Vietnam Journal of Marine Science and Technology 2023, 23(3) 223–232

Vietnam Academy of Science and Technology

Vietnam Journal of Marine Science and Technology

journal homepage: vjs.ac.vn/index.php/jmst

Optimizing the Long Short-Term Memory (LSTM) model by Bayesian method for salinity intrusion forecasting: a study at Dai Ngai station, Soc Trang province, Vietnam

Tien Giang Nguyen[1,*], Trung Duc Tran[2], Cong Thanh Nguyen[3]

[1]*Faculty of Hydrology, Meteorology & Oceanography, VNU University of Science, Vietnam*
[2]*School of Civil and Environmental Engineering, University of Ulsan, Republic of Korea*
[3]*Southern Regional Hydrometeorological Center, Hanoi, Vietnam*

Received: 17 March 2023; Accepted: 22 June 2023

ABSTRACT

Salinity intrusion forecasting is essential and challenging for hydrometeorology, especially in climate change. Employing machine learning (ML) algorithms and conventional forecasting techniques are gaining popularity and providing high performance. This study presents a method to optimize a machine learning model based on the Long Short-Term Memory (LSTM) algorithm for multistep-ahead salinity forecasting (up to 7 days) at Dai Ngai station, Soc Trang province. The optimization method based on the Bayesian algorithm for hyperparameters optimization and input predictors optimization has been highly effective for predicting salinity with a lead time of 1 day to 7 days. Specifically, the forecast results evaluated by the R^2 and RMSE indexes both give satisfactory results on the test data set (with lead time from 1 day to 7 days, R^2 ranges from 0.9 to 0.54, and RMSE ranges from 0.27 to 0.53). This study is a premise for improving machine learning models for short-term and long-term salinity intrusion prediction in the Mekong delta and Vietnam.

Keywords: Long Short-Term Memory (LSTM), Bayesian method, multistep-ahead salinity forecasting.

[*]Corresponding author at: Faculty of Hydrology, Meteorology & Oceanography, VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam. *E-mail addresses:* giangnt@vnu.edu.vn

https://doi.org/10.15625/1859-3097/18174
ISSN 1859-3097; e-ISSN 2815-5904/

INTRODUCTION

Salinity intrusion is an important issue because of its adverse effects on economic activities and social life [1]. Especially in the Mekong delta, salinity intrusion is increasingly complicated by human activities in the upstream Mekong river and sea level rise due to climate change [2]. Salinity intrusion forecasting is one of the crucial tasks in responding to its adverse impacts.

There are two main approaches to predicting salinity intrusion: mathematical models based on equations that simulate the physical processes of flows and material propagation and transport (physical-based) and data-driven models. With the physical-based approach, in the Mekong delta region, many studies have used the Mike model for short-term forecasting and salinity intrusion simulation for climate change scenarios [3–5]. The advantage of this approach is that it helps researchers to understand the relative physical processes and drivers of salinity intrusion, thereby providing a basis for responding solutions. Furthermore, this methodology offers high reliability due to its extensive development history and notable achievements. However, the above approach's limitations are difficult to solve in practice, including (i) large data requirements and (ii) expensive computational resources. Some critical data required for these models (e.g., topography, cross-sections, structures, hydrometeorology, etc.) are always difficult to collect and measure accurately. Specifically, these data are subject to alterations resulting from human activities such as dredging operations, sand mining, and river construction projects, thus increasing the margin of error in the model. Besides, model calculation complexity and simulation time challenge require extensive computational resources.

The data-driven approach is promising because of its advantages. Previously, this approach had many limitations because statistical models could not give good results for time series forecasting, even basic machine learning models such as Autoregressive integrated moving average (ARIMA), Decision tree, and Random Forest [6]. However, with the vigorous development of deep learning algorithms, the limitations of the data-driven approach have been improved and brought about high efficiency, especially in tasks such as streamflow and dam inflow forecasting [7–9], water level forecasting [10], regional wave height forecasting [11]. This approach has been applied to predict saline intrusion in the Mekong delta with promising results [12, 13].

Prominent among these algorithms is Long Short-Term Memory (LSTM), a particular Recurrent Neural Network (RNN) structure. The unique structure of LSTM allows this neural network to have the ability to remember long-term information in time series, through which the model built on LSTM can predict with a long period and high accuracy.

A deep learning model's training efficiency and quality depend on how the hyperparameters are tuned. However, optimizing the hyperparameters of any deep-learning model is always challenging [14]. In addition, the error accumulation problem in multistep-ahead prediction always worries both the modelers and model users [15].

This study aims to propose and test a hybrid model using the Bayesian optimization method and a deep learning model based on the LSTM network to give multi-step-ahead predictions of daily maximum salinity series at Dai Ngai station, Soc Trang province. The investigation outcomes have showcased the method's substantial efficacy, thus establishing the potential for its practical implementation. The model building and hyperparameter optimization process in this study is also a suggestion for other studies on building models based on applicable deep machine learning algorithms.

METHODOLOGY AND DATASET

Long-Short Term Memory (LSTM)

Long Short-Term Memory Network (LSTM) is a unique structure of Recurrent Neural Network (RNN) capable of remembering long-term information of time series data [16] and overcoming the disadvantages of the basic

RNN structure. The remarkable difference in the LSTM structure is the addition of a long-term "memory" (cell state C_t). This memory is updated and transmitted in the LSTM network to help the network remember the information in the long term. Besides, a short-term "memory" (hidden state h_t) is still retained as an RNN structure to calculate for each specific time step (Figure 1). These two memories are informed by equations called gates. These gates decide to discard information that is not useful to the network (forget gate), retain important information (input gate), and provide information to calculate output results. (output gate). The basic equations of the LSTM network structure are shown in Equations (1–6).

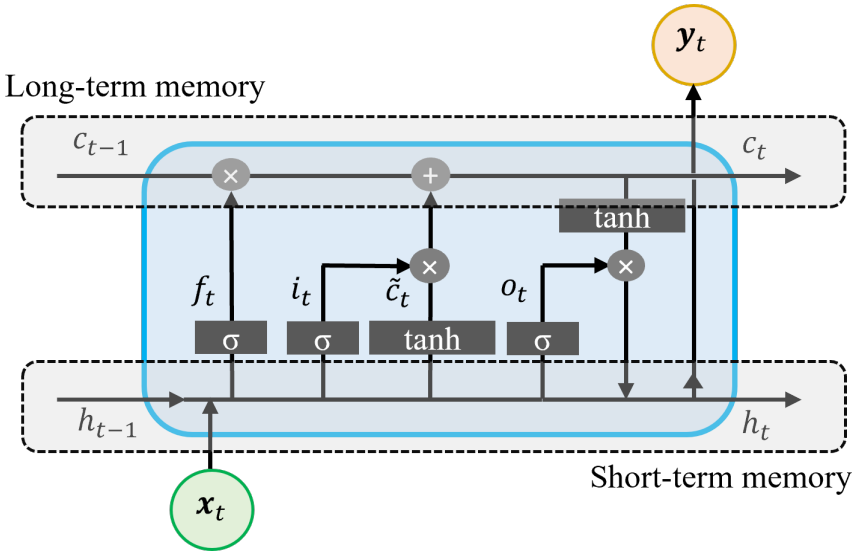

Figure 1. The basic structure of LSTM cell

$$f_t = \sigma\left(W_f.x_t + U_f.h_{t-1} + b_f\right) \quad (1)$$

$$i_t = \sigma\left(W_i.x_t + U_i.h_{t-1} + b_{\tilde{C}}\right) \quad (2)$$

$$\widetilde{C}_t = \tanh\left(W_{\tilde{C}}.x_t + U_{\tilde{C}}.h_{t-1} + b_{\tilde{C}}\right) \quad (3)$$

$$O_t = \sigma\left(W_o.x_t + U_o.h_{t-1} + b_o\right) \quad (4)$$

$$C_t = C_{t-1} \odot f_t + i_t \odot \widetilde{C}_t \quad (5)$$

$$h_t = O_t \odot \tanh\left(C_t\right) \quad (6)$$

f_t represents the forget gate at time step t. The sigma function in this equation gives a value from 0 to 1 that determines what information is removed from the network; if the value of the function is 0, then all information is removed, and vice versa if the function is equal to 1, all information is kept in the network. i_t represents the input gate, which determines the amount of information to feed into the network through the standby state equation $\widetilde{C}_t$. O_t represents the output gate, providing information for calculating outputs. Finally, two long-term memory states, C_t and short-term h_t, are updated. Equations (1) to (4) contain 12 parameters (W_f, U_f, b_f, W_i, U_i, b_i, $W_{\tilde{C}}$, $U_{\tilde{C}}$, $b_{\tilde{C}}$, W_o, U_o, b_o) that an LSTM network needs to train.

A time series prediction model based on the LSTM network structure needs to be implemented by 5 main hyperparameters, including the Number of Layers, the Number of Hidden Units, the Dropout Rate, the Batch Size, and the Number of Epochs. Hyperparameters significantly affect the model's accuracy, therefore, need to be optimized so that the LSTM model has the best results [9]. This study uses the method of hyperparameter optimization by Bayesian optimization. Besides, some general settings

were used for the models in this study, including The loss function used is Mean Square Error (MSE), the parameter initialization method of the model is the Xavier method [14], and the loss function optimization method used the Adam [17].

Hyperparameter optimization method

Before training the model, it is necessary to choose the hyperparameter setting because it affects the model's accuracy, but this is a time-consuming and confusing step [9, 18]. Optimization of hyperparameters is imperative in achieving the maximal accuracy of the model. Table 1 displays the hyperparameters that require determination in this study. The hyperparameter selection process's main challenge is finding the best set of hyperparameters among many different combinations that save computation time and cost. Grid search (GS) and random search (RS) are two popular methods widely used because of their simplicity and convenience [19, 20]. The operation of the above two search methods is depicted in Figures 2a and 2b. The advantage of these two methods is that they are simple, accessible, and give relatively good results compared to the ad-hoc selection method. Nevertheless, the GS method requires extensive computation time to evaluate all possible hyperparameter combinations. In contrast, the RS technique conserves computational resources by randomly exploring potential combinations among hyperparameters but may neglect optimal sets of hyperparameters.

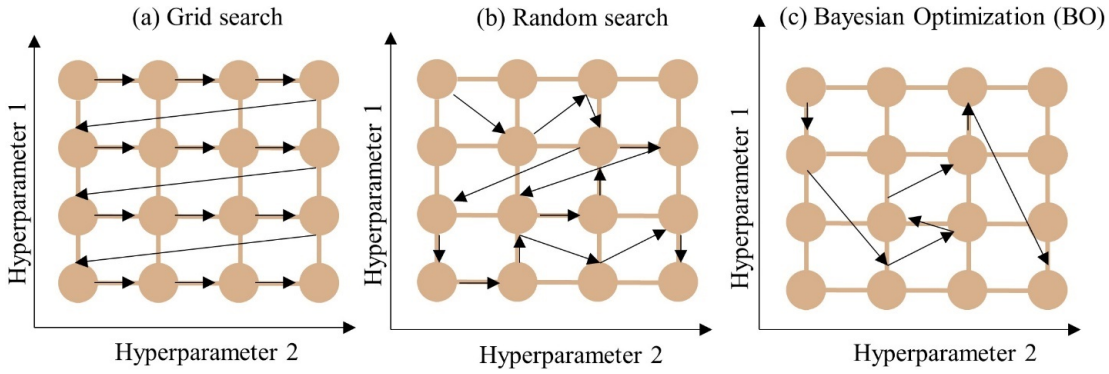

Figure 2. Hyperparameter optimization methods for machine learning models

Optimizing hyperparameters by the Bayesian method has outstanding advantages compared to the above two methods in terms of accuracy and efficiency, and this has been researched and proven in the field of hydrometeorology [21–24]. The Bayesian-based hyperparameter optimization process is shown in equation (7):

$$H^* = \arg\max f(H), H \in \mathbb{H} \quad (7)$$

which: H^* is the optimal set of hyperparameters H; $\mathbb{H}$ is the a priori distribution of H; and f is Gaussian processes ($\mathcal{GP}$). The Bayesian method uses Gaussian distribution for its hyperparameter optimization process. Details of this method are presented in a number of studies such as [22, 23].

Study area and dataset

The area of interest for this study is the end of the Hau river, which is also the Mekong estuary in Soc Trang province, which is likely to be heavily affected by the impacts of climate change (Fig. 3). According to the climate change scenario, if the sea level rises by 1 m, about 43.7% of the area of Soc Trang province will be affected by salinity intrusion and affecting about 450,000 people (35% of the total population of Soc Trang province). In addition, agricultural production accounts for a large proportion of the province's economic structure and people's income. For the above reasons, research on salinity intrusion forecasting and climate change response research is vital and urgent.

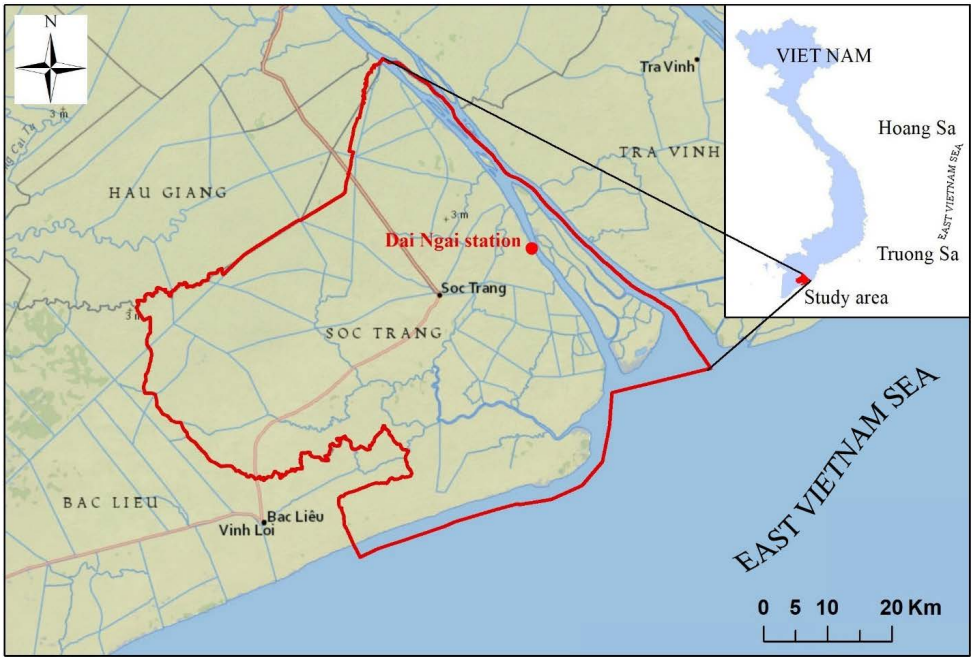

Figure 3. Study area

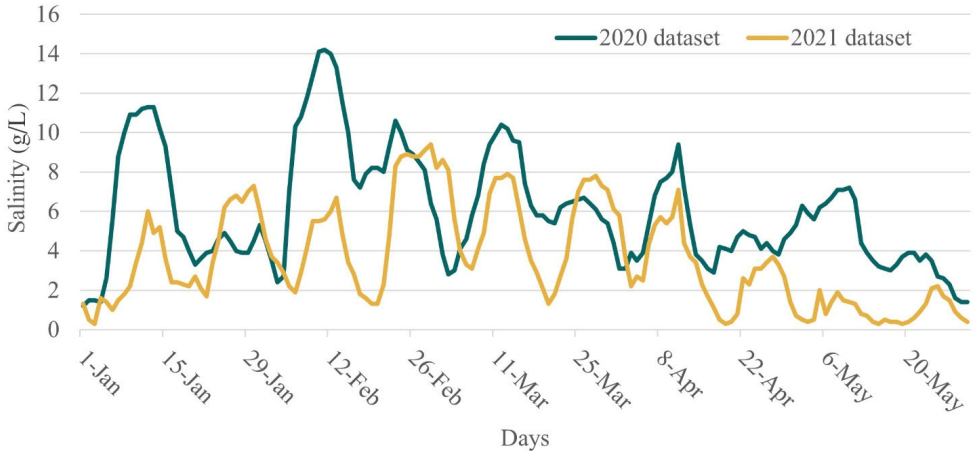

Figure 4. Salinity data (g/L) daily observed at Dai Ngai station, Soc Trang province in the period from January to June of 2020 and 2021

This study uses the highest salinity monitoring data (g/L) per day at Dai Ngai station, Soc Trang province, during the dry season of 2020 and 2021 (January to June). The data presented in Figure 4 of the study reveals a substantial discrepancy between the observation periods of 2020 and 2021. As a result, connecting the two data series for model training purposes may result in significant errors. The study builds two models for two data periods (2020 and 2021) to evaluate the method's effectiveness. Both models use 90% of the data length for training and 10% to test the model with a 7-day lead time.

In a time series forecasting model, data about the delay of that series (lag-time) plays an important role, affecting the model's accuracy [26]. Several methods can be used to select the amount of delay of the sequence as input to the model, such as Autocorrelation

(ACF) and Partial autocorrelation (PACF) [9, 26, 27]. The disadvantage of the above method is that although it calculates the correlation between the lagged time series and the original data series, it is impossible to choose how much the correlation threshold is enough to choose the number of lags. This study does not use the above methods but selects the number of time lags based on the test method to find the most suitable lag time for each model.

RESULTS AND DISCUSSION

Result of lag-time selection for input predictor

Figure 5 shows the accuracy of two LSTM models for two-time series 2020 and 2021 (calculated on the test dataset). The figure shows that with the 2020 dataset, the model will give the slightest RMSE error (about 0.32) and the highest R^2 index (about 0.7) when the lag time used is 7. Like the 2021 data with a lag time of 2, the model gives the smallest RMSE error (about 0.3) and the largest R^2 accuracy (about 0.63). Note that these two LSTM models are not hyperparameter-optimized; the hyperparameters of these models are randomly selected and fixed to evaluate the effect of lag time series. Thus, a time series with a lag time of 7 is used for 2020 data and 2 for 2021.

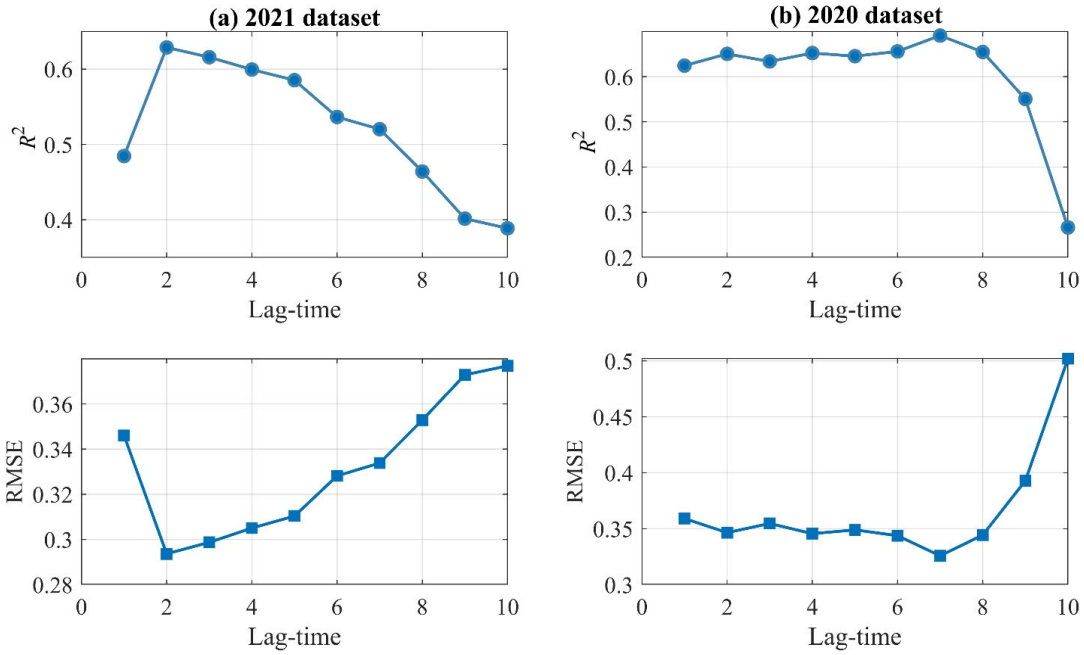

Figure 5. Optimal lag time for input predictors

The results of the optimal LSTM models

Table 1. The range of hyperparameters to optimize for the LSTM model

Hyperparameter	Range
The number of hidden layers	1–7
The number of hidden units	10–1,000
The number of epochs	10–1,000
Dropout rate	0.1–0.9
Batch size	2–2,048

With reliable input data selected in the above section, the LSTM model for two data series (2020, 2021) is set up with hyperparameters automatically optimized by the Bayesian method. The ranges of the hyperparameters' values are shown in Table 1.

Table 2 displays the mean values of the RMSE and the R^2, calculated based on the test data set, for two data series of the observation periods of 2020 and 2021. The evaluation was

conducted across different lead times ranging from 1 to 7. The results show that both models for 2 data series have relatively good accuracy with the 1-day forecast value and even with the 7-day forecast term; the model's accuracy is also acceptable.

Table 2. Results of optimal LSTM models for lead time from 1 to 7 days

Lead time (day)	2020		2021	
	RMSE (g/L)	R^2	RMSE (g/L)	R^2
1	1.16	0.90	0.29	0.89
2	1.42	0.89	0.52	0.76
3	1.49	0.86	0.64	0.60
4	1.36	0.85	0.73	0.59
5	1.48	0.82	0.82	0.57
6	1.66	0.81	0.79	0.50
7	2.08	0.72	0.85	0.51

Figures 6 and 7 compares the measured and predicted values in the test set of 1-day lead time to the 7-day lead time of the 2020 and 2021 data set, respectively. The findings highlight that the accuracy of predictions for a 1-day lead time is superior to those made for a 7-day lead time. However, it is noteworthy that the error associated with a 7-day lead time is not prohibitively large and still yields favorable outcomes.

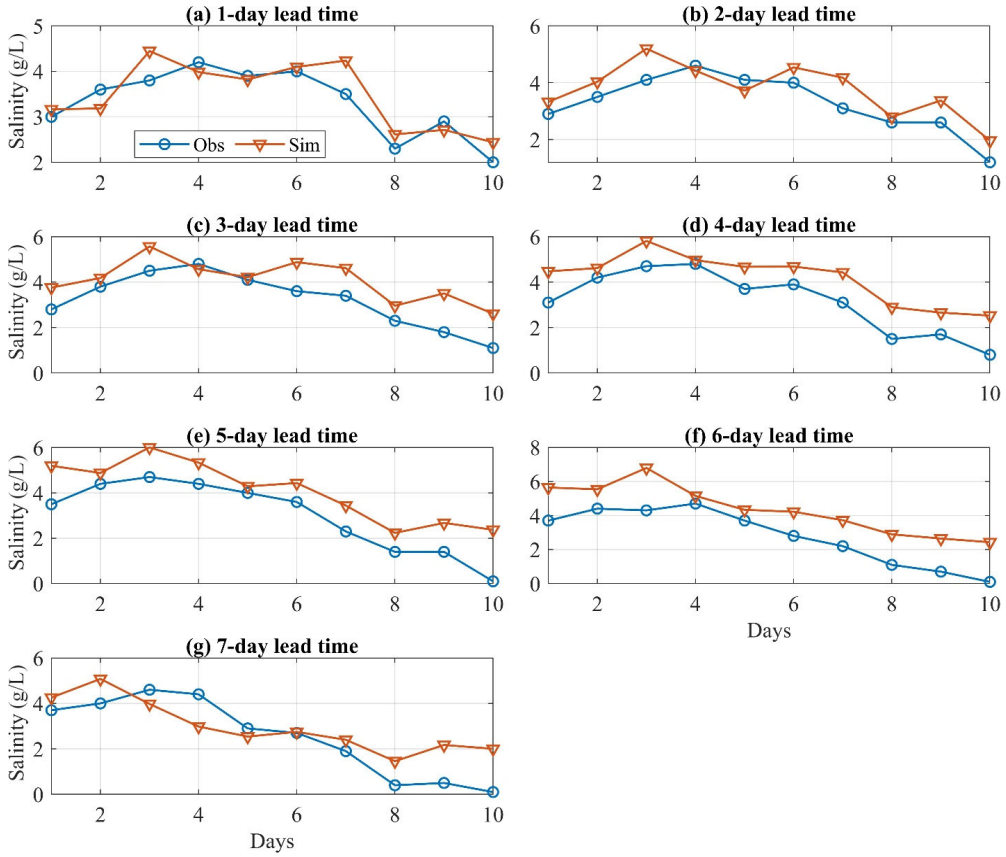

Figure 6. Comparison of actual value and forecasted value of 1-day lead time to 7-day lead time forecasting of the 2020 data set

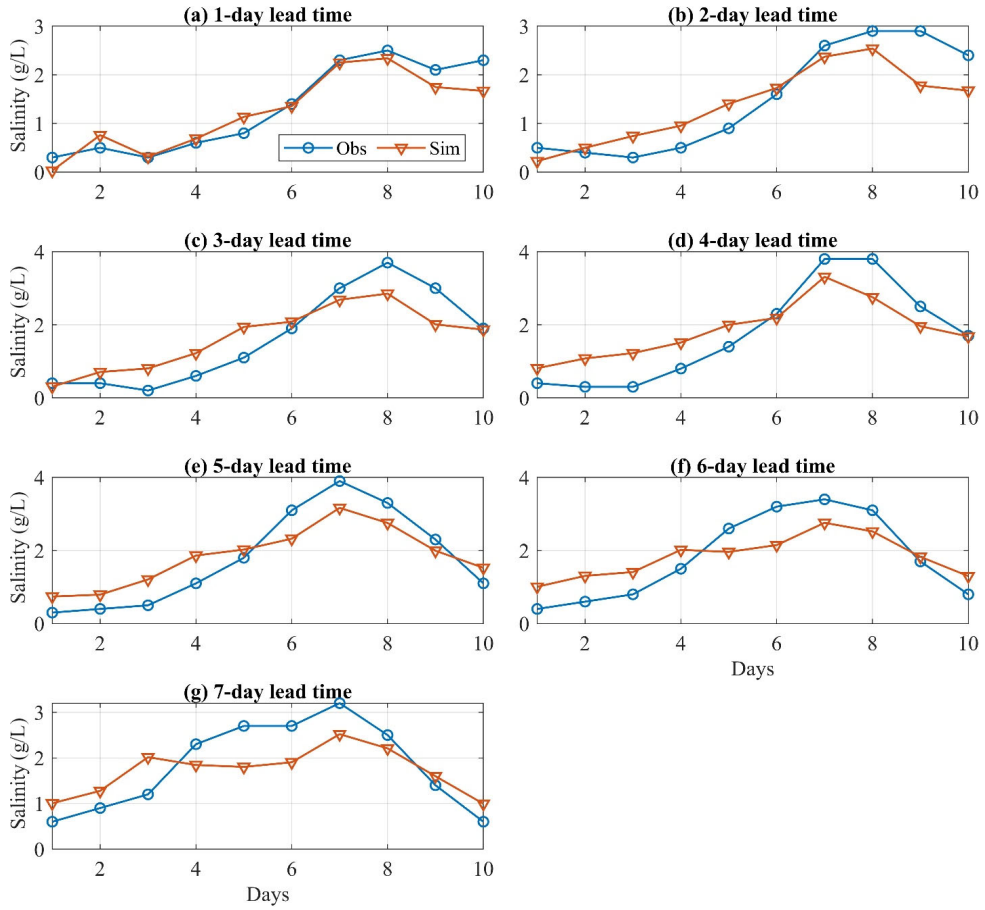

Figure 7. Comparison of actual value and forecasted value of 1-day lead time to 7-day lead time forecasting of the 2021 data set

CONCLUSION

Through the results of the prediction tests with a 7-day lead time using the optimal LSTM model, the research has proven the feasibility of the LSTM model for the salinity forecasting problem within 7 days. The study also shows that the Bayesian method works well for the hyperparameter optimization problem for the LSTM model. Furthermore, it has been demonstrated that optimizing input predictors can significantly impact on the accuracy of a predictive model. Based on the conclusions mentioned above, this study recommends that investigations utilizing LSTM models or other machine learning approaches carry out a meticulous calculation step and carefully select input parameters while also optimizing hyperparameters to enhance the accuracy of the predictive models. This study uses two separate LSTM models with two years of separate data without combining the two data series into one to evaluate the model. In the future, this study intends to conduct experiments utilizing numerous data series of dry seasons from multiple years in the past. The aim is to develop a model that can effectively extract valuable insights from historical data utilizing the LSTM model. Additionally, the study intends to investigate and test various techniques for enhancing the performance of the LSTM model in the context of salinity prediction.

Recently, there has been a remarkable surge in the interest surrounding physics-informed neural networks. This approach stands out due to its incorporation of hydrodynamic information in conjunction with neural networks, resulting in enhanced accuracy of

machine learning models [28–30]. Applying this methodology to predict saline intrusion in the Mekong delta holds great promise and will be a focal point of future research endeavors.

REFERENCES

[1] Soc Trang Statistical Yearbook, 2021. (in Vietnamese).

[2] Van Binh, D., Kantoush, S. A., Saber, M., Mai, N. P., Maskey, S., Phong, D. T., and Sumi, T., 2020. Long-term alterations of flow regimes of the Mekong river and adaptation strategies for the Vietnamese Mekong Delta. *Journal of Hydrology: Regional Studies*, *32*, 100742. https://doi.org/10.1016/j.ejrh.2020.100742

[3] Hoang Lam, D., Huy Phuong, N., Dinh Dat, N., Tien Giang, N., 2022. A setup of Mike 11 model for hydrological and saline intrusion forecast in Ben Tre province. *Vietnam Journal of Hydrometeorology*, *740*(1), 38–49. doi: 10.36335/vnjhm.2022(740(1)).38-49 (in Vietnamese).

[4] Quang Tri, D., 2016. Application Mike 11 model on simulation and calculation for saltwater intrusion for the Southern region. *Vietnam Journal of Hydrometeorology*, *671*, 39–46. (in Vietnamese).

[5] Van Dung, D., Dinh Phuong, T., Thi Oanh, L., Thanh Cong, T., 2018. The effectiveness of the Mike 11 ad model for forecasting and warning the salinity intrusion in the Mekong delta. *Vietnam Journal of Hydrometeorology*, *693*, 48–58. (in Vietnamese).

[6] Xiong, L., and Lu, Y., 2017. Hybrid ARIMA-BPNN model for time series prediction of the Chinese stock market. In *2017 3rd International conference on information management (ICIM)* (pp. 93–97). *IEEE*. doi: 10.1109/INFOMAN.2017.7950353

[7] Zhang, D., Peng, Q., Lin, J., Wang, D., Liu, X., and Zhuang, J., 2019. Simulating reservoir operation using a recurrent neural network algorithm. *Water*, *11*(4), 865. https://doi.org/10.3390/w11040865

[8] Ni, L., Wang, D., Singh, V. P., Wu, J., Wang, Y., Tao, Y., and Zhang, J., 2020. Streamflow and rainfall forecasting by two long short-term memory-based models. *Journal of Hydrology*, *583*, 124296. doi: 10.1016/j.jhydrol.2019.124296

[9] Tran, T. D., Tran, V. N., and Kim, J., 2021. Improving the accuracy of dam inflow predictions using a long short-term memory network coupled with wavelet transform and predictor selection. *Mathematics*, *9*(5), 551. https://doi.org/10.3390/math9050551

[10] Yoo, H. J., Kim, D. H., Kwon, H. H., and Lee, S. O., 2020. Data driven water surface elevation forecasting model with hybrid activation function—A case study for Hangang River, South Korea. *Applied Sciences*, *10*(4), 1424. https://doi.org/10.3390/app10041424

[11] Ahn, S., Tran, T. D., and Kim, J., 2022. Systematization of short-term forecasts of regional wave heights using a machine learning technique and long-term wave hindcast. *Ocean Engineering*, *264*, 112593. https://doi.org/10.1016/j.oceaneng.2022.112593

[12] Cong Thanh, N. and Tien Giang, N., 2022. Building LSTM (Long Short-Term Memory) machine learning model for water salinity forecasting in Dai Ngai. *Vietnam Journal of Hydrometeorology*, *740*(1), 98–104. doi: 10.36335/vnjhm.2022(740(1)).98-104 (in Vietnamese).

[13] Dau Hoang, N., Ngoc Tan, N., and Thi Hue, N., 2022. Building a warning and forecasting model by supervised machine learning method and testing saltwater intrusion prediction for Hau river basin. https://tainguyenvamoitruong.vn/xay-dung-mo-hinh-canh-bao-du-bao-theo-phuong-phap-hoc-may-co-giam-sat-thu-nghiem-du-bao-xam-ngap-man-cho-luu-vuc-song-hau-cid11297.html, accessed 01 March 2023 (in Vietnamese).

[14] Glorot, X., and Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and*

statistics (pp. 249–256). *JMLR Workshop and Conference Proceedings*.

[15] Cheng, H., Tan, P. N., Gao, J., and Scripps, J., 2006. Multistep-ahead time series prediction. In *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9–12, 2006. Proceedings 10* (pp. 765–774). *Springer Berlin Heidelberg*.

[16] Hochreiter, S., and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

[17] Kingma, D. P., and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[18] Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M., 2018. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

[19] Bergstra, J., and Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*, 281–305.

[20] Mantovani, R. G., Rossi, A. L., Vanschoren, J., Bischl, B., and De Carvalho, A. C., 2015. Effectiveness of random search in SVM hyper-parameter tuning. In *2015 international joint conference on neural networks (IJCNN)* (pp. 1–8). *IEEE*. doi: 10.1109/IJCNN.2015.7280664

[21] Brochu, E., Cora, V. M., and De Freitas, N., 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*. https://doi.org/10.48550/arXiv.1012.2599

[22] Saad, S., Javadi, A. A., Chugh, T., and Farmani, R., 2022. Optimal management of mixed hydraulic barriers in coastal aquifers using multi-objective Bayesian optimization. *Journal of Hydrology*, *612*, 128021. https://doi.org/10.1016/j.jhydrol.2022.128021

[23] Du, L., Gao, R., Suganthan, P. N., and Wang, D. Z., 2022. Bayesian optimization based dynamic ensemble for time series forecasting. *Information Sciences*, *591*, 155–175. doi: 10.1016/j.ins.2022.01.010

[24] Alizadeh, B., Bafti, A. G., Kamangir, H., Zhang, Y., Wright, D. B., and Franz, K. J., 2021. A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. *Journal of Hydrology*, *601*, 126526. doi: 10.1016/j.jhydrol.2021.126526

[25] Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., and Deng, S. H., 2019. Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40. doi: 10.11989/JEST.1674-862X.80904120

[26] Ahmad, S. K., and Hossain, F., 2019. A generic data-driven technique for forecasting of reservoir inflow: Application for hydropower maximization. *Environmental Modelling & Software*, *119*, 147–165. https://doi.org/10.1016/j.envsoft.2019.06.008

[27] Denić-Jukić, V., Lozić, A., and Jukić, D., 2020. An application of correlation and spectral analysis in hydrological study of neighboring karst springs. *Water*, *12*(12), 3570. https://doi.org/10.3390/w12123570

[28] Raissi, M., Perdikaris, P., and Karniadakis, G. E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

[29] Mahesh, R. B., Leandro, J., and Lin, Q., 2022. Physics informed neural network for spatial-temporal flood forecasting. In *Climate Change and Water Security: Select Proceedings of VCDRR 2021* (pp. 77–91). *Springer Singapore*.

[30] Liu, W., and Pyrcz, M. J., 2023. Physics-informed graph neural network for spatial-temporal production forecasting. *Geoenergy Science and Engineering*, *223*, 211486. doi: 10.1016/j.geoen.2023.211486