DOI:10.7522/j.issn.1000-0240.2017.0065
CHEN Hao, NING Chen, NAN Zhuotong, et al. Correction of the daily precipitation data over the Tibetan Plateau with machine learning models[J]. Journal of Glaciology and Geocryology, 2017, 39(3): 583-592. [陈浩, 宁忱, 南卓铜, 等. 基于机器学习模型的青藏高原日降水数据的订正研究[J]. 冰川冻土, 2017, 39(3): 583-592.]

基于机器学习模型的青藏高原日降水数据的订正研究

陈 浩[1], 宁 忱[1], 南卓铜[2*], 王玉丹[3], 吴小波[3], 赵 林[3]
(1. 宝鸡文理学院 地理与环境学院, 陕西 宝鸡 721013; 2. 南京师范大学 地理科学学院, 江苏 南京 210023;
3. 中国科学院 西北生态环境资源研究院, 甘肃 兰州 730000)

摘 要: 选择了 5 种机器学习模型, 即 k 最近邻方法(KNN)、多元自回归样条方法(MARS)、支持向量机(SVM)、多项对数线性模型(MLM)和人工神经网络(ANN), 利用海拔、相对湿度、坡向、植被、风速、气温和坡度等因子订正 ITPCAS 和 CMORPH 两种常用的青藏高原日降水数据集。五折交叉验证表明, KNN 的订正精度最高。在三个验证站点(唐古拉、西大滩和五道梁)的误差分析, 以及对青藏高原年降水量的空间分析均表明, KNN 对 CMORPH 的订正效果显著, 对 ITPCAS 在局部区域有一定订正效果, ITPCAS 及其订正值的降水空间分布准确度高于 CMORPH 的订正值。主成分分析法表明降水订正是气象和环境因子综合作用的结果。

关键词: 机器学习模型; 降水数据; 订正; 青藏高原
中图分类号: P407 **文献标志码:** A **文章编号:** 1000-0240(2017)03-0583-10

0 引言

青藏高原位于我国西部, 是世界上海拔最高的高原, 其独特的自然条件和气候特征对周边地区的气候和水文系统具有重大的影响[1-2]。在当前全球变暖的背景下, 研究可用于水文模型和气候模型的高时空分辨率降水数据集, 对于模拟青藏高原的气候变化和水文过程具有重要意义[3]。

青藏高原面积广大, 降水的时空格局十分复杂。目前常用的青藏高原地区降水产品是通过气象站点观测数据插值[4]、遥感降水资料反演与订正[5]、数据同化或者气候模式运算[6]等手段生成的。青藏高原气象站数量稀少且分布不均, 基于气象观测数据插值的数据集的精度很难满足模型模拟需求[7-8]。通过卫星或者气候模式获取的降水数据能够反映高原降水的空间分布特征, 但是降水量误差较大。TRMM (tropical rainfall measuring mission)[9] 和 CMORPH (climate prediction center morphing technique)[10-11] 等降水数据在青藏高原地区都存在明显的高值高估、低值低估的现象[12]。Shen 等[5] 的研究表明, TRMM 和 CMORPH 等 6 套降水产品, 在中国湿润地区和温暖季节的精度较高, 但在青藏高原和高海拔区域需要进一步的订正。

有研究采用基于降水发生发展规律的统计算法来对遥感产品进行订正, 如宇婧婧等[13] 通过概率密度函数匹配法(probability density function matching method, PDF 法)加最优插值的两步融合方案, 订正了 CMORPH 降水数据, 但是在青藏高原地区的订正效果较差。有研究通过融合地面站点降水观测和卫星反演的降水资料的方法, 来获取精度较高的降水数据集, 如 Chen 等[14] 融合 TRMM 和地面观测站等多种降水资料, 形成了中国区域高时空分辨率地面气象要素驱动数据集中的 ITPCAS (Institute of Tibetan Plateau Research, Chinese Academy of Sciences)降水数据。多项研究表明 ITPCAS 降水

收稿日期: 2016-12-20; **修订日期:** 2017-02-10
基金项目: 国家自然科学基金项目(41471059); 陕西省科技统筹创新计划项目(2016KTCL03-17); 陕西省教育厅重点实验室项目(16JS006)资助
作者简介: 陈浩(1978-), 男, 陕西宝鸡人, 讲师, 2015 年在中国科学院寒区旱区环境与工程研究所获博士学位, 从事降水与气象灾害研究. E-mail: chenhao@bjwlxy.cn
*通讯作者: 南卓铜, E-mail: nanzt@njnu.edu.cn.

在青藏高原的多年冻土水热过程和流域水文模拟中有较为成功的应用[15-16]。阚宝云等[17]的研究指出，ITPCAS 降水在中国西部的空间分布要优于 TMPA3B42 v6 和 CMORPH 等降水数据，是当前青藏高原地区准确度最高的降水数据集之一。然而，ITPCAS 降水数据在测站稀疏的叶尔羌河流域的降水时空分布和径流模拟中仍然存在一定的偏差。许多研究表明这些降水数据在不同时空尺度下，仍然存在一定的数据质量问题[18-19]，不同区域的降水时空分布和数值准确度不同，需要进一步的评估和订正。

有研究通过地形和植被等环境因子[20-22]，以及气温、风速、湿度、气压等气象因子[23]来订正青藏高原地区的日降水数据，如王玉丹等[24]采用 7 种气象因子和环境因子，基于 k-最近邻算法[25]（k-nearest neighbor，KNN），订正了青藏高原地区的 CMORPH 日降水数据，效果优于基于 PDF 法订正的 CMORPH 日降水数据。但是不同机器学习模型在青藏高原日降水订正研究中的适用性，机器学习模型对不同区域、不同降水数据集的订正效果，以及气象和环境因子在模型订正中的贡献率，均有待进一步的评估。本文在以上研究的基础上，采用地学领域常用的多元自适应样条（multivariate adaptive regression splines，MARS）[26]、KNN[25]、支持向量机（support vector machine，SVM）[27]、多项对数线性模型（multinomial log-linear models，MLM）[28]和人工神经网络（artificial neural networks，ANN）[29]等 5 种机器学习模型，考虑多个环境因子（海拔、坡度、坡向、植被）和气象因子（气温、湿度、风速），选取中国区域常用的日降水数据集（ITPCAS 和 CMORPH）为订正对象，评估机器学习模型在青藏高原地区日降水订正中的订正效果，并以三个验证站点（唐古拉、西大滩、五道梁）的观测数据，评估机器学习模型对不同日降水数据集的订正效果和区域差异，并讨论以上环境因子和气象因子在降水订正中的贡献率。

1 研究区域和数据

1.1 研究区域

本文的研究范围为整个青藏高原地区（26.01°~39.69° N，75.73°~104.33° E），高原地势自西北向东南倾斜，周边被巨大的山系环绕[30]。青藏高原为典型的高原大陆性气候，夏季平均气温约 10.5 ℃，冬季平均气温约 -6.4 ℃；大部分地区年降水量在 400 mm 以下，降水主要集中在夏季。降水空间异质性强，由东南向西北减少，东南部降水充沛地区的年降水可达 800~1 000 mm，西北部降水稀少地区的年降水则在 50~100 mm 之间[31]。

1.2 数据

将中国气象科学数据共享服务网（http://cdc.nmic.cn/）的 112 个标准气象站的日降水观测资料作为训练数据集，以五折交叉验证选取最优机器学习模型。中国科学院青藏高原冰冻圈观测研究站建设的唐古拉、西大滩和五道梁等站点的实测日降水数据作为对模型模拟结果的验证数据集。

唐古拉站点观测场（33°04′ N，91°56′ E，海拔 5 100 m）是在青藏公路附近建成的综合性观测场。研究区的地表主要由冰碛物、洪积物、块石和砾石构成，年平均气温为 -6~-4 ℃，降水集中在 5-8 月，多年平均降水量在 400 mm 左右。

西大滩综合观测场（35°43′ N，94°49′ E，海拔 4 538 m）是建设于青藏高原连续多年冻土北界的综合观测场。植被为以小蒿草为主的高寒草甸，覆盖度在 60%~70% 之间，年平均气温约 -4 ℃，降水多集中于 5-10 月，年降水量在 300~400 mm 之间。

五道梁观测站（35°13′ N，93°05′ E，海拔 4 735 m），是建设在青藏公路沿线五道梁南的低山丘陵区的综合观测场，下垫面为高寒荒漠草原，植被稀疏，年平均气温约 -5.1 ℃，降水集中在夏季，年降水量在 250~350 mm 之间。

唐古拉、西大滩和五道梁气象站，以及 112 个标准气象站的空间分布如图 1 所示。

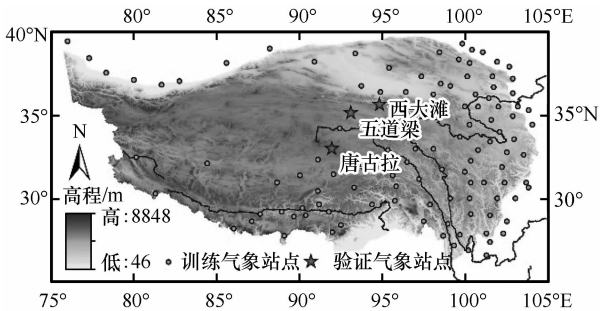

图 1 研究区位置及标准气象站分布

Fig. 1 Map showing the topography of the study area and the distribution of meteorological stations

由寒旱区科学数据中心（http://westdc.westgis.ac.cn/）下载得到中国高时空分辨率气象数据集，从中提取得到网格化的气温、湿度和风速等气象因子数据及 ITPCAS 降水数据的原始值。ITP-

CAS 数据集是以多种再分析资料为背景场，融合了中国气象局气象站点的常规气象观测数据[14]制作而成，时间分辨率为 3 h，空间分辨率为 10 km，时间范围为 2009 – 2012 年。使用 MicroMet[32]插值，改为空间分辨率为 8 km 的日尺度数据。计算 ITPCAS 降水网格数据的日累计值，得到 ITPCAS 日降水的原始值。

CMORPH 降水数据来自网络（ftp://ftp.cpc.ncep.noaa.gov/precip/CMORPH_V1.0/），提取了青藏高原 2009 – 2012 年内时间分辨率为 0.5 h、空间分辨率为 8 km 的 CMORPH 卫星反演降水数据，通过计算日累计值得到 CMORPH 日降水数据。

高程、坡度和坡向数据来自“中国 1 km 分辨率数字高程模型数据集”，由寒旱区科学数据中心下载。植被来自 MODIS NDVI 数据，时间分辨率为 16 d，空间分辨率为 250 m，并将这些环境因子重采样为 8 km 分辨率。

2　方法

2.1　机器学习模型

机器学习是一种基于统计学的分析建模手段，可以利用不同的算法，确定系统输入特征值和输出变量之间的依赖关系，并根据这种关系，对新的输入特征值做出尽可能准确的输出预测[33]。青藏高原日降水具有空间异质性强，样本量较大的特点。本文选取了在地统计学中常用的 5 种机器学习模型进行日降水订正研究，其特点如下。MARS[26]能对环境变量进行分段分析，自动优化环境因子的数量，提高交叉验证的精度。KNN[25]赋予不同距离的相邻样本以不同的权重，根据权重计算样本属性值，从而实现对样本的分类和回归，在处理空间分布不平衡的样本集时效果较优。SVM[27]结构简单，适用于小样本、非线性的模拟问题。MLM[28]通过分析因变量的期望发生比来检验自变量与因变量之间的关系，适用于变量间具有非线性关系的样本。ANN[29]可处理大样本、多变量的数据，但是易出现不能收敛的情况。

本文以 R 语言编写 5 种模型，选取对日降水影响较大的海拔、坡度、坡向以及每日的植被、气温、湿度、风速和遥感降水数据（CMORPH 或 ITPCAS）等 8 个特征值作为输入变量，计算降水订正值。通过调参试验，以及对订正值的 RMSE 分析来确定模型运行的最优参数，表 1 给出了最终得到的 5 种模型的最优运行参数。

2.2　日降水量订正与验证

模型运行在 8 km × 8 km 的格网上，因此采用 MicroMet[32]将点尺度的气象站降水数据升为网格尺度。MicroMet 是一套旨在为区域模型提供高分辨率气象要素数据的气象模型，被广泛应用于寒区的降水数据制备和尺度转换等方面的研究中[34 – 36]。本研究首先找到 112 个标准气象站和 3 个验证气象站所在的 8 km × 8 km 的网格，利用 MicroMet 将每个站点的降水数据升尺度为 1 km × 1 km 的 64 个网格数据，再求出 64 个网格降水数据的平均值作为站点所在 8 km × 8 km 网格的降水值。

模型的误差验证采用基于 R 语言的五折交叉验证法。五折交叉验证随机将 112 个训练气象站点分成均等的 5 份，依次以 1 份作为训练数据集，其余 4 份气象站点作为验证数据集，计算降水订正值与验证数据集中气象站点实测值的 RMSE，然后以 5 次计算得到的 RMSE 均值作为最终的 RMSE。

本文以模拟误差（RMSE）最小的机器学习模型为最优模型，分别计算 ITPCAS 和 CMORPH 日降水数据的降水订正值，计算日降水的多年平均累计值得到 ITPCAS 和 CMORPH 年降水数据。本文采用以下方法分析 CMORPH 订正值，ITPCAS 原始值和 ITPCAS 订正值在青藏高原地区的误差情况。① 误差的时间分布特征。计算2009 – 2012年间，各

表 1　MARS、SVM、MLM、ANN、KNN 等模型的运行参数
Table 1　Parameters of MARS, SVM, MLM, ANN and KNN models

模型名称	函数包	核函数	其他参数
MARS	earth	分段函数	每一行的权重参数、优化过程参数设为默认值
SVM	e1071	径向基函数	* type = eps-regression
MLM	nnet		* contrasts = contr. treatment
ANN	nnet		* size = 2；* maxit = 200
KNN	kknn	Optimal 最优函数	* kmax = 15；* distance = 1

注：* type 表示分类或回归的类型；* contrasts 用于部分或全部因素出现在模型中变量的公式；* size 表示隐含层的数据；* maxit 表示最大的迭代次数；* kmax 表示最大临近点数目；* distance 表示到目标点的距离。

数据集在唐古拉、西大滩和五道梁三个验证站点处的偏差和相关系数，分析误差的季节变化趋势。②误差的空间分布特征。对比 2009 – 2012 年间，各数据集在青藏高原 8 个典型区年降水量与实际多年平均年降水量的误差，分析误差的空间分布情况。

2.3 气象环境因子的敏感性分析

本文通过主成分分析(principal component analysis，PCA)法判断气象环境因子对降水订正的敏感性。海拔、坡度、坡向、植被、气温、相对湿度和风速等 7 种用于降水订正的影响因子相互之间存在相关性，难以满足传统的回归分析方法对独立变量的要求。因此，本文首先采用主成分分析法将 7 个影响因子转换成相互独立的数个主成分，计算得出主成分特征值的方差贡献率和累计方差贡献率，也就是主成分对降水订正结果的贡献率。再通过计算影响因子与主成分之间的荷载，分析 7 个影响因子对降水订正结果的贡献率。

本文用 R 语言实现主成分分析，其基本原理[37]如下所示。

假设有与日降水值相关的 i 组变量，每组变量有 j 个因子，则先按照式(1)进行标准化处理，得到标准化矩阵，

$$y_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j} \tag{1}$$

式中：y_{ij} 为标准化后的因子值；x_{ij} 为原始因子值；$\bar{x}_j$ 为第 j 个因子的算术平均值；S_j 为样本标准差。

计算各个因子之间的相关系数矩阵、特征向量和矩阵特征值，得到：

$$\beta_p = \gamma_p / \sum_{p=1}^{k} \gamma_p \tag{2}$$

$$\beta_{(p)} = \sum_{i=1}^{p} \gamma_i / \sum_{i=1}^{k} \gamma_i \tag{3}$$

式中：γ_p 为每个主成分的特征值；γ_i 为主成分特征值；β_p 为每个主成分的贡献率；$\beta_{(p)}$ 为主成分累积贡献率。

3 结果与分析

3.1 五折交叉验证结果

计算 5 种机器学习模型订正 ITPCAS 和 CMORPH 降水数据的五折交叉验证 RMSE，结果如图 2、表 2 所示。

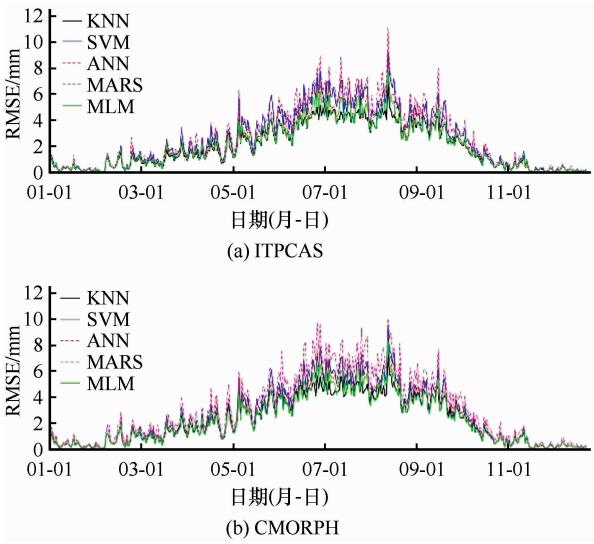

图 2 5 种机器学习模型对 ITPCAS 和 CMORPH 日降水值进行五折交叉验证的 RMSE

Fig. 2 Variations of the RMSE of the ITPCAS and the CMORPH daily precipitation corrected by the 5 machine learning models using the 5-fold cross validation

图 2 显示，在利用 5 种机器模型对 ITPCAS 和 CMORPH 两种降水产品的订正中，RMSE 随时间变化的特征基本一致。5 种机器学习模型的 RMSE 的最大值均发生在 8 月 18 日，最小值均发生在 2 月 5 日，说明 ITPCAS 和 CMORPH 两种降水产品的误差分布很接近，RMSE 极值的发生时间只受降水本身分布特征的影响，五折交叉验证中的模型选

表 2 5 种机器学习模型对 ITPCAS 和 CMORPH 日降水订正值进行五折交叉验证的 RMSE

Table 2 The RMSE of the ITPCAS and the CMORPH daily precipitation corrected by the 5 machine learning models using the 5-fold cross validation

回归模型	RMSE 最大值/mm		RMSE 最小值/mm		RMSE 均值/mm	
	ITPCAS	CMORPH	ITPCAS	CMORPH	ITPCAS	CMORPH
MLM	11.18	10.09	0.0029	0.0026	2.80	3.13
ANN	8.86	9.39	0.0032	0.0032	2.53	2.54
MARS	9.14	9.01	0.0031	0.0032	2.21	2.57
SVM	7.80	8.41	0.0031	0.0031	2.05	2.21
KNN	7.61	7.63	0.0029	0.0026	2.00	2.14

择不能改变 RMSE 极值的发生时间，只能影响 RMSE 的数值大小。表 2 显示，SVM 和 KNN 的 RMSE 的均值、最大值和最小值均较小，且比较接近，其中 KNN 的 RMSE 值最小。文献也表明，KNN 更适合于处理空间分布不平衡的大样本集[25]，因此 KNN 为青藏高原日降水订正的最优模型。5 种模型对 ITPCAS 进行五折模型交叉验证的 RMSE 均值比对 CMORPH 的小，说明 ITPCAS 降水数据在青藏高原地区的准确度较高。

3.2　不同降水数据集日降水误差的时间分布

ITPCAS 和 CMORPH 原始值和订正值的日降水误差分布如图 3、表 3、表 4 所示。

从图 3 可以看出，在唐古拉、西大滩和五道梁三个验证站点处，ITPCAS 原始值、ITPCAS 订正值和 CMORPH 订正值的偏差主要发生在降水量较多的 5 – 10 月，偏差与降水量有正相关性，但是偏差最大值与降水量最大值的发生时间并不一致，说明偏差的产生受包括降水量在内的多种因素影响。从表 3 和表 4 可以看出，经过 KNN 的订正，ITPCAS 和 CMORPH 原始值的偏差最大值和最小值均有所下降，订正值的高估和低估天数分布较为均衡，不存在明显的高估或者低估的趋势。ITPCAS 和 CMORPH 订正值与实测值的相关系数均在 0.7 以上，ITPCAS 原始值与实测值的相关系数也在 0.5 以上。以上结果表明，在唐古拉、西大滩和五道梁等站点，ITPCAS 和 CMORPH 订正值的准确度接近，略高于 ITPCAS 原始值，而 CMORPH 原始值的偏差较大，尤其是在冬季。KNN 对 CMORPH 日降水的订正效果较为显著，对 ITPCAS 日降水有一定订正效果，但不显著。

3.3　不同降水数据集年降水误差的空间分布

根据 2.2 节的方法，得出 ITPCAS 和 CMORPH 原始值、订正值的多年平均降水量在青藏高原的空间分布，如图 4 所示。

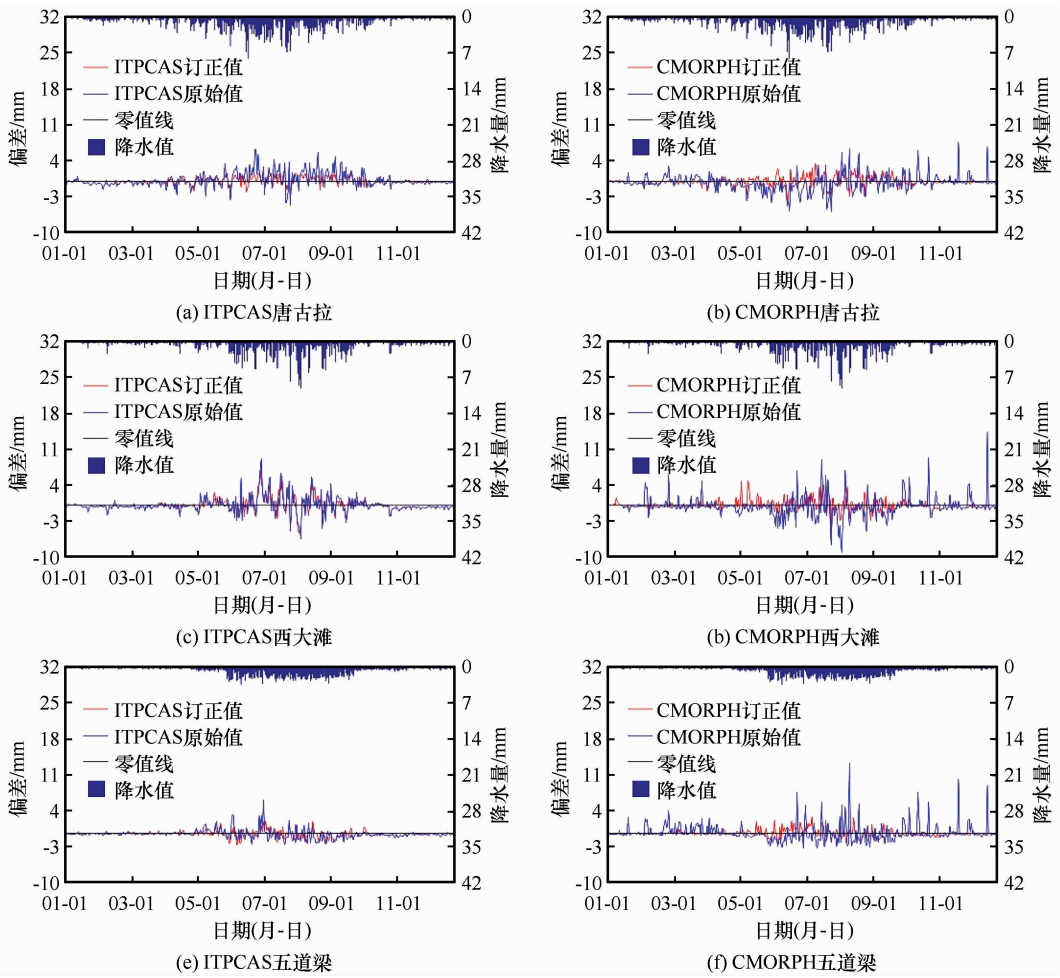

图 3　ITPCAS 和 CMORPH 日降水数据集在验证站点处与观测值对比的偏差

Fig. 3　Variations of the deviations of ITPCAS and CMORPH daily precipitation data sets relative to the observed data at the three validation sites

表 3 ITPCAS 和 CMORPH 日降水数据集在验证站点处与观测值对比的偏差

Table 3 The deviation of the ITPCAS and the CMORPH daily precipitation data sets with the observed data at the validation sites

站点	数据集	多年平均日降水数据偏差					
		高估/d	低估/d	高估最大值/mm	低估最大值/mm	高估最大值发生日期	低估最大值发生日期
唐古拉	ITPCAS 原始值	163	201	6.17	4.72	6 月 26 日	7 月 29 日
	ITPCAS 订正值	148	217	3.94	3.54	7 月 18 日	6 月 21 日
	CMORPH 原始值	101	260	7.67	6.01	11 月 25 日	7 月 29 日
	CMORPH 订正值	134	231	3.52	4.67	7 月 14 日	6 月 18 日
西大滩	ITPCAS 原始值	131	232	9.17	6.57	7 月 2 日	8 月 8 日
	ITPCAS 订正值	147	218	6.89	6.41	7 月 1 日	8 月 8 日
	CMORPH 原始值	78	266	14.40	9.21	12 月 22 日	8 月 8 日
	CMORPH 订正值	114	240	4.90	4.44	5 月 12 日	7 月 29 日
五道梁	ITPCAS 原始值	96	265	6.43	2.23	7 月 5 日	7 月 25 日
	ITPCAS 订正值	120	245	3.02	2.01	6 月 30 日	6 月 12 日
	CMORPH 原始值	107	241	13.53	2.93	8 月 16 日	7 月 4 日
	CMORPH 订正值	131	227	3.21	2.21	7 月 12 日	6 月 10 日

表 4 ITPCAS 和 CMORPH 日降水数据集在验证站点处与观测值的相关系数

Table 4 The correlation coefficients between ITPCAS and CMORPH daily precipitation data sets and observed data at the validation sites

数据集	唐古拉	西大滩	五道梁
ITPCAS 原始值	0.65	0.52	0.60
ITPCAS 订正值	0.71	0.65	0.76
CMORPH 原始值	0.31	0.19	0.18
CMORPH 订正值	0.75	0.72	0.74

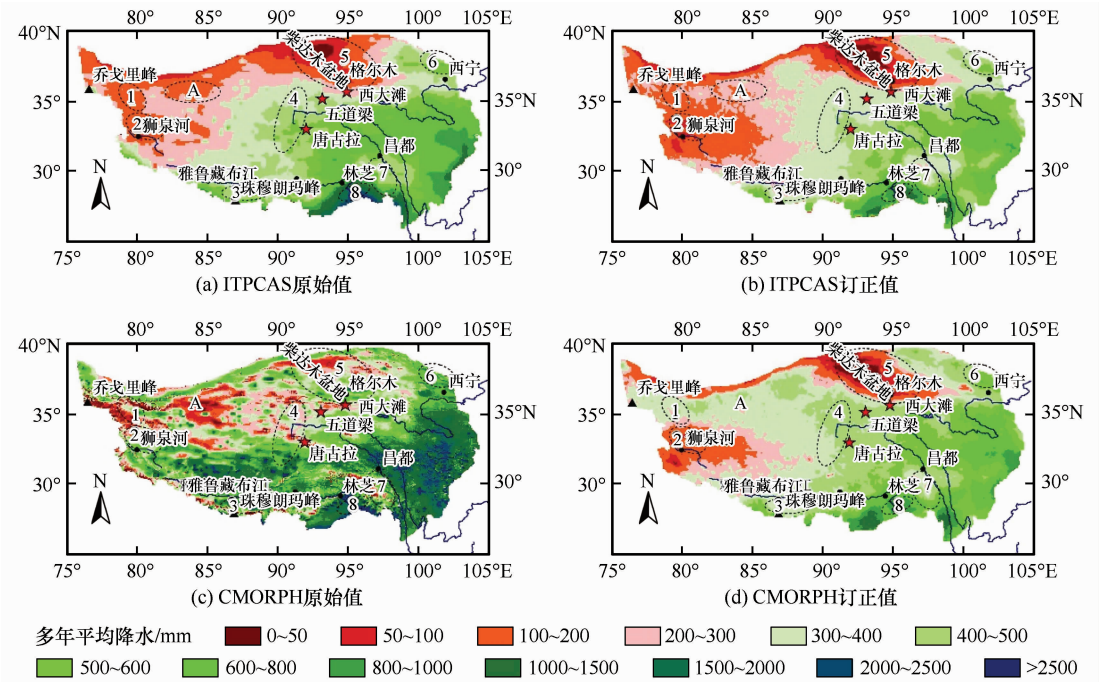

图 4 青藏高原地区 ITPCAS 年降水原始值(a)和订正值(b)以及 CMORPH 年降水原始值(c)和订正值(d)的空间分布

Fig. 4 The spatial distributions of ITPCAS original (a), corrected (b), and CMORPH original (c), corrected (d) of precipitation

研究表明[9]，青藏高原降水的空间分布存在以下 8 个典型区：1 青藏高原寒旱核心少雨区（一般为 20 ~ 40 mm，不超过 130 mm）；2 阿里喀喇昆仑山少雨区（北部 70 ~ 90 mm，大部分区域不超过 100 mm）；3 喜马拉雅北坡雨影区（350 mm 左右）；4 高原腹地的羌塘高原相对多雨区（一般为 300 ~ 400 mm，在各拉丹冬峰附近则为 400 ~ 450 mm）；5 柴达木盆地干旱区（西北部小于 20 mm，不超过 100 mm）；6 祁连山南坡相对多雨区（500 ~ 600 mm）；7 横断山区中心相对少雨区（300 ~ 400 mm，向东西两侧增多）；8 雅鲁藏布江大拐弯多雨区（1 000 ~ 1 500 mm，某些区域在 2 000 ~ 3 000 mm）。

图 4（a）、（b）显示，ITPCAS 的原始值和订正值均能较好地反映以上 8 个典型区的降水分布特征。经过 KNN 订正，ITPCAS 的年降水量在青藏高原的西部和中部均显著降低，体现为典型区 4 年降水量的下降和典型区 2 周边干旱区范围的扩大，这与实际情况较为符合，但是在北部个别区域如区域 A，降水量有所升高，说明模型的订正具有区域选择性。唐古拉站点位于典型区 4 附近，站点观测值表明 ITPCAS 订正值在该区域准确度较高。

图 4（c）、（d）显示，CMORPH 年降水量原始值的空间分布误差较大，订正值则有了显著的改善，在观测站点较多的东部和南部与实际情况比较符合，在观测站点稀疏的西部和北部的一些干旱区域存在较大误差。总体而言，CMORPH 年降水量订正值的空间分布准确度低于 ITPCAS，这是因为 CMORPH 年降水量原始值作为特征值所反映的降水空间分布特征不如 ITPCAS 准确。在青藏高原西部和北部的一些干旱区域，CMORPH 原始值的降水量偏大，而这些区域的观测站点稀疏，不足以订正原始值作为模型输入变量造成的误差。

总体而言，青藏高原地区的 ITPCAS 和 CMORPH 年降水量订正值的空间分布与实际情况较为符合，但 KNN 对于青藏高原西部和北部区域的日降水订正效果并不显著，误差较大。KNN 对 ITPCAS 年降水量的订正在局部区域有效，大部分区域不显著，这与 ITPCAS 已经融合了站点观测数据有关。青藏高原跨越各个气候区带，各个气候带的影响因子不尽相同，如果对青藏高原根据气候特征进行分区订正，可能会取得更好的效果。但是由于青藏高原西部和北部的标准气象站数量太少且分布不均，代表性较差，不能够满足对于区域降水数据进行订正的要求。因此，未来在取得青藏高原西部和北部更多的降水观测数据的情况下，可以考虑对青藏高原进行分区降水订正，以便提高模型在青藏高原西部和北部的模拟精度。

3.4　气象环境因子对降水订正结果的敏感性分析

设置主成分分析的主成分个数为 7，计算得到 7 个主成分的特征值依次为 5. 93、0. 59、0. 35、0. 07、0. 04、0. 009、0. 0001。前 4 个特征值较大，取前 4 个主成分进行分析。前 4 个主成分的特征值及贡献率如表 5 所示。荷载矩阵如表 6 所示。

从表 5 和表 6 可以看出，前 4 种主成分贡献了总方差的 99. 29%，而第 1 主成分贡献率高达 84. 83%，说明第 1 主成分具有较强的代表性。分析 7 种气象和环境因子对第 1 主成分的荷载矩阵，它们对降水订正敏感程度依次为海拔、相对湿度、坡向、植被、风速、气温、坡度。前 5 种因子的贡献率相差不大，说明降水订正是气象和环境因子综合作用的结果，8 km 分辨率下整个青藏高原的降水特征不具有单因子依赖性。

表 5　特征值及主成分贡献率

Table 5　The eigenvalues and the rate of contribution of the principle components

主成分	特征值	贡献率	累计贡献率
第 1 主成分	5.93	84.83%	84.83%
第 2 主成分	0.59	8.44%	93.27%
第 3 主成分	0.35	5.01%	98.28%
第 4 主成分	0.07	1.01%	99.29%

表 6　气象环境因子对第 1 主成分的荷载矩阵

Table 6　The correlation coefficients between the first principle component and the meteorological and environmental factors

项　目	海拔	相对湿度	坡向	植被	风速	气温	坡度
与第 1 主成分的相关系数	0.514	0.496	0.483	0.471	0.467	0.342	0.280

4　结论

本文利用 KNN、MARS、SVM、MLM、ANN 等 5 种机器学习模型，对 ITPCAS 和 CMORPH 两种降水数据集在青藏高原地区的日降水量进行了订正研究，得到了以下结论：

(1) 对 5 种机器学习模型的五折交叉验证表明, KNN 的 RMSE 较小, 更适合于做青藏高原地区日降水数据的订正研究。

(2) 对 ITPCAS 原始值、ITPCAS 订正值和 CMORPH 订正值在三个验证站点处的误差研究表明, KNN 对 CMORPH 日降水的订正效果较为显著, 对 ITPCAS 日降水的订正效果不显著。

(3) 通过在青藏高原 8 个典型区的降水空间分布对比发现, KNN 对 CMORPH 年降水量空间分布的订正效果较为显著, 对 ITPCAS 在局部区域有一定的订正效果, CMORPH 日降水订正值的空间分布准确度低于 ITPCAS 原始值和订正值。

(4) 通过主成分分析法研究了 7 种气象和环境因子对降水订正的贡献率, 贡献率从大到小依次为海拔、相对湿度、坡向、植被、风速、气温、坡度。前 5 种因子的贡献率相差不大, 说明降水订正是气象和环境因子综合作用的结果, 8 km 分辨率下整个青藏高原的降水特征不具有单因子依赖性。

致谢: 感谢中国科学院青藏高原冰冻圈观测试验研究站为本文提供观测数据。

参考文献(References):

[1] Wu Guoxiong, Wang Jun, Liu Xin, et al. Numerical modeling of the influence of Eurasian orography on the atmospheric circulation in different season[J]. Acta Meteorologica Sinica, 2005, 63(5): 603 - 612. [吴国雄, 王军, 刘新, 等. 欧亚地形对不同季节大气环流影响的数值模拟研究[J]. 气象学报, 2005, 63(5): 603 - 612.]

[2] Wang Tongmei, Wu Guoxiong, Wan Rijin. Role of the Tibetan Plateau thermal forcing in the summer climate patterns over subtropical Asia[J]. Plateau Meteorology, 2008, 27(1): 1 - 9. [王同美, 吴国雄, 万日金. 青藏高原的热力和动力作用对亚洲季风区环流的影响[J]. 高原气象, 2008, 27(1): 1 - 9.]

[3] Moulin L, Gaume E, Obled C. Uncertainties on mean areal precipitation: assessment and impact on streamflow simulations[J]. Hydrology and Earth System Sciences, 2009, 13(2): 99 - 114.

[4] He Hongyan, Guo Zhihua, Xiao Wenfa. Review on spatial interpolation techniques of rainfall[J]. Chinese Journal of Ecology, 2005, 24(10): 1187 - 1191. [何红艳, 郭志华, 肖文发. 降水空间插值技术的研究进展[J]. 生态学杂志, 2005, 24(10): 1187 - 1191.]

[5] Shen Yan, Xiong Anyuan, Wang Ying, et al. Performance of high-resolution satellite precipitation products over China [J/OL]. Journal of Geophysical Research: Atmospheres, 2010, 115(D2) [2016-01-30]. http://onlinelibrary.wiley.com/doi/10.1029/2009JD012097/full.

[6] Xie Pingping, Xiong Anyuan. A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses [J/OL]. Journal of Geophysical Research: Atmospheres, 2011, 116(D21) [2016-01-30]. http://onlinelibrary.wiley.com/doi/10.1029/2011JD016118/full.

[7] Han Zhenyu, Zhou Tianjun. Assessing the quality of APHRODITE high-resolution daily precipitation dataset over contiguous China[J]. Chinese Journal of Atmospheric Sciences, 2012, 36(2): 361 - 373. [韩振宇, 周天军. APHRODITE 高分辨率逐日降水资料在中国大陆地区的适用性[J]. 大气科学, 2012, 36(2): 361 - 373.]

[8] Liu Shiwei, Wu Jinkui, Zhang Wenchun, et al. Comparison analysis of sampling methods to estimate the regional precipitation based on Kriging interpolation method: a case study in Gansu Province[J]. Journal of Glaciology and Geocryology, 2015, 37(3): 650 - 657. [刘世伟, 吴锦奎, 张文春, 等. 基于克里金插值估算区域降水量的抽样方法对比分析: 以甘肃省为例[J]. 冰川冻土, 2015, 37(3): 650 - 657.]

[9] Qi Wenwen, Zhang Baiping, Pang Yu, et al. TRMM-data-based spatial and seasonal patterns of precipitation in the Qinghai-Tibet Plateau[J]. Scientia Geographica Sinica, 2013, 33(8): 999 - 1005. [齐文文, 张百平, 庞宇, 等. 基于TRMM 数据的青藏高原降水的空间和季节分布特征[J]. 地理科学, 2013, 33(8): 999 - 1005.]

[10] Xu Shiguang, Niu Zheng, Shen Yan, et al. A research into the characters of CMORPH remote sensing precipitation error in China[J]. Remote Sensing Technology and Application, 2014, 29(2): 189 - 194. [许时光, 牛铮, 沈艳, 等. CMORPH 卫星降水数据在中国区域的误差特征研究[J]. 遥感技术与应用, 2014, 29(2): 189 - 194.]

[11] Xu Shiguang, Niu Zheng, Shen Yan, et al. Evaluation and modification of CMORPH multi-satellite precipitation estimates in summer over Tibetan Plateau[J]. Remote Sensing Information, 2015, 30(1): 71 - 76. [许时光, 牛铮, 沈艳, 等. CMORPH 对青藏高原地区夏季降水的模拟精度研究与修正[J]. 遥感信息, 2015, 30(1): 71 - 76.]

[12] Li Qiong, Yang Meixue, Wan Guoning, et al. Analysis of the accuracy of TRMM 3B43 precipitation data in the source region of the Yellow River[J]. Journal of Glaciology and Geocryology, 2016, 38(3): 620 - 633. [李琼, 杨梅学, 万国宁, 等. TRMM 3B43 降水数据在黄河源区的适用性评价[J]. 冰川冻土, 2016, 38(3): 620 - 633.]

[13] Yu Jingjing, Shen Yan, Pan Yang, et al. Improvement of satellite-based precipitation estimates over China based on probability density function matching method[J]. Journal of Applied Meteorological Science, 2013, 24(5): 544 - 553. [宇婧婧, 沈艳, 潘旸, 等. 概率密度匹配法对中国区域卫星降水资料的改进[J]. 应用气象学报, 2013, 24(5): 544 - 553.]

[14] Chen Yingying, Yang Kun, He Jie, et al. Improving land surface temperature modeling for dry land of China[J/OL]. Journal of Geophysical Research: Atmospheres, 2011, 116(D20) [2016-01-30]. http://onlinelibrary.wiley.com/doi/10.1029/2011JD015921/full.

[15] Guo Donglin, Wang Huijun. Simulation of permafrost and seasonally frozen ground conditions on the Tibetan Plateau, 1981 - 2010[J]. Journal of Geophysical Research: Atmospheres, 2013, 118(11): 5216 - 5230.

[16] Xue Baolin, Wang Lei, Yang Kun, et al. Modeling the land surface water and energy cycles of a mesoscale watershed in the central Tibetan Plateau during summer with a distributed hydrological model[J]. Journal of Geophysical Research: Atmospheres, 2013, 118(16): 8857 - 8868.

[17] Kan Baoyun, Su Fengge, Tong Kai, et al. Analysis of the applicability of four precipitation datasets in the upper reaches of the Yarkant River, the Karakorum[J]. Journal of Glaciology and Geocryology, 2013, 35(3): 710 - 722. [阚宝云, 苏凤阁, 童

凯，等. 四套降水资料在喀喇昆仑山叶尔羌河上游流域的适用性分析[J]. 冰川冻土，2013，35(3)：710－722.]

[18] Jiang Zhihong，Lu Yao，Ding Yuguo. Analysis of the high-resolution merged precipitation products over China based on the temporal and spatial structure score indices[J]. Acta Meteorologica Sinica，2013，71(5)：891－900.［江志红，卢尧，丁裕国. 基于时空结构指标的中国融合降水资料质量评估[J]. 气象学报，2013，71(5)：891－900.]

[19] Shen Yan，Pan Yang，Yu Jingjing，et al. Quality assessment of hourly merged precipitation product over China[J]. Transactions of Atmospheric Sciences，2013，36(1)：37－46.［沈艳，潘旸，宇婧婧，等. 中国区域小时降水量融合产品的质量评估[J]. 大气科学学报，2013，36(1)：37－46.]

[20] Jia Shaofeng，Zhu Wenbin，Lu Aifeng，et al. A statistical spatial downscaling algorithm of TRMM precipitation based on NDVI and DEM in the Qaidam Basin of China[J]. Remote Sensing of Environment，2011，115(12)：3069－3079.

[21] Daly C，Neilson R P，Philips D L. A statistical topographic model for mapping climatological precipitation over mountainous terrain[J]. Journal of Applied Meteorology，1994，33(2)：140－158.

[22] Liu Yaqin，Fan Guangzhou，Zhou Dingwen，et al. Variability of NDVI in winter and spring on the Tibetan Plateau and their relationship with summer precipitation[J]. Acta Meteorologica Sinica，2007，65(6)：959－967.［刘雅勤，范广洲，周定文，等. 青藏高原冬、春植被归一化指数变化特征及其与高原夏季降水的联系[J]. 气象学报，2007，65(6)：959－967.]

[23] Beuchat X，Schaefli B，Soutter M，et al. Toward a robust method for subdaily rainfall downscaling from daily data[J]. Water Resources Research，2011，47(9)：24－42.

[24] Wang Yudan，Nan Zhuotong，Chen Hao，et al. Correction of CMORPH daily precipitation data over the Qinghai-Tibetan Plateau with k-nearest neighbor mode[J]. Remote Sensing Technology and Application，2016，31(3)：607－616.［王玉丹，南卓铜，陈浩，等. 基于k最近邻模型的青藏高原CMORPH日降水数据的订正研究[J]. 遥感技术与应用，2016，31(3)：607－616.]

[25] Hart P. The condensed nearest neighbor rule[J]. IEEE Transactions on Information Theory，1968，14(3)：515－516.

[26] Friedman J H. Multivariate adaptive regression splines[J]. The Annals of Statistics，1999，19(1)：121－141.

[27] Joachims T. Making large-scale SVM learning practical：LS-8 report 24[R]. Dortmund，Germany：Computer Science Department of University of Dortmund，1998.

[28] Christensen R. Log-linear models and logistic regression[M]. 2nd ed. New York：Springer，1997.

[29] Jain A K，Mao J，Mohiuddin K M. Artificial neural networks：a tutorial[J]. Computer，1996，29(3)：31－44.

[30] Fielding E，Isacks B，Barazangi M，et al. How flat is Tibet?[J]. Geology，1994，22(2)：163－167.

[31] Xu Shiguang，Niu Zheng，Kuang Da，et al. Estimating summer precipitation over the Tibetan Plateau with geostatistics and remote sensing[J]. Mountain Research and Development，2013，33(4)：424－436.

[32] Liston G E，Elder K. A meteorological distribution system for high-resolution terrestrial modeling (MicroMet)[J]. Journal of Hydrometeorology，2006，7(2)：217－234.

[33] Bishop C M. Pattern recognition and machine learning[M]. New York：Springer，2006.

[34] Pan Xiaoduo，Li Xin，Yang Kun，et al. Comparison of downscaled precipitation data over a mountainous watershed：a case study in the Heihe River basin[J]. Journal of Hydrometeorology，2014，15(4)：1560－1574.

[35] Mernild S H，Liston G E，Hasholt B. Snow-distribution and melt modelling for glaciers in Zackenberg River drainage basin，northeastern Greenland[J]. Hydrological Processes，2007，21(24)：3249－3263.

[36] Wu Xuejiao，Pan Xiaoduo，Shen Yongping，et al. Validation of WRF model on simulating forcing data for Kayiertesi River basin，Xinjiang[J]. Journal of Glaciology and Geocryology，2016，38(2)：332－340.［吴雪娇，潘小多，沈永平，等. WRF模式制备的气象驱动数据在新疆喀依尔特斯河流域的验证[J]. 冰川冻土，2016，38(2)：332－340.]

[37] Gu M，Eisenstat S C. A stable and efficient algorithm for the rank-one modification of the symmetrical eigenproblem[J]. SIAM Journal on Matrix Analysis and Applications，1994，15(4)：1266－1276.

Correction of the daily precipitation data over the Tibetan Plateau with machine learning models

CHEN Hao[1], NING Chen[1], NAN Zhuotong[2], WANG Yudan[3], WU Xiaobo[3], ZHAO Lin[3]

(1. School of Geography and Environment, Baoji University of Science and Art, Baoji 721013, Shaanxi, China; 2. School of Geography Science, Nanjing Normal University, Nanjing 210023, China; 3. Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China)

Abstract: In this paper, five machine learning models, namely k-nearest neighbor (KNN), multivariate adaptive regression splines (MARS), support vector machine (SVM), multinomial log-linear models (MLM) and artificial neural networks (ANN), are selected to correct two commonly used precipitation datasets, ITPCAS (Institute of Tibetan Plateau Research, Chinese Academy of Sciences) and CMORPH (climate prediction center morphing technique), over the Tibetan Plateau by establishing the relationship between daily precipitation and environmental data (elevation, slope, aspect, vegetation), as well as meteorological factors (air temperature, humidity, wind speed). The 5-fold cross validation shows that the KNN has the highest accuracy. The error analysis over the Tanggula, Xidatan and Wudaoliang Stations and the spatial analysis on annual precipitation over the plateau show that the KNN model can significantly correct the CMORPH over the plateau and the correction on the ITPCAS is significant locally. The KNN-corrected CMORPH has lower accuracy than the two ITPCAS precipitation. Principal component analysis indicates that the correction is the comprehensive effects of both environmental and meteorological factors.

Key words: machine learning model; precipitation data; correction; Tibetan Plateau

（本文编辑：武俊杰）