文章编号:1006 - 7639(2012) - 03 - 0459 - 06

年降水量数据的正态变换方法对比分析

陈学君[1],苏仲岳[2],李仲龙[1],韩 涛[3]

(1. 甘肃省气象信息与技术装备保障中心,甘肃 兰州 730020;
2. 兰州大学数学与统计学院,甘肃 兰州 730000;3. 西北区域气候中心,甘肃 兰州 730020)

摘 要:对于长时间尺度的降水序列变异性强、偏度大、不符合正态分布年降水数据,直接进行非均一性检验、气候统计分析等处理会产生较大误差,需首先选择合适的正态变换方法进行稳健处理。以甘肃省定西、敦煌、武都、镇原4个气象站点年降水资料为例,对于不服从正态分布的站点分别采用Box - Cox变换和Johnson变换进行正态变换。对比结果显示Box - Cox变换与Johnson变换使数据接近正态分布均是有效的。其中,Johnson变换对于异常数据的正态变换效果比Box - Cox变换更优。

关键词:甘肃省;降水量;正态变换;统计检验
中图分类号:P468.0 +24 **文献标识码**:A

引 言

长时间尺度的降水序列是水资源和水循环气候及气候变化研究的基础,但由于台站的迁移、仪器的变更及观测规范的变化等因素,导致了资料序列中的非均一性。国内外许多气候学家对气候序列的非均一性检验及其订正极为关注,作了大量工作,取得了十分重要的进展[1-3];同时,在气候统计分析和预测中,诸如回归分析、判别分析等统计方法均要求预报对象服从正态分布,在其预报区间估计和显著性检验上也均使用这一假定。由于均一性检验、气候统计理论建立在固有假设或内蕴假设的基础上,所以均要求进行分析计算的数据服从正态分布[4-5]。而对于一般降水数据而言,具有明显非正态分布性质,如何将数据变为正态化(或者准正态化)成为非常重要的科学问题。曹杰等[6]对全国160个测站的月降水资料是否符合正态分布进行了检验分析。杨观竹[7]、陶云[8]、胡文东[9]、方建刚[10]、王纪军[11]分别就不同区域降水正态分布进行了相关研究,主要用到的数据变换为对降水量序列进行开平方或开立方处理,以提高降水量序列的正态性。

数据正态变换方法是在有效保留原有信息的基础上使数据服从正态分布(或者准正态化)的方法,目前常用的数据正态变换有对数变换(Logarithmic)和Box - Cox变换,其中Box - Cox变换由于其可以针对不同的数据选择最优的幂参数,所以对于某些无法应用对数变换的数据有较好的变换效果[12];近年来,Johnson变换作为一种高级数据变换方法,在工业产品质量控制领域应用广泛[13],由于Johnson变换包含了一组复杂的变换曲线,理论上具有更强的正态变换能力。因而文中选取甘肃省4个典型站的年降水资料,对于不服从正态分布的站点采用不同的正态变换方法进行数据正态变换(主要为Box - Cox变换和Johnson变换),并通过Shapiro - Wilk和Kolmogorov - Smirnov正态检验对其变换效果进行分析,以期为年降水资料的正态化处理提供依据。

1 方 法

1.1 Box - Cox 正态变换

Box - Cox变换是Box和Cox提出的可使线性回归满足良好性质又不丢失信息的变换。属于幂变换族[12],其中包含对数变换($\lambda = 0$)、平方根变换(λ

收稿日期:2012 - 03 - 02;**改回日期**:2012 - 05 - 22
基金项目:国家公益性行业(气象)科研专项项目"中国近60年地面关键气候要素均一性检验与订正技术及站址变动影响研究"(GYHY201206013)、科技部农业科技成果转化资金项目"西北区域农业干旱监测预警技术推广应用"(2011GB24160005)、甘肃省气象局科研项目"基于多源卫星资料的甘肃省雪盖及雪深监测方法研究"(2011 - 02)共同资助
作者简介:陈学君(1973 -),男,甘肃天水人,博士,高级工程师,主要从事数据分析与处理、气象数据质量控制工作. E - mail:xuejunchen1971@163.com

=1/2)和倒数变换(λ = -1)等常用变换。具体变换公式为:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{(\lambda)-1}}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \quad (1)$$

其中,y 为原始数据,y^{λ} 为变换后数据,式中 λ 可按极大似然估计得到[9]。其中 λ 为待定参数,对不同的 λ,所做的变换就不同。虽然此变换要求 $y>0$,但当此条件不满足时,只要作一个整体平移即可。另外,由于 Box - Cox 包含对数变换、平方根变换和立平方根变换,故而文中不再单独提出上述变换。

1.2 Johnson 正态变换

1949 年,Johnson 提出了关于变量 x 的 3 个分布族很容易转换为标准正态分布。这些分布具体分别表示为 S_B(bounded)、S_L(lognormal) 和 S_U(unbounded) 的 3 种转换类型,一般可由下式表示:

$$z = \gamma + \delta f(x) - \varepsilon\lambda \quad (2)$$

其中:z 为标准正态分布变量;x 为非正态分布变量;参数 γ 和 δ 控制 x 分布的形状;ε 为位置因子,λ 为尺度因子。根据不同的偏度和峰度,变换函数将从 Johnson 函数曲线系统中选择(表 1)。Johnson 函数曲线系统中的参数 δ、ε 和 λ 可参照 Hill[13]、Chou 等[14-16]提出的理论与算法。

表 1 Johnson 分布系统

Tab. 1 Johnson distribution system

Johnson 系统	Johnson 曲线	正态转换	参数约束	X 约束
S_B	$k_1 = \ln\left[\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right]$	$z = \gamma + \eta\ln\left[\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right]$	$\eta,\lambda>0$ $-\infty<\gamma<+\infty$ $-\infty<\varepsilon<+\infty$	$\varepsilon<x<\varepsilon+\lambda$
S_L	$k_2 = \ln(x-\varepsilon)$	$z = \gamma + \eta\ln(x-\varepsilon)$	$\eta>0$ $-\infty<\gamma<+\infty$ $-\infty<\varepsilon<+\infty$	$x>\varepsilon$
S_U	$k_3 = \operatorname{arcsinh}\left[\frac{x-\varepsilon}{\lambda}\right]$	$z = \gamma + \eta\operatorname{arcsinh}\left[\frac{x-\varepsilon}{\lambda}\right]$	$\eta,\lambda>0$ $-\infty<\gamma<+\infty$ $-\infty<\varepsilon<+\infty$	$-\infty<x<+\infty$

注:表中 $\sinh\mu=\frac{e^{\mu}-e^{-\mu}}{2}$,$\operatorname{arcsinh}\mu=\ln(\mu+(\mu^2+1)^{\frac{1}{2}})$

1.3 正态性检验

正态性检验方法的原假设一般为 H_0:数据服从正态分布;相应的备择假设为 H_1:数据不服从正态分布。在这种意义下,这类检验有时也称非正态性检验(non - normality test)。正态性检验方法方法有许多种[17],这里主要采用 W 检验(Shapiro - Wilk 检验)[18]和 Kolmogorov - Smirnov[19] 检验。文中设置信度 $\alpha=0.05$,若 2 种正态性检验方法均检验的 $P<0.05$,则否定原假设,断定总体呈非正态分布。

1.3.1 Shapiro - Wilk 检验

W 检验是 Shapiro 和 Wilk 在 1965 年提出来的,是一种基于相关的检验,属于小样本数据的正态性检验。W 检验的基本思想是在数据服从正态分布的原假设下,通过数据的顺序统计量对经标准化后的顺序统计量的期望值线性回归,得出拟合优度。拟合优度越大,表示 2 个变量的相关程度越高,数据越近似服从正态分布。W 统计量的值夹在 0 和 1 之间,数据越接近 1 则表示越服从正态分布。

1.3.2 Kolmogorov - Smirnov 检验

Kolmogorov - Smirnov 检验方法通过样本的经验分布函数(ECMF)与给定分布函数的比较,推断该样本是否来自给定分布函数的总体。容量 n 的样本的经验分布函数记为 $F_n(x)$,可由样本中小于 x 的数据所占的比例得到,给定分布函数记为 $G(x)$,构造的统计量为 $D_n = \max_x |F(x_n) - G(x)|$,2 个分布函数之差的最大值。对于假设 H_0:总体服从给定的分布 $G(x)$,及给定的 α,根据 D_n 的极限分布确定统计量关于是否接受 H_0 的数量界限。

2　数据和结果分析

2.1　资料选取

甘肃省地处黄土高原、青藏高原和蒙古高原交汇地带,地形复杂,气候差异大。这里主要从甘肃省干旱、半干旱、半湿润、湿润 4 个气候区中各选取 1 站进行数据正态变换研究与分析。选站基本原则为各站代表不同气候区且尽量选取站点降水量序列不服从正态分布的站点。具体选择站点为定西(代表半干旱气候)、敦煌(代表干旱气候)、镇原(代表半湿润气候)、武都(代表湿润气候)站。

甘肃省 4 个典型气象站点年降水资料时间为从建站起至 2010 年,资料经过质量控制(气候界限值检查、极值检查、时间一致性检查、空间一致性检查、内部一致性检查),对于个别缺测数据,由于不影响数据分布的分析,故而未加考虑。该数据来源于甘肃省气象信息与技术装备保障中心。

资料处理过程为:首先对资料进行正态性检验,对于不符合正态分布的站点资料分别采用不同的正态变换方法进行数据正态变换(Box – Cox 变换和 Johnson变换),并通过正态性检验对其变换效果进行分析。下面以定西站年降水资料为例说明其具体处理过程,其他站点的处理类似不再详细说明,只给出最后结果。

2.2　原始降水数据正态检验

定西站原始年降水资料如图 1 所示。

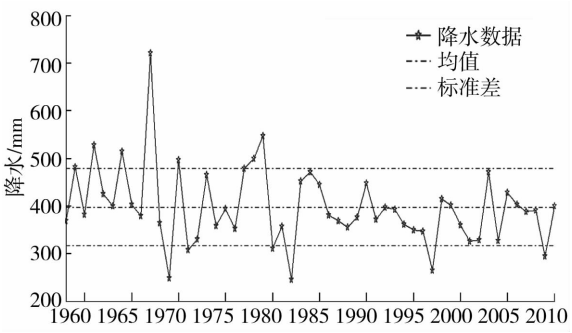

图 1　定西站 1958 ~ 2010 年降水时间序列图

Fig. 1　Annual precipitation from 1958 to 2010 at Dingxi station of Gansu Province

可以明显看出,定西站的年降水数据在 1958 ~ 2010 年间呈波动式变化。其中年平均降水量基本上 <400 mm,其基本统计信息如表 2 所示。

表 2　定西站年降水量描述统计信息表

Tab. 2　Annual precipitation statistical information about the Dingxi station

变量	平均值 /mm	中位数 /mm	最小值 /mm	最大值 /mm	标准方差 /mm	变异系数 /%	偏度	峰度
年降水	397.41	245.70	245.70	720.10	81.33	204.6%	1.1443	6.2368

统计结果显示年降水数据的变异系数达 204.6%,说明数据中极可能存在很大的样本值。由 Kolmogorov – Smirnov 正态检验可知:其检验值 $P<0.01$ ($p=4.0340e-048$),而 Shapiro – Wilk 正态检验也可得到其检验值 $P<0.01$,说明数据总体不符合正态分布(参见正态概率图 2)。且年降水量分布都表现出一定程度的正偏(偏度 >0),这在其直方图中(图 3)有更直观的表现。此外,在年降水量数据直方图右侧存在较长的拖尾。

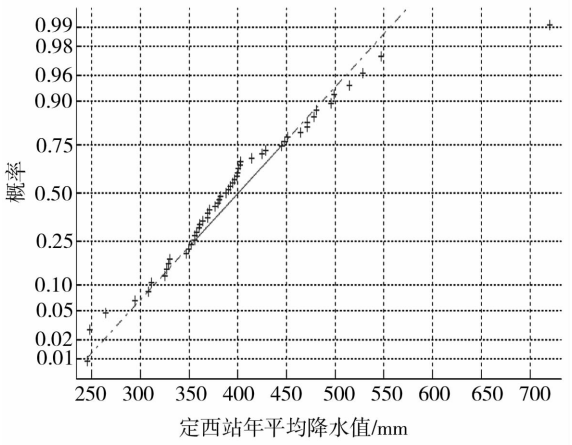

图 2　定西站年降水量正态概率图

Fig. 2　Normal probability figure of annual precipitation for Dingxi station

2.3　降水数据 Box – Cox 变换

利用 Box – Cox 数据变换方法,为寻找合适的参数 λ,先给出一系列的 λ 值,最终最优 λ 值可按极大似然估计得到。从图 4 中可以得到 λ 为 –0.13 时,其标准差最小,故在此降水数据 Box – Cox 变换中,参数 λ 取值为 –0.13。

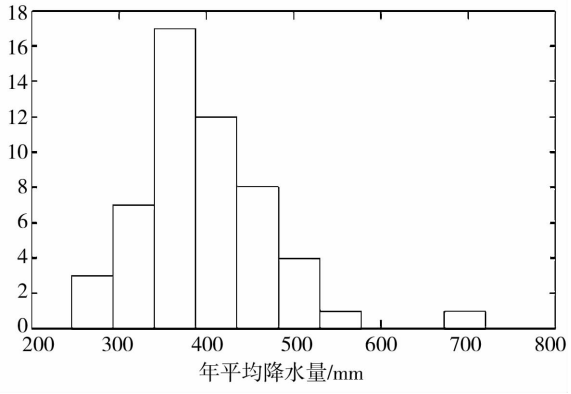

图 3 定西站年降水量直方图

Fig. 3 Frequency histogram of annual precipitation for Dingxi station

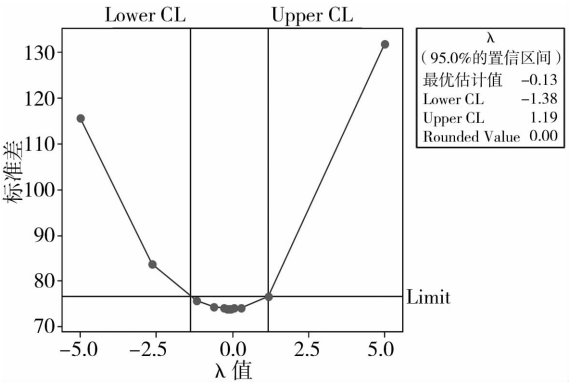

图 4 定西站年降水资料的 Box – Cox 变换 λ 值选择

Fig. 4 λ value choice about Box – Cox transformation of annual precipitation for Dingxi station

得到参数 λ 值后,利用 Box – Cox 数据变换方法即可产生变换结果。计算可知变换数据的偏度为 0.1780,峰度为 4.0770。通过 Shapiro – Wilk 正态检验可得到其检验值 $P > 0.1$,通过了 Shapiro – Wilk 检验。而 K – S 正态检验,可得到其 P 值仍 < 0.01(p = 4.0340e – 048)。综合 2 种检验方法,说明数据总体仍不符合正态分布,参见图 5,图中同时划出超出置信区间为 95% 的分布线。可见,Box – Cox 变换后的数据虽然大部分点都依附于正态分布线周围,但仍有头尾 2 端的数据出现在 95% 的置信区间以外。

2.4 降水数据 Johnson 变换

利用 Johnson 数据变换方法,确定变换类型为 S_U 型。变换公式为:

$$Y = -0.672539 + 1.27396 \cdot A\sinh((X - 355.238)/69.1397) \quad (3)$$

计算变换数据的偏度为 – 0.0446,峰度为 2.7443。通过 Shapiro – Wilk 正态检验可得到其检验值 $P > 0.1$,通过了 Shapiro – Wilk 检验。同时,用 Kolmogorov – Smirnov 正态检验可得检验值 $P > 0.30$(p = 0.3936),说明数据总体经变换后符合正态分布,参见图 6,图中同时划出超出置信区间为 95% 的分布线。从图 6 中可以看出,Johnson 变换后的数据不仅使中间(均值周围)的数据更加吻合正态分布线,而且使更多 2 端的数据落在 95% 置信区间之内。

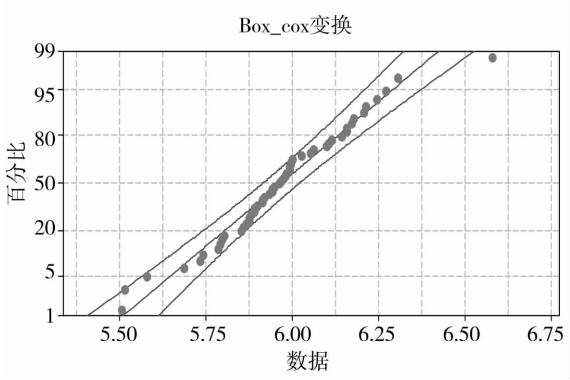

图 5 定西站年降水资料的 Box – Cox 变换正态分布概率图

Fig. 5 Normal distribution probability figure about Box – Cox transformation of annual precipitation for Dingxi station

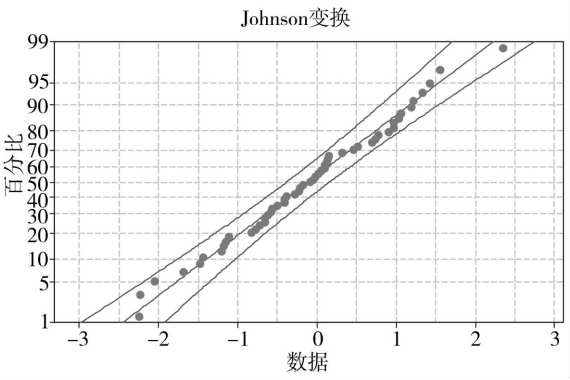

图 6 定西站降水资料的 Johnson 变换正态分布概率图

Fig. 6 Normal distribution probability figure about Johnson transformation of annual precipitation for Dingxi station

2.5 结果分析

甘肃省 4 个典型站定西、敦煌、武都、镇原的处理结果见表 3,其中武都站原始序列呈正态分布,其他 3 站均不服从正态,说明干旱对降水的分布是有

明显影响的。

从表 3 中可以看出,定西站原始数据经 Box – Cox 变换后其虽然通过了 Shapiro – Wilk,但 K – S 正态检验未通过,且偏度值由 1.1443 降为 0.1780,峰度值由 6.2368 降为 4.0770,说明数据总体的分布较原始数据接近正态;定西站原始数据经 Johnson 变换后其偏度值由 Box_cox 变换的 0.1780 降为 – 0.0446,峰度值由 Box_cox 变换的 4.0770 降为 2.7443,Shapiro – Wilk 和 K – S 正态性检验均通过,表明变换后的数据总体的分布为正态。敦煌、镇原站原始数据分别经 Box – Cox 变换和 Johnson 变换后偏度值与峰度值有明显降低,且 2 种变换方法均有效,变换后的数据均通过了 Shapiro – Wilk 检验和 K – S 检验。武都站位于湿润气候区,其原始数据已是正态分布,故而没有再进行相应数据变换。

总的来说,Box – Cox 变换和 Johnson 变换对于数据的正态性的变换均是有效的,其中 Johnson 变换更能使数据接近正态分布,尤其对于 2 端的数据(右端常为异常数据)效果明显优于 Box – Cox 变换。

表 3　不同数据变换方法的结果

Tab. 3　Results from different normal transform methods

站名	序列	偏度	峰度	W 检验(P 值)	K – S 检验(P 值)	结论
定西(半干旱区)	原始序列	1.1443	6.2368	<0.01,未通过	<0.01,未通过	非正态
	Box_cox 变换	0.1780	4.0770	>0.1,通过	<0.01,未通过	非正态
	Johnson 变换	– 0.0446	2.7443	>0.1,通过	>0.30,通过	正态
敦煌(干旱区)	原始序列	1.0849	4.3692	<0.01,未通过	<0.01,未通过	非正态
	Box_cox 变换	0.0107	3.1835	>0.1,通过	>0.15,通过	正态
	Johnson 变换	0.0471	2.6552	>0.1,通过	>0.15,通过	正态
镇原(半湿润)	原始序列	0.7761	3.2578	0.035,未通过	0.069,通过	非正态
	Box_cox 变换	0.0287	2.2849	>0.1,通过	>0.15,通过	正态
	Johnson 变换	0.1400	3.0482	>0.1,通过	>0.15,通过	正态
武都(湿润)	原始序列	0.3586	3.1477	>0.1,通过	>0.15,通过	正态
	Box_cox 变换					
	Johnson 变换					

3　结论与讨论

从甘肃省气象站点中选取代表干旱、半干旱、半湿润、湿润 4 个气候区的 4 个站点中有 3 个站不服从正态分布,通过 Box – Cox 变换和 Johnson 变换使其均符合正态分布。一般而言,对于变异性强、偏度大、不符合正态分布年降水数据,直接进行非均一性检验、气候统计分析等处理会产生较大误差,应选择合适的正态变换方法进行稳健处理。利用 Box – Cox 变换和 Johnson 变换使数据接近正态分布均是有效的,但 Johnson 变换对于异常数据的正态变换效果更优。

当然,数据变换不可能将任何非正态数据变换为正态数据,但经过数据变换后将原始数据变换为正态或准正态是有科学意义的。文中对于长时间尺度的降水序列利用正态变换方法进行数据正态变换(Box – Cox 变换和 Johnson 变换)将为降水数据的后续应用(如均一性检验、气候统计分析等气象领域)提供坚实的技术支撑,具有重要的实用价值和科学价值。

参考文献:

[1] 曹丽娟,严伟中. 地面气候资料均一性研究进展[J]. 气象科技,2011,7(2):130 – 133.

[2] 李庆祥,刘小宁,张洪政,等. 定点气候序列的均一性研究[J]. 气象科技, 2003,31(1):2 – 12.

[3] 岳文泽,徐建华,徐丽华. 基于地统计方法的气候要素空间插值研究[J]. 高原气象,2005,24(6):974 – 980.

[4] Alexanderson H. A homogeneity test applied to precipitation data [J]. International Journal of Climatology, 1986,6:661 – 675.

[5] Easterling D R, Peterson T C. A new method for detecting and adjusting for undocumented discontinuities in climatological time series [J]. International Journal of Climatology, 1995,15:369 – 377.

[6] 曹杰,陶云. 中国的降水量符合正态分布吗? [J]. 自然灾害学报,2002,11(3):115 – 120.

[7] 杨观竹. 陕西省年、月降水量的理论频数分配[J]. 高原气象,

1986,2(2):36 - 41.
[8] 陶云,段旭. 云南降水正态分布特征的初探[J]. 气象科学,2003,23(2):161 - 167.
[9] 胡文东,陈晓光,李艳春,等. 宁夏月、季、年降水量正态性分析[J]. 中国沙漠,2006,26(6):963 - 968.
[10] 方建刚,毛明策,程肖侠. 陕西降水的正态分布特征分析[J]. 西北大学学报(自然科学版),2009,39(1):131 - 136.
[11] 王纪军,任国玉,匡晓燕,等. 河南省月和年降水量正态性分析[J]. 气候与环境研究,2010,15 (4):522 - 528.
[12] BOX G,COX D. An analysis of transformations[J]. The Royal Statistical Society . Series B (Methodological),1964,26(2):211 - 252.
[13] HILL I,HILL R,HOLDER R. Fitting Johnson curves by moments [J]. Applied Statistics,1976,25(2):180 - 189.
[14] CHOU Y,POLANSKY A,MASON R. Transforming non - normal data to normality in statistical process control[J]. Quality Technology,1998,30(2):133 - 141.
[15] SLIFKER J,SHAPIRO S. The Johnson system: Selection and parameter estimation[J]. Technometrics,1980,22(2):239 - 246.
[16] MANDRACCIA S,HALVERSON G,CHOU Y. Control chart design strategies for skewed data[A]. Process ,Equipment, and Materials Control in Integrated Circuit Manufacturing II[C]. USA Austin:TX,1996. 196 - 205.
[17] 梁小筠. 正态性检验[M]. 北京:中国统计出版社,1997.
[18] 国家标准局. GB/T4882 - 1985 正态性检验[S]. 北京:中国标准出版社,1985.
[19] LILLIEFORS H. On the Kolmogorov - Smirnov test for normality with mean and variance unknown[J]. The American Statistical Association,1967,62(318):399 - 402.

Comparative Analysis of the Normal Transformation Methods about Annual Precipitation

CHEN Xuejun[1], SU Zhongyue[2], LI Zhonglong[1], HAN Tao[3]

(1. *Meteorological Information & Technique Support & Equipment Centre of Gansu Province*, *Lanzhou* 730020, *China*;2. *College of Mathematics and Statistics*, *Lanzhou University*, *Lanzhou* 730000, *China*; 3. *Regional Climate Center*, *Lanzhou* 730020, *China*)

Abstract: For the annual precipitation data, non - homogeneity test and climate statistical analysis directly will lead to considerable errors, so at first it needs to choose appropriate normal transform method. In this paper, based on precipitation data from four typical stations (Dingxi, Dounhuang, Zhenyuan and Wudu station) in Gansu province, the normal transform methods about Box - Cox transformation and Johnson transform to non - normal data were adopted. The result shows that both the Box - Cox transformation and Johnson transformation made precipitation data close to normal distribution, and the Johnson transformation was better than the Box - Cox transformation about normal transformation effect to abnormal data.

Key words: Gansu province; precipitation; normal transformation; statistics test

(上接第 458 页)

Advance in Research and Application About Temperature Forecast Method

XUE Zhilei[1], ZHANG Shuyu[2]

(1. *College of Atmospheric Sciences*, *Lanzhou University*, *Lanzhou* 210044, *China*;
2. *Gansu Provincial Meteorological Bureau*, *Lanzhou* 730020, *China*)

Abstract: As we know, temperature forecast is important to the weather prediction. Advance in research and operational application of temperature forecast methods is briefly reviewed in this paper. All the mentioned methods have been contrasted and estimated, and a new fine forecast idea is pointed out which can be used as a reference on improving the operational forecast of temperature.

Key words: temperature forecast; method; progress; fine forecast