Creation of a representative climatological database for Hungary from 1870 to 2020

Beatrix Izsák1*, 2, Tamás Szentimrey3, Mónika Lakatos1, Rita Pongrácz4 and Olivér Szentes1

1 Hungarian Meteorological Service, Kitabel P. s 1, H-1024, Budapest, Hungary
2 ELTE Faculty of Science, Doctoral School of Earth Sciences Budapest, Hungary
3 Varimax Limited Partnership Budapest, Hungary
4 ELTE Institute of Geography and Earth Sciences, Department of Meteorology Budapest, Hungary

izsak.b@met.hu
In order to have a spatially and temporally representative database, the first step is to homogenize, check and complete the missing values in the station data series. Since we have data series of different lengths this is not such an easy task. Creating the representative database we have to use as much data and as long data series as possible!

The research is supported by the ÚNKP-20-3 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.
A spatially and temporally representative climatological database is created at HMS using MASH and MISH software:

MASHv3.03  
(Multiple Analysis of Series for Homogenization; Szentimrey, T.)  
For homogenization, quality control and missing value completion of station daily data series

MISHv1.03  
(Meteorological Interpolation based on Surface Homogenized Data Basis; Szentimrey, T. and Bihari, Z.)  
For gridding (interpolation) of homogenized daily data series

The main properties of the version MASHv3.03

Advantages of MASHv3.03 in the homogenization of monthly series:
- It is a relative homogeneity test procedure.
- **It is a step-by-step iteration procedure**: the role of series (candidate, reference) changes step by step in the course of the procedure.
- An additive (e.g. temperature) or multiplicative (e.g. precipitation) model can be used depending on the distribution.
- It includes quality control and missing data completion.
- It provides the homogeneity of the seasonal and annual series as well.
- Metadata (probable dates of break points) can be used automatically.
- The homogenization results can be evaluated on the basis of verification tables generated automatically during the procedure.

In the homogenization of daily series:
- The procedure is **based on the detected monthly inhomogeneities**.
- It includes quality control and the completion of missing data in daily data.
The verification statistics in MASH

The test statistics generated automatically during the procedure:

TEST STATISTICS FOR SERIES INHOMOGENEITY
- Test statistics after homogenization
- Test statistics before homogenization
- Statistics for estimated inhomogeneities

CHARACTERIZATION OF INHOMOGENEITY
- Relative estimated inhomogeneities
- Relative modification of series
- Lower confidence limit for relative residual inhomogeneities

REPRESENTATIVITY OF STATION NETWORK

EVALUATION OF META DATA
- Test statistics
- Representativity of META data
The modeling subsystem for climate statistical (local and stochastic) parameters:

- This is based on long homogenized data series and supplementary deterministic model variables. The model variables may include such elements as height, topography, distance from the sea etc. There is neighbourhood modeling, correlation model for each grid point.
- It is also a benchmark study, a cross-validation test for interpolation error or representativity.
- It should be noted that the modeling procedure must be executed only once before the interpolation applications.

The interpolation subsystem:

- Additive (e.g. temperature) or multiplicative (e.g. precipitation) model and interpolation formula can also be used, depending on the climate elements.
- Daily or monthly values and the means from a number of years can be interpolated.
- Just a few predictors are sufficient for the interpolation, and no problem arises if the greater part of daily precipitation predictors is equal to 0.
- Representativity is also modelled.
- Supplementary background information (stochastic variables) e.g. satellite, radar, forecast data can also be used.
- Data series completion, that is, missing value interpolation, completion for monthly or daily station data series is possible.
- Interpolation, the gridding of monthly or daily station data series for given predictand locations is made possible. In case of gridding the predictand locations are the nodes of a relatively dense grid.
Problem: how do we homogenize three or more datasets of different lengths together?

When the station network is upgraded and we have short datasets besides the long sets, the common section must be homogeneous together with the long as well as with the short datasets, while the three or more systems have to be homogeneous themselves too. MASH is able to fulfill these criteria, as it is based on hypothesis testing and it involves an iteration procedure. The solution is that we synchronize the common part’s inhomogeneities within three or more different MASH processing for the three or more datasets with different length.

**MASH is an iteration procedure**, the series are examined and adjusted many times, therefore the homogenization of the new system can be considered as a continuation of the earlier homogenization procedure.

The **test of hypothesis** and throughout this test, the **test statistics** enable us to use the former results.
Daily mean temperature

Location of the stations in case of temperature
The steps of homogenization of temperature station data series

1. MASH1: homogenization of monthly data
2. Cut out the inhomogeneities of the common part and insert them into the other three MASHs
3. MASH2: homogenization of monthly data
4. Cut out the inhomogeneities of the common part and insert them into the other three MASHs
5. MASH3: homogenization of monthly data
6. Cut out the inhomogeneities of the common part and insert them into the other three MASHs
7. MASH4: homogenization of monthly data
8. Cut out the inhomogeneities of the common part and insert them into the other three MASHs
9. If statistics are acceptable in MASH1: go to point 10, if not, go to step 1.
10. If statistics are acceptable in MASH2: go to point 11, if not, go to step 3.
11. If statistics are acceptable in MASH3: go to point 12, if not, go to step 5.
13. Gathering the homogenized data sets from the different MASH systems.
The most important verification statistics in case of annual mean temperature

<table>
<thead>
<tr>
<th></th>
<th>MASH</th>
<th>MASH1</th>
<th>MASH2</th>
<th>MASH3</th>
<th>MASH4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significance level:</td>
<td></td>
<td>Critical</td>
<td>Critical</td>
<td>Critical</td>
<td>Critical</td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>value: 22.05</td>
<td>value: 21.76</td>
<td>value: 21.31</td>
<td>value: 20.86</td>
</tr>
<tr>
<td>Test Statistics</td>
<td></td>
<td>1253.08</td>
<td>877.95</td>
<td>357.2</td>
<td>257.49</td>
</tr>
<tr>
<td>Before Homogenization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Statistics</td>
<td></td>
<td>44.89</td>
<td>29.1</td>
<td>24.83</td>
<td>21.81</td>
</tr>
<tr>
<td>After Homogenization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative</td>
<td></td>
<td>0.43</td>
<td>0.48</td>
<td>0.39</td>
<td>0.38</td>
</tr>
<tr>
<td>Modification of Series</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Representativity of</td>
<td></td>
<td>0.83</td>
<td>0.87</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>station network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpolation (gridding) with MISH

• Because far fewer stations were considered in the previous modeling procedure, we updated the results of the MISH model using the new homogenized datasets.

• These results were applied to gridding in all four cases.

• Finally the four gridded datasets are merged to a common database by taking grid point values from the dataset which consists of most stations in the given time period.
Gridpoints (resolution 0.1° x 0.1°)
Gridding by MISH

NEW modeled parameters

Homogenized data series
MASH1

Homogenized data series
MASH2

Homogenized data series
MASH3

Homogenized data series
MASH4

Gridded data series
MISH1
1870-1900

Gridded data series
MISH2
1901-1950

Gridded data series
MISH3
1951-1974

Gridded data series
MISH4
1975-2020

Gridded data series
1870-2020
ANOVA (Analysis Of Variance)

\[ Z(s_j, t) \ (j = 1, \ldots, N; \ t = 1, \ldots, n) \] – data series \ (s_j : \text{location}; \ t : \text{time})

\[ \hat{E}(s_j) = \frac{1}{n} \sum_{t=1}^{n} Z(s_j, t) \ (j = 1, \ldots, N) \] – temporal mean at location \( s_j \)

\[ \hat{D}^2(s_j) = \frac{1}{n} \sum_{t=1}^{n} \left( Z(s_j, t) - \hat{E}(s_j) \right)^2 \ (j = 1, \ldots, N) \] – temporal variance at location \( s_j \)

\[ \hat{E}(t) = \frac{1}{N} \sum_{j=1}^{N} Z(s_j, t) \ (t = 1, \ldots, n) \] – spatial mean at moment \( t \)

\[ \hat{D}^2(t) = \frac{1}{N} \sum_{j=1}^{N} \left( Z(s_j, t) - \hat{E}(t) \right)^2 \ (t = 1, \ldots, n) \] – spatial variance at moment \( t \)

\[ \hat{E} = \frac{1}{N \cdot n} \sum_{j=1}^{N} \sum_{t=1}^{n} Z(s_j, t) = \frac{1}{N} \sum_{j=1}^{N} \hat{E}(s_j) = \frac{1}{n} \sum_{t=1}^{n} \hat{E}(t) \] – total mean

\[ \hat{D}^2 = \frac{1}{N \cdot n} \sum_{j=1}^{N} \sum_{t=1}^{n} \left( Z(s_j, t) - \hat{E} \right)^2 \] – total variance
Partitioning of Total Variance (Theorem)

\[ \hat{D}^2 = \frac{1}{N} \sum_{j=1}^{N} (\hat{E}(s_j) - \hat{E})^2 + \frac{1}{N} \sum_{j=1}^{N} \hat{D}^2(s_j) = \frac{1}{n} \sum_{t=1}^{n} (\hat{E}(t) - \hat{E})^2 + \frac{1}{n} \sum_{t=1}^{n} \hat{D}^2(t) \]

\[ \frac{1}{N} \sum_{j=1}^{N} (\hat{E}(s_j) - \hat{E})^2 \quad \text{– spatial variance of temporal means} \]

\[ \frac{1}{N} \sum_{j=1}^{N} \hat{D}^2(s_j) \quad \text{– spatial mean of temporal variances} \]

\[ \frac{1}{n} \sum_{t=1}^{n} (\hat{E}(t) - \hat{E})^2 \quad \text{– temporal variance of spatial means} \]

\[ \frac{1}{n} \sum_{t=1}^{n} \hat{D}^2(t) \quad \text{– temporal mean of spatial variances} \]
Total standard deviation:

\[ \hat{D} = \sqrt{\frac{1}{N \cdot n} \sum_{j=1}^{N} \sum_{t=1}^{n} (Z(s_j, t) - \hat{E})^2} \]

Spatial standard deviation of temporal means:

\[ \sqrt{\frac{1}{N} \sum_{j=1}^{N} (\hat{E}(s_j) - \hat{E})^2} \]

Root spatial mean of temporal variances:

\[ \sqrt{\frac{1}{N} \sum_{j=1}^{N} \hat{D}^2(s_j)} \]

Temporal standard deviation of spatial means:

\[ \sqrt{\frac{1}{n} \sum_{t=1}^{n} (\hat{E}(t) - \hat{E})^2} \]

Root temporal mean of spatial variances:

\[ \sqrt{\frac{1}{n} \sum_{t=1}^{n} \hat{D}^2(t)} \]
The most important ANOVA results for the gridded annual mean temperature series for the different station systems computed for the common time period 1975-2020

<table>
<thead>
<tr>
<th></th>
<th>MISH1</th>
<th>MISH2</th>
<th>MISH3</th>
<th>MISH4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mean</td>
<td>10.47</td>
<td>10.46</td>
<td>10.47</td>
<td>10.46</td>
</tr>
<tr>
<td>Total standard deviation</td>
<td>3.25</td>
<td>3.23</td>
<td>3.24</td>
<td>3.23</td>
</tr>
<tr>
<td>Spatial standard deviation of temporal means</td>
<td>1.04</td>
<td>1.10</td>
<td>1.07</td>
<td>1.08</td>
</tr>
<tr>
<td>Root spatial mean of temporal variances</td>
<td>0.62</td>
<td>0.72</td>
<td>0.66</td>
<td>0.69</td>
</tr>
<tr>
<td>Temporal standard deviation of spatial means</td>
<td>0.83</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>Root temporal mean of spatial variances</td>
<td>0.82</td>
<td>0.83</td>
<td>0.83</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Spatial mean series of annual mean temperature for the different systems (in °C)

Spatial standard deviation series of annual mean temperature for the different systems (in °C)
Temporal mean values of annual mean temperature (top, in °C) and temporal standard deviation values of annual mean temperature (bottom, in °C) for the period 1975-2020 for the four different station systems
The difference between the temporal mean values of gridded annual mean temperature based on 114 or 11 stations for the common period 1975-2020 (in °C)
Spatial mean series of annual mean temperature for Hungary from 1870 to 2020 with the estimated linear trend
The steps of homogenization of precipitation station data series

1. **MASH1**: homogenization of monthly data
2. Cut out and insert the inhomogeneities of the common part into MASH2 and MASH3
3. **MASH2**: homogenization of monthly data
4. Cut out the inhomogeneities of the common part and insert them into MASH3 and MASH1
5. **MASH3**: homogenization of monthly data
6. Cut out the inhomogeneities of the common part and insert them into MASH1 and MASH2
7. If statistics are acceptable in MASH1: go to 8, if not, go to step 1.
8. If the statistics are acceptable in MASH2: go to 9, if not, go to step 3.
9. Homogenization of daily data in MASH1, MASH2 and MASH3
10. Gathering the homogenized data sets from the different MASH systems.
The most important verification statistics for annual precipitation sum

<table>
<thead>
<tr>
<th>MASH</th>
<th>1870-2020</th>
<th>1901-2020</th>
<th>1951-2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significance level: 0.01</td>
<td>Critical value: 28.00</td>
<td>Critical value: 28.00</td>
<td>Critical value: 30.00</td>
</tr>
<tr>
<td>Test statistics before homogenization</td>
<td>45.46</td>
<td>63.42</td>
<td>42.19</td>
</tr>
<tr>
<td>Test statistics after homogenization</td>
<td>18.79</td>
<td>27.91</td>
<td>26.75</td>
</tr>
<tr>
<td>Relative modification of series</td>
<td>0.23</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>Representativity of station network</td>
<td>0.46</td>
<td>0.63</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Gridding by MISH

Modeled parameters (previous results)

Homogenized data series MASH1

Homogenized data series MASH2

Homogenized data series MASH3

Gridded data series MISH1 1870-1900

Gridded data series MISH2 1901-1950

Gridded data series MISH3 1951-2020

Gridded data series 1870-2020
The most important ANOVA results for the gridded annual precipitation sum series for the different station systems

<table>
<thead>
<tr>
<th></th>
<th>MISH1</th>
<th>MISH2</th>
<th>MISH3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mean</td>
<td>591.42</td>
<td>603.74</td>
<td>601.65</td>
</tr>
<tr>
<td>Total standard deviation</td>
<td>132.51</td>
<td>135.02</td>
<td>135.48</td>
</tr>
<tr>
<td>Spatial standard deviation of temporal means</td>
<td>68.85</td>
<td>66.23</td>
<td>65.46</td>
</tr>
<tr>
<td>Root spatial mean of temporal variances</td>
<td>113.22</td>
<td>117.66</td>
<td>118.62</td>
</tr>
<tr>
<td>Temporal standard deviation of spatial means</td>
<td>99.05</td>
<td>101.38</td>
<td>101.07</td>
</tr>
<tr>
<td>Root temporal mean of spatial variances</td>
<td>88.02</td>
<td>89.17</td>
<td>90.22</td>
</tr>
</tbody>
</table>
Spatial mean series of annual precipitation sum for the different systems (in mm)

Spatial standard deviation series of annual precipitation sum for the different systems (in mm)
Temporal mean values of annual precipitation sum (top, in mm) and temporal standard deviation values of annual precipitation sum (bottom, in mm) for the period 1975-2020 for the three different station systems.
The difference between the temporal mean values of gridded annual precipitation sum based on 461 or 11 stations for the common period 1975-2020 (in mm)
Spatial mean series of annual precipitation sum for Hungary relative to the 1991-2020 normal, from 1870 to 2020.
Summary

Creation of a representative climatological database:

1. Homogenization, quality control, missing value completion with software MASH, based on adequate mathematical base!

2. Interpolation (gridding) with software MISH, based on adequate mathematical base! MISH was developed specifically for the interpolation of meteorological elements.

3. Finally the four gridded datasets are merged to a common database by taking grid point values from the dataset which consists of most stations in the given time period.
Conclusion

• Homogenization can be said to be successful in both cases, the datasets can be considered homogeneous at the appropriate level of significance. (Temperature 0.05, Precipitation 0.01 significance level)

• Analyzing the ANOVA results of both the daily precipitation sum and the mean temperature data series, it can be seen that the different systems are the same as the national average due to the modeling with the long homogenized series. The standard deviations are slightly different, as we reduce them by averaging.
References

• Izsák, B., Szentimrey, T. (2020): To what extent does the detection of climate change in Hungary depend on the choice of statistical methods?. Int J Geomath 11, 17
• https://library.wmo.int/index.php?lvl=notice_display&id=21861#.YHAv2j-8qUm