Interactive comment on "Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression" by G. Ibarra-Berastegi et al.

Anonymous Referee #2

Received and published: 18 April 2011

Summary: The aim of the paper is to explore the utility of analogues for downscaling, and assess that approach relative to adding additional steps to the downscaling including analogues and multiple linear regression. To that extent it is reasonably approached,and the methods generally appear to support the conclusions. What is missing principally is the tie to downscaling GCMs, and generalizing the results to future conditions, which presumably is a motivation for this effort.

1) p. 1952, lines 21-26 and top of following page: This sets the stage for the experiment, where the challenge of using large scale (1.5-4 degrees) GCM output to estimate 'site-specific scenarios' requires some type of downscaling. There are two main issues separating the analysis presented in the manuscript from this, namely 1) by downscaling ERA products at about 1 degree spatial resolution there is much greater spatial resolution than most existing GCMs, and 2) by using reanalysis products, even though precipitation observations are not directly assimilated, other observations are, giving it much higher skill than any GCM will exhibit. For development of downscaling approaches this type of analysis is common, but a discussion of the implications of the findings in the context of much coarser, lower skill input to the downscaling procedure is missing.

The focus of this paper was to test the potential capabilities of a downscaling approach based on a start-of-the-art and at the same time, highly non-linear tool like random forests. To that end, all the literature mentioned in the "Introduction" chapter of our paper, represents the previous works our approach tries to compare with. This set of scientific works represents the frame in which our study tries to make an contribution. In this sense, all the literature that we cite on downscaling uses reanalysis data (either NCEP or ERA) and therefore, our work should be understood in this context. In this line, it is worth mentioning that our group has experience in precipitation downscaling using ERA40 data and analogues although for short-term prognostic purposes and not necessarily only for climate (Fernández-Ferrero et al., 2009; Fernández-Ferrero et al., 2010).

Coming to the GCM issue, we agree with the reviewer that in general their resolution is coarser than for example ERA40. However, GCM do not assimilate observations so performance must be assessed comparing PDF of observations and predictions. Our group has got experience in the evaluation of AR4 models (Errasti et al., 2011) and the evaluation methodology is different, mainly because in reanalises, observations are assimilated and in GCM they are not. Since GCM are always run beyond the limit corresponding to first kind predictability, in the verification process, it cannot be expected that the part of observations corresponding to the high frequency variability is accuretely represented. Therefore, a choice has to be made before downscaling is carried out: either reanalysis data are used as input or data from GCM are used as inputs. As a result, evaluation procedures are different and downscaling results from reanalyses such as ERA40 and GCM are not directly comparable.

As mentioned before, the present work uses ERA40 data as inputs and several models' abilities for precipitation and surface mositure flux downscaling are evaluated. Although comparison with results obtained using GCM models with coarser resolutions is not straightforward, it seems reasonable to expect a similar ranking of performance for the different models tested. As such, this study is of the sensitivity test kind. We

consider several factors fixed and we only vary the methodology followed for downscaling. If the editor gives us the opportunity to prepare a revised version of our paper and suggests us so, we can incorporate a brief discussion in these terms.

2) p. 1957, lines 4-9, some important source regions providing teleconnections to IP climate are noted. However, on p. 1959, lines the domain for large-scale predictors is limited to the predictand domain alone. What is the justification for this?

This is a good suggestion by the reviewer and we appreciate it. Our previous experience shows that bigger domains do not necessarily imply a better performance when downscaling. To that end, see for instance Fig. 13 or Table 1 in a work by one of the coauthors (Fernandez and Saenz, 2003).
There are two reasons behind this behaviour. The first one is that one of the previous steps almost always consists (as in our paper) of a dimensionality reduction by means of EOF. If the domain is too big, there might be areas of it where the variance of predictor fields is high but unrelated to the variability at the local sites.
Since leading EOF always point in the directions of highest variability, the leading directions might end up by being rather irrelevant in terms of local variability in the domain is too big.
The second reason is that if we consider a big domain (for instance Northern Hemisphere), the variability of every predictor can be described (up to the limit of truncation) as a linear combination of the EOF. Therefore, the variability of circulation (geopotential, wind and so on) over our local domain is also a combination of the hemispheric EOF. Only those EOF which represent significant fractions of variance over our local domain are able to originate significant anomalies over the smaller domain over the study area. Therefore, both sources of information are not independent. Conversely, we find that our approach is more robust in the sense that it does not depend on subjective a priori decisions on which teleconnection indexes should be considered.
As a reply to the other reviewer (Dr. Rasmus Benestad) we have performed a sensitivity analysis on the domain size and the results show that this factor is not important in terms of performance. (see reply to reviewer #1, and ANNEX).
If the editor gives us the opportunity to prepare a revised version of our paper and suggests us so, we can include the results, tables and calculations shown in the ANNEX, that in our opinion justify that for this case, the size of domain is not relevant.

3) p. 1961, lines 12-14, the GPCP dataset is used for reference values. Since the GPCP product is a merged data set of observations from a variety of sources, it is unclear why "any downscaling effort on precipitation could only be justified if better results than local persistence and/or raw GPCP data ...were obtained." Does the GPCP data include the observations stations for which downscaling is being performed? Why should any downscaling effort be expected to provide a better estimate of local conditions than an observationally-based data set? Something is not clear here.

GPCP provides gridded precipitation data with a 1ºx1º resolution. This precipitation data is readily available, http://jisao.washington.edu/data/gpcp/daily/ and if it yielded good estimations of observed data at the closest gridpoints (Zaragoza 59 and Tortosa 36 km away) downscaling would not bet necessary. For this reason, we think that evaluating GPCP performance along with the rest of models, can make it possible to evaluate whether the effort needed to dowsncale precipitation values is worthwhile or not. It is the same idea for persistence: if current day value can be accurately estimated with yesterday's record, downscaling is not needed. As can be seen in the results shown in our original paper, performance of both persistance and GPCP, was worse than for any other model. The conclusion is that performance is so poor that downscaling cannot be avoided.

4) p. 1964, line 4, express the RMSE as a percent of the mean value either instead of or in addition to the raw RMSE.

Raw RMSE is a very widely used statistical indicator and in our opinion, provides a good description of models' error but we have no problem with this suggestion. If the editor gives us the opportunity to prepare a revised version of our paper and holds this same point of view, we can either express RMSE as a percent of the mean or in addition to the raw RMSE.

5) section 2.3, the list of statistics for evaluating the methods could be more inclusive of extremes. One of the motivations for downscaling of daily data is to capture better extreme values. For precipitation, are extreme values captured more successfully by one method than another? Since RMSE is heavily affected by high values there may be an implicit assessment of this, but comparing estimates of heavy rain events would be interesting.

In our paper we use 5 years (1826 daily cases, 1997-2001 period) of data for models' evaluation. The focus of our paper was to carry out an overall evaluation of downscaling performance by different methods. Carrying out a thorough analysis of extremes would only be possible with a higher number of years and cases.
However, regarding your concern on precipitation extreme values, we have considered three extreme boundaries (percentiles 95, 97.5 and 99) and graphically represented observations vs predictions. It can be concluded that all models tend to underpredict extremes (Fig.1-Fig.2).
We have also tried to calculate the same set of statistical indicators for these three extremes. However, due to the low number of cases belonging to each of these groups of extremes, the statistical indicators used (R, RSD, RMSE, FA2, RM, D) have very wide 95% confidence boundaries as obtained after bootstrap resampling (5000 times) making it impossible a statistically meaningful model intercomparison. This can be seen in figures 3 and 4 for the correlation coefficient. With 5 years of data we can only state that for the case of precipitation extremes, all models tend to clearly underestimate observations.
I

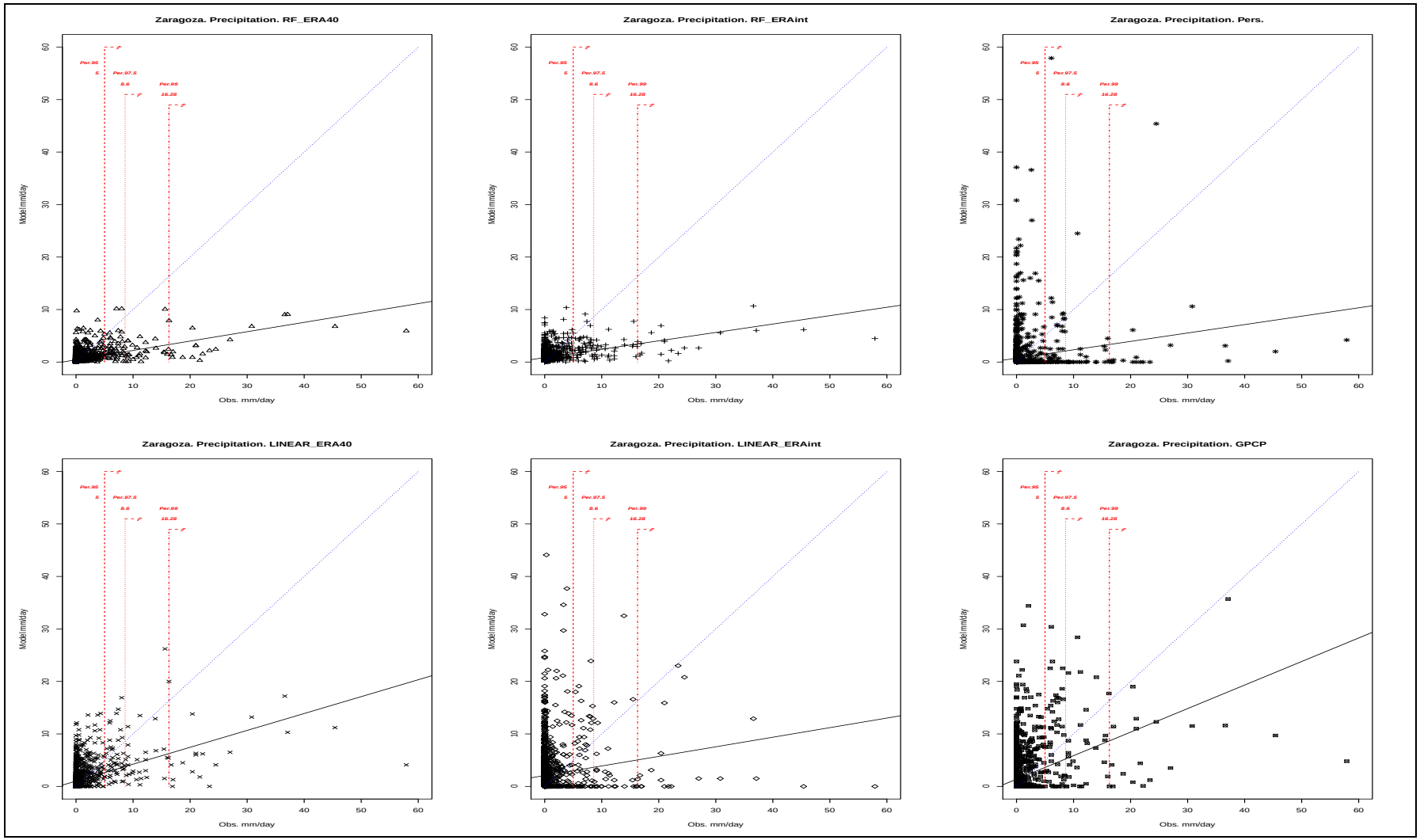

Figure 1. Zaragoza. 1997-2001. 1826 cases belonging to the test data set. Observed vs predicted by the different models. In red, 95, 97.5 and 99 percentiles.

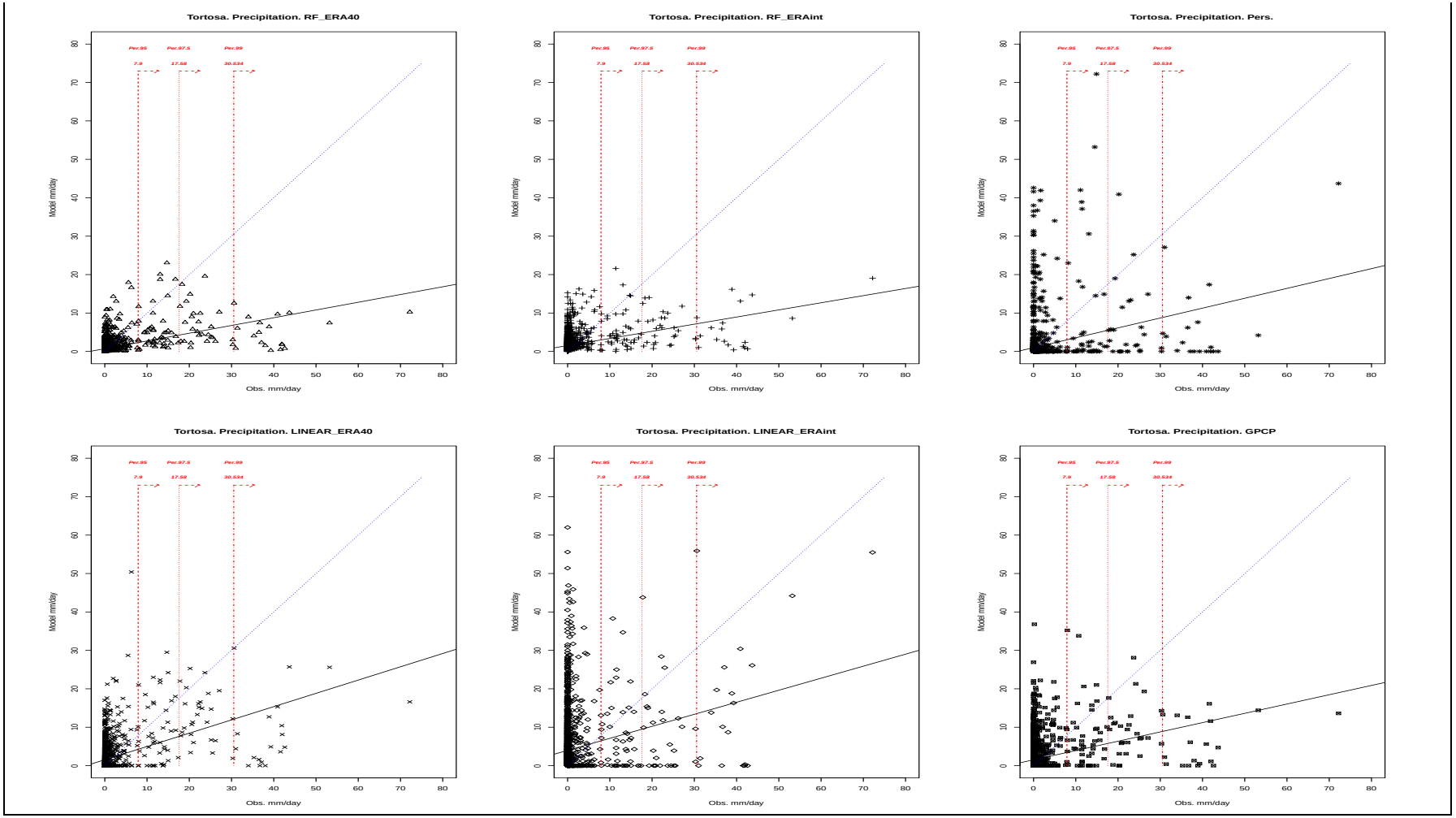

Figure 2. Tortosa. 1997-2001. 1826 cases belonging to the test data set. Observed vs predicted by the different models. In red, 95, 97.5 and 99 percentiles.

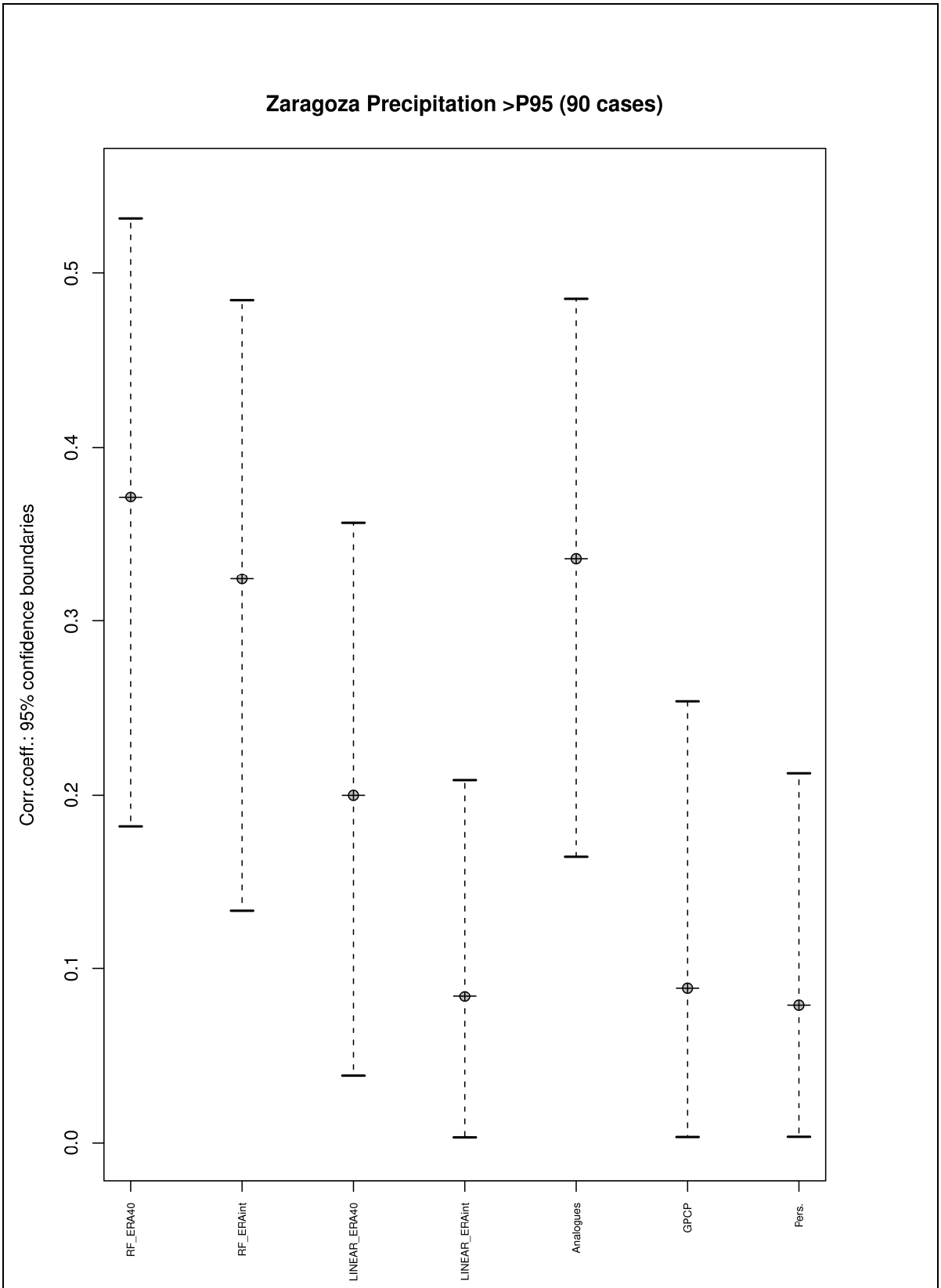

Figure 3. Zaragoza. Precipitation. 95% confidence boundaries of the correlation coefficient among predictions for the different models and observations > percentil 95.

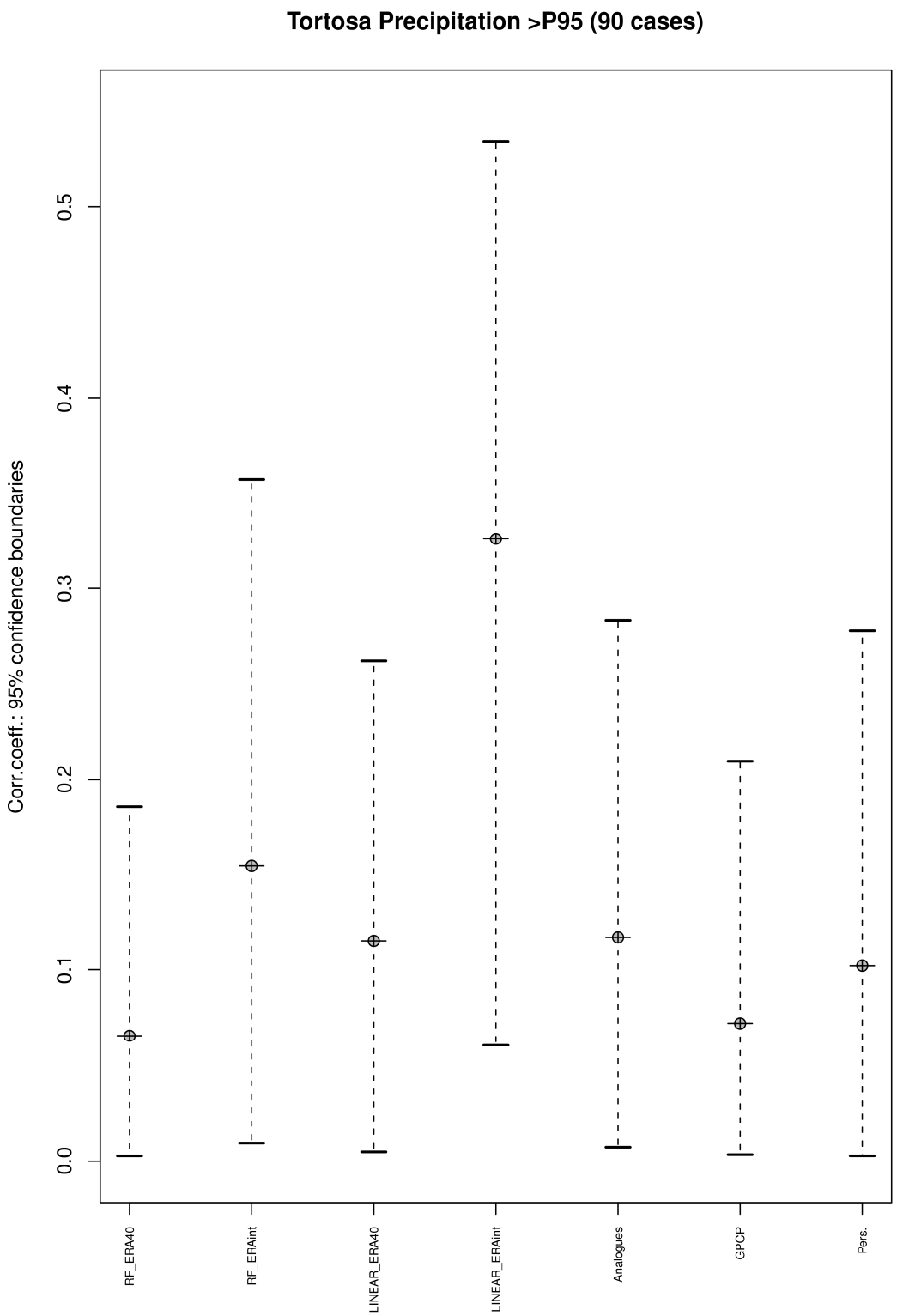

Figure 4. Tortosa. Precipitation. 95% confidence boundaries of the correlation coefficient among predictions for the different models and observations > percentil 95.

6) One issue in using historical analogues is that, when applied to future climates that may bear less resemblance to historic climate, the number of available analogues and their correspondence to simulated patterns may decrease significantly. Some comment on the range of projections and how that might affect the applicability of this method in future climates would be helpful.

We agree with the reviewer at this point, at least partially. The first reviewer also raises the same difficulty and a longer answer is given in reply to Rev#1. However, the main lines of our reply will be repeated here.

1. Some studies exist *(M.D. Frias, E. Zorita, J. Fernandez and C. Rodriguez-Puebla, 2006. Testing statistical downscaling methods in simulated climates, GRL 33; L19807)* which show that this result is dependent on area, kind of averaging and predictor
2. Statistical downscaling and particularly, analogues, can also be used for climate prediction (Fernandez-Ferrero et al., 2009, Fernandez-Ferrero et al., 2010)