Statistical modelling of the present climate by the interpolation method MISH
- theoretical considerations -

Tamás Szentimrey

Varimax Limited Partnership

Budapest

Theoretical considerations

- The climate can be formulated as the probability distribution of the meteorological events or variables.

- The purpose of the statistical climatology is to estimate or model the climate probability distribution or equivalently the climate statistical parameters.

- Furthermore the meteorological data series make possible to estimate or model the climate statistical parameters in accordance with the establishments of statistical climatology principles.

- Our method MISH (Meteorological Interpolation based on Surface Homogenized Data Basis; Szentimrey and Bihari) was developed for spatial interpolation of meteorological elements.

- According to the mathematical theorems the optimal interpolation parameters are known functions of certain climate statistical parameters, which fact means we could interpolate optimally if we knew the climate.

- Consequently according to the principles of climatology the modelling part of software MISH is based on long meteorological data series.

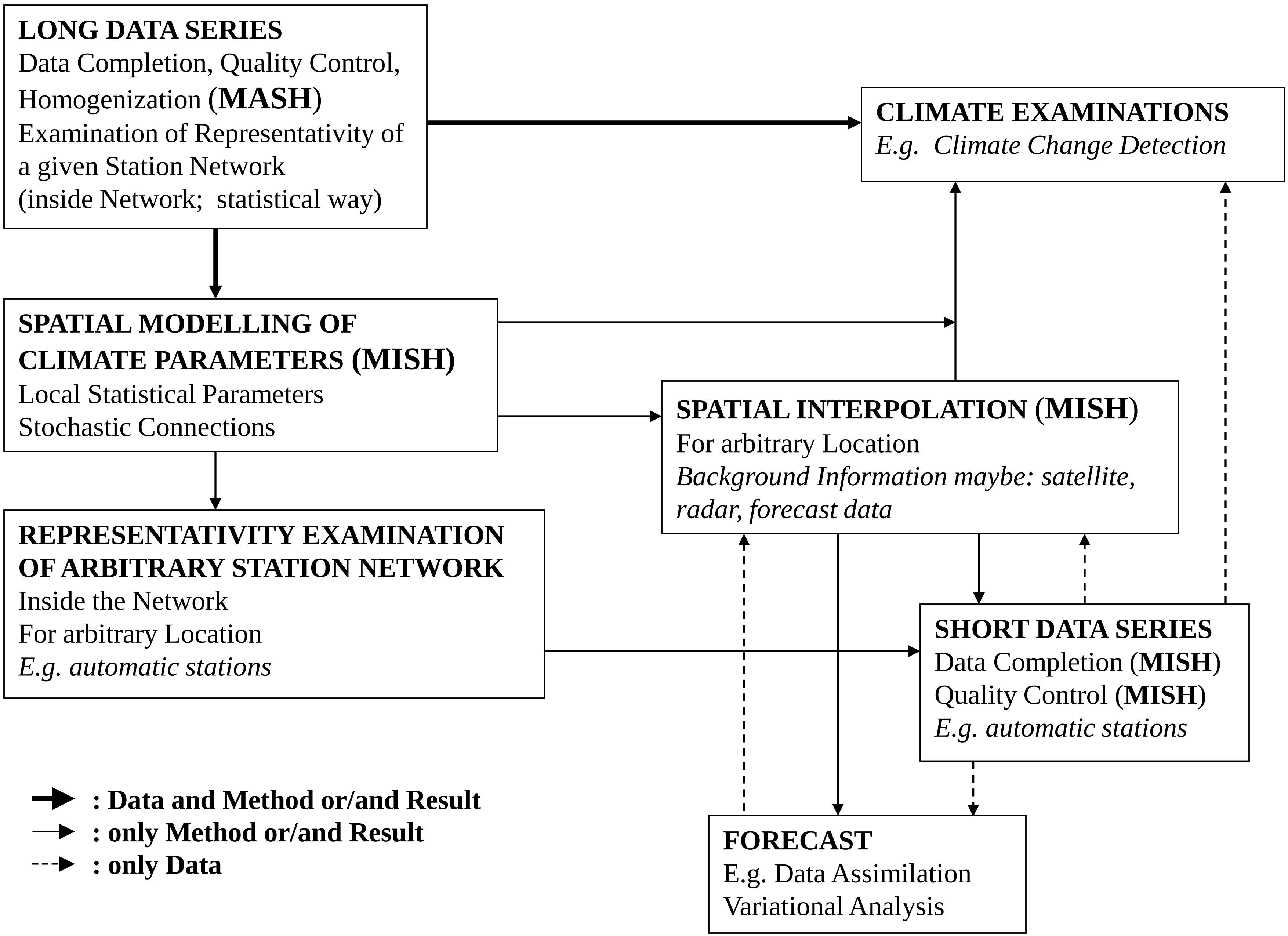
- The main difference between MISH and the geostatistical interpolation methods built in GIS is that the sample for modelling at GIS methods is only the predictors.

- We focus the methodology of the modelling subsystem built in MISH.

- This subsystem was developed to model the following climate statistical parameters for half minutes grid: monthly, daily expected values, standard deviations and the spatial, temporal correlations.

- Consequently the modelling subsystem of MISH is completed for all the first two spatiotemporal moments.

Possible Connection of Topics and Systems



Additive model of spatial interpolation (normal distribution, temperature)

Daily or monthly mean data for a given date

Predictand: $Z(\mathbf{s}_0)$ Predictors: $Z(\mathbf{s}_i)\ (i = 1,\dots,M)$

Statistical Parameters

Expected values (spatial trend): $E(\mathbf{s}_i) = \mathrm{E}(Z(\mathbf{s}_i))\ (i = 0,\dots,M)$

$\left(\text{Temporal trend: } \mathrm{E}(Z(\mathbf{s}_i,t)) = E(\mathbf{s}_i) + \mu(t)\ (\text{year } t;\, i = 0,\dots,M)\right)$

Standard deviations: $D(\mathbf{s}_i) = \mathrm{D}(Z(\mathbf{s}_i))\ (i = 0,\dots,M)$

$\mathbf{r}$: predictand-predictors correlation vector,

$\mathbf{R}$: predictors-predictors correlation matrix.

Additive (Linear) Interpolation

Interpolation Formula:

$$\hat{Z}(\mathbf{s}_0) = \lambda_0 + \sum_{i=1}^{M} \lambda_i \cdot Z(\mathbf{s}_i) \ , \quad \text{where} \ \sum_{i=1}^{M} \lambda_i = 1 \ .$$

Root Mean Square Error: $RMSE(\mathbf{s}_0) = \sqrt{\mathrm{E}\left(\left(Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)\right)^2\right)}$

Representativity Value: $REP(\mathbf{s}_0) = 1 - \dfrac{RMSE(\mathbf{s}_0)}{D(\mathbf{s}_0)}$

Optimal Interpolation Parameters : $\lambda_i \ (i = 0,\ldots,M)$

minimize RMSE.

Theorem 1: the Optimal Interpolation Parameters are known functions of climate statistical parameters!

Optimal constant term: $\lambda_0 = \sum_{i=1}^{M} \lambda_i \left(E(\mathbf{s}_0) - E(\mathbf{s}_i) \right)$

Vector of optimal weighting factors: $\boldsymbol{\lambda} = [\lambda_1, .., \lambda_M]^{\mathrm{T}}$

Vector $\boldsymbol{\lambda}$ can be written as function of parameters:

$D(\mathbf{s}_0)/D(\mathbf{s}_i) \;\; (i = 1,..., M), \; \mathbf{r}, \; \mathbf{R}$.

Conclusion

The expected values (spatial trend), the standard deviations and the correlations (stochastic part) are climate statistical parameters in meteorology. That means:

We could interpolate optimally if we knew the climate well!

Modelling of Monthly Climate Statistical Parameters

The obtained optimal interpolation formula:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{M} \lambda_i \left(E(\mathbf{s}_0) - E(\mathbf{s}_i) \right) + \sum_{i=1}^{M} \lambda_i Z(\mathbf{s}_i)$$, where the weighting factors:

$$\boldsymbol{\lambda} = \mathbf{R}^{-1} \left(\mathbf{r} + \frac{\left(1 - \mathbf{1}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{r}\right)}{\mathbf{1}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{1}} \mathbf{1} \right), \quad \text{if} \quad D(\mathbf{s}_0) / D(\mathbf{s}_i) \approx 1 \quad (i = 1, \ldots, M)$$

Unknown statistical parameters: $E(\mathbf{s}_0) - E(\mathbf{s}_i)$ $(i = 1, \ldots, M)$ (do not depend on temporal trend of climate change) and correlations $\mathbf{r}$, $\mathbf{R}$.

Modelling: can be based on long monthly mean data series $Z(\mathbf{S}_k, t)$ $(t = 1, \ldots, n)$ belonging to the stations $\mathbf{S}_k$ $(k = 1, \ldots, K)$.

Sample in space and time!

Difference between Geostatistics and Meteorology

Amount of information for modelling the statistical parameters.

Geostatistics

Information: only the actual predictors $Z(\mathbf{s}_i)$ $(i = 1,..., M)$.

Single realization in time!

Meteorology

Information: Stations with long data series. Sample in space and time!

Consequently the climate statistical parameters in question (expectations, standard deviations, correlations) for the stations are essentially known.

Much more information for modelling!

Modelling of Monthly Climate Statistical Parameters
(for a half minutes grid)

The monthly climate statistical parameters belonging to the stations $\mathbf{S}_k \ (k = 1,..,K)$ can be used for modelling the correlation structure as well as the spatial variability of local statistical parameters. The basic principle is as follows. Let $P(\mathbf{s}),\ Q(\mathbf{s}),\ r(\mathbf{s}_1, \mathbf{s}_2)$ $(\mathbf{s}, \mathbf{s}_1, \mathbf{s}_2 \in D)$ be certain model functions depending on different model variables with the following properties:

(a) Modelling of correlations: $r(\mathbf{S}_{j1}, \mathbf{S}_{j2}) \approx \mathrm{corr}\left(Z(\mathbf{S}_{j1}), Z(\mathbf{S}_{j2})\right) \quad (j_1, j_2 = 1,..,K)$

(b) Modelling of difference of means (*E*): $P(\mathbf{S}_{j1}) - P(\mathbf{S}_{j2}) \approx E(\mathbf{S}_{j1}) - E(\mathbf{S}_{j2})$

(c) Modelling of ratio of st. deviations (*D*): $\dfrac{Q(\mathbf{S}_{j1})}{Q(\mathbf{S}_{j2})} \approx \dfrac{D(\mathbf{S}_{j1})}{D(\mathbf{S}_{j2})}$

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

Modelling of Monthly Climate Statistical Parameters by Interpolation
(for a half minutes grid)

Predictand location: $\mathbf{s}_0$ Predictor station locations: $\mathbf{S}_{0i}\ \left(i=1,\ldots,M\right)$

The weighting factors: $\boldsymbol{\lambda}=\mathbf{R}^{-1}\left(\mathbf{r}+\dfrac{\left(1-\mathbf{1}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}\right)}{\mathbf{1}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{1}}\mathbf{1}\right)$, where $\mathbf{r}$, $\mathbf{R}$ contain modelled predictand-predictors, predictors-predictors correlations.

Modelling of means, expected values (E) by additive interpolation:

$$E\left(\mathbf{s}_0\right)=\sum_{i=1}^{M}\lambda_i\left(P\left(\mathbf{s}_0\right)-P\left(\mathbf{S}_{0i}\right)\right)+\sum_{i=1}^{M}\lambda_i E\left(\mathbf{S}_{0i}\right)$$

Modelling of st. deviations (D) by multiplicative interpolation:

$$D\left(\mathbf{s}_0\right)=\prod_{i=1}^{M}\left(\frac{Q\left(\mathbf{s}_0\right)}{Q\left(\mathbf{S}_{0i}\right)}\cdot D\left(\mathbf{S}_{0i}\right)\right)^{\lambda_i}$$

Representativity and interpolation error RMSE

(to charactarize quantitatively the uncertainties of interpolation)

$$REP(\mathbf{s}_0) = 1 - \frac{RMSE(\mathbf{s}_0)}{D(\mathbf{s}_0)}$$ depends on the parameters:

$$D(\mathbf{s}_0)/D(\mathbf{s}_i) \ \left(i = 1,\dots, M \right), \ \mathbf{r}, \ \mathbf{R}.$$

If $D(\mathbf{s}_0)/D(\mathbf{s}_i) = 1 \ \left(i = 1,\dots, M \right)$ then,

$$REP(\mathbf{s}_0) = 1 - \sqrt{\left(1 - \mathbf{r}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{r}\right) + \left(1 - \mathbf{1}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{r}\right)^2 \cdot \frac{1}{\mathbf{1}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{1}}}$$

$$RMSE(\mathbf{s}_0) = D(\mathbf{s}_0) \cdot \left(1 - REP(\mathbf{s}_0)\right)$$

Theorem 2

Let us assume for the daily values within a month:

i, Expected values and standard deviations:

$$E_t(\mathbf{s}_0) - E_t(\mathbf{s}_i) = e_{0i}, \quad D_t(\mathbf{s}_0)/D_t(\mathbf{s}_i) = d_{0i} \quad \left(i = 1,\ldots,M;\ t = 1,..,30\right)$$

ii, Correlations: $\operatorname{corr}\left(Z_{t_1}(\mathbf{s}_{i_1}), Z_{t_2}(\mathbf{s}_{i_2})\right) = r^S_{i_1 i_2} \cdot r^T_{t_1 t_2} \quad \left(i_1, i_2 = 1,.,M;\ t_1, t_2 = 1,.,30\right)$

$r^S_{i_1 i_2}$: correlation structure in space, $r^T_{t_1 t_2}$: correlation structure in time.

Then the Spatial Correlation Structure and the Optimum Interpolation Parameters for the daily and monthly mean values are identical:

$$\lambda_{i,t} = \lambda_{i,month} \quad \left(i = 0,\ldots,M;\ t = 1,..,30\right).$$

Moreover the Representatvity Values for the daily and monthly mean values are also identical: $REP_t(\mathbf{s}_0) = REP_{month}(\mathbf{s}_0) \ \left(t = 1,..,30\right)$.

Special modelling parts for daily data (for a half minutes grid)

Modelling of temporal daily autocorrelations $\rho(\mathbf{s})$ and daily standard deviations $D_{daily}(\mathbf{s})$ per months.

Let us assume the daily data of a given month constitute an AR(1) process with common standard deviation $D_{daily}(\mathbf{s})$ and temporal first-order autocorrelation $\rho(\mathbf{s})$.

Modelling of autocorrelation $\rho(\mathbf{s})$ by additive interpolation:

$$\rho(\mathbf{s}_0) = \sum_{i=1}^{M} \lambda_i \rho(\mathbf{S}_{0i})$$

where autocorrelations $\rho(\mathbf{S}_{0i})$ belonging to the stations.

Then $D_{daily}(\mathbf{s})$ can be estimated by using the monthly standard deviation $D_{month}(\mathbf{s})$: $D_{daily}(\mathbf{s}) \approx \sqrt{30 \cdot \frac{1-\rho(\mathbf{s})}{1+\rho(\mathbf{s})}} \cdot D_{month}(\mathbf{s})$

Modelled monthly, daily spatiotemporal statistical parameters in MISH
(for a half minutes grid)

i. Spatial expected values (spatial trend) $E(\mathbf{s})$
ii. Spatial standard deviations $D(\mathbf{s})$
iii. Spatial correlations $r(\mathbf{s}_1, \mathbf{s}_2)$
iv. Temporal first-order autocorrelations $\rho(\mathbf{s})$

Consequently the first two spatiotemporal moments can be modelled for daily and monthly data by MISH! The normal distribution is uniquely determined by these moments.

Interpolation applications for monthly and daily data

$$\hat{Z}(\mathbf{s}_0) = \lambda_0 + \sum_{i=1}^{M} \lambda_i \cdot Z(\mathbf{s}_i) \ , \quad RMSE(\mathbf{s}_0) = D(\mathbf{s}_0) \cdot (1 - REP(\mathbf{s}_0))$$

The Optimum Interpolation Parameters $\lambda_i \ (i = 0,...,M)$, St. Deviation $D(\mathbf{s}_0)$ and Representativity Value $REP(\mathbf{s}_0)$ can be calculated from the above modelled parameters.

Modelling of monthly, daily spatiotemporal statistical parameters in MISH

Modelling is based on long station data series. Sample in space and in time!
There is a substantial difference between Geostatistics and Meteorology
that is the amount of information for modelling the statistical parameters.

Spatial Interpolation by Modelled Climate Statistical Parameters

&

Climate Modelling by Interpolation of Station Climate Statistical Parameters

Conclusion:

We should know the present climate well! (E.g. data assimilation?)

Not the future climate only!

Special advanced MATHEMATICS is needed!

Example for Modelling by MISH (results are on a half minutes grid that can be downloaded)

Mean temperature in September for 10 **arbitrary** locations somewhere in Hungary.

Input: the location coordinates only without any temperature data.

Output: modelled climate statistical parameters

```
Location indices:
      1        2        3        4        5        6        7        8        9       10

 Monthly Expected Values:
  14.59   14.99   14.95   15.06   15.16   15.16   15.13   15.08   15.01   15.05

 Daily Expected Values:
  14.59   14.99   14.95   15.06   15.16   15.16   15.13   15.08   15.01   15.05

 Monthly Standard Deviations:
   1.34    1.62    1.68    1.67    1.68    1.66    1.72    1.66    1.61    1.64

 Daily Standard Deviations:
   2.84    3.44    3.47    3.46    3.47    3.60    3.73    3.58    3.48    3.46

 Temporal Daily Autocorrelations:
   0.74    0.74    0.75    0.75    0.75    0.73    0.73    0.73    0.73    0.74

 Matrix of Spatial Correlations:
   1.00    0.99    0.99    0.98    0.97    0.96    0.97    0.97    0.98    0.98
   0.99    1.00    0.99    0.99    0.98    0.95    0.96    0.96    0.97    0.98
   0.99    0.99    1.00    0.99    0.99    0.94    0.95    0.95    0.96    0.97
   0.98    0.99    0.99    1.00    0.99    0.91    0.93    0.93    0.95    0.96
   0.97    0.98    0.99    0.99    1.00    0.90    0.91    0.91    0.93    0.94
   0.96    0.95    0.94    0.91    0.90    1.00    0.99    0.99    0.98    0.98
   0.97    0.96    0.95    0.93    0.91    0.99    1.00    0.99    0.99    0.98
   0.97    0.96    0.95    0.93    0.91    0.99    0.99    1.00    0.99    0.99
   0.98    0.97    0.96    0.95    0.93    0.98    0.99    0.99    1.00    0.99
   0.98    0.98    0.97    0.96    0.94    0.98    0.98    0.99    0.99    1.00
```

The main features of MISHv2.01

(under development; last shared version MISHv1.03)

I. Modelling system for climate statistical parameters in space

(expected values, standard deviations, spatiotemporal correlations)

- Based on long homogenized data series and model variables.
- Modelling procedure must be executed only once before the interpolation applications.

II. Spatial 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, annual values and many years’ means can be interpolated.
- The expected interpolation error RMSE is modelled too.
- Real time Quality Control for daily and monthly data (additive model).
- Capability for application of background information such as satellite, radar forecast data. (with QC: data assimilation)
- Capability for gridding of data series.

There is no royal road!

(Archimedes)

Thank you for your attention!

Multiplicative Interpolation Formula of MISH

Optimum Interpolation Formula depends on the probability distribution.

Multiplicative Formula based on lognormal distribution for precipitation sum:

Predictand: $Z(\mathbf{s}_0, t)$ Predictors: $Z(\mathbf{s}_i, t) \quad (i = 1,..,M)$

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

where $\vartheta > 0$, $q_i > 0$, $\sum_{i=1}^{M} \lambda_i = 1$ and $\lambda_i \geq 0 \quad (i = 1,...,M)$,

are the interpolation parameters.

The optimum interpolation parameters are uniquely determined by certain climate statistical parameters.