Environmental Science and Ecotechnology 23 (2025) 100522

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Explainable deep learning identifies patterns and drivers of freshwater harmful algal blooms

Shengyue Chen [a, b], Jinliang Huang [a, *], Jiacong Huang [c], Peng Wang [a], Changyang Sun [a], Zhenyu Zhang [a, d], Shijie Jiang [b, e]

[a] Fujian Key Laboratory of Coastal Pollution Prevention and Control, Xiamen University, Xiamen, 361102, China
[b] Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, 07745, Germany
[c] Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, 73 East Beijing Road, Nanjing, 210008, China
[d] School of Geographical Sciences, Fujian Normal University, Fuzhou, 350007, China
[e] ELLIS Unit Jena, Jena, 07745, Germany

ARTICLE INFO

Article history:
Received 15 May 2024
Received in revised form
25 December 2024
Accepted 26 December 2024

Keywords:
Harmful algal blooms
China's lakes and reservoirs
Explainable deep learning
Sensitivity analysis
Regional transferability

ABSTRACT

The escalating magnitude, frequency, and duration of harmful algal blooms (HABs) pose significant challenges to freshwater ecosystems worldwide. However, the mechanisms driving HABs remain poorly understood, in part due to the strong regional specificity of algal processes and the uneven data availability. These complexities make it difficult to generalize HAB dynamics and effectively predict their occurrence using traditional models. To address these challenges, we developed an explainable deep learning approach using long short-term memory (LSTM) models combined with explanation techniques that can capture complex patterns and provide explainable insights into key HAB drivers. We applied this approach for algal density modeling at 102 sites in China's lakes and reservoirs over three years. LSTMs effectively captured daily algal dynamics, achieving mean and maximum Nash-Sutcliffe efficiency coefficients of 0.48 and 0.95 during testing phase. Moreover, water temperature emerged as the primary driver of HABs both nationally and in over 30% of localities, with stronger water temperature sensitivity observed in mid-to low-latitudes. We also identified regional similarities that allow for the successful transferability in modeling algal dynamics. Specifically, using fine-tuned transfer learning, we improved the prediction accuracy in over 75% of poorly gauged areas. Overall, LSTM-based explainable deep learning approach effectively addresses key challenges in HAB modeling by tackling both regional specificity and data limitations. By accurately predicting algal dynamics and identifying critical drivers, this approach provides actionable insights into the mechanisms of HABs, ultimately aids in the implementation of effective mitigation measures for nationwide and regional freshwater ecosystems.

1. Introduction

The magnitude, frequency, and duration of harmful algal blooms (HABs) have escalated globally across freshwater ecosystems in recent years [1,2]. This phenomenon is particularly pronounced in regions heavily impacted by climate change and anthropological activities [3], making it one of the most serious environmental issues worldwide [4,5]. Excessive algal proliferation due to eutrophication, as evidenced by dense surface scum, can create large hypoxic zones that directly or indirectly affect aquatic ecosystems [6]. Importantly, cyanotoxins produced by HABs threaten the health of aquatic animals and humans who use surface water for drinking and recreational activities, leading to significant negative environmental and socioeconomic impacts [7]. According to Wang et al. [8], 63.1% of the 2,058 inland water bodies surveyed worldwide are eutrophic. In China, lake eutrophication and HAB events have increased in frequency over the past decade [9]. The most recent nationwide assessment showed that 27.3% of the 205 lakes and reservoirs surveyed in 2023 exhibited varying degrees of eutrophication (http://www.cnemc.cn/jcbg/). Adequate characterization of algal dynamics in lakes and reservoirs is required to address the

* Corresponding author.
E-mail address: jlhuang@xmu.edu.cn (J. Huang).

https://doi.org/10.1016/j.ese.2024.100522

growing threat of HABs to water security [10,11].

With increasing data availability and computational power in the big-data era, artificial intelligence (AI)-based machine learning (ML) and deep learning (DL) models have increasingly been used to provide predictive insights into the environmental behavior of pollutants [12,13], including HABs [9]. These data-driven models flexibly effectively capture dynamic patterns of objects in predictions and offer insights into data that complement our current understanding of the underlying mechanisms for relationships between inputs and outputs [14,15]. Among the available ML/DL models, a specific recurrent neural network (RNN) with a unique internal structure—long short-term memory (LSTM)—has become an extremely popular DL model for prediction tasks [16]. Recent studies have demonstrated the unrivaled popularity and robustness of LSTM for capturing streamflow [17], nutrient [18], and algal cell density [19] dynamics at different scales. Nevertheless, numerous DL-based applications improve prediction accuracy by minimizing discrepancies between predictions and observations, while largely overlooking the processes underpinning the predicted variables [20]. Using the first n steps of a predicted variable as the input to predict its next k steps is a common modeling strategy to time-series prediction in many studies [21,22]. Due to strong temporal autocorrelations among predicted variables, this modeling strategy can typically achieve high predictive performance while failing to learn the essential relationships with environmental factors that drive the variables, limiting mechanical understanding and the potential of applications such as scenario simulation.

Explainability has recently become a crucial aspect of DL model development, focusing on enhancing transparency and understanding by clarifying how these black-box models generate outputs from inputs [23]. In this context, explainable DL models offer useful and practical tools for understanding the complex nonlinear relationships between variables that traditional analyses often struggle to address [24]. Explainable DL models would, therefore, help identify the potential drivers of and underlie processes underpinning the temporal dynamics of HABs, which is particularly important for addressing the increasing risk of HABs. Effective characterization of the prevalence of individual mechanisms and interregional differences at larger spatial scales is vital for developing effective regional HAB control strategies [9,25]. Despite the prevalence of watershed hydrology and biogeochemistry studies [26], few researchers have explored the potential drivers of algal dynamics at large spatial scales using explainable DL modeling. Therefore, developing explainable LSTM models to characterize HAB dynamics on a large scale (e.g., national level) represents a meaningful and innovative attempt.

A significant challenge to deploying deep learning (DL) applications in lakes and reservoirs globally is the limited availability of training data, driven by the high costs of continuous, high-frequency *in-situ* algal monitoring [27,28]. A promising solution is to leverage comparable algal information from different adequately monitored areas (source domains) to support DL modeling in data-scarce areas (target domains). Transfer learning (TL)—an important ML concept and paradigm—which involves transferring and retraining models based on information from different or related source domains to improve predictions in a target domain, has recently received increasing attention in scientific research [29,30]. This technique helps alleviate the dependence of ML/DL on large amounts of target domain data and has great advantages for small sample modeling [31,32]. Recent studies have reported the broad and creative use of TL-based DL modeling in various fields, such as biomedicine [33], remote sensing [34], and energy systems [35]. In water environment field, TL-based modeling can improve local predictions in areas where water temperature (WT) [36], dissolved oxygen (DO) [37], and ammonium nitrogen (NH_3-N) [38] data are lacking. However, few researchers have evaluated the efficacy of TL for algal prediction under data limitation conditions. Peng et al. [39] demonstrated the advantages of a TL-based transformer model for predicting four water quality indicators in 120 rivers and lakes in China. Ma et al. [40] demonstrated that the transferability of TL-based LSTM models across continents could benefit streamflow simulations in data-scarce areas. It remains to be seen whether the observed similarity and transferability between regions can be extended to improve predictions of algal dynamics, which exhibit more complex biogeochemical mechanisms.

This study aims to characterize daily algal density dynamics in China's lakes and reservoirs, despite varying degrees of data scarcity based on water environmental and meteorological variables, with the goal of exploring the complex processes typically associated with HABs and proposing solutions to the current dilemma of inadequate algal monitoring. Specific research questions to be addressed include: (1) What are the key potential drivers of predictability of algal density dynamics in lakes and reservoirs, and how sensitive are algal dynamics to their changes? (2) Are there consistent or similar patterns of algal dynamics among lakes and reservoirs, and how could DL models utilize these potential common patterns to transfer knowledge from data-rich areas and improve algal density prediction in data-scarce areas? Overall, results of this study may provide promising insights into the usefulness of data-driven models for revealing algal dynamic processes and controlling HABs in lakes and reservoirs.

2. Materials and methods

2.1. Study area and datasets

In 2019, in China, there were 3,051 lakes (with surface areas larger than 1 km^2) covering approximately 73.38 $\times$ 10^3 km^2, 2,194 large reservoirs covering around 16.35 $\times$ 10^3 km^2, and numerous smaller lakes and reservoirs [41]. For this study, we compiled two national datasets: (1) a lake and reservoir water environmental dataset and (2) a meteorological dataset.

2.1.1. Lake and reservoir water environmental dataset

Since 2021, the Chinese National Environmental Monitoring Centre introduced four-hourly water environment monitoring for lakes and reservoirs across China (https://www.cnemc.cn/). By 2023, the program covered over 170 identified lakes and reservoirs. Monitored indicators, as specified in China's Environmental Quality Standards for Surface Waters (GB 3838−2002), include algal density, chlorophyll-a (Chl-a), and nine water quality variables: WT, DO, pH, electrical conductivity (EC), turbidity, permanganate (COD_{Mn}), total nitrogen (TN), total phosphorus (TP), and NH_3-N. We averaged the four-hourly data to obtain daily values for each site after removing erroneous records (e.g., algal densities below 10,000 cells L^{-1} or TN concentrations below NH_3-N levels. Additionally, we excluded outliers, defined as values below the first quartile minus 1.5 $\times$ interquartile range (IQR) or above the third quartile plus 1.5 $\times$ IQR. The compiled dataset covered January 2021−December 2023 and included 161 lakes and reservoirs with precise geographic coordinates (Fig. 1). Most of the monitoring sites were in the southern region of the Hu Huanyong Line (90.1%). Based on the HAB threshold (15 $\times$ 10^6 cells L^{-1}) proposed by Ma et al. [42], 20.5% of the sites with three-year average algal density exceeded this threshold, underscoring the pressing need for HAB management.

2.1.2. Meteorological dataset

To investigate the main meteorological conditions at each site,

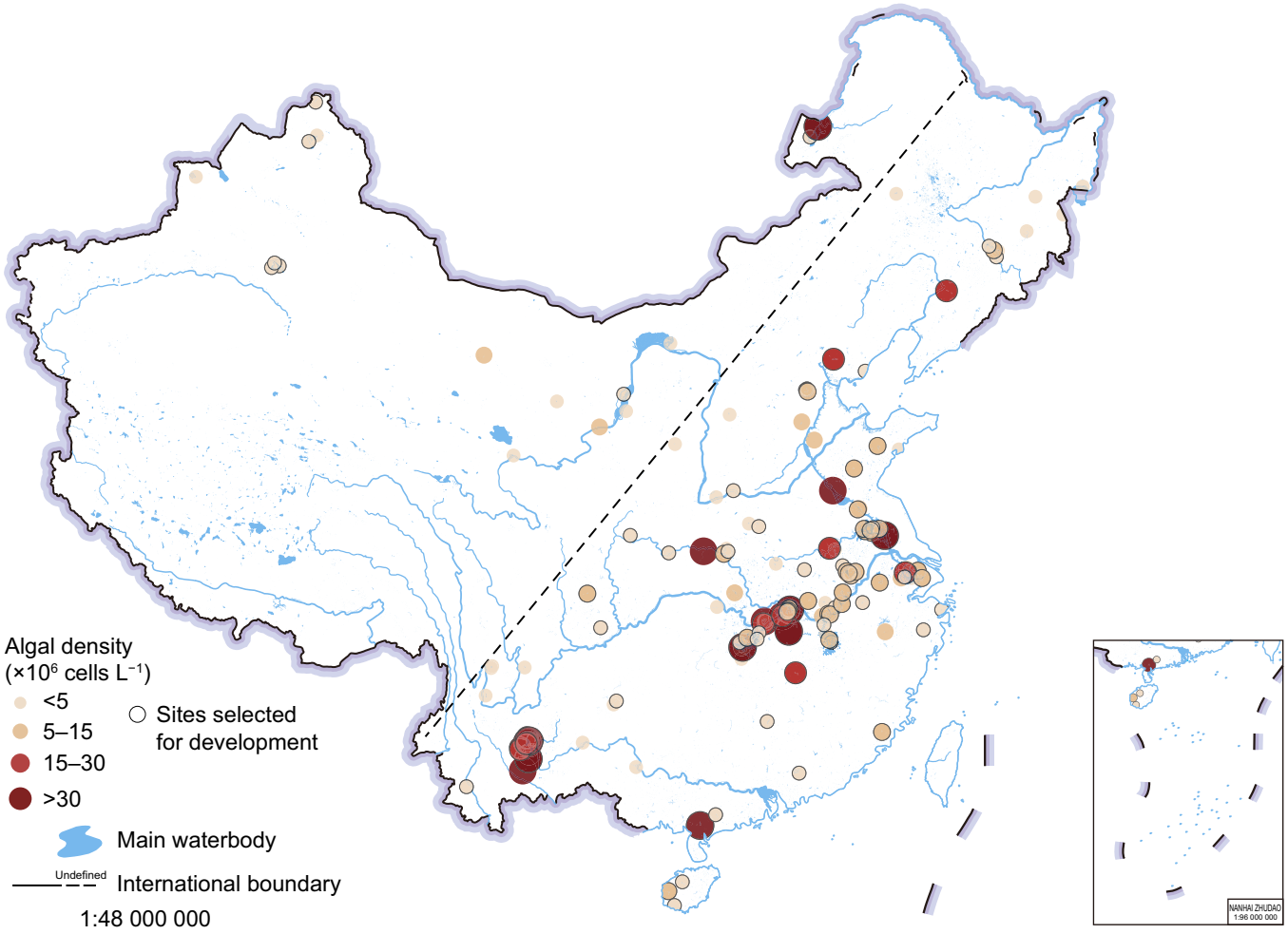

Fig. 1. Spatial patterns of 161 monitored lakes and reservoirs in China during 2021–2023. The scatters with black borders indicate sites selected for model development. The darker and larger scatter indicates higher mean algal density. The black dashed line is the Hu Huanyong Line.

we processed five variables: air pressure, evaporation, precipitation, solar radiation, and wind speed (calculated using $\sqrt{wind_u{}^2 + wind_v{}^2}$), from the European Centre of Medium-range Weather Forecasts Reanalysis 5th Generation (ERA5) product [43] with a spatial resolution of 0.25°. Due to the strong positive correlation between air temperature and WT (Pearson's correlation coefficient, $r > 0.93$), air temperature was excluded from the subsequent modeling. We characterized the meteorological conditions of each site using ERA5 grid data corresponding to the center of site.

2.2. Development of LSTM and baseline models

The DL neural network LSTM is a special form of RNN that addresses the shortcomings of traditional RNNs [44], including gradient explosion, vanishing during the computation of time-series tasks, and the inability to learn long-term dependencies [45]. The LSTM layer consists of recurrently connected memory blocks that store and transfer sequential information. Each memory block has three gates (i.e., the input, forget, and output gates) and two states (i.e., the block and hidden states) to control the details of inflow, forgetting, and memorization across time steps [46]. Such a special structure makes LSTM suitable for processing and predicting important events in a time series with long intervals and delays (i.e., streamflow). Given its generality, LSTM was used to characterize HAB dynamics. Additionally, we also selected eight conventional and widely used ML models—adaptive boosting (AdaBoost), artificial neural network (ANN), Bayesian ridge regression (Bayesian), decision tree (DT), gradient boosting decision tree (GBDT), K-nearest neighbor (KNN), random forest (RF), and support vector machine (SVM) models—as baseline models for comparative evaluation. We developed the LSTM and ANN models in this study using the PyTorch framework [47] and implemented the other seven MLs using the Scikit-Learn framework [48]. Descriptions and hyperparameter values for these ML models are presented in Supplementary Material Text S1.

From the compiled dataset, we selected a core group of 102 sites (Fig. 1) for the model performance evaluation. The number of samples available for modeling at these sites ranged from 154 to 960, with a mean of 587. Based on a previous study, we unified the structure of the LSTM developed for each site to combine an LSTM layer and a fully connected layer based on a previous study [22]. Model inputs included 11 water environment variables (WT, DO, pH, EC, turbidity, COD_{Mn}, TN, TP, N/P, NH_3-N, and Chl-a) and five meteorological variables (air pressure, evaporation, precipitation, solar radiation, and wind speed) and we defined the output as algal density. Prior to the model training, we conducted Pearson's correlation analysis on the inputs and outputs for each site. The analysis revealed significant variability in r values across sites, indicating the varying complexity of algal dynamics

(Supplementary Material Fig. S1). For each site, we divided the first 75% of the dataset into a training set to iteratively optimize the model weights and the remaining 25% into a testing set to evaluate the final performance beyond the training samples. We applied min-max normalization to the training and testing sets using the statistics of the training set (equation (1)) to eliminate unit differences and accelerate convergence. We used a randomly selected 20% of the training set as a validation set during each training epoch to evaluate loss convergence. We used a trial-and-error approach to determine the optimal combination of LSTM hyperparameters that minimized the validation set loss across all sites. Specifically, the selected hyperparameters included a learning rate of 0.001, 32 neurons, a dropout rate of 0.2, an epoch of 100, and a sequence length of 16 (i.e., we used the variables for the current day and the previous 15 days as inputs). Additionally, we employed a learning rate dynamics strategy (i.e., warmup−cosine annealing) to enhance the model convergence to an optimal solution. We saved the LSTM weights whenever the validation loss was lower than that of the previous epochs during training. We used the mean squared error (MSE, equation (2)) as the loss function to update the LSTM weights and biases during training and the Nash−Sutcliffe efficiency coefficient (NSE, equation (3)) to evaluate the model performance.

$$X_i' = \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \tag{1}$$

where X_i' and X_i are the normalized and original variable i, respectively; and $\max(X_i)$ and $\min(X_i)$ are the maximum and minimum values of variable i in the training set, respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2 \tag{2}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{3}$$

where n is the sample amount, and O_i and P_i are the observed and predicted sample i, respectively. $\overline{O}$ represents the mean observed value. An MSE close to 0 and an NSE close to 1 indicate better optimized model convergence and performance.

2.3. Model explanation and sensitivity analysis

Explaining a complex DL model aims to make the black box model more transparent and provide useful insights into the underlying mechanisms for relationships between inputs and outputs. In this study, we employed Shapley additive explanations (SHAP) to quantify the contribution of individual input features in the developed LSTM model to predicted algal density [49]. The SHAP method, grounded in game theory, calculates the Shapley value of each feature based on its marginal contribution to the model output, offering flexibility in interpreting ML/DL models [18]. Larger absolute Shapley values indicate greater predictive importance (contributions) of the corresponding features. Accordingly, we calculated the global importance (GI, %, averaged absolute Shapley value of a feature divided by the sum of all features) for individual features at each site.

$$\phi_i = \sum_{S \lhd N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{i\}) - f(S)] \tag{4}$$

$$I_i = \frac{1}{n}\sum_{j=1}^{n}\left|\phi_i^{(j)}\right| \tag{5}$$

$$GI_i = \frac{I_i}{\sum_{i=1}^{N} I_i} \times 100\% \tag{6}$$

where ϕ_i , I_i , and GI_i are the Shapley value of feature i for a given sample j, the averaged absolute Shapley value of feature i, and the global importance of feature i, respectively. N is the total number of input features, n is the sample size of feature i, S is a subset of N that does not contain feature i, and $f(S \cup \{i\}) - f(S)$ is the marginal contribution of feature i.

Subsequently, we analyzed the sensitivity of algal density dynamics to key potential drivers based on SHAP results. Specifically, for each site, we constructed simple linear regression models for the values of particular drivers and their corresponding SHAP values [50] and then evaluated the spatial variability in the linear slopes across sites and regions. We considered the slope to be an indicator of the response pattern of algal density dynamics to a potential driver, with the positive/negative sign of the slope representing the direction of the effect of that driver on algal density, and a larger absolute value indicating greater sensitivity to the driver.

2.4. Transferring knowledge across lakes and reservoirs

Building on the research question of whether consistent algal dynamics patterns exist across lakes and reservoirs, we hypothesized that at least one data-rich site (potential source domain) would exhibit patterns similar to those of the target site. This similarity would enable knowledge transfer from data-rich to data-scarce areas, enhancing DL-based algal density predictions. For each site in the core group, it was designated as the target domain, while the remaining sites served as potential source domains to cross-validate the hypothesis. We kept the last 25% of the target domain dataset as the testing set, assuming the front 20% would be the scarce available data. We also designed a new source domain selection approach that employed the available algal density record for the target domain to match the records from potential source domains during the same period. Among the matched records, we selected the site with the optimal fit to the target domain as the source domain using r as a criterion (Supplementary Material Fig. S2). The rationale for selecting r over NSE was that it would be more consistent with above hypothesis since a high r value could indicate a high degree of similarity in algal dynamics between the target and source domains, regardless of a large magnitude difference. We selected the most appropriate source domain for each target domain using the above approach, and the fitted r for the matching data from both domains ranged from 0.27 to 0.95, with a mean of 0.75. The spatial distances between the target and source domains within the core group were inversely proportional to the magnitude of the similarity ($r = -0.24$, $p < 0.05$). We pretrained the LSTM model with the complete dataset for the source domain and then transferred it to the target domain. We conducted the control experiments in three groups, as follows:

(1) Based on the fine-tuned TL strategy, we froze the weights of the LSTM layer in the pretrained model to prevent them from changing during the TL process and retrained the fully connected layer using the available data for the target domain. The fine-tuning strategy maximized the retention of knowledge learned from the source domain, reducing the

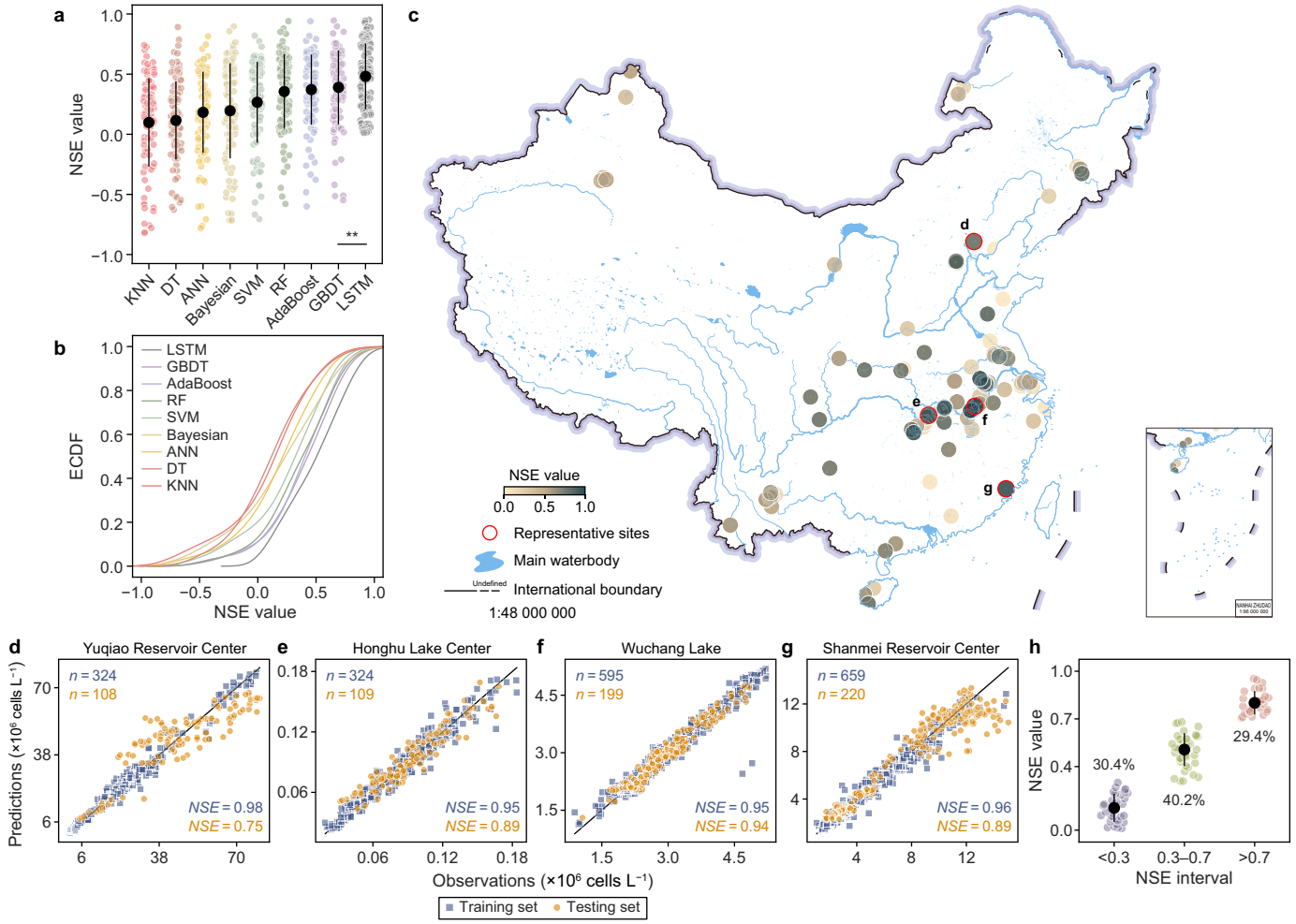

Fig. 2. Model performances. **a**–**b,** The Nash-Sutcliffe efficiency coefficient (NSE) values (**a**) and empirical cumulative distribution function (ECDF, **b**) of long short-term memory (LSTM) and eight baseline models during the testing phase. KNN, k-nearest neighbor; DT, decision tree; ANN, artificial neural network; Bayesian, bayesian ridge regression; SVM, support vector machine; RF, random forest; Adaboost, adaptive boosting; GBDT, gradient boosting decision tree. Significance (**) between LSTM and GBDT in panel **an** indicates $^{**}p < 0.01$. **c,** Spatial performance of the LSTM model for the core group. **d**–**g**, Scatter plots of observations and predictions for four representative sites: Yuqiao Reservoir Center (**d**), Honghu Lake Center (**e**), Wuchang Lake (**f**), and Shanmei Reservoir Center (**g**). The top-left and bottom-right corners of the plots for representative sites show the sample amounts and NSE values during the training and testing phases, respectively. **h,** Percentages of three NSE intervals using LSTM. The percentage represents the ratio of the corresponding interval of NSE value. The black dots in panels **a** and **h** indicate mean values, and the black whiskers indicate the standard deviation.

amount of data needed for fine-tuning while decreasing the likelihood of overfitting [40].

(2) We used the pretrained LSTM model to directly predict algal density for the target domain without retraining.

(3) We trained a new LSTM model using locally available samples.

3. Results and discussion

3.1. Performance of ML and DL models

We evaluated the predictive performance of the LSTM and baseline ML models for the core group of 102 sites (Fig. 2a and b). During the testing phase, the LSTM model showed significantly superior performance, ranking first with a mean, median, and highest NSE of 0.48, 0.52, and 0.95. The three ensemble models (the GBDT, AdaBoost, and RF models) ranked second to fourth in overall performance, with close mean NSEs of 0.39, 0.37, and 0.36, respectively. The SVM and ANN models ranked fifth and sixth, with mean NSEs of 0.26 and 0.21, respectively. The mean NSEs for the Bayesian, DT, and KNN models were below 0.2, with the KNN model performing the worst among the ML models, with a mean NSE of only 0.09. These results are consistent with previous studies showing that LSTM offers advantages over baseline ML, while ensemble models are the leading choice among conventional ML models [51,52]. Compared to the optimal performance of the LSTM model (NSE = 0.95), the highest NSEs of the baseline ML models ranged from 0.74 to 0.94, showing that conventional ML models can sometimes match the LSTM model. However, the LSTM model outperformed others at sites where baseline ML models failed, demonstrating its overall robustness. This comparison confirms the LSTM model's superior suitability for predicting algal dynamics.

The LSTM model performed effectively in lakes and reservoirs across central, northern, and southern China; more intensive well-performing sites were observed in the Hubei and Anhui provinces in the middle and lower reaches of the Yangtze River Basin (Fig. 2b). The percentages of sites where model performance was classified as excellent (NSE > 0.7), fair ($0.7 \geq$ NSE ≥ 0.3), and poor (NSE < 0.3) in the core group were 29.4%, 40.2%, and 30.4%, respectively. The mean NSE values for these categories were 0.80, 0.51, and 0.14, respectively. We found that the LSTM model's performance during the testing phase across sites showed an increasing and flattening

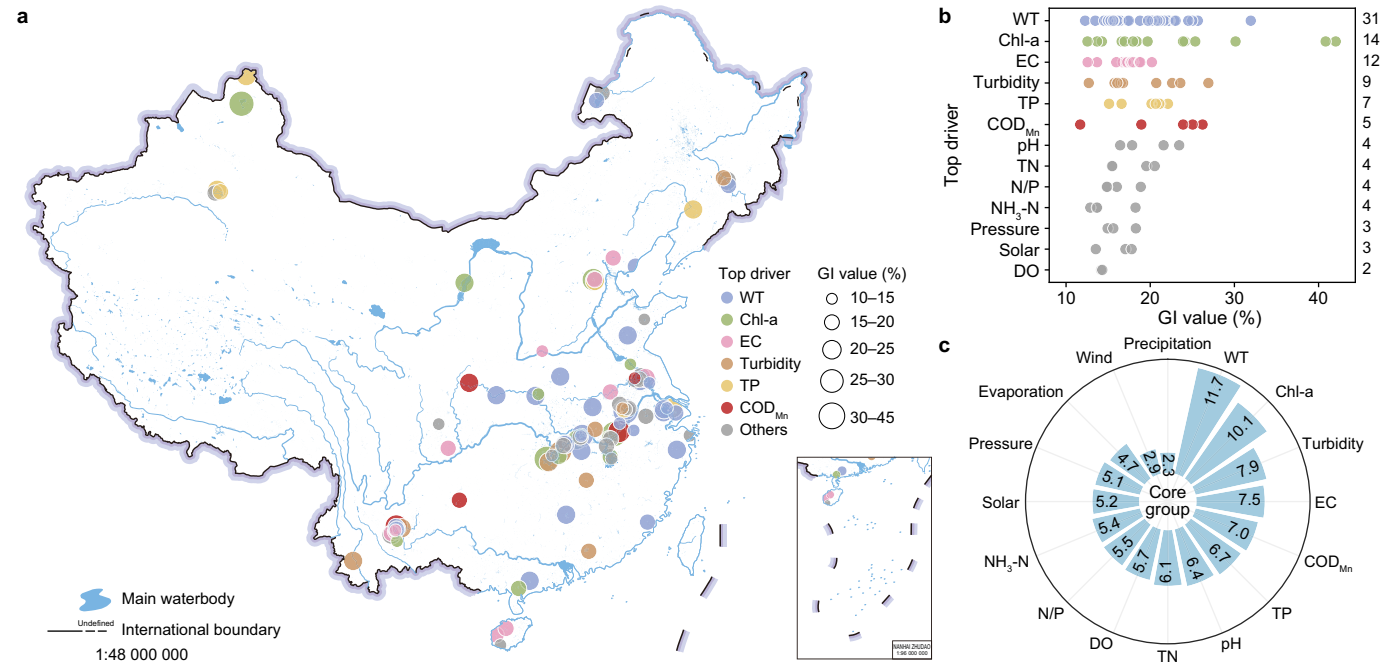

Fig. 3. The post hoc Shapley additive explanations results for input features. **a,** Spatial performance of the top driver with the highest global importance (GI, %) for the core group. **b,** Counts and GI values of the 13 top drivers. The number on the right side represents the count of the corresponding top driver, and the total count is 102. **c,** The averaged magnitude of GI value for input features in the core group. WT, water temperature; Chl-a, chlorophyll-a; EC, electrical conductivity; TP, total phosphorus; COD_{Mn}, permanganate; N/P, ratio of nitrogen−phosphorus; TN, total nitrogen; DO, dissolved oxygen; NH_3-N, ammonia. Solar, Pressure, and Wind are meteorological indicators: solar radiation, air pressure, and wind speed.

pattern as the amount of local training data increased (Supplementary Material Fig. S3a), similar to the power function curve [22]. Although many studies have indicated that larger training datasets generally improve model performance [53–55], the model performed equally well at certain sites with limited training data, such as the representative site shown in Fig. 2d–f. Differences in model performance between sites may have depended on the distributional similarity between the local training and testing sets (Supplementary Material Fig. S3b); the more similar the distributions, the better the model performance. Further comparison with other studies on HAB prediction based on ML/DLs (Supplementary Material Table S1) demonstrated the excellent potential of the LSTM model to accurately capture large-scale, high-frequency algal dynamics.

3.2. Potential drivers of algal dynamics in China's lakes and reservoirs

We evaluated the importance of the input variables in predicting algal density using the LSTM model and the SHAP approach. The most important factors influencing algal dynamics varied between lakes and reservoirs (Fig. 3a). We identified 13 variables as the factors that most influenced algal dynamics for more than one site (Fig. 3b). Among these, WT was the top potential driver at 31 sites, with GI values ranging from 0.7% to 31.9%. Except for WT, Chl-a, EC, turbidity, TP, and COD_{Mn} were the top potential drivers for at least five sites. The five most important variables in the national context were WT, Chl-a, turbidity, EC, and COD_{Mn}, which collectively accounted for 44.2% of algal density dynamics (Fig. 3c). Among these, the most important potential driver, WT, had a GI value of 11.7%, its GI values ranging from 6.3% to 17.4% across the six selected lakes (Supplementary Material, Fig. S4). The SHAP analysis, considering seasonality, revealed WT as the most critical potential driver of HABs in all seasons except spring (where Chl-a dominated), with the strongest effect occurring during the coldest winter (Supplementary Material Fig. S5). In addition, the order of magnitude of WT's importance for HABs within different climate zones was subtropical > temperate > tropical (Supplementary Material Fig. S6). The SHAP-based explanations revealed strong relationships between HABs and temperature, indirectly explaining why severe HABs, especially toxic cyanobacterial blooms, tend to occur in the summer when temperatures are high. These results align with those of previous studies reporting the adaptation and maximum growth rates of cyanobacteria under high temperature conditions [56,57]. As global warming intensifies, the frequency and severity of harmful algal blooms (HABs) are likely to rise, making HAB monitoring and control a priority [1].

Previous studies have demonstrated the relatively high importance of Chl-a, COD_{Mn}, and turbidity for HAB prediction [19,58]. All meteorological variables had lower GI values than water quality variables. The most important meteorological variable was solar radiation, with a GI value of 5.2%, whereas precipitation ranked last. Research has shown the negative and positive effects of air pressure and solar radiation on HABs [59,60], which are supported by the Pearson's correlations shown in Supplementary Material Fig. S1. Notably, because it is relatively easy to measure compared to algal density, Chl-a has been used as a surrogate variable to indirectly characterize HABs in certain studies [21,61–63]. High Chl-a is one of the main symptoms of eutrophication, which closely correlates with the presence of blooms [9,64,65]. Nevertheless, this study showed that the frequency of the top potential driver and the GI value for Chl-a were lower than for WT (Fig. 3), and the mean correlation between Chl-a and algal density was lower than that between WT and COD_{Mn} (Supplementary Material Fig. S1). A previous study also showed that Chl-a has a relatively weak linear correlation with algal density owing to the presence of an inflection point in the pattern of algal density responses to Chl-a (i.e., a clear positive correlation at lower Chl-a magnitudes with a less

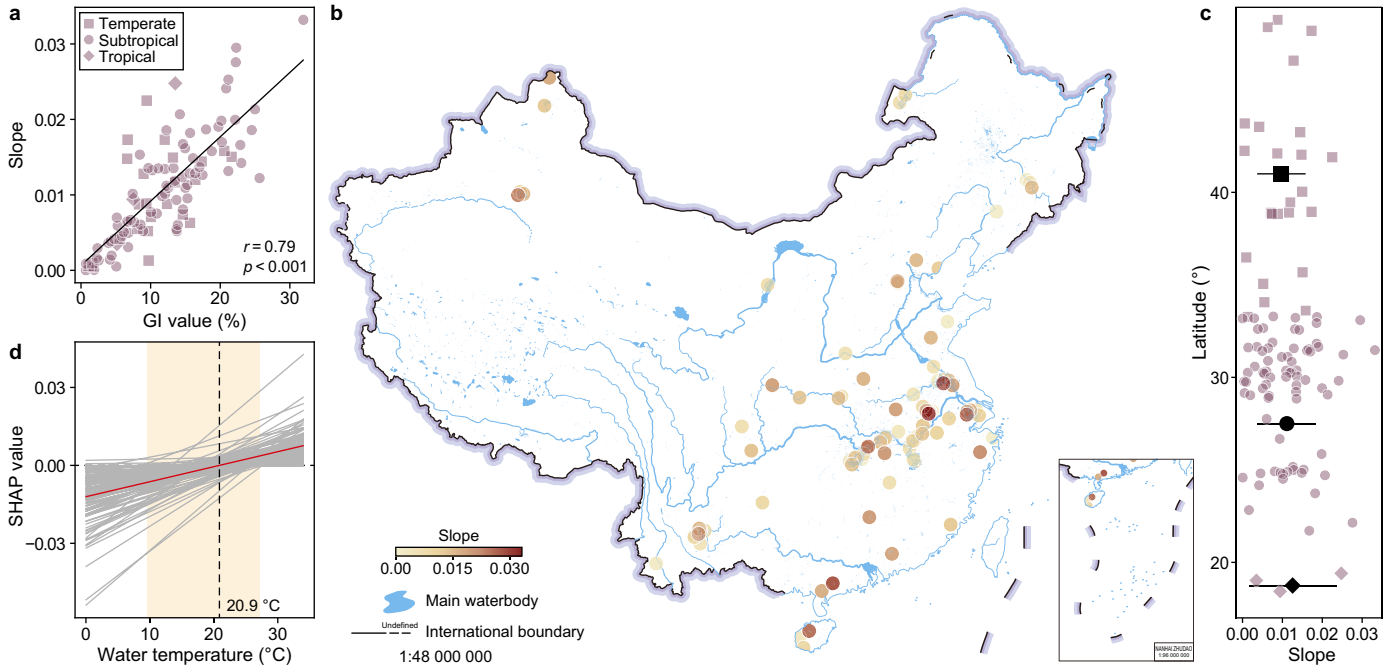

Fig. 4. Sensitivity analysis of algal density in response to water temperature (WT) for sites in the core group. **a,** Relationship of global importance (GI) values and linear slopes between WT series and corresponding Shapley additive explanations (SHAP) values. **b,** Spatial performance of the slope of the fitted line for the core group. **c,** Relationship between slope and latitude. The shapes of scatters in panel **c** have the same meanings as in panel **a**. The differently shaped black scatters indicate the mean slopes for the corresponding climate zones, and the black whiskers indicate the standard deviation. **d**, Relationship between WT and its SHAP value. The shaded interval indicates the 95% confidence intervals for inflection points, the red fitted line indicates the average level of relationships of WT-SHAP value for sites within the 95% confidence intervals, and the vertical dashed line shows the average level of the inflection point of WT.

pronounced trend of increasing algal density at higher Chl-a magnitudes) [19]. These findings suggest that Chl-a is a vital factor influencing algal density, but it does not fully reflect its magnitude. Therefore, we believe that algal density is a more direct and manifest indicator of the magnitude of HABs in freshwater [66]. Overall, LSTM combined with SHAP adequately quantified the potential water environment and meteorological drivers of algal density dynamics at the local and national scales, providing insights for the management of HABs.

3.3. Sensitivity of HABs to WT

We evaluated the sensitivities of HABs to WT variations, as WT is a key response indicator of climate warming and the most essential potential driver of HABs in the national context. The algal density dynamics responded positively to WT, but the magnitude of sensitivity (represented by the linear slope between the feature series and the corresponding SHAP series) varied within sites in the core group. GI and sensitivity were strongly positively correlated (Fig. 4a), indicating that the high importance of WT for algal dynamics is usually, but not necessarily, accompanied by a high sensitivity of algal responses to WT. Sensitivity to WT was weakly negatively correlated with latitude, with high WT sensitivity being more common at mid–low latitude sites (Fig. 4b and c). Notably, the order of WT sensitivity within different climate zones was tropical > subtropical > temperate. This pattern was not consistent with the GI pattern (Supplementary Material Fig. S6), suggesting that the HAB magnitude in low-latitude lakes and reservoirs may be more susceptible to climate warming. Additionally, the SHAP value inflection point for WT varied considerably across sites, ranging from 9.5 to 27.1 °C, with a mean of 20.9 °C. WT higher than the inflection point corresponded with a positive SHAP value (i.e., WT promotes HABs; conversely, it inhibits HABs). Sensitivity analyses based on the constraints of existing observations complemented the SHAP-based exploration of potential drivers, revealing further insights into the sensitivity of regional HABs to climate change. Although our study did not directly address the long-term effects of climate change, the lessons learned about the sensitivity of HABs to WT under current conditions remain highly relevant. Our results clearly indicate that WT is a key driver of HAB dynamics, which is critical as climate change continues to increase global temperatures. These findings can guide the development of more targeted and effective HAB management strategies. For example, water management agencies could use this information to identify which water bodies are most at risk under future climate scenarios and prioritize monitoring and mitigation efforts accordingly.

3.4. Model transferability and uncertainty in data-scarce areas

The application of fine-tuned TL proved to be an effective strategy for significantly ($p < 0.001$) improving the prediction of algal dynamics in data-scarce areas, with an NSE of 0.20 higher than that achieved by using locally limited data to train new LSTMs (Fig. 5a). However, the pretrained LSTM used for direct prediction without retraining showed similar performance to the locally trained new LSTM, with an average NSE reduction of only 0.04. This finding contrasts with previous studies on streamflow simulation in ungauged watersheds [67]. There are two potential reasons for this discrepancy: (1) the drivers of algal dynamics in different lakes and reservoirs are complex and unique (compared to streamflow, which is primarily driven by precipitation), making interregional similarity relatively difficult to capture, and (2) the source domain selection process prior to TL in this study was conducted on a national scale, further amplifying interregional differences in algal dynamics. A major advantage of TL is that it allows the LSTM to

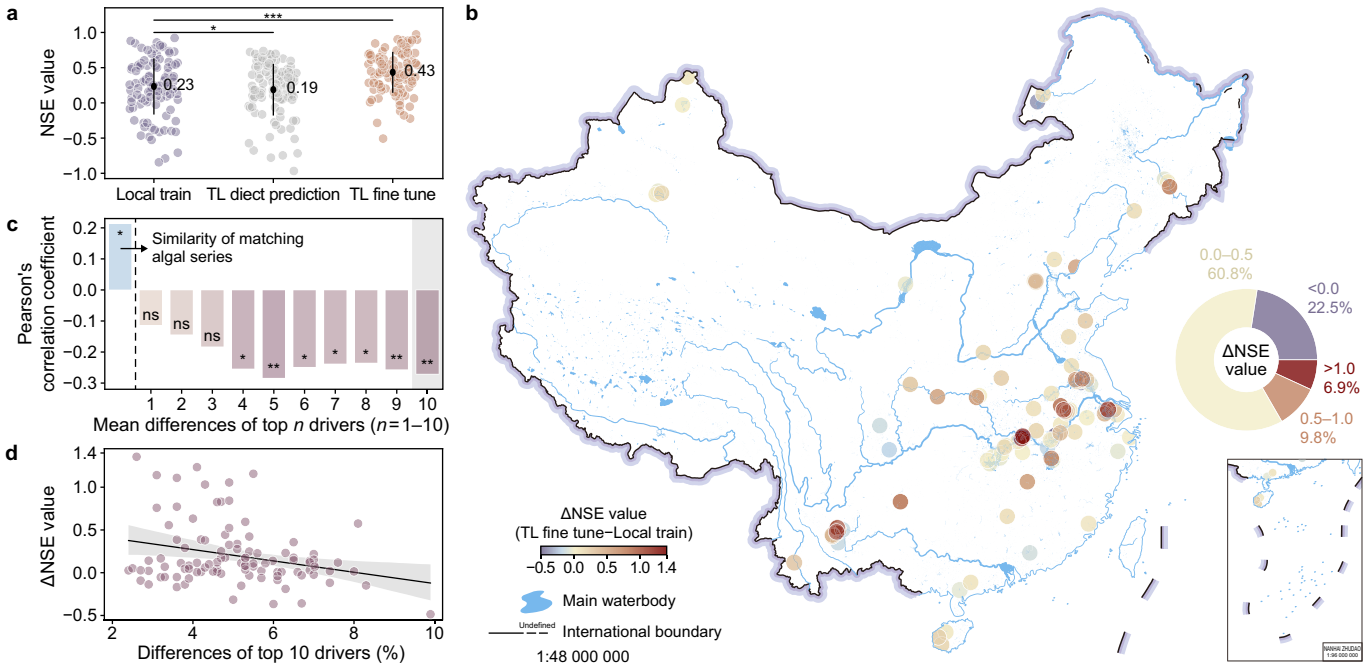

Fig. 5. Transferability and uncertainty of long short-term memory model in data-scarce areas. **a,** Nash-Sutcliffe efficiency coefficient (NSE) value of different modeling strategies during the testing phase. The larger black dots indicate the mean values and the black whiskers indicate the standard deviation. * ($p < 0.05$), *** ($p < 0.001$). **b,** Differences in NSE value between fine-tuned transfer learning (TL) and local training (ΔNSE). The pie plot shows the percentages of ΔNSE in four intervals. **c,** Pearson's correlation coefficient values of ΔNSE and metrics calculated between target and source domains. Ns ($p \geq 0.05$), * ($p < 0.05$), ** ($p < 0.01$). **d,** Relationship between ΔNSE and the representative metrics, i.e., differences of the top ten potential drivers between target and source domains. The black lines with gray intervals are the linear fitting and 95% confidence intervals.

learn basic knowledge about algal dynamics from areas with abundant information, specifically the patterns of algal density responses to water environmental and meteorological variables. However, this knowledge may differ from or even conflict with the available information from a target domain. Following local fine-tuning with a limited amount of data for the target domain, the LSTM model significantly reduced errors and uncertainties due to differences in algal dynamics between the source and target domains, making it suitable for predicting the target domain. Fine-tuned TL could practically improve the prediction accuracy (ΔNSE > 0) for algal density in the context of local data scarcity compared to training a new LSTM model (Fig. 5b). Furthermore, 60.8% of the sites exhibited increased NSE values of 0.0−0.5, while 16.7% of the sites had increased NSE values above 0.5. However, 22.5% of the sites produced varying degrees of "negative transfer" after fine-tuning and local optimization, resulting in decreased accuracy. This phenomenon has rarely been reported in water environment modeling studies compared to studies on remote sensing and energy [35,68]. Due to the very limited data available for the target domain (assumed to be 20% of the samples in the entire dataset, ranging from 31 to 192 samples, with an average of 120 in this study), even a fine-tuned TL strategy suitable for small-sample modeling could not optimize the pretrained LSTM model to a state that fully captured the dynamics of local algal density. Regardless, the benefits of a fine-tuned TL strategy for prediction in data-scarce regions warrant the attention of modelers.

We also assessed the uncertainty of TL and found that the high similarity of algal series between the source and target domains enhanced the transfer prediction of the LSTM model. However, differences in potential algal drivers between the source and target domains hindered transfer, and accounting for differences in additional potential drivers led to more pronounced negative correlations (Fig. 5c). High ΔNSEs occurred more frequently, with smaller differences in potential driver patterns (Fig. 5d). Uncertainty analysis helped explain the emergence of "negative transfer." We accounted for the similarities in algal density between source and target domains in our source domain selection approach. However, differences in environmental variables between the two domains may undermine the effectiveness of the pretrained LSTM model, making the established input-output relationships either ineffective or counterproductive. Therefore, similar input and output variables between source and target domains should be considered when selecting suitable source domains for more robust TL-based modeling. Despite a small number of conflicting results, we affirmed that there are varying degrees of similarity in algal dynamics between China's lakes and reservoirs, and fine-tuned TL can effectively benefit prediction in data-scarce areas.

4. Conclusions and prospects

Characterizing HABs for the maintenance of aquatic ecosystem functioning is essential but challenging. In this study, we developed an explainable LSTM model for lakes and reservoirs with varying degrees of data scarcity in a national context to comprehensively identify the complex relationships between algal densities and multiple environmental factors. Explainable LSTM quantified the importance and sensitivity of potential environmental drivers while providing accurate predictions. For instance, WT was the most important potential driver of HABs, although patterns regarding its importance and sensitivity were inconsistent across climate zones. We also calculated the WT inflection points on HABs based on explainable LSTM, offering specific tipping point reference data for decision-makers. In addition, explainable LSTM helped us analyze the uncertainty in transferring the model to data-scarce

areas, which provided insights for creative approaches to solving "negative transfer" due to the selection of an appropriate source domain from many potential domains and consideration of the similarities among potential driving patterns. Explainable LSTM facilitated the acquisition of evidence for the prevalence of HAB mechanisms and interregional differences, and their explanation could be further improved by incorporating hydrological attributes, such as water-level dynamics and hydraulic residence time, into the input features. Furthermore, explainable DLs can be extended to larger scales to achieve synergistic predictions of continental and even global algal dynamics and to reveal the underlying complex ecological patterns. Nevertheless, before this can be achieved, there is an urgent need to expand algal monitoring networks, especially in poorly gauged waterbodies, to meet the requirements of HAB control and modeling.

CRediT authorship contribution statement

Shengyue Chen: Writing - Original Draft, Visualization, Methodology, Conceptualization, Funding Acquisition. **Jinliang Huang:** Writing - Review & Editing, Supervision, Funding Acquisition, Resources. **Jiacong Huang:** Writing - Review & Editing, Methodology. **Peng Wang:** Software, Data Curation. **Changyang Sun:** Data Curation, Visualization. **Zhenyu Zhang:** Writing - Review & Editing. **Shijie Jiang:** Writing - Review & Editing, Methodology, Funding Acquisition, Validation.

Data availability statement

The data used in this study are available upon reasonable request.

Declaration of competing interest

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.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (Grant No. 42376225), China Scholarship Council (Grant No. 202406310083), Fieldwork Funds for graduate students of Xiamen University (Grant No. 2023FG008), and Google Climate Action Student Research Grants (China) (Grant No. PJ240067). Additional support was provided by the Carl Zeiss Foundation (Junior Research Group "Knowledge integration for spatio-temporal environmental modeling").

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2024.100522.

References

[1] J.C. Ho, A.M. Michalak, N. Pahlevan, Widespread global increase in intense lake phytoplankton blooms since the 1980s, Nature 574 (7780) (2019) 667–670.
[2] B.W. Brooks, J.M. Lazorchak, M.D.A. Howard, M.V.V. Johnson, S.L. Morton, D.A.K. Perkins, E.D. Reavie, G.I. Scott, S.A. Smith, J.A. Steevens, Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? Environ. Toxicol. Chem. 35 (1) (2016) 6–13.
[3] X. Hou, L. Feng, Y. Dai, C. Hu, L. Gibson, J. Tang, Z. Lee, Y. Wang, X. Cai, J. Liu, Global mapping reveals increase in lacustrine algal blooms over the past decade, Nat. Geosci. 15 (2) (2022) 130–134.
[4] Y. Dai, S. Yang, D. Zhao, C. Hu, W. Xu, D.M. Anderson, Y. Li, X.-P. Song, D.G. Boyce, L. Gibson, Coastal phytoplankton blooms expand and intensify in the 21st century, Nature 615 (7951) (2023) 280–284.
[5] G.M. Hallegraeff, D.M. Anderson, C. Belin, M.-Y.D. Bottein, E. Bresnan, M. Chinain, H. Enevoldsen, M. Iwataki, B. Karlson, C.H. McKenzie, Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts, Communications Earth & Environment 2 (1) (2021) 117.
[6] B.D. Turley, M. Karnauskas, M.D. Campbell, D.S. Hanisko, C.R. Kelble, Relationships between blooms of karenia brevis and hypoxia across the west Florida shelf, Harmful Algae 114 (2022) 102223.
[7] W.W. Carmichael, G.L. Boyer, Health impacts from cyanobacteria harmful algae blooms: implications for the North American Great Lakes, Harmful Algae 54 (2016) 194–212.
[8] S. Wang, J. Li, B. Zhang, E. Spyrakos, A.N. Tyler, Q. Shen, F. Zhang, T. Kuster, M.K. Lehmann, Y. Wu, D. Peng, Trophic state assessment of global inland waters using a MODIS-derived Forel-Ule index, Rem. Sens. Environ. 217 (2018) 444–460.
[9] J. Huang, Y. Zhang, G.B. Arhonditsis, J. Gao, Q. Chen, J. Peng, The magnitude and drivers of harmful algal blooms in China's lakes and reservoirs: a national-scale characterization, Water Res. 181 (2020).
[10] H. Li, C. Qin, W. He, F. Sun, P. Du, Improved predictive performance of cyanobacterial blooms using a hybrid statistical and deep-learning method, Environ. Res. Lett. 16 (12) (2021).
[11] G. Hamilton, R. McVinish, K. Mengersen, Bayesian model averaging for harmful algal bloom prediction, Ecol. Appl. 19 (7) (2009) 1805–1814.
[12] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, Prabhat, Deep learning and process understanding for data-driven Earth system science, Nature 566 (7743) (2019) 195–204.
[13] S. Chen, Z. Zhang, J. Lin, J. Huang, Machine learning-based estimation of riverine nutrient concentrations and associated uncertainties caused by sampling frequencies, PLoS One 17 (7) (2022) e0271458.
[14] X. Liu, D. Lu, A. Zhang, Q. Liu, G. Jiang, Data-driven machine learning in environmental pollution: gains and problems, Environ. Sci. Technol. 56 (4) (2022) 2124–2133.
[15] R.M. Adnan, Z. Liang, S. Heddam, M. Zounemat-Kermani, O. Kisi, B. Li, Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs, J. Hydrol. 586 (2020).
[16] Z. Xiang, J. Yan, I. Demir, A rainfall-runoff model with LSTM-based sequence-to-sequence learning, Water Resour. Res. 56 (1) (2020).
[17] D. Feng, K. Fang, C. Shen, Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, Water Resour. Res. 56 (9) (2020).
[18] R. Xiong, Y. Zheng, N. Chen, Q. Tian, W. Liu, F. Han, S. Jiang, M. Lu, Y. Zheng, Predicting dynamic riverine nitrogen export in unmonitored watersheds: leveraging insights of AI from data-rich regions, Environ. Sci. Technol. 56 (14) (2022) 10530–10542.
[19] W. Rao, X. Qian, Y. Fan, T. Liu, A soft sensor for simulating algal cell density based on dynamic response to environmental changes in a eutrophic shallow lake, Sci. Total Environ. 868 (2023).
[20] H. Cao, L. Han, L. Li, A deep learning method for cyanobacterial harmful algae blooms prediction in Taihu Lake, China, Harmful Algae 113 (2022) 102189.
[21] M. Liu, J. He, Y. Huang, T. Tang, J. Hu, X. Xiao, Algal bloom forecasting with time-frequency analysis: a hybrid deep learning approach, Water Res. 219 (2022) 118591.
[22] S. Chen, J. Huang, P. Wang, X. Tang, Z. Zhang, A coupled model to improve river water quality prediction towards addressing non-stationarity and data limitation, Water Res. 248 (2024).
[23] S.W. Fleming, V.V. Vesselinov, A.G. Goodbody, Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach, J. Hydrol. 597 (2021) 126327.
[24] S. Jiang, Y. Zheng, C. Wang, V. Babovic, Uncovering flooding mechanisms across the contiguous United States through interpretive deep learning on representative catchments, Water Resour. Res. 58 (1) (2022) e2021WR030185.
[25] J.C. Ho, A.M. Michalak, Exploring temperature and precipitation impacts on harmful algal blooms across continental US lakes, Limnol. Oceanogr. 65 (5) (2020) 992–1009.
[26] S. Jiang, E. Bevacqua, J. Zscheischler, River flooding mechanisms and their changes in Europe revealed by explainable machine learning, Hydrol. Earth Syst. Sci. 26 (24) (2022) 6339–6359.
[27] R. Marce, G. George, P. Buscarinu, M. Deidda, J. Dunalska, E. de Eyto, G. Flaim, H.P. Grossart, V. Istvanovics, M. Lenhardt, E. Moreno-Ostos, B. Obrador, I. Ostrovsky, D.C. Pierson, J. Potuzak, S. Poikane, K. Rinke, S. Rodriguez-Mozaz, P.A. Staehr, K. Sumberova, G. Waajen, G.A. Weyhenmeyer, K.C. Weathers, M. Zion, B.W. Ibelings, E. Jennings, Automatic high frequency monitoring for improved lake and reservoir management, Environ. Sci. Technol. 50 (20) (2016) 10780–10794.
[28] A.B.G. Janssen, J.H. Janse, A.H.W. Beusen, M. Chang, J.A. Harrison, I. Huttunen, X. Kong, J. Rost, S. Teurlincx, T.A. Troost, How to model algal blooms in any lake on earth, Curr. Opin. Environ. Sustain. 36 (2019) 1–10.
[29] Z. Chen, H. Xu, P. Jiang, S. Yu, G. Lin, I. Bychkov, A. Hmelnov, G. Ruzhnikov, N. Zhu, Z. Liu, A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system, J. Hydrol. 602 (2021).

[30] G.K. Saha, F. Rahmani, C. Shen, L. Li, R. Cibin, A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds, Sci. Total Environ. 878 (2023) 162930.
[31] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[32] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (1) (2021) 43–76.
[33] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021).
[34] Y. Ma, S. Chen, S. Ermon, D.B. Lobell, Transfer learning in environmental remote sensing, Rem. Sens. Environ. 301 (2024) 113924.
[35] N. Wei, C. Yin, L. Yin, J. Tan, J. Liu, S. Wang, W. Qiao, F. Zeng, Short-term load forecasting based on WM algorithm and transfer learning model, Appl. Energy 353 (2024).
[36] J.D. Willard, J.S. Read, A.P. Appling, S.K. Oliver, X. Jia, V. Kumar, Predicting water temperature dynamics of unmonitored lakes with meta-transfer learning, Water Resour. Res. 57 (7) (2021).
[37] N. Zhu, X. Ji, J. Tan, Y. Jiang, Y. Guo, Prediction of dissolved oxygen concentration in aquatic systems based on transfer learning, Comput. Electron. Agric. 180 (2021).
[38] Y. Zhou, Real-time probabilistic forecasting of river water quality under data missing situation: deep learning plus post-processing techniques, J. Hydrol. 589 (2020).
[39] L. Peng, H. Wu, M. Gao, H. Yi, Q. Xiong, L. Yang, S. Cheng, TLT: recurrent fine-tuning transfer learning for water quality long-term prediction, Water Res. 225 (2022) 119171.
[40] K. Ma, D. Feng, K. Lawson, W.P. Tsai, C. Liang, X. Huang, A. Sharma, C. Shen, Transferring hydrologic data across continents – leveraging data-rich regions to improve hydrologic prediction in data-sparse regions, Water Resour. Res. 57 (5) (2021).
[41] X. Wang, X. Xiao, Y. Qin, J. Dong, J. Wu, B. Li, Improved maps of surface water bodies, large dams, reservoirs, and lakes in China, Earth Syst. Sci. Data 14 (8) (2022) 3757–3771.
[42] J.R. Ma, J.M. Deng, B.Q. Qin, S.X. Long, Progress and prospects on cyanobacteria bloom-forming mechanism in lakes, Acta Ecol. Sin. 33 (2013) 3020–3030.
[43] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, ERA5 hourly data on single levels from 1979 to present, Copernicus climate change service (c3s) climate data store (cds) 10 (10.24381) (2018).
[44] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[45] C. Shen, E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer, Z. Fang, K. Fang, D. Li, X. Li, W.-P. Tsai, HESS Opinions: incubating deep-learning-powered hydrologic science advances as a community, Hydrol. Earth Syst. Sci. 22 (11) (2018) 5639–5656.
[46] A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, Phys. Nonlinear Phenom. 404 (2020).
[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga Pytorch, An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[49] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[50] W. Li, M. Migliavacca, M. Forkel, J.M.C. Denissen, M. Reichstein, H. Yang, G. Duveiller, U. Weber, R. Orth, Widespread increasing vegetation sensitivity to soil moisture, Nat. Commun. 13 (1) (2022) 3959.
[51] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, H. Ren, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, Water Res. 171 (2020) 115454.
[52] F. Farooq, W. Ahmed, A. Akbar, F. Aslam, R. Alyousef, Predictive modeling for sustainable high-performance concrete from industrial wastes: a comparison and optimization of models using ensemble learners, J. Clean. Prod. 292 (2021) 126032.
[53] J.S. Read, X. Jia, J. Willard, A.P. Appling, J.A. Zwart, S.K. Oliver, A. Karpatne, G.J.A. Hansen, P.C. Hanson, W. Watkins, M. Steinbach, V. Kumar, Process-guided deep learning predictions of lake water temperature, Water Resour. Res. 55 (11) (2019) 9173–9190.
[54] M.K. Thomas, S. Fontana, M. Reyes, M. Kehoe, F. Pomati, The predictability of a lake phytoplankton community, over time-scales of hours to years, Ecol. Lett. 21 (5) (2018) 619–628.
[55] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, Chem. Mater. 29 (12) (2017) 5090–5103.
[56] S.A. Wood, H. Borges, J. Puddick, L. Biessy, J. Atalah, I. Hawes, D.R. Dietrich, D.P. Hamilton, Contrasting cyanobacterial communities and microcystin concentrations in summers with extreme weather events: insights into potential effects of climate change, Hydrobiologia 785 (1) (2016) 71–89.
[57] Y. Wang, L. Feng, X. Hou, Algal blooms in lakes in China over the past two decades: patterns, trends, and drivers, Water Resour. Res. 59 (10) (2023).
[58] K. Shan, T. Ouyang, X. Wang, H. Yang, B. Zhou, Z. Wu, M. Shang, Temporal prediction of algal parameters in Three Gorges Reservoir based on highly time-resolved monitoring and long short-term memory network, J. Hydrol. 605 (2022).
[59] Y. Zhang, K. Shi, J. Liu, J. Deng, B. Qin, G. Zhu, Y. Zhou, Meteorological and hydrological conditions driving the formation and disappearance of black blooms, an ecological disaster phenomena of eutrophication and algal blooms, Sci. Total Environ. 569 (2016) 1517–1529.
[60] J. León-Muñoz, M.A. Urbina, R. Garreaud, J.L. Iriarte, Hydroclimatic conditions trigger record harmful algal bloom in western Patagonia (summer 2016), Sci. Rep. 8 (1) (2018) 1330.
[61] J. Shen, Q. Qin, Y. Wang, M. Sisson, A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading, Ecol. Model. 398 (2019) 44–54.
[62] L. Zheng, H. Wang, C. Liu, S. Zhang, A. Ding, E. Xie, J. Li, S. Wang, Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models, J. Environ. Manag. 295 (2021) 113060.
[63] M. Alizamir, S. Heddam, S. Kim, A.D. Mehr, On the implementation of a novel data-intelligence model based on extreme learning machine optimized by bat algorithm for estimating daily chlorophyll-a concentration: case studies of river and lake in USA, J. Clean. Prod. 285 (2021).
[64] Y.S. Kwon, J. Pyo, Y.-H. Kwon, H. Duan, K.H. Cho, Y. Park, Drone-based hyperspectral remote sensing of cyanobacteria using vertical cumulative pigment concentration in a deep reservoir, Rem. Sens. Environ. 236 (2020) 111517.
[65] H. Li, X. Li, D. Song, J. Nie, S. Liang, Prediction on daily spatial distribution of chlorophyll-a in coastal seas using a synthetic method of remote sensing, machine learning and numerical modeling, Sci. Total Environ. 910 (2024) 168642.
[66] L. Lai, Y. Zhang, Z. Cao, Z. Liu, Q. Yang, Algal biomass mapping of eutrophic lakes using a machine learning approach with MODIS images, Sci. Total Environ. 880 (2023) 163357.
[67] S. Chen, J. Huang, J.-C. Huang, Improving daily streamflow simulations for data-scarce watersheds using the coupled SWAT-LSTM approach, J. Hydrol. 622 (2023).
[68] Y. Ma, Z. Yang, Z. Zhang, Multisource maximum predictor discrepancy for unsupervised domain adaptation on corn yield prediction, IEEE Trans. Geosci. Rem. Sens. 61 (2023) 1–15.