许立兵,孔扬,周峥,等. 基于机器学习的风场预报订正方法研究[J]. 陕西气象,2023(1):15-20.
文章编号:1006-4354(2023)01-0015-06

基于机器学习的风场预报订正方法研究

许立兵[1,2],孔　扬[3,4],周　峥[1,2,5],王安喜[1,2],梁逸爽[1,2,5]
(1. 国家超级计算无锡中心,江苏无锡　214011;2. 无锡九方科技有限公司,江苏无锡　214011;
3. 宁波市气象局,浙江宁波　315012;4. 宁波市气象灾害应急预警中心,浙江宁波　315012;
5. 清华大学地球系统科学系,北京　100084)

摘　要:为了实现更准确的站点风预报,结合中尺度数值模式 WRF 预报结果和自动气象站观测数据,采用反距离加权插值法,将模式网格和观测站点的数据进行融合构建训练集,利用 3 种机器学习方法对 WRF 预报的风场结果进行订正,优化风场预报准确率。其中随机森林模型实现风速的预报均方根误差(RMSE)平均降低了 18.22%,风向降低了 15.97%;LightGBM 模型对于风速、风向的 RMSE 平均降低了 18.60%和 17.56%;深度神经网络模型对于风速、风向的 RMSE 平均降低了 5.53%和 9.10%。对 2020 年宁波市 9 个大风过程进行检验,利用 LightGBM 模型对于 3 个站点预报进行订正,结果表明风速的 RMSE 从 4.61 m/s 下降到 2.14 m/s,平均降低了 53.58%,风向的 RMSE 从 30.31°下降到 18.20°,平均降低了 39.95%。

关键词:WRF 模式;机器学习;模型解释;大风检验
中图分类号:P457.5　　**文献标识码**:A

传统的大气和海洋预报方法主要依赖数值模式,该方法旨在利用超级计算机,通过求解动力方程,结合初始场信息,对未来的状态进行预报。随着相关理论的不断发展、观测系统的完善以及计算机运算能力的不断提升,预报水平取得了较大的提高。尽管如此,数值天气预报数据与真实的天气情况依旧存在一定的差别,这种误差主要产生于初始场建立时的误差及模型的模式误差[1]。可通过改进模式分辨率和模型参数化方案减少预报模式误差,但会增加很大的计算量,同时也很难消除估计地表气象参数时的初始化或系统误差[2]。近年来,针对数值预报结果开展了大量的后处理工作,形成了多种数值预报订正技术。(1)统计订正方法,如基于多元线性回归的模式输出统计订正方法拟合降尺度因子与降水量分布之间的关系实现对降水的预报[3];相似集合预报方法实现对风和温度的预报[4];卡尔曼滤波方法结合数值预报产品,实现气温的预报[5]。(2)机器学习方法,如基于支持向量机实现对模式预报风的订正[6];人工神经网络方法能够对非线性系统进行建模使得降水预报的准确率大大提高[7],将机器学习技术与传统模式预报技术相结合以提升预报准确度的可行性也得到了证实[8]。芦华等[9]利用站点的观测数据、WRF(weather research forecast)模式和 CMAQ 模式(community multiscale air quality model)的预报数据,实现对成渝区域 $PM_{2.5}$ 的精细化预报,利用随机森林传统的机器学习回归模型和 RNN-LSTM 深度网络回归模型对

收稿日期:2021-10-25
作者简介:许立兵(1988—),男,汉族,江苏无锡人,硕士,工程师,主要从事大数据处理、机器学习的研究。
通信作者:孔扬(1988—),男,汉族,浙江宁波人,硕士,高级工程师,主要从事气象预报服务、大数据处理研究。
基金项目:国家重点研发计划项目(2018YFB0505000);宁波市"科技创新 2025"重大专项(2019B10025);宁波市气象科技计划项目(NBQX2020003B)

$PM_{2.5}$进行预报订正，实验结果表明，订正后的均方根误差(RMSE)减小 70%左右，相关系数能够达到 90%左右。任萍等[10]研发了一套基于机器学习方法 XGBoost 且考虑地形特征影响的数值预报多模式集成技术，预报系统可提供准确性更高的多模式集成确定性预报产品，其中对 2 m 温度集成的误差可降低 11.02%～18.09%，10 m 风速集成误差可降低 31.23%～33.22%，10 m 风向集成误差可降低 4.1%～8.23%。

本文利用在宁波市气象局业务预报中实际应用的 8 km 分辨率 WRF 模式的预报数据，采用机器学习方法对模式的预报风场进行订正，实现对宁波舟山港 3 个港区风场更精准的预报。对比随机森林、LightGBM 与深度神经网络 3 种方法在订正预报结果所取得的不同效果，通过 LightGBM 模型分析得到不同输入特征对预报要素的影响程度，分析关键的预报订正因子。本方案不仅可以提高预报的精度，而且可以保留数值模式预报的物理参数化方案[11-13]。更精准的站点风场预报，可服务于站点区域的港口调度、防灾减灾决策、码头生产作业等。

1 数据来源

1.1 站点数据

利用 2015—2020 年宁波市气象局提供的北仑山 B(121.854°E,29.938°N)、梅山(122.001°E,29.766°N)、远东(122.069°E,29.886°N)3 个站点的逐小时风速、风向观测数据，订正 WRF 模式的风场预报数据。

1.2 模式数据

WRF 是集数值天气预报、大气模拟及数据同化于一体的模式预报系统，能够对中尺度天气进行模拟和预报[14]。模式预报数据来自宁波市气象局的业务化预报数据，采用三维变分进行同化。该数据集在每个时刻预报 19 个要素，包括：纬向风、经向风、温度、能见度等，空间分辨率为 8 km。使用的数据集的时间范围是 2015—2020 年，经纬度范围为：102.839° E～140.153° E、15.126 1°N～45.874 1°N，时间分辨率为 1 h，预报时长为 78 h。

1.3 数据融合

采用反距离加权插值的方式，将模式的网格点数据插值到观测站点。按照预报日数，依据时间分辨率对数据进行划分，将模式预报的纬向风速 U 和经向风速 V 转化为站点的风向、风速，每一个要素，按照超前预报日数训练一个模型，共计 8(2×4)个。2015—2019 年数据用于构建机器学习算法模型，采用随机抽取的方式选择训练集与测试集，其中 80%划分为训练集，20%划分为测试集；2020 年的数据用于检验模型对于大风订正的效果。

1.4 相关性分析

WRF 模式输出的预报因子共计 19 个，其中在气压层(925、850、700 hPa)输出相对湿度、温度。因订正的是 10 m 的风，选择近地面 925 hPa 气压层的的要素作为输入特征，选择与风相关的要素包括：站点观测的风向、风速，模式预报的风向、风速、2 m 相对湿度、近地面层相对湿度、能见度、2 m 温度、近地面层温度。采用 Person 相关分析法，计算每个要素与其他要素的相关系数，结果发现：与站点观测风速最相关的前 3 个要素为站点观测的风向，模式预报的风向、风速，相关系数分别为 0.27、0.24、0.60；与站点观测风向最相关的前 3 个要素为站点观测的风速，模式预报的风向、风速，相关系数分别为 0.27、0.26、0.16。

2 方法与模型

模式的预报数据插值到对应站点后的数据作为机器学习算法的输入值，站点的观测作为真值，利用随机森林(Ramdom Forest)、LightGBM 及深度神经网络(DNN)3 种机器学习方法对模式的预报结果进行订正。

2.1 随机森林

随机森林是一种高度灵活的机器学习算法，采用集成分类器的形式进行分类。随机森林分类器是由很多决策树分类模型组成的集成分类器模型，分类器对参数不敏感，不易过拟合，训练速度快[15]。比较适合多分类以及回归问题，拥有广泛的应用前景，对于大部分的数据，它的拟合效果比较好。算法示意如图 1 所示，构建不同决策树的示意过程。

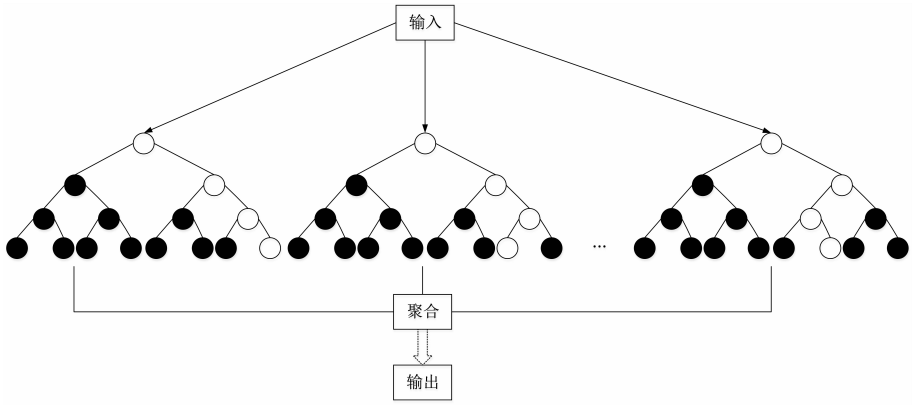

图 1　随机森林结构示意图(图中实心和空心的圈表示选择不同的特征)

随机森林采用 scikit-learn 库实现,版本为 0.23.2,决策树的个数范围是[20,50],步长为 5;树的深度范围是[10,30],步长为 2,数据划分的随机数种子为 50,其他参数采用系统自带的默认值。对树的个数和深度进行循环遍历,选择测试集上均方根误差(RMSE)为最小值的个数及深度。

2.2　LightGBM

LightGBM 是一个基于决策树型的、分布式、高效的梯度提升框架,每次从当前的所有叶子中,找到分裂增益最大的一个叶子作为分裂节点,然后按照此策略,不断进行分裂[16-17]。同按“层生长”策略相比,在设定相同分裂次数参数的条件下,按“叶子生长”策略,可以有效降低误差,得到更好的预测效果。算法如图 2 所示,按“叶子生长”策略主要缺点是可能会得到比较深的决策树,容易造成过拟合,在实际业务应用中,LightGBM 在对树的最大深度进行了限制,保证高效率的同时,也可以有效防止树的深度过深、带来模型的过拟合问题[18]。本实验中使用的 LightGBM 的版本为 2.2.2,数据划分的随机数种子为 50,提升类型设置为 gbdt,学习速率设置为 0.001,叶子节点个数设置为 512,最大循环迭代次数为 10 000,其他采用系统的默认参数。

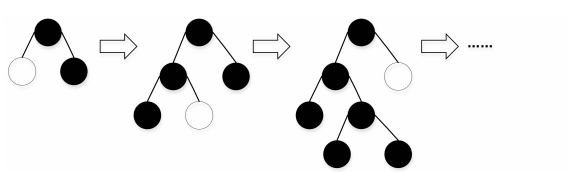

图 2　LightGBM 算法示意图

2.3　深度神经网络

深度神经网络技术在 2012 年后得到了快速发展,在多种任务中,表现出良好的效果[19-20]。针对风向、风速两个气象要素及超前预报的时长分别设计全连接的深度神经网络(DNN),共计 8(2×4)个。其中,输入层包含 7 个节点,模型的输出层为一个节点,网络模型结构如图 3 所示。

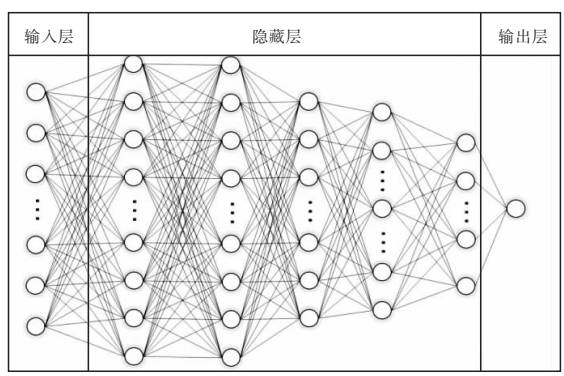

图 3　DNN 模型结构示意图

其中,每个隐藏层的神经元个数分别为 32、64、128、256、32,输入层个数与输入数据的特征相同为 7 个。数据划分的随机数种子为 50,隐藏层使用 ReLU 作为激活函数,学习率设为 0.001,损失函数采用均根方误差,采用自适应矩估计算法进行模型参数的权重优化。在训练过程中,批量放入模型的样本个数为 128,样本迭代次数设置为 20 轮。训练结束的条件为:当验证集的 RMSE 在 10 轮中未发生降低或者满足迭代的最大次数为 10 000[21]。

3　结果分析

对风速、风向两个预报要素,依据时间分辨率

对数据进行划分，对于 3 个站点在测试集上的1～4 d 的平均 RMSE 如表 1 所示。其中原始 RMSE 指的是 WRF 模式插值到对应站点的值与站点观测值之间的均方根方误差(用 $E_{\text{RMS模式}}$ 表示)，机器学习的 RMSE 指的是机器学习模型的预测值与站点观测值之间的均方根方误差(用 $E_{\text{RMS机器学习}}$ 表示)。对于订正的 2 个预报要素的 RMSE 降幅统计，采用公式(1)进行计算：

$$\Delta E_{\text{RMS}}=((E_{\text{RMS模式}}-E_{\text{RMS机器学习}})/E_{\text{RMS模式}})\times 100\%。\quad(1)$$

表 1 WRF 模式和 3 种订正模型对宁波市 3 个站 2015—2019 年测试集上风速、风向预报的平均均方根误差表

站点	WRF 模式		Ramdom Forest		LightGBM		DNN	
	风速/(m/s)	风向/(°)	风速/(m/s)	风向/(°)	风速/(m/s)	风向/(°)	风速/(m/s)	风向/(°)
北仑山 B	2.93	65.59	2.47	54.23	2.46	53.12	2.74	59.83
梅山	2.16	59.92	1.70	51.89	1.67	50.79	1.87	55.30
远东	2.65	73.28	2.17	60.91	2.16	59.97	2.48	65.56
均值	2.58	66.26	2.11	55.68	2.10	54.63	2.36	60.23

测试集上每个站点的 1～4 d 的 RMSE 情况如图 4 所示。从图 4 可以看出，随着预报时长增加，误差也呈现上升趋势，而加入了 3 种订正算法的 RMSE 远低于 WRF 模式，LightGBM 模型订正效果最好，随机森林次之，WRF 最差，深度神经网络略优于 WRF，可以看出，3 种订正算法可以满足实际业务的稳定性需求。

LightGBM 模型对风速、风向预报的 RMSE 较 WRF 模式平均降低了 18.60%和 17.56%，随机森林模型平均降低了 18.22%和 15.97%。两种机器学习算法模型都是基于决策树策略的机器学习模型，达到相同的预测效果；但 LightGBM 使用了基于直方图做差加速的决策树算法并采用了带深度限制的“叶子生长”策略，只需要随机森林五分之一左右的时间，节省了模型训练时间[22]。基于深度神经网络订正结果的风速、风向的 RMSE 平均降低了 5.53%和 9.10%。本实验中基于决策树型的机器学习算法对小样本数据的适应性更强，预报效果表现得更好，深度神经网络对数据量比较敏感，本文中进行实验的数据量较少，未取得良好的预报效果。

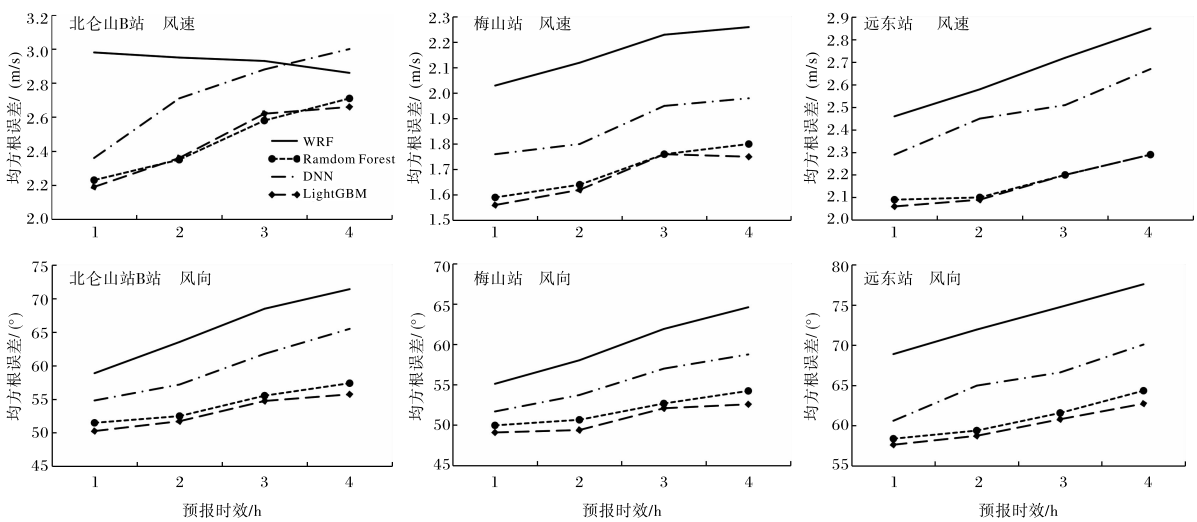

图 4 WRF 模式和 3 种订正模型对宁波市 3 个站 2015—2019 年测试集上的风速、风向预报的均方根误差

4　大风检验

数值模式预报在极端天气现象的模拟效果上仍然存在不足。对于超过 8 级的大风天气,WRF 模式预报结果与实际观测值之间的均方根误差远超模式预报的平均水平。因此选取 3 个站点 2020 年发生的 9 次大风天气过程(风力≥8 级或风速为 7.2～20.4 m/s)进行检验,仅利用订正效果最佳的 LightGBM 模型进行检验。从图 5 可以看出,加入订正模型后,可以有效降低预报的 RMSE。对于北仑山 B 站,订正模型将风速的 RMSE 从超过 7 m/s 降低到不到 3 m/s。还可以看出,3 个站点订模型风速的 RMSE 均在 2 m/s 左右,说明订正模型的稳定性较好。

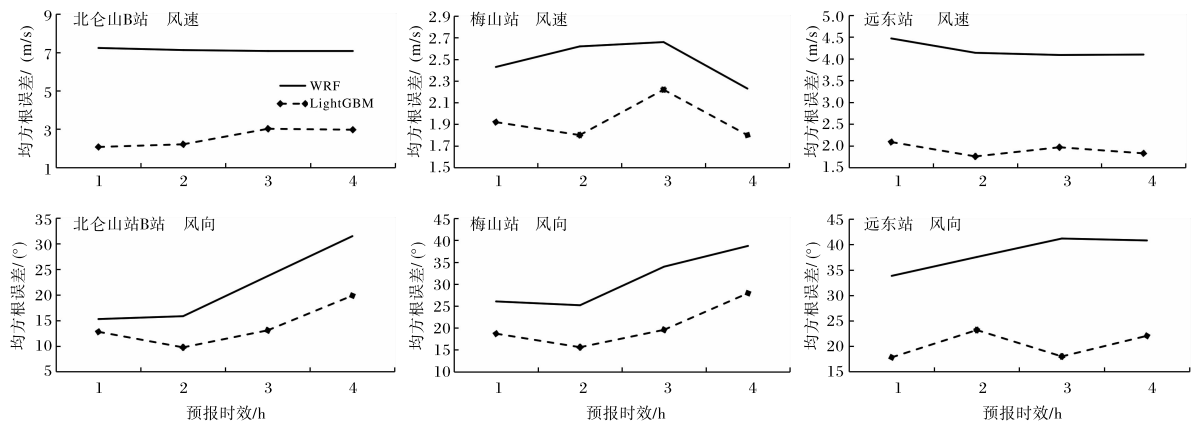

图 5　2020 年 WRF 模式和 LightGBM 模型对宁波 3 个站大风风速、风向预报的均方根误差

对 2020 年 3 个台站 9 个大风个例的具体数值(表 2)进行分析,经 LightGBM 订正后,3 个站点风速均方根误差的平均值从 4.61 m/s 下降到 2.14 m/s,降幅达到 53.58%。风向均方根误差的平均值从 30.31° 下降到 18.20°,降幅为 39.95%。由此可见,利用机器学习订正模型,能够有效弥补数值模式对于极端天气的模拟的不足,更好地为业务部门的预报提供支撑。

表 2　2020 年 WRF 模式和 LightGBM 模型对宁波 3 个站大风检验平均均方根误差

站点	WRF 模式		LightGBM	
	风速/(m/s)	风向/(°)	风速/(m/s)	风向/(°)
北仑山 B	7.14	21.58	2.57	13.88
梅山	2.48	31.00	1.94	20.46
远东	4.20	38.36	1.91	20.25
均值	4.61	30.31	2.14	18.20

5　结论与讨论

(1)本实验设计了基于人工智能的预报订正方案,通过融合站点和模式的预报数据,实现对风向、风速的预报数据订正。

(2)随机森林及 LightGBM 模型的订正效果基本接近,风速均方根误差较 WRF 模式平均降低了 18%左右,风向平均降低了 16%左右;深度神经网络的风速、风向均方根误差平均降低了 5.53%和 9.10%。

(3)对利用 2020 年 9 个大风个例的订正检验结果表明,LightGBM 订正的风速、风向均方根误差较 WRF 模式降低了 53.58%和 39.95%,表明机器学习订正模型能够有效弥补数值模式对大风预报的不准确问题。

(4)本实验的工作仅针对宁波区域的 WRF 预报结果开展相关的研究工作,此方法在其他地区的订正效果有待于进一步开展研究。

参考文献:

[1] ZHANG K, MU M, WANG Q. Identifying the sensitive area in adaptive observation for predicting the upstream Kuroshio transport variation in a 3-D ocean model[J]. Science China Earth Sciences,

2017，60(5)：866-875.

[2] 丛智慧，刘倩，刘永前，等．基于敏感性因素分析的数值天气预报修正方法[J]．分布式能源，2020，5(5)：8-15.

[3] 龚奏，曹宗元，刘菡，等．一种利用浮标站资料改进海浪模式有效波高预报的方法[J]．海洋预报，2020，37(1)：50-54.

[4] DELLE M，ECKEL F，RIFE D，et al. Probabilistic weather prediction with an analog ensemble[J]. Mon Weather Review，2013，141(10)：3498-3516.

[5] 蔡凝昊，俞剑蔚．基于数值模式误差分析的气温预报方法[J]．大气科学学报，2019，42(6)：864-873.

[6] 钱斌凯，何彩芬，金炜，等．基于支持向量机的多因子风速预测[J]．宁波大学学报(理工版)，2018，31(3)：14-19.

[7] 方巍，庞林，王楠，等．人工智能在短临降水预报中应用研究综述[J]．南京信息工程大学学报(自然科学版)，2020 ,12(4)：406-420.

[8] HWANG J，ORENSTEIN P，COHEN J，et al. Improving subseasonal forecasting in the western U S with machine learing[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York：ACM，2019：2325-2335.

[9] 芦华，谢旻，吴钰，等．基于机器学习的成渝地区空气质量数值预报 $PM_{2.5}$ 订正方法研究[J]．环境科学学报，2020，40(12)：4419-4431.

[10] 任萍，陈明轩，曹伟华，等．基于机器学习的复杂地形下短期数值天气预报误差分析与订正[J]．气象学报，2020，78(6)：1002-1020.

[11] 李佰平，智协飞．ECMWF 模式地面气温预报的四种误差订正方法的比较研究[J]．气象，2012，38(8)：897-902.

[12] LIN Y，KRUGER U，ZHANG J，et al. Seasonal analysis and prediction of wind energy using random forests and ARX model structures[J]. IEEE Transactions on Control Systems Technology，2015，23(5)：1994-2002.

[13] 曹渝昆，朱萌．基于主成分分析和 LightGBM 的风电场发电功率超短期预测[J]．上海电力学院学报，2019，35(6)：562-566.

[14] 王晓君，马浩．新一代中尺度预报模式(WRF)国内应用进展[J]．地球科学进展，2011，26(11)：1191-1199.

[15] 李垒，任越美．基于随机森林的高光谱遥感图像分类[J]．计算机工程与应用，2016，52(24)：189-193.

[16] CAI J，LI X，TAN Z，et al. An assembly-level-neutronic calculation method based on LightGBM algorithm[J]. Annals of Nucear Energy，2021，150.

[17] 刘海青，李智桥，李元诚．基于 C-lightGBM 的用户窃电检测[J]．计算机应用研究，2020，37(增刊)：298-300.

[18] 王志宇．基于 LightGBM 框架的上海市大气能见度预报订正研究[D]．上海：华东师范大学，2019.

[19] HAM Y，KIM J，LUO J. Deep learning for multi-year ENSOforecasts[J]. Nature，2019，573，568-572.

[20] HOSSAIN M，REKABDAR B，LOUIS S，et al. Forecasting the weather of Nevada：A deep learning approach[C]//2015 international joint conference on neural networks. New York：IEEE，2015：1-6.

[21] 许立兵，王安喜，汪纯阳，等．基于机器学习的海洋环境预报订正方法研究[J]．海洋通报，2020，39(6)：695-704.

[22] 陈维刚，张会林．基于 RF-LightGBM 算法在风机叶片开裂故障预测中的应用[J]．电子测量技术，2020，43(1)：162-168.