Spatiotemporal prediction of shallow water table depths in continental China

Xing Yuan,[1,2] Zhenghui Xie,[1] and Miaoling Liang[1,3]

Received 21 August 2006; revised 26 September 2007; accepted 27 December 2007; published 12 April 2008.

[1] Macro-scale spatiotemporal distribution of shallow water table depths is important for terrestrial-ecosystem and climate research, and management of water resources. In this paper, an approach is presented to predict the water table depths from climate forcing at a large scale for a region such as continental China. This is achieved by adopting transfer function-noise (TFN) models and parameter regionalization methods. The parameters of the TFN models, which use precipitation surplus (precipitation minus potential evapotranspiration) as the input and water table depth as the output, are calibrated in gauged areas by the Kalman filter method coupled with the global optimization algorithm SCE-UA (shuffled complex evolution method developed at The University of Arizona), and the calibrated parameters are then regionalized to ungauged areas for the region within the same classified zones based on Gaussian Mixture Model (GMM) clustering method. The region such as continental China is classified into several zones by the GMM clustering method according to soil and climate characteristics. Verification and cross-validation show that the proposed methods for calibration and regionalization of the TFN-models for estimation of the water table depths are effective. The spatial and temporal variations of water table depths at a macro-scale in continental China are predicted by TFN models with the calibrated parameters for gauged areas and the regionalized parameters for ungauged areas.

Citation: Yuan, X., Z. Xie, and M. Liang (2008), Spatiotemporal prediction of shallow water table depths in continental China, *Water Resour. Res.*, *44*, W04414, doi:10.1029/2006WR005453.

1. Introduction

[2] In continental China, the average values of shallow water table depths vary from 1 to 10 m according to the available records from monitoring wells, and their temporal standard deviations vary from 0.05 to 2 m. Such spatial-temporal distributions of shallow water table depths are mainly determined by different climate conditions, topographies, and land covers. On the other hand, the fluctuation of the shallow water table results in variation of soil water content, surface and subsurface runoff, and the energy and water balances between the land surface and atmosphere. Therefore estimation of the water table depth is essential to the research of terrestrial ecosystem and land-atmosphere interaction, which provides initial conditions for land surface models [*Salvucci and Entekhabi*, 1995; *Liang and Xie*, 2003; *Liang et al.*, 2003; *Yang and Xie*, 2003; *Tian et al.*, 2006] and climate models. Research results can also be applied in studies on the variation of water resources and land use management.

[3] A number of studies have shown that time series models [e.g., *Box and Jenkins*, 1976; *Tankersley et al.*, 1993; *Crosbie et al.*, 2005] provide an empirical method for stochastic simulation, prediction and forecasting of the behavior of hydrological systems such as water table fluctuation and for quantifying the expected accuracy of the forecast. Moreover, several studies discussed the physical basis of time series models, such as *Parlange et al.* [1992] and *Knotters and Bierkens* [2000]. A particularly useful class of time series models for water table prediction are the transfer function-noise (TFN) models [e.g., *Tankersley et al.*, 1993; *Van Geer and Zuur*, 1997; *Bierkens et al.*, 1999; *Yi and Lee*, 2004]. In general, the primary recharge and discharge of groundwater are from precipitation and evapotranspiration, respectively. Therefore the water table fluctuation in response to various stresses such as surface retention, infiltration, evapotranspiration, percolation, and soil moisture storage can be modeled reasonably well by variation of precipitation surplus (precipitation minus potential evapotranspiration) using TFN models even though there are many uncertainties.

[4] To simulate the variation of the water table depth in data sparse areas, *Van Geer and Zuur* [1997] regionalized parameters of a TFN model by using kriging interpolation for a small area (6 km $\times$ 10 km) with 24 groundwater monitor wells. However, the method mentioned above may not be suitable for a large region such as continental China with very sparse data. Another approach is to relate the optimized model parameters to catchment characteristics through regression. However, *Huang et al.* [2003] pointed

[1]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China.
[2]Graduate University of the Chinese Academy of Sciences, Beijing, China.
[3]Now at National Meteorological Centre, China Meteorological Administration, Beijing, China.

0043-1397/08/2006WR005453$09.00

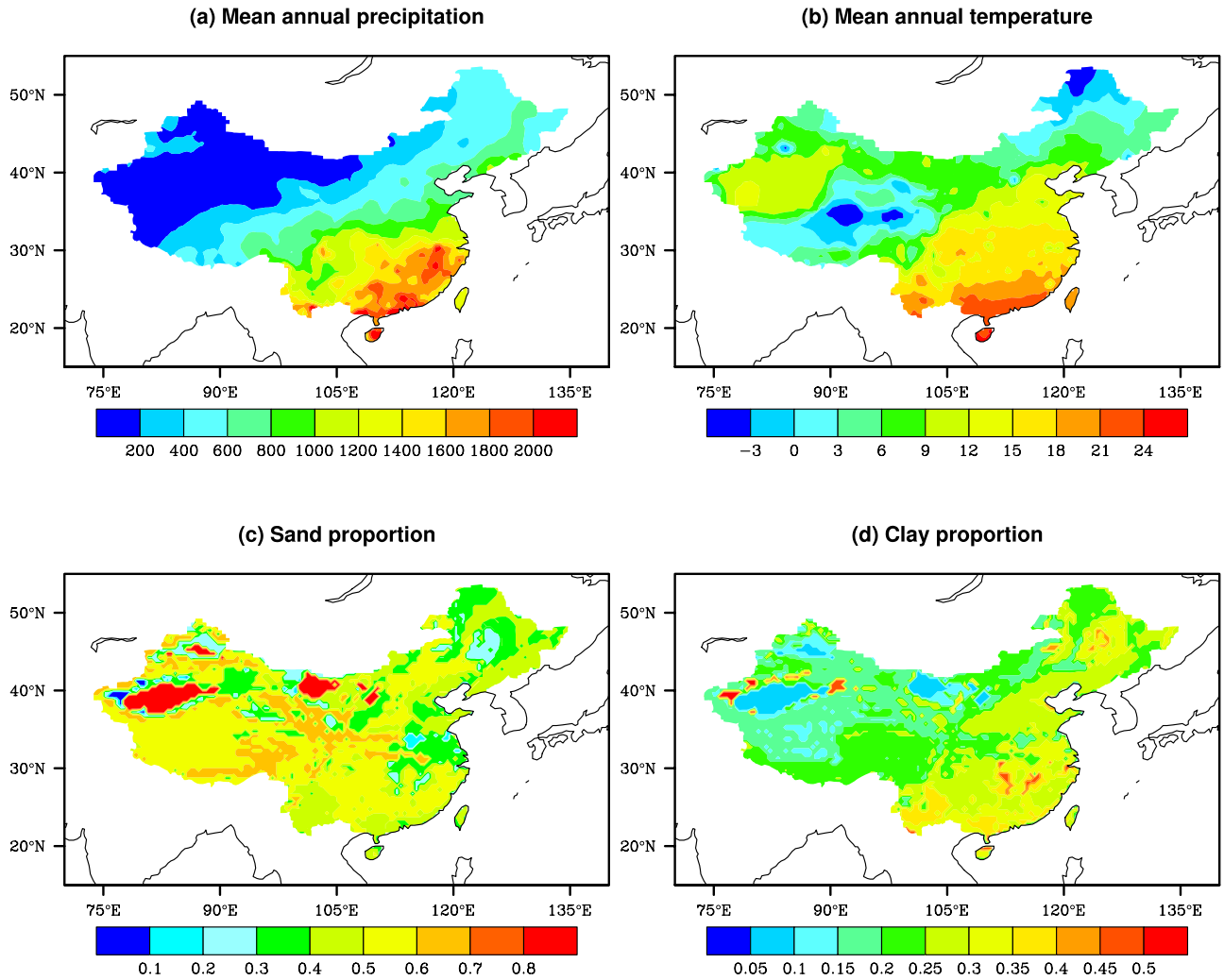

Figure 1. Climate and soil data. (a) Mean annual precipitation (mm) and (b) temperature (°C) from 1961 to 2002; (c) sand proportion and (d) clay proportion data (%) from *FAO* [1998].

out that it is difficult to describe the relationship between the model parameters and physical characteristics (e.g., soil, vegetation, and climatology information) by using multiple linear regression equations. One type of combination method is the use of time series models with regionalized parameters. For instance, *Knotters and Bierkens* [2001] developed a regionalized autoregressive exogenous variable model (RARX) for the relationship between precipitation surplus and water table depth by guessing parameters from auxiliary physical information such as digital elevation models (DEM) and catchment boundaries at ungauged locations. Furthermore, *Knotters and Bierkens* [2002] found that the Kalman filter method, which uses DEM as an external drift, outperformed the alternative methods with respect to spatiotemporal prediction of water table depths. The domain in their studies is a lowland area with elevation not more than 11 m in most parts of the region. The method described above needs to be verified for estimation of water table depths in continental China where the elevations vary from less than −100 m to more than 6000 m. Therefore it is necessary to develop a proper method that is able to avoid establishing linear regression equations and can effectively assimilate proxy data such as soil property data, meteorologic data and elevation data to predict both spatial and temporal distribution of the macro-scale water table depths in a large region such as continental China.

[5] In this paper, TFN models, together with parameter calibration and regionalization methods, are applied to predict the water table depth in continental China. In gauged areas, a method is developed that is able to calibrate the transfer function-noise (TFN) models when the time series of output is irregularly observed. Compared to the calibration approach adopted by *Bierkens et al.* [1999], the method introduced here couples the Kalman filter [*Kalman*, 1960] with SCE-UA (shuffled complex evolution method developed at The University of Arizona) method [*Duan et al.*, 1992, 1994], which combines the strengths of the simplex procedure [*Press et al.*, 1996] with the concepts of controlled random search, competitive evolution and complex shuffling [*Duan et al.*, 1992]. Many studies on parameter regionalization have been discussed, such as *Nijssen et al.* [2001], *Huang et al.* [2003], *McIntyre et al.* [2005], *Wagener and Wheater* [2006], and *Xie et al.* [2007]. We apply the Gaussian Mixture Model (GMM) clustering method [*Bilmes*, 1998] to classify continental China into eight zones according to data on soil texture and climate to measure the

Figure 2. Locations of the water table depth observations.

similarities of parameters in different regions. The clustering method does not need to establish a regression equation, and can reflect to some extent macro-scale spatial variability of physical characteristics such as precipitation and temperature, sand and clay proportion of the soil. On the basis of the classified zones, the parameters in gauged areas can be regionalized to ungauged areas through interpolation using the elevation data in each zone as auxiliary information.

[6] The outline of this paper is as follows. The data used in this study are briefly described in section 2. Section 3 discusses the methodology, including calibration and regionalization. The results of validation and prediction are presented in section 4. Summary and discussion are given in section 5.

2. Data

[7] The relationship between monthly precipitation surplus and the water table depth is considered to be a linear dynamic system for statistical analysis [*Changnon et al.*, 1988]. In this study, we use monthly precipitation surplus as input and the water table depth as output of the TFN models. A 42-year (1961–2002) time series of precipitation and temperature is obtained by interpolating the observed data at 753 meteorological stations, and the soil texture information is derived from the 5-min Food and Agriculture Organization data set [*FAO*, 1998]. Figure 1a shows the mean annual precipitation in China, which decreases from the southeast to the northwest. Figure 1b shows the mean annual temperature, which decreases as the latitude increases in most area of China except the Tibetan Plateau due to its high altitude. Figures 1c and 1d show the sand and clay proportion of the soil in China, respectively.

[8] The daily potential evapotranspiration is calculated according to *Jarvis and McNaughton* [1986]:

$$dpet = \frac{2k\left(\frac{s}{s+\gamma}\right)(uu * hn + vv * \sin(hn))}{\lambda},$$

where *dpet* is the daily potential evapotranspiration (mm/day); k is the conversion factor from solar angular units to seconds; γ and λ are the psychrometer constants; s (Pa/K) is the rate of increase of saturated vapor pressure with mean daily temperature; *uu*, *vv* and *hn* are coefficients calculated from percentage sunshine hours and latitude.

[9] The monthly data of water table depths are supplied by the Ministry of Water Resources and Institute for Geo-Environmental Monitoring in China. Figure 2 shows the locations of 571 monitoring wells. The wells in western China are more sparsely distributed than those in eastern China, and those in southern China are less densely distributed than those in northern China. The sampling frequencies and time series lengths are presented in Figure 3, which shows that the intervals of samples are not constant and their lengths vary from 6 to 288 months. Wells 1 to 229 are in the northeast of China, wells 230 to 418 are in northern China, and wells 419 to 452 are in Gansu and Ningxia provinces. These 452 wells are densely distributed in space, and 396 wells among them have more than two years (24 months) observations. The remaining wells (453 to 571)

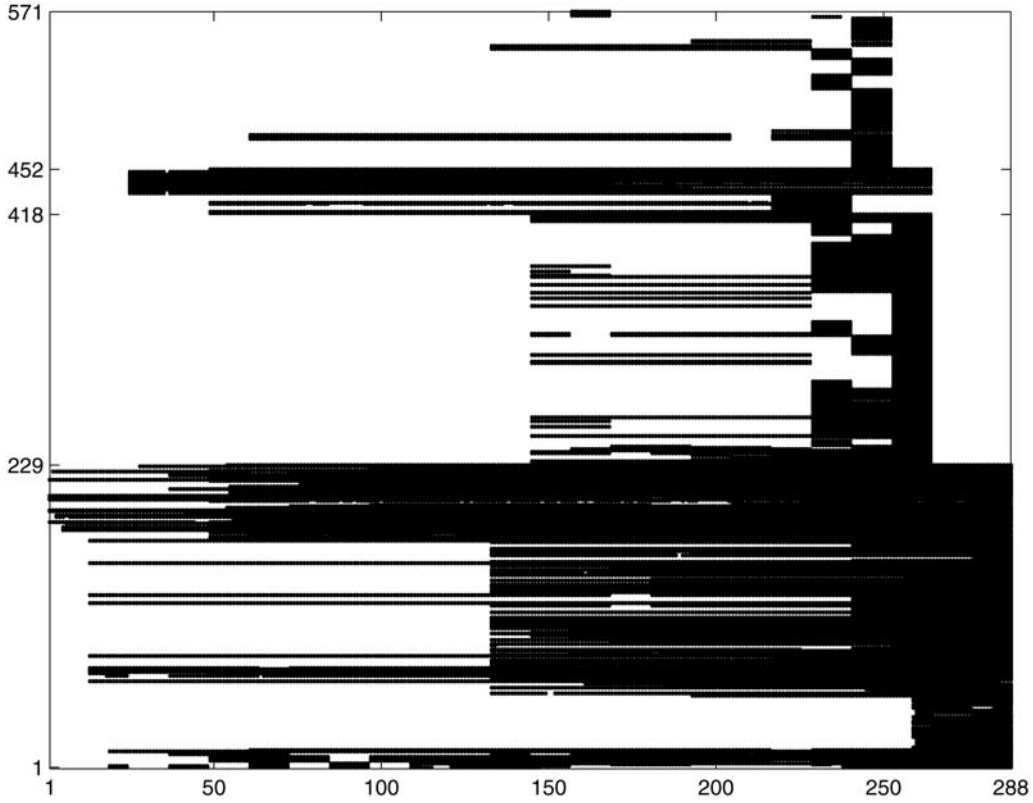

Figure 3. Sampling frequencies and time series lengths. Vertical axis: well number. Horizontal axis: month number (month 1 = January, 1979).

are sparsely distributed in western China, central China and southern China, and more than 90 percent of the 119 wells have only one-year records.

3. Methodology

[10] In this section, we describe a general method for the calibration of TFN models in gauged areas and the regionalization in ungauged areas.

3.1. Calibration of TFN Models Based on Kalman Filter and SCE-UA Method

[11] The single input-output transfer function-noise (TFN) model is given below:

$$G_t = G_t^* + n_t, \tag{1}$$

$$G_t^* = \sum_{i=1}^{r} \delta_i G_{t-i}^* + \sum_{j=0}^{s} \omega_j P_{t-j}, \tag{2}$$

$$(n_t - c) = \sum_{k=1}^{p} \phi_k (n_{t-k} - c) + a_t + \sum_{l=1}^{q} \theta_l a_{t-l}, \tag{3}$$

where G_t is the output series; G_t^* is the part of output series that can be explained by the input; P_t is the input series; n_t is the unexplained component, i.e., noise series; a_t is white noise series with constant variance σ_a^2; c is the expected value of n_t; δ_i are the autoregressive parameters of the transfer model up to order r; ω_j are the moving average parameters of the transfer model up to order s; ϕ_k are the autoregressive parameters of the noise model up to order p; and θ_l are the moving average parameters of the noise model up to order q.

[12] Equations (2)–(3) in the TFN model contain $r + s + p + q + 3$ parameters (i.e., $\boldsymbol{\alpha} = (\delta_1, \cdots, \delta_r, \omega_0, \cdots, \omega_s, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, c, \sigma_a^2)$), which can be written in vector notation as follows:

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{B}\mathbf{U}_t + \mathbf{D}\mathbf{W}_t, \tag{4}$$

where $\mathbf{X}_t = (G_t^*, \cdots, G_{t-r+1}^*, n_t, \cdots, n_{t-p+1})^{\mathrm{T}}$; $\mathbf{U}_t = (P_t, \cdots, P_{t-s}, 1)^T$; $\mathbf{W}_t = (a_t, \cdots, a_{t-q})^T$; $\mathbf{B}$ is a $(r+p) \times (s+2)$ matrix, and the first row of $\mathbf{B}$ is $(\omega_0, \cdots, \omega_s, 0)$, the $(r+1)th$ row of $\mathbf{B}$ is $(0, \cdots, 0, c - \sum_{k=1}^{p} c\phi_k)$, other rows of $\mathbf{B}$ are zero vectors; $\mathbf{D}$ is a $(r+p) \times (q+1)$ matrix, and the $(r+1)th$ row of $\mathbf{D}$ is $(1, \theta_1, \cdots, \theta_q)$, other rows of $\mathbf{D}$ are zero vectors; $\mathbf{A} = \begin{pmatrix} \mathbf{A_d} & \mathbf{0} \\ \mathbf{0} & \mathbf{A_s} \end{pmatrix}$, and

$$\mathbf{A_d} = \begin{pmatrix} \delta_1 & \delta_2 & \cdots & \delta_{r-1} & \delta_r \\ 1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$\mathbf{A_s} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

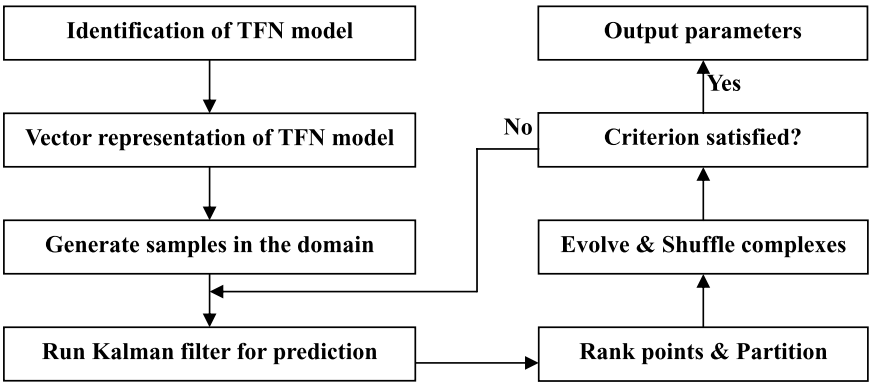

Figure 4. Flowchart of the calibration method coupling the Kalman filter and SCE-UA methods.

[13] Equation (4) is the state equation of a linear discrete stochastic system. Another important equation is the measurement equation that relates the state to observation:

$$y_t = \mathbf{C}\mathbf{X}_t + \varepsilon_t, \tag{5}$$

where y_t is a 1×1 vector of observations; $\mathbf{C}$ is a $1 \times (r + p)$ vector relating the observations to the state, and in our case the first and the $(r + 1)th$ element of $\mathbf{C}$ are 1, other elements of $\mathbf{C}$ are 0; and ε_t is random measurement error.

[14] Given the initial conditions and a parameter set $\boldsymbol{\alpha}$, the state equation (4) and measurement equation (5) consist of the recursive application of the Kalman filter equation as described by *Bierkens et al.* [1999]. Let v_i ($i = 1, \cdots, N$) be the prediction errors of the TFN model, and $\sigma^2_{v,i}$ ($i = 1, \cdots, N$) be the variances of v_i. Under the assumption that the errors obeys the normal distribution, we will get the likelihood function for a parameter set $\boldsymbol{\alpha}$ as follows:

$$L(N;\boldsymbol{\alpha}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{v,i}(\boldsymbol{\alpha})} e^{-\frac{v_i^2(\boldsymbol{\alpha})}{2\sigma^2_{v,i}(\boldsymbol{\alpha})}}.$$

Let $J(N; \boldsymbol{\alpha}) = -2\ln L(N; \boldsymbol{\alpha})$, we obtain the objective function:

$$J(N;\boldsymbol{\alpha}) = N\ln(2\pi) + \sum_{i=1}^{N} \ln(\sigma^2_{v,i}(\boldsymbol{\alpha})) + \sum_{i=1}^{N} \left(\frac{v_i^2(\boldsymbol{\alpha})}{\sigma^2_{v,i}(\boldsymbol{\alpha})} \right). \tag{6}$$

[15] In order to obtain the optimal parameter set, we should maximize the likelihood function $L(N; \boldsymbol{\alpha})$, which is equivalent to minimizing the objective function $J(N; \boldsymbol{\alpha})$. In this paper, we use the SCE-UA algorithm [*Duan et al.*, 1992] to minimize the objective function (6), which does not require derivatives of the function, and can avoid being trapped by small pits and bumps on the function surface. SCE-UA is based on a synthesis of four concepts: (1) combination of deterministic and probabilistic approaches; (2) systematic evolution of a complex of points spanning the parameter space, in the direction of global improvement; (3) competitive evolution; and (4) complex shuffling. The synthesis of the four concepts makes the SCE-UA method more effective and robust, more flexible and efficient, and less sensitive to initial values of parameters than the Simplex method. Figure 4 is the flowchart of the proposed calibration method coupling Kalman filter with SCE-UA method.

3.2. A Parameter Regionalization Scheme by Gaussian Mixture Model Clustering

[16] To estimate parameters in ungauged areas, the calibrated parameters are regionalized according to the GMM clustering method under the assumption that model parameters are related to various auxiliary factors such as climate statistics, soil texture, drainage resistance, and storage capacity. Unlike classification or regression, clustering is a kind of unsupervised training method. In fact, there is no objective function with respect to the accuracy of clustering method, so it is only a procedural method. The only principle about clustering is to minimize the divergence within each group and maximize divergence between groups. GMM transforms a clustering problem into a probabilistic density estimation problem, i.e., it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different group or cluster [*Bilmes*, 1998]. The schematic representation of the parameter regionalization method based on clustering is shown in Figure 5.

[17] Assuming that the observation data can be divided into K groups, we study the following probabilistic model with parameter vector Θ:

$$P(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{K} \pi_j P(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

$$P(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}} \mathbf{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)}, \tag{7}$$

where the parameter vector Θ consists of the mixing proportions π_j ($\pi_j \geq 0$, $\sum_{j=1}^{K} \pi_j = 1$), the mean vectors $\boldsymbol{\mu}_j$ ($j = 1, \cdots, K$), and the covariance matrices Σ_j ($j = 1, \cdots, K$); $\mathbf{x}$ is the observation vector used for clustering, and d is the dimension of the vector $\mathbf{x}$.

[18] Given the component number K and N independent, identically distributed samples $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, we obtain the following log likelihood:

$$L(\boldsymbol{\Theta}) = \ln \prod_{i=1}^{N} P(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{i=1}^{N} \ln P(\mathbf{x}_i|\boldsymbol{\Theta}), \tag{8}$$

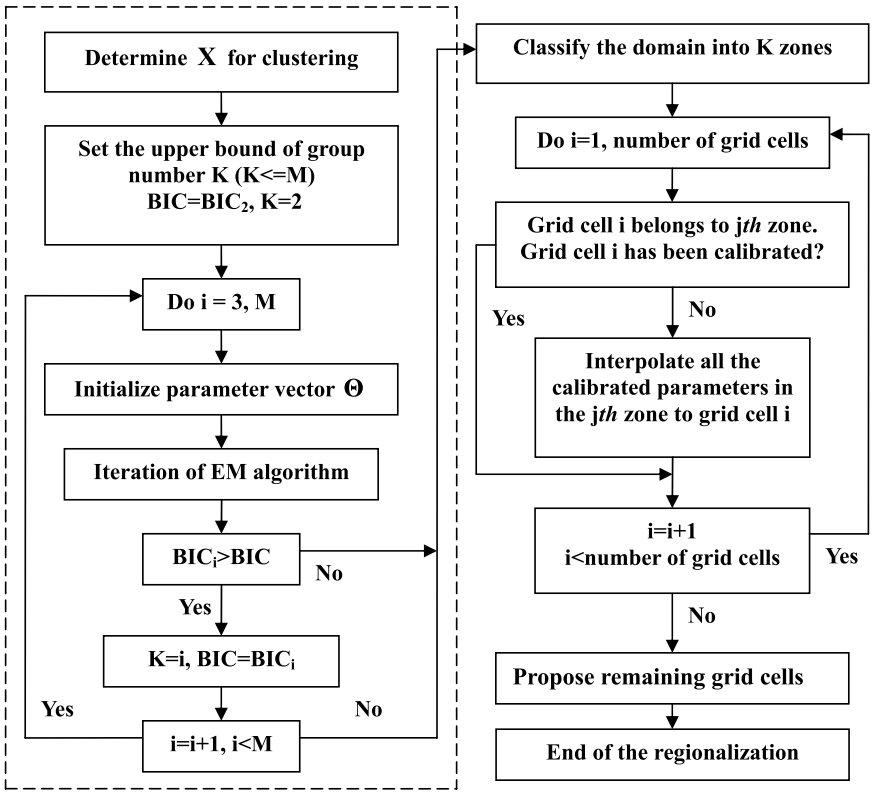

Figure 5. Schematic representation of the parameter regionalization method. The components in the dashed frame constitute the flowchart of the GMM clustering method.

which can be optimized by the following iterative algorithm called the Expectation Maximum likelihood (EM) algorithm [*Dempster et al.*, 1977]:
Posterior probabilities are calculated as expectation of log likelihood in the E-step:

$$\alpha_{ij}^{(t)} = \frac{\pi_j^{(t)} P(\mathbf{x}_i|\boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\sum_{r=1}^{K} \pi_r^{(t)} P(\mathbf{x}_i|\boldsymbol{\mu}_r^{(t)}, \boldsymbol{\Sigma}_r^{(t)})}; \tag{9}$$

Parameters are updated through maximizing the expectation of log likelihood in the M-step:

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{N} \alpha_{ij}^{(t)}}{N}, \tag{10}$$

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{N} \alpha_{ij}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{N} \alpha_{ij}^{(t)}}, \tag{11}$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{N} \alpha_{ij}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{N} \alpha_{ij}^{(t)}}. \tag{12}$$

[19] *Xu and Jordan* [1996] developed the mathematical connection between the EM algorithm and gradient-based approaches for maximum likelihood learning of finite GMM, and presented a comparative discussion of the advantages and disadvantages of EM and other algorithms. Moreover, because GMM is a model-based clustering approach, it allows the use of approximate Bayes factors to compare models [*Fraley and Raftery*, 1998]. This gives a systematic means of selecting not only the parameterization of the mixture-model, but also the number of clusters (i.e., the component number K). The Bayes factor is the posterior probability for one model against the other assuming neither is favored a priori. When EM is used to find the maximum mixture likelihood, a more reliable approximation to twice the log Bayes factor called the Bayesian Information Criterion (BIC) is applicable [*Schwarz*, 1978]:

$$2 \ln p(x|\boldsymbol{\Omega}) + const. \approx 2 l_{\boldsymbol{\Omega}}(x, \hat{\theta}) - m_{\boldsymbol{\Omega}} \ln(N) \equiv \mathrm{BIC}, \tag{13}$$

where $p(x|\boldsymbol{\Omega})$ is the integrated likelihood of the data for the model $\boldsymbol{\Omega}$, $l_{\boldsymbol{\Omega}}(x, \hat{\theta})$ is the maximized mixture log likelihood for the model with parameter $\hat{\theta}$, m_{Ω} is the number of independent parameters to be estimated in the model, and N is the number of observation data. *Fraley and Raftery* [1998] claimed that if each model is equally likely a priori, then $p(x|\boldsymbol{\Omega}$ is proportional to the posterior probability that the data conform to the model $\boldsymbol{\Omega}$. *Dasgupta and Raftery* [1998] proposed a general approach together with the BIC for selecting clusters, but they pointed out that it is important to avoid applying their procedure to a larger number of components than necessary because it often performs poorly when the number of cluster is large, so the first local maximum value of BIC will be considered as the best.

Table 1. Information for 12 Selected Wells and Their Calibration Results

	Well Number	Longitude	Latitude	Missing Ratio,%	δ	ω	ϕ	σ_a^2, cm^2	c, cm	*MAE*, cm
(a)	7	123.15°	47.32°	17.9	0.30	−3.00	0.74	188.93	201.79	20.6
(b)	2	123.75°	47.65°	51.3	0.60	−6.01	0.97	162.18	496.93	70.8
(c)	89	122.62°	45.20°	0.00	0.99	−1.01	0.96	156.17	355.90	6.86
(d)	105	123.51°	45.51°	0.00	0.31	−2.86	0.97	999.95	524.15	91.5
(e)	210	122.35°	42.54°	0.00	0.94	−2.38	0.74	129.01	471.38	9.02
(f)	194	122.77°	40.83°	4.20	0.49	−2.04	0.99	133.19	657.64	59.9
(g)	420	106.06°	38.45°	0.00	0.30	−1.00	0.65	423.14	239.80	19.8
(h)	427	106.50°	38.62°	1.10	0.30	−1.00	0.61	999.99	386.95	51.2
(i)	436	100.46°	38.97°	5.00	0.65	−1.65	0.82	49.729	15.591	8.32
(j)	439	100.48°	38.91°	0.00	0.30	−2.00	0.77	999.99	235.97	44.0
(k)	476	116.75°	24.75°	6.30	0.80	−1.64	0.82	130.57	392.67	10.1
(l)	544	104.25°	29.75°	0.00	0.53	−1.00	0.73	617.00	270.10	20.7

[20] As the entire grid cells in the study domain are classified into several zones, the calibrated parameters in gauged grid cells are transferred to ungauged grid cells according to the following procedure: (1) If grid cell i belongs to one type of zone, find all the gauged grid cells which belong to the same zones as the grid cell i, interpolate parameters of these findings to grid cell i by using elevation data as an external drift as described by *Knotters and Bierkens* [2002]:

$$\hat{a}_1(\mathbf{u}) = b_0 + b_1 z_{\mathrm{DEM}}(\mathbf{u}) + \sum_{i=1}^{n} \lambda_i(\mathbf{u}) \cdot [a_1(\mathbf{u}_i) - (b_0 + b_1 z_{\mathrm{DEM}}(\mathbf{u}_i))],$$

where $\hat{a}_1(\mathbf{u})$ is the estimated parameter at location $\mathbf{u}$, $b_0 + b_1 z_{\mathrm{DEM}}(\mathbf{u})$ is a drift function depending on the DEM, $a_1(\mathbf{u}_i)$ are parameter values calibrated on time series observed at n locations $\mathbf{u}_i$ ($i = 1, \cdots, n$), $\lambda_i(\mathbf{u})$ are interpolation weights depending on the location $\mathbf{u}$ (we used the inverse quadratic distance weighting method for interpolation in this work); (2) If there are still any grid cells left because they belong to a type that does not include any of the gauged grid cells, we have to interpolate the parameters of all the available grid cells (the calibrated or transferred) into these remaining grid cells.

4. Verification, Cross-Validation and Prediction

[21] In this section, the calibration and regionalization methods will be verified, cross-validated and applied to predict the spatial-temporal variability of the water table depths in continental China.

4.1. Verification of TFN Models

[22] We will consider only a limited class of TFN models in our case, i.e., those with $r = 1$, $s = 0$, $p = 1$, $q = 0$ according to *Bierkens et al.* [1999]. Thus equations (2)–(3) can be written as

$$G_t = c + \frac{\omega}{1 - \delta B} P_t + \frac{a_t}{1 - \phi B},$$

where G_t and P_t represent time series of the water table depth and precipitation surplus, respectively; a_t represents the white noise series with constant variance (σ_a^2); c represents the mean water table depth in case $P_t = 0$; and parameters δ and ϕ represent the memory of water table anomaly and noise, respectively; B represents the backshift operator (i.e., $B^n P_t = P_{t-n}$). In addition, 95% confidence intervals can be estimated from the noise parameters:

$$\left[\hat{G}_t - 1.96\sqrt{\sigma_a^2/(1-\phi^2)}, \hat{G}_t + 1.96\sqrt{\sigma_a^2/(1-\phi^2)}\right].$$

[23] In order to verify the calibration method in different regions, we select 12 gauged stations which have at least five years observations for detailed analysis. However, most of the gauged stations located in western China only have one-year records, so the representativeness of the verifications in this region is limited. Table 1 provides the information for these selected wells, such as the well number, location (longitude and latitude), percentage of missing data (missing ratio), calibrated parameters and mean absolute errors (*MAE*), which show that: (1) when the percentage of missing data is less than a given level (not too large), it will not affect the accuracy of estimation significantly (for example, the missing ratio of well 436 is greater than well 439, but the *MAE* of well 436 is much smaller than that of well 439); (2) large values of error are mainly due to the uncertainties of the models, which result in large values of parameter σ_a^2 (wells 105, 427, and 439). The average value of *MAE* for all the monitoring wells (571 wells) is 25.8 cm. The observed and simulated time series of water table depths for the selected wells are plotted in Figure 6, which shows that most of the observed values lie within the confidence intervals.

4.2. Parameter Regionalization and Cross-Validation

[24] In this paper, mean annual precipitation, mean annual temperature, sand proportion, and clay proportion as shown in Figure 1 are applied to constitute the sample vector $\mathbf{x}$ (i.e., $d = 4$ in equation (7)), and the series of the four variables in $\mathbf{x}$ are standardized to make equal weights in the clustering. Since there are 4196 grid cells at the $0.5° \times 0.5°$ resolution in continental China, we set $N = 4196$ in equation (8). In order to reduce the uncertainty from initial values when applying GMM clustering method, we use K-mean clustering [*Huang et al.*, 2003] to initialize $\boldsymbol{\mu}_j$ and Σ_j in equation (7), then obtain the parameter Θ through recursive application of the EM algorithm described in equations (9)–(12). The calculated BIC in equation (13) for the clusters are listed in Table 2, and the first local maximum value BIC = −26290 ($K = 8$) is selected

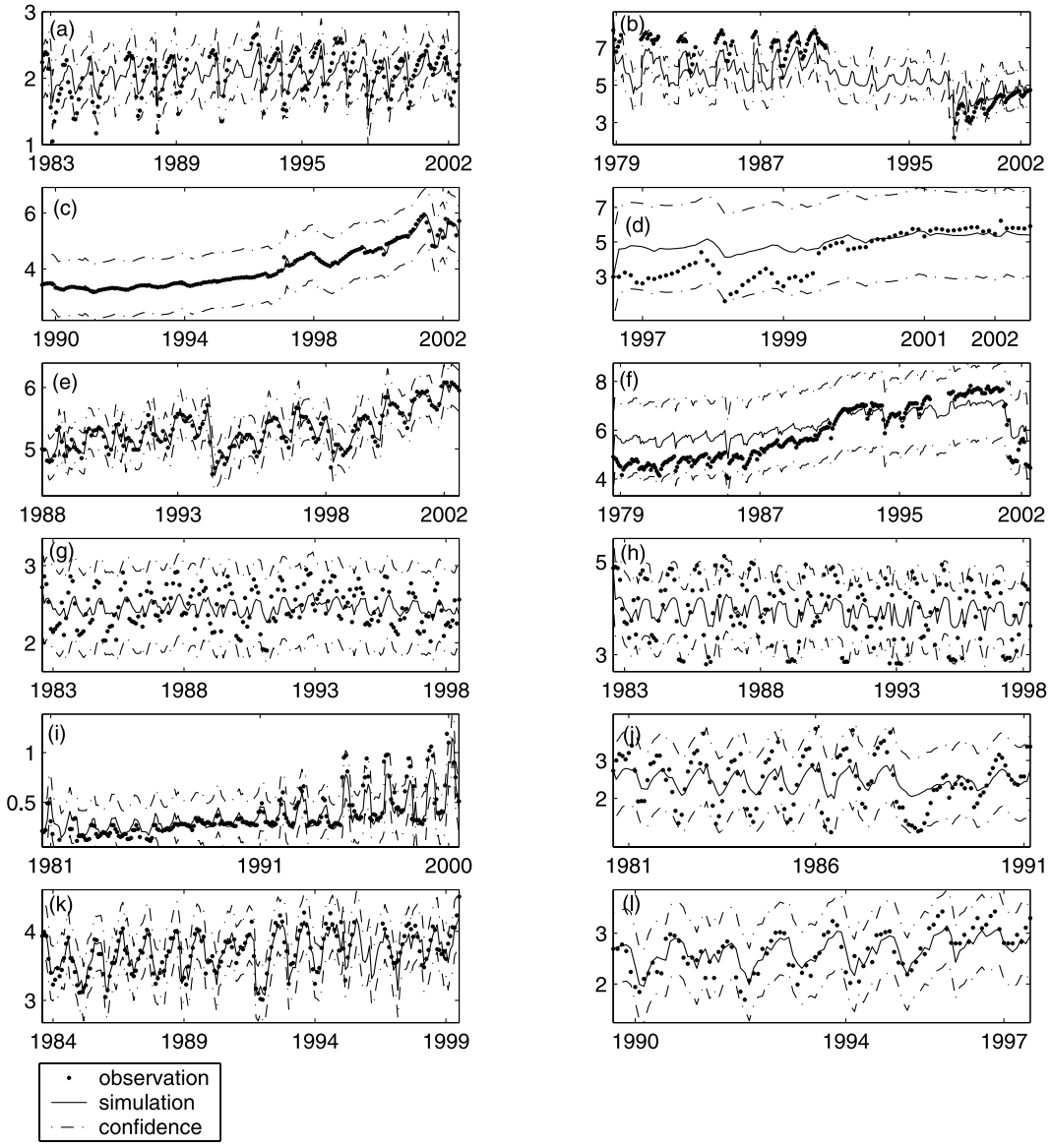

Figure 6. Verification of the TFN model in different regions. Vertical axes are water table depths (m). (a)∼(f) are located in the northeast of China, (g)∼(j) are located in northern China, (k) is located in southeastern China, and (l) is located in southwestern China.

according to *Dasgupta and Raftery* [1998]. The clustering result is shown in Figure 7.

[25] To check the effectiveness of the parameter regionalization scheme, water table depths are predicted at the 571 validation stations by means of a cross-validation procedure. All observed water table depths are used in the prediction, except those observed at the selected validation station. This procedure is repeated 571 times, each time leaving one validation station out for which prediction errors are calculated. The residuals from the cross-validation procedure are used to calculate the following validation measures [*Knotters and Bierkens*, 2001]: the systematic error

$$ME(\mathbf{u}_i) = \frac{1}{n(\mathbf{u}_i)} \sum_{j=1}^{n(\mathbf{u}_i)} e_j(\mathbf{u}_i),$$

where $n(\mathbf{u}_i)$ is the number of observed water table depths at the ith validation location, and $e_j(\mathbf{u}_i)$ is the difference

Table 2. Bayesian Information Criterion of Clusters

K	2	3	4	5	6	7	8	9	10	11	12
BIC	−36245	−35236	−31528	−30856	−28562	−28380	−26290	−26590	−25586	−21775	−22147

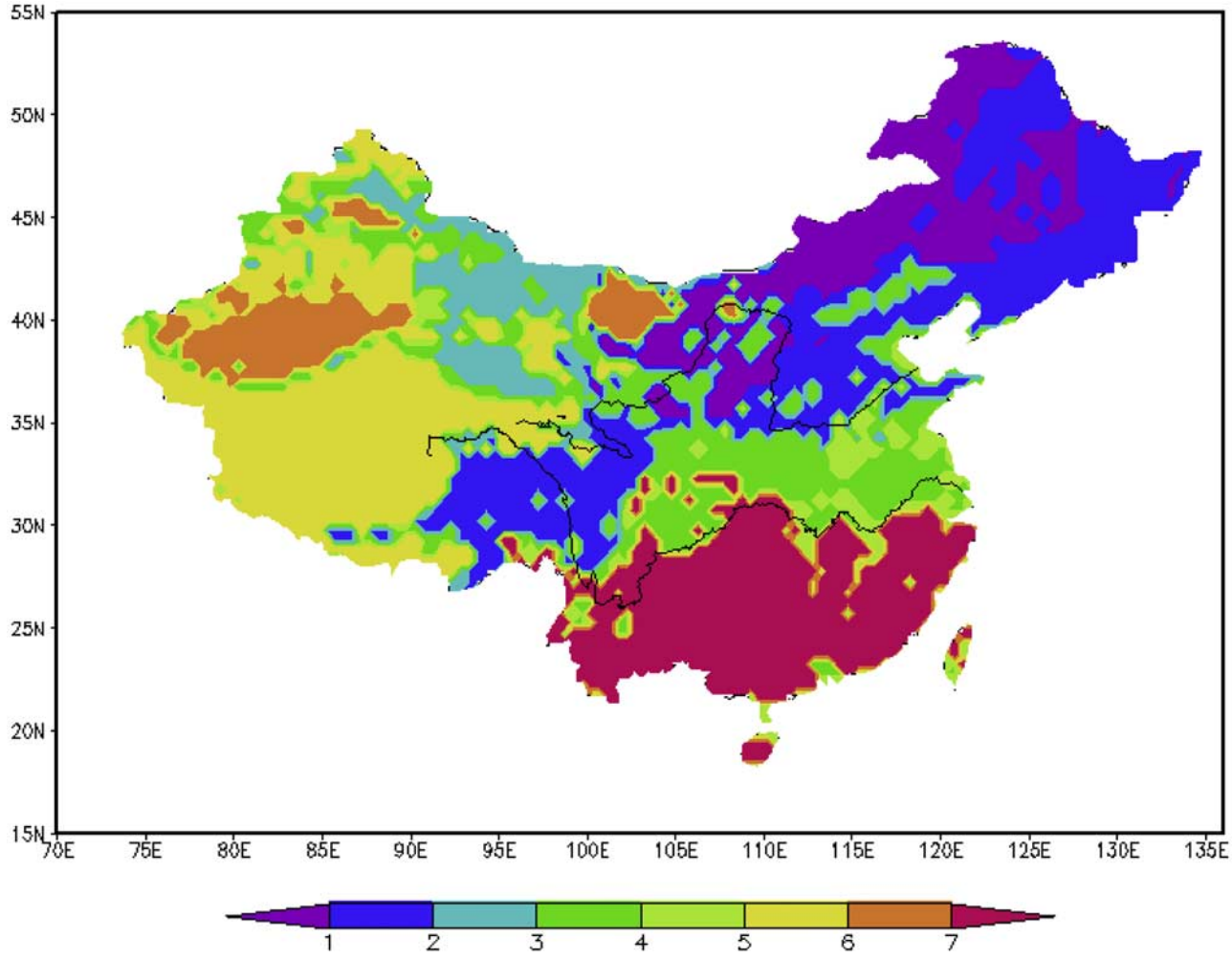

Figure 7. Eight zones resulting from the GMM clustering method based on climate statistics and soil information in China.

between observed and predicted water table depth resulting from the cross-validation; the standard deviation of error

$$SDE(\mathbf{u}_i) = \sqrt{\frac{1}{n(\mathbf{u}_i) - 1} \sum_{j=1}^{n(\mathbf{u}_i)} (e_j(\mathbf{u}_i) - ME(\mathbf{u}_i))^2},$$

which is a measure for the closeness of predicted to observed temporal fluctuation; the root mean squared error

$$RMSE(\mathbf{u}_i) = \sqrt{\frac{1}{n(\mathbf{u}_i)} \sum_{j=1}^{n(\mathbf{u}_i)} e_j(\mathbf{u}_i)^2},$$

which is a measure for the overall closeness of prediction to observation; and the mean absolute error

$$MAE(\mathbf{u}_i) = \frac{1}{n(\mathbf{u}_i)} \sum_{j=1}^{n(\mathbf{u}_i)} |e_j(\mathbf{u}_i)|,$$

which is also a measure of accuracy, but is less sensitive to large errors than *RMSE*.

[26] Figures 8a∼8d show the cross-validation results for the proposed parameter regionalization method in predicting time series of water table depths. In general, 68.7% of the biases are between -40 cm and 40 cm (Figure 8a). The average values of *RMSE* (Figure 8b) and *MAE* (Figure 8c) are 53.1 cm and 45.8 cm, respectively. Some of the *RMSE* and *MAE* values are large due to the large systematic errors (*ME*). The percentage of variance accounted for by the prediction is 63.8% on average, which is calculated from the *SDE* values (Figure 8d). The areal means of *ME*, *RMSE*, *MAE*, and *SDE* of the prediction in each clustering region are listed in Table 3 (region 7 is not listed because none of the observations are located in this region). The poor results with respect to *RMSE* and *MAE* are for region 6 (the yellow color in Figure 7) and region 8 (the magenta color in Figure 7), where observations are limited. By considering the size of our study domain (China has a land area of about 9.6 million km^2) and the variation range of DEM (about −100 m to 6000 m), we can say that the validation results described above are reasonable and acceptable.

4.3. Spatiotemporal Prediction of the Water Table Depths

[27] Assessing the parameter regionalization method through cross-validation shows that the scheme has reason-

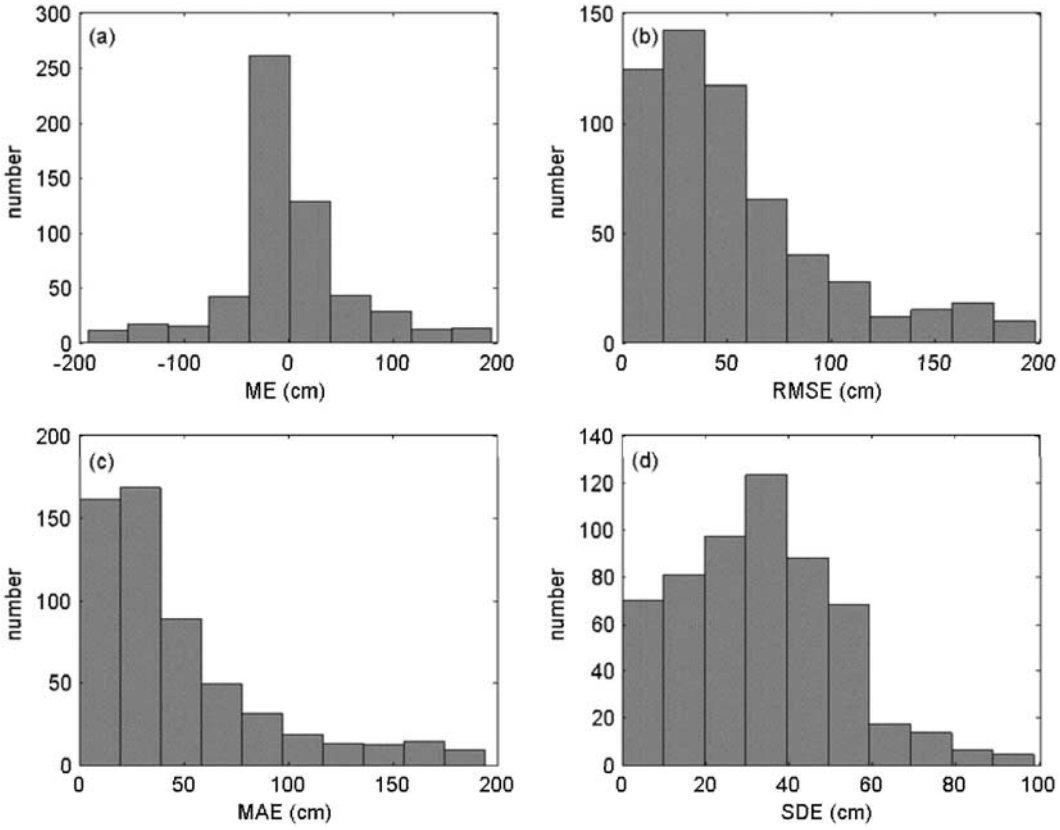

Figure 8. Cross-validation results of the parameter regionalization method. (a) Mean error (*ME*); (b) root mean squared error (*RMSE*); (c) mean absolute error (*MAE*); (d) standard deviation of error (*SDE*).

able accuracy in the validated locations. Therefore the method is adopted to predict the spatiotemporal distribution of shallow water table depths for continental China.

[28] Figure 9 shows the spatial distribution of the estimated average values of water table depths (parameter c) for continental China. Figure 10 shows the distribution of seasonal differences from the long-term average of predicted water table depths in 1998 for continental China. We can see that the anomalies of water table vary between −0.5 m to 0.5 m in most regions of China in 1998. However, a large negative anomaly (about −2 m) occurs near the middle reaches of Yangtze River due to severe precipitation events in August 1998 (Figure 10c), which resulted in a severe flooding event in the Yangtze River basin.

5. Summary and Discussion

[29] The goal of this paper is to develop a general statistical method to estimate the spatial and temporal distribution of the shallow water table depths in continental China, which is important for the land-atmosphere interaction research and water resources management. Therefore the TFN models are calibrated by the SCE-UA method coupled with Kalman filter in the gauged areas, and the calibrated parameters are regionalized to ungauged areas based on GMM clustering method. Subsequently, the shallow water table depths for the continental China are predicted by the TFN models with parameters calibrated or regionalized. In addition, the method described above also could be applied to modeling the spatiotemporal distribution of soil moisture, groundwater recharge, surface

Table 3. Mean Values of Error Measures (m) in Different Regions[a]

	1	2	3	4	5	6	8
L	164	224	7	114	33	10	19
$m_A(ME)$	5.35	−0.68	4.44	6.70	4.46	−38.27	−10.22
$m_A(RMSE)$	56.84	54.14	49.03	47.66	41.01	61.29	58.23
$m_A(MAE)$	49.55	45.92	47.17	41.62	32.98	56.26	51.85
$m_A(SDE)$	33.83	34.63	32.06	27.68	29.83	26.09	38.29

[a] $m_A(y) = \sum_{i=1}^{L} y(\mathbf{u}_i)$ is the areal mean of y, L is the number of observations in a region.

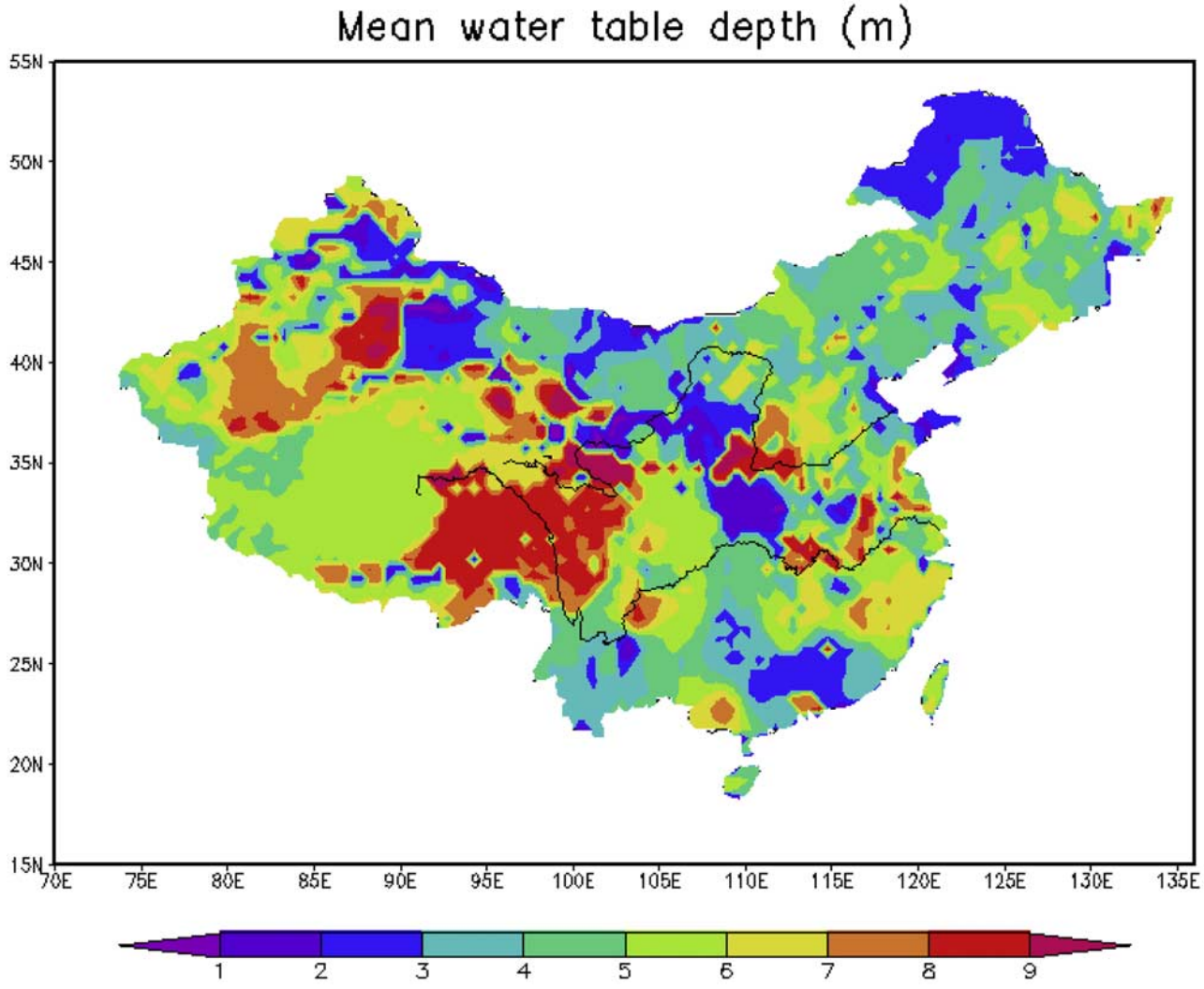

Figure 9. Spatial distribution of the estimated mean water table depth (parameter *c*) in continental China.

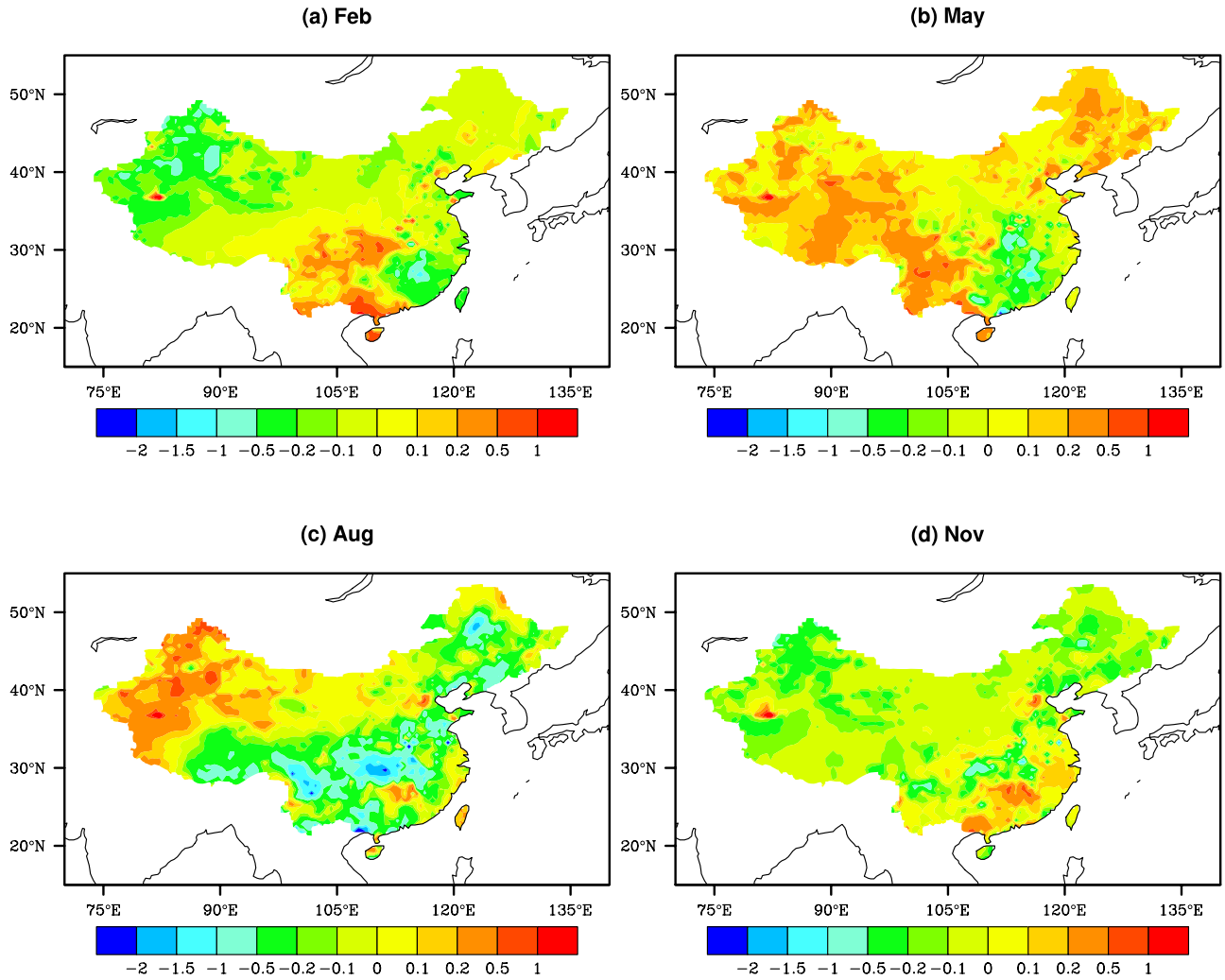

Figure 10. Anomaly relative to average values of the spatiotemporal predicted water table depth in 1998 over continental China.

temperature, etc. The following conclusions are drawn from our study:

[30] (1) The state equation of a general TFN model which contains $r + s + p + q + 3$ parameters is developed. These TFN models not only provide results with high accuracy (average value of *MAE* is 25.8 cm), but also can quantify the prediction uncertainty reasonably well through the calibration method.

[31] (2) Cross-validation shows that the parameter regionalization scheme carried out by GMM clustering method based on meteorologic data, soil texture data and elevation data is an effective way for the estimation of macro-scale water table depths. For instance, the average values of *RMSE* and *MAE* are 53.1 cm and 45.8 cm, respectively.

[32] However, there are still a number of improvements that could be made in predicting the spatiotemporal distribution of water table depths, such as considering the drainage resistance and storage capacity as the factors in the regionalization and quantifying the error induced from DEM.

[33] **Acknowledgments.** The authors would like to thank Zhiyu Liu and Huizhen Xu for providing groundwater data. This manuscript benefited significantly from the comments and suggestions of Xiaobing Feng, Marc B. Parlange, Dara Entekhabi and two anonymous reviewers. This work was supported by the National Natural Science Foundation of China under grant 90411007, the Knowledge Innovation Project of Chinese Academy of Sciences under grant KZCX2-YW-217 and KZCX2-YW-126-2, the National Basic Research Program under the grant 2005CB321704, and CAS International Partnership Creative Group "The Climate System Model Development and Application Studies."

References

Bierkens, M. F. P., M. Knotters, and F. C. Van Geer (1999), Calibration of transfer function-noise models to sparsely or irregularly observed time series, *Water Resour. Res.*, *35*(6), 1741–1750.

Bilmes, J. (1998), *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models,* TR-97-021, International Computer Science Institute, Berkeley, California.

Box, G. E. P., and G. M. Jenkins (1976), *Times Series Analysis: Forecasting and Control*, Holden Day, Merrifield.

Changnon, S. A., F. A. Huff, and C. F. Hsu (1988), Relations between precipitation and shallow groundwater in Illinois, *J. Clim.*, *12*(6), 1239–1250.

Crosbie, R. S., P. Binning, and J. D Kalma (2005), A time series approach to inferring groundwater recharge using the water table fluctuation method, *Water Resour. Res.*, *41*, W01008, doi:10.1029/2004WR003077.

Dasgupta, A., and A. E. Raftery (1998), Detecting features in spatial point processes with clutter via model-based clustering, *J. Am. Stat. Assoc.*, *93*, 294–302.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc.*, *B39*, 1–38.

Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff Models, *Water Resour. Res.*, *28*(4), 1015–1031.

Duan, Q., S. Sorooshian, and V. K. Gupta (1994), Optimal use of SCE-UA global optimization method for calibrating watershed models, *J. Hydrol.*, *158*, 265–284.

Food and Agriculture Organization (FAO) (1998), *Digital Soil Map of the World and Derived Soil Properties [CD-ROM]*, Land Water Digital Media Ser., vol. 1, Rome.

Fraley, C., and A. E. Raftery (1998), How many clusters? Which clustering methods? Answers via model-based cluster analysis, *Comput. J.*, *41*, 578–588.

Huang, M., X. Liang, and Y. Liang (2003), A transferability study of model parameters for the variable infiltration capacity land surface scheme, *J. Geophys. Res.*, *108*(D22), 8864, doi:10.1029/2003JD003676.

Jarvis, P. G., and K. G. McNaughton (1986), Stomatal control of transpiration: Scaling up from leaf to region, *Adv. Ecol. Res.*, *15*, 1–50.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *J. Basic Eng.*, *82*(Series D), 35–45.

Knotters, M., and M. F. P. Bierkens (2000), Physical basis of time series models for water table depths, *Water Resour. Res.*, *36*(1), 181–188.

Knotters, M., and M. F. P. Bierkens (2001), Predicting water table depths in space and time using a regionalised time series model, *Geoderma*, *103*, 51–77.

Knotters, M., and M. F. P. Bierkens (2002), Accuracy of spatio-temporal RARX model predictions of water table depths, *Stochastic Environ. Res. Risk Assess.*, *16*, 112–126.

Liang, X., and Z. H. Xie (2003), Important factors in land-atmosphere interactions: Surface runoff generations and interactions between surface and groundwater, *Global Planet. Change*, *38*, 101–114.

Liang, X., Z. H. Xie, and M. Y. Huang (2003), A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model, *J. Geophys. Res.*, *108*(D16), 8613, doi:10.1029/2002JD003090.

McIntyre, N., H. Lee, H. Wheater, et al. (2005), Ensemble predictions of runoff in ungauged catchments, *Water Resour. Res.*, *41*, W12434, doi:10.1029/2005WR004289.

Nijssen, B., G. M. O'Donnell, and D. P. Lettenmaier (2001), Predicting the discharge of global rivers, *J. Clim.*, *14*, 3307–3323.

Parlange, M. B., G. G. Katul, and R. H. Cuenca (1992), Physical basis for time series model of soil water content, *Water Resour. Res.*, *28*(9), 2437–2446.

Press, W. H., S. A. Teukolsky, and W. T. Vetterling (1996), *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing*, 2nd ed., pp. 1208–1210, Cambridge Univ. Press, New York.

Salvucci, G., and D. Entekhabi (1995), Ponded infiltration into soils bounded by a water table, *Water Resour. Res.*, *31*(11), 2751–2759.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464.

Tankersley, C. D., W. D. Graham, and K. Hatfield (1993), Comparison of univariate and transfer function models of groundwater fluctuations, *Water Resour. Res.*, *29*(10), 3517–3533.

Tian, X. J., Z. H. Xie, S. L. Zhang, and M. L. Liang (2006), A subsurface runoff parameterization with water storage and recharge based on the Boussinesq-storage equation for a land surface model, *Sci. China*, *49*(D), 622–631.

Van Geer, F. C., and A. F. Zuur (1997), An extension of Box-Jenkins transfer/noise models for spatial interpolation of groundwater head series, *J. Hydrol.*, *192*(5), 65–80.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty, *J. Hydrol.*, *320*, 132–154.

Xie, Z. H., F. Yuan, Q. Y. Duan, et al. (2007), Regional parameter estimation of the VIC land surface model: Methodology and application to river basins in China, *J. Hydrometeorol.*, *8*, 447–468.

Xu, L., and M. I. Jordan (1996), On convergence properties of the EM algorithm for Gaussian mixtures, *Neural Comput.*, *8*, 129–151.

Yang, H. W., and Z. H. Xie (2003), A new method to dynamically simulate groundwater table in land surface model VIC, *Prog. Nat. Sci.*, *13*(11), 819–825.

Yi, M. J., and K. K. Lee (2004), Transfer function-noise modelling of irregularly observed groundwater heads using precipitation data, *J. Hydrol.*, *288*(3), 272–287.

M. Liang, National Meteorological Centre, China Meteorological Administration, Beijing 100081, China.

Z. Xie and X. Yuan, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China. (zxie@lasg.iap.ac.cn)