基于机器学习算法的西部方向气候模式预测订正研究

杨理智，张栌丹，王俊锋，张　帅，严渝昇
（中国人民解放军 31308 部队，四川　成都　610031）

摘　要：基于气候预测对西部方向环境保障的重要性，针对高原地区气候模式准确度不高的现实困境，采用大数据挖掘技术，充分处理气候系统非线性统计特征。首先利用随机森林，对气候模式融合网格数据进行订正；而后将订正网格进行 EOF 分解，采用信息流算法挖掘环流因子与时间序列因果关系，构建不同模态下高影响因子集；采用随机森林进行建模，预报不同模态的时间序列；最后还原预报的格点场，完成模式格点数据修订。结果表明，经机器学习算法修订后的气候模式预报准确度、预报技巧显著提高，同时，模型预报的稳定度也有较大提升。本研究基于机器学习算法进行气象大数据挖掘，提升气候模式预测效能，旨在为提升西部方向气候预测水平提供方法思路。

关键词：气候预测；大数据挖掘；信息流；随机森林
中图分类号：P468.1/.7　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2023.11.006
引用格式：杨理智，张栌丹，王俊锋，等. 基于机器学习算法的西部方向气候模式预测订正研究［J］. 网络安全与数据治理，2023，42(11)：29-34.

Prediction correction of western climate model based on machine learning algorithm

Yang Lizhi，Zhang Ludan，Wang Junfeng，Zhang Shuai，Yan Yusheng
(Unit 31308 of the People's Liberation Army，Chengdu 610031，China)

Abstract：Based on the importance of climate prediction to support the battlefield environment in the western，and aimed at the realistic dilemma of low accuracy in plateau-climate model，this paper adopts big data mining technology to fully deal with the nonlinear statistical characteristics of the climate system. Firstly，the random forest is used to correct the data of climate model fusion grid. Then，EOF is used to analyze the corrected grid，and the information flow algorithm is also used to mine the causal relationship between circulation factors and time series，in order to construct the high-impact factor subsets in different modes. Finally，it models with random forest predicts time series of different modes，then restores the predicted grid field and completes the revision of model grid data. The results suggest that the forecasting accuracy and skills of modified climate model by machine learning algorithm have been significantly improved，as well as the stability of model prediction. This research based on machine learning algorithm for big data mining improves the efficiency of prediction model. It Aims at providing methods and ideas for improving the level of climate prediction in the western.

Key words：climate prediction；big-data mining；information flow；random forest

0　引言

气候预测方法有统计学、动力学和动力统计相结合三类方法。统计学方法由于指数因子过多且各因子相互作用过程复杂，难以基于简单的人工分析把握主要统计要素，因此不确定性较高。动力学方法基于数值预报模式，受初始扰动和大气可预报性影响，气候预测技巧有限，特别是青藏高原地区海拔高且地形复杂，气候动力模式难以精准捕捉气候过程，从而表现出了明显偏差[1-2]。动力统计结合方式为现在主流方式，能弥补统计和动力方法各自的不足，明显提升预测准确度[3-5]。因此，利用统计学方法订正西部方向气候模式，以提升预报准确度是值得探索的一个方向。

近年来，大数据分析挖掘技术——机器学习正腾飞发展，也在对数据关键信息的提取、识别和预测上取得了巨大成就。充分利用大数据分析挖掘技术，优化传统统计预测方法，是提升高原地区气候预测准确度的重要途径。气候预测准确性的影响因子众多，包含不同起报时间的模式场数据以及前期环流特征等，因子数量多、

呈现显著的非线性。机器学习算法能够挖掘大数据规律，区别于传统统计方法，它从数据出发进行学习，具有很强的处理非线性问题的能力[6]，能够从地气系统大数据中发现并挖掘分析相互关联信号，提升气候预测技巧[7-8]。

机器学习已经被广泛应用于气候预测中，涌现出大量创新创造性成果[9-11]。机器学习方法常与数值模式融合，Gentine 等[11]用神经网络模拟云和对流中热量、水汽的垂直输送以及辐射与云和水蒸气的相互作用，更有效地改进数值模式的模拟性，对气候模式的发展和预测水平的提高带来深远影响。机器学习也被广泛用于订正动力模式偏差，Moghim 和 Bras[12]使用 ANN 模型对 CCSM3 的南美洲北部降水进行订正，效果显著优于线性回归模型；Wang 等[13-14]用随机森林、支持向量、贝叶斯模型等工智能模型订正偏差，从而提高动力模式预测水平。机器学习算法对提升气候预测业务水平也有极大的贡献，黄超[15]等采用随机森林挑选因子、多层前馈神经网络、支持向量回归和自然梯度算法建立模型，有效提升了湖南夏季降水的预测能力；邓居昌等[7]用多种机器学习算法构建广西月降水量预测统计订正，结合动力模式方法，极大提升了预测准确率；向波等创造性地将机器学习算法融入多省市的气候预测业务中，成功优化预测效果。

上述研究在气候预测中机器学习算法的应用领域做出了较大贡献，但由于模式表现差、测站少等原因，鲜有研究关注西部方向。因此，本文利用西部方向 240 个区域站 30 年观测数据、国内外主流气候模式数据、前期环流特征等大数据样本，基于 EOF 分解的时间系数，采用信息流算法分析挖掘数据因果特征，运用机器学习算法构建高影响因子集与时间系数的预报模型，以优化模式预报场，最后将模式数据、重构预报数据插值回 240 个区域站，分析对比模型预报效果，探索基于机器学习算法的气候模式订正方法在西部方向的适用性。

1　订正模型构建

1.1　模型构建

本文基于机器学习算法实现气候模式订正，技术方案如图 1 所示。

具体步骤如下：

（1）融合格点场。取 EC 预报场和 NCC 预报场的均值，形成融合格点场。

（2）订正格点场。将融合格点场插值到站点，并作为输入，站点实测数据作为输出，利用随机森林训练订正模型，从而订正融合格点场，形成订正预报场。

（3）EOF 分解。对订正后的格点场进行 EOF 分解，得到前 N 个模态的空间系数、时间系数，将前 N 个模态的时间系数作为模型的预测因子。

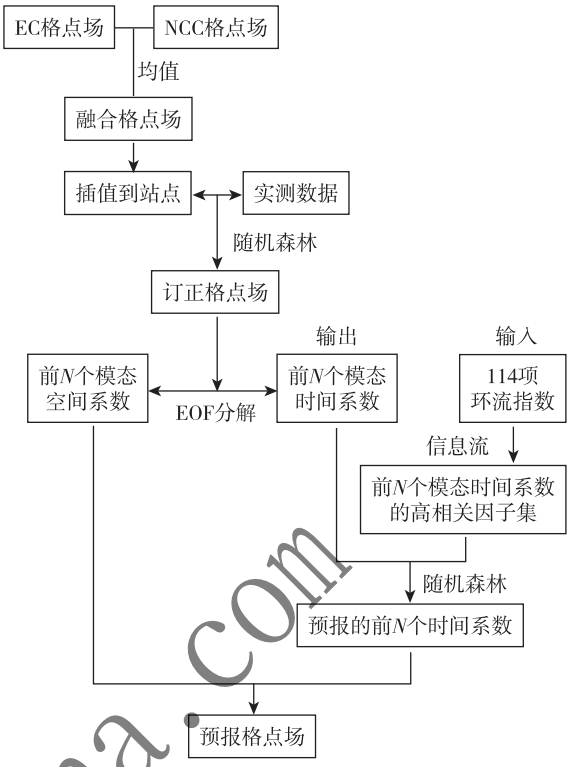


图 1　建模流程

（4）影响因子选取。采用信息流算法，寻找前 N 个模态时间系数与起报前 M 个月环流指数的因果关系，构建高影响因子集。

（5）机器学习建模。采用机器学习算法，对不同模态的高影响因子集与前 N 个模态的时间系数进行建模，得到预测的时间系数。

（6）重构格点场。利用前 N 个模态空间系数及预测的时间系数，得到预报的格点场。

（7）预测效果评估。将预报的格点数据插值到西部方向 240 个站点上，通过 RMSE 评估预报准确度；通过 PS 评分、ACC 评分分析模型预报技巧。

1.2　核心算法简介

1.2.1　随机森林

随机森林（Random Forest，RF）算法是 2001 年 Breiman[15]基于 Bagging 思想首次提出的一种分类和回归算法，它由相互独立的单棵决策树组成，使用多棵决策树样本进行训练和预测，最后利用投票机制来实现最终的分类。具体算法流程如下：

（1）矩利用自助重采样方法，从原始样本集 S 中采用有放回的采样方式，随机抽取 N 个样本子集。

（2）从 N 个样本子集中建立相应的 N 棵决策树：

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

其中 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。

（3）训练决策树模型节点时，随机选取 m（$m \leqslant M$）个预测因子作为树节点划分特征（M 为预测因子总个数），以其中最优的一个特征来划分决策树的左右子树。

（4）训练结束后，将投票得到的所有模态的平均值作为输出，得到随机森林模型预测结果。

$$h(x) = \frac{1}{N}\sum_{n=1}^{N} h(x,\theta_n) \tag{2}$$

1.2.2 信息流

区别于以往因果分析方法，近几年 Liang[16-18] 突破性地证明了因果关系实际上具有严格的物理意义和理论基础：因果关系可以被一规范方程运用最大似然估计所推得的闭合解（定义为信息流，Information Flow，IF）来度量。信息流不仅被证实在线性系统中能够快捷有效地探明因果信息交换情况，还在非线性系统的因果分析中展示了明显优于 Granger 因果测试法和转移熵的表现[19]。

针对两两时间序列 X_2 和 X_1，Liang[20] 运用最大似然估计推得从 X_2 向（注意方向性）传输的信息流可用如下公式计算：

$$T_{2\to1} = \frac{C_{11}C_{12}C_{2,d1} - C_{12}^2 C_{1,d1}}{C_{11}^2 C_{22} - C_{11}C_{12}^2} \tag{3}$$

其中 C_{ij} 指 X_i 和 X_j 之间的协方差，而 $C_{i,dj}$ 由如下算法可得：令 $\dot{X}_j$ 是 $\frac{\mathrm{d}X_j}{\mathrm{d}t}$ 的欧拉前项有限差分项（一般取 $k=1$，只有高度混沌和极端密集项才取 $k=2$），即：

$$\dot{X}_{j,n} = \frac{X_{j,n+k} - X_{j,n}}{k\Delta t} \tag{4}$$

其中 Δt 是时间步长，因此 $C_{i,dj}$ 实际上是序列 X_i 和 $\dot{X}_j$ 的协方差。理论意义上（实际应用时须结合假设检验），如果 $T_{2\to1}=0$ 则表明 X_2 不能引起 X_1（或者说 X_2 不是 X_1 的因），否则说明 X_2 有作用于 X_1（或者说 X_2 是 X_1 的因）。

1.3 预测结果评价方法

1.3.1 趋势异常综合评分 PS

PS 评分计算公式：

$$\mathrm{PS} = \frac{2\times N_0 + 2\times N_1 + 4\times N_2}{N + N_0 + 2\times N_1 + 4\times N_2 + M}\times 100 \tag{5}$$

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）。其中，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。

1.3.2 空间距平相关系数 ACC

ACC 计算公式：

$$\mathrm{Acc} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(o_i - \bar{o})^2}} \tag{6}$$

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

1.3.3 均方根误差

均方根误差（RMSE）又称标准误差，是评估预测结果好坏的常用指标，用来衡量一组数自身的离散程度，能更好地反映模型预测值与真实值之间的偏差，值越小，表明模型的预测能力越好。计算公式为：

$$X_{\mathrm{RMSE}} = \sqrt{\frac{\sum_{i=1}^{n}(X_{\mathrm{obs},i} - X_{\mathrm{model},i})^2}{n}} \tag{7}$$

其中 $X_{\mathrm{obs},i}$ 为每一个真实值，$X_{\mathrm{model},i}$ 表示对应的预测值，n 为样本量。

2 订正模型在气候预测中的应用

2.1 数据来源

本文选取西部方向 7 省（市/区）区域站历史数据、欧洲中期天气预报中心 EC 气候模式数据、国家气候中心 NCC 气候模式数据、114 项气候系统监测指数（88 项大气环流指数、26 项海温指数）。

（1）区域站数据选择 1985 年 1 月至 2021 年 9 月 240 个地面气象观测站逐月的降水资料（西部方向共 243 个区域站，其中 56 666 攀枝花、57 503 东兴区、51 747 塔中站点因缺测资料较多，不计入此次建模）；

（2）EC 气候模式历史回算时间范围为 1993 年 2 月至 2022 年 9 月，空间范围为 25°N − 50°N、70°E − 140°E，空间分辨率 1°×1°，时间分辨率 1 month；

（3）NCC 气候模式历史回算时间范围为 1991 年 1 月至 2022 年 12 月，空间范围为 25°N − 50°N、70°E − 140°E，空间分辨率 1°×1°，时间分辨率 1 month；

（4）114 项气候系统监测指数包含副高、东亚槽、欧亚环流型等 88 项大气环流指数，及厄尔尼诺、暖池等 26 项海温指数的逐月平均值，时间范围为 1951 年 1 月至 2023 年 2 月。

2.2 气候模式订正实验

2.2.1 模式数据订正

将 NCC 模式数据与 EC 模式数据进行均值融合，得到融合网格 NEC，将其插值到 240 个站点。令插值的降水数据为输入，各测站实测的降水数据为输出，利用随机森林构建订正模型。建模完成后，输入均值融合网格

数据进入订正模型，得到利用实测站点订正后的模式网格数据 R_NEC。

将 R_NEC、NCC、EC 插值到站点，与站点实测数据进行误差分析，经订正后的 R_NEC 模式网格数据能够有效减少原有模式数据的均方根误差，如图 2 所示。

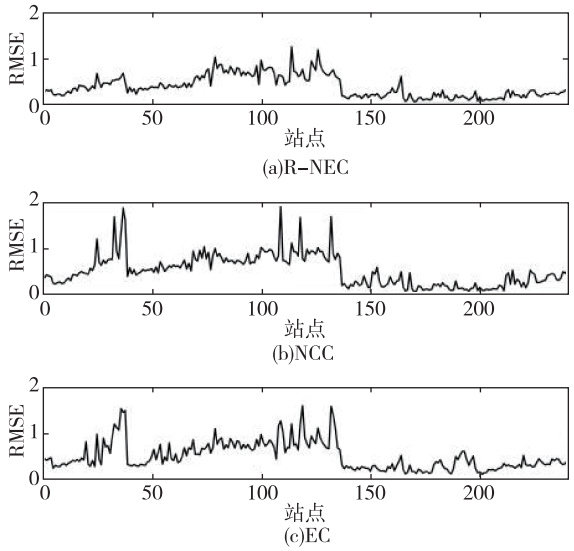


图 2　R_NEC、NCC、EC 模式均方根误差

2.2.2　因果分析

对 R_NEC 格点场降水进行 EOF 分解，得到空间系数和时间系数。根据累积方差贡献率，各模态累计方差贡献率如图 3 所示。

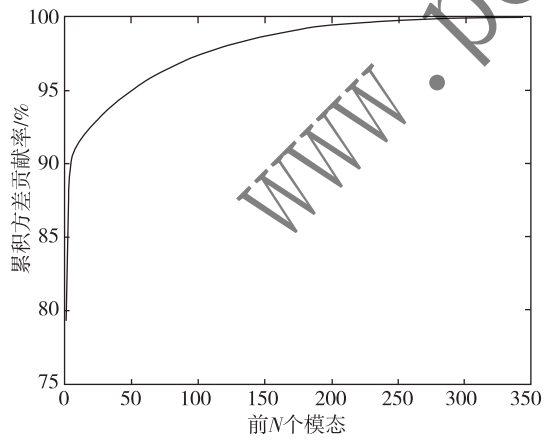


图 3　各模态累计方差贡献率

选取前 4 个模态（方差贡献率 90%）时间系数分别与起报前 1 ~ 6 mon 共 684 项环流指数进行相关性分析。采用信息流分析各模态高影响环流因子，取信息流值 $\tau_{Col\to Row}\geqslant 1\%$ 表示因子之间具有强因果关系[20]，高影响因子集数量如表 1 所示。

表 1　各模态影响因子概况

模态	1	2	3	4
影响因子数量	421	390	391	303

2.3　结果分析

基于随机森林构建模型，以不同模态高影响因子集为输入，以其时间系数为输出，得到各模态预报的时间系数；利用模型预报的时间系数和 EOF 分解出的空间系数还原成格点场。

使用 1993 年 2 月 ~ 2014 年 7 月 258 个样本进行训练，2014 年 8 月 ~ 2021 年 9 月 86 个样本进行检验。将预报数据和优化后的格点数据插值到站点，计算各站点 RMSE 以及逐月 ACC、PS 评分，随机森林模型 RF、融合网格 R_ NEC、欧洲中心气候模式 EC、国家气候中心模式 NCC 四种模式结果如图 4 所示。

随机森林模型 RF、融合网格 R_ NEC、欧洲中心气候模式 EC、国家气候中心模式 NCC 四种模式平均 RMSE、ACC 得分、PS 得分及其对应的方差如表 2 所示。

表 2　各站点预报评分

	RF	R_ NEC	EC	NCC
RMSE	0. 402 3	0. 408 3	0. 528 6	0. 547 1
RMSE 方差	0. 085 0	0. 089 0	0. 134 5	0. 124 4
ACC	0. 580 7	0. 549 7	0. 527 3	0. 532 8
ACC 方差	0. 142	0. 167 6	0. 226 8	0. 090 0
PS	86. 34	85. 94	85. 60	83. 70
PS 方差	8. 01	11. 48	24. 70	17. 59

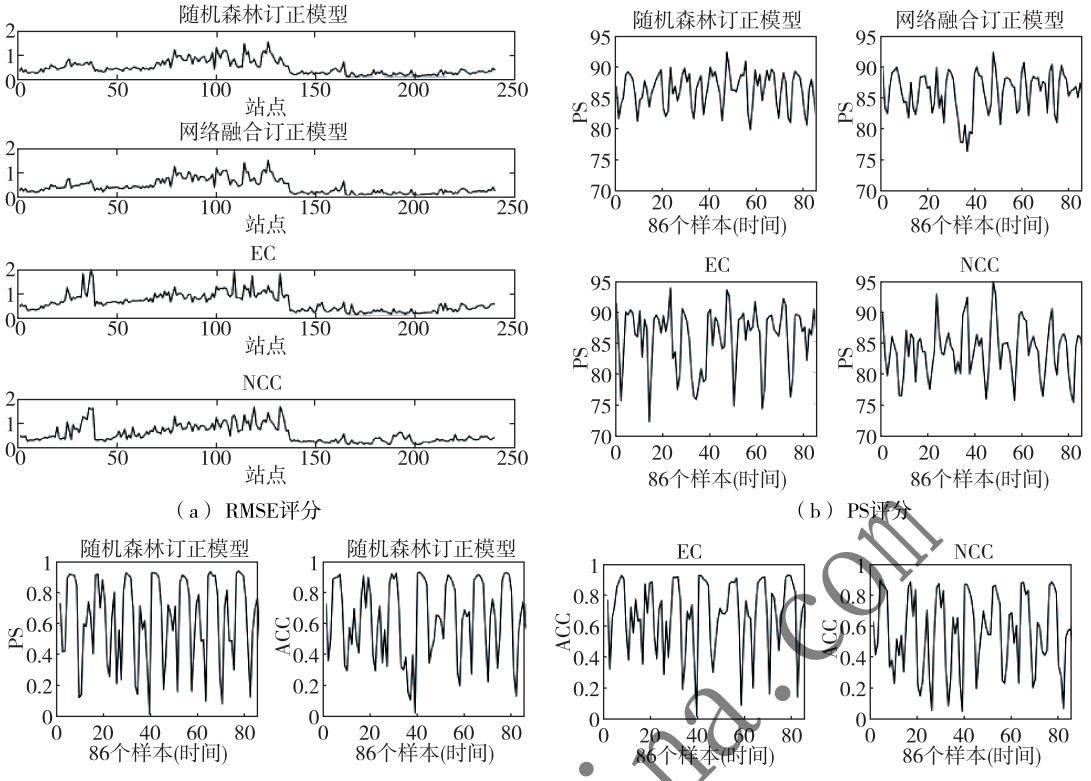
对比来看，RF 模型预报效果最优，R_ NEC 网格次之，均比 NCC、EC 模式预报效果有较大提升。

（1）模型精度方面，RF 模型预报 RMSE 均值最低、RMSE 方差最小。

（2）预报技巧方面，RF 模型 ACC 评分、PS 评分及 PS 评分方差均为最小，特别是 PS 方差较模式预报大幅减小，说明模型在异常降水预报的表现稳定度较模式大幅提升。但 RF 模型 ACC 方差大于 NCC 模式，离散度较高，说明 RF 模型在降水预报空间场表现有待提高。

3　结论

西部方向气候预测影响因子众多、非线性特征凸显，本文利用机器学习算法，分析挖掘地气系统内部规律，提高了气候模式在西部方向的预报准确度，为机器学习算法在西部气候预测中的应用提供了思路。对比国内外主流气候预测模式，本文所建立的模式订正方案能够有效降低预测误差，并具有更好的预报技巧。



（a）RMSE评分　（b）PS评分

（c）ACC评分

图4　不同预测模型预报效果评价

然而，由于模式在高原地区表现极为不佳，加之初始场扰动剧烈，模型优化后的预报准确度依然不高，下一步将探索使用不同机器学习算法修订对比预报结果；此外，将探索数学算法与地气系统物理模型相结合的研究途径，提升模式预报精度，从而为算法模型构建提供更好的初始场。

参考文献

[1] 陈炜，姜大膀，王晓欣. CMIP6模式对青藏高原气候的模拟能力评估与预估研究［J］. 高原气象，2021，40（6）：1455－1469.

[2] 徐蓉蓉. CWRF模式对青藏高原气温和降水的模拟评估［D］. 南京：南京信息工程大学，2021.

[3] 李维京，郑志海，孙丞虎. 近年来我国短期气候预测中动力相似预测方法研究与应用进展［J］. 大气科学，2013，27（2）：10.

[4] 苏海晶. 基于动力—统计的中国夏季温度和降水预测方法的研究［D］. 扬州：扬州大学，2014.

[5] 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）：17.

[6] 张泽. 机器学习算法及其工程应用研究［D］. 天津：天津大学，2023.

[7] 邓居昌，覃卫坚，韦文山. 机器学习算法在气候模式降水预测中的订正研究［J］. 计算机与数字工程，2022，50（11）：2428－2434.

[8] 贺圣平，王会军，李华，等. 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，2021，44（1）：26－38.

[9] 臧晓旭. 基于机器学习算法的降水量预测［D］. 武汉：华中科技大学，2021.

[10] 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）：8－13.

[11] GENTINE P，PRITCHARD M，RASP S，et al. Could machine learning break the convection parameterization deadlock［J］. Geophysical Research Letters，2018，45（11）.

[12] MOGHIM，SANAZ，BRAS，et al. Bias correction of climate modeled temperature and precipitation using artificial neural networks［J］. Journal of hydrometeorology，2017，18（7）：1867－1884.

[13] WANG，BIN，ZHENG，LIHONG，LIU，DE LI，et al. Using multi-model ensembles of CMIP5 global climate models to reproduce observed monthly rainfall and temperature with machine learning methods in Australia［J］. International Journal of Climatology：A Journal of the Royal Meteorological Society，2018，38（13）：4891－4902.

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

[15] 黄超，李巧萍，谢益军，等. 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，2022，45 (2)：12.

[16] LIANG X S. Information flow within stochastic dynamical systems [J]. Physical Review E, 2008, 78 (3): 031113.

[17] LIANG X S. Unraveling the cause - effect relation between time series [J]. Physical Review E, 2014, 90 (5): 052150.

[18] LIANG X S. Information flow and causality as rigorous notions ab initio [J]. Physical Review E, 2016, 94 (5): 052201.

[19] 白成祖. 信息不完备与知识不确定条件下风险评估与决策支持研究及其海上战略支点应用示范 [D]. 长沙：国防科技大学，2018.

[20] STIPS A, MACIAS D, COUGHLAN C, et al. On the causal structure between CO2 and global temperature [J]. Scientific reports, 2016 (6): 21691.

（收稿日期：2023 - 08 - 17）

作者简介：

杨理智（1990 - ），通信作者，女，博士研究生，工程师，主要研究方向：军事气象、气候预测。

张栌丹（1994 - ），女，硕士研究生，助理工程师，主要研究方向：气候分析、气候预测。

王俊锋（1993 - ），男，硕士研究生，助理工程师，主要研究方向：气候分析、气候预测。

（上接第 14 页）

[19] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv: 1412. 3555, 2014.

[20] SIAMI-NAMINI S, TAVAKOLI N, NAMIN A S. The performance of LSTM and BiLSTM in forecasting time series [C]// 2019 IEEE International Conference on Big Data. IEEE, 2019: 3285 - 3292.

[21] LIANG X, SHEN X, FENG J, et al. Semantic object parsing with graph lstm [C]//Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11 - 14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 125 - 143.

[22] KIM D, KIM E, CHA S K, et al. Revisiting binary code similarity analysis using interpretable feature engineering and lessons learned [J]. IEEE Transactions on Software Engineering, 2022, 49 (4): 1661 - 1682.

（收稿日期：2023 - 08 - 17）

作者简介：

李涛（1995 - ），男，硕士研究生，主要研究方向：系统安全。

王金双（1978 - ），男，博士，副教授，主要研究方向：系统安全。

（上接第 19 页）

[16] 焦利，孙松周，刘天须，等. 元数据驱动的分布式数据资源管理技术 [J]. 计算机与现代化，2019 (3)：7.

[17] REN D, GUI X, ZHANG K, et al. Adaptive request scheduling and service caching for MEC-assisted IoT networks: an online learning approach [J]. IEEE Internet of Things Journal, 2022, 9 (18): 17372 - 17386.

（收稿日期：2023 - 08 - 17）

作者简介：

任德旺（1989 - ），通信作者，男，博士，工程师，主要研究方向：大数据技术、数据管理与应用。

周俊鹏（1992 - ），男，硕士，助理工程师，主要研究方向：大数据分析、数据库管理。

倪鑫（1993 - ），男，博士，工程师，主要研究方向：大数据应用、数据库技术。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部