Recommendation: Major Revision

The manuscript presents CEDDAR v1.0.2, a generative downscaling framework based on the Elucidated Diffusion Model (EDM) designed to downscale daily precipitation from ERA5 (~30 km) to 2.5 km DANRA resolution over Denmark. Developing generative AI architectures for precipitation downscaling is a rapidly evolving front and a key area of interest.

In this context, the manuscript makes a valuable contribution by introducing σ*, a scale-aware inference-time parameter. By scaling only the late-stage injected noise in the stochastic sampler during inference, this parameter acts as an interpretable bias-variance regulator, allowing users to modulate the balance between large-scale coherence and fine-scale texture. The systematic component-wise ablation study is also structured logically and provides a useful overview of how different empirical constraints interact.

Despite these conceptual strengths, several fundamental architectural choices, methodological limitations, and constraints regarding computational scalability are insufficiently addressed, which weakens the manuscript's broader conclusions.

My main concerns are detailed below.

## Exclusion of Residual Learning based on under-conditioned baseline (B0)

In Section 2.3 and Supplement Section S1.4, the authors justify their decision to discard residual learning by showing that it degraded performance and "over-constrained the denoising process". However, looking at the experimental setup in Table S10, residual prediction (_R) was only evaluated on the minimal baseline configuration B0, which completely lacks geographic conditioning (topography and land-sea mask) and seasonal conditioning.

This is a significant methodological shortcut. The precipitation residual field is inherently composed of high-frequency spatial structures that are physically forced by coastlines and orographic lift. Asking the baseline model (B0) to resolve these residual features without giving it access to the physical spatial anchors (DEM, coastlines) that dictate where these features occur is mathematically and physically under-conditioned. In complex spatial downscaling, orographic and coastal priors are what allow a residual model to structurally align its high-resolution corrections. Without such a test, the dismissal of the residual framework remains incomplete and potentially misleading.

## Absence of State-of-the-Art Deep Learning Baselines and Insufficiently Demonstrated QM Comparison

The evaluation framework is missing a comparison with established deep learning downscaling architectures. The authors only compare variations of their own model (CEDDAR variants) against bilinear interpolation and Quantile Mapping (QM). To convincingly demonstrate the superiority of this architecture, the authors should include at least one standard deep learning baseline. Furthermore, the criticism of QM as "spectrally overfitted" is not rigorously demonstrated in the main text. Since QM matches or outperforms CEDDAR on several bulk metrics, the authors must provide quantitative, object-based spatial verification (e.g., SAL or Fractions Skill Score) in the main manuscript to prove that QM's spatial coherence is indeed unphysical.

## Missing Scalability and Computational Cost Analysis

CEDDAR is formulated as a pixel-space diffusion model trained on a very small subdomain (180x180 pixels). Pixel-space diffusion is notoriously computationally expensive and does not scale easily. While Section S3 of the Supplement mentions the hardware used (LUMI supercomputer), the manuscript lacks a formal scalability analysis. The authors must address:

- What is the computational cost (training and inference runtime per snapshot) compared to simpler baselines?

- How does a pixel-space EDM scale computationally if applied to larger, continental-scale domains (e.g., CORDEX Europe)?
- Why was a pixel-space approach chosen over other computationally efficient models, such as latent diffusion models, which operate in a compressed latent space?

### Limitations of the Physics-Guided Design in Flat Topography

The "physics-guided" aspects of the model are heavily tailored to Denmark, which is flat and coastline-dominated. However, regional precipitation downscaling is most critical in regions with high, complex orography (e.g., Alpine or Andean regions), where orographic lift dominates the rain signal. Testing the model solely on flat terrain limits the generalizability of the framework. To act as a true "diagnostic testbed" for future operational solutions, the authors need to discuss how their topographic conditioning and SDF-weighted loss would translate to complex mountainous terrains.

### Deterministic Evaluation and Probabilistic Biases

The evaluation heavily relies on spatial realism and spectral properties while omitting standard deterministic metrics for ensemble summaries. Furthermore, the probabilistic results exhibit persistent dry biases and under-dispersion that require more rigorous analysis and standard reporting.

Addressing these concerns will be crucial to elevating the manuscript from a visually realistic regional showcase to a robust, generalizable, and physically validated diagnostic downscaling testbed. Additionally, the presentation is often difficult to follow because essential technical details and extended diagnostics are heavily split across documents, forcing the reader to continuously jump back and forth between the main paper and the supplement.