Review for “CEDDAR v1.0.2: Bridging physics and generative modelling for regional precipitation with controllable diffusion-based downscaling” By Quistgaard et al.

The manuscript presents CEDDAR, a diffusion-based generative model for daily precipitation downscaling from ERA5 to DANRA resolution over Denmark. The topic is timely, and the use of diffusion models for probabilistic precipitation downscaling is potentially interesting. However, I do not think the manuscript currently provides sufficient evidence to support its main claims. My main concern is that the proposed model is presented as improving downscaling performance, but the results do not convincingly show improvement over much simpler baselines, especially when the ensemble mean or PMM is treated as a deterministic product. The probabilistic results are also not fully convincing because the generated ensemble appears biased and under-dispersed, and the target distribution is not consistently captured by the ensemble distribution.

For these reasons, I recommend **rejection** in its current form. A substantially revised manuscript would need to provide stronger deterministic evaluation, clearer probabilistic calibration evidence, stronger event-based comparison against QM, and a more cautious framing of the model’s claimed advantages.

My major concerns are as follows:

1. Deterministic performance is not convincingly better than the baselines

The manuscript does not demonstrate that the proposed method gives more accurate deterministic estimates when the ensemble mean, ensemble median, or PMM is used as a single downscaled product. This is a critical issue because many applications require a best-estimate precipitation field, even when ensembles are available. In Figure 7, the ensemble mean shows a systematic dry bias in annual precipitation accumulation, while ERA5 is described as smoother but wet-biased. This does not clearly establish that the proposed model is better in deterministic terms. The authors also state that the generated fields have consistently lower intensity than HR/DANRA and that PMM degrades skill relative to the ensemble mean by smoothing scale-dependent structure. These statements weaken the claim that the proposed model provides an improved deterministic downscaled estimate. The manuscript should report standard deterministic metrics for ensemble mean, ensemble median, and PMM, including MAE, RMSE, bias, correlation, KGE or NSE, wet-day accuracy, and extreme-event detection metrics. These should be compared directly against bilinear ERA5 and QM. Without this, the deterministic value of the proposed method is not established.

2. The probabilistic results are doubtful because the ensemble appears biased and under-dispersed.

The paper claims probabilistic realism and ensemble calibration, but the evidence is mixed. In the supplement, the PIT histogram is described as showing a dry bias, with observations often falling in the upper PIT bins, and slight under-dispersion. Reliability also deteriorates for rarer precipitation thresholds. This is a serious limitation for a probabilistic downscaling model, especially because precipitation extremes are one of the main motivations for using a generative framework. For a well-calibrated probabilistic model, the observed/reference distribution should generally be statistically consistent with the ensemble distribution. In Figures 5–6, the proposed model does not appear to match the target DANRA distribution better than QM. The authors themselves state that the EDM

underestimates precipitation from the mid-range up to approximately P99.9. If the target distribution frequently lies outside the ensemble envelope, then the ensemble is not adequately representing predictive uncertainty. The manuscript should show distributional uncertainty bands across ensemble members, such as 5–95% or 10–90% ensemble envelopes, and explicitly test whether the DANRA distribution is contained within the generated ensemble distribution. Quantile coverage, rank histograms, PIT diagnostics, spread–error relationships, and reliability diagrams should be discussed more directly in the main text, not only in the supplement.

3. QM is criticized, but the criticism is not adequately proven.

The authors criticize quantile mapping by arguing that its favorable spectral behavior may reflect inflated high-wavenumber variability or "spectral overfitting," and that QM may fail to reproduce physically meaningful spatial organization. However, this criticism is not sufficiently demonstrated in the paper. In several diagnostics, QM appears competitive with or better than the proposed method, especially for marginal distributions and some scale-based metrics. The supplement even states that QM consistently achieves high ISS values across scales and thresholds, often exceeding the EDM-generated ensemble, but then argues that these values should not be interpreted as superior spatial performance because ISS rewards marginal distribution matching and smoothing. This explanation may be plausible, but it is not enough. If the authors argue that QM performs well only superficially, they need to provide direct quantitative evidence that QM has worse event morphology, spatial coherence, or hydrologically meaningful structure. The manuscript should include object-based or event-based diagnostics comparing QM, PMM, ensemble mean, and DANRA. Without evidences, the criticism of QM remains speculative.

4. "Physics-guided" is overstated

The model includes topography, land–sea mask, seasonal conditioning, SDF-weighted loss, and a RainGate wet/dry auxiliary head. These are useful physically motivated features, but they are soft empirical constraints rather than physical laws. The manuscript explicitly states that the model is fully data-driven and does not impose hard conservation rules. Therefore, the term "physics-guided" should be used more cautiously. The authors should distinguish between "physically motivated conditioning" and "physically constrained modeling." The current wording risks overstating the degree of physical consistency.

5. The manuscript organization makes the methodology and results difficult to follow.

Important information is split between the main manuscript and the supplement, including EDM formulation, preprocessing, architecture experiments, hyperparameter sweeps, σ^* analysis, and extended evaluation metrics. The supplement contains extensive methodological details that are necessary for understanding the model and interpreting the results. Currently, the results section mixes architecture selection, final evaluation, probabilistic diagnostics, spatial diagnostics, and inference-time control, making it difficult to identify the actual evidence supporting the final claims.