Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR037138

Special Collection:
Advances in Machine Learning for Earth Science: Observation, Modeling, and Applications

Forecasting Lakes' Chlorophyll Concentrations Using Satellite Images and Generative Adversarial Networks

Nikolaos Nagkoulis[1], Giorgos Vasiloudis[1], Anastasia Moumtzidou[1], Ilias Gialampoukidis[1], Stefanos Vrochidis[1], and Ioannis Kompatsiaris[1]

[1]Centre for Research and Technology Hellas (CERTH), Information Technologies Institute (ITI), Thessaloniki, Greece

Key Points:

- Use of Sentinel 2 data for creating inland water bodies' Chlorophyll-α time-series
- Generative Adversarial Networks are trained to recognize Chlorophyll-α spatio-temporal patterns
- A continental-scale model is created for Chlorophyll-α short term predictions

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

A. Moumtzidou,
moumtzid@iti.gr

Citation:

Nagkoulis, N., Vasiloudis, G., Moumtzidou, A., Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2024). Forecasting lakes' chlorophyll concentrations using satellite images and generative adversarial networks. *Water Resources Research*, *60*, e2024WR037138. https://doi.org/10.1029/2024WR037138

Received 6 FEB 2024
Accepted 24 AUG 2024

Abstract Satellite data are extensively used for water quality monitoring purposes, offering a significantly reduced cost compared to in situ data sampling. Using past measurements to predict future conditions remains a challenging task, because of the complexity of the natural phenomena that are involved, with great potential in terms of water resources management. This paper proposes a model that can be used to forecast Chlorophyll-α (Chl-α) values in water bodies, which are a common water quality indicator. The operation of the model lays on the fact that typically Chl-α increases and decreases periodically. First, we apply C2RCC, which is a common atmospheric correction algorithm, to Sentinel-2 images to get Chl-α maps for 15 lakes for 12 consecutive months around Europe. Then, we use this data set (~1,000 Sentinel-2 images) to train a Generative Adversarial Network (GAN) to recognize spatiotemporal patterns. To accomplish this task, pix2pix algorithm is employed, matching consecutive past and current Chl-α maps to future Chl-α maps. This model has been applied to 3 water bodies around Europe that are not included in the 15-lakes training data set and has been found to perform accurately, achieving high Pearson and Spearman correlations and low RMSE values. Overall, the model can be used to make Chl-α maps' predictions with low computational cost and without using any in situ data and without the requirement of training for every water body.

1. Introduction

Algal blooms are physical phenomena that are caused by high concentrations of algae in water bodies and can be dangerous for both ecosystems and human life. As algae is not always dangerous, scientists are mostly concerned about when it exceeds some thresholds, or when its synthesis has some particular characteristics. A harmful algal bloom (HAB) occurs when toxin-producing algae grow excessively in a body of water. There are also numerous examples in geographic regions, where increases in nutrient loading have been linked with the development of large biomass blooms, leading to anoxia and even toxic or harmful impacts on fisheries resources, ecosystems, and human health or recreation. Many of these regions have witnessed reductions in phytoplankton biomass (quantified as chlorophyll-*a*) or HAB incidence when nutrient controls were put in place (Anderson et al., 2002). Many studies have shown that algal blooms are prone to occur under the conditions of low wind speed, suitable air temperature, and sunshine (Fitch & Moore, 2007; Mu et al., 2021). One of the variables that has grasped the attention of the scientific community is Chlorophyll-α (Chl-α), which has been associated with a variety of water quality indicators, including algal biomass concentrations (Ferral et al., 2017).

The monitoring of the quality of water bodies has been significantly aided during the last decades by the employment of satellite data. Satellite data can be used to easily obtain water quality indicators, offering a cost reduction compared to in situ data sampling. This cost reduction can have a significant impact in many regions (Sheffield et al., 2018). Moreover, satellite data can become a source of extensive time series' data sets (Ross et al., 2019) that can be used to apply machine learning techniques. Apart from typical Chlorophyll indices, a variety of algorithms is currently used to obtain current Chl-α values, mostly based on Neural Networks trained using in situ data (Pahlevan et al., 2020). In this paper, C2RCC (Brockmann et al., 2016), a bio-optical model widely used for atmospheric correction purposes, is used to get Chl-α values from Sentinel 2 SAFE images. These Chl-α values are used in this paper in order to train an algorithm to recognize the spatio-temporal variations of Chl-α and make short-term predictions.

Algal concentrations appear to have some seasonality, which is mainly due to the life cycle of organisms such as Cyanobacteria (Hieronymus et al., 2021). As detecting cyanobacteria and other types of organisms using remote sensing methods is extremely difficult, researchers use Chl-α as an indicator of the quality of a water body. Falconer et al. (1999) provided guidance using Chl-α concentrations mapping effects with toxin concentrations.

The cycles of cyanobacterial growth present some seasonality, resulting in some high values during specific months in some locations (Hieronymus et al., 2021). One of the parameters that affect that seasonality is the temperature. Harmful cyanobacteria such as Microcystis have been found to have an optimal temperature for growth and photosynthesis at, or above, 25°C (Davis et al., 2009). At the start of a cycle, the population remains stable, but then it tends to rise gradually until it reaches a certain steady state. Following a period of constancy, the population eventually begins to decline (Wang et al., 2015). This behavior does not only appear in laboratories but also in actual water bodies. Unmanned aircraft system (UAS) hyperspectral imagery has been successfully used to monitor and model cyanobacteria life cycle (Pokrzywinski et al., 2022). Similarly in many inland water bodies, Chl-α tends to increase during the months when temperature rises, and decrease afterward. In this paper we take advantage of this seasonality to make prediction about future Chl-α concentrations.

In terms of forecasting, in situ Chl-α concentration time-series have been used to create models that can be used to make future projections (German et al., 2020). Moreover, in Lee and Lee (2018) an LSTM model is proposed for producing short-term algal bloom predictions. Similarly, in the work of Abdul Wahid and Arunbabu (2022), the authors demonstrate the improvement in the chlorophyll-a estimation in the Krishnagiri reservoir in India when integrating remote sensing and in situ measurements by using multiple regression equations. Although these methods can offer accurate results, they are spatially limited. At the same time, in situ monitoring stations may be biased systematically toward high or low chlorophyll concentrations, because of the non-homogeneous distribution of chlorophyll values in most water bodies (Lehmann et al., 2021). Hydrodynamic models have also been used in order to estimate the spatial and temporal distribution of water quality parameters (Romas et al., 2018).

In this paper, we propose Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) to make short-term Chl-α forecasts, using Sentinel 2 images. In remote sensing GANs have been used for a variety of problems including among others cloud removal (Darbaghshahi et al., 2021), fog detection (Huang et al., 2021) and change detection (Chen et al., 2020). GANs have recently been used for water quality monitoring. A CycleGAN has been developed to monitor water distribution networks water quality and inform an alarming system (Li et al., 2023). Combined sewer overflows have been considered using GANs to detect anomalies and generate missing values regarding water depth (Koochali et al., 2024). A PGGAN has also been used to create high resolution river images (Gautam et al., 2022). Regarding remote sensing GANs applications in water resources, a cGAN has been designed for water bodies to create high resolution images from lower resolution images, focusing on river boundaries (Filali Boubrahimi et al., 2024). To enhance algal prediction performance generative adversarial networks (GANs) and self-attention mechanisms are used to monitor algae concentrations, providing improved accuracy (Huynh et al., 2022). Even though that the aforementioned literature indicates that GANs can be effectively used to address water resources research questions, to our knowledge this is the first time that GANs are used for short term water quality forecasting, using satellite data.

To perform predictions, pix2pix (Isola et al., 2017), a conditional GAN (cGAN) is used in this paper. Pix2pix is mostly used to translate an image to another image recognizing the patterns between the images, for example, translating a historical map to a map similar to satellite maps (Andrade & Fernandes, 2020). Taking advantage of the model's simplicity, in this paper we use pix2pix to translate a current image to a future image, enabling us to do short term predictions. Contrary to the common Chl-α approaches that produce forecasts based on specific points (monitoring stations), we propose a new method that can be used to create forecasted Chl-α concentration maps, providing values for the whole water body examined, by applying an image-to-image translation Generative Adversarial Network (Pix2Pix). The model is trained using Chl-α concentration maps and it is applied to maps that are not included in the training data set. To the best of our knowledge, this is the first time that Chl-α concentration maps are created based on short term predictions from satellite data.

The paper is organized as follows. In Section 2, we discuss the proposed methodology, the data set and the objective of the methodology. In Section 3, we present the results of the model, while Section 4 discusses the limitations and future work. Finally, Section 5 presents some concluding remarks.

2. Materials and Methods

2.1. General Overview

The model created is able to create future Chl-α maps in lakes around the world, using the 3 last Chl-α maps before the target date. To train the algorithm it is necessary to build data sets of consecutive Chl-α maps from a

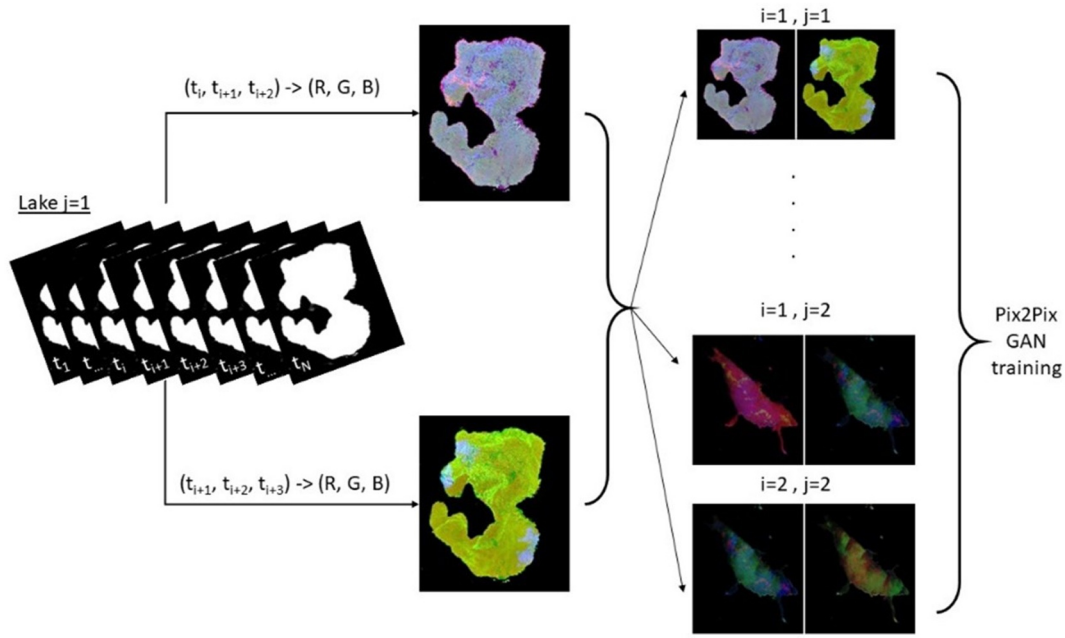

Figure 1. Overview of the method proposed. A data set of images is arranged so that Chl-α maps for a number of lakes are matched to consecutive Chl-α maps. The GAN is trained using sets of RGB images. In this image i is used as a time variable and j as a space variable.

variety of water bodies. This variety allows us to apply the model to new water bodies, which are not included in the training data set. The data sets are grouped into time-series image stacks and imported in a typical pix2pix model (Figure 1). The basic idea is that the model learns to recognize the spatiotemporal patterns between the images, by capturing the cyclic phenological patterns.

Given 3 consecutive Chl-α images that can be represented using a time triplet (t_i, t_{i+1}, t_{i+2}), the model learns to match them with 3 consecutive Chl-α images represented as $(t_{i+1}, t_{i+2}, t_{i+3})$. Considering that the overall number of available Chl-α images is N, the aforementioned process makes it possible to make the following representation using a simple induction: $(C_{t_{N-2}}, C_{t_{N-1}}, C_{t_N}) \rightarrow (C_{t_{N-1}}, C_{t_N}, C_{t_{N+1}})$, with C representing a Chl-α 2D plane (raster). This way, the target image $C_{t_{N+1}}$ that represents a future Chl-α map, can be obtained using the last three images.

2.2. Data Set Formation

To create the Chl-α Data sets that are needed to train the model, a pipeline is built using R and Python. The algorithm takes spatiotemporal input and creates Chl-α maps. The process is illustrated in Figure 2.

The steps are the following:

1. Spatiotemporal inputs. During this step a bounding box is defined for every lake. To create a model that will be able to be applied in many regions, a data set of 15 lakes distributed in Europe is created (Figure 3). We have chosen lakes from a variety of geographical latitudes and longitudes to increase climatic diversity. The climatic diversity improves the ability of the algorithm to provide accurate results to water bodies that are not included in the training data set. Regarding temporal discretization, 12 months are chosen for each body. Choosing 12 months makes it possible to have a full seasonal period for each water body, allowing the forecasting model to learn the temporal Chl-α patterns.
2. Download Sentinel 2 data. The Sentinel 2 L1C products, containing 12 bands are downloaded from the Copernicus Hub using the sen2r R package (Ranghetti et al., 2020). The process is aided by EODAG python package (CS GROUP, 2024) to access data older than 6 months from the date accessed. The Sentinel 2 products are available approximately every 5 days. For a data set of 15 water bodies and a duration of 12 months, using Sentinel 2 data we get a training data set of 1,080 images. Approximately 10% of them are removed from the training data set because of very high cloud coverage.

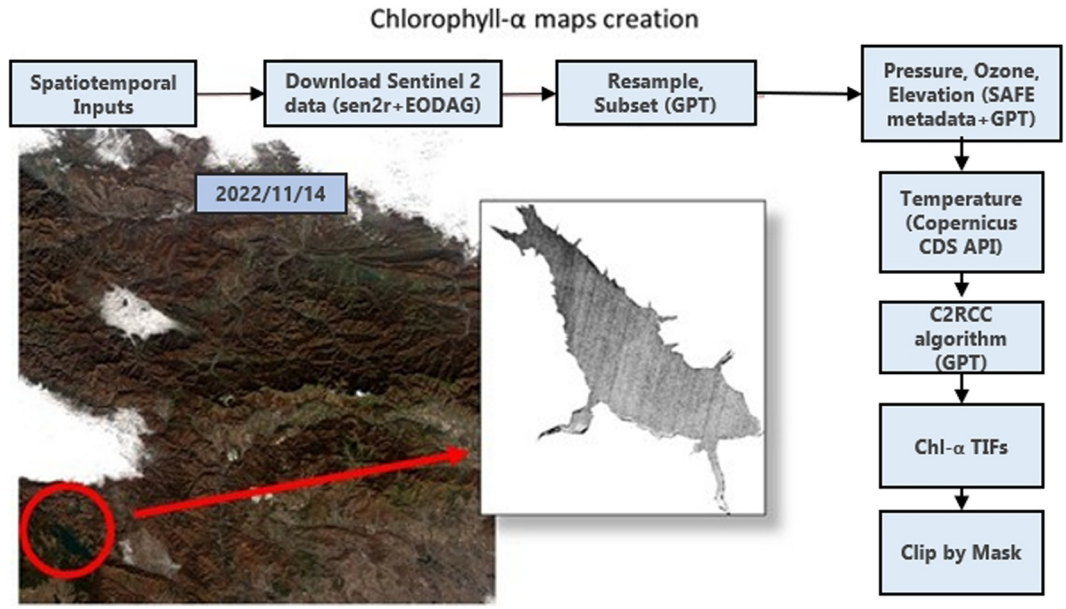

Figure 2. The process of creating the input data. The algorithm takes as inputs a spatial object and a temporal frame and gives an output of a TIFF raster with Chl-α values. Water body: Iskar, Bulgaria, used in the training data set.

3. Resample, Subset. Then, a resampling and subsetting process takes place, using GPT from Snap. Specifically, resampling and subsetting Snap graphs are created, edited through R and executed. The spatial resolution of the final products is 10 m.
4. Pressure, Ozone, Elevation data. The necessary pressure, ozone, elevation and temperature values to get the Chl-α outputs are obtained from the SAFE Sentinel 2 data.
5. Temperature data. Temperature data are downloaded from Copernicus CDS API.
6. C2RCC algorithm. The C2RCC model is used to get Chl-α maps. C2RCC is based on Neural Networks, trained in order to perform the inversion of spectrum for the atmospheric correction, as well as retrieval of inherent optical properties of the water body (Brockmann et al., 2016). C2RCC is a model that is widely used

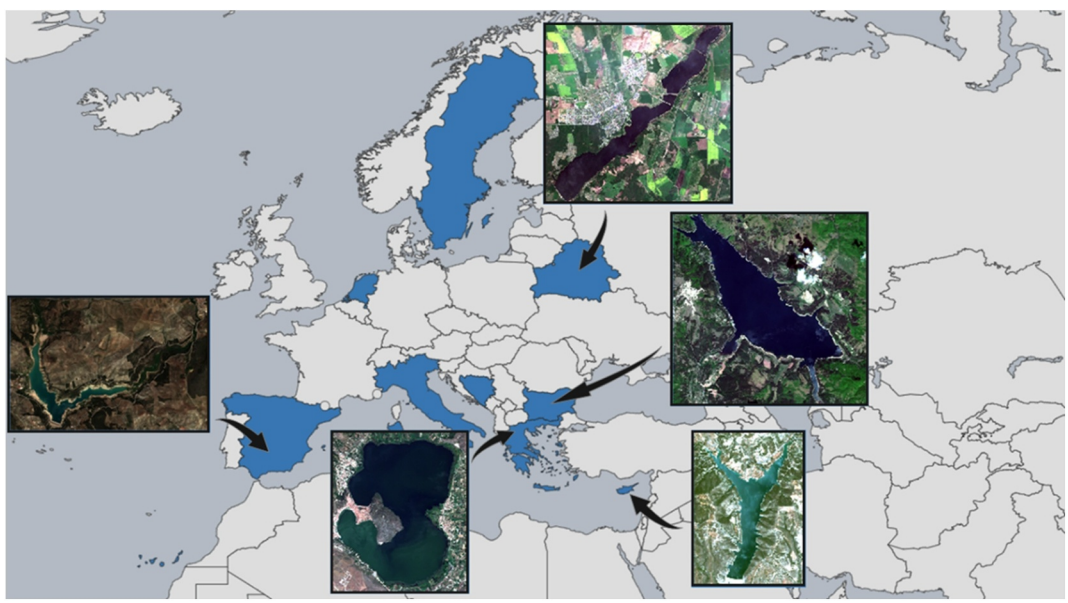

Figure 3. A visual representation of 15 water bodies, used to train the model. 5 of the 15 water bodies are presented in the plot. The water bodies chosen are located in different regions around Europe to maximize climatic diversity.

for water bodies and can be accessed through Snap interface and GPT model. C2RCC offers different levels of accuracy and can produce outputs even when all input environmental data are not available.

7. Chl-a TIFs. The Chl-α maps are grouped into sets of 3 to create RGB images and TIFS as presented in Section 2.1.
8. Clip by Mask. The output Chl-α maps obtained from C2RCC are clipped and masked, excluding neighboring water bodies, so that they correspond only to the examined water bodies.

To test the model, we have used the following 3 water bodies that are not included in the training data set. Similarly to the train data set, the test data set water bodies are obtained from different European regions: one from France (Der Chantecoq), one from Romania (Dridu) and one from Bulgaria (Pancharevo).

2.3. Data Processing

The outputs of the C2RCC model are geoTIFF files that are loaded as images in a Python environment and formulated in stacks of consecutive time frames as already mentioned. Each stack of Chl-α instances is matched to an actual color resulting in an RGB image. This makes it possible to use pix2pix model in order to map an RGB image to a consecutive RGB image.

Pix2pix is a conditional generative adversarial network (cGAN). It operates similarly to other cGANs, creating outputs conditionally to some inputs. However, contrary to most cGANs that create images "conditionally" to text descriptions, pix2pix creates images "conditionally" to other images. Supposing that a sequence of Sentinel 2 images are available for an area, pix2pix can use these images as a "condition" to generate new images, namely to operate short term predictions. Specifically, pix2pix consists of a generator G and a discriminator D. The generator is often mentioned in literature as "artist" and the discriminator as "art critique," because the generator creates images and the discriminator tries to find if these images are "real" or not. The generator and the discriminator typically in GANs play a zero-sum game, inspired from game theory (Goodfellow et al., 2020). An unconditional GAN can be expressed using the following function:

$$L_{\mathrm{GAN}}(G, D) = \mathbb{E}_y[\log D(y)] + \mathbb{E}_z[\log(1 - D(G(z)))] \tag{1}$$

where y represents a set of actual data and z represents a set of pseudo-data generated from G from random noise. D tries to maximize L, while G tries to minimize it. Finally, a Nash Equilibrium is reached in the minimax game: $G^* = \arg\min_G \max_D L_{\mathrm{GAN}}(G,D)$. In addition to random noise z conditional GANs use some conditional data x. This way cGANs learn a mapping from x and z to y: $G : x,z \rightarrow y$. This way Equation 1 is transformed into the following relationship:

$$L_{\mathrm{GAN}}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_z[\log(1 - D(x, G(x, z)))] \tag{2}$$

In this work, we follow the main process proposed by pix2pix GANs paper (Isola et al., 2017). Therefore, we use a distance parameter L_1 as follows:

$$L_{\mathrm{L1}}(G) = \mathbb{E}_{x,y,z}\left[||y - G(x, y, z)||_1\right] \tag{3}$$

And the equilibrium is reached when:

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{4}$$

When $G = G^*$, the discriminator can no longer distinguish if the generated images are real or not. In relationship Equation 4, λ represents a constant value. In this research, x stands for the RGB images that correspond to the (t_i, t_{i+1}, t_{i+2}) triplet and y stands for the RGB images that correspond to the $(t_{i+1}, t_{i+2}, t_{i+3})$ triplet. This way, the generator creates short-time predictions that are conditional to the existing Chl-α instances.

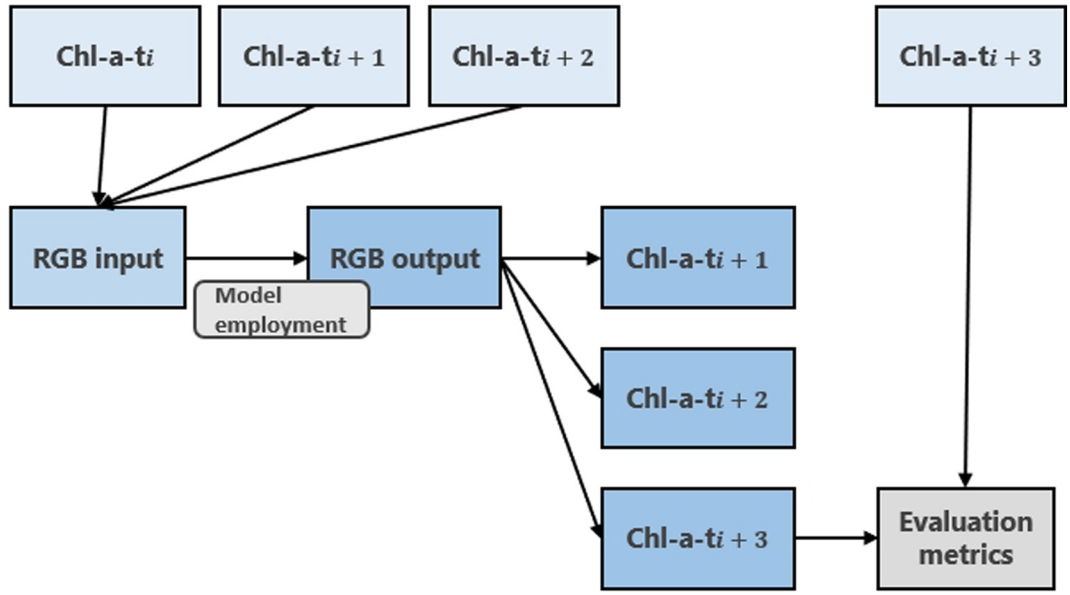

Figure 4. Evaluation of the forecasting ability of the algorithm during the testing phase by comparing its outputs to consecutive Chl-α instances.

2.4. Training and Testing

The pix2pix model is implemented based on the parameters proposed in the main pix2pix paper (Isola et al., 2017). The various Chl-α maps are resized so that their dimensions are $256 \times 256 \times 3$ px. and normalized to $[-1, 1]$.

The generator of the pix2pix is a U-Net that consists of an encoder and a decoder. The generator loss is a sigmoid cross-entropy loss of the generated images. A mean absolute error ($L1$ loss) between the generated image and the target image is also employed. The parameters have been set according to the initial pix2pix paper. The discriminator loss is the sum of two inputs. The first is a sigmoid cross-entropy loss of the real images and the second one is a sigmoid cross-entropy loss of the generated images.

Adam optimizer is used with a learning rate of $2 \cdot 10^{-4}$ and momentum parameter $\mu = 0.5$. The testing is performed in other lakes than those used for testing, in order to test how well the model can operate in European scale. 200 k steps are used for training the model. Pearson correlation coefficient, Spearman's Correlation coefficient (-1 for perfect negative correlation, $+1$ for perfect positive correlations) and Root Mean Square Error (RMSE) (between predicted and actual Chl-α values) are used to examine the similarity between the actual and the predicted images.

The testing phase and the comparison between the images is represented in Figure 4.

3. Results

This section presents the outputs of the pix2pix model for one of the water bodies of the training data set and also the evaluation metrics calculated for three water bodies of the testing data set. In general, even though the algorithm is trained so that it can produce Chl-α maps for all t_{i+1}, t_{i+2}, t_{i+3} dates, the prediction we are interested in is date t_{i+3}, because the other two are already known.

Kouris Lake in Limassol, Cyprus, is chosen as the water body from the training data set. In Figure 5, we can see the input image, the ground truth (GT) and the predicted image for a random date.

The evaluation metrics for specific case are the following: Pearson: 0.452, Spearman: 0.465, RMSE: 0.547. The evaluation metrics for Kouris dam are satisfactory and the predicted and the GT images look very similar due to ability of the model to understand the spatiotemporal patterns.

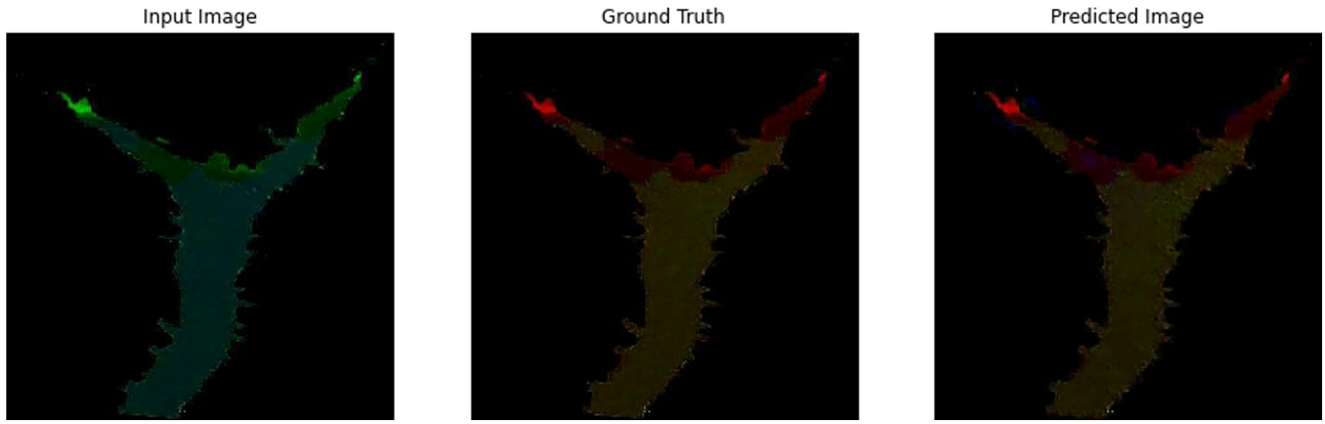

Figure 5. Results for Kouris lake. Input image contains 3 Chl-α bands that correspond to time periods t_i, t_{i+1}, t_{i+2}.

In the sequel, we present the evaluation metrics calculated for the three water bodies of the testing data set. The time period chosen is July 2022. In Figure 6, we can see the outputs of the pix2pix model for the Der-Chantecoq lake located in France.

In Figure 7, we can see the GT and the predicted images as well as the scatter plots between two Chl-α maps that refer to the time t_{i+3} for the three lakes comprising the testing data set (i.e., Pancharevo lake in Bulgaria, Der-Chantecoq lake in France, Dridu lake in Romani). In Table 1, we can see the metrics calculated from the results of the model. Specifically, the table presents for each lake, the values of Pearson correlation coefficient, Spearman's Correlation coefficient and Root Mean Square Error (RMSE) examining the similarity between the actual and the predicted images on a pixel basis along with Supporting Information S1 including the area of each lake, and the average actual and predicted Chl-α. The Chl-α maps are read as matrices so that the scatter plots can be formed matching each forecasted pixel to its GT equivalent.

In all cases the algorithm can successfully predict the shape of the water bodies and the spatial distribution of the Chl-α values. Also, we notice that there are high correlations between the GT chlorophyll values and the forecasted chlorophyll values. RMSE is lower for Pancharevo, because Pancharevo has generally lower chlorophyll values and it is smaller than the other 2 water bodies. Moreover, it is worth pointing out that the largest water body presents better results than the other two. Finally, it needs to be mentioned that July is chosen because we noticed that the accuracy of the model decreases in winter, when cloud coverage is higher.

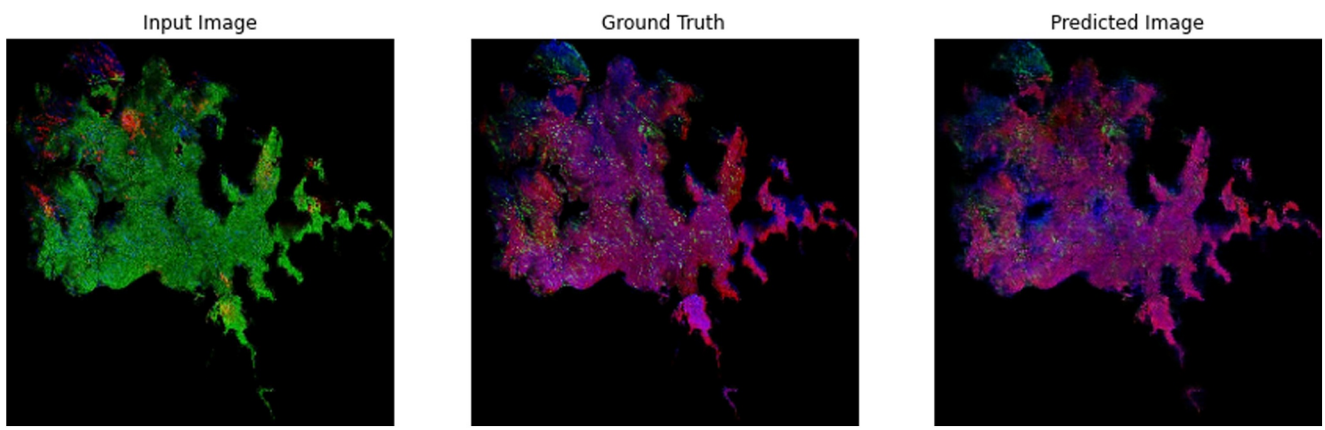

Figure 6. Results for Der-Chantecoq lake. The input image contains 3 Chl-α bands that correspond to time periods t_i, t_{i+1}, t_{i+2}. The algorithm uses these 3 dates to produce the predicted image containing the time period: $t_{i+1}, t_{i+2}, t_{i+3}$. The GT image corresponds to the actual $t_{i+1}, t_{i+2}, t_{i+3}$, used to test the accuracy of the prediction.

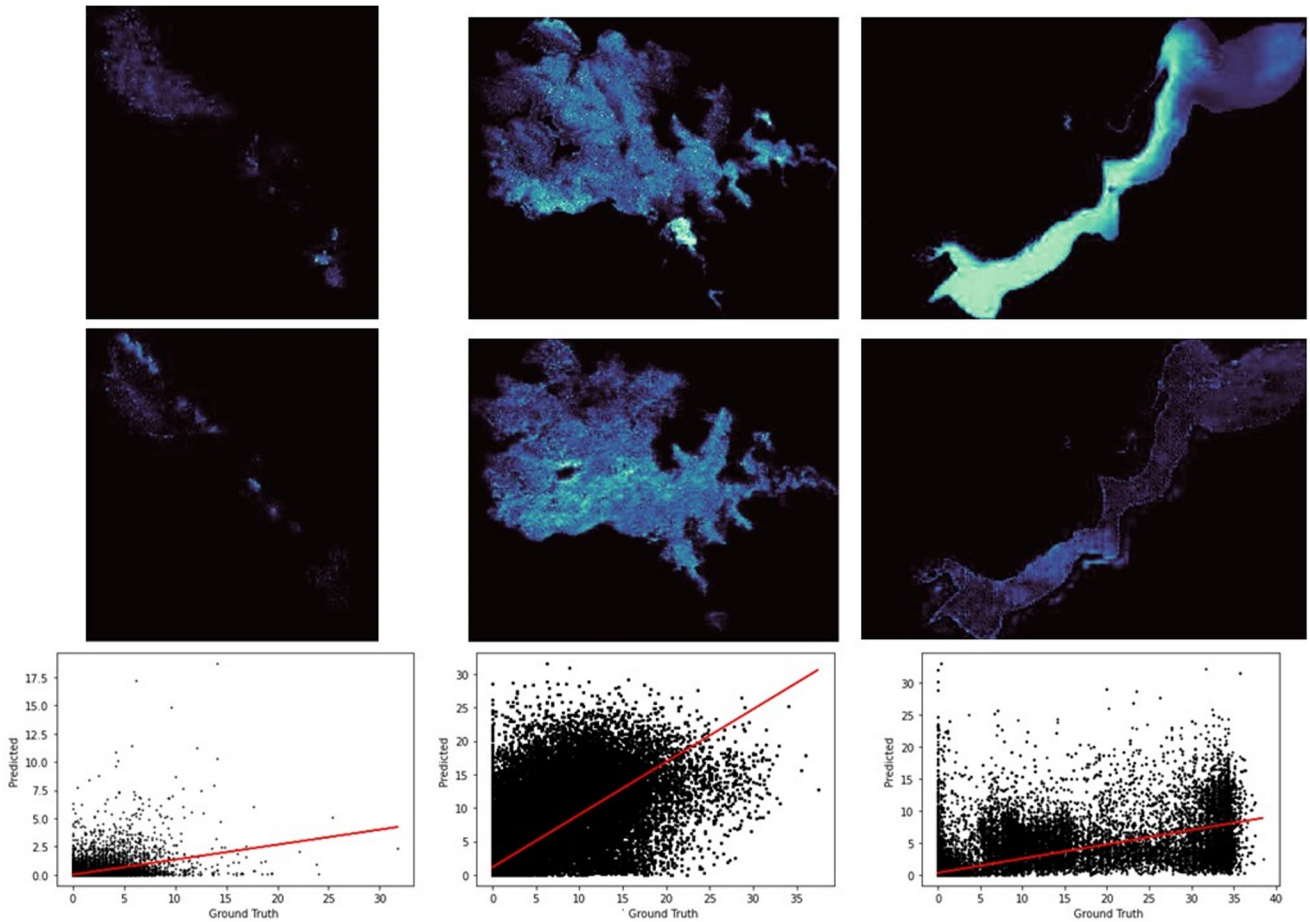

Figure 7. Chl-α maps and scatter plots for three water bodies (Pancharevo left, Der-Chantecoq middle and Dridu right). Top: GT images, middle: Predictions, bottom: Scatter plots. The coloring scales are different for each water body, but they are the same for the two rasters of each water body.

4. Discussion

The model presented can learn Chl-α spatiotemporal patterns and forecast Chl-α spatial concentrations. It has been trained using yearly data in order to understand Chl-α seasonality. However, under climate change conditions, this seasonality will most likely be affected. Organisms that are responsible for algal blooms, such as Cyanobacteria, prefer specific temperatures; thus, the temperature variations will affect the seasonality of the Chl-α values detected in water bodies. Training the model using wider time frames (e.g., 5 years) will allow the algorithm to more effectively adapt to seasonality variations. Similarly, training the algorithm using water bodies from different regions will enhance its ability to recognize more complex spatiotemporal Chl-α phenomena, thereby increasing its accuracy and broadening its spatial applicability.

Table 1
Lakes' Characteristics and Evaluation Metrics

	Pancharevo	Der-Chantecoq	Dridu
Pearson	0.4528	0.7536	0.7062
Spearman	0.5433	0.8480	0.7035
RMSE (mg/m^3)	1.2770	4.0090	7.6029
Area (km^2)	0.875	43.893	6.732
Actual Chl-α (mg/m^3)	0.388	3.255	3.540
Predicted Chl-α (mg/m^3)	0.103	3.654	1.112

Note. "Actual Chl-α" represents the actual average chlorophyll values of the lake and "Predicted Chl-α" represents the predicted average chlorophyll values of the lake.

The model's accuracy can be significantly improved through data fusion. Many disturbances in ecosystems are human induced. Human activities can result in sudden Chl-α increases and decreases, which can reduce the model's accuracy when they do not appear periodically. Incorporating human effects as input layers will reduce unexpected variations of the Chl-α values, improving the model's forecasting accuracy. Another parameter affecting the model outputs is cloud coverage. Although cloud coverage can act as a natural noise, helping the algorithm during training, extremely high cloud coverage can be equivalent to having no information, making forecasting impossible.

Removing highly cloudy images can improve pattern recognition; however, this also reduces the training data set, potentially decreasing the model's accuracy. Data fusion and cloud removal methods can be used to solve that dilemma (Shen et al., 2014). Future fine-tuning approaches can increase the model's accuracy by determining the cloud coverage level that maximizes the accuracy of the model. Finally, combining remote sensing with other environmental data is expected to enhance the algorithm's accuracy, allowing it to adopt more effectively to diverse climatic environments.

Although the model is trained using lakes, it has also been tested to one river. This test is not included in the main paper (but can be found in the supplementing material), as it should not be considered representative of the model's performance in rivers. The model is expected to present high accuracy when examining rivers due to the clearer underlying flow pattern compared to lakes (downstream). This makes rivers an interesting case study for future research. Similarly to river applications, flood forecasting using pix2pix and remote sensing data could be possible, as forecasting floods using DCGANs and simulated data has provided very interesting results (Cheng et al., 2021). Furthermore, the C2RCC atmospheric correction model used to obtain the current Chl-α concentrations can be easily adjusted for marine environments by proper, allowing to test the algorithm's ability to predict algal blooms in marine environments.

Other potential applications include forecasting other water quality values, assuming there is an underlying seasonality in the phenomenon. The proposed model differs from other forecasting approaches, such as LSTM, which are more applicable to multi-point observations rather than raster images used in this paper. However, future research could monitor specific points using satellite data and forecast their values to compare the algorithms' performance. Overall, this paper illustrates the potential of using pix2pix for spatiotemporal (non-water related) forecasting.

5. Conclusions

In this paper, we proposed a water bodies' Chlorophyll forecasting model using a pix2pix GAN and Sentinel 2 images. The model learns to recognize the spatiotemporal seasonality of the Chl-α values, being able to produce Chl-α maps. The accuracy of the model can increase should it be trained using more water bodies and using more metadata, such as temperature, as inputs. We believe that except from lakes, the same architecture can also be applied to monitor coastal water bodies and rivers around the world. Also, the main idea, which is using pix2pix model to handle with spatiotemporal data, could be useful in other disciplines as well. In summary, this study enables the prediction of Chl-α maps on a continental scale for the first time using a pretrained model, eliminating the need to download extensive sets of images and large data sets.

Data Availability Statement

Sentinel 2 mission belongs to the Copernicus programme which is funded by the European Union Space Programme. Sentinel-2 data are freely provided via the Copernicus Data Space, which can be accessed at https://dataspace.copernicus.eu/. The Chlorophyll-*a* values produced by the C2RCC processor were obtained using the SNAP—ESA Sentinel Application Platform v9.0.0, https://step.esa.int/main/. *Code Availability Statement*: The Python 3.7 code used for this publication can be found in the GitHub repository at https://github.com/M4D-MKLab-ITI/Chlorophyll-a-forecasting-with-Sentinel2, and is available under the Creative Commons license BY-NC-ND (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode.txt). After downloading the Sentinel-2 data and running C2RCC on the files, the Python file "run_forecast_on_many_lakes.py" available in the repository allows reproducing the model presented in this article. The required libraries for running the Python file are indicated in the file "dep_venv.yaml."

Acknowledgments
This work has been supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-883484 PathoCERT, and Horizon Europe-101118286 FUELPHORIA.

References

Abdul Wahid, A., & Arunbabu, E. (2022). Forecasting water quality using seasonal ARIMA model by integrating in-situ measurements and remote sensing techniques in Krishnagiri reservoir, India. *Water Practice and Technology*, *17*(5), 1230–1252. https://doi.org/10.2166/wpt.2022.046

Anderson, D. M., Glibert, P. M., & Burkholder, J. M. (2002). Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries*, *25*(4), 704–726. https://doi.org/10.1007/bf02804901

Andrade, H. J., & Fernandes, B. J. (2020). Synthesis of satellite-like urban images from historical maps using conditional GAN. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–4. https://doi.org/10.1109/lgrs.2020.3023170

Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S., & Ruescas, A. (2016). Evolution of the C2RCC neural network for sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. In *Living planet symposium* (Vol. 740, p. 54).

Chen, J., Zhao, W., & Chen, X. (2020). Cropland change detection with harmonic function and generative adversarial network. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/lgrs.2020.3023137

Cheng, M., Fang, F., Navon, I., & Pain, C. (2021). A real-time flow forecasting with deep convolutional generative adversarial network: Application to flooding event in Denmark. *Physics of Fluids*, *33*(5). https://doi.org/10.1063/5.0051213

CS GROUP. (2024). EODAG: Earth observation data access gateway. Retrieved from https://github.com/CS-SI/eodag

Darbaghshahi, F. N., Mohammadi, M. R., & Soryani, M. (2021). Cloud removal in remote sensing images using generative adversarial networks and SAR-to-optical image translation. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–9. https://doi.org/10.1109/tgrs.2021.3131035

Davis, T. W., Berry, D. L., Boyer, G. L., & Gobler, C. J. (2009). The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of microcystis during cyanobacteria blooms. *Harmful Algae*, *8*(5), 715–725. https://doi.org/10.1016/j.hal.2009.02.004

Falconer, I., Bartram, J., Chorus, I., Kuiper-Goodman, T., Utkilen, H., Burch, M., & Codd, G. (1999). Safe levels and safe practices. *Toxic Cyanobacteria in Water*, 155–178.

Ferral, A., Solis, V., Frery, A., Orueta, A., Bernasconi, I., Bresciano, J., et al. (2017). Spatio-temporal changes in water quality in an eutrophic lake with artificial aeration. *Journal of Water and Land Development*, *35*(1), 27–40. https://doi.org/10.1515/jwld-2017-0065

Filali Boubrahimi, S., Neema, A., Nassar, A., Hosseinzadeh, P., & Hamdi, S. M. (2024). Spatiotemporal data augmentation of MODIS-landsat water bodies using adversarial networks. *Water Resources Research*, *60*(3), e2023WR036342. https://doi.org/10.1029/2023wr036342

Fitch, D. T., & Moore, J. K. (2007). Wind speed influence on phytoplankton bloom dynamics in the southern ocean marginal ice zone. *Journal of Geophysical Research*, *112*(C8), C08006. https://doi.org/10.1029/2006jc004061

Gautam, A., Sit, M., & Demir, I. (2022). Realistic river image synthesis using deep generative adversarial networks. *Frontiers in Water*, *4*, 784441. https://doi.org/10.3389/frwa.2022.784441

German, A., Andreo, V., Tauro, C., Scavuzzo, C. M., & Ferral, A. (2020). A novel method based on time series satellite data analysis to detect algal blooms. *Ecological Informatics*, *59*, 101131. https://doi.org/10.1016/j.ecoinf.2020.101131

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Hieronymus, J., Eilola, K., Olofsson, M., Hense, I., Meier, H., & Almroth-Rosell, E. (2021). Modeling cyanobacteria life cycle dynamics and historical nitrogen fixation in the Baltic Proper. *Biogeosciences*, *18*(23), 6213–6227. https://doi.org/10.5194/bg-18-6213-2021

Huang, Y., Wu, M., Guo, J., Zhang, C., & Xu, M. (2021). A correlation context-driven method for sea fog detection in meteorological satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/lgrs.2021.3095731

Huynh, N. H., Böer, G., & Schramm, H. (2022). Self-attention and generative adversarial networks for algae monitoring. *European Journal of Remote Sensing*, *55*(1), 10–22. https://doi.org/10.1080/22797254.2021.2010605

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Koochali, A., Bakhshipour, A. E., Bakhshizadeh, M., Habermehl, R., Dilly, T. C., Dittmer, U., et al. (2024). Water depth prediction in combined sewer networks, application of generative adversarial networks. *Discover Applied Sciences*, *6*(3), 123. https://doi.org/10.1007/s42452-024-05787-4

Lee, S., & Lee, D. (2018). Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models. *International Journal of Environmental Research and Public Health*, *15*(7), 1322. https://doi.org/10.3390/ijerph15071322

Lehmann, M. K., Schütt, E. M., Hieronymi, M., Dare, J., & Krasemann, H. (2021). Analysis of recurring patchiness in satellite-derived chlorophyll a to aid the selection of representative sites for lake water quality monitoring. *International Journal of Applied Earth Observation and Geoinformation*, *104*, 102547. https://doi.org/10.1016/j.jag.2021.102547

Li, Z., Liu, H., Zhang, C., & Fu, G. (2023). Generative adversarial networks for detecting contamination events in water distribution systems using multi-parameter, multi-site water quality monitoring. *Environmental Science and Ecotechnology*, *14*, 100231. https://doi.org/10.1016/j.ese.2022.100231

Mu, M., Li, Y., Bi, S., Lyu, H., Xu, J., Lei, S., et al. (2021). Prediction of algal bloom occurrence based on the naive Bayesian model considering satellite image pixel differences. *Ecological Indicators*, *124*, 107416. https://doi.org/10.1016/j.ecolind.2021.107416

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., et al. (2020). Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sensing of Environment*, *240*, 111604. https://doi.org/10.1016/j.rse.2019.111604

Pokrzywinski, K., Johansen, R., Reif, M., Bourne, S., Hammond, S., & Fernando, B. (2022). Remote sensing of the cyanobacteria life cycle: A mesocosm temporal assessment of a microcystis sp. bloom using coincident unmanned aircraft system (UAS) hyperspectral imagery and ground sampling efforts. *Harmful Algae*, *117*, 102268. https://doi.org/10.1016/j.hal.2022.102268

Ranghetti, L., Boschetti, M., Nutini, F., & Busetto, L. (2020). "sen2r": An R toolbox for automatically downloading and preprocessing Sentinel-2 satellite data. *Computers & Geosciences*, *139*, 104473. https://doi.org/10.1016/j.cageo.2020.104473

Romas, E., Tzimas, A., Kandris, K., Pechlivanidis, I., Boultadakis, G., Giannakoulias, A., et al. (2018). Operational short-term water quantity and quality forecasting in reservoirs intended for potable water production. In *EGU general assembly conference abstracts* (p. 7090).

Ross, M. R., Topp, S. N., Appling, A. P., Yang, X., Kuhn, C., Butman, D., et al. (2019). AquaSat: A data set to enable remote sensing of water quality for inland waters. *Water Resources Research*, *55*(11), 10012–10025. https://doi.org/10.1029/2019wr024883

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., & Verbist, K. (2018). Satellite remote sensing for water resources management: Potential for supporting sustainable development in data-poor regions. *Water Resources Research*, *54*(12), 9724–9758. https://doi.org/10.1029/2017wr022437

Shen, H., Li, H., Qian, Y., Zhang, L., & Yuan, Q. (2014). An effective thin cloud removal procedure for visible remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *96*, 224–235. https://doi.org/10.1016/j.isprsjprs.2014.06.011

Wang, L., Fan, D., Chen, W., & Terentjev, E. M. (2015). Bacterial growth, detachment and cell size control on polyethylene terephthalate surfaces. *Scientific Reports*, *5*(1), 15159. https://doi.org/10.1038/srep15159