KNN方法在11—3月中国近海测站日最大风速预报中的应用

涂小萍[1,2] 赵声蓉[1] 曾晓青[3] 刘还珠[1]

(1.国家气象中心,北京 100081; 2.宁波市气象台; 3.兰州大学大气科学学院)

提 要:应用自组织神经网络方法对欧洲中心(ECMWF)2003年1月1日至2006年12月31日逐日数值预报产品分析场进行天气形势分型,发现11—3月影响我国的天气形势基本属于同一类型。对2004—2007年11—3月ECMWF逐日数值预报产品进行动力诊断,提取与中国近海16个测站日最大风速相关较好的预报因子,将改进后的KNN方法作为预报手段,建立11—3月近海测站日最大风速预报模型,并对2007年1—3月16个测站进行逐日检验,结果表明该方法对近海测站日最大风速有较好的预报能力。

关键词:KNN 近海测站 日最大风速预报 交叉验证

Application of an Updated KNN Method to Daily Maximum Wind Forecast for Coastal Weather Station from November to March

Tu Xiaoping[1,2] Zhao Shengrong[1] Zeng Xiaoqing[3] Liu Huanzhu[1]

(1. National Meteorological Center, Beijing 100081;
2. Ningbo Meteorological Observatory; 3. Atmospheric Science College)

Abstract: Self-organizing neural network method is applied to classify weather patterns based on daily NWPs of ECMWF from Jan. 1,2003 to Dec. 31,2006. It shows that the weather pattern is similar over China from November to March. Dynamic diagnosis is applied to daily NWPs of ECMWF in November to March in year 2004－2007 to pick up predictors which have good correlation coefficients with daily maximum wind speed at 16 coastal weather stations. An updated KNN method is used to set up wind speed forecast models for November to March. Daily wind speed forecast for January to March of 2007 is carried out. Results show that KNN method is of good ability in daily maximum wind forecast.

Key Words: KNN method coastal weather stations maximum daily wind speed forecast cross verification

收稿日期:2008年1月24日; 修定稿日期:2008年3月12日

引 言

提高数值预报模式时空分辨率有助于提高客观预报水平[1-2]，但发展数值预报产品的解释应用技术也同样必要。目前数值预报的解释应用技术已在常规气象要素预报中取得了较好的效果[3]，但是对于风速、降水等要素预报效果还不是十分理想。这一方面是由于风速、降水等要素本身非连续和非正态分布的特点所致，另一方面也由于导致这些天气现象的因素具有复杂多变和局域性强的特点。早在1990年代中期，范淦清[4]将风向风速进行预先处理，利用数值预报产品使用MOS方法制作江苏省各站点风的预报，为定点定量的风向风速预报作了很好的尝试。而后不少气象预报工作者用MOS[5]、神经元网络[6-7]等方法直接建立风向风速预报模型或分别建立U、V分量预报。

近年来有人应用相似方法来制作风预报。毛卫星等[8]用波谱分析相似法、邵明轩等[9]用K邻近域方法制作全国600多站点的风向风速预报，取得了较好的效果。最近，曾晓青等采用改进的KNN(K-nearest neighbor)方法[10]来解决降水这类非连续性气象要素的预报问题，该方法在搜索K邻近域的过程中，考虑天气事件出现的概率不同，分别求取有天气事件的正样本K^+值和无天气事件的负样本K^-值，使该方法选择的最邻近域中的K值取得更为合理。本文尝试将此方法应用到中国近海测站的日最大风速客观预报中。

1 资料与处理

以我国近海测站日最大风速作为预报对象，用每天接收的ECMWF预报产品作为预报基本因子。所用资料来源于国家气象中心的MICAPS系统。

1.1 资料的分型处理

不同风速是不同天气形势产生的结果。首先应用自组织神经网络方法进行天气分型分析[11]。参与分型的资料为2003年1月1日至2006年12月31日逐日ECMWF数值产品的分析场(12UTC)，共计1389个有效样本。要素场包括海平面气压场、500hPa高度场、200hPa和850hPa风场，分型范围：100～150°E、10～55°N。分型结果表明：11—3月的样本基本被分为同一类型。该类型平均环流形势基本特点是：500hPa东亚大槽槽底伸展到30°N附近，地面上则表现为25°N以北的中国大陆受到高压控制。这是我国冬半年的天气形势。自组织神经网络分析表明：冬半年影响我国的天气形势与其他季节不同，这一时段渤海、黄海、东海海域地面处于高压环流控制下，盛行偏北风，台湾海峡及以南海域盛行东北风。

1.2 预报对象的处理

预报对象是测站日最大风速。通常日最大风速是不能实时得到的，以MICAPS资料中逐日8个时次地面观测风速的最大值代替。有必要对不同区域近海测站的逐日最大风速与常规8个时次地面观测风的最大值作一比较。表1是2003—2006年11—3月共605天逐日8个时次地面观测风的最大值与实际逐日最大风速的差异。

表1 11—3月日最大风速与8时次地面观测最大风速比较

海区	平均最大风速/m·s^{-1}	平均差值/m·s^{-1}	差值<2.0 m·s^{-1}比例/%	差值≥4.0 m·s^{-1}比例/%
54区	7.0	1.3	74.9	3.4
58区	8.2	1.3	74.5	4.5
59区	5.6	0.5	97.8	0.1

表中可见：54 区和 58 区 11—3 月地面观测逐日 8 时次风速的最大值平均比日最大风速小 1.3m·s^{-1}左右，二者差值小于 2.0m·s^{-1}的天数达到 75%，差值≥4.0 m·s^{-1}的天数仅占 4%左右。而 59 区两者平均差异仅 0.5m·s^{-1}，且差值小于 2.0m·s^{-1}的天数达到 98%，因此 8 时次地面观测风速的最大值可以近似地代表日最大风速。

这里预报对象都以逐日地面 8 个时次观测风速的最大值代表。将预报对象临界值分为 18、15、12、10、8 和 6m·s^{-1}共 6 个不同级别，并对每个级别的预报对象进行 0、1 化处理。共挑选了 16 个预报站点(表 2)，其中 54 区和 58 区各 5 个测站，59 区 6 个测站。

表 2　预报站点

站号	站名	所属省份	站号	站名	所属省份
54776	成山头	山东	58569	石浦	浙江
54646	平台	天津	58760	洞头	浙江
54751	长岛	山东	59792	东沙岛	海南
54579	长海	辽宁	59559	恒春	台湾
54945	日照	山东	59567	兰屿	台湾
58472	嵊泗	浙江	59981	西沙	海南
58666	大陈岛	浙江	59985	珊瑚岛	海南
58974	彭佳屿	台湾	59997	南沙岛	海南

1.3　预报因子的处理

以欧洲中心 2004—2007 年 11—3 月逐日数值预报产品作为基本因子资料，其中 2004—2006 年资料用于建模，2007 年 1—3 月资料用于预报检验。数值预报产品包括 5 层(海平面、850hPa、700hPa、500hPa、200hPa)8 个时效(00、24、48、72、96、120、144、168 小时)5 个基本气象要素(温度、高度、纬向风、经向风、海平面气压)。利用这些基本气象要素通过动力诊断得出如涡度、散度、位温、垂直厚度以及一些物理量的平流、水平梯度等与风场相关的 71 个扩展物理量因子，用双线性插值方法将这些基本要素和扩充物理量插值到预报站点上，建成站点因子库。

建模时直接选取与测站日最大风速相关系数≥0.26 的因子。如果因子总个数＜10，则逐渐降低相关系数，以保证入选因子总数至少为 10 个。建模前对因子做了归一化处理。

2　KNN 方法及参数的确定

KNN(K-nearest neighbor)非参数估计技术[12-13]是近几年来在数值预报释用中发展较快的一种求解问题技术。在天气预报中，KNN 方法集天气学预报思路和数值预报结果为一体，避开了建立预报方程需要作的种种假设。它基于历史样本建立模型，认为相似条件下发生的“行为"会产生相似的结果，因此对于风的预报是合理的。

KNN 技术通过计算待预报数据样本 X''_j 与历史数据样本中对应的每个子样本 X'_{ij} 的距离，这里采用欧式距离作为相似判据：

$$Dist(X'',X'_i)_j=\sqrt{\sum_{j=1}^{m}(X''_j-X'_{ij})^2} \tag{1}$$

其中 $D_i\in R$，$i=1,2,\ldots,n$；这样 n 个样本可得到 n 个距离。在所有距离中选择第 k 个最小的距离作为待预报数据的判断标准：

$$DistK=\mathrm{Min}(Dist)_k \tag{2}$$

通过统计训练样本中小于判别距离 $DistK$ 的个数，把预测数据集的类别归到其中个数较多的一类中，从而做出预报。

KNN 技术中的关键之一是 K 值的确定。在过去使用中，通常不考虑预报对象出现与不出现的样本数多寡，而是在所有的历史样本中寻找 K 个最优近邻。事实上，不同海域、不同等级的风速出现概率是很不相同的。考虑到样本数悬殊，本文应用 KNN 客观方法时，对历史样本中风速出现(为正样本)和不出现(为负样本)两种情况分别考虑，以确定各自的 K 值，记为 K^+、K^-。

$$K^+=\frac{N^-}{N^++N^-}K,K^-=\frac{N^+}{N^++N^-}K \quad (3)$$

根据文献[10],K^+、K^-的选择是利用交叉验证方法。在参与建模的样本中,取一部分样本作为预报测试集,剩余部分作为训练集,对预报测试结果进行评估,通过不断的交叉更换预报测试样本,选择模型预测样本中的准确率和正样本的概括率都达到相对最优组合所对应的 K^+、K^-作为最佳选择。

$$\text{Save}K^+=K^+[\text{Min}((1-\text{准确率})+(1-\text{正样本的概括率}))] \quad (4)$$

$$\text{Save}K^-=K^-[\text{Min}((1-\text{准确率})+(1-\text{正样本的概括率}))] \quad (5)$$

其中

$$\text{准确率}=\frac{\text{预报正确的样本}}{\text{所有样本数}} \quad (6)$$

$$\text{正样本的概括率}=\frac{\text{预报正确的正样本数}}{\text{所有正样本数}} \quad (7)$$

实际预报中,将某站点实时预报因子,依据上述确定的 K^+,K^-值,从历史样本中选取最邻近域,K^+和 K^-分别对应不同的距离($DistK^+$,$DistK^-$),通过统计小于 $DistK^+$ 距离的正样本数和小于 $DistK^-$ 的负样本数,而后用预报判别值给出预报结论。

预报判别值 = 小于 $DistK^+$ 距离的样本数 / 小于 $DistK^+$ 距离的正样本数 + 小于 $DistK^-$ 负样本数 (8)

预报判别的阈值是通过历史资料的试预报,经比较判断给出。当计算出来的预报判别大于该给定阈值时,则认为有该类天气事件发生,反之则无。

上述 KNN 方法实现步骤见图 1。

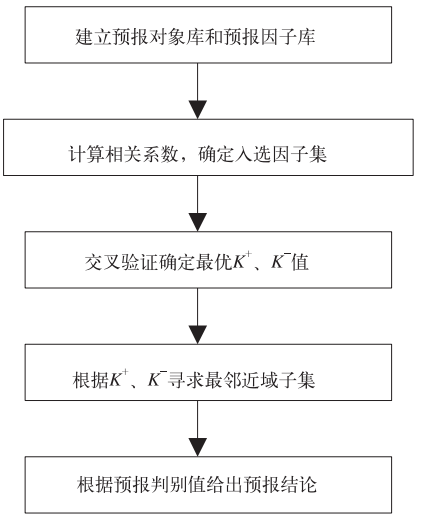

图 1 KNN 方法实现流程

3 结果分析

通过上述改进的 KNN 方法分别建立了11—3月中国16个近海测站不同预报时效的日最大风速预报模型,利用 2007 年 1—3 月 ECMWF 数值预报产品进行了逐日 24～168 小时不同风速级别、不同预报判别阈值下的预报检验,分别统计 TS 评分、空报率、漏报率和概括率。

3.1 区域预报评分

结果表明当临界风速≥12m·s^{-1}时,KNN 方法所建模型仅对部分测站有预报能力,而临界风速≤10m·s^{-1}时所建模型对各测站都有预报能力,因此区域 TS 评分分析仅针对≤10 m·s^{-1}临界风速进行。

图 2 为 54 区 2007 年 1—3 月逐日 10、8、6m·s^{-1} 3 个不同临界风速 24～168 小时预报结果的 TS 评分、空报率和漏报率。分析发现:TS 评分随着临界风速值的降低有整体提高的趋势,且不随预报时效的延长而下降。当临界风速为 10m·s^{-1}时,24～168 小时的 TS 评分为 0.3～0.428,当临界风速减小到 6m·s^{-1}时,TS 评分提高到 0.468～0.572。就空报率和漏报率分析,临界风速偏大时空报率相对较高,24～168 小时预报基本在 0.4 以上,而漏报率一般在 0.20 左右,可见临界风速偏大时空报率是影响 TS 评分的主要原

因。随着临界风速逐渐降低,漏报率对 TS 评分的影响逐渐增大,甚至可能超过空报率的影响。

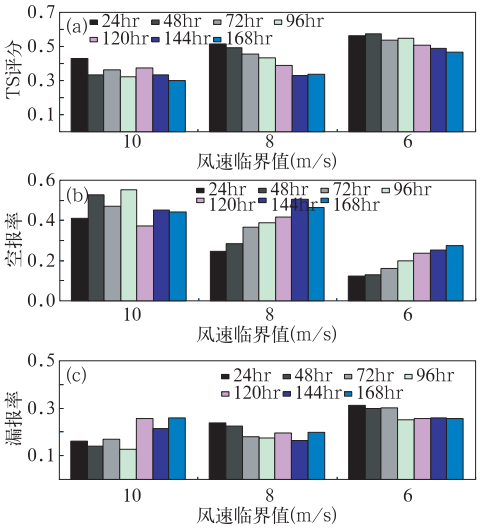

图 2　54 区 24～168 小时预报 TS 评分(a)、空报率(b)、漏报率(c)

与 54 区相比,58 区 TS 评分与之持平,而 59 区的 TS 评分则略低于 54 区和 58 区。就空漏报率分析,当临界风速相对偏大时(8m·s^{-1}以上),对于 58 区来说,空漏报率相当,在 0.2～0.35 之间,二者对 TS 评分几乎有相同的影响。随着临界风速值减小,空报率下降到0.1以下,漏报率仍维持在0.2～0.3 之间,此时漏报率的影响大于空报率。而对于 59 区来说,空漏报率对 TS 评分的影响表现接近于 54 区。

可见 3 个区域的 TS 评分随临界风速的减小都是提高的。在不同临界风速时空漏报率对 TS 评分的影响是不同的。当临界风速相对大(10m·s^{-1})时,54 区和 59 区空报率的影响大于漏报率,58 区则二者相当,但临界风速较小(6m·s^{-1})时,3 个区域的漏报率影响都大于空报率。

3 个区域各测站 TS 评分分析还表明:无论在哪个风速等级,哪个预报时效,54776 站表现都是 54 区是最好的,紧随其后的是 54646 和 54751。而在 58 区各风速等级 TS 评分都大于区域平均值的是 58472、58666 和 58974。59 区相对表现好的测站有 4 个:59792、59559、59567 和 59985。

按照浙江省业务评分标准,浙江沿海日最大风力从 10.8m·s^{-1}(6 级)起评(表 3)。宁波市气象台沿海海面 1—3 月逐日最大风速 24 小时主观预报 TS 评分 2007 年和 2006 年分别为 0.42 和 0.43,而 58 区所建模型 2007 年 1—3 月 10m·s^{-1}临界风速 24 小时预报 TS 评分为 0.47。虽然评分标准和起评风速有差异,但还是表明模型有较好的客观预报能力。

表 3　浙江省海面风力质量评定标准

预报	实况															
	≤4 级			5 级			6 级			7 级			8 级			≥9 级		
≤5 级	不评			不评			F	L	0	F	L	0	F	L	0	F	L	0
5～6 级	F	—	0	不评			F	+	70	F	—	0	F	—	0	F	—	0
6 级	F	K	0	F	K	70	F	+	100	F	+	40	F	—	0	F	—	0
6～7 级	F	K	0	F	K	40	F	+	100	F	+	70	F	—	0	F	—	0
7 级	F	K	0	F	K	0	F	+	70	F	+	100	F	+	40	F	—	0
7～8 级	F	K	0	F	K	0	F	+	40	F	+	100	F	+	70	F	—	0
8 级	F	K	0	F	K	0	F	—	0	F	+	70	F	+	100	F	+	40
8～9 级	F	K	0	F	K	0	F	—	0	F	+	40	F	+	100	F	+	70
≥9 级	F	K	0	F	K	0	F	—	0	F	—	0	F	+	70	F	+	100

F:预报,K:空报,L:漏报,+:预报正确,−:预报错

3.2 临界风速≥12m·s^{-1}的站点预报评分

当临界风速为 18m·s^{-1}时，模型仅对 4 个站(54776、54646、58666、59567)具有预报能力。从 24 小时 TS 评分看，54776 和 54646 站较另外 2 个站好，TS 评分超过了 0.4，而 58666 和 59567 都在 0.2 以下。当临界风速从 15m·s^{-1}减小到 12m·s^{-1}时，各站 TS 评分总体趋势是提高的(图 3)，其中 54776 和 54646 站的 TS 评分仍是较好的，24 小时预报 TS 评分能达到 0.5 以上，58666 站和 59567 站 TS 评分随着临界风速的减小而提高，由 18m·s^{-1}时不足 0.2 提高到 12m·s^{-1}的 0.5 左右。

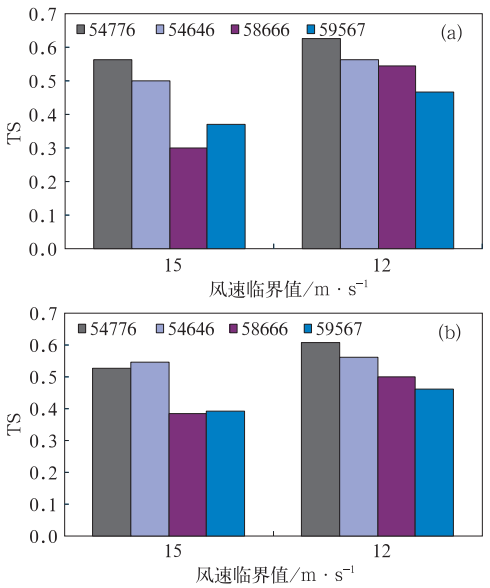

图 3 较大临界风速 24 小时(a)、48 小时(b)预报 TS 评分

与区域 TS 评分有所不同的是，对于临界风速≥12m·s^{-1}的站点，TS 评分随预报时效延长是减小的。在 24～72 小时预报时效内，15m·s^{-1}各站 TS 评分都在 0.3～0.5，96 小时降到 0.25～0.35，120 小时以后则都不到 0.3，而 12m·s^{-1}各站 24～96 小时的 TS 评分变化不大，在 0.4～0.6 之间，120～168 小时则降到 0.3～0.5。

3.3 入选因子分析

模型建立时是根据相关系数大小自动筛选因子。分析 54 区、58 区和 59 区中 TS 评分相对好的 3 个测站(54776、58666、59792)24 小时预报模型入选因子可以看出，3 个站入选的相同因子有 7 个(地面气压的水平梯度、850hPa 温度，x 方向的地面气压水平梯度、850hPa V 偏差风，850hPa 经向风和 850hPa 风速)，这些因子都与风直接相关。54776 和 58666 站入选因子基本均匀分布在从地面到 500hPa 各个层次，而 59792 则集中在 700hPa 及其以下的中低层次，但 3 个站对日最大风速有影响的共同因子都集中在 850hPa 及以下层次。

从不同预报时效的入选因子看，3 站 24～96 小时入选因子变化不大，只是随着预报时效的延长，相关系数大小逐渐降低，入选因子个数逐渐减少；在 120～168 小时预报时效时，入选因子差异才逐渐增大。表现出入选预报因子较好的稳定性。

对于各区域内 TS 评分相对差的测站(54579、58760、59981)，分析发现其入选因子与 TS 评分较好的测站很不相同，表现在要么入选因子相关系数较低，要么入选因子及所在层次变化大，且入选因子随预报时效的变化明显，表现出预报因子的不确定性很大。

3.4 预报判别阈值讨论

预报判别阈值可以进一步控制预报样本与历史样本相似程度，其大小的选定应针对不同测站在不同的临界风速下选定不同的值。预报试验中，分别选取从 0.5～0.8 之间每间隔 0.05 的预报判别阈值在不同测站不同预报时效时的 TS 评分，发现不同临界风速时要使 TS 评分达到最大，则相应的预报判别阈值设定是有所不同的。如 54776 站，当临界风速为 18m·s^{-1}时，试报结果表明预报判别阈值≥0.70 时 TS 评分能达到较好的值，而对于 15m·s^{-1}的临界风速，则最小预

报判别阈值至少设定在 0.60。总之,针对不同测站不同等级的风速预报,其预报判别阈值应当有所不同。分析发现:如果设定的临界风速对于测站来说发生概率相对小,则可以适当提高预报判别阈值,以减少空报率,提高 TS 评分。

4　结论与讨论

本文将 KNN 技术应用到近海测站日最大风速预报时,对不同测站、不同风速等级、不同判别阈值进行了试报。结果表明:

(1) 不同海区入选因子层次分布是不同的,但预报效果较好的站点所选因子基本符合预报员对产生风速影响因素的认识,这些关系应当是稳定的。而预报效果较差的站点,入选因子就比较乱,可以认为他们的关系不够稳定,还有待积累更多的样本资料给予进一步考察。

(2) 临界风速$\leqslant$10m·s^{-1}的区域预报效果分析表明:3 个区域 TS 评分随着临界风速的减小有升高趋势,但 TS 评分随预报时效变化不大。当临界风速相对大(10m·s^{-1})时,54 区和 59 区空报率的影响大于漏报率,58 区则二者相当,随着临界风速减小到 6 m·s^{-1}时,3 个区域的漏报率影响都大于空报率。与主观预报 TS 评分相比,模型表现出较好的客观预报能力。

(3) 临界风速$\geqslant$12m·s^{-1}时模型仅对部分站点有预报能力,站点 TS 评分随着预报时效的增加是减小的。

(4) 无论临界风速等级大小,模型对 54776、54646、58666 和 59567 的预报效果始终是各区域最好的,可以作为日常预报中多加参考的测站。

(5) 当风速预报等级对于测站相对发生概率较小,可以适当预报判别阈值,以控制空报率,提高 TS 评分。

(6) 11—3 月影响我国的天气形势类型相似,因此预报对象为近海测站日最大风速,没有考虑风向的问题。

本文所建预报模型还有待完善。如建模时资料还不是足够长(ECMWF 数值产品中的流场资料从 2004 年 10 月 1 日开始才比较全面),因子筛选完全根据相关系数大小自动筛选,入选因子可能并不相互独立。随着资料长度的累加,在今后的改进工作中,有必要进一步加强研究和试验,使筛选的因子更加合理完善,提高 KNN 方法的预报效果。

参考文献

[1] 张建海,王国强.客观预报中多时刻因子的应用及其效果[J].气象,2005,31(5),62-65.

[2] 龚强,袁国恩,汪宏宇,等.应用 MM5 模式对地面风速过程的模拟试验[J].气象,2005,31(4):53-57.

[3] 刘还珠,赵声蓉,赵翠光,等.国家气象中心气象要素的客观预报——MOS 系统[J].应用气象学报,2004,4,15(2):181-191.

[4] 范淦清.风预报的数值产品释用研究[J].气象,1995,21(10):47-50.

[5] 林良勋,程正泉,张兵,等.完全预报方法在广东冬半年海面强风业务预报中的应用[J].应用气象学报,2004,5(4):485-490.

[6] 胡波,杜惠良.浙江省沿海海面日极大风预报[J].海洋预报,2006,23(B09):64-67.

[7] 杨忠恩,陈淑琴,黄辉.舟山群岛冬半年灾害性大风的成因与预报[J].应用气象学报,2007,18(1):80-85.

[8] 毛卫星,许晨海,何立富,等.多时次多尺度波谱相似预报风要素[J].气象,2005,31(10):28-31.

[9] 邵明轩,刘还珠,窦以文.用非参数估计技术预报风的研究[J].应用气象学报,2006,17(增刊):125-129.

[10] 曾晓青,邵明轩,刘还珠,等.基于交叉验证技术的 KNN 方法在降水预报中的试验[J].应用气象学报,待发表.

[11] 黄卓,杨洪敏,郝为,等.基于智能聚类的综合相似预报[A].暴雨落区预报实用方法[C].北京:气象出版社,2000:53-59.

[12] 翟宇梅,赵瑞星.概率天气预报的 K 近邻非参数估计仿真模型[J].系统仿真学报,2005,17(4):786-788.

[13] 翟宇梅,赵瑞星,肖仁春,等.K 近邻非参数回归概率预报技术及其应用[J].应用气象学报,2005,16(4):453-460.