地球环境学报
JEE
Journal of Earth Environment
Vol.11 No.5 | Oct. 2020
DOI: 10.7515/JEE202011
研究论文
ARTICLE

基于 SSA-XGBoost 方法的降水变化趋势预测模型

徐 磊[1]，王甜莉[2,3]，刘松国[4]，李 东[5]，李 伟[1]，谭亮成[2,6,7]

1. 长安大学 信息工程学院，西安 710064
2. 中国科学院地球环境研究所 黄土与第四纪地质国家重点实验室，西安 710061
3. 中国科学院大学，北京 100049
4. 之江实验室，杭州 310000
5. 长安大学 图书馆，西安 710064
6. 中国科学院第四纪科学与全球变化卓越创新中心，西安 710061
7. 西安交通大学 全球环境变化研究院，西安 710049

摘 要：全球变暖预期下不同区域的未来降水变化，是政府和公众都关心的重要问题，也是全球变化研究的前沿科学问题。预测模型的建立是预测 / 预估研究的重点和难点，现今不稳定的气候背景对预测模型的精准度提出了更高的要求。为了解决传统方法对长期时间序列预测效果欠佳的问题，本文以泰国南部洞穴石笋 $\delta^{18}O$ 重建的过去 270 多年（公元 1773—2004 年）的降水记录为数据集，提出了 SSA-XGBoost 预测模型。对原始数据去趋势预处理后，采用奇异谱分析法（SSA）提取前部分数据（1773—1964 年）的振荡成分以确定数据的最佳谐波个数，并进行准确的周期信号分量分解；之后用去趋势数据减去周期信号得到随机信号，再利用 XGBoost 模型对随机项进行预测；最后将预测的序列、趋势曲线和周期信号延拓结果相叠加得到最终的预测数据（1965—2004 年）。与其他四种模型（XGBoost、ARIMA、SSA-ARIMA、LightGBM）的预测结果相比，SSA-XGBoost 的预测结果与真实值最相近，且 MAE 和 RMSE 均最小，R^2 也更接近 1，说明该模型具有更高的精度和稳定性。该研究对于泰国南部等热带地区未来的降水变化趋势预测具有较好的指导意义，也可为其他长时间序列的预估研究提供借鉴。

关键词：降水变化趋势；预测；机器学习；XGBoost 模型；奇异谱分解

Model based on SSA-XGBoost method for predicting precipitation change trends

XU Lei[1], WANG Tianli[2,3], LIU Songguo[4], LI Dong[5], LI Wei[1], TAN Liangcheng[2,6,7]

1. School of Information Engineering, Chang'an University, Xi'an 710064, China
2. State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an 710061, China
3. University of Chinese Academy of Sciences, Beijing 100049, China
4. Zhejiang Lab, Hangzhou 310000, China
5. Library of Chang'an University, Xi'an 710064, China
6. CAS Center for Excellence in Quaternary Science and Global Change, Xi'an 710061, China
7. Institute of Global Environmental Change, Xi'an Jiaotong University, Xi'an 710049, China

收稿日期：2020-03-24；**录用日期：**2020-06-30；**网络出版：**2020-07-08

Received Date: 2020-03-24; **Accepted Date:** 2020-06-30; **Online first:** 2020-07-08

基金项目：中国科学院战略性先导科技专项（B 类）（XDB40000000）；国家自然科学基金项目（41991252）；国家重点研发计划（2017YFA0603401）

Foundation Item: Strategic Priority Research Program of Chinese Academy of Sciences (XDB40000000); National Natural Science Foundation of China (41991252); National Key Research and Development Program of China (2017YFA0603401)

通信作者：谭亮成，E-mail: tanlch@ieecas.cn; 李 伟，E-mail: grandy@chd.edu.cn

Corresponding Author: TAN Liangcheng, E-mail: tanlch@ieecas.cn; LI Wei, E-mail: grandy@chd.edu.cn

引用格式：徐 磊，王甜莉，刘松国，等. 2020. 基于 SSA-XGBoost 方法的降水变化趋势预测模型 [J]. *地球环境学报*，11(5): 475–485.

Citation: Xu L, Wang T L, Liu S G, et al. 2020. Model based on SSA-XGBoost method for predicting precipitation change trends [J]. *Journal of Earth Environment*, 11(5): 475–485.

Abstract: ***Background, aim, and scope*** Predicting future climate changes is one of the most important scientific questions in the research field of global change. Precise and reliable prediction results can enable us to plan ahead and minimize the impacts of drought and flood disasters on social and economic developments. However, the establishment of prediction models is both necessary and difficult to achieve. Furthermore, owing to the current unstable climate, the requirement for the accuracy of such models has increased. Hence, this study addresses the problems that conventional methods cannot effectively predict in long-term time series, which in turn can be significant for the prediction of periodic series. ***Materials and methods*** In this study, a model based on the singular spectrum analysis (SSA)-XGBoost method was developed by incorporating the precipitation records over the past 270 years (AD 1773—2004) that were reconstructed on the basis of the stalagmite $\delta^{18}O$ content in southern Thailand. After detrending the original data, the oscillation components of the past data (AD 1773—1964) were extracted through SSA for determining the optimal harmonic number of the data and for reconstructing accurate periodic signal components. Subsequently, subtracting the periodic signal from the detrended original data produced a random signal, which was predicted by employing the XGBoost framework for obtaining a forecast sequence. The final forecast results (AD 1965—2004) were obtained by superimposing the obtained sequence and the periodic signal extension results. ***Results*** The obtained results indicate that the proposed model based on the SSA-XGBoost method performs very well while predicting the test date. The proposed model's mean absolute error (MAE) and root mean squared error (RMSE) values are found to be 0.1334 and 0.1678, respectively. Furthermore, the value of R-square (R^2) is found to be 0.536. ***Discussion*** In general, the proposed model has revealed satisfactory performance for the predictions of precipitation. Apart from the machine learning frameworks, XGBoost and LightGBM, conventional models such as ARIMA and SSA-ARIMA were employed for predicting the data to demonstrate the advantages of the proposed model. It is observed that each model provides unique features while predicting the final results. ***Conclusions*** Compared to the prediction results of the other four models (XGBoost, ARIMA, SSA-ARIMA, and LightGBM), the prediction result of the proposed model is closest to the real value. Both the MAE and RMSE of the SSA-XGBoost method are lowest in value, and the value of R^2 is closest to 1, which indicates that the proposed model based on the SSA-XGBoost method has relatively higher precision and stability. ***Recommendations and perspectives*** This study serves as an important directive for future studies on predicting precipitation changes in southern Thailand and other tropical regions, and it also provides a reference for future research on predicting long-term time series.

Key words: precipitation change trend; prediction; machine learning; XGBoost Model; Singular Spectrum Analysis (SSA)

降水变化对地表径流和农作物的生长有直接影响，也与社会经济的稳定发展和人类的生产生活息息相关。已有大量研究表明：历史时期洪水或干旱事件对文明兴衰、朝代更替有着重要影响（Zhang et al，2008；Tan et al，2018a，2019）。近年来，随着全球变暖，极端气候事件频发，严重影响了农业、人口和经济的发展。为了预防旱涝灾害，亟需加强对未来降水变化的预测和预估研究（杜雅娟等，2013；Tan et al，2015）。

预测研究的关键在于预测模型的建立，提高预测结果精准度的重要方法便是选择合适有效的预测模型。目前，在国内外研究中，大多是通过传统的数学方法对序列进行预测。例如：Otok 等人在 2012—2015 年间分别利用 ARIMA 方法和 MGA-ARIMA 方法针对月度降水量数据预测问题展开研究，证实了在处理短时降水量数据中简单方法比复杂方法预测的结果更准确（Otok et al，2012；Gong et al，2013；Zheng et al，2012）。Rizeei et al（2018）针对地表径流和洪水事件，利用 LULC 和降水量变化预测未来地表径流量，提出了 SCS-CN 模型，并利用 Taguchi 方法进行参数的更新从而改进了 ARIMA 模型，成功地预测了 2020 年的降水量变化。同时还有很多研究者使用机器学习方法进行预测。针对光伏问题，Malvoni et al（2016）通过考虑天气变化因素于 2016 年提出了基于数据降维的混合 PCA-LSSVM 预测模

DOI: 10.7515/JEE202011

型，从而验证了主成分分析和 SVM 模型融合的效果。之后至 2017 年，Mustapha 等人利用一种可对线性模型高效求解的梯度提升框架（Extreme Gradient Boosting）对生物分子的活性进行预测，结果表明 XGBoost 模型不仅性能优于随机森林和支持向量机，同时该模型可以支持各种目标函数（包括分类和回归），这为今后回归树（CART）在预测和分类方面的发展和改进都提供了很大的帮助（Mustapha and Saeed，2016；Agapitos et al，2017；Chen et al，2017）。Murari et al（2018）针对故障检测，利用一种基于支持向量机的预测方法，首次将机器学习技术应用到 JET 数据上，表现出额外的灵活性，提高了结果的可解释性。2019 年，Ustuner and Sanli（2019）针对不同的农作物预测分类问题，证明了 LightGBM 模型分类的有效性和可靠性，对今后进行物体分类预测提供了十分明确的指导作用。同年，基于机器学习的 Catboost 模型在时间序列预测和分类方面也实现了很大的进展和突破（Diao et al，2019）。同样深度学习也可以用到时间序列预测上来。Zhang 等人在 2018 年提出了一种基于长短期记忆（LSTM）的新颖方法，并将 LSTM 方法应用于交通流量预测方面，通过对多尺度的交通流量时间序列进行学习，来预测丢失的交通数据。实验结果表明该方法在交通流量预测中具有更高的准确性（Lu et al，2018；Tian et al，2018；Zhang and Zou，2018）。这些工作为基于时间序列的数据预测研究提供了很好的参考价值。

以上介绍的各种方法，大多数是采用传统的数学方法对序列进行预测分析，这对信息资料质量要求较高，历史值取值不当会导致预测值出现较大的偏差。同时预测值与近期相关，与计算期之前的数据无关，这与客观情况不相符。另外，大多数采用机器学习的方法均采用了单一的学习模型，这些方法对长期的周期时间序列进行预测时不能很好地解决非线性和非平稳性的问题，只能粗略显示数据的大致走势，局部详细信息变化表现较差。

由于目前器测记录长度的不足，古气候学家通常会结合各种古气候载体来重建过去的降水变化情况，以便于更好地预测未来的降水变化趋势。其中，洞穴石笋是最常用的古气候载体之一，其氧同位素（δ^{18}O）在一些地区可以指示降水的变化情况（Zhang et al，2008；Tan et al，2015，2018b；Cheng et al，2019）。杜雅娟等（2013）基于时域组合模型对万象洞和黄爷洞石笋 δ^{18}O 时间序列进行了模拟和未来 20 年的趋势预测。Tan et al（2015）同样通过时域组合模型，对汉中大鱼洞石笋 δ^{18}O 序列进行了模拟，并对汉江上游 1982—2042 年的降水进行了趋势预测，结果显示：汉江上游地区未来降水量可能低于过去 500 年的平均值，预测结果得到了 1982—2018 年气象水文数据的验证。最近，Tan et al（2019）通过泰国南部 Klang（可兰）洞石笋 δ^{18}O 记录，高分辨率重建了中印度 - 太平洋北部地区过去 2700 年连续的降水记录。本文选取可兰洞其中一根现代石笋的记录为研究对象，针对传统时间序列预测方法存在的问题，选择 SSA-XGBoost 模型来进行预测研究，探讨该方法在气候变化预测研究中的适用性。

1　研究数据选取及算法流程

本文研究数据来自 Tan et al（2019）采自泰国南部可兰洞的石笋 δ^{18}O 记录。该洞石笋 δ^{18}O 值的减小指示区域降雨量的增加，反之则反。1901—2011 年的观测结果显示该地的年均降雨量为 2760 mm，其中夏半年降雨量（5—10 月）约占全年降水量的 75%，冬半年降水量（11 月—次年 4 月）仅占 25%。本文仅选取了其中一根生长时段为公元 1733 — 2004 年的现代石笋数据进行预测分析。选取该数据的原因是：该洞穴位于热带辐合带（Intertropical Convergence Zone，ITCZ）的核心区，其降水变化可以较好地代表热带地区的降水变化。热带地区作为全球气候变化的关键区域，其降水量变化不仅影响着世界上 40% 的人口和全球生态系统的稳定性，而且对全球水文循环和能量平衡也起着十分重要的作用（Tan et al，2019），因此对热带地区未来的降水变化研究具有重要意义。

算法流程如图 1 所示：首先将原始数据进行去趋势预处理，然后一方面利用奇异谱分析方法（SSA）对数据进行分解，得到周期信号和随机信号，再利用机器学习模型（XGBoost）对随机项进行预测，将预测结果、周期延拓信号和趋势曲线进行叠加得到最终预测结果。另一方面，利用传统的方法 ARIMA 和 SSA-ARIMA 与机器学习方法 XGBoost 和 LightGBM 对该数据进行预测，最后将预测结果同 SSA-XGBoost 方法得到的结果进行对比分析，比较不同模型的预测效果。

DOI: 10.7515/JEE202011

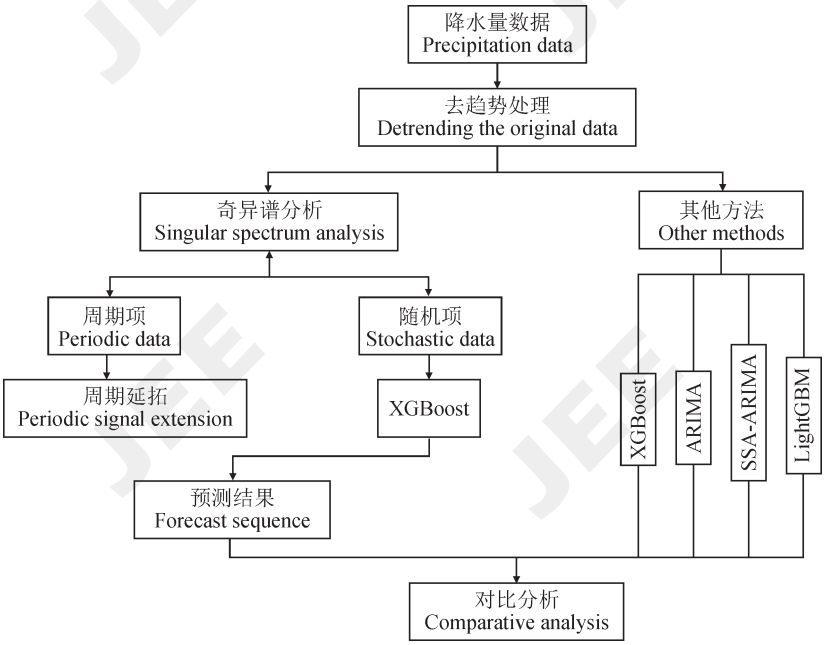

图 1 算法流程
Fig.1 Algorithm flow

2 SSA-XGBoost 预测模型原理

2.1 奇异谱分解（SSA）

奇异谱分解（Singular Spectrum Analysis，SSA）是通过分解原序列中的时间主分量，得到不同层次上的分量序列，然后将分解出来的低频视为序列变化的长期趋势，最终得到数据序列的最佳谐波个数，以此确定时间序列的周期信号。

本文将奇异谱分解分为四个步骤——嵌入、奇异值分解、分组和确定最佳谐波个数（温冬琴和王建东，2014），如下所示：

（1）嵌入。选择适当的窗口长度 m，将一维时间序列 $Y(T)=((y_1),\cdots,(y_T))$ 转换为多维时间序列：$X1,\cdots,Xn$，$(Xi=(yi,\ \cdots,yi+m-1),\ n=T-m+1)$，得到轨迹矩阵，即：

$$X=[X1,\cdots,Xn]=(x_{ij})_{i,j=1}^{n,m}=\begin{Bmatrix} y_1 & y_2 & \cdots & y_m \\ y_2 & y_3 & \cdots & y_{m+1} \\ \cdots & \cdots & \cdots & \cdots \\ y_n & y_{n+1} & \cdots & y_T \end{Bmatrix}$$

（2）奇异值分解。通过对矩阵 SVD 分解，得到 XX^{T} 的 L 个特征值。具体步骤是：将 X 转置得到 X^{T}，然后利用 $X\times X^{\mathrm{T}}$ 得到方阵，再利用方阵的性质求得矩阵的特征值，即利用 $(X^{\mathrm{T}}X)v_i=\lambda_i v_i$，求得 $\sigma_i=\sqrt{\lambda_i}$，$\mu_i=\frac{1}{\sigma_i}Xv_i$。$\sigma$ 即是奇异值，μ 是奇异向量。

（3）分组。假设有 N 个奇异值 $\sigma1,\sigma2,\cdots,\sigma N$，定义第 i 个奇异值的方差贡献率为 $\sigma_i/\sum_{k=1}^{N}\sigma k=p(i)$，由大到小选择前 M 个奇异值，使其方差贡献率之和大于一定阈值。

（4）确定最佳谐波个数。利用前 M 个奇异值的方差贡献率之和大于一定阈值（0.85）来确定最佳谐波个数 M（杜雅娟等，2013）。当 $P(i)\geqslant 0.85$ 时，认为此时的 i 即为最佳的谐波个数，即 $M=i$，然后利用三角函数的性质将数据构造成周期函数。

2.2 周期项和随机项的确定

设 $\{y_t,\ t=1,2,\cdots,N\}$ 为时间序列，若将其看作由周期项和随机项组成，可以用组合模式 $y_t=p_t+x_t$ 描述，其中 p_t 为周期项，x_t 为随机项。p_t 的表达式如下：

$$P_t=a_0/2+\sum_{k=1}^{M}(a_k\cos(2\pi kt/N)+b_k\sin(2\pi kt/N))$$

式中：M 为谐波个数；$a_k=(2/N)\sum_{t=1}^{N}y_t\cos(2\pi kt/N)$，$k=0,1,\cdots,M$）；$b_k=(2/N)\sum_{t=1}^{N}y_t\sin(2\pi kt/N)$，$k=0,1,\cdots,M$。

2.3 XGBoost 模型

XGBoost 模型是在树集成模型基础上对分类回归树（CART 树）进行组合，持续往模型中添加决策树，将多个性能较差的弱学习器集成为一个强学习器，同时支持并行化、正则化技术的一种

DOI: 10.7515/JEE202011

机器学习算法。

本文构建 XGBoost 模型主要分为三步，分别是：模型的集成、损失函数的优化、模型的求解。

（1）模型的集成。以 CART 树的回归树作为基分类器，给定训练数据，利用可加性策略，一次学习一棵树的参数，从第一棵树开始，一直到最后一棵树。通过不断地加入树来降低损失函数。

一般集成预测模型表达式为：$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i)$，其中：$K$ 为树的总个数，f_k 表示第 k 颗树，$\hat{y}_i$表示样本 x_i 的预测结果。

集成模型的损失函数为：$\mathrm{Obj}(\theta)=\sum_{i=1}^{n}l(y_i,\hat{y})+\sum_{k=1}^{K}\Omega(f_k)$，其中：$l(y_i,\hat{y}_i)$为样本 x_i 的训练误差，$\Omega(f_k)$ 表示第 k 颗树的正则项。

XGBoost 利用这种思想通过一种加法模型，将上次预测（由 t-1 棵树组合而成的模型）产生的误差作为参考进行下一棵树（第 t 棵树）的建立。以此通过不断地加入树来降低损失函数，它将模型预测值与实际值残差作为下一颗树的输入数据。设定模型的初始化值，然后往模型中添加第一棵树、第二棵树一直到添加第 t 棵树，如下所示：

$$\hat{y}_i^{(0)}=0,$$
$$\hat{y}_i^{(1)}=f_1(x_i)=\hat{y}_i^{(0)}+f_1(x_i),$$
$$\hat{y}_i^{(2)}=f_1(x_i)+f_2(x_i)=\hat{y}_i^{(1)}=f_2(x_i),$$
$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i)$$

其中：f_k 表示第 k 棵树，$\hat{y}_i^{(t)}$表示组合 t 棵树模型对样本 x_i 的预测结果。

（2）损失函数的优化。通过上述模型的集成可以构建模型的损失函数：$\mathrm{Obj}(\theta)=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t)})+\sum_{k=1}^{t}\Omega(f_k)$，损失函数采用均方误差，将目标函数变为：

$$\mathrm{Obj}(\theta)=\sum_{i=1}^{n}(y_i-(\hat{y}_i^{(t-1)}+f_t(x_i)))^2+\Omega(f_t)+C1$$
$$=\sum_{i=1}^{n}[2(y_i-\hat{y}_i^{(t-1)})f_t(x_i)+f_t(x_i)^2]+\Omega(f_t)+C1$$

本文利用独立思想，对于目标函数中的正则化部分，从每一棵回归树考虑，其模型可表示为：$f_t(x)=w_{q(x)}$, $w\in R^T$, $q:R^d\rightarrow\{1,2\cdots T\}$，其中：$w$ 为叶子节点的得分值，$q(x)$ 表示样本 x 对应的叶子节点，T 为该树的叶子节点个数。

因此，该树的复杂度可表示为：$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$。其中：$\gamma$ 为叶子个数，w_j^2表示 w 的 L2 模平方。最终，XGBoost 的目标函数可写为：

$$\mathrm{Obj}(\theta)=\sum_{i=1}^{n}l(y_i,\hat{y}^{(t-1)}+f_t(x_i))+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2+C。$$

（3）模型的求解。根据泰勒展开式将 $f_t(x_i)$ 看作 Δx 对最终的目标函数进行如下变形：

$$\mathrm{Obj}^t(\theta)\approx\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)})\partial_{y^{t-1}}l(y_i,\hat{y}^{(t-1)})f_t(x_i)+\sum_{i=1}^{n}\frac{1}{2}\partial^2_{y^{t-1}}l(y_i,\hat{y}^{(t-1)})f_t(x_i)^2+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2+C$$

令 $g_i=\partial_{\hat{y}^{(t-1)}}l(y_i,\hat{y}^{(t-1)})$, $h_i=\partial^2_{\hat{y}}l(y_i,\hat{y}^{(t-1)})$；对于第 t 棵树，$l(y_i,\hat{y}^{(t-1)})$为常数；同时去除所有常数项。最终目标损失函数可以写成：

$$\mathrm{Obj}^t(\theta)\approx\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t(x_i)^2]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$

将$\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t(x_i)^2$看作是每个样本在第 t 棵树的叶子节点的分值相关函数的结果之和，上式也可表示为：

$$\mathrm{Obj}^t(\theta)\approx\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t(x_i)^2]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$$
$$=\sum_{i=1}^{T}[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2]+\gamma T$$

式中：T 为第 t 棵树中总叶子节点的个数；$I_j=\{i|q(x_i)=j\}$ 表示在第 j 个叶子节点上的样本；w_j 为第 j 个叶子节点的得分值。

若令 $G_j=\sum_{i\in I}g_i$，$H_j=\sum_{i\in I}h_i$，则：$\mathrm{Obj}^t(\theta)=\sum_{j=1}^{T}[G_jw_j+\frac{1}{2}(H_j+\lambda)w_j^2]+\gamma T$。

此时，对 w_j 求偏导，并使其导函数等于 0，则有：$G_j+(H_j+\lambda)w_j=0$，求解得：$w_j^*=-\frac{G_j}{H_j+\lambda}$，其目标函数可表示为：$\mathrm{Obj}^*=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T$。

另外，需要注意的是，XGBoost 模型中设有多个内置参数，具有特定的含义（表 1）。当参数没有进行赋值时表示该参数使用的是默认值，当参数给予特定的数值表示在该值下模型的效果最佳，即模型的 scoring 参数得分最高。

3　SSA-XGBoost 模型预测分析

3.1　原始数据预处理

数据去趋势的方法有很多，一般由数据对应的特点来决定。本文采用最小二乘法对 1733—2004 年的原始数据进行处理，以消除数据中线性状态的基线偏移和高阶多项式的趋势项。具体步骤如下：

首先利用最小二乘法原理列出趋势多项式的求解方程，然后利用矩阵法求出趋势项系数，得出趋势项拟合曲线，最后用原始信号减去趋势项即可得出有用信号。通过该方法最终得到的原始数据的去趋势图见图 2。趋势曲线公式为 $y=0.0031x-5.5$，其中：y 为 $\delta^{18}O$，x 为年份。

表 1 XGBoost 模型参数及其含义
Tab.1 The parameters of XGBoost model and meanings

模型参数 Model parameters	含义 Meaning
learning_rate	学习率，用于更新叶子节点权重时，乘以该系数，避免步长过大。 Learning rate shrinks the contribution of each tree by learning_rate. There is a trade-off between learning_rate and n_estimators.
max_depth	树的最大深度，取值范围为：[1，∞]，树的深度越大，则对数据的拟合程度越高。 Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
min_child_weight	叶子节点中最小的样本权重和。 Minimum sum of instance weight (hessian) needed in a child.
n_estimators	最佳迭代次数。 Optimal number of iterations.
subsample	用于训练模型的子样本占整个样本集合的比例。 Subsample ratio of the training instances.
colsample_bytree	列采样，对每棵树的生成用的特征进行列采样。一般设置为：0.5—1。 Colsample_bytree is the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed.
gamma	后剪枝时，用于控制是否后剪枝的参数。 Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.
reg_alpha	控制模型复杂度的权重值的 $L1$ 正则项参数，参数越大，模型越不容易过拟合。 $L1$ regularization term on weights. Increasing this value will make model more conservative.
reg_lambda	控制模型复杂度的权重值的 $L2$ 正则项参数，参数越大，模型越不容易过拟合。 $L2$ regularization term on weights. Increasing this value will make model more conservative.
scoring	准确度评价标准，默认 None，这时需要使用 score 函数，根据所选模型不同，评价准则不同。 Model evaluation score and the default value is None. Different models have different evaluation indicators.

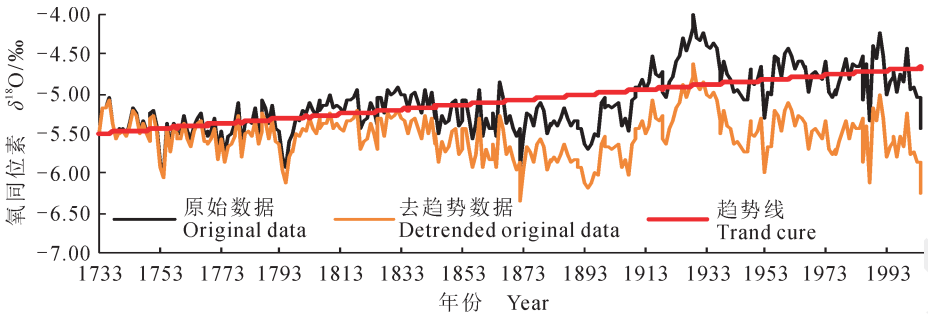

图 2 原始数据 - 去趋势数据对比图
Fig.2 The comparison of row data and detrend data

3.2 对数据进行奇异谱分解

此处将 1733—2004 年的去趋势数据分为两部分，利用 1733—1964 年的数据预测 1965—2004 年的数据并将其与 1965—2004 年的真实数据进行对比。进行窗口长度选取时一般不超出 $N/2$，对前部分数据进行奇异谱分析时分别选取的窗口长度为 1—116（$N/2$），当窗口长度为 80 时，预测精度最高，此处选取窗口长度为 80，可以得到方差贡献率（图 3）。图中黑色曲线为方差贡献率，橙色曲线为方差贡献率的一阶导数，当 $i\geq28$ 时，

DOI: 10.7515/JEE202011

$P(i)\geqslant 85\%$，满足本文的数据处理要求，即当显著谐波个数 M 为 28 时，对应的三角函数信号能表征序列的最主要趋势。

然后，可构造周期函数，表达式详见 2.2 节 p_t。为了显示最佳谐波个数的优点，分别设置 M=20 和 M=40 进行对比分析。图 4 展示了不同谐波个数对应的周期函数。从中可以看出：当谐波个数 M=20 时对应的蓝色曲线不能较好地拟合去趋势数据的局部特征，基本相当于平滑穿过；当 M=28 时，对应的红色虚线与去趋势数据基本完美拟合；当 M=40 时拟合效果同样很好。M=28 对应的红色虚线和 M=40 时的周期函数变化基本一致，但是由于红色曲线谐波个数相对较低，计算比较简单所以证明了 M=28 的合理性。

表 2 详细介绍了周期函数 P_t 中每一个参数的具体含义。根据周期函数的性质，利用周期不变性对周期函数进行延伸得到 1965—2004 年周期函数延拓图（图 5）。然后由 δ^{18}O 记录减去周期项（$x_t=y_t-p_t$），可得到图 6 所示的随机项。

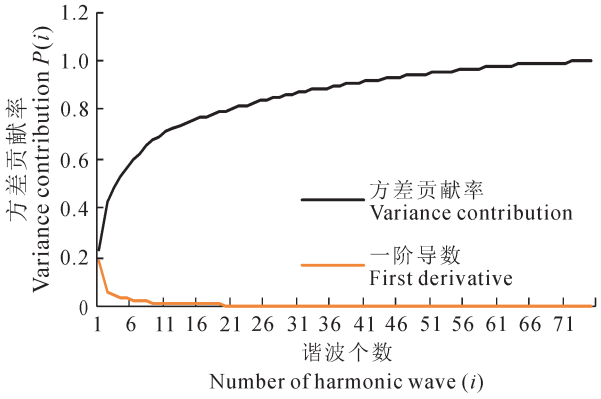

图 3　方差贡献率图

Fig.3　Graph of variance contribution rate

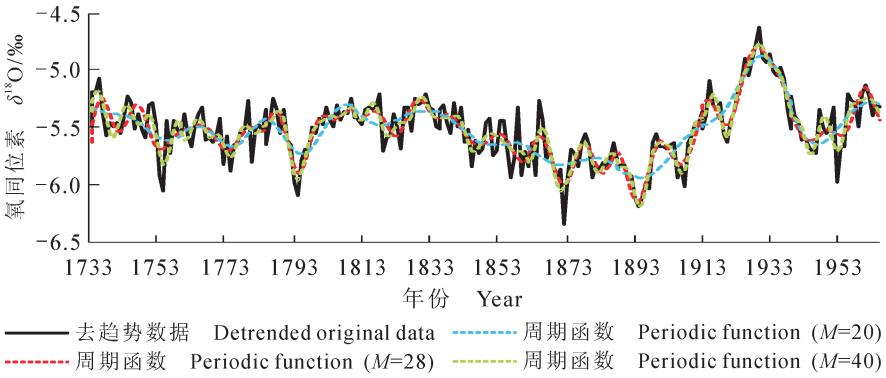

图 4　周期函数对比图

Fig.4　Comparison of periodic functions

表 2　周期函数各参数含义

Tab.2　Meaning of parameters of periodic function

参数 Parameters	含义　Meaning
P_t	t 年周期函数对应的函数值，即 t 年对应的石笋 δ^{18}O 值。 The value of periodic function in t years.
a_0	周期函数初始值，此处为 −11.0543。 Initial value of periodic function, the value is −11.0543.
a_k/b_k	不同谐波个数对应的正余弦函数的系数。 The coefficients of the sine and cosine functions corresponding to different numbers of harmonics.
y_t	t 年时的石笋 δ^{18}O 数值。 The value of δ^{18}O in t years.
M	周期函数的谐波个数，其值等于方差贡献率大于 0.85 时刻的 i 值。 The number of periodic function harmonics, whose value is equal to the value of i when the variance contribution rate is greater than 0.85.
N	降水量序列的长度，此处为 232。 Length of precipitation series, the value is 232.

DOI: 10.7515/JEE202011

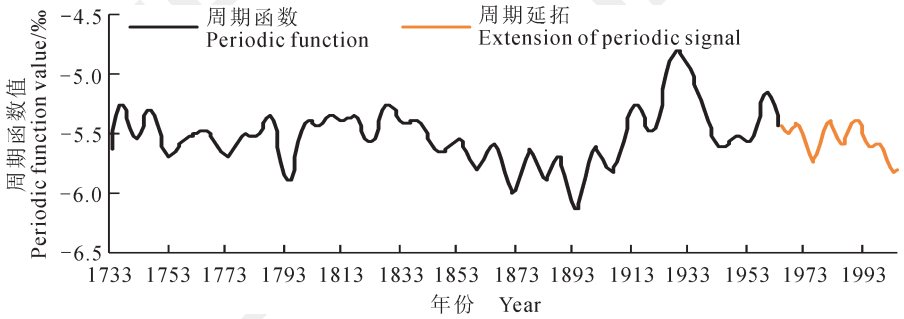

图 5 周期函数延拓图
Fig.5 Periodic function and its extension graph

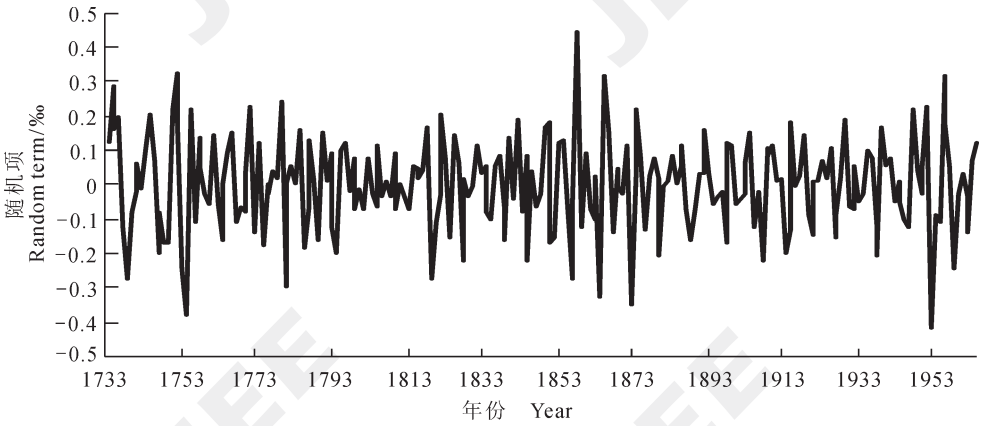

图 6 1733—1964 年氧同位素随机项
Fig.6 Random term of δ^{18}O during 1733—1967

利用机器学习方法（XGBoost 模型）对随机项进行预测。表 3 为该模型各参数的取值，图 7 为预测结果，其中黑色曲线为 1733—1964 年 δ^{18}O 记录随机项对应的数值，橙色曲线即为使用 XGBoost 模型预测的 1965—2004 年的 δ^{18}O 值（XGBoost 内置核函数为开源代码，给出的内置参数取值为调参后的最优值，未给出的为默认值）。

表 3 XGBoost 模型参数取值
Tab.3 The parameters value of XGBoost model

模型参数 Model parameters	取值 Value
learning_rate	0.2
max_depth	3
min_child_weight	1
n_estimators	1500
subsample	0.8
colsample_bytree	0.8
gamma	0.1
reg_alpha	0.05
reg_lambda	3

4 预测效果评价

将前述通过 SSA-XGBoost 模型得到的 1965—2004 年周期函数延拓值、随机项预测值和趋势曲线相叠加得到 1965—2004 年的 δ^{18}O 预测值。为了验证 SSA-XGBoost 模型的预测效果，本文还分别用传统方法 ARIMA、SSA-ARIMA 和机器学习方法 XGBoost 和 LightGBM 对数据进行了预测。将不同预测模型得到的预测结果与原始数据进行对比，得到图 8。图中黑色曲线为 1965—2004 年原始数据，红色曲线为 SSA-XGBoost 模型的预测结果，ARIMA、XGBoost 和 LightGBM 模型的预测结果分别对应黄色、绿色和蓝色曲线，SSA-ARIMA 模型对应蓝色虚线。由于 ARIMA 模型对时序的平稳要求太高，当时间序列平稳性能较差会出现图中预测效果整体趋于均值的现象。SSA-ARIMA 模型虽然采用了周期项，但是由于 ARIMA 预测效果比 XGBoost 预测效果差，导致最终的预测结果仍然不理想。由于数据量本身有限，单独使用 XGBoost 和 LightGBM 模型得到的曲线

DOI: 10.7515/JEE202011

整体比较平滑，不能完整地表示 $\delta^{18}O$ 的局部特征。只有红色曲线 SSA-XGBoost 模型预测结果能够实现较高精度地稳定预测 1965—2004 年的石笋氧同位素变化趋势。

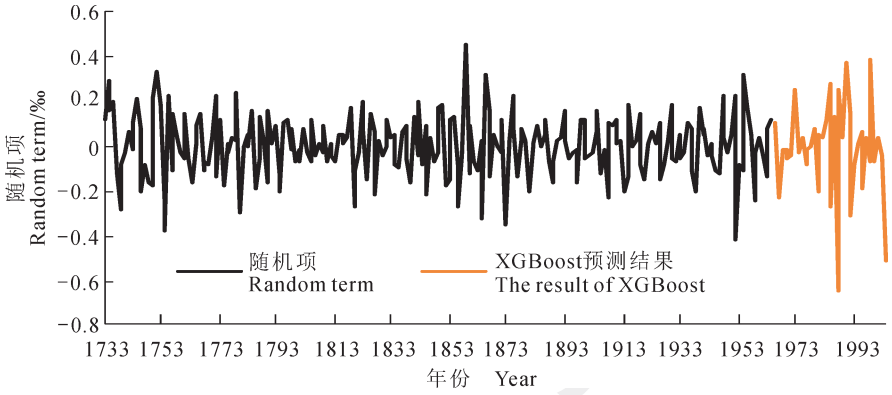

图 7　随机项预测结果
Fig.7　Results of random item prediction

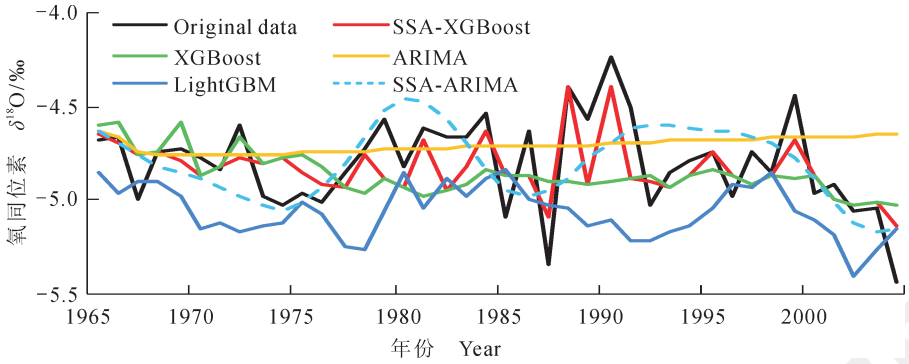

图 8　不同模型预测结果对比图
Fig.8　Comparison of predictive results by different models

另外，本文还采用平均绝对值误差（MAE）、均方根误差（RMSE）和相关系数（R^2）分析评价了不同模型预测结果的精度（表 4），各指标的公式、含义及评价标准见表 5。从表 4 可以发现：SSA-XGBoost 模型的 MAE 最小（0.1334）。就RMSE而言，SSA-XGBoost、XGBoost、ARIMA、SSA-ARIMA和 LightGBM 分别为0.1687、0.2520、0.2632、0.2224 和 0.3641，SSA-XGBoost 模型的 RMSE 依旧是最小的。另外，SSA-XGBoost 模型的 R^2 最接近 1，而其他几种模型的 R^2 均为负数或者比较小。鉴于 MAE 和 RMSE 越小，R^2 越接近 1，预测效果最好，表 4 的对比数据说明 SSA-XGBoost 的预测效果最好，其他四种模型的预测效果较差。

综上，无论是直观的曲线对比（图 8），还是数学统计方法（表 4），均证明 SSA-XGBoost 模型可以较高精度、稳定地预测 1965—2004 年的石笋 $\delta^{18}O$ 变化趋势，即中印度 - 太平洋地区的降水变化趋势。进一步尝试用 SSA-XGBoost 模型对中印度 - 太平洋地区 2005—2030 年的降水量变化趋势进行了预测（图 9）。结果显示：2005 年以后该区降雨整体回升，在 2010 年代降雨整体较多，这和器测记录一致。2020 年代的前半段降水将有所减少，后期回升。

表 4　各模型评价指数
Tab.4　Evaluation results of each model

模型 Model	指标　Norm		
	MAE	RMSE	R^2
SSA-XGBoost	0.1334	0.1687	0.5048
XGBoost	0.2007	0.2520	−0.0360
ARIMA	0.2029	0.2632	−0.2063
SSA-ARIMA	0.1768	0.2224	0.1393
LightGBM	0.3081	0.3641	−1.3079

DOI: 10.7515/JEE202011

表 5 评价指标
Tab.5 Evaluation inde

模型评价指标 The evaluate criteria of model	公式 Formula	意义 Meaning
MAE（平均绝对误差）	$\mathrm{MAE}=\frac{1}{2}\sum_{i=1}^{n}\lvert x_i-x_p\rvert$	计算值与真实值之间的平均绝对误差，MAE 越接近 0，模型越准确。The mean absolute error between true value and prediction value. The closer the value MAE is to 0, the better the prediction effect.
RMSE（均方根误差）	$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-x_p)^2}$	计算值与真实值之间的均方根误差，MSE 越接近 0，模型越准确。The root mean square error between true value and prediction value. The closer the value RMSE is to 0, the better the prediction effect.
R^2（线性相关系数）	$R^2=1-\frac{\mathrm{RSS}}{\mathrm{TSS}}=1-\frac{\sum_{i}^{n}(x_i-x_p)^2}{\sum_{i}^{n}(x_i-\bar{x}_i)^2}$	因变量与自变量之间的线性相关系数，R^2 越接近 1，模型越准确。The R-square between independent and dependent variables. The closer the value R^2 is to 1, the better the prediction effect.

x_i 表示时间序列的真实值，x_p 表示时间序列的预测值，$\bar{x}$表示时间序列的均值，n 表示时间序列的长度。
x_i: the true value of time series, x_p: the forecast of time series, $\bar{x}$: the average of the time series, n: the length of time series.

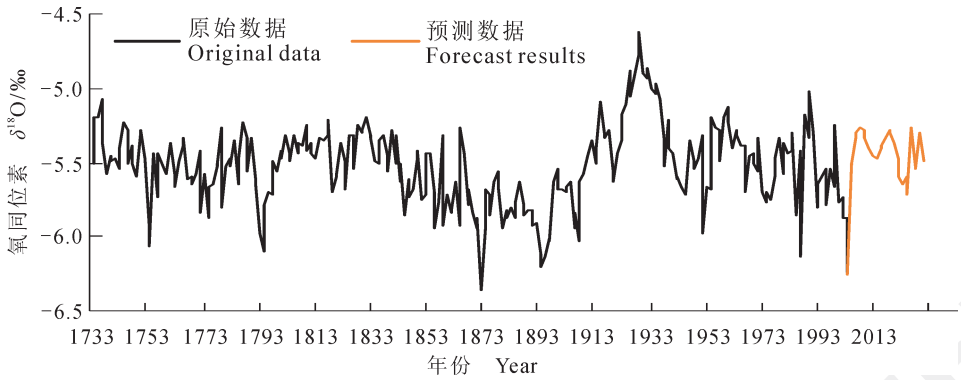

图 9 2005—2030 年降水变化趋势预测结果
Fig.9 Prediction results of precipitation trends in 2005—2030

5 结论

由于传统的预测模型和单纯的机器学习模型对长期时间序列的预测存在较大的局限性，针对此问题，本文采用数学模型（SSA）和机器学习模型结合的思想进行预测研究，并以泰国南部石笋 δ^{18}O 重建的降水变化记录为研究对象（用 1733—1964 年的数据预测 1965—2004 年的数据），提出了 SSA-XGBoost 预测模型。为了评估该模型的预测效果和精度，将其预测结果与 ARIMA、SSA-ARIMA、XGBoost 和 LightGBM 模型的预测结果进行对比，发现 SSA-XGBoost 模型的预测结果更接近真实值，且该模型预测结果的 MAE 均值、RMSE 和 R^2 波动均优于其他四个模型，表明 SSA-XGBoost 模型可以更好地预测泰国南部地区未来的降水趋势变化。这一方法对其他地区的长时间尺度气候预估也具有一定的指导意义。

参考文献

杜雅娟，冯 添，康志海 . 2013. 北方季风边缘区洞穴石笋 δ^{18}O 序列预测初步研究 [J]. *科学通报*，58(3): 254–259. [Du Y J, Feng T, Kang Z H. 2013. Preliminary study on the stalagmite δ^{18}O record prediction of the monsoon boundary zone in the northern China [J]. *Chinese Science Bulletin*, 58(3): 254–259.]

温冬琴，王建东 . 2014. 基于奇异谱分析的机场噪声时间序列预测模型 [J]. *计算机科学*，41(1): 267–270. [Wen D Q, Wang J D. 2014. Prediction model for airport-noise time series based on SSA [J]. *Computer Science*, 41(1): 267–270.]

Agapitos A, Brabazon A, O'Neill M. 2017. Regularised gradient boosting for financial time-series modelling [J]. *Computational Management Science*, 14(3): 367–391.

Chen W B, Fu K, Zuo J W, et al. 2017. Radar emitter classification for large data set based on weighted-xgboost [J]. *IET Radar,*

DOI: 10.7515/JEE202011

Sonar & Navigation, 11(8): 1203–1207.

Cheng H, Zhang H W, Zhao J Y, et al. 2019. Chinese stalagmite paleoclimate researches: a review and perspective [J]. *Science China Earth Sciences*, 62(10): 1489–1513.

Gong S H, Gao Y F, Shi H B, et al. 2013. A practical MGA-ARIMA model for forecasting real-time dynamic rain-induced attenuation [J]. *Radio Science*, 48(3): 208–225.

Diao L, Niu D, Zang Z L, et al. 2019. Short-term weather forecast based on wavelet denoising and catboost [C]// 2019 Chinese Control Conference (CCC), July 27—30, 2019, Guangzhou, China. IEEE: 3760–3764.

Lu K, Sun W X, Wang X, et al. 2018. Short-term wind power prediction model based on encoder-decoder LSTM [J]. *IOP Conference Series: Earth and Environmental Science*, 186(5): 012020. DOI: 10.1088/1755-1315/186/5/012020.

Malvoni M, de Giorgi M G, Congedo P M. 2016. Data on Support Vector Machines (SVM) model to forecast photovoltaic power [J]. *Data in Brief*, 9: 13–16.

Murari A, Lungaroni M, Peluso E, et al. 2018. Adaptive predictors based on probabilistic SVM for real time disruption mitigation on JET [J]. *Nuclear Fusion*, 58(5): 056002. DOI: 10.1088/1741-4326/aaaf9c.

Mustapha I B, Saeed F. 2016. Bioactive molecule prediction using extreme gradient boosting [J]. *Molecules*, 21(8): 983.

Otok B W, Lusia D A, Suhartono, et al. 2012. Ensemble method based on ARIMA-FFNN for climate forecasting [C]// 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE), September 10—12, 2012, Langkawi, Malaysia. IEEE: 1–4.

Rizeei H M, Pradhan B, Saharkhiz M A. 2018. Surface runoff prediction regarding LULC and climate dynamics using coupled LTM, optimized ARIMA, and GIS-based SCS-CN models in tropical region [J]. *Arabian Journal of Geosciences*, 11: 53. DOI: 10.1007/s12517-018-3397-6.

Tan L C, Cai Y J, An Z S, et al. 2015. A Chinese cave links climate change, social impacts and human adaptation over the last 500 years [J]. *Scientific Reports*, 5: 12284. DOI: 10.1038/srep12284.

Tan L C, Cai Y J, Cheng H, et al. 2018a. Centennial- to decadal-scale monsoon precipitation variations in the upper Hanjiang River region, China over the past 6650 years [J]. *Earth and Planetary Science Letters*, 482: 580–590.

Tan L C, Shen C C, Cai Y J, et al. 2018b. Great flood in the middle-lower Yellow River reaches at 4000 a BP inferred from accurately-dated stalagmite records [J]. *Science Bulletin*, 63(4): 206–208.

Tan L C, Shen C C, Löwemark L, et al. 2019. Rainfall variations in central Indo-Pacific over the past 2,700 y [J]. *Proceedings of the National Academy of Sciences*, 116(35): 17201–17206.

Tian Y, Zhang K L, Li J Y, et al. 2018. LSTM-based traffic flow prediction with missing data [J]. *Neurocomputing*, 318: 297–305.

Ustuner M, Sanli F B. 2019. Polarimetric target decompositions and light gradient boosting machine for crop classification: a comparative evaluation [J]. *ISPRS International Journal of Geo-Information*, 8(2): 97. DOI: 10.3390/ijgi8020097.

Zhang P Z, Cheng H, Edwards R L, et al. 2008. A test of climate, sun, and culture relationships from an 1810-year Chinese cave record [J]. *Science*, 322(5903): 940–942.

Zhang R T, Zou Q. 2018. Time series prediction and anomaly detection of light curve using LSTM neural network [J]. *Journal of Physics: Conference Series*, 1061: 012012. DOI: 10.1088/1742-6596/1061/1/012012.

Zheng T C, Gong S H, Shi H B. 2012. The ARIMA prediction model of rain-induced dynamic attenuation characteristics [C]// ISAPE 2012. October 22—26, 2012, Xi'an, China. IEEE: 587–590.