Response to reviewers' comments:

# Response to Reviewer 1

Reviewer #1:Convective wind gusts (GCs) cause great structural damage and serious hazards. This paper designs a physics-constrained model (namely by an improved PhyDNet) for 0-2 h of quantitative CGs nowcasting with a spatial resolution of 0.01° ×0.01° and a 6-minute temporal resolution. The structure of the forecast neural network is designed interesting, which contains a temporal attention module. In addition, this model combines sufficient situ observations and radar data. I admire the author's dedication to such detailed work. I only have some minor points to make before being published.

We appreciated the constructive and detailed comments from the reviewer, which helped us greatly to improve the manuscript. Please find the detailed responses to each comment below.

[Comments]

I **DATA**

① The wind data the authors used is from automatic weather stations (AWSs), and they are few wind observation data on the seacoast (and even over the sea) from Figure 1b. I wonder how to interpolate the wind data over the sea, whether to set it to 0 directly. Please clarify the detailed information on the IDW interpolation method.

Response: Thanks for your suggestion. In this study, we employed the Inverse distance-weighted (IDW) algorithm for wind data interpolation. For each grid point, we used the nearest four stations within a 15 km radius to perform the interpolation. In the case of sea areas, due to the lack of weather stations, the condition of having four stations within a 15 km radius is not met, resulting in many 'nan' values over the sea.

Additionally, since accurate wind speed observations are lacking over the sea, the interpolation in coastal areas may also have significant errors. To address this issue, we used the Natural Earth land-sea dataset to create a mask for the sea areas. We set a special value (here we set to 0) for the sea areas and assigned a weight of 0 to these regions during the training process. This ensures that the sea areas do not contribute to the backpropagation optimization.

② If the wind data over the sea is 0, please give some analyses on whether this approach affects model performance, such as causing underestimations on PWGS to some extent.

Response: As mentioned above, we set the wind data over the sea as 0. During the training process, we mask the wind data on the sea areas by setting the weight of the loss function to 0. This is done to minimize the loss in forecasting gusts on land areas. Although we minimized the loss by masking the wind data during the calculation, errors still exist because the convolution calculation still involves the coastal areas. Setting the wind data on the seacoast and over the sea to 0 or a special value will not affect the model training, as the values on the seacoast and over the sea do not participate in the model's backpropagation optimization. However, during the forward propagation process, which is the forecasting process, the special values on the seacoast (and over the sea) still participate in the convolutional operations. Thus the ASWS or PWGS close to the seacoast may be slightly underestimated.

If the wind data over the sea is set to 0, but the seacoast areas are still included in the loss calculation, it could significantly impact the results and result in more pronounced underestimations.

We have added explanations for this in the revised manscript (Lines 80-81 and 86-91) : “Note that there are limited wind observation data available on the seacoast/over the sea, and as a result, this study focuses on gust forecasting in the land region of eastern China.” “The wind observation data on the seacoast and over the sea was set to 0. Subsequently, during the training process, we masked the wind data on the seacoast and over the sea by setting the weight of the loss function to 0. This ensures that the sea areas do not contribute to the backpropagation optimization. However, the values in the sea areas still participate in the convolutional operations during the forward propagation process, which could lead to a slight underestimation of the ASWS or PWGS values in areas close to the seacoast.”

③ The same interpolation problem also emerges in radar reflectivity. There are 10 weather radar stations in eastern China, and how to interpolate them into grided data. I suggest the authors exhibit some exemplificative image pairs for ASWS/radar data and processed gridded data.

Response: Thanks for your suggestion. The 3 km radar reflecticity/RMOS data was obtained from the operational Doppler weather radar 3-D digital mosaic system developed by the Chinese Academy of Meteorological Sciences (Wang et al., 2009). This system integrates data from multiple Doppler weather radar stations across eastern China, providing a comprehensive and high-resolution representation of the atmosphere at a 3 km altitude. The system can provide quality controlled base data, 3-D reflectivity grid data of single site, 3-D mosaic reflectivity and some derived products base on mosaic base data, which are useful not only for operational work, but also for scientific research.

For specific data processing, it converts the radar scan data from polar coordinates to Cartesian coordinates using a two-step interpolation process. First, it applies a nearest-neighbor interpolation method in the radial direction. Second, it uses vertical linear interpolation in the azimuthal direction.

To combine the gridded reflectivity fields from multiple radar stations, the systerm stitches them together, ensuring appropriate overlap in many regions, particularly in the middle and upper troposphere where data from multiple radars are available. After that, an exponential weighting interpolation method based on the distance between individual grid cells and radar locations is used to obtain interpolated results at an altitude of 3 km. More details of radar data interpolation can be found in (Wang et al., 2009).

We have added the mentioned radar data interpolation into the revised manuscript (Lines 105-107).

Reference:

Wang, H., Liu, L., Wang, G., Zhuang, W., Zhang, Z., and Chen, X.: Development and application of the Doppler weather radar 3-D digital mosaic system, Journal of Applied Meteorological Science, 20, 214–224, 2009 (in Chinese).

④ Is the data the authors used open-source? If so, please clarify the web links of them.

Response: Yes, the data we used is open-source. Please find the wind observation data and radar data files at this site: https://doi.org/10.7910/DVN/PIZU7V. We also have added this in the "Code availability" section of the revised manuscript (Page 24).

II **MODEL**

① Please clarify the detailed structures of the convolutional encoder and decoder in Appendix, and the feature map shapes of h^E, h^p, h^c, and h^m.

Response: Thanks for your detailed suggestion. The detailed structures of the convolutional encoder and decoder are shown in Table S1. The input image shape is batch_size×sequence_length×channels×width×height. Specifically, batch size is set to 2, input sequence length is 10. The feature map shapes showed in the convolutional and deconvolutional layers are just channels×width×height. After each convolutional layer in the encoder, there is a Group Normalization followed by a LeakyReLU activation function. Similarly, in the decoder, a Group Normalization and a LeakyReLU activation function follow each deconvolution layer, except for the fourth deconvolution layer.

The feature map shapes of h^E, h^p, h^c, and h^m are both 128×30×35. We also have added this in supplement of the revised manscript.

Table S1 Parameter settings in encoder and decoder. "Conv" denotes convolutional layer in encoder, "Deconv" denotes deconvolutional layer in decoder.

Encoder	Input size	Kernel size /stride	**Decoder**	Input size	Kernel size /stride
Conv1	2×480×560	3×3/(2,2)	Deconv1	128×30×35	3×3/(2,2)
Conv2	16×240×280	3×3/(2,2)	Deconv2	64×60×70	3×3/(2,2)
Conv3	32×120×140	3×3/(2,2)	Deconv3	32×120×140	3×3/(2,2)
Conv4	64×60×70	3×3/(2,2)	Deconv4	16×240×280	3×3/(2,2)
Encoder output size: 128×30×35			**Decoder output size:** 2×480×560		

② What is the difference between h^D and h^m? In Equation (4), h^D is the summation of h^p and h^c. However, in Figure 3, h^D is from some transformation of h^m, and h^m is the summation of h^p and h^c. Please unify the illustration.

Response: Thanks for pointing out this. After careful checking, h^D is actually h^m. Specifically, h^m is the summation of h^p and h^c, and it is fed into the deconvolutional units to calculate the predictions. We have corrected this mistake and modified Figure 1 (shown below) in the revised manuscript.

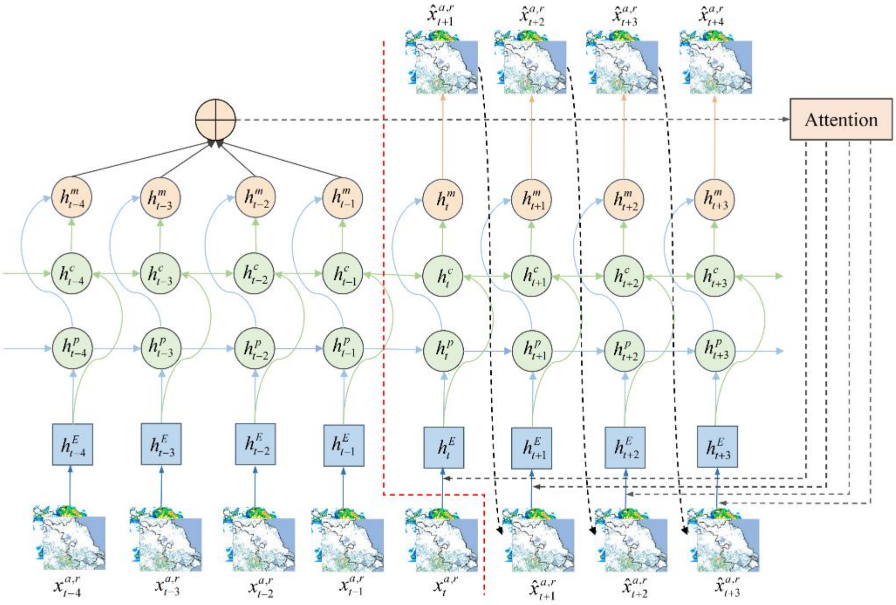


Figure 1. Illustration of CGsNet. The encoder is to the left of the dotted red line, and the decoder is to the right. $x_i^{a,r}$ and $\hat{x}_i^{a,r}$ are the observed and forecasted ASWS/RMOS fields, respectively. h_i^E indicate the input tensors calculated by the convolution units. h_i^c and h_i^p indicate the hidden states of ConvLSTM and PhyCell, respectively. h_i^m represents the hidden state that combines the values from h_i^c and h_i^p .

③ From the example of Figure 3, I wonder if there are 4 individual attention modules corresponding to the output length. And the effect of the attention is to find the most significant historical information from the input sequence. For example, when predicting t , the attention module can assign weights for $\{h^m_{t-4}$, h^m_{t-3}, h^m_{t-2}, $h^m_{t-1}\}$, and when prediction $t + 1$, the attention module can assign weights for $\{h^m_{t-4}$, h^m_{t-3}, h^m_{t-2}, $h^m_{t-1}\}$ as well, or for $\{h^m_{t-3}$, h^m_{t-2}, h^m_{t-1}, $h_t{}^m\}$.

Response: Yes, when predicting $\hat{x}_t^{a,r}$, the attention module can assign weights for $\{h^m_{t-4}$, h^m_{t-3}, h^m_{t-2}, $h^m_{t-1}\}$, and when predicting $\hat{x}^{a,r}_{t+1}$, the attention module can assign weights for $\{h^m_{t-3}$, h^m_{t-2}, h^m_{t-1} , $h_t{}^m\}$. We have added this explanation into the revised manscript (Lines 154-156): "The effect of the attention mechanism operation is to find the most significant historical information from the input sequences, e.g., when predicting $\hat{x}^{a,r}_{t+1}$, the attention module can assign weights for $\{h^m_{t-K}$,, …, $h_t{}^m\}$."

④ There is no k on the right side of Equation (1), how to obtain s_{tk} ($\forall k \in [1, K]$)?

Response:Thanks for pointing out this. After careful checking, we modified Equation (1) as: $s_{tk} = W * h^E_{t-k} + b$, $\forall k \in [1, K]$. Then α_{tk} is calculated by input s_{tk} and it can be interpreted as the relative importance of the k-th h^{m}. Please check this in the Equation (1) of the revised manuscript.

⑤ Please give the detailed calculation process of PhyDNet and ConvLSTM in Appendix.

Response: Thanks for your suggestion. We have added the deailed calculation process of PhyDNet (the main module: PhyCell) and ConvLSTM in supplement, as follows:

1) PhyCell

PhyCell is a physical cell of PhyDNet (Guen and Thome, 2020b), whose dynamics are governed by the PDE response function $\boldsymbol{\mathcal{M}}_p(\boldsymbol{h}^{\boldsymbol{p}},\boldsymbol{u})$:

$$\boldsymbol{\mathcal{M}}_p(\boldsymbol{h},\boldsymbol{u}) \coloneqq \Phi(\mathbf{h}) + \mathsf{C}(\mathbf{h},\mathbf{u}) \qquad (1)$$

where $\Phi(\mathbf{h})$ represents a physical predictor modeling only the latent dynamics, and the $\mathsf{C}(\mathbf{h},\mathbf{u})$ represents a correction term, modeling the interaction between input data and latent state. $\Phi(\mathbf{h})$ can be modeled as:

$$\Phi(\mathbf{h}(\mathrm{t},\mathbf{x})) = \sum_{i,j:i+j\leq q} c_{i,j}\frac{\partial^{i+j}\boldsymbol{h}}{\partial x^i \partial y^j}(t,\mathrm{x}) \qquad (2)$$

$\Phi(\mathbf{h}(\mathrm{t},\mathbf{x}))$ combines the spatial derivatives with coefficients $c_{i,j}$ up to a certain differential order q. A wide range of classical physical models, e.g., the wave equation and heat equation, can be subsumed in this generic class of linear PDEs.

The discrete time PhyCell with the standard forward Euler numerical scheme can be written as:

$$\boldsymbol{h}(t+1) = (1-\mathbf{K}_t)\odot\left(\boldsymbol{h}_t + \Phi(\boldsymbol{h}_t)\right) + \mathbf{K}_t \odot \mathbf{E}(\boldsymbol{u}_t) \qquad (3)$$

where $\odot$ denotes the Hadamard product; $\mathbf{K}_t$ is a gating factor; and $\mathbf{E}(\boldsymbol{u}_t)$ indicates the new observed input. Here we write the equivalent two-steps for Eq (3):

$$\widetilde{\boldsymbol{h}}_{t+1} = \boldsymbol{h}_t + \Phi(\boldsymbol{h}_t) \qquad \text{Prediction} \qquad (4)$$

$$\boldsymbol{h}_{t+1} = \widetilde{\boldsymbol{h}}_{t+1} + \mathbf{K}_t \odot (\mathbf{E}(\boldsymbol{u}_t) - \widetilde{\boldsymbol{h}}_{t+1}) \text{ Correction} \qquad (5)$$

The Eq (4) represents the prediction step, which is a physically-constrained motion in latent space, and it computes the intermediate representation: $\widetilde{\boldsymbol{h}}_{t+1}$. The correction step in Eq (5) incorporates the input data. The decoupling between prediction and correction can be used to robustly train the model in missing data contests and long-term forecasting. Besides, the trade-off between both steps is controlled by $\mathbf{K}_t$, which can be interpreted as the Kalman gain.

The physical predictor Φ in Eq (4) is implemented by using a convolutional neural network, based on the connection between convolutions and differentiations. The 1×1 convolutions are used to linearly combine the derivatives with $c_{i,j}$ coefficients in Eq (2). Moreover, the Kalman gain $\mathbf{K}_t$ is approximated by a gate with learned convolution kernels $\mathbf{W}_h$, $\mathbf{W}_u$ and bias $\mathbf{b}_k$:

$$\mathbf{K}_t = \tanh\left(\mathbf{W}_h * \widetilde{\boldsymbol{h}}_{t+1} + \mathbf{W}_u * \mathbf{E}(\boldsymbol{u}_t) + \mathbf{b}_k\right) \qquad (6)$$

where * respent the convolutional operator. If $\mathbf{K}_t = 0$, the dynamic follows the physical predictor; if $\mathbf{K}_t = 1$, the latent will be reset and only driven by the input.

PhyCell is an atomic recurrent cell used for building physically-constrained RNNs. The PhyDNet here uses one layer of PhyCell, which can also be easily stacked to build more complex models.

2) ConvLSTM

ConvLSTM is a variant of the long short-term memory network (Shi et al., 2015), which is a fundamental and effective spatiotemporal recurrent structure for spatiotemporal modeling. The ConvLSTM uses the forget gate, input gate, and output gate to update its cell and hidden states. The input gate controls how much of the new information will be added to the memory cell. The forget gate is used to control how much of the previous information will be forgotten from the memory cell, while the cell information which will be propagated to the new gate is controlled by the output gate. The calculation processes of the ConvLSTM are as follows:

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \odot c_{t-1} + b_i) \quad (7)$$

$$f_t = \sigma\left(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \odot c_{t-1} + b_f\right) \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (9)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \odot c_t + b_o) \quad (10)$$

$$h_t = o_t \odot \tanh\,(c_t) \quad (11)$$

where σ is the sigmoid activation function. Besides, x_t, c_t, h_t, i_t, f_t, and o_t represent input, memory cell, hidden state, input gate, forget gate, and output gate, respectively. Bias b and weight W both represent learning parameters.

References:

Guen, V. L. and Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 474–11 484, https://doi.org/10.1109/CVPR42600.2020.01149, 2020b.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems, 28, https://doi.org/10.1007/978-3-319-21233-3_6, 2015.

⑥ About the physics-constraint concept mentioned in Introduction and Model architecture (namely PhyDNet), I think the authors should cite the “inductive biases” from doi.org/10.1038/s42254-021-00314-5, which can be interpreted by designing specialized network architectures that implicitly embed prior knowledge and satisfy a set of given physical laws.

Response: Thanks for your suggestion. We have cited the “inductive biases” in the Model architecture section: “PhyDNet steers the learning process toward identifying physically consistent solutions by introducing an appropriate inductive bias (Karniadakis et al., 2021), that is, implicitly embedding prior knowledge in the network architecture and satisfying a given set of physical laws.” (see Lines 126-128 in revised manuscript).

Reference:

Karniadakis, G.E., Kevrekidis, I.G., Lu, L. et al.: Physics-informed machine learning, Nature Reviews Physics, 3, 422–440, https://doi.org/10.1038/s42254-021-00314-5, 2021.

III **EXPERIMENT**

① Please append an ablation study on the proposed attention module.

Response: Thanks for the suggestion. As suggested, we have added an ablation study on the attention module. Specifically, we compared the performance of PhyDNet (without attention) with CGsNet. Table 2, Figure 4-6 show the results of the ablation study, which indicate that the forecasting performance of CGsNet for ASWS is superior to that of PhyDNet. These results suggest that the attention mechanism proposed in CGsNet is effective and can significantly improve the accuracy of ASWS forecasting. Overall, the results confirm that CGsNet is reliable and accurate for ASWS nowcasting. Detailed comparative descriptions can be found in Sections 4.1 and 4.2 of the revised manuscript.

Table 2. Quantitative results of CGsNet and PhyDNet on ASWS nowcasting. 95% CI represent the 95% confidence interval of the indices.

Model	ASWS (m/s)	CSI 95% CI	HSS 95% CI	POD 95% CI	MAE 95% CI	RMSE 95% CI
CGsNet	**8.0**	0.41 (0.33; 0.49)	0.54 (0.47; 0.61)	0.59 (0.51; 0.66)	1.60 (1.46; 1.80)	2.26 (2.00; 2.54)
	10.8	0.31 (0.22; 0.40)	0.42 (0.32; 0.50)	0.47 (0.34; 0.60)	2.19 (1.93; 2.51)	3.07 (2.62; 3.56)
	13.9	0.15 (0.04; 0.21)	0.20 (0.07; 0.24)	0.22 (0.10; 0.31)	2.90 (2.26; 2.66)	4.22 (3.25; 5.10)
PhyDNet	**8.0**	0.39 (0.31; 0.47)	0.52 (0.44; 0.59)	0.55 (0.46; 0.62)	1.71 (1.55; 1.91)	2.40 (2.14; 2.68)
	10.8	0.28 (0.20; 0.38)	0.38 (0.28; 0.47)	0.41 (0.29; 0.53)	2.40 (2.12; 2.76)	3.33 (2.90; 3.83)
	13.9	0.12 (0.03; 0.19)	0.16 (0.05; 0.21)	0.19 (0.09; 0.28)	3.10 (2.39; 3.97)	4.54 (3.59; 5.47)

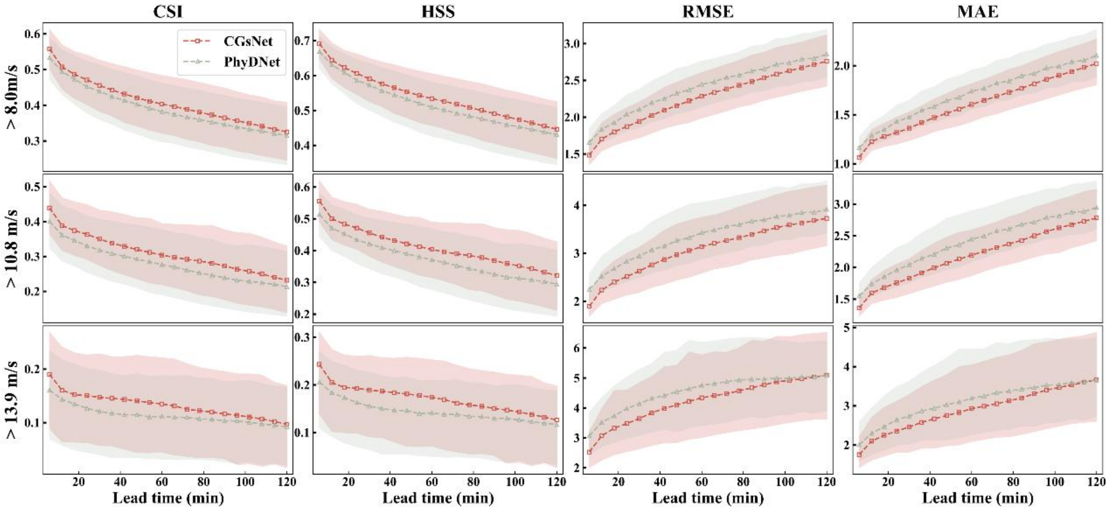


Figure 4. The CGsNet and PhyDNet results for different nowcasting lead times of ASWS at thresholds of 8.0 m/s, 10.8 m/s, and 13.9 m/s.The shaded red and green areas represent the 95% confidence intervals of the CGsNet and PhyDNet indices, respectively.

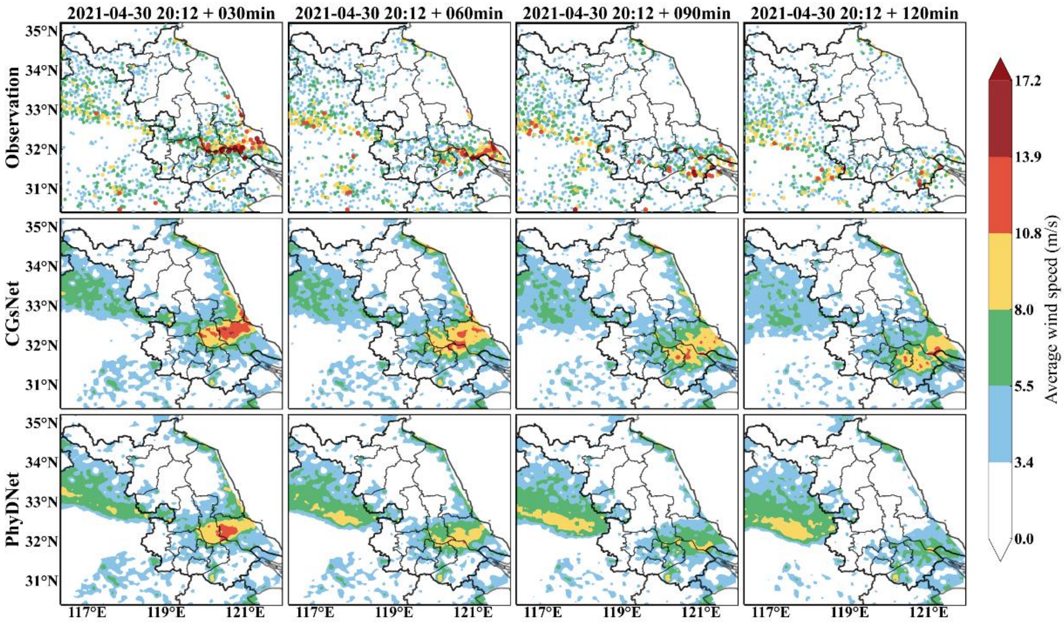


Figure 5. Observations (first row), CGsNet forecasts (second row), and PhyDNet forecasts (third row) of ASWS in eastern China, 30 April 2021, 20:12-22:12 BJT. Note that forecasting started at 20:12, and the observations and forecasts are shown at intervals of 30 min.

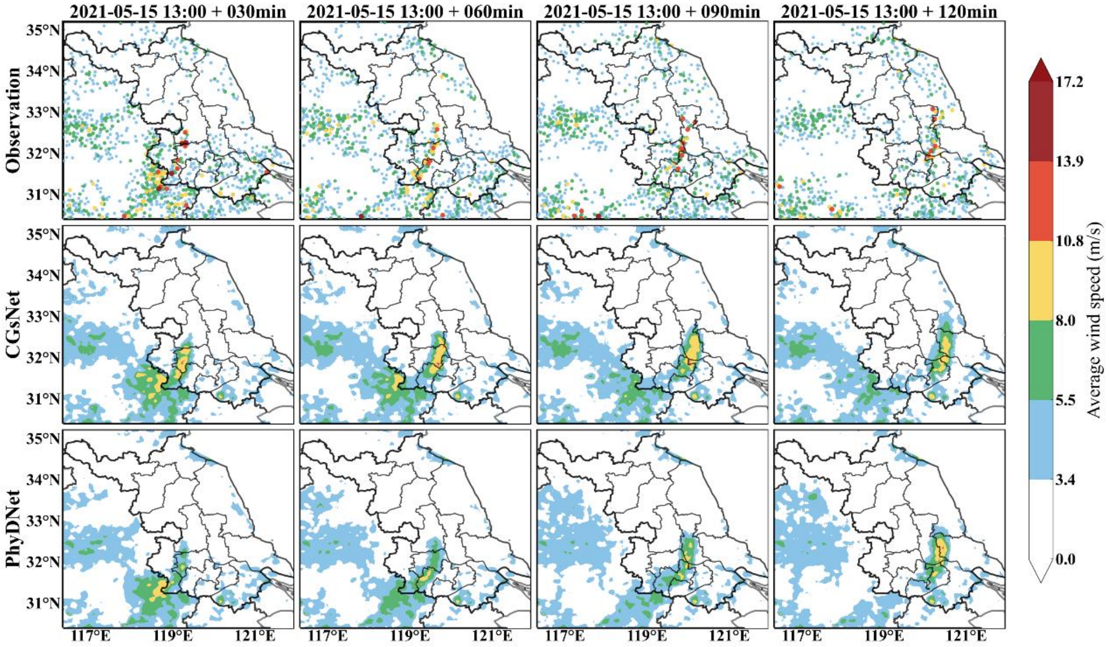


Figure 6. Same as Figure 5, but for 15 May 2021, 13:00-15:00 BJT.

Response to reviewers' comments:

# Response to Reviewer 2

Reviewer #2:This paper describes the use of a neural network for predicting convective wind gusts at lead times of 6-120 minutes. The neural network includes several architectural components that have recently been developed in the field of deep learning. These include pieces of the recurrent neural net (RNN) architecture, where the prediction at the k^{th} lead time becomes an input (predictor) to the neural network to predict at the $(k + 1)^{th}$ lead time; the attention mechanism, where the neural network can automatically determine the most important steps in a time series of predictor values; and PhyCell, which incorporates partial differential equations (PDE). The neural network predicts convective wind gusts on a 1-km grid covering eastern China. The authors compare their results to a baseline model called INCA and demonstrate (albeit without significance-testing) that their neural network, which they call CGsNet, outperforms INCA. Overall, the quality of both the writing and science are very good. I have a handful of major comments, which are made as inline comments in the attached PDF but also summarized below. Minor comments are not summarized below and can be seen only in the attached PDF.

Thanks for the constructive and detailed comments, which help us improve the manuscript significantly. We considered all comments and carefully revised all comments in the revised manuscript. Below is the point-to-point response to major and minor comments.

[Major comments]

1. The abstract contains several unclear/unjustified statements. See inline comments.

Response:Thanks for your detailed suggestion. We have revised the abstract section, addressing previously unclear or unjustified statements. Please check abstract section in the revised manuscript. About several comments, the responses are as follow:

1) Line 5: Do you mean "at 0--2-hour lead time"?

Yes. We have modified this and made it clear (see Line 5).

2) Line 5: What do you mean by "first"? That your approach is the first ever to achieve minute-/km-level forecasts?

In fact, accurately forecasting CGs at the minute-kilometer-level has been a challenge, with few past studies achieving this goal. While fine-scale numerical experiments can predict CGs at this level, they often require long computation times on the order of several hours, making them unsuitable for nowcasting needs. Therefore, we have removed the word "first" in the revised manuscript for the sake of accuracy.

3) Line 7: What do you mean by "spatiotemporally consistent"? This was never discussed in the paper.

The “spatiotemporally consistent” means that the CGs forecasts are continuous in temporal and spatial within the 0-2 hour forecast window. Unlike traditional extrapolation-based forecasting algorithms such as SCIT (Storm Cell Identification and Tracking) (Johnson et al., 1998), which focus on convective cells and require identification of these cells before issuing extrapolation forecasts and rough early warnings for affected areas, our algorithm can perform 2D grid-based forecasts with 6-minute updates in all weather conditions. Therefore, we describe our algorithm as spatiotemporally consistent.

Reference:

Johnson, J. T., MacKeen, P. L., Witt, A., Mitchell, E. D. W., Stumpf, G. J., Eilts, M. D., & Thomas, K. W.: The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm, Weather and forecasting, 13(2), 263-276, https://doi.org/10.1175/1520-0434(1998)013<0263:TSCIAT>2.0.CO;2, 1998.

4) Line 14: "alerts the damaging wind events in meteorological services." --This suggests that your model is used in operations already -- based on the rest of the paper, it sounds like your model is not yet operational.

We have removed this description.

2. Most hyperparameter choices are not justified at all. Whether hyperparameter choices were made via experimentation or *a priori* reasoning, the justification should be included in the paper. If hyperparameters were chosen by experimentation, the experiments should be documented (although I'm okay with putting most of the details in the Appendix or Supplemental Material); if chosen by *a priori* reasoning, the *a priori* reasoning should be explained. For examples of unjustified hyperparameter choices, see lines 84-85, lines 150-152, and Table 1 (and inline comments at these places).

Response:Thanks for the detailed suggestion.

(1) For the Lines 84-85: we used Inverse distance-weighted (IDW) interpolation (radius of influence = 15 km, 4 stations, and power of 2) for interpolation. Here we first briefly explain the IWD calculation steps, as follows:

(a) Calculating the distance d from unknown points to all points;

(b) Calculating the weight λ for each point, the weight is a function of the reciprocal of the distance: $\lambda_i = \frac{d_i^{-p}}{\sum_{i=1}^{K} d_i^{-p}}$, where K is the total number of discrete points. Power (p) is a parameter used to calculate the influence weight of the nearest K discrete points on the interpolation site, which controls the effect of known points on the interpolation value based on their distance from the interpolated point.

(c) Calculating interpolation results: $\hat{z}(x_0, y_0) = \sum_{i=1}^{K} \lambda_i z(x_i, y_i)$, where (x_0, y_0) is the interpolation point coordinate and (x_i, y_i) represents the coordinates of discrete points.

To determine the optimal parameters for the IDW interpolation, we experimented with various combinations of the number of stations (K) and radius of influence (R) (Figure S13). In IDW interpolation, R refers to finding the K nearest discrete points from the interpolation site within a radius range of R. As shown in the Figure S13, it is evident that when R is set to 5 km, the condition of having nearest discrete points/stations within a 5 km radius is not met, resulting in many 'nan' values. This is because the distance between meteorological observation stations is 10 km. Therefore, when R is less than 10 km, the observed field cannot be adequately interpolated.

However, when R is set to 10 km or lager, the interpolation results do not significantly change. For K, when taking different values ranging from 2 to 16, the interpolation results are comparable. As K increases, the value of the interpolated point is progressively smoothed by the surrounding discrete points. Generally, when the value of R is not less than 10 km and the value of K is not too large, the interpolation results of the observed wind field are similar, and the results are effective. Thus using a R of 15 km and K= 4 staions are feasible. The choice of 15 km radius of influence allowed us to capture local variations in the data without being overly influenced by distant stations, which might not have the same meteorological conditions. Using 4 stations ensured that the interpolation incorporated sufficient data points.

For power, as the power value increases, the interpolated value will gradually approach the value of the nearest point. By specifying a small power value, the influence of points farther away will be great, resulting in a smooth plane. Based on a comprehensive analysis, we selected a common value of power=2. The power of 2 provided a balance between the weighting of nearby and distant stations, which has been widely used in many studies.

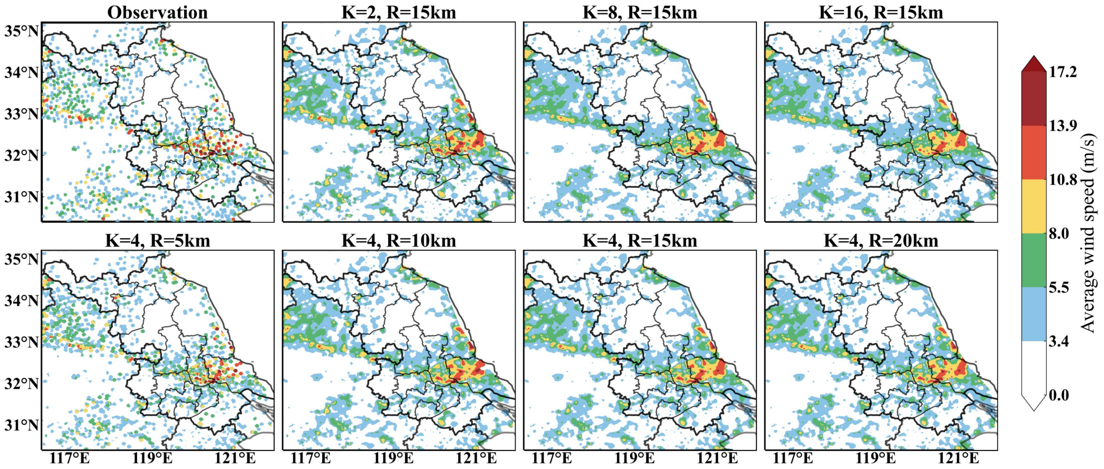


Figure S13. The various combinations of the number of stations (K) and radius of influence (R) for IDW interpolation.

（2）For the lines 150-152：

To ensure the robustness and generalization ability of our model, we carefully selected the hyperparameters through a priori reasoning. The reasons for these hyperparameter choices are as follows:

①Batch size: the batch size of 2 was chosen based on two factors. First, a small batch size also has good generalization ability since each batch is randomly selected, allowing the model to adapt better to new data. Second, a small batch size helps to reduce memory usage and accelerate training, as the samples occupy a significant amount of memory.

②Learning rate: the learning rate was set based on the settings of PhyDNet (Guen and Thome, 2020b), which involved selecting an initial learning rate of 0.001 and decreasing it if the loss function did not decrease after several epochs of training.

③Epoch: setting the number of epochs to 50 is to ensure that the model is fully trained. When saving the model, the one with the lowest validation loss was selected and saved.

（3）Regarding Table 1:

The thresholds parameters of loss function was mainly based on the wind speed classification determined by the China Meteorological Administration (https://www.cma.gov.cn/2011xzt/20120816/2012081601/201208160101/201407/t20140717_252607.html), as shown in the following Table S1. High weights were assigned to high wind speeds and RMOS values since accurate predictions in these ranges are crucial for ensuring life safety and minimizing potential damage caused by extreme weather events. For example, wind speed greater than 20.8 m/s are considered to be particularly hazardous, and thus, predictions in this range were given a high weight. Besides, the issue of imbalance between high wind speed data and low wind speed data can be partially addressed by assigning different weights, as there is a scarcity of high wind speed data samples and an abundance of low wind speed data samples. This approach can help alleviate the problem of the insufficient number of high wind speed data samples and enable the model to forecast areas where strong gusts occur, which is a significant concern. The specific weights were set with reference to the settings of Shi et al (2015) and expert advice.

Table S1. The wind speed levels defined in China

Wind level	Wind speed (m/s)
0	[0.0-0.3）
1	[0.3-1.6)
2	[1.6-3.4)
3	[3.4-5.5)
4	[5.5-8.0)
5	[8.0-10.8)
6	[10.8-13.9)
7	[13.9-17.2)
8	[17.2-20.8)
9	[20.8-24.5)
……	

Reference:

Guen, V. L. and Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 474–11 484, https://doi.org/10.1109/CVPR42600.2020.01149, 2020b.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems, 28, https://doi.org/10.1007/978-3-319-21233-3_6, 2015.

We have also added the parameter settings mentioned above in the supplement of the revised manuscript.

3. It seems like the authors used some kind of subsetting to create their dataset, but the subsetting methodology is unclear. See inline comment on lines 86-88.

Response: Sorry for not clear. Given the extreme nature of CGs, not all observation periods may contain CGs. Therefore, a preliminary data selection is necessary. CGs are identified by the China Meteorological Administration (CMA) as having a Peak Wind Gust Speed (PWGS) of at least 17.2 m/s caused by severe atmospheric convection. Based on the relationship between ASWS and PWGS, we selected wind speed observations with ASWS greater than 10.8 m/s, which satisfies the condition of ASWS > 17.2/1.77 m/s (gust factor = 1.77). Additionally, precipitation often occurs during CG events, thus wind observations with precipitation events were further selected from the dataset with ASWS greater than 10.8 m/s. After selecting the wind observation data, radar data corresponding spatiotemporally to the wind gust data were selected to form a dataset. This dataset was further divided into training, validation, and testing sets.

We have clarified this in the revised manuscript (lines 92-94) as: "To further select ASWS data associated with CGs, the samples selection need meet two concurrent principles: 1) more than 2% of stations recorded ASWS > 10.8 m/s; and 2) the precipitation at more than 5% of the stations within an hour was greater than 0.1 mm."

4.It is unclear how the authors handle radar data at different heights. See inline comment on lines 96-97.

Response: Thanks for the suggestion. We only used radar data (base reflectivity factor) at 3 km, and we didn't create a grid of composite (column-maximum) reflectivity. The use of 3 km altitude radar mosaic data offers several advantages for identifying and forecasting convective gusts:

(1) Radar data at a 3 km altitude is less affected by ground clutter, which can obscure low-level features and lead to false detections.

(2) It provides comprehensive and continuous coverage of the study area, which is particularly beneficial in regions with complex terrain or large spatial gaps between radar stations.

(3) It offers a more accurate representation of storm structure compared to radar data at other altitudes, which is essential for detecting severe weather features such as bow echo and hook echo. These features can be crucial for improving forecast accuracy.

We have added this clarification in lines 101-104 of the revised manuscript: "The constant-altitude radar RMOSs at 3 km are less susceptible to ground clutter compared to radar data at other altitudes. This allows them to provide continuous and comprehensive coverage of the study area, as well as a more accurate representation of storm structure. Thus, these constant-altitude radar RMOSs at 3 km were used as auxiliary data along with observed wind data to forecast the speed of CGs."

5. The target variables (what the CGsNet is predicting) eventually become clear, but quite late in the paper. The target variables should be clear from the beginning. See inline comment on line 107.

Response:Thanks for your suggestion. The CGsNet model can generate predictions for both ASWS and RMOS. We primarily focuses on the forecasting results for ASWS, with the RMOS predictions serving as auxiliary information. Therefore, the analysis is limited to ASWS only. We have modified the description on line 107 as follows: "the outputs were the forecasted ASWS with lead times of 6-120 minutes ahead"(see line 118); "Note that we just focus on analyzing the forecasting results of ASWS, and the forecated RMOS is a secondary result." (see lines 158-159).

6. Section 3.1 needs to be written more clearly. See inline comments for specific points that are unclear.

Response: Thanks for the suggestion. About several comments, the responses are as follow:

1) What does K refer to here? The number of hidden states, or the number of time steps in the input sequence? Or does num_hidden_states =num_time_steps_in_sequence somehow?

Yes, K is the number of time steps in the input sequence, and we have added this into the revised manuscript (line 139). The num_hidden_states in CGsNet is not equal to num_time_steps_in_sequence.

2) Why do the predictions have both the superscripts a (for ASWS) and r (for RMOS)? I'm confused about what you're predicting -- is it just ASWS, or RMOS? This should be clear.

Sorry for not clear. The CGsNet can predict both ASWS and RMOS, forming a sequence-to-sequence architecture that is well-suited for multistep prediction. It can output T future ASWS (a) and RMOS (r) predictions, as clarified by the addition of "This forms a sequence-to-sequence architecture suited for multistep prediction, outputting T future ASWS (a) and RMOS (r) predictions ..." in the revised manuscript (lines 140-141).

3)What is m? Why is this superscript needed?

Here we use "m" to represent the hidden state h_i^m that combines the values from h_i^c and h_i^p . h_i^c and h_i^p indicate the hidden states of ConvLSTM and PhyCell, respectively.

4) Where do the weights come from?

This weight is refer to the α_{tk}, each weight α_{tk} is computed by taking h_{t-k}^E as input, followed by a softmax operation.

5) The caption of Figure 3: This part does not make sense. "The input and output tensors calculated by the conv and deconv units" = 4 things: input tensor for conv unit, output for conv unit, input for deconv unit, output for deconv. But you match this description with 2 variables (h_i^E and h_i^D), not 4 variables.

After careful checking, h^D is actually h^m. Thus, we have modified this as: "h_i^E indicate the input tensors calculated by the convolution units, respectively".

We have revised Section 3.1 as mentioned. Please check the Section 3.1 in the revised manuscript.

7. I don't understand how the authors split their data. Specifically, (a) I don't understand how the training and validation data are split; (b) I don't understand why they have such an incredibly small testing dataset for wind gusts. See inline comments on lines 146-149.

Response:Thanks for pointing out this. After the carefully checking, (a) the ASWS and RMOS grid fields from 2016 to 2020 were used for model's training, and the fields from June to September 2021 and April to May 2021 were employed for ASWS validation and testing, respectively. The fileds from May to July 2022 were used for PWGS testing. We have clarified this in the revised manuscript in lines 163-166.

(b) We understand your concern regarding the small testing dataset for wind gusts and would like to address them as follows:

1) Observation resolution：the PWGS observations are collected at an hourly resolution, while the ASWS are sampled at a minute level. This difference in resolution results in a lower observation frequency for PWGS. Therefore, within the same time frame, there are fewer PWGS data available for testing.

2) Obtaining high-quality wind gust data is often challenging due to difficulties in accurately measuring wind gusts and maintaining reliable measurement equipment. As a result, there is limited availability of such data, which can lead to small testing datasets.

3) Meteorological and hydrological drought: during the year 2022, the Yangtze River Basin experienced an extreme drought, which had a significant impact on the wind gust data. The wind observations during this period may have significantly differed from the data distribution under normal conditions. Therefore, we took care to select the PWGS samples, and samples from PWGS were taken from typical CGs events recorded by the meteorological bureau (we have added this in the revised manuscript, see lines 96-97).

To enhance the model's generalization capabilities and enable good performance during convective gusts nowcasting, we set a low threshold for selecting ASWS samples before training the model. This approach ensures that the model covers a broader range of wind speed data during the learning process, improving its generalization ability under different wind speed conditions. Meanwhile, this strategy is particularly useful when facing challenging meteorological phenomena such as convective gusts.

In conclusion, although the testing dataset for PWGS was small, the available PWGS samples were typical CG events and therefore highly representative. Moreover, we employed rigorous techniques to ensure model robustness, and the use of various indices allowed us to effectively evaluate our model's performance.

8. It seems like the authors inappropriately used their testing data for model selection; only the validation data should be used for model selection. See inline comment on line 153.

Response: As you said, the validation data (not testing data) should be used for model selection. Actually, we did. We described this in lines 152-153:“the model weight with the minimum loss on the validation set was saved. The comparison and evaluation experiments were implemented on the testing set.” The saved model means the trained model, which is the CGsNet. “The comparison and evaluation experiments were implemented on the testing set” means that we evaluated the trained model using the testing set and the comparison experiments with INCA also used the testing set. The dataset was divided into three sets - training, validation, and testing - in chronological order, with no overlap between them, ensuring their independence. The testing set served as one final, independent assessment of the models' performance. For clarity, we have modified the above sentence as: “the model weight with the minimum loss on the validation set was selected and saved. Then the evaluation experiments were implemented on the testing set.” (lines 169-170)

9. The comparison of a custom loss function to data augmentation is highly erroneous. See inline comment on line 164.

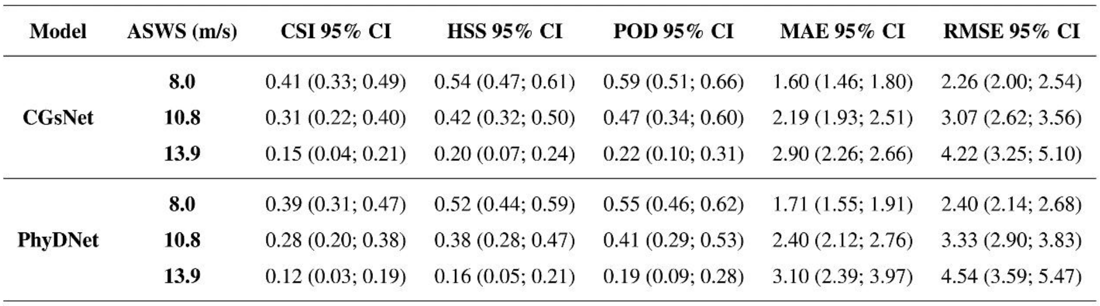
Response: You’re right. It’s a mistake. In fact, the weighted loss function is designed to ameliorate the issue of data imbalance caused by the scarcity of strong gust data. The custom loss function can give a large weight to strong gusts and a small weight to weak gusts. Thus the model can effectively learn the strong gust area during the training process and improve CG forecasting accuracy. We have modified this in the revised manuscript (lines 182-184), as also shown in follows.

“The weighted MAE loss is designed to ameliorate the problem of data imbalance caused by the scarcity of strong gust data, as it gives a large weight to strong gusts and a small weight to weak gusts.”

10. The results contain no confidence intervals or significance-testing -- in other words, no measure of uncertainty. I emphasize this comment the most, since results without uncertainty are nearly meaningless. All results in Figure 4 and Tables 2-4 should be accompanied by at least confidence intervals, if not also significance tests. See inline comments for details.

Response: Thanks for the constructive suggestion. Following the suggestion, we employed the bootstrapping method to calculate the confidence intervals for the indices in tables and figures mentioned in our manuscript, as shown below. The results demonstrate that the evaluation indices we computed all fall within the 95% confidence intervals, which indicates that our findings are reliable. Additionally, we have updated the description of these tables and figures in the revised manuscript to include information about the confidence intervals. Please check it in the revised manuscript.

Table 2. Quantitative results of CGsNet and PhyDNet on ASWS nowcasting. 95% CI represent the 95% confidence intervals of the indices.



Model	ASWS (m/s)	CSI 95% CI	HSS 95% CI	POD 95% CI	MAE 95% CI	RMSE 95% CI
CGsNet	**8.0**	0.41 (0.33; 0.49)	0.54 (0.47; 0.61)	0.59 (0.51; 0.66)	1.60 (1.46; 1.80)	2.26 (2.00; 2.54)
	10.8	0.31 (0.22; 0.40)	0.42 (0.32; 0.50)	0.47 (0.34; 0.60)	2.19 (1.93; 2.51)	3.07 (2.62; 3.56)
	13.9	0.15 (0.04; 0.21)	0.20 (0.07; 0.24)	0.22 (0.10; 0.31)	2.90 (2.26; 2.66)	4.22 (3.25; 5.10)
PhyDNet	**8.0**	0.39 (0.31; 0.47)	0.52 (0.44; 0.59)	0.55 (0.46; 0.62)	1.71 (1.55; 1.91)	2.40 (2.14; 2.68)
	10.8	0.28 (0.20; 0.38)	0.38 (0.28; 0.47)	0.41 (0.29; 0.53)	2.40 (2.12; 2.76)	3.33 (2.90; 3.83)
	13.9	0.12 (0.03; 0.19)	0.16 (0.05; 0.21)	0.19 (0.09; 0.28)	3.10 (2.39; 3.97)	4.54 (3.59; 5.47)

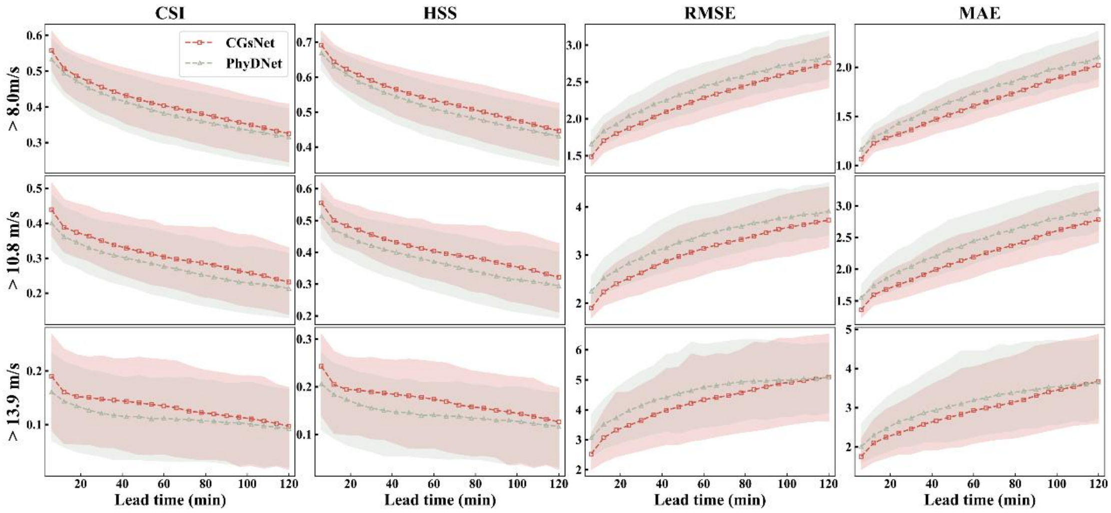


Figure 4. The CGsNet and PhyDNet results for different nowcasting lead times of ASWS at thresholds of 8.0 m/s, 10.8 m/s, and 13.9 m/s.The shaded pink and green areas represent the 95% confidence intervals of the CGsNet and PhyDNet indices, respectively.

Table 3. Quantitative results of CGsNet and INCA on PWGS nowcasting with 95% confidence intervals (in brackets). Note that the values in the first row of each metric represent CGsNet results and the second row is INCA results.

Evaluation Metrics	10.8 m/s	13.9 m/s	17.2 m/s	20.8 m/s
CSI	0.27 (0.25; 0.28)	0.22 (0.18; 0.26)	0.15 (0.09; 0.19)	0.06 (0.02; 0.08)
	0.11 (0.09; 0.12)	0.06 (0.03; 0.08)	0.03 (0.01; 0.05)	0.02 (0.00; 0.03)
POD	0.35 (0.32; 0.36)	0.31 (0.26; 0.37)	0.25 (0.15; 0.34)	0.12 (0.04; 0.18)
	0.30 (0.28; 0.32)	0.13 (0.09; 0.18)	0.07 (0.03; 0.12)	0.04 (0.01; 0.07)
BIAS	0.63 (0.61; 0.65)	0.73 (0.68; 0.78)	0.94 (0.76; 1.13)	1.21 (0.96; 1.54)
	2.07 (1.88; 2.42)	1.51 (1.44; 1.68)	1.27 (1.21; 1.37)	1.14 (1.07; 1.30)
FAR	0.45 (0.44; 0.46)	0.57 (0.52; 0.62)	0.73 (0.69; 0.80)	0.90 (0.87; 0.96)
	0.85 (0.83; 0.88)	0.91 (0.87; 0.95)	0.94 (0.91; 0.99)	0.96 (0.94; 1.00)

Table 4. The PWGS evaluation results from CGsNet and INCA for different nowcasting lead times at thresholds of 10.8 m/s, 13.9 m/s, 17.2 m/s and 20.8 m/s. Note that the values in the brackets represent 95% confidence intervals, and values in the first row of each metric represent CGsNet results and the second row is INCA results.

Lead time of 0-1 h				
Evaluation Metrics	**10.8 m/s**	**13.9 m/s**	**17.2 m/s**	**20.8 m/s**
CSI	0.31 (0.23; 0.40)	0.26 (0.16; 0.34)	0.17 (0.10; 0.24)	0.06 (0.02; 0.09)
	0.13 (0.10; 0.16)	0.06 (0.03; 0.09)	0.04 (0.01; 0.06)	0.02 (0.01; 0.04)
POD	0.40 (0.28; 0.52)	0.37 (0.24; 0.49)	0.28 (0.15; 0.41)	0.14 (0.05; 0.26)
	0.31 (0.23; 0.37)	0.13 (0.06; 0.19)	0.07 (0.02; 0.13)	0.03 (0.01; 0.07)
BIAS	0.69 (0.50; 0.89)	0.79 (0.56; 0.99)	0.99 (0.59; 1.36)	1.43 (0.69; 2.36)
	1.74 (1.37; 2.25)	1.30 (0.89; 1.78)	1.10 (0.56; 1.83)	1.00 (0.33; 1.93)
FAR	0.42 (0.34; 0.52)	0.53 (0.45; 0.64)	0.71 (0.63; 0.83)	0.90 (0.86; 0.96)
	0.82 (0.78; 0.87)	0.90 (0.83; 0.96)	0.93 (0.86; 0.98)	0.97 (0.94; 1.00)
Lead time of 1-2 h				
Evaluation Metrics	**10.8 m/s**	**13.9 m/s**	**17.2 m/s**	**20.8 m/s**
CSI	0.22 (0.13; 0.31)	0.18 (0.09; 0.26)	0.13 (0.04; 0.20)	0.05 (0.01; 0.09)
	0.09 (0.06; 0.12)	0.05 (0.02; 0.08)	0.03 (0.00; 0.06)	0.02 (0.00; 0.04)
POD	0.28 (0.16; 0.41)	0.26 (0.12; 0.39)	0.21 (0.06; 0.36)	0.09 (0.01; 0.19)
	0.29 (0.21; 0.37)	0.14 (0.06; 0.22)	0.07 (0.01; 0.12)	0.05 (0.00; 0.10)
BIAS	0.56 (0.35; 0.77)	0.66 (0.41; 0.89)	0.88 (0.47; 1.26)	0.98 (0.36; 1.72)
	2.43 (1.82; 3.36)	1.73 (1.18; 2.46)	1.45 (0.81; 2.23)	1.29 (0.60; 2.20)
FAR	0.49 (0.39; 0.61)	0.61 (0.52; 0.75)	0.76 (0.66; 0.90)	0.91 (0.83; 1.00)
	0.88 (0.84; 0.92)	0.92 (0.86; 0.97)	0.95 (0.87; 1.00)	0.96 (0.87; 1.00)

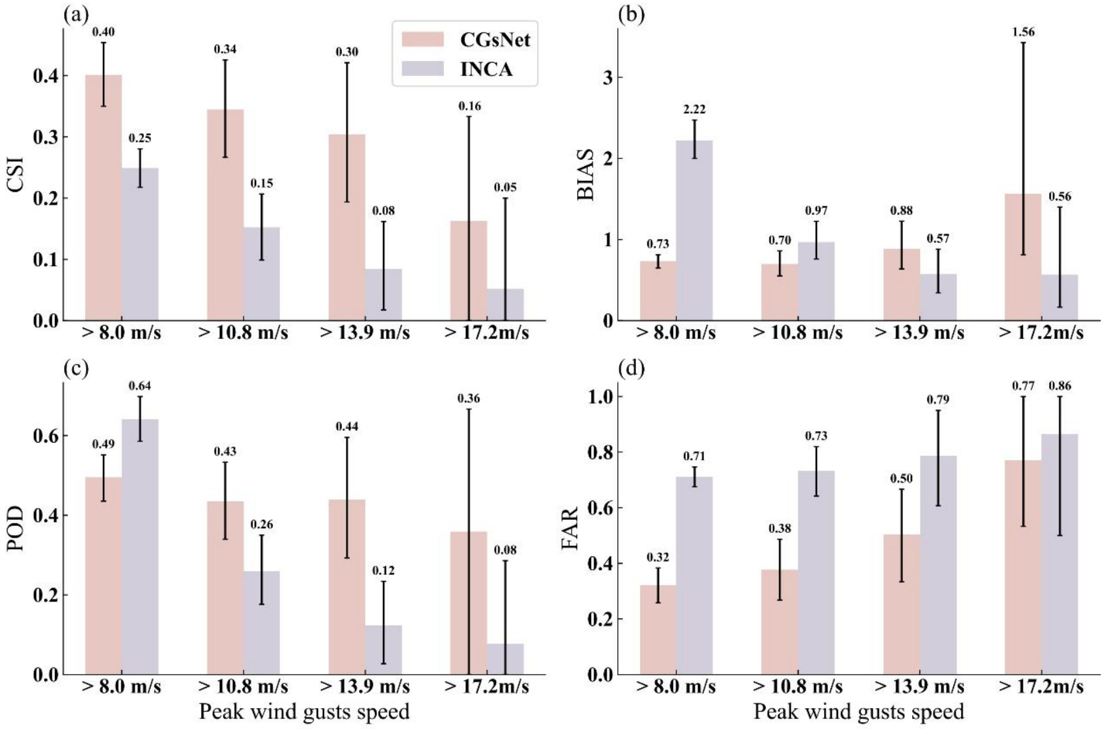


Figure 12. Comparison results of CGsNet and INCA on PWGS at thresholds of 10.8 m/s, 13.9 m/s, and 17.2 m/s on June 23, 2022, 18:00-20:00 BJT (Black error bars represent 95% confidence intervals).

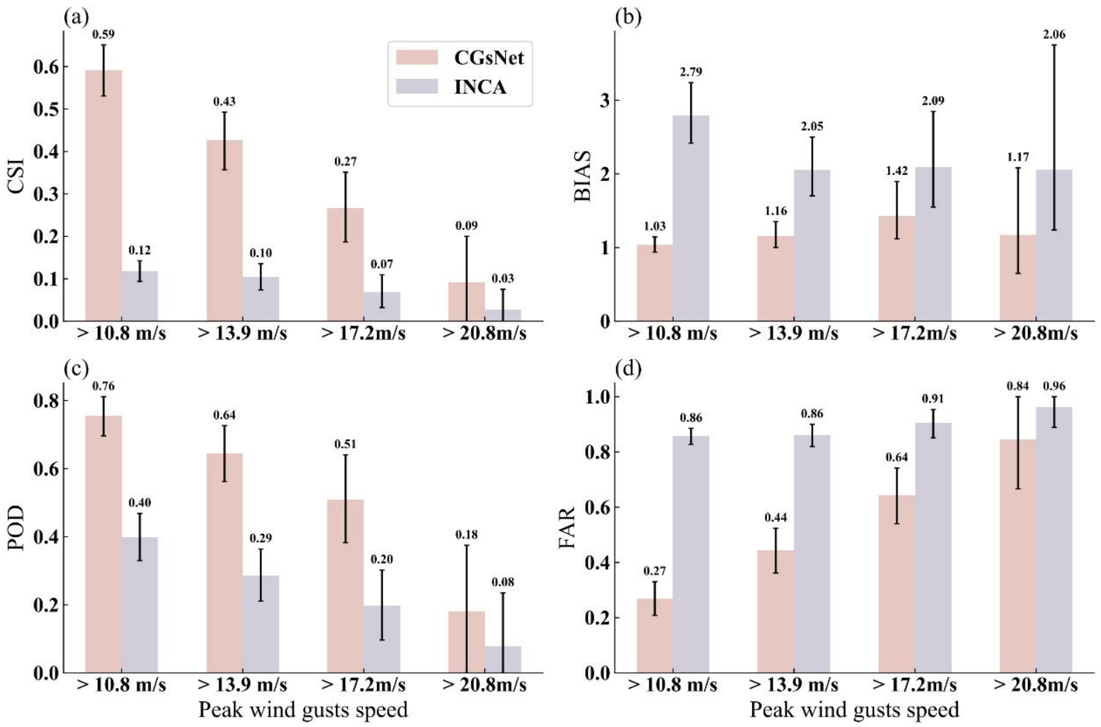


Figure 14. Same as Fig. 12, but for July 26, 2022, 14:00-16:00 BJT.

11. The model evaluation is lacking in detail. The authors present evaluation metrics based on the full testing dataset (least granularity) and some individual case studies (most granularity) -- but nothing in between (intermediate scales of granularity). The paper should include evaluation metrics as a function of time of day (e.g., by hour), time of year (e.g., by month), geographic location (e.g., maps with POD, CSI, FAR, etc. at each 1-by-1-km grid cell), and predicted wind speed (i.e., the reliability curve). At the very least, I want to see reliability curves and some division by time (either time of day or time of year). Ideally, I would like to see the evaluation metrics broken down in all 4 ways that I have listed. Presenting evaluation metrics at varying levels of granularity is crucial for understanding a model, especially for understanding its strengths and weaknesses.

Response: Thanks for the detailed suggestion. For the mentioned evaluations, we have conducted supplementary experiments: 1) evaluation of metrics as a function of time of day; 2) evaluation of metrics as a function of time of year; 3) evaluation of metrics as a function of geographic location. As for 4) the reliability curve of predicted wind speed, since CGsNet is a regression model and its output cannot be transformed into probabilities, it remains as continuous values. Therefore, the reliability curve may not be an appropriate method to evaluate CGsNet's performance. Instead, we used a scatter plot as an alternative method to evaluate the model's performance. The results of the supplementary experiments are presented below.

Figure 7 illustrates the CSI results of PWGS forecasts from CGsNet and INCA at different hours of a day. The results for POD, BIAS, and FAR are shown in Figure S4-S6. The PWGS samples were mainly obtained during the afternoon and night periods, as CGs events tend to occur during these times, particularly in the late afternoon and evening (Firouzabadi et al., 2019). The results demonstrate that CGsNet outperforms INCA in forecasting PWGS at different thresholds, with overall superior forecast performance. However, the performance of both models declines as the PWGS threshold increases. Despite CGsNet's superiority, its performance is less stable than that of INCA across different hours, exhibiting significant variability. For example, at 21:00 (BJT), the performance of CGsNet declines, and even when PWGS > 10.8 m/s, its performance (CSI and POD) is worse than INCA. At PWGS > 17.2m/s, the confidence intervals of CGsNet and INCA are both wide, while some CSI and POD values of INCA fall outside the confidence intervals and exhibit almost no predictability for CGs during 20:00-24:00 (BJT) (CSI=0, POD=0). Additionally, when the PWGS threshold is 20.8 m/s, neither CGsNet nor INCA is skillful in CGs nowcasting between 20:00 and 24:00 (BJT), with CGsNet showing a higher FAR than INCA (Figure S6). These may be attributed to the fact that strong gusts occur less frequently during these times, resulting in fewer high-value PWGS samples and highlighting the imbalance of wind data.

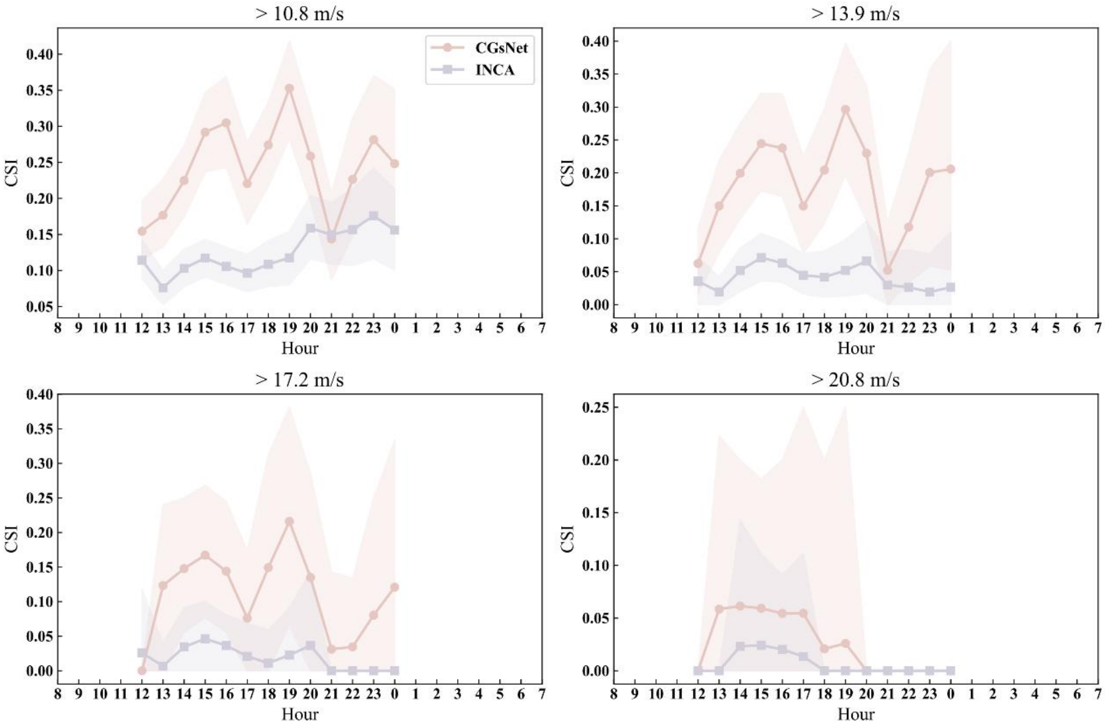


Figure 7. The CSI results of PWGS forecasts from CGsNet and INCA for different hours of a day at thresholds of 10.8 m/s, 13.9 m/s 17.2 m/s, and 20.8 m/s.The shaded pink and purple areas represent the 95% confidence intervals of the CGsNet and INCA indices, respectively.

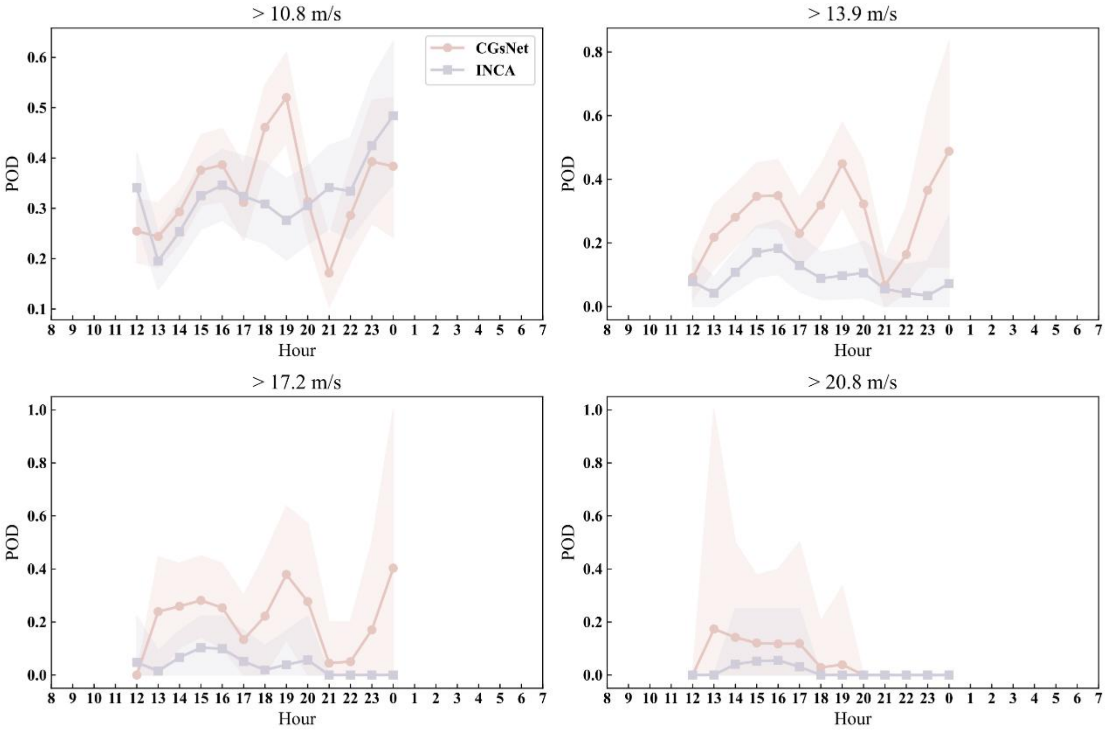


Figure S4. Same as Figure 7, but for POD.

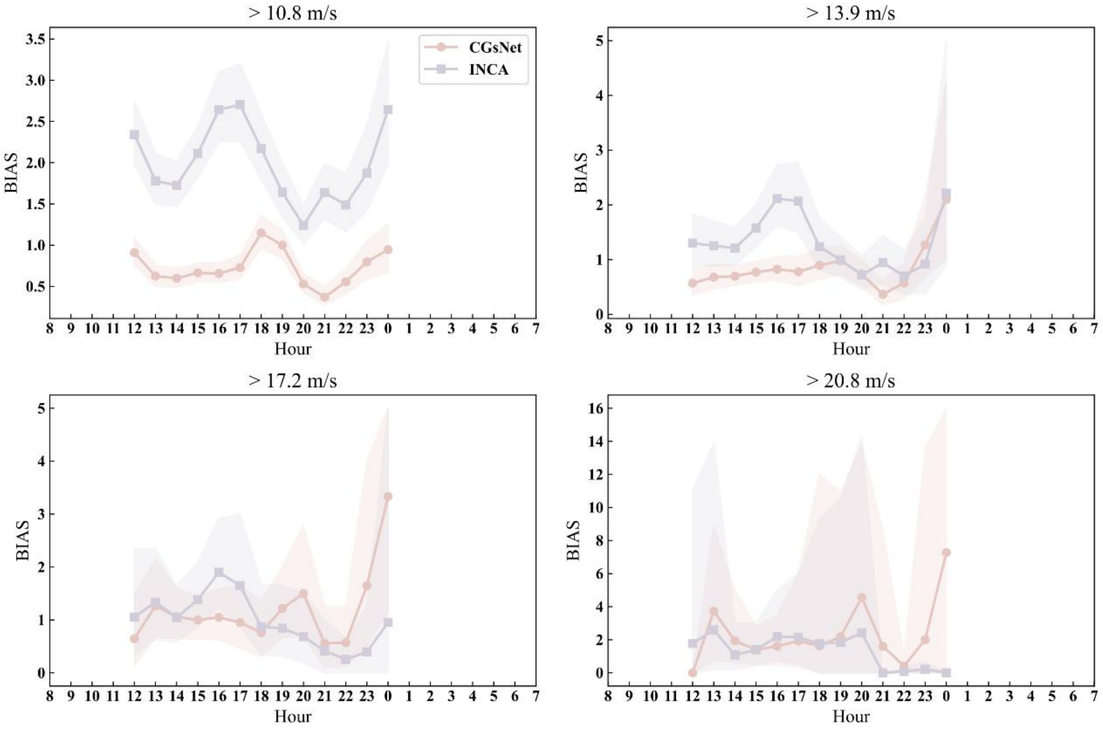


Figure S5. Same as Figure 7, but for BIAS.

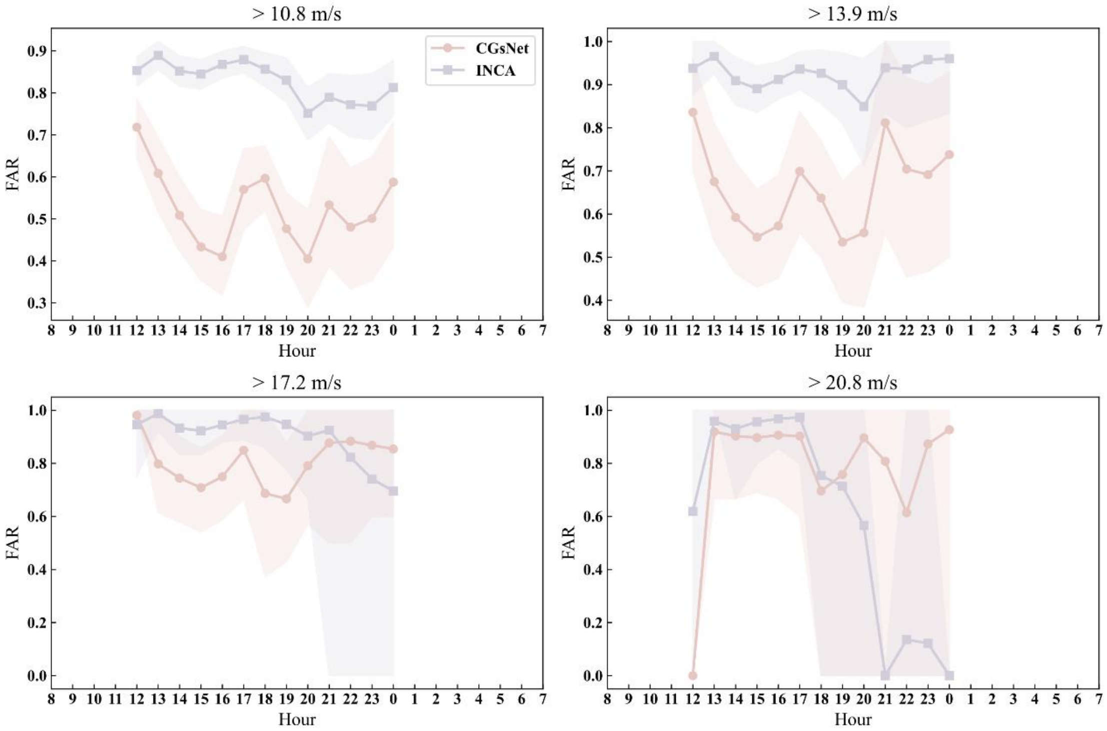


Figure S6. Same as Figure 7, but for FAR.

The evaluation results for the PWGS forecasts of CGsNet and INCA in different months of a year are illustrated in Figure 8 and Figure S7-S9. At different PWGS thresholds, CGsNet outperforms INCA in terms of CSI and POD for PWGS forecasts from May to July, while FAR is lower than INCA. However, there are large confidence intervals in the calculated evaluation metrics at PWGS thresholds of 17.2 m/s and 20.8 m/s due to the low number of strong gusts samples, leading to uncertainty. Moreover, for PWGS > 20.8 m/s, both CGsNet and INCA exhibit almost no predictability for PWGS forecasts in May and June. Additionally, CGsNet's forecast performance generally improves with increasing months, while INCA exhibits a fluctuation in June PWGS forecasting and overall better performance in July PWGS forecasting compared to May PWGS forecasting.

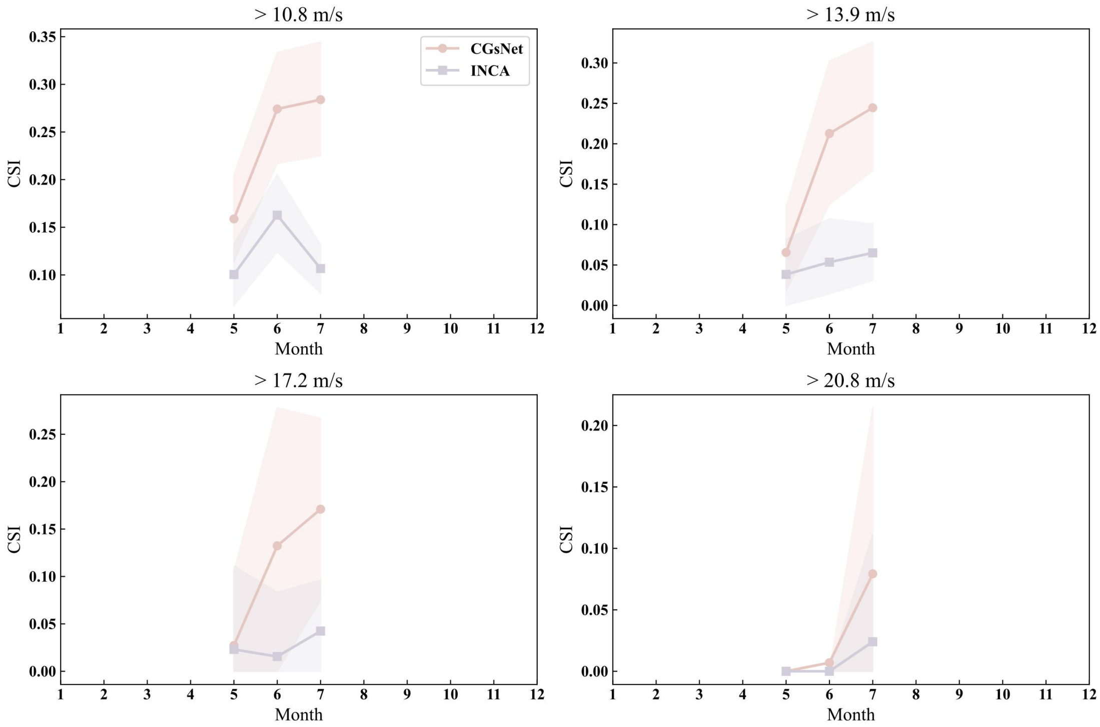


Figure 8. The CSI results of PWGS forecasts from CGsNet and INCA for different months of a year at thresholds of 10.8 m/s, 13.9 m/s 17.2 m/s, and 20.8 m/s.The shaded pink and purple areas represent the 95% confidence intervals of the CGsNet and INCA indices, respectively.

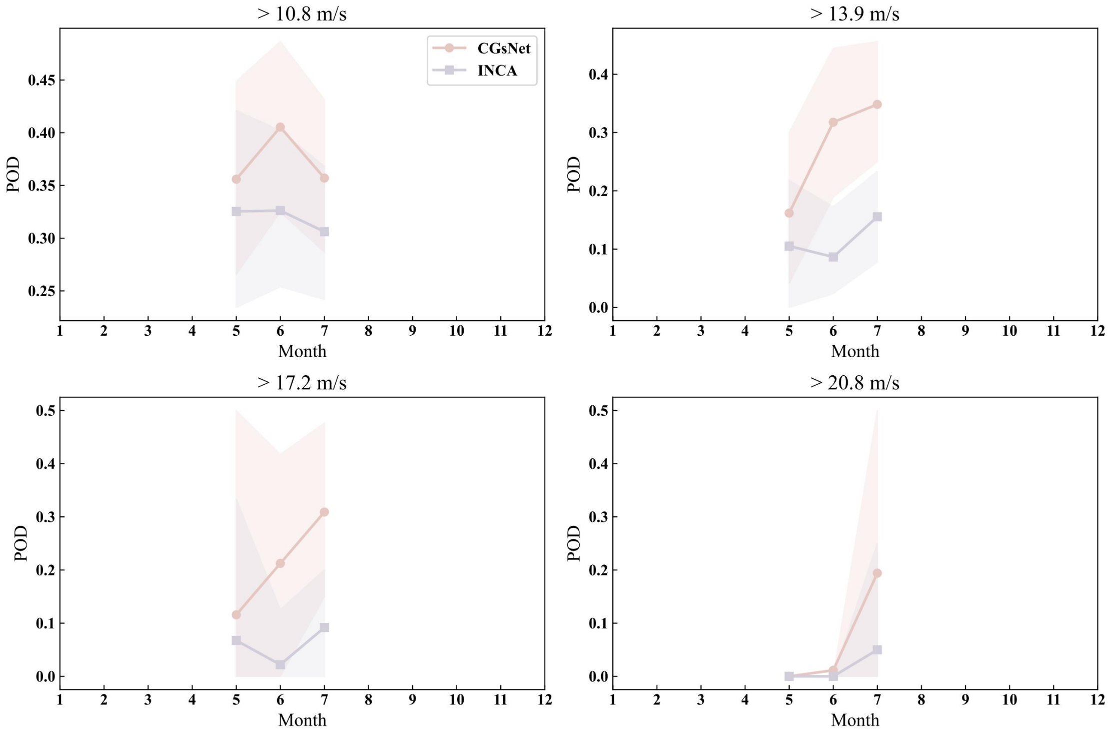


Figure S7. Same as Figure 8, but for POD.

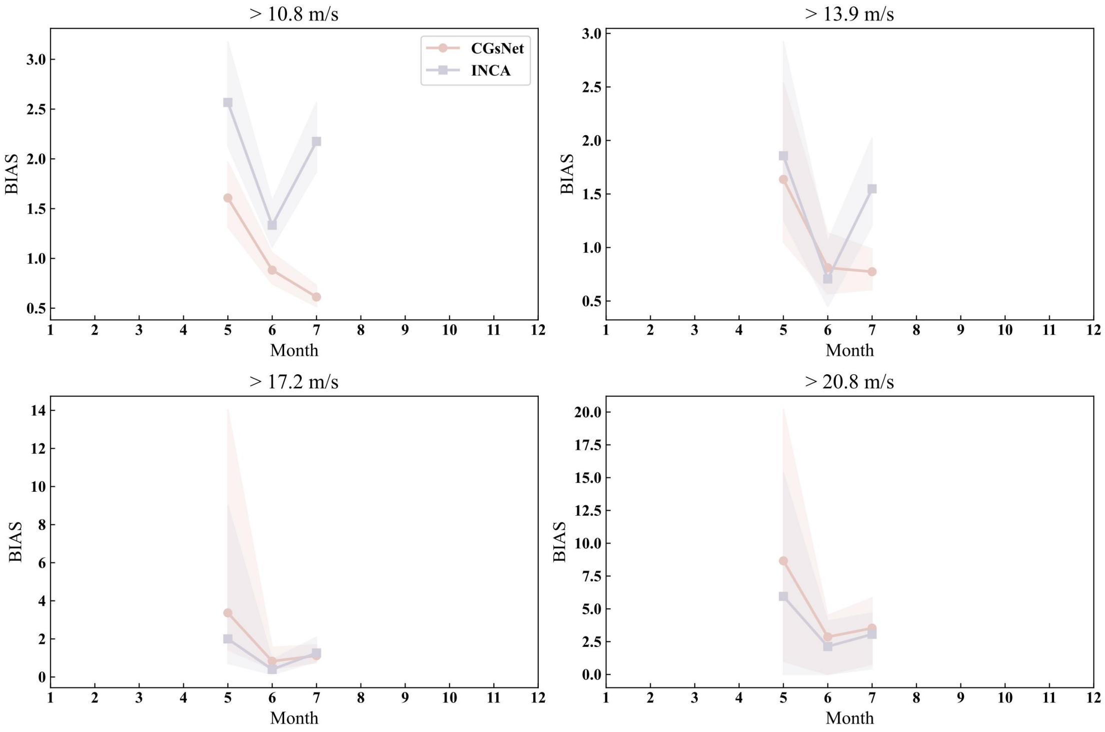


Figure S8. Same as Figure 8, but for BIAS.

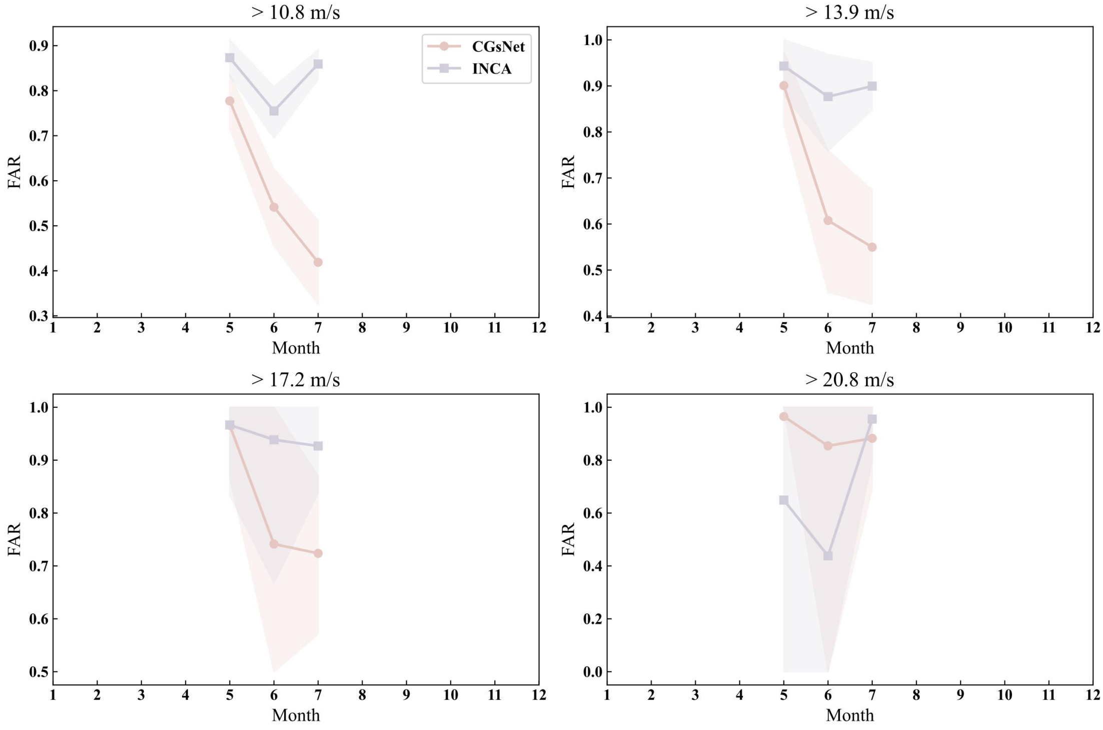


Figure S9. Same as Figure 8, but for FAR.

To better understand the performance of the model, we evaluated the forecasting performance of CGsNet and INCA for PWGS using different indicators across different geographic locations within the study area. Figure 9 displays the results for CSI, while the results for POD, BIAS, and FAR are provided in Figure S10-S12. CGsNet shows better performance than INCA in forecasting

PWGS at different thresholds. As the threshold increases, the forecasting performance of CGsNet decreases. This may be due to two reasons. Firstly, there is a reduction in the number of observed PWGS samples, as some areas do not experience CGs events, resulting in an evaluation value of 0 for those areas. Secondly, CGsNet has limitations in its forecasting ability, particularly for extreme PWGS (>20.8 m/s), which is mainly owing to the imbalance of observed strong and weak wind data. Specifically, for the PWGS thresholds are at 10.8 m/s and 13.9 m/s, CGsNet exhibits a outstanding ability to forecast PWGS for diverse regions in Jiangsu (excluding the edge areas), with most areas achieving CSI and POD values above 0.8. Conversely, INCA's forecasting performance is poor for most areas, with relatively better results in the southwest of Jiangsu. However, even in these regions, INCA's performance is still inferior to that of CGsNet. Although INCA outperforms CGsNet in terms of CSI and POD evaluation results for a PWGS threshold of 10.8 m/s in the Anhui Province region, but INCA exhibits high FAR and poor BIAS in this area. For PWGS > 17.2 m/s, CGsNet shows good forecasting results in central, northern, and southern Jiangsu, while INCA only performs well in a few stations in southwestern Jiangsu. When PWGS > 20.8 m/s, both CGsNet and INCA exhibit poor forecasting skill, and only a few stations can be effectively forecasted. Additionally, CGsNet and INCA show poor forecasting performance in regions outside of Jiangsu, due to the dominance of CGs occurring in Jiangsu in the PWGS test dataset, with limited samples from outside the Jiangsu. Obtaining more PWGS evaluation samples outside of Jiangsu in future studies could address this issue.

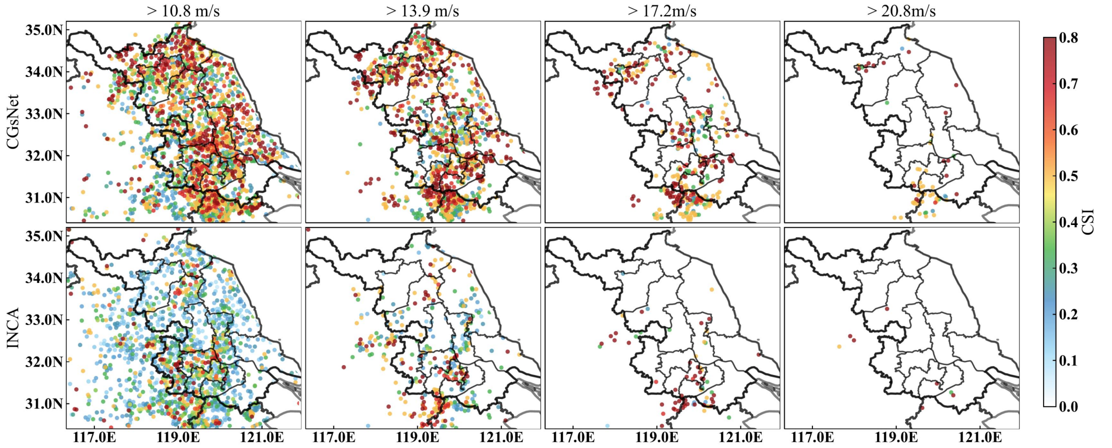


Figure 9. The CSI results of PWGS forecasts from CGsNet and INCA for different areas at thresholds of 10.8 m/s, 13.9 m/s 17.2 m/s, and 20.8 m/s.

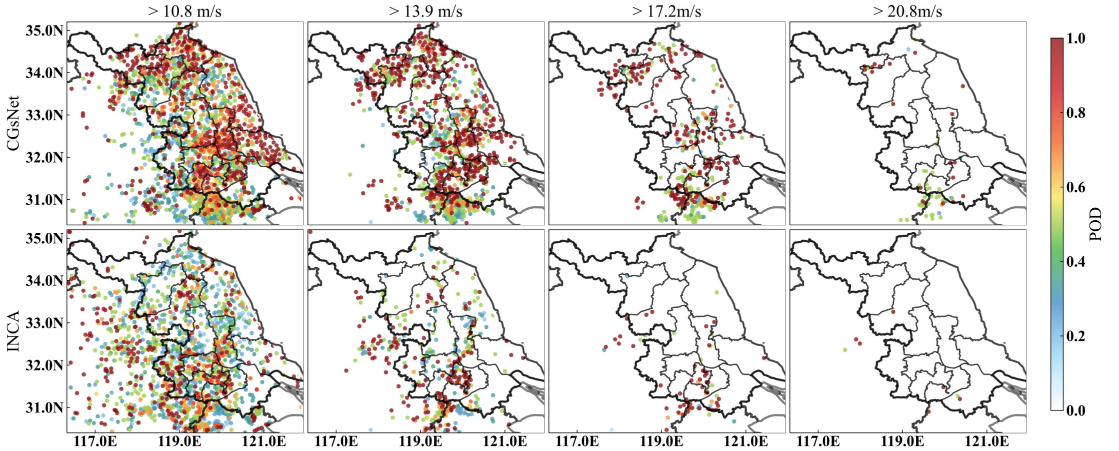


Figure S10. Same as Figure 9, but for POD.

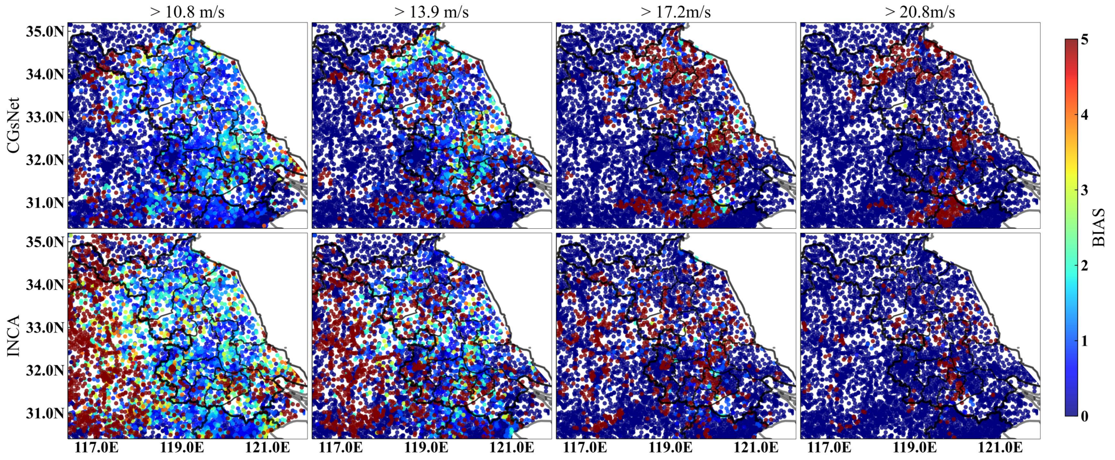


Figure S11. Same as Figure 9, but for BIAS.

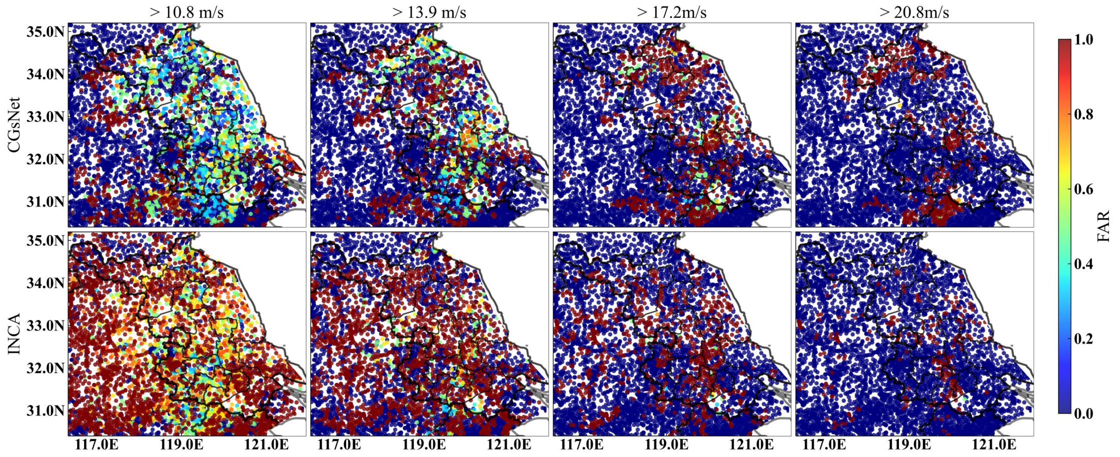


Figure S12. Same as Figure 9, but for FAR.

A scatter plot of observed and forecasted PWGS is presented in Figure 10 for further comparison. The results indicate that although CGsNet slightly underestimates PWGS, it performs well for PWGS values less than about 12 m/s. However, its performance decreases, and a bias in PWGS forecasts is observed for PWGS values greater than approximately 12 m/s. However, its performance decreases and a bias in PWGS forecasts is observed for PWGS greater than about 12 m/s. In contrast, INCA has a obvious overestimation for PWGS < 12 m/s, and the PWGS forecasts have a large deviation, not corresponding well to the observations. Additionally, the performance of INCA for high PWGS values is also poor, which is of great concern. In summary, the results indicate that the developed CGsNet is helpful to improve the accuracy of CGs nowcasting and more skillful than INCA, although it tends to underestimate strong gusts.

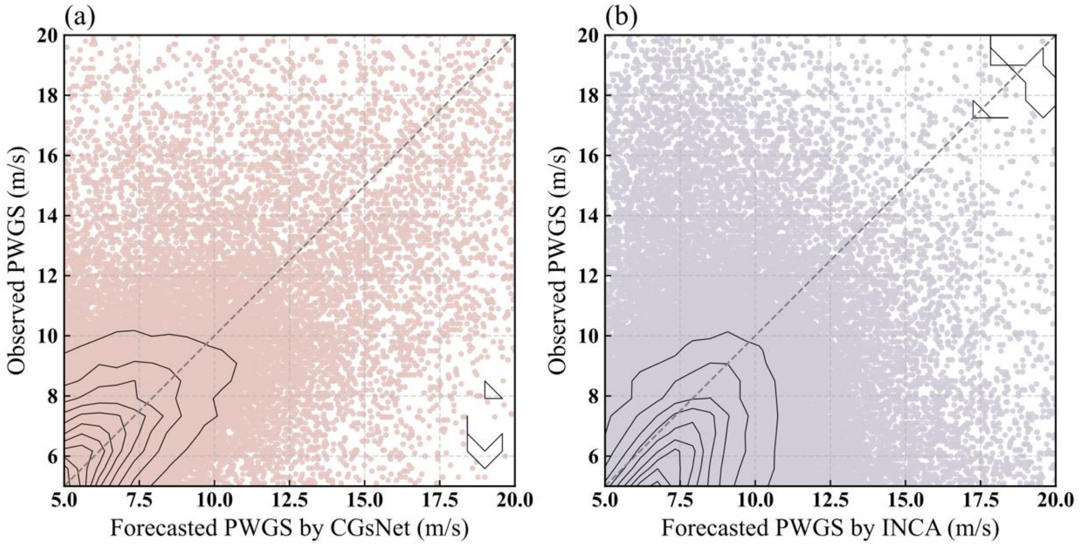


Figure 10. Scatter plot of observation and forecasted PWGS, (a) CGsNet forecasts vs. observation, (b) INCA forecasts vs. observation.

We also have added the mentioned results of the supplementary experiments into the Section 4.3.2 of the revised manuscript, and some figures (e.g., Figure S4, S5, S6, …) are in supplement of the revised manuscript.

Reference:

Firouzabadi, M., Mirzaei, M., and Mohebalhojeh, A. R.: The climatology of severe convective storms in Tehran, Atmospheric Research, 221, 34–45, https://doi.org/10.1016/j.atmosres.2019.01.026, 2019

12. The authors' interpretation of Figure 5 does not seem to be justified. See inline comments on Figure 5 itself and also lines 226-227.

Response: Done as suggested. We have modified and added the interpretation of Figure 5 in the revised manuscript (lines 263-267): "Specifically, the CGsNet model struggles to accurately forecast regions with strong gusts (ASWS > 10.8 m/s) and produced nowcasts that are slightly northward (approximately 50 km) of the observed locations. Additionally, in regions with strong gusts (ASWS > 10.8 m/s), both CGsNet and PhyDNet underestimate the ASWS values, with PhyDNet showing a larger underestimation. Some false reports are also found from both models in the areas where 8.0 m/s < ASWS < 10.8 m/s, indicating that the modeling ability of CGsNet is limited."

We have also addressed this issue in the conclusion section: "In addition to these achievements, there are some points requiring further discussion and investigation. For example, the intensity of ASWS and PWGS sometimes is underestimated (especially for strong gusts, i.e., PWGS>20.8 m/s) or there is sometimes a spatial offset between the forecasted and observed strong gusts, which may be caused by several different factors. …" (lines 430-432 in the revised manuscript)

[Minor comments]

The other minor comments suggested in the PDF:

1.The typos and grammatical errors, e.g., line 15: Typo. Replace with "nontornadic", line 83: Replace with "Thus".

Response: We have carefully checked the manuscript and corrected the pointing out typos and grammar errors.

2.Lines 40-42:"Many previous studies have mainly focused on potential severe convective weather (SCW) forecasting (McNulty, 1995; Doswell et al., 1996) or the possibility of classified SCW forecasting (Zhou et al., 2019; Lagerquist et al., 2017), while quantitative CGs nowcasting has rarely been reported." I don't understand the difference between the two approaches you're highlighting: "potential SCW-forecasting or classified SCW-forecasting" versus "quantitative CG-nowcasting". Do you mean that the previous approaches were doing classification (probability of wind exceeding a threshold) while you're doing regression (predicting the exact wind speed in m/s)?

Response:Yes. Many previous studies have primarily focused on classification (probability of wind exceeding a threshold) while we predict the exact wind speed in m/s.

3.Line 59: "objective-possibility forecasting"

Response: Thanks for pointing out this mistake. We modified this as "probabilistic forecasting".

4.Line 104 and line 146: How exactly do you separate the three datasets? Do training, validation, and testing all contain different years (example: 2016-2019 for training, 2020-2021 for validation, 2022 for testing)? Or what?

Response: Thanks for the comment. The ASWS and RMOS grid fields from 2016 to 2020 were used for model's training, and the fields from June to September 2021 and April to May 2021 were employed for ASWS validation and testing, respectively. The fields from May to July 2022 were used for PWGS testing. We have clarified this in the revised manuscript in lines 163-166.

5. Line 192: Why these specific ranges? Also, why is there an upper bound on these ranges? Why not just evaluate ASWS in the range > 8.0 m/s and PWGS in the range > 10.8 m/s?

Response:Actually, we evaluated ASWS in the range > 8.0 m/s and PWGS in the range > 10.8 m/s. Originally, our intention was to express graded evaluation here, not just evaluate this specific range. The evaluation of wind speed levels is primarily based on the wind speed classification determined by the China Meteorological Administration. We have modified this in the revised manuscript (lines 216-217) as:" In particular, we concentrate on evaluating ASWS at thresholds of 8.0-13.9 m/s (e.g., ASWS > 8.0 m/s or ASWS > 13.9 m/s) and PWGS at thresholds of 10.8-20.8 m/s in CGs events."

6. Lines 214-215: I'm unsure what to think about all your subjective assessments of skill (the adjectives "outstanding," "decent," "skillful," etc.). Is there a baseline, i.e., another method you can compare against? Because a lot of these results -- especially for the thresholds of 10.8 and 13.9 m/s, but also for the longer lead times at all thresholds -- seem quite poor.

Response: Thanks for reminding this. We have revised the sentences to avoid subjective evaluation as much as possible. Furthermore, we have included an ablation study on the proposed attention module, with PhyDNet serving as the baseline. A detailed analysis and description of this can be found in Section 4.1 and 4.2 of the revised manuscript.

7. Lines 219 and 284: Northeast Cold Vortex--This is not a proper name, so it should not be capitalized.

Response: Thanks for pointing out this. After carefully checking, we modified "Northeast Cold Vortex" as "northeast China cold vortex" (Niu et al., 2021) in the revised manscript.

Reference:

Niu, Z., Zou, X., & Li, D.: Northeast China Cold Vortex Observed by FY-3 MWTS-2 and MetOp AMSU-A, Journal of Geophysical Research: Atmospheres, 126(23), e2021JD035471, 2021.

8. Line 275 "Meanwhile, the BIAS and FAR of CGsNet are also lower than those of INCA.": I have two problems with the statement about bias:(1) This is not true. For the threshold of 20.8 m/s, CGsNet has a greater bias than INCA. (2) Lower bias is not always better. For example, suppose that the two bias values are 0.50 and 1.00.

Response: Thanks for pointing out this. We have modified this as "The FAR of CGsNet is also lower than that of INCA. Meanwhile, except for the threshold of 20.8 m/s, CGsNet consistently exhibits better BIAS than INCA." in the revised manscript (lines 315-316).

9.Line 304: 'the movement and development derecho'.

Response: A derecho (pronounced similar to "deh-REY-cho") is a widespread, long-lived wind storm that is associated with a band of rapidly moving showers or thunderstorms (https://www.weather.gov/lmk/derecho). The CGs case on 26 July 2022 was influenced by a derecho.

10. Line 342:Or you could compute site-specific GFs, direction-specific GFs, etc.

Response:Thanks for the suggestion. We have added this into the revised manuscript in line 441.