Chlorophyll Prediction Using Ensemble Deep Learning Technique

Ashapurna Marndi and G. K. Patra

Abstract Chlorophyll is an essential component of phytoplankton and plays an important role in food chain and nutrient cycle required for survival of marine creatures. Getting suitable fishing zone is one of the commercial usages of liveliness measurement of marine ecosystem. Optimal sustainability of marine ecosystems needs an accurate prediction of chlorophyll. Dynamical models to predict the chlorophyll are challenged by the complex physical, chemical, and biological processes. Numerous researchers have attempted to address this problem using various computationally intelligent methods such as neural networks. However, normal neural networks have failed to provide a reliable forecast. This paper proposes a novel ensemble forecasting using Long Short Term Memory (LSTM) and a deep learning (DL) approach for time series data analysis. The methodology was tested to predict chlorophyll in Arabian Sea and found satisfactory result. Improved capabilities of the proposed method are also been demonstrated through various statistical analyses.

Keywords Ensemble forecasting · Long short-term memory · Artificial intelligence · Deep neural networks · Chlorophyll prediction

1 Introduction

Chlorophyll prediction plays an important role in the algal bloom predictions. The concentrations of the plant pigment, i.e., "chlorophyll-a", occur in all marine phytoplankton and provide useful proxy indicator of the amount of nutrients incorporated into phytoplankton biomass. This is because phytoplankton have predictable

A. Marndi (✉) · G. K. Patra
Academy of Scientific and Innovative Research, Ghaziabad 201002, Uttar Pradesh, India
e-mail: asha@csir4pi.in

G. K. Patra
e-mail: gkpatra@csir4pi.in

Council of Scientific and Industrial Research, Fourth Paradigm Institute, Bengaluru 560037, Karnataka, India

H. Das et al. (eds.), *Progress in Computing, Analytics and Networking*,
Advances in Intelligent Systems and Computing 1119,
https://doi.org/10.1007/978-981-15-2414-1_34

nutrient-to-chlorophyll ratios. Chlorophyll-a is the most commonly used parameter for monitoring phytoplankton biomass and nutrient status, as an index of water quality.

Increased nutrient availability, for example, from human activity such as agricultural runoff, soil erosion, discharges of sewage and aquaculture waste, etc. usually leads to a rise in chlorophyll concentrations in coastal waters because of increased phytoplankton biomass. Phytoplankton can rapidly deplete nutrients to levels which would be difficult to sample and analyze directly. Monitoring chlorophyll levels is a direct way of tracking algal growth. Amount of algal availability in a sea zone is usually proportionate to marine creature in that zone. Suitable fishing zones are detected based on extent of chlorophyll availability.

There are various conventional approaches used to solve different prediction problems. However, there are very few traditional techniques used to predict chlorophyll and in the recent years the biogeochemical models helped in assimilating the chlorophyll. Along with traditional approaches, few artificial intelligence (AI) approaches have also been attempted to predict the chlorophyll in near time. Nowadays, subfield of AI, machine learning (ML), is getting significant attention in ocean research fields. ML specifically deep learning (DL) algorithms are capable of discovering hidden important patterns from massive data leading to build insight for concise and reliable analysis. Deep learning technology has been successfully used not only in data analytics of commercial fields but also in scientific fields [1, 2].

LSTM has unique capability to utilize past learned information in best way to predict future values. It has gained significant success in predicting specifically time series data. In this experiment, LSTM technique is used to provide better solution in predicting chlorophyll for future timestamp. Further, this paper describes an enhancement upon on LSTM and found to be better than normal LSTM solving this problem. The merits of the proposed technique are demonstrated comparing with basic LSTM to predict chlorophyll in Arabian Sea. The result shows that the proposed algorithm has made significant improvement on basic LSTM that enhances the accuracy of long-term prediction of chlorophyll.

2 Related Work

Several researchers have tried to solve the problem of Chlorophyll-a (Chl-a) concentration prediction in different ways. In [3], it is attempted to use support vector machine for regression (SVR) and random forest (RF) to predict Chl-a concentration based on multiple variables and concluded that the RF model had a higher predictive ability than the SVR model. Random forest model with a sliding window strategy is used in [4] to predict Chl-a concentration in freshwater. Basic long short-term memory was used for predicting chlorophyll in 2 days and 4 days lead time based on daily measured water quality data as input. Again in [5], long short-term memory was used for forecasting air pollutant concentration. In [6], they combine ANN and

generic algorithm to predict chlorophyll. In [7], it has suggested two model parameters, e.g., electrical conductivity (EC) and turbidity, and it has showed the efficiency of predictive model increased by adding them. In [8], comparison of the predictive performance was made using four types of multiple linear regression (MLR) and principal component regression (PCR) models were compared for chlorophyll-a (Chl-a) prediction. In [9], two chlorophyll predictors, ANN model and SVM models, are compared for chlorophyll prediction.

3 Methodology

3.1 Design Consideration for Model

An efficient model needs all the important parameters to be considered carefully. To solve a problem with no deterministic steps or fixed and finite input data, it is very important to consider all aspects that affect output and draw a fine balance of adjustments among them to derive conclusive result. Possible all such aspects are discussed below that can influence the outcome of this experiment:

Input Parameters

As chlorophyll concentration is influenced by various oceanographic parameters such as sea surface temperature (SST), sea surface salinity (SSS), and sea surface height (SSH) [10]. They have been considered along with chlorophyll as input to the model.

Range of Prediction

Range of prediction plays an important role for solving prediction model. Though it varies application to application, it is required to select the range of prediction carefully based on requirement of the application. However, it is obvious that prediction accuracy decreases with increase of time range of prediction. After considering usability of the application along with tolerable mismatch level, the 2-month time ahead period is been fixed for this.

Fitting Base Model

There are several variants of artificial intelligence techniques available to solve such prediction problems. The long short-term memory (LSTM) in deep learning approach is one of such techniques which can remember past important information for long time and conveniently forgets the less useful patterns. The ability to remember long back pattern leads to predict time series data efficiently. Since data used in this problem are all time series in nature, the solution in this paper is considered to be upon LSTM with suitable modification.

3.2 *Long Short-Term Memory (LSTM)*

The LSTM [11] model consists of forget gate (f_n), input gate (i_n), and output gate (o_n) as shown in Fig. 1. The forget gate is responsible for deciding unwanted information to be discarded. The input gate is responsible for deciding what new information is to be stored in the cell state and what new information to be added by *tanh* layer. Old cell state is updated to new one by removing the information required to be forgotten from the previous state and adding new information to the current state. Finally, the output is based on cell state but filtered through sigmoid layer and then multiply the parts of sigmoid layer output by *tanh* of cell state. Following Eqs. (1)–(5) have represented them mathematically.

$$f_n = \sigma\left(W_f I_n + U_f h_{n-1} + b_f\right) \tag{1}$$

$$i_n = \sigma\left(W_i I_n + U_i h_{n-1} + b_i\right) \tag{2}$$

$$o_n = \sigma\left(W_o I_n + U_o h_{n-1} + b_o\right) \tag{3}$$

$$C_n = f_n * C_{n-1} + i_n * \tanh(W_c I_n + U_c h_{n-1} + b_c) \tag{4}$$

$$h_n = \tanh(C_n) * o_n \tag{5}$$

Input to LSTM network is denoted by I_n. Based on values of forget gate, input gate, and output gate mentioned in Eqs. (1)–(3), cell state (C) and hidden state (h) of LSTM are being updated by Eqs. (4) and (5). W_f, W_i, W_o, W_c and U_f, U_i, U_o, U_c are the weight matrixes at current state and previous state and b_f, b_i, b_o, b_c are the bias

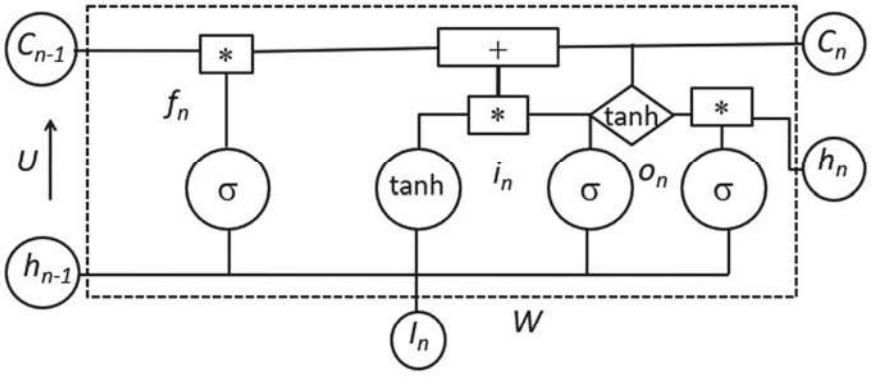

Fig. 1 Architecture of a LSTM network

vectors of forget gate, input gate, output gate, and cell state, h_{n-1} is the hidden unit of previous state, and σ is the activation function.

3.3 Proposed Model (Moving Window LSTM)

The proposed prediction algorithm is based on multilevel LSTM by treating the outputs from first level of LSTMs as the input for the second level. The first-level LSTMs take different ensemble datasets as input divided based on moving window of fixed size. The purpose of distributing input dataset into multiple time windows is that if a pattern strives out to become an impactful pattern for final output, it should be present in all windows of recurring timeframe, which is usually year in this case. Second, the time windows are made to be moving window of more than a year, by which it ensures that the impactful patterns can be in any month or covering multiple months without breaking across time boundaries of a year, though the seasonal pattern usually repeats over years. Since the algorithm is built upon moving window ensembled data with multilevel LSTM, it is named as "Moving Window LSTM" (MW-LSTM).

The whole training dataset is divided into moving window of 4 years duration and the test data is also kept for same duration. As shown in Fig. 2, the moving window period for training data starts from 2004–2007 and continues till 2009–2012 period with gap of a year in subsequent windows. The data during 2013–2016 are considered as test data. There are six LSTMs in the first level, one for each ensemble training dataset of moving window period, and are trained to output optimal result. In second level, outputs from first-level LSTMs are combined and used as inputs as shown by

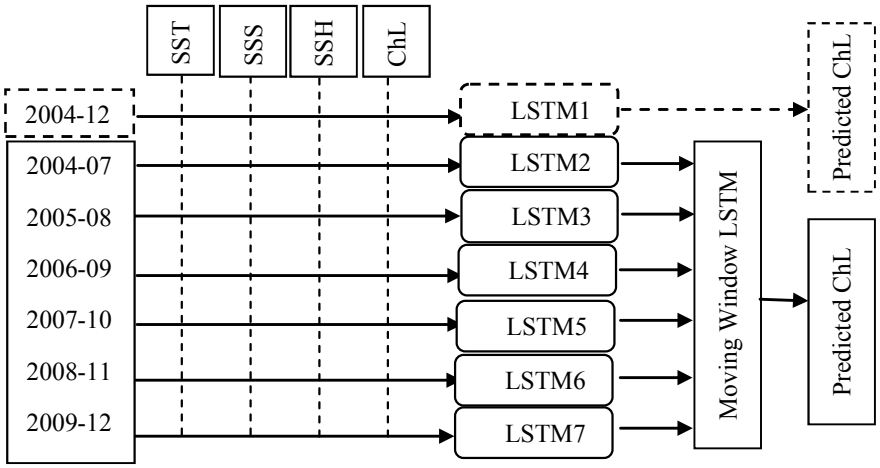

Fig. 2 Block diagram of proposed model depicting multilevel LSTMs with moving window input datasets

LSTM2 to LSTM7 in Fig. 2. In parallel, the whole input dataset 2004–2012 is also fed to a normal LSTM, i.e., LSTM1 in the figure and the output is compared with the output of second-level LSTM in the proposed model.

3.4 Study Area and Experimental Dataset

For this experiment, it has been chosen a portion of the Arabian Sea (Long-65E:72E and Lat-12. 5 N:15 N) as the study area for duration of 2004–2016. The whole data is divided into two sets such as data during 2004–2012 as training data and that of during 2013–2016 as testing data. The subsequent data are captured in 5 days intervals. Sea surface temperature (SST), sea surface salinity (SSS), and sea surface height (SSH) are collected from open-source live access server of Indian National Centre for Ocean Information Services Site (INCOIS) [12], and the chlorophyll data are taken from the open-source merged ocean color data [13].

3.5 Experiment Setup

In this experiment, LSTMs were configured with same set of hyperparameter values. In first level, all LSTMs were trained with 6 hidden layers with 50 neurons each layer. The experiment was initiated with 10 neurons at the first layer and then kept increasing by 10 more neurons in the same layer until the network gave satisfactory result. Once the number of neurons were fixed, the networks were tuned further by adding additional layer starting from second layer till optimum network was achieved. The network is treated as optimal when no further improvement was found by addition of any more neurons or layers. Thus, the LSTM configuration was fixed with 6 hidden layers and 50 neurons in each layer. Following similar approach for the second-level LSTM, it was observed that in epoch 900, it was optimized with 10 hidden layers and 50 neurons in each hidden layer.

The experiment was carried out on intel(R) Xeon(R) CPU E3-1203 v3 @ 3.30 GHz with 8 cores and 32 GB RAM.

4 Result and Discussion

Different statistical tests were carried out to find out the accuracy of predictability of the proposed method with the normal LSTM. The statistical parameters that were used to test are root mean square error (RMSE) and correlation coefficient (CC). These two parameters give a very good estimation on closeness of two time series patterns.

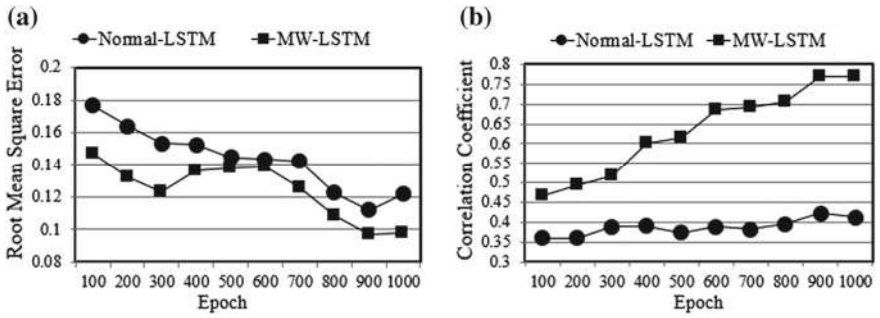

Fig. 3 For normal LSTM and MW-LSTM **a** RMSE versus epoch, **b** CC versus epoch

Figure 3a shows the RMSE of normal LSTM and proposed MW-LSTM for different number of epochs (100–1000). From this figure, it is observed that the RMSE for MW-LSTM is better compared to normal LSTM. This indicates that the average error in case of ensemble forecasting is less than the non-ensemble forecasting. However, it is not good enough to establish the superiority of the proposed method with only RMSE efficiency. Hence, the correlations between the two time series datasets were also compared. Figure 3b depicts the correlation coefficients in both the approaches, i.e., normal LSTM and proposed MW-LSTM between the two time series data that are observed and predicted for different epoch sizes. For both RMSE and CC, the values are found to be saturated after 900 epoch values and thus epoch 1000 was considered as threshold for this experiment by keeping slightly safe margin.

As shown in Fig. 4a–d, the chlorophyll values predicted using MW-LSTM are better than the normal LSTM. The outcome of this experiment is due to multiple factors such as dividing inputs into multiple ensemble datasets based on moving window period and then optimizing the output in multiple level of LSTMs. The more efficiency of this algorithm is demonstrated not only by visual graphs but also through statistical parameters such as root mean square error (RMSE) and correlation coefficient (CC) values.

5 Conclusion

Ensemble forecasting is a well-known methodology in atmospheric sciences using dynamical models. However, uses of ensemble forecasting using artificial intelligence (AI) technique are relatively new and have shown good promises. Use of ensemble forecasting using multilevel LSTMs over moving window data is novel approach giving better result. Chlorophyll prediction in ocean, especially in the Arabian Sea, is a challenging as well as important requirement for the sustainability of the marine ecosystem. This work demonstrated that using the deep learning architecture of LSTM, in an ensemble methodology, a reliable and usable prediction can be made. It has demonstrated how a set of individual LSTMs in the first layer, with

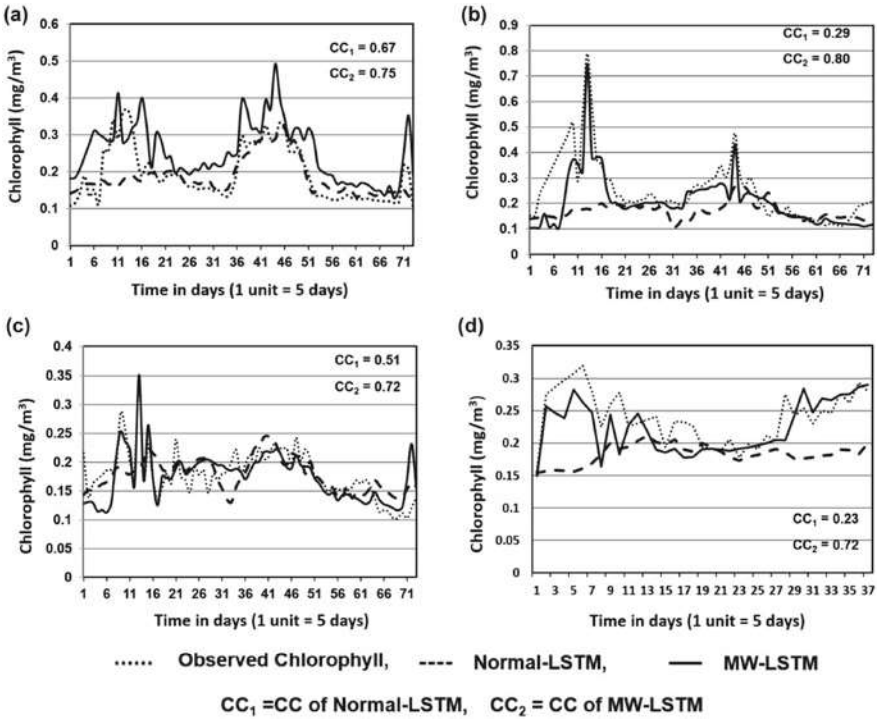

Fig. 4 Predicted results in both normal LSTM and moving Window LSTM versus observed Chl-a during **a** 2013, **b** 2014, **c** 2015, and **d** 2016

overlapping subset of data, is capable of extracting hidden important patterns that can be fed to next level of LSTM to give an optimal result, especially in this case of chlorophyll prediction. The statistical parameters such as correlation coefficient and RMSE indicate the superiority of this proposed method. In coming days, such enhancement over normal LSTM can be useful solving more complex scientific problems in variety of domains.

References

1. Das, H., Naik, B., Behera, H.S.: Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) approach. In: Progress in Computing, Analytics and Networking, pp. 539–549. Springer, Singapore (2018)
2. Sahoo, A.K., Mallik, S., Pradhan, C., Mishra, B.S.P., Barik, R.K., Das, H.: Intelligence-based health recommendation system using big data analytics. In: Big Data Analytics for Intelligent Healthcare Management, pp. 227–246. Academic (2019)
3. Li, X., Sha, J., Wang, Z.-L.: Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. Environ. Sci. Pollut. Res. 1–11 (2018)

4. Yajima, H., Derot, J.: Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. J. Hydro Inf. **20**, 206–220 (2018)
5. Cho, H.: Deep: Learning Application to Time Series Prediction of Daily Chlorophyll-a Concentration (2018)
6. X. et al.: Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. **231**, pp. 997–1004 (2017)
7. Lee, G., Bae, J., Lee, S., Jang, M., Park, H.: Monthly chlorophyll-a prediction using neuro-genetic algorithm for water quality management in Lakes. Desalin. Water Treat. **57**, 26783–26791 (2016)
8. Lee, G., Othman, F., Ibrahim, S., Jang, M.: Determination of the forecasting-model parameters by statistical analysis for development of algae warning system. Desalin. Water Treat. **57**, 26773–26782 (2016)
9. Cho, K.H., Kang, J.-H., Ki, S.J., Park, Y., Cha, S.M., Kim, J.H.: Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan reservoir. Korea. Sci. Total Environ. **407**, 2536–2545 (2009)
10. Krasnopolsky, V., Nadiga, S., Mehra, A., Bayler, E., Behringer, D.: Neural networks technique for filling gaps in satellite measurements: application to ocean color observations. Comput. Intell. Neurosci. **2016**, 9. Article ID 6156513. http://dx.doi.org/10.1155/2016/6156513
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
12. ESSO—Indian National Centre for Ocean Information Services. https://las.incois.gov.in
13. NASA Ocean Color. https://oceancolor.gsfc.nasa.gov