IDŐJÁRÁS

QUARTERLY JOURNAL
OF THE HUNGARIAN METEOROLOGICAL SERVICE

Special Issue: Spatial interpolation techniques in climatology and meteorology
Guest Editor: **Sándor Szalai**

CONTENTS

****

http://www.met.hu/Journal-Idojaras.php

VOL. 115 * NO. 1–2 * JANUARY–JUNE 2011

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service

Editor-in-Chief

LÁSZLÓ BOZÓ

Executive Editor

MARGIT ANTAL

Editorial Office: Gilice tér 39, H-1182 Budapest, Hungary
P.O. Box 39, H-1675 Budapest, Hungary
E-mail: bozo.l@met.hu or *antal.e@met.hu*
Fax: (36-1) 346-4809

Indexed and abstracted in Science Citation Index Expanded™ and Journal Citation Reports/Science Edition
Covered in the abstract and citation database SCOPUS®

Subscription by

mail: IDŐJÁRÁS, P.O. Box 39, H-1675 Budapest, Hungary
E-mail: kenderesy.k@met.hu or *antal.e@met.hu*

Editorial

Spatial interpolation techniques in climatology and meteorology

Interpolation plays a growing role in the meteorology and climatology. Reconstruction of meteorological fields, developed data quality control procedures, and gridded databases require interpolation methods. The increasing needs indicate two directions for development. From one side, the diverse use of interpolation demands more accurate and complex methods, and from the other side, the common everyday's application has a request of simple useable software, usually as an option of free or commercial software. To overbridge this situation, a COST Action was implemented (COST 719: The Use of Geographic Information System in Climatology and Meteorology, end date 2006). The first Conference on the Spatial Interpolation Techniques in Climatology and Meteorology was organized in the frame of this Action. The proceedings were published by the COST Office.

Since there, several international projects deal with interpolation problems, at least partly. The request was arisen for an open meeting to overview the developments in the interpolation techniques. Therefore, the Hungarian Meteorological Service organized the conference second time in 2009.

The participants of the conference agreed, that the presentations could have a possibility to be published in a special issue of the quarterly journal of the Hungarian Meteorological Service, additionally to the abstract volume distributed widely among the other scientists working on the field of interpolation methods.

Finally, eight articles were gathered, and accepted for publication, which are covering wide range of topics on methodological issue, interpolation processes in international data bases, data quality control, applied research like hydrology, gridded data bases, interpolation in climate projections. These papers give about one-fourth of the conference presentations.

We strongly believe, that similar workshops and conferences are needed to avoid the misuse of interpolation method, understand and follow the development of interpolation methods, give new ideas for further scientific developments, involve new applied areas, and show new practices. Dissemination of best practices has benefit not only for the adopting, but the donor parties as well, they are for the common use and development.

Therefore, we are extremely grateful to the Editor-in-Chief of IDŐJÁRÁS supporting the progress on the field of interpolation, thank to the authors of the articles for their high scientific level work, and also to the reviewers supporting the improvement of papers with their critical comments and recommendations keeping the high standards of the journal. We have to underline the hard work of the Executive Editor of the journal, the present volume could not be published without it. Therefore, we express our thanks together with the authors of the papers for that.

Sándor Szalai
Guest Editor

Szent István University, Gödöllő, Hungary
szalai.sandor@mkk.szie.hu

IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 1–11

Mathematical, methodological questions concerning the spatial interpolation of climate elements

Tamás Szentimrey[*], **Zita Bihari, Mónika Lakatos,** and **Sándor Szalai**

Hungarian Meteorological Service,
P.O. Box 38, H-1525 Budapest, Hungary

[*]*Corresponding author; E-mail: szentimrey.t@met.hu*

(Manuscript received in final form January 13, 2011)

Abstract—The paper focuses on the basic mathematical and theoretical questions of spatial interpolation of meteorological elements. Nowadays, in meteorology the most often applied procedures for spatial interpolation are the geostatistical interpolation methods built also in GIS software. The mathematical basis of these methods is the geostatistics that is an exact but special part of the mathematical statistics. However, special meteorological spatial interpolation methods for climate elements also exist, such as Gandin optimum interpolation as well as the MISH method developed at the Hungarian Meteorological Service in the last few years. These meteorological interpolation methods are also based on the mathematical statistical theory. Therefore, the basic type of the interpolation formulas applied by the geostatistical and meteorological methods are similar. One of our intentions is to present some comparison of the various kriging formulas, such as ordinary, universal, regression, residual, detrended, etc., ones. In general, these formulas can be derived from the multiple linear regression formula by using the generalized-least-squares estimation for certain unknown parameters. But the main difference between the geostatistical and meteorological interpolation methods can be found in the amount of information used for modeling the necessary statistical parameters. In geostatistics, the usable information or the sample for modeling is only the system of predictors, which is a single realization in time, while in meteorology we have spatiotemporal data, namely the long data series which form a sample in time and space as well. The long data series is such a speciality of the meteorology that makes possible to model efficiently the statistical parameters in question.

Key-words: spatial interpolation, geostatistics, statistical climatology, data series, geostatistical interpolation, meteorological interpolation

1. Introduction

First let us consider the abstract scheme of the meteorological examinations. The initial stage is the meteorology that means the qualitative formulation of the given problem. The next stage is the mathematics in order to formulate the problem quantitatively. The third stage is to develop software on the basis of the mathematics. Finally, the last stage is again the meteorology that is the application of the developed software and evaluation of the obtained results. In the practice, however, the mathematics is sometimes neglected. Instead of adequate mathematical formulation of the meteorological problem, ready-made software are applied to solve the problem. Of course, in this case the results are not authentic either. Allow me a not word for word citation from John von Neumann: without quantitative formulation of the meteorological questions, we are not able to answer the simplest qualitative questions either.

Concerning our topic we have the following question. What kind of mathematics of spatial interpolation is adequate for meteorology? Nowadays, the geostatistical interpolation methods built in GIS software are applied in meteorology. The mathematical basis of these methods is the geostatistics that is an exact but special part of the mathematical statistics. The speciality is connected with the assumption that the data are purely spatial. To illustrate this problem, here are some quotations from the valuable book of Noel A.C. Cressie: "Statistics for Spatial Data" (*Cressie*, 1991). On page 29: "The first part of this book is concerned with modeling data as a (partial) realization of a random process $\{Z(\mathbf{s}):\mathbf{s}\in D\}$….". Explanation of the sentence is that the data are purely spatial data, since D is a space domain. On page 30: "It is possible to allow for spatiotemporal data by considering the variable $Z(\mathbf{s}, t)$, but for most of this book it will be assumed that the data are purely spatial…". Last, on page 53: "Statistically speaking, some further assumptions have to be made. Otherwise, the data represent an *incomplete* sampling of a *single* realization, making inference impossible." It means "*incomplete* sampling" in space, "*single* realization" in time.

Consequently, as we see it, the geostatistical methods can not efficiently use the meteorological data series, while the data series make possible to obtain the necessary climate information for the interpolation in meteorology.

2. Mathematical statistical model of spatial interpolation

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 $Z(\mathbf{s}_i, t)$ $(i=1,...,M)$, where the location vectors $\mathbf{s}$ are the elements of the given space domain D, and t is the time. The vector form of predictors is $\mathbf{Z}^{\mathrm{T}}(t)=[Z(\mathbf{s}_1,t),.....,Z(\mathbf{s}_M,t)]$. The type of the adequate interpolation formula depends on the probability distribution of the meteorological element in question. In this paper only the linear or additive formula is described in detail, which is appropriate in case of normal probability distribution. However, perhaps it is worthwhile to remark that for case of a quasi lognormal distribution (e.g., precipitation sum), we deduced a mixed additive multiplicative formula which is used also in our MISH system, and it can be written in the following form,

$$\hat{Z}(\mathbf{s}_0,t)=\vartheta\cdot\left(\prod_{q_i\cdot Z(\mathbf{s}_i,t)\geq\vartheta}\left(\frac{q_i\cdot Z(\mathbf{s}_i,t)}{\vartheta}\right)^{\lambda_i}\right)\cdot\left(\sum_{q_i\cdot Z(\mathbf{s}_i,t)\geq\vartheta}\lambda_i+\sum_{q_i\cdot Z(\mathbf{s}_i,t)<\vartheta}\lambda_i\cdot\left(\frac{q_i\cdot Z(\mathbf{s}_i,t)}{\vartheta}\right)\right),\qquad(1)$$

where the interpolation parameters are $\vartheta>0, q_i>0, \lambda_i\geq0\ (i=1,...,M)$, and $\sum_{i=1}^{M}\lambda_i=1$.

2.1. Statistical parameters

In general, the interpolation formulas have some unknown interpolation parameters which are known functions of certain statistical parameters. At the linear interpolation formulas the basic statistical parameters can be divided into two groups, such as the deterministic and the stochastic parameters.

The deterministic or local parameters are the expected values $\mathrm{E}(\mathbf{Z}(\mathbf{s}_i,t))(i=0,...,M)$. Let $\mathrm{E}(\mathbf{Z}(t))$ denote the vector of expected values of predictors, i.e., $\mathrm{E}(\mathbf{Z}(t))^{\mathrm{T}}=[\mathrm{E}(Z(\mathbf{s}_1,t)),....,\mathrm{E}(Z(\mathbf{s}_M,t))]$.

The stochastic parameters are the covariance or variogram values belonging to the predictand and predictors, such as

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

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

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

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

The covariance is preferred in mathematical statistics and meteorology, while the variogram is preferred in geostatistics. Here is a quotation from the chapter "Geostatistics" of the mentioned book of Noel A.C. Cressie (*Cressie*, 1991, p. 30.). "The cornerstone is the variogram, a parameter that in the past has been either unknown or unfashionable among statisticians." In our opinion, the main reason of this reluctance is that the covariance is a more general statistical parameter than the variogram. The variogram values, can be written as functions of the covariance values and it is not true inversely.

2.2. Linear meteorological model for expected values

At the statistical modeling of the meteorological elements we have to assume, that the expected values of the variables are changing in space and time alike. The spatial change means that the climate is different in the regions. The temporal change is the result of the possible global climate change. Consequently, in case of linear modeling of expected values, we assume that

$$\mathrm{E}(Z(\mathbf{s}_i,t))=\mu(t)+E(\mathbf{s}_i) \quad (i=0,\ldots,M), \tag{2}$$

where $\mu(t)$ is the temporal trend or the climate change signal and $E(\mathbf{s})$ is the spatial trend. We emphasize, that this spatiotemporal model for expected values is different from the classic models used in geostatistics or by the multivariate statistical methods. As regards the geostatistics, there are purely spatial data assumed in general.

2.3. Linear regression formula

In essence, the multiple linear regression formula is the theoretical basis of the various linear interpolation methods. The multiple linear regression formula between predictand $Z(\mathbf{s}_0,t)$ and predictors $\mathbf{Z}(t)$ can be written as

$$\hat{Z}_{LR}(\mathbf{s}_0,t)=\mathrm{E}(\mathrm{Z}(\mathbf{s}_0,t))+\mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z}(t)-\mathrm{E}(\mathbf{Z}(t))) \tag{3}$$

and $\hat{Z}_{LR}(\mathbf{s}_0,t)$ is the best linear estimation that minimizes the mean-square prediction error. Consequently, the linear regression formula would be the optimal linear interpolation formula concerning the mean-square prediction error. In respect of application, however, problems arise from the unknown statistical parameters $\mathrm{E}(\mathrm{Z}(\mathbf{s}_0,t))(i=0,\ldots,M)$ and $\mathbf{c}$, $\mathbf{C}$. Assuming the meteorological model, Eq. (2), for the expected values, Eq. (3) can be written as

$$\hat{Z}_{LR}(\mathbf{s}_0,t)=(\mu(t)+E(\mathbf{s}_0))+\mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z}(t)-(\mu(t)\mathbf{1}+\mathbf{E})), \tag{4}$$

where $\mathbf{E}^{\mathrm{T}}=[E(\mathbf{s}_1),\ldots..,E(\mathbf{s}_M)]$ and vector $\mathbf{1}$ is identically one. As it can be seen, the main problem is the estimation of the unknown climate change signal $\mu(t)$, if we want to apply the optimal linear regression interpolation formula.

3. Geostatistical interpolation methods

The various geostatistical interpolation formulas can be obtained from the linear regression formula, Eq. (3), by the application of the generalized-least-squares estimation for the expected values. The type of kriging formulas depends on the model assumed for the expected values.

3.1. Ordinary kriging formula

The ordinary kriging formula is a special case of the universal kriging formula. The assumed model for the expected values is $\mathrm{E}(Z(\mathbf{s}_i,t)) \equiv \mu(t)\,(i=0,...,M)$, thus, there is no spatial trend. The generalized-least-squares estimation for $\mu(t)$ by using only the predictors $\mathbf{Z}(t)$ may be expressed in the form $\hat{\mu}_{gls}(t) = (\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Z}(t)$. Substituting the estimate $\hat{\mu}_{gls}(t)$ into the linear regression formula, Eq. (3), we obtain the ordinary kriging formula as

$$\hat{Z}_{OK}(\mathbf{s}_0,t) = \hat{\mu}_{gls}(t) + \mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z}(t) - \hat{\mu}_{gls}(t)\mathbf{1}) = \sum_{i=1}^{M}\lambda_i Z(\mathbf{s}_i,t), \tag{5}$$

where $\sum_{i=1}^{M}\lambda_i = 1$.

The vector of weighting factors $\boldsymbol{\lambda}^{\mathrm{T}} = [\lambda_1,..,\lambda_M]$ can be written in covariance form

$$\boldsymbol{\lambda}^{\mathrm{T}} = \left(\mathbf{c}^{\mathrm{T}} + \mathbf{1}^{\mathrm{T}}\frac{(1-\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c})}{\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}}\right)\mathbf{C}^{-1}, \tag{6}$$

or equivalently in variogram form

$$\boldsymbol{\lambda}^{\mathrm{T}} = \left(\boldsymbol{\gamma}^{\mathrm{T}} + \mathbf{1}^{\mathrm{T}}\frac{(1-\mathbf{1}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{1}}\right)\boldsymbol{\Gamma}^{-1}. \tag{7}$$

The unknown variogram values $\boldsymbol{\gamma}$, $\boldsymbol{\Gamma}$ preferred in geostatistics are modeled according to the Section 3.3.

3.2. Universal kriging formula

The universal kriging formula is the generalized case of the ordinary kriging formula. The model assumption is that the expected values may be expressed as $\mathrm{E}(\mathrm{Z}(\mathbf{s}_i,t)) = \sum_{k=1}^{K}\beta_k(t)x_k(\mathbf{s}_i)\ (i=0,..., M)$, that is in vector form

$\mathrm{E}(Z(\mathbf{s}_0,t))=\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(t)$, $\mathrm{E}(\mathbf{Z}(t))=\mathbf{X}\boldsymbol{\beta}(t)$, where $\mathbf{x},\mathbf{X}$ are given supplementary deterministic model variables.

The generalized-least-squares estimation for coefficient vector $\boldsymbol{\beta}(t)$, by using only the predictors $\mathbf{Z}(t)$, can be written in the form $\hat{\boldsymbol{\beta}}_{gls}(t)=(\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Z}(t)$. It is to be remarked, that in this way the spatial trend $E(\mathbf{s})$ according to Eq. (2) is modeled also by using only the predictors $\mathbf{Z}(t)$. Substituting the estimates $\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{gls}(t)$, $\mathbf{X}\hat{\boldsymbol{\beta}}_{gls}(t)$ into the linear regression formula, Eq. (3), we obtain the universal kriging formula as

$$\hat{Z}_{UK}(\mathbf{s}_0,t)=\mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{gls}(t)+\mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z}(t)-\mathbf{X}\hat{\boldsymbol{\beta}}_{gls}(t))=\sum_{i=1}^{M}\lambda_i Z(\mathbf{s}_i,t), \tag{8}$$

where $\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{X}=\mathbf{x}^{\mathrm{T}}$.

The vector of weighting factors $\boldsymbol{\lambda}^{\mathrm{T}}=[\lambda_1,...,\lambda_M]$ can be written in covariance form

$$\boldsymbol{\lambda}^{\mathrm{T}}=\left\{\mathbf{c}+\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{X})^{-1}(\mathbf{x}-\mathbf{X}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c})\right\}^{\mathrm{T}}\mathbf{C}^{-1},$$

or equivalently in variogram form

$$\boldsymbol{\lambda}^{\mathrm{T}}=\left\{\boldsymbol{\gamma}+\mathbf{X}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}(\mathbf{x}-\mathbf{X}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})\right\}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}.$$

The unknown variogram values $\boldsymbol{\gamma}$, $\boldsymbol{\Gamma}$ preferred in geostatistics are modeled according to Section 3.3.

3.3. Modeling of unknown statistical parameters in geostatistics

In geostatistics, only the predictors $Z(\mathbf{s}_i,t)(i=1,...,M)$ constitute the usable information or the sample for modeling of variogram values $\boldsymbol{\gamma}$, $\boldsymbol{\Gamma}$. It means we have only a single realization in time for modeling of the statistical parameters in question. In order to solve the problem of absence of temporal data, some assumptions about the statistical structure are made that is some simplification of the problem. For example, such assumptions are the intrinsic stationarity or second-order (weak) stationarity, semivariogram $\gamma(Z(\mathbf{s}_i),Z(\mathbf{s}_j))=\gamma(\mathbf{s}_i-\mathbf{s}_j)$, etc.

4. Meteorological interpolation

Similarly to the geostatistical interpolation formulas, an appropriate meteorological interpolation formula can be obtained from the linear regression formula, Eq. (3), by the application of the generalized-least-squares estimation for the expected values. The key-question is the model assumption for the expected values.

4.1. Meteorological interpolation formula

The meteorological model, Eq. (2), is assumed namely $\mathrm{E}(\mathrm{Z}(\mathbf{s}_i,t))=\mu(t)+E(\mathbf{s}_i)\ (i=0,...,M)$, where $\mu(t)$ is the temporal trend and $E(\mathbf{s})$ is the spatial trend. Supposing that the spatial trend $E(\mathbf{s})$ is known, we apply the generalized-least-squares estimation for temporal trend $\mu(t)$ by using the predictors $\mathbf{Z}(t)$ and the spatial trend $\mathbf{E}^{\mathrm{T}}=[E(\mathbf{s}_1),.....,E(\mathbf{s}_M)]$. In this case, the generalized-least-squares estimate can be written in the form as $\hat{\mu}_{gls}^{E}(t)=(\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z}(t)-\mathbf{E})$. Substituting the estimate $\hat{\mu}_{gls}^{E}(t)$ into the linear regression formula, Eq. (4), rewritten from Eq. (3) according to Eq. (2), we obtain the following interpolation formula:

$$\hat{Z}_{MI}(\mathbf{s}_0,t)=(\hat{\mu}_{gls}^{E}(t)+E(\mathbf{s}_0))+\mathbf{c}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z}(t)-(\hat{\mu}_{gls}^{E}(t)\mathbf{1}+\mathbf{E}))=$$

$$=E(\mathbf{s}_0)+\sum_{i=1}^{M}\lambda_i(Z(\mathbf{s}_i,t)-E(\mathbf{s}_i)), \qquad (9)$$

where $\sum_{i=1}^{M}\lambda_i=1$.

The vector of weighting factors $\boldsymbol{\lambda}^{\mathrm{T}}=[\lambda_1,..,\lambda_M]$ can be written equivalently in covariance and variogram form according to Eqs. (6) and (7). The obtained interpolation formula is a detrended or residual interpolation formula that includes the spatial trend and the theoretical ordinary kriging weighting factors. However, it is not identical with the detrended or residual interpolation method, because the interpolation formula as well as the modeling methodology of the necessary statistical parameters together defines an interpolation method. For example, at the detrended interpolation methods applied in the practice, the modeling procedure for the statistical parameters is based on only the predictors $Z(\mathbf{s}_i,t)(i=1,...,M)$.

4.2. Possibility for modeling of unknown statistical parameters in meteorology

According to Eq. (9), where the sum of weighting factors is equal to one, we have the following appropriate meteorological interpolation formula

$$\hat{Z}_{MI}(\mathbf{s}_0,t)=\sum_{i=1}^{M}\lambda_i\left(E(\mathbf{s}_0)-E(\mathbf{s}_i)\right)+\sum_{i=1}^{M}\lambda_i Z(\mathbf{s}_i,t), \quad (10)$$

where $\sum_{i=1}^{M}\lambda_i=1$ and the covariance form of weighting factors is defined by Eq. (6). Consequently, the unknown statistical parameters are the spatial trend differences $E(\mathbf{s}_0)-E(\mathbf{s}_i)(i=1,...,M)$ and covariances $\mathbf{c},\mathbf{C}$. In essence, these parameters are climate parameters which in fact means that we could interpolate optimally if we knew the climate. The special possibility in meteorology is to use the long meteorological data series for modeling of the climate statistical parameters in question. The data series make possible to know the climate in accordance with the fundamentals of statistical climatology!

4.3. Difference between geostatistics and meteorology in respect of spatial interpolation

The main difference can be found in the amount of information used for modeling the statistical parameters. In geostatistics, the usable information or the sample for modeling is only the predictors $Z(\mathbf{s}_i,t)(i=1,...,M)$ which belong to a fixed instant of time, that is a single realization in time. „Statistically speaking, some further assumptions about Z have to be made. Otherwise, the data represent an *incomplete* sampling of a *single* realization, making inference impossible.” (*Cressie*, 1991, p. 53.). The assumptions are, e.g., intrinsic stationarity or second-order (weak) stationarity, semivariogram $\gamma(Z(\mathbf{s}_i),Z(\mathbf{s}_j))=\gamma(\mathbf{s}_i-\mathbf{s}_j)$, covariogram $\mathrm{cov}(Z(\mathbf{s}_i),Z(\mathbf{s}_j))=\mathbf{C}(\mathbf{s}_i-\mathbf{s}_j)=C(\mathbf{0})-\gamma(\mathbf{s}_i-\mathbf{s}_j)$, which are some simplifications in order to solve the problem of absence of temporal data. While in meteorology, we have spatiotemporal data, namely long data series which form a sample in time and space as well make the modeling of the climate statistical parameters in question possible. If the meteorological stations $\mathbf{S}_k\ (k=1,..,K)\ (\mathbf{S}\in D)$ have long data series, then spatial trend differences $E(\mathbf{S}_k)-E(\mathbf{S}_l)\ (k,l=1,...,K)$ as well as the covariances $\mathrm{cov}(\mathrm{Z}(\mathbf{S}_k),\mathrm{Z}(\mathbf{S}_l))$ $(k,l=1,...,K)$ can be estimated statistically. Consequently, these parameters are essentially known and provide much more information for modeling than the predictors $Z(\mathbf{s}_i,t)(i=1,...,M)$ only.

5. Software and connection of topics

Our method MISH (Meteorological Interpolation based on Surface Homogenized Data Basis) for the spatial interpolation of surface meteorological elements was developed (*Szentimrey* and *Bihari*, 2007a,b) according to the mathematical background that is outlined in Section 4. 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.

The new software version MISHv1.02 consists of two units that are the modeling and the interpolation systems. The interpolation system can be operated on the results of the modeling system. In the following paragraphs we summarize briefly the most important facts about these two units of the developed software.

Modeling system for climate statistical (deterministic and stochastic) parameters:

- Based on long homogenized data series and supplementary deterministic model variables. The model variables may be height, topography, distance from the sea, etc.. Neighborhood modeling, correlation model for each grid point.
- Benchmark study, cross-validation test for interpolation error or representativity.
- Modeling 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 there is no problem if the greater part of daily precipitation predictors is equal to 0.
- The interpolation error or representativity is modeled too.
- Capability for application of supplementary background information (stochastic variables), e.g., satellite, radar, forecast data.
- 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.

As it can be seen, modeling of the climate statistical parameters is a key issue to the interpolation of meteorological elements, and that modeling can be based on the long homogenized data series. The necessary homogenized data series can be obtained by our homogenization software MASHv3.02 (Multiple

Analysis of Series for Homogenization; *Szentimrey*, 1999, 2007). Similarly to the connection of interpolation and homogenization, in our conception the meteorological questions can not be treated separately. We present a block diagram (*Fig. 1*) to illustrate the possible connection between various important meteorological topics.

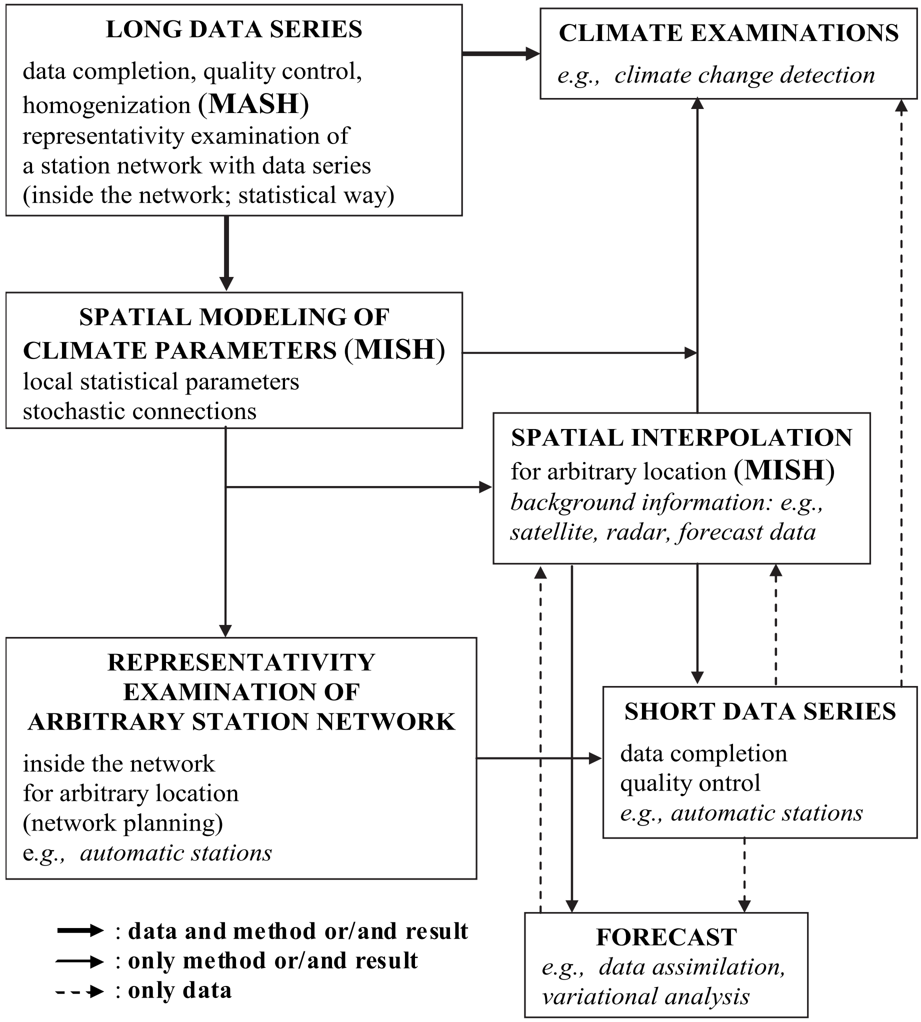

LONG DATA SERIES
data completion, quality control,
homogenization (MASH)
representativity examination of
a station network with data series
(inside the network; statistical way)
CLIMATE EXAMINATIONS
e.g., climate change detection
SPATIAL MODELING OF
CLIMATE PARAMETERS (MISH)
local statistical parameters
stochastic connections
SPATIAL INTERPOLATION
for arbitrary location (MISH)
background information: e.g.,
satellite, radar, forecast data
REPRESENTATIVITY
EXAMINATION OF
ARBITRARY STATION NETWORK
inside the network
for arbitrary location
(network planning)
e.g., automatic stations
SHORT DATA SERIES
data completion
quality ontrol
e.g., automatic stations
FORECAST
e.g., data assimilation,
variational analysis
: data and method or/and result
: only method or/and result
: only data

Fig. 1. Block diagram for the possible connections between various basic meteorological topics and systems.

References

Cressie, *N*., 1991: *Statistics for Spatial Data*. Wiley, New York, 900 pp.

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*, 27-46.

Szentimrey, T., 2007: Manual of homogenization software MASHv3.02. Országos Meteorológiai Szolgálat (Hungarian Meteorological Service), Budapest, p. 65.

Szentimrey, T. and *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* (eds.: *Ole Einar Tveito, Martin Wegehenkel, Frans van der Wel* and *Hartwig Dobesch*), 54-56.

Szentimrey, T. and *Bihari, Z.*, 2007a: 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,* 17-27.

Szentimrey, T. and *Bihari, Z.*, 2007b: Manual of interpolation software MISHv1.02. Országos Meteorológiai Szolgálat (Hungarian Meteorological Service), Budapest, p. 32.

IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 13–30

Aspects regarding the uncertainty of spatial statistical models of climate parameters

Patriche Cristian Valeriu

Romanian Academy, Department of Iasi, Geography Group,
8 Carol I, 700505, Iasi, Romania; E-mail: pvcristi@yahoo.com

(Manuscript received in final form April 26, 2010)

Abstract — Any transformation of a discrete variable into a continuous one is subject to uncertainty. Consequently, the identification and assessment of errors is essential for avoiding misinterpretations of models describing the spatial distribution of climatic parameters. Our study attempts to identify the main sources of errors affecting the statistical spatial models of climatic parameters and to assess their impact on the accuracy of these models. In particular, we focus on georeference errors, the representativeness of the stations network and the related extrapolation problem, the outliers problem, error propagation from simple to complex variables, the problems aroused by heterogeneous regions.

Key-words: uncertainty, spatial statistical models, climate parameters, georeference errors, stations network representativeness, extrapolation, outliers, heterogeneous regions, error propagation

1. Introduction

Our study derives from previous attempts to model the spatial distribution of various climate parameters, which were based, in most cases, on small samples of meteorological stations/rain gauges (*Patriche*, 2007; *Patriche et al.*, 2008). Therefore, our conclusions are applied especially to outputs achieved from such samples, knowing that the degree of uncertainty rises significantly as the sample size used for statistical modeling decreases.

There are many potential sources of uncertainty, which may be grouped into two broad categories:

- **Errors from data pre-processing stage (data quality)**
  - Data recording errors / data series gaps;
  - Instrumental errors;
  - Changes in measurements standards;

- Change in the location of the station / changes in land use around the station.

- **Errors from data processing stage**
  - Georeference errors;
  - Errors derived from the spatial representativeness of the stations network;
  - Errors induced by the presence of outliers;
  - Errors derived from the heterogeneity of the region;
  - Statistical errors;
  - Cumulated errors from computation of complex parameters (error propagation).

Our study focuses on the errors from the data processing stage.

2. Georeference errors

Although simple, the georeference stage is very important. Georeference errors refer to errors of the *x*, *y*, *z* coordinates. Misplacements of stations / rain gauges points on the map may induce significant errors, especially in highly fragmented terrain, when predictors' values are extracted from raster layers or when local interpolators, such as kriging, are used for spatial modeling. The former will lead to wrong predictors' values and, therefore, inaccurate regression models, while the latter will generate locally displaced climatic fields.

The correlation between the stations / rain gauges altitudes and the respective DEM (Digital Elevation Model) altitudes may be used for identifying possible georeference errors or errors in recording the stations / rain gauges altitudes. The correlation should be very good, although not perfect for several reasons: the DEM generalizes the altitude information according to its resolution; the stations / rain gauges latitude and longitude values are generally given in degrees and minutes. Following up the latter issue, if we suppose that the seconds are rounded up or down to the closest minute, it actually means that we may have a coordinate error of up to 30 seconds, meaning about 900 m for latitude and 600 m for longitude, for middle latitudes. These errors double if no coordinate rounding was performed and the seconds were just disregarded.

In the example shown in *Fig. 1*, extracted from a study attempting to model the spatial distribution of mean annual precipitations in Vrancea County, Romania (*Patriche et al.*, 2008), we notice one point (Groapa Tufei) situated outside the correlation cloud indicating a possible georeference error. The recorded altitude for this rain gauge is 125 m, while the DEM altitude for this particular location is 355 m. We can see how far the 125 m altitude isoline is, along which the rain gauge should be located. There are two possible explanations for this error: either the horizontal coordinates of Groapa Tufei are wrong, or the recorded altitude is incorrect. Let us now see the potential negative impact of such a georeference error on spatial statistical models of precipitations. If the real altitude of Groapa Tufei is 125 m, so the recorded altitude is correct, but the horizontal coordinates are wrong, then this point may be used for regression analysis, provided that neither DEM altitude values nor other derived predictors' values are used for models computation. In a geostatistical approach (ordinary kriging, residual kriging, etc.) it is not advisable to include such misplaced points, because they will misplace, in their turn, the precipitation values. Still, if the value of a misplaced point is similar to those of the neighbouring points, as it is in our case, the error induced by the georeference error may be small enough, and the respective point may be kept.

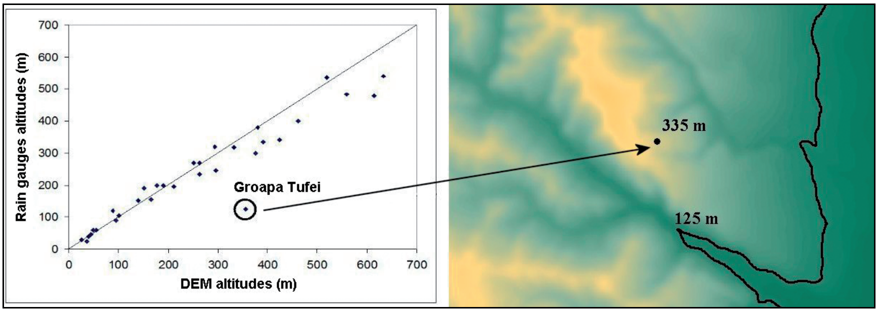


Fig. 1. Revealing two georeference errors for a sample of rain gauges situated in Vrancea County, Romania (*Patriche et al.*, 2008).

3. *Spatial representativeness of the stations network and the extrapolation problem*

The spatial representativeness of the meteorological stations / rain gauges network is an important issue which needs to be addressed in a preliminary stage, as it constitutes a potential source of errors. Theoretically, the spatial distribution of the meteorological network should be well-balanced, in order to grasp all the meteorological and climatological aspects of a territory. However, in most cases, the spatial representativeness of the stations network is more or less inappropriate, due to both its feeble density and its biased location, mainly in valley bottoms.

The representativeness of the meteorological network in relation with the potential predictors may be visualized and evaluated by comparing the predictors' histograms with the histograms of the same predictors, which are based on the predictors' values associated to the meteorological stations / rain gauges.

An example is given in *Fig. 2* for the altitudinal representativeness for a sample of meteorological stations situated in eastern Romania. In an ideal situation, the curves of the cumulated histograms, derived from the DEM and the stations' altitudes, should overlap. However, we notice the shortage of stations between 300 m and 350 m of altitude. Also, we observe the lack of stations at lower altitudes (< 59 m) and especially at higher altitudes, where the highest meteorological station is situated at 391 m of altitude, while the terrain altitudes go as high as 1071 m. As a consequence, we are forced to extrapolate the altitude-based regression models in these areas. As the extrapolation may induce errors, we need to give a special attention to these areas and to consider carefully the reliability of the estimated values.

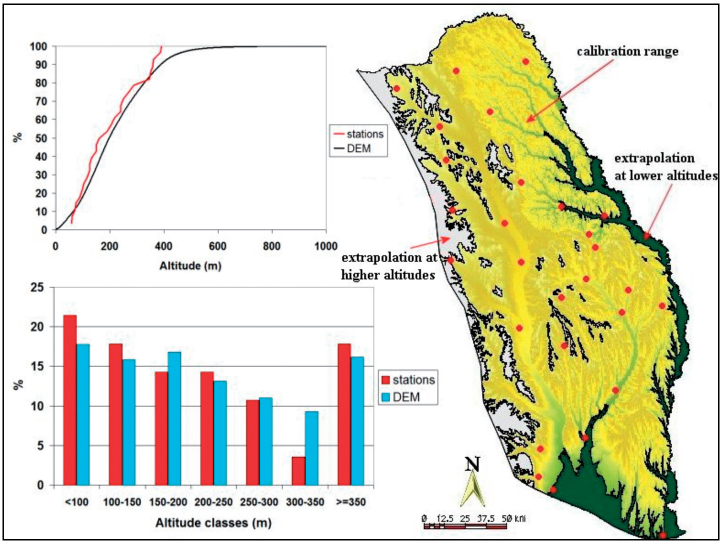


Fig. 2. Assessment of spatial representativeness of stations network by comparing frequencies of predictors' values for station points and for the whole region. Example from eastern Romania (Moldavia) for altitude representativeness for a sample of 28 stations.

Fig. 3 shows an example in which the extrapolation of the regression model should be avoided (*Patriche et al.*, 2008). The mean annual precipitation – altitude regression model, elaborated for Vrancea County (Romania), was based on a sample of 34 rain gauges. The westernmost mountainous part of the region is uncovered by rain gauges, meaning that we must extrapolate our regression model there, if we want to estimate the mean annual precipitation values for this part as well. Performing the extrapolation up to 1770 m of altitude, we estimate precipitation values of up to 1463 mm. Such estimated values are, in our opinion, unrealistic. If the extrapolation is unreliable, then we should confine ourselves with the calibration area of our model. Taking into account that the highest rain gauge altitude is 540 m, we recommend that the study region should not extend over 700 m (*Fig. 3*, black line). Therefore, the entire westernmost part of our region should be excluded from the final map because of extrapolation uncertainty.

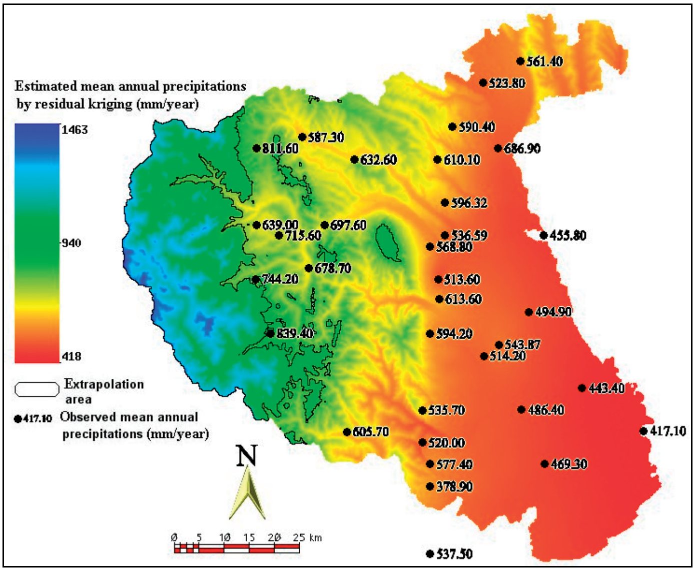


Fig. 3. Avoiding extrapolation. An example from Vrancea County (Romania) for mean annual precipitations.

4. The outliers' problem

An outlier is a point value showing a significant deviation from the statistical model (therefore, marked by a high residual value), corresponding to points (meteorological stations, rain gauges) which mark spatial anomalies for the analyzed parameter's distribution (e.g., foehnization areas, areas of orographic enhancement of precipitations, temperature inversion areas, etc.). Such a "rebel" value may be also an error value, and this possibility must be checked out. If no error is identified then we should proceed to the assessment of the degree in which this value is altering the statistical models, mainly regression models. This is happening in the case of the regression analysis, because it is used mainly as a global interpolation method, and the regression itself is incapable to render spatial anomalies. If such spatial anomalies exist, then the integration within the statistical model of values describing these anomalies may significantly alter the regression equations, which, therefore, become unreliable.

From the viewpoint of their influence over the regression models, we may distinguish two types of outliers:

- Outliers showing high residuals but with similar values of the real residuals and deleted residuals (also known as jackknife error and computed without taking into account the anomaly point). Because such outliers do not modify significantly the regression models, they can be included in the analysis.

- Outliers showing high residuals but with significant differences between the values of the real residuals and those of the deleted residuals. Such outliers modify the regression model and must be, therefore, taken into consideration if the induced modifications are proved to be significant.

There are many statistical procedures aimed towards the identification of outliers. Good syntheses of these procedures are provided by *Maimon* and *Rokach* (2005), and *Wilcox* (2002).

Our approach is a simple one. In order to identify the outliers, we should first inspect the configuration of the correlation cloud between the dependent variable and the predictor, or between the real and predicted values in the case of multiple predictors, looking for points situated significantly outside the cloud. If such points exist, we should further inspect the magnitude of their residual values and see if they are located outside the ± 2.5 RMSE (root mean square error) interval. If such points exist, we should then test their influence on the regression models. The most common way to do this is to perform a cross-validation, the analysis of the differences between the actual residual values and the deleted residuals (jackknife error). If these differences are important, then the exclusion of the respective points significantly changes the regression model, which is, therefore, unstable. Next, we should actually see these changes by elaborating the models with and without the outliers and finally decide whether to keep or eliminate the respective points.

Fig. 4 shows the correlation between the mean annual precipitation and the altitude for a sample of 28 meteorological stations situated in eastern Romania (Moldavia). The chart indicates at least 2 suspect points situated outside the correlation cloud, one with a lower precipitation value than expected for the respective altitude (Cotnari station), another with significantly higher precipitation amounts than expected (Barnova station). These deviations are related to local terrain conditions influencing the pluviometry. Cotnari station is situated in a foehnization area of western air masses descending the eastern slopes of Dealul Mare – Harlau Hill. Here, the real mean annual precipitation value is 121.3 mm lower than the value predicted by the altitude regression model using all stations. On the contrary, Barnova station is situated in an area of orographic enhancement of precipitations caused by the presence of a high energy slope (Iasi Cuesta) facing the more humid western air masses and by the shape of the Barnova-Voinesti depression, which causes the convergence of the western air masses. Another factor is related to the location of Barnova station within a well-forested area. Being the only station from our sample situated within forested areas, it is impossible for us to assess the relative importance of these factors and to state which of them, the local topography or the presence of the forest, is more responsible for the high precipitation values recorded at this location. The real mean annual precipitation value at Barnova station is 172.7 mm higher than the predicted value.

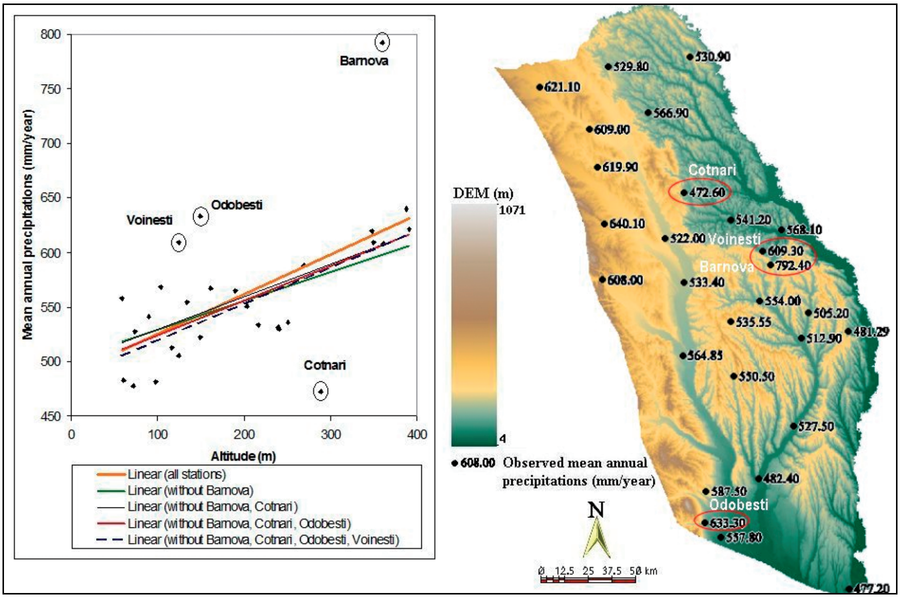


Fig. 4. Correlation chart between altitude and mean annual precipitations for a sample of 28 meteorological stations situated in eastern Romania (Moldavia), indicating the presence of four possible outliers.

If the visual inspection of the correlation charts gives us a first guess on the presence of possible outliers, other methods provide more insight. Our next step is to inspect the magnitude of the residuals. Generally, if some value goes out the interval limited by ±2.5 RMSE (equivalent with the standard deviation of the residuals for large samples), then it is possible that this value is an outlier. From *Fig. 5c* (left), we notice that the residue from Barnova station goes beyond the +2.5 RMSE, while the residue from Cotnari station is very close to the –2.5 RMSE limit. If we eliminate only Barnova station, we find that the residual value at Cotnari goes also beyond the specified limit. Thus, the conclusion is that both stations must be excluded to ensure stability for the regression model. But if we exclude these two stations and rebuild our regression model, we shall find that yet another station (Odobesti) displays residues greater than the +2.5 RMSE limit. Furthermore, if we chose to eliminate Odobesti station, we obtain another high residual value for Voinesti station, situated in the same area of orographic enhancement of precipitations as Barnova station, only at a lower altitude.

So far we have established that we have some poor estimated points in our sample, displaying high residual values. Thus, we are certain that we have some points acting like the first type of outliers (referring to the above classification). The problem now is to decide whether it is necessary to eliminate them from the regression model that is, if this elimination would significantly improve the model.

To answer this question, one must test the influence of these outliers on the regression models and find out whether or not we are dealing with outliers of type two.

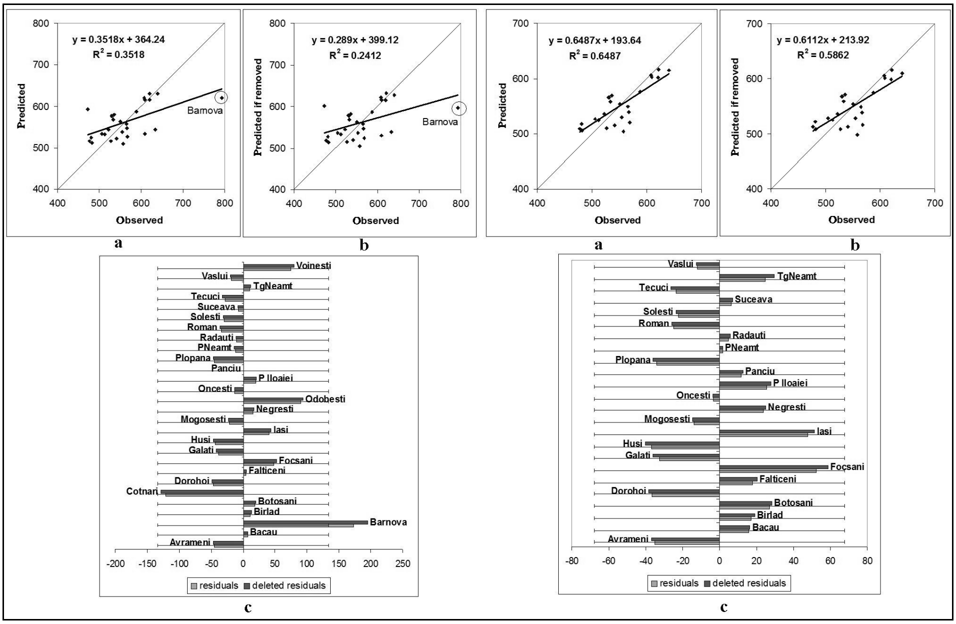


Fig. 5. Correlation between observed and predicted mean annual precipitation (a), cross-validation (b), and comparison of residuals vs. deleted residuals with bars showing the ± 2.5 RMSE (c), using all stations (left) and without four possible outliers (right).

One way to establish the answer is to perform cross-validation, that is, to compare the observed values with the predicted values obtained by successive elimination of the sample points. If the regression models are stable, one should find that the cross-validation charts are similar to the correlation charts between the observed and predicted values. In our case, we may notice that the differences between the observed vs. predicted correlations and the cross-validation correlations decrease as the outliers are removed from the models, from about 11%, in the case of all stations model, to about 6%, in the case of the regression model obtained by removing all of the four possible outliers (*Fig. 5a*,*b*). The slight difference is hampering us so far to state that the removal of the 4 stations significantly improves the regression models.

The comparison between the observed vs. predicted values and the cross-validation charts tells us only something about the stability of the regression models. In order to investigate the influence of particular values, we may find it useful to compare the regression residuals with those obtained by eliminating the suspect point (deleted residuals, jackknife error). If the suspect point is not an outlier, then the magnitude of the residues should be very similar. In our case, we notice that the difference between the actual and deleted residuals is the greatest in the case of Barnova (22.5 mm), which means that its exclusion from the model significantly changes the altitude – precipitation relationship (*Fig. 5c*). The next greatest difference can be found in the case of Cotnari station (7.8 mm). Even if this is not such an important difference, keeping Cotnari station without Barnova station generates an even poorer regression model than the one using all stations. This is due to the fact, that these two points, one above, the other below the regression line, have opposite effects, balancing the regression line to the extent that if one point is removed, the other will "attract" the line towards it. This means that if we chose to eliminate Barnova station, we must eliminate Cotnari station as well.

If we construct our model without these two stations and analyze the residuals, we find that yet 2 other stations display high residuals, going beyond the +2.5 RMSE: Odobesti and Voinesti stations, the latter being situated within the same area of orographic enhancement of precipitations as Barnova station. However, the difference between the actual and deleted residuals is not very significant. The elimination of all these 4 stations leads to a regression model, where no more points display residuals outside the ±2.5 RMSE interval (*Fig. 5c*, right).

Table 1 shows how significant is the influence of the 4 outliers on the regression models. We notice, that the regression quality parameters (correlation coefficients, standard error of estimate) improve by excluding these outliers. However, one should bear in mind that even if there is an overall improvement of the regression models excluding the outliers, these models will still perform poor in the case of the outliers themselves. It is necessary for us to assess if the altitude – precipitation relationship is significantly changing. As we stated before, the regression model without Barnova only is not reliable due to the "attraction" effect of the Cotnari station, and we can clearly see that this model is the most different from the others, showing the highest intercept and the lower pluviometric vertical gradient (regression coefficient). The other models display quite similar parameters: intercepts ranging from 485.6 mm to 498.9 mm and gradients from 30.1 mm/100 m to 36.2 mm/100 m.

Table 1. Comparison of the regression models using and excluding the outliers

Regression model	**Intercept**	**Regression coefficient**	**R^2**	**Standard error of estimates**
All stations	489.21	0.362	0.352	54.472
Without Barnova	501.82	0.265	0.321	41.678
Without Barnova, Cotnari	498.90	0.301	0.450	36.190
Without Barnova, Cotnari, Odobesti	492.72	0.315	0.547	31.697
Barnova, Cotnari, Odobesti, Voinesti	485.64	0.335	0.649	27.626

From *Fig. 6* we may see that 31% of the station sample displays the lowest residuals under the 2nd model (without Barnova and Cotnari stations). A similar percent (30%) is found for the 4th model (without all of the 4 outliers).

To sum up, our conclusion is that in the particular case of our sample, the elimination of the identified 4 outliers improves the regression model even though the differences among the various models are not very important.

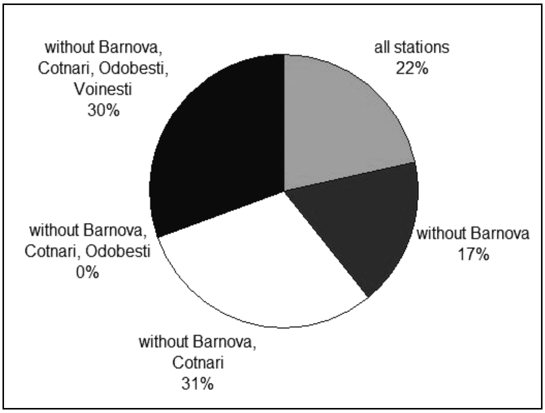


Fig. 6. The optimum altitude regression model (lowest values of actual residuals minus deleted residuals) for each station.

The problem is that we can not just exclude some real values from the analysis, because then we would obtain an incomplete image of the spatial distribution of the analyzed climatic parameter.

Some of the possible solutions could be:

- data transformation (logarithms);
- derivation of new predictors to account for spatial anomalies;
- application of robust regression methods (*Wilcox*, 2002);
- application of regression as a local interpolator (e.g., geographically weighted regression method);
- application of residual kriging.

A common solution is to derive one or more predictors (*Lhotellier* and *Patriche,* 2007) capable to explain the anomaly associated to the outlier point (e.g., the west-east aspect component combined with terrain local altitudinal range could theoretically explain the precipitations anomaly identified at Cotnari, Barnova, and Voinesti stations from the previous example). Practically, we are often hampered in our analysis by the poor spatial representativeness of the stations network, especially when we have to work with small stations samples, which is, in most cases, unable to fully account for all terrain aspects relevant for the spatial distribution of the analyzed climatic parameter.

The application of residual kriging is also a common approach (*Lhotellier*, 2005; *Dobesch et al.,* 2007; *Hengl*, 2007; *Silva et al.*, 2007). Thus, what regression is unable to explain (the residuals), is interpolated using ordinary kriging, then the spatial trend, derived by regression, is added to the spatial anomalies, resulting in the final spatial model of the climate parameter. The output of this approach is still influenced by the quality of the regression model. If the model is significantly influenced by the outliers, then we can not attempt to interpret the predictors-predictand relations.

An alternative solution could be the elaboration of the regression model without the values identified as outliers, the spatialization of the residuals by ordinary kriging, including the residuals associated with the anomaly points, followed by the addition of the spatial trend with the interpolated residuals so as to obtain the final spatialization. We notice, that this is a residual kriging approach, which eliminates the outliers during the regression stage, if these belong to the type two mentioned above, but includes the residuals from these points during the kriging interpolation stage (*Fig. 7*).

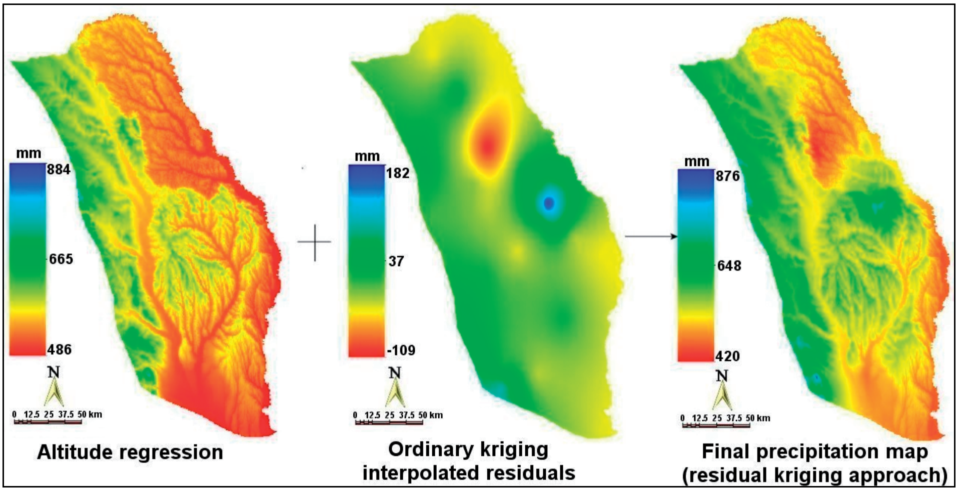


Fig. 7. Mapping the optimum solution: residual kriging approach leaving out the outliers during the regression stage, but taking the outliers' residuals into account during the kriging stage.

A better approach consists in the application of regression as local interpolator (e.g., geographically weighted regression, *Fotheringham et al.*, 2002), taking into account the spatial anomalies (*Engen-Skaugen* and *Tveito*, 2007; *Maracchi et al.*, 2007). The local regression can be further included into a residual kriging approach in order to improve the quality of the output. The main drawback to this approach is the need of a sufficiently large stations sample in order to be capable to derive local regression models.

Let us now see a situation, in which the outliers may indicate possible data errors or different recording intervals. The example is extracted from a study attempting to model the spatial distribution of mean annual precipitations in Vrancea County (Romania) on the basis of 34 rain gauges (*Patriche et al.*, 2008).

Figs. 8 and *9* show 2 points situated significantly outside the altitude – precipitation correlation cloud, namely Pufesti (686.9 mm) and Slobozia Bradului (378.9 mm), therefore, indicating the presence of two possible outliers. In the case of Pufesti rain gauge, the mean annual precipitation regime is characterized by a secondary maximum in August. Taking into account, that all other rain gauges display a single maximum in June, we are inclined to believe that either the August data is incorrect or the Pufesti data represent a shorter time frame, corresponding to a more humid period. On the other hand, the mean annual value recorded at Slobozia Bradului rain gauge is obviously too small for the climatic conditions of our region. Because the monthly values display a normal annual distribution, we are inclined to believe, as before, that the data correspond to a shorter time frame from a drier period.

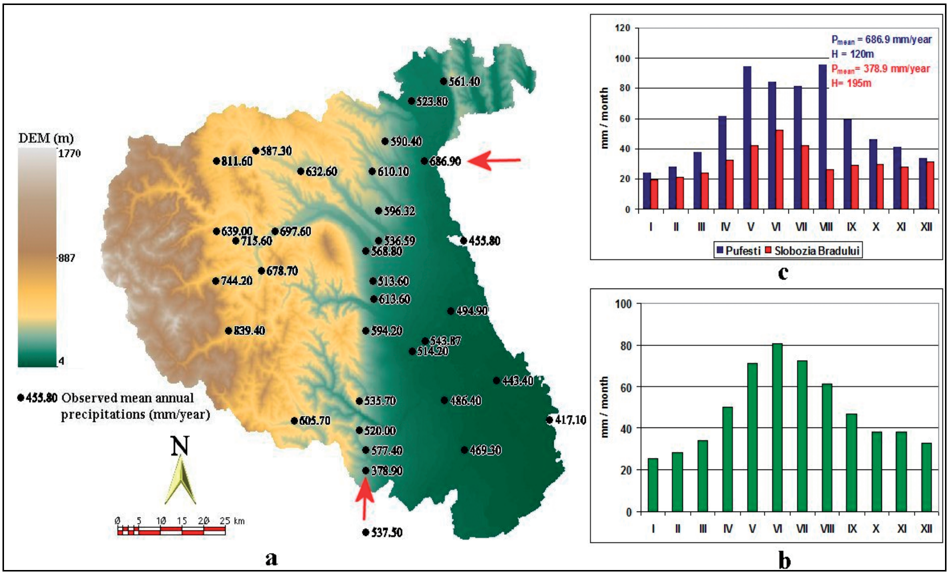


Fig. 8. Observed mean annual precipitations in Vrancea County, Romania (a), mean annual precipitations regime for all stations (b), and for the two suspect points (c).

From *Fig. 9b*, we notice that even though these two points are associated with the highest residuals, the difference between the actual and deleted residuals (jackknife error) is small, meaning that their removal from analysis does not significantly change the altitude regression model. This is happening because the points are situated on opposite sides as compared to the regression line (*Fig. 9a*) and, therefore, they have opposite effects, balancing the regression line. Their removal increases the correlation coefficient but does not significantly change the direction of the regression line, meaning that the regression equations are very similar with or without these points. This can also be grasped, if one notices that the altitude – precipitation correlation coefficient (0.66) is quite similar with the cross-validation correlation coefficient (0.62), meaning that the one by one removal of all sample points does not significantly change the altitude – precipitation relationship (*Fig. 9c*).

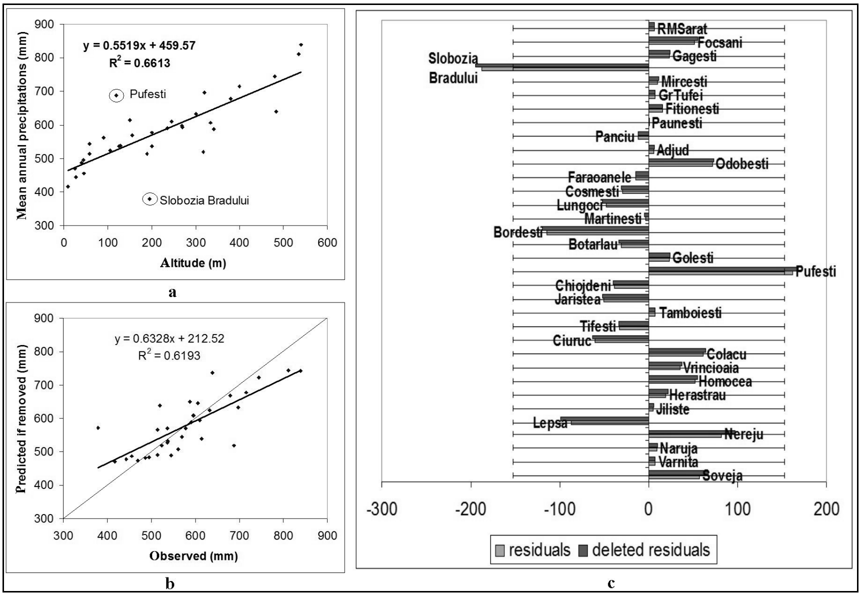


Fig. 9. The altitude – mean annual precipitation relationship (a) and the comparison between actual and deleted residuals (c) showing the presence of two possible outliers. Cross-validation of the altitude model using all stations (b).

Let us see the effects on other predictors. We must mention that, apart from altitude, we also used latitude and longitude as predictors, and at first we obtained a good regression model using both altitude and latitude. Looking further into details, we noticed that the latitude – precipitation correlation is a false correlation, induced by the presence of the two outliers (*Fig. 10*), one with a higher precipitation value situated in the northern part of our region (Pufesti), the other one with a lower precipitation value situated in the south (Slobozia Bradului). If one eliminates these two points, the latitudinal correlation is no longer statistically significant.

For this reason and because of our intention of using also kriging for spatialization, in which case the great residual values of the two suspect points would be represented on the map, we decided to eliminate them from analysis.

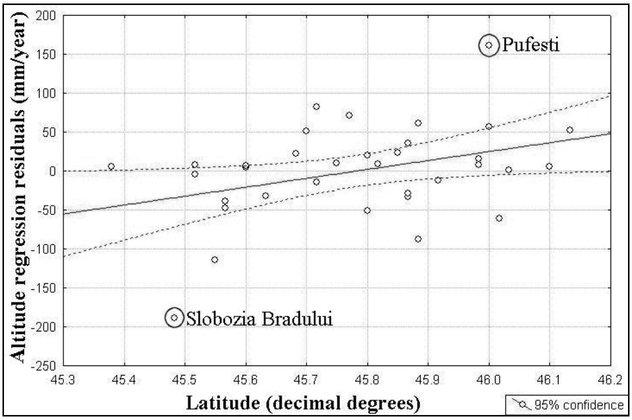


Fig. 10. An unwanted effect of outliers: false precipitation – latitude correlation.

5. *Error propagation*

Statistically based spatial models are usually computed for elementary variables, such as temperature or precipitations. In order to describe the climate of a region, we also need to compute complex variables, derived from the elementary ones, such as the de Martonne index, potential and real evapotraspiration, etc.

Spatial models of complex variables may be achieved either by computing the complex variable at stations' locations and then interpolating the results or by integrating the spatial models of the elementary variables in order to obtain the complex one. Using the first approach, we are able to quickly compute the errors as well. In this case, we cannot speak of error propagation. Still, in our opinion, this approach is conceptually less appropriate, because the computation of the complex variable is deterministic, according to a physical model. For instance, computing the potential evapotranspiration according to Penman-Monteith approach involves the computation of the net shortwave radiation, which depends on terrain slopes and expositions and on land use. If one computes this parameter at stations' locations and then interpolates the results, neither of these control factors will be taken into account.

The second approach, namely the integration of elementary variables, each of them displaying certain errors, has the disadvantage of inducing invariably in the propagation of these errors to the derived, complex variable. Knowing these errors is important for the assessment of the accuracy of the derived variable's spatial distribution.

A simple example is presented in *Table 2*. The example refers to the derivation of the de Martonne aridity index, for the territory of Moldavia (eastern Romania), on the basis of the mean annual temperatures and precipitations statistically modeled by regression. The mean annual temperature model uses altitude and latitude as predictors, the computed standard error of estimate is ±0.215 °C, meaning that the real temperature differs from the estimated one with ±0.215 °C in about 68% of the cases. If we consider, for exemplification, an estimated mean annual temperature of 10 °C, then the real temperature will most probably be found within the interval of 9.8 – 10.2 °C. On the other hand, the mean annual precipitation model uses altitude as predictor and has a standard error of estimate of ± 54.472 mm/year, which means that, for an estimated value of 500 mm, the real precipitation values will most probably lie within the interval of 445 – 554 mm/year. Considering the two estimated temperature (10 °C) and precipitation (500 mm/year) values, it results an aridity index of 25. Taking into account the possible errors for the estimated input parameters, it results that the real value of the aridity index will be most likely found between 22 and 28.

Table 2. Exemplification of error propagation

Statistical parameters		Mean annual precipitation	Mean annual temperature	Aridity index
		Real values		
Exemplification values		500	10	25
Mean		561.95	8.90	29.21
Standard deviation		66.395	0.734	3.136
Standard error		54.472	0.215	–
Confidence interval	Lower limit	445.528	9.785	22.039
	Upper limit	554.472	10.215	28.025
	Range	108.944	0.431	5.986
		Standardized values		
Standardized standard error		0.820	0.294	–
Confidence interval	Lower limit	–1.754	1.206	–2.287
	Upper limit	–0.113	1.793	–0.378
	Range	1.641	0.587	1.909

We are, however, unable to compare these confidence intervals, because the 3 parameters are expressed in different measurement units. One solution is to compute the regression models using the standardized values of the input parameters. We find that the size of the confidence intervals is 1.641 for mean annual precipitation and 0.587 for mean annual temperature. The resulting aridity index distribution for the stations sample is characterized by a mean value of 29.21 and a standard deviation of 3.136. Therefore, the lower limit of the confidence interval (22) corresponds to a standardized value of –2.287 and the upper limit (28) corresponds to a standardized value of – 0.378, resulting a range of 1.909. This value is greater than the ones of the input parameters, indicating the propagation and enhancement of the errors, from the elementary variables to the derived, complex variables.

6. *Homogeneous vs. heterogeneous regions*

Another issue we address in our study is that of heterogeneous regions. Generally, the greater a region, the more heterogeneous it is. A certain level of heterogeneity is necessary for the spatialization of climate parameters. For instance, within a small region, in which the altitudinal range does not exceed, for example, 100–200 m, the spatial variation of the climate fields may be too feeble for us to correctly infer the spatial variation rules. On the other hand, within a large region, the climatic heterogeneity may be too high for a single statistical model to explain it.

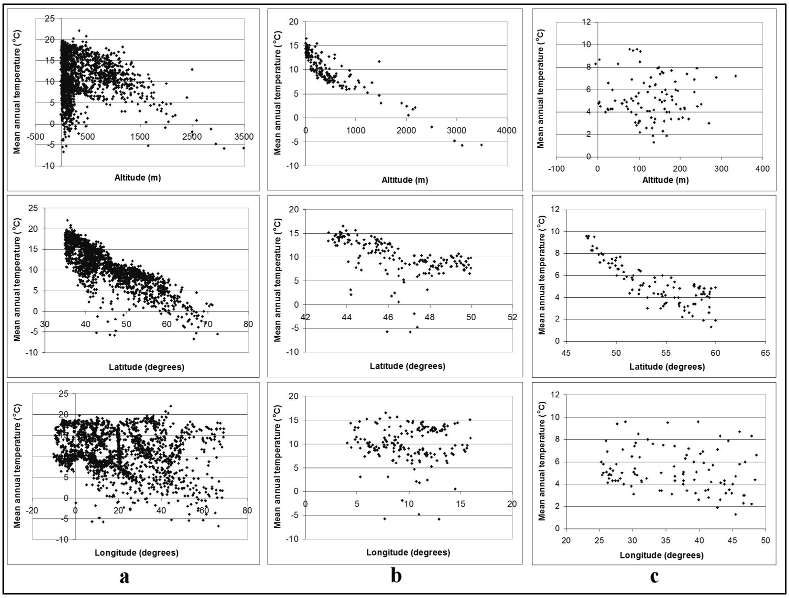


Fig. 11. Changes of the relationships between the mean annual temperatures and the altitude, latitude, and longitude for Europe (a) and for two different subregions: the Alps (b) and the Russian Plain (c). Source of data: *FAO*, 2003.

An example is shown in *Fig. 11* for the relationship between the mean annual temperature (*FAO*, 2003) and 3 predictors: altitude, latitude, and longitude. At continental scale, the territory of Europe is very heterogeneous. We may notice, that the altitude – temperature relationship changes form one region to another to such an extent that a single regression equation for the whole European territory cannot be constructed. A region like the Alps displays a very good altitude – temperature correlation, while the temperature variation within the flat relief of the Russian Plain is statistically independent of the altitude, as temperature inversions are frequent. Here, the latitude comes forward to explain a good part of the temperature spatial distribution.

In such situations, when we deal with large heterogeneous regions, it becomes necessary to divide it into smaller, more homogeneous sub-regions, for which the predictors-predictand relationships do not change. A possible approach could consist in the examination of regression parameters and residuals as we extend or reduce the area of our region and establish the sub-regions limits according to the most stable regression model (maximum correlation, minimum residuals). Another possible approach could be the application of regression as a local interpolator.

7. Conclusions

When applying statistical methods for deriving digital spatial models of climatic variables, one must take great care in identifying and assessing the sources of uncertainty, especially in the case of small stations samples. There are many such sources of different nature, which can easily mislead us towards wrong unrealistic conclusions. Consequently, a good knowledge of data quality, statistical methods, and, needless to say, climatology is imperative for the achievement of sound results. Although simple, the georeference stage is very important. The misplacement of one or more meteorological stations on the map may generate an unwanted chain of errors, because the predictors' values are automatically drawn from the raster maps in GIS environment. The representativeness of the stations network is another important issue, which needs to be analyzed in a preliminary stage of climate parameters spatialization. Theoretically, the spatial distribution of the stations network should be in agreement with terrain complexity, so as to be able to account for all climatic aspects. The extrapolation problem is tightly related to this issue. Unfortunately, in most cases, the stations network is biased, therefore, not sufficiently representative for the terrain. The extrapolation of the spatial models is correct as far as the predictors-predictand relationships do not significantly change outside the calibration area. The outliers problem, meaning the problem of values evading a certain spatial variation rule, is another aspect we analyzed in our study. This is another aspect of the representativeness of the stations network in respect to predictors, which needs to be preliminary addressed in order to minimize the potential errors. Statistical modeling is generally performed on simple, elementary variables, such as temperature or precipitation. For a more thorough investigation of a region's climate, we need to dispose of complex variables, derived from the elementary ones, such as the de Martonne aridity index, potential evapotranspiration, etc. The integration of elementary variables, each having its own statistical errors, into complex variables leads to error propagation. Knowing these errors is very important in order to assess the accuracy of the modeled spatial distribution of the complex variable. Another issue we address in our study is that of the heterogeneous regions. Generally, the greater a region, the more heterogeneous it is. A certain level of heterogeneity is necessary for the spatialization of climate parameters. On the other hand, within a large region, the climatic heterogeneity may be too high for a single statistical model to explain it. In such a situation, it becomes necessary to divide our large region into smaller, more homogeneous sub-regions, for which the predictors-predictand relationships do not change.

Acknowledgment — This study was carried out with support from project POSDRU/89/1.5/S/49944, coordinated by "Alexandru Ioan Cuza" University of Iaşi (Romania).

References

Dobesch, H., Dumolard, P., and *Dyras, I.* (eds.), 2007: *Spatial Interpolation for Climate Data. The Use of GIS in Climatology and Meterology*. Geographical Information Systems Series, ISTE, London and Newport Beach.

Engen-Skaugen, T., and *Tveito, O.E.,* 2007: Spatially distributed temperature lapse rate in Fennoscandia, in COST Action 719: *Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology* (eds.: *Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M.*), Budapest, 25-29 October 2004. Luxembourg: Office for Official Publications of the European Communities, EUR 22596, 93-100.

FAO, 2003: *FAOCLIM-2. World-wide agroclimatic database v.2.02*, FAO/SDRN.

Fotheringham, S., Brunsdon, C., and *Charlton, M.,* 2002: *Geographically Weighted Regression. The Analysis of Spatially Varying Relationships*. Wiley.

Hengl, T., 2007: *A Practical Guide to Geostatistical Mapping of Environmental Variables*. JRC Scientific and Technical Research series. Office for Official Publications of the European Comunities, Luxembourg, EUR 22904 EN.

Lhotellier, R., 2005: *Spatialisation des températures en zone de montagne alpine*, thèse de doctorat, SEIGAD, IGA, Univ. J. Fourier, Grenoble, France.

Lhotellier, R., and *Patriche, C.V.,* 2007: Dérivation des paramètres topographiques et influence sur la spatialisation statistique de la temperature. *Actes du XXème Colloque de l'Association Internationale de Climatologie*, 3-8 septembre 2007, Carthage, Tunisie, 357-362.

Maimon, O.Z., Rokach, L. (eds.), 2005: *Data Mining and Knowledge Discovery Handbook*, Chapter 7. Outlier detection, Ben-Gal, I. Springer.

Maracchi, G., Ferrari, R., Magno, R., Bottai, L., Crisci, A., and *Genesio, L.,* 2007: Agrometorological GIS products through meteorological data spatialization, in COST Action 719: *Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology* (eds.: *Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M.*), Budapest, 25-29 October 2004. Luxembourg: Office for Official Publications of the European Communities, 2007, EUR 22596, 9-16.

Patriche, C.V., 2007: About the influence of space scale on the spatialisation of meteo-climatic variables. *Geographia Technica*, no. 1, Cluj University Press, 70-76.

Patriche, C.V., Sfîcă, L., and *Roşca, B.,* 2008: About the problem of digital precipitations mapping using (geo)statistical methods in GIS. *Geographia Technica*, no. 1, Cluj University Press, 82-91.

Silva, Á.,P., Sousa, A.J., and *Espírito Santo, F.,* 2007: Mean air temperature estimation in mainland Portugal: test and comparison of spatial interpolation methods in Geographical Information Systems, in COST Action 719: *Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology* (eds.: *Szalai, S., Bihari, Z., Szentimrey, T., Lakatos, M.*), Budapest, 25-29 October 2004. Luxembourg: Office for Official Publications of the European Communities, EUR 22596, 37-44.

Wilcox, R., 2002: *Applying Contemporary Statistical Techniques*. Academic Press.

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 31–49

Operational maps of monthly mean temperature for WMO Region VI (Europe and Middle East)

Florian Hogewind[1] and **Peter Bissolli**[2]

[1]*Karlsruhe Institute of Technology,*
Institute of Geography and Geoecology,
Reinhard-Baumeister-Platz 1, 76128 Karlsruhe, Germany
E-mail: Florian.Hogewind@kit.edu

[2]*German Meteorological Service, Offenbach,*
Frankfurter Straße 135, 63067 Offenbach, Germany
E-mail: Peter.Bissolli@dwd.de

(Manuscript received in final form October 30, 2010)

Abstract—A spatial interpolation method for the construction of operational climate maps for the WMO RA VI Region (Europe and Middle East) is presented on the example of monthly mean temperature. The method is suitable for an in situ data base with relatively low data coverage in a relatively large and climatically heterogeneous area, and considers the classical geographical parameters latitude, longitude, and altitude by multi-dimensional linear regression, but improved by continentality, using a new continentality index. A comparison of several interpolation methods reveals that radial basis functions (subtype multiquadratic) seems to be the most appropriate approach. Separate regressions for land and sea areas further improve the results.

Key-words: spatial interpolation, multi-dimensional linear regression, climate maps, RA VI, Europe

1. Introduction

Climate monitoring requires an operational analysis of the variability of climatic quantities in space and time. For this purpose, operational maps, generated for regular time intervals (days, months, seasons, years) are very useful to see at a glance the spatial variability of climate elements and its change with time. Such maps are often used by national meteorological and hydrological services as a basis for climate reviews and interpretation of outstanding features of climate variability. Maps are available for various spatial areas from the catchment scale to the whole globe.

For some recent years, the German Meteorological Service (Deutscher Wetterdienst, DWD) develops methods for generating such operational maps. These methods are not exactly the same for all climate elements due to various databases, their special nature of variability, and the data availability. Some are based on satellite data (e.g., cloud and radiation parameters), others are based only on in situ data, because in that cases the in situ data have a relatively good quality compared to satellite data (e.g., temperature, precipitation, sunshine duration, snow depth). Examples of these maps can be seen on the DWD website (www.dwd.de/rcc-cm, www.dwd.de/snowclim, www.dwd.de/satklim).

On the other hand, it is desirable to use consistently the same method for each climate element to achieve consistent maps, at least the same basic principle of a method. Our present strategy is to develop a basic approach which is at least applicable for most of the in situ data. The process of map generation is still under further development.

Usually, maps are a result of gridding or spatial interpolation of point data into the area. Nowadays, a large variety of mathematical and geostatistical methods for spatial interpolation is available. However, in practice, it has turned out that pure mathematics and geostatistics are necessary, but not sufficient for construction of climate monitoring maps; instead it has been found that the consideration of geographical conditions and climate processes can much improve the results. Nevertheless, the impact of such additional parameters and processes depends highly on the extent and topography of the area of interest, and also on data density. Therefore, the choice of the gridding method depends on the selected area, and the selected climate element as well.

This paper refers specifically to spatial interpolation of monthly mean temperature and its anomalies from the reference period 1961–1990 in a relatively large area, the WMO (World Meteorological Organization) Region VI (covering nearly the whole Europe and the Middle East). The next chapter describes this area and the motivation for the choice of this area. After a short review of previous literature, the data and the succeeding steps of the method applied in this paper are described and compared with a number of alternatives. Results of the comparison and the mapping are presented in Section 6, followed by some conclusions in Section 7.

The main goal of this paper is to propose a method of spatial interpolation of monthly temperature data in WMO Region VI which is suitable for an operational generation of monthly climate monitoring maps. However, it is intended that this approach is applicable to other climate elements as well to receive maps of various elements which are consistent to each other as far as possible, at least for in situ data. Other data sources, like satellite data which already have a large spatial coverage certainly require a different approach.

2. *The WMO Region VI and the Regional Climate Centre (RCC) network*

Recently, a new network of so-called Regional Climate Centres (RCCs) has been established under the auspices of the World Meteorological Organization (WMO) (http://www.wmo.int/pages/prog/dra/eur/RAVI_RCC_Network.php). The term "regional" refers to the six WMO Regions which cover roughly (but not exactly) the various continents and the surrounding sea areas on the globe.

Nearly the whole of Europe (except the easternmost parts of European Russia from 50°E to the Ural) belongs to the WMO Region VI (often referenced as "RA VI", indicating the Regional Association of the WMO in Region VI). Beside Europe, this region also covers parts of the Middle East which belong geographically to Asia, and also large sea areas, namely large parts of the northern and central North Atlantic, the Norwegian Sea, the European part of the Arctic, and the whole Mediterranean. The RA VI area is displayed in *Fig. 1*.

The border of the Region VI (Europe and Middle East) is not rectangular, because it is defined by the borders of single countries, which means largely by political conditions. Over European Russia, the eastern border runs along the 50°E meridian. In the south and west, the border crosses the Mediterranean Sea and the Atlantic Ocean to the Davis Strait and the Baffin Bay between Greenland and Canada.

Thus, that Region covers quite a large and climatically very heterogeneous area, spanning a wide range of latitude, longitude, and altitude and strong contrasts between land and sea climates.

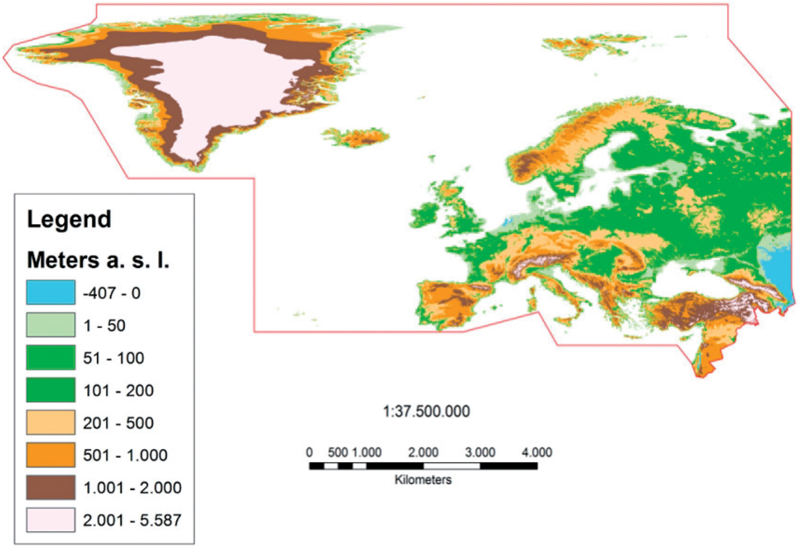


Fig. 1. Map of the Region VI with the height above sea level. The kilometer bar refers to Central Europe.

In RA VI, presently three RCCs (so-called nodes of the RCC network) are already preliminarily established and are operating in a pilot phase since June 2009: one RCC node on climate data, one on climate monitoring, and one on long-range forecasting. The DWD has taken over the lead function of the RCC node on climate monitoring in RA VI, within a consortium consisting of some more members (national meteorological and hydrological services) of RA VI. To fulfill this function, the generation of climate maps for various climate elements in RA VI is a very important task.

3. Previous approaches

A large number of papers dealing with spatial interpolation of climate data have already been published. Basic information about spatial interpolation methods can be found in various books, especially for the widely used kriging technique, which was very popular already in the 1990s (e.g., *Lang*, 1995; *Stein*, 1999). In the 2000s, geographical information systems (GIS) came more and more into operation for climate mapping. Commercial GIS software has made it technically very easy to apply spatial interpolation methods on any geographically defined data points. In 2001, the COST Action 719 was launched (COST= European Cooperation on Science and Technology, an intergovernmental framework for research coordination in Europe, supported by the European Union). The goal of COST 719 was to review and assess the use of GIS for spatial interpolation in meteorology and climatology. The Action had been finished in 2006, resulting in an overview of spatial interpolation methods and their application in climatology by GIS software (*Thornes*, 2005; *Tveito et al.*, 2008) and many related papers (e.g., *Ustrnul* and *Czekierda*, 2005; *Dobesch et al.*, 2007).

Until now, there are several more recent papers. Various methods are applied to national data, some also to larger areas, e.g., the Alps, some to global data, but in coarse resolutions. Many investigators used ordinary or residual kriging techniques for monthly, seasonal, or annual data, e.g., *Bjornsson et al.* (2007) for temperature in Iceland, *Ustrnul* and *Czekierda* (2005) for temperature in Poland, *Dolinar* (2006) for sunshine duration in Slovenia, *Perčec Tadić* (2010) for climate normal values of various elements (including temperature) for Croatia, *Alsamamra et al.* (2009) for solar radiation in southern Spain. Others just used multiple regression techniques, but in a dense station network and with many geographical predictors, e.g., *Hiebl et al.* (2009) for monthly temperature in the Alps or *Claps et al.* (2008) for monthly temperature in Italy. Non-linear instead of linear statistical relationships between terrain variables as predictors and climate variables lead to an improvement at least for special variables like, e.g., fog frequency as shown by *Vicente-Serrano et al.* (2010) for northeast Spain. In some cases, circulation types were used as predictor, e.g., the well-known “Grosswetterlagen” catalogue from *Hess* and *Brezowsky* (1952) for the temperature in Poland (*Ustrnul*, 2006). Other authors included remote sensing data for statistical temperature modeling (e.g., *Cristóbal et al.*, 2008 for northeast Spain).

In contrast, there have been very few attempts to look for a method which is specifically appropriate for an area like RA VI. Recently, *Haylock et al.* (2008) presented a new European high-resolution gridded data set of daily precipitation and surface temperature for the period 1950–2006 on four spatial resolutions (the so-called E-OBS data set). Although this data set is widely known and used, the authors themselves pointed to limitations of their gridded data due to inhomogeneities and interpolation uncertainties (*Hofstra et al.*, 2009). *Hofstra et al.* (2008) also compared several interpolation methods for various variables in some parts of Europe and found that the main controlling factor on the skill of interpolation is rather the density of the station network than the interpolation method. Only recently, another investigation used the spatial variability from past observations of a denser network to improve the interpolation skill, in this case applied to precipitation in the complex terrain of Switzerland (*Schiemann et al*., 2010).

Monthly, seasonal, and annual maps are frequently used for operational climate monitoring activities. The monitoring of the WMO RA VI Regional Climate Centre on Climate Monitoring (WMO RAVI RCC-CM) can be found on the web: http://www.dwd.de/rcc-cm, including links to national maps of many national meteorological and hydrological services. For global climate monitoring, monthly temperature maps are displayed, e.g., on the website of the National Oceanic and Atmospheric Administration (NOAA) in the USA:
http://www.ncdc.noaa.gov/climate-monitoring/index.php#global-icon .

4. Data and data quality

Since the goal is to generate monthly maps for RA VI in the operational environment of DWD, it is essential to use monthly in situ data which are available at DWD soon after the end of month, but, nevertheless, of good quality. National data sets exist for each country in RA VI. Mainly they are under the responsibility of the public national and hydrological services. Due to this national responsibility of the data, each country has its own data policy, and in most cases there are restrictions in data distribution beyond the national services. For this reason, only a limited number of all existing data can be used in the DWD environment. However, there are some data which are distributed internationally and regularly via the Global Telecommunication System (GTS) of the WMO. Two important data sets in this case are the SYNOP and CLIMAT data. SYNOP data are data from synoptical stations, distributed several times a day (often hourly), containing also the air temperature at two meters height over ground. They are mainly intended for usage in weather forecasting. CLIMAT data are distributed only monthly and the number of stations is much smaller than for SYNOP, but the selection of CLIMAT stations was done for usage of climate analyses. Monthly mean temperature is one of the climate elements which are reported in the CLIMAT bulletin. DWD has taken over the task to check the quality of the CLIMAT data each month in various steps. The quality check consists of two steps: a quick automatical check soon after data arrival and a more thoroughly manual check later. More details about the CLIMAT data archive and quality control method can be found on the DWD website (www.dwd.de – click on Climate and Environment – Climate Data Centers – ACD). The first step of quality check is normally done within 10 days after the end of each month. The check of SYNOP data would be more time consuming, and a complete routine quality control for SYNOP temperature data at DWD is only performed for German data, but not for the whole of the RA VI area. For this reason it was decided to use the CLIMAT data of monthly mean temperature for spatial interpolation, which means a data basis which is timely available in good data quality, but relatively poor data coverage (*Fig. 2a*). Around 800 CLIMAT stations are currently available for RA VI each month, and the area has an extension of several 1000 km in both zonal and meridional directions. This decision means to invest into an appropriate and reasonable interpolation procedure, which also takes the diverse topography of RA VI into account.

CLIMAT stations are available only for the land areas, but not for the sea. However, there exist weather reports from ships which are summarized into a 2.5°×2.5° latitude-longitude grid and are archived at DWD. Altogether, around 130 sea grid points are used for each month. Although the grid points are uniformly distributed over the area, the underlying ship reports are not equally distributed. The best data coverage can be found along the main shipping routes such as between Europe and the eastern coast of the USA or Brasilia, and the main route to the Mediterranean Sea, but in other areas ship data are quite rare (*Fig. 2b*). Thus, the quality of ship data is strongly dependent on ship observation coverage. They are most reliable along the main shipping routes where a large number of ship observations during the whole month are considered for gridding, but quite poor in those regions where only very few ship observations are available, e.g., over the Arctic Sea. Long-term averages for the 1961–1990 reference data (CLIMAT and ship data, as far as data available) are also quality controlled and included in the DWD archive, and anomalies (monthly means minus long-term averages) are computed each month as well.

For using the topography in the interpolation procedure, grid data for altitude are needed. Data for the height above sea level are taken from the GTOPO30 altitude raster from the U.S. Geological Survey (www.usgs.gov). The data are available in a spatial resolution of 30 seconds of degree in latitude and longitude (it means about 1 km for middle latitudes). For the operational maps, a spatial resolution of 0.1° was taken; thus, the GTOPO30 data were averaged into a 0.1° grid.

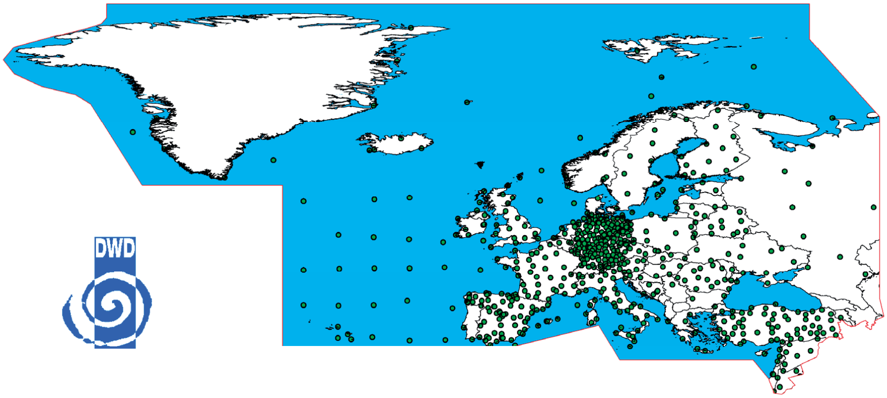


Fig. 2a. Spatial distribution of CLIMAT stations and ship data points available at DWD for September 2010 as an example. Ship data of the whole month are arithmetically averaged into a 2.5°×2.5° grid.

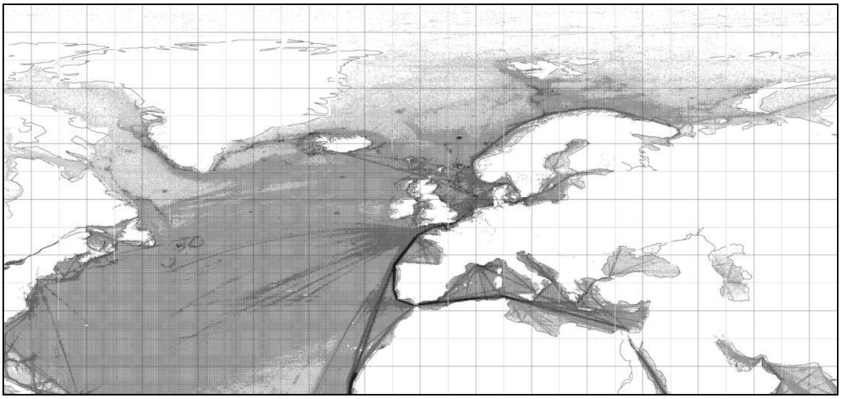


Fig. 2b. Ship data coverage, data from DWD (white = land area, light grey to dark grey = more travel on sea, if the color is darker, more ships travel on this route).

5. *Methods*

In principle, the spatial interpolation method for monthly averages used here consists of three steps. The first step is a multi-dimensional linear reduction of the station data, which means a multiple linear regression of latitude, longitude, altitude, and other parameters to zero level. The linear regression model is subtracted from the original data; the results of the subtraction are often called “residuals”.

The second step is the interpolation of residuals with the method radial basis functions, using the version of the software ESRI Arc GIS 9.2 within the tool geostatistical analyst. The last step is recomputing the interpolated residuals to the original values of latitude, longitude, altitude, and other parameters. This computing is achieved by using the raster calculator of Arc GIS 9.2 within the tool spatial analyst. The three steps are described in more detail in Sections 5.1–5.6.

For the anomalies there is no reduction, just a spatial interpolation is necessary, assuming that they do not depend very strongly on geographical parameters. Spatial resolution is 0.1°; this corresponds to about 10 km over Central Europe. The number of grid points in the RA VI area roughly amounts to nearly one million.

At the borders of Region VI, the problem of extrapolation appears. For this reason, the interpolation is computed for an extended area (from 85°W to 70°E and from 20°N to 90°N), but only the Region VI itself is displayed. For this purpose, some more climate stations beyond Region VI are added to the data pool. The additional climate stations are located in the east part of the USA and Canada, the North African states, and in the part of the Middle East, which belongs to the Region II Asia.

5.1. General approach of multiple regression in latitude, longitude, altitude

The assumption of the multi-dimensional reduction is that the spatial variability of monthly averaged climate is dominated by a very limited number of impact factors.

The general approach is

$$Y = a\, f_1\,(x_1\,, x_2\,, \ldots, x_n\,) + b\, f_2\,(x_1\,, x_2\,, \ldots, x_n\,) + \ldots + k\,, \qquad (1)$$

where Y is a climate state variable like temperature, x_1, x_2, … are impact factors like latitude etc., $f_1()$, $f_2()$, … are functions of impact factors, which are not necessarily linear, and a,b, …, k are constant values.

This approach is used to find the dominating impacts, x_1, x_2, and the functions of impacts, $f_1()$, $f_2()$, for each Y. The functions of the impact factors must be linearly independent from each other. Then, a linear regression can be computed.

5.2. Multiple linear regression in latitude, longitude, altitude

We start with latitude, longitude, and altitude as predictors. These factors are reasonable because of the following reasons: latitude characterizes the climate due to the solar angle, which is, by far, the most dominating factor for Region

VI. The longitude is the alternative for land-sea contrasts or continentality, which explains much of the seasonal variations. Finally, the altitude is included, because all climate state variables increase or decrease more or less with height above sea level. For monthly mean temperature, this factor is generally weak within Region VI compared to latitude and longitude but not negligible, especially in mountainous areas. The amount of variation of temperature as function of altitude varies largely from month to month, depending on season and the prevailing weather type during the month. In most cases, monthly mean temperature decreases with altitude, but in winter months, when inversion weather types are prevailing, a slight increase with altitude can also happen. For this reason, the regression model is fitted for each month separately.

The linear approach in this special case yields:

$$Y = a \cdot latitude + b \cdot longitude + c \cdot altitude + k \,. \qquad (2)$$

This is a specialization of the general approach Eq. (1). The three predictors (latitude, longitude, and altitude) represent the three spatial dimensions which are obviously orthogonal and, therefore, independent from each other. The coordinates are mostly well known for each station, thus, these predictors are mostly easily available. The fitting of the multi-linear regression has been done using the method of least squares (see, e.g., *Mosteller* and *Tukey*, 1977).

5.3. Continentality impact

For improving the approach, the longitude is replaced by a suitable continentality index. The continentality is a function of latitude and the annual temperature amplitude, which is calculated by the difference of the long-term means (1961–1990) of the maximum temperature in summer (June to August) and that of the minimum temperature in winter (from December to February). That calculation of the annual temperature amplitude is only an approximation for simplifying the computation, but does not reflect exactly the real annual amplitude. For example, March, which belongs to spring, is sometimes the coldest month in the year because of the drifting ice in bays near Finland in the Baltic Sea. In the literature, there are various versions of continentality indices (see, e.g., *Blüthgen,* 1980). Many equations show that the continentality for Europe can be described by a function of latitude and the annual temperature amplitude. One example is the approach by *Iwanow* (1959). Hogewind (*Hogewind*, 2010) modified this index to obtain a better suitability for the Region VI:

$$k = 260 \cdot \frac{annual \;\; amplitude}{latitude\, \varphi}, \qquad \textit{Iwanow}\ (1959)\ \text{and}$$

$$k = \frac{110 \cdot annual \;\; amplitude}{(latitude \;\varphi + 6)}, \qquad \textit{Hogewind}\ (2010).$$

This modification results in four classes in the range of the index: between 0 and 25 (highly maritime), from 26 to 50 (maritime), from 51 to 75 (continental), and from 76 to 100 (highly continental) over Region VI and its surroundings, and a threshold of around 50 between prevailing maritime and prevailing continental areas (*Fig. 3*).

Taking the continentality into account, the modified regression approach reads:

$$Y = a \cdot latitude + b \cdot altitude + c \cdot annual \;\; amplitude + d \cdot continenta\,lity + k\,. \qquad (3)$$

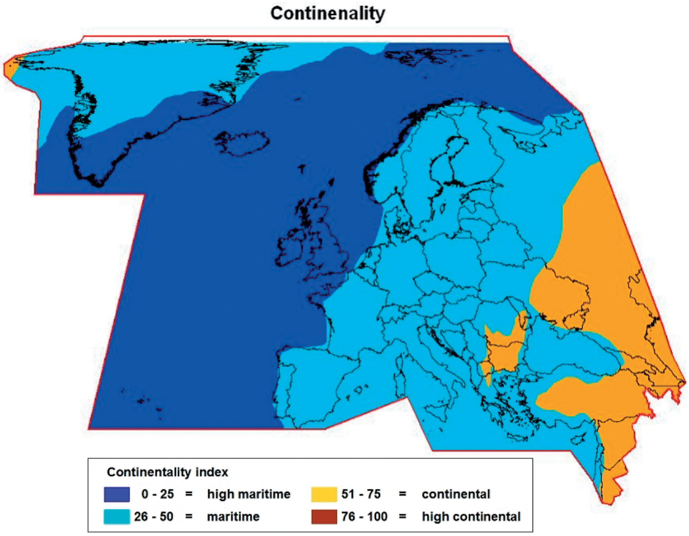


Fig. 3. Continentality (*Hogewind*, 2010).

This is now a non-linear approach in the explanatory variables, because continentality is a non-linear function of latitude, but the multiple regression is still linear, because a non-linear data transformation has been done (*Wilks*, 2006).

To get the residuals (T_{red}, the part of variability which is not explained by the regression model), the linear regression is subtracted from the original monthly mean temperature value for each station:

$$T_{red} = T_m - (a \cdot latitude) - (b \cdot altitude) - (c \cdot annual\, amplitude) \\ - (d \cdot continenta\,lity) - k\,. \qquad (4)$$

The considered parameters are now latitude, altitude, annual amplitude, and the newly created continentality index.

5.4. Interpolation of residuals

The reduced climate state variable, also known as residual (here T_{red}) is given for each measurement station and has to be interpolated into the area. The main question is now: which is the best suitable interpolation method?

The used software ArcGIS 9.2 Geostatistical Analyst offers a number of methods: inverse distance weighted, global polynomial interpolation, local polynomial interpolation, radial basis function, kriging, cokriging, and subtypes for each. From these methods a number of alternative approaches, which seem reasonable, are taken and applied to the computed residuals.

All these methods are described in the literature. An overview can be found in *Tveito et al.* (2008) including the mathematical background, the implementation in GIS software, and further references. The method "radial basis functions", which has been used for the final construction of maps in this paper, is described in the next section.

5.5. Radial basis functions

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that $\Phi(x)=\Phi(||x||)$, or, alternatively, on the distance from some other point c, called a center, so that $\Phi(x,c)=\Phi(||x-c||)$. Any function φ that satisfies the property $\Phi(x)=\Phi(||x||)$ is a radial function. The norm is usually the Euclidean distance, although other distance functions are also possible. For example, by using the Lukaszyk-Karmowski metric, for some radial functions it is possible to avoid problems with ill conditioning of the matrix solved to determine coefficients ω_i (see below), since the $||x||$ is always greater than zero.

Sums of radial basis functions are typically used to approximate given functions. This approximation process can also be interpreted as a simple kind of neural network.

The radial basis functions type used in this paper is multiquadratic ($r=||x-c_i||$):

$$\varphi(r)=\sqrt{r^2+\beta^2} \qquad \text{for some } \beta > 0. \tag{5}$$

Radial basis functions are typically used to build up function approximations of the form:

$$y(x)=\sum_{i=1}^{N}\omega_i \; \Phi(||x-c_i||), \tag{6}$$

where the approximating function $y(x)$ is represented as a sum of N radial basis functions, each associated with a different center c_i, and weighted by an appropriate coefficient ω_i. The weights ω_i can be estimated using the matrix methods of linear least squares, because the approximating function is linear in the weights.

Approximation schemes of this kind have been particularly used in time series prediction and control of non-linear systems exhibiting sufficiently simple chaotic behavior, and 3D reconstruction in computer graphics (*Lukaszyk*, 2004; *Buhmann*, 2003).

Eq. (6) can also be interpreted as a rather simple single-layer type of an artificial neural network, called a radial basis function network, with the radial basis functions taking the role of the activation functions of the network. It can be shown that any continuous functions on a compact interval can in principle be interpolated with arbitrary accuracy by a sum of this form, if a sufficiently large number of radial basis functions is used.

The approximant $y(x)$ is differentiable with respect to the weights ω_i. The weights could thus be learned using any of the standard iterative methods for neural networks.

There is a lot of literature about radial basis functions for further reading (e.g., *Baxter*, 1992; *Beatson et al.,* 2000; *Bors*, 2001; *Buhmann*, 2003; *Wei*, 1998).

5.6. Recomputing interpolated residuals

For recomputing the interpolated residuals to original data, the same regression equation, Eq. (4), as for reduction is used (Section 5.3). The difference is that this time the computing is not carried out for stations, but for the interpolated grid for the Region VI

$$\begin{aligned} T = T_{red} &+ (a \cdot latitude) + (b \cdot altitude) + (c \cdot annual\ amplitude) \\ &+ (d \cdot continentality) + k. \end{aligned} \tag{7}$$

Therefore, gridded data for latitude, altitude, annual temperature amplitude, and continentality are needed. Latitude is just a linear interpolation in meridional direction. For altitude, the grid GTOPO30 from U.S. Geological Survey is used with a recalculated resolution in 0.1°. The annual amplitude is interpolated by the interpolation method radial basis functions from station data, and finally, the continentality is computed from latitude and annual amplitude for each grid point (see Section 5.3).

5.7. Cross validation and root mean square error (RMSE)

To assess the quality of the spatial interpolation, a cross validation of the residuals has been carried out. This means that the spatial interpolation has been repeated after omitting one of the residual station values, and this has been done for each station value. Then, for all station points the difference between the residual station value and the corresponding interpolated value at this point has been computed. Finally, the root mean square error (RMSE) has been computed over the differences for all points, and then for each month and each of various interpolation methods, among them radial basis functions, several kriging approaches, and inverse distance weighted interpolation. Therefore, the RMSE is a quantity for estimating the mean interpolation error. However, it has to be kept in mind that the RMSE only can represent the information at the station points, but not for the whole area, and therefore, it does not exactly give the real mean interpolation error. Nevertheless, the estimate should be near to reality if the stations are representative for the area. As most stations are located in Central Europe, where the interpolation error is expected to be lower than in other more data sparse regions, the real mean interpolation error should be greater than the RMSE, which means that the RMSE can only give a minimum estimation. However, as the data base is the same for each method and each month (except for a few stations missing from month to month), the RMSE is a comparable measure of skill for each interpolation method.

6. Results

6.1. Results of the multiple regression

For the first approach (Eq. (2)), the three predictors (latitude, longitude, and altitude) explain a large part of the variance, generally over 70% for monthly mean temperature in Region VI for all months (*Table 1*).

Table 1. Explained variance in % for each of the predictors in Eqs. (2) and (3), for all months of the 1991–2000 average. Other periods have similar results

Month	**Latitude**	**Longitude**	**Altitude**	**Annual amplitude**	**Continen-tality**	**Lat+lon+alt (Eq. (2))**	**Lat+alt+amp+ cont (Eq. (3))**
Jan	60.98	5.36	3.90	50.78	10.30	70.29	93.83
Feb	66.21	4.85	2.52	43.93	6.13	73.74	93.44
Mar	74.32	2.60	1.17	30.75	1.42	80.07	91.33
Apr	83.92	0.68	0.23	13.73	0.76	89.08	90.72
May	85.02	0.00	0.09	3.36	7.50	90.12	89.02
Jun	80.85	0.54	0.42	0.00	18.63	88.04	88.93
Jul	79.83	0.31	0.86	0.25	23.21	85.74	90.03
Aug	83.78	0.01	0.55	0.12	17.56	88.56	91.33
Sep	88.20	0.84	0.04	5.96	5.17	92.54	93.82
Oct	86.28	2.39	0.64	16.47	0.33	91.73	94.94
Nov	75.07	5.86	2.91	35.66	2.74	83.81	95.89
Dec	65.77	6.95	3.98	47.29	8.08	75.72	95.07

The largest part is explained by latitude, especially in the warmer half year, due to the large variability of temperature as function of the solar angle. The explained part of variance in Eq. (3), using the predictors latitude, altitude, annual amplitude, and continentality, is considerably larger, especially during the colder months (from November to March) compared to Eq. (2), over 90%, due to the high impact of the annual amplitude particularly in winter, which has a high spatial variability within Europe. In the warmer half year, there is practically no or only a slight improvement concerning the explained variance by Eq. (3) compared to Eq. (2). However, the explained variance by Eq. (3) is within a range between 89 and 96% (rounded) for each month.

6.2. Results of the spatial interpolation

Results of the comparison between the various interpolation methods are shown in *Fig. 4*. For some of the interpolation methods and subtypes, unwanted interpolation islands appear (so-called bulls eyes), in particular for inverse distance weighting, global and local polynomial interpolation. Some kriging and cokriging subtypes are not exact at the station points and smooth too much. The interpolation method cokriging needs a second variable with the same resolution as the climate variable. This cannot be an impact variable, because this has already been removed by reduction. Some methods, especially cokriging, need quite a high computing time depending on spatial resolution, and thus, they are not convenient for operational use.

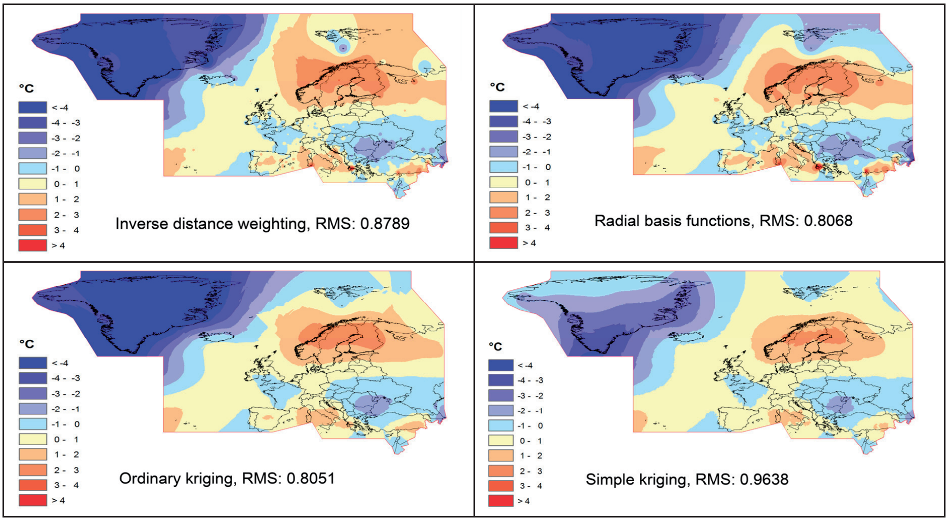
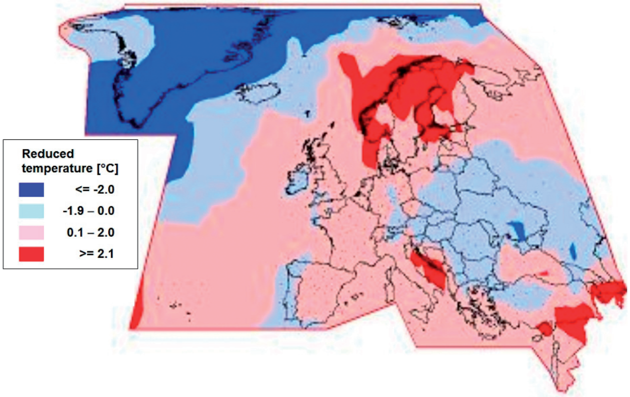


Fig. 4. Interpolated residuals (after subtracting the linear regression model) of annual temperature normal values of the period 1961–1990 using the following interpolation methods: inverse distance weighted (upper left), radial basis functions (upper right), ordinary kriging (lower left), simple kriging (lower right). RMS errors given for each method in K refer to the residuals.

Fig. 4 also shows the root mean square errors (RMSE) for various interpolation methods. The highest RMS errors of these show the interpolation methods simple kriging (*Fig. 4*, lower right) and inverse distance weighted (upper left). For inverse distance weighted, obvious interpolation islands can be clearly seen. Simple kriging amplifies point-to-point differences too much. The other two methods (ordinary kriging, *Fig. 4*, lower left, and radial basis functions, upper right), although basically different from each other, produce more or less the same results. The difference between these two interpolation methods is the longer computing time and the more difficult calibration of ordinary kriging because of every individual interpolation for the climate variable and period. As a result of this comparison of the different methods, the radial basis functions method with the subtype multiquadratic appears as the most suitable method for meeting our demands on operational map generation. The main advantages are exactness at data points (values at the data points are not changed after interpolation, except due to different altitudes and locations of the stations compared to the grid points), no smoothing, but no unrealistic interpolation islands either. The exactness at data points is also good to detect suspicious data on the map. The RMS error for the selected method is one of the lowest, the results are similar to ordinary kriging, but the computing is faster than kriging. Kriging, on the other hand, offers more possibilities of error assessment, but they are more difficult to interpret as they are not comparable with error assessments of other methods. Generally, the choice of the interpolation method matters only in data sparse areas. Otherwise, it is more important, when the regression error is higher or the data quality is worse.

The results of the described process need a further development which is described by *Hogewind* (2010). The different thermal conditions between land and sea require a separate regression over land and sea with separate regression coefficients for land and sea, but each applied to the whole RA VI area (*Fig. 5*). To consider coastal effects, the climate stations near the coast are used for both computing processes for overlapping land-sea areas. Furthermore, the data pool is increased by including the stations from the European Climate Assessment Dataset (ECA&D, www.knmi.nl). To study the space-time variability, the procedure has also been carried out for 10-year subperiods of the period 1951–2000. Examples of recomputed temperature fields for land and sea are shown in *Figs. 6a* and *6b* (for recomputation, a land-sea mask was used). These fields are overlaid to one complete map for the whole Region VI like a puzzle (*Fig. 7*). The effect of the thermal contrast between land and sea can be seen in various places, e.g., for Turkey.



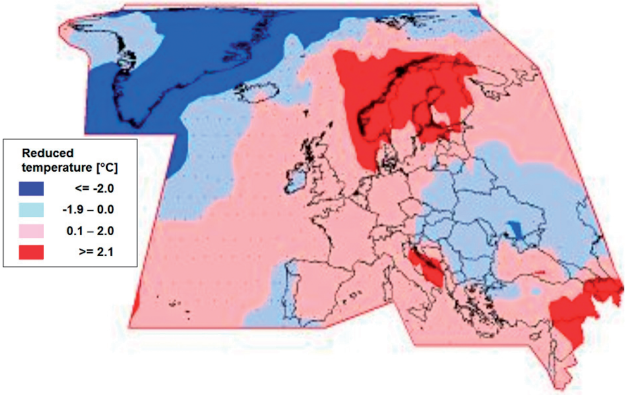


Fig. 5. Reduced temperature (residuals) over land (upper) and sea (lower) in September for the period of 1991–2000. Separate regression coefficients for land and sea are used, but applied to the whole RA VI area.

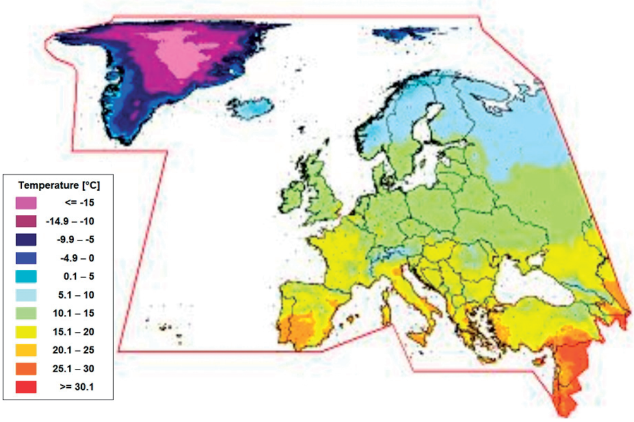


Fig. 6a. Recomputed temperature over land in September for the period of 1991–2000 White areas are excluded by using a land-sea mask.

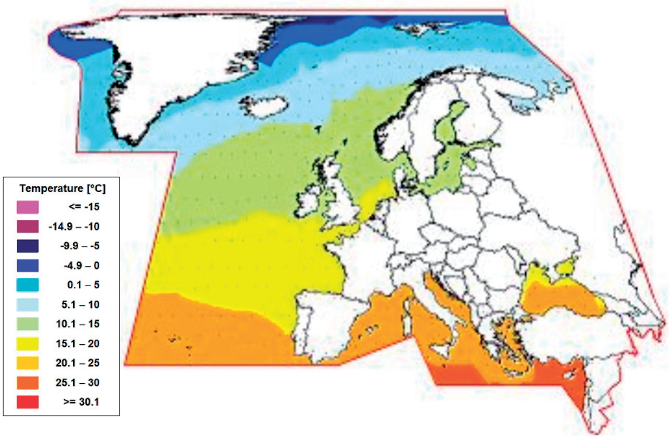


Fig. 6b. Recomputed temperature over sea in September for the period of 1991–2000 White areas are excluded by using a land-sea mask.

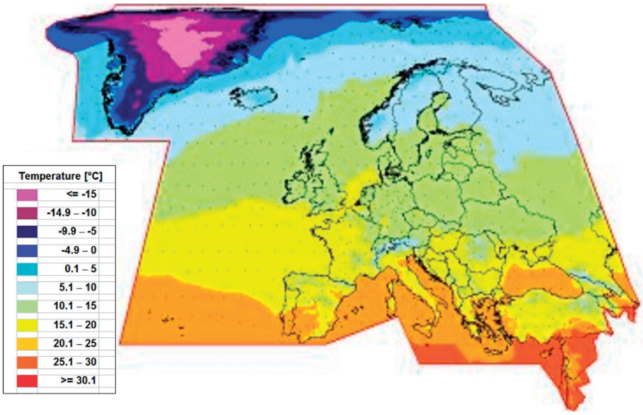


Fig. 7. Recomputed temperature in September for the period of 1991–2000 for the whole RA VI Region (consisting of separate calculations over land and sea as in Fig. 6).

7. Conclusions

The newly created continentality index (*Hogewind*, 2010) improves the regression model in comparison to longitude. The separate land-sea regressions improve the regression model, too. Nevertheless, the most important parameter for Region VI is still the latitude because of the strong influence of the angle of solar radiation. Residuals up to 2 K (RMSE 0.9 K) do not change, due to small scale effects. Radial basis functions turned out to be the most suitable interpolation method at the moment, at least for operational map production of monthly mean temperature in WMO Region VI. The method is exact, has a relatively low RMSE, can be realized very easily by using GIS software, and the interpolation can be computed in reasonable time. Probably the most promising effort to improve the results further is to enlarge and improve the data base and the regression model. Another challenge will be the application of this method to daily instead of monthly data.

Acknowledgments—First of all, we thank our colleagues of the district office of the DWD in Hamburg to prepare the CLIMAT data and data over the sea area, especially *C. Lefebvre* and *J. Busche*. Our technician *V. Zins* at DWD was very helpful in providing some of the figures.

We also thank the Hungarian Meteorological Service for the possibility of realizing this publication, and the language corrector of IDŐJÁRÁS for improvements of wording.

References

Alsamamra, H., Ruiz-Arias, J.A., Pozo-Vázquez, D., Tovar-Pescador, J., 2009: A comparative study of ordinary and residual Kriging techniques for mapping global solar radiation over southern Spain. *Agr. Forest. Meteorol. 149*, 1343-1357.

Baxter, B.J.C., 1992: The interpolation theory of radial basis functions. Cambridge University.

Beatson, R.K., Light, W.A., Billings, S., 2000: Fast solution of the radial basis function interpolation equations: Domain decomposition methods. *Society for Industrial and Applied Mathematics 22*, 1717-1740.

Bjornsson, H., Jonsson, T., Gylfadottir, S.S., Olason, E.O., 2007: Mapping the annual cycle of temperature in Iceland. *Meteorol. Z. 16,* 45-56.

Blüthgen, J., 1980: *Allgemeine Klimageographie*. 3. new revised edition. De Gruyter, Berlin (in German).

Bors, A.G., 2001: Introduction of the Radial Basis Function (RBF) Network. Online Symposium for Electronics Engineers.

Buhmann, M.D., 2003: *Radial Basis Functions: Theory and Implementations*. Cambridge University Press.

Claps, P., Giordano, P., Laguardia, G., 2008: Spatial distribution of the average air temperatures in Italy: quantitative analysis. *J. Hydrolog. Eng. 13*, 242-249.

Cristóbal, J., Ninyerola, M., Pons, X., 2008: Modelling air temperature through a combination of remote sensing and GIS data. *J. Geophys. Res. 113*, D13106.

Dobesch, H., Dumolard, P., Dyras, I., 2007: Spatial interpolation for climate data. The use of GIS in climatology and meteorology. *Geographical Information Systems Series*. ISTE, London.

Dolinar, M., 2006: Spatial interpolation of sunshine duration in Slovenia. *Meteorol. Appl. 13*, 375-384.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *J. Geophys. Res. 113*, D20119.

Hess, P., Brezowsky, H., 1952: Katalog der Grosswetterlagen Europas. *Ber. Deutscher Wetterdienst in US Zone*, Nr. 33.

Hiebl, J., Auer, I., Böhm, R., Schöner, W., Maugeri, M., Lentini, G., Spinoni, J., Brunetti, M., Nanni, T., Perčec Tadić, M., Bihari, Z., Dolinar, M., Müller-Westermeier, G., 2009: A high-resolution 1961-1990 monthly temperature climatology for the greater Alpine region. *Meteorol. Z. 18*, 507-530.

Hofstra, N., Haylock, M., New, M., Jones, P.D., 2009: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. *J. Geophys. Res. 114*, D21101.

Hofstra, N., Haylock, M., New, M., Jones, P., Frei, C., 2008: Comparison of six methods for the interpolation of daily European climate data. *J. Geophys. Res. 113*, D21110.

Hogewind, F., 2010: Raum-zeitliche Analyse der Klimavariabilität anhand von hochaufgelösten interpolierten Klimakarten am Beispiel von Europa und dem Nahen Osten (Region VI der Weltorganisation für Meteorologie). Preliminary results (in German, to be published in PhD thesis, University of Mainz, Germany).

Iwanow, N.N., 1959: Belts of continentality on the globe. (Izwest. Wsejoj. Geogr. Obschtsch. 91, 410-423 (in Russian). Cited in *Blüthgen* (1980), p. 590 (in German).

Lang, C., 1995: *Kriging Interpolation*. Department of Computer Science, Cornell University.

Lukaszyk, S., 2004: A new concept of probability metric and its applications in approximation of scattered data sets. *Computational Mechanics 33,* 299-304.

Mosteller, F., Tukey, J.W., 1977: *Data Analysis and Regression* – a second course in statistics. Addison-Wesley, Reading MA, ISBN 020104854X.

Perčec Tadić, M., 2010: Gridded Croatian climatology for 1961-1990. *Theor. Appl. Climatol. 102*, 87-103.

Schiemann, R., Liniger, M.A., Frei, C., 2010: Reduced space optimal interpolation of daily rain gauge precipitation in Switzerland. *J. Geophys. Res. 115*, D14109.

Stein, M. L., 1999: Interpolation of spatial data: Some theory for kriging. In *Springer Series in Statistics*. New York, Berlin; Springer.

Thornes, J.E., 2005: Special Issue on the use of GIS in Climatology and Meteorology. *Meteorol. Appl. 12,* i-iii.

Tveito, O.E., Wegenhenke, M., van der Wel, F., Dobesch, H., 2008: The use of geographic information systems in climatology and meteorology. *Office for Official Publications of the European Communities*, 245 pp.

Ustrnul, Z., 2006: Spatial differentiation of air temperature in Poland using circulation types and GIS. *Int. J. Climatol. 26*, 1529-1546.

Ustrnul, Z. and *Czekierda, D.*, 2005: Application of GIS for the development of climatological air temperature maps: an example from Poland. *Meteorol. Appl. 12*, 43-50.

Vicente-Serrano, S.M., López-Moreno, J.I., Vega-Rodríguez, M.I., Begueria, S., Cuadrat, J.M., 2010: Comparison of regression techniques for mapping fog frequency: application to the Aragón region (northeast Spain). *Int. J. Climatol. 30*, 935-945.

Wei, G.-Q., 1998: Three-dimensional reconstruction by radial basis function networks: Shape from shading and stereo vision. Munich, Dissertation.

Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2. Ed. Elsevier, Academic Press.

IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 51–70

Time variation of the effect of geographical factors on spatial distribution of summer precipitation over the Czech Republic

Jacques Célestin Moliba Bankanza

Institute of Atmospheric Physics,
Boční II 1401, 141 31 Prague, Czech Republic; E-mail: moliba@ufa.cas.cz

(Manuscript received in final form March 26, 2011)

Abstract—This study deals with modeling of spatial distribution of summer (JJA) precipitation over the Czech Republic. The aim is to analyze the time variation of the relationships between geographical factors and precipitation during summer. Various candidates geographical predictors are evaluated in the stepwise regression models for summer precipitation, namely: (1) a set of omnidirectional parameters of the elevation that characterize an area of 3×3 km around meteorological stations, (2) various cross products calculated on the basis of geographical coordinates and elevation or topographic parameters, (3) slope and four facets of slope aspect characterizing the orographic regimes in the Czech Republic, (4) land cover parameters describing an area of 10 × 10 km around meteorological stations, and (5) geographical coordinates. The orographic parameters are derived from the 1 km resolution digital elevation model (DEM); the land cover parameters are derived from the 1 km resolution CORINE (COoRdination of INformation on the Environment) land cover data. Daily precipitation data for the period 1971–2003 have been used. The precipitations were collected from 203 stations throughout the country. Stepwise regression models of summer precipitation are generated for each year, and each overlapping decade from 1971 to 2003. To ensure the stability of the regression equations and comparability of regression models in time, similar suitable and stable independent variables in time should be selected. Therefore, orthogonally rotated principal component analysis (PCA) and frequency of significant predictors entering models are used to select them. Multivariate regression precipitation models are generated using definitive (PCA or stepwise based) selected predictors. Ten independent geographical variables have been selected as the most important predictors for precipitation regression models. They consist of latitude, longitude, slope aspect of the grid westward from the central grid, slope aspect of the grid northward from the central grid, slope of the grid northeastward from the central grid, slope of the grid eastward from the central grid, slope of the grid northward from the central grid, maximum value of elevation (percentile 95%) of northwestern grid from the central grid, minimum value of elevation (percentile 5%) of the central grid, and vegetation. The relationships between these significant predictors and precipitation are stable in time. No significant trend in regression coefficients has been found during 1971–2003.

Key-words: multivariate regression, precipitation models, summer precipitation, geographic influence, environmental variables, temporal variability, Czech Republic

1. Introduction

Precipitation is a climatic variable that has a high spatial and temporal variability. Its spatial distribution is influenced by various factors such as terrain height, slope, etc. For mapping purposes, it is practical to estimate the effect of such factors on precipitation distribution. Spatial modeling is a suitable tool that allows to explore the relationship between the target variable and predictors, and to get continuous information on precipitation over a targeted area. Many studies have been undertaken recently to assess and model the relationship between the climatic variables, and independent factors. Several geographical variables, including land cover (*Joly et al.*, 2003), proximity to the water bodies (*Weisse* and *Bois*, 2001; *Vicente-Serrano et al.*, 2003; *Marquinez et al.*, 2003; *Daly et al.* 2002), atmospheric circulation (*Johnson* and *Hanson*, 1995; *Basher* and *Zheng*, 1998; *Courault* and *Monestiez*, 1999), and topography (*Johnson* and *Hanson*, 1995; *Goodale et al.*, 1998; *Daly et al.*, 2002) have been frequently used as relevant independent variables to model spatial patterns of precipitation. The latter has a significant influence on spatial variability of precipitation (*Joly et al.*, 2003; *Weisse* and *Bois*, 2001, etc.). Therefore, numbers of these studies have been focused on modeling the influence of topographic features on the spatial variability of climate variables (*Prudhomme* and *Reed*, 1999; *Johnson* and *Hanson*, 1995; *Drogue et al.*, 2002; *Weisse* and *Bois*, 2001; *Diodatto*, 2005, etc.).

According to its geographical position in Central Europe, the Czech Republic is subject to both oceanic and continental influences. Topographically, the inner part of the country is dominated by lowlands and surrounded by highlands. Such topographic feature contributes to modifying airflow over the country and can induce a strong convective precipitation, especially in the mountains (Moravsko-slezské Beskydy, Jeseníky, Krkonoš, Jizerské hory, and Krušné hory). Extreme precipitation events are more frequent and intense over these highlands due, among other factors, to the influence of exposition to airflow (*Kakos*, 2001).

In this study, the relationships between geographical factors and summer precipitation are examined through a stepwise regression model. Summer precipitation is analyzed instead of other seasons for several reasons. First, the annual cycle of precipitation in the Czech Republic is characterized by a tendency for maximum rainfall during summer. Therefore, summer precipitation contributes significantly on the character of the precipitation fluctuation (*Tolasz et al.*, 2007). Second, summer precipitation, usually of shorter duration and greater intensity (*Tolasz et al.*, 2007), is characterized by the high frequency of occurrence of extreme precipitation events (*Kaspar* and *Muller*, 2008), which are often connected with several natural hazards including hydrological flood and soil erosion. The need of precipitation information during summer time is crucial for the risk management. Moreover, spatial models examine spatial dependence of climate variables using a single time realization of the variable, i.e., they widely use the mean values for a given period, as input. However, the performance of a model depends not only on the density of the station network and the choice of methods, but also on the temporal variability (*Hulme et al.*, 1997). Therefore, the choice of the period of study can bias the results of interpolation (*Hulme et al.*, 1997). The spatial variability of environmental variables is commonly a result of complex processes working at the same time and over long periods of time, rather than an effect of a single realization of a single factor (*Hengl*, 2007). Geostatistics are less powerful than the statistical climatology based on sample in time, because they are based on single realization in time *(Szentimrey* and *Bihari*, 2007). The temporal variability seems to be an important task in modeling spatial variation of climate variables. This aspect has received substantial attention in several studies: *Basher* and *Zheng* (1998) take into account seasonal behavior of precipitation (ENSO variations) for mapping precipitation patterns of a data-sparse tropical south-west Pacific Ocean region. *Brown* and *Comrie* (2002) created the 39-year time series of maps and datasets of winter temperature and precipitation for the southwest US by comparing 30 years (1961–1990) modeled means with 39 observed winter temperature and precipitation values. *Johnson* and *Hanson* (1995) modeled the relative contribution of topographical and meteorological variability to regional precipitation variability. In order to improve interpolation of spatially generated weather data, *Baigorria et al.* (2007) analyzed changes in spatial correlations and compared spatial correlation on daily and monthly basis. Therefore, using seasonal rainfall amounts, temporal analysis is needed to find and determine: (1) how the relationships between independent variables and precipitation vary within years and decades; (2) how the model is affected by temporal changes. The aims of this study are: (1) to model spatial pattern of summer (JJA) precipitation in the Czech Republic at year and ten-year time steps from 1971 to 2003 using geographical variables *as independent variables;* (2) to analyze the time variation of the relationships between geographic variables, and the summer precipitation during 1971–2003.

2. Datasets

Digital Elevation Models (DEM)

DEM with the resolutions of 100 m and 1 km have been used. The fine spatial resolution of topographic and elevation variables have been retained in this study, because large-scale topographic features at a resolution of 1–15 km yield a high correlation with precipitation (*Daly et al.*, 1994; *Daly*, 2006). The 100 m DEM resolution data have been used to calculate the following smoothed elevation parameters: (1) the upper (percentile 95%) and (2) lower (percentile 5%) percentiles of elevation for a grid of 1 km resolution, and (3) mean elevation for each grid with 1 × 1 km resolution. On the other hand, slope, and slope aspect are obtained directly from the 1 km resolution DEM data.

Land cover data

Land cover data are obtained from the CORINE (COoRdinate INformation on the Environment) land cover dataset. These data are available in the following link: http://www.dataservice.eea.europa.eu/dataservice. They describe the land cover units in Europe. According to the CORINE land cover classification, four main types of landscape characterizing the Czech Republic were identified, and used as candidate geographical independent variables. They are related to vegetation, agricultural area, water bodies, and artificial areas.

Precipitation data

Daily precipitation data for the period 1971–2003 have been used. The dataset consists of 203 stations distributed over the whole country (*Fig. 1*). Meteorological stations are unevenly distributed across these different land cover units and topographic patterns. Most of them are distributed across urban (towns, small cities, villages) and agricultural areas. Only few stations are located in vegetation-covered areas. Considering ground elevation, about 80% of meteorological stations are located below 600 m. Only 12% of them are located above 600 m on the highlands or mountainous regions that have a significant influence on precipitation distribution (*Table 1*). The lack of observation in forested and mountainous areas shows how much it is important to model the relationships between rainfall, and elevation and/or other geographical variables.

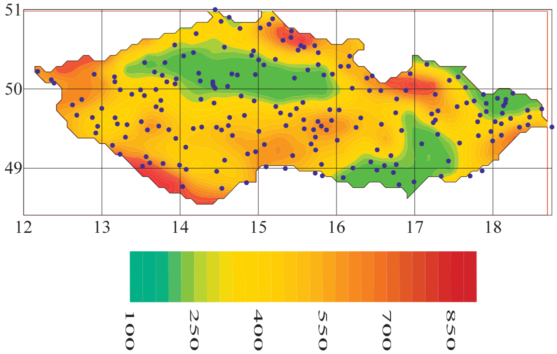


Fig. 1. Spatial distribution of meteorological stations with regard to DEM data (in m). Dots represent geographic position of meteorological stations.

Table 1. Elevation of meteorological stations

Elevation (m)	Number of stations	% of total area
< 200	12	6
200 – 400	85	42
400 – 600	82	40
> 600	24	12

In order to perform time variation analysis (see Section 4), two subsets of summer precipitation series are derived from the summer precipitation dataset: (I) yearly summer precipitation amounts and (II) the overlapping decade precipitation mean, with a shift of one year from 1971–1980 to 1994–2003. The lengths of both subsets of summer precipitation series are 33 and 24, respectively.

3. Methods: model development

3.1. Independent variables

Spatial fields of precipitation are correlated with many environmental or geographic factors especially elevation and geographic coordinates. In this study, 54 candidate independent variables, which can explain spatial variability in the climate data, have been evaluated, and then selected (*Table 2*). A large number of geographical variables are evaluated, because none is a priori the most important. They are related to:

- Omnidirectional variables describing elevation (27 variables), topographic features (slope and slope aspect: 13 variables). Those morpho-topographic variables represent the values from the grids omnidirectionally oriented around central grids. The central grids can be defined as grids in which stations are located. Eight directions around the central grid have been defined as: (1) north, (2) east, (3) south, (4) west, (5) north-east, (6) south-east, (7) south-west, and (8) northwest.
- Cross products involving (eight variables): (a) geographical coordinates and elevation variables (maximum: percentile 95%, minimum: percentile 5%, and average elevation to north and east) and (b) geographical coordinates and topographic features (slope). Cross products were calculated (as indicated in *Table 2*) to obtain west-east or south-north gradient of elevation and topography (*Brown* and *Comrie*, 2002; *Vicente-Serrano et al.*, 2003).
- Land cover parameters were selected from a grid data with spatial resolution of 1 km resolution. An area of 10×10 km around meteorological stations was delimited and four indexes (Iwat, Iveg, Iagr, Iurb) characterizing the main units of landscapes (water bodies,

vegetation cover, agricultural area, and urban and artificial area) in this area were calculated. Indexes are calculated as a ratio (%) of the area covered by land cover units to the total area (grid) of 10 × 10 km resolution.

- Geographic coordinates: latitude (Lat) and longitude (Long).

Table 2. Candidate independent variables for stepwise regression models. They are related to the Elevation variables, topographic variables (slope and slope aspect), cross products involving geographical position and morpho-topographic variables (elevation and topographic variables), geographical coordinates, and land use or cover variables. Abbreviations Emax, Emin, Eavg are related to maximum, minimum, and average elevation. Slope is related to slope values in %, while Oreg abbreviates slope aspect. The numbers after letters indicate orientation of grid from which slope or elevation values have been taken out, inside a bound of 3 × 3 km around the central grid. Central grid is defined as a grid in which meteorological stations are located. For elevation and slope eight directions have been defined (see section 3.1), while for slope aspect 4 have been taken into account (1 for east, 2 for west, 3 for north, and 4 for south). Abbreviations ending with "-grad" are related to cross product involving longitude (with number 1 on the end of letters) or latitude (with number 2 on the end of letters) and elevation variables (Egrad, Exgrad, Engrad) or slope (Sgrad). Iagr, Iveg, Iurb, and Iwat abbreviate four Indexes of landscape units (agriculture, vegetation, urban area, water bodies)

Candidate independent variables	**Abbreviations**	**Candidate independent variables**	**Abbreviations**
Central grid average elevation	Eavg0	Central grid slope values	Slope0
Central grid minimum elevation (percentile 5%)	Emin0	Slope in the north	Slope1
Central grid maximum elevation (percentile 95%)	Emax0	Slope in the east grid	Slope2
Average elevation in the north from the central grid	Eavg1	Slope in the south	Slope3
Minimum elevation (percentile 5%) in the north	Emin1	Slope in the west	Slope4
Maximum elevation (percentile 95%) in the north	Emax1	Slope in the north-east	Slope5
Average elevation in the east from the central grid	Eavg2	Slope in the south-east	Slope6
Minimum elevation (percentile 5%) in the east	Emin2	Slope in the south-west	Slope7
Maximum elevation (percentile 95%) in the east	Emax2	Slope in the north-west	Slope8
Average elevation in the south from the central grid	Eavg3	Slope facet east	Oreg1
Minimum elevation (percentile 5%) in the south	Emin3	Slope facet west	Oreg2
Maximum elevation (percentile 95%) in the south	Emax3	Slope facet north	Oreg3
Average elevation in the west from the central grid	Eavg4	Slope facet south	Oreg4
Minimum elevation (percentile 5%) in the west	Emin4	Cross product Long × average elevation	Egrad1
Maximum elevation (percentile 95%) in the west	Emax4	Cross product Lat × average elevation	Egrad2
Average elevation in the north-east from the central grid	Eavg5	Cross product Long × slope	Sgrad1
Minimum elevation (percentile 5%) in the north-east	Emin5	Cross product Lat × slope	Sgrad2
Maximum elevation (percentile 95%) in the north-east	Emax5	Cross product Long × 95% percentile of elevation	Exgrad1
Average elevation in the south-east from the central grid	Eavg6	Cross product Long × 5% percentile of elevation	Engrad1
Minimum elevation (percentile 5%) in the south-east	Emin6	Cross product Lat × 95% percentile of elevation	Exgrad2
Maximum elevation (percentile 95%) in the south-east	Emax6	Cross product Lat × 5% percentile of elevation	Engrad2
Average elevation in the south-west from the central grid	Eavg7	Index for the ratio of Agricultural area	Iagr
Minimum elevation (percentile 5%) in the south-west	Emin7	Index for the ratio of the Vegetation covered area	Iveg
Maximum elevation (percentile 95%) in the south-west	Emax7	Index for the ratio of Urban area	Iurb
Average elevation in the north-west from the central grid	Eavg8	Index for the ratio of the Water bodies	Iwat
Minimum elevation (percentile 5%) in the north-west	Emin8	Longitude	Long
Maximum elevation (percentile 95%) in the north-west	Emax8	Latitude	Lat

3.2. Selection of suitable predictors for regression model and temporal analysis

Analysis of time variation of relationship between precipitation and geographical factors mentioned in *Table 2* was the objective of this study. Relationships are analyzed through regression models that are performed at various time steps. In order to assure stability of the regression equations and comparability of models in time, it was necessary to select similar suitable and stable independent variables. Selection of suitable independent variables was based on two approaches: stepwise-based approach (STW), and principal component analysis (PCA)-based approach, both for yearly-based precipitation models (STW I / PCA I) and overlapping decade-based precipitation models (STW II / PCA II).

3.2.1. Stepwise regression based models (STW)

Stepwise selection of suitable predictors has been made in two steps. At the first step, the significant candidate independent variables have been selected for each model at various time steps. At the second step, only the most frequently selected significant predictors have been taken into account.

(a) It is important to remind that the set of geographical variables used as predictors (*Table 2*), particularly topographic and elevation parameters, are collinear. The choice of suitable predictors from this set has a significant influence on the behavior of models. Hence, forward stepwise linear regression was used to model summer precipitation as function of the collinear geographical factors at time steps of annual (STW I) and overlapping decades (STW II) from 1971–2003. All predictors mentioned in *Table 2* are used. A p-value of 0.05 has been used to force out of the model any non-significant effects, and to select significant, and non-collinear independent variables. Stepwise regression has been used in many studies (*Ninyerola et al.*, 2000, 2007; *Marquinez et al.*, 2003; *Vicente-Serrano et al.*, 2007) as an accurate method in examining relations between precipitation and collinear independent variables.

(b) On the second step, the most frequently selected significant predictors by both STW I and STW II-based models was considered. A threshold frequency value of 20 – 40% was defined to select them. Geographical variables of which frequency value does not reach at least 20% are considered as improper for temporal analysis, and are discarded; while independent variables exceeding the defined threshold are retained. If the retained variables are collinear, the operation is repeated (using a higher threshold value, i.e., 30 – 40%) until no co-linearity is found among them. Using this procedure, six predictors have been retained for both STW I and STW II-based precipitation models (see *Table 3* and *Figs. 2* and *3*). For the STW I-based models, the selection was ended at the first step, where frequency of significant independent variables exceeded 30%. However, for the STW II-based models, 11 collinear predictors have been selected at the first step (*Fig. 3a*). The operation was repeated for the 11 predictors to select a final series of six non-collinear predictors exceeding 40% (*Fig. 3b*).

Table 3. Definitive selected independent variables for all models

I. Year by year time series			**II. Moving long-term precipitation mean**		
		Models			
STW I	**PCA-A I**	**PCA-B I**	**STW II**	**PCA-A II**	**PCA-B II**
Slope2	PCI	Emin0	Slope2	PCI	Emin0
Slope5	PCII	Slope1	Slope5	PCII	Slope1
Emax8	PCIII	Vegetation	Emin8	PCIII	Vegetation
Lon	Lon	Lon	Lon	Lon	Lon
Lat	Lat	Lat	Lat	Lat	Lat
Oreg2	Oreg2	Oreg2	Oreg2	Oreg2	Oreg2
–	–	–	–	Oreg3	Oreg3

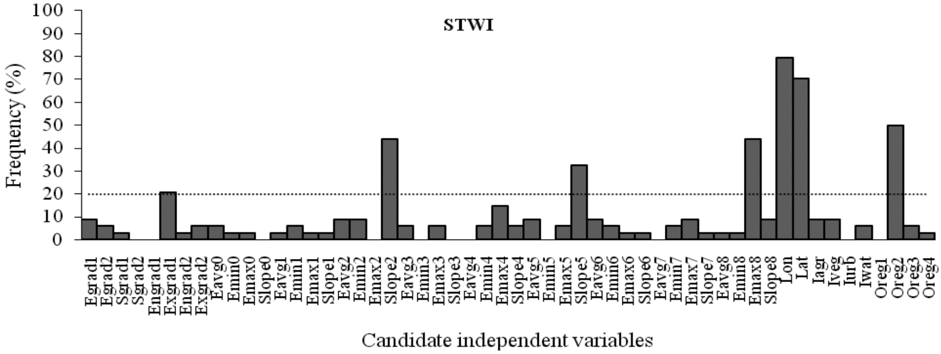


Fig. 2. Frequency of significant variables entering models for STWI-based models.

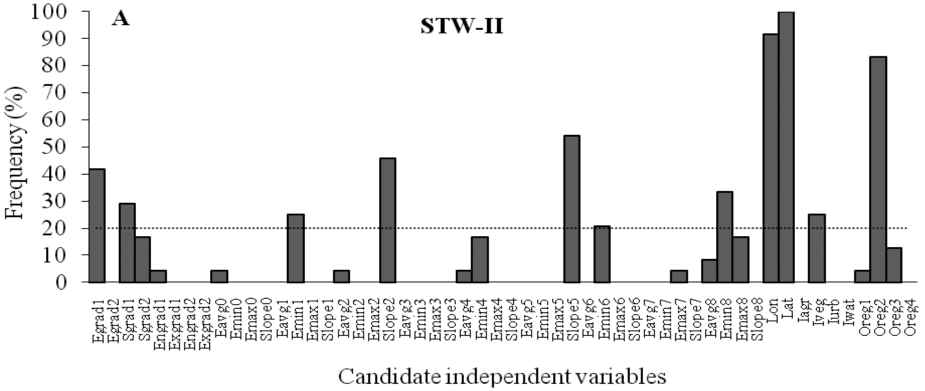


Fig. 3a. Frequency of significant independent variables for STW II-based models. Dotted line represents the threshold value for selecting more frequent predictors. First step (A) of selection.

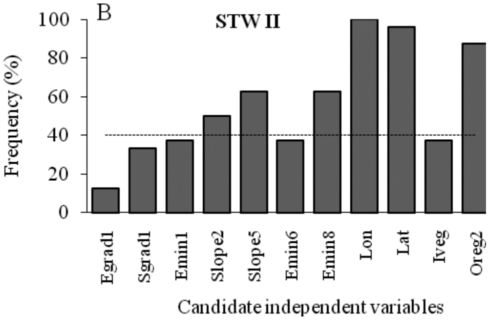


Fig. 3b. Frequency of significant independent variables for STW II-based models. Dotted line represents the threshold value for selecting more frequent predictors. Second step (B) of selection.

3.2.2. PCA – based models (PCA)

Principal component analysis (PCA) has been used to eliminate random variability in independent factors, and to generate stable models in time. The sets of independent variables have been checked for correlation before performing PCA. Eight independent variables, which were less correlated or uncorrelated with other independent variables, were discarded for the PCA. They were: geographical coordinates (longitude and latitude), two units of land cover (urban area and water bodies), and four orographic facets (slope aspect). PCA was performed for the remaining 46 variables. The number of principal components (PCs) to retain for rotation was given using screen test. Three PCs have been retained. They are related to: (1) characteristics of elevation, (2) topographic features, (3) land use and land cover parameters: agricultural and vegetation covers. The three PCs explain about 90.3% of total variability. Morpho-topographic variables that have the highest loadings with those retained PCs were then selected. For the further regression models, both PCs scores (three variables), and independent variables (three morpho-topographic variables), that were selected assuming highest loading, were considered as candidate independent variables. These candidate predictors selected using PCA were recombined with the 8 discarded variables before performing PCA. Then the stepwise regression has been performed to select significant and noncollinear variables. The frequencies of variables (*Fig. 4*) have been analyzed (considered as in Section 3.2.1.b). The frequent variables have been considered as stable and, therefore, suitable for multivariate regression model. This approach helped to avoid the problem encountered during interpretation of the PCs that involve independent uncorrelated variables. The final selected independent variables are shown in *Table 3*. Using this approach, six significant independent variables were selected for yearly-based precipitation models (PCA I) and seven independent variables were selected for decade-based precipitation models (PCA II) (see *Table 3*). PCs scores (PCA-A) or selected variables assuming the highest loadings (PCA-B) have been used to compare their effect on the models. *Table 3* shows important geographical variables that influence significantly the spatial patterns of precipitation in the Czech Republic. All approaches selected the geographical coordinates (including continentality) and the westward slope aspect.

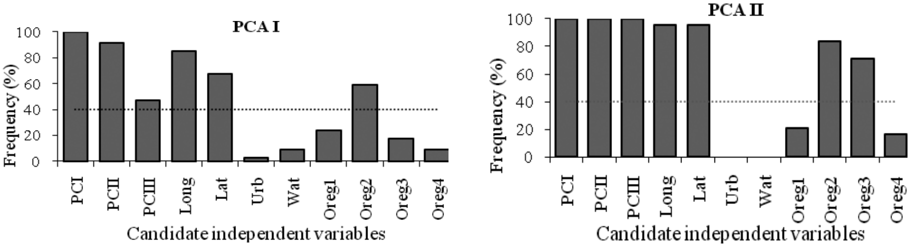


Fig. 4. Frequency of significant variables entering models for PCA I and II-based models. Dotted lines represent the threshold value for selecting more frequent predictors.

3.3. Regression models and trend detection

Once stable significant variables were selected, multivariate regression precipitation models were then performed using them as predictors. The multiple regression relationship is obtained through the following equations:

$$\begin{cases} P_{est}(t_1)=b_0+b_1(x_1)+...+b_n(x_n) \\ P_{est}(t_2)=b_0+b_1(x_1)+...+b_n(x_n) \\ ... \\ P_{est}(t_n)=b_0+b_1(x_1)+...+b_n(x_n) \end{cases}, \qquad (1)$$

where $b_{1\text{–}n}$ is the multiple regression coefficient adjusted for each retained independent variable x_n; P_{est} represents the predicted rainfall, and $t_{1\text{–}n}$ is the time step (i.e., year or decade).

In order to carry out temporal analysis of the relationship between significant geographical factors and precipitation, time series of regression coefficients from each time resolution and approaches-based models were built. The linear trend in those series was estimated using a least-squares regression. The significance of trends was determined using the confidence interval (*CI*) given by the following equation:

$$CI=t+/-\frac{2.042\cdot\sigma e}{\sqrt{n}\cdot\sqrt{\sigma x^2}}, \qquad (2)$$

where t is the trend value, n is the length of the time series of the regression coefficients, σe is the standard deviation of the residuals, and σx is the standard deviation of the independent variable.

4. *Results*

4.1. Model performance

Several standard statistical measures of models performance and accuracy were calculated. The goodness of fit of the model and the proportion of the variation of summer precipitation explained by the model are measured by the coefficient determination (R^2). The magnitude and sign of errors of the regression model are given by mean absolute error (MAE) and root mean square error (RMSE).

The time variation of the coefficients of determination for all approaches and time resolution-based precipitation models is displayed in *Fig. 5*. The magnitude of errors is measured by the rootmeansquare error (*Fig. 6*) and mean absolute error (*Fig. 7*).

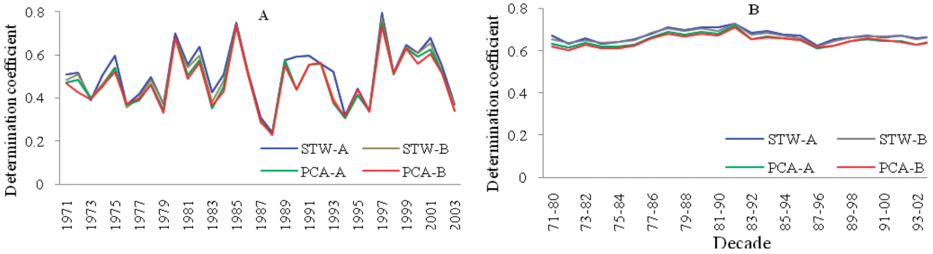


Fig. 5. Time variation of determination coefficients of STW and PCA-based models for year (A) and decade (B) time resolution.

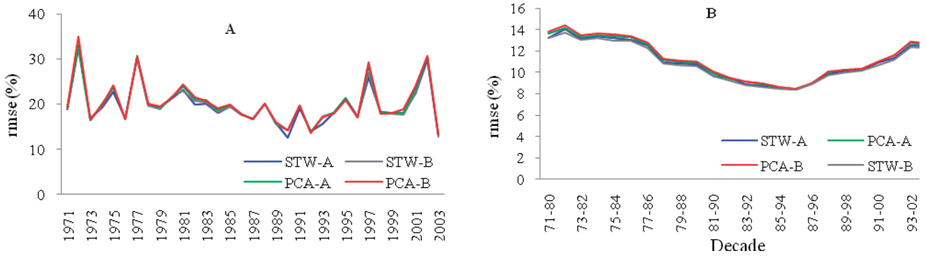


Fig. 6. Time variation of the root-mean-square errors of STW and PCA-based models for year (A) and decade (B) time resolution.

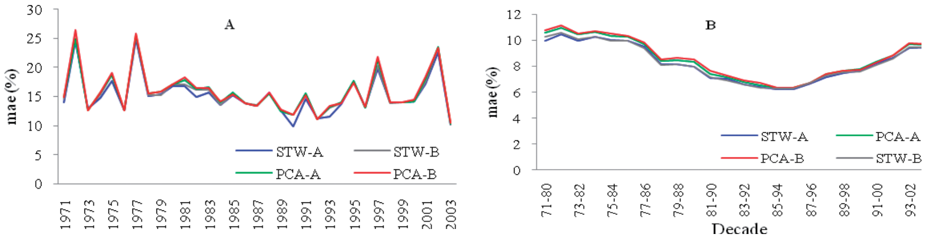


Fig. 7. Time variation of the mean absolute errors of STW and PCA-based models for year (A) and decade (B) time resolution.

The coefficients of determination and errors for model based on different approaches vary consistently in time, whereas their time variations are inconsistent for time resolution-based models. In fact, the STW-based models are more efficient than PCA-based models. Nevertheless, the difference between both approaches-based models is minor. Model accuracy varies widely among models generated at various time resolutions. Models based on long-term time resolution are more accurate and explain more variation than those based on shorter time resolution. An important variability in summer precipitation (> 0.61 for both approaches-based models) is consistently captured by ten-year models for both PCA and STW-based approaches. The magnitude of errors (RMSE, MAE) does not reach 15% (37 mm) of summer precipitation for ten-year models. On the contrary, terrain influences did not always account for an important variability of summer precipitation for the annual precipitation models. The coefficients of determination of these models describe an important inter-annual variability; they vary from 0.23 (1988) to 0.7 (1997). Similarly, the magnitude of errors is fluctuant. The larger (35% – 86 mm) and the smaller RMS errors (13% – 32 mm) have occurred in 1972 and 2003, respectively. Unlike other years, where model errors are inversely proportional to the explained variance, both variance explained and error yielded by the models are large in 1997 and 2002.

Several models were unable to capture more than 50% of variability in precipitation and to generate small error. Considering that a good precipitation model must capture at least 50% of precipitation variability (*Ninyerola et al.*, 2000) and yield very small prediction error, we can conclude that numbers of these models failed. The temporal variability of summer precipitation combined with the ability of the set of predictors to capture variation in data can explain it. *Fig. 8* displays the relationships between rainfall departure (from the long-term precipitation average) and coefficient of determination. It reveals that the stronger the negative rainfall anomalies for a given year, the smaller the variation explained in precipitation data. Indeed, the coefficients of determination fall under 0.5 or rarely (once) overtake this value during dry years. Therefore, additional predictors (different from the used set of predictors) or additional analysis on removing temporal variability were needed to improve models. If it is necessary to add other auxiliary variables, then it should be important to consider that not only geographical factors but also other factors such as atmospheric circulation affect the spatial distribution of precipitation. However, their mechanisms are more complex and not easy to evaluate their statistical relationships.

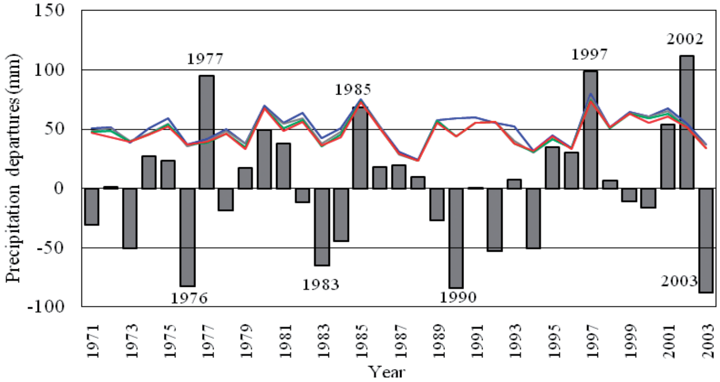


Fig. 8. Comparison of determination coefficients (color lines) with precipitation departures (histogram).

4.2. Time variation of effect of significant geographical factors

The relationships between summer precipitation and geographic variables, selected using STW approach and PCA, have been examined through regression model. Resulting coefficients of regression for each approach and time resolution-based model are plotted in *Figs. 9–12*. Six independent variables have been selected using STW approach: slopes of grids eastward and northeastward from the locations, maximum elevation northwestward from locations, geographic coordinates (longitude and latitude), and slope aspect westward.

At the annual time resolution, spatial patterns of summer precipitation show an increase of precipitation with a growing value of elevation and slope aspect. However, they show an increase and a decrease of precipitation to increasing latitude, longitude, and slope. Considering the inter-annual variability of the relationships between precipitation and each significant geographical variable, it can be pointed out in some extreme cases of strong relationships. They are found between precipitation and elevation (emax8), latitude, slope (slope5), and slope aspect (oreg2), respectively, during 1980, 2002, 1997, and 1972. The precipitation models based on STW approach reveal that the spatial pattern of precipitation during the heavy precipitation events of 1997 and 2002 were strongly related to the slope northeastward and latitude, respectively. The fields of intense precipitation during such years have a significant influence on the spatial pattern of precipitation across the whole country. For example, the decrease of precipitation with an increasing longitude during 1997 is mainly due to the fields of intense precipitation in the northeastern part of the country. Spatial patterns of summer precipitation and precipitation amount fluctuate in time according to different factors such as atmospheric conditions.

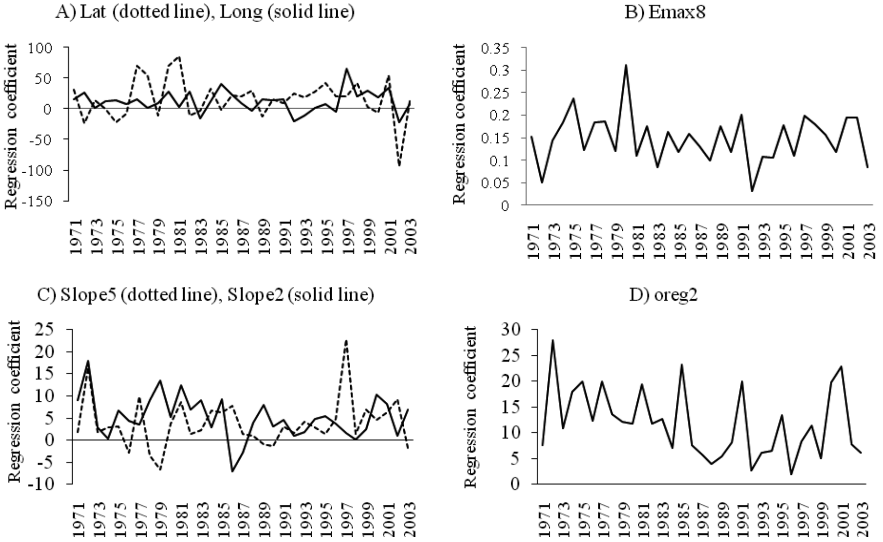


Fig. 9. Time variation of the regression coefficients of significant predictors (as indicated in A, B, C, D) for STW I-based models.

During the whole period under study, the precipitations have had a positive relationship with elevation variables. The well-known effect of elevation on the spatial pattern of summer precipitation, which is related to an increase of precipitation with growing elevation, is observed along the entire period. The highlands are seemingly rainier during summer than lowlands all over the country. However, during some years (1976, 1986, 2003, etc.) spatial pattern of precipitation is characterized by a decrease of rainfall with a growing value of slope (slope2, slope5). Thus, if multiple factors of mountain regions are taken into account, their relationships with precipitation become more complex than a simple increase with growing elevation. During the period 1971–2003, the precipitations increased strongly in 1980 and slightly in 1992 with maximum elevation northwestward from the locations. The slope orientated eastward from the locations has influenced considerably the summer precipitation during 1972.

For the decade-based models, minimum elevation northwestward (Emin8) has been selected as predictor instead of Emax8. Unlike the yearly-based models, no geographical variable was related to any decreases of summer precipitation. Although the effect of minimum elevation (Emin8) on precipitation is less variable in time, the influences of the topographic and geographical position fluctuate largely in time. The variation of the effect of topographic features (slope5, slope2, and oreg2) shows two peaks at the beginning and at the end of the period under study. The influence falls in the middle of this study period (i.e., during 1981–1997). This can be related to drought that occurred in the Czech Republic in this period (*Kaspar* and *Muller*, 2008).

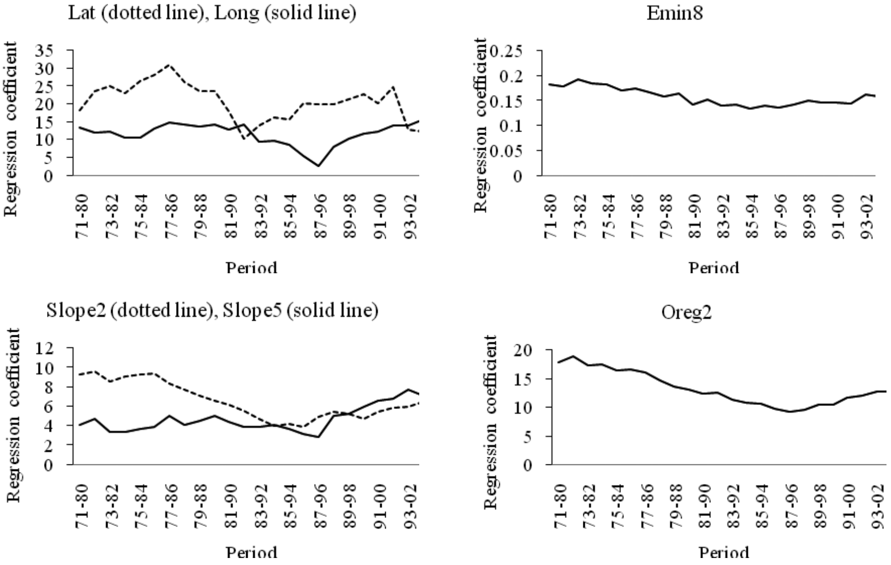


Fig. 10. Time variation of the regression coefficients of the significant predictors for STW II-based models.

Time variation of the relationships between spatial patterns of summer precipitation and significant geographical variables selected using PCA-based approach has been also analyzed (*Figs. 11* and *12*). New variables were selected using the loading of the components: slope northward from the locations (Slope1), vegetation (Veg), and minimum elevation at the locations. An additional geographical variable was specifically selected for the ten-year models: slope aspect oriented northward. These additional predictors influence the spatial patterns of precipitation as well. They are related to the increase of precipitation, except the vegetation during some years (1984, 1992, and 2002). A strong relationship with precipitation is observed during 1972 for slope2 and in 1995, 2002 for the vegetation.

The PCA-based models show an increase of summer precipitation for the two PCs scores during the period considered by this study (*Figs. 11* and *12*). In particular, spatial patterns of summer precipitation in 1972 and 1980 have been related to PC1 and PC2, respectively. Similarly to STW-based models, PCA-based models reveal that spatial distribution of precipitation during the summer of 1997 is related to topographic features and geographical coordinates, while spatial distribution of summer precipitation in 2002 is influenced by urban effect, elevation, and especially latitude.

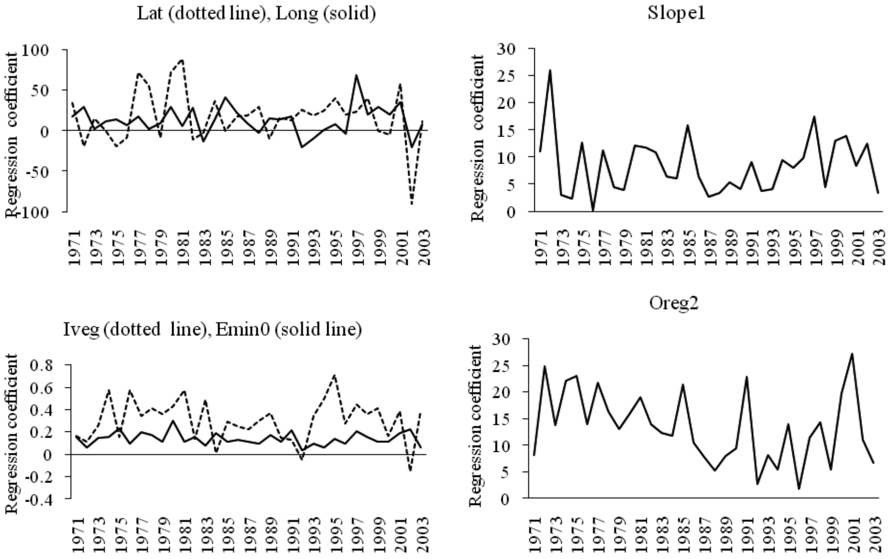


Fig. 11. Time variation of the regression coefficients of the significant predictors for PCA I-based models.

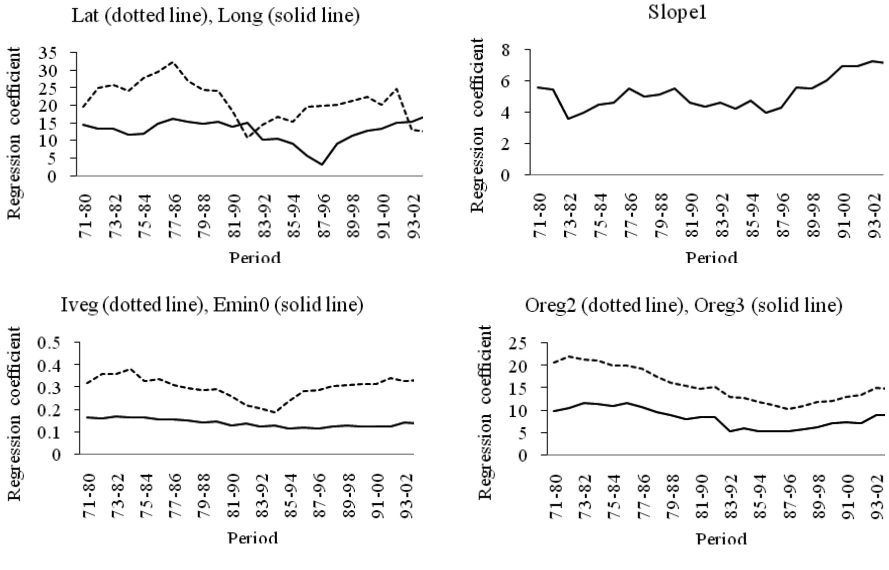


Fig. 12. Time variation of the regression coefficients of the significant predictors for PCA II-based models.

The extremeness of precipitation during the summer of 1997, especially during the two heavy precipitation episodes in July, is associated with the atmospheric advection of moist air stream (*Rezacova et al.*, 2005). Meteorological features were specifically characterized by an intensive influx of moisture into Central Europe and intensive upward motions in the precipitation area. Finally, the regression coefficients of precipitation models reveal that topographic feature and geographic coordinates have stronger influence on spatial distribution of summer precipitation over the Czech Republic than other geographical factors.

4.3. Trends in coefficients of regression of significant predictors

The trends in regression coefficients for the yearly-based models during the period 1971–2003 are displayed in *Fig. 13* (*a,b,c*). It is obvious that the trends in regression coefficients are negative for almost all selected independent variables. The only exception concerns the slope of grids northeastward from the locations, which has positive trend. The magnitude of the trends is higher for latitude and west slope (reaching -0.37 yr^{-1} and -0.23 yr^{-1}) than for other independent variables, especially elevation, vegetation, and longitude. The sign and magnitude of trends are similar for the same predictors independently selected from different model-based approaches. The positive and negative trends detected are insignificant. Therefore, the relationships between the spatial patterns of summer precipitation and geographical variables, during the relatively short period considered in this study, are stable in time.

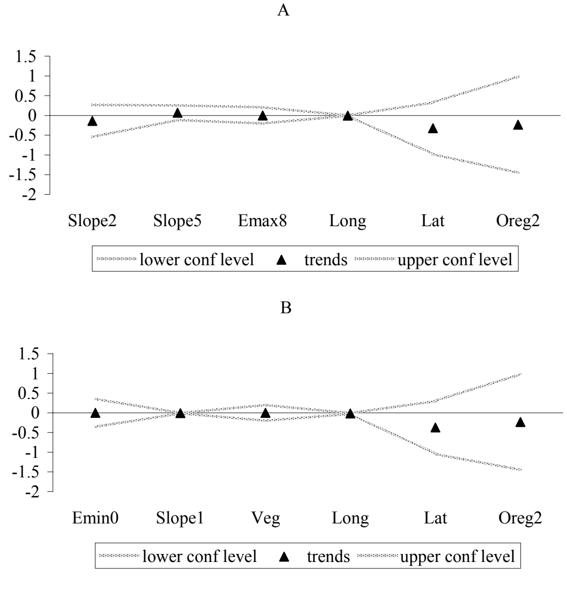


Fig. 13a,b. Trends in coefficient of regression estimated by STW-based (A) and PCA-B (B) precipitation models. Models are generated at the annual time step.

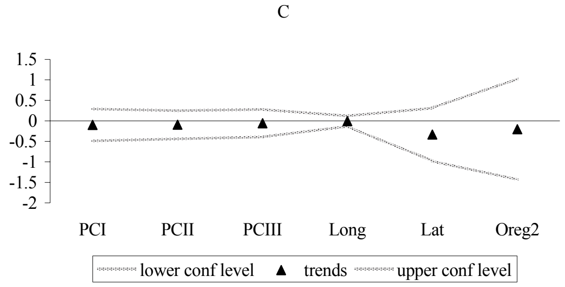


Fig. 13c. Trends in coefficient of regression estimated by PCA A-based precipitation models. Models are generated at the annual time step.

5. Conclusions

Relationships between geographical variables (land cover, geographical positions, morpho-topographical features) and rainfall spatial pattern were analyzed in this study using regression models. Stepwise regression and rotated PCA identified several significant geographical variables: slope2, slope5, emax8, emin0, vegetation, oreg3, oreg2, and geographical coordinates. Geographical coordinates (including continentality) and slope orientation westward are the most significant predictors for modeling the spatial distribution of summer precipitation over the Czech Republic. Except the minimum elevation (Emin0), all important topographical factors correspond to the outside grids at the distance of 1 km from the location. Therefore, the terrain characteristics had more significant influence when a larger area than a station location is taken into account for selecting independent variables. Similarly, the models working with independent variables selected from stations are slightly inaccurate in comparison with models working with independent geographical variables describing a large area around the location. Furthermore, PCA and STW-based approaches for selecting significant geographical variables to model spatial patterns of precipitation over the Czech Republic are consistent. Nevertheless, PCA-based models seem to be less powerful than STW-based models. They yield a small explained variance and a large prediction error. In the same way, models based on a one-year time resolution are notably less powerful than models based on long-term time resolution.

The time variation of explained variance indicated that precipitation variability has an influence on the variance accounted by models and on its performance. Thus, smaller explained variance is accounted by models during dry years. A temporal analysis of regression coefficients from ten-year precipitation models showed positive relationships between precipitation and geographical factors. Spatial patterns of averaged summer precipitation at a decade time resolution are modeled as increasing with continentality, morpho-topographic features, and latitude. On the other side, at the smaller time-resolution, negative relationships were found, especially between precipitation and topographical features, vegetation, and geographical positions. The spatial pattern of precipitation during the summer of 2002, for example, is strongly related to decreasing latitude.

Trend analysis of regression coefficients revealed that relationships between summer precipitation patterns and morpho-topographical features, land cover, and geographical positions are stable in time. No significant trend in the model parameters (effect of geographical factors) has been found during 1971–2003. Model parameters are stable in time. Therefore, spatial prediction of precipitation based on single realization in time (i.e., long term average) is not biased by the length of the sample period.

Spatial precipitation patterns vary in time according to the effect of geographical factors, as well as performance of models. The models failed to capture the relationships between the precipitation patterns and the geographical factors during dry years, namely 1987 and 1988. These years have been dominated by intensive drought (*Kaspar* and *Muller*, 2008). Precipitation patterns during these years could be well modeled using other auxiliary independent variables such as the dominant mode of atmospheric circulation that are linked with both spatial and temporal variability of precipitation. Although this was not the aim of this analysis, the conclusions of this study show the necessity to investigate further on the relationships between precipitation pattern and atmospheric circulation.

References

Baigorria, G.A., Jone, J.W. and *O'Brien, J.J.*, 2007: Understanding rainfall spatial variability in southeast USA at different timescales. *Int. J. Climatol. 27*, 749-760.

Basher, R.E., Zheng, X., 1998: Mapping rainfall fields and their ENSO variation in data-sparce tropical south-west Pacific Ocean region. *Int. J. Climatol. 18*, 237-251.

Brown, D.P., Comrie, A.C., 2002: Spatial modeling of winter temperature and precipitation in Arizona and New Mexico, USA. *Clim. Res. 22*, 115-128.

Courault, D., Monestiez, P., 1999: Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast France. *Int. J. Climatol. 19,* 365-378.

Daly, C., 2006: Guidelines for assessing the suitability of spatial climate data sets. *Int. J. Climatol. 26,* 707-721.

Daly, C., Neilson, R.P., Philip,s D.L., 1994: A statistical-topographic model for mapping climatological precipitation in mountainous terrain. *J. Appl. Meteorol 33,* 140-158.

Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L., Pasteris, P., 2002: A knowledge-based approach to the statistical mapping of climate. *Clim. Res. 22*, 99-113.

Diodatto, N., 2005: The influence of topographic co-variables on the spatial variability of precipitation over small regions of complex terrain. *Int. J. Climatol. 25*, 351-363.

Drogue, G., Humbert, J., Deraisme, J., Mahr, N., Freslon, N., 2002: A statistical – topographic model using an omnidirectional parametrization of the relief for mapping orographic rainfall. *Int. J. Climatol. 22*, 599-613.

Goodale, C.L., Aber, J.D., Ollinger, S.V., 1998: Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model. *Clim. Res. 10*, 35-49.

Hengl, T., 2007: A practical guide to geostatistical mapping of environmental variables. *JRC Scientific and Technical Report* - European Communities, Luxembourg, 143 pp.

Hulme, M., New, M., 1997: Dependence of large-scale precipitation climatologies on temporal and spatial sampling. *J. Climate 10*, 1099-1113.

Johnson, G.L., Hanson, C.L., 1995: Topographic and Atmospheric influences on precipitation variability over a mountainous watershed. *J. Appl. Meteorol. 34,* 68-87.

Joly, D., Nilsen, L., Fury, R., Elvebakk, A., and *Brossard, T.*, 2003: Temperature interpolation at a large scale: test on a small area in Svalbard. *Int. J. Climatol. 23,* 1637–1654.

Kakos, V., 2001: Maximum precipitation in the Czech Republic in terms of synoptic meteorology (in Czech). In *Sbornik přednášek ze semináře výsledkům grantového projektu VaV/510/97,* 46-60.

Kaspar, M., Muller, M., 2008: Selection of historic heavy large-scale rainfall events in the Czech Republic. *Nat. Hazards Earth Syst. Sci. 8,* 1359–1367.

Marquinez, J., Lastra, J., Garcia, P., 2003: Estimation models for precipitation in mountainous regions: the use of GIS and multivariate analysis. *J. Hydrol. 270*, 1-11.

Ninyerola, M., Pons, X., Roure, J.M., 2000: A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques. *Int. J. Climatol. 20*, 1823-1841.

Ninyerola, M., Pons, X., Roure, J.M., 2007: Objective air temperature mapping for Iberian Peninsula using spatial interpolation and GIS. *Int. J. Climatol. 27,* 1231-1242.

Prudhomme, C., Reed, D.C., 1999: Mapping extreme rainfall in mountainous region using geostatistical techniques: a case study in Scotland. *Int. J. Climatol. 19,* 1337-1356.

Rezacova, D., Kaspar, M., Muller, M., Sokol, Z., Kakos, V., Hanslian, D., Pesice, P., 2005: A comparison of the flood precipitation episode in August 2002 with historic extreme precipitation events on the Czech territory. *Atmos. Res. 77*, 354-366.

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

Tolasz, R. et al. (eds.), 2007: *Climate Atlas of Czechia.* Czech Hydrometeorological Institute, Prague, Hardback ISBN 978-80-86690-1.

Vicente-Serrano, S.M., Saz, M.A., Cuadrat, J.M., 2003: Comparative analysis of interpolation methods in the middle Ebro valley (Spain): application to annual precipitation and temperature. *Clim. Res. 24*, 161-180.

Weisse, A.K., Bois, P., 2001: Topographic effects on statistical characteristics of heavy rainfall and mapping in the French Alps. *J. Appl. Meteorol. 40,* 720-740.

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 71–85

Geostatistical modeling of high resolution climate change scenario data

Ladislav Vizi[1], **Tomáš Hlásny**[2,3*], **Aleš Farda**[4], **Petr Štepánek**[4], **Petr Skalák**[4], and **Zuzana Sitková**[2]

[1]*Technical University Košice, Faculty of BERG,*
Park Komenského 19, Košice 04200, Slovakia; E-mail: ladislav.vizi@tuke.sk

[2]*National Forest Centre – Forest Research Institute,*
T. G. Masaryka 22, Zvolen – 96001, Slovakia
E-mail: sitkova@nlcsk.org

[3]*Czech University of Life Sciences, Faculty of Forestry and Environment,*
Kamýcká 1176, Praha 6 – Suchdol 165 21, Czech Republic

[4]*Czech Hydrometeorological Institute,*
Na Šabatce 17, 14306, Prague, Czech Republic
E-mails: ales.farda@chmi.cz; skalak@chmi.cz; stepanek@chmi.cz

[*]*Corresponding author; E-mail: hlasny@nlcsk.org*

(Manuscript received in final form November 22, 2010)

Abstract—Air temperature and precipitation data organized within a 10 × 10 km grid covering the whole of Slovakia were subject to analysis. The source data are produced by the ALADIN Climate/CZ regional climatic model. The output of the global climatic model ARPEGE-Climat (Meteo-France) provided the driving data for the regional model. The IPCC A1B scenario provides the information on the future development of greenhouse gas emissions. Such scenario was developed within the 6th Framework Programme project CECILIA (Central and Eastern Europe Climate Change Impacts and Vulnerability Assessment).

Geostatistical prediction of annual mean air temperature and precipitation data was carried out for the reference (1961–1990) and distant future (2071–2100) climates. The experimental data were non-stationary and significantly correlated with elevation. Therefore, we used non-stationary multivariate geostatistical techniques allowing for the integration of such information. In particular, we used kriging of residuals, universal kriging with external drift, and external drift kriging in the scope of IRF-*k* (intrinsic random functions of order *k*). Prediction based on linear regression of elevation data was used as a complementary technique. Accuracy assessment was based upon the mean square errors produced by cross-validation in case of kriging-based predictions and upon the mean square residual in case of linear regression-based prediction.

We found that all kriging-based techniques outperformed the linear regression-based approach, yielding mean square error lower by 53–75%. External drift kriging in the scope of IRF-*k* produced slightly better results for most of the climate variables analyzed. The poorest results were achieved in the case of annual mean air temperature for the period 1961–1990, where the variogram of residuals was very erratic.

External drift kriging-based techniques were found to be very efficient for interpolating annual mean air temperature and annual precipitation data organized in regular grids. Accuracy assessment indicated that the three predictors used yielded almost identical results for a single variable, while significant differences in mean square error were observed in a between-variables comparison.

Key-words: mean annual air temperature, annual precipitation totals, ALADIN regional climatic model, external drift kriging, non-stationary modeling, Slovakia

1. Introduction

Maps of various climate elements produced by spatial interpolation of point-distributed data are frequently used to improve understanding of climate's spatio-temporal variability as well as for various studies of climate impacts on society and ecosystems (*Haines et al.*, 2006; *Trnka et al.*, 2004; *Hlásny* and *Turčáni*, 2009).

Recent availability of large amounts of climate data produced by global and regional climate models (GCMs / RCMs) has drawn attention to the need for optimizing the spatial interpolation of such data (e.g., *Haylock et al.*, 2008). The data are primarily organized in regular grids with spacing depending on the respective GCM / RCM. However, follow-up studies on agriculture, forestry, air pollution, and other areas often ask for seamless information on climate rather than point-distributed data. Therefore, the search for optimal interpolation techniques is a timely task (*Mulugeta*, 1996; *Dobesch et al.*, 2007). A growing number of recent works comparing interpolation techniques and identifying optimal data- or region-specific methods is testimony to this issue's importance (*Goovaerts*, 2000a; *Haberland*, 2007).

In addition to the frequently used non-model-based techniques (not using a variogram, such as inverse distance weighting or spline interpolation, e.g., *Hancock* and *Hutchinson*, 2006), there exists a range of geostatistical techniques allowing for specific improvements of spatial interpolation, mainly by integrating heterogeneous data (*Isaak* and *Srivastava*, 1989). In this paper, we demonstrate the use of several external drift kriging-based techniques (EDK, hereinafter) (*Matheron*, 1973) for interpolating high resolution climate change scenario data. These techniques allow for flexible integration of point-distributed climatic data with correlated grid-distributed predictor variables, such as elevation and solar insolation.

An early paper on EDK's use for predicting air temperature and precipitation in Scotland was published by *Hudson* and *Wackernagel* (1994). Later, EDK's ability to integrate heterogeneous data prompted many other climatology studies. *Carera-Hernandez* and *Gaskins* (2007) found that the use of elevation as a secondary variable improves the prediction, even if the correlation is low. The influence of such other terrain-related parameters as relief slope and aspect was investigated by *Attore et al.* (2007), who found that universal kriging with external drift performed the best for 17 out of 21 climatic variables analyzed. *Goovaerts* (2000a) tested the efficiency of several approaches to spatial interpolation of rainfall data (linear regression, ordinary cokriging, kriging with external drift, simple kriging with local means) and stressed the benefits of incorporating the elevation data. That author found that the latter two named techniques yield slightly better results than did the others.

The purpose of this paper is to analyze the climatic data produced by the ALADIN-Climate /CZ regional climate model (*Farda et al.*, 2010) for the whole of Slovakia in 10×10 km spatial resolution. Maps of mean annual air temperature (hereinafter just air temperature) and mean annual precipitation totals (hereinafter just precipitation) for the reference (1961–1990) and distant future (2071–2100) climates were to be produced. In particular, we focused on:

(1) describing and preprocessing the data,

(2) using several EDK-based techniques and a linear regression-based approach for spatial prediction of air temperature and precipitation data for the reference and distant future climates, and

(3) assessing the accuracy of the maps produced and discussing the results.

2. Data

The reference and future climate data were originally calculated using the GCM ARPEGE–Climat V4 (*Déqué*, 2007) in an experiment performed by CNRM/-Météo-France. Because of rather coarse resolution of the GCM (~50 km over Central Europe), the RCM ALADIN-Climate /CZ (*Farda et al*., 2010) was used for additional downscaling of the GCM data. The IPCC A1B emission scenario was adopted to provide information on future development of greenhouse gas emissions. The data were developed as part of the CECILIA (Central and Eastern Europe Climate Change Impacts and Vulnerability Assessment, www.cecilia-eu.org) project under the European Union's 6th Framework Programme. The RCM covers Central Europe with a resolution of 10 km. Such resolution allows for better representation of the driving physical processes (e.g., more accurate resolution of geographical features and thus, various interactions with the surface), thus, leading to better description of local climate and positively affecting the quality of the simulations.

The data used in this study comprise a subset of ALADIN's entire integration domain covering the Slovak Republic. The 10 km resolution grid, with rotation 6^{o} azimuth, is extended beyond the country's borderline by approximately one grid point in order to reduce interpolation errors in the edge locations (*Fig. 1*). In total, 644 grid points are used for the analysis. Source data statistics are given in *Table 1*.

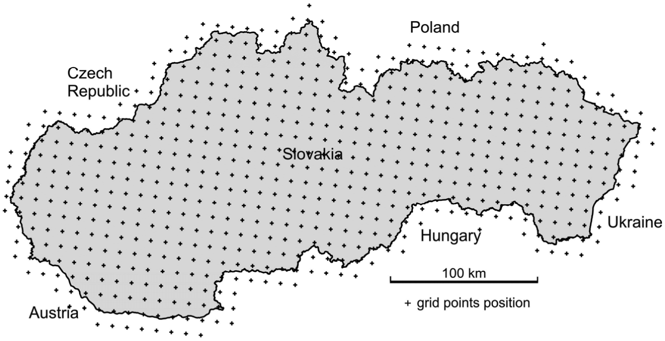


Fig. 1. Spatial arrangement of 10×10 km grid of the ALADIN-Climate/CZ regional climate model in Slovakia.

Table 1. Source data statistics. Abbreviations: N – number of observations, Min – minimum, Max – maximum, Avg – average, Med – median, SD – standard deviation, IQR – inter-quartile range, Skew – skewness, Kurt – kurtosis. Variables: T – mean annual air temperature for the given period, P – mean annual precipitation totals for the given period

Variable	N	Min	Max	Avg	Med	SD	IQR	Skew	Kurt
T 1961–1990	644	0.6	10.7	7.3	7.4	2.0	3.1	−0.32	−0.42
T 2071–2100	644	5.2	13.4	10.7	10.8	1.8	3.2	−0.40	−0.90
P 1961–1990	644	416.5	1206.7	697.3	675.6	145.1	207.8	0.83	0.12
P 2071–2100	644	472.5	1175.0	671.3	629.7	154.5	214.9	0.94	0.05

Elevation of the study area is used as a supportive variable (*Fig. 1*). It is organized in a 180 m resolution grid, which is more than 55 times denser than the ALADIN grid.

3. Methods

The climate data used are clearly non-stationary, as they have a global elevation-controlled trend in the south-north direction. Therefore, we describe here the concepts for multivariate non-stationary geostatistical modeling that are used. All steps of the geostatistical analysis were carried out in the ISATIS v.9 environment (*Geovariances, Centre de Géostatistique in Fontainebleau*). For regression modeling, STATISTICA v.7 (*StatSoft, Inc.,* 2004) was used.

3.1. Stationary spatial models

Stationarity of spatial data, i.e., the presence of a stable mean for an analyzed variable, is the simplest and most frequently documented case of geostatistical analysis. This allows for straightforward modeling of the variogram, which measures the spatial correlation of the studied variable, as well as for an optimal estimation using kriging. In a stationary case, where drift $m(\mathbf{x})$ is a constant, the variogram γ for distance **h** is estimated as:

$$2\gamma_Z(\mathbf{h}) = \mathrm{E}[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})]^2 . \tag{1}$$

For a regionalized variable, as one realization of a random function, the variogram is estimated by forming the average dissimilarities for all $N(\mathbf{h})$ pairs of data $z(\mathbf{x}_\alpha)$ and $z(\mathbf{x}_\alpha + \mathbf{h})$ available at sample points $\mathbf{x}_\alpha$ that are linked by the vector $\mathbf{h} = \mathbf{x}_\alpha - \mathbf{x}_\beta$ (*Hudson* and *Wackernagel*, 1994):

$$2\gamma_Z(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} (z(\mathbf{x}_\alpha) - z(\mathbf{x}_\beta))^2 . \tag{2}$$

Usually, we observe that the average dissimilarities between the couples of values increase when **h** is increased, up to a value of the variable autocorrelation (range of influence). Beyond this value, the dissimilarities become more or less constant around an upper asymptote (sill of the variogram) that is approximately equal to the data variance.

3.2. Non-stationary spatial models

In a non-stationary case, there is a definite trend in the data, being a gradient in a given direction (*Hayet et al.*, 2000). The non-stationary approach to spatial modeling considers the phenomenon under study as a sum of two terms:

$$Z(\mathbf{x}) = Y(\mathbf{x}) + m(\mathbf{x}), \tag{3}$$

where $Y(\mathbf{x})$ describes the local variation of $Z(\mathbf{x})$, and it is assumed to be stationary with constant mean. The term $m(\mathbf{x})$ describes a large-scale variation of $Z(\mathbf{x})$ (drift). It is assumed that the drift can be represented by a polynomial of order L:

$$m(\mathbf{x}) = \sum_{l=0}^{L} a_l f_l(\mathbf{x}), \tag{4}$$

where a_l are unknown coefficients of known functions $f_l(\mathbf{x})$ of the spatial coordinates. Note that for $L=1$, Eq. (4) reduces to a constant term, a_0, which indicates no trend in the spatial coordinates. The term $Y(\mathbf{x})$ in Eq. (3) represents the residual, i.e., the amount of variability remaining after the drift has been removed. The residuals have a stationary covariance (variogram) function between any pairs of random variables $\{Y(\mathbf{x}),Y(\mathbf{x}+\mathbf{h})\}$. The drift is essentially the mean value of the variable as a function of the location at which the variable is measured. In a non-stationary case, we can rewrite Eq. (1) as follows:

$$\begin{aligned}2\gamma_Z(\mathbf{h}) &= \mathrm{E}[Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})]^2 \\ &= \mathrm{E}[Y(\mathbf{x}+\mathbf{h})+m(\mathbf{x}+\mathbf{h})-Y(\mathbf{x})-m(\mathbf{x})]^2 \\ &= \mathrm{E}[Y(\mathbf{x}+\mathbf{h})-Y(\mathbf{x})]^2-[m(\mathbf{x}+\mathbf{h})-m(\mathbf{x})]^2 \\ &= 2\gamma_Y(\mathbf{h})-[m(\mathbf{x}+\mathbf{h})-m(\mathbf{x})]^2,\end{aligned} \tag{5}$$

where the trend values $m(\mathbf{x})$ and $m(\mathbf{x}+\mathbf{h})$ are unknown. The second term on the right side of Eq. (5) provides the drift estimate in a particular direction. The most straightforward approach to non-stationary modeling is based on computation of a residual variogram $2\gamma_Y(\mathbf{h})$. A proper use of this technique is documented by *Dowd* (1984) and *Goovaerts* (2000b), who suggested several ways for coping with certain shortcomings of this technique, as discussed, for example, by *Hayet et al.* (2000).

Another approach to non-stationary modeling used in this paper is the method of increments based on the theory of intrinsic random function of order *k* (IRF-*k*) (*Matheron*, 1973). It defines a linear combination of *Z* data that filters out the drift component $m(\mathbf{x})$. In a stationary case, the first order difference, or increment $[Z(\mathbf{x}+\mathbf{h})-Z(\mathbf{x})]$, filters out the constant drift *m*. In a non-stationary case, higher order differentiation is required to filter out the higher orders of the polynomial drift. This approach leads to a so-called generalized covariance model $K(\mathbf{h})$ instead of a variogram $\gamma(\mathbf{h})$. The most widely used models for generalized covariances are polynomial in form (*Matheron*, 1973):

$$K(\mathbf{h})=a_0+\sum_{k=0}^{K}(-1)^{k+1}a_{k+1}|\mathbf{h}|^{2k+1}. \tag{6}$$

More information on this technique can be found in *Dowd* (1984) or *Chiles* and *Delfiner* (1999).

3.2.1. Case of external drift(s)

In case of a non-stationary spatial model, we consider the trend $m(\mathbf{x})$ of the variable $Z(\mathbf{x})$ to be a function of spatial coordinates. For some applications, exhaustive data for one or more regionalized variables $s_j(\mathbf{x})$ may be available in the studied domain (representing, e.g., elevation). If such data are available, it is worthwhile to use them as additional constraints to the interpolation.

If we assume that $Z(\mathbf{x})$ is on average equal to $s_j(\mathbf{x})$ up to linear way and with coefficients a_0 and b_1, then:

$$\mathrm{E}[Z(\mathbf{x})]=m(\mathbf{x})=a_0+\sum_{j=1}^{J} b_j s_j(\mathbf{x}). \qquad (7)$$

Because variables $s_j(\mathbf{x})$ are exhaustively available, they reflect the average shape of $Z(\mathbf{x})$, where just the scaling is different (*Hudson* and *Wackernagel*, 1994).

3.3. Interpolation techniques

In this study, we include several techniques under the term external drift kriging. Their common feature is that the elevation acts as an external drift correlated with the primary climatic variables. In addition, a linear regression of the elevation data was used to predict the climate data.

Residual kriging, known also as regression kriging (*Odeh et al.*, 1994) or kriging after detrending (*Goovaerts*, 2000b), predicts the residuals at all nodes of the interpolated grid $\mathbf{x}_0$, $Y^*(\mathbf{x}_0)$. Residual kriging uses the drift $m^*(\mathbf{x})$ calculated by a polynomial of a selected degree by the least squares method. Residuals $Y(\mathbf{x}_\alpha)$ are calculated as the differences between $Z(\mathbf{x}_\alpha)$ and $m^*(\mathbf{x}_\alpha)$ at all sample points. Using the variogram of residuals, the kriging system for weights $\omega(\mathbf{x}_\alpha)$, $\alpha=1,\ldots,n$ includes $n+1$ linear equations:

$$\begin{cases} \sum_{\beta=1}^{n} \omega(\mathbf{x}_\beta)\gamma_Y(\mathbf{x}_\alpha,\mathbf{x}_\beta)+\lambda_0=\gamma_Y(\mathbf{x}_0,\mathbf{x}_\alpha) \quad \text{for} \quad \alpha=1,\ldots,n, \\ \sum_{\alpha=1}^{n} \omega(\mathbf{x}_\alpha)=1. \end{cases} \qquad (8)$$

The residual kriging estimator is a linear combination of available *n* data $y(\mathbf{x}_\alpha)$ for only *n* random variables $Z(\mathbf{x}_\alpha)$:

$$y^*(\mathbf{x}_0)=\sum_{\alpha=1}^{n} \omega(\mathbf{x}_\alpha)y(\mathbf{x}_\alpha). \qquad (9)$$

Finally, the estimated drift, Eq. (4) and kriged residuals, Eq. (9) are added together.

Universal kriging provides an unbiased estimation, which considers drift $m(\mathbf{x})$ as a continuous and regular function (Eq. (4)), usually restricted to polynomials up to the order of 2. It uses a model representing both local and global variability of the variable in space. It determines the underlying variogram of $Y(\mathbf{x})$ and estimates the degree of drift. We modeled the drift by Eq. (4), including elevation as the external drift. The simultaneous system of equations for the universal kriging estimator, considering both internal and external drift, is as follows:

$$\begin{cases} \sum\limits_{\beta=1}^{n} \omega(\mathbf{x}_\beta)\gamma_Z(\mathbf{x}_\alpha,\mathbf{x}_\beta)+\lambda_0+\sum\limits_{l=1}^{L}\lambda_l f_l(\mathbf{x}_\alpha)+\sum\limits_{j=1}^{J}\lambda_j s_j(\mathbf{x}_\alpha)=\gamma_Z(\mathbf{x}_0,\mathbf{x}_\alpha) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } \alpha=1,\dots,n, \\ \sum\limits_{\alpha=1}^{n} \omega(\mathbf{x}_\alpha)=1, \\ \sum\limits_{\alpha=1}^{n} \omega(\mathbf{x}_\alpha)s_j(\mathbf{x}_\alpha)=s(\mathbf{x}_0) \qquad \text{for } j=1,\dots,J, \\ \sum\limits_{\alpha=1}^{n} \omega(\mathbf{x}_\alpha)f_l(\mathbf{x}_\alpha)=f_l(\mathbf{x}_0) \qquad \text{for } l=1,\dots,L. \end{cases} \tag{10}$$

The kriging system for IRF-*k* is similar to the universal kriging system, Eq. (10). The only difference is that it uses the generalized covariance model $K(\mathbf{h})$ (Eq. (6)) instead of the variogram $\gamma(\mathbf{h})$ (Eq. (2)). More details about IRF-*k* can be found in *Dowd* (1984) or *Chiles* and *Delfiner* (1999).

3.4. Linear regression-based estimation

The generally recognized relationship between the climate variables addressed and elevation allows for a simple prediction of climate data at all positions for which elevation data are available. There exists a set of collocated climate $z(\mathbf{x}_\alpha)$ and elevation $s(\mathbf{x}_\alpha)$ data $[z(\mathbf{x}_\alpha), s(\mathbf{x}_\alpha)]$; $\alpha=1,\dots,n$, where n is the number of observations. The prediction $z^*(\mathbf{x}_0)$ is based on a linear relationship:

$$z^*(\mathbf{x}_0)=a_0^*+b_1^*s(\mathbf{x}_\alpha), \tag{11}$$

where coefficients a_0^* and b_1^* are estimated from the collocated climate and elevation data. A major shortcoming of this type of prediction is that the climate data at a particular grid node are derived only from the collocated elevation, regardless of the surrounding observed climate data (*Goovaerts*, 2000a).

3.5. Accuracy assessment

Two techniques were used to assess the accuracy of the maps produced and the performance of the predictors used. A cross-validation procedure was used in case of geostatistical predictions (*Isaaks* and *Srivastava*, 1989; *Clark*, 1986). The technique temporarily removes one observation at a time from the data set and "re-estimates" this value from the remaining data using a given predictor. Such procedure produces couples of values, the differences between which yield cross-validation residuals. The main criterion for assessing accuracy is mean square error (MSE), which measures the average squared difference between the observed $z(\mathbf{x}_\alpha)$ and predicted $z^*(\mathbf{x}_\alpha)$ values:

$$\mathrm{MSE} = \frac{1}{n} \sum_{\alpha=1}^{n} [z(\mathbf{x}_\alpha) - z^*(\mathbf{x}_\alpha)]^2 , \qquad (12)$$

where n is the number of observations.

Correlation coefficients of observed versus predicted values, normality of residuals distribution, mean value of residuals (criterion that the mean is approaching zero), and degree of randomness of spatial distribution of residuals can also be used.

Another approach was used in the case of linear regression-based prediction. The MSE was computed as the average square residual value for the linear model fitted using all observations:

$$\mathrm{MSE} = \frac{1}{n} \sum_{\alpha=1}^{n} [z(\mathbf{x}_\alpha) - (a_0^* + b_1^* s(\mathbf{x}_\alpha))]^2 . \qquad (13)$$

4. Results

4.1. Drift identification

To identify an optimal global trend, the polynomials of order one (linear) and two (quadratic) plus one external drift (elevation) were tested by the cross-validation procedure for the lowest mean square error. Other criteria, such as mean of residuals approaching zero, minimal variance, normal distribution, and well-structured directional experimental variograms, were used as well. We found that the linear drift along the x and y coordinates (internal drift), together with the elevation (external drift),

$$m^*(\mathbf{x}) = a_0 + a_1 \mathrm{x} + a_2 \mathrm{y} + b_1 s(\mathbf{x}) , \qquad (14)$$

performed the best for all climate variables.

4.2. Kriging-based predictions

To perform the residuals kriging, we used the trend functions described above to filter out the residuals $y(\mathbf{x}_\alpha)$ from the regionalized variable $z(\mathbf{x}_\alpha)$, then estimated the residual variogram models $\gamma_Y(\mathbf{h})$ for all the climate variables analyzed (*Fig. 2*). Estimation of the variogram model for the variable T 1961–1990 was problematic, because there were erratic directional experimental variograms without clear spatial structure. Therefore, an omnidirectional model was fitted to the experimental variogram values in this case. The variogram's origin was estimated from the directional variogram constructed in the azimuth 6^{o} that rises from the variogram's value at about 0.1 ($^{\mathrm{o}}$C)^2. Directional experimental variograms are presented in *Fig. 2* to demonstrate that no anisotropy can be modeled in this case.

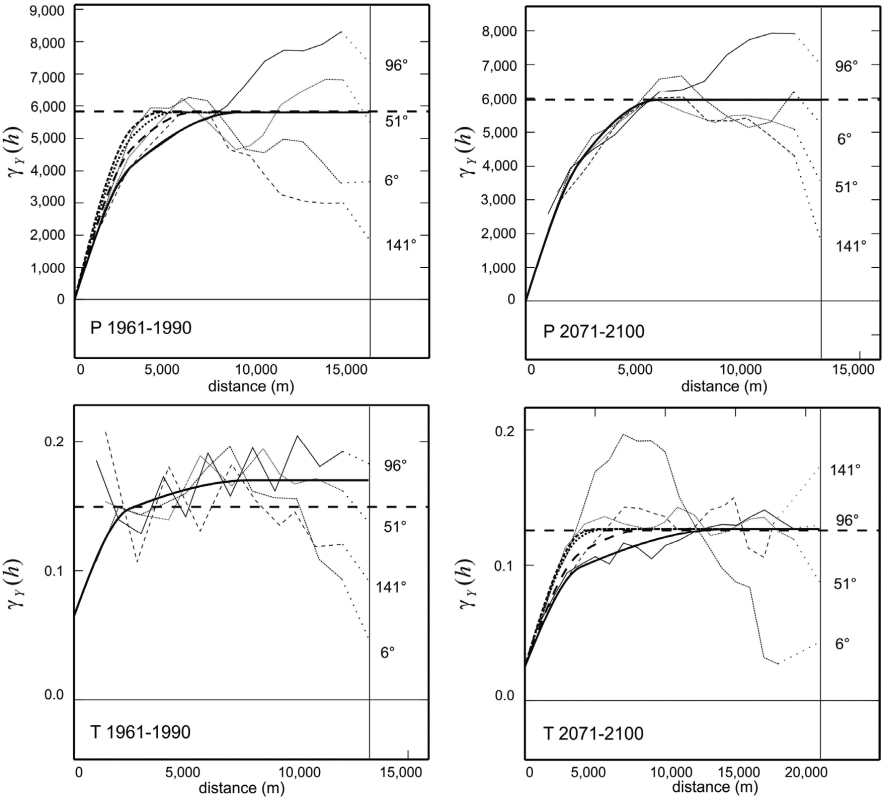


Fig. 2. Directional experimental residual variograms (thin lines) and respective variogram models (thick lines). The numbers on the right side indicate the angles at which the variograms were calculated. Abbreviations: P 1961–1990 – mean precipitation totals during the period 1961–1990, P 2071–2100 – mean precipitation totals during the period 2071–2100, T 1961–1990 – mean annual air temperature during the period 1961–1990, T 2071–2100 – mean annual air temperature during the period 2071–2100.

The estimation of kriging weights $\omega(\mathbf{x}_\alpha)$ was based on Eq. (8). Estimation of residuals $y^*(\mathbf{x}_0)$ was based on the linear combination of available data according to Eq. (9). Finally, the kriged residuals were summed with the trend model according to Eq. (3).

In case of universal kriging, the trend component is directly included into the kriging system according to Eq. (10) for the drift estimation (Eq. (14)). The final estimation was performed directly using the raw variable $Z(\mathbf{x})$.

In case of IRF-*k*, the automatic fitting procedure of the ISATIS environment was used to determine both the degree *k* of the drift and the generalized covariance. For all variables, the degree of the drift was 1 (linear in X and Y directions) plus the external drift represented by the elevation. The generalized covariance of order 1 (similar to the linear model of the variogram) without nugget effect was used for all climate variables.

For interpolation neighborhood definition (*Isaaks* and *Srivastava*, 1989), we used a so-called unique neighborhood, i.e., all available data were used to estimate a value at a particular grid node. We also tested several designs for a moving neighborhood, such as a first ring neighborhood (4 adjacent samples), second ring neighborhood (16 adjacent samples), and third ring neighborhood (36 adjacent samples). The cross-validation tests indicated that the unique neighborhood was performing the best for all climate variables. In addition, the use of moving neighborhoods resulted in "radial" artefacts in the maps produced, due to the resolution of the estimated grid which is more than 55 times higher than that of the ALADIN grid.

The maps of both variables for both time slices produced by EDK in the scope of IRF-*k* can be seen in *Fig. 3*. We can see that the elevation pattern is much stronger in the case of temperature than in that of precipitation data due to the different correlation of climate variables with elevation (*Table 2*).

4.3. Linear regression-based prediction

Linear regression-based prediction was used to provide the reference value for assessing the accuracy of the kriging-based techniques. Regression parameters from elevation and the respective climatic variables are based on all 644 observed values (*Table 2*). Mean square error was calculated using Eq. (13).

4.4. Accuracy assessment

Accuracy assessment was based on comparison of the MSE yielded by kriging-based predictions (*Table 3*) with that from the linear regression-based prediction (*Table 2*) (*Goovaerts*, 2000a). The latter technique provided the MSE reference value for evaluating the performance of kriging techniques. Proportional values of MSE are illustrated in *Fig. 4*. Such an approach allows for evaluating the performance of respective predictors for a single variable as well as for between-variable comparison. The results are discussed below.

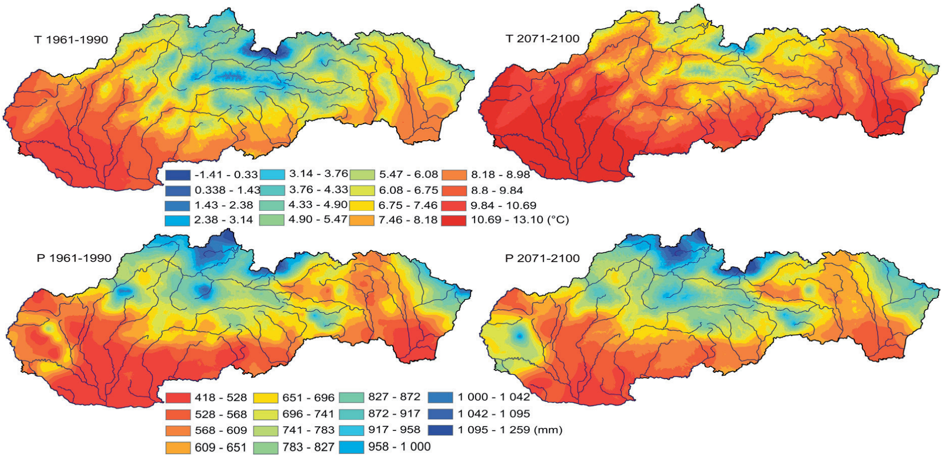


Fig. 3. Maps of mean annual air temperature and mean annual precipitation totals for the reference (1961–1990) and distant future (2071–2100) climates produced using external drift kriging in the scope of IRF-*k*.

Table 2. Linear regression parameters between elevation (X) and respective climatic variables (Y). Abbreviations: a_0^* – intercept, b_1^* – slope, R – correlation coeficient, R^2 – coefficient of determination, MSE – mean square error

X	**Y**	a_0^*	b_1^*	**R**	$\mathbf{R^2}$	**MSE**
Elevation	P 1961–1990	526.86	0.39180	0.75	0.563	9141
Elevation	P 2071–2100	488.27	0.42060	0.76	0.578	10123
Elevation	T 1961–1990	10.26	–0.00688	–0.95	0.903	0.394
Elevation	T 2071–2100	13.34	–0.00606	–0.94	0.887	0.359

Table 3. Results of the cross-validation based accuracy assessment. Abbreviations: KR – residuals kriging, UK – universal kriging with external drift, IRF-*k* – external drift kriging in the scope of IRF-*k*, R – correlation coefficient between observed and predicted values, R^2 – coefficient of determination, MSE – mean square error of prediction, AVG – average value of residuals

Variable	**Interpolator**	**R**	$\mathbf{R^2}$	**MSE**	**AVG**
T 1961–1990	KR	0.978	0.956	0.17457	0.000286
T 1961–1990	UK	0.978	0.957	0.17165	0.000286
T 1961–1990	IRF-*k*	0.976	0.953	0.184628	0.000410
T 2071–2100	KR	0.990	0.981	0.056014	–0.000449
T 2071–2100	UK	0.991	0.981	0.055269	–0.000327
T 2071–2100	IRF-*k*	0.990	0.981	0.054161	–0.000306
P 1961–1990	KR	0.954	0.909	1588.9	–0.1220
P 1961–1990	UK	0.954	0.910	1564.6	–0.0280
P 1961–1990	IRF-*k*	0.955	0.912	1536.9	–0.0099
P 2071–2100	KR	0.939	0.881	2329.5	–0.0180
P 2071–2100	UK	0.940	0.884	2273.9	0.0470
P 2071–2100	IRF-*k*	0.943	0.889	2163.9	0.0660

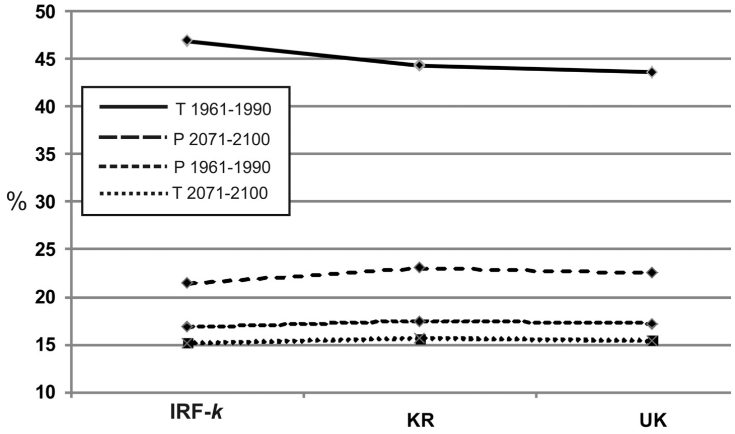


Fig. 4. Results of accuracy assessment for predictions produced by three interpolation techniques for mean annual air temperature and mean annual precipitation totals for the periods 1961–1990 and 2071–2100. The figures indicate the proportions of the MSE (mean square error) yield by cross-validation in the case of EDK-based techniques to MSE yield using the linear regression approach.

5. Conclusions and discussion

We performed a series of analyses of high resolution RCM data covering Slovakia. Both air temperature and precipitation data are well correlated with elevation (*Table 2*), and thus, we focused on the integration of that variable into the interpolation. Such supportive variable is presumed to reduce the amount of uncertainty in the maps produced. We used three external drift kriging-based techniques: residuals kriging, universal kriging with external drift, and external drift kriging in the scope of IRF-*k*. We described in details the particular steps of the geostatistical analysis to allow for a deeper understanding of those techniques used.

All kriging-based techniques produced comparable results for a single climate variable. The reason for this evidently lies in the high correlation of climate data with elevation, which covers the impact of different interpolation algorithms. In the cases of the variables T 2071–2100, P 1961–1990, and P 2071–2100, EDK in the scope of IRF-*k* yields slightly better results than do the remaining kriging-based techniques. This can reflect the benefit of using an automatized procedure in generalized covariance calculation for regularly distributed data in comparison to manual variogram fitting (the case of residuals kriging). The poorest results were reached in the case of the variable T 1961–1990, where the residuals were very erratic, and thus, they influenced the shape of the respective variograms (*Fig. 2*). This applies also for the remaining techniques, because drift parameters remain more or less stable.

Subsequently, we tested the ratio of mean square errors produced by kriging-based techniques to those from linear regression-based estimation. All kriging techniques significantly outperformed the linear regression-based estimation, which yields a mean square error 15 – 47% higher (depending on the variable). This means that, despite high correlation between climate data and elevation, information about the configuration of the surrounding data significantly improved the estimation. The accuracy assessment indicated that the three predictors used yielded almost identical results for a single variable, while significant differences in mean square error were observed by between-variables comparison.

Geostatistical techniques, in general, require a certain extent of user intervention and cannot be fully automatized. In any case, large amounts of climate data produced by various instruments require at least a semi-automatized approach when producing series of climate maps for various time slices. External drift kriging in the scope of IRF-*k* is a candidate technique for this. It yielded slightly better results than did the remaining EDK-based techniques for three out of four variables analyzed, and the underlying generalized covariance may be calculated automatically (see implementation in the ISATIS environment used in this paper). By contrast, residuals kriging requires a series of user interventions, which were not, however, compensated by improved accuracy of the prediction.

Acknowledgements —This research was conducted as part of the 6th Framework Programme project CECILIA (Central and Eastern Europe Climate Change Impacts and Vulnerability Assessment), and within projects of the Ministry of Agriculture of the Czech Republic, No. QH91097 "Analysis of the climate change impacts on the distribution and voltinism of *Ips typographus* in spruce forests of the Czech Republic as underlying information for their sustainable management", and the Scientific Grant Agency of the Ministry of Education of the Slovak Republic, VEGA 1/0222/08 "Application of geostatistical tools for multivariate analysis and data integration of regionalized variables".

References

Attore, F., Alfo, M., De Sanctis, M., Francesconi, F., Bruno, F., 2007: Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale. *Int. J. Climatol. 27*, 1825-1843.

Carrea-Hernandez, J.J., Gaskin, S.J., 2007: Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. *J. Hydrol. 336*, 231-249.

Chiles, J.P., Delfiner, R., 1999: *Geostatistics – Modeling Spatial Uncertainty*. Oxford University Press.

Clark, I., 1986: The art of cross validation in geostatistical applications. In *19*th *Application of Computers and Operational Research in the Mineral Industry* (ed.: *V. Ramani*), 211-220.

Déqué, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. *Global Planet. Change 57*, 16-26.

Dobesch, H., Dumolard, P., Dyras, I. (eds.), 2007: *Spatial Interpolation for Climate Data: The Use of GIS in Climatology and Meteorology. Geographical Information Systems Series*. Wiley-ISTE, New York.

Dowd, P.A., 1984: *MINE5260: Non-Stationarity. MSc notes in Mineral Resources and Environmental Geostatistics*. Dpt. of Mining and Mineral Engineering, University of Leeds, United Kingdom.

Farda A., Déqué, M., Somot S., Horányi A., Spiridonov V., Tóth. H., 2010: Model ALADIN as a Regional Climate Model for Central and Eastern Europe. *Studia Geophysica et Geodaetica, Journal of the Czech Academy of Sciences 54*, 313-332.

Geovariances: ISATIS v.9, Centre de Géostatistique in Fontainebleau.

Goovaerts, P., 2000a: Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *J. Hydrol. 228*, 113-129.

Goovaerts, P., 2000b: Using elevation to aid the geostatistical mapping of rainfall erosivity. *Catena 34*, 227-242.

Haberland, U., 2007: Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. *J. Hydrol. 332*, 144-157.

Haines, A., Kovats, R. S., Campbell-Lendrum, D., Corvalan, C., 2006: Climate change and human health: impacts, vulnerability, and mitigation. *Public Health 367*, 2101-9.

Hancock, P.A., Hutchinson, M.F., 2006: Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. *Environ.Modell. Softw. 21*, 1684-1694.

Hayet, Ch., Galli, A., Ravenne, C., Tesson, M., de Marsily, G., 2000: Estimating the Depth of Stratigraphic Units from Marine Seismic Profiles Using Nonstationary Geostatistics. *Nat. Resour. Res. 9*, 77-95.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res. 113*, D20119.

Hlásny, T., Turčáni, M., 2009: Insect pests as climate change driven disturbances in forest ecosystems. In Bioclimatoloy and Natural Hazards (eds.: *K. Střelcová, C. Mátyás, A. Kleidon, M. Lapin, F. Matejka, M. Blaženec, J. Škvarenina, J. Holécy*). Springer, Netherlands, 165–178.

Hudson, G., Wackernagel, H., 1994: Mapping temperature using kriging with external drift: Theory and example from Scotland. *Int. J. Climatol. 14*, 77-91.

Isaaks, H. E., Srivastava, R. M., 1989: *Introduction to Applied Geostatistics*. Oxford University Press, New York.

Matheron, G., 1973: The intrinsic random function and their applications. *Adv. Appl. Probab. 5*, 439-469.

Mulugeta, G., 1996: Manual and automated interpolation of climatic and geomorphic statistical surfaces: An evaluation. *Annals of the Association of American Geographers 86*, 324-342.

Odeh, I., McBratney, A., Chittleborough, D., 1994: Spatial prediction of soil properties from land-form attributes derived from a digital elevation model. *Geoderma 63*, 197-214.

StatSoft, Inc., (2004). STATISTICA Cz (Software system for statistical analyses), Version 7. www.StatSoft.Cz

Trnka, M., Dubrovský, M., Semerádova, D., Žalud, Z., 2004: Projections of uncertainties in climate change scenarios into expected winter wheat yields. *Theor. Appl. Climatol. 77*, 229-249.

IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 87–98

Interpolation techniques used for data quality control and calculation of technical series: an example of a Central European daily time series

Petr Štěpánek[1*], **Pavel Zahradníček**[1], and **Radan Huth**[2]

[1]*Czech Hydrometeorological Institute,*
Regional office Brno, Kroftova 43, 61667, Brno, Czech Republic

[2]*Institute of Atmospheric Physics, Academy of Sciences of the Czech Republic,*
Boční II 1401, 14131 Praha 4, Czech Republic

[*]*Corresponding author; E-mail: petr.stepanek@chmi.cz*

(Manuscript received in final form August 27, 2010)

Abstract—For various studies, it is necessary to work with a sufficiently long series of daily data that is processed in the same way for the whole area. National meteorological services have their own tools for data quality control; data are usually available non-homogenized (with respect to artificial changes in the series due to relocations, change of observers, etc.). In the case of areas across borders of individual countries, researchers from both sides of a frontier can obtain quite different results depending upon the data they use. This was one of the reasons for processing stations from the area along borders of four countries in the Central European region within the international CECILIA project (Central and Eastern Europe Climate Change Impact and Vulnerability Assessment, project of EC No. 037005). For the processing of the series, quality control has been carried out, gaps have been filled and, in the end, a series at a new position (grid points of RCM output) were calculated. An interpolation technique which is able to deal with all these tasks is described in this work and then applied to a series of various meteorological elements in Central Europe.

Key-words: data quality control, filling missing values, interpolation techniques, climatological time series

1. Introduction

During validation of regional climate model (RCM) outputs, its values are compared with the values of observations. Whereas the observations are located in the station network, which is irregular in its nature, the dynamical model (GCM, RCM) outputs are provided on a regular grid (statistical downscaling procedures can yield output either at stations or grid points, depending on what they were trained on). Dynamical models thus provide area-aggregated, rather than point-specific data, which makes a direct comparison between station data and gridded model output less straightforward, especially for variables with a short correlation distance, such as precipitation (e.g., *Skelly* and *Henderson-Sellers*, 1996). Therefore, validation has the potential to be truer to dynamical models if the observations are transformed from stations to a grid. This was one of the reasons that such a task was carried out within the CECILIA project.

For the development and calibration of statistical downscaling methods, and for the use of outputs from dynamical as well as statistical downscaling in climate change impact studies, a common observed dataset needed to be created. It was decided that the common dataset would extend over the area along the boundaries of the Czech Republic, Austria, Slovakia, and Hungary (this region is hereinafter called the CECILIA Central European domain). The main intention was to cover the majority of the impact target areas in Central Europe. Another deciding factor in this decision was that it would be easier to obtain meteorological data from meteorological services for only relatively small parts of the countries than for their large parts or even whole countries.

To achieve such a goal, it was necessary to prepare observation data in a way that they would be homogeneous, free of erroneous values, and they gaps would be filled. Ideally, they should also be available in the location of the used model output. For this reason, two versions of the dataset were created, one located at the stations, the other located on the grid of the regional climate model, in this case ALADIN-Climate/CZ (details about the model can be found, e.g., in *Farda et al.*, 2007). To create series at given locations, interpolation methods, which are described further in this paper, have been used. The techniques for data quality control, carried out upon the data prior to any further processing, and for filling the missing values in the station series are, in principal, identical to that used for the calculation of series at a new position, mentioned above. For this reason, the quality control is described in this paper as well.

2. Central European dataset, data preparation

The area of interest covered by the dataset can be seen in *Figs. 1* and *2*. It includes:

- in the Czech Republic: the southern and southeastern part, consisting of the regions of České Budějovice, the Highlands (Vysočina), South Moravia, Zlín, and minor southern parts of Central Bohemia;
- in Austria: the federal states of Lower Austria, Upper Austria, Vienna, and Burgenland;
- in Slovakia: the western part, consisting of the regions of Bratislava, Trnava, Nitra, Trenčín, and Banská Bystrica;

- in Hungary: the regions of Győr-Moson-Sopron and Komárom-Esztergom.

The Central European area covers the following impact target areas (processed in the CECILIA project): agriculture – Lower Austria (AT), southern Moravia (CZ), the Danube lowlands (SK), and the northwestern part of Hungary (HU); forestry – southern central Slovakia (SK); hydrology – the Dyje and upper Vltava catchments (CZ), the Hron catchment (SK).

The dataset itself consists of daily data for the period of 1961–2000. Variables available in the dataset are given in *Table 1*. Potential evapotranspiration is not included, since there are several ways it can be calculated and it can also be derived from the available elements by individual users.

Table 1. Meteorological elements available in the common dataset

Abbreviation	**Description**	**Unit**
TMI	Maximum temperature	°C
TMA	Minimum temperature	°C
H	Relative humidity	%
SRA	Precipitation	mm
SSV	Sunshine duration	h

The following comments on the variables selected and not selected should be made:

- Daily mean temperature was not included because of regional differences in its calculation and a change in the practice of its calculation in Austria in the early 1970s, which could induce an inhomogeneity in the time series and inconsistency along the state boundaries.
- Relative humidity, and not another measure of atmospheric moisture unaffected by daily temperature cycle, such as specific humidity, was selected, because some of the impact models require only relative humidity as their input.
- Wind speed and direction were not subjected to gridding and the creation of technical series because of the necessity of working separately with the two wind components, which would cause considerable complications, making the resultant technical series doubtful and unreliable.
- Solar radiation can easily be approximated from the sunshine duration data. Solar radiation was not included among the final products, since meteorological services apply different approaches for its calculation (e.g., the Angström formula or regression models based on altitudes).

Even incomplete time series were allowed into the database. The data were prepared and provided by the following partners: the Czech Hydrometeorological Institute (CHMI) for the Czech Republic, the Forest Research Institute (NFC) for Slovakia (40 stations), the University of Natural Resources and

Applied Life Sciences (BOKU) for Austria (30 stations), and the Hungarian Meteorological Service (OMSZ) for Hungary. The data policy of some of the involved meteorological services does not allow the distribution of raw station data. This was another reason for creating technical series from the station data available, which were distributed among the project participants. Technical series of two kinds were constructed: (i) gridded datasets covering the area where station data are available; this was regarded as a primary dataset; (ii) station technical series, which have the advantage of better homogeneity and completeness over the raw data.

In the CECILIA Central European domain, about 150 climatological stations are available – see *Fig. 1*, in comparison with 832 grid points of the ALADIN-CLIMATE/CZ RCM – see *Fig. 2*. The number of stations available in the individual countries and meteorological elements are given in *Table 2*.

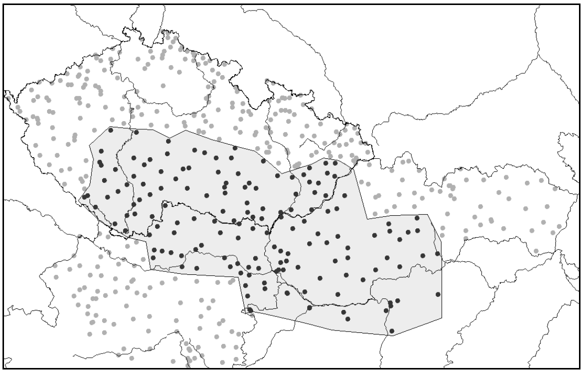


Fig. 1. CECILIA Central European domain (shaded area) with available climatological stations (dark / light dots for stations inside / outside the domain).

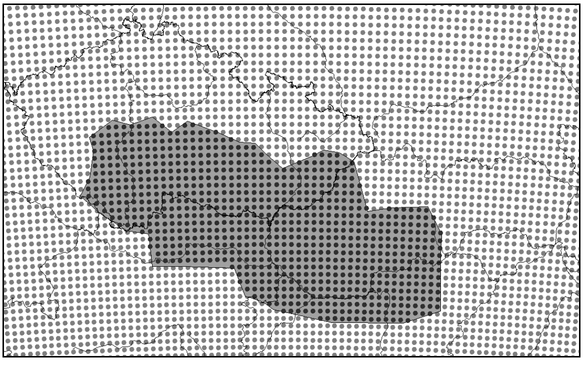


Fig. 2. Grid points of ALADIN-CLIMATE/CZ (dark / light dots) available within / outside the CECILIA Central European domain (shaded area).

Table 2. Number of stations, available per individual country (AT – Austria, CZ – Czech Republic, SK – Slovakia, HU – Hungary) and meteorological element (see Table 1 for explanatory notes)

Country	**Element**				
	TMA	**TMI**	**SRA**	**SSV**	**H**
AT	33	33	35	11	30
CZ	90	90	90	68	91
SK	39	39	39	39	40
HU	11	11	11	6	11
Total	173	173	175	124	172

3. Data quality control

Before the station technical series and gridded dataset were calculated, raw station data had been subjected to thorough quality control using AnClim and ProClimDB softwares (*Štěpánek*, 2007; more details can be found in the documentation of the softwares at www.climahom.eu). Tools available in the softwares were designed so that they could be used for the automated finding of errors in datasets. The outliers were found by a combination of several methods: the percentage of neighbor stations which are significantly ($p = 0.05$) different from the base station (found from standardized differences between neighbors and base station, the limit value is more than 75%); the difference of the base station value and the median calculated from values of neighbors standardized to the base station altitude (using linear regression) divided by standard deviation of the base station, expressed as CDF of normal distribution (the limit value is more than 0.95); the coefficient (multiple) of distance of the base station value above (below) the upper (lower) quartile calculated from the standardized (to the base station altitude) values of neighbor stations (the higher the value, the more similar neighbor values are compared to the base station value, the limit value is a coefficient higher than 5); the difference from the expected value (details on its calculation are given in Section 4); and the median calculated from the original values of neighbor stations divided by the standard deviation of the base station values (expressed as CDF of normal distribution, the value should be low, otherwise it indicates that the calculation of the expected value is probably wrong, the limit value is less than 0.75). The calculation was carried out for each meteorological element and individual day separately (*Štěpánek et al*., 2009).

Table 3 shows an example of the suspicious values found. Such values were found in all the available raw datasets (Austria, Czech Republic, Slovakia, and Hungary, their numbers are given in *Table 4*) and were withdrawn from further processing, replaced with a code for missing value.

Table 3. Output from the ProClimDB software with an example of suspicious values found in the raw dataset (gray column) compared to values of five neighbor stations (five rightmost columns)

Element	Station				Suspected	Expected	Remark	Neighboring stations				
	ID	Year	Month	Day	value	value		9900	13301	9811	15900	16000
TMIN	10000				492.0		Altitude	648.0	480.0	695.0	810.0	842.0
TMIN							Distance	22.0	43.1	50.1	56.9	62.7
TMIN	**10000**	**1961**	**3**	**18**	**8.0**	**–1.8**		**–2.9**	**–1.7**	**–1.5**	**–1.8**	**–2.0**
TMIN	**10000**	**1962**	**4**	**22**	**10.0**	**2.9**		**1.1**	**3.2**	**3.8**	**3.1**	**4.0**
TMIN	**10000**	**1962**	**4**	**23**	**13.0**	**0.9**		**0.1**	**1.3**	**1.8**	**0.6**	**2.8**
TMIN	**10000**	**1962**	**5**	**22**	**7.0**	**1.1**		**1.3**	**0.8**	**2.9**	**0.7**	**1.4**
TMIN	**10000**	**1962**	**7**	**21**	**13.0**	**8.4**		**7.4**	**8.6**	**9.1**	**8.5**	**9.0**
TMIN	**10000**	**1963**	**5**	**30**	**10.6**	**3.3**		**3.1**	**3.3**	**4.1**	**2.7**	**3.2**
TMIN	**10000**	**1964**	**1**	**5**	**–10.0**	**–18.5**		**–19.7**	**–18.4**	**–16.5**	**–16.4**	**–17.0**
TMIN	**10000**	**1968**	**4**	**15**	**5.0**	**–0.6**		**–1.3**	**–0.5**	**0.6**	**–1.4**	**–1.4**
TMIN	**10000**	**1975**	**4**	**6**	**9.4**	**4.0**			**4.2**	**2.1**	**2.1**	**2.2**
TMIN	**10000**	**1976**	**2**	**8**	**–1.2**	**–8.9**			**–9.0**	**–7.9**	**–6.9**	**–8.3**

Table 4. Numbers of suspicious values (evident errors) per country and meteorological element (see Table 1 for explanatory notes)

	Absolute numbers						**Relatively per number of stations**				
Country	**Element**					**Country**	**Element**				
	TMA	**TMI**	**SRA**	**SSV**	**H**		**TMA**	**TMI**	**SRA**	**SSV**	**H**
AT	28	74	195	309	118	AT	0.85	2.24	5.57	28.09	3.93
CZ	36	157	489	910	498	CZ	0.40	1.74	5.43	13.38	5.47
SK	8	37	72	975	346	SK	0.21	0.95	1.85	25.00	8.65
HU	1	10	33	374	201	HU	0.09	0.91	3.00	62.33	18.27
Total	**73**	**278**	**789**	**2568**	**1163**	**Total**	**0.42**	**1.61**	**4.51**	**20.71**	**6.76**

The data quality checked datasets were further used in the calculation of the station technical series and the gridded dataset.

4. Calculation of station technical series and gridded dataset

Several methods can be used to calculate the values of a given meteorological element at a certain geographical position (e.g., at a grid point). Inverse distance weighting is among the more simple methods, but it still gives good results, even when compared to modern geostatistical methods such as kriging, co-kriging, and universal kriging (*Kliegrová et al.*, 2007). As weights, inverse distance or correlation may be used (*Isaaks* and *Srivastava*, 1989), possibly powered to account for lower or higher spatial correlations of a given meteorological element. Applying geostatistical methods to time series is not an easy task (mainly due to the computational demands), but some attempts that combine time and spatial analysis already exist (e.g., *Szentimrey*, 2002; *Květoň* and *Tolasz*, 2003), and such methods have recently begun to be more widely used.

As mentioned above, daily series of several meteorological elements for hundreds of locations (grid points) were to be calculated. Utilizing a GIS environment for a task such as this would be advantageous, because it provides the potential for choosing from a variety of interpolation methods. Nonetheless, current GIS environments (e.g., ArcMap, ESRI ArcView, ArcGIS) are not designed for the easy retrieval of information for time series (calculation for each time step). This is why we needed to create our own tool with enough automation to carry out the task. The software ProClimDB (*Štěpánek*, 2007) was extended for the computation. This software is freely available.

After quality control (see the previous section), the technical series of daily values at a particular grid point (station location) were calculated from up to 6 neighboring (nearest) stations within a distance of 300 km, with an allowed maximum difference in altitude of 500 m. Before applying inverse distance weighting, data at the neighbor stations were standardized relatively to the altitude of the base grid point (station location). The standardization was carried out by means of linear regression and dependence of values of a particular meteorological element on altitude for each day, individually and regionally. Each standardized value was checked to ascertain it did not differ excessively from the original value (providing CDF did not exceed 0.99; in such a case, linear regression was not regarded a good model and an original, i.e., not standardized value, was used for further calculation). In the case of precipitation, neighbors with original values equal to zero were not standardized. For the weighted average (using inverse distances as weights), the power of weights equal to 1 (all meteorological elements except precipitation) and 3 (precipitation) were applied. In the case of temperatures, standardized neighbor values outside the 20% to 80% percentile range were not considered in the calculation of final values (i.e., trimmed mean was applied).

Originally, the "raw" station data (but with suspicious values removed), i.e., series with gaps and also series not available in the whole period of 1961–2000, were used for the calculation of technical series at both stations and grid points. Even if the statistical properties of the original measured data were preserved (like moments) in calculated technical series (calculated for each day separately), some of the time series showed inhomogeneities, which could be resulted from either the inhomogeneity of the original station data or from the method of calculation: if some stations measured only for a short time, the selection of neighbors varies in time. To avoid inhomogeneities of this kind, we proceeded as follows: first, missing values were filled in original station data series; second, for station series with filled gaps, station technical series were calculated, applying standardization of neighbors to base station altitude (estimated using linear regression for the neighboring region, for each month individually), thus, all stations were extended to have values in the whole period of 1961–2000; third, only these equally long station technical series were used for the calculation at grid points.

The altitudes applied in the calculation of grid point series were the actual altitudes, read from a 1 km resolution model of the terrain. However, for the purposes of RCM validation, it would be better to read altitudes of a smoothed terrain (e.g., low-pass filter smoothing for a square of 20×20 km or 10×10 km) to characterize the vicinity of a grid point, much the same as in RCMs. The same is valid for the power of weights (inverse distances). Applying the power of about 0.5 (square root) better characterizes a wider vicinity of a grid point. The goal was, however, to create technical series at a station or grid point and to preserve the statistical characteristic of the particular point. Thus, it is reasonable to say that the calculated series provide point-specific data rather than area-aggregated data. Another reason is that the area of aggregation varies among different climate models (model resolution). The technical series should be used for validation of RCMs with caution.

The settings of parameters of the technical series calculation differ among individual meteorological elements. The next section describes the best solution for each meteorological element with an example of selected stations in the Czech Republic.

5. The best settings in the calculation of station technical series and gridded datasets

The parameter settings for station technical series and the gridded dataset differ for various meteorological elements. The "ideal" setting of parameters was determined by using four base stations in the area of the Czech Republic. Because stations were chosen so they would represent different climatological conditions, both lowland and highland stations were chosen, as well as stations both at the eastern and western edge of the area so as to capture differences between the more maritime and continental weather regimes which manifest across the Czech Republic. The four selected base stations, with their neighbor stations, are displayed in *Fig. 3*, the information on the base stations is provided in *Table 5*. The parameters were tuned by comparing original and calculated values using various verification criteria.

Altogether, 11 various parameters were tested in ProClimDB individually to find the "ideal" setting for all the required meteorological elements: maximum and minimum temperature, relative humidity, precipitation, and sunshine duration. Daily values of the meteorological elements in the period of 1991–2007 were used. The changed (controlled) parameters were: transformation of input values (log, square root, etc.); standardization of neighbor station values to monthly averages (and/or standard deviations) at a base station, standardization of neighbor stations to the altitude of the base station (this case can also be controlled by calculating regression for the whole period – monthly, or for each time step individually, to set the behavior in the case of only one station being present in a given time, and the correction coefficient for regression to control the dependence on altitude); a check whether standardized values become outliers or not; the power of weights for calculation of a new (”expected”) value; applying trimmed mean when a new value is calculated (and setting the limits in such a case).

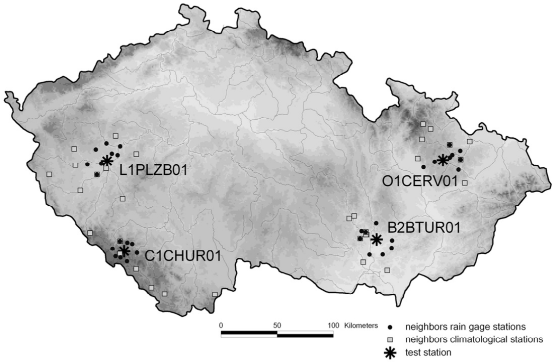

L1PLZB01
O1CERV01
B2BTUR01
C1CHUR01
0
50
100 Kilometers
● neighbors rain gage stations
□ neighbors climatological stations
✱ test station

Fig. 3. Four base stations (marked with an asterisk) and their neighbors (different for precipitation and climatological stations shown in black and grey, respectively) used for the verification of calculated technical series.

Table 5. Base stations used for the verification of calculated technical series

Name	**ID**	**Latitude**	**Longitude**	**Altitude**
Brno-Tuřany	B2BTUR01	49.1597	16.6956	241.00
Plzeň-Bolevec	L1PLZB01	49.7892	13.3867	328.00
Červená	O1CERV01	49.7772	17.5419	750.00
Churáňov	C1CHUR01	49.0683	13.6131	1118.00

It was more difficult to find a solution for precipitation and relative humidity than for the other meteorological elements. Unfortunately, it seems impossible to get 100% realistic values during the calculation (e.g., non-negative relative humidity and precipitation). The unrealistic values are caused mainly by poor quality of raw station data, insufficient length of series at neighbor stations (time gaps simultaneous at several neighbor stations diminish the number of values used for regression), and a greater difference in altitudes of stations used in the regression model. These factors can be controlled to some extent. The input data were controlled for quality before calculation (see previous section). Stations allowed for the calculation can be filtered to retain only those with a certain minimum length and without longer time gaps. The third factor – the difference in altitudes – is not easy to cope with, since we selected the nearest neighbours for the calculation, which, e.g. in the case of precipitation, seems to be the only solution (the selection of the nearest and best correlated stations is the same, while for temperatures, one could also select neighbor stations according to correlations). Problems were especially evident with the mountain station (Churáňov), since its altitude is higher than that of its neighbors and, thus, extrapolation instead of interpolation must be used.

The setting of parameters for maximum temperature, minimum temperature, relative humidity, and sunshine duration is similar to some extent. For station technical series, the neighbor station values were standardized to the base station average and standard deviation using the whole period, within each month individually (in this case we fill gaps in station measurements and this helps to avoid the introduction of inhomogeneities into the series), whereas for the gridded dataset, values were standardized to the altitude of the base station using linear regression estimated for each day individually (which is a better solution, e.g., in case of days with inversions). During the calculation, checks were done to determine that standardized values do not differ too much from the original values. For a value larger than 0.99 (CDF), the original values were used for further calculations: lower settings of 0.95 or 0.90 lead to much worse results. The power for weights (inverse distance) was taken as 1. For maximum and minimum temperature, trimmed mean was applied for calculations of the "expected" value with quantile limits of 20% and 80%. An example of the difference between the original and calculated values of the maximum temperature is shown in *Fig. 4*. It is evident, that stations in lower altitudes show a weak annual cycle of RMSE (root mean square error applied on the calculated and original values). On the contrary, the mountain station of Churáňov reaches very high values of RMSE during winter; the different behavior can be explained by the frequent occurrence of temperature inversions when the lowland stations used for the calculation have substantially different weather conditions.

For the calculation of the technical series of precipitation, a standardization to altitude for the whole period (station technical series), or applied individually for each day (gridded dataset) was again carried out. The difference from previous settings is that the power for weight was set to 3 to reflect lower spatial correlations of precipitation, and a trimmed mean is not applied. No transformation of input values (e.g., logarithms) was performed, since it gave poorer results. The average difference (bias between original and calculated values) for precipitation at Brno-Tuřany is 0.0 mm; in most months it does not exceed 0.1 mm. The highest difference occurs for June, 0.27 mm. RMSE values are highest for summer months as well. Precipitation is influenced by local effects much more than the other meteorological elements, and even at adjacent sites, there can be great differences (in some cases, a 30 to 60 mm precipitation amount is observed at two neighbor stations, while the other two stations record no precipitation at all). For this reason, the correlation coefficient is lower, only 0.875. From the scatter plot (*Fig. 5*, left) we can see several outliers which influence the value of the correlation coefficient. Looking at the histogram (*Fig. 5*, right), we can see that 62% of values differ only negligibly.

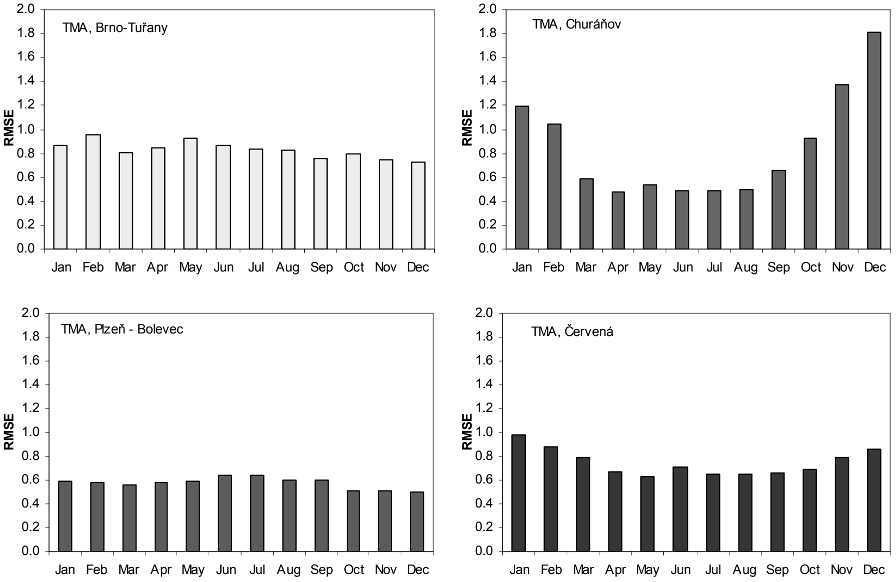


Fig. 4. RMSE (in °C) for four base (tested) stations and maximum temperature.

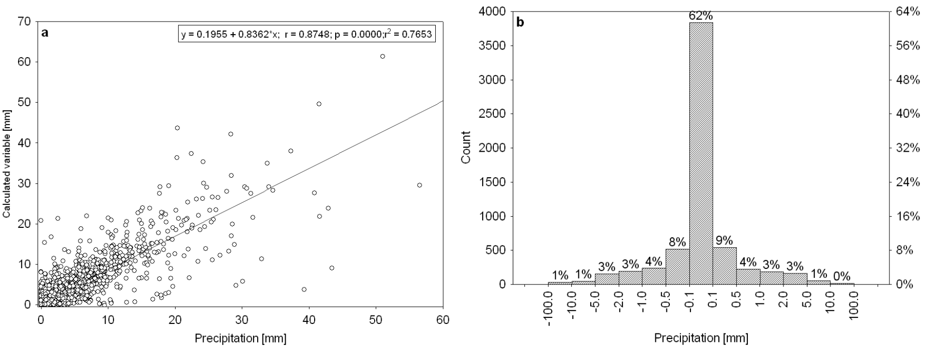


Fig. 5. Scatter plot for calculated and original values of precipitation (left) and histogram of differences between the calculated and original values (right) at station Brno-Tuřany.

More detailed information on the optimal settings found and used in the ProClimDB software is contained within the ProClimDB software documentation, which can be downloaded together with the software itself.

6. Summary

Interpolation techniques can solve many tasks required during data processing. In this work we have shown their application to daily data for various meteorological elements. The technique described is quite general, so that it can be applied to different tasks, such as data quality control (finding suspicious values), filling gaps in the series, or calculation of a new series for a new location. As it can be seen from the given examples of verification results, the calculated station technical series and gridded datasets do very well at reflecting the behavior of the measured values of the processed meteorological elements (maximum and minimum temperature, relative humidity, precipitation, sunshine duration), which make the series capable of being utilized for various purposes, such as a development and calibration of various methods of statistical downscaling, usage in impact studies (since the final network density is much higher than that of the original station network and is, moreover, regular), for a comparison with national datasets (border discrepancies), where available, etc.

Acknowledgements — The present work was supported by the European Commission under Contract 037005 (CECILIA) and the Czech Science Foundation under project 205/08/1619. The following persons are acknowledged for assistance in obtaining the original data: *Tomáš Hlásny*, Forest Research Institute, Zvolen, Slovakia; *Herbert Formeyer*, University of Environmental Sciences, Vienna, Austria, and *Monika Lakatos*, Hungarian Meteorological Service, Budapest, Hungary.

References

Farda, A., Štěpánek, P., Halenka, T., Skalák, P., Belda, M., 2007: Model ALADIN in climate mode forced with ERA40 reanalysis (coarse resolution experiment). *Meteorologický časopis 10,* 123–130.

Isaaks E., Srivastava R., 1989: *An Introduction to Applied Geostatistics.* Oxford University Press, New York, 561 pp.

Kliegrová, S., Dubrovský, M., Metelka, L., 2007: Interpolation methods of weather generator parameters. Program & Abstracts. *10th International Meeting on Statistical Climatology*, Beijing, China, August 20–24, 2007, 122–123.

Květoň, V., Tolasz, R., 2003: Spatial *Analysis of Daily and Hourly Precipitation Amounts with Respect to Terrain*. http://www.map.meteoswiss.ch/icam2003/468.pdf

Skelly, W.C., Henderson-Sellers, A., 1996: Grid box or grid point: What type of data do GCMs deliver to climate impacts researchers? *Int. J. Climatol. 16,* 1079–1086.

Štěpánek, P., 2007: *ProClimDB – software for processing climatological datasets*. CHMI, regional office Brno. http://www.climahom.eu/ProcData.html.

Štěpánek, P., Zahradníček, P., Skalák, P., 2009: Data quality control and homogenization of the air temperature and precipitation series in the Czech Republic in the period 1961–2007. *Adv. Sci. Res. 3,* 23–26.

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

IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 99 –109

Application of gridded daily data series for calculation of extreme temperature and precipitation indices in Hungary

Mónika Lakatos[*], **Tamás Szentimrey,** and **Zita Bihari**

Hungarian Meteorological Service,
P.O. Box 38, H-1525 Budapest, Hungary
E-mails: szentimrey.t@met.hu; bihari.z@met.hu

[*]*Corresponding author; E-mail: lakatos.m@met.hu*

(Manuscript received in final form March 8, 2011)

Abstract—The calculation of extreme climate indices defined by several international projects requires homogeneous time series. To this effect, long term daily extreme temperatures and daily precipitation sums were homogenized, quality controlled, and further processed by the method MASH (Multiple Analysis of Series for Homogenization). After homogenization of station observation series, a gridding procedure was performed on the daily observations by the method MISH (Meteorological Interpolation based on Surface Homogenized Data Basis). The idea behind the MISH interpolation scheme stems from the following principles: gridded data can be created (interpolated) at higher quality with respect to certain climate statistical parameters; and these parameters can be modeled by using the long climate data series. In the MISH procedure, the modeling of the statistical parameters for a given location is based on the long term homogenized monthly data of neighboring stations.

In this paper, we present the computations of extreme temperature and precipitation climate indices for the period of 1901–2009 using datasets which were processed by the above homogenization and gridding algorithms. The obtained trends of several extreme indices as well as their spatial distributions are demonstrated on graphs and maps.

Key words: extreme climate indices, data homogenization, data interpolation, climate of Hungary

1. Introduction

The study of climate extremes in a changing climate has come to the fore in recent years. The most common way to detect changes is the analysis of extreme climate indices, which are defined by several international projects (*Alexander et al.*, 2006). Groups, such as the World Meteorological Organization (WMO)

CCI/CLIVAR Expert team on Climate Change Detection and Indices (ETCCDI), the European Climate Assessment (ECA), and the Asia-Pacific Network (APN) have aimed to provide a framework for defining and analyzing the observed climate extremes. Climate index calculations require quality controlled, homogeneous time series, and the analysis of the results requires a consistent approach (*Wijngaard et al.*, 2003). The global and also the regional studies have focused primarily on the analysis of long term daily temperature and precipitation data (*Frich et al.*, 2002; *Haylock et al.*, 2008), as these climate variables are the most widely available ones. The majority of long data series is inhomogeneous, and often contains shifts in the mean or in the variance due to non-climatic factors, such as site-relocations, changes in instrumentation or in observing practices. Inhomogeneities can distort the true climatic signal, homogeneity testing is important for climate change studies (*van Engelen et al.*, 2008). Amongst the observation series there are good quality data as well, but sorting them out requires the execution of a homogenization procedure first (*Aguilar et al.*, 2003). Neglecting the inhomogeneous series causes a huge loss of valuable information.

Studying the spatio-temporal changes of extremes can be implemented through the analysis of observations reliable in time and space. The spatial interpolation of extreme indices is a difficult task as the distribution functions of the several derived values are unknown. However, the basic variables, such as temperature and precipitation can be gridded by the knowledge of their statistical properties, thus, higher quality gridded datasets can be constructed for further analysis. The main steps of creating the homogenized, gridded dataset for computation of extreme indices are presented in this paper. The changes of such indices for Hungary from the mid-20th century to present are illustrated and shortly analyzed on graphs and trend maps.

2. *Data and methods*

2.1. Homogenization

The computations implemented in this work are based on long term daily data in the period of 1901–2009. Daily maximum and minimum temperatures of 15 observation stations and daily precipitation sum of 58 precipitation stations were taken into account in the analysis. In the preparation phase, homogenization and quality control of the daily measurements were carried out. The homogenization of data was performed with the procedure MASH (*Szentimrey*, 1999). All the MASH options except the metadata information were used in this paper.

2.2. The main features of MASHv3.02

The MASHv3.02 *(Szentimrey,* 2007) software consists of two parts.

Part 1: Quality control, missing data completion, and homogenization of monthly series:

- 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 (e.g., temperature) or multiplicative (e.g., precipitation) model can be used depending on the climate elements.
- Providing the homogeneity of the seasonal and annual series as well.
- Metadata (probable dates of break points) can be used automatically.
- Homogenization and quality control (QC) results can be evaluated on the basis of verification tables generated automatically during the procedure.

Part 2: Homogenization of daily series:

- Based on the detected monthly inhomogeneities.
- Including quality control (QC) and missing data completion for daily data. The quality control results can be evaluated by test tables generated automatically during the procedure.

The importance of homogenization is demonstrated in *Fig. 1* which show the annual number of frost days (daily minimum is below zero) for Szeged station using original and the homogenized daily minimum temperatures. Both the magnitude and the sign of the estimated linear trend are different in the two cases.

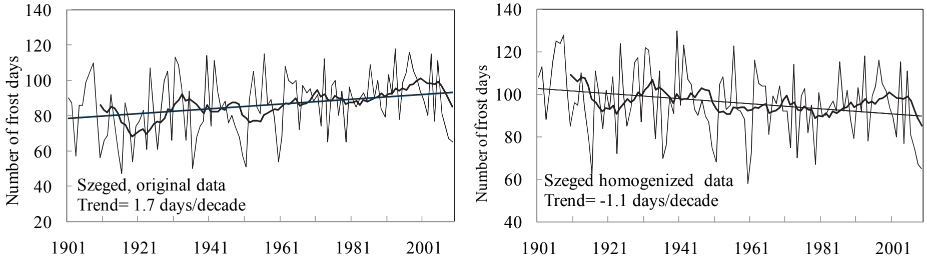


Fig. 1. Annual number of frost days for Szeged station with the fitted linear trend as well as the 10-year moving average in the period of 1901–2009 using the original (left) and homogenized (right) data.

2.3. Gridding

To obtain the high quality, good resolution dataset, a gridding procedure was performed on the homogenized daily series. According to the representativity examinations in the interpolation section, which were performed during this work, the expected interpolation errors may be accepted with using the predictor network of 15 temperatures and 58 precipitation station data series. The MISH interpolation method is a proper choice for this purpose. The MISH procedure was developed at the Hungarian Meteorological Service especially for interpolation of meteorological data (*Szentimrey* and *Bihari*, 2007a). It is based on the principles that the gridded data can be derived (interpolated) at a higher quality if we know certain climate statistical parameters. For example, in the case of normal distribution the means and the covariance structure unambiguously determine the optimal interpolation formula. Long climate data series allow modeling of these statistical parameters. Thus, the modeling for a given location is based on the statistical features of the long term homogenized monthly data of neighboring stations.

2.4. The main features of MISHv1.02

The software MISHv1.02 (*Szentimrey* and *Bihari*, 2007b) consists of two units, the modeling and the interpolation systems. The interpolation system can be operated on the output of the modeling system. The attributes of the MISHv1.02 software can be summarized as follows:

Modeling system for climate statistical parameters in space:

- Based on long homogenized data series and supplementary deterministic model variables, e.g., topography.
- Cross-validation test for interpolation error or representativity.
- Modeling 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.
- The interpolation error or representativity is modeled too.
- Capability for application of supplementary background information (stochastic variables), e.g., satellite, radar, forecast data.
- Capability for gridding of data series.

Gridding system:

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

Contrary to geostatistical methods, the values of variograms must be modeled for each interpolating processes (*Szentimrey et al.*, 2007). One of the most important advantages of the MISH is that the modeling part must be executed only once before the gridding of the data on different timescales, such as daily, monthly, seasonal, or other. Additionally, different station networks can be used in the modeling and in the gridding parts. The modeling part of the MISH procedure is executed on a relative dense, 0.5’×0.5’ resolution grid.

In order to calculate extreme indices, the MISH gridding part was performed on homogenized daily observations for a 0.1°×0.1° grid. The implementation of the MISH gridding procedure resulted in a high quality, homogenized, gridded daily maximum and minimum temperature and daily precipitation datasets with a ~10 km spatial resolution (1104 grid points) in the period of 1901–2009 for Hungary.

3. Climate indices calculations on the gridded dataset

The extreme indices used in this study are based on the CECILIA (Central and Eastern Europe Climate Change Impact and Vulnerability Assessment) project definitions. In the framework of CECILIA project, numerous indices were defined (74 temperature and 55 precipitation indices) on different time scales, i.e., yearly, seasonal and monthly (*Hirschi*, 2008). All of them were implemented for Hungary on homogenized data for observation stations and also for gridded datasets for the whole examined long period. A few selections of the CECILIA extreme indices are presented in this work on yearly scale (*Table 1*).

Table 1. Extreme indices used in this study

Index	**Unit**
Summer days: Tmax > 25 °C	%
Hot days: Tmax ≥ 30 °C	%
Frost days: Tmin < 0 °C	%
Warm nights: Tmin > 20 °C	%
Number of wet days: daily precipitation > 1 mm	days
Percentage of days > 20 mm precipitation	%
Greatest 1-day total rainfall	mm
Greatest 5-day total rainfall	mm
Simple daily intensity: precipitation sum/number of wet days	mm/day
Consecutive dry days: maximum number of consecutive days when the daily precipitation < 1 mm	days

The computational techniques used in the course of index calculations can lead to differences in the results. To obtain comparable results for larger regions, we have to make sure to use the same definition and algorithm. It is particularly important in the case of indices based on percentiles (*Alexander* and *Arblaster*, 2009).

With the help of homogenization, gridding, and extreme index calculation procedures, a high quality, good resolution dataset of the long-term series of indices can be generated and stored. These index datasets can form the basis of further examinations, such as trend estimation and mapping of changes.

4. Graphs and maps based on homogenized gridded data

The course of several temperature and precipitation extreme indices, from the beginning of the 20th century can be followed up in *Figs. 2–4*. Grid point averages represent the countrywide average. The increasing warm temperature extremes coincide with the warming tendencies in the region (*van Engelen*, 2008). The percentage of hot days and that of the warm nights have intensely increased since the early eighties. The presence of more warm nights is also obvious from 1901. The greatest 5-day total rainfall and the days with above 20 mm precipitation show a slight increasing in the last intense warming from eighties. The simple daily intensity index indicates that the rate of the intense rainfall events has increased in summer. The length of the longest dry spell became shorter recently, but considering the whole period, some increase is apparent.

The IPCC Fourth Assessment Report (*IPCC*, 2007) established the features of recent trends of extreme weather events from the late 20th century, in some cases typically after 1960. The trend maps in *Figs. 5–10*, which illustrate the changes of some extreme indices in Hungary, cover the time period 1961–2009 to allow the comparability with other well-known international studies like IPCC. The estimated grid point changes are depicted by linear trend fitting on the corresponding maps. The fitted linear trends were tested on station data and grid point series data as well. In extensive regions of the country, the number of frost days decreased (*Fig. 5*). White areas in *Fig. 5* represent the regions where the changes are not significant at 0.1 probability level. The obvious warming trend is indicated in the percentage of summer days (*Fig. 6*). Beside the point estimation of the slope, confidence intervals were constructed to the estimated trend at different significance levels. *Fig. 7* consists of two maps, according to the bounds of the 0.1 significance confidence interval. The lower bound illustrates the minimum change and the upper bound signifies maximum change occurred in the examined period. Maps of *Figs. 8–10* show the spatial trend of some extreme precipitation climate indices. The number of wet days decreased in Hungary, except for a small region of the country in the northeast (*Fig. 8*). The change in the greatest 1-day total rainfall varies from –15 mm to +10 mm. Regions with growing 1-day precipitation lie mainly to the East from the Danube. The daily precipitation intensity increased in summer. It means that the proportion of the heavy precipitation events in the total rainfall increases over most areas in Hungary

(*Fig. 10*). Regarding the past 50 years the precipitation changes were not significant in extensive regions of the country, according to the applied hypothesis testing.

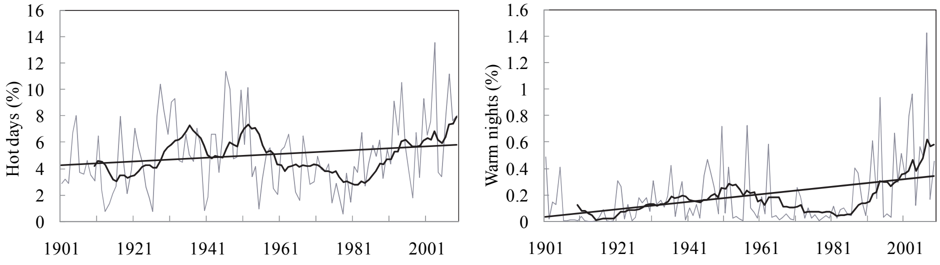


Fig. 2. Grid point average of the yearly percentage of hot days (left) and warm nights (right) with the fitted linear trend as well as the 10-year moving average in the period of 1901–2009 for Hungary.

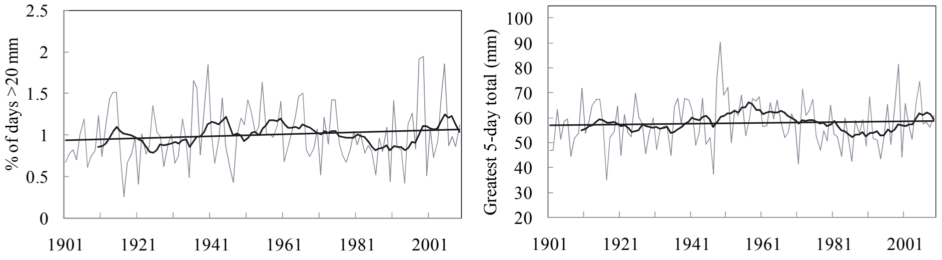


Fig. 3. Grid point average of the yearly percentage of days with above 20 mm (left) and the greatest 5-day precipitation (right) with the fitted linear trend as well as the 10-year moving average in the period of 1901–2009 for Hungary.

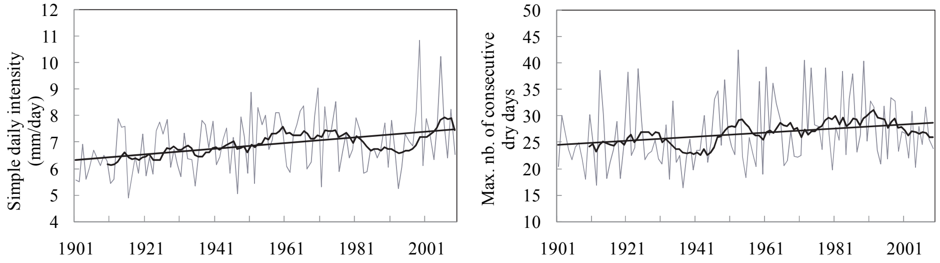


Fig. 4. Grid point average of the daily precipitation intensity index in summer (left) and the maximum number of consecutive dry days (right) with the fitted linear trend as well as the 10-year moving average in the period of 1901–2009 for Hungary.

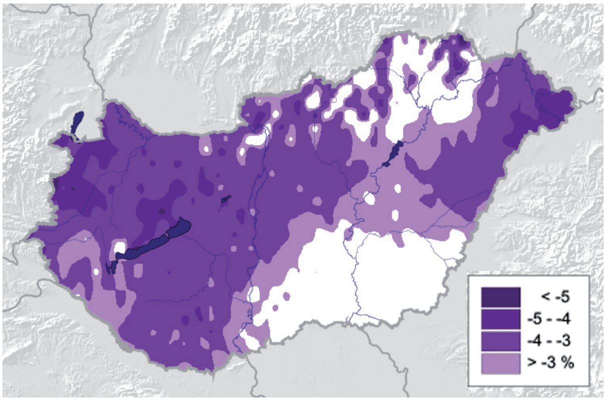


Fig. 5. Change (%) in the number of frost days in the period of 1961–2009. White color indicates no significant change on 0.1 confidence level.

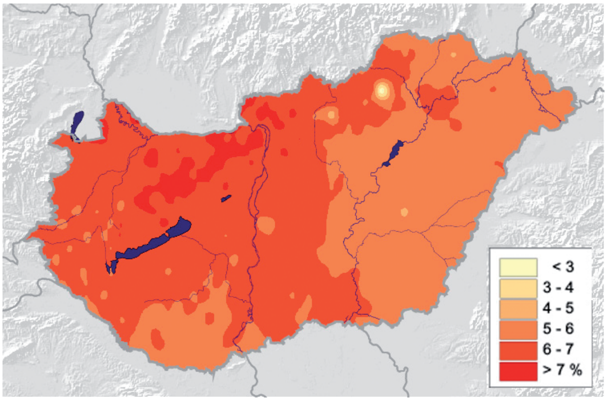


Fig. 6. Change (%) in the number of summer days in the period of 1961–2009.

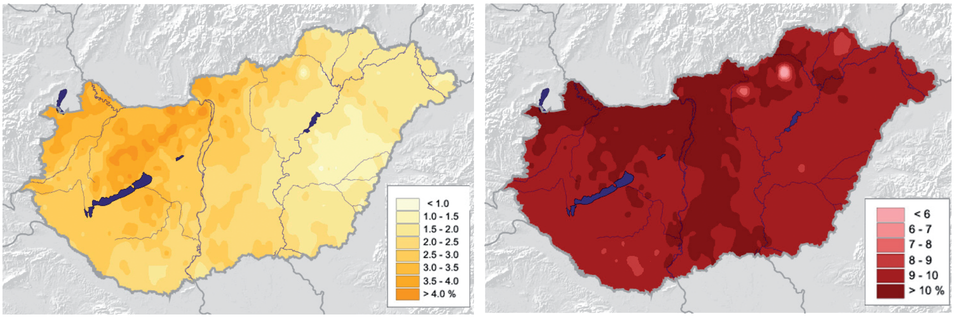


Fig. 7. Change (%) in the minimal (left) and maximal (right) number of summer days according to the 0.1 confidence interval bounds in the period of 1961–2009.

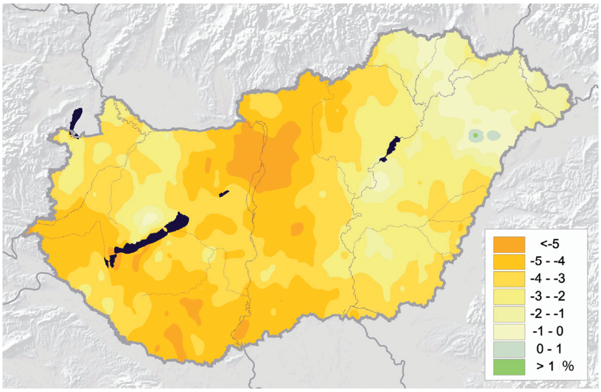


Fig. 8. Change (%) in the number of wet days in the period of 1961–2009.

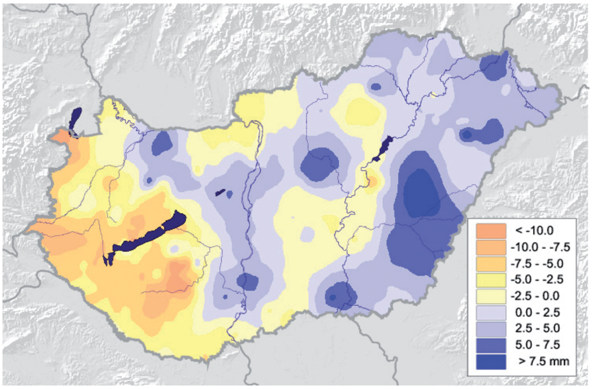


Fig. 9. Change (mm) in the greatest 1-day total rainfall in the period of 1961–2009.

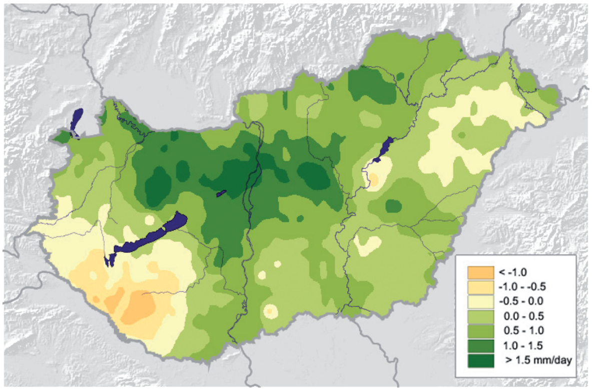


Fig. 10. Change (mm/day) in the summer simple daily precipitation intensity in the period of 1961–2009.

5. *Conclusions*

The preparation of a high quality, homogenized, and gridded daily datasets was presented in this study. Long term daily temperature extremes and precipitation data were quality controlled, homogenized, and gridded in the period of 1901–2009, in order to analyze the extreme climate indices. Instead of the interpolation of extreme indices, the gridding of the basic variables (daily maximum and minimum temperatures and daily precipitation) is recommended, as the probability distributions of the indices are unknown. Time series of the grid point averages for a few selected indices are demonstrated from 1901. The spatial distribution of changes from the mid-20th century is illustrated on trend maps.

The gridded dataset introduced in this work is updated by homogenization and interpolation on the beginning of the new calendar year regularly to serve as an 'as long as possible' time series for climate change studies. The WMO statement on the status of the global climate in 2009 (*WMO*, 2010) underlined that peer reviewed scientific methods for quality control, homogenization, and interpolation to constitute high-quality global climate datasets should be used in the examinations. The created datasets could be relevant contribution to the expected high quality global system of datasets.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.,* 2003*:* Guidelines on climate metadata and homogenization. In *WCDMP* No. 53 – *WMO/TD*-No. 1186.

Alexander, L.V. and *Arblaster, J.M.,* 2009: Assessing trends in observed and modeled climate extremes over Australia in relation to future projections. *Int. J. Climatol. 29*, 417-435.

Alexander, L.V., Zhang, X., Peterson, T.C., Caesar, J., Gleason, B., Klein Tank, A.M.G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Ambenje, P., Rupa Kumar, K., Revadekar, J., and *Griffiths, G.,* 2006*:* Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res. 111*, doi:10.1029/2005JD006290.

Frich, P., Alexander, L.V., Della-Marta, P., Gleason, B., Haylock, M., Klein Tank, A.M.G., and *Peterson, T.,* 2002: Observed coherent changes in climatic extremes during 2nd half of the 20th century. *Climate Research 19*, 193-212.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., and *New, M.,* 2008: European daily high-resolution gridded dataset of surface temperature and precipitation for 1950-2006. *J. Geophys. Res. 113*, D20119, doi:10.1029/2008JD010201.

Hirschi, M., Stepanek, P., Seneviratne, S.I., Christensen, O.B., and *CECILIA WP4 members,* 2008: CECILIA climate indices: Analysis of temperature extremes in Central and Eastern Europe. In *General Assembly of the European Geosciences Union (EGU),* EGU2008-A-09640.

IPCC, 2007: Climate Change 2007. The Scientific Basis, Contribution of Working Group Ito the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). *Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor* and *H.L. Miller* (eds.), Cambridge University Press, UK, 946 pp.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). In *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data,* WMO, WCDMP-No. 41, 27-46.

Szentimrey, T., 2007: Manual of homogenization software MASHv3.02, Országos Meteorológiai Szolgálat (Hungarian Meteorological Service), Budapest, p. 61.

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

Szentimrey, T. and *Bihari, Z.*, 2007b: Manual of interpolation software MISHv1.02. Országos Meteorológiai Szolgálat (Hungarian Meteorological Service), Budapest, p. 32.

Szentimrey, T., Bihari, Z., Szalai, S., 2007: Comparison of geostatistical and meteorological interpolation methods (what is what?). In *Spatial Interpolation for Climate Data – The Use of GIS in Climatology and Meteorology* (eds.: *Hartwig Dobesch, Pierre Dumolard* and *Izabela Dyras*). ISTE Ltd., London, UK, 284 pp, ISBN 978-1-905209-70-5, pp. 45-56.

van Engelen, A., Klein Tank, A., van der Schrier, G., and *Klok, L.,* 2008: European Climate Assessment & Dataset (ECA&D) Report 2008. In *Towards an Operational System for Assessing Observed Changes in Climate Extremes*. KNMI, De Bilt, The Netherlands.

Wijngaard, J.B., Klein Tank, A.M.G., Konnen, G.P., 2003: Homogeneity of 20th century european daily temperature and precipitation series. *Int. J. Climatol. 23*, 679-692.

WMO, 2010: WMO statement on the status of the global climate in 2009. WMO-No. 1055.

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 115, No. 1–2, January – June 2011, pp. 111–120

Spatial differentiation of the climatic water balance in Poland

Agnieszka Wypych[1] and **Zbigniew Ustrnul**[1,2]

[1]*Department of Climatology, Jagiellonian University,*
7 Gronostajowa Str., 30-387 Krakow, Poland; E-mail: agnieszka.wypych@uj.edu.pl
[2]*Institute of Meteorology and Water Management,*
14 Borowego Str., 30-215 Krakow, Poland; E-mail: zbigniew.ustrnul@uj.edu.pl

(Manuscript received in final form October 18, 2010)

Abstract — Recent developments in GIS techniques have produced a wide range of powerful methods for capturing, modeling, and displaying of climate data. The main aim of the study was to identify an optimal interpolation method to describe the spatial differentiation of the climatic water balance in Poland based on meteorological data (temperature, precipitation, solar radiation) collected at 15 weather stations from 1986 to 2006. A climatic water balance index (*CWB*) was created based on a simplified definition, where it is interpreted as the difference between the precipitation total (*P*) and potential evapotranspiration (*PE*). The latter was calculated using the so-called Turc Equation. Four different spatial interpolation methods were used: (1) inverse distance weighting (IDW), (2) local polynomial (LP), (3) radial basis function (RBF), and (4) ordinary kriging. A subjective visual analysis of maps, root mean square error values, and coefficients of correlation indicated that the best *CWB* interpolation methods are the radial basis function method and the ordinary kriging method. However, spatial interpolation results suggest that the problem is more complex. Calculations performed for selected points of reference suggest that local geographic factors play an important role in the shaping of *CWB*. Such results also confirm the need to perform spatial climatic water balance analysis with special attention being paid to local conditions. Further research is needed that takes into account different temporal and spatial scales and aims to test established methods in other regions in Europe.

Key-words: spatial analysis, GIS, interpolation methods, climatic water balance, Poland

1. Introduction

Recent developments in GIS techniques have produced a wide range of powerful methods for capturing, modeling, and displaying of climate data. Advanced data processing methods allow for detailed analysis of climate elements on different temporal and spatial scales. GIS techniques designed to map temperature and atmospheric precipitation fields have received the most attention thus far. However, researchers are often interested not in the meteorological elements themselves but in the information that can be extracted from them in the form of various indices, which are useful in the environmental and social sciences (*Tveito et al.*, 2008).

The *CWB* is a complex index that shows a climate-based assessment of the water resources in a given area. It focuses mainly on the difference between precipitation and potential evapotranspiration. Values of the index depend on many different variables such as solar radiation, relief, land use, and urban development. Spatial distribution of the climatic water balance appears to be very important in spatial management, agriculture, and hydro-climatological modeling. Since 2007, the Drought Monitoring System for Poland has been provided by the Institute of Soil Science and Plant Cultivation – State Research Institute in Pulawy. In the system, meteorological conditions that are causing drought are evaluated by the climatic water balance expressed by the difference between the precipitation and potential evapotranspiration (by Penman formulae). Nevertheless, it has not been the subject of detailed analysis thus far. Data covering any longer period is not readily available – especially evapotranspiration data – which creates the problem of index interpretation, especially due to its reliance on spatial differentiation. Therefore, the main aim of the study was to identify an optimal interpolation method to describe the spatial differentiation of the water balance in Poland taking into account a number of scale-based variables.

2. Source material and methods

Analyses of the climatic water balance are usually developed for regions where input data, mainly air temperature and precipitation, can be readily obtained from meteorological stations. The research presented herein is based on mean monthly values of air temperature as well as monthly solar radiation and precipitation totals. The data were obtained from 61 meteorological stations (temperature and precipitation) and 23 actinometric stations (solar radiation) for the 1951–2006 and 1986–2006 time periods, respectively. Not all meteorological stations collect actinometric data, which is why data was obtained from only 15 stations and covers the period from 1986 to 2006 (*Fig. 1*).

The climatic water budget was introduced into the research literature in the middle of 20th century by *Thornthwaite* (1948). He described the budget as the balance of precipitation, potential evapotranspiration, and actual evapotranspiration, taking into account both soil moisture utilization and soil moisture recharge (*Oliver* and *Fairbridge*, 1987). According to Thornthwaite and his colleagues (*Thornthwaite* and *Mather*, 1957), an average climatic water budget model can be expressed using two interrelated equations:

$$P = ET + S, \qquad PE = ET + D, \tag{1}$$

where P is the precipitation, ET is the evapotranspiration, PE is the potential evapotranspiration, S is the moisture surplus, and D is the moisture deficit. The first equation describes water inflow, outflow, and storage, and the second equation describes energy demands. The procedure designed by *Thornthwaite* and *Mather* (1957) to calculate climatic water balance parameters is still widely used in CWB research (e.g., *Kar* and *Verma*, 2005; *Tateishi* and *Ahn*, 1996).

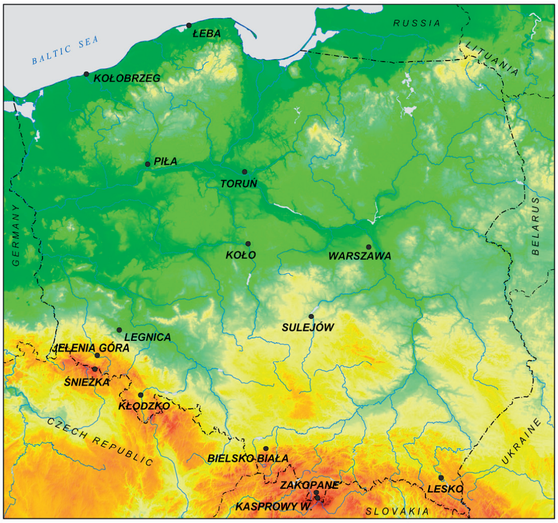


Fig. 1. Locations of the meteorological stations used in the research study.

Evapotranspiration process is the principal component of the climatic water balance, as it returns 60% to 80% of precipitation back into the atmosphere. In order to determine the value of the *CWB* index, the magnitude of evapotranspiration must be properly estimated. Owing to the difficulty of obtaining accurate field measurements, evapotranspiration is commonly computed from weather data using empirically derived formulas. A large number of more or less empirical methods have been developed over the last 50 years and are designed to estimate actual and potential evapotranspiration from different climatic variables. Some of the methods are only valid under specific climatic and agronomic conditions and cannot be applied under conditions different from those under which they were originally developed (*Allan et al.*, 2004). As a result, the FAO Penman-Monteith Method is now recommended as the standard method for the definition and computation of the reference evapotranspiration, ET_o. The reference evapotranspiration provides a standard to which evapotranspiration at different periods of the year or in different regions can be compared (*Allan et al.*, 2004).

The subject of *CWB* spatial interpolation is very complex. It is, first and foremost, a subject associated with the problem of the spatial interpolation of evapotranspiration, which varies considerably with changes in the natural environment. The second complexity has to do with the availability of data. Given the complicated nature of the subject, it is no wonder that there exist many methods that attempt to model the spatial differentiation of evapotranspiration (e.g., *Nováky*, 2002; *Xinfa et al.*, 2002; *Kar* and *Verma*, 2005; *Loheide* and *Gorelick*, 2005; *Fernandes et al.*, 2007). Remote sensing is becoming more commonly used to address this research issue and often supplements ground-based observations (*Woolhizer* and *Wallace*, 1984; *Rosema*, 1990; *Kalma et al.*, 2008).

In Poland, most evapotranspiration and climatic water balance research is focused on the identification of a model that would best suit weather conditions in Poland. The following formulas were used in existing research: *Turc* (1961) method for potential evapotranspiration, *Bac* (1970) reference evaporation formulae for local index, and Penman modified to Polish conditions (*Sarnacka et al.*, 1983). All three methods were applied to the analysis of the measurements data (Wild scale and GGI-3000 pan evaporimeter). Although the Turc method produced the largest differences between evapotranspiration totals measured *in situ* and values derived empirically (also shown by Hungarian research by *Nováky*, 2002), the method proved to be useful because of data availability issues. It was also selected because of other research that has shown that it is best at determining relationships with elevation (*Kowanetz*, 1998, 2000).

The climatic water balance index (*CWB*) was created based on a simplified definition where it is interpreted as the difference between the precipitation total (*P*) and potential evapotranspiration (*PE*). The final formula was the following:

$$CWB = P - 0.4\frac{t}{t+15}I + 50, \tag{2}$$

where *P* is the monthly precipitation totals, *t* is the monthly average air temperature [°C], and *I* is the monthly sum of total solar radiation [cal cm^{-2} day^{-1}].

Given the limited nature of the source data (15 data points only) and the existence of strong relationships between potential evapotranspiration and geographic factors (the same is true for *CWB*), geographic parameter regression models were used to produce grid data consisting of annual *CWB* totals (CWB_{yr}) and vegetation period (April – September) *CWB* totals (CWB_{veg}) at a 0.2 degree spatial resolution (latitude and longitude). The resolution was chosen as the best for regional scale studies. Moreover, the DTM resolution of 250 × 250 meters was available for calculations. Simple and multiple regression models were used, taking into account the dependence of *CWB* on elevation above sea level

(*H*), longitude (*λ*), and latitude (*φ*). The following formulas were used to perform calculations:

$$CWB = f(H) + b, \quad CWB = f(H) + f(\lambda) + f(\varphi) + b, \tag{3}$$

where *b* is the constant value.

Table 1 includes coefficients of correlation between geographic coordinates and *CWB* values on an annual as well as seasonal basis. The coefficients are very large – generally above 0.9 – and statistically meaningful at $\alpha = 0.05$. The coefficients of correlation tend to be somewhat larger when a multiple regression model is used. The value for the growing season is 0.95 and the value for the entire year is 0.94 (*Table 1*). The large size of the coefficient of correlation made it possible to use the regression method in order to calculate *CWB* for individual grids. The calculated values were then used in spatial analysis based on a variety of interpolation (spatialization) methods.

Table. 1. Coefficients of correlation (CC) between geographic parameters (*H*, *φ*, *λ*) and climatic water balance values (*CWB*) for the growing season (April – September) and for an entire year

CC (simple)	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Apr – Sep**	**Year**
H vs. *CWB*	0.95	0.91	0.91	0.90	0.90	0.86	0.92	0.92

CC (multiple)	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Apr – Sep**	**Year**
$H + \varphi + \lambda$ vs. *CWB*	0.97	0.94	0.92	0.92	0.91	0.89	0.95	0.94

There is a dearth of publications on optimal spatial CWB analysis methods, which has led to the testing of a variety of methods based on experiences with the interpolation of individual climate elements (*Dobesch et al.*, 2007; *Tveito et al.*, 2008). RMSE (root mean square error) analysis was used to assess the influence of interpolation methods on the analysis of spatial *CWB* differentiation. The source material available – 15 data points – was used as a source of reference. The relationship between results obtained during the spatialization process and values calculated based on field measurement data were also investigated.

3. Results and discussion

Four different spatial interpolation methods were used: (1) inverse distance weighting (IDW), (2) local polynomial (LP), (3) radial basis function (RBF), and (4) ordinary kriging. The first three are so-called deterministic methods. The fourth method, kriging, is used the most often and it is a geostatistical method.

Spatial interpolation was performed for different seasons and for the entire year, for Poland as a whole, using all four methods. RMSE values and coefficients of correlation as well as a subjective visual analysis of maps produced results that do not differ very much. However, the coefficient of correlation and RMSE suggest a somewhat more accurate interpolation based on RBF and kriging (*Table 2*).

Table 2. Validation results for different interpolation methods used in *CWB* calculations

Interpolation method (simple regression)	**Year**			**Vegetation period**		
	r	**RMSE**	**σ**	**r**	**RMSE**	**σ**
IDW	0.79	602	78	0.83	122	67
LP	0.83	683	207	0.87	186	177
RBF	0.84	641	145	0.88	147	125
Kriging	0.84	637	138	0.87	143	118
Interpolation method (multiple regression)	**r**	**RMSE**	**σ**	**r**	**RMSE**	**σ**
IDW	0.77	637	102	0.79	122	95
LP	0.83	710	196	0.75	177	174
RBF	0.84	673	144	0.86	141	131
Kriging	0.84	669	139	0.85	138	127

The *CWB* maps generated using the above methods can be found in *Figs. 2a,b*. The maps present annual *CWB* values as well as *CWB* values for the growing season (April – September). Differences between the annual spatial distribution and the growing season distribution are readily apparent. Annual *CWB* values range from 430 mm to 1200 mm, with maxima in the southern part of the country (mountains and uplands) and minima in the central part of the country (*Figs. 2a,b*). *CWB* fluctuates the most during the growing season (April – September), with positive values being recorded only in the mountains (up to 200 mm) and negative values (moisture deficit) across the rest of the country – as low as –230 mm in central Poland.

At the same time, *Figs. 2a,b* also show differences in spatial distribution resulting from the interpolation of input data using the simple regression method and the multiple regression method. *Table 2* shows validation results for the interpolation methods used in the study.

The interpolation results generated for Poland as a whole may be considered good, as the differences produced by different methods are small. However, a closer look at the problem on a local scale points to a great deal of complexity. Calculation results for different locations indicate that geographic influence is a factor that does affect *CWB*.

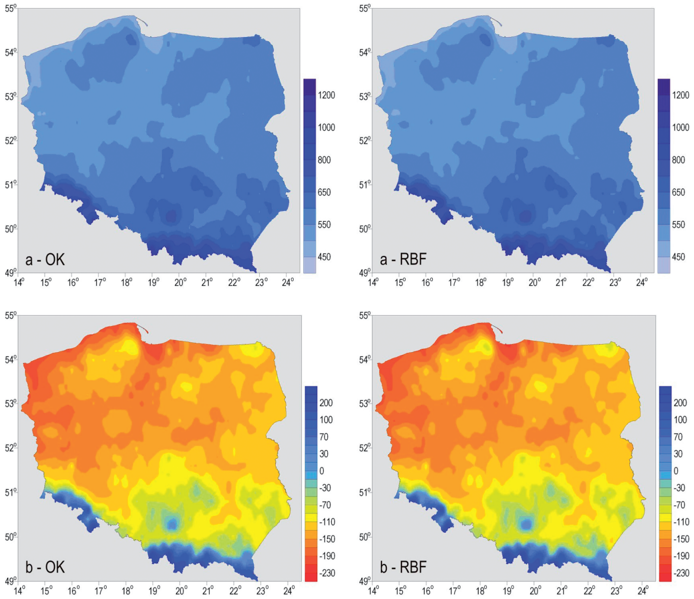


Fig. 2a. Spatial distribution of the *CWB* (mm) in Poland according to different interpolation methods: radial basis function (RBF) and ordinary kriging (OK) (simple regression model); a – annual values, b – vegetation period (April – September) values.

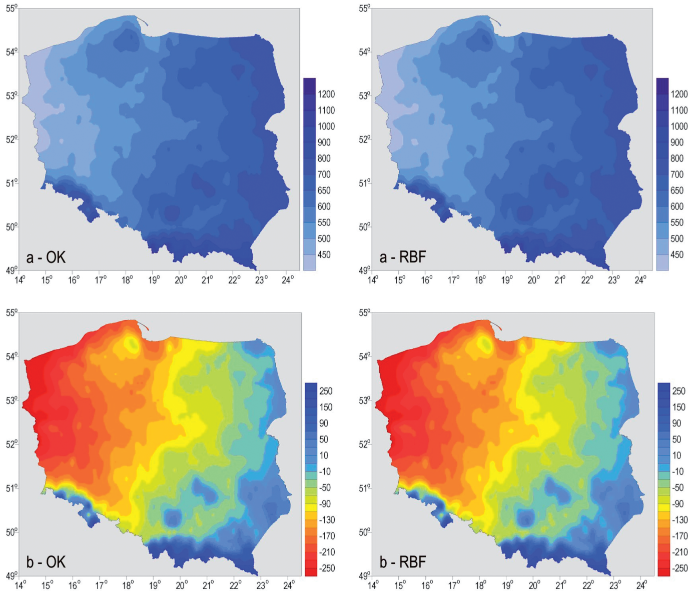


Fig. 2b. Spatial distribution of the *CWB* (mm) in Poland according to different interpolation methods: radial basis function (RBF) and ordinary kriging (OK) (multiple regression model); a – annual values, b – vegetation period (April – September) values.

Decidedly larger differences between values calculated based on field measurements and those produced by the model in question can be observed for the growing season. Using the simple regression model, errors exceed 100% of values calculated for the Jelenia Góra Basin and the Kłodzko Basin (*Fig. 3*). The multiple regression model performs the worst for coastal locations (Łeba, Kołobrzeg) and points near the eastern border of Poland – Lesko (*Fig. 3*). The uncertainty of the results obtained suggests that it is necessary to use supplemental descriptive variables.

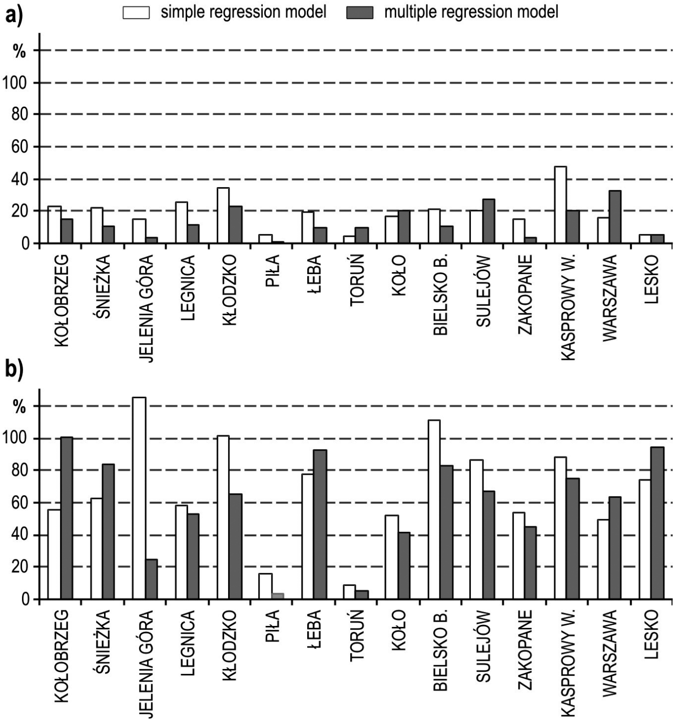


Fig. 3. Mean percent error of estimated and calculated *CWB* values for the meteorological stations used in the research study.

Component variables such as atmospheric precipitation and potential evapotranspiration make the climatic water balance strongly dependent on local conditions. Any analysis of data must take into account local relief and land cover (biological and soil factors). Elevation above sea level and geographic coordinates are not enough to perform an accurate spatial analysis of climatic water balance.

As elevation above sea level is a key component of spatial differentiation analysis of atmospheric precipitation (*Bac-Bronowicz*, 2003; *Łupikasza et al.*, 2007), the choice of descriptive variables is a key factor in spatial *CWB* analysis. Errors may also occur as a result of poor spatial coverage provided by weather stations as well as the interpolation and mapping techniques used. Regarding evapotranspiration estimates mapping, it is usually affected by modeling errors resulting from the derivation of *ET* values (*Climatic Atlas…*, 2001). Hence, the complexity of the evapotranspiration process demands the consideration of local conditions.

The state of current understanding of microclimate differentiation, especially that in mountain areas, suggests that other geographic variables should be taken into consideration. In order to accurately describe the spatial distribution of *CWB*, it is necessary to take into account slope, aspect, land use, and soil type – all of which determine how much solar radiation is available and, consequently, the value of the air temperature (*Ustrnul* and *Czekierda*, 2005). Both solar radiation and air temperature affect the degree of evapotranspiration. Furthermore, the parameters must be calculated independently for smaller regions – especially regions characterized by specific mesoclimate conditions such as those found in coastal or mountain areas.

4. Conclusions

The spatial interpolation results presented herein, based on four different interpolation methods, prove the hypothesis that spatial differentiation analysis of the climatic water balance should take local conditions into account. The validation results are sufficient to fully assess results obtained on a national scale (Poland only), but insufficient with respect to individual geographic locations, where local differences can be quite significant.

Reducing the size of the research area appears to be a reasonable next step. A solid understanding of the causes and effects of particular component elements, such as the natural environment, on the *CWB* index should help in the process of selecting descriptive variables. Existing research suggests the use of distance from a body of water, land cover, and relief as supplemental factors. Another key factor is the selection process of the evapotranspiration (potential and/or actual) calculation method, as this appears to be the main source of possible *CWB* errors.

Given the difficult nature of the analytical process involved, the accuracy of the spatialization method is less important. The most important objectives for further research are validation of the obtained results using different evapotran-spiration formulas and the optimization and testing of spatialization techniques. A few other descriptive variables should also be considered (e.g., circulation types, air masses). Further research will be designed to focus on different temporal and spatial scales as well as the validation in other areas of Europe.

References

Allan, R.G., Pereira, L.S., Raes, L., Smith, M., 2004: Guidelines for computing crop water requirements. *FAO Irrigation and Drainage Papers 56*, p. 328

Bac, S., 1970: Studies on the correlation between free water surface evaporation, areal and potential evapotranspiration (in Polish). Prace *i Studia Komit. Gosp. Wodnej i Sur. PAN 10*, 287-366.

Bac-Bronowicz, J., 2003: Methods of the visualisation of precipitation based on various observation measurement periods in GIS. In *Man and Climate in the 20th Century*. *Studia Geograficzne 75*, Wyd. Uniw. Wrocławskiego, Wrocław, 559-563.

Climatic Atlas of Australia, 2001: Evapotranspiration (ET). Bureau of Meteorology, Australia, p. 45.

Dobesch, H., Dumolard, P., Dyras, I. (eds.), 2007: *Spatial Interpolation for Climate Data, ISTE –* Geographical Information Systems Series. London – Newport Beach, p. 284.

Fernandes, R., Korolevych, V., Wang, S., 2007: Trends in land evapotranspiration over Canada for the period 1960–2000 based on in situ climate observations and a land surface model. *Journal of Hydrometeorology 8*, 1016-1030.

Kalma, J D., McVicar, T.R., McCabe, M.F., 2008: Estimating land surface evaporation: A review of methods using remotely sensed surface temperature data. *Surv. Geophys.* 29, 421-469.

Kar, G., Verma, H.N., 2005: Climatic water balance, probable rainfall, rice crop water requirements and cold periods in AER 12.0 in India. *Agr. Water Manage. 72*, 15-32.

Kowanetz, L., 1998: *Climatic Water Balance in Upper Vistula River Basin* (in Polish). PhD thesis, Jagiellonian University.

Kowanetz, L., 2000: On the method of determining the climatic water balance in mountainous areas, with the example from Polish Carpathians. *Zeszyty Naukowe UJ, Prace Geograficzne 105*, 137–164.

Loheide, S.P., Gorelick, S.M., 2005: A local-scale, high-resolution evapotranspiration mapping algorithm (ETMA) with hydroecological applications at riparian meadow restoration sites. *Remote Sens. Environ. 98*, 182-200.

Łupikasza, E., Ustrnul, Z., Czekierda, D., 2007: The role of explanatory variables in spatial interpolation of selected climate elements. *Roczniki Geomatyki 5,* 1, 55-64.

Nováky, B., 2002: Mapping of mean annual actual evaporation on the ex ample of Zagyva catchment area. *Időjárás 106,* 227-238.

Oliver, J.E., Fairbridge, R.W., 1987: *The Encyclopedia of Climatology*. Van Nostrand Reinhold Company Inc., New York, p. 986.

Rosema, A., 1990: Comparison of Meteosat-based rainfall and evapotranspiration mapping in the Sahel region. *Int. J. Remote Sens. 11,* 2299-2309.

Sarnacka, S., Brzeska, J., Świerczynska, H., 1983: Selected methods in distinguishing potential evapotranspiration (in Polish). *Mat. Bad. Ser. Gosp. Wodn. i Ochr. Wód, IMGW*, 1-35

Tateishi, R., Ahn, C.H., 1996: Mapping evapotranspiration and water balance for global land, ISPRS. *Journal of Photogrammetry and Remote Sensing 51,* 4, 209-215.

Thornthwaite, C.W., 1948: An approach. towards a rational classification of climate. *Geogr. Rev. 38*, 55-94.

Thornthwaite, C.W., Mather, J.R., 1957: Instructions and tables for computing potential evapotranspiration and the water balance. *Publ. Climatol. 10*, 185-311.

Turc, L., 1961: Evaluation des besoins en eau d'irrigation, évapotranspiration potentielle. *Annales Agronomiques 12,*1,13-49.

Tveito, O.E., Wegehenkel, M., Wel van der, F., Dobesch, H. (eds.), 2008: The use of geographic information systems in climatology and meteorology. *Final Report COST Action 719*, COST Office, p. 246.

Ustrnul, Z., Czekierda, D., 2005: Application of GIS for the development of climatological air temperature maps: an example from Poland. *Meteorol. Appl. 12*, 43-50.

Woolhizer, D.A., Wallace, D.E., 1984: Mapping average daily pan evaporation. *J. Irrig. Drain. Eng. 110,* 246-250.

Xinfa, Q., Yan, Z., Changming, L., 2002: A general model for estimating actual evaporation from non-saturated surfaces. *J. Geogr. Sci. 12,* 479-484.

NEWS

Farewell to the Executive Editor, Margit Antal

IDŐJÁRÁS (Weather in English) is one of the oldest meteorological journals in Europe. Since the end of the nineteenth century it has published papers on meteorology as well as other news in the Hungarian language. After the Second World War it was felt that other languages must also be used in order to make publication for foreign authors possible. However, this caused a rather chaotic situation since the language of the publication was not determined. Thus, papers of quite different quality were published in at least five languages.

At the end of the seventies and at the beginning of the eighties it became clear that papers should be accepted only in one language, namely in the world-common language of natural sciences: in English. One also believed that the manuscript reviewing procedure must be similar to that applied in leading scientific journals. The philosophy was to publish papers of Hungarian specialists and other scientists from the neighboring countries in a way that would be understandable all over the world. Further, we believed at that time, that an international forum on atmospheric science was necessary for strengthening the relationship between scientists behind and outside the iron curtain.

As the chairman of the editorial board (later the Editor-in-Chief) I needed a new board consisting of internationally well-known scientists and, also very important, a secretary for fulfilling the administration necessary. It was an obvious solution to choose for this purpose my own secretary, Ms *Margit Antal* (her married name is *Dr Antal Emánuelné*), called by everybody "*Pimpi*", who had already proved that she is a very good person to arrange correspondence and typing not only in Hungarian, but also in English, French, and German. It turned out later that this choice was an excellent one.

I have to emphasize that Pimpi did not finish any course to be secretary. She began to work at the Department on Solar Radiation of the Hungarian Meteorological Service in 1963. After some years she continued similar activity at the Department on Heat and Water Balance. Her bosses were always very satisfied with her work (mostly calculations and observations), she quickly became famous because of her accuracy and precision. This was obviously the reason why the executive editor of IDŐJÁRÁS at that time (*J. Kakas*) used Pimpi's ability to help him in editing of the journal. And, more important for myself, this motivated me to ask her to be my secretary, when I was appointed director of the research institute (Central Institute for Atmospheric Physics) of the Hungarian Meteorological Service. In the life of anyone among us there are good and bad decisions. Anyway, this was a good one for me. Pimpi, as secretary, helped my work in an excellent and correct way until leaving the service in 1992.

In the same year Pimpi became officially the technical editor of the journal. In 1996 the president of the Hungarian Meteorological Service nominated her to the post of executive editor. By this time she already knew everything concerning the editorial work. She played a decisive role in introducing the up-to-date computerized journal production, giving to IDŐJÁRÁS the present attractive aspect. IDŐJÁRÁS has been indexed and abstracted in Science Citation Index Expanded and Journal Citation Reports/Science Edition since 2007. This milestone could not be reached without her perfect editorial activity. She has executed the editorial work to the satisfaction of everybody, including members of the editorial board, authors, and readers. For acknowledging her work during decades for IDŐJÁRÁS and the Hungarian Meteorological Service, the Minister of the Environment Protection and Water Management rewarded her in 2006 with the medallion "Pro Meteorologia".

Dear Pimpi! On the occasion of your retirement, on behalf of all the Editor-in-Chiefs with whom you worked, I want to thank you for your efforts to make this journal an international forum. I wish you further health and a long and peaceful life after almost fifty years of professional work!

Ernő Mészáros
Former Editor-in-Chief
Member of the Hungarian Academy of Sciences

INSTRUCTIONS TO AUTHORS OF *IDŐJÁRÁS*

The purpose of the journal is to publish papers in any field of meteorology and atmosphere related scientific areas. These may be

- research papers on new results of scientific investigations,
- critical review articles summarizing the current state of art of a certain topic,
- short contributions dealing with a particular question.

Some issues contain "News" and "Book review", therefore, such contributions are also welcome. The papers must be in American English and should be checked by a native speaker if necessary.

Authors are requested to send their manuscripts to

Editor-in Chief of IDŐJÁRÁS
P.O. Box 39, H-1675 Budapest, Hungary
E-mail: antal.e@met.hu

including all illustrations. MS Word format is preferred in electronic submission. Papers will then be reviewed normally by two independent referees, who remain unidentified for the author(s). The Editor-in-Chief will inform the author(s) whether or not the paper is acceptable for publication, and what modifications, if any, are necessary.

Please, follow the order given below when typing manuscripts.

Title page: should consist of the title, the name(s) of the author(s), their affiliation(s) including full postal and e-mail address(es). In case of more than one author, the corresponding author must be identified.

Abstract: should contain the purpose, the applied data and methods as well as the basic conclusion(s) of the paper.

Key-words: must be included (from 5 to 10) to help to classify the topic.

Text: has to be typed in single spacing on an A4 size paper using 14 pt Times New Roman font if possible. Use of S.I. units are expected, and the use of negative exponent is preferred to fractional sign. Mathematical formulae are expected to be as simple as possible and numbered in parentheses at the right margin.

All publications cited in the text should be presented in the *list of references*, arranged in alphabetical order. For an article: name(s) of author(s) in Italics, year, title of article, name of journal, volume, number (the latter two in Italics) and pages. E.g., *Nathan, K.K.,* 1986: A note on the relationship between photo-synthetically active radiation and cloud amount. *Időjárás 90,* 10-13. For a book: name(s) of author(s), year, title of the book (all in Italics except the year), publisher and place of publication. E.g., *Junge, C.E.,* 1963: *Air Chemistry and Radioactivity.* Academic Press, New York and London. Reference in the text should contain the name(s) of the author(s) in Italics and year of publication. E.g., in the case of one author: *Miller* (1989); in the case of two authors: *Gamov* and *Cleveland* (1973); and if there are more than two authors: *Smith et al.* (1990). If the name of the author cannot be fitted into the text: *(Miller,* 1989); etc. When referring papers published in the same year by the same author, letters a, b, c, etc. should follow the year of publication.

Tables should be marked by Arabic numbers and printed in separate sheets with their numbers and legends given below them. Avoid too lengthy or complicated tables, or tables duplicating results given in other form in the manuscript (e.g., graphs).

Figures should also be marked with Arabic numbers and printed in black and white or color (under special arrangement) in separate sheets with their numbers and captions given below them. JPG, TIF, GIF, BMP or PNG formats should be used for electronic artwork submission.

Reprints: authors receive 30 reprints free of charge. Additional reprints may be ordered at the authors' expense when sending back the proofs to the Editorial Office.

More information for authors is available: antal.e@met.hu

Published by the Hungarian Meteorological Service

Budapest, Hungary

INDEX 26 361

HU ISSN 0324-6329