Response to Interactive discussion

Hydrology and Earth System Sciences (HESS)

Title: A Nonparametric Statistical Technique for Combining Global Precipitation Datasets: Development and Hydrological Evaluation over the Iberian Peninsula

Md Abul Ehsan Bhuiyan,[1] Efthymios. I. Nikolopoulos,[1,2] Emmanouil. N. Anagnostou,[1] Pere Quintana-Seguí,[3] Anaïs Barella-Ortiz,[3,4]

We would like to thank Reviewer for his insightful discussion and constructive suggestions. Below we provide a point-by-point response to his/her comments. Reviewer's comments are in red and our responses in black font.

-Major Comments:

1) The authors provide detailed information about the products used in the QRF method and explain the method itself very well. But no information is provided on how the training and validation of the method is performed. How much of the data is used for training? How much used for validation and testing? Please also include the temporal coverage of the data.

Ans:

First, we would like to clarify that the precipitation error statistics are based on hold-one-out validation. That means, each data-pair used in the validation statistics was not included in the training of the non-parametric model. This approach gives better estimates of the model performance because it trains and tests based on the entire data set. The temporal resolution of all precipitation products used in this study is three hours. This aspect will be better clarified in the revised manuscript.

2) There is also no information on avoiding overfitting. One of the challenges in data driven methods is overfitting (i.e. the method is so fine tuned to the training data, and has larger errors when applied to new datasets). I don't see any discussion of this in the paper. For example how did you choose to use 1000 trees in the model? Are there noticeable differences between the performance of the method during training and validation?

Ans:

Thank you for bringing the overfitting aspect in the discussion.

First we would like to note that the results showing in section 4 of the original manuscript are based on a validation dataset of 11 years and found to be prominent results for precipitation estimation as well as stream flow simulation. That means our model is successfully calibrated and is able to predict well the independent data.

Quantile Regression Forests (QRF) uses bagged version of decision trees and obtains a lower test error by variance reduction (Meinshausen, 2006). Higher number of trees reduces the variance of the model. So, increasing the number of trees in the ensemble won't have any impact on the bias of the model. Furthermore, a higher variance reduction can be achieved by decreasing the correlation between trees in the ensemble. Therefore, QRF utilizes the optimal number 'mtry' (size of the random subset of predictors) for split point selection at each node. It will introduce some randomness in to the ensemble to reduce the correlation between trees which helps to avoid overfitting (Meinshausen, 2006). In general, prominent '*mtry'* is obtained by cross validation methods in extending the sample size. The application of the machine learning tools can manipulate the training data in such a way that the actual results expected from the unseen data can be quite different from the evaluated results using the training data set, which is called overfitting. Therefore, in this analysis, we used hold-one-out cross validation method which prevent overfitting by producing reliable results. Applying this validation technique, the model has good skill on both the training dataset and the unseen test data.

To strengthen the validation results, in the revised version, we will present validation using one-year-leave-out cross validation. Namely, for each year of the database hold out for validation, we will be calibrating on the rest of the years (ten years). The performance of the combined product, based on the one-year-leave-out cross validation (presented in Figures 1 and 2 below), is found to be very similar to the results shown in the original manuscript (Figures 5 and 6) determined based on the hold-one-out cross validation.

Both validation approaches demonstrated that our model is able to reduce significantly the systematic and random error and is not overfitting.

As we discussed, higher number of trees would reduce the variance of the model and help to avoid overfitting. Therefore, the size of the forest should be relatively large for the stabilizing effect of many trees. For the Quantile Regression Forests, trees are grown as in the standard random forests algorithm and bagged versions of the training data are used for each of the k = 1000 trees to determine the optimal number *mtry* (Meinshausen, 2006). Therefore, to demonstrate the stability of QRF, the default value (k = 1000) is chosen throughout all simulations.

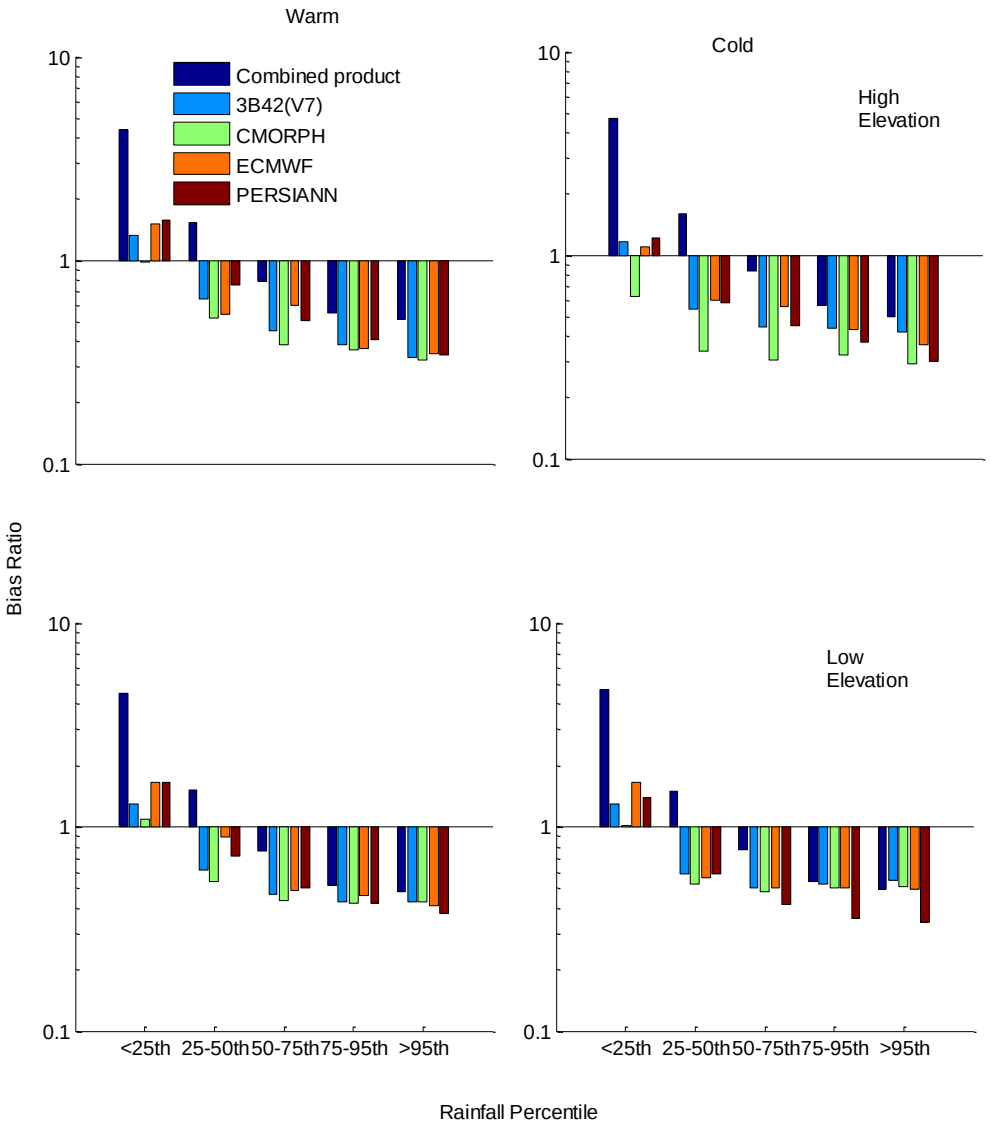

Figure1: Bias ratio for warm and cold season.

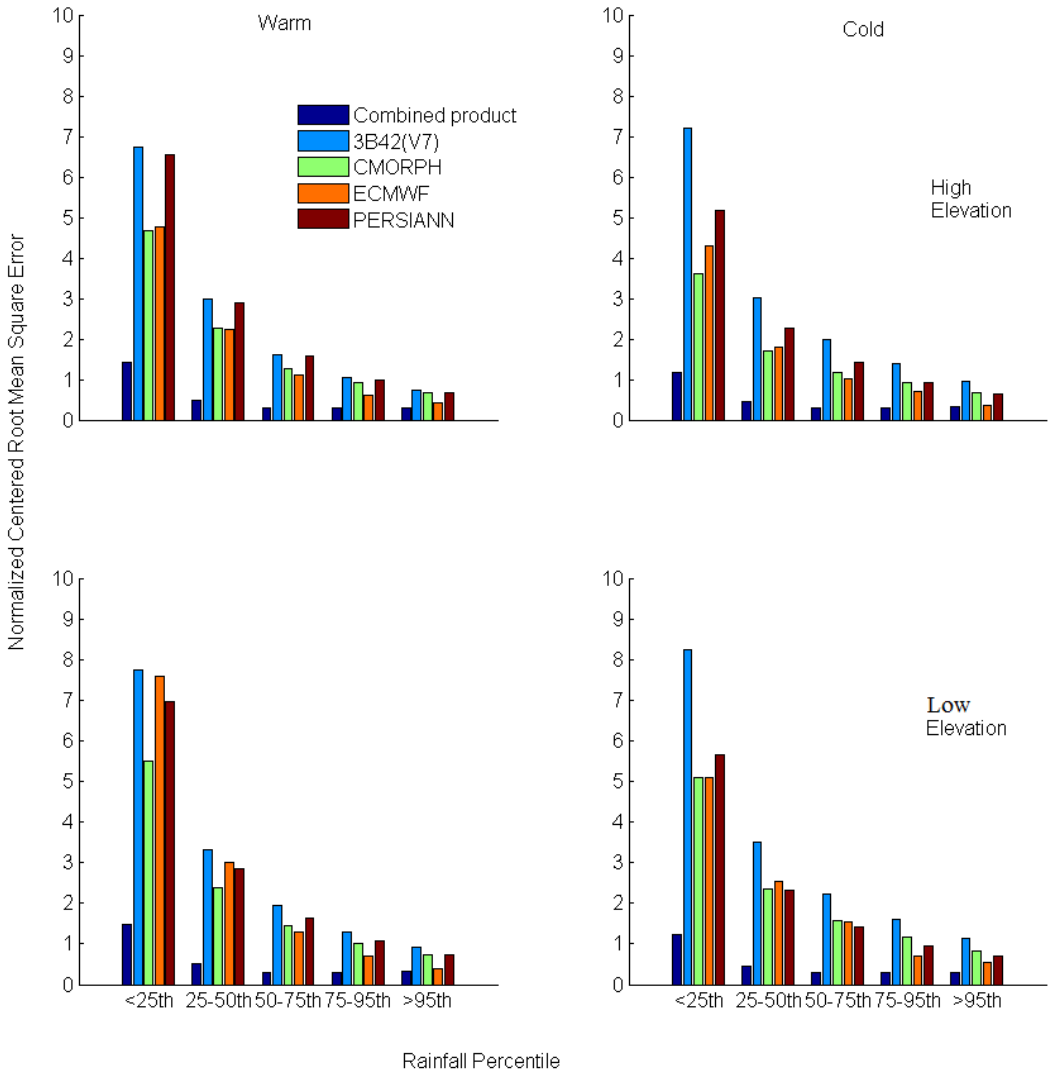

Figure2: Normalized Centered Root Mean Square error for warm and cold season.

3) How are the ensembles generated? No information is provided on how each ensemble member is initialized and generated using the QRF trained on the data.

Ans:

This is part of the statistical machine learning model. For QRF model, we initialize a random forest of 1,000 trees for each terminal node of each of the classified dataset. We calculate 95% prediction intervals for each grid. QRF utilizes the same weights to calculate the empirical distribution function. When, X is predictor variable and Y is response variable, QRF utilizes a

weighted average of all trees for the predicted expected response values to calculate the empirical distribution. To conduct the hydrological simulations in this study, we resampled from the empirical distribution function for 20 times per grid cell to obtain "reference"-like rainfall ensemble members. We will clarify this aspect in the revised manuscript.

4) The results provided in section 4 needs to be clarified whether they are based on the data used in training or the data used in validation, or a mixture of both.

Ans:

The results provided in section 4 are based on hold-one-out validation as explained in our response to question 1. Our study period spans eleven years (2000–2010) and we validated all those 11 years belonging in our dataset. Validation results are presented in section 4.

5) The low value of NCRMSE for the small basins report in Page 11, Line 6 is a signal of overfitting in the algorithm. This is another indication that overfitting should be analyzed in depth.

Ans:

It's actually not overfitting here. Generally, Overfitting depends on the inconsistency of training and validation model results. Overfitting refers good performance on the training data, poor generalization to validation data. Generalization indicates how well the concepts learned by a training model apply to new dataset. So, if we produce validation and training model for particular group of dataset and find inconsistency between two results, then we can justify overfitting. As we said in section 4, all the results are based on only validation results, there is no way of knowing whether overfitting or under fitting without comparing training results. So it is not possible to justify overfitting in algorithm to examine from the validation results only. Results for NCRMSE are shown in Figure 9, which are consistent in terms of the reduction of the random error for all the subbasins as well as precipitation and streamflow percentile ranges. This is the indication that how we successfully trained our model instead of overfitting.

6) Page 6, Lines 10-18: Please clarify if different trees are developed for the three groups that you introduce at the beginning of the paragraph. You have introduced four groups at the end (warm-high, warm-low, cold-high and cold-low) but there is no reference to the categorization of products based on their rain detection (group 1-3 in lines 11-12).

Ans:

Yes, we developed different trees for the three groups. If we grow similar kind of trees, every sampling will be equal that affects the model results. As we mentioned, (QRF) uses bagged version (bootstrapped aggregating) of decision trees by randomly sampling from bootstrapped sample which reduces variance and helps to avoid overfitting to improve the stability and accuracy of our proposed machine learning algorithms. That is the whole idea of choosing

ensemble method where trees grow independently because of the combination of bootstrap samples and random drawing of variables.

Actually, we classified available rainfall estimates from all the products (three satellite and reanalysis) into three subsets: (1) all rainfall products that report rainfall greater than zero (2) all rainfall products that report zero rainfall; and (3) at least one product that reports nonzero rainfall. Then, for each subset, we created 4 groups: warm period-high elevation, warm period-low elevation, cold period-high elevation, cold period-low elevation for the error model. Finally, we prepared total 12 groups from all threes subsets (each one has 4 groups) for the error model.

All these classification we created by our own justification to keep similar types of dataset together. If we keep different kinds of dataset together, our model will not be efficient in accurate prediction due to the lack of uniformity in dataset. Generally, the QRF model is expected not to capture well very low and extremely high values due to the weakness of the empirical distribution function to model probabilities close to 0 or 1. The distribution of proper sample size plays an important role in empirical distribution function. Therefore, very large sample sizes required for low and extremely high values to quantify the rate of convergence to the underlying cumulative distribution function. This is the reason we categorized our dataset from above mentioned procedure.

-Minor Comments:

1) Why did you choose to use PERSIANN product instead of the newer version PERSIAN-CCS?

Ans:

In this study we chose to use gauge adjusted satellite precipitation products: 3B42 (V7), CMORPH and PERSIANN. The gauge-adjusted PERSIAN-CCS is not available over the Iberian Peninsula. Using the PERSIAN-CCS in precipitation error analysis is a good suggestion that we could investigated in a future research.

2) In section 2.3, please include details on how you have downscaled the 0.5 degree reanalysis product to 0.25 degree to be consistent with other products.

Ans:

The dataset was interpolated in space and time using the nearest neighbor interpolation technique for every time steps so as to match the other products. We will add text about this aspect in the revised manuscript.

3) In section 2.4, please include the version number of the ESA-CCI product.

Ans:

The version number of the ESA-CCI product is v02.0 which we will add in in the revised manuscript.

4) Page 8, Line 2: What does actual uncertainty mean? Do you mean uncertainty in the reference product? If so, please explain how a UR=1 will provide the best estimate of the uncertainty in the reference product.

Ans:

Thank you for raising this question. Here, actual uncertainty indicates the maximum possible uncertainty of the prediction interval, which is 1. It is not the uncertainty in reference product. Uncertainty Ratio (UR) quantifies the prediction interval width relative to the magnitude of the predicted variable. UR value close to 1, indicates confidence intervals being in the order of magnitude of the predicted values. We will clarify about this aspect in the revised manuscript.