Research on the Optimization of Model Parameters Based on Particle Filter

Weinan Cui[1], Chong Chen[2,3,*], Xiangli Cui[4], Qian He[1] and Mingda Zhu[2,3]
[1]China Academy of Information and Communications Technology
[2]College of Geophysics and Information Engineering, China University of Petroleum – Beijing
[3]State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, China
[4]Research Institute of Petroleum Exploration & Development-Northwest(NWGI), PetroChina
[*]Corresponding author

Abstract—**With the significant advancements in Information and Communications Technology (ICT), computer simulations have been widely used in natural sciences and engineering analysis. Algorithms for parameters optimization in computer models have been necessity in computer modeling due to the accuracy requirements. As a principal branch of computer models, hydrological models have been a fundamental method for researching hydrological processes. In this paper, a computer model of the groundwater system in the middle reaches of the Heihe River Basin was established. Geological features of different regions were characterized by sub-zones of parameters in the model which were optimized by Particle Filter. The effectiveness for Particle Filter of optimizing the parameters of numerical models was verified. The results indicated significant improvements of parameters after 100 time-steps which converged to optimal value. Meanwhile, the difference between simulated and observed groundwater level was reduced along with the parameters convergence.**

Keywords—computer simulation; parameter optimization; particle filter

I. INTRODUCTION

The purpose of parameters optimization in computer modeling is to match model outputs and observations as much as possible by minimizing objective function. Direct and indirect optimization algorithms are the most common optimization algorithms. Differentiability is the principal limitation of indirect algorithm (such as Newton's method and its improvements). Compared to indirect algorithm, differentiability is not a necessity for direct algorithm by calculating the value of objective functions. Therefore, although the computation cost of direct algorithm is larger then the indirect one, the direct algorithm is able to avoid the indifferentiability in indirect algorithms and have been widely used in the applications of environmental simulations. Currently, groundwater cycle and associated processes which could be megascopic or microscopic can not be fully simulated by groundwater models which are based on prior measurements and parameters in other documents. Therefore, parameters optimization based on measurements are necessary for groundwater models.

In a Bayesian framework, the mean and variance of state variables could be obtained from the posterior probability density function of system states and parameters only if the posterior probability density function of system states and parameters could be estimated from measurements. The Kalman Filter (KF) is the most popular algorithm to estimate posterior probability density function for Linear Gaussian Systems. Later in 1960s, the Extended Kalman Fileter (ETF) and Gaussian Sum Filter were introduced to approximate the posterior probability density function of states and parameters for Nonlinear non-Gaussian Systems. However, the accuracy, stability and convergence of these algorithms cannot satisfy the requirements. Early in 1950s, Sequential Importance Sampling (SIS) method which was based on Monte Carlo method was introduced by [1]. However, SIS was not evolved until 1993 with the proposal of resampling technique and improvements of computation resources, after which, SIS have been applied to many fields. Recent years, researches have proposed many nonlinear filters based on Bayesian estimation (Bootstrap Filter [2], Condensation Tracks [3], Particle Filter [4], Sequential Monte Carlo Method [5], Interacting Particle Approximations [6]) in the fields of fault diagnosis, signal transmission and compression, target tracking, navigation and financial data analysis. Particle Filter (PF) was proposed as a general name for "bootstrap", "Survival of the fittest" or "Sequential Monte Carlo".

KF is able to approach the optimal solution by recursively estimating state variables from measurements under the assumption of Linear Gaussion System. In PF, particles (samples) are used to approximate the prior probability density function and Bayesian Theory is used to update the particles in order to estimate posterior probability density function which avoids the assumption of Linear Gaussion System. Therefore, PF has intrinsic advantages to solve nonlinear problems. With the rapid improvement of computation capacity and the resolving of particle degeneracy, PF has been widely used in many fields [7].

In this paper, we established a groundwater model in the middle reaches of the Heihe River Basin using MODular three-dimensional finite-difference ground-water FLOW model (MODFLOW). Based on the observed groundwater level, parameters of the groundwater model were estimated and optimized using Particle Filter

II.   Method

A.   Particle Filter

Particle Filter (PF) is a kind of Sequential Monte Carlo methods based on Bayesian Theory and has been widely used in resolving problems related to Sequential State Estimation. The basic idea of PF is that the posterior probability density function of system states is approximated by a set of particles (samples) and the corresponding weights for each particle. First, n particles ($\{x_{k-1}^{(1)}, x_{k-1}^{(2)}, ..., x_{k-1}^{(N)}\}$) are sampled from the posterior probability density function of system states ($P(x_{k-1}|z_{k-1})$) in time k-1. Second, the particles ($\{x_k^{(1)}, x_k^{(2)}, ..., x_k^{(N)}\}$) and weights ($\{w_k^{(1)}, w_k^{(2)}, ..., w_k^{(N)}\}$) in time k are estimated and updated based on the given state transition function and the measurement function when the observation of time k becomes available. These particles are then used to approximate the posterior probability density function of system states. As the number of particles becomes large enough, the estimation of PF reaches the optimal solution.

There are three foundations of PF which are: (1) Recursive Beyesian Estimation, (2) Sequential Importance Sampling and (3) Sampling Importance Resampling.

1)   Recursive beyesian estimation
We consider the following nonlinear system:

$$X_k = M(X_{k-1}) + w_k \tag{1}$$

where X represents the vector of all state variables and parameters with length N_x, M(•) represents the nonlinear state transition function (model operator). w_k denotes the noise processes at time step k.

After the measurements at time step k become available. The parameters and states variables can be transferred to measurement space using (2).

$$Y_k = H(X_k) + \varepsilon_k \tag{2}$$

where Y_k represents measurements at time step k. H denotes observation operator as a function of model parameters and forecasted state variables. ε_k is the observational noise.

Let $D_n = \{Y_i; i = 1, 2, ..., n\}$ be the set of all the available observations at time step k. Assume the posterior probability distribution of state variables $p(X_{k-1}|D_{k-1})$ at time step k-1 is known. By applying Chapman-Kolmogoroff (CK) equation, the prior probability distribution of state variables at time step k can be calculated as:

$$p(X_k | D_{k-1}) = \int p(X_k | X_{k-1}) p(X_{k-1} | D_{k-1}) dX_{k-1} \tag{3}$$

where $p(X_k|X_{k-1})$ is state transition probability density function which is determined by (1). Equation (3) indicates that the prior probability distribution of state variables at time step k summarize all the information of state variables. When the observations at Y_k becomes available, the posterior probability distribution of state variables can be calculated by applying Bayes' rule:

$$p(X_k | D_k) = \frac{p(Y_k | X_k) p(X_k | D_{k-1})}{p(Y_k | D_{k-1})} \tag{4}$$

2)   Sequential importance sampling
In PF, the posterior probability distribution of state variables is approximated by particles (samples) and their corresponding weights:

$$p(X_{0:k} | Y_{1:k}) \approx \sum_{i=1}^{N_p} w_k^i \delta(X_{0:k} - X_{0:k}^i) \tag{5}$$

where, $\{X_k^i, w_k^i\}$ represent the ith particle and its weight at time step k, respectively. N_p denotes the total number of particles. $\delta(•)$ represents the Dirac delta function.

The weights of particles should satisfy:

$$w_k^i \propto \frac{p(X_{0:k}^i | Y_{1:k})}{q(X_{0:k}^i | Y_{1:k})} \tag{6}$$

The weights of particles are updated following [5, 8]:

$$w_k^i = w_{k-1}^i \frac{p(Y_k | X_k^i) p(X_k^i | X_{k-1}^i)}{q(X_{k-1}^i | X_{0:k-1}^i, Y_{1:k})} \tag{7}$$

3)   Sampling importance resampling
Resampling is introduced to resolving the particle degeneracy which is able to eliminate particles with lower weights and increase particles with higher weights.

4)   Brief implementation of PF
Step 0: Initialization

$K = 0$

$N_{th} = N_0$

For $i = 1, 2, ..., N$, sample particles $x_0^i \sim p(x_0)$.

Let $k = 1$.

Step 1: Sequential Importance Sampling

For $i = 1, 2, ...$, sample particles $x_0^i \sim p(x_k | x_{k-1}^i)$.

Let $x_{0:k}^i = (x_{0:k-1}^i, x_k^i)$

For $i = 1, 2, ..., N$, estimates the weights for particles

$$w_k = w_{k-1} \frac{p(\mathrm{y}_k \mid \mathrm{x}_k)p(\mathrm{x}_k \mid \mathrm{x}_{k-1})}{q(\mathrm{x}_k \mid \mathrm{x}_{0:k-1}, \mathrm{D}_k)}$$

Normalize the weights,

$$w_k^i = \frac{w_k^i}{\sum_{j=1}^{N} w_k^j}$$

Step 2: Sampling Importance Resampling

Eliminate the particles with lower weights and resample particles with higher weights

Step 3: Output

Step 4: Let $k = k + 1$, repeat Step 0 to Step 5.

B. Groundwater Model

Modular three-dimensional Finite-difference ground-water Flow Model (MODFLOW) was used to establish the groundwater model. MODFLOW was developed by U.S. Geological Survey (USGS) in 1980s and had been widely used all over the world since then. There were many packages in MODFLOW which could simulate hydrology processes such as streamflow, drain, evapotranspiration, precipitation, irrigation and recharge. The concept of Stress Period was used in the partial-differential equation for the three-dimensional movement of ground water (refer to (8)). The whole simulation period was separated into several Stress Periods which were further separated into several Time Steps. During one Stress Period, all the parameters, boundaries and model settings were identical.

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right) - W = S_s\frac{\partial h}{\partial t} \quad (8)$$

where K_{xx}, K_{yy}, K_{zz} are values of hydraulic conductivity along the x, y and z coordinate axes [LT^{-1}]; h represents the potentiometric head [L]; W is a volumetric flux per unit volume representing sources and/or sinks of water, with W < 0.0 for flow out of the groundwater system, and W > 0.0 for flow into the system [T^{-1}]; S_s is the specific storage of the porous material [L^{-1}]; and t is time [T]. In MODFLOW, the finite-difference method is applied to solve the equation approximately. As shown in Figure I, the aquifer system is discretized to a grid of blocks called cells, the locations of which are described in terms of rows, columns and layers. An i, j, k indexing system is used.
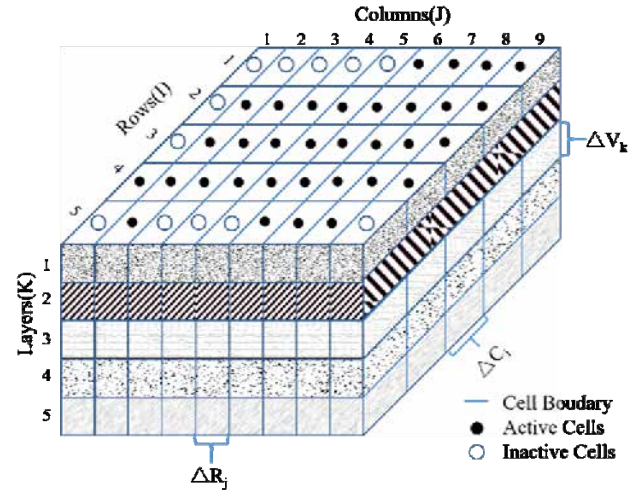

FIGURE I. SPACIAL DISCRETIZATION OF AN AQUIFER.

C. Data Collection and Model Settings

The data which were collected from the middle reaches of the Heihe River Basin (HRB) were used to construct the groundwater model. The middle reaches of the HRB which located in the northwest of China (38°38' ~ 39°53' N, 98°53' ~ 100°44' E; Figure II (a)) was dominated by a typically arid continental climate with limited rainfall, but strong evaporation. The data include Digital Elevation Model (DEM), land use, groundwater pumping yields, observed groundwater level and streamflow rates. Land use data were obtained through visual interpretation of Landsat TM/ETM+ images in 1986 [9], 2000 [10], 2007 [11]. The cultivated area was clipped directly from the 1:100,000-scale land use database developed by the Chinese Academy of Sciences (CAS) using the ArcGIS software. The observed groundwater levels from 42 boreholes which were collected by the Gansu Provincial Bureau of Hydrology were used to calibrate the model. We set the stress period as one month due to the availability of data. All the above-mentioned data were obtained from the WestDC [12].

The study area was discretized to 132 rows and 165 columns with a uniform cell size 1 × 1 km which were also shown in Figure II (a) and Figure II (b). The boundaries of the study area were coincided with earlier studies [13-15] and defined by the natural boundaries. A no-flow boundary was defined between A-E, because a groundwater divide was present at this boundary. E was the outlet of HeiHe River in the middle reach which was coincident with the Zhengyi Gorge reservoir. The groundwater flow from the mountain to the model domain through D-E cannot be exactly quantified. No-flow boundary was defined between D-E as the hydraulic conductivity in the hard rock was significantly smaller than that of the basin sediments according to a previous study [16]. Detail settings of the model boundaries and parameters refer to [17].
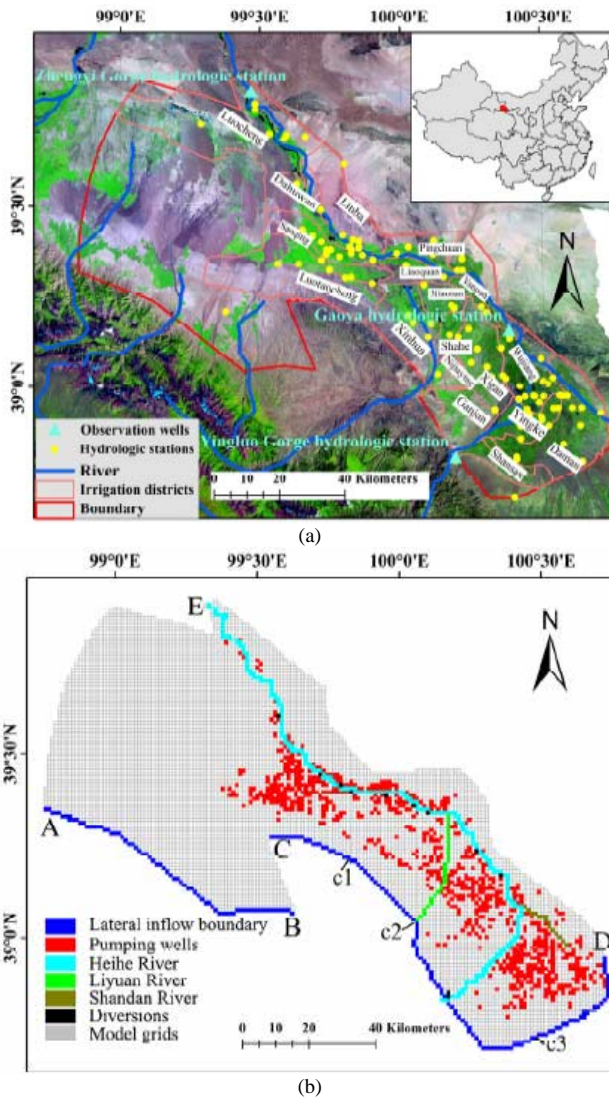
FIGURE II. (A)THE STUDY AREA AND (B) MODEL DOMAIN

III. RESULTS AND ANALYSIS

A. Simulation

The simulation results of groundwater level at three typical boreholes were used to demonstrate and analyze due to the relatively large number of boreholes. In the initialization step, one hundred ($N = 100$) particles (samples) were randomly generated from normal distribution for each parameter. The simulated groundwater levels were calculated based on each set of parameters. The red curve in Figure Ⅲ which indicated a decline trend showed the observed groundwater level at each borehole. The green curves were the mean value of the groundwater level which were calculated from random p=articles (the blue curves). In the beginning of the execution, the range of the simulated groundwater levels were wide because of the random generation of parameters. With the operation of the algorithm, the simulated groundwater levels converged to the observed value because of the inclusion of more observations into PF. The difference between the mean

value of the simulated and observed groundwater level were small which indicated the randomness generating particles.
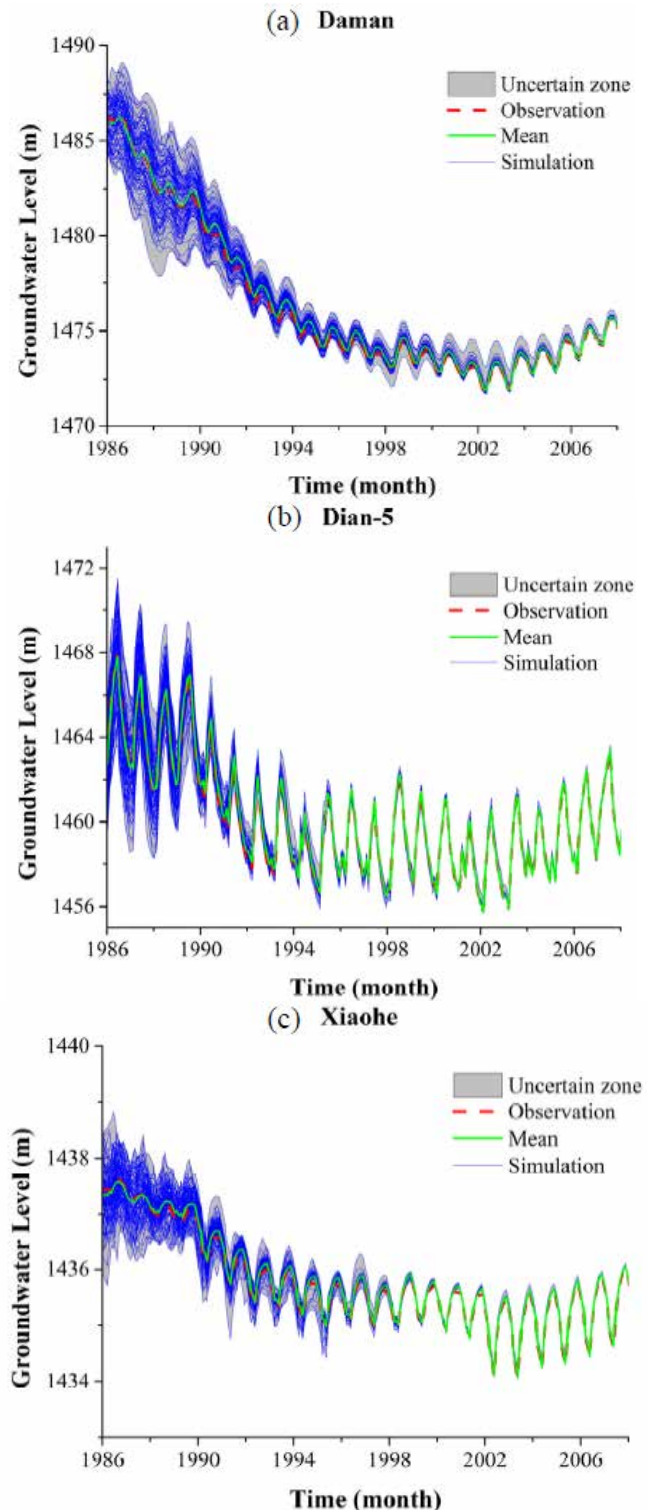

FIGURE III. THE SIMULATED GROUNDWATER LEVELS (BLUE CURVES), MEAN VALUE OF SIMULATED GROUNDWATER LEVELS (GREEN CURVES) AND OBSERVED GROUNDWATER LEVELS (RED CURVES) FOR (A) DAMAN, (B) DIAN-5 AND (C) XIAOHE BOREHOLE
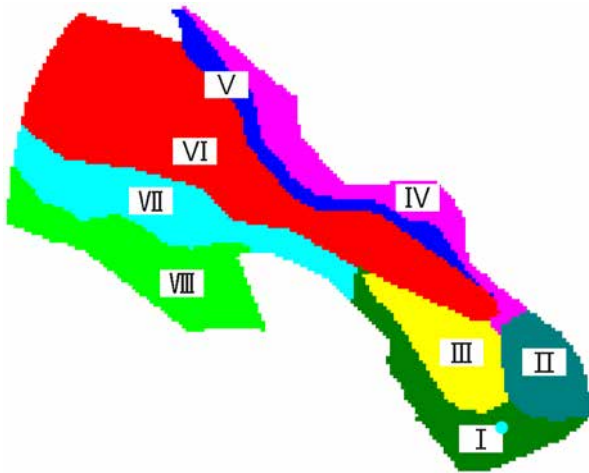
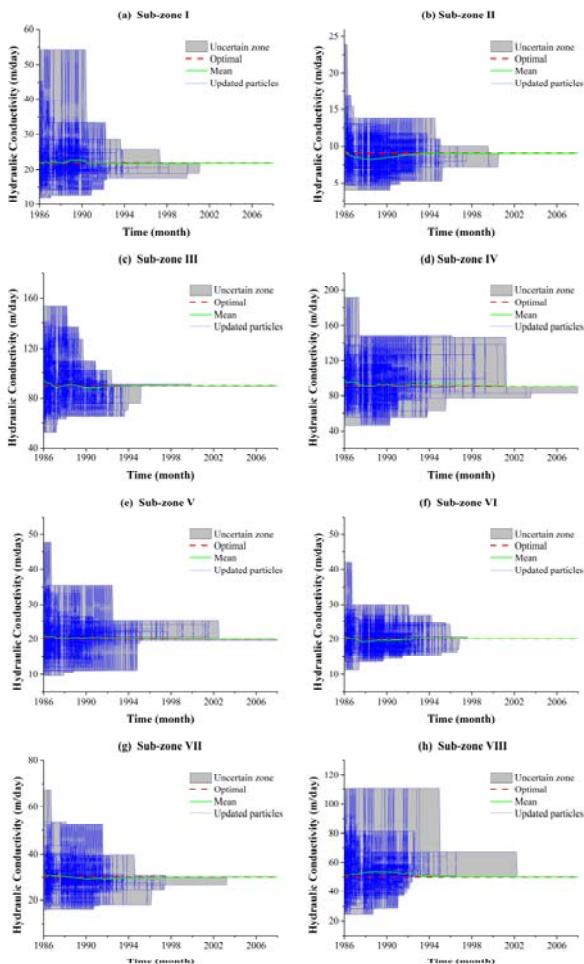FIGURE IV. SUB-ZONES OF HYDRAULIC CONDUCTIVITY.

FIGURE V. (A~H) CONVERGENCE OF MODEL PARAMETERS IN DIFFERENT SUB-ZONES (BLUE CURVES FOR DIFFERENT PARTICLES; GREEN CURVES FOR THE MEAN VALUE OF PARTICLES; RED LINE FOR THE OPTIMAL VALUE)

B. Parameters Optimization

The hydraulic conductivity for the study area was conceptualized by separating to 8 sub-zones (shown in Figure Ⅳ) based on the hydrogeological map [18]. The hydraulic conductivity in each sub-zone is identical. Figure Ⅴ demonstrated the distribution of parameters in each time step which were optimized by PF. In the beginning, one hundred particles were randomly sampled from normal distribution. The PF executed once after the observation of the current time step became available. The particles were resampled based on their importance by SIR. By iterating these procedures, the range of particles converged to the optimal value. In Figure Ⅴ, one could notice that the elapsed time of each parameter for convergence was different. This may be caused by the small difference between particle weights.

IV. CONCLUSIONS

In this study, Particle Filter was used to optimize the parameters in a groundwater model which was constructed for the middle reaches of Heihe River Basin, northwestern China. The results indicated the Particle Filter as a valuable algorithm which was applicable for various non-linear systems. However, there are still many issues should be addressed in the future work. First, although the SIR improved the problem, particle degeneracy still need further research. Second, the accuracy and computation cost of PF were determined by the number of particles. Therefore, further studies should focus on the number of particles which could balance the accuracy and computation cost. Furthermore, it is necessary to find an approach to detect the speed of convergence and convergence itself in the application of PF.

REFERENCES

[1] R. Daley, "Atmospheric data analysis," ed: Cambridge University Press (Cambridge and New York), 1991.

[2] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *International Journal of Computer Vision,* vol. 39, no. 1, pp. 57-71, 2000.

[3] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *IEE Proceedings F (Radar and Signal Processing)*, 1993, vol. 140, no. 2, pp. 107-113: IET.

[4] J. Carpenter, P. Clifford, and P. Fearnhead, "Improved particle filter for nonlinear problems," *IEE Proceedings-Radar, Sonar and Navigation,* vol. 146, no. 1, pp. 2-7, 1999.

[5] A. Smith, A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.

[6] D. Crisan, P. Del Moral, and T. Lyons, *Discrete filtering using branching and interacting particle systems*. Citeseer, 1998.

[7] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadic, "Particle methods for change detection, system identification, and control," *Proceedings of the IEEE,* vol. 92, no. 3, pp. 423-438, 2004.

[8] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing,* vol. 50, no. 2, pp. 174-188, 2002.

[9] Y. Wang, C. Yan, and J. Wang, "Landuse/Landcover data of the Heihe river basin in 1986," *Heihe Plan Science Data Center,* 2011.

[10] Y. Wang, C. Yan, and J. Wang, "Landuse/Landcover data of the Heihe river basin in 2000," *Heihe Plan Science Data Center,* 2011.

[11] J. Wang and X. Hu, "Landuse/Landcover data of Zhangye city in 2007," *Heihe Plan Science Data Center,* 2011.

[12] C. A. o. S. Cold and Arid Regions Environmental and Engineering Research Institute. (2006-2014). *Cold and Arid Regions Science Data Center at Lanzhou*. Available: http://card.westgis.ac.cn/

[13] J. Zhou, B. X. Hu, G. Cheng, G. Wang, and X. Li, "Development of a three-dimensional watershed modelling system for water cycle in the middle part of the Heihe rivershed, in the west of China," *Hydrological Processes,* vol. 25, no. 12, pp. 1964-1978, 2011.

[14] X. Wen, Y. Wu, L. J. E. Lee, J. Su, and J. Wu, "Groundwater flow modeling in the zhangye basin, northwestern china," (in English), *Environmental Geology,* Article vol. 53, no. 1, pp. 77-84, Sep 2007.

[15] L. Hu, C. Chen, J. J. Jiao, and Z. Wang, "Simulated groundwater interaction with rivers and springs in the Heihe river basin," *Hydrological Processes,* vol. 21, no. 20, pp. 2794-2806, 2007.

[16] J.-S. Zhang, E.-S. Kang, Y.-C. Lan, R.-S. Chen, and M.-X. Chen, "Studies of the transformation between surface water and groundwater and the utilization ratio of water resources in Hexi region," *Journal of Glaciology and Geocryology,* vol. 23, no. 4, pp. 375-382, 2001.

[17] C. Chen, J. Shen, G. Zhang, R. Zhao, J. Liu, and Q. Zhou, "A Groundwater Management Tool for Solving the Pumping Yields Minimization Problem: A Case Study in the Heihe River Basin," in *4th International Conference on Advanced Cloud and Big Data, CBD 2016, August 13, 2016 - August 16, 2016*, Chengdu, Sichuan, China, 2016, pp. 289-295: Institute of Electrical and Electronics Engineers Inc.

[18] W. Jianhua, Z. Jun, W. Xiaomin, and F. Bin, "Hydrogeological of the Heihe River Basin," ed, 2013.