RC2 Comments and responses:

We thank the Reviewer for 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.

Comment#1: The study's title and the introduction question whether deep neural networks can efficiently emulate dynamical downscaling. However, the authors employ an image-to-image approach with separate models for the 2m temperature and the 10m-wind. This approach challenges temporal and inter-variable consistency - an aspect inherently maintained with dynamical downscaling using on physics-based numerical models. The authors should address this aspect in the Introduction and discuss future prospects in their Conclusion to align with the scope of the paper. Referring to the recent study by Mardani et al. (2024) [2], which presents promising results for multi-variate downscaling, would strengthen the discussion.

Reply: We agree with the Reviewer that these two specific aspects needed to be better underlined in our manuscript, especially considering the manuscript's title opening question.

Action: As suggested, we added considerations on these two aspects in various parts of the manuscript:

- The Conclusions section now ends with the following paragraph:
  "Nevertheless*, two primary limitations remain in the proposed deep learning approach: (i) inter-variable consistency and (ii) temporal consistency of the generated fields. These fundamental aspects, which are inherently preserved in dynamical downscaling performed using physically-based numerical models, are not guaranteed by the design of this study. Specifically, LDM_res is applied and trained separately for 2-meter temperature and 10-meter horizontal wind components, performing an image-to-image downscaling. Regarding inter-variable consistency, results from this study, as well as findings from other works, such as Mardani et al. (2023), suggest that a multi-variable approach is feasible and could provide significant benefits by better leveraging all available parameters within the network. Concerning temporal consistency, although results briefly showcased in the presented case study indicate that the DL model generates consecutive timesteps that form rather consistent timeseries, further investigation and testing are necessary to thoroughly address this aspect, as discussed in the future work section.*"
- The following paragraphs has been added to the "Related work and contribution" section (answering both this comment and to the other reviewer's minor comment#3):
  "*Specifically, we focus on generating 2-m temperature and 10-m horizontal wind components high-resolution fields. The downscaling of 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).

In light of this evidence, we designed our study to train separate models for 2-meter temperature and 10-meter horizontal wind components, unlike other approaches found in the literature Mardani et al. (2023). This design choice facilitates the interpretation of the model outputs, enabling a clearer understanding of the strengths and limitations of the tested models when applied to individual target variables. However, this approach also imposes a limitation compared to dynamical downscaling, as it introduces uncertainties regarding inter-variable consistency."

- The Future work section has been updated in response to this comment and to comment#8 (please see the updated text under comment#8 response).
- In response to the other reviewer's comment#5 we added a new analysis in the new section "6.3 On the reconstruction of extreme events: a case study of strong winds", where we show, preliminarily, reconstructions of time series from image-to-image downscaling.

Comment#2: In Section 2, the authors mention that the LDM can be applied to larger domains during inference compared to the training step. While this flexibility is appealing, spatial dimensionality constraints on the input data are common for U-Nets due to downsampling operations. For example, three average pooling operations with a factor of 2 would require that input dimensions of the U-Net must be multiplier of 8. Do these constraints also apply to the Denoiser network in the LDM? Additionally, Figure 5 indicates that the input data comprises 512x512 grid points during the inference. Isn't the model evaluated on the larger target domain with 576x672 grid points as mentioned in the text? A brief assessment of any performance degradation, such as in RMSE, when applied to the full target domain would be insightful.

Reply: The denoiser network in LDM does have spatial dimensionality constraints similar to UNets, requiring input dimensions to be multiples of 8 due to the three levels of downsampling. However, the latent diffusion architecture is designed to handle arbitrary input sizes during inference through its patch-based approach. During training, we use 512x512 patches at high resolution (corresponding to 64x64 patches in latent space), but during inference, the model can process the full 576x672 domain: this is possible because the convolutional nature of both the VAE and denoiser allows them to operate on inputs of varying sizes (all transformations are local). The Reviewer is indeed correct, Figure 5 contained erroneous tensor dimensions in the Inference box. We thank the Reviewer very much for spotting this.

Action: Figure 5 has been updated accordingly, with the corrected tensor sizes for inference, corresponding to: 18x672x576 for the input static data, 14x84x72 for the input predictors, 256x84x72 for the embedded conditioning (latent space), (v*32)x84x72 for the noisy target variable (latent space), and vx672x576 for the output target variable. The figure has now been moved to the new Appendix A “Architecture and training procedure of the DL models”, as suggested and discussed in other comments. To clarify and address the reviewer’s concern, indeed all the evaluation results in the paper are computed on the full the full 576x672 domain.

Comment#3: The spatial resolution of the ERA5 input data is increased to 16 km for data alignment. How does this impact the input data quality (e.g. potential aliasing effects)? Can this affect or degrade the downscaling process? Please provide a brief explanation on the necessity for this adjustment.

Reply: The reprojection and interpolation of ERA5 data to the 16km EEA reference grid represents a careful balance of several considerations. The EEA reference grid provides a standardized coordinate system that aligns perfectly with our target domain and high-resolution COSMO-CLM data, which is essential for training deep learning models as it ensures precise spatial correspondence between input and target data. In choosing the interpolation approach, we prioritized preserving as much of the original ERA5 information as possible. The 16km resolution was selected because it's the closest possible grid spacing to ERA5's native resolution (~22km at the domain's latitude) that also aligns with the EEA reference grid system and provides a clean 8x downscaling factor to the 2km target resolution. We acknowledge that nearest-neighbor interpolation may introduce some aliasing artifacts. However, we chose this method over more sophisticated interpolation approaches because it offers a key advantage: it preserves the original values without introducing artificial intermediate values. This is particularly important when dealing with multiple physical quantities that may have different underlying statistical properties and physical constraints. While other interpolation methods might produce smoother fields, they could potentially create artificial values that violate physical relationships between variables, or, even worse, new maximum/minimum values with the risk of violating physical constraints. Since the interpolation distance is relatively small (from ~22km to 16km), we found that the benefits of having aligned grids and preserved original values outweigh the potential drawbacks of any aliasing effects. The subsequent downscaling models are trained on this slightly adjusted data, effectively learning to account for any systematic effects introduced by this preprocessing step. This standardization step ultimately enhances the downscaling process by ensuring consistent spatial relationships between input and target data while maintaining the essential characteristics of the original ERA5 fields.

Action: We added the following pararaph to section “3.2 Data alignment and preprocessing”:

“T*he choice of grid resolution and interpolation method reflects careful consideration of several factors. The 16-km resolution was selected as the closest resolution to ERA5’s native grid (~22 km at domain latitude) that maintains alignment with the EEA reference grid system while preserving the original large-scale information. While nearest-neighbor interpolation may introduce some aliasing artifacts, this method was chosen over more sophisticated approaches because it preserves the original values without creating artificial intermediates. The downscaling models can effectively learn to account for any systematic effects introduced by this preprocessing step.*”

Comment#4: Section 4 describes the tested deep neural network architectures. However, the authors should provide more details on the training process, i.e. learning rate schedule, configuration of the optimizer, number of training epochs and other relevant hyperparameters. The details could be included in the Appendix and would help to ensure reproducibility of the findings.

Reply: We agree with the Reviewer’s comment, deeper insights into the training procedure we followed can be provided.

Action: We added to the manuscript all the missing details regarding the training procedure of all the DL models, including the configuration of the optimizers, learning rate schedulers, batch sizes, patch training information, losses, and training epochs. For conciseness, as also suggested by the Reviewer (in this comment and comment #7), 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: "T*he 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 (λKL = 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: "A*s 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).*"

Comment#5: Section 5.4 provides valuable insight into the reconstruction of the frequency distribution of the target data. However, the results are rather qualitative. I suggest to quantify the difference, in particular with the Integrated Quadratic Distance, a proper divergence score on the underlying cumulative Distribution function (CDF) of the data, see Thorarinsdottir et al. (2013) [3].

Reply: We fully agree with the Reviewer’s comment, a quantitative evaluation of the results in terms of Integrated Quadratic Distance adds robustness to the obtained results.

Action: We calculated the Integrated Quadratic Distance for each model, for both 2-m temperature and wind speed, and added these results in both former Figure 9/now Figure 7 and former Figure D1/now Figure 10. We also added the following in the text, at the end of section 5.4 Frequency distributions, to comment on these results:

“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.”

Additionally, we added details on the computation of this score in the Appendix B:

“In Section 5.4 we discuss the Integrated Quadratic Distance score, which is calculated following Thorarinsdottir et al. (2013) and measures the similarity between two distributions by the integral over the squared difference between the two distribution functions: SEE FORMULA

where: F (t) and G(t) are the empirical Cumulative Distribution Function of the modeled and reference-truth data, respectively, computed following the implementation outlined in Pulkkinen et al. (2019); min and max are [-30,45]°C and [0,25] m/s, for 2-m temperature and wind speed respectively; dt has been fixed to 0.5 °C and 0.25 m/s, for 2-m temperature and wind speed respectively.”

Comment#6: Similarly, the Power Spectra Analysis can be further quantified by calculating the Radially Averaged Log Spectral Distance (RALSD)-score, see Eq. 8 in Harris et al. (2022) [4].

Reply: We fully agree with the Reviewer's comment, a quantitative evaluation of the results in terms of Radially Averaged Log Spectral Distance (RALSD)-score adds robustness to the obtained results.

Action: We calculated the Radially Averaged Log Spectral Distance (RALSD)-score for each model, for both 2-m temperature and wind speed, and added these results in both former Figure 10/now Figure 8 and former Figure D2/now Figure 9. We also added the following in the text, in section 5.5 Radially Averaged Power Spectral Density (RAPSD), to comment on these results:

> *"[...]. To provide a quantitative evaluation of the results in terms of power spectra, we calculated the log spectral distance of the RASPDs, referred to as the Radially Averaged Log Spectral Distance (RALSD)-score, as proposed in Harris et al. (2022) (see Appendix B). For each model, values of RALSD score are indicated in Figure 8."*
>
> *"[2mT ...] Overall, all DL models effectively reconstruct the 2-m temperature power spectra down to wavelengths of 10 km. However, LDM_res consistently outperforms both the UNET and the GAN, as evident from panels (b) and (c) of Figure 8 and the RALSD scores. The UNET and the GAN yield similar results, with marginal yet consistent enhancements originating from the adversarial training of the UNET, as also proved by the similar values of RALSD score."*
>
> *"[WS ...] The RALSD scores for LDM_res and the GAN are comparable, with the GAN exhibiting a slightly lower value. This marginal improvement in the GAN is primarily attributed to enhanced performance at the smallest scales (below 4.2 km), where the model is more prone to generating artifacts."*

Additionally, we added details on the computation of this score in the Appendix B:

> *"In Section 5.5 we discuss results in terms of the log spectral distance of the RASPDs, referred to as the Radially Averaged Log Spectral Distance (RALSD)-score, following Harris et al. (2022):*
> *RALSD = SEE FORMULA*
> *where: Ptrue and Pgen are the Radially Averaged Power Spectral Densities of the reference-truth and modeled data, respectively, computed following the implementation outlined in Pulkkinen et al. (2019); and N is the number of frequencies (i.e., 336)."*

Comment#7: The effect of the LDM approach on the fine-scale variability via the Variational Autoencoder to compress and decompress the high-resolved data is very interesting and relevant for the chosen framework. Also consider to refer to the effective resolution of numerical model data, see e.g. Skamarock et al. (2014) [5]. It is suggested to add this analysis to the main text. To compensate for the increased length of the text, parts of the model architecture description may be moved to the Appendix (together with the details on the training and hyperparameters mentioned above)

Reply: We agree with the Reviewer's comment and followed his suggestions to give more relevance to the analysis of the contribution of the VAE in our model.

Action: As suggested, we rearranged the structure and contents of the manuscript as follows:

- We created a new section: "6 Discussion"
- We moved the analysis presented in the former "Appendix C: On the contribution of the VAE" to the new subsection "6.1 On the contribution of the VAE" under the new section "6 Discussion", including a comment on the effective resolution, as suggested.
- 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 a new appendix (as suggested also in comment#4)

Comment#8: The Section on Future Work is very short and more details on the various mentioned prospects mentioned (temporal and inter-variable consistency, precipitation downscaling, ensemble generation) should be presented.

Reply: We agree with the reviewer and updated the Future Work section based on this comment and minor comment#2 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."