doi:10.6041/j.issn.1000-1298.2017.03.038

基于随机森林算法的参考作物蒸发蒸腾量模拟计算

王　升　付智勇　陈洪松　丁亚丽　吴丽萍　王克林
(中国科学院亚热带农业生态研究所, 长沙 410125)

摘要: 选取西南喀斯特地区 4 个气象站点(都安、河池、百色和融安)5 a(2008—2012 年)的逐日气象数据,包括日最高气温 T_{max}、日最低气温 T_{min}、相对湿度 R_H、日照时数 n 和风速 u_2 这 5 个气象因子的不同组合作为输入,并以 FAO 56 Penman－Monteith 法(FAO P－M)的计算结果作为标准值,建立基于随机森林(Random forest, RF)算法和基因表达式编程(Gene expression programming, GEP)算法的 ET_0 模型,并将模拟结果与传统 Hargreaves 模型的计算结果进行比较。结果表明,不同气象因子组合下建立的 RF 模型均能较好地反映气象因子与 ET_0 之间的非线性关系。随着气象因子的增加,RF 模型模拟的精度随之提高。在仅有气温数据时,RF 模型仍具有足够的精度(R^2 为 0.875,R_{MSE} 为 0.546 mm/d),与传统 Hargreaves 模型相比 R^2 平均增加了 1.98%, R_{MSE} 平均减小了 22.88%,因此在仅有气温数据时可用 RF 模型代替 Hargreaves 模型。RF 算法对气象因子的重要性评估表明,在该区域对 ET_0 最重要的气象因子依次为 T_{max}、n、T_{min}、R_a、R_H 和 u_2。相同气象因子输入下,RF 模型精度高于 GEP 模型。

关键词: 参考作物蒸发蒸腾量; 随机森林; 基因表达式编程; Penman－Monteith 模型; 西南喀斯特地区
中图分类号: S161.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2017)03-0302-08

Simulation of Reference Evapotranspiration Based on Random Forest Method

WANG Sheng　FU Zhiyong　CHEN Hongsong　DING Yali　WU Liping　WANG Kelin
(*Institute of Subtropical Agriculture*, *Chinese Academy of Sciences*, *Changsha* 410125, *China*)

Abstract: Accurate estimation of reference evapotranspiration (ET_0) is very important in hydrological cycle research, and it is also essential in agricultural water management and allocation. Using less meteorological parameters to estimate ET_0 is necessary in areas with limited data. The ability of random forest (RF) and gene expression programming (GEP) algorithm in modeling ET_0 was investigated and compared by using fewer meteorological parameters collected from four weather stations of Duan, Hechi, Baise and Rong'an, in karst region of southwest China, over a five-year period (2008—2012). Daily climatic data of the four stations, including maximum temperature (T_{max}), minimum temperature (T_{min}), sunshine duration (n), relative humidity (R_H) and wind speed (u_2) were employed to model ET_0 by using FAO 56 Penman－Monteith equation as the reference, and their performances were evaluated using determination coefficient (R^2) and root mean square error (R_{MSE}). From the statistical results, the derived RF-based (R^2 was ranged from 0.809 to 0.991, and R_{MSE} was ranged from 0.158 mm/d to 0.678 mm/d) and GEP-based (R^2 was in range of 0.830 ~ 0.977, and R_{MSE} was in range of 0.225 ~ 0.645 mm/d) ET_0 models were successfully applied to model ET_0 with different input combinations. When only the temperature data can be used, the RF models produced satisfactory results (R^2 = 0.875, R_{MSE} = 0.546 mm/d), which can be used as an alternative to the conventional Hargreaves model. The relative importance of meteorological variables for ET_0 can be assessed by RF method, the order of the relative importance of meteorological variables was: T_{max}, n, T_{min}, R_a, R_H and u_2. In most cases, the RF models were found to perform better than the GEP models. The results were expected to be useful to guide rehabilitation strategies and agricultural water management in karst region of Southwest China.

Key words: reference evapotranspiration; random forest; gene-expression programming; Penman－Monteith model; karst region of Southwest China

收稿日期: 2016－07－18　修回日期: 2016－09－07
基金项目: 国家重点基础研究发展计划(973 计划)项目(2015CB452703)和国家自然科学基金项目(41171187、31100294)
作者简介: 王升(1987—),男,博士生,主要从事坡地水文研究,E-mail: hjdx@ foxmail.com
通信作者: 陈洪松(1974—),男,教授,博士生导师,主要从事生态水文研究,E-mail: hbchs@ isa.ac.cn

引言

参考作物蒸发蒸腾量（Reference evapotranspiration，ET_0）是表征大气蒸散发能力的因子，主要用于计算作物需水量，也是灌溉制度设计、水资源管理、流域水量平衡研究关键参数[1]。目前，ET_0计算的标准方法是世界粮食和农业组织（FAO）推荐采用的的 FAO 56 Penman－Monteith 模型（以下简称 FAO P－M），该模型综合了辐射项和空气动力学项，具有充分的理论基础，已经在世界各地、各种气候类型下通过蒸渗仪进行了验证，它也常被应用于校准其他 ET_0模型的参数[1-2]。然而 FAO P－M 需要较为完备的气象数据（太阳辐射、气温、风速和相对湿度），即使在发达国家能同时测量这些气象数据的气象站点也很有限，而且其对数据质量有严格的要求，所以它的应用受到一定限制[3]。其他需要较少气象参数的传统 ET_0经验/半经验模型（蒸发皿法、基于温度或辐射的方法等）准确度较低，如 DJAMAN 等[4]评估了 16 种 ET_0模型，发现这些方法高估或低估了 ET_0，因此在使用时需要根据具体研究区进行参数校正[5]。发展利用较少气象因子得到足够精度 ET_0的模型仍然是一个值得研究的课题。

近年来随着计算能力的大幅提高和大数据时代的崛起，各种机器学习方法被广泛应用于多个领域[6]，也被应用于 ET_0模拟计算，为资料缺乏地区 ET_0计算提供了新途径。如人工神经网络（Artificial neural network，ANN）[7]，然而 ANN 容易过拟合且收敛速度慢。SHIRI 等[2]研究表明基因表达式编程算法（GEP）的 ET_0模拟结果优于自适应模糊推理系统（ANFIS）、Priestley－Taylor 法和 Hargreaves 法。WEN 等[8]和侯志强等[9]研究了支持向量机（SVM）模拟 ET_0，冯禹等[10]用极限学习机模拟川中丘陵区 ET_0，均取得较好的效果。FDRNANDEZ-DELGADO 等[11]评估了 179 种机器学习算法在 121 个数据集上的性能，结果表明随机森林（Random forest，RF）算法的性能最好，其次是 SVM（采用高斯核函数）。随机森林是一种组合式机器学习方法，其通过对大量分类树的汇总提高了模型的预测精度，因其具有更高的准确性和稳健性而在各行业得到越来越多的应用[12]，如用于构建小麦叶片叶绿素相对含量的遥感反演模型[13]、苹果树冠叶面积指数估测模型[14]等。而目前应用随机森林算法模拟 ET_0的研究较少。

中国西南喀斯特地区由于特殊的地质背景和强烈的岩溶作用，以及近代人类不合理的土地开发利用，导致植被退化、水土流失严重，石漠化不断加剧[15-16]。尽管该地区降雨充沛（年平均降水量大于 1 200 mm），但由于地表地下的二元结构发育，土层浅薄且不连续，土壤入渗能力强，使得土壤储水能力低[17]。且喀斯特系统普遍存在溶洞、溶沟、溶隙、漏斗和落水洞，使得该地区水文过程变化迅速，地表水漏失严重，形成了喀斯特小生境特殊的岩溶干旱现象，因此水分亏缺依然是西南喀斯特石漠化地区植被恢复和重建的主要限制性因子[18]。在全球气候变化大背景下，近 60 a 来该地区降水量呈下降趋势（－1.14 mm/a），极端气候事件频繁[19]。准确估算该地区 ET_0有助于指导制定植被恢复策略以及合理开发利用水资源。因此，本文使用桂西北喀斯特地区 4 个气象站点 5 a（2008—2012 年）的逐日气象数据，采用不同的气象因子组合，以 FAO P－M 法的结果为标准，比较基于 RF、GEP 算法的 ET_0模型以及传统 Hargreaves 模型对 ET_0的模拟效果，旨在探讨随机森林算法模拟计算 ET_0的可行性，以期获得使用较少气象参数而计算精度接近于 FAO P－M 且高于传统经验公式的 ET_0模型，为该区域植被恢复重建和农业用水管理提供科学依据。

1　材料与方法

1.1　数据来源

气象数据来自桂西北喀斯特地区的 4 个气象站点：百色、都安、河池和融安，气象要素包括：日最高气温 T_{max}、日最低气温 T_{min}、相对湿度 R_H、日照时数 n 和 2 m 高度处风速 u_2的 5 a 逐日数据（2008—2012 年），如表 1 所示。因为 ET_0与其他气候因素相比变化程度较小，由 5 a 的逐日 ET_0对 ET_0模型进行建立及检验是可接受的[2]。

表 1　气象站点位置及气象因素平均值

Tab.1　Weather station locations and climatic data averages

站点	纬度（N）/（°）	经度（E）/（°）	海拔高度/m	平均值（2008—2012 年）					
				T_{min}/℃	T_{max}/℃	T_{mean}/℃	u_2/（m·s^{-1}）	R_H/%	n/h
百色	23.90	106.60	173.5	18.7	27.6	22.1	1.5	72.5	1 543.1
都安	23.93	108.10	170.8	18.9	26.0	21.6	2.6	75.8	1 380.0
河池	24.70	108.03	260.2	17.9	25.2	20.4	1.6	73.8	1 310.3
融安	25.22	109.40	121.3	16.4	24.5	19.4	1.4	75.1	1 417.0

1.2 参考作物腾发量计算模型

1.2.1 FAO P-M 模型

FAO P-M 模型是计算 ET_0 的标准方法[1],其表达式为

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \tag{1}$$

式中 R_n——太阳净辐射通量,MJ/(m²·d)

G——土壤热通量,MJ/(m²·d)

Δ——饱和水汽压-温度曲线的斜率,kPa/℃

γ——湿度计常数,kPa/℃

e_s——饱和水汽压,kPa

e_a——实际水汽压,kPa

T——2 m 高度处的平均气温,℃

1.2.2 Hargreaves 模型

Hargreaves 模型仅需要日最低和最高气温,其表达式为

$$ET_0 = 0.0023 R_a \left(\frac{T_{\max} + T_{\min}}{2} + 17.8 \right) \sqrt{T_{\max} - T_{\min}} \tag{2}$$

式中 R_a——大气顶层太阳辐射,由气象站点纬度及日序数算得[20],mm/d

1.3 随机森林模型概述

RF 是由 BREIMAN 等[21]在 2001 年提出的一种集成学习算法,具有需要调整的参数较少、不易过拟合、能有效处理大数据集并且可以给出变量的重要性估计等特点[12]。RF 通过多次 bootstrap 抽样获得多个随机样本,然后使用这些样本建立相对应的决策树,从而构成随机森林用于分类和回归分析。对于回归问题,则是由这些树的结果的平均值得到因变量的预测值。用 RF 算法进行回归模拟有 2 个参数需要确定:每个树节点随机变量的数量(m_{try})和森林中树的数量(n_{tree})。假定原始数据有 m 个变量,对于回归问题,通常取 $m_{try} = m/3$。随着随机森林中决策树的数量的增加,森林的总误差率会趋向一个稳定的有限上界,YANG 等[22]发现 n_{tree} 的默认取值(500)不足以产生稳定的结果,因此本文取 $n_{tree} = 2\,000$。

RF 模型判定变量重要性的方法是在每一棵决策树的变量中加入随机噪声,然后检验袋外误差的增减,如果误差增加,则该变量比较重要,反之则不重要[25]。计算方法为

$$I_{VIM_i} = \frac{\sum (E_{errOOB2} - E_{errOOB1})}{n_{tree}} \tag{3}$$

式中 I_{VIM_i}——变量 i 的重要性

$E_{errOOB1}$——袋外数据(Out of bag,OOB)误差

$E_{errOOB2}$——随机对袋外数据 OOB 所有样本的变量 i 加入噪声干扰,再次计算的袋外误差

1.4 基因表达式编程算法概述

关于基因表达式编程算法结构及其模拟计算 ET_0 过程详见文献[3]。

1.5 评价指标

采用 2 个评价指标,即决定系数 R^2 和均方根误差 R_{MSE}。其计算式分别为

$$R^2 = \frac{\left[\sum_{i=1}^{N} (y_i - \bar{y})(y_i' - \bar{y}') \right]^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2 \sum_{i=1}^{N} (y_i' - \bar{y}')^2} \tag{4}$$

$$R_{MSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - y_i')^2}{N}} \tag{5}$$

式中 y_i、y_i'——FAO P-M 和其他模型计算的 ET_0 值

$\bar{y}$、$\bar{y}'$——y_i 和 y_i' 的平均值

N——检验模型时的天数,取 731 d

2 结果与分析

将 4 个站点 5 a 的气象数据分为 2 部分,其中 2008—2010 年的日气象资料及 FAO P-M 模型计算的 ET_0 为训练样本,用 2011—2012 年的日气象资料及 FAO P-M 模型计算的 ET_0 为检验样本,分别采用 RF 和 GEP 算法,得到不同气象因子组合下的 ET_0 模型。4 个气象因子(气温、风速、相对湿度和日照时数)中,气温是各个气象站的常规观测项目,而相对湿度、风速和日照时数只有较少的站点能同时观测[23]。因此气温($T_{\max}$ 和 $T_{\min}$)应用于所有输入组合中。基于 15 个气象因子组合方案,分别建立 15 个基于 RF 和 GEP 算法的模型,输入的气象因子组合及模型精度如表 2、3 所示。

2.1 检验期不同气象因子输入组合下 RF 模型计算结果对比

由表 2 可见(Har 表示 Hargreaves 模型),4 个站点 R^2 的变化范围为 0.809 ~ 0.991,R_{MSE} 的变化范围为 0.168 ~ 0.678 mm/d。RF1 为随机森林温度模型,其输入因子仅为 $T_{\max}$ 和 $T_{\min}$,能够达到足够精度,R^2 平均值为 0.842,R_{MSE} 平均值为 0.603 mm/d。与传统 Hargreaves 模型相比(R^2 = 0.858,R_{MSE} = 0.708 mm/d),精度差异不大,而且从 R_{MSE} 角度来看,RF1 的精度高于传统 Hargreaves 模型。

TRAORE 等[24]研究表明仅利用气温数据不足以得到足够准确的 ET_0 计算结果,增加额外的气象

表 2　不同气象因子组合下验证期 RF 模型精度

Tab. 2　Statistical performance of RF models and Hargreaves model during test period

模型	输入参数	R^2				R_{MSE}/(mm·d^{-1})			
		都安	河池	百色	融安	都安	河池	百色	融安
RF1	T_{min}、T_{max}	0.809	0.834	0.857	0.869	0.678	0.617	0.583	0.533
RF2	T_{min}、T_{max}、R_H	0.914	0.906	0.901	0.912	0.459	0.456	0.485	0.438
RF3	T_{min}、T_{max}、u_2	0.848	0.874	0.908	0.907	0.617	0.529	0.478	0.455
RF4	T_{min}、T_{max}、n	0.906	0.918	0.922	0.938	0.493	0.451	0.434	0.368
RF5	T_{min}、T_{max}、R_H、u_2	0.927	0.918	0.924	0.923	0.436	0.441	0.435	0.414
RF6	T_{min}、T_{max}、u_2、n	0.931	0.948	0.959	0.957	0.437	0.352	0.329	0.314
RF7	T_{min}、T_{max}、R_H、n	0.948	0.950	0.946	0.953	0.366	0.342	0.364	0.323
RF8	T_{min}、T_{max}、R_a	0.827	0.876	0.892	0.903	0.656	0.554	0.514	0.461
RF9	T_{min}、T_{max}、R_a、R_H	0.925	0.925	0.921	0.936	0.434	0.414	0.438	0.381
RF10	T_{min}、T_{max}、R_a、u_2	0.852	0.887	0.919	0.918	0.624	0.517	0.460	0.432
RF11	T_{min}、T_{max}、R_a、n	0.921	0.957	0.956	0.967	0.460	0.362	0.337	0.271
RF12	T_{min}、T_{max}、R_a、R_H、u_2	0.935	0.928	0.933	0.939	0.422	0.415	0.414	0.374
RF13	T_{min}、T_{max}、R_a、u_2、n	0.947	0.973	0.981	0.979	0.399	0.283	0.251	0.225
RF14	T_{min}、T_{max}、R_a、R_H、n	0.966	0.977	0.973	0.980	0.310	0.251	0.270	0.218
RF15	T_{min}、T_{max}、R_a、R_H、u_2、n	0.984	0.990	0.991	0.990	0.218	0.170	0.168	0.158
Har	T_{min}、T_{max}、R_a	0.814	0.868	0.876	0.875	0.693	0.687	0.754	0.697

表 3　不同气象因子组合下验证期 GEP 模型精度

Tab. 3　Statistical performance of GEP models and Hargreaves model during test period

模型	输入参数	R^2				R_{MSE}/(mm·d^{-1})			
		都安	河池	百色	融安	都安	河池	百色	融安
GEP1	T_{min}、T_{max}	0.830	0.849	0.872	0.877	0.646	0.580	0.554	0.517
GEP2	T_{min}、T_{max}、R_H	0.906	0.908	0.899	0.917	0.485	0.456	0.496	0.423
GEP3	T_{min}、T_{max}、u_2	0.799	0.878	0.918	0.900	0.696	0.511	0.450	0.469
GEP4	T_{min}、T_{max}、n	0.895	0.919	0.916	0.924	0.511	0.442	0.450	0.415
GEP5	T_{min}、T_{max}、R_H、u_2	0.889	0.911	0.911	0.924	0.523	0.440	0.474	0.407
GEP6	T_{min}、T_{max}、u_2、n	0.900	0.902	0.928	0.956	0.505	0.458	0.418	0.312
GEP7	T_{min}、T_{max}、R_H、n	0.932	0.945	0.934	0.943	0.411	0.370	0.401	0.353
GEP8	T_{min}、T_{max}、R_a	0.835	0.840	0.890	0.885	0.633	0.616	0.512	0.500
GEP9	T_{min}、T_{max}、R_a、R_H	0.914	0.908	0.907	0.866	0.457	0.459	0.482	0.539
GEP10	T_{min}、T_{max}、R_a、u_2	0.865	0.855	0.927	0.899	0.579	0.556	0.419	0.472
GEP11	T_{min}、T_{max}、R_a、n	0.901	0.946	0.946	0.941	0.491	0.375	0.361	0.365
GEP12	T_{min}、T_{max}、R_a、R_H、u_2	0.859	0.914	0.916	0.934	0.611	0.461	0.468	0.379
GEP13	T_{min}、T_{max}、R_a、u_2、n	0.920	0.951	0.976	0.961	0.449	0.334	0.241	0.297
GEP14	T_{min}、T_{max}、R_a、R_H、n	0.919	0.974	0.970	0.974	0.442	0.235	0.273	0.240
GEP15	T_{min}、T_{max}、R_a、R_H、u_2、n	0.942	0.972	0.970	0.977	0.375	0.244	0.272	0.225
Har	T_{min}、T_{max}、R_a	0.814	0.868	0.876	0.875	0.693	0.687	0.754	0.697

因子能够提高计算精度。因此，RF2、RF3 和 RF4 分别在 RF1 的基础上引入 R_H、u_2和 n，与 RF1 相比，其精度均有所提高，其中引入日照时数的 RF4 提高最显著，R^2平均值由 0.842 增加到 0.921，R_{MSE}的平均值由 0.603 mm/d 降低到 0.437 mm/d。其次为引入相对湿度的 RF2 模型(平均 R^2 = 0.908，平均 R_{MSE} = 0.460 mm/d)，引入风速的 RF3 精度提升最小(平均 R^2 = 0.884，平均 R_{MSE} = 0.520 mm/d)。这说明不同气象因子对 ET_0的重要性不同，桂西北喀斯特地区 ET_0除了气温变量，受 n 的影响最大，其次为 R_H和 u_2。RF 模型可以给出输入变量的重要性估计，可帮助理解主要影响因变量的变量[12]。图 1 给出了 6 个变量重要性排序，4 个站点气象因子相对重要性依次为：T_{max}、n、T_{min}、R_a、R_H和 u_2，结果与模型 RF2、RF3 和 RF4 的平均精度表现一致。

模型 RF5、RF6 和 RF7 分别在模型 RF2、RF3 和 RF4 的基础上引入 u_2、n 和 R_H得到，精度均得到提高，平均 R^2分别增加了 1.65%、7.35%和 3.04%，平均 R_{MSE}分别减小了 6.09%、31.15% 和 20.14%。RF7 模型(平均 R^2 = 0.949，平均 R_{MSE} = 0.349 mm/d)的精度明显高于 RF6(平均 R^2 = 0.949，平均 R_{MSE} = 0.358 mm/d)和 RF5(平均 R^2 = 0.923，平均 R_{MSE} = 0.432 mm/d)，这也可以由气象因子的重要性来解

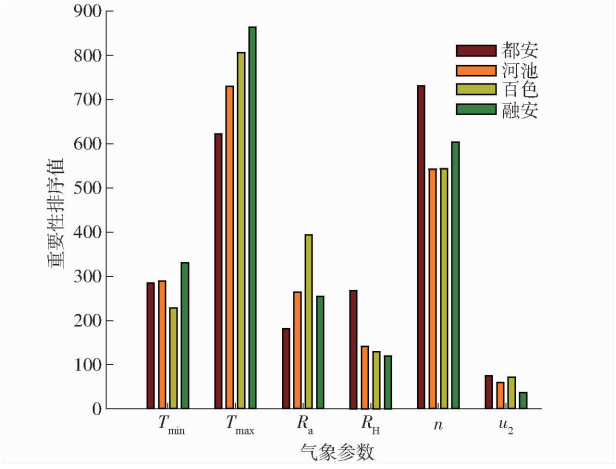

图 1 影响 ET_0 的气象因子重要性排序

Fig. 1 Rank of importance of meteorological variables influencing ET_0

释,重要性依次为 $R_H + n$、$n + u_2$ 和 $R_H + u_2$。

模型 RF8 ~ RF14 分别在模型 RF1 ~ RF7 的基础上引入 R_a 得到,由表 2 可见模型精度均显著提高,平均 R^2 分别增加了 3.92%、2.09%、1.13%、3.15%、1.19%、2.21% 和 2.63%,平均 R_{MSE} 分别减小了 9.45%、9.35%、2.31%、18.08%、6.02%、18.99% 和 24.93%。这是因为蒸散发的能量来源为太阳辐射,理论太阳辐射由日地相对距离和太阳高度角决定,日地相对距离是日序数的函数,太阳高度角是地理纬度的函数,而大气顶层太阳辐射 R_a 是地理纬度和日序数的函数,综合反映了这 2 个因素,因此引入 R_a 可提高 ET_0 模型的精度[3]。而且 R_a 不需要观测,由计算得到,避免了观测导致的误差。我国目前大多数气象站点未能观测太阳辐射数据,只能由 n 转换得到,因此在缺少太阳辐射值而用日照时数来计算太阳净辐射时,有必要引入 R_a。

模型 RF8 的输入因子(T_{max}、T_{min} 和 R_a)和传统 Hargreaves 模型相同,4 个站点 RF8 模型的精度均高于 Hargreaves 模型,平均 R^2 增加了 1.98%,平均 R_{MSE} 减小了 22.88%,因此在仅有气温数据时,可以使用 RF 模型代替传统 Hargreaves 模型,以提高 ET_0 计算精度。

模型 RF15 使用了与 FAO P-M 相同的所有气象因子,其模拟结果也是最准确的(平均 R^2 = 0.989,平均 R_{MSE} = 0.179 mm/d)。尽管 RF15 的输入因子与 FAO P-M 相同,但结果仍然有差异,所需气象因子完备情况下并不能替代 FAO P-M 模型,这是由于 FAO P-M 包含了辐射项和空气动力学项,每一项的中间参数都由气象因子、地理位置信息、仪器安装高度以及日序数等得到,计算过程会引入误差,而 RF 算法未能捕捉到这些误差,造成了 2 种模型的计算差异。

2.2 检验期不同气象因子输入组合下 GEP 模型计算结果对比

表 3 给出了不同气象因子组合下基于 GEP 算法的 ET_0 模型精度,可见 GEP 模型能够很好地模拟 ET_0 和气象因子之间的非线性关系,4 个站点平均 R^2 的变化范围为 0.857 ~ 0.965,平均 R_{MSE} 的变化范围为 0.279 ~ 0.574 mm/d。比较 GEP1、GEP2、GEP3 和 GEP4,可见与 RF 模型相同,引入 n 的 GEP4 模型精度高于引入 R_H 和 u_2 的 GEP2 和 GEP3 模型。同样,在模型 GEP2、GEP3 和 GEP4 的基础上引入 u_2、n 和 R_H 得到的 GEP5、GEP6 和 GEP7 模型的精度也明显提高,平均 R^2 分别增加了 0.11%、5.49% 和 2.74%,平均 R_{MSE} 分别减小了 0.86%、20.49% 和 15.60%。比较 GEP8 ~ GEP14 与 GEP1 ~ GEP7,可见与 RF 模型一样,引入 R_a 使得模型精度均显著提高。

GEP8 模型和传统 Hargreaves 模型使用的气象因子相同,在 4 个站点 GEP8 模型的精度均优于 Hargreaves 模型,平均 R^2 提高了 0.58%,平均 R_{MSE} 减小了 20.20%。相比于其他算法产生的是黑箱模型,GEP 算法的一个优点是能够生成明确的代数表达式,便于应用。因此在仅有气温,缺乏其他气象资料时,GEP8 模型可以代替 Hargreaves 模型。生成 4 个站点(都安、河池、百色和融安)3 个参数(T_{min}、T_{max}、R_a)GEP 模型表达式分别为

$$ET_0 = 1.001\,3^{\left(\frac{R_a}{T_{min}+T_{max}}\right)^2} + \ln(\ln(T_{min} + T_{min}R_a)) - \sqrt[3]{\frac{T_{min}}{T_{max}-T_{min}}} - e^{\sqrt{T_{min}}} \tag{6}$$

$$ET_0 = 0.857T_{max} + \frac{0.981}{T_{max}} - \frac{15.009}{R_a} - 0.557\sqrt{T_{max}} - 1.375 \tag{7}$$

$$ET_0 = 0.192\exp\left(\frac{T_{max}R_a}{T_{min}+10.476R_a}\right) - \frac{R_a + T_{max} + 2T_{min}^2 + 13.993T_{min}}{R_aT_{min}} + 2.209 \tag{8}$$

$$ET_0 = \left(\frac{T_{min}R_a}{T_{min}R_a - 71.185T_{max}}\right)^3 + \frac{\ln^8 T_{max}}{5\,876.43} + \frac{R_a - T_{min}}{18.72} \tag{9}$$

2.3 RF 和 GEP 模型对比

比较表 2、3 可见,除了输入组合 1(T_{max} 和 T_{min})和输入组合 10(T_{min}、T_{max}、R_a 和 u_2)产生的 RF 模型计算精度低于 GEP 模型外,其他输入组合下 RF 模型精度均高于相应 GEP 模型,平均 R_{MSE} 分别减小了 1.20%、2.26%、4.12%、6.84%、18.23%、10.04%、3.48%、16.20%、11.33% 和 56.30%。图 2 给出了

检验期(2011—2012 年)RF8、RF11 和 RF15 以及 GEP8、GEP11 和 GEP15 模型模拟结果与 FAO P－M 计算结果的散点图,根据线性回归 R^2 以及数据点分散程度可见,RF8、RF11 和 RF15 模型精度要高于相应 GEP8、GEP11 和 GEP15 模型,与表 2 和表 3 统计结果一致。尽管 GEP 算法产生的模型达到足够精度所需的运算时间远大于 RF 算法,但 GEP 模型能够产生自变量和因变量之间的算术表达式,这一方面便于挖掘自变量和因变量之间的理论关系,另一方面也便于应用。

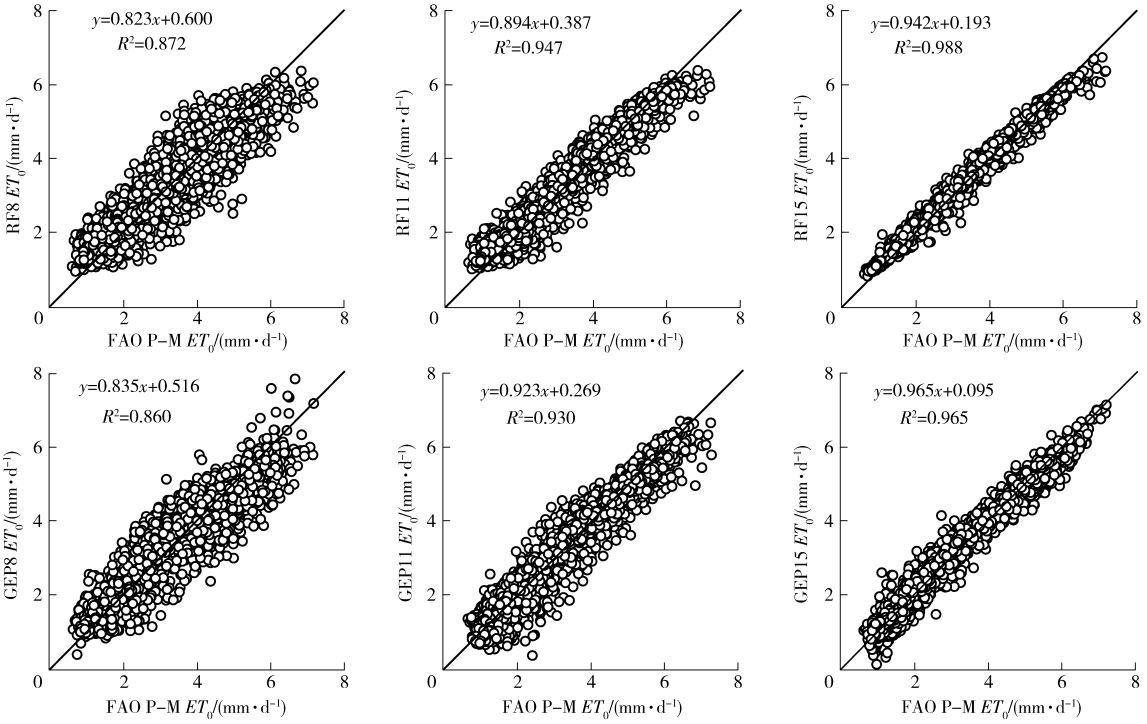

图 2　检验期 RF8、RF11、RF15 和 GEP8、GEP11、GEP15 模型结果与 FAO P－M 结果比较的散点图

Fig. 2　Comparative scatter plots between ET_0 of RF8, RF11, RF15, GEP8, GEP11, GEP15 and ET_0 from FAO P－M during test period

3　讨论

我国西南喀斯特生态脆弱区二元结构发育,雨水通过地表漏斗进入地下暗河,加之土层浅薄,水分涵养能力差,因此尽管降水较丰富,但主要埋藏于地下,地表耕地易于干涸而出现旱情[15,17]。随着社会经济的发展,工业用水比例增加,相应农业可用水量减小,因此精确计算 ET_0 有助于合理分配和管理利用有限的农业水资源。然而在大多数地区气象站所观测的气象数据并不满足 FAO P－M 模型的要求(关键气象因子缺失、数据质量不可靠等),如内蒙古自治区 118 个气象站中能观测太阳净辐射数据的仅有 7 个,贵州省 80 个气象站中仅有 1 个能观测该数据,且仅有 19 个站点能观测用于计算太阳净辐射的日照时数[26],因此有必要研究基于机器学习算法建立所需气象因子少、精度足够高的 ET_0 计算模型。由于随机森林算法在不同数据集上表现稳健、预测准确性高、所需用户指定参数少、不容易过拟合且能够计算预测因子的重要性,因此它在很多领域得到应用[6,11－13]。本研究首次探讨了随机森林回归算法在 ET_0 计算方面的能力,发现在相同输入下,其计算精度明显高于 GEP 模型和径向基函数神经网络模型[3],其中 RF8 仅利用最大和最小气温而模型精度高于相同输入因子的传统 Hargreaves 模型,证明了随机森林算法计算 ET_0 的可行性,为后续研究建立该区基于随机森林的通用 ET_0 模型奠定了基础。以峰丛洼(谷)地为基本景观单元的西南喀斯特地貌“十里不同天”,气象因子(降水量、温度、湿度和光照)时空变异大,然而不可能针对每个灌区建立气象站,这种情况下使用基于简单易测的气象因子的 ET_0 模型(如本文建立的基于气温的 RF8 模型),既能满足生产需求,又能达到节水的目的。

随机森林算法的另一个优良特征是能计算自变量对因变量的重要性,因此其结果可解释性强于其他机器学习算法。具体在本研究中表现为给出各个气象因子对计算 ET_0 的重要性程度,一方面有助于选取关键因子以提高 ET_0 计算精度,另一方面有助于评估未来气候变化条件下农业水资源需求量的变化。

本研究所建立的不同气象站点、不同输入因子

的 ET_0 模型可能仅适应于该站点,下一步需要根据研究区更多气象站点的数据,建立适用于整个研究区的 RF 泛化模型。

4 结论

(1)随机森林算法成功应用于亚热带季风气候区的西南喀斯特地区 ET_0 模拟计算,不同气象因子组合输入下建立的 RF 模型均能较好地反映气象因子与 ET_0 之间的非线性关系。随着气象因子的增加,RF 模型模拟的精度随之提高。在仅有气温数据时,RF 模型仍具有足够的精度(R^2 = 0.875,R_{MSE} = 0.546 mm/d),与传统 Hargreaves 模型相比平均 R^2 增加了 1.98%,平均 R_{MSE} 减小了 22.88%,因此在仅有气温数据时可用 RF 模型代替 Hargreaves 模型。

(2)不同气象因子的引入对模型精度提升不同,在气温基础上引入日照时数对精度提升最显著(平均 R^2 为 0.921,平均 R_{MSE} 为0.437 mm/d),其次为引入相对湿度(平均 R^2 为 0.908,平均 R_{MSE} 为 0.460 mm/d)和风速(平均 R^2 为 0.874,平均 R_{MSE} 为 0.532 mm/d)。RF 算法对气象因子的重要性评估表明,在该区域对 ET_0 重要的气象因子依次为 T_{max}、n、T_{min}、R_a、R_H 和 u_2。引入不需要观测的大气顶层太阳辐射 R_a 能够明显提高模型精度。

(3)总体来说,相同气象因子组合输入下,基于 RF 算法建立的模型精度要高于基于 GEP 算法建立的模型。然而 GEP 模型具有明确的算术表达式,因此对计算机技术并不精通的灌溉工作人员建议使用 GEP 模型,水资源管理及水量平衡研究中建议使用 RF 模型。

参 考 文 献

1 ALLEN R G, PEREIRA L S, RAES D, et al. Crop evapotranspiration—guidelines for computing crop water requirements. FAO irrigation and drainage paper 56[R]. Rome: FAO, 1998, 300(9): D5109.

2 SHIRI J, KISI Ö, LANDERAS G, et al. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain)[J]. Journal of Hydrology, 2012, 414: 302 - 316.

3 王升,陈洪松,聂云鹏,等. 基于基因表达式编程算法的参考作物腾发量模拟计算[J/OL]. 农业机械学报, 2015, 46(4): 106 - 112. http://www.j-csam.org/jcsam/ch/reader/create_pdf.aspx? file_no=20150416&year_id=2015&quarter_id=4&falg=1. DOI:10.6041/j.issn.1000-1298.2015.04.016.
WANG Sheng, CHEN Hongsong, NIE Yunpeng, et al. Simulation of evapotranspiration based on gene-expression programming method[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(4): 106 - 112. (in Chinese)

4 DJAMAN K, BALDE A B, SOW A, et al. Evaluation of sixteen reference evapotranspiration methods under sahelian conditions in the Senegal River Valley[J]. Journal of Hydrology: Regional Studies, 2015, 3: 139 - 159.

5 胡庆芳,杨大文,王银堂,等. Hargreaves 公式的全局校正及适用性评价[J]. 水科学进展, 2011,22(2): 160 - 167.
HU Qingfang, YANG Dawen, WANG Yintang, et al. Global calibration of Hargreaves equation and its applicability in China[J]. Advances in Water Science, 2011, 22(2): 160 - 167. (in Chinese)

6 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

7 LANDERAS G, ORTIZ-BARREDO A, LÓPEZ J J. Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain)[J]. Agricultural Water Management, 2008, 95(5): 553 - 565.

8 WEN X, SI J, HE Z, et al. Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions [J]. Water Resources Management, 2015, 29(9): 3195 - 3209.

9 侯志强,杨培岭,苏艳平,等. 基于最小二乘支持向量机的 ET_0 模拟计算[J]. 水利学报, 2011, 42(6): 743 - 749.
HOU Zhiqiang, YANG Peiling, SU Yanping, et al. Simulation of ET_0 based on LS - SVM Method[J]. Journal of Hydraulic Engineering, 2011, 42(6): 743 - 749. (in Chinese)

10 冯禹,崔宁博,龚道枝,等. 基于极限学习机的参考作物蒸散量预测模型[J]. 农业工程学报,2015, 31(增刊 1): 153 - 160.
FENG Yu, CUI Ningbo, GONG Daozhi, et al. Prediction model of reference crop evapotranspiration based on extreme learning machine[J]. Transactions of the CSAE, 2015, 31(Supp. 1): 153 - 160. (in Chinese)

11 FERNANDEZ-DELGADO M, CERNADAS E, BARRO S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. The Journal of Machine Learning Research, 2014, 15(1): 3133 - 3181.

12 张雷,王琳琳,张旭东,等. 随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例[J]. 生态学报, 2014,34(3): 650 - 659.
ZHANG Lei, WANG Linlin, ZHANG Xudong, et al. The basic principle of random forest and its applications in ecology: a case study of *Pinus yunnanensis*[J]. Acta Ecologica Sinica, 2014, 34(3): 650 - 659. (in Chinese)

13 王丽爱,马昌,周旭东,等. 基于随机森林回归算法的小麦叶片 SPAD 值遥感估算[J/OL]. 农业机械学报, 2015, 46(1): 259 - 265. http://www.j-csam.org/jcsam/ch/reader/create_pdf.aspx? file_no=20150136&flag=1&journal_id=jcsam. DOI: 10.6041/j.issn.1000-1298.2015.01.036.

WANG Li'ai, MA Chang, ZHOU Xudong, et al. Estimation of wheat leaf SPAD value using RF algorithmic model and remote sensing data[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1): 259 – 265. (in Chinese)

14 韩兆迎,朱西存,房贤一,等. 基于 SVM 与 RF 的苹果树冠 LAI 高光谱估测[J]. 光谱学与光谱分析, 2016, 36(3): 800 – 805.
HAN Zhaoying, ZHU Xicun, FANG Xianyi, et al. Hyperspectral estimation of appple tree canopy LAI based on SVM and RF regression[J]. Spectroscopy and Spectral Analysis, 2016, 36(3): 800 – 805. (in Chinese)

15 陈洪松,杨静,傅伟,等. 桂西北喀斯特峰丛不同土地利用方式坡面产流产沙特征[J]. 农业工程学报, 2012, 28(16): 121 – 126.
CHEN Hongsong, YANG Jing, FU Wei, et al. Characteristics of slope runoff and sediment yield on karst hill-slope with different land-use types in northwest Guangxi[J]. Transactions of the CSAE, 2012, 28(16): 121 – 126. (in Chinese)

16 JIANG Z, LIAN Y, QIN X. Rocky desertification in Southwest China: impacts, causes, and restoration[J]. Earth-Science Reviews, 2014, 132: 1 – 12.

17 付同刚,陈洪松,张伟,等. 喀斯特小流域土壤含水率空间异质性及其影响因素[J]. 农业工程学报, 2014, 30(14): 124 – 131.
FU Tonggang, CHEN Hongsong, ZHANG Wei, et al. Spatial variability of soil moisture content and its influencing factors in small karst catchment during dry period[J]. Transactions of the CSAE, 2014, 30(14): 124 – 131. (in Chinese)

18 CHEN H, ZHANG W, WANG K, et al. Soil moisture dynamics under different land uses on karst hillslope in northwest Guangxi, China[J]. Environmental Earth Sciences, 2010, 61(6): 1105 – 1111.

19 LIU M, XU X, SUN A Y, et al. Is southwestern China experiencing more frequent precipitation extremes? [J]. Environmental Research Letters, 2014, 9(6): 64002.

20 刘钰,蔡林根. 参照腾发量的新定义及计算方法对比[J]. 水利学报, 1997,24(6): 27 – 33.
LIU Yu, CAI Lin'gen. Update definition and computation of reference evapotranspiration comparison with former method[J]. Journal of Hydraulic Engineering, 1997,24(6):27 – 33. (in Chinese)

21 BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5 – 32.

22 YANG R, ZHANG G, LIU F, et al. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem[J]. Ecological Indicators, 2016, 60: 870 – 878.

23 DROOGERS P, ALLEN R G. Estimating reference evapotranspiration under inaccurate data conditions[J]. Irrigation and Drainage Systems, 2002, 16(1): 33 – 45.

24 TRAORE S, WANG Y, KERH T. Artificial neural network for modeling reference evapotranspiration complex process in Sudano – Sahelian zone[J]. Agricultural Water Management, 2010, 97(5): 707 – 714.

25 王茵茵,齐雁冰,陈洋,等. 基于多分辨率遥感数据与随机森林算法的土壤有机质预测研究[J]. 土壤学报, 2016, 53(2): 342 – 354.
WANG Yinyin, QI Yanbing, CHEN Yang, et al. Prediction of soil organic matter based on multi-resolution remote sensing data and random forest algorithm[J]. Acta Pedologica Sinica, 2016, 53(2): 342 – 354. (in Chinese)

26 XU J. Proper methods and its calibration for estimating reference evapotranspiration using limited climatic data in Southwestern China[J]. Archives of Agronomy & Soil Science, 2014, 61(3):415 – 426.