DOI:10.20040/j.cnki.1000-7709.2023.20221334

基于EMD分解的黄河流域兰州水文站日径流预测

路 炜[a],魏霖静[b]

(甘肃农业大学 a. 理学院; b. 信息科学技术学院,甘肃 兰州 730030)

摘要:为探索提高径流预测精度,基于兰州水文站2001年8月~2019年12月的逐日径流数据,在控制变量法的基础上,应用LSTM、ARIMA、SVR、XGBoost四种模型,建立了单一模型、EMD分解重构、剔除噪声模态分量后的EMD分解重构等三类处理方式共12种模型方案,并对12种方案的评价指标进行对比。结果表明,EMD序列分解重构技术和基于Hurst指数的噪声模态分量剔除有助于提升预测精度,与单一模型相比,前者构建的模型的均方根误差(R_{RMSE})平均下降了15.16%,后者平均下降了28.49%;12种方案中,预测效果较好的方案是剔除噪声模态分量后的"EMD-SVR-ARIMA"模型。

关键词:日径流预测;EMD分解;兰州水文站;机器学习模型

中图分类号:P338$^+$.1;TV121$^+$.4 **文献标志码**:A **文章编号**:1000-7709(2023)08-0019-04

1 引言

河流径流量预测对于防洪减灾、水利利用具有重要意义。近年来,随着计算机性能的提升和机器学习、深度学习的快速发展,开始使用非线性模型预测河流径流量,如树模型[1,2]、支持向量机(SVM)[3]、长短期记忆神经网络(LSTM)[4]。同时,由于对径流预测精度要求越来越高,小波分解、经验模态分解(EMD)、集合经验模态分解(EEMD)等分解方法受到关注,"分解-预测-重构"模式作为一种新的、有效的预测思路被广泛研究[5,6],但大多数研究在使用EMD等分解方法时缺乏控制变量法的对比试验,且鲜有研究在分解重构过程中尝试剔除噪音分量。黄河是中国第二长河,其上中游水资源短缺较严重。其中,黄河上中游的兰州水文站控制流域面积22.26×10^4 km^2,占黄河总面积的29.60%,平均月流量为989 m^3/s,9、2月多年平均流量分别为1 665、420 m^3/s。对此,本文以兰州水文站日径流量序列为例,应用LSTM、ARIMA、SVR、XGBoost四种模型,构建单一模型、EMD分解预测模型、EMD分解+剔除噪声模态分量的预测模型并进行对比,探索"分解—预测—重构"方法和剔除噪声模态分量对模型精度提升的影响,并确定了较好的径流预测方法,为流域水资源开发利用和洪水预测提供了科学依据和解决方法。

2 研究方法

2.1 LSTM

LSTM是一种特殊的RNN网络,适合于处理和预测时间序列中间隔和延迟非常长的重要事件[7],其内部结构见图1。图1中,F_t为遗忘门;I_t为输入门;O_t为输出门;C_{t-1}、C_t分别为单元的上一步、当前状态;$\boldsymbol{H}_{t-1}$、$\boldsymbol{X}_t$均为输入向量;$\boldsymbol{H}_t$为输出向量;σ、tanh均为激活函数。遗忘门决定

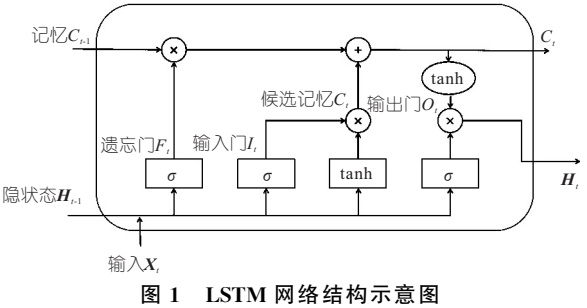

图1 LSTM网络结构示意图

Fig. 1 Schematic diagram of LSTM network structure neuron

收稿日期:2022-06-29,**修回日期**:2022-10-19

基金项目:2020年甘肃农业大学研究生教育研究项目(2020-19);2021年度兰州市人才创新创业项目(2021-RC-47);2021年教育部产学研合作协同育人项目(202102326036)

作者简介:路炜(1998-),女,硕士研究生,研究方向为机器学习,E-mail:tongji_luwei@163.com

通讯作者:魏霖静(1977-),女,教授,研究方向为智能计算、农业信息化,E-mail:wlj@gsau.edu.cn

前一时刻的状态中有多少信息被遗忘；输入门决定是否忽略掉该时刻输入的信息；输出门决定当前输出是否使用前一时刻隐状态中的信息。LSTM 神经网络通过将序列输入后作用于遗忘门、输入门和输出门达到记忆长期状态的目的。

2.2 EMD 分解

经验模态分解(EMD)算法主要参考文献[6]。

2.3 Hurst 指数

R/S 分析法作为计算 Hurst 指数最常用的方法，通常被用来分析时间序列的分形特征和长期记忆过程。R/S 分析研究表明，降水、温度、树木年轮、冰川纹泥、太阳黑子等自然现象均具有 Hurst 效应[8]。

Hurst 指数(H)取值范围在 0～1 之间，根据不同取值分为 3 种形式分别代表了不同的时间序列相关性。其中，$H<0.5$ 表示该时间序列负相关；$H>0.5$ 表示该时间序列具有长期相关特征，越接近于 1，相关性越强；$H=0.5$ 表示该时间序列不相关，为随机序列。

使用 R/S 分析方法计算 Hurst 指数的基本原理为对于任意正整数 $\tau \geqslant 1$，长度为 τ 时间序列的平均值 $\bar{\varepsilon}$ 为：

$$\bar{\varepsilon}=\frac{1}{\tau}\sum_{t=1}^{\tau}\varepsilon(t) \tag{1}$$

式中，$\varepsilon(t)$ 为 t 时刻时间序列值。

用 $X(t)$ 表示累计离差为：

$$X(t,\tau)=\sum_{u=1}^{t}[\varepsilon(u)-\bar{\varepsilon}] \tag{2}$$

式中，$\varepsilon(u)$ 为 u 时刻时间序列值。

将同一个 τ 值所对应的最大值 $X(t)$ 与最小值 $X(t)$ 之差称为极差 $R(\tau)$，并记为：

$$R(\tau)=\max_{1\leqslant t\leqslant\tau}X(t,\tau)-\min_{1\leqslant t\leqslant\tau}X(t,\tau) \tag{3}$$

Hurst 利用的标准偏差 $S(\tau)$ 为：

$$S(\tau)=\left\{\frac{1}{\tau}\sum_{t=1}^{\tau}[\varepsilon(\tau)-\bar{\varepsilon}]^2\right\}^{1/2} \tag{4}$$

式中，t_1、t_2、t_3、…、t_n 为时间点；ε_1、ε_2、ε_3、…、ε_n 为 t_1、t_2、t_3、…、t_n 处取得的响应时间序列。

对于时间序列存在如下关系：

$$R/S=(\tau/2)^H \tag{5}$$

3 基于 EMD 分解的黄河流域兰州水文站日径流预测

3.1 研究数据

数据源于国家地球系统科学数据中心(http://www.geodata.cn)，数据集包括兰州水文站 2001 年 8 月 10 日～2019 年 12 月 31 日的逐日径流量，共 6 716 条数据，见图 2。

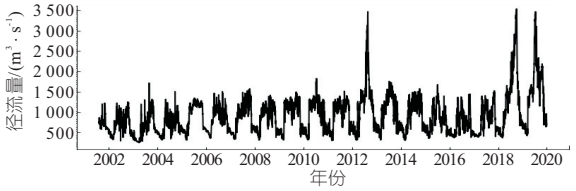

图 2 兰州水文站逐日径流量

Fig. 2 Daily runoff of Lanzhou Hydrometric Station

3.2 评价指标

选择均方根误差(R_{RMSE})和纳什效率系数(N_{NSE})两种模型评价指标，其中均方根误差用来衡量预测值与真实值之间的偏差，R_{RMSE} 越小表示定量误差越小；纳什效率系数一般用来验证水文模型的好坏，其取值为负无穷到 1，N_{NSE} 越接近 1 表明径流预测值与观测值的过程吻合性越好[7]。

3.3 模型建立

3.3.1 EMD 分解

EMD 分解适合于非线性、非平稳时间序列的处理，能使复杂信号分解为有限个本征模函数 IMF，所分解出来的各 IMF 分量包含了原信号的不同时间尺度的局部特征。由于兰州水文站逐日径流量数据属于非平稳、非线性的时间序列，因此选择 EMD 分解对数据进行处理，使用 matlab 软件分解后得到 11 个本征模态分量(IMF)和 1 个残差序列(RES)，结果见图 3。

3.3.2 高低频分量定义

由于分解出的 IMF 数量较多，对每个 IMF 进行建模再重构较复杂，且易过拟合。所以将 IMF 重组为高频分量和低频分量，再分别对高频分量、低频分量及剩余的残差序列(RES)预测后进行重构。使用 R/S 分析法计算每个 IMF 的 Hurst 指数，结果见表 1。将最低的 Hurst 指数值作为噪声模态分量的分界点、0.5 作为高频分量和低频分量的分界点[9]，即在本文日径流序列分解出的 IMF 中，噪声模态分量为 IMF1，高频分量为 IMF1、IMF2、IMF3、IMF4、IMF5、IMF6、IMF7、IMF8，低频分量为 IMF9、IMF10、IMF11。重组后的高频分量、低频分量和 RES 见图 4。

3.3.3 无控制变量的预测模型

选择 XGBoost、SVR、ARIMA(AR)、LSTM 四个模型进行径流预测，其中 ARIMA 是时间序列中最基本的模型，AR 模型适合预测平稳时间序列；SVR 为机器学习模型中的有监督学习模型，具有鲁棒性的优势；XGBoost 属于集成学习模型，泛化性较好；LSTM 为深度学习模型，适合于时间序列预测。

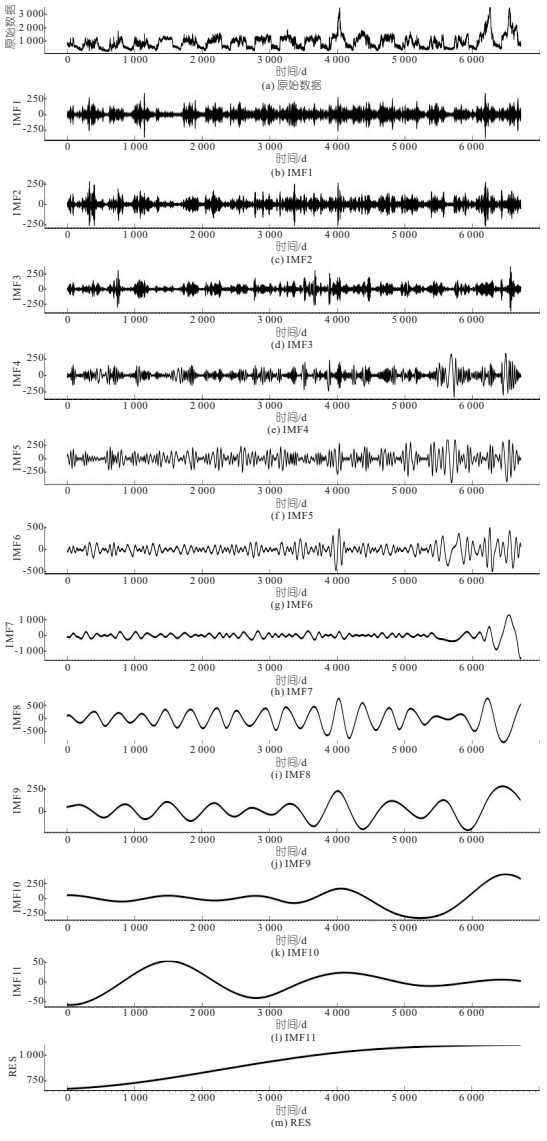

图 3　兰州水文站逐日径流量 EMD 分解结果

Fig. 3　EMD decomposition results of daily runoff of Lanzhou Hydrometric Station

表 1　各 IMF 分量对应的 Hurst 指数值

Tab. 1　Hurst index value corresponding to each IMF component

IMF	Hurst 指数	IMF	Hurst 指数
IMF1	0.132 4	IMF7	0.306 1
IMF2	0.139 3	IMF8	0.383 1
IMF3	0.211 5	IMF9	0.542 4
IMF4	0.212 7	IMF10	0.810 6
IMF5	0.245 7	IMF11	0.898 3
IMF6	0.399 8		

利用四个单一模型分别拟合高频分量、低频分量、残差序列 RES,结果见表 2。由表 2 可知,低频分量和残差序列拟合后的 R_{RMSE} 较小;对高频分量进行拟合的 R_{RMSE} 较大,远大于低频分量和残差序列。由于对高频分量拟合误差较大会导致重构后的误差也较大,可推断高频分量的预测误差是决定最终误差大小的主要来源。

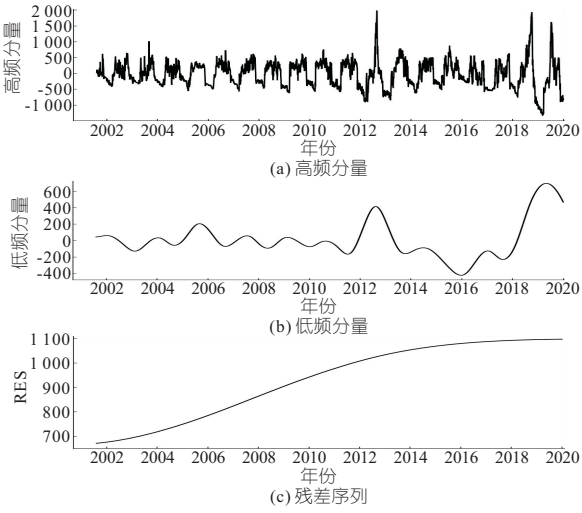

图 4　高频分量、低频分量、残差序列示意图

Fig. 4　Schematic diagram of high frequency component, low frequency component and residual sequence

表 2　无控制变量的预测模型误差

Tab. 2　Prediction model results without control variables

单一模型	高频分量	低频分量	残差序列
XGBoost	111.229	0.061	0.052
SVR	56.150	0.059	0.040
LSTM	143.562	1.230	0.990
ARIMA(AR)	56.607	0.069	0.056

3.3.4　控制变量的预测模型

在表 2 的基础上,基于控制变量的思路设计试验。首先对原始序列分别应用四个单一模型进行预测,计算其预测结果的评价指标,得到单一模型预测结果(表 3)。为了检验“分解－预测－重构”思路对预测精度提升的影响,分别对分解重组后的各分量进行预测并重构。由于高频分量的预测误差是决定最终误差大小的主要来源,因此对低频分量和残差序列(RES)均使用支持向量机进行拟合,然后应用四个单一模型分别对高频分量进行拟合重构,计算重构后的预测精度指标,得到 EMD 组合模型预测结果(表 3)。最后为了检验剔除噪声模态分量后预测精度有无提升,在尝试对以上不同预测模型剔除噪声模态分量(IMF1)后再次进行预测与重构,计算重构后的预测精度指标,得到剔除噪声模态分量的 EMD 组合模型预测结果(表 3)。

表 3　不同径流预测方案及其对应的预测精度指标

Tab. 3　Different runoff prediction schemes and their corresponding prediction accuracy indexes

模型组合	方案	模型	R_{RMSE}	N_{NSE}
单一模型	1	XGBoost	209.24	0.917 2
	2	SVR	186.13	0.934 4
	3	LSTM	200.92	0.912 7
	4	ARIMA(AR)	170.57	0.940 0
EMD 组合模型	5	resSVR+低频 SVR+高频 XGBoost	187.47	0.924 0
	6	resSVR+低频 SVR+高频 SVR	147.06	0.953 2
	7	resSVR+低频 SVR+高频 LSTM	198.45	0.998 9
	8	resSVR+低频 SVR+高频 ARIMA(AR)	122.78	0.967 4
剔除噪声模态分量的 EMD 组合模型	9	在方案 5 中剔除 IMF1 分量	160.07	0.944 7
	10	在方案 6 中剔除 IMF1 分量	131.38	0.962 6
	11	在方案 7 中剔除 IMF1 分量	175.29	0.999 7
	12	在方案 8 中剔除 IMF1 分量	88.21	0.983 1

对于调参问题，SVR、XGBoost、ARIMA 模型均通过网格搜索进行调参，找到最适合的超参数，以保证模型的效果最好。由于 LSTM 模型较复杂，因此使用手动调参，对不同超参数组合进行尝试，结合预测精度指标，最终确定的网络结构为一个 LSTM 层加上一个全连接层，LSTM 层神经网络节点数为 16 个，学习率为 1×10^{-4}，模型进行完整训练的最大次数为 1 000。为了防止过拟合，在其中加入了 drop out 层，drop out 的比率设置为 0.3。

将 2001～2015 年的数据划分为训练集，2016～2019 年的数据划分为测试集。在 SVR、XGBoost、LSTM 模型中设置的输入为前一天的径流量，预测值为第二天的径流量。表 3 中的预测精度指标均为测试集上的模型结果。

3.4 结果与分析

由表 3 可看出：①对比单一模型与 EMD 组合模型，发现将径流分解重组后分别预测不同频率的 IMF 和残差序列 RES 有助于提升预测精度。单一模型的 R_{RMSE} 分别为 209.24、186.13、200.92、170.57。EMD 组合模型的 R_{RMSE} 分别为 187.47、147.06、198.45、122.78，误差分别下降了 10.4%、20.99%、1.23%、28.02%，平均下降了 15.16%。单一模型对应的纳什效率系数 N_{NSE} 分别从 0.917 2、0.934 4、0.912 7、0.940 0 上升到了 0.924 0、0.953 2、0.998 9、0.967 4。②对比剔除噪声模态分量的 EMD 组合模型与 EMD 组合模型及单一模型，发现径流序列分解重构过程中，在剔除了噪声模态分量 IMF1 后，预测精度有了显著提升。剔除 IMF1 后的模型与单一模型相比，R_{RMSE} 平均下降了 28.49%，与 EMD 模型分解后的预测结果相比，R_{RMSE} 平均下降了 20.36%。除了 R_{RMSE} 外，N_{NSE} 也有一定程度的提升。③剔除噪声模态分量后再进行重构的方案 12 为所有方案中误差最小的，R_{RMSE} 仅有 88.21，远低于其他方案，N_{NSE} 为 0.983 1，相对其他方案也较高。图 5 为方案 12 在测试集上的预测结果。由图 5 可看出，方案 12 整体的拟合结果极好，真实值和预测值基本对应。

图 5 方案 12 预测结果示意图

Fig. 5 Schematic diagram of prediction results of scheme 12

在单一模型、组合模型、剔除噪声模态分量的组合模型中，ARIMA(AR)的模型结果均为最好，这证明 ARIMA 模型十分适合于预测径流时间序列。其次是支持向量回归模型，在对剔除噪声模态分量后的模型进行预测重构时，支持向量回归的 R_{RMSE} 降至 131.38。排名第三的是 XGBoost 模型，预测效果较为一般，主要原因可能在于输入的预测变量维度较少及数据量较小，从而不能发挥出 XGBoost 模型的优势。LSTM 的误差是四种模型中最高的，但使用了 EMD 进行分解和剔除掉噪声模态分量 IMF1 后，模型的纳什效率系数 N_{NSE} 均达 0.99，这表明模型虽然误差较大但结果仍可信。

4 结论

a. 将径流时间序列值使用 EMD 进行分解重组预测、使用 Hurst 指数作为筛选标准剔除掉噪声模态分量两种处理方式均有助于提升预测精度，与单一模型相比，前者构建的模型 R_{RMSE} 平均下降了 15.16%，后者平均下降了 28.49%。

b. EMD-SVR-ARIMA、EMD-SVR-LSTM 模型均较适合日径流量预测。

c. 本文仅研究了单一径流时间序列，未来可加入其他径流影响因素，如降水量、上游水文站径流量等进行拓展研究。

参考文献：

[1] 夏润亮，刘启兴，李涛，等. 基于集成学习的黄河未控区径流预测研究[J]. 应用基础与工程科学学报，2020，28(3)：740-749.

[2] ORELLANA-ALVEAR JOHANNA, MUÑOZ PAUL, BENDIX JÖRG. Influence of random forest hyperparameterization on short-term runoff forecasting in an andean mountain catchment[J]. Atmosphere, 2021, 12(2): 238-254.

[3] 李伶杰，王银堂，胡庆芳，等. 基于时变权重组合与贝叶斯修正的中长期径流预报[J]. 地理科学进展，2020，39(4)：643-650.

[4] YUAN XIAOHUI, CHEN CHEN, LEI XIAOHUI, et al. Monthly runoff forecasting based on LSTM-ALO model[J]. Stochastic environmental research and risk assessment, 2018, 32(8): 2199-2212.

[5] ZHANG JINPING, XIAO HONGLIN, FANG HONG YUAN. Component-based reconstruction prediction of runoff at multi-time scales in the source area of the Yellow River based on the ARMA model[J]. Water resources management, 2022(prepublish): 433-448.

[6] 张洪波，余荧皓，孙文博，等. 面向 EMD 分解的径流分量重构方法对比研究[J]. 南水北调与水利科技，2017，15(1)：60-66，166.

[7] 胡庆芳，曹士圮，杨辉斌，等. 汉江流域安康站日径流预测的 LSTM 模型初步研究[J]. 地理科学进展，2020，39(4)：636-642.

[8] 潘雅婧，王仰麟，彭建，等. 基于小波与 R/S 方法的汉江中下游流域降水量时间序列分析[J]. 地理研究，2012，31(5)：811-820.

[9] 龚云，信杰，南守珊. 一种引入 Hurst 指数的 MEMS 陀螺仪去噪模型[J]. 大地测量与地球动力学，2022，42(5)：457-461.

(下转第 9 页)

Responses of Growing Period NDVI (G-NDVI) to Meteorological Factors Spatio-temporal Variations in Lhasa River Basin

ZHANG Yang, ZHANG Run-run, GUO Ming-chen, WANG Zhao
(College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract: The Lhasa River Basin is a typical arid and semi-arid basin in the Qinghai-Tibet Plateau, where the ecosystem is extremely fragile. It is of great significance to study the spatio-temporal variation of vegetation index (NDVI) in response to the changes of meteorological factors, and to explore the adaptability of vegetation on the Qinghai-Tibet Plateau to the meteorological factors under the background of climate change. Based on the monthly NDVI, precipitation (P), and average temperature (T) time series dataset in the Lhasa River Basin from 1982 to 2017, using Pettitt, Mann-Kendall trend test and Pearson correlation analysis, this paper analyzed the spatio-temporal variation characteristics of growing period NDVI (G-NDVI) and meteorological factors, and identified responses patterns of G-NDVI to climate factors. The results show that the G-NDVI changed abruptly in 1997, and there was a trend shift from increasing to decreasing. The climate of the watershed changed from "wetting-colding" before the abrupt point to "drying-warming" after the abrupt point, and its effect on vegetation growth changed from promoting to inhibiting. There are two zones in the watershed where the responses patterns of G-NDVI to meteorological factors changed before and after the abrupt point. In the western permafrost areas, G-NDVI shows significant correlation with P and T in the second phase after 1997, i.e., emerging the "responding" function on meteorological factors variation. In the southern seasonal frozen zone, after the abrupt point, time lags of G-NDVI to P were elongated, and meanwhile the "responding" of G-NDVI to T has been triggered. The latency of NDVI response to P in the western permafrost region is longer than that in the southern seasonal frozen zone. In the second phase after 1997, the response area of NDVI to meteorological factors increased compared with that before the abrupt point, and the effect of meteorological factors on vegetation growth in the Lhasa River basin was enhanced.

Key words: NDVI; climate change; lag time; spatio-temporal dynamics; Lhasa River Basin

(上接第 22 页)

Daily Runoff Prediction of Lanzhou Hydrological Station in Yellow River Basin Based on EMD Decomposition

LU Wei[a], WEI Lin-jing[b]
(a. College of Science; b. College of Information Science and Technology,
Gansu Agricultural University, Lanzhou 730030, China)

Abstract: In order to improve the accuracy of runoff prediction, based on the control variable method and the daily runoff data of Lanzhou hydrometric station from August 2001 to December 2019, the models of the LSTM, ARIMA, SVR and XGBoost were used to establish 12 model schemes, including single model, EMD decomposition and reconstruction, EMD decomposition and reconstruction after removing noise components, and evaluation indicators of the 12 schemes were compared. The results show that the EMD sequence decomposition and reconstruction technology and noise component elimination based on Hurst exponent are helpful to improve the prediction accuracy. Compared with the single model, the R_{RMSE} of the model constructed by the former decreased by 15.16% on average, and that of the latter decreased by 28.49% on average. Among the 12 schemes, EMD-SVR-ARIMA with noise components removed is the best model.

Key words: daily runoff forecasting; EMD decomposition; Lanzhou hydrological station; machine learning model

(上接第 34 页)

Effects of Different Layout Forms of Rigid Vegetation Groups on Hydraulic Characteristics of Continuous Bends

ZHANG Chao-yu[1], SUN Xue-lan[1], JI Zi-qing[2], DUAN Jing-jing[1]
(1. College of Water Resources Science and Engineering, Taiyuan University of Technology, Taiyuan 030024, China;
2. State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Tianjin 300072, China)

Abstract: In order to understand the impact of vegetation groups on the hydraulic characteristics of natural meandering river, this paper designed two kinds of layout forms in which the cluster layout simulated the natural cluster vegetation group, and the uniform layout simulated the artificial regular planting vegetation group. Through the flume experiment, the impact of different layout forms of vegetation on the distribution of flow field, hydrodynamic axis, turbulent kinetic energy and transverse circulation structure in continuous bends was explored. The results show that the flow field in the bend appears obvious velocity partition in which the high velocity area is close to the convex bank, and the low velocity area is attached to the concave bank. Vegetation layout makes the distribution range of high and low flow rate areas have the opposite change rule: the high flow rate area increases, while the low flow rate area decreases. Under the action of the bend, the hydrodynamic axis gradually swings to the convex bank on the upstream side of the bend top, close to the convex bank near the bend top and swings to the concave bank on the downstream side of the bend top, completing a cycle of motion changes. Vegetation layout makes it swing to the middle of the flume and the deviation degree of cluster layout is greater than that of uniform layout. Under the action of the bend, the distribution of turbulent kinetic energy of water flow is small on both sides and large in the middle. On the whole, the layout of vegetation increases the turbulent kinetic energy on the downstream side of the bent top, and the impact of cluster layout is greater than the uniform. The vegetation layout changes the circulation structure of the section, which is manifested by the swing of the vortex core position and the change of the distribution range.

Key words: continuous bends; rigid vegetation; flow field; hydrodynamic axis; turbulent kinetic energy; circulating current