International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: 2https://doi.org/10.62051/ijcsit.v2n1.46
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Wind Speed Prediction Method Based on Feature Fusion

Shengjie Zhou, Fei Luo*

College of Software Engineering, Chengdu University of Information Technology, Chengdu, China

ABSTRACT

Wind speed occupies a central position in the study of atmospheric circulation and climate change, playing a crucial role in the accuracy and reliability of meteorological forecasting models. The current meteorological observation network, especially in remote and mountainous areas, faces challenges due to insufficient data resulting from low station density. This limitation hampers a comprehensive understanding of wind field dynamics, subsequently affecting the performance of forecasting models. To enhance the performance of meteorological prediction models, this study adopts an innovative approach that combines boxplot difference indexing and feature-level fusion techniques to filter key predictive features and reduce information redundancy. Furthermore, by integrating spatial location information, the contribution of geographic data to wind field prediction is enhanced, improving the quality of the fused data and thereby increasing the accuracy of model predictions. Experimental comparisons demonstrate that the integrated wind field dataset, post feature fusion, achieves better results across various deep learning models.

KEYWORDS

Wind speed prediction; Multisource data; Feature fusion; Deep learning

1. INTRODUCTION

Wind field forecasting plays a crucial role in human activities such as aviation meteorological services and disaster early warnings. Initially, traditional wind field prediction relied primarily on singular historical wind speed data[1], derived from fixed meteorological stations and high-altitude balloon observations, which are sparsely distributed and difficult to obtain, especially over maritime regions. For areas lacking meteorological stations, data assimilation[2] can fill these regional gaps by interpolating observational data within models, thus enhancing global forecast coverage. Subsequent in-depth research led to Sasaki's[3,4] method of optimizing model parameters or initial values by minimizing error functions, aiming to align model outputs more closely with actual observational data. Then, LeDimet[5] and others introduced variational methods, marking a new chapter in the field by using variational principles and the theory of conjugate equations to process observational data at different times, obtaining an optimal initial field. With the advent of supercomputers, computational capacity has further improved, making it possible to solve entire systems of equations[6].

Advances in meteorological observation techniques, such as the increasing use of weather radars and wind profilers, have made observational data richer and more complex, significantly impacting the accuracy of weather predictions. Data fusion technology has gradually come into public view. Researchers[7] have constructed fused datasets based on various meteorological and environmental data, developing models aimed at enhancing prediction accuracy. Multisource data fusion involves integrating data from diverse channels to create a comprehensive information assembly. In current

research practices, multisource data integration methods are increasingly employed with the aim of synthesizing multidimensional information to achieve more in-depth and reliable research outcomes.

With access to rich meteorological datasets, deep learning techniques have been applied to extract complex information within the data. Yang et al.[8] used a BP neural network method based on historical wind speed data and surrounding airspace wind data for short-term wind speed prediction, proving to be an effective and precise approach. However, this algorithm can easily become trapped in local minima or stasis points, and the network's fixed learning rate leads to slow convergence. To avoid the issue of BP neural networks falling into local optima, researchers have proposed RNN neural networks with connections established between neurons across layers. But RNNs encounter significant difficulties in dealing with long-term dependencies with gradients[9]. To solve this issue of long-term dependencies, Long Short-Term Memory networks were proposed, performing exceptionally well in handling time-series data problems[10]. Zhang[11] utilized LSTM networks to capture historical information from wind speed measurement series. Wang[12] and others introduced a Seq2Seq network with multiple layers of RNNs, incorporating uncertainty quantification into wind speed predictions. Li[13] applied LSTM and Bi-directional Long Short-Term Memory (Bi-LSTM) models with different configurations and activation functions for wind speed prediction, finding that LSTM networks have better application effects in long-term wind forecasting. Chen[14] developed a combined model based on LSTM and BP neural networks, effectively improving the predictive performance of nonlinear wind speed data. Concurrently, SHEN[15] and others combined CNN with LSTM to improve model accuracy, and through comparisons with other models, they validated the superiority of the mixed model in multi-step wind speed prediction. As deep learning technology continues to evolve, it not only enhances the accuracy of wind speed prediction but also paves the way for the development of future meteorological models.

2. METHODS

This paper selects multi-source data: the first is historical observation data, acquired from the National Comprehensive Meteorological Information Sharing Platform, consisting of historical measured data from meteorological stations; the second is forecast data, primarily derived from fine grid data provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Initially, data alignment is necessary to obtain time series data; then, this paper employs the boxplot difference index for feature selection and performs feature-level fusion to acquire a comprehensive wind field dataset. Furthermore, the integration of geographical location information from meteorological stations is introduced to further improve data quality, with the goal of providing richer information to deep learning models.

2.1. Data Alignment

Historical observation data is a type of time series data, which is usually recorded at hourly intervals. Forecast data from ECMWF offers grid data with a spatial resolution of 0.25°×0.25°. ECMWF conducts two forecast updates daily, scheduled at 08:00 and 20:00 (Beijing time), covering a forecast period from 0 to 72 hours, with an interval of every 3 hours. The first step is to perform effective interpolation on grid data to extract time series type data for meteorological observation sites.

In this study, the widely recognized Inverse Distance Weighting (IDW) spatial interpolation technique is adopted as the primary method of interpolation. It aims to estimate attribute values for unsampled points, with its main advantages being the intuitiveness of the algorithm and its sensitivity to local features of the data. The IDW method is based on a key assumption that a known point's influence on an unknown point diminishes as the distance between them increases, and this diminishing influence is inversely proportional to the power of the distance. Within this framework, observation points closer to the target point have greater weights. When applying IDW interpolation, the geographical coordinates and corresponding attribute values of each known data point are first

determined, followed by the selection of an appropriate power index k based on the inverse power law of distance to optimize the weight distribution process, thus accurately estimating the attribute values of unknown posints. This method not only enhances the model's adaptability to the spatial distribution of data but also improves the precision and reliability of the interpolation results. The formula is as follows.

$$v(p) = \frac{\sum_{i=1}^{n} \frac{v_i}{d(p, p_i)^k}}{\sum_{i=1}^{n} \frac{1}{d(p, p_i)^k}} \tag{1}$$

Where v(p) is the attribute value to be interpolated at point p, v_i is the attribute value at known point p_i, and $d(p,p_i)$ is the distance between the interpolation point p and known point p_i. In this study, the power index k is set to 2 to control the influence of distance. The interpolated ECMWF data corresponding to the meteorological elements of the weather stations are extracted and formed into a time series to construct continuous feature data. A 0.25°×0.25° grid refined through interpolation to a 0.125°×0.125° grid means that each grid cell becomes an area approximately 13.875 km × 13.875 km in size. The increased interpolation resolutio

n implies that the data are finer in spatial detail, allowing for a better representation of variations within small regions. On the temporal scale, linear interpolation is used, employing the average of adjacent data points to increase the time resolution from a three-hour to a one-hour interval. Subsequently, the grid data is extracted according to the existing location information of the meteorological stations, ensuring consistency with the format of the historical observation data.

2.2. Feature Selection

The boxplot difference index (I_{bd}) is applied to the field of meteorology to effectively evaluate and select key features within meteorological data. In the context of wind speed elements, I_{bd} helps to distinguish which features contribute most significantly to the variability of wind speed and to the predictive model. For historical observation data, the meteorological elements obtained include: wind speed, wind direction, air pressure, temperature, dew point temperature, relative humidity, and hourly precipitation; for forecast data, the elements obtained are: forecast albedo, convective precipitation, 2-meter dew point temperature, large-scale precipitation, 2-meter temperature, accumulated precipitation, relative humidity, and surface pressure. The I_{bd} for each feature is then calculated by quantifying the differences in their distributions. The formula is as follows, where a larger absolute value of I_{bd} indicates a greater relevance of that feature to the variability of wind speed.

$$I_{bd} = \frac{m_1 - m_2}{\sigma_1 + \sigma_2} \tag{2}$$

In the formula, the ground measured wind speed data samples in the dataset and the non wind speed data samples in the reanalysis data are considered as two datasets. m_1 and m_2 represent the mean values of these two datasets for this feature σ_1 and σ_2 are the standard deviations of the dataset on this feature. The larger the absolute value of I_{bd}, the more relevant the feature is to changes in wind speed

2.3. Feature Fusion

Feature fusion is primarily achieved through two methods: Concat (concatenating different features) and Add (stacking feature vectors), catering to different data processing needs. These methods enable the model to integrate various types of data, such as upper-air pattern forecasts, radar data, and

observation data. Features with a significant I_{bd} are considered more influential and are chosen for fusion. They are selected for further analysis and as inputs for the model.

This study opts for the Concatenation operation to perform feature fusion and simultaneously stitches the geographical location information of the sites as a feature, resulting in a complex data structure with enhanced comprehensiveness. The specific fusion formula is as follows.

$$X^t = \left[X^t_{pos}, X^t_{feature}, X^t_{ecwmf} \right] \tag{3}$$

In the formula, $X^{'}_{feature}$ represents the feature vector of the meteorological element series at time step t; X^t_{pos} is the position feature vector that includes longitude, latitude, and altitude, and its shape matches the dimension of the $X^t_{feature}$ feature vector; X^t_{ecwmf} indicates the feature vector of the upper-air numerical forecast element series at time step t with the features selected using the method from the previous section; ',' represents the Concatenation operation. The time steps for all these features are the same. In this way, a comprehensive feature vector is created at each time step, reflecting both wind speed information and location information.

3. EXPERIMENT

3.1. Calculation result of I_{bd}

For the 14 meteorological elements of historical observation data and model forecast data, the boxplot difference index is calculated for each meteorological element data sample forecasted by the model and the actual wind speed data sample. These are then sorted, and the results are shown in Table 1. Different combinations are selected, for instance, when the combination number is 10, it means that the top ten meteorological elements with the highest I_{bd} values are chosen for feature fusion.

Table 1. Sample Sorting Of I_{bd}

Feature	Meaning	Unit	I_{bd}
FAL	Forecast albedo	-	0.36
CP	Convective precipitation	m	0.40
2D	2 meters dew point temperature	K	0.41
LSP	Large-scale precipitation	m	0.45
2T	2 m temperature	K	0.47
TP	Accumulated precipitation	m	0.50
R	Relative humidity	-	0.52
SP	Surface pressure	Pa	0.62

From the arrangement of this table, it can be observed that surface pressure (SP), relative humidity (R), large-scale precipitation (LSP), and 2-meter temperature (2T) are closely related to the occurrence of high wind weather. Through feature selection, we can reduce data redundancy and as much as possible filter out features closely related to wind field variability for use in wind field forecast models.

3.2. Model Training

To find the optimal K value, it is necessary to determine the features of the dataset before training. This study uses seven features from the historical observational data as a baseline dataset. On this basis, different combination numbers K are selected to choose features from the ECMWF dataset. Subsequently, the feature fusion method from section 3.3 is used to form the fused dataset. At the same time, the fused dataset is applied with min-max normalization, reshaping the input into a tensor format composed of three dimensions: weather variables, time steps, and meteorological stations, which are then fed into the LSTM model for training.

3.3. Evaluation Metrics

This paper selected three key statistical indicators: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The formulas for these metrics are as follows. These indicators are widely used in evaluating the accuracy of forecasting models.

$$e_{MAE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{4}$$

$$e_{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{5}$$

$$e_{RMSE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{6}$$

Here, N represents the number of hourly wind speed data points, x_i is the actual value, and y_i is the predicted value.

4. RESULTS ANALYSIS

To validate the selection of the K value, that is, how the choice of the number of features affects the wind speed prediction performance, the following four experiments were designed. K values from 1 to 5 were selected, and the datasets resulting from the fusion of these five features were tested on LSTM and Bi-LSTM models, respectively. The table below shows the experimental results.

Table 1. Comparison Results Of Various Models

Model	K	MAE	RMSE	MSE
LSTM	1	1.39	1.57	2.21
	2	1.12	1.35	1.58
	3	0.79	1.03	1.22
	4	0.94	1.17	1.34
	5	1.17	1.40	1.52
Bi-LSTM	1	1.23	1.46	1.84
	2	1.09	1.24	1.36
	3	0.73	0.98	1.17
	4	0.91	1.13	1.29
	5	1.12	1.25	1.35

For the LSTM model: The lowest MAE is observed at K=3, indicating that a combination of the top three most correlated features from the ECMWF data yields more accurate predictions. At K=3, the MSE and RMSE also reach their lowest levels, suggesting that this set of features not only minimizes the average error but also reduces the frequency of large prediction errors. For the Bi-LSTM model: The lowest MAE is achieved at K=3, mirroring the LSTM model, which further confirms that K=3 is the optimal choice for feature selection.

Given that the lowest error rates for all metrics in both models occur at K=3, it can be concluded that selecting the top three features based on the boxplot difference index for data fusion provides the most reliable and accurate wind speed predictions for both LSTM and Bi-LSTM models. This suggests that more features do not necessarily enhance the predictive power of the models; instead, they may introduce noise or redundancy

5. CONCLUSION

In this study, we addressed the pivotal role of wind speed within atmospheric and climate research by tackling the challenge posed by sparse meteorological data, particularly in remote regions, which hampers a comprehensive understanding of wind dynamics. To enhance the performance of forecasting models, we implemented an innovative method that merges the boxplot difference index with feature-level fusion techniques to refine essential predictive features, thereby drastically reducing information redundancy and bolstering the contribution of geographical data to wind field predictions. Our research confirmed that the optimal K value for feature fusion on this multisource data is 3. These findings not only propel the field of wind prediction forward but also provide a methodological blueprint for the effective utilization of deep learning in meteorological applications.

ACKNOWLEDGEMENTS

Key R & D Program for Social Development in Yunnan Provincial (in China) (202203AC100006).

REFERENCES

[1] Kalnay E. Atmospheric modeling, data assimilation and predictability[M]. Cambridge university press, 2003. [M].https://doi.org/10.1198/tech.2005.s326

[2] Gustafsson N, Janjić T, Schraff C, et al. Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres[J]. Quarterly Journal of the Royal Meteorological Society, 2018, 144(713): 1218-1256.https://doi.org/10.1002/qj.3179

[3] Sasaki Y. An ObJective Analysis Based on the Variational Method[J]. Journal of the Meteorological Society of Japan. Ser. II, 1958, 36(3): 77-88.https://doi.org/10.2151/jmsj1923.36.3_77

[4] Sasaki Y. Some basic formalisms in numerical variational analysis[J]. Monthly Weather Review, 1970, 98(12): 875-883.https://doi.org/10.1175/1520-0493(1970)098<0875:SBFINV>2.3.CO;2

[5] Dimet F X L, Talagrand O. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects[J]. Tellus A, 1986, 38 A(2): 97-110.https://doi.org/10.1111/j.1600-0870.1986.tb00459.x

[6] Lynch P. The origins of computer weather prediction and climate modeling[J]. Journal of Computational Physics, 2008, 227(7): 3431-3444.https://doi.org/10.1016/j.jcp.2007.02.034

[7] Di, R., Wang, X., Meng, X., et al. Research on visibility prediction method based on multisource feature fusion. China New Technologies and Products, 2022(11): 13-16.10.13612/j.cnki.cntp.2022.11.020.

[8] Yang, X., Xiao, Y., Chen, S. Study on Wind Speed and Power Generation Prediction in Wind Farms. Proceedings of the Chinese Society for Electrical Engineering, 2005, 25(11): 1-5..https://doi.org/10.3321/j.issn:0258-8013.2005.11.001

[9] Maddix D C, Wang Y, Smola A. Deep Factors with Gaussian Processes for Forecasting[M]. arXiv, 2018.http://arxiv.org/abs/1812.00098

[10] Shin H, Rüttgers M, Lee S. Effects of spatiotemporal correlations in wind data on neural network-based wind predictions[J]. Energy, 2023, 279: 128068.https://doi.org/10.1016/j.energy.2023.128068

[11] Lipton Z C, Kale D C, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks[J]. arXiv preprint arXiv:1511.03677, 2015. http://arxiv.org/abs/1511.03677

[12] Agrawal S, Barrington L, Bromberg C, et al. Machine Learning for Precipitation Nowcasting from Radar Images[M]. arXiv, 2019.http://arxiv.org/abs/1912.12132

[13] Li X, Gao H, Zhang M, et al. Prediction of Forest Fire Spread Rate Using UAV Images and an LSTM Model Considering the Interaction between Fire and Wind[J]. Remote Sensing, 2021, 13(21): 4325.https://doi.org/10.3390/rs13214325

[14] Chen G, Tang B, Zeng X, et al. Short-term wind speed forecasting based on long short-term memory and improved BP neural network[J]. International Journal of Electrical Power & Energy Systems, 2022, 134: 107365.https://doi.org/10.1016/j.ijepes.2021.107365

[15] Shen Z, Fan X, Zhang L, et al. Wind speed prediction of unmanned sailboat based on CNN and LSTM hybrid neural network[J]. Ocean Engineering, 2022, 254: 111352.https://doi.org/10.1016/j.oceaneng.2022.111352