Response to Referee 1

We sincerely appreciate the referee for the thorough review and valuable suggestions on our manuscript. These comments are highly constructive and helpful for improving the quality and clarity of our work. We have carefully addressed all the major and minor comments point-by-point, and revised the manuscript accordingly. The detailed responses are as follows:

Major comment 1: To ensure a fair comparison with the dynamically downscaled results from CMA-MESO, the authors should provide more details about the simulation conditions used for CMA-MESO. For example, was CMA-MESO nudged toward the global reanalysis fields through data assimilation?

Response: CMA-MESO is not nudged toward global reanalysis fields during simulation. Instead, it is driven by CMA-GFS, which provides its initial and lateral boundary conditions. The primary data assimilation scheme employed in CMA-MESO is a three-dimensional variational system, which assimilates high-resolution in-situ and remote sensing observations. These details have been added to the revised manuscript to clarify the experimental setup and ensure transparency in the comparison.

Major comment 2: Using aggregated metrics such as mean absolute error (MAE) and continuous ranked probability score (CRPS) may be insufficient for assessing generative models such as diffusion models. To provide a more complete assessment, additional diagnostics would be helpful, such as spatial spectra or structural similarity metrics. Without these diagnostics, it is difficult to verify claims such as "…our data-driven models can potentially outperform CMA-MESO for most variables."

Response: We sincerely appreciate the referee's valuable suggestion regarding comprehensive evaluation of the diffusion-based generative model. We agree that conventional aggregated metrics including MAE and CRPS are insufficient to fully characterize the performance of downscaling generative models. To strengthen the assessment, we have additionally calculated power spectral density and structural similarity index measure and added corresponding analyses and figures in the revised manuscript.

Figure 1 presents the power spectral density of downscaled meteorological fields. To improve legibility, a zoomed-in view focusing on the zonal wavenumber range [0.005, 0.05] is provided in Fig. 2. First, the CorrDiff model exhibits higher power spectral density than the regression model, which aligns with the inherent ability of diffusion models to generate richer high-frequency details. For most variables, at a zonal wavenumber of approximately 0.2, the power spectral density of CMA-MESO and the regression model is closer to the 3 km reanalysis (which is used as the ground truth) than that of the CorrDiff model, suggesting that CorrDiff may produce some non-physical high-frequency components. As shown in Fig. 2, the power spectral density derived from the CorrDiff model shows a closer agreement with the ground truth than that from CMA-MESO for most variables, particularly those near the surface, within the zonal wavenumber

range of [0.005, 0.05]. For example, the power spectral density of zonal and meridional winds at 700, 850, and 925 hPa from CorrDiff is significantly closer to the ground truth than those from CMA-MESO and the regression model. Notably, for mean sea level pressure, all downscaling models outperform CMA-MESO. In Fig. 3, for radar composite reflectivity, the CorrDiff model achieves clearly superior power spectral density performance relative to CMA-MESO, indicating that diffusion-based models are particularly suitable for representing fine-scale structures of radar reflectivity.

Figures 4 and 5 present the structural similarity index measure values for the downscaled meteorological variables and radar composite reflectivity, respectively. The regression model achieves higher structural similarity index measure than the corresponding CorrDiff counterparts. For nearly all variables and all low-resolution input data, the regression model outperforms CMA-MESO. Among the three classes of outputs including Regression CMA-GFS, Regression SFF (ERA5) and Regression SFF (CRA), the performance follows a consistent order. Regression SFF (ERA5) performs best, followed by Regression SFF (CRA) and then Regression CMA-GFS. This ranking is consistent with the results obtained from the MAE metric.

These additional diagnostics support a more robust and complete comparison between our data-driven models and CMA-MESO, and we have revised the relevant statements in the manuscript accordingly.

Major comment 3: The results indicate that CorrDiff predictions have larger MAE than those of the regression models. The authors should discuss this in greater detail. For example, could the CorrDiff model be adding stochastic noise rather than learning physically meaningful corrections?

Response: Thanks for this important and insightful comment. Figures 6 and 7 illustrate the correction fields generated by the CorrDiff models. In these difference plots (CorrDiff minus regression), red and blue regions represent positive and negative corrections, respectively. We observe that the corrections are predominantly positive or negative for some variables, such as 2m temperature, while positive and negative corrections are nearly balanced for other variables such as 10m wind. The magnitudes of corrections also vary spatially. For 10m wind, corrections are notably stronger over the Qinghai-Tibet Plateau and oceanic regions than over other areas. For radar composite reflectivity, corrections are concentrated in limited regions and nearly zero elsewhere. Nevertheless, we notice that corrections for certain variables in some regions resemble random noise. These results suggest that corrections should be applied selectively rather than uniformly across the entire domain, and a more refined correction strategy could further reduce prediction errors.

Major comment 4: Given the large spatial domain used for training, the authors should discuss the number of model parameters, computational cost estimates, and memory requirements. These details are important for assessing the feasibility of the proposed approach.

Response: We have supplemented detailed information on model parameters, training cost, and memory usage. All experiments in this work are conducted using 8 NVIDIA H20 GPUs. The

computational costs are summarized below.

- Training time per epoch:
  - Regression 1 and 3-1: approximately 4 hours.
  - Regression 2: approximately 3.6 hours.
  - Regression 3-2: approximately 1.3 hours.
  - Regression 4: approximately 3.3 hours.
  - CorrDiff 1, 2, and 4: approximately 2 hours.
- Pytorch model file size:
  - Regression 1, 2, and 3-1: 3.35 GB.
  - Regression 3-2: 312 MB.
  - Regression 4: 2.74 GB.
  - CorrDiff 1, 2, and 4: 6.21 GB.
- Peak GPU reserved memory (with one batch per GPU):
  - Regression 1, 2, and 3-1: approximately 85 GB.
  - Regression 3-2: approximately 20 GB.
  - Regression 4: approximately 65 GB.
  - CorrDiff 1: approximately 77 GB.
  - CorrDiff 2: approximately 70 GB.
  - CorrDiff 4: approximately 60 GB.
- UNet embedding size:
  - Regression 1, 2, and 3-1: 128, 256, 512, 512, 1024.
  - Regression 3-2: 32, 64, 128, 256, 256.
  - Regresison 4: 96, 192, 384, 768, 768.
  - CorrDiff 1, 2, and 4: 128, 256, 512, 512, 1024.

These details have been added to the revised manuscript to fully demonstrate the feasibility and practicality of the proposed framework.

Major comment 5: The manuscript makes heavy use of abbreviations, which may reduce readability. The authors may consider limiting the number of abbreviations.

Response: We have carefully revised the manuscript to limit the number of abbreviations, removed unnecessary abbreviations, ensured that all remaining abbreviations are spelled out at their first occurrence with clear definitions, and replaced some low-frequency abbreviations with their full names (including PSD, SSIM, NWP, SFNO, SISL, and FSS). We hope these changes meet the referee's requirements.

Minor comment 1: The manuscript is generally understandable but contains several language and stylistic issues.
Minor comment 2: Line 20: Please spell out the abbreviation "km" at first occurrence.

Response: We appreciate the reviewer's careful reading and constructive suggestions. We have carefully revised the language and stylistic issues throughout the manuscript to improve clarity and fluency. In addition, "km" has been spelled out as "kilometers" at its first occurrence in accordance with the requirement. All relevant revisions have been marked in the revised

manuscript.

Minor comment 3: Lines 20-21: The manuscript states that computational time limits high-resolution NWP forecasts. Please clarify what these limitations are. Providing an example of the computational resources required for km-scale NWP would strengthen this statement.

Response: Increasing the spatial resolution results in a quadratic or cubic growth in the number of grid points, which drastically raises floating-point operations and computational time. This often makes it impossible to deliver forecasts within the time window required for real-time operational weather services. To alleviate this burden, massive computational resources are typically required. As a representative example, an ECMWF team carried out a global 1 km resolution simulation on the Summit supercomputer at Oak Ridge National Laboratory. Even using a subset of the machine—960 Summit nodes, with 5760 MPI tasks and 28 threads per task—the model achieved only about one simulated week per wall-clock day, with a per-time-step cost of 5.4 seconds (Wedi et al., 2020). This example clearly illustrates the extreme computational demands of km-scale NWP.

Wedi, N.P., Polichtchouk, I., Dueben, P., Anantharaj, V.G., Bauer, P., Boussetta, S., Browne, P., Deconinck, W., Gaudin, W., Hadade, I. and Hatfield, S., 2020. A baseline for global weather and climate simulations at 1 km resolution. Journal of Advances in Modeling Earth Systems, 12(11), p.e2020MS002192.

Minor comment 4: Line 22 and throughout the manuscript: Please ensure that citations follow the correct citation format consistently.

Response: We have carefully checked and uniformly standardized the citation format throughout the entire manuscript to ensure consistency and correctness in all references.

Minor comment 5: Line 32: The manuscript states that statistical downscaling can be more accurate than dynamical downscaling. Please provide supporting references or evidence for this claim.

Response: We appreciate the reviewer's careful comment. We have added relevant supporting references and evidence to justify the statement that statistical downscaling can achieve higher accuracy than dynamical downscaling. The corresponding literature has been cited in the revised manuscript to strengthen the claim and improve the reliability of the introduction.

[1] Legasa, M. N., A.Casanueva, and R.Manzanas. 2026. "Strengths and Limitations of Statistical and Dynamical Downscaling for the Representation of Compound Dry and Hot Events Over Spain." International Journal of Climatology. https://doi.org/10.1002/joc.70183.

[2] Diez, E., Primo, C., Garcia-Moya, J.A., Gutiérrez, J.M. and Orfila, B., 2005. Statistical and dynamical downscaling of precipitation over Spain from DEMETER seasonal forecasts. Tellus A: Dynamic Meteorology and Oceanography, 57(3), pp.409-423.

Minor comment 6: Line 39: I guess the authors intended to use “similar”.

Response: We have revised the inappropriate word to similar to ensure accuracy and readability of the manuscript.

Minor comment 7: The introduction section briefly mentions conclusions before defining the problem and presenting the relevant results. The section would benefit from restructuring to provide a clearer overview of the state of the art in downscaling and to identify the research gap addressed by the manuscript.

Response: We sincerely appreciate the reviewer’s constructive suggestion on improving the structure of the introduction. We have improved the introduction by providing a clearer overview of the state of the art in deep learning-based downscaling methods.

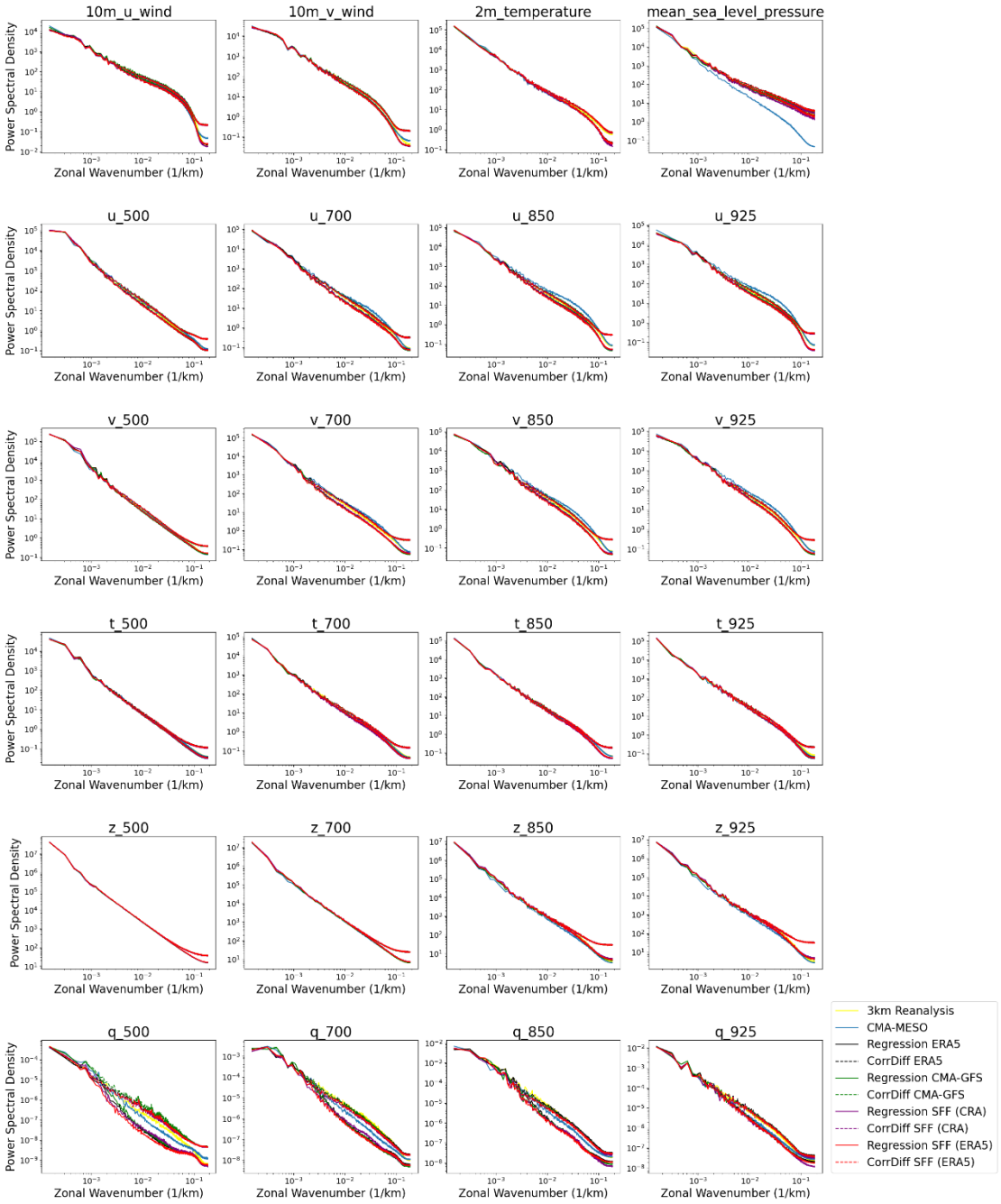

Figure 1. Power spectral density of the downscaling variables. Results are shown for Regression 4 and CorrDiff 4. The 3 km reanalysis and ERA5 data used in this figure are valid at 2023-03-02-00 UTC. The CMA-GFS, CMA-MESO and SFF forecasts are 24-hour forecasts initialized at 2023-03-01-00 UTC.

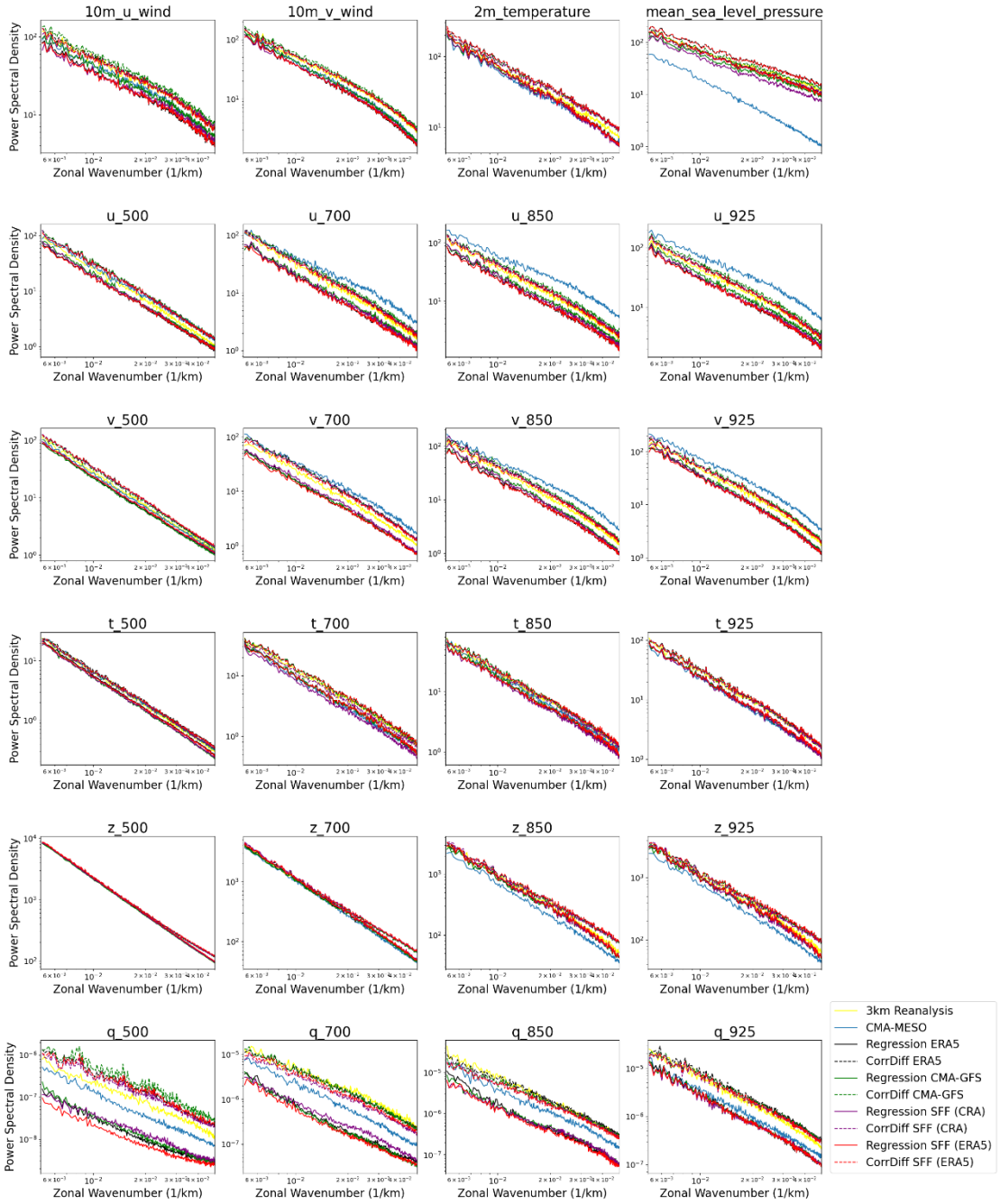

Figure 2. Same as Figure 1, but focused on the zonal wavenumber interval [0.005, 0.05] for improved visualization of detailed spectral features.

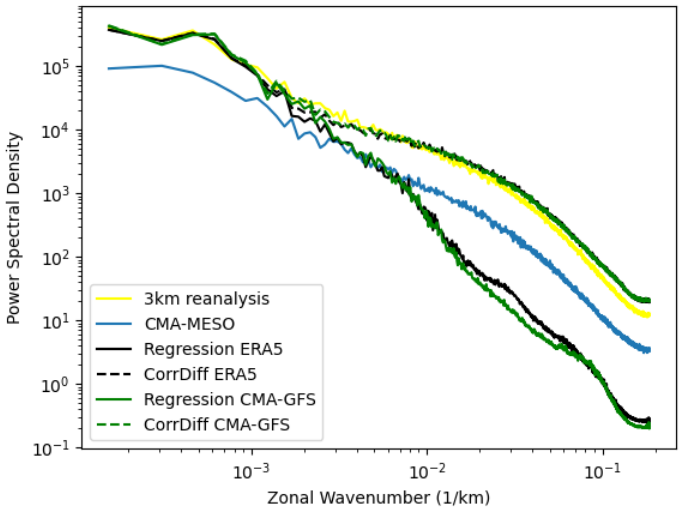

Figure 3. Power spectral density of radar composite reflectivity. Results are presented for Regression 2 and CorrDiff 2. The 3 km reanalysis and ERA5 data correspond to 2023-07-31-00 UTC. The CMA-GFS and CMA-MESO forecasts are 24-hour forecasts initialized at 2023-07-30-00 UTC.

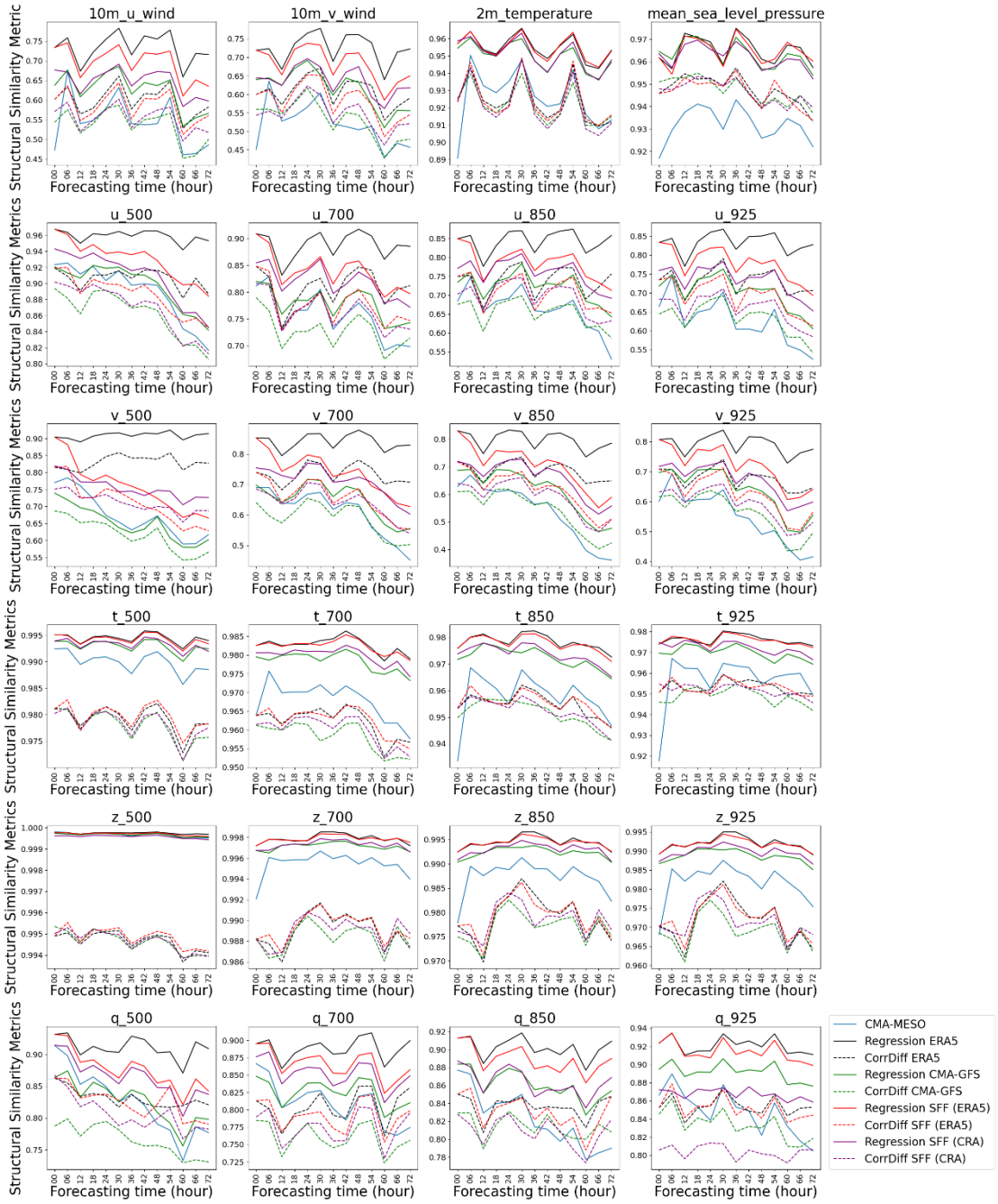

Figure 4. The structural similarity index measure of 3 km forecasts, including downscaled SFF forecasts initialized from CRA1.5 or ERA5, downscaled CMA-GFS forecasts, and CMA-MESO forecasts, validated against 3 km reanalysis as the ground truth. Downscaled ERA5 results are also included for comparison. The downscaling models are Regression 4 and CorrDiff 4, with an initialization time of 2023-03-01-00 UTC.

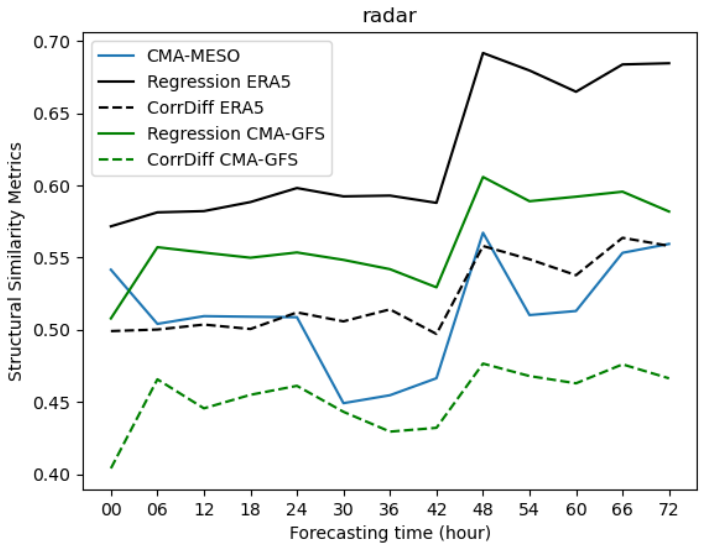

Figure 5. The structural similarity index measure of 3 km radar composite reflectivity forecasts, including downscaled CMA-GFS forecasts, CMA-MESO forecasts, and downscaled ERA5 outputs, evaluated against 3 km reanalysis as the ground truth. The downscaling models are Regression 2 and CorrDiff 2, initialized at 2023-07-30-00 UTC.

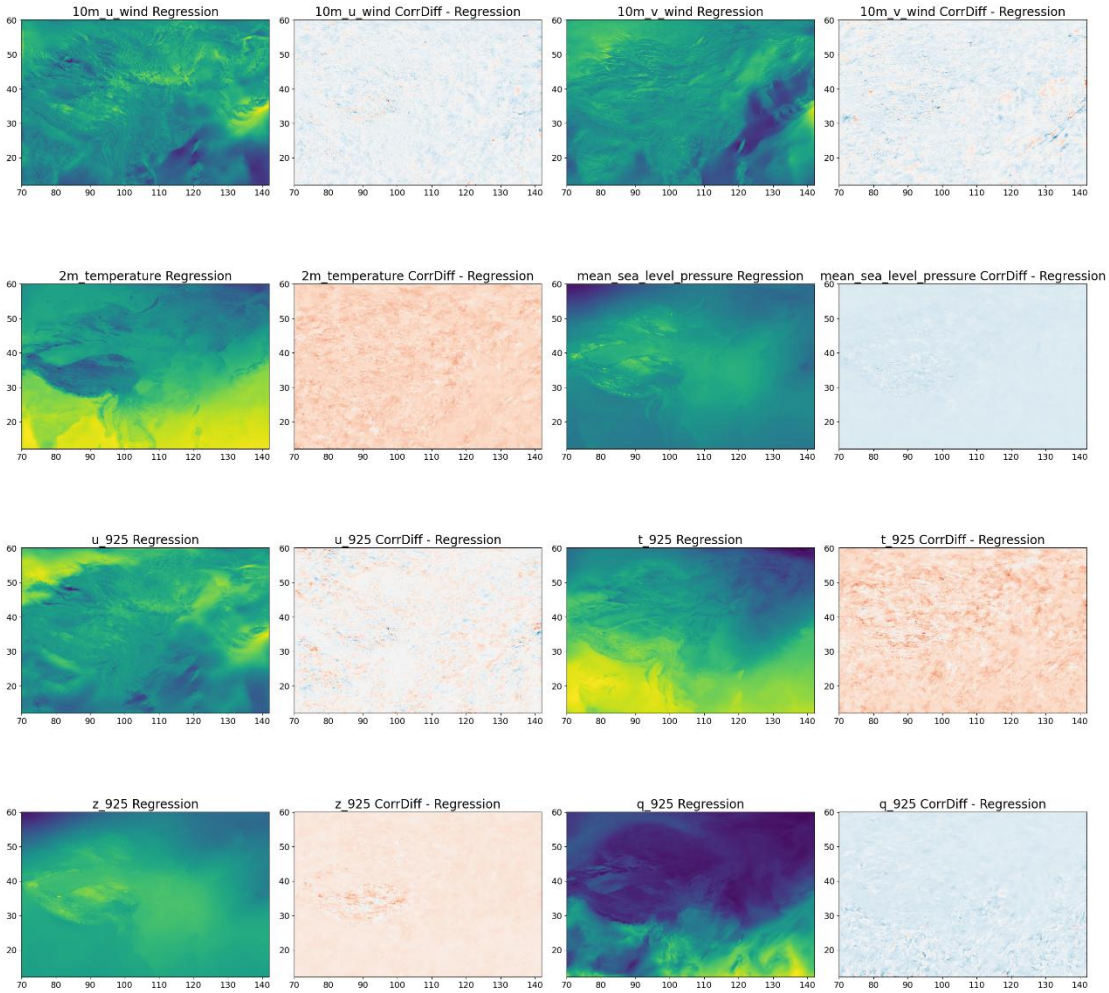

Figure 6. Outputs of Regression 4 and the corresponding differences between CorrDiff 4 and Regression 4 (CorrDiff 4 −Regression 4) for selected downscaled variables. Inputs are derived from ERA5 data at 2023-03-02-00 UTC.

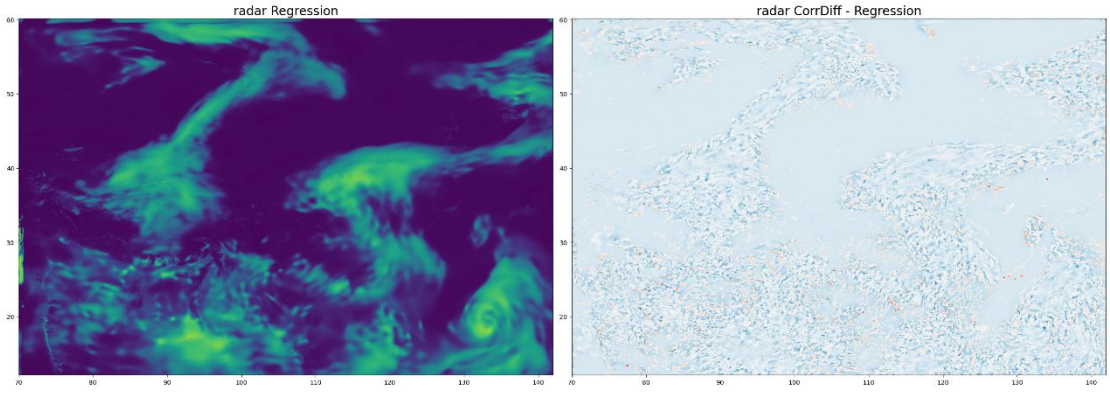

Figure 7. Outputs of Regression 2 and the differences between CorrDiff 2 and Regression 2 (CorrDiff 2 −Regression 2) for radar reflectivity. Inputs are obtained from ERA5 data at 2023-07-30-00 UTC.