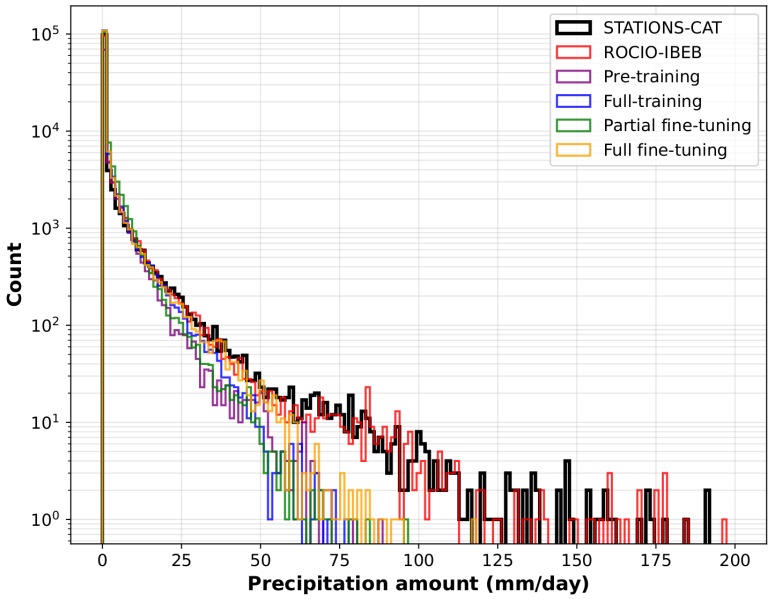
# Review 1

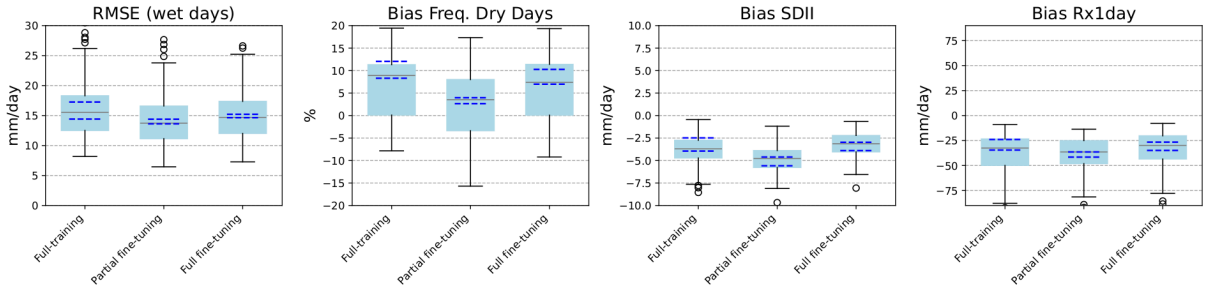
*We thank the reviewer for the detailed and constructive assessment. In the following, we address each concern through new experiments directly aligned with the reviewer's suggestions, focusing on precipitation given its more diverse and informative saliency patterns compared to temperature for which local-relationships tend to dominate [1].*

**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.**

*We agree with the reviewer that a necessary condition for practical usefulness is that fine-tuning should not degrade performance relative to training directly on station data. For STATIONS-IBEB, this is already demonstrated in Figures 3 and 4 of the manuscript, where fine-tuned models show results comparable to the fully-trained model across mean and extreme-related metrics. To address this concern for the regional dataset, we provide two additional figures for STATIONS-CAT that we show following this paragraph. The first one shows the distribution of extreme precipitation values (log-scale y-axis) comparing observations (STATIONS-CAT), the nearest ROCIO-IBEB grid points, the pre-trained model, and the three training regimes (in the STATIONS-CAT test set). The second figure presents the equivalent of Figure 4 for STATIONS-CAT (RMSE over wet days, bias in the frequency of dry days, SDII, and Rx1day). These results confirm that fine-tuned models preserve station-scale skill. Some underestimation of extremes is apparent across all regimes, but this is a known limitation of deterministic loss functions, not specific to fine-tuning [2].*




**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.**

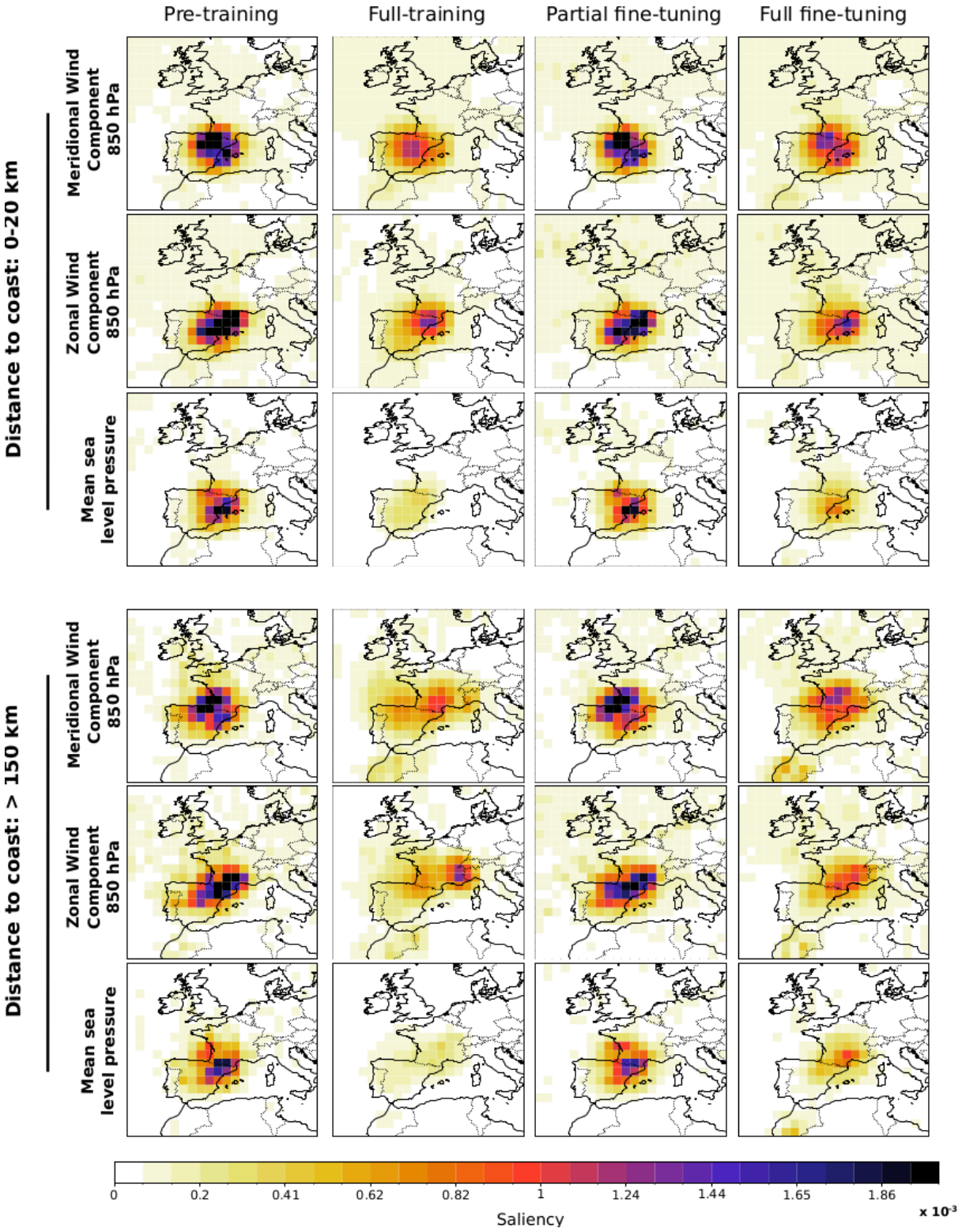
*We have conducted all suggested analyses. We detail the results below.*

*(i) Saliency maps stratified by coastal distance and elevation*

*We compute saliency maps (analogous to Figure 11 of the updated manuscript) for the STATIONS-CAT dataset stratified by coastal distance and by elevation, thus addressing the reviewer's request for stratified ASM diagnostics and for examination of how geographic controls are expressed in the learned relationships. We focus on the meridional and zonal wind components at 850 hPa and mean sea level pressure, some of the most informative predictors based on Figure 10 of the updated manuscript.*

*Following this paragraph we show the equivalent of Figure 10 of the updated manuscript but for stations (from STATIONS-CAT) close to the coast (< 20 km, top) and far from the coast (> 150 km, bottom). The saliency is aggregated over the corresponding stations in each group. For coastal stations, all models (including the full-trained one) concentrate saliency over the coastal region for all three variables, which is consistent with the dominant role of the Mediterranean in driving coastal precipitation. For inland stations, the picture changes. In the pre-trained and fine-tuned models, the meridional wind component at 850 hPa shows a saliency displacement beyond the inland station locations toward the Bay of Biscay region. This displacement exceeds the geographic shift of the stations themselves, as confirmed by the zonal wind and mean sea level pressure variables, where saliency shifts only modestly and in proportion to the station displacement. This indicates that the model is capturing a physical signal: for inland Catalan stations, meridional flow from the northwest (channeled*

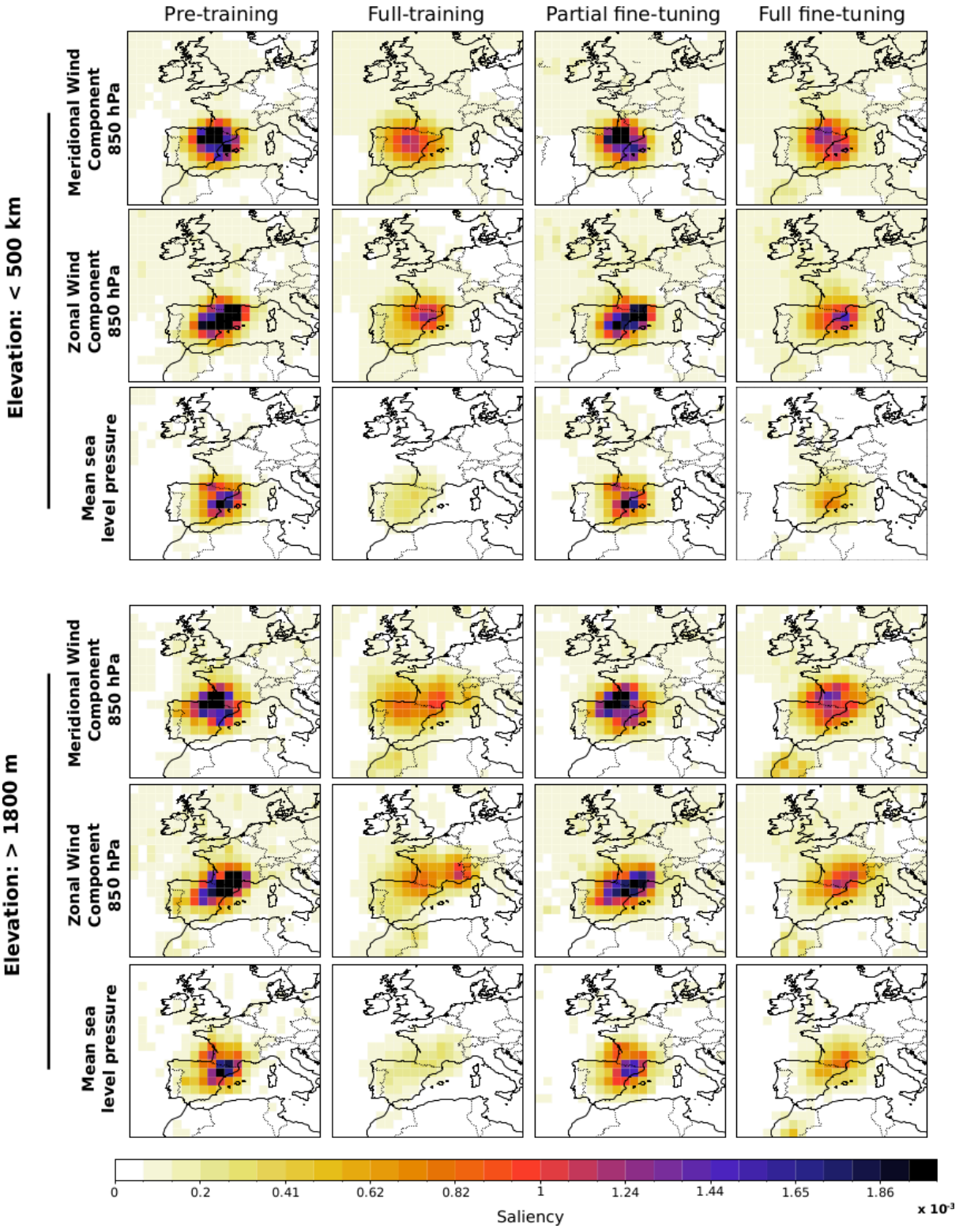
*through the Bay of Biscay) plays a distinct role in precipitation that is not equally important for coastal stations. This behavior is not observed for the full-training model, which instead shows saliency for the zonal wind component extending into western France, Italy, and Switzerland (regions with no direct physical link to daily precipitation in Catalonia).*




*In the next figure we show the same analysis but for stations with low elevation (< 500 m) and high elevation (> 1800 m). The results closely mirror those of the previous figure, including the physically misaligned saliency of the full-training model and the focus of the*

*pre-trained and fine-tuned models on the Bay of Biscay region for the high-elevation stations. This similarity is expected: in Catalonia, inland and high-elevation stations largely overlap geographically, as the highest stations are located in the Pyrenees, far from the coast. Since the DeepESD architecture receives only large-scale atmospheric fields and does not incorporate local co-variables such as elevation or distance to coast, the model cannot differentiate between these two station properties and instead learns a common large-scale pattern relevant for both. In this case, the model identifies the Bay of Biscay as an important region for these stations, which is physically coherent: northwesterly flow channeled through this area is a moisture pathway for precipitation events in the Pyrenees and interior Catalonia. The full-training model, by contrast, relies on spatially dispersed saliency reaching remote areas, suggesting spurious correlations that may fit the limited training data but lack a physically plausible link to precipitation at these stations.*

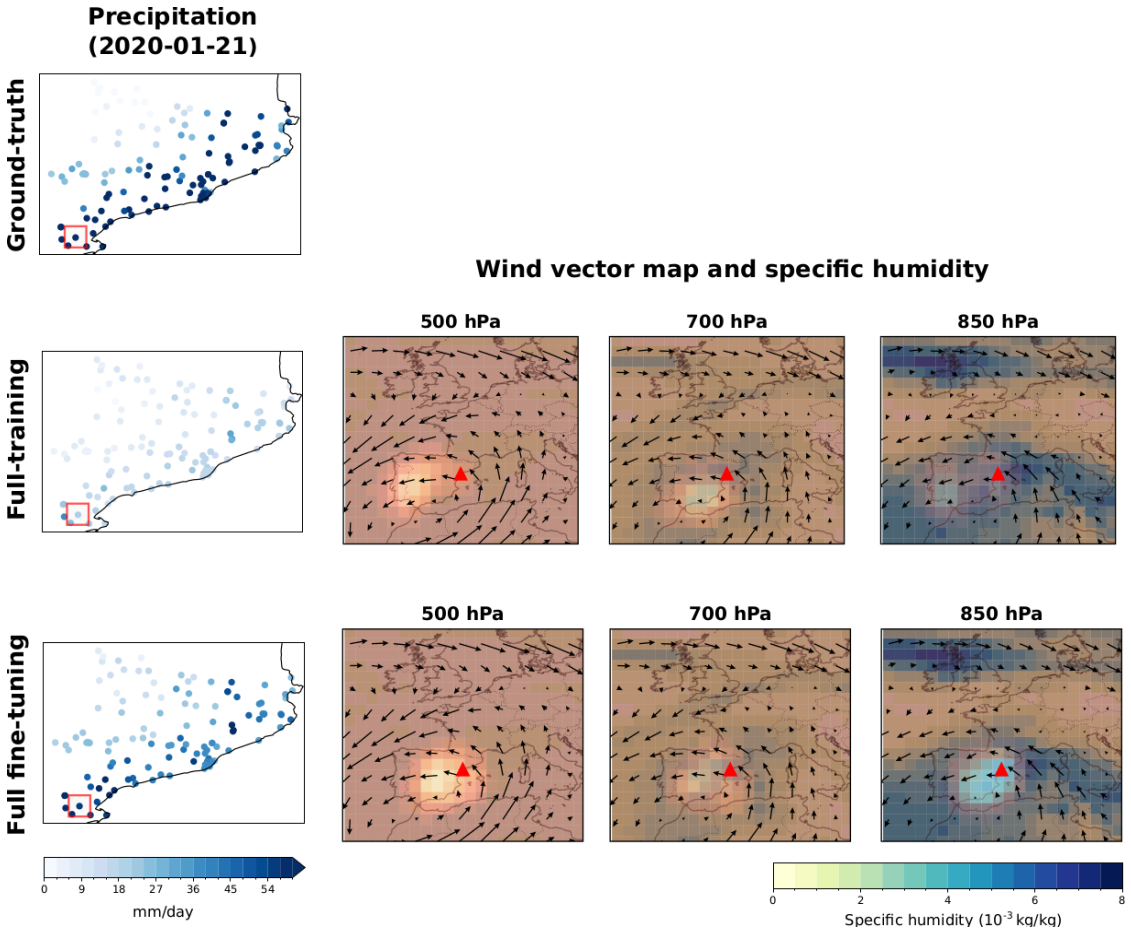
*This lack of co-variable information has not hindered the performance of the DeepESD model across diverse regions in the literature (see references in the manuscript), however, incorporating local geographic co-variables is an active area of development in our group and a natural extension of the DeepESD architecture. We appreciate the reviewer for raising this point, as it motivates a clear direction for future improvements.*




*(ii) Case-wise saliency for extreme events*

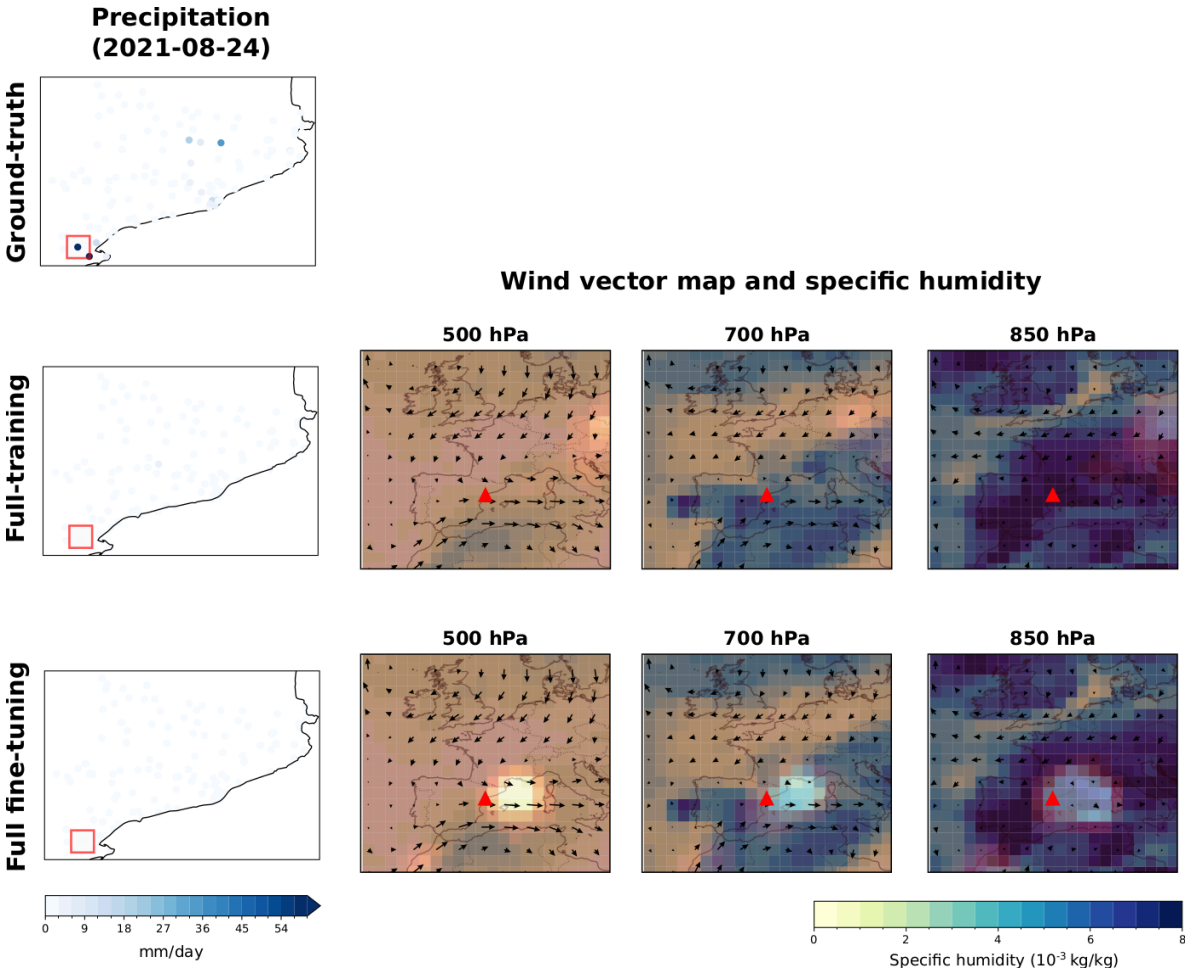
*We compare predictions and saliency maps from the full-training and full fine-tuning models (following the reviewer's suggestion) for two different extreme precipitation events. These events were selected as they exceed the 99th percentile for the analyzed station.*

*In the following figure we show the observed (ground-truth) precipitation field and the precipitation downscaled by the full-training and full fine-tuning models (first column) for 2020-01-21 (test set). This day corresponds to a widespread precipitation event affecting the broader Catalonia region. For each model, we show the wind vector maps (arrows) at 500, 700, and 850 hPa computed from the zonal and meridional wind components, with specific humidity displayed in blue in the background. The highlighted (brighter) regions indicate the areas most influential for the model when downscaling precipitation at a specific station (marked by a red box in the precipitation maps and a red triangle in the saliency maps). This highlighted region corresponds to the aggregated saliency from the wind components used to compute the wind vectors. Both models underestimate extreme precipitation, but the fine-tuned model shows less underestimation. More importantly, the full-training model attributes relevance to a region where the wind field does not point toward the target station, a spatial misalignment with the expected synoptic forcing. The fine-tuned model concentrates saliency near the station, with wind vectors oriented toward it, likely consistent with moisture advection driving the event.*




*In the following figure we show the same analysis but for the day 2021-08-24. This event is confined to the target station and not observed regionally, characteristic of convective*

*precipitation. As expected, neither model captures the event, since isolated convective processes arise at scales finer than the large-scale predictor fields. However, the saliency patterns are instructive. The full-training model exhibits highly dispersed, nearly unstructured saliency across the domain (hence the absence of a clear bright region). Even at 850 hPa, it shows a localized bright region in the northeastern part of the domain, whereas the fine-tuned model maintains a more concentrated pattern near the station across all levels. This indicates that even when both models fail, fine-tuning preserves physically plausible spatial saliency rather than collapsing into dispersed, uninformative patterns.*




*(iii) Station-scale performance for extremes*

*As discussed above in our response to the reviewer's first concern, we demonstrate that differences between training regimes do not concentrate on degraded skill for extreme-related metrics. The fine-tuned models match the full-training model across all evaluated indices (RMSE, dry-day frequency bias, SDII, Rx1day) for STATIONS-CAT, confirming that the pre-training framework preserves the performance to predict local phenomena.*

**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.**

*We believe the analyses presented above directly address the suggestions made by the reviewer. In summary:*

- *Saliency maps stratified by coastal distance and elevation show that fine-tuned models maintain more localized and physically coherent relationships than models trained from scratch on STATIONS-CAT, with the full-training model exhibiting spurious saliency to remote regions.*
- *Case-wise saliency for extreme events confirms that fine-tuning tends to tie the model to physically plausible relationships, even when predictions fail.*
- *Fine-tuning does not degrade station-scale performance, including extreme-related metrics.*

*We believe these results collectively establish the link between cross-dataset consistency and reliable downscaling that the reviewer identified as essential. The stratified saliency analyses demonstrate that the pre-trained relationships remain physically interpretable across geographically heterogeneous station subsets, while the performance evaluation and case-wise comparisons confirm that fine-tuned models preserve station-scale performance and exhibit physically-based behavior for extremes. Together, this evidence supports that the proposed framework enables effective station-based downscaling beyond representational consistency alone.*

*We would like to acknowledge that the analyses presented here approach the practical limits of current XAI techniques: saliency-based methods quantify input sensitivity but do not provide causal explanations, and precise interpretation of what a model understands remains an open challenge. We appreciate the reviewer for pushing us to work at the frontier of these methods. However, we would like to note that further interpretation of the model's internal representations would require a newer generation of interpretability techniques, most of which are still active areas of research and would fall outside the scope of this work, effectively requiring an almost entirely new study to address them.*

*To close, we summarize the changes introduced in the manuscript in light of this response. We have added the following:*

- *The histogram of the precipitation distribution and the corresponding discussion, confirming that fine-tuning preserves station-scale predictive performance for extremes in STATIONS-CAT.*

- *The figure of saliency maps stratified by distance to the coast and its analysis, highlighting the physically coherent shift in spatial saliency between coastal and inland station groups and the contrasting behavior of the full-training model. For the sake of clarity, and given the already large number of figures in the manuscript, the equivalent figure stratified by elevation has not been included.*

- *The figure of case-wise saliency maps for the extreme event of 2020-01-21 and its analysis. The results from the localized event are briefly discussed in the text without showing the figure.*

- *Expanded discussion and conclusions sections reflecting the analyses developed in this response.*

*[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.*

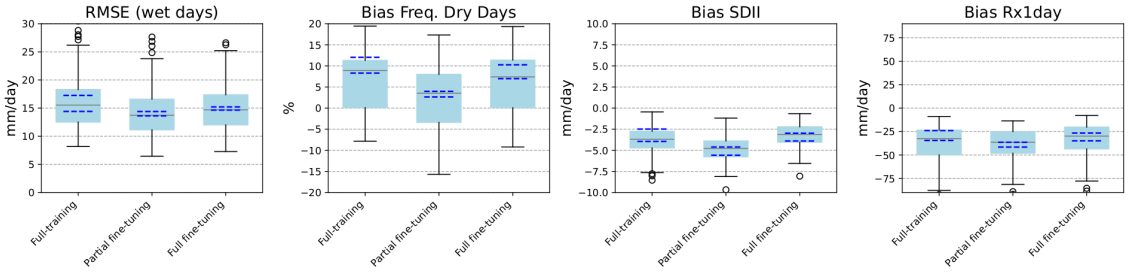
*[2] González-Abad, J., & Gutiérrez, J. M. (2025). Are deep learning methods suitable for downscaling global climate projections? An intercomparison for temperature and precipitation over Spain. Artificial Intelligence for the Earth Systems, 4(4), 240121.*

# Review 2

**I would like to thank the authors for including results for the additional dataset, STATIONS-CAT. The revised manuscript addresses most of my initial concerns, and I am happy to recommend it for publication pending a few minor remaining points. Please note that all line numbers in the following comments refer to the manuscript with tracked changes.**

**1. In the discussion (L433), the authors conclude that "models with limited data availability may fit noise and learn relationships that are less aligned with known dynamics." If this were the case for STATIONS-CAT, the fully trained model would likely exhibit higher RMSEs or biases in the test set compared to the pre-trained models. It would strengthen the manuscript if the authors could support this hypothesis with a result for STATIONS-CAT similar to Figures 3, 4, or 8 (adding just one of these would suffice).**

*We thank the reviewer for this suggestion, as properly supporting such claims is important in the context of this work. In the following, we provide a figure showing the equivalent of Figure 4 (conclusions are similar for temperature) but for STATIONS-CAT (RMSE over wet days, bias in the frequency of dry days, SDII, and Rx1day).*




*Focusing on the RMSE, the partial and full fine-tuned models tend to achieve slightly lower values compared to the fully-trained model. This trend is consistent for the bias in the frequency of dry days and the bias of Rx1day, though in these cases only the partially fine-tuned model outperforms the fully-trained one, suggesting that inheriting the pre-trained relationships from ROCIO-IBEB helps the model learn more physically grounded associations. For the SDII, however, the best result is obtained by the full fine-tuning regime, indicating that certain aspects of precipitation downscaling may benefit from allowing greater freedom in parameter updating.*

*We argue that the learning of physically less grounded relationships in this regime is already demonstrated by the saliency-based analysis in the manuscript, which has been further expanded in this revision in response to comments from reviewer 1. While we agree that including this figure would add clarity, given the already large number of figures in the manuscript we believe the text benefits from keeping the saliency analysis as the primary evidence for this claim, as it provides a more direct and interpretable demonstration of the underlying relationships. Nonetheless, we have added a sentence in the discussion noting that this is also reflected in the performance metrics:*

*"[...] This is also reflected in standard metrics (not shown), such as those presented in Figures 3 and 4, where fine-tuned models tend to achieve slightly lower errors and biases."*

**2. In Figure 6, it is currently difficult to extract meaningful information from the temperature fields because the color patterns across all models appear nearly identical. This is problematic, as Figure 6 provides crucial insight into whether pretraining helps constrain climate change signals more consistently. I recommend reducing the temperature range in the colorbars for TNn and TXx to make the differences between models more discernible. For instance, ranges of 1 to 4 °C for TNn and 4 to 8 °C for TXx seem to capture most of the variability and would likely provide much better contrast.**

*We agree with the reviewer that this figure could benefit from improved colorbar design. Accordingly, we have updated it following the reviewer's advice. We believe that, with these changes (and those related to the comment below), the figure now communicates its message more clearly.*

**3. Some qualitative statements in the text would be more compelling if backed by quantitative analysis. For example, regarding the reduced extreme precipitation in northern Spain or the changes in the Duero River (L324): could relative errors or biases be provided to support these observations?**

*We agree with the reviewer that some statements regarding differences in the climate change signal between models would benefit from more quantitative support. Accordingly, following the reviewer's suggestion, we have added an extra row to Figure 6 for each index, showing the differences between the climate change signal of the pre-trained model and the three training regimes. This addition helps support and visualize the statements made, for instance, for RX1day, where the underestimation and overestimation over the Duero River basin by the full-training and full fine-tuning models, respectively, can be clearly observed. We also report the spatial mean of the absolute differences for these maps to provide a clearer quantitative summary of the overall discrepancies.*

**Minor comments:**

**L104: "as well as global high-res grids" -> This phrase should probably be deleted since the corresponding reference has been removed.**

*We agree with the reviewer and have removed the term "global", as we are referring to high-resolution grids over specific sub-domains.*

**L386: I believe the reference here should be to fine-tuning on STATIONS-CAT, not STATIONS-IBEB.**

*Yes, we have corrected this in the revised version of the manuscript.*

**L392: Should "Fig 5" be "Fig 10"?**

*Yes, we have corrected it.*