水文模拟替代模型方法对比探究

李小兰,曾献奎,王 栋,吴吉春
(南京大学地球科学与工程学院,江苏 南京 210023)

摘要:传统的水文模型参数识别需要多次调用水文模型,从而导致严重的计算负荷。替代模型具有和原始水文模型几乎相同的模拟精度且运行时间可以忽略不计,从而有效解决参数识别中的计算负荷问题。以长江流域上游水文模拟为案例分析,系统对比了当前几种常用的替代模型方法,如稀疏网格方法、Elman-NN方法、RBF-NN方法,为大尺度水文模拟的替代模型构建提供一定的参考依据。

关键词:水文模型;稀疏网格;神经网络
中图分类号:[TV11] **文献标识码:**B **文章编号:**1007-7839(2022)08-0057-0005

Comparative study of surrogate model methods for hydrological simulation

LI Xiaolan, ZENG Xiankui, WANG Dong, WU Jichun
(*School of Earth Science and Engineering, Nanjing University, Nanjing* 210023, *China*)

Abstract: Traditional parameter identification of hydrological models usually needs to call the hydrological model many times, which leads to the problem of computational load. The surrogate model has almost the same simulation accuracy as the original model, and the running time can be ignored. Therefore, the surrogate model can solve the problem of excessive calculation load. This study takes the hydrological simulation in the upper reaches of the Yangtze River Basin as a case study, and compares several commonly systematically used surrogate model methods, such as sparse grid method, Elman-NN method and RBF-NN method, which can provide a certain reference basis for the construction of surrogate models for large-scale hydrological simulation.

Key words: hydrological model; sparse grid; neural network

水文模型是定量描述水文过程和认识水文要素响应机制的重要工具。随着计算机技术的迅速发展,分布式水文模型受到广泛重视,已在全球范围内的水资源、水环境与水生态领域得到成功应用[1]。随着流域水安全问题受到越来越多的关注,大尺度水文模拟成为流域水问题综合整治的重要工具。通常情况下,大尺度水文模拟是指研究空间尺度面积大于10 000 km^{2}或长度大于100 km的水文模拟[2],其具有水文过程复杂、模型运行耗时长、参数多等特点。参数识别是进行水文模拟的重要环节,通常需多次调用水文模型,如几千至几万次,而大尺度水文模拟运行一次一般需几小时至几天,从而导致严重的计算负荷问题[3]。

替代模型是指具有和原始模型几乎相同的模

收稿日期:2022-03-16
基金项目:江苏省水利科技项目(2020038);国家重点研发计划(2016YFC0402802)
作者简介:李小兰(1998—),女,硕士研究生,主要从事水文学与水资源研究。E-mail:mg1929057@smail.nju.edu.cn
通信作者:曾献奎(1985—),男,副教授,博士,主要从事地下水数值模拟研究。E-mail: xiankuizeng@nju.edu.cn

拟精度并且运行时间可以忽略不计的模型，用于替代原始的水文模型，是解决水文模拟参数识别计算耗时问题的有效手段。替代模型方法已广泛用于水文领域，如水文模型的校正、多目标优化、参数敏感性分析[4]、不确定性分析[5]。本次研究选择当前主要的3种替代模型方法，如稀疏网格（Sparse Grid，SG）、Elman-NN、RBF-NN，以长江流域上游水文模拟为案例，从替代成本、替代精度等方面系统对比分析了不同替代模型方法的特点。研究成果可为大尺度水文模拟的替代模型的构建及参数识别过程提供参考。

1 研究方法

1.1 VIC模型

VIC（Variable Infiltration Capacity）模型是由Liang等[6]开发的基于物理机制的大尺度分布式水文模型，已广泛应用于全球范围内的径流模拟、气候变化影响和水文变异性研究。该模型将空间分布网格化，每个网格具有对应的地表高程、土壤性质、植被覆盖、降水、气温等信息。VIC模型的模拟过程分为产流和汇流2个阶段，产流阶段每个网格独立计算天气、土壤、地形、植被综合作用下的径流和基流，汇流阶段将各个网格的径流深转化成流域出口断面流量。

1.2 稀疏网格替代模型方法

稀疏网格（Sparse Grid，SG）技术是一种基于Smolyak规则的分层拉格朗日插值算法，最早由Smolyak在1963年提出[7]。SG的基本原理为在参数分布空间生成插值节点，再进行拉格朗日插值。

1.2.1 维数局部自适应稀疏网格

维数局部自适应稀疏网格（Dimensional Adaptive-Local Adaptive-SG，LA-DA-SG）是维数自适应和局部自适应的耦合技术[8]。维数自适应的原理是不断找出对替代对象有显著影响的级数向量，并生成对应的插值节点，但当目标函数变化区域集中在较小参数空间区域时效率较低。局部自适应的原理是不在已满足替代精度的父节点生成子节点，其对于线性程度高的区域效率高，但不能考虑不同维度级数向量的敏感性。因此，将两种自适应方法耦合可以显著提高替代模型的构建效率。DA-LA-SG的应用步骤是先对替代对象进行维数自适应，再对生成的插值节点进行局部自适应。

1.2.2 优化自适应稀疏网格

优化自适应稀疏网格（optimized-DA-LA-SG，O-DA-LA-SG）是在DA-LA-SG基础上，将RPSO（斥力粒子群优化）算法用于识别替代对象的关键区（如极值区），进行针对性的SG插值节点分布，利用RPSO获取的极值点来定义极值区域的范围。

替代对象$f(x)$的极大值区域Γ_δ可表示为

$$\Gamma_\delta=\left\{x\in\Gamma\left|\frac{f(x)}{\max_{x\in\Gamma}f(x)}\geqslant\delta\right.\right\} \tag{1}$$

式中：δ为阈值，一般取0.001；Γ为参数x的分布空间；$\max_{x\in\Gamma}f(x)$为$f(x)$的最大值。

使用O-DA-LA-SG技术构建替代模型的过程包括初期和后期2个阶段，在初期不启动优化算法RPSO，即相当于DA-LA-SG技术。当全局替代误差超过某个阈值后进入后期，开始启动RPSO程序并搜寻替代对象的极值区域，分区实行局部自适应操作。在极值区域外设置较低的局部自适应标准，在极值区域内设置较高的局部自适应标准，进一步优化SG替代模型的节点分布，从而提高替代效率。

1.3 神经网络替代模型方法

神经网络（Neural Net，NN）是由大量神经元即节点相互连接、相互传递构成的复杂网络系统，是一种模仿生物神经网络的数学模型。每个节点代表一个输出函数，称为激励函数。

1.3.1 拉丁超立方抽样方法

采用拉丁超立方抽样（Latin Hypercube Sampling，LHS）方法来获取神经网络的训练样本。LHS是一种多维分层采样方法，LHS方法从变量$x(x_1,x_2,\cdots,x_i,\cdots,x_n)$（$1\leqslant i\leqslant n$）中抽取样本的过程如下。

（1）根据各变量的分布区间以及概率密度函数，将每个向量分量的子空间划分为m个不相交的层空间；

（2）根据各个变量的概率密度函数，从各个变量的独立层空间中随机抽取一个样本，每个变量获得m个样本；

（3）各变量x_i抽取的m个样本之间随机组成1个m组n维样本，共获得$(m!)^{n-1}$种组合方式；

（4）从$(m!)^{n-1}$种组合方式中以特定的方式筛选出指定数量的样本组合。

1.3.2 Elman神经网络

Elman神经网络（Elman-Neural Net，Elman-NN）是一种动态递归神经网络，其包含4层结构，分别是输入隐、隐含层、承接层和输出层。其中输入层和输出层均为1层，隐含层和承接层的层数一致且可设置为多层。承接层用来记忆隐含层前一时刻的输出值，从而使网络具有适应变化的能力，增加了

网络的全局稳定性。隐含层的输出通过承接层自联到隐藏层的输入，使其对历史数据具有记忆性，从而进行动态建模[9]。

1.3.3　RBF 神经网络

RBF神经网络(Radial basis function-Neural Net，RBF-NN)是一种高效多层前馈神经网络，运算速度快，具有较强的非线性映射能力。RBF-NN可以进行局部调整、相互覆盖接受域的局部逼近，同时训练方法快速易行，克服了局部最优问题，这些优点使得RBF神经网络广泛应用于很多领域，如分类、模型识别以及信号处理等领域。

RBF-NN结构主要包含3层，即输入层、隐含层和输出层，输入层到隐含层之间没有权值连接，输入向量直接被反馈到隐含层，通过激活函数进行非线性映射。RBF-NN的激活函数是具有多变量插值功能的径向基函数。径向基函数是一种沿径向对称的标量函数，是表示样本到数据中心之间的径向距离的单调函数，如高斯函数。隐含层到输出层之间有权值连接，为线性映射关系，即输出层的结果是隐含层结果的线性加权和。

1.4　替代模型精度的评价指标

(1)相对均方根误差N_{RMSE}

相对均方根误差N_{RMSE}的计算式为

$$N_{\mathrm{RMSE}}=\frac{\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/n}}{|f|_{\max}} \tag{2}$$

式中：n为测试点的数量；y_i和$\hat{y}_i$分别为第i(i=1，2，…，n)个测试点的原始模型输出值和替代模型输出值；$|f|_{\max}$为所有测试点替代模型输出的最大绝对值，N_{RMSE}越小表示替代模型的精度越高。

(2)决定系数R^2

决定系数R^2的计算式为

$$R^2=\frac{\left[\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})\right]^2}{\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]\left[\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2\right]} \tag{3}$$

式中：n为测试点的数量；y_i和$\hat{y}_i$分别为第i(i=1，2，…，n)个测试点的原始模型输出值和替代模型输出值；$\bar{y}$和$\bar{\hat{y}}$分别表示测试点的原始模型和替代模型输出值的平均值；R^2值为0到1之间，越接近1表示替代模型精度越高。

2　水文模型的构建

2.1　研究区概况

本次研究选择长江源头至干流寸滩站的区域为研究区，面积约95万km²，范围约为90.5°E~108.5°E、25.0°N~36.0°E，属于长江流域上游区域。本研究将建立该区域的VIC水文模型，进行替代模型方法的对比研究。

2.2　模型数据

高程数据来源于地理空间数据云平台的“SRTMDEMUTM 90M分辨率数字高程数据产品”。气象数据包括日降水量、最高气温、最低气温和风速数据等4项，本次研究使用区域内气象站(1961—1975年)气象数据。土壤参数数据包括饱和导水系数、田间持水量、土壤水扩散系数等多个用来代表土壤质地类型、土壤颗粒配比的参数，本文中的土壤参数数据来自空间分辨率为5′的全球土壤数据库。植被覆盖数据包括植被类型数目、比例、根系分布和逐月叶面积指数(LAI)，这些数据采用陆面覆盖类型资料。

2.3　模型构建

2.3.1　流域提取及网格剖分

根据研究区的数字高程数据和研究区内的水文站点坐标，借助ARCGIS软件，依次通过填洼、流向计算、汇流累积、捕捉倾泻点来提取研究流域范围。模型单元格剖分大小设置为0.5°×0.5°，研究区总共被划分为324个网格。

2.3.2　模型输入

气象输入数据包括研究区内每个网格1961—1975年逐日的降水、最高气温、最低气温和风速数据。采用反距离加权法将原始气象数据插值到0.5°×0.5°网格单元。土壤数据输入包含各网格的土壤参数，其中大多数参数可以直接获取或通过计算获取，如水力系数、饱和导水率、田间持水量、凋萎含水率。其中，深层土壤深度、可变下渗曲线方程的幂指数、基流最大流速、非线性基流产生的因子值和非线性基流产生时最大土壤含水量因子等参数为待识别参数。植被覆盖输入数据包括各网格的植被覆盖种类数目、各种类植被覆盖比例、叶面积指数等信息。根据Maryland大学全球1 km的陆面覆盖类型资料，计算各个网格内的植被类型及其在网格内所占的比例。

3　替代模型方法对比分析

3.1　替代对象

本次案例分析中，VIC模型的待识别参数包括可变下渗曲线方程的幂指数b、基流最大流速D_{smax}、非线性基流产生的因子值Ds、非线性基流产生时土壤含水量因子Ws、第二层土壤厚度$D2$、第三层

土壤厚度$D3$共6个参数，参数的识别范围如表1所示。

表1 待识别的VIC参数及其分布范围

参数	范围
b	0.200~0.600
Ds	0.010~1.000
D_{smax}/(mm·d^{-1})	8.000~18.00
Ws	0.100~1.000
$D2$/m	0.100~0.800
$D3$/m	0.100~1.000

通过贝叶斯方法（如MCMC等）识别模型参数的概率分布时，需要计算参数的似然函数L，用于搜索参数的概率分布空间，计算式为

$$L(y|\theta)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}[y-f(\theta)]^T\Sigma^{-1}[y-f(\theta)]\right\} \tag{4}$$

式中：y为观测数据；θ为模型的待识别参数；$f(\theta)$为模型模拟值；n为观测数据的个数；Σ为观测误差的协方差矩阵；$|\Sigma|$为其行列式。

因此，本次研究所建立的替代模型为6个待识别参数θ与对应似然函数L的响应关系。传统方法是通过运行水文模型$f(\theta)$获得模型输出，进而计算$L(\theta)$，而替代模型是直接构建$L(\theta)$~θ关系，省去运行模型的环节，从而解决模型运行计算耗时问题。

3.2 DA-LA-SG替代模型

DA-LA-SG替代模型的插值级数设置为L=9，采用500个样本测试替代模型的精度。替代模型的N_{RMSE}和R^2随样本数量的变化如图1所示，其中样本数量表示用于构建替代模型所需的运行原始水文模型的次数。随着替代成本的增加，替代精度逐渐升高，当样本数量为1 998个时，R^2= 0.9920，对应的N_{RMSE}=0.0134。

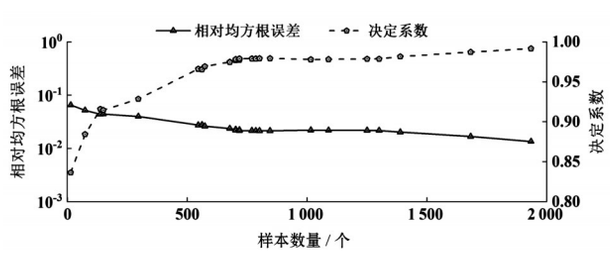

图1 DA-LA-SG替代模型的相对均方根误差和决定系数随样本数量的变化

3.3 O-DA-LA-SG替代模型

O-DA-LA-SG替代模型的插值级数设为L=9，全局N_{RMSE}≤0.025时开始启动优化程序。替代模型的N_{RMSE}和R^2随样本数量的变化如图2所示，随着替代成本的增加，替代精度在不断升高，当样本数量为1 469个时，R^2=0.9945，对应的N_{RMSE}=0.0115。

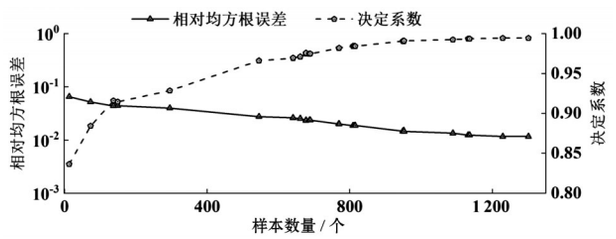

图2 O-DA-LA-SG替代模型的相对均方根误差和决定系数随样本数量的变化

3.4 Elman-NN替代模型

利用LHS方法抽取神经网络替代模型的训练样本，参数范围中每次抽取10个样本点，连续抽取200次，总共抽取2 000个点作为训练样本。Elman-NN替代模型的$NRMSE$和R^2随样本数量的变化如图3所示，随着替代成本的增加，替代精度在不断升高。当样本数量达到790个时，R^2= 0.9936，对应的N_{RMSE}=0.0143；当样本数量为1 205个时，R^2=0.9971，对应的N_{RMSE}=0.0095；当样本数量为1 695个时，R^2=0.9982，对应的N_{RMSE}=0.0075。

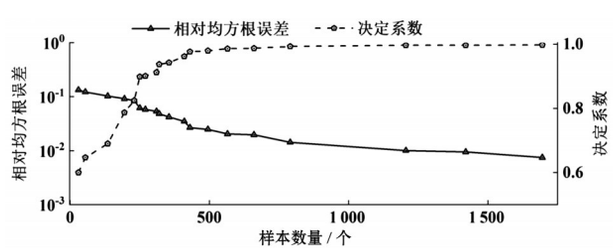

图3 Elman-NN替代模型的相对均方根误差和决定系数随样本数量的变化

3.5 RBF-NN替代模型

RBF-NN替代模型的样本抽样同Elman-NN，其$NRMSE$和R^2随样本数量的变化如图4所示，随着用于替代模型构建的样本增加，前期替代精度迅速升高，后期替代精度变化趋于平缓。当样本数量为850个时，R^2=0.96628，对应的N_{RMSE}=0.0275；当

样本数量为1 950个时，R^2=0.9752，对应的N_{RMSE}=0.0236。

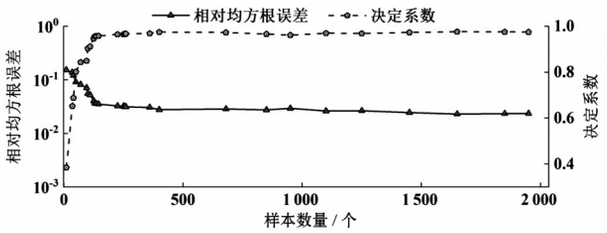

图4　RBF-NN替代模型的相对均方根误差和决定系数随样本数量的变化

3.6　替代模型方法对比分析

图5所示为DA-LA-SG、O-DA-LA-SG、Elman-NN和RBF-NNp这4种方法对于长江流域上游水文模拟替代模型表现的对比。从图5中可以看出，当样本数较少时（如400个），RBF-NN替代模型的替代误差降低最快，DA-LA-SG与O-DA-LA-SG替代模型的误差相对RBF-NN较大，Elman-NN替代模型的误差降低最慢。当样本数较大时（如1 000个），Elman-NN替代模型的误差快速降低，最先达到目标精度*NRMSE*=0.01，所需的替代成本最少；O-DA-LA-SG达到替代精度所需的成本小于DA-LA-SG，两者达到目标精度N_{RMSE}=0.01时的成本均小于RBF-NN。

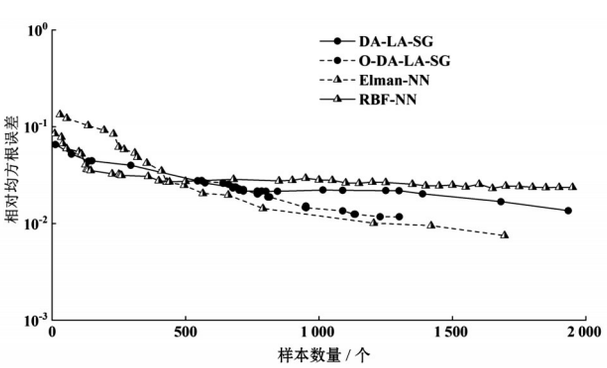

图5　4种替代模型的替代精度对比

4　结　论

本次研究以长江流域上游大尺度水文模拟为案例，系统对比分析了DA-LA-SG、O-DA-LA-SG、Elman-NN、RBF-NN等4种常见的替代模型方法。

（1）在样本数量相对较小的条件下，针对水文模拟替代模型的精度，随着样本数增加，RBF-NN替代模型的误差降低最快，DA-LA-SG与O-DA-LA-SG替代模型的误差下降速度居中，Elman-NN替代模型的误差降低最慢。

（2）在样本数量相对较多的条件下，针对水文模拟替代模型的精度，RBF-NN替代模型的误差降低相对平缓，O-DA-LA-SG替代模型的误差逐渐小于DA-LA-SG替代模型，并优于RBF-NN，Elman-NN替代模型的误差降低相对最快。

（3）根据实际条件下水文模拟的计算耗时特征，判断所能承担的用于构建替代模型的成本，从而选择相应最合适的替代模型方法，解决水文模拟参数识别中的计算耗时问题。

参考文献：

[1] PARK D, MARKUS M. Analysis of a changing hydrologic flood regime using the Variable Infiltration Capacity model [J]. Journal of Hydrology, 2014(515): 267-280.

[2] BECKER A. Criteria for a hydrologically sound structuring of large scale land surface process models [M]//O'Kane J P. Advances in Theoretical Hydrology. Amsterdam: Elsevier, 1992: 97-111.

[3] COUSQUER Y, PRYET A, ATTEIA O, et al. Developing a particle tracking surrogate model to improve inversion of ground water Surface water models [J]. JOURNAL OF HYDROLOGY, 2018(558): 356-365.

[4] HU Y, GARCIA-CABREJO O, CAI X, et al. Global sensitivity analysis for large-scale socio-hydrological models using Hadoop [J]. Environmental Modelling & Software, 2015(73): 231-243.

[5] TSOUKALAS I, MAKROPOULOS C. Multi objective optimisation on a budget: Exploring surrogate modelling for robust multi-reservoir rules generation under hydrological uncertainty [J]. Environmental Modelling & Software, 2015(69): 396-413.

[6] LIANG X, LETTENMAIER D P, WOOD E F, et al. A simple hydrologically based model of land surface water and energy fluxes for general circulation models [J]. Journal of Geophysical Research-Atmospheres, 1994, 99(D7): 14415-14428.

[7] SMOLYAK S. Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions [J]. DokladyAkademiinauk SSSR, 1963, 4(5): 240-243.

[8] 高鑫宇，曾献奎，吴吉春. 基于改进稀疏网格替代模拟的地下水DNAPLs运移不确定性分析[J]. 水文地质工程地质, 2020,47(1):1-10.

[9] ELMAN J L. Finding Structure in time [J]. Cognitive Science, 1990, 14(2): 179-211.