Review of the manuscript

entitled: “Predicting Forecast Errors with Diffusion Model for Uncertainty Quantification in Wind Speed Nowcasting”

of authors: Yanwei Zhu, Aitor Atencia, Markus Dabernig, Yong Wang, Shuyan Zhou

Suggestion: reconsidered after major revisions

General comments:

The manuscript describes and evaluates an ensemble nowcasting system, which has been constructed with use of a denoising diffusion probabilistic model (DDPM). Though diffusion models (DM) are recently intensively studied in meteorology and weather forecasting, their ability to simulate the wind uncertainty has got less attention. Thus, the presented results can be very informative for those who would like to build a complex probabilistic nowcasting system based on a diffusion model. The authors succeeded to demonstrate with both statistical evaluation and case studies that such a model could be used for nowcasting of wind speed. The DDPM itself and methods used for its evaluation are generally well described and justified, with some exceptions (which will be commented below). It must be also appreciated that the performance of several noise schedules (Linear, Cosine, Sigmoid) was thoroughly tested and analyzed. Still, there are several aspects where the model could be improved, above all concerning extreme events – as it is admitted in the conclusion of the manuscript. The authors claim several times that one of the biggest advantages of the DM-s against other AI/ML techniques is the maintenance of physical consistency. But the evaluation is done only with respect to their own analyses and reference forecasts, one cannot find comparisons of the DDPM outputs with performance of AI models based on different principles. Thus, these statements would need some clarification. Also, the presented case studies could be analyzed and discussed with more details.

The manuscript can be considered as an interesting pilot study, introducing the application of DDPM in wind nowcasting. Despite the above-mentioned weaknesses and some deficiencies in the presentation of the schemes and results, the manuscript exhibits an overall good quality and after corrections and clarifications, it could be appropriate for publication.

Specific comments:

Between the Lines 80-85: “The learning objective of DDPM is the error of SIVA nowcast, defined as the difference between forecast and the corresponding analysis field.“

Reviewer (R): *Usually, objective analysis also contains errors, above-all in grid-points laying far from observations, in complex terrain, etc. How it is in the case of SIVA nowcast, how much could the error of the analysis influence the DDPM training and the magnitude of forecast errors? Can you estimate somehow the accuracy of analyses used (e.g. have you done cross-validation)? Can you comment on this?*

85: *The Fig. S1 in the supplementary materials shows the domain only very schematically. One has no information about the orography of the area (e.g. are there also high mountains?), which can be important from the wind nowcast point of view. Also, it is rather unusual that the domain is presented in the supplementary material and it is not one of the figures of the manuscript.*

85-90: „The data was spilt into nonoverlapping parts for training (1 October 2021–30 April 2023), validation (1 May 2023–31 May 2023), and testing (remainder).“

R: Later in the conclusion of the manuscript (lines 350-355) you mention that DDPM forecasts are still too smooth to reproduce some extreme events. Is it perhaps related to too short (~1,5 year) training period? Was there any particular reason for choosing the period of this length?

100-105: *In the Section 3.1., the meaning of variables „q“ and „p“, which are products of Equations (1) and (2) is not explicitly mentioned, though one could guess that these are probability distributions. Also „I“ is not explained (Identity matrix?). I would recommend to specify it, for bigger clarity.*

Figure1: *The figure might help the reader to understand the training process. But I do not find it to be self-explanatory enough and it takes some time to interpret it. Hence I would propose to improve this Figure and to relate it better with the corresponding text in 3.2. and equations in 3.1. For example, one should easily find, where is the start of the training (this is the creation of „Errors“ in the upper left corner of the Figure). You could highlight this step somehow (e.g. with a letter or number) and refer to it also in the text of 3.2. Similarly also some other parts of the DDPM framework, the cycle of the sampling process, etc. as one might understand in detail, what is depicted on the Figure. There is also a 3D graph (maybe a distribution function, product of the Denoiser?) on the middle, right hand side of the Figure, which has no name and explanation. Please, amend that as well.*

150: „For verification, the SIVA forecast errors are designated as the reference for evaluating the DDPM, while the analysis fields serve as benchmark for the ensemble nowcasts.“

R: *Did you use every grid-point of the SIVA domain and the analysis for verification? This is not clear, as in the caption of the Fig. 2 you denote observations as „Truth“, while in the caption of the Fig. 4. the analysis field is denoted „Truth“. It should be clarified in the Metrics description that where are you using point observations and where nowcasting software analysis for verification and why. It must be also taken into account that the analysis can already exhibit errors so it can be considered only as a near description of the real state of the atmosphere.*

155-160: „Given the negligible sensitivity of key verification metrics to ensemble size, an ensemble with 16 members was adopted to ensure statistical robustness while maintaining computational tractability.“

R: *Can you support somehow your statement that the key verification metrics is not sensitive to ensemble size? Have you verified that? In the introduction, lines 35-45 you mention that „The ensemble often fails to fully characterize the true probability distribution“ or „However, a finite number of members cannot fully represent the true distribution, which inevitably leads to under-dispersion.“ This is seemingly in contradiction with your statement in the Evaluation Metrics part. How did you choose the 16 members then? Upon which criterion?*

165-170: „The agreement of both the joint and marginal probability distributions with the benchmark (Fig. 2) demonstrates that the errors predicted by DDPM are physically consistent and statistically robust.“

R: *Why do you think that if the statistical distribution of the evaluated model's forecasts and forecast errors matches the benchmark (which is considered to be physically consistent), then it must be physically consistent, too? Can you prove it? Or can you cite examples that if the physical consistency would not be fulfilled (e.g. using different approaches) then the forecast error distribution would be different?*

205-210: First sentence of 4.2: „Evaluations of the generated errors reveal that DDPM captures the physical characteristics of forecast errors, thereby learning more than just their statistics.“

R: *As for 165-170, I am not convinced that this was really shown as there was no comparison with other methods, where we would expect “only” learning statistics.*

255-265 and a part of the Caption in Figure 7:

R: *The discussion on the Brier Score and Brier Skill Score should be moved to 3.3 (Evaluation Metrics). Also the note on the use of climatology probability in BSS calculation. You could also mention, why the climatology probability was chosen to be a reference. Though a standard practice, it may have some implications for verification of rare events, especially when you used only the training dataset (~1,5 year).*

285-295: Figure S2: *You have quite a lot of description (one paragraph) concerning the output of Fig. S2 (Probability diagram of wind ensemble nowcasting in different lead time for three schedules with threshold 1 ms^{-1}). It would be fair to present it as a part of standard Figures of the manuscript, e.g. as Fig. 9a. The other diagram concerning the higher threshold of 10.8 ms^{-1} could be the Fig. 9b.*

295: *You probably erroneously refer to Fig. S2, while the diagram in the current Fig.9 (for the threshold 10.8 m/s) is not referenced in the text at all. As mentioned above, consider moving Fig. S2 to Fig. 9a and denote current Figure 9 as Fig. 9b.*

315-325 and Figure 10: *You mention Comparison with "truth" or "Ground Truth" although this is probably only the analysis of your deterministic nowcasting system and it may contain errors as I have mentioned earlier. It would be probably better to denote it as Analysis.*

Figure 10: *A mistype occurred in the caption of the figure: Instead of "Ground Turth" there should be "Ground Truth" or even better "Analysis".*

Figure 10: "(c) Ensemble Mean forecast"

R: *I do not understand why you did not show the forecasts of ensemble maxima, not even in the supplementary materials. Though, this is a parameter, which is often used in weather forecasting, especially for severe weather warnings. And you even note the "smoothing effect on the ensemble mean" in the text below the Figure (line 330). How could we know that the "strong wind band" considered as false alarm (319-320) and appearing in the deterministic nowcast (REF) on Fig. 10b would be not reproduced by some of the DDPM ensemble members? Only comparison with ensemble maximum of wind could exclude that. In the supplementary Fig. S3 one can see that although less expressed than in REF, but certain members (e.g. 9,10,12) show a signal for such wind band in that area.*

330-335 “The smoothing effect on the ensemble mean was also reduced, and the probabilistic forecast shows finer and more accurate details.”

R: *It would be nice to mention an example. In addition, it would be perhaps noteworthy to highlight as an interesting feature that there is an area of reduced wind speed on the right edge of the domain (nearly in the middle, over the sea), visible in all outputs valid for the +1h time (Figure 11). Although it vanishes in both analysis and reference as the wind strengthens in time, it remains in the DDPM forecast until +6 hour, which suggests that the DDPM can exhibit certain “inertia” in some cases, even against its reference.*

350-355 “However, for extreme events, the model still suffers from the issue of excessive smoothness. This limitation may stem from the use of a basic diffusion model architecture.”

R: *Is there not an additional problem that you used a relatively short (~1.5 year) training period? See my previous comment for the lines 85-90.*

370-375 “… while maintaining computational efficiency and physical consistency.”

R: *Can you specify how much is DDPM computationally efficient? E.g. against a traditional nowcast (e.g. of your reference nowcast or some different nowcasts/ensembles).*

Review criteria:

1. Does the paper address relevant scientific modelling questions within the scope of GMD? Does the paper present a model, advances in modelling science, or a modelling protocol that is suitable for addressing relevant scientific questions within the scope of EGU?

   R: Yes, The question of using AI/ML methods and DM-s for high resolution nowcasting systems is a very relevant topic in current weather forecasting.

2. Does the paper present novel concepts, ideas, tools, or data?

   R: Yes, several ones (use of DM-s for wind nowcasting, analysis of results with respect to different schedules, etc.).

3. Does the paper represent a sufficiently substantial advance in modelling science?

R: Yes, the use of DM-s in wind nowcasting as presented is advantageous due to its computational efficiency and ability of direct estimation of forecast errors. This could enable to construct forecast ensembles, which would be difficult and computationally very demanding with classic NWP approaches at such high spatial and temporal resolution.

4. Are the methods and assumptions valid and clearly outlined?

   R: Mostly yes, see the specific comments above for the exceptions.

5. Are the results sufficient to support the interpretations and conclusions?

   R: Mostly yes, see the specific comments above for the exceptions. Some doubts are concerning physical consistency of the produced forecast errors and results of the case study.

6. Is the description sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? In the case of model description papers, it should in theory be possible for an independent scientist to construct a model that, while not necessarily numerically identical, will produce scientifically equivalent results. Model development papers should be similarly reproducible. For MIP and benchmarking papers, it should be possible for the protocol to be precisely reproduced for an independent model. Descriptions of numerical advances should be precisely reproducible.

   R: I think that a reviewer could hardly verify this in reasonable time. But from the description of the model and methods and upon the claim in the "Code and data availability" (all the links were available by 17 May 2026) I believe that most of the results, if not all, are reproducible.

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution?

   R: Yes.

8. Does the title clearly reflect the contents of the paper? The model name and number should be included in papers that deal with only one model.

   R: Yes.

9. Does the abstract provide a concise and complete summary?

   R: Yes.

10. Is the overall presentation well structured and clear?

    R: Mostly yes, see some suggestions how to improve it in the specific comments.

11. Is the language fluent and precise?

R: Yes.

12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?

    R: I missed the description of some variables (though probably obvious for experts in DM modeling area) and asked for amendment. See the specific comments above.

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?

    R: I made suggestions for some rearranging of the figures (Fig. S2, Fig. 9, etc.) – see the specific comments.

14. Are the number and quality of references appropriate?

    R: Yes.

15. Is the amount and quality of supplementary material appropriate? For model description papers, authors are strongly encouraged to submit supplementary material containing the model code and a user manual. For development, technical, and benchmarking papers, the submission of code to perform calculations described in the text is strongly encouraged.

    R: There are links to code and data availability after conclusion, which might be sufficient for readers, who are highly interested in testing the DDPM. I suggested to move certain figures from the supplementary material to the main part of the manuscript as these were mentioned and discussed in the text and seemed to be important (see the specific comments). As a reader, I would personally prefer only such figures among supplementary materials, which are for "further reading", as a kind of appendix, and are not as important and not discussed in the main part of the text. To search for both "standard" figures and supplementary figures during reading is a little bit disturbing to me.