DOI:10.28974/idojaras.2024.2.5

IDŐJÁRÁS

Quarterly Journal of the HungaroMet Hungarian Meteorological Service
Vol. 128, No. 2, April – June, 2024, pp. 219–235

Development of new version MASHv4.01 for homogenization of standard deviation

Tamás Szentimrey

Varimax Limited Partnership, Budapest, Hungary

Author E-mail: szentimrey.t@gmail.com

(Manuscript received in final form October 31, 2023)

Abstract— The earlier versions of our method MASH (Multiple Analysis of Series for Homogenization; *Szentimrey*) were developed for homogenization of the daily and monthly data series in the mean, i.e., the first order moment. The software MASH was developed as an interactive automatic, artificial intelligence (AI) system that simulates the human intelligence and mimics the human analysis on the basis of advanced mathematics. This year we finished the new version MASHv4.01 that is able to homogenize also the standard deviation, i.e., the second order moment. The problem of standard deviation is related to the monthly and daily data series homogenization.

Key-words: climate data series, homogenization, mathematical formulation, normal distribution, adjustment of standard deviation, AI system, MASH, MISH

1. Introduction

In essence, the theme of homogenization can be divided into two subgroups, such as monthly and daily data series homogenization. These subjects are in strong connection with each other of course, for example the monthly results can be used for the homogenization of daily data. In the practice, the monthly series are homogenized in the mean only, while there exist some trials to homogenize the daily series also in higher order moments. These procedures are based on a popular assumption that is the adjustment of mean is sufficient for monthly series,

and the adjustment 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 inhomogeneity in the higher order moments. However, this assumption is false, since it can be proved if there is a common inhomogeneity in the standard deviation of daily data then we have the same inhomogeneity in the monthly data. Therefore, we developed a mathematical procedure for the homogenization of mean and standard deviation together. This developed procedure was incorporated into our new software version MASHv4.01 (Multiple Analysis of Series for Homogenization; *Szentimrey*, 2023b), which is based on the examination of different type of monthly series, and the monthly results are applied for the homogenization of daily series. We remark if the data are normally distributed (e.g., mean temperature) then the homogenization of mean and standard deviation is sufficient, since in case of normal distribution if the first two moments are homogenous then the higher order moments are also homogeneous. The most important novelty of this paper is the methodology for adjusting the mean and standard deviation together, as detailed in Sections 5.3 and 5.4. However, first we need to review the mathematical, methodological background.

As for the MASH as an artificial intelligence system, it is not just a "buzzword", as MASH has been developed along these lines for many years. To illustrate this, here is a quote from the proceedings of the 4th Homogenization Seminar (*Szentimrey*, 2004):

"Programmed Statistical Procedure (Software: MASHv2.03)

EXAMPLE

Let us assume that there is a difficult stochastic problem.

In case of having relatively few statistical information:

– an intelligent human is possibly able to solve the problem, but it is time-consuming,

– the solution of the problem cannot be programmed.

In case of increasing the amount of statistical information:

– one is unable to discuss and evaluate all the information,

– but then the solution of the problem can be programmed (as in chess expert systems).

AIM, REQUIREMENT

– Development of mathematical methodology in order to increase the amount of statistical information.

– Development of algorithms for optimal using of both the statistical and the metadata information."

In essence, the Deep Blue chess expert system, which defeated Garry Kasparov in 1997, motivated the development of MASH.

In our conception, the meteorological questions and topics cannot be treated separately. Therefore, we present a block diagram (*Fig. 1*) to illustrate the possible

connection between various important meteorological topics. The software MASH and MISH (Meteorological Interpolation based on Surface Homogenized Data Basis; *Szentimrey* and *Bihari*, 2014) were developed by us. These software were applied also in the CARPATCLIM project (*Szentimrey et al.*, 2012a,b; *Lakatos et al.*, 2013). The paper of *Izsák et al.* (2022) presents another application to create a representative database for Hungary.

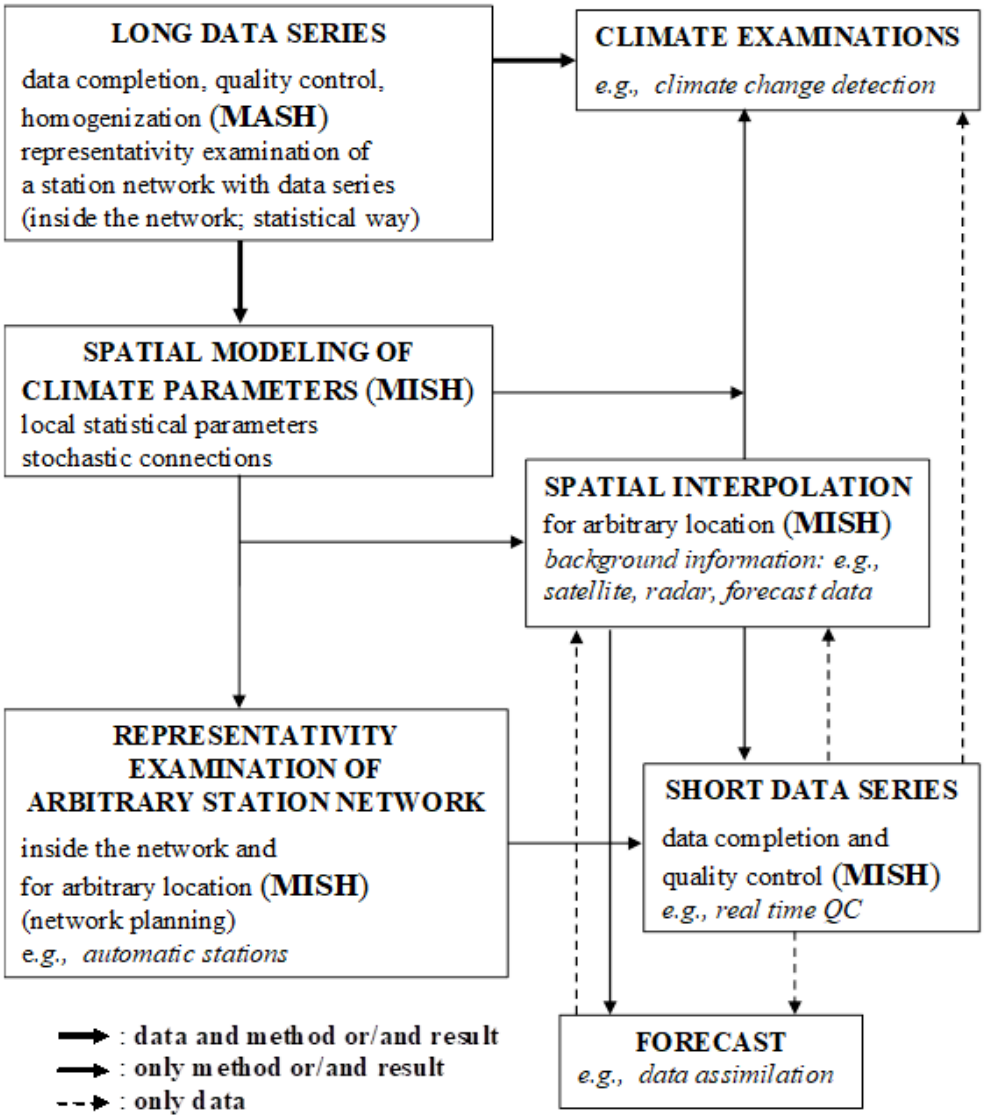

Fig. 1. Block diagram for the possible connection, between various basic meteorological topics and systems.

Finally, last but not least, in Section 7 we reflect to some incorrect sentences about MASH from a book.

2. Mathematical formulation of climate data homogenization

Unfortunately, the exact theoretical, mathematical formulation of the problem of homogenization is neglected at the meteorological studies in general. Therefore, we try to formulate this problem in accordance with the mathematical conventions. First we emphasize that the homogenization is a distribution problem and not a regression one (*Szentimrey*, 2013).

2.1. General mathematical formulation

Notation

Let us assume we have daily or monthly climate 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.
$1 \le T < n$: changepoint, 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 changepoint.

The appropriate theoretical cumulative distribution functions (CDF) are:

$$F_{1,t}(y) = P(Y_1(t) < y) \ , \quad F_{2,t}(y) = P(Y_2(t) < y) \quad y \in (-\infty, \infty) \ , t = 1,2,..,n$$

It is very important to remark that as a consequence of some natural changes - e.g. annual cycle, climate change - the series of distribution functions $F_{1,t}(y)$, $F_{2,t}(y)$ $(t = 1,2,..,n)$ may change in time! In the statistical climatology the climate change is equivalent with the changing probability of the meteorological events. The inhomogeneity of data series can be defined on the basis of the distribution functions.

Definition 1

The merged series $Y_2(t)$ $(t = 1,2,..,T)$, $Y_1(t)$ $(t = T+1,..,n)$ is inhomogeneous, if the identity of the distribution functions $F_{2,t}(y) \equiv F_{1,t}(y)$ $(t = 1,2,..,T)$ is not true.

Definition 2

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

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

This definition means the equality in distribution, i.e.: $Y_{1,2h}(t) \stackrel{d}{=} Y_1(t)$ $(t = 1,2,..,T)$. ("d over =", simply means that the two distribution functions are the same.)

Theorem 1

Let us assume about the random variables Y_1,Y_2 and their distribution functions $F_1(y)$,$F_2(y)$, that $P\left(Y_j \in (a_j, b_j)\right) = 1$ and $F_j(y)$ is a strictly increasing continuous function on the interval (a_j, b_j) $(j = 1,2)$. Then applying the transfer function $Y_{1,2h} = F_1^{-1}(F_2(Y_2))$ we obtain that the variable $Y_{1,2h}$ has the same distribution as Y_1, i.e: $\mathrm{P}(Y_{1,2h} < y) = \mathrm{P}(Y_1 < y) = F_1(y)$.

Definition 3

Transfer function: $F_{1,t}^{-1}\left(F_{2,t}(y)\right)$ and quantile function: $F_{1,t}^{-1}(p)$.
Theoretical formulation of homogenization of $Y_2(t)$ $(t = 1,2,..,T)$:
$Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}(Y_2(t))\right)$, then $\mathrm{P}(Y_{1,2h}(t) < y) = F_{1,t}(y)$.

Remark

The basis of the Quantile Matching methods can be integrated into the general theory. However, these methods developed in practice mainly for daily data are very weak empiric methods.

2.2 Mathematical formulation for normal distribution

The homogenization problem is very complicated in the general case, however in the case of normal distribution, a much simpler mathematical formula can be obtained. We emphasize that the normal distribution is a special case, but it is a basic one in mathematical statistics as well as in the meteorology. For example, the normal distribution model can be accepted for temperature variables in general.

Theorem 2

Let us assume the elements of data series $Y_1(t)$, $Y_2(t)$ are normally distributed, that is,
$Y_1(t) \in N(E_1(t), D_1(t))$, $Y_2(t) \in N(E_2(t), D_2(t))$ $(t = 1,2,..,n)$,
where $\mathrm{E}(Y_1(t)) = E_1(t)$, $\mathrm{E}(Y_2(t)) = E_2(t)$ are the means or expected values and $\mathrm{D}(Y_1(t)) = D_1(t)$, $\mathrm{D}(Y_2(t)) = D_2(t)$ are the standard deviations.

Then the transfer formula of homogenization:

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

In case of normal distribution, we have a much simpler transfer formula for adjustment than the general form $Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}(Y_2(t))\right)(t = 1,2,..,T)$.
This simple linear formula means that, if the data series have normal distribution, it is sufficient to homogenize the means (*E*) and standard deviations (*D*) only that is equivalent with the homogenization of the first two moments. We emphasize that the normal distribution is a basic model in the mathematical statistics as well as in the meteorology, and there is no "tail distribution" problem (*Szentimrey*, 2023a) for this important distribution according to the *Theorem 2*! For normal distribution, if the means and standard deviations are homogenous then the higher order moments are also homogeneous, and there is not any inhomogeneity in the tails of the distributions.

3. Relation of daily and monthly data homogenization

The monthly and daily data series homogenization are in strong connection with each other of course, for example the monthly results can be used for the homogenization of daily data.
If we have daily data series, the general way of homogenization is

- calculation of monthly series,
- homogenization of monthly series taking advantage of the larger signal to noise ratio,
- homogenization of daily series using the detected monthly inhomogeneities.

So we have the question: how can we use the valuable information of detected monthly inhomogeneities for the daily data homogenization?

4. Methodology for homogenization of monthly series

This section is an overview of some various theoretical aspects of monthly series homogenization (*Szentimrey*, 1999, 2008, 2021, 2023a,b; *Venema et al.*, 2012). The aim of these homogenization procedures is to detect the inhomogeneities of monthly series and to adjust the series. In connection with such homogenization methods, we have to give solutions for the following mathematical problems: relative models, statistical spatiotemporal modeling of the series, methodology for comparison of series, breakpoint (changepoint), and outlier detection, methodology for adjustment of series, quality control procedures, missing data completion, usage of metadata, relation of daily and

monthly homogenization, manual versus automatic methods, evaluation of methods (theoretical, benchmark validation). The following Sections 4.1–4.2 are related to the Chapter 5 of the WMO Guidelines on Homogenization (*WMO*, 2020).

4.1. General structure of the relative spatiotemporal models

Relative methods can be applied if there are more station data series given, which can be compared mutually. In this case, the statistical spatiotemporal modeling of the series is a basic question. The adequate comparison, breakpoint detection, and adjustment procedures depend on the chosen statistical model. Depending on the climate elements, additive or multiplicative models are applied.

4.1.1. The additive spatiotemporal model

This model is based on the normal distribution (Section 2.2) and it can be used if the data series are quasi normally distributed (e.g., temperature). For this model we assume inhomogeneity of mean (E), i.e., expected value.

In case of relative methods, a general form of additive model for more monthly series belonging to the same month in a small climate region can be written as follows (*WMO*, 2020):

$$X_j(t) = \mu(t) + E_j + IH_j(t) + \varepsilon_j(t) \quad (j = 1,2,\dots,N\ ;\ t = 1,2,\dots.,n)\ , \qquad (3)$$

where $\mu(t)$ is the common and unknown climate change signal, E_j are the spatial expected values, $IH_j(t)$ are the inhomogeneity signals, and $\varepsilon_j(t)$ are normal white noise series.

The type of inhomogeneity$IH(t)$ is in general a step-like function with unknown breakpoints T and artificial shifts $IH(T) - IH(T+1) \neq 0$. The normal distributed vector variables $\boldsymbol{\varepsilon}(t) = [\varepsilon_1(t),..,\varepsilon_N(t)]^T \in N(\mathbf{0},\mathbf{C})(t = 1,\dots,n)$ are totally independent in time. The spatial covariance matrix $\mathbf{C}$ describes the spatial structure of the series, which is important for comparison of series.

4.1.2. The multiplicative spatiotemporal model

If the data series are quasi lognormal distributed (e.g., precipitation) then the multiplicative model can be used. According to this model, the monthly series belonging to the same month in a small climate region can be written as follows:

$$X_j(t) = \mu(t) \cdot E_j \cdot IH_j(t) \cdot exp\big(\varepsilon_j(t)\big)\ \ (j = 1,2,\dots,N\ ;\ t = 1,2,\dots.,n)\ . \qquad (4)$$

This multiplicative model can be transformed into the additive one by certain logarithmic procedure, where the little values near zero are increased slightly before logarithmization. Therefore, the homogenization methodology

(Section 4.2) developed for additive model can be used with some modification also for the multiplicative model.

4.2. Methodological questions for additive spatiotemporal model

4.2.1. Methodology for comparison of series

The problem of comparison of series is related to the following questions: reference series creation, difference series constitution, multiple comparisons of series, etc. This topic is very important for detection as well as for adjustment, because the efficient series comparison can increase both the significance and the power. The development of efficient comparison methods can be based on the examination of the spatial covariance structure of data series. The examined series $X_j(t)$ have to be taken as candidate and reference series alike, furthermore, the homogeneity of the reference series is not assumed!

4.2.2. Methodology for breakpoint (changepoint) detection

One of the basic tasks of the homogenization is the examination of the difference series in order to detect the breakpoints and to attribute the appropriate ones for the candidate series.

The more sophisticated multiple breakpoints detection procedures were developed for joint detection of the breakpoints. There may be different principles of the detection methods that are classical ways in the mathematical statistics.
Multiple breakpoints detection procedures for difference series are as follows.

a) Bayesian approach (model selection, segmentation), penalized likelihood methods:
   PRODIGE (*Caussinus* and *Mestre*, 2004), HOMER (*Mestre et al.*, 2013), ACMANT (Adapted Caussinus-Mestre Algorithm for Networks of Temperature series; *Domonkos*, 2011).
b) Multiple breakpoints detection based on test of hypothesis and confidence intervals for the breakpoints, that make possible, automatic use of metadata: MASH (*Szentimrey*, 1999, 2023b).

4.2.3. Methodology for adjustment of series

Beside the detection, another basic task of the homogenization is the adjustment of series. Calculation of the adjustment factors can be based on the examination of difference series for estimation of shifts at the detected breakpoints. In general, the methods use point estimation for the shifts at the detected breakpoints.

There are methods that use the standard least squares technique after breakpoint detection procedure for joint estimation of the shifts of all the examined series, for example the methods PRODIGE, HOMER, ACMANT. Probably the generalized least squares estimation technique based on spatial

covariance matrix would be more efficient, and it would be equivalent with the maximum likelihood estimation for the shifts in the case of normal distribution.

Another way is that the calculation of the adjustment factors is based on some confidence intervals given for the shifts at the detected breakpoints, as in the method MASH. The confidence intervals given for the breakpoints and shifts make the automatic use of metadata possible.

5. Methodology for homogenization of daily series

The basic question is what would be the appropriate, exact methodology for the daily data homogenization. According to Sections 3 and 4, how can we use the valuable information of detected monthly inhomogeneities for the daily data homogenization (*Szentimrey*, 2008, 2013, 2017, 2021, 2023a,b)? How can we use the methodology developed for monthly series?

5.1. A popular procedure for daily data, e.g., the variable correction methods

The typical steps of the procedure are as follows.

1. Calculation of monthly series from daily series.
2. Homogenization of monthly series:
   Breakpoints detection, adjustment in the first order moment (mean (*E*)).
   Assumption: homogeneity of higher order moments (e.g., standard deviation (*D*)).
3. Homogenization of daily series:
   A trial to homogenize also the higher order moments.
   (Quantile matching (*Wang* and *Feng*, 2013), spline methods (*Mestre et al.*, 2011))

The used monthly information are only the detected breakpoints.

However, the following questions are arising at this procedure:

– Is it an adequate model that we have inhomogeneity in higher moments only at daily series but not at monthly ones? Can this model be accepted according to the probability theory? No, it can be proved, if there is a common inhomogeneity in the standard deviation (*D*) of daily data, we may have the same inhomogeneity in monthly data.
– Why are the monthly adjustment factors not used for daily homogenization? It seems to lose some valuable information obtained during the monthly homogenization.

5.2. Problem of inhomogeneity of the standard deviation

According to the former assumption, which is popular in practice, the adjustment in mean (*E*) is sufficient for monthly and annual series, and the adjustment 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, it can be proved, if there is a common inhomogeneity in the standard deviation (*D*) of daily data, we may have the same inhomogeneity in monthly data.

Theorem 3

Let us assume $Y(t)$ $(t = 1,..,30)$ are daily data and the monthly mean is $\bar{Y} = \frac{1}{30}\sum_{t=1}^{30} Y(t)$.

The monthly variable for examination of the inhomogeneity of standard deviation (*D*) is

$$S = \sqrt{\frac{1}{2\cdot 29}\sum_{t=2}^{30}(Y(t) - Y(t-1))^2}\ . \qquad (5)$$

Let us introduce some inhomogeneity of the mean (*E*) and the standard deviation (*D*) for the daily data by a linear function:

$Y_{ih}(t) = \alpha \cdot \big(Y(t) - \mathrm{E}(Y(t))\big) + \mathrm{E}(Y(t)) + \beta \qquad (t = 1,..,30).$

Then the expected values and the standard deviations are:

$\mathrm{E}(Y_{ih}(t)) = \mathrm{E}(Y(t)) + \beta\ ,\ \mathrm{D}(Y_{ih}(t)) = \alpha \cdot \mathrm{D}(Y(t)) \quad (t = 1,..,30).$

The appropriate monthly variables are:

$\bar{Y}_{ih} = \frac{1}{30}\sum_{t=1}^{30} Y_{ih}(t), \qquad S_{ih} = \sqrt{\frac{1}{2\cdot 29}\sum_{t=2}^{30}(Y_{ih}(t) - Y_{ih}(t-1))^2}.$

i) Then the monthly mean is also inhomogeneous in mean (*E*) and standard deviation (*D*) with the same measure like the daily values: $\mathrm{E}(\bar{Y}_{ih}) = \mathrm{E}(\bar{Y}) + \beta$ and $\mathrm{D}(\bar{Y}_{ih}) = \alpha \cdot \mathrm{D}(\bar{Y})$.

ii) Moreover, variables S, S_{ih} can be used to estimate the inhomogeneity α of the standard deviation (*D*): $\mathrm{E}(S_{ih}) = \alpha \cdot \mathrm{E}(S)$

5.3. The alternative procedure for daily and monthly data developed in MASH

We suggest an alternative procedure to homogenize both the daily and the monthly series.
The steps of the procedure in case of quasi normal distribution (e.g., temperature) are as follows.
First we examine both the monthly mean series $\bar{Y}(t)$ for the inhomogeneity of expected values (*E*) and the monthly series $S(t)$ derived to *Theorem 3* for the inhomogeneity of standard deviations (*D*). The proper inhomogeneity characteristics are the difference of the expected values and the ratio of the standard deviations. Therefore, we apply different models for these parameters.

1. Homogenization of monthly series $S(t)$, $\bar{Y}(t)$.
   Homogenization of series $S(t)$ by the multiplicative model (4.1.2): breakpoints detection, estimation of inhomogeneity of standard deviation (*D*).
   Adjustment of standard deviation of series $\bar{Y}(t)$ is detailed in Section 5.4.
   Homogenization of adjusted series $\bar{Y}(t)$ by additive model (4.1.1): breakpoints detection, estimation of the inhomogeneity of mean (*E*). It is detailed in Section 5.4.
   Assumption: after homogenization of *E*, *D*, there is no inhomogeneity in the higher order (>2) moments of adjusted series $\bar{Y}(t)$. This assumption is always right in case of normal distribution according to *Theorem 2*.
2. Homogenization of daily series.
   Homogenization of mean (*E*) and standard deviation (*D*) on the basis of the monthly results. The used monthly information are the breakpoints and the monthly adjustments of the mean (*E*) and standard deviation (*D*). The adjustment is based on the transfer formula (Eq (2)) considering *Theorem 3*. If the daily data are normally distributed then after homogenization of *E*, *D* there is no inhomogeneity in the higher order moments according to *Theorem 2*.

5.4. Adjustment of monthly mean series, daily series, transfer formula

The adjustment of monthly mean series $Y_2(t)$ in mean (*E*) and standard deviation (*D*) is based on the transfer formula according to Eq. (2), i.e.:

$$Y_{1,2h}(t) = \frac{D_1(t)}{D_2(t)}(Y_2(t) - E_2(t)) + E_1(t) \quad (t = 1,2,..,T)\,. \tag{6}$$

5.4.1. Adjustment of $Y_2(t)$ in standard deviation (D)

The theoretical formula is $Y_{1,2hD}(t) = \frac{D_1(t)}{D_2(t)}(Y_2(t) - E_2(t)) + E_2(t)$ $(t = 1,2,..,T)$ but $E_2(t)$ $(t = 1,2,..,T)$ are unknown.
Therefore, the applied formula is $Y_{1,2hD}(t) = \frac{D_1(t)}{D_2(t)}(Y_2(t) - \bar{E}_2) + \bar{E}_2$,
where $\bar{E}_2$ is the mean value of $E_2(t)$ $(t = 1,2,..,T)$. $\bar{E}_2$ can be estimated by mean $\bar{Y}_2$.
Inhomogeneity of standard deviation $IH_D(t) = \frac{D_2(t)}{D_1(t)}$ can be estimated by homogenizing the monthly standard deviation series $S(t)$ using the multiplicative model.
The adjustment of $Y_2(t)$ is $Y_{1,2hD}(t) = \frac{Y_2(t)}{IH_D(t)} - IH_{D,E}(t)$,
where $IH_{D,E}(t) = \left(\frac{D_1(t)}{D_2(t)} - 1\right)\bar{E}_2$.

5.4.2. Adjustment of $Y_{1,2hD}(t)$ in mean (E)

According to Eq. (6), the inhomogeneity of $Y_{1,2hD}(t)$ in mean is $IH_{E,D}(t) = \mathrm{E}\left(Y_{1,2hD}(t)\right) - E_1(t)$, and it can be estimated by homogenizing the monthly mean series $Y_{1,2hD}(t)$ using additive model.
$IH_{E,D}(t) = \left(\frac{D_1(t)}{D_2(t)} - 1\right)(E_2(t) - \bar{E}_2) + E_2(t) - E_1(t)$.

5.4.3. Summary of the adjustment of $Y_2(t)$ and the daily data series

The adjustment of $Y_2(t)$ in mean (*E*) and standard deviation (*D*) can be written in the following linear function form:
$Y_{1,2h}(t) = \frac{Y_2(t)}{IH_D(t)} - IH_E(t)$, where $IH_E(t) = IH_{D,E}(t) + IH_{E,D}(t)$.
For homogenization of daily data series in mean (*E*) and standard deviation (*D*) we also use this linear function form, in accordance with the *Theorem 3*. In this case the estimated inhomogeneity values $IH_D(t)$, $IH_E(t)$ are smoothed, as it was developed in MASH for homogenization of daily data (*Szentimrey*, 2008, 2013).

6. Summary of software MASH

6.1. General comments

The new version MASHv4.01 (Multiple Analysis of Series for Homogenization; *Szentimrey* 1999, 2004, 2008, 2013, 2017, 2021, 2023a,b,c) has been developed

for homogenization of daily and monthly series. The most important novelty of this version is the homogenization in standard deviation (*D*) beside the mean (*E*), see Sections 5.3, 5.4. The basic concept of the MASH system is first to homogenize the monthly series derived from the daily series, and then to homogenize the daily series based on the detected monthly inhomogeneities. The procedures depend on the distribution of climate elements, and additive or multiplicative models can be used.

6.1.1. Quasi normal distribution (e.g., temperature)

Beside the monthly mean series, another type of monthly series are also derived to estimate the inhomogeneity of standard deviation (*D*). These latter series can be homogenized by the multiplicative model (4.1.2), and the monthly mean series can be adjusted with the estimated inhomogeneity in standard deviation (*D*). The adjusted monthly mean series can be homogenized in mean (*E*) by the additive model (4.1.1).

6.1.2. Quasi lognormal distribution (e.g., precipitation)

Monthly mean or sum series can be homogenized by the multiplicative model (4.1.2).
The multiplicative model can be transformed into the additive one (4.1.1) by certain logarithmic procedures.

6.2. The most important features of the MASH system

Homogenization of monthly series:

- Relative homogeneity test procedure.
- Step by step iteration procedure: the role of series (candidate, reference) changes step by step in the course of the procedure.
- Interactive automatic, artificial intelligence (AI) system (see Section 6.3).
- Additive or multiplicative model can be used depending on the distribution.
- Including automatic quality control and missing data completion.
- Providing the homogeneity of the seasonal and annual series as well.
- Metadata (probable dates of breakpoints) can be used automatically.
- The homogenization results and the metadata can be verified.

Homogenization of daily series:

- Based on the detected monthly inhomogeneities (*E, D*).
- Including automatic quality control and missing data completion for daily data.

The elder version of MISH and MASH software can be downloaded from:
http://www.met.hu/en/omsz/rendezvenyek/homogenization_and_interpolation/software/
We plan to share the new version MASHv4.01 this year (2023).

6.3. Some verification results for homogenization in mean (E) and standard deviation (D)

The aim of MASH is not the full automation, and we also are sceptical in such an aspect. Our intention was to develop a flexible, interactive automatic, artificial intelligence (AI) system that simulates the human intelligence and mimics the human analysis on the basis of advanced mathematics. The mechanic, labor-intensive sub-procedures are fully automated, moreover, the operating process can be controlled simply, and the accidental mistakes can be corrected interactively. The basic idea of this concept is controlling the results via the verification tables generated automatically during the automatic procedures (*Szentimrey*, 2004, 2023b). Interactive decisions also can be made based on the analysis of the verification tables.

Some examples for verification tables related to the inhomogeneities are presented in *Fig. 2*. In the example, 15 Hungarian July mean temperature series (1901–2015) were homogenized by MASH in mean (*E*) and standard deviation (*D*). The estimated inhomogeneities can be characterized by the following statistics.

i) For mean (*E*, additive model):
$IHE = \frac{1}{n}\sum_{t=1}^{n}|IHE(t)|$, where $E_{ih}(t) = E(t) + IHE(t) \quad (t = 1,\dots,n)$, and $E_{ih}(t)$, $E(t)$ are the means before and after homogenization.

ii) For standard deviation (*D*, multiplicative model):
$IHD = \frac{100}{n}\sum_{t=1}^{n}|IHD(t) - 1|$, where $D_{ih}(t) = D(t) \cdot IHD(t)$ $(t = 1,\dots,n)$, and $D_{ih}(t)$, $D(t)$ are the standard deviations before and after homogenization.

These *IHE* and *IHD* statistics can be seen in *Fig. 2*.

Estimated Inhomogeneities for Mean (E) (°C)

Series	IHE	Series	IHE	Series	IHE
3	0.80	8	0.55	15	0.53
7	0.52	12	0.48	10	0.48
14	0.31	6	0.31	5	0.29
11	0.24	1	0.23	4	0.14
9	0.13	2	0.09	13	0.08

AVERAGE: 0.35

Estimated Inhomogeneities for St. Deviation (D) (%)

Series	IHD	Series	IHD	Series	IHD
8	8.05	9	7.98	4	6.73
12	4.88	7	4.08	11	3.59
6	3.33	2	2.43	15	2.22
5	2.16	13	2.02	10	1.70
1	1.57	14	1.34	3	0.54

AVERAGE: 3.51

Fig. 2. Characterization of inhomogeneities for mean (E) and standard deviation (D).

7. Incorrect sentences about MASH from a book

During the last 11th Seminar for Homogenization, I was obliged to make some comments to the following book of Elsevier:
P. Domonkos, R. Tóth and *L. Nyitrai*, 2022: "Climate observations: data quality control and time series homogenization" (*Domonkos et al.*, 2022).
My comments published in the seminar proceedings (*Szentimrey*, 2023c) were as follows.
"The last sentence is on page 200 is: "MASHv3 is better than MASHv4." Sorry, but it is an absolutely misleading untrue statement! Publication of the book was in 2022 while publication of MASHv4 was later in 2023. The authors could not know MASHv4!
In Section (c) "Novelties in MASHv4", on page 200 is:"… the proposed algorithm easily detects false breaks of the standard deviation around the breakpoints for the means. It is because the empirical standard deviation is elevated for periods including shifts in the means." It is also an incorrect statement! We do not use the empirical standard deviation at all!
There is a funny personal note about me as the creator of MASH on page 198:
"The creator often chose unique mathematical solutions differing both from the traditional tools of climate data homogenization and from those suggested by other statisticians." Yes, because I am a mathematician!
Conclusion: The credibility of the content of this book is doubtful for me!"
After the seminar, in the same proceedings, Peter Domonkos responded to my objections and he accepted them (*Domonkos*, 2023).

References

Caussinus, H. and *Mestre, O.*, 2004: Detection and correction of artificial shifts in climate series. *Appl. Statist. 53*, Part 3, 405–425.

Domonkos, P., 2011: Adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT). *Int. J. Geosci. 2*, 293–309. https://doi.org/10.4236/ijg.2011.23032

Domonkos, P., Tóth, R. and *Nyitrai, L.* 2022: Climate observations: Data quality control and time series homogenization. Elsevier.
https://www.elsevier.com/books/climate-observations/domonkos/978-0-323-90487-2

Domonkos, P., 2023: Response to Tamás Szentimrey regarding the presentation of MASHV4 in "Climate observations" by Domonkos, P., Tóth, R., and Nyitrai, L., Proceedings of the 11th Seminar for Homogenization and Quality Control in Climatological Databases and 6th Conference on Spatial Interpolation Techniques in Climatology and Meteorology (Ed. *Lakatos M, Puskás M, Szentimrey T*), Budapest, Hungary, 2023, WCDMP-No. 87, 14.
https://library.wmo.int/idurl/4/68452

Izsák, B., Szentimrey, T., Lakatos, M., Pongrácz, R., and *Szentes, O.*, 2022: Creation of a representative climatological database for Hungary from 1870 to 2020. *Időjárás 126*, 1–26.
https://doi.org.10.28974/idojaras.2022.1.1

Lakatos, M., Szentimrey, T., Bihari, Z., and *Szalai, S.*, 2013: Creation of a homogenized climate database for the Carpathian region by applying the MASH procedure and the preliminary analysis of the data. *Időjárás 117*. 143–158.

Mestre, O. et al., 2011: SPLIDHOM: A method for homogenization of daily temperature observations. *J. Appl. Meteorol. Climatol.* 50, 2343–2358. https://doi.org/10.1175/2011JAMC2641.1

Mestre, O. et al., 2013: HOMER: A homogenization software – Methods and applications. *Időjárás 117*, 47–67.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH), Proceedings of the Second Seminar for Homogenization of Surface Climatological Data, Budapest, Hungary; WMO, WCDMP-No. 41, 27–46.

Szentimrey, T., 2004: Multiple Analysis of Series for Homogenization (MASH); Verification procedure for homogenized time series, Proceedings of the Fourth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary; WMO, WCDMP-No. 56, 193–201.

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, 2006, WCDMP-No. 71, WMO/TD-NO. 1493, 123–130.

Szentimrey, T. et al., 2012a: Final report on quality control and data homogenization measures applied per country, including QC protocols and measures to determine the achieved increase in data quality. Carpatclim Project, Deliverable D1.12.
http://www.carpatclim-eu.org/docs/deliverables/D1_12.pdf

Szentimrey T. et al., 2012b: Final report on the creation of national gridded datasets, per country. Carpatclim Project, Deliverable D2.9. http://www.carpatclim-eu.org/docs/deliverables/D2_9.pdf

Szentimrey, T., 2013: Theoretical questions of daily data homogenization, *Időjárás 117*. 113–122.

Szentimrey, T. and *Bihari, Z.*, 2014: Manual of interpolation software MISHv1.03, Hungarian Meteorological Service.

Szentimrey, T., 2017: Some theoretical questions and development of MASH for homogenization of standard deviation, Proceedings of the 9th Seminar for Homogenization and Quality Control in Climatological Databases and 4th Conference on Spatial Interpolation Techniques in Climatology and Meteorology (Eds. *Szentimrey T, Lakatos M, Hoffmann L*), Budapest, Hungary, 2017, WCDMP-No. 85, 63–73.

Szentimrey, T., 2021: Mathematical questions of homogenization and summary of MASH, Proceedings of the 10th Seminar for Homogenization and Quality Control in Climatological Databases and 5th Conference on Spatial Interpolation Techniques in Climatology and Meteorology (Eds. *Lakatos M, Hoffmann L, Kircsi A, Szentimrey T*), Budapest, Hungary, 2020, WCDMP-No. 86, pp. 4-17

Szentimrey, T., 2023a: Overview of mathematical background of homogenization, summary of method MASH and comments on benchmark validation. *Int. J. Climatol.43*, 6314–6329.
https://doi.org/10.1002/joc.8207

Szentimrey, T., 2023b: Manual of homogenization software MASHv4.01, Varimax Limited Partnership

Szentimrey, T., 2023c: Development of new version MASHV4.01 for homogenization of standard deviation (extended abstract), Proceedings of the 11th Seminar for Homogenization and Quality Control in Climatological Databases and 6th Conference on Spatial Interpolation Techniques in Climatology and Meteorology (Eds. Lakatos M, Puskás M, Szentimrey T), Budapest, Hungary, 2023, WCDMP-No. 87. 8–13. https://library.wmo.int/idurl/4/68452

Venema et al., 2012: Benchmarking monthly homogenization algorithms. *Climate Past 8*, 89–115.

Wang, X.L. and *Y. Feng*, 2013: RHtestsV4 User Manual, Climate Data and Analysis Section – Environment and Climate Change Canada. Published online July 2013. https://github.com/ECCC-CDAS.

WMO, 2020: World Meteorological Organization Guidelines on Homogenization, WMO-No. 1245.