

王庆荣, 刘心康, 朱昌锋, 等. 基于 LASSO - MEMD - SSA - LSTM 的中国碳排放预测研究[J]. 安全与环境学报, 2026, 26(6): 2475 - 2484.

WANG Q R, LIU X K, ZHU C F, et al. Research on China's carbon emission prediction based on LASSO - MEMD - SSA - LSTM[J]. Journal of Safety and Environment, 2026, 26(6): 2475 - 2484.

基于 LASSO - MEMD - SSA - LSTM 的中国碳排放预测研究*

王庆荣¹, 刘心康¹, 朱昌锋², 王俊杰¹

(1 兰州交通大学电子与信息工程学院, 兰州 730070; 2 兰州交通大学交通运输学院, 兰州 730070)

摘要:针对传统碳排放预测中数据质量差、模型训练效率低的问题,建立套索回归(Least Absolute Shrinkage and Selection Operator, LASSO)、多元经验模态分解(Multivariate Empirical Mode Decomposition, MEMD)、麻雀搜索算法(Sparrow Search Algorithm, SSA)和长短时记忆网络(Long Short-Term Memory network, LSTM)相结合的模型,以提高碳排放量的预测精度。首先,使用可拓展的随机性环境影响回归模型(Stochastic T Impacts by Regression on PAT, STIRPAT)选取年末总人口数量、人均GDP等影响因素;其次,通过皮尔逊相关系数和 LASSO 回归模型进行初步和二次筛选;再次,对数据进行预处理并进行 MEMD 分解,进而对分解出的不同频率的 IMF 分量以及残差项 RES 分别建立预测模型;最后,使用 SSA 对 LSTM 模型的神经元个数、迭代次数、学习率等参数进行寻优,并将数据汇总整合,对比分析其他预测模型的预测结果。结果显示, LASSO - MEMD - SSA - LSTM 模型的预测值与实际值最为接近,其平均绝对误差、均方根误差、平均相对百分比误差分别为 8.45%、8.32%、0.85%,均优于其他模型。

关键词:环境工程学;碳排放预测;套索回归;多元经验模态分解;麻雀搜索算法;长短时记忆网络

中图分类号:X22 **文献标志码:**A **文章编号:**1009-6094(2026)06-2475-10

DOI: 10.13637/j.issn.1009-6094.2025.1346

0 引言

中国由于人口与经济发展而高度重视气候变化问题,其中碳排放预测作为关键研究内容,对实现“双碳”目标具有重要意义。然而,碳排放数据具有强波动性与非线性,会严重制约预测精度^[1-2],且特征选取亦对预测效果影响显著^[3-4]。当前研究多基于可拓展的随机性环境影响回归模型(Stochastic T Impacts by Regression on PAT, STIRPAT)等识别碳排放影响因素,如刘淳森等^[5]借助该模型筛选变量并预测达峰时间。杜涵蓓等^[6]结合 Kaya 恒等式与长期能源替代规划系统(Long-range Energy Alternatives Planning System, LEAP)进行多情景模拟。

在预测方法方面,既有研究广泛采用自回归积分滑动平均模型(AutoRegressive Integrated Moving Average, ARIMA)^[7]、灰色理论^[8]、机器学习融合模型^[9-10]及分解集成框架^[11-12]等。尽管这些模型在一定程度上提升了预测能力,但仍存在局限性:多数研究未能充分处理数据波动性与非线性;部分模型忽视环境因素的影响;特征选取与多特征序列的协同

处理亦有不足;现有方法多侧重于短期预测,难以满足长期战略规划需求。为此,本文基于 STIRPAT 模型与两步特征筛选法选取影响因素,引入多元经验模态分解(Multivariate Empirical Mode Decomposition, MEMD)处理原始数据,并将其分解为多个本征模态分量(Intrinsic Mode Functions, IMFs)和残差(Residuals, RES),进而构建麻雀搜索算法(Sparrow Search Algorithm, SSA)优化长短时记忆网络(Long Short-Term Memory network, LSTM)预测模型,对各分量分别进行预测,以提升碳排放中长期预测的精度与稳健性。

1 碳排放影响因素选取

1.1 影响因素的选取与筛选

STIRPAT 模型作为一种经典的环境影响评价模型,具有较强的应用潜力和可塑性,在碳排放领域的研究中具有重要意义。因此,本文采用 STIRPAT 模型选取碳排放因素。根据王珂珂等^[13]对中国碳排放的预测研究,发现中国碳排放受人口、经济、技术水平影响较大。

* 收稿日期: 2025 - 08 - 01

作者简介: 王庆荣,教授,硕士,从事交通大数据与碳排放、机器学习的应用研究,329046272@qq.com。

基金项目:国家自然科学基金项目(72161024);甘肃省教育厅“双一流”重大研究项目(GSSYLXM-04)

用于预测碳排放量所使用的影响因素十分重要,会直接影响最后的预测结果,研究运用皮尔逊相关系数(Pearson Correlation Coefficient, PCC)分析各影响因素与碳排放量之间的相关性,并依据相关系数初步筛选影响因素。皮尔逊相关系数是衡量两个变量相关程度的一种统计学式,通常用协方差与方差的商来表示^[14],见式(1)。皮尔逊相关强度见表1。

$$r(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{D(X)} \sqrt{D(Y)}} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} \quad (1)$$

为了进一步研究影响因素与碳排放量之间的相关性,采用 LASSO 回归模型筛选第二轮变量。LASSO 作为一种含 L1 正则化的压缩估计方法,其特性在于通过压缩系数(使部分系数趋于 0)自动筛选变量,以有效简化模型结构并防止过拟合,尤其适合处理影响因素间可能存在的多重共线性问题。

1.2 碳排放预测模型构建

1.2.1 基于 MEMD 的碳排放数据分解

碳排放量受多种因素影响,如科技水平、经济的发达程度和人口规模等,导致碳排放量表现出非线性和波动性特征。碳排放数据的非线性和波动性特点会降低预测准确性,因此需要进行数据处理。为此,本文采用黄愕在 1998 年首次提出的经验模态分解(Empirical Mode Decomposition, EMD)方法^[15]进行研究,以处理非线性和非平稳过程。该方法的有效性在多个领域得到了验证,张茵栋等^[16]利用 EMD 算法消除了变量端点效应和时序非平稳性的影响。秦进等^[17]通过经验模态分解(EMD)分解高速铁路客流量序列,得到了反映周期性及线路客流内在特性的模态分量。

鉴于标准的经验模态分解方法通常难以适用于多元化数据,2010 年,Rehman 等^[18]提出了多元经验模态分解方法,即将传统经验模态分解方法扩展至

多维信号处理。该方法有效地解决了传统方法在处理多维信号时 IMF 分量数量和频率尺度不对齐的问题,实现了多元信号的同步联合处理。2024 年, Huang 等^[19]最先使用一种基于多变量经验模态分解的动态图深度学习模型,以用于网络范围内的地铁客流预测。贺毅岳等^[20]通过 LASSO 方法从 30 个指标中筛选变量,构建沪金价格影响因素体系,并对变量与价格序列进行 MEMD 同步分解。获得了各频率尺度下若干组 IMF 分量,并分别建立了每一频率尺度下 IMF 分量的深度融合多频率尺度同步分解功能与多元时序复杂特征高效提取的集成预测模型(Wasserstein Generative Adversarial Network with Gradient Penalty, WGAN-GP)。这表明 MEMD 可以有效提升时序预测的精度。

MEMD 分解过程如下。

(1) 在 $n-1$ 维球面上,本文使用 Hammersley 序列采样法进行均匀采样,并获得 K 个方向向量。

(2) 给定输入信号 $v(t) = \{v_1(t), v_2(t), \dots, v_n(t)\}$, 每个方向向量 $\mathbf{x}^k = (x_1^k, x_2^k, \dots, x_n^k)$, 对应球面中方向角 $\theta_k = (\theta_{k1}, \theta_{k2}, \dots, \theta_{k(n-1)})$ 。其中, $t = 1, 2, \dots, T$, T 为信号长度, n 为通道数。

(3) 计算信号 $v(t)$ 在 K 个方向向量 $\mathbf{x}^k$ 上的投影,以得到投影集 $|p^k(t)|_1^k$ 。其中, p 表示投影值, θ 表示方向角, t 表示时间点。

(4) 找出每个投影集 $|p^k(t)|_1^k$ 的极值点对应的时刻 $\{t_i^k\}$ (i 为极值点序号, θ 为方向角)。

(5) 对数据点 $\{t_i^k, v(t_i^k)\}$ 进行多元样条插值,得到对应的包络曲线集 $\{e^k(t)\}_1^k$ 。

(6) 方向向量包络曲线的均值 $m(t)$ 由式(2)定义。其中, K 是方向向量组数, $e^k(t)$ 为第 k 个包络值。

$$m(t) = \frac{1}{K} \sum_{k=1}^K e^k(t) \quad (2)$$

(7) 定义剩余信号 $d(t) = v(t) - m(t)$ 。IMF 的判定依据是上下包络均值的零近似性或特定准则满足性,若未达到要求则循环计算。使用 MEMD 分解输入信号 $v(t)$, 分解生成 M 个多频 IMF, 见式(3)。

表1 皮尔逊相关强度表

Table 1 Pearson correlation intensity table

相关系数	[0.8, 1.0]	[0.6, 0.8)	[0.4, 0.6)	[0.2, 0.4)	[0.0, 0.2)
相关强度	极强相关	强相关	中等程度相关	弱相关	极弱或无相关

$$v(t) = \sum_{m=1}^M \bar{C}_m(t) + \bar{r}(t) \quad (3)$$

式中 $\bar{C}_m(t)$ 为第 m 个 IMF 值; $\bar{r}(t)$ 为分解残差。

1.2.2 麻雀搜索算法

麻雀搜索算法 (SSA) 是 Xue 等^[21] 在 2020 年提出的一种新兴群体智能算法,具有收敛速度快、寻优能力强等优点。麻雀搜索算法是一种基于麻雀觅食行为而提出的启发式优化算法。SSA 将种群分为 3 类:探索者负责全局寻优并向最优位置靠近;跟随者利用探索者信息进行局部开发或随机移动;警戒者随机选取,当感知危险时向安全中心移动或随机扰动以逃离局部最优。通过三者的协同作用,SSA 高效平衡了全局探索与局部开发能力。

1.2.3 LSTM

LSTM 作为特殊的循环神经网络 (Recurrent Neural Network, RNN),通过输入门、遗忘门、输出门和细胞态结构解决了传统 RNN 的梯度消失/爆炸难题。通过这些门控单元,LSTM 网络通过其门控机制,能够自主决定在每个时间步中保留或筛选哪些信息,从而更好地处理长期依赖关系。LSTM 单元结构见图 1。

若 LSTM 参数设置不当,易引发过拟合或欠拟合,最终导致模型在新数据中表现不佳。因此,本文使用 SSA 算法来确定 LSTM 的参数以确保模型性能达到最佳水平。

1.2.4 SSA 优化 LSTM

LSTM 模型在多维碳排放预测中的性能主要受神经元数、迭代次数、学习率等超参数影响。传统人工调参效率低、精度差,尤其面对复杂多维输入时更显不足。为此,引入 SSA 实现超参数自动寻优,以有效解决传统方法调参效率低、精度不足的问题,进而显著提升多维碳排放预测模型的性能。

2 基于 MEMD 分解的 SSA - LSTM 模型的构建

为了提高碳排放预测的精度,本文提出一种基

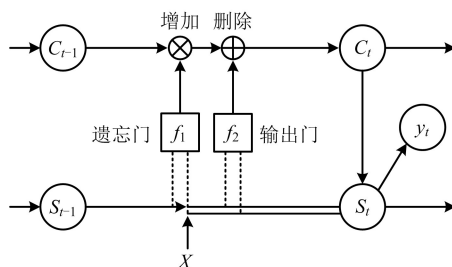


图 1 LSTM 内部结构

Fig.1 Internal structure of LSTM

于 MEMD 分解的 SSA 优化 LSTM 模型的预测方法,主要结构见图 2。

本文构建 MEMD - SSA - LSTM 集成预测模型以对中国 2000—2021 年碳排放量进行实证分析。本文基于 STIRPAT 模型初步筛选碳排放影响因素,综合运用皮尔逊相关分析法和 LASSO 回归模型进行二次筛选,并使用 SSA 算法对 LSTM 模型的超参数进行寻优。为了降低碳排放数据序列的波动性和非线性影响,本文将筛选出的数据进行 MEMD 分解,并对分解出的不同频率的 IMF 分量以及残差项 RES 分别建立碳排放预测模型。本文按比例划分训练集与测试集输入模型,整合模型预测结果获得最终碳排放量预测值,完整预测流程见图 3。

本文选用平均绝对误差 (MAE)、均方根误差 (RMSE) 和平均相对百分比误差 (MAPE) 作为指标综合评估模型性能,计算分别见式 (4) ~ (6)。

$$Y_{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

$$Y_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$Y_{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (6)$$

式中 y_i 为碳排放实际值; $\hat{y}_i$ 为碳排放预测值; n 为样本总数。RMSE、MAE、MAPE 可用于衡量预测值与实际值之间的误差,值越小,预测精度越高。

3 实例分析

3.1 影响因素选取

STIRPAT 模型拓展自环境压力模型 (IPAT),可纳入多维度独立变量,计算见式 (7)。

$$I = a P^b A^c T^d e \quad (7)$$

式中 I 为环境压力; a 为模型系数; b 、 c 、 d 分别为人口、经济、技术水平的弹性系数; e 为误差; P 为人口维度; A 为经济维度; T 为技术水平维度。本文基于 STIRPAT 模型初步确定年末总人口、人均 GDP、居民人均消费支出、城镇单位就业人员工资总额、城镇人口、平均每天能源消耗量、科技成果登记数和出口总额等 8 个变量为我国碳排放评价指标,见表 2。

3.2 影响因素筛选

通过皮尔逊相关系数法分析得知 (图 4): 平均每天能源消耗量与碳排放量的皮尔逊系数最高,为 0.986; 城镇单位就业人员工资总额与碳排放量的皮尔逊系数最低,为 0.852。所有影响因素的皮尔逊相关系数均大于 0.8,表明所选因素均与碳排放量

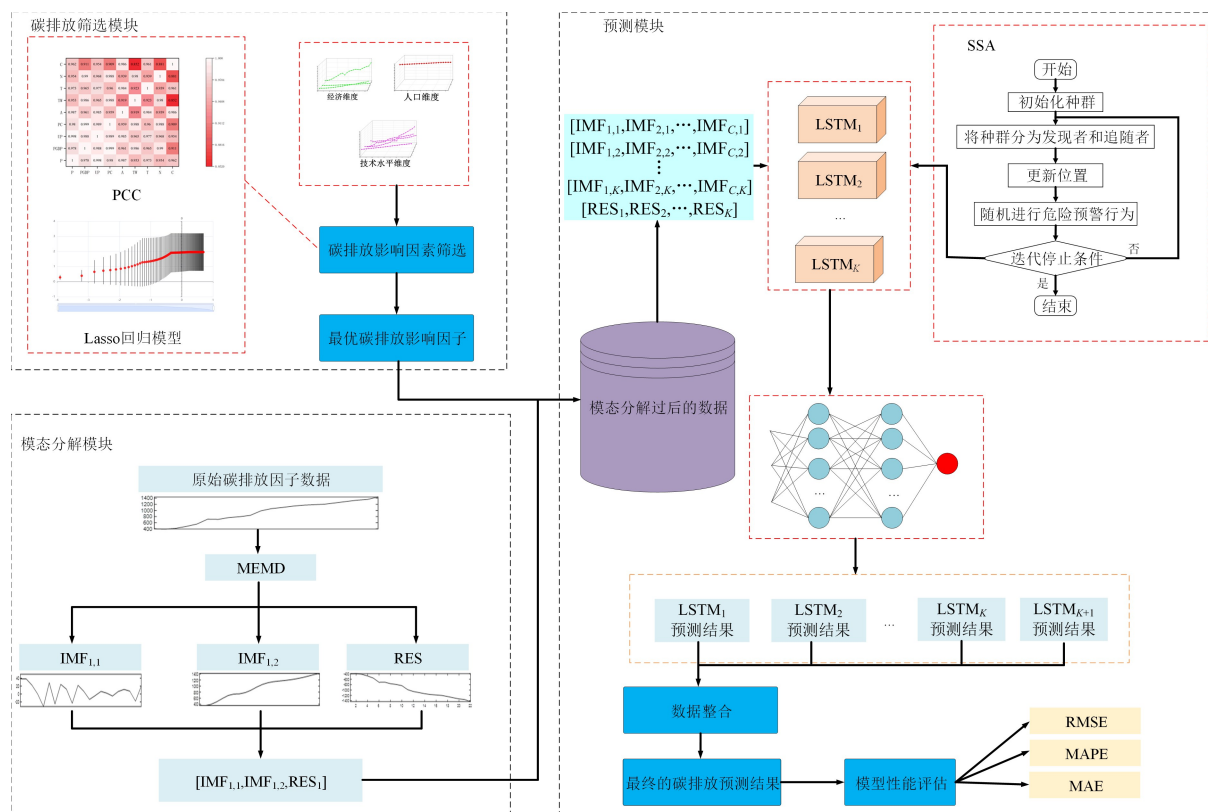


图2 模型框架图

Fig. 2 Model framework diagram

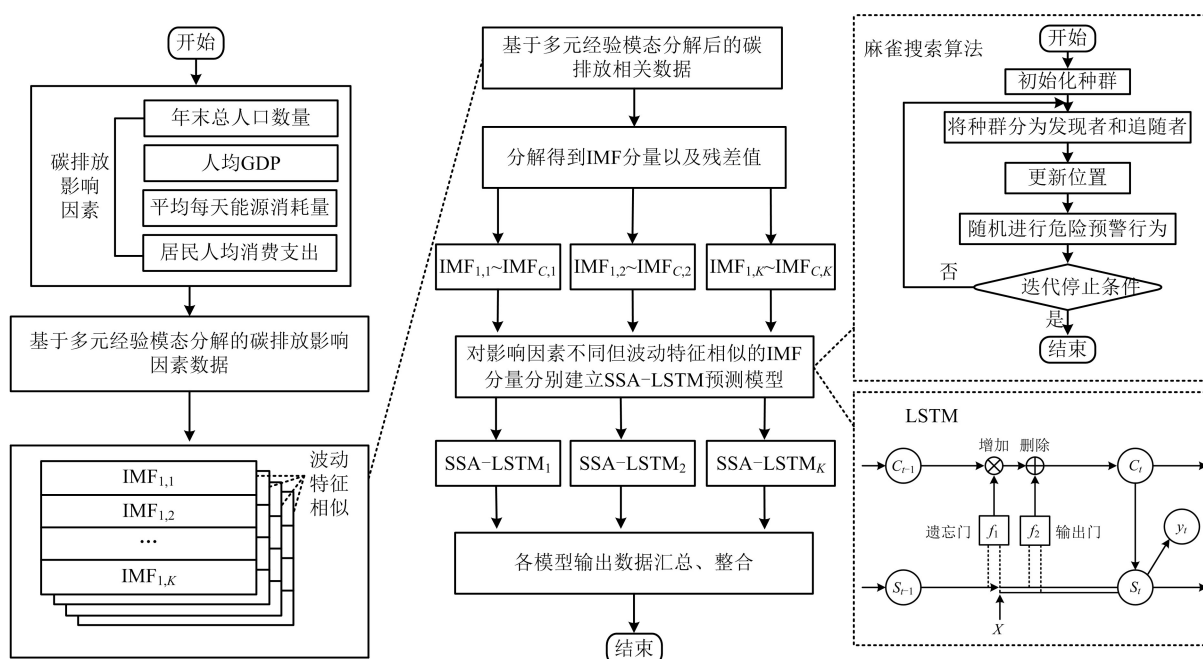


图3 预测流程图

Fig. 3 Forecast flowchart

具有极强相关性,因此本文使用 LASSO 对其进一步筛选。

采用 LASSO 回归(L1 正则化)筛选关键预测变量。通过交叉验证确定最优 λ , 而 λ 取 LASSO 模型

表 2 碳排放影响因素

Table 2 Factors affecting carbon emissions

因素种类	具体因素	缩写	单位
人口维度	年末总人口数量	P	万人
	人均 GDP	PGDP	亿元
经济维度	出口总额	T	亿元
	居民人均消费支出	PC	元
	城镇单位就业人员工资总额	TW	亿元
技术水 平维度	城镇人口	UP	万人
	平均每天能源消耗量	A	万 t 标准煤
	科技成果登记数	N	项

中均方误差最小时所对应的数值。本文依据 LASSO 回归轨迹图(图 5)确定最优 λ ,将最优 λ 参数值代入 LASSO 回归模型中,然后剔除标准化系数为 0 的变量,保留非零系数变量以构建最终模型。影响因素的标准化系数从大到小排序见表 3。其中,正数为正向影响,负数为负向影响。

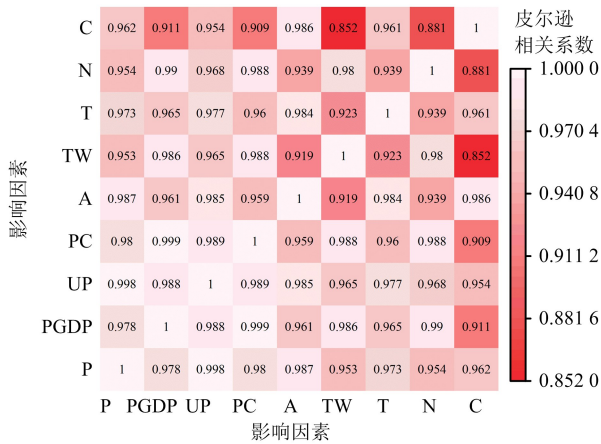


图 4 皮尔逊相关系数热力图

Fig. 4 Pearson correlation coefficient heat map

LASSO 回归模型通过对系数进行惩罚来进行特征选择,因此本文计算的系数可以直观地解释为特征的重要性。据此,筛选出 4 个关键影响因素:年末

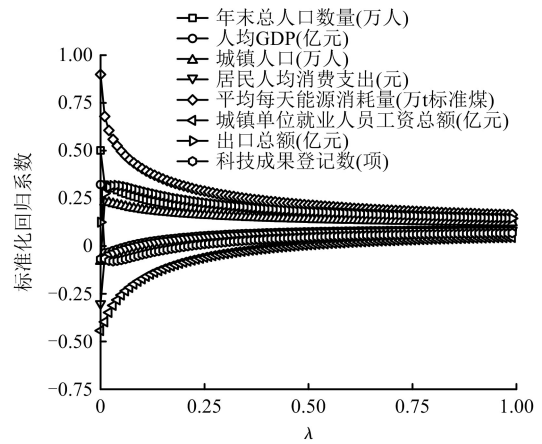


图 5 LASSO 回归轨迹图

Fig. 5 LASSO regression trajectory plot

总人口、人均 GDP、日均能源消耗量、居民人均消费支出。

3.3 数据来源及预处理

本文选用 min - max 标准化方法对原始数据实施归一化预处理。如式(8)所示,本文将所有数据归到(0,1)区间。

$$X_i = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (8)$$

式中 X_i 为归一化后的结果; X 为原始数据; $X_{\max}$ 和 $X_{\min}$ 分别为各序列原始数据的最大值和最小值。

主要数据源于国家统计局官方平台(<https://data.stats.gov.cn/index.htm>)与中国碳核算数据库(CEADs)。其中,2000—2021 年中国碳排放量数据引自 CEADs 数据库,其完整时间序列变化趋势见图 6。

3.4 基于 MEMD 的碳排放数据分解结果

MEMD 可以对多变量时序进行多频率尺度的同步分解。其中,具有复杂波动特征各个变量会被分解为不同频率尺度下波动特征相似、数量相同的多个 IMFs 分量,以及一个 RES 残差分量,这可以有效减少数据的非线性和波动性对时序预测精度的影响。将所选取的 4 个碳排放影响因素进行 MEMD 分解,每个原始数据都被分解为 K 个 IMF 分量, C 维

表 3 影响因素标准化系数

Table 3 Standardized coefficients of factors influencing carbon emissions

变量名	平均每天 能源消耗量	人均 GDP	居民人均 消费支出	年末总 人口数量	城镇单位就业 人员工资总额	科技成果 登记数	城镇 人口	出口 总额
标准化系数	0.95	0.764	-0.728	0.558	-0.405	-0.114	-0.112	0.085

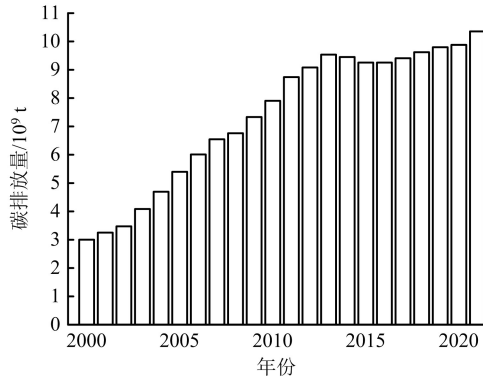


图6 2000—2021年碳排放数据图

Fig. 6 Carbon emissions data chart from 2000 to 2021

原始数据被分解为 $C \times K$ 维特征相似的 IMF 分量, 即 $IMF_{1,1}, IMF_{1,2}, \dots, IMF_{1,K}, \dots, IMF_{C,K}$ 。以年末总人口数 x_1 举例, 可分解为 $IMF_{1,1}, IMF_{1,2}, \dots, IMF_{1,K}$ 。其中, 特征相似是指年末总人口数分解得到的不同 IMF 分量中与人均 GDP 分解得到的不同 IMF 中有一个分量特征相似, 即 $IMF_{1,1}, IMF_{2,1}, \dots, IMF_{C,1}$ 组成 IMF_1 的特征相似, 但 $IMF_1, IMF_2, \dots, IMF_K$ 特征不同。在碳排放预测模块, 针对特征相似但是由不同碳排放特征数据分解得出的 IMF 分量进行建模, 即对 $IMF_1, IMF_2, \dots, IMF_K$ 分别建立 K 个预测模型。图 7 呈现了平均每天能源消耗量 MEMD 方法分解的解析结果。

通过 MEMD 分解算法对通过皮尔逊系数法和 LASSO 综合选择的影响因素数据进行处理。为了提升试验效率, 采用 MEMD 算法将碳排放影响因子解构为 2 个 IMF 及 1 个 RES, 然后对每组进行单独建模。

$[IMF_{1,1}, IMF_{2,1}, IMF_{3,1}, IMF_{4,1}]$

$[IMF_{1,2}, IMF_{2,2}, IMF_{3,2}, IMF_{4,2}]$

$[RES_1, RES_2, RES_3, RES_4]$

3.5 SSA 优化 LSTM 超参数

模型采用 2000—2016 年数据进行训练, 2017—2021 年数据用于测试验证。基于麻雀搜索算法 (SSA) 对 LSTM 实施超参数优化 (种群规模为 40, 迭代次数为 50)。为了评估算法性能, 选取灰狼优化器 (Grey Wolf Optimizer, GWO)、粒子群算法 (Particle Swarm Optimization, PSO) 与 SSA 进行对比分析, 三者的收敛过程见图 8。

3.6 模型预测结果分析

在预测时, 首先利用 MEMD 对 4 个影响因素进行分解, 并将得到的模态分量与残差合理分组; 其

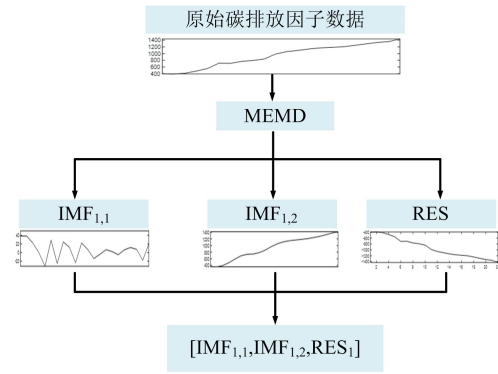


图7 平均每天能源消耗量 MEMD 分解结果

Fig. 7 MEMD decomposition results of the average daily energy consumption

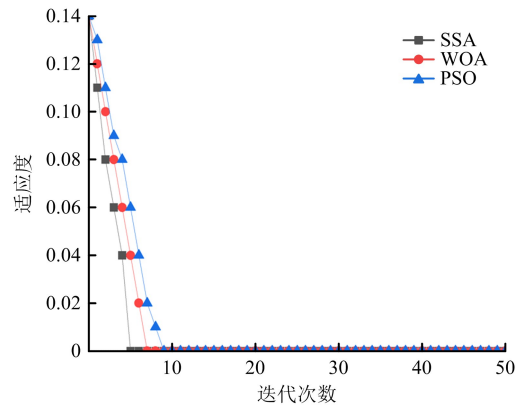


图8 迭代适应度曲线

Fig. 8 Iterative fitness curve

次, 为每个分组单独构建 SSA-LSTM 模型以进行预测; 最后, 综合各分组模型的预测结果得到碳排放总量预测值。根据前人研究, 引入反向传播 (Back Propagation, BP)^[22]、灰狼优化算法-核极限学习机混合模型 (Grey Wolf Optimizer - Kernel Extreme Learning Machine Hybrid Model, GWO-KELM)^[9] 两种模型一起进行预测, 本文试验在 MATLAB R2020b 环境下运行。

按照上述步骤对模型进行训练以及预测, 并将输出的预测结果进行数据整合以对结果进行分析, 见图 9。图 10 与表 4 的对比结果表明: 相较于其他碳排放预测模型, MEMD-SSA-LSTM 模型具有最优拟合优度与最高预测精度, 其输出结果与实际碳排放变化趋势高度吻合。

本文采用 MAE、RMSE 和 MAPE 作为模型性能的核心评价指标进行研究, 分析发现, 单一 LSTM 模型在预测碳排放时存在显著不足, 其预测误差较大, 这在表 5 与图 10 的实证结果中清晰呈现。为了提

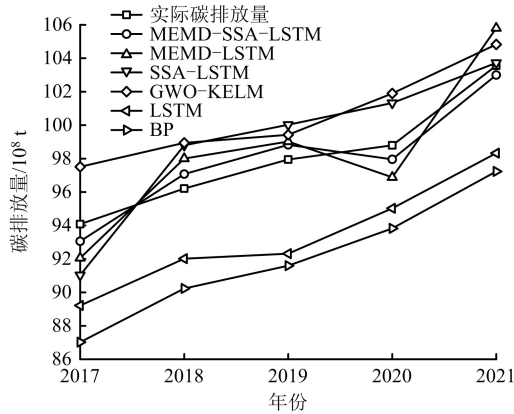


图9 预测结果对比

Fig. 9 Comparison of prediction results

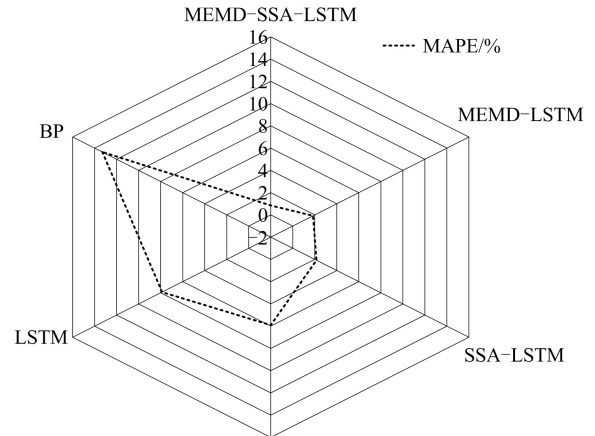
升预测精度,引入 MEMD 对原始碳排放序列进行分解处理,该模型有效提取了数据内在特征并降低了非平稳性。结合 LSTM 预测经 MEMD 分解后的数据,模型误差得到显著改善,具体表现为 MAPE 下降 6.04 百分点,MAE 下降 59.3%,RMSE 下降 59.4%。在此基础上,进一步利用 SSA 优化 LSTM 超参数,构建的 SSA-LSTM 模型性能超越单一 LSTM,MAPE、MAE 和 RMSE 分别再降 5.73 百分点、56.6% 和 54.8%。综合比较表明,融合 MEMD 分解与 SSA 优化的 MEMD-SSA-LSTM 组合模型在碳排放预测中表现最优。

4 结 论

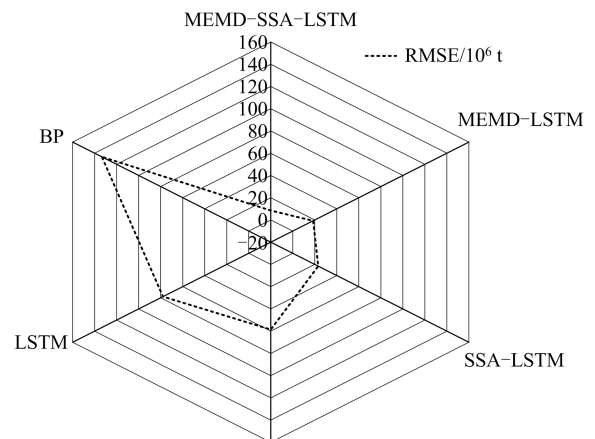
本文围绕中国碳排放预测精度提升的关键问题,构建了基于 LASSO - MEMD - SSA - LSTM 的混合预测模型,并通过实证分析与对比验证,得出以下结论。

(1) LASSO - MEMD - SSA - LSTM 模型显著提高了中国碳排放预测的科学性与准确性。LASSO 回归有效识别并筛选出驱动碳排放的关键影响因素,增强了模型输入变量的代表性与解释性;MEMD 技术通过多尺度分解充分提取了原始序列中的复杂波动与趋势信息,显著降低了非平稳序列对预测建模的干扰,使各子序列更具规律性和可预测性;SSA 算法实现了对 LSTM 超参数的智能优化,克服了传统人为调参的随机性与低效性问题,显著提升了模型收敛速度与泛化能力。三者有机结合,共同保障了最终预测结果的可靠性与稳健性。

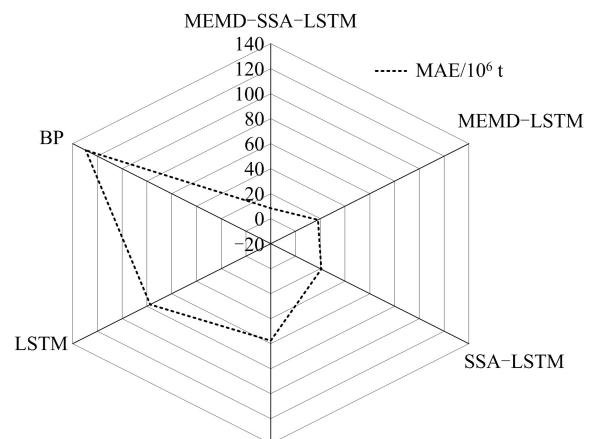
(2) 模型验证结果显示,本文提出的混合模型在拟合优度、误差控制及趋势捕捉方面均优于单一 LSTM 模型以及未分解优化的基准模型,其平均绝对



(a) 各模型预测MAPE对比图



(b) 各模型预测RMSE对比图



(c) 各模型预测MAE对比图

图10 各模型预测误差雷达对比

Fig. 10 Radar chart of prediction errors by various models

误差(MAE)、均方根误差(RMSE)显著降低,表明该模型能更科学合理地处理碳排放数据的非线性、非平稳性及高噪声特性。

(3) 基于本文模型的预测结果,中国碳排放量预计在 2030 年前实现达峰,这与国家“双碳”目标高

表4 实际碳排放量与各个模型预测对比

Table 4 Comparison between actual carbon emissions and predictions from various models

年份	实际碳排放量/ 10 ⁶ t	预测碳排放量/10 ⁶ t					
		MEMD - SSA - LSTM	MEMD - LSTM	SSA - LSTM	GWO - KELM	LSTM	BP
2017	9 408. 17	9 305. 99	9 205. 82	9 103. 12	9 751. 24	8 921. 32	8 702. 32
2018	9 621. 12	9 707. 32	9 800. 21	9 879. 32	9 894. 42	9 201. 54	9 023. 24
2019	9 794. 76	9 882. 56	9 902. 21	10 002. 34	9 942. 53	9 231. 65	9 159. 43
2020	9 879. 75	9 795. 93	9 688. 21	10 132. 43	10 189. 53	9 502. 42	9 382. 72
2021	10 356. 27	10 300. 27	10 580. 37	10 372. 31	10 482. 87	9 832. 43	9 723. 43

表5 6种模型的误差仿真结果

Table 5 Error simulation results of 6 kinds of models

模型	RMSE/10 ⁶ t	MAE/10 ⁶ t	MAPE/%
MEMD - SSA - LSTM	8. 45	8. 32	0. 85
MEMD - LSTM	18. 50	18. 10	1. 84
SSA - LSTM	23. 10	20. 80	2. 15
GWO - KELM	58. 60	57. 60	5. 91
LSTM	77. 90	77. 40	7. 88
BP	133. 50	129. 40	13. 30

度吻合,也进一步印证了模型在服务于中长期政策制定与评估方面的应用潜力。

未来研究可进一步引入社会行为、政策强度等隐性变量,拓展影响因素体系;同时,探索其他先进信号分解方法与深层神经网络结构,以更好地应对极端波动与复杂时序的依赖关系,进一步提升预测模型的适应性与前瞻性。此外,本文目前主要基于国家级年度数据验证模型性能。在更高时空分辨率(如小时级、城市)层面验证本模型的适用性,是未来的重要研究方向。这一工作具有双重意义:一方面,能够严格评估 MEMD - SSA - LSTM 框架在处理高频、强波动数据时的稳健性与预测效能;另一方面,研究成果将直接推动该模型在智慧环保动态监测、城市级碳排放实时核算等精细化场景中的实际应用。

参考文献 (References):

- [1] WANG M. Predicting China's energy consumption and CO₂ emissions by employing a novel grey model [J]. Energies, 2024, 17(21): 5256 - 5256.
- [2] 姜丽娜,高亚锋,江欣妍,等. 城市建设碳排放核算规则与数据质量管理策略[J/OL]. 安全与环境学报: 2482

1 - 12 [2025 - 09 - 28]. <https://doi.org/10.13637/j.issn.1009-6094.2025.0163>.

JIANG L N, GAO Y F, JIANG X Y, et al. Carbon emission accounting rules and data quality management strategies for urban construction[J/OL]. Journal of Safety and Environment: 1 - 12 [2025 - 09 - 28]. <https://doi.org/10.13637/j.issn.1009-6094.2025.0163>.

- [3] 杨青,江宇航,吴婵媛,等. 基于 CNN - LSTM - Attention 的中国省域交通运输业碳达峰预测研究[J/OL]. 安全与环境学报: 1 - 12 [2025 - 09 - 16]. <https://doi.org/10.13637/j.issn.1009-6094.2024.1993>.

YANG Q, JIANG Y H, WU C Y, et al. Prediction of carbon emission peak in China's provincial transportation sector based on CNN - LSTM - Attention model[J/OL]. Journal of Safety and Environment: 1 - 12 [2025 - 09 - 16]. <https://doi.org/10.13637/j.issn.1009-6094.2024.1993>.

- [4] 张新生,张红文,聂达文. 基于二次分解与 MAML - MHA - DELM 的电力行业碳排放预测模型研究[J]. 安全与环境学报, 2025, 25(9): 3386 - 3399.

ZHANG X S, ZHANG H W, NIE D W. A carbon emission prediction model for the power industry based on dual decomposition and MAML - MHA - DELM [J]. Journal of Safety and Environment, 2025, 25(9): 3386 - 3399.

- [5] 刘淳森,曲建升,葛钰洁,等. 基于 LSTM 模型的中国交通运输业碳排放预测[J]. 中国环境科学, 2023, 43(5): 2574 - 2582.

LIU C S, QU J S, GE Y J, et al. Prediction of carbon emissions from China's transportation industry based on LSTM model[J]. China Environmental Science, 2023, 43(5): 2574 - 2582.

- [6] 杜涵蓓,赵立君,刘臣炜,等. 基于 LEAP 模型和 KAYA 模型的主城区碳达峰预测及不确定性分析[J]. 生态与农村环境学报, 2022, 38(8): 983 - 991.

DU H B, ZHAO L J, LIU C W, et al. Carbon peak

- prediction and uncertainty analysis of main urban area based on LEAP model and KAYA model[J]. *Journal of Ecology and Rural Environment*, 2022, 38 (8): 983 - 991.
- [7] YANG H, O'CONNELL F J. Short-term carbon emissions forecast for aviation industry in Shanghai[J]. *Journal of Cleaner Production*, 2020, 275: 122734 - 122745.
- [8] 王阳,唐朝晖,王紫勋,等. 选用改进高斯过程回归模型的碳排放短期预测[J]. *计算机工程与应用*, 2018, 54(23): 246 - 251, 258.
- WANG Y, TANG Z H, WANG Z X, et al. Short-term carbon emission forecasting using improved Gaussian process regression model[J]. *Computer Engineering and Applications*, 2018, 54(23): 246 - 251, 258.
- [9] 张新生,魏志臻,陈章政,等. 基于 LASSO - GWO - KELM 的工业碳排放预测方法研究[J]. *环境工程*, 2023, 41(10): 141 - 149.
- ZHANG X S, WEI Z Z, CHEN Z Z, et al. Research on industrial carbon emission prediction method based on LASSO - GWO - KELM[J]. *Environmental Engineering*, 2023, 41(10): 141 - 149.
- [10] 池小波,续泽晋,贾新春,等. 基于 WPD - ISSA - CA - CNN 模型的电厂碳排放预测[J]. *控制工程*, 2025, 32(8): 1387 - 1394.
- CHI X B, XU Z J, JIA X C, et al. Carbon emission prediction of power plant based on WPD - ISSA - CA - CNN model[J]. *Control Engineering*, 2025, 32(8): 1387 - 1394.
- [11] 张雯,吴志彬,徐玖平. 基于 EMD - PSO - LSSVM 的碳排分解集成预测方法[J]. *控制与决策*, 2022, 37(7): 1837 - 1846.
- ZHANG W, WU Z B, XU J P. Carbon emission decomposition ensemble forecasting method based on EMD - PSO - LSSVM[J]. *Control and Decision*, 2022, 37(7): 1837 - 1846.
- [12] 李柚洁,杨萍,赵顺昱,等. 基于 VMD - RE - LSTM 的碳排放发展趋势捕捉及短期预测[J]. *环境工程*, 2023, 41(增刊 2): 279 - 285, 249.
- LI Y J, YANG P, ZHAO S Y, et al. Capturing development trend and short-term prediction of carbon emission based on VMD - RE - LSTM[J]. *Environmental Engineering*, 2023, 41(Suppl. 2): 279 - 285, 249.
- [13] 王珂珂,牛东晓,甄皓,等. 基于 WOA - ELM 模型的中国碳排放预测研究[J]. *生态经济*, 2020, 36(8): 20 - 27.
- WANG K K, NIU D X, ZHEN H, et al. Research on China's carbon emission prediction based on WOA - ELM model[J]. *Ecological Economy*, 2020, 36(8): 20 - 27.
- [14] 陶洋,祝小钧,杨柳. 基于皮尔逊相关系数和信息熵的多传感器数据融合[J]. *小型微型计算机系统*, 2023, 44(5): 1075 - 1080.
- TAO Y, ZHU X J, YANG L. Multi-sensor data fusion based on Pearson correlation coefficient and information entropy[J]. *Journal of Chinese Computer Systems*, 2023, 44(5): 1075 - 1080.
- [15] LOH C, WU T, HUANG E N. Data files from "Application of the Empirical Mode Decomposition-Hilbert Spectrum Method to Identify Near-Fault Ground-Motion Characteristics and Structural Responses"[J]. *Bulletin of the Seismological Society of America*, 2001, 91(5): 1387 - 1387.
- [16] 张茵栋,陈丽贤,张德康,等. 基于 EMD - LSTM 多变量输入的极端海况预报模型[J]. *太阳能学报*, 2024, 45(3): 193 - 200.
- ZHANG H D, CHEN L X, ZHANG D K, et al. Extreme sea state prediction model based on EMD - LSTM multivariate input[J]. *Acta Energiæ Solaris Sinica*, 2024, 45(3): 193 - 200.
- [17] 秦进,胡冉,毛成辉,等. 高速铁路日常客运量的 EMD - Informer 组合预测方法[J]. *铁道学报*, 2024, 46(2): 1 - 11.
- QIN J, HU R, MAO C H, et al. A combined EMD - Informer forecasting method for daily high-speed railway passenger flow[J]. *Journal of the China Railway Society*, 2024, 46(2): 1 - 11.
- [18] REHMAN N, MANDIC P D. Multivariate empirical mode decomposition[J]. *Proceedings: Mathematical, Physical and Engineering Sciences*, 2010, 466(2117): 1291 - 1302.
- [19] HUANG H, MAO J, KANG L, et al. A dynamic graph deep learning model with multivariate empirical mode decomposition for network-wide metro passenger flow prediction[J]. *Computer-Aided Civil and Infrastructure Engineering*, 2024, 39(17): 2596 - 2618.
- [20] 贺毅岳,陈倩倩,高妮,等. 基于 MEMD - WGAN - GP 的黄金期货价格集成预测研究[J]. *系统工程理论与实践*, 2025, 45(6): 2033 - 2049.
- HE Y Y, CHEN Q Q, GAO N, et al. Integrated forecasting research on gold futures price based on MEMD - WGAN - GP model[J]. *Systems Engineering: Theory & Practice*, 2025, 45(6): 2033 - 2049.
- [21] XUE J, SHEN B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. *Systems Science & Control Engineering*, 2020, 8(1): 22 - 34.
- [22] 郑琰,夏朝泽,肖玉杰,等. 基于 BP 神经网络的江苏省多维度碳排放预测[J]. *环境科学*, 2025, 46

(6): 3485–3495.

ZHENG Y, XIA C Z, XIAO Y J, et al. Multi-dimensional carbon emission prediction in Jiangsu

Province based on BP neural network[J]. Environmental Science, 2025, 46(6): 3485–3495.

Research on China's carbon emission prediction based on LASSO – MEMD – SSA – LSTM

WANG Qingrong¹, LIU Xinkang¹, ZHU Changfeng², WANG Junjie¹

(1 School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;

2 School of Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: This study presents a novel hybrid prediction model that integrates the Least Absolute Shrinkage and Selection Operator (LASSO), Multivariate Empirical Mode Decomposition (MEMD), Sparrow Search Algorithm (SSA), and Long Short-Term Memory (LSTM) network to significantly enhance the accuracy of carbon emission forecasting. The methodological framework was implemented in a structured, multi-stage process. Initially, within the Stochastic T Impacts by Regression on PAT (STIRPAT) framework, key influencing factors such as year-end population, per capita GDP, daily energy consumption, and per capita consumption expenditure were identified. These variables were then refined using a two-step statistical approach: Pearson correlation analysis was conducted to identify relationships among the variables, followed by LASSO regression, which penalized coefficients and eliminated multicollinearity, thereby ensuring the selection of the most robust predictors. Subsequently, the refined multivariable time series data were processed using the MEMD technique. MEMD adaptively decomposed the complex, non-stationary signals into a finite set of simpler and more stable sub-series components, specifically several Intrinsic Mode Functions (IMFs) representing different frequency bands, along with a single monotonic Residual (RES) component that denotes the long-term trend. This critical step effectively mitigated the inherent nonlinearity and non-stationarity of the data, thus simplifying the forecasting task. Each IMF and the RES component then served as inputs for individual LSTM networks. To optimize the performance of each LSTM model, key hyperparameters—including the number of hidden neurons, training iterations, and learning rate—were automatically and intelligently tuned using the Sparrow Search Algorithm (SSA). This optimization aimed to minimize prediction errors for each component, enhancing convergence speed and preventing overfitting. The final carbon emission forecast was obtained by aggregating the predictions from all the IMF and RES components. The model was rigorously validated using China's carbon emission data from 2000 to 2021. Its performance was benchmarked against several other models, including the BP neural network, standalone LSTM, GWO – KELM, SSA – LSTM, and MEMD – LSTM. The proposed LASSO – MEMD – SSA – LSTM model demonstrated superior performance, achieving the lowest error metrics: a Mean Absolute Error (MAE) of 8.32, a Root Mean Square Error (RMSE) of 8.45, and an impressively low Mean Absolute Percentage Error (MAPE) of 0.85%. An ablation analysis underscored the contribution of each component: the MEMD decomposition alone led to reductions in MAPE, MAE, and RMSE by 6.04 percentage points, 59.3%, and 59.4%, respectively, compared to a standalone LSTM, with further significant enhancements achieved through SSA optimization. In conclusion, the integrated model demonstrates high effectiveness in managing the volatility and complexity of carbon emission data, providing a reliable and accurate tool for medium- and long-term forecasting. This capability establishes a solid foundation for strategic energy and environmental policy planning. Future studies will focus on incorporating socio-political variables and expanding the model's application to high-resolution spatial-temporal analyses.

Key words: environmental engineering; carbon emission prediction; LASSO regression; multivariate empirical mode decomposition; sparrow search algorithm; long short-term memory network