孙全德,焦瑞莉,夏江江,等,2019. 基于机器学习的数值天气预报风速订正研究[J]. 气象,45(3):426-436. Sun Q D, Jiao R L, Xia J J, et al, 2019. Adjusting wind speed prediction of numerical weather forecast model based on machine learning methods[J]. Meteor Mon, 45(3):426-436(in Chinese).

基于机器学习的数值天气预报风速订正研究*

孙全德[1] 焦瑞莉[1] 夏江江[2] 严中伟[2] 李昊辰[3]
孙建华[2] 王立志[2] 梁钊明[4]

1 北京信息科技大学,北京 100101
2 中国科学院大气物理研究所,北京 100029
3 北京大学,北京 100871
4 中国气象科学研究院灾害天气国家重点实验室,北京 100081

提 要:对风速进行准确预测是精细化天气预报服务(如风能发电、冬季奥运会赛场条件保障等)的重要环节。本文基于三种机器学习算法(LASSO 回归、随机森林和深度学习),对数值天气预报模式 ECMWF 预测的华北地区近地面 10 m 风速进行订正。首先利用 LASSO 回归算法提取对 10 m 风速有重要影响的气象要素特征集,将其作为三种机器学习算法的输入,建立相应模型对 ECMWF 预测的风速进行订正。用提取后的气象要素特征集建模有助于减少计算量和存储开销,并减小模型的复杂性,从而提高模型的泛化能力。将订正结果与传统订正方法模式输出统计(model output statistics, MOS)得到的订正结果进行对比。结果表明,三种机器学习算法的订正效果均好于 MOS 方法,显示了机器学习方法在改善局地精准气象预报方面的潜力。

关键词:ECMWF 模式,机器学习,模式输出统计(MOS),风速,华北
中图分类号:P456 **文献标志码**:A **DOI**:10.7519/j.issn.1000-0526.2019.03.012

Adjusting Wind Speed Prediction of Numerical Weather Forecast Model Based on Machine Learning Methods

SUN Quande[1] JIAO Ruili[1] XIA Jiangjiang[2] YAN Zhongwei[2]
LI Haochen[3] SUN Jianhua[2] WANG Lizhi[2] LIANG Zhaoming[4]

1 Beijing Information Science and Technology University, Beijing 100101
2 Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029
3 Peking University, Beijing 100871
4 State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081

Abstract: Accurate prediction of wind speed is crucial for local weather forecasting services (e.g., dealing with wind power industry and the Olympic Winter Game). Based on three machine learning algorithms (LASSO regression, random forest and deep learning), this paper demonstrates three models for adjusting the 10 m wind speed in North China predicted by the numerical weather forecast model of ECMWF. Firstly, the LASSO regression algorithm is applied to identify the features which significantly affect the near-surface wind speed, among all the available meteorological elements. The extracted feature set is used as input for each machine learning algorithm to establish a model to adjust the ECMWF-predicted wind speed.

* 中国科学院战略性先导科技专项(A 类-XDA19030403)和北京信息科技大学 2017 年度"实培计划"共同资助
2018 年 5 月 28 日收稿; 2018 年 10 月 15 日收修定稿
第一作者:孙全德,主要从事机器学习研究. Email:41898090@qq.com
通信作者:夏江江,主要从事大气物理和天气预报研究. Email:xiajj@tea.ac.cn

Feature extraction helps to reduce the amount of computation, storage overhead and the complexity of the model, hence to facilitate the generalization of the model. The results of the three machine learning algorithms are compared with that of the traditional MOS method. All the three machine learning methods show a better performance in adjusting the wind speed than that of MOS, indicating great potential of the machine learning methods in improving local weather forecast.

Key words: ECMWF model, machine learning, MOS (model output statistics), wind speed, North China

引　言

提高风速的精细化预报水平是很多行业对精细化天气预报服务的需求。例如，对风速的精确预测是风电场风能预测的基础（张颖超等，2016；叶小岭等，2017），目前对风的精准预报水平不高在一定程度上制约了风力发电的发展，做好对风速的精确预报有助于高效利用风能这种可再生资源。又如，北京 2022 年冬季奥运会赛区地面风场是冬季奥运会组委会十分关注的气象条件之一（张治国等，2017），基于赛场赛道风速的预测才能对雪上项目的进行、基础设施（如缆车的使用）等提前做风险评估和应对准备。

目前风速预测方法大致可分为物理方法和统计方法。物理方法，如数值天气预报，主要考虑到影响风场的物理因素（如地形特征、气压和环境温度等）及其间的物理相互作用规律来对风场进行预测。这类方法需要对实际物理过程有清晰的认识和重现能力，但由于模式的物理参数化方案的不完善和很多参数的不确定性等，使得对近地面风场的预报存在较大的误差。统计方法通常利用大量历史数据来构建预测的统计模型，包括传统统计方法（Erdem and Shi，2011；Ren et al，2016；胡海川等，2017）和机器学习方法（Li and Shi，2010；杨薛明等，2016；López et al，2018；Wang et al，2018）。实践表明，现有单一模型很难准确地进行局地风速预测。近年来一些学者尝试基于机器学习方法对数值天气预报模式结果进行订正，达到对风速的精细化预报。传统的订正模型大多采用线性方法（肖擎曜等，2017），不足以捕捉风速变化中隐藏的非线性特征。基于机器学习方法的订正模型则能捕捉非线性风速变化，在风速预报上表现出良好的性能，例如针对风速预报订正的人工神经网络（孙军波等，2010；Zjavka，2015；邓华等，2018）、支持向量机（戚双斌等，2009；孔令彬等，2014）、随机森林（Lin et al，2015）等方法。

本文采用目前较为常用的三种机器学习（LASSO 回归、随机森林和深度学习）以及 MOS 方法（经典天气预报的统计订正方法），对 ECMWF 数值天气预报模式预测的近地面（10 m）风速进行订正。首先基于 ECMWF 模式计算所得的各种要素特征进行特征选择，即通过机器学习算法自适应地获得相关要素特征集，再以选择的特征集进行机器学习建模，对 ECMWF 预测的未来 1～15 d 华北地区逐日格点风速进行订正。以此评估各订正方法的能力，为实现风速的精细化预报提供新的方法思路。

1　数据与方法

1.1　数据

本文采用的数据来源于欧洲中期天气预报中心（European Centre for Medium-Range Weather Forecasts，ECMWF）网站公开的数值模式输出数据。数据包括逐日 00 时（对应北京时间为 08 时）的分析场（0 时刻场）和 24～360 h（逐日）的预报场。时间范围为 2012 年 1 月至 2016 年 12 月，空间范围为 38°～43°N、113°～119°E（华北地区），空间分辨率为 0.5°×0.5°。ECMWF 模式输出数据共包括 23 个气象要素场（表 1）。

首先将 10 m 纬向风分量（U）和 10 m 经向风分量（V）合成为 10 m 风速（W），合成公式如下：

$$W = \sqrt{U^2 + V^2} \tag{1}$$

故现在共有 24 个要素。本研究将采用研究时段内模式 0 时刻 10 m 风速的分析场作为标记（机器学习算法中的真值），将 ECMWF 预测的对应标记所处时刻的所有 24 个气象要素作为机器学习算法的输入，以此构建机器学习风速订正模型。

表 1 ECMWF 数值预报 23 个气象要素场

Table 1 Twenty-three meteorological element fields of ECMWF numerical prediction

序号	ECMWF 数值预报 23 个气象要素场
1	2 m 温度(2 m temperature)
2	降雪水当量(snow fall water equivalent)
3	日照时间(sunshine duration)
4	地表潜热通量(surface latent heat flux)
5	地表净太阳辐射(surface net solar radiation)
6	地表净热辐射(surface net thermal radiation)
7	地表感热通量(surface sensible heat flux)
8	最高净热辐射(top net thermal radiation)
9	总降水量(total precipitation)
10	过去 6 h 2 m 最高温度(maximum temperature at 2 m in the last 6 hours)
11	过去 6 h 2 m 最低温度(minimum temperature at 2 m in the last 6 hours)
12	10 m 纬向风分量(10 m U wind component)
13	10 m 经向风分量(10 m V wind component)
14	2 m 露点温度(2 m dewpoint temperature)
15	地形(orography)
16	对流有效位能(convective available potential energy)
17	平均海平面气压(mean sea level pressure)
18	体感温度(skin temperature)
19	积雪深度水当量(snow depth water equivalent)
20	地面气压(surface pressure)
21	总云量(total cloud cover)
22	总柱水汽量(total column water)
23	海陆(land-sea mask)

1.2 方法

1.2.1 MOS 预报方法

MOS 方法是在数值天气预报模式的预报产品和相应时次的预报对象间建立统计关系(预报方程)(吴启树等,2016)。本文以 ECMWF 数值预报模式在某一预报时效的风速的预报场和与其对应的分析场建立一元线性回归方程:

$$S_i = a + bF_i \tag{2}$$

式中,S_i 为第 i 时刻回归订正值,F_i 为该时刻模式预报值,a 为常数项,b 为回归系数(采用最小二乘法求解)。用得到的回归方程对所有时效的风速进行订正。

1.2.2 LASSO 回归

LASSO(least absolute shrinkage and selection operator)回归通过构造一个惩罚函数得到一个较为精炼的模型,使得它压缩一些系数,同时设定一些系数为 0。因此保留了子集收缩的优点,是一种处理具有复共线性数据的有偏估计。LASSO 回归通过放弃最小二乘法的无偏性,以损失部分信息和降低精度为代价获得回归系数更为符合实际、更可靠的回归方法。

在本文风速预测中,给定有 m 个气象因素(特征)的特征向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$,其中 x_i 为 x 在第 i 个特征上的取值,通过 m 个气象特征的线性组合来进行预测风速,即公式如下:

$$f(\boldsymbol{x}) = w_1x_1 + w_2x_2 + \cdots + w_mx_m \tag{3}$$

式中 $w=(w_1,w_2,\cdots,w_m)$是气象特征的权重。

损失函数定义为

$$loss(\boldsymbol{w}) = \| f(\boldsymbol{x}) - y \|^2 + \alpha \| \boldsymbol{w} \| \tag{4}$$

式中,y 表示风速实测值,$\alpha \| \boldsymbol{w} \|$ 是正则化项,不仅有助于降低过拟合风险,还具有特征选择的作用(周志华,2016)。通过对 $loss(\boldsymbol{w})$进行求最小值,模型学习得到 $\boldsymbol{w}$,从而 LASSO 回归模型得以确定。

1.2.3 随机森林

随机森林(random forest)算法是由多个决策树集成的,在进行随机森林过程中,其输出值是随机森林中所有决策树结果的平均值。

在本文风速预测中,就是以气象要素建立特征向量作为输入,以该特征向量对应的风速大小作为预测结果,通过训练样本进行拟合得到预测模型的过程(李丽辉等,2017)。随机森林建模过程如下:

(1) 定义风速预测训练集合 $\boldsymbol{X}_i \rightarrow Y_i$。其中,$Y_i$ 为随机森林预测模型中的真实值,映射为资料中第

i 个样本的风速实值;$\boldsymbol{X}_i$ 为资料中第 i 个样本的气象要素取值所建立的特征向量,以$\{I_{i1},I_{i2},\cdots I_{in}\}\rightarrow X_i$ 表示第 i 个样本的 n 个影响因子。

(2) 在确定了训练集的基础上,建立单棵回归决策树。通过训练样本中的特征向量 $\boldsymbol{X}$ 和其对应的真实值 Y,对分裂变量和分裂值进行搜索,回归决策树将整个向量空间分为 m 个分区$\{\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_m\}$。对于其中任意分区可以映射为模型 C_m,通过某个特征的取值将向量空间分为两部分,表达式为

$$\boldsymbol{R}_1(j,s)=\{\boldsymbol{I}\mid I_j\leqslant s\} \tag{5}$$

$$\boldsymbol{R}_2(j,s)=\{\boldsymbol{I}\mid I_j>s\} \tag{6}$$

式中,j 代表一个影响因子,s 代表进行分裂时的值。进行向量空间分裂变量和分裂值搜索的目标函数为

$$z:\min_{j,s}\Big[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\Big] \tag{7}$$

式中,z 为风速实值的最小方差;y_i 为第 i 个样本的风速实值;x_i 为第 i 个样本的影响因子向量的对应值;c_1 为第一部分风速实值均值;c_2 为第二部分风速实值均值。

(3) 在单棵决策树的构建基础上,构建整个随机森林。生成的随机森林是多元非线性回归分析模型,随机森林预测值是所有决策树预测值的平均值。

1.2.4　*深度学习*

深度学习是含有多个神经元层的深层神经网络。深度学习基本模型如图 1 所示,深度学习是由输入层、隐藏层和输出层三部分组成,隐藏层可以包含很多层。相对比于浅层学习,深度学习显然在计算层次上更为复杂。在模型训练方面,深度学习采用反向传播算法。反向传播算法的核心思想是求导的链式法则,常被用来求解神经网络中的最优化问题。输入层神经元个数等于样本的特征量,隐藏层的层数、隐藏层的神经元数、学习率等参数,是通过大量的训练与验证而确定。

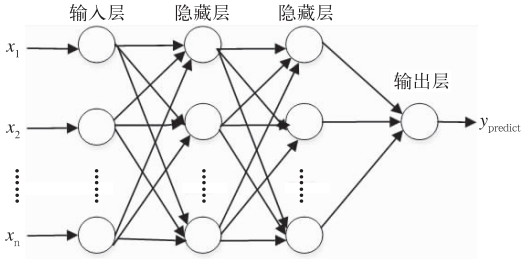

图 1　深度学习基本模型

Fig. 1　The basic model of deep learning

结合图 1 和本文研究内容,由于本文构建的是回归模型,故输出层不设激活函数,即将第四个隐藏层的输出数据加权平均后直接输出。公式如下:

$$\boldsymbol{y}^{(1)}=\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)} \tag{8}$$

$$\boldsymbol{y}^{(2)}=\boldsymbol{W}^{(2)}\varphi_1(\boldsymbol{y}^{(1)})+\boldsymbol{b}^{(2)} \tag{9}$$

$$\boldsymbol{y}^{(3)}=\boldsymbol{W}^{(3)}\varphi_2(\boldsymbol{y}^{(2)})+\boldsymbol{b}^{(3)} \tag{10}$$

$$\boldsymbol{y}^{(4)}=\boldsymbol{W}^{(4)}\varphi_3(\boldsymbol{y}^{(3)})+\boldsymbol{b}^{(4)} \tag{11}$$

$$\boldsymbol{y}_{\text{predict}}=\boldsymbol{W}^{(5)}\varphi_4(\boldsymbol{y}^{(4)})+\boldsymbol{b}^{(5)} \tag{12}$$

式中,$\boldsymbol{x}$ 是有气象要素组成的特征向量,是输入层的输入信号。$\boldsymbol{y}^{(i)}$ 为第 i 个隐藏层的输入信号,$\boldsymbol{W}^{(i)}$ 为第 $i-1$ 层到第 i 层的连接权重,$\boldsymbol{b}^{(i)}$ 为第 $i-1$ 层到第 i 层的连接偏差,φ_i 为第 i 个隐层的激活函数。$\boldsymbol{y}_{\text{predict}}$(风速预测值)为输出层的输出信号。隐层神经元输出的激活函数采用 ReLU 函数,它的数学表达式如下:

$$\varphi(z)=\max(0,z) \tag{13}$$

优化算法采用 Adam 算法,该算法是随机梯度下降算法的扩展式,它对超参数的选择相当鲁棒。

1.2.5　*检验方法*

均方根误差(root mean square error, RMSE)是风速预报中最常用的性能度量指标,均方根误差越小,风速整体预报就越准确。公式如下:

$$RMSE(f;D)=\sqrt{\frac{1}{K}\sum_{k=1}^{K}[f(\boldsymbol{X}_k)-y_k]^2} \tag{14}$$

式中,f 为算法模型,D 为数据集,K 为数据集 D 的样本总数,$\boldsymbol{X}_k$ 为第 k 个样本的输入,y_k 为第 k 个样本的标记。

风速预报准确率(F_a)是风速预报绝对偏差不大于 1 m·s^{-1}的百分率,公式如下:

$$F_a=\frac{N_r}{N_f}\times 100\% \tag{15}$$

式中,N_r 为风速预测值与分析场风速值之差不大于 1 m·s^{-1}的样本数,N_f 为预报的样本数。

2　预测模型构建

将数据分为三部分:训练集、验证集和测试集。训练集和验证集用于确定预测模型,测试集进行实际预测。从 2012 年 1 月至 2015 年 12 月 ECMWF 预报数据中随机抽取 85%作为训练集,剩下 15%作为验证集;2016 年 1—12 月 ECMWF 预报数据作为测试集。基于机器学习的风速预测模型流程图如图 2 所示。

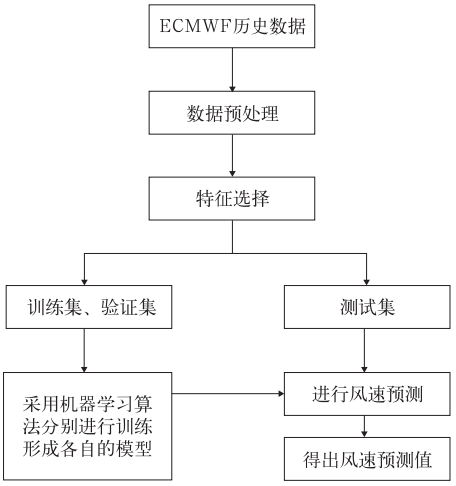

图 2 机器学习风速预测流程图

Fig. 2 Flow chart of machine learning predicting wind speed

(1) 数据预处理。针对数据矩阵中的空值和乱码进行处理,也对整体数据进行拆分和采样等操作。同时,对各要素数据进行标准化处理,不仅避免训练时由于各要素数值小而贡献小的问题(袁翀等,2016),而且提高运算速率。本文预报时效为 24～360 h(时效间隔为 24 h)的预报场数据分别对应未来第 1～15 d(间隔为 1 d)的分析场数据;数据集为 2012 年 1 月至 2016 年 12 月一共有 1827 d;华北地区(38°～43°N、113°～119°E)模式预报数据,水平分辨率为 0.5°×0.5°的网格,共计 143 个格点;因此每个预报时效的原始数据集由一个大小为 1827×143 个样本组成,每个样本有 24 个特征。

(2) 特征选择。从输入的原始数据(共 24 个特征)中,利用 LASSO 回归算法提取出对 10 m 风速有影响的气象要素特征集。结果详见第 3 节。

(3) 将选择选出的特征组合成新的输入数据,采用机器学习算法(LASSO 回归、随机森林和深度学习)分别进行训练形成模型。

(4) 将测试集数据输入到已训练好的模型中,输出即为订正后的风速数据,评估预测模型的准确性。

3 特征选择

特征选择是从原始特征中选择出一些最有效特征以降低数据集维度的过程,学习任务的难度会有所降低,涉及的计算和存储开销会减少,学习得到模型的可解释性也会提高。常见的特征选择法有过滤式选择、包裹式选择和嵌入式选择。过滤式选择过程与后续学习器无关。从最终学习器性能来看,包裹式特征选择比过滤式特征选择更好,但包裹式特征选择计算开销大(周志华,2016)。本文是基于 LASSO 回归的嵌入式选择法进行特征选择,LASSO 回归的学习方法是其特征选择过程与学习器训练过程融为一体,两者在同一个优化过程中完成,并且 LASSO 回归计算效率高。

结合 1.2.2 节,使用 LASSO 模型后,由于加入了正则化因子 $\alpha \| \boldsymbol{w} \|$,不显著的变量被收缩为 0。随着惩罚力度的加强(超参数 α 变大),越来越多的变量会被收缩为 0。LASSO 不仅可以降低过拟合风险,还具有稀疏作用(韩耀风等,2017)。

以预报时效为 216 h 为例,如图 3 所示,基于 LASSO 模型对测试集的风速预测,横坐标特征的数目是通过调节 LASSO 模型中超参数 α 的大小来控制,纵坐标是 RMSE。当 LASSO 筛选出 10 个特征时,就足以将 RMSE 降低到平稳的边界。为了检验这 10 个特征(10 m 纬向风分量、10 m 经向风分量、10 m 风速、积雪深度水当量、地表净热辐射、地面气压、地面感热通量、最高净热辐射、总柱水汽量和海陆)选择结果的稳健性,分别采用随机森林和深度学习算法进行风速预测。并把 ECMWF 风速预测(EC)、输入 24 个特征的随机森林风速预测(RF-all)、输入选择出这 10 个特征的随机森林风速预测(RF-select)、输入 24 个特征的深度学习风速预测(DL-all)和输入选择出这 10 个特征的深度学习风速预测(DL-select)。图4显示,随机森林和深度学

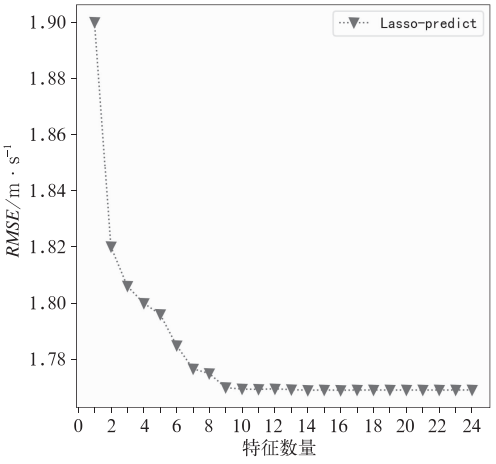

图 3 基于 LASSO 模型的风速预测

Fig. 3 Wind speed forecast based on LASSO model

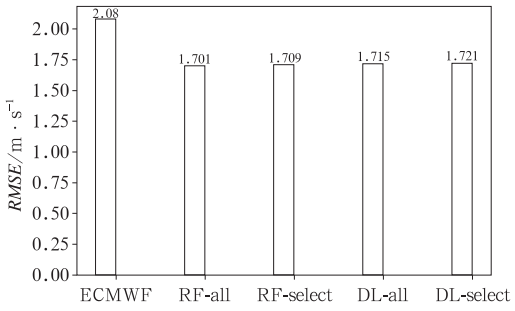

图 4　特征选择前后的风速预测对比

Fig. 4　Comparison of wind speed prediction before and after feature selection

习算法的预测性能明显优于 ECMWF;并且 RF-select 的 RMSE 基本等于 RF-all,DL-select 的 RMSE 基本等于 DL-all,充分证明 LASSO 能够筛选出对风速有关的主要特征。

按照上例特征选择方法对所有时效(24～360 h 逐 24 h)进行特征选择。选择结果如表 2 和表 3 所示,表中的数值是各预报时效中特征权重。特征对风速预测的影响程度与特征权重的大小成正比,即某个特征的权重越大说明这个特征对风速预测的影响就越大。对预报时效 24～360 h 中的平均权重由大到小依次特征排列为:10 m 风速、海陆、地表净热辐射、积雪深度水当量、地面气压、10 m 纬向风分量、10 m 经向风分量、最高净热辐射、地面感热通量、总柱水汽量、2 m 露点温度、降雪水当量、平均海平面气压、总云量。在以上对风速预测有影响的特征中:(1)10 m 风速的权重随着预报时效的增加而减小;(2)海陆、地表净热辐射、积雪深度水当量、地面气压、地面感热通量的权重是随着预报时效的增加而增大的;(3)10 m 纬向风分量、10 m 经向风分量、最高净热辐射、总柱水汽量、2 m 露点温度、降雪水当量、平均海平面气压、总云量的权重与预报时效的增加没有明显的相关关系。

风是一个二维矢量,即 10 m 风速由 10 m 纬向风分量和 10 m 经向风分量组成。ECMWF 预测的 10 m 纬向风分量、10 m 经向风分量和 10 m 风速,显然与分析场风速相关。当太阳福射在地球表面上后,地表会向大气支出热量(地表净热辐射和最高净热辐射),地表的空气受热膨胀变轻而往上升。热空气上升后,造成气压分布不均,低温的冷空气横向流入,这种空气的流动就产生了风。地表感热通量是由于湍流运动从地面向大气传输的热量通量,一般

表 2　各预报时效中特征权重分布(预报时效为 24～192 h)

Table 2　Distribution of feature weights in each forecast period (forecast lead-time is 24－192 h)

特征	预报时效/h							
	24	48	72	96	120	144	168	192
10 m 纬向风分量	0.075	0.107	0.107	0.113	0.122	0.147	0.109	0.080
10 m 经向风分量	0.024	0.046	0.031	0.054	0.007	0	0	0
10 m 风速	1.708	1.588	1.476	1.336	1.202	1.011	0.910	0.780
2 m 露点温度	0	0	0.021	0.045	0.034	0	0	0
2 m 温度	0	0	0	0	0	0	0	0
对流有效位能	0	0	0	0	0	0	0	0
过去 6 h 2 m 最高温度	0	0	0	0	0	0	0	0
平均海平面气压	0	0	0	0.004	0.022	0	0.010	0
过去 6 h 2 m 最低温度	0	0	0	0	0	0	0	0
地形	0	0	0	0	0	0	0	0
体感温度	0	0	0	0	0	0	0	0
积雪深度水当量	0	0	0	0.002	0.026	0.069	0.089	0.083
降雪水当量	0	0	0	0	0	0	0	0
日照时间	0	0	0	0	0	0	0	0
地表潜热通量	0	0	0	0	0	0	0.003	0
地表净太阳辐射	0	0	0	0	0	0	0	0
地表净热辐射	0	0	0.008	0.017	0.053	0.094	0.120	0.143
地面气压	0	0.001	0.007	0.018	0.034	0.063	0.07	0.107
地面感热通量	0	0	0	0	0	0	0.007	0
最高净热辐射	0	0	0.010	0.019	0.043	0.060	0.085	0.108
总云量	0	0	0.015	0.001	0	0	0	0
总柱水汽量	0.013	0.028	0.014	0	0	0.010	0	0
总降水量	0	0	0	0	0	0	0	0
海陆	0.026	0.068	0.111	0.163	0.214	0.285	0.318	0.364

表 3 同表 2,但为 216～360 h

Table 3 Same as Table 2, but for 216－360 h

特征	预报时效/h							
	216	240	264	288	312	336	360	平均
10 m 纬向风分量	0.040	0.010	0	0.030	0	0	0	0.063
10 m 经向风分量	0.020	0.083	0.104	0.047	0.138	0.131	0.162	0.056
10 m 风速	0.691	0.59	0.506	0.451	0.437	0.446	0.425	0.900
2 m 露点温度	0	0	0	0	0	0	0	0.007
2 m 温度	0	0	0	0	0	0	0	0
对流有效位能	0	0	0	0	0	0	0	0
过去 6 h 2 m 最高温度	0	0	0	0	0	0	0	0
平均海平面气压	0	0.001	0.004	0	0	0	0	0.003
过去 6 h 2 m 最低温度	0	0	0	0	0	0	0	0
地形	0	0	0	0	0	0	0	0
体感温度	0	0	0	0	0	0	0	0
积雪深度水当量	0.120	0.129	0.193	0.209	0.201	0.214	0.190	0.102
降雪水当量	0	0.007	0.018	0.004	0.027	0.003	0	0.004
日照时间	0	0	0	0	0	0	0	0
地表潜热通量	0	0	0	0	0	0	0	0
地表净太阳辐射	0	0	0	0	0	0	0	0
地表净热辐射	0.171	0.147	0.247	0.264	0.290	0.301	0.286	0.143
地面气压	0.121	0.122	0.173	0.177	0.177	0.166	0.169	0.094
地面感热通量	0.019	0.020	0.059	0.071	0.099	0.089	0.083	0.030
最高净热辐射	0.088	0.049	0.067	0.043	0.041	0.051	0.073	0.049
总云量	0	0	0.001	0	0	0	0	0.001
总柱水汽量	0.013	0.087	0	0	0	0	0	0.011
总降水量	0	0	0	0	0	0	0	0
海陆	0.382	0.417	0.440	0.462	0.458	0.455	0.470	0.309

地风速越大,感热通量越大(阳坤等,2010)。当阳光照向地球表面上时,云、地面积雪从中吸收一部分热辐射,所以云和地面积雪的分布会影响地表热量分布。2 m 露点温度和总柱水汽量是衡量空气湿度的重要指标,风是表征大气运动的变量,湿度的分布因大气的运动而改变,在数值预报中,风会引起湿度场的迅速改变(陶祖钰等,2016)。海陆是区分陆地和海洋的重要特征,因海洋和陆地受热不均匀造成地面与海平面气压不同,从而在海岸附近形成的一种有日变化的风系。以预报时效为 216 h 为例对海陆因素的影响进行进一步的对比分析,它们的区域预测 RMSE 空间分布如图 5 所示,去掉海陆后,海域及沿海地区的 RMSE 明显上升,说明海陆对风速预测确实有重要作用,从一定程度上反映了使用 LASSO 方法进行特征选择的合理性。

4 订正结果

4.1 预报时效订正

分别采用机器学习(LASSO 回归、随机森林和深度学习)和 MOS 方法,对华北地区进行风速格点预测,预报时效为 24～360 h(时效间隔为 24 h)。并结合 ECMWF 分析场对预报能力进行客观检验。在测试集进行测试,对每个预测时效(样本量 360×143 个)计算 RMSE 和精准度,结果如图 6 所示:机器学习、MOS 法和 ECMWF 的预报精度都在随着预报时效的增加而下降;MOS 方法和机器学习的预报效果优于 ECMWF,并且在未来第 8～15 d,MOS 方法和机器学习方法预报效果明显优于 ECMWF;在未来第 1～5 d,机器学习方法略优于 MOS 方法,在未来第 6～15 d,机器学习方法明显优于 MOS 方法。总之,机器学习对 ECMWF 在不同时效的10 m 风速预报有不同程度的订正,尤其对未来第 8～15 d 有明显的订正。

4.2 区域订正

分别采用 LASSO 回归、随机森林、深度学习和 MOS 方法,对华北地区进行风速格点预测,预报时效为 24～360 h(时效间隔为 24 h)。并结合 ECMWF 分析场对预报能力进行客观检验。在测试集进行测试,分别对预报时效为72、192和

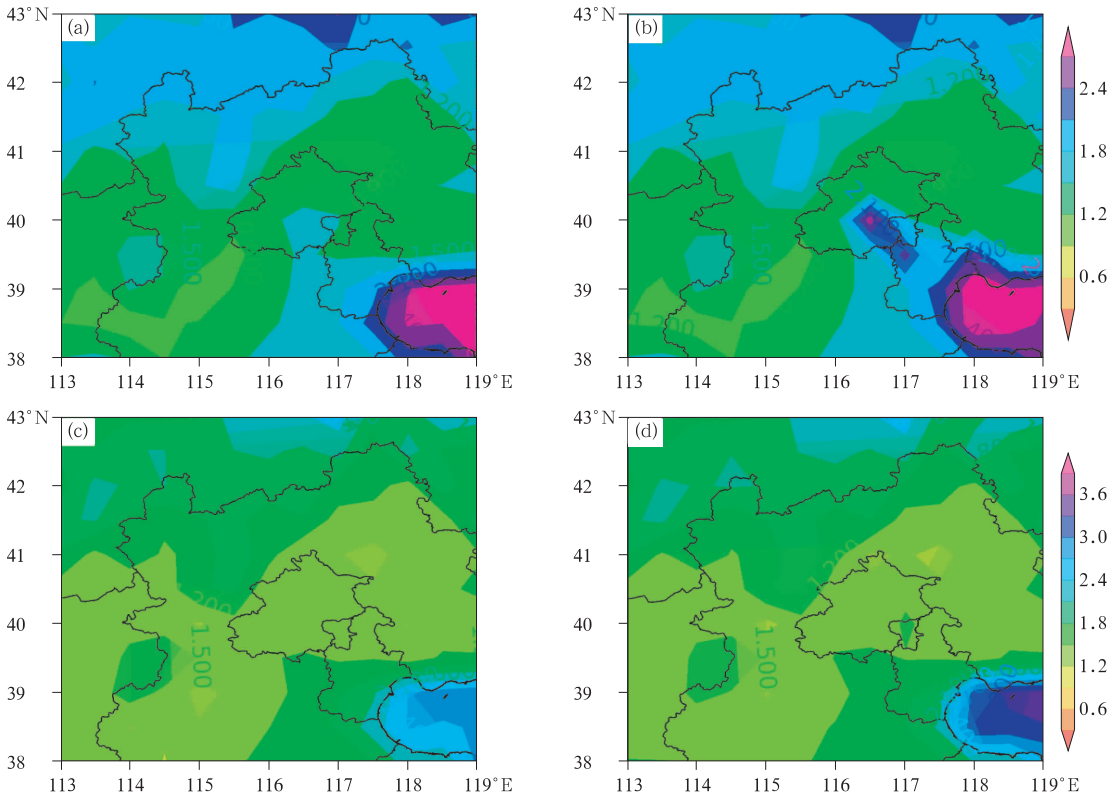

图 5　去掉海陆因素前后的风速预测空间分布(单位:m·s^{-1})

(a)基于随机森林的订正(有海陆因素),(b)基于随机森林的订正(没有海陆因素),

(c)基于深度学习的订正(有海陆因素),(d)基于深度学习的订正(没有海陆因素)

Fig. 5　Prediction of spatial distribution of wind speed before and after removing the land-sea mask (unit: m·s^{-1})

(a)correction based on random forest (with land-sea mask), (b) correction based on random forest (without land-sea mask), (c) correction based on deep learning (with land-sea mask), (d) correction based on deep learning (without land-sea mask)

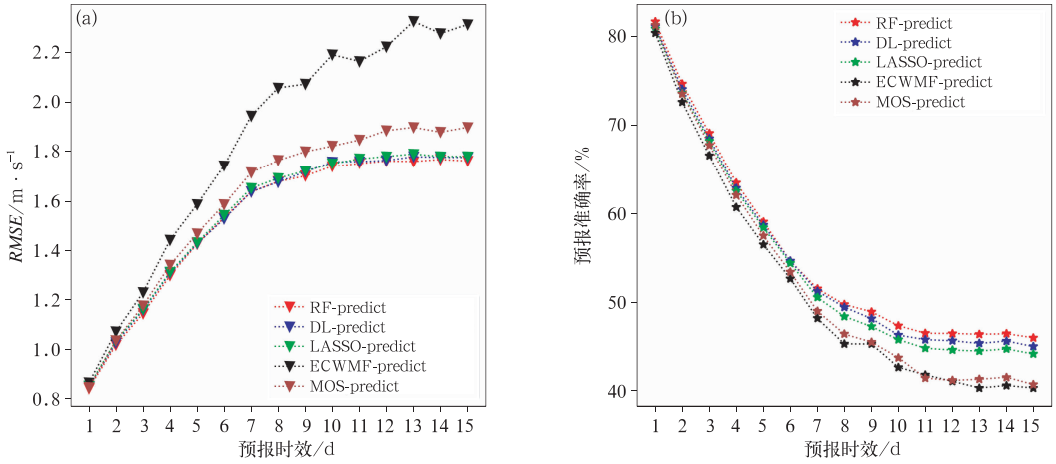

图 6　ECMWF、三种机器学习模型、MOS 在预报时效上的预测效果

(a)RMSE,(b)预报准确率

Fig. 6　Comparison of ECMWF, three machine learning models and MOS prediction

(a)RMSE,(b)forcast accuracy

312 h 中的每个格点(样本量 360 个)计算 RMSE,结果如图 7 所示:随着预报时效的增加,机器学习、MOS法和 ECMWF 在各格点预测的误差都呈逐渐上升趋势;机器学习和MOS法在格点上对ECMWF都有不

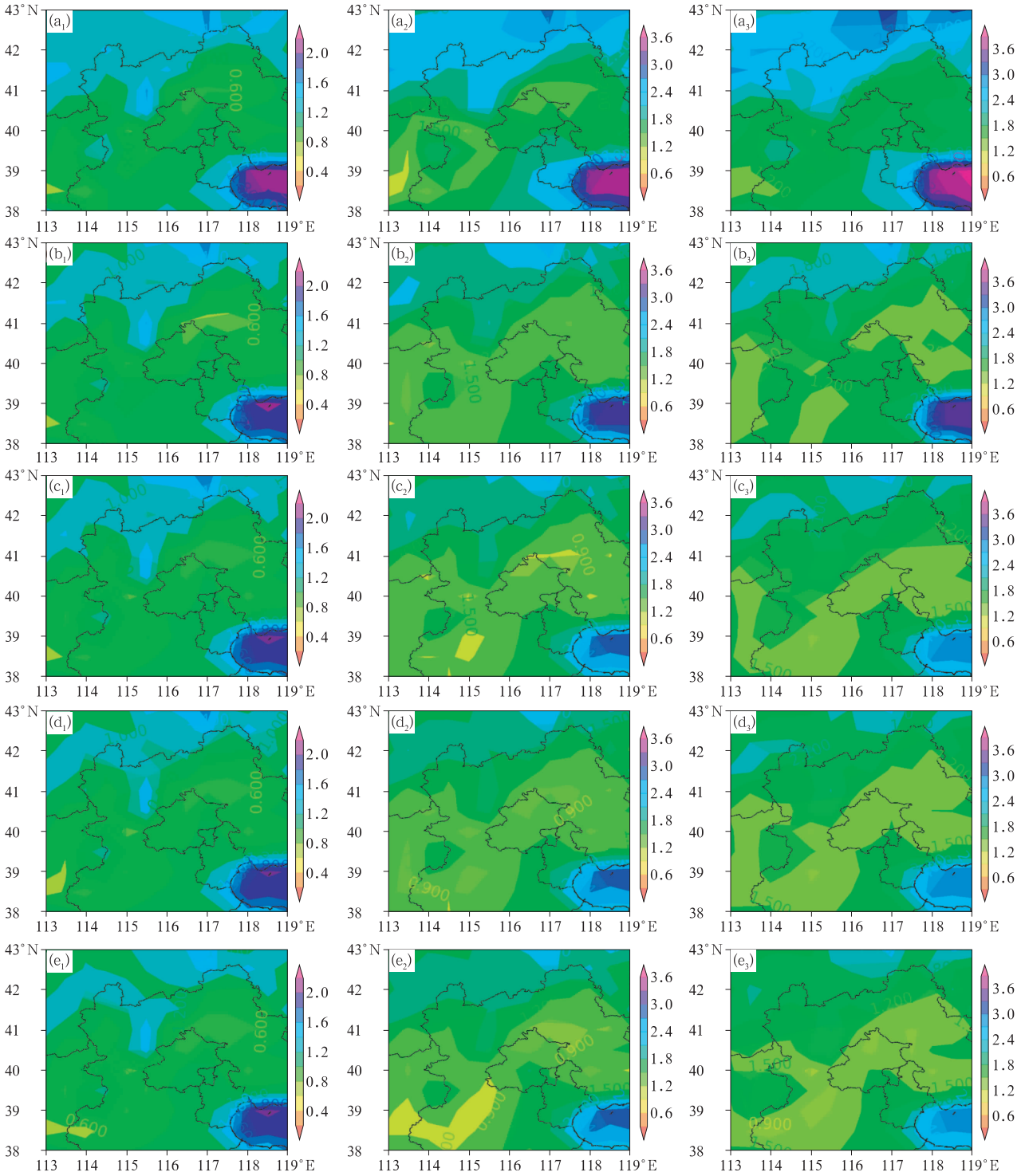

图 7 ECMWF(a)、MOS(b)、LASSO(c)、深度学习(d)和随机森林(e)在区域上的
72 h(a_1,b_1,c_1,d_1,e_1)、192 h(a_2,b_2,c_2,d_2,e_2)、312 h(a_3,b_3,c_3,d_3,e_3)预测结果(RMSE,单位:$m \cdot s^{-1}$)

Fig. 7 ECMWF (a), MOS (b), LASSO (c), deep learning (d) and random forest (e) prediction results (RMSE) distribution in area (unit: $m \cdot s^{-1}$)
(a_1,b_1,c_1,d_1,e_1) 72 h, (a_2,b_2,c_2,d_2,e_2) 192 h, (a_3,b_3,c_3,d_3,e_3) 312 h

同程度订正；机器学习在北京地区的预测精度优于其他地区；ECMWF 和 MOS 法对海域预测较差，机器学习明显改善了对海域及沿海地区预测的精度；总之，机器学习对 ECMWF 在不同区域的 10 m 风速预报有不同程度的订正，尤其对海域及其沿海地区有明显的订正。

5　结　论

本文基于机器学习方法针对数值天气预报模式 ECMWF 预测的 10 m 风速进行了订正。较之于传统 MOS 订正方法，本研究中机器学习模型构建中综合考虑了多种气象要素特征，获得了更完善的风速订正模型。结果表明，机器学习对不同时效的 10 m 风速预报有不同程度的订正，随着预报时效的增加，订正的力度越来越大。尤其是 LASSO 回归算法不仅提高了风速订正的准确性，还在特征选择上表现非常出色；特征选择降低机器学习任务的难度，减少了计算和存储开销，从而优化了学习模型。研究表明，对 10 m 风速预测有影响的特征主要（按平均权重由大到小排序）是 10 m 风速、海陆、地表净热辐射、积雪深度水当量、地面气压、10 m 纬向风分量、10 m 经向风分量、最高净热辐射、地面感热通量、总柱水汽量、2 m 露点温度、降雪水当量、平均海平面气压、总云量。通过机器学习算法自适应获得的气象要素特征集从一定程度上也有助于加深对以往建立的风速预测物理模型所使用的气象要素组合的认识。

参考文献

邓华，张颖超，顾荣，等，2018. 基于 PCA-RBF 的风电场短期风速订正方法研究[J]. 气象科技，46(1)：10-15. Deng H，Zhang Y C，Gu R，et al，2018. Research on Short-term wind speed correction method of wind farm based on PCA-RBF[J]. Meteor Sci Tech，46(1)：10-15(in Chinese).

韩耀风，覃文锋，陈炜，等，2017. adaptive LASSO logistic 回归模型应用于老年人养老意愿影响因素研究的探讨[J]. 中国卫生统计，34(1)：18-22. Han Y F，Qin W F，Chen W，et al，2017. Study on the application of adaptive LASSO logistic regression model to the influencing factors of the elderly's willingness to support the elderly[J]. Chin J Health Statis，34(1)：18-22(in Chinese)

胡海川，黄彬，魏晓琳，2017. 我国近海洋面 10 m 风速集合预报客观订正方法[J]. 气象，43(7)：856-862. Hu H C，Huang B，Wei X L，2017. An objective correction method for the ensemble prediction of 10 m wind speed near the ocean surface in China [J]. Meteor Mon，43(7)：856-862(in Chinese).

孔令彬，赵艳茹，王聚杰，等，2014. 基于支持向量机风速订正方法的研究[J]. 西南大学学报(自然科学版)，36(5)：194-200. Kong L B，Zhao Y R，Wang J J，et al，2014. Research on wind speed correction method based on support vector machine[J]. Southwest Univ(Nat Sci)，36(5)：194-200(in Chinese).

李丽辉，朱建生，强丽霞，等，2017. 基于随机森林回归算法的高速铁路短期客流预测研究[J]. 铁道运输与经济，39(9)：12-16. Li L H，Zhu J S，Qiang L X，et al，2017. Research on short-term passenger flow forecasting of high-speed railway based on random forest regression algorithm[J]. Rail Way Transport and Economy，39(9)：12-16(in Chinese).

戚双斌，王维庆，张新燕，2009. 基于支持向量机的风速与风功率预测方法研究[J]. 华东电力，37(9)：1600-1603. Qi S B，Wang W Q，Zhang X Y，et al，2009. Research on wind speed and wind power prediction based on support vector machine[J]. East China Electric Power，37(9)：1600-1603(in Chinese).

孙军波，钱燕珍，陈佩燕，等，2010. 登陆台风站点大风预报的人工神经网络方法[J]. 气象，36(9)：81-86. Sun J B，Qian Y Z，Chen P Y，et al，2010. Artificial neural network method for landing typhoon site gale forecast[J]. Meteor Mon，36(9)：81-86(in Chinese).

陶祖钰，范俊红，李开元，等，2016. 谈谈气象要素(压、温、湿、风)的物理意义和预报应用价值[J]. 气象科技进展，6(5)：59-64. Tao Z Y，Fan J H，Li K Y，et al，2016. Talk about the physical meaning and forecasting application value of meteorological elements (pressure，temperature，humidity，wind)[J]. Adv Meteor Sci Technol，6(5)：59-64(in Chinese).

吴启树，韩美，郭弘，等，2016. MOS 温度预报中最优训练期方案[J]. 应用气象学报，27(4)：426-434. Wu Q S，Han M，Guo H，et al，2016. Optimal training period scheme in MOS temperature prediction[J]. J Appl Meteor，27(4)：426-434(in Chinese).

肖擎曜，胡非，范绍佳，等，2017. 风能数值预报的模式输出统计(MOS)研究[J]. 资源科学，39(1)：116-124. Xiao Q Y，Hu F，Fan S J，et al，2017. Model output statistics (MOS) of wind energy numerical prediction[J]. Res Sci，39(1)：116-124(in Chinese).

阳坤，郭晓峰，武炳义，2010. 青藏高原地表感热通量的近期变化趋势[J]. 中国科学：地球科学，40(7)：923-932. Yang K，Guo X F，Wu B Y，2010. Recent changes in surface sensible heat flux over the Qinghai-Tibet Plateau[J]. Sci China：Earth Sci，40(7)：923-932(in Chinese).

杨薛明，边继飞，朱霄珣，等，2016. 基于最大熵混沌时间序列的支持向量机短期风速预测模型研究[J]. 太阳能学报，37(9)：2173-2179. Yang X M，Bian J F，Zhu X X，et al，2016. Research on short-term wind speed prediction model of support vector machine based on maximum entropy chaotic time series[J]. Acta Energiae Solaris Sin，37(9)：2173-2179(in Chinese).

叶小岭，顾荣，邓华，等，2017. 基于 WRF 模式和 PSO-LSSVM 的风电场短期风速订正[J]. 电力系统保护与控制，45(22)：48-54. Ye X L，Gu R，Deng H，et al，2017. Short-term wind speed correc-

tion of wind farm based on WRF mode and PSO-LSSVM[J]. Power System Protection and Control, 45(22): 48-54(in Chinese).

袁翀,戚佳金,王文霞,等,2016.采用正则化极限学习机的短期风速预测[J].电网与清洁能源,32(11):62-68. Yuan C, Qi J J, Wang W X, et al, 2016. Short-term wind speed prediction using regularized limit learning machine[J]. Power System and Clean Energy, 32(11): 62-68(in Chinese).

张颖超,肖寅,邓华,2016.基于ELM的风电场短期风速订正技术研究.气象,42(4):466-471. Zhang Y C, Xiao Y, Deng H, 2016. Research on short-term wind speed correction technology of wind farm based on ELM[J]. Meteor Mon, 42(4): 466-471(in Chinese).

张治国,崔炜,白雪涛,等,2017.第24届冬奥会海坨山赛区近两年冬季地面风场特征[J].干旱气象,35(3):433-438. Zhang Z G, Cui W, Bai X T, et al, 2017. Characteristics of ground wind field in the winter of Haishu Mountain in the 24th Winter Olympic Games[J]. Arid Meteor, 35(3): 433-438(in Chinese).

周志华,2016.机器学习[M].北京:清华大学出版社:248-261. Zhou Z H, 2016. Machine Learning[M]. Beijing: Tsinghua University Press: 248-261(in Chinese).

Erdem E, Shi J, 2011. ARMA based approaches for forecasting the tuple of wind speed and direction[J]. Appl Energy, 88(4): 1405-1414.

Li G, Shi J, 2010. On comparing three artificial neural networks for wind speed forecasting[J]. Appl Energy, 87(7): 2313-2320.

Lin Y J, Kruger U, Zhang J P, et al, 2015. Seasonal analysis and prediction of wind energy using random forests and arx model structures[J]. IEEE Trans Control Syst Technol, 23(5): 1994-2002.

López E, Valle C, Allende H, et al, 2018. Wind power forecasting based on echo state networks and long short-term memory[J]. Energies, 11(3): 526.

Ren Y, Suganthan P N, Srikanth N, 2016. A novel empirical mode decomposition with support vector regression for wind speed forecasting[J]. IEEE Trans Neural Netw Learn Sys, 27(8): 1793-1798.

Wang J J, Wang Y F, Li Y N, 2018. A novel hybrid strategy using three-phase feature extraction and a weighted regularized extreme learning machine for multi-step ahead wind speed prediction[J]. Energies, 11(2): 321.

Zjavka L, 2015. Wind speed forecast correction models using polynomial neural networks[J]. Renew Energy, 83: 998-1006.