RC1 Comments and responses:

We thank the Reviewer for her/his insightful comments. We appreciate the time and effort invested in providing detailed suggestions. Below, we address each comment in detail and outline the corresponding actions we have taken.

In light of the reviewers' comments, some major modifications to the structure of the manuscript have been introduced. Namely: (i) the introduction of a 'Discussion' section which now hosts material previously shown in two appendixes, (ii) the introduction of a new Appendix that hosts a wide part of the architectural design of the models, previously shown in the Methods section, and (iii) the addition of a case study section in the discussion section which shows an additional analysis performed for the reviews.

Specific Comments

Comment#1: The objective of applying LDM for dynamical downscaling is timely and relevant, but it should be made clearer to a general audience. Explicitly outline what key challenges the LDM addresses that previous models have struggled with. Highlight the main contributions more prominently in the abstract and methods sections. Emphasize the specific advancements over existing downscaling techniques, particularly the practical significance of LDM compared to UNET and GAN approaches.

Reply: We understand the feedback of the Reviewer and have revised the manuscript to address the Reviewer's concerns with the following specific actions:

Action:

- The abstract has been mainly re-written and now reads as follows:

"*Deep learning (DL)-based downscaling is a key application in Earth System Modeling, enabling the generation of high-resolution fields from coarse numerical simulations at reduced computational costs compared to traditional regional models. Additionally, generative DL models can potentially provide uncertainty quantification through ensemble-like scenario generation, a task prohibitive for conventional numerical approaches. In this study, we apply a Latent Diffusion Model (LDM) to demonstrate that recent advancements in generative modeling enable DL to deliver results comparable to those of numerical dynamical models, given the same input data, preserving the realism of fine-scale features and flow characteristics at reduced computational costs. We apply our LDM to downscale ERA5 data over Italy up to a resolution of 2 km. The high-resolution target data consists of 2-m temperature and 10-m horizontal wind components from a dynamical downscaling performed with COSMO_CLM. A selection of predictors from ERA5 is used as input and a residual approach against a reference UNET is leveraged in applying the LDM. The performance of the generative LDM is compared with reference baselines of increasing complexity: quadratic interpolation of ERA5, a UNET, and a Generative Adversarial Network (GAN) built on the same reference UNET. Results highlight the improvements introduced by the LDM architecture combined with the residual approach, outperforming all the baselines in terms of spatial error, frequency distributions, and power spectra. These findings point out the potential of LDMs as cost-effective, robust*

alternatives for downscaling applications (e.g. downscaling of climate projections) where computational resources are limited but high-resolution data is critical."

- The last part of the Introduction has been re-written and now reads as follows:

"In this study, we develop and evaluate a Latent Diffusion Model (LDM), which represents a novel approach for atmospheric downscaling tasks. This method offers two key advantages: first, the diffusion-based framework ensures significantly more stable training and more realistic generations compared to GAN models while retaining the capability to generate fine-scale features and enabling ensemble generation. These attributes have demonstrated superior performance in image processing applications compared to GANs (Saharia et al., 2023; Dhariwal and Nichol, 2021). Second, the latent-space approach improves upon pixel-space diffusion methods by substantially reducing computational costs for both training and inference (Rombach et al., 2021), making it especially suitable for scaling downscaling tasks to larger spatial domains and longer temporal scales. Lastly, the high-resolution output from a numerical dynamical downscaling simulation serves as our target-reference dataset allowing us to assess whether a well-trained LDM can emulate the accuracy of dynamical downscaling. If successful, this approach would provide a highly efficient alternative to traditional numerical methods by drastically reducing computational demands while maintaining accuracy, making it a promising tool for a wide range of critical downscaling applications."

- The central body of Section '3. Related work and Contribution' has been adjusted/re-written and now reads as follows:

"On the other hand, diffusion models have recently overtaken the GANs in the computer vision domain for super-resolution applications because they are easier and more stable to train and can produce more realistic samples (Moser et al., 2024; Saharia et al., 2023; Dhariwal and Nichol, 2021). Indeed, diffusion models explicitly model the probability distribution of the data through a diffusion process, ensuring that fine details are preserved while generating diverse outputs. On the contrary, the adversarial training of the GANs sometimes leads to artifacts or limited variability in results. In the Earth system domain, diffusion models introduce a relatively younger approach but have already been proven very effective in weather forecasting and nowcasting applications (e.g., Leinonen et al., 2023; Li et al., 2024). Diffusion models have yet to be widely tested and evaluated on the atmospheric downscaling task, but their characteristics and capabilities are undoubtedly promising for this application as shown for example in Addison et al. (2022) for precipitation, or in Mardani et al. (2023) for 2-m temperature, surface wind speed, and precipitation, or Merizzi et al. (2024) for wind speed."

Comment#2: The paper provides a detailed account of the architectures, but more information is needed on hyperparameter selection. Clarify how hyperparameters were optimized for each model to ensure a fair comparison. For instance, how is the timestep embedding in Figure 4 implemented? The residual approach is a strong point, but a deeper explanation of why the residual

method was adopted would strengthen the paper. What specific insights led to using a residual method, and how does it enhance the model's training stability?

Reply:

- **Hyperparameters:** We also received a similar comment from reviewer 2 and therefore understand that we need to assess this point more properly. We did not perform an automatic search for hyperparameters because we used the same hyperparameters chosen by the reference papers for each model. Specifically, for the GAN we referred to the network setting tuned by Esser et al. (2021) who already performed the search and optimization of the relevant hyperparameters of the patch-GAN we applied. Similarly, for the Latent Diffusion Model, we used hyperparameters derived from the LDCast implementation, proposed by Leinonen et al. (2023) but applied different architectural adjustments already described in the manuscript (e.g. the v-parameterization).

- **Figure 4:** We understand the reviewer's comment on the timestep embedding, as the former Figure 4 could be misleading. The LDM_res model does not embed any date/time information and does not take any extra information as input compared to the other models. The 'timestep embedding' in the former Figure 4 referred to the denoising timestep information which is intrinsic to the diffusion model.

- **Residual approach:** We agree with the Reviewer's comment. The residual approach is a strong point of the manuscript, and we, therefore, followed his suggestions to give it more relevance. The analysis of the residual approach's role is now presented in the main text, compared against the performance of the latent diffusion model applied without the residual approach. This follows exactly the tests that we performed developing the model: we first applied the LDM, obtaining unsatisfactory results, and then moved to the residual approach, finally achieving a much better performance. Nevertheless, the residual approach does not enhance the model's training stability: the training of the LDM model is stable and effective both with and without the residual approach. The residual has been introduced only to improve model performance.

Action:

- **Hyperparameters:** We improved the description of our training procedure for all the models following both the reviewers' suggestions. For conciseness, as many additional results were added in the main text considering the reviews, we reorganized the Methods section moving a wide part of the architectural design and training procedure to a new appendix. The new Appendix A "Architecture and training procedure of the DL models" reports on all the requested additional information regarding the training procedure/hyperparameters, as follows:

  - UNET: "*The loss function used for training is the Mean Squared Error (MSE) loss with mean reduction, which is suitable for regression-based tasks and ensures smooth convergence. The Adam optimizer (Kingma and Ba, 2015) is employed with a learning rate of 10−3 and no weight decay, chosen for its effectiveness in handling non-stationary objectives and sparse gradients. Training is conducted up to 100 epochs with early stopping enabled to terminate training if validation performance does not improve over 10 consecutive epochs. Each epoch involved iterating through the dataset with a batch size of 16, chosen to balance memory constraints and training efficiency. The network is constantly fed with the whole target domain and no patch-training is applied.*"

- GAN: "*The hyperparameters set for the training are derived from the search and optimization already performed by Esser et al. (2021), while the parameters we manually fine-tuned are described in the following. The pixel loss we used is the Mean Absolute Error (MAE), while the discriminator loss is the hinge loss. The discriminator is activated after 50000 training steps, giving the generator time to learn to generate consistent outputs and thus stabilizing the adversarial training (Esser et al., 2021). After activating the discriminator, the network is trained by updating alternatively the gradients of the generator and the discriminator. The network is constantly fed with the whole target domain and no patch-training is applied. The Adam optimizer (Kingma and Ba, 2015) is used for both the generator and the discriminator, with a base learning rate equal to 4.5 * 10−6 multiplied by the number of the GPU and the batch size used for the training (i.e. 4.5 * 10−6 x 1GPU x 4batch size) (Goyal et al., 2018), and beta parameters set to 0.5 and 0.9. Training is conducted for up to 100 epochs, with early stopping enabled to terminate training if validation performance does not improve over 10 consecutive epochs. Each epoch involved iterating through the dataset with a batch size of 4, chosen to balance memory constraints and training efficiency.*"

- VAE: "*The VAEs are trained on random 512x512 pixel patches of high-resolution target variables, with a 16 batch size. The training process for the VAEs leverages AdamW optimizer (Loshchilov and Hutter, 2019), with a base learning rate of 1 * 10−3, beta parameters set to 0.5 and 0.9, and a weight decay of 1 * 10−3. The loss function combines a reconstruction loss, computed as the Mean Absolute Error (MAE) between predicted and target outputs, with the KL divergence term. The KL divergence, scaled by a weight factor ($\lambda KL = 0.01$), enforces a standard normal distribution on the latent space. A ReduceLROnPlateau scheduler (PyTorch, 2023) is employed to dynamically adjust the learning rate by reducing it by a factor of 0.25 if the validation reconstruction loss does not improve for three consecutive epochs. Training is conducted for up to 100 epochs, with early stopping enabled to terminate training if validation performance does not improve over 10 consecutive epochs*"

- DENOISER/CONDITIONER: "*As shown in Figure A3, the conditioner and the denoiser are trained together, minimizing the Mean Square Error (MSE), and feeding the network with random patches of ERA5 predictors (64x64 pixels) and static data (512x512 pixels) for the conditioning, and high-resolution target variables (512x512 pixels) for the ground truth. The batch size is set to 4 and 8 for the 2-m temperature and 10-m wind components models respectively, tuned to balance memory constraints and training efficiency. The training is performed using the AdamW optimizer (Loshchilov and Hutter, 2019), with a base learning rate of 1 * 10−4, beta parameters set to 0.5 and 0.9, and a weight decay of 1 * 10−3. A ReduceLROnPlateau scheduler (PyTorch, 2023) is employed to dynamically adjust the learning rate by reducing it by a factor of 0.25 if the validation loss does not improve for three consecutive epochs. Additionally, Exponential Moving Averaging (EMA) is applied to the network weights, following Rombach et al. (2021). Training is conducted for up to 100 epochs, with early stopping enabled to terminate training if validation performance does not improve over 10 consecutive epochs.*

The other hyperparameters set for the training are derived from the implementation of the LDCast model by Leinonen et al. (2023)."

- **Figure 4**: We updated former Figure 4 specifying the wording 'denoising timestep embedding' instead of 'timestep embedding'. The figure is now Figure A2, as it has now been moved to the new Appendix A "Architecture and training procedure of the DL models".

- **Residual approach:** In order to give more relevance and insights into the role of the residual approach, we improved the structure and contents of the manuscript as follows:
  - We created a new section: "6 Discussion"
  - We moved and enriched the analysis presented in the former "Appendix D: LDM residual versus non-residual results" to the new subsection "6.2 On the contribution of the residual approach" under the new section "6 Discussion", providing a direct comparison of results from the LDM with and without the residual approach. We also added quantitative information on the results in terms of pdf and spectra as suggested by the second reviewer. The new section now reads as follows:

    "In this section, we compare the performance of the LDM trained with and without the residual approach, highlighting the significant improvements introduced by the residual methodology. The comparison focuses on the frequency distribution and the Radially Averaged Power Spectral Density (RAPSD), as shown in Figures 10 and 9, respectively, along with their associated metrics, IQD and RASPL. The analysis reveals notable differences between the two models, particularly in: (i) accurately estimating the most frequent values of 2-m temperature, (ii) reconstructing the full frequency distribution of wind speed, (iii) reconstructing the 2-m temperature power spectra at small scales, where the non-residual LDM underperforms compared to the quadratic interpolation of ERA5, and (iv) reconstructing the 10-m wind speed power spectra across all scales, with the non-residual LDM exhibiting a quasi-constant lag across all wavelengths. The corresponding VAEs for the two models (VAE and VAE_res) show comparable performance except at the smallest scales of the 2-m temperature power spectra (not shown). Consequently, the diminished performance of the non-residual LDM can be attributed to the VAE only in this specific case (iii). All other deficiencies are solely due to the diffusion process. Training the diffusion model to reconstruct a residual field instead of the original target field significantly enhances performance, improving the reconstruction of frequency distributions and power spectra across all wavelengths, particularly for chaotic variables such as wind speed."
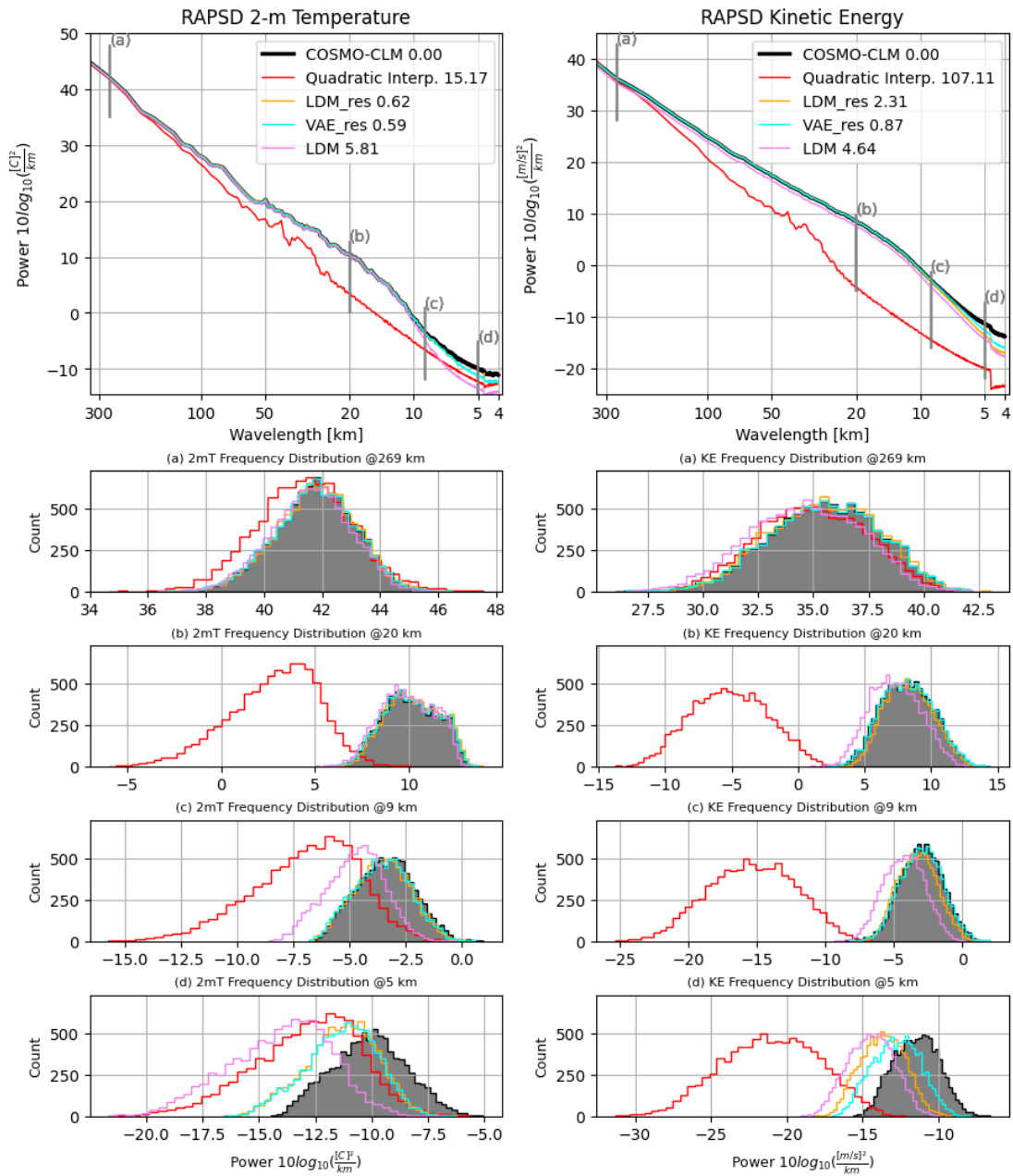
Figure 9. Comparison of Radially Averaged Power Spectral Density (RAPSD) distributions for results from LDM_res, VAE_res, and LDM against COSMO-CLM reference-truth and ERA5 quadratic interpolation. The left column refers to the 2-m temperature, and the right column refers to the 10-m horizontal wind speed. The first row shows the averaged-in-time spectra, across the whole test dataset. Notice that in the first row y-axes are logarithmic to highlight the tail of the distributions, hence the high frequencies. The bottom rows show the distributions of single-time RAPSD values for fixed wavelengths, namely 269, 20, 9, and 5 km
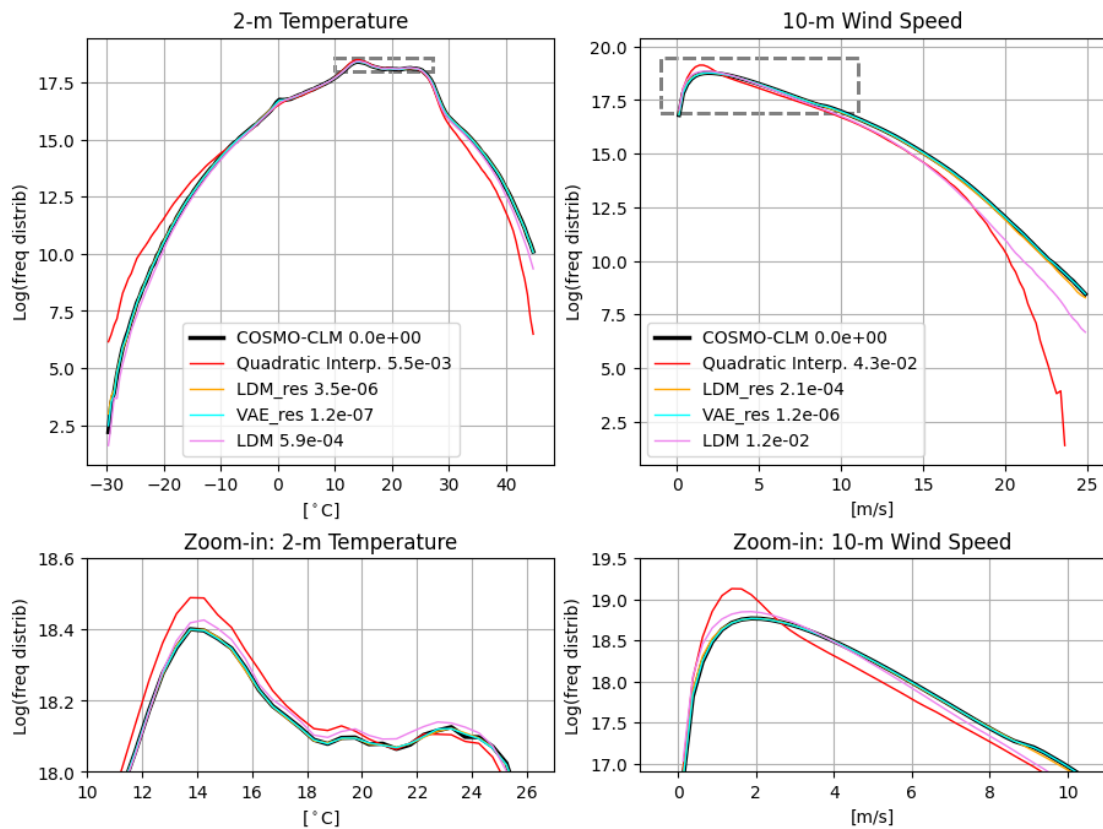
Figure 10. Comparison of frequency distributions for results from LDM_res, VAE_res, and LDM against COSMO-CLM reference-truth and ERA5 quadratic interpolation. The left column refers to the 2-m temperature, and the right column refers to the 10-m wind speed. Counting of pixel-wise data is cumulated for the yearly test dataset over bins of 0.5 °C and 0.05 m/s for temperature and wind speed, respectively. Notice that y-axes are logarithmic to highlight the tails of the distributions, hence the extreme values. The top row focuses on the tails of the distributions, i.e. on extreme values, while the bottom row focuses on the most frequent values and is a zoom-in on the dashed boxes for each variable

- o To compensate for the increased length of the main text of the manuscript we reduced the Methods section, moving a wide part of the architectural design and training procedure to the new Appendix A "Architecture and training procedure of the DL models".

Comment#3: The comparisons with UNET, GAN, and interpolation are useful, but the study lacks a broader context regarding recent advancements. It would improve the paper if comparisons with other state-of-the-art downscaling methods were included, such as recent transformer-based approaches or hybrid models that combine CNNs with probabilistic methods.

Reply: We understand the concerns of the Reviewer and agree that a wide comparison with other state-of-the-art DL downscaling approaches would be beneficial. However, we find it difficult to incorporate such comparisons for several reasons.

- **Resource Limitations:** Conducting a fair comparison would require retraining other models using our specific data, which would involve considerable time and resources that are currently beyond the scope of this study.
- **Data Preprocessing:** Even running other pre-trained models, without re-training them, would necessitate significant additional data preprocessing, which would require further resources and time, as alternative models often rely on different sets of predictors than those we use in this work.

- **Lack of a Recognized Standard:** while we would be very interested in comparing our results against a well-established "golden standard" deep learning model, at present, there is no universally accepted reference model for the downscaling tasks. The choice of model would therefore be somewhat arbitrary, which makes meaningful comparisons challenging.

In summary, while we acknowledge the importance of comparing with other DL models, the constraints of time, resources, and the scope of our work prevent us from undertaking such comparisons at this stage. Additionally, we believe that the presented comparison against the reference UNET and GAN already brings credibility to the manuscript and the presented results, showing different performances against well-known networks. We hope that this explanation clarifies our position. Of course, a downscaling benchmark would be highly valuable in this field. We are aware that there is ongoing work by Langguth et al. (2024), who is developing a "Benchmark Dataset for Statistical Downscaling of Meteorological Fields with Deep Neural Networks". This would allow a direct and fair comparison of different DL downscaling approaches over an out-of-domain testbed. Nevertheless, as far as we know, this project is still an ongoing work and not yet available for direct comparison. We would be very happy to join a model-score board once available.

Action: We included citation on transformer-based architecture in our Introduction to give a wider overview of recent DL applications:

"A*dditionally, the use of transformer-based architectures for downscaling is an emerging approach but remains relatively new within the Earth system science domain (Zhong et al., 2024)."*

Comment#4: The study used 21 years of data, but the explanation regarding "70% for training, 15% for validation, and 5% for testing (corresponding to 15, 3, and 1 year, respectively)" is unclear. Is the one year of test data randomly selected from a specific year, or does it include a few days from each season across multiple years? If only data from a particular year was used, it may not account for interannual variations. Additionally, have factors such as La Niña effects been considered in analyzing the temperature and wind fields?

Reply: We appreciate the reviewer's feedback and agree that our explanation of the dataset splitting lacked sufficient detail. The complete dataset comprises 21 years of hourly data. Specifically, the training dataset includes 128,873 samples (~15 years), the validation dataset consists of 27,616 samples (~3 years), and the test dataset contains 8,760 samples (~1 year). These samples were randomly selected from the entire 21-year dataset.

The figure below illustrates the distribution of samples across various time-based aggregations (i.e., years, months, and hours of the day) for the training, validation, and test datasets. As depicted, the random selection ensured a uniform distribution of samples across all temporal aggregations in each dataset.

As for the La Niña effects, we did not perform a sensitivity test of the model's performance for different phases of the Nino/Nina indexes. The evaluation of the DL models is performed on the whole test dataset, which presents a distribution of El Nino/La Nina index values consistent with the distribution of the training and validation datasets (last row of the following Figure). To produce these plots, we used Nino 4 monthly Anomalies released by the Climate Prediction Center of the US National Weather Service (calculated using ersstv5, https://psl.noaa.gov/data/climateindices/list/).

Action: Section 3.5 'Dataset splitting strategy' has been updated and now reads:

"*The experimental database consists of hourly data spanning from 2000 to 2020, totaling approximately 184,000 hourly samples, for both low and high-resolution data. The dataset was randomly divided into three subsets: 70% for training (~15 years, 128,873 samples), 15% for validation (~3 years, 27,616 samples), and 5% for testing (~1 year, 8,760 samples). This random splitting ensures a uniform distribution of samples across years, months, and hours of the day in all three datasets. The testing dataset was limited to 1 year (5% of the total dataset) to address time constraints associated with running all models during evaluation, particularly the diffusion model.*"

Comment#5: The evaluation is thorough, but it could be strengthened by including the following: A case study of an extreme weather event to showcase the model's performance in challenging scenarios. A discussion of the model's performance across different seasons within the study area to illustrate robustness in various conditions.

Reply: We agree with the Reviewer that evaluating the models' performance in a case study of an extreme event and across different seasons would enhance the manuscript. In response, we conducted both these analyses and reported the results in the manuscript.

Action:

- EXTREME EVENT: We performed an analysis of the reconstruction of an extreme event and added a new subsection "6.3 On the reconstruction of extreme events: a case study of strong winds" under the new section "6 Discussion", which reads as follows:

  "*To evaluate the performance of the deep learning models in a challenging scenario, we selected February 7, 2022, as a case study. This analysis represents a preliminary*

investigation of the deep learning models' ability to reconstruct a single strong wind event and their performance in reproducing time series when applied to consecutive timesteps. The selected event provides independent data, separate from the training, validation, and test datasets previously discussed. February 7, 2022, is particularly noteworthy due to widespread strong winds across Italy, prompting weather alerts in various regions and causing significant wind-related damage. On this day, the Italian peninsula was affected by a pronounced pressure gradient, resulting from the southward descent of a low trough from Northeastern Europe toward the Ionian Sea and the simultaneous presence of a high-pressure system centered over the Bay of Biscay (see Figure 11). This synoptic configuration generated widespread föhn conditions, with strong northerly to northwesterly winds affecting Northern regions and areas downwind of the Apennine ridges.
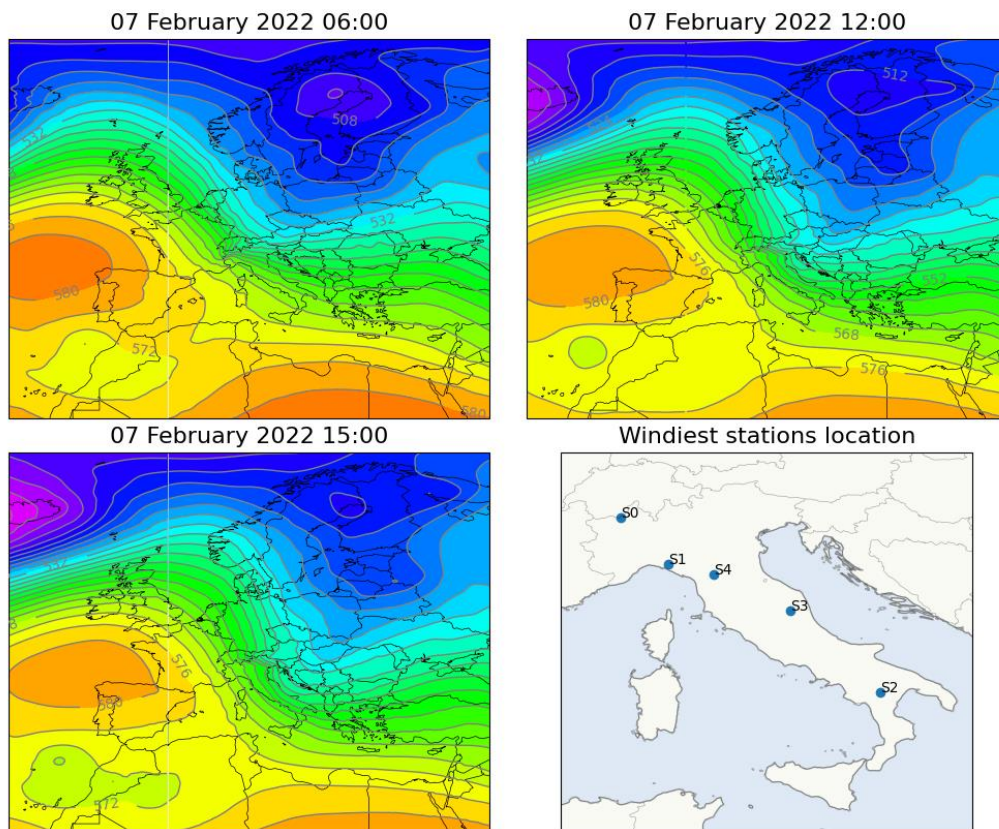
Figure 11. First three panels show the evolution of geopotential height at 500 hPa [dam] from 06 UTC on Feb. 7, 2022, to 15 UTC on Feb. 7, 2022, (ERA5 data). Last panel shows the location of 5 weather stations used for the analysis.

The performance of the models was assessed at five locations of interest, corresponding to Italian weather stations that recorded hourly wind speeds exceeding 20 m/s during the case study. These stations, situated in complex terrain, are highlighted in the final panel of Figure 11. Figure 12 presents the time series of 10-m wind components and wind speeds at each target location. For 10-m wind speed, observational data collected by the weather stations are included as a reference for comparison. As illustrated in Figure 13, significant differences are observed between ERA5 and COSMO-CLM data across all reference stations. The dynamical downscaling approach of COSMO-CLM produces substantially higher wind speeds compared to the low-resolution ERA5 data, with discrepancies reaching up to 10–13 m/s. While COSMO-CLM still underestimates wind speeds compared

to observations, its temporal evolution of wind flux generally aligns well with measurements. The deep learning models demonstrate remarkable performance, with the target ground truth being the COSMO-CLM output. All three models effectively reconstruct wind intensities for both components, showing minimal dependency on the underestimation present in the input low-resolution data. Notably, the models accurately capture the temporal evolution of wind components, frequently correcting the phase discrepancies in wind speed trends (increases or decreases) present in the low-resolution data. This capability is particularly noteworthy given that the models are trained exclusively for image-to-image downscaling without access to temporal information from adjacent timesteps. Among the deep learning models, results are generally comparable. However, the GAN and UNET models tend to produce smoother temporal signals compared to the LDM_res model and the COSMO-CLM baseline."
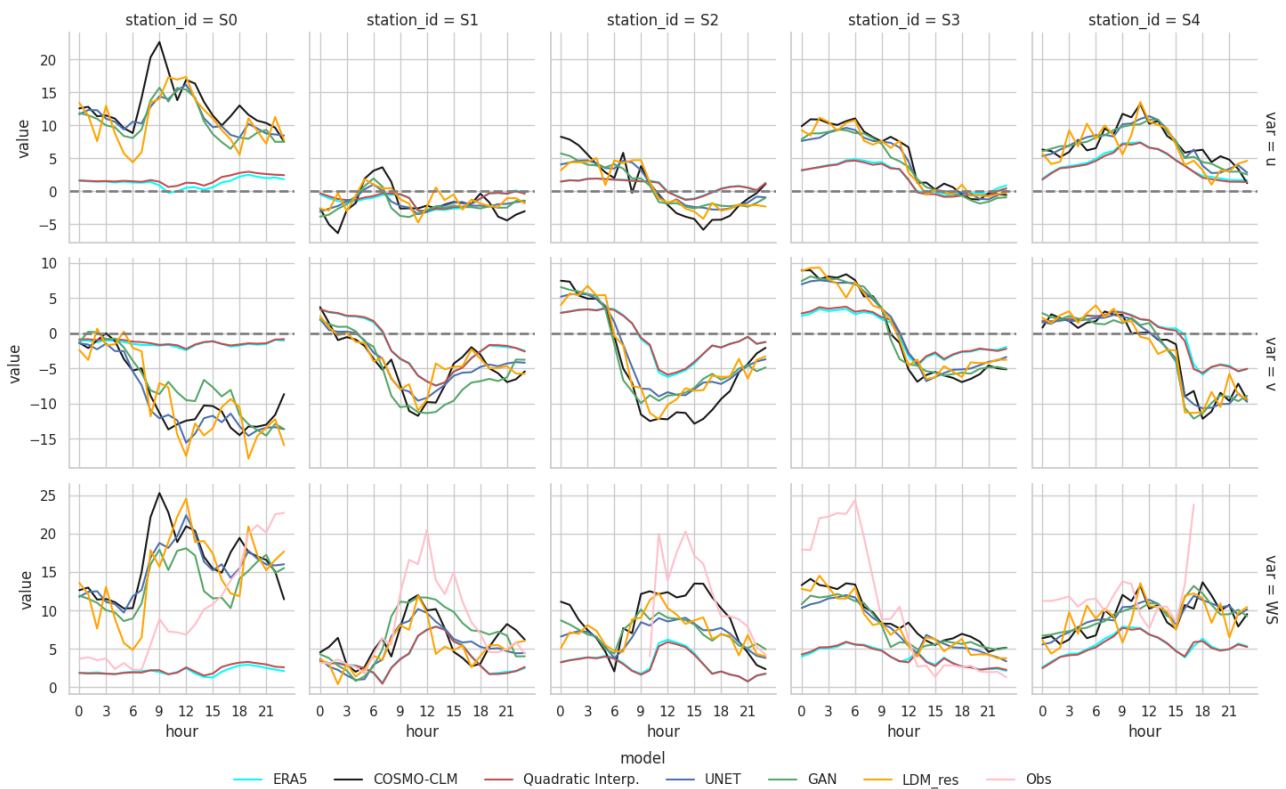
Figure 12. Hourly evolution of 10-m wind components and 10-m wind speed as reconstructed by all models on February 7, 2022, in the selected target locations. Available 10-m wind speed observations from each weather station are also reported.

- SEASONAL EVALUATION: To investigate the models' performance across different seasons we calculated all the metrics already presented in the manuscript for each season, splitting the test dataset accordingly. Each season subset is composed of ~2200 hourly samples. In addition to RMSE, BIAS, R2, and PCC we also calculated two additional metrics, leveraging the suggestions from Reviewer2- comments #5 and #6-, to quantify results in terms of frequency distribution and power spectra: the Integrated Quadratic Distance, and the Radially Averaged Log Spectral Distance (RALSD)-score. In the manuscript, we added

Figure 5 summarizing results for each metric, for each model (see below), and we discussed these results in each related section.

Text added to section 5.2 Verification deterministic metrics:

"*A more detailed evaluation of the models' performance using these metrics is provided in Figure 5, where the test dataset is divided into meteorological seasons for seasonal analysis. The figure demonstrates that the results remain consistent across seasons, reinforcing the evaluations previously discussed. Notably, the summer season exhibits slightly lower metric values across all models. Additionally, deep learning models display more stable performance across different seasons compared to quadratic interpolation.*"

Text added to section 5.4 Frequency distributions:

"*These qualitative evaluations can be quantified by calculating a divergence score on the underlying empirical Cumulative Distribution function (CDF) of the data such as the Integrated Quadratic Distance (IQD), as proposed by Thorarinsdottir et al. (2013) (see Appendix B for details). Values of IQD score are indicated in Figure 7. Consistently with the associated frequency distributions, values of IQD scores are lower for 2-m temperature compared to wind speed across all models. Notably, the LDM_res model achieves the best performance, with IQD scores two orders of magnitude lower than the UNET and the GAN for 2-m temperature, and one order of magnitude lower than the GAN for wind speed, highlighting its superior accuracy. IQD scores were also computed for each season within the test dataset, with the results presented in Figure 5. The models' scores align with the yearly analysis across all seasons. In particular, LDM_res exhibits the smallest score variations across seasons, indicating minimal sensitivity to seasonal changes.*"

Text added to section 5.5 Radially Averaged Power Spectral Density (RAPSD):

"*The RALSD scores were also computed for each season within the test dataset, with the results presented in Figure 5. As for IQDs, the models' RALSD scores align with the yearly analysis across all seasons. In particular, LDM_res exhibits the smallest score variations across seasons (together with the GAN for wind speed), indicating minimal performance sensitivity to seasonal changes.*"
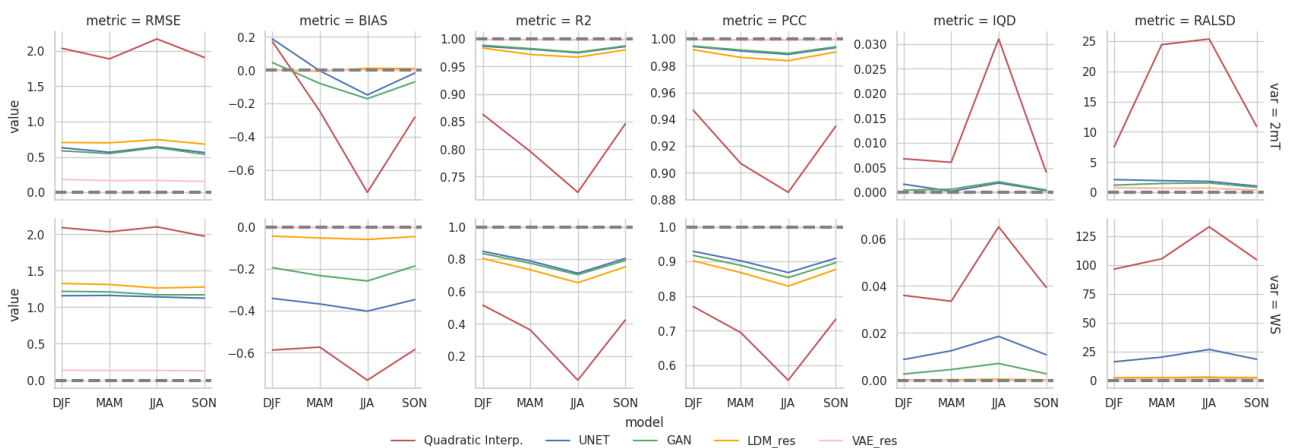
Figure 5. Comparison of different metrics/scores across seasons for the analyzed models (top row refers to the 2-m

temperature and bottom row refers to the 10-m wind speed). Notice that y axes are not shared between panels. The dashed line highlights the reference value for each metric

Minor Comments

Comment#1: In Figure 6, the second column zooms in on Sardinia Island. Consider resizing this image to match the first column, as this would make the figure more aesthetically pleasing.

Reply: We fully agree with the Reviewer's comment.

Action: We recreated former Figure 6, Figure A1, and Figure A2 to make them more aesthetically pleasing and substituted them in the manuscript. After the reviews, the numbers of these images are now Figure 3, Figure B1, and Figure B2, respectively.

Comment#2: The future work section is informative. It would be beneficial to add some thoughts on how this approach could be integrated into existing modeling frameworks or operational systems, which could help bridge the gap between research and practical applications.

Reply: We agree with the reviewer and updated the Future Work section based on this comment and comment#8 from the other reviewer.

Action: We re-wrote the Future work section, which now reads as follows:

"The results presented in this work suggest several promising directions for further investigation into the application of latent diffusion models for downscaling. Addressing the primary limitations of our DL approach, namely (i) inter-variable consistency and (ii) temporal consistency of the generated fields, is a key priority. Applying LDM_res with a multi-variable approach requires no architectural adjustments and could yield valuable insights, such as whether additional variables necessitate larger network architectures to optimize performance. The potential and efficiency of a multi-variable approach have already been demonstrated in pixel-space diffusion downscaling by Mardani et al. (2023). Further evaluation of the temporal consistency of downscaled data in this version of LDM_res is also relevant. Enhancements in this area could involve conditioning the diffusion process on a (short-) temporal sequence of low-resolution fields or incorporating previous high-resolution outputs in an auto-regressive approach. Additionally, the generative capabilities of LDM_res need to be explored, to assess the potential added value of a DL-generated ensemble.

Future developments could explore the integration of latent diffusion models into existing modeling frameworks and operational systems, such as using them as post-processing tools for real-time weather forecasts, seasonal forecasts, and climate projections. These integrations would require tailored training procedures, alignment with operational inputs and reference data, and rigorous validation to ensure robustness and compatibility with practical applications. For example, ongoing research, funded as an Innovation project on this research's funding, is already investigating the effectiveness of LDM_res in downscaling precipitation and temperature, in a multi-variable approach, from climate projections predictors. This exploration is expected to further elucidate the versatility and robustness of the proposed approach and showcase its practical applications."

Comment#3: Wind speed downscaling is evidently more challenging than temperature, and while some explanations are given, providing references would lend more credibility to the discussion. Additionally, consider exploring whether other variables, such as boundary layer height or atmospheric vorticity, could be used to improve downscaling performance for wind speed.

Reply:
- **References**: We agree with the reviewer that the manuscript lacked literature references on this specific point.
- **Additional predictors:** At the beginning of our study, we considered a wider set of input predictors, including, for example, the BLH. We ended up with a smaller set of predictors on the basis of two considerations:
  o Literature review for the selection of the parameters (already cited in the manuscript). For example, regarding specifically the BLH, the Analysis of feature importance presented by Höhlein et al. (2020) surprisingly showed that "Perturbations in BLH also have only a slight impact on prediction performance [of 100-m wind speed components]" for a wide range of tested DL models.
  o The model we developed and tested on matching reanalysis is meant to be further tested and applied to different applications, such as the downscaling of numerical seasonal predictions and climate projections. The selection of predictors has therefore been guided by the availability of the variables in a wide range of these alternative input sources.

Action:
- **References**: We added a paragraph to the "Related work and contribution" section, which reads as follows:
  "*Downscaling temperature and wind poses distinct challenges due to their inherent differences as meteorological variables (De et al., 2023; Höhlein et al., 2020). The 2-m temperature is generally easier to predict, being a scalar variable and predominantly aligning with well-established patterns, such as dependence on terrain elevation and diurnal cycles. In contrast, wind is a vector field, comprising both magnitude and direction, is influenced by small-scale processes (such as turbulence and localized interactions) and therefore exhibits greater variability, and strong scale dependency, especially over complex terrain (Serafin et al., 2018; Rotach and Zardi, 2007). These characteristics make wind considerably more challenging to downscale, regardless of the downscaling methodology applied, as widely acknowledged for example by Pryor and Hahmann (2019). The difference in downscaling the two variables is also clear in the already proposed deep learning-based approaches, tackling the downscaling of both these variables (Mardani et al., 2023).*"