

地下水位机器学习模型中的特征筛选及应用效果分析

郭敏丽¹, 刘天航², 毕二平³, 胡晓斌⁴, 肖颖⁴, 刘春时⁴

(1. 北京市水科学技术研究院, 北京 100048; 2. 中国农业大学人文与发展学院, 北京 100083;
3. 中国地质大学(北京)水资源与环境学院, 北京 100083; 4. 北京市水资源调度管理事务中心, 北京 100195)

摘要:为提高地下水位机器学习模型模拟效果,采用偏相关分析方法、皮尔逊相关系数法、最大相关-最小冗余(mRMR)法和随机森林(RF)法,对构建的密怀顺区域3种地下水位机器学习模型的输入参数进行特征筛选,并对参数特征筛选前后的模型模拟效果进行了比较。结果表明:不同参数适合不同的特征筛选方法,地下水位及滞后值特征参数可由偏相关系数法获取,人工回补量及滞后值、降水量及滞后值特征参数需由mRMR法和RF法联合确定,其中mRMR法侧重于降水量及滞后值的筛选,RF法侧重于人工回补量及滞后值的筛选;特征筛选有效提高了极限学习机(ELM)模型和RF模型的模拟精度,提升了带有外部输入的非线性自回归神经网络(NARX)模型的运行速度;密怀顺区域3种地下水位机器学习模型应用经过特征筛选后的参数进行模拟,ELM模型的均方根误差、纳什效率系数和决定系数分别提升了63%、98%和45%,RF模型分别提升了49%、6%和2%,NARX模型的运行速度提升了11倍。

关键词:地下水位;机器学习模型;特征筛选;密怀顺区域

中图分类号:TV213.4 **文献标志码:**A **文章编号:**1004-6933(2025)03-0179-08

Feature selection in machine learning models of groundwater level and its application effect analysis//GUO Minli¹, LIU Tianhang², BI Erping³, HU Xiaobin⁴, XIAO Ying⁴, LIU Chunshi⁴(1. *Beijing Water Science & Technology Institute, Beijing 100048, China*; 2. *College of Humanities and Development Studies, China Agricultural University, Beijing 100083, China*; 3. *School of Water Resources and Environment, China University of Geosciences, Beijing 100083, China*; 4. *Beijing Water Resources Dispatching Center, Beijing 100195, China*)

Abstract: To improve the simulation performance of machine learning models for groundwater levels, four feature selection methods, including partial correlation analysis, Pearson correlation coefficient, maximum relevance-minimum redundancy (mRMR), and random forest (RF) methods, were employed to screen input parameters for three groundwater level machine learning models in the Mihuaishun Area. The simulation results before and after parameter feature selection were compared. The results show that different parameters require different feature selection methods. Groundwater level and its lagged values can be determined using partial correlation analysis, while artificial recharge and its lagged values, as well as precipitation and its lagged values, require a combination of mRMR and RF methods. Specifically, the mRMR method is more effective for selecting precipitation and its lagged values, whereas the RF method is better suited for screening artificial recharge and its lagged values. Feature selection significantly improved the simulation accuracy of the extreme learning machine (ELM) and RF models while enhancing the computational speed of the nonlinear autoregressive neural network with exogenous inputs (NARX) model. When applied to the three groundwater level machine learning models in the Mihuaishun Area, the parameter feature selection led to notable improvements that the ELM model showed a 63% reduction in root mean square error (RMSE), a 98% increase in the Nash-Sutcliffe efficiency coefficient (NSE), and a 45% improvement in the coefficient of determination (R^2). The RF model achieved a 49% reduction in RMSE, a 6% increase in NSE, and a 2% improvement in R^2 , while the NARX model demonstrated an 11-fold increase in computational speed.

Key words: groundwater level; machine learning model; feature selection; the Mihuaishun Area

随着监测技术现代化的飞速发展和大数据分析技术的不断深入,基于机器学习算法的模型开始广泛应用于地下水研究领域^[1-3]。在地下水位机器学习模型构建中,通过特征筛选能够剔除掉不相关或

冗余的特征,减少特征的数量,进而提高模型模拟精度和运行效率^[4-5]。目前,经常使用的特征筛选方法有3种:过滤法、包裹法和嵌入法^[6-7]。过滤法通过统计学方法筛选出对分类或回归任务有重要作用的特征,快速实现特征降维,常用的方法有Relief算法、卡方统计、相关性分析等^[8-9];包裹法将机器学习模型的性能作为特征子集的评价标准,目的是为特定的机器学习模型选择最有利于其性能的特征子集,常用方法有遗传算法、退火算法等^[10];嵌入法将特征筛选作为模型一部分,直接将特征筛选与模型训练过程相结合,通过对特征权重的调整来选择对模型性能最有帮助的特征,常用的方法包括随机森林(random forest, RF)法、人工神经网络等^[11-12]。为提高地下水位机器学习模型的模拟效率,本文应用偏相关分析、皮尔逊相关系数法、最大相关-最小冗余(minimum redundancy maximum relevance, mRMR)法、RF法等对地下水位机器学习模型的参数进行特征筛选,并构建密怀顺区域地下水位机器学习模型,对特征筛选前后的模拟效果进行对比分析,从而研究特征筛选在密怀顺区域地下水位机器学习模型构建中的重要性。

1 研究区概况与监测井布设

1.1 研究区概况

密云、怀柔、顺义是北京东北部3个相邻的行政区,合称密怀顺区域,地理位置为40°00'N~40°48'N、116°17'E~117°30'E。区域地处燕山山脉与华北平原过渡带,兼具生态屏障和城市发展功能,是连接京津冀生态屏障与城市发展的重要走廊,也是北京重要的水源保护地。区域多年平均温度12.6℃,最高温度32.7℃,最低温度-16.4℃;多年平均湿度55.3%,最大湿度99%,最小湿度7%。区域降水量年内分布不均,79%降雨集中在6—9月。区域主要河流有潮白河和怀河,河道水流方向为自北向南,怀河在潮白河牛栏山橡胶坝处汇入潮白河(图1)。区域内河道为自然河道,没有采取人工防渗措施,河道底部介质主要为砂卵砾石,受汛期河水冲刷影响,部分区域有厚度不一的黏土覆盖。2021年在河道开展的现场双环试验结果显示,河道含水层底部渗透系数为0.67~17.57 m/d^[13],渗透系数从北向南逐渐减小。

2015—2020年密怀顺区域降水均较为丰沛,年均降水量625 mm,最大降水量发生在2018年,为748 mm;最小降水量发生在2015年,为535 mm。2015年前,密怀顺区域内河流为季节性河流,除汛期有少量雨水外,其他时间均为干涸状态。2015年

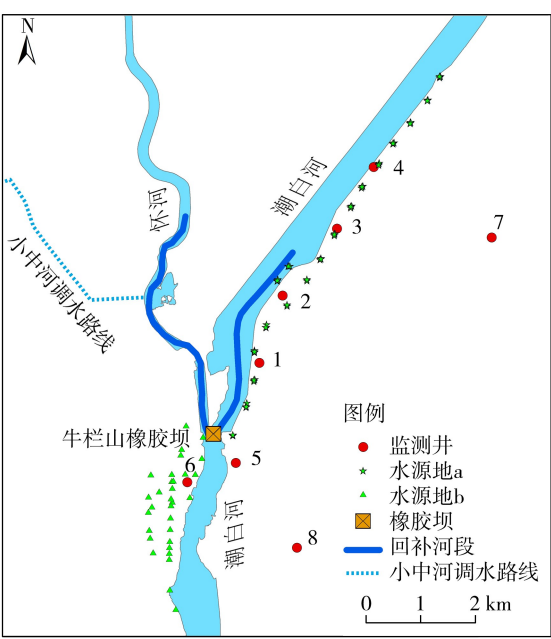


图1 密怀顺区域概况及监测井布设位置
Fig.1 Overview of the Mihuaishun Area and location of monitoring wells

后,随着南水北调中线水进京,建设了小中河调水路线,将京密引水渠的南水北调水通过怀河引入潮白河,涵养补充地下水,2015—2020年补水时间累计795 d,补水量累计3.25亿m³^[14]。补水期间潮白河上的牛栏山橡胶坝为充气状态,河水水面主要汇集于橡胶坝以上3~4 km河段。区域内分布有重要的地下水源地a和b(图1),其中水源地a位于橡胶坝以上区域,水源地b位于橡胶坝以下区域。图2为2015—2020年密怀顺区域河道补水量和地下水开采量,密怀顺区域2个水源地年均地下水开采量为0.77亿m³,累积地下水开采量为4.6亿m³^[15]。

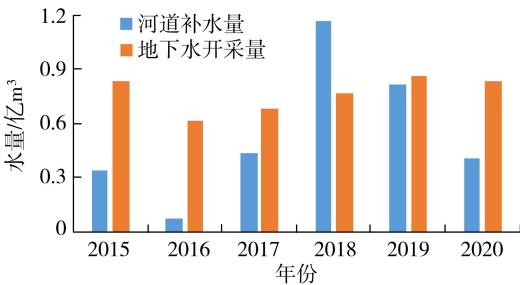


图2 2015—2020年密怀顺区域河道补水量和地下水开采量
Fig.2 River water supplement and groundwater mining output in the Mihuaishun Area from 2015 to 2020

1.2 监测井布设

研究区布设了8眼地下水监测井,位置如图1所示。1~6号监测井紧邻河道两侧,与河岸距离不超过100 m;7、8号监测井与河道岸边距离2~3 km。1、2号监测井附近河道补水期间为有水河段,3、4号监测井附近河道在补水量大时才有水面形成,5、

6号监测井位于补水河道下游,监测期间水源只有汛期雨水,河道大部分时间处于干涸状态。地下水位监测时间为2015年1月1日至2020年10月31日,监测频率为1次/d。

2 研究方法

2.1 特征筛选

2.1.1 地下水位滞后期确定

在关于时间序列的机器学习建模中,识别地下水位滞后期是一项基本任务^[16]。地下水位的有效滞后期使用自相关函数和偏自相关函数计算,偏自相关函数表示去掉其他滞后期水位的影响,当期滞后水位对当前水位的影响。自相关系数和偏自相关系数的计算公式为

$$r_A = \text{cov}(x_t, x_{t-m}) / \text{var}(x_t) \tag{1}$$

$$r_{PA} = \text{corr}(x_t, x_{t-m} | x_{t-1}, x_{t-2}, \cdots, x_{t-m+1}) \tag{2}$$

式中: r_A 为当前水位与其滞后水位之间的相关系数; x_t 为当前地下水位; x_{t-m} 为滞后 m 时的历史水位; t 为当前时间; m 为滞后时间; r_{PA} 为去掉其他滞后期地下水位的影 响,当期滞后水位与当前水位的相关系数, $r_{PA} > 0$ 表示该参数与地下水位呈正相关关系, $r_{PA} < 0$ 表示该参数与地下水位呈负相关关系。

2.1.2 特征筛选方法

a. 皮尔逊相关系数法。选用皮尔逊相关系数来评价地下水位与各参数的相关性,计算公式为

$$r_P = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{3}$$

式中: r_P 为皮尔逊相关系数; $\bar{x}$ 为地下水位平均值; y_i 为当前评价参数值; $\bar{y}$ 为评价参数平均值; n 为时段数。 $r_P > 0$ 表示该参数与地下水位呈正相关关系, $r_P < 0$ 表示该参数与地下水位呈负相关关系, $r_P > 0.2$ 或者 $r_P < -0.2$ 称为参数与地下水位具有相关性。

b. mRMR法。mRMR法通过最大化特征与目标变量的相关性以保留有效信息,同时最小化特征之间的冗余性以去除重复或相似信息,从而筛选出最具判别力且彼此互补的特征子集^[17],计算公式为

$$P_m = \max_S \left[\frac{1}{|S|} \sum_{f_i \in S} I(f_i; c) - \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i; f_j) \right] \tag{4}$$

式中: P_m 为mRMR法指数值; f_i 为参数 i 的特征; f_j 为参数 j 的特征; c 为拟合值; S 为参数特征集; $I(x; y)$ 为变量 x 和 y 的互信息, I 值越大说明两变量包含的信息越多,则相关性越强,越有利于拟合。

c. RF法。应用MATLAB 2016a中的RF工具箱进行特征重要性分析,该方法是在节点随机选择

一个包含 k 个特征的特征子集,再从这个特征子集中挑选一个新的评分高的特征进行划分。模型中的参数 k 决定特征子集大小^[18]。本文RF法的决策树棵数设定为500棵,每次随机选择的特征数量设定为14,通过importance函数输出各特征的重要性参数,并对参数进行归一化处理,将各参数在RF模拟中累积贡献以占比的方式计算。

2.2 模型构建与数据处理

2.2.1 地下水位机器学习模型构建

a. 极限学习机(extreme learning machine, ELM)模型。ELM模型是由Huang等^[19]于2005年提出的一种特殊前向反馈神经网络算法,包括输入层、隐含层、输出层3层结构。作为输入层和隐含层之间的连接的输入权重和隐含层中的偏置值是随机生成的,生成后不再调整;通过隐含层输出矩阵的简单广义逆运算解析计算隐含层和输出层之间的输出权重矩阵。该模型具有学习时间短、泛化误差小、操作简单等优点。本文ELM模型输入数据为地下水位、降水量、人工回补量、地下水开采量、温度、湿度等。根据文献[20]设定模型参数的范围,隐含层节点个数在20~80选择,激活函数为sig、sin和hardlim。应用MATLAB 2016a软件,采用试错法^[20]对模型参数进行率定,最终确定各监测井的率定参数组合。

b. 带有外部输入的非线性自回归神经网络(nonlinear auto-regressive model with exogenous inputs, NARX)模型。NARX模型是一种特殊类型的循环神经网络,其在输出层和输入层之间使用全局反馈连接^[21]。NARX模型为神经网络与带有外部输入的自回归模型的组合,神经网络用于帮助带有外部输入的自回归模型捕捉非线性关系。与其他递归网络相比,NARX模型收敛速度更快,泛化能力更好^[22-23]。除此之外,NARX模型可用监测值直接作为输入数据对参数进行训练,使得训练更有效。本文NARX模型输入数据为地下水位、降水量、人工回补量、地下水开采量、温度、湿度等,采用MATLAB R2016a的神经网络工具箱中默认的参数组合(滞后期为2 d、隐含层神经元数量为10)进行建模,同时设置最大迭代次数为1 000次^[24]。

c. RF模型。本文RF模型输入参数同样为地下水位、降水量、人工回补量、地下水开采量、温度、湿度等,需要率定的参数为决策树数量和 k 值。应用MATLAB 2016a中的RF工具箱建立模型,经过反复试验,确定决策树数量为500棵, k 值为5。

2.2.2 数据预处理

地下水位、降水量、人工回补量、地下水开采量、温度、湿度等指标均选择统一时间范围内的日数据。

降水量、温度和湿度选取当地水文站监测数据,人工回补量选用实际补水期间统计的日补水量,地下水开采量只有月数据,将数据平均处理后作为日开采量。为了减少机器学习算法受输入变量数值变动范围的影响,在应用机器学习模型模拟前,首先使用最小-最大归一化方法对输入变量进行归一化处理,使其值范围为0~1。

2.2.3 评价指标选择

为定量评价参数特征筛选前后机器学习模型的模拟精度,采用均方根误差(RMSE)、决定系数(R^2)和纳什效率系数(NSE)作为评价指标。 R^2 和NSE越接近1,RMSE越接近0时,预测值与实测值越接近,模型精度越高,性能越好;当 R^2 和NSE为1,RMSE为0时,可认为该模型为理想模型。

3 结果与分析

3.1 特征筛选结果

根据 Rajae 等^[25]的研究,利用机器学习算法建立的地下水位模型输入参数主要为地下水位、降水量、温度、河流流量、蒸发量、地表水位和湿度。研究区地下水埋深在补水前为37~54 m,补水后为15~28 m,均大于4 m,因此模拟不考虑地下水蒸发量。河道补水期间河流有流量和水位数据,但在非补水期间,河流基本为干涸状态,没有流量和水位数据,因此模拟不考虑河流流量和地表水位。另外,研究区内有大型地下水源地,地下水开采量对地下水位影响很大。因此确定地下水位、降水量、人工回补量、地下水开采量、温度和湿度作为本次地下水位机器学习模型的参数。

在机器学习模型建模和预测中,参数滞后期识别非常重要^[26]。地下水开采量数据为月数据,不考虑地下水开采量的滞后影响;根据 Zanotti 等^[27]的研究,不考虑湿度和温度的滞后期;人工回补和降水过程发生后,研究区地下水位的抬升不是瞬间产生的,因此需要考虑地下水位、人工回补量和降水量的

滞后影响。本次模拟所用数据量为2 125,根据 Rajae 等^[25]的研究选择的特征参数数量应控制在10~20,因此确定特征筛选后的输入参数数量为20。

3.1.1 地下水位有效滞后期确定

根据研究区包气带厚度及介质情况,设置地下水位滞后期上限为20 d。首先计算2号监测井的自相关函数,结果为0.956~0.997,相关性较高。可见,采用自相关函数无法识别出研究区的地下水位有效滞后期。产生这种现象的主要原因为监测井地下水位受河道补水和水源地地下水开采等人为因素影响显著,河道补水方案会根据补水水源、降雨等情况调整,每年的补水时间和补水量均不同,地下水位变化没有周期性,需进一步计算偏自相关系数^[28]。对地下水位时间序列进行二阶差分后,进行偏自相关分析,得到1~8号监测井的偏自相关系数,结果如表1所示,置信区间选为 $r_{PA}>0.05$ 及 $r_{PA}<-0.05$ 区间,超过置信区间的地下水位滞后期被认为是机器学习模型预测中的有效地下水位滞后期。由表1可见,1~8号监测井地下水位有效滞后期为5~11 d,地下水位滞后期均较小,这主要与研究区水文地质条件、近年来开展的地下水回补工作有关。研究区河道底部介质渗透系数为0.67~17.5 m/d,2015—2020年区域内潮白河地下水回补河段年均补水量0.54万 m^3 ,河道渗透性良好,回补期间河道存在水面,回补结束约一个月后随着河道的入渗蒸发,水面逐渐消失。

3.1.2 地下水位主要影响参数的特征筛选

对主要影响参数与地下水位进行相关性分析。研究区地下水位变化主要受降水量、人工回补量、地下水开采量、温度和湿度影响,其中地下水位抬升需要考虑带有滞后的人工回补量、降水量^[29]。将滞后14 d的人工回补量和降水量作为模型输入,同时考虑地下水开采量、降水、温度、湿度等参数,计算1~8号监测井的地下水位变化与各参数的皮尔逊相关系数,结果如表2所示,表中 Q 为人工回补量, P 为

表1 不同监测井地下水位与滞后时间的偏自相关系数

Table 1 Partial autocorrelation coefficients between groundwater level and lag time of different mointoring wells		不同滞后时间下的偏自相关系数																			
监测井 编号	0d	1d	2d	3d	4d	5d	6d	7d	8d	9d	10d	11d	12d	13d	14d	15d	16d	17d	18d	19d	20d
1	1.00	1.00	-0.99	0.99	-0.95	0.40	-0.18	0.04	-0.04	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.03	0.03
2	1.00	1.00	-0.60	-0.39	-0.16	0.04	0.16	-0.12	-0.13	-0.08	-0.06	-0.16	-0.04	0.05	0.06	0.01	0.01	-0.03	-0.02	-0.01	-0.06
3	1.00	1.00	-0.91	0.84	-0.66	0.10	-0.26	0.27	-0.27	0.02	-0.02	-0.13	0.02	0.05	-0.03	-0.14	0.01	-0.10	-0.09	-0.04	0.10
4	1.00	1.00	-0.93	0.80	-0.63	0.19	-0.18	0.07	-0.25	0.17	-0.07	0.04	-0.14	-0.04	0.01	-0.06	-0.07	0.02	-0.04	-0.08	0.05
5	1.00	1.00	-1.00	1.00	-0.96	0.41	-0.18	0.05	-0.03	0.00	-0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.03	0.04	0.04
6	1.00	1.00	-0.63	-0.34	-0.20	-0.13	-0.08	-0.05	-0.04	-0.02	-0.02	-0.01	-0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02
7	1.00	0.59	0.25	0.00	-0.07	-0.09	-0.04	0.02	0.05	0.07	0.06	0.04	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01
8	1.00	0.92	-0.76	0.53	-0.32	0.06	-0.06	0.16	-0.18	0.08	-0.06	0.04	-0.06	-0.01	-0.03	0.01	0.05	0.03	0.01	0.02	0.01

表 2 不同监测井地下水位与参数的皮尔逊相关系数

Table 2 Pearson correlation coefficients between groundwater level and parameters of different monitoring wells																	
参数	1 号	2 号	3 号	4 号	5 号	6 号	7 号	8 号	参数	1 号	2 号	3 号	4 号	5 号	6 号	7 号	8 号
Q	0.31	0.25	0.22	0.25	0.34	0.36	0.06	0.23	P_2	0.05	0.00	-0.03	-0.02	0.02	0.02	0.02	0.01
Q_1	0.31	0.25	0.23	0.25	0.34	0.36	0.07	0.23	P_3	0.05	0.01	-0.03	-0.02	0.03	0.02	0.02	0.02
Q_2	0.31	0.25	0.23	0.26	0.35	0.37	0.08	0.23	P_4	0.05	0.01	-0.03	-0.02	0.03	0.02	0.02	0.02
Q_3	0.31	0.26	0.23	0.26	0.35	0.37	0.08	0.23	P_5	0.05	0.01	-0.03	-0.02	0.03	0.02	0.03	0.02
Q_4	0.31	0.26	0.23	0.26	0.35	0.37	0.09	0.24	P_6	0.05	0.01	-0.03	-0.02	0.03	0.03	0.03	0.02
Q_5	0.32	0.27	0.23	0.26	0.36	0.38	0.10	0.24	P_7	0.05	0.01	-0.03	-0.02	0.03	0.03	0.03	0.01
Q_6	0.32	0.27	0.23	0.26	0.36	0.38	0.10	0.24	P_8	0.05	0.01	-0.03	-0.02	0.03	0.03	0.03	0.01
Q_7	0.32	0.27	0.24	0.26	0.36	0.38	0.11	0.24	P_9	0.05	0.02	-0.03	-0.02	0.03	0.03	0.03	0.01
Q_8	0.32	0.28	0.24	0.26	0.37	0.38	0.12	0.24	P_{10}	0.06	0.02	-0.03	-0.02	0.04	0.03	0.03	0.01
Q_9	0.32	0.28	0.24	0.27	0.37	0.38	0.13	0.25	P_{11}	0.06	0.02	-0.02	-0.02	0.04	0.03	0.03	0.01
Q_{10}	0.32	0.29	0.24	0.27	0.37	0.39	0.14	0.25	P_{12}	0.06	0.02	-0.02	-0.02	0.04	0.03	0.04	0.00
Q_{11}	0.32	0.29	0.25	0.27	0.38	0.39	0.14	0.25	P_{13}	0.06	0.02	-0.02	-0.02	0.04	0.04	0.04	0.00
Q_{12}	0.32	0.29	0.25	0.27	0.38	0.39	0.15	0.25	P_{14}	0.06	0.02	-0.02	-0.02	0.04	0.04	0.04	0.00
Q_{13}	0.33	0.30	0.25	0.27	0.38	0.39	0.16	0.25	R_a	0.36	0.49	0.44	0.44	0.31	0.30	0.49	0.91
Q_{14}	0.33	0.30	0.25	0.27	0.38	0.39	0.17	0.25	R_b	0.78	0.77	0.84	0.86	0.85	0.86	0.48	0.05
P	0.04	0.00	-0.03	-0.03	0.02	0.02	0.01	0.01	T	-0.30	-0.06	-0.13	-0.10	0.10	0.09	-0.07	-0.06
P_1	0.04	0.00	-0.03	-0.02	0.02	0.02	0.01	0.02	H	0.12	0.03	-0.11	-0.08	0.05	0.01	0.11	-0.43

降水量, Q 和 P 的下标数字代表滞后天数, 如 Q_1 代表滞后 1 d 的人工回补量; R_a 为水源地 a 开采量; R_b 为水源地 b 开采量; T 为温度; H 为湿度, 下同。置信区间为 $r_p > 0.2$ 及 $r_p < -0.2$ 区间, 置信区间外的相关系数绝对值越大, 地下水位与该参数的相关性越大。由表 2 可见, 地下水开采量和人工回补量与各监测井地下水位的相关系数均较大, 尤其是距离水源地或回补河道较近的 1~6 号监测井, 相关系数均大于 0.2, 河道外围的 7~8 号监测井相关系数相对较小, 但仍大于降水量、温度和湿度, 说明地下水开采量和人工回补量是研究区地下水位的主要影响因素, 降水量、温度和湿度影响则相对较小。

为提高地下水位机器学习模型模拟精度和运行效率, 应用 mRMR 法对地下水位主要影响参数进行冗余数据筛选^[30]。将滞后 14 d 的人工回补量和降水量、两个水源地的地下水开采量、温度、湿度作为模型的备选输入, 分别对 1~8 号监测井地下水位主要影响参数进行评价, 获得各参数的影响排序。

表 3 应用 mRMR 法获得的不同监测井排名前 20 的参数

Table 3 Top 20 parameters of different monitoring wells obtained by applying mRMR method																	
排名	1 号	2 号	3 号	4 号	5 号	6 号	7 号	8 号	排名	1 号	2 号	3 号	4 号	5 号	6 号	7 号	8 号
1	Q_6	R_a	R_a	R_a	R_a	R_a	R_a	R_a	11	P_8	P_9	P_3	P_5	Q_3	Q_{11}	P_{10}	P
2	R_a	H	H	Q_3	Q_{11}	Q_9	Q_8	R_b	12	Q_{13}	Q_4	P_4	Q_{11}	P_4	P_{13}	Q_{11}	P_4
3	R_b	R_b	R_b	R_b	R_b	R_b	R_b	H	13	P	P_7	P	P_{13}	Q_{12}	P_{11}	P_{13}	Q_{10}
4	H	Q_{14}	Q_9	T	H	H	T	Q_1	14	Q_1	Q_{13}	P_{13}	Q_1	P_{12}	Q_6	Q_1	P_{12}
5	Q_{14}	T	T	Q_{14}	T	P_{12}	P	T	15	P_{13}	P_5	Q_{12}	P_{11}	P_3	P_3	P_5	P
6	T	Q_1	P_{12}	H	P	Q_{14}	H	Q_{14}	16	Q_7	P_2	P_6	P_6	Q_7	Q_{13}	P_{11}	Q_{13}
7	P	P_6	Q_1	P_{14}	P_7	T	P_{14}	P_9	17	P_3	P	P_{11}	Q_6	P_{13}	P_9	Q_9	P_{10}
8	P_{14}	P_4	P_{14}	P	Q_{14}	P	Q_{14}	Q_3	18	P_{12}	P_{14}	P_2	P_1	Q	P_4	P_4	Q_5
9	P_2	Q_{12}	Q_{14}	P_{12}	P_2	P_{14}	P_{12}	P_{14}	19	Q_{12}	P_{12}	Q_8	Q_{13}	P_5	Q_1	Q_{13}	P_1
10	Q_3	P_8	P_5	Q_8	P_{14}	P_{10}	Q_3	P_2	20	P_1	Q_{11}	P_9	P_{10}	P	P_2	P_2	P_{13}

表 3 为应用 mRMR 法获得的排名前 20 的参数。由表 3 可见, 应用 mRMR 法特征筛选后的参数中降水量滞后数据多于人工回补量数据, 而从地下水位与各参数的相关性分析结果可知, 地下水位与人工回补量的相关性远大于降水量, 可见 mRMR 法更侧重于降水量及滞后量, 忽略了人工回补量的筛选。因此, 应用 RF 法对地下水位主要影响参数重要性进行分析, 结果如表 4 所示, 进而筛选出重要性排名前 20 的参数。由表 4 可见, 应用 RF 法筛选后的参数保留了大量人工回补量数据, 但降水量数据又偏少, 可见该法侧重于筛选人工回补量及滞后量, 但忽略了降水量的筛选。因此, 本文参数的特征筛选综合 mRMR 法与 RF 法评价结果, 最终确定出各监测井的模型输入参数, 结果见表 5, 其中 L 为地下水位, 下标数字代表滞后天数。由表 5 可见, 采用 mRMR 法和 RF 法联合筛选后的参数既考虑了人工回补量滞后的影响, 又考虑了降水量滞后的影响, 有效避免了单一方法的缺陷。

表 4 应用 RF 法获得的不同监测井参数特征重要性占比

Table 4 Proportion of parameter characteristic importance of different monitoring wells obtained by RF method																	
参数	不同监测井参数重要性占比/%								参数	不同监测井参数重要性占比/%							
	1 号	2 号	3 号	4 号	5 号	6 号	7 号	8 号		1 号	2 号	3 号	4 号	5 号	6 号	7 号	8 号
Q	1.29	1.02	1.43	0.37	0.11	0.90	0.31	0.69	P_2	0.54	0.69	0.63	0.07	0.05	0.48	0.10	0.04
Q_1	1.23	0.93	1.28	0.35	0.09	0.79	0.29	0.53	P_3	0.61	0.85	0.81	0.08	0.06	0.56	0.16	0.07
Q_2	1.76	1.54	1.70	0.71	0.14	1.33	0.38	1.56	P_4	0.43	0.56	0.48	0.06	0.04	0.34	0.08	0.04
Q_3	1.54	1.19	1.64	0.65	0.14	1.16	0.37	1.42	P_5	0.42	0.42	0.40	0.05	0.04	0.31	0.06	0.03
Q_4	1.28	0.99	1.40	0.36	0.10	0.81	0.30	0.59	P_6	0.31	0.38	0.33	0.05	0.02	0.24	0.05	0.03
Q_5	1.52	1.10	1.62	0.45	0.13	1.14	0.36	1.37	P_7	0.43	0.55	0.45	0.06	0.04	0.34	0.08	0.04
Q_6	38.02	29.94	34.89	39.96	90.27	61.76	69.82	32.57	P_8	0.38	0.38	0.36	0.05	0.03	0.30	0.06	0.03
Q_7	1.06	0.91	1.15	0.30	0.08	0.74	0.17	0.41	P_9	0.43	0.50	0.44	0.06	0.04	0.33	0.07	0.03
Q_8	1.17	0.93	1.20	0.32	0.09	0.78	0.26	0.49	P_{10}	0.48	0.65	0.53	0.07	0.05	0.39	0.08	0.04
Q_9	1.34	1.07	1.52	0.43	0.11	1.14	0.33	1.30	P_{11}	0.60	0.79	0.76	0.08	0.06	0.54	0.12	0.07
Q_{10}	1.33	1.06	1.49	0.42	0.11	1.03	0.32	0.95	P_{12}	0.55	0.77	0.64	0.07	0.05	0.54	0.10	0.05
Q_{11}	2.67	4.45	2.43	3.39	0.30	1.93	0.82	6.47	P_{13}	0.40	0.39	0.39	0.05	0.04	0.32	0.06	0.03
Q_{12}	2.04	1.82	1.79	0.72	0.14	1.48	0.39	1.65	P_{14}	0.45	0.57	0.50	0.07	0.05	0.34	0.08	0.04
Q_{13}	2.45	1.89	2.11	0.72	0.15	1.48	0.40	2.81	R_a	6.90	8.27	2.76	10.79	0.45	3.09	1.20	7.42
Q_{14}	2.60	2.31	2.22	0.77	0.16	1.64	0.70	5.31	R_b	15.17	16.87	21.61	26.26	5.99	6.85	17.02	22.42
P	0.48	0.62	0.50	0.07	0.05	0.39	0.08	0.04	T	7.16	13.08	7.88	11.38	0.62	4.61	4.84	7.76
P_1	0.42	0.50	0.41	0.05	0.04	0.33	0.07	0.03	H	2.55	2.02	2.22	0.72	0.15	1.57	0.47	3.65

表 5 不同监测井经过特征筛选后的模型输入参数

Table 5 Input parameters for models after feature selection of different monitoring wells								
监测井 编号	地下水位参数				主要影响参数			
	数量	名称			数量	名称		
1	6	$L_1、L_2、L_3、L_4、L_5、L_6$			14	$R_a、R_b、H、T、Q、Q_1、Q_3、Q_6、Q_{11}、Q_{13}、Q_{14}、P_2、P_8、P_{14}$		
2	11	$L_1、L_2、L_3、L_4、L_5、L_6、L_7、L_8、L_9、L_{10}、L_{11}$			9	$R_a、R_b、H、T、Q_1、Q_9、Q_{14}、P_4、P_6$		
3	11	$L_1、L_2、L_3、L_4、L_5、L_6、L_7、L_8、L_9、L_{10}、L_{11}$			9	$R_a、R_b、H、T、Q_1、Q_9、Q_{14}、P_{12}、P_{14}$		
4	9	$L_1、L_2、L_3、L_4、L_5、L_6、L_7、L_8、L_9$			11	$R_a、R_b、H、T、Q_3、Q_8、Q_{10}、Q_{14}、P、P_{12}、P_{14}$		
5	6	$L_1、L_2、L_3、L_4、L_5、L_6$			14	$R_a、R_b、H、T、Q_3、Q_{10}、Q_{11}、Q_{12}、Q_{14}、P、P_2、P_4、P_7、P_{14}$		
6	6	$L_1、L_2、L_3、L_4、L_5、L_6$			14	$R_a、R_b、H、T、Q_6、Q_9、Q_{11}、Q_{14}、P、P_{10}、P_{11}、P_{12}、P_{13}、P_{14}$		
7	5	$L_1、L_2、L_3、L_4、L_5$			15	$R_a、R_b、H、T、Q_1、Q_3、Q_9、Q_{11}、Q_{14}、P_5、P_{10}、P_{11}、P_{12}、P_{13}、P_{14}$		
8	8	$L_1、L_2、L_3、L_4、L_5、L_6、L_7、L_8$			12	$R_a、R_b、H、T、Q_1、Q_3、Q_{13}、Q_{14}、P、P_2、P_9、P_{14}$		

3.2 特征筛选对地下水位机器学习模型模拟效果的影响

3.2.1 特征筛选对模型模拟精度的影响

将未经过特征筛选的 34 个参数和经过特征筛选的 20 个参数分别输入到已经构建好的密怀顺区域 ELM、NARX、RF 模型中,对 1~8 号监测井的地下水位进行模拟,对输出结果进行对比分析。由于篇幅限制,只列出 1 号和 7 号监测井应用不同机器学习模型在参数特征筛选前后的模拟值与监测值的对比,如图 3 所示。由图 3 可见,无论是距离河道较近的 1 号监测井,还是距离河道较远的 7 号监测井,参数经过特征筛选后,ELM 模型模拟值与监测值吻合效果明显优于筛选前,特征筛选对于 ELM 模型的模拟精度有明显提升。RF 模型的输入参数经过特征筛选后,模拟值也更接近监测值,筛选效果较好,特征筛选对于 RF 模型模拟精度有一定程度的提升。NARX 模型在 1 号监测井进行模拟时,特征筛选前后模拟值均与监测值吻合较好,模型模拟精度没有明显改变;在 7 号监测井的模拟中,特征筛选后的模拟值与监测值反而出现了一定差距,特征筛选在一

定程度上降低了模拟精度。可见,在不同地下水位机器学习模型中,特征筛选对模型模拟精度的影响不同。

分别计算 1~8 号监测井在 3 种地下水位机器学习模型中特征筛选前后的 RMSE、NSE 和 R^2 值,结果见表 6。由表 6 可见,特征筛选对 ELM 模型的 3 项评价指标均有很大提升,RMSE、NSE 和 R^2 分别提升了 63%、98%和 45%,特征筛选前 ELM 模型的模拟结果几乎不可信,而经过特征筛选后的模拟结果可信度得到了很大提高。这可能是由于 ELM 模型结构简单,更善于处理低维度的简单数据,特征筛选对于 ELM 模型构建十分必要。NARX 模型特征筛选前后的 RMSE、NSE 和 R^2 变化不大,其中 RMSE、NSE 出现小幅度下降,分别下降了 14%、9%。NARX 模型特征筛选前的模拟精度已经达到较高水平,筛选后出现模拟精度下降现象,推测是由于 NARX 模型中的神经网络结构更善于从复杂数据中找寻规律,特征维数增加会进一步提高精度。RF 模型特征筛选后的 RMSE、NSE 和 R^2 分别提升了 49%、6%和 2%,评价指标除 RMSE 外,其他指标提

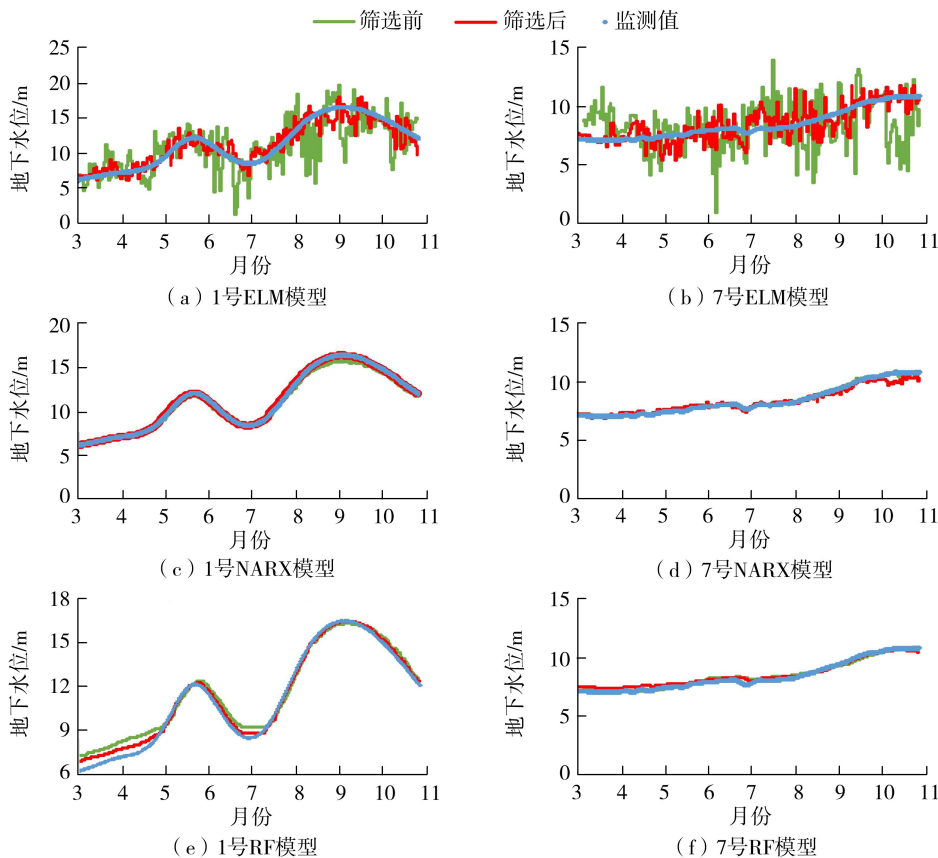


图3 1号和7号监测井地下水位机器学习模型中特征筛选前后的模拟效果对比

Fig.3 Comparison of simulation effects before and after feature selection in machine learning models for groundwater level of monitoring wells 1 and 7

表6 3种模型特征筛选前后的模拟效果评价指标平均值

Table 6 Average of evaluation indicators calculated by three models before and after feature selection

模型	时段	RMSE/m	NSE	R^2
ELM	筛选前	2.050	0.018	0.477
	筛选后	0.753	0.836	0.874
NARX	筛选前	0.316	0.959	0.982
	筛选后	0.361	0.880	0.981
RF	筛选前	0.704	0.917	0.956
	筛选后	0.357	0.972	0.972

升不明显。这主要是由于RF模型在迭代模拟过程中内置了特征筛选这一过程,在模型运行时对高维数据筛选能力较强,不易受到冗余信息干扰,因此在输入特征维度较高时也体现了较高精确度。

综上所述,对于3种机器学习模型,特征筛选在模拟精度方面提升效果由大到小依次为ELM模型、RF模型、NARX模型,特征筛选极大地提升了ELM模型模拟精度,对RF模型精度也有一定提高,但对NARX模型没有明显效果。可见,特征筛选对于ELM模型构建具有重要意义。

3.2.2 特征筛选对模型运行速度的影响

对3种地下水位机器学习模型特征筛选前后的运行速度进行比较,模型均在型号为Inter Core i7-

4970的8核CPU的计算机中运行,结果显示,特征筛选前后ELM模型的运行速度均为5s左右,RF模型则为10s左右,特征筛选对ELM模型和RF模型运行速度几乎没有影响。特征筛选前NARX模型运行时间为18min,特征筛选后运行时间为1.5min,模型运行速度提升了11倍,可见特征筛选极大程度提升了NARX模型的运行速度,减少了模型的运算负担,同时也保证了模型精度,可见特征筛选对NARX模型构建具有重要意义。

4 结论

a. 密怀顺区域地下水位机器学习模型中特征参数包括地下水位及滞后值、地下水开采量、人工回补量及滞后值、降水量及滞后值、温度、湿度等20个参数。地下水位及滞后天数可利用偏自相关系数法获取。人工回补量及滞后值、降水量及滞后值由mRMR法和RF法联合确定,其中mRMR法侧重筛选人工回补量及滞后值的数据冗余,RF法则侧重筛选降水量及滞后值的数据冗余。

b. 特征筛选可以有效提高ELM模型和RF模型的模拟精度,但对NARX模型作用较小。特征筛选后,ELM模型的RMSE、NSE和 R^2 分别提升了

63%、98%和45%,RF模型分别提升了49%、6%和2%,NARX模型的RMSE和NSE则分别下降了14%和9%。

c. 特征筛选可以有效提升NARX模型运行效率,特征筛选后NARX模型运行速度提升了11倍。特征筛选对ELM模型和RF模型运行速度没有明显作用。

参考文献:

[1] SAHOO S, RUSSO T A, ELLIOTT J, et al. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U. S. [J]. Water Resources Research,2017,53(5):3878-3895.

[2] 陈云天. 基于机器学习的测井曲线补全与生成研究 [D]. 北京:北京大学,2020.

[3] 牛欣怡,鲁程鹏,卢佳赞,等. 机器学习模型在地下水埋深模拟中的适应性分析[J]. 河海大学学报(自然科学版),2022,50(4):74-82. (NIU Xinyi,LU Chengpeng,LU Jiayun,et al. Adaptability analysis of machine learning model in groundwater depth simulation [J]. Journal of Hohai University (Natural Sciences), 2022, 50 (4): 74-82. (in Chinese))

[4] 陈珺,黄燕华,洪朋,等. 基于机器学习模型的河道水位预测[J]. 水利水电科技进展,2023,43(3):9-14. (CHEN Jun, HUANG Yanhua, HONG Peng, et al. Prediction of river water level based on machine learning model[J]. Advances in Science and Technology of Water Resources,2023,43(3):9-14. (in Chinese))

[5] 张晓婷. 基于特征筛选及机器学习方法的混合区域土壤制图[D]. 武汉:华中农业大学,2023.

[6] BARZEGAR R, FIJANI E, MOGHADDAM A A, et al. Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models[J]. Science of the Total Environment,2017,599-600:20-31.

[7] 刘浩淇. 基于机器学习的胃癌特征基因筛选及预测 [D]. 郑州:河南大学,2023.

[8] KHOSRAVI K,PANAHI M,BUI D T. Spatial prediction of groundwater spring potential mapping based on an adaptive neuro-fuzzy inference system and metaheuristic optimization[J]. Hydrology and Earth System Sciences, 2018,22(9):4771-4792.

[9] TANG Yandong,ZANG Cuiping,WEI Yong,et al. Data-driven modeling of groundwater level with least-square support vector machine and spatial-temporal analysis[J]. Geotechnical and Geological Engineering,2019,37(3):1661-1670.

[10] 李邳琴,杜建强,聂斌,等. 特征选择方法综述[J]. 计算机工程与应用,2019,55(24):10-19. (LI Zhiqin,DU Jianqiang,NIE Bin,et al. Summary of feature selection

methods [J]. Computer Engineering and Applications, 2019,55(24):10-19. (in Chinese))

[11] OUEDRAOGO I, DEFOURNY P, VANCLOOSTER M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale[J]. Hydrogeology Journal,2019,27(3):1081-1098.

[12] 钟雅婷,林艳梅,陈定甲,等. 多组学数据整合分析和应用研究综述[J]. 计算机工程与应用,2021,57(23):1-17. (ZHONG Yating,LIN Yanmei,CHEN Dingjia,et al. Review on integration analysis and application of multi-omics data [J]. Computer Engineering and Applications, 2021,57(23):1-17. (in Chinese))

[13] 北京市水科学技术研究院. 北京市水科学技术研究院助力2021年春季潮白河试验性生态补水工作[J]. 北京水务,2021(5):8-9. (Beijing Water Science and Technology Institute. Beijing Institute of Water Science and Technology supports the experimental ecological water replenishment work of Chaobai River in spring 2021 [J]. Beijing Water,2021(5):8-9. (in Chinese))

[14] 郭敏丽,陈鹏,马岩川,等. 地下水回补对密怀顺地下水资源环境的影响分析[J]. 北京水务,2022(3):14-18. (GUO Minli,CHEN Peng,MA Yanchuan,et al. Impact analysis of groundwater recharge on groundwater resources and environment in Mihuaishun area [J]. Beijing Water, 2022(3):14-18. (in Chinese))

[15] 北京市水科学技术研究院. 北京水问题研究与实践 [M]. 北京:中国水利水电出版社,2023:3-8.

[16] JONES A L,SMART P L. Spatial and temporal changes in the structure of groundwater nitrate concentration time series (1935-1999) as demonstrated by autoregressive modelling[J]. Journal of Hydrology,2005,310(1/2/3/4):201-215.

[17] PENG H C, LONG F H, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. Ieee Transactions on Pattern Analysis and Machine Intelligence,2005,27(8):1226-1238.

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

[19] HUANG Guangbin,ZHU Qinyu,SIEW C K. Extreme learning machine: theory and applications [J]. Neurocomputing,2006,70(1/2/3):489-501.

[20] 郭敏丽,刘天航,毕二平,等. 3种机器学习模型在地下水位动态模拟中的效果比较[J]. 环境工程学报,2024,18(5):1406-1414. (GUO Minli,LIU Tianhang,BI Erping,et al. Comparison of three machine learning models in dynamic simulation of groundwater level [J]. Chinese Journal of Environmental Engineering,2024,18(5):1406-1414. (in Chinese))

(下转第221页)

水 $\delta^{18}\text{O}$ 的特征及水汽来源[J]. 科学通报, 2009, 54(22): 3521-3531. (LIU Jianrong, SONG Xianfang, YUAN Guofu, et al. Characteristics of $\delta^{18}\text{O}$ in precipitation over eastern monsoon China and the water vapor sources[J]. Chinese Science Bulletin, 2009, 54(22): 3521-3531. (in Chinese))

[32] XIE Luhua, WEI Gangjian, DENG Wenfeng, et al. Daily $\delta^{18}\text{O}$ and δD of precipitations from 2007 to 2009 in Guangzhou, south China: implications for changes of moisture sources[J]. Journal of Hydrology, 2011, 400(3-4): 477-489.

[33] CRAWFORD J, HOLLINS S E, MEREDITH K T, et al. Precipitation stable isotope variability and subcloud evaporation processes in a semi-arid region [J]. Hydrological Processes, 2017, 31(1): 20-34.

[34] VIMEUX F, GALLAIRE R, BONY S, et al. What are the climate controls on δD in precipitation in the Zongo Valley (Bolivia)? Implications for the Illimani ice core interpretation[J]. Earth and Planetary Science Letters, 2005, 240(2): 205-220.

[35] ADHIKARI N, GAO Jing, YAO Tandong, et al. The influence of moisture transport processes on the stable isotopic compositions in precipitation on the South slope of the Himalayas[J]. Global and Planetary Change, 2024, 237: 104453.

[36] TAN Ming. Circulation effect: response of precipitation $\delta^{18}\text{O}$ to the ENSO cycle in monsoon regions of China[J]. Climate Dynamics, 2014, 42(3-4): 1067-1077.

[37] 宋献方, 唐瑜, 张应华, 等. 北京连续降水水汽输送差异的同位素示踪[J]. 水科学进展, 2017, 28(4): 488-495. (SONG Xianfang, TANG Yu, ZHANG Yinghua, et al. Using stable isotopes to study vapor transport of continuous precipitation in Beijing [J]. Advances in Water Science, 2017, 28(4): 488-495. (in Chinese))

[38] WANG Di, TIAN Lide, CAI Zhongyin, et al. Indian monsoon precipitation isotopes linked with high level cloud cover at local and regional scales[J]. Earth and Planetary Science Letters, 2020, 529: 115837.

[39] WEI Zhongwang, LEE Xuhui, LIU Zhongfang, et al. Influences of large-scale convection and moisture source on monthly precipitation isotope ratios observed in Thailand, southeast Asia[J]. Earth and Planetary Science Letters, 2018, 488: 181-192.

[40] CAI Zhongyin, TIAN Lide. Atmospheric controls on seasonal and interannual variations in the precipitation isotope in the east Asian monsoon region[J]. Journal of Climate, 2016, 29(4): 1339-1352.

[41] GUO Rong, YU Wusheng, ZHANG Jingyi, et al. Different dynamics drive Indian Ocean moisture to the southern slope of central Himalayas: an isotopic approach [J]. Geophysical Research Letters, 2024, 51(11): e2024GL109359.

[42] ZHOU Hui, ZHANG Xinping, YAO Tianci, et al. Variation of $\delta^{18}\text{O}$ in precipitation and its response to upstream atmospheric convection and rainout: a case study of Changsha Station, south-central China[J]. Science of the Total Environment, 2019, 659: 1199-1208.

(收稿日期: 2024-08-29 编辑: 胡新宇)

(上接第 186 页)

[21] WUNSCH A, LIESCH T, BRODA S. Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX)[J]. Journal of Hydrology, 2018, 567: 743-758.

[22] IZADY A, DAVARY K, ALIZADEH A, et al. Application of NN-ARX model to predict groundwater levels in the neishaboor plain, Iran[J]. Water Resources Management, 2013, 27(14): 4773-4794.

[23] LIN T, HORNE B G, TINO P, et al. Learning long-term dependencies in NARX recurrent neural networks [J]. IEEE Transactions on Neural Networks, 1996, 7(6): 1329-1338.

[24] SAMARASINGHE S. Neural networks for applied sciences and engineering[M]. New York: Taylor and Francis, 2007.

[25] RAJAEI T, EBRAHIMI H, NOURANI V. A review of the artificial intelligence methods in groundwater level modeling[J]. Journal of Hydrology, 2019, 572: 336-351.

[26] 刘天航. 基于机器学习的密怀顺地区短期地下水位动态模型研究[D]. 北京: 中国地质大学(北京), 2022.

[27] ZANOTTI C, ROTIROTI M, STERLACCHINI S, et al. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system [J]. Journal of Hydrology, 2019, 578: 124015.

[28] DU Kongchang, ZHAO Ying, LEI Jiaqiang. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series[J]. Journal of Hydrology, 2017, 552: 44-51.

[29] ZARE M, KOCH M. Groundwater level fluctuations simulation and prediction by ANFIS- and hybrid wavelet-ANFIS/fuzzy C-means (FCM) clustering models: application to the Miandarband plain [J]. Journal of Hydro-Environment Research, 2018, 18: 63-76.

[30] PENG Hanchuan, LONG Fuhui, DING Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

(收稿日期: 2024-07-16 编辑: 王芳)