E. Zehe (Referee)

erwin.zehe@kit.edu
Received and published: 30 May 2011

Reply to referee E. Zehe

Dear Erwin Zehe, the authors thank you very much for your valuable comments which help to improve our paper significantly! We hope that we could fully address all points. Concerning the discussion about our bias correction procedure, we will reduce its weight in the paper and we will set a focus on index interpretation relative to catchment properties and input data.

This is an interesting study that defines and uses smart signature indices based on the flood frequency curve and on event runoff coefficient to assess change impacts on catchment response behavior. The study is based on a sound data base, the manuscript is well structured manuscript - though some copy editing by a native speaker is still necessary. The presented findings are potentially of high interest for the audience of HESS. Unfortunately the authors miss many opportunities to work out potentially very interesting findings of the proposed study. I strongly encourage the authors to better work out the hydrological insights within major revisions. I hope the authors find the following major and minor points useful for this task:

The revised paper will much more focus on the hydrological insights gained by the signature indices.

Major points:

- The author should better discuss the potential of the proposed indices to assess quality of model predictions where model is defined as code, parameter set for a given catchment and meteorological input data. Even with observed rainfall LARSIM performs poor for instance with respect to peaks flows, and very poor during summer events (30 -40 % overestimation of runoff coefficients for catchments Kronweiler and Kellerbach). Model performance is not acceptable for Gensingen. These findings should be better stressed and discussed. Especially the poor performance in Gensingen needs to be discussed in as it indicates that soil moisture accounting in LARSIM might be ill suited for Loess soils.

This is correct: The indices reveal quite poor model performance for Gensingen. For this gaging station, it has to be assumed that the water balance is not closed. Therefore, the corresponding indices show large differences in runoff behavior.

- The author miss the opportunity to interpret and discuss the obvious shortcoming of the bias correction in Cosmo. Bias corrected rainfall data do not necessarily cause a better model performance. At least this cannot be inferred from figure 6.

This is correct: The indices reveal the shortcomings of the bias correction of the COSMO data: We will discuss this in more detail in the revised manuscript.

The authors should always compare both bias corrected and uncorrected rainfall to simulations with the observed rainfall input and to the signature derived from observed discharge. This will show,

which COSMO version performs better in comparison to observed rainfall input and how close/far COSMO driven simulations are from the real integral catchment behavior.

Our point of view is, that comparison with observed discharge makes not very much sense because of the large differences between measured and simulated discharge shown in Figure 4 (resulting from model errors & water balance problems at Gensingen). Therefore, we use the simulated discharge form measured rainfall as our baseline for figure 5. Figure 6 then gives an estimate on the effect of bias correction on hydrologic behavior, wheras in Figure 7, we extract the climate change signal. Direct comparison to measured discharge would only be reasonable, if model errors and errors introduced by bias correction are smaller. In our case, we have to conclude that our data sets are not applicable to really quantify the effect of climate change. This will be done in a future study, where we have to improve the bias correction method.

- The proposed bias correction is furthermore a non linear transformation and will destroy the spatial covariance and the temporal autocorrelation of the simulated rainfall field (which is only invariant under linear transformations). For me such a procedure is a quick fix to improve the model, good engineering but no science! It destroys the main advantage of dynamic models i.e. to create rainfall with an auto correlation structure and spatial correlation structure that is consistent with atmospheric physics. This seems to be not wise, and does not necessarily increase the hydrological value of the data.

The two issues, whether the bias correction destroys the spatial covariance/correlation and the temporal autocorrelation of the CCLM precipitation or not, are tested with the following procedure. To address the problem of the spatial covariance/correlation, the time series of every grid point is extracted and Pearson's correlation coefficient is calculated between the extracted time series and all other time series of the remaining grid boxes. The result is a correlation map for every extracted time series with the dimension of the original precipitation field (figure 1 shows an example). With the dimension of x and y equals 20 each, there are 400 correlation maps generated. To answer the question from above, the differences of the fields before and after the bias correction are calculated for an absolute comparison. In one case the whole field is considered and in a second case, only the area ± 5 grid points of the origin of the extracted time series are considered. These absolute differences are then averaged to yield a more compact format.

In addition, two-sampled t-tests are carried out for every grid point, to test if there are differences between the mean correlation coefficients on a 95% significance level. For the calculation, the 400 correlation maps before and after the bias correction are used. Every grid point in the 20x20 domain consists of a vector of 400 correlation coefficients in each map. The t-test now uses these 400 values before the bias correction and compares the mean of them with the mean of the 400 values after the bias correction (H0: mean before = mean after, H1: mean before ≠ mean after, significance level=95%). This procedure is repeated for every grid point, resulting in 400 t-tests.

The results of the difference are shown in figure 2. The left boxplot with the whole domain considered and the right boxplot for the ± 5 grid points region. The example in figure 1 is calculated for the extracted time series in the box [1,1] on the lower most left corner. There the correlation coefficient r=1. Near this origin are the highest values of r, as expected, but there are considerably high values throughout the rest of the domain around r = 0.6-0.7. This is valid for the C20 and the A1B scenario. The mean differences of the C20 correlation coefficients (shown in table 1) range from -2.22*10-16 to + 2.22*10-16 for the whole domain (figure 2: left boxplot) and from -0.034 to 0.039 for the region covered by the origin ± 5 grid points (figure 2: right boxplot). The differences for the A1B scenario are given in the boxplots of figure 3. They range from -2.22*10-16 to + 2.22*10-16 for

the whole domain as well and from -0.38 to 0.05 for the ± 5 grid points region. In summary, there are only minor changes in the spatial correlation caused by the bias correction.

The results of the t-tests are shown in figure 4. The black dots indicate grid boxes with a significant change in correlation terms after the bias correction. From 400 gridboxes there are 78 (20%) grid boxes significant for the C20 scenario and 86 (22%) grid boxes for the A1B scenario.

Table 1: Mean spatial differences of the Pearson correlation coefficients (r after bias correction minus r before bias correction).

	whole domain			origin ± 5 grid boxes		
	min	**max**	**sd**	**min**	**max**	**sd**
C20	$-2.22*10^{-16}$	$2.22*10^{-16}$	1.006114e-16	-0.034	0.039	0.01
A1B	$-2.22*10^{-16}$	$2.22*10^{-16}$	9.987485e-17	-0.38	0.05	0.01

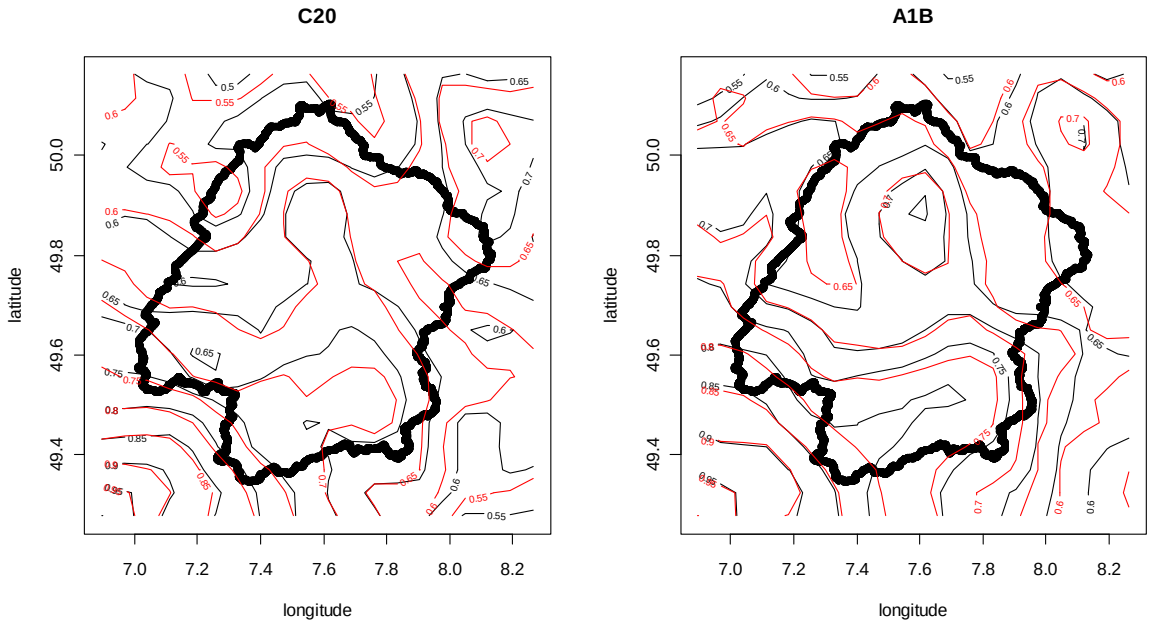

Figure 1: Example of a correlation map for the C20 (left) and the A1B scenario (right) 20x20 grid boxes. Pearson's correlation coefficient is calculated for the extracted time series [1,1] in the lower most left corner and all other time series. The black contours show the correlation coefficients before the bias correction, the red contours after the bias correction. Only for the same grid box where the time series has been extracted results r=1.

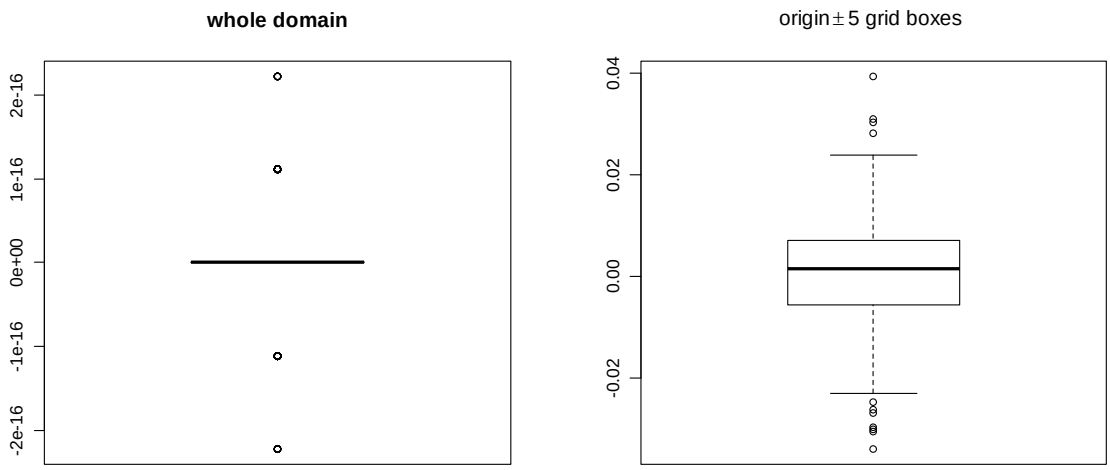

Figure 2: Boxplots of the averaged differences between the C20 correlation maps with Pearsons's rho.

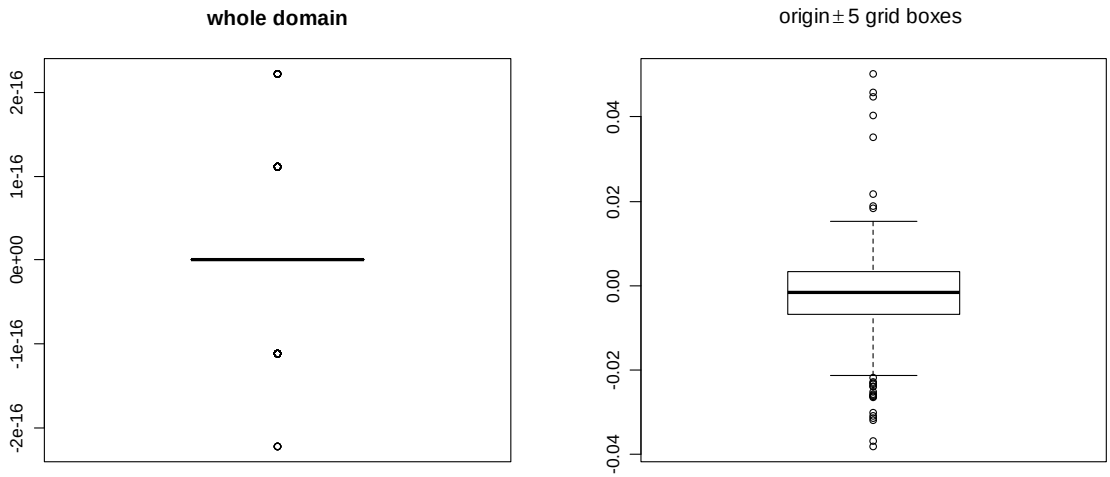

Figure 3: Boxplots of the averaged differences between the A1B correlation maps with Pearsons's rho.

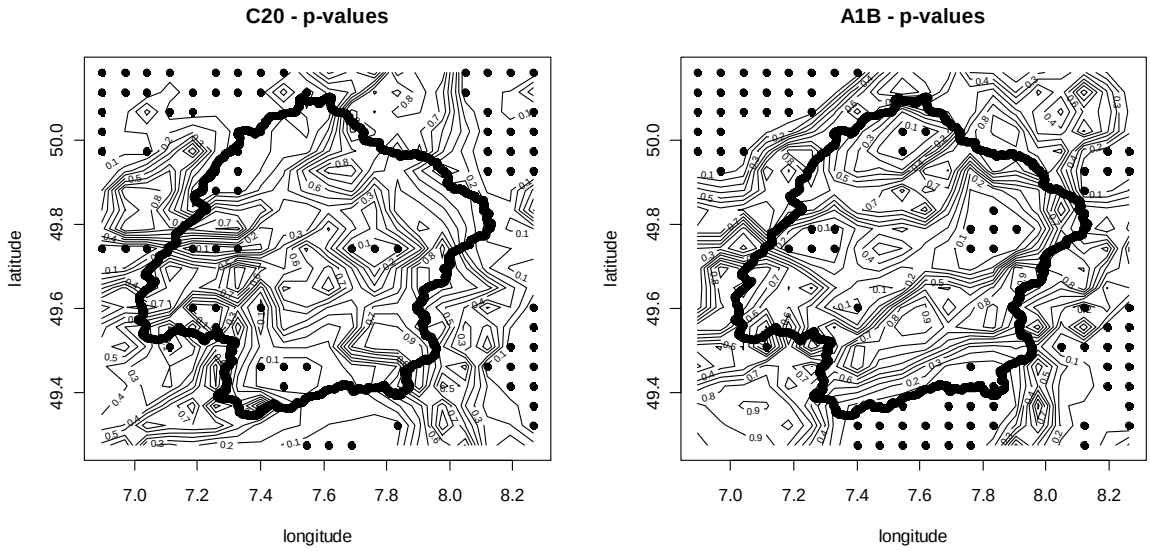

Figure 4: Contour maps of the p-values from the two-sampled t-tests at the 95%-significance level for the C20 (left) and the A1B scenario (right). Black dots mark the grid boxes with significant changes.

Because precipitation is not normal distributed, one of the assumptions of **Pearson's rho** is violated. Therefore, the same procedure is repeated with **Spearman's rho**, which does not assume any specific distribution. In figure 5 the same example of the correlation map is shown as in figure 1 but with Spearman's rho. There are differences in the spatial structure compared to figure 1 and the correlation coefficients are higher throughout the domain. The deviations after the bias correction are of the magnitude -0.1. Averages of the mean deviations are shown in the boxplots in figure 6 and 7. They all show that the differences are close to zero in the average. So far, there are no major differences to the results of the Pearson correlation, but the t-test results support the opposite hypothesis (figure 8). The results of the t-test favor the H1 hypothesis (the means are different) this time for every grid point, although the absolute values of the differences of the means are not very large (as seen from the boxplots). In summary, the spatial correlation is significantly changed for every grid point.

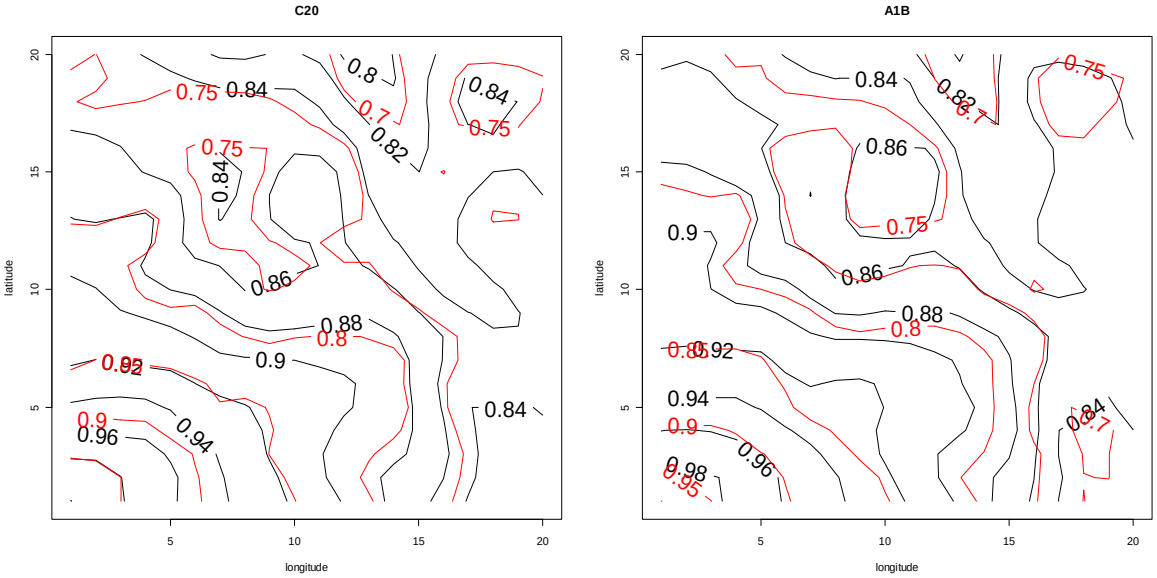

Figure 5: Example of a correlation map for the C20 (left) and the A1B scenario (right) 20x20 grid boxes. Spearman's correlation coefficient is calculated for the extracted time series [1,1] in the lower most left corner and all other time series. The black contours show the correlation coefficients before the bias correction, the red contours after the bias correction. Only for the same grid box where the time series has been extracted results r=1.

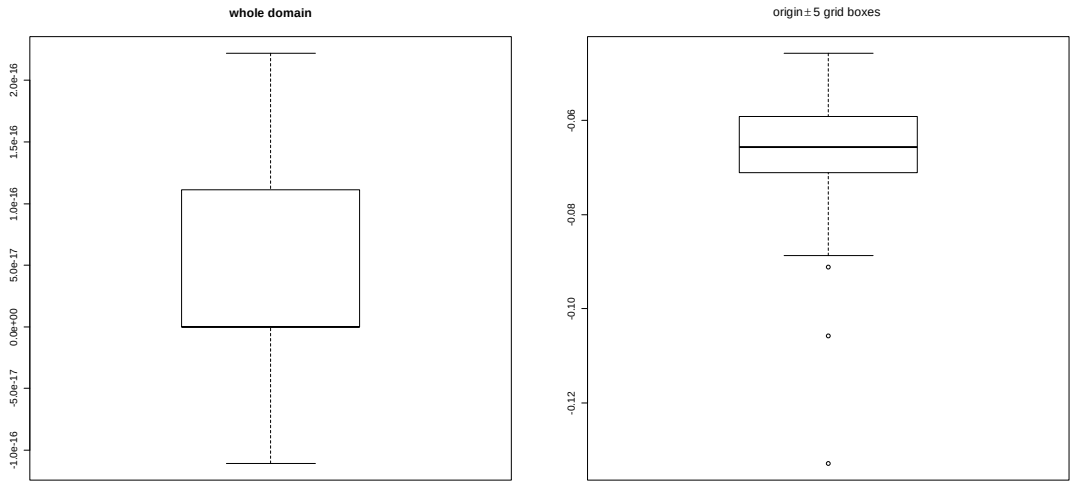

Figure 6: Boxplots of the averaged differences between the C20 correlation maps with Spearman's rho.

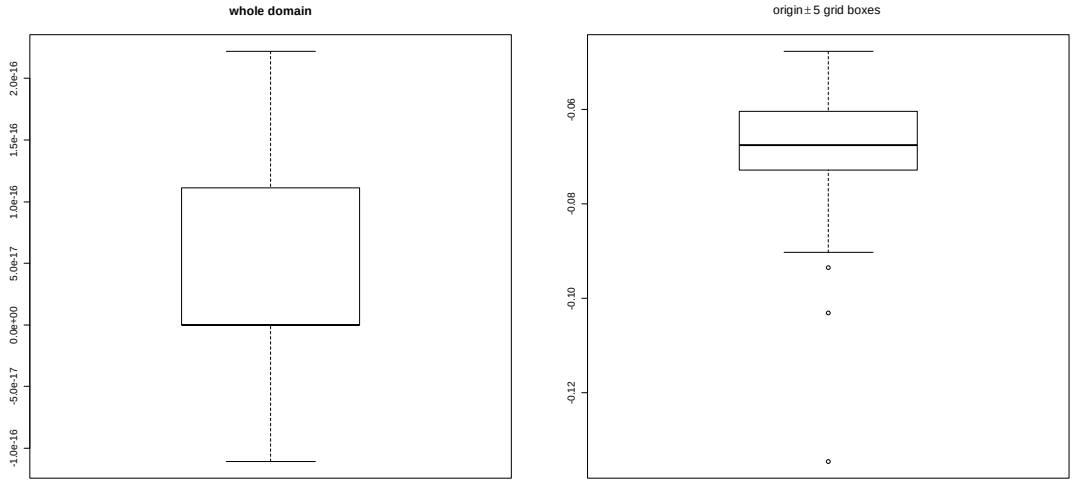

Figure 7: Boxplots of the averaged differences between the A1B correlation maps with Spearman's rho.

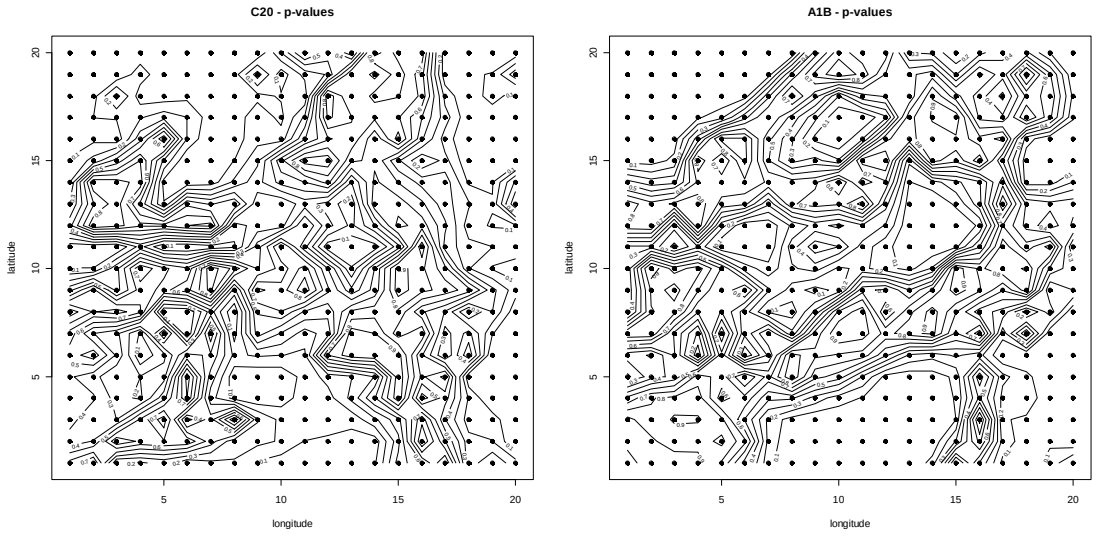

Figure 8: Contour maps of the p-values from the two-sampled t-tests at the 95%-significance level for the C20 (left) and the A1B scenario (right). Black dots mark the grid boxes with significant changes.

Comparison of the temporal auto-correlation

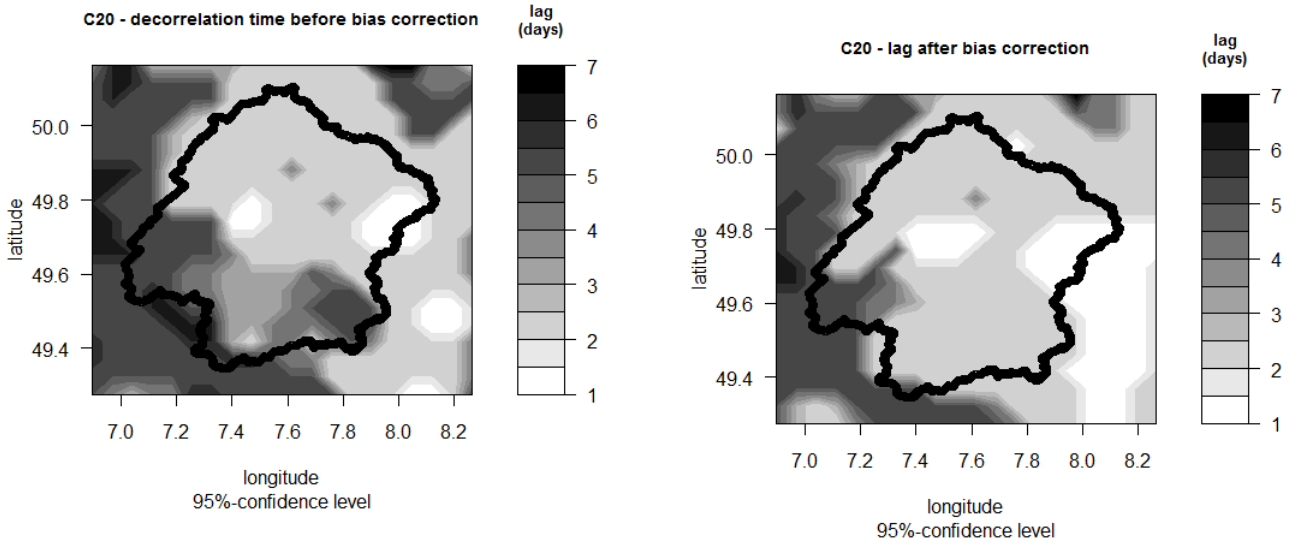

Figure 9: Contour maps of the time lag until the auto-correlation coefficients of precipitation are significant on the 95%-confidence level, the C20 before bias correction (left) and the C20 scenario after bias correction(right). The black contour lines enclosed the Nahe catchment.

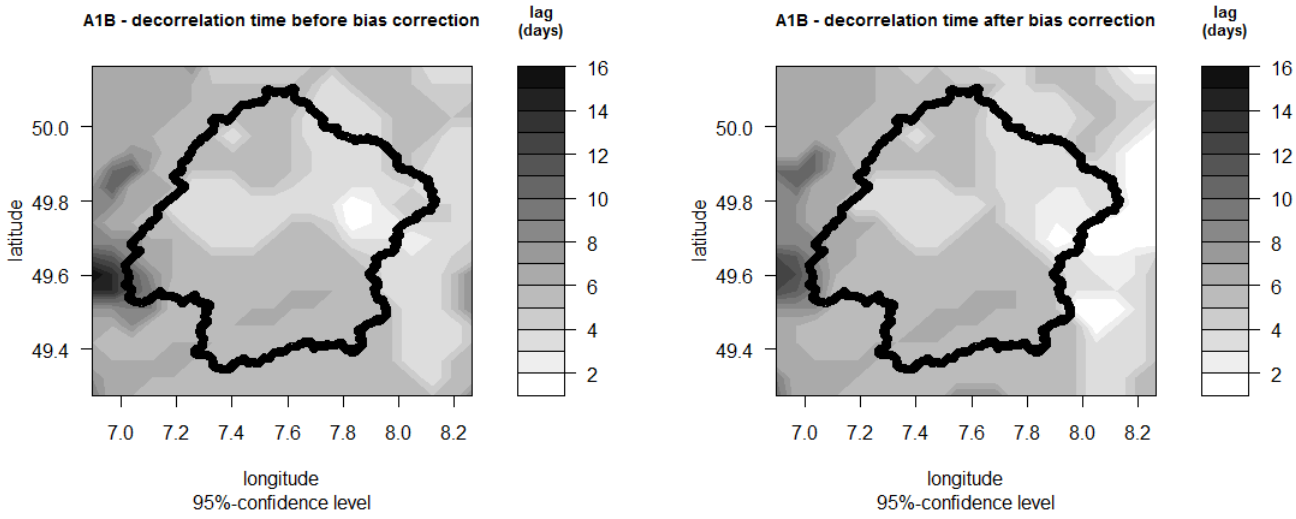

Figure 10: Contour maps of the time lag until the auto-correlation coefficients of precipitation are significant on the 95%-confidence level, the A1B before bias correction (left) and the A1B scenario after bias correction(right The black contour lines enclosed the Nahe catchment.

In figure 9, we show the spatial distribution of the decorrelation time (on the 95%-confidence level) of precipitation for the C20 scenario . The range of the decorrelation time within the Nahe catchment is from 1 to 5 days. The highest decorrelation time occurs in the western and southern part of the catchment while the decorrelation times in the northern and eastern part are only around 1 to 2 days.

After applying the bias correction, the south-western decorrelation times drop from >5 days to 2 days (in the southern part) and reduce more slightly in the western part.

For the A1B scenario there are neither major changes in the spatial distribution nor in the magnitudes (figure 10).

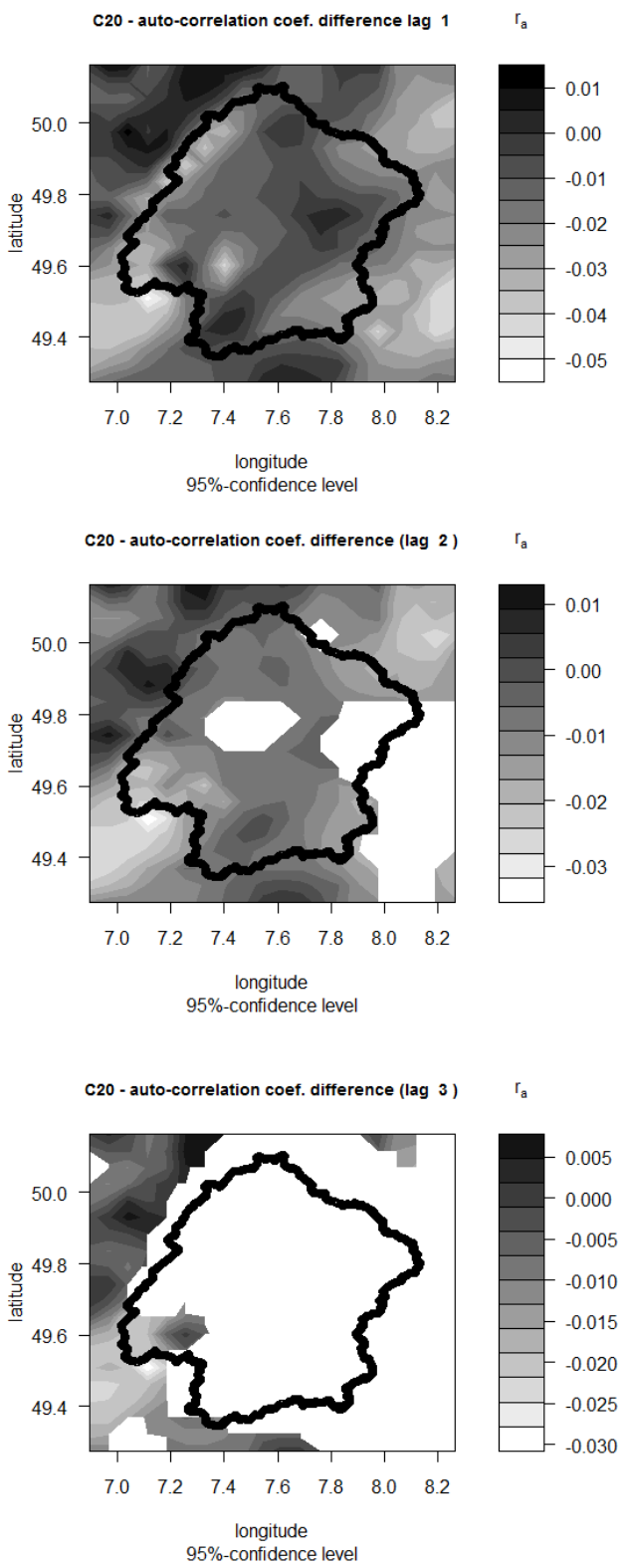

Figure 11: Spatial distribution of the auto-correlation coefficient (r_a) differences (after bias correction minus before bias correction). White color indicates missing values due to non-significance. Top: lag=1 day, middle: lag=2 days, bottom: lag=3 days.

The impact of the bias correction on the auto-correlation coefficients (figure 11) is a minor one, although the differences vary in space. As the magnitudes of the differences of ra indicate, there are only changes of about 1%-5%.

Assuming this bias correction scheme as stationary under change is to me a very unlikely assumption. I see that the authors cannot fully address all these issues within their study. They should however show that they are aware of these shortcomings.

Of course, this assumption can hardly be hold for the future. But the same is true for linear correction method, such as scaling methods, e.g. ratios of means may also alter in the future and may not account for possible (non-linear) dynamic changes in the future (discussed in e.g. Maraun et al. 2010:27). Another limitation is given by the fact that the LandCaRe2020 CCLM runs are not driven with reanalysis-data (ECHAM5 was at the beginning of the nesting chain), which restricts the bias correction to distribution-wise methods (e.g. Maraun et al. 2010:27).

Stationary transfer functions are presumed by many other as well (e.g. vanRoosmalen 2011:141), but of course this is associated with a relatively high degree of uncertainty.

- The authors need to address the problem how to judge when a change in one of their signature indices is significant. What is noise and what is a real fingerprint of change.

This is a good point, but it cannot be fully addressed in this paper, because we mainly wanted to show the high sensitivity of our indices on changes in model input. Especially the indices derived from distribution of ERC depend on a partly subjective event separation procedure, which may lead to a visible change in index pattern. But in contrast, indices from FDC are not very sensitive to noise.

Minor points

- The future paper should clearly stress the scope of the study, which is for me clearly the use of these signatures to address for different purposes as outlined above.

We will clearly focus the paper on the interpretation of our results

- Line 25 what is meant with vertical distribution of soil moisture, infiltration and percolation as processes or the soil moisture profiles? The lateral pattern of soil moisture can be of equal importance.

This only means the distribution of water along the different reservoirs (in the catchment). Our model is a quite simple storage approach with no soil moisture profiles simulated.

- The catchments are not small but lower mesoscale catchments, please comment on apparent soils and provide data on specific runoff behavior (independent of catchment size).

Runoff behavior of the three catchments can be characterized by long term runoff rates and mean runoff coefficients as a result of soils, geology, topology and climate. Kronweiler, a catchment with steep slopes, low water storage capacity of soils and geology, high precipitation and low potential evapotranspiration shows the highest runoff rate and runoff coefficients and a high seasonality of discharge. Catchment Kellenbach with more moderate slopes and a slightly higher field capacity shows a lower runoff rate and runoff coefficients caused by lower MAP. Gensingen, the catchments with the flattest topology and a very high field capacity but low MAP and high ET, has the lowest runoff rate and runoff coefficients of the three catchments with a low seasonality of discharge too.

Table 1: Selected catchment properties

Catchment property		Kronweiler (65 km²)	Kellenbach (362 km²)	Gensingen (196 km²)
MAP (mm/y)		928	673	545
potET (mm/y)		535	540	614
Long term runoff rate (l/s*km²)	year	14.0	7.2	2.3
	winter	23.1	10.9	2.8
	summer	5.2	3.6	1.7
Mean runoff coefficient (93-08)	year	0.23	0.17	0.04
	winter	0.41	0.28	0.08
	summer	0.09	0.07	0.03
Field capacity of total soil column (mm)		288	356	605
Land use		43 % arable land 2.6 % built-up area 54 % forest	58 % arable land 2.5 % built-up area 39 % forest	74 % arable land, orchards, vineyards 5 % built-up area 20 % forest
Soils (FAO85)		Gleyic and humic podzols with cambisols	Gleyic podzols with few cambisols, luvisols and fluvisols at headwaters	Humic podzols and cambisols at upper reaches and luvisols, gleysols and regosols at lower reaches

The first part of the table will be included in the revised manuscript.

- The model description is far too brief to grasp the underlying concepts. It should provide the main concepts and model structure.

The model description will be extended. The text will be changed to: "

"The water balance model LARSIM (Large Area Runoff Simulation Model) allows a continuous process- and area-detailed simulation of the medium-scale mainland water cycle (Ludwig and Bremicker, 2006). In simplistic terms, the watershed is subdivided into 1-D elements linked by a flood routing scheme. To account for sub-grid variability, interception, snow accumulation and melt, evapotranspiration and soil water movement (including runoff generation) are simulated separately for 16 distinct land-use classes within each element. The land-use specific soil column is simulated by a modified form of the Xinanjiang approach (Todini 1996). Runoff concentration within each element is simulated by three parallel linear reservoirs: one for groundwater discharge, one for interflow, and one for direct runoff. Direct runoff comprises fast subsurface runoff and overland flow. Flood routing within the river sections is performed with a kinematic wave approach.The temporal resolution of the water balance calculation is one hour. The model needs as meteorological input spatial fields of precipitation, temperature, air pressure, wind speed, global radiation and relative humidity."

- Bias correction: from a statistical point of view rainfall simulation is a mixed problem, the occurrence of a rainy day is a discrete event (yes or no), given the condition that rainfall occurs, the distribution of rainfall amounts per time step is a continuous. I do not see that this scheme accounts for this mixed nature of the problem. There is, by the way, a huge bunch of literature that suggests alternatives by means of statistical downscaling (compare to the dynamical downscaling of GCM output used) here for instance by Buerger or Bardossy. These alternatives should be clearly acknowledged.

This problem do not occur while using empirical distribution within the quantile matching, because 0mm are part of this distribution in contrast to the parametric gamma distribution. So if the probability of 0mm in the CCLM is for example lower compared to observations, then 0mm remains (normally the case for modeled precipitation). The very low precipitation values from the CCLM (drizzle error) are all corrected to 0mm as well, because their probability is still lower than the probability of 0mm in the observations.

- Is there any spatial structure in the bias between COSMO and observed precipitation time series? If yes the proposed invers distance scheme might be not the best way to interpolate the scheme to all grid points. I miss at least one statement that within the bias correction data at different spatial aggregation levels are compared (jump in scales).

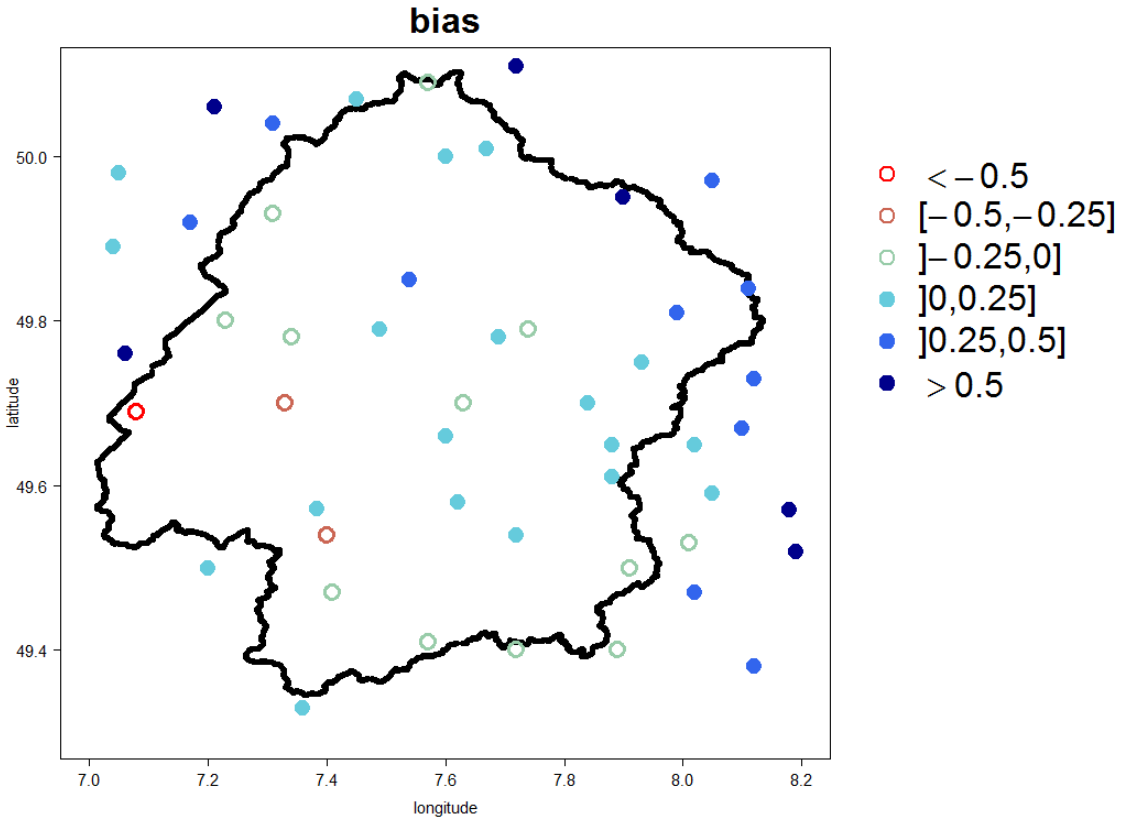

Figure 12: Spatial distribution of the precipitation bias in mm/day. The points are the location of precipitation stations and the colours indicate the magnitude of the bias. Filled points have a positive bias, unfilled points a negative.

The bias of the difference in mean daily precipitation (CCLM minus Station, figure 12) tends to be positive in the northern and eastern part of the Nahe catchment and rather negative in the western (Hunsrück) and negative in the southern region (Palatinate Forest). Thus, the assumption of homogeneity within the spatial distribution of the bias is violated. A linear interpolation is therefore maybe not appropriate. It seems that the bias depends on orography (negative in western Hunsrück and Palatinate Forest in the south, positive in the flatter north-east area); however, the bias is positive in the northern Hunsrück. Considering the sub-catchments Gensingen (east), Kellenbach (north) and Kronweiler (west), only Kronweiler may be affected by the assumption of homogeneity. The bias changes sign at the vicinity of the Kronweiler catchment, whereas it is more or less constant (positive) for the other two catchments.

- Please briefly provide details how the end point of the event is defined.

p. 3579, l. 3-5: Text will be changed to:

"Second, to define events, each runoff time series was screened starting from the largest peak flow and proceeding to the second largest peak flow and so forth according to the following method. A peak flow was assumed to be the peak flow of a potential event, if the ratio of direct runoff to baseflow at time of the peak was larger than 2 and there was no larger flow in the previous and following 12 h. For each peak flow, the start and end of an event was searched within a given time period by finding the time where the direct runoff becomes lower than a given threshold, which depends on the direct runoff at the time of the peak flow. If no starting point was found, the search was repeated by gradually increasing the time period and the threshold. With this iterative approach, the direct runoff at the beginning and end of an event is as small as possible (Merz et al. 2006, Norbiato et al. 2009)."

- Notation and indices in Eq. 10 should be explained.

BiasFDCMidslope: percent bias in slope of the mid-segment:

$$BiasFDCmidslope = \frac{\left(\log\left(FDC_{1,0.2}\right) - \log\left(FDC_{1,0.7}\right)\right) - \left(\log\left(FDC_{2,0.2}\right) - \log\left(FDC_{2,0.7}\right)\right)}{\left(\log\left(FDC_{2,0.2}\right) - \log\left(FDC2_{1,0.7}\right)\right)} \cdot 100,$$

where $FDC_{i,p}$ is the runoff with exceedance probability *p* of FDC number *i* (red and blue triangles in Fig. 2), e.g. $FDC_{1,0.2}$ is the runoff with exceedance probability p=0.2 of FDC number 1.

- The Bias FHV is defined based on the difference in integral upper two percent. Please explain the choice.

Because of the short time series (10 years), it is not possible to apply extreme value statistics. But, by evaluating the upper 2% of all values we focus on the rainfall driven periods with high discharges without being too sensitive on single extreme peaks.

References

Ludwig, K., Bremicker, M. (Eds.): The Water Balance Model LARSIM. Freiburger Schriften zur Hydrologie, Band 22 Institut für Hydrologie der Universität Freiburg, 2006.

Merz, R., Blöschl, G. and Parajka, J.: Spatio-temporal variability of event runoff coefficients. J. Hydrol.331, 591-604, 2006.

Norbiato, D., Borga, M., Merz, R., Blöschl, G. and Carton, A.: Controls on event runoff coefficients in the eastern Italian Alps. J. Hydrol.375, 312–325, 2009.

Todini, E. (1996) The ARNO rainfall-runoff model. J. Hydrol. 175, 339–385.

van Roosmalen, L., Sonnenborg, T.O., Christensen, J.H.: Comparison of hydrological simulations of climate change using perturbation of observation and distribution-based scaling. Vadose Zone Journal, 10(1), 136-150. doi: 10.2136/vzj2010.0112, 2011.