李梦杰, 刘琨, 牟海磊, 等. 结合多重假设检验的随机森林长期降水预测方法及应用[J]. 南水北调与水利科技(中英文), 2024, 22(5): 920-926. LI M J, LIU K, MOU H L, et al. A random forest long-term precipitation prediction method combined with multiple hypothesis testing and its application[J]. South-to-North Water Transfers and Water Science & Technology, 2024, 22(5): 920-926. (in Chinese)

结合多重假设检验的随机森林长期降水预测方法及应用

李梦杰[1], 刘琨[1], 牟海磊[2], 殷兆凯[1], 刘志武[1], 吴迪[2], 梁犁丽[1]

(1. 中国长江三峡集团有限公司科学技术研究院, 北京 101199; 2. 中国三峡国际股份有限公司, 北京 101199)

摘要: 为解决随机森林方法经验性选取预测因子时存在的错误发现率问题, 引入多重假设检验领域控制错误发现率的方法对预测因子的筛选进行质量控制, 将因子筛选由经验依赖转化为数据依赖, 从而提出一种基于多重假设检验的随机森林方法长期降水预测方法。以巴西巴拉那河上游流域为研究区, 利用逐月气候系统指数, 应用提出的方法对研究区 2018—2020 年 54 个雨量站点的逐月降水量进行模拟预测、检验和交叉验证。结果表明: 与传统的随机森林方法相比, 该方法预报精度更高, 对不同站点 1—12 月的预测平均合格率达到 64%, 其中 6 月预测合格率达到 84%, 表明该方法可以作为流域长期降水预测的有效工具之一。

关键词: 随机森林方法; 长期降水预测; 预测因子筛选; 质量控制; 多重假设检验

中图分类号: TV11 **文献标志码:** A **DOI:** 10.13476/j.cnki.nsbdqk.2024.0091

长期降水预测是指预见期 1 个月以上的降水预测, 它是水资源综合管理的重要依据。由于多种不确定性因素影响, 长期降水预测准确度较低。传统的长期降水预测方法主要分为动力数值方法和数理统计方法[1]。动力数值方法借助于海陆热动力模型模拟未来天气状况来进行降水预测, 其物理机制明确, 但模型计算复杂[2-7]。数理统计方法则从统计学角度出发, 模拟降水与预测因子之间的相关关系, 建立长期预测模型[8-12]。然而, 基于数理统计方法的降水预测研究多集中在模型的改进, 对于如何选取预测因子的研究相对较少。实际上, 预测因子的合适与否影响着模型预测的准确率[13-14]。因此, 如何进行预测因子的筛选是降水预测的重点和难点。随机森林方法进行因子筛选运行灵活高效, 因此应用较为广泛[15-18]。然而, 随机森林方法只能依据因子重要性排序经验性的筛选部分重要因子, 导致筛选的因子存在一定的错误率[19-20], 而多重假设检验作为分析高维数据的重要方法, 可以使用错误发现率(false discovery rate, FDR)来表征这一错误率[21-23]。

本研究采用多重假设检验领域控 FDR 的“Model-X Knockoff”方法[24] 来控制随机森林变量筛选的错误率, 实现对巴西巴拉那河流域降水预测因子筛选的有效质控, 最终利用筛选的降水预测因子结合随机森林方法建立逐月的长期降水预测模型, 以期为长期降水预测提供技术支撑。

1 研究区概况

巴拉那河发源于巴西高原东南缘的曼蒂凯拉山脉北坡, 主源为格兰德河, 汇合巴拉那伊巴河后, 始称巴拉那河。本研究区域属于巴拉那河上游, 位于巴西中南部地区, 面积 78.3 万 km^2, 位于 15°27′S~25°39′S 和 43°35′W~55°56′W。该区域具有多种地貌地形, 其中, 大西洋高原海拔高于 2 000 m, 巴拉那河谷海拔 100 ~ 350 m, 年平均降雨量约为 1 543 mm[25]。区域内降水具有很大的空间变异性, 北部受南美季风的影响, 夏季的降雨量会超过 800 mm, 而在冬季

收稿日期: 2023-12-07 **修回日期:** 2024-06-06 **网络出版时间:** 2024-07-18
网络出版地址: https://link.cnki.net/urlid/13.1430.TV.20240716.1657.004
基金项目: 中国长江三峡集团有限公司自主科研项目(NBZZ20210055); 国家科技基础资源调查专项项目(2021xjkk0405)
作者简介: 李梦杰(1991—), 女, 山东济宁人, 工程师, 博士, 主要从事数理统计方法与应用研究。E-mail: limengjie_sky@163.com
通信作者: 梁犁丽(1982—), 女, 河南许昌人, 正高级工程师, 博士, 主要从事水文预报、水库调度及水资源配置研究。E-mail: liangli0921@163.com

降雨量低至 30 mm，南部受热带辐合带、冷锋和中尺度对流复合体影响，降雨主要在春季和夏季。根据研究区域内 54 个雨量站 1980—2020 年的观测降水量估算，多年平均月降水量约 123 mm，其中：汛期 10 月—次年 3 月约为 191 mm；非汛期 4—9 月约为 56 mm。

2　数据与方法

2.1　数据来源

以巴西巴拉那河上游流域 54 个雨量站点 1980—2020 年的实测月降水序列（资料来自巴西国家水务局 http://www3.ana.gov.br/）作为研究资料。气候因子使用中国气象局国家气候中心 1980—2020 年的 130 项逐月气候系统指数，包含北半球副高面积指数、北非副高面积指数等 88 项大气环流指数，NINO 1+2 区海表温度距平指数、NINO 3 区海表温度距平指数等 26 项海温指数，以及太阳黑子指数、南方涛动指数等 16 项其他指数。

实测日降水量数据来源于巴西国家水务局（http://www3.ana.gov.br/），时间跨度为 1980—2020 年。首先对站点实测降水数据进行质量控制，如删除重复记录、修正数值错误以及站点位置校核；随后将日降水合并为月降水数据；考虑建模对数据时间序列长度的要求，筛选出 1980—2020 年数据缺失小于 100 条的站点，并利用线性插补的方法对缺失数据进行插补，最终得到 54 个雨量站点的逐月降水数据用于模型构建。

2.2　研究方法

2.2.1　灰色关联分析

灰色关联分析作为多因素统计分析方法，可以逐月量化不同气候因子与该月降水之间的关联程度。对于数列X_0有若干个比较数列：$X_1,X_2,\cdots,X_n$，每个数列X_i长度为 N，i=1,$\cdots$,n。各比较数列与参考数列在各个时刻的关联系数$\xi_i(k)$的计算公式为

$$\xi_i(k)=\frac{\min\limits_i\min\limits_k|x_0(k)-x_i(k)|+\rho[\max\limits_i\max\limits_k|x_0(k)-x_i(k)|]}{|x_0(k)-x_i(k)|+\rho[\max\limits_i\max\limits_k|x_0(k)-x_i(k)|]} \tag{1}$$

式中：k=1,$\cdots$,N，N 为时刻个数；ρ为分辨系数，根据经验和测试验证取$\rho=0.01$。将相应月份的降水量设置为参考数列，各种气候因子设置为比较数列。通过计算气候因子与降水的关联度$r_i=\frac{1}{N}\sum_{k=1}^{N}\xi_i(k)$，逐月选取灰色关联度最高的前 20 个气候因子构建该月份建模所需的样本集，样本集分为解释变量与因变量，其中，解释变量包括当年每月降水量以及影响该月降水量的前 20 个关键气候因子在当年每月的观测值（共$12+20\times12=252$个解释变量），因变量则为次年该月降水量。

2.2.2　基于 Model-X Knockoff 的因子筛选

对于灰色关联分析得到的第 i(i=1,2,$\cdots$,12) 个月的解释变量和因变量分别记为X_i和Y_i，记X_i每个分量为$X_{ij}(j=1,2,\cdots,252)$，记$X_{i,-j}=\{X_{i1},\cdots,X_{ip}\}\backslash\{X_{ij}\}$，则变量$X_{ij}$对应的原假设是$X_{ij}|X_{i,-j}$与$Y_i$独立（null），备择假设是：$X_{ij}|X_{i,-j}$与$Y_i$非独立。对于每个解释变量$X_{ij}(j=1,2,\cdots,252)$，利用近似半正定规划（approximate semi-definite program, ASDP）方法构造其相应的二阶“仿冒”（Knockoff）变量$\tilde{X}_{ij}$，将所有的X_{ij}以及$\tilde{X}_{ij}$作为自变量，次年该月份的降水作为因变量，利用随机森林进行建模，通过残差平方和来度量节点杂质，获取每个变量X_{ij}以及$\tilde{X}_{ij}$的重要性打分Z_{ij}、$\tilde{Z}_{ij}$，随后构造竞争性统计量$W_{ij}=|Z_{ij}|-|\tilde{Z}_{ij}|$，使得对于任意的打分$t>0$，满足：

$$\#\{j:W_{ij}\leqslant-t\}\geqslant\#\{\text{null } j:W_{ij}\leqslant-t\}\overset{d}{=}\#\{\text{null } j:W_{ij}\geqslant t\} \tag{2}$$

则所有$W_t\geqslant t$的假设检验对应的 FDR 为

$$\mathrm{FDR}(t)=E\left(\frac{\#\{\text{null } j:W_{ij}\geqslant t\}}{\#\{j:W_{ij}\geqslant t\}}\right) \tag{3}$$

对于给定的错误率水平α，计算竞争性打分的打分阈值 T 为

$$T=\min\left\{t\in\omega:\frac{1+\#\{j:W_j\leqslant-t\}}{\#\{j:W_j\geqslant t\}}\leqslant\alpha\right\} \tag{4}$$

式中：$\omega=\{|W_j|:j=1,2,\cdots,252\}\backslash\{0\}$。计算$S_i=\{j:W_{ij}\geqslant T\}$，则将所有的$X_{ik},k\in S_i$作为第 i 个月份降水模型最终的预测因子，则上述过程可以严格控制所筛选因子的FDR$\leqslant\alpha$。

2.2.3　降水预测模型搭建

基于上述筛选得到的影响次年 i 月份降水量的预测因子$X_{ik},k\in S_i$和次年相应 i 月份降水量进行随机森林建模，训练得到次年相应月份降水量的长期降水预测模型φ_i，i=1,$\cdots$,12。最终获得次年 1—12 月共 12 个降水预测模型为

$$\hat{Y}_i=\varphi_i(X_{ik}),k\in S_i \tag{5}$$

式中：$\hat{Y}_i$为次年i月份的降水预测值，mm；φ_i为次年 i 月份的降水预测模型，$X_{ik},k\in S_i$代表影响次年 i 月份降水的预测因子。

考虑到逐月样本数据量较少，因此仅将原始数

据划分为训练期以及测试期两类。设置模型训练期为 1980—2017 年,训练期共有 38 年的月尺度数据;模型测试期为 2018—2020 年,测试期共有 3 年的月尺度资料。利用训练期样本进行建模,测试期样本进行模型测试。建模中,利用提出的基于 Model-X Knockoff 的随机森林方法进行变量筛选时,均设置错误率水平阈值$\alpha = 0.2$。对于原始的随机森林方法,对解释变量进行变量重要性打分,筛选出重要性打分前 5 的预测因子重新进行随机森林建模。

为进一步验证结合多重假设检验的长期降水方法相较于传统的随机森林方法的改进,此处分别利用 10 折交叉验证,以及 2018—2020 年降水模拟预测检验结果来验证本文所提出方法的应用效果。其中 10 折交叉验证的方法,即将每个站点逐月的样本随机平均分成 10 份,依次选取其中 9 份样本(36 个)作为训练集,剩余 1 份样本(4 个)作为测试集,对于每个训练集以及测试集样本,建模对测试集样本的逐月降水进行预测,重复 10 次,遍历每个站点逐月的所有样本,对 10 次交叉验证结果取平均作为该站点逐月的结果,最终对所有站点的逐月结果求平均。采用 10 折交叉检验的方法一方面可以弥补数据量的不足,另一方面可以利用模型均值的形式验证模型的泛化能力。

2.2.4　方法性能评估

考虑到现行《水文情报预报规范(SL 250—2000)》规定中长期降水的定量预报以多年变幅的 20% 作为许可误差,若预报误差小于许可误差则为合格,否则为不合格。对于预测结果,分别采用合格率(P)、均方根误差(E_{RMS})和平均绝对误差(E_{MA})来评估定量降水的预报精度,具体计算公式为

$$P = \frac{m_1}{m} \times 100\% \tag{6}$$

$$E_{\mathrm{EMS}}(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[h(x_i) - y_i]^2} \tag{7}$$

$$E_{\mathrm{MA}}(X,h) = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i| \tag{8}$$

式(6)~(8)中:m_1 为合格的样本个数;m 为样本总数;y_i 为样本 i 的实际降水量,mm;$h(x_i)$ 为样本 i 的预测降水量,mm。

3　结果与分析

10 折交叉验证的结果见表 1,其中的许可误差一列为 54 个站点逐月的许可误差均值。结果显示,基于 Model-X Knockoff 的随机森林方法(RF+Knockoff)在 1—12 月份的模型预测合格率均高于传统的随机森林方法(RF)。其中:6 月份的合格率最高,达到了 77%,RF 方法则为 74%;对于 1—12 月的平均合格率,RF+Knockoff 方法的随机森林模型为 63.9%,比 RF 方法提升了 2.1%。E_{RMS} 以及 E_{MA} 的结果显示,RF+Knockoff 方法的预测模型在 1—12 月的拟合误差均低于 RF 方法。

表 1　10 折交叉验证的结果比较

Tab. 1　Comparison of results of 10 fold cross-validation

月份	许可误差/mm	RF+ Knockoff			RF		
		P/%	E_{RMS}/mm	E_{MA}/mm	P/%	E_{RMS}/mm	E_{MA}/mm
1	109	57	131	103	57	137	109
2	84	59	99	80	58	102	81
3	85	63	97	76	60	102	80
4	49	55	62	50	52	67	54
5	43	67	49	37	67	52	38
6	35	77	40	27	74	42	27
7	28	76	31	21	74	32	21
8	31	72	35	24	69	37	26
9	46	65	52	40	63	54	42
10	60	56	74	59	55	79	62
11	71	57	85	69	54	91	74
12	85	63	98	77	59	106	83
平均	60.5	63.9	71.1	55.3	61.8	75.1	58.1

利用传统随机森林方法和本研究提出的基于 Model-X Knockoff 的随机森林方法,对 2018—2020 年的月降水进行预测,结果见表 2。2 种方法对 5—9 月的降水预测效果优于汛期,预测合格率均大于 60%。RF+Knockoff 方法的预测模型除在 3 月、5 月、7 月的预测合格率低于 RF 方法之外,其余月份降水预测合格率均高于 RF 方法。其中,6 月的合格率最高,达到了 84%。对于 1—12 月的平均合格率,RF+Knockoff 方法的预测结果为 65.7%,RF 方法的预测结果为 63.9%。E_{RMS} 以及 E_{MA} 的结果显示,RF+Knockoff 方法 1—12 月的拟合效果多数情况下显著优于 RF 方法。

为进一步展示 RF+Knockoff 方法与 RF 方法所筛选的预测因子的差异,从 54 个雨量站中选取 3 个代表站(分别位于流域的北部,西北部和东北部)展示次年 1 月份以及次年 7 月份建模时 2 种方法所筛

选的预测因子。如表 3 和表 4 所示，RF+Knockoff 方法与 RF 方法所筛选的预测因子存在部分交集，但差异较大。对于次年 1 月份的降水预测模型，A 站点 RF+Knockoff 方法过滤出 7 个预测因子，只有当年 1 月份的西太平洋暖池面积指数以及当年 6 月份的西太平洋副高脊线位置指数与 RF 方法建模筛选的预测因子一致，其余预测因子均不一致。RF+Knockoff 方法筛选的预测因子除各类气候因子之外，还包含当年 12 月份的降水。B 站点 RF+Knockoff 方法过滤出的 4 个预测因子与 RF 方法筛选的预测因子均不一致，其中 RF 方法筛选的预测因子包含当年 12 月份的降水。C 站点 RF+Knockoff 方法过滤出 5 个预测因子，其中除当年 7 月份的东亚槽位置指数与 RF 方法一致，其余预测因子均不重叠，其中 RF+Knockoff 方法筛选的预测因子中包含当年 3 月份的降水。

表 2　2 种方法对 2018—2020 年月降水预测结果的比较

Tab. 2　Comparison of precipitation prediction results from 2018 to 2020 by two methods

月份	RF + Knockoff			RF		
	P/%	E_{RMS}/mm	E_{MA}/mm	P/%	E_{RMS}/mm	E_{MA}/mm
1	57	118	103	51	126	110
2	64	103	84	60	108	87
3	59	84	74	62	90	80
4	54	56	50	49	59	53
5	66	44	38	70	46	39
6	84	26	24	81	24	21
7	71	22	20	75	21	18
8	73	33	27	70	34	27
9	79	37	31	75	40	33
10	57	71	62	57	76	65
11	57	80	70	55	85	73
12	67	79	69	62	85	74
平均	65.7	62.8	54.3	63.9	66.2	56.7

表 3　3 个站点 2 种方法次年 1 月份建模筛选的预测因子

Tab. 3　Predictors of modelling selection in January of the following year at three sites by two methods

站点	RF+Knockoff筛选的预测因子	RF筛选的预测因子
A站	当年1月份的西太平洋暖池面积指数； 当年2月份的欧亚经向环流指数； 当年3月份的大西洋欧洲区极涡面积指数； 当年6月份的西太平洋副高脊线位置指数； 当年9月份的太平洋区极涡面积指数； 当年11月份的西藏高原-1指数； 当年12月份的降水、太平洋区极涡强度指数；	当年1月份的西太平洋暖池面积指数； 当年5月份的北太平洋副高脊线位置指数； 当年6月份的西太平洋副高脊线位置指数、北大西洋.欧洲区极涡强度指数； 当年10月份的太平洋区极涡强度指数；
B站	当年8月份的西藏高原-2指数、北非副高脊线位置指数、北大西洋.欧洲区极涡强度指数；	当年6月份的北美区极涡面积指数； 当年8月份的北非副高脊线位置指数、北大西洋.欧洲区极涡强度指数、印度副高脊线位置指数； 当年12月份的降水；
C站	当年3月份的降水； 当年7月份的东亚槽位置指数； 当年10月份的北半球极涡中心强度指数、北半球极涡面积指数、东亚槽位置指数；	当年2月份的西藏高原-2指数； 当年4月份的北太平洋副高脊线位置指数； 当年7月份的东亚槽位置指数； 当年10月份的大西洋欧洲区极涡面积指数； 当年12月份的北半球极涡强度指数；

对于表 4 所示的次年 7 月份的降水预测模型，A 站点 RF+Knockoff 方法过滤的 6 个预测因子与 RF 方法筛选的预测因子均不一致。RF+Knockoff 方法的结果显示除各类气候因子之外，当年 8 月、9 月的降水对次年 1 月的降水也影响较大。B 站点 RF+Knockoff 方法过滤出的 7 个预测因子中包含当年 12 月的降水，此外只有当年 4 月的欧亚经向环流指数与 RF 方法一致。C 站点 RF+Knockoff 方法过滤出的 3 个预测因子中当年 5 月的南海副高脊线位置指数和当年 10 月的亚洲区极涡强度指数与 RF 方法过滤的预测因子结果一致，其余预测因子 2 种方法均不一致，此外，RF 方法筛选的预测因子包含当年 11 月的降水。

由于 2 种方法因子筛选的机理不同，RF+Knockoff 方法借助统计多重假设检验方法筛选的预测因子与 RF 方法依赖经验筛选的预测因子存在较大的差异，进而导致 2 种方法进行降水预测的效果存在一定差异。总体来看，通过 10 折交叉验证以

及对测试期进行降水预测的对比发现，相较于传统的随机森林方法，结合多重假设检验的 RF+Knockoff 方法预测的合格率更高，且模型的稳定性更好。

表 4　3 个站点 2 种方法次年 7 月份建模筛选的预测因子

Tab. 4　Predictors of modelling selection in July of the following year at three sites by two methods

站点	RF+Knockoff筛选的预测因子	RF筛选的预测因子
A站	当年4月份的850 hPa东太平洋信风指数； 当年6月份的太阳黑子指数、亚洲经向环流指数； 当年8月份、9月份的降水； 当年11月份的太阳黑子指数；	当年8月份的850 hPa东太平洋信风指数、北大西洋.欧洲区极涡强度指数； 当年10月份的亚洲区极涡强度指数； 当年12月份的西太平洋副高西伸脊点指数、北半球极涡中心经向位置指数；
B站	当年3月份的亚洲经向环流指数、亚洲纬向环流指数； 当年4月份的欧亚经向环流指数； 当年6月份的东太平洋副高面积指数； 当年11月份的欧亚经向环流指数、印度洋暖池强度指数； 当年12月份的降水；	当年4月份的欧亚经向环流指数； 当年6月份的欧亚纬向环流指数； 当年8月份的亚洲经向环流指数； 当年11月份的北半球副高面积指数、西太平洋副高强度指数；
C站	当年3月份的北半球极涡面积指数； 当年5月份的南海副高脊线位置指数； 当年10月份的亚洲区极涡强度指数；	当年1月份的亚洲经向环流指数； 当年2月份的北半球极涡中心纬向位置指数； 当年5月份的南海副高脊线位置指数； 当年10月份的亚洲区极涡强度指数； 当年11月份的降水；

4　结 论

针对随机森林方法经验性选取降水预测因子时存在的错误率问题，本文提出利用多重假设检验领域控制错误发现率的方法对预测因子的筛选进行质控，将因子筛选由经验依赖转化为数据依赖，最终利用随机森林方法结合筛选的降水预测因子进行长期降水预测。以巴西巴拉那河上游流域为研究区，对实测数据的 54 个雨量站点的降水量以及 130 项气候系统指数进行分析，利用影响次年相应月份降水量的预测因子与次年相应月份降水之间建立遥相关关系，随后利用 10 折交叉验证以及对 2018—2020 年月降水预测结果检验的方法验证所提出方法的有效性。结果表明，相较于传统的随机森林方法，结合多重假设检验进行预测因子筛选质控的方法可以有效地提升降水预测的准确性和可靠性。

鉴于长期降水预测的复杂性和不确定性，如何从物理机制解释所筛选的预测因子，从而更适应生产服务的需求，是下一步需要深入研究的内容；另外，还需要充分利用和挖掘其余实测站点的雨量信息，进一步提高长期降水预测精度。

参考文献：

[1] GHIMIRE S, YASEEN Z M, FAROOQUE A A, et al. Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks[J]. Scientific Reports, 2021, 11: 17497. DOI: 10.1038/s41598-021-96751-4.

[2] 巴欢欢, 郭生练, 钟逸轩, 等. 考虑降水预报的三峡入库洪水集合概率预报方法比较 [J]. 水科学进展, 2019, 30(2): 186-197. DOI: 10.14042/j.cnki.32.1309.2019.02.004.

[3] STRAZZO S, COLLINS D C, SCHEPEN A, et al. Application of a hybrid statistical-dynamical system to seasonal prediction of North American temperature and precipitation[J]. Monthly Weather Review, 2019, 147: 607-625. DOI: 10.1175/MWR-D-18-0156.1.

[4] 黄泽青, 刘洋, 杨振华, 等. 北美多模型集合预报实验(NMME)全球降水预报对长江上游流域夏季降水适用性研究 [J]. 水文, 2020, 40(6): 16-23. DOI: 10.19797/j.cnki.1000-0852.20190246.

[5] 夏达忠, 刘艳, 罗锡斌, 等. 基于 CFS 的中长期降水预报适用性研究 [J]. 水力发电, 2021, 47(12): 19-22. DOI: 10.3969/j.issn.0559-9342.2021.12.005.

[6] 黄赛男, 李文韬, 段青云. GEFSv12 降水再预报数据在淮河流域的适用性评估 [J]. 南水北调与水利科技(中英文), 2022, 20(5): 925-934. DOI: 10.13476/j.cnki.nsbdqk.2022.0092.

[7] 赵胤懋, 廖卫红, 田雨, 等. 西江流域 CMORPH 降水产品精度评估及水文效应研究 [J]. 水利水电技术, 2019, 50(2): 88-94. DOI: 10.13928/j.cnki.wrahe.2019.02.012.

[8] ZHANG X Q, WU X L, HE S Y, et al. Precipitation forecast based on CEEMD-LSTM coupled model[J]. Water Supply, 2021, 21(8): 4641-4657. DOI: 10.2166/ws.2021.237.

[9] REDDY B S N, PRAMADA S K, ROSHNI T. Monthly surface runoff prediction using artificial intelligence: A study from a tropical climate river basin[J]. Journal of Earth System Science, 2021, 130: 35. DOI: 10.1007/s12040-020-01508-8.

[10] 王永灿, 魏加华, 李琼, 等. 基于雷达回波的临近降水预报卷积循环神经网络模型研究 [J]. 水利水电技术 (中英文), 2023, 54(1): 24-41. DOI: 10. 13928 / j. cnki. wrahe. 2023. 01. 003.

[11] 徐冬梅, 张一多, 王文川. 基于小波包分解的 LS-SVM-ARIMA 组合降水预测 [J]. 南水北调与水利科技 (中英文), 2020, 18(6): 71-77. DOI: 10.13476/j.cnki.nsbdqk.2020.0116.

[12] 杨琼波, 崔东文. WPD-COA-ELM 模型在汛期月降水量时间序列预测中的应用研究 [J]. 水文, 2023, 43(1): 17-23. DOI: 10.19797/j.cnki.1000-0852.20210314.

[13] 李诒路. 基于 ECMWF System 4 的中长期径流集合预报研究 [D]. 北京: 清华大学, 2021. DOI: 10.27266/d.cnki.gqhau.2018.000892.

[14] 温馨, 孙艳, 李昱, 等. 流域年径流预报方法及因素影响分析 [J]. 水利水电技术 (中英文), 2023, 54(11): 113-123. DOI: 10.13928/j.cnki.wrahe.2023.11.010.

[15] 黄朝君, 贾建伟, 秦赫, 等. 基于 Copula 熵-随机森林的中长期径流预报研究 [J]. 人民长江, 2021, 52(11): 81-85. DOI: 10.16232/j.cnki.1001-4179.2021.11.013.

[16] 刁艳芳, 王蒙, 王昊, 等. 基于随机森林的水库防洪调度研究 [J]. 中国农村水利水电, 2022(3): 8-12. DOI: 10.3969/j.issn.1007-2284.2022.03.002.

[17] 陈雪怡, 陈元芳, 王文鹏, 等. 月径流预报建模方法对比分析：以嘉陵江北碚站为例 [J]. 人民长江, 2022, 53(9): 80-86. DOI: 10.16232/j.cnki.1001-4179.2022.09.013.

[18] 万育生, 王栋, 黄朝君. 丹江口水库来水情势分析与径流预测 [J]. 南水北调与水利科技 (中英文), 2021, 19(3): 417-426. DOI: 10.13476/j.cnki.nsbdqk.2021.0045.

[19] 牛勇, 李华鹏, 刘阳惠, 等. 超高维数据特征筛选方法综述 [J]. 应用概率统计, 2021, 37(1): 69-110. DOI: 10.3969/j.issn.1001-4268.2021.01.007.

[20] BARBER R F, CANDES E J. A knockoff filter for high-dimensional selective inference[J]. The Annals of Statistics, 2019, 47(5): 2504-2537. DOI: 10.1214/18-AOS1755.

[21] WANG R, RAMDAS A. False discovery rate control with *E*-values[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2022, 84(3): 822-852. DOI: 10.1111/rssb.12489.

[22] DAI C, LIN B. XING X, et al. A scale-free approach for false discovery rate control in generalized linear models[J]. Journal of the American Statistical Association, 2023, 118(543): 1551-1565. DOI: 10.1080/01621459.2023.2165930.

[23] HE K, LI M J, FU Y, et al. Null-free false discovery rate control using decoy permutations[J]. Acta Mathematicae Applicatae Sinica, English Series, 2022, 38: 235-253. DOI: 10.1007/s10255-022-1077-5.

[24] CANDES E J, FAN Y, JANSON L, et al. Panning for gold: 'Model-X' Knockoffs for high dimensional controlled variable selection[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2018, 80(3): 551-577. DOI: 10.1111/rssb.12265.

[25] ABOU RAFEE S A, UVO C B, MARTINS J A, et al. Large-scale hydrological modelling of the upper Paraná River basin[J]. Water, 2019, 11(5): 882. DOI: 10.3390/w11050882.

A random forest long-term precipitation prediction method combined with multiple hypothesis testing and its application

LI Mengjie[1], LIU Kun[1], MOU Hailei[2], YIN Zhaokai[1], LIU Zhiwu[1], WU Di[2], LIANG Lili[1]

(1. *Institude of Science and Technology, China Three Gorges Corporation, Beijing* 101199, *China*;
2. *China Three Gorges International Corporation, Beijing* 101199, *China*)

Abstract: Long-term precipitation prediction refers to forecasting precipitation over a period of more than one month. This is a crucial aspect of integrated water resources management. The accuracy of long-term precipitation predictions is low due to various uncertainties. Traditional long-term precipitation prediction methods are mainly divided into dynamical numerical methods and mathematical statistical methods. Dynamical numerical methods simulate future weather conditions using sea-land thermodynamic models for precipitation prediction. This approach has a clear physical mechanism, but the model calculations are complex. Data-driven mathematical-statistical

methods simulate the correlation between precipitation and predictors from a statistical perspective to establish a long-term prediction model. However, research on precipitation prediction based on mathematical statistical methods mainly focuses on improving the model, with relatively little emphasis on how to select the predictors. In fact, the predictors affect the accuracy of model predictions. Therefore, the focus and challenge of precipitation prediction lie in selecting the necessary predictors for modeling from the relevant factors. Random forest, as a flexible, efficient, and easy-to-use machine learning algorithm, has been widely used in hydrological prediction. The random forest method calculates the importance scores of various related factors and then selects predictors for the model based on empirical experience. This process can result in a certain error rate issue with the selected predictors.

To address the issue of false discovery rate in the random forest algorithm when selecting key predictors, this study employs the false discovery rate control method in multiple hypothesis testing to ensure quality control in predictor selection. This transformation shifts variable selection from being experience-dependent to becoming data-dependent. Finally, the random forest algorithm is used to construct a long-term precipitation prediction model by integrating the selected precipitation predictors. Taking the upper basin of the Parana River in Brazil as the study area, the precipitation from 54 measured rainfall stations and 130 climate system indices was analyzed. The predictors influencing precipitation in the corresponding months of the following year were obtained using the "Model-X Knockoff" method. A monthly precipitation prediction model is established based on the predictors that influence the precipitation for the corresponding month of the following year. The top 5 predictors with the highest importance scores are directly selected for random forest modeling using the traditional random forest method. The validity of the proposed method is subsequently verified using 10-fold cross-validation and a test of the monthly precipitation prediction results from 2018 to 2020.

The effect of 10-fold cross-validation for 54 rainfall stations shows that the model prediction pass rate of the method introduced is higher than that of the traditional random forest method from January to December, with the highest pass rate of 77% in June. The results of precipitation prediction from 2018 to 2020 indicate that our method achieved an average pass rate of 66% from January to December, outperforming the traditional random forest method, which scored 64%.

In summary, our research combines multiple hypothesis testing with predictor selection and quality control to establish a long-term precipitation prediction model, which differs from the traditional random forest method. This model exhibits a higher prediction pass rate and improved stability, suggesting that this approach can serve as an effective tool for long-term precipitation prediction in a basin.

Key words: random forest; long-term precipitation prediction; predictor selection; quality control; multiple hypothesis testing