

## Data validation procedures in agricultural meteorology – a prerequisite for their use

J. Estévez<sup>1</sup>, P. Gavilán<sup>2</sup>, and A. P. García-Marín<sup>1</sup>

<sup>1</sup>University of Córdoba, Projects Engineering, Córdoba, Spain

<sup>2</sup>IFAPA Center “Alameda del Obispo”, Junta de Andalucía, Córdoba, Spain

Received: 16 December 2010 – Revised: 27 April 2011 – Accepted: 13 May 2011 – Published: 20 May 2011

**Abstract.** Quality meteorological data sources are critical to scientists, engineers, climate assessments and to make climate related decisions. Accurate quantification of reference evapotranspiration ( $ET_0$ ) in irrigated agriculture is crucial for optimizing crop production, planning and managing irrigation, and for using water resources efficiently. Validation of data insures that the information needed is been properly generated, identifies incorrect values and detects problems that require immediate maintenance attention. The Agroclimatic Information Network of Andalusia at present provides daily estimations of  $ET_0$  using meteorological information collected by nearly of one hundred automatic weather stations. It is currently used for technicians and farmers to generate irrigation schedules. Data validation is essential in this context and then, diverse quality control procedures have been applied for each station. Daily average of several meteorological variables were analysed (air temperature, relative humidity and rainfall). The main objective of this study was to develop a quality control system for daily meteorological data which could be applied on any platform and using open source code. Each procedure will either accept the datum as being true or reject the datum and label it as an outlier. The number of outliers for each variable is related to a dynamic range used on each test. Finally, geographical distribution of the outliers was analysed. The study underscores the fact that it is necessary to use different ranges for each station, variable and test to keep the rate of error uniform across the region.

### 1 Introduction

Meteorological information is one of the most important tools used by agriculture producers in decision making (Weiss and Robb, 1986). Some of the applications for these climate data include: crop water-use estimates, irrigation scheduling, integrated pest management, crop and soil moisture modeling, design and management of irrigation and drainage system and frost and freeze warnings and forecasts (Meyer and Hubbard, 1992).

Andalusia is located in the south of the Iberian Peninsula. This region is situated between the meridians  $1^\circ$  and  $7^\circ$  W and the parallels  $37^\circ$  and  $39^\circ$  N, with an extension around 9 Mha. The climate is semiarid, typically Mediterranean, with very hot and dry summers. In Andalusia 900 000 ha are irrigated (around 20 % of the cultivated area) under very different conditions (Gavilán et al., 2006).

The Agroclimatic Information Network of Andalusia (RIAA in Spanish) was deployed to provide coverage to most of the irrigated areas of the region and to improve irrigation water management (De Haro et al., 2003). Its exploitation and maintenance are carried out by the IFAPA (Agricultural Research Institute of Regional Government of Andalusia). This network provides at present daily estimations of reference evapotranspiration ( $ET_0$ ) using meteorological information collected by nearly one hundred automatic weather stations (Gavilán et al., 2008). This information is easily accessible due to it is published in the Web: <http://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/>.

Meteorological data validation is very important for hydrological designs and agricultural decision makings, concretely to estimate irrigation schedules. The quality control system discussed herein was applied to 85 stations, summarized in Table 1. The rest of the stations have been recently installed and their data series were too short. Quality control system consists of procedures or tests against which data are tested, setting data flags to provide guidance to end users. These flags give information about which tests have been applied satisfactorily or not to meteorological data.



Correspondence to: J. Estévez  
(jestevez@uco.es)

**Table 1.** Summary of automated weather stations used in the study.

Stations (Province)	Elevation (m)	Latitude (°)	Longitude (°)
Basurta-Jerez (CÁDIZ)	60	36.75	-6.01
Jerez Frontera (CÁDIZ)	32	36.64	-6.01
Villamartín (CÁDIZ)	171	36.84	-5.62
Conil Frontera (CÁDIZ)	26	36.33	-6.13
Vejer Frontera (CÁDIZ)	24	36.28	-5.83
Jimena Frontera (CÁDIZ)	53	36.41	-5.38
Puerto Sta. María (CÁDIZ)	20	36.61	-6.15
La Mojenera (ALMERÍA)	142	36.78	-2.70
Almería (ALMERÍA)	22	36.83	-2.40
Tabernas (ALMERÍA)	435	37.09	-2.30
Fiñana (ALMERÍA)	971	37.15	-2.83
V. Fátima-Cuevas (ALMERÍA)	185	37.39	-1.76
Huércal-Overa (ALMERÍA)	317	37.41	-1.88
Cuevas Almanz. (ALMERÍA)	20	37.25	-1.79
Adra (ALMERÍA)	42	36.74	-2.99
Níjar (ALMERÍA)	182	36.95	-2.15
Tíjola (ALMERÍA)	796	37.37	-2.45
Bélmex (CÓRDOBA)	523	38.25	-5.20
Adamuz (CÓRDOBA)	90	37.99	-4.44
Palma del Río (CÓRDOBA)	134	37.67	-5.24
Hornachuelos (CÓRDOBA)	157	37.72	-5.15
El Carpio (CÓRDOBA)	165	37.91	-4.50
Córdoba (CÓRDOBA)	117	37.86	-4.80
Santaella (CÓRDOBA)	207	37.52	-4.88
Baena (CÓRDOBA)	334	37.69	-4.30
Baza (GRANADA)	814	37.56	-2.76
Puebla D.Fadriq. (GRANADA)	1110	37.87	-2.38
Loja (GRANADA)	487	37.17	-4.13
Pinos Puente (GRANADA)	594	37.26	-3.77
Iznalloz (GRANADA)	935	37.41	-3.55
Jerez Marques. (GRANADA)	1212	37.19	-3.14
Cádiar (GRANADA)	950	36.92	-3.18
Zafarraya (GRANADA)	905	36.99	-4.15
Almuñécar (GRANADA)	49	36.74	-3.67
Padul (GRANADA)	781	37.02	-3.59
Tojalillo-Gibraleón (HUELVA)	52	37.31	-7.02
Lepe (HUELVA)	74	37.24	-7.24
Gibraleón (HUELVA)	169	37.41	-7.05
Moguer (HUELVA)	87	37.14	-6.79
Niebla (HUELVA)	52	37.34	-6.73
Aroche (HUELVA)	299	37.95	-6.94
Puebla Guzmán (HUELVA)	288	37.55	-7.24
El Campillo (HUELVA)	406	37.66	-6.59
Palma Condado (HUELVA)	192	37.36	-6.54
Almonte (HUELVA)	18	37.15	-6.47
Moguer-Cebollar (HUELVA)	63	37.24	-6.80
Huesa (JAÉN)	793	37.74	-3.06
Pozo Alcón (JAÉN)	893	37.67	-2.92
S.José Propios (JAÉN)	509	37.85	-3.22
Sabiote (JAÉN)	822	38.08	-3.23
Torreblascopedro (JAÉN)	291	37.98	-3.68
Alcaudete (JAÉN)	645	37.57	-4.07
Mancha Real (JAÉN)	436	37.91	-3.59
Úbeda (JAÉN)	358	37.94	-3.29
Linares (JAÉN)	443	38.06	-3.64
Marmolejo (JAÉN)	208	38.05	-4.12
Chiclana Segura (JAÉN)	510	38.30	-2.95
Higuera Arjona (JAÉN)	267	37.95	-4.00

**Table 1.** Continued.

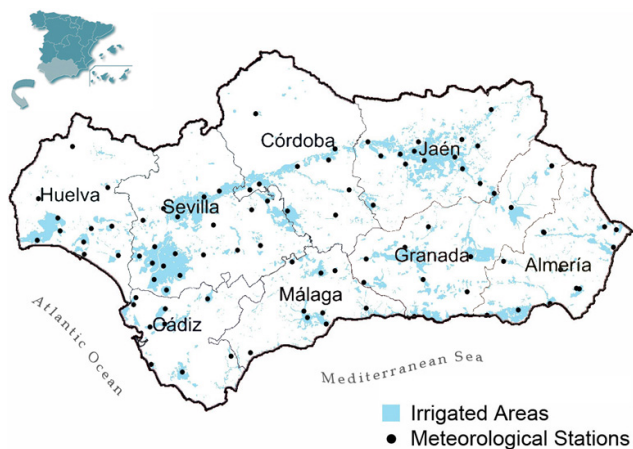
Stations (Province)	Elevation (m)	Latitude (°)	Longitude (°)
Santo Tomé (JAÉN)	571	38.03	-3.08
Jaén (JAÉN)	299	37.89	-3.77
Palacios-Villafran. (SEVILLA)	21	37.18	-5.93
Cabezas S. Juan (SEVILLA)	25	37.01	-5.88
Lebrija 2 (SEVILLA)	40	36.90	-6.00
Aznalcázar (SEVILLA)	4	37.15	-6.27
Puebla del Río II (SEVILLA)	41	37.08	-6.04
Écija (SEVILLA)	125	37.59	-5.07
La Luisiana (SEVILLA)	188	37.52	-5.22
Osuna (SEVILLA)	214	37.25	-5.13
La Rinconada (SEVILLA)	37	37.45	-5.92
Sanlúcar la Mayor (SEVILLA)	88	37.42	-6.25
Villan.Río-Minas (SEVILLA)	38	37.61	-5.68
Lora del Río (SEVILLA)	68	37.66	-5.53
Los Molares (SEVILLA)	90	37.17	-5.67
Guillena (SEVILLA)	191	37.51	-6.06
Puebla Cazalla (SEVILLA)	229	37.21	-5.34
Carmona-Tomejil (SEVILLA)	79	37.40	-5.58
Málaga (MÁLAGA)	68	36.75	-4.53
Vélez-Málaga (MÁLAGA)	49	36.79	-4.13
Antequera (MÁLAGA)	457	37.05	-4.55
Estepona (MÁLAGA)	199	36.44	-5.20
Archidona (MÁLAGA)	516	37.07	-4.42
Sierra Yeguas (MÁLAGA)	464	37.13	-4.83
Churriana (MÁLAGA)	32	36.67	-4.50
Pizarra (MÁLAGA)	84	36.76	-4.71
Cártama (MÁLAGA)	95	36.71	-4.67

## 2 Materials and methods

### 2.1 Source of data

The dataset used in the present study was obtained from the daily database of the RIAA and it was from 2004 to 2009. Each station is controlled by a CR10X datalogger (Campbell Scientific) and is equipped with sensors to measure air temperature and relative humidity (HMP45C probe, Vaisala), solar radiation (pyranometer SP1110 Skye), wind speed and direction (wind monitor RM Young 05103) and rainfall (tipping bucket rain gauge ARG 100). Air temperature and relative humidity are measured at 1.5 m and wind speed at 2 m above soil surface. Data from stations are transferred to the data-collecting seat (Main Center) by using GSM modems. This information is saved in a database. The Main Center is responsible for quality control procedures that comprise the routine maintenance program of the network, including sensor calibration and data validation.

Accuracy of  $ET_0$  calculations depends on the quality and the integrity of meteorological data used (Allen, 1996), being necessary data quality control application. Different procedures for quality assurance have been described by Meek and Hatfield (1994), Allen (1996), Shafer et al. (2000) and Feng et al. (2004). These tests are based on some rules proposed



**Figure 1.** Agroclimatic Information Network of Andalusia (85 meteorological stations).

by O’Brien and Keefer (1985). However, the tests applied in this study are based on statistical decisions and they were conducted for 84 stations (Fig. 1), using data only from a single site. Three procedures were tuned to the prevailing climate: seasonal thresholds, seasonal rate of change and seasonal persistence (Hubbard et al., 2005). These tests are related to station climatology at the monthly level, using dynamic limits for each variable. The tests were applied to the following variables: maximum, minimum and mean air temperature (Tx, Tn, Tm), maximum, minimum and mean relative humidity (RHx, RHn, RHm), and precipitation (Preci).

### 2.2 Theory

The THRESHOLD test is a quality control approach that checks whether the variable  $x$  falls in a specific range for the month in question. The equation is

$$\bar{x} - f\sigma_x \leq x \leq \bar{x} + f\sigma_x \tag{1}$$

where  $\bar{x}$  is the daily mean (e.g., mean of maximum daily temperature for December) and  $\sigma_x$  is the standard deviation of the daily values for the month in question. This relationship indicates that with larger values of  $f$ , the number of potential outliers decreases.

The STEP CHANGE test compares the change between successive observations. This test checks if the difference value of the variable falls inside the climatologically expected lower and upper limits on daily rate of change for the month in question. The step change test for variable  $x$  is given in Eq. (2):

$$\bar{d}_i - f\sigma_{d_i} \leq d_i \leq \bar{d}_i + f\sigma_{d_i} \tag{2}$$

where  $d_i = x_i - x_{i-1}$ ,  $i$  is the day and  $\sigma_{d_i}$  is the standard deviation of  $d_i$ .

The PERSISTENCE test checks the variability of the measurements. When the variability is too high or too low, the

data should be flagged for further checking. If the sensor fails it will often report a constant value and the standard deviation ( $\sigma$ ) will become smaller. When the sensor is out for an entire period,  $\sigma$  will be zero. If the instrument works intermittently and produces reasonable values interspersed with zero values, thereby greatly increasing the variability for the period. This test compares the standard deviation for the time period being tested to the limits expected as follows:

$$\bar{\sigma}_j - f\sigma_{\sigma_j} \leq \sigma_j \leq \bar{\sigma}_j + f\sigma_{\sigma_j} \tag{3}$$

where  $\sigma_j$  is the standard deviation from daily values for each month ( $j$ ) and year and  $\sigma_{\sigma_j}$  is the standard deviation of  $\sigma_j$  for the month in question.

When the datum is valid and is rejected by the tests, a Type I error is committed. If the datum is not valid but it is accepted by the quality control procedures, a Type II error is committed. The results discussed in this paper only show the potential outliers of Type I error.

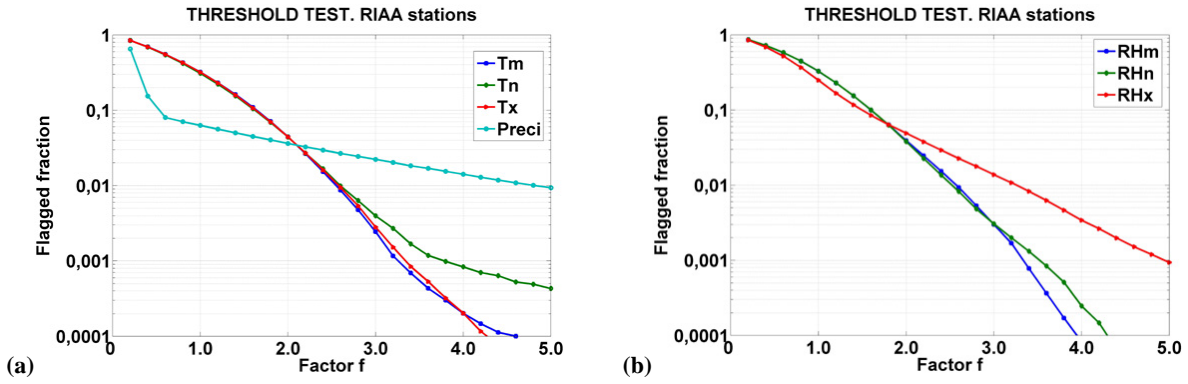
This system was developed in open source code, using GNU GPL (General Public License) support and it can be installed on any platform: Linux, Windows, Unix, Mac OS, Solaris, etc. PostgreSQL, PostGIS and PLpgSQL are the selected free technologies under the quality procedures were developed.

PostgreSQL is an object-relational database management system (ORDBMS) based on POSTGRES version 4.2, developed at the University of California at the Berkeley Computer Science Department (Stonebraker and Kemnitz, 1991). It supports a large part of the SQL standard and offers many modern features: complex queries, foreign keys, triggers, views, functions, procedures languages, etc. PostGIS is an extension to PostgreSQL which allows GIS (Geographic Information Systems) objects to be stored in the database. It includes support for a range important GIS functionality, including full OpenGIS support, advanced topological constructs (coverages, surfaces, networks), desktop user interface tools for viewing and editing GIS data, and web-based access tools. Finally, PLpgSQL is a powerful procedure language used to specify a sequence of steps that are followed to procedure an intended programmatic result. The use of SQL within PLpgSQL increases the power, flexibility, and performance of the quality tests. The most important aspect of using this language is its portability. Its functions are compatible with all the platforms that can operate de PostgreSQL database system.

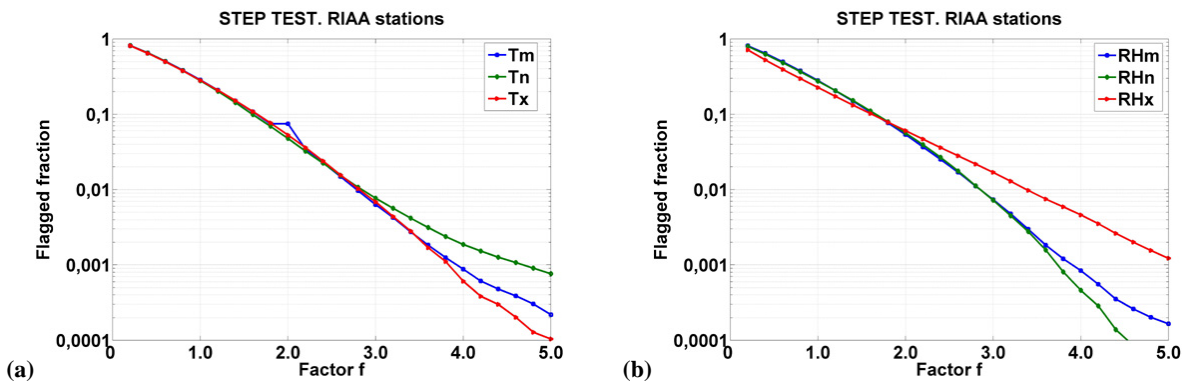
These three tests were applied to data from selected stations, following Eqs. (1), (2) and (3).

### 3 Results and discussion

The next figures show the number of potential Type I errors that would occur when using the specified tests with various  $f$  factors. The fraction data flagged is represented on a log scale and related to the all the network tested (85 stations).



**Figure 2.** (a) Threshold Test – Maximum (Tx), minimum (Tn) and mean temperature (Tm) and Precipitation (Preci). (b) Threshold Test – Maximum (RHx), minimum (RHn) and mean relative humidity (RHm).



**Figure 3.** (a) Step Test – Maximum (Tx), minimum (Tn) and mean temperature (Tm). (b) Step Test – Maximum (RHx), minimum (RHn) and mean relative humidity (RHm).

The general shape of the relationship between  $f$  and the fraction of data flagged is shown in Figs. 2, 3 and 4. The results obtained in this work are similar to the results of Hubbard et al. (2005). The results for the threshold analysis indicate that approximately 2 % of the data would be flagged for maximum, minimum and mean temperature if an  $f$  value of 2.3 is used. For precipitation, 2 % of the data were flagged in this test for an  $f$  value of 3.1. These results are shown in Fig. 2a. The results on Fig. 2b show the same fraction data flagged for minimum and mean relative humidity when  $f$  value of 2.2 is used. In this figure and for maximum relative humidity, this percentage of data would be flagged with an  $f$  value of 2.7. Similar figures are shown for the step change test (Fig. 3a and b) and the persistence test (Fig. 4a and b). The results for the persistence analysis indicate that approximately 1 % of the data would be flagged for all the variables if an  $f$  value less than 2.0 is used. This is consequence of the need for longer series of data to calculate the variability from daily values for each month and year. For precipitation, the step test was not applied because of the discontinuous nature of rainfall. These results are related to the three tests applied to 85 automatic weather stations of the RIAA. It is impor-

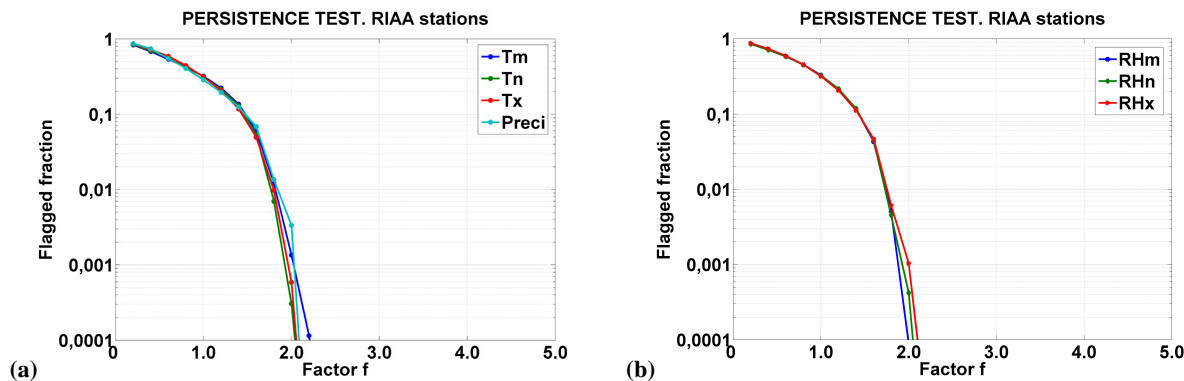
tant to remark that the fraction flagged for each  $f$  value was different for each station. These results show that it will be possible to select dynamic  $f$  values for each station and temporal scale and to fix a specific rate of Type I errors across the region.

The spatial distribution of the fraction data flagged for an  $f$  value of 3 in threshold and step tests was estimated using GIS techniques for all the variables. This analysis is very useful to visually study the distribution of outliers across the region. The results for threshold test using ordinary krigging interpolation for maximum temperature are shown in Fig. 5. This map shows that the fraction data flagged is higher in coastal weather stations than in inland locations. This is caused by the different climate regime between them. The maximum temperatures are lower in locations near the coast than in inland locations where the air masses are not influenced by a nearby and large water body (Mediterranean Sea or Atlantic Ocean).

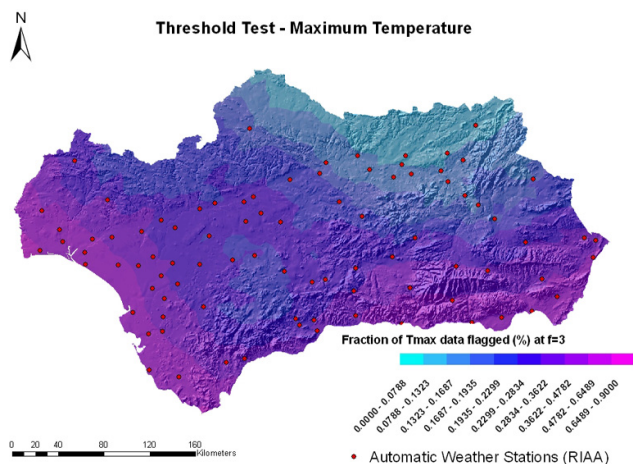
The quality control system can dynamically generate this type of maps using any GIS software at any time.

Sometimes, for scientific or other purposes we cannot reject too much data. It can be very useful to fix a rate of





**Figure 4.** (a) Persistence Test – Maximum (Tx), minimum (Tn) and mean temperature (Tm) and Precipitation (Preci). (b) Persistence Test – Maximum (RHx), minimum (RHn) and mean relative humidity (RHm).



**Figure 5.** Fraction of maximum temperature data flagged at  $f = 3$  for threshold test.

potential outliers for not considering them in our model or study. For fixing a specific rate of fraction flagged in this example of maximum temperature (Tx), we should use different  $f$  values for each station. As it can be seen in Fig. 5, using  $f = 3$ , the fraction of Tx data flagged ranged from nearly 0 (station located at northeast of Jaén) to 0.6–0.9 approximately (coastal stations) across Andalusia region.

These automated validation procedures should be accompanied by other tasks such as: field visits for maintenance routines, sensors calibration and manual inspection (Feng et al., 2004; Shafer et al., 2000). This manual inspection is crucial and necessary for ensuring an appropriate flagging process, providing human judgment to it, catching subtle errors that automated techniques may miss (Shafer et al., 2000).

#### 4 Summary and conclusions

In this study, the validation tests applied to daily climatic data from 85 automatic weather stations varied modestly with climate type and significantly with the variable tested. It is essential to test the capability of validation procedures because of quality control is a major prerequisite for using meteorological information. Several tests based on statistical decisions have been applied to meteorological data from the Agroclimatic Information network of Andalusia (RIAA). The validated variables were maximum, minimum and mean air temperature (Tx, Tn, Tm), maximum, minimum and mean relative humidity (RHx, RHn, RHm) and precipitation (Preci). Although daily precipitation is known to follow a gamma distribution, it was included in these tests to give a reference point. Results obtained from running the quality control procedures showed a high variability when different  $f$  values are used. It is essential to test the capability of these tests to produce flags if data are out of range or are internally or temporally inconsistent.

The use of open source code and General Public License technologies (GNU GPL) to develop the procedures allows any meteorological network to implement a similar system with zero cost. All the functions and algorithms can be read and rewritten or adapted for future users.


The possibility of dynamically mapping the percentage of errors for any variable is a powerful tool to visually study the spatial distribution of the fraction data flagged. These results show that it necessary to select dynamic  $f$  values for each station and test to preselect a fixed rate of error detection across the Andalusia region.

This quality control system can easily be used with any conventional GIS software. The treatment of the meteorological data like geographical variables using GIS techniques can be very useful for maintenance routines and sensors calibration.

Future works of the authors should include spatial consistency procedures and to introduce seeded random errors to examine the Type II errors detection.

Edited by: B. Lalic

Reviewed by: V. Vucetic and two other anonymous referees

**sc | nat**  The publication of this article is sponsored by the Swiss Academy of Sciences.

## References

- Allen, R. G.: Assessing integrity of weather data for reference evapotranspiration estimation, *J. Irrig. Drain. Eng.*, 122(2), 97–106, 1996.
- De Haro, J. M., Gavilán, P., and Fernández, R.: The Agroclimatic Information Network of Andalusia, Proceeding of the Third International Conference on Experiences with Automatic Weather Stations, Torremolinos, Spain, 19–21 February, 1–12, 2003.
- Feng, S., Hu, Q., and Qian, Q.: Quality control of daily meteorological data in China, 1951–2000: a new dataset, *Int. J. Climatol.*, 24, 853–870, 2004.
- Gavilán, P., Lorite, I. J., Tornero, S., and Berengena, J.: Regional calibration of Hargreaves equation for estimating reference ET in a semiarid environment, *Agric. Water Manag.*, 81, 257–281, 2006.
- Gavilán, P., Estévez J., and Berengena, J.: Comparison of standardized reference evapotranspiration equations in southern Spain, *J. Irrig. Drain. Eng. ASCE*, 134(1), 1–12, 2008.
- Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., and Osugi, T. T.: Performance of quality assurance procedures for an applied climate information system, *J. Atmos. Oceanic Technol.*, 22, 105–112, 2005.
- Meek, D. W. and Hatfield, J. L.: Data quality checking for single station meteorological databases, *Agric. For. Meteorol.*, 69, 85–109, 1994.
- Meyer, S. J. and Hubbard, K. G.: Nonfederal automated weather stations and networks in the United States and Canada: a preliminary survey, *B. Am. Meteorol. Soc.*, 73(4), 449–457, 1992.
- O'Brien, K. J. and Keefer, T. N.: Real-time data verification, Proc. ASCE Special Conf., Buffalo, NY, American Society of Civil Engineers, 764–770, 1985.
- PostGIS: <http://postgis.refractions.net> (last access: 5 December 2009), 2009.
- PostgreSQL: <http://www.postgresql.org> (last access: 5 December 2009), 2009.
- Shafer, M. A., Fiebrich, C. A., Arndt, D. S., Fredrickson, S. E., and Hughes, T. W.: Quality assurance procedures in the Oklahoma Mesonet, *J. Atmos. Oceanic Technol.*, 17, 474–494, 2000.
- Stonebraker, M. and Kemnitz, G.: The Postgres next-generation database-management system, *Communicat. ACM.*, 34, 78–92, 1991.
- Weiss, A. and Robb, J. G.: Results and interpretations from a survey on agriculturally related weather information, *B. Am. Meteorol. Soc.*, 67(1), 10–15, 1986.