

刘永和, 张万昌, 朱时良, 等. 2010. 基于广义线性模型和 NCEP 资料的降水随机发生器 [J]. 大气科学, 34 (3): 599–610. Liu Yonghe, Zhang Wanchang, Zhu Shiliang, et al. 2010. A stochastic precipitation generator based on generalized linear models and NCEP reanalysis data [J]. Chinese Journal of Atmospheric Sciences (in Chinese), 34 (3): 599–610.

基于广义线性模型和 NCEP 资料的降水随机发生器

刘永和^{1, 4, 5} 张万昌² 朱时良³ 孙成武³

1 中国科学院东亚区域气候-环境重点实验室, 全球变化东亚区域研究中心, 中国科学院大气物理研究所, 北京 100029

2 南京大学水科学研究中心, 南京 210093

3 山东省临沂市气象局, 临沂 276004

4 河南理工大学资源环境学院, 焦作 454000

5 中国科学院研究生院, 北京 100049

摘 要 天气发生器可以用来插补历史缺测气象数据或生成未来天气情境, 近年来被普遍应用于对气象变量的降尺度研究, 为陆面的水文、生态模拟提供外强迫输入。广义线性模型 (GLM) 是近年来用于建立大尺度气象变量与地面气象因子之间的一种有效方法, 基于 GLM 的天气发生器具有一定的应用前景。本文以 NCEP 再分析资料中的单格点气温、500 hPa 位势高度、位温、相对湿度、海平面气压等 5 个变量作为影响降水变化的大尺度因子建立模拟逐日降水量的广义线性模型。模型中对降水概率的描述采用 Logistic 模型模拟, 而对降水量则分别试用 Gamma 分布、指数分布、正态分布和对数正态分布来模拟, 试图比较和揭示这些基于不同理论分布的模型的能力。模型中待定参数的估计及对研究区逐日降水量的模拟采用了完全相同的实测逐日降水数据和同期 NCEP 再分析资料。参数的最大似然估计用遗传算法来实现, 对山东省临沂地区 10 个主要气象观测站降水资料的研究表明, Gamma 分布模型的拟合效果最好, 对数正态分布次之, 指数分布再次, 正态分布最差; 参数估计分月获取的拟合效果略好于不分月的。模型逐日降水模拟表明, 对降水发生概率的模拟会低估各月的多年平均值, 基于指数分布的 GLM 会低估各月总降水量期望 (为月内每日降水量期望之和) 的多年平均值, 而基于对数正态分布的 GLM 则会在降水量较大时产生高估现象。由对应的天气发生器模型生成的随机模拟降水序列表明, 基于对数正态分布的模型会高估月降水量较大时的多年平均, 而基于指数分布及 Gamma 分布的模型则模拟效果较好。总体上看, 这种基于 NCEP 再分析资料和 GLM 的天气发生器对降水变率具有很强的解释和模拟能力。

关键词 广义线性模型 天气发生器 降尺度 NCEP 再分析资料 遗传算法

文章编号 1006-9895 (2010) 03-0599-12

中图分类号 P413

文献标识码 A

A Stochastic Precipitation Generator Based on Generalized Linear Models and NCEP Reanalysis Data

LIU Yonghe^{1, 4, 5}, ZHANG Wanchang², ZHU Shiliang³, and SUN Chengwu³

1 Key Laboratory of Regional Climate-Environment Research for Temperate East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029

2 Center for Hydro-Sciences Research, Nanjing University, Nanjing 210093

收稿日期 2009-05-02, 2009-09-18 收修定稿

资助项目 国家重点基础研究发展规划项目 2006CB400502, 中国科学院“百人计划”择优支持项目 8-057493, 中国气象局气象行业专项 GYHY (QX) 2007-6-1, 国家自然科学基金资助项目 40971024

作者简介 刘永和, 男, 1976 年出生, 博士研究生, 主要从事水文气象及遥感与地理信息系统及其在全球变化中应用研究。

E-mail: sucksis@163.com

通讯作者 张万昌, E-mail: zhangwc@tea.ac.cn

3 Linyi Meteorological Bureau, Shandong Province, Linyi 276004

4 Institute of Resources and Environment, Henan Polytechnic University, Jiaozuo 454000

5 Graduate University of Chinese Academy Sciences, Beijing 100049

Abstract Weather generators were useful for generating incomplete history weather records and weather scenarios in future. Recently they were used to study downscaling climate variables and provide weather forcings for hydrological and ecological simulations. Generalized linear models (GLM) are powerful tools for relating large-scale weather variables with high resolution variables at the earth surface. The application of generators based on GLM is promising. In this paper, by using five variables derived from single grid NCEP reanalysis data, such as air temperature, 500-hPa geopotential height, potential temperature, relative humidity, sea-level pressure, as large-scale independent variables that affect precipitation variations, GLM for downscaling and simulating daily precipitation were constructed. For determination of precipitation occurrence probability, the logistic model was used, and for simulating daily precipitation amounts, the Gamma, exponential, normal, and lognormal distributions were used respectively. The observed daily precipitation series and the selected NCEP reanalysis data were used for parameter-estimation and simulation. The maximum likelihood estimate of the model parameters was performed by genetic algorithms. The results of estimation showed that the fitting effect of the gamma distribution-based models was the best, the lognormal distribution was the second, the exponential distribution was the third, and the fitting effect of the normal distribution-based models was the worst; on the other hand, the fitting effect of the models with their parameters estimated for each month separately was a little better than that without separation of months. Both the observed and simulated daily precipitation-occurrence probabilities were added for each month and the corresponding precipitation amounts were also averaged for each month in order to verify the models' ability. Results showed that the precipitation-occurrence probability was underestimated slightly by the logistic model, and precipitation-amounts expectations were underestimated by the exponential and gamma distribution-based GLM but were overestimated by the lognormal distribution-based GLM when monthly total precipitation amounts were large. The stochastic simulation of precipitation showed that the lognormal distribution-based GLM overestimated the yearly averaged monthly total precipitation and the exponential or gamma distribution-based GLM had a good simulation effect. On the whole, these stochastic models based on GLM and the NCEP reanalysis data are powerful tools for explaining and simulating the occurrence probability and amounts of precipitation.

Key words generalized linear models, weather generator, downscaling, NCEP reanalysis data, genetic algorithms

1 引言

随机天气发生器 (stochastic weather generator) 是用于随机模拟生成在统计特征上类似于实测气象数据的一类统计模型 (Woolhiser and Pegram, 1979; Wilks, 1999a; Kou et al., 2007)。天气发生器可以用来插补历史缺测气象数据或生成未来天气情境, 其优点是可以生成任意长度的时间序列。功能强大的天气发生器甚至可以同时生成一个区域内多个站点或区域中所有格网上的气象数据, 具有时间及空间降尺度模型的基本功能。由于天气发生器生成的数据可以用于流域水文、生态环境、作物生长方面的模拟。常用天气发生器生成的气象数据主要有降水、气温、风速、相对湿度、气压等。其中, 降水数据由于它在空间和时间上的高度不连

续性 (Wilks, 1998) 是各种模拟应用中最重要, 但也是最难模拟的变量, 因此, 有关降水的随机发生器模拟一直以来是学者们研究的一大热点。

用于模拟逐日降水的方法分为有参方法、半经验方法和无参方法三类 (Wilks, 1998)。有参方法以 Richardson (1981) 的 WGEN 为代表, 降水发生的概率由一阶或二阶马尔可夫链模型来模拟, 而降水量则由假设的某种概率分布模型来模拟, 如指数分布 (Richardson, 1981)、Gamma 分布 (Woolhiser and Roldan, 1982)、混合指数模型 (Woolhiser and Roldan, 1982)、对数正态分布等。半经验方法是一种由统计直方图表示降水发生概率及降水量分布的方法, 它以 Semenov and Barrow (1997) 在 1997 年提出的 LARS-WG 为代表。无参方法是一种根据历史数据与当前数据的相似性从历史数据中

进行取样的方法, 它以 Brandsma and Buishand (1998)、Zorita and von Storch (1999)、Buishand and Brandsma (2001)、Obled et al. (2002)、Beersma and Buishand (2003)、Wetterhall et al. (2005) 提出的方法为代表, 该方法的最大优点是它能够以简单的方法实现空间降尺度或多点随机模拟, 并保持降水的空间相关性。以上三种方法中, 有参方法具有易于通过参数来控制降水概率分布的特点, 因而它经常被扩展为具有更强功能的天气发生器, 使其成为一种实用的降尺度技术, 如 Wilks (1999b) 提出了通过给定一组具有空间相关性的随机数来驱动 WGEN, 以生成接近实测空间相关性的降水, Brissette et al. (2007) 又对这种方法进行了进一步改进; Hughes and Guttorp (1994)、Bellone et al. (2000)、Charles et al. (2004) 和 Greene et al. (2008) 则使用隐马尔可夫模型 (HMM), 使降水的随机生成与大气环流形势相联系。

广义线性模型 (GLM) 是另一种能够使地面气象要素与大尺度环流形势相联系的模型。GLM 是普通线性模型在分析非正态分布变量时的推广。与普通线性模型只适合分析正态分布的因变量不同, GLM 可适用于对一些非正态分布假设的因变量的分析。降水量本身是一种偏态分布的变量, 采用 GLM 来对其建模比采用普通线性模型能够获取更为合理的模型参数。Coe and Stern (1982) 认为 Markov 模型实质上是广义线性模型 (GLM) 的实例, 他们首次将传统的由基于 Markov 链的天气发生器用广义线性模型来表达。GLM 有着比马尔可夫链更灵活的表达能力, 更易于把要分析的变量与可能的影响因子建立某种联系。比如, Fealy and Sweeney (2007) 使用广义线性模型来实现对降水的降尺度, Buishand et al. (2004) 研究了基于 GLM 降尺度的时间尺度选择, Yang et al. (2005) 则使用 GLM 建立了一种时空降水发生模型, Chandler and Wheeler (2002)、Chandler (2005) 则用广义线性模型来分析和解释气候波动的因果关系。Yan et al. (2002) 使用广义线性模型分析了某些气候变化因子与西欧和北欧日最大风速的关系。这些研究都充分体现了 GLM 是一种研究气候变化和天气波动的强有力工具。

本文试图用 GLM 建立地面降水分布与 NCEP 再分析资料之间的关系, 尝试用部分 NCEP 资料作

为大尺度气象相关因子建立 GLM, 并探讨、验证基于这种 GLM 的随机发生器对研究区降水发生概率及降水量变化的模拟能力。

2 模拟降水的广义线性模型

设要分析的 $n \times 1$ 随机因变量向量为 $\mathbf{W} = (W_1, W_2, \dots, W_n)^T$, 所有因变量都受到 p 个自变量的影响, 这些自变量用 $n \times p$ 的矩阵 $\mathbf{U}$ 来表示 (其每个元素用 u_{ij} 表示相应于 W_i 的第 j 个自变量值), $\mathbf{W}$ 的均值向量为 $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^T$, 则自变量 $\mathbf{U}$ 与因变量 $\mathbf{W}$ 之间的关系用广义线性模型 (Chandler, 2005) 可表示为

$$g(\boldsymbol{\mu}) = \mathbf{U}\boldsymbol{\beta}, \quad (1)$$

其中, $g(\cdot)$ 为一个单调且可微的函数, 被称作 GLM 模型的连接函数, $\boldsymbol{\beta}$ 为 $p \times 1$ 的系数向量。可见, 也是通过自变量的线性组合来预测的。与传统的线性模型相比, 广义线性模型的广义性体现在 (Chandler, 2005): ① 所研究的因变量可以不是正态分布的, 为便于计算和推导, 一般要求因变量为指数分布族, 如 Poisson 分布、Gamma 分布、指数分布和二项分布等; ② 不要求因变量和自变量之间存在线性关系, 二者之间的非线性关系通过 $g(\boldsymbol{\mu})$ 的变换作用使其纳入到线性框架下; ③ 不要求方差为常量, 方差可以是均值的函数。传统的线性模型实际上是广义线性模型在因变量为正态分布、 $g(\boldsymbol{\mu})$ 采用恒等连接函数、方差固定情况下的特例。

公式 (1) 给出了 GLM 模型的最一般表示。而对于降水变量, 本文中采用类似于 Coe and Stern (1982) 提出的方法。与通常基于 Markov 链的降水模型类似, 该方法也是分两步进行的。第一步使用 Logistic 回归模拟一个站点上降水是否发生的概率, 公式为

$$g[\pi(i)] = \ln \left[\frac{\pi(i)}{1 - \pi(i)} \right] = a_0 + a_1 x_1(i) + \dots + a_p x_p(i), \quad (2)$$

其中, $g(\cdot)$ 表示连接函数, $\pi(i)$ 为第 i 个日期的降水发生概率, $x_1, x_2, \dots, x_p$ 为 p 个对降水发生概率有影响的观测自变量, $a_0, a_1, \dots, a_p$ 为 $p+1$ 个系数, i 为日期序号。

第二步是对有雨天的降水量用指数分布 (或 Gamma 分布、正态分布、对数正态分布) 进行拟合。设第 i 日的降水量均值为 $\mu(i)$, GLM 将降水

量均值 $\mu(i)$ 与影响降水量的所有自变量相联系。为保证由模拟得到的 $\mu(i)$ 为正值, 取

$$\ln[\mu(i)] = b_0 + b_1 x_1(i) + \cdots + b_p x_p(i), \quad (3)$$

其中, $b_0, b_1, \cdots, b_p$ 是 $p+1$ 个系数。

Gamma 分布的概率密度函数为

$$f_{k, \theta}(y) = y^{k-1} \frac{\exp(-y/\theta)}{\Gamma(k)\theta^k}, \quad y > 0 \text{ 且 } k, \theta > 0, \quad (4)$$

其中, y 表示所研究的变量, 在本文中指降水量 (单位: mm)。与下面列出的其他概率密度函数中的 y 均具有相同的意义。 k 为形状参数, θ 为尺度参数。其理论均值和方差分别为 $k\theta$ 和 $k\theta^2$ 。在用 Gamma 分布拟合降水量时, 可以假定形状参数 k 为一个不受自变量影响的变量, 只有 θ 受自变量的影响, 即 $\theta(i) = \mu(i)/k$ 。

指数分布是 Gamma 分布在 $k=1$ 时的特例, 其概率密度函数为

$$f_{\theta}(y) = \lambda e^{-\lambda y}, \quad x \geq 0, \quad (5)$$

其中, λ 相当于 Gamma 函数中 θ 的倒数, 它是指数分布函数中唯一的待定参数。而 $1/\lambda$ 就是指数分布变量的理论均值, 所以在本文的 GLM 中该参数的变化依赖于自变量的变化。

正态分布的概率密度函数为

$$f_{\mu, \sigma}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \quad (6)$$

μ, σ 分别是正态分布变量的均值和方差, 且这里的 μ 为公式 (3) 中的 $\mu(i)$ 的简写形式。在本文中假定 σ 与 μ 成比例, 即假定一个不受自变量影响的量 ϕ (也是待定参数), 则 $\sigma = \phi\mu$ 。

对数正态分布是指对数化后的变量符合正态分布的一种偏态分布, 其函数曲线形状类似于 Gamma 分布。其概率密度函数为

$$f_{\mu, \sigma}(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2}\right], \quad (7)$$

这里的 μ, σ 分别指对数正态分布的变量在对数化后的均值和方差。需要指出的是, 该分布由于不属于指数分布族, 无法直接推导出其最大似然法的求解方法。为方便估计参数, 本文中的 GLM 将降水量对数化后的均值 μ 与自变量建立关系。注意这里的 μ 不是直接如公式 (3) 所描述的降水量均值。在参数估计时也假定 σ 与 μ 成比例, 即 $\sigma = \phi\mu$ 。

需要指出的是, 混合指数分布在拟合降水量分

布时常常比 Gamma 分布和指数分布获得更好的拟合效果 (Woolhiser and Roldan, 1982), 但难以从理论上将其与广义线性模型建立联系, 因此本文中没有使用该分布。

3 参数估计与随机数生成

GLM 的参数可以采用最大似然法来估计。公式 (2) 中的系数 $a_0, a_1, \cdots, a_p$ 以及公式 (3) 中的系数 $b_0, b_1, \cdots, b_p$ 是待估计的参数, 此外, 当降水量被假定为不同的分布时, 概率密度函数中不受自变量影响的参数也需要估计, 如 Gamma 分布中的 k , 正态分布与对数正态分布中的 ϕ , 而指数分布中无这样的参数。

估计公式 (2) 的系数时采用的似然公式 (Woolhiser and Roldan, 1982) 为

$$J = \sum_{n=1}^N \{s(n) \ln \pi(n) + [1-s(n)] \ln [1-\pi(n)]\}, \quad (8)$$

其中, N 表示降雨量时间序列的总天数, $s(n)$ 表示实测降雨量时间序列中第 n 天为有降水或无降水的状态, 有雨时值为 1, 无雨时值为 0。 $\pi(n)$ 为时间序列中第 n 天有降水的概率, 根据 GLM 公式 (2), 它实际上是对降水有影响的各个因子的函数。

估计公式 (3) 的系数时采用的似然公式为

$$J = \sum_{n=1}^N s(n) \ln f(r(n)), \quad (9)$$

公式中的 $r(n)$ 为第 n 天的实测降水量值, $f(\cdot)$ 为概率密度函数, 具体形式如公式 (4) ~ (7)。

最大似然法在本质上就是寻找最合适的参数值以使似然公式 (8) 和 (9) 最大化的方法, 因此可采用遗传算法 (Genetic Algorithm) 来求解。遗传算法是一种通过模拟物种的遗传和变异来对问题的解进行逼近的方法。由于遗传算法具有比其他优化算法更好的避免出现局部极值的特点, 在对许多问题的求解中使用遗传算法能够获得较高精度的近似解。遗传算法的另一个优点是参数估计方法简单, 对各种数学问题求解时只需给出最大目标函数的公式并为待求的解空间建立基因编码即可。有关遗传算法的原理已有众多文献论述, 在此不再赘述。本研究中涉及到的最大似然问题求解中, 公式 (8) 和 (9) 均可作为遗传算法的基因适应度函数直接使用。基因编码采用实数编码法, 将待求参数的序列

作为基因序列, 如对于基于 Gamma 分布的 GLM, 其基因序列编码即为 $b_0, b_1, \dots, b_p, k$, 而基于指数分布的 GLM, 其基因序列编码即为 $b_0, b_1, \dots, b_p$ 。

下面给出遗传算法中基因适应度的计算步骤。对于似然公式 (8), 计算步骤为:

(1) 将遗传得到的基因序列的值作为公式 (2) 的参数;

(2) 将 $n=1, 2, \dots, N$ 天的观测雨量值及实测影响因子 (即 NCEP 变量) 的值代入公式 (2);

(3) 用公式 (2) 导出所有 $n=1, 2, \dots, N$ 时 $\pi(n)$ 值;

(4) 将所有 $\pi(n)$ 值代入公式 (8) 即可得出适应度。

同样, 对于似然公式 (9), 计算步骤为:

(1) 先要将遗传得到的基因序列的值作为公式 (4) [或公式 (5)、(6) 和 (7)] 的参数;

(2) 将 $n=1, 2, \dots, N$ 天的观测雨量值及实测影响因子的值代入 (4) 或公式 (5)、(6) 和 (7), 即可计算出所有 $f(r(n))$ 的值;

(3) 将所有 $n=1, 2, \dots, N$ 时 $f(r(n))$ 的值代入公式 (9) 即可得出适应度。

采用上述方法估计出各参数的值后, 即可将每日的大尺度气象因子代入公式 (2) 和 (3) 计算当日的降水发生概率 $\hat{\pi}$ 及降水量均值 $\hat{\mu}$ 。对于指数分布 [公式 (5)], 参数 λ 的值就是 $1/\hat{\mu}$; 对于 Gamma 分布, 由于关系 $\mu=k\theta$ 成立, 且参数 k 已由遗传算法估出, 因此参数 θ 的值为 μ/k ; 对于正态分布和对数正态分布, 由于关系 $\sigma=\phi\mu$ 成立, 且参数 ϕ 已由遗传算法估出, 进而可求出方差 σ 。至此, 所用四种分布的参数全部求出, 采用与这些分布对应的随机数生成器即可输出随机的降水时间序列。有关各种分布的随机数生成方法可参阅 Wilks (1998)、<http://zh.wikipedia.org/wiki>。

4 实例分析

4.1 数据资料选取

本文选取的逐日降水数据来自山东省临沂市所辖区域内的 10 个主要气象观测站 (包括临沂市及其所辖 9 个县的气象站点), 这些站点分布在 ($35^\circ\text{N}\sim 36^\circ\text{N}$, $118^\circ\text{E}\sim 119^\circ\text{E}$) 范围内。降水观测时间序列为 1962~2007 年 (共 46 年) 的逐日资料。由于所

在区域为典型的温带季风气候区域, 冬半年降水较少, 夏半年雨量丰富, 降水多集中在每年的 6、7、8 月, 其中 6 月的平均降水量在 100 mm 以上, 7 月的平均降水量超过 200 mm, 8 月的降水量在 160 mm 以上。多年平均降水量在 800 mm 以上。

本文采用美国国家环境预报中心 (NCEP) 和国家大气中心 (NCAR) 联合发布的再分析资料中的要素作为大尺度因子。由于建立 GLM 时因子太多会影响遗传算法参数估计的效率和精度, 本文直接选取了 NCEP 资料中最常用来进行天气变率分析或降尺度研究的 5 种大尺度变量 (地表温度、500 hPa 位势高度、位温、相对湿度、海平面气压) 作为影响研究区降水量发生概率和日降水量期望的因子 (Coulibaly, 2004; Cheng et al., 2008; Hundedcha and Bardossy, 2008; Huth et al., 2008; Kioutsioukis et al., 2008)。本研究拟在分析这 5 种指标与区域降水间的关系的基础上尝试建立用于逐日降水模拟的 GLM 并探讨、验证基于这种 GLM 的随机发生器对研究区降水发生概率及降水量变化的模拟能力。为此, 我们提取了以临沂气象站为中心的周围 25 个格点的 NCEP 变量, 并为了减少数据量, 事先对每个变量的格点数据进行了经验正交函数 (EOF) 分解, 结果发现所有变量的第一特征值占了所有特征值之后的 99% 以上。这说明这些变量在不同的格点上均具有极其相近的值, 这使采用多格点的 EOF 第一特征场与直接采用其中一个格点处的原始变量值具有相同的效果。因此, 本文直接采用在格点位置为 117.5°N 和 35.0°N 处的值作为影响本区域降水量的因子来建立广义线性模型。

4.2 参数拟合效果

4.2.1 多种分布模型的参数拟合效果比较

将临沂地区 10 个气象站点的连续 46 年 (1962~2007) 的逐日降水资料作为分析变量, 以 NCEP 再分析资料中所选的 5 个变量 (时间序列与 46 的逐日降水资料一一对应) 作为影响因子, 对模拟降水发生概率的 Logistic 模型以及假设降水量分布为 Gamma 分布、指数分布、正态分布和对数正态分布的降水量模型进行最大似然法参数估计。参数估计时分两种方式进行, 第一种方式 (即不分月份方式) 是使用所有样本数据进行参数估计, 即由大尺度气象因子来解释各月降水分布的差别, 每个站点上的每种假设分布只得到一组参数; 第二种方式

(即分月份方式)是为每个月估计一组参数,即由各月的参数变化来解释各月的降水差别,每个站点上的每种假设分布都有 12 组参数(每月对应一组参数)。

对第一种方式最终由遗传算法优化得到的最大似然值见表 1。其中,由公式(2)的参数估计过程获得的最大似然值列于表 1 第 2 列,由公式(3)~(7)的参数估计过程获得的最大似然值列于表 1 第 3~6 列。通过对比优化后的最大似然值可发现,四种降水量分布模型的拟合效果按似然值从大到小的排列顺序为:Gamma 分布—对数正态分布—指数分布—正态分布。由于正态分布的概率密度函数

与降水量的实际分布形状差异最大,其拟合效果也显示它在任何一个站点上均是拟合最差的。Gamma 分布除在莒南和临沭两站上比对数正态分布的拟合效果更差外,在其余站点的拟合效果均拟合最好。指数分布在所有站点上均比 Gamma 分布拟合得更好。从总体上来看,Gamma 分布的拟合效果略优于指数分布。

由第二种方式优化得到了基于这 4 个假设分布的 GLM 参数及最大似然值。为了方便与由不分月方式优化得到的全年似然值作比较,将分月方式得到的年内各月的最大似然值求和,得到了表 2 的结果。在表 2 中同样可以看出,正态分布的拟合效果

表 1 不分月的 GLM 参数估计得到的最大似然值

Table 1 The maximum likelihood values derived from parameter estimation of generalized linear models (GLM) (the estimation was performed without reference to different calendar months)

站点名	降水发生概率					似然值最大的降水量分布模型
	[公式(2)]	指数分布	Gamma 分布	对数正态分布	正态分布	
蒙阴	−6352.2	−10382.6	−10314.2	−10326.1	−12582.7	Gamma 分布
平邑	−6256.3	−10038.9	−9986.2	−10032.0	−12019.7	Gamma 分布
费县	−6401.0	−10774.8	−10723.8	−10781.1	−12909.4	Gamma 分布
沂水	−6280.3	−10171.0	−10123.1	−10148.4	−12248.2	Gamma 分布
沂南	−6420.3	−10554.5	−10502.5	−10571.0	−12717.7	Gamma 分布
临沂	−6488.3	−11036.3	−10952.1	−10997.2	−13397.3	Gamma 分布
莒南	−6510.7	−10903.5	−10818.2	−10816.9	−13225.4	对数正态分布
苍山	−6476.2	−10716.6	−10646.8	−10663.3	−13040.3	Gamma 分布
临沭	−6594.4	−11069.8	−10987.2	−10985.1	−13428.4	对数正态分布
郯城	−6611.5	−11041.2	−10963.1	−10964.7	−13373.1	Gamma 分布
平均	−6439.1	−10668.9	−10601.7	−10628.6	−12894.2	Gamma 分布

表 2 分月的 GLM 参数估计得到的多种分布全年累积最大似然值对比

Table 2 The annual cumulative maximum likelihood values derived from the parameter estimations of GLM based on the four different distributions (the estimation was performed for each calendar month respectively)

站点名	降水发生概率模型					似然值最大的降水量分布模型
		指数分布	对数正态分布	Gamma 分布	正态分布	
蒙阴	−6034.7	−10283.4	−10276.4	−10182.1	−12300.7	Gamma 分布
平邑	−5936.7	−9944.6	−9945.5	−9860.2	−11761.4	Gamma 分布
费县	−6073.1	−10715.4	−10714.0	−10633.8	−12730.2	Gamma 分布
沂水	−5972.7	−10084.8	−10093.3	−10006.3	−12019.7	Gamma 分布
沂南	−6097.0	−10473.2	−10490.4	−10396.0	−12444.8	Gamma 分布
临沂	−6124.0	−10956.5	−10920.8	−10841.1	−13135.3	Gamma 分布
莒南	−6166.5	−10816.5	−10767.7	−10696.5	−12963.4	Gamma 分布
苍山	−6128.5	−10635.8	−10596.6	−10533.2	−12749.7	Gamma 分布
临沭	−6264.3	−10985.1	−10930.4	−10870.3	−13153.3	Gamma 分布
郯城	−6240.8	−10960.5	−10912.9	−10848.6	−13106.3	Gamma 分布
平均	−6103.8	−10585.6	−10564.8	−10486.8	−12636.5	Gamma 分布

最差, 对数正态分布优于指数分布, Gamma 分布的拟合效果最好。对比表 1 和表 2 可以发现, 由分月方式优化得到的似然值总体上大于不分月方式优化的结果, 拟合效果更高, 这与理论上的认识是一致的。

4.2.2 GLM 模拟结果与观测资料的对比

为了验证 GLM 的模拟能力, 将 GLM 参数估计时所用的 NCEP 再分析资料再次作为因子, 全部代入已拟合好参数的 GLM 模型来计算从 1962 年至 2007 年这 46 年间每天的降水发生概率以及降水量的分布均值。由于正态分布的拟合效果太差, 本文只对降水发生概率模型及指数分布、Gamma 分布和对数正态分布这三个降水量分布模型进行验证。

图 1 显示了 2006 年 6 月至 8 月每日所选 NCEP 资料对蒙阴站当日降水发生概率及基于 Gamma 分布降水量期望的预报值及实际观测数据的对比情况 (由于图幅限制, 文中没有给出更长的时间序列)。注意, 这里采用的降水发生概率模型和降水量模型均是采用单月方式拟合得到的。可以看出, 实际降水大都发生在预报的降水发生概率处于峰值及附近的日内, 且实际降水量较高时预测的降水量期望也较高。显然预报值与观测值之间具有一定的相关性, 预报降水发生概率与观测降水发生状况之间的相关系数为 0.273, 经检验在 0.01 水平上显著相关; 在有降水的日内, 预报降水量与观测降水量之

间的相关系数是 0.444, 经检验在 0.01 水平上显著相关。

图 2 中显示了各个月份中观测的降水发生概率与模拟得到的降水概率均值的对比状况。图中的每个点表示 1 个站点上 1 个月份内的平均降水概率, 这些点代表了 10 个站点在 12 个月份的所有多年平均降水概率 (文中的图 3、图 4 与此类似)。各月的观测降水发生概率是由 46 年中在该月的有雨天数除以这 46 年中落入该月份的总天数得到; 而模拟得到的多年降水概率均值是将这 46 年中所选的每日 NCEP 再分析资料代入公式 (2), 得到模拟降水发生概率 π , 再对 π 作多年日平均得到。由图 2 可见, 不分月份和分月份方式的参数估计得到的模拟值和实测值构成的点分布在图中的 45°斜线附近, 这表明选中的 NCEP 再分析资料变量能够较好地解释降水发生概率的变化。图 2a 中有较多的点分布在斜线下方, 而图 2b 中的全部点分布在斜线下方, 这表明由 GLM 解释的均值略小于实测均值。对比两种参数估计方式可以发现, 不分月的参数估计得到的模型解释能力较差, 在图 2 中表现为点的分布呈团簇状, 观测与模拟之间的相关系数为 0.92, 而分月的参数估计得到的模拟结果较好, 在图 2 中表现为点的分布更接近一条直线, 观测模拟之间的相关系数为 0.995。

图 3a~d 分别显示了各个月份中基于指数分布

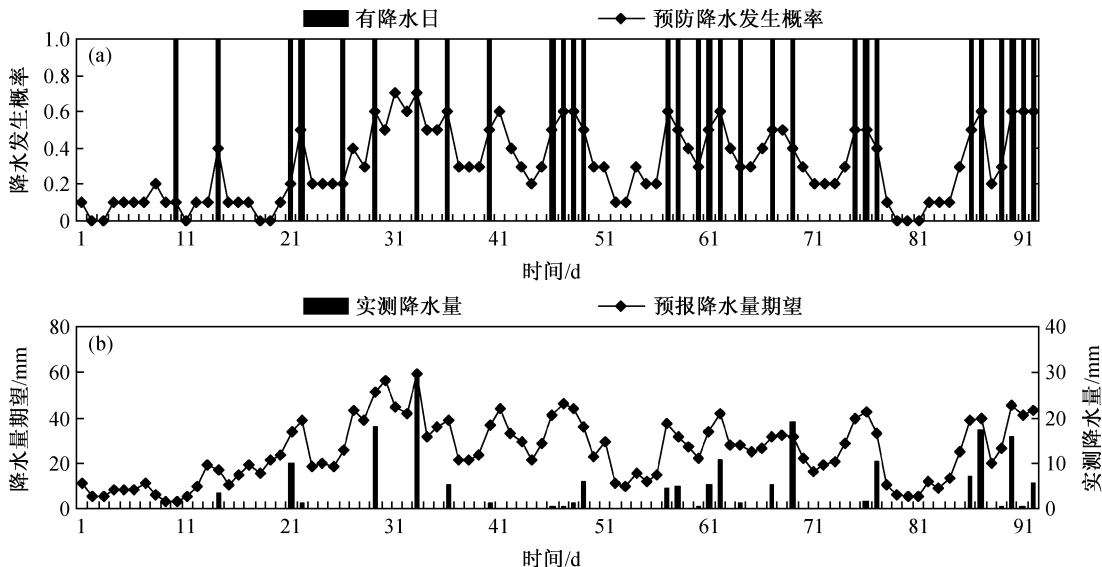


图 1 NCEP 资料通过 GLM 预报的 (a) 降水发生概率及 (b) 降水量期望与实测值对比

Fig. 1 Observed (a) precipitation occurrence and (b) mean of precipitation amounts versus those forecasted by NCEP data through GLM

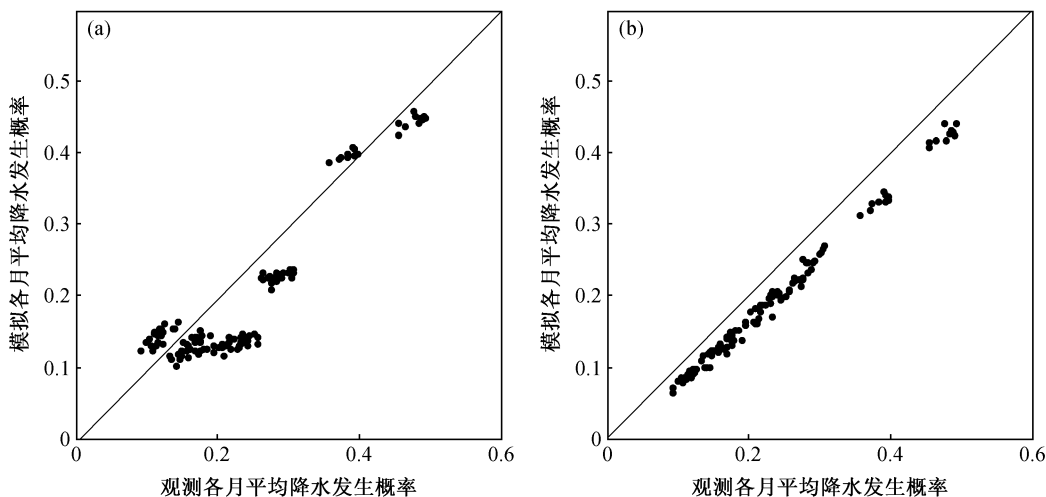


图2 模拟的 (a) 不分月和 (b) 分月的各月平均降水发生概率与观测降水发生概率对比

Fig. 2 Comparison of simulated precipitation-occurrence probability by (a) the models whose parameters were estimated without considering different calendar months and (b) the models whose parameters were estimated for each calendar month versus the observations

和 Gamma 分布的 GLM 模拟的降水量分布均值与观测均值对比。各月的观测降水量分布均值是由该月在 46 年中的降水量总和除以总降水日数得出。各月的模拟降水量分布均值是由 46 年中该月中所选每日的 NCEP 再分析资料代入公式 (3), 进而导出当日的降水量均值 μ , 再对 μ 做多年日平均即可。由这两种分布及两种参数估计方式得到的模拟值与观测值构成的点分布在 45° 斜线附近, 表明所选变量的 NCEP 再分析资料能够较好地解释降水量指数分布均值的变化。由图 3a~d 中点的散布状况可以看出, 对这两种分布分别采用两种参数估计方式得到的 GLM 模拟结果十分接近, 二者与观测值的相关系数均大于 0.98。

图 3e、f 显示了基于对数正态分布的 GLM 模拟的降水量分布均值与观测值的对比情况。基于对数正态分布的模拟降水量分布均值的计算步骤是: 先将所选 NCEP 再分析资料代入公式 (3) 计算出 μ , 再通过事先的假设关系 $\sigma = \phi\mu$ 得出 σ (注意, 这里的 μ 和 σ 是对数化降水量的均值和方差)。根据对数正态分布的均值公式 $m = \exp(\mu + \sigma^2/2)$ 计算出降水量的均值 m ; 最后, 对各月内的 m 做多年日平均计算即可。由图 3e、f 可知, 两种参数估计方式得到的模拟值与观测值构成的点也分布在 45° 斜线附近, 所选变量的 NCEP 再分析资料能够较好地解释降水量对数正态分布均值的变化。由这些图 3e、f 中点的散布状况可知, 二种参数估计方式得

到的 GLM 模拟结果较为相似, 但分月估计方式总体上略好于不分月的估计方式, 两种模拟结果与观测值的相关系数分别为 0.979 与 0.973, 回归系数分别为 1.294 和 1.211; 对大于 12 mm 的降水量均值, 分月估计方式比不分月估计方式高估得更多。

从图 3a~f 都可以发现所有点的散布状况呈现弧形, 在弧线的两头处降水量分布均值与观测均值之比较大, 在弧线的中间处降水量分布均值与观测均值之比较小; 其中在图 3a、b 中中弧线上的绝大部分点位于 45° 斜线下方, 只在弧线两头处较接近于斜线; 总体上看基于指数分布和 Gamma 分布的 GLM 低估了降雨量。而在图 3e、f 中模拟均值小于 12 mm 时, 大部分点位于斜线下方, 当模拟均值大于 12 mm 时, 几乎全部点位于斜线上方且偏离斜线较远, 这说明基于对数正态分布的 GLM 在观测的均值低于 12 mm 时低估了降雨量, 而在观测均值高于 12 mm 时, 模型则高估了降雨量, 这样就会造成基于该 GLM 的天气发生器会生成较大的降水极值。造成图 3 中的弧形点分布的原因可能有以下几个: 一是对降水量分布所假设的理论模型都存在拟合不足的问题, 无论基于哪种分布的 GLM 都难以完美地拟合降水量的真实分布; 二是 GLM 对降水发生概率的低估 (图 2) 造成了对中等规模日降水的低估, 使散点簇中部下凹; 三是指数分布、Gamma 分布及对数正态分布均不同程度的存在对极值降水的高估现象, 使散点簇上端尾部的上扬,

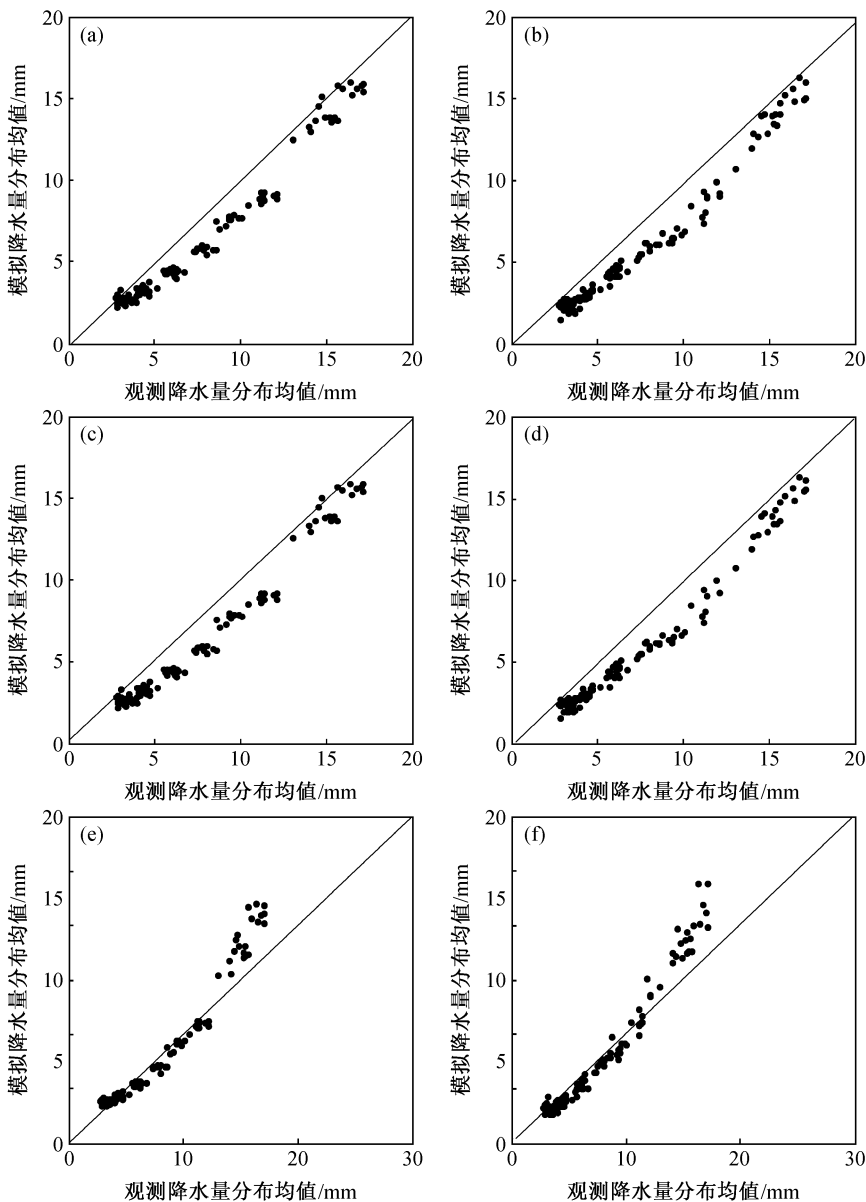


图3 基于 (a、b) 指数、(c、d) Gamma、(e、f) 对数正态分布模型的 GLM 模拟的月内日降水量均值与观测月内日降水量均值对比: (a、c、e) 不分月份; (b、d、f) 分月份

Fig. 3 Comparison of average of daily precipitation amounts in all months observed versus simulated with GLM developed with (a, b) exponential, (c, d) gamma, and (e, f) lognormal distribution functions; (a, c, e) Model parameters were estimated without considering different calendar months; (b, d, f) model parameters were estimated for each calendar month separately

其中对数正态分布的高估现象最为明显。对比图 3a、b 和图 3e、f, 可知基于对数正态分布的 GLM 的模拟结果略好于基于指数分布和 Gamma 分布的模拟结果。

4. 2. 3 基于 GLM 的天气发生器的随机模拟能力

根据 GLM 预报的降水发生概率以及基于以上三种分布的降水量分布均值, 就可以使用相应的随机天气发生器模型来生成降水时间序列。为了对比

随机生成数据对历史数据统计特征的再现能力, 对随机生成的日降水时间序列以及实际观测时间序列都进行了各站月平均降水量的统计。图 4 中显示了由三种分布所生成的年均月总降水量与观测年均月总降水量之对比。由图 4a、b 可见, 不分月的对数正态分布模型参数估计所得到的模型的随机模拟效果要好于分月估计的效果, 主要表现在前者得到的模拟效果在总体上略好, 而后者得到的结果在月降

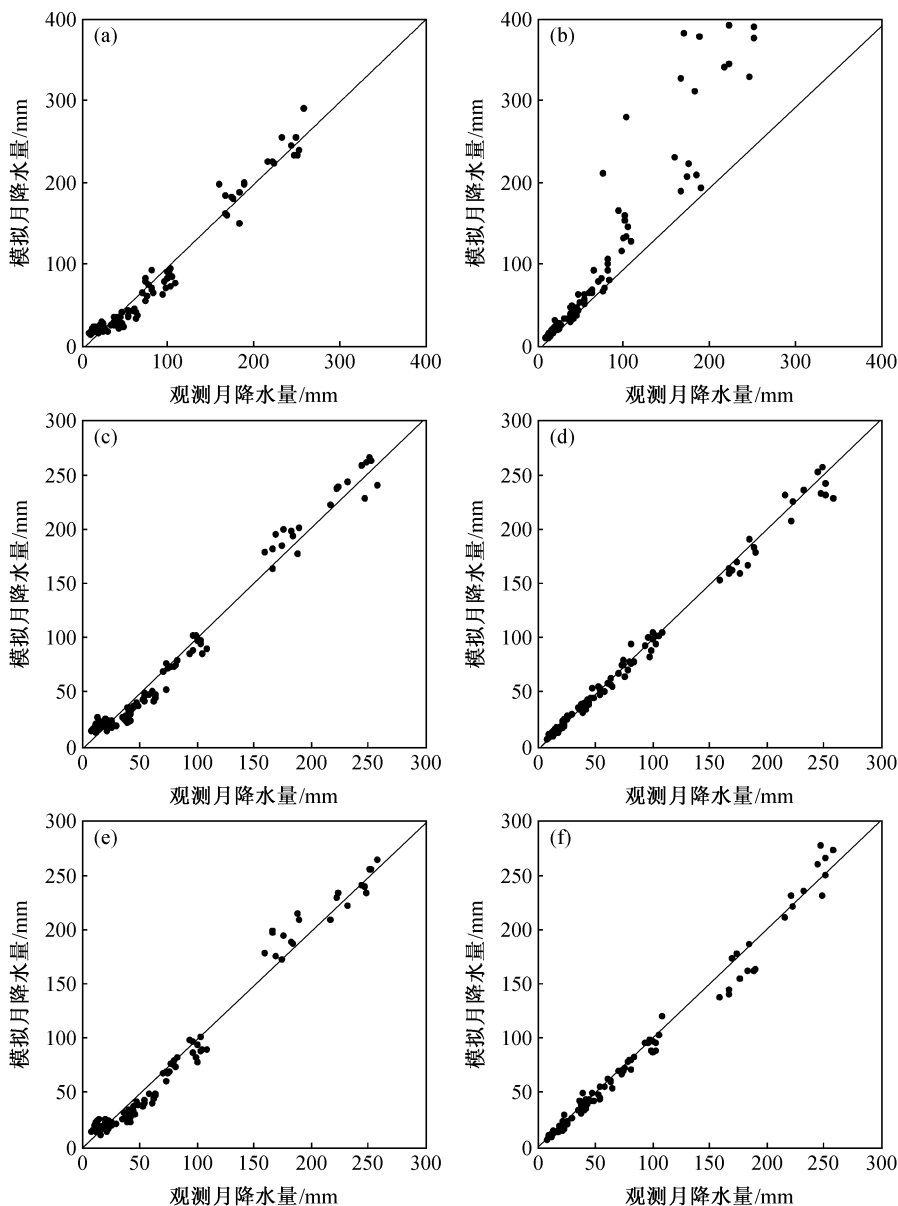


图 4 同图 3, 但为随机发生器产生的各月总降水量与观测各月总降水量的对比

Fig. 4 Similar to Fig. 3, except for stochastically simulated versus observed monthly total precipitation amounts

水量较小时与观测结果较为接近, 当月降水量较大时会产生过度高估的现象。这与前面的结论似乎不符, 但这主要是因为单月参数估计的对数正态分布模型会产生较大的极值, 从而高估了月降水量。由图 4c~f 可见, 无论采用分月还是不分月的参数估计, 基于指数及 Gamma 分布的随机模拟效果能够很好地再现实际观测雨量的多年月平均值, 而且分月参数估计的更优于不分月估计的模型。需要指出的是, 尽管基于指数分布的 GLM 低估了降水发生概率及降水量期望的多年月均值, 但其相应的随机

发生器却能很好地再现多年月平均降水量。这是因为前面对指数分布的降水量期望进行多年月内平均时容易忽略降水极值的影响从而导致低估现象, 而随机模拟的结果就不会出现这样的情况。

5 结论与讨论

广义线性模型是传统线性模型在对非正态分布的变量进行因果分析时的扩展, 特别适合于针对指数型分布变量如降水发生概率及降水量分布的分析, 且该方法还能够将任意多种自变量与因变量之

间建立起关系, 因此是一种强有力的变量分析工具。在气象学领域, 一些学者已将广义线性模型应用于全球或区域气候变化的影响分析, 或实现从 GCM 输出或如 NCEP 再分析资料的降尺度研究, 以及建立降水的随机发生器模型。

本文尝试了用 NCEP 再分析资料中的 5 种气象要素建立了模拟降水时间序列的 GLM 模型, 其中降水发生概率模型采用常用的 logistic 模型, 而对降水量分布则采用 Gamma 分布、指数分布、正态分布及对数正态分布模型。为了考虑季节变化的影响, 传统的基于 Markov 链的随机降水发生器模型总是采用每个月估计一组参数或者将年内每日的参数用傅里叶级数来表达。而本文中的 GLM 在参数估计时也采用了分月参数估计的方式以便反映季节变化。将遗传算法应用于 GLM 参数的最大似然法求解, 无论是分月估计还是不分月估计, 结果发现 Gamma 分布拟合效果最好, 对数正态分布次之, 指数分布再次, 正态分布最差。由分月和不分月的参数估计得到的最大似然值的对比发现, 分月的参数估计的拟合效果更好, 这与理论相一致, 但两种结果差异不大, 这说明 GLM 仅通过表征大尺度天气形势的 5 个单格点 NCEP 资料就可以解释季节差异。在实际模拟时, 采用不分月的参数估计会更为方便, 因为基于指数分布的 GLM 模型参数只有 12 个 (包括 6 个与降水发生概率有关的系数和 6 个与日降水量期望有关的系数), 而基于对数正态分布的 GLM 模型参数只有 13 个 (除了 12 个 GLM 系数外, 还有一个是降水量分布的参数, 如 μ), 若采用分月估计的方式, 则参数个数为以上个数的 12 倍。

将参数拟合时所采用的每日 NCEP 资料全部输入 GLM 模型就可得到相应的降水发生概率及降水量均值, 与观测降水资料的相应统计值作对比可以发现, 基于所选 NCEP 资料的 GLM 对降水的发生概率及降水量均值具有较强的解释能力, 同时表明基于指数和 Gamma 分布的 GLM 对降水量期望的月平均有所低估, 而基于对数正态分布的 GLM 产生的日降水量期望的月均值更接近于观测值, 只是在日降水量期望大于 12 mm 时的降水量期望被高估了。将由 GLM 计算得到的降水发生概率和降水量分布期望用于生成随机模拟降水, 结果表明基于指数分布 GLM 的随机发生器生成的月降水总量的多年平均比基于对数正态分布的更接近于实际观

测, 而基于对数正态分布的模型在降水总量较大时会出现高估现象, 特别是采用分月模拟参数的方式, 高估现象更为明显。

本文研究对于尝试使用 GLM 与大尺度气象变量建立天气发生器和降尺度技术提供了一定的参考。总体上看, 所选 NCEP 再分析资料能够通过基于三种理论分布模型 (指数、Gamma 和对数正态分布) 的 GLM 较好地再现沂河流域的逐日降水序列的统计特征, 且基于这三种分布的 GLM 均存在不同程度的拟合不足的问题。在实际应用中, 可根据需要及这三种分布的拟合能力选择其中一种分布建立 GLM。今后, 进一步的研究工作需采用 NCEP 再分析资料中多种其它大尺度要素来建立 GLM, 通过对比分析不同大尺度要素和带有不同阶次非线性性质的 GLM 来获得更优化的模型。

参考文献 (References)

- Beersma J J, Buishand T A. 2003. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation [J]. *Climate Research*, 25 (2): 121–133.
- Bellone E, Hughes J P, Guttorp P. 2000. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts [J]. *Climate Research*, 15 (1): 1–12.
- Brandsma T, Buishand T A. 1998. Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling [J]. *Hydrology and Earth System Sciences*, 2 (2–3): 195–209.
- Brissette F P, Khalili M, Leconte R. 2007. Efficient stochastic generation of multi-site synthetic precipitation data [J]. *Journal of Hydrology*, 345 (3–4): 121–133.
- Buishand T A, Brandsma T. 2001. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling [J]. *Water Resources Research*, 37 (11): 2761–2776.
- Buishand T A, Shabalova M V, Brandsma T. 2004. On the choice of the temporal aggregation level for statistical downscaling of precipitation [J]. *Journal of Climate*, 17 (9): 1816–1827.
- Chandler R E, Wheeler H S. 2002. Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland [J]. *Water Resources Research*, 38 (10): 1–10.
- Chandler R E. 2005. On the use of generalized linear models for interpreting climate variability [J]. *Environmetrics*, 16 (7): 699–715.
- Charles S P, Bates B C, Smith I N, et al. 2004. Statistical downscaling of daily precipitation from observed and modelled atmospheric fields [J]. *Hydrological Processes*, 18 (8): 1373–1394.
- Cheng C S, Li G, Li Q, et al. 2008. Statistical downscaling of hour-

- ly and daily climate scenarios for various meteorological variables in south-central Canada [J]. *Theoretical and Applied Climatology*, 91 (1–4): 129–147.
- Coe R, Stern R D. 1982. Fitting models to daily rainfall data [J]. *Journal of Applied Meteorology*, 21: 1024–1031.
- Coulibaly P. 2004. Downscaling daily extreme temperatures with genetic programming [J]. *Geophys. Res. Lett.*, 31, L16203, doi: 10.1029/2004GL020075.
- Fealy R, Sweeney J. 2007. Statistical downscaling of precipitation for a selection of sites in Ireland employing a generalised linear modelling approach [J]. *International Journal of Climatology*, 27 (15): 2083–2094.
- Greene A M, Robertson A W, Kirshner S. 2008. Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time-scales using a hidden Markov model [J]. *Quarterly Journal of the Royal Meteorological Society*, 134 (633): 875–887.
- Hughes J P, Guttorp P. 1994. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena [J]. *Water Resour. Res.*, 30 (5): 1535–1546.
- Hundecha Y, Bardossy A. 2008. Statistical downscaling of extremes of daily precipitation and temperature and construction of their future scenarios [J]. *International Journal of Climatology*, 28 (5): 589–610.
- Huth R, Kliegrová S, Metelka L. 2008. Non-linearity in statistical downscaling: Does it bring an improvement for daily temperature in Europe? [J]. *International Journal of Climatology*, 28 (4): 465–477.
- Kioutsoukis I, Melas D, Zanis P. 2008. Statistical downscaling of daily precipitation over Greece [J]. *International Journal of Climatology*, 28 (5): 679–691.
- Kou X J, Ge J P, Wang Y, et al. 2007. Validation of the weather generator CLIGEN with daily precipitation data from the Loess Plateau, China [J]. *Journal of Hydrology*, 347 (3–4): 347–357.
- Obléd C, Bontron G, Garçon R. 2002. Quantitative precipitation forecasts: A statistical adaptation of model outputs through an analogues sorting approach [J]. *Atmospheric Research*, 63 (3–4): 303–324.
- Richardson C W. 1981. Stochastic simulation of daily precipitation, temperature, and solar-radiation [J]. *Water Resour. Res.*, 17 (1): 182–190.
- Semenov M A, Barrow E M. 1997. Use of a stochastic weather generator in the development of climate change scenarios [J]. *Climate Change*, 35 (4): 397–414.
- Wetterhall F, Halldin S, Xu C Y. 2005. Statistical precipitation downscaling in central Sweden with the analogue method [J]. *Journal of Hydrology*, 306 (1–4): 174–190.
- Wilks D S. 1998. Multisite generalization of a daily stochastic precipitation generation model [J]. *Journal of Hydrology*, 210 (1–4): 178–191.
- Wilks D S. 1999a. Interannual variability and extreme-value characteristics of several stochastic daily precipitation models [J]. *Agricultural and Forest Meteorology*, 93 (3): 153–169.
- Wilks D S. 1999b. Multisite downscaling of daily precipitation with a stochastic weather generator [J]. *Climate Research*, 11 (2): 125–136.
- Woolhiser D A, Pegram G G S. 1979. Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models [J]. *Journal of Applied Meteorology*, 18 (1): 34–42.
- Woolhiser D A, Roldán J. 1982. Stochastic daily precipitation models. 2. A comparison of distributions of amounts [J]. *Water Resour. Res.*, 18 (5): 1461–1468.
- Yan Z W, Bate S, Chandler R E, et al. 2002. An analysis of daily maximum wind speed in northwestern Europe using generalized linear models [J]. *Journal of Climate*, 15 (15): 2073–2088.
- Yang C, Chandler R E, Isham V S, et al. 2005. Spatial-temporal rainfall simulation using generalized linear models [J]. *Water Resour. Res.*, 41, W11415, doi: 10.1029/2004WR003739.
- Zorita E, von Storch H. 1999. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods [J]. *Journal of Climate*, 12 (8): 2474–2489.