基于机器学习的长江口含沙量动态预测方法研究

涂俊豪

（交通运输部长江口航道管理局，上海 200003）

摘要：长江口作为中国水量最大河流的出海口，其含沙量变化和预测直接影响河口地区的生态环境、航道维护和防洪安全等。开发一个基于机器学习的模型，用于预测长江口区域的含沙量动态。考虑到含沙量受多种水文环境因素的影响，通过收集长江口区域 1 年的水文数据包括流速、潮位、含沙量等，运用时间序列分析方法，提取关键的特征和模式，选取长短期记忆网络(LSTM)对数据进行训练和测试。分析结果表明，基于 LSTM 的模型在预测长江口区域含沙量方面表现出了较高的准确性，模型的平均绝对误差为 0.146 5，决定系数为 0.931 4。

关键词：长江口；机器学习；含沙量；预测

中图分类号：U611　　**文献标志码：**A　　**文章编号：**1002-4972(2024)12-0206-06

Dynamic prediction method of sediment content in Yangtze River estuary based on machine learning

TU Junhao

(Yangtze Estuary Waterway Administration Bureau, MOT, Shanghai 200003, China)

Abstract: The Yangtze River estuary is the outlet of the largest river in China, the changes and predicions of sediment content directly affect the ecological environment, waterway maintenance and flood control safety in the estuary area. This study develops a machine learning-based model for predicting sediment content dynamics in the Yangtze River estuary region. Considering that sediment content is affected by a variety of hydrological environmental factors, this study collects hydrological data for a year in the Yangtze River estuary area, including flow velocity, tide level, sediment content, etc. and uses time series analysis methods to extract key features and patterns, and a long short-term memory network (LSTM) is selected, trained and tested. The analysis results show that the LSTM-based model shows high accuracy in predicting sediment content in the Yangtze River estuary region. The mean absolute error of the model is 0.146 5, and the coefficient of determination is 0.931 4.

Keywords: Yangtze River estuary; machine learning; sediment content prediction; hydrological parameter

近年来，随着全球气候变化和人类活动的加剧，长江口区域面临着严峻的环境挑战。长江口作为中国水量最大的河流的出海口，其水文特征的变化和预测直接关系到该区域的生态平衡、航道安全以及防洪措施等的有效性。机器学习技术，尤其是长短期记忆网络(long short-term memory network,LSTM)，因其在处理时间序列数据方面的优势，为高精度预测提供了新的可能性。近年来的研究进展表明，利用机器学习方法能有效地处理和预测复杂的水文和河流动态变化，尤其在数据缺失和异常检测方面展现出独特的价值。

Kulanuwat 等[1]开发了一种基于中位数的统计异常检测方法，使用滑动窗口技术针对水位数据进行异常检测，探索用于填补异常值的各种插值

收稿日期：2024-02-26

作者简介：涂俊豪（1995—），男，硕士，助理工程师，从事港口航道工程技术管理及研究。

技术，显示了在处理水文时间序列方面的潜力。研究特别突出了在非周期性数据上使用样条插值方法和在特定潮汐数据模式上使用 LSTM 模型的优越性能。另一项研究中，Ha 等[2]探讨了厄尔尼诺-南方涛动(ENSO)对流量和洪水发生的时空效应，表明将 ENSO 数据纳入机器学习模型可以提高流量预测的准确性，该研究强调了在机器学习模型中集成外部气候变量以更好地预测水文现象的价值。不少学者对于长江口的水文环境和动力条件均开展了研究。薛为[3]针对长江口的水文特征进行分析，揭示了该区域水动力条件的复杂性及其变化趋势；罗大松等[4]探讨长江口上海近岸海域的敌草隆分布特征及其生态风险，强调了对长江口生态系统进行持续监测的重要性。工程建设对长江口生态环境的影响也是近年来研究的重点之一。宋荣华等[5]通过数值模拟研究横沙东滩工程对长江口的影响，提出相关工程对水流和沉积物传输的潜在影响；付桂等[6]讨论了长江口水域现场监测技术的创新与实践，突出技术进步在环境监测中的应用。长江口水文要素的时空分布特征及其变化对于理解和管理该区域的环境具有重要意义。刘传杰等[7]对 2001—2017 年长江口南北港水文要素的变化进行深入分析，提供了有价值的长期变化数据；王淑楠等[8]进一步阐述了长江口水文要素的时空分布特征，为水资源管理提供参考。针对长江口潮位序列的非一致性和预测难题，张悦[9]基于混合分布提出一种新的频率计算方法，为潮位预测提供了新思路。在应对这些挑战的过程中，机器学习技术展现出巨大的潜力。Ren 等[10]及 Zhou 等[11]的研究分别展示了深度学习模型在模拟长江口悬浮沉积物浓度变化和洪水预测中的应用，突显了机器学习方法在水文模型和生态风险评估中的高效性和准确性。

尽管已有研究取得了一定进展，如何准确预测长江口含沙量动态变化仍是一个挑战，需要深入研究和探索。本研究致力于开发一种基于机器学习的模型，以预测长江口区域的含沙量动态。考虑到长江口含沙量的变化受多种水文环境因素的影响，本文通过收集长江口区域 1 年的水文数据，包括流速、潮位、含沙量等，利用时间序列分析方法提取关键的特征和模式。通过选取 LSTM 并进行训练和测试，旨在提高预测的准确性和可靠性，尤其对于监测数据缺失和异常情况的检测与补充，以期为长江口区域的河流管理和环境监测提供支持，填补现有研究的空白。

1　数据准备

研究采用的数据源自长江口水文泥沙监测站，涵盖 2018 年整年度的水文监测数据，如图 1 所示。该数据集包括流速、潮位、含沙量 3 个关键参数，每 10 min 记录 1 次。为了保证数据分析的精确度和可靠性，共收集 5.256 万个样本。这些数据直接反映了长江口区域在 2018 年内的水文环境变化情况，为研究提供了丰富的基础资料。

为了适应机器学习模型的要求，初步的数据准备工作包括数据清洗和特征选择。1）数据清洗过程排除缺失或异常的数据记录，确保分析的准确性。2）特征选择将流速、潮位、含沙量作为关键参数，是基于这些因素对河口含沙量变化的直接影响和内在联系。流速是影响河口含沙量的主要物理因素之一，直接决定了沙粒的输运能力和沉积速率；潮位是外海潮汐和上游径流相互作用的结果，其变化在一定策划程度上反映了潮汐力的作用，影响着河口区域的水动力环境，从而间接影响含沙量的分布和变化；含沙量本身是研究的核心对象，直接表征了河口区域悬浮泥沙的水平。

将数据集划分为训练集和测试集两部分，以验证模型的预测性能。采用 80%的数据作为训练集，用于模型的训练；剩余的 20%数据作为测试集，用于评估模型的准确性和泛化能力，如图 2 所示。该划分方法确保模型可以在未知数据上进行有效的预测，同时也能反映模型在实际应用中的可靠性。

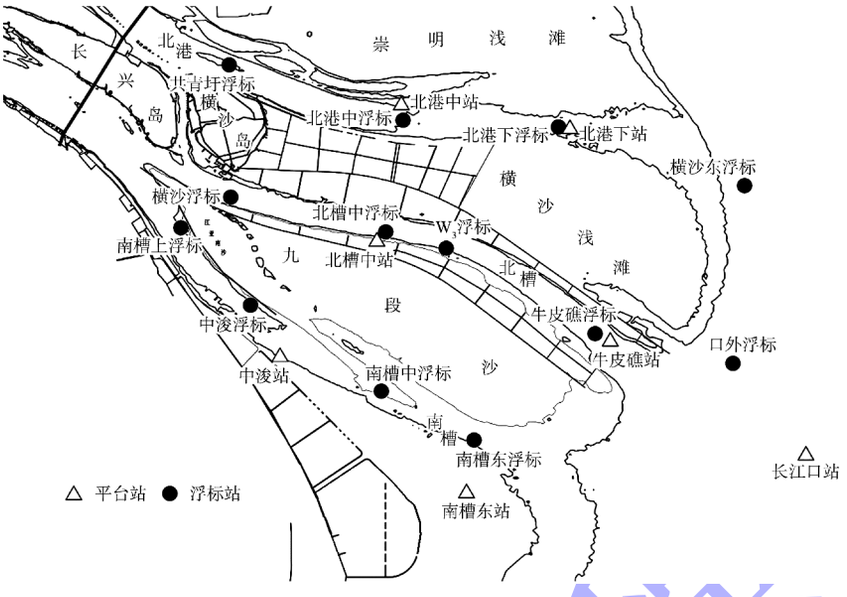
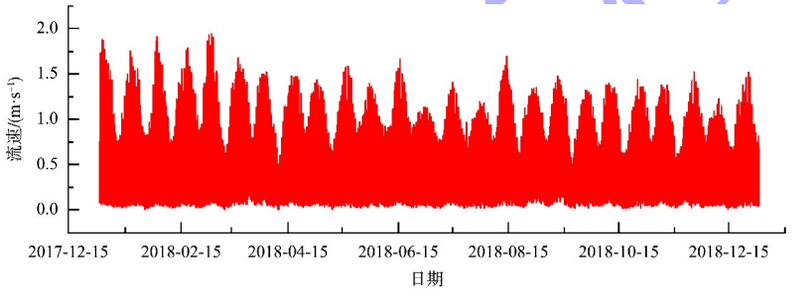
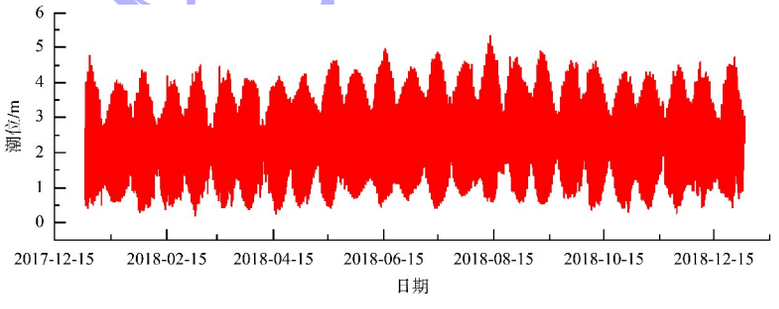


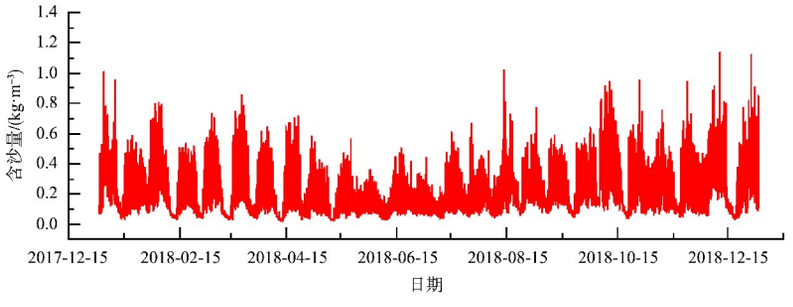
图 1　长江口水文、泥沙、波浪自动监测系统现场观测站点



a) 流速



b) 潮位



c) 含沙量

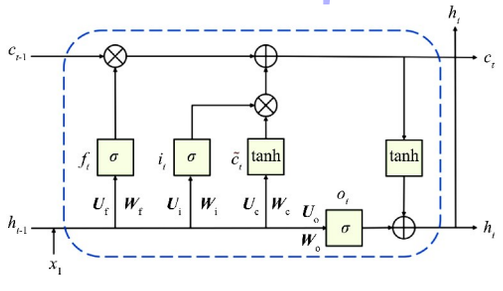
图 2　关键参数数据集

通过上述数据的准备和预处理工作，为使用 LSTM 进行含沙量动态预测研究奠定了基础。这些步骤不仅保证了数据质量，也优化了后续模型训练和测试过程，为达到研究目标提供了有力的支持。

2　预测模型

LSTM 是一种基于循环神经网络(recurrent neural network，RNN)发展优化的网络。RNN 的特点在于通过在序列上定义循环关系处理可变长度的输入序列，打破了传统神经网络中接收固定大小输出与提供固定大小输出的约束，从而为网络提取序列的时间关系特征提供了可能性。LSTM 最大的优势在于能够解决传统 RNN 面临的长期依赖问题，即在处理长序列数据时能够有效记住信息，使得 LSTM 在语音识别、自然语言处理、时间序列预测等领域表现出色。在自然语言处理领域，LSTM 在语言模型、情感分析和机器翻译等任务上取得了突破性进展，如 Sutskever 等[12]利用 LSTM 实现了端到端的机器翻译系统，显著提高了翻译质量。在时间序列预测方面，LSTM 被广泛应用于股票市场预测、气象数据分析等任务，如 Fischer 等[13]展示了 LSTM 在股票价格预测上的有效性。

本文选择 LSTM 作为预测模型的核心，通过引入 3 种不同的门结构(遗忘门、输入门和输出门)实现对信息流的精细控制，如图 3 所示。



注：x_t 和 h_t 分别为 t 时刻的输入和输出（或隐状态）；i_t 为 t 时刻的输入门；f_t 为 t 时刻的遗忘门；o_t 为 t 时刻的输出门；c_t 为 t 时刻的单元状态；$\tilde{c}_t$ 为候选状态，其基于 t-1 时刻的隐状态和 t 时刻的输入，通过 tanh 函数变换；σ 为 sigmoid 函数；W 为输入权重，$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$ 分别为与上一时刻的隐含状态 h_{t-1} 相乘的权重矩阵，$\boldsymbol{W}_f$ 为遗忘门对应的权重矩阵，$\boldsymbol{W}_i$ 为输入门对应的权重矩阵，$\boldsymbol{W}_c$ 为候选记忆单元的权重矩阵，$\boldsymbol{W}_o$ 为输出门对应的权重矩阵；U 为状态权重，$\boldsymbol{U}_f$、$\boldsymbol{U}_i$、$\boldsymbol{U}_c$、$\boldsymbol{U}_o$ 为与当前输入 x_t 相乘的权重矩阵，$\boldsymbol{U}_f$ 为遗忘门对应的权重矩阵，$\boldsymbol{U}_i$ 为输入门对应的权重矩阵，$\boldsymbol{U}_c$ 为候选记忆单元对应的的权重矩阵，$\boldsymbol{U}_o$ 为输出门对应的权重矩阵。

图 3　LSTM 单元

遗忘门负责决定哪些信息应该被丢弃，输入门控制新信息的加入量，而输出门决定下一状态的输出。这些门结构的协同工作使得 LSTM 能够在长序列中有效地保存和访问有用信息，避免了长期依赖中的梯度消失或爆炸问题，从而提高了模型在时间序列数据分析、语言处理和其他需要记忆长期信息的任务中的性能和可靠性[14]。

遗忘门的结构为：

$$f_t=\sigma(\boldsymbol{W}_f[x_t,h_{t-1}]+b_f) \tag{1}$$

门单元更新可以表示为：

$$i_t=\sigma(\boldsymbol{W}_f[h_{t-1},x_t]+b_i) \tag{2}$$

数据更新可以表示为：

$$\tilde{c}_t=\tanh(\boldsymbol{W}_c[h_{t-1},x_t]+b_c) \tag{3}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\tilde{c}_t \tag{4}$$

$$o_t=\sigma(\boldsymbol{W}_o[h_{t-1},x_t]+b_o) \tag{5}$$

$$h_t=o_t\odot\tanh(c_t) \tag{6}$$

式中：b_i、b_f、b_o、b_c 分别为输入门、遗忘门、输出门和候选记忆单元的偏置项；⊙代表矩阵元素依次相乘。

LSTM 在处理时间序列数据方面具有显著优势[15-16]。与传统的机器学习模型相比，LSTM 能够有效处理和记忆长期依赖关系，对于预测受多种因素影响且具有时间序列特征的含沙量变化尤为关键。利用 LSTM 模型能够捕捉到流速、潮位等因素随时间变化对含沙量影响的内在动态模式。

将 3 个输入变量的时间序列数据进行归一化处理，使其值落在同一尺度，以避免训练过程中的梯度问题。为此，对数据进行 Z-score 标准化：

$$Z=(X-\mu)/\sigma \tag{7}$$

式中：X 为原始数据，μ 为原始数据的均值，σ 为原始数据的标准差。

模型预测结果是标准化的，可以通过逆运算将其转换回原始值：

$$X=Z\sigma+\mu \tag{8}$$

标准化可以使得这类特殊的数据适应深度学习模型的输入，并确保数据在整体上呈现零均值和单位方差，从而助力模型更快收敛。此外，标准化还可以增强模型训练的稳定性，避免因原始数据的尺

度或分布差异导致的梯度消失或爆炸问题。

模型使用过去 6 h 数据预测未来 1 h 的含沙量，每个输入窗口包含连续 36 个时间步的数据作为输入(X)和随后 1 h(即未来 6 个时间步)的含沙量作为输出(Y)。模型包含 4 个 LSTM 层，后接 2 个全连接层，最终的输出层用于预测未来 1 h 的含沙量，如图 4 所示。

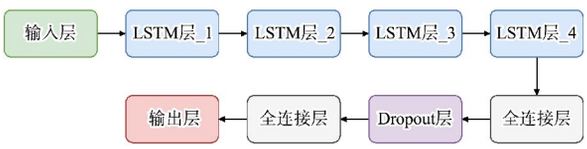


图 4 预测模型结构

模型构建过程包括输入层、多个 LSTM 层和输出层。输入层接收标准化后的流速、潮位、含沙量数据。模型设计了多个 LSTM 层来增强模型的学习能力，每层旨在捕捉数据中不同时间尺度的依赖关系。输出层则负责预测未来的含沙量。在模型构建中，通过调整 LSTM 层的数量和每层的神经元数量，以达到最佳的预测性能，见表 1。模型训练采用回归任务中常用的平均绝对误差 M 作为损失函数，利用 Adam 优化器进行参数优化，Adam 优化器是一种用于深度学习应用中的梯度下降算法，它结合了动量和 RMSprop 两种优化算法的特点。Adam 优化器通过计算梯度的一阶矩估计(即均值)和二阶矩估计(即未中心化的方差)调整每个参数的学习率，因此能够自适应地调整各参数的学习率，使其适用于不同的问题和数据集。训练过程中，通过调整学习率和批量大小，找到了 1 组最优的训练参数，确保模型训练的效率和预测的准确性。

表 1 超参数设置值

学习率	批量大小	LSTM 层数	LSTM 单元数	优化器	损失函数	Epoch 数量
0.001	64	4	100	Adam	M	100

模型的评估基于 2 个主要指标：M 和相关系数 R。M 提供了预测值与实际值之间差异的量化度量，而 R 指标则反映了模型预测值的变异量占总变异量的比例，即模型的解释能力。2 个指标共同衡量了 LSTM 模型在预测长江口区域含沙量方面的性能，确保了评估过程的全面性和准确性。

3 结果分析

如图 5 所示，训练损失和验证损失随着迭代次数的增加而逐渐下降，表明模型在学习过程中持续改进。训练损失的持续降低表明模型在拟合训练数据方面变得更加有效；验证损失的下降和稳定表明模型具有良好的泛化能力，未出现明显的过拟合现象。在 83 个 Epochs 后，2 个损失值趋于稳定，说明模型训练达到了收敛状态。

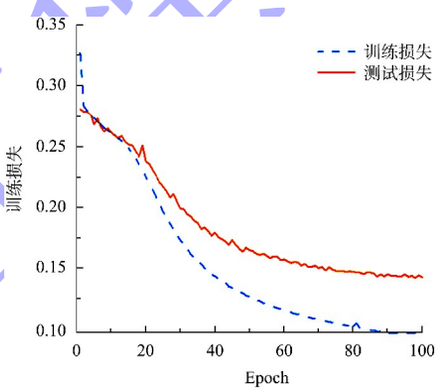


图 5 训练损失曲线

图 6 为模型预测值与实际值的比较。可以看出，模型的预测曲线与实际含沙量的动态吻合度较高，反映了模型对于时间序列数据内在模式的有效捕捉能力。M 为 0.146 5，这一低误差水平进一步证明了模型预测的准确性。此外，决定系数 R^2 为 0.931 4，表示预测值与实际值有较高的拟合度。

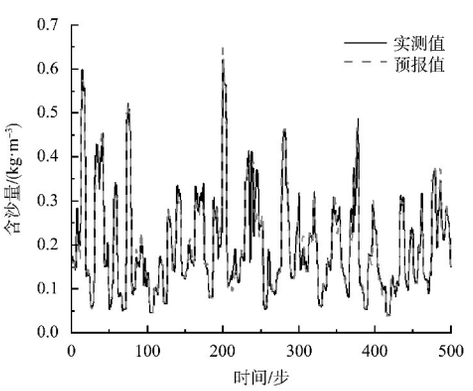


图 6 预测结果

4　结论

1）本研究开发并验证了 1 个基于长短期记忆网络(LSTM)的机器学习模型，用于预测长江口区域的含沙量变化。模型设计考虑了各种环境变量的时间序列数据的特点，通过 1 年期的数据训练和测试，展示了模型在捕捉和预测含沙量变化方面的能力。模型的平均绝对误差为 0.146 5，决定系数为 0.931 4，表明模型具有较好的预测能力。

2）研究成果对于水文学和河口工程领域的实践应用具有参考意义。通过精确预测含沙量的动态变化，可以更有效地进行河口区域的环境监测、资源管理和工程规划，尤其对于监测数据的补缺和异常检测具有重要意义。此外，研究方法和发现为使用深度学习技术处理和分析水文数据提供了视角。

3）后续研究可以集中于整合更多类型的环境因素，以进一步提高预测的准确性和模型的鲁棒性。此外，对模型进行长期的性能监测和调整，确保其在多变的环境条件下仍能保持高效的预测能力，也是未来工作的一个重要方向。

参考文献：

[1] KULANUWAT L, CHANTRAPORNCHAI C, MALEEWONG M, et al. Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series[J]. Water, 2021, 13(13): 1862.

[2] HA S, LIU D R, MU L. Prediction of Yangtze River streamflow based on deep learning neural network with El Niño-Southern Oscillation [J]. Scientific reports, 2021, 11(1): 11738.

[3] 薛为. 长江口南港北槽水域水文特征分析[J]. 水运工程, 2024(2): 107-112, 119.

[4] 罗大松, 杨红, 王春峰, 等. 长江口上海近岸海域敌草隆空间分布特征及生态风险评估[J]. 上海海洋大学学报, 2024, 33(1): 150-160.

[5] 宋荣华, 唐建华. 横沙东滩(六～八期)工程对长江口影响的数值模拟研究[J]. 水电能源科学, 2023, 41(11): 26-30.

[6] 付桂, 刘栋, 李为华. 长江口水域现场监测技术的创新与实践[J]. 中国水运, 2022(S2): 88-95.

[7] 刘传杰, 李保, 李昌生, 等. 2001—2017 年长江口南北港水文要素变化分析[J]. 水利水电快报, 2022, 43(4): 27-30, 37.

[8] 王淑楠, 顾峰峰, 李俊花, 等. 长江口水文要素时空分布特征[J]. 水运工程, 2022(3): 85-92.

[9] 张悦. 基于混合分布的非一致性长江口潮位序列频率计算方法[J]. 水电能源科学, 2020, 38(8): 26-28, 78.

[10] REN Z D, LIU C J, OU Y F, et al. Deep learning-based simulation of surface suspended sediment concentration in the Yangtze Estuary during Typhoon In-Fa[J]. Water, 2024, 16(1): 146.

[11] ZHOU L W, KANG L. A comparative analysis of multiple machine learning methods for flood routing in the Yangtze River[J]. Water, 2023, 15(8): 1556.

[12] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 12: 3104-3112.

[13] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions [J]. European journal of operational research, 2018, 270(2): 654-669.

[14] 孙瑞奇. 基于 LSTM 神经网络的美股股指价格趋势预测模型的研究[D]. 北京: 首都经济贸易大学, 2016.

[15] JIA X Y, JI Q Y, HAN L, et al. Prediction of sea surface temperature in the East China Sea based on LSTM neural network[J]. Remote sensing, 2022, 14(14): 3300-3300.

[16] SHRUTHI R K, SAGAR B P. An LSTM-Based Approach to Predict Stock Price Movement for IT Sector Companies [J]. International journal of cognitive informatics and natural intelligence, 2021, 15(4): 1-12.

（本文编辑　王传瑜）