**Advances in
Science and
Research**

Calculation of climatic reference values and its use for automatic outlier detection in meteorological datasets

B. Téllez, T. Cernocky, and E. Terradellas

Instituto Nacional de Meteorología, CMT de Cataluña, Spain

Abstract. The climatic reference values for monthly and annual average air temperature and total precipitation in Catalonia – northeast of Spain – are calculated using a combination of statistical methods and geostatistical techniques of interpolation. In order to estimate the uncertainty of the method, the initial dataset is split into two parts that are, respectively, used for estimation and validation. The resulting maps are then used in the automatic outlier detection in meteorological datasets.

1 Introduction

In climatology, as well as in geography, in biology or in other fields, it is necessary to have climatic reference values for a specific spot or region. Rather often, there is not a meteorological station in the area of interest. Moreover, there is not any guarantee that data supplied by the nearest station describe accurately enough its climatic conditions. There are numerous methods to build a map to represent the spatial distribution of climatic parameters: deterministic, probabilistic or physical methods, artificial neural networks, etc. (Lam, 1983; Demyanov et al., 1998; Ninyerola et al., 2007).

On the other hand, studies including analysis of meteorological data usually require a preliminary arduous and unpleasant debugging of the original datasets in order to purge wrong values. Most times, this task has been carried out through the comparison of the dataset to analyse with others recorded at nearby stations. Nevertheless, the method is not straightforward, especially when it is applied to magnitudes with strong spatial variations. Such variations are usually more evident in areas with complex physiographic features.

The problem can be mitigated through the use of the difference with the reference value instead of the original value. In Sect. 3, the method used to obtain the spatial distribution of the climatic magnitudes is described. Section 4 refers the attempts to automatically identify possible outliers from the spatial distribution of differences between meteorological data and climatic reference values.

2 Geographic framework and data

The geographic framework for the present study is Catalonia, which is located in the north-eastern corner of the Iberian Peninsula. It has a surface area of $31\,895\,\mathrm{km}^2$ and a significant geographic diversity, from the Mediterranean Sea to peaks over 3000 m above sea level in the Pyrenees. The climatological dataset contains 1961–1990 averages of monthly data (INM, 2000) recorded at 144 thermometric and 302 pluviometric stations. 15 thermometric and 55 pluviometric stations are located over surrounding regions and have been included in the study in order to smooth border effects. Auxiliary physiographic parameters, such as altitude, distance to sea or land slope have been retrieved from a digital elevation model (DEM) using GIS techniques. The DEM, with a 200-m resolution, has been provided by the *Institut Cartogràfic de Catalunya* (ICC).

3 Calculation of the climatic reference values

The climatic magnitudes – average monthly temperature and monthly precipitation – are represented using a combination of statistic methods and geostatistical techniques of interpolation. First, a multiple regression analysis yields a model that relates the climatic variable with several physiographic parameters. Then, the residuals, that is, the difference between the observed values and the predictions of the linear model, are spatially interpolated using an ordinary kriging. The result, a map of residuals, shows the amount of variability that is not explained by the regression model. This variability can be attributed either to errors in the original datasets or to relationships with physiographic or meteorological magnitudes that have not been considered in the

---

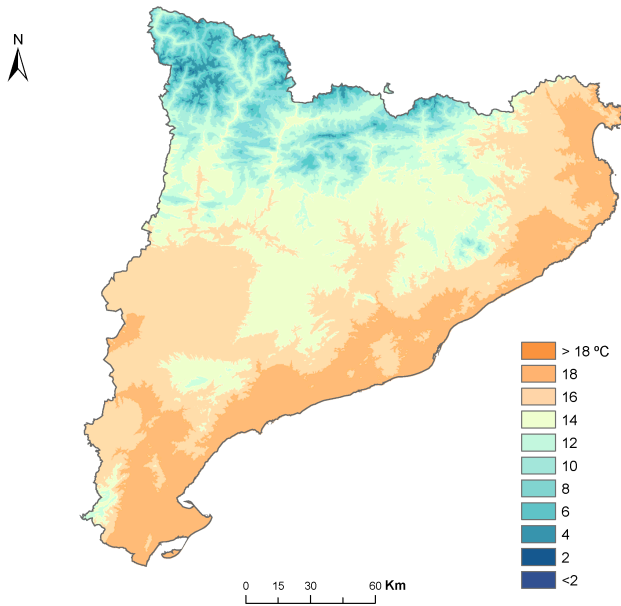*Correspondence to:* B. Téllez
(bea@inm.es)

Figure 1. Reference values for the average temperature in October.

Figure 2. Distribution of the thermal anomaly, that is, the difference between the monthly-averaged temperature and its reference value, in June 2007. In both sides, zoom into areas where the slope exceeds the threshold value of $0.7°C\,km^{-1}$.

regression model. The final result is the addition of the map drawn from the model predictions and the map of residuals (see in Fig. 1 the final map of average temperatures in October).

The physiographic parameters used in the regression are: altitude, latitude, longitude, distance to sea and average and minimum altitudes in circles of 5 to 40 km radius, which are considered, following Dyras et al. (2005), in order to account for topography at different scales. Then, the best predictive parameters for a given variable are chosen using a technique called stepwise regression (Hocking, 1976).

For average monthly temperature, altitude and distance to sea turn out to be the best predictive physiographic variables. The altitude explains most of the variance and it is significant in all cases. On the other hand, the distance to sea is only significant during the cold season, from October to March. The coefficients of the regression equation are displayed in Table 1.

For total monthly precipitation, the best predictive variables are those related to altitude. Between May and September, the average altitude in a 40-km radius circle explains most of the variance. In the other months, the most significant variables are average and/or minimum altitudes for a 10/15-km radius circle. Latitude and longitude have also resulted to be significant in some cases. The coefficients of the regression equation are shown in Table 2.

An ordinary kriging is used to interpolate both temperature and precipitation residuals. The empirical semivariogram, built following Journel and Huijbreghts (1978), is modelled with an exponential curve, with the range set to 40 km. The values of nugget for the different months are in the range
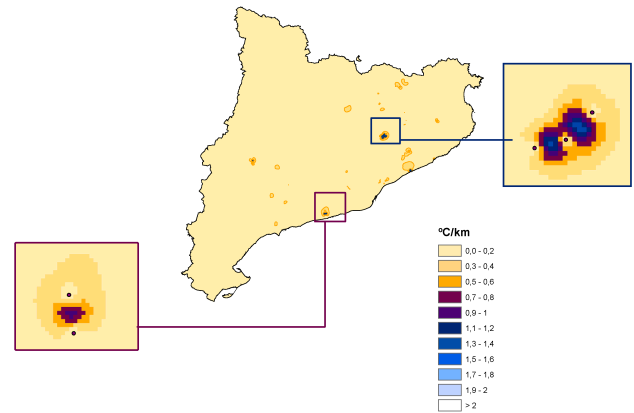
$0.2–1°C^2$ and $5–30\,mm^2$, and those of total sill in the range $0.8–1.6°C^2$ and $70–140\,mm^2$.

The initial meteorological dataset is split into two parts in order to estimate the uncertainty of the method. The full process is first carried out using 70% of the initial data, which are selected at random. Then, the remaining 30% of data are compared with values retrieved from the resulting maps. This linear regression yields adjusted coefficients of determination that are over 0.80 for average temperature (they range from 0.81 in July to 0.89 in November) and over 0.75 for total precipitation (they range from 0.75 in September to 0.90 in July). The highest errors are found in areas of great orographic complexity and sparse observational coverage, where both the data fitting into the regression model and the interpolation of residuals are deficient. Anyway, since the overall results are acceptable, the process is finally carried out using the whole meteorological dataset.

4 Application of the reference values in debugging meteorological datasets

At the *Instituto Nacional de Meteorología*, it has been implemented a method to automatically detect possible errors in the monthly average temperature and total precipitation datasets. It is based on GIS techniques and makes use of the climatic reference values previously computed.

The first step of the method is the calculation of the so-called monthly anomalies. The thermal anomalies are defined as the difference between a monthly averaged temperature and its reference value. The precipitation anomalies are defined in a similar way, but they are expressed as a percent of the reference value.

Then, two different filters are applied in order to find out those stations whose data are suspicious and require further investigation. The first filter simply detects the stations

Table 1. Coefficients of the regression equation for monthly average temperature (T): $T=C+Ax_1+Bx_2$, where x_1 is the altitude of the station, x_2 is the distance to sea. The adjusted coefficient of regression and the root mean square error are also displayed.

	C (°C)	A (°C/km)	B (°C/km)	Adj. r^2	R.M.S.E. (°C)
Jan	7.97	−2.9886	−0.0223	0.78	1.35
Feb	9.56	−4.6524	−0.0132	0.87	1.10
Mar	11.71	−5.4209	−0.0057	0.89	1.00
Apr	14.10	−6.2638		0.89	1.00
May	17.88	−6.0909		0.88	1.04
Jun	22.02	−6.1843		0.86	1.18
Jul	25.21	−6.8277		0.83	1.30
Aug	24.97	−6.3076		0.87	1.16
Sep	22.08	−6.1143		0.88	1.07
Oct	17.48	−5.0825	−0.0099	0.89	1.02
Nov	11.84	−3.7142	−0.0176	0.82	1.20
Dec	8.53	−2.8485	−0.0238	0.77	1.39

Table 2. Coefficients of the regression equation for monthly precipitation (P): $P=C+A'x_1+D'x_2+E'x_3+F'x_4+G'x_5+H'x_6$, where x_1 is the altitude of the station (retrieved from the DEM), x_2 is the difference between the altitude of the station and the average altitude in a circle of a 10 km-radius around it, x_3 is the difference between the altitude of the station and the minimum altitude in a circle of a 15 km-radius around it, x_4 is the difference between the altitude of the station and the average altitude in a circle of 40 km-radius around it, x_5 is the eastern longitude of the station and x_6 is the northern latitude (minus 40 degrees). The adjusted coefficient of regression and the root mean square error are also displayed.

	C' mm	A' mm/m	D' mm/m	E' mm/m	F' mm/m	G' mm/deg	H' mm/deg	Adj. r^2	R.M.S.E. mm
Jan	21.79	−0.0241	−0.0682	0.0878		5.7213		0.67	9.89
Feb	18.37	−0.0162	−0.0574	0.0681		5.3990		0.68	8.54
Mar	24.40	−0.0152	−0.0465	0.0624		7.1333		0.65	8.10
Apr	29.02		−0.0525	0.0593		4.3991	5.9033	0.75	10.43
May	35.24	0.0556			−0.0280	7.2358		0.70	14.39
Jun.	20.62	0.0560			−0.0334	6.1848		0.78	12.56
Jul	−5.05	0.0513			−0.0337	9.0187		0.82	10.12
Aug	11.56	0.0496			−0.0349	15.1225	−0.8466	0.74	12.66
Sep	53.08	0.0398			−0.0290	10.2760	−13.5816	0.44	13.65
Oct	66.50	−0.0293	−0.0722	0.1004		16.5413	−19.5442	0.60	12.88
Nov	34.41	−0.0263	−0.0803	0.1113		9.6314		0.66	12.35
Dec	24.78	−0.0251	−0.0704	0.0946		6.1082		0.67	9.54

whose anomalies exceed a predefined threshold, which is specific for every variable. The second filter is a little more complex, since it detects groups of nearby stations with incompatible data. The presence of areas with sharp spatial variations of anomalies is supposed to be related with the existence of incompatible data.

To analyse the spatial variations of anomalies, they are first interpolated to 400-m spaced grid-points using an inverse-distance weighting method that avoids excessive smoothing. Then, the slopes in this grid are calculated and the second filter may detect the areas with the steepest slopes. The thresholds for suspicious values have been initially set to $0.7°\text{C km}^{-1}$ for thermal anomalies and $7\,\text{km}^{-1}$ for anomalies in precipitation, although they can be changed according to the particular data distribution.

Figure 2 illustrates the application of the method to monthly averaged temperatures recorded in June 2007. It shows a zoom into the areas where the slope in the distribution of thermal anomalies exceeds the above-defined threshold. Further investigation revealed the presence of a wrong value in both areas.

5 Conclusions

A reliable map of climatic reference values allows precise estimations of the represented climatic magnitudes in any area, even if there is not a meteorological station in it. It also may refer meteorological data gathered in a station to its corresponding climatic value, even if there is not a long and homogeneous time series of such data in the station.

It has been implemented a method to automatically detect possible errors in meteorological datasets. It performs very well for monthly precipitation in months with a smooth spatial distribution of anomalies. Sharper spatial distributions of anomalies (occurrence of heavy local showers) lead to higher false alarm rates. The method also presents a good performance for average monthly temperature. Nevertheless, it is expected to improve the method skill through its separated application to monthly averaged minimum and maximum daily temperatures.

Edited by: M. Dolinar
Reviewed by: L. de Salas, J. A. Fernandez,
                     and another anonymous referee

References

Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., and Timonin, V.: Neural network residual kriging application for climatic data, J. Geogr. Inform. Decis. Anal., 2, 215–232, 1998.

Dobesch, H., Dumolard, P., and Dyras, I. (Eds.): Spatial Interpolation for Climate data, Iste Publishing Co., 2007

Dyras, I., Dobesch, H., Grueter, E., Perdigao, A., and Tveito, O.: The use of Geographic Information Systems in climatology and meteorology: COST 719, Meteorol. Appl., 12, 1–5, 2005

Hocking, R. R.: The Analysis and Selection of Variables in Linear Regression, Biometrics, 32, 1–51, 1976.

Instituto Nacional de Meteorología: Valores normales de precipitación y temperatura de la Red Climatológica (1961–1990), Serie Monografías, INM, Madrid, 2000.

Journel, A. G. and Huijbregts, C. J.: Mining Geostatistics, Academic Press, London, 601 pp., 1978.

Lam, N. S.: Spatial interpolation methods: A review, Amer. Cart., 10, 129–149, 1993.

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