Author Comments on egusphere-2026-223 (Round 3)

We thank the reviewers for their continued engagement and helpful feedback. Below we provide our response to the final comments.

Reply to RC1 (Anonymous referee #1)

RC1-C1 *"I appreciate the author's thorough and reflective response to the comments and concerns raised in my first review. The response is adequate and well justified. The revised version of this paper provides a clearer and more concise presentation of methods, data and results than the original submission, which in my opinion makes it a more valuable contribution to the applied climatology community. I do not have any further comments or remarks on the paper."*

AC1-1: We sincerely thank the reviewer for their time, their constructive feedback throughout the review process, and their positive final assessment of the manuscript.

Reply to RC2 (Referee #2: Karandeep Singh)

RC2-C1 — Baseline design and DL comparisons > *"Thank you for the responses and revisions. The authors have addressed the majority of my concerns. I have a couple of comments: > > The added DL based baseline results are helpful. To avoid confusion, could the authors clarify what the model training data look like in practice. If the inputs are essentially tabular station/day records with coordinates and covariates, then a strong tree based baselines, especially xgboost, is important. I recommend including at least xgboost with a reasonable hyperparameter tuning effort, even if uncertainty quantification is evaluated separately."*

AC2-1: Thank you for this suggestion. Following the reviewer's recommendation, we performed an additional XGBoost experiment to check whether a strong tree-based model would change the conclusions of the study. The model used the same geographic, topographic, and temporal predictors, supplemented with spatial-lag features calculated from the k nearest stations using inverse-distance weighting.

The results are reported here for the reviewer as an additional sensitivity check. XGBoost achieved an MAE of 1.07 °C for temperature and 1.13 mm day^{-1} for precipitation. For comparison, SMACNP (Localized) obtained 0.80 °C and 1.05 mm day^{-1}, respectively, while RK obtained 0.94 °C and 1.24 mm day^{-1}.

Thus, XGBoost is competitive, particularly for precipitation, but it does not outperform SMACNP (Localized). In addition, inspection of the gridded fields revealed lower spatial coherence than for RK and SMACNP, including localized station-centred patterns and relatively abrupt spatial transitions (Figure R1). These features are not strongly reflected in station-wise MAE but are relevant for a climatological gridding product, where the spatial structure of the resulting field is also important.

We therefore retained the main benchmark of the manuscript unchanged and provide the XGBoost results here specifically in response to the reviewer's request. To ensure clarity, we have added a statement to Section 3.3 clarifying that SMACNP inputs are daily context sets rather than independent station-day records, and we have updated the Introduction to restrict our conclusions strictly to the methods formally evaluated in the manuscript.

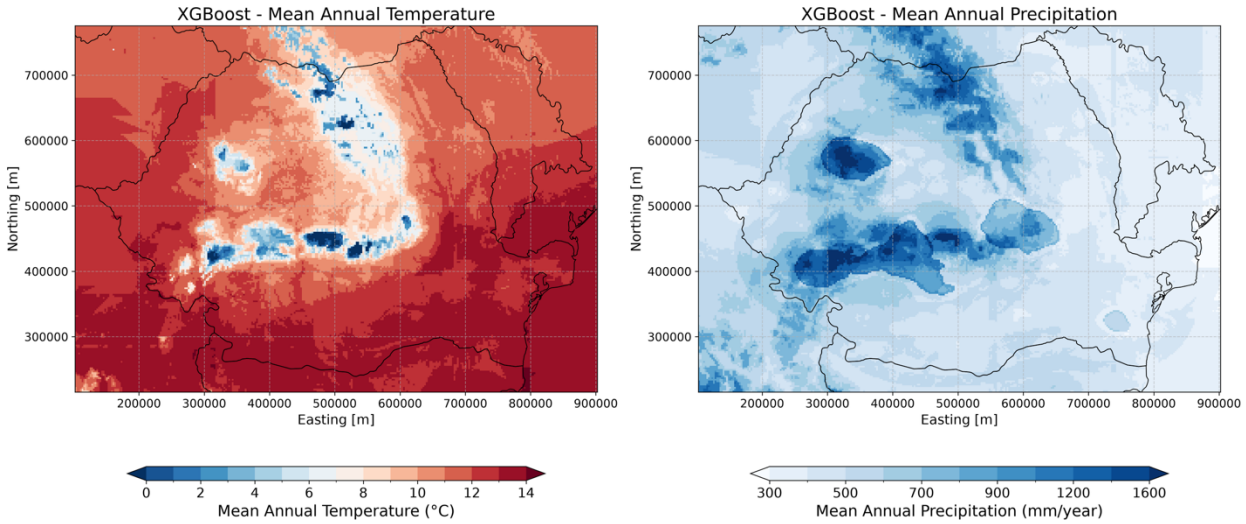

Figure R1. Mean annual temperature (a) and precipitation (b) fields obtained with the additional XGBoost sensitivity experiment. While providing competitive point-wise accuracy, XGBoost shows stronger localized spatial features (station-centred artifacts) and sharper transitions in the gridded fields compared to the geostatistical and neural-process methods evaluated in the main text.

RC2-C2 — Validation protocol clarity > *"The clarification on the held out test stations is good. I just want to further clarify whether"available days (2020–2023)" in authors' reply mean available days from training."*

AC2-2: Yes, the final evaluation uses all days in the 2020–2023 period. To clarify the complete protocol: the available days are split temporally into 80% for training and 20% for validation (used exclusively for early stopping and model selection). The final evaluation at the 26 held-out test stations is then performed across **all** days (both training and validation days), since the generalization axis is entirely spatial — the 26 test stations were never used during training, validation, hyperparameter tuning, or data normalization. For each day, the model receives the observations from the 130 training-pool stations as context and predicts at the 26 held-out locations. The model has thus seen the "weather of the day" at the training locations during training, but has never encountered the specific topographic-climate relationships at the test locations. We have updated the manuscript accordingly (Sections 1, 2.1, 3.3, and 5) to make this explicit and to clarify the scope of the baseline comparison.