# 三水源新安江模型参数不确定性分析 PAM 算法

程春田, 李向阳
(大连理工大学土木水利学院,辽宁 大连 116024)

[摘要] 针对水文模型参数不确定性分析常用方法收敛速度缓慢,容易陷入参数空间局部最优区域等问题,提出了 PAM (parallel adaptive metropolis) 算法;对三水源新安江模型参数不确定性进行分析研究。实例研究表明显著提高了计算速度和求解质量,参数后验分布结果为区间预报提供了条件。

[关键词] 水文模型; 参数不确定性分析; MCMC; PAM; 并行计算
[中图分类号] TV123 [文献标识码] A [文章编号] 1009 - 1742(2007)09 - 0047 - 05

## 1 引言

"等效性" (equifinality) 是流域水文模型参数率定过程中经常碰到的现象[1],给洪水预报应用带来选择上的困难,因而水文模型参数不确定性分析成为近十多年国际水文学研究的热点和难点问题,并取得了一些国际上较有影响的成果,主要包括 Beven 等人提出的 GLUE (generalized likelihood uncertainty estimation) 方法[1],Thiemann 等人提出 BaRE (bayesian recursive estimation) [2]方法以及马尔可夫链蒙特卡罗 (Markov Chain Monte Carlo, MCMC)[3~5]方法等。其中自适应 MCMC 方法因其简单易用、遍历性强以及模拟复杂问题的能力,已成为水文模型参数不确定性分析领域一个活跃的研究方向。然而,如果 MCMC 方法的初始样本及转移概率密度选得不好,往往会导致算法收敛速度缓慢,且易于陷入参数空间局部最优区域。特别是对于参数较多的水文模型,MCMC 方法通常需要经过很多次的循环抽样才能达到收敛,其时间开销往往十分惊人。为了提高 MCMC 方法的计算速度和求解质量,有效地对我国广泛应用的三水源新安江模型参数及预报不确定性进行分析研究,笔者提出了 parallel adaptive metropolis (PAM) 算法,并以湖南省双牌水库为背景,在 Windows XP 微机集群上以并行计算的方式对三水源新安江模型参数及预报不确定性进行分析。

## 2 PAM 算法

### 2.1 贝叶斯方法与 MCMC 抽样

贝叶斯理论为水文模型参数不确定性分析提供了一个方法性框架,该方法能将有关参数的先验知识和实际观察样本数据相结合,结果以参数空间概率分布的形式表示,即参数后验分布。按照贝叶斯理论,水文模型参数为参数空间内由概率密度表示的某种分布,而不是单一的参数组合。根据贝叶斯公式:

$$p(\theta|Q) = \frac{p(\theta)L(Q|\theta)}{\int p(\theta)L(Q|\theta)\mathrm{d}\theta} \qquad (1)$$

式中 $\theta$ 为模型参数向量,$\theta \in \Theta \subseteq R^n$,$R^n$ 为 $n$ 维欧式空间;$p(\theta)$ 为参数先验分布,表示关于模型参数的先验知识,可以通过对历史资料的分析得到,或者参考气候、地质、植被条件相似且具有详细水文资料的流域,在缺少先验知识的情况下一般假定为参数空间内的均匀分布;$L(Q|\theta)$ 为参数 $\theta$ 的似然值,表示

[收稿日期] 2006 - 01 - 19;修回日期 2006 - 10 - 24
[基金项目] 国家自然科学基金资助项目(50479055)
[作者简介] 程春田(1965 - ),男,湖北孝感市人,大连理工大学教授,博士生导师,主要研究方向:数字防汛减灾与水电站及电网经济运行研究,E-mail: ctcheng@dlut.edu.cn

水文模型输出序列和实测序列的吻合程度;$p(\theta|Q)$为参数后验分布。

假定水文模型预报残差系列相互独立、同方差、正态分布,则参数似然函数可以定义为

$$L(Q|\theta) = (2\pi\sigma^2)^{-N/2}\prod_t^N \exp\left\{-\frac{|Q_t - f(x_t;\theta)|^2}{2\sigma^2}\right\} \tag{2}$$

式中 $N$ 为序列长度;$Q_t$ 为时段 $t$ 的实测流量;$x_t$ 为时段 $t$ 的模型输入(包括降雨、蒸发等);$f(x_t;\theta)$为时段 $t$ 的模型输出流量;$\sigma$ 为残差序列方差。

如果最后的后验密度函数不是已知的分布时,可利用 MCMC 抽样方法求解该密度函数的近似分布。其基本思想是,首先构造一个合适的 Markov 链,然后使用 Monte Carlo 方法对此 Markov 链进行抽样,使得到的样本序列的极限概率分布收敛于后验分布 $p(\theta|Q)$。MCMC 方法的关键是针对特定的问题构造合适的转移概率密度,以保证按照转移概率密度抽取的样本快速有效地收敛于目标后验分布。

## 2.2 AM 算法

Haario 等人提出的 adaptive metropolis (AM)[6] 算法是一种有效的自适应 MCMC 算法,其主要特点是将转移密度定义为参数空间内多维正态分布形式,并以此为依据对参数空间进行随机抽样。在进化过程中,根据 Markov 链的历史抽样信息自适应的调整转移密度(即协方差矩阵),从而大大提高算法的收敛速度。AM 算法主要步骤如下:

1) 初始化,令 $i=0$。

2) 由下式计算转移密度的协方差矩阵

$$\boldsymbol{C}_i = \begin{cases} C_0 & i \leqslant t_0 \\ s_d Cov(\theta_0, \theta_1, \cdots, \theta_{i-1}) + s_d \varepsilon I_d & i > t_0 \end{cases} \tag{3}$$

式中 $C_0$ 为初始协方差;$s_d$ 为仅依赖模型参数 $\theta$ 维数 $d$ 的比例参数,其作用是保证新产生的样本具备一定的接受概率,$s_d$ 一般取 $(2.4)^2/d$;$\varepsilon > 0$ 为一较小的数,其作用是避免 $\boldsymbol{C}_i$ 过于单一;$\theta_0, \theta_1, \cdots, \theta_{i-1}$ 为参数的样本序列;$t_0$ 为初始段长度。

3) 由转移概率分布 $\theta \sim N(\theta_i, C_i)$ 随机生成一个样本 $\theta^*$。

4) 计算样本 $\theta^*$ 的接受概率

$$\alpha = \min\left\{1, \frac{L(Q|\theta^*)p(\theta^*)}{L(Q|\theta_i)p(\theta_i)}\right\} \tag{4}$$

式中 $p(\theta)$为 $\theta$ 的先验分布;$L(Q|\theta)$为似然函数。

5) 生成一均匀分布随机数 $u \sim U[0, 1]$。

6) 如果 $u < \alpha$,则 $\theta_{i+1} = \theta^*$,否则 $\theta_{i+1} = \theta_i$。

7) $i = i+1$,重复步骤 2~7。经过多次抽样以后,得到的样本序列 $\theta_0, \theta_1, \theta_2, \cdots$收敛于后验分布。

## 2.3 PAM 算法的实现

AM 算法的收敛及遍历效果在很大程度上取决于初始样本及初始协方差矩阵 $\boldsymbol{C}_0$ 的选取,如果初始样本及协方差矩阵选的不好,会导致算法的收敛速度缓慢,特别是对于具有多个局部最优值的优化问题,容易陷入参数空间局部最优区域[4]。为了避免这一情况,笔者提出了 PAM 算法来改进初始选择随机性对优化结果的影响。

PAM 算法的原理和并行遗传算法相差不大,因此其实现方式可参照并行遗传算法。并行遗传算法的实现模型主要分为主从式、粗粒度和细粒度模型。PAM 算法采用主从式模型,其拓扑结构如图 1 所示。整个种群分成若干子种群,并分配给不同子线程 Markov 链,各子线程在不同的处理器上以相对独立的方式并发执行 AM 进化计算,线程间的通信采用同步控制方式,并利用 JPVM 消息传递接口进行消息传递。即每经过一定的进化代,由主线程负责对各子线程进行数据交换,以达到各子线程一定程度上的信息共享。PAM 算法不但加速了 AM 算法的搜索过程,而且由于种群规模扩大和各子线程内样本相对隔离,使样本多样性得以丰富和保持,减少了收敛于局部最优区域的可能性,从而提高了求解质量和计算速度。PAM 算法的基本流程如图 2 所示。

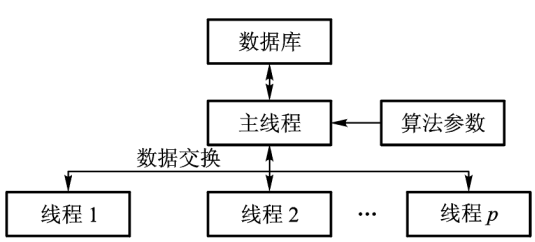


图 1 PAM 算法拓扑结构

**Fig. 1 Topology of the parallel adaptive metropolis algorithm**

## 2.4 收敛条件

从理论上讲,PAM 算法在 $i \to \infty$ 时必将达到收敛,但是在实际应用当中,必须确定 PAM 算法收敛于稳定的后验分布需要进行的抽样次数,即给出收敛判断条件。收敛判断是 MCMC 抽样方法的重要内容,Gelman 等提出比例换算系数 (scale reduction

factor)判断多序列收敛性[7],得到了广泛的应用。

$$\sqrt{R} = \sqrt{\frac{i-1}{i} + \frac{p+1}{pi}\frac{B}{W}} \tag{5}$$

$$B/i = \sum_{j=1}^{p} (\mu_j - \bar{\mu})^2/(p-1) \tag{6}$$

$$W = \sum_{j=1}^{p} s_j^2/p \tag{7}$$

式中 $i$ 为每个 Markov 链进化次数;$B/i$ 为各 Markov 链参数样本均值 $\mu_j$ 的方差;$W$ 为各 Markov 链参数样本方差 $s_j$ 的均值;$\bar{\mu}$ 为 $\mu_j$ 的均值。Gelman 等建议将比例换算系数 $\sqrt{R} < 1.2$ 作为多序列进化算法收敛判断条件。

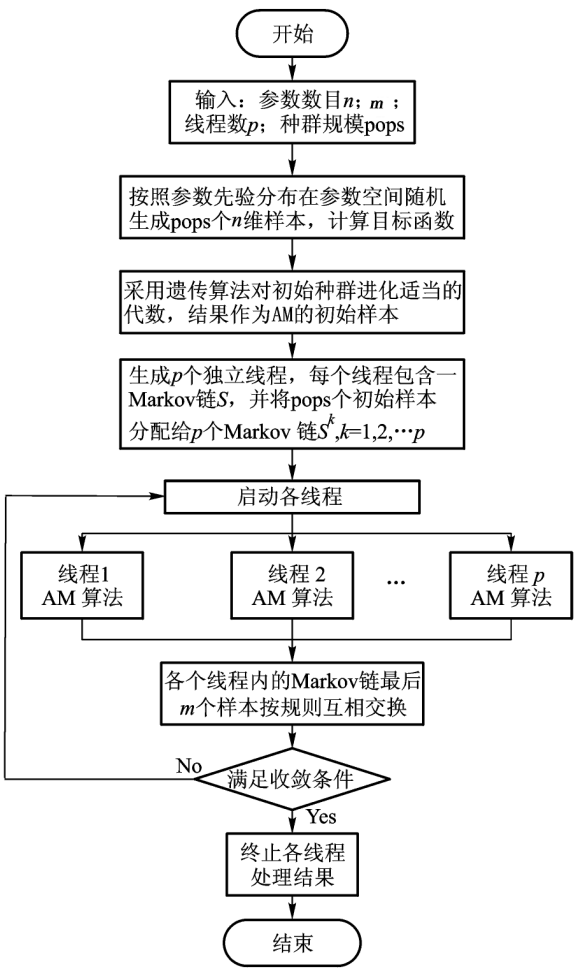


图 **2** **PAM** 算法流程图

**Fig. 2 Flowchart of the parallel adaptive metropolis algorithm**

## 3 实例分析

以湖南省双牌水库 1984 年～1995 年 30 场历史洪水作为实测资料,采用本文提出的 PAM 算法对三水源新安江模型参数不确定性进行研究,并对 PAM 算法效率同另外一种高效的自适应 MCMC 算法—SCEM－UA 算法进行比较。

PAM 算法运行的计算机软硬件环境为:Pentium 4 处理器 2.4GHZ,Windows xp 系统、SQL Server 数据库、Java 程序设计语言。PAM 算法采用的参数为:三水源新安江模型参数个数 $n = 17$;种群规模 pops = 400;线程数 $p = 8$;初始进化代数 $g = 10$;最大循环次数 $L = 50\ 000$;AM 算法进化代数 $m = 10$。比例换算系数的进化过程如图 3(b)所示,作为比较,图 3(a)给出了 SCEM－UA 算法的进化过程,表 1 给出了 PAM 算法与 SCEM－UA 算法效率比较。

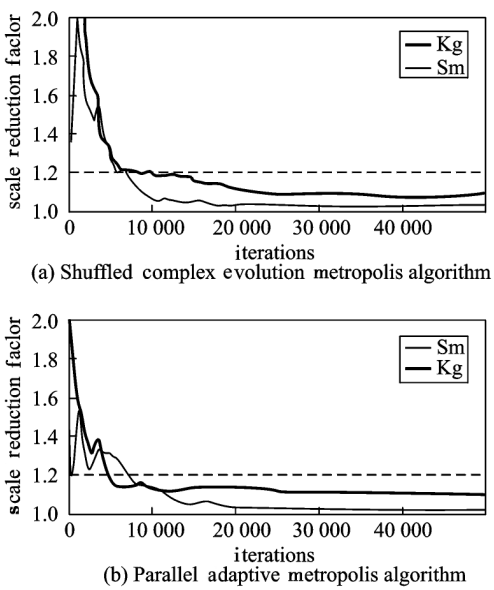


图 **3** 比例换算系数进化过程

**Fig. 3 Evolution of the Gelman and rubin scale reduction facto**

为了使比较更加直观,图中只列出了三水源新安江模型 Kg, Sm 两个参数的进化过程。SCEM－UA 算法的采用的参数为:分区数 $q = 8$;种群规模 $s = 400$ (SCEM－UA算法详见文献[4])。由图 3 及表 1 可以看出 PAM 算法比 SCEM － UA 算法达到收敛( $\sqrt{R} < 1.2$)所需的进化次数略少,由于采用

了微机集群并行进化计算方法,PAM 算法的运行时间比 SCEM – UA 算法大大缩短(从 SCEM – UA 算法的 2 780 min 缩短到 PAM 算法的 283 min)。

表 1　SCEM – UA 算法和 PAM 算法效率比较

**Table 1　Efficiency comparison between SCEM – UA and PAM algorithm**

| 比较项目 | SCEM – UA 算法 | | PAM 算法 | |
|---|---|---|---|---|
| | Kg | Sm | Kg | Sm |
| 达到收敛需要的模型运算次数 | 9 760 | 6 880 | 4 760 | 7300 |
| 总运行时间/min | 2 780 | | 283 | |

图 4 给出了由 PAM 算法抽样得到的双牌水库三水源新安江模型各参数边缘分布,每刻度对应的纵坐标标值表示该参数值落在该刻度与前一刻度值之间的概率。由图 4 可以看出,张力水容量 $W_m$、蒸发能力折算系数 $K$、表土自由水蓄水容量 $S_m$、地下水库的消退系数 $C_g$、河网蓄水量的消退系数 $C_s$、马斯京干法的单元河段的两个参数 $K_e$,$X_e$ 有明显的分布规律,而模型另外一些参数在有效空间内分布比较均匀,即不同的参数组合均能取得较好的效果,三水源新安江模型参数存在较大的不确定性。导致这一结果的原因可能是模型结构本身存在缺陷,如模型参数有冗余或相关性太强;历史水文资料的测量误差、初始条件设置的误差等。另外由于双牌水库三水源新安江模型时段长采用 3 h,且以场次洪水作为计算单位的水文模拟方式,导致了场次洪水间的误差累加,增加了模型参数的不确定性。

虽然通过 PAM 算法可以获得如图 4 所示模型各参数的边缘后验分布,但是实际水文预报的时候,真正具有意义的是模型各参数的组合,而不是单个模型参数。采用 PAM 算法达到收敛后的抽取的1 000个参数组样本分别对历史洪水进行模拟,每个时段对应各参数组可生成 1 000 个流量数据,然后根据这些流量数据求得该时段的流量分布函数,并求得该分布 5 % 和 95 % 的分位数作为水文预报的不确定性区间。图 5 给出了过去时间内 19900530 号洪水实测流量过程、水文预报 90 % 不确定性区间。由图 5 可以看出该场洪水大部分时段实测流量落在区间之内,因此,采用本文提出的 PAM 算法结果进行不确定性预报是可行的。另外也有少数几个时段实测流量位于区间之外,这主要是因为实测流量过程存在误差或者新安江模型结构需要改善。

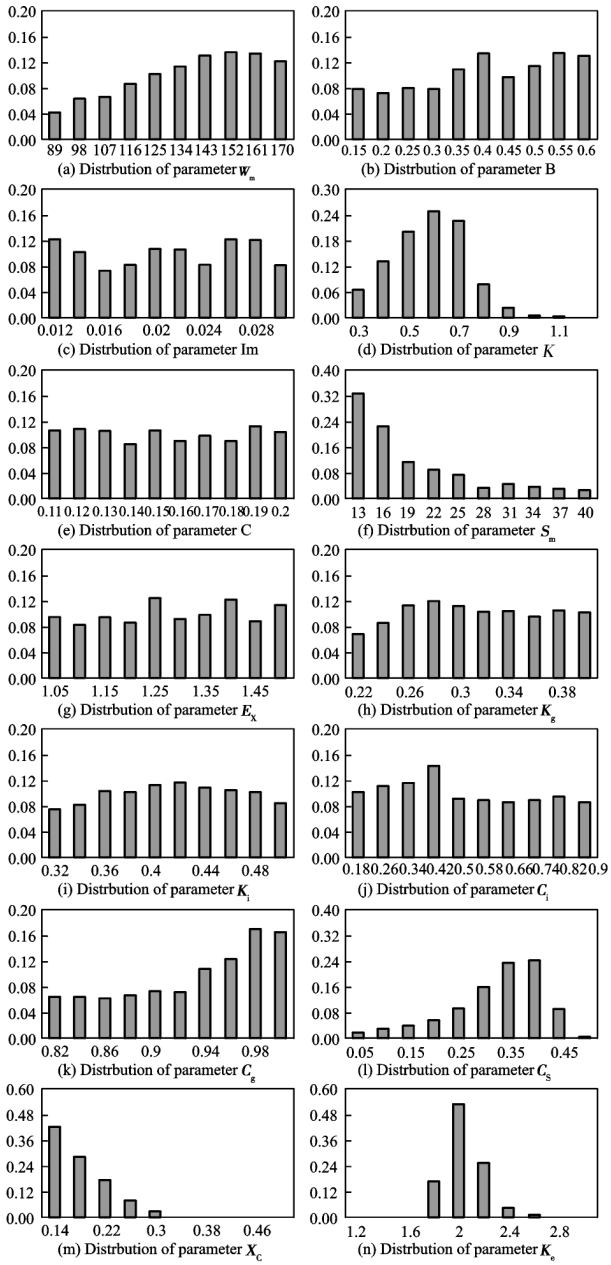


图 4　三水源新安江模型各参数边缘后验分布

**Fig. 4　Marginal posterior distributions of the Xin'anjiang model parameters**

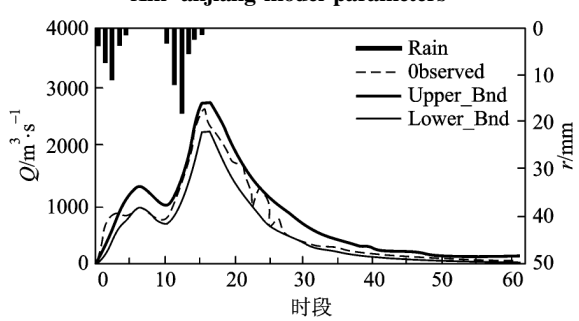


图 5　19900530 号洪水实测过程和 90% 预报区间

**Fig. 5　The observed flow and predictive uncertainty for flood 19900530**