Interactive comment on "A statistically based seasonal precipitation forecast model with automatic predictor selection and its application to Central and South Asian headwater catchments" *by* Lars Gerlitz et al.

Anonymous Referee #1

Received and published: 8 April 2016

I only have some training in statistics, and don't have education background in hydrology or climate. My viewpoint may be quite different from the HESS community.

Technically, the forecast model is well-designed using various domain knowledge. I would like to know how kmeans is carried out, e.g., whether the correlation is used in the clustering, how to determine the random seeds or initial cluster centers. As we know, different random seeds will lead to different clustering results in K-means. Simply aggregating neighboring grids to a big region does not make sense if these grids have quite different correlation with the precipitation in the target region.

How can simply aggregating 'poor' monthly forecast to seasonal forecast lead to some good forecast results? Before using random forest, we should explore a single regression tree for monthly forecast first in order to identify which kinds of predictors are more important as well as its performance.

Reasonable forecast comparison should be carried out and reported. For example, what is difference if only those widely accepted climate indexes are used as predictors. How about comparing random forecast with a single regression tree. How about using 7 clusters instead of 7?

The sensitive analysis to me does not fit the key statement of paper that much. Actually from the appearance frequency of predictors in random forest can give us some ideas how important is a predictor.

Line 42, Page 15 has some typo or overclaim. AUC values are not always >0.7.

There are too many references. Reference Chen et al (2012) needs some correction.