Statistical Control Charts for Quality Control of Weather Data for Reference Evapotranspiration Estimation

by Simon, O. Eching
Office of Water Use Efficiency
CA Department of Water Resources
P.O. Box 94236-0001
Sacramento, CA 94236-0001

and Richard, L. Snyder
Dept. of Land Air and Water Resources
U.C. Davis
University of California, Davis
Davis, CA 95616

Keywords: Time variant control chart, evapotranspiration, CIMIS, quality control, normality, control limit.

Abstract

Data quality control is a necessary component of any weather station network used for estimating reference evapotranspiration (ETo). The absence of a quality control program can result in poor quality ETo data that severely limits it usefulness for irrigation scheduling. Statistical quality control criteria are developed for assessing quality and reasonableness of hourly and daily weather data for the California Irrigation Management Information System (CIMIS) weather stations. The quality control criteria, based on means ($\bar{x}$) and standard deviations (σ), are developed from historical CIMIS weather station data. Two statistical quality control limits, 3σ and 2σ upper control limit and lower control limit, are developed. The two control limits are integrated into existing data screening rules forming new CIMIS data quality control criteria. A new version of a control chart, time variant control chart is introduced. Statistical control charts have been widely used in the manufacturing industry for process mean or variability monitoring and quality control. Control limits developed herein are similar to those used in the manufacture of products. Unlike in manufacturing where one seeks to attain a state of statistical control, these control limits are used to identify data that fall outside the control limits. Such data are then flagged with a quality control flag.

INTRODUCTION

Recent improvements in automated weather station sensors and reference evapotranspiration (ETo) estimation techniques has made real time or near real time (ETo) readily available, which allows farmers to adopt ETo based water budget irrigation scheduling techniques. The usefulness of ETo data, however, is dependent upon the quality of data used to estimate the ETo-solar radiation, air temperature, wind speed, and vapor pressure. Statistical quality control lends itself as a convenient means to screen some of these data.

Shewhart is credited for being the first to apply statistical methods to quality control. In 1924, he proposed the concept of a control chart (Montgomery, 2000). A control chart shows the value of the quality characteristics of interest as a function of time or sample number. Generally, a control chart is made of a centerline which represents the mean value for the in-control process, and two horizontal lines, the upper control limit (UCL) and the lower control limit (LCL) as shown in figure 1. In 1931, Shewhart published "Economic Control of Quality of Manufactured Product," a book that outlines statistical methods for use in production and control charts methods. A summary of the historical background of statistical quality control is found in "Quality Control and Industrial

Statistics" (Duncan, 1986).

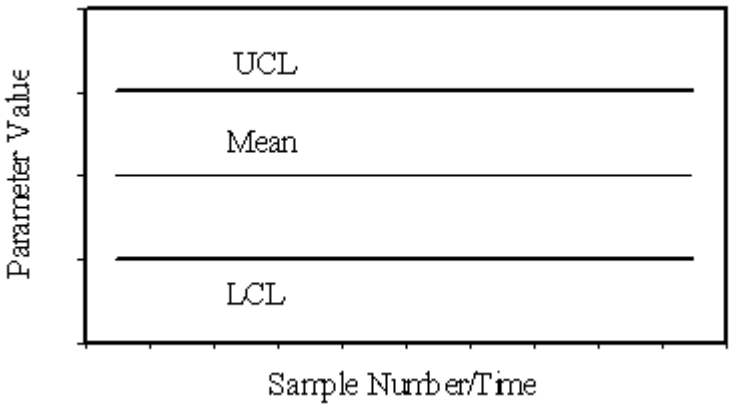

Fig. 1 - Shewhart Quality Control Chart.

The objective of this research is to develop quality control criteria for assessing quality and reasonableness of hourly and daily air temperature, wind speed, and vapor pressure data collected at the California Irrigation Management Information System (CIMIS) weather stations. In addition to data used for calculating ETo, CIMIS collects other data as well, therefore, this study includes data other than those used for calculating ETo. The quality control approach used herein is different than traditional quality control, where statistical quality control is used to ensure the production of a sequence of objects having quality characteristics that are within some specified limits-that is, to attain a state of statistical control. Control limits in this research are used to develop a set of quality control criteria for the purpose of identifying data that fall outside the control limits. The control limits are set in terms of the statistical parameter's mean and standard deviation as suggested by Snyder and Pruitt (1992). Although solar radiation is a key data in the estimation of ETo, the technique is not suited for screening solar radiation data.

MATERIALS AND METHODS

The Davis CIMIS weather station was selected for this study because of its relatively long period of data (since 1982). The following hourly data were used: air and soil temperature, wind speed, and vapor pressure. Daily data used were: average wind speed; maximum, minimum, and average air temperature, soil temperature, and vapor pressure. Not all data values were used. Data were screened based on previous CIMIS quality control criteria.

Statistical quality control generally requires the design of Shewhart control charts. Designing Shewhart control charts supposes that the probability density function (pdf) of the variable to be tested follows the normal distribution or is approximately normal (Castagliola, 2000). Moreover, whenever means and standard deviations are calculated, it is assumed that the distribution of the underlying population is known. Very often it is assumed a priori that environmental data or the logarithm of the data follows the normal distribution. The logic behind such an assumption, at least in cases where data are averages of smaller time steps, is that the central limit theory applies. In this research,

data are qualitatively tested for normality. As occurs in most statistical applications, it is impossible or impractical to observe the entire population. The concept of statistical inference was used; the distribution of the population was inferred from the distribution of samples used to establish the quality control criteria.

For each hourly variable, a graphical test for normality was performed for selected hours between 0100 and 2400 for the months of February, April, June, August, October, and December, for the period 1989 through 1999. For daily weather variables, out of brevity, the central limit theory is evoked and the data are assumed to follow a normal distribution. Daily data used were also for the period 1989 through 1999.

The normal probability plot (Chambers 1983) technique was used for assessing whether a particular data variable is approximately normally distributed. The procedure is outlined below:

1. Sample data are ordered from smallest to largest.
2. Normal order statistic medians are calculated.
3. Ordered sample data are plotted as a function of the corresponding normal ordered statistic medians, which represent a theoretical normal distribution.
4. Normally distributed data should form an approximately straight line. Departure from a straight line indicates departure from normality.

The normal ordered statistic medians are defined as:

$$N(i) = G(U(i)) \qquad (1)$$

where

G = percent point function of the normal distribution, the value of variable x corresponding to a given cumulative distribution function.

U(i) = the uniform order statistics medians or plotting position.

The uniform order statistics medians are defined as:

$$U(i) = 1 - U(n) \qquad i = 1$$

$$U(i) = 0.5^{\frac{1}{n}} \qquad i = n \qquad (2)$$

$$U(i) = \left(\frac{i - 0.3175}{n + 0.365}\right) \qquad elsewhere$$

where

i = rank of the ordered sample data

n = number of the sample data

Based on the test for normality, monthly means and standard deviations were calculated separately for hours 1, 2, 3, . . ., 24 for each weather station variable that was determined to follow a normal distribution. Monthly means and standard deviations were also calculated for daily data for each weather station variable using archived data. Unlike

the test for normality where only selected data were used, the hourly and daily means and standard deviations were calculated from the entire data for the period 1989 through 1999. All calculations were performed using Structured Query Language (SQL) on the CIMIS Oracle database.

The concept of quality control limits was then used to develop new data screening rules. That is, for each weather variable, the UCL and the LCL are calculated as follows:

$$UCL_t = \bar{x}_t + ks_t$$

(3)

$$LCL_t = \bar{x}_t - ks_t$$

where

$\bar{x}$ = sample mean

k = an integer multiplier

s = sample standard deviation

t = a given month

Two screening rules were developed based on the value of k as follows:

Flag	Description
R	Data value > 3 σ from $\bar{X}$ (k=3)
Y	Data value > 3 σ from $\bar{X}$ (k=2)

To facilitate graphical screening of data quality via a control chart, we introduce a new version of a control chart- Time Variant (TV) control chart- it accommodates the change in the value of the centerline and the control limits with time.

RESULTS AND DISCUSSION

The April normal probability plot for selected Davis hourly air temperature are shown in Figure 2. Overall, the plots show close fit to a linear pattern, which indicate the data follow a normal distribution. The excellent straight-line pattern is verified by excellent correlation as shown by values of R^2 greater than 0.95. Air temperature plots for other months show similar linear patterns with the exception of the August plot. The August probability plot shows marked departure from a linear pattern at the lower and upper range of the data, which may indicate that a distribution other than the normal distribution would better fit these data. Probability plots for vapor pressure and wind speed are not presented; however, they are described below.

Most of the hourly vapor pressure normal probability plots also show a strong linear pattern with most R^2 values greater than 0.97. Similar to the air temperature plots, the August plots show the greatest departure from the fitted line, particularly the first and the last few points.

The number of plots that show a strong linear pattern is sufficiently high to make an assumption that these data follow a normal distribution and thus allow the calculation of

mean and standard deviation based on that assumption. Hourly wind speed plots for several months show very strong non-linear patterns, an indication that the normal distribution is not a good model for hourly wind speed data. No attempt was made to determine the type of distribution that best fits hourly wind speed data. In other words, the probability density function is unknown. Were the distribution known, hourly wind speed statistical quality control based on such a distribution would have been developed.

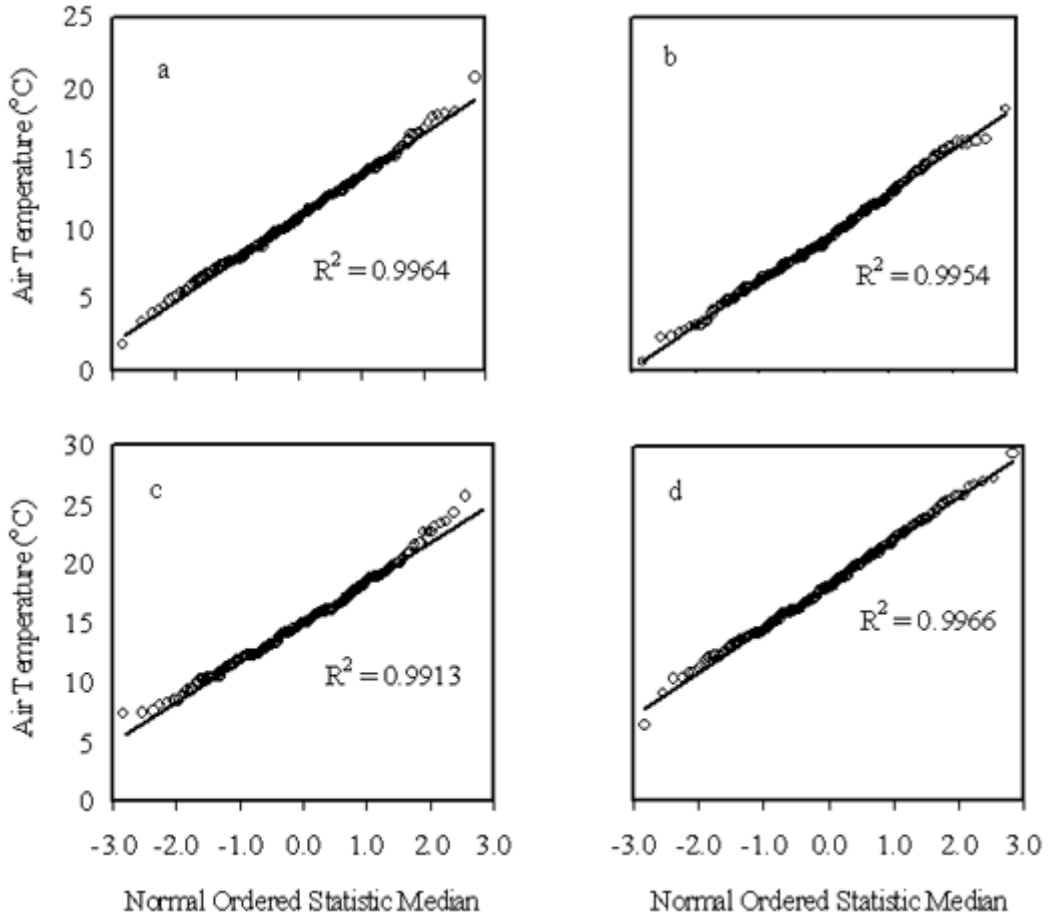

Figure 2 - Normal probability plot of April hourly air temperature from Davis for hours (a) 1:00 am (d) 5:00 am (c) 9:00 am (d) 11:00 am.

The new screening rules or statistical quality control limits, as previously explained, are developed from the mean and standard deviation. For example, the January 1:00 am 3σ UCL, 3σ LCL, 2σ UCL, and 3σ LCL for hourly air temperature are: 16.99 °C, –5.26 °C, 13.28 °C, and –1.55 °C respectively. During data quality screening, January 1:00 a.m. air temperature data that fall outside these limits are assigned specific quality control flags to alert users of possible data problems. That is, data would be flagged R if the value is greater than 16.99 °C or less than –5.26 °C; it would be flagged Y if greater than 13.28 °C or less than –1.55 °C. Some hourly vapor pressure 3σ LCL values are contrary to theoretical values; they are less than zero. We, therefore, felt that hourly vapor pressure data should not be tested based on statistical control limits.

Screening rules and data flagging for daily data are similar to that for hourly data; they are based on means and standard deviations with 3σ and 2σ control limits. A sample of

the criteria for daily minimum air temperature is presented in Table 1. Like hourly data, these daily data flagging criteria are part of an extensive quality control criteria. Control limits can be cast in terms of a percentage confidence interval (CI) as well. For example, a 96% confidence interval can be used instead of the 2σ limit.

Table 1. Minimum Daily Air temperature (°C) Statistical Quality Control Parameters for the Davis CIMIS station

		Standard	+ 3 x σ	- 3 x σ	+ 2 x σ	- 2 x σ
Month	Mean	Deviation	UCL	LCL	UCL	LCL
Jan	3.45	3.76	14.72	-7.82	10.96	-4.06
Feb	4.75	3.67	15.75	-6.26	12.08	-2.59
Mar	5.94	3.09	15.21	-3.33	12.12	-0.24
April	7.82	3.02	16.88	-1.23	13.86	1.79
May	10.16	2.86	18.74	1.59	15.88	4.45
Jun	12.63	2.56	20.30	4.95	17.74	7.51
July	13.95	2.36	21.03	6.87	18.67	9.23
Aug	13.45	2.36	20.54	6.36	18.18	8.73
Sep	12.29	2.60	20.09	4.48	17.49	7.08
Oct	9.46	3.06	18.63	0.29	15.58	3.35
Nov	5.57	3.24	15.28	-4.15	12.04	-0.91
Dec	2.61	3.60	13.41	-8.19	9.81	-4.59

Although exact numerical values are convenient for computer programs that automatically scan data for conformance with control limits, graphical presentation is preferable for day-to-day human monitoring of data quality. Data that fall outside the control limits can easily be identified. Another advantage of the TV control chart is that an entire year of data can be scanned quickly for conformance with the quality control limits. Fig. 3 are TV Control Charts for hourly air temperature (Fig 3a), and daily maximum air temperature (Fig. 3b). The hourly TV chart as presented in Fig. 3a is constructed from the entire 24 hour data for each month. If the chart were constructed for each individual hour, it would have taken the stair case appearance of the daily TV chart shown in Fig 3b.

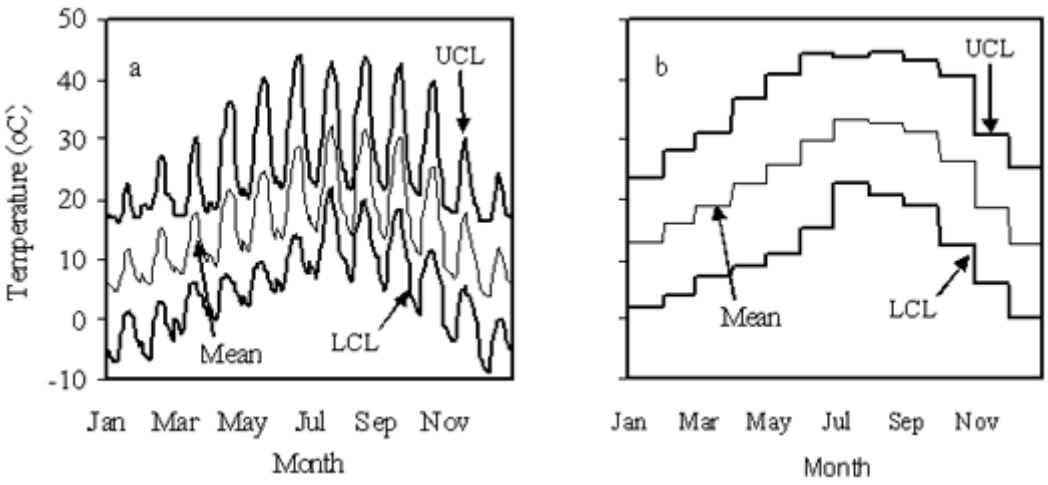

Fig. 3 - Davis 3σ time variant control charts (a) hourly air temperature (b) daily maximum air temperature.

Data quality testing using TV Control Charts is demonstrated in Fig. 4. Hourly air temperature for the 26th day of each month in 2002 is screened using a 2σ control limit in Fig. 4a. It clearly shows that during March, May, June, July, August, September, and October, there were data on or outside the control limit. Use of the TV control chart to screen daily data is illustrated in Fig. 4b. Similar to the hourly air temperature screening (Fig. 4a), the entire 2002 daily average air temperature data quality is tested. In this case, it shows that only a few of the daily average air temperature data were outside of the 2σ control limits.

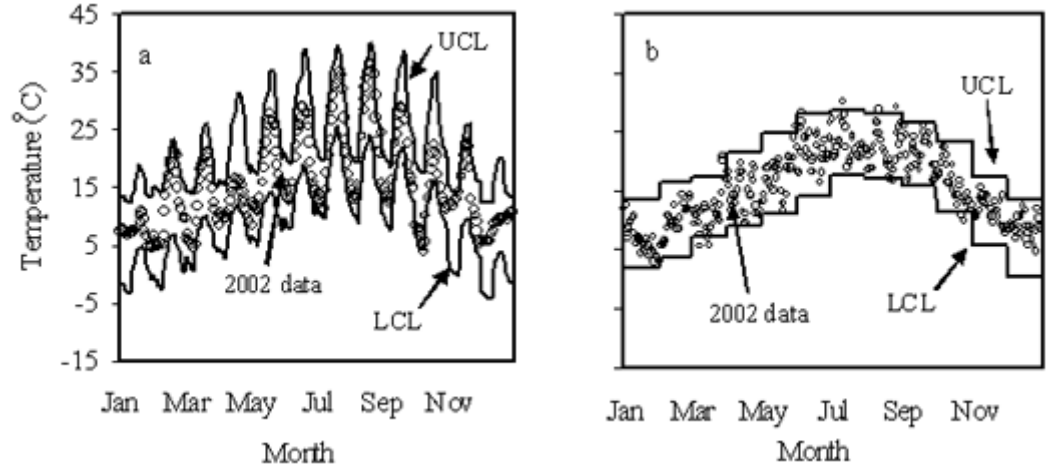

Fig. 4 - Davis 2002 data 2σ quality control (a) hourly air temperature for the 26th day of the month (b) daily average air temperature.

The fact that data are flagged R or Y by itself does not necessarily mean the data are bad. The purpose of the R and Y quality control flags is to inform users of data credibility as related to the set of quality control limits. Users can then evaluate the data based on their experience and decide whether or not they should discard it. Persistently flagged data, however, could indicate potential problems with a sensor.

CONCLUSION

Quality control criteria are developed based on means and standard deviation. The criteria consists of two control limits 3σ and the 2σ limits. Control limits developed herein are similar to those used in the manufacture of products. Unlike in manufacturing where one seeks to attain a state of statistical control, these control limits are used to identify data that fall outside the control limits. Data that fall outside the 3σ control limit are flagged R and those that fall outside the 2σ but are within the 3σ limit are flagged Y. These two screening or flagging rules are integrated with previous CIMIS quality control criteria to form new CIMIS quality control criteria. Because the statistical quality control criteria capture statistical properties of data and change with time, they ensure better data quality integrity. Maintaining the statistical properties of data will allow data integrity problems to be identified more readily and will help to plan unscheduled maintenance visits.

Prior to calculating the means and standard deviations, probability plots were used to test whether hourly data follow a normal distribution. The normality test was carried out on hourly data. Test for normality results show that Davis hourly air temperature and vapor pressure data follow a normal distribution, but wind speed does not.

A new version of a Control Chart, named TV control chart, is introduced. As the name suggests, control limits for the TV chart vary with time. It is simply a graphical representation of the 3σ and the 2σ control limits. The main advantage of the TV control chart over the numerical control limits is that it provides an efficient means for inspecting an entire year of data. When current data are superimposed on the control chart, potential data quality problems can be easily identified.

Literature Cited

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. 1983. Graphical Methods for Data Analysis, Wadsworth.

Duncan, A.J. 1986. Quality Control and Industrial Statistics, Fifth Edition, Irwin, Homewood, Illinois.

Montgomery, D.C. 2000. Introduction to Statistical Quality Control, Fourth Edition, John Wiley and Sons Inc., New York.

NIST/SEMATECH, 2002. Engineering Statistics Internet Handbook, www.itl.nist.gov/div898/handbook.html

Shewhart, W.A. 1931. Economic Control of Quality of Manufactured Product, Van Nostrand, New York.

Shewhart, W.A. 1987. Statistical Method from the Viewpoint of Quality Control, Edited by Deming, W.E., Dover Publications, Inc., New York.

Snyder, R.L., and Pruitt, W.O. 1992. Evapotranspiration data management in California. Irrigation and Drainage Session Proceedings/Water Forum '92 EE,HY,IR WR Div/ASAE, Baltimore, Maryland, August 2-6, 1992.