

DOI: 10.20040/j.cnki.1000-7709.2025.20242186

基于“优选—组合—修正”策略的月径流预测模型研究

王许彭¹, 罗军刚¹, 董洪涛¹, 张上要², 万 勇¹, 张晴阳¹

(1. 西安理工大学旱区水工程生态环境全国重点实验室, 陕西 西安 710048;
2. 中国电建集团贵阳勘测设计研究院有限公司规划研究院, 贵州 贵阳 550000)

摘要:为解决非线性径流序列及单一预测模型稳定性差导致的预测精度下降问题,在“分解—预测”模型的基础上,提出了一种“优选—组合—修正”模型构建策略,首先选择 DNN、SVM、LSTM、TCN、GBRT 5 种模型,建立了基于 EMD、CEEMDAN、VMD 的 15 种耦合模型,并对模型进行优选;其次将优选出的模型作为基础模型,并将基础模型各时期的预测结果进行处理,输入多层感知机,构建新的组合模型;然后针对组合模型的测试期构建残差修正方程,进一步提高模型预测精度;最后将其应用于渭河流域华县站和汉江流域洋县站测试研究中。结果表明,多层感知机构建的组合模型相比单一模型有更高的预测精度,并能够结合其他模型的优点,提高模型泛化能力。通过残差修正技术的模型在各方面均优于组合模型,尤其在径流峰值拟合效果上进一步提高了预测精度。

关键词: 信号分解; 径流预测; 残差修正; 模型优选; 多层感知机

中图分类号: TV121 **文献标志码:** A **文章编号:** 1000-7709(2025)09-0001-05

1 引言

受变化环境的影响,流域的径流特性呈现出非线性和非平稳的特征,这使得水文预测变得更加复杂。在这种背景下,传统的水文预测方法面临一定挑战^[1]。近年来,机器学习模型逐渐应用于径流预测中,因其能够从数据中挖掘规律,无需依赖复杂的物理过程,已广泛应用于径流预测中,并取得较好的结果^[2]。然而,机器学习模型的预测性能受限于输入数据质量和特征。为此,信号分解技术作为一种有效手段得到了关注。该技术通过分离不同尺度和频率的成分,去除噪声干扰并提取更多有价值的特征信息,从而提高预测模型的精度和可靠性^[3]。如张上要等^[4]提出的 VMD-TCN 耦合模型,能有效降低非平稳径流序列的影响,结果表明该模型在不同预见期下的预测精度高于其他模型。然而,单一预测模型容易受到自身算法限制,可能出现欠拟合训练或对异常值敏感等问题,从而导致预测精度下降。为了克服这些局限性,多模型组合成为提升预测精度的有效方法^[5]。周婷等^[6]通过贝叶斯模型平均方法对三种模型的预测结果进行组合,从而提高了

预测模型的准确性。此外,预测结果的修正也逐渐成为研究热点,该方法对模型输出进行处理,从而弥补模型预测中的偏差。如王文川等^[7]利用灰色模型和 BP 神经网络对径流预测,然后采用马尔可夫链推求状态概率进行修正,再利用最小二乘法对修正模型的结果进行耦合,结果表明修正模型拟合效果更好。鉴此,本文提出了一种“优选—组合—修正”的模型构建策略,并以渭河流域华县站和汉江流域洋县站为例验证所提模型的有效性,通过该研究以期能有效提升水文径流预测的精度和可靠性,为未来中长期水文预测模型的构建提供新的思路与方法。

2 研究方法

2.1 策略方法研究

为分析径流波动大且复杂的问题,先后形成了“分解—预测”、模型组合、优化算法等一系列解决方案。在已有研究的基础上,本文提出了一种“优选—组合—修正”策略,其中优选是指从多种单一模型中进行初筛,选择出适合不同流域断面的几种基础模型,旨在提高后续组合模型的适用性;组合是指在优选的几种基础模型内,采用人工

收稿日期: 2024-10-21, 修回日期: 2024-12-04

作者简介: 王许彭(1999-),男,硕士研究生,研究方向为工程管理、水文水资源, E-mail: 1759140254@qq.com

神经网络,通过多层结构、非线性激活函数等捕捉模型预测与实测之间的复杂非线性关系,旨在降低模型自身算法影响,同时结合各个模型的优势;修正是指在组合模型的基础上,构建测试集的预测值与实测值之间的回归方程,修正预测结果。

2.2 分解—预测模型构建

构建“分解—预测”模型时,选择经验模态分解(EMD)、自适应噪声完备集合经验模态分解(CEEMDAN)和变分模态分解(VMD)三种信号分解方法,其中VMD是一种自适应、完全非递归的信号处理方法,自适应确定模态分解的个数,通过变分框架在EMD的基础上对径流序列进行迭代优化,减少模态混叠问题;CEEMDAN是在传统的EEMD基础上,引入了自适应噪声和多次分解,通过在原始信号中多次添加不同的噪声,生成新的信号集合,然后多次迭代分解,逐步去除序列中的噪声干扰。

预测模型选择深度神经网络(DNN)、支持向量机(SVM)、长短期记忆网络(LSTM)、梯度提升回归树(GBRT)^[8]、时间卷积网络(TCN),其中DNN^[9]通过多层神经元连接,能处理大规模的高维数据,SVM^[10]则适合于小样本数据处理,二者对于复杂的非线性径流数据均有较好效果;LSTM^[11]引入了细胞状态的概念用来储存信息,能够更好地处理长序列数据,捕捉长期依赖关系。梯度提升回归树(GBRT)通过多个决策树的集成,逐步提升模型的预测精度;TCN^[12]则通过卷积操作而非传统的递归结构,处理时序数据,特别适用于具有长时间依赖关系和长期趋势的水文序列。这些模型具备多样性和互补性,能够有效捕捉径流数据的复杂性和多样性^[13],将它们与三种信号分解方法相结合,构建出多种“分解—预测”模型,将其应用于径流预测,为后续优选模型提供充足支撑。

2.3 基于MLP的模型组合模型

采用模型组合方法兼顾基础模型的优势,提高稳定性,同时还需权衡模型的复杂性和过拟合的风险,因此组合模型应选择简单组合方法。本文选用多层感知机(MLP)作为组合模型的算法。MLP^[14]是一种人工神经网络,其结构包括输入层、隐藏层和输出层。每一层的神经元通过全连接方式连接到下一层,每个连接均有相应的权重,用来调整输入影响。隐藏层和输出层之间的神经元通过激活函数进行非线性变换。相较于线性组合,MLP组合模型能够捕捉基础模型中的复杂非线性关系,从而结合各个模型优势,提高组合

模型的稳定性和预测精度。

2.4 基于最小二乘法的模型修正

为进一步提高预测精度,需建立回归误差方程来修正模型的预测结果。其原理基于最小二乘法,将组合模型的预测值作为自变量,模型与实测值之间的残差作为因变量,从而调整模型的预测效果。其步骤如下。

步骤1 计算残差。即:

$$e_i = y_i - \hat{y}_i \quad (1)$$

式中, e_i 为残差序列; y_i 、 \hat{y}_i 分别为实测序列、预测序列。

步骤2 建立残差修正方程。即:

$$e_i = \beta_0 + \beta_1 \hat{y}_i + \varepsilon_i \quad (2)$$

式中, β_0 、 β_1 均为回归系数; ε_i 为随机误差,服从正态分布 $N(0, \sigma^2)$ 。

步骤3 求解回归系数。使用最小二乘法拟合残差修正模型,通过对残差修正模型的损失函数进行最小化求解,即对损失函数求偏导数,并令偏导数等于零,可得到回归系数 β_0 、 β_1 的估计值。损失函数为残差修正模型残差的平方和,即:

$$R_{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n (e_i - \beta_0 - \beta_1 \hat{y}_i)^2 \quad (3)$$

步骤4 得到残差修正方程。

步骤5 根据残差值修改预测值。即:

$$q_i = \hat{y}_i - e_i \quad (4)$$

式中, q_i 为修正序列。

3 实例应用

3.1 研究区径流演变情况

以渭河流域华县水文站和汉江流域洋县水文站为例,通过两个不同流域水文站来验证方法策略的可行性。采用1970年1月~2018年12月共计49年的月径流数据。序列按照8:1:1划分为训练集、校验集和测试集。采用线性趋势和MK趋势检验进行趋势分析,采用MK检验、Pettitt检验、滑动t检验进行突变分析,结果见表1。由表1可知,水文序列存在非平稳特征。

表1 水文序列趋势分析与突变点检验结果

Tab. 1 Hydrological sequence trend analysis and change point detection results

水文站	时期	多年平均径流量 / 10^8 m^3	MK 检验 (Z 统计量)	突变点 检验	趋势 分析	时间
华县	训练集	66.04	-2.48	1993 年	下降	1970~2008
	验证和测试集	62.17	-0.12	不存在	下降	2009~2018
洋县	训练集	60.50	-3.28	1991 年	下降	1970~2008
	验证和测试集	59.57	-1.24	不存在	下降	2009~2018

3.2 信号分解技术

对径流序列采取信号分解技术处理。以华县站为例,使用 python 的 pyvmd 库对径流序列进行 VMD 分解预试验。当模态数 $K=9$ 时,中心频率出现模态混叠现象,当 $K=8$ 时,分解效果最佳,同时无模态混叠的情况发生。对于 EMD 分解,CEEMDAN 分解则采用 python 的 PyEMD 库进行分解,得到分解结果。

3.3 评价指标

为了保证组合模型的预测精度,选取纳什效

率系数(N_{NSE})、标准均方根误差(N_{NRMSE})、平均绝对百分比误差(M_{MAPE})、百分比峰值趋势(P_{PPTS})和相关系数(R^2)5 种评价指标来评估模型预测能力。

3.4 单一模型优选

根据评价指标对模型的模拟结果进行评价,从中筛选出预测精度较高的基础模型,为组合模型做准备,预测结果见表 2。

由表 2 可知,不同的分解技术处理的预测模型精度有较大差距,VMD 分解的效果远优于 EMD、

表 2 测试期不同分解算法的模型性能评价指标

Tab. 2 Model performance evaluation metrics of different decomposition algorithms during the testing periods

水文站	模型	SVM		LSTM		DNN		TCN		GBRT	
		N_{NSE}	N_{NRMSE}	N_{NSE}	N_{NRMSE}	N_{NSE}	N_{NRMSE}	N_{NSE}	N_{NRMSE}	N_{NSE}	N_{NRMSE}
华县	VMD	0.912	0.285	0.895	0.310	0.886	0.323	0.858	0.360	0.738	0.489
	EMD	0.617	0.592	0.563	0.633	0.811	0.416	0.596	0.608	0.551	0.641
	CEEMDAN	0.862	0.355	0.811	0.416	0.813	0.415	0.769	0.459	0.744	0.485
洋县	VMD	0.926	0.304	0.820	0.314	0.906	0.367	0.800	0.526	0.758	0.578
	EMD	0.701	0.638	0.725	0.532	0.708	0.636	0.583	0.675	0.688	0.657
	CEEMDAN	0.857	0.361	0.753	0.511	0.779	0.553	0.909	0.353	0.758	0.471

CEEMDAN 分解。华县站数据中,以 SVM、LSTM 模型为例,采用 EMD、CEEMDAN 分解的模型的纳什系数分别为 0.617、0.862 和 0.563、0.811,而采用 VMD 分解的模型在验证期的纳什系数分别为 0.912、0.895,相较于 EMD、CEEMDAN 分解的模拟结果分别提高了 47.6%、5.7% 和 58.8%、10.4%。这表明 VMD 分解能够更好地适应不同场景、不同模型的需求,有效解决了模态混叠的问题,提高了模型的泛化能力和适应性。GBRT 模型预测性能最差,其平均纳什系数仅为 0.706,预测精度未达到丙级,也低于其他模型的平均预测精度 0.811、0.761、0.817、0.753。其中,DNN 模型受分解算法影响最小,模拟效果稳定,VMD 分解效果最佳。因此,对于华县站,基础模型选择基于 VMD 分解的 SVM、LSTM、DNN 模型。对于洋县站,基础模型选择基于 VMD 分解的 SVM、DNN 模型和基于 CEEMDAN 分解的 TCN 模型。

3.5 MLP 组合模型

为防止过拟合情况发生,将 MLP 神经网络配置为 3 层,包括输入层、一个隐藏层和输出层。采用 ReLU 作为激活函数,能够有效地解决梯度消失问题并加速模型训练。将多个优选出基础模型预测结果,作为 MLP 模型的训练集和测试集,得到组合模型预测结果,见表 3。

3.6 残差修正

基于 1970~2013 年华县站实测径流序列和

表 3 组合模型训练期和测试期性能评价

Tab. 3 Performance evaluation of the combined model during the training and testing periods

水文站	时期	N_{NSE}	N_{NRMSE}	M_{MAPE}	P_{PPTS}
华县	训练期	0.996	0.073	19.03	0.86
	测试期	0.927	0.258	28.15	6.88
洋县	训练期	0.983	0.167	26.92	2.61
	测试期	0.936	0.256	36.16	6.95

组合模型的预测序列,构建残差修正方程。通过 scipy 库中的 minimize 函数,根据最小二乘法原理,求解使损失函数最小化的参数值,即回归系数,将回归系数代入残差修正方程,得到残差修正方程,见表 4。

表 4 残差修正方程

Tab. 4 Residual correction equation

水文站	残差回归方程	β_0	β_1
华县	$x=0.00398y-35.02$	0.00398	-35.02
洋县	$x=0.01339y-214.77$	0.01339	-214.77

注: x 为残差, y 为组合模型的预测值。

3.7 对比试验及结果分析

将 2014~2018 年组合模型预测值经过残差修正后得到最终的修正结果,见表 5。图 1 为测试集模型性能评价雷达图。

由表 5、图 1 可知,与“分解—预测”模型相比,组合模型融合了单一模型的优势,使得模型在预测精度、鲁棒性和泛化能力方面均有提升,以华县站为例,组合模型的 N_{NSE} 达到了 0.927, P_{PPTS} 、 N_{NRMSE} 、 M_{MAPE} 分别为 6.88、0.328、29.26。修正模型预测精度的提升则更加明显,对比

表 5 测试期修正模型与单一模型性能评价指标
Tab. 5 Performance evaluation metrics of correction models and single models during the testing periods

水文站	模型	N_{NSE}	N_{NRMSE}	M_{MAPE}	P_{PPTS}
华县	VMD-SVM	0.912	0.285	29.81	6.72
	VMD-LSTM	0.895	0.310	28.27	8.08
	VMD-DNN	0.886	0.323	29.66	13.02
	组合模型	0.927	0.258	28.15	6.88
	修正模型	0.932	0.249	27.85	6.21
洋县	VMD-SVM	0.926	0.304	41.30	6.24
	VMD-DNN	0.906	0.367	43.99	6.17
	CEEMDAN-TCN	0.909	0.353	39.42	13.50
	组合模型	0.936	0.256	36.16	6.95
	修正模型	0.937	0.253	33.93	6.88

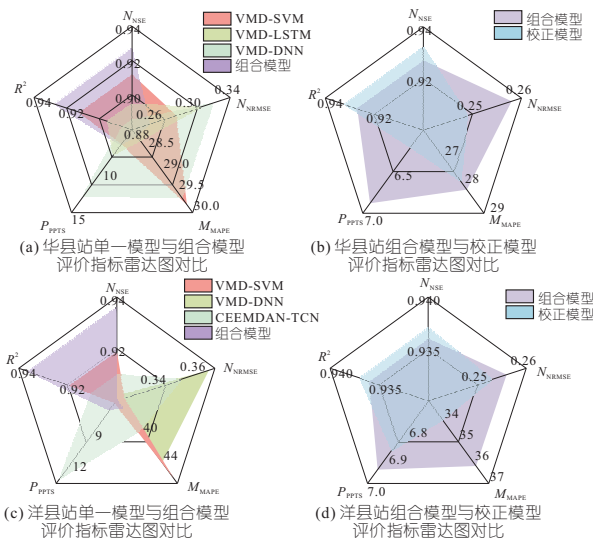


图 1 测试集模型性能评价雷达图

Fig. 1 Radar chart of model performance evaluation during the testing periods

VMD-DNN 模型,组合模型的纳什系数提高了 4.6%,达到了 0.927,预测精度为甲级, N_{NRMSE} 、 M_{MAPE} 则分别降低了 20.1%、5.1%。特别是 P_{PPTS} 值降低了 47.2%,峰值预测精度明显提高,有效弥补了 DNN 模型对于极值预测的缺陷。对比组合模型,修正模型有更高的 N_{NSE} 和更低的 N_{NRMSE} 、 P_{PPTS} 、 M_{MAPE} ,表明组合模型可通过残差修正方程减少误差,进一步提高模型的预测性能。

图 2 为测试集多模型实测与预测月径流量折线图。由图 2 可知,修正模型的预测精度最高,组合模型次之,其余模型预测效果一般。观察极值分布,修正模型的峰值最接近于实测峰值,介于组合模型数据和实测数据之间,经过修正的模型相较于单一模型的峰值和谷值误差明显降低,修正模型对峰值和谷值的拟合更加精确。结合表 5 可知,组合模型通过多层结构和反向传播算法,自适应的调整各模型的权重,结合多个基础模型的优势,提高了模型预测精度。在组合模型的基础上,修正模型通过构建训练集实测值与预测值

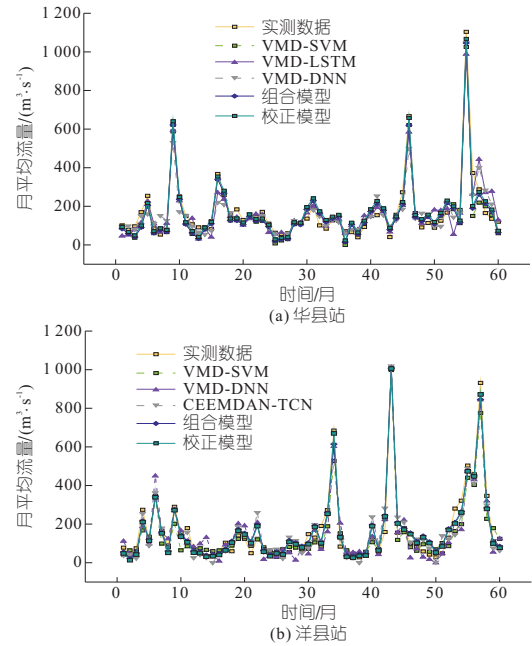


图 2 测试集多模型实测与预测月径流量折线分布图
Fig. 2 Line and scatter plot of measured and predicted monthly average flow of multiple models

之间的回归方程,修正测试集的预测结果,进一步减小误差。

图 3 为多模型实测与预测月径流量散点分布图。图 3 中黑色直线为基准线,其相关系数为 1。由图 3 可知,相较于 VMD-SVM、VMD-LSTM、VMD-DNN 模型,组合模型的拟合线距离基准线最近,其散点也更均匀的分布在基准线附近,在峰值处也更接近实测值。经过残差修正技术的模型相关系数达到了 0.932,高于组合模型,同时修正模型的拟合线与基准线夹角最小,峰值更加贴合。这表明修正技术是有效的,预测值和实测值相关性进一步提高。

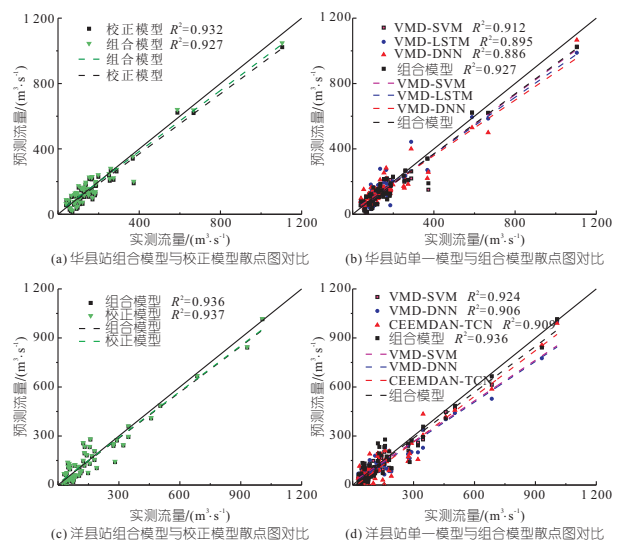


图 3 多模型实测与预测月径流量散点分布图
Fig. 3 Multi-model observed and predicted runoff scatter plot

4 结论

a. 信号分解技术挖掘了更多的特征信息,减少了模态混叠现象。

b. 组合模型能够捕捉基础模型之间的复杂非线性关系和峰值变化趋势,减少了各预测模型误差带来的不稳定性,提高了模型的泛化能力。

c. 通过残差修正技术的修正模型在各个方面均优于组合模型,具有更小的 N_{NRMSE} 、 M_{MAPE} 、 P_{PPTS} 和更大的 N_{NSE} ,能够进一步提高预测精度,在预测峰值方面也最接近实测值。

参考文献:

- [1] 陈克宇. 变化环境下河川径流中长期规律与预测研究——以陕西省典型流域为例[D]. 西安: 长安大学, 2015.
- [2] 张力,王红瑞,郭玮楠,等. 基于时序分解与机器学习的非平稳径流序列集成模型与应用[J]. 水科学进展, 2023, 34(1): 42-52.
- [3] 梁浩,黄生志,孟二浩,等. 基于多种混合模型的径流预测研究[J]. 水利学报, 2020, 51(1): 112-125.
- [4] 张上要,罗军刚,石国栋,等. 基于 VMD-TCN 模型的渭河流域月径流量预测研究[J]. 人民黄河, 2023, 45(10): 25-29.
- [5] 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.
- [6] 周婷,温小虎,冯起,等. 基于 BMA 多模型组合的疏勒河径流预测研究[J]. 冰川冻土, 2022, 44(5): 1606-1619.
- [7] 王文川,李文锦,徐冬梅,等. 基于马尔可夫链校正 GM-BP 模型的径流预测[J]. 南水北调与水利科技, 2019, 17(5): 44-49.
- [8] 杨文忠,张志豪,吾守尔·斯拉木,等. 基于时间序列关系的 GBRT 交通事故预测模型[J]. 电子科技大学学报, 2020, 49(4): 615-621.
- [9] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.
- [10] ASEFA T, KEMBLAWSKI M, MCKEE M, et al. Multi-time scale stream flow predictions: The support vector machines approach[J]. Journal of hydrology, 2006, 318(1/2/3/4): 7-16.
- [11] 王鑫,吴际,刘超,等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报, 2018, 44(4): 772-784.
- [12] BAI S J, KOLTER J Z, KOLTUN V. Trellis networks for sequence modeling[EB/OL] <https://arxiv.org/abs/1810.06682v2>, 2018: 1810.
- [13] 鲍振鑫,张建云,王国庆,等. 基于水文模型与机器学习集合模拟的水沙变异归因定量识别——以黄河中游窟野河流域为例[J]. 水科学进展, 2021, 32(4): 485-496.
- [14] 韩玲. 基于人工神经网络——多层感知器(MLP)的遥感影像分类模型[J]. 测绘通报, 2004(9): 29-30, 42.

Monthly Runoff Prediction Based on Selection-Combination-Correction Strategy

WANG Xu-peng¹, LUO Jun-gang¹, DONG Hong-tao¹, ZHANG Shang-yao²,
WAN Yong¹, ZHANG Qing-yang¹

(1. State Key Laboratory of Water Engineering Ecology and Environment in Arid Area, Xi'an University of Technology, Xi'an 710048, China; 2. Planning Research Institute, PowerChina Guiyang Engineering Corporation Limited, Guiyang 550000, China)

Abstract: To solve the problem of decreasing prediction accuracy caused by nonlinear runoff sequence and instability of single prediction model, this paper proposes a "selection-combination-correction" modeling strategy based on the "decomposition-prediction" model. Firstly, five models including DNN, SVM, LSTM, TCN, and GBRT are used to establish 15 coupled models based on EMD, CEEMDAN, and VMD, and the models are selected. Then, the selected model is used as the base model, and the predicted results of each period of the base model are processed and input into a multi-layer perceptron to construct a new combination model. A residual correction equation is constructed for the test period of the combination model to further improve the prediction accuracy. Finally, the method is applied to the test studies of Huaxian Station in Weihe River Basin and Yangxian Station in Hanjiang River Basin. The results show that the combination model constructed by the multi-layer perceptron has higher prediction accuracy than the single model, and can integrate the advantages of other models to improve the model's generalization ability. The model with residual correction technology is superior to the combination model in all aspects, especially in the fitting of peak discharge, further improving the prediction accuracy.

Key words: signal decomposition; runoff prediction; residual correction; model optimization; multilayer perceptron