Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-642-RC2, 2022


# Comment on hess-2021-642

Oscar Manuel Baez Villanueva (Referee)

Referee comment on "A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China" by Huajin Lei et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-642-RC2, 2022

The article titled "A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China" assesses the effectiveness of three machine learning-based algorithms to merge precipitation products over China. The article is very well written, concise, relevant, and I enjoyed reading it. I believe that it fulfills the requirements to be published in HESS. In the following points, I describe some points and suggestions that I think can be considered to increase the quality of the manuscript.

ERA5 has already superseded ERA-Interim; therefore, I recommend using ERA5 instead. Similarly, the authors used TMPA 3B42 (with IMERG), which is no longer in production. For reproducibility purposes, it is essential to use products that can still be acquired.

I suggest including a figure showing the number and location of stations used in the GPCC product in the appendix or supplement material. This information will be beneficial for the readers.

I liked that the authors obtained the meteorological data from GLDAS and not from ERA-Interim or ERA5. This can help avoid introducing biases towards the ERA product (although they used the cloud cover from the ERA product).

It would be helpful to describe whether the ground-based data were quality controlled.

The readers would benefit substantially from an improved description of the methodology. I liked Section 3.1 as it is very clear and informative, as well as the description of the

machine learning algorithms and their respective figures. However, I believe that the description of the two-step merging strategy (Section 3.2) can be further improved. The authors mention that first, the gauge observations are classified into wet and dry days to be used as the benchmark for classification models. Can the authors explain what the reason behind this is? It is not clear how the dry days were separated from wet days. How is a day classified as dry? Is the classification performed by grid-cell and by day? What is the final result of this classification? How can data scarcity affect this classification? Solving these questions is crucial for the understanding of the manuscript. Additionally, the authors mention that a regression process was applied to the wet days for the cold and warm periods. How were the models trained independently for the warm and cold periods? Perhaps it would be helpful to do a diagram for the first and second steps where the process is clearly explained. This will increase the manuscript's impact and is essential as the article presents these novel merging procedures.

I liked the idea of using the semivariogram as a spatial autocorrelation variable :). Can the authors discuss the influence that the selection of a particular semivariogram model can have in applying the method?

For reproducibility purposes, please mention the parameters of the RF, GBDT, and XGBoost that were used while training the models during the warm and cold periods.

The authors evaluated the performance of two categorical indices (CSI and HSS) over different precipitation intensities. This is a very good idea as the detection of no-precipitation events mainly masks the categorical performance. In this sense, as the objective of this article is to assess and compare the effectiveness of merging precipitation products using different ML techniques, I suggest evaluating all categorical metrics over these precipitation intensities. This separation into rain intensities will provide additional insights regarding the performance of these merged products.

I want to congratulate the authors for their figures. They are very nice and informative :)

The first two sections of the Results are a bit puzzling for me. In Section 4.1, "Performance assessment for classification results", the results obtained during the first step of the merging procedure are shown. However, in this section, the authors evaluate precipitation intensities (see Fig 6), which I believe, according to the methodology, are the result of applying the regression models over the wet days. Later on, in Section 4.2, "Performance assessment for regression results", the authors mention that regression models predict the final results presented in this section. By improving the explanation in the methodology, these two sections will be much clearer, and this issue can be solved.

I liked Section 5.2. I think that it will be helpful to perform a simple statistical test to assess whether the performance of the different methods is statistically significant or not.

Can the authors discuss the influence of the gauge density in the performance of the merging methods? This will be very helpful for the readers as in many regions, there are fewer rain gauges than those used in this study. This can give an insight on how to apply these techniques over diverse areas with different gauge densities.

I also liked the experiment performed in Section 5.1, where different approaches are evaluated and inter-compared.

The Discussion section could be improved by comparing the results of the present research with the wider literature. This is important because it will help to highlight the novelty of the approach. Similarly, the authors can discuss the implication of using products with different spatial resolutions and the representativeness of the rain-gauge stations with their respective grid-cells.

L355: significantly is a statistical loaded term, which must be accompanied by its respective p-value. If the p-value is not provided, I suggest using the word "substantially" instead. Please apply this throughout the manuscript.

L395: The authors mention the following "...although Kriging exhibits better performance than original MSPs, its accuracy is strongly dependent on gauge density. This only gauge-based interpolation method would have limited in complex mountainous areas with few gauges." I would be cautious with this statement. Although I agree with it, the ML algorithms cannot predict values outside of their training range, which could be translated into plausible underestimating precipitation over high elevations. Additionally, these techniques are also affected by the size of the training sample. Therefore, the ML techniques, in a sense, have limited performance in complex mountainous areas with few gauges as well.

L400: CC or r is also a component of the KGE. Also, see L412. I suggest including it inside of the KGE.

Figure7: Nice Figure! I think that it should be KGE instead of KEG.