

基于多模型集成学习的对流云覆盖率预测研究

赵润华

民航西南空管局气象中心, 四川 成都

收稿日期: 2026年8月1日; 录用日期: 2026年8月27日; 发布日期: 2026年9月3日

摘要

基于2016~2024年5~9月ERA5再分析资料的逐小时97个大气物理特征变量数据, 以实况的组合反射率大于35 dBZ的终端区覆盖率为目标变量, 利用随机森林(RandomForest)、极端随机树(ExtraTrees)、梯度提升决策树(GBDT)、XGBoost、CatBoost和LightGBM共6种机器学习模型。采用Pearson相关性分析进行特征筛选, 利用MinMax标准化进行数据预处理, 通过3折交叉验证评估模型泛化能力, 并引入SHAP (SHapley Additive exPlanations)方法对在所测模型中表现最优模型进行全局和局部可解释性分析。结果表明: ExtraTrees模型预测性能表现最好, 测试集决定系数(R^2)为0.1761, 平均绝对误差(MAE)为0.0242, 均方根误差(RMSE)为0.0627, 交叉验证 R^2 为0.1837; 大气可降水量(pw)、降水总量(tp)和850 hPa露点温度(td_850)是预测云覆盖率最重要的三个特征, 模型学习到的特征影响模式与气象学物理机制高度一致; 学习曲线分析表明模型尚未达到性能平台期, 增加训练数据量有望进一步提升预测精度。

关键词

云覆盖率, 机器学习, 回归预测, 集成学习, ExtraTrees, SHAP可解释性

Predictive Research of Convective Cloud Cover Rate Using Multi-Model Ensemble Learning

Runhua Zhao

Meteorological Center of Southwest Regional Air Traffic Management Bureau, CAAC, Chengdu Sichuan

Received: August 1, 2026; accepted: August 27, 2026; published: September 3, 2026

Abstract

Based on hourly atmospheric physical characteristic variables from ERA5 reanalysis data from May to September 2016 to 2024, the target variable was the terminal area coverage rate with a combined

reflectance greater than 35 dBZ. Six machine learning models were used: Random Forest, ExtraTrees, Gradient Boosting Decision Tree (GBDT), XGBoost, CatBoost, and LightGBM. Pearson correlation analysis was employed for feature screening, MinMax normalization for data preprocessing, and 3-fold cross-validation for model generalization assessment. The SHAP (SHapley Additive exPlanations) method was introduced for global and local interpretability analysis of the optimal model. The main conclusions are as follows: The ExtraTrees model achieved the best prediction performance with a test R^2 of 0.1761, MAE of 0.0242, RMSE of 0.0627, and cross-validation R^2 of 0.1837. SHAP feature importance analysis identified precipitable water (pw), total precipitation (tp), and 850 hPa dew-point temperature (td_850) as the three most important predictors of cloud cover rate, with the learned feature impact patterns being highly consistent with meteorological physical mechanisms. Learning curve analysis indicates that the model has not yet reached a performance plateau, suggesting that increasing training data volume could further improve prediction accuracy.

Keywords

Loud Cover Rate, Machine Learning, Regression Prediction, Ensemble Learning, ExtraTrees, SHAP

Copyright © 2026 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

强对流天气对航空运输的安全和效率有重要的影响,业务运行中,通常将组合反射率 35 dBZ 作为一个关键阈值,当在终端区或机场范围内出现一定覆盖率的大于 35 dBZ 的强回波时,往往会导致该空域拥堵、航空器绕飞,产生安全运行风险,是管制运行关注的重点指标。云覆盖率的预测主要依赖于数值天气预报和统计经验模型两大类。由于数值天气预报模式对云微物理过程的参数化存在较大不确定性,且对初始条件极为敏感,对流云团的生成、发展和消散过程存在较大偏差[1]。统计方法如多元线性回归和自回归移动平均模型虽然计算高效,但难以捕捉大气变量之间复杂的非线性关系和交互效应。

近年来,集成学习方法因其能够有效建模高维非线性关系取得了显著进展。随机森林(Random Forest)通过 Bagging 策略和随机特征子空间选择构建多棵决策树,有效降低了模型方差[2]。梯度提升机(Gradient Boosting Machine)通过迭代拟合残差可以有效逐步提升预测精度[3]。在此基础上,Chen 和 Guestrin [4]开发的 XGBoost 框架引入正则化项和列采样机制,在多项数据挖掘竞赛中表现卓越。Ke 等[5]提出的 LightGBM 采用基于直方图的决策树算法和互斥特征捆绑技术,大幅提升了训练效率。Prokhorenkova 等[6]提出的 CatBoost 通过有序提升(Ordered Boosting)算法有效处理类别特征,减少了梯度偏差。Geurts 等[7]提出的极端随机树(ExtraTrees)通过在节点分裂时引入更激进的随机化策略,进一步降低了模型方差。

国内学者在气象要素机器学习预报方面开展了大量研究工作:黄小刚等[8]基于 XGBoost 算法建立了西南地区短临降水预报模型,取得了优于传统方法的预报效果。李晓琳等[9]利用随机森林方法对华东地区能见度进行预测研究,发现模型能够较好地捕捉低能见度事件的主要影响因子。王咏薇等[10]将 LightGBM 应用于地面气温预报,显著提升了预报精度。张志薇等[11]采用 SHAP 方法对降水预报模型进行了可解释性分析,揭示了模型决策的物理机制。现有研究多集中于单一模型对气象要素的预测应用,缺乏对不同集成学习框架的系统性比较和物理可解释性分析。特别是利用高分辨率再分析数据构建多维大气特征空间进行对流云覆盖率回归预测的研究尚不充分。基于上述背景,本研究通过构建 2016~2024 年 5~9 月包含多气压层的热力、动力和水汽等变量的 97 个大气物理特征与双流机场终端区大于 35 dBZ

对流云团覆盖率的高维空间；全面比较随机森林(RandomForest)、极端随机树(ExtraTrees)、梯度提升决策树(GBDT)、XGBoost、CatBoost 和 LightGBM 等 6 种代表性的机器学习模型在对流云覆盖率回归预测任务中的性能差异；引入 SHAP 方法对在所测模型中表现模型进行全局特征重要性分析和局部样本解释，揭示影响对流云覆盖率的关键物理因子及其非线性影响机制；最后通过学习曲线和交叉验证系统评估模型的泛化能力和数据需求特征，为模型的业务化应用提供可靠依据。

2. 数据与方法

2.1. 数据与回归模型

本研究使用的数据集来源于 ERA5 再分析资料，包含 2016~2024 年 5~9 月的 24,706 条高分辨率大气再分析记录，时间间隔为 1 小时。每条记录包含 97 个物理特征变量和 1 个大于 35 dBZ 的对流云覆盖率目标变量(cover_rate_35)。特征变量涵盖了多个气压层(925 hPa, 850 hPa, 700 hPa, 500 hPa, 200 hPa)的热力学变量(温度、露点温度、比湿)、动力学变量(风场分量 u/v、涡度、散度)以及复合指数(K 指数 kx、对流有效位能 cape、可降水量 pw 等)。目标变量 cover_rate_35 数据的分布呈显著正偏态(均值 0.0183、标准差 0.0669、最小值 0.0000、最大值 0.9757)。这一高度偏态的分布特征对回归模型的损失函数选择和预测性能评估提出了特殊挑战。本文采用 MinMax 标准化将特征值缩放到[0, 1]区间，以消除量纲差异对距离敏感模型的影响，数据集按 8:2 比例随机划分为训练集(80%，19,764 条)和测试集(20%，4942 条)，划分过程采用分层抽样策略以保证目标变量分布的一致性。

研究选取 6 种代表性机器学习回归模型：RandomForest、ExtraTrees、GBDT、XGBoost、CatBoost 和 LightGBM 等 6 种代表性的机器学习模型进行对比评估，并对 RandomForest、XGBoost、CatBoost 和 LightGBM 四个模型采用 Optuna 框架结合 TPE (Tree-structured Parzen Estimator)贝叶斯优化算法进行自动化超参数调优，以最大化模型在 5 折验证集上的平均 R^2 得分。

2.2. SHAP 可解释性分析

为深入理解模型的决策机制，引入 SHAP (SHapley Additive exPlanations)方法对在所测模型中表现最优模型进行可解释性分析。SHAP 基于博弈论中的 Shapley 值理论，通过计算各特征对所有可能特征组合边际贡献的加权平均，为每个样本的每个特征分配一个 SHAP 值，量化了该特征对模型输出的正负影响方向和大小[12]。

3. 结果与分析

3.1. 模型性能对比

表 1 展示了 6 种回归模型在测试集上的性能比较及经贝叶斯优化后的超参数。ExtraTrees 模型在所有评估指标上均表现最好，测试集 R^2 为 0.1761，MAE 为 0.0242，RMSE 为 0.0627，交叉验证 R^2 为 0.1837。经超参数优化的 XGBoost 和 RandomForest 分别位列第二和第三，测试集 R^2 分别为 0.1455 和 0.1573。值得注意的是，CatBoost 和 LightGBM 作为梯度提升框架的新成员，同样展现出较强的竞争力，测试集 R^2 分别为 0.1694 和 0.1668，进一步验证了集成学习方法在云覆盖率预测任务中的有效性。

从过拟合程度分析(图 1)，ExtraTrees 的训练集 R^2 (0.3920)与测试集 R^2 (0.1761)差异较大，表明模型在训练数据上存在一定的过拟合现象。然而，得益于其较强的双重随机化机制(在特征选择和分割阈值两个层面同时引入随机性)，ExtraTrees 的交叉验证 R^2 (0.1837)仍为所有模型中最高的，表明其泛化能力相对最佳。相比之下，GBDT 模型虽然训练集 R^2 达到 0.3003，但测试集 R^2 仅为 0.1243，泛化性能下降最为明显，说明其在处理云覆盖率这类高度偏态分布的目标变量时更容易过拟合。

Table 1. Performance comparison and optimal hyperparameters of six regression models
表 1. 6 种回归模型性能对比及超参数

Model	R ² _train	R ² _test	MAE_test	RMSE_test	R ² _CV	最佳超参数
ExtraTrees	0.3920	0.1761	0.0242	0.0627	0.1837	n_estimators = 300, max_depth = 25
CatBoost	0.3210	0.1694	0.0243	0.0630	0.1594	iterations = 500, depth = 8
LightGBM	0.3299	0.1668	0.0239	0.0631	0.1671	n_estimators = 400, num_leaves = 31
RandomForest	0.2532	0.1573	0.0242	0.0634	0.1573	n_estimators = 250, max_depth = 20
XGBoost	0.1963	0.1455	0.0244	0.0639	0.1402	n_estimators = 350, max_depth = 6
GBDT	0.3003	0.1243	0.0248	0.0646	0.1200	n_estimators = 200, max_depth = 5

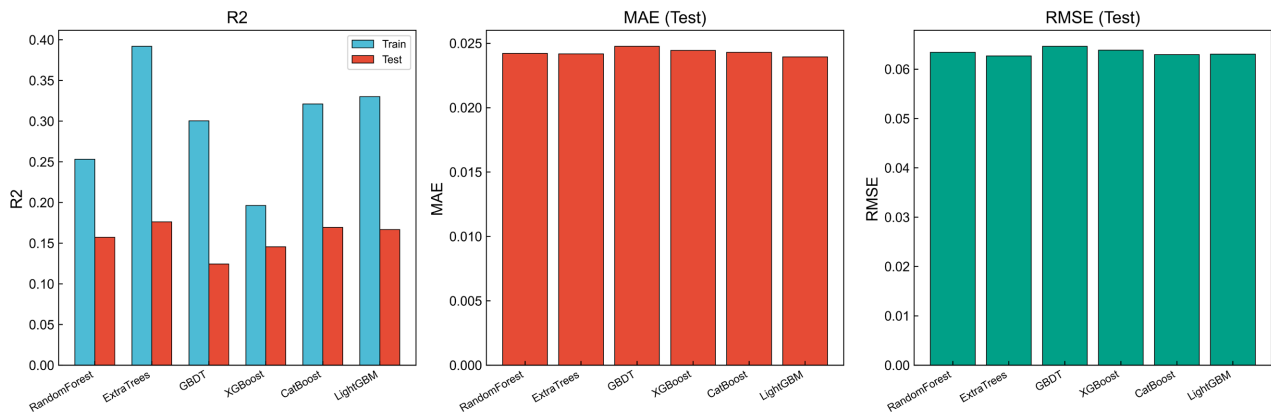


Figure 1. Comparison of training R², test MAE and RMSE among six models
图 1. 6 种模型训练集 R²、测试集 MAE 和 RMSE 对比

3.2. 预测值与残差分析

图 2 展示了各模型预测值与观测值的散点分布(含 1:1 理想线和 95%置信区间)。从散点图可以观察到,所有模型在低云覆盖率区间(0~0.1)的预测较为集中,反映了训练数据中低云覆盖样本占主导地位的数据分布特征。在高云覆盖率区间(>0.3),各模型的预测误差显著增大,尤其是 ExtraTrees 和 XGBoost 表现出一定程度的系统性低估倾向,即模型倾向于预测较低的云覆盖率值,难以准确捕捉极端高云覆盖事件。

图 3 为残差分析图(含 LOWESS 平滑趋势线),展示了各模型残差与预测值之间的关系。残差分布呈现明显的异方差性——在低预测值区间残差较为集中,随着预测值增大残差散点迅速扩散,呈现出典型的喇叭形分布。这一模式表明,云覆盖率预测的难度随云量增加而显著增大,模型的不确定性在高云量条件下急剧上升。

3.3. 交叉验证稳定性分析

图 4 的 3 折交叉验证 MAE 箱线图揭示了各模型泛化性能的稳定性差异。ExtraTrees 的交叉验证 MAE 中位数最低且箱体最为紧凑,表明其预测精度高且波动小,泛化稳定性最佳。LightGBM 和 CatBoost 的交叉验证表现同样较为稳定,箱体高度较窄,抖动散点分布集中。相比之下,GBDT 的交叉验证箱体最宽,且存在明显的离群点,表明其泛化性能不够稳定,对训练集的具体划分方式较为敏感。这一结果进一步印证了在模型选择中应优先考虑交叉验证性能而非仅关注单次训练-测试划分的指标。

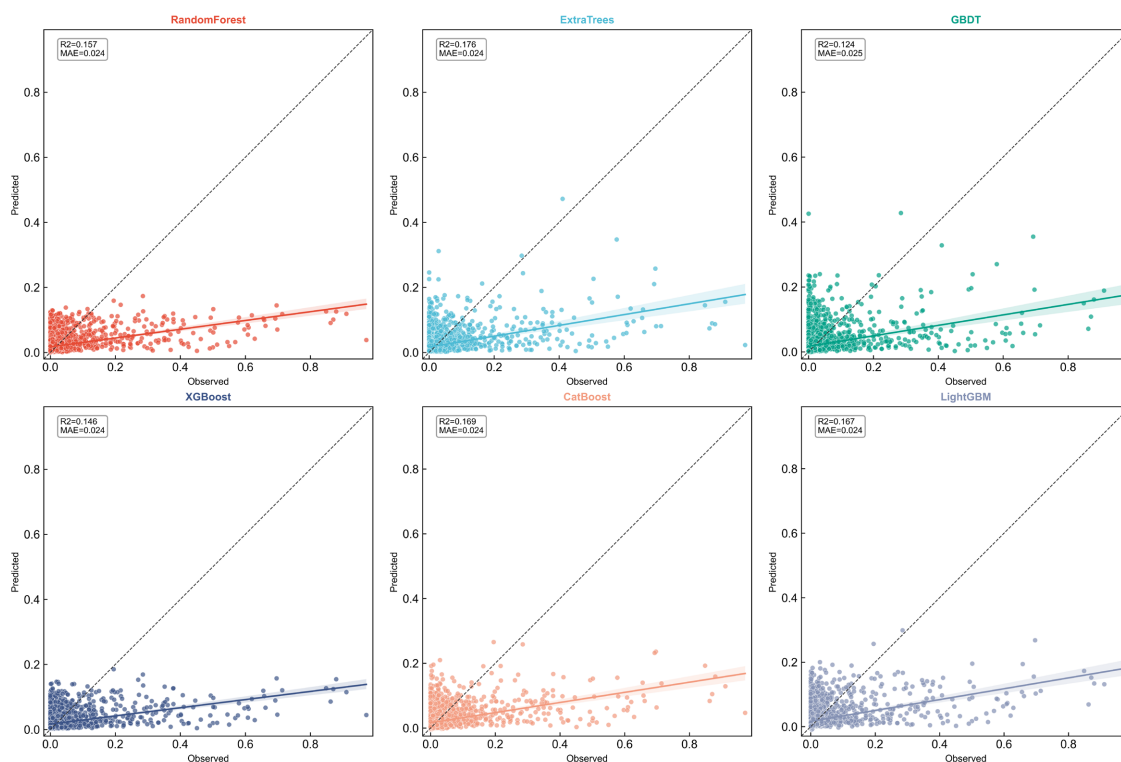


Figure 2. Scatter plots of predicted vs. observed cloud cover rate for six models
图 2. 6 种模型预测值与观测值散点图

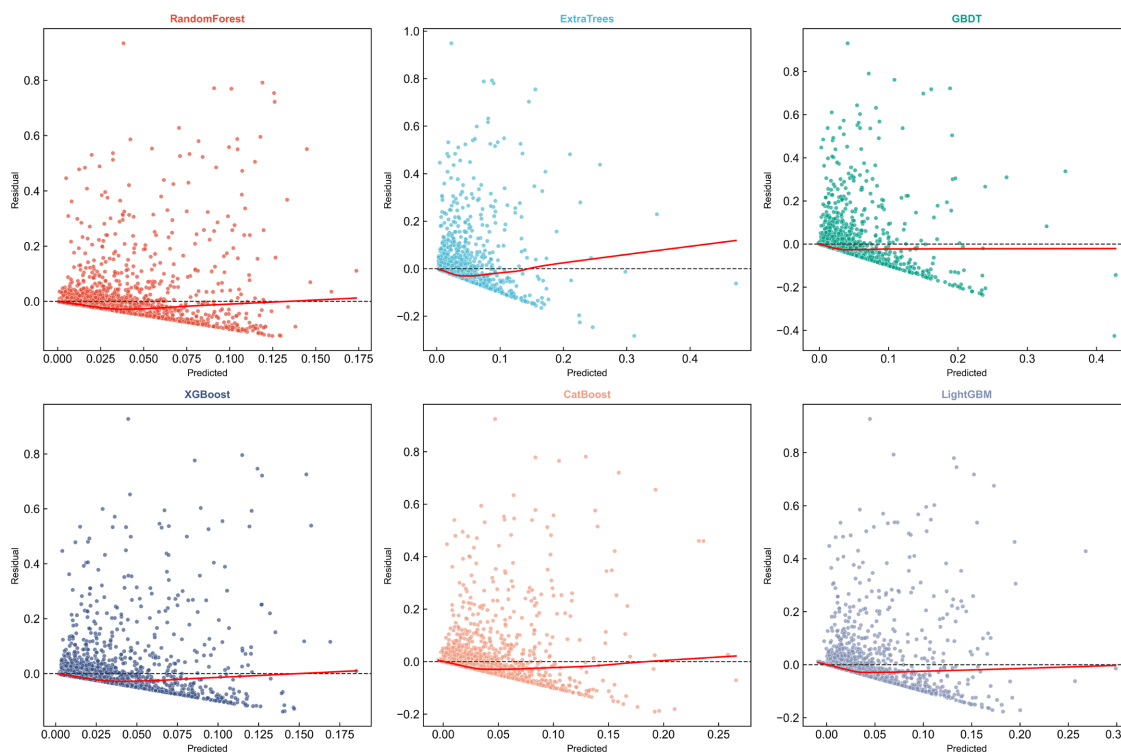


Figure 3. Residuals vs. predicted values with LOWESS smoothing for six models
图 3. 6 种模型残差与预测值散点图(含 LOWESS 平滑线)

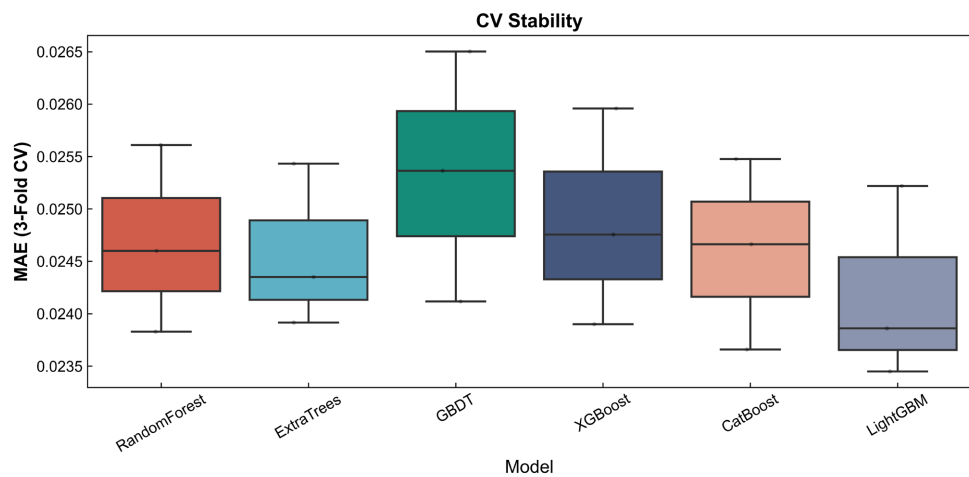


Figure 4. Boxplot of 3-fold CV MAE for six models
图 4. 3 折交叉验证 MAE 箱线图

3.4. 特征重要性分析

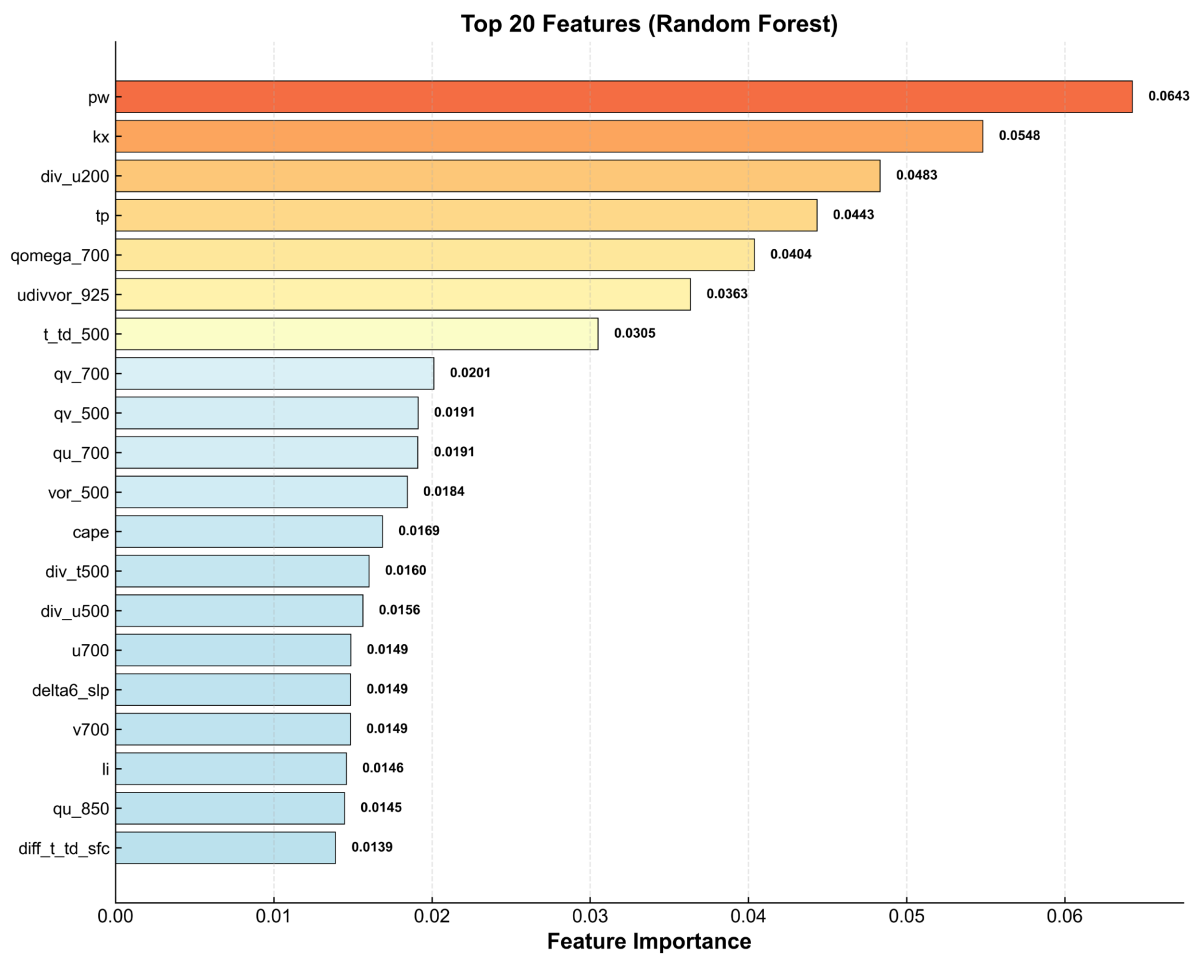


Figure 5. Top 20 feature importance ranking based on random forest
图 5. 随机森林 Top 20 特征重要性排名

图 5 展示了基于随机森林模型计算的特征重要性排名(Top 20)。最重要的前 5 个特征分别为: pw(大气可降水量, 0.0643)、kx(K 指数, 0.0548)、div_u200(200 hPa 散度, 0.0483)、tp(降水总量, 0.0443)和 qomega_700(700 hPa 准地转垂直速度, 0.0404)。

从物理机制角度分析, 重要性排名靠前的特征涵盖了云形成的关键条件: (1) 充足的水汽供应 pw(可降水量)的重要性显著高于其他特征, 这是因为云的形成和维持与大气中的水汽含量直接相关, 水汽是成云致雨的物质基础[13]; (2) 大气不稳定层结 K 指数作为表征大气不稳定度的综合指标, 其高重要性反映了层结不稳定对云发展的促进作用; (3) 动力强迫机制 200 hPa 散度反映了高层大气辐散对垂直上升运动的强迫效应, 是影响云覆盖的重要动力因子。此外, 降水总量(tp)的高重要性体现了云-降水过程的紧密耦合关系。这一特征重要性结构与气象学中云与降水形成的经典理论框架高度一致, 验证了机器学习模型在云物理应用中的可解释性。

3.5. SHAP 可解释性分析

3.5.1. 全局特征重要性

图 6 的 SHAP 蜂群图揭示了比传统特征重要性更为细致的特征影响模式。SHAP 重要性 Top 5 为: pw(可降水量, SHAP 均值 = 0.0043)、tp(降水总量, 0.0022)、td_850(850 hPa 露点温度, 0.0022)、q_850(850 hPa 比湿, 0.0018)和 q_925(925 hPa 比湿, 0.0016)。可降水量(pw)以绝对优势位居第一, 其高值对应正 SHAP 值, 推动预测值升高, 这与水汽充足是云形成必要条件的物理直觉完全一致。降水总量(tp)和 850 hPa 露点温度(td_850)紧随其后, 分别反映了水汽凝结输出和低层水汽饱和度对云覆盖的重要影响。值得注意的是, SHAP 重要性排名与传统随机森林特征重要性排名存在一定差异——SHAP 方法更能反映特征在模型决策中的实际贡献, 而传统特征重要性可能受特征间相关性的影响[14]。

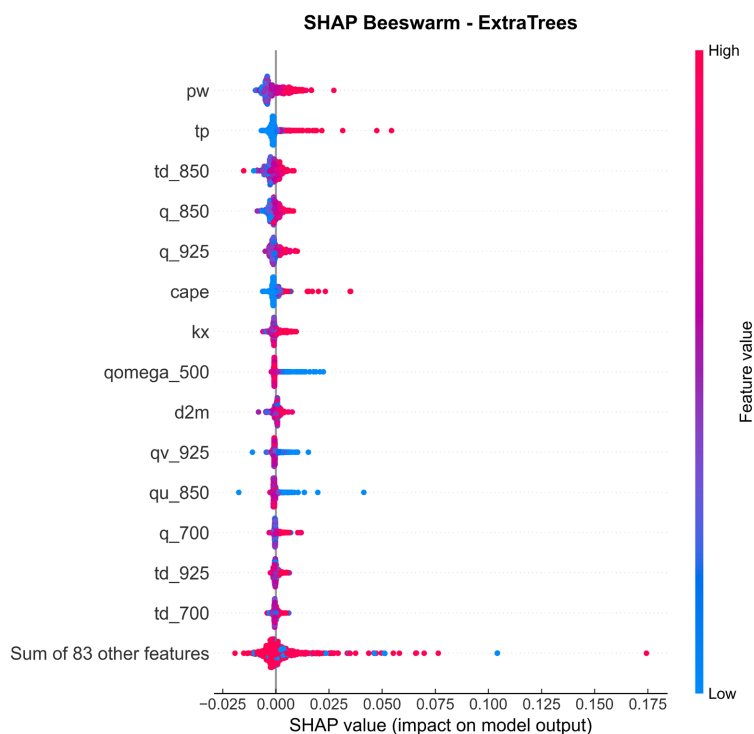


Figure 6. SHAP beeswarm plot showing global feature impact direction and magnitude
图 6. SHAP 蜂群图——全局特征影响方向与大小

3.5.2. 训练 - 测试分布一致性

图 7 对比了训练集和测试集的 SHAP 重要性排名。两者排序高度一致——pw 在两个数据集上均为最重要特征，tp、td_850、q_850、q_925 的相对排序也基本保持一致。这种高度一致性表明模型在训练和测试集上依赖相同的特征模式进行预测，未出现因数据泄露或分布偏移导致的特征重要性反转现象，进一步验证了模型在未见数据上泛化的可靠性。

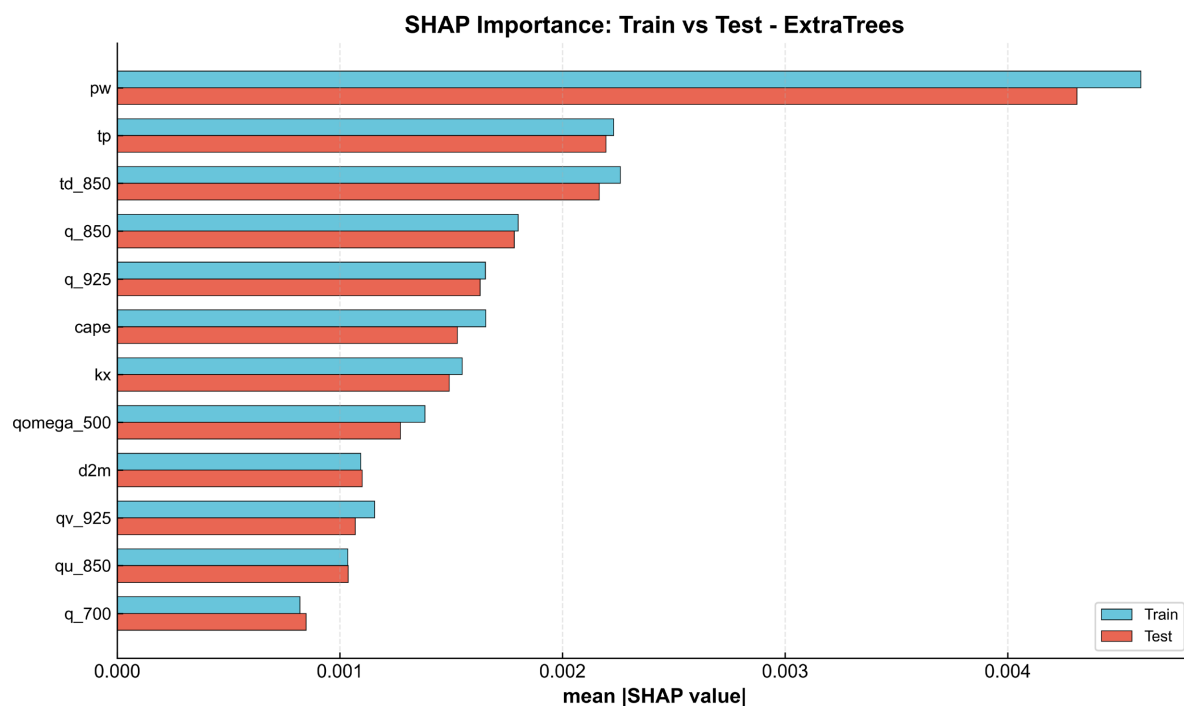


Figure 7. SHAP importance comparison between training and test sets

图 7. SHAP 重要性：训练集与测试集对比

4. 结论与讨论

本研究系统评估了 6 种机器学习模型在对流云覆盖率回归预测任务中的性能，并通过 SHAP 方法进行了深度可解释性分析。结果表明：

(1) 6 种模型中，ExtraTrees 表现在所测模型中表现最优(测试集 $R^2 = 0.1761$ ， $MAE = 0.0242$ ， $RMSE = 0.0627$)，XGBoost 和 CatBoost 紧随其后。ExtraTrees 模型的表现可归因于其独特的双重随机化机制，在特征选择和阈值选择两个层面同时引入随机性，有效抑制了方差，使其在面对高度偏态分布的目标变量时展现出更强的鲁棒性。

(2) SHAP 分析进一步揭示，模型学习到的特征影响模式与气象学物理机制高度一致。水汽供应(pw, q_925, q_850)、热力条件(td_850, d2m)和动力条件(qomega_500, div_u200)构成了预测的核心因子体系，这些特征分别对应云形成的三大必要条件：充足的水汽、不稳定层结和抬升触发机制。验证了机器学习方法在大气科学应用中的物理合理性。

(3) 学习曲线分析表明模型尚未达到性能平台期，扩大训练数据规模和引入时序/空间特征有望进一步改善预测精度。目标变量的高度偏态分布对标准回归方法提出了挑战，需探索面向不平衡分布数据的专门建模策略。

(4) 机器学习模型在业务运行中存在一定的局限性,主要体现在: R^2 值整体偏低(最优模型仅 0.176),反映了云覆盖率预测的固有困难性;目标变量的高度偏态分布使得标准回归损失函数对高值样本的预测不足;极端样本的特征分布与训练集主流分布偏离较大时,模型决策趋于不稳定。

本研究聚焦于对机器学习方法在云覆盖预测中的探索和对问题难度的量化评估。为基于机器学习的气象要素预测提供了方法论框架和基准性能参考,对数据驱动的云物理参数化方案改进具有一定指导意义。

参考文献

- [1] Boukabara, S.A., Krasnopolsky, V., Penny, S.G., Stewart, J.Q., McGovern, A., Hall, D., et al. (2021) Outlook for Exploiting Artificial Intelligence in the Earth and Environmental Sciences. *Bulletin of the American Meteorological Society*, **102**, E1016-E1032. <https://doi.org/10.1175/bams-d-20-0031.1>
- [2] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. <https://doi.org/10.1023/a:1010933404324>
- [3] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- [4] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>
- [5] Ke, G.L., Meng, Q., Finley, T., et al. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: *Advances in Neural Information Processing Systems* 30, Curran Associates, 3146-3154.
- [6] Prokhorenkova, L., Gusev, G., Vorobev, A., et al. (2018) CatBoost: Unbiased Boosting with Categorical Features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates, 6638-6648.
- [7] Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely Randomized Trees. *Machine Learning*, **63**, 3-42. <https://doi.org/10.1007/s10994-006-6226-1>
- [8] 黄小刚, 陈明轩, 秦睿, 等. 基于 XGBoost 的西南地区短临降水预报方法研究[J]. *气象学报*, 2020, 78(4): 666-682.
- [9] 李晓琳, 牛生杰, 吕晶晶, 等. 基于随机森林的华东地区能见度预报方法[J]. *大气科学学报*, 2021, 44(2): 331-342.
- [10] 王咏薇, 朱磊, 张录军. 基于 LightGBM 的地面气温预报方法研究[J]. *气象*, 2022, 48(5): 612-624.
- [11] 张志薇, 况雪源, 李栋梁. 基于 SHAP 方法的降水预报模型可解释性分析[J]. *高原气象*, 2023, 42(3): 729-742.
- [12] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* 30, Curran Associates, 4765-4774.
- [13] 程正泉, 陈联寿, 李英, 等. 中国近海热带气旋云和降水结构特征的 TRMM 卫星观测分析[J]. *热带气象学报*, 2017, 23(4): 323-332.
- [14] 赵增月, 管兆勇, 王慧. 基于 SHAP 的机器学习模型可解释性方法在气象中的应用[J]. *气象科学*, 2022, 42(2): 129-139.