Review 1

We thank the reviewer for the thorough and insightful review. As previously mentioned, the referee provided concrete suggestions that have been very helpful in developing the revised paper. Below we respond to the overall assessment and all major comments regarding the changes introduced in the revised version of the manuscript.

In this study, the authors aim to explore the benefits of transfer learning in the context of ML-based statistical downscaling for climate applications. Specifically, the goal is to understand the benefits of pre-training the latent representations of ML-based downscaling models on a core dataset, to achieve improved skill or higher consistency when the downscaling model is fine-tuned for a different downscaling task.

The reviewer provides an accurate overview of this work (particularly highlighting the objective to achieve consistency across downscaling methods for different tasks), although the motivation and main objective were not sufficiently clear in the original manuscript. We have therefore revised the title and introduction to better articulate the focus of the work, namely assessing the benefits of transfer learning as a means to increase the consistency and credibility of regional climate projections produced by downscaling models trained on different observational datasets (e.g. national-scale versus regional datasets).

Statistical downscaling models rely on the availability of observational datasets (either gridded or station-based) over the region of interest for training. In practice, national-scale and regional-scale downscaling models are often developed independently using datasets of differing quality and characteristics, with regional datasets typically offering higher spatial resolution but shorter temporal coverage. As a consequence, users are frequently confronted with multiple, and sometimes conflicting, sets of regional climate projections which can undermine confidence in their use for decision-making. In this study, we demonstrate that models trained on long, high-quality observational records can be effectively transferred to downscale climate information over a different regional observational dataset (usually of shorter duration) while preserving the large-scale features identified as relevant predictors while fine-tuning the results to the new observational reference.

This topic is certainly of practical and scientific interest for climate modelers, and a good fit for the journal. However, I find the implementation of this study to be largely uninformative regarding the questions posed by the authors in the abstract and introduction. In my view, this is due to the choices made by the authors for the pre-training and fine-tuning datasets, as well as the forms of fine-tuning explored for the DeepESD model. For these reasons, which I detail below, I find this manuscript in its current form unsuitable for publication. I encourage the authors to find a more effective lens through which the questions that they set out to study can be answered.

We have clarified the motivation and objectives. We refer the reviewer to the diff file for the exact modifications introduced. Please note that we have also expanded the Discussion and Conclusions sections to further elaborate on these points.

Regarding the choice of the pre-training and fine-tuning datasets we have addressed these issues and included the corresponding clarifications in our responses to the major comments below.

Major comments:

- Transfer learning is a well-established way to fine-tune large ML models, pre-trained on an extensive pre-existing dataset, on a smaller dataset that is more representative of some final task. The manuscript instead explores pre-training roughly 17,000 parameters of the final DeepESD models, out of a total of ~4.4M parameters for temperature and ~7.5M parameters for precipitation, respectively. This can hardly be called fine-tuning, when the pre-trained parameters represent less than 1% of the total number of parameters in both cases. This partly explains why all the variants yield statistically equivalent results (Figures 4-7), and why no conclusions can be drawn from this experimental setup.

We thank the reviewer for this comment. Following the response we included in the previous round of responses, we have now included in the manuscript, more precisely in the Pre-training and Fine-tuning section, a mention to why, in the case of DeepESD, more important than the number of parameters is the role they play within the model. Following this, the calibrator is a linear regression (per grid point) fitted over the relationship learned by the feature extractor:

"Although the feature extractor contains a relatively small fraction of the total model parameters, parameter count alone does not reflect functional importance. The calibrator's large parameter count arises from its scaling with the number of output locations, but functionally it performs location-dependent linear mappings from the learned representations. In contrast, the feature extractor learns the nonlinear representations that determine what information is available to the final layer. In addition, prior work shows that transferability depends strongly on layer role/depth, with earlier representation layers often being more transferable than later task-specific layers."

In addition to this, we have emphasized in the Discussion section this fact by saying that, when only the calibrator is fine-tuned, the spread across replicas is the lowest, thus emphasizing that the source of variation relies on the feature extractor, despite its lower amount of parameters with respect to the calibrator:

"[...] the spread across replicas becomes nearly zero when the feature extractor is frozen, demonstrating that, once representations are fixed, tuning the calibrator has minimal impact on variability."

In addition, following the extended analysis using XAI techniques suggested by other reviewer, we show that when only the calibrator is fine-tuned while the feature extractor is kept fixed (partial fine-tuning), the learned relationships remain largely unchanged (see Figure 10 in the revised manuscript). This further supports the previous point that, despite representing a substantially smaller fraction of the total parameters, the feature extractor plays the dominant role in shaping the model's learned relationships.

- The final target dataset of interest, STATIONS-IBEB, is used to construct the pre-training dataset, ROCIO-IBEB. This setup omits the most important practical aspect of transfer learning: can we train on one dataset to improve predictive skill on a different dataset? This is an important question for some of the applications cited by the authors: pre-train on a national-level dataset, and fine-tune on a local dataset with different statistics (Taboada et al, 2024). The current setup is too idealized and not representative of the situations where transfer learning may actually be useful, in my opinion.

We thank the reviewer for this important point. We emphasize that the primary objective of our work is not to maximize predictive skill on individual datasets, but rather to improve consistency and robustness across regional downscaling products, particularly when working with datasets that have limited spatial and/or temporal coverage. As clarified throughout the revised manuscript, the proposed methodology aims to ensure that regional models inherit physically meaningful relationships from a reliable national-scale baseline, thereby reducing epistemic uncertainty.

We acknowledge the reviewer's concern regarding the relationship between ROCIO-IBEB and STATIONS-IBEB. To address this limitation and demonstrate the practical applicability of our approach, we have added STATIONS-CAT as an independent regional dataset:

"STATIONS-CAT provides an independent validation of our fine-tuning approach, as these stations were not used in the construction of ROCIO-IBEB, unlike STATIONS-IBEB"

This dataset, based on the daily blended station data from the European Climate Assessment & Dataset (ECA&D), covers only Catalonia, contains around 110 stations, spans a shorter temporal period (2009-2021), and is not used in the construction of ROCIO-IBEB. Section 4.5 presents a comprehensive analysis showing that fine-tuning successfully transfers physically consistent relationships to this independent regional dataset, produces more robust climate change signals, and avoids the spurious patterns learned when training exclusively on limited regional data. These results directly address the practical scenario outlined by the reviewer: adapting a national-scale model to an independent regional dataset.

- Figure 3 shows that pre-training does not yield improved results, only faster training for a relatively inexpensive model where training cost is not really an issue. The rest of the results also fail to show any positive effects of pre-training. I think this is because at the level of the high-level representations of the data learned by the convolutional layers, the datasets ROCIO-IBEB and STATIONS-IBEB are largely indistinguishable (since they share the same data sources). This leads me to believe that the improved skill of fine-tuned models when the STATIONS-IBEB dataset is artificially shrunk (Fig 9) is due to the fact that you are actually showing a very similar version of the omitted samples to the model through ROCIO-IBEB. This is another example of why useful conclusions cannot be drawn given the similarity of the two datasets considered.

We thank the reviewer for this valuable feedback. In response, we have removed the original Figure 3 and its associated analysis of convergence speed, as we agree this aspect is not the most critical contribution of our work.

We appreciate the reviewer's concern regarding the similarity between ROCIO-IBEB and STATIONS-IBEB. To address this important point and demonstrate the practical applicability of our approach in scenarios where datasets are truly independent, we have added STATIONS-CAT as an additional case study. This dataset is completely independent of ROCIO-IBEB construction, covers only Catalonia with 110-114 stations, and spans a shorter temporal period (2009-2021) representing the realistic operational scenario where pre-training would be most beneficial.

The results on this independent dataset provide clear evidence of pre-training benefits:

- *Figure 9 shows that the ASM diagnostic reveals substantial deviations in learned relationships for the full-training model, while fine-tuned models preserve patterns aligned with the pre-trained baseline.*
- *Figure 10 demonstrates that full-training models exhibit different spatial distributions of relevance (focusing on the central Iberian Peninsula rather than the expected Mediterranean region), while fine-tuned models maintain physically plausible attention patterns consistent with known precipitation dynamics over Catalonia.*
- *Figure 11 shows that fine-tuning substantially reduces variability in climate change signals across training replicas compared to training from scratch.*

These results address the reviewer's concern by demonstrating that when transferring to a truly independent regional dataset with limited coverage, pre-training provides tangible benefits: improved physical consistency, enhanced robustness and avoidance of spurious patterns that emerge when training exclusively on limited data.

Minor comments:

- L38-40: "Diverging outcomes, which may confuse users". Improving consistency at the expense of capturing the true uncertainty of regional climate projections is actually a disservice to the downstream users, because it leads to biased estimates of risk.

We agree with the reviewer that diversity may at times be attributable to the inherent uncertainty of regional climate projections. In this instance, however, we use the term to describe diversity resulting from limitations in the construction of certain datasets, including restricted temporal or spatial coverage.. We have rephrased the sentence to clarify this point:

“Such methodological diversity, arising from limitations in the techniques used to construct these datasets or from restricted temporal or spatial coverage, can lead to conflicting outcomes that may confuse end users.”

- L82: Calling a 1km resolution downscaled dataset spanning thousands of years and supported by extremely sparse observations is a stretch (Karger et al, 2023).

We agree with the reviewer and have therefore removed the reference to this dataset from the revised manuscript.

- L123: "the most widely used in the downscaling literature". I don't think this method is that well established (60 studies reference it), so this needs to be toned down. There are studies from 2024 on downscaling with more references (e.g., CorrDiff), and deterministic ML-based downscaling models are not representative of the state of the art anymore.

We agree that DeepESD is not state-of-the-art. However, it remains the most widely used model for generating operational GCM-based climate projections, despite newer models showing superior performance in research settings. The model's simplicity also facilitates the XAI analysis central to our transfer learning study. We have revised the manuscript to tone down this claim while highlighting DeepESD's widespread operational adoption:

"We select the DeepESD architecture as the basis for the standard DL model. This choice is motivated by several factors. First, while not representing the current state-of-the-art in ML-based downscaling, this model is among the most widely adopted for generating climate projections from GCMs, having been applied to various regions [...]"

- The version of DeepESD used in this paper is different than the one introduced by Bano-Medina et al (2022) for temperature. The MSE loss assumes a homogeneous uncertainty estimate, unlike the original Gaussian log-likelihood where the variance is explicitly modeled. I would also say that the deterministic MSE is no longer a "widely adopted" loss in downscaling due to its tendency to smooth fields in space and underestimate extremes.

This is correct: in Bano-Medina et al. (2022), the DeepESD model for temperature was trained using the negative log-likelihood of a Gaussian distribution. However, we follow more recent studies, such as [1], which compared different loss functions for temperature and concluded that the MSE works well enough for downscaling climate projections. Moreover, models that predict independent probability distributions can encounter additional issues, such as spatially inconsistent projections when sampling, due to the independence between distributions. Previous results, using XAI techniques, also showed that for temperature, the relationships learned for the standard deviation parameter of the modelled Gaussian distribution did not correspond to any dynamical patterns, resembling mostly white noise [2].

[1] 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.

[2] González-Abad, J. (2024). Towards explainable and physically-based deep learning statistical downscaling methods.

- Fig 2: The legend reads "full-tuning" for the right column It should be full fine-tuning.

Yes, we have made the correction.

- Fig 3: Is this the training loss or the validation loss? If the former, please change to the validation loss, which is more representative of operational skill. Otherwise, please show both.

This figure has been removed from the revised manuscript. Please refer to our response to one of the major comments for further details.

- L212: "take about half the number of epochs": Certainly not to reach the best final skill, since the fully trained model is better. How are you defining a common final time for all models to assert this?

See the comment above.

- The results shown in Figures 4 and 5 for the ROCIO-IBEB dataset are not comparable to those in the STATIONS-IBEB dataset, since the former is a smoothed interpolation of the latter. Errors on the ROCIO-IBEB dataset will always be lower.
Yes, we agree with the reviewer. We considered removing the ROCIO-IBEB results from these figures, however, we believe they are useful to illustrate the differences between the two datasets. To avoid potential misinterpretation, we have added clarifying statements in this section emphasizing the lack of direct comparability between the results from the two datasets:

"[...] results on ROCIO-IBEB are shown for reference to illustrate the characteristics of this gridded dataset and are consistent with previous studies. However, these results are not directly comparable to the STATIONS-IBEB models, as they involve different predictand datasets with distinct spatial characteristics."

- Figure 9: I think the legend should refer to different versions of STATIONS-IBEB, not ROCIO-IBEB.

Yes, you are right. We have corrected it.

- Discussion: I do not know where the study demonstrates "the potential of pre-training" (L303), or the affirmation that "fine-tuning the extractor appears to be beneficial". Beyond Figure 9, which has some issues I raised before, the other results are largely equivalent for all variants.

We hope this comment has been adequately addressed in our responses to some of the major comments correctly raised by the reviewer.

Review 2

We thank the reviewer for their thorough and insightful review. Below, we address the overall assessment and all major comments on the revisions to the manuscript.

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 this important comment. As emphasized in the revised manuscript, the primary objective of this study is not to improve predictive skill for individual datasets, but to enhance consistency and robustness across national and regional downscaling products, particularly in low-data and heterogeneous observational settings.

That said, we agree that improvements in efficiency and generalization should not come at the expense of degraded performance, especially for localized extremes characteristic of station-based observations. Our results indicate that this trade-off does not occur in practice. For the STATIONS-IBEB dataset, Figures 3 and 4 (in the revised manuscript) show that fine-tuning preserves performance across all evaluated metrics, including those related to extremes (TNn, TXx, Rx1day), while improving robustness and consistency with the pre-trained national-scale model.

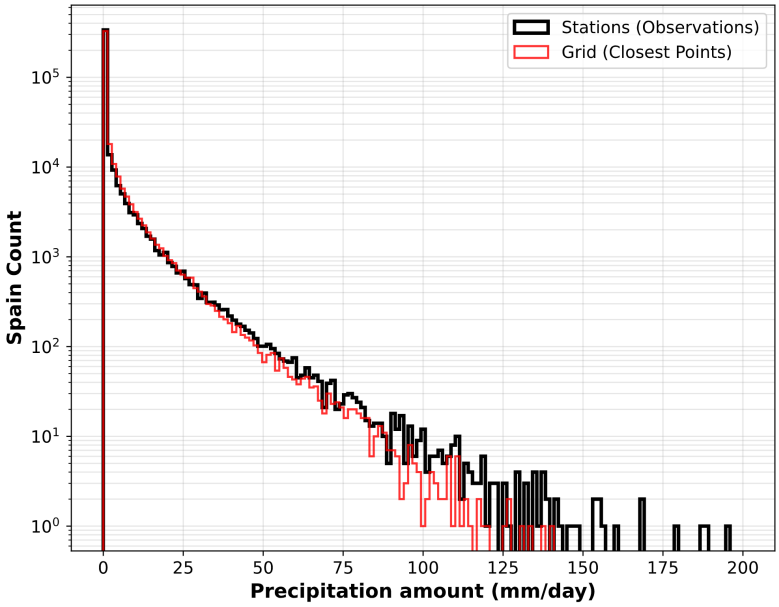
A similar behavior is observed for the independent STATIONS-CAT dataset introduced in this revision: fine-tuning does not lead to a loss of skill relative to training from scratch, while enhancing robustness. For conciseness, we did not include figures analogous to Figures 3 and 4 for STATIONS-CAT in the manuscript, but we have verified that the same conclusions hold and can provide these results as supplementary material if needed.

Overall, pre-training promotes physically consistent large-scale relationships while retaining sufficient flexibility to represent localized variability and extremes.

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 thank the reviewer for this insightful comment. The rationale for using STATIONS-IBEB as the downscaling target is twofold: station-based datasets are commonly used in operational regional climate services but typically have limited temporal/spatial coverage compared to national gridded products and this setup allows us to evaluate whether relationships learned from gridded data can effectively transfer to point-based observations, which have distinct spatial characteristics.

To address the reviewer's comment regarding distribution tails and the treatment of extreme values, we analyzed the precipitation histogram by comparing ROCIO-IBEB grid points with their corresponding STATIONS-CAT stations (which are independent of the ROCIO-IBEB construction), placing particular emphasis on extremes by applying a logarithmic scale to the y-axis.



This comparison shows that the interpolation technique used in ROCIO-IBEB adequately preserves the distribution of extremes present in station-based datasets. Combined with the independent validation on STATIONS-CAT, these results demonstrate that the physical relationships learned from ROCIO-IBEB successfully transfer to station-based datasets with different spatial characteristics, supporting the validity of our pre-training approach for operational applications where regional station networks are the primary observational resource.

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 this interesting suggestion. However, we note that the differences in TXx and TNn performance between fine-tuned and fully-trained models in Figure 3 (in the revised manuscript) are minimal and almost fall within the variability across training replicas. More importantly, the objective of our proposed paradigm is not to maximize predictive skill but to improve consistency and robustness without degrading performance, a goal that the current results support.

The key finding is that fine-tuned models achieve comparable skill to fully-trained models while simultaneously providing enhanced physical consistency (Figures 5, 9, 10) and reduced projection uncertainty (Figures 7, 11). This trade-off is particularly favorable given that the small differences in extreme metrics are negligible compared to the substantial gains in robustness and interpretability.

Furthermore, the independent validation on STATIONS-CAT reinforces these conclusions: fine-tuning preserves performance while learning physically meaningful relationships that models trained from scratch fail to capture. Given the current length of the manuscript and the clear demonstration of our core findings, we believe that an extensive additional analysis of extremes would not substantially alter the main conclusions regarding the benefits of the proposed pre-training approach for operational regional downscaling applications.

We have added clarifications emphasizing these points in the revised manuscript.

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 focused on Aggregated Saliency Maps (ASM) as a deliberate choice to provide a robust and directly comparable diagnostic of the predictor–predictand relationships learned under the different training regimes. ASM summarizes spatial saliency information in a way that reduces noise from individual events and facilitates systematic comparison across datasets and models.

However, we agree with the reviewer that regional or event-specific saliency maps can provide additional insight into the learned relationships. For this reason, in the revised manuscript we leverage the additional experiment introduced for the STATIONS-CAT dataset and include the saliency maps computed over the test period without aggregating for all grid points in the predictand. These maps allow a more detailed exploration of how relevance is spatially distributed across the predictor domain for the different variables.

The results support the conclusions drawn from the ASM analysis in the previous version, and further highlight that, particularly for datasets with limited temporal coverage, fine-tuning

helps inherit the physically meaningful relationships learned by the pre-trained model. Specifically, the fine-tuned models exhibit coherent relevance patterns over the Mediterranean region (known to play a key role in precipitation over Catalonia) whereas the fully trained model shows more spatially fragmented relevance patterns that, while explaining the data, are less consistent with the underlying physical dynamics. We refer the reviewer to Figure 10 in Section 4.5 for details.

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 thank the reviewer for this insightful comment. As emphasized in the revised manuscript, the primary goal of our proposed methodology is not to maximize predictive skill, but rather to improve the consistency and robustness of deep downscaling methods when applied to regional datasets. In line with this focus, we have reframed the experiments in Section 4.4 to highlight how the pre-trained model captures relationships that are particularly useful in regimes with sparse data, when fine-tuning is performed on such datasets.

We agree that further evaluation of model performance under extreme conditions is interesting. However, we believe that our current results already demonstrate that the pre-trained model does not degrade performance compared to models trained on full datasets. This is illustrated through metrics such as the bias of the RX1Day index. These findings support our goal of improving robustness without compromising the fidelity of extremes.

While additional analysis of extreme events could provide further insights, we consider it beyond the main scope of this study.

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.