

Final author comments – Response to Referee 1

Submitted by Thea Quistgaard on behalf of all co-authors.

We thank the referee for the detailed assessment and for identifying important weaknesses in the evidence chain supporting some of the manuscript's main claims. We agree that the current framing overstates the demonstrated comparative performance and that several essential diagnostics are insufficiently prominent.

We will substantially reduce claims of comparative superiority and restructure the manuscript around its principal methodological contributions, with clearer emphasis on the inference-time scale-control parameter σ^* and the component analysis. We will further clarify that CEDDAR is primarily a conditional stochastic model whose primary output is an ensemble, and distinguish explicitly between spatial realism, climatological fidelity, deterministic skill, and probabilistic calibration. Finally, we will substantially restructure the manuscript and supplement to improve readability and clarity.

The revised manuscript will also clarify that its scientific contribution extends beyond demonstrating uniformly superior predictive performance. The study introduces and investigates methodological elements specific to diffusion-based stochastic downscaling, including the inference-time control parameter σ^* and the systematic component analysis. The revised manuscript will therefore distinguish more clearly between comparative predictive performance and the methodological insight obtained from analysing the behaviour, trade-offs, and limitations of a generative downscaling framework.

We also note that the two referee reports identify several overlapping concerns, particularly regarding evaluation, QM interpretation, terminology, and manuscript organisation. We believe addressing these consistently will substantially strengthen the manuscript.

In the following, we will respond to each concern, referring to R1.X as the point X raised by Reviewer 1 (R1).

R1.1: Deterministic performance is not convincingly better than the baselines

The reviewer raises concerns regarding standard metrics evaluation for single-field products, asking for ensemble mean, median, and Probability Matched Mean (PMM) comparisons against bilinear ERA5 and Quantile Mapping (QM), including error, bias, correlation, wet-day, and extreme-event metrics.

We agree that deterministic ensemble-derived products were insufficiently evaluated. In the original manuscript, we presented only PMM as a deterministic summary, since we intended any ensemble summary to be a secondary product rather than the full generative output. Nevertheless, we agree that such summaries are relevant for applications requiring a single-field product. We will, in the revised manuscript, therefore, add ensemble mean, median, and PMM, comparing them directly with bilinear ERA5 and QM. We will evaluate the ensemble mean, median, and PMM using a compact set covering complementary aspects of

deterministic performance, including bias, MAE and/or RMSE, correlation, wet-day occurrence/detection, and heavy-precipitation detection. We will interpret the results as product-specific trade-offs and will not imply deterministic superiority where it is not supported.

R1.2: The probabilistic results are doubtful because the ensemble appears biased and under-dispersed

The reviewer notes that PIT, reliability, and distribution results show dry bias and underdispersion, inconsistent with the manuscript's stronger calibration language. Further, they point out that many key probabilistic metrics are placed in the supplement and should be moved to the main paper.

We agree that the use of “*well calibrated*” is too strong and is not fully supported by the results. In the revised manuscript, we will more clearly distinguish ensemble variability, probabilistic skill, and calibration. We will further move probabilistic results, including PIT/rank, reliability, and spread-error diagnostics, into the main text. We will additionally show ensemble distributional uncertainty bands (e.g., 10-90 % or 5-95 %) as a descriptive diagnostic and quantify the empirical coverage of the DANRA reference by corresponding ensemble intervals. The revised manuscript will also explicitly discuss the dry bias, and investigate its sources through a decomposition of, e.g., wet-day occurrence and conditional precipitation intensity, as well as the underdispersion, and present these results as clear limitations. All language across the manuscript, specifically in the abstract, discussion, and conclusion, will be revised accordingly.

R1.3: QM is criticized, but the criticism is not adequately proven

The reviewer notes that the manuscript does not adequately provide evidence for the criticism of Quantile Mapping, and asks for direct quantitative evidence of, for example, event morphology and spatial coherence.

We agree that the claim of “*spectral overfitting*” is currently insufficiently demonstrated, and that the original manuscript infers structural deficiencies in QM without sufficient direct evidence. In the revised manuscript, we will add object-based and event-based comparisons (e.g., SAL and connected-component properties) across QM, ERA5, PMM, ensemble mean, and DANRA. We will revise our interpretation according to these results and remove the term “*spectral overfitting*” unless it can be quantitatively substantiated. If the new diagnostics do not clearly support the current criticism, the criticism will be strongly revised or removed.

R1.4: “Physics-guided” is overstated

The reviewer argues that the implemented conditions and architectural choices (topography, land-sea mask, seasonality, SDF weighting, and RainGate) are physically motivated features and not direct physical constraints.

We agree that the current terminology overstates what is implemented, and we will distinguish the physically motivated or geographically informed conditioning from actual physical constraints. We will revise the terminology throughout the manuscript, with particular attention to the title, abstract, and methodology sections. Further, we will more explicitly state that CEDDAR remains a data-driven statistical emulator.

R1.5: The manuscript organisation makes the methodology and results difficult to follow

The reviewer notes that essential information and diagnostics are split between the main manuscript and the supplement, making the arguments and contents difficult to follow. Further, the results and discussion contain too many narratives, muddling the actual contributions of the work.

We agree that the manuscript would benefit from a restructuring. We will more clearly separate the architecture-selection analysis, evaluation of the final configuration, probabilistic and spatial diagnostics, and the inference-time σ^* investigation. We will restructure the Results section around a much smaller number of scientific questions and move essential information into the main manuscript: a concise EDM formulation, essential preprocessing information, the central architecture-study logic, the complete σ^* formulation and key sampling choices, and the main calibration diagnostics. We will retain implementation and exhaustive diagnostic detail (e.g., hyperparameters and sampling sweeps) in the supplement, and we will consolidate figures and tables rather than simply adding new ones. We will restructure and compress the architecture study around two main findings: (1) conditioning and auxiliary objectives interact nonlinearly, and (2) tail and accumulation metrics are more seed-sensitive than bulk metrics.