HENRY
Hydraulic Engineering Repository

Ein Service der Bundesanstalt für Wasserbau

Article, Published Version

Chen, Cheng; Lin, Yuqing; Feng, Tao; Chen, Qiuwen

Data assimilation and deep learning in lake algal global bloom forecasting

Hydrolink

Verfügbar unter/Available at: https://hdl.handle.net/20.500.11970/109570

Vorgeschlagene Zitierweise/Suggested citation:
Chen, Cheng; Lin, Yuqing; Feng, Tao; Chen, Qiuwen (2022): Data assimilation and deep learning in lake algal global bloom forecasting. In: Hydrolink 2022/2. Madrid: International Association for Hydro-Environment Engineering and Research (IAHR). S. 56-58. https://www.iahr.org/library/download-paper-file?code=Zve5b36xip.

Standardnutzungsbedingungen/Terms of Use:

Die Dokumente in HENRY stehen unter der Creative Commons Lizenz CC BY 4.0, sofern keine abweichenden Nutzungsbedingungen getroffen wurden. Damit ist sowohl die kommerzielle Nutzung als auch das Teilen, die Weiterbearbeitung und Speicherung erlaubt. Das Verwenden und das Bearbeiten stehen unter der Bedingung der Namensnennung. Im Einzelfall kann eine restriktivere Lizenz gelten; dann gelten abweichend von den obigen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Documents in HENRY are made available under the Creative Commons License CC BY 4.0, if no other license is applicable. Under CC BY 4.0 commercial use and sharing, remixing, transforming, and building upon the material of the work is permitted. In some cases a different, more restrictive license may apply; if applicable the terms of the restrictive license will be binding.

Data assimilation and deep learning in lake algal bloom forecasting

By Cheng Chen, Yuqing Lin, Tao Feng and Qiuwen Chen

Due to the combined impact of human activity and the effects of climate change, harmful algal blooms (HABs) have become a serious socio-ecological problem for sustaining water and food security in many parts of the world. Oxygen depletion from the decay of biomass and higher organisms are aggravating global water shortages and can cause serious ecological disasters, that are often detrimental to food security. The outbreak of HABs in Lake Taihu, about 80 km west of Shanghai, in 2007 caused a drinking water crisis and impacted on the health of millions of people in the nearby Wuxi city. Although eutrophication control strategies have already been in place for decades, the challenge of avoiding HABs remains largely unresolved. Forecasting and early warning of potential HABs have therefore become essential for risk management in eutrophic lakes, aimed at providing adequate water security. This article shares some of the techniques that have been developed for this purpose.

A variety of data-driven and process-driven models is available to forecast the occurrence of HABs in lakes[1]. The performance of these models highly depends on the spatial-temporal resolution of available measured data. The process-based models have the potential to track the entire development of a bloom event, allowing early warning of timing, position, magnitude, and duration[1]. However, the intensive computation time and complex configuration as well as adequate calibration, limit their applicability.

Data acquisition technologies, online monitoring and remote sensing are nowadays being widely used for HAB management, although it may be difficult to obtain high-resolution and high frequency measurements. Since both model predictions and multi-source observations have their advantages and shortcomings, it remains a hot topic and challenging task to effectively synergize different technologies to improve the capability of forecasting HABs, thereby securing water and food supply in the region.

The ecohydraulics research group at the Nanjing Hydraulic Research Institute (NHRI) has recently developed a robust system for HABs prediction, by fusing data-driven and process-based models as well as assimilating multi-sourced data (**Figure 1**).

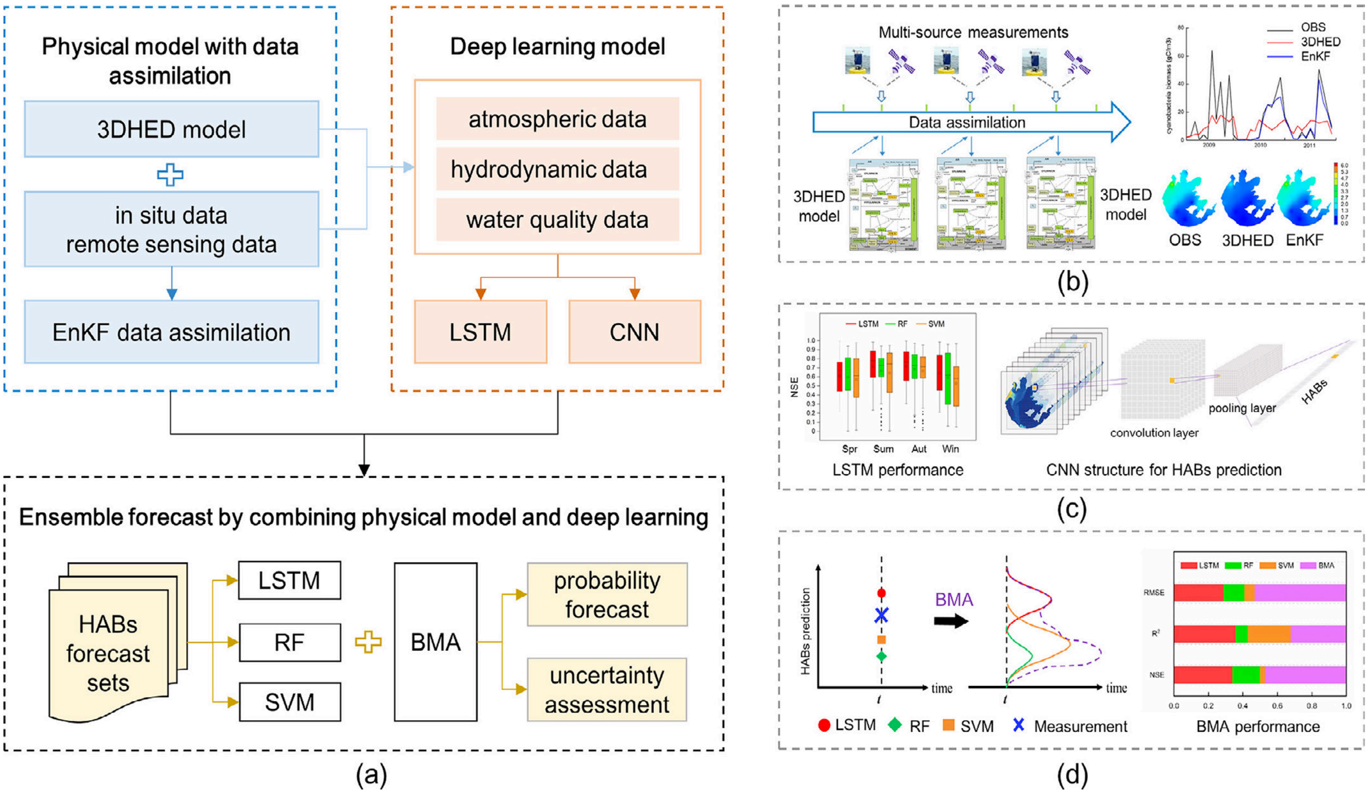

Figure 1 | (a) flowchart of data assimilation and deep learning in lake algal bloom forecasting; (b) schematic diagrams and results of physical modelling with data assimilation; (c) deep learning model; (d) ensemble forecast by combining physical models and deep learning.

Physical-based modelling with data assimilation

A physical-based three-dimensional hydro-ecological model (3DHED) was developed to predict cyanobacteria biomass by coupling the models SALMO (Simulation by means of an Analytical Lake Model) and SELFE (Semi-implicit Eulerian-Lagrangian Finite Element model). Such a combination avoids solving the complex 3D partial differential equations for both mass and momentum conservation and thereby improves computational efficiency. Moreover, the 3DHED model can distinguish different functional groups of algae. Based on careful calibration and validation, the 3DHED model is able to adequately simulate the spatial-temporal dynamics of DO, PO_4-P, NO_3-N and green algae concentrations. However, the cyanobacteria biomass, especially the peak values, were not well captured in the initial application of the model.

Data assimilation is one of the best ways to improve the simulation accuracy of physical modelling by merging measured data into model predictions. Multi-source data (i.e., in situ observations and remote sensing data) are assimilated into 3DHED to update the simulated cyanobacterial biomass (**Figure 1b**). Assimilating *in situ* measurements not only better simulates the dynamics of cyanobacterial biomass, but also better captures the peak values. Using multi-source data assimilation further improves the simulation accuracy of the model with respect to the spatial-temporal distribution of cyanobacteria. However, continuously updating parameters from different data sources could trigger oscillations in the model predictions, leading to occasionally unexpected model performance. These problems can be remedied by filtering and fusing data to obtain high-quality data sources[2, 3].

Deep learning model

LSTM (long short-term memory) is an Artificial RNN (regression neural network) architecture frequently used in the field of deep learning. Compared with ordinary machine learning models, LSTM can take into account the influence of historical information on current state variables. Chlorophyll-a concentrations are usually selected as the indicator for predicting HABs due to their close relationship to the abundance and biomass of aquatic phytoplankton[2]. The data dislocation processing method is used to establish the relationship between chlorophyll-a before and after the moment, and the known monitoring data at the current time is substituted into the established relationship to obtain short-term forecast predictions in the future. LSTM usually has a better performance when compared to the Random Forest (RF) and Support Vector Machine (SVM) learning models (**Figure 1c**).

CNN (convolution neural network) models mainly include a convolution layer and a pooling layer (**Figure 1c**). The convolution layers induce the input's internal features by performing dot-product multiplication of the input vector and the learnable weights and biases. The pooling layers, usually located between two convolution layers, are then used to extract distinctive spatial dimension to reduce the number of parameters and the amount of computation in the network. CNN models utilize the kernel matrices of convolutional layers to capture the rich features of the input data, leading to the substantial consideration of both the spatial and spectral information of the multi-dimensional image. The physical-based model can provide input maps of atmospheric, hydrodynamic, and water quality variables. The in situ observations or remote sensing data can then be used as the output true value[4]. A point-centred approach is used for CNN regression, in which a small image window patch scans the in situ observations or whole remote sensing image pixel by pixel to generate the training datasets.

Ensemble forecast by combining physical model and deep learning

The physically-based model could provide long-term and relatively high frequency chlorophyll-a data. NHRI have developed three machine learning-based short-term forecast sets of HABs by using the results from the physically-based model. In addition, considering the uncertainty of single model forecast results, a multi-model ensemble forecast method for HABs was established based on Bayesian Model Averaging (BMA). The overall performances of the four models are in the order BMA > LSTM > RF > SVM (**Figure 1d**). In general, a BMA multi-model ensemble forecast can effectively integrate the forecast results for a single model. Meanwhile, the uncertainty of forecast results can also be quantified.

Practical implications

A modelling approach can well simulate the hydrodynamics, water quality and algae dynamics of Taihu Lake, but the predicted results of peak values of cyanobacterial biomass are not satisfactory. By assimilating multi-source data, the accuracy of spatial-temporal patterns and peak values of cyanobacterial biomass were eminently improved. The LSTM deep learning approach takes into account the influence of historical information on defining current state variables, outperforming other conventional techniques like SVM and RF for time-series HABs prediction. A short-term Bayesian ensemble forecast method combing physical-based model results and deep learning techniques can be used to further improve model performance, providing better ways to issue advanced warnings for possible disasters, thus helping to secure adequate water and food security.

References

1 | Chen, C., Huang, J., Chen, Q., Zhang, J., Li, Z., Lin, Y. (2019). Assimilating multi-source data into a three-dimensional hydro-ecological dynamics model using Ensemble Kalman Filter. Environmental Modelling & Software, 117, 188-199.

2 | Chen, C., Chen, Q., Li, G., He, M., Dong, J., Yan, H., Wang, Z., Duan, Z. (2021). A novel multi-source data fusion method based on Bayesian inference for accurate estimation of chlorophyll-a concentration over eutrophic lakes. Environmental Modelling & Software, 141, 105057.

3 | Li, G., Chen, C., He, X., He, M., D, Y., Chen, Q. (2021). Modify of data assimilation model for lake algae dynamic model. Water Resources Protection, 37(4), 156-165.

4 | Pyo, J., Park, L. J., Pachepsky, Y., Baek, S. S., Kim, K., Cho, K. H. (2020). Using convolutional neural network for predicting cyanobacteria concentrations in river water. Water Research, 186, 116349.

Cheng Chen

Cheng Chen is a researcher in the Center for Eco-Environmental Research of Nanjing Hydraulic Research Institute. His main research interest cover environmental hydroinformatics and remote sensing of water quality. He is skilled at the combinnation of muti-source data observations and numerical modelling techniques for lake algal bloom forecasting.

Yuqing Lin

Yuqing Lin is a researcher in the Center for Eco-Environmental Research of Nanjing Hydraulic Research Institute. Her main research interests cover ecohydraulics and hydroinformatics. She has been long engaged in the fundamental research and engineering practice oriented to eco-environmental conservation for sustainable river and lake development.

Tao Feng

Tao Feng is a researcher in the Center for Eco-Environmental Research of Nanjing Hydraulic Research Institute. His main research interest is the development of Euler-Lagrangian coupled numerical model to describe the impacts of physical processes on water quality in large shallow lake and coastal archipelago.

Qiuwen Chen

Qiuwen Chen is a professor in Nanjing Hydraulic Research Institute, where he leads the Center for Eco-Environmental Research. His main interests cover ecohydraulics and environmental hydroinformatics, particularly numerical modeling of water quality, biogeochemical cycling of biogenic elements and aquatic ecosystem in rivers, lakes and reservoirs. He has bee long engaged in the fundamental research and engineering practice for eco-environmental conservation.