基于机器学习模型的兰州市月降水量预测研究

沈梓诣　班文超
（浙江海洋大学海洋工程装备学院，舟山 316000）

摘要：月降水量的精确预测对于国民生产、防灾减灾有重大意义，然而单独的模型难以完成准确预测降水的任务。本文把自适应噪声完备集合经验模态分解（CEEMDAN）分别与误差反向传播模型（BP）和长短期记忆神经网络（LSTM）结合起来，以兰州市降水数据为例，与单一的LSTM模型、差分整合移动平均自回归模型（ARIMA）和BP模型的性能进行比较。结果表明，两个复合模型有效提高了观测值和预测值的拟合度，克服了峰值预测精度不高的问题，显著优于对比模型。

关键词：兰州市，月降水量，预测，CEEMDAN，LSTM，ARIMA，BP

中图分类号：P4　　**文献标志码：**A　　**DOI：**10.3969/j.issn.2095-1973.2024.03.010

Research on Monthly Precipitation Prediction in Lanzhou City Based on Machine Learning Model

Shen Ziyi, Ban Wenchao
(School of Ocean Engineering Equipment, Zhejiang Ocean University, Zhoushan 316000)

Abstract: Accurate forecasting of monthly precipitation is of great significance for national production as well as disaster prevention and mitigation. However, it is difficult for a single model to complete the task of accurate precipitation prediction. We combine CEEMDAN with the error reverse communication model (BP) and with the long short-term memory neural network (LSTM), respectively. And we compare Lanzhou's precipitation data with the precipitation predictions with a single LSTM model, ARIMA model and BP model. The research results show that the two composite models effectively improve the fitting of observation values and prediction values. Thus, the problem of low accuracy in peak prediction is overcomed, showing significantly higher performance than the comparative models.

Keywords: Lanzhou City, monthly precipitation, prediction, CEEMDAN, LSTM, ARIMA, BP

0　引言

月降水量的准确预测对于国民生产有重大意义，不论是在农业生产、水资源利用还是防灾减灾方面[1-3]。因为气象过程多变和随机因素复杂，导致精确地预测出降水量的难度很大[4-8]。最近几年，由于深度机器学习技术的发展，给降水预测研究带来了全新的方式[9-12]。如支持向量机、差分整合移动平均自回归模型（ARIMA）等开始运用在降水预报中并取得了良好的成效。不过这些方法也很难确定与降水量数值间的长期依赖关系，无法保证预测精度。

方楠等[13]于1997年首次提出长短期记忆神经网络（LSTM）。它是在循环神经网络（RNN）的基础上，专门针对于处理时间依赖问题而设计的。LSTM的模型中添加了单元状态和门结构来控制信息的传递，可以通过复杂性结构的多非线性变换处理方法对数据分析进行高度的抽象化设计，同时支持对时间序列信息的长短期记忆，具有良好的时间序列问题解决功能，可应用于降水量的预测。总的来说，LSTM模型的预报准确度超过一般模型，但因为降水的不确定性，在有些时刻会发生极端降水现象，这个时候单纯的LSTM模型也很难精确预报雨量序列的突变，会有较大的误差。将自适应噪声完备集合经验模态分解（CEEMDAN）算法[14-17]与LSTM模型结合，能够有效解决对于突变数据的预测，得到更高精度的预测结果。CEEMDAN算法会将信号分解成一系列IMF分量和Res残差，显著降低降水数据的复杂性，而后使用LSTM模型进行预测的难度会有所下降。本文拟构建一个CEEMDAN-LSTM系统模型，对兰州市月降水量进行预测；与此同时，为检验CEEMDAN分解算法的可靠性，又构建了另一个BP模型结合CEEMDAN算法的复合模型，即CEEMDAN-BP模型，并将这两个复合模型与LSTM模型、ARIMA模型和BP神经网络模型的预测结果进行对比，检验其优越性。

收稿日期：2022年11月11日；修回日期：2023年3月28日
第一作者：沈梓诣（2000— ），Email：imshenziyi@163.com

1 研究区概况

兰州市是甘肃省的省会城市，位于我国西北地区的东部。兰州市属于中温带大陆性气候，由于降水大多集中于夏季，而其余季节的降水极少，所以很容易发生干旱灾害。本文选取兰州站1960年1月—2020年12月的月均降水量资料，共有732组降水数据，采样间隔为1个月，对其中前90%的降水数据进行了模拟训练，后10%的降水数据用于检验预测结果。

兰州市月降水量序列如图1所显示，由图可见兰州市降水量具有很大的多变性和波动性。

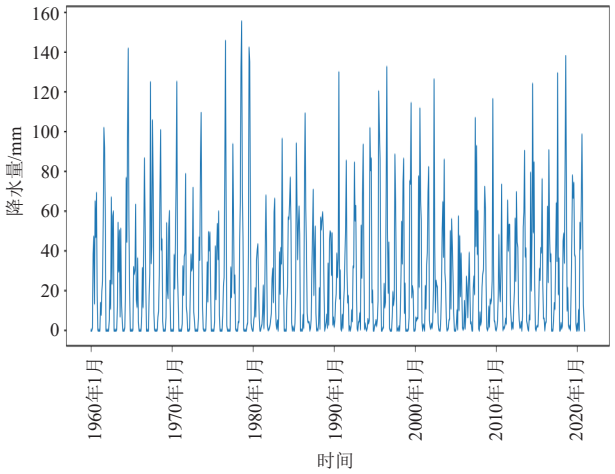

图1 1960—2020年兰州市月降水量序列

Fig. 1 Monthly precipitation series in Lanzhou during 1960—2020

2 研究方法

2.1 CEEMDAN算法

CEEMDAN算法多用于预测，它可以把原始数据进行分解，将波动强烈的原始数据分解成数个较为平稳的成分数据，降低数据的多变性，以此达到提高预测精度的效果。这种算法是集合经验模态分解（EEMD）算法和经验模态分解（EMD）算法的继续发展算法，很好地解决了EEMD中存在的模态混叠现象。该算法一方面增加了经过EMD分解的IMF成分，代替了原先直接把高斯白噪声信号放到原始信号中的方法，另一方面抛弃了原始的总体平均的计算方法，选择在得到一阶IMF成分后就进行总体平均计算，生成最终的一阶IMF成分，而后再对剩下部分重复以上方法，有效解决了白噪声的转移传递问题。

设$E_i(\cdot)$为经过EMD分解后得到的第i个本征模态成分；经过CEEMDAN分解后获得的第i个本征模态成分为$\overline{C_i(t)}$；V^j表示符合标准正态分布的高斯白噪声信号，$j=1,2,\cdots;N$为加入白噪声的频率；ε为标准的白噪声；$y(t)$为待分解数据。CEEMDAN分解步骤如下。

1）将高斯白噪声添加到待分解信号$y(t)$上得到新信号$y(t)+(-1)^q\varepsilon^j(t)$，其中$q=1, 2$。对新信号进行EMD分解，得到一阶本征模态成分：

$$E(y(t)+(-1)^q\varepsilon v^j(t))=C_1^j(t)+r^j。\quad(1)$$

2）把产生的第N个模态成分加以总体平衡，得出CEEMDAN分解的第1个本征模态成分：

$$\overline{C_1(t)}=\frac{1}{N}\sum_{j=1}^{N}C_1^j(t)。\quad(2)$$

3）计算减去第一个模态成分后的残差：

$$r_1(t)=y(t)-\overline{C_i(t)}。\quad(3)$$

4）在$r_1(t)$中加入正负成对的高斯白噪声的新信号，用新信号作为媒介开始EMD分解，可以得出一阶模态成分D_1，从而生成CEEMDAN分解的第2个本征模态成分：

$$\overline{C_2(t)}=\frac{1}{N}\sum_{j=1}^{N}D_1^j(t)。\quad(4)$$

5）计算结果减去第二个模态成分后的残差：

$$r_2(t)=r_1(t)-\overline{C_2(t)}。\quad(5)$$

6）重复以上过程，直到所获取的残差信号一直是单调函数，并且无法进一步分解，则计算完毕。此时得到的本征模态成分的数量为K，则原始信号被分解为：

$$y(t)=\sum_{k=1}^{K}\overline{C_k(t)}+r_K(t)。\quad(6)$$

2.2 LSTM模型

LSTM有较强的延时记忆数据的能力，在其掌握当前内容的时候，会获取更长时间跨度数据信息间的关系，达到长期记忆，能够对事物的发展做出比较精准的预测。一个LSTM细胞有3个门，分别叫做遗忘门（f），输入门（i）和输出门（O）。图2为神经元结构的单元体。

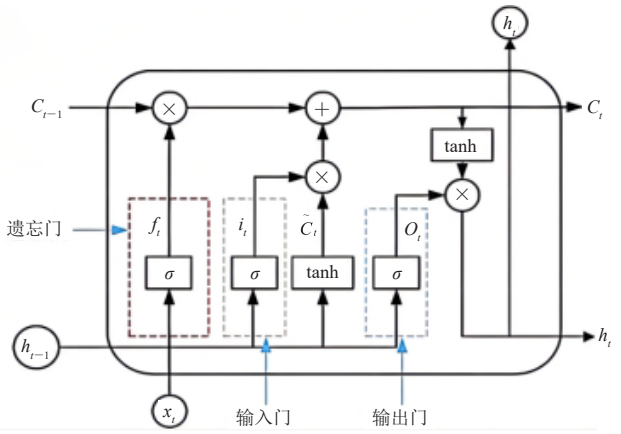

图2 LSTM模型的神经元结构

Fig. 2 Neuronal structure of LSTM model

$$f_t=\sigma(W_f*[h_{t-1},\ x_t]+\boldsymbol{b}_f),\quad(7)$$

式中，f_t为遗忘门；σ为标准的sigmoid激活函数；h_{t-1}表示上一时间的单元数值；W_f为权重数值；x_t为时间为t时的输入值；$\boldsymbol{b}_f$为偏置值的向量。

$$i_t = \sigma(W_i * [h_{t-1},\ x_t] + \boldsymbol{b}_i), \tag{8}$$

$$\tilde{C}_t = \tanh(W_c * [h_{t-1},\ x_t] + \boldsymbol{b}_c), \tag{9}$$

式中，i_t为输入门；$\tilde{C}_t$是目前输入的状态单元；tanh(·)是双曲正切激活函数。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{10}$$

式中，C_t为隐含层在t时刻的状态单元；f_t为上一次遗忘关于信息C_{t-1}的程度；i_t为要将$\tilde{C}_t$加入的程度，最后得到了本细胞的状态C_t。

$$O_t = \sigma(W_0[h_{t-1}, x_t] + \boldsymbol{b}_0), \tag{11}$$

$$h_t = O_t * \tanh(C_t), \tag{12}$$

式中，O_t为当前的输出门。

本文建模输入时间步为数据集前90%的数据步长，输出时间步为10步，进行累次叠加，输出总量为数据集后10%的数据。

本文所采用的LSTM模型层数为3层。第一、二、三层的神经元个数分别为50、50、100，使用的优化器类型为Adam，激活函数为relu，迭代次数是300次，学习率为0.01。本文建模主要使用Python，学习包为TensorFlow。

2.3 CEEMDAN的两组复合模型

因为降水量数据存在着很大的多变性和波动性，用简单的模型很难准确地预测出降水数值。本文选择先使用CEEMDAN算法把原本波动性很大的降水数据分解成相对较为平稳的几个部分；再通过BP模型和LSTM网络分别预测分解后每个部分的数值；最后叠加求和还原成降水量的预测值。其中CEEMDAN-LSTM模型的主要运行过程如图3所示。

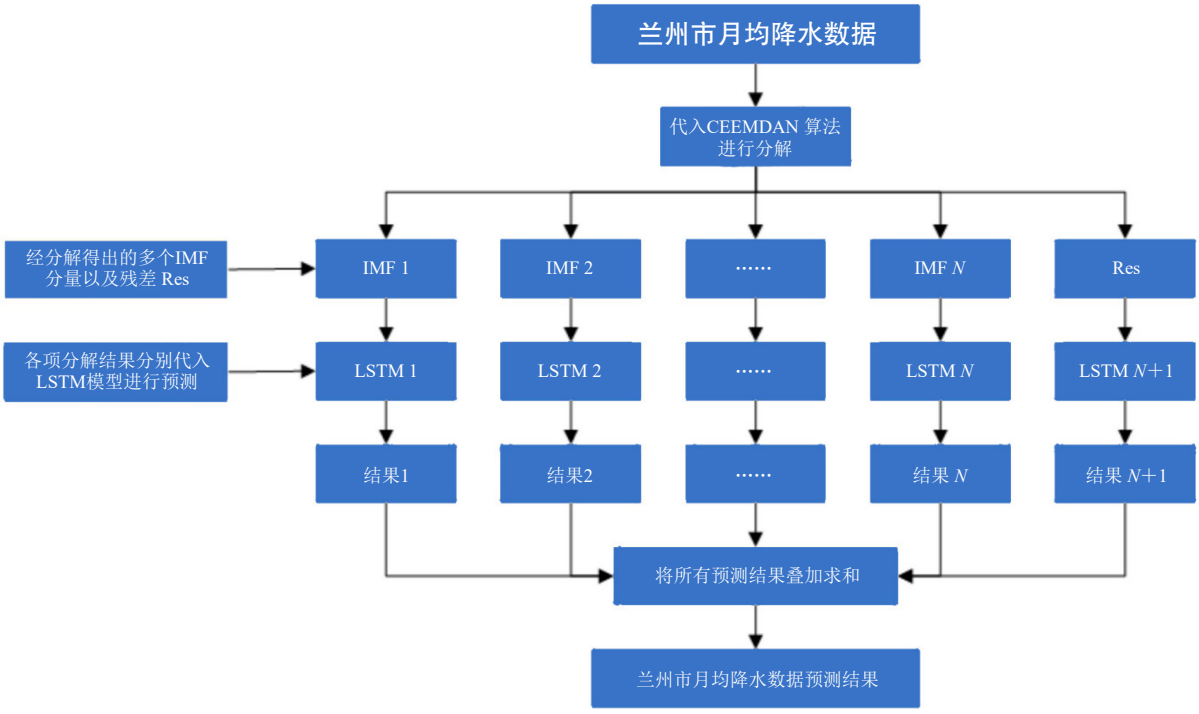

图3 CEEMDAN-LSTM模型运行流程图

Fig. 3 Flow chart of CEEMDAN-LSTM model run

2.4 ARIMA模型

ARIMA模型又叫做差分整合移动平均自回归模型，是最广泛使用的需求预测模型之一。ARIMA模型是由自回归（AR）模型和滑动平均（MA）模型所构成。它与自回归移动平均（ARMA）模型同属于自回归模型，但是二者对于数据的要求大相径庭。ARMA模型适用于平稳时间序列的数据，而ARIMA模型则是更加适用于差分后为平稳时间序列的数据。ARIMA模型会把所需预测的原始时序数据看作一个等长随机的数据集，然后使用数学的方法去识别这个数据集的特征，并且用数模的方式描述它，建模完毕后便可以达成通过已知数值预测未来数值的效果。图4为ARIMA模型的流程图。

ARIMA(p，d，q)模型的表达式：

$$(1-\sum_{i=1}^{p}\varphi_i L^i)(1-L^i)X_t=(1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t, \quad (13)$$

$$d=(1-L^i), \quad (14)$$

式中，L为滞后算子；p，d，q是模型中的三个参数，p代表ARIMA模型中原始数据的滞后数，d代表数据需要进行几阶差分化达到稳定，d必须为正整数参数，q代表预测模型中采用预测误差的滞后数；φ代表自回归系数；θ代表移动平均系数；X_t代表时间t的序列值；ε_t表示时间t的误差项。

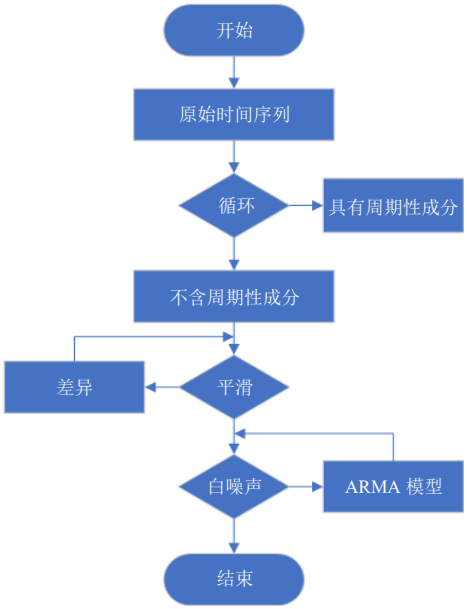

图4 差分整合移动平均自回归模型（ARIMA）流程图
Fig. 4 Flow chart of autoregressive integrated moving average model (ARIMA) with difference

2.5 BP模型

BP神经网络，即误差反向传播算法，是一类使用误差反向传播训练的多层前馈神经网络，是最常用的神经网络架构之一。其网络结构可分为输入层、隐含层和输出层。其实质内容是将样本数据从输入输出的问题转化为非线性优化问题，并采用一定的方法使权值沿着误差函数的负趋势变化。其数学表达式：

$$\begin{aligned} y=&w_{11}^{(2,3)}*\text{tansig}(w_{11}^{(1,2)}*x_1+w_{21}^{(1,2)}*x_2+b_1^{(2)})+\\ &w_{21}^{(2,3)}*\text{tansig}(w_{12}^{(1,2)}*x_1+w_{22}^{(1,2)}*x_2+b_2^{(2)})+\\ &w_{31}^{(2,3)}*\text{tansig}(w_{13}^{(1,2)}*x_1+w_{23}^{(1,2)}*x_2+b_3^{(2)})+b_1^{(3)} \end{aligned} \quad (15)$$

式中，w为权重值；b为阈值；$w_{11}^{(2,3)}$表示该权重值为第二层第1个节点到第三层第1个节点的权重值；$b_1^{(2)}$表示这个阈值是第二层第1个节点的阈值。

2.6 模型评价指标

为简单地比较模型的参数，可以使用均方根误差(RMSE)、均方误差（MSE）、平均绝对误差（MAE）和拟合度（R^2）来判断模型的预测精度，表达式：

$$\text{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}, \quad (16)$$

$$\text{MSE}=\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2, \quad (17)$$

$$\text{MAE}=\frac{1}{m}\sum_{i=1}^{m}|y_i-\hat{y}_i|, \quad (18)$$

$$R^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2}, \quad (19)$$

式中，m为试验样本的数量；y_i和z_i分别为降水量的观测值和预测值。

3 结果与分析

3.1 单一的LSTM模型

采用单一的LSTM模型的降水预测结果见图5。由图可见，采用单独的LSTM模型可以预测出未来降水量的基本走势，但是在波峰处会有较大的误差，尤其是在出现极端降水的情况下，很难预测出高精度的降水数据。

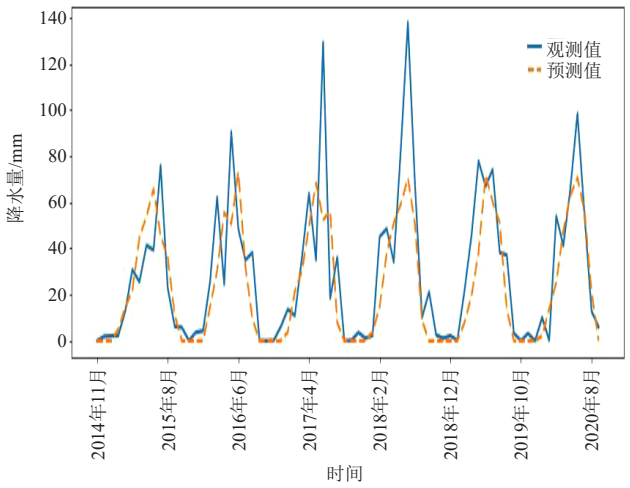

图5 2014—2020年单一LSTM模型降水预测结果
Fig. 5 Precipitation predictions with a single LSTM model during 2014–2020

3.2 ARIMA模型

采用ARIMA模型的降水预测结果见图6。由图可见，采用ARIMA模型同样也只能预测出未来降水数据的基本走势，波峰处的观测值和预测值并不能很好地拟合，对于极端降水情况无法有效预测。

3.3 BP模型

采用BP模型的降水预测结果见图7。由图可见，采用BP模型同样也只能预测出未来降水数据的基本走势，波峰处的观测值和预测值并不能很好地拟合，对

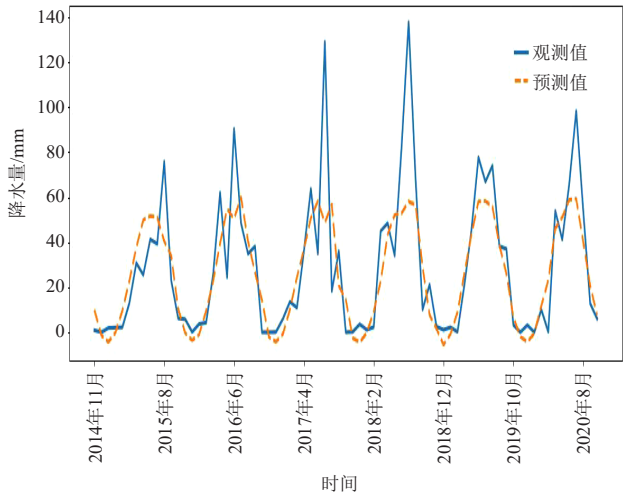

图6 2014—2020年ARIMA模型降水预测结果

Fig. 6 Precipitation predictions with ARIMA model during 2014–2020

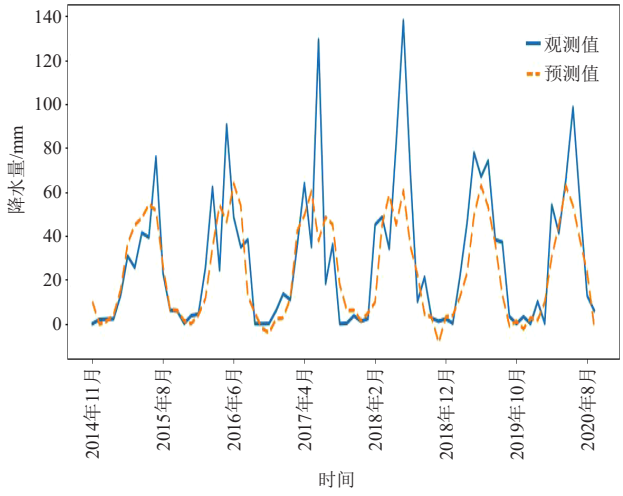

图7 2014—2020年BP模型降水预测结果

Fig. 7 Precipitation predictions with BP model during 2014–2020

于极端降水情况无法准确预测。

3.4 CEEMDAN的两组复合模型

首先对原始数据进行CEEMDAN分解，共得出7个IMF成分和1个趋势项，分解信号见图8。图中IMF1～IMF7分别为7组IMF分量，Res为残差。7个模态成分的频率依次下降，但是波动性明显减少。和原始降水量数据比较，建模的更为简单。特别是原始降水量数据的峰值变化较大，分布不对称且出现了较多的极端值，相较而言，现在各成分的分布接近于对称，且极端数值也更少，因此能够给BP模型和LSTM模型带来更稳定的输入。

采用CEEMDAN-BP和CEEMDAN-LSTM模型的降水预测结果分别见图9。由图可见，CEEMDAN-BP

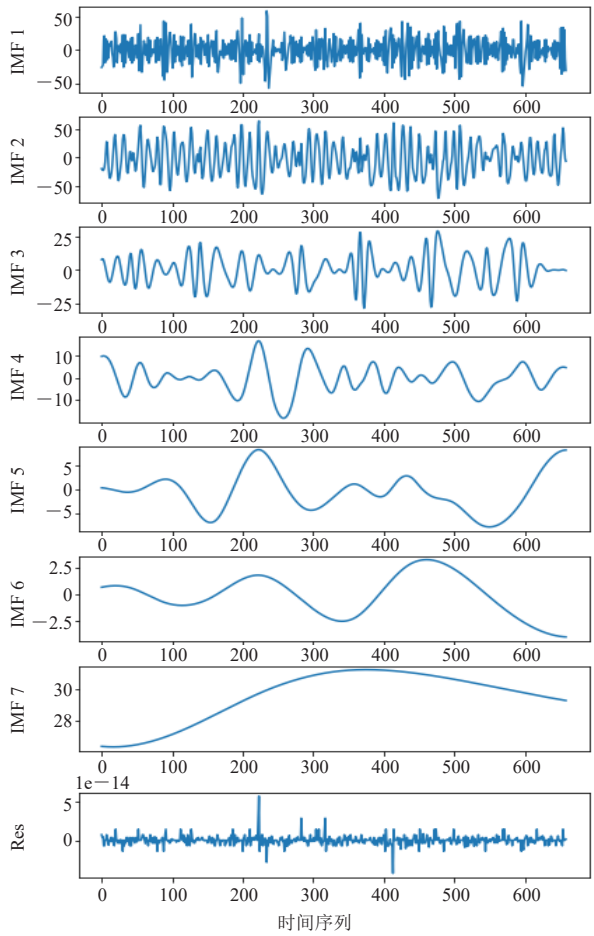

图8 兰州市降水量CEEMDAN分解结果

Fig. 8 CEEMDAN decomposition results of precipitation in Lanzhou

和CEEMDAN-LSTM可以较好地预测出未来降水量的走势，尤其在波峰处产生的误差也很小，说明在出现极端降水的情况下，也能比较好地预测出未来的降水量。

3.5 误差分析

使用RMSE、MSE、R^2和MAE作为评估模型的指标（表1），复合模型的各项误差指标相较于单模型都有较大的提升，其中CEEMDAN-BP模型的RMSE值较LSTM模型、ARIMA模型和BP模型分别降低了20.2%、20.9%和22.5%；CEEMDAN-LSTM模型的RMSE值较LSTM模型、ARIMA模型和BP模型分别降低了15.9%、

表1 各模型的预测误差对比

Table 1 Comparison of prediction errors from different models

模型	LSTM	ARIMA	BP	CEEMDAN-BP	CEEMDAN-LSTM
RMSE/mm	21.50	21.69	22.14	17.15	18.09
MSE/mm	462.25	470.46	490.18	294.13	327.25
R^2	0.60	0.59	0.57	0.71	0.71
MAE/mm	14.13	14.29	14.31	12.92	12.43

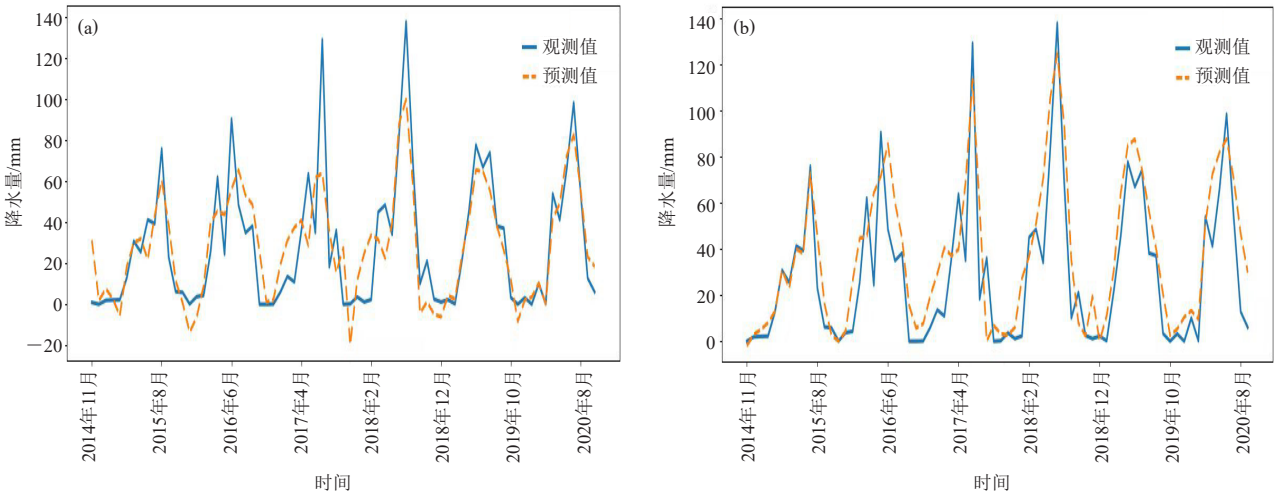

图9 2014—2020年CEEMDAN-BP模型（a）和CEEMDAN-LSTM模型（b）降水预测结果

Fig. 9 Precipitation predictions with CEEMDAN-BP model (a) and CEEMDAN-LSTM model (b) during 2014-2020

16.6%和18.3%；并且CEEMDAN的两个复合模型的R^2相较于三个单模型更接近于1，误差更小，预测结果更精准。

4 结论

1）运用ARIMA模型、LSTM模型和BP模型单独对兰州市降水进行预测时，预测结果与观测值的基本趋势大体一致，但不能解决极端降水难以准确预测的问题。为克服此问题，本文对兰州市的降水数据进行了CEEMDAN算法分解，使得原本极端的数据波动性显著减小且分布更均匀，降低了下一步进行模型预测的难度；然后使用BP模型和LSTM模型对分解出来的各个部分分别进行预测且叠加求和。结果表明，复合模型提高了对极端降水预测的精度，与实际观测值之间拟合良好，总体性能较好。

2）CEEMDAN的引入可以使得原始降水数据得到更好的处理，有效分解出更为稳定的模态分量，从而提高模型的预测精度，为降水预测提供了新的研究思路和方法。未来可以考虑把CEEMDAN算法引入到更多数据的处理中，运用更多资料验证该算法。

参考文献

[1] 曹枝俏, 王国利, 梁国华, 等. 基于随机模拟信息的神经网络洪水预报模型[J]. 水力发电学报, 2010, 29(4): 63-69.
[2] 常新雨, 周建中, 方威, 等. 黄龙滩水库中长期径流预报方法研究[J]. 水力发电, 2021, 47(8): 10-14, 93.
[3] 叶陈雷, 徐宗学, 雷晓辉, 等. 城市社区尺度降雨径流快速模拟——以福州市一排水小区为例[J]. 水力发电学报, 2021, 40(10): 81-94.
[4] 沈皓俊, 罗勇, 赵宗慈, 等. 基于LSTM网络的中国夏季降水预测研究[J]. 气候变化研究进展, 2020, 16(3): 263-275.
[5] 方巍, 庞林, 王楠, 等. 人工智能在短临降水预报中应用研究综述[J]. 南京信息工程大学学报(自然科学版), 2020, 12(4): 406-420.
[6] 李智强, 邹红霞, 齐斌, 等. 基于EEMD-ARIMA的年降水预测拟合模型研究[J]. 计算机应用与软件, 2020, 37(11): 46-50, 78.
[7] 孟锦根. 基于PSO-LSSVM的干旱区中长期降水预测模型研究[J]. 长江科学院院报, 2016, 33(10): 36-40.
[8] 沙世琨. 基于随机森林算法的陈垓灌区降水量预测模型[J]. 水利技术监督, 2020(5): 134-137.
[9] 黄春艳, 韩志伟, 畅建霞, 等. 基于EEMD和GRNN的降水量序列预测研究[J]. 人民黄河, 2017, 39(5): 26-28.
[10] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM[J]. Chaos, Solitons & Fractals, 2020, 140: 110212.
[11] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[12] Deng L, Yu D. Deep learning: methods and applications[J]. Foundations and Trends in Signal Processing, 2014, 7(3/4): 197-387.
[13] 方楠, 谢国权, 阮小建, 等. 长短期记忆神经网络(LSTM)模型在低能见度预报中的应用[J]. 气象与环境学报, 2022, 38(5): 34-41.
[14] 顾云青, 苏玉香, 沈晓群, 等 基于改进的CEEMDAN排列熵和GWO-SVM的滚动轴承故障诊断[J]. 组合机床与自动化加工技术, 2022(8): 62-66.
[15] 肖俊青, 金江涛, 岳敏楠, 等. 改进CEEMDAN算法与分形融合的深度学习轴承故障分析[J]. 动力工程学报, 2022, 42(6): 522-529.
[16] 沈富鑫, 邴其春, 张伟健, 等. 基于CEEMDAN-ABC-LSTM组合模型的短时交通流预测[J]. 青岛理工大学学报, 2022, 43(5): 96-103, 119.
[17] Hu Y J, Ouyang Y, Wang Z L, et al. Vibration signal denoising method based on CEEMDAN and its application in brake disc unbalance detection[J]. Mechanical Systems and Signal Processing, 2023, 187: 109972.

（编辑：郑秋红）