Final author comments – Response to Referee 2

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

We thank the referee for the detailed and constructive assessment and for identifying σ* and the component study as useful conceptual contributions. We also appreciate the concerns raised regarding the broader architectural claims, computational scalability, geographical scope, and evaluation framework.
In the revised manuscript, we will narrow or remove unsupported general claims and restructure the manuscript around the component study and the σ* inference-time control, while strengthening the evidence supporting these contributions. The revised Methods will clarify the exact σ* implementation and the inference range examined in this study. We will further expand the deterministic and probabilistic evaluation, move key probabilistic diagnostics from the supplement to the main paper, strengthen the Quantile Mapping comparison with object- and event-based verification, and expand the discussion of computational scalability and geographical scope.
The substantial overlap between the two referee reports also provides a clear basis for revision, particularly regarding evaluation, comparison with QM, terminology, and manuscript organisation. We believe that addressing these points while sharpening the scope of the manuscript will substantially improve both the strength and clarity of the study.

In the following, we will respond to each concern, referring to R2.X as the point X raised by Reviewer 2 (R2).

R2.1: Exclusion of residual learning based on B0

The reviewer correctly observes that residual learning was tested only in the B0 setting, but that the manuscript makes broader statements about its unsuitability. The supplement currently interprets the B0 result as showing that residual prediction "over-constrains" denoising and suppresses stochastic expressiveness.

We agree that the original wording generalised beyond the tested experiment. Residual prediction was evaluated only within the B0 configuration, and in the revised manuscript, we will clarify this limitation. We will further withdraw the general conclusion that residual learning is unsuitable and instead identify residual learning under richer seasonal and geographic conditioning as an untested alternative. A meaningful extension would require matched conditioned configurations and sufficient training-seed replication. Given the seed sensitivity already observed, a single additional run would not resolve the methodological uncertainty, and we argue that a full analysis of the residual architecture is out of the current scope. We will therefore either reinterpret the proposed mechanistic explanations explicitly as hypotheses or remove them where they are not supported by the evidence.

R2.2: Absence of state-of-the-art Deep Learning baseline and insufficiently demonstrated QM comparison

We divide the reviewer's concerns regarding baselines into two distinct issues: (1) the absence of a DL baseline, and (2) the QM comparison.

R2.2a: Deep-learning baseline

The reviewer asks for an established deep-learning (DL) baseline to demonstrate architectural superiority.

We acknowledge that an established DL downscaling architecture baseline would be required to support claims of architectural superiority. This study, however, was not designed as a state-of-the-art architecture benchmark, and we will therefore remove language implying such superiority. Instead, the revised manuscript will restrict its claims to what is directly supported by the experiments: the behaviour of a conditional diffusion framework, the interactions between its investigated components, and the inference-time response associated with σ^*.
Bilinear ERA5 and QM will be retained as transparent baselines testing resolution-only interpolation and marginal-distribution correction, respectively. The absence of a neural baseline will be stated explicitly as a limitation, and no claim of superiority over established deep-learning downscaling methods will be made.

R2.2b: QM comparison

The reviewer points out the lack of evidence for the Quantile Mapping (QM) criticism posed in the original manuscript.

We fully agree that the original manuscript infers structural deficiencies in QM without sufficient direct evidence. In the revised main manuscript, we will add object-based and event-based comparisons (e.g., SAL and connected-component properties) across QM, bilinear ERA5, Probability Matched Mean (PMM), ensemble mean, individual generated members, and DANRA. We will revise the interpretation according to these results and remove the term "s*pectral overfitting*" unless it can be supported by the new analyses and properly quantified. If the additional diagnostics do not support the current criticism, we will revise or remove that interpretation accordingly.

R2.3 Missing scalability and computational cost analysis

The reviewer asks for training and inference cost, explanation of how the model scales to larger domains, and justification of pixel-space training rather than latent diffusion. We will address each of these concerns individually below.

R2.3a Training and inference cost

We agree that computational cost would be a valuable addition to the manuscript. In the revised manuscript, we will report the following: (1) training wall time and/or total GPU-hours

where recoverable, (2) inference time per daily member, (3) representative ensemble-generation cost (32 members), (4) a qualitative or quantitative comparison with the substantially cheaper bilinear and QM baselines, as appropriate.

R2.3b Scaling to larger domains

Although experiments on larger domains are outside the scope of the current work, we will add a discussion of approximate computation scaling and memory constraints. We will discuss first-order computational scaling with spatial domain size, number of sampling steps, and ensemble size, while noting that realised runtime and memory requirements also depend on network architecture, batching, hardware, and distributed execution. Continental-scale use has not been demonstrated and may require tiling, distributed training/inference, or architectural changes.

R2.3c Pixel versus latent diffusion

In the revised manuscript, we will explain why pixel-space diffusion was reasonable for the present regional diagnostic study, while also acknowledging latent diffusion as a potentially more scalable alternative, if the model were to be implemented for larger domains. Pixel-space diffusion was feasible for the present domain and allowed the study to focus directly on the downscaling and sampling behaviour without introducing a separate latent representation. We will, however, not claim any superiority of either approach because this study does not provide a direct comparison.

R2.4: Limitations of physics-guided design in flat topography

The reviewer raises two issues in this comment: (1) the terminology used throughout the manuscript, and (2) the lack of discussion of geographic generalisation. We will address each issue individually.

R2.4a: Terminology

The reviewer notes that the "physics-guided" aspects of the model are not general and largely focused on the geographical aspects of Denmark.

We agree that the present evidence supports geographically informed and physically motivated conditioning within the Danish domain, rather than a generally validated "physics-guided" framework. We will adopt more precise terminology throughout and avoid implying that these conditioning choices impose physical laws or have demonstrated effectiveness across geographical regimes.

R2.4b: Geographic generalisation

The reviewer points out that the tested Denmark-only setting does not establish behaviour in strongly orographic terrain, and that the manuscript needs a discussion of how the geographic conditions translate to mountainous terrains.

We agree that the current work provides no evidence for model performance in substantially different geographical regimes. We will clarify in the revised manuscript that any present evidence is Denmark-specific and that the current work does not establish behaviour under complex orography. We will add a specific limitation on the lack of generalisation and transferability to different geographical domains. We will discuss that application to strongly orographic regions would require retraining and renewed evaluation of the topographic representation, conditioning strength, and SDF formulation rather than direct transfer of the Danish configuration. We will correspondingly narrow any language implying general operational applicability or geographically general validation of CEDDAR as a diagnostic testbed.

R2.5: Deterministic evaluation and probabilistic biases

The reviewer raises two relevant points in the same paragraph, and we address them individually as: (1) lack of deterministic evaluation, and (2) lack of rigorous analysis of probabilistic properties.

R2.5a: Deterministic evaluation

The reviewer argues that the evaluation is based heavily on spatial realism and spectral properties, while omitting deterministic metrics for ensemble summaries.

We agree that the original manuscript currently lacks standard deterministic metrics for ensemble summaries, thus lacking sufficient characterisation of the model when reduced to a deterministic product. We presented only PMM as a deterministic summary, as we intended any ensemble summary to be a secondary product rather than the full generative output. Nevertheless, we agree that adding more standard deterministic metrics would provide important information relevant to many applications. We will therefore, in the revised manuscript, add ensemble mean, median, and PMM, comparing them directly with bilinear ERA5 and QM. We will perform this comparison across a compact metric set, including bias, MAE and/or RMSE, correlation, wet-day occurrence/detection, and heavy-precipitation detection. We will interpret the results accordingly, adjust the manuscript to the results, and we will not imply deterministic superiority unless it is supported by these results.

R2.5b: Probabilistic calibration

The reviewer notes that the probabilistic results exhibit persistent dry biases and underdispersion and asks for more rigorous analysis and standard reporting.

We agree that the current probabilistic behaviour is insufficiently characterised and that the wording in the original manuscript of "*well-calibrated*" ensembles was overstated. To support the probabilistic investigation and ease readability, we will move PIT/rank, reliability, and spread-error analyses into the main text. We will further investigate the dry bias through decomposition into wet-day occurrence and conditional precipitation intensity. We will also quantify empirical interval coverage and show ensemble distributional uncertainty bands (e.g., 10-90 % or 5-95 %) as a descriptive diagnostic. The dry bias and underdispersion will be explicitly treated as limitations, and the manuscript language will be adjusted accordingly.

R2.6: Presentation and main/supplement split

The reviewer comments that the current presentation is often difficult to follow, because essential methodological information and diagnostics are split between the main manuscript and supplement.

We agree that the current organisation is suboptimal, and the revised manuscript will go through a thorough restructure, moving key diagnostics and essential methodology into the main text: concise EDM formulation, essential preprocessing, key sampling choices, central component-study logic, and main probabilistic diagnostics. We will further compress the architecture narrative to focus on only a few main findings, reducing the number of competing narratives. We will further consolidate figures and tables and leave exhaustive implementation and secondary diagnostics not key to the main narrative in the supplement.