

基于VMD-HMM-BiLSTM的降雨量预测

王 欣, 李俊刚

北方工业大学理学院, 北京

收稿日期: 2025年6月22日; 录用日期: 2025年7月12日; 发布日期: 2025年7月25日

摘 要

降雨量预测对农业生产、水资源管理及气象灾害防御具有重要意义。本文主要从地理单元全覆盖、气候类型代表性和观测需求导出, 选择河北省的8个气象站点(保定、乐亭、青龙、邢台、石家庄、蔚县、围场和张家口), 时间范围为1974年1月1日至2023年12月31日的日数据, 经过预处理, 建立VMD-HMM-BiLSTM模型, 并与BiLSTM、VMD-BiLSTM模型进行对比。实验结果表明: 1) 变分模态分解(VMD)方法的引入有效降低了降雨量数据的非平稳性, 减轻了模型在提取时序特征的压力, 与单一模型BiLSTM模型相比, R^2 最大有效提升率达21%; 2) VMD-HMM-BiLSTM模型具有更低的误差, R^2 整体平均值达到0.825, 说明HMM与BiLSTM模型的融合也能有效提升降雨量预测的精度, 体现了组合模型在降水量预测方面的优势。

关键词

BiLSTM, 变分模态分解, 隐马尔可夫模型, 降水量预测, 组合模型

Rainfall Forecasting Based on VMD-HMM-BiLSTM

Xin Wang, Jungang Li

College of Science, North China University of Technology, Beijing

Received: Jun. 22nd, 2025; accepted: Jul. 12th, 2025; published: Jul. 25th, 2025

Abstract

Rainfall prediction is of great significance for agricultural production, water resource management, and meteorological disaster prevention. This article mainly focuses on the comprehensive coverage of geographical units, representativeness of climate types, and observation needs. Eight meteorological stations in Hebei Province (Baoding, Leting, Qinglong, Xingtai, Shijiazhuang, Weixian, Weichang, and Zhangjiakou) were selected, with daily data from January 1, 1974 to December 31, 2023. After preprocessing, a VMD-HMM BiLSTM model was established and compared with BiLSTM and VMD

BiLSTM models. The experimental results show that: 1) The introduction of the variational mode decomposition (VMD) method effectively reduces the non-stationary nature of rainfall data and reduces the pressure of the model in extracting time series features. Compared with the single model BiLSTM model, the maximum effective lifting rate of R^2 is 21%; 2) The VMD-HMM BiLSTM model has lower errors, with an overall average R^2 of 0.825, indicating that the fusion of HMM and BiLSTM models can effectively improve the accuracy of rainfall prediction, reflecting the advantages of the combined model in precipitation prediction.

Keywords

BiLSTM, VMD, Hidden Markov Model, Precipitation Prediction, Combinatorial Model

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

降水是地球水循环的核心环节,而降雨量作为衡量降水的关键指标,通常指在特定时间段内降落在水平地表上的未经蒸发、渗透或流失的水层深度,其单位为毫米。精准的降雨量预测不仅能够为农业生产提供灌溉决策依据,避免因降水不足导致的干旱减产或过量降水引发涝渍灾害;还能在水资源管理部门优化水库调度、城市防洪排涝规划、生态环境修复等提供关键的数据支撑。因此,进行降雨量预测的研究有利于把握未来降水趋势,对破解区域水资源制约瓶颈、保障社会经济高质量发展和生态环境可持续性具有不可替代的战略意义。

传统降水预测模型主要基于统计学、动力学理论等基础方法,通过分析历史气象数据、物理机制或者变量间的相关性从而实现降水量的预测。但它只能进行中小尺度的降水量预测、局部区域气候分析。郭廓[1]对阜新地区建立 ARIMA 模型进行降雨量的预测,通过对模型的适用性分析发现 ARIMA 模型适用于降水量较高月份的预测,且预测评估误差在 $\pm 15\%$ 以内,但不适合较低月份的降水量预测。陈聪[2]指出数值天气预报(numerical weather prediction, NWP)可以通过模拟大气运动的过程来预测灾害性气象,但该方式需要大型设备辅助,预测成本很高,并且模式仿真与实际情况存在误差,需要进行相关修正才能满足预报精度。在此之后,机器学习、深度学习的引入减少了复杂的计算成本,Khan [3]指出新兴的机器学习技术因其能够提高降雨预测的准确性和灵活性。由于降雨时间序列的特征包含非线性和非平稳,单一的模型很难完全捕捉到气象序列中的非线性因子,以至于模型的预测能力有所降低。所以 Menggang K [4]通过实验和分析发现,与 ARIMA、BP 和 LSTM 等单个模型相比,所提出的组合模型在所有预测步骤中的预测精度都得到了显著提高。隐马尔可夫模型(HMM)作为一种机器学习模型,将其与其他模型混合能提高模型的预测能力,且混合模型在多个领域运用广泛。陈子浩等人[5]建立 Bayesian-NHMM 有效地识别出京津冀地区日降雨的季节性变化规律,显著提高了模型的预测精度。李方圆等人[6]运用 HMM-XGBoost 模型分析了国内的股票市场,发现混合模型与其他基础模型相比确实效果有所提升。

基于众多研究发现,当前降水量预测模型仍面临诸多挑战,首先是传统模型在处理非平稳、非线性的降水量数据时存在特征提取不足的问题,导致预测精度低;其次是现有的模型通常无法同时考虑时间序列过去和未来的时间信息,从而限制了模型的预测能力;再者是单一的模型难以兼顾降雨量的多尺度特征与状态依赖性;最后是隐马尔可夫模型与深度学习模型的组合在降水量预测应用研究上研究略少。因此本文以月降水量为研究对象,建立 BiLSTM 模型、VMD-BiLSTM 模型、VMD-HMM-BiLSTM 模型

来分析未来月降水量的变化趋势, 将结果进行对比分析, 找到效果较好的模型, 进行后续的应用。

2. 数据介绍和预处理

2.1. 研究区域

河北省地处华北平原北部, “环绕京畿”, 是连接东北、华北、西北的重要枢纽, 总面积约 18.88 万平方千米, 是中国唯一兼具高原、山地、丘陵、平原、湖泊和海滨的省份, 境内水资源跨海河、滦河等四大流域, 海岸线长 487 千米。1) 该省份港口众多, 交通网络发达, 是“京津冀协同发展”核心区域; 2) 气候属温带大陆性季风气候, 气候要素南北、东西差异明显, 光照热量条件优越; 3) 降水时空分布不均, 地形影响显著, 年际波动大, 旱涝灾害频发, 在 2023 年 7 月河北受冷暖空气和台风杜苏芮的共同影响, 大部分地区出现强降雨过程; 4) 2024 年末常住人口 7378 万人, 地区生产总值达 47526.9 亿元, 三大产业协调发展, 装备制造业增长迅速, 第三产业对经济增长贡献率最高。

2.2. 数据集的构建

本文主要研究河北省的 8 个气象站点, 选择依据主要从地理单元全覆盖、气候类型代表性和观测需求导出, 分别为保定、乐亭、青龙、邢台、石家庄、蔚县、围场和张家口。数据来自 NCEI, 该数据集的气象指标包括: 平均温度、平均露点、平均海平面压力、平均观测站压力、平均能见度、平均风速、最大持续风速、最高气温、最低气温、降水量、积雪深度的气象情况指示, 见表 1。选择数据的时间范围为 1974 年 1 月 1 日至 2023 年 12 月 31 日, 共 163,935 条数据。其中训练集与测试集划分比例为 7:3。

Table 1. Description of structural equation model variables

表 1. 气象指标解释说明

气象指标	变量说明	单位
TEMP	平均温度	华氏度
DEWP	平均露点	华氏度
SLP	平均海平面压力	毫巴/百帕
STP	平均观测站压力	毫巴/百帕
VISIB	平均能见度	英里
WDSP	平均风速	节
MXSPD	最大持续风速	节
TEMPDIF	最高温度与最低温度之差	华氏度
PRCP	降水量	英寸

2.3. 数据预处理

由于天气复杂多变, 检测到的数据会由于测量误差包含一些缺失数据以及异常数据, 为了使训练模型得到更好的效果, 利用 Python 工具进行数据预处理步骤。

1) 数据清洗。通过生成完整日期序列并重新索引, 实现时间序列完整性处理; 紧接着识别并删除缺失值异常占比为 80%的气象指标 STP、SLP, 对连续型变量运用时间序列插值法[7]填充缺失值, 收尾残留缺失值采用线性插值法[8]完善; 最后使用四分位距(IQR)法检测异常值, 将超出 $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ 区间的数据视为离群值, 针对降水量(PRCP)的极端高值, 通过百分位数截断策略将超过 99%分位数的观测值截断, 以提升模型性能。

2) 数据归一化。在进行数据归一化前, 先对原始数据中非标准单位进行转换, 包括华氏度转摄氏度、

英里转公里等, 并将数值型变量保留三位小数以统一精度; 同时, 将日尺度数据聚合为月尺度, 降水量按月累加, 其余气象指标通过先累加再除以当月天数获取平均值, 消除天数差异干扰。由于数据量纲差异会影响模型拟合与预测精度, 采用归一化(Min-Max Scaling)方法, 基于线性变换将原始数据映射到[0, 1]区间, 公式为:

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中, $x_{\max}$ 代表原始数据中的最大值, $x_{\min}$ 代表原始数据中的最小值。

2.4. 降水量影响因素分析

2.4.1. 降水量自身影响

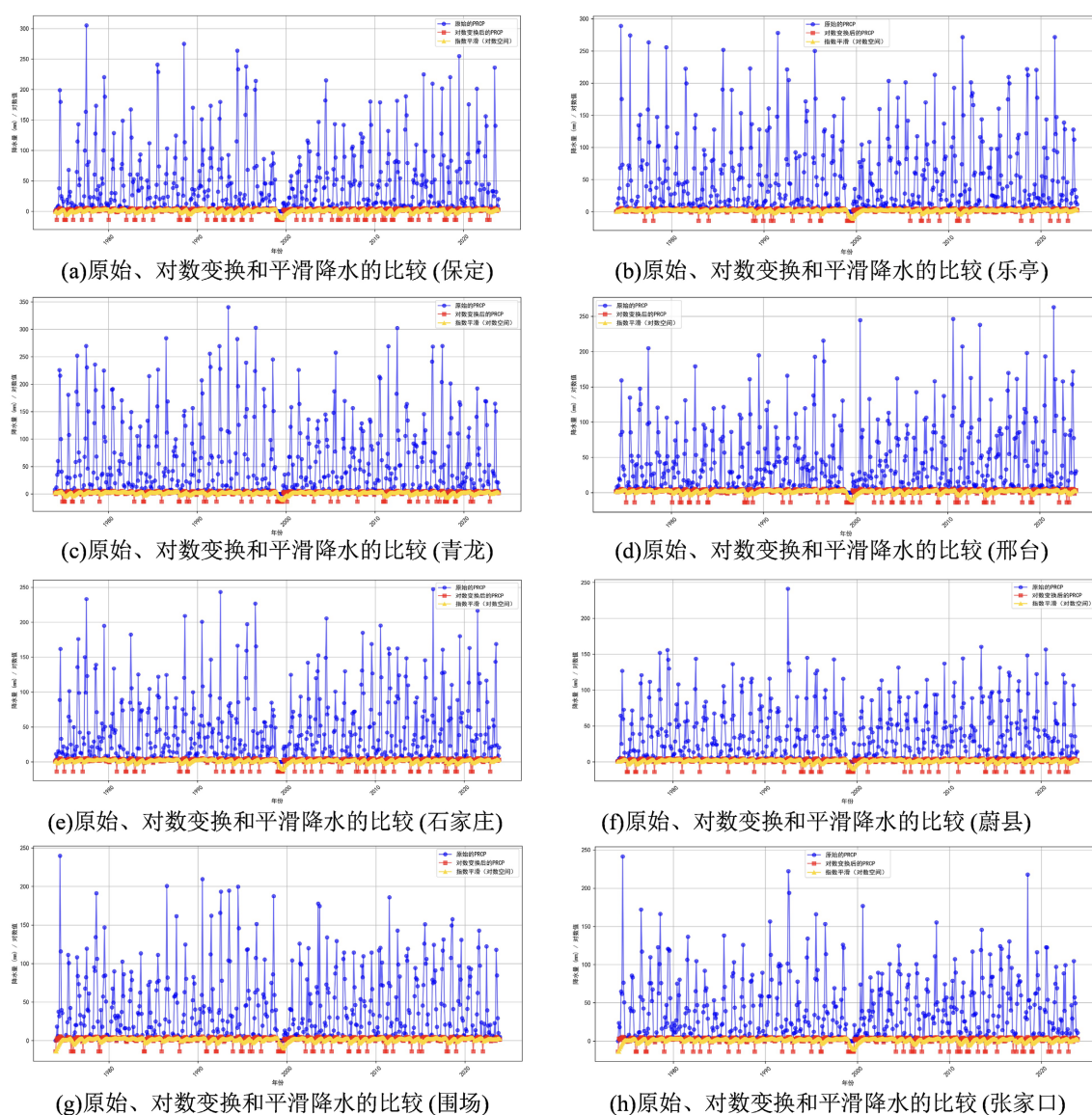


Figure 1. Comparison of original, logarithmic transformed and smoothed precipitation at 8 stations

图 1. 8 个站点原始、对数变换和平滑降水的比较

为探究降水量自身因素影响, 本文对降水量数据进行对数变换和指数平滑处理, 由于对数变换要求数据为正实数, 而原始数据中存在降水量为 0 的情况, 为避免对数变换时出现数学错误, 将降水量小于等于 0 的值替换为一个极小的正数(1×10^{-6}), 保证数据对数变换的条件, 同时最小化对原始数据特征的影响。进一步, 在对数变换过程中, 由于对数变换能够有效压缩数据的尺度, 降低极端高降水量值对整体数据分布的影响, 使数据分布更加均匀, 稳定数据的方差, 从而为后续分析提供更具稳定性和规律性的数据形态, 便于揭示数据的潜在特征。最后进行简单指数平滑方法对对数变换后的数据进行处理。

$$S_t = \alpha y_t + (1 - \alpha) S_{t-1}$$

(2)

其中, S_t 为 t 时刻的平滑值, y_t 为 t 时刻的实际值(此处为对数变换后的降水量值), S_{t-1} 为 $t-1$ 时刻的平滑值, α 为平滑系数(取值范围 $0 < \alpha < 1$)。 α 设定为 0.2, 该取值在减少短期随机噪声与保留长期趋势之间取得平衡。以便突出数据的长期变化趋势, 弱化短期波动的干扰, 使降水量数据的潜在规律更清晰地呈现。本文选取月作为时间尺度。[图 1](#) 展示了本文所选取八个站点的原始、对数变换和平滑降水的比较。

根据[图 1](#), 原始降水数据表现出强烈的季节性与随机性的叠加特征, 其中年内降水呈现规律性波动, 表明夏季月份存在集中性强降水事件, 青龙、保定等站点能观测到显著的峰值。通过对数变换后, 数据尺度被有效压缩, 极端降水值对整体分布的偏态影响显著降低, 方差趋于稳定。在采用指数平滑法对对数空间数据进行处理后, 成功分离出降水的长期趋势成分, 高频随机波动。年内季节性周期与年际缓变特征比较明显, 石家庄、邢台等站点经平滑后均显示出稳定的雨季 - 旱季交替模式。

2.4.2. 其它气象要素影响

降水量受到自身周期性和季节性的影响外, 还可能受到其它气象要素如温度、露点的影响, 以石家庄站点为例进行研究。

1) 相关性分析

首先进行皮尔逊(Pearson)相关性[\[9\]](#)分析初步筛选特征指标, 计算结果见[表 2](#)。

Table 2. The correlation between meteorological indicators and precipitation
表 2. 气象指标与降水量的相关性

特征	Pearson 相关性	P 值(双尾)
DEWP	0.716	<0.001
TEMP	0.628	<0.001
SLP	-0.628	<0.001
TEMPDIF	-0.341	<0.001
MXSPD	-0.175	<0.001
WDSP	-0.176	<0.001
VISIB	-0.044	0.279

根据[表 2](#), 平均能见度(VISIB)与降水量(PRCP)无统计学上的显著相关性, 可以剔除, 判断依据为当 Pearson 相关系数的绝对值越接近 1 时, 变量之间的相关性强度越强, 并且当绝对值超过 0.5 时, 变量之间存在显著的相关性。因此, 平均露点(DEWP)、平均温度(TEMP)和平均海平面压力(SLP)与降水量具有显著的相关性。

2) 多重共线性分析

为进一步验证特征指标间潜在的多重共线性并增强分析的可视化解释, 绘制了气象指标间的相关热力图, 如[图 2](#) 所示。

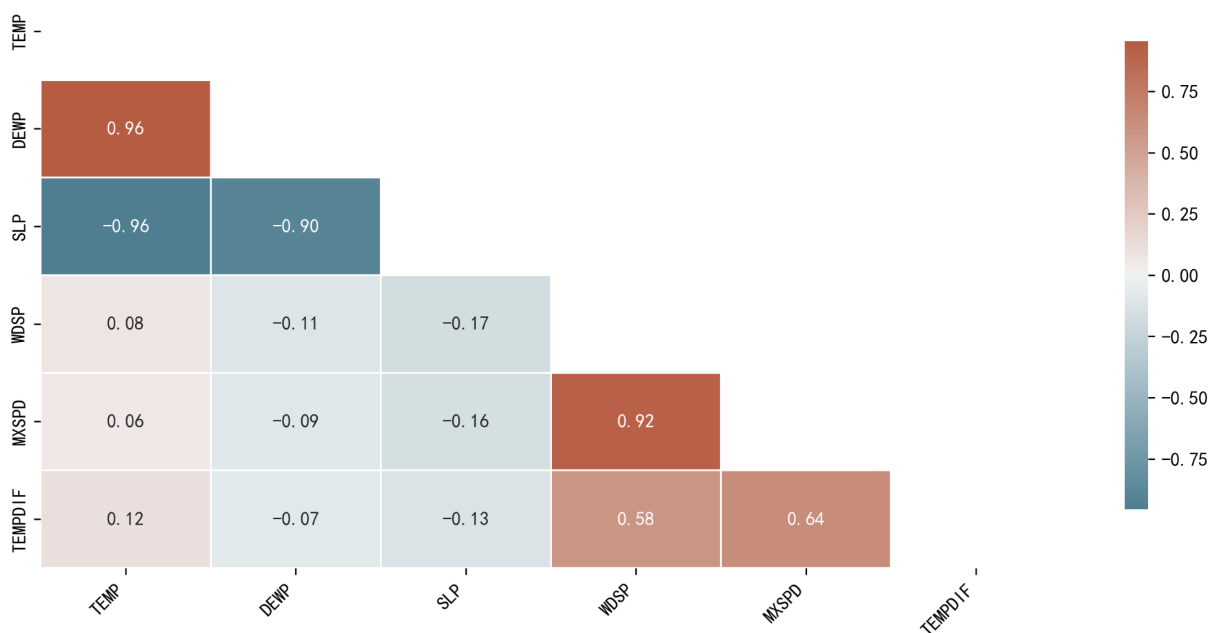


Figure 2. Model of influencing factors of purchase intention

图 2. 显著性气象指标相关性热力图

根据图 2, TEMP (平均温度)与 DEWP (平均露点)、SLP (平均海平面压力)呈现显著的相关性, 故剔除 TEMP、SLP, 保留 TEMPDIF (温差), WDSP (平均风速)和 MXSPD (最大持续风速)具有较强的正相关性, 两者保留一个即可, 故剔除 MXSPD, 保留 WDSP。

3) 气象因素影响程度分析

为进一步分析筛选后的特征指标(DEWP, TEMPDIF, WDSP)对降水量预测模型的具体影响机制, 需要借助 LightGBM 模型进行下一步的特征筛选。LightGBM 模型[10]是一个梯度提升决策树(GBDT)的实现, 该模型的本质原理是利用基分类器决策树训练集成, 得到最优的模型。基于 LightGBM 梯度提升框架, 参数配置为控制模型复杂度为 31, 学习率为 0.05, 树的数量为 100, 数据集划分按照 8:2 比例随机分割训练集/测试集, 经过性能评估, 测试集 $R^2 = 0.6306$, 表明模型可解释 63.06%的降水量方差。

基于此, 利用 SHAP 图解释气象因素对降水量的特征重要性排序, 如图 3 所示。

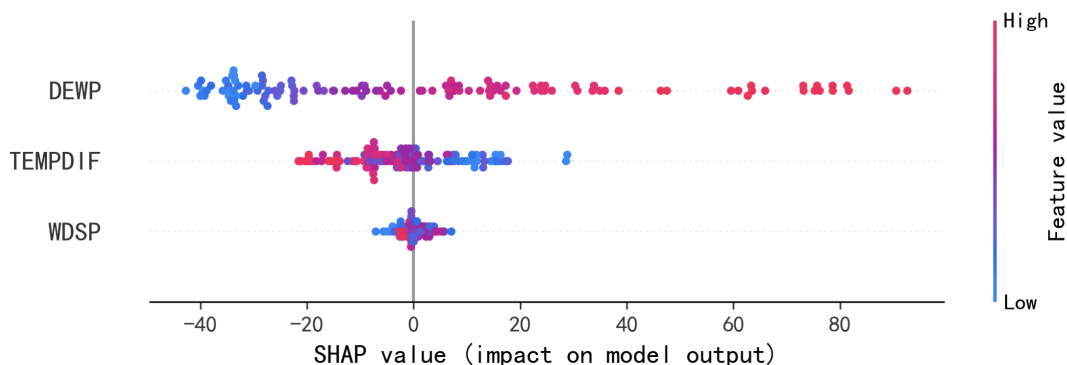


Figure 3. Model of influencing factors of purchase intention

图 3. SHAP 值解释图

根据图 3, DEWP 具有显著的趋势依赖性影响, TEMPDIF 的影响程度较为混杂, WDSP 基本聚集在

零值附近, 表明对模型输出的影响极小, 无论其值是高还是低。特征重要性排序: $DEWP > TEMPDIF > WDSP$ 。

故最终选择 $DEWP$ (平均露点)、 $TEMPDIF$ (温差)、 $WDSP$ (平均风速)作为后续构建模型的潜变量特征来提升模型预测精度。

3. 研究方法

3.1. 变分模态分解

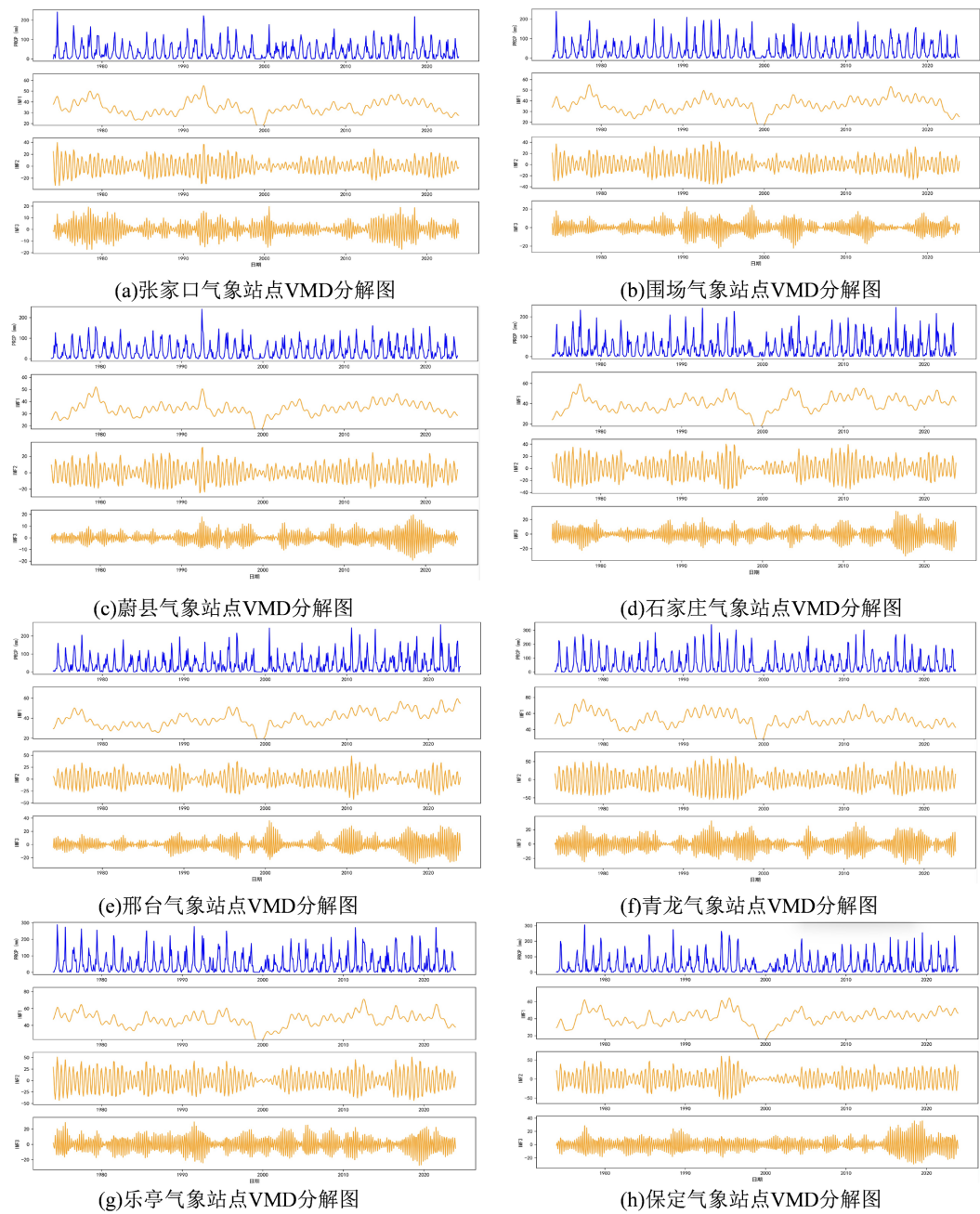


Figure 4. VMD decomposition diagrams of 8 meteorological stations
图 4. 8 个气象站点 VMD 分解图

变分模态分解(Variational Mode Decomposition, VMD) [11]是一种基于变分模型的自适应信号分解方法, 通过非递归的方式将复杂的多分量信号分解为若干个不同频率和带宽的本征模态函数(Intrinsic Mode Function, IMF)。与传统的经验模态分解(Empirical Mode De-composition, EMD)相比, VMD 的核心思想是通过构建和求解变分优化问题, 从而实现信号的自适应分解, 尤其在处理非平稳、非线性信号时能够显著提高分解的稳定性。

VMD 主要包括以下步骤:

1) 模态更新。对每个模态 $u_k(t)$, 其频域形式 $\hat{u}_k(\omega)$ 的更新公式为:

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i^n(\omega) + \left(\frac{\hat{\lambda}^n(\omega)}{2} \right)}{1 + 2\alpha(\omega - \omega_k^n)^2} \quad (3)$$

2) 中心频率更新。中心频率 ω_k^{n+1} 被更新为模态频谱的中心:

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}_k^{n+1}(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k^{n+1}(\omega)|^2 d\omega} \quad (4)$$

3) 拉格朗日乘子更新。对于所有的 $\omega \geq 0$, 拉格朗日乘子 $\hat{\lambda}^{n+1}$ 通过下式更新, 以确保满足约束。

$$\hat{\lambda}^{n+1} = \hat{\lambda}^n + \tau \left(\hat{f} - \sum_k \hat{u}_k^{n+1} \right) \quad (5)$$

直到满足 $\sum_k \hat{u}_k^{n+1} - \hat{u}_k^{n2} < \varepsilon$ 的收敛条件。

八个气象站点的 VMD 分解图如图 4 所示。

3.2. 隐马尔可夫模型

隐马尔可夫模型(Hidden Markov Model, HMM)是一种生成式概率统计模型[12], 它基于马尔可夫链, 通过引入隐藏状态来描述具有不确定性和动态变化的过程。其核心思想是假设存在不可直接观测的、隐含状态的马尔可夫链, 但可以通过模型计算这些隐含状态的概率, 从而生成可观测的输出值。HMM 是一种生成式的概率模型。由初始状态概率向量 π 、状态转移概率矩阵 A 和观测概率矩阵 B 。

1) HMM 模型的定义: $\lambda = (A, B, \pi)$ 。

2) HMM 模型的三个基本问题分别为评估问题、解码问题和学习问题。本研究主要利用 HMM 的解码问题进行模型组合。

3) 解码问题(Decoding)

给定模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = \{o_1, o_2, \dots, o_T\}$, 找到最有可能产生该观测序列的隐藏状态序列 $S^* = \{s_1^*, s_2^*, \dots, s_T^*\}$ 。

维特比算法定义维特比变量 $\delta_t(i)$ 为在时刻 t 到达状态 q_i 的所有路径中, 概率最大的路径的概率值, 即

$$\delta_t(i) = \max_{s_1, s_2, \dots, s_{t-1}} P(s_1, s_2, \dots, s_{t-1}, s_t = q_i, o_1, o_2, \dots, o_t | \lambda) \quad (6)$$

通过递归公式 $\delta_{t+1}(j) = \max_{1 \leq i \leq N} [\delta_t(i) a_{ij}] b_j(o_{t+1})$ 找到概率最大的路径, 计算复杂度为 $O(N^2T)$ 。

3.3. 双向长短时记忆神经网络

双向长短期记忆网络(Bidirectional Long Short-Term Memory Network, BiLSTM)是长短期记忆网络

(LSTM)的改进版本[13], 它的核心思想是通过双向循环结构突破单向循环神经网络, 由于 LSTM、GRU 仅能捕捉单向时间依赖的局限, 实现对序列数据过去和未来上下文信息的同时建模, 从而更全面地捕捉时间序列中的复杂依赖关系, 如图 5 所示。

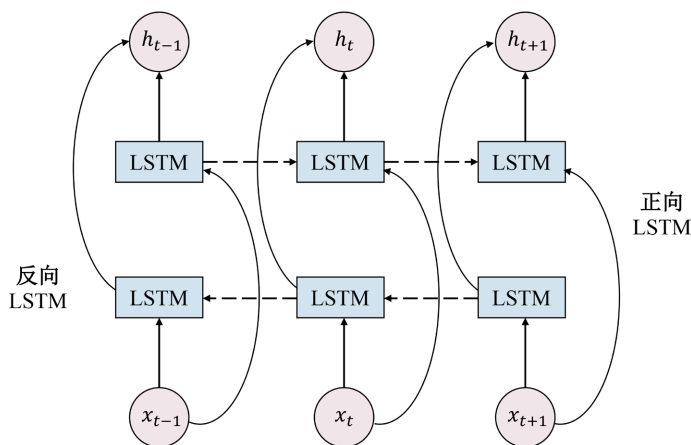


Figure 5. BiLSTM network structure diagram
图 5. BiLSTM 网络结构图

BiLSTM 网络中正向 LSTM 结构计算过程与单个 LSTM 计算过程相似, 将正向隐藏状态和反向隐藏状态组合得到 BiLSTM 网络的隐含层状态, 其计算公式为:

$$\bar{h}_t = LSTM(h_{t-1}, x_t) \quad (7)$$

$$\bar{h}_t = LSTM(h_{t+1}, x_t) \quad (8)$$

$$h_t = \alpha \bar{h}_t + \beta \bar{h}_t \quad (9)$$

其中, x_t , $\bar{h}_t$, $\bar{h}_t$ 分别为 t 时刻的输入数据、正向 LSTM 隐含层的输出和反向 LSTM 隐含层的输出; α , β 均为常数, 分别表示 $\bar{h}_t$, $\bar{h}_t$ 的权重。

4. 基于 VMD-HMM-BiLSTM 的降水量预测模型

4.1. 模型的建立

为进一步挖掘月降水量数据的时序特征, 建立了基于 VMD-HMM-BiLSTM 的降水预测模型, 预测流程步骤如下:

1) 使用 VMD 方法对原始降水序列进行变分模态分解, 通过时间序列交叉验证自动优化分解模态数 $K \in [2, 8]$, 带宽参数 $\alpha = 2000$ 确保频率分离精度, 通过 VMD 将非平稳的降水序列分解为多个本征模态函数 IMF 分量, 每个 IMF 代表不同时间尺度的降水波动特征;

2) HMM 模型将原始降水序列划分为训练集与测试集, 划分比例为 7:3, 确保训练充分性和评估客观性。通过网络搜索在 $n \in [3, 8]$ 范围内优化。采用高斯型 HMM 基于模型对数似然评分选择最优状态数。之后采用 Viterbi 算法求解概率最大的隐藏状态序列。为将离散状态信息融入预测模型, 对状态序列进行独热编码(One-Hot-Encoding), 生成二进制特征矩阵 $H \in \mathbb{R}^{T \times n_{components}}$ 。该矩阵与原始气象特征、VMD 分解分量共同构成多维输入特征集;

3) 将 HMM 状态特征与 VMD 分解后的 IMF 分量、气象变量(平均露点、风速、温差)进行归一化处理, 构建时空特征张量 $X \in \mathbb{R}^{T \times (d_{base} + K + n_{components})}$, 作为后续 BiLSTM 网络的输入, 其中 d_{base} 为基础特征维

度, K 为 VMD 分解模态数。

VMD-HMM-BiLSTM 模型的建模流程如图 6 所示。

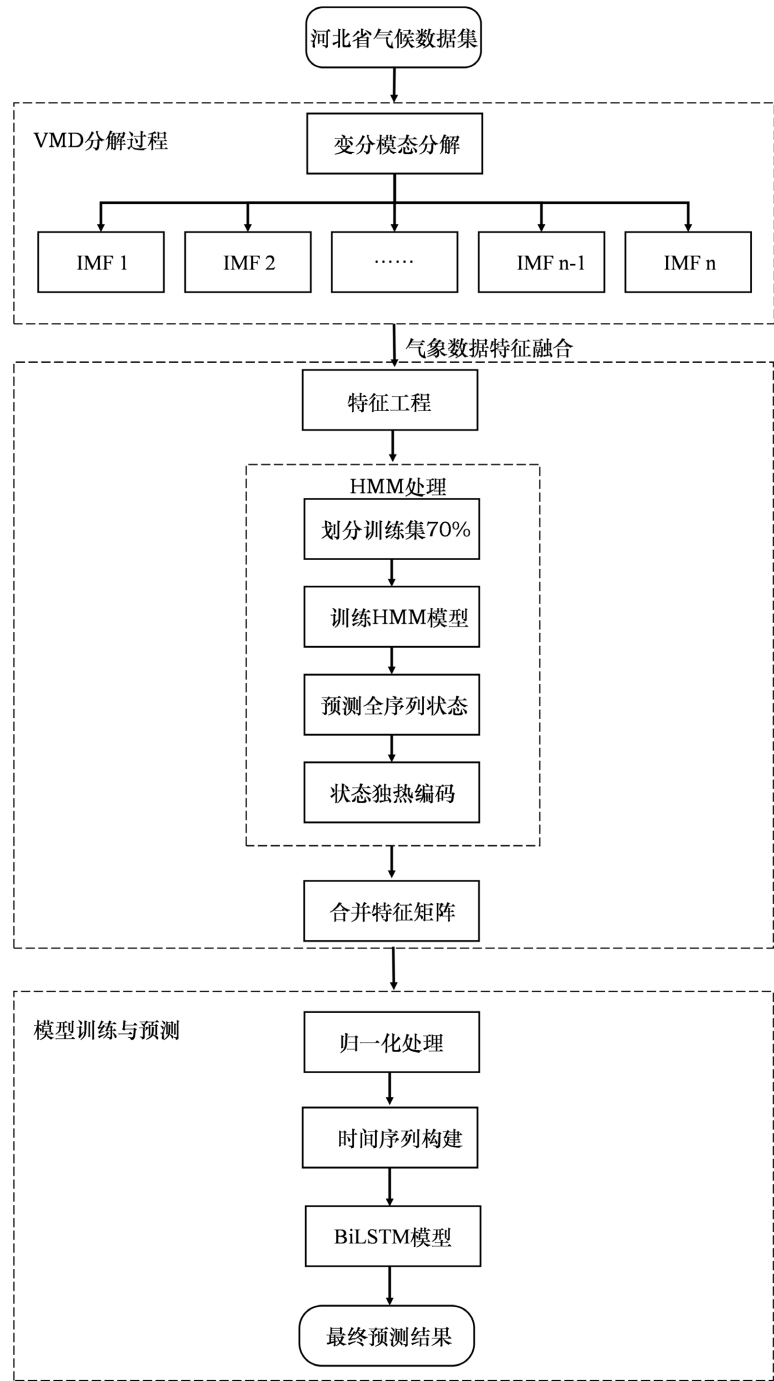


Figure 6. Model framework for VMD-HMM-BiLSTM
图 6. VMD-HMM-BiLSTM 的模型框架

4.2. 评价指标

在进行预测工作时, 为了更加直观清晰地评判每种模型的预测效果, 使用均方根误差 RMSE (Root

Mean Square Error)、平均绝对误差 MAE (Mean Absolute Deviation)和决定系数 R^2 (Coefficient of Determination)作为评价降水预测模型精度的指标。

其中, RMSE 反映了模型预测值与真实值之间的偏离程度, 其单位与原始数据的单位相同, 因此它在解释模型预测精度时具有直观性。当 RMSE 值较小时, 表明模型的预测更为精准。MAE 是所有单个观测值与真实值之间绝对差值的平均值。与 RMSE 相比, MAE 更侧重于描述误差的实际大小。决定系数 R^2 是衡量回归模型拟合优度的一个常用指标, 它反映了自变量对因变量的解释程度, 取值范围在 0 到 1 之间, R^2 越接近 1, 模型就拟合得越好。

4.3. 实验结果分析

4.3.1. 实验结果对比分析

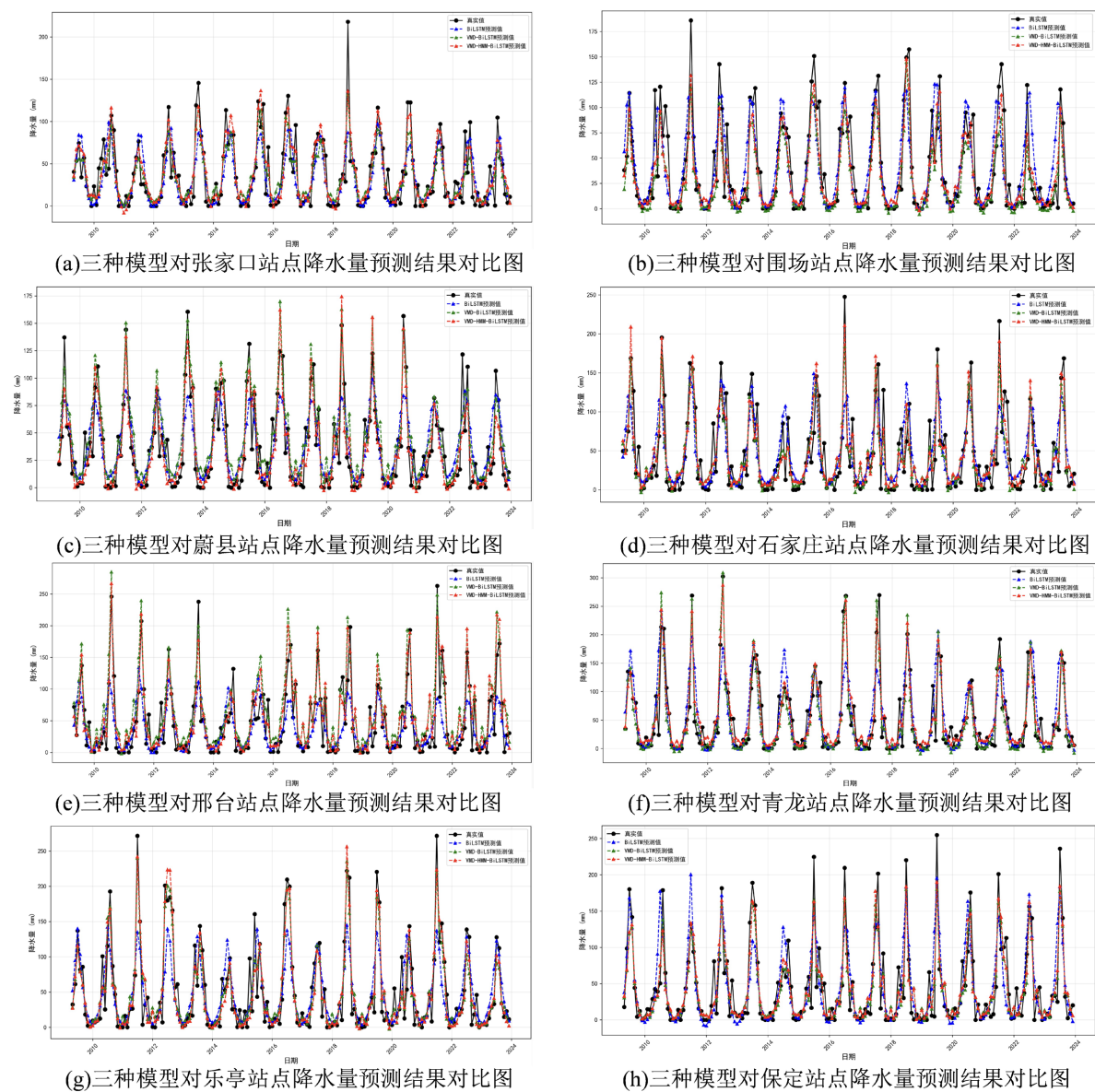


Figure 7. Comparison of precipitation prediction results of three models for 8 stations
图 7. 三种模型对 8 个站点降水量预测结果对比

为全面评估不同模型在河北省降水量预测中的性能, 本研究对 BiLSTM、VMD-BiLSTM 和 VMD-HMM-BiLSTM 三种模型进行多维度对比分析。其中, BiLSTM 模型作为基础时序模型, VMD-BiLSTM 模型通过变分模态分解(VMD)增强数据特征提取能力, VMD-HMM-BiLSTM 模型进一步引入隐马尔可夫模型(HMM)刻画降水状态的时序依赖性与随机性。

8 个站点 3 种模型的预测结果对比图如图 7 所示。

进一步从各站点误差指标和整体性能展开分析, 如表 3 所示。

Table 3. The performance evaluation index values of the three models in 8 sites
表 3. 3 种模型在 8 个站点中的性能评价指标值

站点	模型	均方根误差	平均绝对误差	绝对系数
张家口	BiLSTM	24.58	17.33	0.60
	VMD-BiLSTM	18.37	13.78	0.78
	VMD-HMM-BiLSTM	17.68	13.16	0.79
围场	BiLSTM	28.36	18.75	0.59
	VMD-BiLSTM	22.14	14.71	0.75
	VMD-HMM-BiLSTM	19.43	13.33	0.81
蔚县	BiLSTM	23.90	18.55	0.63
	VMD-BiLSTM	18.87	15.17	0.77
	VMD-HMM-BiLSTM	17.18	12.58	0.81
石家庄	BiLSTM	34.31	23.90	0.57
	VMD-BiLSTM	24.76	17.69	0.78
	VMD-HMM-BiLSTM	22.26	16.98	0.82
邢台	BiLSTM	38.88	25.89	0.51
	VMD-BiLSTM	30.04	22.75	0.71
	VMD-HMM-BiLSTM	25.55	18.50	0.79
青龙	BiLSTM	38.48	25.18	0.66
	VMD-BiLSTM	24.61	16.83	0.86
	VMD-HMM-BiLSTM	21.41	16.26	0.89
乐亭	BiLSTM	37.60	25.56	0.63
	VMD-BiLSTM	24.21	17.45	0.84
	VMD-HMM-BiLSTM	22.38	15.95	0.87
保定	BiLSTM	35.38	24.29	0.62
	VMD-BiLSTM	28.25	19.25	0.76
	VMD-HMM-BiLSTM	24.58	17.99	0.82

根据表 3, 综合比较八个站点各模型的预测结果, VMD-HMM-BiLSTM 模型在均方根误差(RMSE)、平均绝对误差(MAE)和绝对系数(R^2)三项关键指标上均一致优于基础的 BiLSTM 模型及其改进型 VMD-BiLSTM 模型。具体表现为: 在每个站点, 该模型的 RMSE 和 MAE 均为最低值, 而 R^2 均为最高值。例如, 在邢台站, 它的 RMSE 为 25.55, 较 BiLSTM 模型的 38.88 显著降低, 且它的 R^2 为 0.59, 较基础模

型的 0.51 有大幅度提升; 在青龙站点, 该站点的 R^2 达到最优, 为 0.89, 充分证明了 VMD-HMM-BiLSTM 模型在预测精度何拟合优度上的良好性能。

VMD-HMM-BiLSTM 模型展现出的性能优势具有显著的普遍性和稳健性。基础模型 BiLSTM 的 RMSE 范围从 24.58 到 38.88, 该模型均能稳定地实现预测误差达到最大程度地降低, 其中邢台站的 RMSE 最大降幅达到 34.3%, 并且石家庄地拟合度显著提升, R^2 从 0.57 提升至 0.82。更重要的是, 在所有站点上, VMD-HMM-BiLSTM 模型优于 VMD-BiLSTM 模型, 例如围场站的 RMSE 从 22.14 降低到 19.53, 蔚县站的 MAE 从 15.17 降低到 12.58, 这强有力地验证了变分模态分解 VMD、隐马尔可夫模型 HMM 和双向长短期记忆网络 BiLSTM 策略的有效性和广泛的普适性, 表明该模型架构能有效适应不同环境下的时序数据特征。

4.3.2. 模型泛化能力测试

首先引入对称平均绝对百分比误差(SMAPE, Symmetric Mean Absolute Percentage Error)来增加模型的评估效果, 对称平均绝对百分比误差(SMAPE)是一种用于衡量预测值与真实值之间百分比误差平均值的指标。该指标考虑了预测值和真实值的规模, 通过对差值进行归一化来减少规模差异引起的偏差。增加新的气象站点(郑州站点), 经过多次实验验证模型的泛化能力, 模型评估效果如表 4 所示。

Table 4. The performance evaluation index values of the three models in Zhengzhou site
表 4. 3 种模型在郑州站点中的性能评价指标值

次数	模型	均方根误差	平均绝对误差	对称平均绝对百分比误差	绝对系数
1	BiLSTM	47.83	33.76	60.23%	0.39
	VMD-BiLSTM	17.91	14.20	42.71%	0.91
	VMD-HMM-BiLSTM	16.15	12.58	34.82%	0.93
2	BiLSTM	47.36	33.38	62.97%	0.4
	VMD-BiLSTM	18.89	14.77	40.93%	0.91
	VMD-HMM-BiLSTM	17.49	12.73	31.61%	0.92
3	BiLSTM	47.54	33.49	59.74%	0.40
	VMD-BiLSTM	19.01	14.85	44.02%	0.90
	VMD-HMM-BiLSTM	15.74	12.09	34.01%	0.93
4	BiLSTM	48.06	34.17	60.95%	0.39
	VMD-BiLSTM	18.44	14.53	47.79%	0.91
	VMD-HMM-BiLSTM	17.70	13.39	35.74%	0.92
5	BiLSTM	48.54	34.45	62.32%	0.37
	VMD-BiLSTM	20.39	15.87	46.31%	0.89
	VMD-HMM-BiLSTM	16.83	12.73	36.57%	0.92
5 次均值	BiLSTM	47.87	33.85	61.24%	0.39
	VMD-BiLSTM	18.93	14.84	44.35%	0.90
	VMD-HMM-BiLSTM	16.78	12.70	34.55%	0.92

根据表 4, 为了验证该模型的泛化能力, 在郑州站点进行了 5 次预测分析, 并对 5 次结果求取平均值。结果显示, 整体模型效果的有效性: VMD-HMM-BiLSTM > VMD-BiLSTM > BiLSTM。

5. 结论

本文对河北省八个气象站点的降雨量数据进行分析, 考虑到 PRCP (降水量) 自身因素和 DEWP (平均露点)、TEMPDIF (温差)、WDSP (平均风速) 其它气象因素的影响, 将其作为输入变量进行后续降水量的预测。本文首先采用了 VMD 方法将原始降雨量数据分解为多个子序列, 有效降低了数据的非平稳性, 进一步通过 HMM 模型确定最优参数, 将 IMF 子序列和其它气象因素进行归一化作为 BiLSTM 模型的输入, 建立了 VMD-HMM-BiLSTM 月降水量预测模型, 再与 BiLSTM、VMD-BiLSTM 模型进行对比, 实验结果表明:

1) VMD 方法的引入有效降低了降雨量数据的非平稳性, 减轻了模型在提取时序特征的压力, 与单一模型 BiLSTM 模型相比, R^2 最大有效提升率达 21%;

2) VMD-HMM-BiLSTM 模型与 VMD-BiLSTM 模型、BiLSTM 模型相比, 具有更低的误差, R^2 整体平均值达到 0.825, 说明 HMM 与 BiLSTM 模型的融合也能有效提升降雨量预测的精度, 体现了组合模型一定的优势, 有一定的参考价值, 且模型有一定的泛化能力, 可应用于其他地区。

参考文献

- [1] 郭廓. ARIMA 模型在阜新地区降水量预测中的适用性分析[J]. 黑龙江水利科技, 2024, 52(4): 37-39+79.
- [2] 陈聪. 基于机器学习的灾害性气象预测[D]: [硕士学位论文]. 天津: 天津理工大学, 2023.
- [3] Khan, S.A., Zhang, C. and Zhang, H. (2022) Review of Machine Learning Techniques for Hydrological Time Series Forecasting. *Water Resources Management*, **36**, 3801-3823.
- [4] Kou, M., Zhang, K., Zhang, W., Ma, J., Ren, J. and Wang, G. (2023) Application Research of Combined Model Based on VMD and MOHMO in Precipitable Water Vapor Prediction. *Atmospheric Research*, **292**, Article ID: 106841. <https://doi.org/10.1016/j.atmosres.2023.106841>
- [5] 陈子浩, 闫磊, 董旺, 等. 基于贝叶斯非齐次隐马尔科夫模型的京津冀日降雨模拟研究[J]. 水文, 2025, 45(3): 48-53+100.
- [6] 李方圆, 张涛. 基于 HMM-XGBoost 的股价预测[J]. 桂林航天工业学院学报, 2021, 26(4): 484-488.
- [7] 赵子鹏, 魏新奇, 唐龙, 等. 基于 ARIMA 模型对定西天气数据的分析与预测[J]. 现代信息科技, 2024, 8(9): 140-143.
- [8] 杨波, 靳壮壮, 刘洋. 基于有限元线性插值法的泡沫混凝土配合比优化[J]. 建筑结构, 2023, 53(S2): 1376-1380.
- [9] 战帅. 基于皮尔逊相关分析的公益筹款平台质量影响因素研究[J]. 国际公关, 2022(19): 71-73.
- [10] 王伟, 党海龙, 康胜松, 等. 基于 LightGBM 和 SHAP 算法的致密油储层孔隙度预测[J/OL]. 油气地质与采收率, 1-10. <https://doi.org/10.13673/j.pgre.202403007>, 2025-06-07.
- [11] Dragomiretskiy, K. and Zosso, D. (2014) Variational Mode Decomposition. *IEEE Transactions on Signal Processing*, **62**, 531-544. <https://doi.org/10.1109/tsp.2013.2288675>
- [12] Eddy, S.R. (1996) Hidden Markov Models. *Current Opinion in Structural Biology*, **6**, 361-365. [https://doi.org/10.1016/s0959-440x\(96\)80056-x](https://doi.org/10.1016/s0959-440x(96)80056-x)
- [13] Schuster, M. and Paliwal, K.K. (1997) Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, **45**, 2673-2681. <https://doi.org/10.1109/78.650093>