基于特征优选与优化算法的干热河谷区柑橘蒸腾模拟

吴明清[1,2,3], 李伟康[1,2,3], 王 静[1,2,3], 李 靖[1,2,3], 裴仁涛[1], 黄海涛[1,2], 邱皇皇[1,2], 雷楚靖[1,2], 赵 多[1], 黄锦焘[4], 高志永[1,2,3,5*]

(1. 云南农业大学水利学院,云南昆明 650201,中国; 2. 云南省智慧农业与水安全国际联合研发中心,云南昆明 650201,中国; 3. 滇池湖泊生态系统云南省野外科学观测研究站,云南昆明 650228,中国; 4. 云南农业大学建筑工程学院,云南昆明 650201,中国; 5. 陆良山间盆地土地利用云南省野外科学观测研究站,云南曲靖 655600,中国)

摘 要: [**目的/意义**] 干热河谷正加快建设智慧灌区,蒸腾量是量化作物耗水过程和需水状况的关键指标,其准确模拟是优化灌区水分调控与配水策略、提高灌溉决策可靠性的基础。当前灌区仍以传统机理模型为主,但此类模型参数需求高、对部分参数高度敏感且误差来源复杂,难以在数据受限与地形复杂的实际场景中应用。[**方法**] 本研究以10年生柑橘树为研究对象,收集了2020—2022年期间的树干液流与环境因子数据。采用随机森林递归特征消除(Random Forest-Recursive Feature Elimination, RF-RFE)识别蒸腾的关键驱动因子,并对这些关键因子的贡献进行量化分析。随后,基于RF-RFE筛选出的因子,构建了三种机器学习模型:随机森林、支持向量回归与极端梯度提升(Extreme Gradient Boosting, XGB)。最后,引入粒子群优化、贝叶斯优化与异步连续减半(Asynchronous Successive Halving Algorithm, ASHA)对模型进行超参数优化,并系统评估了12种模型的模拟性能。[**结果和讨论**] 在RF-RFE中,17个候选变量被精简为5个关键因子(实际水汽压、土壤体积含水率、风速、平均气温和日序数),精简后的模型性能不降反升。其中日序数特征重要性最高,有效整合了区域季节性气候变化与柑橘物候特征。综合对比表明,XGB-ASHA模型具有最优的预测性能,其决定系数为0.91,均方根误差为0.19 mm/d。[**结论**] 基于特征优选和代理变量的建模方法,可以实现高精度、低参数依赖的柑橘蒸腾量精准估算,为该地区农业水管理提供智慧决策的科学依据。

关键词: 蒸腾耗水; 机器学习; 特征选择; 智慧农业; 水资源管理

中图分类号: S274.4;TP391.9;TP274 **文献标志码:** A **文章编号:** SA202511007

引用格式:吴明清, 李伟康, 王静, 李靖, 裴仁涛, 黄海涛, 邱皇皇, 雷楚靖, 赵多, 黄锦焘, 高志永. 基于特征优选与优化算法的干热河谷区柑橘蒸腾模拟[J]. 智慧农业(中英文), 2026, 8(4): 192-203. DOI: 10.12133/j.smartag.SA202511007

WU Mingqing, LI Weikang, WANG Jing, LI Jing, PEI Rentao, HUANG Haitao, QIU Huanghuang, LEI Chujing, ZHAO Duo, HUANG Jingtao, GAO Zhiyong. Estimation of Citrus Transpiration in a Savanna Valley Based on Feature Selection and Optimization Algorithms[J]. Smart Agriculture, 2026, 8(4): 192-203. DOI: 10.12133/j.smartag.SA202511007 (in Chinese with English abstract)

0 引 言

全球水资源紧缺是21世纪最紧迫的挑战之一,农业用水消耗约70%的可用淡水资源,且随着人口增长和气候变化,这一需求仍在不断增加[1]。蒸腾作用是陆地生态系统中水循环的关键环节[2,3],也是全球大陆最主要的水分通量,占总蒸散量的80%~90%[4]。蒸腾作为连接水分、能量与碳循环的关键过程,在调节土壤-植物-大气连续体(Soil-Plant-Atmosphere Continuum, SPAC)中的能-水通量中发挥着关键作用[5]。因此,准确估算蒸腾量对于深入理解作物水分利用机制、优化区域水资源管理具有重要理论与实践意义[6]。

实测和模拟是获取蒸腾的两种有效途径,蒸腾

收稿日期:2025-11-09

基金项目:云南省农业基础研究联合专项面上项目(202301BD070001-181);国家重点研发计划项目(2023YFD1901203)

作者简介:1. 吴明清,硕士研究生,研究方向为作物高效用水。E-mail:Wmqing1015@163.com;2. 李伟康,硕士研究生,研究方向为水钾一体化调控作物产量品质。E-mail:liw1328131531@163.com。吴明清、李伟康为共同一作

*通信作者:高志永,博士,讲师,研究方向为经济林耗水规律及调控。E-mail:GZYstruggling@163.com

copyright©2026 by the authors

的实测可以通过树干液流法[7]、气体交换室法[8]、微型蒸渗仪法[9]和同位素法[10]直接监测，但空间代表性有限，且需要大量的人力、物力和时间投入，难以满足大尺度应用需求[11]；模拟则借助经验模型、物理模型与机器学习（Machine Learning, ML）模型来估算蒸腾量[5]。其中多元线性回归（Multiple Linear Regression, MLR）等传统经验模型形式简单，但因缺乏物理机制约束，其参数往往不具备普适性，在新环境下的泛化能力受限[5]。以Penman-Monteith（P-M）[12]和Shuttleworth-Wallace（S-W）[13]为代表的传统物理模型虽机理清晰，但需要大量的变量且对这些参数误差敏感，此外，关键参数的率定也较为困难，例如空气动力学、气孔和土壤表面阻力。更重要的是，这些传统模型在捕捉蒸腾量与其驱动因子间复杂非线性耦合关系方面存在局限[14]。

近年来，ML因其无需预设物理方程，能够直接从数据中学习并精准刻画因子间高度复杂的非线性关系，已被广泛用于构建高精度作物水分相关预测模型[5]。为进一步提升模型性能，许多研究通过融合多源异构数据。例如，HUANG等[15]在气象和土壤数据的基础上，融入遥感植被指数可有效提升随机森林（Random Forest, RF）模型的精度。同样，整合叶面积指数（Leaf Area Index, LAI）、感热通量等更多维度的生物物理特征，也被证实能够获得比仅使用气象数据更优的模拟结果[16]。然而，单纯增加特征数量并非最优策略。冗余、低相关性或含噪声的变量反而会增加模型复杂度、引入干扰，甚至削弱模型的泛化能力。因此，特征选择已成为构建高性能ML模型的关键环节。PAGANO等[1]在对意大利柑橘园的蒸腾进行建模时，通过特征重要性排序将输入特征从10个精简至7个，RF模型的性能反而得到了提升。此外，关键驱动因子的重要性并非一成不变，而是具有高度的时空动态性。一方面，其重要性随作物物候阶段而变化。例如，XING等[17]在中国黄土高原的苹果树蒸腾研究中发现，在冠层稀疏阶段LAI和饱和水汽压差（Vapor Pressure Deficit, VPD）最为关键，而在冠层茂密阶段则变为VPD和净辐射（Net Radiation, R_n）。另一方面，关键因子组合也表现出显著的生态系统特异性。温带气候森林[18]、温带气候温室番茄[19]，以及南非半干旱区的亚热带果园[20]的蒸腾关键特征因子均不相同。这种普遍存在的时空异质性与多源特征输入，凸显了针对特定区域和特定作物进行建模与关键特征识别的必要性。柑橘是全球广泛种植的水果之一，根据联合国粮食及农业组织，中国已成为全球最大的柑橘生产国，约占全球总产量的21%[21]。干热河谷区凭借其优越的光热条件，成为中国柑橘的主要优质产区。然而，该区域典型的季节性干旱与水资源时空分布不均的问题，使得粗放的灌溉管理成为制约柑橘产业可持续发展的关键瓶颈。为突破此瓶颈，在地方政府主导下，以精准作物耗水估算为核心的智慧灌溉果园建设正在大力推进，以保障干热河谷柑橘产业的稳产增收。然而，精准的耗水估算是实现智慧灌溉的技术前提，但现有模型在该地区面临严峻挑战。传统的P-M模型由于缺乏本地化的作物系数，难以准确估算柑橘耗水。尽管CHEN等[22]通过引入动态冠层消光系数改进了P-M模型，但其多达10个参数的复杂本地化校准过程，使其在智慧农业中的推广与应用潜力受到极大限制。因此，能否利用ML对特征间耦合能力的挖掘，在显著降低模型对复杂参数依赖的同时，为干热河谷这一独特气候区的柑橘建立高精度的蒸腾量模拟模型，既是一个关键的科学问题，也是一个明确的研究空白。

本研究旨在应对此挑战，通过筛选易于获取的多维度耦合因子，对比多种ML模型（RF、支持向量回归（Support Vector Regression, SVR）、极端梯度提升（Extreme Gradient Boosting, XGB）及超参数优化算法（粒子群优化（Particle Swarm Optimization, PSO）、贝叶斯优化（Bayesian Optimization, BO）、异步连续减半算法（Asynchronous Successive Halving Algorithm, ASHA））的组合效果，以开发一种兼具高精度与实用性的柑橘树蒸腾量估算新方法。具体研究目标如下：1）利用随机森林-递归特征消除（Random Forest-Recursive Feature Elimination, RF-RFE），筛选出驱动干热河谷柑橘蒸腾量的关键特征组合；2）系统评估不同ML模型与超参数优化算法的组合性能，以确定适用于本区域的最优模型。

1　材料和方法

1.1　研究区概况

本研究试验点设于中国云南省元江干热河谷核心区的玉溪市新平彝族傣族自治县（23°58′6″N，101°38′56″E，海拔高度550 m），属于典型的亚热带低纬度高原季风气候区，主要气候特征表现为光热资源充裕与季节性干旱并存。该区年均相对

湿度为67%，年日照时数为2 350 h，潜在蒸发量为1 750 mm。年均降水量为733 mm，其中77.9%的降水发生在6月至10月的雨季。该地区年均气温为24.9 ℃，其中4月至6月为高温季节，最高气温可达43 ℃（超过100天的气温超过35 ℃），最低气温为3.6 ℃。该地区土壤为典型燥红壤，土壤表层薄、含砾量多、粗骨性强（0～1 m颗粒分布占比：粒径（0.05，2.00］ mm为35.2%、(0.02，0.05］ mm为30%、(0.002，0.02］ mm为23.32%、(0，0.002］ mm为11.48%），pH为6.1，田间持水量为29.80 cm^3/cm^3，土壤容重为1.29 g/cm^3。

1.2 供试材料

研究对象为元江干热河谷区大面积种植的典型人工经济林木柑橘树，品种为甜橙（*Citrus sinensis*（L.）Osbeck），试验柑橘树于2012年种植，种植密度为1 665株/hm^2（株距×行距：2 m×3 m）。果园总面积为9.33 hm^2，其中0.13 hm^2作为实验样地，具体可见［22］。柑橘树嫁接在三叶橙砧木上，以增强其抗逆性并改善果实品质。在果园中，管理措施包括灌溉、施肥和修剪，均按照该地区普遍采用的常规方法进行。树木定期修剪，以维持均匀的高度和冠幅；同时，疏果和疏花有助于保持合理的果实负荷。在此条件下，树木表现出相对一致的生长，包括主干直径和叶面积。因此，果园中的柑橘树没有显著的大小分类差异。雨季期间，自然降水基本可满足柑橘树需水，未采取人工灌溉；在旱季，为保证柑橘树平稳度过环境胁迫阶段，每年根据干旱程度采取灌溉措施三次（2021年1月27日灌水量为25 mm；2021年2月11日灌水量为25 mm；2021年2月25日灌水量为30 mm；2022年2月19日灌溉25 mm；2022年3月11日灌溉25 mm；2022年4月2日灌溉25 mm）。田间试验从2020年5月20日持续到2022年7月7日。

1.3 监测指标与数据采集

1.3.1 冠层蒸腾量

由于试验区果园采用相同的管理条件，树木生长情况较为一致，因此随机选取7株柑橘树作为重复进行研究。本研究采用热扩散式探针（Thermal Dissipation Probe, TDP）（SF-G，Ecomatik GmbH，德国）实时监测试验柑橘树的树干液流。采用CR1000X型数据采集器每15 min自动记录1次数据，树干液流速率计算详情请参考［23］。

1.3.2 气象要素

研究期内，采用自动气象站（AG1000型，Campbell Scientific Inc., USA）实时监测气象数据，监测时间步长为15 min，测量指标包括降水量（Precipitation, *P*, mm）、净辐射（Net radiation, R_n, W/m^2）、太阳总辐射（Total solar radiation, R_s, W/m^2）、光合有效辐射（Photosynthetically Active Radiation, PAR, μmol/s/m^2）、气温（Air temperature, ℃），以及相对湿度（Relative humidity, *RH*, %）、2 m处风速（u_2, m/s）。水汽压差（VPD, kPa）由饱和水汽压（e_s, kPa）和实际水汽压（e_a, kPa）计算得出[24]。

1.3.3 土壤水分

研究期内，土壤水分通过时域反射仪（TDR，CS650型，Campbell Scientific Inc., USA）进行监测。考虑到土壤水分的空间异质性，在柑橘园内沿坡度梯度随机选取行内和行间各两个监测点位（共4个点位），每个点位安装4个探头。因柑橘树水分吸收的主要根区在0～60 cm土层[25]，本研究中监测0～60 cm土层土壤体积含水率（Volumetric Water Content, VWC, %）。具体来说，2个剖面位于行内，用于测量冠层空间下的土壤水分（分别距离树干0.5 m和1.0 m）。另外2个剖面位于行间，用于测量冠层荫蔽区外的土壤水分（距离树干1.25 m，一个在内坡位置，另一个在外坡位置）。在每个剖面的10、20、40、60 cm深度处各安装一个TDR探头，所有土壤含水量数据均通过数据采集器（CR1000X，Campbell Scientific）每15 min进行1次数据采集。在试验期间，为确保TDR测定结果的准确性，每隔2个月利用烘干法进行校准。

1.4 数据处理和划分

1.4.1 数据收集及预处理

本研究所使用的环境特征共16个，包括气温（最大值、最小值和平均值；T_{max}、T_{min}和T_a）、相对湿度（最大值、最小值和平均值；RH_{max}、RH_{min}和RH_a）、太阳辐射（*R*s）、净辐射（R_n）、PAR、日照时数（*n*）、大气风速（u_2）、饱和水汽压差（e_s）、实际水汽压（e_a）、VPD、降雨量（*P*）和0～60 cm的VWC，以监测柑橘树的蒸腾量为因变量（图1）。

因干热河谷地区独特的地形和气候因素及其受“焚风效应”的影响，区域气象因子全年表现出明显的季节性变化趋势，尤其在旱季与雨季交替期间更为明显。此外，前人研究表明，影响作物蒸腾的

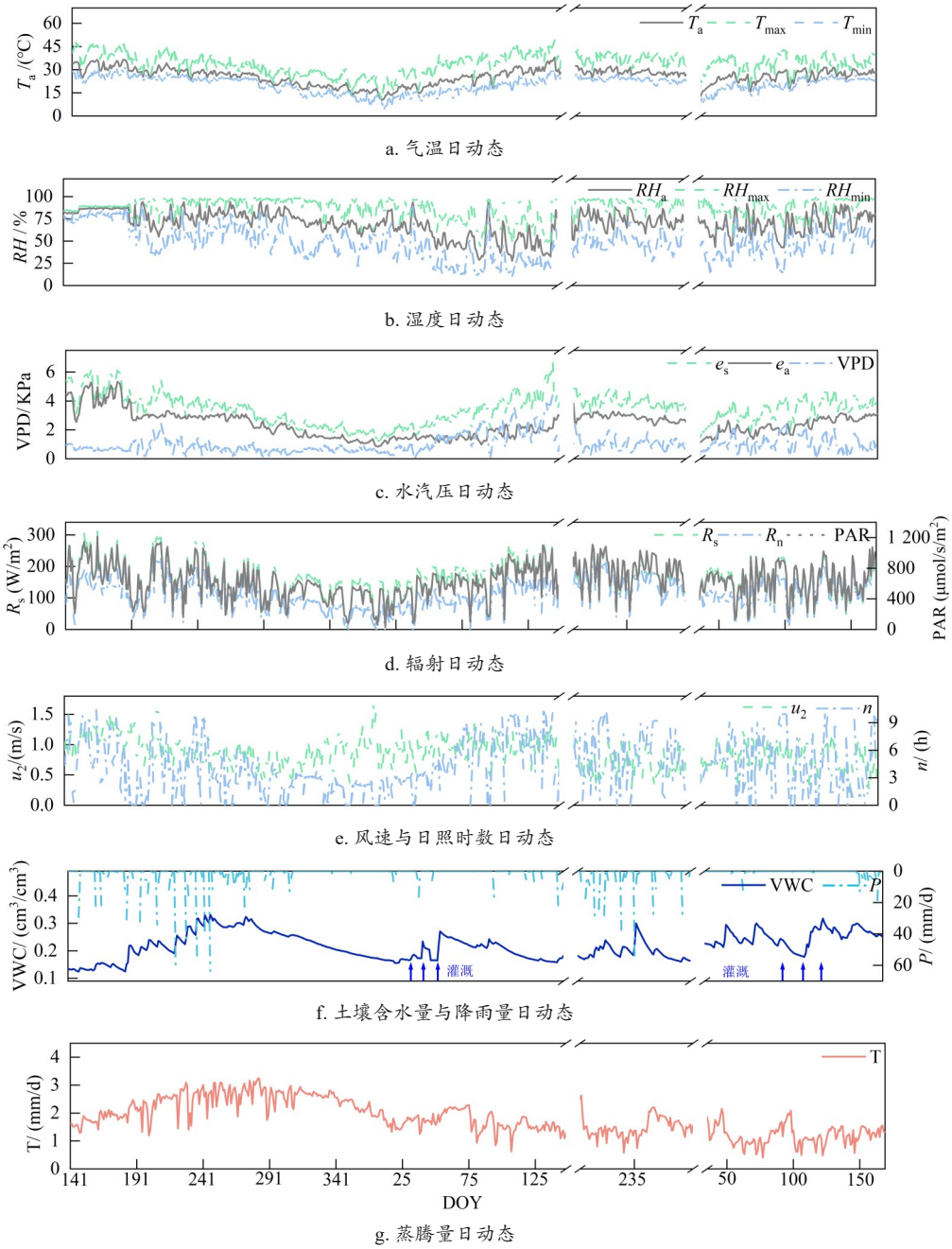


图 1　2020—2022 年试验区域的环境因子与柑橘蒸腾量的日尺度动态变化

Fig. 1　Daily-scale dynamic variations of environmental factors and citrus transpiration in the experimental area during 2020—2022

气象因子之间存在非线性耦合与交互效应[26]，为了使模型能够更充分挖掘时间因素与气象因子之间的潜在关联，从而更准确地反映蒸腾在不同季节背景下的动态响应特征，本研究加入了反映作物的物候变化特征的年积日（Day of Year, DOY）。为识别并保留蒸腾过程中的关键特征因子，在兼顾模型性能的前提下，去除冗余、无关或噪声特征，通过 RF-RFE 因子间的共线性，递归地移除重要性最低的特征来选择最优特征子集，随机森林基准模型使用默认的超参数配置，当特征数量减少至 1 时，迭代停止。此外，通过皮尔逊相关热力图展示因子间的线性相关性，以便更充分展现因子间的共线性，理解特征优选对冗余特征的剔除。在数据收集过程中，受设备断电的影响，部分数据存在缺失。为确保模型训练的有效性，仅保留当日特征数据均完整的样本用于建模；缺失任意变量的数据将在预处理

阶段予以剔除，不纳入模型的建立与验证流程。

1.4.2 数据划分

在ML模型建模中，数据集通常被划分为训练集和测试集两个相互独立的子集。然而，这种简单的划分方式在存在明显的作物物候变化或季节性气候变化的农业或水文研究中，可能导致数据分布不均，削弱模型对特征季节性变化的学习能力。本研究针对干热河谷地区季节性干旱、区域气候季节性明显的特点，根据ML建模中常用的8∶2的划分比例，采取系统抽样的数据划分方式，即按时间序列每5天为一个单元，将第1天划入测试集，后4天划入训练集，从而将数据划分为80%的训练集与20%的测试集。该方法确保了训练集和测试集在全年的时间跨度上分布均匀，涵盖了相似的气象变化与蒸腾强度范围，以减轻环境因子季节性趋势带来的潜在的抽样偏倚。

1.5 机器学习模型

本研究采用了三种ML模型（RF、SVR和XGB）[5, 27]进行蒸腾建模。每种模型不仅在原始形式下应用，同时也与三种优化算法（PSO、BO和ASHA）[26, 28, 29]结合，共形成12种模型组合。建模所用的输入变量为特征选择后保留的关键特征。在模型训练过程中，在各模型中输入预处理好的80%数据集。同时，为了避免模型过拟合并充分利用数据集，在超参数调优与最优模型训练阶段使用十折交叉验证，在十折交叉验证中，训练集被划分为10个等量子集，每轮迭代中选取其中一个子集作为验证集，其余9个子集用于模型训练。该过程重复10次，确保每个子集都作为一次验证集，从而最大程度减少因数据划分带来的偏差，在样本量有限的情况下提高模型的鲁棒性。每轮交叉验证中记录对应的自适应函数平均绝对误差（Mean Absolute Error, MAE），最终计算其平均值作为模型性能的综合评估指标，并据此比较不同参数组合的表现，从而确定最优超参数配置。表1展示了模型参数的具体设置。在本研究中所有模型均使用Python 3.9.13中的scikit-learn软件包进行训练。

表1 不同模型及优化算法的超参数搜索空间

Table 1 Hyperparameter search space for different models and optimization algorithms

模型及优化算法	参数名称	参数空间	优化算法	参数名称	参数设置
RF	n_estimators	(10, 500)	BO	n_trials	100
	max_depth	(2, 10)		n_startup_trials	20
	min_samples_split	(2, 10)			
SVR	kernel	(rbf,linear,sigmoid)	PSO	swarm_size	50
	c	(0.1, 100.0)		max_iter	100
	epsilon	(0.01, 0.3)		w	0.6
	gamma	(0.01, 1.0)		c1/c2	1.7/1.7
XGB	n_estimators	(10, 500)	ASHA	n_trials	40
	max_depth	(2,10)		step	(1, 10)
	learning_rate	(1e-3, 3e-1)		HyperbandPruner	(min_=1;max=500, reduction_factor=2)
	min_child_weight	(1, 8)			

1.6 模型性能评价指标

在ML模型中，采用多维度的模型性能评估方法有助于全面评估模型性能，防止评价指标单一造成的模型性能误判。在本研究中，为评估所构建的ML模型的性能，采用了拟合度指标、误差度量指标和不确定性评估指标进行模型评估。其中，拟合度指标R^2用于评估模型的拟合程度；误差度量指标均方根误差（Root Mean Square Error, RMSE）与MAE用于衡量模型预测值与实测值之间的偏差和误差程度。为了衡量模型预测结果的不确定性和可靠性，采用了t统计检验（T_{stat}）、95%置信区间的预测误差（U_{95}）。上述评价指标的计算如公式（1）～公式（5）所示。

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

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

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

$$T_{stat} = \sqrt{\frac{(n-1)\times \mathrm{MBE}^2}{\mathrm{RMSE}^2 - \mathrm{MBE}^2}} \tag{4}$$

$$U_{95} = 1.96\sqrt{\mathrm{SD}^2 + \mathrm{RMSD}^2} \tag{5}$$

式中：y_i、$\hat{y}_i$和$\bar{y}$分别为第i个实测值、第i个预测值和所有实测值的平均值；n为样本数量；SD（Standard Deviation）表示蒸腾量实测值与预测值之差的标准差；MBE（Mean Bias Error）与RMSD（Root Mean Square Deviation）分别为平均偏差误差与均方根偏差。

根据JIERULA等[30]的研究结果，当R^2≥0.9时，模型性能为“优秀”，当0.7≤R^2<0.9时为“良好”，当0.5≤R^2<0.7时为“一般”，当R^2<0.5时为“差”；MAE、RMSE越小（最小为0），说明模型预测值与实测值之间的偏差越小，模型性能越好；U_{95}表示约有95%的残差落在可信区间内，因此，U_{95}值越小表示残差离散程度越低、随机误差波动越小，模型预测结果越稳定可靠；T_{stat}可判断模型预测是否存在显著系统偏差，当T_{stat}小于临界t值（$|t|$）时，认为模型平均偏差不显著；反之则说明模型存在显著的高估或低估倾向。在本研究中，设显著性水平为α=0.05，根据双尾检验临界值查表，得出对应的$|t|$约为1.98。

2　结果与讨论

2.1　蒸腾关键因子的筛选与识别

图1展示了研究时段内（2020—2022年）各气象变量的日动态，其中包括了经RF-RFE筛选出的五个关键输入因子（T_a、e_a、VWC、u_2和DOY）。监测期内，T_a在10.1 ℃至38 ℃间波动，均值为24.5 ℃。与T_a相比，e_a表现出与T_a相似的年际变化趋势，其范围为0.84到5.33 kPa，均值为2.38 kPa。u_2维持在0.11至1.64 m/s的较低水平，日均为0.83 m/s。在本研究中，为保证柑橘树度过干旱胁迫阶段，根据土壤干旱情况在试验期间共灌溉6次，降雨是本研究中补充土壤水分的主要来源（图1f）。强降雨过后，VWC会有一个明显的上升，之后便逐渐回落，直到下一次降雨来临；相比之下，小雨对深层土壤的水分补充效果则不太明显。VWC变化范围为11.1%至33.2%，均值为20.2%。蒸腾量变化范围在0.48～3.25 mm/d之间，均值为1.85 mm/d。

图2展示了特征选择前各输入变量与蒸腾量之间的皮尔逊相关系数，众多因子与蒸腾量之间及诸因子之间均存在相关性。随后通过RF-RFE逐步去除重要性最小的特征因子（图3）。如图3a所示，随着特征数量的递减，模型性能在特征数量为5时达到最优（较高的R^2与较低的MAE和RMSE），说明在该点之前去除的是冗余特征，进一步减少特征将造成性能下降，特征数量为5时，模型兼具预测精度与特征空间精简性。图3b进一步展示了RFE最终筛选出的5个关键特征在随机森林模型中的重要性排序，明确了本研究中特征选择后的各特征因子对蒸腾量预测的重要性大小。按重要性由高到低依次为DOY、VWC、T_a、u_2和e_a。

图3b可见，影响柑橘蒸腾的关键因子DOY的贡献性达54%，这表明DOY在本研究中并非只是一个简单的时间标记，而是扮演了一个信息高度集成的代理变量角色。作为关键的时间代理变量，DOY表征了作物内在生理节律与外部环境季节性动态的耦合关系。具体而言，在本研究区，特定的DOY范围分别对应柑橘的春梢萌发、开花坐果及果实膨大等关键生长期。这些物候阶段的生理活动差异巨大：例如，春梢萌发期伴随着新叶的快速生长与扩张（展叶），直接导致冠层总蒸腾面积增加；而在生长旺盛期，为保证光合作用效率，气孔保持高度开放。研究表明，DOY常被用作构建作物LAI动态变化模型的核心变量[31]，在玉米物候监测中DOY能够有效表征关键生长阶段[32]，而在森林生态系统中DOY与LAI呈显著的季节性同步变化[33]。这些研究均证实了DOY作为作物生理状态代理变量的有效性，支持了本研究中DOY高特征重要性的发现。除了整合作物生理信息外，驱动蒸腾的关键气象因子本身也随DOY呈现强烈的季节性周期[34]，本研究也证实了DOY与温度、湿度等因子存在显著相关（图2）。它们蕴含了物候变化、根区水分供应、热力驱动、大气边界层扰动和空气干湿度五个方面，从多维度刻画了蒸腾主控机制。

2.2　不同模型优化的有效性评估

图4展示了不同ML模型预测蒸腾量与实测蒸腾量的散点图。表2和表3分别展示了柑橘园蒸腾量模拟中12种模型在训练集和测试集上的性能指标。在训练集结果中，RF及其优化后的模型均表现出良好的模型性能（R^2>0.7）。在测试集中，与RF模型相比，在优化算法调参后的RF-PSO、RF-BO与RF-ASHA模型均表现出更好的性能。与单一RF模型相比，引入优化算法后的模型拟合精度均有提升。其中，RF-PSO、RF-BO和RF-ASHA模型的R^2分别提高了0.07、0.09和0.12。在误差指标方面（MAE和RMSE），RF模型的MAE、RMSE分别为0.22和0.30，而三种优化后的模型的MAE、

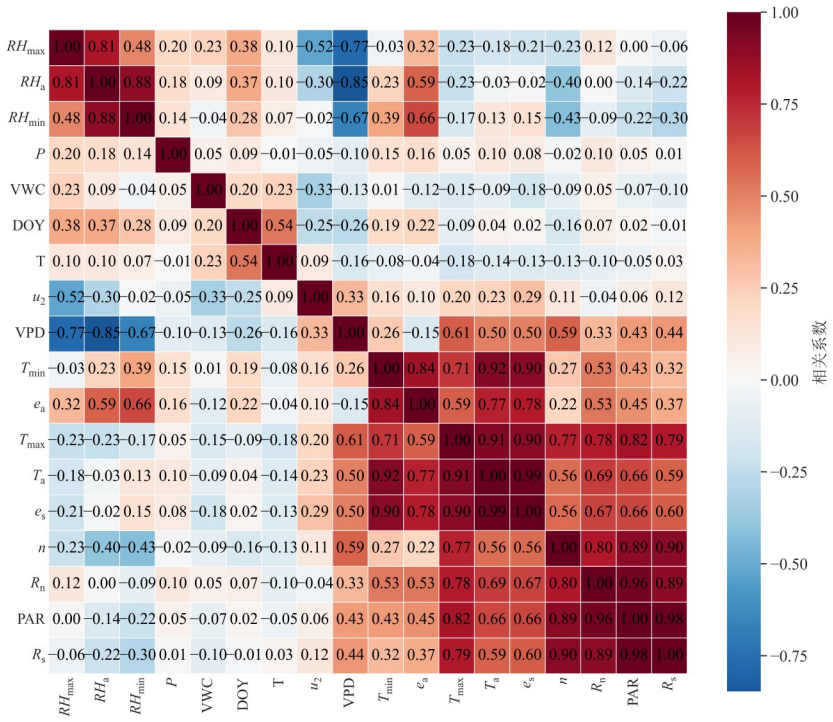
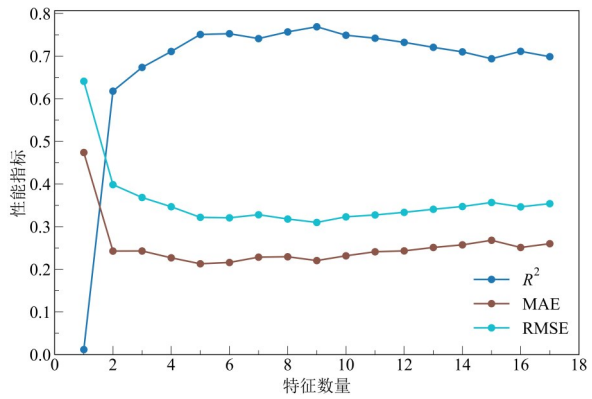
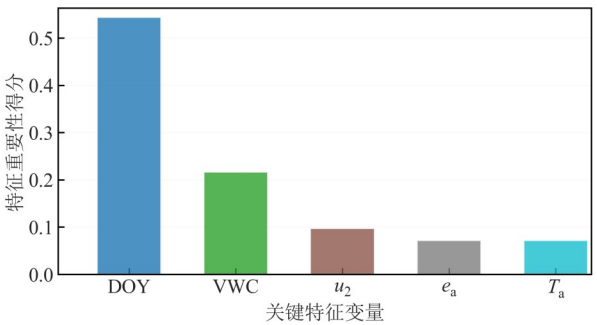


图2 环境因子与蒸腾作用关系的皮尔逊相关系数

Fig. 2 Pearson correlation coefficients between environmental factors and transpiration



a. RF-RFE特征选择过程中最优特征数与模型性能变化



b. 最优特征子集重要性得分

图3 基于RF-RFE方法筛选柑橘日蒸腾量的关键特征及其重要性得分

Fig. 3 Key features and their importance scores for daily citrus transpiration based on the RF-RFE method

RMSE分别在0.15～0.17和0.21～0.25的范围内。上述结果表明，PSO、BO和ASHA优化算法能够有效提升RF模型的预测性能。图4a～图4d实测值与预测值均接近于1∶1拟合线，且图4b～图4d相较于图4a离散程度较小，再次验证了优化算法对RF模型性能的提升。

在SVR模型中，从表2和表3中训练集和测试集结果的对比可以看出，准确性指标和误差性指标的结果在训练集和测试集中差异较小，说明SVR、SVR-PSO、SVR-BO和SVR-ASHA这4个模型在展示出较好的模型性能的同时具有优秀的泛化能力。在训练集与测试集结果的对比上，训练集R^2范围在0.80～0.81之间，测试集为0.79～0.80；SVR原始模型及其3个优化模型在训练集中的MAE和RMSE均为0.18和0.27，测试集的MAE和RMSE范围分别为0.20～0.21和0.28～0.29。图4e～图4h中，SVR及其优化后的SVR模型的训练集和测试集在1∶1拟合线附近的表现与上述结果相符。

对于XGB模型，XGB及其PSO、BO和ASHA优化模型在训练集上均表现出优异的性能（表2）。在测试集中，XGB-ASHA模型在各项指标上均表现最佳。在拟合度指标中，XGB模型R^2=0.78，在经过PSO、BO、ASHA算法调参后的XGB模型R^2提升范围为0.07～0.13，其中XGB-ASHA模型的R^2

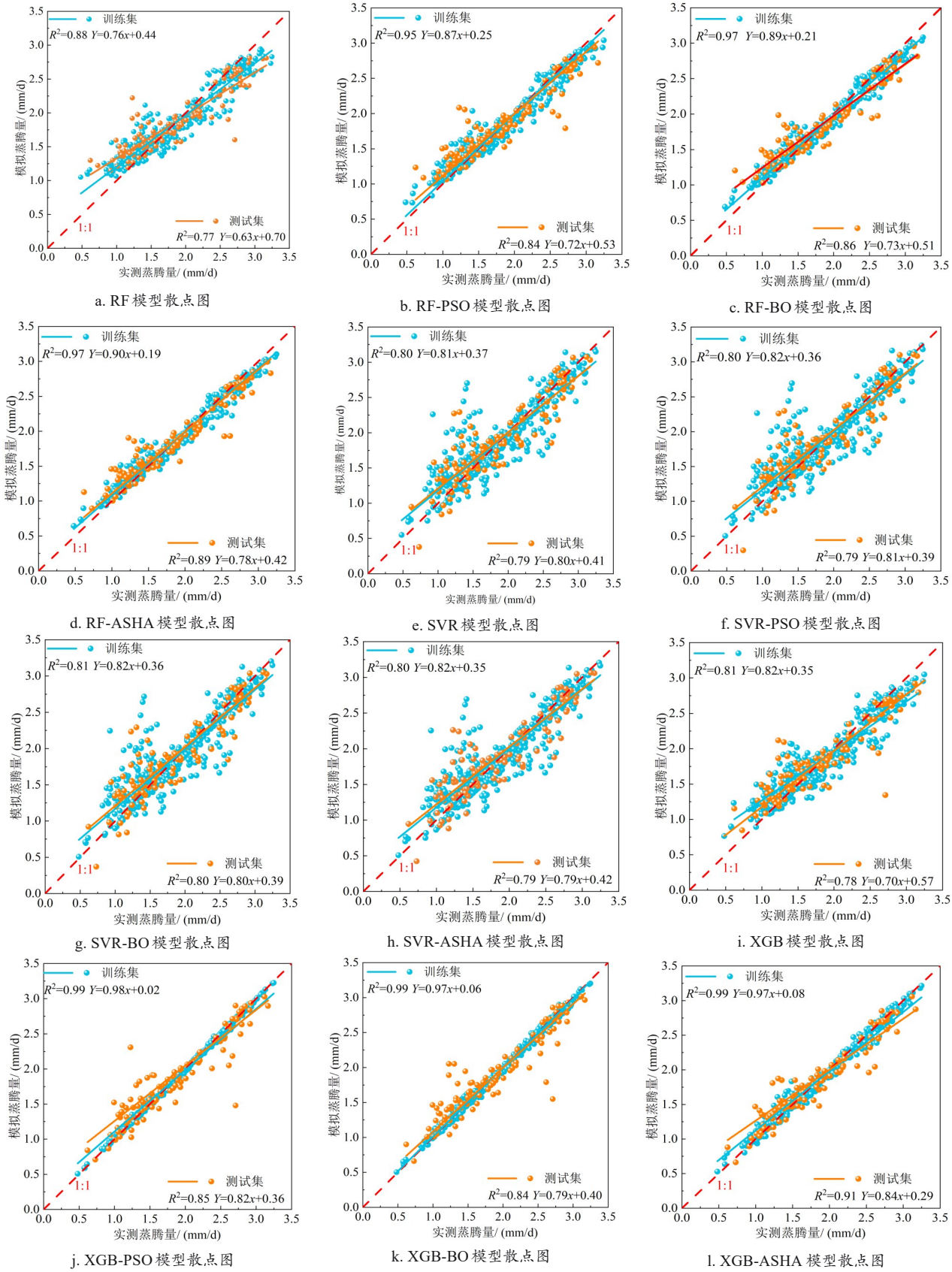


图 4 12 种所构建模型在训练期与测试期柑橘日蒸腾量的估测值与实测值散点图

Fig. 4 Scatter plots of estimated versus observed daily citrus transpiration for the 12 constructed models during the training and testing periods

表2 原始及优化后RF、SVR、XGB模型在训练集的柑橘蒸腾预测性能

Table 2 Performance of original and optimized RF, SVR, and XGB models for citrus transpiration prediction on the training set

模型	R^2	MAE/(mm/d)	RMSE/(mm/d)	T_{stat}	U_{95}
RF	0.88	0.16	0.21	0.24	0.58
RF-PSO	0.95	0.10	0.14	0.38	0.38
RF-BO	0.97	0.08	0.11	0.02	0.31
RF-ASHA	0.97	0.07	0.11	0.02	0.29
SVR	0.80	0.18	0.27	1.61	0.74
SVR-PSO	0.80	0.18	0.27	1.65	0.74
SVR-BO	0.81	0.18	0.27	1.73	0.73
SVR-ASHA	0.81	0.18	0.27	1.05	0.74
XGB	0.88	0.15	0.21	0.12	0.57
XGB-PSO	0.99	0.02	0.02	0.07	0.06
XGB-BO	0.99	0.04	0.05	0.14	0.13
XGB-ASHA	0.99	0.05	0.07	0.04	0.48

表3 原始及优化后RF、SVR、XGB模型在测试集的柑橘蒸腾预测性能

Table 3 Performance of original and optimized RF, SVR, and XGB models for citrus transpiration prediction on the test set

模型	R^2	MAE/(mm/d)	RMSE/(mm/d)	T_{stat}	U_{95}
RF	0.77	0.22	0.30	0.87	0.83
RF-PSO	0.84	0.17	0.25	0.61	0.70
RF-BO	0.86	0.17	0.23	0.96	0.64
RF-ASHA	0.89	0.15	0.21	0.95	0.57
SVR	0.79	0.21	0.29	1.52	0.79
SVR-PSO	0.79	0.21	0.29	1.44	0.79
SVR-BO	0.80	0.20	0.28	1.26	0.77
SVR-ASHA	0.79	0.21	0.29	1.14	0.79
XGB	0.78	0.20	0.29	0.72	0.80
XGB-PSO	0.85	0.15	0.24	0.66	0.67
XGB-BO	0.84	0.17	0.25	0.94	0.69
XGB-ASHA	0.91	0.13	0.19	0.33	0.52

为0.91，高于其他模型；在误差性指标方面，XGB模型的MAE、RMSE分别为0.20和0.29，3种优化后的XGB模型的误差指标范围分别是0.13～0.17和0.19～0.25，其最低值均为XGB-ASHA模型。以上结果说明了优化算法对XGB模型性能的提升是有效的，且其中提升最大的是运用了ASHA算法的XGB模型。

2.3 模型性能的综合评估与优选

由表3可知，在测试集中，XGB-ASHA模型的拟合度指标最高（R^2= 0.91），误差性指标最低（MAE和RMSE分别为0.13和0.19）。然而，单一的拟合度指标无法全面反映模型的稳定性。为此，本研究在评估模型表现时，除了拟合度指标外，还引入了t统计检验（T_{stat}）和不确定性指标（U_{95}），以评估模型的一致性和可靠性，从而更全面地衡量模型的稳定性。表3展示了所应用模型的t统计量和U_{95}指数，可以看到XGB-ASHA模型的t统计量在95%的显著性水平下表现极佳，且其U_{95}指数最小，从而进一步验证了该模型在误差控制和精度上的优越性。此外，与双尾检验下的$|t|$对比，所有模型的t统计指数在95%显著性水平下都小于$|t|$（1.98），说明这些模型对柑橘树蒸腾量的实际值与预测值没有显著差异。

由表3可知，t检验指数最小的三个值分别为0.33、0.61和0.66，对应的模型分别是XGB-ASHA、RF-PSO和XGB-PSO模型，为进一步对比这3个模型的预测误差特征，通过箱线图（图5）展示了3个模型实测值与预测值之间的残差分布，正误差和负误差分别代表偏高预测和偏低预测。在该箱线图中，与XGB-ASHA模型的异常值范围（最大值0.57和最小值−0.58）相比，RF-PSO和XGB-PSO模型的异常值点更多，且分布范围更广，其中RF-PSO和XGB-PSO的异常值最大值分别为0.92和1.23，最小值分别为−0.86和−1.08。这些结果表明，U_{95}指数最小的XGB-ASHA模型，残差离散程度最低，预测误差的波动性相对最小。此外，图6显示了所应用模型的泰勒图。泰勒图是一种用于展示预测结果与观测数据匹配程度的可视化工具[35]。图6通过3个统计量（相关系数（r）、SD和RMSE）来衡量柑橘树蒸腾量预测值与实测值之间的匹配程度，模型所代表的点离观测值越近则说明模型性能越好。可以看到，最接近观测点的前3个模型分别为XGB-ASHA、RF-PSO和XGB-PSO模型。与T_{stat}中所对应的结果一致。其中，性能最佳的为XGB-ASHA模型，通过对比XGB-ASHA和XGB模型在泰勒图中的表现，可以明显看出，XGB-ASHA在相关性、标准差和均方根误差这三个统计指标上均表现出较高的优势，验证了ASHA算法在提升XGB模型性能方面的有效性。在国内外蒸发蒸腾的模拟研究中，与意大利地中海气候柑橘园实际蒸散发预测研究相比[1]，其最优随机森林模型R^2= 0.84、RMSE=0.39 mm/d，并使用7个输入特征；本研究最佳模型在相同评价指标下，R^2=

0.91、RMSE=0.19 mm/d，且本研究仅需4个监测的环境因子（DOY可直接获取），在柑橘园同类任务上表现出明显提升。此外，张晶莹等[31]先前在本研究区域使用Jarvis－Stewart机理模型并采用贝叶斯优化对柑橘蒸腾进行了模拟，但该传统模型不仅所需参数众多，且最佳模型性能也远低于本研究（R^2=0.69），证明了本研究中所使用建模方法的优越性。需要指出的是，本研究基于干热河谷区单一区域与特定作物样本构建模型，其在跨区域、跨作物情景下的蒸腾量模拟性能尚待验证。未来可结合蒸散过程机理，引入物理约束的深度学习方法，构建机理与数据融合的灰箱模型，以提升模型的可解释性与泛化能力。总的来说，准确估算作物蒸腾量对于实现作物的高效用水和提高产量至关重要，本研究推荐在干热河谷区使用XGB-ASHA模型对柑橘树蒸腾量进行模拟。

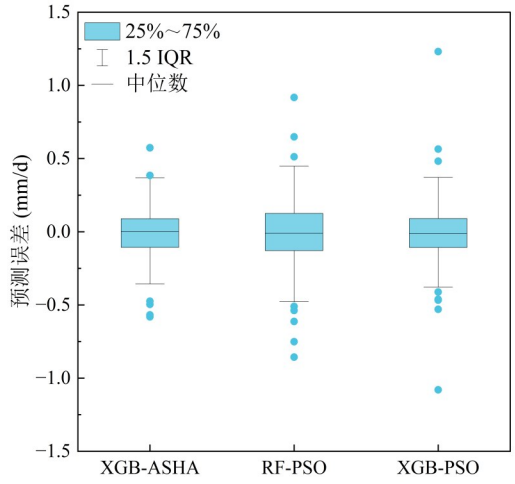


图5　XGB-ASHA、RF-PSO和XGB-PSO模型蒸腾量真实值与预测值的预测误差比较

Fig. 5　Comparison of prediction errors between measured and predicted transpiration values for XGB-ASHA, RF-PSO, and XGB-PSO models

3　结　论

面向干热河谷柑橘园的智慧化灌溉需求，本研究构建了基于特征优选的机器学习蒸腾量模拟模型。采用RF-RFE将17个候选因子精简为5个关键变量（T_a、u_2、VWC、e_a、DOY），在保持模型性能的同时显著降低数据采集成本。其中，DOY的特征重要性为54%，能够综合表征季节性气候与作物物候信息，部分替代难以连续监测的生理参数。对三类原始ML模型及其与三种超参数优化策略耦合的12组方案进行系统比较表明，XGB-ASHA在测试集上表现最佳（R^2=0.91，RMSE=0.19 mm/d），整体优于其他组合。基于此，本研究提出了一种融合特征优选与模型优化的建模策略。该策略仅需4个易获取的环境变量（T_a、u_2、VWC、e_a）及DOY，即可实现对蒸腾量的高精度模拟。在减少输入变量数量和数据需求的同时，该策略还可与灌溉阈值设定、分区配水及用水预警等环节相衔接，为干热河谷柑橘园智慧灌溉提供技术支撑和决策依据。

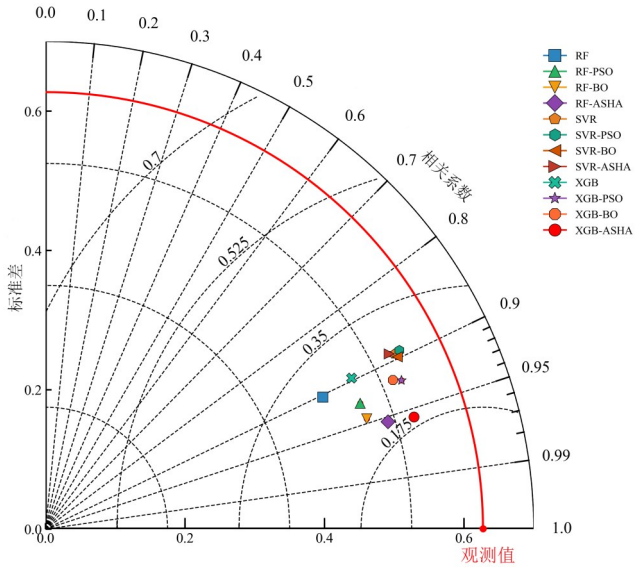


图6　12种模型对柑橘日蒸腾量的模拟的泰勒图

Fig. 6　Taylor diagram of the 12 models' simulations of daily citrus transpiration

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] PAGANO A, AMATO F, IPPOLITO M, et al. Machine learning models to predict daily actual evapotranspiration of *Citrus* orchards under regulated deficit irrigation[J]. Ecological Informatics, 2023, 76: 102133.

[2] STOY P C, EL-MADANY T S, FISHER J B, et al. Reviews and syntheses: Turning the challenges of partitioning ecosystem evaporation and transpiration into opportunities[J]. Biogeosciences, 2019, 16(19): 3747-3775.

[3] GAO Z Y, SHI W J, WANG X, et al. Non-rainfall water contributions to dryland jujube plantation evapotranspiration in the Hilly Loess Region of China[J]. Journal of Hydrology, 2020, 583: 124604.

[4] ZHOU S, YU B F, ZHANG Y, et al. Partitioning evapotranspiration based on the concept of underlying water use efficiency[J]. Water Resources Research, 2016, 52(2): 1160-1175.

[5] FAN J L, ZHENG J, WU L F, et al. Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models[J]. Agricultural Water Management, 2021, 245: 106547.

[6] LI W Y, TARTAKOVSKY D M. Fast and accurate estimation of evapotranspiration for smart agriculture[J]. Water Resources Research, 2023, 59(4): e2023WR034535.

[7] 张菊, 江朝晖, 李博, 等. 基于热扩散探针的便携式植物液流监测仪研究[J]. 农业机械学报, 2020, 51(7): 237-243.
ZHANG J, JIANG ZH, LI B, et al. Design and experiment of portable plant sap flow meter based on TDP[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(7): 237-243.

[8] SONG Q F, XIAO H, XIAO X L, et al. A new canopy photosynthesis and transpiration measurement system (CAPTS) for canopy gas exchange research[J]. Agricultural and Forest Meteorology, 2016, 217: 101-107.

[9] TEZZA L, HÄUSLER M, CONCEIÇÃO N, et al. Measuring and modelling soil evaporation in an irrigated olive orchard to improve water management[J]. Water, 2019, 11(12): 2529.

[10] WEN X F, YANG B, SUN X M, et al. Evapotranspiration partitioning through *in-situ* oxygen isotope measurements in an oasis cropland[J]. Agricultural and Forest Meteorology, 2016, 230/231: 89-96.

[11] DING J Y, YIN J L, ENGEL B A, et al. Estimating cropland evapotranspiration based on remote sensing models: A global meta-analysis[J]. European Journal of Agronomy, 2025, 170: 127758.

[12] 张宇, 赵春江, 林森, 等. 基于Penman-Monteith模型和路径排序算法相结合的草莓灌溉方法与验证[J]. 智慧农业(中英文), 2021, 3(3): 116-128.
ZHANG Y, ZHAO C J, LIN S, et al. Irrigation method and verification of strawberry based on penman-monteith model and path ranking algorithm[J]. Smart Agriculture, 2021, 3(3): 116-128.

[13] 高冠龙, 冯起, 刘贤德. 基于改进的双源模型模拟荒漠河岸胡杨林蒸散发[J]. 生态学报, 2020, 40(10): 3462-3472.
GAO G L, FENG Q, LIU X D. Simulating evapotranspiration of the desert riparian *Populus euphratica* Olive. forest based on an improved dual-source model[J]. Acta Ecologica Sinica, 2020, 40 (10): 3462-3472.

[14] 桂泽春, 赵思健. 人工智能在农业风险管理中的应用研究综述[J]. 智慧农业(中英文), 2023, 5(1): 82-98.
GUI Z C, ZHAO S J. Research application of artificial intelligence in agricultural risk management: A review[J]. Smart Agriculture, 2023, 5(1): 82-98.

[15] HUANG J J, ZHANG S, ZHANG J H, et al. Integrating meteorological and remote sensing data to simulate cropland nocturnal evapotranspiration using machine learning[J]. Sustainability, 2024, 16(5): 1987.

[16] ZHANG J, YANG S S, WANG J W, et al. Evapotranspiration partitioning for croplands based on eddy covariance measurements and machine learning models[J]. Agronomy, 2025, 15(3): 512.

[17] XING L W, CUI N B, LIU C W, et al. Estimation of daily apple tree transpiration in the Loess Plateau region of China using deep learning models[J]. Agricultural Water Management, 2022, 273: 107889.

[18] KABALA J P, MASSARI C, NICCOLI F, et al. Reconstruction of the dynamics of sap-flow timeseries of a beech forest using a machine learning approach[J]. Agricultural and Forest Meteorology, 2025, 362: 110379.

[19] GE J K, ZHAO L F, YU Z H, et al. Prediction of greenhouse tomato crop evapotranspiration using XGBoost machine learning model[J]. Plants, 2022, 11(15): 1923.

[20] DANGARE P, MASHIMBYE Z E, CRONJE P J R, et al. Evapotranspiration partitioning in selected subtropical fruit tree orchards based on Sentinel-2 data using a light gradient-boosting machine (LightGBM) learning model in Malelane, South Africa [J]. Hydrology, 2025, 12(7): 189.

[21] CANTON H. Food and Agriculture Organization of the United Nations—FAO[M]// 23rd ed. London: Routledge, 2021: 297-305.

[22] CHEN D Y, HU X T, DUAN X W, et al. Incorporating dynamic schemes of canopy light extinction coefficient improves transpiration model performance for fruit plantations[J]. Journal of Hydrology, 2023, 627: 130397.

[23] CHEN D Y, WANG Y K, WANG X, et al. Effects of branch removal on water use of rain-fed jujube (*Ziziphus jujuba* Mill.) plantations in Chinese semiarid Loess Plateau region[J]. Agricultural Water Management, 2016, 178: 258-270.

[24] GROSSIORD C, BUCKLEY T N, CERNUSAK L A, et al. Plant responses to rising vapor pressure deficit[J]. The New Phytologist, 2020, 226(6): 1550-1566.

[25] ZHONG Y, HUANG Z K, HAO K, et al. Effects of water deficit on fruit quality and water productivity of Citrus under plastic film mulching in Western Hubei, China[J]. Frontiers in Plant Science, 2025, 16: 1498798.

[26] 陈绍民, 李晓丽, 杨启良, 等. 基于机器学习的遮荫设施内参考作物蒸散量估算[J]. 农业工程学报, 2022, 38(11): 108-116.
CHEN S M, LI X L, YANG Q L, et al. Estimation of reference evapotranspiration in shading facility using machine learning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(11): 108-116.

[27] IMANI M, BEIKMOHAMMADI A, ARABNIA H R. Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels[J]. Technologies, 2025, 13(3): 88.

[28] 商志根, 段小汇. 基于PSO与LS-SVM的作物需水量预测[J]. 计算机与现代化, 2018(10): 44-47.
SHANG Z G, DUAN X H. Predicting crop water requirements based on particle swarm optimization and least square support vector machine[J]. Computer and Modernization, 2018(10): 44-47.

[29] AACH M, SARMA R, NEUKIRCHEN H, et al. Resource-adaptive successive doubling for hyperparameter optimization with large datasets on high-performance computing systems[J]. Future Generation Computer Systems, 2026, 175: 108042.

[30] JIERULA A, WANG S H, OH T M, et al. Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data [J]. Applied Sciences, 2021, 11(5): 2314.

[31] 张晶莹, 陈滇豫, 马永胜, 等. 基于贝叶斯分析的干热河谷区橙子林冠层蒸腾耗水模拟[J]. 农业机械学报, 2024, 55(1): 305-315.
ZHANG J Y, CHEN D Y, MA Y S, et al. Canopy transpiration water consumption simulation of orange forest in dry and hot valley area based on Bayesian analysis[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(1): 305-315.

[32] GUO Y H, XIAO Y, LI M W, et al. Identifying crop phenology using maize height constructed from multi-sources images[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 115: 103121.

[33] SHABANOV N V, EGOROV V A, MIKLASHEVICH T S, et al. Utility of leaf area index for monitoring phenology of Russian forests[J]. Remote Sensing, 2023, 15(22): 5419.

[34] LIU W H, ZHANG B Z, WEI Z, et al. Heterogeneity analysis of main driving factors affecting potential evapotranspiration changes across different climate regions[J]. Science of the Total Environment, 2024, 912: 168991.

[35] SÁENZ J, CARRENO-MADINABEITIA S, ESNAOLA G, et al. The Sailor diagram: A new diagram for the verification of two-dimensional vector data from multiple models[J]. Geoscientific Model Development, 2020, 13(7): 3221-3240.

Estimation of Citrus Transpiration in a Savanna Valley Based on Feature Selection and Optimization Algorithms

WU Mingqing[1,2,3], LI Weikang[1,2,3], WANG Jing[1,2,3], LI Jing[1,2,3], PEI Rentao[1], HUANG Haitao[1,2], QIU Huanghuang[1,2], LEI Chujing[1,2], ZHAO Duo[1], HUANG Jingtao[4], GAO Zhiyong[1,2,3,5*]

（1. *College of Water Conservancy*, *Yunnan Agricultural University*, *Kunming* 650201, *China*; 2. *Yunnan International Joint R&D Center of Smart Agriculture and Water Security*, *Kunming* 650201, *China*; 3. *Dianchi Lake Ecosystem Observation and Research Station of Yunnan Province*, *Kunming* 650228, *China*; 4. *College of Architecture and Civil Engineering*, *Yunnan Agricultural University*, *Kunming* 650201, *China*; 5. *Luliang Mountain Basin Land Use Field Scientific Observation Station of Yunnan Province*, *Qujing* 655600,*China*）

Abstract: [Objective] As the core physiological process determining crop water demand, transpiration profoundly influences irrigation-district water regulation and allocation strategies. During the dry season, water stress is particularly acute in the Yunnan savanna region. For smart, sensor-driven irrigation management, reliable orchard-scale transpiration estimation is essential for real-time irrigation scheduling and water allocation. However, traditional mechanistic models face challenges such as difficulties in parameter acquisition, high sensitivity to specific parameters, and limited applicability in complex terrains. To address these issues in the development of smart irrigation districts, a high-precision citrus transpiration estimation model was constructed with low parameter dependence. **[Methods]** A citrus orchard with 10-year-old trees in the savanna region was selected as the study site. Three years of stem sap flow observations and synchronous environmental-factor data were collected. Sap flow was accumulated to the daily scale and used to estimate field-scale transpiration. A Random Forest–based wrapper feature selection method was applied to identify an optimal subset of key predictors from 16 original environmental parameters and the day of year (DOY). Specifically, the wrapper method iteratively removed the variable with the lowest feature importance and rebuilt the model, continuing until only one variable remained. During this process, model performance under different numbers of variables was recorded for each iteration, and the best feature subset was determined based on performance. In addition, a heatmap was used to quantify linear correlations among variables to better understand redundant parameters removed by the model. After identifying the optimal feature combination, the built-in feature importance method of Random Forest was used for importance evaluation. Based on these parameters, extreme gradient boosting (XGB), support vector regression (SVR), and random forest (RF) models were established. Particle swarm optimization (PSO), bayesian optimization (BO), and the asynchronous successive halving algorithm (ASHA) were introduced for hyperparameter tuning, resulting in 12 simulation models in total. The dataset was split into 80% for training and 20% for validation. During training, 10-fold cross-validation was adopted, and mean absolute error (MAE) was used as the fitness function. Model performance was comprehensively evaluated using the coefficient of determination (R^2), root mean square error (RMSE), MAE, and uncertainty metrics. **[Results and Discussions]** The RF wrapper feature selection reduced the 16 meteorological input parameters to four: actual vapor pressure (e_a), soil water content (VWC), wind speed at 2 m (u_2), and mean air temperature (T_a). In addition, the DOY was readily obtainable and was identified as the most critical feature, effectively characterizing the unique seasonal climatic variations of the Savanna Valley and citrus phenological characteristics; meanwhile, e_a, VWC, T_a, and u_2 mainly characterize short-term atmospheric evaporative demand and soil moisture constraints. While substantially reducing data requirements, the proposed method still achieved high prediction accuracy. Among all model combinations, XGB-ASHA performed best, reaching R^2=0.91 and RMSE=0.19 mm/d for daily-scale transpiration prediction. Hyperparameter optimization improved model robustness, further reducing RMSE and MAE compared with non-optimized baselines. Compared with commonly used mechanistic models in the region, the proposed machine-learning approach does not require complex parameter calibration, is less sensitive to input errors, and shows better application feasibility. Uncertainty evaluation further indicates that this framework can provide not only accurate point predictions but also reliable confidence information for operational decision-making. **[Conclusions]** The proposed model requires monitoring only four environmental parameters to achieve high-accuracy, operational daily-scale transpiration estimation for citrus orchards in the savanna region. The optimal configuration, XGB-ASHA, is suitable for embedding into smart irrigation systems, supporting water-saving irrigation scheduling and irrigation-district-scale water management. Future work will integrate multi-source data and explore more advanced learning models to further improve transferability and generalization across orchards and years.

Key words: transpiration water consumption; machine learning; feature selection; smart agriculture; water resource management

Foundation items: Yunnan Province Agricultural Fundamental Research Joint Special Project (202301BD070001-181); National Key Research and Development Program of China (2023YFD1901203)

First authors: 1. WU Mingqing, E-mail: Wmqing1015@163.com; 2. LI Weikang E-mail: liw1328131531@163.com

***Corresponding author:** GAO Zhiyong, E-mail: GZYstruggling@163.com