

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.

Changes in manuscript:

New content has been added to the first paragraph of Section 3.3 “Downscaling the SFF and CMA-GFS Forecasts”.

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.

Fig. 1 (of this file) 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 (of this file). 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 (of this file), 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 (of this file), 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 (of this file) 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.

Changes in manuscript:

New figures (Figures 12, 13, 14, 15 and 16 of the revised manuscript) and corresponding discussions (the last three paragraphs of Section 3.3.3 “3km Forecast Evaluation”) have been added.

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: Thank you for this important and insightful comment. This behavior is expected because the regression model is trained to minimize a deterministic pointwise loss (MAE), whereas CorrDiff is trained to model the conditional distribution of the residual through diffusion. Consequently, a single stochastic sample generated by CorrDiff is not expected to minimize MAE and can therefore produce a larger deterministic error than the regression prediction. This observation is also consistent with the original CorrDiff paper.

To further investigate whether CorrDiff learns physically meaningful corrections or merely injects stochastic noise, we analyzed the correction fields generated by the diffusion model. Figures 6 and 7 (of this file) illustrate the difference between CorrDiff and regression predictions (CorrDiff

minus regression), where 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 the corrections also exhibit clear spatial dependence. For 10m wind, corrections are notably stronger over the Qinghai–Tibet Plateau and oceanic regions than over other areas. For radar composite reflectivity, the corrections are concentrated in limited regions and remain close to zero elsewhere. These characteristics indicate that the diffusion model does not simply add random perturbations uniformly across the domain. Instead, it learns spatially organized corrections that are associated with meteorologically active regions and variables with strong small-scale variability.

At the same time, we also observe that, for some relatively smooth variables and in certain regions, part of the correction field resembles stochastic noise. This suggests that the current CorrDiff framework may over-correct weak residual signals, which could partially explain the degradation in deterministic MAE. We therefore believe that diffusion-based corrections should be applied selectively rather than uniformly across all variables and regions. Developing region-aware or variable-dependent correction strategies is an important direction for future work and may further improve deterministic prediction accuracy while preserving the ability of CorrDiff to generate realistic fine-scale structures and probabilistic forecasts.

Changes in manuscript:

New figures (Figures 5 and 6 of the revised manuscript) and corresponding discussions (third paragraph of Section 3.2.2 “CorrDiff Models”) have been added.

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.
- The number of model parameters:
 - Regression 1, 2, and 3-1: $8.99e8$.

- Regression 3-2: $8.18e7$.
- Regression 4: $7.35e8$.
- CorrDiff 1, 2, and 4: $1.67e9$.
- 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.
 - Regression 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.

Changes in manuscript:

New content has been added (the last paragraph of Section 3.2.2 “CorrDiff Models”).

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.

Changes in manuscript:

Some abbreviations have been removed.

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.

Changes in manuscript:

New content has been added to the first paragraph of Section 1 “Introduction”.

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.

Changes in manuscript:

New references have been added to the second paragraph of Section 1 “Introduction”.

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.

Changes in manuscript:

New content has been added to the fourth paragraph of Section 1 “Introduction”.

Response to Referee 2

Comments: The original CorrDiff framework relies on the assumption of spatially stationary and homogeneous residual distributions, which is theoretically invalid over a large domain with highly heterogeneous surfaces (e.g. the Tibetan Plateau, deserts, coasts, and mountains in China). Furthermore, the uncertainly estimated from diffusion ensemble variance is empirically motivated and lacks a solid theoretical link to atmospheric intrinsic predictability, initial condition errors, or model error sources.

In this sense, the authors should justify the application of the stationary residual assumption over complex regions, for example by spatially stratified error analysis or regional subdomain modeling.

Response:

We sincerely appreciate the reviewer’s thorough and constructive comments, which greatly help improve the quality of this manuscript. Detailed responses are presented below.

First, regarding the assumption of spatially stationary and homogeneous residual distributions. As pointed out, the study area of China features highly heterogeneous underlying surfaces including the Tibetan Plateau, deserts, coasts and mountains, so the residual field naturally violates the strict spatial stationarity and homogeneity assumption. This spatial heterogeneity is well validated by Figures 8 and 9 and Table 1 (of this file), which exhibit differentiated prediction performances at different vertical levels and geographical subdomains.

Technically, the residual correction model used in this work is a conditional generative model rather than an unconditional residual sampler. The diffusion model is conditioned on the interpolated low-resolution fields, the regression-model output, and high-resolution orography. Therefore, although a single set of network parameters is shared over the whole domain, the learned residual distribution is conditional on spatially varying meteorological states and terrain information. This design allows the model to represent spatially dependent residual structures to some extent.

Nevertheless, we recognize that the current conditional structure cannot fully resolve the challenges caused by regional heterogeneity. To quantify this issue, we have added a spatially stratified error analysis in the revised manuscript. The results show that the prediction errors and CorrDiff corrections exhibit clear regional and vertical dependence. For example, stronger corrections are observed over complex-terrain and oceanic regions for near-surface wind. These results confirm the reviewer's concern that the residual statistics are not uniform over the domain, but they also show that the present conditional CorrDiff model can partially adapt to such heterogeneity through its meteorological and orographic conditioning.

We have supplemented relevant discussions to clarify this inherent limitation. In future work, we will investigate more region-aware CorrDiff variants, including spatially stratified training/evaluation, terrain- or climate-zone-dependent normalization, regional expert modules, and mixture-of-experts or domain-adaptive diffusion heads for different subdomains and variables. On this basis, we will further implement targeted differentiated modeling strategy for different subdomains and meteorological variables at different pressure levels, such as employing distinct neural network modules for diverse regions and variables, to better account for the distinct residual distributions over the Tibetan Plateau, arid regions, coastal areas, and mountainous regions.

Second, with respect to uncertainty estimation based on diffusion ensemble variance. We agree that the variance of the diffusion ensemble should not be interpreted as a rigorous measure of atmospheric intrinsic predictability, initial-condition uncertainty, or model-error covariance. The core goal of this study is to demonstrate the feasibility of the residual corrective diffusion model for high-resolution downscaling and explore whether the ensemble spread from the diffusion process can serve as a practical empirical proxy for predictive uncertainty. This approach is empirically inspired by the successful applications of diffusion ensembles in other fields such as computer vision, where stochastic sampling during inference can produce diverse and physically plausible outputs. We have never intended to claim that the ensemble spread can quantify classical predictability limits or error covariance in a dynamically consistent manner.

In the revised manuscript, we have softened the relevant claim and now describe the diffusion ensemble variance as an empirical proxy for predictive uncertainty. Our analysis shows that grid points with lower CorrDiff ensemble variance generally have lower absolute error, and that variance-based grouping can separate regions with different error levels. Thus, the ensemble spread provides useful diagnostic information about the reliability of the downscaled prediction,

but it does not by itself provide a complete dynamical uncertainty decomposition. We have clarified this distinction in the revised manuscript.

Finally, we have summarized two key research directions for follow-up work: (1) Distinct prediction patterns across different variables and subdomains are observed in this work. The CorrDiff framework could be optimized via subdomain- and variable-specific customized modeling to adapt to heterogeneous regions; (2) The current uncertainty estimate is empirically motivated; future work should aim to establish theoretical connections between diffusion-based uncertainty quantification and atmospheric predictability to improve the physical interpretability of uncertainty estimation.

Changes in manuscript:

A new paragraph (last paragraph of Section 3.2.3 “Uncertainty Estimation based on CorrDiff”) has been added to soften the relevant claim about the uncertainty estimation based on diffusion ensemble variance and to describe the diffusion ensemble variance as an empirical proxy for predictive uncertainty.

New figures (Fig. 7 of the revised manuscript), new table (Table 4 of the revised manuscript) and corresponding discussions (the fourth paragraph of Section 3.2.2 “CorrDiff Models”) have been added regarding the assumption of spatially stationary and homogeneous residual distributions.

Two new directions of future work have been added to the Conclusion section.

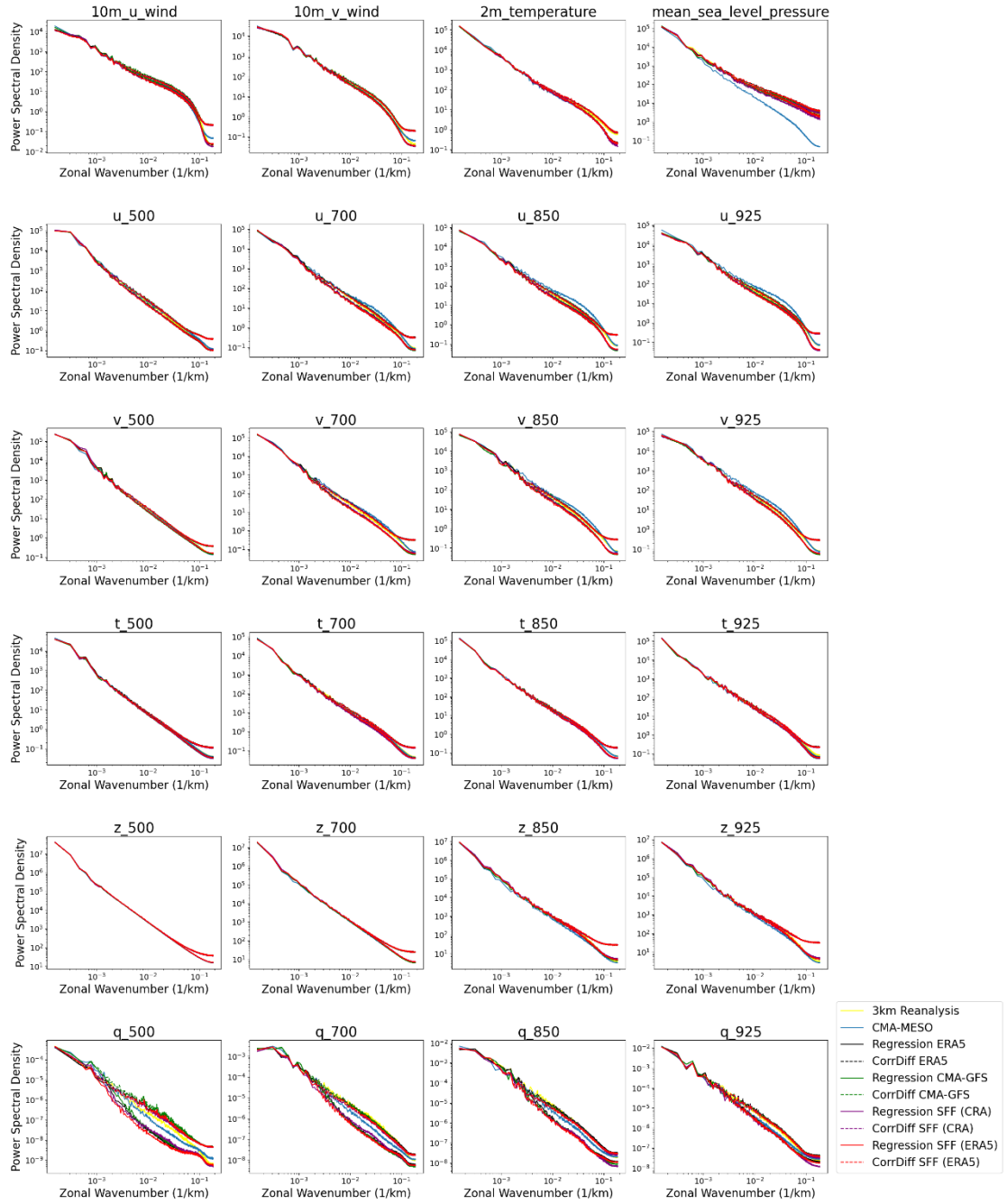


Figure 1. Power spectral density of the downscaling variables. Results are shown for Regression 4 and CorrDiff 4. The 3 km reanalysis and ERA5 data used in this figure are valid at 2023-03-02-00 UTC. The CMA-GFS, CMA-MESO and SFF forecasts are 24-hour forecasts initialized at 2023-03-01-00 UTC.

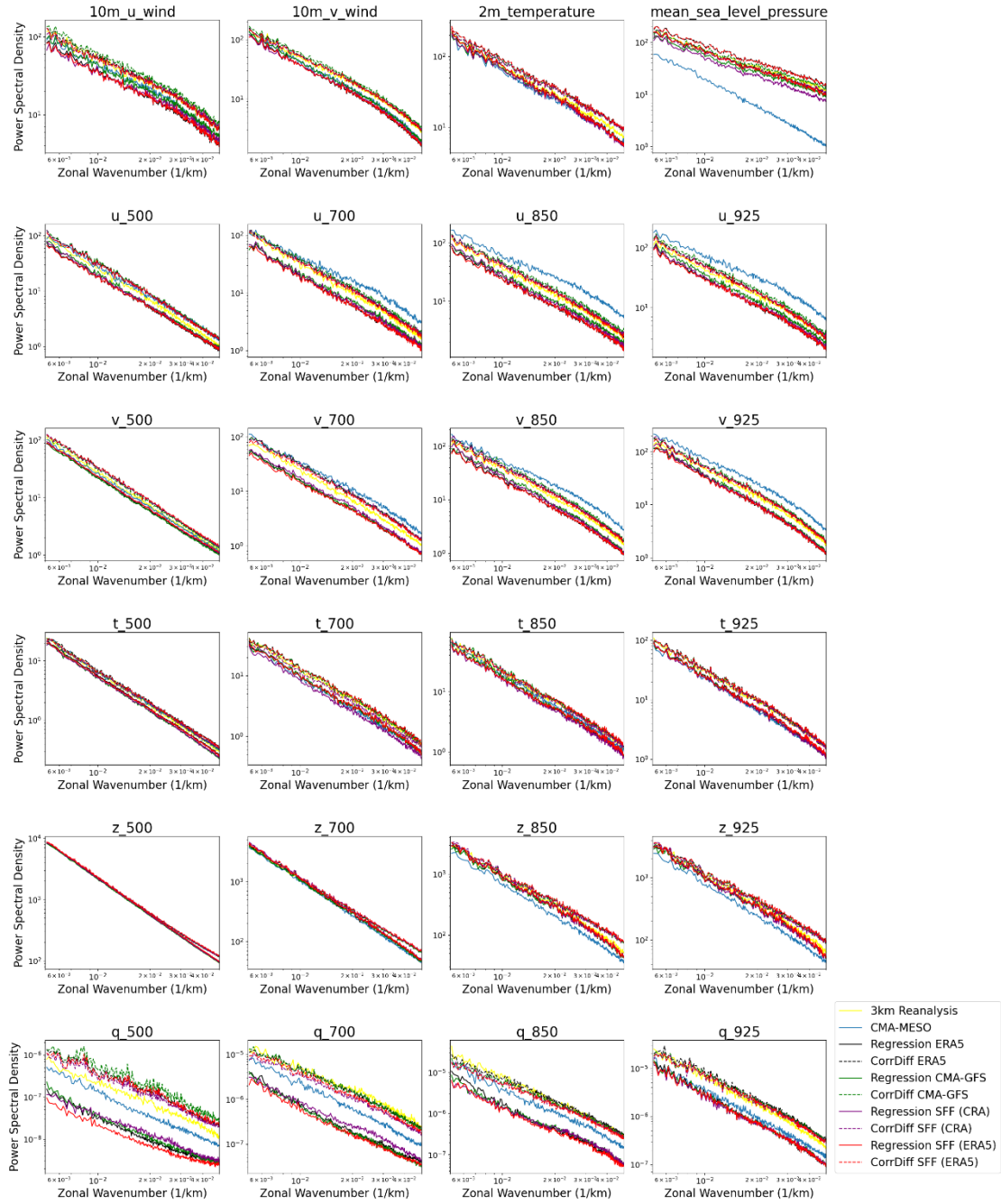


Figure 2. Same as Figure 1, but focused on the zonal wavenumber interval $[0.005, 0.05]$ for improved visualization of detailed spectral features.

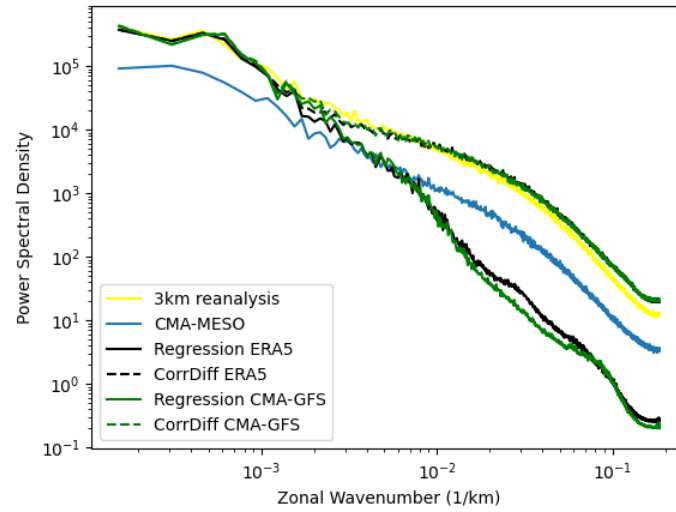


Figure 3. Power spectral density of radar composite reflectivity. Results are presented for Regression 2 and CorrDiff 2. The 3 km reanalysis and ERA5 data correspond to 2023-07-31-00 UTC. The CMA-GFS and CMA-MESO forecasts are 24-hour forecasts initialized at 2023-07-30-00 UTC.

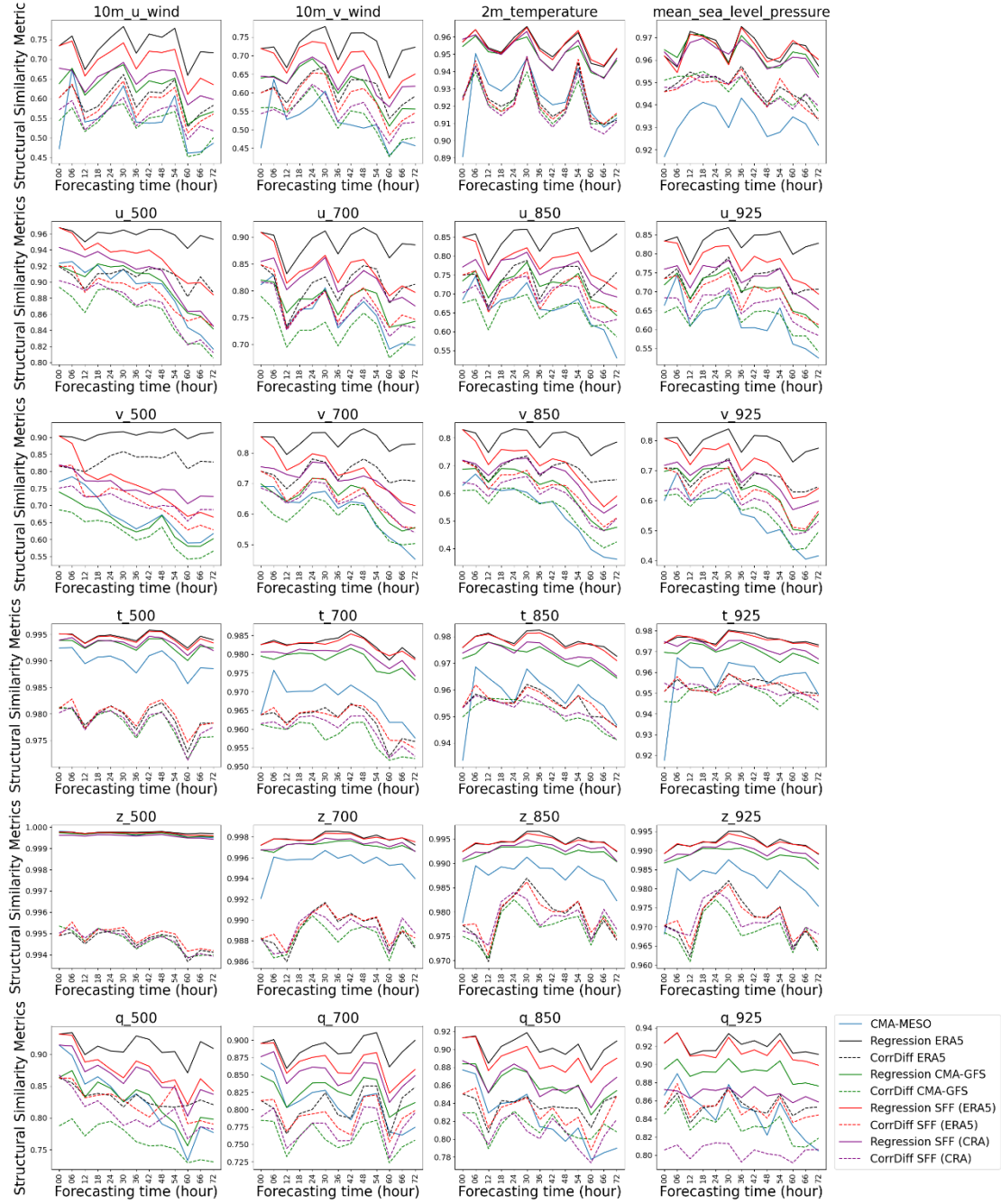


Figure 4. The structural similarity index measure of 3 km forecasts, including downscaled SFF forecasts initialized from CRA1.5 or ERA5, downscaled CMA-GFS forecasts, and CMA-MESO forecasts, validated against 3 km reanalysis as the ground truth. Downscaled ERA5 results are also included for comparison.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

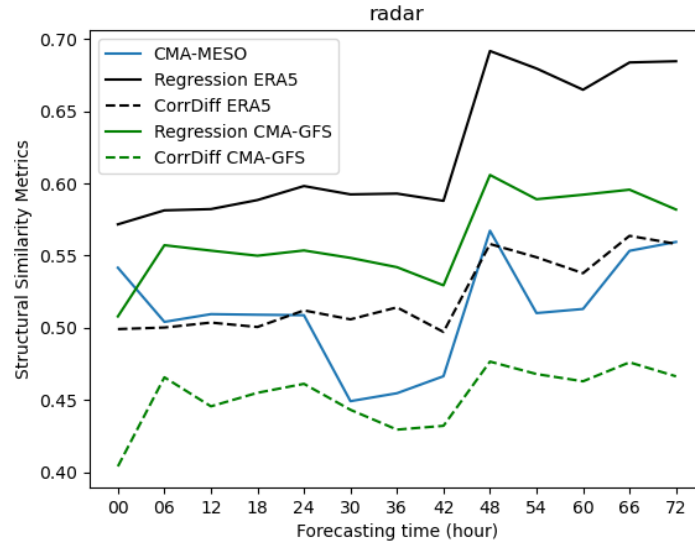


Figure 5. The structural similarity index measure of 3 km radar composite reflectivity forecasts, including downscaled CMA-GFS forecasts, CMA-MESO forecasts, and downscaled ERA5 outputs, evaluated against 3 km reanalysis as the ground truth. The downscaling models are Regression 2 and CorrDiff 2, initialized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

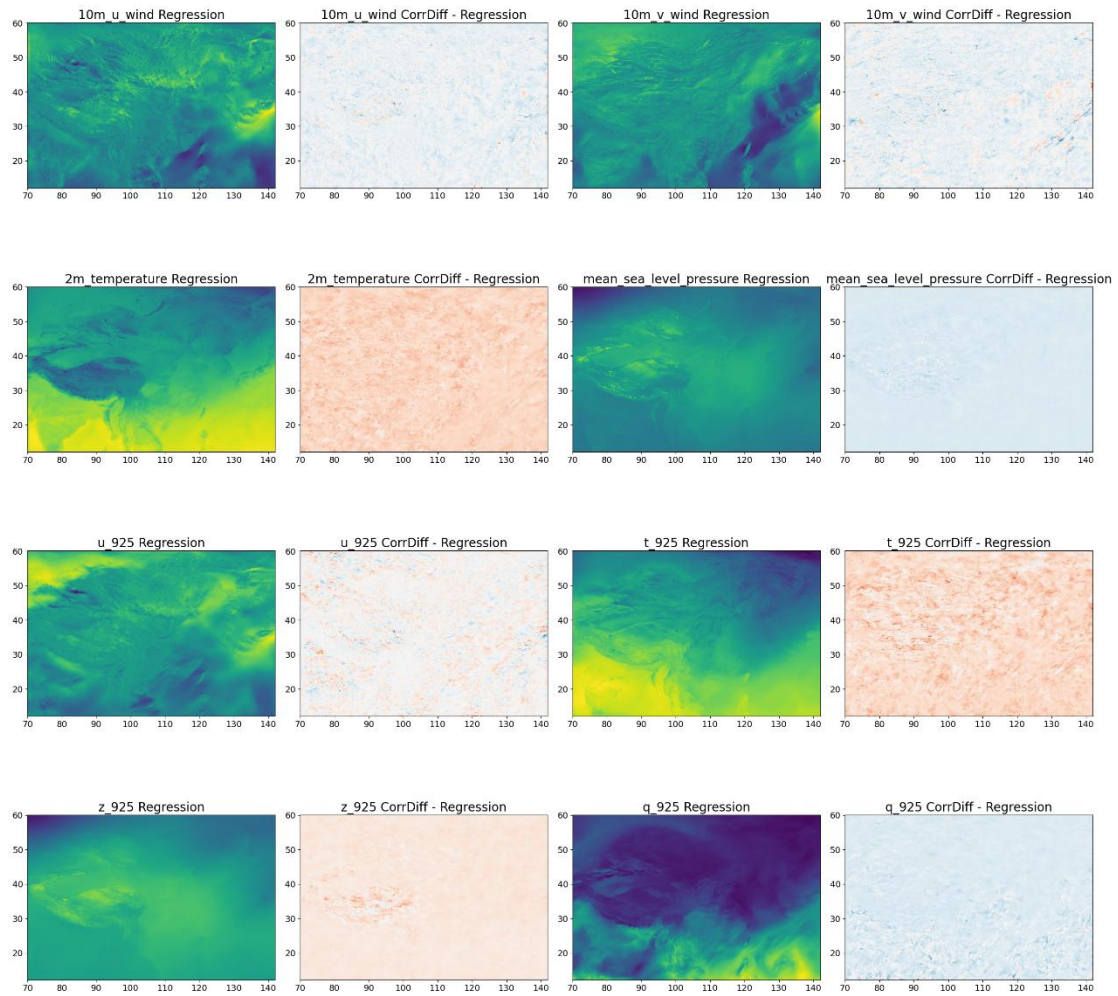


Figure 6. Outputs of Regression 4 and the corresponding differences between CorrDiff 4 and Regression 4 ($\text{CorrDiff4} - \text{Regression4}$) for selected downscaled variables. Inputs are derived from ERA5 data at 2023-03-02-00 UTC.

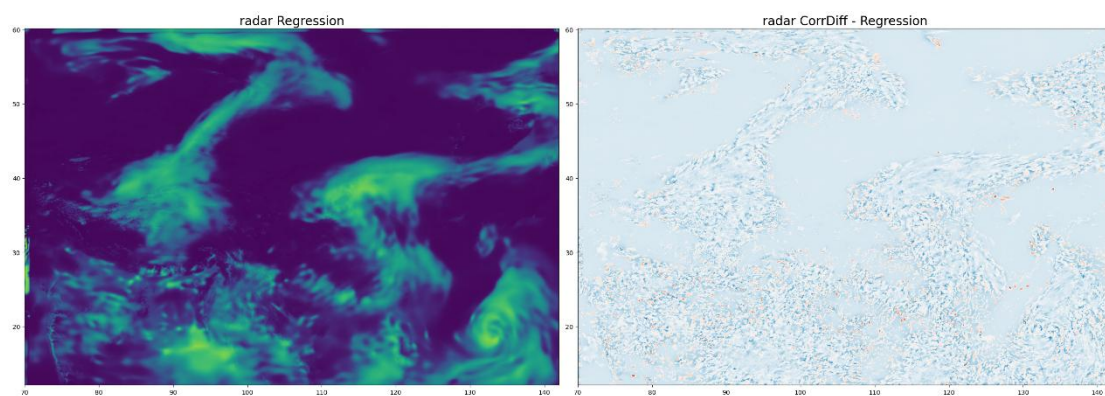


Figure 7. Outputs of Regression 2 and the differences between CorrDiff 2 and Regression 2 ($\text{CorrDiff2} - \text{Regression2}$) for radar reflectivity. Inputs are obtained from ERA5 data at 2023-07-30-00 UTC.

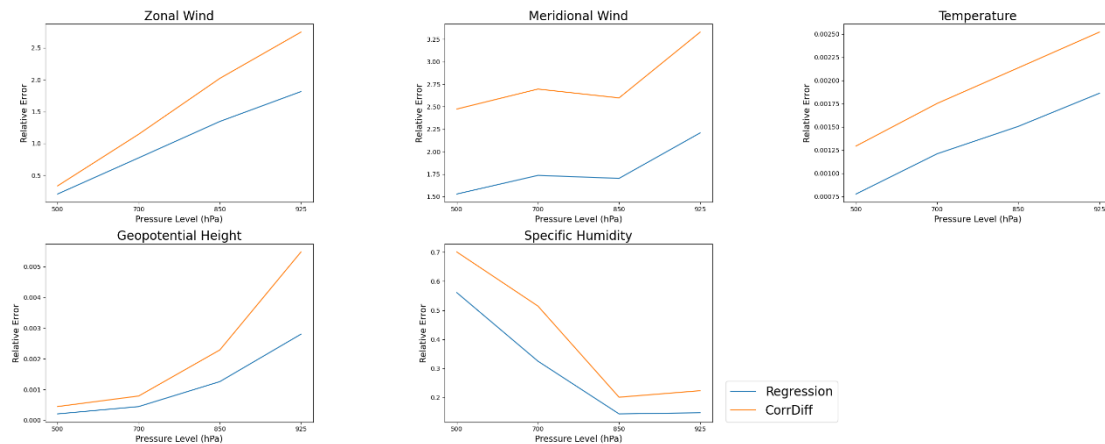


Figure 8. Relative error at different pressure levels, based on data from 00, 06, 12, and 18 UTC on 2023-03-01, 2023-03-02, and 2023-03-03. Downscaling models are Regression 4 and CorrDiff 4.

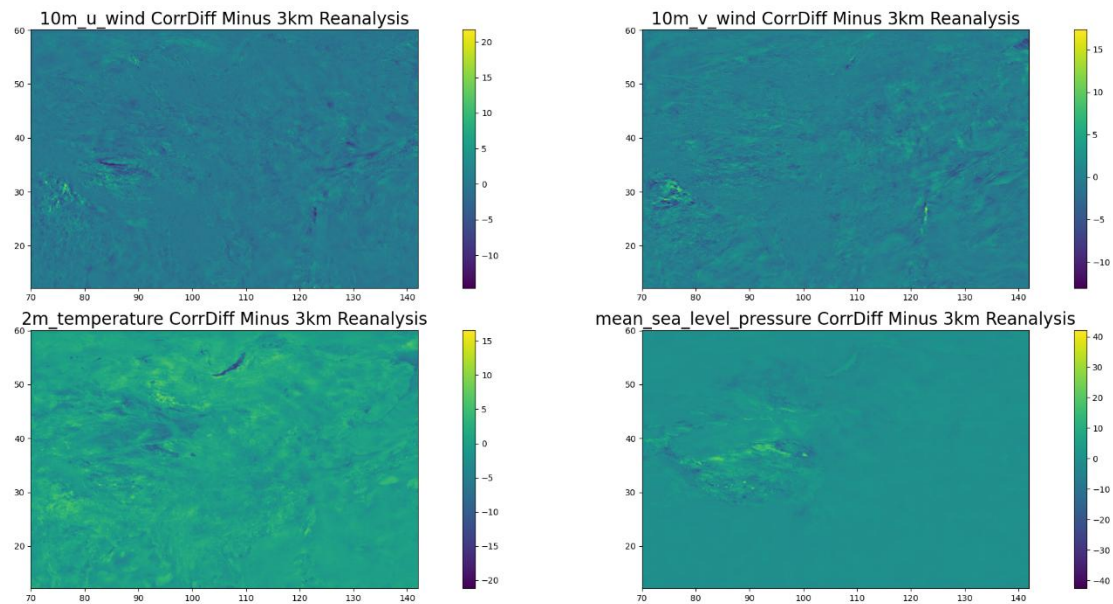


Figure 9. Spatial distribution of error, based on data from 00 UTC on 2023-03-01.

Variable	Metric	Tibetan Plateau	East China	South China coast	Northwest arid region
	MAE (Regression)	0.9359	0.6284	0.5828	0.8952
10m Zonal Wind (m/s)	MAE (CorrDiff)	1.3407	0.8771	0.8155	1.2313
	SSIM (Regression)	0.5582	0.6332	0.6790	0.6078
	SSIM (CorrDiff)	0.4079	0.4851	0.5402	0.4523
2m Temperature (K)	MAE (Regression)	1.1950	0.8553	0.9491	1.2752
	MAE (CorrDiff)	1.3726	1.1482	0.9491	1.4615
	SSIM (Regression)	0.9414	0.8499	0.8364	0.9200
	SSIM (CorrDiff)	0.9047	0.7960	0.7898	0.8486

Table 1. MAE and the structural similarity index measure (SSIM) of 10m zonal wind and 2m temperature downscaled by Regression 4 and CorrDiff 4, evaluated over four regions: Tibetan Plateau (25.99-40N, 70-104.68E), East China (21.01-38.23N, 112.99-122.98E), South China coast (18.01-25.51N, 95.98-118E) and Northwest arid region (34.99-49.99N, 73-106E), using data from 00, 06, 12, and 18 UTC on 2023-03-01, 2023-03-02, and 2023-03-03.