大气科学学报 2022年 第45卷 第2期:191-202
论 著 http://dqkxxb.cnjournals.org

机器学习方法在湖南夏季降水预测中的应用

黄超[①②],李巧萍[③*],谢益军[①②],彭嘉栋[①②]
① 湖南省气候中心,湖南 长沙 410118;
② 气象防灾减灾湖南省重点实验室,湖南 长沙 410118;
③ 中国气象局地球系统数值预报中心,北京 100081
*联系人,E-mail:liqp@cma.gov.cn

2021-09-03 收稿,2021-12-10 接受
湖南省气象局研究型业务预报预测专项(XQKJ21C011);中国气象局预报员专项(CMAYBY2020-087);国家重点研发计划项目(2018YFC1505806)

摘要 利用湖南97个国家站的逐月降水资料、国家气候中心130项气候指数集以及国家气候中心和美国国家环境预报中心两套季节预测模式的降水预测资料,采用递归特征消除法确定预测因子并使用多层前馈神经网络、支持向量回归和自然梯度提升三种算法建立了两种湖南夏季降水统计预测方案的模型,检验了预测效果。结果表明:基于机器学习的预测模型对湖南夏季雨型分布有较好的预测能力,两种统计方案提前1~6 mon起报的夏季降水平均距平相关系数分别为0.15和0.19,相比于NCEP和NCC模式有较大提升,平均PS评分分别为69.3和69.2,高于NCC模式的63.1,略低于NCEP模式的71.5;进一步分析表明,3—5月起报的机器学习模型的预测技巧可能来源于前冬极地和中高纬环流,12—2月起报的模型预测技巧则可能来自海温的前兆信号。

关键词
机器学习;
夏季降水;
预测

湖南地形具有三面环山、南高北低的特点,气候复杂多变,夏季旱涝转换(李易芝等,2017),易出现洪涝、干旱等气象灾害。在全球变暖背景下,湖南夏季极端降水明显增加(周莉等,2018),因此进一步提升夏季降水预测水平对湖南防灾减灾具有重要现实意义。

目前降水季节趋势预报主要分为统计学、动力学和动力统计相结合三类方法。统计方法充分利用历史资料规律,选取有明确物理意义和显著相关的因子进行建模。范可等(2007)通过前期因子建立统计模型对长江中下游夏季降水年际增量进行预测,显著提高了业务预测技巧。杜良敏等(2016)针对不同气候分区建立统计模型对我国夏季降水进行预测。李春晖等(2018)采用时空投影方法建立广东省降水统计预测方法。Yim et al.(2014)使用统计模型对中国南方夏季降水进行预测。但由于各预测因子相互作用过程复杂,不同时间尺度的预测信号不一致,加大了预测的难度。随着数值模式的发展,动力模式成为气候预测的主要工具,许多国家建立了数值预报模式(丁一汇,2011)。近年来,我国季节预测模式对大气环流、ENSO(El Niño-Southern Oscillation)现象、亚洲夏季风等的预测能力已有明显提升(吴捷等,2017),但对降水预测技巧依然有限,特别是对东亚地区夏季降水的预报技巧相对较低(王予等,2021)。在这样的现实情况下,专家学者在此基础上发展了动力和统计相结合的预测方法(封国林等,2013),充分利用历史资料并考虑大气海洋物理机制,进一步提高了降水预测准确率。柯宗建等(2009)提出了最优子集回归方法。贾小龙等(2010)发展了变形典型相关分析(Combination of Empirical Orthogonal Function and Canonical Correlation Analysis,BP-CCA)方法。舒建川等(2019)在此基础上使用BP-CCA方法在西南地区进行了应用。组合统计降尺度方法(Liu and Fan,2014;刘颖

引用格式:黄超,李巧萍,谢益军,等,2022.机器学习方法在湖南夏季降水预测中的应用[J].大气科学学报,45(2):191-202.
Huang C,Li Q P,Xie Y J,et al.,2022.Prediction of summer precipitation in Hunan based on machine learning[J].Trans Atmos Sci,45(2):191-202.doi:10.13878/j.cnki.dqkxxb.20210903001.(in Chinese).

等,2020)也能够提升一定的降水预测技巧。此外,国家气候中心多模式解释应用集成预测系统(Multi-model Downscaling Ensemble Prediction System,MODES)(刘长征等,2013)和动力-统计相结合的季节预测系统(Forecast System on Dynamic-Analogue Combined Skills,FODAS)(王启光等,2011)的研发对我国夏季降水预测业务水平提升起到了关键作用。

机器学习强调从历史数据中学习规则,对新数据进行推理和预测。区别于传统统计方法,机器学习擅长处理非线性问题,利用机器学习的优势可以从地球系统中发现并提取新的相互关联信号(贺圣平等,2021)。近年来,机器学习在气象领域的应用越来越广泛,常用的机器学习算法有支持向量机、贝叶斯算法、神经网络、决策树算法等(冯汉中和陈永义,2004;孙照渤等,2013;张宇彤等,2013;苗春生等,2017)。随着计算能力的提高和深度学习理论的发展,以卷积神经网络(Convolutional Neural Networks,CNN)和长短期时间记忆网络(Long Short-Term Memory,LSTM)为代表的深度学习方法在气候领域得到应用,例如 CNN 算法对 ENSO 指数的预测技巧超过了主流动力模式(Ham et al.,2019),沈皓俊等(2020)采用的 LSTM 算法对中国夏季降水预测评分超过了同期业务模式。

湖南夏季降水时空分布不均,影响因子复杂,当前对其机理和预测的研究还存在短板,动力模式预测水平与业务服务需求存在差距,有必要利用机器学习的优势进一步提高当地预测水平。考虑到湖南降水观测资料年份较少,不适合深度学习方法,因此本文采用随机森林算法进行递归特征消除来挑选预测因子,使用多层前馈神经网络、支持向量回归和自然梯度提升方法建模,结合动力模式降水预测结果,建立适用于湖南本地的夏季降水统计预测方法。

1　资料和方法

1.1　数据来源和预处理

预报因子资料来源于国家气候中心提供的气候系统监测指数集(下载地址:http://cmdp.ncc-cma.net/Monitoring/cn_index_130.php),共包含 130 项气候系统指数的月平均值。其中大气环流指数 88 项,主要包括副高、东亚槽、极涡、欧亚环流型、遥相关、太平洋信风等大气环流指数。海温指数 26 项,主要包括厄尔尼诺(各区及类型)、暖池、印度洋、亲潮区、黑潮区等海温指数。其他指数 16 项,主要包括冷空气、台风、南方涛动、北太平洋年代际振荡、准两年振荡、次表层海温等指数。时间尺度为 1980 年 1 月—2020 年 12 月,若出现缺测,直接将该因子剔除。

美国国家环境预报中心(National Centers for Environmental Prediction,NCEP)和国家气候中心(National Climate Center,NCC)气候预测模式数据来自 MODES 系统,空间分辨率均为 1°×1°,NCEP 模式历史回算时间范围为 1982—2020 年(其中 2011 年资料缺失),模式气候态取 1982—2010 年,NCC 模式历史回算时间范围为 1991—2020 年,气候态取 1991—2010 年。分别计算模式不同起报时间的夏季(6—8 月)降水距平百分率,并采用双线性插值将网格数据插值到站点上。实况夏季降水资料来自湖南省 97 个国家站 1981—2020 年的观测数据,夏季降水没有明显的线性趋势,因此未做去趋势处理,直接处理成降水距平百分率进行分析。

样本集共包含 1981—2020 年共 40 a、6 个起报时间、10 个模态共计 2 400 个样本(40×6×10);根据起报时间和模态划分为 60 个子样本集,每个子样本集 40 个样本;训练集时间段为 1981—2010 年,测试集为 2011—2020 年。建模时挑选对应起报时间和模态的样本集,其中训练集 30 个样本进行训练和交叉验证,测试集 10 个样本进行独立检验。

1.2　评估方法

对湖南夏季降水评价指标采用趋势异常综合评分 P_S(PS)和空间距平相关系数 A_{CC}(Anomaly Correlatiom Coefficent,ACC)。PS 评分的计算公式为:

$$P_S=\frac{N_0+P_1\times N_1+P_2\times N_2}{N+P_1\times N_1+P_2\times N_2}\times 100。$$

其中:N 为总站数,本研究中取 97;$P_1=0.5$,$P_2=1.0$;N_0 为预报与实况距平符号相同站数或符号不同但相差只有 1 级站数之和;N_1 为预报与实况同为 2 级 5 级的站数,N_2 为预报与实况同为 1 级、6 级的站数。

表 1　降水趋势预报分级标准

Table 1　Grading standards for precipitation anomaly

级别	降水距平百分率/%
1	$\nabla R\leqslant -50$
2	$-50<\nabla R\leqslant -20$
3	$-20<\nabla R<0$
4	$0\leqslant \nabla R<20$
5	$20\leqslant \nabla R<50$
6	$\nabla R\geqslant 50$

ACC 的计算公式为:

$$A_{CC}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(o_i-\bar{o})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(o_i-\bar{o})^2}}。$$

其中:n 为站点数,y_i 和 o_i 分别表示预测值和观测值;$\bar{y}$ 和 $\bar{o}$ 分别表示预测值和观测值的平均值。

1.3 建模方法

为了减少建模过程中机器学习算法的随机性影响,本文采用多层前馈神经网络、支持向量机和决策树集成三种不同的机器学习方法进行建模,这三种算法均属于机器学习中的分类和回归方法,对数据的识别和拟合过程具有一定差异。

1)多层前馈神经网络

本文使用的神经网络算法为多层前馈神经网络,其一般包含输入层、隐含层和输出层(韩力群,2006;LeCun et al.,2015)。隐含层越多,模型数据表示能力越强,更易造成过拟合,因此本文仅采用两层隐含层,神经元个数也不超过预报因子个数。模型的表达式为:

$$P_k=g_2\left[\sum_{j=1}^{m}w_{kj}g_1\left(\sum_{i=1}^{n}w_{ji}x_i+w_{j0}\right)+w_{k0}\right]。$$

其中:x_i 为节点 i 的输入值;P_k 为节点 k 的输出值;g_1 为隐含层激活函数;g_2 为输出层激活函数;m 和 n 分别为输入层和输出层神经元个数;w_{j0}为隐含层第 j 个神经元的偏差;w_{k0}为输出层第 k 个神经元的偏差;w_{kj}为输出节点 k 与隐含节点 j 的权重;w_{ji}为输入节点 i 与隐含节点 j 的权重。

2)支持向量回归

支持向量回归是支持向量机的拓展,算法通过核函数在高维或有限维空间中构造一个或一组超平面使数据与其距离最小(陈永义等,2004),在处理小样本、高维和非线性问题上具有优势。本文选用高斯核函数,因此表达式为:

$$f(x)=\sum_{i=1}^{L}(a_i-a_i^*)K(x,x_i)+b。$$

$$K(x,x_i)=e^{-\frac{\|x-x_i\|^2}{2\sigma^2}}。$$

其中:L 为支持向量的个数;a_i、a_i^*、b 为通过训练样本确定的最优超平面参数;x_i 为预报因子;σ 为控制高斯核参数宽度的参数。

3)决策树集成

决策树是机器学习中的分类回归算法,对于回归问题,算法目标是尽量使划分同一类别的平方误差最小,但也易造成过拟合,可通过决策树集成方法克服。本文使用的随机森林和自然梯度提升树均属于决策树集成算法。随机森林回归算法通过对训练集重复随机采样进行决策树建模,取多个决策树平均值作为预测结果(Breiman,2001);而自然梯度提升树算法通过梯度提升方法进行预测,不断对预测残差进行建模并集成多个决策树,从而达到减少预测误差的目(Peng et al.,2020)。

4)递归特征消除法

递归特征消除法是机器学习中常用的特征处理方法,起到挑选重要因子的作用。该方法通过反复构建模型剔除重要程度最低的因子,并遍历所有因子达到确定因子重要程度的目的。本文采用的重要性衡量方法为基尼重要性,在随机森林内部节点中通过反复将数据集分为两个独立的集合,计算每次分类后的集合内部方差,依据分类前后集合的方差差值确定气候因子的重要性,方差差值越大表示因子重要性越高。

2 机器学习在降水预测中的应用

2.1 湖南夏季降水预报方案

将机器学习方法应用预测因子筛选及湖南夏季降水预测建模中,图 1 给出了降水预测的主要流程:

1)资料处理:分为三部分,第一部分获取前期因子集,将起报时间前 3 mon 的 130 项气候系统指数作为前期因子集,例如 5 月起报的模型使用的是 2、3、4 月的气候因子。第二部分是降水数据,对 1981—2010 年湖南夏季观测降水的距平百分率采用经验正交函数分析方法(EOF)进行分解,时间系数为预测目标。第三部分是模式预报降水场,使用观测降水场 EOF 分解后的空间系数对模式降水场进行投影,得到模式预报的时间系数。

2)筛选关键因子组合:基于机器学习的递归特征消除思想,采用随机森林算法获取重要的气候因子,再通过交叉验证选取合适的关键因子组合。

3)建模:分为两个方案,方案一直接使用筛选的关键因子组合与 EOF 时间系数进行建模;方案二将 NCEP 模式预报的降水场时间系数与方案一中的因子共同作为预报因子进行建模。

4)输出预报结果:利用模型预报的时间系数和观测降水 EOF 分解的空间系数还原成预报降水场,对不同机器学习模型的预报结果进行等权集合平均作为最终的预报结果。

2.2 确定预测因子及 EOF 模态个数

使用随机森林算法进行递归特征消除来筛选预

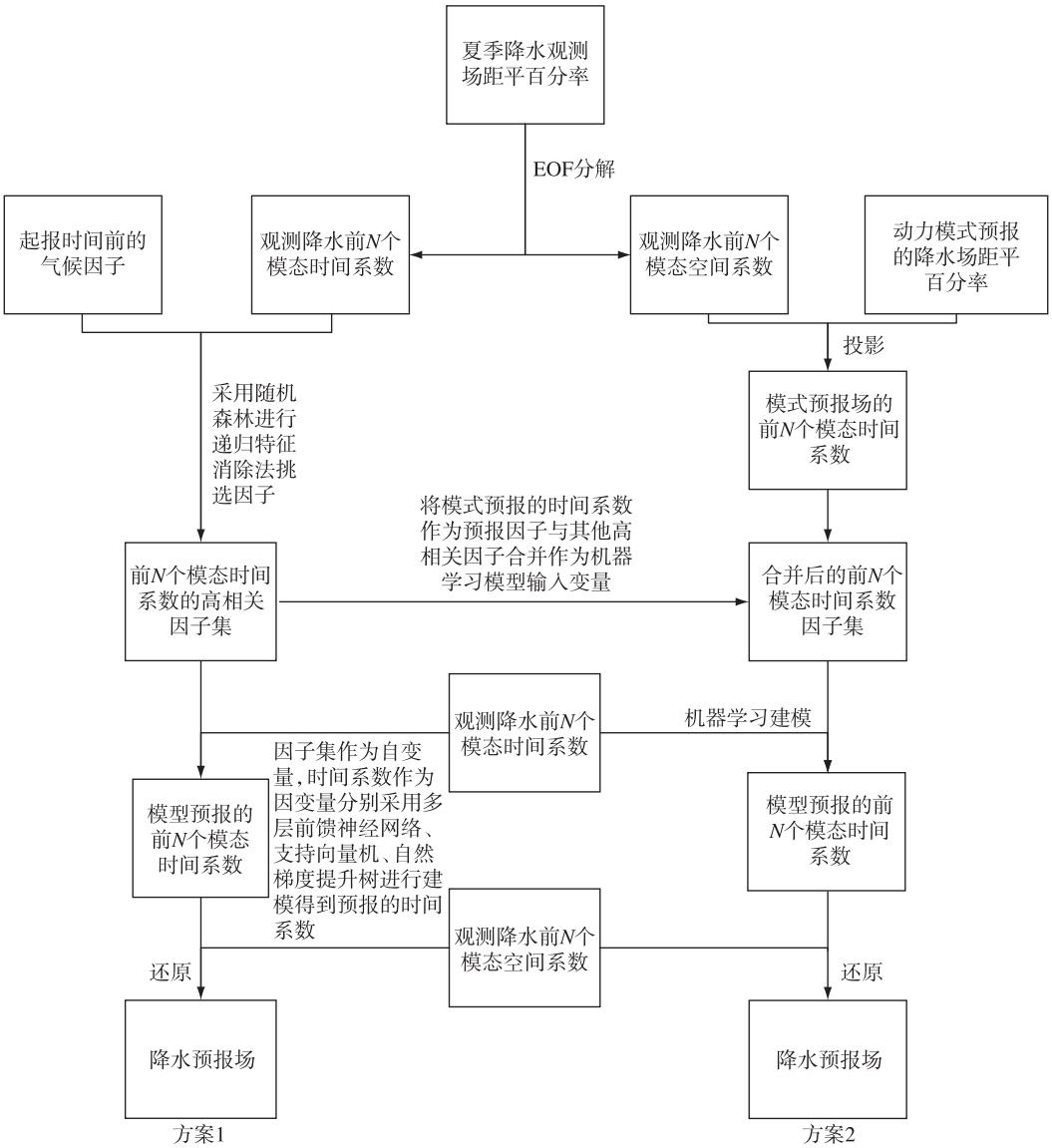

图 1　利用机器学习方法建立夏季湖南降水预测模型流程图

Fig.1　Flowchart for building the prediction model with the machine learning method

测因子,随机森林算法通过计算预测因子的基尼重要性对其进行排序,从而剔除不重要的预测因子,达到降维的目的。将所有候选气候因子与前 10 个 EOF 模态时间系数分别进行递归特征消除(决策树数量参数设置为 100,持续增大后误差并无显著减少),采用五折交叉验证进行误差分析。图 2 给出了 5 月起报的前 10 个模态通过递归特征消除法剔除因子后均方根误差,当因子数达到某一阈值,误差趋于平稳。在因子重要性排序之后,选取该阈值之前的关键性因子即能在降低模型误差的同时达到降维的目的,依据此方法便得到提前 1~6 mon 起报和不同模态的预测因子(表 2)。

采用交叉验证方法分析不同 EOF 模态个数对预测结果的影响。图 3 给出了利用 1981—2010 年湖南夏季降水数据分别截取前 1~20 个不同 EOF 模态进行五折交叉验证的结果,ACC 和 PS 评分均为提前 1~6 mon 起报的平均值。从图中可以看出,EOF 模态个数超过 6 个,PS 和 ACC 变化趋于平稳,当 EOF 个数取 8 和 10 时,ACC 和 PS 分别达到最大值。结合图 3 中所示的不同 EOF 模态的累计方差贡献率以及还原后降水场与观测场的相关系数来看,EOF 模态个数越多,越能反映降水的时空变化,因此这里将 EOF 模态个数定为 10,此时累计方差贡献率为 89.1%,ACC 和 TCC 分别达到 0.86 和 0.94,能够反映降水的时空分布特征。

2.3　参数设置及模型建立

表 3 中给出了多层前馈神经网络、支持向量机回归、自然梯度提升树三种算法的参数范围,为避免

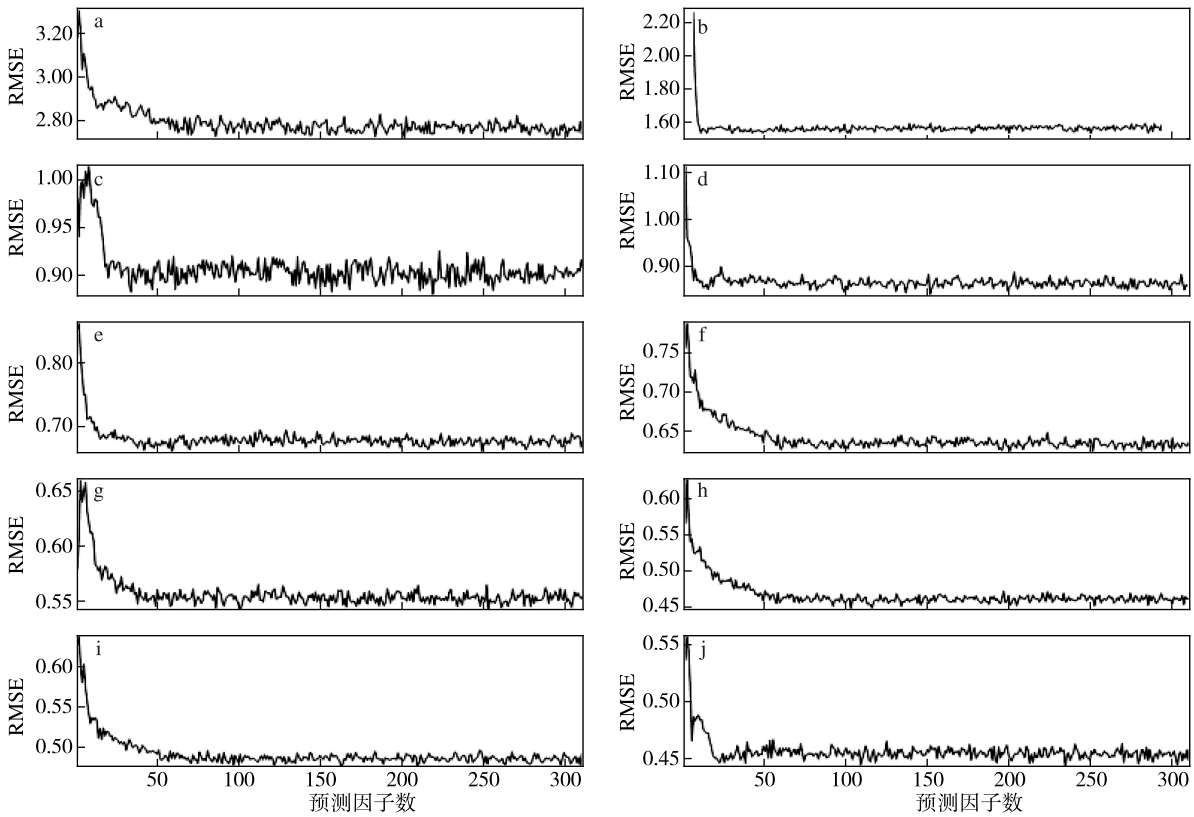

图 2 机器学习算法挑选的 5 月起报的前 10 个(a—j)EOF 模态的因子数量对应的均方根误差

Fig.2 RMSE corresponding to the number of factors of (a—j) different EOF patterns selected by the machine learning method initialized in May

表 2 不同模态提前 1~6 mon 起报的机器学习算法选取的因子个数

Table 2 Numbers of factors selected by the machine learning method with the first 10 EOF modes at different lead times(months)

EOF 模态	Lead-1	Lead-2	Lead-3	Lead-4	Lead-5	Lead-6
1	70	59	15	9	40	9
2	6	13	3	40	45	30
3	34	12	14	40	10	5
4	15	14	3	26	31	38
5	42	26	63	39	50	17
6	60	40	44	36	40	54
7	38	29	50	34	24	5
8	49	40	30	2	6	11
9	51	38	36	24	20	25
10	22	20	22	36	13	26

过拟合,参数设置尽量简单,降低模型复杂度,所有数据进行标准化处理。神经网络层数为 2 层,节点数取 20~50,激活函数使用 Relu;支持向量机使用高斯核;自然梯度提升树的决策树数量在 20~500 间取值。建模时取训练集对参数取值范围内的不同参数组合分别建模。例如对 5 月起报的 EOF 第一模态时间系数使用多层前馈神经网络建模时,隐含层数量为 2,对应隐含层节点数分别为{50,50}、{40,40}、{30,30}、{20,20}共 4 组,正则化参数分别为 0.000 1、0.001、0.01、0.1、1 共 5 个,不同参数组合共计 20 个;然后采用五折交叉验证方法计算得到 20 个模型的平均均方根误差,其中隐含层节点数

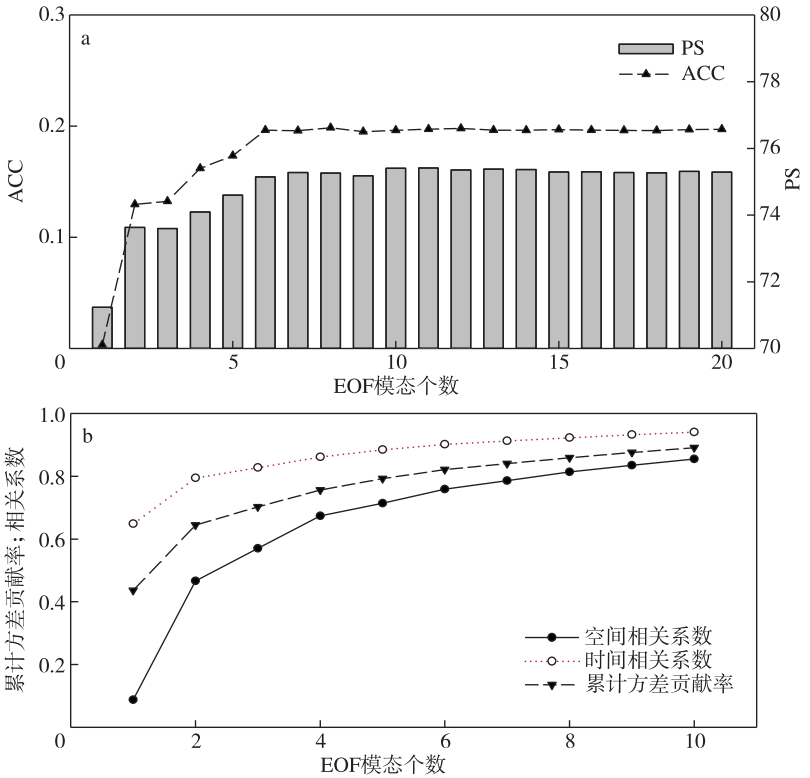

图 3　不同 EOF 模态个数交叉验证评分及与观测降水的相关系数:(a)分别取前 1~20 个 EOF 模态的交叉验证 ACC 和 PS 评分;(b)取前 1~10 个 EOF 模态的累计方差贡献率及其还原的降水与观测降水之间的距平相关系数(ACC)和时间相关系数(TCC)

Fig.3　Cross-validation scores of different EOF modes and correlation coefficients with observed precipitation:(a) ACC and PS scores of the first 1 to 20 EOF modes;(b) cumulative variance contribution rate of the first 1 to 10 EOF modes, and the anomaly correlation coefficient and temporal correlation coefficient between the restored and observed precipitation

为{40,40}、正则化参数为 0.01 的模型误差最小,作为最终预测模型;重复该步骤即得到 2 种方案 3 种算法 10 个模态提前 1~6 mon 起报的共计 360 个预测模型。

对比三种算法不同起报时间的平均均方根误差(图 4a),提前 1、3、4 mon 起报的模型中支持向量回归误差最小,提前 2、5、6 mon 起报的模型中,自然梯度提升树误差最小;通过不同模态的平均均方根误差来看(图 4b),预测误差主要位于前两个模态。

表 3　模型参数设置

Table 3　Model parameters

模型	参数取值范围	数据标准化
多层前馈神经网络	隐含层数量:2 节点数:{50,40,30,20} 正则化参数:{0.000 1,0.001,0.01,0.1,1} 激活函数:Relu	Z-Sore
支持向量机回归	核:高斯核 高斯核宽度:{0.001,0.01,0.1,1,10,100} 正则化参数:{0.001,0.01,0.1,1,10,100}	
自然梯度提升树	决策树数量:{500,400,300,200,100,50,20}	

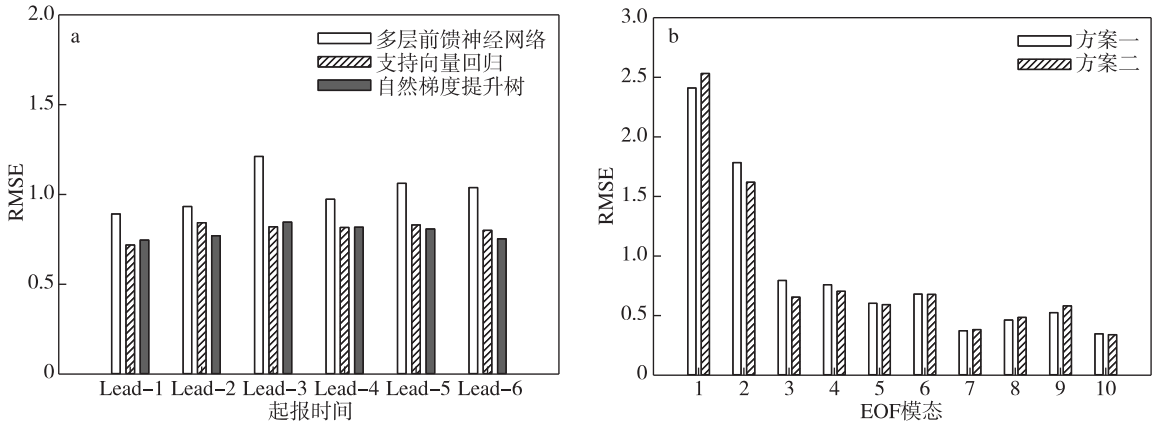

图 4 三种算法不同起报时间(a)和两种预测方案前 10 个模态(b)的均方根误差

Fig.4 RMSE of (a) three algorithms with lead times of 1—6 mon and (b) first 10 EOF modes of two prediction schemes

3 模型的预报技巧评估

采用基于上述机器学习方法建立的湖南夏季降水预报模型,两种统计方案分别使用 2011—2020 年和 2012—2020 年数据进行独立样本检验,得到对应年份 97 个国家站夏季降水距平百分率数据并评估预报技巧。图 5 分别给出气候模式本身及两种方案的统计模型提前 1~6 mon 起报的降水距平空间相关系数(ACC)和 PS 评分。整体来看,方案一提前 1~6 mon 起报的 ACC 分别为 0.25、0.15、0.09、0.23、0.15、0.05,平均为 0.15;方案二提前 1~6 mon 起报的 ACC 分别为 0.25、0.23、0.19、0.26、0.24、-0.01,平均为 0.19;NCEP 和 NCC 模式预报的平均 ACC 分别为 0.08 和 0.02,统计方案有明显提高;两种方案提前 1~6 mon 起报的平均 PS 评分分别为 69.3 和 69.2,相比 NCEP 模式的 71.5 略低,但优于 NCC 模式的 63.1。从不同起报时间来看,2 月起报(Lead-4)的 ACC 最高,4 月起报(Lead-2)的 PS 评分最高。与动力模式结果相比,机器学习模型的平均 ACC 比 NCEP、NCC 模式高,这种优势在提前 3~6 mon 起报的模型上更加明显,两套动力模式在提前 3 mon 以上预报夏季降水几乎没有技巧,但在 PS 评分上,NCEP 模式则更具优势。上述结果说明两种基于机器学习的预测方法在降水空间分布的预测技巧上有优势,并且方案二比方案一效果更好,但在降水异常级的预测能力上比 NCEP 的动力模式要差,可能因为统计方法更加倾向于预测平均状态,对降水异常级的预测能力不足。

图 6 给出了方案一的 2011—2020 年和方案二的 2012—2020 年逐年夏季降水预测 ACC 和 PS 评分,可见预测评分表现出明显的年际差异。两种方案的预测模型在 2012、2013、2016、2018 年提前 1~6 mon 起报的平均 PS 评分均超过 70,预测效果较好,对应的 ACC 评分在上述年份也较高。相比之下,方

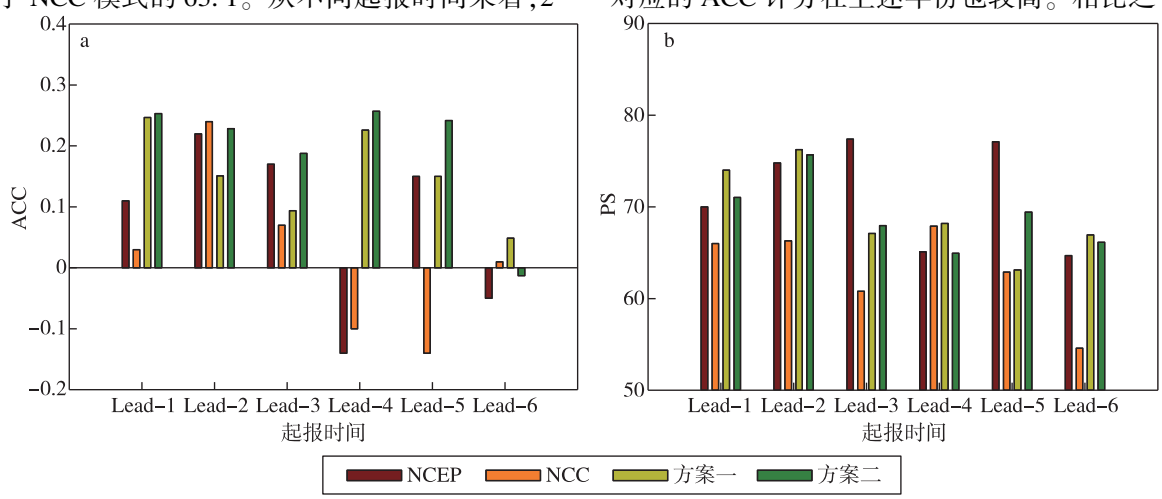

图 5 不同方案预测的湖南夏季降水 ACC(a)和 PS 评分(b)

Fig.5 (a) ACC and (b) PS scores of summer precipitation in Hunan under different prediction schemes

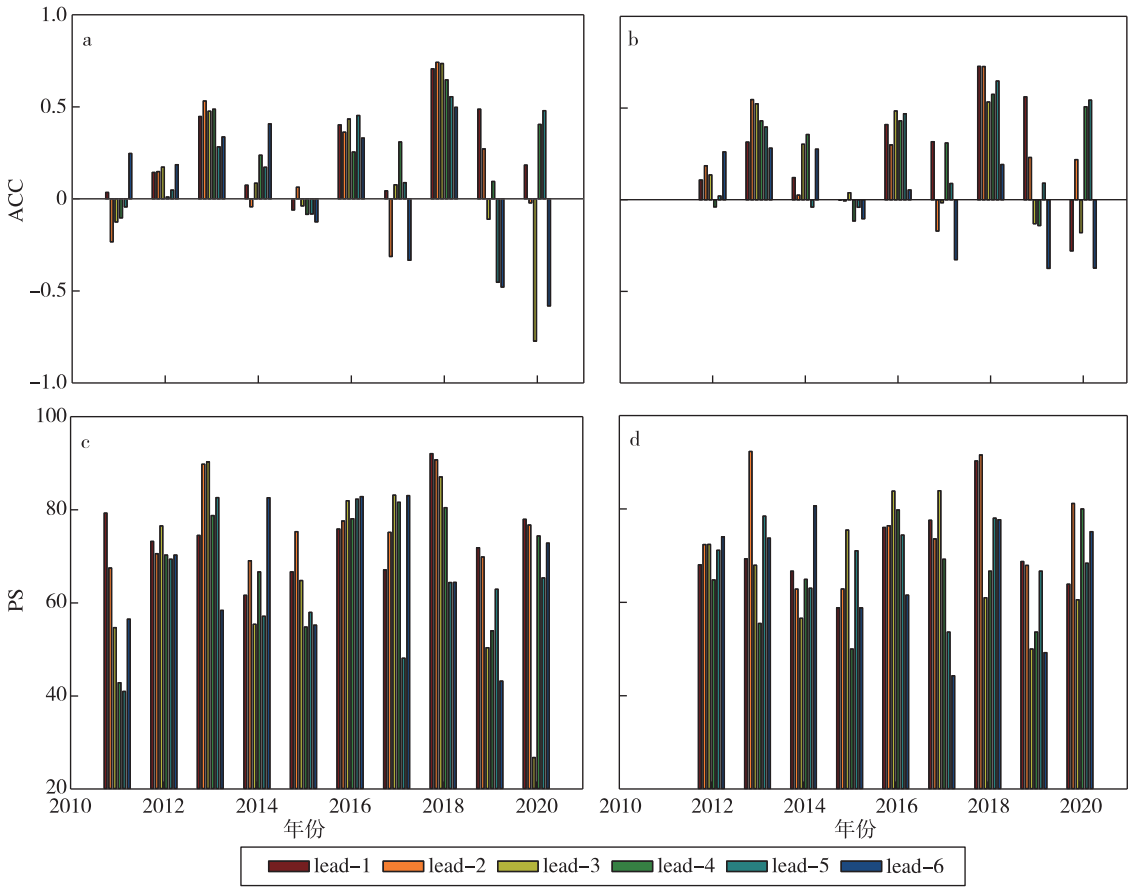

图 6　方案一(a、c)和方案二(b、d)不同起报时间的机器学习模型预测的历年湖南夏季降 ACC(a,b)和 PS(c,d)评分

Fig.6　(a,b) ACC and (c,d) PS scores of summer precipitation in Hunan predicted by the machine learning models with lead times of 1 to 6 months in (a,c) scheme Ⅰ and (b,d) scheme Ⅱ

案二的平均 ACC 除 2015 年为负值,其余年份均为正值,整体预测技巧高于方案一。众所周知,ENSO 是热带太平洋地区海气系统年际气候变率最强信号(宗海锋等,2010;Wen el al.,2015),ENSO 对湖南降水预测有重要指示意义,分析机器学习模型在 ENSO 年的预测情况可以进一步了解其预测水平。2016 和 2020 年为典型的厄尔尼诺衰减年,两种方案在 2016 年不同起报时间的预测均有较高的正技巧,平均 ACC 分别达到 0.37 和 0.36,PS 评分分别达到 79.8 和 75.3,但 2020 年的预测并不稳定,3 月起报的降水预测为评分较低。从拉尼娜衰减年预测来看,2013 和 2018 年的预测均有较高的正技巧;整体来看,机器学习建模方法在一定程度上能够识别 ENSO 对湖南降水的影响。

4　讨论

上述结果表明机器学习方法能够改善湖南夏季降水空间分布的预测技巧,但机器学习算法通常属于黑箱模型,在解决气候问题时难以给出合理的物理过程解释,为了能够进一步了解机器学习方法预报技巧的来源,同时考虑到相近起报月份的预测因子相近,这里仅给出 3 月和 12 月起报的预测因子中前 4 个模态通过显著性检验的因子(表 4)。可以看出,3 月起报模型的前 3 个模态相关显著的预测因子以前冬极地和中高纬环流指数为主,第四个模态中的南方涛动和赤道中东太平洋 200 hPa 纬向风指数均反映与 ENSO 的高相关性,并且 4 月和 5 月起报的预测因子具有相似特点。12 月起报的模型因子第一模态与前期东亚槽和西太副高位置有显著相关,后 3 个模态与海温相关显著,2 月和 1 月起报的预测因子也与海温显著相关。这可能说明,3—5 月起报的模型预报技巧主要来自前冬极地和中高纬环流的信号,而 12 月—次年 2 月起报的模型预报技巧主要来自前期海温,而这些因子如何影响湖南降水还需要进一步研究。

表 4 机器学习方法挑选的 EOF 前 4 个模态中相关系数通过置信度为 95%的显著性水平检验的预测因子

Table 4 Factors of the first four EOF modes selected by the machine learning method and its correlation coefficients that are significant at the 95% confidence level

EOF 模态	3 月起报		12 月起报	
	因子名称	相关系数	因子名称	相关系数
1 全区一致型	2 月西藏高原-2 指数	-0.43	10 月东亚槽位置指数	-0.45
	2 月太平洋区极涡面积指数	-0.38	10 月西太平洋副高西伸脊点指数	0.32
	1 月太平洋区极涡面积指数	-0.33		
	1 月北美区极涡面积指数	-0.31		
2 南北反向型	12 月东大西洋-西俄罗斯遥相关型指数	-0.40	11 月 Niño 1+2 区海表温度距平指数	-0.47
	2 月太平洋-北美遥相关型指数	-0.39	11 月 850 hPa 东太平洋信风指数	0.43
	2 月亚洲区极涡强度指数	0.32	11 月 850 hPa 中太平洋信风指数	0.43
			11 月大西洋欧洲区极涡面积指数	-0.37
			11 月欧亚纬向环流指数	-0.36
			10 月 850 hPa 东太平洋信风指数	0.31
3 西北-东南三极子型	12 月欧亚纬向环流指数	0.36	9 月西风漂流区海温指数	-0.61
	1 月大西洋多年代际振荡指数	-0.34	11 月热带南大西洋海温指数	0.36
	12 月西藏高原-1 指数	0.34		
	1 月 50 hPa 纬向风指数	0.33		
	12 月斯堪的纳维亚遥相关型指数	-0.32		
4 西北-西南三极子型	1 月赤道中东太平洋 200 hPa 纬向风指数	0.33	10 月北太平洋副高北界位置指数	0.50
	2 月北大西洋-欧洲区极涡强度指数	0.35	10 月西太平洋副高脊线位置指数	0.40
	2 月南极涛动指数	0.34	9 月类 ENSO 指数	-0.39
			10 月黑潮区海温指数	0.38
			10 月北半球副高北界位置指数	0.35
			9 月 850 hPa 西太平洋信风指数	0.34
			11 月大西洋副高脊线位置指数	-0.34

本文仅考虑了起报时间前 3 mon 的气候因子,相关研究表明湖南夏季降水与前冬的大气海洋状态有重要联系(李瑜等,2015;赵俊虎等,2016;高辉等,2017;余荣和翟盘茂,2018;谢傲和罗伯良,2020),而 5 月起报模型的预测因子并未包含整个冬季,本文尝试将预测因子时间扩大至起报时间前 6 mon 的范围,结果表明 5 月起报的方案一和方案二模型对于湖南夏季降水预测的平均 ACC 分别为 0.12 和 0.15,PS 分别为 65.1 和 68.7,效果并不如前者,将其他起报时间的预测因子范围也扩大至前 6 mon,整体来看二者的平均 ACC 分别为 0.16 和 0.17,PS 分别为 68.8 和 69.1,相比前者也并没有显著的提升,说明机器学习模型挑选的预测因子也存在一定的局限性,通过简单增加预测因子数量的方式对于机器学习模型的预测效果并不会有显著的提升。此外,地形的动力和热力作用对降水的发生有重要影响,湖南三面环山的特殊地形是影响湖南区域性降水的因素之一,本文基于大尺度气候信号构建的模型没有考虑地形因素,对降水异常级预测能力有限,如何在机器学习模型中加入地形因素的影响还需进一步研究。

5 结论

本文采用机器学习算法筛选预测因子并结合动力模式的降水预报建立了湖南夏季降水预测模型。主要结论如下:

采用随机森林算法进行递归特征消除确定预测因子,通过交叉验证确定最优 EOF 模态个数后,使用多层前馈神经网络、支持向量回归以及自然梯度提升分别建模并对预测结果进行集合平均,比较了

两种方案的预测模型及两套动力模式对于湖南夏季降水的预测性能。评估结果表明基于机器学习的预测模型对湖南夏季雨型分布有较好的预测能力，ACC 技巧优于动力模式，但对降水异常级的预测不如 NCEP 模式，两种方案的预测模型不同起报时间的平均 ACC 分别为 0.15 和 0.19，平均 PS 评分分别为 69.3 和 69.2；并且机器学习建模方法能够较好地识别 ENSO 对湖南降水的影响。

进一步分析机器学习模型挑选的预测因子与降水关联，3—5 月起报的机器学习模型的预测技巧可能来源于前冬极地和中高纬环流，12 月—次年 2 月起报的模型预测技巧则可能来自海温的前兆信号，由于机器学习的黑箱特点，很难了解这些因子之间相互调制的物理过程，有待通过诊断分析及模式敏感性试验等方法进一步研究。

参考文献（References）

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

陈永义，俞小鼎，高学浩，等，2004.处理非线性分类和回归问题的一种新方法（Ⅰ）：支持向量机方法简介[J].应用气象学报，15(3)：345-354. Chen Y Y，Yu X D，Gao X H，et al.，2004.A new method for non-linear classify and non-linear regression Ⅰ：introduction to support vector machine[J].J Appl Meteor，15(3)：345-354.(in Chinese).

丁一汇，2011.季节气候预测的进展和前景[J].气象科技进展，1(3)：14-27. Ding Y H，2011.Progress and prospects of seasonal climate prediction[J].Adv Meteor Sci Technol，1(3)：14-27.(in Chinese).

杜良敏，柯宗建，刘长征，等，2016.基于聚类分区的中国夏季降水预测模型[J].气象，42(1)：89-96. Du L M，Ke Z J，Liu C Z，et al.，2016.Summer precipitation prediction models based on the clustering regionalization in China[J].Meteor Mon，42(1)：89-96.(in Chinese).

范可，王会军，Choi Y J，2007.一个长江中下游夏季降水的物理统计预测模型[J].科学通报，52(24)：2900-2905. Fan K，Wang H J，Choi Y J，2007.A physical statistical prediction model for summer precipitation in the middle and lower reaches of the Yangtze River[J].Chin Sci Bull，52(24)：2900-2905.(in Chinese).

封国林，赵俊虎，支蓉，等，2013.动力-统计客观定量化汛期降水预测研究新进展[J].应用气象学报，24(6)：656-665. Feng G L，Zhao J H，Zhi R，et al.，2013.Recent progress on the objective and quantifiable forecast of summer precipitation based on dynamical-statistical method[J].J Appl Meteor Sci，24(6)：656-665.(in Chinese).

冯汉中，陈永义，2004.处理非线性分类和回归问题的一种新方法（Ⅱ）：支持向量机方法在天气预报中的应用[J].应用气象学报，15(3)：355-365. Feng H Z，Chen Y Y，2004.A new method for non-linear classify and non-linear regression Ⅱ：application of support vector machine to weather forecast[J].J Appl Meteor，15(3)：355-365.(in Chinese).

高辉，袁媛，洪洁莉，等，2017.2016 年汛期气候预测效果评述及主要先兆信号与应用[J].气象，43(4)：486-494. Gao H，Yuan Y，Hong J L，et al.，2017.Overview of climate prediction of the summer 2016 and the precursory signals[J].Meteor Mon，43(4)：486-494.(in Chinese).

Ham Y G，Kim J H，Luo J J，2019.Deep learning for multi-year ENSO forecasts[J].Nature，573(7775)：568-572.doi：10.1038/s41586-019-1559-7.

韩力群，2006.人工神经网络教程[M].北京：北京邮电大学出版社：29-36. Han L Q，2006.Artificial neural network tutorial[M].Beijing：Beijing University of Posts and Telecommunications Press：29-36.(in Chinese).

贺圣平，王会军，李华，等，2021.机器学习的原理及其在气候预测中的潜在应用[J].大气科学学报，44(1)：26-38. He S P，Wang H J，Li H，et al.，2021.Machine learning and its potential application to climate prediction[J].Trans Atmos Sci，44(1)：26-38.doi：10.13878/j.cnki.dqkxxb.20201125001.(in Chinese).

贾小龙，陈丽娟，李维京，等，2010.BP-CCA 方法用于中国冬季温度和降水的可预报性研究和降尺度季节预测[J].气象学报，68(3)：398-410. Jia X L，Chen L J，Li W J，et al.，2010.Statistical downscaling based on BP-CCA：predictability and application to the winter temperature and precipitation in China[J].Acta Meteorol Sin，68(3)：398-410.(in Chinese).

柯宗建，张培群，董文杰，等，2009.最优子集回归方法在季节气候预测中的应用[J].大气科学，33(5)：994-1002. Ke Z J，Zhang P Q，Dong W J，et al.，2009.An application of optimal subset regression in seasonal climate prediction[J].Chin J Atmos Sci，33(5)：994-1002.(in Chinese).

李春晖，潘蔚娟，王婷，2018.广东省降水的多尺度时空投影预测方法[J].应用气象学报，29(2)：217-231. Li C H，Pan W J，Wang T，2018.A multi-scale spatial-temporal projection method for monthly and seasonal rainfall prediction in Guangdong[J].J Appl Meteor Sci，29(2)：217-231.(in Chinese).

李易芝，罗伯良，霍林，2017.湖南夏季旱涝转折异常特征分析[J].暴雨灾害，36(4)：339-347. Li Y Z，Luo B L，Huo L，2017.Analysis on anomalous characteristics of the summer drought-flood transitions in Hunan[J].Torrential Rain Disasters，36(4)：339-347.(in Chinese).

李瑜，李维京，任宏利，等，2015.长江中下游地区冬夏干湿韵律特征分析[J].气象学报，73(3)：496-504. Li Y，Li W J，Ren H L，et al.，2015.Analysis of dry/wet rhythms in winter and summer precipitations over the midlower reaches of the Yangtze River Basin[J].Acta Meteorol Sin，73(3)：496-504.(in Chinese).

LeCun Y，Bengio Y，Hinton G，2015.Deep learning[J].Nature，521(7553)：436-444.doi：10.1038/nature14539.

刘长征，杜良敏，柯宗建，等，2013.国家气候中心多模式解释应用集成预测[J].应用气象学报，24(6)：677-685. Liu C Z，Du L M，Ke Z J，et al.，

2013.Multi-model downscaling ensemble prediction in national climate center[J].J Appl Meteor Sci,24(6):677-685.(in Chinese).

Liu Y,Fan K,2014.An application of hybrid downscaling model to forecast summer precipitation at stations in China[J].Atmos Res,143:17-30.doi:10.1016/j.atmosres.2014.01.024.

刘颖,任宏利,张培群,等,2020.中国夏季降水的组合统计降尺度模型预测研究[J].气候与环境研究,25(2):163-171. Liu Y,Ren H L,Zhang P Q,et al.,2020.Application of the hybrid statistical downscaling model in summer precipitation prediction in China[J].Clim Environ Res,25(2):163-171.(in Chinese).

苗春生,何东坡,王坚红,等,2017.基于C4.5算法的长江中下游地区夏季降水预测模型研究及应用[J].气象科学,37(2):256-264. Miao C S,He D P,Wang J H,et al,2017.Research and application of summer rainfall prediction model in the middle and lower reaches of the Yangtze River based on C4.5 algorithm[J].J Meteor Sci,37(2):256-264.(in Chinese).

Peng T,Zhi X F,Ji Y,et al,2020.Prediction skill of extended range 2-m maximum air temperature probabilistic forecasts using machine learning post-processing methods[J].Atmosphere,11(8):823.doi:10.3390/atmos11080823.

沈皓俊,罗勇,赵宗慈,等,2020.基于LSTM网络的中国夏季降水预测研究[J].气候变化研究进展,16(3):263-275. Shen H J,Luo Y,Zhao Z C,et al.,2020.Prediction of summer precipitation in China based on LSTM network[J].Clim Change Res,16(3):263-275.(in Chinese).

舒建川,蒋兴文,黄小梅,等,2019.中国西南夏季降水预测的统计降尺度建模分析[J].高原气象,38(2):349-358. Shu J C,Jiang X W,Huang X M,et al.,2019.Statistical downscaling modeling analysis of summer precipitation in southwest China[J].Plateau Meteor,38(2):349-358.(in Chinese).

孙照渤,谭桂容,赵振国,等,2013.中国东部夏季雨型的人工神经网络集合预测[J].大气科学学报,36(1):1-6. Sun Z B,Tan G R,Zhao Z G,et al.,2013.Ensemble prediction of summer rainfall patterns over Eastern China based on artificial neural networks[J].Trans Atmos Sci,36(1):1-6.doi:10.13878/j.cnki.dqkxxb.2013.01.001.(in Chinese).

王启光,封国林,郑志海,等,2011.长江中下游汛期降水优化多因子组合客观定量化预测研究[J].大气科学,35(2):287-297. Wang Q G,Feng G L,Zheng Z H,et al.,2011.A study of the objective and quantifiable forecasting based on optimal factors combinations in precipitation in the middle and lower reaches of the Yangtze River in summer[J].Chin J Atmos Sci,35(2):287-297.(in Chinese).

王予,李惠心,王会军,等,2021.CMIP6全球气候模式对中国极端降水模拟能力的评估及其与CMIP5的比较[J].气象学报,79(3):369-386. Wang Y,Li H X,Wang H J,et al.,2021.Evaluation of CMIP6 model simulations of extreme precipitation in China and comparison with CMIP5[J].Acta Meteorol Sin,79(3):369-386.(in Chinese).

Wen N,Liu Z Y,Liu Y H,2015.Direct impact of El Niño on East Asian summer precipitation in the observation[J].Climate Dyn,44(11/12):2979-2987.doi:10.1007/s00382-015-2605-2.

吴捷,任宏利,张帅,等,2017.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,41(6):1300-1315. Wu J,Ren H L,Zhang S,et al.,2017.Evaluation and predictability analysis of seasonal prediction by BCC second-generation climate system model[J].Chin J Atmos Sci,41(6):1300-1315.(in Chinese).

谢傲,罗伯良,2020.湖南夏季降水与前期北太平洋海温异常的关系[J].气象与环境科学,43(4):49-57. Xie A,Luo B L,2020.Relations between the preceding SSTA in northern Pacific Ocean and summer precipitation in Hunan[J].Meteor Environ Sci,43(4):49-57.doi:10.16765/j.cnki.1673-7148.2020.04.007.(in Chinese).

Yim S Y,Wang B,Xing W,2014.Prediction of early summer rainfall over South China by a physical-empirical model[J].Climate Dyn,43(7/8):1883-1891.doi:10.1007/s00382-013-2014-3.

余荣,翟盘茂,2018.厄尔尼诺对长江中下游地区夏季持续性降水结构的影响及其可能机理[J].气象学报,76(3):408-419. Yu R,Zhai P M,2018.The influence of El Niño on summer persistent precipitation structure in the middle and lower reaches of the Yangtze River and its possible mechanism[J].Acta Meteorol Sin,76(3):408-419.(in Chinese).

张宇彤,矫梅燕,陈静,2013.基于模式先验信息的贝叶斯集合降水概率预报试验[J].气象,39(10):1233-1246. Zhang Y T,Jiao M Y,Chen J,2013.Bayesian ensemble probabilistic forecasting model experiment of precipitation based on model priori information[J].Meteor Mon,39(10):1233-1246.(in Chinese).

赵俊虎,杨柳,曾宇星,等,2016.夏季长江中下游和华南两类雨型的环流特征及预测信号[J].大气科学,40(6):1182-1198. Zhao J H,Yang L,Zeng Y X,et al.,2016.Analysis of atmospheric circulation and prediction signals for summer rainfall patterns in Southern China[J].Chin J Atmos Sci,40(6):1182-1198.(in Chinese).

周莉,胡瑞卿,李伟,等,2018.湖南省夏季极端降水异常时空特征及其成因分析[J].气象科学,38(6):838-848. Zhou L,Hu R Q,Li W,et al.,2018.Characteristics of summer extreme precipitation anomaly and the cause of concurrent anomaly pattern in Hunan Province[J].J Meteor Sci,38(6):838-848.(in Chinese).

宗海锋,陈烈庭,张庆云,2010.ENSO与中国夏季降水年际变化关系的不稳定性特征[J].大气科学,34(1):184-192. Zong H F,Chen L T,Zhang Q Y,2010.The instability of the interannual relationship between ENSO and the summer rainfall in China[J].Chin J Atmos Sci,34(1):184-192.(in Chinese).

Prediction of summer precipitation in Hunan based on machine learning

HUANG Chao[1,2], LI Qiaoping[3], XIE Yijun[1,2], PENG Jiadong[1,2]

[1] *Hunan Climate Center, Changsha* 410118, *China*;

[2] *Hunan Key Laboratory of Meteorological Disaster Prevention and Reduction, Changsha* 410118, *China*;

[3] *CMA Earth System Modeling and Prediction Centre (CEMC), Beijing* 100081, *China*

Against the background of global warming, summer extreme precipitation in Hunan has increased significantly. Therefore, improving the prediction accuracy of precipitation is of great practical significance for disaster prevention and mitigation in Hunan Province. Using the monthly precipitation data from meteorological stations in Hunan, the climate index sets from the National Climate Center (NCC) and the precipitation data from the hindcast experiments are performed using seasonal prediction models of NCC and NCEP (National Centers for Environmental Prediction). The recursive feature elimination (RFE) method is used to determine the key factors, and two statistical prediction schemes of summer precipitation in Hunan are established by three algorithms: multilayer feedforward neural network (FNN), support vector regression (SVR) and natural gradient boosting (NGBoost). The results show that the prediction model based on machine learning (ML) has superior ability to predict the distribution pattern of summer precipitation in Hunan. The respective average ACC skills of the two statistical schemes with lead times of 1 to 6 months are 0. 15 and 0. 19, which is a great improvement compared with the dynamic model. The respective average PS scores are 69. 3 and 69. 2, which are higher than the NCC model. The further analysis indicates that the preceding winter polar and mid-and high-latitude latitude circulation may be the main predictability sources of ML models with lead times of 1 to 3 months. Finally, the prediction skills of models with lead times of 4 to 6 months are likely derived from the precursory signal of sea surface temperature.

machine learning; summer precipitation; forecast

doi: 10. 13878/j.cnki.dqkxxb.20210903001

(责任编辑: 刘菲)