**Advances in Marine Sciences** 海洋科学前沿, 2018, 5(3), 108-117
Published Online September 2018 in Hans. http://www.hanspub.org/journal/ams
https://doi.org/10.12677/ams.2018.53013

Hans汉斯

# Bayesian MCMC Method for the Solution of Data Assimilation

**Xiaoqun Cao, Junqiang Song, Bainian Liu, Hongze Leng, Shuo Ma, Weimin Zhang**

College of Meteorology and Oceanography, National University of Defense Technology, Changsha Hunan
Email: caoxiaoqun@nudt.edu.cn

Received: Aug. 27th, 2018; accepted: Sep. 11th, 2018; published: Sep. 18th, 2018

## Abstract

**In the framework of Bayesian theorem, a new method is proposed to find the solution of data assimilation problem based on Markov Chain Monte Carlo (MCMC) algorithm, which can quantify the posterior probability density function (PPDF) for initial states and model errors of nonlinear dynamical system. Firstly, the PPDF for unknown initial states and model errors which are derived with the Bayesian method and parameters to be estimated can be thought as the mathematic expectation of corresponding marginal PPDF. Secondly, taking the posterior probability as the invariant distribution, the Adaptive Metropolis algorithm is used to construct the Markov Chains of unknown initial states and model errors, respectively. That is, importance sampling of the posterior distribution is carried out. And the converged samples are used to calculate the mathematic expectation. So far, initial states and model errors of nonlinear dynamical system are estimated by Bayesian MCMC method successfully. Then, one and two-dimensional posterior distributions are constructed from the converged samples of initial states and model errors. And two-dimensional posterior distributions depict the interactions and correlations quantitatively between two different and arbitrary parameters. Finally, the results of numerical experiments show that the new data assimilation method can estimate initial conditions of nonlinear dynamical system very conveniently and accurately.**

## Keywords

**Nonlinear Model, Data Assimilation, Markov Chain Monte Carlo Method, Model Error**

# 基于贝叶斯MCMC方法的资料同化技术研究

**曹小群，宋君强，刘柏年，冷洪泽，马　烁，张卫民**

国防科技大学气象海洋学院，湖南 长沙
Email: caoxiaoqun@nudt.edu.cn

收稿日期：2018年8月27日；录用日期：2018年9月11日；发布日期：2018年9月18日

文章引用: 曹小群, 宋君强, 刘柏年, 冷洪泽, 马烁, 张卫民. 基于贝叶斯 MCMC 方法的资料同化技术研究[J]. 海洋科学前沿, 2018, 5(3): 108-117. DOI: 10.12677/ams.2018.53013

## 摘 要

**在贝叶斯理论框架下，提出基于马尔科夫链蒙特卡罗(MCMC)算法估计非线性模型初始状态和模式误差概率密度分布的一种新方法。首先利用贝叶斯方法，导出了非线性动力系统中未知初始状态和模式误差分布规律的后验概率密度函数(PDF)，将每个参数的后验边缘PDF的数学期望当作未知参数估计值。其次采用自适应Metropolis算法以后验PDF分布为极限不变分布来构造Markov链，即对未知参数进行重要性抽样，并利用收敛后的样本序列计算数学期望，从而得到初始状态和模式误差的估计值。然后利用初始状态和模式误差样本序列定量计算了未知参数的一维后验分布和相互之间的二维后验分布，后者定量描述了初始状态和模式误差之间的相关关系。最后通过数值试验结果说明该方法能有效地估计非线性动力系统的初始条件，具有较好的同化效果。**

## 关键词

**非线性模型，资料同化，马尔科夫链蒙特卡罗方法，模式误差**



## 1. 引言

资料同化是一种将观测资料融合于数值模式的分析技术，目的是通过有效利用一切信息(包括观测数据、背景场、模式以及相应的误差统计等)为非线性动力预报模型提供最优初始值[1] [2] [3] [4]。资料同化是一种具有普适性的通用技术，在海洋预报、气象预报、地震预测、电离层建模和流体等领域有广泛应用，因此对其方法的研究具有重要价值[5]-[12]。目前先进同化方法主要分为两大类：一类是变分资料同化方法[3] [4] [10] [11] [12]，另一类是顺序资料同化方法[5] [7] [8] [9]。顺序同化方法主要包括卡尔曼滤波(KF)、变形卡尔曼滤波、集合卡尔曼滤波(EnKF)和粒子滤波等。变分同化方法主要包括三维和四维变分资料同化方法(3D/4D-Var)，后者是前者在时间维上的扩展且同化效果更好。随着高性能计算机能力的提高和精细化数值预报需求增大，混合资料同化技术开始出现，例如：集合变分资料同化方法、集合最优插值同化方法、集合卡尔曼滤波和变分同化的混合方法，等等。虽然 4D-Var 是目前国际上气象和海洋业务预报中最先进和应用最成功的同化方法，但有效求解 4D-Var 问题时需要引入切线性和伴随模式，利用自动微分工具或手工编码方法仍然无法开发出完美无缺的伴随模式[13] [14] [15]。因此，众多的科研工作者一直都在探索和研究新资料同化技术[2]-[12]。

资料同化问题是典型的反问题，和正问题主要研究解的性质和数值求解方法等不同，反问题是通过试验或运行中的观测资料反求模型的未知参数：模式初始值、模型参数和模式误差等，从而使模型预测尽量准确或接近观测资料。因为观测量与未知参数之间常常不存在显式的直接关系，同时由于观测不准确、不充分和系统非线性等特征，所以导致反问题求解经常是不适定的。即解不一定存在、即使解存在也不唯一、在解存在唯一条件下也不稳定(即解不连续依赖于观测数据) [1]。因此非线性动力系统资料同化问题的求解必须采用特殊方法[4]-[9]。本文在贝叶斯理论的基础上，提出基于马尔科夫链蒙特卡罗(Markov Chain Monte Carlo，简称 MCMC)算法[16]来定量计算资料同化中初始值和模式误差的概率密度分布。综合利用贝叶斯方法和 MCMC 算法求解资料同化问题，具有以下优点：1) 能方便地将各种先验

信息融合到资料同化问题的求解过程中，减小问题不确定性；2) 与确定性算法不同，反问题的不适定性不再是 MCMC 算法要考虑的问题，且计算获得的是全局最可能解，而在变分资料同化方法中，如果背景场准确性不够高，那么最优化算法可能陷入目标函数局部极小值；3) 能对定义在高维空间且复杂的概率分布密度函数进行数值计算，而确定性方法无法解决此类问题；4) MCMC 算法通过构造 Markov 链在初始状态和模式误差所构成的参数空间进行重要性采样，最后获得的初始状态样本序列之间是平等关系，因此能提供给集合预报系统作为初始场集合；而变分资料同化极小化计算是一个逐步寻优过程，后面的迭代值要优于初始或中间值，即最终的分析场在理论上是唯一最优的。

本文首先利用贝叶斯公式推导了非线性动力系统需要估计的初始状态和模式误差分布规律的后验概率密度函数，参数估计值被认为是对应一维边缘后验分布的数学期望。接着采用自适应 Metropolis 算法[17]以后验概率分布为极限不变分布来构造 Markov 链，即对未知参数进行重要性抽样，并截取收敛后的样本序列计算数学期望，从而得到初始状态和模式误差的估计值；利用初始条件和模式误差样本集合定量计算了每个未知参数的一维边缘后验密度分布以及两个参数之间的二维边缘后验分布，后者定量描述了初始状态值和模式误差之间的相互关系。最后进行了计算机模拟，数值试验结果表明：贝叶斯 MCMC 方法能有效地估计非线性动力模型的初始条件，即具有较好的同化效果。

## 2. 贝叶斯理论

首先考虑如下非线性动力系统的资料同化问题：

$$\begin{cases} \dot{\boldsymbol{x}}(t)=\mathbf{F}\left[\boldsymbol{x}(t),t\right]+\boldsymbol{w}, \quad t\in[0,T] \\ \boldsymbol{x}\big|_{t=0}=\boldsymbol{x}_0, \end{cases}$$

其中，$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\in R^n$ 表示随时间变化的状态向量，$\boldsymbol{w}=(w_1,w_2,\cdots,w_n)^{\mathrm{T}}\in R^n$ 表示不随时间变化的常数型模式误差向量，$\boldsymbol{x}_0$ 是初始状态向量，同时在 $[0,\mathrm{T}]$ 时间段内分布有一系列观测量 $\boldsymbol{y}\in R^m$。资料同化的目标是利用非线性动力预报模式(1)提取和融合观测信息，从而有效地估计由初始状态和模式误差参数构成的未知参数向量 $\mathbf{m}=(\boldsymbol{x}_0,\boldsymbol{w})^{\mathrm{T}}$。

贝叶斯推理方法的基本思想是：认为整体分布中未知参数 $\mathbf{m}$ 本身就是随机向量，在对 $\mathbf{m}$ 进行统计推断时，除了使用样本信息外，还需要对 $\mathbf{m}$ 设置一个先验分布 $p(\mathbf{m})$。先验分布是在采集观测前就已存在的关于未知参数信息的概率表示。在获取观测资料以后，通过利用观测信息和贝叶斯公式可将未知参数的先验分布改进为后验分布。根据贝叶斯理论，参数的先验分布、观测样本信息和后验分布具有如下关系：

$$p(\mathbf{m}|\boldsymbol{y})=\frac{p(\boldsymbol{y}|\mathbf{m})\,p(\mathbf{m})}{p(\boldsymbol{y})} \tag{2}$$

其中 $\mathbf{m}^{\mathrm{T}}=(m_1,m_2,\cdots,m_N)$ 表示维数为 $N$ 的未知随机向量，$\boldsymbol{y}^{\mathrm{T}}=(d_1,d_2,\cdots,d_M)$ 是包含 $M$ 个观测量的列向量。$p(\mathbf{m})$ 表示未知随机向量的先验概率密度，$p(\boldsymbol{y}|\mathbf{m})$ 是条件概率密度，$p(\mathbf{m}|\boldsymbol{y})$ 是后验概率密度。因为观测量已经给出，所以 $p(\boldsymbol{y})$ 是一个与 $\mathbf{m}$ 无关的常数，于是(2)式可以写成：

$$p(\mathbf{m}|\boldsymbol{y})\propto p(\boldsymbol{y}|\mathbf{m})\,p(\mathbf{m}) \tag{3}$$

(3)式是进行贝叶斯推理的基础，各项的物理或统计意义解释如下：先验概率分布 $p(\mathbf{m})$ 包含了观测获取前就存在的未知随机向量先验信息，它是在进行贝叶斯推理时的一种必要信息。从反问题的角度，先验信息的引入增加了信息容量、减小了不确定性范围，从而可以部分克服反问题不适定性。先验信息主要来源于历史观测资料、经验和主观判断等，例如：变分资料同化中的背景场及其误差协方差矩阵就

是一种典型先验信息。条件概率密度 $p(\boldsymbol{y}|\mathbf{m})$ 又称似然(Likelihood)函数，也可以写作 $L(\boldsymbol{y}|\mathbf{m})$，包含了观测信息，即在有观测条件下未知参数的似然度信息。通俗地说，其表示了所估计的模式初始状态和模式误差与观测量之间的拟合程度，值越大表明拟合程度越好，反之越差。通过融合先验信息和观测信息后，就得到反映未知随机向量 $\mathbf{m}$ 整体信息的后验概率密度函数 $p(\mathbf{m}|\boldsymbol{y})$，它定义在初始状态和模式误差的整个解空间，表示问题的“完全”解。

在资料同化中，通过贝叶斯公式导出初始状态和模式误差等未知随机向量的后验分布表达式后，理论上可以获得未知参数的统计特征，例如：边缘分布、均值和方差等。但在实际应用中除了极简单情形，后验概率密度函数都无明确的数学表达式。另外，采用一般的数值积分方法(例如，Monte Carlo 方法)也存在极大困难：必须对整个未知向量空间进行随机抽样以获得代表性样本，计算量将随向量维数增加而呈指数增长。上述原因使得直接利用贝叶斯方法无法解决复杂的实际问题，但是随着马尔可夫链蒙特卡洛算法的新发展，该情形不断被改善。本文结合贝叶斯理论和 MCMC 算法对资料同化问题中的初始状态和模式误差参数进行估计。

## 3. MCMC 算法

通过上面的分析易知，虽然利用贝叶斯公式可以导出资料同化中初始状态和模式误差等未知参数的后验概率密度函数公式，但仍然无法求解资料同化问题。另外，考虑到观测误差和模式误差等不确定性因素的影响，单一“最可能”解存在巨大的局限性；而利用后验分布函数对解进行整体推断更具有合理性。换言之，研究重点不应该是通过最优化获得单一解，而是对解空间中最可能区域进行重要性抽样(importance sampling) [16]后，然后基于参数样本集合计算解的估计值和置信区间。马尔可夫链蒙特卡洛算法是一种直接模拟后验概率密度函数的方法[16]，通过自动搜索概率大值区域而对未知向量进行随机抽样，然后由所得抽样序列对每个未知参数进行各种整体性推断。

MCMC (Markov chain Monte Carlo)方法是现代统计计算中最重要的算法之一，通过 MCMC 方法可以得到复杂模型中众多物理参数的有效范围。MCMC 算法基于马尔科夫链的采样机制，可以对定义在高维随机向量空间 $\mathbf{M}$ 上无明确数学表达式的概率分布 $\boldsymbol{\pi}$ 进行有效抽样，基本思想是产生大量服从分布 $\boldsymbol{\pi}$ 的随机向量序列 $\{\mathbf{m}_1,\mathbf{m}_2,\cdots,\mathbf{m}_I\}$，其中 $I$ 为抽样数。如果向量序列满足马尔可夫性质：向量 $\mathbf{m}_{i+1}$ 的产生仅依赖于前一个向量 $\mathbf{m}_i$，而与过去时刻 $i-1,i-2,\cdots,1$ 的状态向量 $\mathbf{m}_{i-1},\mathbf{m}_{i-2},\cdots,\mathbf{m}_1$ 都无关，则该向量序列称为马尔可夫链。马尔可夫性质的另一种表述是：若抽样算法当前访问的是 $\mathbf{m}_j$ 点，则下一步访问另一点 $\mathbf{m}_i$ 的概率只依赖于 $\mathbf{m}_j$，而与先前访问的点无关。马尔科夫性质意味着抽样算法完全可由转移概率矩阵 $\boldsymbol{\pi}$ 描述，矩阵元素 $\pi_{ij}$ 表示算法在当前访问 $\mathbf{m}_j$ 的条件下接着将要访问 $\mathbf{m}_i$ 的条件概率。按照构造 Markov 链所用转移概率矩阵的不同，MCMC 方法的主要抽样算法有：Gibbs 抽样器算法、Metropolis-Hastings 算法和自适应 Metropolis 算法[17]。

本文采用自适应 Metropolis 算法以非线性动力预报系统初始状态和模式误差参数的后验分布为不变极限分布来构造 Markov 链。与传统的 Metropolis-Hastings 算法相比，自适应 Metropolis 算法不需要预先确定参数的推荐分布，而是由后验参数的协方差矩阵来估算参数分布[17]，从而大大提高计算效率。后验参数的协方差矩阵在每一次迭代后需要自适应地调整。如此，第 $i$ 步参数的推荐分布就是均值为 $\bar{m}_i$、协方差矩阵为 $\mathbf{C}_i$ 的多元正态分布[17]。协方差的计算公式如下面(4)式所示，在初始 $i_0$ 次迭代中，协方差矩阵 $\mathbf{C}_i$ 取固定值 $\mathbf{C}_0$，之后进行自适应更新。$\hat{m}_i$ 是未知模式参数向量 $\mathbf{m}$ 中某个元素在第 $i$ 次迭代的估计值。

$$\mathbf{C}_i=\begin{cases}\mathbf{C}_0, & i\le i_0\\ s_d Cov(\hat{m}_0,\cdots,\hat{m}_{i-1})+s_d\varepsilon\mathbf{I}_d, & i\ge i_0\end{cases} \tag{4}$$

其中，$\varepsilon=10^{-6}$，其引入是为了确保$\mathbf{C}_i$不成为奇异矩阵；$s_d$是比例因子，依赖于未知随机向量空间的维数$d$，目的是保证接受率在一个合适范围内，在本文中$s_d=(2.4)^2/d$。$\mathbf{I}_d$为$d$维单位矩阵。当进行第$i+1$次迭代时，由公式(4)可导出协方差的计算公式(5)。

$$\mathbf{C}_{i+1}=\frac{i-1}{i}\mathbf{C}_i+\frac{s_d}{i}\left(i\bar{m}_{i-1}\bar{m}_{i-1}^{\mathrm{T}}-(i+1)\bar{m}_i\bar{m}_i^{\mathrm{T}}+m_i m_i^{\mathrm{T}}+\varepsilon\mathbf{I}_d\right) \tag{5}$$

其中，$\bar{m}_{i-1}$和$\bar{m}_i$是前面$i-1$次和$i$次迭代的参数均值。自适应 Metropolis 算法的计算流程[17]如下：

1) 设定$i=0$，对所有未知参数变量进行初始化；

2) 随机量的生成和接受，构造 Markov 链：

a) 利用公式(4)计算协方差矩阵$\mathbf{C}_i$；

b) 产生服从高斯正态分布的推荐参数值$m^*\sim N(m_i,\mathbf{C}_i)$；

c)计算接受概率$\alpha=\min\left\{1,\dfrac{p(\mathbf{d}|m^*)p(m^*)}{p(\mathbf{d}|m_i)p(m_i)}\right\}$；

d) 产生服从均匀分布的随机数$u\sim U(0,1)$；

e) 若$u<\alpha$，则接受$m_{i+1}=m^*$，否则$m_{i+1}=m_i$。

3) 重复上面的步骤(a)~(e)，直到产生预先指定数量的样本为止。

## 4. 资料同化问题求解

下面以典型的海洋生物种群动力学模型为例，说明利用贝叶斯 MCMC 方法对资料同化问题进行求解的有效性。该非线性动力系统由如下的常微分方程组表示：

$$\begin{cases}\mathrm{d}N_1/\mathrm{d}t=aN_1-cN_1N_2+w_1,\\ \mathrm{d}N_2/\mathrm{d}t=bN_2+dN_1N_2-eN_2^2+w_2,\\ (N_1,N_2)|_{t=0}=(N_1(0),N_2(0)),\end{cases} \tag{6}$$

系统(6)表示了海洋生物二种群动力学模型，主要模拟种群之间捕食与被捕食的相互作用，刻画的是两个种群数量密度$N_1$、$N_2$随时间$t$的演化，其中$(a,b,c,d,e)$为已知的模型参数。数值试验中采用四阶龙格-库塔算法求解微分方程组，步长设置为$h$。观测量仿真过程如下：首先设置模式状态初值，然后任其演化至$Th$时刻，从而得到系统在离散时间序列$\{0,h,2h,\cdots,Th\}$上的状态量序列$(N_1(t_i),N_2(t_i))$，$i=0,1,2,\cdots,T$。状态量可直接作为观测量$(y_1(t_i),y_2(t_i))$，也可加入随机噪声得到更真实的模拟观测资料。本文资料同化的目标是在已知部分观测资料的条件下，采用贝叶斯 MCMC 方法估计非线性动力系统(6)的初始状态和模式误差参数。

下面利用第 1 部分中的贝叶斯公式推导系统(6)中未知参数的后验概率密度函数公式。由于系统(6)中影响模式状态量的四个参数$(N_1(0),N_2(0),w_1,w_2)$都是未知的，即需要对多参数进行联合反演，且只使用部分观测资料。根据贝叶斯公式(3)，且不考虑分母表示的常数项，则未知参数的后验分布$p(\mathbf{m}|\boldsymbol{y})=p[N_1(0),N_2(0),w_1,w_2|\boldsymbol{y}]$可以通过下式进行计算：

$$\begin{aligned}&p[N_1(0),N_2(0),w_1,w_2|\boldsymbol{y}]\\&=p[\boldsymbol{y}|N_1(0),N_2(0),w_1,w_2]\,p[N_1(0),N_2(0),w_1,w_2]\end{aligned} \tag{7}$$

公式(7)已经假设$\boldsymbol{y}=M(\boldsymbol{x}_0,\boldsymbol{w})+\omega$，其中$M(\boldsymbol{x}_0,\boldsymbol{w})$表示系统(6)的离散数值模式，$\omega$为包含观测误差的独立分布随机噪声。假定随机噪声$\omega$服从均值为零、标准偏差为$\sigma_o$的正态分布，即$\omega\sim N(0,\sigma_o^2)$。同

时假设由数值模式 $M$ 引进的离散等模拟误差包含在模式误差向量 $\boldsymbol{w}$ 中。通过上述假设，可以得到以下形式的似然函数：

$$L(\boldsymbol{y}|\mathbf{m})=L(\boldsymbol{y}|\boldsymbol{x}_0,\boldsymbol{w})=\frac{1}{\left(2\pi\sigma_o^2\right)^{n/2}}\exp\left[-\left\|\boldsymbol{y}-M(\boldsymbol{x}_0,\boldsymbol{w})\right\|^2\Big/\left(2\sigma_o^2\right)\right] \tag{8}$$

式中 $n$ 表示观测数量，$\|\cdot\|$ 表示欧几里得范数。

在贝叶斯推理中，认为未知向量 $\mathbf{m}$ 是随机向量，首先需要设定未知参数的先验分布。在本文中设定所有先验分布都是独立的均匀分布，初始状态和模式误差参数分别满足均匀分布：$U\left[N_1(0)\right]$、$U\left[N_2(0)\right]$、$U(w_1)$ 和 $U(w_2)$，则总的先验分布表示为：

$$p(\mathbf{m})=U\left[N_1(0)\right]U\left[N_2(0)\right]U(w_1)U(w_2) \tag{9}$$

均匀分布是一种最简单的先验分布，虽然只能指定未知参数变化的上下界，但是可以缩小对参数随机抽样的目标区域，有利于提高参数的估计精度。一般来说，可以通过经验知识和历史统计确定更复杂和准确的先验分布。自适应 Metropolis 算法的一个优点是对于 $\mathbf{m}$ 的任何先验分布都能够收敛于目标分布。由式(7)、(8)和(9)可得，在给定观测条件下资料同化问题未知参数的后验概率密度函数表示为：

$$p(\boldsymbol{x}_0,\boldsymbol{w}|\boldsymbol{y})=\frac{1}{\left(2\pi\sigma_o^2\right)^{n/2}}\exp\left[-\frac{\left\|\boldsymbol{y}-M(\boldsymbol{x}_0,\boldsymbol{w})\right\|^2}{2\sigma_o^2}\right]U\left[N_1(0)\right]U\left[N_2(0)\right]U(w_1)U(w_2) \tag{10}$$

## 5. 数值试验结果

在数值试验中，首先设定海洋生物种群动力模式初始状态量 $\left(N_1(0),N_2(0)\right)$、模式误差参数和时间积分区间 $[0,4]$，取积分步长 $h=0.01$ 和利用四阶龙格-库塔算法对模式进行积分，就可以得到系统在离散时间序列 $0h,1h,2h,\cdots,400h$ 上的标准状态值。试验中只抽取 $0h,20h,40h,\cdots,400h$ 时刻的状态量作为 $n=21$ 个观测数据，同时在所选状态量上叠加高斯型观测噪声 $N(0,\sigma_o)$。噪声的均值和标准偏差分别取为 0 和 $\sigma_o$。未知参数 $N_1(0)$、$N_2(0)$、$w_1$ 和 $w_2$ 的先验分布分别取如下形式：

$$U\left[N_1(0)\right]=\begin{cases}1/30 & -15<N_1(0)<15\\ 0 & \text{otherwise}\end{cases}, \tag{11}$$

$$U\left[N_2(0)\right]=\begin{cases}1/30 & -15<N_2(0)<15\\ 0 & \text{otherwise}\end{cases}, \tag{12}$$

$$U(w_1)=\begin{cases}1/20 & -10<w_1<10\\ 0 & \text{otherwise}\end{cases}, \tag{13}$$

$$U(w_2)=\begin{cases}1/20 & -10<w_2<10\\ 0 & \text{otherwise}\end{cases}。 \tag{14}$$

针对(10)式表示的后验概率密度函数，利用 MCMC 方法中的自适应 Metropolis 算法按照其计算流程中的步骤(a)~(e)构造每个未知参数的 Markov 链。未知参数初始值取对应先验均匀分布中的随机值，为了提高准确性，采用两次重要性采样，首先利用样本数为 1000 的 Markov 链获得一个比较准确的参数估计值。然后执行第二次重要性采样，进一步提高参数估计值。构造 Markov 链的过程实际上是在由先验分布界定的区间内对未知参数进行随机最优搜索。抽样过程中未知参数的每次更新，都要对非线性动力系统(6)式进行一次数值积分，以便计算后验概率密度函数的大小。图 1~图 3 显示了初始状态和第一个模式误差参数的 Markov 链，可以看出经过第一次重要性采样后马尔科夫链基本上达到收敛；取 Markov 链中序

号 2000~5000 之间的样本值计算数学期望，从而得到各个参数的估计值。

表 1 给出了在不同观测噪声水平条件下，新同化方法对海洋生物种群非线性动力系统(6)式初始状态和模式误差的估计值。从表中可知，在不存在观测误差时，未知初始状态的估计精度最高，初始状态量可以精确到小数点后第 3 位，表明了基于贝叶斯 MCMC 的资料同化方法估计非线性物理系统未知初始状态的有效性。另外，随着观测误差增加，资料同化结果的精度下降，但迭代估计结果仍然收敛到真实值附近；即使当观测误差的标准偏差为$\sigma_o = 0.2$ 时，虽然模式误差参数的估计误差较大，但是初始状态估计值仍然较接近真实值，可以精确到小数点后 1 位，说明基于贝叶斯 MCMC 的资料同化方法具有较强的抗噪声性能。图 1、图 2 和图 3 分别给出了初始状态 $N_1(0)$、$N_2(0)$ 和第一个模式误差参数 $w_1$ 的 Markov 链。从图中可以看出，经过第一阶段马尔科夫链 1000 次迭代后，第二阶段初始状态的马尔科夫链变化幅度较小，基本达到收敛。取 2000~5000 步的样本序列并采用后验均值方法计算参数估计值，分别得到 $N_1(0) = 8.0010$ 和 $N_2(0) = 1.0023$。分析图中所示的试验结果可得如下结论：采用贝叶斯 MCMC 方法估计非线性动力模式初始状态值具有较高的准确性和稳定性。对于模式误差参数，从图 3 中可知，相对于初始状态的 Markov 链，变化幅度增大，但基于 2000~5000 步 Markov 链值计算的后验均值非常接近真实值(见表 1)。模式误差参数 $w_2$ 的 Markov 链情形与 $w_1$ 类似，受篇幅所限不再给出。采用贝叶斯 MCMC 方法估计出初始状态和模式误差值后，代入海洋生物种群非线性动力系统(6)的数值模式后可以计算出状态量 $(N_1, N_2)$ 的预报轨迹。图 4 和图 5 分别显示了 $N_1$ 和 $N_2$ 的预报轨迹、观测值和离散时间点上的真实值，从图中可知，预报轨迹与观测结果、真实值吻合得很好，验证了用贝叶斯 MCMC 同化方法对多个未知参数进行联合估计的正确性。

**Table 1.** The results of data assimilation for nonlinear dynamical system (6) under different levels of observation noise
**表 1.** 不同观测噪声水平情况下非线性动力系统(6)的资料同化结果

| | $N_1(0)$ | $N_2(0)$ | $w_1$ | $w_2$ |
|---|---|---|---|---|
| 真实值 | 8.0 | 1.0 | 2.0 | 1.0 |
| $\sigma_o = 0.0$ | 8.0010 | 1.0023 | 7.9946 | 1.0017 |
| $\sigma_o = 0.1$ | 7.9260 | 0.9943 | 2.0687 | 0.9193 |
| $\sigma_o = 0.2$ | 8.0218 | 1.0129 | 2.1173 | 1.1769 |

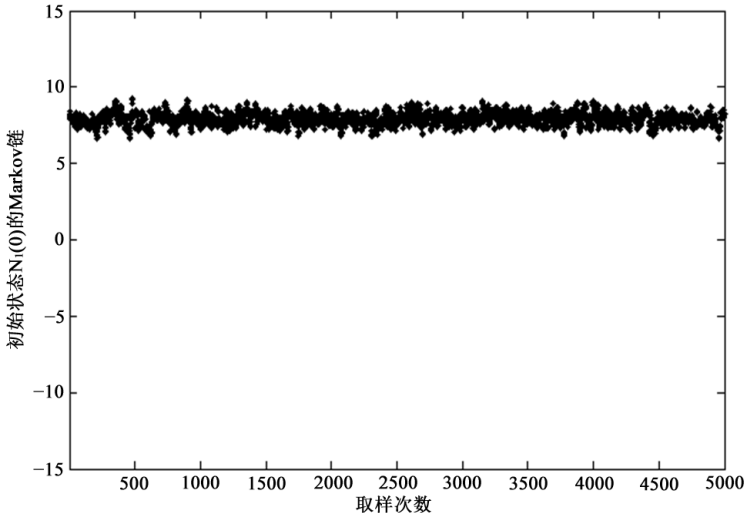


**Figure 1.** The Markov Chain of initial state variable $N_1(0)$
**图 1.** 初始状态 $N_1(0)$ 的 Markov 链

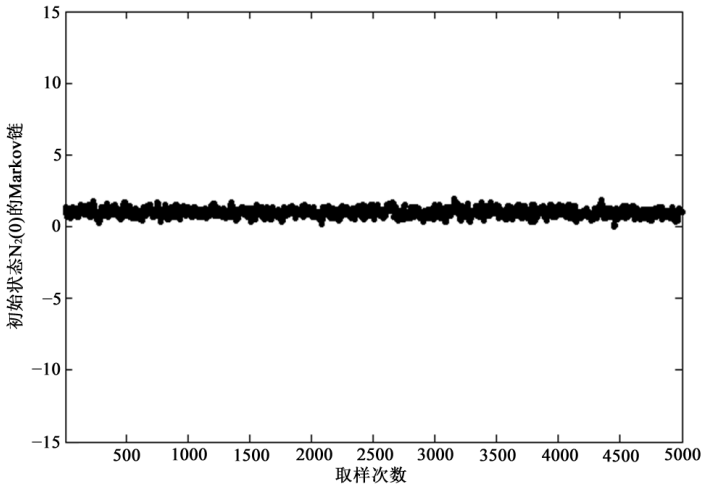


**Figure 2.** The Markov Chain of initial state variable $N_2(0)$

**图 2.** 初始状态 $N_2(0)$ 的 Markov 链

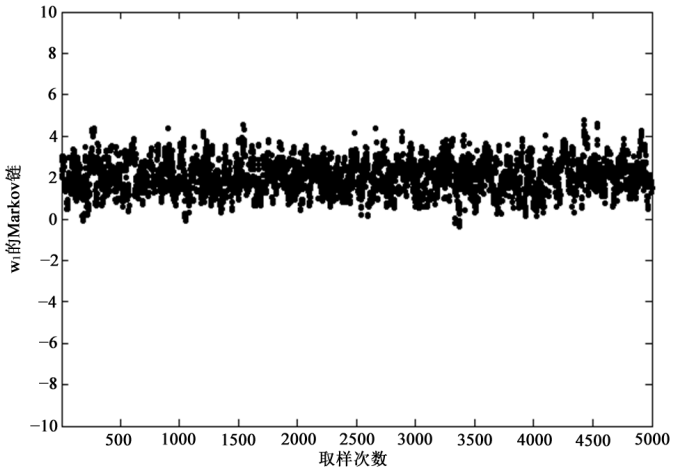


**Figure 3.** The Markov Chain of model error parameter $w_1$

**图 3.** 模式误差参数 $w_1$ 的 Markov 链

图 6 展示了在观测误差标准偏差取 $\sigma_o = 0.1$ 时初始状态和模式误差等 4 个自由参数的边缘后验分布。从图中可以看出所有参数的一维后验分布(对角线位置)都是类高斯分布，参数样本分布在一定的置信区间，没有出现发散情形，而且频率曲线的最高峰位置对应横坐标上参数真实值，这种情况说明观测资料对初始状态和模式误差具有很强的限制和约束作用，也就是贝叶斯 MCMC 资料同化算法能很好地将观测信息传递给未知自由参数，在逐次重要性采样中不断调整未知参数值，最终收敛到真实值附近。一维后验分布给出了资料同化问题的“完整解”，即各个参数的概率分布数值，通过后验均值方法能计算出准确的初始状态和模式误差(如表 1 第三行所示)。图 6 下三角每个二维后验分布子图表示了不同参数之间的相关关系，从图中可知：任意两个参数之间都具有较大的相关性，即初始状态和模式误差存在显著的相关作用，说明在资料同化过程中模式误差对初始状态值有重大的影响；任意两个未知参数之间的相关形态都不相同，即两个参数之间的相互影响和作用不相同；显示的相关函数呈现出近似椭圆状，即是类二维高斯分布，说明观测信息也对二维后验分布具有很强的限制和约束作用。

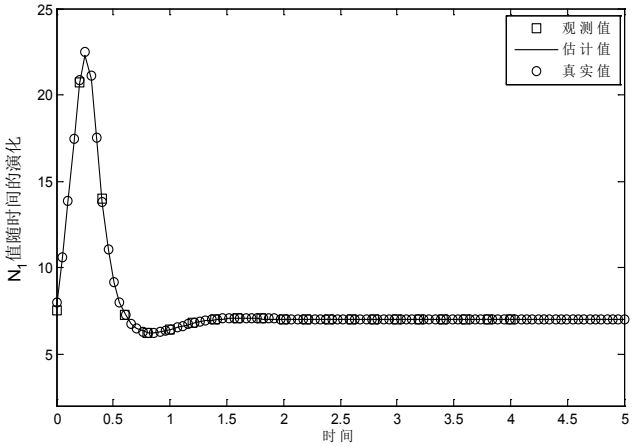


**Figure 4.** Comparison of the state variable $N_1$

**图 4.** 状态变量 $N_1$ 的比较图

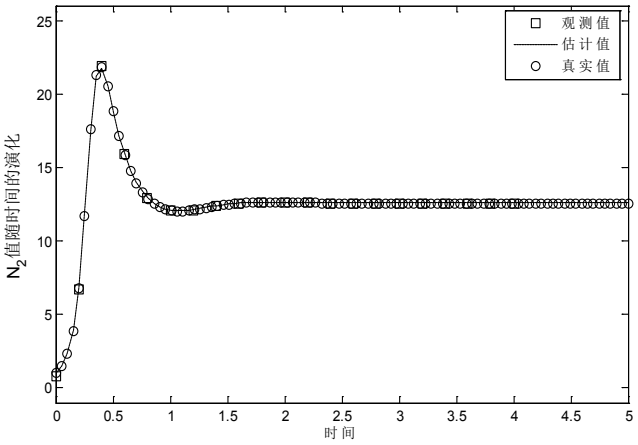


**Figure 5.** Comparison of the state variable $N_2$

**图 5.** 状态变量 $N_2$ 的比较图

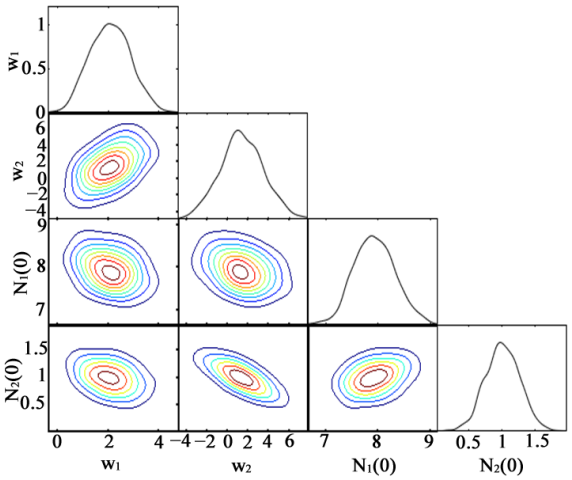


**Figure 6.** The marginal posterior distribution of model errors and initial state variables

**图 6.** 模式误差和初始状态的边缘后验分布

## 6. 结束语

本文在贝叶斯理论框架下，提出了一种基于 Bayesian MCMC 算法[16] [17]从观测信息中同时估计非线性模型初始状态和模式误差等未知变量的新同化方法。主要结论如下：1) 贝叶斯 MCMC 方法在不需要切线性/伴随模式的情况下，能准确地估计出模式初始状态和模式误差参数，验证了新方法的有效性。2) 与确定性同化(例如：变分资料同化)方法只能获得单一最优解不同，贝叶斯 MCMC 同化方法不但能获得作为分析场的后验均值，而且能定量计算出估计参数的一/二维后验分布。其中一维后验分布给出了同化结果的分布范围和取值频率，而二维后验分布定量刻画了不同参数之间的相关关系。3) 新同化方法获得的收敛初始状态样本序列之间是平等关系，因此能提供给集合预报系统作为初始场集合。综上所述，与基于伴随模式的变分资料同化等确定性方法[6] [7] [8] [9]一样，Bayesian MCMC 方法具有求解资料同化问题的能力：能准确地估计出初始状态和模式误差，同时具有较好的抗噪声性能。

## 基金项目

国家自然科学基金资助项目(41475094; 41105063; 41375105)。

## 参考文献

[1] 黄思训, 伍荣生. 大气科学中的数学物理问题[M]. 北京: 气象出版社, 2001.

[2] 邹晓蕾. 资料同化理论与应用(上册) [M]. 北京: 气象出版社, 2009.

[3] 曹小群, 黄思训, 杜华栋. 变分同化中水平误差函数的正交小波模拟新方法[J]. 物理学报, 2008, 57(3): 1984-1989.

[4] 张卫民, 曹小群, 宋君强. 以全球谱模式为约束的四维变分资料同化系统 YH4DVAR 的设计和实现[J]. 物理学报, 2012, 61(24): 249202-249213.

[5] 韩月琪, 张耀存, 王云峰. 一种新的顺序数据同化方法[J]. 中国科学 E 辑: 技术科学, 2009, 39(8): 1472-1482.

[6] 曹小群, 宋君强, 张卫民. 一种基于复数域微分的资料同化新方法[J]. 物理学报, 2013, 62(17): 170504-170509.

[7] Evensen, G. (1994) Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics. *Journal of Geophysical Research*, **99**, 10143-10162. https://doi.org/10.1029/94JC00572

[8] Houtekamer, P.L. and Mitchell, H.L. (1998) Data Assimilation Using an Ensemble Kalman Filter Technique. *Monthly Weather Review*, **126**, 796-811. https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2

[9] Houtekamer, P.L., Mitchell, H.L., Pellerin, G., *et al.* (2005) Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Monthly Weather Review*, **133**, 604-620. https://doi.org/10.1175/MWR-2864.1

[10] Rabier, F. (2005) Overview of Global Data Assimilation Developments in Numerical Weather-Prediction Centres. *Quarterly Journal of the Royal Meteorological Society*, **131**, 3215-3233. https://doi.org/10.1256/qj.05.129

[11] Rabier, F., Jarvinen, H., Klinker, E., *et al.* (2000) The ECMWF Operational Implementation of Four Dimensional Variational Assimilation. Part I: Experimental Results with Simplified Physics. *Quarterly Journal of the Royal Meteorological Societys*, **126**, 1143-1170. https://doi.org/10.1002/qj.49712656415

[12] Courtier, P., Thepaut, J.N. and Hollingsworth, A. (1994) A Strategy for Operational Implementation of 4D-Var, Using an Incremental Approach. *Quarterly Journal of the Royal Meteorological Society*, **120**, 1367-1387. https://doi.org/10.1002/qj.49712051912

[13] Giering, R. and Kaminski, T. (1998) Recipes for Adjoint Code Construction. *ACM Transactions on Mathematical Software*, **24**, 437-474. https://doi.org/10.1145/293686.293695

[14] 程强, 曹建文, 王斌, 等. 伴随模式生成器[J]. 中国科学 F 辑: 信息科学, 2009, 39(5): 545-558.

[15] 刘永柱, 张林, 金之雁. GRAPES 全球切线性和伴随模式的调优[J]. 应用气象学报, 2017, 28(1): 62-71.

[16] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

[17] Tierney, L. (1994) Markov-Chains for Exploring Posterior Distributions. *Annals of Statistics*, **22**, 1701-1762. https://doi.org/10.1214/aos/1176325750

**知网检索的两种方式：**

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2376-4260，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ams@hanspub.org