Comments on "Pre-training for Deep Statistical Climate Downscaling: Enhancing Consistency and Robustness Across Regional Datasets"

While the previous version emphasized efficiency and generalization of pre-training for station downscaling, the revision shifts the focus to improving cross-dataset consistency and the robustness of learned relationships across national and regional products, rather than maximizing predictive skill for individual station datasets. The introduction of the independent STATIONS-CAT dataset and the expanded interpretability analysis reinforce this emphasis on transferable representations and coherent climate-change signals across diverse observational products. This repositioning clarifies the intended scope of the proposed framework and enhances its relevance for multi-dataset downscaling consistency.

However, this reframing raises an important concern regarding the interpretation of robustness and the practical motivation for transferring a model trained on gridded data to station observations. In statistical downscaling, adapting a model pre-trained on a spatially smoothed gridded dataset is typically motivated by the need to recover localized variability and extremes attenuated by interpolation. While the manuscript argues that fine-tuning improves cross-dataset consistency without degrading overall performance, evaluation of effectiveness for station-scale extremes remains limited, and predictive skill is explicitly de-emphasized. It therefore remains unclear whether the transfer-learning framework preserves or improves the ability to infer local extreme phenomena relative to models trained directly on station data. Clarifying the relationship between representational consistency and predictive effectiveness at the station scale would strengthen the practical relevance of the approach.

If cross-dataset consistency and transferable physical relationships are the central goals, the reliability of the pre-trained model and the physical validity of the learned relationships become critical assumptions. In this respect, the interpretation of the Aggregated Saliency Map (ASM) diagnostics appears preliminary and does not yet demonstrate that the pre-trained model captures the dominant geographic controls governing station-scale climate variability. Minimum temperature and precipitation are strongly modulated by topography, elevation, coastal proximity, and land–sea

contrast, yet the ASM discussion largely focuses on large-scale atmospheric predictors without examining how such geographically mediated effects are represented when training on the smoothed ROCIO-IBEB grid. Because spatial interpolation reduces local gradients and extremes, it remains uncertain whether the pre-trained relationships adequately encode the geographic conditioning present in station observations.

This uncertainty also affects the claimed transferability of the learned relationships. The manuscript implicitly assumes that consistency of large-scale sensitivities in the gridded domain implies feasibility of transfer to station-scale variability and extremes, but this assumption is not yet substantiated by the ASM analysis. Further examination of how geographic controls are expressed in the learned relationships—particularly across regions with strong orographic or coastal gradients—would strengthen the physical basis of the proposed framework. The ASM results in Figures 5 and 9 provide useful initial indications of regime-dependent sensitivities, yet deeper analysis is needed. For example, stratifying ASM by geographic station subsets (elevation, coastal distance, geographic region) could reveal whether the learned relationships remain physically consistent across heterogeneous environments. In addition, Figure 10 shows that full training and full fine-tuning converge to distinct large-scale sensitivities despite identical architectures, suggesting different local minima in the loss landscape. It would therefore be informative to assess whether these models exhibit systematically different behavior for individual events or regional station clusters, particularly for extremes. Such differences could indicate regime-dependent overfitting to localized processes or alternative physically meaningful relationships.

Concretely, the authors could (i) compute ASM separately for stations stratified by elevation, coastal distance, or physiographic region; (ii) compare case-wise errors and saliency patterns between full training and full fine-tuned models for selected extreme events or regional subsets; and (iii) evaluate whether differences between regimes concentrate in stations or periods associated with known local extreme processes (e.g., orographic precipitation, coastal convection, cold-air pooling). These analyses would clarify whether divergence between training strategies reflects physically interpretable regime dependence or optimization variability, and would provide stronger evidence that the pre-trained relationships are both physically

grounded and transferable to station-scale downscaling.

Overall, demonstrating that the pre-trained relationships remain physically interpretable across geographically stratified station subsets and that fine-tuned models preserve or improve skill for station-scale variability and extremes is essential to establish that the proposed framework supports effective station-based downscaling beyond representational consistency. Such evidence would be necessary to align the study with typical expectations for methodological contributions in statistical climate downscaling and to substantiate the current title and scope. Given that these aspects are not yet adequately addressed in the present revision, I recommend rejection of the manuscript in its current form under the stated framing. However, the proposed approach has clear potential, and I strongly encourage the authors to undertake the suggested station-scale performance and physical interpretability analyses and resubmit a more comprehensive and fully validated manuscript. Establishing this link between cross-dataset consistency and reliable inference of localized extremes would substantially strengthen both the scientific credibility and practical relevance of the work for publication in GMD.