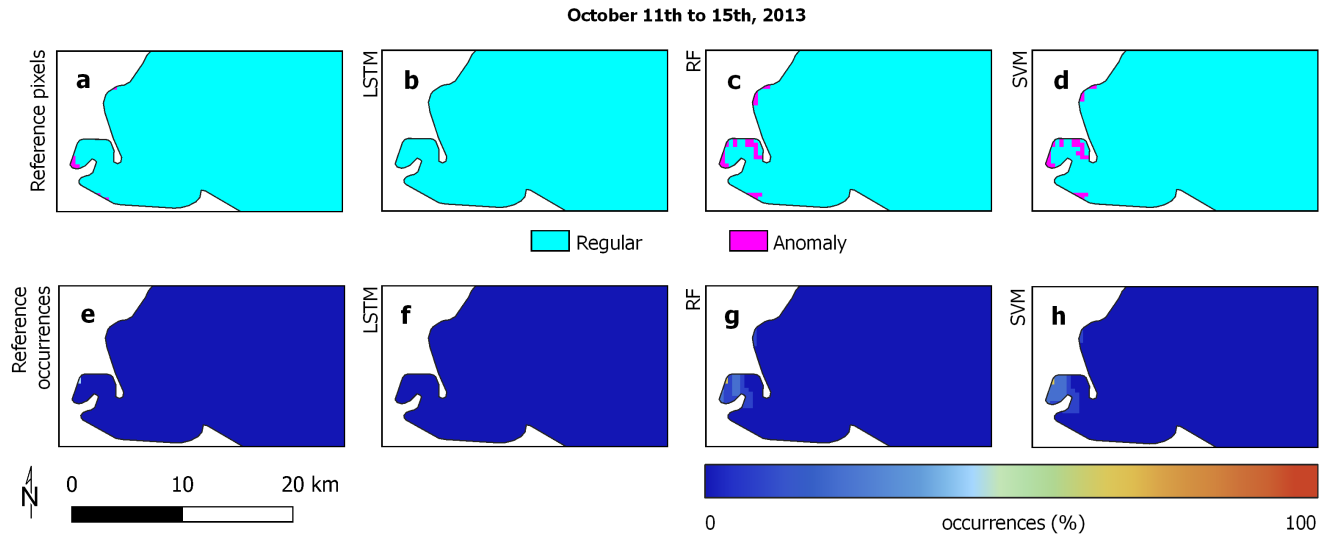


Figure 12: Period trend forecasting by pixels (a)-(d) and occurrences (e)-(h) from October 11th–15th, 2013, related to Lake Erie. In this sense, (a) and (b) represent the mapping of the reference data.

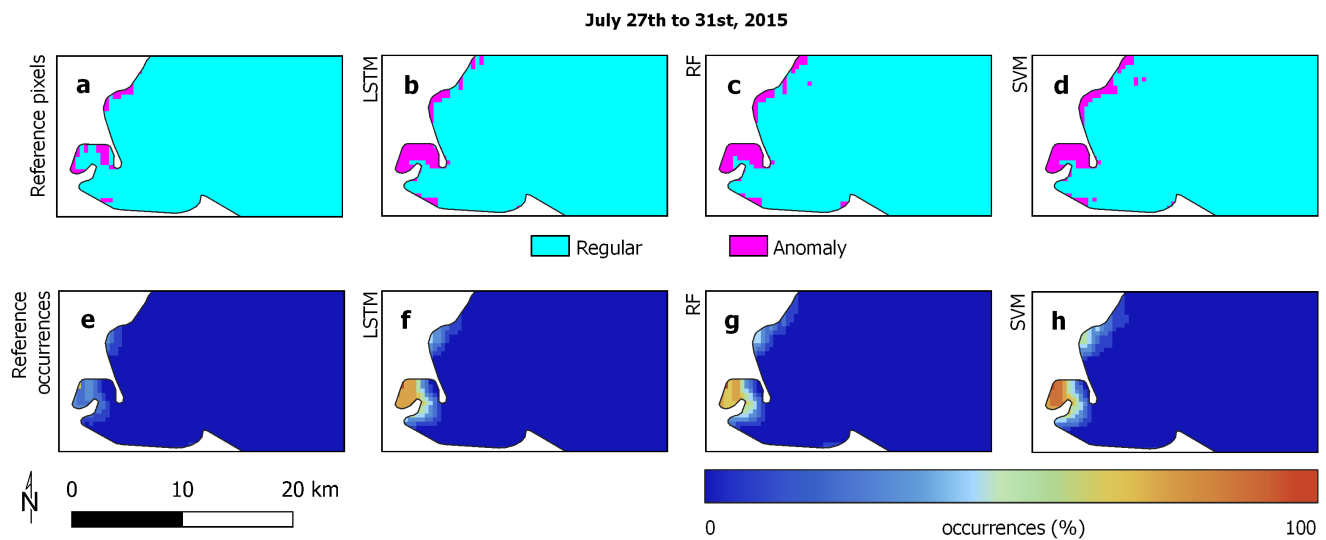


Figure 13: Period trend forecasting by pixels (a)-(d) and occurrences (e)-(h) from July 27th–31st, 2015, related to Lake Erie. In this sense, (a) and (b) represent the mapping of the reference data.

- [33] H. Q. Liu and A. Huete. A feedback based modification of the ndvi to minimize canopy background and atmospheric noise. *IEEE Transactions on Geoscience and Remote Sensing*, 33(2):457–465, 1995.
- [34] R. Ma. Lake Taihu Chlorophyll Inversion Product Data Set (2016). National Earth System Science Data Center, National Science and Technology Infrastructure of China, 2016. Accessed 2020-06-13.
- [35] R. Ma. Lake Taihu Chlorophyll Inversion Product Data Set (2017). National Earth System Science Data Center, National Science and Technology Infrastructure of China, 2017. Accessed 2020-06-13.
- [36] F. Martínez-Álvarez and D. Tien Bui. Advanced machine learning and big data analytics in remote sensing for natural hazards management. *Remote Sensing*, 12(2), 2020.
- [37] S. K. McFeeeters. The use of the normalized difference water index (ndwi) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432, 1996.
- [38] W. McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.
- [39] W. McKinney et al. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

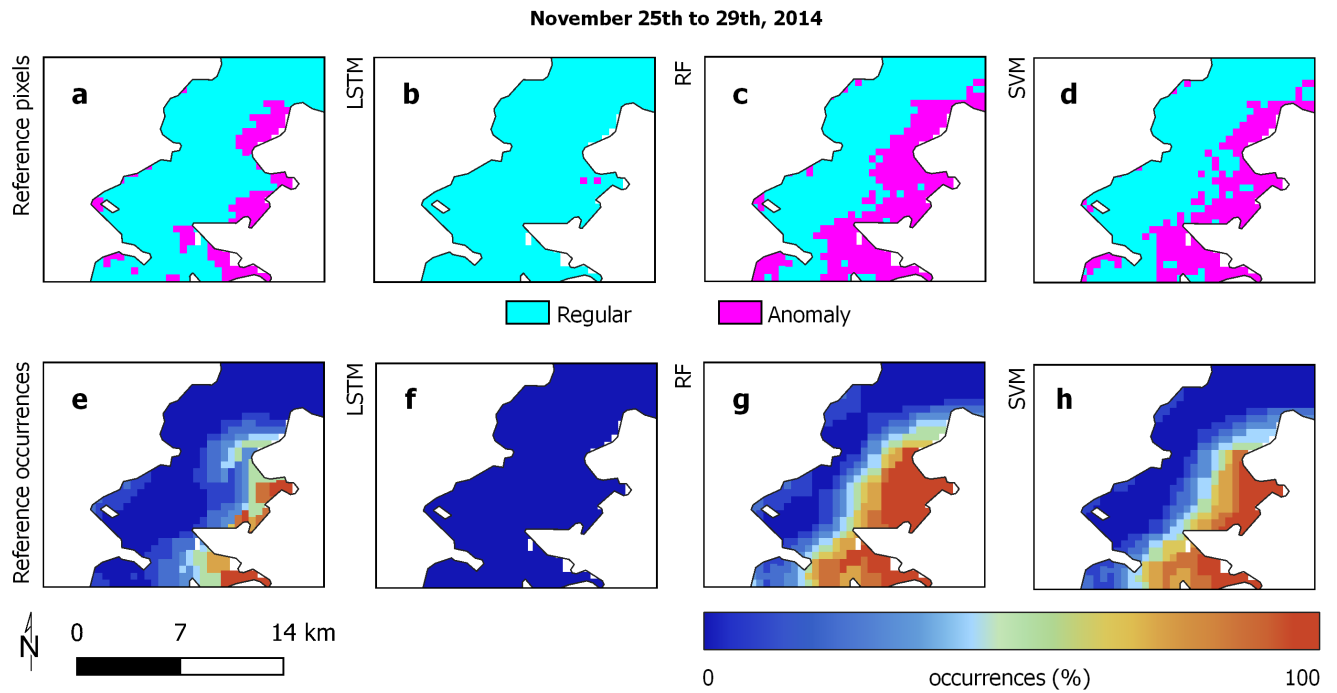


Figure 14: Period trend forecasting by pixels (a)-(d) and occurrences (e)-(h) from November 25th–29th, 2014, related to Lake Chilika. In this sense, (a) and (b) represent the mapping of the reference data.

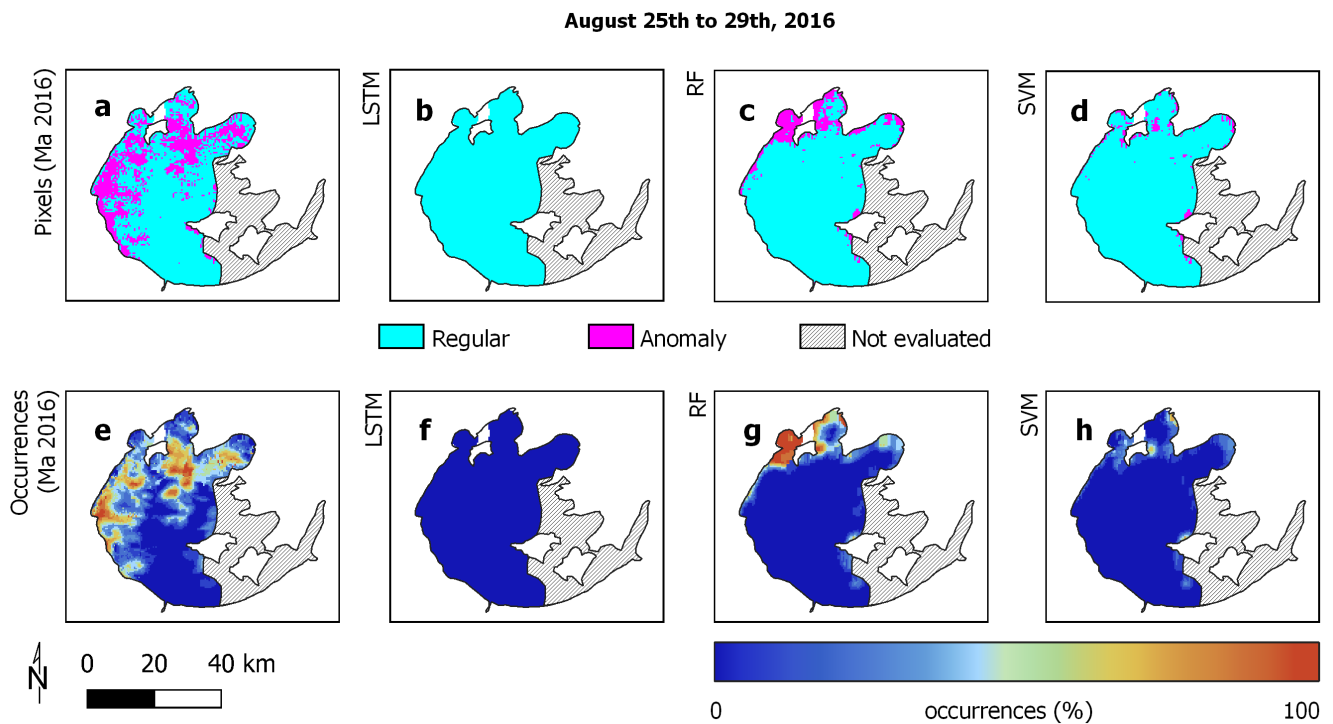


Figure 15: Period trend forecasting by pixels (a)-(d) and occurrences (e)-(h) from August 25th–29th, 2016, related to Lake Taihu. In this sense, (a) and (b) represent the mapping of the reference data.

[40] A. K. Mishra and P. Coulibaly. Developments in hydrometric network design: A review. *Reviews of Geophysics*, 47(2), 2009.

July 25th to 29th, 2017

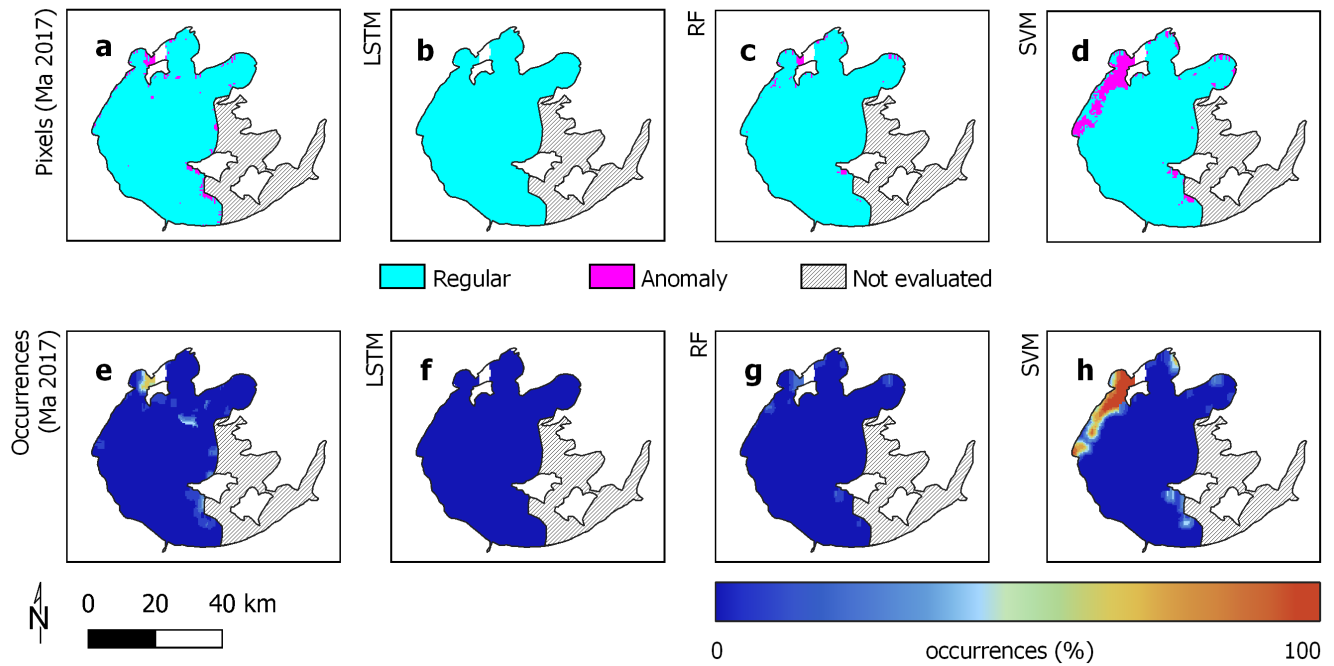


Figure 16: Period trend forecasting by pixels (a)-(d) and occurrences (e)-(h) from July 25th–29th, 2017, related to Lake Taihu. In this sense, (a) and (b) represent the mapping of the reference data.

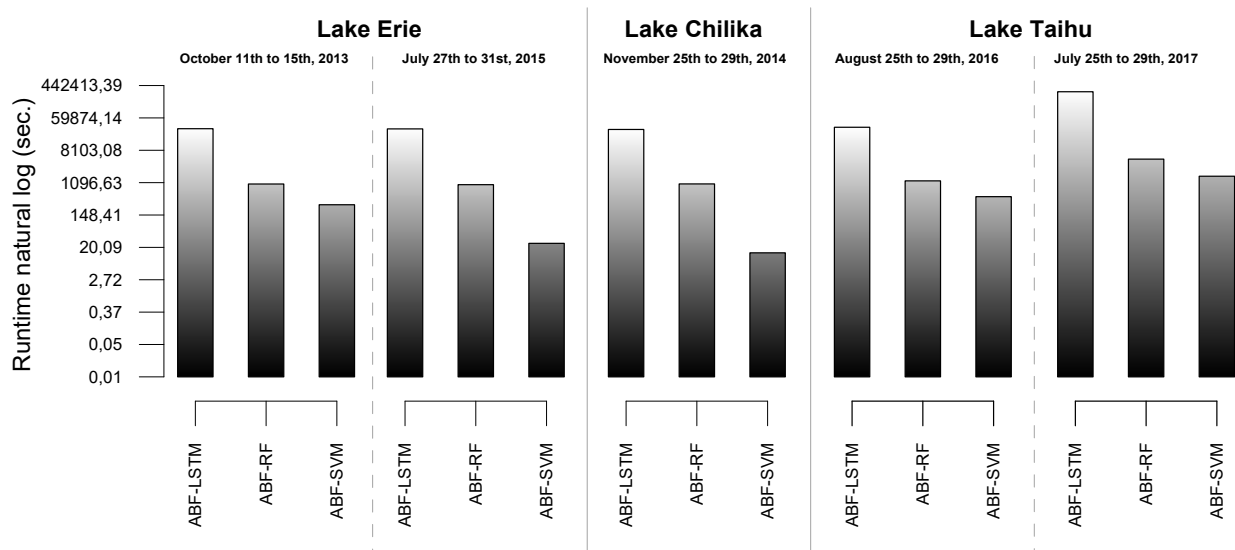


Figure 17: Median computational run-times for the three areas and periods evaluated.

- [41] G. Mountrakis, J. Im, and C. Ogole. Support Vector Machines in Remote Sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing Society*, 66(3):247–259, 2011.
- [42] Y. Oyama, T. Fukushima, B. Matsushita, H. Matsuzaki, K. Kamiya, and H. Kobinata. Monitoring levels of cyanobacterial blooms using the visual cyanobacteria index (vci) and floating algae index (fai). *International Journal of Applied Earth Observation and Geoinformation*, 38:335–348, 2015.
- [43] J. K. Panigrahi. Water quality, biodiversity and livelihood issues: A case study of chilika lake, india. In *2007 Atlanta Conference on Science, Technology and Innovation Policy*, pages 1–6, 2007.
- [44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

- [45] L. Qi, C. Hu, H. Duan, B. B. Barnes, and R. Ma. An eof-based algorithm to estimate chlorophyll a concentrations in taihu lake from modis land-band measurements: Implications for near real-time applications and forecasting models. *Remote Sensing*, 6(11):10694–10715, 2014.
- [46] R. Ranjan. A forestry-based pes mechanism for enhancing the sustainability of chilika lake through reduced siltation loading. *Forest Policy and Economics*, 106:101944, 2019.
- [47] L. Rastrigin. The convergence of the random search method in the extremal control of a many parameter system. *Automaton & Remote Control*, 24:1337–1342, 1963.
- [48] M. Rodell, P. R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.-J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, and et al. The global land data assimilation system. *Bulletin of the American Meteorological Society*, 85(3):381394, Mar 2004.
- [49] J. W. Rouse Jr, R. H. Haas, J. Schell, and D. Deering. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. *Texas A&M University College Station*, 1973.
- [50] B. Z. Rouso, E. Bertone, R. Stewart, and D. P. Hamilton. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Research*, 182:115959, 2020.
- [51] M. Sengupta, R. Anandurai, S. Nanda, and A. A. Datti. Geospatial identification of algal blooms in inland waters: A post cyclone case study of chilka lake, odisha, india. *RASAYAN Journal of Chemistry*, 10(1):234–239, 2017.
- [52] J. Shawe-Taylor, N. Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- [53] K. Shi, Y. Zhang, B. Qin, and B. Zhou. Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. *Science Bulletin*, 64(20):1540–1556, 2019.
- [54] L. Silveira Kupssinskü, T. Thomassim Guimarães, E. Menezes de Souza, D. C. Zanotta, M. Roberto Veronez, L. Gonzaga, and F. F. Mauad. A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. *Sensors*, 20(7), 2020.
- [55] R. P. Stumpf, T. T. Wynne, D. B. Baker, and G. L. Fahnenstiel. Interannual variability of cyanobacterial blooms in lake erie. *PLOS ONE*, 7(8):1–11, 08 2012.
- [56] J. Takaku, T. Tadono, M. Doutsu, F. Ohgushi, and H. Kai. Updates of aw3d30 alos global digital surface model with other open access datasets. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B4-2020:183–189, 2020.
- [57] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, San Diego, 4th edition, 2008.
- [58] USGS. MODIS/Terra Surface Reflectance Daily L2G Global 1 km and 500 m, 2020. Accessed: 2020-03-23.
- [59] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, 2011.
- [60] C. J. van Rijsbergen. *Information retrieval*. Butterworths, London, 2 edition, 1979.
- [61] G. van Rossum and F. L. Drake. *The Python Language Reference Manual*. Network Theory Ltd., 2011.
- [62] V. Vapnik. Estimation of dependences based on empirical data berlin, 1982.
- [63] J. V. Vilán, J. A. Fernández, P. G. Nieto, F. S. Lasheras, F. de Cos Juez, and C. D. Muñiz. Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the trasona reservoir (northern spain). *Water resources management*, 27(9):3457–3476, 2013.
- [64] Z. Wan, S. Hook, and G. Hulley. Mod11a1 modis/terra land surface temperature/emissivity daily 13 global 1km sin grid v006. *NASA EOSDIS Land Processes DAAC*, 2015. Type: dataset.
- [65] M. Wang, M. Stokral, P. Burek, C. Kroeze, L. Ma, and A. B. Janssen. Excess nutrient loads to lake taihu: Opportunities for nutrient reduction. *Science of The Total Environment*, 664:865–873, 2019.
- [66] X. Wang, H. Yan, C. Huo, J. Yu, and C. Pant. Enhancing pix2pix for remote sensing image classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2332–2336, 2018.
- [67] Z. Wang, C. Liu, and A. Huete. From avhrr-ndvi to modis-evi: Advances in vegetation index research. *Acta ecologica sinica*, 23(5):979–987, 2003.
- [68] M. L. Wells, V. L. Trainer, T. J. Smayda, B. S. Karlson, C. G. Trick, R. M. Kudela, A. Ishikawa, S. Bernard, A. Wulff, D. M. Anderson, et al. Harmful algal blooms and climate change: learning from the past and present to forecast the future. *Harmful algae*, 49:68–93, 2015.
- [69] H. Xu. Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14):3025–3033, 2006.
- [70] H. Xu. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International journal of remote sensing*, 27(14):3025–3033, 2006.
- [71] H. Xu, H. W. Paerl, B. Qin, G. Zhu, N. S. Hall, and Y. Wu. Determining critical nutrient thresholds needed to control harmful cyanobacterial blooms in eutrophic lake taihu, china. *Environmental Science & Technology*, 49(2):1051–1059, 2015. PMID: 25495555.
- [72] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020.
- [73] T. Zhang, M. Huang, and Z. Wang. Estimation of chlorophyll-a concentration of lakes based on svm algorithm and landsat 8 oli images. *Environmental Science and Pollution Research*, pages 1–14, 2020.
- [74] D. Zhao. Application of NDVI to detecting algal bloom in the Bohai Sea of China from AVHRR. In *Ocean Remote Sensing and Applications*, volume 4892, pages 241–246. International Society for Optics and Photonics, 2003.
- [75] M. Zhu, G. Zhu, L. Zhao, X. Yao, Y. Zhang, G. Gao, and B. Qin. Influence of algal bloom degradation on nutrient release at the sediment–water interface in lake taihu, china. *Environmental Science and Pollution Research*, 20(3):1803–1811, July 2012.

Confirming submission to Environmental Modelling and Software

Environmental Modelling and Software <em@editorialmanager.com>

23 de julho de 2021 18:37

Responder a: Environmental Modelling and Software <envsoft@elsevier.com>

Para: Pedro Henrique Moraes Ananias <pedro.ananias@unesp.br>

This is an automated message.

Forecasting algal bloom events in inland waters using remote sensing data and machine learning methods

Dear Mr. Ananias,

We have received the above referenced manuscript you submitted to Environmental Modelling and Software.

To track the status of your manuscript, please log in as an author at <https://www.editorialmanager.com/envsoft/>, and navigate to the "Submissions Being Processed" folder.

Thank you for submitting your work to this journal.

Kind regards,
Environmental Modelling and Software

More information and support

You will find information relevant for you as an author on Elsevier's Author Hub: <https://www.elsevier.com/authors>

FAQ: How can I reset a forgotten password?

https://service.elsevier.com/app/answers/detail/a_id/28452/supporthub/publishing/kw/editorial+manager/

For further assistance, please visit our customer service site: <https://service.elsevier.com/app/home/supporthub/publishing/>. Here you can search for solutions on a range of topics, find answers to frequently asked questions, and learn more about Editorial Manager via interactive tutorials. You can also talk 24/7 to our customer support team by phone and 24/7 by live chat and email.

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: <https://www.editorialmanager.com/envsoft/login.asp?a=r>). Please contact the publication office if you have any questions.

3 CONSIDERAÇÕES FINAIS

Diante da necessária gestão de recursos hídricos e em face da problemática de proliferação de algas potencialmente tóxicas, bem como seus efeitos adversos para a sociedade, sejam relacionados à saúde ou às perdas econômicas, é fundamental o desenvolvimento de estudos que visem seu monitoramento e controle. Motivados por essas premissas, foram desenvolvidos dois algoritmos que se baseiam em abordagens de detecção e previsão de anomalias com estudos de caso sobre a ocorrência de proliferação de algas em águas interiores. Ambas as propostas foram apoiadas por SR, séries multi-temporais de imagens, índices espectrais, atributos meteorológicos/climáticos e modelos de AM.

O *Anomalous Behaviour Detection* (ABD) realiza a detecção do fenômeno com a aplicação do modelo OC-SVM em imagens extraídas dos sensores multi-espectrais Landsat-8 e MODIS. Em sua validação, foram realizados dois estudos de caso com dados *in situ*. Nesse sentido, comparações com propostas semelhantes e baseadas na limiarização dos índices NDVI e FAI foram incluídas nas análises. Ainda, a proposta original foi alterada a fim de utilizar o método RF no processo de classificação, tendo, por sua vez, seus resultados comparados à abordagem OC-SVM. Adicionalmente, em uma aplicação no mundo real, um mapeamento de 5 anos da ocorrência de proliferação de algas no Lago Taihu foi apresentado, demonstrando o comportamento espaço-temporal do fenômeno no corpo d'água.

O método proposto apresenta níveis de acurácia superiores aos concorrentes analisados, com valores médios de coeficiente kappa, acurácia global e F1-Score de 0.68, 87% e 86%, respectivamente; contra 0.47, 78% e 75% para ABD-RF; 0.28, 67% e 57% para NDVI; e - 0.04, 53% e 46% para FAI. Além disso, foi observado que os métodos testados superestimam a não ocorrência de proliferação de algas. Por outro lado, deve-se mencionar que a proposta é altamente dependente de conexão de rede e imagens com boa visibilidade (i.e. menos de 50% de nuvens) no conjunto de treinamento e requer um maior esforço computacional. Ainda assim, dada a falta de dados de verdade terrestre, não foi possível validar o comportamento da proposta em regiões com presença de neve ou gelo. Como mencionado, mudanças repentinas na resposta espectral do alvo podem ser detectadas como anomalias. O oposto é verdadeiro quando a ocorrência do fenômeno é muito sutil, sendo detectado como regular. Adicionalmente, é esperada uma variação positiva ou negativa de desempenho em diferentes instantes, tanto para a proposta quanto para os índices avaliados.

O *Anomalous Behaviour Forecasting* (ABF), diferentemente do primeiro, prevê (também de forma totalmente automática) a ocorrência de proliferação de algas em ambientes aquáticos utilizando-se de imagens Modis, atributos meteorológicos/climáticos e algoritmos de AM. Paralelamente, a proposta foi explorada considerando três variantes diferentes, as quais fazem uso dos métodos RF, SVM e LSTM como modelos de classificação.

Os resultados obtidos apontaram que a variante ABF-RF como de maior regularidade e menor custo computacional, obtendo valores médios de acurácia global, F1-Score, coeficiente

kappa, R^2 -Score e RMSE entre 95%, 94%, 0.51, 0.38 e 8.22, respectivamente, contra 90%, 92%, 0.51, 0.61 e 20.98 para o ABF-SVM; e 95%, 93%, 0.0, 0.0 e 8.05 para o ABF-LSTM. Ainda, foram avaliados os pontos fracos do presente trabalho. O primeiro refere-se à necessidade de imagens com boa visibilidade (i.e. menos de 50% de nuvens/sombras). Caso o conjunto de imagens seja extremamente pequeno, não haverá uma extração suficiente das características temporais do fenômeno estudado. Ainda, a extensão da área de estudo deve impactar significativamente no custo computacional. Por fim, a execução do algoritmo está condicionada à disponibilidade de conexão de rede e dados do sensor NASA GLDAS 2.1, já que este apresenta defasagem média de dois meses nos dados disponibilizados pela plataforma GEE em comparação com o produto Modis MOD09GA.006.

Ambos os algoritmos ABD e ABF, através dos resultados discutidos anteriormente, demonstram grande potencial na redução de custos e aumento de performance em detecção, monitoramento e previsão de floração de algas nos ambientes de águas interiores globais. Embora o custo computacional seja elevado, o benefício é evidente quando se leva em consideração a saúde dos seres vivos e outros potenciais impactos ambientais que o fenômeno pode causar.

Finalmente, perspectivas futuras para o presente estudo incluem: (i) o desenvolvimento de estratégias alternativas para reduzir o custo computacional dos modelos; (ii) avaliar o comportamento de outros índices espectrais no processo de detecção e previsão a fim de ampliar as suas aplicações; (iii) para o ABF, minimizar o efeito da defasagem de dados no sensor NASA GLDAS 2.1; (iv) avaliar a aplicação dos algoritmos em novos estudos de caso, como queimadas, desmatamento e alagamentos.

REFERÊNCIAS

- ALAWADI, F. Detection of surface algal blooms using the newly developed algorithm surface algal bloom index (SABI). In: JR., C. R. B. et al. (Ed.). **Remote Sensing of the Ocean, Sea Ice, and Large Water Regions 2010**. [S.l.]: SPIE, 2010. v. 7825, p. 45 – 58.
- ALLEN, J. I. et al. How well can we forecast high biomass algal bloom events in a eutrophic coastal sea? **Harmful Algae**, Elsevier, v. 8, n. 1, p. 70–76, 2008.
- BARZEGAR, R.; AALAMI, M. T.; ADAMOWSKI, J. Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. **Stochastic Environmental Research and Risk Assessment**, Springer, p. 1–19, 2020.
- BEYRUTH, Z. Periodic disturbances, trophic gradient and phytoplankton characteristics related to cyanobacterial growth in Guarapiranga reservoir, São Paulo State, Brazil. In: **The Trophic Spectrum Revisited**. [S.l.]: Springer, 2000. p. 51–65.
- BINDING, C. et al. An analysis of satellite-derived chlorophyll and algal bloom indices on Lake Winnipeg. **Journal of Great Lakes Research**, Elsevier, v. 44, n. 3, p. 436–446, 2018.
- BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.
- CARVALHO, M. d. C. et al. Manual de cianobactérias planctônicas: legislação, orientações para o monitoramento e aspectos ambientais. **Governo do Estado São Paulo. Secretaria do Meio Ambiente**, 2013.
- CASTRO, N. O. de; MOSER, G. A. de O. Florações de algas nocivas e seus efeitos ambientais. **Oecologia Australis**, v. 16, n. 2, p. 235–264, 2012.
- CHAWLA, I.; KARTHIKEYAN, L.; MISHRA, A. K. A review of remote sensing applications for water security: Quantity, quality, and extremes. **Journal of Hydrology**, Elsevier, p. 124826, 2020.
- CHO, H.; CHOI, U.; PARK, H. Deep learning application to time-series prediction of daily chlorophyll-a concentration. **WIT Trans. Ecol. Environ**, v. 215, p. 157–63, 2018.
- CORTES, C.; VAPNIK, V. Support vector machine. **Machine learning**, v. 20, n. 3, p. 273–297, 1995.
- DUAN, H. et al. Distribution and incidence of algal blooms in Lake Taihu. **Aquatic Sciences**, Springer, v. 77, n. 1, p. 9–16, 2015.
- FLAGELLATES, T. Physiological ecology of red tide flagellates. **Biochemistry and physiology of protozoa**, Academic Press, v. 1, p. 357, 1979.
- FOWLER, H. G.; AGUIAR, A. M. D. D. Environmental impact assessment in Brazil. **Environmental Impact Assessment Review**, Elsevier, v. 13, n. 3, p. 169–176, 1993.
- FRANKLIN, J. B. et al. A novel approach to predict chlorophyll-a in coastal-marine ecosystems using multiple linear regression and principal component scores. **Marine Pollution Bulletin**, Elsevier, v. 152, p. 110902, 2020.

- GOWER, J. et al. Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer. **International Journal of Remote Sensing**, Taylor & Francis, v. 26, n. 9, p. 2005–2012, 2005.
- GRIMMOND, S. U. Urbanization and global environmental change: local effects of urban warming. **Geographical Journal**, Wiley Online Library, v. 173, n. 1, p. 83–88, 2007.
- HAN-QIU, X. A study on information extraction of water body with the modified normalized difference water index (mndwi). **Journal of remote sensing**, v. 5, p. 589–595, 2005.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- HU, C. A novel ocean color index to detect floating algae in the global oceans. **Remote Sensing of Environment**, Elsevier, v. 113, n. 10, p. 2118–2129, 2009.
- JR, J. W. R. et al. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. 1973.
- KUPSSINSKÜ, L. S. et al. A method for chlorophyll-a and suspended solids prediction through remote sensing and machine learning. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 20, n. 7, p. 2125, 2020.
- LECUN, Y. et al. Generalization and network design strategies. **Connectionism in perspective**, Elsevier Zurich, Switzerland, v. 19, p. 143–155, 1989.
- LEE, M.-S. et al. Red tide detection using deep learning and high-spatial resolution optical satellite imagery. **International Journal of Remote Sensing**, Taylor & Francis, p. 1–23, 2019.
- MARTÍNEZ-ÁLVAREZ, F.; BUI, D. T. **Advanced Machine Learning and Big Data Analytics in Remote Sensing for Natural Hazards Management**. [S.l.]: Multidisciplinary Digital Publishing Institute, 2020.
- MATSUMURA-TUNDISI, T.; TUNDISI, J. G. Plankton richness in a eutrophic reservoir (Barra Bonita reservoir, SP, Brazil). **Hydrobiologia**, Springer, v. 542, n. 1, p. 367–378, 2005.
- MATTHEWS, M. W.; BERNARD, S.; ROBERTSON, L. An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters. **Remote Sensing of Environment**, Elsevier, v. 124, p. 637–652, 2012.
- MILENKOVIĆ, S. M. et al. The identification of chlorophyll and its derivatives in the pigment mixtures: Hplc-chromatography, visible and mass spectroscopy studies. **Adv. Technol**, v. 1, n. 1, p. 16–24, 2012.
- MISHRA, A. K.; COULIBALY, P. Developments in hydrometric network design: A review. **Reviews of Geophysics**, Wiley Online Library, v. 47, n. 2, 2009.
- MUNOZ-MARI, J. et al. Semisupervised one-class support vector machines for classification of remote sensing data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 48, n. 8, p. 3188–3197, Aug 2010. ISSN 0196-2892.
- NAGENDRA, H. et al. Remote sensing for conservation monitoring: Assessing protected areas, habitat extent, habitat condition, species diversity, and threats. **Ecological Indicators**, Elsevier, v. 33, p. 45–59, 2013.

- NOBRE, P. et al. Mudanças climáticas e desertificação: os desafios para o estado brasileiro. **Desertificação e mudanças climáticas no semiárido brasileiro. Campina Grande: INSA-PB**, p. 25–36, 2011.
- OGASHAWARA, I. et al. Cyanobacteria detection in Guarapiranga reservoir (São Paulo state, Brazil) using Landsat TM and ETM+ images. **Revista Ambiente & Água**, SciELO Brasil, v. 9, n. 2, p. 224–238, 2014.
- OYAMA, Y. et al. Monitoring levels of cyanobacterial blooms using the Visual Cyanobacteria Index (VCI) and Floating Algae Index (FAI). **International Journal of Applied Earth Observation and Geoinformation**, Elsevier, v. 38, p. 335–348, 2015.
- QI, L. et al. An EOF-based algorithm to estimate chlorophyll a concentrations in Taihu Lake from MODIS land-band measurements: Implications for near real-time applications and forecasting models. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 6, n. 11, p. 10694–10715, 2014.
- ROUSSO, B. Z. et al. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. **Water Research**, v. 182, p. 115959, 2020. ISSN 0043-1354.
- SATHYENDRANATH, S. et al. Remote sensing of phytoplankton pigments: a comparison of empirical and theoretical approaches. **International Journal of Remote Sensing**, Taylor & Francis, v. 22, n. 2-3, p. 249–273, 2001.
- SCHÖLKOPF, B. et al. Support vector method for novelty detection. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2000. p. 582–588.
- SHI, K. et al. Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. **Science Bulletin**, Elsevier, 2019.
- SONG, W. et al. Learning-based algal bloom event recognition for oceanographic decision support system using remote sensing data. **Remote Sensing**, Multidisciplinary Digital Publishing Institute, v. 7, n. 10, p. 13564–13585, 2015.
- STUMPF, R. P. et al. Interannual variability of cyanobacterial blooms in Lake Erie. **PloS one**, Public Library of Science, v. 7, n. 8, p. e42444, 2012.
- SUN, D.; LI, Y.; WANG, Q. A unified model for remotely estimating chlorophyll-a in Lake Taihu, China, based on SVM and in situ hyperspectral data. **IEEE Transactions on Geoscience and Remote Sensing**, IEEE, v. 47, n. 8, p. 2957–2965, 2009.
- VILÁN, J. V. et al. Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain). **Water resources management**, Springer, v. 27, n. 9, p. 3457–3476, 2013.
- WATANABE, F. et al. Remote sensing of the chlorophyll-a based on OLI/Landsat-8 and MSI/Sentinel-2A (Barra Bonita reservoir, Brazil). **Anais da Academia Brasileira de Ciências**, SciELO Brasil, v. 90, n. 2, p. 1987–2000, 2018.
- WELLS, M. L. et al. Harmful algal blooms and climate change: learning from the past and present to forecast the future. **Harmful algae**, Elsevier, v. 49, p. 68–93, 2015.

XU, H. Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. **International Journal of Remote Sensing**, Taylor & Francis, v. 27, n. 14, p. 3025–3033, 2006.

YI, H.-S. et al. Algal bloom prediction using extreme learning machine models at artificial weirs in the Nakdong River, Korea. **International Journal of Environmental Research and Public Health**, Multidisciplinary Digital Publishing Institute, v. 15, n. 10, p. 2078, 2018.

YUAN, Q. et al. Deep learning in environmental remote sensing: Achievements and challenges. **Remote Sensing of Environment**, Elsevier, v. 241, p. 111716, 2020.

ZHANG, T.; HUANG, M.; WANG, Z. Estimation of chlorophyll-a concentration of lakes based on SVM algorithm and Landsat 8 OLI images. **Environmental Science and Pollution Research**, Springer, p. 1–14, 2020.

ZHAO, D. Application of NDVI to detecting algal bloom in the Bohai Sea of China from AVHRR. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Ocean Remote Sensing and Applications**. [S.l.], 2003. v. 4892, p. 241–246.