China Regional 3 km Downscaling Based on Residual Corrective Diffusion Model

The manuscript proposes a deep-learning framework for statistical downscaling of meteorological forecasts over China. The approach builds on the Corrective Diffusion (CorrDiff) model, combining a deterministic regression model (UNet), trained to predict the mean of high-resolution meteorological variables, with a diffusion model trained to correct the residual errors of the regression model.

The model maps 25 km resolution global data to 3 km regional fields. Training is performed using ERA5 reanalysis as low-resolution input and CMA-RRA 3 km reanalysis as high-resolution targets. The trained models are applied to forecasts from two global systems: CMA-GFS (a numerical weather prediction model) and SFF (a deep-learning weather model).

The resulting downscaled forecasts are compared with those from CMA-MESO, a regional numerical weather prediction model. The authors claim that the proposed framework produces lower mean absolute error (MAE) for most variables and generates more realistic fine-scale features compared with deterministic regression models.

Below are my comments, divided into two categories.

Major comments

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?
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."
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?
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.
5. The manuscript makes heavy use of abbreviations, which may reduce readability. The authors may consider limiting the number of abbreviations.

Minor comments:

1. The manuscript is generally understandable but contains several language and stylistic issues.
2. Line 20: Please spell out the abbreviation “km” at first occurrence.
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.
4. Line 22 and throughout the manuscript: Please ensure that citations follow the correct citation format consistently.
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.
6. Line 39: I guess the authors intended to use “similar”.
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.