Dear Reviewer,

Thanks for your comments on our paper. Detailed comments and responses are as follows.

The paper applies a machine learning technique for downscaling and calibration of precipitation based on remotely sensed inputs that also aims to incorporate the spatial structure of rainfall using spatial autocorrelation. The idea of paper is interesting and it also has a organized structure which is generally well-written. However, based on the methods applied and discussion of the results, the paper has several shortcomings that need to be addressed and further explained prior to publication.

Major comments:

[1] Several aspects of the OK based interpolated maps at 1k and 10k resolutions are not fully convincing. First, the accuracy of the OK-derived maps should be reported in order to determine reliability of the maps. Errors in the interpolated maps are going to be propagated to the errors in the spatial RF model because it is one of the covariates used, so they are important. It would be interesting to see if the large RMSE's in the middle part of the study area in fig.7 also show up with large errors or variance in the OK maps.

[Reply]

The average error maps of OK for interpolating 1 km and 10 km precipitation products will be given in the revised paper, which could give a comparison to the RMSE maps in Fig. 7.

[2] Related to this, the authors also need to further clarify the interpolation of a 1km image based on a 10km IMERG images using OK, which is a raster-to-raster interpolation performed (lines 273-284). A coarse to fine raster-based interpolation seems unusual, so that authors need to further describe this step.

[Reply]

For IMERG interpolation, the raster-based values were transformed into point-based values with the form of spatial coordinates (e.g. *x* and *y*) and precipitation values, and then the scattered points were interpolated by OK to produce a map with the given resolution.

[3] The parameters tested and chosen for all the models, including the semi-variogram should be reported otherwise the study is not reproducible.

[Reply]

Kriging was used to produce the 10 km and 1 km satellite-based precipitation products and to interpolate the rain gauge observations for each month of the 5 years. Namely, 3*12*5=180 variograms were used in this study. The information is so much that cannot be listed in the paper due to the page limitation.

[4] It appears that the sRF model (also for the other ML techniques applied) did not include a separate testing phase . This is a standard approach applied when assessing the accuracy of a ML methods. I would suggest to also validate the models using an independent test that is not used in the training phase. Or re-configure the ML methods to split the total data into a training and a test set.

[Reply]

All the methods were assessed with separate testing points. the detailed information is as follows:

To quantitatively analyze the performance of all the methods, all rain gauge observations were randomly divided into *l* folds (e.g. *l*=10), where the *l*-1 folds (i.e. training/validation data) was used to construct the model, while the remaining one set (i.e. testing data) to assess the performance of the model (Xu and Goodacre, 2018). During model construction, the *l*-1 folds were randomly divided into training and validation datasets with the proportions of 80% and 20%, respectively, where the former was used to train the model and the latter to validate the model for tuning parameters. Then, the model with the optimized parameters was assessed using the testing data. The aforementioned process was repeated *l* times until all folds were taken as the testing data.

It can be found that the testing points were not used to construct the model.

[5] Discussion of the results focuses more on the positive aspects of using sRF but the authors do not give a balanced view by providing a critical analysis of the results of sRF. For instance, the accuracy metrics presented highlight that sRF performs well compared to the other models. However, visual comparison of the boxplots of these metrics alone in figs. 8-9 shows comparable accuracies all the models based on their range and median. Significant differences between the accuracies obtained, particularly in relation to sRF, should be reported to provide gravitas on the authors claim that sRF outperforms the other models.

[Reply]

In the revised paper, a balanced view will be added to the revised paper to give a fair assessment on the performance of the proposed method.

[6] There is an underestimation of precipitation values regardless of the model used based on fig 5 . This should be further elaborated in addition to the three accuracy metrics provided, so the bias of the estimates should also be reported. Furthermore, for very high precipitation values (e.g. >400mm), the scatter of the points in fig.5 becomes larger, indicating that all the models tested perform poorly at v. high rainfall amounts. It could be insightful to assess separately how the models compare for v. high rainfall conditions, since prediction of these extreme cases need to be generally improved.

[Reply]

Results illustrate that all models seem to underestimate the precipitation, especially for very high precipitation values (e.g. >400mm). This is because high precipitations are often caused by complex environmental factors, resulting in complicated predictors-precipitation relationships. Thus, more important land surface characteristics should be included into the model to improve the estimation accuracy.

The above information will be added to the revised paper.

[7] It is unclear how the importance measures are calculated from fig. 13, so this should also be included in the methodology of the paper. Furthermore, discussion of the rankings could be made more in depth by determining whether they agree or deviate (and why they do) from known controls on rainfall distribution.

[Reply]

To measure the importance of the *i*th predictor, its values are permuted while the values of the other predictors remain unchanged. Then, the OOB error based on the permuted samples is computed. Next, the importance score of the *i*th predictor is computed by averaging the difference between the OOB errors before and after the permutation. With the scores, the importance of each variable can be ranked.

Moreover, detailed analysis on the variable rank will be added in the revised paper.

[8] The authors already indicate that there is a delayed response of vegetation to rainfall. It is perhaps expected that the NDVI is one of the least important factors in the sRF model. But actually, this also provides an opportunity to also explore the lagged values of the predictors (and not only NDVI) with known delayed responses to rainfall.

[Reply]

This might be a reason for the least importance of NDVI. The above information will be added to the revised paper.

Minor comments:

[9] The captions of the figures need to be improved. Some of the features in multi-plot figures are hard to understand because of the captions are highly simplified.

[Reply]

The captions of the figures will be added to better understand in the revised paper.

[10] The final version of the manuscript will benefit for another round a English check as some sentences a phrased a bit vaguely (e.g. line 150-151)

[Reply]

Our paper will be polished by a naïve English speaker.

Best wishes,

Chuanfa Chen

Baojian Hu

Yanyan Li