

## Article

# A Deep Learning Model Using Satellite Ocean Color and Hydrodynamic Model to Estimate Chlorophyll-*a* Concentration

Daeyong Jin <sup>1</sup>, Eojin Lee <sup>1</sup>, Kyonghwan Kwon <sup>2</sup> and Taeyun Kim <sup>1,\*</sup><sup>1</sup> Environment Data Strategy Center & Environmental Assessment Group, Korea Environment Institute, Sejong 30147, Korea; dyjin@kei.re.kr (D.J.); eojinlee@kei.re.kr (E.L.)<sup>2</sup> Ocean Environment Group, Oceanic, Seoul 07207, Korea; khkwon@oceanicnt.com

\* Correspondence: kimty@kei.re.kr; Tel.: +82-44-415-7415

**Abstract:** In this study, we used convolutional neural networks (CNNs)—which are well-known deep learning models suitable for image data processing—to estimate the temporal and spatial distribution of chlorophyll-*a* in a bay. The training data required the construction of a deep learning model acquired from the satellite ocean color and hydrodynamic model. Chlorophyll-*a*, total suspended sediment (TSS), visibility, and colored dissolved organic matter (CDOM) were extracted from the satellite ocean color data, and water level, currents, temperature, and salinity were generated from the hydrodynamic model. We developed CNN Model I—which estimates the concentration of chlorophyll-*a* using a  $48 \times 27$  sized overall image—and CNN Model II—which uses a  $7 \times 7$  segmented image. Because the CNN Model II conducts estimation using only data around the points of interest, the quantity of training data is more than 300 times larger than that of CNN Model I. Consequently, it was possible to extract and analyze the inherent patterns in the training data, improving the predictive ability of the deep learning model. The average root mean square error (RMSE), calculated by applying CNN Model II, was 0.191, and when the prediction was good, the coefficient of determination ( $R^2$ ) exceeded 0.91. Finally, we performed a sensitivity analysis, which revealed that CDOM is the most influential variable in estimating the spatiotemporal distribution of chlorophyll-*a*.

**Keywords:** deep learning; convolutional neural network; chlorophyll-*a*; satellite; hydrodynamic model



**Citation:** Jin, D.; Lee, E.; Kwon, K.; Kim, T. A Deep Learning Model Using Satellite Ocean Color and Hydrodynamic Model to Estimate Chlorophyll-*a* Concentration. *Remote Sens.* **2021**, *13*, 2003. <https://doi.org/10.3390/rs13102003>

Academic Editor: Giacomo De Carolis

Received: 6 April 2021

Accepted: 17 May 2021

Published: 20 May 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Marine environments experience continuous deterioration owing to the influx of pollutants from rivers and various infrastructure projects including breakwater construction, dredging, and reclamation. To restore marine environments, numerous mitigation plans have been established using various prediction and evaluation techniques. Nevertheless, several limitations still remain: first, the ocean is a complex three-dimensional system that is difficult to model accurately; second, sea water constituents exhibit dynamic movements due to external forces such as wind, tides, currents, density, etc.; third, a significant amount of time and effort is required to observe oceanic trends; and finally, despite significant developments in marine environment prediction technology, several assumptions and additional research area information are still required [1–4].

The water quality model has been widely employed in marine environment prediction, although professional knowledge and experience, various input data, and model validation procedures are required to utilize it. However, owing to the complex and interconnected nature of marine environments, major problems such as eutrophication, harmful algal blooms (HABs), and hypoxia, are difficult to identify and solve. Consequently, considerable research has been conducted on the development of efficient and reliable prediction techniques. Since 2015, deep learning technology that makes predictions using big data has been widely used in various atmospheric, financial, medical, and scientific fields [5–8].

Marine research using deep learning technology can be divided into prediction-related research, classification-related research, and research on methods to correct missing values. Prediction-related research has been applied to various topics, such as the El Niño Index, chlorophyll-*a* time series, and sea surface temperature [9–11]. Classification-related research has been conducted to classify marine life using image data. For example, studies have been conducted to identify the harmful algae that adversely affect marine ecosystems and to classify coral reefs and monitor aquatic ecosystems [12–14]. However, observations using sensors can contain a significant amount of missing data. Consequently, various methods have been developed to estimate the missing data using deep learning techniques [15].

In addition to water quality modeling and deep learning studies, significant research has also been conducted to evaluate the status of plankton and other environmental factors related to marine environments using remote sensing. Ocean color sensors have been used in remote sensing satellites for decades. Those currently in operation include the Chinese Ocean Color and Temperature Scanner (COCTS) onboard HY-1D; Geostationary Ocean Color Imager (GOCI) onboard the Communication, Ocean, and Meteorological Satellite (COMS); Moderate Resolution Imaging Spectroradiometer (MODIS) onboard Aqua; Multi-Spectral Instrument (MSI) onboard Sentinel-2A and Sentinel-2B; Ocean and Land Color Instrument (OLCI) onboard Sentinel-3A and Sentinel-3B; Visible Infrared Imaging Radiometer Suite (VIIRS) onboard Suomi NPP; and Second-Generation Global Imager (SGLI) onboard GCOM-C [16,17]. Ocean color sensors provide vast amounts of spatial data that cannot be obtained from in situ measurements, and consequently, various analyses of spatiotemporal trends are possible. Therefore, extensive research has been conducted to retrieve marine inherent optical properties from ocean color remote sensing and verify ocean color data [18–21]. The data obtained from ocean color sensors are calibrated and verified by comparing them with in situ measurements and the results of existing ocean color sensors [22,23]. Recently, the measurement of ocean color data products such as colored dissolved organic matter (CDOM), chlorophyll-*a*, and total suspended sediment (TSS) has been improved using various neural network methods [24–26].

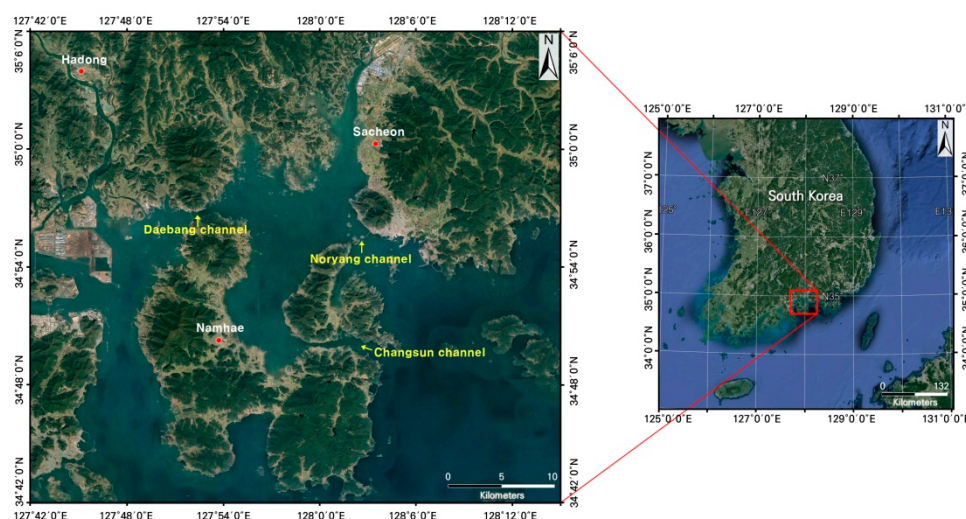
Another significant problem is the occurrence of HABs, which induce hypoxia and kills fish in marine environments. An HAB is caused by complex external environmental processes and factors such as eutrophication, currents, and salinity gradients [27,28]. Monitoring and predicting the spatiotemporal distribution of chlorophyll-*a* are vital to minimize the damage of HABs [29]. A variety of spatial information is required to predict the spatiotemporal distribution of chlorophyll-*a*, owing to the complex interaction of various physical, chemical, and biological factors. Although CDOM, TSS, and chlorophyll-*a* data can be obtained using ocean color sensors, the extraction of physical information such as currents, velocity, and salinity is limited, and in situ measurements can only provide some information. The continued development of hydrodynamic models has significantly improved their prediction ability, providing physical information with a root mean square error (RMSE) of  $\pm 10\%$ ,  $\pm 10\%$  to  $\pm 20\%$ ,  $\pm 0.5\text{ }^{\circ}\text{C}$ , and  $\pm 1\text{ psu}$  for water level, velocity, temperature, and salinity, respectively [30].

In this study, we aim to develop a tool that can estimate the spatial distribution of chlorophyll-*a* using deep learning technology. Satellite ocean color and hydrodynamic model data are used as the training data for the deep learning model. The CDOM, TSS, visibility, and chlorophyll-*a* data recorded on an hourly basis were extracted from a geostationary satellite. The hydrodynamic model data include temperature, salinity, water level, and velocity. The developed tool estimates the spatial distribution of chlorophyll-*a* using the spatial information of CDOM, TSS, visibility, water level, velocity, temperature, and salinity. The accuracy and applicability of the developed prediction tool is demonstrated by comparing the predicted results against the satellite data. As the variables applied to the prediction of chlorophyll-*a* contribute both individually and collectively, the contribution of each variable to the estimation of chlorophyll-*a* is examined as well.

## 2. Material and Methods

### 2.1. Study Area

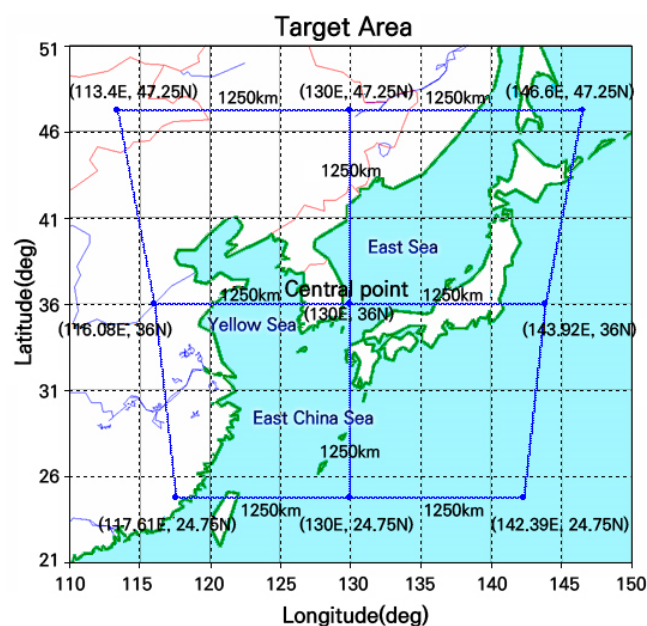
The study area is a semi-closed maritime region surrounded by Hadong-gun, Sacheon, and Namhae-gun in South Korea, and is connected to the sea through the Daebang channel to the east, the Noryang channel to the west, and the Changsun channel to the south, as shown in Figure 1. The study area is approximately 19 km long along the north–south direction, and 13 km long along the east–west direction. The length of the coastline is approximately 136 km and the bounded area is approximately 180 km<sup>2</sup>. The average depth is approximately 3.6 m, the depth of the central area is approximately 10 m, and the deepest area—in the channels—is approximately 30–40 m. In summer, a large volume of river water flows into the study area through the channels due to high rainfall. Consequently, although it is a semi-closed sea area, seawater exchange occurs. Sprayed shellfish farming is actively carried out in the region, gradually increasing from 230 tons in 2000, to 730 tons in 2010, and 2410 tons in 2014 [31]. Consequently, sustainable water quality management is vital in such semi-closed marine environments with active aquaculture.



**Figure 1.** Study area in South Korea (Source: Google Earth).

### 2.2. Satellite Ocean Color

Various satellites with ocean color sensors have been launched from around the world, and Korea launched COMS in 2010 for ocean observation [32,33]. COMS performs meteorological and ocean observations and provides communication services. Ocean color observations are made using the GOCI. The GOCI observes an area of 2500 km × 2500 km, centered on the Korean Peninsula. The resolution of each grid is 500 m, both in width and height, as shown in Figure 2. As COMS is a geostationary satellite, the GOCI records data eight times a day (from 9:00 to 16:00), with images recorded for 30 min every hour. The primary role of the GOCI is to monitor the marine ecosystems around the Korean Peninsula, including long- and short-term marine environmental and climatic changes, coastal and marine environmental monitoring, coastal and marine resource management, and the generation of marine and fishery information [34,35].

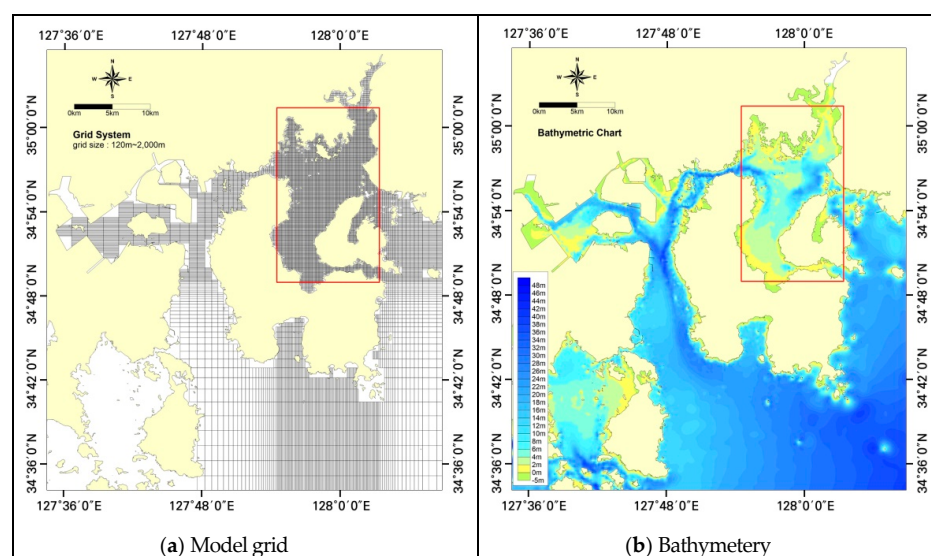


**Figure 2.** Spatial information observed by the GOCI [http://kosc.kiost.ac.kr/p20/kosc\\_p21.html](http://kosc.kiost.ac.kr/p20/kosc_p21.html) (accessed on 26 May 2020).

The GOCI has six visible bands with band centers of 412 nm (B1), 443 nm (B2), 490 nm (B3), 555 nm (B4), 660 nm (B5), and 680 nm (B6), and two near-infrared bands with band centers of 745 nm (B7) and 865 nm (B8). Bands B1–B5 are used to record the water quality parameters. The main applications of each band are B1 for yellow substances and turbidity; B2 for chlorophyll absorption maximum; B3 for chlorophyll and other pigments; B4 for turbidity and suspended sediment; and B5 for baseline of fluorescence signal, chlorophyll, and suspended sediment [36]. The amount of light recorded by the optical sensor onboard the satellite is converted to an electronic value and stored in the satellite image. Radiometric calibration is used to precisely define the relationship between the amount of light and the electronic value, and geometric correction is performed to correct the positional information of each pixel in the image. Subsequently, first-order outputs, such as the top-of-atmosphere radiance, and secondary outputs, such as the remote sensing reflectance, chlorophyll-*a*, TSS, and CDOM concentrations, are verified. Various calibration and validation studies have been performed on the GOCI data to improve its accuracy [35,37–39]. The ocean data products used herein were obtained from the GOCI using a software GDPS including atmospheric correction and ocean environment analysis algorithms. The GDPS enables real-time data processing using a Windows-based GUI. The data products obtained from the GDPS include the water leaving radiance (Lw), normalized water leaving radiance (nLw), chlorophyll-*a*, TSS, and CDOM [40].

### 2.3. Hydrodynamic Model

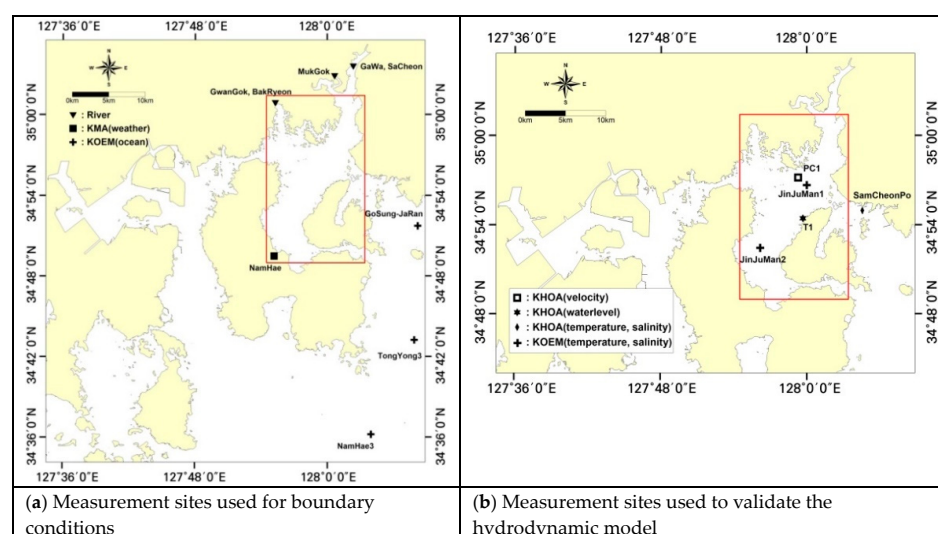
A hydrodynamic model was used to generate marine physical factors, such as the currents, water level, salinity, and temperature, in the study area. The Delft 3D model, which has been applied in several research areas, was used to simulate three-dimensional hydrodynamics [41–44]. The model domain extended for 58 km along the north–south direction and 53 km along the east–west direction, to sufficiently cover the study area. The model grid contained  $155 \times 245$  horizontal cells and, to optimize the computational time, fine and coarse grids were formed in the study area and open sea area, respectively. A total of five vertical layers were modeled to replicate the interaction between the vertical layers and the vertical distribution of salinity and water temperature. Bathymetry for the study area was obtained from the latest navigational charts and the survey data of the Korea Hydrographic and Oceanographic Agency (KHOA). As shown in the bathymetry chart in Figure 3, the bay has a relatively shallow depth and the channels are relatively deep.



**Figure 3.** (a) Model grid in the study area; (b) Bathymetry in the study area.

The boundary conditions of the study area must be defined to execute the hydrodynamic model. The water levels, salinity, and temperatures observed at different measurement sites (GoSung-JaRan, TongYong3, NamHae3) by the Korea Marine Environment Management Corporation (KOEM) were set as the sea boundary conditions, and the monthly average flow rates at GwanGok, BakRyeon, MukGok, GaWa, and SaCheon were set as the river boundary conditions. Meteorological data, such as the wind direction, wind speed, air temperature, and relative humidity, measured at the NamHae site of the Korea Meteorological Administration (KMA), were also used as model input data. The initial conditions of the water level and velocity were set to zero, and the initial conditions of temperature and salinity were derived from the measured data at the five KOEM stations shown in Figure 4. The hydrodynamic model was simulated for a total of five years from 1 January 2015 to 31 December 2019. As the data used in the deep learning model include the water level, current, salinity and temperature, these data were verified. The water level was verified using the data observed at the T1 site operated by KHOA, which is located inside the bay. The current was validated against the data recorded at the PC1 site operated by KHOA, between 24 July 2015 and 26 August 2015. The salinity and water temperature were validated against the data measured at the JinJuMan 1 and JinJuMan 2 sites, operated by KOEM, and the SamCheonPo site, operated by KHOA, as shown in Figure 4.

The water levels in the study area fluctuated by approximately 3 m and were primarily affected by the tides. The average difference in the water level between the hydrodynamic model and the observed values was approximately 10 cm, and the absolute error was within 8–10%, with slight differences every year. The currents observed between 24 July 2015 and 26 August 2015 were classified into a U-component—moving east–west—and a V-component—moving north–south. As shown, the U-component was the dominant current in the study area. The U-component current flowed as fast as 0.5 m/s and fluctuated based on the tidal cycle. Although the hydrodynamic model results appear to underestimate the current patterns, the results are reproduced well. The temperature was below 10 °C during winter and almost 30 °C during summer, with clearly noticeable seasonal variations. The water temperature varied between 13 °C and 20 °C during spring and autumn, with the lowest temperature in February and the highest temperature in August. Considering the predicted daily temperatures, the hydrodynamic model adequately reproduced the annual temperature-change pattern, and the average RMSE of the temperature was 0.862 °C. The salinity was highly influenced by the river flow, i.e., during spells of high rainfall, the salinity temporarily decreased before increasing to approximately 32–33 psu. The average RMSE of the salinity was 0.6 psu, as shown in Figure 5.



**Figure 4.** Locations of the KMA, KOEM, KHOA, and river monitoring stations in the study area. (a) Measurement sites used for boundary conditions. (b) Measurement sites used to validate the hydrodynamic model.

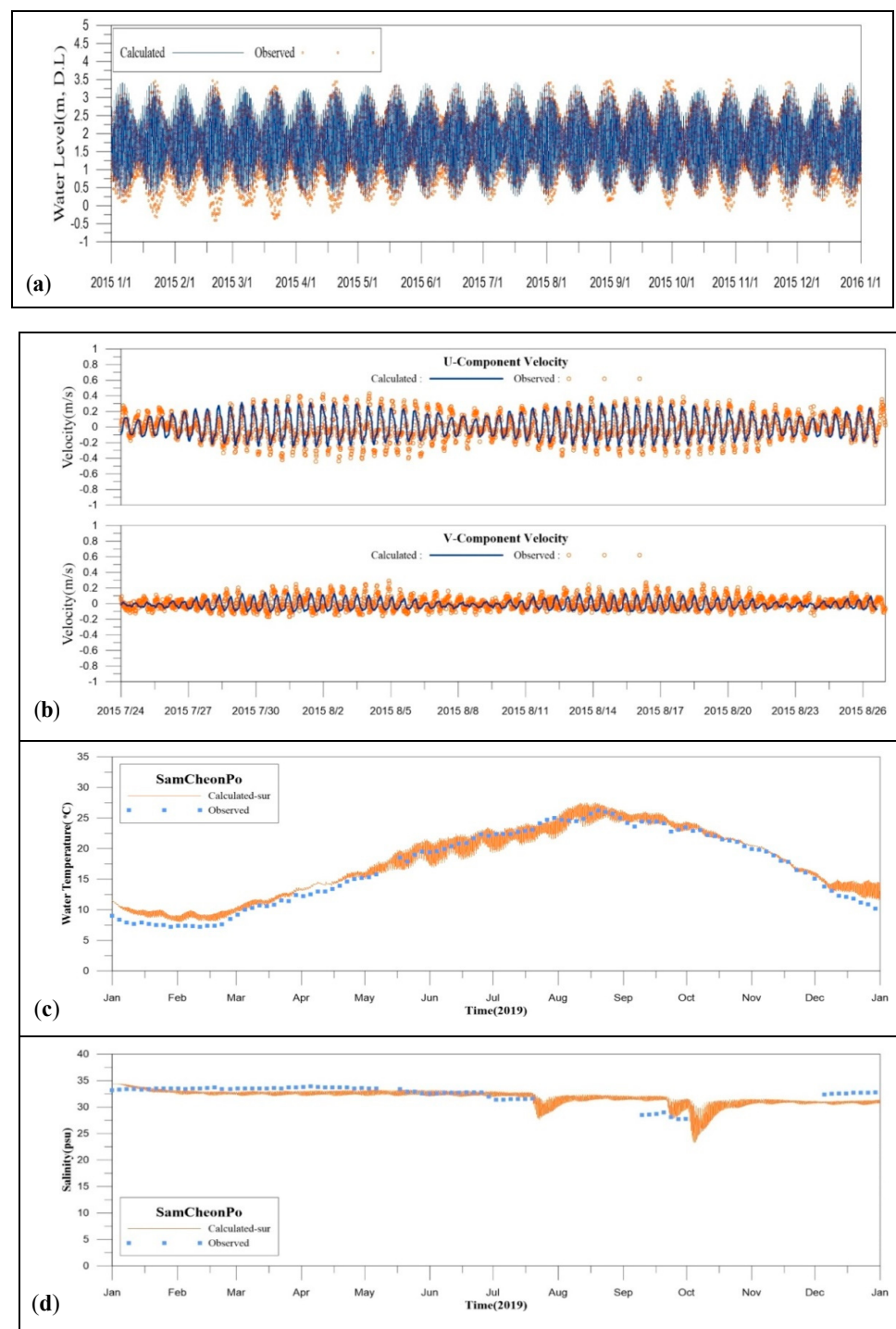
#### 2.4. Data Structure for Deep Learning Model

The satellite data of the study area, which was required to construct the deep learning model, was provided by the Korea Ocean Satellite Center (KOSC) in the Korea Institute of Ocean Science and Technology. The data were recorded eight times per day between 9:00 and 16:00, from January 2015 to December 2019. The data obtained included the entire Korean Peninsula, and the total size of the data was approximately 14 TB. No satellite data could be extracted when the study area was covered by clouds. The total number of extracted data was 391 in 2015, 276 in 2016, 266 in 2017, 271 in 2018, and 128 in 2019. Generally, a large amount of data were recorded during winter, when the weather was good, and a small amount of data were recorded during summer, owing to the increased rainfall and typhoons.

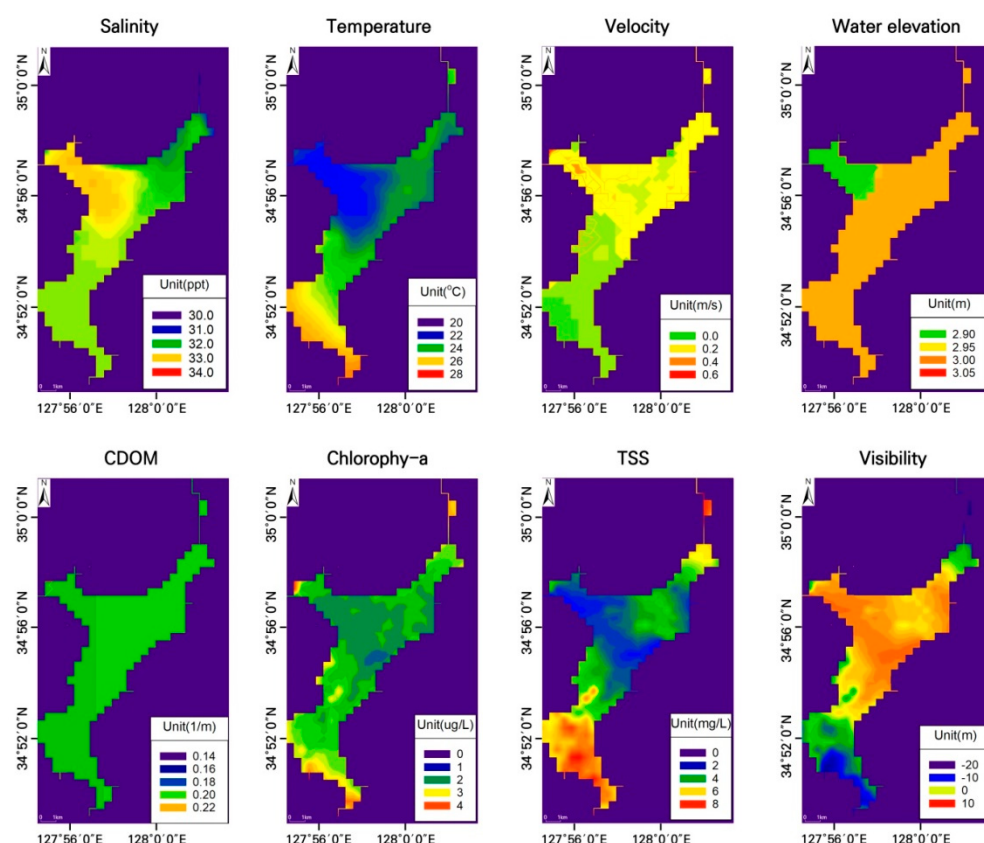
The hydrodynamic model results were extracted for the same area as the satellite measurements, as shown in Figure 6. The hourly salinity, temperature, currents, and water levels between 2015 and 2019 were converted into a grid format. As the resolution of the satellite data was 500 m, the data from the area adjacent to the coastline could not be obtained. Therefore, only the data pertaining to the sea area 500 m away from the coastline were used to train the deep learning model. Accordingly, the hydrodynamic model results of the area adjacent to the coastline were also neglected.

#### 2.5. Deep Learning Model Structure

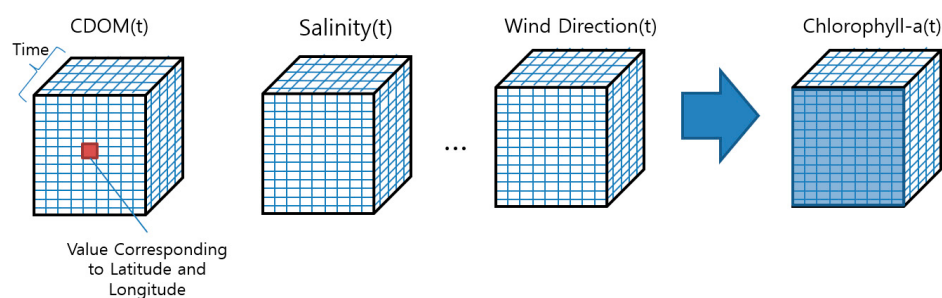
As the satellite and hydrodynamic model data were in the form of a  $48 \times 27$  grid, they could be treated as image data. Consequently, an image-based deep learning method was applied herein. Each  $48 \times 27$  grid was referred to as an ‘image,’ and each point in the image was referred to as the ‘data’ or ‘point’. The satellite chlorophyll-*a* data were treated as ground-truth data, as several studies have shown a high correlation between the ground-truth chlorophyll-*a* data and satellite chlorophyll-*a* data. Accordingly, we constructed a deep learning model to estimate the temporal and spatial distribution of chlorophyll-*a* using both the satellite and the hydrodynamic model data. Specifically, the deep learning model estimated the temporal and spatial distribution of chlorophyll-*a* at a given time (*t*) by integrating the satellite data, such as the CDOM, TSS, and visibility, and the hydrodynamic model data, such as the currents, water level, temperature, and salinity, at the same time (*t*), as illustrated in Figure 7.



**Figure 5.** (a) Temporal variations of water level; (b) Temporal variation of currents; (c) Temporal variation of salinity; (d) Temporal variation of temperature (points are observations and lines are model results).



**Figure 6.** Spatial distribution of training data in the study area: salinity, temperature, currents, and water levels from the hydrodynamic model, and CDOM, chlorophyll-*a*, TSS, and visibility from the satellite ocean color data.

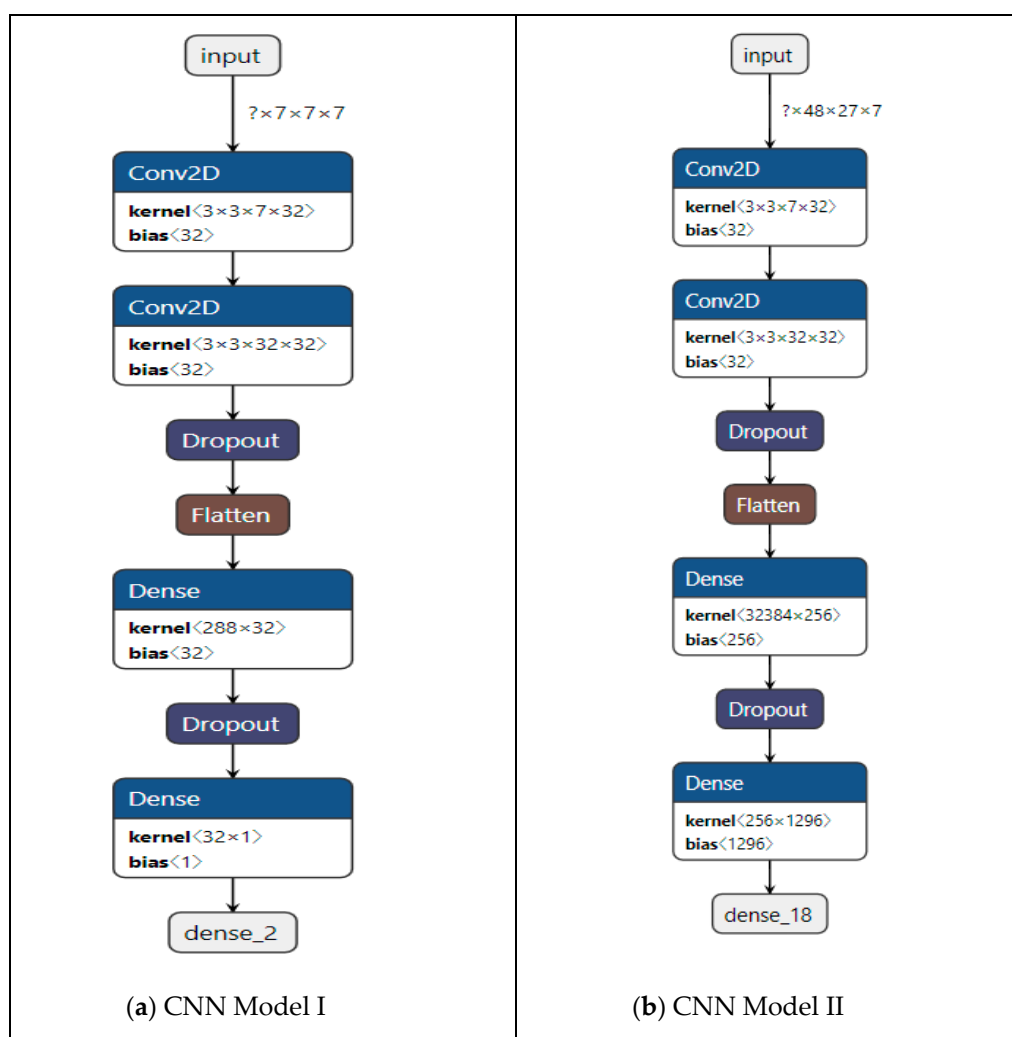


**Figure 7.** Construction of the deep learning model for estimating the temporal and spatial distribution of chlorophyll-*a*. To utilize spatial information, the input data were organized in a matrix accumulated over time. The value corresponding to each row and column corresponds to the latitude and longitude of each data.

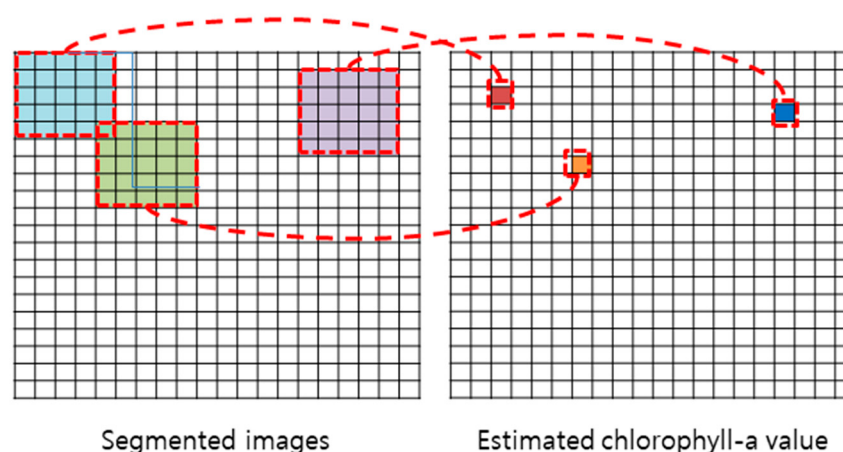
A convolutional neural network (CNN) is a well-known deep learning model that is suitable for image data processing. A CNN model consists of multiple convolutional layers that extract features from an image and pool the layers through subsampling, leaving only the important patterns behind. Classification and estimation are performed through iterative convolutional and pooling operations. We designed two approaches to estimate chlorophyll-*a* based on a CNN. The first CNN model, called 'CNN Model I', estimates the chlorophyll-*a* concentration from an image in a  $48 \times 27$  grid format by integrating a total of seven images—three images from the satellite data, such as the CDOM, TSS, and visibility, and four images from the hydrodynamic model data, such as the currents, water level, temperature, and salinity—as shown in Figure 8. Notably, as the image size

was small, there was no pooling layer. Consequently, the pooling layer for information compression was ineffective. The second CNN model, called 'CNN Model II', predicted the chlorophyll-*a* concentration using segmented images.

Additional preprocessing is required to use segmented images as the model input. For example, in the case of  $7 \times 7$  segmented images, the chlorophyll-*a* value is estimated by using segmented images of seven individual input variables. The difference between CNN Model I and CNN Model II is that the former estimates one chlorophyll-*a* image by integrating the images of seven individual input variable changes, whereas the latter estimates the chlorophyll-*a* value by integrating segmented images of seven individual input variables, as shown in Figure 9. As CNN Model II estimates the chlorophyll-*a* value using the data around a point of interest, we believe that it also reflects the local characteristics well.



**Figure 8.** (a) Algorithm of CNN Model I and (b) CNN Model II. CNN Model I uses seven images of  $48 \times 27$  grid size and estimates the chlorophyll-*a* value in a  $48 \times 27$  grid format. CNN Model II uses segmented images in a  $7 \times 7$  grid format and estimates the chlorophyll-*a* value.



**Figure 9.** Schematic diagram of the application of segmented images in the CNN Model II; segmented images are generated by iteratively moving the window cell-by-cell. The CNN Model II estimates a chlorophyll-*a* value integrating segmented images of seven individual input variables.

To verify the reliability of the deep learning model, the data were divided into training data, validation data, and test data, considering the seasonal characteristics over an entire year. For CNN Model I, 932 images were used for training, 271 images for validation, and 128 images for testing. For CNN Model II, the images in a  $48 \times 27$  grid format were divided into segmented images with a  $7 \times 7$  grid format. Consequently, the number of images used for training, validation, and testing increased to 293,580, 85,365, and 40,320, respectively. As CNN Model II did not have the segmented images required to estimate the values of three columns and three rows at the edge of each image, the values related to these regions were not predicted. The quantity of available data varied from one year to another as the satellite measurements could not be obtained on days with poor weather. In particular, the quantity of data obtained during summer was relatively small compared to that obtained during the other seasons owing to increased rainfall and typhoons, as shown in Table 1.

**Table 1.** Information of training data, validation data, and test data in the CNN Model I and CNN Model II.

| Category                                                 | Training Data | Validation Data | Test Data |
|----------------------------------------------------------|---------------|-----------------|-----------|
| Period (year)                                            | 2015–2017     | 2018            | 2019      |
| CNN Model I (# of images)                                | 932           | 271             | 128       |
| CNN Model II<br>(# of segmented images ( $7 \times 7$ )) | 293,580       | 85,365          | 40,320    |

### 3. Results

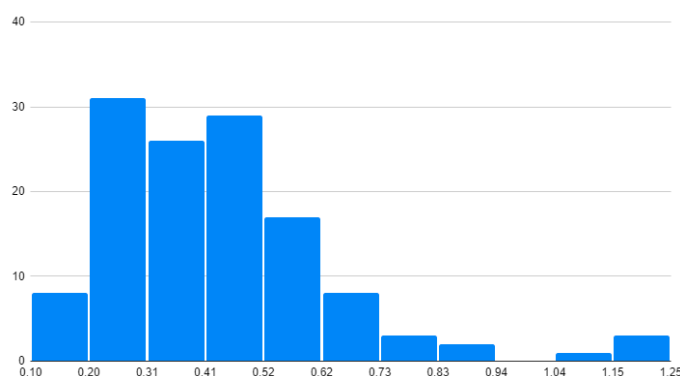
#### 3.1. CNN Model I

The RMSE, which is the difference between the predicted chlorophyll-*a* and the satellite chlorophyll-*a* values, was used to evaluate the accuracy of the CNN models designed herein. The RMSE was calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{pred}(i) - \text{target}(i))^2} \quad (1)$$

where  $\text{pred}(i)$  represents the predicted chlorophyll-*a* pixel value for of the  $i$ th point and  $\text{target}(i)$  represents the satellite chlorophyll-*a* pixel value for the  $i$ th point in each image.

CNN Model I was used to estimate the chlorophyll-*a* value of 128 images recorded in 2019. In most cases, the RMSE was approximately 0.2–0.6 and the average RMSE was 0.436, as shown in Figure 10. The minimum RMSE was 0.106 and the maximum RMSE was 1.242, which is a significant gap. Therefore, specific analyses were performed for the cases with RMSE = 0.106, RMSE = 0.506, and RMSE = 1.209, as shown in Figure 11.



**Figure 10.** RMSE distribution for 128 images using CNN Model I: histogram with the range of RMSE values on the X-axis and the number of images on the Y-axis.

In the case with the lowest RMSE (RMSE = 0.106), the model results showed that there was a slight predictive error in the image, but the overall trend was well estimated. In the case with the RMSE close to the average value (RMSE = 0.506), the overall change in chlorophyll-*a* in the entire image was clearly estimated, but the accuracy of the estimation of the local changes in chlorophyll-*a* was limited. In the case with the high RMSE (RMSE = 1.209), the model was unable to estimate the satellite chlorophyll-*a* value. The measured values clearly indicate a change in the spatial chlorophyll-*a* values, whereas the estimated values tend to converge to the average value at most points. Thus, the model appeared to have a tendency to approximate the average value as the estimated value when the training data were insufficient, as shown in Figure 11. Consequently, the coefficient of determination ( $R^2$ ), which represents how well the model results fit the satellite data, was applied herein.  $R^2$  is represented by a value of 0.0–1.0, where a value of 1.0 indicates a perfect fit. When the RMSE was relatively low, the  $R^2$  was around 0.673, and when the RMSE was high,  $R^2 < 0.5$ . When  $R^2 < 0.5$ , the higher the chlorophyll-*a* value of the satellite data, the lower the predictive ability, as shown in Figure 12.

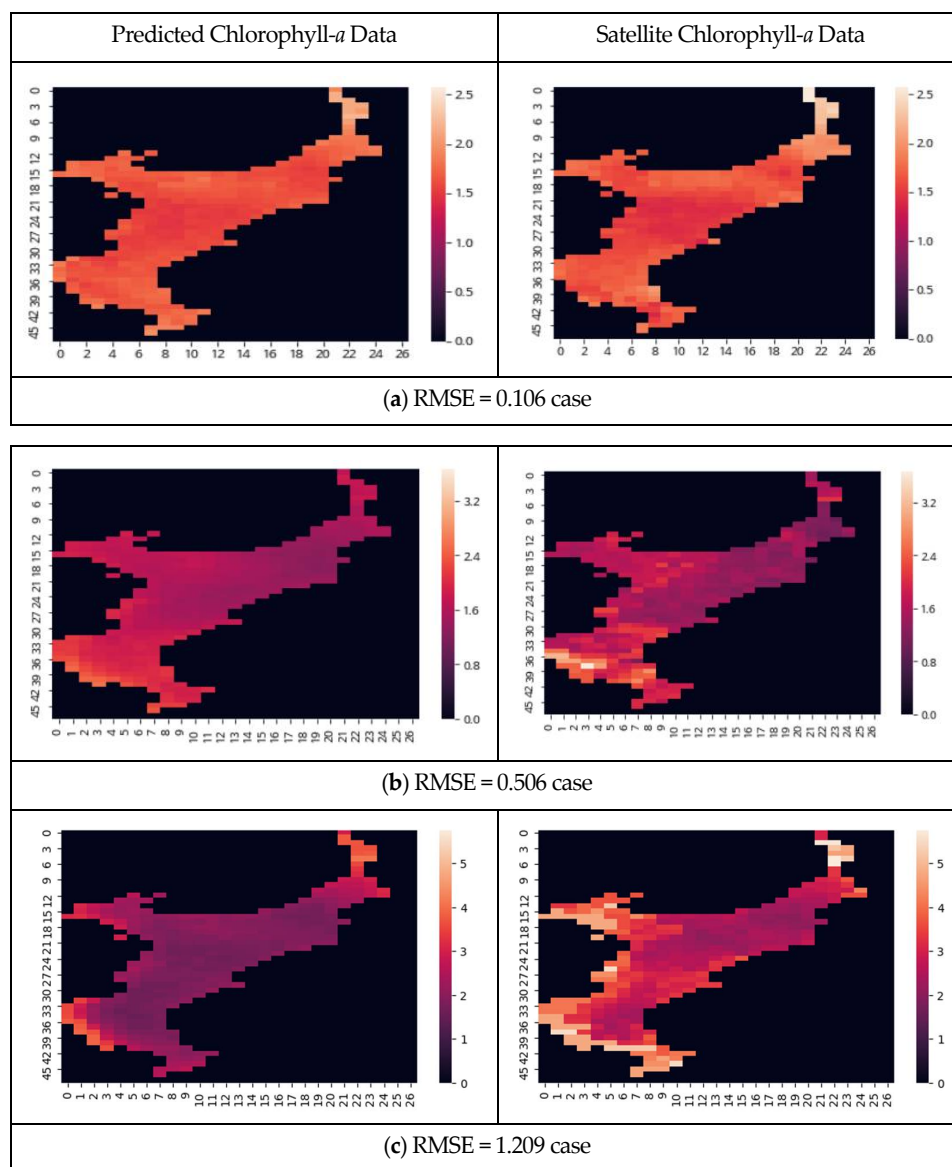
The results of CNN Model I tended to be averaged by assimilating the surrounding values instead of estimating local changes. As deep learning models such as a CNN estimate values by analyzing patterns from training data, the prediction patterns could not be determined from insufficient training data. Therefore, CNN Model I, which was trained using only 1203 training and validation images, could predict the overall trends but failed to predict local changes. Notably, if additional training data is provided, the prediction accuracy of CNN Model I can be improved.

### 3.2. CNN Model II

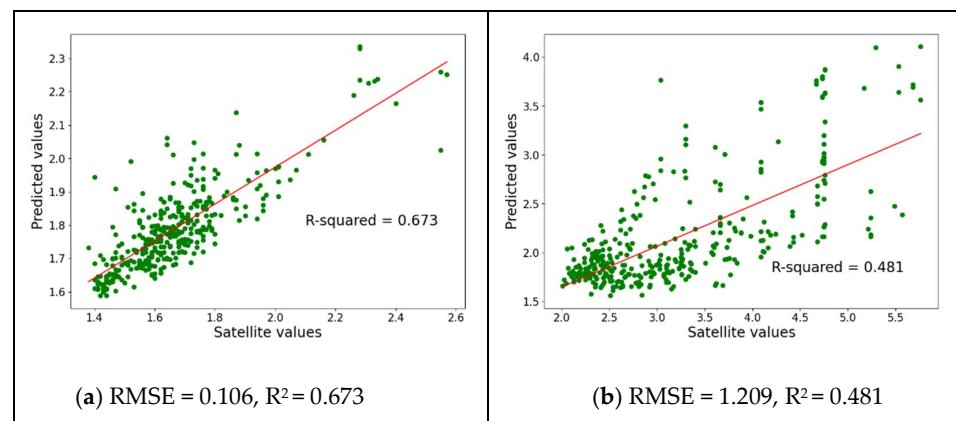
Chlorophyll-*a* estimation was also performed using CNN Model II, which utilized 300 times more training and validation data than CNN Model I, owing to the use of segmented images. The RMSE values of CNN Model II were around 0.05–0.8. Most of the RMSE values were less than or equal to 0.2, with an average of 0.167. Compared to the results of CNN Model I, the RMSE values of CNN Model II were significantly lower, confirming the excellent predictive ability of the latter. Notably, RMSE was less than or equal to 0.12 in almost half the total number of predictions. A detailed analysis was performed by classifying the RMSE values of CNN Model II into good, average, and bad cases, as shown in Figure 13.

In the case of a low RMSE value (RMSE = 0.055), the predicted chlorophyll-*a* values were almost the same as those of the satellite chlorophyll-*a* values. Furthermore, the spatial variations of chlorophyll-*a* concentration were properly estimated. The case with an RMSE value close to the average value (RMSE = 0.204) also demonstrated similar results to the observed values. In particular, the changes in the spatial concentration were estimated accurately. In the case of a high RMSE value (RMSE = 0.775), the model accurately reproduced the spatial concentration pattern but tended to underestimate the concentration at some

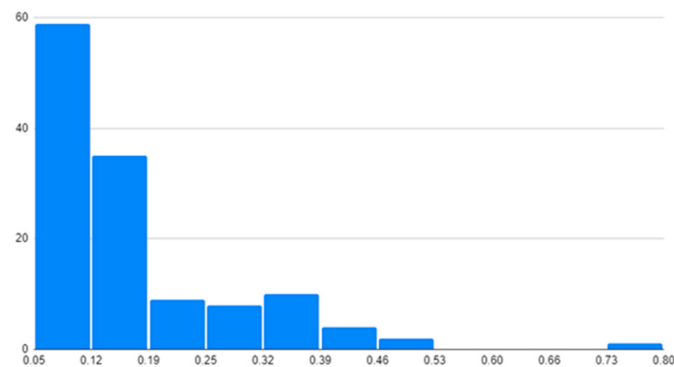
points. The satellite data exhibited large variations in the concentration between adjacent points, whereas the deep learning model corrected this drastic change and estimated it smoothly in space, as shown in Figure 14.



**Figure 11.** Chlorophyll-*a* results estimated using the CNN Model I: The left section shows the predicted chlorophyll-*a* values and the right section shows the satellite chlorophyll-*a* values corresponding to the left section. The RMSE values for the three cases are (a) 0.106, (b) 0.506, and (c) 1.209, respectively.

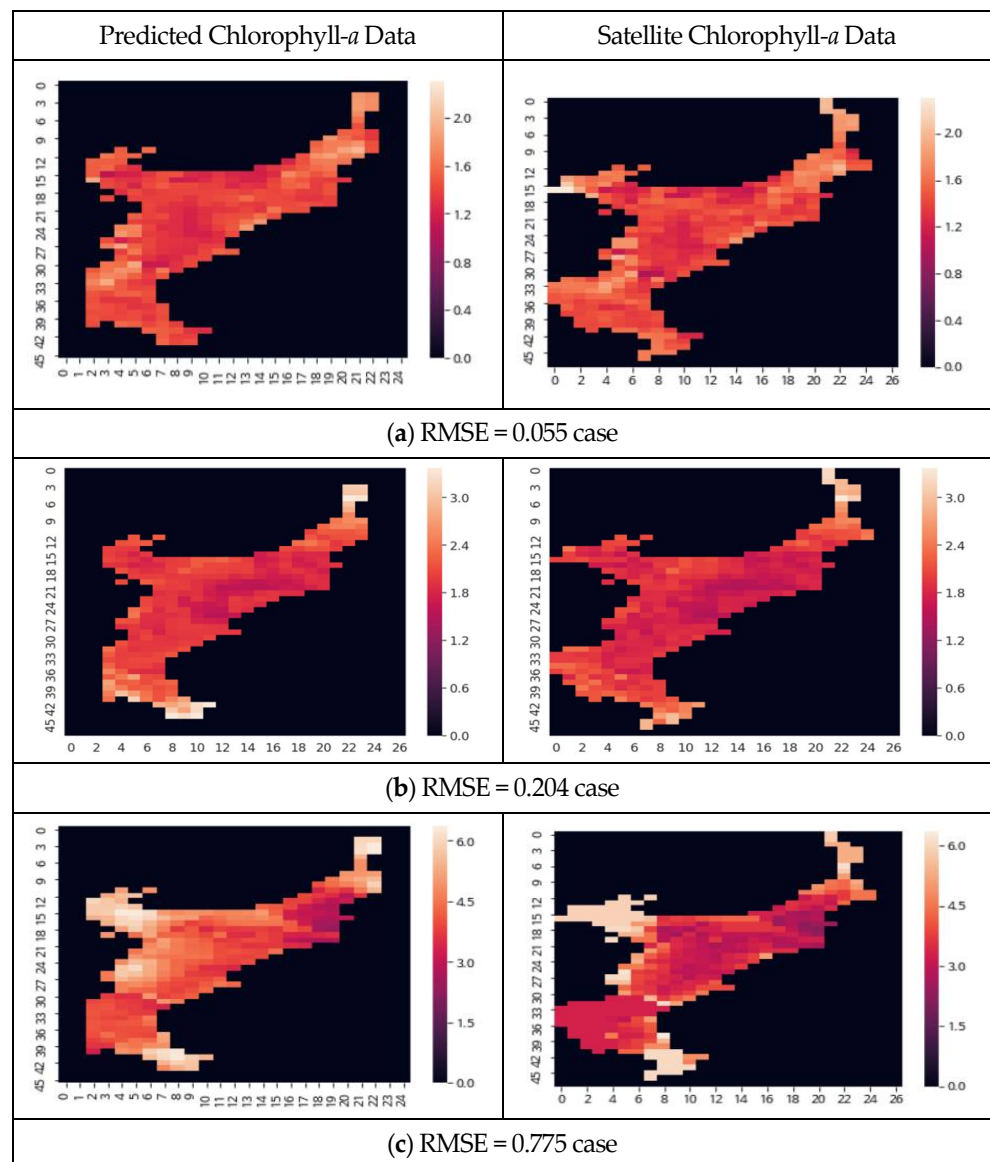


**Figure 12.** Examples of (a) good  $R^2$  and (b) bad  $R^2$  values among the results of the CNN Model I.

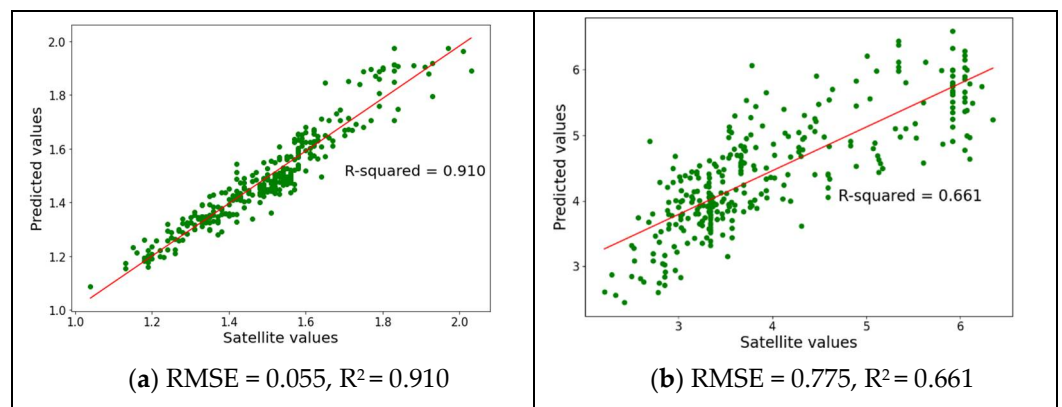


**Figure 13.** RMSE distribution for 128 images using the CNN Model II: histogram with the range of RMSE values on the X-axis and the number of images on the Y-axis.

Compared to CNN Model I, CNN Model II has significantly better chlorophyll-*a* estimation ability, and the spatial change pattern of chlorophyll-*a* was successfully estimated in all the model results. Furthermore, the coefficient of determination ( $R^2$ ) improved significantly. When  $RMSE = 0.055$ ,  $R^2 = 0.91$ , and when  $RMSE = 0.775$ , which suggests a high degree of error, the overall trend was reproduced well and  $R^2 = 0.661$ , as shown in Figure 15. Although both models used the same CNN technique, the difference in their estimation abilities is likely due to the large difference in their respective training data volumes.



**Figure 14.** Chlorophyll-*a* results estimated using CNN Model II. The left section shows the predicted chlorophyll-*a* values and the right section shows the corresponding satellite chlorophyll-*a* image values. The corresponding RMSE values are (a) 0.055, (b) 0.204, and (c) 0.775, respectively.



**Figure 15.** Examples of (a) good  $R^2$  and (b) bad  $R^2$  values among the results of the CNN Model II.

#### 4. Discussion

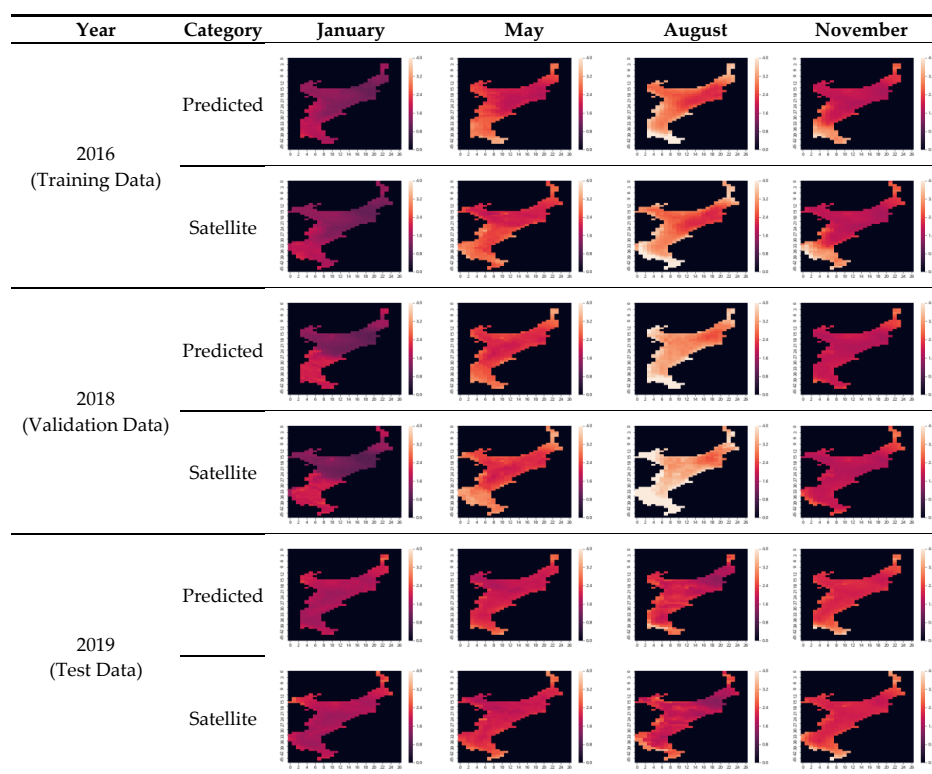
Plankton growth is affected by various factors such as water flow, water temperature, nutrients, and light. The concentration of plankton is relatively high in shallow water coastal areas and upwelling regions, as they have a rich supply of nutrients. The surface salinity and temperature of the study area change significantly as high salinity and low temperature seawater flows through the Daebang channel, located in the northeast. The satellite data reveals that the seawater flowing in from the Daebang channel contains low concentrations of chlorophyll-*a*, resulting in a relatively low chlorophyll-*a* concentration in the center of the study area. Moreover, the study area is connected to a river, and large amounts of river water flow into the study area during the rainy summer season, affecting the growth of plankton. As the growth of each type of plankton depends on the water temperature, it is important to predict the seasonal changes in plankton concentration.

The monthly averaged satellite data and model data were compared to determine whether the prediction model developed herein can adequately estimate the seasonal changes in plankton concentration. In 2016 and 2018, the chlorophyll-*a* concentration was low in January—the winter season—but high during spring and summer. The concentration decreased again in November, which clearly demonstrates the seasonal fluctuations in plankton concentration in the study area. The developed model successfully estimated the seasonal fluctuations in plankton concentration in 2016 and 2018. Notably, although the seasonal fluctuations in 2019 were relatively small compared to those in 2016 and 2018, the developed model accurately estimated the small seasonal and local concentration changes, as shown in Figure 16.

We performed a sensitivity analysis to determine the influence of each input variable in the model results. To do so, the performance of the model was investigated by only using individual input variables as training data for the deep learning model. The results of the sensitivity analysis (Table 2) indicated that CDOM contributes significantly to the estimation of chlorophyll-*a*, with an RMSE of 0.231. The visibility, TSS, and temperature are also relatively important variables, whereas the remaining input variables have a relatively low contribution to the improvements in model performance. Notably, when all the input variables, except for CDOM, were integrated, the RMSE increased to 0.330. Thus, although the individual input variables have a negligible effect on the model performance, the integration of the input variables has a complementary effect and improves model prediction. When all the input variables were used, the RMSE was 0.191, which represents the best model performance.

Predictive studies on plankton concentrations have been conducted for decades using various water quality models. However, there are numerous challenges and limitations owing to the complex interactions between water quality parameters, uncertainty of hydrodynamic information, and lack of boundary nutrient loadings and validation data. For example, the results of studies that predicted the level of chlorophyll-*a* in Chesapeake Bay by employing a 3D water quality model had a correlation coefficient of less than 0.5 [45,46]. The main objective of this study was to develop a prediction tool that can be used in combination with existing water quality models, wherein the currents, water level, salinity, and temperature calculated from the hydrodynamic model were used to predict chlorophyll-*a* concentration. As the hydrodynamic model results have an error of only 10–20%, they can be used as training data for deep learning models [30]. Accordingly, satellite data such as CDOM, TSS, and visibility, which were validated through various studies, were used as training data to develop a chlorophyll-*a* prediction tool. The prediction model developed herein—CNN Model II—has good accuracy in the estimation of chlorophyll-*a* concentration, as evidenced by an  $R^2$  of 0.66–0.91 and an RMSE of 0.055–0.775. Although the data used in the model are not in situ measurements, satellite data and hydrodynamic model data have continuously improved in recent years, and provide spatiotemporal data that cannot be obtained from in situ measurements. In addition, the developed model can predict the spatiotemporal chlorophyll-*a* concentration based on changes in individual parameters such as an increase in water temperature due to climate change, an increase in

CDOM due to land development, and an increase in TSS as a result of poor flushing due to the presence of coastal structures, etc.



**Figure 16.** Monthly averaged spatial distribution of model results and satellite chlorophyll-*a* images (CNN Model II).

**Table 2.** Sensitivity analysis results showing RMSE values corresponding to input variables.

| Input Variables | RMSE  |
|-----------------|-------|
| CDOM            | 0.231 |
| TSS             | 0.526 |
| Visibility      | 0.492 |
| Currents        | 0.651 |
| Salinity        | 0.648 |
| Temperature     | 0.545 |
| Water level     | 0.653 |
| All except CDOM | 0.330 |
| All             | 0.191 |

The model results must be compared to real-world measurement data to validate the performance of the model. However, spatiotemporal chlorophyll-*a* data cannot be obtained through in situ measurements. The performance of the chlorophyll algorithms used for the GOCI radiometric data were evaluated using in situ measurements collected at 491 stations [47]. The evaluation results of the coincident in situ pairs of Rrs and chlorophyll measurements demonstrated that the mean uncertainty was <35%, with a correlation of around 0.8. Therefore, assuming that the data from GOCI are close to the real-world values, the model results were validated by comparing them against the satellite data. To improve the developed model, it is necessary to conduct a validation study with the measurement data of the study area and a comparative study with the state-of-the-art methods.

## 5. Conclusions

In this study, we developed a deep learning model using a CNN to predict the spatiotemporal changes in chlorophyll-*a* in a bay in Korea. The data used to train the deep learning model were the spatial data of chlorophyll-*a*, total suspended sediment (TSS), visibility, and colored dissolved organic matter (CDOM) obtained from the Geostationary Ocean Color Imager (GOCI) on board COMS, and the water level, currents, temperature, and salinity calculated by a verified hydrodynamic model. CNN Model I, which estimates chlorophyll-*a* images in a  $48 \times 27$  grid format, was developed using the same  $48 \times 27$  grid size of the CDOM, TSS, visibility, water level, currents, temperature, and salinity data. The RMSE between the satellite image and the predicted image from the model was calculated, and was between 0.2 and 0.6 in most cases. Although CNN Model I was able to estimate the overall trend, there were significant differences between the predicted results and the satellite data in some cases. As the deep learning model improves the predictive ability of the model by extracting and analyzing the inherent patterns in the training data, if the training data is insufficient, the predictive ability of the model decreases significantly.

To solve the problem of insufficient data, we designed another deep learning model—CNN Model II—using segmented images in a  $7 \times 7$  grid format. CNN Model II estimates target values only using the data around the point of interest and, consequently, the volume of training data used in CNN Model II is around 300 times more than that of CNN Model I. Therefore, CNN Model II can extract and analyze inherent patterns in the training data more accurately. The average RMSE of CNN Model II was 0.191, which is significantly lower than that of CNN Model I, which was 0.463. Moreover, the spatial concentration of chlorophyll-*a* was well estimated by CNN Model II, thereby proving the efficacy of the deep learning model.

A sensitivity analysis was performed to determine the influence of each input variable on the model performance, and CDOM was found to have the most influence on the prediction of chlorophyll-*a*. The visibility, TSS, and temperature were also relatively important variables. The input variables with a strong influence on the model performance have a direct relationship with nutrients, photosynthesis, and temperature, which influence plankton growth. Therefore, the data-based deep learning model considers the major factors related to the growth of plankton and makes predictions. Additionally, the predictive accuracy of the deep learning model was improved if the training data also included the currents, velocity, and salinity.

**Author Contributions:** Conceptualization, D.J. and T.K.; methodology, E.L. and T.K.; software, D.J., T.K. and K.K.; validation, D.J. and K.K.; data curation, E.L.; writing—original draft preparation, D.J. and T.K.; writing—review and editing, D.J. and T.K.; visualization, D.J. and K.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing is not applicable to this article.

**Acknowledgments:** This paper was written following the research work “A Study on Marine Pollution Using Deep Learning and its Application to Environmental Impact Assessment (II)” (RE2021-08), funded by the Korea Environment Institute (KEI).

**Conflicts of Interest:** The authors declare that they have no conflict of interest.

## References

1. Beck, M.B. Water quality modeling: A review of the analysis of uncertainty. *Water Resour. Res.* **1987**, *23*, 1393–1442. [[CrossRef](#)]
2. Zheng, L.; Chen, C.; Zhang, F.Y. Development of water quality model in the Satilla River Estuary, Georgia. *Ecol. Model.* **2004**, *178*, 457–482. [[CrossRef](#)]
3. Jia, H.; Xu, T.; Liang, S.; Zhao, P.; Xu, C. Bayesian framework of parameter sensitivity, uncertainty, and identifiability analysis in complex water quality models. *Environ. Model. Softw.* **2018**, *104*, 13–26. [[CrossRef](#)]

4. Yan, J.; Xu, Z.; Yu, Y.; Xu, H.; Gao, K. Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing. *Appl. Sci.* **2019**, *9*, 1863. [\[CrossRef\]](#)
5. Vargas, M.R.; de Lima, B.S.L.P.; Evsukoff, A.G. Deep learning for stock market prediction from financial news articles. In Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Annecy, France, 26–28 June 2017. [\[CrossRef\]](#)
6. Razzak, M.I.; Naz, S.; Zaib, A. Deep learning for medical image processing: Overview, challenges and the future. In *Classification in BioApps*; Dey, N., Ashour, A., Borra, S., Eds.; Springer: Cham, Switzerland, 2018. [\[CrossRef\]](#)
7. Matsuoka, D.; Watanabe, S.; Sato, K.; Kawazoe, S.; Yu, W.; Easterbrook, S. Application of deep learning to estimate atmospheric gravity wave parameters in reanalysis data sets. *Geophys. Res. Lett.* **2020**, *47*, e2020GL089436. [\[CrossRef\]](#)
8. Singh, R.; Agarwal, A.; Anthony, B.W. Mapping the design space of photonic topological states via deep learning. *Opt. Express* **2020**, *28*, 27893. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Xiao, C.; Chen, N.; Hu, C.; Wang, K.; Xu, Z.; Cai, Y.; Xu, L.; Chen, Z.; Gong, J. A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. *Environ. Model. Softw.* **2019**, *120*, 104502. [\[CrossRef\]](#)
10. Guo, Y.; Cao, X.; Liu, B.; Peng, K. El Niño index prediction using deep learning with ensemble empirical mode decomposition. *Symmetry* **2020**, *12*, 893. [\[CrossRef\]](#)
11. Shin, Y.; Kim, T.; Hong, S.; Lee, S.; Lee, E.; Hong, S.; Lee, C.; Kim, T.; Park, M.; Park, J.; et al. Prediction of chlorophyll-*a* concentrations in the Nakdong River using machine learning methods. *Water* **2020**, *12*, 1822. [\[CrossRef\]](#)
12. Park, S.; Kim, J. Red tide algae image classification using deep learning based open source. *Smart Media*. **2018**, *7*, 34–39.
13. Lumini, A.; Nanni, L.; Maguolo, G. Deep learning for plankton and coral classification. *Appl. Comput. Inform.* **2019**. [\[CrossRef\]](#)
14. Raphael, A.; Dubinsky, Z.; Iluz, D.; Benichou, J.I.C.; Netanyahu, N.S. Deep neural network recognition of shallow water corals in the Gulf of Eilat (Aqaba). *Sci. Rep.* **2020**. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Velasco-Gallego, C.; Lazakis, I. Real-time data-driven missing data imputation for short-time sensor data of marine systems: A comparative study. *Ocean Eng.* **2020**, *218*, 108261. [\[CrossRef\]](#)
16. ICCG. Current Ocean-Colour Sensors. Available online: <https://ioccg.org/resources/missions-instruments/current-ocean-colour-sensors/> (accessed on 16 April 2021).
17. McKinna, L.I.W. Three decades of ocean-color remote-sensing *Trichodesmium* spp. In The World's oceans: A review. *Prog. Oceanogr.* **2015**, *131*, 177–199. [\[CrossRef\]](#)
18. Hu, S.B.; Cao, W.X.; Wang, G.F.; Xu, Z.T.; Lin, J.F.; Zhao, W.J.; Yang, Y.Z.; Zhou, W.; Sun, Z.H.; Yao, L.J. Comparison of MERIS, MODIS, SeaWiFS-derived particulate organic carbon, and in situ measurements in the South China Sea. *Int. J. Remote Sens.* **2016**, *37*, 1585–1600. [\[CrossRef\]](#)
19. Werdell, P.J.; McKinna, L.I.W.; Boss, E.; Ackleson, S.G.; Craig, S.E.; Gregg, W.W.; Lee, Z.; Maritorena, S.; Roesler, C.S.; Rousseaus, C.S.; et al. An overview of approaches and challenges for retrieving marine inherent optical properties from ocean color remote sensing. *Prog. Oceanogr.* **2018**, *160*, 186–212. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Scott, J.P.; Werdell, P.J. Comparing level-2 and level-3 satellite ocean color retrieval validation methodologies. *Opt. Express* **2019**, *27*, 30140–30157. [\[CrossRef\]](#)
21. Niroumand-Jadidi, M.; Bovolo, F.; Bruzzone, L. Novel spectra-derived features for empirical retrieval of water quality parameters: Demonstrations for OLI, MSI, and OLCI sensors. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10285–10300. [\[CrossRef\]](#)
22. Crout, R.L.; Ladner, S.; Lawson, A.; Martinolich, P.; Bowers, J. Calibration and validation of multiple ocean color sensors. In Proceedings of the OCEANS 2018 MTS/IEEE Conference, Charleston, SC, USA, 22–25 October 2018; pp. 1–6. [\[CrossRef\]](#)
23. Wang, M.; Ahn, J.H.; Jiang, L.; Shi, W.; Son, S.H.; Park, Y.J.; Ryu, J.H. Ocean color products from the Korean Geostationary Ocean Color Imager (GOCI). *Opt. Express* **2013**, *21*, 3835–3849. [\[CrossRef\]](#)
24. Hieronymi, M.; Muller, D.; Doerffer, R. The OLCI neural network swarm (ONNS): A Bio-Geo-Optical algorithm for open ocean and coastal waters. *Front. Mar. Sci.* **2017**, *4*, 140. [\[CrossRef\]](#)
25. Brockmann, C.; Doerffer, R.; Peters, M.; Stelzer, K.; Embacher, S.; Ruescas, A. Evolution of the C2RCC neural network for SENTINEL 1 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. In Proceedings of the Living Planet Symposium 2016, Prague, Czech Republic, 9–13 May 2016.
26. Xie, F.; Tao, Z.; Zhou, X.; Lv, T.; Wang, J.; Li, R. A prediction model of water in situ data change under the influence of environment variables in remote sensing validation. *Remote Sens.* **2021**, *13*, 70. [\[CrossRef\]](#)
27. Cui, A.; Xu, Q.; Gibson, K.; Liu, S.; Chen, N. Metabarcoding analysis of harmful algal bloom species in the Changjiang Estuary, China. *Sci. Total Environ.* **2021**, *782*, 146823. [\[CrossRef\]](#)
28. Breitburg, D. Effects of hypoxia, and the balance between hypoxia and enrichment, on coastal fishes and fisheries. *Estuaries* **2002**, *26*, 767–781. [\[CrossRef\]](#)
29. Zhao, N.; Zhang, G.; Zhang, S.; Bai, Y.; Ali, S.; Zhang, J. Temporal-Spatial Distribution of Chlorophyll-*a* and Impacts of Environmental Factors in the Bohai Sea and Yellow Sea. *IEEE Access* **2019**, *7*, 160947–160960. [\[CrossRef\]](#)
30. Williams, J.J.; Esteves, L.S. Guidance on setup, calibration, and validation of hydrodynamic, wave, and sediment models for shelf seas and estuaries. *Adv. Civ. Eng.* **2017**, *2017*, 5251902. [\[CrossRef\]](#)
31. Lee, G.; Hwang, H.J.; Kim, J.B.; Hwang, D.W. Pollution status of surface sediment in Jinju bay, a spraying shellfish farming area, Korea. *J. Korean Soc. Mar. Env. Saf.* **2020**, *26*, 392–402. [\[CrossRef\]](#)

32. Groom, S.; Sathyendranath, S.; Ban, Y.; Bernard, S.; Brewin, R.; Brotas, V.; Brockmann, C.; Chauhan, P.; Choi, J.; Chuprin, A.; et al. Satellite ocean colour: Current status and future perspective. *Front. Mar. Sci.* **2019**, *6*, 485. [\[CrossRef\]](#)
33. Minnett, P.J.; Alvera-Azcárate, A.; Chin, T.M.; Corlett, G.K.; Gentemann, C.L.; Karagali, I.; Li, X.; Marsouin, A.; Marullo, S.; Maturi, E.; et al. Half a century of satellite remote sensing of sea-surface temperature. *Remote Sens. Environ.* **2019**, *233*, 111366. [\[CrossRef\]](#)
34. Kim, D.K.; Yoo, H.H. Analysis of temporal and spatial red tide change in the south sea of Korea using the GOCI Images of COMS. *J. Korean Assoc. Geogr. Inf. Stud.* **2014**, *22*, 129–136.
35. KIOST. Korea Ocean Satellite Center. Available online: <https://www.kiost.ac.kr/eng.do> (accessed on 30 November 2020).
36. Choi, J.K.; Park, Y.J.; Ahn, J.H.; Lim, H.S.; Eom, J.; Ryu, J.H. GOCI, the world's first geostationary ocean color observation satellite, for the monitoring of temporal variability in coastal water turbidity. *J. Geophys. Res.* **2012**, *117*, C09004. [\[CrossRef\]](#)
37. Lee, K.H.; Lee, S.H. Monitoring of floating green algae using ocean color satellite remote sensing. *J. Korean Assoc. Geogr. Inf. Stud.* **2012**, *15*, 137–147. [\[CrossRef\]](#)
38. Huang, C.; Yang, H.; Zhu, A.; Zhang, M.; Lu, H.; Huan, T.; Zou, J.; Li, Y. Evaluation of the geostationary ocean color imager (GOCI) to monitor the dynamic characteristics of suspension sediment in Taihu Lake. *Int. J. Remote Sens.* **2015**, *36*, 3859–3874. [\[CrossRef\]](#)
39. Concha, J.; Mannino, A.; Franz, B.; Bailey, S.; Kim, T. Vicarious calibration of GOCI for the SeaDAS ocean color retrieval. *Int. J. Remote Sens.* **2019**, *40*, 3984–4001. [\[CrossRef\]](#)
40. Ryu, J.H.; Han, H.J.; Cho, S.; Park, Y.J.; Ahn, Y.H. Overview of geostationary ocean color imager(GOCI) and GOCI Data Processing System(GDPS). *Ocean Sci. J.* **2012**, *47*, 223–233. [\[CrossRef\]](#)
41. Lesser, G.R.; Roelvink, J.A.; van Kester, J.A.T.M.; Stelling, G.S. Development and validation of a three-dimensional morphological model. *Coast. Eng.* **2004**, *51*, 883–915. [\[CrossRef\]](#)
42. Hu, K.; Ding, P.; Wang, Z.; Yang, S. A 2D/3D hydrodynamic and sediment transport model for the Yangtze estuary, China. *J. Mar. Syst.* **2009**, *77*, 114–136. [\[CrossRef\]](#)
43. Dissanayake, P.; Hofmann, H.; Peeters, F. Comparison of results from two 3D hydrodynamic models with field data: Internal seiches and horizontal currents. *Inland Waters* **2019**, *9*, 239–260. [\[CrossRef\]](#)
44. Ramos, V.; Carballo, R.; Ringwood, J.V. Application of the actuator disc theory of Delft3D-FLOW to model far-field hydrodynamic impacts of tidal turbines. *Renew. Energy* **2019**, *139*, 1320–1335. [\[CrossRef\]](#)
45. Xia, M.; Jiang, L. Application of an unstructured grid-based water quality model to Chaeapeake Bay and its adjacent coastal ocean. *J. Mar. Sci. Eng.* **2016**, *4*, 52. [\[CrossRef\]](#)
46. Hartnett, M.; Nash, S. An integrated measurement and modelling methodology for estuarine water quality management. *Water Sci. Eng.* **2015**, *8*, 9–19. [\[CrossRef\]](#)
47. Kim, W.K.; Moon, J.E.; Park, Y.J.; Ishizaka, J. Evaluation of chlorophyll retrievals from Geostationary Ocean Imager(GOCI) for the North-East Asian region. *Remote Sens. Environ.* **2016**, *184*, 482–495. [\[CrossRef\]](#)