Adv. Sci. Res., 10, 85–90, 2013
www.adv-sci-res.net/10/85/2013/
doi:10.5194/asr-10-85-2013

Advances in Science & Research
Open Access Proceedings

Multivariate testing of spatio-temporal consistence of daily precipitation records

H. Mächel[1] and A. Kapala[2]

[1]Deutscher Wetterdienst, Klima und Umwelt, Frankfurter Str. 135, 63067 Offenbach, Germany
[2]Meteorological Institute, University Bonn, Auf dem Hügel 20, 53121 Bonn, Germany

Correspondence to: H. Mächel (hermann.maechel@dwd.de)

Received: 24 January 2013 – Revised: 22 April 2013 – Accepted: 3 May 2013 – Published: 10 June 2013

Abstract. The project KLIDADIGI of the German Meteorological Service (DWD) systematically rescues historical daily climate data of Germany by keying and imaging. Up to now, daily nearly gap-free precipitation time series at 118 locations for the period 1901–2000 are collected and extended by digitalization of hand-written protocols. To screen the spatio-temporal consistence of these raw data, we apply principal component analysis (PCA) in S (spatial) mode for daily precipitation records as well as for indices such as the number of rainy days above a certain threshold, intensity and absolute daily maximum in monthly, seasonal or annual resolution. Results of this screening test indicate that the PCA is a useful tool for detection of questionable stations and data preprocessing for further quality control and homogenization.

1 Introduction

For an operational quality control of climate records numerous standard methods are used at weather services and in climate research. The common procedural steps are: selection of outlier values using different thresholds (e.g. ±3 or 4 standard deviations), pairwise comparison of the target station with nearest neighbours using differences, ratios, linear regressions and frequency distributions (e.g. Feng et al., 2004; Vincente-Serrano et al., 2010, and their references). However, quality control of historical, especially of precipitation data is still a challenge if only few series with a sparse (irregular) spatial coverage are available.

Up to now, multivariate methods such as principal component analysis (PCA) are not used in the data preprocessing as a first step in quality control and homogenization.

In climatological studies the PCA is commonly applied to data reduction and to detection of leading spatial patterns/regions of similar temporal variability in different climate parameters. Detailed information to the method as well as to the discussion of the interpretation of PCA results can be found in Jolliffe (2002), Wilks (2006), and Compagnucci and Richmann (2008), for example. For analysis of monthly and daily precipitation variability and changes the PCA is used in several studies (e.g. White et al., 1991; Bonell and Sumner, 1992; and Brunetti et al., 2004, 2006a, b).

For example, Widmann and Schär (1997) classified daily precipitation records from Switzerland for the period 1901–1990 by means of unrotated and rotated PCA. The purpose of their study was, in parallel to the trend analysis, the homogenization of the precipitation series. According to the varimax rotation, three subregions in Switzerland were detected. Recently, Brienen et al. (2012) analysed the variability and change in several precipitation indices for winter (DJF) and summer (JJA) in Germany in the period 1901–2000 applying PCA in S (spatial) mode including varimax rotation to identify regions with similar precipitation behaviour.

This contribution will show the applicability of the PCA for identifying questionable stations in raw daily precipitation time series and derived indices by testing all stations of the network simultaneously.

This paper is subdivided in four parts: in Sect. 2 the database is presented. Section 3 shows the results of the PCA classification and some examples of erroneous data are given in Sect. 4. The results are discussed in Sect. 5.

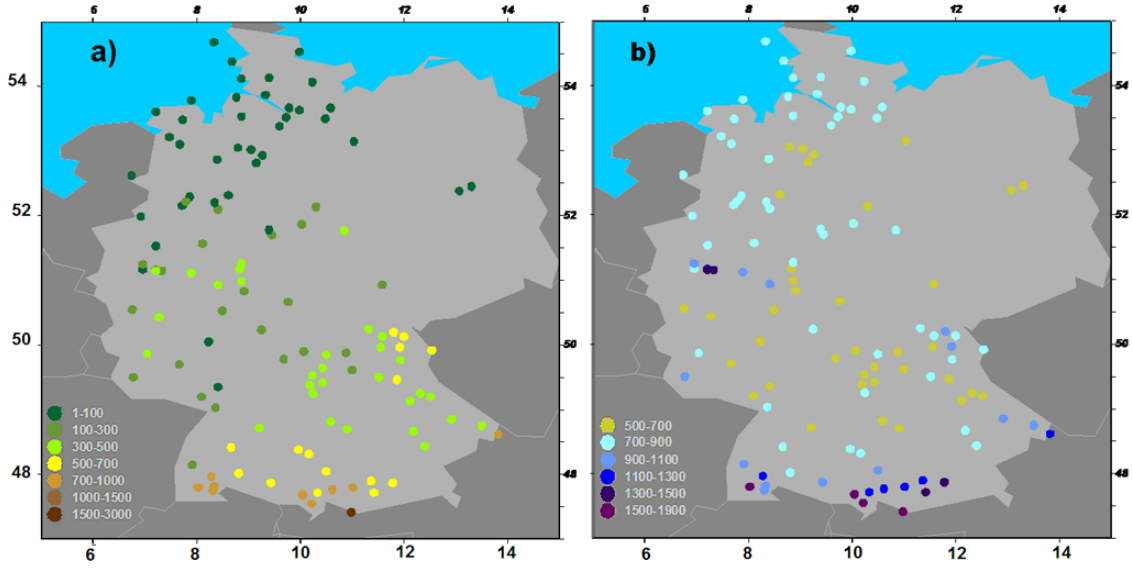

Figure 1. Digitized precipitation series: **(a)** location of the digitized records (the colours of the dots indicate the station elevation in metres), **(b)** mean yearly precipitation sum for individual stations for the period 1901–2000 (in mm yr^{-1}).

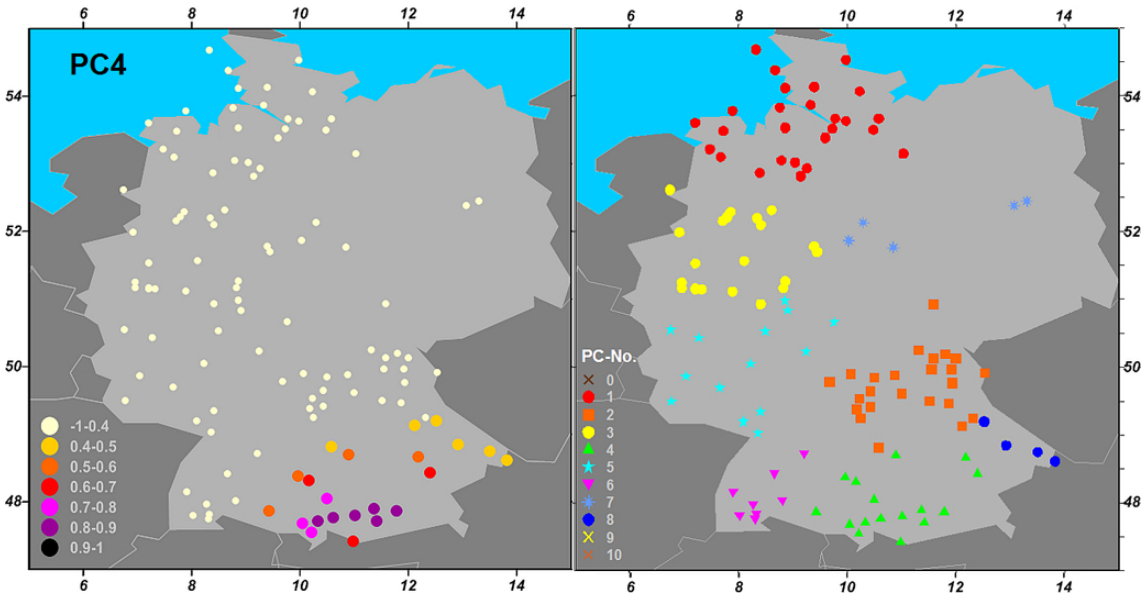

Figure 2. PCA of daily precipitation of all 36 525 days in the period 1901–2000. Examples for spatial patterns of the extracted rotated principal components which correspond to the loadings (here: correlation coefficients) of the original records. Left: PC Nr. 4, right: all 8 rotated PCs (PC9 and PC10 contain only the rest variances with loadings < 0.4).

2 Database

Most of the digitally available daily precipitation series in Germany start in 1931 or later. Therefore the German Meteorological Service (DWD) decided to digitize hand-written protocols to extend the database. 118 daily precipitation time series for the period 1901–2000 with few gaps (lower than 3 yr or 1095 days, mainly in 1945–1946) are selected to test the PCA for detecting outliers and erroneous data. The location, the station elevation above sea level and the mean yearly precipitation sum for these 118 stations are depicted in Fig. 1. In Eastern Germany the digitally available daily precipitation series starts in 1951 or 1969 and only few series match the above mentioned requirements.

The DWD precipitation stations were equipped since the 1890s with Hellmann rain gauges with a 200 cm^2 receiving orifice and brass ring (reading accuracy of 0.1 mm). All rain gauges were protected against evaporation. In the cold season

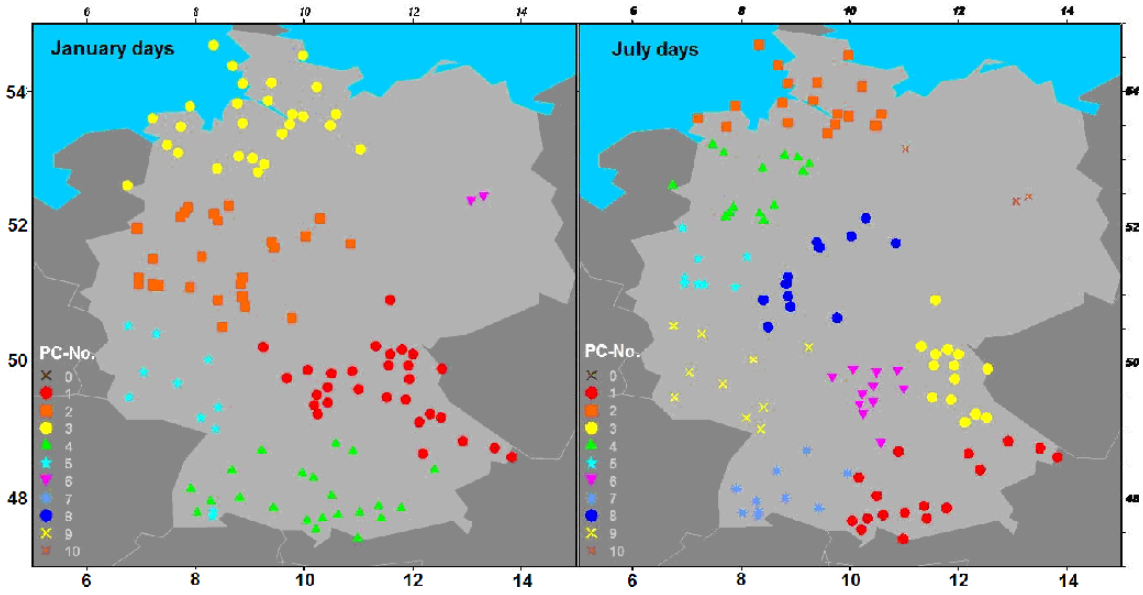

Figure 3. Same as Fig. 2, but for all January days (left) and all July days (right) in the period 1901–2000. In January the two stations within the PC4 pattern seems to be questionable and their original protocols should be checked.

a snow cross had to be placed in the gauge funnel. In the 1990s they were replaced at main stations by tipping-buckets combined with drop-counting (for low rain rates) rain gauges with the same orifice of 200 cm^2. The accuracy of these automatic precipitation measurements were 0.01 mm min^{-1}. The observational time was at the beginning 07:00 LT (local time), from 1 January 1979 the observational time changed to 07:30 CET (Central European Time) (06:30 UTC, Coordinated Universal Time). According to Hellmann (1906), "the changes in observational time and instrumentation are less pronounced than re-locations"; furthermore, "the accuracy of the measurements depends more strongly on the location of the gauges and on the special diligence of the observer than on the used gauge type." On average, the observer changes occur every 10–15 yr often combined with station relocations. A station is continued after relocation with the same identification number if the horizontal distance is less than 5 km, the altitude differs less than 50 m and both locations belong to the same river catchment.

The systematic and documented quality assurance of the data began in 1979 with the introduction of IT (information technology)-supported verification methods (QUALKO, used also in other weather services). With this procedure the daily data was checked automatically for inner, temporal and spatial consistency and flagged. If necessary the erroneous data was manually corrected and flagged accordingly. Since the beginning of the meteorological service in Germany in 1848, the observational protocols are checked manually with neighbouring stations in real time, i.e. in the following month. The newly digitized data is not quality controlled and the following sections should demonstrate a possibility to do this (Mächel et al., 2009).

3 Classification of daily precipitation records by means of PCA

To check the spatio-temporal consistence of the precipitation records from the above mentioned 118 stations and to identify questionable stations in this data set the PCA including varimax rotation of the PCs (principle components) is applied to the correlation matrix of different precipitation characteristics (e.g. daily records of different months, single days and total monthly or yearly amount) for the period 1901–2000.

Since the PCA is widely used this method will not be described here, for details see Richman and Gong (1999), Jolliffe (2002), Wilks (2006), Compagnucci and Richmann (2008) and many others.

As shown by Brunetti et al. (2006b), Widman and Schär (1997), and Brienen et al. (2012), for example, the varimax rotation of the PCs leads to a selection of well-separated spatial patterns (regions) of similar precipitation behaviour based on PC-loadings. In this step, single questionable stations can be easily detected. Questionable stations are: outstanding stations within coherent spatial patterns and stations which could not be assigned to one of the extracted PCs.

In the following, the detected outlier stations are candidates for a manual checking of their precipitation records with original reports. The most frequent inconsistencies in the raw precipitation records result from typing errors, inaccurate specification of the days with no precipitation and not-observed (missing) precipitation and measurements over several days (accumulated precipitation amounts).

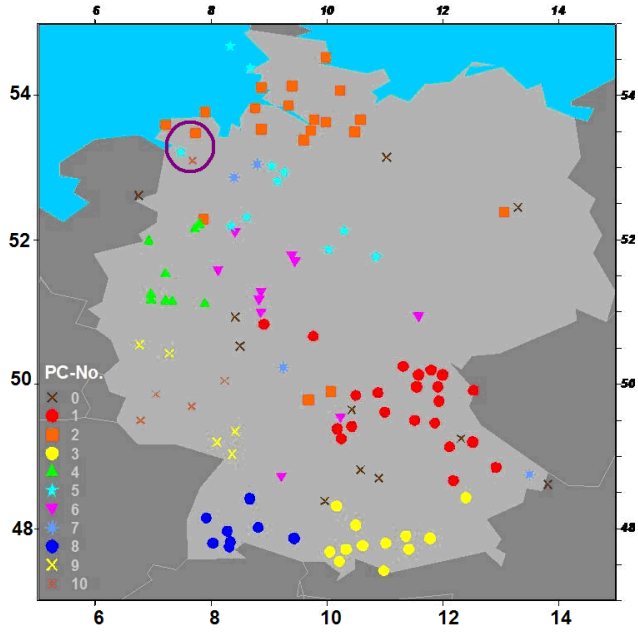

Figure 4. Same as Fig. 2 but, for mean precipitation intensity in July of the period 1901–2000.

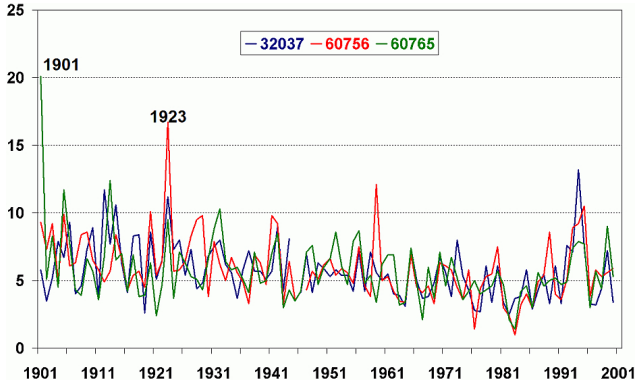

Figure 5. Example of 3 neighbouring time series of mean precipitation intensity (mm day^{-1}) in July which are classified to different PCs (marked by the circle in Fig. 4).

According to Kaiser's criterion (eigenvalue = 1) and screen plots (Wilks, 2006) we select a maximum of 10 PCs for the varimax rotation of the components. To simplify the interpretation of these patterns the loadings are rescaled according to Wilks (2006) by multiplying them with the square root of the eigenvalue. Accordingly, the loadings are expressed in correlation coefficients of the original data at individual stations with the time series of the corresponding principal components (scores). The single station is assigned to one PC according to the maximum correlation. Figure 2 illustrates this for the extracted 8 PCs for daily precipitation (without annual cycle and standardized) of all 36 525 days in the period 1901–2000. The stations with correlation coefficients above 0.5 in the left panel in Fig. 2 correspond to the green triangles (PC4) of the right panel. In this case we obtain well-separated patterns (regions) which correspond with the topography that influences the precipitation. In some cases, however, neighbouring stations are assigned to different PCs. This seems to suggest that the nearest neighbour is not necessarily adequate for paired comparison of the precipitation records.

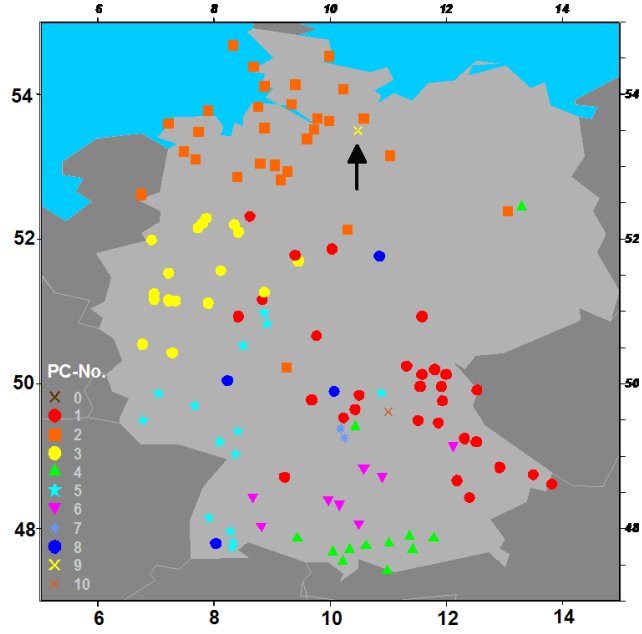

Figure 6. Same as Fig. 2, but for precipitation at individual days; here 9 February for the years 1901–2000.

In the following Figs. 2–6 the loadings (correlation coefficients) < 0.4 are cut off (cf. Richman and Gong, 1999). Coloured symbols indicate the PC's rank according to their explained variance. A station is classified to one of the 10 PCs if the PC time series (PC scores) explains at minimum 16 % of the variance of the original time series (loadings/correlation coefficients $r = 0.4$). If a station does not pass this criterion, it is marked by PC-No. = 0.

If the PCA is applied to daily precipitation of individual months the lower variability of the daily precipitation in winter is manifested through only 6 PCs in January (Fig. 3). In July 10 PCs are needed to classify all stations.

4 Examples of detected errors

If an individual station does not fit in the well-separated regions, it can be assumed that the data contain either errors or real extraordinary events. In such cases we manually compare the data with original reports.

For the detection of a possible existence of over several days accumulated precipitation measurements, the error sensitive index precipitation intensity, defined as total

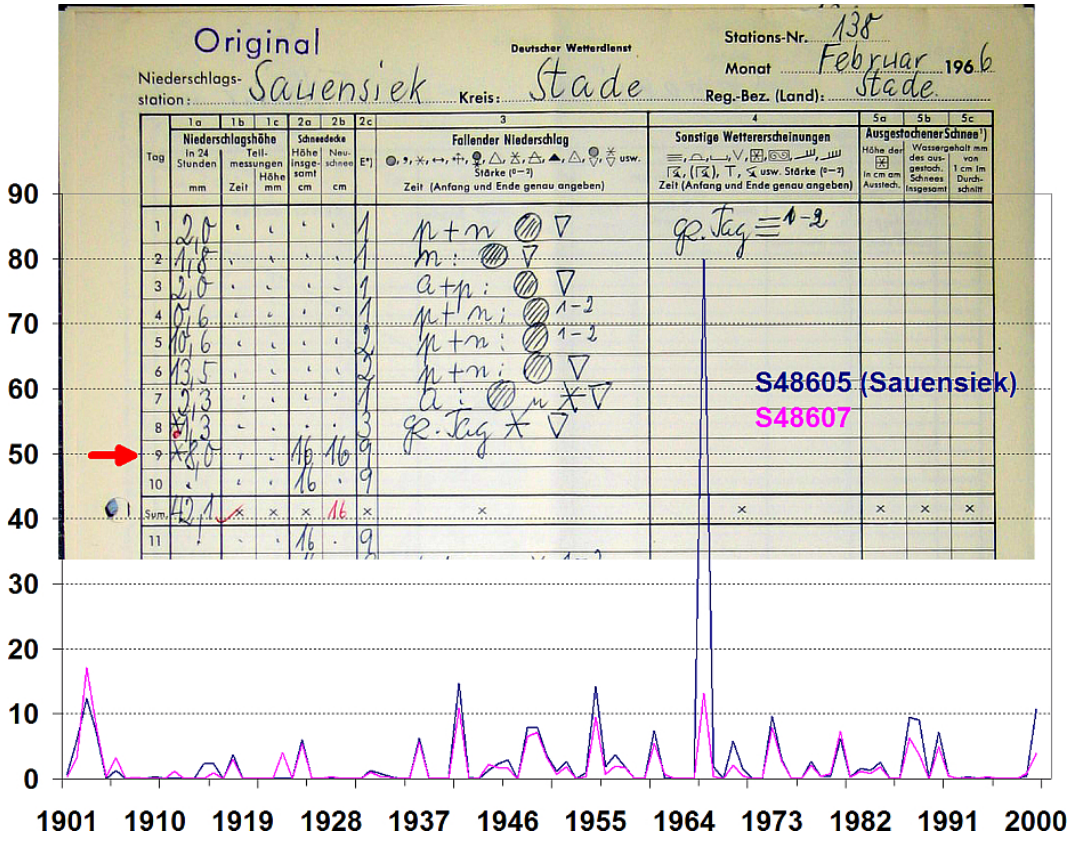

Figure 7. Precipitation time series (mm day^{-1}) of two neighbouring stations for 9 February for the years 1901–2000 as an example of an outlier station. Station 48605 and its neighbours are classified into different PCs (see arrow in Fig. 6). The image shows the original report of station 48605 for February 1966.

precipitation divided by the number wet days (precipitation amount > 0.0) for single months, seasons or a year, is used.

As can be seen in Fig. 4, the delineation of regions of precipitation intensity in July 1901–2000 is not as clear as in the case of daily precipitation amounts in July (cf. Fig. 3, right). Some of the stations differ from their neighbours and several of the stations have not passed the criterion of 0.4 correlation coefficient. This suggests that the number of wet days and/or the precipitation amounts at some stations may be incorrect. For example, 3 stations are striking in the northwestern corner of Germany marked by the purple circle. Their time series of precipitation intensity show (Fig. 5) some peaks that are verified with the original reports; on 22 July 1901 station 60765 reported a daily value of 78.5 mm, the other two stations only 16 mm (station 32037) and 24 mm (station 60756). Whether the value of 78.5 mm is real or not, cannot be clearly decided without additional information. In fact, some neighbouring stations with shorter time series (not digitally available at this time) also show high values. Therefore the detected peak value of 78.5 mm seems to be real. The high value in 1923 at station 60756 is the result from 2, over 3 days of accumulated values and therefore the number of wet days is lower than at the neighbouring stations. This erroneous amount should be disaggregated in proportion to the neighbouring stations.

The PC-analysis is also applied to a single day in a year for the period 1901–2000. An example for 9 February is shown in Fig. 6. The PCA of individual days enables easy identification of outlier stations and checking of their records if original reports are available. One station in northern Germany, marked by the arrow, is striking. Figure 7 shows the solution. In the time series of daily precipitation on 9 February 1966 an outlier value of 80.0 mm is found which is extraordinary compared to other neighbours. According to the original report only 8.0 mm were measured on this day.

5 Conclusions

The quality control of historical daily data is very time consuming. Due to the high daily precipitation variability only large errors can be found. To reduce time and efforts, the screening of the spatio-temporal consistence of the considered data set by means of the PCA can be used as a first step of quality control.

One advantage of the PCA is that this method provides regions with similar behaviour of the data or select stations

with similar variability and, therefore, errors could easily be detected. As show in the few examples, the probability of finding questionable stations in the considered data is higher by checking records of every individual day and different indices calculated from the daily records than the precipitation amount itself (cf. Figs. 4 and 5).

An error sensitive index is the precipitation intensity, the ratio between the monthly precipitation sum and the number of precipitation days above a certain threshold (zero or 1.0 mm). By means of this index it is easy to find months with accumulated values, if an observer did not measure continuously every day (perhaps only on work days or not during holidays). Such values can be disaggregated in proportion to their neighbouring stations.

In general, the PCA is an useful tool for a quick check of spatio-temporal coherence of climate records including identification of conspicuous stations. It can be used as a first step in quality control and homogenization of the data. After the first correction of the erroneous data it can be repeated to assess the improvement in the spatio-temporal consistence. Furthermore, the highest loadings of each PC enables one to identify the leading or most representative station in the selected region that can be used as a reference station for a homogenization procedure. However, the disadvantage of PCA is that the data has to be without gaps, in other words, the missing values in one of the series causes that this date be omitted from the analysis. To handle this problem, in a first step monthly data can be analysed for the whole country to find regions. In the second step only stations within such a region can be used for testing the spatio-temporal consistence of daily data or different indices.

However, neither PCA nor other quality control procedure (e.g. regression with a neighbouring station) can guarantee that simultaneously occurred errors/inconsistencies in several time series can be detected.

Acknowledgements. The authors would like to thank Jennifer Lenhardt for her assistance during the conference and the reviewers.

Edited by: M. Brunet-India
Reviewed by: J. Sigro and one anonymous referee

References

Bonell, M. and Sumner, G.: Autumn and winter daily precipitation areas in Wales, 1982–1983 to 1986–1987, Int. J. Climatol., 12, 77–102, doi:10.1002/joc.3370120108, 1992.

Brienen, S., Kapala, A., Mächel, H., and Simmer, C.: Regional centennial precipitation variability over Germany from extended observation records, Int. J. Climatol., doi:10.1002/joc.3581, in press, 2012.

Brunetti, M., Maugeri, M., Monti, F., and Nanni, T.: Changes in daily precipitation frequency and distribution in Italy over the last 120 years, J. Geophys. Res., 109, D05102, doi:10.1029/2003JD004296, 2004.

Brunetti, M., Maugeri, M., Monti, F., and Nanni, T.: Temperature and precipitation variability in Italy in the last two centuries from homogenised instrumental time series, Int. J. Climatol., 26, 345–381, doi:10.1002/joc.1251, 2006a.

Brunetti, M., Maugeri, M., Nanni, T., Auer, I., Böhm, R., and Schöner, W.: Precipitation variability and changes in the Greater Alpine Region over the 1800–2003 period, J. Geophys. Res., 111, D11107, doi:10.1029/2005JD006674, 2006b.

Compagnucci, R. H. and Richman, M. B.: Can principal component analysis provide atmospheric circulation or teleconnection patterns?, Int. J. Climatol., 28, 703–726, doi:10.1002/joc.1574, 2008.

Feng, S., Hu. Q., and Qian, W.: Quality control of daily meteorological data in China, 1951–2000: a new dataset, Int. J. Climatol., 24, 853–870, 2004.

Hellmann, G.: Die Niederschläge in den Norddeutschen Stromgebieten, Band I, Dietrich Reimer Verlag, Berlin, 425 pp., 1906.

Jolliffe, I. T.: Principal Component Analysis, Springer, New York, doi:10.1007/b98835, 2002.

Mächel, H., Kapala, A., Behrendt, J., and Simmer, C.: Rettung historischer Klimadaten in Deutschland: das KLIDADIGI-Projekt des DWD, Klimastatusbericht 2008 (Climate Status Report), Deutscher Wetterdienst, Offenbach, 103–118, http://www.ksb.dwd.de/ (last access: 12 January 2013), 2009.

QUALKO: www.dwd.de/nkdz – Quality assurance, last access: 20 April 2013.

Richman, M. B. and Gong, X.: Relationships between the definition of the hyperplane width to the fidelity of principal component loading patterns, J. Climate, 12, 1557–1576, doi:10.1175/1520-0442(1999)012<1557:RBTDOT>2.0.CO;2, 1999.

Vicente-Serrano, S. M., Begueria, S., Lopez-Moreno, J. I., Garcia-Vera, M. A., and Stepanek, P.: A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity, Int. J. Climatol., 30, 1146–1163, doi:10.1002/joc.1850, 2010.

White, D., Richman, M., and Yarnal, B.: Climate regionalization and rotation of principal components, Int. J. Climatol., 11, 1–25, doi:10.1002/joc.3370110102, 1991.

Widmann, M. and Schär, C.: A principal component and long-term trend analysis of daily precipitation in Switzerland, Int. J. Climatol., 17, 1333–1356, doi:10.1002/(SICI)1097-0088(199710)17:12<1333::AID-JOC108>3.0.CO;2-Q, 1997.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Elsevier, San Diego, 2006.