Introduction on homogenization, quality control, spatial interpolation, gridding

Tamás Szentimrey

Hungarian Meteorological Service

Background

The first eight Seminars for Homogenization and Quality Control as well as the first three Conferences on Spatial Interpolation were held in Budapest and hosted by HMS and supported by WMO.

The specialty of both series was the **Mathematical Methodology**!

In 2014 the 8th Homogenization Seminar and the 3rd Interpolation Conference were organized together considering certain theoretical and practical aspects.

Theoretically there is a strong connection between these topics since the homogenization and quality control procedures need spatial statistics and interpolation techniques for spatial comparison of data. On the other hand the spatial interpolation procedures (e.g. gridding) require homogeneous, high quality data series to obtain good results.

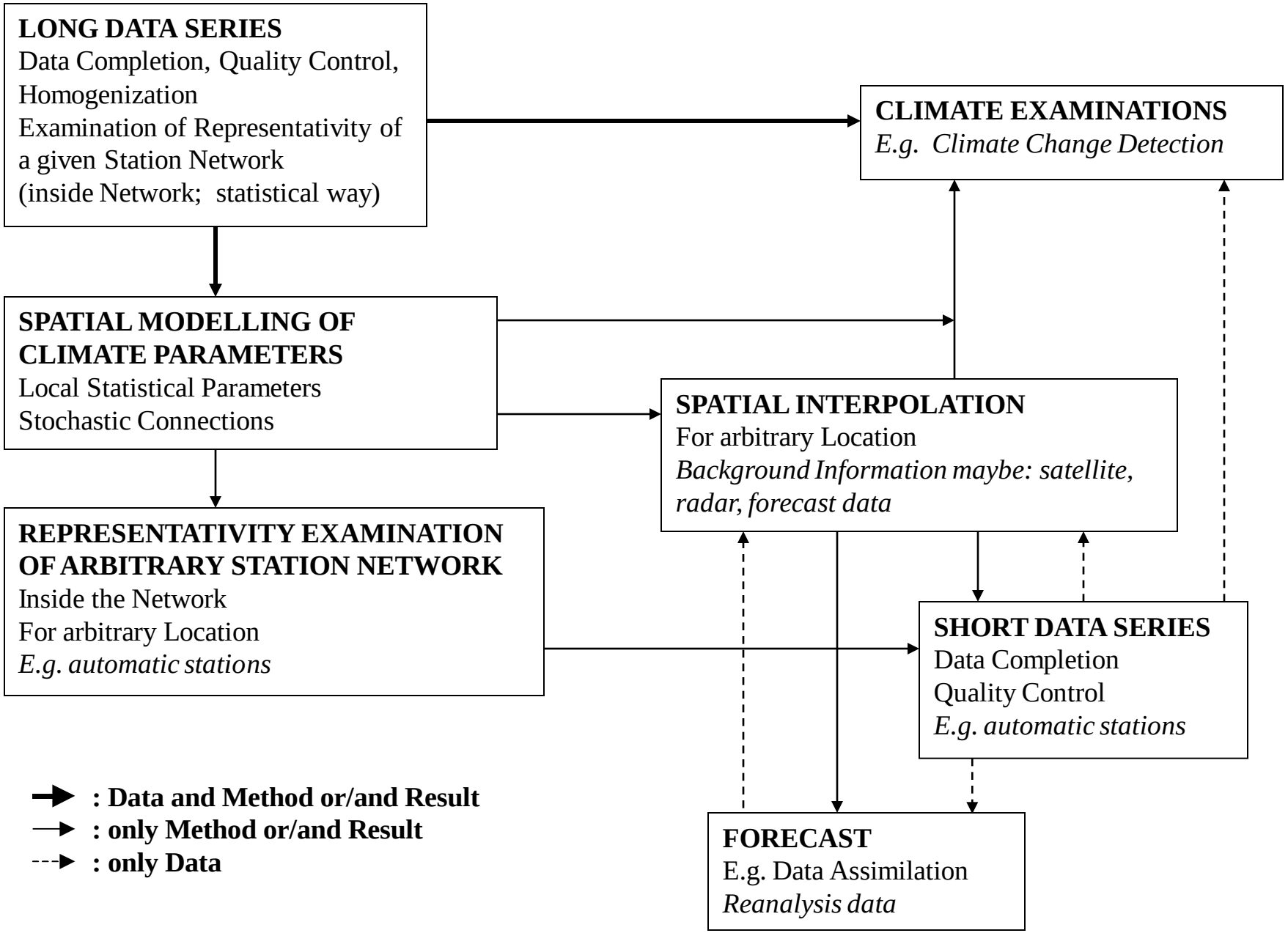
The main topics of homogenization and quality control are the following:

- Theoretical, mathematical questions. There is not any exact mathematical theory of the homogenization.
- Relation of monthly and daily homogenization, mathematical formulation of homogenization for climate data series generally.
- Methods for homogenization and quality control (QC) of monthly data series, missing data completion.
- Spatial comparison of series, inhomogeneity detection, correction of series.
- Methods for homogenization and quality control (QC) of daily data series, missing data completion, examination of parallel measurements.
- Usage of metadata.
- Manual versus automatic methods.
- Theoretical evaluation and benchmark for methods, validation statistics.
- Applications of different homogenization and quality control methods, experiences with different meteorological variables.

The main topics of spatial interpolation are the following:

- Theoretical, mathematical questions.
- Interpolation formulas and loss functions depending on the spatial probability distribution of meteorological variables.
- Estimation and modelling of statistical parameters (e.g.: spatial trend, covariance or variogram) for interpolation formulas using spatiotemporal sample and auxiliary model variables (topography).
- Characterization, modelling of interpolation error.
- Real time data quality control (QC) procedures based on spatial comparison, interpolation.
- Use of auxiliary co-variables, background information (e.g.: forecast, satellite, radar data) for spatial interpolation, relation with data assimilation, reanalysis.
- Applications of different interpolation methods for the meteorological and climatological data, experiences with different meteorological variables.
- Gridding of data series, gridded databases.

Possible Connection of Topics and Systems



Schema of Meteorological Examinations

1. **Meteorology:** Qualitative formulation of the problem.

2. **Mathematics:** Quantitative formulation of the problem.

3. **Software:** Based on Mathematics.

4. **Meteorology:** Application of Software.

In general the Mathematics is neglected!

Mathematics of homogenization of climate data series?

There are several methods and software in meteorology but

- there is no exact mathematical theory of homogenization!

Moreover,

- the mathematical formulation is neglected in general,

-“mathematical statements” without proof are in the papers,

- unreasonable dominance of the practice over the theory.

No solution without advanced mathematics!

Spatial Interpolation Mathematics for Meteorology?

– Nowadays the geostatistical interpolation methods built in **GIS** are applied in meteorology.

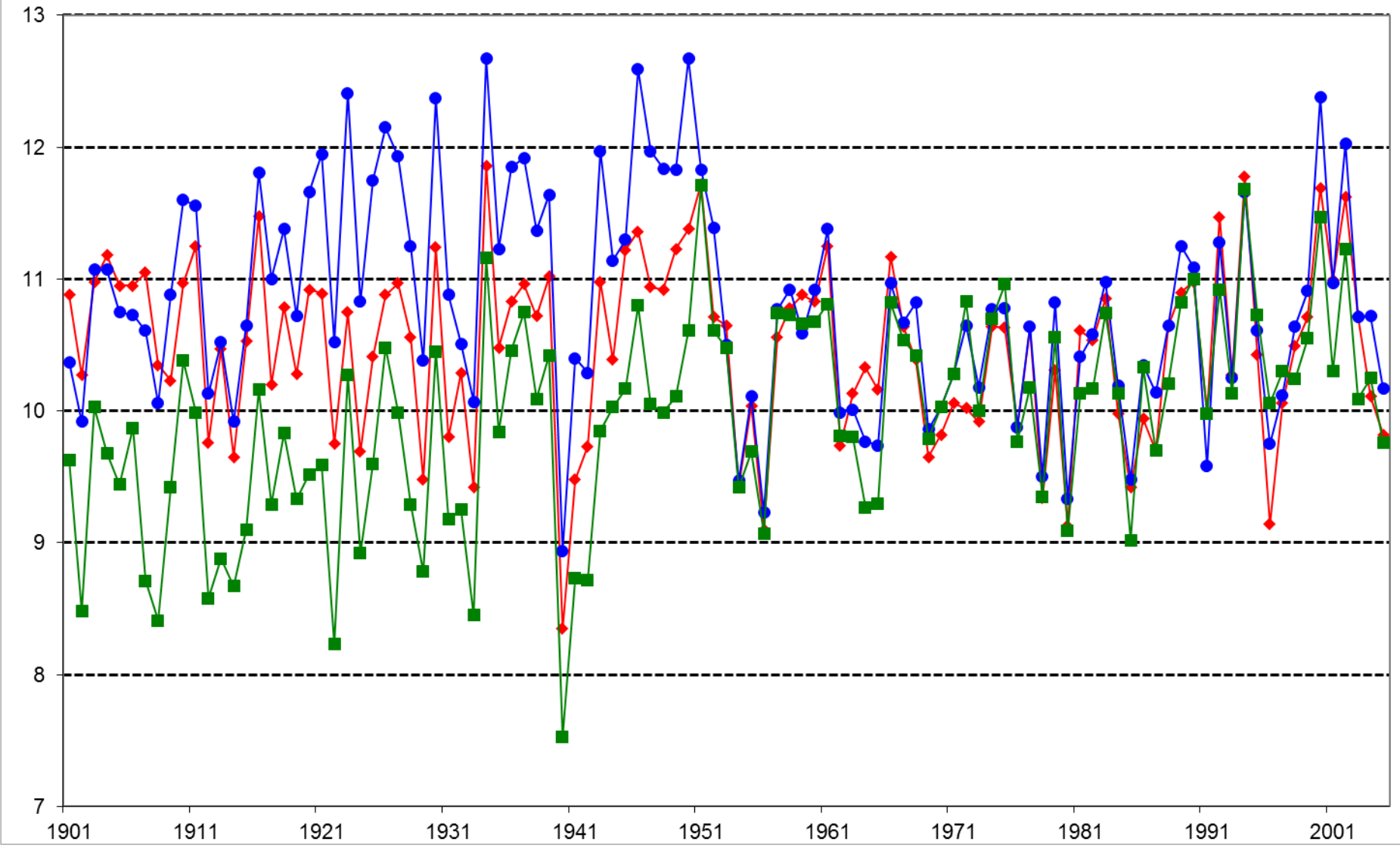
– The exact mathematical basis of the geostatistical interpolation methods: **Geostatistics**

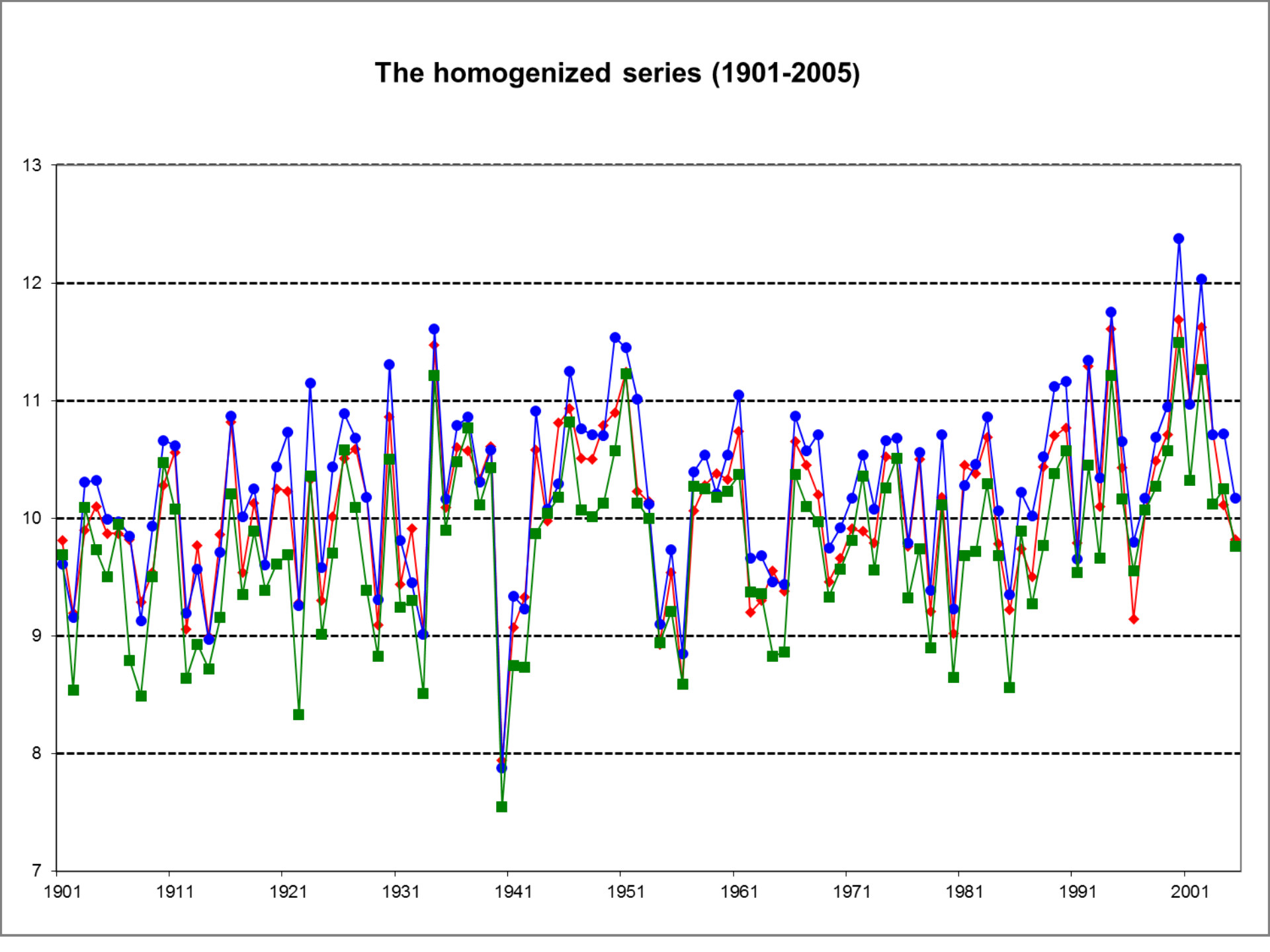
– But the geostatistical methods can not efficiently use the meteorological data series.

– While the data series make possible to obtain the necessary climate information.

HOMOGENIZATION

Annual mean temperature series of 3 stations in Hungary (1901-2005)




The homogenized series (1901-2005)
13
12
11
10
9
8
7
1901
1911
1921
1931
1941
1951
1961
1971
1981
1991
2001

MATHEMATICAL FORMULATION OF HOMOGENIZATION

(Distribution problem)

Let us assume we have daily or monthly data series.

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

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

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

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

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

Theoretical cumulative distribution functions (CDF):

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

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

Theoretical formulation of homogenization

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

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

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

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

The correction formula: $Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right) \quad \left(t = 1,2,..,T\right)$

Problems

Estimation, detection of change point(s) T ?

Estimation of distribution functions $F_{1,t}(y)$, $F_{2,t}(y)$ $\left(t = 1,2,..,T\right)$?

i, $F_{1,t}(y)$, $F_{2,t}(y)$ change in time (annual cycle, climate change)

ii, No sample for $F_{1,t}(y)$ $\left(t = 1,2,..,T\right)$

The problem is insolvable in general case!

Only relative methods can be used with some assumptions.

Statistically speaking, some assumptions have to be made!

Relation of daily and monthly homogenization

If we have daily series the general way is,

- calculation of monthly series
- homogenization of monthly series (larger signal to noise ratio)
- homogenization of daily series based on monthly inhomogeneities

Question

How can we use the valuable information of estimated monthly inhomogeneities for daily data homogenization?

Overview on homogenization of monthly data in mean (normal distribution, temperature)

Statistical spatiotemporal modelling of the series

Relative models and methods

Methodology for comparison of series

Break point (changepoint) and outlier detection (QC)

Methodology for correction of series

Missing data completion

Usage of metada

Manual versus automatic methods

Relation of monthly, seasonal, annual series

Benchmark for methods

Statistical spatiotemporal modelling of monthly series

Relative Additive Model (e.g. temperature)

Monthly series for a given month in a small region:

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

μ : unknown climate change signal; E : spatial expected value;
IH : inhomogeneity signal in mean; ε : normal noise

Type of $\mu(t)$: No assumption about the shape of this signal

Type of inhomogeneity $IH(t)$ in general: 'step-like function'
with unknown break points T and shifts $IH(T) - IH(T+1)$.

Noise $\varepsilon(t) = [\varepsilon_1(t),\ldots,\varepsilon_N(t)]^T \in N(\mathbf{0}, \mathbf{C})$ $(t = 1,\ldots,n)$ are independent

$\mathbf{C}$: spatial covariance matrix, very important!

Methodology for comparison of series

Related to the questions: reference series creation,

difference series constitution, multiple comparison of series etc.

All the examined series $X_j(t)\ \left(j = 1,..., N\right)$:

candidate and reference series alike.

Reference series are not assumed to be homogeneous!

Aim: to filter out $\mu(t)$ and to increase signal to noise ratio (power)

The spatial covariance matrix $\mathbf{C}$ may have a key role in

methodology of comparison of series.

Break point (changepoint) detection

Examination (more) difference series to detect the break points and to attribute (separate) for the candidate series.

Key question of the homogenization software:

Automatic procedures for attribution of the break points for the candidate series!!!

Remark

What is the aim of the homogenization?

- It is not the precise break point detection. (tool)
- The aim is good estimation of the inhomogeneity $IH(t)$!!!!

Multiple break points detection for a difference series

Possibilities, principles for joint estimation of break points:

(Classical ways in mathematical statistics!)

a, Bayesian Aproach (model selection, segmentation), penalized likelihood methods

Example: HOMER (Caussinus&Mestre), ACMANT (Domonkos)

b, Multiple break points detection based on **Test of Hypothesis,** confidence intervals for the break points
(make possible automatic use of metadata)

Example: MASH (Szentimrey)

Methodology for correction of series

Examination of (difference) series for estimation of shifts (correction factors) at the detected break points.

Possibilities, principles

a, In general: **Point Estimation**

a1, Least-Squares (joint) estimation (ANOVA):
HOMER, ACMANT

a2, Maximum Likelihood method, Generalized-Least-Squares (joint) estimation (based on spatial covariance matrix **C**)

b, Estimation is based on **Confidence Intervals**
(Test of Hypothesis): MASH

Automation of methods and software

Manual versus interactive or automatic methods?

In the practice numerous stations series must be examined!

Flexible automatic systems are necessary wherein the mechanic, labour-intensive procedures must be automated.

But not pushing button systems! The problem is much more complex.

Key questions for the methods and software:

- quality of homogenized data
- quantity of stations (automation!)

Necessary conditions for automation of methods, software:

- automatic attribution of break points for the candidate series
- automatic use of metadata

Evaluation of the methods applied in practice

1. Theoretical evaluation

2. Benchmark (to test the methods)

However the benchmark results depend on:

- Methods (quality, manual or automatic?)
- Benchmark dataset (quality, adequacy?)
- Testers (skilled or unskilled?)
- Mathematics of evaluation (validition statistics?)

Additive model of Spatial Interpolation
(normal distribution, temperature)

Predictand: $Z(\mathbf{s}_0, t)$

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

(**s**: space, t: time)

Statistical Parameters

<u>Deterministic Parameters:</u>

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

<u>Linear meteorological model for expected values:</u>

$$\mathrm{E}(Z(\mathbf{s}_i, t)) = \mu(t) + E(\mathbf{s}_i) \quad (i = 0,.., M)$$

Temporal trend (unknown climate change): $\mu(t)$, Spatial trend: $E(\mathbf{s})$

Stochastic parameters

Covariance preferred in mathematical statistics and meteorology:

$\mathbf{c}$: predictand-predictors covariance vector

$\mathbf{C}$: predictors-predictors covariance matrix

Variogram preferred in geostatistics:

$\boldsymbol{\gamma}$: predictand-predictors variogram vector

$\boldsymbol{\Gamma}$: predictors-predictors variogram matrix

Additive (Linear) Interpolation

Interpolation Formula: $\hat{Z}(\mathbf{s}_0) = \lambda_0 + \sum_{i=1}^{M} \lambda_i \cdot Z(\mathbf{s}_i)$,

where $\sum_{i=1}^{M} \lambda_i = 1$, because of unknown $\mu(t)$.

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

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

minimize RMSE.

The Optimal Interpolation Parameters are known functions of 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: $\boldsymbol{\lambda} = [\lambda_1, .., \lambda_M]^{\mathrm{T}}$

i, $\boldsymbol{\lambda} = \mathbf{C}^{-1} \left(\mathbf{c} + \frac{\left(1 - \mathbf{1}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{c}\right)}{\mathbf{1}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{1}} \mathbf{1} \right)$ (covariance form)

ii, $\boldsymbol{\lambda} = \boldsymbol{\Gamma}^{-1} \left(\boldsymbol{\gamma} + \frac{\left(1 - \mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}\right)}{\mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} \mathbf{1}} \mathbf{1} \right)$ (variogram form)

Conclusion

The expected values (spatial trend) and the covariances (stochastic part) are climate statistical parameters in meteorology.

That means:

We could interpolate optimally if we knew the climate well!

Remark

Problematic formulas:

- Inverse Distance Weighting (IDW), $\lambda_0 = 0$ and $\lambda_i \left(i = 1,..., M \right)$ not optimal
- Ordinary kriging, $\lambda_0 = 0$

Adequate formulas:

- Universal kriging,
- Regression (residual, detrended) kriging

But in geostatistics: modelling of statistical parameters is based on only the actual predictors

Modelling of climate statistical parameters

The obtained optimal interpolation formula:

$$\hat{Z}(\mathbf{s}_0,t)=\sum_{i=1}^{M}\lambda_i\left(E(\mathbf{s}_0)-E(\mathbf{s}_i)\right)+\sum_{i=1}^{M}\lambda_i Z(\mathbf{s}_i,t) \quad ,$$

where the weighting factors: $\boldsymbol{\lambda}^{\mathrm{T}}=\left(\mathbf{c}^{\mathrm{T}}+\mathbf{1}^{\mathrm{T}}\frac{\left(1-\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}\right)}{\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}}\right)\mathbf{C}^{-1}$

Unknown statistical parameters: $E(\mathbf{s}_0)-E(\mathbf{s}_i)\left(i=1,...,M\right),\ \mathbf{c},\ \mathbf{C}$

Modelling: can be based on long station data series $Z(\mathbf{S}_k,t)\ \left(t=1,...,n\right)$ belonging to the stations $\mathbf{S}_k\ \left(k=1,..,K\right)$. Sample in space and in 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 in time!

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

Much more information for modelling!

Interpolation error RMSE
(to charactarize quantitatively the uncertainties of interpolation)

$$RMSE(\mathbf{s}_0) = \sqrt{\left(D^2(\mathbf{s}_0) - \mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}\right) + \left(1 - \mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}\right)^2 \cdot \frac{1}{\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}}}$$

Modelling of RMSE!

Real time Quality Control

Test schema of QC procedure at additive, normal model is:

$$\frac{Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)}{RMSE(\mathbf{s}_0)} \in N(0,1),$$

where $Z(\mathbf{s}_0)$ is the predictand to be controlled, $\hat{Z}(\mathbf{s}_0)$ is the interpolated value and $RMSE(\mathbf{s}_0)$ is the modelled interpolation error.

Interpolation with Background Information

Background information can decrease the interpolation error.
For example: forecast, satellite, radar data

$Z(\mathbf{s}_0, t)$: predictand

$\hat{Z}(\mathbf{s}_0, t) = \lambda_0 + \sum_{i=1}^{M} \lambda_i Z(\mathbf{s}_i, t)$: interpolation

$\mathbf{G} = \left\{ G(\mathbf{s}, t) \,\middle|\, \mathbf{s} \in D \right\}$: background information on a dense grid

<u>Principle of interpolation with Background Information</u>

$$\hat{Z}_G(\mathbf{s}_0, t) = \hat{Z}(\mathbf{s}_0, t) + \mathrm{E}\left(Z(\mathbf{s}_0, t) - \hat{Z}(\mathbf{s}_0, t) \,\middle|\, \mathbf{G} \right)$$

where $\mathrm{E}\left(Z(\mathbf{s}_0, t) - \hat{Z}(\mathbf{s}_0, t) \,\middle|\, \mathbf{G} \right)$ is the conditional

expectation of $Z(\mathbf{s}_0, t) - \hat{Z}(\mathbf{s}_0, t)$, given $\mathbf{G}$.

Reanalysis data

Based on Data Assimilation, variational analysis

Minimization of the variational cost function:

$$J(\mathbf{z}) = (\mathbf{z}-\mathbf{g})^{\mathrm{T}}\mathbf{Q}^{-1}(\mathbf{z}-\mathbf{g}) + (\mathbf{y}_0-\mathbf{F}\mathbf{z})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{y}_0-\mathbf{F}\mathbf{z}) \ ,$$

$\mathbf{z}$: analysis field, predictand (grid),

$\mathbf{g}$: background field (forecast), assumption $\mathrm{E}(\mathbf{z}|\,\mathbf{g}\,) = \mathbf{g}$,

$\mathbf{y}_0$: observations, predictors; $\mathbf{F}\mathbf{z} = \mathrm{E}(\mathbf{y}_0|\mathbf{z})$,

$\mathbf{Q}$, $\mathbf{P}$: covariance matrices

In essence:
Interpolation with background information + Quality control

Problem with Reanalysis data

i, Inhomogeneous predictor station data series

ii, Few stations, little spatial representativity

iii, Problem with the data assimilation formula:

- Lack of good climate statistical parameters in matrix **Q**

- Assumption: $\mathrm{E}\left(\mathbf{z}\middle|\mathbf{g}\right)=\mathbf{g}$?

Szentimrey, T. (2016): Analysis of the data assimilation methods from the mathematical point of view. In: *Mathematical Problems in Meteorological Modelling*, Springer International Publishing, Switzerland, 193–205

Importance of gridded databases with good quality!

- Homogenization of dense station data series
- Interpolation, gridding of homogenized series
- Comparison of gridded datasets?

Important question:

Homogeneity of satellite datasets?

There is no royal road!

Thank you for your attention!