Adv. Sci. Res., 8, 99–104, 2012
www.adv-sci-res.net/8/99/2012/
doi:10.5194/asr-8-99-2012

Advances in Science & Research
Open Access Proceedings

Advanced information criterion for environmental data quality assurance

A. Düsterhus and A. Hense

Meteorological Institute of the University of Bonn, Auf dem Hügel 20, 53121 Bonn, Germany

Correspondence to: A. Düsterhus (andue@uni-bonn.de)

Received: 19 December 2011 – Revised: 11 April 2012 – Accepted: 16 April 2012 – Published: 7 May 2012

Abstract. A new method for testing time series of environmental data for internal inconsistencies is presented. The method divides the dataset into several disjunct blocks. By means of a comparison of the blocks' estimated probability density distributions, each block is compared with the others. In order to judge the differences, four different measures are used and compared: Kullback-Leibler Divergence, Jensen-Shannon Divergence, Earth Mover's Distance and the Root Mean Square. By looking at the resulting patterns, conclusions on possible inconsistencies in the data can be drawn.

This paper shows some sensitivitiy tests and gives an example for an application to real data. Furthermore, it is shown, in which cases of errors (shift in mean, shift in variance and rounding), which measure performs best.

1 Introduction

When using data measured from natural systems to draw conclusions about the observed system, data quality assurance is a very important factor. In this quality assurance process not only the metadata, but also the data itself should be controlled. For this kind of control, there are several generic methods available, which means that they may be applied to any data set without detailed knowledge of its specifics.

These generic methods are mostly based on rules described by Meek and Hatfield (1994). These rules control datasets separately for each data point, on whether specified limits are exceeded (LIM), the number of successive elements, which are not changing, exceeds a predefined number (NOC) or whether the rate of change between two successive data points exceeds a limit (ROC). Those rules were applied to several datasets with different methods for defining the parameters of the tests (e.g. Hubbard et al., 2005; Zahumensky, 2007; Jiménez et al., 2010; Durre et al., 2010). For data sets that are available on a regular basis, like meteorological networks, methods like "Complex Quality Control" (CQC), developed by Gandin (1988), homogenization (Peterson et al., 1998) or Mathes et al. (2008) might be more useful.

When the sources of data are unknown, a more general procedure is required. In this paper a newly developed method for this problem is presented. It is based on the analysis of the estimated probability density of data, for which two basic forms are possible. One takes a look at statistical moments (like mean, standard deviation or percentiles) and their development within the whole dataset. The other approach investigates all the distribution information. This is what will be pursued in this paper.

To avoid any preconditioning of the results, a non-parametric density estimation shall be the starting point. With the help of these estimates an evaluation of the two probability densities is performed. The assumption used is that two probability densities, characterizing the data in two time windows, are identical. A strict hypothesis test in the statistical sense is not performed.

In the next step a distance measure between two densities is defined. Standard methods like the Kolmogorov-Smirnov test are very sensitive to sample variability (Owen, 1995). Therefore, we would like to use more robust measures. These have to take into account the full structure of the estimated densities or their integrals, the probability distributions. This is in contrast to the Kolmogorov-Smirnov test which only

take into account the difference at one point, namely the maximum deviation. There are several divergences available for comparing distributions. Those used in this paper are described in Sect. 2: Kullback-Leibler (Kullback and Leibler, 1951), Jensen-Shannon (Endres and Schindelin, 2003) and the Earth Mover's distance (Rubner et al., 2000). In Sect. 3 some sensitivity studies are performed before an application is shown in Sect. 4. The paper ends with a discussion in Sect. 5 and is summarized in Sect. 6.

2 Method

The basic methods rely on a division of the dataset into blocks of blocksize s_b and a comparison of every block to the others. This is carried out by comparing the blocks' normalized histograms as estimators of the underlying probability density, which uses a number of bins n_b. These bins are uniformly distributed between the maximum and minimum of both blocks.

To determine the difference between both histograms ($f, g \in \mathbb{R}^{n_b}$) the following distance measures are used:

Kullback-Leibler Divergence (KLD). KLD is based on the work described in Kullback and Leibler (1951). It is an unsymmetric function between two histograms and defined by Lin (1991) as follows:

$$D_{KL}(f\|g) = \sum_{i=1}^{n_b} f(x_i) \cdot \log_2 \frac{f(x_i)}{g(x_i)}. \quad (1)$$

It is obvious that a problem occurs when $g(x) = 0$ for any $x \in [1, n_b]$. To prevent this, a prior estimation a_p is introduced for every bin of both estimated probability densities:

$$h_i = \frac{a_i + a_p}{s_b + n_b \cdot a_p} \quad (2)$$

where h_i is the resulting bin of the histogram, a_i is the number of observations in bin i and s_b is the total number of observations in the block. To couple a_p to the number of observations s_b, a_p depends on a small factor a_f and s_b and is defined by the following equation:

$$a_p = \frac{1}{a_f \cdot s_b}. \quad (3)$$

Jensen-Shannon Divergence (JSD). JSD is a symmetrization of the KLD and can be defined as follows (Endres and Schindelin, 2003):

$$D_{JS}(f\|g) = \frac{1}{2} D_{KL}\left(f \left\| \frac{1}{2}(f+g)\right.\right) + \frac{1}{2} D_{KL}\left(g \left\| \frac{1}{2}(f+g)\right.\right). \quad (4)$$

JSD and KLD are positive definite functionals, but neither the first nor the second are "real" distance measures because they do not obey the triangle inequality.

Earth Mover's Distance (EMD). EMD was developed as a solution of a transportation problem (Rubner et al., 2000). In contrast to KLD and JSD it does not rely on a bin-wise ratio. Rather, it figures out how to transform one histogram to the other. To do this the probability of every bin is seen as a mass, which has to be transported. The EMD measures the minimal work that has to be invested for this task. Important here is that the distance between two bins is not neglected, but defined as $d(i, j) = \frac{|i-j|}{n_b}$. For a one-dimensional histogram this leads to (Rabin et al., 2008):

$$D_{EM}(f\|g) = \frac{1}{n_b} \sum_{i=1}^{n_b} |F(x_i) - G(x_i)| \quad (5)$$

where F and G are the cumulative distribution functions of f and g. EMD is a true distance measure being positive definite, symmetric and obeying the triangle inequality. EMD is a special case of the more general Wasserstein distance of probability density functions (Levina and Bickel, 2001).

Root Mean Square (RMS). RMS is only used as a reference in this paper. The well known definition is given by:

$$D_{RMS}(f\|g) = \frac{1}{n_b} \left(\sum_{i=1}^{n_b} (f(x_i) - g(x_i))^2 \right)^{\frac{1}{2}}. \quad (6)$$

When such a method is used to evaluate a dataset, a typical resulting plot consists of a two dimensional array. Each entry is the result of a comparison of two parts of the dataset. On the diagonal, each part of the dataset is compared to itself and the value should be zero. This condition is fulfilled by all of the four distance measures. The rest of the array is filled with the distances between the histograms of every part to the others. Also, all but the KLD deliver symmetric arrays.

In the next section sensitivity tests are performed in order to simulate the influence of the different distance measures on this method. Because this method delivers only relative results, it is necessary to define a measure that makes the different measures comparable. Therefore the dataset will be separated into two parts. Each part gets a different characteristic. When the blocks are compared to each other it is now known, which comparison looks at blocks with the same characteristics and which at blocks with different characteristics. To determine the difference between blocks of differences of same and different characteristics the definition of x_{sd} is introduced as follows:

$$x_{sd} = \frac{\mu_{diff} - \mu_{same}}{\sigma_{diff} + \sigma_{same}}. \quad (7)$$

μ_{same} and σ_{same} are the mean and standard deviation of those distances, which compare sections with the same characteristics of the dataset. The same is valid for the sections with different characteristics (diff). The distances, which are zero are neglected in the calculation of x_{sd}.

It is plausible that higher values for x_{sd} means that the differences in the data set are easier to detect than lower ones.

3 Sensitivity tests

In this section some characteristics of the methodology using simulated observations are discussed and the different distance measures are compared. For the simulation a sample of

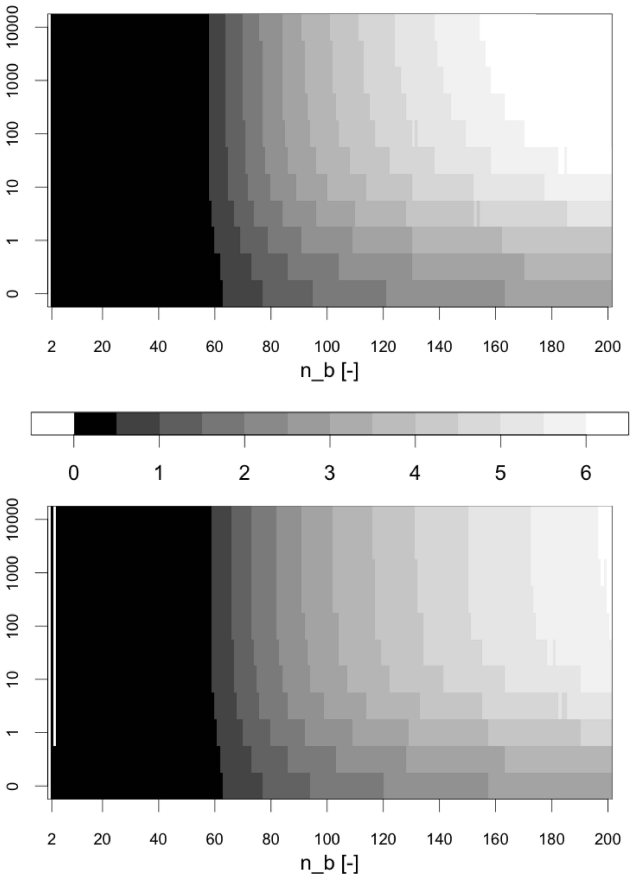

Figure 1. Influence of a_f on the results measured in x_{sd}, displayed as shadings of gray, of the KLD (upper figure) and JSD (lower figure) for different number of bins n_b.

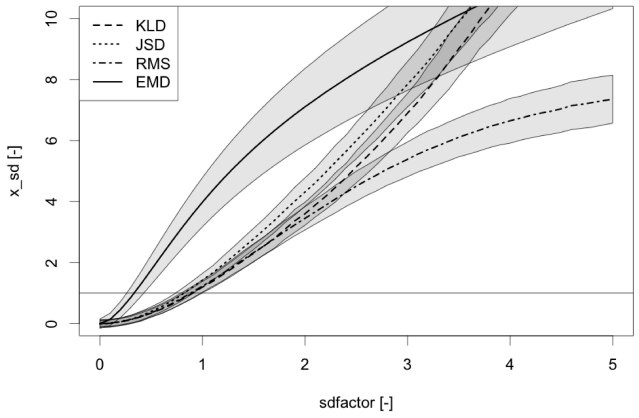

Figure 2. Results of the regime shift sensitivity test, with artificially included shifts in the mean. The shift is measured in terms of the standard deviation and shown on the x-axis. The curves shows the average and their respective uncertainties of the measure x_{sd} in Eq. (7) for 100 randomly generated data sets (without shift normally distributed with expectation mean = 0, sd = 1) for the four different measures.

2000 realizations of a Gaussian distributed and normalized (mean zero, variance 1) random variable is used. The sample is split into two equally large subsamples where the second sample is subjected to a change. Afterwards, the method is applied with a blocksize $s_b = 100$ and x_{sd} is calculated. In this calculation the comparisons with "different characteristics" are represented by the influence of the first (block 1 to 10) on the second half (block 11 to 20). For the comparison with the "same characteristics" the influence of the second half on itself is used. The treatment of the second half in the next section is a rounding on the first digit.

3.1 Influence of a_f

In the definition of KLD and JSD the value a_f is used to incorporate the amplitude of the prior for each bin. In Fig. 1 the results for x_{sd} are calculated for 100 different randomly drawn vectors and the mean is shown for 200 different n_b and eleven different a_f, which are distributed on an logarithmic scale.

Principally, better values are achieved for a higher number of bins. It is also better to use higher a_f values, what is equivalent to a lower prior a_p for each bin. For values higher than $a_f = 100$ no further significant difference is detectable in comparison to higher values. That is why this value is chosen for the next tests with the KLD and the JSD. As a next step, x_{sd} have to be chosen, whereby two sections with different characteristics are clearly distinguishable. This can be defined, when x_{sd} exceeding 1. In Sects. 3.2 and 3.3 $x_{sd} = 1$ is also used as a detection limit of inconsistencies within a dataset. The condition of x_{sd} exceeds 1 is also used here to determine the number of bins n_b of the histograms. It is fulfilled at approximately $n_b = 65$, which is used throughout the remainder of the paper.

3.2 Shift in mean

As a second sensitivity test a detection of a regime shift is used. Unlike before the second half of the tested dataset is not rounded, but a factor of y_{sd} standard deviations is added. This y_{sd} is now selected in the range of 0 to 5 and the evaluation is carried out like before. The mean results for 100 vectors and their standard deviation with $n_b = 65$ are shown in Fig. 2. Here, the results measured in x_{sd} are plotted against the added value measured in standard deviations y_{sd}. The detection limit chosen before is indicated by a line at $x_{sd} = 1$. Since KLD is asymmetric it delivers different results, if the ingoing histograms are transposed. Therefore, only the better result of the KLD is shown.

The best detection result is achieved by the EMD. This distance measure is highly sensitive for low values of y_{sd} and reaches the detection limit at about $y_{sd} \approx 0.4$. The three other distance measures are less sensitive and reach their detection limit at about twice the value of the EMD $y_{sd} \approx 0.9$. For higher values of y_{sd}, the JSD measure detects shifts slightly

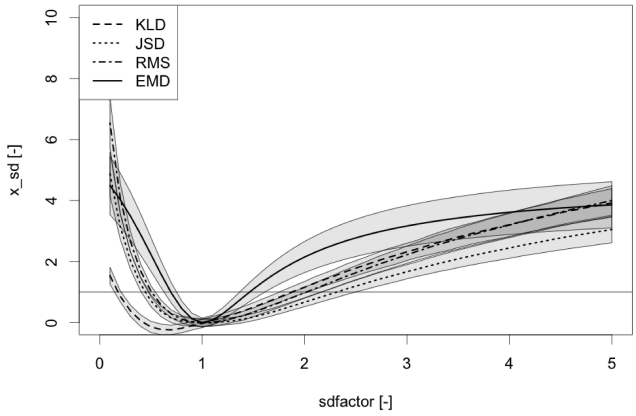

Figure 3. Results of the regime shift sensitivity test, with artificially included shifts in the variance. The shift is measured in terms of the standard deviation and shown on the x-axis. The curves shows the average and their respective uncertainties of the measure x_{sd} in Eq. (7) for 100 randomly generated data sets (without shift normally distributed with expectation mean = 0, sd = 1) for the four different measures.

better than the KLD. RMS proves to be worst in detecting the shifts.

Increasing the number of bins n_b deteriorates these results except for EMD (not shown).

3.3 Shift in variance

Like before the second half of the dataset is manipulated, but now the y_{sd} is not added but multiplied increasing the variance. The results are shown in Fig. 3 and are constructed as specified in Sect. 3.2. Once again EMD delivers the best results, by reaching the detection limit with the smallest deviation of $y_{sd} = 1.5$. The next is the KLD, where the better of the both possibilities to calculate this measure reach the detection limit at around $y_{sd} = 2.0$. The RMS follows with $y_{sd} = 2.1$ and the worst results are delivered by the JSD with reaching the detection limit at around $y_{sd} = 2.5$. At this point it is necessary to mention briefly that the asymmetry property of the KLD plays a huge role in this test. While the differences of choosing $D_{KL}(f||g)$ or $D_{KL}(g||f)$ can be neglected when a change in the mean occur, in the case of a variance shift, the detection limit of the inferior was not reached under $y_{sd} = 5.0$.

4 Application

As an example a time series measured by the German National Meteorological Service DWD is used (available from http://www.dwd.de/klimadaten). This time series shows the daily maximum wind data from Lindenberg/Germany (station id: 10393/52°21′ N/14°12′ E, elev. 98 m), consisting of twenty years (1991–2010) of data.

As previously, the tests are performed with all four different divergence measures. The parameters are set to $n_b = 65$ and $s_b = 365$. The latter choice serves to eliminate seasonal effects. This prevents a bias of taking into account a season more often than an other into one block. The results are shown in Fig. 4.

Especially KLD (Fig. 4a) and JSD (Fig. 4b) show a pattern of higher values in the years 1991, 1999 and 2000. RMS (Fig. 4c) also delivers such an indication, but in the result produced with the EMD (Fig. 4d) no evidence of these special time periods can be found.

A reason for these higher values are demonstrated for the period 1999 and 2000. In Fig. 4e displaying the time series for the period July 1998 till July 2000. Obviously, at 1 December 1998 there was a change in the recording procedure initiated with the data stored only to the nearest integer. This period ends at the beginning of April 2001, when another change in the recording procedure has occurred. The same rounding of the data can be found in the dataset up to 1 June 1992, which explains the high values in 1991. This shows that EMD is apparently insensitive to this type of change in data in contrast to the remaining measures. The reason will be discussed in the next section.

5 Discussion

The method for testing data quality presented in this paper offers a simple way to detect potential errors and discrepancies to data users. We propose to use a set of measures derived from estimated probability densities (histograms). These have been tested on artificial data with the tests showing a clear advantage in most situations of the EMD, which is a distance measure for probability densities.

It is shown that different measures of these changes react differently to distinct types of these changes. For example, the EMD is much more sensitive to potential regime shifts or changes in the variance of the data than KLD, JSD and RMS. This is rooted in the definition of EMD as a solution of the minimal work for the transportation problem. The focus is set on the distance, when the probability of one bin is "transported" to another. KLD, JSD and RMS are simply comparing the difference between the bins, without looking at the range. The same argumentation holds for the better results of KLD and JSD in rounding problems. Because the range is so small between the bins with different probabilities, the difference in value matters more than the distance between the bins.

The regime and variance shifts are a common phenomenon in observational data sets. Therefore, a number of tests are available for these kinds of potential errors (Ducré-Robitaille et al., 2003). In contrast, rounding problems are mostly neglected, although they deliver a good indication for changes in measurement techniques. The presented method with the

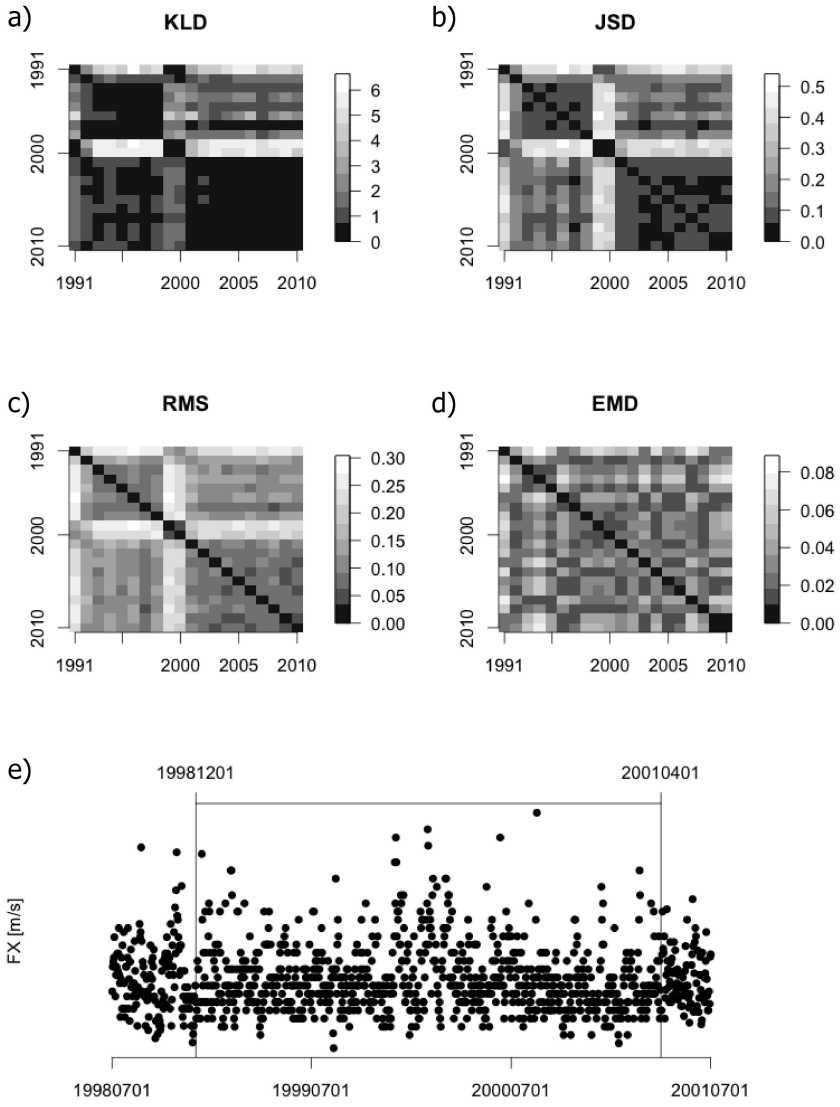

Figure 4. Analysis of the maximum wind at Lindenberg station between 1991 and 2010 with the four different measures (panel **a–d**). Also shown in panel **(e)** is the relevant section in the data between July 1998 and July 2001, where KLD, JSD and RMS show higher values.

KLD or JSD as a measure delivers a good test for such changes.

Tests on internal consistency are an important part of a data quality assurance workflow. If it is known what type of data is under review, simple rules can be applied to highlight the problematic parts of a dataset. Examples are the ROC and NOC rules by Meek and Hatfield (1994). Others can be found in the framework of a complex quality assurance (Gandin, 1988; Graybeal et al., 2004) or homogenization (Peterson et al., 1998).

If there is no prior information on the data that is actually checked, the task will become more complicate. Of course, normalized limits can be checked (Hubbard et al., 2005).

All these tests only validate one value to check against one or more recently measured values of the same measurement or measurement type. The approach presented here is different, because it evaluates complete datasets.

An additional advantage is the flexibility of choosing the blocks within a dataset. This enables the possibility to perform these checks on two or more dimensional data like model outputs.

6 Conclusions

In this paper a new method for data quality assurance is presented. It divides the dataset to be tested into disjunct blocks, before each block is compared to the others. This works by a comparison of the blocks' estimated probability density. In order to determine the differences, four different distance measures are applied. While the Earth Mover's Distance

delivers good results for detection of regime and variance shifts in data, the Kullback-Leibler and Jensen-Shannon Divergences are best at rounding problems.

Acknowledgements. This work is funded by the Deutsche Forschungsgemeinschaft DFG (Literatur- und Informationssysteme) under the number He1916/18-1. Additionally, we like to thank the contributors of the R project and its packages (R Development Core Team, 2011; Furrer et al., 2011). We also acknowledge the data support by the DWD. Furthermore, we would like to thank two anonymous reviewers for their constructive comments.

Edited by: M. Brunet-India
Reviewed by: two anonymous referees

The publication of this article is sponsored by the European Meteorological Society.

References

Ducré-Robitaille, J.-F., Vincent, L. A., and Boulet, G.: Comparison of techniques for detection of discontinuities in temperature series, Int. J. Climatol., 23, 1087–1101, 2003.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S.: Comprehensive Automated Quality Assurance of Daily Surface Observations, J. Appl. Meteorol. Clim., 49, 1615–1633, 2010.

Endres, D. M. and Schindelin, J. E.: A new metric for probability distributions, IEEE T. Inform. Theory, 49, 1858–1860, 2003.

Furrer, R., Nychka, D., and Sain, S.: Fields: Tools for spatial data, http://CRAN.R-project.org/package=fields, R package version 6.6.2, 2011.

Gandin, L. S.: Complex Quality Control of Meteorological Observations, Mon. Weather Rev., 116, 1137–1156, 1988.

Graybeal, D. Y., DeGaetano, A. T., and Eggleston, K. L.: Complex Quality Assurance of Historical Hourly Surface Airways Meteorological Data, J. Atmos. Ocean. Tech., 21, 1156–1169, 2004.

Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., and Osugi, T. T.: Performance of Quality Assurance Procedures for an Applied Climate Information System, J. Atmos. Ocean. Tech., 22, 105–112, 2005.

Jiménez, P. A., González-Rouco, J. F., Navarro, J., Montávez, J. P., and Garcia-Bustamante, E.: Quality Assurance of Surface Wind Observations from Automated Weather Stations, J. Atmos. Ocean. Tech., 27, 1101–1122, 2010.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, The Annals of Mathematical Statistics, 22, 79–86, 1951.

Levina, E. and Bickel, P.: The Earth Mover's distance is the Mallows distance: some insights from statistics, Eighth IEEE International Conference on Computer Vision, 2001, ICCV 2001, Proceedings, 251–256, 2001.

Lin, J.: Divergence measures based on the Shannon entropy, IEEE T. Inform. Theory, 37, 145–151, 1991.

Mathes, A., Friederichs, P., and Hense, A.: Towards a quality control of precipitation data, Meteorol. Z., 17, 733–749, 2008.

Meek, D. W. and Hatfield, J. L.: Data Quality Checking for Single Station Meteorological Databases, Agr. Forest Meteorol., 69, 85–109, 1994.

Owen, A. B.: Nonparametric Likelihood Confidence Bands for a Distribution Function, J. Am. Stat. Assoc., 90, 516–521, 1995.

Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L. A., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and Parker, D.: Homogeneity adjustments of in situ atmospheric climate data: a review, Int. J. Climatol., 18, 1493–1517, 1998.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, R Foundation for Statistical Computing, http://www.R-project.org/, ISBN 3-900051-07-0, 2011.

Rabin, J., Delon, J., and Gousseau, Y.: Circular Earth Mover's Distance for the comparison of local features, 19th International Conference on Pattern Recognition, 2008.

Rubner, Y., Tomasi, C., and Guibas, L. J.: The Earth Mover's Distance as a Metric for Image Retrieval, Int. J. Comput. Vision, 40, 99–121, 2000.

Zahumensky, I.: Guidelines on Quality Control Procedures for Data from Automatic Weather Stations, World Meteorological Organization, WMO-No. 488, Appendix VI.2, 2007.