ESL-HH-02-05-26

Restoration of Short Periods of Missing Energy Use and Weather Data Using Cubic Spline and Fourier Series Approaches: Qualitative Analysis

Juan Carlos Baltazar and David E. Claridge
Energy Systems Laboratory
Texas A&M University
College Station, TX. 77843-3581

ABSTRACT
The paper presents seventeen approaches that use cubic splines and Fourier series for restoring short term missing data in time series of building energy use and weather data. The study is based on twenty samples of hourly data, each at least one year long. In order to differentiate the approaches, two comparisons were carried out. The first comparison was made between the estimated and actual values, as time series and as cross check plots. The second comparison is based on the fraction of the total data that can be estimated by an approach within specific ranges or error. Thus for the cooling and heating data, the fraction of the data set estimated within 1%, 5%, and 10% of the measured values was determined. For the dew point and the dry-bulb temperature samples, the performance is based on the amount of data that are within 1, 3, 5 and 10 °F from the actual data. From the results of this analysis, it appears that linear interpolation is a better approach for filling gaps one to three hours long. The cubic splines approach gave better performance for gaps between four and six.

INTRODUCTION

The measurement of savings is an important element of successful energy conservation programs. The success of the savings determination process is heavily dependent on the quality of the data that is collected and available. However, the determination of retrofit savings in the presence of missing data or large amounts of bad data has received only limited attention. There are many reasons why data might be lost. Missing data may result from sensor failure, maintenance needs of instrumentation hardware, telephone line problems, or because of bad data processing.

The Energy Systems Laboratory of Texas A&M operates the LoanSTAR monitoring program (Turner, et al. 1990), which has supported the development, maintenance and use of a large database. This database is one of the largest depositories of building energy use and weather data. It currently records data from more than 50 weather stations (from the National Weather Service, NWS), and also includes comparisons were carried out. The first comparison was made between the estimated and actual values, as time series and as cross check plots. The second comparison is based on the energy use data from over 600 buildings (Haberl, et al., 1998). Unfortunately, there are missing data in the LoanSTAR database for all of the reasons indicated above. Currently the LoanSTAR database contains an essentially random distribution of missing data. Between 50% and 70% of the missing data in the LoanSTAR database correspond to gaps of six consecutive hours or less (Chen, 1999).

A large number of studies have been carried out for predicting weather variables. Most of them are based on knowledge of the behavior of the variables for long periods (Hittle and Perdensen, 1981, Hokoi et al., 1990, Hansen and Driscoll, 1977, and McCutchan, 1979). Though reliable and accurate, these approaches present mathematical complexities and require large amounts of data. Other procedures for filling missing data come from other environmentally related fields, such as air quality dispersion (Atkinson and Lee, 1992), rainfall records (Tang, et al., 1996, and Makhuvha, et al., 1997), and hydrologic data (Bennis et al. 1997). These cases typically treat daily data and rarely involve hourly data. Colliver et al. (1995), in their investigation of occurrences of extreme dew point and mean coincident dry-bulb temperatures, tried to recover missing data and made comparisons between the interpolated hourly temperatures and the actual recorded temperatures. They found that the linear interpolation technique was the best technique to fit the dew-point temperature, but interpolation via cubic splines provided a better fit to dry-bulb data. However, they did not attempt to fill gaps greater than two hours. This paper analyzes the application of the cubic spline and Fourier series techniques for

filling in short periods of missing building energy use and weather data. The short periods investigated in this paper are those between 1-6 consecutive hours of missing data. The two mathematical methodologies are compared with the simple linear interpolation technique, which has been found by Chen and Claridge (2000) to be reliable for this task.

Table 1. Total number of data points in each sample and the number of missing data points in each sample.

BUILDING NAME	AREA (ft^2)	DATA TYPE	UNITS	ANALYZED PERIOD start	end	MISSING DATA	AVAILABLE DATA	LENGTH OF DATA SET	MEAN	AHU
Zachry Engng. Center, Texas A&M University	324,400	wbcool	MMBTU/hr	09/14/89	09/14/90	365	8419	8784	5.588	DDCAV
		wbcool	MMBTU/hr	01/01/92	12/31/92	532	8252	8784	3.749	VAV
Geology Building, University of Texas	127,000	wbcool	MMBTU/hr	02/01/96	02/01/97	145	8663	8808	3.400	DDCAV
		wbheat	MMBTU/hr	02/01/96	02/01/97	145	8663	8808	2.162	DDCAV
Main Building, University of Texas	81,000	wbcool	MMBTU/hr	06/04/93	06/04/94	238	8546	8784	1.900	MZRH, CVRH
Taylor Hall, University of Texas	100,773	wbcool	MMBTU/hr	07/17/97	07/17/98	0	8784	8784	0.901	VAV
		wbheat	MMBTU/hr	07/17/97	07/17/98	0	8784	8784	0.604	VAV
		wbcool	MMBTU/hr	06/22/96	06/23/97	1	8807	8808	0.863	DDVAV, DDCAV
		wbheat	MMBTU/hr	06/22/96	06/23/97	1	8807	8808	0.507	DDVAV, DDCAV
Jester Hall, University of Texas	157,270	wbheat	MMBTU/hr	04/07/96	05/05/97	105	9351	9456	0.621	DDMZ, CVRH
		wbheat	MMBTU/hr	03/18/97	05/07/98	2	9982	9984	0.404	VAV
NWS: Houston, TX	N. A.*	Tdb	°F	01/01/95	12/31/95	143	8617	8760	69.81	N. A.*
	N. A.*	Tdp	°F	01/01/95	12/31/95	144	8616	8760	60.13	N. A.*
NWS: Washington, DC	N. A.*	Tdb	°F	01/01/97	12/31/97	152	8608	8760	54.98	N. A.*
	N. A.*	Tdp	°F	01/01/97	12/31/97	152	8608	8760	44.11	N. A.*
NWS: Minneapolis, Min	N. A.*	Tdb	°F	04/01/96	04/01/97	1037	7747	8784	43.04	N. A.*
	N. A.*	Tdp	°F	04/01/96	04/01/97	1037	7747	8784	34.00	N. A.*
NWS: El Paso, TX	N. A.*	Tdb	°F	01/01/97	12/31/97	123	8637	8760	64.46	N. A.*
	N. A.*	Tdp	°F	01/01/97	12/31/97	123	8637	8760	38.86	N. A.*
College Station, TX	N. A.*	Tdb	°F	01/01/94	12/31/94	147	8613	8760	67.99	N. A.*

*N. A. means "Not Applicable"

DESCRIPTION OF THE SAMPLES

The samples used for this study were separated according to variable and the total sample set included five locations for dry-bulb temperature, four for dew-point temperature, six buildings for whole building cooling, and five sites for whole building heating.

The samples of National Weather Service dry–bulb temperature and dew point temperature data contain extreme weather conditions. For example, the Minneapolis, MN site has occurrences of very low ambient temperatures in winter and El Paso, TX has some very high summer temperatures. The temperatures in the total dry-bulb temperature sample range from –20 °F to 110°F. The cooling and heating consumption data have been collected from monitored buildings in the LoanSTAR database. These data come from buildings of different sizes, different air conditioning systems, (see Table 1), and different levels of energy use. Some of the data sets come from the same building in two periods; in such cases, the system has been retrofitted between the data sets. Table 2 presents the number of locations and length of the samples that are considered in this study.

The mean values of the whole building cooling use data sets vary from 0.86 to 5.58 MMBTU/hr, while those for whole building heating use vary from 0.40 to 2.16 MMBTU/hr. In the case of the temperature data, the averages of the individual samples range from 43 °F to 69.8 °F for the dry-bulb temperature, and from 34 °F to 60.1 °F for the data sets of dew-point temperature.

Table 1. Type and length of the data sets used in the numerical experiment.

Variable	Units	Locations	Number of Data Records
Tdb	[°F]	5	43,824
Tdp	[°F]	4	35,064
Wbcool	[MMBtu/hr]	6	52,752
Wbheat	[MMBtu/hr]	5	45,840
TOTAL		**20**	**177,840**

Table 3 shows a summary of all the mathematical interpolation approaches included in this study. 114 simulations were carried out for each sample, giving a total of 2280 simulations, to determine which technique and approach seems most suitable for filling gaps in the LoanSTAR database.

Table 2. Summary of the approaches tested for filling in missing data.

APPROACH	DATA Before	DATA After	TERMS	APPROACH	DATA Before	DATA After	TERMS
Linear	1	1	-	Fourier	12	12	6
Spline	4	1	-	Fourier	12	12	8
Spline	4	4	-	Fourier	12	12	10
Spline	4	7	-	Fourier	24	24	4
Spline	6	4	-	Fourier	24	24	6
Spline	6	6	-	Fourier	24	24	8
Fourier	6	6	4	Fourier	24	24	10
Fourier	6	6	6	Fourier	24	24	12
Fourier	6	6	8				
Fourier	6	6	9				

METHODOLOGY

The mathematical methodologies for filling data gaps analyzed in this paper are cubic splines and Fourier series. Neither methodology is primarily an interpolation technique, but they have been implemented to evaluate the available time series data as a piecewise function, where it is possible to locally utilize them as interpolating techniques. A complete description of these mathematical methodologies and their implementation is presented in Baltazar (2000).

Seventeen approaches, 5 related to the cubic spline and 12 belonging to the Fourier series, were tested on 20 samples of energy use and weather data, each of them at least one year in length (see Tables 1 and 2).

In general, the procedure used to evaluate the performance of the approaches is based on the creation of artificial missing data, which are called pseudo-gaps. Given the random characteristic of the actual missing data in the LoanSTAR database, from which the data were taken, it is necessary to identify the actual missing data points present in each data set and mark them. Another important factor is the extent to which neighboring data affects the pseudo-gaps. For this reason, the techniques were tested with a different number of points preceding and following the pseudo-gaps. These considerations influence the number of pseudo gaps that can be defined and evaluated with each approach within a given data set.

ANALYSIS OF RESULTS

Interpolation depends on the behavior of the variable in the neighborhood of the point to be estimated. In most cases, the best estimation of a particular value comes from models that do not include high degree terms. That is why smoothness techniques, like the spline or linear combinations of trigonometric functions such as the Fourier series, may help in tasks of interpolation. These techniques appear to give a soft continuity through the points where they are applied.

In general, the spline technique can be defined in different degrees, according to the number of times that it is mathematically differentiable. Thus, cubic, quadratic, or "linear" (linear piecewise curve) splines may be defined. The spline technique that was used in this study is the cubic spline, because it is widely used, and it has proven to give a better representation of physical phenomena. Fourier series is just a mathematical way of representing a function through a linear combination of sine and cosine functions. This same principle can be discretized to evaluate the series under suitable conditions and to follow the pattern of some data. The Fourier series is the base of many engineering applications, and its characteristics are very well known.

In order to analyze the influence of the points around the missing values, different numbers of points before and after a pseudo-gap were tested. Fourier series can be developed with many variants; in this paper 6, 12, and 24 hours before and after each pseudo-gap of missing data were evaluated. Note that in this technique only a symmetric number of data points around the pseudo-gap were considered. The number of terms was also varied in each approach (see Table 3) in order to observe their influence on the evaluation of pseudo-gaps.

The nomenclature for identification of the Spline approaches is as follows: Spline (before, after), with the numbers inside the parenthesis expressing the number of points utilized before and after the pseudo-gap of continuous missing data. The nomenclature used to identify all the approaches for Fourier series is similar to that of the splines, but includes an extra entry for the number of terms. This is represented as Fourier (before, after, terms).

Performance of the Approaches for Filling in Missing Data

It is possible to observe the behavior of one approach for filling in missing data by directly comparing the generated points with the measured data set. This kind of comparison is shown in Figure 1 for the ambient temperature data set of College Station, TX. Each of the four plots in this figure presents two weeks of hourly measured data, filled data for this period, and the respective residuals. The four plots each correspond to one of the four seasons in the year. For the cases shown, reading from left to right and top to bottom, the figure presents the behavior in winter, spring, summer, and fall respectively. It is easy to observe in this figure how the linear approach behaves and how it follows the trend of the data in each season of the year. While the figure gives a good qualitative understanding of the

behavior, the statistical correlation between the estimated and the real values might be helpful. Figure 2 shows a crosscheck plot between the estimated ambient temperature and the measured data set. For this case, the regression coefficient is 0.9518, which represents a high degree of correlation. Since these plots are very helpful for explaining the performance of any approach for a particular gap size, plots for all the approaches and all the gap sizes were created and analyzed (Baltazar, 2000).

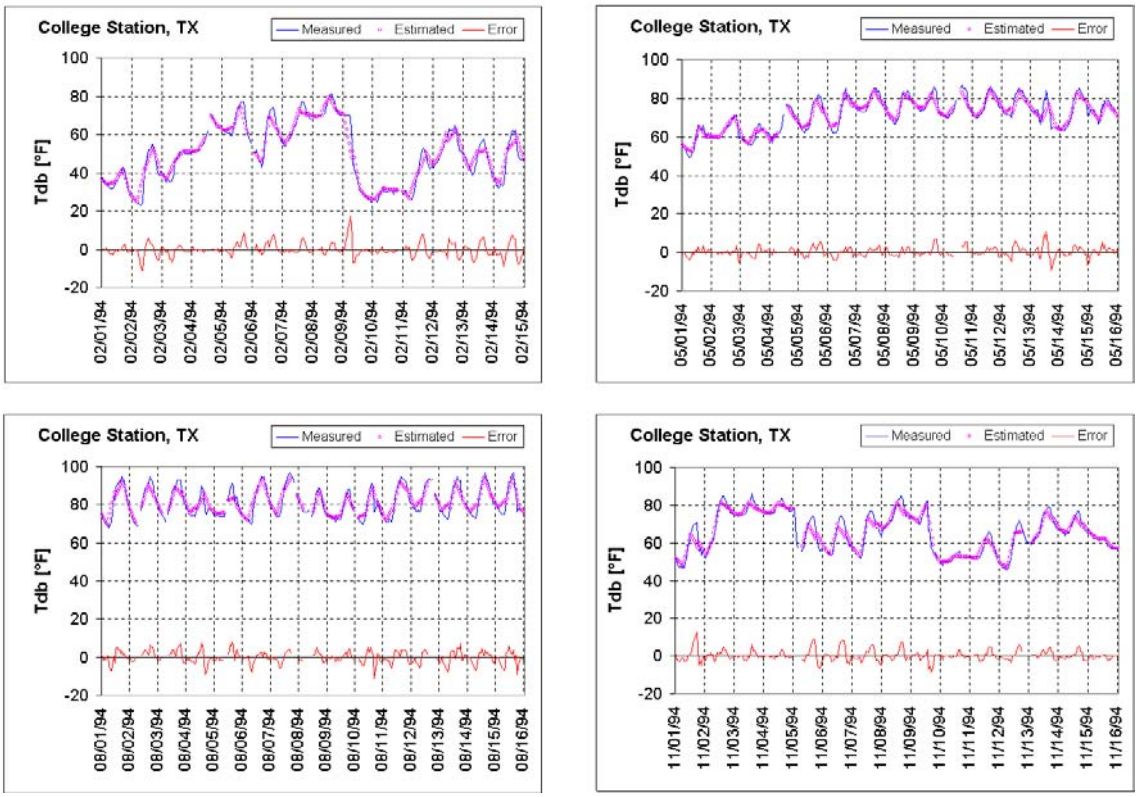

Figure 1. Comparison of estimated and measured dry-bulb temperature time series for College Station, TX in different seasons. Interpolation technique: linear, pseudo-gap=6.

Another qualitative way for evaluating the performance of any approach is to assess the fraction of the data set that can be accurately estimated to within some specified difference or percentage of the measured values. These fractions help to identify differences between techniques, or validate how close in aggregate the interpolation technique is to the real sample. Figure 3 shows the percentage of the total filled points that are within 1°F, 3°F, 5°F, and 10°F of the measured dry bulb temperature for College Station, TX for a spline approach. For whole building energy use data sets, instead of grouping by differences from measured values, the grouping is done by percentages of difference from the actual values (i.e. within 1%, 5% and 10%) of the data set. The plots were generated by approach and by length of pseudo-gap. The number of pseudo-gaps evaluated includes all those for which this approach can be directly compared with others. The number of pseudo-gaps differs due to actual gaps in the data and the fact that different methods require different numbers of consecutive points before and after a gap starts. A description of this fact can be found in Baltazar (2000). From the distribution presented in Figure 3, it is possible observe that this index is quite stable for different numbers of points in the sample in each pseudo-gap length analyzed, but does show small variation.

Using this index is possible to compare the approaches for each sample. Figures 4 to 6 compare the linear, Spline (4,4) and Fourier (12,12,10) approaches for dry-bulb temperature data of College Station, TX within 1°F, 3°F and 5°F, respectively. It is observed that the linear interpolation technique has a clear advantage over the other techniques for pseudo-gaps of three hours or less. However, for pseudo-gaps longer than three hours, splines generally give a slightly better performance than the

other techniques. The Fourier series technique is, in all but two of the study cases, the option with the poorest performance.

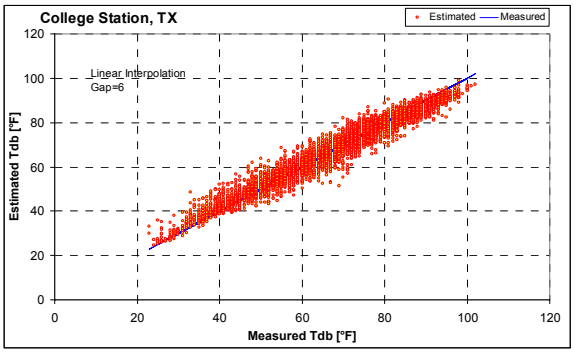

Figure 2. Crosscheck plot of total estimated and measured dry-bulb temperature data set for College Station, TX. Interpolation technique: linear, pseudo-gap=6.

Analysis of all the samples produced several trends: it is observed that the smaller the pseudo-gap length, the smaller the spread among the approaches for any difference (or percentage) between estimated and measured data. In addition, the greater the difference between the interpolated and the real values, the smaller the differences found among the approaches.

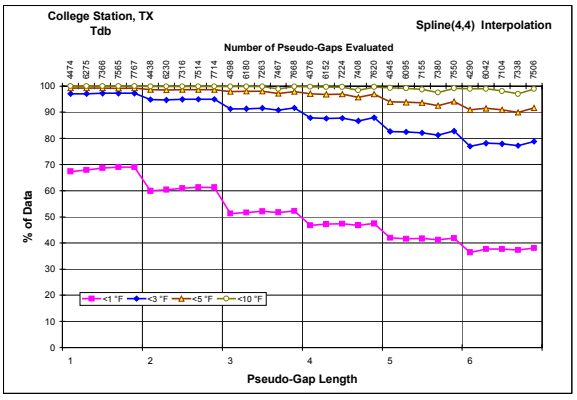

Figure 3. Percent of the ambient temperature data within the indicated difference of the measured values as a function of pseudo-gap length.

It is evident that this kind of graph may offer a qualitative understanding of the accuracy that is being obtained from the application of any interpolation technique.

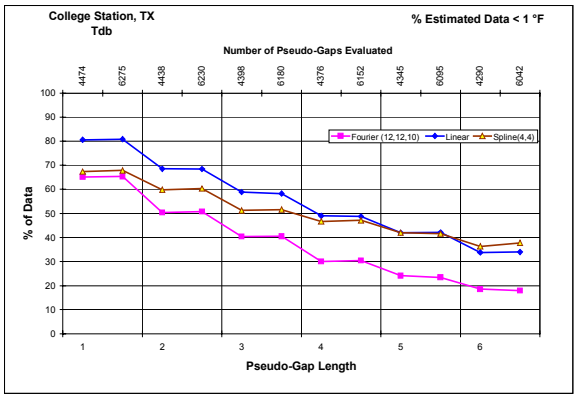

Figure 3. Performance of linear, spline (4,4), and Fourier(12,12,10) approaches for a difference of 1°F of the measured values for different pseudo-gap lengths. Dry bulb temperature data for College Station, TX.

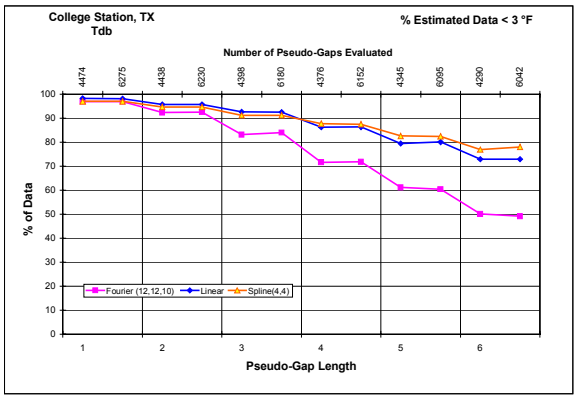

Figure 4. Performance of linear, spline (4,4), and Fourier(12,12,10) approaches for a difference of 3°F of the measured values for different pseudo-gap lengths. Dry bulb temperature data for College Station, TX.

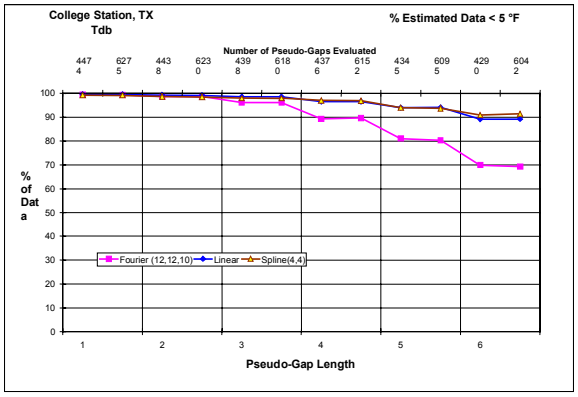

Figure 5. Performance of linear, spline (4,4), and Fourier(12,12,10) approaches for a difference of 5°F of the measured values for different pseudo-gap lengths. Dry bulb temperature data for College Station, TX.

CONCLUSIONS

To define a single "best" technique for evaluation of missing data in time series information for weather and for building energy use data is a complex task.

The qualitative analysis techniques shown give a general understanding of how each approach behaves, and allows the observation of general trends. It is observed that the smaller the pseudo-gap length, the smaller the spreading among the approaches for any difference (or percentage) between estimated and measured data. Also the greater the difference between the interpolated and the real values, the smaller the differences found among the approaches.

From the results of this study, it appears that linear interpolation is a better approach for filling gaps one to three hours long. The cubic splines approach gave better performance for gaps between four and six

Since statistical (or quantitative) parameters are generally more accepted for the assessment of the behavior of data analysis, the use of a normalizing technique to determine the best approach for this application is recommended. An investigation of this approach is reported elsewhere (Baltazar and Claridge 2002).

REFERENCES

Atkinson. D., and Lee, R. F., 1992, "Procedures for Substituting Values for Missing NWS Meteorological Data for Use in Regulatory air Quality Models", Available Internet: http://earth1.epa.gov/scram001/surface/missdata

Baltazar-Cervantes, J. C., 2000, "Study of Cubic Splines and Fourier Series as Interpolation Techniques for Filling in Short Periods of Missing Building Energy Use and Weather Data" M.S. Thesis, Mechanical Engineering Department, Texas A&M University, College Station, TX, December.

Baltazar, J.C. and Claridge, D.E., 2002, "Study of Cubic Splines and Fourier Series as Interpolation Techniques for Filling in Short Periods of Missing Building Energy Use and Weather Data," to be published in *Solar Engineering 2002 – Proc. of ASME International Solar Energy Conference*.

Bennis, S., Berrada, F., and Kang, N., 1997, "Improving Single-variable and Multivariable Techniques for Estimating Missing Hydrological Data", *Journal of Hydrology*, Vol. 191, pp. 87-105.

Chen, H., 1999, "Rehabilitating Missing Energy Use and Weather Data When Determining Retrofit Energy Savings in Commercial Buildings", M.S. Thesis, Mechanical Engineering Department, Texas A&M University, College Station, TX, December.

Chen, H., and Claridge, D. E., 2000, "Procedures for Filling Short Gaps in Energy Use and Weather Data", *12th Symposium on Improving Building Systems in Hot and Humid Climates*, pp. 314-326, San Antonio, TX, May.

Colliver, D. G., Zhang, H., Gates, R. S., and Priddy, K. T., 1995, "Determination of the 1%, 2.5%, and 5% Occurrences of Extreme Dew-Point Temperatures and Mean Coincident Dry-Bulb Temperatures", *ASHRAE Transactions*, Vol. 101, part 2, pp. 265-286.

Haberl, J. S., Thamilseran S., Reddy T. A., Claridge, D. E., O'Neal, D., and Turner, W. D., 1998, "Baseline Calculations for Measurement and Verification of Energy and Demand Savings in a Revolving Loan Program in Texas", *ASHRAE Transactions*, Vol. 104, Pt. 2, pp. 841-858.

Hansen, J. E., and Driscoll, 1977, "A Mathematical Model for the Generation of Hourly Temperatures", *Journal of Applied Meteorology*, Vol. 16, September, pp. 935-948.

Hittle, D. C., Pedersen, C. O., 1981, "Periodic and Stochastic Behavior of Weather Data", *ASHRAE Transactions,* Vol. 87, Part 2, pp 173-194.

Hokoi, S., Matsumoto, M., and Kagawa, M., 1990,"Stochastic Models of Solar Radiation and Outdoor Temperature", *ASHRAE Transactions*, Vol. 87, Part 2, pp. 245-252.

Makhuvha, T., Pegram, G., Sparks, R., and Zucchini, 1997, "Patching Rainfall Data Using Regression Methods: 1. Best Subset Selection, EM and Pseudo-EM Methods: Theory", *Journal of Hydrology*, Vol. 198, pp. 289-307.

McCutchan, M. H., 1979, "Determining the Diurnal Variation of Surface Temperature in Mountainous Terrain", *Journal of Applied Meteorology*, Vol. 5, September, 1224-1229.

Tang, W. Y., Kassim, A. H. M., and Abubakar, S. H., 1996, "Comparative Studies of Various Missing Data Treatment Methods- The Malaysian Experience", *Atmospheric Research*, Vol. 42, pp. 247-262.

Turner, W. D., 1990, "Overview of the LoanSTAR Monitoring Program", *7th Symposium on Improving Building Systems in Hot and Humid Climates*, pp.28-34, Fort Worth, TX, October.