Dear Prof. Eric Gaume,

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

---

[1] The proposed article is focused on an interesting question: the improvement of satellite-based precipitation products for the estimation of month, seasonal or annual precipitation amounts. It presents an original method aiming at improving the IMERG monthly precipitation product at local scale. The original aspect of the proposal, if compared to previously published methods consists in including spatial input variables in a random forest model: longitude, latitude and above all spatially interpolated rain gauge measurements based on ordinary kriging. The authors call therefore their method “spatial random forest”.

The article is interesting and overall well written and structured, but could be improved in several ways. Moreover, it suffers from an evaluation flaw that has to be corrected to provide accurate estimates of the real performances of the tested methods and fair conclusions: i.e. the performance of the proposed method should not be evaluated on a validation set, but on a test set, totally independent from the model calibration and selection step. The confusion between validation and testing is a common error in the implementation of IA methods when cross-validation procedures are implemented for the calibration and selection of the models. This pitfall has been pointed out by numerous authors and generally leads to substantially overrate the performances of the IA models (See ref 1. and 2, hereafter). The authors should not

split theirs samples into two, but three subsample: a calibration set (a) and a validation set (b) (used for the cross-validation model adjustment procedure) but also an independent test set (c) used in the final step of model assessment. This has absolutely to be modified to my opinion to provide sensible results, before the manuscript can be published in HESS. I am not convinced that if really tested on an independent data set, the performances of the proposed method remain higher than the performances of the kriging method…

[Reply]

Based on the aforementioned comment, the following scheme will be adopted in the revised paper:

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 (i.e. parameter optimization). 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.

[2] Some other aspects of the method and of its presentation could be improved (see

also the attached annotated manuscript):

[Reply]

The paper will be polished according to the comments.

[3] Some implementation information is missing and could be added in the manuscript such as the nuggets and ranges of the variograms used for the spatial interpolation.

[Reply]

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

[4] The authors should provide the names of the software and possible libraries they have used for the implementation of RF.

[Reply]

In the study, the RF regression model was performed with the freely available codes, downloaded from the website (https://code.google.com/archive/p/randomforest-matlab/downloads).

The above information will be added to the revised paper.

[5] Figure 13 gives an interesting insight into the calibrated model and the driving input variables. It would be interesting, to provide an even clearer insight, to test the real added value of the input variables in the SFR model. I have the impression that the dominant variable is the kriging result and not the spatial coordinates. Could the

performances of the model based on the spatial coordinate only or the kriging result only be provided for a more complete discussion. I have the impression that removing the spatial coordinates from the input variables, as well as all other terrain characteristics will have little consequences on the model result.

[Reply]

The relative importance rank of all the variables was obtained as follows:

RF can evaluate the relative importance of the predictors by means of the out-of-bag (OOB) observations, i.e. the samples without being used for model construction. Specifically, 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.

It can be seen from figure 13 that except for aspect, all variables have an effect on the precipitation estimation. The dominant variable is the kriging result and not the spatial coordinates. This result is expected, since the kriging result is precipitation-related value, whereas the spatial coordinates is not directly related to precipitation.

[6] The proposed model finally mostly consists in an intelligent merging between spatially interpolated rain gauge measurements and satellite downscaled precipitation. By the way, was the downscaling step really useful (see comments in the manuscript)?

[Reply]

Downscaling can decrease the scale mismatch problem, since the original IMERG has the resolution of 10 m, while the downscaled one has the resolution of 1 km. Moreover, the 1 km IMERG show more detailed information and variation of precipitation patterns than the original one, as shown in Fig. 12.

[7] Likewise, ordinary kriging is a relatively basic interpolation approach. I wonder if co-kriging or kriging of residuals approaches, popular for spatial rainfall interpolation, could also have been tested. But this is probably not feasible for the revised version of this manuscript but a suggestion for future developments.

[Reply]

In our further researches, we will assess the performance of co-kriging or kriging of residuals approaches.

Best wishes,

Chuanfa Chen

Baojian Hu

Yanyan Li