

王敏, 李晓莉, 范广洲, 等. 区域集合预报系统 2 m 温度预报的校准技术. 应用气象学报, 2012, 23(4): 395-401.

# 区域集合预报系统 2 m 温度预报的校准技术

王 敏<sup>1)2)</sup> 李晓莉<sup>2)\*</sup> 范广洲<sup>1)</sup> 李泽椿<sup>3)</sup>

<sup>1)</sup>(成都信息工程学院, 成都 610225)

<sup>2)</sup>(中国气象局数值预报中心, 北京 100081) <sup>3)</sup>(国家气象中心, 北京 100081)

## 摘 要

采用非齐次高斯回归(NGR)技术对国家气象中心区域集合预报系统的 2 m 温度预报结果开展了一阶偏差和二阶离散度的校准研究。对预报结果比较详尽的检验分析表明:校准后的 2 m 温度预报可靠性和预报技巧均显著提高,表现为校准后集合预报成员的均方根误差与离散度更为接近;原 Talagrand 直方图中的“L”形分布现象得到有效改善;Brier 评分、最小连续分级概率评分(CRPS)明显减小,相对作用特征(ROC)面积增大,说明校准后的 2 m 温度预报表现出更好的预报技能。此外,NGR 技术与自适应误差订正技术的对比试验表明,NGR 在消除集合平均偏差和提高集合离散度两个方面均有优势。

**关键词:** 集合预报; 偏差校准; 非齐次高斯回归(NGR); 检验评分

## 引 言

集合预报技术在过去的二十多年里取得了巨大进步,现已成为数值天气预报的重要组成部分,并在水文预报等领域也得到了广泛应用。但是集合预报仍受模式误差及初值扰动技术的限制,使得预报结果离散度不足,预报分布存在偏差。为了提高集合预报系统的预报技巧,近几年,在传统的数值模式后处理技术如 MOS (Model Output Statistics) 方法<sup>[1-2]</sup>的基础上,一系列基于各种统计方法的校准技术被广泛用于国外集合预报概率产品的校准。这些方法包括贝叶斯模式平均(BMA)<sup>[3]</sup>,非齐次高斯回归(NGR)<sup>[4]</sup>, Logistic 回归<sup>[5-6]</sup>, Analog 技术<sup>[7-8]</sup>, 类似 Kalman 滤波的自适应误差订正技术<sup>[9]</sup>等。国内也开展了针对集合预报后处理技术的研究及应用,林春泽等<sup>[10]</sup>基于 TIGGE 资料,利用超级集合平均、多模式集合平均和消除偏差集合平均 3 种方法对多模式集合预报的地面气温预报进行集成;马清等<sup>[11]</sup>采用自适应误差订正技术进行了 B08RDP 多中心区域集合预报系统 2 m 温度预报结果的偏差订正。这些技术各具特点,但校准效果仍受到相关

条件的限制,使其具有一定的适用性,如 BMA 在校正多模式集合预报时才能体现其优势,训练数据的大小对 Analog 技术和 Logistic 回归技术的校准效果有明显的影 响等。

本文研究对象是国家气象中心区域集合预报系统的 2 m 温度预报,已有相关研究<sup>[12]</sup>对 2008 年北京奥运会期间包括国家气象中心在内的 B08RDP 各参与成员的 2 m 温度预报进行了较全面的评估,发现大部分系统对 2 m 温度的预报都存在着偏差和离散度不够的问题。Alexander 等<sup>[13]</sup>研究结果表明,NGR 技术在训练期较短的情况下,仍能有效校准具有高斯分布特征的连续变量(如 2 m 温度)的集合平均偏差和集合离散度,且校准过程简单。因此,本文将采用 NGR 技术进行区域集合预报系统 2 m 温度预报的校准研究,并通过综合的集合预报检验评分来检验校准前后的预报效果,最后初步开展 NGR 技术与自适应误差订正技术的比较研究。

## 1 区域集合预报系统及资料简介

国家气象中心区域集合预报系统是依托于 B08-RDP 项目而发展的<sup>[14]</sup>,该系统共有 15 个集合成员,

2011-08-05 收到, 2012-05-10 收到再改稿。

资助项目:公益性行业(气象)科研专项(GYHY201006015)

\* 通信作者, E-mail: lixl@cma.gov.cn

模式系统为中尺度模式 WRF, 模式水平分辨率为  $15\text{ km} \times 15\text{ km}$ , 垂直层次为 35 层; 初值扰动技术采用增长模繁殖法, 侧边界来自于 T213 模式的全球集合预报系统; 模式扰动技术采用的是多物理过程参数化方案方法。B08RDP 期间, 区域集合预报系统的预报范围是华北地区, 此后对该系统进行了加入陆面过程 NOAH 方案及拓展预报区域到全国范围的升级试验, 并于 2010 年 7 月实现了升级后的区域集合预报系统的准业务运行。

本文将国家气象中心区域集合预报系统 2008 年 7 月 20 日—8 月 20 日每日 12:00(世界时, 下同)起报的 2 m 温度预报为研究资料, 开展校准技术应用和分析试验。检验的观测资料采用预报区域 ( $30 \sim 45^\circ\text{N}$ ,  $105 \sim 125^\circ\text{E}$ ) 内 400 个地面站每 3 h 的 2 m 温度观测资料。为了便于直接使用, 先采用双线性插值法将格点预报结果插值到站点。

## 2 非齐次高斯回归校正技术原理及应用

非齐次高斯回归(NGR)是 Gneiting 等<sup>[4]</sup>基于标准线性回归而发展的针对集合预报的后处理技术, 与其他集合预报校准技术相比, 其显著优点是在训练期较短的情况下, 仍能有效校准具有高斯分布特征的连续变量。

2 m 温度预报是区域集合预报系统的重要预报变量之一, 其预报准确率被预报员所关注。研究表明: 2 m 温度的观测统计近似服从正态分布, 因此从理论上讲, 校准后的集合平均和方差可以用来表示 2 m 温度正态分布的位置和形状。在 NGR 方法中, 集合平均和方差由线性回归技术获得, 而回归过程中方差随着预报因子的值而发生变化, 因此, 这种回归又是非均匀的。

NGR 技术所模拟的具有正态分布形式的 2 m 温度预报概率密度函数为

$$\rho = N(a + b\mu, c + d\sigma^2). \quad (1)$$

式(1)中,  $\mu$  是集合平均,  $\sigma$  是标准差(集合离散度)。  $a, b, c, d$  是待定的回归系数, 其中系数  $a$  和  $b$  表示偏差的特征, 系数  $c$  和  $d$  表示集合预报离散度-技巧关系。本文采用最小连续分级概率评分(简称 CRPS)来确定变量概率密度函数的参数。

CRPS 是一种定量比较预报累计分布概率与观测累计分布概率距离差异的集合预报概率评分方法。对于一个连续型预报量  $x$ , 其概率密度函数为

$\rho(x)$ , 观测值为  $x_a$ , 则连续分级概率评分计算公式可表示为

$$S_{\text{crp}}(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx. \quad (2)$$

$P, P_a$  分别表示集合概率预报和观测值的累计分布函数:

$$P(x) = \int_{-\infty}^x \rho(y) dy, \quad (3)$$

$$P_a(x) = H(x - x_a). \quad (4)$$

$$H(x - x_a) = \begin{cases} 0 & x < x_a, \\ 1 & x \geq x_a. \end{cases} \quad (5)$$

当预报变量满足高斯分布时, 一定训练期内的 CRPS 可以由式(1)中的系数  $a, b, c$  和  $d$  解析地表示出:

$$S_{\text{crp}} = \frac{1}{k} \sum_{i=1}^k (c + d\sigma_i^2)^{1/2} \times \left\{ Z_i [2\Phi(Z_i) - 1] + 2\phi(Z_i) - \frac{1}{\sqrt{\pi}} \right\}, \quad (6)$$

$$Z_i = \frac{Y_i - (a + b\mu_i)}{(c + d\sigma_i^2)^{1/2}}, \quad Z_i \sim N(0, 1). \quad (7)$$

这里,  $\Phi$  是累积分布函数,  $\phi$  是概率密度函数,  $k$  是训练期的长度(单位:d),  $Y_i$  是第  $i$  天的观测值。

采用 Nelder 等<sup>[15]</sup>提出的单纯形法来迭代估算回归系数, 然后由回归系数可计算出式(1)中校准后的集合平均和离散度。为了便于比较校准前后每个成员的预报结果, 必须利用正态分布函数计算出校准后的每个成员值。当集合成员为  $n$  时, 一种可选方法是, 由标准正态分布的第  $i/(n+1)$  个分位函数计算第  $i$  个集合成员值<sup>[16]</sup>, 但是这种方法校准后的集合离散度偏大。为了克服这一缺陷, Alexander 等<sup>[13]</sup>发展了依赖于预报概率的正态分布的分位函数技术, 通过限定校准后的离散度范围, 从而重新调整各集合成员的分位函数值, 使得校准后的离散度更合理, 然后根据校准后的集合平均和离散度来确定不同集合成员的预报。本文采用上述方法来计算校准后的集合成员的预报值:

$$N(i) = Q(p(i))(c + d\sigma^2) + (a + b\mu). \quad (8)$$

这里,  $Q(p(i))$  是标准正态分布的分位函数, 第  $i$  个成员的概率  $p(i)$  为

$$p(i) = \frac{1-z}{2} + (i-1)\left(\frac{z}{n-1}\right). \quad (9)$$

$z$  是调整后的集合预报的百分位面积。其值根据限

制条件  $\sigma \leq fE$  循环计算而得到。其中,  $\sigma$  为标准差,  $f$  为离散度因子,  $E$  为均方根误差。  $f$  由最小的 CRPS 结果确定, 本文将在后面章节讨论  $f$  的敏感性。经过试验分析, 当前采用  $f=1$ 。

图 1 是校准前后集合平均与观测场之差的地理分布图。整体上看, 校准前预报区域西北部集合平

均偏小, 南部偏大, 并且在四川东部, 陕西、山西和河南三省交界处, 山东中部, 合肥南部有些区域预报存在异常偏大(红色)现象, 校准后整个预报区域与观测值更接近, 预报区域西北部、南部的预报得到明显改善, 且原异常区域的预报值也均得到合理改善。

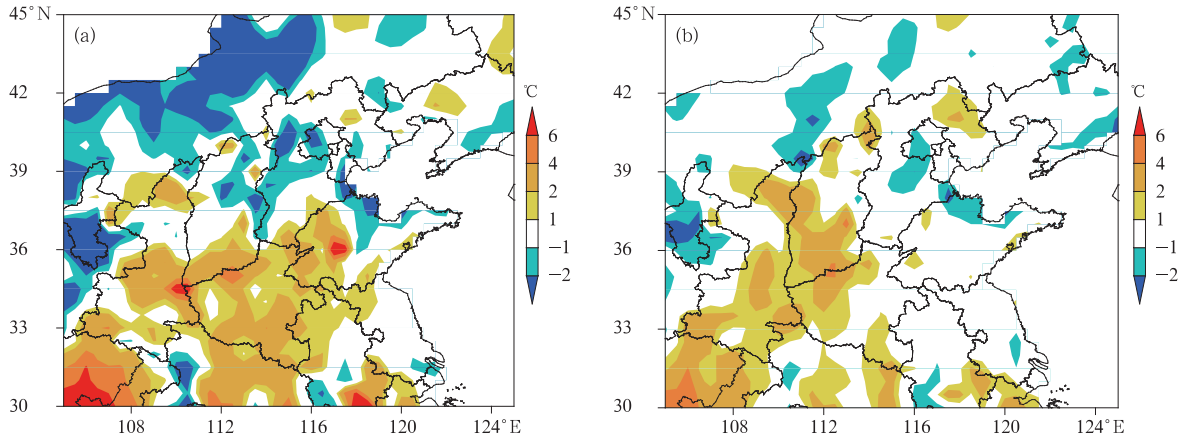


图 1 2008 年 7 月 20 日 12:00 起报的 2 m 温度集合平均(预报时效为 30 h)与观测场之差的空间分布 (a)校准前, (b)校准后

Fig. 1 Distribution of the difference of 30-hour 2 m temperature forecast at 1200 UTC 20 July 2008 (a)the difference between ensemble average of model direct output and observation, (b)the difference between calibrated ensemble average and observation

### 3 校准结果

#### 3.1 集合平均均方根误差和集合离散度的关系

集合平均均方根误差和集合预报离散度的关系是衡量集合预报系统可靠性的一个重要标准, 一个好的集合预报系统的集合平均均方根误差和集合离散度应该基本一致。图 2 给出了校准前后 2 m 温度预报的集合平均均方根误差和集合离散度的预报时效演变图。从图 2 可看出, 校准前模式直接输出的预报存在明显的偏差并且离散度过小, 而校准后的均

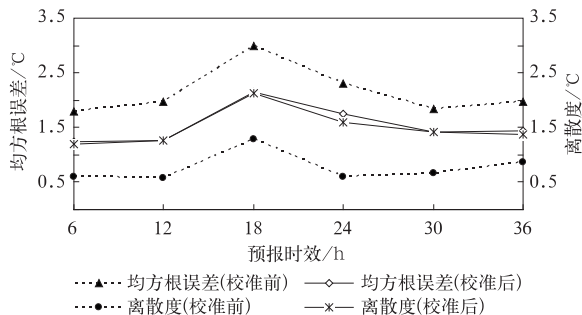


图 2 2 m 温度预报的均方根误差和离散度  
Fig. 2 RMSE and ensemble spread of 2 m temperature

方根误差明显减小, 减小幅度为 0.6°C 左右, 同时校准后的集合离散度也明显增大, 与集合平均均方根误差非常接近。

#### 3.2 Talagrand 直方图分析

Talagrand 直方图是衡量集合预报成员分布与观测值分布是否一致的评分, 能够反映出集合预报系统可靠性及系统偏差的特征。对于具有完美离散度的集合预报系统, Talagrand 表现出比较平坦的分布特征。图 3 给出了校准前后 6 h 预报的 2 m 温

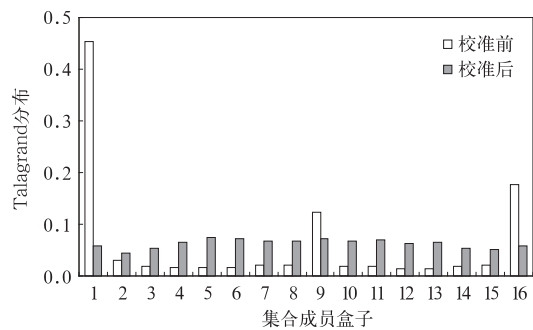


图 3 2 m 温度预报 Talagrand 分布 (预报时效为 6 h)

Fig. 3 Talagrand diagrams of raw and calibrated 2 m temperature (forecast lead time is 6 hours)

度 Talagrand 直方图。由图 3 可以看出,校准前的 Talagrand 分布表现出明显的“L”型分布,这说明集合系统的离散度偏小,并有暖偏差的特征,另外校准前 Talagrand 直方图中的第 9 个集合成员盒子表现出了比较显著的高频率,这可能和集合预报系统的系统偏差有关。校准后 Talagrand 的分布比较平缓,观测值基本以相同的概率落在各盒子中,预报系统的集合离散度得到显著改进,同时,原来出现在第 9 个盒子的明显偏差也得到了改进,说明集合成员分布更合理。

### 3.3 Brier 评分

Brier 评分(简称 BS)是计算预报概率误差的一种概率评分,计算公式可表示为

$$S_B = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2. \quad (9)$$

其中,  $N$  是被检验的二分类事件的样本数;  $f_i$  是事件发生的预报概率;  $o_i$  是事件观测概率,当事件发生时  $o_i = 1$ ,不发生时  $o_i = 0$ 。由此可知,  $S_B$  取值范围是  $0 \sim 1$ ,评分越小预报效果越好,  $S_B = 0$  表示预报完全正确,  $S_B = 1$  表示预报无效。本文以 2 m 温度距平大于 1 个标准差为二分类事件,计算出校准前后各预报时效的 BS 评分(图 4)。相比于原模式直接输出,校准后温度的预报技巧有显著提高,其中 12 h 预报的 BS 评分提高最显著。整体上,校准后各时效预报的 BS 评分比较接近,表明校准后系统的预报能力更为稳定。

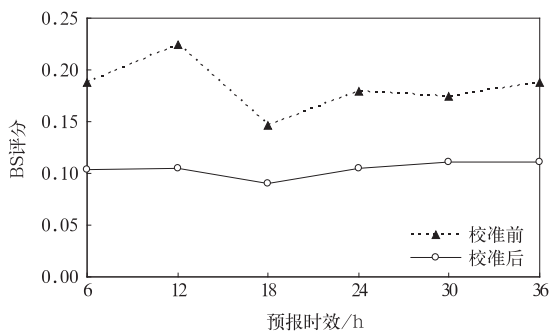


图 4 2 m 温度预报 BS 评分

Fig. 4 BS of raw and calibrated 2 m temperature

### 3.4 CRPS 评分

CRPS 是评价集合预报系统整体性能的定量评分,其值越小表示预报的概率密度和观测的差值越小,预报系统预报能力越高。图 5 给出了校准前后 2 m 温度预报的 CRPS 评分,可以发现经过 NGR 方法校准后的 CRPS 显著减小,减小幅度为  $0.4^{\circ}\text{C}$  左右。

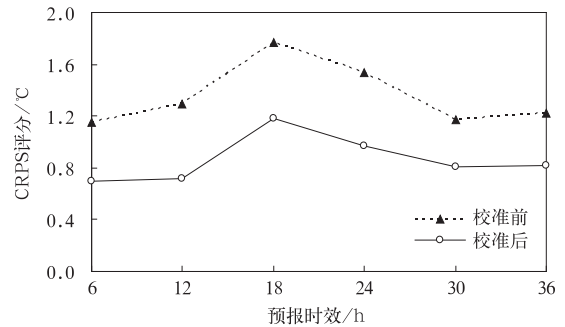


图 5 2 m 温度预报 CRPS 评分

Fig. 5 CRPS of raw and calibrated 2 m temperature

### 3.5 ROC 分析

相对作用特征(ROC)是信号探测理论在数值天气预报中的一种应用<sup>[17]</sup>。用观测值检验预报发生或不发生两种状态,计算出的预报假警报率和命中率在平面直角坐标系中构成的曲线为 ROC 曲线。一个完美的预报假警报率为 0,命中率为 1,将命中率沿假警报率增加的方向积分就可得到 ROC 面积。ROC 面积越大,预报技巧越高,若面积小于 0.5,则预报无技巧。图 6 中给出了二分类事件距平大于 1 个标准差在不同预报时效的 ROC 面积。由图 6 可以看出,校准后各时效预报的 ROC 面积相对校正前均有了明显提高,说明 NGR 技术也能有效提高集合系统对事件的辨别能力,另外,校准后的 ROC 面积曲线更平缓,说明系统预报能力稳定。

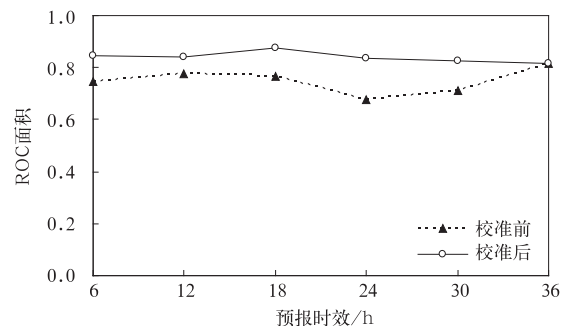


图 6 2 m 温度预报 ROC 面积

Fig. 6 Area under ROC curves for 2 m temperature

## 4 敏感性试验

### 4.1 不同训练期的校准结果对比

训练期的选取直接影响校准技术对集合预报系统的校准效果<sup>[5-7]</sup>。图 7 给出了不同训练期的 2 m 温度 CRPS 评分结果,从图 7 中可看出 20 d, 30 d 训练期的 CRPS 差别很小,10 d 训练期的校正效果

最好。因此,对于 NGR 技术而言,当训练期的时间序列较短时,NGR 技术仍能有效校准 2 m 温度预报,甚至可能取得更好的校准效果,上述结果与前人的研究结果<sup>[13,16]</sup>类似。

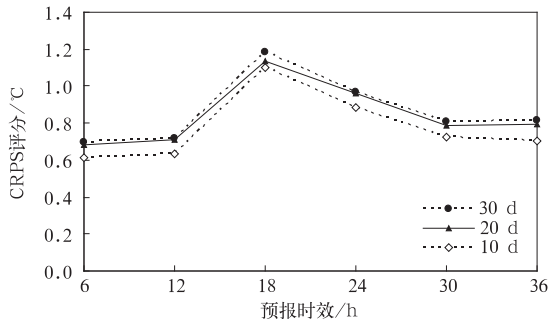


图 7 不同训练期的 2 m 温度预报 CRPS 评分  
Fig. 7 CRPS of 2 m temperature from different training length

#### 4.2 不同离散度因子 $f$ 对校准结果的影响

由第 2 章可知,离散度因子  $f$  决定了校准后的预报概率密度函数的形状,因此,本文开展了利用不同离散度因子校准预报的敏感试验。图 8 是分别采用  $f$  等于  $1/2$ ,  $2/3$ ,  $3/4$  和 1 调整集合离散度后的预报结果的 CRPS 评分。对比发现,当  $f=1/2$  时,CRPS 评分相对偏大,而其他取值 CRPS 评分相差很小,因此本文选取离散度因子  $f=1$ 。

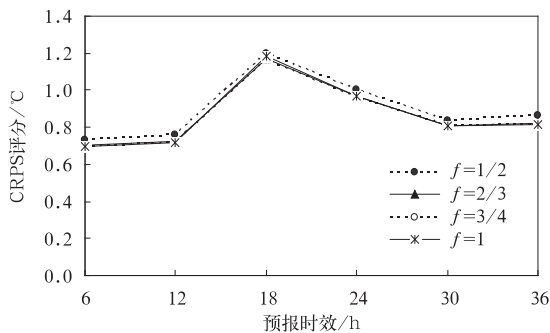


图 8 不同离散度因子的 2 m 温度预报 CRPS 评分  
Fig. 8 CRPS of 2 m temperature from different spread rescaling

### 5 NGR 技术与自适应误差订正技术的对比试验

上述结果证明了 NGR 技术能够显著校准区域集合预报系统 2 m 温度预报的系统偏差和提高集合离散度。在此,将通过 NGR 技术与自适应误差订正技术的对比试验,进一步说明该方法的适用性。

为了便于直接比较校准结果,自适应误差订正

技术也将以相同的区域集合预报结果和观测资料为研究资料。校准过程中,对修正系数实行冷启动<sup>[8]</sup>,相应地使用  $w=3.3\%$  为权重系数。结果表明,自适应误差订正技术校准一阶矩系统误差效果明显,但对二阶离散度校准作用甚微(图略),这与 Cui 等<sup>[9]</sup>的研究结果一致。因此下面主要对两种技术校准系统误差的效果进行比较。图 9 给出了原模式直接输出及两种校准技术校准后的集合平均误差变化状况。由图 9 可以看出,NGR 技术和自适应误差订正技术分别在不同程度校准了 2 m 温度预报的系统误差,但 NGR 校正效果更好。从 BS 评分(图 10)、CRPS 评分(图 11)及 ROC 面积(图 12)各项检验指标可以看出,相对于自适应误差订正技术而言,NGR

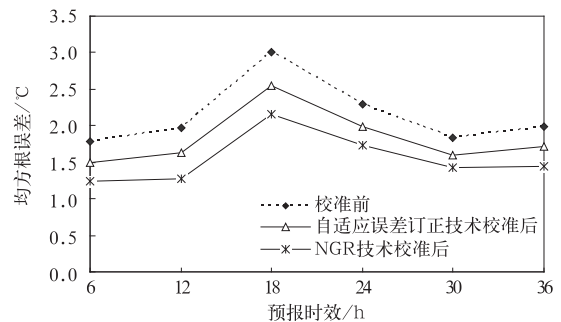


图 9 2 m 温度预报的均方根误差  
Fig. 9 RMSE of 2 m temperature

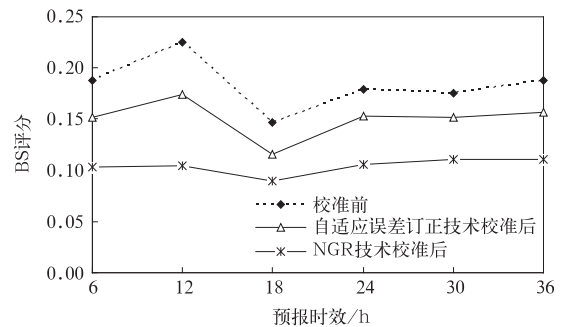


图 10 2 m 温度预报的 BS 评分  
Fig. 10 BS of 2 m temperature

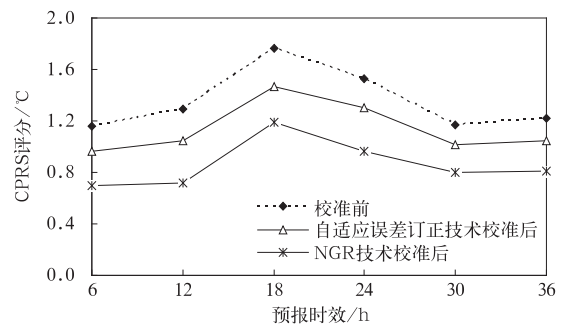


图 11 2 m 温度预报的 CRPS 评分  
Fig. 11 CRPS of 2 m temperature

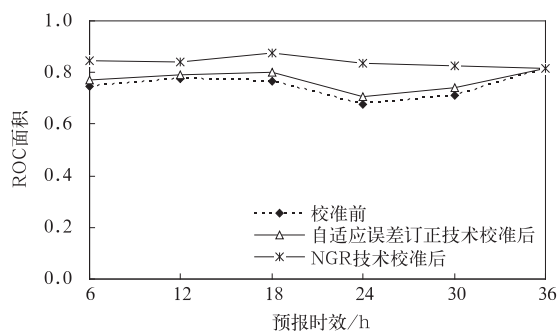


图 12 2 m 温度预报 ROC 面积

Fig. 12 Area under ROC curves  
for 2 m temperature

技术能更加显著地提高 2 m 温度预报技巧。

## 6 小 结

本文基于 NGR 技术开展了区域集合预报系统 2 m 温度预报的后处理校准技术研究,实现了对其一阶偏差和二阶离散度的校准,并初步开展了 NGR 技术与自适应误差订正技术的比较试验。通过对校准前后 2 m 温度预报较全面的检验分析,得到以下主要结论:

1) NGR 技术对 2 m 温度预报的集合离散度和均方根误差校准效果显著,校准后的均方根误差减小,集合离散度增大,两者之间的离散度-预报技巧关系显著提高。Talagrand 直方图的检验结果表明,NGR 技术能有效改进 2 m 温度预报离散度不足的缺陷,并且显著减小了原 Talagrand 直方图中的暖偏差现象和第 9 个排序区间的明显偏差,使得校准后的集合成员分布更均匀。

2) 在相同的温度阈值下,BS 评分、ROC 面积和 CRPS 评分的检验结果均表明,经 NGR 技术校准后的 2 m 温度的预报能力显著提高,其中包括了对特定阈值的鉴别能力。

3) 不同训练期的校准结果对比试验表明,时间序列较短时,NGR 技术对原模式预报校准效果仍然显著。而目前的校准技术大多都是训练期越长校准效果越明显,因此 NGR 技术校准效果受训练数据影响小的特点使其更具有普遍适用性。

4) NGR 技术与自适应误差订正技术的对比试验表明,NGR 技术对集合平均误差的校准效果更显著。同时,BS 评分和 CRPS 评分等也表明,NGR 技术要优于自适应误差订正技术。

## 参 考 文 献

- [1] Carter G M, Dallavalle J P, Glahn H R. Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea Forecasting*, 1989, 4: 401-412.
- [2] Vislocky R L, Fritsch J M. Performance of an advanced MOS system in the 1996-97 national collegiate weather forecasting contest. *Bull Amer Meteor Soc*, 1997, 78:2851-2857.
- [3] Raftery A E, Gneiting T, Balabdaoui F, et al. Using Bayesian model averaging to calibrate forecast ensembles. *Mon Wea Rev*, 2005, 133:1155-1174.
- [4] Gneiting T, Raftery A E, Westveld A H, et al. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon Wea Rev*, 2005, 133:1098-1118.
- [5] Hamill T M, Whitaker J S, Wei X. Ensemble reforecasting: Improving medium range forecast skill using retrospective forecasts. *Mon Wea Rev*, 2004, 132:1434-1447.
- [6] Hamill T M, Whitaker J S. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon Wea Rev*, 2006, 134: 3209-3229.
- [7] Hamill T M, Whitaker J S. Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon Wea Rev*, 2007, 135:3273-3280.
- [8] Hamill T M, Whitaker J S, Mullen S L. Reforecasts: An important dataset for improving weather predictions. *Bull Amer Meteor Soc*, 2006, 87:33-46.
- [9] Cui B, Toth Z, Zhu Y, et al. The Trade-off in Bias Correction Between Using the Latest Analysis/Modeling System with a Short, Versus an Older System with a Long Archive. Proc First THORPEX Int Science Symp, Montreal, Canada, World Meteorological Organization, 2006;281-284.
- [10] 林春泽,智协飞,韩艳,等. 基于 TIGGE 资料的地面气温多模式超级集合预报研究. *应用气象学报*, 2009, 20(6):706-712.
- [11] 马清,龚建东,李莉,等. 超级集合预报的误差订正与集成研究. *气象*, 2008, 34(3):42-48.
- [12] Kunii M, Saito K, Seko H, et al. Verifications and intercomparisons of mesoscale ensemble prediction systems in B08RDP. *Tellus*, 2011, 63A: 531-549.
- [13] Alexander Kann, Wittmann Christoph, Wang Yong. Calibrating 2 m temperature of limited-area ensemble forecasts using high-resolution analysis. *Mon Wea Rev*, 2009, 137:3373-3387.
- [14] 邓国,龚建东,邓莲堂,等. 国家级区域集合预报系统研发和性能检验. *应用气象学报*, 2010, 21(5):513-523.
- [15] Nelder J A, Mead R. A simplex method for function minimization. *Comput J*, 1965, 7:308-313.
- [16] Hagedorn R, Hamill T M, Whitaker J S. Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Temperature. *Mon Wea Rev*, 2008, 136: 2608-2619.
- [17] 皇甫雷官. 国家气象中心集合数值预报检验评价. *应用气象学报*, 2002, 13(1):29-36.

## Calibrating 2 m Temperature Forecast for the Regional Ensemble Prediction System at NMC

Wang Min<sup>1)2)</sup> Li Xiaoli<sup>2)</sup> Fan Guangzhou<sup>1)</sup> Li Zechun<sup>3)</sup>

<sup>1)</sup> (*Chengdu University of Information Technology, Chengdu 610225*)

<sup>2)</sup> (*CMA Numerical Prediction Center, Beijing 100081*)

<sup>3)</sup> (*National Meteorological Center, Beijing 100081*)

### Abstract

It's known that ensemble forecasts provide a flow-dependent sample of the probability distribution of possible future atmospheric states instead of the single and deterministic prediction. Ideally, the probability of any event could be skillfully estimated directly from the relative event frequency in the ensemble. Unfortunately, although the quality of ensemble prediction systems (EPS) has been improved greatly, the direct output of EPS generally is subject to the systematic deficiencies, especially for surface variables. They are under-dispersive and lack of reliability. Therefore, statistical post-processing methods have been developed to improve direct model output. The nonhomogeneous Gaussian regression (NGR) is used to calibrate 2 m temperature forecast of the regional EPS at NMC/CMA. The NGR is the statistical correction method with the first and the second moment (mean bias and dispersion) for Gaussian-distributed continuous variable. This method is based on the multiple linear regression technique and provides a predictive probability density function (PDF) in terms of the normal distribution. The method of minimum continuous ranked probability score (CRPS) estimation is used to fit the regression coefficients of PDF. It can be found that NGR method can greatly improve 2 m temperature forecast compared with the raw ensemble output, and the improvement is as follows: The mean bias is reduced and the spread of ensemble members is increased reasonably; the L-shaped Talagrand diagram of the direct ensemble output has been improved and the calibration reduces the number of outliers, especially in the 9th bin; the probabilistic scores (the brier score, continuous ranked probability score, area under relative operating characteristic curves) all show the significant forecast skill improvement in the calibrated forecasts.

In addition, the sensitive study is performed to investigate the effect of the training length, and the results show that the training length plays a minor role, at least for the chosen verification period. Finally, the comparison by using the time-decaying average bias correction method and NGR is performed, showing that NGR not only has advantages in reducing ensemble mean bias and increasing ensemble spread, but improves the forecast skill in terms of probabilistic scores.

**Key words:** ensemble forecast; bias correction; nonhomogeneous Gaussian regression; probabilistic scores