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*.

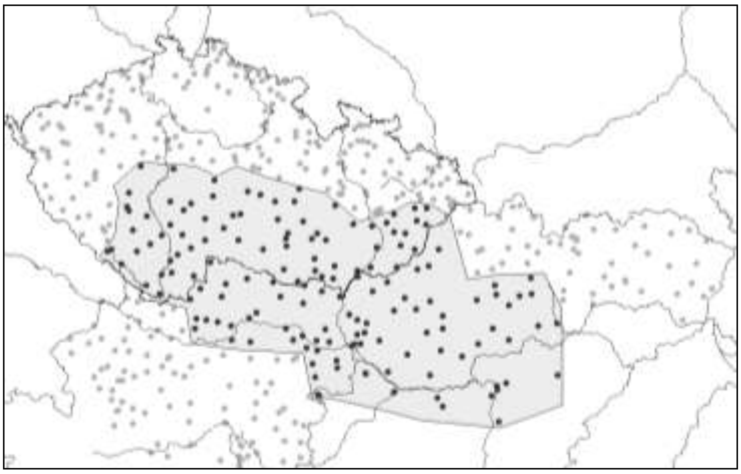

Fig. 1. CECILIA Central European domain (shaded area) with available climatological stations (dark / light dots for stations inside / outside the domain).

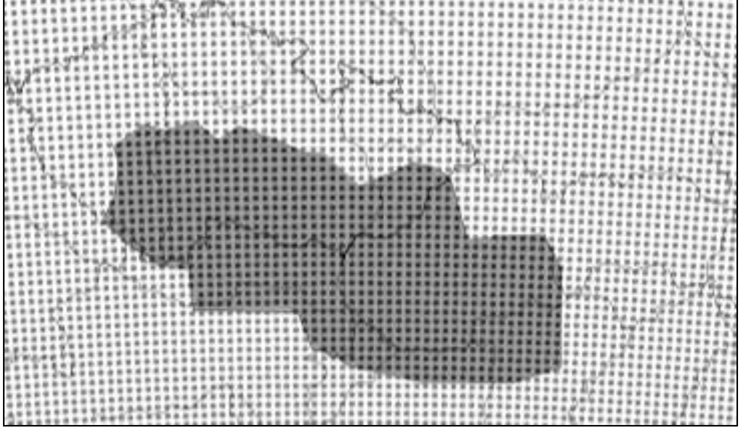

Fig. 2. Grid points of ALADIN-CLIMATE/CZ (dark / light dots) available within / outside the CECILIA Central European domain (shaded area).

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).

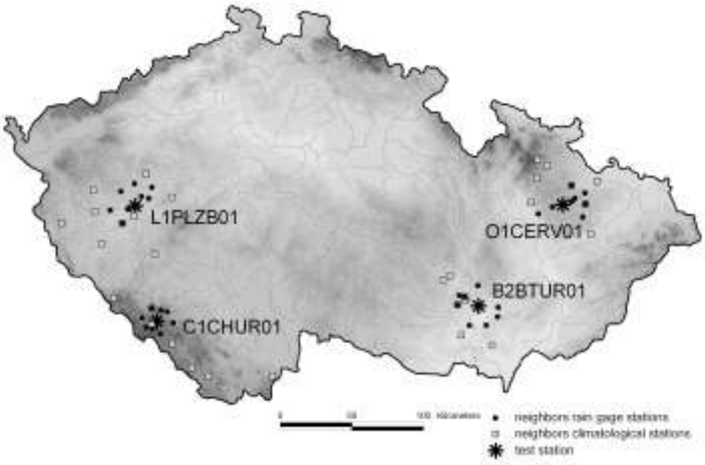
L1PLZB01
O1CERV01
B2BTUR01
C1CHUR01
0
50
100 Kilometers
● neighbors rain gage stations
□ neighbors climatological stations
test station

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.

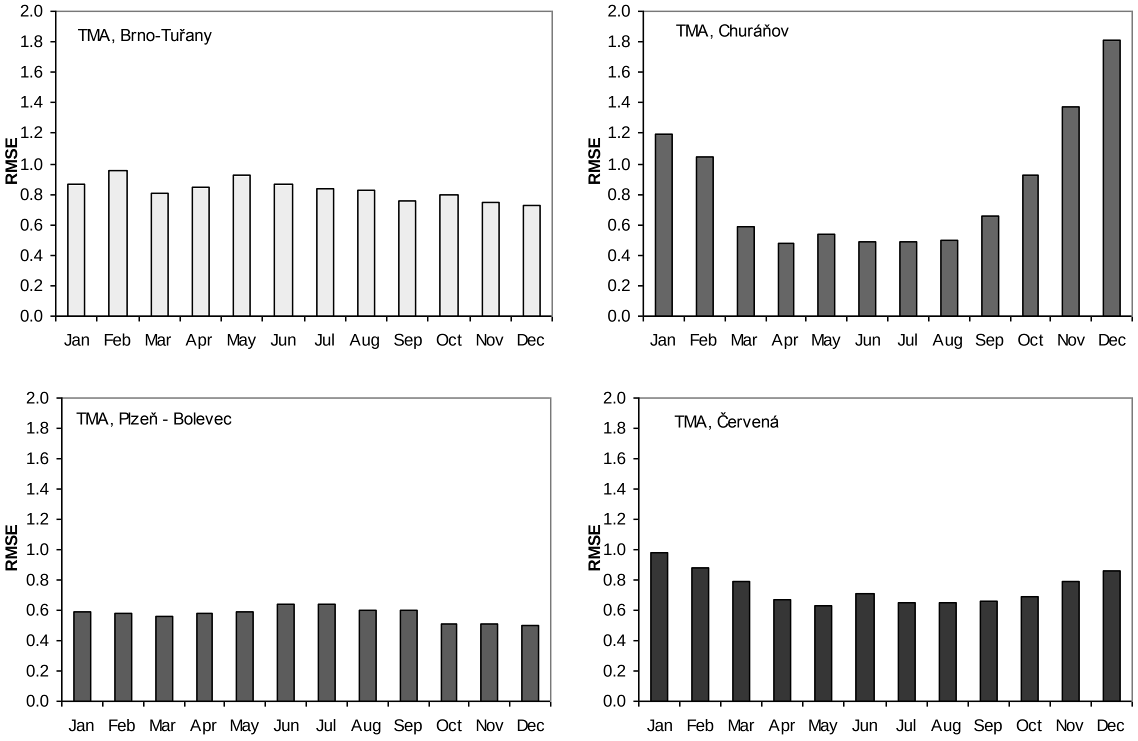

Fig. 4. RMSE (in °C) for four base (tested) stations and maximum temperature.

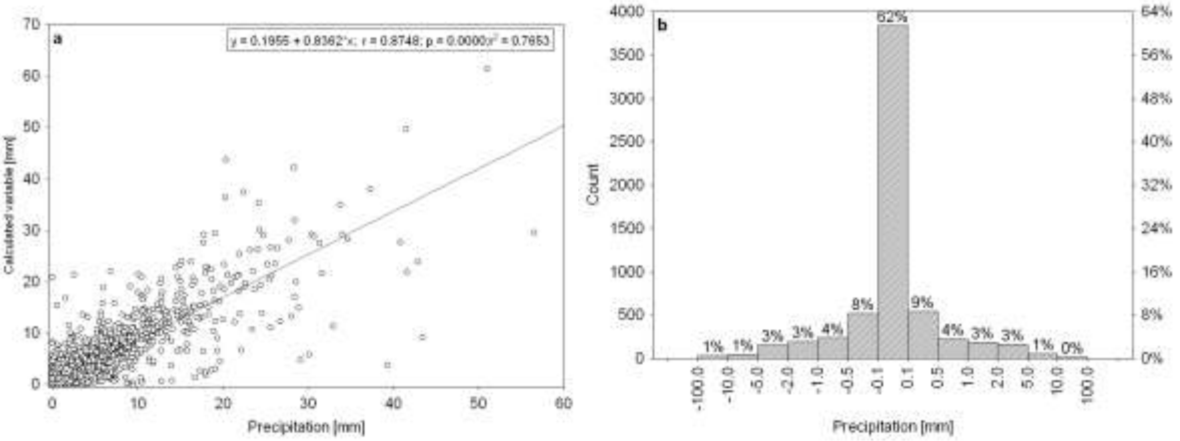

Fig. 5. Scatter plot for calculated and original values of precipitation (left) and histogram of differences between the calculated and original values (right) at station Brno-Tuřany.

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