Ecological Indicators 110 (2020) 105879

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Deep learning-based retrieval of cyanobacteria pigment in inland water for in-situ and airborne hyperspectral data

Inhyeok Yim[a,1], Jihoon Shin[a,1], Hyuk Lee[b], Sanghyun Park[b], Gibeom Nam[b], Taegu Kang[b], Kyung Hwa Cho[c,*], YoonKyung Cha[a,*]

[a] School of Environmental Engineering, University of Seoul, Dongdaemun-gu, Seoul 130-743, Republic of Korea
[b] Water Quality Assessment Research Division, National Institute of Environmental Research, Environmental Research Complex, Incheon, Republic of Korea
[c] School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan 689-798, Republic of Korea

ARTICLE INFO

Keywords:
Cyanobacteria
Phycocyanin
Hyperspectral imaging
Deep learning
Deep neural networks
Stacked autoencoder

ABSTRACT

Worldwide proliferation of cyanobacteria blooms in inland waters not only affects the intended use of water but potentially threatens human and animal health. In this study, a stacked autoencoder-deep neural network (SAE-DNN) was developed to estimate phycocyanin (PC) concentration by using *in situ* reflectance spectra in productive inland water. The estimated PC using the SAE-DNN was in close agreement with the measured PC, with an R^2 of 0.87, root mean square error (RMSE) of 14.45 μg/L, and relative RMSE of 86.42%. The performance of the SAE-DNN was superior to that of the DNN and band-ratio algorithms. An analysis on the deep spectral features extracted using the SAE yielded the most useful spectral bands, namely 538, 596, and 735 nm, for the retrieval of PC. The estimation accuracy of the SAE-DNN_{Peaks}, using only the aforementioned spectral bands as input variables, was comparable to that of the SAE-DNN, demonstrating that the high-level of abstraction using the SAE facilitated the improvement in feature learning. The application of the SAE-DNN_{Peaks} to airborne hyperspectral image data resulted in an acceptable estimation accuracy, despite a bias toward underestimation, potentially arising from uncertainty associated with atmospheric correction, at high PC concentrations. Our results suggest that simple, empirical-based approaches, such as the SAE-DNN_{Peaks}, have the potential to serve as a rapid assessment tool for the abundance and spatial distribution of cyanobacteria.

1. Introduction

Cyanobacteria blooms are an increasing phenomenon in inland waters worldwide (Huisman et al., 2018; Paerl, 2017; Watson et al., 2016; Yan et al., 2019). Byproducts of the cyanobacterial bloom, such as scum and odorous compounds, hamper the designated uses of the water and cause aesthetic problems (Carmichael and Boyer, 2016). In particular, toxins produced by a number of cyanobacteria genera can pose a threat to human and animal health (Paerl and Otten, 2013). From a management perspective, monitoring cyanobacteria abundance plays a crucial role in taking timely actions against cyanobacteria blooms at an early stage. For monitoring the excess of cyanobacteria, remote sensing-based techniques has the potential to complement traditional field-based monitoring methods, which have disadvantages such as high time expenditure and limited spatial coverage. Recently, a program for remotely monitoring cyanobacteria blooms using airborne hyperspectral sensors has been employed and operated on a trial basis in South Korea. Hyperspectral imagery consists of contiguous, narrow spectral bands; in contrast, multispectral sensors provide images in discrete, broad spectral bands. Although the use of hyperspectral sensors has the advantage of obtaining informative spectral bands that precisely represent the optical properties of a targeted water constituent, selecting an optimal combination of spectral bands among the hundreds of contiguous, correlated bands is challenging.

Different types of models have been developed to predict cyanobacteria abundance in inland waters, often using phycocyanin (PC) pigment as a surrogate (Heddam, 2016; Song et al., 2013; Woźniak et al., 2016). PC is a pigment specific to cyanobacteria and some cryptophytes (Vincent et al., 2004). Bio-optical models based on the theories of optics, aiming at inherent optical properties (IOPs) such as bulk absorption or backscattering, have the potential to be robust and transferable (Dekker, 1993; Randolph et al., 2008). However, the

* Corresponding authors.
E-mail addresses: dmdnr89@gmail.com (I. Yim), sjh3473@uos.ac.kr (J. Shin), ehyuk72@korea.kr (H. Lee), pbaby75@korea.kr (S. Park), gbnam@korea.kr (G. Nam), taegu98@korea.kr (T. Kang), khcho@unist.ac.kr (K.H. Cho), ykcha@uos.ac.kr (Y. Cha).
[1] Co-first author.

https://doi.org/10.1016/j.ecolind.2019.105879
Received 15 April 2019; Received in revised form 25 October 2019; Accepted 29 October 2019
Available online 09 November 2019
1470-160X/

requirement for such models, the measurement of IOPs, restricts their immediate applicability to less studied waterbodies, such as the Geum River, where sufficient measurements of IOPs are not made. Semi-empirical and empirical models have shown successful performance in predicting PC concentrations in inland waters (Dash et al., 2011; Hunter et al., 2010; Hunter et al., 2009; Randolph et al., 2008; Schalles and Yacobi, 2000). Among them, both semi-empirical and empirical band-ratio algorithms using the remote sensing reflectance ($R_{rs}(\lambda)$) for the bands of 709 and 620 nm have been widely applied and evaluated (Duan et al., 2012; Hunter et al., 2010; Pyo et al., 2018; Randolph et al., 2008; Simis et al., 2005). However, the $R_{rs}(709)/R_{rs}(620)$ algorithms are subject to overestimation at relatively low PC concentrations, even if the interference effect of chlorophyll-*a* (Chl-a) absorption around 620 nm is accounted for in the nested semi-empirical structure (Simis et al., 2005). An alternative empirical band-ratio algorithm using 700 and 600 nm was proposed and claimed to be less affected by Chl-a interference (Mishra et al., 2009; Ogashawara et al., 2013).

In this study, a stacked autoencoder-deep neural network (SAE-DNN) is proposed as a methodological framework for selecting the optimal spectral bands and predicting PC concentrations in inland waters. Deep learning (DL), a multi-layered artificial neural network with the capability to learn features from raw data without domain knowledge, has excelled in terms of prediction accuracy in many different tasks (Bengio et al., 2013; Goodfellow et al., 2016; LeCun et al., 2015). Among a variety of DL architectures, the SAE-DNN was selected owing to its suitability for the analytic purpose of this study: automatically extracted useful spectral features in the SAE are fed into a single-layered, fully connected network to simultaneously predict PC concentration. Compared to the previously published machine learning methods (Keller et al., 2018; Song et al., 2014), the integrated architecture of the SAE-DNN removes the need for independently reducing dimensionality, and enables the weight parameters of the network to be updatable with new data. The objectives of this study were to 1) develop the SAE-DNN for the retrieval of PC concentration from in-situ hyperspectral images in Baekje Reservoir of the Geum River, 2) evaluate and compare its performance with the band-ratio algorithms using $R_{rs}(709)/R_{rs}(620)$ and $R_{rs}(700)/R_{rs}(600)$, and 3) select the minimum number of useful spectral bands for predicting PC concentration. Furthermore, the applicability of the SAE-DNN to airborne hyperspectral imagery was examined.

2. Materials and methods

2.1. Region of study and data sets

Baekje Reservoir (36°32′N 126°94′E), formed as the result of the construction of Baekje Weir in 2012, is located approximately 55 km upstream of the Geum River estuary (Fig. 1). The basin area for the reservoir is 7976 km^2, which corresponds to approximately 80% of the entire Geum River basin area. The reservoir has the total storage capacity of 24.2×10^6 m^3. Shifting from the lotic to lentic system, the reservoir has experienced reoccuring algal blooms frequently dominated by cyanobacteria genera, such as *Microcystis*, *Anabaena*, *Aphanizomenon*, or *Oscillatoria*, during warm seasons (Cha et al., 2017; Park et al., 2017).

Field trips were conducted across the reservoir to collect in-situ R_{rs} and water samples. From a total of 13 trips, eight and five trips were undertaken during the periods of June–October of 2016 and September–November of 2017, respectively. Note that no trip was conducted in July of either year owing to the lasting rainy weather affected by the East Asian monsoon. Each trip, starting from the Baekje Weir extending to 10–15 km upstream of the weir, included 12–20 sampling sites. In-situ R_{rs} was measured on the water surface using a FieldSpec HandHeld 2 spectroradiometer (ASD Inc., Boulder, CO, USA), which had a wavelength range of 325–1075 nm. The sampling and analytical methods for PC and Chl-a are described in detail in Pyo et al. (2017).

For the period during June–November of 2016 and 2017, cyanobacteria cell count data were obtained from the online Water Environment Information Systems (http://water.nier.go.kr/publicMain/mainContent.do) of the National Institute of Environmental Research. The data were monitored weekly to biweekly and the monitoring site was located at 500 m upstream of Baekje Weir (Fig. 1).

Airborne imagery data were acquired on the same days when the field campaigns were undertaken. Each airborne campaign began at 8:30 a.m. and lasted two to three hours. The hyperspectral images, which were measured using an AISA eagle sensor (SPECIM Inc., Finland), had 127 wavelength bands from 404 nm to 996 nm with a spectral resolution of 4–5 nm and a spatial resolution of 2×2 m. Atmospheric correction was performed using the atmospheric and topographic correction 4 (ATCOR 4) software.

2.2. Model development

2.2.1. Deep neural networks with pre-training using stacked autoencoder (SAE-DNNs)

For the model comparison, a deep neural network (referred to as DNN) and a deep neural network with pre-training using a stacked autoencoder (referred to as SAE-DNN) were developed for the retrieval of PC concentration (Fig. 2). The DNN is a type of DL algorithm, consisting of input and output layers linked by multiple hidden layers (Fig. 2a). In the DNN, input values are fed to the hidden layer to automatically extract useful features (i.e., feature learning), and the extracted features are used to estimate output values (i.e., predictive learning) (Fig. 2b).

To improve the feature learning ability of the DNN, the greedy layer-wise pre-training method using a stacked autoencoder (SAE) was employed (Bengio et al., 2007). The SAE-DNN procedure includes local greedy layer-wise pre-training followed by global supervised fine-tuning (Fig. 3). In the SAE consisting of multiple autoencoders, each hidden layer (from bottom to top) is pre-trained in an unsupervised manner to capture the variation in its input, and the last hidden layer is connected to the simple perceptron (SP), whereby the parameters, weights, and biases of the SP are pre-trained in an supervised manner to estimate output values (Fig. 3a). Then, all pre-trained hidden layers and the output layer are stacked to construct a SAE-DNN (Fig. 3a). During the fine-tuning step, the SAE-DNN is fine-tuned in a supervised learning manner for improved generalization (Fig. 3b). Note that the decoders, in which the extracted features are mapped back to have the same dimensionality of the input variables, which are a component of the SAE architecture, were not necessary for the modeling purpose, and therefore, were discarded (Fig. 2). The parameters of the SAE-DNN are initialized during the greedy layer-wise pre-training step layer by layer, and are optimized globally during the fine-tuning step (Bengio et al., 2007; Erhan et al., 2010; Larochelle et al., 2007; Vincent et al., 2008; Vincent et al., 2010).

The calibration and validation were implemented using the MATLAB 2018a neural network toolbox (Mathworks, 2013). Four built-in functions (trainAutoencoder.m, perceptron.m, stack.m, and train.m) and two built-in functions (fitnet.m and train.m) in the neural networks toolbox, were used for the DNN and the SAE-DNN, respectively. The entire dataset was divided into calibration and validation datasets based on the interleaved selection method. As a result, 70% (n = 142) and 30% (n = 61) of the data points were assigned to calibration and validation, respectively.

To minimize the errors caused by different scales of variables, both the in-situ reflectance and measured PC concentration were normalized. Each variable was rescaled to have the range from 0 to 1 by subtracting the mean from each value and dividing it by the difference between maximum and minimum values. Note that the maximum and minimum values of each variable are selected from the calibration

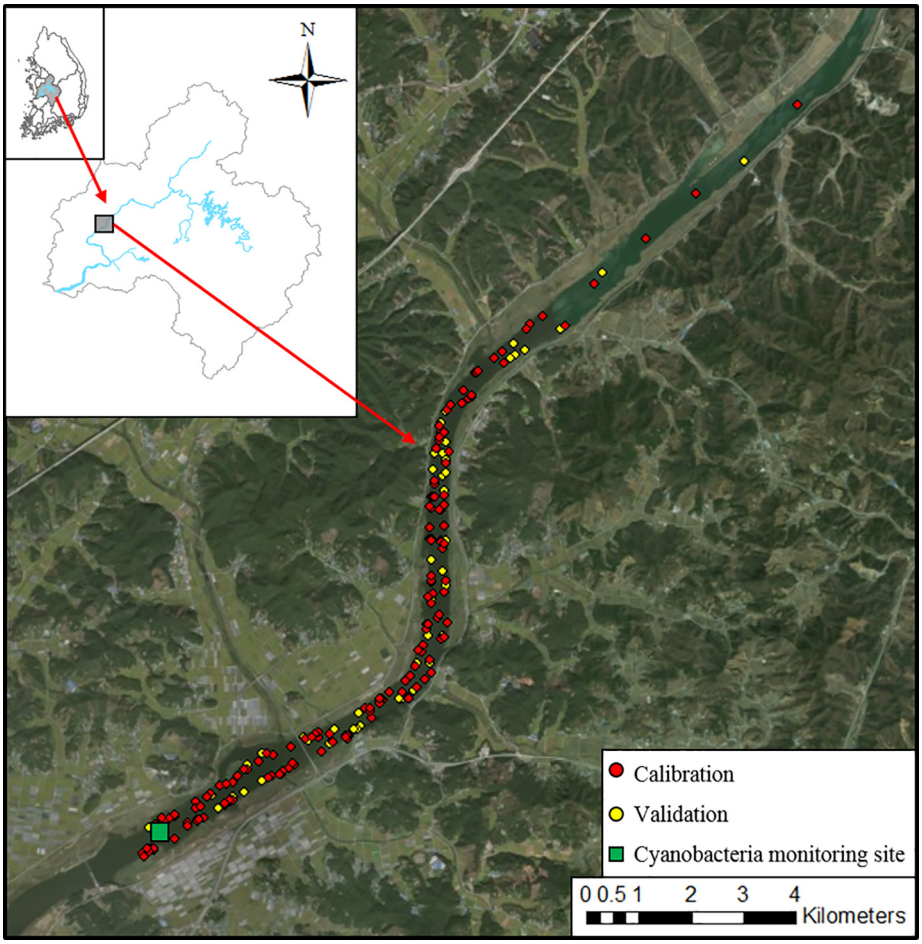

Fig. 1. Study region and locations of field campaigns. Red and yellow circles indicate monitoring data used for calibration and validation, respectively. Green square indicates the monitoring location for cyanobacteria cell counts.

dataset, not to inform the validation dataset of prior knowledge.

$$y_i^{nor} = \frac{y_i - \min(y_{cal})}{\max(y_{cal}) - \min(y_{cal})} \quad (1)$$

$$R_{rsi}^{nor}(\lambda_k) = \frac{R_{rsi}(\lambda_k) - min(R_{rs,cal}(\lambda))}{max(R_{rs,cal}(\lambda)) - min(R_{rs,cal}(\lambda))} \quad (2)$$

where y_{cal} and y_i^{nor} are the PC concentration of the calibration dataset and normalized PC concentration, respectively, $R_{rs,cal}(\lambda)$ and $R_{rsi}^{nor}(\lambda_k)$ are the R_{rs} values of the calibration dataset and normalized R_{rs} values,

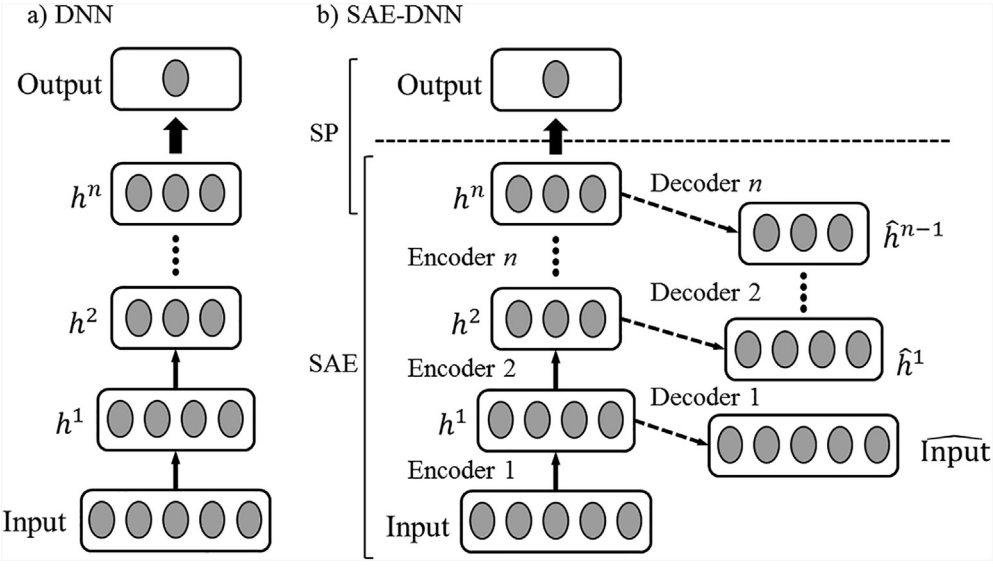

Fig. 2. Architectures of a) DNN and b) SAE-DNN. SAE and SP denote the stacked autoencoder and simple perceptron, respectively.

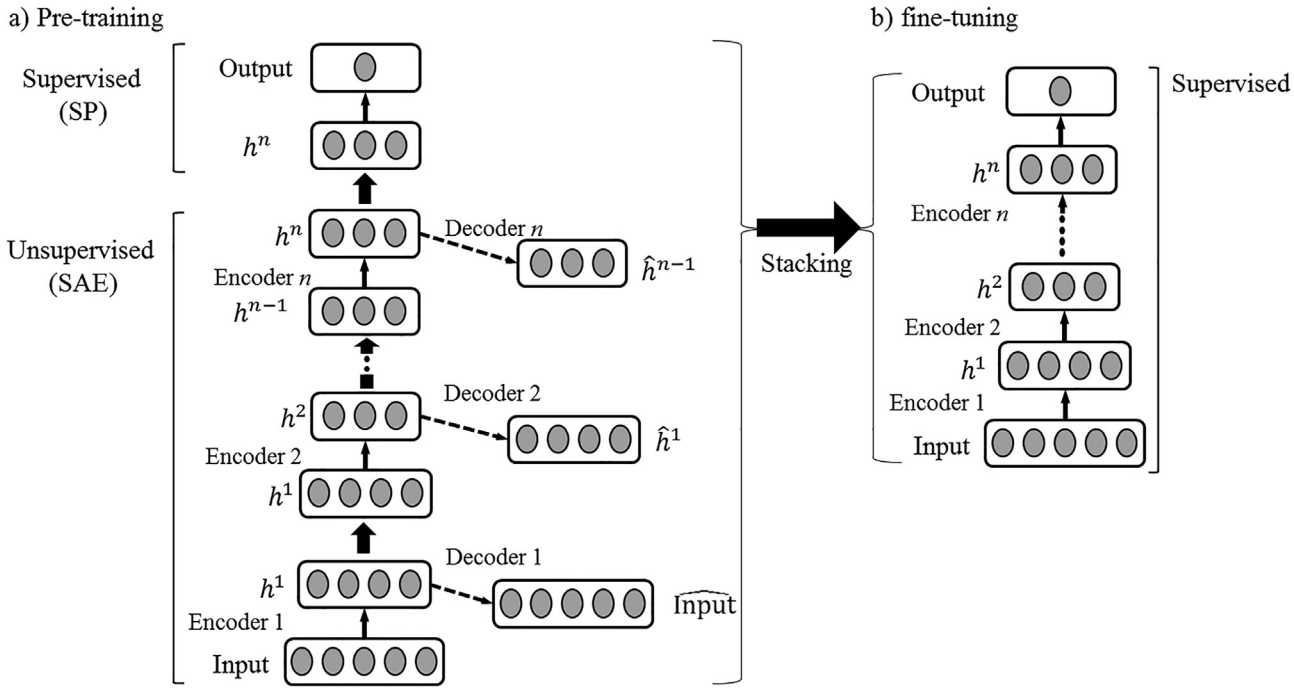

Fig. 3. Schematic procedure of training SAE-DNN with a) greedy layer wise pre-training and b) fine tuning process.

respectively, and λ_k is the kth spectral band.

2.2.2. Model set-up

For the calibration process, hyperparameters in the model architecture and optimizer were determined. The number of hidden neurons (*nH*) in each layer was selected based on the geometric pyramid rule (Masters, 1993). The log-sigmoid function was selected as the activation function in all layers (both hidden and output) owing to its capability to address non-linearity. The scaled-conjugate-gradient (SCG) algorithm was selected as the training algorithm because it facilitates faster optimization by avoiding the iteration of line search per learning and selection of hyperparameters, such as the learning rate and momentum constant (Møller, 1993). To evaluate the effects of using only pre-training, other hyperparameters, which can also affect the performance of DNNs, were not tuned.

Optimization results converge to a different stage every time the model is run because the training process of a DNN is non-deterministic (Goodfellow et al., 2016). Therefore, each one of the DNN and the SAE-DNN models was run 100 times to check the variability of the performance.

2.3. Model evaluation

2.3.1. Relative importance analysis

The relative importance (RI) of input variables measures the relative contribution of each input variable to the variation in the output variable (Gevrey et al., 2003; Olden and Jackson, 2002; Olden et al., 2004; Pianosi et al., 2016). Therefore, the RI analysis can be used to evaluate whether the extracted features from a DNN are useful for conducting specific tasks and for gaining meaningful insights. The algorithm for calculating the weight parameters and RI of input variables (Gevrey et al., 2003) was modified in this study to make it suitable for neural networks with multiple hidden layers (i.e., $nL > 2$). The modified equations can be expressed as

$$\mathbf{M} = W^{(0)}\left(\prod_{l=2}^{nL} W^{(l)}\right)W^{(1)} \tag{3}$$

$$\boldsymbol{a} = \text{abs}(\mathbf{M}) \tag{4}$$

$$\boldsymbol{b} = \text{sum}(\boldsymbol{a}) \tag{5}$$

$$\boldsymbol{RI}(\%) = 100 \times \text{rdivide}(\boldsymbol{a}, \boldsymbol{b}) \tag{6}$$

where $\mathbf{M}$ is the $D^y \times D^x$ matrix generated by multiplying all weight matrices for the output and input layers, respectively; $\boldsymbol{a}$ is the $D^y \times D^x$ matrix taking values in the matrix $\mathbf{M}$ and returning their absolute values; $\boldsymbol{b}$ is the row vector returning the sum of each column in the matrix $\boldsymbol{a}$; j is an index denoting output variables (j = 1, …, J, J = number of output variables); k is an index denoting input variables (k = 1, 2, 3, …, K, K = number of input variables); $\boldsymbol{RI}$ is the $D^y \times D^x$ matrix containing the relative importance of RI_{jk}, where RI_{jk} is the relative importance of x_k with respect to the j^{th} output variable; and *rdivide* is the function dividing each element of $\boldsymbol{a}$ by the corresponding element of $\boldsymbol{b}$.

2.3.2. Estimation accuracy

The performance of the model was evaluated by the coefficient of determination (R^2), root mean square error (RMSE), relative root mean square error (rRMSE), and relative residual. The R^2, RMSE, rRMSE, and relative residual were calculated as

$$\text{rRMSE} = 100 \times \frac{\text{RMSE}}{\bar{y}} \tag{7}$$

$$\text{relative residual} = \frac{|\widehat{y} - y_i|}{y_i} \tag{8}$$

where y_i is the measured PC concentration (μg/L), $\widehat{y_i}$ is the estimated PC concentration, and $\bar{y}$ is the average measured PC concentration.

3. Results

3.1. Characteristics of water constituents

The concentrations of phytoplankton pigments, Chl-a and PC, for the study sites exhibited substantial monthly variations during the summer and fall of 2016 and 2017 (Fig. 4). Chl-a concentrations ranged from 8.45 to 130.37 μg/L, with a mean of 32.43 μg/L. Chl-a concentrations were generally higher in August, the warmest month the

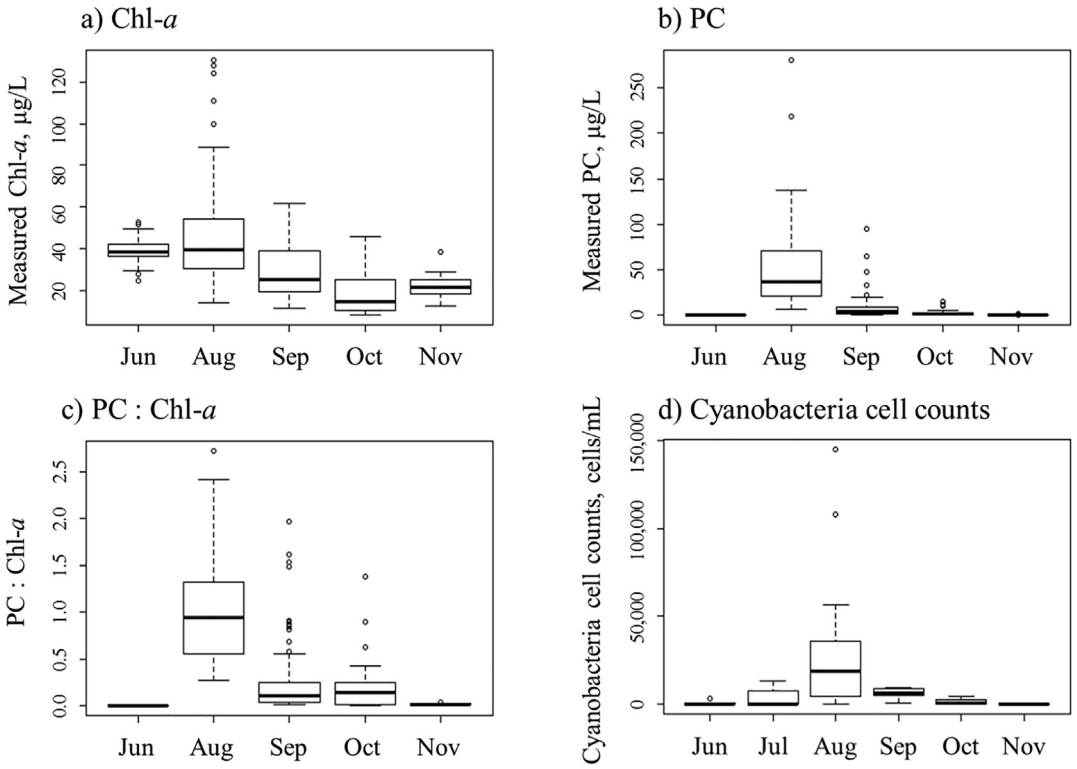

Fig. 4. Monthly variations of a) measured Chl-*a* concentration (µg/mL), b) measured PC concentration (µg/mL), c) PC:Chl-*a*, and d) cyanobacteria cell counts (cells/mL) during the field monitoring period.

year, than in other months (Fig. 4a). The temporal changes in PC concentrations, ranging from 0.02 to 280.87 µg/L (mean = 16.41 µg/L), were more dynamic than those in Chl-a concentrations (Fig. 4b). The PC concentrations, which were extremely low in June with a monthly maximum of 0.45 µg/L, peaked in August with a monthly maximum of 280.87 µg/L, followed by a rapid decrease toward November.

The seasonal variations of PC concentrations were reflected in the PC:Chl-a ratios, which had a range of 0–2.41 and a mean of 0.38 (Fig. 4c). The PC:Chl-a ratio > 0.5, which is an indicator of cyanobacteria dominance in the phytoplankton population (Hunter et al., 2009), comprised ~ 28% (58 out of 204) of the total samples, and ~ 81% of the PC:Chl-a ratio > 0.5 occurred during August. In contrast, the PC:Chl-a ratios observed in June and November did not exceed 0.5. The monthly patterns of cyanobacteria cell counts were similar to those of PC and PC:Chl-a (Fig. 4d). Ranging from 0 to 145,198 cells/mL, cyanobacteria cell counts were distinctly higher in August than in other months. The low levels of cyanobacteria cell counts along with low PC:Chl-a ratios in June and November suggested that the contribution of cyanobacteria abundance to phytoplankton abundance was minor in those months.

PC concentrations tended to increase with increasing Chl-a concentrations, with a determination coefficient (R^2) of 0.47 (Fig. 5). The relationship between PC and Chl-a was weak for PC:Chl-a < 0.5 (R^2 = 0.20), whereas a positive relationship between the two pigments was clear for PC:Chl-a ≥ 0.5 (R^2 = 0.70) (Fig. 5).

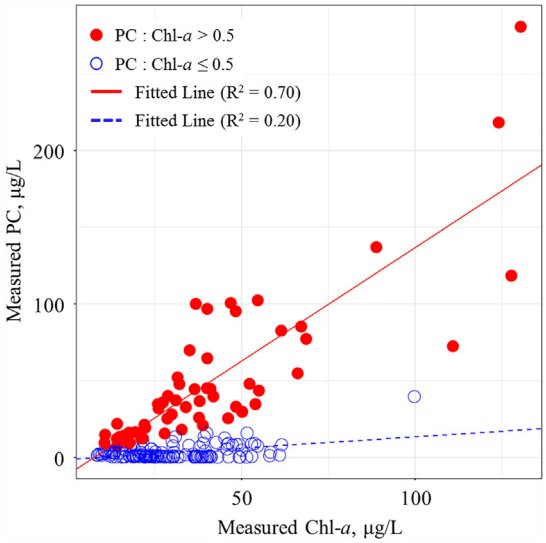

Fig. 5. Relationship between PC and Chl-*a* concentrations.

3.2. Model performances

3.2.1. Performance of DNNs

The SAE-DNN appeared more successful than the DNN for the retrieval of PC from Baekje Reservoir data. The SAE-DNN yielded an RMSE of 8.78 µg/L and rRMSE of 55.32% for the calibration data set and an RMSE of 14.45 µg/L and rRMSE of 86.42% for the validation data set, whereas the estimation accuracy of the DNN was marginally lower, with an RMSE of 11.83 µg/L and rRMSE of 74.55% for the calibration data set and an RMSE of 16.24 and rRMSE of 97.11% for the validation data set (Table 1). The SAE-DNN resulted in a strong linearity between the estimated and measured PC, with an R^2 of 0.92 and 0.87 for the calibration and validation data sets, respectively (Fig. 6b). The DNN resulted in a slightly weaker relationship (R^2 = 0.85 and 0.84 for the calibration and validation data sets, respectively), exhibiting larger deviations of the estimated PC from the 1:1 line at a measured PC range of approximately 50–100 µg/L (Fig. 6a).

The differences in model performance may arise from the different characteristics of feature learning between SAE-DNN and DNN. The RI analysis during the model calibration suggested that the SAE-DNN provided greater differentiation for determining important input variables (Fig. 7). The RI values from the DNN had a narrow range, from

Table 1
Summary of estimation accuracy of DNNs and band-ratio algorithms.

Category	Algorithm	Calibration			Validation		
		R^2	RMSE (μg/L)	rRMSE (%)	R^2	RMSE (μg/L)	rRMSE (%)
DNN	DNN	0.85	11.83	74.55	0.84	16.24	97.11
	SAE-DNN	0.92	8.78	55.32	0.87	14.45	86.42
	SAE-DNN_{Peaks}	0.84	12.25	77.22	0.88	14.36	85.88
	LM_{Peaks}	0.61	19.01	119.82	0.52	28.11	168.07
Band-ratio	$R_{rs}(709)/R_{rs}(620)$	0.64	18.27	115.14	0.77	19.50	116.60
	$R_{rs}(700)/R_{rs}(600)$	0.57	19.91	125.47	0.63	24.64	147.38

0.18 to 0.33 (Fig. 7a), whereas the SAE-DNN yielded a substantially wider RI range, from 0.04 to 0.63, with a number of pronounced peaks (Fig. 7b). Five input variables corresponded to RI peaks, which had significantly higher importance ($p < 0.05$) than the average RI (=0.25%). The most pronounced two peaks, located at 538 and 596 nm, were within the visible light range of 400–800 nm, and the less distinct three peaks at 735, 757, and 776 nm were within the near-infrared (NIR) range (Fig. 7b).

A simplified network model was developed by using only the three spectral bands located at the RI peaks, 538, 596, and 735 nm, as input variables of the SAE-DNN (referred to as the $SAE\text{-}DNN_{peaks}$). Note that replacing 735 nm with either 757 or 776 nm resulted in insignificant differences, and therefore only the results using 735 nm were reported. In addition to the $SAE\text{-}DNN_{peaks}$, a multiple linear regression model (referred to as the ML_{peaks}) for PC was developed using $R_{rs}(538)$, $R_{rs}(596)$, and $R_{rs}(735)$ as input variables to examine the linearity between the output and input variables. For the calibration data set, the estimation accuracy of $SAE\text{-}DNN_{peaks}$ (R^2 = 0.84, RMSE = 12.25 μg/L, and rRMSE = 77.22%) was slightly lower than that of SAE-DNN but was comparable to that of the DNN (Table 1 and Fig. 6c). The validation results suggested that the performance of the $SAE\text{-}DNN_{peaks}$ (R^2 = 0.88, RMSE = 14.36 μg/L, and rRMSE = 85.88%) was comparable to that of the SAE-DNN and marginally better than that of the DNN (Table 1 and Fig. 6c).

The fitted regression line for LM_{Peaks} was as follows:

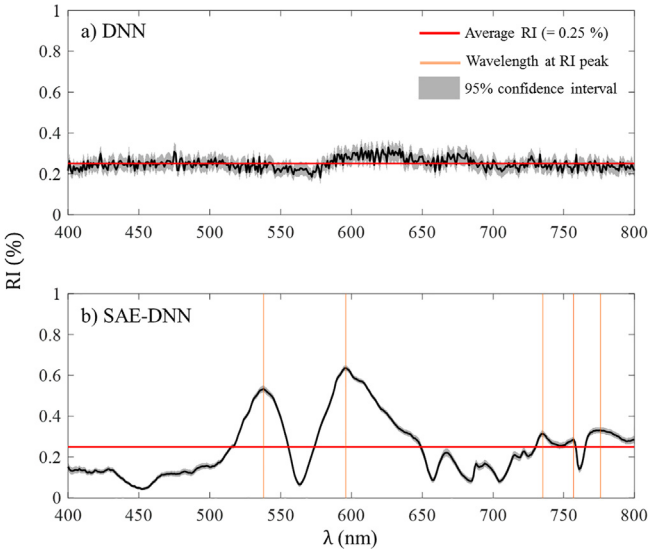

Fig. 7. Results of the RI analysis based on a) DNN, and b) SAE-DNN.

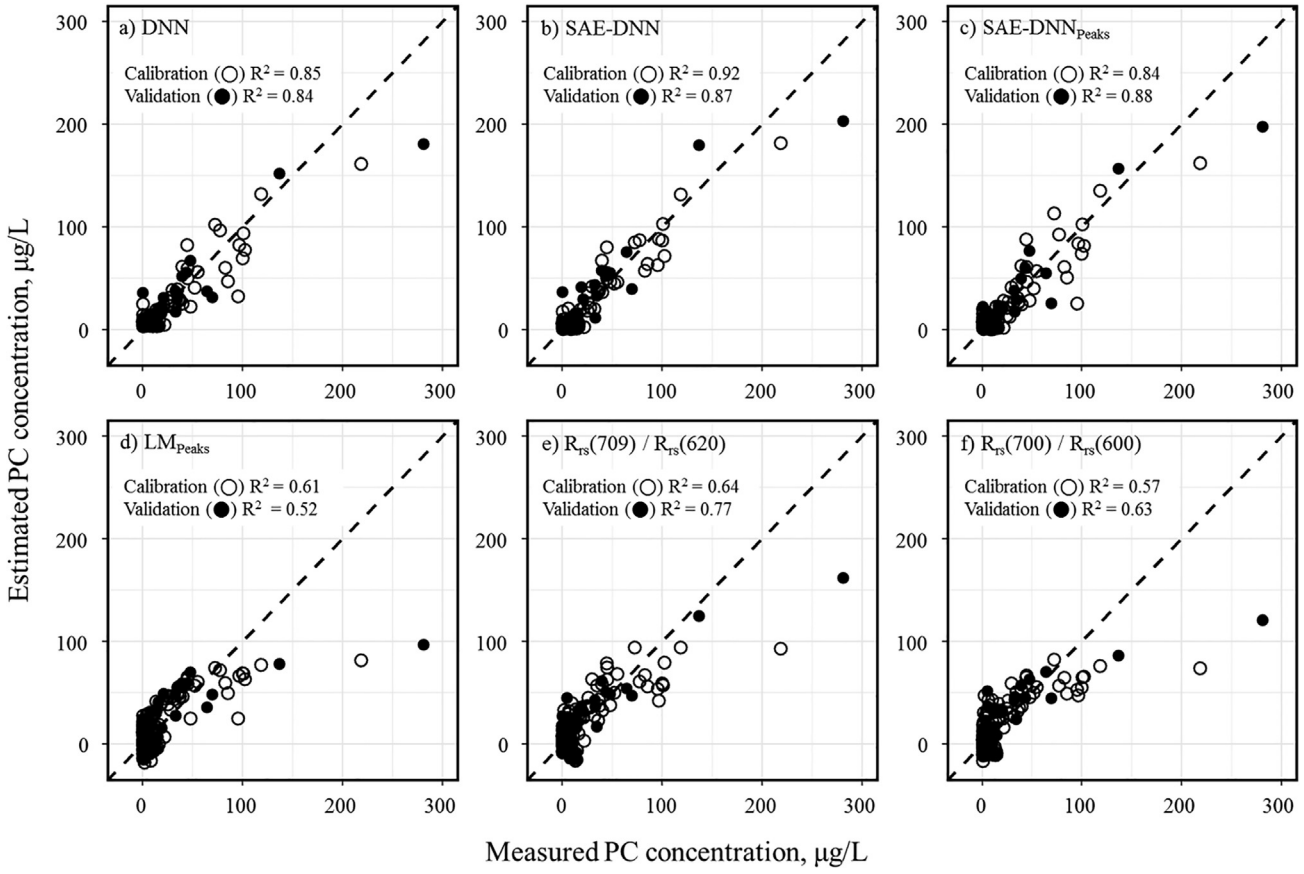

Fig. 6. Relationships between estimated and measured PC using a) DNN, b) SAE-DNN, c) $SAE\text{-}DNN_{Peaks}$, d) LM_{Peaks}, e) $R_{rs}(709)/R_{rs}(620)$, and f) $R_{rs}(700)/R_{rs}(600)$.

$$\widehat{PC} = 12493.29 - 13734.88 \cdot Rrs(538) + 2939.36 \cdot Rrs(596) + 19.21 \cdot Rrs(735) \quad (9)$$

where $\widehat{PC}$ indicates the estimates for PC concentration. A comparison of the LM_{Peaks} with the SAE-DNN_{peaks} suggested that the use of the linear regression as a methodology for predicting PC concentration yielded a decrease in estimation accuracy, as demonstrated by R^2 (0.59 and 0.56 for calibration and validation, respectively), RMSE (19.01 and 28.11 μg/L), and rRMSE (119.82 and 168.07%) (Table 1 and Fig. 6d).

3.2.2. Performance of band-ratio algorithms

The published band-ratio algorithms, using the ratios $R_{rs}(709)/R_{rs}(620)$ (Simis et al., 2005) and $R_{rs}(700)/R_{rs}(600)$ (Mishra et al., 2009), were tuned with the calibration data set resulting in the following equations:

$$\widehat{PC} = 222.4 - 173.06 \cdot R_{rs}(720)/R_{rs}(620) \quad (10)$$

$$\widehat{PC} = 164.42 - 141.14 \cdot R_{rs}(700)/R_{rs}(600) \quad (11)$$

where $\widehat{PC}$ indicates the estimates for PC concentrations. The band-ratio algorithm using $R_{rs}(709)/R_{rs}(620)$ performed better ($R^2 = 0.64$ and 0.77, RMSE = 18.27 and 19.50 μg/L, and rRMSE = 115.14 and 116.60% for calibration and validation, respectively) than the algorithm using $R_{rs}(700)/R_{rs}(600)$ ($R^2 = 0.57$ and 0.63, RMSE = 19.91 and 24.64 μg/L, and rRMSE = 125.47 and 147.38%) (Table 1 and Fig. 6e and f). The estimation accuracy of linear regression-based algorithms, i.e., the ML_{peaks} and the two band-ratio algorithms, was lower than that of DNN-based algorithms (Table 1 and Fig. 6). Particularly, the linear regression-based algorithms had a distinct tendency to overestimation at high PC > ~75 μg/L (Fig. 6d–f).

4. Discussion

4.1. Performance comparison

The obtained results using the calibration and validation data sets for Baekje Reservoir illustrate that both the DNN-based approaches and band-ratio algorithms performed reasonably well for predicting PC concentrations in inland waters (Table 1 and Fig. 6). However, the SAE-DNN outperformed the DNN model, suggesting that the pre-training of network parameters through multiple levels of abstraction facilitated the extraction of useful spectral features. The efficacy of the deep feature learning by the SAE yielded an improved estimation accuracy. Further, the satisfactory performance of SAE-DNN_{Peaks} confirms that the extracted spectral features had the sufficient level of information for estimation of PC concentrations, demonstrating the feasibility of using a simplified network structure with substantially reduced computation load.

The performance of DNNs (DNN, SAE-DNN, and SAE-DNN_{Peaks}) was generally superior to that of linear regression models (LM_{Peaks} and band-ratio algorithms) (Table 1 and Fig. 6). The difference in prediction accuracy suggests that the nature of the relationship between cyanobacterial pigment and reflectance spectra is non-linear, and the DNNs were able to address the complexity and non-linearity in spectral features.

4.2. Interpretability of selected spectral bands

Based on the RI analysis of the SAE-DNN results, 538 and 596 nm were the spectral bands corresponding to the most pronounced two RI peaks (Fig. 7), indicating that the R_{rs} of these wavelengths were the most relevant to the variation in PC concentrations. Owing to the distinct absorption feature of PC, the band of 620 nm has been the most commonly used in the models for retrieval of PC concentration (Hunter et al., 2010; Hunter et al., 2009; Lyu et al., 2013; Randolph et al., 2008; Schalles and Yacobi, 2000; Simis et al., 2005). However, a few attempts to use the wavelength range of 500–600 nm for quantifying cyanobacteria abundance also showed promising results. Dash et al. (2011) used the spectral slopes of bands 4 (510.6 nm) and 5 (556.4 nm) from the Ocean-1 satellite Ocean Colour Monitor to quantify PC concentration. Without the correction for the effects of other optically active components (OACs), the model showed good performance with an R^2 of 0.7248. The results from the partial least squares regression developed by Jin et al. (2017) indicated that among 15 MERIS bands, three bands, including B5 (560 nm), had the largest contribution to explain the variance of cyanobacteria abundance. The predicted PC concentration from the band-ratio algorithm proposed by Mishra et al. (2009), which used 600 nm rather than 620 nm, showed good agreement with the measured PC concentration, and the model performance was less affected by the presence of Chl-a.

Our results indicate that the R_{rs} around the bands of 538 and 596 nm contains useful information for predicting cyanobacteria abundance. Along the spectral reflectance curve, 538 nm belongs to the reflectance shoulder induced by the absorption maxima, 470–530 nm, of carotenoids, a pigment suite of cyanobacteria, which exhibit a high correlation with PC (Schalles and Yacobi, 2000). Thus, the R_{rs} around 538 nm reflects the cyanobacteria abundance together with PC concentration relative to the green reflectance peak, where the absorption by Chl-a and carotenoids is minimal (Mishra et al., 2009; Schalles and Yacobi, 2000). In addition, the 596-nm band is located within the reflectance shoulder induced by the maximum PC absorption at approximately 620 nm. The R_{rs} around 596 nm, where the proportion of PC absorption to Chl-a absorption is higher than that at 620 nm, may better represent the PC absorption not interfered with Chl-a than the absorption at 620 nm (Mishra et al., 2009; Ogashawara et al., 2013).

4.3. The effects of Chl-a interference on model performance

Previous studies reported that the estimation accuracy of empirical or semi-empirical models are apt to be hindered by the presence of Chl-a; the errors of estimated PC increased with decreasing PC:Chl-a ratio, suggesting that an increase in the relative contribution of Chl-a to total pigment absorption interfered with model performance (Hunter et al., 2010; Mishra et al., 2009; Ogashawara et al., 2013; Randolph et al., 2008; Simis et al., 2005). In agreement with previous results, the estimation of PC using the SAE-DNN, the best-performing model in this study, yielded an inverse relationship of relative residual with PC concentration and also with PC:Chl-a (Fig. 8). The absolute value of relative residual had a range of 0–387.17, with the maximum value occurring at an extremely low PC concentration of 0.015 ug/L and PC:Chl-a ratio < 0.01. For the absolute values of relative residual > 10, the mean PC concentration was merely 0.24 and the mean PC:Chl-a ratio was as lower than 0.01. The model error decreased rapidly and stabilized with increasing PC:Chl-a ratio (Fig. 8b), with a mean absolute value of relative residual reaching 0.36 at PC:Chl-a > 0.5. However, the SAE-DNN results did not show any tendency of overestimation at low PC concentrations (Fig. 6b), indicating that high relative residuals at low PC concentrations may not arise from Chl-a interference, but may arise from other sources, such as measurement errors.

4.4. Applicability to airborne hyperspectral images

The feasibility of using the simplified network, SAE-DNN_{Peaks}, as a tool for monitoring cyanobacteria abundance in inland waters, was examined by applying the model to airborne hyperspectral image data. Airborne imagery with the three spectral bands (538, 596, and 735 nm) used in the SAE-DNN_{Peaks} was fed to the trained SAE-DNN_{Peaks} as inputs. The overall estimation accuracy of the model was acceptable, with an RMSE of 15.31 μg/L and rRMSE of 124.11% (Fig. 9). The estimated PC using the SAE-DNN_{Peaks} from the airborne images was generally in line with the measured PC, with an R^2 of 0.47, but exhibited a tendency

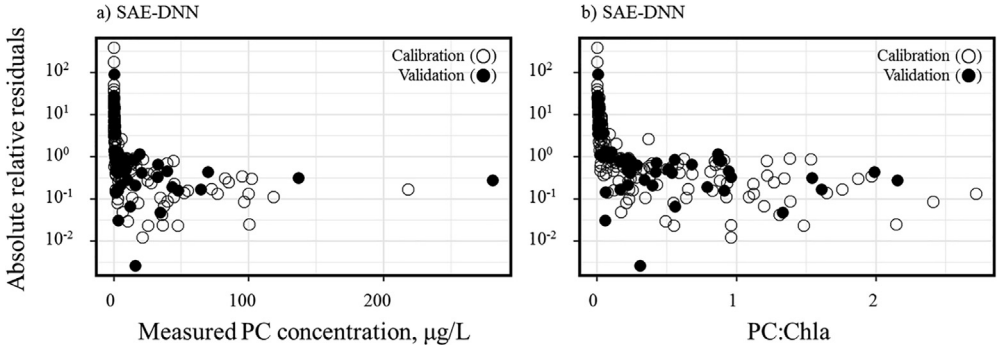

Fig. 8. Relationships between absolute relative residuals of estimated PC and a) measured PC concentration and b) PC:Chl-*a*.

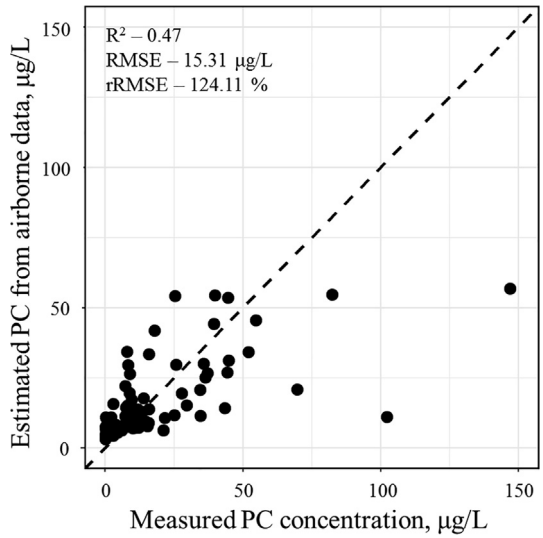

Fig. 9. Relationship between estimated PC using the SAE-DNN$_{Peaks}$ for airborne hyperspectral data and measured PC.

of underestimation at high PC concentrations (Fig. 9). A convincing candidate for the source of error may be the lack of an appropriate sample size, especially at high PC concentrations. Despite the fact that high PC concentrations and cyanobacteria dominance in Baekje Reservoir occur most frequently during August, the field and airborne monitoring campaigns were conducted only three times in August of 2016 and were not conducted until September in 2017. Another potential cause for the error arises from the uncertainty associated with the atmospheric correction. A recent study using the hyperspectral imagery for Baekje Reservoir indicated that among the different methods for atmospheric correction, ATCOR 4 performed poorer than MODTRAN 6 and the artificial neural network (Pyo et al., 2018). Furthermore, the finding of Pyo et al. (2018), namely the negative Nash–Sutcliffe efficiency (NSE) of the estimated PC using ATCOR 4, adds to the evidence that the errors associated with atmospheric correction are attributable to the bias toward underestimation with increasing PC (Fig. 8).

5. Conclusions

The SAE-DNN was proposed as a modeling approach for the retrieval of PC in inland waters. The results highlight that the SAE-DNN outperforms the DNN and band-ratio algorithms, proving the efficacy of using the SAE for deep feature learning and improved predictive learning. The increases in model errors, indicated by the relative residual error, with decreasing PC:Chl-a ratio did not result in the overestimation of PC at low PC concentrations, indicating that Chl-a absorption may not interfere with the model performance. The SAE-DNN provided improved differentiation of the RI among the spectral bands for estimating PC, and as a result, 538, 596, and 735 nm were selected as the important bands. The development and successful performance of the SAE-DNN$_{Peaks}$, using only the selected bands as inputs, made it possible to apply the model to airborne hyperspectral images. The estimated PC using the SAE-DNN$_{Peaks}$ for airborne data showed a moderate correlation with the measured PC but exhibited underestimation at high PC concentrations.

To corroborate the feasibility of the SAE-DNN for the retrieval of PC in inland waters, future research should be focused on the following: 1) supplementing with IOP measurements to account for the effects of OACs on model performance; 2) acquiring more field data that represent the seasonality and interannual variability of pigment concentration to better validate the model and improve the estimation accuracy; 3) extending the application of the SAE-DNN to the other three major rivers of South Korea to increase model transferability; and 4) improving atmospheric correction to better examine the applicability of the SAE-DNN$_{Peaks}$ to airborne hyperspectral imagery.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by National Institute of Environmental Research funded by Ministry of Environment, South Korea [NIER-2018-03-01-005].

References

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.
Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks. Adv. Neural Inf. Process. Syst, 153–160.
Carmichael, W.W., Boyer, G.L., 2016. Health impacts from cyanobacteria harmful algae blooms: implications for the North American Great Lakes. Harmful Algae 54, 194–212.
Cha, Y., Cho, K.H., Lee, H., Kang, T., Kim, J.H., 2017. The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers. Water Res. 124, 11–19.
Dash, P., Walker, N.D., Mishra, D.R., Hu, C., Pinckney, J.L., D'Sa, E.J., 2011. Estimation of cyanobacterial pigments in a freshwater lake using OCM satellite data. Remote Sens. Environ. 115, 3409–3423.
Dekker, A.G., 1993. Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing (Ph.D. thesis). Earth and Life Sciences, Amsterdam, The Netherlands. Proefschrift Vrije Universiteit (Free University).
Duan, H., Ma, R., Hu, C., 2012. Evaluation of remote sensing algorithms for cyanobacterial pigment retrievals during spring bloom formation in several lakes of East China. Remote Sens. Environ. 126, 126–135.
Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11, 625–660.
Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Model. 160,

249–264.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge.
Heddam, S., 2016. Multilayer perceptron neural network-based approach for modeling phycocyanin pigment concentrations: case study from lower Charles River buoy, USA. Environ. Sci. Pollut. Res. 23, 17210–17225.
Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M., Visser, P.M., 2018. Cyanobacterial blooms. Nat. Rev. Microbiol. 16, 471.
Hunter, P.D., Tyler, A.N., Carvalho, L., Codd, G.A., Maberly, S.C., 2010. Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes. Remote Sens. Environ. 114, 2705–2718.
Hunter, P.D., Tyler, A.N., Gilvear, D.J., Willby, N.J., 2009. Using remote sensing to aid the assessment of human health risks from blooms of potentially toxic cyanobacteria. Environ. Sci. Technol. 43, 2627–2633.
Jin, Q., Lyu, H., Shi, L., Miao, S., Wu, Z., Li, Y., Wang, Q., 2017. Developing a two-step method for retrieving cyanobacteria abundance from inland eutrophic lakes using MERIS data. Ecol. Indic. 81, 543–554.
Keller, S., Maier, P., Riese, F., Norra, S., Holbach, A., Börsig, N., Wilhelms, A., Moldaenke, C., Zaake, A., Hinz, S., 2018. Hyperspectral data and machine learning for estimating CDOM, chlorophyll a, diatoms, green algae and turbidity. Int. J. Environ. Res. Public Health 15, 1881.
Larochelle, H., Erhan, D., Courville, A., Bergstra, J., Bengio, Y., 2007. An empirical evaluation of deep architectures on problems with many factors of variation. In: Proceedings of the 24th International Conference on Machine Learning, pp. 473–480.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lyu, H., Wang, Q., Wu, C., Zhu, L., Yin, B., Li, Y., Huang, J., 2013. Retrieval of phycocyanin concentration from remote-sensing reflectance using a semi-analytic model in eutrophic lakes. Ecol. Inf. 18, 178–187.
Møller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. 6, 525–533.
Masters, T., 1993. Practical Neural Network Recipes in C++. Morgan Kaufmann, San Francisco, CA.
Mishra, S., Mishra, D., Schluchter, W., 2009. A novel algorithm for predicting phycocyanin concentrations in cyanobacteria: a proximal hyperspectral remote sensing approach. Remote Sens. 1, 758–775.
Ogashawara, I., Mishra, D., Mishra, S., Curtarelli, M., Stech, J., 2013. A performance review of reflectance based algorithms for predicting phycocyanin concentrations in inland waters. Remote Sens. 5, 4774–4798.
Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. 154, 135–150.
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178, 389–397.
Paerl, H.W., 2017. Controlling harmful cyanobacterial blooms in a climatically more extreme world: management options and research needs. J. Plankton Res. 39, 763–771.
Paerl, H.W., Otten, T.G., 2013. Harmful cyanobacterial blooms: Causes, consequences, and controls. Microb. Ecol. 65, 995–1010.
Park, Y., Pyo, J., Kwon, Y.S., Cha, Y., Lee, H., Kang, T., Cho, K.H., 2017. Evaluating physico-chemical influences on cyanobacterial blooms using hyperspectral images in inland water, Korea. Water Res. 126, 319–328.
Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ. Model. Softw. 79, 214–232.
Pyo, J., Ligaray, M., Kwon, Y., Ahn, M.-H., Kim, K., Lee, H., Kang, T., Cho, S., Park, Y., Cho, K., 2018. High-spatial resolution monitoring of phycocyanin and chlorophyll-a using airborne hyperspectral imagery. Remote Sens. 10, 1180.
Pyo, J., Pachepsky, Y., Baek, S.-S., Kwon, Y., Kim, M., Lee, H., Park, S., Cha, Y., Ha, R., Nam, G., 2017. Optimizing semi-analytical algorithms for estimating chlorophyll-a and phycocyanin concentrations in inland waters in Korea. Remote Sens. 9, 542.
Randolph, K., Wilson, J., Tedesco, L., Li, L., Pascual, D.L., Soyeux, E., 2008. Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin. Remote Sens. Environ. 112, 4009–4019.
Schalles, J.F., Yacobi, Y.Z., 2000. Remote detection and seasonal patterns of phycocyanin, carotenoid and chlorophyll pigments in eutrophic waters. Ergeb. Limnol. 55, 153–168.
Simis, S.G., Peters, S.W., Gons, H.J., 2005. Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water. Limnol. Oceanogr. 50, 237–245.
Song, K., Li, L., Li, Z., Tedesco, L., Hall, B., Shi, K., 2013. Remote detection of cyanobacteria through phycocyanin for water supply source using three-band model. Ecol. Inform. 15, 22–33.
Song, K., Li, L., Tedesco, L.P., Li, S., Hall, B.E., Du, J., 2014. Remote quantification of phycocyanin in potable water sources through an adaptive model. ISPRS J. Photogram. Remote Sens. 95, 68–80.
Vincent, R.K., Qin, X., McKay, R.M.L., Miner, J., Czajkowski, K., Savino, J., Bridgeman, T., 2004. Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial blooms in Lake Erie. Remote Sens. Environ. 89, 381–392.
Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408.
Watson, S.B., Miller, C., Arhonditsis, G., Boyer, G.L., Carmichael, W., Charlton, M.N., Confesor, R., Depew, D.C., Höök, T.O., Ludsin, S.A., Matisoff, G., McElmurry, S.P., Murray, M.W., Peter Richards, R., Rao, Y.R., Steffen, M.M., Wilhelm, S.W., 2016. The re-eutrophication of Lake Erie: harmful algal blooms and hypoxia. Harmful Algae 56, 44–66.
Woźniak, M., Bradtke, K., Darecki, M., Krężel, A., 2016. Empirical model for phycocyanin concentration estimation as an indicator of cyanobacterial bloom in the optically complex coastal waters of the Baltic Sea. Remote Sens. 8, 212.
Yan, D., Xu, H., Yang, M., Lan, J., Hou, W., Wang, F., Zhang, J., Zhou, K., An, Z., Goldsmith, Y., 2019. Responses of cyanobacteria to climate and human activities at Lake Chenghai over the past 100 years. Ecol. Indic. 104, 755–763.