DOI: 10.19797/j.cnki.1000-0852.20210311

SMA-LSTM模型在径流预报中的适用性研究

李 佳[1], 曲 田[1], 牟时宇[1], 陶思铭[2], 胡义明[2]

(1.国能大渡河流域水电开发有限公司, 四川 成都 610041;
2.河海大学 水文水资源学院, 江苏 南京 210098)

摘 要: 径流预报对于防洪、发电和生态调度等具有重要意义。以大渡河丹巴以上流域为研究区域,采用黏菌优化算法(SMA)对长短期记忆神经网络(LSTM)的隐藏层输出维度进行优化,构建SMA-LSTM模型对未来10日径流过程进行预报,以探讨深度学习方法对流域径流预报的适用性。基于2012—2017年的日降雨量和日流量资料,构建了预见期为10天的逐日径流SMA-LSTM预报模型,以2018—2019年的资料进行模型验证;采用最大1日径流量相对误差和10日总径流量相对误差作为SMA-LSTM模型精度的评价指标,并与未优化的LSTM模型和新安江模型结果进行对比。结果表明:SMA-LSTM模型具有较高的模拟和预报精度,无论是在率定期还是验证期,两种指标均控制在±10%以内,且两种指标的绝对值平均都不超过7%;整体而言,SMA-LSTM模型精度更高,预报的径流过程与实测过程更为贴近。研究成果可供流域径流预报实际工作参考。

关键词: 长短期记忆神经网络(LSTM);黏菌优化算法(SMA);径流预报

中图分类号: TV124 **文献标识码:** A **文章编号:** 1000-0852(2023)01-0047-05

0 引言

径流预报是保证防洪度汛安全,提高水资源综合利用效率的重要技术基础。但随着预见期的增长,径流的不确定性影响因素随之增多,给径流的精准预报带来困难。因此,研究径流预报模型与方法具有重要意义。

现有的径流预报模型包括基于物理过程驱动和基于数据驱动两种类型。前者对流域水文过程适当概化,再采用数学物理方程和逻辑方程进行描述,该类方法具有模型结构清晰、物理基础较强等特点,如新安江模型[1]、水箱模型[2]等概念性模型以及SWAT模型[3]和VIC模型[4]等分布式模型。但物理过程驱动的水文模型对资料要求高,有时会限制其实际应用。基于数据驱动的预报模型通过分析众多影响因子与预报量的因果关系,筛选出起主要作用的因子,并建立预报变量与影响因子之间的统计或数据关系以实现预报。该类方法不直接考虑产汇流过程的物理关系,而是根据数据间的统计或逻辑关系建立预报模型,在实际应用中可根据资料条件确定预报因子,具有建模简单、使用方便等特点,因此也可获得较好效果[5]。数据驱动类模型包括传统的统计模型如线性回归[6]和现代的机器学习模型如支持向量机、随机森林[7]和神经网络[8]等。其中,循环神经网络因其循环结构的设定更加适合处理序列数据,因此可应用于解决水文时序数据的模拟与预测问题。但水文时序数据(如逐日径流过程数据)前后之间往往具有较强的关联性,当日径流与多日前的影响因素均有一定的相关性,特别对较大流域更是如此,而传统的循环神经网络结构对久远信息(长记忆)的利用受到一定限制。近年来提出的长短期神经网络[9](Long Short- Term Memory, LSTM),在传统循环神经网络的基础上通过改变神经

收稿日期:2021-09-02 网络首发日期:2023-01-19
网络首发地址:https://kns.cnki.net/kcms/detail//11.1814.p.20230117.1818.014.html
基金项目:国家自然科学基金(41730750);国能大渡河流域水电开发有限公司科技项目(CEZB200505212)
作者简介:李佳(1983-),女,四川德阳人,硕士研究生,主要从事径流预报及梯级水库优化调度方面的工作。E-mail:156032658@qq.com
通信作者:胡义明(1986-),男,江苏宿迁人,副教授,主要从事水文水资源方面工作。E-mail: yiming.hu@hhu.edu.cn

元结构，可以有效解决长记忆丢失问题，因此对水文时间序列的预报亦具有适用性[10]。

LSTM中有多个需人工设定的超参数，不同的超参数组合会影响模型的最终结果[11]。为充分展示LSTM模型的模拟效果，需选取能达到最优或接近最优模拟效果的超参数组合。Li等[12]根据黏菌在觅食过程中的行为和形态变化中收获灵感，对其完整生命周期进行重建进而构造的黏菌优化算法（Slime Mould Algorithm, SMA）可用于挑选LSTM的超参数组合。因此结合两种算法构建的SMA-LSTM模型用于水文时间序列预报具有相对LSTM模型更精确的模拟效果，以期获得更优的预报结果。

本文以大渡河丹巴断面为研究对象，采用SMA-LSTM模型构建预见期为10天的逐日径流量预报模型，并与LSTM模型和新安江模型结果进行对比分析，为大渡河径流预报提供一种新的方案。

1 模型与方法

1.1 LSTM神经网络模型

神经网络是一种受到大脑神经突触联接结构处理信息的启发而诞生的数学模型，其中循环神经网络（Recurrent Neural Network, RNN）的结构更加适合处理序列数据。理论上，RNN可以利用任意长的序列信息，但因梯度消失、梯度爆炸等缺陷，实际运行过程中仅仅可记忆之前几步的输入信息。

LSTM作为RNN的变体，通过改变隐藏层结构，增加类似“传送带”的细胞状态设计和让信息选择性通过“门”的设计来控制记忆信息的流动，有效解决了信息的长期传输和记忆问题[9]。因此，LSTM比RNN更适合处理时间序列问题，例如水文预报、水文数据的插补等。LSTM神经网络概念图如图1所示。

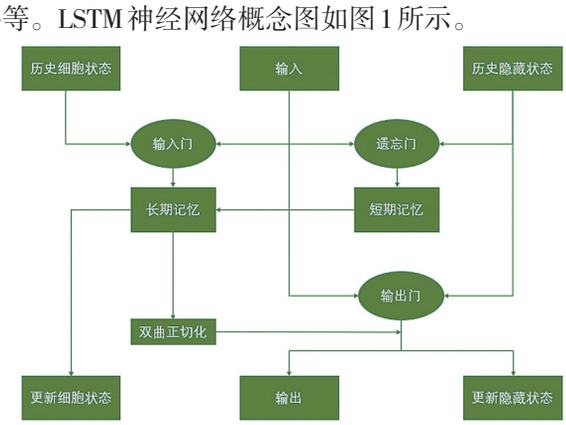

图1 LSTM神经网络概念图
Fig.1 Concept map of LSTM

1.2 黏菌优化算法

针对黏菌在发现食物时会振荡收缩特性，可利用权值的变化模拟觅食过程中黏菌本体产生的正反馈和负反馈过程，进而模仿觅食形态[12]。黏菌的位置公式为：

$$x(t+1)=\begin{cases}x_b(t)+vb\times\left(W\times x_{rand1}(t)-x_{rand2}(t)\right), r>p\\ vc\times x(t), r<p\end{cases} \quad (1)$$

式中：t为迭代次数；$x(t)$为在第t次迭代时黏菌所在位置；r为[0, 1]内的随机数；x_b(t)为当前适应度最优的个体位置；vb为在[−a, a]内的随机数；vc为由1到0线性减小的参数；x_{rand1}(t)、x_{rand2}(t)表示随机个体位置。

其中控制参数p, vb绝对值的限制a以及权重系数W的公式如下：

$$p=\tanh\left|S(i)-DF\right| \quad (2)$$

$$a=\operatorname{arctan}h\left(1-\frac{t}{t_{\max}}\right) \quad (3)$$

$$W\left(SmellIndex(i)\right)=\begin{cases}1+r\times\log\left(\left(b_F-w_F\right)+1\right), condition\\ 1-r\times\log\left(\left(b_F-w_F\right)+1\right), others\end{cases} \quad (4)$$

$$SmellIndex=sort(S) \quad (5)$$

式(2)~(5)中：i∈1, 2, 3, ..., N, N为种群数量；$S(i)$为个体适应度值；DF为当前取得的最佳适应值；t_{max}为最大迭代次数；$condition$为种群中适应度排在前一半的个体；$others$为剩下的个体；b_F为当前迭代获取的最佳适应度值；W_F为当前迭代最差适应度值；$SmellIndex$为排序后的适应度值序列。

通过上述迭代，尽管已找到局部最优的食物源，黏菌还是会分离部分个体用于寻找更优的食物源，新的位置公式如下：

$$x(t+1)=\begin{cases}rand\times\left(UB-LB\right)+LB, rand<z\\ x_b(t)+vb\times\left(W\times x_{rand1}(t)-x_{rand2}(t)\right), r>p\\ vc\times x(t), r<p\end{cases} \quad (6)$$

式中：$rand$为在[0, 1]内的随机数，用于形成任意角度的搜索向量；UB和LB分别为搜索范围的上限和下限；z为自定义参数。另外，在寻找更优食物源过程中，vb和vc随迭代次数增加逐渐接近0。

计算时，首先初始化种群，设定相应参数；然后计算适应度值$S(i)$并且排序，通过式(6)更新种群位置；然后重新计算适应度值并更新全局最优位置和当前最优位置；最后判断是否满足结束条件，满足输出最优结果，否则，重复上述过程直至满足条件。

因各层LSTM神经网络中不同的输出维度会对应不同的结果,则可通过黏菌优化算法对输出维度值进行最优化判断,将各层LSTM神经网络中的输出维度分别设为不同类别的黏菌,利用上述算法获取最优组合。

1.3　新安江模型

新安江模型是基于蓄满产流理论建立的模型,适用于湿润和半湿润地区[1]。新安江模型的结构主要分为4个层次,一是采用三层蒸散发模式进行蒸散发计算;二是采用蓄满产流概念计算模式产流;三是采用自由水蓄水库结构将径流划分为地表、壤中及地下径流(三水源结构);四是将汇流分为坡面、河网和河道汇流三阶段。新安江模型参数的物理意义比较明确,且都有相对应的合理的取值范围,在国内湿润和半湿润区水文预报中得到了普遍使用。

2　应用研究

2.1　流域概况

大渡河是长江上游岷江水系的最大支流,丹巴站是大渡河流域上游的一个重要节点。丹巴以上流域范围为东经99°35′~102°55′,北纬30°17′~33°27′,控制面积为5 2763 km^2,约占大渡河流域总面积的68%。丹巴以上流域属川西高原气候区,干湿季分明、气温日变化大。径流由降雨形成,部分为融雪和冰川补给。多年平均年降水量一般仅600~700 mm,且集中在6—9月,多年平均年径流量为400~500 mm。流域内以丘状高原地貌为主,山顶平缓浑圆,谷底宽阔平坦,河流迂回曲折。丹巴站流量大小基本可反映大渡河河源区对中下游水库入库径流量的贡献情况,因此,丹巴断面流量的准确预测对大渡河流域的发电、防洪和水资源综合利用具有至关重要作用。丹巴以上流域水系及主要水文站网分布如图2所示。

2.2　数据及处理

选用丹巴流域内2012—2019年间共8年的日降雨和日流量数据进行预报模型的率定和验证,其中,输入的降雨量数据为日部、一林场、马尔康、绰斯甲、足木足、大金、布科、东谷和小金共9个子流域内观测站收集到的实测降雨,并通过泰森多边形法计算得到面平均日雨量。径流资料为丹巴、布科、东谷、小金、大金站共5个站的同期流量数据。将2012—2017年作为模型率定期(对于LSTM模型和SMA-LSTM模型,2016—2017年用来检验模型是否过拟合);2018—2019年作为模型验证期。

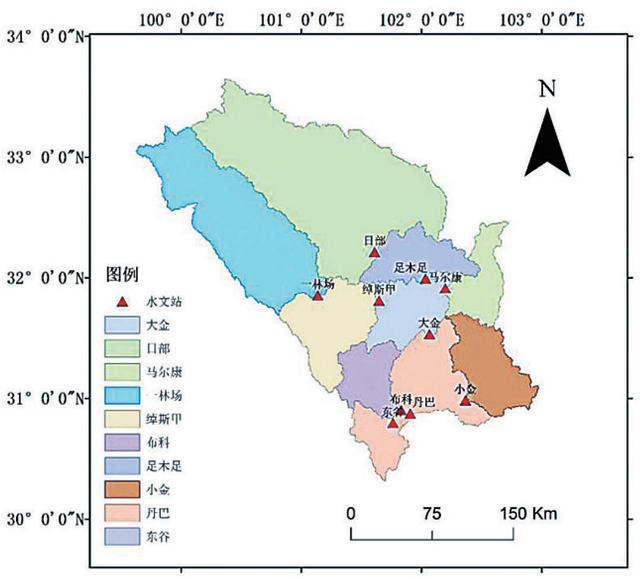

图2　丹巴以上流域水系及站网分布

Fig.2　Distribution of watershed system and station network of the River Basin above Danba

将9个子流域的面平均雨量和5个水文站流量作为LSTM模型和SMA-LSTM模型的输入时,因降雨量和径流量在量纲不同且量级上差别较大,先将数据进行归一化处理,即将数据变换到[0, 1]区间内,本文使用最大最小归一化处理,公式为:

$$X_{\text{norm}}=\frac{X-X_{\min}}{X_{\max}-X_{\min}} \tag{7}$$

式中:X为原始的实测数据,$X_{\max}$为实测数据中最大值,$X_{\min}$为实测数据的最小值,X_{norm}为归一化的数据。

2.3　LSTM模型和SMA-LSTM模型的构建

2.3.1　模型构建方式

LSTM模型和SMA-LSTM模型的模型构建方式一致,以下以SMA-LSTM模型为例,构建丹巴站预见期为K(本次K=10)天的逐日流量过程预报模型。设从t_0日开始预报,对未来任一日t_n($1\leqslant n\leqslant K$)的预报径流量$W(t_n)$,建立$W(t_n)$与t_0前L日逐日径流量、前L日逐日面雨量以及t_n($1\leqslant n\leqslant K$)预见期内逐日预报面雨量的SMA-LSTM模型,记为SMA-LSTM-n,模型构造图如图3所示。不同预见期初设的损失函数、隐藏层层数等超参数相同,但数据之间相关关系不同,共需构建10个不同的SMA-LSTM模型。

$$\begin{aligned}
&\text{SMA-LSTM-1}:W_{t+1}=f(W_{t-L},\ldots,W_{t-1},W_t,P_{t-L+1},\ldots,P_t,P_{t+1})\\
&\text{SMA-LSTM-2}:W_{t+2}=f(W_{t-L},\ldots,W_{t-1},W_t,P_{t-L+2},\ldots,P_{t+1},P_{t+2})\\
&\vdots\\
&\text{SMA-LSTM-}K:W_{t+K}=f(W_{t-L},\ldots,W_{t-1},W_t,P_{t-L+K},\ldots,P_t,\ldots,P_{t+K})
\end{aligned} \tag{8}$$

式中:W_m在$m\leqslant t$时为实测径流量,$m>t$时为预报径流量,P_n在$n\leqslant t$时为实测降水量,$n>t$时为预报降水量。

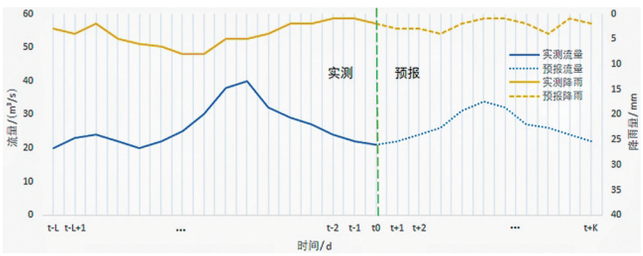

图3 模型构造示意图

Fig.3 Model construction diagram

2.3.2 模型参数设定

SMA模型中择优的超参数为两层隐藏层中各自的输出维度(units),*UB*设为30,*LB*设为10,迭代次数设为200,种群数量设为30,*z*值设为0.4,用于计算适应度值的函数设定为平均绝对误差(Mean Absolute Deviation, MAE),其公式为:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_0(i)-y_f(i)\right| \tag{9}$$

式中:n为样本个数;$y_0(i)$为归一化后的实测流量值;$y_f(i)$为模型计算值。

LSTM模型中两层隐藏层中各自的输出维度均设置为20,其他超参数SMA-LSTM模型和LSTM模型设置一致。

神经网络训练过程的本质在于最大化模型模拟值与实测值贴近程度,与SMA模型中适应度值的函数相同,选取MAE作为损失函数,模型一次训练所选取的样本数(batch-size)设定为512,模型针对完整训练集重复训练的次数(epochs)设定为600次,隐藏层为两层LSTM和一层全连接层,全连接层输出维度设置为1。

2.4 新安江模型构建

根据水文站点空间位置和水系组成,将丹巴以上流域划分为9个子流域单元(见图2),各子流域分别进行新安江模型参数率定。预报时,先进行上游子流域的产汇流计算,并将其出口断面的流量过程采用马斯京根法进行河道汇流演算,再与区间子流域经产汇流计算得到的流量过程进行叠加,得到下游断面的流量过程,最终得到丹巴站的流量过程。

2.5 模型评价指标

本文以确定性系数和相对误差为指标,对模型精度进行评定。

(1)最大1日径流量相对误差RE_{1d}。

$$RE_{1d}=\frac{W_m-W_o}{W_o}\times 100\% \tag{10}$$

式中,W_m为10日预报日径流量中最大1日的径流量、W_O为同期范围内实测流量最大值。RE_{1d}越接近0,表示模型精度越高,尤其是汛期时,该指标更能体现模型预测峰值的能力。

(2)10日径流量相对误差RE_{10d}。

$$RE_{10d}=\frac{\sum_{i=1}^{10}W_m(i)-\sum_{i=1}^{10}W_o(i)}{\sum_{i=1}^{10}W_o(i)}\times 100\% \tag{11}$$

式中:$W_m(i)$为预报径流量、$W_O(i)$为实测径流量。RE_{10d}越接近0,表示模型精度越高,对预见期内径流总量的预报就更准确。

2.6 结果分析

优选的SMA-LSTM模型中两层隐藏层中神经元个数信息见表1。SMA-LSTM模型及新安江模型(XAJ)的精度统计结果见表2,包含汛初(4月份)、汛中(6月份)与汛末(10月份)的部分场次结果(率定期18场,验证期6场)。

表1 SMA-LSTM-K模型隐藏层神经元个数对应表

Table1 number of neurons in SMA-LSTM-K model's hidden layers

隐藏层 \ K	SMA-LSTM-K									
	1	2	3	4	5	6	7	8	9	10
第一层	25	29	27	28	22	27	30	26	29	30
第二层	12	25	25	19	20	15	22	17	24	28

表2 丹巴以上流域SMA-LSTM、LSTM与XAJ模型10天预见期径流预报精度统计

Table2 Precision comparison of SMA-LSTM and XAJ models in 10 day forecast period of Danba River Basin

时期	预报日期	SMA-LSTM		LSTM		XAJ	
		RE_{1d}	RE_{10d}	RE_{1d}	RE_{10d}	RE_{1d}	RE_{10d}
率定期	20120420	-10.36	3.1	2.36	15.32	21.69	40.83
	20120608	0.19	1.13	1.85	-2.61	-4.19	-6.56
	20121022	0.08	2.3	2.64	5.64	5.39	12.77
	20130423	-6.31	-14.92	-10.96	-20.31	42.7	32.78
	20130616	-1.34	5.22	6.97	10.94	7.62	22.18
	20131006	-0.94	0.75	8.75	4.29	17.88	18.93
	20140414	-9.2	-9.87	8.66	-12.11	13.1	21.25
	20140614	1.64	0.32	6.12	1.3	-5.93	16.28
	20141018	1.59	1.29	5.31	1.26	12.13	18.53
	20150423	-4.66	-3.05	2.98	9.82	36.97	60.9
	20150611	1.37	0.24	7.29	7.61	-6.2	12.74
	20151005	2.8	-1.37	-0.91	2.27	1.97	0.67
	20160426	1.89	-0.84	6.67	-5.19	38.7	29.37
	20160617	-6.87	-10.11	-11.95	-5.14	3.94	9.9
	20161002	3.51	5.85	1.34	6.31	-3.7	6.6
	20170406	7.6	6.67	8.25	11.07	7.16	16.82
	20170616	-0.73	-2.34	2.1	7.12	-4.88	7.53
	20171025	4.46	0.59	1.27	4.48	7.04	7.43
验证期	20180412	-2.54	1.82	15.25	3.61	3.52	26.67
	20180608	3.01	12.29	2.95	15.73	4.53	17.42
	20181007	2.56	-9.09	1.92	-12.31	-3.28	-7.48
	20190422	-15.56	-13.99	-22.17	-23.87	18.66	16.71
	20190611	0.43	-6.64	-7.07	-5.87	-6.73	-10.06
	20191029	0.58	0.26	7.86	17.50	32.26	26.67

由表2可知,10天预见期内,不论是SMA-LSTM模型还是LSTM模型,精度均高于新安江模型,SMA-LSTM模型优于LSTM模型。SMA-LSTM模型的最大1日径流量相对误差RE_{1d}和10日总径流量的相对误差RE_{10d}基本均在10%以内,其中,率定期的最大1日$|RE_{1d}|$平均值和10日$|RE_{10d}|$平均值分别为3.64%和3.89%,在验证期分别为4.11%和7.35%。LSTM模型的最大1日径流量相对误差RE_{1d}和10日总径流量的相对误差RE_{10d}均在20%以内,其中,率定期的最大1日$|RE_{1d}|$平均值和10日$|RE_{10d}|$平均值分别为5.35%和7.38%,在验证期分别为9.82%和13.15%。对新安江模型而言,有5场的最大1日RE_{1d}超出20%,有7场的10日RE_{10d}超出20%,最大1日$|RE_{1d}|$平均值和10日$|RE_{10d}|$平均值在率定期分别为13.40%和19.00%,在验证期分别为11.50%和17.50%。

从验证期中挑选4场径流过程,将SMA-LSTM、LSTM和XAJ模型的预报结果与实测过程(SC)进行对比,如图4所示。

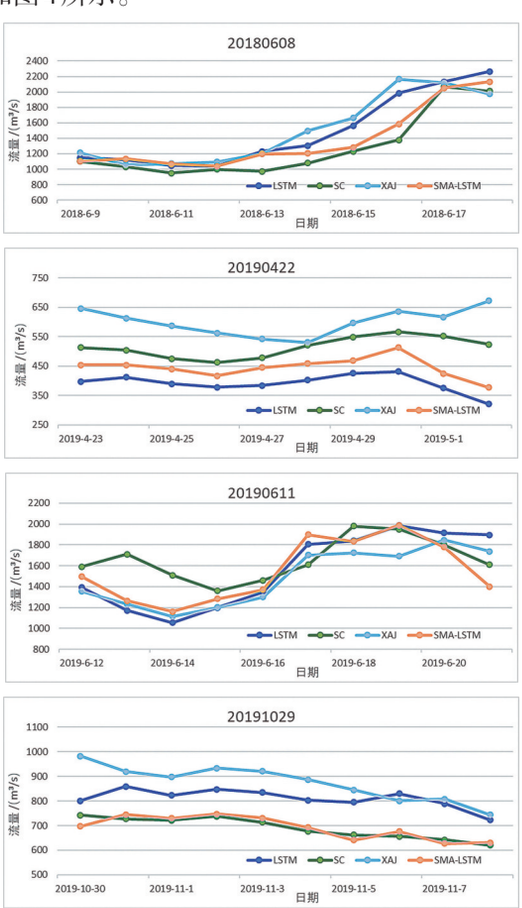

图4　新安江模型与SMA-LSTM模型径流过程模拟结果对比

Fig.4　Comparison of simulation between XAJ model and SMA-LSTM model in runoff process

从图4的四场径流过程可以看出,SMA-LSTM模型预报的径流过程与实测过程拟合的更好,整体而言,具有更高的精度。

3　主要结论

本文使用黏菌优化算法(SMA)对长短期记忆神经网络(LSTM)的超参数进行优化,构建了SMA-LSTM模型用于丹巴断面径流预报,并与LSTM模型、新安江模型预报结果进行对比分析,主要结论如下:

(1)SMA-LSTM模型在10天预见期内具有较高的预报精度。在对汛初(4月份)、汛中(6月份)与汛末(10月份)的预报中,SMA-LSTM模型的最大1日径流量和10日径流量预报相对误差,不管在率定期还是验证期均在±10%以内,两种指标的绝对值平均都不超过7%。SMA-LSTM模型对研究区的径流预报表现出较好的适用性。

(2)SMA-LSTM模型与新安江模型对比结果表明,SMA-LSTM模型预报精度整体高于新安江模型,预报的径流过程更接近实测过程。

本文仅是SMA-LSTM模型在大渡河丹巴站径流预报的一个初步探讨,对其他研究区域的适用性还需要进一步深入研究。

参考文献:

[1]　刘金涛, 宋慧卿, 张行南, 等. 新安江模型理论研究的进展与探讨[J]. 水文, 2014, 34(1):1-6.

[2]　汤成友, 项祖伟, 缪韧, 等. 水箱模型在大尺度流域实时洪水预报模型研制中的应用[J]. 水文, 2007.

[3]　张爱玲, 王韶伟, 汪萍, 等. 基于SWAT模型的资水流域径流模拟[J]. 水文, 2017, 37(5):38-42.

[4]　韩潇, 张亚萍, 周国兵, 等. 基于VIC模型的涪江流域径流模拟[J]. 水文, 2022, 42(5):76-81.

[5]　黄华平, 郦于杰, 王栋, 靳高阳. 三峡水库中长期径流预测及不确定性分析研究[J]. 中国农村水利水电, 2022(3):80-85.

[6]　蓝羽栖, 张尹, 农振昌, 韦永江. 基于统计模型的西江枯季中长期径流预报研究[J]. 人民珠江, 2022, http://kns.cnki.net/kcms/detail/44.1037.TV.20221011.0911.008.html

[7]　LIANG Z, TANG T, LI B, et al. Long-term streamflow forecasting using SWAT through the integration of the random forests precipitation generator: case study of Danjiangkou Reservoir[J]. Hydrology Research, 2017, 49(5):1513-1527.

[8]　陶思铭, 梁忠民, 陈在妮, 等. 长短期记忆网络在中长期径流预报中的应用[J]. 武汉大学学报(工学版), 2021, 54(1):21-27.

[9]　HOCHREITER S, SCHMIDHUBER J . Long Short-Term Memory [J].

(下转第56页)