This paper described a new Gridded Meteorological Ensemble Tool (GMET) for probabilistic hydro-meteorological estimations. The new tool was developed with Python, ensuring the improved flexibility and wider application ranges. This work would be a big contribution to the community for ensemble studies. However, there are a few concerns need to be addressed before the paper can be published.

Response: Thank you for the comprehensive review and insightful feedback on our manuscript. We have addressed each of your comments in detail below and revised the manuscript accordingly.

Input data intensity & model performance evaluation: The method relies heavily on available gauge data, and authors use US area as shown studies in this paper. I wonder how the method perform when applied to data-sparse area? LOO cross-validation method might not be enough, but author can evaluate for instance different proportion of calibration/validation gauges, to check the change in model performance.

Response: We agree that the data intensity has a notable impact on the accuracy of spatial estimation. However, the primary focus of this paper is on the software development, and the experiments in this study serve as demonstrations of the software's capabilities rather than a comprehensive assessment of the method's performance across different data scenarios. We show the cross validation results because it is the foundation of empirical/data-driven prediction and is provided as an intermediate output by GPEP.

The GPEP provides numerous spatial estimation choices, and their performance all depends on their data support, as with any numerical/statistical modeling effort. The GPEP itself, as a collection of various geospatial estimation methods, cannot be as easily quantitatively analyzed concerning its sensitivity to data intensity as would be a single-method software. While the topic of data intensity and its influence on model performance is important, it may be more suitable for a separate study that dives deeper into the methodological aspects.

L152, one of the largest advantages of translating GMET to GPEP is the easy extension of python libraries. For instance, the supervised learning method, however, authors didn't show how this could be operated in the current model or how results are differing compared to conventional methods.

Response: We have added an experiment using random forest in Section 4.2. Please see the new Figure 6 and relevant analyses. For the experiment, we only use the default settings of the sklearn package. The efficiency of random forest is influenced by factors like hyperparameters and feature combinations, but a deep dive into these is beyond the scope of this paper due to the rationale provided in our response to the previous comment.

Authors state that the spatial interpolation method could be different from investigated variables, however, details about the method selection and explanation are missing. I would like to know which method is more ideal for different basic variables and the reason.

Response: This problem you proposed is pertinent and scientifically challenging. In the context of this manuscript, we explore four variables: precipitation, Tmean, Trange, and SWE. GPEP provides many interpolation methods, ranging from global to local, and linear to non-linear regression techniques. There are also numerous parameters to account for, such as weight functions, the number of neighboring stations, transformation exponentials, and predictor combinations. These factors are intricately interlinked, adding to the complexity of the problem.

Furthermore, the ideal method may vary based on regional characteristics. This means that even with extensive research in this study in a specific region, the generalizability of our conclusions to other regions may be limited. Given these complexities, it's beyond the scope of this manuscript to recommend an optimal method for each basic variable.

However, we posit that GPEP is a promising tool to delve deeper into this scientific question. Given the depth and breadth of the subject, multiple papers would be required to comprehensively address and report the findings. The author team will no doubt investigate some aspects of this problem in the future, given that GPEP was developed to support new applications, but the purpose of this paper is to document the conversion of GMET and GPEP, describe and give examples of some of its features, rather than any detailed application-specific analysis of the ramifications of different available choices (such as interpolation method).

L315-316. Why using the standard deviation of KGE is better for temperature?

Response: This is because this method avoids the impact of units (e.g., Kelvin vs Celsius) and the amplified bias around zero temperature (when Celsius is used). We have added some explanation in the manuscript.

Figure 6. Why the uncertainty range is so large?

Response: A large uncertainty range arises when the statistical model does not capture strong relationships between the predictors and the predictands, thus the calibration and predictive errors of the relationship (e.g. regression) are relatively large. In this case, the predictands are daily weather parameters in complex, mountainous terrain, from the foothills of the eastern Rocky Mountains to the Continental Divide, which is a notoriously difficult estimation problem and a long-standing challenge in hydrometeorology. Estimating this extreme precipitation case (which occasioned the Boulder floods of 2013) using geospatial methods based on ground station networks, as depicted in Figure 7a (formerly Figure 6a), is particularly challenging. One of the strengths of probabilistic estimates is their ability to unveil this uncertainty through ensemble members. The large uncertainty range also means that a traditional deterministic estimation for the same event would likely be prone to substantial error (as is somewhat documented in the Gochis et al reference).

Figure 8. the scale of the two maps should be unified.

Response: We have unified their scale.