Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/hess-2021-332-RC3, 2021


# Comment on hess-2021-332

Eric Gaume (Referee)

Referee comment on "Easy-to-use spatial random-forest-based downscaling-calibration method for producing precipitation data with high resolution and high accuracy" by Chuanfa Chen et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-332-RC3, 2021

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 raingauge 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…

Some other aspects of the method and of its presentation could be improved (see also the attached annotated manuscript):

- 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.
- The authors should provide the names of the software and possible libraries they have used for the implementation of RF.
- 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. The proposed model finally mostly consists in an intelligent merging between spatially interpolated raingauge measurements and satellite downscaled precipitation. By the way, was the downscaling step really useful (see comments in the manuscript)?
- 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

Cited references:


- Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 2009 (3rd edition). page 709. "*Peeking is a consequence of using test-set performance to both choose a hypothesis and evaluate it. The way to avoid this is to really hold the test set out—lock it away until you are completely done with learning and simply wish to obtain an independent evaluation of the final hypothesis. (And then, if you don't like the results … you have to obtain, and lock away, a completely new test set if you want to go back and find a better hypothesis.)*
- Xu, Y., Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Anal. Test.* **2,** 249–262 (2018). https://doi.org/10.1007/s41664-018-0068-2. *"There is a significant gap between the performance estimated from the validation set and the one from the test set for the all the data splitting methods employed on small datasets"*

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2021-332/hess-2021-332-RC3-supplement.pdf