覃卫坚,何莉阳,蔡悦幸. 基于两种机器学习方法的广西后汛期降水预测模型[J]. 气象研究与应用,2022,43(1):08-13.
Qin Weijian,He Liyang,Cai Yuexing. Prediction model of post-flood season precipitation in Guangxi based on two machine learning methods[J]. Journal of Meteorological Research and Application,2022,43(1):08-13.

基于两种机器学习方法的广西后汛期降水预测模型

覃卫坚，何莉阳，蔡悦幸
（广西壮族自治区气候中心，南宁 530022）

摘要：使用 1991—2021 年 7—9 月广西 90 个地面气象观测站降水量、NCEP/NCAR 月再分析资料和国家气候中心 BCC_CSM1.1 气候模式回报资料，研究建立基于粒子群-神经网络、随机森林算法的广西后汛期降水气候预测模型，并对 2016 年—2021 年预测进行应用试验。结果表明，基于粒子群-神经网络、随机森林算法的后汛期降水预测 Ps 得分比逐步回归方法分别提高了 2.78 分、2.5 分，比气候模式分别提高了 29.22 分、28.94 分，预测能力有明显的提升。

关键词：粒子群-神经网络；随机森林算法；经验正交函数；气候预测；汛期降水
中图分类号：P457.6 **文献标识码**：A **doi**：10.19849/j.cnki.CN45-1356/P.2022.1.02

引言

汛期气候预测为各级政府防灾减灾决策部署工作提供了技术支撑，是每年气象部门气候预测服务重中之重的工作任务。气候变暖背景下极端异常降水事件频发，进一步加大了旱涝预测的难度，因此开展汛期气候预测方法的研究具有重要的科学意义和应用价值。目前我国短期气候预测的科技水平和业务能力已从传统的统计分析发展到了动力-统计相结合的预测技术和方法，发展动力-统计相结合的气候预测方法是现阶段及未来很长时期内提高气候预测准确率的重要途径[1-4]。国家气候中心第二代季节预测模式系统(BCC_CSM1.1)预测能力较第一代得到了很大的提高，对大尺度环流预报能力较高[5-6]，对华南地区夏季降水量预测能力偏弱[7-8]。如何利用更有效的气候模式预测信息，就这个问题统计降尺度方法在气候预测中得到了应用，对气候模式具有较高预测技巧的大尺度环流信息和局地气象要素进行相关统计，建立预测模型，从而提高了气候预测能力[9-10]，如顾伟宗等[11]分别计算了预报对象和模式资料的预报因子场以及再分析资料的预报因子场的相关系数，利用最优回归方法建立预测模型，降水预测效果远高于模式直接输出的预测结果；封国林等[12]利用气候模式回报资料筛选出能反映模式预报误差分布特征的关键预报因子，通过计算检验得到最优多因子配置，建立汛期降水集成预测模型，提高了降水预测能力；郭渠等[13]利用 BCC_CSM 模式环流预测资料，建立多元回归预报模型，提高了夏季降水的预报技巧。以上统计降尺度方法主要使用传统的回归方法和集成建模预测，而把粒子群-神经网络和随机森林算法等机器学习方法应用其中还不多见。粒子群-神经网络等机器学习方法具有较强的处理非线性问题的能力，在气象预报中有了很好的应用效果，如陆虹等[14]、覃卫坚等[15-16]、孔庆燕等[17]、吴建生等[18]、田心如等[19]把粒子群-神经网络方法应用在广西冷湿天气、寒露风日数、降水量、夏季空调负荷预报中，预报准确度较线性回归方法有明显提高；Kim H L 和 Kim B H[20]把随机森林方法应用于城市洪水灾害等级预测中，预测准确率得到了提高。因此，利用 BCC_CSM1.1 气候模式回算资料，对广西后汛期降水距平百分率进行 EOF 分解，分别计算各模态时间系数和气候模式预测回算资料、气候模式回算资料和再分析资料的相关，得到高相关区域，使用逐步回归方法计算筛选得到预报因子，利用

收稿日期：2021-11-10
基金项目：广西自然科学基金(2019GXNSFAA245048)、广西科技计划项目(桂科 AB21075005)
作者简介：覃卫坚(1971—)，男，广西上林人，正研级高工，博士，主要从事短期气候预测方法研究。E-mail:qinweijian2008@126.com

粒子群–神经网络和随机森林算法进行建模预测，为提高后汛期降水预测率提供新的思路。

1 资料和方法

1.1 资料来源

资料包括：(1)1991—2021 年后汛期（7—9 月）广西 90 个地面气象观测站逐月降水距平百分率资料；(2)1991—2015 年 NCEP/NCAR 2.5°×2.5°格点月再分析资料，包括高度场、风场等；(3)1991—2021 年 BCC_CSM1.1 气候模式 6 月起报 7—9 月逐月回报数据，包括高度场、风场、降水距平值等。BCC_CSM1.1 气候模式是第二代季节气候预测模式系统，为一个包含海陆冰气系统、植被和碳循环的全耦合气候系统模式，模式分辨率为 2.5°×2.5°。

1.2 气候预测评分方法

Ps 评分计算公式：

$$Ps=\frac{2\times N_0+2\times N_1+4\times N_2}{N+N_0+2\times N_1+4\times N_2+M}\times 100 \quad (1)$$

其中：N_0 为气候趋势预测正确的站数，N_1 为一级异常预测正确的站数，N_2 为二级异常预测正确的站数，M 为没有预报二级异常而实况出现降水距平百分率≥100%或等于–100%的站数（称漏报站）。20%≤降水距平百分率绝对值<50%为一级异常，降水距平百分率绝对值≥50%为二级异常。

同号率指各站降水距平值实况和预报正负符号相同的站数占总站数的百分比。

1.3 粒子群–神经网络方法

Kennedy J and Eberhart R[21]1995 年提出了粒子群算法，粒子群–神经网络最优解的数学函数[17]：

$$\min_{a\leqslant\omega\leqslant b} E(\omega,\theta)=\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad (2)$$

$$\hat{y}_k=[1+e^{-(\sum_{i=1}^{n}x_i\omega_i+\theta_i)}]^{-1} \quad (3)$$

(2)–(3) 式中，ω 为网络权值，n 为样本数，θ 为网络阀值，x_i 为训练样本的输入，θ 实际输出，y_i 期望输出。粒子的位置更新调整公式[22]：

$$v_{ij}(k+1)=\omega\cdot(v_{ij}(k)+dir\times(c_1\times rand1\times(P_{ij}(k)-x_{ij}(k))+c_2\times rand2\times(P_{gj}(k)-x_{ij}(k)))) \quad (4)$$

式中，$dir=\begin{cases}-1 & f(s)<d_{low}\\ +1 & f(s)>d_{high}\end{cases}$

$$x_{ij}(k+1)=x_{ij}(k)+v_{ij}(k+1)\quad 1\leqslant i\leqslant m,1\leqslant j\leqslant n \quad (5)$$

$$f(s)=\frac{1}{S|R(k)|}\sum_{i=1}^{S}\sqrt{\sum_{j=1}^{n}(x_{ij}(k)-\bar{x}_j(k))^2} \quad (6)$$

(6)式中，$f(s)$为种群多样性指数，S 为种群中粒子总数，$|R(k)|=\max(|x_{ij}(k)|\,1\leqslant i\leqslant m,1\leqslant j\leqslant n)$，$n$ 为维数。$\bar{x}_j$ 为粒子第 j 维的平均值。

当 $f(s)<d_{low}$ 时，$dir=-1$，种群远离最优位置；当 $f(s)>d_{high}$ 时，$dir=1$，种群向整体最优位置靠拢。具体计算步骤[23]如下：

(1)初始化粒子群；

(2)计算每个粒子的适应度；

(3)随机输入个体最佳初始值及全局最佳初始值，再根据粒子的适应度进行更新；

(4)使用权重系数矩阵控制着网络权值和阀值的大小；

(5)连接结构矩阵变量矩阵控制着隐节点数，计算更新位置矩阵中的连接结构矩阵。

(6)反复进行(2)–(5)步骤的计算，当迭代次数达到了最大训练次数或满足最小训练误差时，停止计算，并输出最优解。

1.4 随机森林算法

Breiman[24]2001 年提出基于 bagging 思想的随机森林算法，是一种使用多棵决策树对样本进行训练和预测的分类器，它由不完全相同的单棵决策树组成，利用多棵决策树投票机制来决定最终的分类[25]。随机森林算法具有分类速度快、可调节参数少、计算效率高、减少过拟合现象等特征。设定含有 N 个样本的原始样本集，从原始样本集中随机抽样，组成多个训练集，建立 N 棵决策树[26]：

$$\{h(x,\theta_n),n=1,2,\cdots N\} \quad (7)$$

x 为输入的自变量和因变量，θ_n 为服从独立同分布随机向量。

在训练决策树模型的节点时，随机从节点上所有样本特征中选择一部分样本特征，以其中最优的一个特征来划分决策树的左右子树，训练结束后进行投票得到所有模态的平均值作为输出：

$$h(x)=\frac{1}{N}\sum_{n=1}^{N}h(x,\theta_n) \quad (8)$$

2 因子选取

2.1 EOF 分解

对 1991—2015 年广西 90 站后汛期降水距平百分率进行 EOF 分解，得到主要空间模态和各模态的时间系数。各特征向量能够反映出后汛期降水变化的空间结构，第一模态是后汛期降水变化最具有代表性的分布场，其次为第二模态、第三模态等，前三

个模态的方差贡献率分别为 51.4%、13.7%、7.2%，前三个模态累计方差贡献率达到了 72.3%，第四个模态方差贡献率仅为 4.1%，相对前三个模态方差较小，为了减少计算量，只计算前三个模态的时间系数。第一模态特征向量值基本为正值(图 1a)，体现了广西后汛降水的一致性变化这一重要特征。第二模态特征向量值呈桂北为正值、桂南为负值的空间分布特征(图 1b)，说明了桂南和桂北降水存在反相的变化特征。第三模态特征向量值桂西为正值、桂东为负值的空间分布特征(图 1c)，即桂西和桂东降水存在反相变化特征。从各模态时间系数历年变化来看，第一时间系数(PC1)1990—2010 年呈现出减小趋势，2011 年以后为增大趋势；第二时间系数(PC2)为减小趋势，其中 1991—2003 年变化幅度比较大，2003 年之后变化趋于平缓；第三时间系数(PC3)1991—2000、2010—2015 年变化比较平缓，2000—2010 年变化幅度大(图1d)。

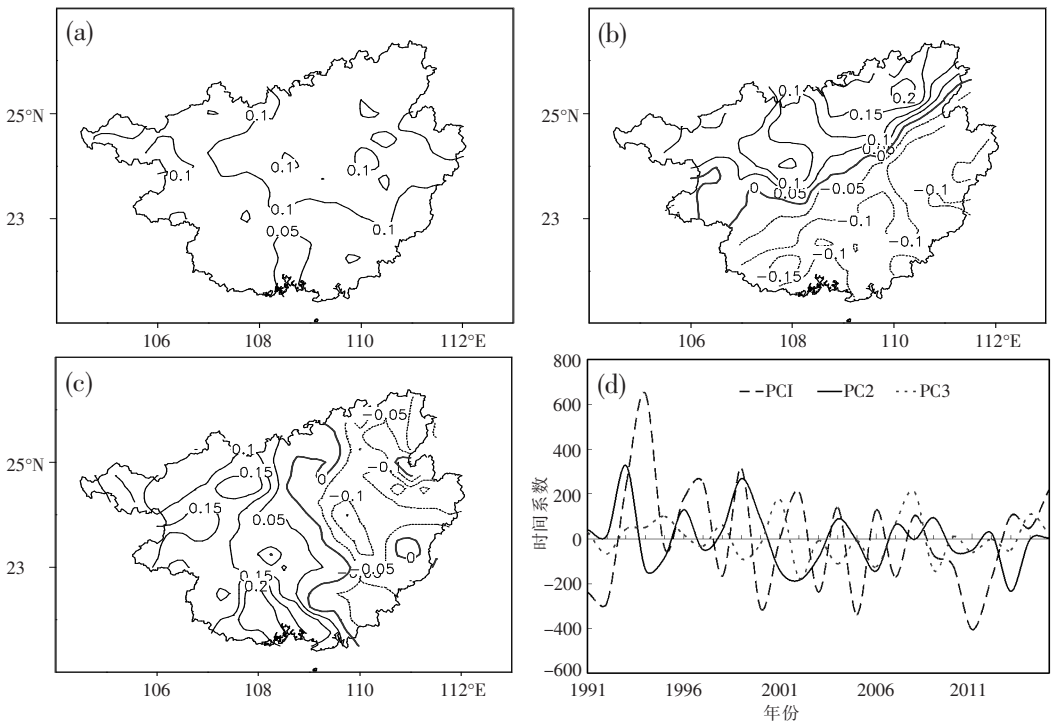

图 1 广西后汛期降水 EOF 前三个模态空间分布及时间系数

(a)EOF1；(b)EOF2；(c)EOF3；(d)PC1-3

2.2 预测因子的查找和筛选

因子查找从两个方面入手：一方面，计算后汛期降水距平百分率的前三个模态时间系数与 BCC_CSM1.1 模式6 月起报的环流预测资料的相关系数，得到显著相关的区域；另一方面，计算 BCC_CSM1.1 模式环流预测和 NCEP/NCAR 实况场的相关系数，得到模式环流预测的高技巧区，即相关系数通过水平为 0.05 的显著性检验区域。选出各模态时间系数与模式预测资料的相关显著区域，同时这区域也是模式预测高技巧区，把区域格点值进行平均后作为预选因子。为了保证在已选定的一批因子中得到最优的因子，使用逐步回归方法再进一步筛选，建立第一模态时间系数逐步回归预报方程：

$$y=515.521-194.793x_1+75.38x_2+315.965x_3$$

$$(F=4,\sigma=150.923,R=0.826) \quad (9)$$

式(9)中，x_1、x_2、x_3 分别为巴尔喀什湖和贝加尔湖之间区域、秘鲁西海岸附近、南非的 200hPa 经向风，如图 2a 所示。

第二模态时间系数逐步回归预报方程：

$$y=78904.57-3.023x_1-7.588x_2-240.934x_3$$

$$(F=2,\sigma=96.65,R=0.714) \quad (10)$$

式(10)中，x_1 为南非以南地区 200hPa 高度场，如图 2b 所示；x_2、x_3 分别为巴尔喀什湖以南附近地区、美国和墨西哥交界地区 500hPa 高度场，如图 2c 所示。

第三模态时间系数逐步回归预报方程：

$$y=61760-0.617x_1+119.974x_2-51.316x_3$$

$$(F=2,\sigma=73.629,R=0.678) \quad (11)$$

式(11)中，x_1 为贝加尔湖东部地区海平面气压，

如图 2d 所示；x_2 为澳大利亚南部 850hPa 纬向风，如图 2e 所示；x_3 为南美洲西部沿海 200hPa 经向风，如图 2f 所示。式(9)—式(11)σ 表示剩余标准差，R 表示复相关系数。

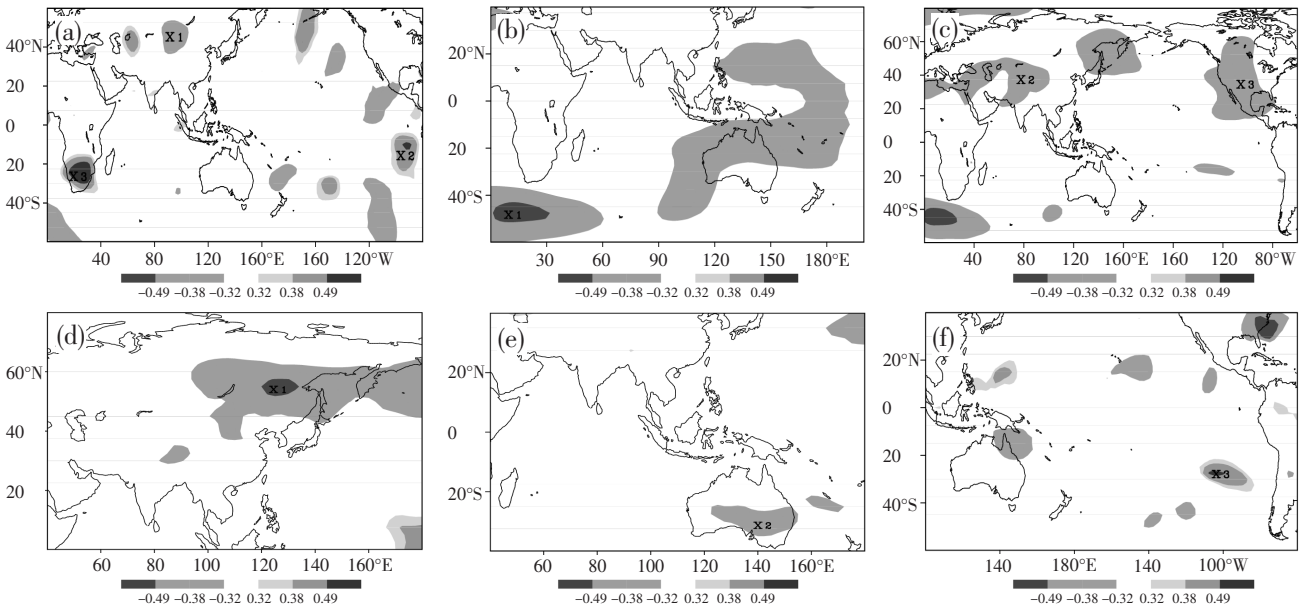

图 2　1991—2015 年第一模态时间系数与模式 200hPa 经向风的相关(a)，第二模态时间系数与模式 200hPa 高度场(b)、500hPa 高度场(c)的相关，第三模态时间系数与模式海平面气压(d)、850hPa 纬向风(e)和 200hPa 经向风(f)预测值的相关分析(阴影为通过 0.1 显著性水平检验的区域)

3　预测结果对比分析

通过逐步回归方程筛选得到预测因子，使用粒子群–神经网络和随机森林算法建立预报模型。粒子群–神经网络预报模型输出节点个数为 1，隐节点下限为 0.3，隐节点上限为 1.5，目标误差为 0.01，学习速率为 0.5，动量因子为 0.75，训练次数为 200，个体最优导向系数为 2，全局最优导向系数为 2，粒子位置下限为–3，粒子位置上限为 3，种群规模为 50，最大迭代次数为 100。随机森林算法策略树的数量为 50，构建决策树时对于节点数量没有限制，没有限制计算量，利用最大资源建模直至得到最优解。

利用粒子群–神经网络、随机森林算法、逐步回归方法对三个模态时间系数进行预测，得到 2016—2021 年各模态时间系数预报值，再与对应的特征向量相乘，最后合成得到降水距平百分率的预报场。表 1 给出了 2016—2021 年粒子群–神经网络、随机森林算法、逐步回归方法和气候模式的后汛期降水预测 Ps 得分和同号率，两种机器学习方法预测得分均高于逐步回归方法和气候模式，其中粒子群–神经网络方法平均得分最高，为 81.53，较逐步回归方法和气候模式分别提高了 2.78、29.22；其次为随机森林算法，平均得分为 81.25，较逐步回归方法和气候模式提高了 2.5、28.94；两种机器学习方法预测和

表 1　粒子群–神经网络、随机森林算法、逐步回归方法、气候模式后汛期降水预测 Ps 得分和同号率

年份	粒子群–神经网络		随机森林算法		逐步回归		气候模式	
	Ps	同号率	Ps	同号率	Ps	同号率	Ps	同号率
2016	74.13	0.59	75	0.6	75	0.6	69.57	0.53
2017	90.48	0.84	90.48	0.84	90.48	0.84	26.42	0.16
2018	81.58	0.69	81.58	0.69	66.67	0.5	74.13	0.59
2019	77.33	0.64	75.68	0.62	69.5	0.54	69.5	0.54
2020	68.53	0.52	67.63	0.52	73.68	0.52	63.7	0.48
2021	97.14	0.94	97.14	0.94	97.14	0.94	10.53	0.06
平均	81.53	0.70	81.25	0.70	78.75	0.66	52.31	0.39

实况同号率比逐步回归方法提高了 0.04、比气候模式预测提高了 0.31。从 6a 的预测试验来看,2017 和 2021 年气候模式预测误差较大,2017 年广西降水实况为偏多，而模式预测降水偏少;2021 年气候模式预测广西降水偏多,而实况是偏少。可见,利用模式有效的环流预测信息来建模预测，能够明显的提高降水的预测能力。

4 结论和讨论

利用 BCC_CSM1.1 气候模式预测等资料，使用相关方法查找和筛选得到预测因子，建立基于粒子群-神经网络、随机森林算法的广西后汛期降水气候预测模型。在 2016—2021 年业务预测试验应用中,基于粒子群-神经网络、随机森林算法的后汛期降水预测 Ps 得分较逐步回归方法分别提高了 2.78 分、2.5 分，较气候模式分别提高了 29.22 分、28.94 分,预测能力有明显的提升。

本文利用粒子群-神经网络、随机森林算法机器学习方法对气候模式降水进行订正，做了初步的预测试验,取得了良好的预测效果。这得益于本研究充分利用了气候模式有效的预测信息，在建模预测中机器学习算法具有自学习能力，较传统线性统计方法对复杂的非线性模型能够更准确的描述。在后续的研究中,将增加更多气候模式资料,做进一步的试验和研究。

参考文献:

[1] 魏凤英. 我国短期气候预测的物理基础及其预测思路[J].应用气象学报,2011,22(1):1-11.

[2] 宋连春,肖风劲,李威.我国现代气候业务现状及未来发展趋势[J].应用气象学报,2013,24(5):513-520.

[3] 贾小龙,陈丽娟,高辉,等.我国短期气候预测技术进展[J].应用气象学报,2013,24(6):641-655.

[4] 王会军,任宏利,陈活泼,等.中国气候预测研究与业务发展的回顾[J].气象学报,2020,78(3):317-331.

[5] 吴捷,任宏利,张帅,等.BCC 二代气候系统模式的季节预测评估和可预报性分析[J].大气科学,2017,41 (6):1300-1315.

[6] 刘芸芸,王永光,龚振淞,等.2020 年汛期气候预测效果评述及先兆信号分析[J].气象,2021,47(4):488-498.

[7] 张丹琦,孙凤华,张耀存.基于 BCC 第二代短期气候预测模式系统的中国夏季降水季节预测评估[J].高原气象,2019,38(6):1229-1240.

[8] 程智,高辉,朱月佳,等.BCC 第二代气候系统模式对东亚夏季气候预测能力的评估[J]. 气象,2020,46(11):1508-1519.

[9] 孙建奇,马洁华,陈活泼,等.降尺度方法在东亚气候预测中的应用[J].大气科学,2018,42 (4): 806-822.

[10] 陈丽娟,顾伟宗,伯忠凯,等.黄淮地区夏季降水的统计降尺度预测[J].应用气象学报,2017,28 (2):129-141.

[11] 顾伟宗,陈丽娟,李维京,等.降尺度方法在中国不同区域夏季降水预测中的应用[J].气象学报,2017, 70 (2):202-212.

[12] 封国林,赵俊虎,杨杰,等.中国汛期降水动力-统计预测研究[M].北京:科学出版社,2015:1-330.

[13] 郭渠,刘向文,吴统文,等.基于 BCC_CSM 模式的中国东部夏季降水预测检验及订正[J].大气科学,2017,41 (1):71-90.

[14] 陆虹,翟盘茂,覃卫坚,等.低温雨雪过程的粒子群-神经网络预报模型[J].应用气象学报,2015,26(5):513-524.

[15] 覃卫坚,陆虹,黄志,等.粒子群-神经网络法在广西寒露风日数预报中的应用[J].气象与环境学报,2015,31(6):158-162.

[16] 覃卫坚,李耀先,陈思蓉,等.粒子群-神经网络在华南夏季降水短期气候预测中应用研究[J].气象研究与应用,2015,36(2):1-7.

[17] 孔庆燕,史旭明,金龙.基于粒子群-支持向量机定量降水集合预报方法[J].数学的实践与认识,2017,47(5):219-225.

[18] 吴建生,刘丽萍,金龙. 粒子群-神经网络集成学习算法气象预报建模研究[J].热带气象学报,2008,24(6):679-686.

[19] 田心如,蔡凝昊,张志薇.基于气象因子及机器学习回归算法的夏季空调负荷预测[J] .气象科学,2019,39(4):548-555.

[20] Kim H L, Kim B H. Flood hazard rating prediction for urban areas using random forest and LSTM[J]. Journal of Civil Engineering,2020,24(12):3884-3896.

[21] Kennedy J, Eberhart R. Particle Swarm Optimization[C]//International Conference on Neural Networks, IEEE, 1995:1942-1948.

[22] Zhao H S,Jin L,Huang Y. An Objective Prediction Model for Typhoon Rainstorm Using Particle Swarm Optimization - Neural Network Ensemble[J]. Natural Hazards,2014,73 (2):427-437.

[23] 吴建生.基于粒子群算法的神经网络短期降水预报建模研究[J].智能系统学报,2006(2):67-73.

[24] Breiman L. Random Forests[J].Machine Learning,2001,45 (1):5-32.

[25] 唐宇迪,李琳,候惠芳,等.人工智能数学基础[M].北京:北京大学出版社,2020:1-7.

[26] 魏一钊,陈军锋.基于随机森林算法的冻融期土壤蒸发预报模型研究[J] .水电能源科学,2021,39(4):20-23.

Prediction model of post-flood season precipitation in Guangxi based on two machine learning methods

Qin Weijian, He Liyang, Cai Yuexing
(Guangxi Climate Center, Nanning 530022, China)

Abstract: Using the precipitation of 90 surface meteorological observation stations in Guangxi from July to September 1991 to 2021, NCEP/NCAR monthly reanalysis data and the National Climate Center BCC_CSM1.1 climate model return data, this paper established the precipitation climate prediction model of Guangxi in post-flood season based on particle swarm optimization neural network and random forest algorithm. An application test on the 2016-2021 forecasts were conducted. The results show that the PS score of precipitation prediction in post-flood season based on particle swarm optimization neural network and random forest algorithm is 2.78 and 2.5 points higher than that of stepwise regression method, 29.22 and 28.94 points higher than that of climate model, and the prediction ability is significantly improved.

Key words: particle swarm-neural network; random forest algorithm; empirical orthogonal function; climate prediction; precipitation in flood season