



DOI:10.12404/j.issn.1671-1815.2405660

引用格式:高明,李虎,刘鑫锦,等.基于距离-属性混合聚类结合 ConvLSTM 模型的岩溶地层地下水预测方法[J].科学技术与工程,2025,25(20):8424-8434.

Gao Ming, Li Hu, Liu Xinjin, et al. Prediction method of groundwater in karst strata based on distance-attribute hybrid clustering combined with ConvLSTM model[J]. Science Technology and Engineering, 2025, 25(20): 8424-8434.

基于距离-属性混合聚类结合 ConvLSTM 模型的 岩溶地层地下水预测方法

高明¹, 李虎^{1,2*}, 刘鑫锦², 张康², 韩健勇¹

(1. 山东建筑大学土木工程学院, 济南 250101; 2. 济南轨道交通集团有限公司, 济南 250014)

摘要 为解决因地下水相关因素未考虑充分而导致的模型对地下水位预测不准确的问题,提出观测井的空间位置距离聚类方法、水文地质属性聚类方法和距离-属性混合聚类方法,验证观测井间连通性在地下水位预测中的重要性。设计 4 种模型并分别对济南岩溶水层的地下水位进行模拟和预测并与实际观测值对比。预测结果表明:考虑岩溶含水层连通性特征的联合模型 ConvLSTM(convolution-long short term memory)要优于传统的长短期记忆网络模型(long short term memory, LSTM)。其中考虑距离-属性混合聚类结果的同类别井(连通性强)的模型(mix-multivariate-convolution-long short term memory, M-MV-ConvLSTM)预测结果精度最高、误差最小,其平均均方根误差约为 0.457,纳什效率系数约为 0.216,预测准确度高于传统的 LSTM 预测模型。研究成果可为岩溶水层的实时地下水位预测提供借鉴。

关键词 地下水位预测;长短期记忆网络(LSTM);聚类;岩溶含水层

中图分类号 P641.1;

文献标志码 A

Prediction Method of Groundwater in Karst Strata Based on Distance-attribute Hybrid Clustering Combined with ConvLSTM Model

GAO Ming¹, LI Hu^{1,2*}, LIU Xin-jin², ZHANG Kang², HAN Jian-yong¹

(1. School of Civil Engineering, Shandong Jianzhu University, Jinan 250101, China;

2. Jinan Rail Transit Group Co., Ltd., Jinan 250014, China)

[Abstract] To address the issue of inaccuracies in groundwater level predictions due to the insufficient consideration of groundwater-related factors, clustering methods for observation wells based on spatial distance, hydrogeological attributes, and a hybrid of distance and attributes were proposed. The significance of inter-well connectivity in groundwater level prediction was validated. Four models were designed, which were applied to simulate and predict groundwater levels in the karst water region of Jinan and compared with actual observations. The prediction results indicate that the combined model incorporating the connectivity characteristics of karst aquifers, known as convolution-long short-term memory (ConvLSTM), outperforms the traditional long short-term memory (LSTM) model. Among the models, the mix-multivariate-convolution-long short-term memory (M-MV-ConvLSTM) model, which accounts for wells of the same category based on the hybrid distance-attribute clustering results (characterized by strong connectivity), achieves the highest prediction accuracy and the smallest error. The average root mean square error is approximately 0.457, and the Nash-Sutcliffe efficiency is approximately 0.216, demonstrating a higher prediction accuracy than the traditional LSTM model. The research results is positioned to serve as a reference for real-time groundwater level prediction in karst regions.

[Keywords] groundwater level prediction; long short term memory network(LSTM); cluster; karst aquifer

一直以来地下水都与人们有着密不可分的联系,不管是对地下水的合理开发与可持续利用,还是由于水位变化引起的城市内涝等灾害,均体现了地下水对人类生活的重要影响。目前,学者们主要关注对地下水位实时监测的研究,忽略了地下水位预测的重要

性。不同位置地下水位之间存在着联系,但由于不同地质条件与周边环境造成地下水位间的关系十分复杂,通常不能有效准确地对地下水位预测。

近年来,深度学习技术不断革新发展。其中长短期记忆网络模型(long short term memory, LSTM)

收稿日期:2024-07-28; 修订日期:2025-04-15

基金项目:国家外国专家项目(G2022023020L);山东省重大科技创新工程(2019JZZY020105);甘肃省重点研发计划(22YF7FH224)

第一作者:高明(1998—),男,汉族,山东临沂人,硕士研究生。研究方向:隧道工程。E-mail:ming.8660@foxmail.com。

*通信作者:李虎(1983—),男,汉族,山东济南人,博士,高级工程师。研究方向:轨道交通工程建设。E-mail:Lihu1007@163.com。

模型^[1-4]适用于处理和预测时间序列中间隔和延迟非常长的重要事件, 并且解决了循环神经网络 (recurrent neural network, RNN) 存在的梯度消失和爆炸问题^[5-6]。因此, 诸多学者基于 LSTM 模型对地下水位预测展开研究。闫佰忠等^[7]为解决以往模型未考虑地下水位相关影响因素的问题, 探讨 LSTM 在地下水位预测中的应用, 利用 LSTM 模型, 采用多变量输入的方式, 构建了基于多变量 LSTM 的地下水位预测模型。汪云等^[8]利用 LSTM 构建地下水位预测模型, 解决了传统神经网络预测模型处理时序数据时未考虑时间序列的问题, 同时采用多影响变量输入的方式弥补了简单时序模型处理数据时过于依赖时间的缺点。冯希尧等^[9]为解决未考虑最优因子组合, 会对地下水潜在性制图产生不利影响等问题, 提出了遗传算法优化支持向量机的地下水潜在性预测方法, 优化后模型的准确度为 0.777, 验证集 (area under curve, AUC) 为 0.806, 结果表明, 遗传算法优化的支持向量机模型的预测准确度高、可靠性好。胡飞跃等^[10]构建了滑动窗口-长短期记忆神经网络模型 (sliding window algorithm-long short term memory, SWA-LSTM), 利用地下水位变化于降雨量的滞后性关系对地下水位进行预测, 弥补了简单 LSTM 模型过度依赖周期性规律而忽视降雨量对地下水位变化规律的影响的缺点。侯金霄等^[11]针对岩溶地下水水位预测精度较差的问题, 提出一种 (empirical mode decomposition-long short term memory, EMD-LSTM) 耦合模型, 具有较强的可靠性和稳定性, 可为岩溶地下水水位的精确预测提供借鉴。郭艺等^[12]运用时间序列分析方法分析了济南市岩溶泉水位动态变化规律, 建立的泉水位预测模型可以较好地预测该地区岩溶泉水位。

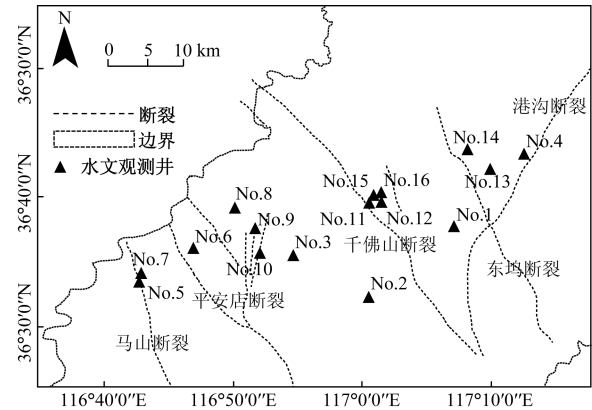
针对地下水位的预测, 传统的神经网络模型已无法满足对预测准确度需求。鉴于此, 把卷积神经网络 (convolutional neural network, CNN)^[13-15] 和 LSTM 网络相结合构成 ConvLSTM 模型, 其不仅具有 LSTM 的时间建模能力, 而且还可以描绘局部特征, 并利用 K-means 聚类算法对观测点的空间位置和水文地质属性聚类, 聚类结果可用于岩溶含水层连通性分析, 在地下结构复杂多变的岩溶地层, 聚类结果与模型结合以期获得合理有效的预测结果。

1 岩溶含水层连通性分析

岩溶含水层水量丰富而不均一, 不同观测井间因水流特征可能会有连通关系, 若忽视观测井之间的连通性, 预测结果准确性将会下降。对观测井间的连通性分析有效避免了模型对地下水位预测的不准确不精确问题。选取山东省济南市马山断裂

与港沟断裂之间的碳酸盐岩岩溶含水层区域。研究区范围和观测井的位置如图 1 所示。

岩溶含水层连通性分析主要包括观测井基本信息、断层断裂、地下水位监测序列和数字高程模型 (digital elevation model, DEM) 数据, 提取各观测井坐标和水文地质因子按特征聚类后得出聚类结果。按条件结合两类结果最后进行连通性分析得出含水层连通性专题图, 连通性分析总流程图如图 2 所示。



基于自然资源部标准地图服务网站审图号鲁 SG(2024)035 号的标准地图制作, 底图无修改; No. 1 ~ No. 16 为水文观测井编号

图 1 研究区范围和观测井位置

Fig. 1 Scope of the study area and the location of observation wells

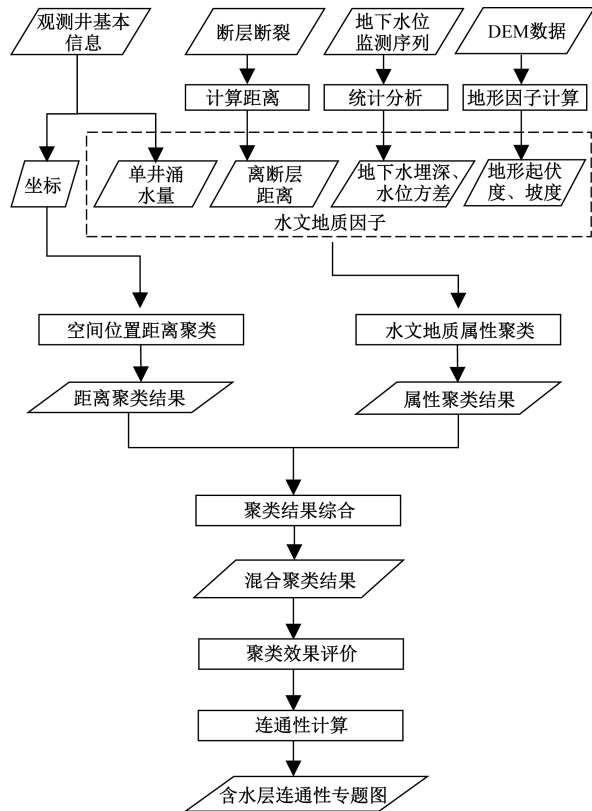


图 2 连通性分析总流程图

Fig. 2 Total flow chart of connectivity analysis

数据源包括该区域内布设的 16 口观测井及其长期地下水位监测序列、30 m 分辨率 DEM 数据如图 3 所示。

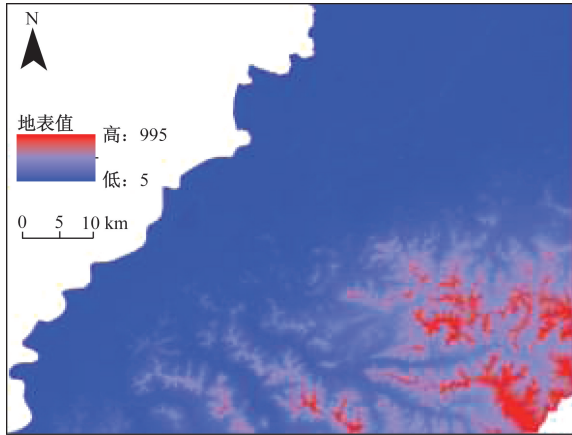


图 3 地表 DEM
Fig. 3 Surface DEM

1.1 聚类变量的选取

读取和计算 16 口观测井的坐标位置、地下水埋深、水位方差、离断层距离,以及观测井处的地形起伏度、坡度等聚类变量,并分别存入观测井位置集合和水文地质属性集合,可分别表示为

$$W_{\text{ellp}} = \{w_{\text{pi}}(x_i, y_i) \mid i = 0, 1, \dots, n_w\} \quad (1)$$

$$W_{\text{Hydro}} = \{w_{\text{hdi}}(g_{\text{di}}, g_{\text{vi}}, w_{\text{ydi}}, w_{\text{tri}}, w_{\text{si}}, w_{\text{fi}}) \mid i = 1, 2, \dots, n_w\} \quad (2)$$

式中: x_i 为观测井的横坐标; y_i 为观测井的纵坐标; w_{pi} 为观测井的位置属性集; n_w 为观测井数量, w_{hdi} 为观测井的水文地质属性集; g_{di} 为地下水埋深因子; g_{vi} 为地下水位方差因子; w_{ydi} 为单井涌水量因子; w_{tri} 为地形起伏度值; w_{si} 为坡度因子, w_{fi} 为离断层距离因子。

选取聚类变量的步骤如下。

步骤 1 创建观测井集合。

$$W_{\text{ell}} = \{w_i(w_{\text{pi}}, w_{\text{hdi}}) \mid i = 1, 2, \dots, n_w\} \quad (3)$$

步骤 2 读取观测井基本信息表中观测井的编号、坐标位置,得到观测井位置属性集合。

$$W_{\text{ellp}} = \{w_{\text{pi}}(x_i, y_i) \mid i = 1, 2, \dots, n_w\} \quad (4)$$

步骤 3 读取观测井的长期地下水位监测时间序列数据,获取和计算观测井的地下水埋深因子 g_{di} , 地下水位方差因子 g_{vi} 。

步骤 4 读取由抽水试验获得的观测井的单井涌水量因子 w_{ydi} 。

步骤 5 读取 DEM 数据,计算观测井所处位置的地形起伏度值 w_{tri} 。

步骤 6 计算观测井所处位置的坡度值,得到坡度因子 w_{si} 。

步骤 7 读取断裂数据文件,计算观测井距最近的断裂的距离,得到离断层距离因子 w_{fi} 。

步骤 8 循环执行步骤 3 ~ 7,得到所有观测井的水文地质属性信息,并存入集合 W_{Hydro} 中。

基于 W_{ellp} 、 W_{Hydro} , 更新观测井集合,可表示为

$$W_{\text{ell}} = \{w_i(w_{\text{pi}}, w_{\text{hdi}}) \mid i = 1, 2, \dots, n_w\} \quad (5)$$

得到更新后的观测井数据如表 1 所示。

1.2 数据标准化

为了消除指标之间的量纲影响,需要进行数据标准化处理,以解决数据指标之间的可比性。原始数据经过数据标准化处理后,各指标处于同一数量级,适合进行综合对比评价。

对观测井的水文地质属性集合 W_{Hydro} 中的各个因子,分别进行 Z-Score 标准化处理,得到标准化处理后的集合 W_{Hydro} 。

表 1 观测井位置和水文地质属

Table 1 Observation well location and hydrogeological properties

编号	X	Y	水位埋深/m	水位方差/m	单井涌水量/(m ³ ·d ⁻¹)	起伏度/m	坡度/(°)	离断层距离/m
No. 1	117.12	36.63	38.47	179.02	500	70	4.55	3 502.63
No. 2	117.01	36.54	37.44	2.99	500	34	3.43	5 943.78
No. 3	116.91	36.59	26.95	3.53	2 500	39	1.18	4 031.99
No. 4	117.21	36.72	29.73	1.14	2 500	31	2.56	838.63
No. 5	116.71	36.56	1.47	1.08	7 500	22	7.34	160.50
No. 6	116.78	36.60	5.99	1.37	7 500	21	1.97	1 558.72
No. 7	116.71	36.57	11.67	1.39	7 500	26	3.43	427.73
No. 8	116.84	36.65	3.20	0.98	7 500	23	3.71	3 654.12
No. 9	116.86	36.62	12.76	1.00	7 500	22	0.66	895.40
No. 10	116.87	36.59	22.97	3.53	7 500	43	6.29	208.02
No. 11	117.01	36.66	8.67	0.24	7 500	44	1.18	912.59
No. 12	117.03	36.66	22.55	0.28	7 500	31	3.03	1 751.44
No. 13	117.17	36.70	23.61	1.93	2 500	22	1.86	2 441.75
No. 14	117.14	36.73	0.82	0.44	2 500	26	6.87	1 465.39
No. 15	117.01	36.67	13.75	0	2 500	25	12.86	1 526.24
No. 16	117.02	36.67	6.74	0	2 500	23	3.84	1 926.57

1.3 空间距离聚类

基于空间位置因子和水文地质因子中所提供的数据, 通过利用 K -means 聚类算法^[16-18] 对数据进行计算分类从而得到相应的聚类结果。聚类方法流程图如图 4 所示。

空间位置距离聚类由 K -means 聚类算法分类得出, 它基于距离要求把符合要求的相应观测井分到一种类别组中^[19], 此分类对后文地下水位预测中起到关键作用。空间位置距离聚类流程如图 4 所示。具体聚类方法如下。

读取观测井位置属集合 W_{allp} , 在 W_{allp} 中随机选取 k 个观测井作为距离聚类初始中心集合, 可表示为

$$D_{\text{CP}} = \{d_{\text{cpj}}(d_{\text{cpjx}}, d_{\text{cpjy}}) \mid j = 1, 2, \dots, k\} \quad (6)$$

并创建距离聚类分组集合:

$$D_{\text{C}} = \{d_{\text{iscatj}}(w_j^m) \mid j = 1, 2, \dots, k; m = 1, 2, \dots, n_j\} \quad (7)$$

式中: d_{cpj} 为距离聚类中心点; d_{cpjx} 为横坐标; d_{cpjy} 为纵坐标; d_{iscatj} 为距离聚类的第 j 个分组; w_j^m 为第 j 个分组中的第 m 个观测井; k 为分组数; n_j 为第 j 个分组中观测井的数目。

开始 K -means 空间位置距离聚类, 读取距离聚类中心点集合, 即

$$D_{\text{CP}} = \{d_{\text{cpj}}(d_{\text{cpjx}}, d_{\text{cpjy}}) \mid j = 1, 2, \dots, k\} \quad (8)$$

根据式(8)计算观测井 w_i 与聚类中心 d_{cpj} 的欧式距离, 计算公式为

$$d_{ij}(w_i, d_{\text{cpj}}) = \sqrt{(w_{xi} - d_{\text{cpjx}})^2 + (w_{yi} - d_{\text{cpjy}})^2} \quad (9)$$

式(9)中: w_{xi} 为观测井横坐标; w_{yi} 为观测井纵坐标。

得到观测井 w_i 与 D_{CP} 中所有聚类中心点的欧式

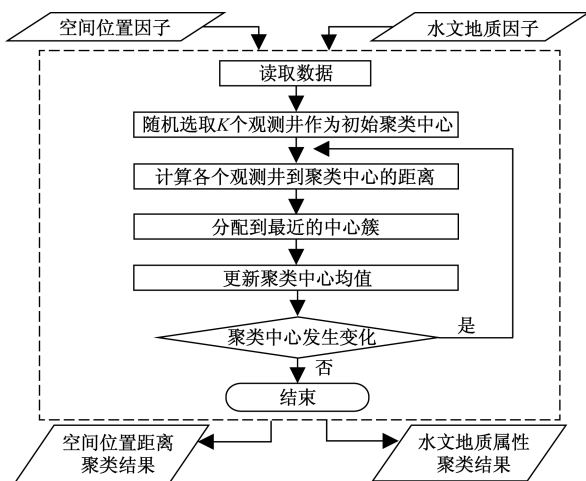


图 4 聚类方法流程图
Fig. 4 Flowchart of the clustering method

距离, 获取与观测井距离最小的聚类中心点编号 j , 将观测井存入 d_{iscatj} 。重复上述操作, 将所有的观测井都分类到距其最近的聚类中心点簇。

更新聚类中心, 计算 d_{iscatj} 中包含的所有观测井的质心, 将其作为新的聚类中心点, 添加至新聚类中心点集合, 即

$$N_{\text{DCP}} = \{n_{\text{depj}}(n_{\text{depxj}}, n_{\text{depyj}}) \mid j = 1, 2, \dots, k\} \quad (10)$$

完成所有分组的中心点更新。并判断更新后的聚类中心点集合 N_{DCP} 相比原聚类中心集 D_{CP} 是否发生改变, 如果两集合相比发生变化, 则将更新后的新聚类中心点集 N_{DCP} 赋值给 D_{CP} 继续利用 K -means 聚类; 如果未发生变化, 则聚类结束, 得到观测井的空间距离聚类结果如图 5 所示。

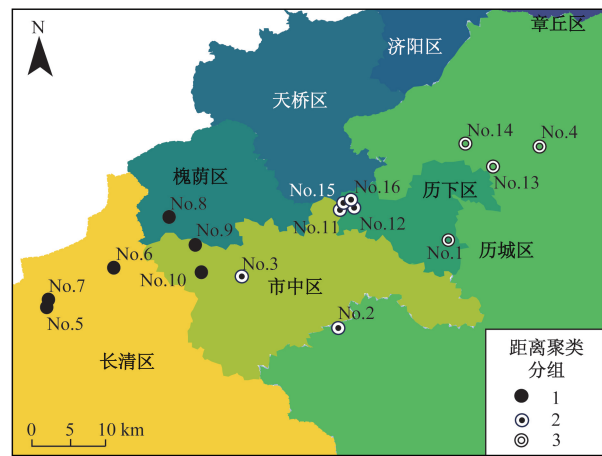


图 5 空间位置距离聚类结果
Fig. 5 Spatial location distance clustering results

1.4 水文地质属性聚类

水文地质属性聚类流程如图 4 所示。水文地质属性聚类与空间位置距离聚类的聚类方法是一样的, 其不同之处在于其属性聚类分为 5 个组, 且水文地质属性的距离表达式变化为

$$h_{\text{ydij}}(w_i, h_{\text{epi}}) = [(g_{di} - g_{dj})^2 + (g_{vi} - g_{vj})^2 + (w_{ydi} - w_{ydj})^2 + (w_{tri} - w_{trj})^2 + (w_{si} - w_{sj})^2 + (w_{fi} - w_{fj})^2]^{\frac{1}{2}} \quad (11)$$

水文地质属性聚类结果如图 6 所示。

1.5 距离-属性混合聚类

基于上述得到的空间距离聚类结果和水文地质属性聚类结果, 采用交叉合并规则进行聚类结果的综合, 得到混合聚类结果。混合聚类规则如下。

规则 1 w_i 和 w_j 属于空间位置距离聚类 D_{C} 中的同一组。

规则 2 w_i 和 w_j 属于水文地质属性聚类 H_{C} 中的同一组。

任意两个观测井 w_i 和 w_j ($i, j = 1, 2, \dots, n_w; i \neq j$), 如不符合规则 1, 则认为观测井 w_i 和 w_j 之间不连通并将其分为不同类; 如符合规则 1 而不符合规则 2, 则认为观测井 w_i 和 w_j 间存在弱连通性并将其分为不同类; 如同时符合规则 1 和规则 2, 则认为观测井 w_i 和 w_j 之间存在强连通性并将其分为同一类。满足规则后得出混合聚类结果 M_C , 同一类的观测井之间为强连通, 不同类之间的观测井为不连通或弱连通。混合聚类结果如图 7 所示。

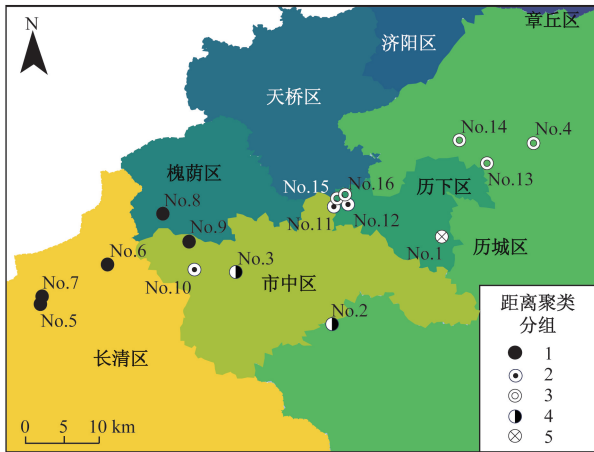


图 6 水文地质属性聚类结果

Fig. 6 Hydrogeological attribute clustering results



图 7 混合聚类结果

Fig. 7 Mixed clustering results

1.6 聚类结果评价

进行聚类结果的效果评价, 分组有效性通过 Calinski-Harabasz 伪 F 统计量 (简称 CH 指标) 来测量, 是反映组内相似性和组间差异性的比率。依据式 (12) ~ 式 (15) 来计算 CH 的指标。

$$CH = \frac{R^2}{\frac{1 - R^2}{n_w - k}} \quad (12)$$

$$R^2 = \frac{SST - SSE}{SST} \quad (13)$$

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} \sum_{v=1}^{n_v} (w_{ji}^v - \bar{w}^v)^2 \quad (14)$$

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} \sum_{v=1}^{n_v} (w_{ji}^v - \bar{w}_j^v)^2 \quad (15)$$

式中: R^2 为决定系数; SST 反映组间差别; SSE 反映组内相似性; n_w 为观测井数目; n_j 为组 j 中的观测井数目; k 为类 (组) 数目; n_v 为用于将观测井进行分组的变量数目; w_{ji}^v 为第 j 组中第 i 个观测井的第 v 个变量值; \bar{w}^v 为所有观测井第 v 个变量值的平均值; \bar{w}_j^v 为第 j 组中第 v 个变量值的平均值。

使用计算完成的 CH 指标值对聚类效果进行评估, 该值越大表示聚类效果越好, 用于评估不同聚类分组数目情况下的聚类效果。经过聚类效果评价, 当距离聚类分为 3 组、属性聚类分为 5 组、综合后的混合聚类为 7 组时, 混合聚类效果最佳, 观测井的最终分组如表 2 所示。

表 2 观测井混合聚类结果

Table 2 Mixed clustering results of observation wells

类 1	类 2	类 3	类 4	类 5	类 6	类 7
No. 5	No. 4	No. 2	No. 15	No. 11	No. 1	No. 10
No. 6	No. 13	No. 3	No. 16	No. 12		
No. 7	No. 14					
No. 8						
No. 9						

在空间位置距离的聚类结果中 (图 5), 观测井被分为三类, 平均轮廓系数值最大。水文地质特征的聚类结果 (图 6), 与图 5 相比有部分差异; 如距离聚类结果中的 11 号、12 号、15 号、16 号, 由于这 4 口观测井的实际距离较近, 感觉上会被归为同一类。但在基于水文地质特征的聚类结果中, 它们的单井产水量和地下水位差异较大, 因此分为两类。

混合聚类方法有效避免了两种错误: 一是避免只考虑井距, 而忽略观测井是否真正连通; 二是避免了仅考虑观测井的水文地质特征时, 聚类结果在空间上没有聚类的现象。

1.7 连通性强弱等级

岩溶含水层连通性的强弱会影响到模型的精度与准确性, 因此需要对连通性进行强弱的划分。基于观测井点状数据构建不规则三角网; 基于混合聚类结果, 观测井之间不连通、弱连通和强连通关系分别使用不同类型的直线符号进行表示, 制作观测井连通性专题图如图 8 所示。

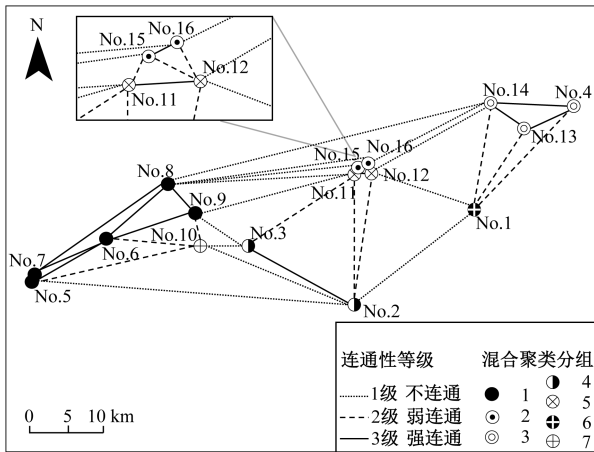


图 8 岩溶含水层观测井连通性专题图
Fig. 8 Thematic map of connectivity of karst aquifer observation wells

2 地下水位预测模型构建

2.1 水位预测指标的选择与确定

地下水位的动态变化受到自然因素和人为因素影响,影响过程如表 3 所示。现有的地下水位预测研究表明,使用人工智能模型来预测地下水位时,地下水位时间序列的过去步长和降水是最常用的变量输入。根据研究区的实际情况和数据可用性,本研究采用套索回归方法,基于回归权重从原始数据中提取重要变量^[20]。

结合岩溶含水层连通性分析结果,考虑到岩溶含水层的水流特征与其空间连通性密切相关,在预测目标井的地下水位时,还需要考虑到与其连通性

表 3 地下水位影响因素

Table 3 Influencing factors of groundwater level

影响因素	指标	影响过程
自然因素	降雨	降水入渗补给使水位上升
	河流水位	地下水动态受地表水的明显影响。河水位上升时,近岸处的水位上升最快,上升幅度最大;远离河岸,水位变化幅度变小,反应时间滞后
	气温	气温影响地下水的蒸发速度,气温越高,蒸发越快
地质条件	蒸散发	地下水受温度等因素蒸散发
	地质条件	地质因素影响地下水位的变化幅度与变化速度。例如,承压含水层受到上覆隔水层的限制,补给区动态变化强烈而迅速,远离补给区则变得微弱而滞后
	人为因素	人为的开采或补给地下水,导致地下水位短时间内急剧下降或上升
人为因素	地表水漫灌	地表水大水漫灌而不加强排水,导致灌入水入渗,水位上升

较好的其他观测井对其影响程度。因此,最终选择的输入变量是地下水位、地下水位时间序列的过去步长、降雨量、降雨量序列的过去步长和同类别井(相互间具有强连通性)的地下水位序列。

2.2 数据集构建

在使用神经网络模型进行预测前,需要将输入的数据集进行转化。预测井的地下水位和相关变量用一维向量表示为

$$w = (G_p, R_p, G_{cat}) \quad (16)$$

式(16)中: G_p 为目标预测井的地下水位; R_p 为降雨量; G_{cat} 为与预测井属于同一类别的其余观测井的平均地下水位值。

运用式(17)将多个时间步长的一维向量组成一个二维矩阵,用于表示一段时间内的输入数据,即作为时间窗口。运用式(18)将该时间窗口从时间序列方向滑动预测井的完整数据序列,可以得到该预测井的多个时间窗数据。

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} G_p^1 & R_p^1 & G_{cat}^1 \\ G_p^2 & R_p^2 & G_{cat}^2 \\ \vdots & \vdots & \vdots \\ G_p^m & R_p^m & G_{cat}^m \end{bmatrix} \quad (17)$$

$$P_k = \begin{bmatrix} W_1^k \\ W_2^k \\ \vdots \\ W_n^k \end{bmatrix} = \begin{bmatrix} w_1^1 & w_2^1 & \cdots & w_m^1 \\ w_1^2 & w_2^2 & \cdots & w_m^2 \\ \vdots & \vdots & & \vdots \\ w_1^n & w_2^n & \cdots & w_m^n \end{bmatrix} \quad (18)$$

式中: W 为时间窗内的数据; m 为时间窗的大小; G_p^m, R_p^m 和 G_{cat}^m 分别为时间窗内时刻 m 的 G_p, R_p 和 G_{cat} ; P_k 为目标预测井 k 的数据; w_n^k 为目标预测井 k 的时间窗 n ; w_m^n 为第 n 个时间窗口中 m 时间处的 w 。

将 16 口观测井的数据整合到一个输入矩阵 I_{PD} 中,进入 ConvLSTM 模型,见式(19)。该网络可以学习区域内观测井的地下水位共性特征,以单一的预测模型实现对区域内多个观测井的实时高效的地下水位预测效果。

$$I_{PD} = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_{16} \end{bmatrix} = \begin{bmatrix} w_1^1 & w_2^1 & \cdots & w_n^1 \\ w_1^2 & w_2^2 & \cdots & w_n^2 \\ \vdots & \vdots & & \vdots \\ w_1^{16} & w_2^{16} & \cdots & w_n^{16} \end{bmatrix} \quad (19)$$

2.3 神经网络预测模型构建

在众多随机性模型中,神经网络模型由于具有大规模并行和分布式处理、自适应、自组织和自学习功能,以及良好的容错性和强大的自我调节能力等特点,被众多学者用以进行预测工作。并且人工

神经网络有着适应于水文资源系统的独特优越性；神经网络可在事先不知道数据中潜在规律性的情况下从提供的实例中学习、总结数据中内在的规律性，并且调整自身行为以适应新信息或环境，从而具有从已有状态向新的环境状态进化的概括能力。

2.3.1 LSTM 模型

神经网络模型中，LSTM 适用于处理和预测时间序列中间隔和延迟非常长的重要事件，并且解决了 RNN 存在的梯度消失和梯度爆炸问题。因此选用 LSTM 作为本实验的基本预测模型。

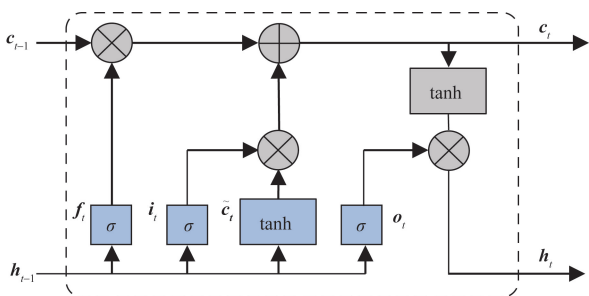
对于给定序列 $x = (x_1, x_2, \dots, x_n)$ ，应用一个标准的 RNN 模型，可以通过迭代式(20)和式(21)计算出一个隐藏层序列 $h = (h_1, h_2, \dots, h_n)$ 和一个输出序列 $y = (y_1, y_2, \dots, y_n)$ 。

$$h_t = f_a(W_{xh}x_t + W_{hh}h_{t-1} + b_n) \quad (20)$$

$$y_t = W_{hy}h_t + b_y \quad (21)$$

式中： h_t 为 t 时刻的隐藏状态向量； x_t 为 t 时刻的输入； y_t 为 t 时刻的输出； W_{xh} 、 W_{hy} 、 W_{hh} 为权重系数矩阵； b_n 、 b_y 为偏置向量； f_a 为激活函数；下标 t 为时刻。

LSTM 模型是将隐藏层的 RNN 细胞替换为 LSTM 细胞，使其具有长期记忆能力。经过不断地演化，目前应用最为广泛的 LSTM 模型细胞结构如图 9 所示。



f_t 为遗忘门； i_t 为输入门； \tilde{c}_t 为候选记忆细胞； o_t 为输出门； σ 为非线性激活函数； \tanh 为激活函数； c_{t-1} 为长时记忆输入； c_t 为长时记忆输出； x_t 为输入； h_{t-1} 为隐藏层输入； h_t 为输出

图 9 LSTM 内部结构图

Fig. 9 Diagram of the internal structure of the LSTM

2.3.2 ConvLSTM 模型

虽然 LSTM 网络可以有效提取时间序列的时间特征，但网络无法捕捉数据的空间特征。ConvLSTM 模型包括卷积神经网络和 LSTM 网络，其不仅具有 LSTM 的时间建模能力，而且还可以描绘局部特征。该模型通过多维数据中进行卷积运算来捕捉基本的空间特征，并用 LSTM 单元中每个门的卷积运算代替矩阵乘法步骤，提取数据的时空特征。本研

究设计的 ConvLSTM 模型的内部结构如图 10 所示，其既具备 LSTM 网络的时间建模能力，同时具备 CNN 描绘数据的空间特征能力。

ConvLSTM 计算过程见式(22)~式(26)。

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} C_{t-1} + b_i) \quad (22)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} C_{t-1} + b_f) \quad (23)$$

$$C_t = f_t \odot C_{t-1} + i_t \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (24)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot C_t + b_o) \quad (25)$$

$$H_t = o_t \odot \tanh(C_t) \quad (26)$$

式中： i_t 、 f_t 、 C_t 、 o_t 分别为 LSTM 结构中的输入门、遗忘门、控制单元和输出门； σ 为非线性激活函数； X_t 为 t 时刻的输入； H_t 为 t 时刻的输出； W_{xi} 、 W_{hi} 、 W_{ci} 、 W_{xf} 、 W_{hf} 、 W_{cf} 、 W_{xc} 、 W_{hc} 、 W_{xo} 、 W_{ho} 、 W_{co} 为权重矩阵参数； \odot 为 Hadamard 积符号； $*$ 为卷积运算符号； b_i 、 b_o 、 b_f 、 b_c 为偏置向量； \tanh 为激活函数。

本项目使用的 ConvLSTM 模型如图 11 所示。它由两个卷积层和两个 LSTM 层组成。

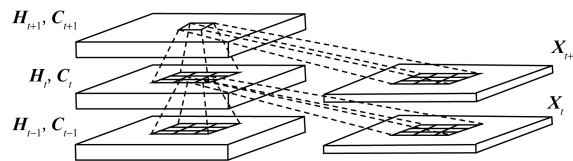
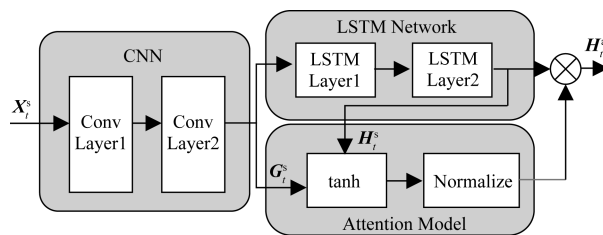


图 10 ConvLSTM 内部结构

Fig. 10 ConvLSTM internal structure



X_t^s 为空间信息输入数据； G_t^s 为空间信息候选细胞； H_t^s 为空间信息输出； H_t^f 为注意力加权后的输出；CNN 为卷积层；LSTM Network 为长短期记忆网络；Attention Model 为注意力机制

图 11 ConvLSTM 模型示意图

Fig. 11 Schematic diagram of the ConvLSTM model

3 模型模拟与精度验证

3.1 实验设计

选取山东省济南市马山断裂与港沟断裂之间的碳酸盐岩岩溶含水层作为研究区。数据包括该区域内 2009—2012 年布设的 16 口观测井及其长期地下水位监测序列、30 m 分辨率数字高程模型 (digital elevation model, DEM) 数据。将地下水位、

地下水位时间序列的过去步长、降雨量、降雨量序列的过去步长和同类别井(相互间具有强连通性)的地下水位序列作为输入变量。在输入数据集中,将2009年1月—2011年12月阶段划分为训练集,2012年1—12月为预测集。并且在每批次训练中,将训练集的后20%划分为验证集用于数据验证和调整参数,数据集划分如表4所示。

为了验证考虑基于混合聚类结果的连通性特征的 ConvLSTM 神经网络模型 M-MV-ConvLSTM 在预测地下水位时的有效性,还将与单变量 SV-LSTM 模型、多变量 MV-LSTM 模型、仅考虑空间距离的多变量 D-MV-ConvLSTM 模型进行对比分析。

3.2 模型参数

模型参数通过大量案例调整情况下给出。隐藏节点的数量、卷积层的核大小和学习率共包含45种组合,因此在验证阶段通过试错法确定一个最优组合。每个模型的最佳参数组合如表5所示。

3.3 结果分析

对于 SV-LSTM、MV-LSTM、D-MV-ConvLSTM 和 M-MV-ConvLSTM 模型,平均均方根误差值 (root mean square error, RMSE) 分别为 1.38、0.75、0.56 和 0.46。实验结果表明,考虑降雨因素 MV-LSTM 模型在每口井的表现均优于 SV-LSTM 模型。考虑

了 GWL_d 的 D-MV-ConvLSTM 模型比 SV-LSTM 和 MV-LSTM 模型表现更好。此外, M-MV-ConvLSTM 模型考虑了观测井之间的连通性,其精度得到进一步提高。表6列出了4个地下水位预测结果的 RMSE 和纳什效率系数 (Nash-Sutcliffe efficiency, NSE)。图12为预测结果的 RMSE 直方图。

从表6中 NSE 可以看出, D-MV-ConvLSTM 和 M-MV-ConvLSTM 模型的 NSE 与 SV-LSTM 模型和 MV-LSTM 相比都有一定程度的改进。与 D-MV-ConvLSTM 相比, M-MV-ConvLSTM 模型的 NSE 值进一步提高。使用该模型,在所有井中,50%的井的 NSE 大于 0.9,80% 井的 NSE 大于 0.85,说明同时考虑了气象因素和混合聚类结果连通性强弱的 M-MV-ConvLSTM 模型预测效果最好,可信度高。

表5 4个模型的输入变量和模型参数
Table 5 Input variables and model parameters of the four models

模型	输入变量	卷积核大小	隐藏层神经元个数	学习率	验证误差
SV-LSTM	GWL, GWL_{t-1}	—	8	0.003	0.023 6
MV-LSTM	GWL, R, GWL_{t-1}, R_{t-1}	—	6	0.003	0.013 8
D-MV-ConvLSTM	GWL, $R, GWL_{t-1}, R_{t-1}, GWL_d$	5	7	0.003	0.011 1
M-MV-ConvLSTM	GWL, $R, GWL_{t-1}, R_{t-1}, GWL_c$	5	5	0.005	0.010 0

注:GWL (groundwater level) 为地下水位序列; GWL_{t-1} 为地下水位时间序列的过去步长; R 为降雨量; R_{t-1} 为降雨时间序列过去步长; GWL_d 为考虑了空间位置距离聚类结果中同类别井的地下水位序列; GWL_c 为考虑了混合聚类结果中同类别井的地下水位序列。

表4 模型数据集划分

Table 4 Model data set partition

数据集类型	数量	比例
训练集	3 456	75%
验证集	每批次中训练集的 20%	
预测集	1 152	25%

表6 4种模型中16口井得到的地下水位预测结果的 RMSE 和 NSE

Table 6 RMSE and NSE of groundwater level prediction results obtained from 16 wells in the four models

观测井编号	SV-LSTM		MV-LSTM		D-MV-ConvLSTM		M-MV-ConvLSTM	
	RMSE	NSE	RMSE	NSE	RMSE	NSE	RMSE	NSE
No. 1	9.192	-0.146	4.836	0.677	3.518	0.832	2.954	0.881
No. 2	5.012	-24.152	2.339	-1.063	0.770	0.406	0.935	0.124
No. 3	0.507	0.854	0.259	0.895	0.373	0.921	0.257	0.962
No. 4	1.129	-0.212	0.915	0.056	0.916	0.202	0.960	0.122
No. 5	0.342	0.807	0.313	0.768	0.315	0.836	0.261	0.888
No. 6	0.285	0.845	0.269	0.784	0.263	0.867	0.217	0.910
No. 7	0.367	0.766	0.284	0.774	0.295	0.848	0.221	0.916
No. 8	0.262	0.746	0.208	0.686	0.216	0.827	0.174	0.888
No. 9	0.387	0.927	0.218	0.939	0.214	0.978	0.112	0.994
No. 10	0.550	0.798	0.401	0.837	0.456	0.861	0.358	0.914
No. 11	0.341	0.075	0.181	0.502	0.155	0.809	0.088	0.939
No. 12	0.346	0.144	0.187	0.545	0.162	0.813	0.094	0.937
No. 13	0.707	0.670	0.556	0.813	0.567	0.788	0.428	0.879
No. 14	0.786	-0.765	0.283	0.508	0.218	0.864	0.155	0.931
No. 15	1.055	-498.447	0.504	-8.943	0.274	-32.608	0.032	0.540
No. 16	0.792	-1 261.636	0.176	-331.881	0.230	-105.330	0.068	-8.361
平均值	1.379	-111.171	0.746	-20.819	0.559	-7.943	0.457	0.216

图 13 为选定的部分井使用 4 种模型得到的预测结果。预测数据为 2012 年 1—12 月的地下水位

观测数据,每月采样 6 次,采样间隔为 5 d。可以看出,4 个模型的预测值和实际观测值的趋势是一致的

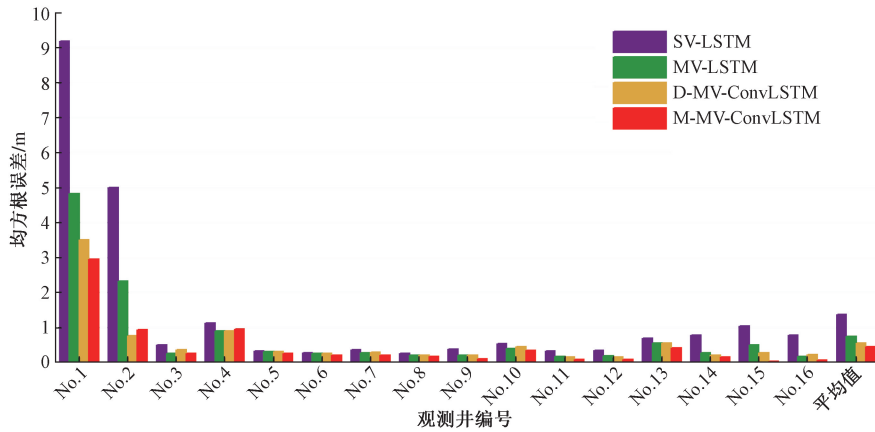


图 12 4 种模型中 16 口井得到的地下水位预测结果的 RMSE

Fig. 12 RMSE of the predicted groundwater level results obtained from 16 wells in four models

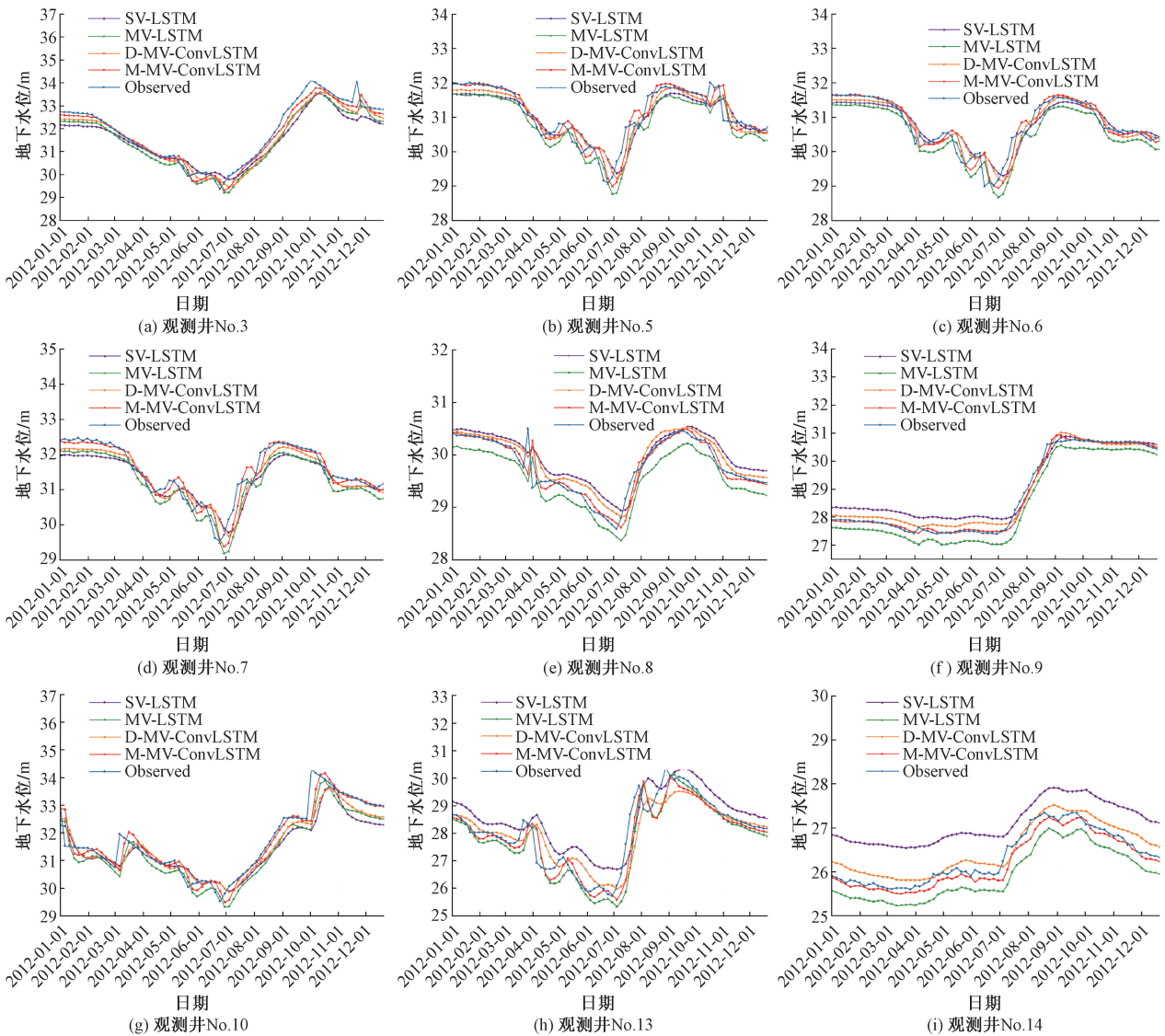


图 13 4 个模型中部分井的模拟预测结果

Fig. 13 Simulation results of some of the wells in the four models

的。由于 M-MV-ConvLSTM 模型不仅考虑了目标观测井本身的地下水位数据和气象数据,还考虑了相关观测井的地下水位数据。因此 M-MV-ConvLSTM 模型的预测结果比其他 3 个模型的预测结果更符合实际观察结果。

与单一结构的 LSTM 模型相比,ConvLSTM 模型综合分析了数据的时空特征,对预测模型的时滞问题有一定的改善。从图 13 所示的各井预测结果可以看出,M-MV-ConvLSTM 模型大大改善了预测滞后问题。考虑到预测精度和滞后问题的改善,所提出的考虑观测井连通性的 M-MV-ConvLSTM 预测模型对地下水位的预测具有较好的适用性。

4 结论

针对地下水位预测考虑因素不全面的问题,提出一种混合聚类方法,其考虑了观测井间连通性因素。并构建 ConvLSTM 地下水预测模型,与传统的 LSTM 预测模型对比,得到以下结论。

(1) 根据观测井间的距离和含水层水文地质特征,提出了基于距离-属性的混合聚类方法,综合分析了岩溶含水层的连通性,并在岩溶含水层地下水位预测中有效证明观测井间连通性的强弱会对预测结果产生明显差异,体现了在预测过程中考虑连通性因素的重要性。

(2) 把 LSTM 模型与 CNN 模型相结合,提出了 ConvLSTM 模型,使其不仅具有 LSTM 的时间建模能力,而且还可以描绘局部特征,相比传统的 LSTM 预测模型而言其精确度与可靠性得到了进一步提高。

(3) 验证连通性因素在地下水预测中的有效性,考虑基于混合聚类结果的连通性特征的 M-MV-ConvLSTM 模型与单变量 SV-LSTM 模型、多变量 MV-LSTM 模型和仅考虑空间距离模型 D-MV-ConvLSTM 对比分析。分析得出平均 RMSE 为 0.457, NSE 为 0.216,结果表明,考虑观测井间连通性特征的 M-MV-ConvLSTM 模型具有更高的预测精度,体现了连通性因素和 M-MV-ConvLSTM 模型的有效性、创新性。

参 考 文 献

- [1] 张召悦,张红波. 基于 ARMA-AE-LSTM 模型的进场交通流预测方法[J]. 科学技术与工程, 2024, 24(27): 11919-11927.
Zhang Zhaoyue, Zhang Hongbo. Approach traffic flow prediction method based on ARMA-AE-LSTM Model[J]. Science Technology and Engineering, 2024, 24(27): 11919-11927.
- [2] 孙虹洁,赵振华,黄林显,等. 多变量 LSTM 神经网络模型在地下水位预测中的应用[J]. 人民黄河, 2022, 44(8): 69-75.
Sun Hongjie, Zhao Zhenhua, Huang Linxian, et al. Application of multivariate LSTM neural network model in groundwater level prediction[J]. Yellow River, 2022, 44(8): 69-75.
- [3] 张朝逢,陈皓锐,岳中奇. 基于长短时记忆网络(LSTM)的地下水埋深模拟预测——以关中平原为例的实例分析[J]. 中国农村水利水电, 2020(9): 127-131, 137.
Zhang Chaofeng, Chen Haorui, Yue Zhongqi. Simulation and prediction of groundwater depth based on long short-term memory network (LSTM): a case study of Guanzhong Plain[J]. China Rural Water and Hydropower, 2020(9): 127-131, 137.
- [4] 陈守开,蒋海峰,李海瑞,等. 基于 RST-LSTM 模型的堤坝渗压动态预测分析[J]. 应用基础与工程科学学报, 2021, 29(1): 112-122.
Chen Shoukai, Jiang Haifeng, Li Hairui, et al. Dynamic prediction analysis of dam seepage pressure based on RST-LSTM model[J]. Journal of Basic Science and Engineering, 2021, 29(1): 112-122.
- [5] 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.
- [6] 陈聪,候磊,李乐乐,等. 基于 GRU 改进 RNN 神经网络的飞机燃油流量预测[J]. 科学技术与工程, 2021, 21(27): 11663-11673.
Chen Cong, Hou Lei, Li Lele, et al. Prediction of aircraft fuel flow based on recurrent neural network[J]. Science Technology and Engineering, 2021, 21(27): 11663-11673.
- [7] 闫佰忠,孙剑,王昕洲,等. 基于多变量 LSTM 神经网络的地下水位预测[J]. 吉林大学学报(地球科学版), 2020, 50(1): 208-216.
Yan Baizhong, Sun Jian, Wang Xinzhou, et al. Groundwater level prediction based on multivariate LSTM neural network[J]. Journal of Jilin University (Earth Science Edition), 2020, 50(1): 208-216.
- [8] 汪云,杨海博,徐建,等. 基于长短时记忆神经网络模型的地下水位预测研究[J]. 节水灌溉, 2019(10): 73-77.
Wang Yun, Yang Haibo, Xu Jian, et al. Research on groundwater level prediction based on long short-term memory neural network model[J]. Water Saving Irrigation, 2019(10): 73-77.
- [9] 冯希尧,苟俊程,刘瑞,等. 基于遗传算法优化机器学习模型的地下水潜在性预测[J]. 科学技术与工程, 2024, 24(19): 7988-7998.
Feng Xiyao, Gou Juncheng, Liu Rui, et al. Prediction of groundwater potential based on genetic algorithm optimized machine learning model[J]. Science Technology and Engineering, 2024, 24(19): 7988-7998.
- [10] 胡飞跃,王忠忠. 基于降雨量的 SWA-LSTM 深度学习模型在地下水位预测中的应用研究[J]. 地下水, 2022, 44(6): 70-72, 113.
Hu Feiyue, Wang Zhongzhong. The application of SWA-LSTM deep learning model based on rainfall in groundwater level prediction[J]. Ground Water, 2022, 44(6): 70-72, 113.
- [11] 侯金霄,黄林显,胡晓农,等. 基于 EMD-LSTM 耦合模型的趵突泉岩溶地下水水位预测应用[J]. 水资源与水工程学报, 2023, 34(4): 92-98.
Hou Jinxiao, Huang Linxian, Hu Xiaonong, et al. Application of

- karst groundwater level prediction in Baotu Spring based on EMD-LSTM coupling model[J]. *Journal of Water Resources and Water Engineering*, 2023, 34(4): 92-98.
- [12] 郭艺, 秦大军, 王枫, 等. 基于时间序列分析法的岩溶泉水位预测[J]. *中国岩溶*, 2021, 40(4): 689-697.
Guo Yi, Qin Dajun, Wang Feng, et al. Prediction of karst spring water level based on time series analysis method[J]. *Carsologica Sinica*, 2021, 40(4): 689-697.
- [13] 冯鹏宇, 金韬, 沈一选, 等. 基于 CNN-Transformer 的城区地下水位预测[J]. *计算机仿真*, 2023, 40(4): 492-498.
Feng Pengyu, Jin Tao, Shen Yixuan, et al. Prediction of urban groundwater level based on CNN-Transformer[J]. *Computer Simulation*, 2023, 40(4): 492-498.
- [14] 张研, 郭道静, 张树光, 等. 岩溶区灰岩溶蚀程度卷积神经网络识别及敏感性分析[J]. *应用基础与工程科学学报*, 2023, 31(4): 961-976.
Zhang Yan, Guo Daojing, Zhang Shuguang, et al. Convolutional neural network identification and sensitivity analysis of limestone dissolution degree in karst area[J]. *Journal of Basic Science and Engineering*, 2023, 31(4): 961-976.
- [15] 任青阳, 王彦丁, 施俭. 卷积神经网络目标检测算法研究进展[J]. *科学技术与工程*, 2024, 24(32): 13665-13677.
Ren Qingyang, Wang Yanding, Shi Jian. Advances in target detection algorithms for convolutional neural networks[J]. *Science Technology and Engineering*, 2024, 24(32): 13665-13677.
- [16] 郭靖. 对 K-means 聚类算法欧氏距离加权系数的研究[J]. *网络安全技术与应用*, 2016(10): 74-75.
Guo Jing. Research on the Euclidean distance weighted coefficient of K-means clustering algorithm[J]. *Network Security Technology & Application*, 2016(10): 74-75.
- [17] 金福喜, 袁权威, 秦帅帅. 基于 GIS 和 K-means 聚类方法的滑坡气象预警模型[J]. *金属矿山*, 2018(6): 162-166.
Jin Fuxi, Yuan Quanwei, Qin Shuaishuai. Landslide meteorological early warning model based on GIS and K-means clustering method[J]. *Metal Mine*, 2018(6): 162-166.
- [18] Yuan C H, Yang H T. Research on K-value selection method of K-means clustering algorithm [J]. *Multidisciplinary Scientific Journal*, 2019, 2(2): 226-235.
- [19] 程亚南, 王晓峰, 刘淞佐, 等. 求解多起点多旅行商问题的 K-means 聚类信息传播算法[J]. *科学技术与工程*, 2022, 22(23): 10146-10154.
Cheng Yanan, Wang Xiaofeng, Liu Songzuo, et al. K-means clustering information propagation algorithm for multiple depots multiple traveling salesman problem[J]. *Science Technology and Engineering*, 2022, 22(23): 10146-10154.
- [20] 蒋溢, 伍书平, 胡昆, 等. 基于 Lasso 和构造性覆盖算法的不均衡数据分类方法[J]. *计算机应用*, 2023, 43(4): 1086-1093.
Jiang Yi, Wu Shuping, Hu Kun, et al. Imbalanced data classification method based on Lasso and constructive covering algorithm [J]. *Journal of Computer Applications*, 2023, 43(4): 1086-1093.