文章编号:1674-2974(2019)06-0120-13 DOI:10.16339/j.cnki.hdxbzkb.2019.06.017

地下水污染监测井优化设计及污染源识别

张双圣[1,3],刘汉湖[1],强静[2†],刘喜坤[3],朱雪强[1]
(1.中国矿业大学 环境与测绘学院,江苏 徐州 221116;2.中国矿业大学 数学学院,江苏 徐州 221116;
3.徐州市城区水资源管理处,江苏 徐州 221018)

摘 要:在地下水污染源识别过程中,针对监测井监测值信息量不充分或监测值与模型参数关联性较弱的问题,提出一种基于贝叶斯公式与信息熵的监测井优化方法.构建二维地下水溶质运移模型,并运用 GMS 软件进行数值求解.为减少监测井优化设计及污染源识别过程中反复调用数值模型的计算负荷,采用克里金法建立数值模型的替代模型.以信息熵作为优化指标,筛选出不同监测类型的最优监测方案,并以监测成本和反演精度为参考因素,选定相应监测方案, 最后运用差分进化自适应 Metropolis 算法进行污染源识别.算例研究表明:7 口监测井的克里金替代模型的决定系数均大于 0.98, 可较好地替代原数值模型.基于监测成本最小的方案 1(3 号单井),其信息熵为 12.772;兼顾监测成本和反演精度的方案 2(井(2,3)组合),其信息熵为 9.723;基于反演精度较高的方案 3(3 井(2,3,5)组合),其信息熵为 9.377.方案 1 到方案 3 参数后验分布范围及标准差均逐渐减小,验证了信息熵是参数后验分布不确定性的有效量度.

关键词:监测井优化;污染源识别;贝叶斯方法;信息熵;最优拉丁超立方抽样;差分进化自适应 Metropolis 算法;克里金
中图分类号:X523 **文献标志码**:A

Optimization Design of Groundwater Pollution Monitoring Wells and Identification of Contamination Source

ZHANG Shuangsheng[1,3],LIU Hanhu1,QIANG Jing[2†],LIU Xikun[3],ZHU Xueqiang[1]
(1. School of Environment Science and Spatial Informatics,China University of Mining and Technology,Xuzhou 221116,China;
2. School of Mathematics,China University of Mining and Technology,Xuzhou 221116,China;
3. Xuzhou City Water Resource Administrative Office,Xuzhou 221018,China)

Abstract: In the process of identifying groundwater pollution sources,a monitoring well optimization method based on Bayesian formula and information entropy is proposed for the problem that the monitoring value of monitoring wells is insufficient or the correlation between monitoring values and model parameters is weak. The two-dimensional groundwater contaminant transport model was numerically solved by GMS software. To reduce the computational load of the numerical model repeatedly in the optimization design of the monitoring wells and the identification

* **收稿日期**:2018-11-19
基金项目: 国家水体污染控制与治理科技重大专项基金资助项目 (2015ZX07406005), Major Science and Technology Program for Water Pollution Control and Treatment(2015ZX07406005)
作者简介:张双圣(1983—),男,山东昌邑人,中国矿业大学博士研究生
† 通讯联系人,E-mail:57591340@qq.com

process of the pollution source, the Kriging method was used to establish the surrogate model of the numerical model. As an optimization index, the optimal monitoring schemes of different monitoring types were selected, and the monitoring cost and inversion accuracy were taken as reference factors for the corresponding monitoring schemes. Then, the differential evolution adaptive Metropolis algorithm was used to identify the pollution source. The case study results show that: The determination coefficient of the Kriging surrogate models of the 7 monitoring wells was greater than 0.98, which indicated that the Kriging surrogate models can well replace the original numerical model. The scheme 1(single well No. 3) based on the lowest monitoring cost has an information entropy of 12.772; The scheme 2 (the combination of well No.2 and No.3) taking the monitoring cost and inversion accuracy into account has an information entropy of 9.723; The scheme 3(the combination of well No.2, 3 and 5) with higher inversion precision has an information entropy of 9.377. Both the posterior distribution ranges and the standard deviation of model parameters from scheme 1 to scheme 3 were gradually reduced, which verifies that the information entropy is an effective measure of the uncertainty of the posterior distribution of the parameters.

Key words: monitoring well optimization; contamination source identification; Bayesian approach; information entropy; optimal Latin hypercube sampling; differential evolution adaptive Metropolis algorithm; Kriging

地下水污染具有隐蔽性、发现滞后性及修复难度大、费用高的特点,能否准确地得到污染源的相关信息,对于地下水污染的治理具有重要的现实意义. 地下水污染源识别反问题是指通过建立地下水溶质运移模型,利用监测井处的污染物浓度监测值反求污染源位置、污染源释放强度及释放时间等信息. 由于建设监测井成本高昂, 且存在监测数据包含的信息量不充分或者监测值和未知参数的关联性较弱的问题,需要对现有监测方案进行优化设计. 通常以信噪比(Signal-noise Ratio,SNR)[1],基于贝叶斯公式的相对熵[2-3]作为监测井方案信息量的量度指标. 信噪比仅考虑监测误差对监测数据的干扰影响,相对熵未考虑参数先验分布对后验分布的影响. 为此引入信息熵概念[4],信息熵是信息不确定性的度量,不确定性越大,信息熵越大.

地下水污染源识别的求解方法主要包括贝叶斯统计方法[5-6]、地质统计学方法[7],微分进化算法[8]、遗传算法[9]和模拟退火算法[10]等. 其中,贝叶斯统计方法应用较为广泛,在运用该方法对模型参数进行反演识别时,经常需要求解参数的后验估计值或者后验分布,在参数维数不是特别高时可以采用数值积分或者正态近似的方法求解[11]. 但是,随着参数维数的增加, 数值积分算法的计算量将呈指数增长,求解过程复杂而且难度较大,往往需要借助独立抽样的蒙特卡罗方法 (Monte Carlo 方法, 简称 MC 方法)[12]进行近似求解,其中马尔科夫链蒙特卡罗方法(Markov chain Monte Carlo 方法,简称 MCMC 方法)[13-18]作为一种经典抽样方法得到广泛应用. 近些年,比较流行 MCMC 算法主要包括经典 Metropolis 算法[13-14]、延迟拒绝算法 (delayed rejection ,DR)[15-16], 自适应 Metropolis 算法(AM)[17],延迟拒绝自适应 Metropolis 算法(DRAM)[18]等. 但是这些算法均是单链 MCMC 算法,容易出现反演结果不收敛,或者局部最优的问题.多链 MCMC 算法适用于参数维度高,有多个局部最优值点,搜索量大的参数空间,能够更好地解决 Markov chain 局部收敛的问题 [19]. 常用多链 MCMC算法有 DE-MC 算法 (Differential Evolution Markov Chain)[20],DREAM 算法(Differential Evolution Adaptive Metropolis)[21] 等. DREAM 算法是 DE-MC 算法的改进版本, 相比于 DE-MC 算法,DREAM 算法采用自适应随机子空间采样技术及能够自适应调整的交叉概率,并且运用 IQR 统计方法去除无用链,这几个方面提高了 DREAM 算法的搜索效率和解的精度[21].

此外在监测井优化设计及污染源识别过程中需要多次调用地下水溶质运移数值模型,计算代价非常高,而替代模型的应用能够有效地减少计算量. 常用替代模型方法有多项式回归法[22],径向基函数法[23-24],人工神经网络法[25],Kriging 法[26-28]等,其中克里金法(Kriging)作为多项式回归分析的一种改进方法,包含多项式和随机过程两部分,同时具有局部和全局的统计特性, 是一种监督式学习算法. 而且

在 MATLAB 软件中可用专门的 DACE 工具箱[29]建立Kriging 替代模型,方便实用,因此得到广泛应用.

综合上述研究进展,本文建立二维地下水溶质运移数值模型,在确定初始监测时刻,监测间隔时间及监测次数条件下利用最优拉丁超立方抽样方法及 Kriging 法建立数值模型的替代模型,以信息熵为评价指标分别计算不同监测类型下的最优监测井方案,并筛选出兼顾监测成本与监测精度的监测方案, 然后以筛选出的监测方案运用 DREAM 算法进行污染源识别,为地下水污染源识别及监测方案优化研究提供借鉴.

1 研究方法

1.1 贝叶斯公式

贝叶斯公式如下:

$$p(\alpha|d)=\frac{p(d|\alpha)p(\alpha)}{p(d)}\propto p(d|\alpha)p(\alpha) \tag{1}$$

式中:α 为模型的未知参数;d 为实测数据;$p(\alpha|d)$为参数的后验概率密度函数;$p(\alpha)$为参数的先验概率密度函数;$p(d|\alpha)$为条件概率密度函数;$p(d)=\int p(d|\alpha)p(\alpha)\mathrm{d}\alpha$ 为归一化的积分常数,同时也称其为监测数据 d 出现的概率.

假设模型中未知参数共有 m 个,则 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_m)$. 水文地质模型中的参数都分布在一个特定的范围内, 可以认为每个参数都服从均匀分布,且 $\alpha_1,\alpha_2,\cdots,\alpha_m$ 相互独立. 模型参数 α_i 的先验概率密度函数可定义为:

$$p(\alpha_i)=\begin{cases}\dfrac{2}{B_i-A_i}, & \alpha_i\in[A_i,B_i]\\ 0, & \text{其他}\end{cases} \tag{2}$$

则总的先验分布 $p(\alpha)$可表示为:

$$p(\alpha)=\prod_{i=1}^{m}p(\alpha_i) \tag{3}$$

假设模型中监测值记为 $d=(d_1,d_2,\cdots,d_n)$. d 表示监测数据,$F(\alpha)$表示参数 α 模型的计算值,则 $\varepsilon=d-F(\alpha)$为整体误差. 假设整体误差 $\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)$均值为 0(即 $E(\varepsilon)=0$),服从 n 维正态分布,且 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立,则整体误差 $\varepsilon=d-F(\alpha)$条件概率密度函数 $p(d|\alpha)$可表示如下:

$$p(d|\alpha)=\frac{1}{(2\pi)^{n/2}|\boldsymbol{C}(\varepsilon)|^{1/2}}\times \exp\left\{-\frac{1}{2}(d-F(\alpha))^{\mathrm{T}}\boldsymbol{C}(\varepsilon)^{-1}(d-F(\alpha))\right\} \tag{4}$$

式中:协方差矩阵 $\boldsymbol{C}(\varepsilon)=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix}$;$\boldsymbol{C}(\varepsilon)$表示矩阵 $\boldsymbol{C}(\varepsilon)$的行列式;$\boldsymbol{C}(\varepsilon)^{-1}$ 表示矩阵 $\boldsymbol{C}(\varepsilon)$的逆矩阵;$\sigma_i>0(i=1,2,\cdots,n)$.

联合式(1)(2)(3)(4),可得 α 的后验概率密度函数 $p(\alpha|d)$为:

$$p(\alpha|d)=\frac{\prod_{i=1}^{m}p(\alpha_i)}{p(d)(2\pi)^{n/2}|\boldsymbol{C}(\varepsilon)|^{1/2}}\times \exp\left\{-\frac{1}{2}(d-F(\alpha))^{\mathrm{T}}\boldsymbol{C}(\varepsilon)^{-1}(d-F(\alpha))\right\}= \lambda\exp\left\{-\frac{1}{2}(d-F(\alpha))^{\mathrm{T}}\boldsymbol{C}(\varepsilon)^{-1}(d-F(\alpha))\right\} \tag{5}$$

式中:$\lambda=\dfrac{\prod_{i=1}^{m}p(\alpha_i)}{p(d)(2\pi)^{n/2}|\boldsymbol{C}(\varepsilon)|^{1/2}}$是一固定值,与参数 α 的选取无关.

在实测数据 d 固定的条件下,式(5)是关于参数 α 的函数. 通过积分求解参数 α 的后验分布很难得出显式表达式,本研究采用 MCMC (Markov Chain Monte Carlo)算法[21]对式(5)的后验分布进行求解.

1.2 基于贝叶斯公式与信息熵的监测井优化设计

监测方案 Monitoring Proposal (MP)的优化设计主要包括对监测井的数量、位置以及监测频率的优化.假设初始监测时间为 t_1(固定值),由监测方案 MP 所得监测值仍记为 d,此时贝叶斯公式可以改写成:

$$p(\alpha|d,\mathrm{MP})=\frac{p(\alpha|\mathrm{MP})p(d|\alpha,\mathrm{MP})}{\int p(\alpha|\mathrm{MP})p(d|\alpha,\mathrm{MP})\mathrm{d}\alpha} \tag{6}$$

由于参数的先验分布 $p(\alpha|\mathrm{MP})$表示对未知参数的初步认识,不受监测设计方案 MP 的影响,即 $p(\alpha|\mathrm{MP})=p(\alpha)$,式(6)变为:

$$p(\alpha|d,\mathrm{MP})=\frac{p(\alpha)p(d|\alpha,\mathrm{MP})}{\int p(\alpha)p(d|\alpha,\mathrm{MP})\mathrm{d}\alpha} \tag{7}$$

归一化常数 $\int p(\alpha)p(d|\alpha,\mathrm{MP})\mathrm{d}\alpha$ 表示由监测设计方案 MP 所得监测数据 d 出现的概率, 可简记为 $p(d|\mathrm{MP})$,即

$$p(d|\mathrm{MP})=\int p(\alpha)p(d|\alpha,\mathrm{MP})\mathrm{d}\alpha \tag{8}$$

设一维连续型随机变量 Θ 的概率密度函数为 $f(\theta)$,Θ 在区间$[a,b]$上信息熵[4]定义如下:

$$H(\boldsymbol{\Theta}) = -\int_a^b f(\theta)\ln f(\theta)\mathrm{d}\theta \tag{9}$$

运用污染物浓度监测值 d 反演未知参数 α,α 后验概率密度函数 $p(\alpha|d,\mathrm{MP})$,α 的后验分布的信息熵类似定义如下:

$$H(\mathrm{MP},d) = -\int p(\alpha|d,\mathrm{MP})\ln p(\alpha|d,\mathrm{MP})\mathrm{d}\alpha \tag{10}$$

式(10)左端项含有监测值 d,但在进行监测井方案设计时,并未真正得到 d,可认为监测值 d 是随机变量,其概率密度函数为 $p(d|\mathrm{MP})$. 为了得到一个仅包含监测方案 MP 的函数,对式(10)两侧乘以 $p(d|\mathrm{MP})$,再对 d 积分,求出信息熵 $H(\mathrm{MP},d)$ 的期望为:

$$E(H(\mathrm{MP},d)) = -\int\left[\int p(\alpha|d,\mathrm{MP})\ln p(\alpha|d,\mathrm{MP})\mathrm{d}\alpha\right]p(d|\mathrm{MP})\mathrm{d}d = -\iint p(\alpha|d,\mathrm{MP})p(d|\mathrm{MP})\ln p(\alpha|d,\mathrm{MP})\mathrm{d}\alpha\mathrm{d}d \tag{11}$$

式(11)中 $E(H(\mathrm{MP},d)$只受到监测方案 MP 的影响,可简记为 $E(\mathrm{MP})$. 通过求解 $E(\mathrm{MP})$的最小值,就可以获得最优监测井设计方案 MP^*. 根据信息熵概念,利用由 MP^* 获得的污染物浓度监测值 d^* 反演未知参数 α,此时 α 后验分布的信息熵最小,表示 α 的不确定性也最小,反演效果最优.

式(11)的求解算法较为复杂,难以得出显示表达式,本文运用文献[2]中蒙特卡罗方法近似求解.

首先利用式(7)对式(11)改写如下:

$$E(\mathrm{MP}) = -\iint p(d|\alpha,\mathrm{MP})p(\alpha)\ln p(\alpha)\mathrm{d}\alpha\mathrm{d}d - \iint p(d|\alpha,\mathrm{MP})p(\alpha)[\ln p(d|\alpha,\mathrm{MP}) - \ln p(d|\mathrm{MP})]\mathrm{d}\alpha\mathrm{d}d \tag{12}$$

先验分布 $p(\alpha)$由式(2)(3)给定,故式(12)中 α 的先验分布信息熵

$$-\iint p(d|\alpha,\mathrm{MP})p(\alpha)\ln p(\alpha)\mathrm{d}\alpha\mathrm{d}d = -\ln p(\alpha)$$

可知 α 先验范围越大,α 的先验分布信息熵越大,α 的不确定性越大. 在先验概率分布 $p(\alpha)$不变时,$-\ln p(\alpha)$保持不变,因此,要得到 $E(\mathrm{MP})$的最小值,只需计算$-\iint p(d|\alpha,\mathrm{MP})p(\alpha)[\ln p(d|\alpha,\mathrm{MP}) - \ln p(d|\mathrm{MP})]\mathrm{d}\alpha\mathrm{d}d$ 的最小值. 为叙述方便,记

$$U(\mathrm{MP}) = -\iint p(d|\alpha,\mathrm{MP})p(\alpha)[\ln p(d|\alpha,\mathrm{MP}) - \ln p(d|\mathrm{MP})]\mathrm{d}\alpha\mathrm{d}d \tag{13}$$

据文献[2],$U(\mathrm{MP})\leqslant 0$,因此由信息熵概念和式(12)可知在贝叶斯公式中监测值 d 的使用降低了参数 α 的不确定性.

运用蒙特卡罗方法求解式(13):

$$U(\mathrm{MP}) \approx -\frac{1}{\Pi}\sum_{i=1}^{\Pi}[\ln p(d^i|\alpha^i,\mathrm{MP}) - \ln p(d^i|\mathrm{MP})] \tag{14}$$

其中,首先从未知参数 α 先验分布 $p(\alpha)$中随机采集 Π 个样本,记为 $\alpha^i(i=1,2,\cdots,\Pi)$;对于每一个 $i\in\Pi$,根据式(4)从条件概率密度函数 $p(d|\alpha^i,\mathrm{MP})$中随机采集 1 个样本 d^i,共得到 Π 个;将每组 α^i 和 d^i 代入式(4)得到式(14)中 $p(d^i|\alpha^i,\mathrm{MP})$. 由式(8)可知式(14)中 $p(d^i|\mathrm{MP}) = \int p(\alpha)p(d^i|\alpha,\mathrm{MP})\mathrm{d}\alpha$,仍采用蒙特卡罗方法求此积分,故

$$p(d^i|\mathrm{MP}) \approx -\frac{1}{\Pi}\sum_{j=1}^{\Pi}p(d^i|\alpha^i,\mathrm{MP}) \tag{15}$$

因此,只要监测设计方案 MP 固定,通过式(12)(14)(15)就可以得到此种监测方案的信息熵近似值.

1.3 DREAM 算法

1.3.1 DREAM 算法具体步骤

DREAM 算法[21]是在 DE-MC 算法[20]的基础上提出的. DREAM 算法步骤如下:

1)在模型未知参数 α 先验范围内随机产生 N_P 个初始样本 $X_i(t)=(X_{i,1}(t),X_{i,2}(t),\cdots,X_{i,m}(t))^{\mathrm{T}}(i=1,2,\cdots,N_P,t=0)$,作为 N_P 条马尔科夫链的起始点.

2)针对第 $i(i=1,2,\cdots,N_P)$条马尔科夫链,运用 DE 方法产生参数的变异样本 $Z_j(t)$.

$$Z_i(t) = X_i(t) + (\boldsymbol{I}+\boldsymbol{e})\gamma(\delta,d')\times\left(\sum_{j=1}^{\delta}X_{r_1(k)}(t) - \sum_{k=1}^{\delta}X_{r_2(n)}(t)\right) + \boldsymbol{\varepsilon} \tag{16}$$

式中:$\boldsymbol{I}$ 表示 m 阶单位矩阵;$\boldsymbol{e}$ 表示 m 阶方阵, 其对角线上元素服从均匀分布 $U(-b,b)$,b 为自定义的极小值;δ 表示用于产生候选样本的平行链数的二分之一,$r_1(j),r_2(k)\in\{1,2,\cdots,N_P\}$为随机选取的平行链编号,且当 $j,k=1,2,\cdots,\delta$ 时,$r_1(j)\neq r_2(k)\neq i$;$\gamma(\delta,d')$为比例因子;$\boldsymbol{\varepsilon}$ 表示 m 维向量,其 m 个元素均服从正态分布 $N(0,b^*)$,b^* 为自定义的极小值.

3)引入交叉概率 Cr,交叉混合 $X_i(t)$和 $Z_i(t)$得

到候选样本 V_i 如下:

对 $\forall i \in \{1,2,\cdots,N_P\}$,当 $j=1$ 时,令 $d'=d$;对 $j=1,2,\cdots,m$,有

$$X_{i,j}(t)=\begin{cases}X_{i,j}(t) & \text{if } U\leqslant 1-Cr, d'=d'-1\\ Z_{i,j}(t) & \text{其他}\end{cases} \tag{17}$$

其中 $0\leqslant Cr\leqslant 1$,U 是区间[0,1]上的随机数.

4)产生区间[0,1]上随机数 u;计算接受概率 $\min\left\{1,\frac{p(V_i(t)|d}{p(X_i(t)|d}\right\}$;其中 $p(V_i(t)|d)$与 $p(X_i(t)|d)$由式(5)计算得到. 如果 $u<\min\left\{1,\frac{p(V_i(t)|d}{p(X_i(t)|d}\right\}$,接受 $V_i(t)$,即 $X_i(t+1)=V_i(t)$;否则,拒绝 $V_i(t)$,即 $X_i(t+1)=X_i(t)$.

5)根据 IQR 统计方法[21]去除无用链.

6)判断收敛性.如果马尔科夫链满足 Gelman-Rubin 收敛准则[30],若满足条件则计算终止,否则继续进化平行序列.

根据文献[21],为了确保平行序列样本的多样性和提高收敛速度,建议 δ 取为集合$\{1,2,\cdots,n_\delta\}$中的随机数,其中 $n_\delta\leqslant(N_P-1)/2$;建议 $\gamma(\delta,d')=2.38/\sqrt{2\delta d'}$,且当进化代数 $t/5$ 为整数时,令 $\gamma(\delta,d')=1.0$;建议 $0\leqslant Cr\leqslant 1$,集合$\{1,2,3\}$中的随机数记为 n_{Cr},取 $Cr=1/n_{Cr}$.

1.3.2 DREAM 算法的收敛性判断

本文采用 Gelman-Rubin 收敛诊断方法[30]对 DREAM 算法后 50%采样过程的收敛性进行判断,判断指标为:

$$\hat{R}_j=\sqrt{\frac{g-1}{g}+\frac{q+1}{q}\cdot\frac{B_j}{W_j}} \tag{18}$$

式中:$\hat{R}_j(j=1,2,\cdots,m)$表示第 j 个参数判断指标;g 表示多链 DRAM 算法中马尔科夫链长度的 1/2;q 表示用于评价的马尔科夫链条数;B_j 表示第 j 个参数的 q 条马尔科夫链后 50%样本的平均值的方差;W_j 表示第 j 个参数的 q 条马尔科夫链后 50%样本的方差的平均值.

当 $\hat{R}_j<1.2$ 时,表示马尔科夫链已经收敛;当 $\hat{R}_j\geqslant 1.2$ 时,表示马尔科夫链未收敛.

1.4 最优拉丁超立方抽样方法和 Kriging 替代模型

1.4.1 最优拉丁超立方抽样方法

拉丁超立方抽样是一种多维分层随机抽样方法,但抽样结果有很大地随机性,每一次抽样得到的结果差别很大. 假设要在 m 维参数 α 取值范围 $[0,1]^m$ 内运用拉丁超立方抽样方法抽取 q 组样本,q 组样本数据用矩阵 $\boldsymbol{\Phi}_{qm}$ 表示; 符合要求的矩阵 $\boldsymbol{\Phi}_{qm}$ 共有$(q!)^m$种,总共可以得到$(q!)^m$种拉丁超立方抽样方案,$(q!)^m$种抽样方案 $\boldsymbol{\Phi}_{qm}$ 组成的集合记为 Λ. 这些抽样方案对整个抽样空间的覆盖填充程度有很大地差异[28]. 本文运用中心化 L_2 偏差[31]作为优化指标寻求覆盖填充程度最优的抽样方案,中心化 L_2 偏差表达式如下:

$$CL_2(\boldsymbol{\Phi}_{qm})=\left(\frac{13}{12}\right)^m-\frac{2}{q}\sum_{k=1}^{q}\prod_{i=1}^{m}\left(1+\frac{1}{2}\left|\phi_{ki}-\frac{1}{2}\right|-\frac{1}{2}\left|\phi_{ki}-\frac{1}{2}\right|^2\right)+$$
$$\frac{1}{q^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\prod_{i=1}^{m}\left(1+\frac{1}{2}\left|\phi_{ki}-\frac{1}{2}\right|+\frac{1}{2}\left|\phi_{li}-\frac{1}{2}\right|-\frac{1}{2}|\phi_{ki}-\phi_{li}|\right)$$

式中:m 为参数 α 的维数;q 为拉丁超立方抽样方案抽取样本数;ϕ_{ki} 为矩阵 $\boldsymbol{\Phi}_{qm}$ 中第 k 行 i 列元素;ϕ_{li} 表示矩阵 $\boldsymbol{\Phi}_{qm}$ 中第 l 行 i 列元素.

求解 $\min\limits_{\boldsymbol{\Phi}_{qm}\in\Lambda} CL_2(\boldsymbol{\Phi}_{qm})$即可得到最优拉丁超立方抽样.

$$\boldsymbol{\Phi}_{qm}^{*}=\begin{bmatrix}\phi_{11}^{*} & \phi_{12}^{*} & \cdots & \phi_{1m}^{*}\\ \phi_{21}^{*} & \phi_{22}^{*} & \cdots & \phi_{2m}^{*}\\ \vdots & \vdots & & \vdots\\ \phi_{q1}^{*} & \phi_{q2}^{*} & \cdots & \phi_{qm}^{*}\end{bmatrix}$$

假设要在 m 维模型参数 α 先验范围 $[A_i,B_i](i=1,2,\cdots,m)$内抽取 q 组最优拉丁超立方样本,样本矩阵形式如下:

$$\boldsymbol{\Theta}=\begin{bmatrix}A_1+(B_1-A_1)\phi_{11}^{*} & A_2+(B_2-A_2)\phi_{12}^{*} & \cdots & A_m+(B_m-A_m)\phi_{1m}^{*}\\ A_1+(B_1-A_1)\phi_{21}^{*} & A_2+(B_2-A_2)\phi_{22}^{*} & \cdots & A_m+(B_m-A_m)\phi_{2m}^{*}\\ \vdots & \vdots & & \vdots\\ A_1+(B_1-A_1)\phi_{q1}^{*} & A_2+(B_2-A_2)\phi_{q2}^{*} & \cdots & A_m+(B_m-A_m)\phi_{qm}^{*}\end{bmatrix}$$

1.4.2 Kriging 替代模型

Kriging 法[26]是由 Matheron 等人发明的一种优化插值方法,其具体工作原理可查阅文献[27],基本步骤如下:

Kriging 替代模型响应值 y 与自变量 $\boldsymbol{k}$ 之间的关系可用下式表示:$y(\boldsymbol{k})=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{B}+Z$,式中 $\boldsymbol{k}\in \mathrm{IR}^{m_k}$($m_k$ 表示自变量 k 维数),$\boldsymbol{f}(\boldsymbol{k})=(f_1(\boldsymbol{k}),f_2(\boldsymbol{k}),\cdots,f_p(\boldsymbol{k}))^{\mathrm{T}}$,其中 $f_i(\cdot)(i=1,2,\cdots,p)$表示事先确定的多项式函数, 常常是 0 阶,1 阶或者 2 阶多项式;$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 为待定参数;Z 为误差随机变量,$E(Z)=0$,$\mathrm{Var}(E)=\sigma^2$.

给定训练样本集 $\boldsymbol{k}=(\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_{n_k})^{\mathrm{T}}$ 及响应集 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_{n_k})^{\mathrm{T}}$,其中 $\boldsymbol{k}_i\in\mathrm{IR}^{m_k}$,$y_i\in\mathrm{IR}$. 令 $\boldsymbol{F}=[F_{ij}]_{n_k\times p}=[\boldsymbol{f}(k_1)\ \boldsymbol{f}(k_2)\cdots\boldsymbol{f}(k_{n_k})]^{\mathrm{T}}$,$\boldsymbol{R}=[R(\boldsymbol{k}_i,\boldsymbol{k}_j)]_{n_k\times n_k}$,

$R(\boldsymbol{k}_i,\boldsymbol{k}_j)$为任意两个训练样本点 $\boldsymbol{k}_i,\boldsymbol{k}_j$ 之间的空间相关函数，本文采用 Gauss 型相关函数：$R(\boldsymbol{k}_i,\boldsymbol{k}_j)=\exp\left[-\sum_{k=1}^{m_k}\theta_k(\boldsymbol{k}_i^k-\boldsymbol{k}_j^k)^2\right](i,j=1,2,\cdots,n_k)$，其中 $\boldsymbol{k}_i^k$、$\boldsymbol{k}_j^k$ 分别表示训练样本点 $\boldsymbol{k}_i$、$\boldsymbol{k}_j$ 的第 k 个分量，θ_k（$k=1,2,\cdots,m_k$）为待定参数.

任意一个预测点 $\boldsymbol{k}_{new}$ 的估计响应值为 $\hat{y}(\boldsymbol{k}_{new})=\boldsymbol{f}(\boldsymbol{k}_{new})^{\mathrm{T}}\beta^*+\boldsymbol{r}(\boldsymbol{k}_{new})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{F}\beta^*)$，式中运用最优线性无偏估计可以得到待定参数 $\boldsymbol{\beta}^*=(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{Y}$；令 $\boldsymbol{r}(\boldsymbol{k}_{new})=[R(\boldsymbol{k}_{new},\boldsymbol{k}_1),R(\boldsymbol{k}_{new},\boldsymbol{k}_2),\cdots,R(\boldsymbol{k}_{new},\boldsymbol{k}_{n_k})]^{\mathrm{T}}$，$\boldsymbol{r}(\boldsymbol{k}_{new})$表示预测点 $\boldsymbol{k}_{new}$ 与训练样本点之间的相关向量；待定参数 θ_k 根据极大似然估计给出，即选取满足 $\max_{\theta_k>0}\left[-\frac{n_k}{2}\ln(\sigma^2)-\frac{1}{2}\ln|\boldsymbol{R}|\right]$ 的，其中 $\sigma^2=\frac{(\boldsymbol{Y}-\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\beta}^*)}{n_k}$.

2002 年，Lophaven S N 等人[29]基于 Matlab 平台创建了 DACE 工具箱，用于生成 Kriging 替代模型，本文中 Kriging 替代模型的建立就是基于此工具箱.

2 算例应用

2.1 模型建立及问题概述

2.1.1 模型建立

假定研究区域为矩形区域，长 1 000 m，宽 600 m，含水层为厚度 35 m 的砂质潜水含水层（水文地质参数见表 1），西部边界 Γ_1 与东部边界 Γ_3 为给定水头边界，其中东部水头为 25 m，西部水头为 30 m，北部边界 Γ_2 与南部边界 Γ_4 为隔水边界. 天然状态下地下水流为自西向东流动的二维均质各向同性的非承压稳定流.

表 1 研究区域已知水文地质参数

Tab.1 Known hydrological parameters in the studied area

纵向弥散度 $D_{\alpha x}$/m	横向弥散度 $D_{\alpha y}$/m	渗透系数 K/(m·d^{-1})	有效孔隙度 n	给水度 μ	含水层水位 H/m	含水层底板高程 B/m
20	5	20	0.22	0.20	35	0

研究区内共有 7 眼监测井，地下水水质背景值较好，含水层污染物的初始浓度为零，某日研究区下游发现污染物 X，初步断定污染源 S 在上游的某一区域范围（先验范围）内，且在某时间段内以注水井的形式（200 m^3/d）持续恒定地向含水层中排放污染物. 研究区平面示意图如图 1 所示.

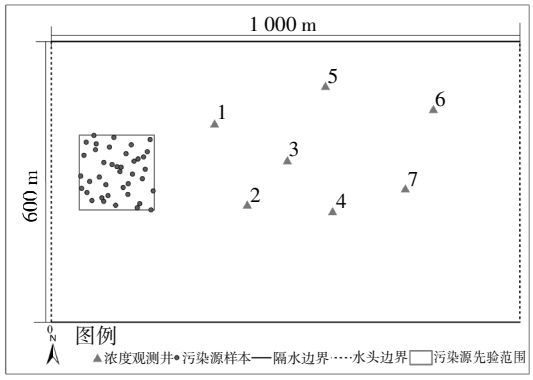


图 1 算例模型示意图

Fig.1 Sketch of example model

以西南角为坐标原点建立坐标系，根据研究区水文地质条件，建立地下水水流数值模型：

$$\begin{cases}\dfrac{\partial}{\partial x}[K(H-B)\dfrac{\partial H}{\partial x}]+\dfrac{\partial}{\partial y}[K(H-B)\dfrac{\partial H}{\partial y}]+w=\mu\dfrac{\partial H}{\partial t} & (x,y)\in S,t\geqslant 0\\ H(x,y,t)|_{t=0}=H_0(x,y) & (x,y)\in S,t=0\\ H(x,y,t)|_{\Gamma_1,\Gamma_3}=H_1(x,y,t) & (x,y)\in\Gamma_1,\Gamma_3,t\geqslant 0\\ K_n(H-B)\dfrac{\partial H}{\partial\vec{n}}\Big|_{\Gamma_2,\Gamma_4}=0 & (x,y)\in\Gamma_2,\Gamma_4,t\geqslant 0\end{cases}$$

式中：K 为含水层渗透系数，m/d；H 为潜水含水层水位，m；B 为潜水含水层底板高程，m；w 为源汇项；μ 为含水层给水度，无量纲；S 为模拟区范围；$H_0(x,y)$ 为初始水位，m；H_1 为一类边界上的已知水头，m；Γ_1，Γ_3 为给定水头边界；Γ_2，Γ_4 为零流量边界；K_n 为边界法向量上的渗透系数，m/d；$\vec{n}$ 为二类边界的外法线方向.

在地下水流数值模型基础上，可建立地下水溶质运移数值模型，模拟区各边界可概化为：Γ_1 为零浓度边界，Γ_3 为对流弥散通量边界，Γ_2、Γ_4 视为零弥散通量边界. 则研究区地下水的溶质运移模型为：

$$\begin{cases}n\dfrac{\partial c}{\partial t}=\dfrac{\partial}{\partial x}(nD_x\dfrac{\partial c}{\partial x})+\dfrac{\partial}{\partial y}(nD_y\dfrac{\partial c}{\partial y})-\dfrac{\partial}{\partial x}(v_x c)-\dfrac{\partial}{\partial y}(v_y c)+C_{inj}Q_{inj} & (x,y)\in S,t\geqslant 0\\ c(x,y,t)|_{t=0}=0 & (x,y)\in S,t=0\\ c(x,y,t)|_{\Gamma_1}=0 & (x,y)\in\Gamma_1,t>0\\ D\dfrac{\partial c}{\partial\vec{n}}\Big|_{\Gamma_2,\Gamma_4}=0 & (x,y)\in\Gamma_2,\Gamma_4,t>0\\ (-nD_x\dfrac{\partial c}{\partial x}+cv_x|_{\Gamma_2}=f(x,y,t) & (x,y)\in\Gamma_3,t>0\end{cases}$$

式中：D_x、D_y 分别为水动力弥散系数在 x、y 方向上的分量，m^2/d；v_x、v_y 分别为 x、y 方向上地下水的渗流速

度,m/d;n 为含水层介质的孔隙度,无量纲;c 为污染物质量浓度,mg/L;Q_{inj} 为向含水层的注入液体量,m^3/d;C_{inj} 为随液体进入含水层的污染物质量浓度,mg/L;$f(x,y,t)$表示仅在水动力弥散作用下,单位时间通过实际过水断面单位面积的溶质质量.

建立的地下水水流及溶质运移模型运用 GMS 软件中的 MODFLOW 和 MT3DMS 模块进行计算求解. 为保证每个网格中心对应一个潜在污染源（污染源样本）的位置，将研究区域剖分为 150 行 200 列的正方形有限差分网格,基本单元格边长为 4 m.

2.1.2 问题概述

针对潜在的污染源范围,要求在现有 7 眼监测井的基础上,制定优化的监测井方案,以此进行污染源的识别(包括污染源的位置,污染物投放和结束时间,以及污染物的投放质量浓度),即求解污染源未知参数 $\alpha=(X_s,Y_s,T_1,T_2,C_s)$,其中$(X_s,Y_s)$为污染源位置,m;$T_1$、$T_2$ 分别为污染源开始排放和结束排放的时间,d;C_s 是注入污染物的质量浓度,mg/L. 表 2 为 7 眼监测井的坐标位置.

表 2 7 眼监测井的坐标位置

Tab.2 The coordinate positions of 7 monitoring wells

序号	坐标
1	(348.5,424.9)
2	(418.4,251.7)
3	(503.9,346.1)
4	(600.4,237.3)
5	(585.2,505.3)
6	(815.6,455.8)
7	(755.8,286.1)

以未发生污染某确定时刻作为初始时间,此时 t=0. 要求从第 750 d 至第 900 d 内完成污染源识别任务（期间每 10 d 进行一次监测，每眼井共监测 16 次),假设上述 5 个参数的先验分布为均匀分布 $\alpha=(X_s,Y_s,T_1,T_2,C_s)$,5 个待求参数的先验范围分别如下:

60 m≤X_s≤220 m,240 m≤Y_s≤400 m,8 d≤T_1≤12 d,18 d≤T_2≤22 d,2 800 mg/L≤C_s≤3 300 mg/L.

2.2 Kriging 替代模型的建立

1)运用最优拉丁超立方抽样方法从污染源未知参数 α 的先验范围内抽得均匀分布的 100 组参数作为 Kriging 替代模型训练输入值（此时 $CL_2(\boldsymbol{\Phi}_{100,5})$=0.002 462),如表 3 所示.

表 3 从待求参数的先验分布中得到 100 组训练输入数据

Tab.3 100 sets of training input data obtained from the prior distribution

序号	X_s	Y_s	T_1	T_2	C_s
1	106.21	361.1	9.0344	21.461	2 918.8
2	192.44	399.55	11.377	19.595	3 056.8
3	174.86	310.88	10.952	18.533	3 183.7
4	167.33	300.34	8.051 7	18.001	3 152.8
5	209.20	278.73	11.830	18.461	3 072.0
6	65.497	281.98	9.496 7	19.773	3 118.8
7	211.57	287.51	8.697 1	20.216	2 914.2
8	145.15	244.03	9.264 4	18.768	2 816.7
9	63.006	292.82	9.386 0	18.289	3 128.3
10	94.898	372.43	8.261 0	19.945	2 931.1
⋮	⋮	⋮	⋮	⋮	⋮
100	82.607	388.8	8.239	20.773	3 019.9

2)分别对 7 眼监测井建立 Kriging 替代模型. 将表 3 中 100 组参数分别代入 GMS 软件中,得到 7 眼监测井不同监测时间点的污染物浓度值，作为 Kriging 替代模型输出值. 将 100 组输入值与输出值作为训练样本代入 MATLAB 软件,利用 DACE 工具箱对各监测井的 Kriging 替代模型进行训练.

在未知参数的先验范围内运用拉丁超立方抽样方法重新得到 10 组参数作为检验样本的输入值(表 4),再次代入 GMS 软件中,得到 7 眼监测井不

表 4 从待求参数的先验分布中得到 10 组检验输入数据

Tab.4 10 sets of testing input data obtained from the prior distribution

序号	X_s	Y_s	T_1	T_2	C_s
1	130.45	335.27	11.48	21.513	3 200.4
2	71.545	386.01	10.945	20.959	3 109.5
3	191.75	349.01	9.323 4	21.938	2 986.3
4	174.211 4	293.623 3	11.921	18.165	3 095.5
5	151.16	379.6	8.652 8	18.678	3 045.7
6	106.86	276.58	9.068 7	19.006	2 870.5
7	206.3	258.9	9.924 3	20.254	3 278.7
8	85.136	316.25	8.304 4	19.443	3 168.5
9	179.27	361.59	10.255	19.945	2 834.2
10	121.09	254.18	10.608	20.8	2 925.4

同监测时间点的污染物质量浓度值,作为检验样本的输出值，记为 $y_{i,\text{out}}$,$y_{i,\text{out}}=(y_{i,1},y_{i,1},\cdots,y_{i,160})$,$i=1,2,\cdots,7$,表示 i 号监测井. 将 10 组检验样本的输入值代入到替代模型中，得到替代模型输出值,记为$\hat{y}_{i,\text{out}}$,$\hat{y}_{i,\text{out}}=(\hat{y}_{i,1},\hat{y}_{i,2},\cdots,\hat{y}_{i,160})$,$i=1,2,\cdots,7$,表示 i 号监测井.

以 7 号监测井为例,以检验样本的输出值作为横坐标,替代模型的输出值作为纵坐标,绘制替代模型输出值与检验样本输出值的对比图（图 2). 由图 2 可知，数据散点均集中分布于 $y=x$ 直线上,表明替代模型可较好地对数值模型进行替代.

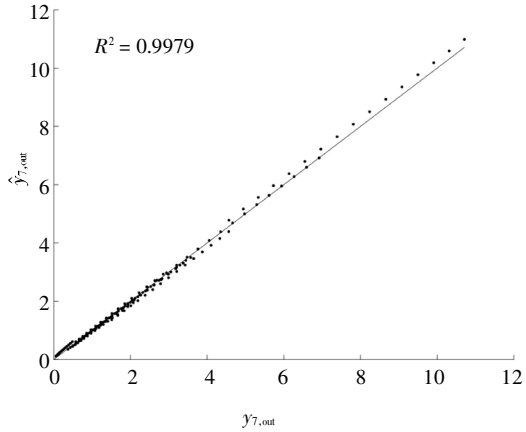


图 2　Kriging 替代模型与数值模型输出结果对比

Fig.2　Output comparision between Kriging surrogate model and numerical model

为进一步检验替代模型的精度,分别运用决定系数,平均绝对误差及均方根误差 3 个指标对 7 眼监测井的替代模型进行检验评价,结果如表 5 所示.

①决定系数:$R_i^2=1-\dfrac{\sum_{j=1}^{160}(y_{ij}-\hat{y}_{ij})^2}{\sum_{j=1}^{160}(y_{ij}-\bar{y}_i)^2}$,$i=1,2,\cdots,7$.

其中$\bar{y}_i=\sum_{j=1}^{160}y_{i,j}/160$ 表示数值模型输出的均值.

②平均绝对误差:

$$\text{MAE}_i=\sum_{j=1}^{210}|y_{i,j}-\hat{y}_{i,j}|/210,\ i=1,2,\cdots,7.$$

③均方根误差

$$\text{RMSE}_i=\sqrt{\sum_{j=1}^{160}(y_{i,j}-\hat{y}_{i,j})^2/(160-1)},\ i=1,2,\cdots,7.$$

表 5　7 眼监测井 Kriging 替代模型的决定系数及平均绝对误差

Tab.5　The determination coefficient and average absolute errors of Kriging surrogate models of the 7 monitoring wells

井号 i	决定系数	平均绝对误差	均方根误差
1	0.995 3	0.19	0.29
2	0.998 1	0.19	0.24
3	0.997 2	0.23	0.33
4	0.997 6	0.22	0.30
5	0.984 8	0.03	0.03
6	0.996 7	0.01	0.02
7	0.997 9	0.08	0.11

由表 5 数据可知,Kriging 替代模型预测精度较高，且 7 眼备选监测井平均绝对误差的平均值为 0.14,平均决定系数 $R^2=0.996\,8$. 在参数反演及监测井优化问题中,由于运用 Kriging 替代模型,整体误差包括替代模型误差及测量误差. 假设第 i（i=1,2,$\cdots$,7）眼备选监测井的 Kriging 替代模型误差 ε_i'满足正态分布 $N(0,(\sigma_i')^2)$,且均值 $E(\varepsilon_i')=0$,均方差 $\sigma_i'=\text{RMSE}_i$；假设测量误差 ε''满足正态分布 $N(0,(\sigma'')^2)$,且均值 $E(\varepsilon'')=0$,均方差 $\sigma''=0.01$. 由于替代模型误差 ε_i'及测量误差 ε''相互独立，故第 i（i=1,2,$\cdots$,7)口监测井整体误差 $\bar{\varepsilon}_i=\varepsilon_i'+\varepsilon''$满足正态分布 $N(0,(\sigma_i')^2+(\sigma'')^2)$,据此可以确定 1.1 节中式(4)和式(5)中协方差矩阵 $\boldsymbol{C}(\varepsilon)$.

2.3　监测方案优化设计

7 眼井可得到 7 组监测数据,任取 i(i=1,2,$\cdots$,7)组监测数据作为参数反演的监测数据 d,共有 C_7^i(i=1,2,$\cdots$,7)种组合形式,每种组合形式代表一种监测方案,从而监测方案共有 C_7^i(i=1,2,$\cdots$,7)种. 由于 i=1,2，$\cdots$,7，按照选取的监测井数量进行划分,监测类型共有 7 类. 根据 1.2 节可知,每类监测方案优化问题其实就是在 C_7^i(i=1,2,$\cdots$,7)种监测方案里面选取信息熵最小的监测方案,信息熵最小的监测方案可视为最优监测方案. 问题可概化如下:

$$\min E(\text{MP})=-\ln p(\alpha)+\min U(\text{MP}) \qquad (19)$$

式中:MP 表示监测方案.

根据式(12)(14)(15)分别求得不同监测方案的信息熵 E(MP).为了验证基于贝叶斯公式与信息熵的监测井优化设计效果,以信息熵 E(MP)与反演结果相对误差均值 MRE(MP)为指标对 C_7^i种监测方案

进行评价. 从参数 α 的先验分布里运用拉丁超立方抽样方法随机并且均匀地得到 20 组参数作为真实参数(表 6),对应于 C_7^i 种监测方案,20 组真实参数通过 Kriging 替代模型产生了 $20\times C_7^i$ 组浓度监测值. 利用产生的监测值,运用 DREAM 算法(初始平行链数为 10)反演参数 α,其中每条马尔科夫链长度为 12 000,在马尔科夫链长度 10 000 时所有参数的收敛性判断指标 $\hat{R}$<1.2.为了保证精度,只将马尔科夫链趋于稳定后的最后 2 000 组样本进行后验统计,得出参数后验均值估计 M_{MP}. 并将表 6 中的真实参数 α 一并代入 MRE(MP)表达式,如下:

$$\mathrm{MRE(MP)}=\left(\sum_{j=1}^{20}\sum_{k=1}^{5}\frac{|M_{\mathrm{MP}}(j,k)-\alpha(j,k)|}{\alpha(j,k)}\right)/100 \tag{20}$$

式中:j 为第 j 组真实参数;α,k 表示参数 α 的第 k 个分量.由式(20)得出 C_7^i 种监测方案的反演结果 MRE(MP),如表 7 所示.

表 6 从参数的先验分布中得到 20 组真实参数

Tab.6 20 sets of real parameters obtained from the prior distribution

序号	X_s	Y_s	T_1	T_2	C_s
1	81.445	266.32	11.202	19.934	3 011.6
2	154.68	272.08	9.106 1	21.293	2 905.6
3	211.85	306.9	9.856	18.914	3 153.5
4	68.409	243.52	10.413	19.752	3 202.8
5	136.61	368.58	10.714	20.395	3 186.7
6	173.72	390.85	11.569	18.36	2 840.1
7	190.65	291.92	9.262 2	19.551	2 967.2
8	140.44	283.7	11.952	18.445	3 037.8
9	100.69	324.16	11.132	21.105	3 275.6
10	61.521	376.01	10.282	21.679	3 240.5
11	196.38	347.49	8.128 5	20.685	3 265.5
12	90.543	263.18	8.809 1	21.54	2 815.4
13	113.57	300.91	10.111	21.966	2 939.1
14	124.81	319.86	8.674 5	19.082	3 103.5
15	121.8	335.98	9.494 3	20.135	3 073.7
16	96.349	367.22	8.350 7	20.461	2 888.5
17	215.98	341.87	10.833	19.306	2 864.6
18	167.95	396.6	11.749	18.061	2 993.6
19	158.91	252.55	8.459 2	18.613	3 150.0
20	185.84	355.8	9.637	20.92	3 089.4

表 7 各监测类型下各监测方案的 E(MP)和 MRE(MP)值

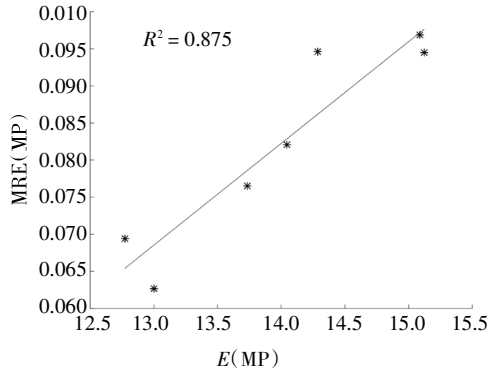
Tab.7 E(MP) and MRE(MP) of every monitoring scheme within different monitoring types

类型	MP	E(MP)	MRE(MP)	MP	E(MP)	MRE(MP)
1 眼井	1	15.123	0.094	5	14.286	0.095
	2	13.004	0.063	6	15.090	0.097
	3	12.772	0.069	7	14.047	0.082
	4	13.734	0.076			
	最优方案			3		
2 眼井	(1,2)	10.862	0.054	(3,4)	9.898	0.044
	(1,3)	10.847	0.046	(3,5)	10.475	0.054
	(1,4)	11.221	0.053	(3,6)	10.785	0.052
	(1,5)	12.680	0.088	(3,7)	10.075	0.043
	(1,6)	12.652	0.071	(4,5)	10.658	0.050
	(1,7)	11.174	0.047	(4,6)	11.101	0.052
	(2,3)	9.723	0.045	(4,7)	11.066	0.052
	(2,4)	11.314	0.058	(5,6)	12.660	0.074
	(2,5)	10.379	0.052	(5,7)	10.822	0.052
	(2,6)	10.643	0.052	(6,7)	11.303	0.055
	(2,7)	10.428	0.048			
	最优方案			(2,3)		
3 眼井	(1,2,3)	9.478	0.044	(2,3,7)	9.397	0.043
	(1,2,4)	9.855	0.050	(2,4,5)	9.591	0.053
	(1,2,5)	9.873	0.047	(2,4,6)	9.817	0.057
	(1,2,6)	9.840	0.044	(2,4,7)	10.039	0.046
	(1,2,7)	9.546	0.046	(2,5,6)	9.864	0.048
	(1,3,4)	9.549	0.046	(2,5,7)	9.447	0.046
	(1,3,5)	10.200	0.049	(2,6,7)	9.647	0.048
	(1,3,6)	10.268	0.049	(3,4,5)	9.399	0.046
	(1,3,7)	9.671	0.045	(3,4,6)	9.481	0.043
	(1,4,5)	9.967	0.043	(3,4,7)	9.549	0.043
	(1,4,6)	10.006	0.043	(3,5,6)	10.190	0.053
	(1,4,7)	9.823	0.044	(3,5,7)	9.498	0.047
	(1,5,6)	11.666	0.068	(3,6,7)	9.615	0.043
	(1,5,7)	10.134	0.047	(4,5,6)	10.092	0.050
	(1,6,7)	10.256	0.046	(4,5,7)	9.669	0.046
	(2,3,4)	9.461	0.046	(4,6,7)	10.074	0.049
	(2,3,5)	9.377	0.044	(5,6,7)	10.369	0.055
	(2,3,6)	9.431	0.044			
	最优方案			(2,3,5)		

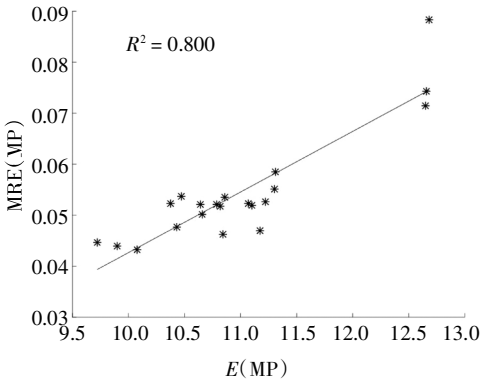
续表

类型	MP	E(MP)	MRE(MP)	MP	E(MP)	MRE(MP)
4 眼井	(1,2,3,4)	9.325	0.044	(1,4,6,7)	9.520	0.041
	(1,2,3,5)	9.344	0.044	(1,5,6,7)	9.918	0.049
	(1,2,3,6)	9.368	0.043	(2,3,4,5)	9.277	0.045
	(1,2,3,7)	9.302	0.042	(2,3,4,6)	9.305	0.044
	(1,2,4,5)	9.404	0.046	(2,3,4,7)	9.344	0.044
	(1,2,4,6)	9.412	0.045	(2,3,5,6)	9.333	0.045
	(1,2,4,7)	9.407	0.045	(2,3,5,7)	9.264	0.042
	(1,2,5,6)	9.624	0.044	(2,3,6,7)	9.296	0.042
	(1,2,5,7)	9.343	0.041	(2,4,5,6)	9.423	0.050
	(1,2,6,7)	9.371	0.044	(2,4,5,7)	9.325	0.043
	(1,3,4,5)	9.356	0.043	(2,4,6,7)	9.475	0.047
	(1,3,4,6)	9.383	0.044	(2,5,6,7)	9.374	0.046
	(1,3,4,7)	9.368	0.043	(3,4,5,6)	9.344	0.045
	(1,3,5,6)	9.994	0.050	(3,4,5,7)	9.301	0.044
	(1,3,5,7)	9.446	0.045	(3,4,6,7)	9.358	0.043
	(1,3,6,7)	9.498	0.044	(3,5,6,7)	9.442	0.047
	(1,4,5,6)	9.703	0.044	(4,5,6,7)	9.560	0.048
	(1,4,5,7)	9.432	0.041			
	最优方案			(2,3,5,7)		
5 眼井	(1,2,3,4,5)	9.263	0.042	(1,3,4,5,7)	9.284	0.041
	(1,2,3,4,6)	9.269	0.042	(1,3,4,6,7)	9.298	0.043
	(1,2,3,4,7)	9.280	0.043	(1,3,5,6,7)	9.393	0.046
	(1,2,3,5,6)	9.319	0.043	(1,4,5,6,7)	9.389	0.042
	(1,2,3,5,7)	9.256	0.040	(2,3,4,5,6)	9.259	0.043
	(1,2,3,6,7)	9.265	0.042	(2,3,4,5,7)	9.252	0.042
	(1,2,4,5,6)	9.330	0.044	(2,3,4,6,7)	9.275	0.042
	(1,2,4,5,7)	9.281	0.040	(2,3,5,6,7)	9.259	0.043
	(1,2,4,6,7)	9.300	0.043	(2,4,5,6,7)	9.297	0.043
	(1,2,5,6,7)	9.310	0.042	(3,4,5,6,7)	9.280	0.044
	(1,3,4,5,6)	9.322	0.042			
	最优方案			(2,3,4,5,7)		
6 眼井	(1,2,3,4,5,6)	9.253	0.041	(1,2,4,5,6,7)	9.265	0.041
	(1,2,3,4,5,7)	9.247	0.039	(1,3,4,5,6,7)	9.270	0.042
	(1,2,3,4,6,7)	9.255	0.042	(2,3,4,5,6,7)	9.249	0.043
	(1,2,3,5,6,7)	9.250	0.041			
	最优方案			(1,2,3,4,5,7)		
7 眼井	(1,2,3,4,5,6,7)	9.247	0.041			

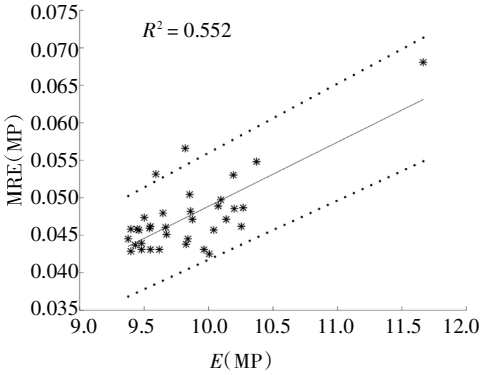
分别将表 7 中 1 眼井监测方案,2 眼井组合监测方案以及 3 眼井组合监测方案下的 MRE(MP)与 E(MP)进行线性拟合,结果见图 3 和表 8. 由于 4 眼井组合监测方案,5 眼井组合监测方案以及 6 眼井组合监测方案下 E(MP)数值接近,同时 MRE(MP)数值也十分接近，并且考虑到计算误差的影响,不再对这 3 种类型监测方案下 MRE（MP）与 E(MP)进行线性拟合.



(a) 1 眼井监测方案



(b) 2 眼井监测方案



(c) 3 眼井监测方案

图 3 不同监测类型下 E(MP) 和 MRE(MP)的线性拟合图

Fig.3 Linear fitting diagrams of the relationship between E(MP) and MRE(MP) within different monitoring types

表 8 不同监测类型下 MRE(MP)与 E(MP)线性拟合方程

Tab.8 Linear fitting equations between MRE(MP) and E(MP) within different monitoring types

监测类型	拟合方程	决定系数 R^2
1 眼井	MRE(MP)=0.014E(MP)−0.110	0.875
2 眼井	MRE(MP)=0.012E(MP)−0.076	0.800
3 眼井	MRE(MP)=0.009E(MP)−0.037	0.552

由于图 3(c)中决定系数 R^2=0.552 比较小,运用 F 检验对图 3(c)中线性方程的显著性进行检验[32].

检验假设 H_0: E(MP) 和 RT(MP)之间没有真正的线性关系,H_1: E(MP)和 RT(MP)之间有线性关系;显著性水平 α = 0.05.

经计算可得 F 检验统计量 F = 40.739, 而 F_α(1,35−2) = F_α(1,33) = 4.139,因此 $F > F_\alpha$(1,33),拒绝 H_0. F 检验表明 E(MP) 和 RT(MP)之间存在线性关系. 另外图 3(c)中实线表示拟合直线,两侧虚线标出其 95%置信区间,只有 2 个点落在置信区间外面,也充分说明了 E(MP) 和 RT(MP)之间具有正线性关系.

综上,由图 3 和表 8 可知,MRE(MP)与 E(MP)呈较好的正相关关系,说明 E(MP)是参数反演结果精度的有效量度,E(MP)越小,参数反演结果精度越高. 但表 7 中 E (MP) 取得最小值的监测方案的 MRE(MP)未必是最小值(如 1 眼监测井情况下,选取 3 监测方案时 E (MP) 为 12.772,MRE (MP)为 0.069;选取 2 监测方案时 E(MP)为 13.004,MRE (MP)为 0.063),主要原因在于选取 20 组参数真值(表 6)时,虽然运用拉丁超立方抽样方法使 20 组参数真值尽量均匀分布在参数先验范围内,但由于数据较少, 无法在真正意义上均匀分布在先验范围内,而信息熵的求解运用蒙特卡洛方法(MC 方法)在参数先验范围内运用拉丁超立方抽样方法抽取样本 20 000 次,二者相比,因此认为以 E(MP)最小值作为选取最优监测方案的指标更加可信,而不能以 MRE(MP)最小值作为选取最优监测方案的指标.

另由表 7 可知,并不是任何条件下监测井数量越多,信息熵越小,如两眼监测井情况下,(2,3)组合方式下的 E (MP) 为 9.723,3 眼监测井情况下,(1,5,6)组合方式下的 E(MP)为 11.666. 表 7 显示每种监测类型下均存在信息熵最小的最优监测井方案. 因此从表 7 中各监测类型中筛选出信息熵最小的监测方案作为相应监测井组合方案中的最优方案,并绘制不同监测类型中最优方案的信息熵 E(MP)随监测井数量的变化曲线,如图 4 所示.

由图 4 可知, 在各监测类型的最优方案条件下,信息熵随着监测井数量的增加而减小,其中两眼井的信息熵显著小于一眼井的信息熵,但是与其他数量的监测井的信息熵相差不大.

通常情况下, 对于利用监测井进行污染源识别,既要考虑反演精度,又要考虑监测成本. 因此本算例选取 3 种监测方案进行污染源识别. 方案 1:监测成本最小的 1 眼监测井方案 (3 号单井). 方案 2:兼顾反演精度与监测成本的监测井方案: 由于 2 眼井的信息熵与 3~7 眼井的信息熵相差不大,但随着监测井数量的增加,监测费用显著增加,因此该方案选 2 眼井监测方案(监测井(2,3)组合). 方案 3:基于较高反演精度的监测井方案,由于 3~7 眼井的信息熵相差不大,认为反演精度相近,因此该方案选 3 眼井监测方案(监测井(2,3,5)组合).

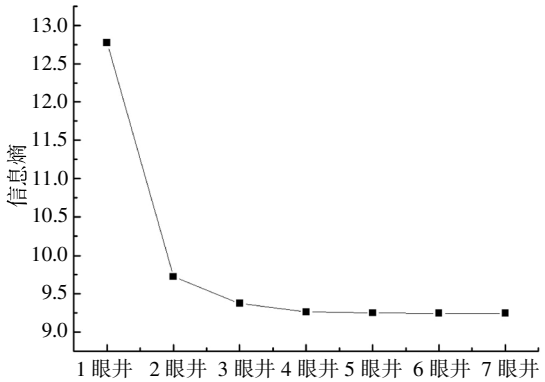


图 4 不同监测类型中最优方案的信息熵随监测井数量的变化曲线

Fig.4 The change curve of the information entropy of the optimal monitoring schemes under different monitoring types with the number of monitoring wells

2.4 基于优化监测方案的污染源识别

以表 4 中第 1 组参数真值 α=(X_s,Y_s,T_1,T_2,C_s)=(81.4,266.3,11.2,19.9,3 011.6)为例,分别利用 2.3 节选取的 3 种监测方案,运用 DREAM 算法(初始平行链数为 10,反演稳定后平行链数记为 n)进行污染源参数反演,其中每条马尔科夫链长度是 12 000.

通过计算可知,各监测方案下,在平行链长度为 10 000 时,5 个参数的收敛性判断指标$\hat{R}_i$<1.2(i = 1,2,…,5),此时 DREAM 算法所得马尔科夫链都已收敛. 剔除平行马尔科夫链前 10 000 组样本后,稳定后的剩余 2 000 组样本的后验分布范围如表 9 所示,分布直方图(竖直线代表参数真值)如图 5 所示.

表 9 3 种监测方案下模型参数后验分布范围

Tab.9 The posterior distribution ranges of model parameters under three monitoring schemes

方案	信息熵	反演稳定后平行链数 n	参数后验分布范围				
			X_s	Y_s	T_1	T_2	C_s
1	12.772	8	[60,124]	[240,288]	[8,12]	[18,22]	[2 800,3 300]
2	9.723	8	[60,92]	[256,272]	[8.5,12]	[18,21.5]	[2 800,3 300]
3	9.377	8	[60,92]	[256,272]	[9.5,12]	[18,20.5]	[3 000,3 300]

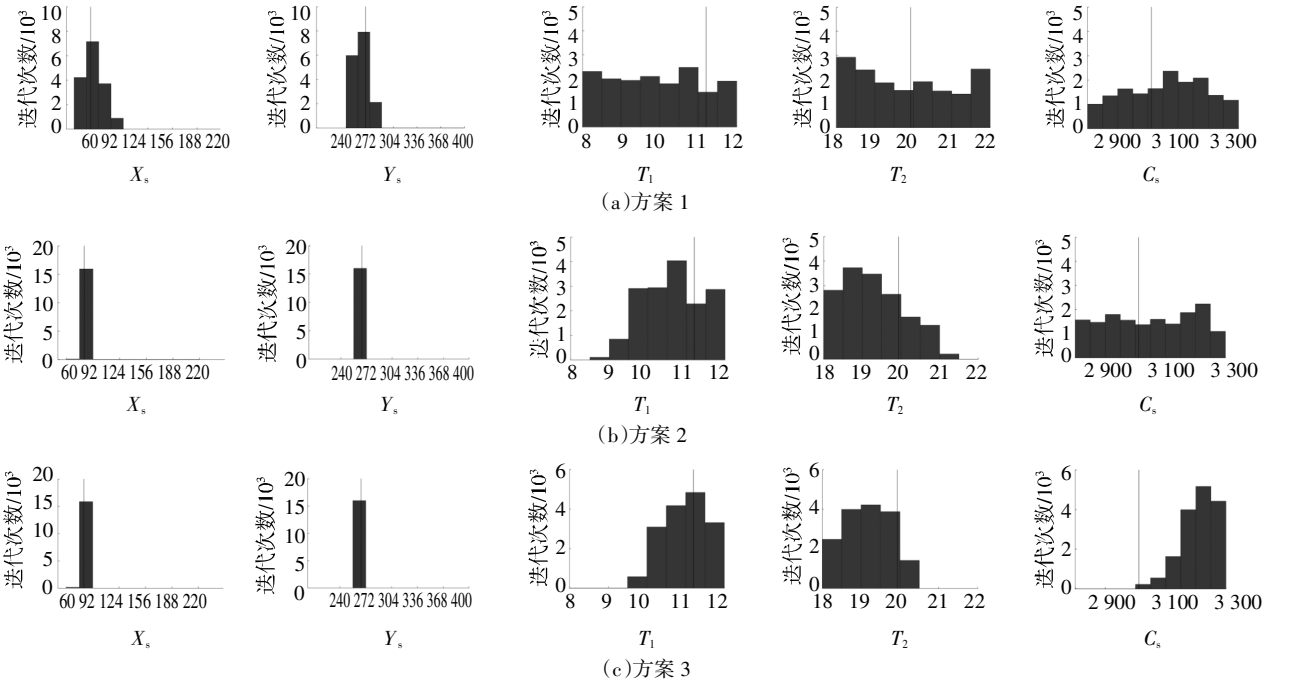


图 5　3 种监测方案下模型参数后验分布直方图

Fig.5　The posterior distribution histograms of model parameters under three monitoring schemes

由表 9 和图 5 可知,方案 1 到方案 3 参数后验分布范围逐渐减小,与相应监测方案信息熵的变化趋势一致(图 4),且参数真值均在参数后验分布范围内.进一步表明监测方案信息熵越小,参数后验分布不确定性越小.

为进一步说明 3 种监测方案的参数反演效果,对马尔科夫链稳定后的剩余 2 000 个样本点进行后验统计分析,结果见表 10.

表 10　3 种监测方案下模型参数后验统计结果及收敛性判断指标

Tab.10　Posterior statistical results of model parameters and convergence judgment indicators under three monitoring schemes

参数真值	方案 1					方案 2					方案 3				
	均值	均值相对误差/%	标准差	均值 95% 置信区间	$\hat{R}_i$	均值	均值相对误差/%	标准差	均值 95% 置信区间	$\hat{R}_i$	均值	均值相对误差/%	标准差	均值 95% 置信区间	$\hat{R}_i$
X_s=81.4	86.14	5.77	4.28	[85.95,86.33]	1.01	82.15	0.87	1.16	[82.10,82.20]	1.03	81.13	0.39	0.54	[81.11,81.15]	1.00
Y_s=266.3	259.86	2.42	3.80	[259.70,260.03]	1.05	265.11	0.46	0.42	[265.09,265.13]	1.07	265.20	0.42	0.24	[265.19,265.21]	1.01
T_1=81.4	9.88	11.81	0.39	[9.86,9.90]	1.03	10.63	5.06	0.27	[10.62,10.65]	1.08	11.12	0.75	0.12	[11.11,11.12]	1.02
T_2=81.4	19.97	0.18	0.42	[19.95,19.99]	1.01	19.30	3.18	0.35	[19.29,19.32]	1.10	19.31	3.15	0.15	[19.30,19.31]	1.02
C_s=3 011.6	3 047.89	1.21	50.41	[3 045.68,3 050.10]	1.03	3 053.00	1.37	49.61	[3 050.83,3 055.18]	1.07	3 209.49	6.57	18.17	[3 208.70,3 210.29]	1.01

由表 10 可知，从方案 1 到方案 3,5 个参数的标准差均逐渐减小;方案 2 与方案 3 下 5 个参数后验均值相对误差数值接近，与方案 1 相比,X_s,Y_s 以及 T_1 的后验均值相对误差均大幅减少,但是 T_2、C_s 2 个参数的后验均值相对误差却有所增大.结合图 3、图 4 及表 10 分析，进一步说明参数后验分布的信息熵与 5 个参数后验均值相对误差的平均值成正比,但不能保证参数后验分布的信息熵与每个参数后验均值相对误差成正比,即运用参数后验分布的信息熵作为监测井优化设计的指标,在参数整体反演精度最高的情况下,不能保证每个参数的反演效果均达到最佳.

监测井数越多所需费用越多,单井监测费用最少.方案 2 与方案 3 信息熵差距很小,5 个参数后验样本均值相对误差的平均值也近似相等,却增加了 1 眼井进行监测，费用明显增加.对于本文案例,如果需要综合考虑监测成本及参数后验分布范围大小,认为方案 2 为最佳的监测井方案.

3　结　论

1)与拉丁超立方抽样方法相比,最优拉丁超立方抽样可有效提高参数先验分布抽样的均匀性,避

免抽样结果的随机性及对整个抽样空间的覆盖填充程度的差异性.

2)Kriging 替代模型属于黑箱模型,能以较小的计算量得到和地下水数值模型相近的输入输出关系,在保证模拟精度的条件下,显著降低计算负荷.

3)参数反演结果的相对均方根误差与监测方案信息熵呈现较好的正相关关系. 信息熵是参数后验分布不确定性的有效量度,信息熵越小,参数后验范围越小, 基于贝叶斯公式与信息熵的监测方案优化设计方法是确定地下水污染监测方案的有效方法.

4)并非监测井数量越多越有利于污染源的反演识别,必须以信息熵为筛选指标制定各监测类型下的最优监测井组合方案, 而且可以以监测成本、监测效率、反演精度等为限制条件进行具体工况条件下最优监测方案的选择.

参考文献

[1] CZANNER G,SARMA S V,EDEN U T,*et al*. A signal-to-noise ratio estimator for generalized linear model systems [J]. Lecture Notes in Engineering & Computer Science,2008,2171:1063—1069.

[2] HUAN X,MARZOUK Y M. Simulation-based optimal Bayesian experimental design for nonlinear systems [J]. Journal of Computational Physics,2013,232(1):288—317.

[3] LINDLEY D V. On a measure of the information provided by an experiment [J]. The Annals of Mathematical Statistics,1956,27(4):986—1005.

[4] SHANNON C E. A mathematical theory of communication [J]. The Bell System Technical Journal,1948,27(3):379—423.

[5] SOHN M D,SMALL M J,PANTAZIDOU M. Reducing uncertainty in site characterization using bayes monte carlo methods [J]. Journal of Environmental Engineering-asce,2000,126(10):893—902.

[6] CHEN M,IZADY A,ABDALLA O A,*et al*. A surrogate-based sensitivity quantification and Bayesian inversion of a regional groundwater flow model [J]. Journal of Hydrology,2018,557:826—837.

[7] SNODGRASS M F,KITANIDIS P K. A geostatistical approach to contaminant source identification [J]. Water Resources Research,1997,33(4):537—546.

[8] RUZEK B,KVASNICKA M. Differential evolution algorithm in the earthquake hypocenter location [J]. Pure and Applied Geophysics,2001,158:667—693.

[9] GIACOBBO F,MARSEGUERRA M,Zio E. Solving the inverse problem of parameter estimation by genetic algorithms:the case of a groundwater contaminant transport model [J]. Annals of Nuclear Energy,2002,29(8):967—981.

[10] DOUGHERTY D E,MARRYOTT R A. Optimal groundwater management:simulated annealing [J]. Water Resources Research,1991,27(10):2493—2508.

[11] TANNER M A. Tools for statistical inference:methods for the expectation of posterior distribution and likelihood functions [M]. Berlin:Springer,1996:14—39.

[12] ROBERTS C P,CASELLA G. Monte carlo statistical methods [M]. 2nd ed. Berlin:Springer,2004:79—122.

[13] METROPOLIS N,ROSENBLUTH A W,ROSENBLUTH M N,*et al*. Equation of state calculations by fast computing machines [J]. The Journal of Chemical Physics,1953,21(6):1087—1092.

[14] HASTINGS W K. Monte Carlo sampling methods using Markov chains and their applications [J]. Biometrika,1970,57(1):97—109.

[15] TIERNEY L,MIRA A. Some adaptive Monte Carlo methods for bayesian inference [J]. Statistics in Medicine,1999,18:2507—2515.

[16] MIRA A. Ordering and improving the performance of Monte Carlo Markov Chains [J]. Statistical Science,2002,16:340—350.

[17] HAARIO H,SAKSMAN E,TAMMINEN J. An adaptive metropolis algorithm [J]. Bernoulli,2001,7(2):223—242.

[18] HAARIO H,LAINE M,MIRA A. DRAM:efficient adaptive MCMC [J]. Statistics and Computing,2006,16(4):339—354.

[19] 张江江. 地下水污染源解析的贝叶斯监测设计与参数反演方法 [D]. 杭州:浙江大学环境与资源学院,2017.
ZHANG J J. Bayesian monitoring design and parameter inversion for groundwater contaminant source identification [D]. Hangzhou: College of Environmental and Resource Sciences,Zhejiang University,2017. (In Chinese)

[20] TER BRAAK C J F. A Markov Chain Monte Carlo version of the genetic algorithm differential evolution:easy Bayesian computing for real parameter spaces [J]. Statistics and Computing,2006,16(3):239—249.

[21] VRUGT J A,TER BRAAK C J F,DIKS C G H,*et al*. Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling [J]. International Journal of Nonlinear Sciences and Numerical Simulation,2009,10(3):273—290.

[22] KNILL D L,GIUNTA A A,BAKER C A,*et al*. Response surface models combining linear and Euler aerodynamics for supersonic transport design [J]. Journal of Aircraft ,1999,36(1):75—86.

[23] LI J,CHEN Y,PEPPER D. Radial basis function method for 1-D and 2 -D groundwater contaminant transport modeling [J]. Computational Mechanics,2003,32(1):10—15.

[24] 肖传宁,卢文喜,赵莹,等.基于径向基函数模型的优化方法在地下水污染源识别中的应用 [J]. 中国环境科学,2016,36(7):2067—2072.
XIAO C N,LU W X,ZHAO Y,*et al*. Optimization method of identification of groundwater pollution sources based on radial basis function model [J]. China Environmental Science,2016,36(7):2067—2072. (In Chinese)

[25] CHRISTIE M,DEMYANOV V,ERBAS D. Uncertainty quantification for porous media flows [J]. Journal of Computational Physics,2006,217(1):143—158.

[26] MATHERTON G. Principles of geostatistics [J]. Economic Geology,1963,58(8):1246—1266.

[27] SACKS J,WELCH W J,MITCHELL T J,*et al*. Design and analysis of computer experiments [J]. Statistical Science,1989,4(4):409—435.

[28] 高月华. 基于 Kriging 代理模型的优化设计方法及其在注塑成型中的作用 [D]. 大连: 大连理工大学运载工程与力学学部,2008.
GAO Y H. Optimization methods based on Kriging surrogate model and their application in injection molding [D]. Dalian:Faculty of Vehicle Engineering and Mechanics,Dalian University of Technology,2008. (In Chinese)

[29] LOPHAVEN S N,NIELSEN H B,SONDERGAARD J. Dace:A MATLAB Kriging toolbox [R]. Kongens Lyngby: Technical University of Denmark,Technical Report No. IMM-TR-2002—12.

[30] GELMAN A G,RUBIN D B. Inference from iterative simulation using multiple sequences [J]. Statistical Science,1992,7:457—472.

[31 HICKERNELL F J. A generalized discrepancy and quadrature error bound [J]. Mathematics of Computation of the American Mathematical Society,1998,67(221):299—322.

[32] 周圣武,李金玉,周长新. 概率论与数理统计[M].2 版. 北京:煤炭工业出版社,2007:208—215.
ZHOU S W,LI J Y,ZHOU C X. Probability theory and mathematical statistics [M].2nd ed. Beijin:China Coal Industry Publishing House,2007:208—215.(In Chinese)