Author's comments in reply to referee 1

We thank Prajwal for the valuable comments and suggestions, which gave us useful insights to improve the quality of the manuscript.
Below we provide a detailed response to all comments. Referee comments are in bold, and author's comments are in *blue and italics*.

Jose P. Teran & Jhon Vila,
On behalf of the authors.

The paper presents a comprehensive comparison and evaluation of existing and downscaled precipitation products for the Peruvian Andes. The discussion section is particularly well written and effectively synthesizes the findings. However, the primary novelty of this manuscript lies in generating a regional dataset (downscaling precipitation specifically for the Peruvian Andes) rather than proposing a novel methodological framework. I urge the editor to consider whether this level of regional application aligns with the journal's specific criteria for novelty and scope.
Furthermore, significant methodological clarifications are required regarding the mathematical formulation of the GWR model, baseline dataset selection, and threshold coherency before the manuscript can be considered scientifically sound, which are described below.

Major Methodological Concerns
1. Justification of the Parent Product for Downscaling
The evaluation demonstrates that Rain4PE generally performs better overall, and PISCO performs better during extreme rainfall events than GPM-IMERGF. Given that downscaling primarily enhances spatial resolution (yielding finer spatial patterns) but inherently inherits the accuracy and biases of the parent product, the authors must justify why GPM-IMERGF was selected as the basis for downscaling over the superior-performing Rain4PE or PISCO datasets.

We thank the referee for this comment. Our choice of GPM-IMERGF was driven by the characteristics of our study region, where precipitation observations are extremely scarce and rain-gauge reliability is limited. Previous work in the area (e.g., Padrón et al., 2020) has shown that even neighbouring gauges can exhibit substantial inconsistencies, making gauge-corrected products difficult to interpret as a reference truth.
In this context, GPM-IMERGF provides a globally consistent, non-gauge-adjusted baseline that allows us to evaluate the downscaling procedure without introducing additional corrections derived from sparse or potentially biased ground observations. Our objective was not to replace/choose Rain4PE or PISCO, but to test whether spatial downscaling can improve the representation of precipitation fields when starting from a raw satellite product in data-scarce catchments. Rain4PE and PISCO are included in the evaluation precisely to benchmark the downscaled product against higher-performing alternatives. We also consider this study a first step toward systematic comparisons of ground-based and satellite precipitation datasets in data-scarce catchments, where such evaluations are still largely absent.

2. Ambiguity in Scale Selection for Downscaling (Line 172) The manuscript states that the best-fit exponential regression occurred at the 0.75-degree resolution. If I understood properly, the operational downscaling methodology appears to use only the 0.1-degree resolution. The authors need to clarify the purpose of identifying the best fit if it does not serve as the statistical foundation for subsequent spatial upscaling.

We thank you for noting this. As described on lines 166-172, and following the approach of Immerzeel et al. (2009), the exponential regression is fitted at each of the six tested resolutions. The resolution with the best fit, found at 0.75-degree resolution, provides the coefficients (a,b) used in the downscaling (Eqs. 1 and 3). These coefficients are applied at the native 0.1-degree resolution, necessary to calculate the residual at this resolution, and subsequently at 0.01-degree resolution, later corrected by the interpolated "high-resolution" residual, consistent with Immerzeel et al. (2009). We will revise lines 172-174 to state this more clearly.

3. Mathematical Notation in GWR (Equation 7).

Equation 7 differs from the standard Geographically Weighted Regression (GWR) formulation by imposing a non-linear interaction through the multiplication of two coefficients: one for optimization (δ), and one for regression (φ). If both parameters are intended to be calculated simultaneously, the system becomes mathematically unidentifiable via standard Weighted Ordinary Least Squares, requiring the authors to either write the assumption or explain how these coefficients were calculated. Also, what does this (δ) signify in this case?

Furthermore, to ensure methodological transparency and reproducibility, the authors must expand the methodology section to explicitly state the specific spatial kernel function used (e.g., a fixed Gaussian or adaptive Bi-square kernel) and detail the statistical optimization routine used to determine the bandwidth (e.g., cross-validation scoring or the Akaike Information Criterion).

We thank the referee for this observation. The issue arises from an oversimplified presentation of Equation 7 in the original manuscript. In the revision, we will present the standard GWR formulation, where only spatially varying regression coefficients are estimated. The term δ was not intended to represent a second regression coefficient; it referred to the bandwidth optimisation step. Because bandwidth optimization is performed separately from the local coefficient estimation, the identifiability problem noted by the referee does not occur. We will remove the ambiguous notation and clarify the role of each parameter. We agree with the referee. In the revised manuscript, we explicitly state the kernel function used (bi-square kernel) and describe the bandwidth optimisation routine (golden-section search guided by AICc). These details will be added to the Methods section to ensure full transparency and reproducibility.

4. Incoherent Precipitation Thresholds

If I understood correctly, there is a contradiction in the classification of rainfall events across the methodology. In Section 3 (Line 236), the authors establish a firm threshold of 2.5 mm/day to filter out satellite noise, defining any value below this threshold as "no rain. However, in Table 3, the ETCCDI indices for Consecutive Dry Days (CDD) and Consecutive Wet Days (CWD) are calculated using a threshold of 1mm/day. So, any day with less than 2.5 mm from GPM-IMERGf and the subsequent downscaled products will be classified as no rain, hence affecting CCD and CWD. This lack of coherence invalidates direct comparisons between the general and extreme forecasting indices. The authors must standardize this threshold across the entire study or provide a robust meteorological justification for shifting the definition of a "dry day" mid-analysis.

We thank the referee for this important observation. The difference between the 2.5 mm/day and 1 mm/day thresholds arises because they were used to represent conceptually similar notions of rainfall occurrence, but in different parts of the workflow.

The 2.5 mm/day threshold was employed as an operational definition of a "true" rainfall event, given the known low-intensity noise in GPM-IMERGf. In practice, this value was used to distinguish days with meaningful precipitation from days with spurious low-intensity signals. However, we acknowledge that this choice effectively defines what is considered a "rainy day," and therefore overlaps conceptually with the definition of dry/wet days used in other parts of the analysis.

In contrast, the 1 mm/day threshold used for the ETCCDI indices (CDD and CWD) follows the standard convention established in the ETCCDI framework. These indices are computed after the downscaling process and applied uniformly across all datasets to evaluate their ability to reproduce extreme-event statistics. Nonetheless, using two different thresholds to define rainfall occurrence introduces an inconsistency between the general evaluation metrics and the extreme-event indicators.

We agree with the referee that this inconsistency may affect the interpretability of the comparisons. To ensure conceptual coherence across all analyses, we will revise the manuscript and adopt a single, unified threshold for defining rainfall occurrence throughout both the general evaluation and the ETCCDI-based extreme indices. This adjustment will make the comparisons between indicators more consistent and representative.

Specific Line Comments

• Line 81: The argument that Machine Learning (ML) is limited because it "works only for the calibration period" is fundamentally incorrect and should be revised. Just like other models, ML modeling pipelines inherently account for this by using strict training, validation, and test holdout sets to ensure out-of-sample generalization. A stronger, scientifically valid argument against ML in this context would focus on its highly parameterized, non-linear nature, compared to the linear nature of equations used in this paper.

We thank the referee for this clarification. Our intention was not to imply that machine learning models cannot generalize beyond the calibration period; as the referee notes, ML pipelines explicitly incorporate training, validation, and test splits to ensure out-of-sample performance. The point we intended to convey is that ML approaches are highly parameterized and non-linear, and their extrapolation behaviour depends strongly on the representativeness of the training data. In contrast, the parametric relationships used in this study are simpler and physically interpretable, which can provide more stable behaviour outside the calibration period. We will revise the wording accordingly to avoid the misleading implication and to better reflect the intended argument.

• Line 86: The phrase "performance is adequate at low spatial resolutions" is vague. The authors should define what quantitative metrics or thresholds constitute "adequate" in this climatological context.

We thank the referee for the comment. The phrase "adequate performance" was meant as a summary of conclusions from previous studies, not as a metric defined in our work. We will revise the wording to make clear that this refers to findings in the existing literature and not to a quantitative threshold introduced in our study.

• Lines 195–197: The manuscript states that precipitation uncertainty is particularly acute in the Andes due to sparse station networks and subsequently claims to "address this gap." This is misleading. While the authors produce a high-resolution interpolated map, the fundamental epistemic uncertainty arising from a lack of ground-truth validation data remains unresolved. The phrasing should be tempered to reflect that the paper provides a higher-resolution estimate rather than a solution to the underlying data scarcity.

We thank the referee for the comment. We agree that our study does not solve the underlying data-scarcity problem, which can only be addressed by expanding the station network. Our aim was instead to test and provide a higher-resolution precipitation estimate that operates independently of that limitation. We will adjust the phrasing to make clear that the work offers improved spatial detail, not a solution to the fundamental lack of observations.

Suggestions for Figures and Visualizations

• Figure 3: The image quality is insufficient. The dotted (median) and dashed (quartile) lines inside the violins are nearly impossible to read, making it difficult to verify the authors' claims in the text visually.

We thank the referee for the comment. We will improve the figure quality and adjust the internal violin-plot markings so the median and quartile lines are clearly visible.

• Figure 4 effectively presents the conclusions the author intends, so I suggest keeping it before Figure 3.

Thank you for your suggestion. We will change the order.

• Figures 5 & 7: It is exceedingly difficult to judge the spatial differences between GPM IMERGF, GPM-EXP, and GPM-GWR simply by looking at absolute value maps side by-side. The authors may consider replacing or supplementing these with Difference Maps (e.g., mapping) to highlight where spatial deviations occur explicitly.

We appreciate this observation. Comparing several products without a single reference dataset is indeed challenging. We will revise these figures and explore more informative visualizations, including difference maps, to better highlight spatial deviations between products.

• Figures 8 & 12: Connecting discrete data points with a continuous line implies a sequential relationship between those points that may not exist. A scatter plot or grouped bar chart would be a more statistically honest representation of this data, in my view.

Thank you for your suggestion.We agree that the current line plots may imply relationships that do not exist. We will replace them with bar-based or scatter-based representations to avoid misinterpretation.

• Figure: The presentation of this continuous graph is unclear. Please explicitly define in the methodology or caption how the elevation binning was performed to generate these lines.
Figure 9 is just a continuous line.

We thank the referee for this comment. Figure 9 will be standardized to follow the same visual pattern as Figures 8 and 12, and we will clarify the elevation-binning procedure in the caption to ensure consistency and transparency.

References

Immerzeel, W. W., Rutten, M. M., & Droogers, P. (2009). Spatial downscaling of TRMM precipitation using vegetative response on the Iberian Peninsula. Remote Sensing of Environment, 113(2), 362–370. https://doi.org/10.1016/j.rse.2008.10.004

Padrón, R., Feyen, J., Córdova, M., Crespo, P., & Célleri, R. (2020). Rain gauge inter-comparison quantifies deficiencies in precipitation monitoring. La Granja, 31(1), 7–20. https://doi.org/10.17163/lgr.n31.2020.01