赵华生,黄小燕,黄颖. ECMWF 集合预报产品在广西暴雨预报中的释用. 应用气象学报,2018,29(3):344-353.
DOI: 10.11898/1001-7313.20180308

ECMWF 集合预报产品在广西暴雨预报中的释用

赵华生* 黄小燕 黄 颖
(广西壮族自治区气象减灾研究所, 南宁 530022)

摘 要

基于最大相关最小冗余度算法和随机森林回归算法,该文提出一种对欧洲中期天气预报中心(ECMWF)集合预报产品进行暴雨预报的释用方法。该方法采用最大相关最小冗余度算法,对 ECMWF 集合预报的 51 个成员进行筛选,选取若干个与预报对象相关性最大、相互间冗余度最小的成员作为随机森林回归算法的输入因子。利用 ECMWF 集合预报降水量平均值对建模样本进行分类,使预报模型的建模样本更具有针对性。通过 2012 年 4 月—2015 年 12 月的交叉独立样本试验预报和 2016 年 1—9 月的业务预报试验的统计结果表明:该释用方法的暴雨预报 TS 和 ETS 评分,均比采用 ECMWF 集合预报产品 51 个成员降水量预报进行插值后取平均值的释用方法分别提高了 0.07 和 0.05 以上,显示了较好的数值预报产品释用效果。

关键词: 最大相关最小冗余度算法; 随机森林回归算法; 释用

引 言

随着社会的发展,各行各业对气象服务的需求越来越多,要求也越来越高。同时,随着高时空分辨率的数值预报模式在预报业务中的使用,人们希望气象部门能提供更为准确的天气预报产品。然而,受模式误差、输出误差、模式稳定性以及预报员的主观分析等原因影响,如果直接使用数值预报模式中的输出结果进行预报,其准确率相对较低。为了得到客观化和定量化的预报效果,目前,对数值模式产品进行客观订正的释用是解决该问题的主要方法之一。实践证明:通过对数值预报产品的释用,其预报能力比模式的直接输出预报有明显提高[1]。同时,由于暴雨是我国最重要气象灾害之一,开展客观定量的数值预报产品暴雨预报释用方法研究具有重要意义。

到目前为止,很多学者开展了利用数值预报产品进行客观定量降水预报(quantitative precipitation forecast, QPF)的释用方法研究[2-6],包括强降水的客观定量预报释用方法研究,取得了较多的研究成果,并在实际预报业务中得到应用。如 David 等[7-8]利用分级技术从 NCEP 多模式短期集合预报系统中输出较可靠的 3 h 概率定量降水预报(PQPF),并对此进行分析发现分级技术对降水事件能提供更多预报技巧和可靠性;赵声蓉等[9]以多个模式的降水量预报产品作为神经网络的预报因子,利用神经网络方法建立一种数值预报产品释用的定量降水预报模型,预报效果提高明显;唐晓文等[10]根据深厚湿对流系统长时间维持将产生强降水这一配料法的预报原理以及中国不同区域的天气特点,选取对强降水(不小于 25 mm·d^{-1})有显著影响的 4 类因子(水汽因子、动力因子、不稳定因子及热力因子),在一定的物理条件约束下,利用经验和统计相结合的方法建立配料综合指数与强降水之间的关系;Jerome 等[11-12]利用美国国家气象局气象发展研究室开发的高分辨率全球预报系统(Global Forecasting System, GFS)得到美国本土的4 km 格

2017-07-25 收到, 2018-01-29 收到再改稿。
资助项目: 国家自然科学基金项目(41575051,61562008,41765002),广西重点基金项目(2017GXNSFDA198030),广西青年基金项目(2014GXNSFBA118211)
* 邮箱: 2006zhaohuasheng@163.com

距的高分辨率 MOS 定量降水预报;孙靖等[13]利用分等级消除偏差法,并采用混合训练期和 60 d 滑动训练期方案,对 2012 年欧洲中期天气预报中心(ECMWF)数值模式夏季 1—5 d 的降水预报进行订正试验,试验结果表明:该方法对 25 mm·d^{-1}以上降水预报的 ETS 评分提高明显。

从上述国内外研究现状可以看到,基于数值预报产品释用的客观定量降水预报方法是目前短期客观定量预报的重要发展方向。然而,目前国内外还没有一种成熟、有效并得到预报业务技术人员普遍认同的数值预报产品释用方法。为此,本文尝试提出一种基于最大相关最小冗余度(maximum relevance minimum redundancy,MRMR)算法和随机森林回归算法(random forest regression,RFR)相结合的数值预报产品释用方法。该方法以广西 89 个站点的降水作为预报对象,采用最大相关最小冗余度算法从 ECMWF 集合预报的 51 个成员筛选出若干个成员,这些成员与预报对象具有相关性最大、成员间冗余度最小的优点。将提取出的预报成员作为随机森林回归算法的输入因子进行释用。

1　资料与方法

1.1　资　料

本文所用资料是 ECMWF 集合预报模式的逐日 08:00(北京时,下同)和 20:00 的 24 h 和 48 h 降水量预报产品(格距为 0.5°×0.5°,水平范围为 21°～26.25°N,104.75°～112°E);选取试验样本时段为 2012 年 4 月—2016 年 9 月,共 2176 个样本(去除数据缺失样本),其中 2012 年 4 月—2015 年 12 月为建模样本,2016 年 1—9 月(共 526 个样本)为独立样本进行预报试验。

1.2　方　法

统计预报中,有两方面因素影响其预报效果:一是预报因子的选取,即所选取的模型预报因子与预报对象之间的相关性强弱以及预报因子间的复共线性大小问题;二是预报模型本身的问题,即所选取模型的拟合能力和泛化性能强弱。为此,本文首先采用最大相关最小冗余度(MRMR)算法,从多个预报因子中选取与预报对象相关性大而它们之间的预报信息重叠少(复共线性小)的若干个预报因子作为预报模型的输入。其次,在预报模型的构建上,考虑采用具有泛化能力和抗噪声能力强、训练时间短且不容易陷入过拟合的随机森林回归算法。

本文提出的基于最大相关最小冗余度算法和随机森林回归算法相结合的数值模式预报释用方法主要步骤如下:

① 采用三次多项式插值方法将每个 ECMWF 集合预报成员插值到站点上。

② 利用 ECMWF 集合预报成员插值的平均值,对预报对象的历史样本序列进行分类。

③ 将分类后的预报对象样本对应的因子矩阵采用最大相关和最小冗余度算法,对每个站点的 ECMWF 集合预报 51 个成员进行筛选。

④ 将步骤③选出的预报成员作为随机森林算法的模型输入进行预报建模,并输出预报结果。

1.2.1　最大相关最小冗余度算法

预报因子是统计预报模型的重要组成部分,预报因子的选取直接影响预报效果。本文将 ECMWF 集合预报产品 51 个成员的输出值作为预报模型的输入因子,成员之间的相关性和冗余性会影响预报模型的预报能力,为此本文尝试采用 MRMR 算法对集合预报的成员进行筛选。

MRMR 算法是信息论中典型的特征降维算法[14],其主要思想是从特征空间中寻找与目标类别相关性最大而相互之间冗余性却最少的 m 个特征[15],他们之间的相关性和冗余性利用互信息[16]衡量。互信息用于衡量两个随机变量之间的相互约束程度。对于给定 $\boldsymbol{S}=\{x_i \mid i=1,\cdots,m\}$ 为特征集合和目标类别 $\boldsymbol{Y}$,则特征集合 $\boldsymbol{S}$ 中的特征与目标类别 $\boldsymbol{Y}$ 的最大相关度以及度量特征集合 $\boldsymbol{S}$ 中各特征间的互相关度分别为式(1)和式(2)。

$$\max D(\boldsymbol{S},\boldsymbol{Y}),\ D=\frac{1}{|\boldsymbol{S}|}\sum_{x_i\in\boldsymbol{S}} I(x_i,\boldsymbol{Y}); \quad (1)$$

$$\min R(\boldsymbol{S}),\ R=\frac{1}{|\boldsymbol{S}|^2}\sum_{x_i,x_j\in\boldsymbol{S}} I(x_i,x_j)。 \quad (2)$$

Peng 等[15]定义算子(D,R)结合相关性因子 D 和冗余性因子 R,即将式(1)减式(2)进行组合,从而得到最大相关性最小冗余度准则(MRMR):

$$\max\Phi(D,R)=D(\boldsymbol{S},\boldsymbol{Y})-R(\boldsymbol{S})。 \quad (3)$$

1.2.2　随机森林回归算法

统计预报模型构建方法的选取,对释用效果也有直接影响,本文采用多决策树组合而成的随机森林回归算法(random forest regression,RFR)进行构建预报模型,RFR 算法是由 Breiman 于 2001 年提出的一种非线性统计方法[17],该算法具有抗噪声

强、预报结果稳定的优点。RFR 是利用自举法从原始样本中抽取多个训练样本子集，对每个样本子集分别进行决策树建模。进一步通过组合多棵决策树进行预测，并通过取平均值得到最终预测结果[18]，其本质与气象上的集合预报思想相近，是将多棵决策树建模得到的预测结果进行集成。

随机森林回归算法同时还具有计算速度快、泛化性能好以及参数少等优点，且不容易出现人工神经网络的过拟合现象。目前该方法在农业、水文和医学等众多领域得到广泛应用[19-21]。然而，该方法在降水预测中的应用报道较为少见。为此，本文尝试采用该方法进行数值模式的降水预报产品进行释用。

RFR 算法是通过自举法抽样技术，由随机向量 $\boldsymbol{\theta}_i$（即回归决策树）生长形成 $\{h(\boldsymbol{X},\boldsymbol{\theta}_i),i=1,\cdots,k\}$ 的组合模型。预测变量是一个数值型变量，与其分类模型不同，其预测值是通过 k 棵回归决策树的预测结果取平均值得到的。RFR 算法实现见图 1。

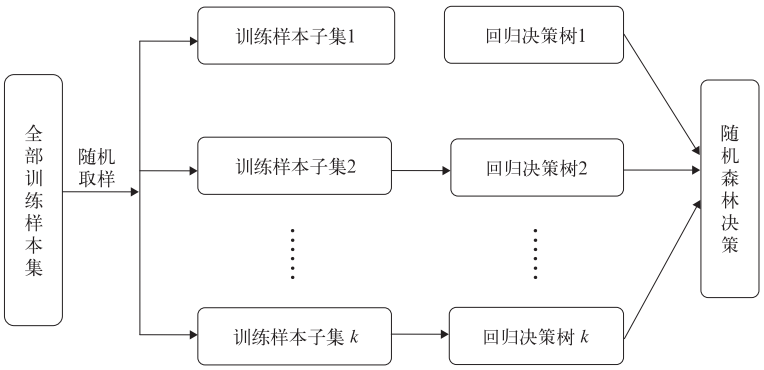


图 1 RFR 算法流程图

Fig. 1 RFR algorithm flow chart

2 建模试验

本文以广西 89 个气象站未来 24 h（每日 2 个预报时次为 08:00 和 20:00）降水量作为预报对象，重点研究暴雨以上量级的降水预报。研究区域和站点分布如图 2 所示。

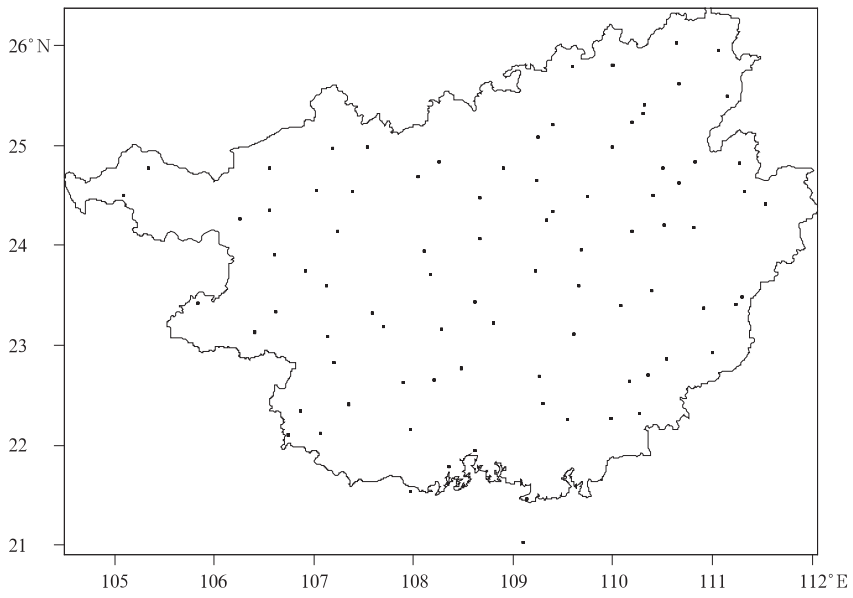


图 2 研究区域和站点分布

Fig. 2 The target area and station distribution

2.1　单站暴雨集合预报建模样本及预报因子处理

经过统计分析,2012 年 4 月—2016 年 9 月共 2176 个样本中(去除资料缺失样本),广西 89 个基本气象站 24 h(统计时段包含当日 08:00—次日 08:00 和前日 20:00—当日 20:00)内出现 1 个站以上、降水量达到暴雨以上的有 4775 个站次。本文的研究对象 24 h 降水量达到暴雨以上量级的情况比较频繁。为提高暴雨预报的准确率,在进行单站暴雨集合预报建模时,对模型的建模样本采取分类处理方式,具体建模步骤如下:

① 对 ECMWF 集合预报的每个成员(共 51 个成员),利用前一日 48 h 累积降水量预报场减去 24 h 累积降水量预报场,从而获得该成员当日的未来 24 h 降水量预报场 $\boldsymbol{R}_{24}$。

② 采用多项式插值方法,将降水量预报场 $\boldsymbol{R}_{24}$ 插值到 89 个气象站。此时,每个预报对象(气象站)都对应有 51 个预报因子(51 个 ECMWF 集合预报成员的插值),记为 $\boldsymbol{F}_{51}$。

③ 对于第 $k(k=1,\cdots,89)$个预报对象 $\boldsymbol{Y}$(气象站),如果该站点的 $\boldsymbol{F}_{51}$平均值大于 α mm,则在预报对象 $\boldsymbol{Y}$ 的历史样本序列(其样本序号记为 $\boldsymbol{\Omega}$)中,选出降水量大于 β mm 的样本序号集 $\boldsymbol{\Omega}_1$($\boldsymbol{\Omega}_1 \subset \boldsymbol{\Omega}$)。若该站点的 $\boldsymbol{F}_{51}$平均值小于 α mm,则令 $\boldsymbol{\Omega}_1=\boldsymbol{\Omega}$。

④ 利用步骤③得到的预报对象 $\boldsymbol{Y}$ 的建模样本序号集 $\boldsymbol{\Omega}_1$,可得到与之对应的由 $\boldsymbol{F}_{51}$组成的因子矩阵 $\boldsymbol{S}'$和预报量序列 $\boldsymbol{Y}'$。

⑤ 将 $\boldsymbol{S}'$和 $\boldsymbol{Y}'$带入 MRMR 算法,求出这 51 个因子中(集合预报成员)与预报量 $\boldsymbol{Y}'$相关性最大和冗余度最小的 n 个因子(成员)。

⑥ 以步骤⑤选出的 $\boldsymbol{S}'$和 $\boldsymbol{Y}'$为基础,采用随机森林回归模型算法建立集合预报模型。

⑦ 将步骤⑤选出的因子(成员)对应的预报样本代入步骤⑥训练好的预报模型中,得到对该站未来 24 h 的降水量预报。

2.2　结果分析

在进行建模试验时,为了能够更好地了解本文提出方法的释用能力,在此采取交叉检验法,即将 2012 年 4 月—2015 年 12 月的样本作为独立样本进行与实际预报相同的独立预报试验,对 2016 年 1—9 月进行业务预报试验。在进行试验计算前,要先确定出 2.1 节计算步骤中的几个阈值参数(n,α,β)。参数 n 用于选取若干个集合预报成员作为模型的预报因子参与最后的训练计算,根据文献[22]的研究结果,参与集合预报计算的成员,其数量控制在 8～10 个比较合适。本文先设定参数 n 为 10,分析另外两个阈值参数 α 和 β 不同取值时对预报模型的影响情况。同时,为了能够将预报结果与数值预报产品进行比较,本文采用常用的多项式插值方法将 ECMWF 集合预报成员的格点资料插值到站点上,并将所有成员插值到站点后取其平均值的预报方法称为 AVI(average value of the interpolation)。表 1 为本文提出的新方法(称为 MRMR-RFR)预报结果与相应预报时段的 ECMWF 集合预报 51 个成员插值到站点后取其平均值(AVI 方法)的 TS 评分,表 1 中 TS 评分越高,表示对应方法预报的准确性越高。考虑到 ETS 评分方法可有效去除随机降水概率对评分的影响,同时也可以避免使用气候概率的情况,本文也对试验预报结果采用 ETS 评分进行对比。

表 1　2012—2015 年暴雨以上量级降水交叉独立预报 TS 评分

Table 1　TS of cross independent sample test forecast of rainstorm from 2012 to 2015

年份	MRMR-RFR						AVI	
	(n=10,α=15,β=10)		(n=10,α=20,β=15)		(n=10,α=25,β=15)			
	TS	ETS	TS	ETS	TS	ETS	TS	ETS
2012	0.12	0.11	0.15	0.12	0.14	0.11	0.07	0.06
2013	0.11	0.10	0.13	0.12	0.13	0.12	0.09	0.08
2014	0.11	0.11	0.17	0.14	0.13	0.12	0.08	0.07
2015	0.12	0.11	0.14	0.12	0.14	0.10	0.06	0.06

对数值模式产品的解释应用,是在承认其具有一定预报能力的基础上进行的。因此,假设当集合预报的成员插值平均达到 15 mm 以上(阈值 α 取 15 mm 以上)时,实况降水量才可能出现暴雨以上降水的情况。同时,由于模式预报结果与实际降水量差异存在,一般设定阈值参数 $\beta<\alpha$。从表 1 的统计结果中可以看到:两种评分方法的统计结果中,ETS 评分整体上略低于 TS 评分,这是因为 ETS 评

分方法对空报和漏报都有惩罚。相比 AVI 方法的 ETS 评分较 TS 评分偏低幅度，MRMR-RFR 释用方法 ETS 评分偏低幅度更大一些，说明 MRMR-RFR 释用方法空报的次数多于 AVI 方法。②本文选取的 3 组阈值参数，其预报结果在暴雨以上量级降水预报(降水量不小于 50 mm)的 TS 和 ETS 评分相差不大，说明这两个阈值参数对其预报能力的敏感性不明显，只要取值在一定范围之内，MRMR-RFR 释用方法 TS 和 ETS 评分均高于 AVI 方法的预报结果，这其中提高幅度最大的是 α 取 20 mm，β 取 15 mm 时，即当集合预报 51 个成员的插值平均预报降水量达到 20 mm 以上时，选取实况降水量大于 15 mm 的样本进行建模，该组参数的预报结果在 2012 年 4 月—2015 年 12 月的交叉独立预报中，其暴雨 TS 评分比 AVI 方法分别提高了 0.08，0.04，0.09 和0.08，而相应的 ETS 评分方法也比 AVI 方法的 ETS 评分分别提高了 0.06，0.04，0.07 和 0.06。

由上述的分析可知，参数 α 和 β 在一定范围内进行取值，MRMR-RFR 方法的预报效果稳定。为了进一步考察参数 n 取值对预报模型稳定性的影响，首先固定参数 α 和 β 的取值(取 TS，ETS 评分最高的参数组合：α=20 mm，β=15 mm)选取不同的 n 值进行上述同样本的试验分析，其中 n 的取值分别尝试取 9 个和 11 个(8～10 个的附近[22])，其统计结果见表 2。

表 2 不同参数 n 下 2012—2015 年暴雨交叉独立预报 TS 评分

Table 2 TS of cross independent sample test forecast of rainstorm under different n from 2012 to 2015

年份	MRMR-RFR					
	(n=9，α=20，β=15)		(n=10，α=20，β=15)		(n=11，α=20，β=15)	
	TS	ETS	TS	ETS	TS	ETS
2012	0.15	0.13	0.15	0.12	0.14	0.12
2013	0.13	0.11	0.13	0.12	0.14	0.13
2014	0.14	0.12	0.17	0.14	0.15	0.13
2015	0.12	0.11	0.14	0.12	0.16	0.14

分析表 2 可知，MRMR-RFR 方法在 3 个不同 n 取值情况下，该释用方法交叉独立样本的试验结果中，3 组参数组合的 TS，ETS 评分互有高低，统计结果相对比较稳定。由此可知，参数 n 在一定范围的取值，MRMR-RFR 释用方法的预报结果也保持稳定。

为了进一步检验模型的预报性能，采用 2012 年 4 月—2015 年 12 月试验结果中 TS，ETS 评分相对较高的参数组合(n=10，α=20，β=15)，对 2016 年 1—9 月进行业务预报试验。预报结果见表 3。

由表 3 可知，本文提出的 MRMR-RFR 释用方法在的逐次独立样本预报检验中，9 个月 TS，ETS 评分全部超过 0 分，并且在出现暴雨站次达到 100 个以上的4—8月，该释用方法的TS，ETS评分均

表 3 2016 年 1—9 月单站暴雨以上量级降水业务预报 TS，ETS 评分

Table 3 TS and ETS of single-station forecast of rainstorm using different methods from Jan 2016 to Sep 2016

月份	MRMR-RFR (n=10，α=20，β=15)		AVI		降水量不小于 50 mm 站次
	TS	ETS	TS	ETS	
1	0.24	0.18	0.00	0.00	60
2	0.50	0.45	0.00	0.00	1
3	0.07	0.05	0.06	0.05	11
4	0.08	0.05	0.03	0.02	105
5	0.17	0.15	0.13	0.11	203
6	0.17	0.14	0.06	0.03	231
7	0.13	0.10	0.01	0.00	114
8	0.24	0.22	0.18	0.15	206
9	0.07	0.04	0.00	0.00	36
1—9	0.15	0.11	0.08	0.06	967

明显高于 AVI 方法。统计可知,MRMR-RFR 释用方法在 5—8 月 TS,ETS 评分均高于 0.10。该释用方法 1—9 月平均的 TS,ETS 评分比 AVI 方法评分分别提高 0.07 和 0.05。

本文选取 2016 年 8 月两次暴雨预报情况进行分析。一次过程选取受 2016 年第 4 号台风妮妲影响而造成的大范围暴雨,具体时段为 2016 年 8 月 2 日 20:00—3 日 20:00,共有 30 个气象站降水量达到 50 mm 以上,图 3 为两种方法预报对比。另一次过程选取 2016 年 8 月 12 日 20:00—13 日 20:00 的非台风类的一般性暴雨强降水过程,共有 27 个气象站降水量达到 50 mm 以上,图 4 为该次过程两种方法预报对比。

由图 3 可以看到,对于 8 月 2 日 20:00—3 日 20:00 预报中,MRMR-RFR 释用方法和 AVI 方法预报降水量大于 50 mm 的落区基本重合,覆盖了全区的大部分地区,然而,MRMR-RFR 释用方法比 AVI 方法在广西西南部和东南部这两个区域更接近实况。

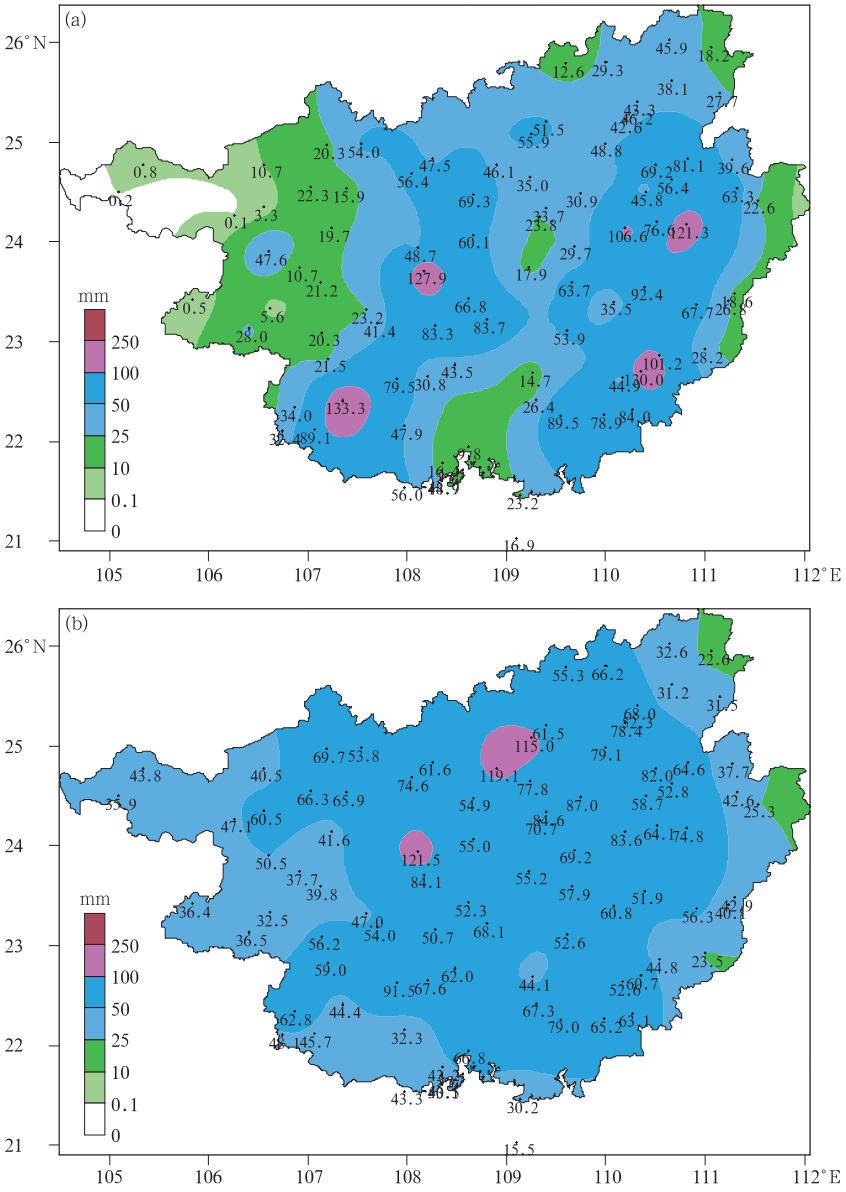
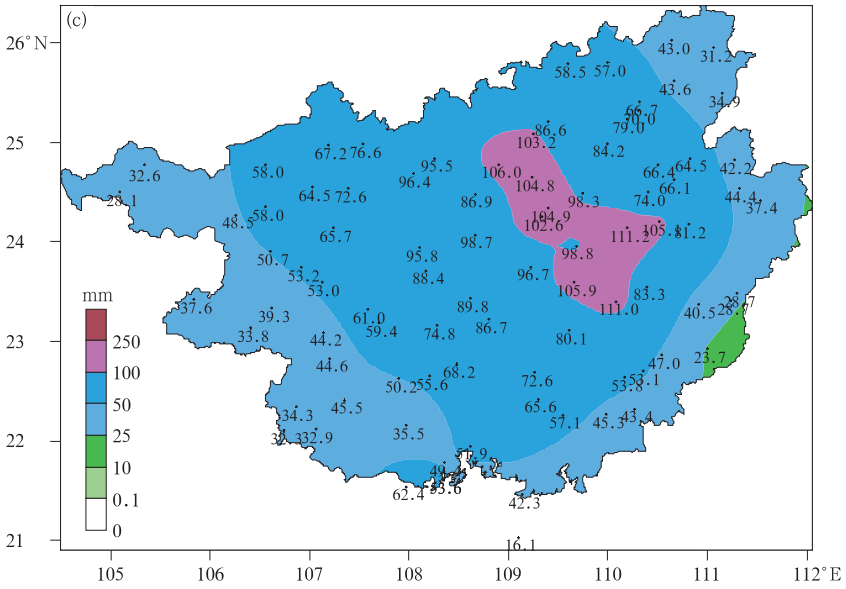


图 3　2016 年 8 月 2 日 20:00—3 日 20:00 24 h 降水实况与预报
(a)实况,(b)MRMR-RFR 释用方法预报,(c)AVI 方法预报

Fig. 3　Observation and prediction of the case from 2000 BT 2 Aug to 2000 BT 3 Aug in 2016
(a)observation,(b)prediction of MRMR-RFR,(c)prediction of AVI



续图 3

由图 4 可知，对于 8 月 12 日 20:00—13 日 20:00 AVI 方法预报出该降水过程的雨带，但所有站降水量均未超过 50 mm。而实况有 27 个站降水量超过 50 mm。MRMR-RFR 释用方法预报出 13 个站降水量超过 50 mm，空报 16 个站，漏报 14 个站，TS 评分为 0.30（ETS 评分为 0.16）。即 MRMR-RFR 释用方法具有正预报技巧。

综上所述，MRMR-RFR 释用方法对 ECMWF 集合预报产品释用，其预报技巧在大部分情况下为正技巧，可以一定程度上提高 ECMWF 集合预报的暴雨预报能力。这与本文采用的因子选取方法、建模样本分类以及预报模型建立的方法有关，通过这些处理，既可以使模型更加专注于强的降水过程，同时又能在预报强降水的建模样本中消除部分降水量很小或较小的样本，减少建模样本中的噪声（干扰）。

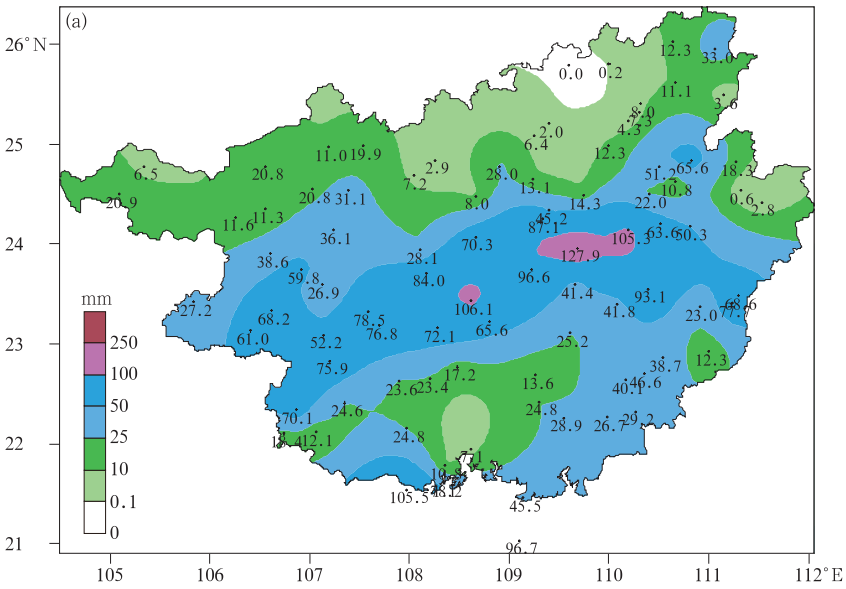
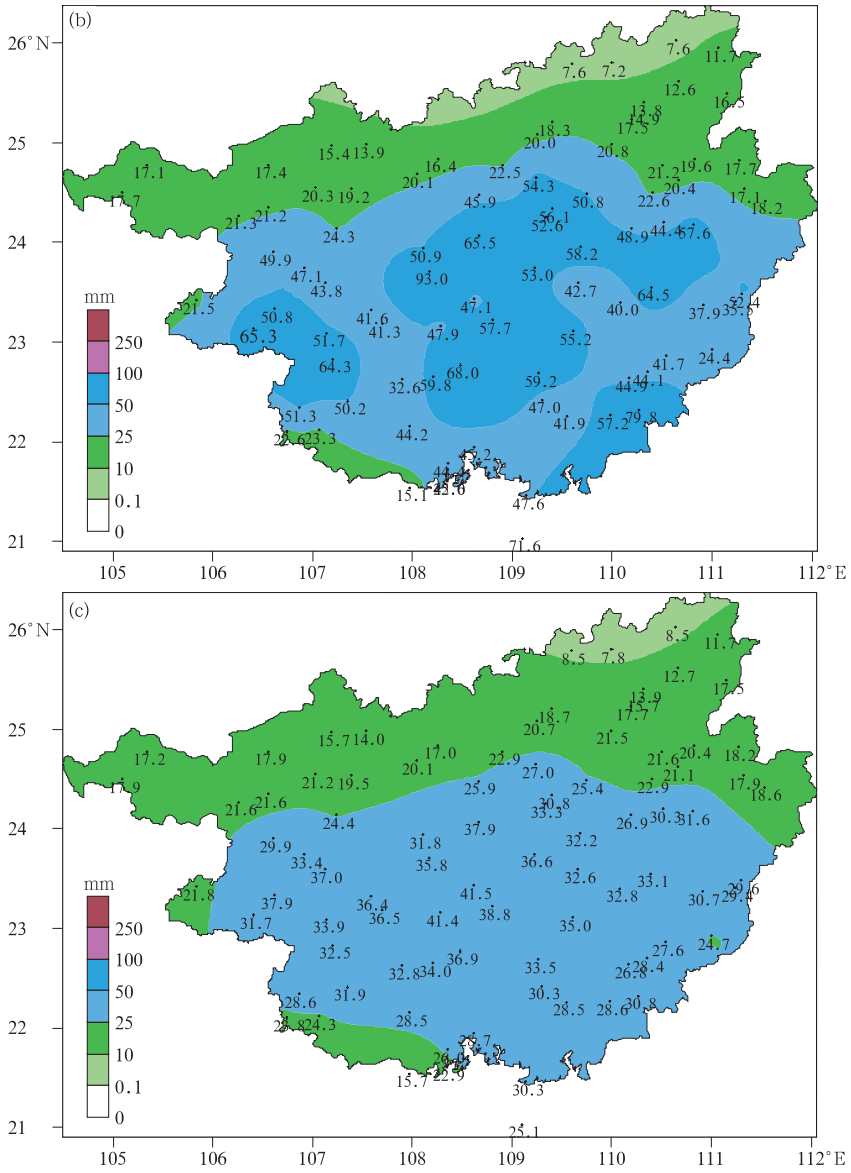


图 4 2016 年 8 月 12 日 20:00—13 日 20:00 24 h 降水实况和预报

(a)实况，(b)MRMR-RFR 释用方法预报，(c)AVI 方法预报

Fig. 4 Observation and prediction of the case from 2000 BT 12 Aug to 2000 BT 13 Aug in 2016

(a)observation, (b)prediction of MRMR-RFR, (c)prediction of AVI



续图 4

3　小　结

1）采用最大相关最小冗余度的互信息技术进行因子选取，可以提高入选因子的预报信息，同时入选因子与预报量具有最大的相关性。构建模型时采用可调参数极少、计算速度快且有很好的非线性拟合能力和泛化性能的随机森林算法。

2）独立样本预报试验表明：本文提出的 MRMR-RFR 释用方法相对于 ECMWF 集合预报产品插值方法（AVI 方法），能更好地判断暴雨的落区及落区范围大小。

本文提出的对数值预报产品的 MRMR-RFR 释用方法较为简单，便于预报人员在业务中使用。

参 考 文 献

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

[2] 毕宝贵，代刊，王毅，等. 定量降水预报技术进展. 应用气象学报，2016，27(5)：534-549.

[3] 岳彩军，寿亦萱，寿绍文，等. 湿 Q 矢量释用技术及其在定量降水预报中的应用. 应用气象学报，2007，18(5)：666-675.

[4] 孔荣，王建捷，梁丰，等. 尺度分解技术在定量降水临近预报检验中的应用. 应用气象学报，2010，21(5)：535-544.

[5] 杨成荫，王汉杰，周林，等. 基于全场信息的数值预报产品释用

方法研究.应用气象学报,2009,20(2):232-239.

[6] 刘长征,杜良敏,柯宗建,等.国家气候中心多模式解释应用集成预测.应用气象学报,2013,24(6):677-685.

[7] David J S,Nusrat Y. Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea Forecasting*,2007,22(1):3-17.

[8] Nusrat Y,David J S. Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system during the 2005/06 cool season. *Mont Wea Rev*,2008,136(6):2157-2172.

[9] 赵声蓉,裴海瑛.客观定量预报中降水的预处理.应用气象学报,2007,18(1):21-28.

[10] 唐晓文,汤剑平,张小玲.基于业务中尺度模式的配料法强降水定量预报.南京大学学报(自然科学版),2010,46(3):277-283.

[11] Jerome P C,Frederick G S. Regionalization in fine-grid GFS MOS 6-h quantitative precipitation forecasts. *Mon Wea Rev*,2011,139(1):24-38.

[12] Jerome P C,Frederick G S. High-resolution GFS-based MOS quantitative precipitation forecasts on a 4-km grid. *Mon Wea Rev*,2011,139(1):39-68.

[13] 孙靖,程光光,张小玲.一种改进的数值预报降水偏差订正方法及应用.应用气象学报,2015,26(2):173-184.

[14] Ding C,Peng H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*,2005,3(2):185-205.

[15] Peng H,Long F H,Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2005,27(8):1226-1238.

[16] Hossain M A,Jia X,Pickering M. Subspace detectionusing a mutual information measure for hyperspectral image classification. *IEEE on Geoscience and Remote Sensing Letters*,2014,2(11):424-428.

[17] Breiman L. Random forests. *Machine Learning*,2001,45(1):5-32.

[18] 方匡南.随机森林组合预测理论及其在金融中的应用.厦门:厦门大学出版社,2012.

[19] 邹亮,黄琼,李骜,等.基于随机森林和富集分析的阿尔茨海默症GWA研究.中国科学(生命科学),2012,42(8):639-647.

[20] 李建更,高志坤.随机森林:一种重要的肿瘤特征基因选择法.生物物理学报,2009,25(1):51-56.

[21] 方匡南,朱建平,谢邦昌.基于随机森林方法的基金收益率方向预测与交易策略研究.经济经纬,2010(2):61-65.

[22] Du J,Mullen S L,Sanders F. Short-range ensemble forecasting of quantitative precipitation. *Mon Wea Rev*,1997,125:2427-2459.

Application of ECMWF Ensemble Forecast Products to Rainstorm Forecast in Guangxi

Zhao Huasheng　Huang Xiaoyan　Huang Ying

(*Guangxi Research Institute of Meteorological Disasters Mitigation*, *Nanning* 530022)

Abstract

Using the maximal correlation minimum redundancy algorithm and random forest regression algorithm, a rainstorm interpretation forecasting method with numerical prediction products is proposed based on the ensemble prediction system of European Center for Medium-Range Weather Forecasts (ECMWF). The precipitation forecast of 51 members in ECMWF ensemble prediction system are interpolated to weather stations, and then, the maximum related minimum redundancy algorithm is used to filter ensemble members. Finally, several member interpolations that have the highest correlation with the predictand and the least redundancy with each other are selected as input factors of the random forest regression algorithm. Furthemore, in order to make modeling samples of the forecast model more pertinence, the modeling samples are classified using the mean rainfall value of ECMWF ensemble prediction products of 51 members. That is, when the mean precipitation using ECMWF ensemble prediction products at a certain station is relatively large and there is a possibility of precipitation above the storm level, only historical samples containing a large amount of precipitation are selected as modeling samples of the forecasting model. Therefore, the forecasting model reduces the influence of the sunny and wet weather samples on the noise of the forecasting model, so that forecasting model focuses on the training of large precipitation samples. When the mean value of the predicted ECMWF ensemble precipitation at a certain weather station is small, all samples of the weather station (including samples of sunny days and heavy precipitation) are modeled so that the training of the forecasting model can reconcile the heavy rain samples and thus as far as possible to avoid the rainstorm of weather station omissions reported. This method is applied to 89 stations in Guangxi, and a 4-year cross-independent sample test forecast for 2012－2015 is carried out. The business test forecast is carried out in 2016. In the 4-year cross-independent sample test results, rainstorm TS and ETS scores of this method are all improved by 0.04－0.09 and 0.04－0.07, respectively, compared with the average value after interpolation using the precipitation forecast of 51 members in ECMWF ensemble prediction products. Results of the business trial in 2016 show that TS and ETS scores of the method for interpretation rainstorms TS and ETS scores are improved by 0.07 and 0.05, respectively, compared with average values of pre-interpolation methods for the precipitation forecast of 51 members in ECMWF ensemble prediction products. It shows that the proposed rainstorm precipitation method of ECMWF ensemble prediction products has advantageous effects on forecasting and practical application forecast.

Key words: maximum relevance minimum redundancy; random forest regression; interpretation