DOI:10.13878/j.cnki.jnuist.2016.06.011

赵滨[1]　李子良[2]　张博[1]

蒙特卡罗方法在降水评估中的应用研究

摘要

预报检验重点关注预报与观测间的综合统计特征用以探讨模式预报性能,而统计显著性检验方法是衡量评估结论的重要指标,是判断预报效果改进与否的有效手段.当前诸多重要检验指标如降水技巧评分等由于不满足正态分布特征均难以采用简单的计算方式获得置信区间以衡量检验指标的误差特征,因此难以正确判断通过统计检验所获得的评估差异是真实反映模式预报效果差异或是由检验样本的不确定性所造成.蒙特卡罗方法可通过样本重构获取正态分布的统计样本从而有效地解决这一问题.采用 2015 年 8 月的 T639 模式及 GRAPES 全球预报模式 24 h 降水预报产品,使用中国区域 2 400 站日降水资料作为实况,重点研究蒙特卡罗方法在统计显著性检验中的应用特征,分析不同蒙特卡罗重构次数对检验结果的收敛性.结果表明 10 000 次蒙特卡罗重构后统计指标可满足正态分布,而通过显著性检验分析后可明显区分预报系统间降水评分差异的统计特征.

关键词

蒙特卡罗方法;降水技巧评分;显著性检验;T639 模式;GRAPES 全球模式

中图分类号 P435

文献标志码 A

收稿日期 2015-12-25

资助项目 国家自然科学基金青年基金(41305091);中国气象局成都高原气象研究所基金(LPM201401);公益性行业专项课题(GYHY201506002)

作者简介

赵滨,男,博士,主要从事数值预报检验评估工作.zhaob@ cma.gov.cn

1　国家气象中心,北京,100081

2　黑龙江黑河市气象局,黑河,164300

0　引言

预报检验关注不同预报系统之间的差异,通常采用模式预报产品与相同实况产品(或分析产品)进行比较获取统计差异特征[1-2].针对此类统计差异,常利用均值检验方法取得预报系统的综合评估信息,但总体检验样本中可能存在的异常样本将严重影响均值检验结果而产生评估误判.当前统计检验中尚未形成一套标准的方法来解决此类预报要素(特别是降水)检验结果的不确定性问题[3-8],而不采用特定的评估手段以确定均值检验结果属性就不能判断检验效果的改进是由预报系统自身优势还是由随机样本"突变"造成的[9-10].Stephenson[11]指出缺乏有效验证的检验评估结论是毫无意义的,即需要采用显著性检验方法确定预报系统间差异的具体属性.

置信区间是一种应用于显著性检验的有效方法[12-13],它利用在一定显著性水平下的总体统计量的区间估计获取均值样本的显著性特征.当前应用于预报检验的气象要素主要分为两类,一类是如温度、风速等连续性变量,另一类是诸如降水预报等非连续性变量.针对这两类预报要素的置信区间计算方法很大程度上取决于检验要素的样本分布.常规连续性变量在统计检验中样本大多满足正态分布特征,可采用简单的置信区间计算方法获取区间估计值以完成显著性检验,而对于降水等非连续性变量,其检验指标大多不满足正态分布特征,需要采用样本重构方法构建正态分布的检验样本.蒙特卡罗方法作为一种简单的样本重构方法可在非正态分布检验要素的显著性检验中使用.

蒙特卡罗方法基于 Markov 链构建,是预报检验中较常使用的标准方法[14-18],它主要是利用随机函数将原始样本序列通过多次样本重构过程,随机组合成满足正态分布的样本序列.当前 NCEP 等业务预报中心在降水均值检验中多采用蒙特卡罗方法进行样本重构,同时,由于蒙特卡罗方法计算相对简单,非常适用于应用到降水显著性检验评估中.本文利用 T639 模式及 GRAPES 全球预报模式 2015 年 8 月整月 24 h 降水预报产品,采用中国区域内 2 400 站实况站点日降水资料作为对比分析资料计算 24 h 降水技巧评分.通过蒙特卡罗方法对降水技巧评分进行样本重构,并采用置信区间计算方法获取区间估计,以了解特定显著性水平下两个模式的降水预报差异性质.

1 资料和方法

本文重点分析2015年8月整月24 h累计降水技巧评分差异的显著性特征.降水实况采用国家气象信息中心提供的2 400站实况资料,模式预报采用中国气象局数值预报中心开发的GRAPES(Global-Regional Assimilation and Prediction System)全球预报模式24 h降水预报产品,分辨率0.25°,以及中国气象局数值预报中心业务运行的T639模式24 h降水预报产品,分辨率为0.281 25°,两模式均采用12UTC启报资料,选取中国区域(70~145°E,15~65°N范围内的中国陆面区域)进行降水统计评估分析.

当前降水检验中主要采用二分类统计检验方法获取降水技巧评分,即通过考察不同降水阈值条件下预报与观测之间的有无关系计算统计指标.表1给出了二分类降水检验分类,其中a表示参与计算的观测站点上预报与观测均出现满足阈值强度降水,b表示预报出现满足阈值强度降水,而实况未出现,c表示实况出现满足阈值强度降水,而预报未出现,d则表示预报和实况均未出现满足阈值强度降水.

表1 二分类降水检验分类

Table 1 Contingency table for precipitation forecasts of binary events

	实况(有)	实况(无)
预报(有)	a	b
预报(无)	c	d

通过对参与统计的所有站点计算结果分析获取降水准确率(T)及偏差(B)技巧评分,其统计公式如下:

$$T = a/(a + b + c), \tag{1}$$

$$B = (a + b)/(a + c), \tag{2}$$

其中,当$B>1$表示模式降水预报以空报为主,$B<1$则表示降水预报以漏报为主.

在统计检验中往往采用均值来衡量预报变量的综合评估性能,但在均值检验中大多未采用显著性检验方法来进一步讨论评估结论是否处于合理范围之内,即评估结论是否可信.本文在均值检验中增加显著性检验方法,利用置信区间的概念来了解不同预报模式间的差异特征.

所谓置信区间是指由样本统计量所构造的总体参数的估计区间.置信区间展现的是这个参数的真实值在某一概率水平(置信水平)下的可信范围.置信区间的两端被称为置信极限(上限及下限).

以单独总体均值的置信区间为例,对于正态分布$N(\mu,\sigma^2)$的总体,样本均数$\bar{X} \sim N(\mu,\sigma^2/n)$,其置信度为$1-\alpha$的置信区间为$\bar{X} \pm t_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$,$\sigma$为样本的标准差.

对于相同样本数的两个总体样本μ_1和μ_2,由于两个总体的样本数相同,其差值可以退化为单独总体样本的均值检验问题,则定义:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - Y_i) - (\bar{X} - \bar{Y})]^2, \tag{3}$$

即偏差标准差.

其置信下限定义为

$$(\bar{X} - \bar{Y}) - t_{\alpha/2}\left(\sqrt{\frac{\sigma^2}{n}}\right), \tag{4}$$

置信上限为

$$(\bar{X} - \bar{Y}) + t_{\alpha/2}\left(\sqrt{\frac{\sigma^2}{n}}\right), \tag{5}$$

其中$\bar{X}$和$\bar{Y}$为两样本总体均值.

对于95%的显著性检验,表明样本的差异有95%的可能性处于置信区间之内.因此当置信下限$(\bar{X} - \bar{Y}) - t_{\alpha/2}\left(\sqrt{\frac{\sigma^2}{n}}\right) > 0$时,证明整个置信区间均大于0,表现为偏差整体为正;当置信上限$(\bar{X} - \bar{Y}) + t_{\alpha/2}\left(\sqrt{\frac{\sigma^2}{n}}\right) < 0$时,证明整个置信区间均小于0,表现为偏差整体为负;当置信下限与置信上限异号,即置信区间处于正负之间时,偏差可正可负,也就是说无法判断样本的均值差异性质.使用置信检验的方法就可以很好地判断统计结论中差异的来源和特性,可以更好地判断和确定所得到的统计检验的性质.

对置信检验公式进一步推导可得:

$$\text{当}(\bar{X} - \bar{Y}) > t_{\alpha/2}\left(\sqrt{\frac{\sigma^2}{n}}\right)\text{时,偏差显著为正,} \tag{6}$$

$$\text{当}(\bar{X} - \bar{Y}) < -t_{\alpha/2}\left(\sqrt{\frac{\sigma^2}{n}}\right)\text{,偏差显著为负,} \tag{7}$$

这就将问题归结为样本平均偏差和置信估计两者的比较问题.

对于具有正态分布的样本总体而言,采用常规的置信区间计算方法即可完成显著性检验,而对于降水技巧评分这种具有明显非正态分布的样本而言,需要采用特定的样本重构方法构建具有正态分

布的总体样本用于显著性检验.蒙特卡罗方法作为一种样本重构方法可用于降水技巧评分的显著性检验以评估预报系统间差异.

对于给定的降水量阈值,以降水准确率评分为例,式(1)改写为

$$T=\sum_{t=1}^{t_m}a^t/\sum_{t=1}^{t_m}(a^t+b^t+c^t), \tag{8}$$

其中 $t=1,2,3,4,\cdots,t_m$ 为时间序列.对于两个不同预报系统,降水评分可改写为

$$T_i=\sum_{t=1}^{t_m}a_i{}^t/\sum_{t=1}^{t_m}(a_i{}^t+b_i{}^t+c_i{}^t), \tag{9}$$

$$T_j=\sum_{t=1}^{t_m}a_j{}^t/\sum_{t=1}^{t_m}(a_j{}^t+b_j{}^t+c_j{}^t), \tag{10}$$

其中 i,j 表示两个预报系统.采用蒙特卡罗方法首先获取随机变量 r^t,设定 r^t 为 0 或者 1,则有:

$$a_{ik}^t=\begin{cases}a_i^t, & r^t=1,\\ a_j^t, & r^t=0,\end{cases}\quad a_{jk}^t=\begin{cases}a_j^t, & r^t=1,\\ a_i^t, & r^t=0.\end{cases} \tag{11}$$

同样方法计算评分值(b、c、d),则基于式(11)重新计算降水评分 T_{ik} 及 T_{jk},其中 k 为蒙特卡罗数,通过 k 次重构可最终获取正态分布的总体样本用于显著性检验以获取不同预报系统间的差异特征.

2 评估分析

利用式(1)、(2)计算 T639 模式及 GRAPES 全球模式 2015 年 8 月的 24 h 累计降水准确率及偏差评分,降水实况选取中国区域 2 400 站降水实况站点资料.图 1 给出了 10 mm 以上降水量阈值条件下两个模式的降水技巧评分时间序列,可以看到,GRAPES 模式与 T639 模式在降水准确率技巧评分上表现基本相当,技巧评分均保持在为 0.3 左右,而 8 月 4 日及 10 日显示出较大的差异,两模式分别表现出较为明显的预报优势.从偏差评分上看,T639 模式基本处于 1 以上,显示出较为明显的空报现象,而 GRAPES 模式整体表现更好(更接近 1),且综合表现是降水预报偏保守,存在一定的漏报现象.

在统计检验中,通常采用均值检验考察模式预报要素的基本性能,即检验模式预报要素的综合预报能力.图 2 给出了不同阈值条件下 24 h 累计降水评分在 8 月时段内的平均分布状况.可以看到 0.1 mm 以上降水预报中,T639 模式及 GRAPES 模式的降水准确率技巧评分分别为 0.53 及 0.50,差异仅为 0.03,随着阈值的增大,两模式综合的技巧评分差异基本保持在 0.03 左右.而偏差评分分布特征存在一定差异,其中 GRAPES 模式在小量级降水(1 mm 以下)空报高于 T639 模式,而在 10 mm 以上空报现象明显减弱,表现出一定的降水漏报现象,漏报方面也较 T639 模式更为明显.从图 2 中可以主观判断两个模式降水预报能力,即 T639 模式相对于 GRAPES 模式在各个降水阈值条件下均显示出一定的预报优势,但此种优势是真实反映模式的整体预报能力还是由某一时段技巧评分"突变"(明显提高或明显降低)的孤立事件引起的无法确定,这种孤立的样本将直接影响总体样本的最终评估结论.因此判断样本均值之间的显著性水平是考察预报差异属性的必然手段.

对降水技巧评分进行显著性检验,首先要确认降水技巧评分本身的分布特征.以 1 及 10 mm 降水阈值为例,图 3 给出了两种阈值条件下两个模式降

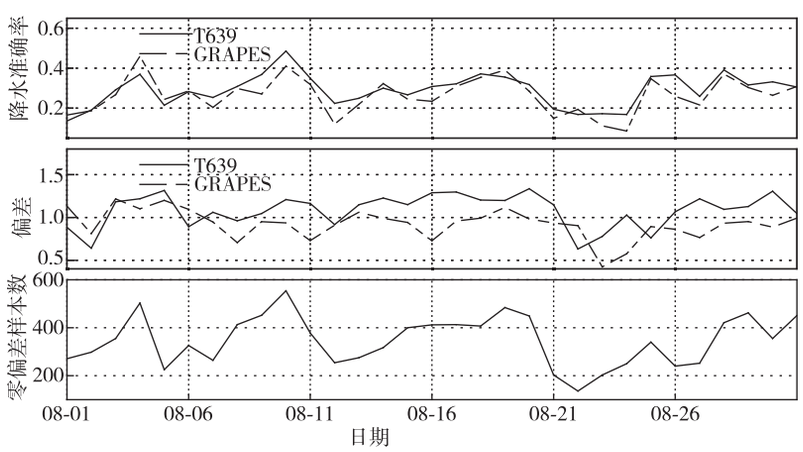

图 1 2015 年 8 月 1—31 日 T639 模式及 GRAPES 全球模式 24 h(12~36 hr)降水技巧评分分布(降水阈值≥10 mm/d)

Fig. 1 Time series of daily(f12-f36) precipitation skill scores from 01 Aug,2015 to 31 Aug, 2015 with T639 and GRAPES(threshold is≥10 mm/d)

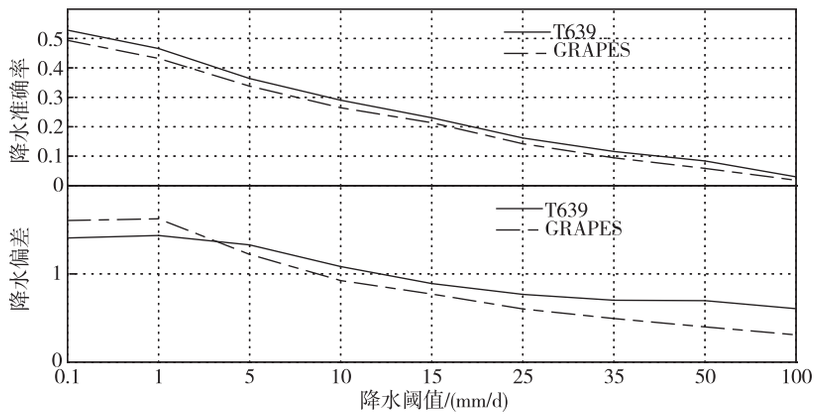

图 2　不同阈值条件下降水技巧评分的平均分布(2015 年 8 月 1—31 日)

Fig. 2　Mean precipitation skill scores over different thresholds from 01 Aug,2015 to 31 Aug,2015

水准确率及偏差对应的样本(天数)分布.可以看到,10 mm 阈值准确率评分差异 0.28 对应的样本数最多,为 51.6%,而偏差评分差异在 1.6 处样本数最多,为 61.5%,均不满足正态分布特征.因此降水技巧评分不能采用常规的显著性检验方法,需要利用蒙特卡罗方法对样本进行重构,产生符合正态分布的总体样本后进行显著性检验.

在采用蒙特卡罗方法进行样本重构前,需要进行敏感试验以确定需要进行的样本重构次数.可通过考察不同样本重构次数所获取评估结果的稳定性确认重构次数的选取,图 4 给出了不同蒙特卡罗重构次数下不同阈值降水评分的标准差分布,可以看到,100~5 000 次重构后,评分标准差持续增加,而 10 000 次后评分标准差基本稳定,其中 5 mm 阈值以上准确率评分的标准差稳定在 0.024 左右,而偏差评分的标准差也基本稳定于 0.2 以内.因此可采用 10 000 次蒙特卡罗重构进行显著性检验.

图 5 给出了 1 和 10 mm 降水阈值下,经过 10 000 次蒙特卡罗样本重构后的样本数分布.可以看到,重构的样本严格满足高斯正态分布特征.大多数样本处于 95%的置信区间之内,而处于置信区间外的样本则表现为显著性差异.通过蒙特卡罗重构后,利用式(9)、(10)即可计算两个模式降水预报技巧评分差异,并获得 95%的置信估计以确定降水评估的预报差异属性.

图 6 给出了通过显著性检验的各降水阈值条件下技巧评分的平均分布,可以看到,图 6 拓展了图 2 所获取的评估信息.T639 模式在中雨(10 mm)以下量级的降水准确率评分相对于 GRAPES 模式的优势通过了显著性检验,其差异均通过了 95%的显著性检验,体现为显著改进.中雨及以上量级降水评分虽在均值分布上体现出相对 GRAPES 模式的优势,但由于样本自身统计指标的波动较大,差异均未通过显著性检验,即可认为改进不显著.同样对于降水偏

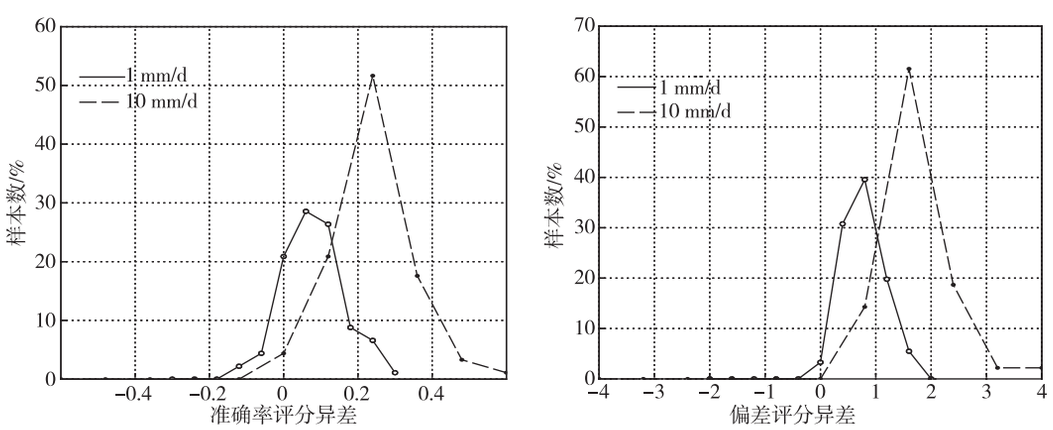

图 3　1 及 10 mm 阈值下不同技巧评分差异所对应的样本数分布

Fig. 3　Comparison of count frequencies for skill score difference above 1 and 10 mm/d thresholds

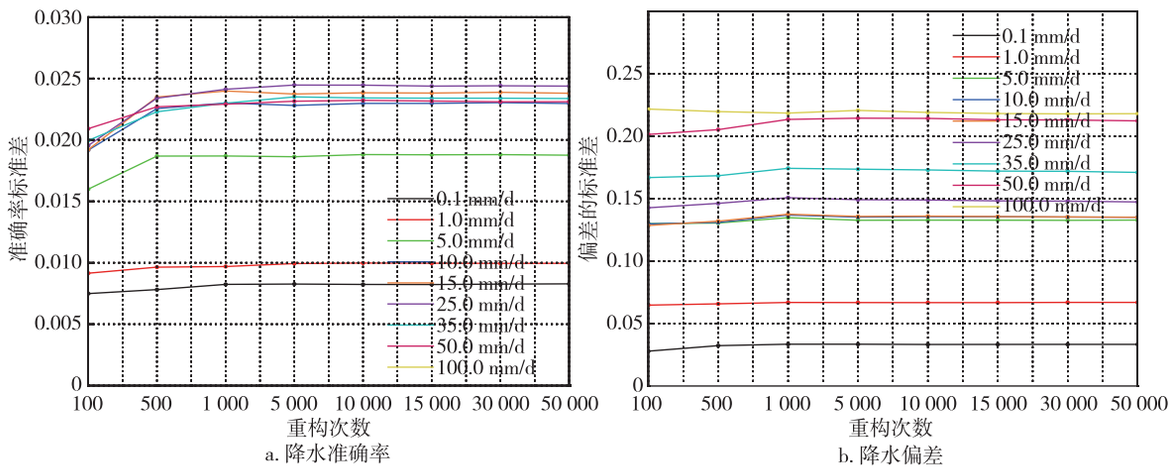

图 4　不同蒙特卡罗样本重构次数下不同阈值降水评分的标准差分布

Fig. 4　Comparison of standard deviations of precipitation skill scores with different resampling tests

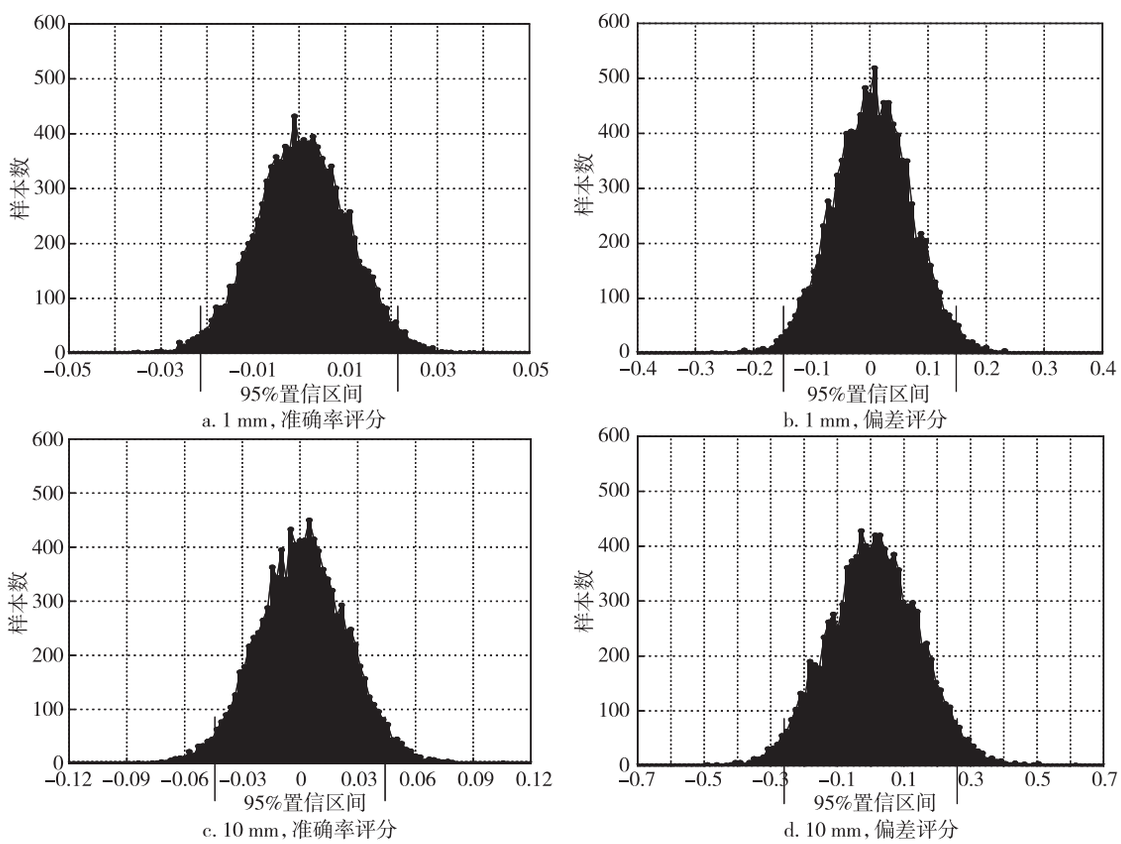

图 5　1 和 10 mm 阈值下不同降水技巧评分所对应的样本数分布

Fig. 5　Count distribution of TS and BIAS differences above 1 and 10 mm/d thresholds

差评分,可以看到,GRAPES 模式相对于 T639 模式在中雨以上量级的漏报及在小雨以上量级的空报现象均通过了显著性检验,表明均值分布中体现出的空报及漏报现象即反映两模式间的真实预报特性.通过上述显著性检验可获取较图 2 中更为确切的评估信息,因此可了解降水技巧评分差异的属性易于准确的评估模式降水预报性能.

3　结论

准确率及偏差降水技巧评分作为重要的降水统

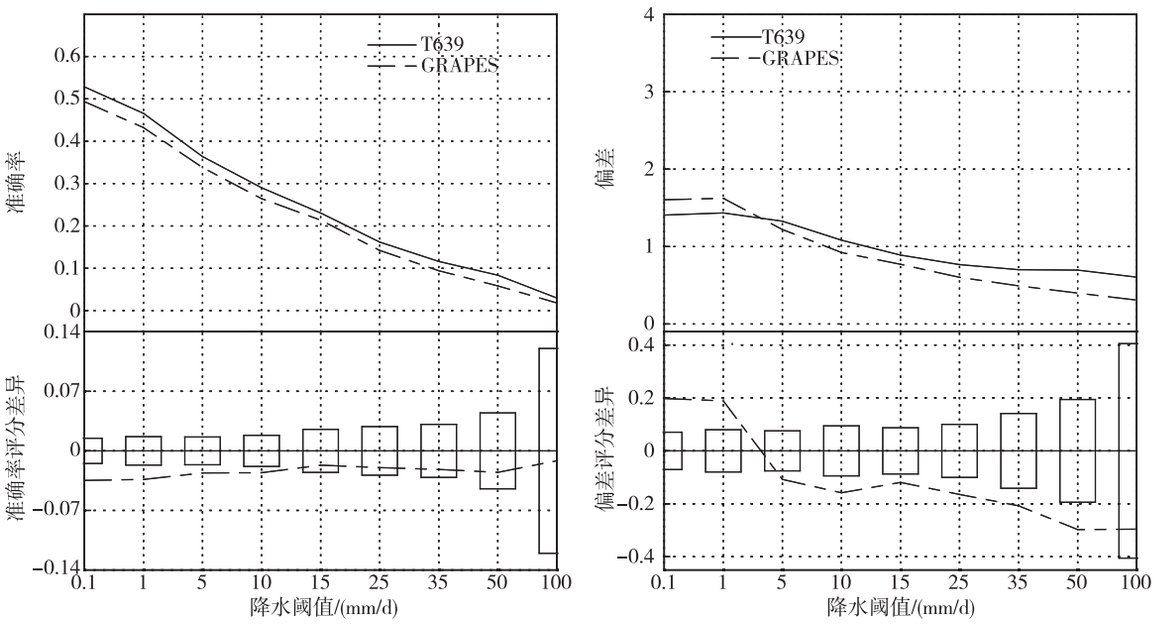

图 6　通过显著性检验的不同阈值条件下降水技巧评分的平均分布

Fig. 6　Mean precipitation skill scores with Monte Carlo significance test from 01 Aug,2015 to 31 Aug,2015

计检验指标是评估模式降水预报效果的主要手段,常规业务检验中通常采用简单的统计平均考察预报要素的综合效果,但由样本不确定性所引起的评估误导问题却很少被涉及.采用均值显著性检验方法可有效获得评估差异的真实属性,即差异是真实反映了模式预报能力差异还是由随机样本的评估"突变"所引起的.由于降水技巧评分不满足正态分布特征,所以无法通过常规的置信区间计算方法获取显著性检验信息.利用蒙特卡罗方法对降水技巧评分进行样本重构可有效获取正态分布总体样本,以此进行的显著性检验可用来评估降水检验特征.

本文利用中国气象局数值预报中心研发的 T639 模式及 GRAPES 全球预报模式 24 h 降水预报及站点实况日降水资料计算准确率及偏差降水技巧评分,基于蒙特卡罗方法获得显著性检验结果以分析评估差异的真实属性.分析表明,经过 10 000 次蒙特卡罗样本重构后即可获取满足正态分布的稳定的检验样本,经过显著性检验发现 T639 模式在小雨以上量级表现出相较 GRAPES 模式的一定优势,同时在中雨以上量级预报效果优势不显著.

本文重点考察降水检验评估的不确定问题,显著性检验方法的使用可有效弥补传统评估方式中难以分析由于随机样本不确定性而引起的评估误导问题.蒙特卡罗方法仅是一种样本重构方法,此方法构建简单,较易在降水评估计算中使用.当前还有诸如 Bootstrap 方法等多种样本重构方法也均可以应用于非正态分布样本分析中,蒙特卡罗方法与其他样本重构方法的分析差异及对评估效果的影响分析将在未来工作中进一步讨论.

参考文献

References

[1] Katz R W. Statistical evaluation of climate experiments with general circulation models: A parametric time series modeling approach[J]. Journal of the Atmospheric Sciences, 1982, 39(7): 1446-1455

[2] Brown B G, Thompson G, Bruintsjes R T, et al. Intercomparison of in-flight icing algorithms. Part II: Statistical verification results[J]. Weather & Forecasting, 1997, 12(4): 890-914

[3] 胜春岩,薛德强,雷霆,等.雷达资料同化与提高模式水平分辨率对短时预报影响的数值对比试验[J].气象学报,2006,64(3):293-307
SHENG Chunyan, XUE Deqiang, LEI Ting, et al. Comparative experiments between effects of Doppler radar data assimilation and increasing horizontal resolution on short-range prediction[J]. Acta Meteorologica Sinica, 2006, 64(3): 293-307

[4] Lang X M. A hybrid dynamical-statistical approach for predicting winter precipitation over Eastern China[J]. Acta Meteorologica Sinica, 2011, 25(3): 272-282

[5] 张冰,魏建苏,裴海瑛.2006 年 T213 模式在江苏的降水和温度检验评估[J].气象科学,2008,28(4):468-472
ZHANG Bing, WEI Jiansu, PEI Haiying. Verification and evaluation of rainfall and temperature forecasting for

T213L31 global numerical model in Jiangsu area in 2006 [J].Scientia Meteorologica Sinica,2008,28(4):468-472
[6] 梁红,王元,钱昊,等.欧洲 ECWMF 模式与我国 T213 模式夏季预报能力的对比分析检验[J].气象科学,2007,27(3):253-258
LIANG Hong,WANG Yuan,QIAN Hao,et al.Verification and comparative analysis of prediction of ECMWF model and T213 model in summer[J].Scientia Meteorologica Sinica,2007,27(3):253-258
[7] 王晨稀.MM5 模式中不同对流参数化方案对降水预报效果影响的对比试验[J].气象科学,2004,24(2):168-176
WANG Chenxi.Comparison experiments on the effects of different cumulus parameterization scheme in MM5 on precipitation[J].Scientia Meteorologica Sinica,2004,24(2):168-176
[8] 王明欢,沈学顺,肖锋.GRAPES 模式中高精度正定保形物质平流方案的研究Ⅱ:连续实际预报试验[J].气象学报,2011,69(1):16-25
WANG Minghuan,SHEN Xueshun,XIAO Feng.A study of the high-order accuracy and positive-definite conformal advection scheme in the GRAPES model Ⅱ:Continuous actual rainfall prediction experiments[J].Acta Meteorologica Sinica,2011,69(1):16-25
[9] Jolliffe I T,Stephenson D B.Forecast verification:A practitioner's guide in atmospheric science[M].2nd ed.Hoboken,New Jersey:Wiley and Sons Ltd,2012
[10] Jolliffe I T.Uncertainty and inference for verification measures[J].Weather & Forecasting,2007,22(3):637-650
[11] Stephenson D B.Use of the 'odds ratio' for diagnosing forecast skill[J].Weather & Forecasting,2000,15(2):211-232
[12] Seaman R,Mason I,Woodcock F.Confidence intervals for some performance measures of yes/no forecasts[J].Aust Meteorol Mag,1996,45:49-53
[13] Wilks D S.Statistical methods in the atmospheric sciences[M].2nd ed.Pittsburgh:Academic Press,2006
[14] Neumann C J,Lawrence M B,Caso E L.Monte Carlo significance testing as applied to statistical tropical cyclone prediction models[J].Journal of Applied Meteorology,1977,16(11):1165-1174
[15] Livezey R E,Chen W Y.Statistical field significance and its determination by Monte Carlo techniques[J].Monthly Weather Review,1983,111(1):46-59
[16] Dixon K W,Shulman M D.A statistical evaluation of the predictive abilities of climatic averages[J].Journal of Applied Meteorology,1984,23(11):1542-1552
[17] Chen W Y.Another approach to forecasting forecast skill[J].Monthly Weather Review,1989,117(2):427-435
[18] Robert C P,Casella G.Monte Carlo statistical methods[J].New York:Springer Verlag,2004

Application of Monte Carlo significance test in precipitation skill score

ZHAO Bin[1] LI Ziliang[2] ZHANG Bo[1]
1 National Meteorological Center,Beijing 100081
2 Heihe Weather Office of Heilongjiang Province,Heihe 164300

Abstract Forecast verification involves exploring and summarizing the relationship between sets of forecast and observation and making comparisons between the performances of different forecasting systems.Statistical significance is one important aspect of measuring the absolute quality of verification results.It is an effective way to judge whether the performance improvement is statistically significant or just arisen by chance. For general meteorological forecast verification, some verification scores, such as precipitation skill scores, can hardly use the standard procedure for confidence interval to measures the difference in performance between different forecast systems.It is not possible to be sure that the apparent differences in skill scores are real and not just due to random fluctuations because of the data uncertainty.The Monte Carlo method is a numerical way to account for this.By resampling process,we can provide an adequate representation of the full underlying population which satisfies normal distribution by the verification samples of random variables.In this paper,some precipitation skill scores of GRAPES global forecast system and T639 models such as Threat Score and Bias Score are calculated from 1Aug to 31 Aug of 2015. The daily precipitation observation data are taken from 2 400 Chinese rain gauges.The Monte Carlo method is used for a statistical significance test and the convergence characteristics with different resampling times are also analyzed.Results show that Monte Carlo test using 10 000 test samples looks sufficient and a real model performance with significant improvement is provided.

Key words Monte Carlo method;precipitation skill score;significance test;T639 mode;GRAPES global mode