DOI: 10.13718/j.cnki.xdzk.2026.05.017
燕振刚，杨发发，王耀东. 基于 LSTM-XGBoost 组合模型的农田 CO_2 排放浓度预测研究[J]. 西南大学学报(自然科学版)，2026，48(5)：209-221.

基于 LSTM-XGBoost 组合模型的农田 CO_2 排放浓度预测研究

燕振刚，　杨发发，　王耀东

甘肃农业大学信息科学技术学院，甘肃 兰州 730070

摘要：针对传统模型在农田 CO_2 排放浓度预测中的局限性，构建了长短期记忆网络(Long Short-Term Memory，LSTM)、极限梯度提升算法(eXtreme Gradient Boosting，XGBoost)、门控循环单元(Gated Recurrent Unit，GRU)和随机森林(Random Forest，RF)4 种机器学习模型，并通过贝叶斯优化和随机搜索优化超参数。结果表明：LSTM 擅长捕捉长期时序特征($R^2=0.86$，$MAPE=5.97$)，XGBoost 在短期非线性拟合中表现优异($R^2=0.84$，$MAPE=4.62$)。为弥补单一模型的局限性，构建了基于 LSTM 时序特征拼接的 XGBoost 组合模型，有效兼顾长期趋势与局部波动拟合能力，提升了预测性能($R^2=0.94$，$RMSE=22.81$，$MAPE=2.21$)。

关　键　词：LSTM-XGBoost 组合模型；农田 CO_2 排放预测；循环神经网络模型；集成树模型

中图分类号：TP18　　**文献标识码**：A

文 章 编 号：1673-9868(2026)05-0209-13

开放科学(资源服务)标识码(OSID)：

Research on Prediction of Farmland CO_2 Emission Concentration Based on an LSTM-XGBoost Hybrid Model

YAN Zhengang，　YANG Fafa，　WANG Yaodong

College of Information Science and Technology，Gansu Agricultural University，Lanzhou Gansu 730070，China

Abstract：Aiming at the limitations of traditional models in predicting farmland CO_2 emission concentration，this study constructed four machine learning models，namely long short-term memory (LSTM)，eXtreme Gradient Boosting (XGBoost)，gated recurrent unit (GRU) and random forest (RF)，and optimized their hyperparameters viaBayesian and random search methods. Results showed that LSTM excelledin capturing long-term temporal patterns ($R^2=0.86$，$MAPE=5.97$)，while XGBoost performed well in short-term nonlinear fitting ($R^2=0.84$，$MAPE=4.62$). To overcome the limitations of a single

收稿日期：2025-07-02
基金项目：甘肃省重点研发计划项目(21YF5FA095)；国家自然科学基金项目(31660347)；甘肃省高等学校创新项目(2021A-057)；甘肃省财政厅项目(GSCZZ-20160909-03)。
作者简介：燕振刚，博士，教授，硕士研究生导师，主要从事农业信息技术应用方向的教学与研究。

model, this paper constructs a combined XGBoost model based on LSTM time series feature splicing. The model effectively balances the fitting performance of long-term trends and local fluctuations, and improves the overall prediction accuracy($R^2=0.94$, $RMSE=22.81$, $MAPE=2.21$).

Key words: LSTM-XGBoosthybrid model; farmland CO_2 emission prediction; recurrent neural network model; ensemble tree model

在农田 CO_2 排放预测研究中，基于过程的生物地球化学模型，如农业技术转移决策支持系统(Decision Support System for Agrotechnology Transfer, DSSAT)、脱氮-分解模型(DeNitrification-DeComposition, DNDC)、环境政策综合气候模型(Environmental Policy Integrated Climate, EPIC)、农业生产系统模拟模型(Agricultural Production Systems sIMulator, APSIM)，虽取得一定成效，但存在参数需求高、建模复杂、计算资源消耗大及环境适应性差等局限[1-2]。相比之下，机器学习方法因其强大的模式识别能力和良好的泛化性能，已成为农业碳排放预测的重要技术路径[3]。机器学习能从高频、长期的农田 CO_2 排放浓度监测数据中挖掘复杂的非线性关联关系，并适应多变的环境和管理条件[4]。种植前，农户可通过预测模型评估碳排放量，调整施肥、灌溉量和耕作措施等。因此，建立基于机器学习的 CO_2 排放浓度预测模型，不仅可弥补传统模型的不足，更可助力精准农业的发展与碳减排目标的实现，推动农业智能化、前瞻化管理迈向新阶段。

近年来，机器学习方法因其强大的非线性拟合与时序建模能力被应用于农田 CO_2 排放浓度预测。文献[5]采用极限梯度提升算法(eXtreme Gradient Boosting, XGBoost)模型提升农田碳交换预测精度，弥补传统模型对非线性关系处理的不足。文献[6]将长短期记忆网络(long short term memory, LSTM)与 XGBoost 应用于 CO_2 与 CH_4 排放趋势预测，验证了二者在处理复杂时序数据中的有效性。文献[7]进一步探索了 XGBoost、卷积神经网络(Convolutional Neural Network, CNN)与 LSTM 模型融合在碳排放预测中的潜力，取得了优异结果。文献[8]构建的 CNN-GRU 模型与文献[9]提出的 CNN-LSTM 模型，通过将擅长局部特征提取的卷积神经网络与擅长时序依赖建模的循环神经网络相结合，能够有效克服单一模型的局限性，在不同领域证实了混合深度学习架构的优越性。这一设计思路为本文解决农田 CO_2 浓度多因子耦合时序预测问题提供了研究思路。总体来看，融合深度学习与树模型优势的组合模型成为主流趋势，尤其在非线性强、时序性显著的农业碳排放预测中表现出更高精度与稳定性。本文构建的基于 LSTM 时序特征拼接的 XGBoost 组合模型正是对该趋势的实践，可显著提升预测精度，为农田 CO_2 排放浓度建模提供了新路径与理论支撑。

1 数据来源与预处理

1.1 数据来源

预测模型所用数据由基于 STM32 的农田 CO_2 排放浓度测量全自动静态箱系统与基于 Elasticsearch(ES)数据库的农田物联网云存储系统采集得到，两套系统于2024年8月和9月在苜蓿大田中部署，采样周期为苜蓿一茬完整生长周期。具体的数据采集现场示意图见图1。基于 STM32 的农田 CO_2 排放浓度测量全自动静态箱系统是一套基于32位微控制器(STM32)的全自动农田数据采集系统，箱体大小为75 cm×75 cm×75 cm，箱体底部配有插入土壤5 cm的密封水槽，系统可以全自动定时采集农田的 CO_2 排放量(图2)、环境温湿度和箱内温湿度数据；基于 ES 数据库构建的农田物联网云存储系统，通过部署

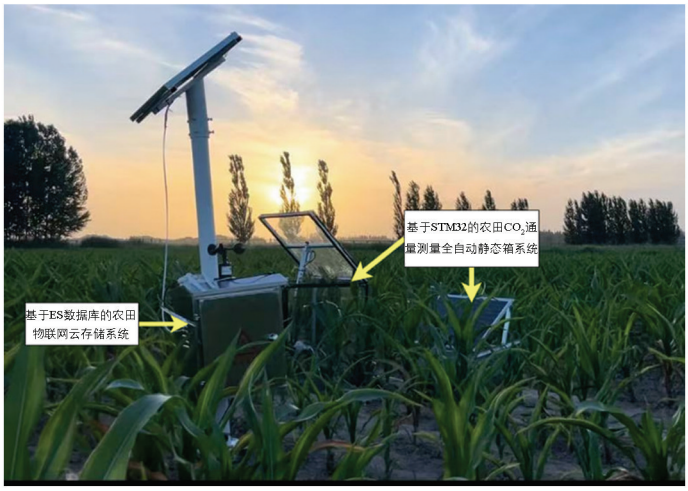


图1 农田 CO_2 排放浓度与环境数据采集现场示意图

光照强度、土壤温湿度、土壤电导率、风速传感器，土壤测量传感器均掩埋在地表 20～40 cm 下，系统实时接收前端采集模块上传的农田环境数据，并利用 ES 实现数据的高效存储。

本研究所采集的数据包含 CO_2 排放浓度、环境温湿度、光照强度、风速、土壤电导率及土壤温湿度。具体数据类型及变量说明见表 1。

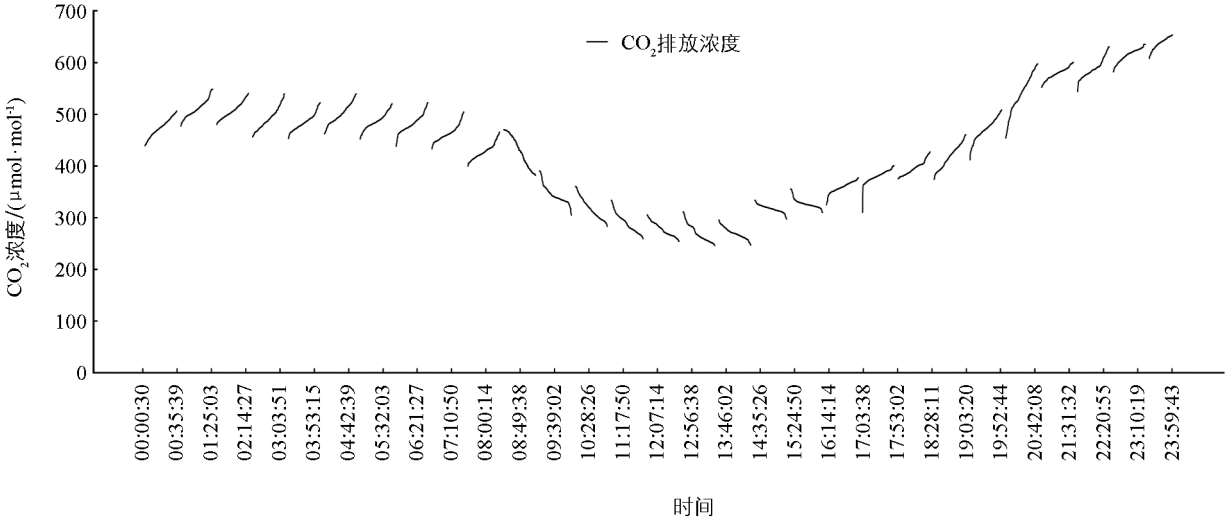


图 2　农田 CO_2 排放浓度日采集数据图

表 1　采集数据类型及变量说明表

变量	数据类型	含义	单位	说明
C	Float	CO_2 排放浓度	$\mu mol \cdot mol^{-1}$	静态箱内的 CO_2 排放浓度变化
T_a	Float	环境温度	℃	监测区域的大气温度
H_a	Float	环境湿度	%	监测区域的大气湿度
T_i	Float	箱内温度	℃	静态箱内部的温度，反映箱内微气候
H_i	Float	箱内湿度	%	静态箱内的湿度，与 CO_2 排放浓度变化密切相关
I	Int	光照强度	Lux	单位面积上接收到的太阳辐射能量
v	Int	风速	$m \cdot s^{-1}$	监测区域的风速
EC	Float	土壤电导率	$mS \cdot cm^{-1}$	土壤电导率反映土壤中溶解盐分浓度，影响植物水分吸收、光合作用和呼吸作用
T_s	Float	土壤温度	℃	土壤温度会影响微生物活性和根系呼吸等生物过程
W_s	Float	土壤湿度	%	土壤湿度会影响植物根系的呼吸作用
t	Datetime	时间步长		数据采集的时间戳

1.2　数据预处理

1.2.1　数据集降频

本研究中两套采集系统经过时间戳对齐后的数据采集周期为 20 min，每个周期间隔 30 min，数据采集频率为 3 s/次。采用均值降频法将采样频率从每 3 s 降频至 1 min，降频后的部分数据如表 2 所示，降频公式如下：

$$x_{\text{mean}}(t)=\frac{1}{N}\sum_{i=1}^{N}x(t_i),\ t_i\in[t-\Delta t,\ t] \tag{1}$$

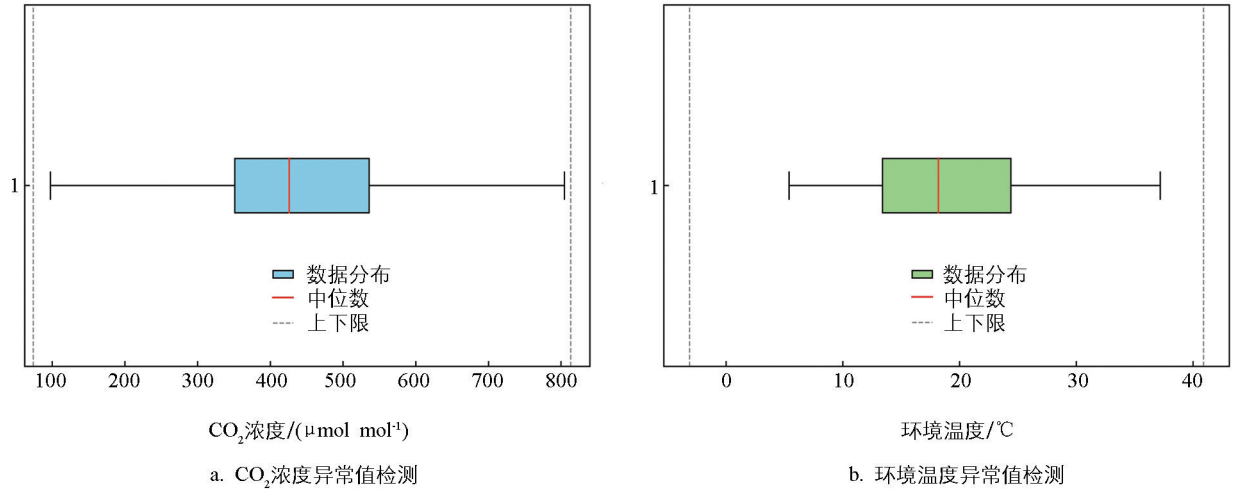
其中：$x_{\text{mean}}(t)$表示降频后的 1 min 平均数据；$N=20$ 表示每分钟的采样点数；t_i 表示采样时间点，$\Delta t=1$ min。按 1 min 的时间窗口划分数据，在每个时间窗口内计算光照强度、土壤湿度、温湿度等变量的均值，将降频后的数据作为预测模型的输入。

表2 降频后的数据示例(部分)

时间	C/(μmol·mol^{-1})	H_a/%	T_a/℃	H_i/%	T_i/℃	I/Lux	EC/(mS·cm^{-1})	T_s/℃	W_s/%
19:58	419	29.2	28	53.8	27.1	1 023	535	21.6	244
19:59	422	28.7	28	56.1	26.9	997	536.5	21.6	244

1.2.2 异常值处理

本研究采用3σ原则判定异常值，超出$\mu\pm3\sigma$的值被视为异常并替换为相邻点的均值[10]。同时，根据实际场景设定环境变量合理范围，对于明显不符合实际的值进行清除。对于异常值的处理，采用邻点均值替换法：若异常点位于数据序列中间位置，则在其前后各取一个最近的正常数据点并计算均值以替换该异常值；若异常点位于序列首尾，则在单侧连续选取最近的2个正常点计算均值进行替换，保障替换值的合理性。本实验中，经均值替换与直接清除的异常值占总数据量的4.91%。对所有数据进行异常值处理，部分数据处理后的检测结果如图3所示。



a. CO_2浓度异常值检测　　b. 环境温度异常值检测

图3 异常值检测图(部分)

1.2.3 数据归一化与标准化

如(2)式所示，采用Min-Max归一化方法将所有特征值映射到[0，1]区间以消除不同变量的量纲差异[11]；如(3)式所示，对训练完整的预测数据进行反归一化操作，将数据还原到真实值[11]。

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

$$x = x' \times (X_{\max} - X_{\min}) + X_{\min} \tag{3}$$

式中：x'为归一化后的值；x代表原始特征值；$x_{\min}$代表特征的最小值；$x_{\max}$代表特征的最大值。

如(4)式所示，采用Z-Score方法对数据进行标准化处理；如式(5)式所示，对预测结果的数据进行反标准化处理。

$$x_{\text{std}} = \frac{x - \mu}{\sigma} \tag{4}$$

$$x = x_{\text{std}} \times \sigma + \mu \tag{5}$$

式中：x_{std}表示标准化后的值；x表示原始特征值；μ表示特征的均值；σ表示特征的标准差。

1.3 数据集构建

本研究的数据经过降频处理后的数据量为3万条，将数据集按时间顺序划分为80%的训练集和20%的验证集，确保时间序列的完整性并模拟实际应用中的预测任务。为避免数据泄露问题并构建适合模型的样本，本研究采用了滑动窗口方法，该方法通过在时间序列上滑动一个固定的窗口，从序列中提取连续的子序列作为样本。每个样本的输入特征只包含当前时间步及其之前的信息，确保未来信息不会被泄露。

2　模型构建

2.1　单一模型选取与优化

为评估不同模型在农田 CO_2 排放浓度预测中的表现，本文选取了 LSTM 、XGBoost 、GRU 和 RF 模型，并分别采用随机搜索与贝叶斯优化进行超参数调优[12-13]。输入变量包括农田 CO_2 浓度、温湿度(环境、箱内、土壤)、土壤电导率、风速与光照。

LSTM 与 GRU 模型用于建模时间序列特征，通过构建滞后变量增强其对时序依赖的学习能力，采用随机搜索在设定参数空间内进行优化，结合 5 折时间序列交叉验证，以均方根误差(Root Mean Square Error，*RMSE*)作为评估指标筛选最优超参数。

RF 与 XGBoost 模型侧重于捕捉输入变量之间的非线性关系，采用贝叶斯优化策略提升搜索效率，同样以 *RMSE* 为核心指标，选择最优配置。最终，4 种模型均在训练集上完成训练，并在验证集上评估性能，为后续混合建模提供基础支持。

2.2　加权组合模型的构建

本研究构建了多种基于 LSTM 和 XGBoost 的加权组合模型进行对比分析。加权组合模型是一种通过对多个基模型的预测结果进行加权平均，从而提升整体模型性能的方法。该方法尤其适用于单一模型性能无法达到理想水平，或者多个模型性能相似但在不同方面具有优势和劣势的情况[14]。本研究采用了基于决定系数(Coefficient of Determination，R^2)的加权方式、基于 *RMSE* 倒数的加权方式、基于平均绝对误差(Mean Absolute Error，*MAE*)倒数的加权方式、基于 R^2、*RMSE* 和 *MAE* 的平均加权方式以及网格搜索方法的加权组合方式，并从中筛选出最佳的加权组合方式。

基于 R^2 值的加权方式以 R^2 值作为评价标准，R^2 值越高，模型对数据的拟合效果越好，模型获得权重越大，计算公式为：

$$W_{R^2}=\frac{s}{\sum_i s_i} \tag{6}$$

式中：s 表示被评估基模型的 R^2 值；s_i 表示第 i 个参与组合的基模型对应的 R^2 值；$\sum_i s_i$ 是所有参与加权组合的基模型 R^2 值的和。通过这种方式，具有较高 R^2 值的模型会在加权组合中占据更大的份额。

基于 *RMSE* 倒数的加权方式是为了衡量误差较小的模型，采用 *RMSE* 的倒数作为权重。对于误差指标，较小的值表示更优的模型，因此将获得更大的权重。基于 *RMSE* 倒数的加权计算公式为：

$$W_{RMSE}=\frac{1}{e_{RMSE}+\epsilon}\times\left(\sum_i \frac{1}{e_{RMSE,i}+\epsilon}\right)^{-1} \tag{7}$$

式中：e_{RMSE} 为被评估基模型的 *RMSE*；$e_{RMSE,i}$ 为第 i 个参与组合的基模型对应的 *RMSE*；ϵ 为防止分母为零所引入的小常数，用于避免除零错误；$\sum_i \frac{1}{e_{RMSE,i}+\epsilon}$ 为所有基模型误差指标倒数的总和。误差越小的模型对应的倒数越大，权重也越大。

同理可得基于 *MAE* 倒数的加权计算公式为：

$$W_{MAE}=\frac{1}{e_{MAE}+\epsilon}\times\left(\sum_i \frac{1}{e_{MAE,i}+\epsilon}\right)^{-1} \tag{8}$$

式中：e_{MAE} 为被评估基模型的 *MAE*；$e_{MAE,i}$ 为第 i 个参与组合的基模型对应的 *MAE*。

计算出每个基模型在 R^2、*RMSE* 和 *MAE* 上的权重后，按照(9)式求出平均权重，以此作为该基模型在最终加权组合模型中的权重。

$$W_{\text{average}}=\frac{W_{R^2}+W_{RMSE}+W_{MAE}}{3} \tag{9}$$

通过这种方式，可以综合考虑加权组合模型在不同评估指标下的表现，从而进一步提升其性能。

除了使用 R^2 加权、误差倒数法和平均加权计算权重外，本研究还通过网格搜索进一步优化加权组合模型的性能[15]。网格搜索的目的是通过对 LSTM 和 XGBoost 两个基模型的权重进行调优，找到最优的权

重组合，从而提升加权组合模型的预测性能。在网格搜索中，两个基模型的权重在区间[0.1，0.9]上均匀取值，共有90个可能的权重组合，确保$w_{\text{lstm}}+w_{\text{xgb}}=1$。采用5折交叉验证，在每个权重组合下进行训练和验证，以减小过拟合风险，并确保加权组合模型在不同数据子集上的泛化能力。选择负均方误差作为评分标准，在网格搜索过程中，选择具有最小误差的权重组合作为最优解[16]。具体的加权组合计算公式为：

$$y_{\text{ensemble}}=w_{\text{lstm\_opt}}\times y_{\text{lstm}}+w_{\text{xgboost\_opt}}\times y_{\text{xgb}} \tag{10}$$

式中：y_{lstm} 和 y_{xgb} 分别为LSTM和XGBoost基模型的预测结果；$w_{\text{lstm\_opt}}$ 和 $w_{\text{xgboost\_opt}}$ 为通过网格搜索得到的最优权重。

2.3 基于LSTM时序特征拼接的XGBoost组合模型设计

混合预测模型是一种将多种不同的预测方法或算法结合起来使用的技术，可以提高预测的准确性、鲁棒性和泛化能力[17]。与单一模型相比，混合预测模型可以有效地克服单一方法存在的局限性，生成更加可靠和稳定的预测结果。

农田CO_2排放浓度变化具有明显的时序特征和非线性特点。LSTM凭借其独特的门控机制，能够捕捉长期依赖与周期性模式[18]。XGBoost作为基于决策树的集成算法，擅长自动挖掘复杂的特征交互，能有效处理高维数据和非线性关系[19]。基于LSTM和XGBoost在各自领域的优势，本研究设计了如图4所示的基于LSTM时序特征拼接的XGBoost组合模型(LSTM-XGBoost)。

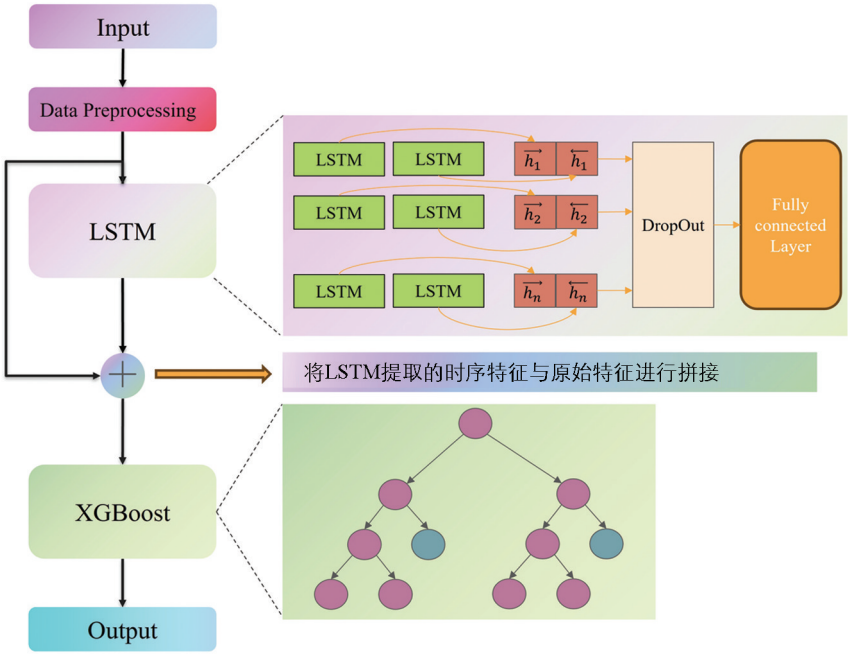


图4 LSTM-XGBoost组合模型结构图

组合模型的输入数据包括CO_2排放浓度、箱内温湿度、土壤温湿度、环境温湿度、土壤电导率、光照和风速。具体的构建步骤如下所示：

1) 使用LSTM模型对采集到的数据集进行训练，并使用随机搜索对LSTM进行超参数优化，以捕捉时间序列中的动态特征。LSTM会从数据集中提取各个数据项的时间特征，捕捉时间序列的周期性变化。时间序列数据如式(11)所示。

$$\boldsymbol{X}_{\text{orig}}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n],\ \boldsymbol{x}_i\in\mathbb{R}^M \tag{11}$$

式中：$\boldsymbol{X}_{\text{orig}}$ 代表原始特征矩阵；n 代表样本数；M 代表每个样本的特征维度；$\boldsymbol{x}_i$ 代表第 i 个样本的原始特征向量。$\boldsymbol{X}_{\text{orig}}$ 通过LSTM网络处理后，会输出每个样本在连续时间步上的隐藏层状态序列，即LSTM提取的时序特征。第 i 个样本的时序特征序列定义如式(12)所示：

$$\boldsymbol{H}_i=[\boldsymbol{h}_{i,1},\boldsymbol{h}_{i,2},\cdots,\boldsymbol{h}_{i,T}],\ \boldsymbol{h}_{i,t}\in\mathbb{R}^D \tag{12}$$

式中：$\boldsymbol{H}_i$ 代表LSTM提取的第 i 个样本的时序特征序列；T 代表时间序列的步数；D 代表LSTM隐藏层状态的维度；$\boldsymbol{h}_{i,t}$ 代表第 i 个样本在第 t 个时间步的隐藏层输出特征。

2）LSTM 训练完成后，将每个样本对应的时序特征与原始特征进行水平拼接。由于 XGBoost 要求输入为一维特征向量，需先将二维的时序特征序列 $\boldsymbol{H}_i$ 展平为一维向量，再拼接到对应原始特征向量 $\boldsymbol{x}_i$ 的末尾。第 i 个样本的拼接特征向量定义如公式(13)所示：

$$\boldsymbol{X}_{\text{concat},i}=[\boldsymbol{x}_i,\boldsymbol{h}_{i,1},\boldsymbol{h}_{i,2},\cdots,\boldsymbol{h}_{i,T}] \tag{13}$$

式中：$\boldsymbol{X}_{\text{concat},i}$ 代表第 i 个样本的拼接特征向量，由原始特征向量 $\boldsymbol{x}_i$ 与展平后的 LSTM 时序特征序列拼接得到。将所有样本的拼接特征向量按行组合，即可得到最终输入 XGBoost 的训练特征矩阵，其定义如式(14)所示：

$$\boldsymbol{X}_{\text{train\_concat}}=\begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{h}_{1,1} & \boldsymbol{h}_{1,2} & \cdots & \boldsymbol{h}_{1,T}\\ \boldsymbol{x}_2 & \boldsymbol{h}_{2,1} & \boldsymbol{h}_{2,2} & \cdots & \boldsymbol{h}_{2,T}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{x}_n & \boldsymbol{h}_{n,1} & \boldsymbol{h}_{n,2} & \cdots & \boldsymbol{h}_{n,T}\end{bmatrix} \tag{14}$$

式中：$\boldsymbol{X}_{\text{train\_concat}}$ 代表最终输入 XGBoost 的训练特征矩阵，维度为 $n\times(M+T\times D)$，$T\times D$ 代表单个样本的 LSTM 时序特征展平后的总维度。

3）将特征矩阵 $\boldsymbol{X}_{\text{train\_concat}}$ 输入到 XGBoost 模型，并采用贝叶斯优化算法对 XGBoost 的超参数进行自适应调整。训练完成后，XGBoost 将输出最终的 CO_2 排放浓度预测值，其预测过程可表示为：

$$\hat{y}_{CO_2}=\sum_{k=1}^{K}f_k(X_{\text{train\_concat}}) \tag{15}$$

式中：K 代表决策树数量，f_k 函数表示第 K 棵树对输入数据 $\boldsymbol{X}_{\text{train\_concat}}$ 的预测结果。为评估模型的性能，采用时间序列交叉验证，确保模型在不同时间段的数据上进行验证，有效避免数据泄漏。同时使用负均方误差作为评估指标，衡量模型的预测精度和稳定性，从而确保组合模型能提供准确的 CO_2 排放预测。

通过以上步骤的构建，LSTM 可以捕捉数据集中时间序列的时间依赖性，将隐藏层特征与原始特征拼接后传递给 XGBoost。XGBoost 负责从特征中挖掘非线性关系进行预测。

3　实验设计与结果分析

3.1　模型评价指标

本研究采用 MAE、$RMSE$、平均绝对百分比误差(Mean Absolute Percentage Error，$MAPE$)和 R^2 作为评价指标。其中，$RMSE$ 表示模型预测值和真实值的平方误差的平均值的平方根，其计算公式如式(16)所示。MAE 表示模型预测值和真实值的绝对误差的平均值，其计算公式如式(17)所示。$MAPE$ 用于评估模型的拟合效果，其计算公式如式(18)所示。R^2 的值通常介于 0 和 1 之间，越接近 1 则代表模型的拟合效果越好。R^2 可以用于评估模型的拟合能力，其中拟合效果好的模型的 R^2 值较大，其计算公式如式(19)所示[20]。通过 $RMSE$、MAE、$MAPE$ 和 R^2 可以衡量模型的预测性能、分析模型的误差来源、预测的准确度以及拟合的质量。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{16}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{17}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\% \tag{18}$$

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

式中：y_i 代表真实值；$\hat{y}_i$ 代表预测值；$\bar{y}$ 代表真实值的平均值；n 是样本数量。

3.2 GRU与LSTM模型对比分析

GRU与LSTM在未优化前均使用默认参数进行训练。对GRU与LSTM使用随机搜索后的超参数搜索空间和最终的最佳参数组合如表3所示。

表3 GRU与LSTM的超参数搜索范围及最优值

模型参数	参数范围	GRU	LSTM
第一层隐藏单元数	16～128	32	64
第二层隐藏单元数	8～64	32	32
丢弃率	0.1～0.4	0.1	0.3
学习率	0.001～0.01	0.005	0.005

优化前后的GRU和LSTM预测趋势如图5所示。从预测趋势分析来看，优化后的GRU和LSTM预测值比优化前更接近真实值。特别是在CO_2排放浓度高值和低值区域，LSTM预测曲线与真实值更加接近。而优化前的GRU和LSTM在高CO_2排放浓度时段存在一定程度的低估，在优化后得到了很大改善，但GRU的预测值比LSTM差。另外，在CO_2排放浓度变化较为剧烈的时刻，优化后的LSTM具有更快的响应速度，滞后性有所改善。实验分析结果表明，LSTM具有更强的时间序列建模能力。

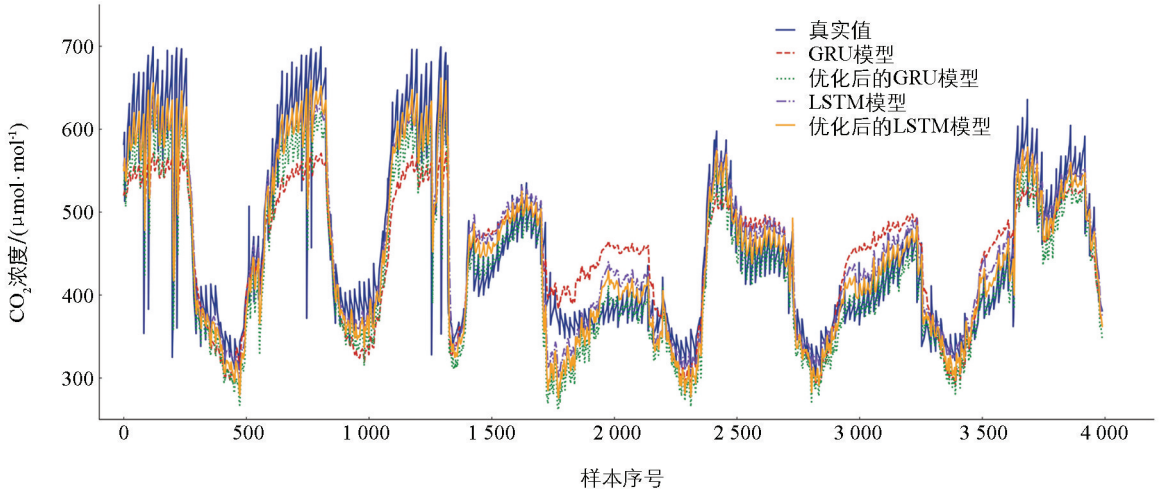


图5 GRU与LSTM预测趋势图

为了进一步分析GRU与LSTM模型的性能，计算出相关评价指标，结果如表4所示。

表4 GRU与LSTM评价指标

模型	*RMSE*	*MAE*	R^2	*MAPE*
GRU	55.64	45.46	0.71	9.83
优化后的GRU	42.21	32.81	0.82	7.16
LSTM	37.70	28.11	0.84	6.27
优化后的LSTM	35.93	26.51	0.86	5.97

从评价指标分析，优化后的LSTM和GRU相比优化前性能均有所提升。优化后的LSTM模型展现出更低的误差值和更高的拟合度，整体预测精度优于优化后的GRU模型。虽然GRU模型在优化后性能有所改善，但其RMSE和MAE明显高于LSTM，表明其预测误差较大。分析结果表明，在相同预测任务下，LSTM在各项指标上均优于GRU，LSTM具备更精准的预测能力、更强的建模能力以及更好的泛化性能。

综合上述误差分布与评价指标分析结果表明，在农田CO_2排放浓度预测任务中，LSTM相较于GRU具有更显著的优势。LSTM在整体趋势跟踪上表现更为稳定，具备更好的长期预测能力。本研究最终选取LSTM作为组合模型的组成部分，以提升预测模型的整体性能。

3.3　RF 与 XGBoost 模型对比与分析

RF 与 XGBoost 在优化前均使用默认参数进行训练。RF 与 XGBoost 模型的超参数搜索空间及最优参数结果如表 5 所示。

表 5　RF 与 XGBoost 超参数寻优范围与最优值

参数意义	寻优范围	XGBoost	RF
学习速率	0.01～0.10	0.03	—
最佳决策树的个数	100～300	152	120
最大回归树深度	1～10	6	8
最小叶子节点样本权重	1～10	8	8
每次分裂时使用的特征比例	0.1～1.0	—	0.8

RF 和 XGBoost 预测趋势如图 6 所示。从趋势图分析，优化前的 RF 和 XGBoost 均能较好地跟随 CO_2 排放浓度的整体变化趋势，但 XGBoost 在拟合实际值方面表现更佳，能更准确地捕捉浓度波动的细节变化。RF 模型在响应快速变化时表现较为迟缓，对 CO_2 排放浓度波动的反应较为平缓，预测精度受限。优化后的 XGBoost 在农田 CO_2 排放浓度预测中的表现明显优于优化后的 RF。XGBoost 不仅在整体趋势上保持良好拟合，而且在数据剧烈波动或复杂区域中依然能够保持较小的误差，预测结果稳定性与可靠性更高。虽然 RF 在优化后预测能力有所提升，但在高波动和快速变化的数据区间内仍存在较大的预测偏差，不能像 XGBoost 那样精确地跟随实际值的波动。

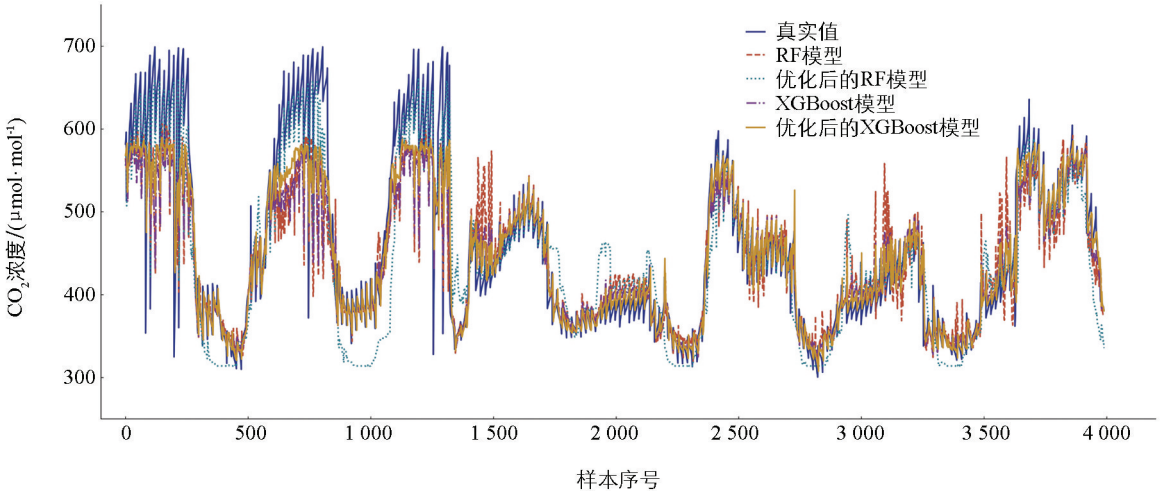


图 6　RF 与 XGBoost 预测趋势图

为了进一步分析 RF 与 XGBoost 模型的性能，计算出相关评价指标，结果如表 6 所示。

表 6　RF 与 XGBoost 评价指标

模型	*RMSE*	*MAE*	R^2	*MAPE*
RF	48.96	31.14	0.75	6.32
优化后的 RF	43.84	30.90	0.81	6.91
XGBoost	46.52	28.47	0.79	5.50
优化后的 XGBoost	39.52	23.33	0.84	4.62

性能分析结果表明，优化后的 XGBoost 和 RF 模型在各项指标上均优于优化前的表现。虽然 RF 优化后在一定程度上降低了 *RMSE*，但其 *MAE* 和 *MAPE* 仍相对较高，说明其误差分布较大，预测稳定性仍不能达到预期要求。相比之下，XGBoost 优化后的误差显著降低，并且拟合度提升，优化后的 XGBoost 整体预测精度优于 RF 优化模型。特别是在面对复杂的农田 CO_2 排放数据时，XGBoost 能够提供更低的误差和更高的预测准确性。

通过以上综合分析对比，在农田 CO_2 排放预测方面 XGBoost 明显优于 RF。XGBoost 在局部波动区域的预测趋势和拟合能力优于 RF，因此本研究选用 XGBoost 作为组合模型的一部分。

3.4 基于 LSTM 与 XGBoost 的加权组合模型分析

各加权组合模型的预测趋势如图 7 所示。从图 7 可知，不同的加权方法在 CO_2 排放预测中均有较高的拟合度，但在 CO_2 排放急剧变化时，所有加权方法的预测仍然存在偏差，误差幅度较小。最优加权组合模型和混合加权组合模型的预测值在大部分情况下能够较好地接近真实值，但在一些波动较大的区域仍存在一定误差。基于 R^2 和 *MAE* 的加权预测模型的预测结果波动较大，不能始终跟随真实值。基于 *RMSE* 的加权预测模型的预测值在某些区域能够较好地反映真实值，在上述几种加权方式中预测性能最佳。

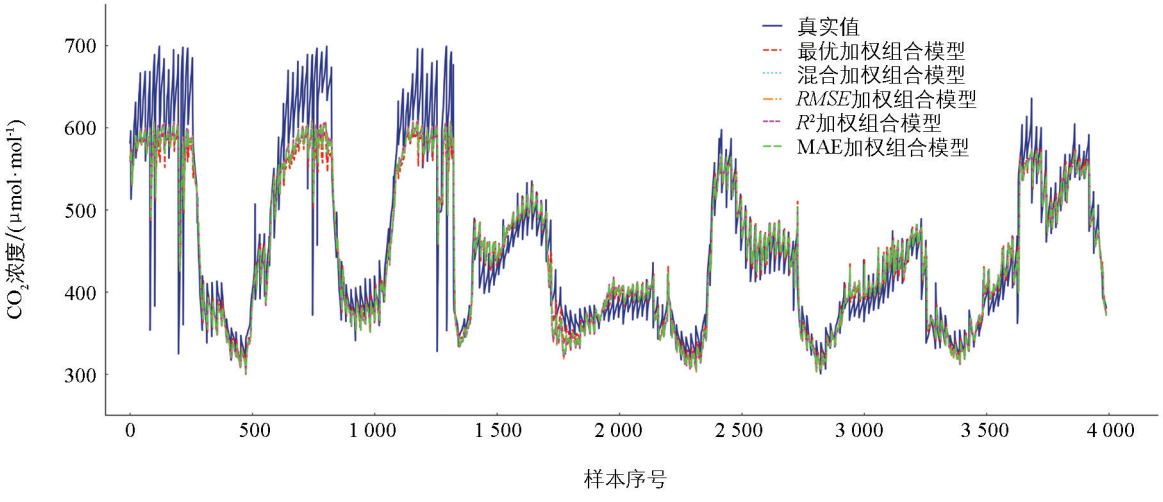


图 7 各加权组合模型预测趋势图

为了进一步分析不同加权组合模型的预测性能，我们计算了各项评价指标，结果如表 7 所示。

表 7 各加权组合模型评价指标

加权方式	权重		*RMSE*	*MAE*	R^2	*MAPE*
	LSTM	XGBoost				
RMSE	0.523	0.476	34.418	23.34	0.883	4.932
MAE	0.468	0.531	34.643	23.123	0.881	4.842
R^2	0.507	0.492	34.474	23.279	0.882	4.905
平均加权	0.499	0.500	34.504	23.247	0.882	4.892
网格搜索	0.310	0.689	35.716	22.686	0.874	4.683

从性能指标来看，加权组合模型的效果明显稳健。*RMSE* 加权组合、*MAE* 加权组合、R^2 加权组合等模型的 *RMSE* 和 *MAE* 普遍较低，表明加权组合模型具有更高的预测稳定性和准确性。在上述加权组合模型中，*RMSE* 加权组合模型表现最优，各项性能指标均表现较好。通过以上综合分析结果，本研究选取了最优的 *RMSE* 加权组合模型与 LSTM 时序特征拼接的 XGBoost 组合模型进行最后的对比分析。

3.5 LSTM 时序特征拼接的 XGBoost 组合模型分析

组合模型中 LSTM 与 XGBoost 的超参数优化后的最佳参数结果如表 8 所示。

表 8 组合模型最佳参数

LSTM		XGBoost	
模型参数	寻优结果	模型参数	寻优结果
第一层隐藏单元数	256	学习速率	0.005
第二层隐藏单元数	128	最佳决策树的个数	389
丢弃率	0.4	最大回归树深度	7
学习率	0.01	最小叶子节点样本权重	5

各模型预测趋势如图 8 所示。LSTM-XGBoost 组合模型结合了 LSTM 对时间序列特征的提取能力和 XGBoost 的非线性回归能力，使得预测曲线在短期和长期趋势上均能较好地拟合真实 CO_2 排放浓度。相比其他单一和常规加权组合模型，LSTM-XGBoost 组合模型在数据波动剧烈的区域预测误差更小，能够更精准地响应 CO_2 排放浓度的突变趋势，组合模型能更好地适应 CO_2 排放浓度的复杂变化，使得预测结果更加稳定。不同于单一模型和加权组合模型在短期剧烈波动和长期趋势变化时不能精准预测的不足，LSTM-XGBoost 组合模型在不同时间段内提供更精准的 CO_2 排放浓度预测。

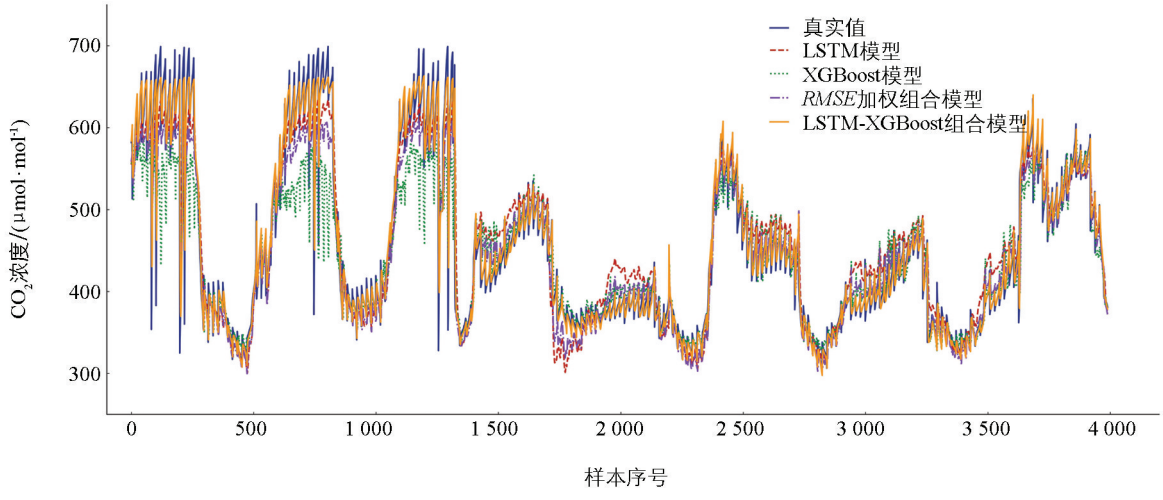


图 8　各模型预测趋势图

各模型误差趋势如图 9 所示。LSTM-XGBoost 组合模型的误差在大部分区间内保持最低水平并且波动较小，表明该模型能够更有效地结合 LSTM 的时序特征提取能力与 XGBoost 的非线性建模能力，在农田 CO_2 排放浓度的预测任务中表现最佳。从综合分析结果来看，LSTM-XGBoost 组合模型不仅降低了预测误差，还提高了预测的稳定性和可靠性，使其在 CO_2 排放浓度预测任务中更具优势。

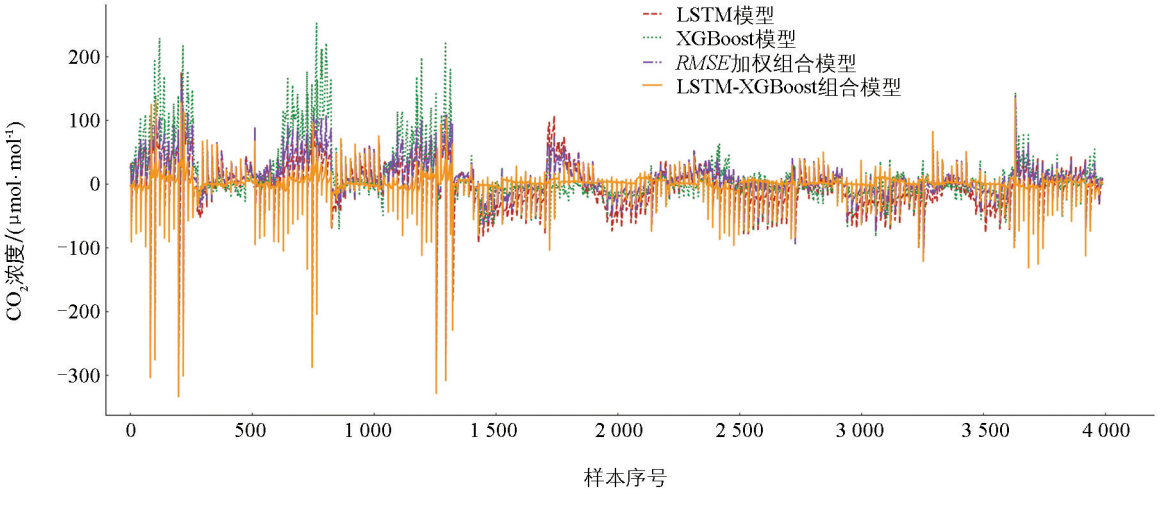


图 9　各模型误差趋势图

使用模型评价指标进一步分析各个模型的预测性能，各模型评价指标如表 9 所示。

根据性能评价指标分析，LSTM-XGBoost 组合模型在所有指标上均表现最佳，显著优于单独的 LSTM 和 XGBoost 模型以及 LSTM 和 XGBoost 的 *RMSE* 加权组合模型，展现出更高的预测精度和稳定性。LSTM-XGBoost 组合模型在各项性能指标上均优于单一模型与加权组合模型。实验结果显示 LSTM-XGBoost 的 *RMSE* 降至 22.81，*MAE* 为 9.97，R^2 提高至 0.94，*MAPE* 仅为 2.21，显著优于其他模型，表明 LSTM-XGBoost 能够更精确地预测 CO_2 排放的动态变化。

通过分析预测趋势、误差趋势以及多项评价指标，LSTM-XGBoost 组合模型在农田 CO_2 排放浓度的

短期波动和长期趋势预测方面均表现优异，特别是在面对剧烈变化区域时依然能够迅速捕捉异常变化并做出响应。与单一模型和传统集成方法相比，组合模型在降低预测误差和增强泛化能力方面具有明显优势，为农田 CO_2 排放浓度预测提供了一种更加高效和精准的解决方案。

表9 各模型评价指标

模型	*RMSE*	*MAE*	R^2	*MAPE*
LSTM	35.93	26.51	0.86	5.97
GRU	42.21	32.81	0.82	7.16
XGBoost	39.52	23.33	0.84	4.62
RF	43.84	30.90	0.81	6.91
RMSE 加权组合模型	34.41	23.34	0.88	4.93
LSTM-XGBoost 组合模型	22.81	9.97	0.94	2.21

4 讨论与结论

4.1 讨论

针对农田 CO_2 排放浓度预测中的建模问题，本文在对比现有单一模型性能的基础上，分析了基于LSTM时序特征拼接的XGBoost组合模型的有效性。文献[21]探讨了多种单个模型在农田 CO_2 排放预测中的研究，研究表明LSTM在所有单个预测模型中的表现最佳($R^2=0.87$)，稍高于本研究构建的LSTM预测模型($R^2=0.86$)，但低于本研究提出的基于LSTM-XGBoost组合模型的 $R^2=0.94$。文献[22]使用模糊基函数回归(Fuzzy Basis Function Regression，FBFR)、支持向量回归(Support Vector Regression，SVR)、CNN及前馈神经网络(Feedforward Neural Network，FNN)对玉米农田 CO_2 排放进行了预测，预测结果对比分析显示FNN的预测性能最佳，FNN预测性能的 $R^2=0.918$ 低于本研究提出的LSTM-XGBoost组合模型的 $R^2=0.94$，$RMSE=67.75$ 高于本研究LSTM-XGBoost组合模型的22.81。文献[23]利用反向传播神经网络(Back Propagation Neural Network，BP)对农田土壤 CO_2 排放进行预测，预测性能的 $R^2=0.918\,8$，低于本研究的LSTM-XGBoost组合模型。通过对比分析，本研究的LSTM-XGBoost组合模型预测精度高、误差小，为农田 CO_2 排放浓度的预测提供了一种新的思路和方法。

4.2 结论

本研究构建了LSTM、GRU、XGBoost和RF 4种单一机器学习模型，并分别对模型进行了随机搜索和贝叶斯优化。实验结果表明，LSTM在长期趋势捕捉方面优于GRU，XGBoost在短期非线性拟合方面优于RF。本研究通过实验验证了不同模型的差异化优势，同时也发现单一模型在面对农田 CO_2 排放浓度数据中的复杂时序与非线性交互特征时，仍然存在一定的预测局限性。针对单一模型的局限性，本研究提出并建立了基于LSTM时序特征拼接的XGBoost混合预测模型，该模型充分融合了LSTM在长期趋势建模与XGBoost在短期非线性特征拟合方面的优势。实验证明该组合模型在预测精度和稳定性方面表现卓越，R^2 提高至0.94，*RMSE* 降低至22.81，*MAPE* 仅为2.21，相较于传统单一模型及常规加权融合模型有显著提高，为农田 CO_2 排放浓度的精准预测提供了新的有效方法。

参考文献：

[1] 马晨光，蔡焕杰，卢亚军. 基于APSIM模型不同水氮处理下 N_2O 的排放研究 [J]. 灌溉排水学报，2020，39(11)：120-129.

[2] 吴梦琴，李成芳，盛锋，等. 基于DNDC模型评估湖北省不同稻作系统不同管理措施温室气体排放的周年变化 [J]. 中国生态农业学报(中英文)，2021，29(9)：1480-1492.

[3] 燕振刚，李薇，Yan Tianhai，等. BP神经网络算法在河西绿洲玉米生产碳排放评估中的应用及算法有效性研究 [J]. 中国生态农业学报(中英文)，2018，26(8)：1100-1106.

[4] BRISCOE N J，MORRIS S D，MATHEWSON P D，et al. Mechanistic Forecasts of Species Responses to Climate

Change: The Promise of Biophysical Ecology [J]. Global Change Biology, 2023, 29(6): 1451-1470.

[5] 吴成秋，曹召丹，赵小二，等. 基于水文气象因子的农田生态系统碳通量预测 [J]. 湖北农业科学，2024，63(8)：267-280.

[6] JOBARTEH B, NEETHIRAJAN S. Leveraging Satellite Data for Greenhouse Gas Mitigation in Canadian Poultry Farming [J]. Smart Agricultural Technology, 2025, 10: 100704.

[7] SEO J Y, LEE S I. CO_2 Emissions Associated with Groundwater Storage Depletion in South Korea: Estimation and Vulnerability Assessment Using Satellite Data and Data-Driven Models [J]. Remote Sensing, 2024, 16(17): 3122-3144.

[8] 张文栋，刘子琨，梁涛，等. 基于 CNN-LSTM 的综合能源系统负荷预测模型 [J]. 重庆邮电大学学报(自然科学版)，2023，35(2)：254-262.

[9] 宋育苗，于金霞. 基于 CNN-GRU 的移动 APP 流行度预测模型 [J]. 重庆邮电大学学报(自然科学版)，2024，36(4)：747-755.

[10] JI X, WANG J C, YAN Z J. A Stock Price Prediction Method Based on Deep Learning Technology [J]. International Journal of Crowd Science, 2021, 5(1): 55-72.

[11] 杨寒雨，赵晓永，王磊. 数据归一化方法综述 [J]. 计算机工程与应用，2023，59(3)：13-22.

[12] 王兴浩. 基于贝叶斯优化的图神经网络架构搜索方法 [D]. 长春：吉林大学，2022.

[13] 任建吉，位慧慧，邹卓霖，等. 基于 CNN-BiLSTM-Attention 的超短期电力负荷预测 [J]. 电力系统保护与控制，2022，50(8)：108-116.

[14] 赖晓莹，钱俊. ARIMA-LSTM-XGBoost 加权组合模型在肺结核发病趋势预测的研究 [J]. 现代预防医学，2021，48(1)：5-9.

[15] 代业明，周琼. 基于改进 Bi-LSTM 和 XGBoost 的电力负荷组合预测方法 [J]. 上海理工大学学报，2022，44(2)：138-147.

[16] 赵阳，范文奕，安佳坤，等. 基于智能加权混合模型的新型电力系统电量预测方法 [J]. 电测与仪表，2022，59(12)：56-63.

[17] 侯慧，吴文杰，魏瑞增，等. 基于注意力机制的 CNN-LSTM-XGBoost 台风暴雨电力气象混合预测模型 [J]. 智慧电力，2024，52(10)：96-102.

[18] 赵宏，王乐，王伟杰. 基于 BiLSTM-CNN 串行混合模型的文本情感分析 [J]. 计算机应用，2020，40(1)：16-22.

[19] 王晓玲，王成，王佳俊，等. 大坝渗压混合预测的 STL 分解-集成学习模型 [J]. 水力发电学报，2024，43(9)：106-123.

[20] 赵明珠，王丹，方杰，等. 基于 LSTM 神经网络的地铁车站温度预测 [J]. 北京交通大学学报，2020，44(4)：94-101.

[21] HAMRANI A, AKBARZADEH A, MADRAMOOTOO C A. Machine Learning for Predicting Greenhouse Gas Emissions from Agricultural Soils [J]. Science of the Total Environment, 2020, 741: 140338.

[22] HARSÁNYI E, MIRZAEI M, ARSHAD S, et al. Assessment of Advanced Machine and Deep Learning Approaches for Predicting CO_2 Emissions from Agricultural Lands: Insights across Diverse Agroclimatic Zones [J]. Earth Systems and Environment, 2024, 8(4): 1109-1125.

[23] FREITAS L P S, LOPES M L M, CARVALHO L B, et al. Forecasting the Spatiotemporal Variability of Soil CO_2 Emissions in Sugarcane Areas in Southeastern Brazil Using Artificial Neural Networks [J]. Environmental Monitoring and Assessment, 2018, 190(12): 741.

责任编辑　张枸