We thank the reviewer for the careful and constructive review. Their comments clearly identified key limitations and provided guidance that has helped us strengthen the revised manuscript. Below we address the overall assessment and all major points raised.

This study presents an important exploration of using pre-trained deep learning (DL) models for climate downscaling, aiming to maintain physical consistency between large-scale predictors and localized datasets. By systematically testing multiple training strategies (pre-training, partial fine-tuning, full fine-tuning, and full training), the authors demonstrate the robustness and efficiency of applying pre-trained models on the station-based dataset. However, as the authors note in the discussion, "this benefit does not necessarily translate into improved accuracy on STATIONS-IBEB, likely due to the presence of higher and more localized extreme values, which are more challenging to model than their smoothed counterparts in the interpolated ROCIO-IBEB gridded dataset." This observation raises a critical issue: while pre-training improves efficiency and generalization, it may limit the model's ability to capture localized extremes that define station-based observations. Clarifying this trade-off would deepen the study's insight into how pre-trained DL models balance physical consistency and predictive reliability in downscaling applications. The following comments aim to clarify and deepen several aspects of this discussion.

We thank the reviewer for highlighting this important trade-off. We agree that pre-training may not necessarily translate into improved skill on STATIONS-IBEB, and clarifying this balance is central. In our study, the primary objective is not to maximize accuracy, but to assess whether pre-training can promote consistency across products and align the learned large-scale dependencies of models trained on different (often sparse) observational targets (we acknowledge that this objective was not stated clearly enough in the current manuscript). Importantly, our results indicate that fine-tuning delivers these benefits without degrading overall skill compared to training from scratch, while improving training efficiency and robustness. We will revise the manuscript to make these objectives explicit and to better frame the consistency-accuracy trade-off in the discussion.

The key distinction between ROCIO-IBEB and STATION-IBEB lies in their treatment of local extremes. Since station-based datasets inherently preserve localized weather phenomena, it would be helpful to elaborate on the rationale for using STATION-IBEB as the downscaling target and to contrast its statistical characteristics—particularly the distribution tails representing extreme events—with those of ROCIO-IBEB. This clarification would highlight the physical implications of transferring knowledge between datasets with distinct spatial and statistical properties.

We agree. In the revised manuscript we will strengthen the motivation for using STATIONS-IBEB as the target: unlike the gridded ROCIO-IBEB product, the station network preserves local variability and extremes, which is precisely the challenging regime we want to test when moving from a "core" product to point-scale observations. We will apply the same characterization to the additional independent station dataset we will introduce in the revision.

Although fine-tuned models converge faster and achieve comparable performance to fully-trained models in terms of RMSE and mean bias, Figure 4 (right column) suggests that fully-trained models perform slightly better for extreme metrics such as TXx and TNn. This raises an important question about the ability of pre-trained models to represent localized extremes, which are critical for reliable high-impact weather downscaling. A focused evaluation of model skill over the extreme subsets of both ROCIO-IBEB and STATION-IBEB would help determine whether performance limitations stem from the coarse representation of extremes in the pre-training data or from the fine-tuning process itself, which may not fully adapt to station-scale variability.

We thank the reviewer for highlighting this important point. We agree that, in principle, fine-tuning from a model pre-trained on a smoother gridded product could limit the ability to fully adapt to the more localized extremes present in station-based targets such as STATIONS-IBEB, and the current results may be an early indication of this trade-off. In the revised manuscript we will examine this more directly by expanding the evaluation to focus on extreme indices and/or extreme subsets and leveraging the additional independent station-based dataset to assess whether any limitation in representing extremes is systematic across station targets. If present, we will incorporate these findings into the discussion as a clearly stated limitation and as guidance for future extensions of the approach.

The aggregated saliency map results reveal differences between full-training and pre-trained models, yet it is unclear whether these reflect meaningful large-scale dependencies capable of inferring local extremes or potential overfitting to dominant features. Providing examples of regional or event-specific saliency maps, rather than only aggregated values, would clarify whether the learned features correspond to physically interpretable meteorological patterns or spurious correlations introduced during training.

We agree with the reviewer. In the current manuscript we mainly report aggregated saliency statistics, and while the resulting predictor-importance ranking is consistent with what has been reported in related downscaling/XAI studies [1], this does not by itself demonstrate that the learned dependencies are physically meaningful at the event/regional scale. In the revised manuscript we will consider adding examples of regional/event-specific saliency maps to assess whether the highlighted patterns are meteorologically interpretable rather than driven by dominant but spurious features. We will also extend the same XAI analysis to the additional independent station dataset introduced in the revision, to examine whether fine-tuning consistently promotes alignment of large-scale dependencies across different station targets (as motivated in our responses above).

[1] González‐Abad, J., Baño‐Medina, J., & Gutiérrez, J. M. (2023). Using explainability to inform statistical downscaling based on deep learning beyond standard validation approaches. Journal of Advances in Modeling Earth Systems, 15(11), e2023MS003641.

In Section 4.4, fine-tuned models trained with datasets containing varying fractions of missing data show lower RMSE values, attributed to pre-learned representations improving generalization. However, such pre-training could potentially smooth out

localized extremes in unseen data. Evaluating performance specifically under extreme conditions in these incomplete datasets would strengthen the interpretation of how pre-training affects robustness and physical fidelity when data coverage is limited.

We agree. In the revised manuscript we will extend the analysis in Section 4.4 to better account for extreme behavior under missing data, in line with the additional extremes-focused evaluation described in our response to a previous comment. This will help clarify whether the robustness gains we observe with pre-training are achieved without compromising the representation of localized extremes when data coverage is limited.

Overall, this study makes a valuable contribution to understanding how pre-trained DL models can be adapted for regional climate applications. Further analysis focusing on extreme events and saliency-based interpretability would enhance confidence in the approach and clarify the trade-offs between maintaining physical consistency and capturing localized, high-impact weather phenomena.