IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 113-

Theoretical questions of daily data homogenization

Tamás Szentimrey

Hungarian Meteorological Service,
P.O. Box 38, H-1525 Budapest, Hungary
E-mail: szentimrey.t@met.hu

(Manuscript received in final form December 10, 2012)

Abstract—The so-called variable correction methods form a special type of methods developed for daily data homogenization. Their common assumption is that in case of daily data series, the corrections for inhomogeneity have to vary according to the meteorological situation of each day in order to represent the extremes. In this paper we express our objections to these variable correction methods, especially to their underlying principles. Since the exact theoretical mathematical formulation of the question of daily data homogenization is generally neglected, we also try to formulate and analyze this problem in accordance with mathematical conventions.

Key-words: daily data series, homogenization, climate extremes, higher-order moments, distribution, mathematical formulation

1. Introduction

During the last years, the interest to the daily data series homogenization has increased dramatically. The main reason of this tendency is that daily data are essential for studying extremes of weather and climate, for example, computing extreme climate indices requires reliable daily data series. However, according to numerous climatologists homogenization of daily data is still in its infancy and is much more difficult problem than homogenization at monthly or annual scales. The essence of this argumentation is that the correction in mean is sufficient for monthly and annual series, but in case of daily data series, the corrections should vary according to the meteorological situation of each day in order to represent the extremes. This idea was published in the paper by *Trewin* and *Trevitt* (1996), where parallel measurements were examined and compared to each other. Since then on the basis of the ideas formulated in the paper, a number of variable correction methods have been developed with the declared aim of being capable of correcting the daily data not only in mean (first moment) but also in the higher order moments. For example, we mention the following methods: higher order moments (HOM) method by *Della-Marta* and *Wanner* (2006) and spline daily homogenization (SPLIDHOM) method by *Mestre et al.* (2011), and there are numerous other similar methods applied in practice. But unfortunately, in this paper we have to make a criticism about these variable correction methods, especially about their underlying principles. In our humble opinion, during the examinations only some physical experiences were considered without any exact theoretical, mathematical formulation of the problem. The empiric interpretation and formulation seem to be a misunderstanding. Moreover, there are some mathematical statements at the description of the methods – e. g., capability to correct the higher order moments – but without any proof, and this practice is of course contrary to the mathematical conventions.

2. Examination of parallel measurements

2.1. Examinations by Trewin and *Trevitt (1996)*

First here is a quotation from the paper of *Della-Marta* and *Wanner* (2006): “One of the most robust methods capable of adjusting the higher-order moments of daily temperature data is that of *Trewin* and *Trevitt* (1996).” *Trewin* and *Trevitt* (1996) intended to homogenize daily data series in order to create composite temperature records. The following sentences are from their paper: “It is therefore necessary to make use of climatological records with inhomogeneities, and to develop a means of removing or minimizing the impact of inhomogeneities on these records. One way of doing this is by adjusting all parts of a record to be comparable with some ‘reference period’. Standard procedures for such adjustments in mean temperatures have relied on the implicit assumption that, if two neighbouring stations both have homogeneous records over some period of time, the difference in daily maximum (or minimum) temperature between them will be a constant for any day in a given month of the year. This implies that the difference in monthly means will be a constant for that month from year to year.” In general it is not true of course, but after some examination of real station data series they obtained the following result: “This is observed at Armidale (P. Burr, pers. comm.), ..,where the difference in minimum temperature between the town centre site used in this study and a second site approximately 2 km to the east, in the outer part of the town, has a mean value of 1.5 to 2 °C , but can increase to 4 °C on cold, clear nights. The assumption that the temperature difference between any two nearby sites is always constant must therefore be questioned."

The above conclusion was all right, but the next conclusion is a little bit surprising for us: "The relationship between the temperature characteristics of the two sites in each pair was examined, with the aim of determining an appropriate method for use in extrapolating records at one site to records at the other."

Probably here is the origin of the methods that apply varying corrections per days, and at this step a regression or interpolation problem was obtained for homogenization instead of the adequate distribution problem. Three interpolation techniques were considered by *Trewin* and *Trevitt* (1996) namely: the 'traditional' constant-difference approach, the 'regression' method, and the frequency distribution matching. The methods will be detailed in Section 4.1.

2.2. Mathematical examinations of parallel measurements

What was the reason of the development of the variable correction methods? Essentially, an observed phenomenon at the extremes, namely the differences of parallel measurements are larger in case of extremes. In our opinion, this observed phenomenon has a simple and logical reason, and it is superfluous to look for some complicated physical explanation for the inhomogeneity. The simple reason is that the extremes may be expected at different moments in case of parallel measurements, or in other words, there may be systematic biases in rank order! It is a natural phenomenon, and for illustration a trivial example is presented according to the probability theory.

Example 2.2

Let $Y_1(t) \in N(0,1))$, $Y_2(t) \in N(0,1)$ $(t = 1,2,..,n)$ be standard normally distributed series with expected values $\mathrm{E}(Y_1(t)) = \mathrm{E}(Y_2(t)) = 0$, with standard deviations $\mathrm{D}(Y_1(t)) = \mathrm{D}(Y_2(t)) = 1$, and with correlation between the series $\mathrm{corr}(Y_1(t), Y_2(t)) = \rho$ $(t = 1,2,..,n)$.

Then the mean difference $\mathrm{E}(Y_1(t) - Y_2(t)) = 0$ of course, however, the difference $Y_1(t) - Y_2(t)$ is not independent from the elements $Y_1(t)$, $Y_2(t)$ if $\rho \neq 1$, and, e.g., the conditional expectation of difference $Y_1(t) - Y_2(t)$ given $Y_1(t)$, or equivalently the regression of difference $Y_1(t) - Y_2(t)$ on $Y_1(t)$ is $\mathrm{E}(Y_1(t) - Y_2(t) | Y_1(t)) = (1 - \rho) \cdot Y_1(t)$.

Consequently, the difference $Y_1(t) - Y_2(t)$ is an expectedly monotonous increasing function of $Y_1(t)$ if $\rho \neq 1$. This is the theory, but it can be demonstrated in practice too. We generated such standard normal series by the Monte Carlo method with parameters $\rho = 0.9$, $n = 1000$. In this case,

$\mathrm{E}\left(Y_1(t)-Y_2(t)|Y_1(t)\right)=0.1\cdot Y_1(t)$ and the difference series $Y_1(t)-Y_2(t)$ as a function of series $Y_1(t)$ is plotted in *Fig. 1*.

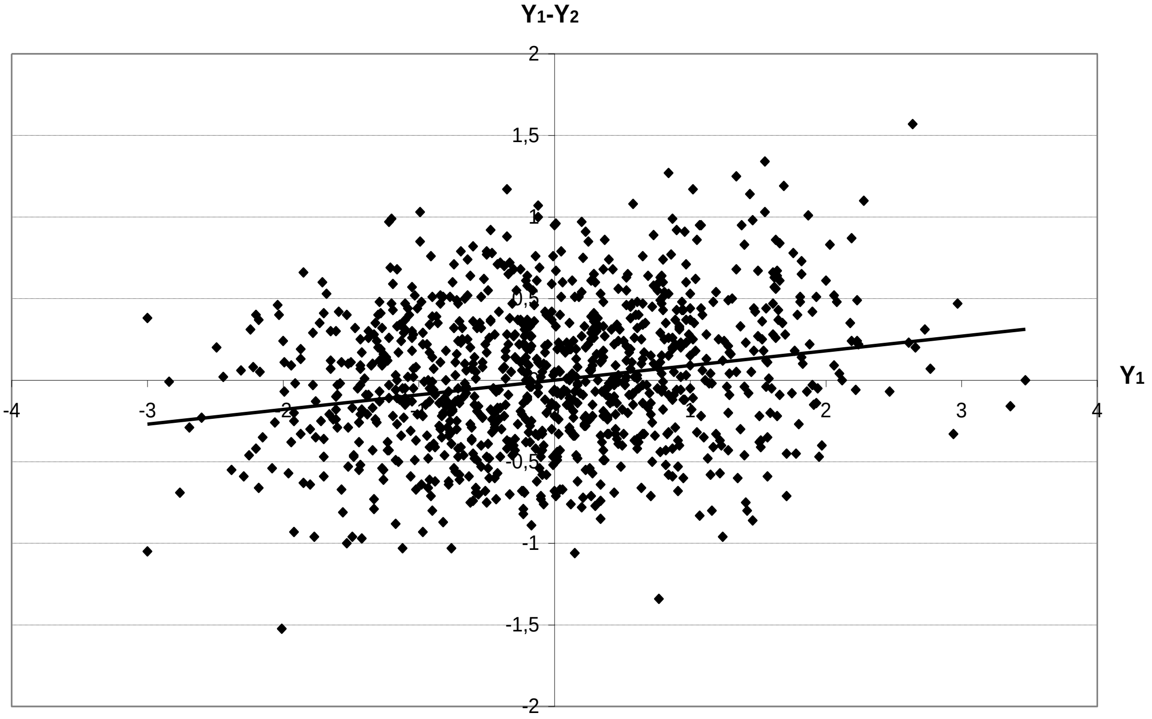

Fig. 1. Difference series $Y_1(t)-Y_2(t)$ as a function of series $Y_1(t)$

It is evident that the conditional expectation of difference $Y_1(t)-Y_2(t)$ is monotonous increasing function of $Y_1(t)$, consequently the difference may be larger mainly in the case of extreme values. It is a general phenomenon not only observed for meteorological measurements. Presumably this experience is the reason for the idea that the correction of daily data should vary according to the meteorological situation of each day, in particular on the basis of some regression models. But it is a misunderstanding of the homogenization problem.

3. Mathematical formulation of the daily data homogenization

Unfortunately, the exact theoretical, mathematical formulation of the problem of homogenization is generally neglected in meteorological studies. Therefore, we try to formulate this problem in accordance with mathematical conventions. First of all it is necessary to emphasize that homogenization is a distribution problem and not a regression one.

Notation

Let us assume we have daily data series:
$Y_1(t)\ (t=1,2,..,n)$: candidate time series of the new observing system.
$Y_2(t)\ (t=1,2,..,n)$: candidate time series of the old observing system.
$1 \le T < n$: change-point, series $Y_2(t)\ (t=1,2,..,T)$ can be used before and series $Y_1(t)\ (t=T+1,..,n)$ can be used after the change-point.

Definition

The aim of homogenization is the adjustment or correction of values $Y_2(t)\ (t=1,2,..,T)$ in order to have the corrected values $Y_{1,2h}(t)\ (t=1,2,..,T)$ with the same distribution as the elements of series $Y_1(t)\ (t=1,2,..,T)$, i.e.:

$$\mathrm{P}\left(Y_{1,2h}(t) < y\right) = \mathrm{P}\left(Y_1(t) < y\right), \quad y \in (-\infty,\infty)\ ,\ t=1,2,..,T\ . \qquad (1)$$

Eq. (1) means the equality in distribution: $Y_{1,2h}(t) \overset{d}{=} Y_1(t)\ \ (t=1,2,..,T)$.

Consequence

Within the same climate area, if the variables $Y_1(t), Y_2(t)\ (t=1,2,..,T)$ have identical distribution, i.e., $Y_2(t) \overset{d}{=} Y_1(t)\ \ (t=1,2,..,T)$, then the merged series $Y_2(t)\ (t=1,2,..,T)$, $Y_1(t)\ (t=T+1,..,n)$ is homogeneous.

Example

Let us assume we have parallel measurements $Y_1(t),\ Y_2(t)\ (t=1,2,..,n)$ within the same climate area with distance 50 m between the locations. Then, as a consequence of micrometeorological processes, the series are probably different, $Y_2(t) \neq Y_1(t)\ (t=1,2,..,n)$, but they may be equal in distribution, $Y_2(t) \overset{d}{=} Y_1(t)$ $(t=1,2,..,n)$. In this case, the mixed series $Y_2(t)\ (t=1,2,..,T)$, $Y_1(t)\ (t=T+1,..,n)$ can be taken as a homogeneous series. This mixed series is equivalent with the homogeneous series $Y_1(t)\ (t=1,2,..,n)$ also in respect of the distribution of extremes.

Returning to the general question, we have to see clearly that the aim of homogenization is to correct the distribution of $Y_2(t)$ according to $Y_1(t)$, instead of the estimation or regression of $Y_1(t)$on $Y_2(t)$! Moreover, the correction of distribution is equivalent in essence with the correction or adjustment of the moments. The aim of the homogenization expressed in k^{th} moments:

$$m_k = \mathrm{E}\left((Y_{1,2h}(t))^k\right) = \mathrm{E}\left((Y_1(t))^k\right) \quad k = 1,2,\ldots\,;\; t = 1,2,..,T\,, \tag{2}$$

where E is the usual notation of the expected value or mean equivalently. Some remarkable formulas for the moments:

$$E = m_1,\;\; D^2 = m_2 - m_1^2 \tag{3}$$

where E denotes the expected value or mean, and D denotes the standard deviation.

In practice, numerous methods indicate the capability to correct the higher order moments but without any exact proof.

4. The variable correction methods

We return to the methods suggested by *Trewin* and *Trevitt* (1996) which was mentioned in Section 2.1. Essentially, the underlying principles of the variable correction procedures developed later were formulated based on these methods. We do not agree with these principles as explained by our argument in Sections 2.1 and 3, but let us see some details and properties of the mathematical consequences.

4.1. The Trewin and Trevitt (1996) methods for parallel measurements

The short description is cited word for word again from the paper of *Della-Marta* and *Wanner* (2006):

"Trewin and Trevitt (1996) present three different methods to build a composite daily temperature series. Essential to the methods is the existence of simultaneous (in time) observations from the *new* and *old* observing system. These parallel measurements had been taken based on the recommendations of Karl et al. (1995), who suggest that a minimum of a 2-yr overlap between the new and old observing systems be made. In Australia, for example, this practice has only become routine since around 1993 and so many inhomogeneities needed to be adjusted using the traditional constant difference techniques with neighboring reference stations. In this way, Trewin (2001) created a homogenized daily temperature dataset that has subsequently been used by Collins et al. (2000) to assess trends in the frequency of extreme temperature events in Australia.

The three methods they intercompared were constant difference, linear regression, and frequency distribution matching.

The constant difference approach simply adjusted the older data with the newer data using the mean of the daily differences in the simultaneous (parallel) measurements.

The linear regression method fitted a linear model to the difference in daily simultaneous measurements between the two observing systems and the temperature at the older station. This model could then be used to adjust daily temperatures at the older station differentially depending on the temperature, thereby adjusting the higher-order moments.

Their third method determines the frequency distribution of each site during the simultaneous measurement period. The adjustment for each desired percentile is calculated simply as the difference between the two percentiles. This method assumes that there is no systematic bias in the rank order of the temperatures at the two sites.

They show that both the regression method and the frequency distribution matching technique have certain advantages; however, if the homogenization of extreme events is most needed, then their frequency distribution matching technique is more accurate."

Our mathematical comments to the methods are as follows.

4.1.1. Constant difference approach

Yes, this approach is correct if the inhomogeneity is in mean or expected value or first moment, which are the same with different names.

4.1.2. Linear regression method

This procedure is absolutely wrong for homogenization. To demonstrate the problem, a trivial counter-example is presented.

Theorem

Let us assume that the different series $Y_1(t)$, $Y_2(t)$ $(t=1,2,..,n)$ have identical distribution, with expected values $\mathrm{E}(Y_1(t))=\mathrm{E}(Y_2(t))=0$, standard deviations $\mathrm{D}(Y_1(t))=\mathrm{D}(Y_2(t))=1$, and correlation between the series $\mathrm{corr}(Y_1(t),Y_2(t))=\rho$ $(t=1,2,..,n)$.

(i) Then the linear regression of difference $Y_1(t)-Y_2(t)$on $Y_2(t)$ is $(\rho-1)\cdot Y_2(t)$, consequently, the homogenized series after the suggested adjustment, $Y_{1,2h}(t)=Y_2(t)+(\rho-1)\cdot Y_2(t)=\rho\cdot Y_2(t)$ and $\rho\cdot Y_2(t)$ is just the linear regression of $Y_1(t)$on $Y_2(t)$.

(ii) Moreover, since the expected values $\mathrm{E}(Y_{1,2h}(t))=\mathrm{E}(Y_1(t))=\mathrm{E}(Y_2(t))=0$, therefore – using Eq. (3) – , the second moment of $Y_{1,2h}(t)$is equal to the variance $\mathrm{D}^2(Y_{1,2h}(t))=\delta^2<1$, while the common second moment of $Y_1(t)$,

$Y_2(t)$ is equal to the variances $\mathrm{D}^2(Y_1(t)) = \mathrm{D}^2(Y_2(t)) = 1$. Therefore, the second moment was decreased from 1 to $\delta^2 < 1$during the regression.
Summing up, according to (i) this procedure is equivalent with the simple linear regression of $Y_1(t)$on $Y_2(t)$. Furthermore, according to (ii) the following statement about the method is absolutely false: "This model could then be used to adjust daily temperatures at the older station differentially depending on the temperature, thereby adjusting the higher-order moments." The truth is just the opposite, since the correct second moment was damaged at our counter-example.

4.1.3. Frequency distribution matching technique

The main problem is the following assumption which is the fundament of the method: "This method assumes that there is no systematic bias in the rank order of the temperatures at the two sites."
Unfortunately, the reality and the mathematics are much more complicated, and the above assumption cannot be accepted as it is demonstrated in *Fig. 1*. The bias in rank order depends on the stochastic connection, and there may be systematic bias, since $Y_1(t)$, $Y_2(t)$ are not monotonous increasing functions of each others. At this method, the adjusted $Y_{1,2h}(t)$ is obtained essentially by a simple exchange $Y_2(t)$ for $Y_1(t)$ according to the rank orders. Why? For example, if $Y_1(t)$, $Y_2(t)$ were equal in distribution then such an exchange would not be necessary.

4.2. The general type of variable correction methods applied in the practice

On the basis of the former principles described in Sections 4.1.2 and 4.1.3 (regression and frequency distribution matching), a number of variable correction methods have been developed during the last years. The new improvement of these methods is that they do not need overlap observations, instead of this they use information from nearby reference stations, for example higher order moments (HOM) method by *Della-Marta* and *Wanner* (2006) and spline daily homogenization (SPLIDHOM) method by *Mestre et al.* (2011). We do not want to criticize the details of these methods however, we express again our skepticism on their common fundamental principles which were based on a pseudo problem demonstrated in *Example 2.2*. Moreover, we repeat the following sources of errors for consideration.

- The assumption of the frequency distribution matching technique, i.e., there is no systematic bias in the rank, cannot be accepted.

- The regression methods are not adequate to correct the higher order moments.

Our last remark is connected also with the higher order moments. In general, the papers about these methods indicate the capability to correct the higher order moments, but this statement is always without any exact mathematical proof. We are skeptic, however if somebody could send us a nice proof, we would be grateful for it.

5. Some remarks about the homogenization in the higher-order moments

We suggest to consider the following remarks when developing homogenization methods with the capability to correct also the higher order moments.

Remark 1

There is a common assumption that the correction in mean is sufficient for monthly and annual series, and that the correction of higher order moments is necessary only in the case of daily data series. In general, it is tacitly assumed that the averaging is capable to filter out the inhomogeneities in the higher order moments. However, this assumption is false, for example, if there is an inhomogeneity in the standard deviation of daily data, we may have the same inhomogeneity in monthly data.

Proof

Daily data are $X(t)\,(t=1,2,..,30)$, monthly average is $\overline{X}=\frac{1}{30}\sum_{t=1}^{30}X(t)$.

Let us introduce an inhomogeneity in the standard deviation for the daily data:
$X_{ih}(t)=\alpha\cdot\left(X(t)-\mathrm{E}\left(X(t)\right)\right)+\mathrm{E}\left(X(t)\right),\ \ (t=1,2,..,30)$.

The expected value is unchanged: $\mathrm{E}\left(X_{ih}(t)\right)=\mathrm{E}\left(X(t)\right)$, but the standard deviation has changed: $\mathrm{D}\left(X_{ih}(t)\right)=\alpha\cdot\mathrm{D}\left(X(t)\right)$.

Let us see the new monthly average: $\overline{X}_{ih}=\frac{1}{30}\sum_{t=1}^{30}X_{ih}(t)$.

The expected value is unchanged: $\mathrm{E}\left(\overline{X}_{ih}\right)=\mathrm{E}\left(\overline{X}\right)$, but the standard deviation changed with the same measure:

$$\mathrm{D}\left(\overline{X}_{ih}\right)=\mathrm{D}\left(\frac{1}{30}\sum_{t=1}^{30}X_{ih}(t)\right)=\mathrm{D}\left(\frac{1}{30}\sum_{t=1}^{30}\alpha\cdot X(t)\right)=\alpha\cdot\mathrm{D}\left(\frac{1}{30}\sum_{t=1}^{30}X(t)\right)=\alpha\cdot\mathrm{D}\left(\overline{X}\right).$$

Remark 2

The correction in the first two moments or, equivalently, in mean and standard deviation can be formulated by using the notations defined in Section 3 as follows:

$$Y_{1,2h}(t) = E_1(t) + \frac{D_1(t)}{D_2(t)}\left(Y_2(t) - E_2(t)\right) \quad \left(t = 1,2,..,T\right) , \qquad (4)$$

where $E_1(t) = \mathrm{E}\left(Y_1(t)\right)$, $E_2(t) = \mathrm{E}\left(Y_2(t)\right)$ are the means, and $D_1(t) = \mathrm{D}\left(Y_1(t)\right)$, $D_2(t) = \mathrm{D}\left(Y_2(t)\right)$ are the standard deviations. Then $\mathrm{E}\left(Y_{1,2h}(t)\right) = E_1(t)$, $\mathrm{D}\left(Y_{1,2h}(t)\right) = D_1(t)$.

In general, the detection of the change points and the estimation of correction factors are suggested to be based on the examination of monthly data series because of the larger signal to noise ratio.

Remark 3

If the joint distribution of the series is normal, $Y_1(t) \in N\left(E_1(t), D_1(t)\right)$, $Y_2(t) \in N\left(E_2(t), D_2(t)\right)$ $\left(t = 1,2,..,n\right)$ and $Y_{1,2h}(t)$ $\left(t = 1,2,..,T\right)$, calculated according to Eq. (4), then $Y_{1,2h}(t), Y_1(t)$ $\left(t = 1,2,..,T\right)$ have identical distribution: $Y_{1,2h}(t) \overset{d}{=} Y_1(t)$ $\left(t = 1,2,..,T\right)$. Consequently, the mixed series $Y_{1,2h}(t)$ $\left(t = 1,2,..,T\right)$, $Y_1(t)$ $\left(t = T+1,..,n\right)$ is homogeneous, that means it is sufficient to correct only the first two moments in case of joint normal distribution.

Proof

Owing to Remark 2 and the joint normal distribution, $Y_{1,2h}(t) \in N\left(E_1(t), D_1(t)\right)$ $\left(t = 1,2,..,T\right)$.

6. Conclusion

It is necessary to define the exact mathematical theory for homogenization of climate data series. Homogenization is a probability distribution problem, and the methods applied in practice should be theoretically evaluated in this respect.

References

Della-Marta, P.M. and *Wanner, H.*, 2006: A Method for homogenizing the extremes and mean of daily temperature measurements. *J. Climate 19*, 4179–4197.

Mestre, O., Gruber, C., Prieur, C., Caussinus, H., and *Jourdain, S.*, 2011: Splidhom: a method for homogenization of daily temperature observations. *J. Appl. Meteor. Climatol. 50*, 2343–2358.

Szentimrey, T., 2008: Development of MASH homogenization procedure for daily data, *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases,* Budapest, Hungary, 2006; *WCDMP-No. 68, WMO-TD NO. 1434*, 116–125.

Trewin, B.C. and *Trevitt, A.C.F.*, 1996: The development of composite temperature records. *Int. J. Climatol. 16*, 1227–1242.