General statement

We would like to thank the editor for coordinating the review of our work and the peer-reviewers for their valuable comments on our study. In the following, we will address the referees' comments and present our plans and ideas for revising the manuscript. For clarity, our responses are highlighted in blue.

Referee comment #1

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.

Reply:

We thank the reviewer for the positive and constructive assessment of our work. We will carefully address all the comments and revise the manuscript accordingly. In particular, we will remove the unsupported claims about physical consistency, add more details to the case studies, and clarify the limitations of the current study. The revisions will be traceable in the revised manuscript.

Specific comments:

#Comments 1

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?

Reply 1:

Thanks for the comments. We agree that the analysis contains errors. To quantify its accuracy, we performed a cross-validation experiment using the same station data that go into the analysis as described in section 2. The stations were split into a training set (90% stations) used to produce the analysis, and a test set (10% stations) as out-of-samples. The cross-validation was performed separately for the training, validation, and test periods, with consistent results across all three. Table 1 shows the results for the test period (the period used to evaluate our DDPM results).

Table 1. Cross-validation scores of the SIVA analysis

Station set	train	test
bias (m s^{-1})	0.149	0.133
mae (m s^{-1})	0.351	0.725
rmse (m s^{-1})	0.454	0.943
Avg. wind (m s^{-1})	1.73	1.74

The results show that the analysis has a mean RMSE of about 0.94 m s^{-1} on out-of-sample stations. The spatial maps (Fig. 1) show that errors are a bit larger in the two mountainous areas (30~31.5°N, 116~117°E and 30~30.9°N, 117~120°E) in the southwest (400–1200 m elevation). Published results from the SIVA system (Zhu et al., 2025) and its older version INCA (Haiden et al., 2011) also support that the analysis is a reasonable approximation of the true state. Given this moderate error level, we treat the analysis as a practical approximation, acknowledging that it may introduce some uncertainty into our results.

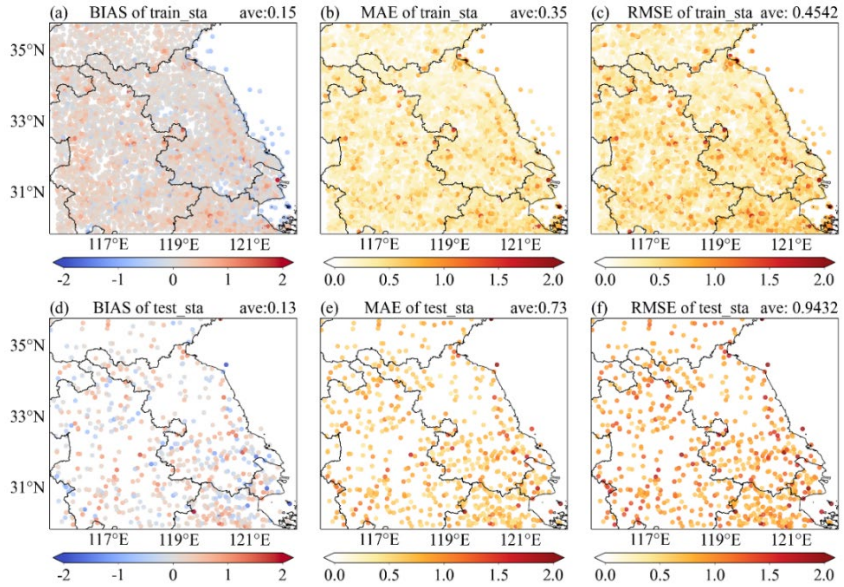

Fig.1. Spatial distribution of bias, MAE, and RMSE for the training stations (a–c) and test stations (d–f) in the test period. (a, d) Bias; (b, e) MAE; (c, f) RMSE.

Regarding the impact on DDPM training and the magnitude of forecast errors: we agree that errors in the analysis could affect both. The cross-validation results show that the analysis has relatively small errors (RMSE ~0.94 m s^{-1} on out-of-sample stations, and 0.45 m s^{-1} on training stations) with a mean wind speed of ~1.74 m s^{-1}. Given this overall error level, we believe it is

unlikely to overturn our main conclusions. Based on this moderate error level, we treat the analysis as a gridded approximation of the true state. We have not deeply investigated how the analysis errors propagate into the DDPM training or the final ensemble nowcasts. However, we fully agree with the comment that if one aims to represent forecast uncertainty against real observations, the influence of analysis errors should be considered. This is a clear limitation of the current study, and we will note it as an important direction for future work, where independent station observations will be used to validate the results and to better understand the role of analysis errors. We will also add a short note in the Conclusions section.

References:

Zhu, Y. W., Atencia, A., Dabernig, M., and Wang, Y.: Quantifying the analysis uncertainty for nowcasting application, Geosci. Model Dev., 18, 1545–1559, 2025.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region, Weather Forecast., 26, 166–183, 2011.

#Comments 2

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.

Reply 2:

Thanks for the suggestion. We agree that presenting the domain map only in the supplementary material is unusual and that orographic information is crucial for wind nowcasting. Accordingly, we will move this figure to the main text as the new Figure 1 and add a subplot for topographic information. The revised figure and its caption will be traceable in the revised manuscript.

#Comments 3

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?

Reply 3:

Thanks for the comment. The training period (1 October 2021 – 30 April 2023) was determined by the dataset that was accessible to us under the terms of our data agreement; a longer continuous period was not available at the time of this study. We agree that a longer training period would likely improve the representation of extreme events and reduce the excessive smoothness noted in the conclusion. We will therefore revise the conclusion to include the short training period as a contributing factor i.e.: “This limitation may stem from the use of a basic diffusion model architecture as well as the relatively short training period (~1.5 years), which likely does not provide sufficient samples of extreme wind events”. The revision will be traceable in the manuscript.

#Comments 4

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.

Reply 4:

Thanks for the suggestion. We will add explicit definitions of q, p, and I in Section 3.1. Specifically, we will state that $q(x_t|x_0)$ is the conditional probability distribution of x_t after adding t steps of Gaussian noise to x_0, with I as the identity matrix; and that $p_\theta(x_{t-1}|x_t)$ is the reverse conditional probability distribution for recovering x_{t-1} from x_t, parameterized by a neural network. The revision will be traceable in the manuscript.

#Comments 5

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.

Reply 5:

Thanks for this detailed suggestion. We agree that the current figure is not self-explanatory enough. We will redesign the figure as follows:

- Add numbered steps (e.g., ①, ②, ⋯) to clearly indicate the start of training (the creation of "Errors" in the upper left) and the main stages of the DDPM framework.
- Include a clear indication of the sampling process cycle.
- Label the 3D graph on the right hand side (e.g., "learned error distribution").
- Reference these numbered steps in the text of Section 3.2 to link the figure with the equations in Section 3.1.

The revised figure will be included in the revised manuscript.

#Comments 6

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.

Reply 6:

Thanks for the comment.

We confirm that all verification uses every grid-point of the SIVA domain (e.g., error distributions, reliability diagrams, CRPS). In this study, the verification is based on the analysis field, which we treat as a gridded approximation of the true state (as noted in the response to #Comment 1) and we do not use point-wise station observations for verification in the present work. We will revise the sentence in Section 3.3 i.e.:

"In this study, we treat the gridded SIVA analysis field as a practical approximation of the true state. The forecast error of the deterministic nowcast is defined on the model grid as the difference between the SIVA forecast and the corresponding analysis field. For verification, we perform two separate evaluations. First, the errors generated by the DDPM are compared against the original SIVA forecast errors to assess how well the DDPM predicts these errors. Second, the ensemble nowcasts (obtained by adding the DDPM-generated errors to the SIVA forecast) are evaluated against the analysis field (as a surrogate for the true state) to assess the quality of the final probabilistic forecasts. All verification is performed on all grid points of the SIVA domain using standard probabilistic metrics. We note that the analysis is not a perfect truth; quantifying the impact of its residual errors on our results is left for future work, which will also include verification against independent station observations."

We will replace all instances of "truth" or "ground truth" with "analysis" throughout the manuscript (including figure captions) to avoid confusion between analysis and observations.

The revision will be traceable in the manuscript.

#Comments 7

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?

Reply 7:

Thanks for the comment. When designing the ensemble, we compared 8, 16, 32, and 50 members using CRPS, RMSE, and the RMSE/spread ratio. The figure 2 shows the results. The 8 members clearly underperforms the others. However, the differences among 16, 32, and 50 are minimal – CRPS differs by less than 0.01, and the RMSE/spread ratio differs by less than 0.02. Therefore, we chose 16 members as a balance between statistical robustness and computational cost (2 minutes per forecast cycle). Larger sizes (32–50) would take 4–8 minutes with virtually no improvement.

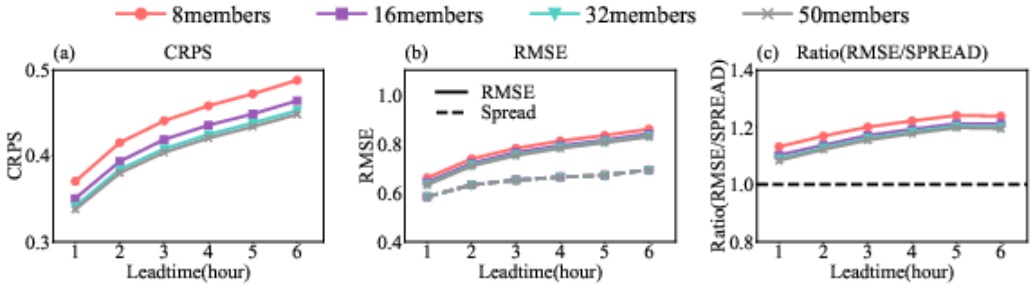

Fig.2. Verification metrics for ensemble nowcasts with different members over the test period. (a) CRPS, (b) RMSE (solid lines) and ensemble spread (dashed lines). (c) Ratio of RMSE to SPREAD.

Regarding the apparent contradiction with the introduction: the description refers to classical ensemble methods where the true distribution is unknown and a finite sample inevitably leads to under-dispersion. But our model is trained to learn the target distribution of forecast errors. Once trained, generating an ensemble member is simply a random draw from that learned distribution. The distribution itself is already captured; increasing the number of draws beyond a certain point (here 16) brings negligible improvement. Thus, there is no contradiction. The introduction describes the general challenge, while our method addresses it by learning the distribution.

We will revise the text in Section 3.3 to clarify this reasoning and remove the phrase "negligible sensitivity" to avoid overstatement, i.e. "A comparison of ensemble sizes 8, 16, 32, and 50 shows that the 8-member ensemble underperforms the larger sizes, while the differences among 16, 32, and 50 members are minimal (e.g., CRPS differs by less than 0.01). Therefore, an ensemble size of 16 was chosen as a balance between statistical robustness and computational cost." The revision will be traceable in the manuscript.

#Comments 8

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?

Reply 8:

Thanks for the comment. We agree that matching the probability distributions alone does not prove physical consistency. Therefore, we will revise this sentence to state that the errors are "statistically consistent" rather than "physically consistent", i.e. "The agreement of both the joint and marginal probability distributions with the benchmark (Fig. 3) demonstrates that the errors predicted by DDPM are statistically consistent." The revision will be traceable in the manuscript.

#Comments 9

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.

Reply 9:

Thanks for the comment. We agree that without comparison to other methods we cannot claim that the DDPM learns “physical characteristics” beyond statistics. Therefore, we will revise the sentence in Section 4.2 to simply state what the evaluations actually show, i.e., that the DDPM reproduces the statistics of the forecast errors. The revised sentence reads: “Evaluations of the generated errors reveal that the DDPM reproduces the statistics of the forecast errors.” The revision will be traceable in the manuscript.

#Comments 10

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).

Reply 10:

Thanks for the suggestions. We will make the following revisions:

1. Move the definitions of Brier Score and Brier Skill Score from the results section to Section 3.3 (Evaluation Metrics).

2. Add a note in Section 4.2 (Verification of Ensemble Nowcasts) i.e.: “It should be noted, however, that the climatological probability used as the reference for BSS is estimated from the training period (~1.5 years). This choice follows standard practice for probabilistic forecast verification (a no-skill benchmark). For rare events such as wind speeds exceeding 10.8 m s^{-1}, this short period yields a limited sample size, which may affect the robustness of the estimated climatology. This is a possible reason why the differences among the three noise schedules are less pronounced compared to the lower threshold. In future work, a longer climatological reference should be used to better evaluate probabilistic forecasts for rare events." The revision will be traceable in the manuscript.

3. Following the suggestion, we will remove the methodological sentence from the caption of Figure 7 (“The reference value in calculating BSS is the climatology probability...”) and move this information to Section 3.3, where the BSS definition is now given. The caption of Figure 7 will only describe the results.

#Comments 11

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.

Reply 11:

Thanks for the suggestion. We will move Figure S2 to the main text and merged it with the original Figure 9. The merged figure will be Figure 10, with panel (a) for the 1 m s^{-1} threshold and panel (b) for the 10.8 m s^{-1} threshold. The cross-references in the text will be corrected accordingly, i.e.: "Figure 10a shows the reliability diagrams for the 1 m s^{-1} wind speed, which lies at the centre of the observed probability distribution." and "The ensembles maintain high reliability at high wind events with threshold 10.8 m s^{-1} (Fig. 10b), exhibiting temporal evolution characteristics consistent with those observed for the 1 m s^{-1} threshold (Fig. 10a)." The revision will be traceable in the manuscript.

#Comments 12

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.

Reply 12:

Thanks for the hint. We will correct the cross-references and now explicitly cite the high-threshold diagram (Fig. 9, now Fig. 10b after merging) in the text. Please see the response to **#comment 11** for details.

#Comments 13

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.

Reply 13:

Thanks for the suggestion. We will replace "truth" and "ground truth" with "analysis" throughout the manuscript.

#Comments 14

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

Reply 14:

Thanks for pointing this out and sorry for the mistake. Following the suggestion, we will replace all instances of "truth" or "ground truth" with "analysis" throughout the manuscript to avoid confusion. For the caption of this figure, we will correct the typo as follows: "Fig. 11. Wind speed from 08:00 to 13:00 UTC on 10 June 2023. (a) Analysis, ..." The revision will be traceable in the manuscript.

#Comments 15

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.

Reply 15:

Thanks for the comment. We agree that stating the smoothing effect of the ensemble mean while calling the strong wind band in the deterministic forecast a false alarm is inconsistent without checking the ensemble maximum. To resolve this, we will examine the ensemble maximum and added it to Figures 11 and 12 (new panels). The ensemble maximum in Figure 11 does show the strong wind band, but with lower probability. Therefore, we will revise the text accordingly i.e.: "The ensemble mean reduced the bias and captured the spatial structure more accurately, but the strong wind band was largely smoothed out (Fig. 11c). The ensemble maximum (Fig. 11d) and individual members (e.g., members 9, 10, 12 in Fig. S1) still show a weaker signal in that area. Thus, the band is not a complete false alarm; rather, the ensemble represents it with a non-zero but lower probability." The revision will be traceable in the manuscript.

#Comments 16

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.

Reply 16:

Thanks for the suggestion. We will add a concrete example to support the statement. The example points to the wind speed centre over land near 33°N, 118°E, where the ensemble mean captures the location and intensity more accurately than the deterministic forecast. We will also add a short discussion of the low-wind area over the sea near 33°N, 121.5°E. This feature fades in the analysis and reference after +1h but persists in the ensemble mean until +6h, while the ensemble maximum evolves more consistently with the analysis. We interpret this as a sign of under-dispersion in that region: the ensemble underestimates the probability of wind speed increase. This is consistent with our earlier observation (Fig. S5) that the model tends to be under-dispersive for extreme events. We will revise the text accordingly i.e. "Over the sea near 33°N, 121.5°E, a low-wind area is visible in all outputs at +1h. It fades in the analysis and deterministic reference after +1 hour as wind speed increases, but the ensemble mean retains it until +6 hour. The ensemble maximum, however, evolves more consistently with the analysis. This suggests that the ensemble members are not equally capturing the wind speed increase in this

region, i.e., the ensemble is under-dispersive there. This behaviour is consistent with the under-dispersion noted for extreme events (Fig. S5).” The revision will be traceable in the manuscript.

#Comments 17

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.

Reply 17:

Thanks for the hint. The relatively short training period (~1.5 years) is indeed an additional factor contributing to the excessive smoothness for extreme events, as it likely contains too few samples of such events. We will revise the sentence in the conclusion to include this point. The sentence now reads: “This limitation may stem from the use of a basic diffusion model architecture as well as the relatively short training period (~1.5 years), which likely does not provide sufficient samples of extreme wind events.” The revision will be traceable in the manuscript.

#Comments 18

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).

Reply 18:

Thanks for the question. The deterministic SIVA nowcast takes about 2 minutes per forecast cycle, and our DDPM generates a 16-member ensemble in about 2 minutes on 2 NVIDIA 5090 GPUs. We have added computational efficiency information in the revised manuscript i.e.: “while maintaining computational efficiency: the deterministic reference nowcast takes about 2 minutes per forecast cycle, whereas the DDPM generates a 16-member ensemble in about 2 minutes on 2 NVIDIA 5090 GPUs.” The revision will be traceable in the manuscript.