This study presents a machine-learning-based approach to estimating gridded rainfall data for Saudi Arabia, an arid region with significant data limitations. The proposed dataset, SaRa, is compared against multiple existing precipitation datasets. The approach uses a combination of random forests and XGboots models. While not very novel, the results suggest superior performance, thus adding value and contributing to the data availability in the region. While the paper is well-structured with a sound methodology, fundamental concerns arise regarding the model accuracy away from training sites, generalizability, and reliability of the identified trends.

We thank the reviewer for their insightful comments.

Major:

1. I appreciate the authors filtering for potentially double precipitation gauges within 2 km, but the paper needs more clarity on how the split training/testing sample was performed. Was it random? Stratified? Distance-based?

Good question. This was done randomly. Since we had so many stations, we decided to allocate a very large amount (50%) to the validation subset, to ensure a robust validation. We have explicitly mentioned this in the text.

2. When applying ML to geospatial datasets, a critical issue is the use of testing sites near training sites that often artificially boost validation statistics. That's because precipitation data is spatially correlated. To enhance transparency and thrust into ML approaches, the accuracy of the ML models should also be evaluated based on their distance from training sites. Please plot the KGE testing accuracy of each testing point vs. its distance from the nearest training site (km). This will evidence how well the proposed ML approach is trusted in distant/ungauged areas. This plot would be informative for the main individual ML models and the ensemble stack.

Thanks for the great suggestions. We will include a new figure in the paper showing KGE values of testing stations versus distance to the nearest training station. This plot indicates that the KGE value does not decrease with distance to the nearest training station, underscoring the generalizability of our approach. We will also add corresponding text to section 3.1:

"A key limitation of ML-based P estimation is poor generalizability; models often fail in regions lacking training data (Xu et al., 2024). To assess whether this applies to our models, we analyzed KGE values of the evaluation stations as a function of distance to the nearest training station (Figure 4). The results show no clear decline in KGE with increasing distance, indicating satisfactory spatial generalizability."

3. The ensemble approach, while interesting, results in a black-box system—there is little discussion on the physical interpretability of the model structures and the predictive power of the inputs. Sklearn Random forests and XGboost have out-of-the-box libraries that can be easily deployed to evaluate model interpretability further. This could improve model understanding and expand the proposed approaches' generalizability.

We have calculated the importance of each predictor for each of the 18 model stacks and each of the four submodels. We cannot present all importances for all model stacks and submodels, as this would overwhelm the reader. However, we agree that physical interpretability of the model is important; therefore, we will add a new table to present importances for model_01 for all of its four submodels (Table 5). We will also add corresponding text to section 3.1 to discuss the results:

"To improve transparency, we computed predictor importances for all four submodels of model_01 (Table 5). IMERG-L V07 consistently ranked higher than GSMaP-MVK V8 in importance, indicating a model preference for IMERG, which aligns with its superior validation performance (Table 4). ERA5 was the most important predictor for the daily submodel, whereas IMERG dominated in the 3-hourly and hourly submodels. This likely reflects the advantage of observational datasets like IMERG in accurately capturing event timing. Static predictors were overall much less important than dynamic ones. Among the static predictors, abs_lat and lat had the highest importances, likely reflecting the latitudinal dependence of P product performance observed in global evaluations (e.g., Beck et al., 2017)."

4. The study does not sufficiently address uncertainty in trend estimations. There are no confidence intervals, no discussion of interannual variability, and no attempt to separate natural variability from long-term trends. Given the known issues with historical precipitation datasets, particularly in arid regions, one must question how much of the trend results from dataset evolution rather than actual climate change.

Thank you for the comment. We already strongly emphasize the uncertainty in the trends and also highlight the interannual variability in the trend section of the paper (section 3.5), however, to emphasize this further, we will add the following text:

"However, it is important to note that these trend estimates are subject to significant uncertainty due to considerable interannual variability, as well as substantial uncertainties in gauge, model, and satellite P estimates (see Section 3.4). In addition to random errors, satellite datasets are affected by transitions in data sources and radar sensors used for calibration (e.g., TRMM to GPM circa 2015; see Huffman, 2019), while reanalyses are affected by updates in data assimilation, such as the progressive inclusion of new satellite datasets (e.g., the TOVS to ATOVS transition in 2000), as well as the concatenation of

different production streams (Hersbach, 2020). These discontinuities propagate through and are reflected in SaRa. Interestingly, future projections from climate models in the sixth phase of the Coupled Model Intercomparison Project (CMIP6) indicate that increases in all three metrics are likely across most regions of Saudi Arabia (Iturbide et al., 2021; Calvin et al., 2023)"

Furthermore, to better convey the uncertainty in the trends, we will add dots to the map (Figure 6) to indicate grid-cells with significant trends ($p<0.05$).

Unfortunately, it is not straightforward to separate natural variability from long-term trends. Since long-term trends can also be natural, we assume the reviewer means anthropogenic impacts. However, these are extremely difficult to quantify for precipitation, requiring long observational records, convection-permitting climate simulations, and advanced attribution analysis.

Regarding the comment on separating "dataset evolution" from actual climate change, as mentioned in the Data and Methods section, time series from different model stacks are harmonized to our reference model stack, model_01. As such, the time series should be relatively homogenous through time.

Moderate:

1. The paper would benefit from a quantitative analysis and discussion of how temporal resolution mismatches in the gauge data impact validation results.

We do not think temporal resolution mismatches exert a major impact on validation results, because all the gridded datasets were either in daily temporal resolution originally, or have been aggregated to the daily resolution.

If the reviewer is referring to mismatches in gauge reporting times, it is true that reporting times may impact the validation, namely degrade the results for gauges with large mismatches. However, this phenomenon affects all datasets equally; hence it does not affect the performance ranking of our datasets. Note that we have highlighted the issue of reporting times in section 3.4:

"Time shifts between daily P totals from gauges and satellite or (re)analysis products (Yang et al., 2020; Beck et al., 2019) further reduce performance scores, especially in arid regions due to the short duration of rainfall events. The boundary between daily totals from satellite or (re)analysis products is midnight UTC, whereas it varies for daily gauge totals depending on regional reporting practices. In Saudi Arabia, the average boundary time was determined to be 05:00 AM UTC (08:00 AM local time; see Section 2.6). Consequently, for a short, hourly event, there is a 100 × 5/24 = 21 % chance that it will be assigned to the 'wrong' day."

Minor:

1. L 143 clarify what are gross errors.

Gross errors are unpredictable mistakes from careless observers when using equipment, reading scales or recording observations (Thapa and Bossler, 1992).

References

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I., Weedon, G. P., ... & Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrology and Earth System Sciences, 21(12), 6201-6217.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., ... & Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. Bulletin of the American Meteorological Society, 100(3), 473-500.

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., ... & Hauser, M. (2023). IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland., 1-34.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... & Thépaut, J. N. (2020). The ERA5 global reanalysis. Quarterly journal of the royal meteorological society, 146(730), 1999-2049.

Huffman, G. J. (2019). The transition in multi-satellite products from TRMM to GPM (TMPA to IMERG). Algorithm Information Document.

Thapa, K., & Bossler, J. (1992). Accuracy of spatial data used information systems. Photogrammetric Engineering & Remote Sensing, 58(6), 835-841.

Xu, Y., Tang, G., Li, L., & Wan, W. (2024). Multi-source precipitation estimation using machine learning: Clarification and benchmarking. Journal of Hydrology, 635, 131195.

Yang, S., Jones, P. D., Jiang, H., & Zhou, Z. (2020). Development of a near-real-time global in situ daily precipitation dataset for 0000–0000 UTC. International Journal of Climatology, 40(5), 2795-2810.