Technical note: Changes of cross- and auto-dependence structures in climate projections of daily precipitation and their sensitivity to outliers

Jan Hnilica[1, 2], Martin Hanel[2] and Vladimír Puš[2]

[1] The Czech Academy of Sciences, Institute of Hydrodynamics, Pod Paťankou 5, Prague 6, 166 12, Czech Republic

5 [2] Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Prague 6 – Suchdol, 165 21, Czech Republic

Correspondence to: Jan Hnilica (hnilica@ih.cas.cz)

Abstract. Simulations of regional or global climate models are often used for climate change impact assessment. To eliminate systematic errors, which are inherent to all climate model simulations, a number of post processing (statistical

10 downscaling) methods have been proposed recently. In addition to basic statistical properties of simulated variables, some of these methods consider also a dependence structure between or within variables. In the present paper we assess the changes in cross- and auto-correlation structures of daily precipitation in six regional climate model simulations. In addition the effect of outliers is explored making a distinction between ordinary outliers (i.e. values exceptionally small or large) and dependence outliers (values deviating from dependence structures). It is demonstrated that correlation estimates can be strongly

15 influenced by few outliers even in large data sets. In turn, any statistical downscaling method relying on sample correlation can therefore provide misleading results. An exploratory procedure is proposed to detect the dependence outliers in multi-variate data and to quantify their impact on correlation structures.

1 Introduction

The investigation of climate change impact on the hydrological cycle is one of the crucial topics in the field of water resources

20 management and planning (Mehrotra and Sharma, 2015). Simulations of regional and global climate models (RCMs and GCMs) represent a fundamental data source for climate change impact studies. It is well known that raw climate model outputs cannot be used directly in impact studies due to inherent biases which are found even for basic statistical properties (Chen et al., 2015). The bias is caused primarily by simplified representation of important physical processes (Solomon et al., 2007), which is often resulting from low spatial resolution of the RCMs.

25 Therefore, many methods have been developed to post-process the climate model outputs in order to move their statistical indicators closer to observations. An overview of these methods is presented e.g. by Maraun et al. (2010). Precipitation is a key input into hydrological climate change impact studies and at the same time it belongs to meteorological variables that are most affected by bias. The comparison of correction methods commonly used for precipitation data is provided by Teutschbein and Seibert (2012). Nevertheless, these standard methods correct only the bias in statistical indicators (mean,

30 variance, distribution function) of individual variables. The bias in persistence parameters of time series as well as the bias in

cross-dependence structures between variables is often neglected. However, the dependence structures of the meteorological variables affect the hydrological response of a catchment (Bárdossy and Pegram, 2012), thus their inadequate representation in the data can impair hydrological impact studies (Teng et al., 2015; Hanel et al., 2017).

In recent years several studies attempted to overcome this limitation. Hoffmann and Rath (2012) and Piani and Haerter (2012) focused on the relationship between precipitation and temperature data from single location. Bárdossy and Pegram (2012) developed two procedures correcting a spatial correlation structure of RCM precipitation. Mao et al. (2015) proposed a stochastic multivariate procedure based on copulas. Johnson and Sharma (2012) developed procedure correcting common statistics (mean, variance) together with lag-1 autocorrelation in multiple-time scales. The procedure was later extended with recursive approach by Mehrotra and Sharma (2015) and subsequently with non-parametric quantile mapping by Mehrotra and Sharma (2016) to correct the bias in auto- and cross-dependence structures across multiple time scales. An approach based on the principal components was presented by Hnilica et al. (2017), correcting bias in cross-covariance and cross-correlation structures.

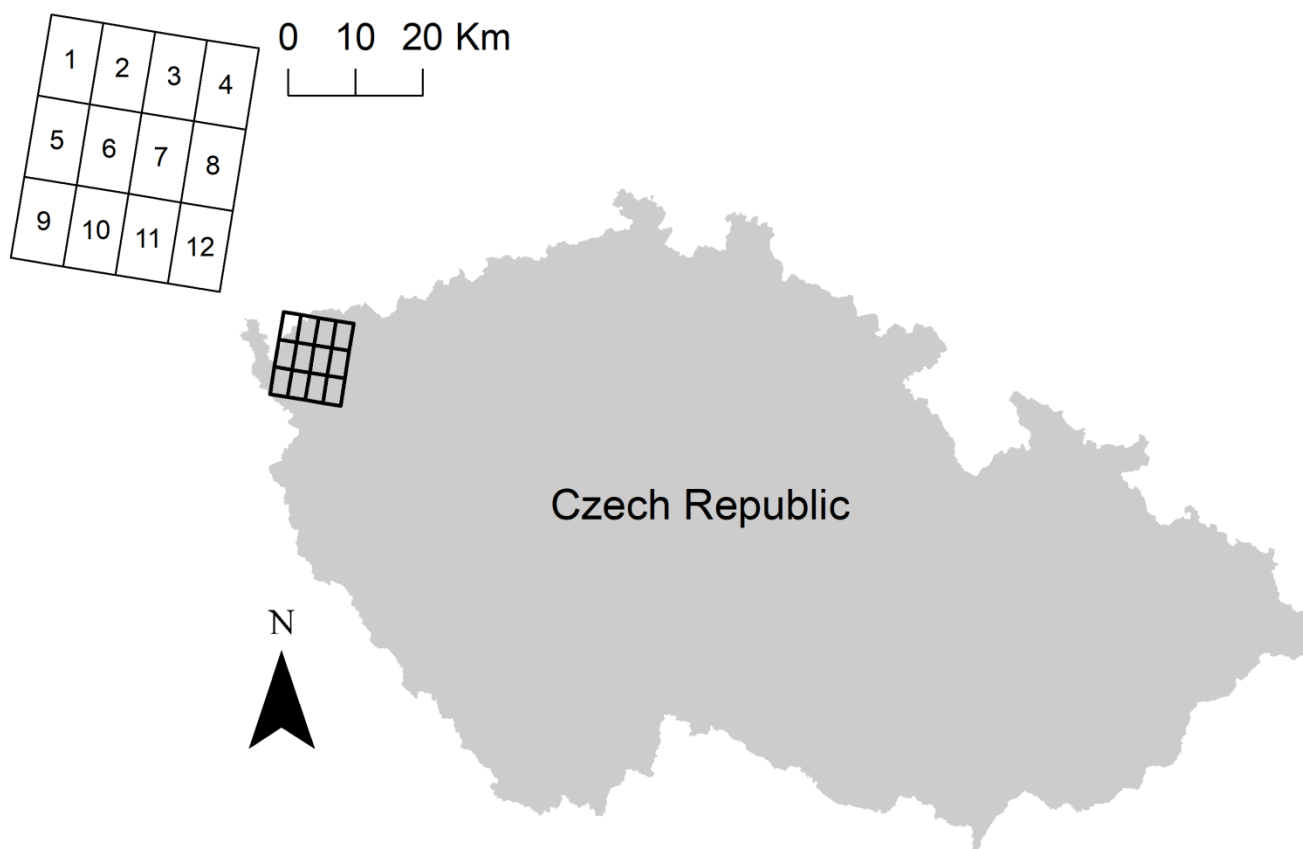
This study is focused on a temporal stability of dependence structures. We evaluate the temporal changes in cross-and auto-correlation structures in multivariate precipitation data simulated by an ensemble of climate models. We further investigate whether the magnitude of the changes exceeds considerably the natural variability. The attention is finally paid to the effect of outlying values, which can significantly affect the correlations and can thus lead to artefacts in bias-corrected time series.

The paper is organised as follows. In Sect. 2 the data used in this study are presented and Sect. 3 describes the methodology. In Sect. 4 the results are reported and in Sect. 5 their consequences for climate changes impact studies are discussed.

2 Data and study area

The daily precipitation sums from six EURO-CORDEX (Giorgi et al., 2009) regional climate models were considered. The ensemble of models was composed of two RCMs (CCLM, RCA) driven by three GCMs (EC_EARTH, HadGEM2-ES and MPI-ESM-LR); see Table 1 for the overview. The simulations with 0.11 degree spatial resolution forced by RCP8.5 were used. The data from twelve model grid-boxes located in the western part of the Czech Republic were analysed, see Fig. 1 for the details of the area. The control period spans the years from 1971 to 2000, the future period the years from 2051 to 2080.

Table 1. Global and regional climate models used in the present study.

GCM	RCM	ID
EC-EARTH	CCLM-4-8-17	1A
	RCA4	1B
HadGEM2-ES	CCLM-4-8-17	2A
	RCA4	2B
MPI-ESM-LR	CCLM-4-8-17	3A
	RCA4	3B

60   **Figure 1. Location of the considered grid-boxes in the Czech Republic.**

3 Methods

The wet and dry periods were treated separately in this study. The cross-correlations were calculated in two stages. Firstly the binary cross-correlations were calculated to assess the correspondence of wet/dry periods, using the time series with the values replaced by 0 (dry day) or by 1 (wet day). In the second stage the cross-correlations of overlapping wet periods were calculated. The auto correlations were analysed through the lag-1 auto-correlation coefficient, where only the non-zero pairs of neighbouring values x_i and x_{i+1} were considered.

The individual grid-boxes were labelled by numbers 1-12, as shown by labels in Fig. 1. The cross-correlation between the grid-boxes i and j is denoted as $r_{i,j}$. The symbol **R** denotes the correlation matrix (i.e. the square matrix with elements $r_{i,j}$). The lag-1 auto-correlation from grid-box i is denoted as r_i^1. If appropriate, the subscripts denoting the grid-boxes are omitted for clarity.

The changes of correlation coefficients were calculated as

$$\Delta r = r_F - r_C \tag{1}$$

where Δr denotes the change of r (cross- or auto-correlation), subscripts F and C denote the future and control periods, respectively.

The sampling variability of individual cross- and auto-correlation was investigated to assess the statistical significance of their changes. The confidence intervals were derived using the block bootstrap approach (Davison and Hinkley, 1997). Specifically, the confidence interval around the correlation $r_{i,j}$ was obtained as follows:

1. one-year blocks from the time series for basins i and j were randomly selected with replacement (30 times to obtain the same sample size as the original data), subsequently the correlation of the 30-year sample was calculated

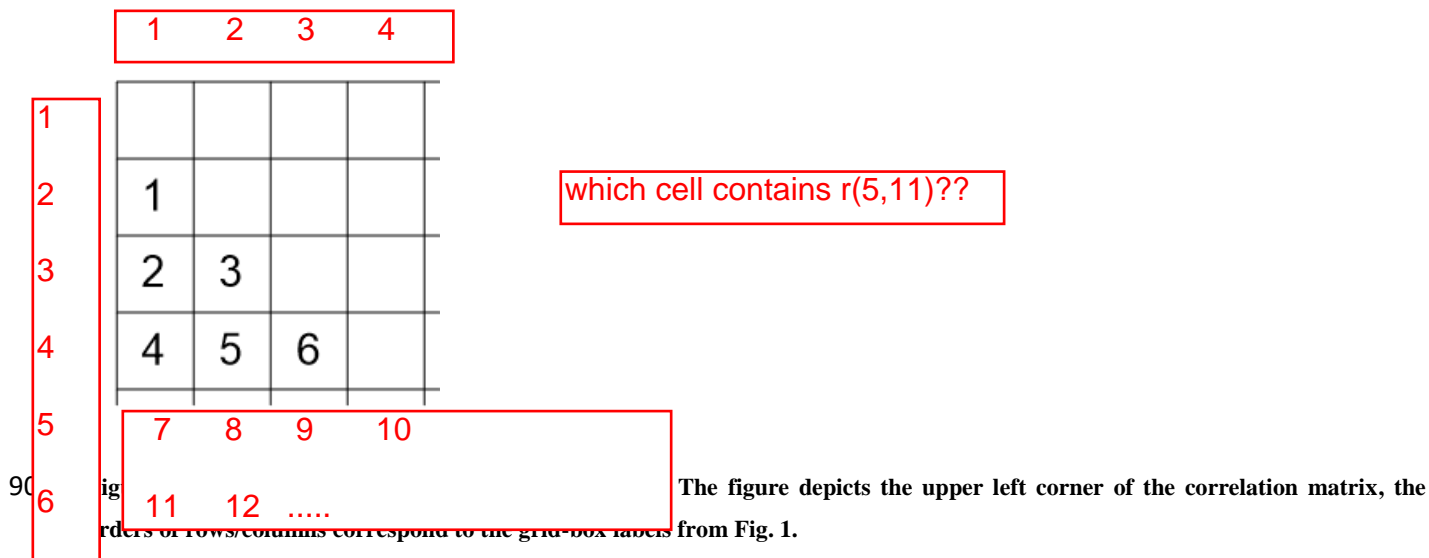
2. step 1 was repeated 1000 times

3. the 95% confidence interval was derived as a range between the 0.025 and 0.975 quantiles of the resampled correlations.

The block approach was chosen to preserve seasonal variability in the bootstrap samples. For the presentation of confidence intervals, the unique identifier (ID) was assigned to each pair of grid-boxes, the numbering was done according to rows of correlation matrix; the scheme is depicted in Fig. 2. The confidence intervals for auto-correlation were derived in the same way using one-year blocks of time series. The joints of the adjacent blocks were not included in the calculation.

I do not understand this sentence; please reword it.

which cell contains r(5,11)??

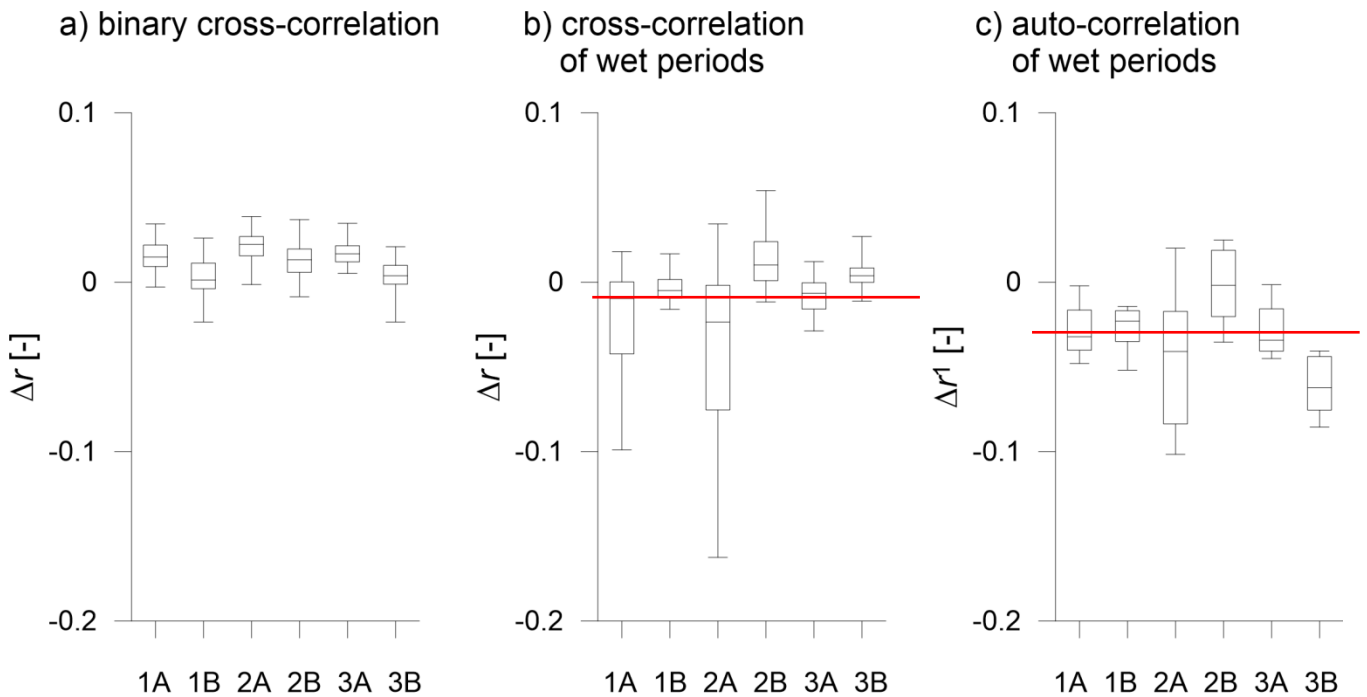
90 Fig ... The figure depicts the upper left corner of the correlation matrix, the orders of rows/columns correspond to the grid-box labels from Fig. 1.

4 Results

4.1 Changes in correlation structures

95 In the case of 12-dimensional data, the change of the cross-correlation structure consists of changes in $r_{i,j}$ coefficients for 66 pairs of grid-boxes (corresponding to the sub-diagonal part of the correlation matrix). For clarity, these 66 changes are presented in the form of box-plots for individual models.

of the medians?

Figures 3a and 3b present the changes in the binary cross-correlations and in the cross-correlations of wet periods, respectively. As seen from the figures, the binary correlations are relatively stable; their changes range approximately from -
100 0.02 to 0.03. Therefore, the correspondence of wet/dry periods between individual grid-boxes remains similar in the control and future periods. The correlations of wet periods change more substantially, the changes range from -0.16 to 0.05. Nevertheless, there are strong differences between individual models, the models 1A and 2A reach noticeably higher changes than other models. Figure 3c presents the changes of lag-1 auto-correlations; the box-plots for individual models are compiled from 12 changes in time series from individual grid-boxes. The changes range from -0.1 to 0.025, the widest range
105 of changes is reached by the model 2A. The maximal relative changes in cross-correlation reach up to 18% of the value from the control period, in the case of auto-correlation it is almost 45%. Note that the auto-correlations are in general markedly lower than cross-correlations (the mean cross-correlation of individual models exceeded 0.8; the mean lag-1 auto-correlation is around 0.23).

surely these values are 0.08 and 0.023??
see my estimates in Figures 3 (b) & (c)

110   **Figure 3. Overview of the changes in correlation structures for all models: (a) the changes of binary cross-correlations, (b) the changes of cross-correlations of overlapping wet periods, (c) the changes of lag-1 auto-correlations.**

The significance of the changes in wet-periods correlations was assessed using a block bootstrap. Figure 4 presents the 95% confidence intervals of individual cross-correlations for all models. The blue dividers identify the successive rows below the

115   diagonal in the correlation matrix. In general, the majority of changes show a little significance, the intervals from control and future periods overlap considerably. Figure 5 shows the same for lag-1 auto-correlations of individual grid-boxes. Also in this case the majority of changes do not exceed the sampling variability; the most significant changes are reached by the model 3B. but the overall trend is a drop in the future

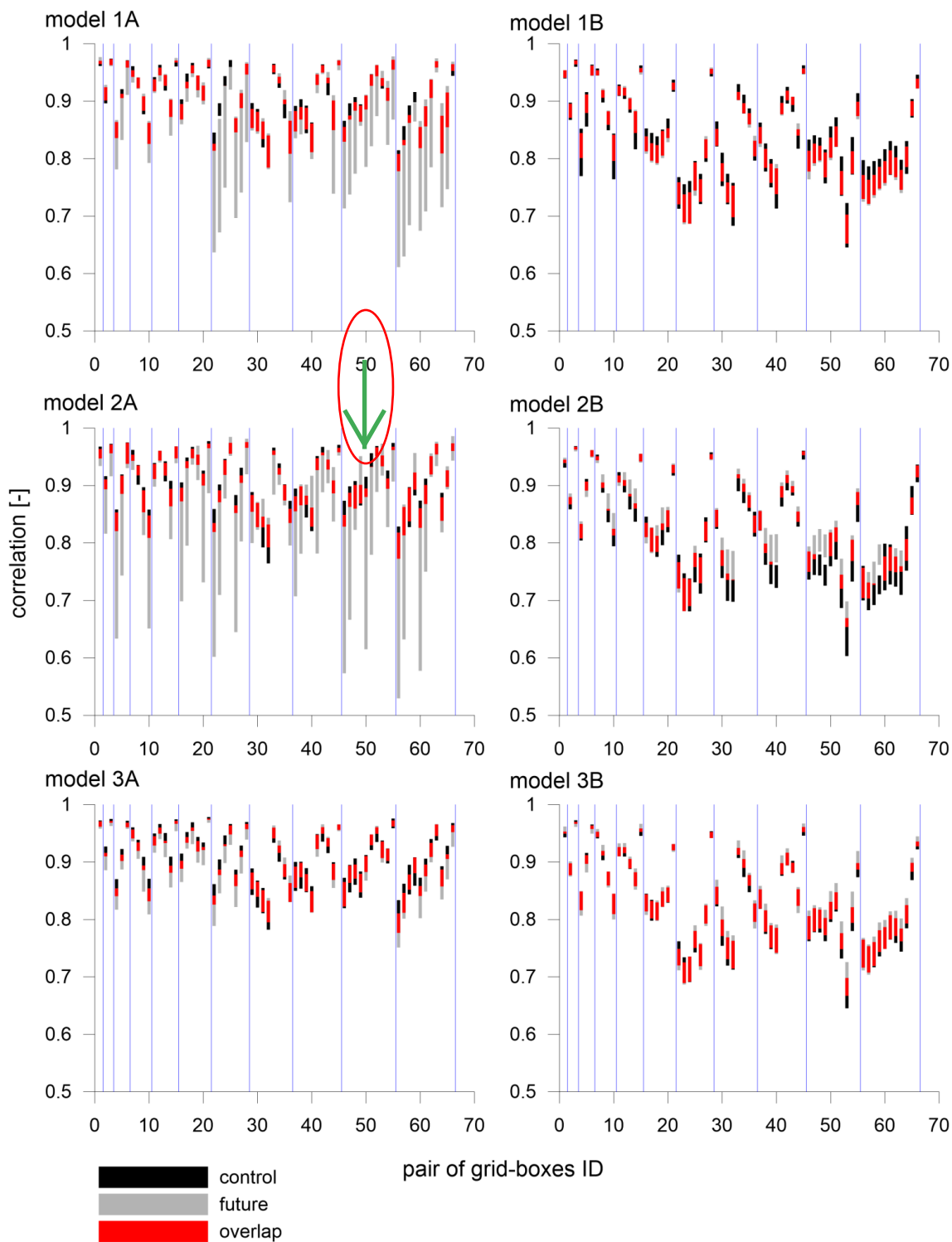
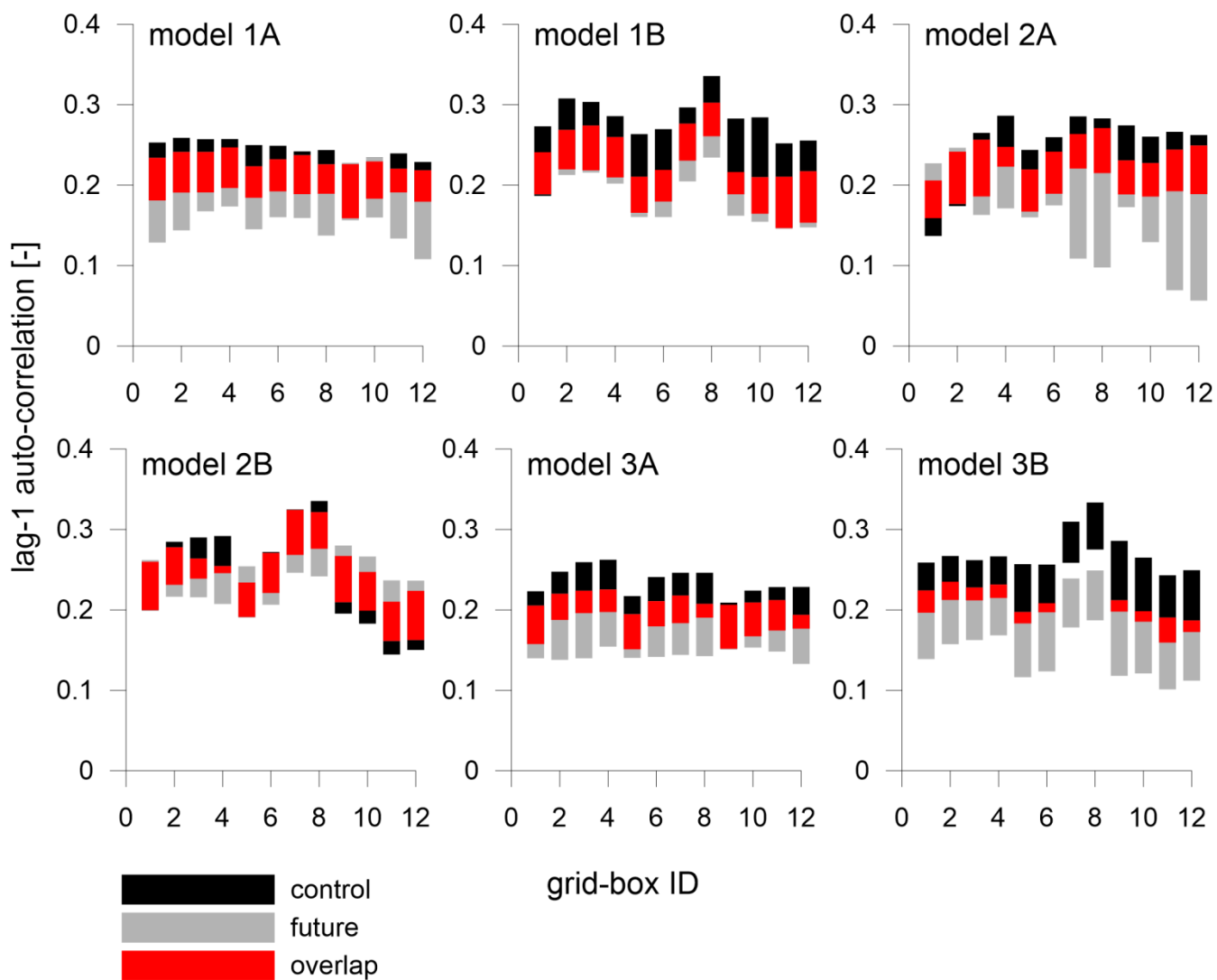
except models 1A and 2A??

120   **Figure 4. The 95% confidence intervals of the cross-correlation for overlapping wet periods for all models. The identifiers (ID) of individual pairs of grid-boxes are explained in Fig. 2. The blue lines separate successive rows below the diagonal in the correlation matrix.**

125

Figure 5. The 95% confidence intervals of the lag-1 auto-correlation for all models.

Hydrol. Earth Syst. S...
Manuscript under rev...
Discussion started: 13...
© Author(s) 2018. C...

4.2 Effect of outliers

The previous section demonstrated that in some cases the changes in cross-correlation show a little significance despite of
130   their high absolute values, which is particularly related to models 1A and 2A. At the same time, it can be seen in Fig. 4 that
some confidence intervals for these models are exceptionally wide. Further analyses showed that this instability of
correlation estimates is introduced by outlying values, which causes seeming changes of the correlation structures.

In the simulation of the model 2A, $r_{5,11}$ decreased from 0.90 in the control period to 0.73 in the future period. Figure 6a
depicts the data from the future period (values from the grid-box 5 plotted against values from the grid-box 11, the data with
135   any zero values are excluded). The decrease is in large part caused by one outlying point, which is circled in the plot. Its
removal from the data increases the correlation in the future period to 0.86, which markedly reduces the change. On the other
hand, high values do not necessarily affect the correlation, as seen in Fig. 6b, where the data from grid-boxes 11 and 12 are
plotted (again the model 2A, future period). The circled outlier does not affect the correlation in this case, since the location
of the point is in accordance with the configuration of data – the point lies approximately in a direction of a potential
140   regression line.

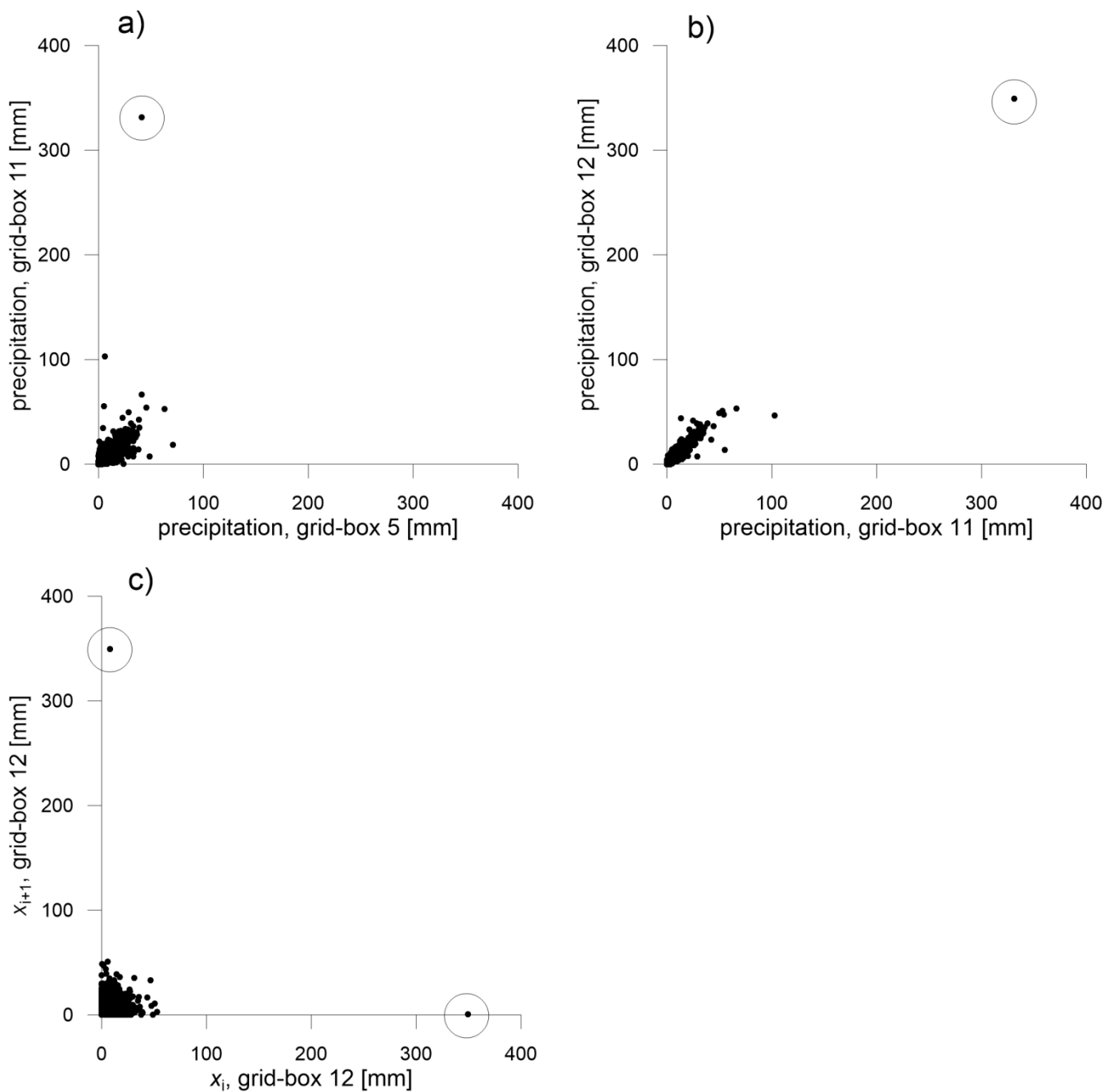
Outlying values affect also the auto-correlation. The largest change of the auto-correlation was achieved by the model 2A,
where r_{12}^1 decreased from 0.23 in the control period to 0.12 in the future period. This decrease in caused by the outlier 349.4
mm in the future data; this extraordinary value was simulated by the model 2A for the 8th May 2080. Figure 6c depicts the
data for the calculation of r_{12}^1, i.e. the values x_i plotted against the values x_{i+1}, where i denotes the order of the value x in the
145   time series. The outlier is employed twice within the calculation (as x_i and as x_{i+1}, circled values in Fig. 6c) which markedly
affects the result. If the outlier is removed from the time series, r_{12}^1 increases from 0.12 to 0.22, which reduces the change
almost to zero. The calculation of other members of the auto-correlation function is affected by the outlier in the same way.
We note, that the effect of an outlier to the auto-correlation strongly depends on values, by which the outlier is surrounded in
the time series. The presence of noticeable outlier thus makes the calculation of the auto-correlation very unstable.

150

Figure 6. The effect of outliers on correlation structures of model 2A in the future period, the outliers are circled: (a) the outlier strongly affecting the correlation, (b) the outlier with no effect on the correlation, (c) the outlier affecting the calculation of r_{12}^1.

cross-

the serial correlation

Discussion started
© Author(s) 2018

4.3 Detection of outliers

155   The examples showed that outliers can distort cross- and auto-correlation structures of a large dataset comprising many thousands values. Nevertheless, it should be realized that not each extreme value necessarily affects the correlation (as seen in Fig. 6b). Therefore, a more specific concept of outliers is presented in this study. Values deviating from the correlation structure are denoted as *dependence* outliers. As well as ordinary outliers, the dependence outliers are values at a long distance from the origin; nevertheless, the difference between them and ordinary outliers consists in a coordinate system in

160   which the distance is measured. Figure 7 illustrates this by an example of synthetic 2-dimensional data. The dashed lines and coordinates in square brackets define the standard (canonical) coordinate system. The ordinary outliers are points in a long distance from the origin [0, 0], measured in the standard coordinates; the point A represents an example. The solid lines and coordinates in round brackets define an alternative coordinate system, which reflects the intensity of linear dependence between the variables X and Y. The dependence outliers are points in a long distance from the origin (0, 0), measured in the

165   alternative coordinates; the point B represents an example. Let us remark that the point B does not represent an extreme value neither in X nor Y data, in contrast to the point A. Nevertheless, the point B deviates from the dependence structure, which comes out when its distance from (0, 0) in alternative coordinates is calculated. The alternative coordinate system is constructed through the covariance matrix of the data. The directions of the axes are given by the eigenvectors of the matrix, the lengths of unit vectors are given by the square root of the corresponding eigenvalues and the origin is located in the mean

170   of the data. The construction of the system is related to the principal component analysis; see for example Wilks (2011) for details.

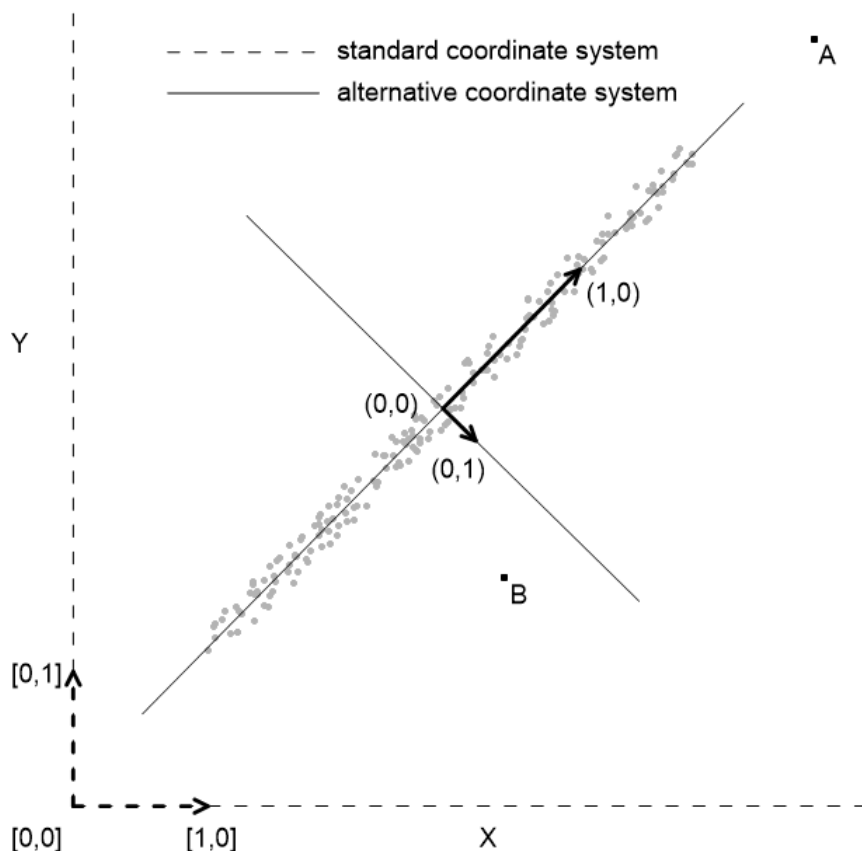



Figure 7. The difference between the ordinary and dependence outliers. The dashed lines define the standard coordinate system; the solid lines define an alternative coordinate system. The points outlying in the standard coordinates are ordinary outliers (point A); the points outlying in the alternative coordinates are denoted as dependence outliers (point deviating from the dependence structure, point B). The construction of alternative coordinates is explained in the text.

The problem is that the presence of outliers is not simply detectable from the changes of dependence structures. It can be indicated ~~either~~ indirectly from the analysis of sampling variability; nevertheless the wide confidence intervals do not necessarily imply the presence of outliers. Or alternatively, it can be found out when the individual pairs of datasets are visually checked. We propose a procedure allowing for identification of significant dependence outliers and assessment of their effect on correlation structure. The procedure consists of three steps:

1. the most outlying (multi-variate) value is found in the data

2. the value is removed from the data and a new correlation matrix is calculated

3. a difference between the new and the previous correlation matrix is calculated and recorded.

185   These three steps are repeated. The difference in step 3 is quantified through

$$\delta\mathbf{R}_m = \|\mathbf{R}_m - \mathbf{R}_{m-1}\| \qquad\qquad (2)$$

where $\mathbf{R}_m$ denotes the correlation matrix of the data from which m largest outliers were removed, the symbol $\|\cdot\|$ denotes the Frobenius matrix norm. The most outlying value in the step 1 is simply defined as the value with the highest distance from origin (measured in the alternative coordinates). A result of the proposed exploratory procedure is a sequence of $\delta\mathbf{R}_m$, which

190   clearly indicates the presence of noticeable outliers. We note that the alternative coordinate system in which the dependence outliers are searched is data-dependent (in contrast to the standard coordinates). It means that after each outlier removal the alternative coordinates slightly change and must be recalculated to correspond to the actual data.

The procedure is demonstrated on two simple 2-dimensional examples. Figure 8a depicts the sequence of $\delta\mathbf{R}_m$ for the data from Fig. 6a. A massive impact of the first outlier is clearly visible; the removal of next outliers already does not affect the

195   correlation matrix substantially (the first member $\delta\mathbf{R}_1$ corresponds to the circled outlier in Fig. 6a). Figure 8b depicts the same for the data from Fig. 6b; a gradual evolution of $\delta\mathbf{R}_m$ indicates that the data do not contain noticeable (dependence) outliers.

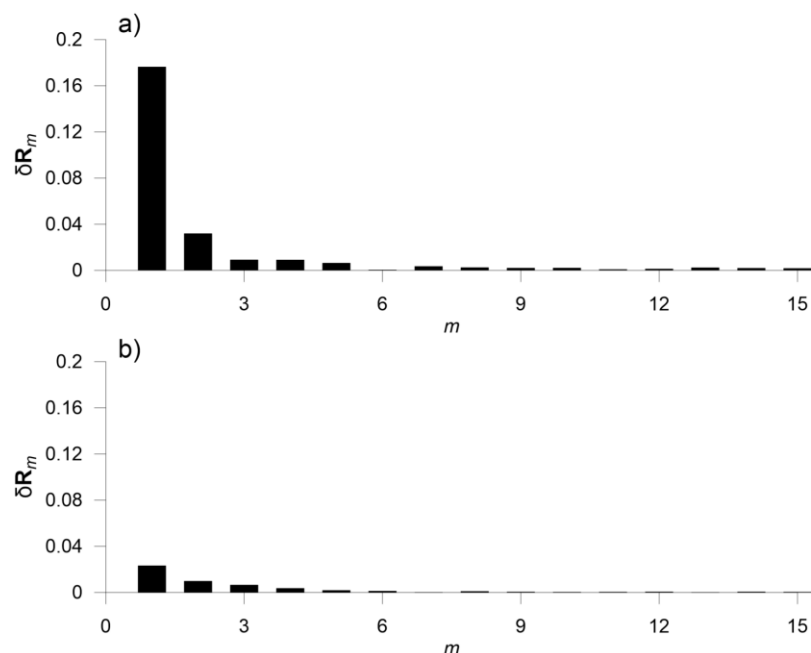



Figure 8. The demonstration of the exploratory procedure: a) the detection of dependence outliers for 2-dimensional data from
200   **Fig. 6a – the plot of $\delta\mathbf{R}_m$ indicates a noticeable outlier in the data, b) the same for the data from Fig. 6b – a gradual evolution of**
$\delta\mathbf{R}_m$ indicates that data do not contain dependence outliers.

usefulness of the procedure is   set of

A real utility of the procedure consists in a fact, that a large multi-variate data can be explored as a whole. The n-dimensional outliers can be searched in the same way as the 2-dimensional outliers in the examples presented above. A result of the procedure is always a one-dimensional plot of $\delta\mathbf{R}_m$, regardless of a dimension of input dataset. Figure 9 shows the plots of

205   $\delta\mathbf{R}_m$ for the complete 12-dimensional data from the future period for all models. The strong outliers in data from 1A and 2A are simply detectable from the plots. Generally, a plot of $\delta\mathbf{R}_m$ enables a simple assessment of the internal structure of the data and a direct evaluation of an importance of individual outliers.

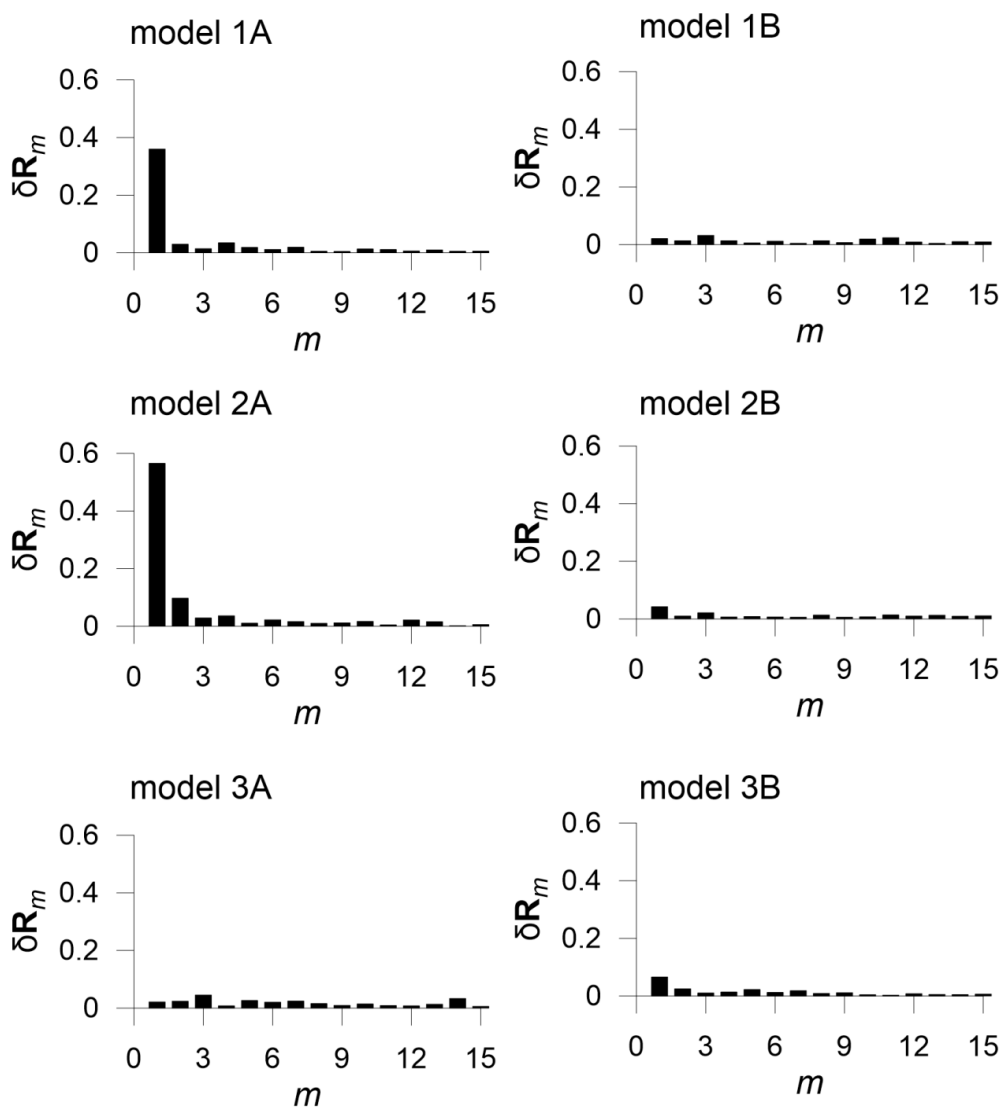



Figure 9. The detection of dependence outliers for complete 12-dimensional data from all models in the future period. The strong

210   outliers in data from the models 1A and 2A are clearly distinguishable.

5 Conclusions

The examples demonstrated that outliers can strongly affect the cross- and auto-correlation structures of the data comprising many thousands of values. In general, it must be stressed that the presence of outliers cannot be considered as a bias. The extreme precipitation values as well as the dependence outliers naturally occur. Nevertheless, although the dependence structures are markedly influenced by a small number of outliers, they characterize the data as a whole. Therefore a substantial bias can arise when the data with noticeable outliers are used to assess the dependence structures or when their dependence structures are used e.g. for calibration of the bias correction functions. The cross- and auto-correlation structures are the key ingredients in several multi-variate bias correction methods, for example in Mehrotra and Sharma (2015) and Mehrotra and Sharma (2016). The results based on these methods can be devalued by outliers. The appropriate tool for testing the presence of outliers is the analysis $\delta\mathbf{R}_m$ of the difference between the new and the previous correlation matrix (Eq 2), presented above; the exploratory procedure can be automatized and included in the modelling chain as a pre-processing step to automatically remove at least the most noticeable outliers.

The analysis of significance showed, that in most cases the correlations are stable in time, their changes are insignificant and caused by outlying values. Therefore the climate projection can be interpreted as a linear transformation of an initial state, because a nonlinear transformation would change the correlations substantially. From this point of view a reasonable scenario of future precipitation can be obtained by the corresponding linear transformation of observations, i.e. by the multiplicative delta method (Déqué, 2007). Such an approach avoids the problems of complex bias correction methods (e.g. their increasing complexity and unclear effect on climate change signal), which are recently the subjects of serious criticism, for example by Ehret et al. (2012) or Maraun et al. (2017).

Data and code availability. The RCM data, the source codes and the plot data are available online at https://doi.org/10.5281/zenodo.1407992, allowing to generate all results and to reproduce all plots.

Author contributions. JH an MH designed the study and wrote the paper. JH wrote the source codes. VP provided theoretical background for the principal component analysis and for the bootstrap. All authors participated in the interpretation of the results.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgement. This study was supported by the Czech Science Foundation (grants no. 16-05665S [JH] and 16-16549S [MH]). Moreover, the financial support from RVO: 67985874 is greatly acknowledged. We acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank to the climate modelling groups of CLM Community and Rossby Centre (Swedish Meteorological and Hydrological Institute) for producing and making available their model outputs.

References

Bárdossy, A. and Pegram, G.: Multiscale spatial recorrelation of RCM precipitation to produce unbiased climate change scenarios over large areas and small, Water Resour. Res., 48, W09502, doi:10.1029/2011WR011524, 2012.

Chen. J., Brissette. F. P. and Lucas-Picher, P.: Assessing the limits of bias-correcting climate model outputs for climate change impact studies, J. Geophys. Res.-Atmos, 120, 1123–1136, doi:10.1002/2014JD022635, 2015.

Davison, A. C. and Hinkley, D. V.: Bootstrap methods and their application, Cambridge University Press, Cambridge, United Kingdom, 1997.

Déqué, M.: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values, Global Planet. Change, 57, 16–26, doi:10.1016/j.gloplacha.2006.11.030, 2007.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K. and Liebert, J.: Should we apply bias correction to global and regional climate model data?, Hydrol. Earth Syst. Sc., 16, 3391–3404, doi:10.5194/hess-16-3391-2012, 2012.

Giorgi, F., Jones, C. and Ghassem, R.: Addressing climate information needs at the regional level: the CORDEX framework, World Meteorological Organization (WMO) Bulletin, 58, 175–183, 2009.

Hanel, M., Kožín, R., Heřmanovský, M. and Roub, R.: An R package for assessment of statistical downscaling methods for hydrological climate change impact studies, Environ. Modell. Softw., 95, 22–28, 2017.

Hnilica, J., Hanel, M. and Puš, V.: Multisite bias correction of precipitation data from regional climate models, Int. J. Climatol., 37, 2934–2946, doi:10.1002/joc.4890, 2017.

Hoffmann, H. and Rath, T.: Meteorologically consistent bias correction of climate time series for agricultural models, Theor. Appl. Climatol., 110, 129–141, doi:10.1007/s00704-012-0618-x, 2012.

Johnson, F. and Sharma A.: A nesting model for bias correction of variability at multiple time scales in general circulation model precipitation simulations, Water Resour. Res., 48, W01504, doi:10.1029/2011WR010464, 2012.

Mao, G., Vogl, S., Laux, P., Wagner, S. and Kunstmann, H.: Stochastic bias correction of dynamically downscaled precipitation fields for Germany through copula-based integration of gridded observation data, Hydrol. Earth Syst. Sc., 19, 1787–1806, doi:l0.5194/hess-19-1787-2015, 2015.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A. and Mearns, L. O.: Towards process-informed bias correction of climate change simulations, Nat. Clim. Change, 7(11), 764-773, doi: 10.1038/NCLIMATE3418, 2017.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J.,Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M. and Thiele-Eich, I.: Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user, Rev. Geophys., 48, RG3003, doi:10.1029/2009RG000314, 2010.

Mehrotra, R. and Sharma, A.: Correcting for systematic biases in multiple raw GCM variables across a range of timescales, J. Hydrol., 520, 214–223, doi:10.1016/j.jhydrol.2014.11.037, 2015.

Mehrotra, R. and Sharma, A.: A multivariate quantile-matching bias correction approach with auto-and cross-dependence across multiple time scales: Implications for downscaling. J. Climate, 29, 3519–3539, doi:10.1175/JCLI-D-15-0356.1, 2016.

Piani, C. and Haerter, J. O.: Two dimensional bias correction of temperature and precipitation copulas in climate models, Geophys. Res. Lett., 39, L20401, doi:10.1029/2012GL053839, 2012.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M. and Miller, H. L. (Eds): Climate Change 2007: The Physical Science Basis Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Teng, J., Potter, N. J., Chiew, F. H. S., Zhang, L., Wang, B., Vaze, J. and Evans, J. P.: How does bias correction of regional climate model precipitation affect modelled runoff?, Hydrol. Earth Syst. Sc., 19, 711–728, doi:10.5194/hess-19-711-2015, 2015.

Teutschbein, C. and Seibert, J.: Bias correction of regional climatemodel simulations for hydrological climate-change impact studies: Review and evaluation of different methods, J. Hydrol., 456, 12–29, doi:10.1016/j.jhydrol.2012.05.052, 2012.

Wilks, D. S.: Statistical methods in the atmospheric science, 3rd edn. Academic, Amsterdam, 2011.