

GPR、XGBoost 和 CatBoost 模拟江西地区 参考作物蒸散量的适应性研究

刘小强^{1,2}, 代智光¹, 吴立峰^{1*}, 张富仓², 董建华³, 陈志月⁴

(1.南昌工程学院 水利与生态工程学院, 南昌 330099;

2.西北农林科技大学 旱区农业水土工程教育部重点实验室, 陕西 杨凌 712100;

3.昆明理工大学 农业与食品学院, 昆明 650500; 4.河海大学 水文水资源学院, 南京 210098)

摘 要:【目的】提高机器学习模型模拟参考作物蒸散量在江西省适应性和精度。【方法】基于江西南昌等 15 个气象站 2001—2015 年日值气象数据(最高气温、最低气温、地表辐射、大气顶层辐射、相对湿度和 2 m 高风速), 以 FAO-56 Penman-Monteith(P-M)公式的计算结果作为对照, 建立了计算 ET_0 的高斯过程回归(GPR)、极限梯度提升(XGBoost)和梯度提升决策树(CatBoost)模型, 并分别与经验模型进行比较。【结果】各气象参数对机器学习模型模拟 ET_0 的精度影响由大到小依次为: R_s 、 T_{max} 和 T_{min} 、 RH 、 U_2 , 且采用 T_{max} 、 T_{min} 、 R_s 和 RH 气象参数组合的机器学习模型($RMSE < 0.2$ mm/d)模拟 ET_0 精度高。此外, 3 种机器学习模型在有限的气象数据时具有较好的适用性, 且优于传统经验模型, 其中 GPR 和 CatBoost 模型的预测精度高, 但 GPR 模型稳定性最好。【结论】考虑到所研究模型调参的复杂性、预测精度和稳定性, GPR 模型可作为江西地区参考作物蒸散量模拟的推荐方法。

关键词: 参考作物蒸散量; 高斯过程回归; 极限提升增强; 梯度提升决策树; 经验模型

中图分类号: S274.1; S274.4

文献标志码: A

doi: 10.13522/j.cnki.gggs.2020056

OSID: 

刘小强, 代智光, 吴立峰, 等. GPR、XGBoost 和 CatBoost 模拟江西地区参考作物蒸散量的适应性研究[J]. 灌溉排水学报, 2021, 40(1): 91-96.

LIU Xiaoqiang, DAI Zhiguang, WU Lifeng, et al. Comparing the Performance of GPR, XGBoost and CatBoost Models for Calculating Reference Crop Evapotranspiration in Jiangxi Province[J]. Journal of Irrigation and Drainage, 2021, 40(1): 91-96.

0 引 言

【研究意义】作物需水量是农田土壤水分循环的关键因子, 对水资源优化配置和灌溉制度的制定有重要意义, 而计算作物需水量的关键是确定参考作物蒸散量(ET_0)^[1]。【研究进展】国内外通常将 FAO-56 Penman-Monteith(P-M)作为估算 ET_0 的标准方法^[2], 而 P-M 法需要的气象数据完整性高, 多数气象观测数据无法达到该方法要求, 使得 P-M 法的应用受到一定程度的限制, 于是利用有限气象数据的经验法就得到了广泛应用, 如基于辐射的 Irmak 法^[3]和 Makkink

法^[4]等。张倩等^[5]比较了基于辐射和温度等 9 种方法在新乡的适用性, 发现辐射法中 Irmak 模型的精度高于温度法。胡兴波等^[6]在青海高寒地区发现 Makkink 法可直接用于计算极端干旱区以外的 ET_0 。

近年来, 神经网络方法^[7]、支持向量机^[8]、基因表达式编程^[9]和随机森林^[10]以及各种优化模型(蝙蝠算法优化极限学习机^[11]和极限学习机优化遗传算法^[12]等)由于输入参数组合灵活以及精度优于经验模型而得到广泛研究, 并且在某些特定区域具有更高的精度^[9-10]。【切入点】江西地处我国华东地区, 水热资源丰富, 但由于经常旱涝急转严重制约了作物的高产稳产。此外, 江西不同区域气候差异较大, 但具有长系列气象观测资料的气象站点却匮乏, 无法满足农业生产对气象资料的需要。因此, 确定适宜的 ET_0 计算方法极其重要。而大多数学者运用机器学习模拟 ET_0 时, 以模型预测精度为研究对象较多^[7-9], 而综合考虑其精度和稳定性^[13]的比较研究在江西地区还缺乏报道。

收稿日期: 2020-02-10

基金项目: 江西省教育厅研究项目青年基金项目(GJJ180952); 江西省科技厅自然科学基金项目(20171BAB216051)

作者简介: 刘小强(1995-), 男, 江西进贤人。硕士研究生, 主要从事节水灌溉理论与技术研究。E-mail: liuxiaoqiangyx@163.com

通信作者: 吴立峰(1985-), 男, 黑龙江阿城人。讲师, 博士, 研究方向为节水灌溉理论与技术研究。E-mail: china.sw@163.com

【拟解决的关键问题】为此,以 FAO-56 P-M 计算的 ET_0 结果为对照,建立基于有限的气象数据的 3 种机器学习模型 (GPR、XGBoost 和 CatBoost), 分析不同气象要素对江西地区 ET_0 预测精度的影响和稳定性; 并将机器学习模型与 Irmak 和 Makkink 模型进行比较, 评估机器学习模型的精度和稳定性, 以便筛选出气象数据不足条件下江西地区最适宜的 ET_0 估算替代方法, 以期为江西地区灌溉制度制定和水资源优化配置提供科学指导。

1 材料与方法

1.1 试验区概况

江西省 (24°29'—30°04'N, 113°34'—118°28'E) 位于长江中下游地区, 属中亚热带湿润季风气候, 全省多年年均气温为 16.3~19.5 °C, 且一般自北向南递增。省内降水丰沛, 主要集中在 4—9 月, 多年平均降水量 1 341~1 940 mm。降水的季节性变化大, 汛期河水暴涨, 易泛滥成灾。

1.2 数据收集与处理

选取江西省修水、宜春、吉安、遂川、赣县、庐山、鄱阳、景德镇、南昌、樟树、贵溪、玉山、南城、广昌、寻乌 15 个气象站 2001—2015 年的地面观测数据中的日值数据集 (包括最高气温 ($T_{\max}$)、最低气温 ($T_{\min}$)、相对湿度 (RH)、2 m 高风速 (U_2)、大气顶层辐射 (R_a)、地表辐射 (R_s))。其中 2001—2010 年用于训练, 2011—2015 年用于验证。

1.3 研究方法

1.3.1 FAO-56 Penman-Monteith 模型

FAO-56 Penman-Monteith (P-M) 公式被联合国粮农组织推荐为最适宜估算参考作物蒸散量的方法^[2], 其具体表达式为:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T + 273} U_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)}, \quad (1)$$

式中: ET_0 为参考作物蒸散量; R_n 为地表净辐射; G 为土壤热通量密度; T 为 2 m 高处的平均气温; U_2 为 2 m 高处风速; e_s 和 e_a 分别为饱和水汽压和实际水汽压; Δ 为蒸汽压曲线的斜率; γ 为温度计常数。

1.3.2 高斯过程回归模型

给定训练集 $D = \{(x_i, y_i) | i = 1, 2, \dots, n\}$, 其中 x 为 D 维输入向量, y 为输出的标量, n 为训练样本数, 输入矩阵 X 为 $D \times n$ 列的向量, Y 为目标输出, 因此记为 $D = (X, Y)$ 。高斯过程回归模型 (GPR) 是给定输入向量时确定目标输出的联合高斯分布, 由均值函数 $\mu(x)$ 和协方差函数 $K(x, x')$ ^[14] 给出:

$$f(x) \sim GP(\mu(X), K(X, X')). \quad (2)$$

1.3.3 极端梯度提升模型

极端梯度提升 (XGBoost) 是由 Chen 和 Guestrin^[15] 于 2016 年提出的一个梯度增强机 (GBMs) 的新型算法。XGBoost 模型旨在防止过度拟合, 同时通过简化和正则化使预测保持最佳计算效率而降低计算成本。XGBoost 算法源于“提升”的概念, 它结合了一组弱学习者的所有预测, 通过特殊训练培养强学习者。其计算式为:

$$f_i^{(t)} = \sum_{k=1}^t f_k(x_i) = f_i^{t-1} + f_t(x_i), \quad (3)$$

式中: $f_t(x_i)$ 为步骤 t 的学习者; $f_i^{(t)}$ 和 f_i^{t-1} 是步骤为 t 和 $t-1$; x_i 是输入变量。

1.3.4 梯度提升决策树模型

梯度提升决策树 (CatBoost) 是一种新的梯度提升决策树 (GBDT) 算法^[16]。它成功地处理了分类特征, 并利用训练过程中对分类特征处理, 而不是预处理。该算法的另一个优点是它在选择树结构时用新模式计算叶值, 这有助于减少过度拟合并允许使用整个训练数据集, 即对每个示例数据集进行随机排列并计算该示例的平均值。该方法对于回归任务, 需要将获取的数据平均值用于先验计算。

令 $\theta = [\sigma_1, \sigma_2, \dots, \sigma_n]^T$ 为置换, 然后用式 (4) 代替:

$$X_{\sigma_p, k} = \frac{\sum_{j=1}^{p-1} [X_{\sigma_j, k = X_{\sigma_p, k}}] \cdot Y_{\sigma_j} + \beta \cdot P}{\sum_{j=1}^{p-1} [X_{\sigma_j, k = X_{\sigma_p, k}}] + \beta}, \quad (4)$$

式中: P 为先验值; 参数 β 是先验值的权重。

1.4 统计指标

本研究使用了 3 个常用的统计指标, 分别为平均绝对误差 (MAE)、均方根误差 (RMSE) 和决定系数 (R^2)。

2 结果与分析

2.1 3 种机器学习模型精度的比较

表 1 为 3 种机器学习模型不同输入组合下的预测 ET_0 的性能评估结果。由表 1 可知, 对于训练期, 组合 1~9 的模型精度表现为 XGBoost>CatBoost>GPR, 而组合 10 表现为 CatBoost>XGBoost>GPR。在验证期, 由于多数组合的 RMSE 和 MAE 的误差都在 2.7% 以内, 故 CatBoost 和 GPR 模型具有相似的精度, 整体上 CatBoost 和 GPR 模型预测 ET_0 的精度比 XGBoost 模型高。

合理的输入参数组合对模型模拟的精度有显著提高, 如采用 $T_{\max}$ 、 $T_{\min}$ 、 R_s 、 RH 、 $T_{\max}$ 、 $T_{\min}$ 、 R_s 、

U_2 和 $T_{\max}$ 、 $T_{\min}$ 、 R_s 作为输入参数的模型比采用 $T_{\max}$ 、 $T_{\min}$ 、 R_a 、 RH 、 $T_{\max}$ 、 $T_{\min}$ 、 R_a 、 U_2 和 $T_{\max}$ 、 $T_{\min}$ 、 R_a 模型模拟的效果好，这表明 R_s 比 R_a 对模型模拟效果影响大。另外，模型 9 和模型 10 的性能优于模型 8，表明 RH 、 U_2 对模型模拟的精度有一定的影响。余下组合则展示 R_s 对于预测 ET_0 的影响最大， $T_{\max}/T_{\min}$ 次

之， U_2 最小。在验证期，模型 CatBoost10 的 $RMSE$ 和 MAE 的值是最低的， R^2 最高（ $R^2=0.998$ ， $RMSE=0.073$ mm/d， $MAE=0.050$ mm/d），与上述情况一致。因此考虑到组合 8 仅有温度和地表辐射资料就可获得较高的模拟精度，推荐模型 8 作为该地区 ET_0 适宜模型。

表 1 GPR、XGBoost 和 CatBoost 模型的平均统计指标

Table 1 Average statistical results of the GPR, XGBoost and CatBoost models

输入组合	模型	训练			验证			
		R^2	$RMSE/$ (mm d ⁻¹)	$MAE/$ (mm d ⁻¹)	R^2	$RMSE/$ (mm d ⁻¹)	$MAE/$ (mm d ⁻¹)	R^2 排序
$T_{\max}$ 、 $T_{\min}$	GPR1	0.838	0.612	0.446	0.837	0.618	0.451	1
	XGBoost1	0.849	0.592	0.436	0.833	0.625	0.458	2
	CatBoost1	0.842	0.606	0.446	0.832	0.625	0.459	3
R_s	GPR2	0.934	0.392	0.289	0.940	0.375	0.279	2
	XGBoost2	0.955	0.338	0.250	0.943	0.368	0.275	1
	CatBoost2	0.942	0.365	0.269	0.939	0.377	0.281	3
RH	GPR3	0.366	1.230	0.979	0.314	1.284	1.007	3
	XGBoost3	0.458	1.227	0.977	0.363	1.284	1.006	1
	CatBoost3	0.375	1.225	0.974	0.315	1.284	1.006	2
U_2	GPR4	0.052	1.507	1.262	0.042	1.519	1.276	3
	XGBoost4	0.175	1.506	1.261	0.107	1.519	1.274	1
	CatBoost4	0.050	1.514	1.269	0.045	1.516	1.272	2
$T_{\max}$ 、 $T_{\min}$ 、 R_a	GPR5	0.862	0.565	0.399	0.857	0.577	0.412	1
	XGBoost5	0.813	0.497	0.359	0.807	0.589	0.422	3
	CatBoost5	0.876	0.535	0.384	0.851	0.588	0.423	2
$T_{\max}$ 、 $T_{\min}$ 、 R_a 、 U_2	GPR6	0.883	0.521	0.369	0.870	0.548	0.390	2
	XGBoost6	0.818	0.377	0.271	0.812	0.550	0.386	3
	CatBoost6	0.910	0.456	0.324	0.871	0.547	0.387	1
$T_{\max}$ 、 $T_{\min}$ 、 R_a 、 RH	GPR7	0.925	0.418	0.282	0.922	0.437	0.296	1
	XGBoost7	0.954	0.304	0.213	0.914	0.448	0.303	3
	CatBoost7	0.943	0.368	0.252	0.919	0.443	0.301	2
$T_{\max}$ 、 $T_{\min}$ 、 R_s	GPR8	0.967	0.270	0.194	0.966	0.277	0.205	1
	XGBoost8	0.983	0.161	0.119	0.959	0.283	0.208	3
	CatBoost8	0.976	0.231	0.167	0.966	0.279	0.205	1
$T_{\max}$ 、 $T_{\min}$ 、 R_s 、 RH	GPR9	0.988	0.167	0.111	0.987	0.179	0.117	1
	XGBoost9	0.994	0.061	0.045	0.982	0.179	0.118	3
	CatBoost9	0.992	0.131	0.091	0.986	0.178	0.118	2
$T_{\max}$ 、 $T_{\min}$ 、 R_s 、 U_2	GPR10	0.967	0.269	0.194	0.966	0.277	0.206	2
	XGBoost10	0.993	0.123	0.092	0.964	0.287	0.211	3
	CatBoost10	0.998	0.052	0.039	0.998	0.073	0.050	1

本研究通过分析 R^2 的大小比较 3 种机器学习模型的差异（表 1），可得，GPR 模型中有 5 个组合预测 ET_0 的 R^2 最高，其中组合 $T_{\max}$ 、 $T_{\min}$ 、 R_s 、 U_2 的最高 R^2 为 0.987；XGBoost 模型有 3 个组合预测 ET_0 的

R^2 最高，这些组合包含 R_s 、 RH 、 U_2 ，而最高 R^2 为 0.943；CatBoost 模型含有风速时预测 ET_0 的 R^2 最高，其 R^2 为 0.998。此外，有 5 个组合预测 ET_0 的 R^2 排在第 2 位。总体上看，在验证期中，XGBoost 模型 R^2 排序

最大, 排第 3 位, CatBoost 模型排第 2 位, 而 GPR 模型 R^2 的排序最小, 排第 1 位。

2.2 3 种机器学习模型的稳定性比较

由表 1 加粗字体可知, 在训练期, 总体上 XGBoost 模型优于 GPR 和 CatBoost 模型, 然而验证期, GPR 模型却优于 CatBoost 和 XGBoost 模型。通过分析机器学习模型验证期相对训练期的平均 RMSE 及其百

分比(表 2)可知: 对于 3 种机器学习模型, XGBoost 模型验证期平均 RMSE 的百分比在各个组合均最大, 其最大百分比是 193.4%; 而 GPR 模型其百分比增长幅度最小, 都在 8%以内; 对于 CatBoost 模型, 在前 5 个组合中, 其百分比在 10%以内, 而后 5 个组合中其介于 20%~41%之间, 说明 GPR 模型模拟时稳定性最好, 其次是 CatBoost 模型, 而 XGBoost 模型最差。

表 2 机器学习模型验证期相对训练期的平均 RMSE 及其百分比

Table 2 The average RMSE and percentage of machine learning models during the texting period relative to the training period

输入组合	训练			验证			百分比/%		
	GPR	XGBoost	CatBoost	GPR	XGBoost	CatBoost	GPR	XGBoost	CatBoost
$T_{\max}$ 、 $T_{\min}$	0.612	0.592	0.606	0.618	0.625	0.625	0.980	5.574	3.135
R_s	0.392	0.338	0.365	0.375	0.368	0.377	-4.337	8.876	3.288
RH	1.230	1.227	1.225	1.284	1.284	1.284	4.390	4.645	4.816
U_2	1.507	1.506	1.514	1.519	1.519	1.516	0.796	0.863	0.132
$T_{\max}$ 、 $T_{\min}$ 、 R_a	0.565	0.497	0.535	0.577	0.589	0.588	2.124	18.511	9.907
$T_{\max}$ 、 $T_{\min}$ 、 R_a 、 U_2	0.521	0.377	0.456	0.548	0.550	0.547	5.182	45.889	19.956
$T_{\max}$ 、 $T_{\min}$ 、 R_a 、 RH	0.418	0.304	0.368	0.437	0.448	0.443	4.545	47.368	20.380
$T_{\max}$ 、 $T_{\min}$ 、 R_s	0.270	0.161	0.231	0.277	0.283	0.279	2.593	75.776	20.779
$T_{\max}$ 、 $T_{\min}$ 、 R_s 、 RH	0.167	0.061	0.131	0.179	0.179	0.178	7.186	193.443	35.878
$T_{\max}$ 、 $T_{\min}$ 、 R_s 、 U_2	0.269	0.123	0.052	0.277	0.287	0.073	2.974	133.333	40.385

表 3 经验模型和机器学习模型的平均统计指标

Table 3 Average statistical results of the empirical and machine learning models

输入组合	模型	训练			验证		
		R^2	RMSE/ (mm d ⁻¹)	MAE/ (mm d ⁻¹)	R^2	RMSE/ (mm d ⁻¹)	MAE/ (mm d ⁻¹)
$T_{\max}$ 、 $T_{\min}$ 、 R_s	Irmak	0.92	0.426	0.340	0.92	0.430	0.342
	GPR8	0.96	0.270	0.194	0.96	0.277	0.205
	XGBoost	0.98	0.161	0.119	0.95	0.283	0.208
	CatBoost	0.97	0.231	0.167	0.96	0.279	0.205
$T_{\max}$ 、 $T_{\min}$ 、 R_s 、 RH	Makkink	0.92	0.447	0.337	0.93	0.440	0.333
	GPR9	0.98	0.167	0.111	0.98	0.179	0.117
	XGBoost	0.99	0.061	0.045	0.98	0.179	0.118
	CatBoost	0.99	0.131	0.091	0.98	0.178	0.118

2.3 3 种机器学习模型与经验模型比较

本研究分析了经验模型与相同输入参数的机器学习模型预测 ET_0 的平均统计指标(表 3), 可得机器学习模型的精度都高于经验模型。在 $T_{\max}$ 、 $T_{\min}$ 和 R_s 的输入组合下, Irmak 模型预测精度最低(验证期 $R^2=0.922$, $RMSE=0.430$ mm/d, $MAE=0.342$ mm/d), 而 GPR8 模型预测精度最高(验证期 $R^2=0.966$, $RMSE=0.277$ mm/d, $MAE=0.205$ mm/d); 在 $T_{\max}$ 、 $T_{\min}$ 、 R_s 和 RH 的输入组合下, 验证期中 Makkink 模型预测 ET_0 的精度最低($R^2=0.931$, $RMSE=0.440$ mm/d,

$MAE=0.333$ mm/d)。

3 讨论

3.1 气象参数输入组合方式

输入气象参数组合方式是机器学习模型预测高精度的 ET_0 的关键因子。本研究中, 当使用相对湿度和风速时, 机器学习模型的模拟值与世界粮农组织推荐的标准方法^[2]计算值偏差最大, 然而使用温度($T_{\max}/T_{\min}$)和辐射数据时, 机器学习模型的模拟值精度高, 与 Fan 等^[10]和 Feng 等^[17]在亚热带季风性湿润地区基于温度和地表辐射的机器学习模型预测 ET_0 的精度高和基于温度和大气顶层辐射模拟精度较高的结果一致。主要是在作物生长过程中, 太阳辐射和温度是不可替代的关键因素。当使用组合 $T_{\max}$ 、 $T_{\min}$ 、 R_s 、 U_2 时, U_2 与 R_s 的耦合作用对 CatBoost 模型预测精度影响巨大, 具体出现的原因还有待进一步研究。此外, 模型预测精度随着输入气象参数个数增加而提高, 与前人研究^[18-20]结果一致。

3.2 机器学习模型的预测精度

本研究 GPR 模型在验证期预测 ET_0 的精度高。Holman 等^[14]发现, 在高原地区高斯过程比最小二乘回归的精度高。Karbasi 等^[21]研究表明: GPR 模型随着使用时间序列的增长其预测的精度越高, 但具体能否在江西地区获得相同的结果, 还有待进一步验证。

Jhaveri 等^[22]在其他领域也应用 CatBoost 和 XGBoost 模型, 由于 XGBoost 模型存在过度拟合的问题, 故 XGBoost 模型精度较差。Huang 等^[23]发现, 由于 CatBoost 模型是将该模型获得最佳的训练精度来获得最优结果, 故 CatBoost 模型的精度较高, 但本研究中 GPR 和 CatBoost 模型在 $T_{\max}$ 、 $T_{\min}$ 、 R_s 、 RH 的组合下 $RMSE$ 和 MAE 的误差都在 0.9% 以内, 当输入 3 个参数时, $RMSE$ 和 MAE 的误差都在 2.7% 内而输入 1 个参数的 $RMSE$ 和 MAE 的误差都在 0.7% 内, 表明 GPR 模型模拟江西地区 ET_0 的精度高。

3.3 机器学习模型的稳定性

机器学习模型的稳定性是预测 ET_0 时需要考虑的关键因素。研究表明, 在机器学习模型中, XGBoost 模型验证期相对训练期的 $RMSE$ 百分比增长最大, 其次是 CatBoost 模型, GPR 模型可能是因为能够处理非线性关系使其增长最小, 但具体原因还有待后续研究。此结果揭示了 XGBoost 模型极不稳定, 且随着使用气象参数个数的增加, XGBoost 模型预测稳定性出现显著下降, 与 Fan 等^[24]利用 XGBoost 模型预测太阳辐射时, 验证期 $RMSE$ 增长幅度比其他模型大, 而 CatBoost 模型对早期预测不正确的点赋予额外的权重后进行加权预测使 CatBoost 模型的 $RMSE$ 百分比增加幅度比 XGBoost 模型小的结果一致。

4 结 论

机器学习模型提高了江西地区参考作物蒸散量的精度, 且各气象要素对机器学习模型模拟效果的影响由大到小依次为: R_s 、 $T_{\max}/T_{\min}$ 、 RH 、 U_2 。

使用 $T_{\max}$ 、 $T_{\min}$ 和 R_s 作为输入组合的 GPR 模型, 验证期 $R^2=0.966$, $RMSE=0.277$ mm/d, $MAE=0.205$ mm/d, 为江西地区适宜的参考作物蒸散量模型。

参考文献:

- [1] MEHDIZADEH S. Estimation of daily reference evapotranspiration (ET_0) using artificial intelligence methods: Offering a new approach for lagged ET_0 data-based modeling [J]. Journal of Hydrology, 2018, 559: 794-812.
- [2] ALLEN R G, PEREIRA L S, RAES D, et al. Crop evapotranspiration (guidelines for computing crop water requirements) [M]. Rome: FAO, 1998.
- [3] IRMAK S, IRMAK A, ALLEN R G, et al. Solar and net radiation-based equations to estimate reference evapotranspiration in humid climates[J]. Journal of Irrigation and Drainage Engineering, 2003, 129(5): 336-347.
- [4] MAKINK G F. Testing the Penman formula by means of lysimeters [J]. Journal of the Institution of Water Engineers, 1957, 11(3): 277-288.

- [5] 张倩, 段爱旺, 高阳, 等. 基于温度资料估算参考作物腾发量的方法比较[J]. 农业机械学报, 2015, 46(2): 104-109.
- ZHANG Qian, DUAN Aiwang, GAO Yang, et al. Comparative analysis of reference evapotranspiration estimation methods using temperature data [J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(2): 104-109.
- [6] 胡兴波, 芦新建, 董梅, 等. 简化参照作物蒸散量(ET_0)计算公式在青海省高寒区的适用性分析[J]. 西北农林科技大学学报(自然科学版), 2013, 41(11): 201-208.
- HU Xingbo, LU Xinjian, DONG Mei, et al. Applicability of simplified reference crop evapotranspiration equations in high altitude and cold area of Qinghai Province[J]. Journal of Northwest A & F University (Natural Science Edition), 2013, 41(11): 201-208.
- [7] 赵文刚, 马孝义, 刘晓群, 等. 基于神经网络算法的广东省典型代表站点 ET_0 简化计算模型研究[J]. 灌溉排水学报, 2019, 38(5): 91-99.
- ZHAO Wengang, MA Xiaoyi, LIU Xiaoqun, et al. Using neural network model to simplify ET_0 calculation for representative stations in Guangdong Province[J]. Journal of Irrigation and Drainage, 2019, 38(5): 91-99.
- [8] YAO Y J, LIANG S L, LI X L, et al. Improving global terrestrial evapotranspiration estimation using support vector machine by integrating three process-based algorithms[J]. Agricultural and Forest Meteorology, 2017, 242: 55-74.
- [9] WANG S, FU Z Y, CHEN H S, et al. Modeling daily reference ET in the Karst area of northwest Guangxi (China) using gene expression programming (GEP) and artificial neural network (ANN)[J]. Theoretical and Applied Climatology, 2016, 126(3): 493-504.
- [10] FAN J L, YUE W J, WU L F, et al. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China[J]. Agricultural and Forest Meteorology, 2018, 263: 225-241.
- [11] DONG J H, WU L F, LIU X G, et al. Estimation of daily dew point temperature by using bat algorithm optimization based extreme learning machine[J]. Applied Thermal Engineering, 2020, 165: 114569.
- [12] WU L F, ZHOU H M, MA X, et al. Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: Application in contrasting climates of China[J]. Journal of Hydrology, 2019, 577: 123960.
- [13] HASSAN M A, KHALIL A, KASEB S, et al. Exploring the potential of tree-based ensemble methods in solar radiation modeling[J]. Applied Energy, 2017, 203: 897-916.
- [14] HOLMAN D, SRIDHARAN M, GOWDA P H, et al. Gaussian process models for reference ET estimation from alternative meteorological data sources[J]. Journal of Hydrology, 2014, 32: 28-35.
- [15] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acmsigkdd international conference on

- knowledge discovery and data mining [EB/OL], 2016(8): 785-794.
- [16] DOROGUSH A V, ERSHOV V, GULIN A. CatBoost: gradient boosting with categorical features support [EB/OL]. 2018: arXiv: 1810.11363[cs.LG]. <https://arxiv.org/abs/1810.11363>
- [17] FENG Y, PENG Y, CUI N B, et al. Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data[J]. Computers and Electronics in Agriculture, 2017, 136: 71-78.
- [18] TORRES A F, WALKER W R, MCKEE M. Forecasting daily potential evapotranspiration using machine learning and limited climatic data[J]. Agricultural Water Management, 2011, 98(4): 553-562.
- [19] TABARI H, KISI O, EZANI A, et al. SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment[J]. Journal of Hydrology, 2012, 444: 78-89.
- [20] ANTONOPOULOS V Z, ANTONOPOULOS A V. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables[J]. Computers and Electronics in Agriculture, 2017, 132: 86-96.
- [21] KARBASI M. Forecasting of multi-step ahead reference evapotranspiration using wavelet- Gaussian process regression model[J]. Water Resources Management, 2018, 32(3): 1 035-1 052.
- [22] JHAVERI S, KHEDKAR I, KANTHARIA Y, et al. Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns[C]//2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019(2): 1 170-1 173.
- [23] HUANG G M, WU L F, MA X, et al. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions[J]. Journal of Hydrology, 2019, 574: 1 029-1 041.
- [24] FAN J L, WU L F, MA X, et al. Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions[J]. Renewable Energy, 2020, 145: 2 034-2 045.

Comparing the Performance of GPR, XGBoost and CatBoost Models for Calculating Reference Crop Evapotranspiration in Jiangxi Province

LIU Xiaoqiang^{1,2}, DAI Zhiguang¹, WU Lifeng^{1*}, ZHANG Fucang², DONG Jianhua³, CHEN Zhiyue⁴

(1.College of water conservancy and ecological engineering, Nanchang Institute of Technology, Nanchang 330099, China;

2. Key Laboratory of Agricultural Soil and Water Engineering in Arid and Semiarid Areas, Ministry of Education, Northwest A&F

University, Yangling 712100, China; 3. Faculty of Agriculture and Food, Kunming University of Science and Technology,

Kunming 650500, China; 4. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract: 【Background】 Alternate drought and waterlogging increasingly occurring in Jiangxi province means that rational irrigation strategies are required to safeguard its agricultural production. 【Objective】 The objective of this paper is to select a suitable machine learning model to calculate reference crop evapotranspiration across the province. 【Method】 Meteorological data - including daily maximum ($T_{\max}$) and minimum ($T_{\min}$) ambient temperature, global solar radiation, extra-terrestrial solar radiation(R_s), relative humidity (RH) and 2m-height wind speed (U_2) - were measured from 2001 to 2015 at 15 stations across the province; they were then used to train and test three models: The gaussian process regression (GPR), the extreme gradient boosting (XGBoost), and the gradient boosting with categorical features support (CatBoost). We compared accuracy with empirical model for estimating the reference evapotranspiration. 【Result】 The meteorological factors that impacted the accuracy of the machine learning model for estimating ET_0 was ranked in the descending order as follows based on their significance: $R_s > T_{\max} > T_{\min} > RH > U_2$. Models using $T_{\max}$, $T_{\min}$, R_s and U_2 gave the most accurate ET_0 estimate with $RMSE < 0.2$ mm/d. All three models have a good applicability by using limited meteorological data, and are superior to the traditional empirical model. In particular, GPR and CatBoost were more accurate, and GPR was most stable. 【Conclusion】 In terms of complexity, accuracy and stability, GPR was the most suitable model for estimating reference crop evapotranspiration in Jiangxi province.

Key words: reference crop evapotranspiration; gaussian process regression; extreme gradient boosting; gradient boosting with categorical features support; empirical model

责任编辑: 韩 洋