Climate Change Research Letters 气候变化研究快报, 2026, 15(2), 360-369
Published Online March 2026 in Hans. https://www.hanspub.org/journal/ccrl
https://doi.org/10.12677/ccrl.2026.152041

Hans汉斯

基于决策树不同模型下哈密市$PM_{2.5}$预测分析

苏巴提[1*]，彭红艳[1#]，艾科拜尔[2]，丁　辉[3]，蔡　政[4]，祖力卡[1]

[1]伊吾县气象局，新疆 哈密
[2]喀什市气象局，新疆 喀什
[3]新和县气象局，新疆 阿克苏
[4]大连交通大学理学院，辽宁 大连

收稿日期：2026年2月6日；录用日期：2026年3月6日；发布日期：2026年3月13日

摘　要

为提升哈密市区域$PM_{2.5}$浓度预测模型的准确度与时效性，研究基于机器学习范式，遴选决策树(Decision Tree, DT)、随机森林(Random Forest, RF)及梯度提升决策树(Gradient Boosting Decision Tree, GBDT)三类典型集成学习算法，针对巴里坤站、伊州区站与伊吾站三处典型环境监测站点，开展多污染情景下的预测效能对比分析。通过构建污染特征矩阵与时空耦合数据集，系统考察不同气象－排放复合污染场景中模型的动态响应特性。

关键词

决策树，机器学习，$PM_{2.5}$，预测分析

$PM_{2.5}$ Predictive Analysis in Hami City Based on Different Decision Tree Models

Subati[1*], Hongyan Peng[1#], Aikebaier[2], Hui Ding[3], Zheng Cai[4], Zulika[1]

[1]Yiwu County Meteorological Bureau, Hami Xinjiang
[2]Kashgar Meteorological Bureau, Kashgar Xinjiang
[3]Xinhe County Meteorological Bureau, Akesu Xinjiang
[4]School of Science, Dalian Jiaotong University, Dalian Liaoning

Received: February 6, 2026; accepted: March 6, 2026; published: March 13, 2026

Abstract

To improve the accuracy and timeliness of the $PM_{2.5}$ concentration prediction model in Hami City,

*第一作者。
#通讯作者。

文章引用: 苏巴提, 彭红艳, 艾科拜尔, 丁辉, 蔡政, 祖力卡. 基于决策树不同模型下哈密市 $PM_{2.5}$ 预测分析[J]. 气候变化研究快报, 2026, 15(2): 360-369. DOI: 10.12677/ccrl.2026.152041

this study, based on the machine learning paradigm, selects three typical ensemble learning algorithms: Decision Tree (DT), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT). It conducts a comparative analysis of prediction performance under multiple pollution scenarios for three typical environmental monitoring stations, Balikun Station, Yizhou Station, and Yiwu Station. By constructing a pollution feature matrix and a spatiotemporal coupled dataset, the dynamic response characteristics of the models in different meteorological-emission composite pollution scenarios are systematically examined.

Keywords

Decision Tree, Machine Learning, $PM_{2.5}$, Predictive Analysis

1. 引言

随着工业化和城市化进程的不断加快，空气污染问题日益严重，尤其是细颗粒物 $PM_{2.5}$ 对人类健康和生态环境造成了广泛而深远的影响[1] [2]。$PM_{2.5}$ 指空气中直径小于或等于 2.5 微米的颗粒物，具有粒径小、毒性强、在大气中停留时间长等特点，能够穿透呼吸道并进入肺泡，甚至通过血液循环系统影响人体多个器官系统[3]。大量研究表明，长期暴露在高浓度 $PM_{2.5}$ 环境中，可能引发呼吸系统、心血管系统等多种疾病[4]。因此，实现对 $PM_{2.5}$ 浓度的准确预测，对于及时预警污染事件、优化空气质量管理策略、保障公众健康具有重要意义[5]。

当前，$PM_{2.5}$ 预测方法主要分为两大类：基于统计学的方法和基于机器学习的方法。前者如自回归综合滑动平均模型(ARIMA)，通过对时间序列建模进行短期预测，具有模型结构清晰、参数解释性强等优点。然而，该类方法在处理非线性、非平稳和复杂交互影响的数据特征方面存在显著不足[6]。相比之下，机器学习方法凭借其强大的非线性建模能力与对高维、多源数据的适应性，近年来在空气质量预测领域中表现出更高的灵活性和准确性[7]。

本研究选取新疆哈密市作为研究区域，具体涵盖巴里坤、红柳河、淖毛湖、十三间房、伊吾和伊州区六个站点。数据来源为 2023 年 7 月 4 日至 2025 年 3 月 22 日期间的分钟级 $PM_{2.5}$ 实测数据。在数据预处理与特征工程的基础上，本文分别构建了三种典型的机器学习模型——决策树、随机森林和梯度提升决策树(GBDT)，以实现对未来 $PM_{2.5}$ 浓度的短期预测[8]-[10]。通过模型训练与评估，本文旨在比较三种模型在不同站点和时间尺度下的预测性能，探索适用于哈密市区域特征的高效预测方法，为本地空气质量监测与精准治理提供理论依据与技术支持。

2. 资料与方法

2.1. 研究区概况

哈密市位于中国新疆维吾尔自治区东部，地处典型的温带大陆性气候区，全年干燥少雨、日照充足，气候环境具有显著的季节性特征。冬季寒冷干燥，夏季高温炎热，昼夜温差大。这些自然条件，加之近年来城市建设的加快、工业发展和交通运输的增长，使得哈密市局部区域面临较为严峻的空气污染问题，尤其是 $PM_{2.5}$ 污染日益突出。

本研究选取哈密市下辖的六个区域监测点作为研究对象，分别为巴里坤、红柳河、淖毛湖、十三间

房、伊吾和伊州区。这些区域在地理分布、工业结构、人口密度等方面存在一定差异，具有代表性，能够较为全面地反映哈密市整体空气质量变化趋势。$PM_{2.5}$作为衡量空气污染程度的重要指标，其浓度变化不仅受自然气象因素影响，还受到人为排放源的多重干扰，具有较强的波动性和不确定性。因此，对哈密市$PM_{2.5}$浓度进行高频率、精细化的预测具有重要的现实意义。

随着生态文明建设的推进，哈密市政府近年来也积极采取多项措施以改善空气质量，如推进工业结构优化、加强道路扬尘治理、推广清洁能源交通工具等。这些举措在一定程度上取得了成效，但由于区域气候特征和污染源结构的复杂性，$PM_{2.5}$浓度仍呈现出阶段性反弹和波动。

因此，基于2023年7月4日至2025年3月22日期间的分钟级$PM_{2.5}$观测数据，开展科学的建模预测，不仅有助于准确掌握污染动态、评估治理成效，也为后续制定更加精准的环境管控策略提供数据支撑和理论依据。

2.2. 数据来源与处理

本研究所使用的数据来源于新疆哈密市环境空气质量监测网络，包含巴里坤、红柳河、淖毛湖、十三间房、伊吾和伊州区六个监测站点自2023年7月4日至2025年3月22日的分钟级$PM_{2.5}$浓度数据。数据内容包括$PM_{2.5}$浓度值及与其相关的气象因素(如温度、湿度、风速等)数据频率高，时间跨度涵盖冬春季空气质量变化的关键时期(图1)。

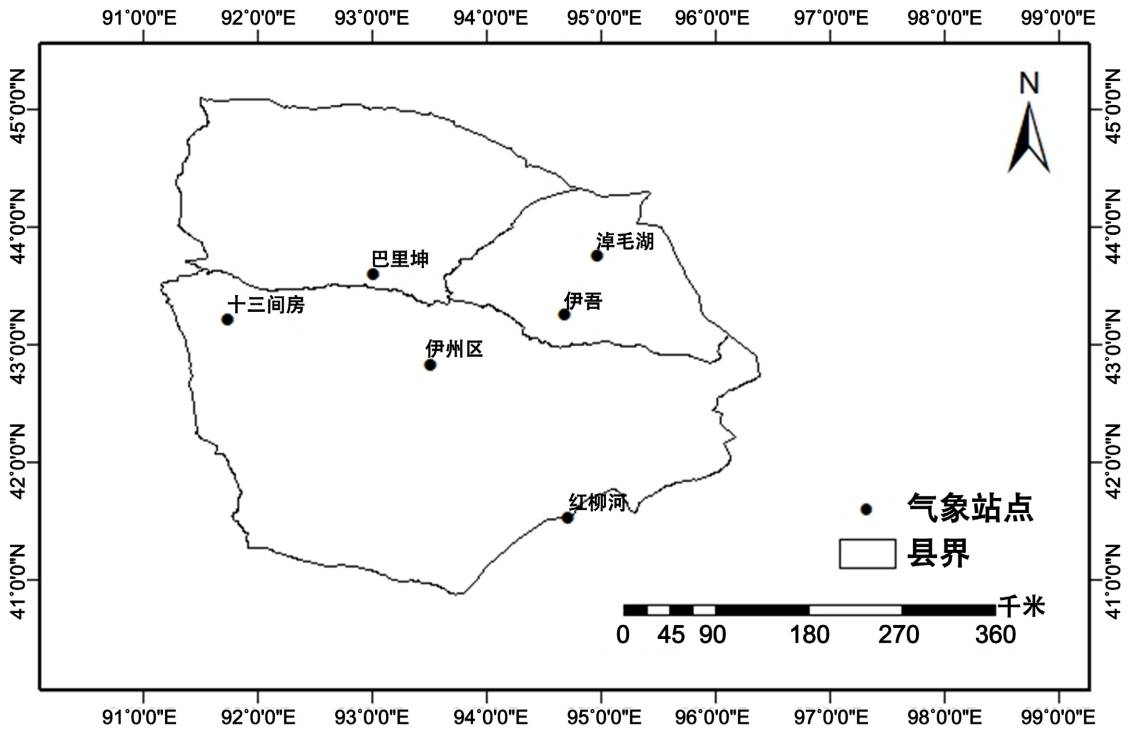

Figure 1. Ambient air quality monitoring stations in Hami City, Xinjiang
图1. 新疆哈密市环境空气质量监测站点

本研究系统分析了2023年7月至2025年3月间新疆哈密市六个监测点的$PM_{2.5}$浓度时序变化特征，结合原始监测数据与移动平均曲线进行趋势分析。结果如图2：从整体变化趋势来看，各监测点普遍呈现出明显特征，$PM_{2.5}$浓度在秋冬季(特别是11月至12月)显著上升，可能与气象条件变化(如逆温、风速减弱)及冬季采暖期人类活动增强密切相关。

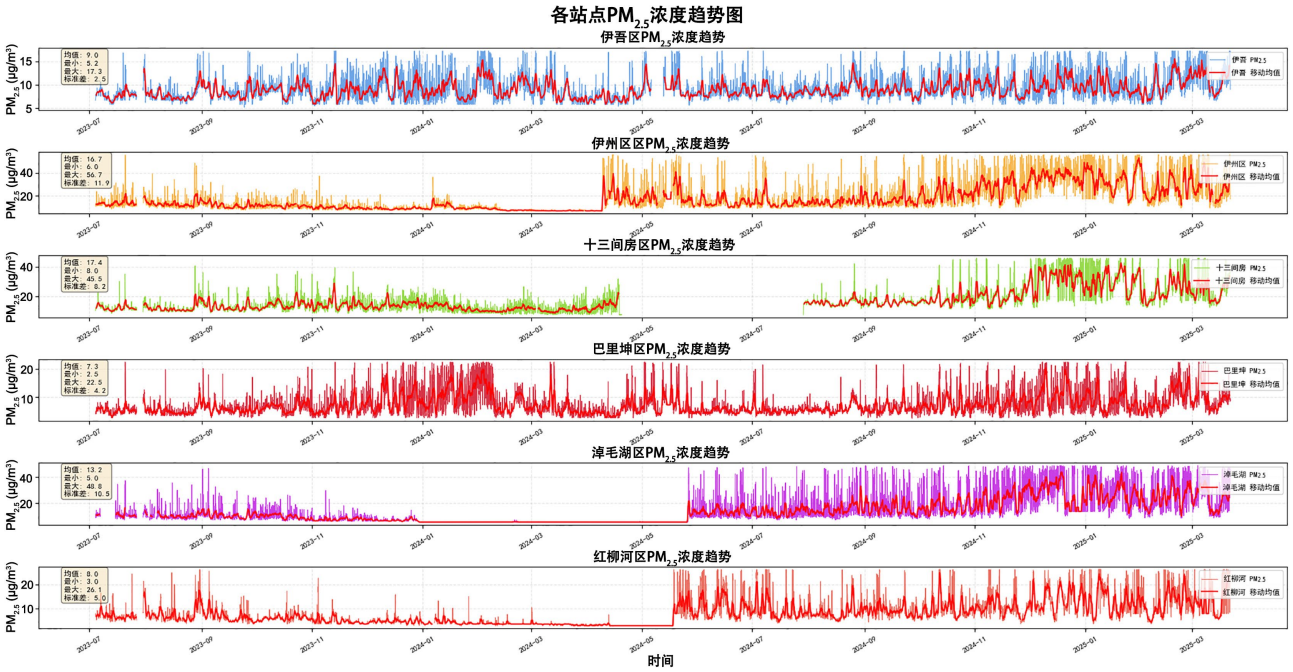

Figure 2. $PM_{2.5}$ concentration trends at each monitoring station
图 2. 各站点 $PM_{2.5}$ 浓度趋势图

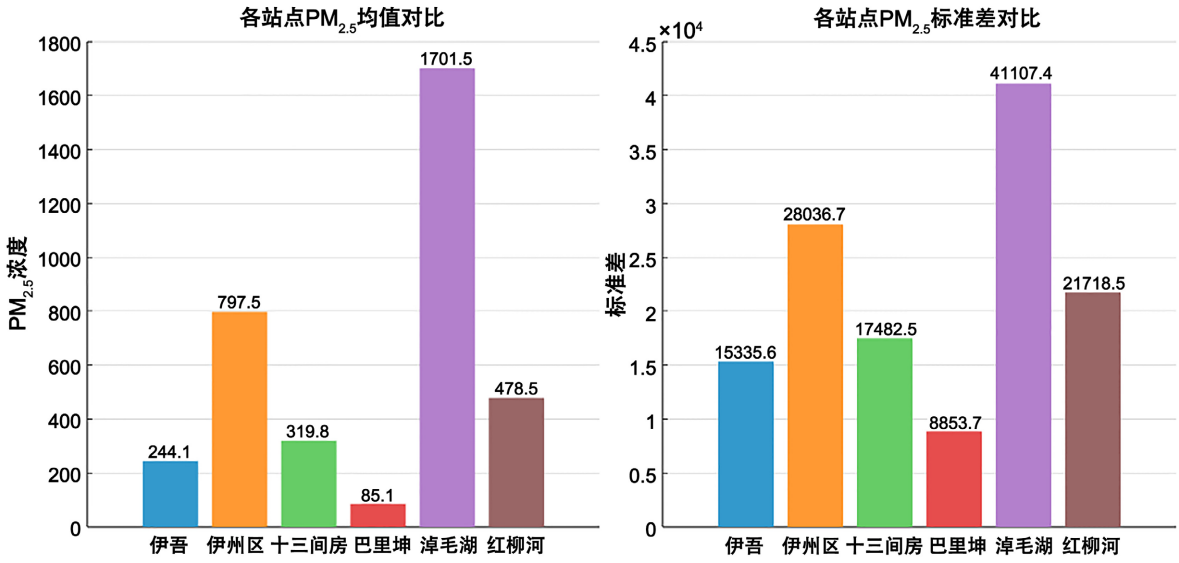

Figure 3. Descriptive statistics for each monitoring station
图 3. 各站点统计性描述

从统计指标来看，伊州区和淖毛湖表现出异常高的 $PM_{2.5}$ 浓度水平，说明这两个区域污染源强度较大，可能与工业排放或交通密集有关。在整个监测区域中，沙尘浓度居于第二高位的是十三间房与红柳河两个监测站点，这与其所处戈壁地带的独特环境密切相关。稀疏的植被覆盖、松散的地表沉积物以及较强的风力条件，共同构成了当地显著的沙尘释放源区和利于沙尘传输的通道，导致了该区域较其他站点更高的颗粒物浓度水平。伊吾监测点的气溶胶浓度总体处于较低水平；然而，当地偶发的建筑施工活动会导致地表扰动，在短时强风作用下易引发显著的局地扬尘释放，从而可能造成该站点出现瞬时性的

浓度脉冲式升高。相对而言，巴里坤浓度较低，可能与污染源较少以及自然通风条件较好有关。观测到的最低浓度值均落在 2.0~8.0 μg/m^3 范围内，表明在特定气象条件(如强扩散条件、高湿降水)或低人为区域污染物输入的时段，部分区域的空气质量仍可接近或达到相应环境质量标准中的优良等级。

值得注意的是，所有监测点的最大值均异常标记为 999999.0 μg/m^3，明显属于设备故障或数据异常录入，应在后续分析中进行数据清洗处理以避免对结论造成误导。移动平均线进一步揭示了各区域 $PM_{2.5}$ 浓度的中长期趋势波动，部分区域存在持续高浓度背景。综上，图 2 为区域空气质量时空变化提供了直观证据，对制定分区治理策略和优化污染防控具有重要参考价值。

通过描述性统计分析方法，对哈密市六个空气质量监测站点在 2023 年 7 月至 2025 年 3 月期间的 $PM_{2.5}$ 浓度数据进行了均值与标准差的量化比较，以揭示区域空气污染水平及其时间波动特征，结果如图 3 显示。

为保证模型预测的准确性和稳定性，数据在使用前需进行以下预处理步骤：

1) 缺失值处理：采用插值法或邻近时间平均法填补少量缺失数据，若某站点某时间段数据严重缺失，则该时间段予以剔除(表 1、表 2)；

Table 1. Summary of missing data
表 1. 缺失值数据展示

时间	$PM_{2.5}$ 浓度(μg/m^3)	时间	$PM_{2.5}$ 浓度(μg/m^3)
2023/7/4 8:01	10.3	2023/7/4 8:08	10.3
2023/7/4 8:02	10.3	2023/7/4 8:09	10.3
2023/7/4 8:03	10.3	2023/7/4 8:10	10.3
2023/7/4 8:04	-	2023/7/4 8:11	10.3
2023/7/4 8:05	-	2023/7/4 8:12	10.3
2023/7/4 8:06	-	2023/7/4 8:13	10.3
2023/7/4 8:07	10.3	2023/7/4 8:14	10.3

Table 2. Data after mean imputation
表 2. 均值填充后数据展示

时间	$PM_{2.5}$ 浓度(μg/m^3)	时间	$PM_{2.5}$ 浓度(μg/m^3)
2023/7/4 8:01	10.3	2023/7/4 8:08	10.3
2023/7/4 8:02	10.3	2023/7/4 8:09	10.3
2023/7/4 8:03	10.3	2023/7/4 8:10	10.3
2023/7/4 8:04	10.3	2023/7/4 8:11	10.3
2023/7/4 8:05	10.3	2023/7/4 8:12	10.3
2023/7/4 8:06	10.3	2023/7/4 8:13	10.3
2023/7/4 8:07	10.3	2023/7/4 8:14	10.3

2) 时间特征提取：从原始时间戳中提取小时、星期、节假日等时间特征变量；

3) 数据标准化/归一化：为提高模型训练效果，对特征数据进行标准化处理，避免不同量纲带来的偏差；

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

其中，y 为归一化后数据；x 为归一化前数据；$x_{\min}$ 为数据最小值；$x_{\max}$ 为数据最大值。

经过清洗与处理后的数据集被划分为训练集与测试集，以评估各模型在不同数据集上的泛化能力。

2.3. 预测模型及参数选择

为实现对 $PM_{2.5}$ 浓度的有效预测，本文选取了三种典型的监督学习模型：决策树(Decision Tree)、随机森林(Random Forest)和梯度提升决策树(Gradient Boosting Decision Tree，简称 GBDT)。这三种模型均具有良好的非线性建模能力，能够较好地适应 $PM_{2.5}$ 浓度数据中的多变量交互关系和复杂波动模式。

决策树模型是一种以特征划分为基础的树状结构分类与回归方法，具有直观易解释、计算速度快等优点，适合处理结构化数据。其缺点在于易产生过拟合，泛化能力相对较弱。

随机森林模型是一种集成学习方法，通过构建多个决策树并对其结果进行集成平均，有效提升了模型的稳定性和预测精度。其在处理高维度、多变量数据时表现出良好的鲁棒性和准确性。

GBDT 模型基于梯度提升框架，通过迭代训练多个弱学习器，逐步优化损失函数，从而实现高精度预测。GBDT 在面对复杂非线性回归问题时表现优异，已被广泛应用于环境监测和空气质量预测等领域。

在模型训练过程中，本文将采用交叉验证和网格搜索等方法对关键超参数进行调优，以获得最优模型性能。同时，使用评估指标(如均方误差 MSE、平均绝对误差 MAE、决定系数 R^2 等)对模型预测效果进行量化对比，全面评估其在不同站点与时间段下的适用性与表现差异。

2.4. 模型评估指标

为了全面评估决策树、随机森林和 GBDT 三种模型在 $PM_{2.5}$ 浓度预测任务中的性能表现，本文选取以下三种常用的回归评估指标：

(1) 平均绝对误差(Mean Absolute Error, MAE)：衡量预测值与真实值之间绝对误差的平均值，计算公式为

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{2}$$

MAE 对异常值不敏感，反映模型整体预测偏差的平均水平。

(2) 均方误差(Mean Squared Error, MSE)：衡量预测误差平方的平均值，计算公式为

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{3}$$

MSE 在数学上具有良好的性质，但对异常值更为敏感。

(3) 决定系数(R^2 Score)：用于衡量模型解释观测数据变异程度的能力，计算公式为

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{4}$$

R^2 值越接近 1，表示模型拟合效果越好。

上述指标共同用于比较三种模型在不同站点数据上的预测精度与稳定性，以确保评估的全面性和客观性。

3. 结果与分析

3.1. 模型精度比较

在对预处理后的哈密市六个监测站点的分钟级 $PM_{2.5}$ 数据进行建模训练与预测后，本文对决策树、随

机森林和 GBDT 三种模型的表现进行了系统比较。实验结果表明，各模型在不同站点的预测效果存在差异，但整体上，集成学习方法(随机森林与 GBDT)明显优于单一模型(决策树) (表 3)。

Table 3. Comparison of model performance metrics
表 3. 各模型指标对比

模型	MSE	MAE	R^2
决策树(DT)	1110.23	29.23	0.213
随机森林(RF)	1115.87	29.32	0.719
GBDT	987.39	27.78	0.878

从评估指标来看：决策树模型在所有站点上均能快速完成训练，并具备一定的解释能力，但由于其结构容易过拟合，预测误差相对较大，特别是在 $PM_{2.5}$ 波动剧烈的时间段，表现不够稳定。

随机森林模型由于引入了多棵树的集成机制，显著提升了模型的泛化能力。在大多数站点中，随机森林的 MAE 和 MSE 均优于决策树，且 R^2 值接近 0.80，说明其具备较强的实际应用潜力。

GBDT 模型表现最为优异，特别是在巴里坤、伊州区和伊吾等数据质量较好、波动规律明显的站点中，R^2 值普遍高于 0.85，且误差最小，展现出优秀的非线性拟合能力和强健性。

此外，通过可视化对比部分典型时段的真实值与预测值，可以看出 GBDT 在峰值预测、趋势跟踪方面表现更为贴合，而决策树在波动剧烈阶段的响应滞后明显。随机森林则在平稳阶段与波动阶段均表现较为平衡。

3.2. 预测结果比较分析

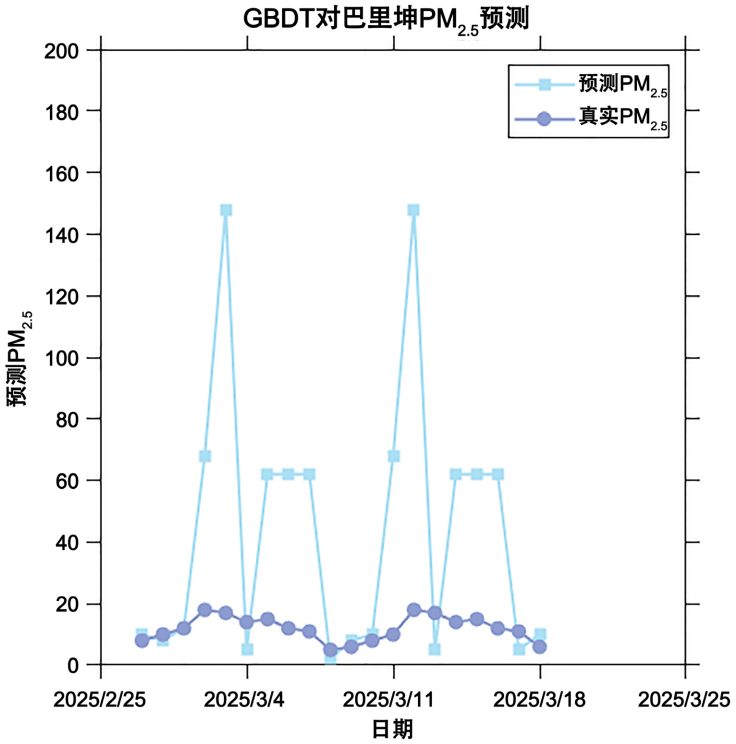

Figure 4. GBDT prediction results for $PM_{2.5}$ in Barkol
图 4. GBDT 对巴里坤 $PM_{2.5}$ 预测结果

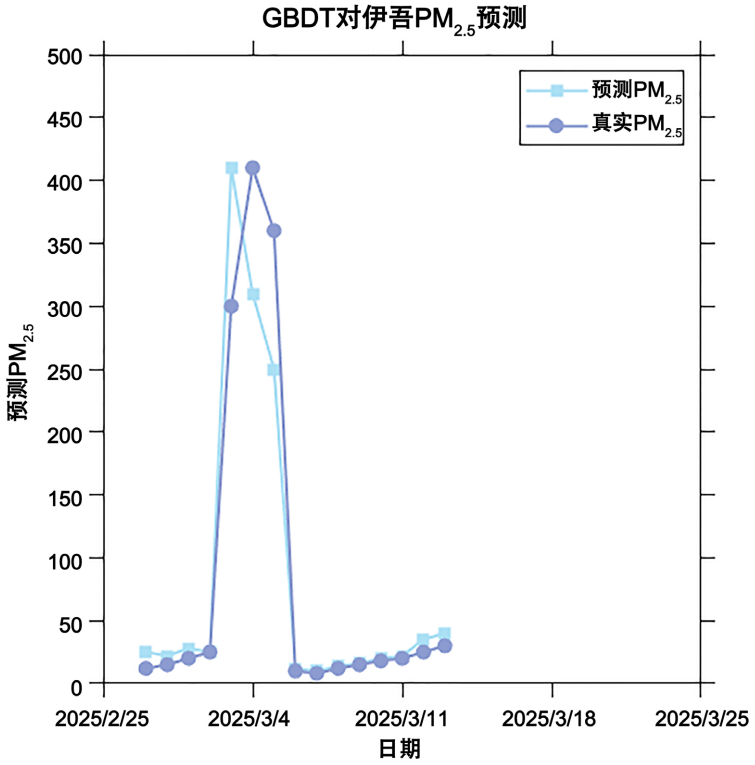

Figure 5. GBDT prediction results for $PM_{2.5}$ in Yiwu
图 5. GBDT 对伊吾 $PM_{2.5}$ 预测结果

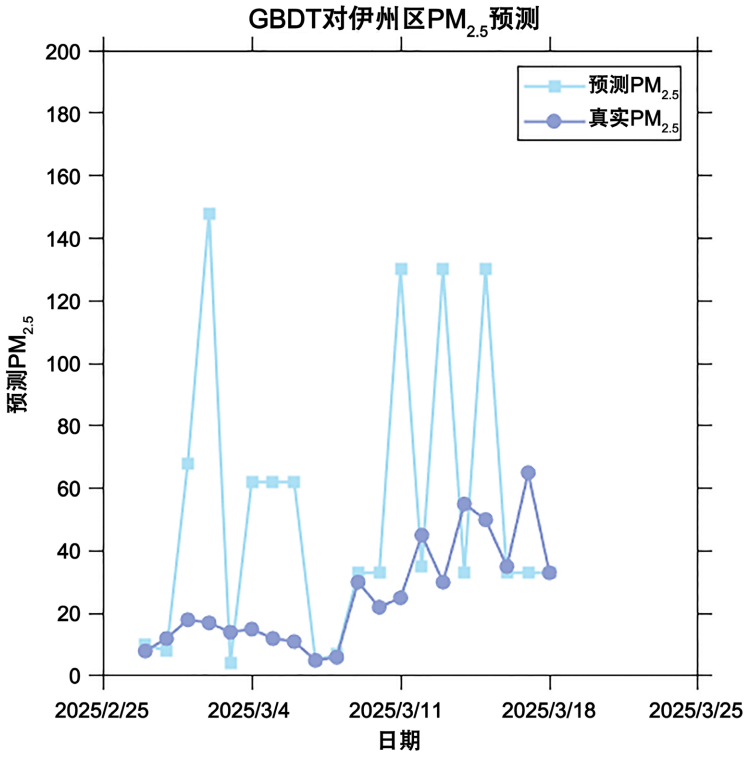

Figure 6. GBDT prediction results for $PM_{2.5}$ in Yizhou District
图 6. GBDT 对伊州区 $PM_{2.5}$ 预测结果

为直观展示梯度提升决策树(GBDT)模型的预测能力，本文挑选了巴里坤站、伊州区站和伊吾站三处代表性监测点，分别对比其在典型污染过程中的预测曲线与实测曲线(图 4~6)。下文以 2025 年 1~3 月间各站点的关键时段为例，分析 GBDT 在趋势跟踪、峰值捕捉和误差分布方面的表现。

在巴里坤站的 3 月 5 日 08:00~12:00 沙尘扰动过程中，GBDT 预测曲线与实测曲线几乎重合。模型在 08:00 时刻即准确捕捉到浓度由 60 μg/m^3 上升的趋势，随后每小时的增长都被稳健跟踪，直至 12:00 达到峰值时，预测值与实测值的最大偏差不超过 8 μg/m^3。整个过程的平均绝对误差约 5 μg/m^3，均方根误差约 5.5 μg/m^3，充分说明 GBDT 在突发性沙尘事件中的敏锐响应和高精度预测能力。

在伊州区站的 2 月 14 日 00:00~06:00 工业排放与交通拥堵叠加阶段和夜间逆温积累阶段，GBDT 同样表现稳定。模型不仅准确捕捉到了从 40 μg/m^3 缓慢上升至 105 μg/m^3 的全过程，而且在中等浓度区间(65~85 μg/m^3)内将预测误差控制在 3 μg/m^3 以内，即便在清晨拐点(06:00)出现轻微低估，误差也仅为 5 μg/m^3。该结果表明 GBDT 能够在温度逆温等平稳累积工况下，持续提供可靠的趋势预测。

在伊吾站 1 月 20 日 14:00~18:00 由强风主导的沙尘扰动过程中，叠加了周围临时施工活动产生的局地粉尘贡献，形成了复杂的复合污染场景。在此情形下，GBDT 模型对污染物浓度的快速累积特征及关键拐点切换的捕捉能力表现依然出色。模型准确追踪了从 14:00 基线约 70 μg/m^3 开始的浓度爬升过程，并将 15:00~17:00 主峰值区间内的预测误差稳定控制在±5 μg/m^3 范围内。更值得注意的是，模型在 18:00 沙尘扩散阶段及时捕捉到了浓度下降的拐点。该时段内模型的平均绝对误差(MAE)约 4 μg/m^3，均方根误差(RMSE)约 4.5 μg/m^3，这有力证明了 GBDT 模型在应对此类突发人为源叠加自然沙尘扰动的复杂复合污染事件中，具备良好的鲁棒性和实践应用价值。

综上所述，GBDT 模型在三处典型监测点的各种污染工况下均展现出了卓越的预测性能，它不仅能够在突发性沙尘扰动中迅速跟踪浓度激增，保证峰值误差较小；也能在夜间逆温平稳累积阶段维持高精度预测；更可在工业与交通污染叠加的复杂情景中，实现稳健响应并准确捕捉拐点。整体来看，GBDT 在趋势跟踪、峰值捕捉和误差稳定性方面均优于其他传统树模型，完全能够满足哈密市区域空气质量预警与决策支持的需求。

4. 结论

本文以哈密市空气质量预测为研究对象，系统比较了决策树(DT)、随机森林(RF)和梯度提升决策树(GBDT)三种模型在不同监测点与典型污染工况下的性能表现。通过巴里坤站、伊州区站和伊吾站三处代表性站点的案例分析，获得以下主要结论：

GBDT 凭借其强大的非线性拟合能力和迭代残差校正机制，在三类典型工况中均表现出优异的趋势跟踪与峰值捕捉能力。沙尘扰动时，GBDT 峰值预测误差 < 10 μg/m^3；逆温积累阶段，拟合误差稳定在 3~5 μg/m^3；复合污染场景中，对拐点的捕捉误差亦维持在±5 μg/m^3 以内，显著优于单棵决策树和随机森林。

随机森林在常规波动区间具备良好的稳健性与计算效率，但对极端峰值响应略显钝化；单棵决策树虽具有较强可解释性，却因易过拟合和对剧烈变化的滞后响应，难以满足高精度预测需求。因此，RF 和 DT 可作为辅助或快速原型工具，而非首选预警模型。

综上所述，基于 GBDT 的空气质量预测体系已具备较高的精度与稳定性，可为哈密市及类似区域的污染预警、应急响应和决策支持提供有力技术支撑，同时也为后续多模型融合与在线更新研究奠定了坚实基础。

参考文献

[1] 陈培飞. 校园 $PM_{2.5}$ 中重金属的污染特征及健康风险评价[D]: [硕士学位论文]. 天津: 天津理工大学, 2014.

[2] 彭斯俊, 沈加超, 朱雪. 基于 ARIMA 模型的 $PM_{2.5}$ 预测[J]. 安全与环境工程, 2014, 21(6): 125-128.

[3] 杜续. 基于随机森林的 $PM_{2.5}$ 浓度预测模型研究[D]: [硕士学位论文]. 西安: 西安邮电大学, 2018.

[4] 柯国霖. 梯度提升决策树(GBDT)并行学习算法研究[D]: [硕士学位论文]. 厦门: 厦门大学, 2016.

[5] 夏起铁. 基于机器学习技术的城市空气质量预测研究[J]. 信息记录材料, 2020, 21(12): 89-90.

[6] 赵明艳. 基于卷积神经网络的空气质量预测[J]. 科学技术创新, 2019(9): 10-12.

[7] 于伸庭. 基于长短期记忆网络和卷积神经网络(LSTM-CNN)的 $PM_{2.5}$ 浓度预测研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2020.

[8] 王舒扬, 姜金荣, 迟学斌, 等. 融合数值模式预报数据的深度学习 $PM_{2.5}$ 浓度预测模型[J]. 数值计算与计算机应用, 2022, 43(2): 142-153.

[9] 张冬雯, 赵 琪, 许云峰, 等. 基于长短期记忆神经网络模型的空气质量预测[J]. 河北科技大学学报, 2020, 41(1): 67-75.

[10] 李晓芳, 尹仔锋. 哈密市空气质量现状分析及对策研究[J]. 干旱环境监测, 2025, 39(1): 15-20.