Earth Syst. Sci. Data Discuss., author comment AC2
https://doi.org/10.5194/essd-2022-299-AC2, 2022

Reply on RC2

Yaozhi Jiang et al.

Author comment on "TPHiPr: a long-term (1979–2020) high-accuracy precipitation dataset (1/30°, daily) for the Third Pole region based on high-resolution atmospheric modeling and dense observations" by Yaozhi Jiang et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-299-AC2, 2022

Response to Reviewer 2

General comments:

High-resolution precipitation over the Tibetan Plateau(TP) region is important in climate science and other related fields. Climate models can simulate high spatial-temporal resolution precipitation datasets but generally overestimate the precipitation amount. The gauge-based rainfall observations are relatively accurate but only short-period, sparse-distribution records. This manuscript tries to take the advantage of both two and generates a high-resolution 1/30° long-term (1979-2020) precipitation dataset (TPHiPr) over the TP. A high-resolution pre-derived precipitation dataset (ERA5-CNN) and a dense gauge-based dataset are used. The manuscript describes the merged procedure and then intercompared the TPHiPr with independent station observations and several global datasets. The TPHiPr will benefit the researchers who are working on the climate or related works. However, before the manuscript was published in the journal, the below comments should be answered or clarified.

Response: *Thanks for the reviewer's comments and we believe that these comments are beneficial for improving our work. We have carefully considered these comments and a point-by-point response is given as follows. A full revision will be given at a later stage.*

Major comments:

1. From the data construction procedure (flow chart) and description in section 3, the RF and Kriging were repeatedly used to convert data between grid cells and gauge stations. However, the manuscript does not provide the reasons and also does not describe the methods in detail. Machine Learning has been used in climate sciences for decades and it includes many different algorisms. The RF is only one of them. Similarly, ordinary Kriging is also one of the interpolation methods. There should be specific reasons to choose those two approaches. It is necessary to provide them clearly in the manuscript.

Response: *Thanks for the comments. The interpolation algorithm used in our study is based on the idea of Regression Kriging, in which the interpolated variable is assigned to*

the spatial trend (deterministic) and the stochastic component (residual). A regression model is applied to predict the spatial trend and the Ordinary Kriging is used to estimate the stochastic component that is expected to be Gaussian distribution. In this method, multiple regression methods can be combined with Kriging. Machine learning-based regression models combined with Kriging were widely applied in earth science and proved to have good performance, as reported in many previous works (Araki et al., 2015; Cellura et al., 2008; Demyanov et al., 1998).

In terms of different machine learning methods, the Random Forest (RF) is an ensemble method based on Decision Tree. It randomly selects samples for training each Decision Trees and aggregates estimates from multiple Decision Trees, therefore, it is less likely to suffer from overfitting and has good generalization capability. Many works have applied the RF in earth science and demonstrated its good performance (Baez-Villanueva et al., 2020; He et al., 2016; Zhang et al., 2021). To demonstrate the reliability of the RF, we compared the performance of four widely-used machine learning methods for estimating the monthly precipitation in 2018. Figure R1 attached in the supplement shows that the RF generally performs better than the other three methods.

In the revised manuscript, we will further clarify the underlying logic of the merging algorithm and introduce more about the RF and Kriging.

.....Figure R1 was attached in the Supplement

Figure R1 Comparison between the monthly precipitation in 2018 estimated by four machine learning models and the observed monthly precipitation. RF: Random Forest; MLP: Multi-layer Perceptron; DT: Decision Trees; LGB: LightGBM.

Araki, S., Yamamoto, K., Kondo, A., 2015. Application of regression kriging to air pollutant concentrations in Japan with high spatial resolution. Aerosol Air Qual. Res. 15, 234–241. https://doi.org/10.4209/aaqr.2014.01.0011

Baez-Villanueva, O.M., Zambrano-Bigiarini, M., Beck, H.E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K., Giraldo-Osorio, J.D., Xuan Thinh, N., 2020. RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements. Remote Sens. Environ. 239, 111606. https://doi.org/10.1016/j.rse.2019.111606

Cellura, M., Cirrincione, G., Marvuglia, A., Miraoui, A., 2008. Wind speed spatial estimation for energy planning in Sicily: A neural kriging application. Renew. Energy 33, 1251–1266. https://doi.org/10.1016/j.renene.2007.08.013

Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., Timonin, V., 1998. Neural Network Residual Kriging Application for Climatic Data 2, 215–232.

He, X., Chaney, N.W., Schleiss, M., Sheffield, J., 2016. Spatial downscaling of precipitation using adaptable random forests. Water Resour. Res. 52, 8217–8237. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., Ge, Y., 2021. Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. J. Hydrol. 594, 125969. https://doi.org/10.1016/j.jhydrol.2021.125969

2. L193-196. "the daily precipitation fields after residual correction (Pd2) are further adjusted to ensure that the sum of the daily precipitation amount in a month…" At a certain station/grid cell in the TP, the non-raining day in a month should be very common. Let's take an assumption. When the above monthly precipitation is greater than "the sum of the daily precipitation amount in a month", how do you perform the "adjust" on both rainy days and non-raining days? If you only add the differences in the amount on rainy days, this would enhance daily extreme. Otherwise, it will increase the frequency of rainfall if both rainy or non-raining days are "adjusted". A detailed "adjust" process is needed.

Response: *Thanks for the comments. We adjust the daily precipitation as follows:*

$P_{a,i}=P_{m1}*P_{o,i}/P_{m2}=P_{m1}*P_{o,i}/\Sigma P_{o,i}$

where $P_{a,i}$ is the adjusted precipitation for the ith day in a month, $P_{o,i}$ is the original precipitation for the ith day, and P_{m1} is the monthly precipitation, P_{m2} is the sum of the daily precipitation.

When the monthly precipitation (P_{m1}) is no-zero but the sum (P_{m2}) of the daily precipitation amount in that month is zero, we will search the nearest grid that has a non-zero P_{m2} and then disaggregate P_{m1} to daily precipitation according to the day-to-day variation of precipitation in the nearest grid. In fact, the differences between P_{m1} and P_{m2} are small in most cases and the adjustment does not increase the daily extreme. We will add more details about the adjustment in the revised manuscript.

3. Figure 2 and section 3 present the data construction procedure based on the ERA5_CNN and observations at gauged stations. Over regions without observation (e.g., northwest TP in Figure 1b), is the TPHiPr directly from ERA5_CNN or another approach? Compared to Figure 3 and Figure 1b, it seems that regions without stations also show non-zero differences between TPHiPr and ERA5_CNN.

Response: *In regions without observation, the correction value is also non-zero. In the merging algorithm, the RF model is trained at gauge locations but the trained model is applied to all grids in the study area, which will result in precipitation changes in ungauged regions. In addition, the Kriging-based residual correction can also change the precipitation amount, although its impact is more evident in regions close to the gauges and less in regions far from the gauges.*

Minor comments:

1. The latitude and longitude labels on both the x-axis and y-axis are needed for all figures with the map.

Response: *Thanks for the comment. We will add the latitude and longitude labels in the revised manuscript.*

2. L124 To correct the biases of gauged precipitation, wind speed and air temperature from ERA 5 are used. Why do you use both variables from ERA5? Do you have any justification?

Response: *The ERA5 is the latest generation of reanalysis which has assimilated lots of in situ data. Our evaluation based on CMA stations showed that the wind speed and air temperature from ERA5 generally have better performance than those from two other datasets in the Third Pole (Figure R2 and R3, attached in the supplement). In addition, the results of Huai et al. (2021) also demonstrated the superiority of the near-surface climate from ERA5 to some other reanalysis datasets in the Third Pole region. Moreover, the ERA5 has a long time series, which can be used for correcting the early gauged precipitation. We will further clarify these details in the revised manuscript.*

Huai, B., Wang, J., Sun, W., Wang, Y., Zhang, W., 2021. Evaluation of the near-surface climate of the recent global atmospheric reanalysis for Qilian Mountains, Qinghai-Tibet Plateau. Atmos. Res. 250, 105401. https://doi.org/10.1016/j.atmosres.2020.105401

…… Figure R2 was attached in the supplement

Figure R2 Error metrics at each station based on daily 10-m wind speed derived from (a–c) ERA5, (d–f) HAR v2 and (g–i) WRF3 versus observation for the period from June to September of 2013.

…… Figure R3 was attached in the supplement

Figure R3 Error metrics at each station based on daily 2-m air temperature derived from (a–c) ERA5, (d–f) HAR v2 and (g–i) WRF3 versus observation for the period from June to September of 2013.

3. What interpolated methods are used to convert the TPHiPr from grid cell to station location when they are intercompared?

Response: *We compared the gauge observations with the precipitation from the nearest TPHiPr grid. Our dataset has a spatial resolution of 1/30°, the spatial scale of our dataset is more close to gauge observations than other coarse datasets. Nevertheless, we have to acknowledge that a spatial scale mismatch still exists between these two datasets. For dealing with this problem, very high-resolution datasets are still needed.*

4. L266-268 it is necessary to explicitly the station location in Figure 1 or in an additional figure. Also, the temporal range/resolution of those rain gauge-based precipitation should be given.

Response: *Thanks for the comment. We will add these details in the revised manuscript.*

5. Figure 7 shows the mean seasonal precipitation amounts from different databases. The spatial patterns of those datasets are very similar and cannot be distinguished by eye. I suggest plotting the differences between the three reference datasets and the TPHiPr.

Response: *That is a good suggestion. We will show the differences between these datasets and the TPHiPr in the revised manuscript.*

Please also note the supplement to this comment:
https://essd.copernicus.org/preprints/essd-2022-299/essd-2022-299-AC2-supplement.pdf