基于 HHO-CNN-LSTM 的 CMAQ 修正模型及其在上海市空气质量预报中的应用

郑鑫楠[1], 林开颜[2], 王孜竞[1], 宋远博[1], 师　洋[1], 路函悦[1], 张亚雷[2,3], 沈　峥[2,*]

(1. 同济大学 电子与信息工程学院, 上海 201804; 2. 同济大学 新农村发展研究院, 上海 201804;
3. 同济大学 环境科学与工程学院, 上海 200092)

摘要:建立空气质量预报模型,预测污染物浓度对人类健康和社会经济发展具有重要意义。然而,传统的空气质量模型 CMAQ 对污染物浓度的预报精度并不理想。对此,本文提出了一种基于卷积神经网络(CNN)和长短期记忆神经网络(LSTM)的空气质量预报修正模型,并使用哈里斯鹰算法(HHO)对模型的超参数进行优化;用 CMAQ 模型对上海市 2022 年 12 月六种大气污染物(SO_2、NO_2、PM_{10}、$PM_{2.5}$、O_3、CO)浓度的预报数据以及监测站的气象数据和污染物浓度实测数据作为 HHO-CNN-LSTM 模型的输入,对 CMAQ 模型预报结果进行修正。使用均方根误差(*RMSE*)、平均绝对误差(*MAE*)和一致性指数(*IOA*)作为评价指标。结果显示,修正模型显著提高了六项污染物浓度的预测精度,*RMSE* 减少了 73.11%~91.31%,*MAE* 减少了 67.19%~89.25%,*IOA* 提升了 35.34%~108.29%。同时针对 HHO 算法陷入局部最优而导致修正模型对 CO 浓度预测效果不佳的问题,使用高斯随机游走策略对 HHO 算法进行改进,显著提高了 CO 浓度的预测精度。相比于改进之前,*RMSE* 减少了 39.55%,*MAE* 减少了 45.93%,*IOA* 提高了 32.43%。

关键词:空气质量预报;CMAQ 模型;卷积神经网络;长短期记忆神经网络;哈里斯鹰优化算法

中图分类号:X823　　**文献标识码**:A

Application of HHO-CNN-LSTM-based CMAQ correction model in air quality forecasting in Shanghai

ZHENG Xinnan[1], LIN Kaiyan[2], WANG Zijing[1], SONG Yuanbo[1], SHI Yang[1], LU Hanyue[1], ZHANG Yalei[2,3], SHEN Zheng[2,*]

(1. *College of Electronics and Information Engineering*, *Tongji University*, *Shanghai* 201804, *China*;
2. *Institute of New Rural Development*, *Tongji University*, *Shanghai* 201804, *China*;
3. *College of Environmental Science and Engineering*, *Tongji University*, *Shanghai* 200092, *China*)

Abstract: With rising levels of air-pollution, air-quality forecasting has become integral to the dissemination of human health advisories and the preparation of mitigation strategies. Traditional air quality models, such as the Community Multi-scale Air Quality (CMAQ) model, have unsatisfactory accuracy. Accordingly, a correction model, which combines convolutional neural network (CNN) and long-short-term memory neural network (LSTM) and optimized by harris hawks optimization algorithm (HHO) was established to enhance the accuracy of CMAQ model's prediction results for six air pollutants (SO_2, NO_2, PM_{10}, $PM_{2.5}$, O_3 and CO). The accuracy of HHO-CNN-LSTM was evaluated using root mean square error (*RMSE*), mean absolute error (*MAE*), and the index of agreement (*IOA*). The results demonstrated a significant improvement in the accuracy of prediction for the six pollutants using the correction model. *RMSE* decreased by 73.11% to 91.31 %, *MAE* decreased by 67.19% to

收稿日期:2023-10-17　　**DOI**:10.20078/j.eep.20231107

基金项目:国家重点研发计划政府间国际合作资助项目(2022YFE0120600);国家自然科学基金面上资助项目(21978224)

作者简介:郑鑫楠(2000—),男,浙江绍兴人,硕士研究生在读,主要研究方向为智慧环境工程。E-mail: zhengxinnan@tongji.edu.cn

通讯作者:沈　峥(1978—),男,浙江宁波人,研究员,主要研究方向为废弃物资源化利用。E-mail: shenzheng@tongji.edu.cn

89.25%, and *IOA* increased by 35.34% to 108.29%. To address the propensity of the HHO algorithm to converge on local optima, leading to poor CO correction performance, this study proposed a method for the HHO algorithm with a Gaussian random walk strategy to improve the CO concentration correction performance.

Keywords: Air quality prediction; CMAQ; Convolutional neural network (CNN); Long-short-term memory neural network (LSTM); Harris hawks optimization algorithm (HHO)

0 引　　言

大气环境污染物一般是由二氧化硫(SO_2)、氮氧化物(NO_x)、臭氧(O_3)、一氧化碳(CO)等工业生产废物,以及 PM_{10}、$PM_{2.5}$等固体粒子组成。这种污染物会引发肿瘤等各种病症,严重危害人们健康。随着中国经济社会发展和人民生活水平的提升,大气环境污染已成为我国目前存在的主要大气环境问题之一。因此,建立空气质量模型预测污染物的浓度对人类健康和环境管理具有重要意义。

目前传统的空气质量模型以区域多尺度空气质量模型(CMAQ)为代表,用数学方程模拟污染物传播时的物理化学反应机制,充分考虑了实际环境中污染物相互之间的变化与影响,因此得到广泛的应用。例如,ZHENG 等[1]用非均相化学更新的 CMAQ 模型研究中国北方次生无机气溶胶的形成;ZHE 等[2]使用 CMAQ 模型分析 2013 年严重雾霾期间河北源部地区和其他地区 $PM_{2.5}$的排放量;HU 等[3]使用 WRF-CMAQ 模型,对 2013 年中国的臭氧和颗粒物进行了模拟;NAPELENOK 等[4]使用 CMAQ-ISAM 模型研究十余种生物质燃烧对美国东南部 $PM_{2.5}$浓度的影响;KOO 等[5]使用 WRF-CMAQ 模型预测了韩国首尔地区的 PM_{10}污染事件发生的时间和污染物的传输路径;WANG 等[6]利用 WRF-CMAQ 模型模拟了香港地区 O_3在不同海拔地区的分布浓度以及其形成、扩散的物理化学过程。

CMAQ 模型的预报需要将污染源排放清单作为数据输入,由于污染源种类繁多、分布面广和变化复杂,所以排放清单的编制工作需要较长的时间以及较多的人力,其制作过程决定了排放清单无法满足实时更新的要求;另外,CAMQ 模型是基于"一个大气"的核心概念建立的,而人们对于大气这个异常复杂的系统的了解十分有限,无法对其中所有的大气传输、污染物扩散和化学反应等过程进行量化处理;污染源位置和高度、大气稳定度以及人口、燃料构成等对大气质量的影响和作用往往是非线性的[7],在应用偏微分方程来描述这些非线性作用时,又使用了大量的近似方法来简化求解过程。这些都是 CMAQ 模型的预测结果存在偏差的主要原因。为提高 CMAQ 模型预报能力,利用监测数据对模型预报结果进行统计修正的方法应用也较为普遍。谢敏等[8]尝试将监测数据直接作为预报初始值,结合 CMAQ 模型预报的增减量建立修正方法;王茜等[9]利用线性回归方法建立预测数据与监测数据之间的关系,降低了由于污染源不确定性产生的预报偏差;芦华等[10]使用多元线性回归方法对 CMAQ 模型的预报结果进行滚动订正,有效提高了模型的预报效果。SAYEED 等[11]利用深度卷积神经网络(DCNN)对 CMAQ 模型进行修正和扩展,提高了模型在颗粒物浓度预测上的准确性。

近年来,由于人工智能的应用,不少深度学习算法也逐渐发展,如深度信念网络(DBN)、卷积神经网络(CNN)和循环神经网络(RNN)等。相比于传统的统计方法,深度学习算法能够处理更多非线性、非结构化的数据,具有更好的性能。一些研究人员已将其应用于空气质量研究,YI 等[12]提出了一种基于深度神经网络的 $PM_{2.5}$浓度预测模型,使用卷积神经网络和循环神经网络进行特征提取和序列建模,并引入了注意力机制和残差连接以增强模型的表达能力;XAYASOUK 等[13]提出了一种基于深度自编码器(DAE)和长短期记忆网络(LSTM)的空气污染物浓度预测方法,使用 DAE 对输入特征进行降维和特征提取,然后利用 LSTM 对时间序列数据进行预测;PAK 等[14]提出了一种基于卷积神经网络和长短期记忆神经网络的混合模型(CNN-LSTM)用于预测臭氧浓度,并证实具有良好的精度。LI 等[15]使用 CNN-LSTM 模型预测北京未来 24 小时 $PM_{2.5}$浓度,并通过比较得出 CNN-LSTM 模型具有误差小、训练时间短的优点。DU 等[16]提出了由多个一维卷积神经网络和一个双向长短期记忆神经网络组成的混合 CNN-

BiLSTM 模型,多个一维卷积神经网络用于提取多个监测站的空间相关性特征,双向长短期记忆神经网络可以学习时间序列数据过去和未来的特征,从而进行更有效的预测。

上述研究表明,CNN-LSTM 模型在大气污染物浓度预测方面具有较好的性能。在此基础上,利用哈里斯鹰优化算法(HHO)寻找 CNN-LSTM 模型的最优超参数,可以使模型拥有更好的预测效果。本文将会在 CMAQ 模型对上海市污染物浓度进行预测的基础上,使用深度学习方法构建基于 HHO-CNN-LSTM 的修正模型。将 CMAQ 模型的预报数据以及影响污染物浓度的气象数据和污染物浓度实测数据作为 HHO-CNN-LSTM 模型的输入,进行污染物浓度再预测,从而实现对 CMAQ 模型预报结果的修正。

1 方　　法

1.1 卷积神经网络

卷积神经网络(CNN)是一种包含卷积结构的深度前馈网络,由于其强大的特征提取能力,卷积神经网络已被广泛用于时间序列数据分析[17]。卷积神经网络可以提取空间结构中多维时间序列数据之间的关系,它由输入层、卷积层、池化层、全连接层和输出层组成。其中,卷积层的特征提取主要是通过卷积核进行的,它可以捕捉污染物数据中存在的时间依赖性[18];池化层主要用于特征降维,减少参数的数量,防止过拟合。经过卷积层和池化层作用后的特征进入全连接层后进行再整合,最终转化成一维向量。在本研究中,可以将模型的输入数据样本看作一个二维矩阵,其中横轴表示时间维度,纵轴表示特征维度。卷积核在时间维度上进行滑动,对每个时间点附近的特征进行卷积操作。通过多层不同大小的卷积核的叠加,卷积神经网络可以不断提取时间维度上的更高级别特征,从而获得更好的预测效果。

1.2 长短期记忆神经网络

长短期记忆神经网络(LSTM)是一种改进的循环神经网络。通过引入门结构(Gate),用门结构决定序列上信息的去留,记住需要长时间记忆的信息,过滤不重要的信息,解决了循环神经网络的长期依赖问题[19]。它被提出后也进行了改良,增加了额外的遗忘门。改良后的长短期记忆神经网络解决了模型训练中“梯度消失”的问题,可以学习时间序列长短期依赖信息,是目前最成功的循环神经网络架构,应用于许多场景中。在本研究中,大气污染物浓度数据和气象数据属于时间序列数据,当前时刻的状态通常与过去时刻的状态有关。通过长短期记忆神经网络的“遗忘门”“输入门”和“输出门”等机制,学习并记忆过往时刻的状态信息,可以有效地对时间序列数据进行预测。

1.3 哈里斯鹰优化算法

神经网络模型包含许多超参数,如神经网络层数、学习率、神经元数量等,选取最优的超参数能显著提高模型的精度和拟合度。传统的超参数选取往往依赖于研究者的个人经验或者每个超参数组合的效果[20],这种做法需要耗费大量的时间。优化算法的应用可以减少超参数搜索的时间,增强模型的预测效果[21]。近年来,基于种群的元启发式算法——群智能优化算法开始应用于神经网络的超参数优化[22-24]。

哈里斯鹰优化算法(HHO)是 Heidari 在 2019 年提出的一种群智能优化算法,具有参数少、搜索精度高和简单易行的优点[25]。该算法由哈里斯鹰对猎物的追捕行为演化而来,其具体流程如图 1 所示。根据猎物能量 E 和捕获概率 r 的变化,哈里斯鹰会执行不同的追逐策略。其中,哈里斯鹰为候选解,猎物为最优解,哈里斯鹰捕捉猎物的过程即为候选解向最优解迭代的过程。

1.4 基于 HHO 优化的 CNN-LSTM 模型

由于卷积神经网络具有较好的特征提取能力,长短期记忆神经网络在处理时间序列问题上有较大的优势,同时也能避免梯度消失的问题,因此本研究选择将卷积神经网络与长短期记忆神经网络相结合构建模型,具体结构如图 2 所示。模型的前半部分是卷积神经网络,用于特征提取,提取的信息经过最大池化层(Max-Pool)和 Dropout 层处理后,可有效防止其过拟合;模型的后半部分是长短期记忆神经网络,用于时间序列数据的预测,LSTM 层的输出结果经过全连接层(FC)的展平操作后,最终变为一维的预测数据进行输出。

CNN-LSTM 混合神经网络有卷积层卷积核大小、卷积核数量、LSTM 层神经元数量、批次大小等超参数,这些超参数的选取会显著影响模型的性能。因此本文使用哈里斯鹰优化算法对 CNN-LSTM 模型进行优化,寻找到最优的超参数,提高模型的预测精度。

哈里斯鹰算法优化 CNN-LSTM 模型的具体步骤如图 3 所示。每个哈里斯鹰个体代表一组超

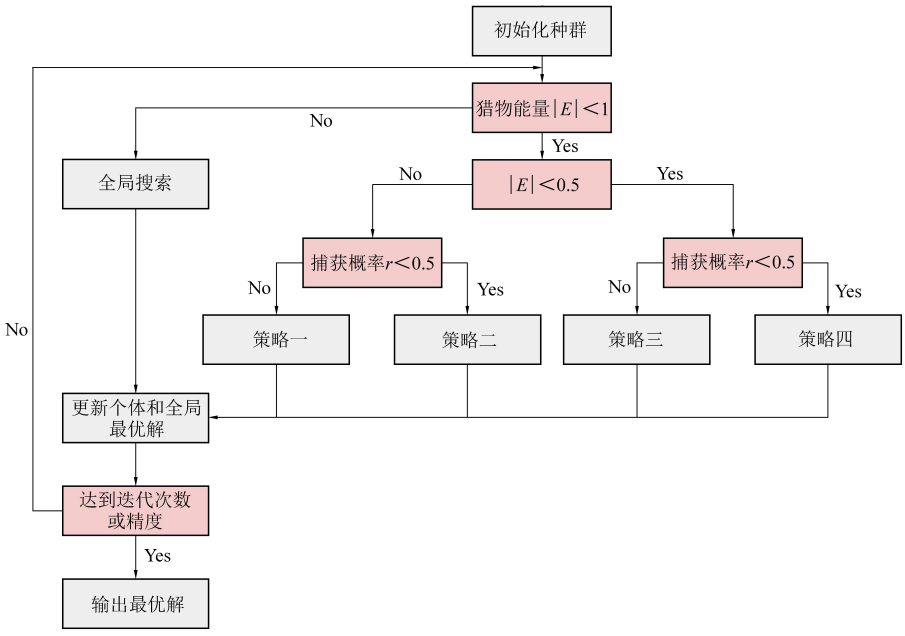

图 1　哈里斯鹰优化算法流程示意图

Fig. 1　Flowchart of Harris Hawks Optimization algorithm

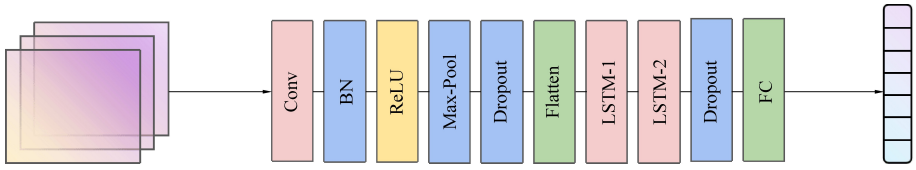

图 2　CNN-LSTM 网络结构示意图

Fig. 2　Schematic diagram of CNN-LSTM network structure

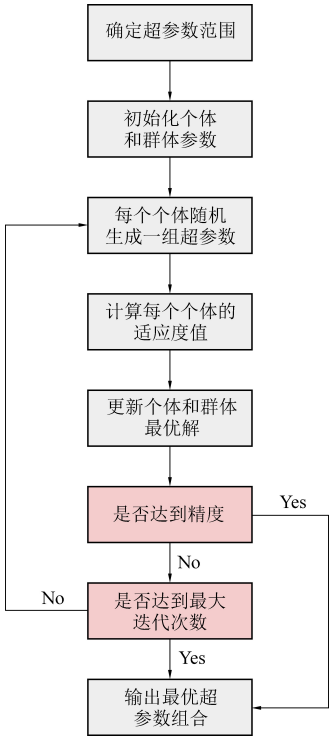

图 3　HHO 优化 CNN-LSTM 模型流程示意图

Fig. 3　Flowchart of HHO-optimized CNN-LSTM model

参数,通过计算适应度值对个体进行排序,选取表现最好的一部分个体,作为新一轮迭代的种群,重复迭代直到达到最大迭代次数或找到满意的超参数组合为止。

2　实　　验

2.1　实验数据

本文研究使用 2022 年 12 月 1 日至 2022 年 12 月 31 日上海市徐汇区上海师范大学空气质量监测站的大气污染物浓度数据和徐家汇气象站的气象数据。大气污染物数据来自中国环境监测总站,包括二氧化硫(SO_2)、二氧化氮(NO_2)、可吸入颗粒物(PM_{10})、细颗粒物($PM_{2.5}$)、一氧化碳(CO)和臭氧(O_3)的逐小时监测数据。我们将其作为实测数据,后续用于修正模型的训练和比较。气象数据来自国家气象信息中心,包括温度、湿度、气压、风速和风向的逐小时监测数据,后续作为影响污染物浓度的气象因子用于修正模型的训练。

2.2　基于 CMAQ 模型的污染物浓度预测

CMAQ 模型是美国环保署(EPA)开发的第三代空气质量预报模型。通过输入气象数据和排放

源清单,CMAQ 模型使用数学算法和反应动力学模型对大气中各种污染物的传输、化学反应、扩散等过程进行建模和模拟,从而预测不同时间和空间范围内污染物的浓度分布情况。化学传输模块是 CMAQ 模型的核心,包括扩散模块、平流模块、气象化学模块、气溶胶模块等,用于模拟和预测污染物的化学反应、输送和扩散过程。

本研究使用 CMAQ 模型对上海市 2022 年 12 月 1 日至 2022 年 12 月 30 日的空气污染物浓度进行逐时预报。空气质量预报模型模拟区域采用 Lambert 投影坐标系,坐标中心点为 31°N、121°E,设置两层嵌套网格,第一层网格水平分辨率为 27 公里,网格数为 100×100;第二层网格水平分辨率为 9 公里,网格数为 103×103。CMAQ 模型自 2022 年 12 月 1 日开始,每日 0 时起报,预报未来 72 小时的污染物浓度。将相同时间点的预报数据进行均值化处理,得到了上海市 2022 年 12 月 1 日至 2022 年 12 月 30 日的空气污染物(SO_2、NO_2、PM_{10}、$PM_{2.5}$、O_3、CO)浓度逐小时预报数据。将其作为 CMAQ 预报数据后续用于修正模型的训练和比较。

2.3 修正模型的训练

本研究使用哈里斯鹰算法优化的 CNN-LSTM 模型作为污染物浓度的修正模型,对 CMAQ 模型的预报结果进行修正。考虑到污染物浓度与气象条件紧密相关,同时污染物之间存在复杂的化学反应[26],修正模型的输入特征包括气象因子(温度、湿度、气压、风速、风向)和除自身外其他 5 项污染物浓度的实测数据以及该项污染物的 CMAQ 预报数据,共计 11 个特征,输出数据为该项污染物浓度的实测数据。实验数据的时间范围为 2022 年 12 月 1 日 0 时至 2022 年 12 月 30 日 23 时,时间步长为 1 小时,共计 720 条数据。对实验数据进行划分,设置训练集、验证集和测试集的比例为 7∶2∶1 并进行归一化处理。

将处理好的数据输入模型后,开始使用哈里斯鹰算法对 CNN-LSTM 的超参数进行寻优迭代。需要优化的超参数包括卷积核大小、卷积核数量、批次大小、第一层 LSTM 神经元个数、第二层 LSTM 神经元个数、最大迭代数和学习率。确定每个超参数的寻优范围,通过 HHO 迭代找到最优的超参数。将最优的超参数组合应用于 CNN-LSTM 模型,当模型完成训练之后便可得到新的污染物浓度预测值,从而实现对 CMAQ 模型预报结果的修正。

3 结果与讨论

3.1 污染物浓度修正结果

基于 HHO 优化的 CNN-LSTM 大气污染物浓度修正模型对 CMAQ 预测数据的修正结果如图 4 所示。选择均方根误差(*RMSE*)、平均绝对误差(*MAE*)和一致性指数(*IOA*)作为评价指标来评价模型的预测效果。均方根误差和平均绝对误差反映预测值与实测值的数值偏差,一致性指数反映预测值与实测值的一致性。三个评价指标的计算公式如下:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - Q_i)^2}, i = 1,2,\cdots,N \tag{1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - Q_i|, i = 1,2,\cdots,N \tag{2}$$

$$IOA = 1 - \frac{\sum_{i=1}^{N}(P_i - Q_i)^2}{\sum_{i=1}^{N}(|Q_i - \overline{Q}| + |P_i - \overline{Q}|)^2}, \quad i = 1,2,\cdots,N \tag{3}$$

式(1~3)中 P_i 为污染物浓度的预测值,Q_i 为污染物浓度的实际值,$\overline{Q}$ 为污染物浓度实际值的平均值。比较修正前后的评价指标,结果见表 1。

由图 4 可以明显看出经过深度学习算法修正后的 CMAQ 预报数据(CMAQ-DL)相比修正前(CMAQ)更贴近实测值。根据表 1 可知,修正后的模型预测结果在均方根误差、平均绝对误差,和一致性指数三项评价指标上均表现得更加优异。六项污染物浓度的预测误差都大幅降低,*RMSE* 减少了 73.11%~91.31%,*MAE* 减少了 67.19%~89.25%。各项污染物浓度的预测值与实际值的一致性也都有显著提升,*IOA* 提升了 35.34%~108.29%。相比之前学者采用的线性回归方法(*IOA* 从 0.564 提升至 0.721)[9],HHO-CNN-LSTM 模型对 CMAQ 预报结果的修正效果更好。这是因为本研究考虑了更多维度的影响因素,且神经网络方法在处理高维度和非线性数据上具有较大的优势。

其中 CO 的 *IOA* 虽有很大提升,但相比于其他污染物,CO 的 *IOA* 仍然较低,一方面可能是因为 CMAQ 模型对 CO 的预测精度较低,从而影响了神经网络的训练;另一方面可能是因为 HHO-CNN-LSTM 模型性能上的问题,接下来将对修正模型的性能进行检验。

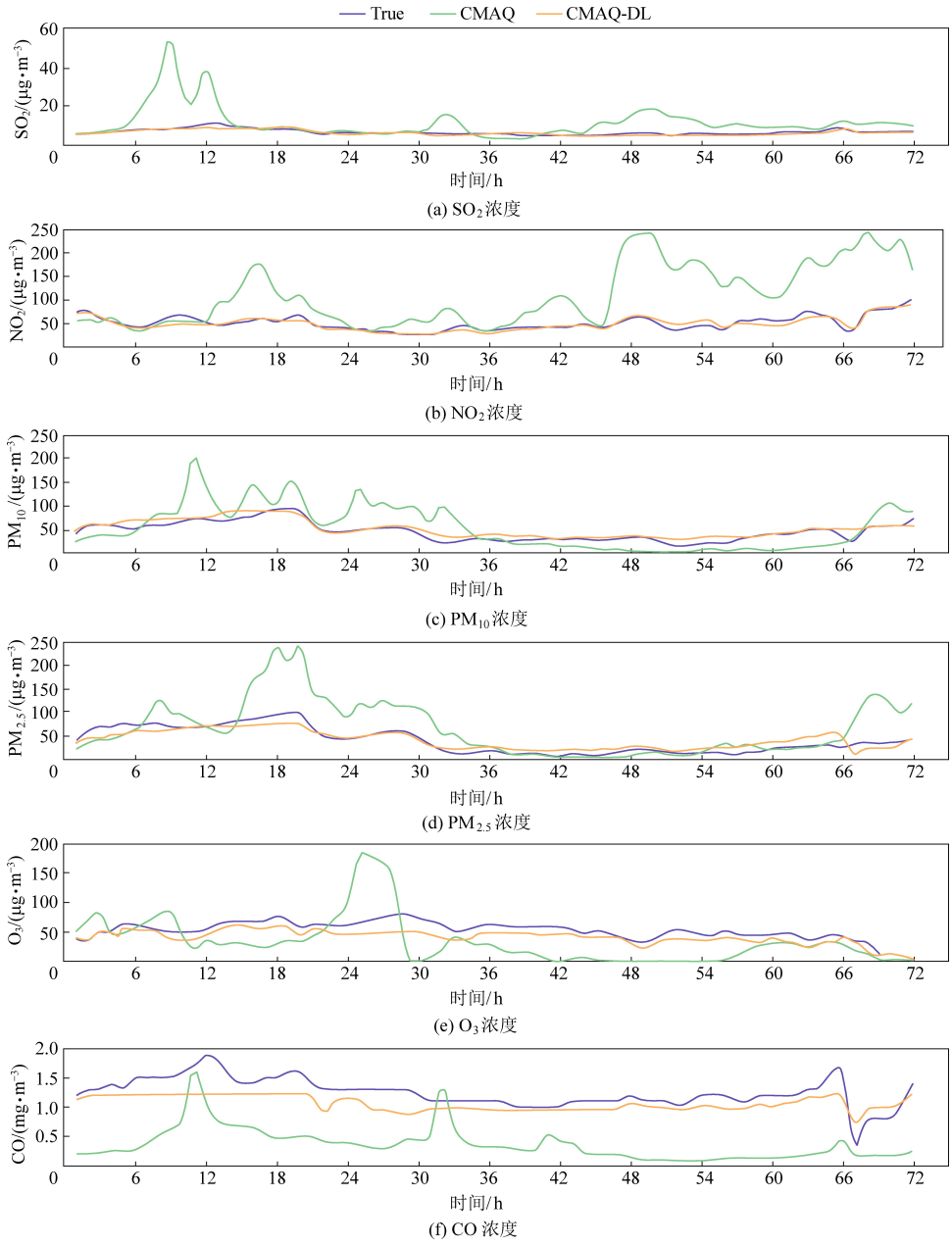

图 4　污染物浓度修正结果

Fig. 4　Correction results of pollutant concentration

表 1　修正前后的评价指标结果

Table 1　Evaluation index results before and after correction

评价指标	SO_2/(μg·m^{-3})		NO_2/(μg·m^{-3})		PM_{10}/(μg·m^{-3})	
	CAMQ	CMAQ-DL	CAMQ	CMAQ-DL	CAMQ	CMAQ-DL
RMSE	9.331 9	0.810 9	82.801 1	7.966 7	37.128 6	9.686 9
MAE	5.214 2	0.649 7	60.971 3	6.555 9	28.266 6	7.436 3
IOA	0.565 2	0.913 2	0.583 6	0.928 5	0.688 5	0.935 4
评价指标	$PM_{2.5}$/(μg·m^{-3})		O_3/(μg·m^{-3})		CO/(mg·m^{-3})	
	CAMQ	CMAQ-DL	CAMQ	CMAQ-DL	CAMQ	CMAQ-DL
RMSE	48.868 9	12.713 4	41.661 0	10.803 0	0.925 3	0.248 8
MAE	33.090 8	10.856 3	34.429 4	9.205 8	0.896 0	0.209 0
IOA	0.681 4	0.922 2	0.461 7	0.840 0	0.318 3	0.663 0

3.2 模型比较

为了检验 HHO-CNN-LSTM 模型的性能,本文将其与 CNN-LSTM、LSTM、支持向量机(SVM)三个模型进行比较。选取 2022 年 12 月的实测数据与 CMAQ 预报数据,使用上述四个模型对 CMAQ 预报结果进行修正,同样选择均方根误差、平均绝对误差和一致性指数作为评价指标,结果见表 2。

表 2 模型比较

Table 2 Model comparison

污染物	模型	*RMSE*	*MAE*	*IOA*
SO_2	HHO-CNN-LSTM	0.810 9	0.649 7	0.913 2
	CNN-LSTM	0.910 8	0.669 2	0.863 7
	LSTM	1.247 2	1.066 7	0.723 6
	SVM	1.546 2	1.412 4	0.716 9
NO_2	HHO-CNN-LSTM	7.966 7	6.555 9	0.928 5
	CNN-LSTM	8.502 4	6.808 3	0.918 4
	LSTM	9.841 7	8.473 5	0.903 9
	SVM	13.692 1	10.377 8	0.859 3
PM_{10}	HHO-CNN-LSTM	9.686 9	7.436 3	0.935 4
	CNN-LSTM	11.522 5	9.634 6	0.897 7
	LSTM	15.887 1	12.829 9	0.869 8
	SVM	15.460 5	10.792 2	0.897 6
$PM_{2.5}$	HHO-CNN-LSTM	12.713 4	10.856 3	0.922 2
	CNN-LSTM	14.022 2	12.223 5	0.906 0
	LSTM	15.658 8	12.970 2	0.854 6
	SVM	15.955 1	14.104 1	0.848 4
O_3	HHO-CNN-LSTM	10.803 0	9.205 8	0.840 0
	CNN-LSTM	13.303 4	11.424 1	0.804 4
	LSTM	15.482 2	13.317 0	0.740 9
	SVM	13.906 0	11.622 7	0.768 6
CO	HHO-CNN-LSTM	0.248 8	0.209 0	0.663 0
	CNN-LSTM	0.176 9	0.144 5	0.836 8
	LSTM	0.203 5	0.173 7	0.801 3
	SVM	0.169 8	0.137 8	0.828 0

比较表 1 和表 2 可以看出,四个模型修正后的预测值均更加接近实际值。其中,CNN-LSTM 模型对六项污染物浓度的预测效果均好于 LSTM 模型,可见卷积层在特征提取方面的优势。HHO-CNN-LSTM 模型在 SO_2、NO_2、PM_{10}、$PM_{2.5}$、O_3 这五种污染物浓度的修正效果上优于其他三个模型,相比于 CNN-LSTM 模型,HHO-CNN-LSTM 模型预测值的 *RMSE* 减少了 6.30%~18.80%,*MAE* 减少了 2.91%~22.82%,*IOA* 提升了 1.10%~5.73%,这是因为哈里斯鹰算法在训练过程中为混合神经网络找到了最优的超参数,提高了模型的预测性能。然而,在 CO 浓度的预测中,HHO-CNN-LSTM 模型的结果并不理想,在三项评价指标的表现上不如其他三个模型,这可能是因为哈里斯鹰算法在超参数迭代过程中陷入了局部最优[27],本文将针对这个问题对哈里斯鹰算法进行改进。

3.3 改进的哈里斯鹰优化算法及其表现

针对哈里斯鹰算法可能在对模型进行优化的过程中陷入了局部最优而导致对 CO 浓度预测效果不佳的问题,本文在算法迭代寻优过程中加入了高斯随机游走策略来对算法进行改进。利用优势种群的平均值来判断算法是否陷入停滞,当优势种群的平均值在连续两次迭代过程中没有变化,则认为算法陷入停滞。此时利用高斯随机游走策略生成新个体进而帮助哈里斯鹰算法跳出局部最优。高斯随机游走策略的公式如下:

$$X(t+1)=Gaussian(X(t),\sigma) \quad (4)$$

$$\sigma=\cos\left(\frac{\pi}{2}\times\left(\frac{t}{T}\right)^2\right)\times(X(t)-X^*(t)) \quad (5)$$

式(4~5)中 σ 为随机游走的步长,X 为从优势种群中随机选择的一个个体,t 和 T 分别为当前迭代次数和最大迭代次数。通过余弦函数在迭代前期施加较大扰动,迭代后期扰动迅速减小,进而平衡了算法的寻优能力。

将使用高斯随机游走策略改进后的哈里斯鹰算法应用于修正模型进行 CO 浓度的预测,结果如图 5 所示。由图 5 可知,基于改进的 HHO 优化的 CNN-LSTM 模型(GHHO-CNN-LSTM 模型)在 CO 浓度的预测效果上有了很大提升,预测值比其他模型更接近实际值。此外,将改进前后的修正模型进行比较(见表 3),发现两者在 SO_2、NO_2、PM_{10}、$PM_{2.5}$、O_3 五种污染物浓度上的预测效果相差无几,可见两个模型都在算法的优化下找到了最优的超参数组合。而在 CO 浓度的预测上,相比于改进前的 HHO-CNN-LSTM 模型,GHHO-CNN-LSTM 模型在三项指标的表现上均有了显著提升,*RMSE* 减少了 39.55%,*MAE* 减少了45.93%,*IOA* 提高了 32.43%。可见加入了高斯随机游走策略的哈里斯鹰算法有效解决了传统哈里斯鹰算法在寻优迭代过程中易陷入局部最优的问题,提高了修正模型在 CO 浓度上的预测精度。

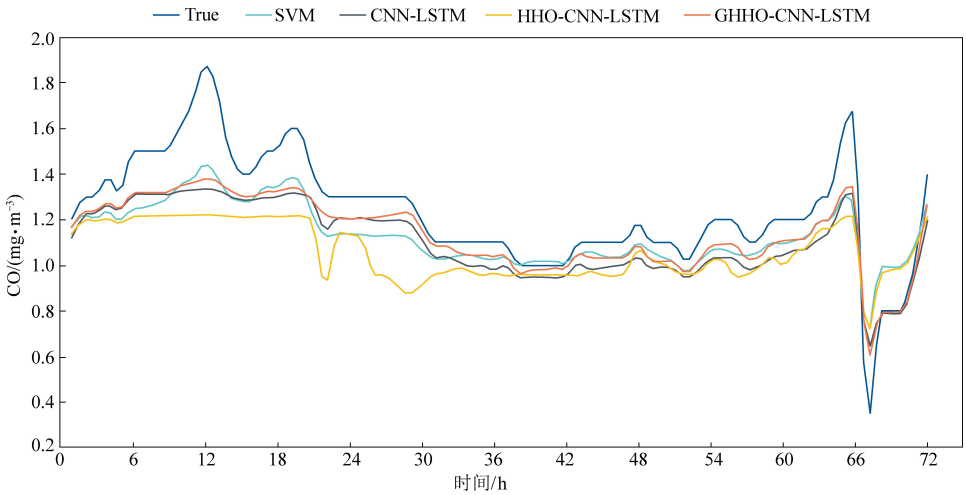

图 5　CO 浓度预测结果

Fig. 5　Prediction results of CO concentration

表 3　改进前后的模型评价指标结果

Table 3　Model evaluation index results before and after improvement

污染物	模型	*RMSE*	*MAE*	*IOA*
SO_2	HHO-CNN-LSTM	0.810 9	0.649 7	0.913 2
	GHHO-CNN-LSTM	0.800 5	0.629 5	0.909 8
NO_2	HHO-CNN-LSTM	7.966 7	6.555 9	0.928 5
	GHHO-CNN-LSTM	7.927 0	6.504 2	0.928 6
PM_{10}	HHO-CNN-LSTM	9.686 9	7.436 3	0.935 4
	GHHO-CNN-LSTM	9.822 4	7.536 7	0.933 7
$PM_{2.5}$	HHO-CNN-LSTM	12.713 4	10.856 3	0.922 2
	GHHO-CNN-LSTM	12.020 2	10.290 6	0.930 6
O_3	HHO-CNN-LSTM	10.803 0	9.205 8	0.840 0
	GHHO-CNN-LSTM	12.726 7	11.024 7	0.817 1
CO	HHO-CNN-LSTM	0.248 8	0.209 0	0.663 0
	GHHO-CNN-LSTM	0.150 4	0.113 0	0.878 0

4　总　　结

在日益加剧的空气污染严重影响人们身体健康和社会经济发展的背景下,建立空气质量模型预测污染物浓度具有重要意义。然而传统的空气质量模型 CMAQ 对污染物浓度的预报精度并不理

想。基于此,本文在使用 CMAQ 模型对上海市 2022 年 12 月六种大气污染物(SO_2、NO_2、PM_{10}、$PM_{2.5}$、O_3、CO)浓度进行预测的基础上,建立 HHO-CNN-LSTM 模型对预报结果进行修正,显著提高了预测精度,*RMSE* 减少了 73.11%~91.31%,*MAE* 减少了 67.19%~89.25%,*IOA* 提升了 35.34%~108.29%。并针对 HHO 算法陷入局部最优而导致对 CO 浓度修正效果不佳的问题,使用高斯随机游走策略对算法进行改进,提高了修正模型在 CO 浓度上的预测精度。与改进前相比,*RMSE* 减少了 39.55%,*MAE* 减少了 45.93%,*IOA* 提高了 32.43%。本文的工作为人工智能技术与传统空气质量模型的结合提供了思路,为大气污染物的防治作出了贡献。

然而,基于深度学习的预测方法也存在其局限性。例如,在中长期预测中可能会出现显著误差以及面临“缺乏可解释性”和“对极端天气条件的不准确预测”等挑战。因此,在未来的研究中,可以用更大时间尺度上的数据训练该模型,并将反映基础物理机制的数学方程式纳入神经网络架构中,以此建立一个具有更强的普适性和可解释性的空气质量预测模型。

参考文献(References):

[1] ZHENG B, ZHANG Q, ZHANG Y, et al. Heterogeneous chemistry: A mechanism missing in current models to explain secondary inorganic aerosol formation during the episode in north China[J]. Atmospheric Chemistry & Physics, 2015, 14(15): 2031-2049.

[2] ZHE W, WANG L, CHEN M, et al. The 2013 severe haze over the Southern Hebei, China: $PM_{2.5}$ composition and source apportionment[J]. Atmospheric Pollution Research, 2014, 5(4): 759-768.

[3] HU J, CHEN J, YING Q, et al. One-year simulation of ozone and particulate matter in china using WRF/CMAQ modeling system[J]. Atmospheric Chemistry & Physics Discussions, 2016, 16(16): 10333-10350.

[4] NAPELENOK S L, VEDANTHAM R, BHAVE P, et al. Source-receptor reconciliation of fine-particulate emissions from residential wood combustion in the southeastern United States[J]. Atmospheric Environment, 2014, 98: 454-460.

[5] KOO Y S, KIM S T, CHO J S, et al. Performance evaluation of the updated air quality forecasting system for Seoul predicting PM_{10}[J]. Atmospheric Environment, 2012, 58(30): 56-69.

[6] WANG N, GUO H, JIANG F, et al. Simulation of ozone formation at different elevations in mountainous area of Hong Kong using WRF - CMAQ model[J]. Science of the Total Environment, 2015, 505(36): 939-951.

[7] 武文琪. 基于灰色 GM(1,1)模型的成都市大气污染物浓度预测[J]. 能源环境保护, 2019, 33(2): 56-58+55.
WU Wenqi. Air pollutant concentration prediction in Chengdu based on grey GM(1,1) model[J]. Energy Environmental Protection, 2019, 33(2): 56-58+55.

[8] 谢敏, 钟流举, 陈焕盛, 等. CMAQ 模式及其修正预报在珠三角区域的应用检验[J]. 环境科学与技术, 2012, 35(2): 96-101.
XIE Min, ZHONG Liuju, CHEN Huansheng, et al. Application and verification of CMAQ model and revision forecast in Pearl River Delta Region[J]. Environmental Science & Technology, 2012, 35(2): 96-101.

[9] 王茜, 吴剑斌, 林燕芬. CMAQ 模式及其修正技术在上海市 $PM_{2.5}$ 预报中的应用检验[J]. 环境科学学报, 2015, 35(6): 1651-1656.
WANG Qian, WU Jianbin, LIN Yanfen. Implementation of a dynamic linear regression method on the CMAQ forecast of $PM_{2.5}$ in Shanghai[J]. Acta Scientiae Circumstantiae, 2015, 35(6): 1651-1656.

[10] 芦华, 吴钲, 刘伯骏, 等. 空气质量模式在重庆主城区预报效果检验订正[J]. 西南大学学报(自然科学版), 2021, 43(7): 176-184.
LU Hua, WU Zheng, LIU Bojun, et al. Test and correction of forecast effect by air quality numerical prediction models in Chongqing urban city[J]. Journal of Southwest University (Natural Science Edition), 2021, 43(7): 176-184.

[11] SAYEED A, LOPS Y, CHOI Y, et al. Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks[J]. Atmospheric Environment, 2021, 253: 118376-118384.

[12] YI X, DUAN Z, LI R, et al. Predicting fine-grained air quality based on deep neural networks[J]. IEEE Transactions on Big Data, 2020, 8(5): 1326-1339.

[13] XAYASOUK T, LEE H M, LEE G. Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models[J]. Sustainability, 2020, 12(6): 2570-2587.

[14] PAK U, KIM C, RYU U, et al. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction[J]. Air Quality, Atmosphere & Health, 2018,11(8): 883-895.

[15] LI T, HUA M, WU X. A hybrid CNN-LSTM model for forecasting particulate matter ($PM_{2.5}$)[J]. IEEE Access, 2020, 8: 26933-26941.

[16] DU S, LI T, YANG Y, et al. Deep air quality forecasting using hybrid deep learning framework[J]. IEEE Transaction on Knowledge and Data Engineering, 2021, 33(6): 2412-2424.

[17] ZHAO B, LU H, CHEN S, et al. Convolutional neural networks for time series classification[J]. Journal of Systems Engineering and Electronics, 2017, 28(1): 162-169.

[18] SHAHHOSSEINI M, HU G, KHAKI S, et al. Corn yield prediction with ensemble CNN-DNN[J]. Frontiers in Plant Science, 2021, 12: 709008-709020.

[19] KANJO E, YOUNIS E, ANG C S. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection[J]. Information Fusion, 2019, 49: 46-56.

[20] LI X, PENG L, YAO X, et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation[J]. Environmental pollution, 2017, 231: 997-1004.

[21] MA J, DING Y, CHENG J C P, et al. A Lag-FLSTM deep learning network based on Bayesian Optimization for multi-sequential-variant $PM_{2.5}$ prediction[J]. Sustainable Cities and Society, 2020, 60: 102237-102246.

[22] HE Q Q, WU C, SI Y W. LSTM with Particle Swam Optimization for sales forecasting[J]. Electronic Commerce Research and Applications, 2022, 51: 101118-101137.

[23] ZHOU Y, SHI J, CHEN H, et al. Interval prediction of photovoltaic output based on WOA-LSTM-LSSVM combined model [C]. Chongqing: 2021 6th Asia Conference on Power and Electrical Engineering (ACPEE). IEEE, 2021: 514-519.

[24] HORA S K, POONGODAN R, PRADO R P, et al. Long short-term memory network-based metaheuristic for effective electric energy consumption prediction[J]. Applied Sciences, 2021, 11(23): 11263-11282.

[25] HEIDARI AA, MIRJALILI S, FARIS H, et al. Harris Hawks Optimization: Algorithm and applications [J]. Future Generation Computer Systems, 2019, 97: 849-872.

[26] TIAN M, WANG H B, CHEN Y, et al. Highly time-resolved characterization of water-soluble inorganic ions in $PM_{2.5}$, in a humid and acidic mega city in Sichuan Basin, China[J]. Science of the Total Environment, 2017, 580: 224-234.

[27] 高岳林, 杨钦文, 王晓峰, 等. 新型群体智能优化算法综述[J]. 郑州大学学报(工学版), 2022, 43(3): 21-30.
GAO Yuelin, YANG Qinwen, WANG Xiaofeng, et al. Overview of new swarm intelligent optimization algorithms[J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(3): 21-30.