Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2020-0006

Vol. **28**, No. 4, pp. 1305-1320
October 2020

A Data-Driven Random Subfeature Ensemble Learning Algorithm for Weather Forecasting

Chen Yu[1], Haochen Li[2], Jiangjiang Xia[3], Hanqiuzi Wen[1,4] and Pingwen Zhang[1,4,*]

[1] *School of Mathematical Sciences, Peking University, Beijing 100871, P.R. China.*
[2] *School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China.*
[3] *Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, P.R. China.*
[4] *National Engineering Laboratory for Big Data Analysis and Application, Peking University, Beijing 100871, P.R. China.*

Received 10 January 2020; Accepted (in revised version) 20 April 2020

Abstract. In this paper, the RSEL (Random Subfeature Ensemble Learning) algorithm is proposed to improve the forecast results of weather forecasting. Based on the classical machine learning algorithms, RSEL algorithm integrates random subfeature selection and ensemble learning combination strategy to enhance the diversity of the features and avoid the influence of a small number of unstable outliers generated randomly. Furthermore, the feature engineering schemes are designed for the weather forecast data to make full use of spatial or temporal context. RSEL algorithm is tested by forecasting the wind speed and direction, and it improves the forecast accuracy of traditional methods and has good robustness.

AMS subject classifications: 62P12, 86A10, 93B15, 97M10

Key words: Weather forecasting, ensemble learning, machine learning, feature engineering.

1 Introduction

Weather forecasting is closely related to various fields, including agriculture, transportation, industry, and energy. In recent years, weather forecasting industry has developed rapidly, which mainly relies on the better theory, the updating of numerical weather prediction (NWP), the increase in the number and accuracy of meteorological observatories, and the improved computational power [1]. A variety of weather prediction methods have been developed in the literature, and they are generally classified into physical methods, statistical methods, machine learning methods, and hybrid methods [2].

*Corresponding author. *Email addresses:* pzhang@pku.edu.cn (P. Zhang), yuchen1995@pku.edu.cn (Y. Chen), lihaochen_bjut@sina.com (H. Li), xiajj@tea.ac.cn (J. Xia), qiuzi.wh@pku.edu.cn (H. Wen)

The physical methods based on the NWP model, which simulates the overall trend of atmospheric motion by solving atmospheric physical equations [3]. However, The NWP models have deficiencies, such as the adaptability of physical equations to local alpine areas, not enough spatial and temporal resolution and bad results of nowcasting and short-term forecasting [4]. Global NWP models include the European Centre for Medium-Range Weather Forecasts (ECMWF), the Global Forecast System (GFS), the Integrated Forecast Model (IFS), etc [5–7]. The statistical method utilizes historical observation data to establish a statistical model for training, which is suitable for short-term prediction. Commonly used statistical methods are Model Output Statistics (MOS) [8–11], Analog Ensemble (ANEN) [12, 13], Kalman Filter (KF) [14, 15] and Markov Chain models [16, 17]. Statistical methods are not available for medium and long-term forecasting, and these methods are not suitable for solving the problem of large data volume.

Machine learning methods can deal with big data in meteorological fields, such as meteorological observations and NWP data. There have been many applications of machine learning in meteorological science [18–20]. The features of big data are diverse in machine learning, thus how to extract useful information from the ever-increasing stream of geoscience data and how to obtain effective features from the NWP models are unavoidable problems [21]. But researches on feature engineering in weather forecasting have received little concerns [22, 23]. Li et al. (2019) proposed model output machine learning (MOML) method to process spatiotemporal features and solved the grid temperature forecasting problem [24]. Nevertheless, due to the spatial and temporal complexity of weather forecasting, it is difficult for current methods to give an optimal scheme directly. A new approach is a hybrid model, coupling physical NWP models with the versatility of data-driven machine learning [21]. Most of the existing hybrid models only mix several statistical methods with weighted strategies and do not form an integrated machine learning algorithm [25–27]. Thus, these methods lack the general optimal strategy. Ensemble learning achieves learning tasks by building and combining multiple base learners. It has superior generalization than a single learner. The representative methods of ensemble learning include boosting and bagging [28–30].

In this paper, an innovative random subfeature ensemble learning algorithm (RSEL) is proposed for weather forecasting. RSEL is a data-driven hybrid ensemble learning algorithm, it brings forth new ideas in the feature engineering scheme and the strategy of ensemble learning algorithm, which also couples the NWP model data and the observational data. To test the application in practical problems, we applied the RSEL algorithm to forecast the wind speed and wind direction at two weather stations that are located in the alpine region, and focused on the next 12-240 h forecasting results. We performed experiments to verify the root mean square error and forecast accuracy of these results and compared them with the ECMWF model, the classical multivariate linear MOS algorithm [10], and MOML algorithm, which has certain innovative meanings [24].

The remainder of the paper is organized as follows. In Section 2, the data concerned in this study are described. Feature engineering scheme and random subfeature ensemble learning algorithm are proposed in Section 3 and Section 4 respectively. Section 5 gives the case studies of RSEL algorithm. The conclusions and future work are finally drawn in Section 6.

2 Data

Meteorological data are typical multi-modal data, mainly including meteorological observation data and NWP model data. These two types of data are data from different sources, with their own data structures and physical variables. The meteorological observation data are local high-frequency data, including several observable meteorological indicators, such as temperature, pressure, wind speed and precipitation. The characteristics of the observational data are real and the time interval is short, but only the information of one weather station on the ground is included, and the spatial and the surrounding information is not included, and only historical data exists. The NWP model data are low-frequency data for global medium-and long-term weather forecasting, which contains numerical prediction grid results of the predictors for the next few days. The advantages of the NWP model data are that these data are time series data covering the region and there are many predictors, such as 2 metre temperature, convective available potential energy, 10 meter U wind component and 10 meter V wind component. However, the time interval is large, and the grid points can not be accurately matched to the weather stations, which may cause errors. These multi-modal data require the algorithm to get useful information from big data and combine them into appropriate formats.

3 Feature engineering

It is difficult to match the multi-modal data of weather forecasting directly, thus feature engineering is the first and an important step before the forecasting method [24]. We constructed a special feature engineering scheme to ensure that the method can better couple the NWP model data and observation data.

3.1 Database

Fig. 1 shows the typical database of weather forecasting at a weather station. The database involves the model data of nine grid points around the weather station and the observational data of the weather station, which are called the original model dataset and the original observational dataset, respectively. The model data of nine grid points add spatial information around the weather station to the original model dataset.

Model dataset. The original model grid data (the nine blue cubes in Fig. 1) is a high-dimensional array. The model output of the predictors on a certain day at each grid point is called a sample, and there are L samples. Each sample has T forecasting time series and M predictors.

In fact, the model dataset is a 5-dimensional $T\times L\times M\times 3\times 3$ array for wind forecasting at a weather station, which contains only nine grid points around the weather station. The coordinate of the closest grid point to the weather station is denoted by (i,j), and the model dataset for this weather station is recorded as $\boldsymbol{X}_0^{(i,j)}$. $\boldsymbol{X}_0^{(i,j)}$ can be written as

$$\boldsymbol{X}_0^{(i,j)}=\left\{x_{t,l,m}^{(p,q)}\right\}_{t=1,2,\cdots,T,\ l=1,2,\cdots,L,\ m=1,2,\cdots,M}^{p=i-1,i,i+1,\ q=j-1,j,j+1}. \tag{3.1}$$

Observation dataset. The observation dataset (the green rectangle in Fig. 1) for each weather station is a $K\times N$ matrix, where the number of rows K is the original observation data with an interval of 1 hour in the corresponding the L-day model dataset, and the number of columns N is N meteorological variables. The observational dataset for this weather station is recorded as $\boldsymbol{Z}_0$. $\boldsymbol{Z}_0$ can be written as

$$\boldsymbol{Z}_0=[z_{k,n}]_{k=1,2,\cdots,K,\ n=1,2,\cdots,N}. \tag{3.2}$$

3.2 Machine learning datasets

We constructed three suitable datasets $D_d=(\boldsymbol{X}_d,\boldsymbol{Y}_d)$ for machine learning algorithm training by feature engineering from the model dataset $\boldsymbol{X}_0^{(i,j)}$, where $\boldsymbol{X}_d$ are features, $\boldsymbol{Y}_d$ are labels and $d=1,2,3$. The labels are all extracted from the observational dataset. For the forecast lead time t, set n_0 as the variable to be predicted and k_l^t is the selected L samples, that is

$$\boldsymbol{Y}_0=[y_{t,l}]_{t=1,2,\cdots,T,\ l=1,2,\cdots,L}\triangleq\left[z_{k_l^t,n_0}\right]_{l=1,2,\cdots,L,\ t=1,2,\cdots,T}. \tag{3.3}$$

Thus, the labels of the three machine learning datasets are equal to $\boldsymbol{Y}_0$, i.e. $\boldsymbol{Y}_1=\boldsymbol{Y}_2=\boldsymbol{Y}_3=\boldsymbol{Y}_0$. The features of these three datasets are different according to the diverse ways of feature engineering.

Dataset 1 directly reshaped the model dataset into the features $\boldsymbol{X}_1$, of size $T\times L\times(9M)$. The dimensions represent T forecast lead times, L samples, and $9M$ predictors, where $9M$ is reshaped by $M\times 3\times 3$. Dataset 1 considers the spatial context in the database without temporal information. The features $\boldsymbol{X}_1$ of $D_1=(\boldsymbol{X}_1,\boldsymbol{Y}_1)$ is denoted as

$$\boldsymbol{X}_1=\{x_{t,l,m^{p,q}}\}_{t=1,2,\cdots,T,\ l=1,2,\cdots,L,\ m=1,2,\cdots,M}^{p=i-1,i,i+1,\ q=j-1,j,j+1}. \tag{3.4}$$

Dataset 2 adds time series of the model dataset to the features. The size of $\boldsymbol{X}_2$ in $D_2=(\boldsymbol{X}_2,\boldsymbol{Y}_2)$ is $T\times L\times[(9M)\times S]$. The additional dimension S represents that dataset 2 adds the model prediction results with the $S-1$ time series before the forecast time to the feature data. Therefore, the feature number of dataset 2 is $(9M)\times S$, and the features $\boldsymbol{X}_2$ of D_2 is denoted as

$$\boldsymbol{X}_2=\left\{x_{t,l,m_s^{p,q}}\right\}_{t=1,2,\cdots,T,\ l=1,2,\cdots,L,\ m=1,2,\cdots,M,\ s=1,2,\cdots,S}^{p=i-1,i,i+1,\ q=j-1,j,j+1}. \tag{3.5}$$

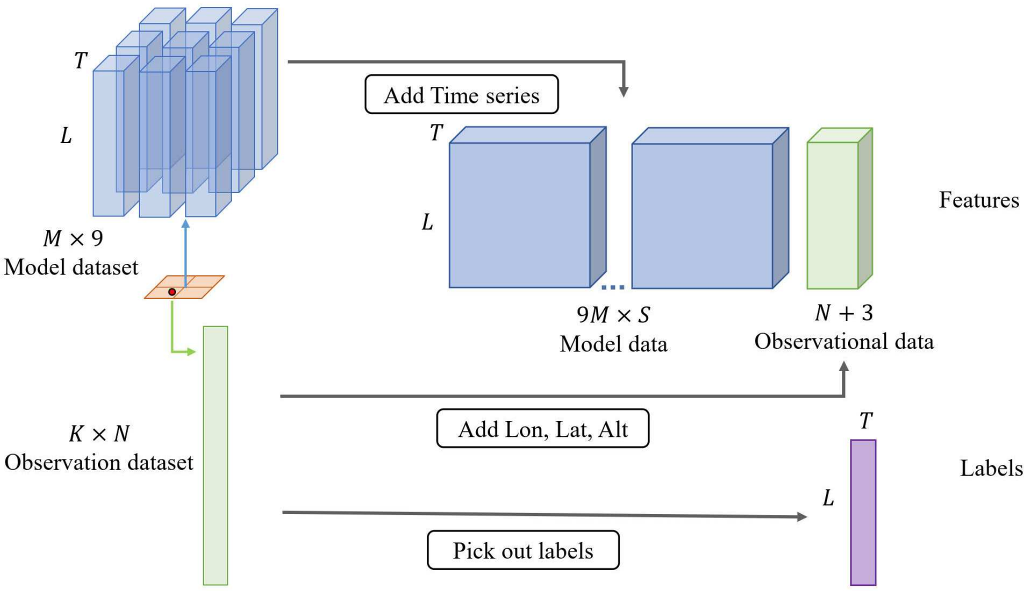


Figure 1: Database of multi-modal data at a weather station is on the left side. The red point is the weather station that needs to be forecasted. The orange 3×3 grid points are the grid point closest to this weather station and the surrounding grid points. Each blue cuboid is the model output data of each grid point, and the green rectangle is the observational data of this site. The machine learning datasets (take dataset 3 as an example) is on the right side. The blue cuboids and the green cuboid on the right are the features, and the purple rectangle is the labels.

Dataset 3 involves observational data and its geographical information, such as the longitude, the latitude and the altitude of the observational station. The size of $\boldsymbol{X}_3$ in $D_3 = (\boldsymbol{X}_3, \boldsymbol{Y}_3)$ is $T\times L\times[(9M)\times S+N+3]$, where

$$\boldsymbol{X}_3 = \{x_{t,l,c}\}_{t=1,2,\cdots,T,\ l=1,2,\cdots,L,\ c\in\{m_s^{p,q},n,Lon,Lat,Alt\}^{p=i-1,i,i+1,\ q=j-1,j,j+1}_{m=1,2,\cdots,M,\ s=1,2,\cdots,S,\ n=1,2,\cdots,N}}. \tag{3.6}$$

Its first two dimensions represent the same meaning as dataset 1 and 2, and the third dimension adds $x_{t,l,n}$ that are selected from the observation dataset $\boldsymbol{Z}_0$, where $x_{t,l,n} \triangleq z_{\hat{k}_l^t,n}$. The historical observation data has N features, which are N meteorological observations. Add the latest observational data and the three geographic features, the feature number of dataset 3 is $(9M)\times S+N+3$. Fig. 1 shows the diagram of dataset 3.

Remark 3.1. For example, assuming that we want to forecast the wind that is offset by 12 hours from UTC 0000, only the model prediction results at this forecast time (UTC 1200) are taken as features in dataset 1. Dataset 2 adds the model prediction results of 3 hours, 6 hours and 9 hours offset (UTC 0300, 0600 and 0900) to the feature data. Based on dataset 2, dataset 3 adds the latest observational data and geographic features.

4 RSEL method

We adopted the integration idea in machine learning and designed a data-driven random subfeature ensemble learning algorithm (RSEL), a novel ensemble learning algorithm. RSEL adopts our new feature engineering scheme and integrates random subfeature selection and a novel ensemble learning combination strategy to get an optimal forecasting solution.

Because the samples from many years ago may differ greatly from those from today in data quality and structure, there are not enough samples but many features, especially the number of features further increased after feature engineering. It is not very effective to apply a machine learning algorithm directly to such datasets, so we proposed a new bootstrap aggregating (Bagging) method, namely random subfeature selection. Random subfeature selection does not sample the samples of the machine learning datasets, but sample the features of the machine learning datasets to obtain a random subfeature dataset. We applied several machine learning algorithms to these random subfeature datasets to obtain preliminary forecasts, and then repeated the process of random subfeature selection and machine learning training.

The forecasting problems can be divided into regression and classification problems. Furthermore, we proposed two ensemble learning combination strategies based on the

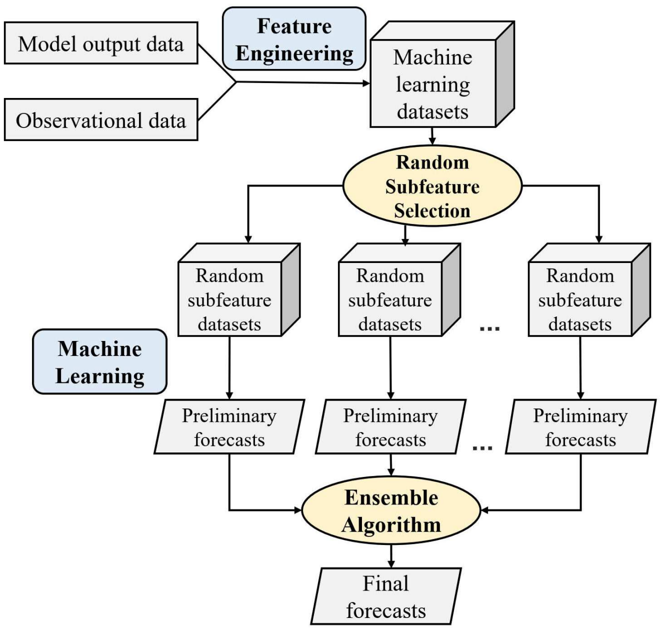


Figure 2: The flow diagram of RESL. The grey rectangle represents the data, the grey cube represents the dataset, the grey parallelogram represents the forecast value, the yellow ellipse represents the algorithm module, and the blue rounded rectangle represents the operation.

preliminary forecast of each machine learning algorithm, namely, adopting the median strategy for regression problems and adopting the majority voting strategy for classification problems. The new ensemble learning combination strategy can avoid the influence of a small number of unstable outliers generated randomly. Finally, the results of random subfeature ensemble learning algorithm are obtained, and the optimal ensemble algorithm scheme is given according to the specific problems and the loss functions. Fig. 2 shows the flow diagram of random subfeature selection in RSEL algorithm. The RSEL algorithm integrates the advantages of model and observation data by feature engineering and enhances the diversity by random subfeature selection. Moreover, the RSEL algorithm enhances the robustness by using the ensemble learning combination strategies. The RSEL algorithm is illustrated by the pseudo-code in Algorithm 1, where the RandomSelect and Map functions work together to perform the random subfeature selection. RandomSelect function samples several feature indexes from all features, and Map function maps this sampling result to the whole dataset. The training set and test set of dataset 1, 2, or 3 are

$$D=\{(\boldsymbol{X},\boldsymbol{y})\}\triangleq\{(\boldsymbol{X}_i,\boldsymbol{y}_i)\,|\,t=t_0\}_{i=1,2,3}=\{(x_{t,l,c},y_{t.l})\}_{l=1,2,\cdots,L_1,\ c=1,2,\cdots,C}^{t=t_0}. \tag{4.1}$$

Algorithm 1 Random subfeature ensemble learning algorithm.

Require:

Training set $D=\{(\boldsymbol{X},\boldsymbol{y})\}$, Test set $\hat{D}=\{(\hat{\boldsymbol{X}},\hat{\boldsymbol{y}})\}$,
Base learners G, number of basic learner categories Λ, maximum iterations Ω,
Number of features d, number of subfeatures d'.

Ensure:

Forecast results of ensemble algorithm $H(\hat{\boldsymbol{X}})$.

```
for ω = 1,2,···,Ω do
    F^ω = RandomSelect(d,d')
    D^ω = Map_{F^ω}(D)
    for λ = 1,2,···,Λ do
        g_{ω,λ} = G_λ(D^ω)
    end for
end for
return H(X̂) = h_{argmin_k Loss(h_k(X̂),ŷ)}(X̂), where
```

1: **for** $\omega=1,2,\cdots,\Omega$ **do**
2: $F^\omega=\text{RandomSelect}(d,d')$
3: $D^\omega=\text{Map}_{F^\omega}(D)$
4: **for** $\lambda=1,2,\cdots,\Lambda$ **do**
5: $g_{\omega,\lambda}=G_\lambda(D^\omega)$
6: **end for**
7: **end for**
8: **return** $H(\hat{\boldsymbol{X}})=h_{\arg\min_k \text{Loss}(h_k(\hat{\boldsymbol{X}}),\hat{\boldsymbol{y}})}(\hat{\boldsymbol{X}})$, where

regression: $h_k(\hat{\boldsymbol{X}})=\text{Median}\Big(\{g_{\omega,\lambda}(\text{Map}_{F^\omega}(\hat{\boldsymbol{X}}))\}_{\omega=1}^{\Omega}\Big)$.

classification: $h_k(\hat{\boldsymbol{X}})=\text{MajorityVote}\Big(\{g_{\omega,\lambda}(\text{Map}_{F^\omega}(\hat{\boldsymbol{X}}))\}_{\omega=1}^{\Omega}\Big)$.

In fact, in these machine learning algorithms, no matter whether it is the regression problem or the classification problem, some Boosting algorithms perform well and become the optimal ensemble algorithm scheme. So we can understand the RSEL algorithm as a novel ensemble learning algorithm integrating Bagging and Boosting algorithms. It has three advantages:

1. Updating the traditional Bagging of the samples to Bagging of the features enhances the diversity of the features.

2. The mature Boosting algorithms are used as the base learners of the new Bagging method, which ensure the stable improvement of the accuracy of the forecast results.

3. At the end of the algorithm, the median or majority voting combination strategy can avoid the influence of a small number of unstable outliers generated by random feature selection.

5 Simulation and case study

To test the application of RSEL algorithm in practical problems, we applied RSEL algorithm to forecast the wind speed and wind direction at Yanqing and Foyeding weather stations that are located in Beijing, and focused on the next 12-240 h forecasting results. The altitude of Foyeding weather station is high, and the accuracy of numerical weather prediction near this weather station is always inaccurate. Yanqing weather station is located with the competition zones of the 2022 Winter Olympic Games, thus it requires very high accuracy in forecasting wind direction and wind speed. Therefore, it is representative and practical to choose these two weather stations to test the effectiveness of our method.

The meteorological observation we used include the observational data from January, 2015 to November, 2017 at the weather stations. Hourly observational data of each weather station contains six weather elements: temperature, pressure, relative humidity, wind direction, wind speed, precipitation. The NWP model data used in this paper are the numerical forecast products of ECMWF from January, 2015 to October, 2017. The range is 35 to 45 degrees north latitude and 110 to 120 degrees east longitude. In fact, the valuable data are some grid data near the weather stations from initialized at UTC 1200 up to lead time of 240 hours. The forecast interval of the first 72 hours is 3 hours, that of 78-240 hours is 6 hours, and the spatial resolution on the ground is $0.125^\circ \times 0.125^\circ$, and at high altitude is $0.25^\circ \times 0.25^\circ$. After deleting some unnecessary predictors (e.g. land-sea mask), the NWP model data include 44 predictors (listed in Table 1).

We set the samples from January 16, 2015, to October 29, 2016, as the training set and the samples from October 30, 2016, to October 30, 2017, as the test set. The specific values of the indicators in the database and the machine learning datasets (Eqs. (3.1)-(3.6)) are $L=1019$, $M=44$, $T=49$, $N=6$, $K=8152$, $S=4$. The base learners used in this case involve Least Absolute Shrinkage and Selectionator Operator (LASSO) [31], Random Forest (RF) [32], Gradient Boosting Decision/Regression Tree (GBDT/GBRT) [33, 34], and eXtreme Gradient Boosting (XGB) [35], including linear algorithms and nonlinear algorithms. We performed experiments to verify the root mean square error and forecast

Table 1: Information of the predictors taken from ECMWF.

Predictors		
100 meter U wind component	Low cloud cover	Sea surface temperature
100 meter V wind component	Large-scale precipitation	Temperature [500 hPa]
10 meter U wind component	Mean sea level pressure	Temperature [850 hPa]
10 meter V wind component	Potential vorticity [1000 hPa]	Total cloud cover
2 meter dewpoint temperature	Potential vorticity [500 hPa]	Total column water
2 metre temperature	Potential vorticity [850 hPa]	Total column water vapour
Convective available potential energy	Specific humidity [1000 hPa]	Total precipitation
Divergence [1000 hPa]	Specific humidity [500 hPa]	U wind component [500 hPa]
Divergence [500 hPa]	Specific humidity [850 hPa]	U wind component [850 hPa]
Divergence [850 hPa]	Relative humidity [1000 hPa]	V wind component [500 hPa]
Zero Degree Level	Relative humidity [500 hPa]	V wind component [850 hPa]
Forecast albedo	Relative humidity [850 hPa]	Vertical velocity [1000 hPa]
Geopotential height [1000 hPa]	Snow depth	Vertical velocity[500 hPa]
Geopotential height [500 hPa]	Snowfall	Vertical velocity [850 hPa]
Geopotential height [850 hPa]	Skin temperature	

accuracy of these results and compared them with the ECMWF model, the classical multivariate linear MOS algorithm [8–10], and MOML algorithm [24].

Remark 5.1. Applying RSEL to the datasets constructed in Section 3.2, we can get various algorithm models. For example, RSEL apply the GBDT algorithm to dataset 2 to get RSEL_GBDT_2 model.

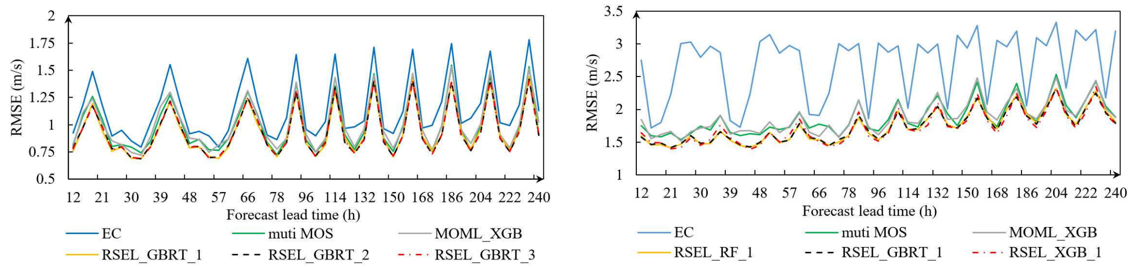
5.1 The regression problem of wind speed forecasting

Wind speed is a continuous variable, so the wind speed forecast problem is suitable to be resolved as a regression problem. The base learners used by RSEL algorithm include GBRT, LASSO, RF, and XGB, which means $\Lambda = 4$. We choose the root mean square error as the loss function of RSEL algorithm.

Table 2 show the performance of wind speed forecast at Yanqing weather station and Foyeding weather station, all the RSEL algorithms can improve the wind speed forecast results of the ECMWF model, multivariate linear MOS and MOML algorithms quite well in the sense of annual mean. The altitude of Foyeding weather station is higher than that of Yanqing weather station, thus the forecast results of ECMWF model is poor that provide RSEL algorithm with significant upside. Selecting the boosting algorithm (GBRT, XGB) as the base learner is superior to the other machine learning algorithms used in RSEL algorithms. It proves that the RSEL algorithm with the appropriate Boosting method can improve the forecast results very well in the wind speed forecasting problem. The root mean square errors (RMSE) of these feature engineering approaches have little difference, so the simplest dataset (dataset 1) is more suitable for wind speed forecasting.

Table 2: The average RMSE of the next 12-240 h wind speed forecast results at Yanqing and Foyeding weather stations.

Station	Evaluation	Dataset	RSEL				EC	MOS	MOML
			GBRT	LASSO	RF	XGB			XGB
Yanqing	RMSE (m/s)	1	0.9373	0.9477	0.9394	0.9436	1.1482	0.9884	1.0079
		2	0.9369	0.9514	0.9427	0.9443			
		3	0.9371	0.9523	0.9426	0.9424			
Foyeding	RMSE (m/s)	1	1.7293	1.7439	1.7300	1.7232	2.6883	1.8676	1.8727
		2	1.7324	1.7709	1.7444	1.7438			
		3	1.7326	1.7694	1.7438	1.7457			

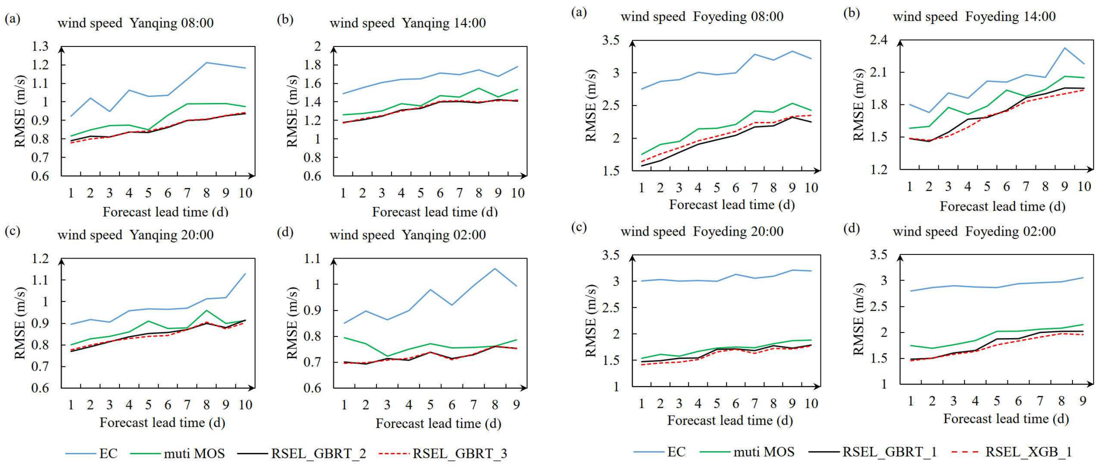


(a) Yanqing weather station (b) Foyeding weather station

Figure 3: RMSE of wind speed forecasting at Yanqing and Foyeding weather stations. The RSEL results are better than multivariate linear MOS algorithm and MOML algorithm.

Fig. 3 shows the variation trend of the RMSE of 12-240 h forecast of several well-performing RSEL algorithms. It can be seen that the RMSE of RSEL algorithms are less than that of the ECMWF model, multivariate linear MOS and MOML algorithms. These figures also show that the RMSE does not increase with the increase of forecast lead time in the medium and long-term forecast, and the values of RMSE have periodicity with the increase of forecast lead time, and the period is about 24 hours. So we compared the results in daily BJT 08:00, 14:00, 20:00 and 02:00 in Fig. 4. The results of RSEL algorithm for wind speed forecast at Yanqing weather station is the best in daily BJT 02:00 and 08:00, and the RMSE of RSEL_GBRT_2 in the daily BJT 02:00 and 08:00 decrease by 23.042% and 19.729% than that of ECMWF model, respectively. The RMSE of RSEL_XGB_1 for 10 days at Foyeding weather station are minimum in daily BJT 02:00 and 20:00, which decrease by 40.514% and 47.784% than that of ECMWF model, respectively. Thus, these two algorithms are the optimal algorithms for wind speed forecasting at the two weather stations.

Remark 5.2. In the simulation experiment, the algorithm is not sensitive to the number of iterations (Ω) if Ω is large enough and we showed the results of $\Omega = 60$. Because of random sampling, the algorithm is not sensitive to the number of subfeatures d'. In this experiment, d' is set at the same magnitude as M, and the result of $d' = 40$ is shown.



(a) Yanqing weather station (b) Foyeding weather station

Figure 4: RMSE of wind speed forecasting at Yanqing and Foyeding weather stations by time. The results of RSEL algorithm for wind speed forecast is the best in daily BJT 02:00, and can fully exceed multivariate linear MOS and ECMWF model.

5.2 The classification problem of wind direction forecasting

The data of wind direction has a large data range, and few categories, thus the wind direction forecast problem is suitable to be solved as a classification problem. The classification problem is not suitable to be evaluated by root mean square error, so we calculate the values of forecast accuracy in this subsection. The wind direction is divided into eight categories and numbered from 0 to 7, and the wind direction forecast accuracy (denoted by F_a) is defined by the forecast score matrix (Table 3) [36].

Table 3: The forecast score matrix of wind direction.

Srore	0.0-22.5°, 337.5-360.0°	22.5-67.5°	67.5-112.5°	112.5-157.5°	157.5-202.5°	202.5-247.5°	247.5-292.5°	292.5-337.5°
0.0-22.5°, 337.5-360.0°	1	0.6	0	0	0	0	0	0.6
22.5-67.5°	0.6	1	0.6	0	0	0	0	0
67.5-112.5°	0	0.6	1	0.6	0	0	0	0
112.5-157.5°	0	0	0.6	1	0.6	0	0	0
157.5-202.5°	0	0	0	0.6	1	0.6	0	0
202.5-247.5°	0	0	0	0	0.6	1	0.6	0
247.5-292.5°	0	0	0	0	0	0.6	1	0.6
292.5-337.5°	0.6	0	0	0	0	0	0.6	1

Table 4: The average forecast accuracy of the next 12-240 h wind direction forecast results at Yanqing and Foyeding weather station.

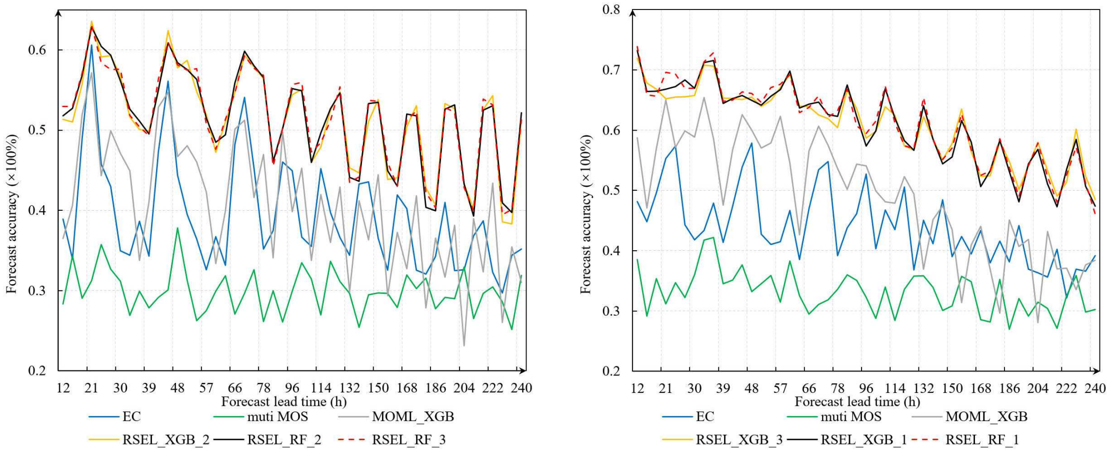
Station	Evaluation	Dataset	RSEL GBDT	RSEL RF	RSEL XGB	EC	MOS	MOML XGB
Yanqing	**Fa** ($\times$100%)	**1**	0.5026	0.5146	0.5090	0.3937	0.2991	0.4045
		2	0.5080	0.5150	0.5135			
		3	0.5096	0.5154	0.5126			
Foyeding	**Fa** ($\times$100%)	**1**	0.6090	0.6131	0.6099	0.4394	0.3312	0.4902
		2	0.6081	0.6089	0.6082			
		3	0.6056	0.6081	0.6094			

For the fixed forecast lead time, the formula of forecast accuracy is Eq. (5.1), where SC_l are the forecast scores that can be found in Table 3, L_{test} is the number of samples in test set, and the forecast score is shown as a percentage and recorded as the forecast accuracy.

$$\mathrm{F_a} = \sum_{l=1}^{L_{test}} \frac{SC_l}{L_{test}} \times 100\%. \tag{5.1}$$

The base learners used by RSEL algorithm include GBDT, RF and XGB, which means $\Lambda=3$, the number of subfeatures $d'=40$ and the maximum iterations $\Omega=60$. The loss function used in this problem is the forecast accuracy. Table 4 show the forecast accuracy of wind direction forecast at the two weather station. The results show that RSEL algorithms has got good classification accuracy for the test set. Both multivariate linear MOS and MOML algorithms are regression algorithms, so their forecast accuracy is not high, which verifies that the wind direction forecasting problem is more suitable to be solved as a classification problem. In RSEL algorithms, selecting dataset 1 and 3 and RF or XGB algorithm is superior to the others in this study. It proves that the RSEL algorithm which combines the new Bagging method with the appropriate Boosting method can improve the accuracy very well in the classification problem.

Fig. 5 shows the variation trend of the forecast accuracy of 12-240 h wind direction forecasting. The figures show the stability of the RSEL algorithm. On the whole, the forecast accuracy of the algorithms have periodicity and decrease with the increase of forecast lead time. Fig. 6 shows the forecast accuracy of the short-term wind direction forecasting. It can be seen in Figs. 5 and 6 that the dataset 3 with observation data is helpful to improve the accuracy of short-term forecast, but has little effect on long-term forecast, which illustrates the necessity of constructing diverse machine learning datasets. We can conclude that RSEL_RF_3 and RSEL_RF_1 are the optimal algorithms for wind direction forecasting at Yanqing and Foyeding weather stations, respectively. Actually, the average forecast accuracy of RSEL_RF_3 and RSEL_RF_1 for 10 days is 21.628% and 28.186% higher than that of multivariate linear MOS algorithm and 11.088% and 12.286% higher than that of MOML algorithm with XGBoost, respectively at these two weather stations.



(a) Yanqing weather station

(b) Foyeding weather station

Figure 5: Forecast accuracy of wind direction forecasting at Yanqing and Foyeding weather stations. The RSEL results are better than multivariate linear MOS algorithm and MOML algorithm.

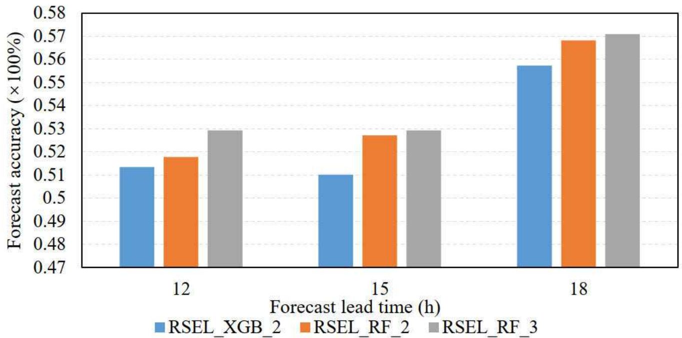


Figure 6: Forecast accuracy of 12-18 h wind direction forecasting at Yanqing weather stations. The dataset 3 with observation data has the advantages of the short-term wind direction forecasting.

6 Conclusions

In this paper, a novel RSEL algorithm is proposed to solve the problems in weather forecasting. In fact, this algorithm can be used to solve similar problems involving multimodal data. With feature engineering for weather forecast data, RSEL integrates random subfeature selection, a new Bagging method to enhance the diversity of the features, and a novel ensemble learning combination strategy to avoid the influence of a small number of unstable outliers generated randomly. The case studies show that regardless of regression or classification, RSEL outperforms other candidate algorithms for wind forecasting. In the sense of annual mean, the RSEL algorithms can improve the wind forecasting results of the ECMWF model at Yanqing and Foyeding weather stations quite well. The advantages of RSEL in the classification problem is greater than that in the regression problem, and if time series and historical observation information are included in the feature, the result can be improved. The forecast accuracy of wind speed increases considerably before dawn and that of wind direction improved greatly throughout the whole day. Boosting algorithms are suitable to be the base learners of RSEL, which ensure the stable improvement of the accuracy of the forecast results.

In summary, RSEL has the ability to improve the results of weather forecasting, and it has better performance than multivariate linear MOS algorithm and MOML algorithm. In addition, as a post-processing method, RSEL can be applied to the weather consultation and other problems involving multi-modal data. This approach has good application prospects and can greatly reduce manpower consumption in the consultation and improve the forecast accuracy. Considering the practical application, the machine learning algorithm in this paper adopts the mature eXtreme Gradient Boosting algorithm and Random Forests algorithm, which do not need a lot of parameter adjustment and are easy to use. Constructing a more accurate and efficient ensemble learning algorithm to solve the problem is our later research work.

Acknowledgments

The authors would like to express sincere gratitude to Lizhi WANG, lve WU and Xiao LOU for unpublished data; and Chongping JI and Yingxin ZHANG for professional guidance. This work is supported by the National Key Research and Development Program of China (Grant Nos. 2017YFC0209804 and 2018YFF0300104), Beijing Academy of Artificial Intelligence (BAAI), the National Natural Science Foundation of China (Grant No. 11421101) and the Open Research Fund of Shenzhen Research Institute of Big Data (Grant No. 2019ORF01001).

References

[1] P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[2] J. Mendes, J. Sumaili, R. Bessa, H. Keko, V. Miranda, A. Botterud, and Z. Zhou. Very short-term wind power forecasting: State-of-the-art. Technical report, Argonne National Laboratory (ANL), 2014.

[3] S. Al-Yahyai, Y. Charabi, and A. Gastli. Review of the use of numerical weather prediction (nwp) models for wind energy assessment. *Renewable and Sustainable Energy Reviews*, 14(9):3192–3198, 2010.

[4] Q. Hu, P. Su, D. Yu, and J. Liu. Pattern-based wind speed prediction based on generalized principal component analysis. *Sustainable Energy, IEEE Transactions on*, 5:866–874, 07 2014.

[5] T. Palmer, J. Barkmeijer, R. Buizza, and T. Petroliagis. The ecmwf ensemble prediction system. *Meteorological Applications*, 4:301–304, 12 1997.

[6] N. P. Wedi and P. K. Smolarkiewicz. A framework for testing global non-hydrostatic models. *Quarterly Journal of the Royal Meteorological Society*, 135(639):469–484, 2010.

[7] J. Cote, S. Gravel, A. Methot, A. Patoine, M. Roch, and A. Staniforth. The operational cmc cmrb global environmental multiscale (gem) model. part i: Design considerations and formulation. *Monthly Weather Review*, 126(6):1373–1395, 1998.
[8] H. R. Glahn and D. A. Lowry. The use of model output statistics (mos) in objective weather forecasting. *Journal of Applied Meteorology*, 11(8):1203–1211, 1972.
[9] B. Glahn, M. Peroutka, and J. Wiedenfeld. Mos uncertainty estimates in an ensemble framework. *Monthly Weather Review*, 137(1):246–268, 2009.
[10] X. N. Zhang, J. Cao, S. Y. Yang, and Q. I. Ming-Hui. Multi-model compositive mos method application of fine temperature forecast. *Journal of Yunnan University*, 33(1):67–66, 2011.
[11] Q. Wu, M. Han, H. Guo, and T. Su. The optimal training period scheme of mos temperature forecast. *Journal of Applied Meteorological Science*, (4):426–434, December 2016.
[12] L. D. Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight. Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10):3498–3516, 2013.
[13] I. O. Plenkovic, L. Delle Monache, K. Horvath, and M. Hrastinski. Deterministic wind speed predictions with analog-based methods over complex topography. *Journal of Applied Meteorology and Climatology*, 57(9):2047–2070, 2018.
[14] L. Delle Monache, T. Nipen, Y. Liu, G. Roux, and R. Stull. Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, 139(11):3554–3570, 2011.
[15] A. Pelosi, H. Medina, J. V. D. Bergh, S. Vannitsem, and G. B. Chirico. Adaptive kalman filtering for postprocessing ensemble numerical weather predictions. *Monthly Weather Review*, 145(12), 2017.
[16] A. Carpinone, R. Langella, A. Testa, and M. Giorgio. Very short-term probabilistic wind power forecasting based on markov chain models. In *IEEE International Conference on Probabilistic Methods Applied to Power Systems*, 2010.
[17] Z. Song, Y. Jiang, and Z. Zhang. Short-term wind speed forecasting with markov-switching model. *Applied Energy*, 130(5):103–112, 2014.
[18] L. Valliappa, G. Eric, M. Amy, and T. Martin. *Machine Learning and Data Mining Approaches to Climate Science*. Springer International Publishing, 2015.
[19] S. E. Haupt and B. Kosovic. Big data and machine learning for applied weather forecasts: Forecasting solar power for utility operations. In *Computational Intelligence, 2015 IEEE Symposium*, pages 496–501, 2016.
[20] W. Woo and W. Wong. Operational application of optical flow techniques to radar-based rainfall nowcasting. *Atmosphere*, 8(8(3)):48, 2017.
[21] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and M. Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204, 02 2019.
[22] D. Liu, J. Wang, and H. Wang. Short-term wind speed forecasting based on spectral clustering and optimised echo state networks. *Renewable Energy*, 78:599–608, 2015.
[23] Q. He, J. Wang, and H. Lu. A hybrid system for short-term wind speed forecasting. *Applied Energy*, 226:756–771, 09 2018.
[24] H. Li, C. Yu, J. Xia, Y. Wang, J. Zhu, and P. Zhang. A model output machine learning method for grid temperature forecasts in the beijing area. *Advances in Atmospheric Sciences*, 36(10):1089–1103, 10 2019.
[25] H. Liu, H.-Q. Tian, C. Chen, and Y.-f. Li. A hybrid statistical method to predict wind speed and wind power. *Renewable Energy*, 35(8):1857–1861, 08 2010.
[26] A. Kaur, H. T. C. Pedro, and C. F. M. Coimbra. Ensemble re-forecasting methods for en-

hanced power load prediction. *Energy Conversion and Management*, 80:582–590, 04 2014.

[27] H. Liu, H.-q. Tian, and Y.-f. Li. Four wind speed multi-step forecasting models using extreme learning machines and signal decomposing algorithms. *Energy Conversion and Management*, 100:16–22, 08 2015.

[28] L. Breiman. Using iterated bagging to debias regressions. *Machine Learning*, 45(3):261–277, 2001.

[29] R. E. Schapire and Y. Freund. Boosting: Foundations and Algorithms. *Kybernetes*, 42(1):164–166, 2013.

[30] Z. H. Zhou. *Ensemble methods: foundations and algorithms.* CRC press, 2012.

[31] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 01 1996.

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

[33] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[34] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 02 2002.

[35] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016.

[36] C. Gao and J. Zeng. Validation of wind forecast based on MOS method at the Ningde coastal region. *Marine Forecasts*, 35(4):19-26, 08 2018.