Improved predictive performance of cyanobacterial blooms using a hybrid statistical and deep-learning method

Hu Li, Chengxin Qin, Weiqi He, Fu Sun and Pengfei Du

1 State Key Laboratory of Environmental Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, People's Republic of China
2 Research Institute of Environmental Innovation (Suzhou), Tsinghua University, Suzhou 215163, People's Republic of China
* Author to whom any correspondence should be addressed.
E-mail: dupf@tsinghua.edu.cn

Keywords: cyanobacterial blooms, time-series prediction, ARIMA, LSTM, hybrid model, predictive performance

Abstract
Cyanobacterial harmful algal blooms (CyanoHABs) threaten ecosystem functioning and human health at both regional and global levels, and this threat is likely to become more frequent and severe under climate change. Predictive information can help local water managers to alleviate or manage the adverse effects posed by CyanoHABs. Previous works have led to various approaches for predicting cyanobacteria abundance by feeding various environmental variables into statistical models or neural networks. However, these models alone may have limited predictive performance owing to their inability to capture extreme situations. In this paper, we consider the possibility of a hybrid approach that leverages the merits of these methods by integrating a statistical model with a deep-learning model. In particular, the autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) were used in tandem to better capture temporal patterns of highly dynamic observations. Results show that the proposed ARIMA-LSTM model exhibited the promising potential to outperform the state-of-the-art baseline models for CyanoHAB prediction in highly variable time-series observations, characterized by nonstationarity and imbalance. The predictive error of the mean absolute error and root mean square error, compared with the best baseline model, were largely reduced by 12.4% and 15.5%, respectively. This study demonstrates the potential for the hybrid model to assist in cyanobacterial risk assessment and management, especially in shallow and eutrophic waters.

1. Introduction
Cyanobacterial harmful algae blooms (CyanoHABs) are excessive cyanobacterial proliferations that have adverse effects on water quality, ecological function and human health across the freshwater-marine continuum (McGowan et al 2017, Paerl et al 2018). Their occurrence poses a serious threat to inland aquatic systems, and this threat is deemed to be intensified by climate change (O’Reilly et al 2015). Obviously, timely monitoring and prediction of CyanoHABs would be of critical value to the development of appropriate mitigation or control strategies (Hamilton et al 2009). Traditionally, the study of CyanoHABs relies heavily on shipborne surveys. However, this method is time-consuming and labor-intensive and, more importantly, cruise observations are often spatially and temporally sparse. The inadequate sampling issue is now alleviated by the technological advancements of automatic high-frequency sensors, which could provide a wealth of real-time information on variations in water quality. These high-frequency time-series observations are critical for capturing and understanding temporal patterns of the ecological status (Marcé et al 2016). The high-frequency observations, in conjunction with time-series forecasting methods, provide the possibility to produce the short-term prediction of CyanoHABs that is needed to control or alleviate their environmental risks.

Although the prediction of CyanoHABs is an important task, this task remains challenging due to the strong stochastic nature and high-level variability of cyanobacteria growth. Various statistical methods have been developed for the prediction of CyanoHABs. For example, simple and multiple
linear regression models were developed to evaluate the risk of CyanoHABs with bloom-related environmental parameters, such as pH, water temperature and water quality (Fornarelli et al 2013, Brasil et al 2016). More recent approaches introduced a hybrid evolutionary algorithm, which could produce a spatially-explicit prediction of cyanobacteria growth with good accuracy regarding the timing and magnitudes of CyanoHABs (Recknagel et al 2014, 2017, Cao et al 2016). The well-known autoregressive integrated moving average (ARIMA) method was also developed to predict daily chlorophyll-a concentrations, due to its requirement for fewer data and simple structure (Chen et al 2015). Apart from these statistical methods, deep learning methods have gained renewed popularity in the prediction of CyanoHABs, fueled by the advancement of data volume and computing capacity. Recknagel et al (1997) first applied a fully connected multilayer feed-forward network to modelling and predicting the succession, timing and magnitudes of algal blooms. This method is further extended by introducing sensitivity analysis (Wei et al 2001), generalized regression (Teles et al 2006) or a self-organizing map (Ahn et al 2011). Recently, attempts have been made to improve the predictive performance of CyanoHABs by better capturing long-term temporal dependency, for example, by using long short-term memory (LSTM) (Lee and Lee 2018), or other sequentially structured neural networks (Pyo et al 2020).

However, these aforementioned models alone may have limited predictive capacity due to their inadequacy in capturing the high levels of temporal variability. This may be particularly true for prediction of CyanoHABs, as phytoplankton dynamics are often characterized by uncertainty and stochasticity (McGowan et al 2017). Phytoplankton can respond very rapidly to the changing environment, such as sudden changes in nutrient levels, temperature or flows (Bowes et al 2016), and their populations can fluctuate over an order of magnitude within days (Fang et al 2019). In addition, the monitoring data of CyanoHABs are often incomplete and imbalanced (Choi et al 2019), which strengthens the variability of the time-series observations. Both theoretical and empirical findings have demonstrated that integration of different models has enormous potential to improve their predictive power, especially when the dataset exhibits high levels of variability (Zhang 2003, Khashei and Bijari 2011). Zhang (2003) proposed a hybrid ARIMA and Artificial Neural Network (ANN) model to better capture temporal patterns in the time-series data, and this methodology has attracted considerable attention in time-series prediction. The benefits of the hybrid method for prediction of CyanoHABs have also been illustrated in several recent studies. For example, Xiao et al (2017) proposed a single-parameter method combining wavelet analysis with neural networks, which achieved better predictive accuracy. Similarly, Qin et al (2017) introduced a hybrid model of ARIMA-DBN (Deep Belief Network) to predict the red tide occurrence, and their method showed good predictive performance and robustness.

Obviously, although the hybrid approach has been proven to be an effective way to improve predictive performance, it is rarely used in the study of CyanoHAB prediction thus far. Motivated by the fact that phytoplankton dynamics are often highly variable and stochastic in nature (Chen et al 2015, Rousso et al 2020), in this study, we demonstrate the utility of a hybrid ARIMA-LSTM model in short-term prediction of CyanoHABs. We improve upon previous modelling frameworks by using LSTM to identify nonlinear temporal patterns. LSTM can better capture long-term temporal dependency and does not undergo gradient vanishing issues; it is thus expected that the proposed ARIMA-LSTM model could potentially outperform the state-of-the-art models. To this end, the ARIMA model was first used to decompose the time-series observations to subseries, and LSTM was subsequently implemented to capture nonlinear relationships and temporal dependencies. Our model showed modest but incremental predictive performance over two state-of-the-art baseline models, suggesting that the proposed ARIMA-LSTM model could be a promising tool to assist in cyanobacterial risk management.

2. Methodology

2.1. Study area
Taihu Lake is a large (surface area 2300 km$^2$), shallow (average depth 1.9 m) and eutrophic waterbody within the Yangtze River drainage basin, which has been plagued by cyanobacterial blooms since the 1980s (Qin et al 2015). Taihu Lake experiences a subtropical humid climate, with annual average precipitation ranging between 1000 and 1400 mm. Due to its large and polymeric nature, the lake has a relatively long hydraulic retention time of about 309 d (Huang et al 2012, Zhang et al 2012). Taihu Lake is surrounded by a dense network of waterways, with more than 219 rivers or canals discharged into the lake. The lake is a vital drinking water supply for surrounding cities, and acts as a natural regulator of lacustrine ecosystems and regional climate. The lake faces myriads of challenges, however, many of which are related to eutrophication and periodic cyanobacterial blooms. Although many efforts have been made to reduce nutrient loadings since 2007, and water quality has generally shown an improved trend, cyanobacterial blooms recur annually and may be promoted by more favorable conditions for bloom formation within the context of profound changes in the physical environment (Zhang et al 2018).
2.2. Data collection
Meteorological and limnological data were acquired with an *in situ* automated monitoring platform at a sampling site anchored at Gonghu Bay (figure 1). From 2017 to 2018, water quality parameters were recorded at 30 min intervals. The multi-parameter sensor measured a standard suite of parameters, including dissolved oxygen, pH, total nitrogen, total phosphorus, chlorophyll-a and algae density. Algae density was used as a proxy measurement for phytoplankton biomass, and was averaged every 4 h for model development and validation. Missing values were imputed using the ‘forward-fill’ method (propagating the last valid values forward along an axis) using the Python Pandas library. The primary characteristics of the limnological parameters are summarized in table 1. Time-series observations from January 2017 to October 2018 were reserved for model fitting, and the rest of the data were left to examine the model performance on unseen samples.

2.3. Methods
2.3.1. The hybrid ARIMA-LSTM model
The ARIMA model was first proposed by Box (1976). ARIMA can be applied to nonstationary data by performing a differencing process, which transforms the nonstationary time series into a stationary one. The ARIMA model is essentially a linear regression model, in which a time series is characterized by a combination of three aspects: an autoregressive term, an integrated term and a moving average term. In contrast, the LSTM network is a type of recurrent neural network (RNN), and was first introduced by Hochreiter and Schmidhuber (1997). The motivation for developing LSTM was to address the vanishing gradient problems that emerge from RNN when learning long-range dependency (Lee and Lee 2018). In contrast to traditional neural networks, LSTM has the ability to capture the dynamics of temporal dependencies within variables due to its memory cell properties.

The environmental factors and processes contributing to cyanobacterial blooms are not merely manifold but involute, hampering our ability to understand the bloom-triggering mechanism. Previous studies (Xiao *et al* 2017, Rousso *et al* 2020) showed that phytoplankton dynamics are nonlinear and non-stationary in nature, due to the myriad interactions between physical, chemical and biological parameters. As indicated by Qin *et al* (2017), this nonlinear pattern needs to be ascertained because traditional linear models like ARIMA do not exhibit satisfactory predictive ability. Neural networks, in contrast, are able to represent nonlinearity because they are characterized by a class of approximation functions. Hence, as stated previously, a hybrid approach may be an appropriate alternative to produce better forecast skills than either model can deliver separately.

In this work, a hybrid model is proposed to improve the prediction of cyanobacterial blooms. In particular, the model incorporates a linear statistical method of ARIMA and a deep-learning method of...
Table 1. Summary statistics of the parameters measured at the monitoring site in Gonghu Bay from 2017 to 2018.

| Variables | WT (°C) | TN (mg l⁻¹) | TP (mg l⁻¹) | Chl-a (ug l⁻¹) | Algae density (×10⁶ Ind. l⁻¹) |
|-----------|---------|--------------|-------------|---------------|-----------------------------|
| Mean      | 19.10   | 2.18         | 0.05        | 6.04          | 7.95                        |
| Min.      | 1.70    | 0.00         | 0.00        | 0.10          | 0.09                        |
| Max.      | 36.20   | 8.31         | 0.32        | 92.8          | 93.17                       |

Note: WT (water temperature), TN (total nitrogen), TP (total phosphorus).

Figure 2. The structure of the hybrid ARIMA-LSTM model for prediction of cyanobacterial bloom: (a)–(c) present the flowchart of the ARIMA-LSTM model, the structure of the stacked LSTM networks and the internal architecture of an LSTM memory cell, respectively.

LSTM. A stepwise procedure for the hybrid ARIMA-LSTM model is illustrated in figure 2. As part of this method, ARIMA serves as a stand-alone statistical model to disentangle the linear and nonlinear components embedded in the observed time series. LSTM is then used to predict the residuals, resulting from the nonlinear processes, that are decomposed by ARIMA predictions. Using this combined method, the hidden patterns and relationships from the time-series observations may be captured, especially when a large degree of complexity is present (Jeong et al 2008, Ömer Faruk 2010). Following the competitive architecture of a hybrid approach (Zhang 2003, Khashei and Bijari 2011), it is assumed that the linear autoregressive components and nonlinear components are additive, expressed as follows:

\[ y_t = L_t + N_t \]

where \( L_t \) and \( N_t \) represent the linear and nonlinear components, respectively. These two components can be estimated from the observed data. ARIMA is firstly applied to fit the linear components \( L_t \), yielding a time series of predictions \( \hat{L}_t \). By comparing the observed value \( y_t \) and the predicted value \( \hat{L}_t \), we obtain a time series of residuals \( e_t = y_t - \hat{L}_t \). Then, the residuals encompassing the nonlinear relationships are fed to the subsequent LSTM networks. The LSTM model serves to produce a series of nonlinear components \( \hat{N}_t \), using the previously deduced residuals \( e_t \) as input variables. That is, the residuals will be estimated by a nonlinear approximation function, as shown below:

\[ e_t = f(e_{t-1}, e_{t-1}, \ldots, e_{t-n}) + \varepsilon_t. \]

Here, \( f \) denotes a nonlinear function determined by the structure of neural networks, and \( \varepsilon_t \) is the
remaining error term. As noted by Zhang (2003), if the \( f \) model is not appropriately identified, the remaining error is not necessarily random. The overall forecast \( \hat{y}_t \) of the hybrid model is given below:

\[
\hat{y}_t = \hat{L}_t + \hat{N}_t.
\]

The ARIMA and LSTM model were implemented with Python. The ARIMA model was realized using the Statsmodel library, and the fit and forecast functions were called to fit the model and produce predictions, respectively. The LSTM model was developed with Keras, a high-level library that could build fast and efficient deep learning models. The input data, fed to LSTM, are in the form of a 3D array with a shape of (samples, time steps, features). The network structure consists of two hidden layers, followed by a fully connected output layer that makes one-step-ahead predictions. Both layers are regularized by a dropout layer, to improve generalization. Also, the network employs the mean squared error as a loss function and the ‘Adam’ method as an optimizer. The details of the model structure and parameters used to configure the ARIMA-LSTM model are presented in table 3.

2.3.2. Model performance evaluation

ARIMA and LSTM are used as baseline models against which the performance of the hybrid ARIMA-LSTM model can be gauged. Generally, using multiple performance measures is advised when performing model calibration, as each measure has its pros and cons (Moriasi 2015). To evaluate the goodness-of-fit, three performance metrics are used herein, including the mean absolute error (MAE), root mean square error (RMSE) and Nash–Sutcliffe efficiency (NSE). These metrics possess different features, which may help to efficiently understand the predictive capabilities of the models from various perspectives. The above-mentioned estimators are given by:

\[
\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}
\]

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

where \( y_i \) is the observed value, \( \hat{y}_i \) is the predicted value and \( \bar{y} \) is the average of the series \( y_i \) in the test dataset. RMSE and MAE quantify the differences between observations and corresponding predictions with the same unit of the variance. NSE is a normalized statistic that determines the relative magnitude of the residual variance compared to the observed variance (Cao et al 2016). It ranges between (-\( \infty \), 1), with NSE = 1 corresponding to a perfect fit.

3. Results

3.1. The highly variable time-series observations

Phytoplankton dynamics are often characterized by convoluted time evolutions, which are generally non-stationary on top of seasonal oscillations and exhibit high variability at both inter-annual and intra-annual timescales (Choi et al 2019, Wilkinson et al 2020). Our results showed a rather high-level of irregularity and volatility in algal biomass (figure 3). The observed time series represented a high fluctuation, spanning a wide range of about ten times the standard deviation during the monitoring period. The results showed a significant negative trend (Kendall’s tau = -0.07; \( p < 0.01 \)) with the Mann–Kendall test, indicating a slow and consistent decline in algal biomass over time. Previous studies (Zhang et al 2012, Shi et al 2019) have shown that algal biomass in Taihu Lake exhibit apparent seasonal cycles, generally with yearly peaking in early summer and declining in late autumn. However, the observed time series showed that there was not a discernable seasonal pattern in algal biomass, as no obvious seasonal periodicity is observed, as shown in figure 3. The lack of seasonality is somewhat unexpected, since it is well established that cyanobacteria populations may wax and wane with the season (Ma et al 2016, Shi et al 2019). Nevertheless, the absence of seasonality may be explained by the relatively short range of the time series, since temporal patterns can be better captured by long-term sequences. Obviously, two years of observation is a rather short-term period in terms of the phenological time scale. Also, we notice that an extremely intense cyanobacterial bloom occurred in early January in 2017. Although it is considered unlikely that cyanobacterial blooms can develop in winter due to low temperature, previous recordings (Ma et al 2016) have confirmed the persistence of cyanobacterial blooms with water temperature below 10 °C throughout winter in Taihu Lake. As indicated by Ma et al (2016), temperature increases together with nutrient enrichment promoted the growth of cyanobacteria, while low temperature lowered the loss rate of *Microcystis*, allowing winter blooms to sustain. This uncommon occurrence of CyanoHABs in turn demonstrated that phytoplankton dynamics in Taihu Lake are subject to stochasticity to a large extent.

To date, there is not a well-defined or clear relationship between cyanobacteria biomass and discriminative thresholds of CyanoHABs. Herein, we employ the guidance value of 1.5 × 10^7 Ind. l^{-1} (algal density) as the bloom threshold, as suggested by Ma et al (2013). About three fourths (71%) of the observations had cell densities below the bloom threshold with the mean greater than the median, leading to a strongly right-skewed distribution (figure 3). Most of the values fall to the right in the distribution curve and the tail is long and slim with few extremes. The skewness and kurtosis values were both positive.
Figure 3. The time series of sub-daily values of the algae density at the monitoring site. A relative frequency histogram of the corresponding observations is shown in the embedded subplot, with the dashed lines indicating the probability density functions of a skewed normal distribution.

3.2. Predictive performance of the models
We validated the proposed ARIMA-LSTM model with the observed time series on a rolling prediction task against two baseline models, i.e. the ARIMA and LSTM models. Figure 4 illustrates the time series of the ARIMA, LSTM and ARIMA-LSTM predictions, as well as the observations in the testing period. A visual inspection of the plot reveals that the correspondence between the ARIMA predictions and observations is, in most cases, similar to the LSTM predictions, indicating that the two baseline models may have a comparable predictive performance. Despite this similarity, pairwise comparison of the performance metrics shows that LSTM has a better performance. It is worth to note that, despite the performance gain achieved by LSTM, our results showed that LSTM is far from being superior to the statistical model, which contradicts findings from some previous studies (Jeong et al 2008, Lee and Lee 2018). The scatter plot in figure 4 depicts the joint distribution of predictions and observations, and it can be seen that both the ARIMA and LSTM models have a tendency to underestimate the extreme values when predicting higher levels of algal biomass. This is particularly the case when predictions are above the bloom threshold, and this feature of the result suggests that both of the baseline models may fail to capture the spiky patterns of the time series when cyanobacterial blooms occur.

In contrast, the proposed ARIMA-LSTM model outperforms the two baseline models by reducing the forecast errors by a noticeable margin. The ARIMA-LSTM model produced reasonable results that were well matched with the observations in terms of both timing and magnitude. In particular, the ARIMA-LSTM model has produced predictions that exhibit a stronger linear association ($R^2 = 0.97$) with the observations, as presented in figure 4(c). Also, the hybrid model can better capture the temporal evolution of bloom extremes, indicating there was an incremental improvement in forecast accuracy when predicting high levels of algal biomass. This improvement was further supported by the fact that the prediction errors (MAE and RMSE), compared with the LSTM model, were largely reduced by 12.4% and 15.5%, respectively. Nevertheless, it is important to note that considerable caution and prudence need to be exercised when discerning a bloom event when using the hybrid model, as it is positively biased with higher values (figure 4) and, hence, is prone to producing overestimated predictions.

In table 2, we report the performance metrics resulting from the ARIMA, LSTM and ARIMA-LSTM models. We can see that the ARIMA-LSTM model performs the best in all metrics, especially in terms of NSE. As noted by Moriasi (2015), NSE is a robust quantitative measure, and can be used to determine how well the model simulates trends for the output response of interest. It appears from table 2 that the hybrid model is better in terms of its ability to capture the structure inherent in the data, due to an incremental improvement in NSE. Further, a close examination of the residuals presented in figure 4 reveals that the residuals spread more evenly and fluctuate within a narrower range, indicating that the hybrid model is less biased. To summarize, it can be stated that the hybrid approach, by coupling the ARIMA and LSTM model, leads to a modest but incremental predictive performance.

3.3. Model sensitivity and uncertainty
Both statistical and deep-learning models have hyperparameters, and a fundamental yet important task is
The observed values were compared with model predictions from: (a) ARIMA, (b) LSTM and (c) ARIMA-LSTM. The left panels show the time series of model predictions, paired with observations and residuals. The right panels show the scatter diagrams of observations versus predictions. Note that $R^2$ is a goodness-of-fit measure for the linear regression model derived from the observed and predicted values.

Figure 4. The observed values were compared with model predictions from: (a) ARIMA, (b) LSTM and (c) ARIMA-LSTM. The left panels show the time series of model predictions, paired with observations and residuals. The right panels show the scatter diagrams of observations versus predictions. Note that $R^2$ is a goodness-of-fit measure for the linear regression model derived from the observed and predicted values.

Table 2. A summary of model performance metrics from the ARIMA, LSTM and ARIMA-LSTM models.

| Metrics | ARIMA | LSTM | ARIMA-LSTM |
|---------|-------|------|------------|
| MAE     | 0.727 | 0.590| 0.517      |
| RMSE    | 1.137 | 1.047| 0.885      |
| NSE     | 0.925 | 0.934| 0.953      |

Table 3. The search space and results of parameter estimation for the ARIMA and LSTM models.

| Models | Parameters | Search space | Best trial |
|--------|------------|--------------|------------|
| ARIMA  | P          | [0, 4]       | 3          |
|        | Q          | [0, 4]       | 4          |
| LSTM   | Time step  | [1, 6]       | 6          |
|        | LSTM units | [20, 80]     | 20         |
|        | Epochs     | [50, 300]    | 100        |

Figure 5 shows the predictive performance with varying hyperparameters of LSTM. There is a gradual decrease in MAE and RMSE in figure 5(a), with a time step size of 6 providing the lowest error. The LSTM units in figure 5(b) show relatively lower error variance, and the MAE initially increases with the number of LSTM units and then seems to plateau after 60. In
Figure 5. Effects of different hyperparameters, (a) time steps, (b) LSTM units and (c) epochs, on the performance measurements of the ARIMA-LSTM model.

In view of this, it may be stated that the results are more sensitive to the time step size than the layer width. This statement is supported by the fact that there is a prominent time-lagged influence in cyanobacteria growth, as reported by previous studies (Recknagel et al. 2013, Zhang et al. 2014). Figure 5(c) shows the learning curve of validation, with the performance metrics plotted against training epochs. It can be observed that the MAE and RMSE initially decreased with epochs and then began to increase at around 100, suggesting that the model may be faced with the overfitting issue after being trained with more than 100 epochs.

Measurement uncertainty and natural variability can be a noteworthy problem when performing the prediction of cyanobacterial blooms. Measurement uncertainty, i.e. the inability to measure precisely, is inevitable because of the inaccuracy of the equipment. This is especially the case for turbid productive waters like Taihu Lake, where measurement of algal biomass may be confounded by the suspended particulates. And, as noted by Ahn et al. (2007), cyanobacterial cell counts may involve substantial uncertainty, as a microscopic enumeration of cyanobacterial cells is often confounded by large algae colonies and intertwined long filaments. As a result, cell counts may vary remarkably, subject to the skillfulness and provisional decision of analysts. Natural variability (i.e. variability inherent in the dataset), as mentioned previously, is another source of observational uncertainty. Natural variability of ecological monitoring is commonly seen, especially in Taihu Lake, as shallow lakes are more sensitive to external perturbations. It is clear that the prediction uncertainty also depends on the model structure. There is no doubt that the hybrid ARIMA-LSTM model increased the size of model complexity, and inevitably may lead to the growth of prediction uncertainty.
4. Discussion

4.1. The dilemma of CyanoHAB prediction

In this paper, we proposed a hybrid ARIMA-LSTM model, capable of extracting the benefits of both statistical and deep-learning models. Based on the results from comparisons of the three models, it can be concluded that the ARIMA-LSTM model marginally outperforms the two state-of-the-art predictive models. Even though the overall improvement in forecast accuracy was modest, this incremental improvement may still have important practical implications. This is because it may be these minor differences in forecast accuracy that determine whether or not any response actions should be launched to prevent and control the bloom. Moreover, the ARIMA-LSTM model performs better in terms of capturing high extremes and local fluctuations in the bloom, suggesting that the model is more robust and reliable when predicting emergent and intensive bloom events. Unlike other limnological variables, the time-series observations of phytoplankton may involve considerable nonlinearity and uncertainty in most cases. Hence, the selection of predictive models should be guided by the nature of the dataset at hand, and priority should be given to the hybrid approach when the dataset is characterized by extreme volatility.

Often, however, modelers are faced with a general difficulty in which the ecological data in aquatic systems are often sparse and incomplete, both spatially and temporally. Such ‘imperfect’ data make it challenging to establish multivariate predictive models, despite the fact that various approaches exist for dealing with missing or gapped data (Chen et al. 2015). On the other hand, with the advent of emerging monitoring technologies, high-frequency sensors provide enormous and consecutive time series that minimize the often lengthy time between sample collection, data analysis and response action. This feature provides researchers and managers with the ability to better capture temporal variability and elucidate bloom-related factors (Bowes et al. 2016). However, it may be prudent to undergo scrutiny of the dataset before blindly using them to develop predictive models of CyanoHABs for two main reasons. Firstly, as noted above, large uncertainty and error exists when using the time-series observations (Ahn et al. 2007, Rousso et al. 2020). Thus, the more variables that are used as model inputs, the higher the chance of introducing uncertainty that propagates within the model (Zhao and Kockelman 2002, Xiao et al. 2017). Secondly, the parsimony and interpretability of the developed models may be undermined with more input variables involved, due to the latent correlation between input variables and increased model complexity. In particular, these variables may contribute little to the explanatory capacity of results interpretation whilst adding unwanted model complexity. Also, another potential issue of predictive modeling with various variables is that the model may have high predictive performance but little ecological meaning, and is thus difficult to understand or interpret (Hamilton et al. 2009). This issue is further compounded by the ongoing controversies related to the controlling strategy of single (P only) versus dual (N and P) nutrients for cyanobacterial bloom mitigation in eutrophic water bodies (Paerl et al. 2016, Nelson et al. 2018). As Xiao et al. (2017) suggest, the single-variable models may be a useful modelling strategy and can be considered for implementation where appropriate.

4.2. Strengths and limitations

The results presented in this study, as well as previous studies involving the hybrid methods of CyanoHAB predictions (Qin et al. 2017, Xiao et al. 2017), indicate that the predictive performance can be promoted by the integration of statistical and deep-learning methods. This claim is also supported by a wide range of applications from many other fields (Zhang 2003, Khashei and Bijari 2011). Several reasons can be suggested to explain why the hybrid method performs better. The most important explanation is that the hybrid model decomposes the dataset to different subseries, and thus can better resolve temporal variability. Shallow lakes like Taihu Lake are vulnerable to variability because of their higher sensitivity to perturbations, and this variability is assumed to be strengthened under climate uncertainty (O’Reilly et al. 2015). In this context, the hybrid ARIMA-LSTM model may be an appealing option for CyanoHAB prediction in shallow and eutrophic lakes. In our results, the hybrid ARIMA-LSTM model performs better than two state-of-the-art baseline models when predicting higher levels of algal biomass. This means that the hybrid model can better capture the short-term local variability induced by transient proliferations of cyanobacteria. For the same reason, this also means that the ARIMA-LSTM model can be a more reliable method when considering whether or not to raise a bloom warning, whereby the algal biomass exceeds the bloom threshold.

Although the results of the hybrid method are promising, it is important to note that several limitations exist in the hybrid model. Firstly, there is no doubt that the hybrid methods add considerably to model complexity accompanied with increased uncertainty. However, to what extent the uncertainty is enlarged remains unknown, and this may be investigated in future study. Secondly, in our model, it is assumed that the linear and nonlinear component of the model are additive, while this is not necessarily the case for the real-world process (Khashei and Bijari 2011). Besides, hyperparameter tuning of the hybrid method can be a challenging and laborious task, since more hyperparameters are introduced and thus more model runs need to be performed to find
the best trial. Thirdly, this study only used the time-series observations of a single site for model development due to limited available data. This means that the developed model is site-specific (i.e. a regionally tuned model), and this may undermine its generalization and usefulness in predicting CyanoHABs in various aquatic systems across heterogeneous environmental gradients. More data would have made it possible to perform further analysis, producing results that may enhance the credibility of this study or highlight potential problems.

4.3. Implications for environmental management
Both simple and complex models have their merits and pitfalls. This fact suggests that modelers should carefully balance the trade-off between complexity and accuracy, parsimony and pragmatism. For CyanoHAB prediction, the hybrid method should be given a higher priority, for their better performance over nonlinear and nonstationary time-series observations. Although the hybrid methods do not necessarily always lead to better performance, they reduce the risk of using an inappropriate model and therefore the risk of inducing spurious results or total prediction failure (Zhang 2003, Khashei and Bijari 2011). With regard to the ARIMA-LSTM model specifically, the model has the potential for deployment over online monitoring and prediction platforms, which is a comparative advantage over models that demand time-consuming sample collection and laboratory analysis. When using the proposed hybrid model, it is assumed that there is some a priori information about the presence of high variability in the time-series observations, which makes the procedure somewhat Bayesian in nature. Our model therefore may be a better alternative for early warning of CyanoHABs in shallow and eutrophic waters, characterized by high-frequency fluctuations of environmental parameters. Furthermore, it can be stated that the distant past is less important for short-term CyanoHAB prediction, as the underlying temporal patterns and response may have changed in the meantime.

5. Conclusions
The prognostic information of temporal variation in cyanobacteria is invaluable to local managers as a means to reduce the environmental risks resulting from CyanoHABs. This paper contributes a methodology for prediction of CyanoHABs with high-frequency time-series observations using a hybrid approach of a statistical and deep-learning method, demonstrated on a case study of a shallow and eutrophic lake, Taihu Lake. The results indicate that the ARIMA-LSTM model is a competitive method, effectively capturing temporal variability of the time series and thereby being able to outperform two state-of-the-art baseline models. Our results also illustrate that the hybrid model may have priority over standalone models for short-term CyanoHAB prediction in shallow and eutrophic waters, which are characterized by highly fluctuating time-series observations. Spatial and temporal variability of the lacustrine environment may be important sources of uncertainty that affect the predictive capability of the models. We suggest that these variabilities should be taken into consideration when exploring and evaluating the hybrid approach in a wider context. With further work to examine the credibility and practicality of this method within a real-world decision context, this approach could be a promising tool to support cyanobacterial risk assessment and management.

Data availability statement
The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments
This research is funded by the Major Science and Technology Program for Water Pollution Control and Treatment (No. 2017ZX07301006), the National Science Foundation of Jiangsu Province (No. BK20191181), and the National Science Foundation of China (No. 51878372). The authors would like to thank the Suzhou Environmental Monitoring Center for providing the monitoring data used in this study. The authors also want to thank the anonymous reviewers for their insightful comments and suggestions, which helped to improve this manuscript substantially.

References
Ahn C Y, Joung S H, Yoon S K and Oh H M 2007 Alternative alert system for cyanobacterial bloom, using phycocyanin as a level determinant J. Microbiol. 45 98–104
Ahn C Y, Oh H M and Park Y S 2011 Evaluation of environmental factors on cyanobacterial bloom in eutrophic reservoir using artificial neural networks J. Phycol. 47 495–504
Bowes M J, Loewenthal M, Read D S, Hutchins M G, Prudhomme C, Armstrong L K, Harman S A, Wickham H D, Gozzard E and Carvalho I 2016 Identifying multiple stressor controls on phytoplankton dynamics in the River Thames (UK) using high-frequency water quality data Sci. Total Environ. 569–570 1489–99
Box G E 1976 Time Series Analysis: Forecasting and Control (San Francisco, CA: Holden-Day)
Brasil J, Attayde J L, Vasconcelos F R, Dantas D D and Huszar V I 2016 Drought-induced water-level reduction favors cyanobacteria blooms in tropical shallow lakes Hydrobiologia 770 145–64
Cao H, Recknagel F and Barkow M 2016 Spatially-explicit forecasting of cyanobacteria assemblages in freshwater lakes by multi-objective hybrid evolutionary algorithms Ecol. Model. 342 97–112
Chen Q, Guan T, Yun L, Li R and Recknagel F 2015 Online forecasting chlorophyll a concentrations by an
auto-regressive integrated moving average model: feasibilities and potentials Harmful Algae 43 58–65
Choi J, Kim J, Won J and Min O 2019 Modelling chlorophyll-a concentration using deep neural networks considering extreme data imbalance and skewness 2019 21st Int. Conf. on Advanced Communication Technology (ICACT) pp 631–4
Fang S, Del Guerude D, Scavia D, Binding C E, Bridgeman T B, Chaffin J D, Evans M A, Guinness J, Johengen T H and Obenour D R 2019 A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent Sci. Total Environ. 695 133776
Fornarrelli R, Galleli S, Castelletti A, Antenucci J P and Marti C L 2013 An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers Water Resour. Res. 49 3626–41
Hamilton G, McVinish R and Mengersen K 2009 Bayesian model averaging for harmful algal bloom prediction Ecol. Appl. 19 1805–14
Hochreiter S and Schmidhuber J 1997 Long short-term memory Neural Comput. 9 1735–80
Huang J, Gao J and Hörmann G 2012 Hydrodynamic-phytoplankton model for short-term forecasts of phytoplankton in Lake Taihu China Limnol. 42 7–18
Jeong K S, Kim D K, Jung J M, Kim M C and Joo G J 2008 Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics Ecol. Modell. 211 292–300
Khashei M and Bijari M 2011 A novel hybridization of artificial neural networks and ARIMA models for time series forecasting Appl. Soft Comput. 11 2664–75
Lee S and Lee D 2018 Improved prediction of harmful algal blooms in four major South Korea’s Rivers using deep learning models Int. J. Environ. Res. Public Health 15 1322
Ma J et al 2016 The persistence of cyanobacterial (Microcystis spp.) blooms throughout winter in Lake Taihu, China Limnol. Oceanogr. 61 711–22
Ma J, Deng J, Qin B and Long S 2013 Progress and prospects on cyanobacteria bloom-forming mechanism in lakes Acta Ecol. Sin. 33 3020–30
Marcus R et al 2016 Automatic high frequency monitoring for improved lake and reservoir management Environ. Sci. Technol. 50 10780–94
McGowan J A, Deyle E R, Ye H, Carter M L, Perretti C T, Seger K D, De Verneil A and Sugihara G 2017 Predicting coastal algal blooms in southern California Ecology 98 1419–33
Moriasi D N 2015 Hydrologic and water quality models: performance measures and evaluation criteria Trans. ASABE 58 1765–83
Nelson N G, Muñoz-Carpena R, Phelps E J, Kaplan D, Susic P and Hendrickson J 2018 Revealing biotic and abiotic controls of harmful algal blooms in a shallow subtropical lake through statistical machine learning Environ. Sci. Technol. 52 3527–35
O'Reilly C M, Sharma S, Gray D K, Hampton S E, Read J S, Rowley R J, Schneider F, Lenters J D, McIntyre P B and Kraemer B M 2015 Rapid and highly variable warming of lake surface waters around the globe Geophys. Res. Lett. 42 10773–81
Ömer Faruk D 2010 A hybrid neural network and ARIMA model for water quality time series prediction Eng. Appl. Artif. Intell. 23 586–94
Paerl H W, Otten T G and Kudela R 2018 Mitigating the expansion of harmful algal blooms across the freshwater-to-marine continuum Environ. Sci. Technol. 52 5519–29
Paerl H W, Scott J T, McCarthy M J, Newell S E, Gardner W S, Havens K E, Hoffman D K, Wilhelm S W and Wurthbaugh W A 2016 It takes two to tangle: when and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems Environ. Sci. Technol. 50 10805–13
Pyo J, Park J I, Pachephy P, Baek S-S, Kim K and Cho K H 2020 Using convolutional neural network for predicting cyanobacteria concentrations in river water Water Res. 186 116349
Qin B, Li W, Zhu G, Zhang Y, Wu T and Gao G 2015 Cyanobacterial bloom management through integrated monitoring and forecasting in large shallow eutrophic Lake Taihu (China) J. Hazards Mater. 287 356–63
Qin M, Li Z and Du Z J 2017 B-S Reed tide time series forecasting by combining ARIMA and deep belief network Knowl.-Based Syst. 125 39–52
Recknagel F, French M, Harkonen P and Yabunaka K 1997 Artificial neural network approach for modelling and prediction of algal blooms Ecol. Model. 96 11–28
Recknagel F, Orr P T, Bartkow M, Swanepeol A and Cao H 2017 Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling Harmful Algae 69 18–27
Recknagel F, Orr P T and Cao H 2014 Inductive reasoning and forecasting of population dynamics of Cylindrospermopsis raciborskii in three sub-tropical reservoirs by evolutionary computation Harmful Algae 31 26–34
Recknagel F, Ostrovsky I, Cao H, Zohary T and Zhang X 2013 Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computation and wavelets Ecol. Modell. 255 70–86
Rousso B Z, Bertone E, Stewart R and Hamilton D P 2020 A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes Water Res. 182 115892
Shi K, Zhang Y, Zhang Y, Li N, Qin B, Zhu G and Zhou Y 2019 Phenology of phytoplankton blooms in a trophic lake observed from long-term MODIS data Environ. Sci. Technol. 53 2324–31
Teles I O, Vasconcelos V, Teles L O, Pereira E, Saker M and Vasconcelos V 2006 Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks Environ. Manage. 38 227–37
Wei B, Sugirua N and Maekawa T 2001 Use of artificial neural network in the prediction of algal blooms Water Res. 35 2022–8
Wilkinson G M, Walter J, Fleck R and Paci M L 2020 Beyond the trends: the need to understand multimannual dynamics in aquatic ecosystems Limnol. Oceanogr. Lett. 5 281–6
Xiao H, Je H, Huang H, Miller T R, Christakos G, Reichwaldt E S, Qin M, Li Z and Du Z J 2017 A novel single-parameter approach for forecasting algal blooms Water Res. 108 222–31
Zhang G P 2003 Time series forecasting using a hybrid ARIMA and neural network model Neurocomputing 50 159–75
Zhang M, Duan H, Shi X, Yu Y and Kong F 2012 Contributions of meteorology to the phenology of cyanobacterial blooms: implications for future climate change Water Res. 46 442–52
Zhang X, Chen Q, Recknagel F and Shi R 2014 Wavelet analysis of time-lags in the response of cyanobacteria growth to water quality conditions in Lake Taihu, China Ecol. Inform. 22 52–57
Zhang Y, Qin B, Zhu G, Shi K and Zhou Y 2018 Profound changes in cyanobacteria bloom-forming mechanism in lakes Limnol. Oceanogr. 70–86
Zhao Y and Kockelman K M 2002 The propagation of uncertainty through travel demand models: an exploratory analysis Ann. Reg. Sci. 36 145–63