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 $0.01^{\circ} \times 0.01^{\circ}$ 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.

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.

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.

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.

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

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.

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.

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 $\{h_{t-4}^m, h_{t-3}^m, h_{t-2}^m, h_{t-1}^m\}$, and when prediction $t+1$, the attention module can assign weights for $\{h_{t-4}^m, h_{t-3}^m, h_{t-2}^m, h_{t-1}^m\}$ as well, or for $\{h_{t-3}^m, h_{t-2}^m, h_{t-1}^m, h_t^m\}$.

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

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

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.

III **EXPERIMENT**

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