CARPATCLIM

Climate of the Carpathian Region

Deliverable D2.10

Final version of metadata per country of all national gridded datasets created within module 2

Annex 3 – Description of MASH and MISH algorithms

Contract number: OJEU 2010/S 110-166082
Deliverable: D2.10
Author: Zita Bihari et al.
Date: 21-01-2013
Version: final

The following pages present the description of the applied data homogenization method MASH and interpolation method MISH. The manuals contain all the information regarding these processes from the mathematical background to the detailed introduction of the program systems and examples of application. Additional methods used for special parameters such as daily wind direction are described in deliverables D1.12, D2.5, D2.8 and D2.9.

Multiple Analysis of Series for Homogenization (M A S H v3.03)

Tamás Szentimrey

Meteorological Interpolation based on Surface Homogenized Data Basis (M I S H v1.02)

Tamás Szentimrey and Zita Bihari

PREFACE of Version MASHv3.03

The developments of the new version MASHv3.03 are connected with four topics.

The first is the input/output operations. At the present version the COST Action ES0601 (HOME) format also can be used and the description of this file format is enclosed. The converting programs can be found by the scheme of program system (see pages 21, 56).

The second is a modification of homogenization procedure of monthly series at SAM system (page 28). At the new version we recommend to start with a preliminary examination of the annual series. Normally MASH procedure would start with the monthly series and later examine the seasonal and annual series. The new possibility is to begin with an examination of annual series and to use the detected breaks as preliminary information (metadata) for the standard application of MASH for monthly data.

The third topic is the new developments for automation that aims to obtain automatic procedures. These automatic 'user friendly' procedures make the homogenization easier for the users. The directions for use can be found at the descriptions of "MASH in Practice" (p. 23), "SAM in Practice" (p. 28-29), which sections are strongly recommended for the users.

As regards the fourth topic, there were also some developments for daily data. Some new program procedures were elaborated for missing data completion and data quality control. The description of these procedures can be found on page 59.

PREFACE of Version MASHv3.02

The MASH procedure was developed originally for homogenization of monthly series. It is a relative method and depending on the distribution of examined meteorological element additive (e.g. temperature) or multiplicative (e.g. precipitation) model can be applied. In the earlier program system MASHv2.03 the following subjects were elaborated for monthly series: series comparison, break point (change point) and outlier detection, correction of series, missing data completion, automatic usage of meta data and last but not least a verification procedure to evaluate the homogenization results.

The next version MASHv3.01 was developed for homogenization of daily data furthermore for quality control of daily data and missing data completion. During the procedure normal distribution and additive model were assumed for daily data that are appropriate for temperature, pressure etc. elements.

The new version MASHv3.02 is extended also for homogenization of daily precipitation data. The procedure developed for daily data is in accordance with the multiplicative (or cumulative) model that is assumed for monthly precipitation sum data. Quality control of daily data and missing data completion are also performed during the procedure.

The program system for homogenization of monthly series was not changed. Only exception is the type of usable coordinates that were changed for filambda type, because perhaps it is more general.

PREFACE of Version MASHv2.03

The MASH method was developed in the Hungarian Meteorological Service (see References). It is a relative homogeneity test procedure that does not assume the reference series are homogeneous. Possible break points and shifts can be detected and adjusted through mutual comparisons of series within the same climatic area. The candidate series is chosen from the available time series and the remaining series are considered as reference series. The role of series changes step by step in the course of the procedure. Depending on the climatic elements, additive or multiplicative models are applied. The second case can be transformed into the first one by logarithmization.

Several difference series are constructed from the candidate and weighted reference series. The optimal weighting is determined by minimizing the variance of the difference series, in order to increase the efficiency of the statistical tests. Providing that the candidate series is the only common series of all the difference series, break points detected in all the difference series can be attributed to the candidate series.

A new multiple break points detection procedure has been developed which takes the problem of significance and efficiency into account. The significance and the efficiency are formulated according to the conventional statistics related to type one and type two errors, respectively. This test obtains not only estimated break points and shift values, but the corresponding confidence intervals as well. The series can be adjusted by using the point and interval estimates.

Since a MASH program system has been developed for the PC, the application of this method is relatively easy, with emphasis on GAME of MASH (see program MASHGAME.BAT), which is a playful version of MASH procedure for homogenization. This version can be developed towards the automation.

Some developments are connected with special problems of the homogenization of climatic time series.

One of them is the relation of monthly, seasonal and annual series. The problem arises from the fact, that the signal to noise ratio is probably less in case of monthly series than in case of derived seasonal or annual ones. Consequently the inhomogeneity can be detected easier at the derived series although we intend to adjust the monthly series (see the SAM system).

Another problem is connected with the usage of Meta Data in the course of homogenization procedure. The developed version of MASH system makes possible to use the meta data information - in particular the probable dates of break points - automatically.

The new version includes a new transformation procedure as well, which has been developed for the multiplicative model on purpose to solve the problem arising from the values coming near to zero.

A new part of MASH system is a verification procedure (MASHVERI.BAT) which makes possible to evaluate the actual or the final stage of the homogenization. We think the verification is an important part of the topic of homogenization since all over the world there are a lot of so called homogenized series however their reliability sometimes is doubtful. The basic conception of the verification procedure is that the confidence in the homogenized series may be increased by the joint comparative mathematical examination of the original and the homogenized series systems.

The last development is connected with certain automation of the procedures (see program: SAMTEST.BAT).

(MOTTO)

PROBLEM of HOMOGENIZATION

Basis: DATA

Tools:

MATHEMATICS	**:**	**abstract formulation**
META DATA	**:**	**historical, climatological information**
SOFTWARE	**:**	**automation**

SOLUTION = MATHEMATICS + META DATA + SOFTWARE

(i) without SOFTWARE:

MATHEMATICS + META DATA = THEORY WITHOUT BENEFIT

(ii) without META DATA:

MATHEMATICS + SOFTWARE = GAMBLING

(iii) without MATHEMATICS:

META DATA + SOFTWARE = “STONE AGE” + “BILL GATES”

BASIC PRINCIPLES of MASH Procedure

- **Relative homogeneity test procedure.**

- **Step by step procedure: the role of series (candidate or reference series) changes step by step in the course of the procedure.**

- **Additive or cumulative model can be used depending on the climate elements.**

- **Monthly, seasonal or annual time series can be homogenized.**

- **In case of having monthly series for all the 12 months, the monthly, seasonal and annual series can be homogenized together.
(SAM procedure: Seasonal Application of MASH)**

- **The daily inhomogeneities can be derived from the monthly ones.**

- **META DATA (probable dates of break points) can be used automatically.**

- **The actual or the final stage of the homogenization can be verified.**

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 man is possibly able to solve the problem, but it is time-consuming;

- the solution of the problem can not 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. (CHESS!!)

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 ‘meta data’ information.

THE MAIN CLIMATOLOGICAL AND STATISTICAL PROBLEMS

Modelling of the stochastic relationship between data series:
additive model, cumulative (multiplicative) model depending on climate elements, distribution of series elements.

Modelling of "inhomogeneity": break points, shifts, outliers etc..

Comparison of the examined series (Relative Test): methods for multiple comparison of the candidate series with more reference series; selection for "good" reference series systems, weighting of reference series, estimation of weighting factors. ***Multiple Comparison by Optimum Interpolation.***

Missing values: methods for closing gaps in the series.

Break points detection:
mathematical formalization according to the statistical conventions:
- first kind error (significance)
- second kind error (efficiency),

point estimation and interval estimation (confidence interval),
procedure for multiple break points and outliers detection.

Correction (adjusting) of candidate series:
separation of the detected break points and outliers for the candidate series, point estimation, interval estimation (confidence interval) for the shifts.

Relation of monthly series, seasonal series, annual series:
SAM (Seasonal Application of MASH).

Homogenization of daily data: methodology.

Meta Data: automatic using of station history.

Automation: interactive, automatic procedures for homogenization.

Verification: procedure to evaluate the homogenization results.

I. THE MATHEMATICAL BASIS OF 'MASH' PROCEDURE

(draft version)

1. STATISTICAL MODELLING

1.1 Additive Model (for example temperature)

Examined series

$$X_j(t) = C_j(t) + IH_j(t) + \varepsilon_j(t) \qquad (j = 1,2,\ldots,N;\ t = 1,2,\ldots.,n)$$

C : climate change; IH : inhomogeneity, ε : noise

1.2 Multiplicative Model (for example monthly or seasonal precipitation)

Examined series

$$X_j^*(t) = C_j^*(t) \cdot IH_j^*(t) \cdot \varepsilon_j^*(t) \qquad (j = 1,2,\ldots,N;\ t = 1,2,\ldots.,n)$$

C^* : climate change; IH^* : inhomogeneity, ε^* : noise

Logarithmization for Additive Model

$$X_j(t) = C_j(t) + IH_j(t) + \varepsilon_j(t) \qquad (j = 1,2,\ldots,N;\ t = 1,2,\ldots.,n)$$

where

$$X_j(t) = \ln X_j^*(t) \quad , \quad C_j(t) = \ln C_j^*(t) \ ,$$

$$IH_j(t) = \ln IH_j^*(t) \quad , \quad \varepsilon_j(t) = \ln \varepsilon_j^*(t)$$

Problem

If $X_j^*(t)$ values are near or equal to 0.

This problem can be solved by a Transformation Procedure which increases slightly the little values. Consequently the Multiplicative Model can be transformed into the Additive One.

2. MULTIPLE COMPARISON OF THE EXAMINED SERIES

Candidate series and its inhomogeneity: $X_c(t)$, $IH_c(t)$ $\quad c \in \{1,2,...,N\}$

Set of indexes of reference series: $R_c \subset \{1,2,...,N\}$ $\left(i \in R_c \quad \text{,if } C_i(t) \approx C_c(t)\right)$

Optimal Difference Series belonging to the subset $R_c^{(m)} \subseteq R_c$ $\left(m = 1,..,2^{|R_c|} - 1\right)$
$(\,|\;|\,$: numerosity)

$$Z_c^{(m)}(t) = X_c(t) - \sum_{i \in R_c^{(m)}} w_i \cdot X_i(t), \quad \text{where} \qquad \sum w_i = 1 \quad , \quad w_i \geq 0$$

and $\quad \mathrm{V}(Z_c^{(m)}) = \text{Variance } (Z_c^{(m)}) = \underset{w}{\text{minimum}}$

Result:

$$Z_c^{(m)}(t) = IH_c(t) \; - \sum_{i \in R_c^{(m)}} w_i \cdot IH_i(t) \; + \; \delta_c^{(m)}(t) = \; IH_c(t) - IH_{R_c^{(m)}}(t) \; + \; \delta_c^{(m)}(t)$$

Example:

If $\mathrm{V}(Z_c^{(m)}) = \text{Variance}(\,\delta_c^{(m)}) = 0$ and $IH_{R_c^{(m)}}(t) \equiv 0$ then $\quad Z_c^{(m)}(t) \equiv IH_c(t)$

Optimal Difference Series System: $Z_c^{(m)}(t)$, $m \in \mathrm{M}^* \subset \left\{\; 1,....,2^{|R_c|} - 1 \right\}$, $\left|\mathrm{M}^*\right| \geq 2$

(i) $Z_c^{(m)}(t)$: Optimal Difference Series belonging to subset $R_c^{(m)}$ (for efficiency)

(ii) $\bigcap_{m \in \mathrm{M}^*} R_c^{(m)} = \varnothing$ (for identification of inhomogeneity of candidate series)

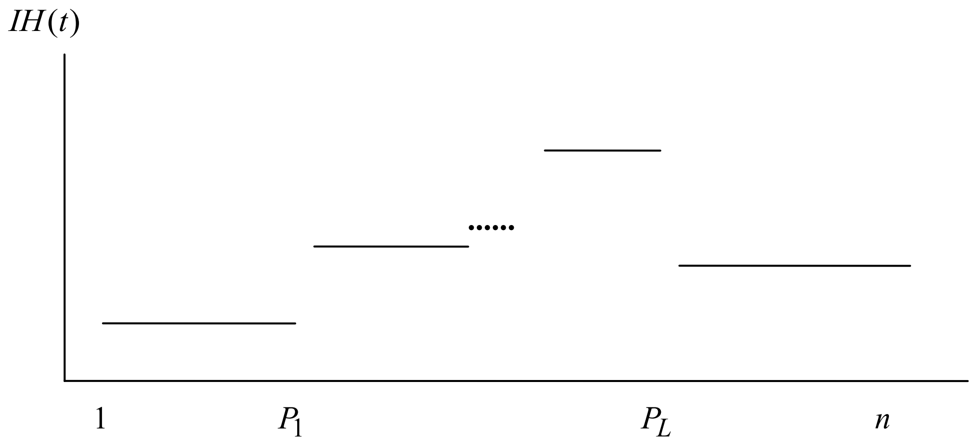
(iii) $\underset{m \in \mathrm{M}^*}{\text{maximum}}$ (Variance($Z_c^{(m)}$)) = $\underset{\mathrm{M}^*}{\text{minimum}}$ (for efficiency)

(iv) If (i), (ii), (iii) are fulfilled then let $\left|\mathrm{M}^*\right|$ be minimal too! (for efficiency)

3. EXAMINATION OF DIFFERENCE SERIES

3.1 Break Points Detection

Difference series: $Z(t) = IH(t) + \delta(t)$ $\quad (t = 1,2,\ldots.,n)$



<u>The real break points (to the left)</u> : $\{1 \le P_1 < P_2 < ..... < P_L < n\}$

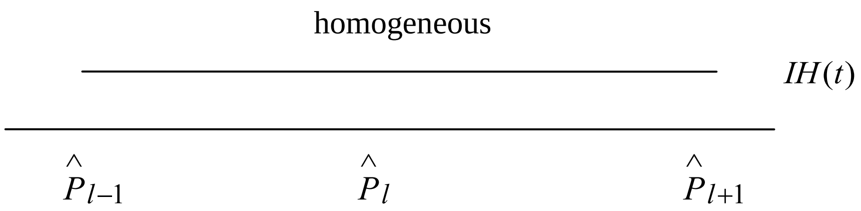
<u>BASIC POSTULATES FOR THE DECISION METHODS (FORMALIZATION)</u>

The detected break points: $\left\{ 1 \le \hat{P}_1 < \hat{P}_2 < ..... < \hat{P}_{\hat{L}} < n \right\}$

(i) <u>Type one error (significance)</u>

There exists such a $\hat{P}_l$:

interval $(\hat{P}_{l\text{-}1}, \hat{P}_{l+1}) \cap$ set$\{P_1 < P_2 < ..... < P_L\} = \varnothing$



We have to intend to give the probability of type one error, i.e. the significance level!

(ii) <u>Type two error (efficiency)</u>

There exists such a real break point that we could not detect. As much as possible!

3.2 Significant Procedure for Break Points Detection

Inhomogeneity measure for all the intervals

Statistics: $\mathrm{INH}([k, l]) \geq 0 \qquad \forall\ k,l : 1 \leq k < l \leq n$

and $\mathrm{INH}([i, j]) \leq \mathrm{INH}([k, l])$, if $[i, j] \subseteq [k, l]$

Test Statistic of difference series

The inhomogeneity of difference series $Z(t)$ can be characterized by the

Test Statistic: $\mathrm{TS} = \mathrm{INH}([1, n])$

The critical value (α) (by Monte Carlo Method)

$$\mathrm{P}(\,\mathrm{TS} > \alpha \ \mid\ \text{if } Z(t) \text{ homogeneous}\,) = \text{sig. level } (= 0.1, 0.05, 0.01)$$

Test Statistic can be compared to the critical value and in case of homogeneity it should be less, on the given significance level.

PROPERTIES OF THE DETECTING PROCEDURE
(FOR THE PURPOSE OF SIGNIFICANCE AND EFFICIENCY)

If the detected break points: $\left\{ 1 \leq \hat{P}_1 < \hat{P}_2 < ..... < \hat{P}_{\hat{L}} < n \right\}$, then

$$\underset{l=1,\ldots,\hat{L}+1}{\text{maximum}} \left(\mathrm{INH}\left((\hat{P}_{l\text{-}1}, \hat{P}_l] \right) \right) \leq \alpha < \underset{l=1,\ldots,\hat{L}}{\text{minimum}} \left(\mathrm{INH}\left((\hat{P}_{l\text{-}1}, \hat{P}_{l+1}] \right) \right)$$

i.e. on the given significance level:

- the intervals $(\hat{P}_{l\text{-}1}, \hat{P}_{l+1}]$ are not homogeneous, consequently the detected break points $\hat{P}_l$ are not superfluous,
- the intervals $(\hat{P}_{l\text{-}1}, \hat{P}_l]$ can be accepted to be homogeneous.

Confidence Intervals

Confidence intervals also can be given for the break points on the confidence level (1-sig. level): $\mathrm{I}_l \qquad l = 1,....,\hat{L}$

3.3 Estimation of Shifts

Point estimation; Confidence intervals for the shifts

4. EVALUATION OF HOMOGENEITY OF CANDIDATE SERIES $X_c(t)$

Based on the Test Statistics (TS) belonging to the Optimal Difference Series:

$$Z_c^{(m)}(t) \qquad (m = 1,..,2^{|R_c|} - 1)$$

5. CORRECTION OF CANDIDATE SERIES $X_c(t)$

Based on the examination of the Optimal Difference Series System:

$$Z_c^{(m)}(t)\,, \quad m \in \mathrm{M}^* \subset \left\{ 1,....,2^{|R_c|} - 1 \right\} \;, \quad \left|\mathrm{M}^*\right| \geq 2$$

<u>BASIC PRINCIPLE OF BREAK POINT DETECTION FOR CANDIDATE SERIES</u>

Let us assume, that

$\hat{P}^{(m)}$ $(m \in \mathrm{M}^*)$: detected Break Points,

$\mathrm{I}^{(m)}$ $(m \in \mathrm{M}^*)$: Confidence Intervals

belonging to the Optimal Difference Series $Z_c^{(m)}(t)$ $(m \in \mathrm{M}^*)$, AND

$$\bigcap_{m \in \mathrm{M}^*} \mathrm{I}^{(m)} \neq \varnothing \qquad \text{as well as} \qquad \forall \; \hat{P}^{(m)} \in \bigcap_{m \in \mathrm{M}^*} \mathrm{I}^{(m)}$$

<u>DECISION</u>

The „most probable” $\hat{P}^{(m)}$ is a Break Point of the Candidate Series $X_c(t)$.

6. USING OF META DATA (Meta Data: probable dates of break points)

BASIC PRINCIPLE OF BREAK POINT DETECTION BY USING OF META DATA

Candidate series and its Meta Data:

$$X_c(t) \quad , \quad \Delta_c = \left\{ 1 \le D_1^{(c)} < D_2^{(c)} < .... < D_{K_c}^{(c)} < n \right\}$$

Optimal Difference Series System: $Z_c^{(m)}(t)$, $m \in \mathrm{M}^*$, $\left|\mathrm{M}^*\right| \ge 2$

Let us assume, that

$\hat{P}{}^{(m)}$ $(m \in \mathrm{M}^*)$: detected Break Points,

$\mathrm{I}^{(m)}$ $(m \in \mathrm{M}^*)$: Confidence Intervals

belonging to the Optimal Difference Series $Z_c^{(m)}(t)$ $(m \in \mathrm{M}^*)$, AND

$$\bigcap_{m \in \mathrm{M}^*} \mathrm{I}^{(m)} \ne \varnothing \qquad \text{as well as} \qquad \forall \ \hat{P}{}^{(m)} \ \in \bigcap_{m \in \mathrm{M}^*} \mathrm{I}^{(m)}$$

BASIC DECISION RULE

(i) If $\mathrm{Q} := \left(\bigcap_{m \in \mathrm{M}^*} \mathrm{I}^{(m)} \right) \bigcap \Delta_c \ne \varnothing$

The „most probable" $D^{(c)} \in \mathrm{Q}$ is a Break Point of the Candidate Series $X_c(t)$.
(Break Point: Meta Data)

(ii) If $\left(\bigcap_{m \in \mathrm{M}^*} \mathrm{I}^{(m)} \right) \bigcap \Delta_c = \varnothing$ but $\left(\bigcup_{m \in \mathrm{M}^*} \mathrm{I}^{(m)} \right) \bigcap \Delta_c \ne \varnothing$

No Decision.

(iii) If $\left(\bigcup_{m \in \mathrm{M}^*} \mathrm{I}^{(m)} \right) \bigcap \Delta_c = \varnothing$

The „most probable" $\hat{P}{}^{(m)}$ is a Break Point of the Candidate Series $X_c(t)$.
(Break Point: is not Meta Data, but „undoubtful")

7. EVALUATION OF META DATA

(Meta Data: probable dates of break points)

THE QUALITY OF META DATA CAN BE VERIFIED BY STATISTICAL TESTS!!!

For example: the problem of Missing Meta Data??

In Practice: the statistical Test Results are often verified with the Meta Data.

BUT: the question may be turned round!

Examined series and their Meta Data

$$X_j(t), \quad \Delta_j = \left\{ 1 \le D_1^{(j)} < D_2^{(j)} < .... < D_{K_j}^{(j)} < n \right\} \qquad (j = 1,2,\ldots,N)$$

Candidate series and its Meta Data: $X_c(t)$, Δ_c $c \in \{1,2,...,N\}$

Optimal Difference Series belonging to the subset $R_c^{(m)} \subseteq R_c$:

$$Z_c^{(m)}(t) = X_c(t) - \sum_{i \in R_c^{(m)}} w_i \cdot X_i(t) = \sum_{i \in R_c^{(m)}} w_i \cdot \left(X_c(t) - X_i(t) \right) = \sum_{i \in R_c^{(m)}} w_i \cdot Z_{ci}(t)$$

Transformation of Difference Series $Z_{ci}(t)$

$$\Delta_{c \vee i} = \Delta_c \bigcup \Delta_i = \left\{ 1 \le D_1^{(c \vee i)} < D_1^{(c \vee i)} < \ldots < D_{K_{c \vee i}}^{(c \vee i)} < n \right\}$$

$$\widetilde{Z}_{ci}(t) = \begin{cases} Z_{ci}(t) - \overline{Z}_{ci}[1, D_1^{(c \vee i)}] & ,\text{if} \quad 1 \le t \le D_1^{(c \vee i)} \\ Z_{ci}(t) - \overline{Z}_{ci}(D_{k-1}^{(c \vee i)}, D_k^{(c \vee i)}] & ,\text{if} \quad D_{k-1}^{(c \vee i)} < t \le D_k^{(c \vee i)} \quad (k = 2,\ldots,K_{c \vee i}) \\ Z_{ci}(t) - \overline{Z}_{ci}(D_{K_{c \vee i}}^{(c \vee i)}, n] & ,\text{if} \quad D_{K_{c \vee i}}^{(c \vee i)} < t \le n \end{cases}$$

$\overline{Z}_{ci}\langle a,b \rangle$: average of $Z_{ci}(t)$ above the interval $\langle a,b \rangle$.

Transformed Optimal Difference Series belonging to the subset $R_c^{(m)} \subseteq R_c$:

$$\widetilde{Z}_c^{(m)}(t) = \sum_{i \in R_c^{(m)}} w_i \cdot \widetilde{Z}_{ci}(t) \qquad (m = 1,..,2^{|R_c|} - 1)$$

are homogeneous if the inhomogeneities can be explained by the Meta Data!

EVALUATION OF META DATA: Based on the Test Statistics (TS) belonging to the Transformed Optimal Difference Series $\widetilde{Z}_c^{(m)}(t)$.

8. SEASONAL APPLICATION OF MASH (SAM)

Monthly difference series: $Z^{(k)}(t)$ $\left(k = 1,2,\ldots.,K \right)$

Expectations and Variances: $\mathrm{E}(Z^{(k)}(t)) = IH^{(k)}(t)$, $\mathrm{V}(Z^{(k)})$

Seasonal mean difference series: $\bar{Z}(t) = \frac{1}{K} \sum_{k=1}^{K} Z^{(k)}(t)$

Expectation and Variance: $\mathrm{E}(\bar{Z}(t)) = \overline{IH}(t) = \frac{1}{K} \sum_{k=1}^{K} IH^{(k)}(t)$, $\mathrm{V}(\bar{Z})$

The test results after the Homogenization of monthly series

H_0: $IH^{(k)}(t) \equiv 0$ $\left(k = 1,2,\ldots.,K \right)$ can be accepted.

BUT! (sometimes) H_0: $\overline{IH}(t) \equiv 0$ can not be accepted!

The reason of the problem

The efficiency of test depends on the signal to noise ratio, and according to the test results

$$\mathrm{R}(\bar{Z}(t)) = \frac{\left|\overline{IH}(t)\right|}{\sqrt{\mathrm{V}(\bar{Z})}} > \mathrm{R}(Z^{(k)}(t)) = \frac{\left|IH^{(k)}(t)\right|}{\sqrt{\mathrm{V}(Z^{(k)})}} \approx 0 \qquad \left(k = 1,2,\ldots.,K \right),$$

as a consequence of the general inequality: $\mathrm{V}(\bar{Z}) <$ $\mathrm{V}(Z^{(k)})$ $\left(k = 1,2,\ldots.,K \right)$

Deviance series and ratios

$$Z^{(k)}(t) - \bar{Z}(t) \quad , \quad \mathrm{R}\left(Z^{(k)}(t) - \bar{Z}(t)\right) = \frac{\left|IH^{(k)}(t) - \overline{IH}(t)\right|}{\sqrt{\mathrm{V}(Z^{(k)} - \bar{Z})}} \qquad \left(k = 1,2,\ldots.,K \right)$$

Lemma 1

If $\mathrm{R}(\bar{Z}(t)) > \mathrm{R}(Z^{(k)}(t)) \quad \left(k = 1,2,\ldots.,K\right)$, then

$$\bar{\mathrm{R}}\left(Z(t)-\bar{Z}(t)\right) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{R}\left(Z^{(k)}(t)-\bar{Z}(t)\right) \leq \max_{k}\left\{\mathrm{R}(Z^{(k)}(t)\right\}\cdot\sqrt{\frac{\bar{\mathrm{V}}(Z-\bar{Z})}{\bar{\mathrm{V}}_{\mathrm{H}}(Z-\bar{Z})}}$$

where

$\bar{\mathrm{V}}(Z-\bar{Z})$: arithmetic mean of the variances $\mathrm{V}(Z^{(k)}-\bar{Z}) \quad \left(k = 1,2,\ldots.,K\right)$,

$\bar{\mathrm{V}}_{\mathrm{H}}(Z-\bar{Z})$: harmonic mean of the variances $\mathrm{V}(Z^{(k)}-\bar{Z}) \quad \left(k = 1,2,\ldots.,K\right)$

Consequently if $\mathrm{R}(\bar{Z}(t)) > \mathrm{R}(Z^{(k)}(t)) \approx 0 \quad \left(k = 1,2,\ldots.,K\right)$, then
the ratios $\mathrm{R}\left(Z^{(k)}(t)-\bar{Z}(t)\right) \quad \left(k = 1,2,\ldots.,K\right)$ are probably near to 0.

Test of Hypothesis

$$\mathrm{H}_0 : \ \mathrm{R}\left(Z^{(k)}(t)-\bar{Z}(t)\right) \equiv 0 \quad (\ \Leftrightarrow \ IH^{(k)}(t) \equiv \overline{IH}(t) \ ) \quad \left(k = 1,2,\ldots.,K\right)$$

The test of hypothesis is based on the examination of the deviance series
$Z^{(k)}(t)-\bar{Z}(t) \quad \left(k = 1,2,\ldots.,K\right)$.

If H_0 can be accepted, then

$$\bar{\mathrm{R}}\left(Z(t)-\overline{IH}(t)\right) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{R}\left(Z^{(k)}(t)-\overline{IH}(t)\right) \approx 0$$

as a consequence of the following lemma.

Lemma 2

$$\overline{\mathrm{R}}\left(Z(t)-\overline{IH}(t)\right)\leq \max_{k}\left\{\mathrm{R}\left(Z^{(k)}(t)-\overline{Z}(t)\right)\right\}\cdot\sqrt{\frac{\overline{\mathrm{V}}(Z)}{\overline{\mathrm{V}}_{\mathrm{H}}(Z)}}$$

where

$\overline{\mathrm{V}}(Z)$: arithmetic mean of the variances $\mathrm{V}(Z^{(k)})$ $\left(k=1,2,\ldots.,K\right)$,

$\overline{\mathrm{V}}_{\mathrm{H}}(Z)$: harmonic mean of the variances $\mathrm{V}(Z^{(k)})$ $\left(k=1,2,\ldots.,K\right)$.

Consequently the ratios

$\mathrm{R}\left(Z^{(k)}(t)-\overline{IH}(t)\right)$ $\left(k=1,2,\ldots.,K\right)$

are probably near to 0, i.e. the monthly inhomogeneities $IH^{(k)}(t)$ $\left(k=1,2,..,K\right)$ can be estimated with the estimation of the seasonal inhomogeneity $\overline{IH}(t)$.

9. VERIFICATION OF HOMOGENIZATION

9.1 Additive Model (for example temperature)

Original Series

$X_{O,j}(t) = C_j(t)+IH_j(t)+\varepsilon_j(t)$ $\left(j=1,2,\ldots,N;\ t=1,2,\ldots.,n\right)$

C : climate change; IH : inhomogeneity, ε : noise

Estimated Inhomogeneity Series $I\hat{H}_j(t)$

Homogenized Series $X_{H,j}(t)=X_{O,j}(t)-I\hat{H}_j(t)$

Residual Inhomogeneity Series $IH_{res,j}(t)=IH_j(t)-I\hat{H}_j(t)$

Optimal Interpolation of Series

Interpolation of Original Series: $\hat{X}_{O,j}(t) = w_0 + \sum_{i \in R_j} w_i \cdot X_{O,i}(t)$,

where R_j is the reference index set, $\sum_{i \in R_j} w_i = 1$ and

$ERR = \mathrm{E}\left(\left(X_{O,j}(t) - \hat{X}_{O,j}(t)\right)^2\right) = \underset{w_0, w_i}{\text{minimum}}$.

Interpolation of Homogenized Series: $\hat{X}_{H,j}(t) = w_0 + \sum_{i \in R_j} w_i \cdot X_{H,i}(t)$,

where $\sum_{i \in R_j} w_i = 1$ and $ERR = \mathrm{E}\left(\left(X_{H,j}(t) - \hat{X}_{H,j}(t)\right)^2\right) = \underset{w_0, w_i}{\text{minimum}}$.

Regression of $I\hat{H}_j(t)$ by Meta Data (probable dates of break points)

Meta Data: $\Delta_j = \left\{1 \le D_1^{(j)} < D_2^{(j)} < .... < D_{K_j}^{(j)} < n\right\}$

$$I\hat{H}_{Mreg,j}(t) = \begin{cases} \left(\overline{I\hat{H}_j}\right)_A [1, D_1^{(j)}] & ,\text{if} \quad 1 \le t \le D_1^{(j)} \\ \left(\overline{I\hat{H}_j}\right)_A (D_{k-1}^{(j)}, D_k^{(j)}] & ,\text{if} \quad D_{k-1}^{(j)} < t \le D_k^{(j)} \quad (k = 2, ..., K_j) \\ \left(\overline{I\hat{H}_j}\right)_A (D_{K_j}^{(j)}, n] & ,\text{if} \quad D_{K_j}^{(j)} < t \le n \end{cases}$$

$\left(\overline{I\hat{H}_j}\right)_A \langle a, b \rangle$: arithmetic mean of $I\hat{H}_j(t)$ above the interval $\langle a, b \rangle$.

9.2 Multiplicative Model (for example monthly or seasonal precipitation)

Original Series

$$X^*_{O,j}(t) = C^*_j(t) \cdot IH^*_j(t) \cdot \varepsilon^*_j(t) \qquad \left(j = 1,2,\ldots,N;\ t = 1,2,\ldots.,n \right)$$

C^* : climate change; IH^* : inhomogeneity, ε^* : noise

Logarithmization for Additive Model

$$X_{O,j}(t) = C_j(t) + IH_j(t) + \varepsilon_j(t) \qquad \left(j = 1,2,\ldots,N;\ t = 1,2,\ldots.,n \right)$$

where

$$X_{O,j}(t) = \ln X^*_{O,j}(t) \ ,\ C_i(t) = \ln C^*_j(t) \ ,\ IH_j(t) = \ln IH^*_j(t) \ ,\ \varepsilon_j(t) = \ln \varepsilon^*_j(t)$$

Problem

If $X^*_{O,j}(t)$ values are near or equal to 0. This problem can be solved by a Transformation Procedure which increases slightly the little values. Consequently the Multiplicative Model can be transformed into the Additive One.

Estimated Inhomogeneity Series

$$I\hat{H}^*_j(t) \ (> 0) \qquad , \qquad I\hat{H}_j(t) = \ln I\hat{H}^*_j(t)$$

Homogenized Series

$$X^*_{H,j}(t) = \frac{X^*_{O,j}(t)}{I\hat{H}^*_j(t)} \qquad , \qquad X_{H,j}(t) = \ln X^*_{H,j}(t) = X_{O,j}(t) - I\hat{H}_j(t)$$

Residual Inhomogeneity Series

$$IH^*_{res,j}(t) = \frac{IH^*_j(t)}{I\hat{H}^*_j(t)} \qquad , \qquad IH_{res,j}(t) = \ln IH^*_{res,j}(t) = IH_j(t) - I\hat{H}_j(t)$$

‘Optimal’ Interpolation (multiplicative)

Interpolation of Original Series: $\hat{X}^*_{O,j}(t) = \exp\left(\hat{X}_{O,j}(t)\right) = e^{w_0} \cdot \prod_{i \in R_j} \left(X^*_{O,i}(t) \right)^{w_i}$

where $\hat{X}_{O,j}(t)$ is the optimally interpolated series of $X_{O,j}(t)$.

Interpolation of Homogenized Series: $\hat{X}^*_{H,j}(t) = \exp\left(\hat{X}_{H,j}(t)\right) = e^{w_0} \cdot \prod_{i \in R_j} \left(X^*_{H,i}(t) \right)^{w_i}$ where

$\hat{X}_{H,j}(t)$ is the optimally interpolated series of $X_{H,j}(t)$.

Regression of $I\hat{H}^*_j(t)$ by Meta Data (probable dates of break points)

Meta Data: $\Delta_j = \left\{ 1 \le D_1^{(j)} < D_2^{(j)} < .... < D_{K_j}^{(j)} < n \right\}$

$$I\hat{H}^{*}_{Mreg,j}(t) = \exp\left(I\hat{H}_{Mreg,j}(t)\right) =$$

$$= \begin{cases} \left(\overline{I\hat{H}_j^{*}}\right)_G [1, D_1^{(j)}] & ,\text{if} \quad 1 \le t \le D_1^{(j)} \\ \left(\overline{I\hat{H}_j^{*}}\right)_G (D_{k-1}^{(j)}, D_k^{(j)}] & ,\text{if} \quad D_{k-1}^{(j)} < t \le D_k^{(j)} \ \ (k = 2,., K_j) \\ \left(\overline{I\hat{H}_j^{*}}\right)_G (D_{K_j}^{(j)}, n] & ,\text{if} \quad D_{K_j}^{(j)} < t \le n \end{cases}$$

$\left(\overline{I\hat{H}_j}\right)_G \langle a,b \rangle$: geometric mean of $I\hat{H}_j(t)$ above the interval $\langle a,b \rangle$.

9.3 Series for Verification Procedure

	‘Additive’ Series	‘Multiplicative’ Series
Original Series:	$X_O(t)$	$X_O^{*}(t) = \exp\left(X_O(t)\right)$
Estimated Inhomogeneity :	$I\hat{H}(t)$	$I\hat{H}^{*}(t) = \exp\left(I\hat{H}(t)\right)$
Homogenized Series:	$X_H(t)$	$X_H^{*}(t) = \exp\left(X_H(t)\right)$
Residual Inhom. (unknown):	$IH_{res}(t)$	$IH_{res}^{*}(t) = \exp\left(IH_{res}(t)\right)$
Opt. Int. of Orig. Series:	$\hat{X}_O(t)$	$\hat{X}_O^{*}(t) = \exp\left(\hat{X}_O(t)\right)$
Opt. Int. of Hom. Series:	$\hat{X}_H(t)$	$\hat{X}_H^{*}(t) = \exp\left(\hat{X}_H(t)\right)$
Regr. of Est. Inh. by Meta:	$I\hat{H}_{Mreg,j}(t)$	$I\hat{H}^{*}_{Mreg,j}(t) = \exp\left(I\hat{H}_{Mreg,j}(t)\right)$

At the additive model we have additive series only, while in case of the multiplicative model we have additive and multiplicative series alike.

9.4 Basic Statistical Functions for Verification Procedure

Statistical Functions for ‘Additive’ Series

Deviaton of series $x(t)$, $y(t)$ $\left(t = 1,2,\ldots,n\right)$: $D(x,y) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x(t) - y(t)\right)^2}$

Standard Deviaton of series $x(t)$ $\left(t = 1,2,\ldots,n\right)$: $S(x) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x(t) - \overline{x(t)}_A\right)^2}$

Deviation Error of estimation $\hat{x}(t)$ $\left(t = 1,2,\ldots,n\right)$: $ERR(x,\hat{x}) = D(x,\hat{x})$

Statistical Functions for ‘Multiplicative’ Series

Fluctuation of series $x(t)(>0)$, $y(t)(>0)$ $\left(t = 1,2,\ldots,n\right)$: $F(x,y) = \left(\prod_{t=1}^{n} \max\left(\frac{x(t)}{y(t)}, \frac{y(t)}{x(t)}\right)\right)^{\frac{1}{n}}$

Standard Fluctuation of series $x(t)(>0)$ $(t=1,2,\ldots.,n)$: $SF(x)=\left(\prod_{t=1}^{n}\max\left(\frac{x(t)}{\bar{x}_G}\,,\,\frac{\bar{x}_G}{x(t)}\right)\right)^{\frac{1}{n}}$

(G: geometric mean)
Fluctuation Error of estimation $\hat{x}(t)(>0)$ $(t=1,2,\ldots,n)$: $FERR(x,\hat{x})=F(x,\hat{x})$

Lemma

Connection between the additive and multiplicative statistical functions:

$SF(y)\approx SF(x)^{\frac{S(\ln y)}{S(\ln x)}}$ and $F(x,y)\approx SF(x)^{\frac{D(\ln x,\ln y)}{S(\ln x)}}$

9.5 The Verification Statistics

For both model the calculation of verification statistics is based on the ‘additive’ series, but in case of multiplicative model the verification statistics can be interpreted for the ‘multiplicative’ series too according to the lemma.

I. Test Statistics for Series Inhomogeneity

I.1. Test Statistic After Homogenization (TSA)

Examined series: $Z_H(t)=X_H(t)-\hat{X}_H(t)$

I.2. Test Statistic Before Homogenization (TSB)

Examined series: $Z_O(t)=X_O(t)-\hat{X}_O(t)$

I.3. Statistic for Estimated Inhomogeneity (IS)

Examined series: $I\hat{H}(t)$
The homogenization can be considered is successful if the Test Statistic After Homogenization is little and the Statistic for Estimated Inhomogeneity is in accordance with the Test Statistic Before Homogenization.

II. Characterization of Inhomogeneity

II.1. Relative Estimated Inhomogeneity: $\mathrm{RI1}=\frac{S(I\hat{H})}{S(X_O)}$

Multiplicative interpretation: $SF(I\hat{H}^{\,*})\approx SF(X_O^{\,*})^{\mathrm{RI1}}$

II.2. Relative Modification of Series: $\mathrm{RI2}=\frac{D(X_O,X_H)}{S(X_O)}$

Multiplicative interpretation: $F(X_O^*,X_H^*)\approx SF(X_O^{\,*})^{\mathrm{RI2}}$

II.3. Lower Confidence Limit (RI3) for Relative Residual Inhomogeneity:

$$\mathrm{P}\left(\frac{S(IH_{res})}{S(X_H)} \geq \mathrm{RI3} \right) \geq 1- \text{ sig. level } \left(= 0.9, 0.95, 0.99\right)$$

Multiplicative interpretation: $\mathrm{P}\left(SF(IH^{*}_{res}) \geq SF(X^{*}_{H})^{\mathrm{RI3}} \right) \geq 1- \text{ sig. level}$

III. Representativity of Station Network

$$\mathrm{RS} = 1 - \frac{ERR(X_H, \hat{X}_H)}{S(X_H)}$$

Multiplicative interpretation: $FERR(X^{*}_{H}, \hat{X}^{*}_{H}) \approx SF(X^{*}_{H})^{1-\mathrm{RS}}$

IV. Test Statistic for Meta Data

Examined series: $Z_O(t) = X_O(t) - \hat{X}_O(t)$ with Meta Data.

V. Representativity of Meta Data

$$\mathrm{RM} = 1 - \frac{ERR(I\hat{H}, I\hat{H}_{Mreg})}{S(I\hat{H})}$$

Multiplicative interpretation: $FERR(I\hat{H}^{*}, I\hat{H}^{*}_{Mreg}) \approx SF(I\hat{H}^{*})^{1-\mathrm{RM}}$

CRITICAL VALUES FOR TEST STATISTICS (by Monte Carlo Method)

Significance level: 0.1
Length of series: critical value for the Test statistic of inhomogeneity

```
 10: 15.902 ;  20: 15.845 ;  30: 16.160 ;  40: 16.765;
 50: 17.156 ;  60: 17.697 ;  70: 18.059 ;  80: 18.369;
 90: 18.655 ; 100: 18.843 ; 110: 19.008 ; 120: 19.101;
130: 19.220 ; 140: 19.397 ; 150: 19.526 ; 160: 19.609;
170: 19.678 ; 180: 19.749 ; 190: 19.789 ; 200: 19.950
```

Significance level: 0.1
Length of series: critical value for the outliers Test statistic

```
 10:  5.495 ;  20:  5.530 ;  30:  5.898 ;  40:  6.126;
 50:  6.330 ;  60:  6.486 ;  70:  6.613 ;  80:  6.719;
 90:  6.802 ; 100:  6.914 ; 110:  7.009 ; 120:  7.089;
130:  7.145 ; 140:  7.234 ; 150:  7.294 ; 160:  7.343;
170:  7.387 ; 180:  7.434 ; 190:  7.512 ; 200:  7.558
```

Significance level: 0.05
Length of series: critical value for the Test statistic of inhomogeneity

```
 10: 23.602 ;  20: 20.924 ;  30: 20.530 ;  40: 20.574;
 50: 20.861 ;  60: 20.914 ;  70: 21.313 ;  80: 21.395;
 90: 21.534 ; 100: 21.599 ; 110: 21.731 ; 120: 21.760;
130: 21.933 ; 140: 21.936 ; 150: 22.052 ; 160: 22.063;
170: 22.078 ; 180: 22.193 ; 190: 22.288 ; 200: 22.362
```

Significance level: 0.05
Length of series: critical value for the outliers Test statistic

```
 10:  9.263 ;  20:  7.445 ;  30:  7.442 ;  40:  7.582;
 50:  7.710 ;  60:  7.797 ;  70:  7.901 ;  80:  7.996;
 90:  8.028 ; 100:  8.076 ; 110:  8.147 ; 120:  8.202;
130:  8.295 ; 140:  8.344 ; 150:  8.403 ; 160:  8.433;
170:  8.484 ; 180:  8.518 ; 190:  8.531 ; 200:  8.607
```

Significance level: 0.01
Length of series: critical value for the Test statistic of inhomogeneity (over-estimated values)

```
 10: 52.000 ;  20: 37.000 ;  30: 33.000 ;  40: 32.000;
 50: 31.000 ;  60: 30.000 ;  70: 30.000 ;  80: 29.000;
 90: 29.000 ; 100: 29.000 ; 110: 29.000 ; 120: 28.000;
130: 28.000 ; 140: 28.000 ; 150: 28.000 ; 160: 28.000;
170: 28.000 ; 180: 28.000 ; 190: 28.000 ; 200: 28.000
```

Significance level: 0.01
Length of series: critical value for the outliers Test statistic (over-estimated values)

```
 10: 32.000 ;  20: 14.000 ;  30: 12.000 ;  40: 12.000;
 50: 12.000 ;  60: 12.000 ;  70: 12.000 ;  80: 11.000;
 90: 11.000 ; 100: 11.000 ; 110: 11.000 ; 120: 11.000;
130: 11.000 ; 140: 11.000 ; 150: 11.000 ; 160: 11.000;
170: 11.000 ; 180: 11.000 ; 190: 11.000 ; 200: 11.000
```

Remark: The critical values are built in the program system.

II. THE STRUCTURE OF PROGRAM SYSTEM

Main Directory **MASHv3.03**:

- Directory **MASHDAILY (**See Page **56)**
- Directory **MASHMONTHLY**:
    - Subdirectory **COSTHOMEINPUT**
    - Subdirectory **COSTHOMEOUTPUT**
    - Subdirectory **SAM**:
        - Subdirectory **SAMPAR**
          (parametrization program)
        - **Main Program Files of SAM**
        - Subdirectory **SAMAUTO**
          (automatic homogenization programs)
        - Subdirectory **SAMMISS**
          (data completion and QC programs)
        - Subdirectory **SAMVERI**
          (verification program)
        - Subdirectory **SAMMANU**
          ("manual" programs)
        - Subdirectory **SAMEND**
          (finishing program)
        - Subdirectory **SAMSUB**
          (do not use it including "subroutines")
        - Subdirectory **MASH**:
            - Subdirectory **MASHPAR**
              (parametrization program)
            - **Main Program Files of MASH**
            - Subdirectory **MASHAUTO**
              (automatic homogenization program)
            - Subdirectory **MASHMANU**
              ("manual" programs)
            - Subdirectory **MASHEND**
              (finishing program)
            - Subdirectory **MASHSUB**
              (do not use it including "subroutines")

General Comments

Monthly, seasonal or annual time series can be homogenized by the program system.

The data series belonging to different stations are compared in the course of the procedure.
The maximal number of the stations: 500
The maximal length of the time series: 200

In case of having monthly series for all the 12 months, the monthly, seasonal and annual series can be homogenized together by the main program files of the subdirectory SAM (Seasonal Application of MASH; see page 27).

In case of having only annual series, or monthly series belonging to a given month, or seasonal series belonging to a given season, the series can be homogenized by the main program files of subdirectory MASH (see page 23).

Depending on the climatic elements, additive (e.g. temperature) or multiplicative (e.g. precipitation) models are applied. The second case can be transformed into the first one by logarithmization. The problem of values being near to zero can be solved by a Transformation Procedure which increases slightly the little values.

III. THE MASH SYSTEM

- Subdirectory **MASH**:

- Subdirectory **MASHPAR** (parametrization program)
- **Main Program Files of MASH**
- Subdirectory **MASHAUTO** (automatic homogenization program)
- Subdirectory **MASHEND** (finishing program)
- Subdirectory **MASHMANU** ("manual" programs)
- Subdirectory **MASHSUB** (don not use it including "subroutines")

MASH IN PRACTICE

I. Parametrization in Subdirectory MASH\MASHPAR

MASHPAR.BAT
Data File, Significance level (0.1, 0.05, 0.01), Table of Reference System OR Table of Filambda Station Coordinates, Table of META DATA

II. The Main Program Steps in Subdirectory MASH

1. Automatic filling of missing values (MASHMISS.BAT)
It is obligatory in case of missing values! It can be repeated!

2. The further steps can be used optionally

MASHVERI.BAT: To verify the actual or the final stage of homogenization.

MASHGAME.BAT: An intensive examination for correction of one of the examined series in a playful way.

MASHCOR.BAT: Possibility for manual correction of examined series.

MASHDRAW.BAT: Graphic series.

MASHLIER.BAT: For automatic correction of outliers.

AUTOMATIC, ITERATIVE application of MASHGAME.BAT i.e.:

Running two **Batch Files in Subdirectory MASH\MASHAUTO:**

i, MAUTOPAR.BAT: Parametrization; input: number of iteration steps

ii, MASHAUTO.BAT: Examination, homogenization

Remark: During running of MASHAUTO.BAT the verification results are generated automatically in the files MASHVERI.RES, MASHVERO.RES.

(The steps (1 -2) can be repeated optionally!!!!!)

III. Finishing in Subdirectory MASHEND

The final results are saved in **MASHEND** by **MASHEND.BAT**

THE MAIN PROGRAM and I/O FILES of Subdirectory MASHPAR

1. Executive File

MASHPAR.BAT : Parametrization and a transformation procedure for the data which are near 0, in case of cumulative model.

2. Input Files and Input Data

Data File:

Format of Data File (maximal number of series: 500, maximal length of series: 200):
row 1: names of series or stations (obligatory!)
column 1: series of dates (I4)
column i+1: series i.
Data Format:
additive model (for example temperature): F6.2
cumulative model: I6 (data must be nonnegative!)
(for example precipitation, values multiplied by ten)
Mark of Missing Values:
additive model:999.99 ; cumulative model:999999
(For example: HUNTEMP.DAT)

Significance level: 0.1 or 0.05 or 0.01

Table of Reference System:
Indexes of reference series belonging to the candidate series. For example: HUNTEMP.REF

OR: Table of Filambda Station Coordinates: For example: HUNCOORD.PAR

Table of META DATA: Probable dates of the Break Points. For example: HUNMETA.DAT

3. Result Files written in Subdirectory MASH

SEE: Data and Result Files of Subdirectory MASH:

MASHPAR.PAR, MASHPAR2.PAR, MASHMETA.DAT, MASHDAT.SER, MASHMISS.SER, MASHINH.SER, MASHHOM.SER

4. Parameter Files

MASHPAR1.PAR, MASHPAR2.PAR

THE MAIN PROGRAM and I/O FILES of Subdirectory MASHEND

1. Executive File

MASHEND.BAT : Finishing and a retransformation procedure in case of cumulative model.

2. Result Files

MASHMISS.SER : Original data series (with missing values).

MASHDAT.SER : Original data series (with filled missing values).

MASHHOM.SER : Homogenized data series.

MASHINH.SER : Inhomogeneity series**.**

3. Parameter File: MASHPAR2.PAR

THE MAIN PROGRAM and I/O FILES of Subdirectory MASH

1.1 Executive Files of MASH

MASHMISS.BAT : Automatic filling of missing values.

MASHLIER.BAT : For automatic correction of outliers.

MASHGAME.BAT:
An intensive examination for correction of one of the examined series in a playful way.

MASHCOR.BAT : Possibility for manual correction of examined series.

MASHDRAW.BAT: Graphic series.

MASHVERI.BAT : Verification of Homogenization, evaluation of Meta Data

1.2 AUTOMATIC Homogenization Procedure in MASH\MASHAUTO:

i, MAUTOPAR.BAT: Parametrization; input: number of iteration steps

ii, MASHAUTO.BAT: Examination, homogenization

1.3 Executive Files of Subdirectory MASH\MASHMANU ("Manual" Program Files)

MASHSELR.BAT: Help for selection of reference series.

MASHEX1.BAT : To examine the optimal series belonging to the candidate series.

MASHEX2.BAT : To examine the optimal series system belonging to the candidate series.

MASHAUTC.BAT: Automatic correction of candidate series .

Remark: The „manual" program files (MASHSELR.BAT, MASHEX2.BAT, MASHAUTC.BAT) have been automatized. Their combined automatic version is the program file MASHGAME.BAT which is recommended to use instead of them.

2. Data and Result Files

MASHPAR.PAR : Parameters, Table of Reference System.

MASHMETA.DAT: Table of META DATA.

MASHMISS.SER : Original data series (with missing values).

MASHMISS.RES : Statistical results of filling missing values.

MASHDAT.SER : Original data series (with filled missing values).

MASHINH.SER : Inhomogeneity series**.**

MASHHOM.SER : Homogenized data series.

MASHEX1.RES : Statistical results: optimal difference series belonging to the candidate series and its detected inhomogeneities.

MASHEX1.SER : Result series: optimal difference series belonging to the candidate series and its inhomogeneity series.

MASHEX2.RES : Statistical results: optimal difference series system belonging to the candidate series and the detected inhomogeneities of the system elements.

MASHEX2.SER : Result series: optimal difference series system belonging to the candidate series and the inhomogeneity series of the system elements.

MASHCOR.RES : Detected break points, outliers and shifts (additive model) or ratios (cumulative model).

MASHSELR.RES : Table for selection of reference series.

MASHVERI.RES : Result of Verification file MASHVERI.BAT .

MASHVERO.RES : Result of Verification file MASHVERI.BAT (ordered statistics).

3. Work and Parameter Files

MASHPAR2.PAR, MASHPRCR.PAR, MASHSTEP.PAR, MASHMETA.PAR, MASHEINH.SER,MASHAUTC.INP, MASHAUTC.IND, GAME1.PAR, GAME2.PAR, GAME3.PAR, GAME4.PAR, GAME5.PAR, GAME6.PAR

FILES of Subdirectory MASHSUB (“Subroutines”)

GAMEAUTA.EXE, GAMEAUTO.EXE, GAMESELA.EXE, GAMESELO.EXE, MASHAUTA.EXE, MASHAUTC.EXE, MASHAUTG.EXE, MASHAUTO.EXE, MASHCOR.EXE, MASHDRAW.EXE, MASHEX1.EXE, MASHEX2.EXE, MASHEX2A.EXE, MASHEX2G.EXE, MASHEX2O.EXE, MASHHELP.EXE, MASHHELX.EXE,MASHINV.EXE, MASHMISS.EXE, MASHPAR.EXE, MASHSELA.EXE, MASHSELG.EXE, MASHSELO.EXE, MASHSELR.EXE, MASHSETA.EXE, MASHSETG.EXE, METAHELP.EXE, MASHTRAN.EXE, MASHVERI.EXE, METAVERI.EXE

IV. THE SAM SYSTEM

The Suggested Step by Step Procedure:

0. Examination of the annual series. The detected break points can be useful information for the further steps 1-3.

1. Examination of the monthly series for all the 12 months. Homogenization of the monthly series.

2. Examination of the seasonal series for residual inhomogeneity. Homogenization of the monthly series.

3. Examination of the annual series for residual inhomogeneity. Homogenization of the monthly series.

THE STRUCTURE OF SAM SYSTEM

- Subdirectory **SAM**:
    - Subdirectory **SAMPAR** (parametrization program)
    - Subdirectory **MASH** (examination of annual series)
    - **Main Program Files of SAM**
    - Subdirectory **SAMMISS** (data completion and QC programs)
    - Subdirectory **SAMVERI** (verification programs)
    - Subdirectory **SAMAUTO** (automatic homogenization programs)
    - Subdirectory **SAMMANU** ("manual" programs)
    - Subdirectory **SAMEND** (finishing program)
    - Subdirectory **SAMSUB** (don not use it including "subroutines")

SAM IN PRACTICE

I. Parametrization in Subdirectory SAM\SAMPAR (SAMPAR.BAT)

Data Files, Significance level (0.1, 0.05, 0.01), Table of Reference System OR Table of Filambda Station Coordinates, Table of META DATA

Remark
In directory **MASHMONTHLY\COSTHOMEINPUT:**
input files in format of COST Action ES0601 (HOME) can be converted.
Possibility for making monthly input files in MASH Format by **COSTHOMEINPUT.BAT**
Input: all.txt files according to COSTHOME Format
Output: m{j}. (j=1,…,12) (data files) and filastat.par (filambda station coordinates), input.par

II. Examination of annual series in Subdirectory SAM\MASH:

(the detected break points can be useful "metadata" information for the monthly series)

- Automatic parametrization in subdirectory **MASH\MASHPAR** by **MASHPARSAM.BAT**
  Input automatically: annual series with missing values
  (Remark: **SAMMISSOUT.BAT** may be used before in subirectory **SAM\SAMMISS** for the purpose of QC of monthly values)
- Homogenization of annual series in directory **MASH** (page 23)
- Finishing of annual examination in subdirectory **MASH\MASHEND:**
  annual metadata for main directory **SAM** by **MASHENDSAM.BAT**
  (Back to main directory SAM)

III. The Main Program Possibilities in Directory SAM

DATA COMPLETION for all the 12 Months together:
(Automatic version of MASHMISS.BAT. It can be repeated!)
Running Batch File **SAMMISS.BAT in Subirectory SAM\SAMMISS.**

DATA COMPLETION and OUTLIER DETECTION for all the 12 Months together:
(It can be repeated!)
Running Batch File **SAMMISSOUT.BAT in Subirectory SAM\SAMMISS.**
(SAMMISS.BAT or SAMMISSOUT.BAT is obligatory in case of having missing values!)

VERIFICATION PROCEDURE for all Monthly, Seasonal and Annual Series:
(Automatic version of MASHVERI.BAT. It can be repeated!)
Running Batch File **SAMVERI.BAT in Subirectory SAM\SAMVERI**
Output Files in Directory SAM: **V{j}.** (j=1,….,17) and **VERISUM** (summary)

AUTOMATIC HOMOGENIZATION for all the 12 Months in SAM\SAMAUTO:
Running two batch files, Sautopar.. (parametrization) and Samauto..(homogenization):
a, Strict decision rule (detected breaks: only metadata)
Running **SAUTOPAR12S.BAT and SAMAUTO12.BAT**
b, Basic decision rule (detected breaks: meta data or "undoubtful" breaks)
Running **SAUTOPAR12B.BAT and SAMAUTO12.BAT**
c, Light decision rule (detected breaks: arbitrary breaks)
Running **SAUTOPAR12L.BAT and SAMAUTO12.BAT**
The steps a,b,c, together: **SAUTOPAR12SBL.BAT and SAMAUTO12.BAT**
Verification results are generated automatically in the files **V{j}.** (j=1,….,12) .

EXAMINATIONS OF CHOSEN MONTHLY, SEASONAL SERIES

1. Taking the chosen monthly or seasonal series In (SAMIN.BAT)

2. The further steps can be used optionally

MASHMISS.BAT : Automatic filling of missing values.

MASHVERI.BAT : To verify the actual or the final stage of homogenization.

MASHGAME.BAT: An intensive examination for correction of one of the examined series in a playful way.

MASHCOR.BAT: Possibility for manual correction of examined series.

MASHDRAW.BAT: Graphic series.

MASHLIER.BAT: For automatic correction of outliers.

AUTOMATIC, ITERATIVE application of MASHGAME.BAT i.e.:
Running two **Batch Files in Subdirectory SAM\SAMAUTO:**
i, SAUTOPAR.BAT: Parametrization; input: number of iteration steps
ii, SAMAUTO.BAT: Examination, homogenization
(possible decision rules: strict, basic, light)

AUTOMATIC PROCEDURE in strict, basic and light ways together:
Running **Batch Files in Subdirectory SAM\SAMAUTO:**
i, SAUTOPARSBL.BAT: Parametrization; input: number of iteration steps
ii, SAMAUTO.BAT: Examination, homogenization

Remark: During running of SAMAUTO.BAT the verification results are generated automatically in the files MASHVERI.RES, MASHVERO.RES.

3. The further step can be used in case of Seasonal Series

SAMTEST.BAT : Test for comparison of the inhomogeneities between the seasonal series and the appropriate monthly series, moreover procedure for selecting stations which have different inhomogeneities between the seasonal series and the appropriate monthly series.

4. Taking the chosen monthly or seasonal series Out (SAMOUT.BAT)

(The steps (1 - 4) can be repeated optionally!!!!!)

IV. Finishing in Subdirectory SAMEND

The final results are saved in **SAMEND** by **SAMEND.BAT**

Remark
In directory MASHMONTHLY\COSTHOMEOUTPUT:
Posssibility for making output files in COSTHOME Format by
COSTHOMEOUTPUT.BAT
Input: from **SAMEND** and **COSTHOMEINPUT**
Output: all .txt files according to COSTHOME Format

THE MAIN PROGRAM and I/O FILES of Subdirectory SAMPAR

1. Executive File

SAMPAR.BAT : Parametrization and a transformation procedure for the data which are near 0, in case of cumulative model.

2. Input Files and Input Data

12 Data Files:

m{**j**} (j=1,....,12): original monthly series

Format of Data Files (maximal number of stations: 500, maximal length of series: 200):
row 1: station names (obligatory!)
  column 1: series of dates (I4)
  column i+1: series i.
Data Format:
  additive model (for example temperature): F6.2
  cumulative model: I6 (data must be nonnegative!)
  (for example precipitation, values multiplied by ten)
Mark of Missing Values:
  additive model:999.99
  cumulative model:999999

Significance level: 0.1 or 0.05 or 0.01

Table of Reference System:
Indexes of reference series belonging to the candidate series.
For example: HUNTEMP.REF

OR: Table of Filambda Station Coordinates:
For example: HUNCOORD.PAR

Table of META DATA:
Probable dates of the Break Points. For example: HUNMETA.DAT

3. Result Files written in Subdirectory SAM

SEE Data and Result Files of Subdirectory SAM:

m{**j**}, **m**{**j**}**h**, **m**{**j**}**i**, **m**{**j**}**c** (j=1,....,12)

s{**j**}, **s**{**j**}**h**, **s**{**j**}**i**, **s**{**j**}**ei**, **s**{**j**}**c** (j=1, 2, 3, 4)

year, **yearh**, **yeari**, **yearei**, **yearc**

SAMPAR.PAR, MASHPAR.PAR, MASHMETA.DAT

4. Parameter Files

SAMPAR4.PAR, SAMPAR5.PAR, SAMPAR6.PAR

THE MAIN PROGRAM and I/O FILES of Subdirectory SAMEND

1. Executive File

SAMEND.BAT : Finishing and a retransformation procedure in case of cumulative model.

2. Result Files

m{**j**} (j=1,....,12): original monthly series (with filled missing values).

s{**j**} (j=1, 2, 3, 4) : original seasonal series (with filled missing values).
(winter = {1, 2, 12 }, spring = {3, 4, 5 }, summer = {6, 7, 8 }, autumn = {9, 10, 11 }).

year : original annual series (with filled missing values).

m{**j**}**h** (j=1,....,12): homogenized monthly series.

s{**j**}**h** (j=1, 2, 3, 4): homogenized seasonal series (based on homogenized monthly series).

yearh : homogenized annual series (based on homogenized monthly series).

m{**j**}**i** (j=1,....,12) : estimated inhomogeneity series for months.

s{**j**}**i** (j=1, 2, 3, 4) : estimated inhomogeneity series for seasons.

yeari : estimated inhomogeneity series for year.

s{**j**}**ei** (j=1, 2, 3, 4): estimated "expectation" of inhomogeneity series for seasons.

yearei : estimated "expectation" of inhomogeneity series for year.

m{**j**}**c** (j=1,....,12) : break points and shifts (add. m.) or ratios (cum. m.) for months.

s{**j**}**c** (j=1, 2, 3, 4) : break points and shifts (add. m.) or ratios (cum. m.) for seasons.

yearc : break points and shifts (add. m.) or ratios (cum. m.) for year.

v{j}. (j=1,….,17): verification statistics for the months and seasons
(winter: 13 spring: 14 summer: 15 autumn: 16 year: 17)

verisum: summary of verification statistics

3. Parameter File

SAMPAR5.PAR

THE MAIN PROGRAM and I/O FILES of Subdirectory SAM

1. Executive Files

1.1 Special Executive Files of SAM System

SAMIN.BAT : Taking the chosen monthly or seasonal series In.

SAMOUT.BAT : Taking the chosen monthly or seasonal series Out.

SAMTEST.BAT : Test for comparison of the inhomogeneities between the seasonal series and the appropriate monthly series. Moreover, procedure for selecting stations that have different inhomogeneities between the seasonal series and the appropriate monthly series.

1.2 Executive Files of MASH System

MASHMISS.BAT : Automatic filling of missing values.

MASHLIER.BAT : For automatic correction of outliers.

MASHGAME.BAT:
An intensive examination for correction of one of the examined series in a playful way.

MASHCOR.BAT : Possibility for manual correction of examined series.

MASHDRAW.BAT: Graphic series.

MASHVERI.BAT : Verification of Homogenization, evaluation of Meta Data.

1.3 Executive Files in SAM\SAMMISS

SAMMISS.BAT: Data completion for all the 12 months together.

SAMMISSOUT.BAT: Data completion and outlier detection for all the 12 months together.

1.3 Executive Files in SAM\SAMVERI

SAMVERI.BAT: Verification Procedure for all monthly, seasonal and annual series:

1.4 Automatic Homogenization Procedures in SAM\SAMAUTO

Running two batch files, Sautopar.. (parametrization) and Samauto..(homogenization):

AUTOMATIC HOMOGENIZATION for all the 12 Months
a, Strict decision rule (detected breaks: only metadata)
Running **SAUTOPAR12S.BAT and SAMAUTO12.BAT**
b, Basic decision rule (detected breaks: meta data or “undoubtful” breaks)
Running **SAUTOPAR12B.BAT and SAMAUTO12.BAT**
c, Light decision rule (detected breaks: arbitrary breaks)
Running **SAUTOPAR12L.BAT and SAMAUTO12.BAT**
The steps a,b,c, together: **SAUTOPAR12SBL.BAT and SAMAUTO12.BAT**

AUTOMATIC HOMOGENIZATION for a chosen Month, Season or Year
The operational way (strict, basic, light) may be chosen:
Running **SAUTOPAR.BAT and SAMAUTO.BAT**
The operational ways (strict, basic, light) together:
Running **SAUTOPARSBL.BAT and SAMAUTO.BAT**

1.5 Executive Files of Subdirectory SAM\SAMMANU (“Manual” Program Files)

MASHSELR.BAT: Help for selection of reference series.

MASHEX1.BAT : To examine the optimal series belonging to the candidate series.

MASHEX2.BAT : To examine the optimal series system belonging to the candidate series.

MASHAUTC.BAT: Automatic correction of candidate series .

Remark: The „manual” program files (MASHSELR.BAT, MASHEX2.BAT, MASHAUTC.BAT) have been automatized. Their combined automatic version is the program file MASHGAME.BAT which is recommended to use instead of them.

2. Data and Result Files

2.1 Special Data and Result Files of SAM System

m{**j**} (j=1,....,12): original monthly series

s{**j**} (j=1, 2, 3, 4) : original seasonal series
(winter = {1, 2, 12 }, spring = {3, 4, 5 }, summer = {6, 7, 8 }, autumn = {9, 10, 11 }).

year : original annual series.

m{**j**}**h** (j=1,....,12): homogenized monthly series.

s{**j**}**h** (j=1, 2, 3, 4): homogenized seasonal series (based on homogenized monthly series).

yearh : homogenized annual series (based on homogenized monthly series).

m{**j**}**i** (j=1,....,12) : estimated inhomogeneity series for months.

s{**j**}**i** (j=1, 2, 3, 4) : estimated inhomogeneity series for seasons.

yeari : estimated inhomogeneity series for year.

s{**j**}**ei** (j=1, 2, 3, 4): estimated "expectation" of inhomogeneity series for seasons.

yearei : estimated "expectation" of inhomogeneity series for year.

m{**j**}**c** (j=1,....,12) : break points and shifts (add. m.) or ratios (cum. m.) for months.

s{**j**}**c** (j=1, 2, 3, 4) : break points and shifts (add. m.) or ratios (cum. m.) for seasons.

yearc : break points and shifts (add. m.) or ratios (cum. m.) for year.

v{j}. (j=1,….,17): verification statistics for the months and seasons
(winter: 13 spring: 14 summer: 15 autumn: 16 year: 17)

verisum: summary of verification statistics

SAMPAR.PAR : Parameters, Table of Reference System.

SAMTEST.RES: Output of SAMTEST.BAT.

2.2 Data and Result Files of MASH System

MASHPAR.PAR : Parameters, Table of Reference System.

MASHMETA.DAT: Table of META DATA.

MASHMISS.SER : Original data series (with missing values).

MASHMISS.RES : Statistical results of filling missing values.

MASHDAT.SER : Original data series (with filled missing values).

MASHINH.SER : Inhomogeneity series**.**

MASHHOM.SER : Homogenized data series.

MASHEX1.RES : Statistical results: optimal difference series belonging to the candidate series and its detected inhomogeneities.

MASHEX1.SER : Result series: optimal difference series belonging to the candidate series and its inhomogeneity series.

MASHEX2.RES : Statistical results: optimal difference series system belonging to the candidate series and the detected inhomogeneities of the system elements.

MASHEX2.SER : Result series: optimal difference series system belonging to the candidate series and the inhomogeneity series of the system elements.

MASHCOR.RES : Detected break points, outliers and shifts (additive model) or ratios (cumulative model).

MASHSELR.RES : Table for selection of reference series.

MASHVERI.RES : Result of Verification file MASHVERI.BAT .

MASHVERO.RES : Result of Verification file MASHVERI.BAT (ordered statistics).

3. Work and Parameter Files

3.1 Special Work and Parameter Files of SAM System

SAMPAR2.PAR, SAMPAR3.PAR, SAMPAR4.PAR, SAMPAR5.PAR, SAMPRCR.PAR, SAMTEST.PAR, SAMORINH.SER, SAMTESTD.SER, SAMTESTI.SER,SAMTIMER.PAR

3.2 Work and Parameter Files of MASH System

MASHPAR2.PAR, MASHPRCR.PAR, MASHSTEP.PAR, MASHMETA.PAR, MASHEINH.SER,MASHAUTC.INP, MASHAUTC.IND, GAME1.PAR, GAME2.PAR, GAME3.PAR, GAME4.PAR, GAME5.PAR, GAME6.PAR

FILES of Subdirectory SAMSUB (“Subroutines”)

SAMHELP1.EXE, SAMHELP2.EXE, SAMHELP3.EXE, SAMIN1.EXE, SAMIN2.EXE, SAMINV.EXE, SAMMISS.EXE, SAMOUT1.EXE, SAMOUT2.EXE, SAMOUT3.EXE, SAMPAR.EXE, SAMTESTC.EXE, SAMTESTS.EXE, SAMTEST1.EXE, SAMTEST2.EXE, SAMTRAN.EXE

V. EXAMPLE FOR APPLICATION OF MASH SYSTEM

Data File: HUNTEMP.DAT

Examined Series: Hungarian annual mean temperature series (1901-1999).

Examined Stations:
1. Budapest (bp), 2. Debrecen (de), 3. Kecskemét (ke), 4. Miskolc (mi), 5. Mosonmagyaróvár (mo), 6. Nyíregyháza (ny), 7. Pécs (pe), 8. Sopron (sr), 9. Szeged (se), 10. Szombathely (so)

Table of Reference System: HUNTEMP.REF

```
TABLE OF REFERENCE SYSTEM (two rows belong to each examined series)
row 1: index of candidate series(I3); number of reference series(I3)
row 2: indexes of reference series(I3)
  1  9
  2  3  4  5  6  7  8  9 10
  2  6
  1  3  4  6  7  9
  3  9
  1  2  4  5  6  7  8  9 10
  4  9
  1  2  3  5  6  7  8  9 10
  5  7
  1  3  4  7  8  9 10
  6  6
  1  2  3  4  7  9
  7  9
  1  2  3  4  5  6  8  9 10
  8  7
  1  3  4  5  7  9 10
  9  9
  1  2  3  4  5  6  7  8 10
 10  7
  1  3  4  5  7  8  9
```

Table of META DATA: HUNMETA.DAT

```
TABLE OF META DATA (one or two rows belong to each examined series)
row 1: index of examined series(I3); number of meta data(I5)
row 2: meta data(I5), if they exist
  1    8
 1909 1960 1986 1987 1988 1991 1992 1993
  2    3
 1950 1954 1955
  3    7
 1943 1944 1945 1946 1947 1969 1970
  4    6
 1922 1930 1938 1950 1964 1965
  5    5
 1950 1960 1966 1969 1970
  6    8
 1950 1951 1960 1965 1966 1967 1991 1992
  7    4
 1950 1957 1958 1960
  8    1
 1973
  9    2
 1950 1951
 10    1
 1950
```

Table of Filambda Station Coordinates: HUNCOORD.PAR

```
index    lambda(x)        fi(y)
1       19.02499960   47.50833510   Budapest
2       21.60833360   47.49166490   Debrecen
.
.
9       20.09166720   46.25833510   Szeged
10      16.63333320   47.26666640   Szombathely
```

____

```
 99 1021.60a12.00 8.08  .05
 Name of Data File: huntemp.dat           MISSING VALUES!
 Model: additive
 Number of series:  10
 Length of series:  99
 Significance level:   .05
 Critical value for break points: 21.60
 Critical value for correction: 12.00
 Critical value for outliers:  8.08

 EXAMINED SERIES AND INDEXES

      bp: 1       de: 2       ke: 3       mi: 4       mo: 5       ny: 6
      pe: 7       sr: 8       se: 9       so:10

 TABLE OF REFERENCE SYSTEM (two rows belong to each examined series)
 row 1: index of candidate series(I3); number of reference series(I3)
 row 2: indexes of reference series(I3)
  1  9
  2  3  4  5  6  7  8  9 10
  2  6
  1  3  4  6  7  9
  3  9
  1  2  4  5  6  7  8  9 10
  4  9
  1  2  3  5  6  7  8  9 10
  5  7
  1  3  4  7  8  9 10
  6  6
  1  2  3  4  7  9
  7  9
  1  2  3  4  5  6  8  9 10
  8  7
  1  3  4  5  7  9 10
  9  9
  1  2  3  4  5  6  7  8 10
 10  7
  1  3  4  5  7  8  9

 File of Meta Data: MASHMETA.DAT
 Original series (with missing values): MASHMISS.SER
 Original series (without missing values): MASHDAT.SER
 Homogenized series: MASHHOM.SER
 Inhomogeneity series: MASHINH.SER
 Automatic filling of missing values: MASHMISS.BAT
 Automatic correction of outliers: MASHLIER.BAT
 GAME of MASH: MASHGAME.BAT
 Non-automatic correction: MASHCOR.BAT
 Verification of Homogenization: MASHVERI.BAT
```

```
Graphics: MASHDRAW.BAT
```

Figure 1. Output of Parametrization (MASHPAR.PAR)

```
                   CANDIDATE SERIES:     bp

VARIANCE & DEVIATION:   .4865   .6975
DATE OF MISSING VALUE:  1916
EXCLUDED REFERENCE SERIES:     de
OPTIMAL POSITIVE WEIGHTING
REFERENCE SERIES, WEIGHTING FACTORS, ERRORS

              ke      ny      sr      so      Variance      std.error
    bp    .16281  .24556  .54768  .04396        .06357        .25214

INTERCEPT:    1.34
ESTIMATED VALUE:   11.73


                   CANDIDATE SERIES:     de

VARIANCE & DEVIATION:   .5411   .7356
DATE OF MISSING VALUE:  1916
EXCLUDED REFERENCE SERIES:     bp
OPTIMAL POSITIVE WEIGHTING
REFERENCE SERIES, WEIGHTING FACTORS, ERRORS

              ke      ny      Variance      std.error
    de    .28617  .71383        .06031        .24559

INTERCEPT:    -.04
ESTIMATED VALUE:   10.46


                   CANDIDATE SERIES:     de

VARIANCE & DEVIATION:   .5411   .7356
DATE OF MISSING VALUE:  1928
THERE IS NO EXCLUDED REFERENCE SERIES
OPTIMAL POSITIVE WEIGHTING
REFERENCE SERIES, WEIGHTING FACTORS, ERRORS

              bp      ke      mi      ny      Variance      std.error
    de    .35121  .12020  .02383  .50476        .04568        .21374

INTERCEPT:    -.41
ESTIMATED VALUE:    9.73


                   CANDIDATE SERIES:     de

VARIANCE & DEVIATION:   .5411   .7356
DATE OF MISSING VALUE:  1996
EXCLUDED REFERENCE SERIES:     pe
OPTIMAL POSITIVE WEIGHTING
REFERENCE SERIES, WEIGHTING FACTORS, ERRORS

              bp      ke      mi      ny      Variance      std.error
    de    .35121  .12020  .02383  .50476        .04568        .21374

INTERCEPT:    -.41
ESTIMATED VALUE:    9.42
```

Figure 2. Part of Statistical Results of Filling Missing Values (MASHMISS.RES)

```
TEST STATISTICS FOR SERIES INHOMOGENEITY
Null hypothesis: the examined series are homogeneous.
Critical value (significance level .05):  21.60
Test statistics (TS) can be compared to the critical value.
The larger TS values are more suspicious!

Series  Index     TS       Series  Index     TS       Series  Index     TS
    bp    1     719.39         de    2     151.68         ke    3     599.99
    mi    4    1180.70         mo    5     160.65         ny    6     137.37
    pe    7     457.81         sr    8     111.60         se    9     828.81
    so   10     100.97
 AVERAGE:    444.90
```

Figure 3. Part of First Output of Verification Program MASHVERI.BAT (MASHVERI.RES)

```
TEST STATISTICS FOR EVALUATION OF META DATA
Null hypothesis: the inhomogeneities can be explained by the Meta Data.
Critical value (significance level .05):  21.60
Test statistics (TSM) can be compared to the critical value.
The larger TSM values are more suspicious!

Series  Index     TSM      Series  Index     TSM      Series  Index     TSM
    bp    1      53.09         de    2      41.63         ke    3      96.93
    mi    4    1180.70         mo    5      88.35         ny    6     120.16
    pe    7     228.76         sr    8      41.62         se    9      92.58
    so   10      77.92
 AVERAGE:    202.17
```

Figure 4. Part of First Output of Verification Program MASHVERI.BAT (MASHVERI.RES)

Application of Program MASHGAME.BAT (one step)

```
HELP: TABLE FOR SELECTION OF REFERENCE SERIES AND/OR CANDIDATE SERIES
Null hypothesis 1: the examined series are homogeneous.
Test Statistics belonging to the null hypothesis 1: TS
Null hypothesis 2: the inhomogeneities can be explained by the Meta Data.
Test Statistics belonging to the null hypothesis 2: TSM
Critical value (significance level .05):  21.60
Test Statistics (both TS and TSM) can be compared to the critical value.
The larger Test Statistics are more suspicious!
Series marked with asterisk(*) are not used for reference series.

Candidate series:     mi   Index:  4     TS: 1155.78*     TSM: 1155.78
Reference series:     bp   Index:  1     TS:  279.07*     TSM:   57.92
Reference series:     de   Index:  2     TS:   68.20      TSM:   49.58
Reference series:     ke   Index:  3     TS:   96.73      TSM:   35.91
Reference series:     mo   Index:  5     TS:   82.01      TSM:   62.51
Reference series:     ny   Index:  6     TS:  177.52*     TSM:   56.82
Reference series:     pe   Index:  7     TS:  512.79*     TSM:  185.26
Reference series:     sr   Index:  8     TS:  104.88      TSM:   56.83
Reference series:     se   Index:  9     TS:  934.22*     TSM:   83.67
Reference series:     so   Index: 10     TS:  162.95*     TSM:  116.41
```

Figure 5. Partial Output of Program MASHGAME.BAT (On the Screen)

```
CANDIDATE SERIES:     mi      (Index:   4)

 NUMBER OF DIFFERENCE SERIES: 2
 REFERENCE SERIES, WEIGHTING FACTORS, VARIANCE OF DIFFERENCE SERIES

             ke      mo      Variance       Deviation
     mi    .66227  .33773        .06984        .26427

             de      sr      Variance       Deviation
     mi    .77541  .22459        .04675        .21621

 NO FORMER ESTIMATED BREAKS

 EXAMINATION OF DIFFERENCE SERIES

 1. DIFFERENCE SERIES
 BREAK POINTS ( critical value:  21.60 )
 Test statistic before homogenization of diff. s.:  420.27
             Date    Conf. Int.    Stat.   Shift      Conf. Int.
                                    8.46       +
         1    1908    [1908,1908]  420.27   -2.03    [ -2.38, -1.69]
                                   19.54       -
         2    1921    [1919,1922]   80.06     .82    [   .52,  1.19]
                                    4.38       +
         3    1931    [1929,1932]   59.00    -.80    [ -1.19,  -.45]
                                    2.11       +
         4    1939    [1937,1940]   38.34     .84    [   .37,  1.30]
                                    5.74       +
         5    1943    [1941,1949]   22.07    -.61    [ -1.24,  -.19]
                                    1.04       -
         6    1950    [1945,1959]   22.73     .47    [   .14,   .88]
                                    4.71       -
         7    1964    [1962,1967]   33.55     .67    [   .27,  1.06]
                                    3.36       +
         8    1969    [1968,1971]   39.69    -.67    [ -1.04,  -.30]
                                   10.64       -
 Test statistic after homogenization of diff. s.:   19.54

 2. DIFFERENCE SERIES
 BREAK POINTS ( critical value:  21.60 )
 Test statistic before homogenization of diff. s.:  895.43
             Date    Conf. Int.    Stat.   Shift      Conf. Int.
                                    2.21       -
         1    1904    [1902,1906]   26.92     .48    [   .22,  1.08]
                                    6.82       -
         2    1908    [1908,1908]  498.17   -2.09    [ -2.42, -1.77]
                                    1.76       +
         3    1916    [1915,1916]   35.92    -.52    [  -.83,  -.22]
                                    2.65       +
         4    1921    [1921,1922]  158.75    1.06    [   .77,  1.35]
                                   12.16       +
         5    1931    [1929,1932]   44.86    -.41    [  -.87,  -.28]
                                    3.74       +
         6    1939    [1933,1940]   25.43     .38    [   .16,   .88]
                                    7.68       -
         7    1944    [1942,1944]   37.20    -.50    [ -1.22,  -.34]
                                   12.18       +
        8    1950    [1950,1950]  194.56     .92    [   .70,  1.16]
                                    13.91      +
 Test statistic after homogenization of diff. s.:   17.64
```

Figure 6. Statistical Partial Results of Program MASHGAME.BAT (MASHEX2.RES)

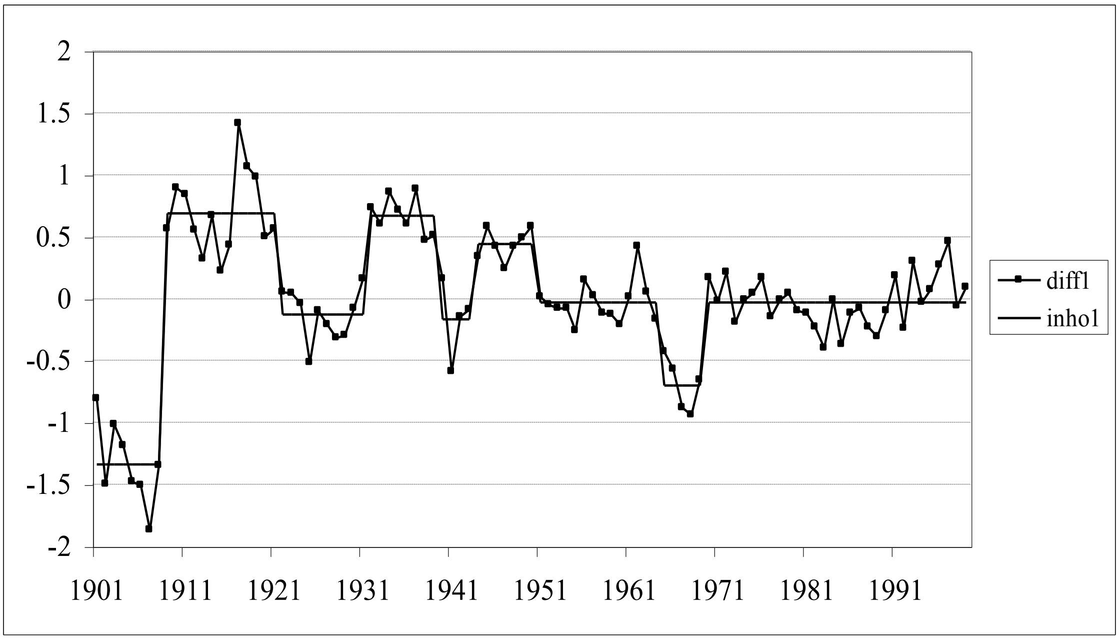


Figure 7. Graphic Partial Results of Program MASHGAME.BAT: Difference series 1 with its estimated Inhomogeneity series (MASHEX2.SER, MASHDRAW.BAT)

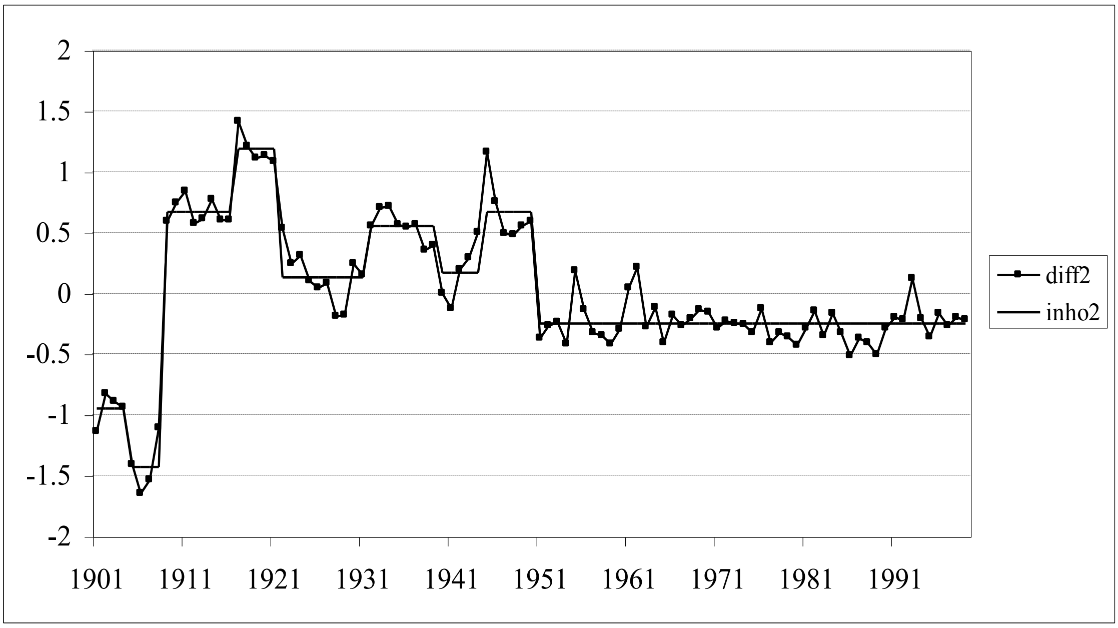


Figure 8. Graphic Partial Results of Program MASHGAME.BAT: Difference series 2 with its estimated Inhomogeneity series (MASHEX2.SER, MASHDRAW.BAT)

```
ESTIMATED BREAK POINTS AND SHIFTS
 (Mark M: META DATA)

   bp:
 No Break Points

   de:
 No Break Points

   ke:
 No Break Points

   mi:
  1908: -1.78/  M1922:   .78/  M1930:  -.41/  M1938:   .38/   1944:  -.35/
 M1950:   .47

   mo:
 No Break Points

   ny:
 No Break Points

   pe:
 No Break Points

   sr:
 No Break Points

   se:
 No Break Points

   so:
 No Break Points
```

Figure 9. Result of Examination made by Program MASHGAME.BAT (MASCOR.RES)

```
HELP: TABLE FOR SELECTION OF REFERENCE SERIES AND/OR CANDIDATE SERIES
Null hypothesis 1: the examined series are homogeneous.
Test Statistics belonging to the null hypothesis 1: TS
Null hypothesis 2: the inhomogeneities can be explained by the Meta Data.
Test Statistics belonging to the null hypothesis 2: TSM
Critical value (significance level .05):  21.60
Test Statistics (both TS and TSM) can be compared to the critical value.
The larger Test Statistics are more suspicious!
Series marked with asterisk(*) are not used for reference series.

Candidate series:     mi   Index:  4     TS:   76.28*     TSM:   57.04
Reference series:     bp   Index:  1     TS:  279.07*     TSM:   57.92
Reference series:     de   Index:  2     TS:   68.20      TSM:   49.58
Reference series:     ke   Index:  3     TS:   96.73      TSM:   35.91
Reference series:     mo   Index:  5     TS:   82.01      TSM:   62.51
Reference series:     ny   Index:  6     TS:  177.52*     TSM:   56.82
Reference series:     pe   Index:  7     TS:  512.79*     TSM:  185.26
Reference series:     sr   Index:  8     TS:  104.88      TSM:   56.83
Reference series:     se   Index:  9     TS:  934.22*     TSM:   83.67
Reference series:     so   Index: 10     TS:  162.95*     TSM:  116.41
```

Figure 10. Last Output of Program MASHGAME.BAT after Automatic Correction (On the Screen)

Verification of Homogenization (MASHVERI.BAT)

```
   I. TEST STATISTICS FOR SERIES INHOMOGENEITY
 Null hypothesis: the examined series are homogeneous.
 Critical value (significance level .05):  21.60
 Test statistics (TS) can be compared to the critical value.
 The larger TS values are more suspicious!

 1. Test Statistics After Homogenization
Series  Index     TSA         Series  Index    TSA         Series  Index    TSA
    bp     1     26.51          de     2     18.01          ke     3     29.93
    mi     4     22.64          mo     5     16.94          ny     6     22.20
    pe     7     26.99          sr     8     30.01          se     9     26.11
    so    10     13.89
 AVERAGE:     23.32
 2. Test Statistics Before Homogenization
Series  Index     TSB         Series  Index    TSB         Series  Index    TSB
    bp     1    719.39          de     2    151.68          ke     3    599.99
    mi     4   1180.70          mo     5    160.65          ny     6    137.37
    pe     7    457.81          sr     8    111.60          se     9    828.81
    so    10    100.97
 AVERAGE:    444.90
 3. Statistics for Estimated Inhomogeneities
(IS statistics can be compared with the TSB ones)
Series  Index     IS         Series  Index     IS         Series  Index     IS
    bp     1    570.28          de     2    212.85          ke     3    296.11
    mi     4   1121.99          mo     5     77.62          ny     6     37.03
    pe     7    438.62          sr     8     54.16          se     9    627.87
    so    10     90.94
 AVERAGE:    352.75

  II. CHARACTERIZATION OF INFHOMOGENEITY

 1. Relative Estimated Inhomogeneities
Series  Index     RI1        Series  Index     RI1        Series  Index     RI1
    bp     1       .36          de     2       .29          ke     3       .32
    mi     4       .52          mo     5       .22          ny     6       .12
    pe     7       .43          sr     8       .14          se     9       .45
    so    10       .21
 AVERAGE:       .30
 2. Relative Modification of Series
Series  Index     RI2        Series  Index     RI2        Series  Index     RI2
    bp     1       .49          de     2       .41          ke     3       .43
    mi     4       .53          mo     5       .26          ny     6       .12
    pe     7       .60          sr     8       .27          se     9       .76
    so    10       .21
 AVERAGE:       .41
 3. Lower Confidence Limit for Relative Residual Inhomogeneities
   (confidence level:  .95)
Series  Index     RI3        Series  Index     RI3        Series  Index     RI3
    bp     1       .01          de     2       .00          ke     3       .03
    mi     4       .00          mo     5       .00          ny     6       .00
    pe     7       .02          sr     8       .02          se     9       .03
    so    10       .00
 AVERAGE:       .01

 III. REPRESENTATIVITY OF STATION NETWORK
(1-relative interpolation error)
Series  Index      RS        Series  Index      RS        Series  Index      RS
    bp     1       .84          de     2       .82          ke     3       .82
    mi     4       .80          mo     5       .82          ny     6       .81
    pe     7       .78          sr     8       .80          se     9       .81
    so    10       .82
 AVERAGE:       .81
```

Figure 11.a, Verification Results after Finishing the Homogenization Procedure (MASHVERI.RES)

```
                          EVALUATION OF META DATA


 IV. TEST STATISTICS
 Null hypothesis: the inhomogeneities can be explained by the Meta Data.
 Critical value (significance level .05):  21.60
 Test statistics (TSM) can be compared to the critical value.
 The larger TSM values are more suspicious!
Series  Index     TSM      Series  Index     TSM      Series  Index     TSM
    bp     1     53.09         de     2     41.63         ke     3     96.93
    mi     4   1180.70         mo     5     88.35         ny     6    120.16
    pe     7    228.76         sr     8     41.62         se     9     92.58
    so    10     77.92
 AVERAGE:    202.17


  V. REPRESENTATIVITY OF META DATA
(Relative part of estimated inhomogeneity can be explained by the Meta Data)
Series  Index      RM      Series  Index      RM      Series  Index      RM
    bp     1       .55         de     2      1.00         ke     3       .33
    mi     4       .04         mo     5       .20         ny     6       .05
    pe     7       .49         sr     8      1.00         se     9       .52
    so    10       .05
 AVERAGE:       .42
```

Figure 11.b, Verification Results for Meta Data after Finishing the Homogenization Procedure (MASHVERI.RES)

VI. EXAMPLE FOR APPLICATION OF SAM SYSTEM

Data Files (monthly series)**: m{j}** (j=1,....,12)

Examined Series: Hungarian monthly mean temperature series (1901-1930).

Examined Stations:
1. Budapest (bp), 2. Debrecen (de), 3. Kecskemét (ke), 4. Miskolc (mi), 5. Mosonmagyaróvár (mo), 6. Nyíregyháza (ny), 7. Pécs (pe), 8. Sopron (sr), 9. Szeged (se), 10. Szombathely (so)

Table of Reference System: HUNTEMP.REF

```
TABLE OF REFERENCE SYSTEM (two rows belong to each examined series)
row 1: index of candidate series(I3); number of reference series(I3)
row 2: indexes of reference series(I3)
  1  9
  2  3  4  5  6  7  8  9 10
  2  6
  1  3  4  6  7  9
  3  9
  1  2  4  5  6  7  8  9 10
  4  9
  1  2  3  5  6  7  8  9 10
  5  7
  1  3  4  7  8  9 10
  6  6
  1  2  3  4  7  9
  7  9
  1  2  3  4  5  6  8  9 10
  8  7
  1  3  4  5  7  9 10
  9  9
  1  2  3  4  5  6  7  8 10
 10  7
  1  3  4  5  7  8  9
```

Table of Filambda Station Coordinates: HUNCOORD.PAR

```
index    lambda(x)        fi(y)
1       19.02499960   47.50833510   Budapest
2       21.60833360   47.49166490   Debrecen
.
.
9       20.09166720   46.25833510   Szeged
10      16.63333320   47.26666640   Szombathely
```

Table of META DATA: HUNMETA.DAT (for the given period)

```
TABLE OF META DATA (one or two rows belong to each examined series)
row 1: index of examined series(I3); number of meta data(I5)
row 2: meta data(I5), if they exist
  1    1
 1909
  2    0
  3    0
  4    1
 1922
  5    0
  6    0
  7    0
  8    0
  9    0
 10    0
```

```
30 1020.53a12.00 7.44  .05
Model: additive
Number of stations:  10
Length of series:  30
Significance level:   .05
Critical value for break points: 20.53
Critical value for correction: 12.00
Critical value for outliers:  7.44
EXAMINED STATIONS AND INDEXES

     bp: 1       de: 2       ke: 3       mi: 4       mo: 5       ny: 6
     pe: 7       sr: 8       se: 9       so:10

TABLE OF REFERENCE SYSTEM (two rows belong to each examined station)
row 1: index of candidate station(I3); number of reference stations(I3)
row 2: indexes of reference stations(I3)

  1  9
  2  3  4  5  6  7  8  9 10
  2  6
  1  3  4  6  7  9
  3  9
  1  2  4  5  6  7  8  9 10
  4  9
  1  2  3  5  6  7  8  9 10
  5  7
  1  3  4  7  8  9 10
  6  6
  1  2  3  4  7  9
  7  9
  1  2  3  4  5  6  8  9 10
  8  7
  1  3  4  5  7  9 10
  9  9
  1  2  3  4  5  6  7  8 10
 10  7
  1  3  4  5  7  8  9

File of Meta Data: MESHMETA.DAT
Original monthly series: M{J}, (J=1,..,12)
Original seasonal series: S{J}, (J=1,2,3,4)
(winter,spring,summer,autumn)
Original annual series: YEAR
Homogenized monthly series: M{J}H, (J=1,..,12)
Homogenized seasonal series: S{J}H, (J=1,2,3,4)
(winter,spring,summer,autumn)
Homogenized annual series: YEARH
Inhomogeneity series for months: M{J}I, (J=1,..,12)
Inhomogeneity series for seasons: S{J}I, (J=1,2,3,4)
(winter,spring,summer,autumn)
Inhomogeneity series for year: YEARI
Break Points and Shifts for months: M{J}C, (J=1,..,12)
Break Points and Shifts for seasons: S{J}C, (J=1,2,3,4)
(winter,spring,summer,autumn)
Break Points and Shifts for year: YEARC
Taking the chosen monthly or seasonal series In: SAMIN.BAT
Taking the chosen monthly or seasonal series Out: SAMOUT.BAT

MONTHS with MISSING VALUES:  7  8
```

Figure 1. Output of Parametrization (SAMPAR.PAR)

1. Taking Month August In (SAMIN.BAT)

```
TAKING SERIES IN

SEASONAL INDEXES
MONTHS: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
WINTER: 13  SPRING: 14  SUMMER: 15  AUTUMN: 16  YEAR: 17

MONTHS with MISSING VALUES:   7  8
MONTHS without FILLING    :   7  8

Index?
```

8 (for example)

```
EXAMINED STATIONS AND INDEXES

     bp: 1       de: 2       ke: 3       mi: 4       mo: 5       ny: 6
     pe: 7       sr: 8       se: 9       so:10


Information, Parameters: SAMPAR.PAR, MASHPAR.PAR
File of Meta Data: MASHMETA.DAT
Original series (with missing values): MASHMISS.SER
Original series (without missing values): MASHDAT.SER
Homogenized series: MASHHOM.SER
Inhomogeneity series: MASHINH.SER
Break Points and Shifts: MASHCOR.RES
Automatic filling of missing values: MASHMISS.BAT
Automatic correction of outliers: MASHLIER.BAT
GAME of MASH: MASHGAME.BAT
Comparing Test for seasonal series: SAMTEST.BAT
Non-automatic correction: MASHCOR.BAT
Verification of Homogenization: MASHVERI.BAT
Graphics: MASHDRAW.BAT

MISSING VALUES!
THE FIRST STEP: MASHMISS.BAT
```

Figure 2. Partial Output of Program SAMIN.BAT on the Screen

2.The Further Steps

Filling of missing values (MASHMISS.BAT); Correction of outliers (MASHLIER.BAT);
Taking month AUGUST Out (SAMOUT.BAT).

Taking month JULY In (SAMIN.BAT); Filling of missing values (MASHMISS.BAT);
Correction of outliers (MASHLIER.BAT); Taking month JULY Out (SAMOUT.BAT).

Taking month JUNE In (SAMIN.BAT); Correction of outliers (MASHLIER.BAT);
Taking month JUNE Out (SAMOUT.BAT).

3.The Next Steps: Examination of Season SUMMER

There is a possibility for examination of the seasonal series instead of the monthly series.
The monthly inhomogeneities can be corrected by usage of the detected seasonal inhomogeneities, if the monthly inhomogeneities are identical within the given season.

3.1 Taking Season SUMMER In (SAMIN.BAT), Application of Program SAMTEST.BAT

```
HELP: TEST TABLE
 TOTAL COMPARISON of the Inhomogeneities between summer Months and Summer
 Null hypothesis: the monthly and the seasonal inhomogeneities are identical.
 Critical value (significance level 0.05):  50.00
 Test statistics (TS) can be compared to the critical value.
 The larger TS values are more suspicious!

 Station  Index     TS      Station  Index     TS      Station  Index     TS
     bp      1     20.25        de      2     24.37        ke      3     51.12
     mi      4     18.65        mo      5     22.01        ny      6     15.74
     pe      7     21.42        sr      8     20.84        se      9     25.89
     so     10     19.83

 Index of excluded station:  3
```

Figure 3. Output of Test Program SAMTEST.BAT (SAMTEST.RES)

It can be seen that the null hypothesis can be accepted for all the stations with exception of station ke (index 3).

3.2 Homogenization of the SUMMER Series by more running of MASHGAME.BAT

```
ESTIMATED BREAK POINTS AND SHIFTS
  (Mark M: META DATA)
    bp:
  M1909:   .26
    de:
  No Break Points
    ke:
   1927:   .39/   1928:  -.39
    mi:
   1908: -3.74
    mo:
  No Break Points
    ny:
   1901:   .93
    pe:
   1918:   .52/   1921: -1.18
    sr:
  No Break Points
    se:
   1918:  -.76
    so:
   1917:   .25
```

Figure 4. Detected SUMMER Inhomogeneities (MASHCOR.RES)

3.3 Evaluation of the Homogenization of SUMMER Series

```
TEST STATISTICS FOR SERIES INHOMOGENEITY, BEFORE HOMOGENIZATION
Null hypothesis: the examined series are homogeneous.
Critical value (significance level .05):  20.53
Test statistics (TS) can be compared to the critical value.
The larger TS values are more suspicious!
Series  Index     TSB       Series  Index     TSB       Series  Index     TSB
   bp     1     97.59          de     2     16.61          ke     3     41.88
   mi     4   1012.34          mo     5     27.64          ny     6     39.53
   pe     7     82.45          sr     8      8.01          se     9     67.78
   so    10     84.04

TEST STATISTICS FOR SERIES INHOMOGENEITY, AFTER HOMOGENIZATION
Null hypothesis: the examined series are homogeneous.
Critical value (significance level .05):  20.53
Test statistics (TS) can be compared to the critical value.
The larger TS values are more suspicious!
Series  Index     TSA       Series  Index     TSA       Series  Index     TSA
   bp     1     10.49          de     2      7.31          ke     3     19.16
   mi     4     27.96          mo     5     19.10          ny     6     10.23
   pe     7     22.70          sr     8      4.91          se     9     16.18
   so    10     12.21
```

Figure 5. Partial Outputs of MASHVERI.BAT Before and After Homogenization (MASHVERI.RES)

```
TEST STATISTICS FOR EVALUATION OF META DATA
Null hypothesis: the inhomogeneities can be explained by the Meta Data.
Critical value (significance level .05):  20.53
Test statistics (TSM) can be compared to the critical value.
The larger TSM values are more suspicious!
Series  Index     TSM       Series  Index     TSM       Series  Index     TSM
    bp     1      9.18          de     2      20.45          ke     3      38.11
    mi     4    971.28          mo     5      17.86          ny     6      39.53
    pe     7     82.45          sr     8       8.01          se     9      71.13
    so    10     68.18
```

Figure 6. Partial Output of MASHVERI.BAT (MASHVERI.RES)

3.4 Taking Season SUMMER Out (SAMOUT.BAT)

Homogenization of summer (June, July, August) monthly series on the basis of the detected summer inhomogeneities with exception of station ke (index 3) as a result of the Test Program SAMTEST.BAT (see Figure 3)

4. Evaluation of the Homogenization of Monthly Series

4.1 Taking month JUNE In (SAMIN.BAT); Application of MASHVERI.BAT

```
 TEST STATISTICS FOR SERIES INHOMOGENEITY
 Null hypothesis: the examined series are homogeneous.
 Critical value (significance level 0.05):  20.53
 Test statistics (TS) can be compared to the critical value.
 The larger TS values are more suspicious!
 1. Test Statistics After Homogenization
 Series  Index      TSA      Series  Index      TSA      Series  Index      TSA
    bp     1       9.71         de     2       9.98         ke     3      47.66
    mi     4      43.60         mo     5      46.04         ny     6      10.28
    pe     7      13.54         sr     8      14.17         se     9      28.10
    so    10      11.97
 2. Test Statistics Before Homogenization
 Series  Index      TSB      Series  Index      TSB      Series  Index      TSB
    bp     1      24.42         de     2       7.78         ke     3     193.87
    mi     4     633.59         mo     5      43.84         ny     6      27.35
    pe     7      28.89         sr     8      14.92         se     9      88.49
    so    10      53.76
```

Figure 7. Part of Verification Output MASHVERI.RES

4.2 Taking Month JULY In (SAMIN.BAT); Application of MASHVERI.BAT

```
  TEST STATISTICS FOR SERIES INHOMOGENEITY
  Null hypothesis: the examined series are homogeneous.
  Critical value (significance level 0.05):  20.53
  1. Test Statistics After Homogenization
  Series  Index      TSA      Series  Index      TSA      Series  Index      TSA
    bp     1      19.71         de     2      20.36         ke     3      21.23
    mi     4      18.29         mo     5      28.06         ny     6      12.60
    pe     7      23.32         sr     8       6.01         se     9      75.59
    so    10      12.09
  2. Test Statistics Before Homogenization
  Series  Index      TSB      Series  Index      TSB      Series  Index      TSB
    bp     1      55.93         de     2      10.03         ke     3      59.38
    mi     4     764.40         mo     5      45.49         ny     6      53.30
    pe     7      54.73         sr     8       5.99         se     9      78.28
    so    10      50.69
```

Figure 8. Part of Verification Output MASHVERI.RES

4.3 Taking Month AUGUST In (SAMIN.BAT); Application of MASHVERI.BAT

```
TEST STATISTICS FOR SERIES INHOMOGENEITY
 Null hypothesis: the examined series are homogeneous.
 Critical value (significance level 0.05):  20.53
 1. Test Statistics After Homogenization
 Series  Index      TSA      Series  Index      TSA      Series  Index      TSA
    bp     1       8.10         de     2       9.16         ke     3      10.21
    mi     4      27.00         mo     5       9.02         ny     6       8.57
    pe     7      10.67         sr     8      14.93         se     9      12.97
    so    10      15.46
2. Test Statistics Before Homogenization
Series  Index      TSB      Series  Index      TSB      Series  Index      TSB
    bp     1     140.42         de     2      39.04         ke     3       8.82
    mi     4    1935.64         mo     5      14.22         ny     6      23.56
    pe     7      53.83         sr     8      18.30         se     9      26.14
    so    10      32.30
```

Figure 9. Part of Verification Output MASHVERI.RES

5. Verification of Homogenization for SUMMER series (MASHVERI.BAT)

```
   I. TEST STATISTICS FOR SERIES INHOMOGENEITY
 Null hypothesis: the examined series are homogeneous.
 Critical value (significance level 0.05):  20.53
 Test statistics (TS) can be compared to the critical value.
 The larger TS values are more suspicious!

 1. Test Statistics After Homogenization
Series  Index     TSA       Series  Index     TSA       Series  Index     TSA
    mi     4     27.96          pe     7     22.70          ke     3     19.16
    mo     5     19.10          se     9     16.18          so    10     12.21
    bp     1     10.49          ny     6     10.23          de     2      7.31
    sr     8      4.91
 AVERAGE:        15.02
 2. Test Statistics Before Homogenization
Series  Index     TSB       Series  Index     TSB       Series  Index     TSB
    mi     4   1012.34          bp     1     97.59          so    10     84.04
    pe     7     82.45          se     9     67.78          ke     3     41.88
    ny     6     39.53          mo     5     27.64          de     2     16.61
    sr     8      8.01
 AVERAGE:       147.79
 3. Statistics for Estimated Inhomogeneities
(IS statistics can be compared with the TSB ones)
Series  Index     IS        Series  Index     IS        Series  Index     IS
    mi     4    889.88          se     9    128.11          pe     7     42.41
    bp     1     19.34          ny     6     17.29          so    10      9.29
    ke     3      3.43          de     2      0.00          mo     5      0.00
    sr     8      0.00
 AVERAGE:       110.97

  II. CHARACTERIZATION OF INFHOMOGENEITY
 1. Relative Estimated Inhomogeneities
Series  Index     RI1       Series  Index     RI1       Series  Index     RI1
    mi     4      0.98          se     9      0.38          pe     7      0.31
    ny     6      0.18          so    10      0.15          bp     1      0.13
    ke     3      0.08          de     2      0.00          mo     5      0.00
    sr     8      0.00
 AVERAGE:         0.22
 2. Relative Modification of Series
Series  Index     RI2       Series  Index     RI2       Series  Index     RI2
    mi     4      1.14          se     9      0.61          pe     7      0.54
    so    10      0.22          ny     6      0.18          bp     1      0.15
    ke     3      0.08          de     2      0.00          mo     5      0.00
    sr     8      0.00
 AVERAGE:         0.29
 3. Lower Confidence Limit for Relative Residual Inhomogeneities
   (confidence level: 0.95)
Series  Index     RI3       Series  Index     RI3       Series  Index     RI3
    mi     4      0.03          pe     7      0.02          bp     1      0.00
    de     2      0.00          ke     3      0.00          mo     5      0.00
    ny     6      0.00          sr     8      0.00          se     9      0.00
    so    10      0.00
 AVERAGE:         0.00

 III. REPRESENTATIVITY OF STATION NETWORK
(1-relative interpolation error)
Series  Index      RS       Series  Index      RS       Series  Index      RS
    sr     8      0.64          mi     4      0.66          ke     3      0.69
    ny     6      0.74          pe     7      0.75          so    10      0.75
    de     2      0.79          mo     5      0.79          se     9      0.79
    bp     1      0.84
 AVERAGE:         0.74
```

Figure 10.a, Verification Results at the actual stage of Homogenization (MASHVERO.RES, ordered statistics)

```
                             EVALUATION OF META DATA


 IV. TEST STATISTICS
 Null hypothesis: the inhomogeneities can be explained by the Meta Data.
 Critical value (significance level 0.05):  20.53
 Test statistics (TSM) can be compared to the critical value.
 The larger TSM values are more suspicious!
Series  Index     TSM       Series  Index     TSM       Series  Index     TSM
    mi     4    971.28          pe     7     82.45          se     9     70.22
    so    10     68.18          ke     3     41.88          ny     6     39.53
    mo     5     17.86          de     2     16.12          bp     1     10.53
    sr     8      8.01
 AVERAGE:    132.60


  V. REPRESENTATIVITY OF META DATA
(Relative part of estimated inhomogeneity can be explained by the Meta Data)
Series  Index      RM       Series  Index      RM       Series  Index      RM
    so    10      0.00          se     9      0.00          pe     7      0.00
    ny     6      0.00          ke     3      0.00          mi     4      0.07
    sr     8      1.00          mo     5      1.00          de     2      1.00
    bp     1      1.00
 AVERAGE:      0.41
```

Figure 10.b, Verification Results for Meta Data at the actual stage of Homogenization (MASHVERO.RES, ordered statistics)

VII. HOMOGENIZATION OF DAILY DATA

MATHEMATICAL BASIS (draft version)

Only the additive model is presented that is appropriate for temperature, pressure etc. elements.

Relation of daily and monthly homogenization

<u>Alternative possibilities</u>
– To use the detected monthly inhomogeneities directly for daily data homogenization
– Direct methods for daily data homogenization
<u>Problems</u>
– The direct use of the detected monthly inhomogeneities is probably not sufficient.
– Direct methods for daily data homogenization is probably not enough efficient thinking of the larger variability (less signal to noise ratio).
<u>The Question</u>
How can we use the valuable information of detected monthly inhomogeneities for daily data homogenization?

Additive model for daily values (e.g. temperature)

$$X^{st}(y,m,d)=\mu(y,m,d)+\mu_0^{st}(m,d)+IH^{st}(y,m,d)+\varepsilon^{st}(y,m,d)$$

μ : climate change signal, μ_0 : spatial expected value,

IH : inhomogeneity , ε : normal noise

st : station, m : month, y : year, d : day

Additive model for monthly means

$X_m^{st}(y)=\mu_m(y)+\mu_{0m}^{st}+IH_m^{st}(y)+\varepsilon_m^{st}(y)$

$IH_m^{st}(y)=\overline{IH^{st}}(y,m)$: inhomogeneity (break points and shifts)

<u>We have:</u> estimated monthly inhomogeneities: $I\hat{H}_m^{st}(y)$

Valuable information! : $\overline{IH^{st}}(y,m)=IH_m^{st}(y)$

But maybe a problem of direct use: the smoothness

<u>Question:</u>

Smooth estimation $I\hat{H}^{st}(y,m,d)$ for daily inhomogeneities

by using the estimated monthly inhomogeneities $I\hat{H}_m^{st}(y)$?

Possible condition for daily estimation $I\hat{H}^{st}(y,m,d)$:

Smoothness and condition for mean: $\overline{I\hat{H}^{st}}(y,m)=I\hat{H}_m^{st}(y)$

<u>Maybe a problem:</u> too strong inhomogeneities can be obtained.

Other train of thought

Not to forget: the monthly values $I\hat{H}_m^{st}(y)$ are only estimations,

stochastic variables. To know the real $IH_m^{st}(y)$ is impossible.

Consequently the monthly estimations $I\hat{H}_m^{st}(y)$ may be modified.
But the modification must be controlled.

<u>The Essence of Procedure</u>

i, Smooth estimation for daily inhomogeneities with
a not too strong condition e.g.: $\exists d_0 : I\hat{H}^{st}(y,m,d_0) = I\hat{H}_m^{st}(y)$

ii, Test of hypothesis to control the new monthly estimations:
$I\tilde{H}_m^{st}(y) := \overline{I\hat{H}^{st}(y,m)}$

The MASH Procedure for Daily Data

1. Monthly means $X_m^{st}(y)$ from daily data $X^{st}(y,m,d)$.
2. MASH homogenization procedure for monthly series $X_m^{st}(y)$, estimation of monthly inhomogeneities: $I\hat{H}_m^{st}(y)$
3. On the basis of estimated monthly inhomogeneities $I\hat{H}_m^{st}(y)$, smooth estimation for daily inhomogeneities: $I\hat{H}^{st}(y,m,d)$.
4. Homogenization of daily data:
$\tilde{X}^{st}(y,m,d) = X^{st}(y,m,d) - I\hat{H}^{st}(y,m,d)$.
5. Qulity Control for homogenized daily data $\tilde{X}^{st}(y,m,d)$.
6. Missing daily data completion.
7. Monthly means $\tilde{X}_m^{st}(y)$ from homogenized, controlled, completed daily data $\tilde{X}^{st}(y,m,d)$.
8. Test of homogeneity for the new monthly series $\tilde{X}_m^{st}(y)$ by MASH.
Repeating steps 2-8 with $\tilde{X}_m^{st}(y)$, $\tilde{X}^{st}(y,m,d)$ if it is necessary.

Interpolation technique for QC and Data Completion

<u>Daily data for a given month:</u>

$X_j(t) \in N(E_j(t), D_j(t))$ $\quad (j = 1,..,M$ station; $t = 1,...,30)$

<u>Candidate data:</u> $X_j(t)$ <u>Reference data:</u> $X_i(t)\,(i \neq j)$

Interpolation: $\hat{X}_j(t) = w_{j0}(t) + \sum_{i \neq j} w_{ji}(t) X_i(t)$ where $\sum_{i \neq j} w_{ji}(t) = 1$.

RMS Error and Representativity: $RMSE_j(t)$, $REP_j(t) = 1 - \frac{RMSE_j(t)}{D_j(t)}$

The Optimum Interpolation Parameters $w_{j0}^{opt}(t)$, $w_{ji}^{opt}(t)\,(i \neq j; t = 1,..,30)$ minimizing $RMSE_j(t)$, are uniquely determined by the expectations, st. deviations and the correlations.

Problem: Estimation of daily statistical parameters.

Assumptions:

i, $E_j(t) - E_i(t) = e_{ji}$, $D_j(t)/D_i(t) = d_{ji}$, $(i \neq j; t = 1,..,30)$

ii, $\mathrm{corr}(X_{j_1}(t_1), X_{j_2}(t_2)) = r^S_{j_1 j_2} \cdot r^T_{t_1 t_2}$ $(j_1, j_2 = 1,...M; t_1, t_2 = 1,..,30)$

$r^S_{j_1 j_2}$: correlation structure in space, $r^T_{t_1 t_2}$: correlation structure in time

$\Leftrightarrow$ Partial corr.: $\mathrm{corr}_{X_{j_1}(t_2)}(X_{j_1}(t_1), X_{j_2}(t_2)) = \mathrm{corr}_{X_{j_2}(t_1)}(X_{j_1}(t_1), X_{j_2}(t_2)) = 0$

Statement: If the assumptions i, ii, are fulfilled then

$$w_{j0}^{opt}(t) \equiv w_{j0}^{opt},\ w_{ji}^{opt}(t) \equiv w_{ji}^{opt},\ REP_j^{opt}(t) \equiv REP_j^{opt} \quad (t = 1,..,30) ,$$

where $w_{j0}^{opt}, w_{ji}^{opt}, REP_j^{opt}$ are the optimal parameters of monthly interpolation: $\hat{\bar{X}}_j(t) = w_{j0} + \sum_{i \neq j} w_{ji} \bar{X}_i$ where $\sum_{i \neq j} w_{ji} = 1$.

Consequence

The monthly statistical parameters can be used for daily interpolation.

i, Data Completion: $\hat{X}_j(t) = w_{j0}^{opt} + \sum_{i \neq j} w_{ji}^{opt} X_i(t)$

ii, Quality Cotrol can be based on the standardized error:

$$Z_j(t) = \frac{X_j(t) - \hat{X}_j(t)}{D_j(t)\left(1 - REP_j^{opt}\right)} \in N(0,1)$$

where $w_{j0}^{opt}, w_{ji}^{opt}, REP_j^{opt}$ are the optimal parameters of monthly interpolation, and $D_j(t)$ is the daily standard deviation.

Test of Hypothesis of the standardized error series $Z(t)\ (t=1,..,n)$

If the data have good quality then $Z(t) \in N(0,1)$ $(\ t=1,..,n)$.

<u>But Problem:</u> $\mathrm{P}\left(\ \max_t |Z(t)| < z\ \right)$ depends on the autocorrelation.

<u>Statement:</u>

i, If $Z(t)\ (\ t=1,..,n)$ is a Markov process, furthemore

ii, and $\mathrm{P}\left(\ |Z(t)| < z\ \middle|\ |Z(t-1)| < z\ \right) \geq \mathrm{P}\left(\ |Z(t)| < z\ \right)$ $\quad (t=2,..,n)$,

then $\mathrm{P}\left(\ \max_t |Z(t)| < z\ \right) \geq \prod_{t=1}^{n} \mathrm{P}\left(\ |Z(t)| < z\ \right)$.

<u>Example:</u>

If $Z(t)\ (\ t=1,..,n)$ is a normal AR(1) process then i, ii, are fulfilled.

Decision according to test of hypothesis

<u>We have wrong data:</u>

If $|Z(t)| > z_p$ where critical value z_p is defined by

the significance level p (e.g.: p=0.01) as,

$\mathrm{P}\left(\ \max_t |Z(t)| < z_p\ \right) \geq \left(\ 2\Phi(z_p) - 1\ \right)^n = 1 - p$,

$\Phi(z)$: standard normal distribution function.

Multiple QC for daily data

More standardized error series are examined without common reference series to separate the wrong data for the candidate station. Correction of the wrong data is based on confidence intervals.

THE STRUCTURE OF MASHDAILY PROGRAM SYSTEM

Main Directory **MASHv3.03**:

Directory **MASHDAILY:**

- Subdirectory **COSTHOMEINPUT**
- Subdirectory **MASHDAMO:**
  - MASHDAMO.BAT
  - Subdirectory **MASHFORMAT**
  - Subdirectory **MASHDAMOSUB**
    (do not use it including "subroutines")
- Subdirectory **MASHDAILY:**
  - Subdirectory **MASHDPAR**
    - Parametrization program: MASHDPAR.BAT
  - Main Program: MASHD.BAT
  - Subdirectory **MASHDMANUQC** (“manual” programs)
  - Subdirectory **MASHDMISSING** (“manual” programs)
  - Subdirectory **MASHDSUB**
    (do not use it including "subroutines")

Directory **MASHMONTHLY** (See Page 21)

MASHDAILY IN PRACTICE

I. Monthly Data from Daily Data in Subdirectory MASHDAILY\ MASHDAMO

MASHDAMO.BAT (see page 57)

II. Homogenization of Monthly Series in Directory MASHMONTHLY\SAM

MASH homogenization procedure for monthly series, estimation of monthly inhomogeneities.
Input Files from **MASHDAMO\MASHFORMAT: M{j} (j=1,....,12), FILASTAT.PAR** (see p. 58; Copy batch File in Subdirectory MASHFORMAT: COPYSAMPAR.BAT)

III. Homogenization of Daily Data in Subdirectory MASHDAILY\ MASHDAILY

1. Parametrization in Subdirectory MASHDAILY\ MASHDAILY\MASHDPAR:
MASHDPAR.BAT (see pages 57-58)

2. Homogenization of Daily Data, Automatic Qulity Control for Homogenized Daily Data, Missing Daily Data Completion in Subdirectory MASHDAILY\ MASHDAILY:
MASHD.BAT (see pages 58-59)

3. Possibility for Manual Program Procedures in MASHDAILY\ MASHDAILY (p. 59):
Quality Control in MASHDMANUQC; Missing data completion in MASHDMISSING

THE MAIN PROGRAM and I/O FILES in Subdirectory MASHDAMO

EXECUTIVE FILE: MASHDAMO.BAT

INPUT:
Original Daily Data: **DAILY.DAT**
Maximal number of stations: 500
Maximal number of years: 200
Format of Data File:
row 1: names of stations (obligatory!), Format: (2x,character*6)?
column 1: date of year (I4)
column 2: month (I2)
column 3: day (I2)
column i+3: series i.
Data Format: F8.2
Mark of Missing Values: 9999.99

File of Filambda Station Coordinates: **FILASTAT.PAR** (see page 60)

Model: additive or multiplicative

RESULT OUTPUT FILES in MASHDAMO\MASHFORMAT:
(Input Files of MASHMONTHLY\SAM)
Files of Monthly Series: **M{J} (J=1,..,12)**
File of Filambda Station Coordinates: **FILASTAT.PAR**

RESULT OUTPUT FILES in MASHDAILY\MASHDPAR:
Original Daily Data: **DAILY.DAT**
Files of Monthly Series: **M{J} (J=1,..,12)**
File of Filambda Station Coordinates: **FILASTAT.PAR**
Parameter File: **MASHDPAR.PAR**

THE MAIN PROGRAM and I/O FILES of Subdirectory MASHDAILY\ MASHDAILY\MASHDPAR

EXECUTIVE BATCH FILE: MASHDPAR.BAT

THE STEPS OF MASHDPAR.BAT:
MASHDTRAN.EXE & MASHDTOP.EXE& MASHDTEXT.EXE

INPUT DATA FILES:
Result Files from **MASHMONTHLY\SAM\SAMEND**:
Homogenized Monthly Series **M{j}h (j=1,....,12)**
Monthly Inhomogeneities **M{j}i (j=1,....,12)**
(Copy batch File in Subdirectory SAMEND: COPYMASHDPAR)

INPUT DATA FILES DAILY.DAT, M{J} (J=1,..,12), MASHDPAR.PAR, FILASTAT.PAR are written in by MASHDAMO.BAT

OUTPUT FILES written in Directory MASHDAILY\MASHDAILY:
MASHDPAR.PAR, FILASTAT.PAR, REFERENCE.PAR, DAILY.DAT, M{j} (j=1,....,12), M{j}h (j=1,....,12), M{j}i (j=1,....,12)
Work and Parameter Files: DMP{j}.PAR (j=1,....,6), M{j}h.tr (j=1,....,12)

THE MAIN PROGRAM and I/O FILES of Subdirectory MASHDAILY\ MASHDAILY

EXECUTIVE BATCH FILE: MASHD.BAT

INPUT DATA FILES are written in by MASHDPAR.BAT

RESULT OUTPUT FILES:

Homogenized, Controlled and Completed daily Data: **DAILYHOMQC.DAT**

Homogenized and Completed daily Data (without QC): **DAILYHOM.DAT**

Daily Inhomogeneities: **DAILYINHOM.DAT**

Result of Quality Control: **ERROR.RES**

THE STEPS OF MASHD.BAT:

1. STAT1.EXE:
Input: MASHDPAR.PAR, M{j}h (j=1,....,12), FILASTAT.PAR
Output: STAT1{j}.PAR (j=1,....,12)

2. DMINHOM1.EXE:
Input: MASHDPAR.PAR, M{j} (j=1,....,12), M{j}i (j=1,....,12)
Output: DMSTAT.RES, DM{j}i.PAR (j=1,....,12)

3. DMINHOM2.EXE:
Input: MASHDPAR.PAR, M{j} (j=1,....,12), M{j}i (j=1,....,12)
Output: DM{j}i (j=1,....,12)

4. DMINHOM3.EXE:
Input: MASHDPAR.PAR, DM{j}i (j=1,....,12)
Output: DM{j}i (j=1,....,12) (Daily inhomogeneities in 12 files)

5. DMDATA.EXE:
Input: MASHDPAR.PAR, DAILY.DAT
Output: DM{j} (j=1,....,12) (Original daily data in 12 files)

6. DMINHCORR.EXE:
Input: MASHDPAR.PAR, DM{j}i.PAR (j=1,....,12), DM{j}i (j=1,....,12), DM{j} (j=1,....,12)
Output: DM{j}h (j=1,....,12) (Homogenized daily data in 12 files), DM{j}d (j=1,....,12) (Daily st. deviations in 12 files)

7. QC.EXE

Input: MASHDPAR.PAR, REFERENCE.PAR, STAT1{j}.PAR (j=1,....,12),
DM{j}i (j=1,....,12), DM{j}h (j=1,....,12), DM{j}d (j=1,....,12),
DM{j} (j=1,....,12), M{j}i (j=1,....,12)
Output: DM{j}hc (j=1,....,12) (Homogenized, controlled daily data in 12 files)

ERROR.RES

Work and Parameter Files: QC1.PAR, QC2.PAR, ERR{j} (j=1,....,12), ERROR.PAR

8. MISSING.EXE:
Input: MASHDPAR.PAR, FILASTAT.PAR,
DM{j}hc (j=1,....,12), STAT1{j}.PAR (i=1,..12)
Output: DM{j}hcm (j=1,....,12) (Homog., controlled, completed daily data in 12 files)
Work and Parameter Files: MISSING1.PAR, MISSING2.PAR

9. DAILYEND.EXE:
Input: MASHDPAR.PAR, DM{j}hcm (j=1,....,12), DM{j}i (j=1,....,12)
Output: **DAILYHOMQC.DAT, DAILYHOM.DAT, DAILYINHOM.DAT**

MANUAL PROGRAM PROCEDURES in Subdirectory MASHDAILY\ MASHDAILY

1. Quality Control programs in subdirectory MASHDMANUQC

1.1 Selection of errors resulted by automatic QC

ERRORSELECT.BAT: Selection Procedure

Input: **ERROR.RES** being result of automatic QC, and a critical value for errors

Output: **ERRORSELECT.RES**

1.2 Correction of the homogenized daily data series

DAILYHOMQCM.BAT: Manual Correction of **DAILYHOM.DAT**

Input: **DAILYHOM.DAT** and
optionally **ERRORSELECT.RES** or typing errors

Output: **DAILYHOMQCM.BAT**

2. Missing data completion in subdirectory MASHDMISSING

MASHDMISS.BAT: Missing Data Completion Procedure

Input File: **DAILYMISS.DAT** with missing values (9999.99)

Output File: **DAILYCOMPL.DAT** with completed series

Remark: It is also a possibility to modify the result of automatic QC.

EXAMPLE FOR APPLICATION OF MASHDAILY SYSTEM

Examined Data: DAILY.DAT

Daily temperature series (1901-1930), 10 Stations in Hungary.

Temperature element: (max+min)/2

Model: additive

File of Filambda Station Coordinates: FILASTAT.PAR

```
index   lambda(x)       fi(y)
   1   18.65308760   47.18486400
   2   20.62684630   46.57159040
   3   17.09815790   47.05921170
   4   20.14969640   47.21730420
   5   17.39957430   47.90058140
   6   20.80181500   46.85120390
   7   16.83021930   47.91734310
   8   19.18956180   45.94506840
   9   16.73989490   47.59620290
  10   19.86032100   46.45614240
```

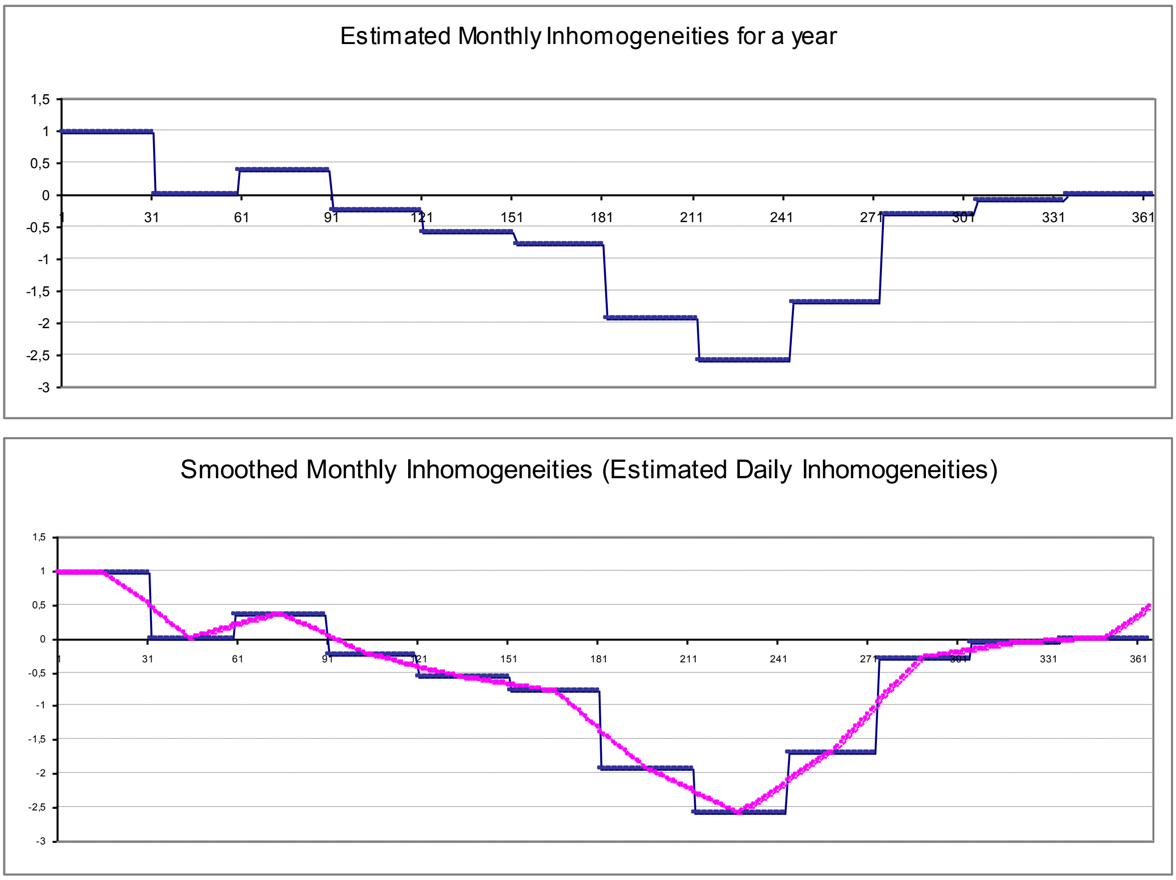


Figure 1. Example for smoothing of Monthly Ihomogeneities

```
Detected errors in September 1903 at Station 10 (ERROR.RES)
          st1  st2  st3  st4  st5  st6  st7  st8  st9 st10
1903 9 1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -3.4
1903 9 2  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -2.2
1903 9 3  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -3.1
1903 9 4  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -5.0
1903 9 5  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -2.6
1903 9 6  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -2.7
1903 9 7  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -4.9
1903 9 8  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -2.9
1903 910  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -1.8
1903 911  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 -5.5
Original Data
1903 9 1 20.5 17.1 20.0 17.5 21.0 18.1 18.7 19.5 19.3 12.3
1903 9 2 19.8 17.9 20.8 15.5 21.5 14.9 18.4 19.3 20.2 13.2
1903 9 3 19.1 17.3 20.8 15.5 21.3 15.8 18.5 17.2 17.5 11.3
1903 9 4 19.7 17.5 19.8 15.3 19.0 15.8 18.8 19.2 17.2 10.4
1903 9 5 20.3 17.8 20.5 16.0 21.0 17.4 19.3 20.4 17.8 13.2
1903 9 6 20.9 18.7 21.3 17.3 20.0 18.6 19.9 21.4 18.8 13.8
1903 9 7 22.9 21.5 22.5 17.8 22.0 19.5 18.9 23.6 19.0 13.9
1903 9 8 22.5 20.9 25.0 19.0 23.0 19.8 19.1 23.5 19.5 15.5
1903 910 17.7 18.4 17.0 13.8 13.6 19.0 14.3 18.9 13.7 12.7
1903 911 16.5 13.7 18.3 11.8 14.5 13.5 13.1 18.8 14.1  6.2
Longterm means in September
         16.7 15.9 16.2 14.9 15.9 15.5 14.6 17.0 14.7 16.6
```

Figure 2. Part Results of Quality Control

```
TEST STATISTICS for ANNUAL SERIES (OUTPUT of MASH)
Critical value (significance level 0.05): 20.53

 1. Test Statistics Before Monthly Homogenization
Station       TSBM     Station       TSBM     Station       TSBM
      4     317.85           6     241.41           2     155.04
      9     127.66           7      91.66          10      68.36
      1      62.55           8      61.84           5      42.06
      3      15.82     AVERAGE:    118.42

 2. Test Statistics After Monthly Homogenization
Station       TSAM     Station       TSAM     Station       TSAM
      7      28.64           5      25.11           9      22.73
      4      18.52           1      18.12           8      15.26
      6      14.96           2      14.82          10      12.41
      3      10.26     AVERAGE:     18.08

 3. Test Statistics After Monthly&Daily Homogenization
Station      TSAMD     Station      TSAMD     Station      TSAMD
      7      28.89           5      25.40           2      25.06
      9      21.98           1      17.60           4      16.52
      8      15.23           6      14.66           3       9.69
     10       9.00     AVERAGE:     18.40
```

Figure 3. Verification Results for the Annual Series (MASHVERI.RES)

```
AVERAGED TEST STATISTICS FOR MONTHLY SERIES (10 Stations)
Average of Test Statistics Before Monthly Homogenization: TSBM
Average of Test Statistics After Monthly Homogenization: TSAM
Average of Test Statistics After Monthly&Daily Homogenization: TSAMD

MONTH       TSBM        TSAM        TSAMD
    1       28.5        12.0         12.1
    2       21.1        16.6         17.0
    3       41.2        24.0         22.4
    4       73.7        17.5         17.8
    5       82.1        15.7         13.4
    6      100.7        14.7         12.5
    7       84.5        16.1         14.2
    8       61.7        16.0         14.3
    9      131.4        12.9         13.1
   10       56.3        14.6         16.0
   11       38.9        10.4         11.2
   12       34.5        18.7         20.4
   SP       90.6        19.9         20.2
   SU       92.6        18.7         17.2
   AU      101.3        17.1         19.6
   WI       32.1        18.3         16.6
    Y      118.4        18.1         18.4

 Critical value (significance level 0.05): 20.53
 Test statistics (TS) can be compared to the critical value.
```

Figure 4. Average of Verification Results for the Monthly Series

References

Szentimrey, T., 1994: "Statistical problems connected with the homogenization of climatic time series", Proceedings of the European Workshop on Climate Variations, Kirkkonummi, Finland, Publications of the Academy of Finland, 3/94, pp. 330-339.

Szentimrey, T., 1995: "Statistical methods for detection of inhomogeneities", Proceedings of the Regional Workshop on Climate Variability and Climate Change Vulnerability and Adaptation, Prague, pp. 293-298.

Szentimrey, T., 1995: "General problems of the estimation of inhomogeneities, optimal weighting of the reference stations", Proceedings of the 6h International Meeting on Statistical Climatology, Galway, Ireland, pp. 629-631.

Szentimrey, T., 1996: "Some statistical problems of homogenization: break points detection, weighting of reference series", Proceedings of the 13th Conference on Probability and Statistics in the Atmospheric Sciences, San Francisco, California, pp. 365-368.

Szentimrey, T., 1997: "Statistical procedure for joint homogenization of climatic time series", Proceedings of the Seminar for Homogenization of Surface Climatological Data, Budapest, Hungary, pp. 47-62.

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., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E.J., Hanssen-Bauer, I., Alexanderson, H., Jones, P. and Parker D., 1998: “Homogeneity adjustments of *in situ* atmospheric climata data: a review”, International Journal of Climatology, 18: 1493-1517

Szentimrey, T., 1998: ”MASHv1.03”, Guide for Software Package, Hungarian Meteorological Service, Budapest, Hungary, p. 25.

Auer, I., Böhm, R., 1998: “Endbericht des Projects ALOCLIM, Teil I-II”, Zentralanstalt für Meteorologie und Geodynamik, Wien**.**

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, pp. 27-46.

Szentimrey, T., 2000: “Multiple Analysis of Series for Homogenization (MASH). Seasonal application of MASH (SAM), Automatic using of Meta Data”, Proceedings of the Third Seminar for Homogenization of Surface Climatological Data, Budapest, Hungary,
Home page:http://omsz.met.hu/ismeretterjesztes/rendezvenyek/rendezveny_hu.html

Auer, I., Böhm, R,. Schöner, W., 2001: “Austrian Long-Term Climate (ALOCLIM) 1767-2000 , Multiple instrumental climate time series from Central Europe”, Österreichische Beiträge zu Meteorologie und Geophysik, Heft 25, Central Institute for Meteorology and Geodynamics, Vienna.

Szentimrey, T., 2002: “Statistical problems connected with the spatial interpolation of climatic time series.”,
Home page:http://www.knmi.nl/samenw/cost719/documents/Szentimrey.pdf

Szentimrey, T., 2003: “Homogenization software MASHv2.03”,
Home page:http://www.wmo.ch/web/wcp/clips2001/html/MASH_software.htm

Szentimrey, T., 2004: “Something like an Introduction”, Proceedings of the Fourth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary; WMO, WCDMP-No. 56, pp. 5-16.

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, pp. 193-201.

Szentimrey, T., Bihari, Z., 2007: “Mathematical background of the spatial interpolation methods and the software MISH (Meteorological Interpolation based on Surface Homogenized Data Basis)”, Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, Hungary, 2004, COST Action 719, COST Office, 2007, pp. 17-27

Szentimrey, T., 2007: “Manual of homogenization software MASHv3.02”, Hungarian Meteorological Service, p. 61.

Szentimrey, T., 2007: “Manual of interpolation software MISHv1.02”, Hungarian Meteorological Service, p. 32

Szentimrey, T., 2008: “An overview on the main methodological questions of homogenization”, Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary, 2006; WCDMP-No. 71, WMO/TD-NO. 1493, 2008, pp. 1-6.

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. 71, WMO/TD-NO. 1493, 2008, pp. 123-130.

Szentimrey, T., Bihari, Z., Lakatos, M., 2010: “Quality control procedures in MISH-MASH systems”, European Conference on Applied Climatology (ECAC), 13-17 September 2010, Zürich, Switzerland

Szentimrey, T., Lakatos, M., Bihari, Z., 2010: “Methodological questions of data series comparison for homogenization”, 11th International Meeting of Statistical Climatology, 12-16 July 2010, Edinburgh, Scotland

Szentimrey, T., 2011: “Methodological questions of series comparison”, Proceedings of COST-ES0601 (HOME) Action Management Committee and Working Groups and Sixth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, 26-30 May 2008.
WCDMP-No. 76, WMO/TD-NO. 1576, 2011, pp. 1-7.

Lakatos, M., Szentimrey, T., Bihari, Z., Szalai, S, 2011: “Homogenization of daily data series for extreme climate indices calculation, Proceedings of COST-ES0601 (HOME) Action Management Committee and Working Groups and Sixth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, 26-30 May 2008. WCDMP-No. 76, WMO/TD-NO. 1576, 2011, pp. 100-109.

PREFACE of Version MISHv1.02

The present version MISHv1.02 is a continued development of the first version MISHv1.01. The new parts built in the interpolation subsystem of MISH procedure are as follows:

– Data series complementing that is missing value interpolation, completion for monthly or daily station data series.

– Interpolation, gridding of monthly or daily station data series for given predictand locations. In case of gridding the predictand locations are the nodes of a relatively dense grid.

The potential interpolation area was also increased. At the present version the maximum number of the rows of the half minutes grid that covers the interpolation area is 600 instead of the earlier 400 ones. This means area with 150 000-300 000 km^2 in Europe.

PREFACE of Version MISHv1.01

The MISH method for the spatial interpolation of surface meteorological elements was developed at the Hungarian Meteorological Service. This is a meteorological system not only in respect of the aim but in respect of the tools as well. It means that using all the valuable meteorological information – climate and supplementary model or background information – is intended. For that purpose developing an adequate mathematical background was also necessary of course.

In the practice many kinds of interpolation methods exist therefore the question is the difference between them. According to the interpolation problem the unknown predictand value is estimated by use of the known predictor values. The type of the adequate interpolation formula depends on the probability distribution of the meteorological elements! Additive formula is appropriate for normal distribution (e.g. temperature) while some multiplicative formula can be applied for quasi lognormal distribution (e.g. precipitation). The expected interpolation error depends on certain interpolation parameters as for example the weighting factors. The optimum interpolation parameters minimize the expected interpolation error and these parameters are certain known functions of different climate statistical parameters e.g. expectations, deviations and correlations. Consequently the modelling of the climate statistical parameters is a key issue to the interpolation of meteorological elements.

The various geostatistical kriging methods applied in GIS are also based on the above mathematical theory. However these methods use only a single realization in time for modelling of the necessary statistical parameters that is neglecting the long data series which series form a sample in time and space alike. The long data series is such a speciality of the meteorology that makes possible to model efficiently the climate statistical parameters in question!

The MISH method has been developed according to the above basic principles. The main steps of the interpolation procedure are as follows.

- To model the climate statistical parameters by using long homogenized data series.
- To calculate the modelled optimum interpolation parameters which are certain known functions of the modelled climate statistical parameters.
- To substitute the modelled optimum interpolation parameters and the predictor values into the interpolation formula.

The software MISH consists of two units that are the modelling and the interpolation systems. The interpolation system can be operated on the results of the modelling system.

Modelling System for climate statistical (deterministic and stochastic) parameters:

- Based on long homogenized monthly series and supplementary model variables. The deterministic model variables may be as height, topography, distance from the sea etc..
- Benchmark study, cross-validation test for representativity.
- Modelling procedure must be executed only once before the interpolation applications!

Interpolation System:

- Additive (e.g. temperature) or multiplicative (e.g. precipitation) model and interpolation formula can be used depending on the climate elements.
- Daily, monthly values and many years' means can be interpolated.
- Few predictors are also sufficient for the interpolation and no problem if the greater part of daily precipitation predictors is equal to 0.
- The representativity values are modelled too.
- Capability for application of supplementary background information (stochastic variables) e.g. satellite, radar, forecast data.

I. MATHEMATICAL BACKGROUND[1]

1. INTRODUCTION

The MISH method was developed at the Hungarian Meteorological Service for the spatial interpolation of surface meteorological elements. This is a meteorological system not only in respect of the aim but in respect of the tools as well. It means that using all the valuable meteorological information – e.g. climate and possible background information – is required. For that purpose an adequate mathematical background is also necessary of course.

2. SURFACE METEOROLOGICAL INFORMATION

The two basic types of information for the surface meteorological values are data measured at the observation stations and certain background information given at the nodes of a relatively dense grid. Fig. 1. is an illustration of the different kinds of utilized information.

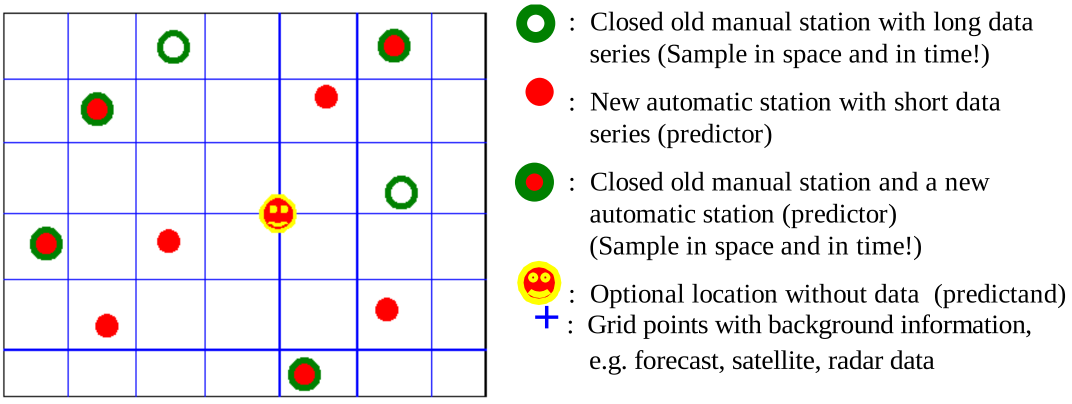

: Closed old manual station with long data series (Sample in space and in time!)
: New automatic station with short data series (predictor)
: Closed old manual station and a new automatic station (predictor) (Sample in space and in time!)
: Optional location without data (predictand)
: Grid points with background information, e.g. forecast, satellite, radar data

Figure 1. Types of information for the surface meteorological values

The long data series can be considered as a sample in space and time for the climate and this sample implies valuable information for the interpolation as well.

[1] Szentimrey, T., Bihari, Z., 2007: „Mathematical background of the spatial interpolation methods and the software MISH (Meteorological Interpolation based on Surface Homogenized Data Basis)", Proceedings of the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, Hungary, COST Action 719, COST Office, 2007, pp. 17-27

3. SPATIAL INTERPOLATION METHODS

In practice many kinds of interpolation methods exist therefore the question is the difference between them. According to the interpolation problem the unknown predictand $Z(\mathbf{s}_0,t)$ is estimated by use of the known predictors $Z(\mathbf{s}_i,t)$ $(i=1,\dots,M)$ where the location vectors $\mathbf{s}$ are the elements of the given space domain D and t is the time. The type of the adequate interpolation formula depends on the probability distribution of the meteorological element.

3.1 Additive Interpolation Formula

Assuming normal distribution (e.g. temperature) the additive formula is adequate, that is, the estimate may be written as

$$\hat{Z}(\mathbf{s}_0,t)=w_0+\sum_{i=1}^{M} w_i\cdot Z(\mathbf{s}_i,t) \quad \text{where} \quad \sum_{i=1}^{M} w_i=1,\ w_i\ge 0 \quad (i=1,\dots,M), \qquad (1)$$

and w_0, w_i $(i=1,\dots,M)$ are the interpolation parameters.

Root Mean Square Interpolation Error and Representativity:

$$ERR(\mathbf{s}_0)=\sqrt{\mathrm{E}\left(\left(Z(\mathbf{s}_0,t)-\hat{Z}(\mathbf{s}_0,t)\right)^2\right)}\quad,\quad REP(\mathbf{s}_0)=1-\frac{ERR(\mathbf{s}_0)}{D(\mathbf{s}_0)},$$

where E is the expectation and $D(\mathbf{s}_0)$ is the standard deviation of the predictand.

The local statistical parameters (expectations, standard deviations) and the stochastic connections (correlations), which are climate statistical parameters in meteorology, uniquely determine the optimum interpolation parameters that minimize the interpolation error. The various geostatistical kriging methods applied in GIS are also based on the above theory. However, these methods use only a single realisation in time for modelling statistical parameters and neglect the long data series which form a sample in time and space as well, while the sample makes it possible to model the climate statistical parameters in question.

3.2 Multiplicative Interpolation Formula

Assuming quasi-lognormal distribution (e.g. precipitation sum) the multiplicative formula is adequate, that is, the estimate may be written as

$$\hat{Z}(\mathbf{s}_0,t)=\vartheta\cdot\left(\prod_{q_i\cdot Z(\mathbf{s}_i,t)\ge\vartheta}\left(\frac{q_i\cdot Z(\mathbf{s}_i,t)}{\vartheta}\right)^{w_i}\right)\cdot\left(\sum_{q_i\cdot Z(\mathbf{s}_i,t)\ge\vartheta} w_i+\sum_{q_i\cdot Z(\mathbf{s}_i,t)<\vartheta} w_i\cdot\left(\frac{q_i\cdot Z(\mathbf{s}_i,t)}{\vartheta}\right)\right)$$

where $\vartheta>0$, $q_i>0$, $\sum_{i=1}^{M} w_i=1$ and $w_i\ge 0$ $(i=1,\dots,M)$,

and q_i, w_i $(i=1,\dots,M)$ are the interpolation parameters.

Similarly to the additive case above the optimum interpolation parameters are uniquely determined by certain climate statistical parameters such as some local statistical parameters and stochastic connections.

4. POSSIBLE CONNECTION OF DIFFERENT TOPICS AND SYSTEMS

As we have seen modelling of the climate statistical parameters is a key issue to the interpolation of meteorological elements and that modelling can be based on the long data series. Before detailing the problem of modelling and interpolation we present a block diagram to illustrate the possible connection between various important meteorological topics.

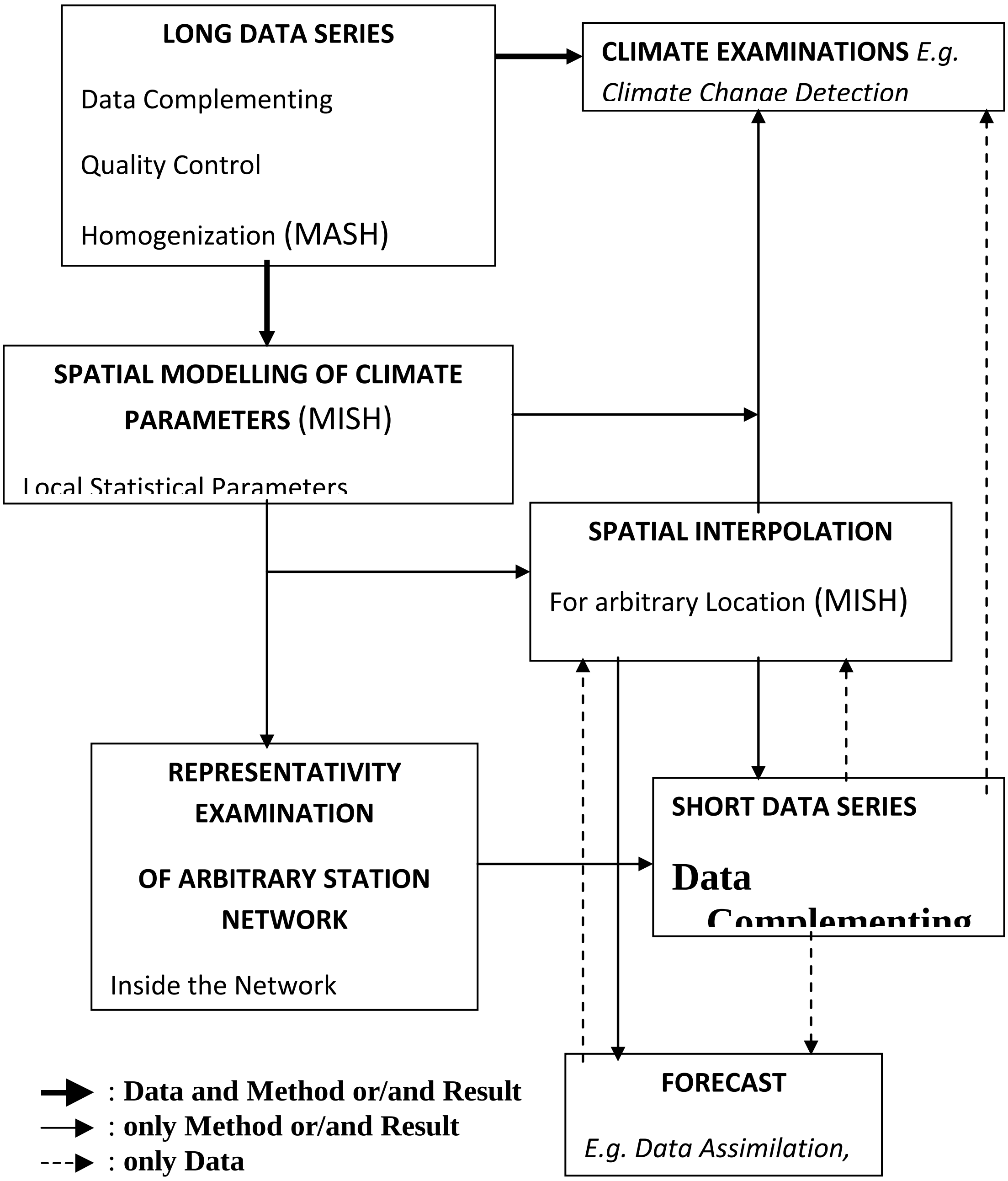

LONG DATA SERIES
Data Complementing
Quality Control
Homogenization (MASH)
CLIMATE EXAMINATIONS E.g.
Climate Change Detection
SPATIAL MODELLING OF CLIMATE PARAMETERS (MISH)
Local Statistical Parameters
SPATIAL INTERPOLATION
For arbitrary Location (MISH)
REPRESENTATIVITY EXAMINATION
OF ARBITRARY STATION NETWORK
Inside the Network
SHORT DATA SERIES
Data Complementing
FORECAST
E.g. Data Assimilation,
: Data and Method or/and Result
: only Method or/and Result
: only Data

Figure 2. Connection of topics and systems

5. SPATIAL INTERPOLATION WITH OPTIMUM PARAMETERS

If we want to obtain appropriate modeled interpolation parameters first we have to examine the optimum interpolation parameters, which can be written as certain functions of the climate statistical parameters. In this paper only the interpolation by additive formula (chapter 3.1) is detailed.

Notation:
$Z(\mathbf{s}_0,t)$: predictand, $Z(\mathbf{s}_i,t)$ ($i=1,...,M$): predictors
$ERR(\mathbf{s}_0)$: root mean square interpolation error
$E(\mathbf{s})$: expectation, $D(\mathbf{s})$: standard deviation, $r(\mathbf{s}_1,\mathbf{s}_2)$: correlation
where location vectors $\mathbf{s}$ are the elements of the given space domain D.

Optimum Interpolation Error and Representativity:

$$ERR_{OP}(\mathbf{s}_0) = \text{minimum } ERR(\mathbf{s}_0) \qquad , \qquad REP_{OP}(\mathbf{s}_0) = 1 - \frac{ERR_{OP}(\mathbf{s}_0)}{D(\mathbf{s}_0)}$$

The Structure of the Optimum Interpolation Parameters:
The minimum error can be obtained by the optimum interpolation parameters.
The optimum constant w_0 depends on the differences $E(\mathbf{s}_0)-E(\mathbf{s}_i)$ $(i=1,...,M)$, furthermore the optimum weighing factors w_i $(i=1,.,M)$ as well as the optimum representativity $REP_{OP}(\mathbf{s}_0)$ depend on the ratios $D(\mathbf{s}_0)/D(\mathbf{s}_i)$ $(i=1,.,M)$ and the correlations $r(\mathbf{s}_i,\mathbf{s}_j)$ $(i,j=0,...,M)$. Thus the optimum interpolation parameters and the optimum representativity depend only on the correlation structure and the spatial variability of local statistical parameters. Therefore the monthly interpolation parameters are applicable for the interpolation of daily values too.

Remark: It can be proved that $w_0 = \sum_{i=1}^{M} w_i\left(E(\mathbf{s}_0)-E(\mathbf{s}_i)\right)$ and the vector of nonzero weighing factors is $\mathbf{w} = \frac{\mathbf{C}_{pr}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\mathbf{C}_{pr}^{-1}\mathbf{1}} + \left(\mathbf{C}_{pr}^{-1} - \frac{\mathbf{C}_{pr}^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{C}_{pr}^{-1}}{\mathbf{1}^{\mathrm{T}}\mathbf{C}_{pr}^{-1}\mathbf{1}}\right)\mathbf{c}_{0,pr}$, where $\mathbf{c}_{0,pr}$ and $\mathbf{C}_{pr}$ are the proper predictand-predictors covariance vector and predictors-predictors covariance matrix respectively and vector $\mathbf{1}$ is identically one.

6. SPATIAL MODELLING OF CLIMATE PARAMETERS

6.1 Known Climate Statistical Parameters for Modelling

The long data series can be used to model the climate statistical parameters. If the stations $\mathbf{S}_j$ ($j=1,...,N$) $(\mathbf{S}\in D)$ have long monthly series then the local parameters $E(\mathbf{S}_j)$, $D(\mathbf{S}_j)$ ($j=1,...,N$) as well as the correlations $r(\mathbf{S}_{j1},\mathbf{S}_{j2})$ $(j_1,j_2=1,...,N)$ can be estimated statistically. Consequently these parameters are essentially known and provide a lot of information for modelling. It is again to be remarked that the geostatistical methods applied in GIS neglect the long data series which leads to a loss of information.

6.2 Neighbourhood Modelling of Climate Statistical Parameters

The known climate statistical parameters can be used for modelling the correlation structure as well as the spatial variability of local statistical parameters. The basic principle of the neighbourhood modelling is as follows. Let $P(\mathbf{s})$, $Q(\mathbf{s})$, $\tilde{r}_{\mathbf{s}_0}(\mathbf{s}_1,\mathbf{s}_2)$ $\left(\mathbf{s},\mathbf{s}_0,\mathbf{s}_1,\mathbf{s}_2 \in D\right)$ be certain model functions depending on different model variables with the following properties:

(a) $P(\mathbf{S}_{j1})-P(\mathbf{S}_{j2}) \approx E(\mathbf{S}_{j1})-E(\mathbf{S}_{j2})$, if $\left\|\mathbf{S}_{j1}-\mathbf{S}_{j2}\right\| < d_0$

(b) $\dfrac{Q(\mathbf{S}_{j1})}{Q(\mathbf{S}_{j2})} \approx \dfrac{D(\mathbf{S}_{j1})}{D(\mathbf{S}_{j2})}$, if $\left\|\mathbf{S}_{j1}-\mathbf{S}_{j2}\right\| < d_0$

(c) $\tilde{r}_{\mathbf{s}_0}(\mathbf{S}_{j1},\mathbf{S}_{j2}) \approx r(\mathbf{S}_{j1},\mathbf{S}_{j2})$, if $\left\|\mathbf{S}_{j1}-\mathbf{s}_0\right\| < d_0$ and $\left\|\mathbf{S}_{j2}-\mathbf{s}_0\right\| < d_0$

The model variables may be height, topography (e.g. AURELHY principal components), distance from the sea etc..

7. SPATIAL INTERPOLATION WITH MODELLED PARAMETERS

According to the chapters 5., 6.2 both the modelled weighting factors $\tilde{\mathbf{w}} = [\tilde{w}_1,....,\tilde{w}_M]^{\mathrm{T}}$ and the modelled optimum representativity $REP_{OP}^{\mathrm{mod}}(\mathbf{s}_0)$ can be derived from the values of $\dfrac{Q(\mathbf{s}_0)}{Q(\mathbf{s}_i)}$ $(i=1,...,M)$, $\tilde{r}_{\mathbf{s}_0}(\mathbf{s}_i,\mathbf{s}_j)$ $(i,j=0,...,M)$. Hence,

Interpolation with Modelled Parameters:

$\hat{Z}(\mathbf{s}_0,t) = \tilde{w}_0 + \sum_{i=1}^{M} \tilde{w}_i Z(\mathbf{s}_i,t) = \sum_{i=1}^{M} \tilde{w}_i \left(P(\mathbf{s}_0)-P(\mathbf{s}_i)\right) + \sum_{i=1}^{M} \tilde{w}_i Z(\mathbf{s}_i,t)$. Furthermore,

Representativity of the Interpolation with Modelled Parameters:

$REP_{MP}(\mathbf{s}_0) = 1 - \dfrac{ERR_{MP}(\mathbf{s}_0)}{D(\mathbf{s}_0)}$, where $ERR_{MP}(\mathbf{s}_0)$ is the root mean square inter-polation error obtained by the modelled parameters.

To model the local statistical parameters we can follow a similar approach, that is,

Modelling of Monthly Expectation (using additive interpolation):

$$E^{\mathrm{mod}}(\mathbf{s}_0) = \sum_{k=1}^{K} \tilde{w}_k \left(P(\mathbf{s}_0)-P(\mathbf{S}_{jk})\right) + \sum_{k=1}^{K} \tilde{w}_k E(\mathbf{S}_{jk})$$

Modelling of Monthly Standard Deviation (using multiplicative interpolation):

$$D^{\mathrm{mod}}(\mathbf{s}_0) = \prod_{k=1}^{K} \left(\frac{Q(\mathbf{s}_0)}{Q(\mathbf{S}_{jk})} \cdot D(\mathbf{S}_{jk}) \right)^{\tilde{w}_k}$$

Examples in Hungary

Hungary: 0.5’x0.5’ resolution, approx. 300 000 grid points.

Example 1

Monthly mean temperature: 57 stations with long homogenized data series (1971-2000). One model for each grid point taking into account the nearest 10 stations.
Examination of approx. 600 combinations of stations.

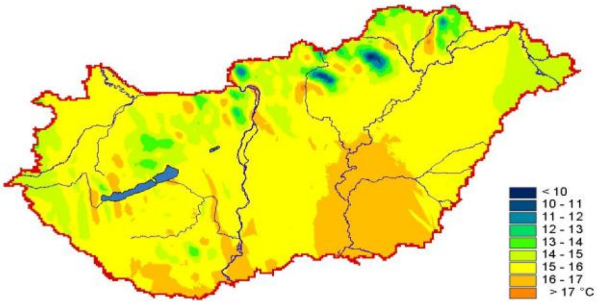


Figure 3. Modelled expectation of monthly mean temperature in September

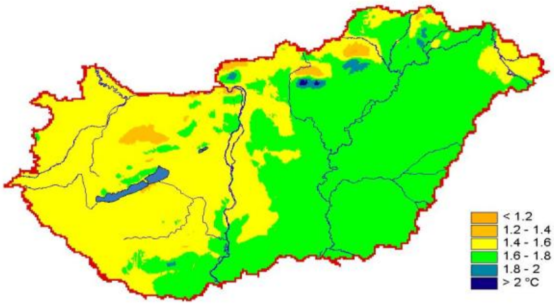


Figure 4. Modelled standard deviation of monthly mean temperature in September

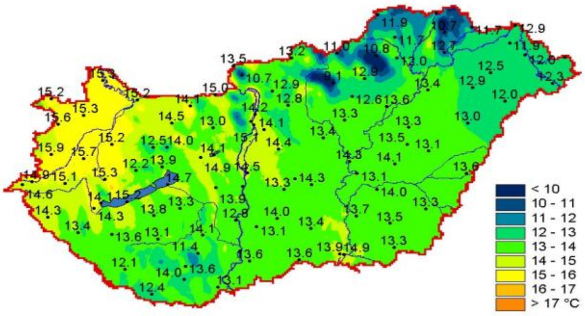


Figure 5. Interpolation of daily mean temperature on 29 September 2004 based on 100 observations

Example 2

Monthly precipitation sum: 500 stations with long homogenized data series (1971-2000). One model for each grid point taking into account the nearest 30 stations.
Examination of approx. 18 000 combinations of stations.

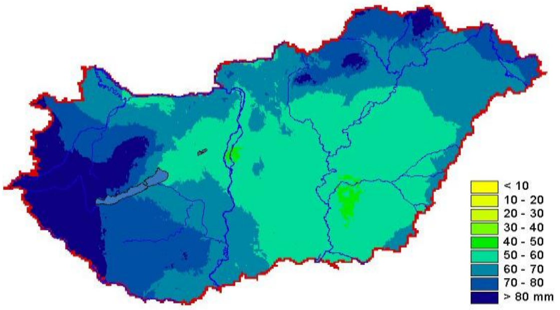


Figure 6. Modelled expectation of monthly precipitation sum in July

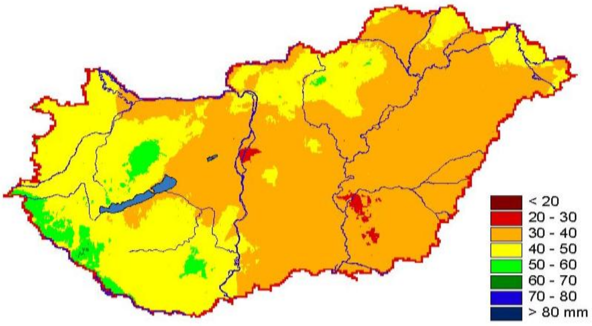


Figure 7. Modelled standard deviation of monthly precipitation sum in July

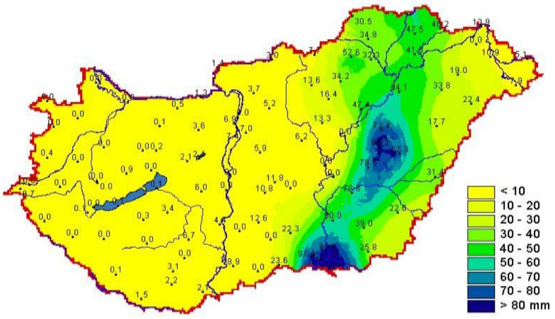


Figure 8. Interpolation of daily precipitation sum on 27 July 2004 based on 103 observations

8. BENCHMARK STUDY TO TEST THE MODELLING RESULTS

The cross-validation test is a possibility to evaluate the interpolation methods. That is interpolation between the station data series and examination of the root mean square interpolation errors $ERR(\mathbf{s}_j)$ or the representativity $REP(\mathbf{s}_j)\ (j = 1,..,N)$.

In our case the interpolation with modelled parameters has been compared to the interpolation with optimum parameters. In Figures 9, 10 we show the mean monthly representativity values that were calculated for both the monthly mean temperature (based on 57 stations) and the monthly precipitation sum (based on 500 stations). For temperature the additive formula (chapter 3.1) while for precipitation the multiplicative formula (chapter 3.2) was applied. For the temperature the inverse distance method which has also an additive interpolation formula was applied too. The notations of the various representativity values are,

REP_{OP}: interpolation with optimum parameters,

REP_{MP}: interpolation with modelled parameters,

REP_{INV}: inverse distance method.

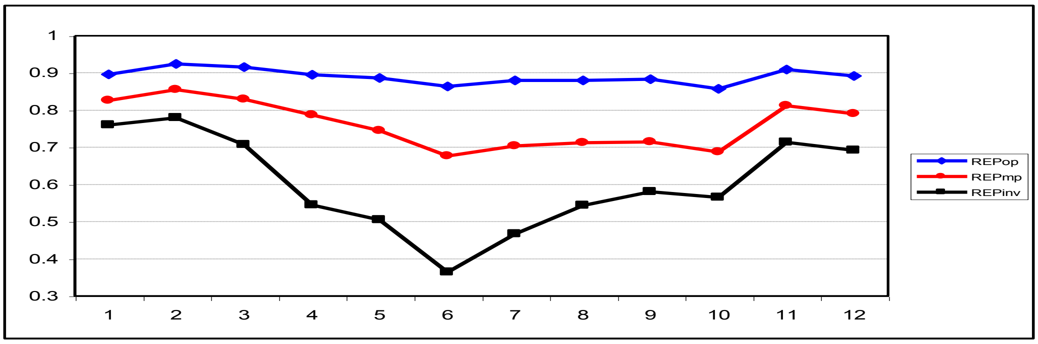


Figure 9. Mean monthly representativity values for monthly mean temperature, 57 stations

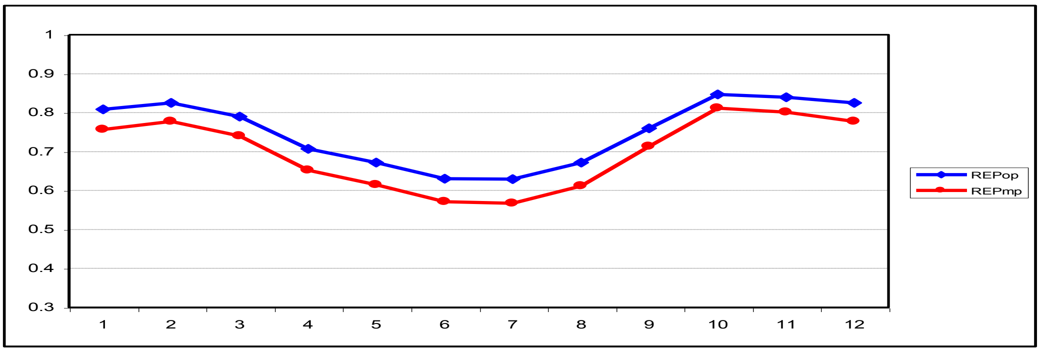


Figure 10. Mean monthly representativity values for monthly precipitation sum, 500 stations

9. MODELLING OF REPRESENTATIVITY REP_{MP}

We can also develop an interpolation procedure for modelling the interpolation error or the representativity. Let $REP_{OP}^{\mathrm{mod}}(\mathbf{s}_0)$, $REP_{OP}^{\mathrm{mod}}(\mathbf{S}_j)\ (j = 1,...,N)$ be the modelled optimum representativity values according to chapter 7, where $\mathbf{s}_0$ is the predictand location, and $\mathbf{S}_j\ (j = 1,...,N)$ are the former station locations. More-over the representativity values $REP_{MP}(\mathbf{S}_j)\ (j = 1,...,N)$ are known as a result of the benchmark study (see chapter 8). Then the representativity of the interpolation with modelled parameters can be interpolated as

$$REP_{MP}^{\mathrm{mod}}(\mathbf{s}_0) = 1 - \prod_{k=1}^{K} \left(\frac{1 - REP_{MP}(\mathbf{S}_{jk})}{1 - REP_{OP}^{\mathrm{mod}}(\mathbf{S}_{jk})} \cdot \left(1 - REP_{OP}^{\mathrm{mod}}(\mathbf{s}_0)\right) \right)^{\widetilde{w}_k}$$

The strength of representativity depends on the predictand-predictors system as well as the quality of modelling. Figures 11,12 are an illustration where the grid points are the predictand locations.

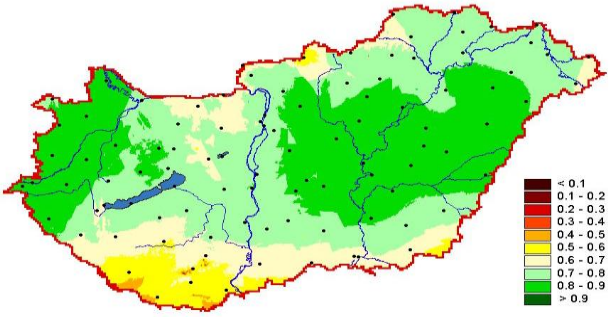


Figure 11. Modelled representativity values REP_{MP}^{mod} for mean temperature in September, the former 100 observing stations are the predictor locations

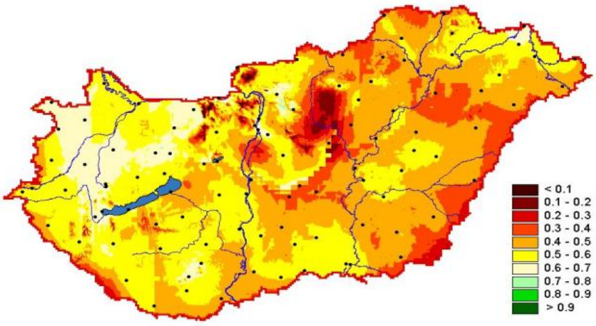


Figure 12. Modelled representativity values REP_{MP}^{mod} for precipitation sum in July, the former 103 observing stations are the predictor locations

10. INTERPOLATION WITH BACKGROUND INFORMATION

The background information e.g. forecast, satellite, radar data (see Fig. 1) can be efficiently used to decrease the interpolation error. In this paper only the interpolation based on additive model or normal distribution is presented.

Let us assume that $Z(\mathbf{s}_j,t)$ $(j=1,...,N)$ are the data measured at the observation stations, $Z(\mathbf{s}_0,t)$ is the predictand and $Z(\mathbf{s}_{ji},t)$ $(i=1,...,M)$ are the predictors where the location vectors $\mathbf{s}$ are the elements of the given space domain D. Furthermore let $G(\mathbf{s},t)$ $(\mathbf{s}\in D)$ be some background information given on a dense grid. The linear model of conditional expectation of $Z(\mathbf{s},t)$, given $G(\mathbf{s},t)$, is

$$\mathrm{E}\big(Z(\mathbf{s},t)\,|\,G(\mathbf{s},t)\big)=E(\mathbf{s})+\gamma_0+\gamma_1\cdot\big(G(\mathbf{s},t)-E(\mathbf{s})\big)\quad,\ (\mathbf{s}\in D)$$

where $E(\mathbf{s})$ is the expectation in space (chapter 5.). The unknown regression parameters γ_0,γ_1 and the correlation $R=\mathrm{corr}\big(Z(\mathbf{s},t),G(\mathbf{s},t)\big)$ can be estimated taking int account the given $Z(\mathbf{s}_j,t)$, $G(\mathbf{s}_j,t)$ $(j=1,...,N)$ and the modelled expectations $E^{\mathrm{mod}}(\mathbf{s}_j)$ $(j=1,...,N)$ formulated in chapter 7. According to chapter 7 again the interpolation without background information can be written as

$$\hat{Z}(\mathbf{s}_0,t)=\tilde{w}_0+\sum_{i=1}^{M}\tilde{w}_{ji}\cdot Z(\mathbf{s}_{ji},t)$$

Applying the same interpolation formula for the background information, we have

$$\hat{G}(\mathbf{s}_0,t)=\tilde{w}_0+\sum_{i=1}^{M}\tilde{w}_{ji}\cdot G(\mathbf{s}_{ji},t)$$

Finally, the formulas in case of using background information are as follows:

Interpolation: $\hat{Z}_G(\mathbf{s}_0,t)=\hat{Z}(\mathbf{s}_0,t)+\gamma_1\cdot\Big(G(\mathbf{s}_0,t)-\hat{G}(\mathbf{s}_0,t)\Big)$

Representativity: $REP_{G,MP}^{\mathrm{mod}}(\mathbf{s}_0)=REP_{MP}^{\mathrm{mod}}(\mathbf{s}_0)+\big(1-REP_{MP}^{\mathrm{mod}}(\mathbf{s}_0)\big)\cdot\Big(1-\sqrt{1-R^2}\Big)$

Figure 13 shows an example. The similarity to Figure 5 is a consequence of the weakness of the correlation.

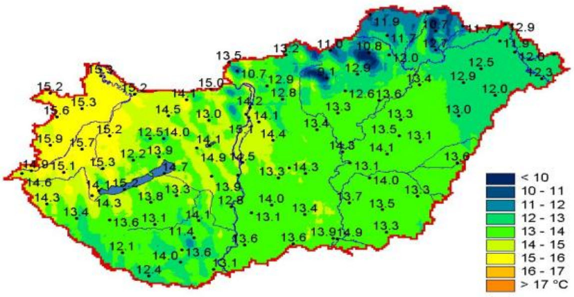


Figure 13. Interpolation of daily mean temperature on 29 September 2004 based on 100 observations and 24 hourly forecast as background information (correlation: *R*=0.48)

11. SOFTWARE: MISHv1.01

We summarize briefly the most important facts about the developed software MISH. Essentially the system consists of two units that are the modelling and the interpolation systems. The interpolation system can be operated on the results of the modelling subsystem.

(a) Modelling Subsystem
(1) Based on long homogenized monthly series.
(2) Benchmark study for interpolation errors or representativity.
(3) Modelling procedure must be executed only once before the interpolation applications.

(b) Interpolation System
(1) Additive (e.g. temperature) or multiplicative (e.g. precipitation) model and interpolation formula can be used depending on the climate elements.
(2) Daily, monthly values and many years' means can be interpolated.
(3) Few predictors are also sufficient for the interpolation.
(4) No problem if the greater part of daily precipitation predictors is zero.
(5) Interpolation error (or rather the representativity) can be modelled too.
(6) Capability for application of background information such as satellite, radar, forecast data.

12. CONCLUSION

To clarify the problem of spatial interpolation in meteorology we have to compare the statistical climatology to the geostatistics in respect of methodology. The statistical climatology based on sample in time is bound to be more powerful than the geostatistics based on only one realisation in time. In meteorology the preference of the geostatistical methods – applied also in GIS – over the statistical climatology leads to a loss of information. Nevertheless appropriate spatial modelling parts must be incorporated into statistical climatology. For that purpose an adequate mathematical background is also necessary of course.

II. THE PROGRAM SYSTEM MISH

II.1 GENERAL COMMENTS

The software MISH consists of two units that are the modelling and the interpolation systems. The interpolation system can be operated on the results of the modelling system!

A, Modelling System

- For monthly climate statistical parameters:
  deterministic parameters (e.g. expectations), stochastic parameters (e.g. correlations)
- Based on long homogenized monthly series and supplementary model variables. The statistical parameters can be modelled per month on the basis of the monthly series. The deterministic model variables may be as height, topography, distance from the sea etc.
- Additive (e.g. temperature) or multiplicative (e.g. precipitation) model can be used depending on the climate elements.
- Benchmark study, cross-validation test for expected interpolation error or representativity.
- Modelling procedure must be executed only once before the interpolation applications!
- The statistical parameters modelled for a month can be used for the interpolation of arbitrary daily and monthly values within the month!

1. Coordinate system: spherical coordinates in decimal degrees ($\varphi^\circ, \lambda^\circ$)

2. To cover the interpolation area with a (rectangle) Grid in decimal degrees ($\varphi^\circ, \lambda^\circ$).
Cell size: equidistant dense scale, scale is the same in decimal degrees ($\varphi^\circ, \lambda^\circ$);
0.5’x0.5’ resolution is suggested ($0.5' \approx 0.0083333333^\circ$)!
The Grid as a matrix: maximum number of rows: 600, maximum number of colums: 900 (e.g. 0.5’x0.5’ resolution, 600 rows, 900 colums: 150 000-300 000 km^2 in Europe).

3. Height data for the Grid (A,2). The height is always model variable.

4. Observation stations with long (homogenized) monthly series within the interpolation area (covered by the Grid (A,2)). Modelling of the statistical parameters for a month is based on the monthly series. However the modelled monthly statistical parameters can be used also to interpolate daily values within the month!
Minimum number of the stations: 10; maximum number of the stations: 500.
Representative station network is suggested.
Minimum length of the series: 20; maximum length of the series: 50.
Length 30-50 is suggested taking into account the temporal representativity as well as the posssible climate change.

5. Other model variables besides the height for the Grid (A,2). The model variables are deterministic variables, e.g. topography, distance from the sea.
Minimum number of model variables besides the height: 0; maximum number of model variables besides the height: 19.

B, Interpolation System

- The interpolation system can be operated on the results of the modelling system!
- Modelling procedure must be executed only once before the interpolation applications!
- Daily, monthly values and many years' means can be interpolated. The statistical parameters modelled for a month can be used for the interpolation of arbitrary daily and monthly values within the month!
- Additive (e.g. temperature) or multiplicative (e.g. precipitation) model and interpolation formula can be used depending on the climate elements.
- Few predictors are also sufficient for the interpolation and no problem if the greater part of daily precipitation predictors is equal to 0.
- The representativity values are modelled too.
- Capability for application of supplementary background information (stochastic variable) e.g. satellite, radar, forecast data.

1. Observations within the interpolation area (covered by the Grid (A,2)) can be daily and monthly values or many years' means.
Minimum number of observations: 1; maximum number of observations: 1000.

2. Interpolation:
a, For given predictand locations (minimum: 1, maximum: 1000) with detailed Results.
Predictand locations: spherical coordinates in decimal degrees ($\varphi^\circ, \lambda^\circ$)
b, For the Grid (A,2) , to obtain Map.

3. Background Information for a relatively dense grid covered by the Grid (A,2).
Background Information Grid: in decimal degrees ($\varphi^\circ, \lambda^\circ$).Cell size: equidistant scale (at our example: $0.15\lambda^\circ$, $0.1\varphi^\circ$); matrix form: maximum number of rows: 600, maximum number of colums: 900. In case of having Background Information the minimum number of observations is 10. The Background Information is appropriate stochastic variable such as satellite, radar or forecast data.

Remark

The modelling and the interpolation systems can be applied directly for interpolation of annual values and many years' annual means as well. In this case the modelling of statistical parameters is based on long homogenized annual series.

The new parts of Interpolation System in Version MISHv1.02

- Missing value interpolation, completion for monthly or daily station data series.
(max. number of series: 500; max. length of series for a given month: 4000)
- Interpolation, gridding of monthly or daily station data series for given predictand locations. In case of gridding the predictand locations are the nodes of a relatively dense grid. (max. number of series: 500; max. length of series for a given month: 4000; max. number of predictand locations, gridpoints: 5000)

These new parts can be also operated on the results of the modelling system! The statistical parameters modelled for a month can be used for arbitrary daily and monthly series values separated for the month!

II.2 THE STRUCTURE OF PROGRAM SYSTEM

Main Directory **MISHv1.02**:

- **MISHMANUAL.PDF**
- Subdirectory **EXAMPLE**
- Directory **MISH**:
    - Subdirectory **MODEL**:
        - **Modelling Program and I/O Files of MISH**
        - Subdirectory **MODPARINTER**
          (Parameter Files for subdirectory **INTERPOL**)
        - Subdirectory **MODELSUB**
          (Executive subroutines for MODEL.BAT, do not run them)
    - Subdirectory **INTERPOL**:
        - **Interpolation Program and I/O Files of INTERPOL**
        - Subdirectory **MODPARINTER**
          (Parameter Files of subdirectory **INTERPOL**)
        - Subdirectory **INTERSUB**
          (Executive subroutines for INTERPAR.BAT, INTERPRED.BAT, INTERGRID.BAT, do not run them)
        - Subdirectory **MISHMISS**
            - **Program and I/O Files of MISHMISS**
            - Subdirectory **MISSSUB**
              (Executive subroutines, do not run them)
        - Subdirectory **MISHINTERSER**
            - **Program and I/O Files of MISHINTERSER**
            - Subdirectory **INTERSERSUB**
              (Executive subroutines, do not run them)

MISH IN PRACTICE

I. Modelling in Subdirectory MODEL

MODEL.BAT: Modelling Procedure. Modelled Statistical Parameters for Interpolation are obtained in Subdirectory MODEL\MODPARINTER.
(To save the Modelled Statistical Parameters is suggested.)

II. Interpolation in Subdirectory INTERPOL

The appropriate monthly Modelled Statistical Parameters for Interpolation must be included by Subdirectory INTERPOL\MODPARINTER.
(Modelled Statistical Parameters must be copied in.)

1. INTERPAR.BAT:
Parametrization and Examination of Observations and Background Information.

2. The further steps can be used optionally

INTERPRED.BAT: Interpolation for given Predictand Locations.

INTERGRID.BAT: Interpolation for the Grid.

Attention: The INTERPAR.BAT must be repeated before the interpolation if the Files of Observations or Background Information are changed!

III. Data Complementing in Subdirectory INTERPOL\MISHMISS

The appropriate monthly Modelled Statistical Parameters must be included by Subdirectory INTERPOL\MODPARINTER. (Modelled Statistical Parameters must be copied in.)

MISHMISS.BAT: Missing Values Completion of Station Data Series.

IV. Interpolation of Series (Gridding) in Subdirectory INTERPOL\MISHINTERSER

The appropriate monthly Modelled Statistical Parameters must be included by Subdirectory INTERPOL\MODPARINTER. (Modelled Statistical Parameters must be copied in.)

MISHINTERSER.BAT: Interpolation of Station Data Series for given Predictand Locations.

Gridding: the locations are the nodes of a relatively dense grid.

The MODELLING PROGRAM and I/O FILES of Subdirectory MODEL

1. Executive Batch File in Directory MODEL

MODEL.BAT: Modelling Procedure

1.1 Subroutines of MODEL.BAT (in MODEL\MODELSUB):

PAR.EXE: Parametrization
STATISTICS.EXE: Estimation of statistical parameters of the data series
COMBIN.EXE: Selection of station combinations for neighbourhood modelling
STOCHMODEL.EXE: Modelling of stochastic parameters
DETMODEL.EXE: Modelling of deterministic parameters
BENCHMARK.EXE: Evaluation of modelling, cross-validation test
MODELGRID.EXE: Modelling results for the grid

2. Input Files and Input Data in Directory MODEL

(See the Data Files of Subdirectory EXAMPLE)

DATASERIES.DAT:
Monthly data series for a given month.
Format of DATASERIES.DAT (max. number of series: 500, suggested length of series: 30):
row 1: indexes or numbers of stations (obligatory!)
column 1: dates or serial indexes
column i+1: series i.

FILAMBDAHST.DAT: Spherical coordinates $\varphi^\circ, \lambda^\circ$ and heights for the stations

HEIGHTGRID.DAT: Determination of the grid ($\varphi^\circ, \lambda^\circ$); heights for the grid

MODVARIST.DAT: Model variables for the stations

MODVARIGRID.DAT: Model variables for the grid determined by HEIGHTGRID.DAT

Question on the screen

Model?: (a)dditive (e.g temperature) or (m)ultiplicative (e.g. precipitation)

3. Output and Result Files

3.1 Result Files 1 written in Subdirectory MODEL\MODPARINTER

(See: Input Files of Directory INTERPOL)
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, INTPAR1.PAR, HEIGHT.PAR(= HEIGHTGRID.DAT)

3.2 Result Files 2 in Directory MODEL

DETMODSTAT.RES: Statistical results of modelling deterministic parameters

BENCHMARK.RES: Benchmark study, evaluation of modelling

MEANGRID.RES: Long term means interpolated for the grid determined by HEIGHTGRID.DAT

4. Parameter Files

TRANS.PAR, MHTR.PAR, STAT1.PAR, TOPOG.PAR, TAVOLSAG.PAR, MAPCOMB.PAR, REFCOMB.PAR, REFSTCOMB.PAR, STAT2ST.PAR, STAT2.PAR, VARST.PAR, VAR.PAR, REPST.PAR

The INTERPOLATION PROGRAM and I/O FILES of Subdirectory INTERPOL

1. Executive Batch Files in Subdirectory INTERPOL

INTERPAR.BAT: Parametrization and Examination of the Background Information

INTERPRED.BAT: Interpolation for given Predictand Locations

INTERGRID.BAT: Interpolation for the grid determined by HEIGHT.PAR

2. Input Files and Input Data

2.1 Input Files 1 in Subdirectory INTERPOL\MODPARINTER

ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, INTPAR1.PAR, HEIGHT.PAR

See: Result Files of Directory MODEL

Attention: Subdirectory MODPARINTER must include the necessary Parameter Files! (Modelled Statistical Parameters must be copied in Subdirectory MODPARINTER.)

2.2 Input Files 2 in Subdirectory INTERPOL

(See the Data Files of Subdirectory EXAMPLE)

OBSERVED.DAT: Observations and coordinates (min. number: 1; max. number: 1000)

BACKINFGRID.DAT: Background Information for a grid inside the grid determined by HEIGHT.PAR

PREDTANDFILA.DAT: Predictand coordinates (min. number: 1; max. number: 1000)
(Input of **INTERPRED.BAT**)

3. Output and Result Files in Subdirectory INTERPOL

INTERPAR.RES: Output of INTERPAR.BAT (if we have Background Information)

INTERPRED1.RES: Output of INTERPRED.BAT (detailed Results)

INTERPRED2.RES: Output of INTERPRED.BAT (less detailed Results)

INTERGRID1.RES: Output of INTERGRID.BAT (Interpolation without Background Information)

INTERGRID2.RES: Output of INTERGRID.BAT (Interpolation with Background Information)

4. Parameter Files

BACKINFH.PAR, BACKINFM.PAR, OBSERVED1.PAR, INTPAR2.PAR, MODPAR.PAR

The PROGRAM and I/O FILES of Subdirectory INTERPOL\MISHMISS

1. Executive Batch File in Directory MISHMISS

MISHMISS.BAT: Data Complementing Procedure

Subroutines of **MISHMISS.BAT** (in **MISHMISS\MISSSUB**):

INTERPAR3.EXE: Parametrization
INTERMISS.EXE: Data Complementing Subroutine

2. Input Files and Input Data

2.1 Input Files 1 in Subdirectory INTERPOL\MODPARINTER

ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, INTPAR1.PAR, HEIGHT.PAR
See: Result Files of Directory MODEL
Attention: Subdirectory MODPARINTER must include the necessary Parameter Files!
(Modelled Statistical Parameters must be copied in Subdirectory MODPARINTER.)

2.2 Input Files 2 in Subdirectory MISHMISS

(See the Data Files of Subdirectory EXAMPLE)

OBSSERIES.DAT:
Observed station data series with missing values.
Format of OBSSERIES.DAT (max. number of series: 500; max. length of series: 4000):
  row 1: indexes or numbers of stations (obligatory!)
  column 1: dates or serial indexes
  column i+1: series i.
Mark of Missing Values: 9999

OBSFILA.DAT: Coordinates of Stations

3. Output and Result File in Subdirectory MISHMISS

OBSSERIES.RES: Complemented station data series

4. Parameter Files

MODPAR.PAR, MISS1.PAR, MISS2.PAR

The PROGRAM and I/O FILES of Subdirectory INTERPOL\MISHINTERSER

1. Executive Batch File in Directory MISHINTERSER

MISHINTERSER.BAT: Series Interpolation Procedure

Subroutines of **MISHINTERSER.BAT** (in **MISHINTERSER\INTERSERSUB**):

INTERPAR4.EXE: Parametrization
INTERSER.EXE: Series Interpolation Subroutine

2. Input Files and Input Data

2.1 Input Files 1 in Subdirectory INTERPOL\MODPARINTER

ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, INTPAR1.PAR, HEIGHT.PAR
See: Result Files of Directory MODEL
Attention: Subdirectory MODPARINTER must include the necessary Parameter Files!
(Modelled Statistical Parameters must be copied in Subdirectory MODPARINTER.)

2.2 Input Files 2 in Subdirectory MISHINTERSER

(See the Data Files of Subdirectory EXAMPLE)

OBSSERIES.DAT:
Observed station data series.
Format of OBSSERIES.DAT (max. number of series: 500; max. length of series: 4000):
row 1: indexes or numbers of stations (obligatory!)
column 1: dates or serial indexes
column i+1: series i.

OBSFILA.DAT: Coordinates of Stations

PREDTANDFILA.DAT: Coordinates of Predictand Locations (max. number: 5000)
Gridding: the locations are the nodes of a relatively dense grid.

3. Output and Result Files in Subdirectory MISHINTERSER

INTERSERIES.RES: Interpolated (Gridded) Series

INTERSERSTAT.RES: Statistical Results for the Predictand Locations

4. Parameter Files

MODPAR.PAR, OBSSER.PAR

III. EXAMPLE FOR APPLICATION OF MISH SYSTEM

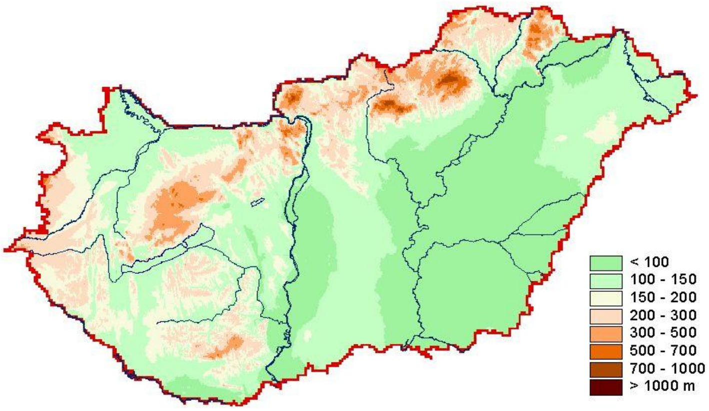


Figure 1. Map of Hungary

Interpolation Area: Transdanubia (the western part from Danube in Hungary)
Modelled Elements: (Monthly, daily) mean temperature in September and (monthly, daily) precipitation sum in November.
Interpolated Elements: Daily mean temperature for a day in September and daily precipitation sum for a day in November.
Missing Values Completion of Station Data Series: Monthly mean temperature series in September and monthly precipitation sum series in November.
Gridding: Monthly mean temperature series in September and monthly precipitation sum series in November.

III.1 EXAMPLE FOR TEMPERATURE

III.1.1 MODELLING PART (Directory MODEL)

Input Data Files

(See the Data Files Format in Subdirectory EXAMPLE\ HUNTEMP\DATA\MODEL)
DATASERIES.DAT: Series of monthly mean temperature in September; 30 stations and 30 years. (Not genuine data.)
FILAMBDAHST.DAT: spherical coordinates in decimal degrees $\varphi^\circ, \lambda^\circ$ and heights for the stations
HEIGHTGRID.DAT: grid (0.5’x0.5’ resolution) covering Transdanubia; heights for the grid
MODVARIST.DAT: 15 model variables (AURELHY principal components) besides the height for the stations
MODVARIGRID.DAT: the model variables for the grid determined by HEIGHTGRID.DAT
MODEL (answer to the question on the screen): (a)dditive

Output and Result Files

Result Files 1: Modelled Climate Statistical Parameters for September
written in Subdirectory MODEL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR
(See: Input Files of Directory INTERPOL)

Result Files 2 (written in Directory MODEL) are the following:

```
MODELLING OF DETERMINISTIC PART (linear regression)

FINAL RESULT:

number of model variables:  6    correlation: 0.921    percentage: 61.1%
model variables and coefficients:
        h        5        9       10       12       14
  -0.0033   0.0284   0.0482   0.0324  -0.0924  -0.0274

(percentage=(1-RMSE/(Standard Deviation))*100%)


DETAILED RESULTS:

number of variables:  1    correlation: 0.814    percentage: 41.9%
       h
  -0.004
number of variables:  2    correlation: 0.879    percentage: 52.4%
       h      12
  -0.004  -0.079
number of variables:  3    correlation: 0.901    percentage: 56.6%
       h      10      12
  -0.004   0.042  -0.081
number of variables:  4    correlation: 0.907    percentage: 57.9%
       h       3      10      12
  -0.004  -0.011   0.047  -0.084
number of variables:  5    correlation: 0.919    percentage: 60.6%
       h       5       9      10      12
  -0.004   0.024   0.045   0.040  -0.079
number of variables:  6    correlation: 0.921    percentage: 61.1%
       h       5       9      10      12      14
  -0.003   0.028   0.048   0.032  -0.092  -0.027
number of variables:  7    correlation: 0.926    percentage: 62.1%
       h       4       5       9      10      12      14
  -0.004  -0.027   0.038   0.061   0.041  -0.086  -0.038
number of variables:  8    correlation: 0.929    percentage: 63.1%
       h       1       4       5       9      10      12      15
  -0.003   0.009  -0.033   0.036   0.045   0.041  -0.059   0.096
      .
      .
```

Figure 2. Statistical results of modelling deterministic parameters (DETMODSTAT.RES)

```
BENCHMARK STUDY: cross-validation test, interpolation between the stations

REPRESENTATIVITY VALUES (REP) FOR THE STATIONS
REP=1-RMSE/(Standard Deviation)
REPop: interpolation with optimum parameters
REPmp: interpolation with modelled parameters

ST. INDEX    REPop    REPmp
        1    0.893    0.795
        2    0.847    0.779
        3    0.860    0.800
        4    0.916    0.638
        5    0.942    0.705
        6    0.869    0.557
        7    0.915    0.586
        8    0.923    0.885
        9    0.891    0.760
       10    0.883    0.829
       11    0.906    0.606
           .
           .
           .
       19    0.828    0.757
       20    0.843    0.818
       21    0.864    0.852
       22    0.879    0.623
       23    0.845    0.827
       24    0.899    0.742
       25    0.919    0.863
       26    0.892    0.848
       27    0.863    0.792
       28    0.902    0.871
       29    0.895    0.813
       30    0.876    0.680
     MEAN    0.885    0.766
```

Figure 2. Benchmark study, evaluation of modelling (BENCHMARK.RES)

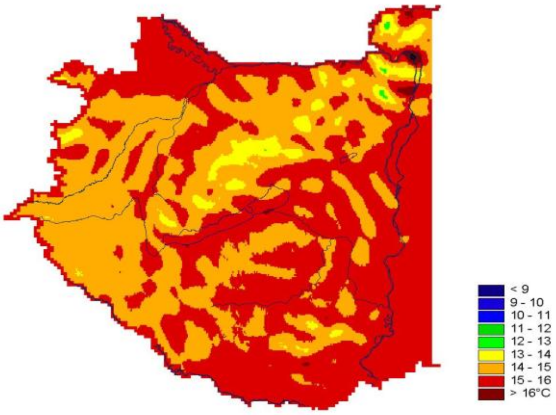


Figure 3. Modelled expectation (or interpolated many years' mean) of monthly mean temperature in September (MEANGRID.RES)

III.1.2 INTERPOLATION PART (Directory INTERPOL)

Input Files and Input Data
(See the Data Files Format in Subdirectory EXAMPLE\HUNTEMP \DATA\INTERPOL)

Input Files 1: Modelled Climate Statistical Parameters for September
in Subdirectory INTERPOL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR
(See: Output Files of Directory MODEL)

Input Data Files 2 (in Directory INTERPOL)
OBSERVED.DAT: 49 daily mean temperature observations for a day in September and the observing locations (coordinates $\varphi^{\circ},\lambda^{\circ}$)
BACKINFGRID.DAT: Forecast data as Background Information for a grid (6’x9’ resolution) inside the grid (0.5’x0.5’ resolution) determined by HEIGHT.PAR
PREDTANDFILA.DAT: 121 predictand coordinates $\varphi^{\circ},\lambda^{\circ}$ (Input of INTERPRED.BAT)

Result Files (written in Directory INTERPOL) are the following

```
EXAMINATION OF BACKGROUND INFORMATION
 Correlation:     0.300
 Constant:       -0.354
 Coefficient:     0.571
 Interpolated Background Information for the Observing Locations:
14.08   13.80   13.23   12.75   13.20   13.59   12.20   11.62   14.04 ………
```

Figure 4. Correlation and regression analysis between obsevations and background information (forecast data) (INTERPAR.RES)

```
Number of Predictands:       121
.
.
Predictand   64:   17.600000   47.400000

 Predictor Indexes :    12    13    10    30    16    11    14
  Weighting Factors: 0.204 0.065 0.199 0.178 0.114 0.129 0.110
 Interpolation without Background Information:
  Predictand Value:     14.77
  Representativity:     0.814
 Interpolation with Background Information:
  Predictand Value:     14.74
  Representativity:     0.822
.
.
```

Figure 5. Detailed result of interpolation for the given predictand locations (INTERPRED1.RES)

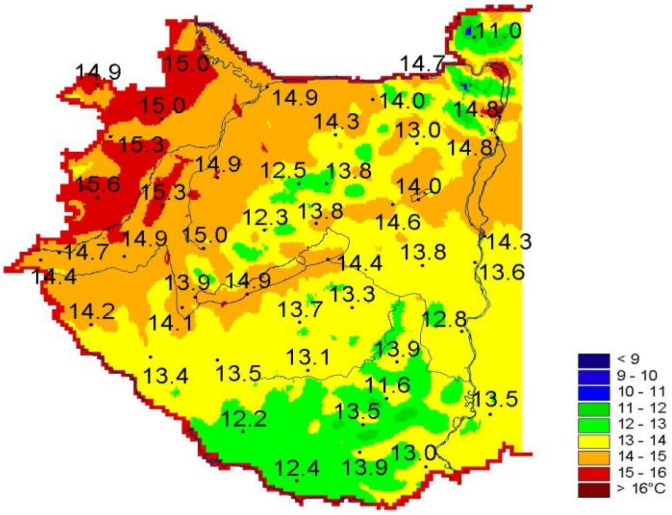


Figure 6. Interpolation without background information for the grid (0.5'x0.5' resolution) determined by HEIGHT.PAR (INTERGRID1.RES)

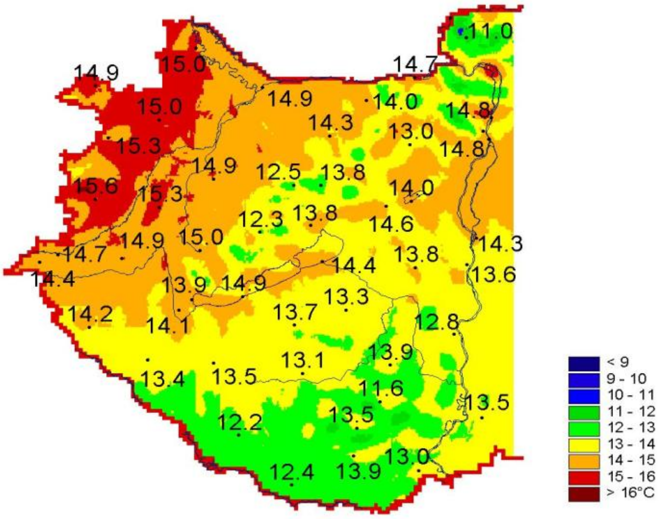


Figure 7. Interpolation with background information (forecast data) for the grid (0.5'x0.5' resolution) determined by HEIGHT.PAR (INTERGRID2.RES)

III.1.3 DATA SERIES COMPLEMENTING (Subdirectory INTERPOL\MISHMISS)

<u>Input Files and Input Data</u> (See the Data Files Format in Subdirectory EXAMPLE\HUNTEMP \DATA\INTERPOL\MISHMISS)

Input Files 1: Modelled Climate Statistical Parameters for September
in Subdirectory INTERPOL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR (See: Output Files of Directory MODEL)

Input Data Files 2 (in Subdirectory INTERPOL\MISHMISS)

OBSSERIES.DAT: Monthly mean temperature series in September with missing values; 30 stations and 29 years. Mark of missing values: 9999.00

OBSFILA.DAT: spherical coordinates in decimal degrees $\varphi^\circ, \lambda^\circ$ for the stations

<u>Result File (written in Subdirectory INTERPOL\MISHMISS) is the following:</u>

OBSSERIES.RES: The complemented data series
(See in Subdirectory EXAMPLE\HUNTEMP\RESULTS\INTERPOL\MISHMISS)

III.1.4 INTERPOLATION OF SERIES, GRIDDING (Subdirectory INTERPOL\MISHINTERSER)

<u>Input Files and Input Data</u> (See the Data Files Format in Subdirectory EXAMPLE\HUNTEMP \DATA\INTERPOL\MISHINTERSER)

Input Files 1: Modelled Climate Statistical Parameters for September
in Subdirectory INTERPOL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR (See: Output Files of Directory MODEL)

Input Data Files 2 (in Subdirectory INTERPOL\MISHINTERSER)
OBSSERIES.DAT: Monthly mean temperature series in September without missing values; 30 stations and 29 years

OBSFILA.DAT: spherical coordinates in decimal degrees $\varphi^\circ, \lambda^\circ$ for the stations

PREDTANDFILA.DAT: spherical coordinates in decimal degrees $\varphi^\circ, \lambda^\circ$ of 768 grid points,
$0.1\lambda^\circ$ x $0.1\varphi^\circ$ resolution.
(Gridding: the grid points are the predictand locations)

<u>Result Files (written in Subdirectory INTERPOL\MISHINTERSER) are the following:</u>

INTERSERIES.RES: Interpolated (Gridded) Series for the 768 grid points

INTERSERSTAT.RES: Statistical Results of the Interpolation for the 768 grid points

(See in Subdirectory EXAMPLE\HUNTEMP\RESULTS\INTERPOL\MISHINTERSER)

III.2 EXAMPLE FOR PRECIPITATION

III.2.1 MODELLING PART (Directory MODEL)

Input Data Files

(See the Data Files Format in Subdirectory EXAMPLE\HUNPREC \DATA\MODEL)
DATASERIES.DAT: Series of monthly precipitation sum in November; 117 stations and 30 years. (Not genuine data.)
FILAMBDAHST.DAT: spherical coordinates in decimal degrees $\varphi^\circ, \lambda^\circ$ and heights for the stations
HEIGHTGRID.DAT: grid (0.5’x0.5’ resolution) covering Transdanubia; heights for the grid
MODVARIST.DAT: 15 model variables (AURELHY principal components) besides the height for the stations
MODVARIGRID.DAT: the model variables for the grid determined by HEIGHTGRID.DAT
MODEL (answer to the question on the screen): (m)ultiplicative

Output and Result Files

Result Files 1: Modelled Climate Statistical Parameters for November
written in Subdirectory MODEL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR
(See: Input Files of Directory INTERPOL)

Result Files 2 (written in Directory MODEL) are the following:

```
MODELLING OF DETERMINISTIC PART (linear regression)
Multiplicative model: logarithmic values are used
FINAL RESULT:
number of model variables:  5    correlation: 0.530    percentage: 15.2%
model variables and coefficients:
        h        3        5        6        9
   0.0004   0.0018  -0.0023  -0.0031   0.0066
(percentage=(1-RMSE/(Standard Deviation))*100%)

DETAILED RESULTS:
number of variables:  1    correlation: 0.391    percentage:  8.0%
       h
   0.001
number of variables:  2    correlation: 0.449    percentage: 10.7%
       h       9
   0.001   0.006
number of variables:  3    correlation: 0.494    percentage: 13.1%
       h       3       9
   0.000   0.002   0.007
number of variables:  4    correlation: 0.521    percentage: 14.7%
       h       3       6       9
   0.000   0.002  -0.003   0.007
number of variables:  5    correlation: 0.530    percentage: 15.2%
       h       3       5       6       9
   0.000   0.002  -0.002  -0.003   0.007
number of variables:  6    correlation: 0.534    percentage: 15.4%
       h       3       5       6       8       9
   0.000   0.002  -0.002  -0.003  -0.002   0.007
```

Figure 8. Statistical results of modelling deterministic parameters (DETMODSTAT.RES)

```
BENCHMARK STUDY: cross-validation test, interpolation between the stations

REPRESENTATIVITY VALUES (REP) FOR THE STATIONS
REP=1-RMSE/(Standard Deviation)
REPop: interpolation with optimum parameters
REPmp: interpolation with modelled parameters

REPRESENTATIVITY VALUES FOR THE STATIONS

ST. INDEX     REPop     REPmp
        1     0.739     0.710
        2     0.742     0.691
        3     0.831     0.720
        4     0.751     0.735
        5     0.714     0.586
        6     0.803     0.768
        7     0.858     0.840
        8     0.684     0.622
        9     0.763     0.693
       10     0.820     0.804
       11     0.814     0.770
           .
      106     0.863     0.836
      107     0.793     0.757
      108     0.818     0.784
      109     0.825     0.746
      110     0.853     0.807
      111     0.860     0.759
      112     0.830     0.790
      113     0.813     0.693
      114     0.743     0.650
      115     0.777     0.769
      116     0.806     0.750
      117     0.764     0.709
     MEAN     0.816     0.766
```

Figure 9. Benchmark study, evaluation of modelling (BENCHMARK.RES)

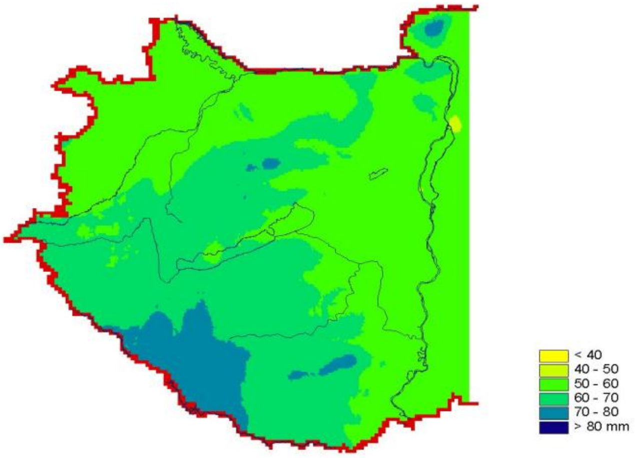


Figure 10. Modelled expectation (or interpolated many years' mean) of of monthly precipitation sum in November (MEANGRID.RES)

III.2.2 INTERPOLATION PART (Directory INTERPOL)

Input Files and Input Data

(See the Data Files Format in Subdirectory EXAMPLE\HUNPREC \DATA\INTERPOL)

Input Files 1: Modelled Climate Statistical Parameters for November
in Subdirectory INTERPOL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR
(See: Output Files of Directory MODEL)

Input Data Files 2 (in Directory INTERPOL)

OBSERVED.DAT: 43 daily precipitation sum observations for a day in November and the observing locations (coordinates $\varphi^{\circ}, \lambda^{\circ}$)

BACKINFGRID.DAT: No

PREDTANDFILA.DAT: 121 predictand coordinates $\varphi^{\circ}, \lambda^{\circ}$ (Input of INTERPRED.BAT)

Result Files (written in Directory INTERPOL) are the following

```
Number of Predictands:        121
.
Predictand   64:   17.600000   47.400000

 Predictor Indexes :     10     8    27    25     9    11    13
  Weighting Factors: 0.422 0.199 0.080 0.191 0.058 0.002 0.048
 Interpolation:
  Predictand Value:     24.40
  Representativity:     0.663
.
.
```

Figure 11. Detailed result of interpolation for the given predictand locations (INTERPRED1.RES)

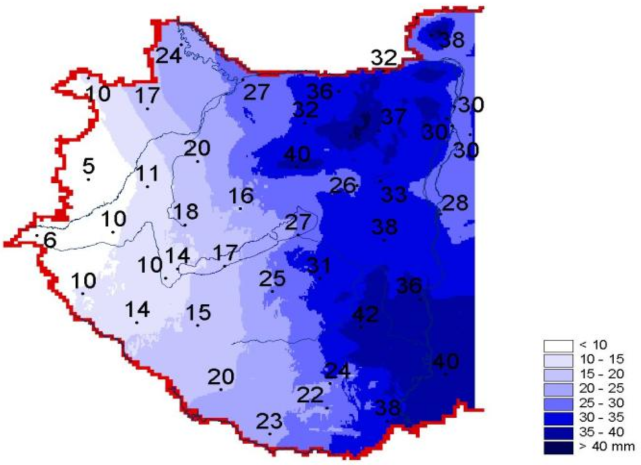


Figure 12. Interpolation without background information for the grid (0.5’x0.5’ resolution) determined by HEIGHT.PAR (INTERGRID1.RES)

III.2.3 DATA SERIES COMPLEMENTING (Subdirectory INTERPOL\MISHMISS)

Input Files and Input Data (See the Data Files Format in Subdirectory EXAMPLE\HUNPREC\DATA\INTERPOL\MISHMISS)

Input Files 1: Modelled Climate Statistical Parameters for November
in Subdirectory INTERPOL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR (See: Output Files of Directory MODEL)

Input Data Files 2 (in Subdirectory INTERPOL\MISHMISS)

OBSSERIES.DAT: Monthly precipitation sum series in November with missing values; 117 stations and 30 years. Mark of missing values: 9999.00

OBSFILA.DAT: spherical coordinates in decimal degrees $\varphi^{\circ}, \lambda^{\circ}$ for the stations

Result File (written in Subdirectory INTERPOL\MISHMISS) is the following:

OBSSERIES.RES: The complemented data series
(See in Subdirectory EXAMPLE\HUNPREC\RESULTS\INTERPOL\MISHMISS)

III.2.4 INTERPOLATION OF SERIES, GRIDDING PART (Subdirectory INTERPOL\MISHINTERSER)

Input Files and Input Data (See the Data Files Format in Subdirectory EXAMPLE\HUNPREC\DATA\INTERPOL\MISHINTERSER)

Input Files 1: Modelled Climate Statistical Parameters for November
in Subdirectory INTERPOL\MODPARINTER:
ALF.PAR, BET.PAR, GAM.PAR, MED.PAR, DEL.PAR, POTPRED.PAR, HEIGHT.PAR, INTPAR1.PAR (See: Output Files of Directory MODEL)

Input Data Files 2 (in Subdirectory INTERPOL\MISHINTERSER)
OBSSERIES.DAT: Monthly precipitation sum series in November without missing values; 117 stations and 30 years

OBSFILA.DAT: spherical coordinates in decimal degrees $\varphi^{\circ}, \lambda^{\circ}$ for the stations

PREDTANDFILA.DAT: spherical coordinates in decimal degrees $\varphi^{\circ}, \lambda^{\circ}$ of 768 grid points,
$0.1\lambda^{\circ}$ x $0.1\varphi^{\circ}$ resolution.
(Gridding: the grid points are the predictand locations)

Result Files (written in Subdirectory INTERPOL\MISHINTERSER) are the following:

INTERSERIES.RES: Interpolated (Gridded) Series for the 768 grid points

INTERSERSTAT.RES: Statistical Results of the Interpolation for the 768 grid points

(See in Subdirectory EXAMPLE\HUNPREC\RESULTS\INTERPOL\MISHINTERSER)

References

Cressie, N., 1991: „Statistics for Spatial Data.", Wiley, New York, 900p.

Benichou, P., Le Breton, O., 1986: „Prise en compte de la topographie pour la cartographie des champs pluviométriqes statistiques." Prix Norbert Gerbier, Direction de la Météorologie Nationale.

Bihari, Z., Szentimrey, T., Lakatos, M., Szalai, S., 2007: „Verification of radar precipitation measurements with interpolated surface data", Advances in Geosciences (submitted)

Gandin, L., 1965: „Objective analysis of meteorological fields", Israel Program for Scientific Translations, Jerusalem, Israel.

Szentimrey, T., 2002: „Statistical problems connected with the spatial interpolation of climatic time series.", Home page:http://www.knmi.nl/samenw/cost719/documents/Szentimrey.pdf

Szentimrey, T., 2004: „Something like an Introduction", Proceedings of the Fourth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary, 2003; WMO, WCDMP-No. 56, pp. 5-16.

Szentimrey, T., Bihari, Z., Szalai, S., 2005: „Meteorological Interpolation based on Surface Homogenized Data Basis (MISH)", European Geosciences Union, General Assembly 2005, Vienna, Austria, 24 - 29 April 2005

Szentimrey, T., Bihari, Z., Szalai, S., 2005: „Limitations of the present GIS methods in respect of meteorological purposes", 5th Annual Meeting of the European Meteorological Society (EMS)/7th ECAM , Utrecht, Netherlands, 12-16 September 2005

Szentimrey, T., Bihari, Z., 2006: „MISH (Meteorological Interpolation based on Surface Homogenized Data Basis)", COST Action 719 Final Report, The use of GIS in climatology and meteorology, Edited by Ole Einar Tveito, Martin Wegehenkel, Frans van der Wel and Hartwig Dobesch, 2006, pp. 54-56

Szentimrey, T., Bihari, Z., 2007: „Mathematical background of the spatial interpolation methods and the software MISH (Meteorological Interpolation based on Surface Homogenized Data Basis)", Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, Hungary, 2004, COST Action 719, COST Office, 2007, pp. 17-27

Szentimrey, T., 2007: „Manual of homogenization software MASHv3.02", Hungarian Meteorological Service, p. 65

Szentimrey, T, Bihari, Z., Szalai,S., 2007: „Comparison of geostatistical and meteorological interpolation methods (what is what?)", Spatial Interpolation for climate data - the use of GIS in climatology and meteorology, Edited by Hartwig Dobesch, Pierre Dumolard and Izabela Dyras, 2007, ISTE ltd., London, UK, 284pp, ISBN 978-1-905209-70-5, pp.45-56

Tveito, O., E., Schöner, W., 2002: „Applications of spatial interpolation of climatological an meteorological elements by the use of geographical information systems (GIS)", Report no. 1/WG2 Spatialisation/ COST-719, DNMI report 28/02 KLIMA, Oslo, Norway