# New version MASHv4.01 for joint homogenization of mean and standard deviation

**Tamás Szentimrey**

**Varimax Limited Partnership**

## Introduction

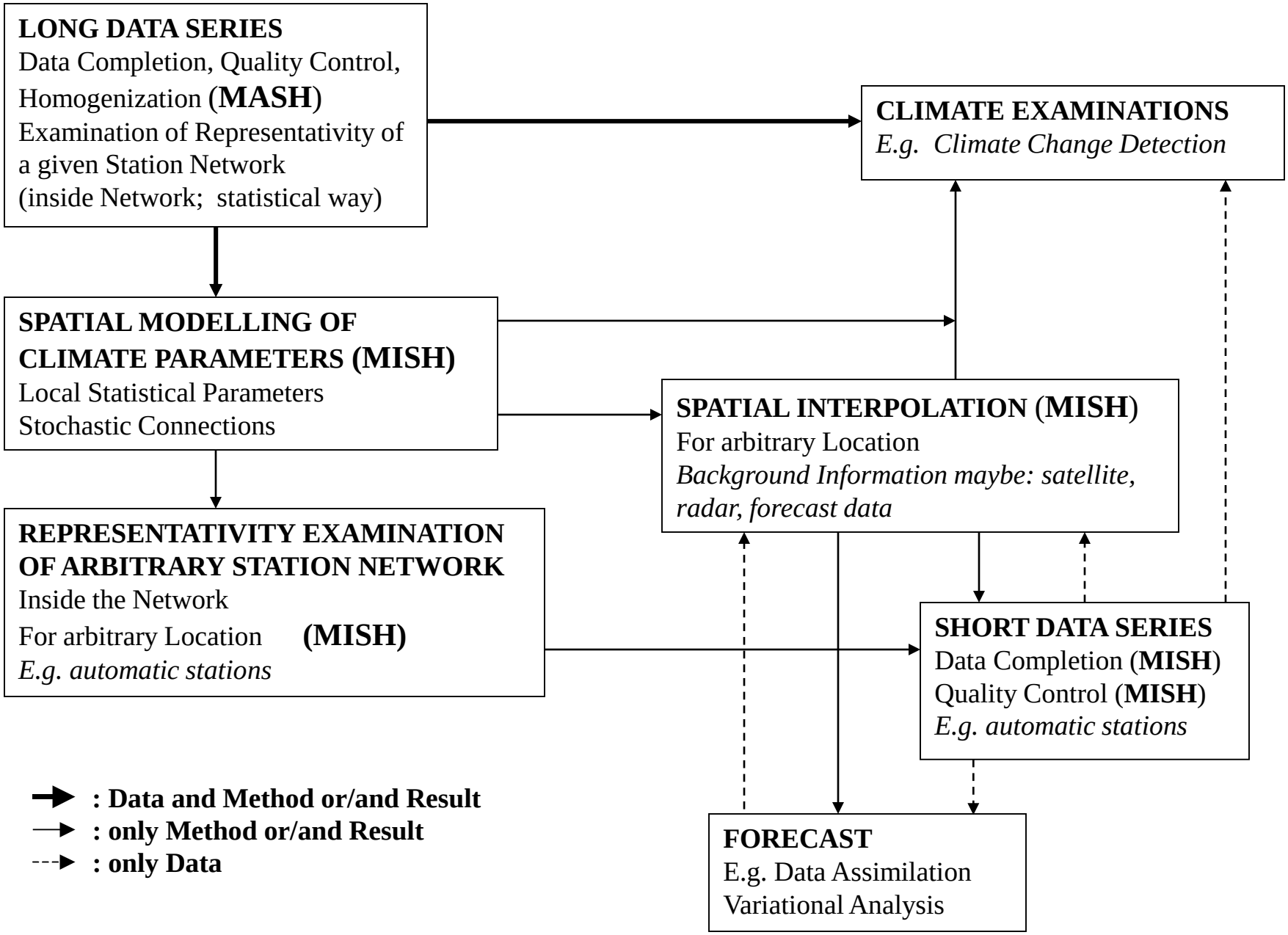
I retired from the Hungarian Meteorological Service.

I plan to continue my activity in VARIMAX Limited Partnership.

Software: the new versions MASHv4.01 and MISHv2.01
are planned to put on the website of VARIMAX this year.

Presentation: summary of my conception on homogenization.

# Possible Connection of Topics and Systems



# MATHEMATICAL FORMULATION OF HOMOGENIZATION

Let us assume we have daily or monthly data series.

$Y_1(t)\ (t = 1,2,..,n)$: candidate series of the new observing system

$Y_2(t)\ (t = 1,2,..,n)$: candidate series of the old observing system

$1 \le T < n$ : change-point

Before $T$: series $Y_2(t)\ (t = 1,2,..,T)$ can be used

After $T$: series $Y_1(t)\ (t = T+1,..,n)$ can be used

**Theoretical cumulative distribution functions (CDF):**

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

Functions $F_{1,t}(y)$, $F_{2,t}(y)$ change in time (e.g. climate change)!

**Theoretical formulation of homogenization**

Inhomogeneity: $F_{2,t}(y) \neq F_{1,t}(y) \quad (t = 1,2,..,T)$

Homogenization of $Y_2(t) \quad (t = 1,2,..,T)$:

$Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right)$ , then $\mathrm{P}\left(Y_{1,2h}(t) < y\right) = F_{1,t}(y)$

Transfer function: $F_{1,t}^{-1}\left(F_{2,t}\left(y\right)\right)$, Quantile function: $F_{1,t}^{-1}\left(p\right)$

**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.

It is not real mathematics! (good heuristics with poor mathematics)

**Special but basic case: Normal Distribution (e.g. temperature)**

**Theorem.**

Let us assume normal distribution,

$$Y_1(t) \in N\left(E_1(t), D_1(t)\right), \quad Y_2(t) \in N\left(E_2(t), D_2(t)\right) \quad \left(t = 1,2,..,n\right)$$

$E_1(t), E_2(t)$: means $\quad D_1(t), D_2(t)$: standard deviations

Then the transfer function of homogenization:

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

**Remarks:**

i, A simple linear function and there is no “tail distribution” problem!

ii, Only the mean (***E***) and standard deviation (***D***) must be homogenized!

**What is in the Practice?**

**A popular procedure**

1. Homogenization of monthly mean series:

   Break points detection, correction of mean (*E*)

   Assumption: homogeneity of higher order moments (e.g. st. deviation (*D*))

2. Homogenization of daily series:

   Trial to homogenize also the higher order moments
   (Quantile Matching, Spline)

   Used monthly information: only the detected break points

**Contradiction**

- Inhomogeneity of higher moments, **daily: yes** versus **monthly: no** ?

  It is not adequate mathematical model for standard deviation (*D*)!

- Why are not used the monthly correction factors for daily homogenization?

**Theorem**

Daily data: $Y(t)\ (t=1,..,30)$ , monthly mean: $\bar{Y}=\frac{1}{30}\sum_{t=1}^{30}Y(t)$

Monthly variable for examination of standard deviation ($D$): $S=\sqrt{\frac{1}{29}\sum_{t=2}^{30}\left(Y(t)-Y(t-1)\right)^2}$

Daily data with inhomogeneity in mean ($E$) and standard deviation ($D$):

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

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

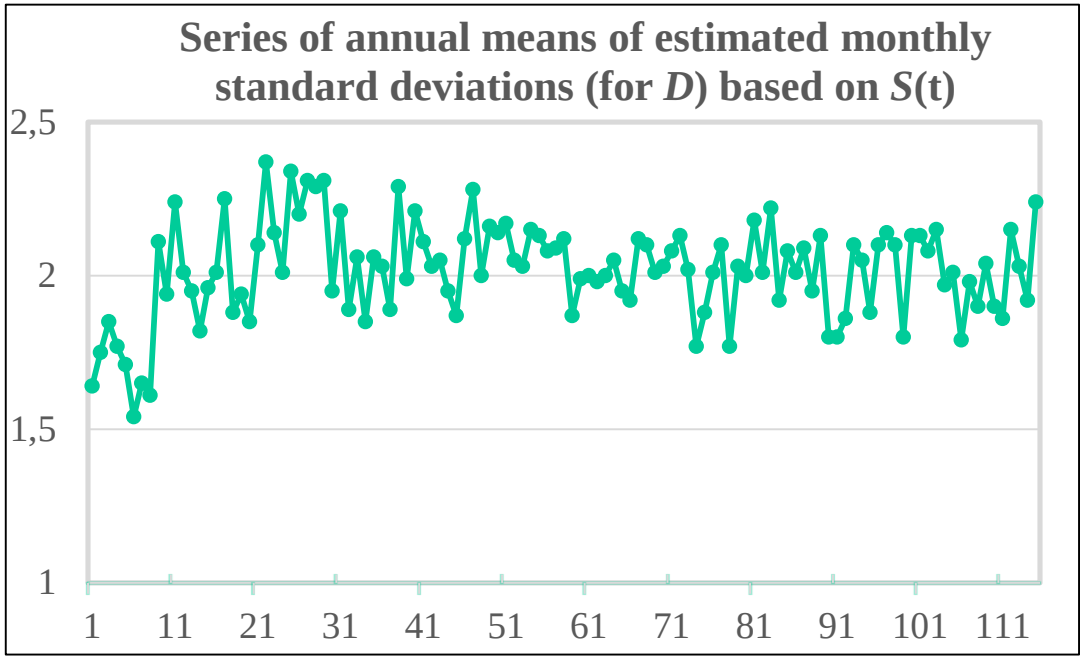
The appropriate monthly variables: $\bar{Y}_{ih}=\frac{1}{30}\sum_{t=1}^{30}Y_{ih}(t)$, $S_{ih}=\sqrt{\frac{1}{29}\sum_{t=2}^{30}\left(Y_{ih}(t)-Y_{ih}(t-1)\right)^2}$

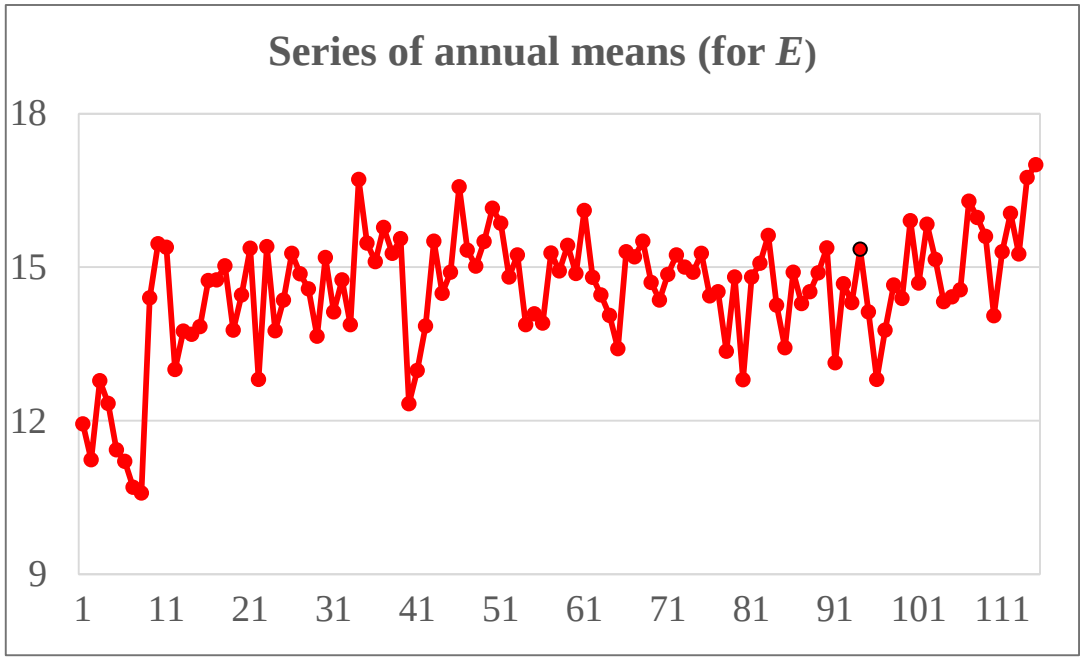
i, Then the monthly mean is also inhomogeneous in mean ($E$) and standard deviation ($D$):

$$\mathrm{E}\left(\bar{Y}_{ih}\right)=\mathrm{E}\left(\bar{Y}\right)+\beta \quad \text{and} \quad \mathrm{D}\left(\bar{Y}_{ih}\right)=\alpha\cdot\mathrm{D}\left(\bar{Y}\right)$$

ii, Moreover variable $S_{ih}$ can be used to estimate the inhomogeneity of standard deviation ($D$):

$$\mathrm{E}\left(S_{ih}\right)=\alpha\cdot\mathrm{E}\left(S\right)$$




Maximum temperature series of Miskolc (in Hungary) 1901-2015

Inhomogeneity in 1901-1908, measured Réaumur: $\mathrm{Re} = 0.8 \cdot \mathrm{C}$

$$\mathrm{E}\left(Y_{ih}(t)\right) = 0.8 \cdot \mathrm{E}\left(Y(t)\right)$$

$$\mathrm{D}\left(Y_{ih}(t)\right) = 0.8 \cdot \mathrm{D}\left(Y(t)\right)$$

**An alternative procedure developed in MASH**

**1. Homogenization of monthly series** $S(t)$, $\bar{Y}(t)$.

Homogenization of series $S(t)$ by multiplicative model.

- Break points detection, estimation of inhomogeneity of st. deviation (*D*).

Correction of standard deviation of series $\bar{Y}(t)$.

Homogenization of corrected series $\bar{Y}(t)$ by additive model.

- Break points detection, estimation of the inhomogeneity of mean (*E*).

Assumption: homogeneity of higher order (>2) moments.

This assumption is always right in case of normal distribution!

**2. Homogenization of daily series**

Homogenization of mean and standard deviation on the basis of the monthly results. The used monthly information are the break points and the monthly corrections of the mean (*E*) and standard deviation (*D*).

**Software MASHv4.01 (Multiple Analysis of Series for Homogenization)**
(*T. Szentimrey*)

The MASH system is based on homogenization of monthly series derived from daily series. The procedures depend on the distribution of climate elements.

**Quasi normal distribution (e.g. temperature)**

Beside the monthly mean series another type monthly series are also derived. These series are homogenized by multiplicative model for standard deviation (*D*). The monthly mean series corrected in standard deviation are homogenized by additive model for mean (*E*).

**Quasi lognormal distribution (e.g. precipitation)**

Monthly mean or sum series are homogenized by multiplicative model.

**Software MASHv4.01 (Multiple Analysis of Series for Homogenization)**
(*T. Szentimrey*)

**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.
– Additive or multiplicative model can be used depending on the distribution.
– Providing the homogeneity of the seasonal and annual series as well.
– Metadata (probable dates of break points) can be used automatically.
– The homogenization results and the metadata can be verified.

**Homogenization of daily series:**

– Based on the detected monthly inhomogeneities.
– Including Quality Control and missing data completion for daily data.

**EXAMPLE: 15 Hungarian July Mean Temperature Series 1901-2015**

**Test Statistics for St. Deviation (D) Before Homogenization**
**Critical value (significance level 0.01): 28.00**

| Series | TSB | Series | TSB | Series | TSB |
|---|---|---|---|---|---|
| 7 | 201.40 | 8 | 168.65 | 13 | 126.68 |
| 9 | 123.38 | 4 | 121.03 | 14 | 94.02 |
| 12 | 83.32 | 2 | 78.07 | 5 | 63.54 |
| 6 | 58.47 | 11 | 44.14 | 15 | 43.91 |
| 10 | 32.60 | 1 | 25.54 | 3 | 17.14 |

**AVERAGE: 85.46**

**Test Statistics for Mean (E) Before Homogenization**
**Critical value (significance level 0.05): 21.76**

| Series | TSB | Series | TSB | Series | TSB |
|---|---|---|---|---|---|
| 12 | 1674.66 | 7 | 388.59 | 8 | 237.88 |
| 3 | 230.71 | 10 | 224.70 | 5 | 211.41 |
| 6 | 188.81 | 11 | 154.68 | 14 | 125.35 |
| 4 | 82.50 | 9 | 72.61 | 15 | 57.57 |
| 1 | 53.55 | 13 | 49.91 | 2 | 32.95 |

**AVERAGE: 252.39**

**15 Hungarian July Mean Temperature Series 1901-2015**

**Test Statistics for St. Deviation (D) After Homogenization**
**Critical value (significance level 0.01): 28.00**

| Series | TSA | Series | TSA | Series | TSA |
|---|---|---|---|---|---|
| 13 | 36.93 | 14 | 32.50 | 4 | 32.29 |
| 8 | 26.56 | 12 | 25.73 | 7 | 23.87 |
| 9 | 23.71 | 5 | 23.37 | 2 | 22.18 |
| 1 | 19.85 | 3 | 19.70 | 11 | 18.49 |
| 6 | 16.55 | 10 | 16.55 | 15 | 14.82 |

**AVERAGE: 23.54**

**Test Statistics for Mean (E) After Homogenization**
**Critical value (significance level 0.05): 21.76**

| Series | TSA | Series | TSA | Series | TSA |
|---|---|---|---|---|---|
| 5 | 25.55 | 3 | 23.24 | 13 | 21.64 |
| 14 | 21.19 | 7 | 19.43 | 9 | 18.53 |
| 6 | 18.02 | 15 | 16.98 | 8 | 16.61 |
| 12 | 16.49 | 11 | 16.25 | 4 | 15.70 |
| 2 | 14.29 | 10 | 13.28 | 1 | 11.69 |

**AVERAGE: 17.93**

**15 Hungarian July Mean Temperature Series 1901-2015**

**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

$$D_{ih}(t) = D(t) \cdot IHD(t) \quad (t = 1,., n), \qquad IHD = \frac{100}{n} \sum_{t=1}^{n} \left| IHD(t) - 1 \right|$$

**Estimated Inhomogeneities for Mean (E)** $(^{o}\mathrm{C})$

| 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

$$E_{ih}(t) = E(t) + IHE(t) \quad (t = 1,., n), \qquad IHE = \frac{1}{n} \sum_{t=1}^{n} \left| IHE(t) \right|$$

**There is no royal road!**

**Thank you for your attention!**