CHAPTER 1

Introduction

There are many components of building design, and with the supply of non-renewable energy resources decreasing, the manner and amount of a building's energy consumption is an increasingly important and cost-significant aspect of the design. Actual testing of building energy systems is not often practical due to the dependence of system performance upon solar radiation, ambient temperature, humidity, and windspeed. These weather variables are neither completely random nor deterministic; they can best be described as irregular functions of time, both on small (e.g., hourly or daily) and large (e.g., seasonally or yearly) time scales. It is this irregular behavior of the weather which complicates the analyses of energy systems and makes experimental determinations of system performance time consuming, expensive, and inconclusive with regard to the manner in which they would have performed under other climatic conditions. An experiment run for one year will yield only the performance for that particular year; it will not give a good indication of the average performance over the life of the system. Two alternative techniques to experimentation are the use of correlation methods and computer simulations.

Correlation methods, such as f-Chart [Klein et al, 1976], permit the designer to quickly obtain an estimate of system performance from several design parameters and the appropriate correlation. For several standard system configurations, simulation results obtained while varying several design parameters are expressed as functions of these design parameters. From these correlations, only the final performance estimate (for f-Chart, the fraction of the load met by solar) is obtained; nothing is learned about the system dynamics. Correlation methods are limited in that they are applicable only for the system configurations and range of parameters for which they were developed.

Simulations are more versatile; the designer is able to model many different systems and configurations without the expense and time required for experimentation. The simulations, like actual testing, are dependent upon weather inputs, therefore the same problem of one year of weather data not providing meaningful results still exists. On the other hand, since simulations are not run in real time, they can be run for a sufficiently long time (e.g. twenty years) to obtain an average performance estimate, assuming long-term weather data exists for the desired location. Twenty years of hourly weather measurements is a lot of data, requiring large amounts of storage. More significantly, such detailed weather records do not exist for many locations; in the U.S., they are available for approximately 26 sites. Such large amounts of data are often awkward to work with; simulations run for that length are time consuming (on small computers) and expensive (on large computers).

Design years, which consist of one year (or some other amount) of data which will supposedly yield a sufficiently accurate estimate of the average long-term performance, are commonly used. They allow designers to consider more options and alternatives by speeding up the simulation process. An example these design years are the "Typical Meteorological Years" (TMY), developed for 26 U.S.

locations [Hall et al, 1978]. They consist of 12 typical months, where the months were chosen from 23 years of recorded data to be representative of the long-term. The cumulative distributions, means, medians, and persistence structure of thirteen variables were compared. Since it is virtually impossible to find one month out of the 23 in which all variables display characteristics identical to long-term, a weighting system was devised and the best possible month was chosen. While the use of TMY data is an improvement over using just one year of actual data, it still cannot accurately represent the statistics of the long-term. It is, in effect, a compromise between using a reduced set of weather data and accurately portraying the statistics represented in long-term data. Other examples of design years are the "Test Reference Years" (TRY) [National Climatic Center,1976] and those developed by [Arens,1980].

These design years do not solve the problem of providing data for other locations; their development requires the existence of long records of data for the location. The need to provide data elsewhere has resulted in the concept of weather data generation. There are two major areas of generation; extrapolation and synthesis. Extrapolation involves using the data at one or more similar or neighboring locations to infer the weather data at a location for which data are not available. Extrapolation of existing data requires great care, as large errors may result [Suckling,1985]. Even so, it is still necessary to have data to extrapolate from.

Synthesis of hourly weather data involves taking monthly-average inputs and combining them with documented long term statistical patterns of weather (and correlations) to "make up" or synthesize weather data. Surprisingly, when observed over a long enough period of time, weather data can often be shown to behave in a location independent manner. This concept is what allows synthesizers to work; however, the accuracy of the data produced from such a synthesizer is limited by the ability of the location independent correlation to reproduce the long-term weather statistics at a particular site. Generated data will not reproduce location specific tendencies, such as morning fog, as these will be erased by the location independent correlations used. Numerous authors have developed models for the generation of radiation sequences (both daily and/or hourly), for example [Amato,1986; Brinkworth,1977; Hittle and Pedersen,1981; Graham,1985]. Degelman [1976] developed a weather generator which requires only limited monthly-average values as input and outputs hourly radiation (both total and direct normal), dry bulb and dewpoint temperatures, relative humidity, and windspeed all with a statistical behavior similar to that observed in long-term weather data records.

There are two basic approaches to the development of a weather data synthesizer. The first, a "Type I" or "realistic" generator, is to combine the existing information concerning distributions, autocorrelations, and crosscorrelations of the meteorological variables into a format which is driven by a random number generator and produces, over the long-term, weather statistics which are indistinguishable from those of recorded data. Techniques commonly used involve modeling of the deterministic trends with Fourier series or other means followed by use of time series analyses for the stochastic component. Most of the generation models for the different weather variables developed so far have been autoregressive models of order 1 or 2 (AR1 or AR2). These generators do not in any way predict the weather; they only provide a statistically possible weather sequence.

The disadvantage of these models, however, is that the generated data is subject to the same variability as real data and consequently simulations would have to be run for long periods in order to reproduce statistics apparent in long-term weather records. From this synthetic data a TMY year could be constructed, but as with actual data, it is difficult to choose one year of data which closely represents all of the statistics observed in the long-term data. However, since the data are being generated, it should be possible to generate a year in which *all* of the statistics match the long-term much more accurately than in the TMY years. This type of generator is an "average" or "Type II" generator, and it is the approach originally taken by Degelman and continued in this study. The only required inputs are latitude, monthly-average radiation, ambient temperature and humidity ratio values.

This study consisted of evaluating the Degelman generator, both by analyzing the techniques involved and the resulting generated data. Long-term statistics were compiled from 22 years of recorded hourly weather data [SOLMET, 1979] at three U.S. locations to provide a basis for comparison. Areas of weakness in the Degelman generator were identified and improved upon with more recent work. In addition, the statistics of the TMY data were compared. The analysis was done for three locations; Madison, WI, Albuquerque, NM, and New York, NY, representing three different climates.

Chapter 2 provides an overview and summary of the statistics used and the observed long-term properties of the various weather variables. Chapter 3 describes the radiation generation process, beginning with the description and evaluation of the Degelman generator, describing modifications, and comparing results. Chapter 4 presents the method of ambient temperature generation, again starting with Degelman's technique, followed by an evaluation, an alternative technique, and the comparison of generated and long-term data. Comparisons of the TMY data for the three locations are also included. Chapter 5 summarizes the Degelman approach to generating relative humidity and windspeed data; no attempt to improve upon his techniques has been made due to time limitations. The appendices include some of the many statistics and graphs collected during the data analysis.

CHAPTER 2

Statistical Overview

Statistics provide the only means for characterizing and comparing large groups of data. In particular, the adequacy of the generated weather data can be evaluated by comparing its statistics with those of the long-term data.

Weather varies greatly from place to place and month to month, and on a short-term basis seems highly random and unpredictable. Over the long-term, however, many of the statistics which describe weather data can be sufficiently represented by location-independent correlations.

The various statistical quantities and documented long-term meteorological properties used throughout this work are summarized in this chapter.

2.1 Statistics

2.1.1 Distributions

To construct a distribution from a set of data, the data must be grouped, or "binned", by placing all values within a small range into a bin and recording the number of values within each bin. For example, in constructing a distribution of ambient temperature, all temperatures between 16 and 18 degrees C might make up one bin and all be considered as 17 degrees C. Doing this with all of the data results in a histogram, where the width of the bins is 2 degrees C. The bin width is not arbitrary; selection of too small of a bin width results in some bins without data, while too large a bin width smooths out the distribution and loses resolution, consequently losing information. After constructing a histogram, the number of values in each bin are divided by the total number of values and then by the bin width, yielding the probability density for each bin. The histogram can be smoothed and a function fit which describes the variation in probability density as a function of the variable. This is called the probability density function (pdf). As an example, the pdf of the normal or Gaussian distribution is the familiar bell-shaped curve (Figure 2.1). The integral of the pdf between the limits negative infinity and positive infinity must always be one (the probability of all possible values must be equal to one); that is

$$\int_{-\infty}^{+\infty} p(y)\,dy = 1 \tag{2.1.1}$$

where p(y) is the pdf. The cumulative distribution function (cdf) is the function obtained by integrating the pdf.

$$F(y) = \int_{\infty}^{y} p(y)\,dy \tag{2.1.2}$$

It can also be constructed directly from the data. The cdf relates the cumulative fraction of occurrence (the fraction of time the variable under consideration takes on a value less than some particular value) to a particular value of the variable. The cdf

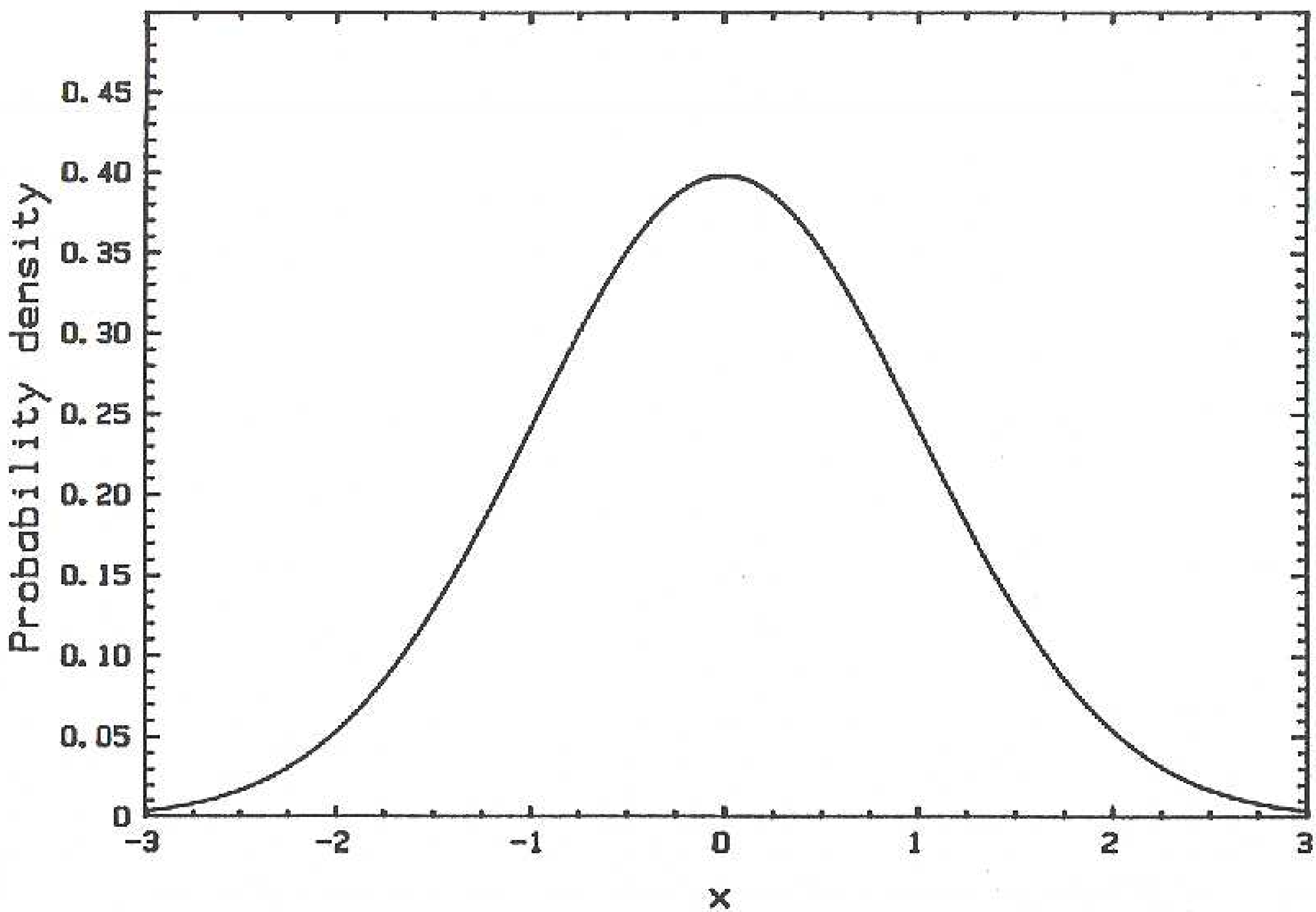

Figure 2.1 Probability Density Function of the Normal Distribution

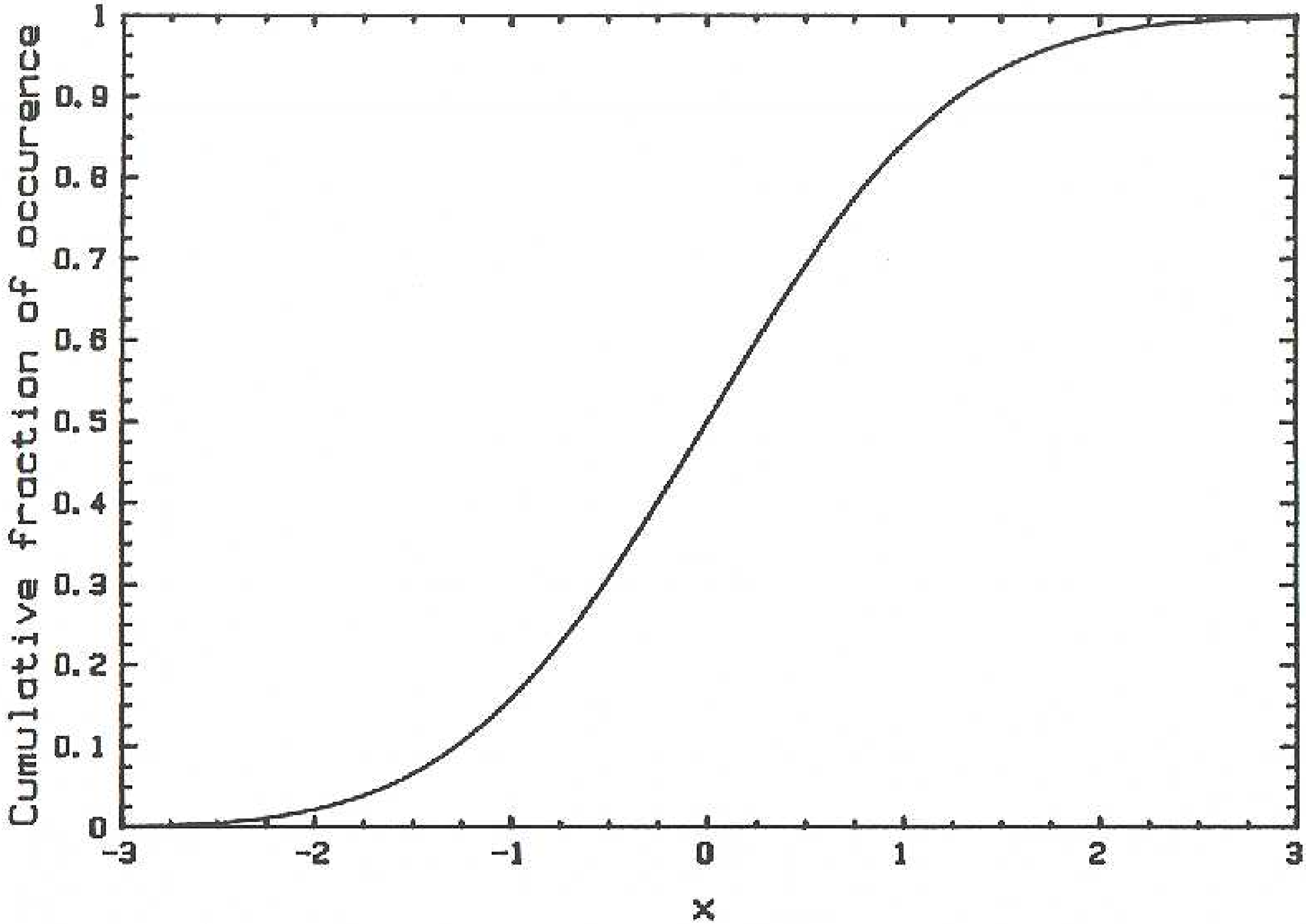

Figure 2.2 Cumulative Distribution Function of the Normal Distribution

as defined is always an increasing function; an example of the cdf corresponding to the normal distribution is shown in Figure 2.2.

A convenient way to characterize a distribution is by its moments. The i^{th} central moment, or i^{th} moment about the mean, is defined as

$$m_i = E\left[(y - \mu)^i\right] \tag{2.1.3}$$

where μ is the mean. The moments of a distribution are unique, so that if all of the central moments and the means of two distributions are the same, then the distributions are equivalent. While there are an infinite number of moments, for the purposes of this study, only the first four will be used to compare distributions.

The sample estimate of the mean is obtained from

$$\mu = E[Y] \approx \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.1.4}$$

where n is the number of data points.

The first four central moments can then be written as

$$m_1 = E\left[Y - \mu\right] = 0 \tag{2.1.5}$$

$$m_2 = E\left[(Y-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 \tag{2.1.6}$$

$$m_3 = E\left[(y-\mu)^3\right] = \frac{1}{n}\sum_{i=1}^{n} y_i^3 - \frac{3\bar{y}}{n}\sum_{i=1}^{n} y_i^2 + 2\bar{y}^3 \tag{2.1.7}$$

$$m_4 = E\left[(Y-\mu)^4\right] = \frac{1}{n}\sum_{i=1}^{n} y_i^4 - \frac{4\bar{y}}{n}\sum_{i=1}^{n} y_i^3 + \frac{6\bar{y}^2}{n}\sum_{i=1}^{n} y_i^2 - 3\bar{y}^4 \tag{2.1.8}$$

Rather than dealing with the moments directly, four related measures with more intuitive meaning are the mean, the standard deviation, the coefficient of skewness, and the coefficient of excess kurtosis. The mean, as defined in Equation (2.1.4), is simply the average of the data. The variance of the data, σ^2, is equal to the second moment, however, the square root of the variance, the standard deviation, is measured in the same units as the data, and as such is intuitively more appealing. The standard deviation, σ, is a measure of the scatter or spread of the data.

$$\sigma^2 = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} y^2 - (\sum_{i=1}^{n} y)^2/n}{n-1}} \tag{2.1.9}$$

The numerator under the square root sign is called the sum of squares of the data; the divisor n-1 is used instead of n due to the loss of a degree of freedom in estimating the mean.

The coefficient of skewness, α_3, measures the lack of symmetry of a distribution. For a symmetrical distribution, α_3=0, while if skewed to the left, α_3 is positive and if skewed to the right, α_3 is negative. The skewness is related to the third moment by

$$\alpha_3 = \frac{m_3}{m_2^{3/2}} \tag{2.1.10}$$

The coefficient of excess kurtosis, α_4, measures the relative peakedness of a distribution. It is defined such that α_4=0 for a normal distribution. Positive values of α_4 reflect a more peaked distribution, while a negative value indicates less peakedness. The relation of the excess kurtosis to the fourth moment is

$$\alpha_4 = \frac{m_4}{m_2^2} - 3 \tag{2.1.11}$$

For a normal, or Gaussian, distribution, approximately 95% of the data is contained within plus or minus two standard deviations from the mean. The skewness and kurtosis are always zero; hence the normal distribution is uniquely defined by its mean and standard deviation.

2.1.2 Correlation

Sometimes it is useful not only to have information about the distribution of a variable, but also how it relates to itself and other variables at different times. A measure of the correlation between two variables, x and y, is the cross-correlation coefficient.

The value of a correlation coefficient ranges between -1 and 1. A value of zero indicates no correlation, -1 indicates perfect negative correlation, (if x increases, y decreases), while +1 indicates perfect positive correlation, (if x increases, y increases).

Correlations can be calculated between variables at different times, e.g., between the present value of x (today's temperature) and the value of y two timesteps previously (the radiation two days ago). The difference between the timesteps is called the lag.

The cross-correlation coefficient between two variables x and y at lag k can be estimated as follows:

$$\rho_{xy,k} = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(y_{t+k} - \bar{y})}{\sqrt{\sum_{t=1}^{n} (x_t - \bar{x})^2 \sum_{t=1}^{n} (y_t - \bar{y})^2}} \tag{2.1.12}$$

where n is the total number of values of x (or y) and x_t indicates the value of x at time t.

Similarly, the correlation between values of one variable at different times can be calculated; this is called the autocorrelation coefficient

$$\rho_{yy,k} = \frac{\sum_{i=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2} \tag{2.1.13}$$

where again, k is the lag, y_t is the value of y at time t, and n is the number of y values. The subscript yy will be subsequently dropped, such that the autocorrelation coefficient will be denoted simply by ρ_k and the crosscorrelation will be indicated by $\rho_{xy,k}$.

Both Equation (2.1.12) and (2.1.13) are approximations; a minimum of 50 consecutive values are usually considered necessary to obtain a sufficiently accurate estimate. An assumption involved is that the joint probability distribution is independent of absolute time, being solely a function of lag. This might not in fact be the case, for example, the autocorrelation between temperature at 2 and 3 am could be higher than the autocorrelation between 1 and 2 pm.

2.1.3 Time Series Models

The modeling of autocorrelated variables is often done with stochastic, or time series, models. The models estimate the current value of a variable as a function of past values and/or past random disturbances. Assuming a time series model is fit correctly, it should reproduce a series with the correct autocorrelation structure. A common technique for identification of a specific model involves the examination of the autocorrelation and partial autocorrelation functions [Box and Jenkins,1976].

To use a time series model, the series to be modeled must be stationary. A stationary model is one in which the the probability structure is independent of time, dependent only upon the lag. This definition is often relaxed, such that any series with a constant mean and variance, and autocovariance dependent solely on lag is considered weakly stationary.

Another requirement of proper model fitting is that the residuals must be normally-distributed random independent values. To obtain this end result, often the modeled series must itself be normally distributed or else transformed in such a way as to obtain normal residuals.

The family of autoregressive models are those in which the present value of a variable, y_t, is estimated from n past values plus a random error term, ε_t, called an ARn model.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_n y_{t-n} + \varepsilon_t \tag{2.1.14}$$

$\phi_1, \phi_2, \dots, \phi_n$ are the autoregressive parameters, estimated by linear regression. For an AR1 model,

$$y_t = \phi\, y_{t-1} + \varepsilon_t \tag{2.1.15}$$

the autoregressive parameter, ϕ_1, can be estimated as being equivalent to the lag one autocorrelation of the series. The mean of the random error term is zero, while an estimation of the variance can be obtained from $\sigma^2 = 1 - \phi_1{}^2$.

Moving average models predict the present value of a variable from the past n values of the random disturbance, called MAn models.

$$y_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \dots - \theta_n\varepsilon_{t-n} \tag{2.1.16}$$

The coefficients, $\theta_1, \theta_2, ..., \theta_n$ are estimated by non-linear regression.

Models which are a combination of the two can also be used to describe a sequence of data, for example an ARMA(2,1) model:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t - \theta_1\varepsilon_{t-1} \tag{2.1.17}$$

2.1.4 Error

One way of testing the accuracy of a model is to compare the values estimated from the model, $\hat{y}$, with the known or experimental values, y. Two general measures of the error between the estimated data and the experimental data are the bias error and the root-mean-square (RMS) error. They can be thought of as the direction and magnitude of the error, respectively.

The bias error, as its name implies, measures the bias of the estimated data with respect to the known data.

$$\text{Bias error} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{2.1.18}$$

While the actual difference between the values may be small, if the model consistently underpredicts values, a negative bias error would result. Likewise, if it consistently overpredicts, the bias error would be positive. If the errors were symmetric about the known values, the bias error would be zero.

On the other hand, while the bias error may be zero, the model might be predicting values significantly different from the known values. The RMS error is an indication of the magnitude of the error:

$$\text{RMS Error} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{2.1.19}$$

2.2 Long-Term Weather Statistics

2.2.1 Radiation

Radiation is often described in a dimensionless form called the clearness index, which is simply the ratio of the total radiation on a horizontal surface over at a specified time to the extraterrestrial global solar radiation on a horizontal surface at the same time. The instantaneous values can be integrated over any time period; commonly used quantities are hourly ($k_t=I/I_o$), daily ($K_t=H/H_o$), and monthly ($\overline{K}_T=\overline{H}/\overline{H}_o$) clearness indices. Equations for calculating the extraterrestrial radiation are presented in [Duffie and Beckman, 1980]. It is important to recognize that the daily clearness index (K_t) is not the same as the average of the hourly clearness indices for the day. Likewise, the monthly-average daily clearness index and the monthly clearness index are not identical (Table 2.1), however, the difference is small since H_o does not vary greatly throughout the month, and the quantities can generally be used interchangeably.

Table 2.1
Comparison of monthly-average daily clearness index and monthly clearness index

	Albuquerque NM		Madison WI		New York NY	
	$\frac{\overline{H}}{\overline{H}_o}$	$\frac{1}{n}\sum K_t$	$\frac{\overline{H}}{\overline{H}_o}$	$\frac{1}{n}\sum K_t$	$\frac{\overline{H}}{\overline{H}_o}$	$\frac{1}{n}\sum K_t$
Jan	0.638	0.637	0.445	0.444	0.389	0.389
Feb	0.667	0.666	0.496	0.496	0.416	0.414
Mar	0.685	0.685	0.500	0.499	0.439	0.437
Apr	0.715	0.713	0.479	0.476	0.450	0.450
May	0.727	0.728	0.506	0.507	0.472	0.474
Jun	0.740	0.739	0.536	0.536	0.473	0.473
Jul	0.702	0.702	0.544	0.543	0.474	0.474
Aug	0.704	0.703	0.552	0.552	0.474	0.474
Sep	0.716	0.716	0.519	0.518	0.473	0.473
Oct	0.702	0.703	0.497	0.493	0.469	0.467
Nov	0.672	0.670	0.396	0.395	0.384	0.384
Dec	0.633	0.633	0.381	0.381	0.350	0.349

The distribution of daily radiation, when expressed in terms of the daily clearness index, K_t, has been shown to be primarily dependent on the monthly-average daily clearness index, $\overline{K}_T$ [Liu and Jordan, 1960]. Analytical expressions for the cumulative distribution function have been proposed by Bendt et al[1981] and Hollands and Huget[1983] (Figure 2.3). There are slight differences between the two in terms of interpretation. The Hollands and Huget expression is a fit to the original work of Liu and Jordan, which defines the distribution in such a way that the mean is the long-term average value for a particular calendar month at a location. Bendt analyzed approximately 20 years of daily radiation data from 90 U.S. locations, completing a more comprehensive study. The distribution in this case,

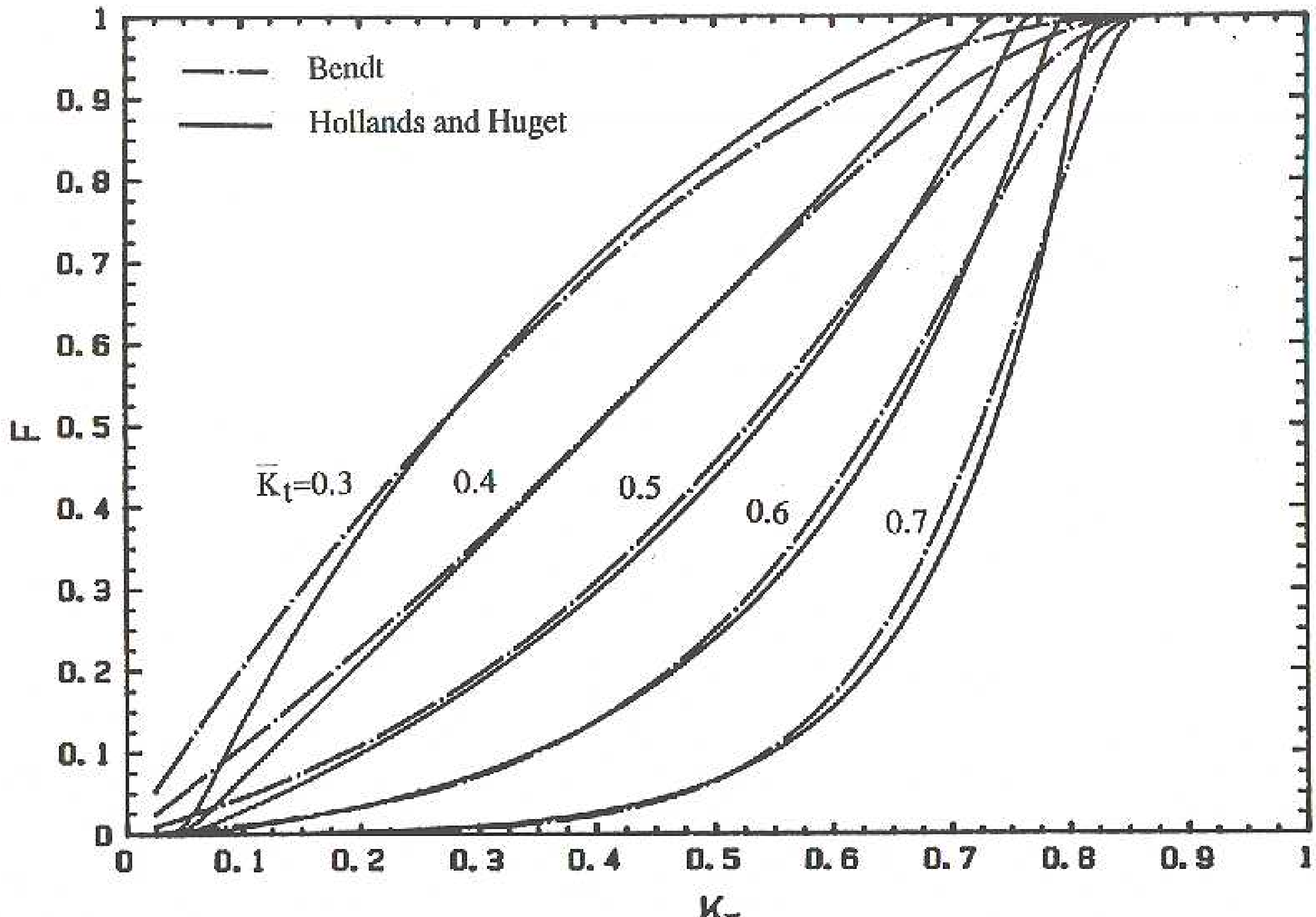

Figure 2.3 Cumulative Distribution Function of the Daily Clearness Index

however, was defined such that all months with a mean of a specified value are part of the same distribution. For example, the Liu and Jordan distribution assumes all Januarys in Madison are part of the same distribution, while the Bendt distribution assumes all individual month's with $\overline{K}_T$ values of 0.5 are part of the same distribution. As it turns out, the two sets of curves are very similar, differing mainly at the high fractional values, which Bendt attributes to Liu and Jordan's use of a constant extraterrestrial radiation throughout the month. A more rigorous examination of the shape's dependence on the mean was made by Graham[1985], who plotted the higher moments against the mean to demonstrate that all of the moments, and therefore the distribution, are in fact primarily functions of the mean.

There are several different distributions describing the variation of the hourly clearness index, k_t. One is the distribution of the hourly values about the monthly-average hourly value. This distribution has been found to be similar to the Liu and Jordan distribution described previously [Liu and Jordan, 1960]. Another type of hourly k_t distribution is that of the k_t values at a particular hour of the month (e.g. 10-11 am) about the average value at that hour. This distribution was observed to be similar in shape to the distribution of daily K_t values, but exhibiting somewhat less variability [Theilacker, 1980].

The long-term average diurnal variation of hourly total radiation (when divided by the average daily total radiation) and the hourly diffuse radiation (when divided by the average daily diffuse radiation) have been shown to be primarily a function of the hour angle, ω, and the sunset hour angle, ω_s. An expression for the average diffuse diurnal variation was developed by Liu and Jordan [1960]

$$r_d = \frac{\pi}{24}\left[\frac{\cos\omega - \cos\omega_s}{\sin\omega_s - (2\pi\omega_s/360)\cos\omega_s}\right] \quad (2.2.1)$$

while a correlation for the average diurnal variation of the total radiation was developed by Collares-Pereira and Rabl [1979]

$$r_t = r_d(a + b\cos\omega) \quad (2.2.2)$$

where

$$a = 0.409 + 0.5016\sin(\omega_s\text{-}60)$$

$$b = 0.6609 - 0.4767\sin(\omega_s\text{-}60)$$

Many authors have considered the estimation of diffuse radiation from total or direct normal radiation. Some of the more recent work has been by Orgill and Hollands[1977] and Erbs et al[1982], who developed expressions for estimating the hourly diffuse fraction (the ratio of the hourly diffuse radiation to the hourly total radiation) as a function of the hourly clearness index. Erbs also included work on a daily and monthly-average basis, in which a slight seasonal variation was observed and accounted for with a dependence on sunset hour angle. Equation (2.2.3) is the hourly expression developed by Erbs.

$$\begin{aligned} I_d/I &= 1.0 - 0.09k_t & k_t &\le 0.22 \\ I_d/I &= 0.9511 - 0.1604k_t + 4.388k_t^2 \\ &\quad - 16.638k_t^3 + 12.336k_t^4 & 0.22 < k_t &\le 0.80 \\ I_d/I &= 0.165 & k_t &> 0.80 \end{aligned} \quad (2.2.3)$$

Hollands and Crha[1987] studied the probability structure of the hourly diffuse fraction, and concluded that the distribution is strongly dependent upon k_t and may be quasi-universal.

The autocorrelation of daily total radiation and daily clearness index, while often considered to be the same, are in fact slightly different (Table 2.2). The sequence of radiation values has an underlying trend due to the changing earth-sun geometry; this trend is absent in the K_t sequence. The lag one autocorrelation of K_t has been observed by many authors to be in the approximate range of 0.15 to 0.35 [Amato,1986; Brinkworth,1977; Graham,1985; Klein, 1987]. Graham[1985] developed an expression which approximates ρ_1 as a function of monthly-average ambient temperature and the difference between $\overline{K}_T$ and the yearly-average $\overline{K}_T$, $\overline{K}_{Ty}$.

$$\rho_{1,Kt} = 0.259 + 0.551\,(\overline{K}_T - \overline{K}_{Ty}) + 0.008\,\overline{T} - 0.086\langle\overline{T}-17\rangle \qquad (2.2.4)$$

The < > around the last term indicates that it should be included only when it is positive.

Table 2.2
Comparison of lag one daily autocorrelation values of total solar radiation (H) and clearness index (K_t)

	Albuquerque NM		Madison WI		New York NY	
	H	K_t	H	K_t	H	K_t
Jan	0.25	0.17	0.20	0.16	0.04	0.03
Feb	0.29	0.22	0.20	0.14	0.10	0.05
Mar	0.26	0.17	0.25	0.22	0.19	0.15
Apr	0.26	0.22	0.18	0.17	0.14	0.12
May	0.20	0.20	0.24	0.23	0.12	0.12
Jun	0.25	0.24	0.16	0.16	0.21	0.21
Jul	0.23	0.21	0.11	0.11	0.13	0.12
Aug	0.29	0.26	0.16	0.14	0.16	0.16
Sep	0.30	0.22	0.24	0.17	0.25	0.22
Oct	0.43	0.32	0.41	0.35	0.24	0.19
Nov	0.29	0.23	0.23	0.19	0.14	0.12
Dec	0.26	0.27	0.18	0.18	0.08	0.07

2.2.2 Ambient Temperature

Erbs[1984] analyzed 22 years of ambient temperature data at 9 U.S. locations and found that both the distribution and the diurnal variation could be characterized by location independent correlations.

The distribution of the daily-average ambient temperature, $\overline{T}$, about the monthly-average daily ambient temperature, $\overline{T}_m$, exhibits wide variations in the mean and standard deviation for different months and locations. Although the distribution is not Gaussian, Erbs showed that when expressed in terms of the normalized variable h, the cumulative distribution can be represented by the following equation

$$F_{temp} = \frac{1 + \tanh(1.698h)}{2} = \frac{1}{1 + \exp(-3.396h)} \tag{2.2.5}$$

where

$$h = (T - \overline{T}_m)/(\sigma_m \sqrt{N/24}) \tag{2.2.6}$$

In Equation (2.2.6), N is the number of hours in the month and σ_m is the standard deviation of $\overline{T}_m$ about its long-term average value; Erbs was able to relate $\overline{\sigma}_m$ to $\overline{T}_m$ and σ_{yr}, the standard deviation of the 12 $\overline{T}_m$'s about the yearly-average temperature:

$$\sigma_m = 1.45 - 0.0290\,\overline{T}_m + 0.0664\,\sigma_{yr} \tag{2.2.7}$$

Erbs observed differences in the skewness for different locations and months, but was not able to correlate it to any other weather variable, and as such recommends Equation (2.2.5) which is a symmetric "average" distribution (Figure 2.4).

Erbs also explains that the expression for the distribution is equally valid for the distribution of the hourly temperatures about their hourly monthly-average values, $\overline{T}_h$, if $\overline{T}_h$ is substituted for $\overline{T}_m$ in Equation (2.2.6). The same relation for σ_m can be used since σ_m is approximately equal to the standard deviation of the hourly temperatures.

The diurnal variation of the hourly monthly-average temperatures, $\overline{T}_h$, has a location and month independent shape when standardized by subtracting the mean

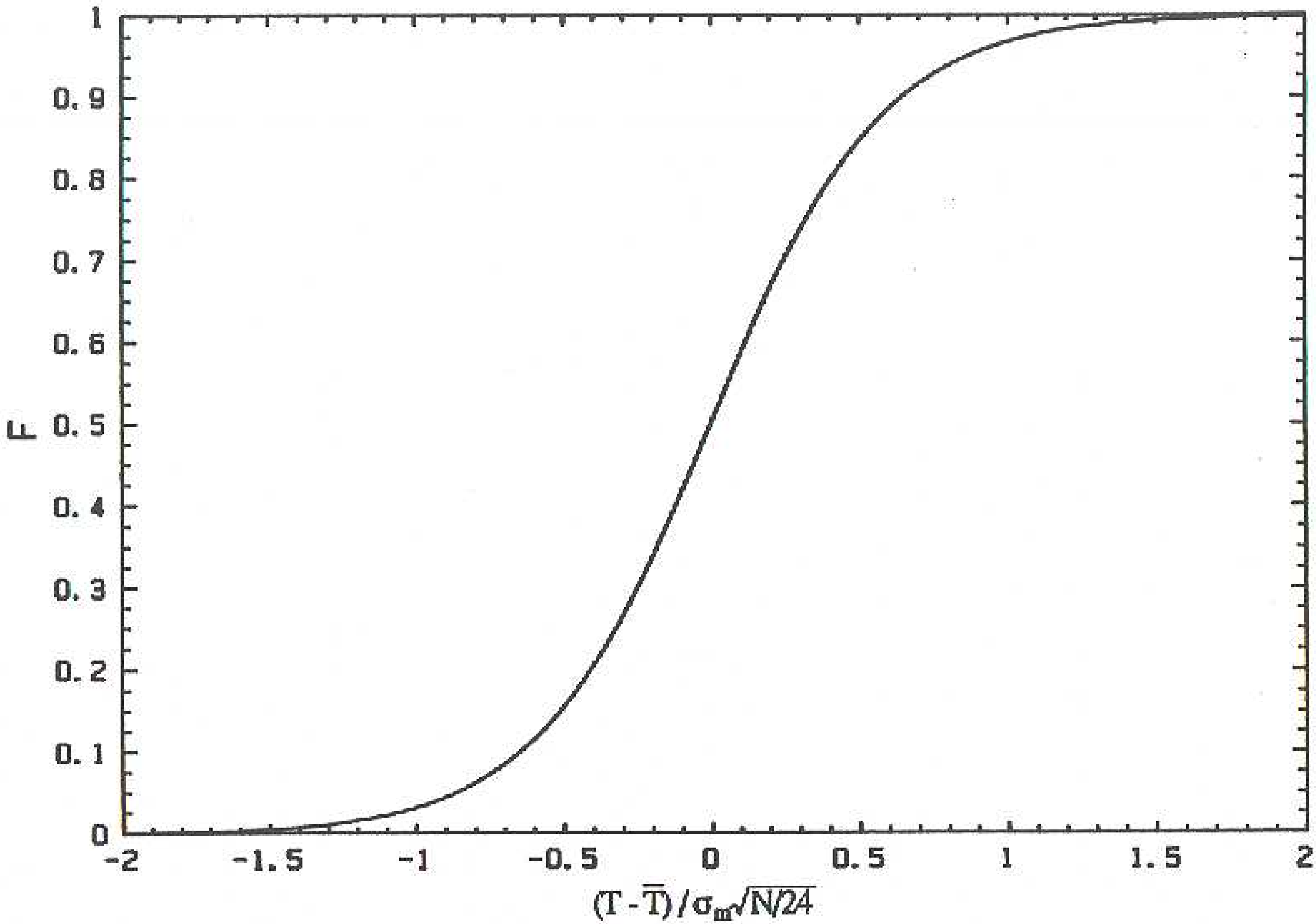

Figure 2.4 Erbs Correlation for the Long-Term Daily-Average Ambient Temperature Cumulative Distribution

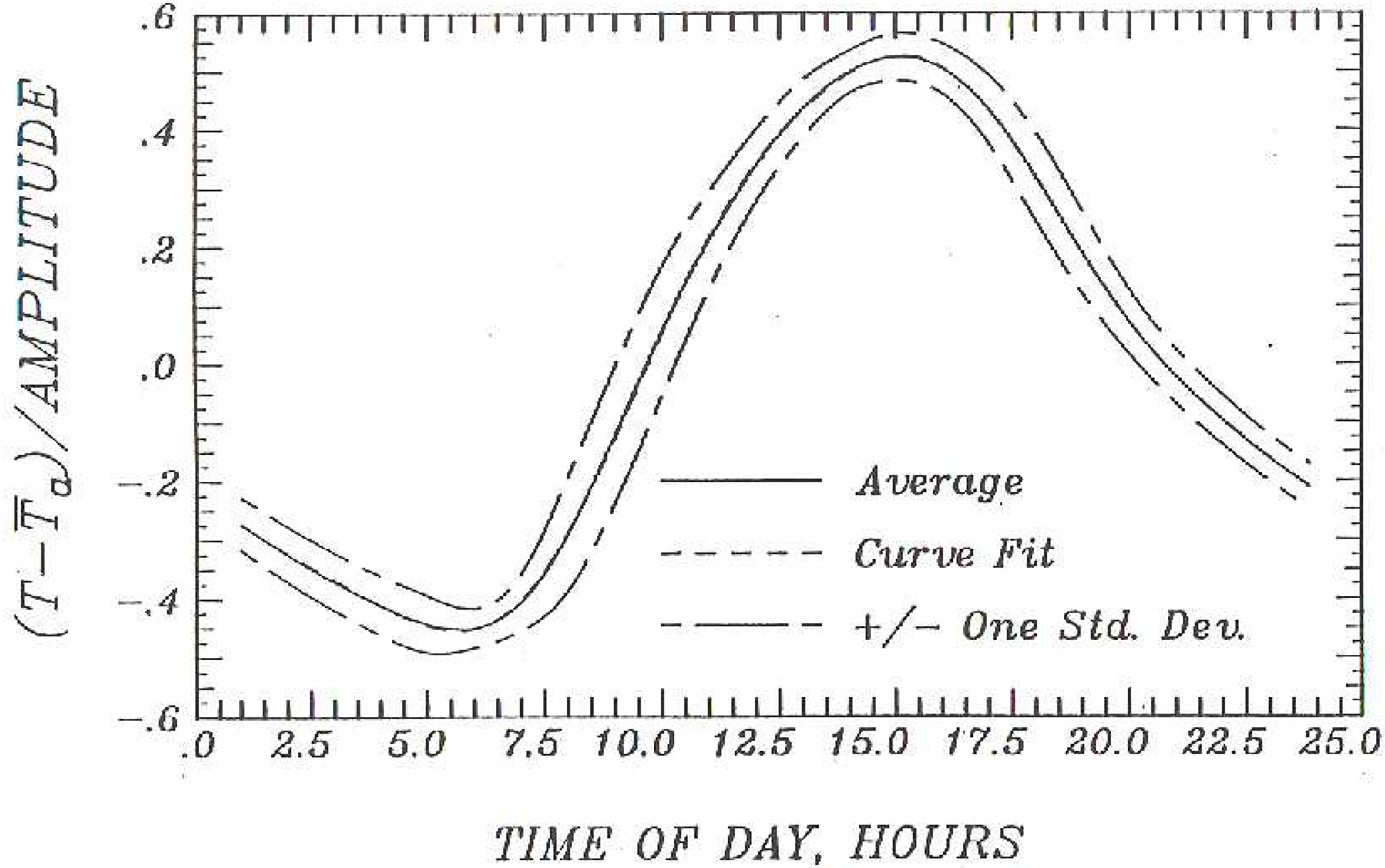

Figure 2.5 Erbs Correlation for the Monthly-Average Diurnal Variation of Ambient Temperature

and dividing by the amplitude (Figure 2.5). Erbs demonstrated that the average normalized diurnal temperature variation is well represented by

$$(\overline{T}_h - \overline{T}_m)/A = 0.4632\cos(t^*-3.805) + 0.0984\cos(2t^*-0.360) + 0.0168\cos(3t^*-0.822) + 0.0138\cos(4t^*-3.513) \tag{2.2.8}$$

$$t^* = 2\pi(t-1)/24$$

where $\overline{T}_m$ is the monthly-average daily ambient temperature and t is the hour of the day defined such that t=1 at 1 am and t=24 at midnight. The amplitude, A, while varying considerably with month and location, was shown by Erbs to be related to $\overline{K}_T$

$$A = 25.8\,\overline{K}_T - 5.21 \tag{2.2.9}$$

where the amplitude is in degrees C.

2.2.3 Relative Humidity

Erbs completed analogous work in the area of relative humidity, developing location-independent expressions for both the distribution and diurnal variation.

The distribution of hourly relative humidity about the monthly-average relative humidity, similar to that of daily clearness index about the monthly-average clearness index, was found by Erbs[1984] to be primarily dependent upon the monthly-average relative humidity. Slight variations in skewness were observed, however the variation was not regarded as significant enough to warrant inclusion in the model. A Wiebull distribution was fit by Erbs to approximately 22 years of data collected from 9 U.S. locations (Figure 2.6). Since relative humidity values only vary between 0 and 1, the integral of the pdf between the limits 0 and 1 must be equal to one, rather than between the limits negative infinity and positive infinity. Erbs included in the expression the division by a parameter-dependent constant in order to force the integral of the pdf between the limits 0 and 1 to be one.

$$F_{RH} = \frac{1 - \exp\left[-(RH)/\theta_1)^{\theta_2}\right]}{1 - \exp\left[-(1/\theta_1)^{\theta_2}\right]} \qquad (2.2.11)$$

where

$$\theta_1 = -0.02691 + 1.2276\,\overline{RH} - 0.14880\,\overline{RH}^2$$

$$\theta_2 = 0.08165 \exp(5.3801\,\overline{RH}) + 2.2747 \exp(-0.59958\,\overline{RH})$$

For the case in which the variable (of whose pdf we are interested) is constrained to vary between 0 and 1 (such as for relative humidity and clearness index), it can also be shown that the integral of the cdf between 0 and 1 is equal to 1 minus the mean, or alternatively, the mean is equal to 1 minus the integral of the cdf between 0 and 1 In other words, when the cdf is plotted with the variable, for example RH, as the independent variable and the cumulative fraction of occurrence, F, as the dependent variable, the area *above* the curve is equal to the mean value of that curve. Possibly due to the difficulties involved in fitting one set of curves to many slightly different ones, the Wiebull distribution when fit, was apparently not constrained so as to ensure the correct mean. That is, the area above the curve (and hence the mean value

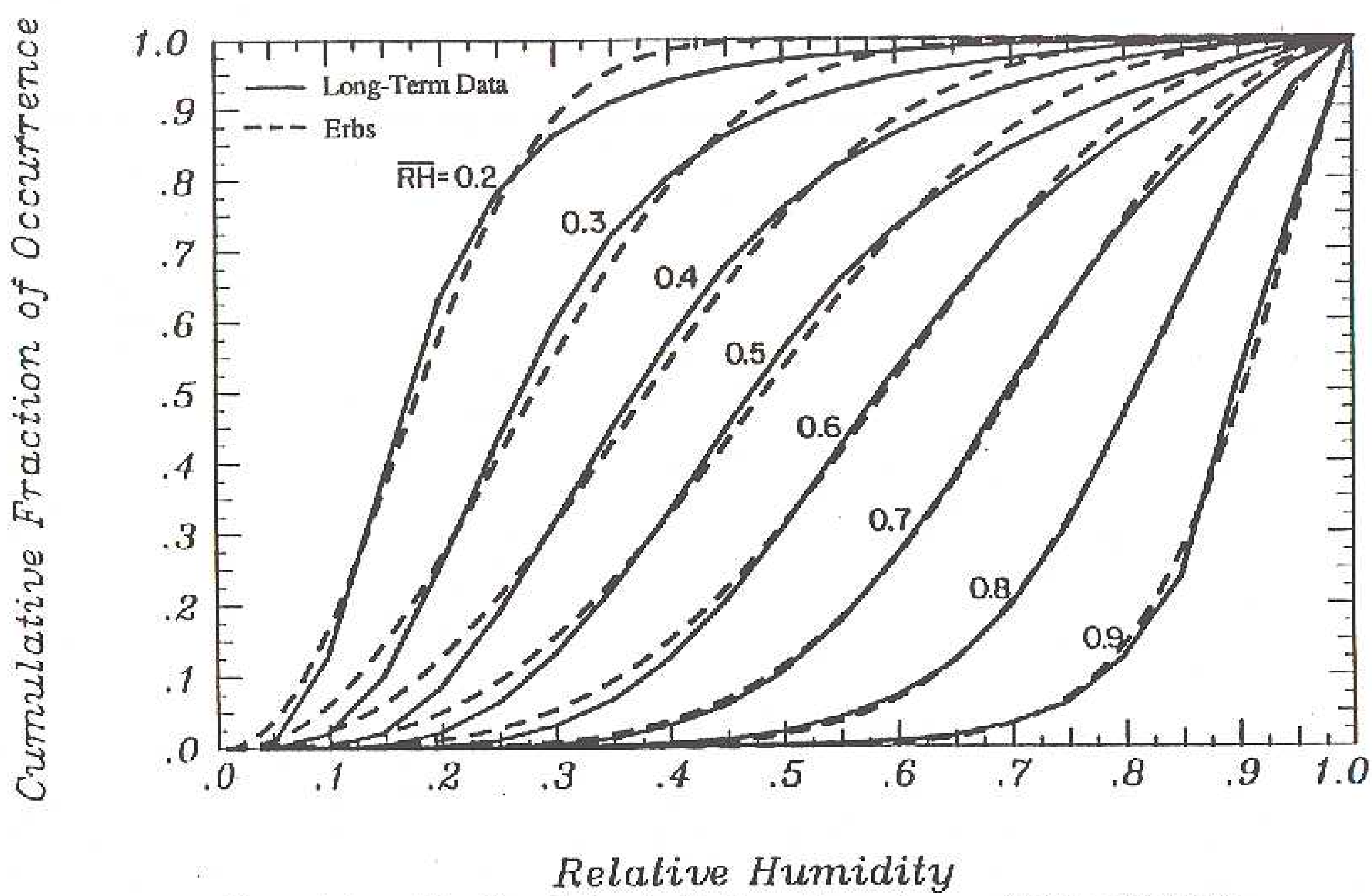

Figure 2.6 Erbs Correlation for the Long-Term Average Relative Humidity Distribution

of that curve) is not exactly the mean associated with that curve. This can be observed in Figure 2.6, taken from Erbs[1984], where the area under the curve representing the data and that representing the correlation are not the same. This discrepancy is particularly noticeable for the curve representing a monthly-average relative humidities of 0.6. The difference in the means is summarized in Table 2.3, where the correlation values were obtained by numerically integrating the expression for the cdf.

Table 2.3
Comparison of monthly-average relative humidities to the monthly means obtained from Erbs' distribution expression

Monthly-average	Correlation Mean
0.10	0.084
0.20	0.188
0.30	0.291
0.40	0.390
0.50	0.488
0.60	0.587
0.70	0.691
0.80	0.796
0.90	0.888

While no study of the distribution of daily-average relative humidity about the monthly-average relative humidity was completed by Erbs, it would be reasonable to expect that the daily distribution would have the same shape as the hourly distribution, as this was observed for both clearness index and ambient temperature. In this study, the distribution of the daily-average relative humidity will be compared

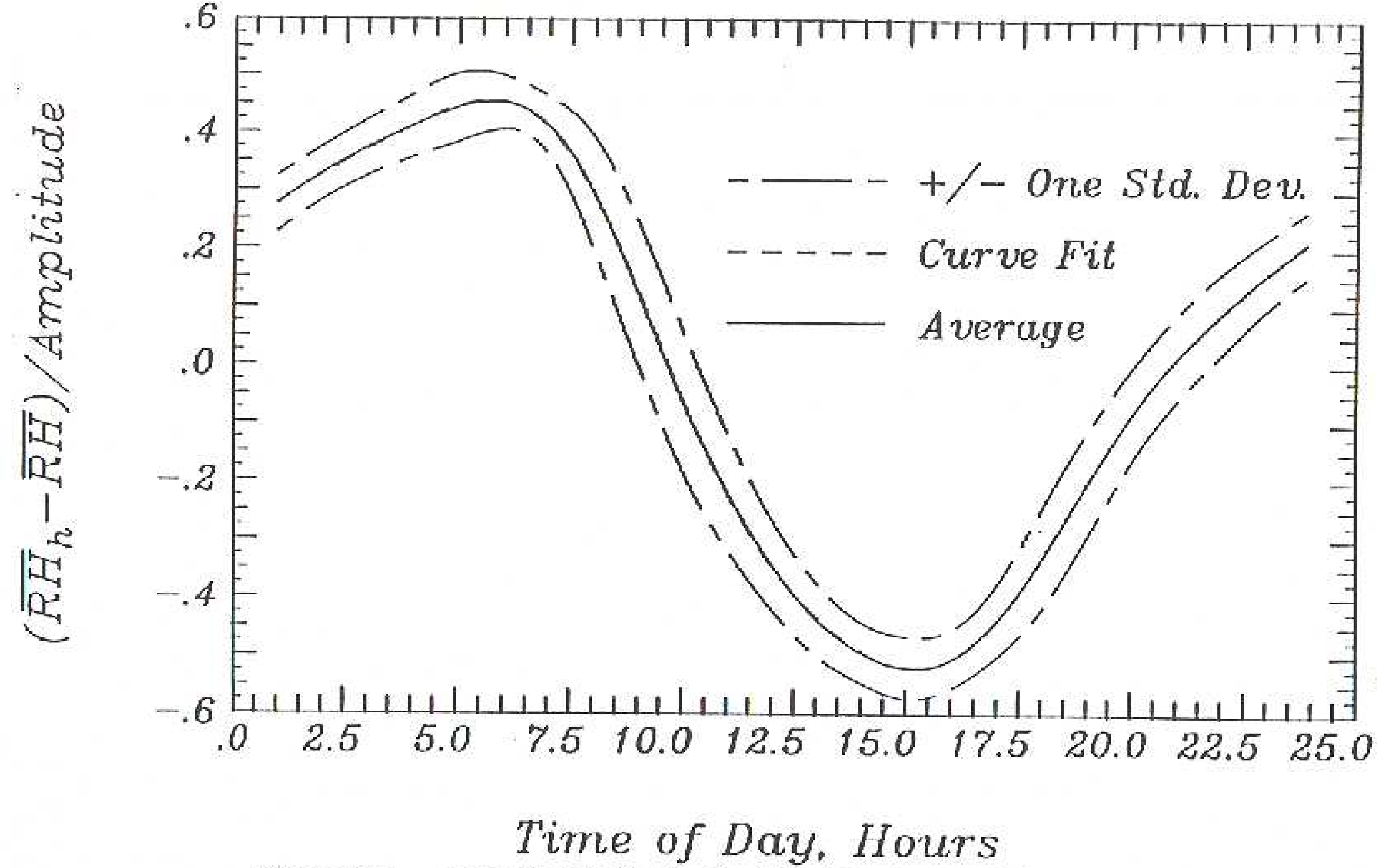

Figure 2.7 Erbs Correlation for the Monthly-Average Diurnal Variation of Relative Humidity

to the hourly expression of Equation (2.2.10) to see if this is indeed the case.

Erbs also demonstrated that the monthly-average diurnal variation of relative humidity, which varied in amplitude and mean for different months and locations, could be represented by a single correlation by subtracting the mean and dividing by the amplitude, as was done with the diurnal variation of ambient temperature (Figure 2.7).

$$(\overline{RH} - \overline{RH}_m)/A_{RH} = 0.4672\cos(t^*-0.666) + 0.0958\cos(2t^*-3.484) + 0.0195\cos(3t^*-4.147) + 0.0147\cos(4t^*-0.452) \qquad (2.2.11)$$

where

$$t^* = 2\pi(t-1)/24$$

The amplitude, A_{RH}, was shown by Erbs to be expressible as a function of the monthly-average ambient temperature and monthly-average clearness index:

$$A_{RH} = -0.516 + 1.933\overline{K}_T - 1.663\overline{K}_T^3 + 0.00669\overline{T}_m - 1.993\times10^{-4}\overline{T}_m^2 \qquad (2.2.12)$$

where $\overline{T}_m$ is in degrees C.

2.2.4 Windspeed

The distribution of hourly windspeed has been investigated by various authors, [Corotis,1977; Exell,1985], and is often described by a Weibull distribution with the parameters estimated from the mean and variance. A different approach is that taken by Balouktsis et al[1986] in which the hourly and daily windspeed values are transformed and analyzed. The daily windspeed is transformed by subtracting the mean value for the day and dividing by the standard deviation for the day, both of which are obtained by interpolating between the monthly means and standard deviations. A Weibull distribution with constant parameters was found sufficient to represent this distribution. The hourly values were transformed in a similar manner, however including a factor to remove the diurnal cycle; this distribution was found to be well approximated by a Gaussian distribution.

2.2.5 Cross-correlations

Cross-correlations between hourly values of ambient temperature and clearness index (both after subtracting their hourly monthly-average values), and ambient temperature and relative humidity (again after subtracting their hourly monthly-average values) were tabulated by Erbs[1984].

In 7 out of the 9 U.S. locations studied by Erbs, negative cross-correlations between k_t and ambient temperature were observed for the winter months, changing to positive correlations for the rest of the year. For some months and some locations, cross-correlation values were significant, while for others they were not.

The cross-correlation of hourly ambient temperature and relative humidity was observed to be negative during the summer months, however the magnitudes varied considerably from place to place. For the other months, the magnitudes were generally small, however there was no consistency in the signs of the cross-correlation values.

Degelman[1976] computed various daily cross-correlation values for Columbia, MO. Significant positive correlations between solar radiation and dew point depression, solar radiation and ambient temperature amplitude were observed. High cross-correlations were also observed between the various temperature variables.

In this study, daily cross-correlation values for the three locations were calculated from the 22 years of weather data and will be analyzed in the comparison sections at the end of each chapter.

Due to the lack of any observed location-independent behavior, no attempt has been made to model any cross-correlations other than what is achieved coincidentally from the independent models.

CHAPTER 3
Radiation Generation

This chapter presents the radiation portion of the generation process, including a description of both Degelman's technique and the modifications developed to improve it. Results of the generation process, long-term data and TMY data for the Albuquerque, NM, Madison, WI, and New York, NY are exhibited and discussed. The radiation model consists of two major parts. First the daily radiation is generated and then it is broken down into hourly values. This is the manner in which it will be presented.

3.1 Degelman Radiation Model

3.1.1 Daily Radiation Generation

Once the cumulative distribution function for a variable is known, it becomes relatively easy to generate a specified number of values from that distribution. Specifically, once the cumulative distribution function of the daily K_t values is known, a value for each day of a month can be found from the distribution; recalling Chapter 2, all that is required to determine the distribution function is the monthly-average daily clearness index [Liu and Jordan,1960], $\bar{K}_T$, which is a required input for generation. A K_t value is needed for each day of a month; or in other words, N K_t values are required for each month, where N is the number of days in the month (31,30 or 28). To truly represent the long-term conditions and make up and "average" month, the K_t values must be selected from the known long-term distribution and must themselves make up a replicate of that long-term distribution.

As described in Chapter 2, the cumulative distribution function relates the cumulative fraction of occurrence, F, to K_t. The cumulative fraction of occurrence specifies the fraction of time the K_t variable will be less than a specified value of K_t. Assuming a 31-day month, there is a value of K_t corresponding to F equal to 1/31, meaning that only 1/31 of the time will a K_t value less than a some particular value occur, or alternatively, only 1 out of 31 days will be less than some particular value of K_t. Additionally, 1 out of 31 days *must* be less than this value of K_t. Where within the range from 0 to this value of K_t is unspecified; a logical way to choose a value is to take the value of K_t at the average of this F-value (1/31) and the previous one (0), for example, F=1/62. The K_t value corresponding to this F-value can then be found from the expression for the cumulative distribution function. Likewise, one and only one K_t value will occur between the K_t value associated with F=1/31 and F=2/31; the K_t value corresponding to an average F value of F=3/62 is found, and so on. 31 days can be generated in this manner. This analysis can easily be extended to a 30 or 28 day month.

While this technique provides N values of K_t for a month, it does not specify the order in which the K_t values should occur. They should not occur in either ascending or descending order, yet neither should they be ordered randomly. As noted in Chapter 2, the lag one autocorrelation of daily K_t values is generally in the range of 0.15 to 0.35, an indication of weak positive correlation. Degelman's approach capitalizes on this similarity in autocorrelation over all locations, and fixes the order in which the days occur so as to approximate the correct lag one daily K_t

autocorrelation. Specifically, the numbers 1 to 31 are assigned to the 31 values obtained from the daily "source" distribution (the Liu and Jordan distribution), with 1 corresponding to the smallest K_t value and 31 to the largest. The numbers 1 to 31 are then placed in an order such that when the K_t values corresponding to the numbers are placed in that order, the lag one autocorrelation value is obtained. Figure 3.1 illustrates this process for a 5 day month.

The same sequence of the numbers 1 to 31 is always used for ordering the daily values of K_t, however, the starting position within the sequence is determined randomly at the beginning of the generation process. For months of other than 31 days, only that number of days (i.e., 28 or 30) are obtained *from the 31-day distribution*, resulting in slight differences between the generated and "source" distribution for these months. The method whereby the daily values are obtained from the distribution is a common one; the method of ordering the days is unique to Degelman. Degelman's method allows the generation of one "long-term" month, falling in with the goals of the "average", or Type II generator, namely that each month has the long-term mean, distribution, and autocorrelation.

3.1.2 Critique of Degelman's Daily Radiation Generation

Several problems were apparent upon examination of the daily portion of the Degelman radiation generator. First, and foremost, his work was completed prior to the work of Bendt et al [1981] and Hollands and Huget [1983] in which expressions for the Liu and Jordan distribution were developed. The form of the weather generator being examined contained an approximation to the Liu and Jordan distribution based somewhat upon the normal distribution. As no reference to the

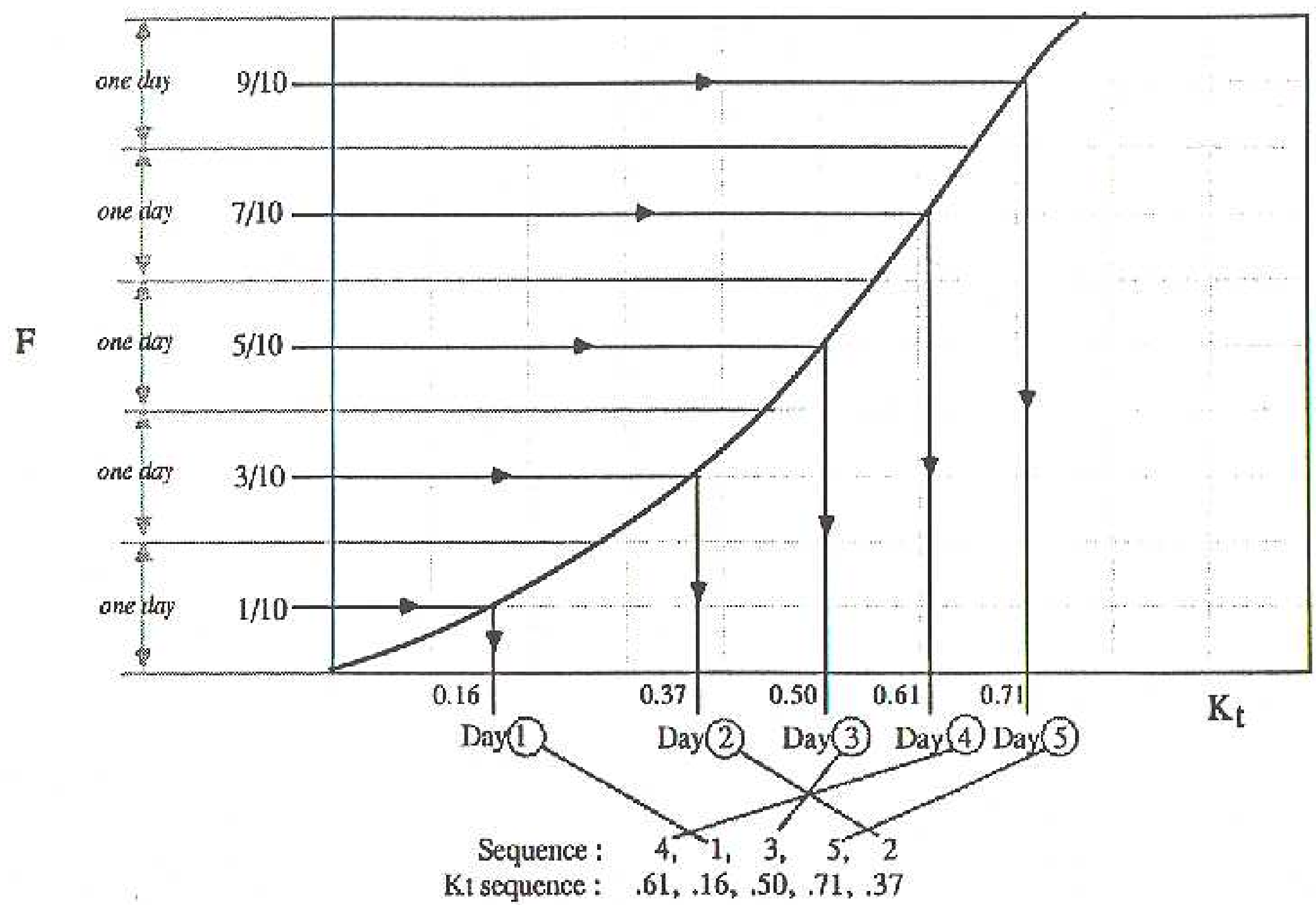

Figure 3.1 Obtaining the Daily K_t Values from the Distribution in a Predetermined Order for a Five Day Month

origin of the algorithm could be found, it was assumed to be an approximation developed by Degelman. While it reasonably reproduces the Liu and Jordan distribution for months with $\overline{K}_T$'s of 0.5 and 0.6, significant differences are apparent for other values of $\overline{K}_T$ (Figure 3.2). Not only is the shape different, but the mean also differs considerably (Figure 3.3). Degelman's algorithm was subsequently replaced by the Bendt expression [Bendt, 1981].

The lag one daily autocorrelation value generated from Degelman's fixed sequence was observed to be somewhat high while also varying as a function of $\overline{K}_T$ (Table 3.1). The autocorrelation value can be lowered simply by changing the sequence slightly, although it will have some effect upon the crosscorrelation with other weather variables. The variation in autocorrelation with $\overline{K}_T$, however, is not so easily removed. Due to the variation and uncertainty in the long-term autocorrelation values, some variation is appropriate, however, the marked variation as a function of $\overline{K}_T$ is a problem. The reason for this variation lies in the differing shapes of the K_t distribution curves. While the F-values are equally spaced, the K_t values are not. For example, for $\overline{K}_T$ of 0.5, the distribution is near normal in shape, with approximately half of the K_t values less than $\overline{K}_T$ and half greater. For $\overline{K}_T$ equal to 0.7 however, there will be many K_t values just slightly greater than $\overline{K}_T$ and somewhat fewer K_t values less than $\overline{K}_T$, although they will be farther from $\overline{K}_T$ in value; this different "spacing" would tend to impart a different autocorrelation value. As no analyzed technique for obtaining a sequence of the numbers 1 to 31 as a function of $\overline{K}_T$ has been found, a trial and error method has

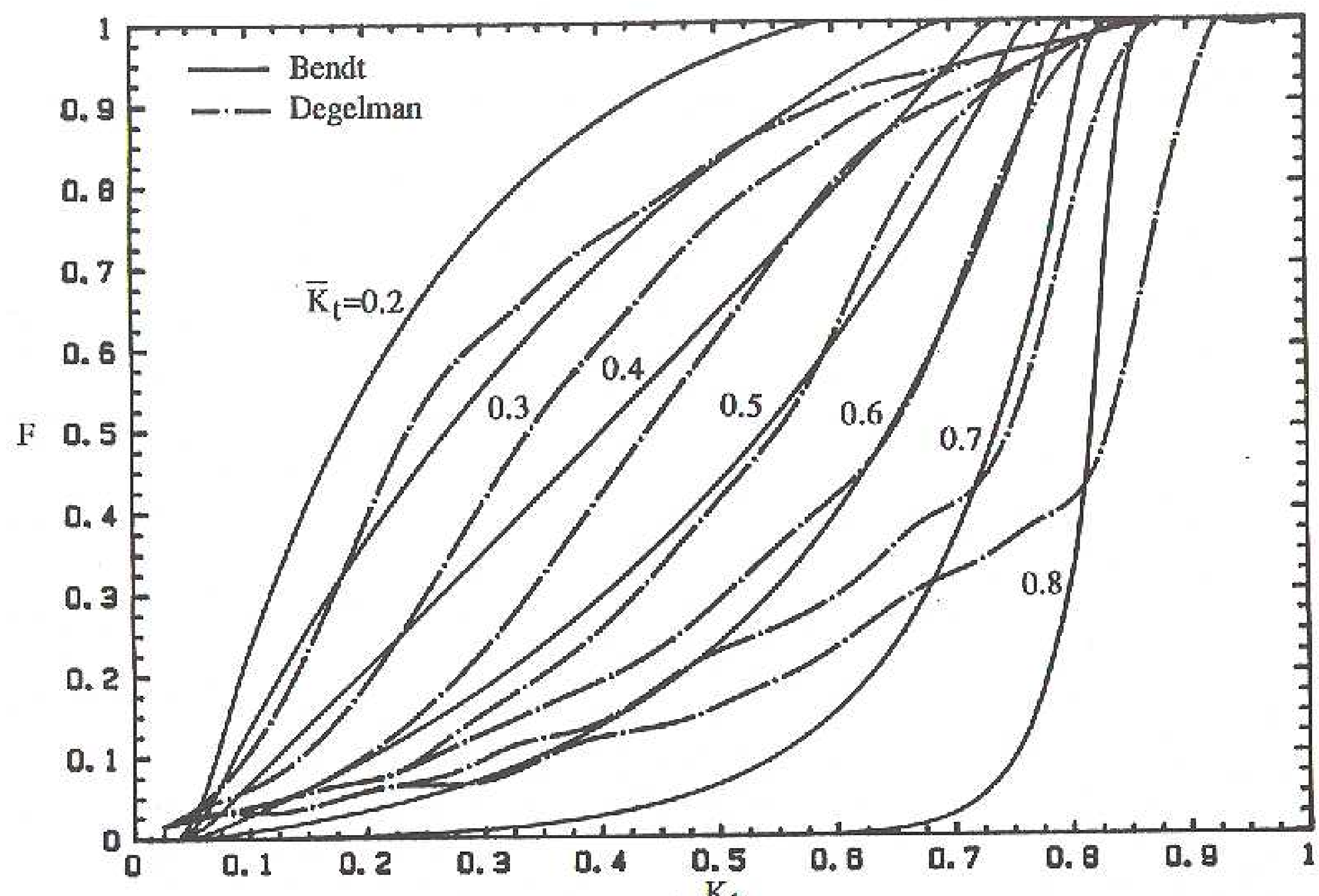

Figure 3.2 Comparison of Degelman's K_t Distribution to the Bendt Correlation for the Liu and Jordan Distribution

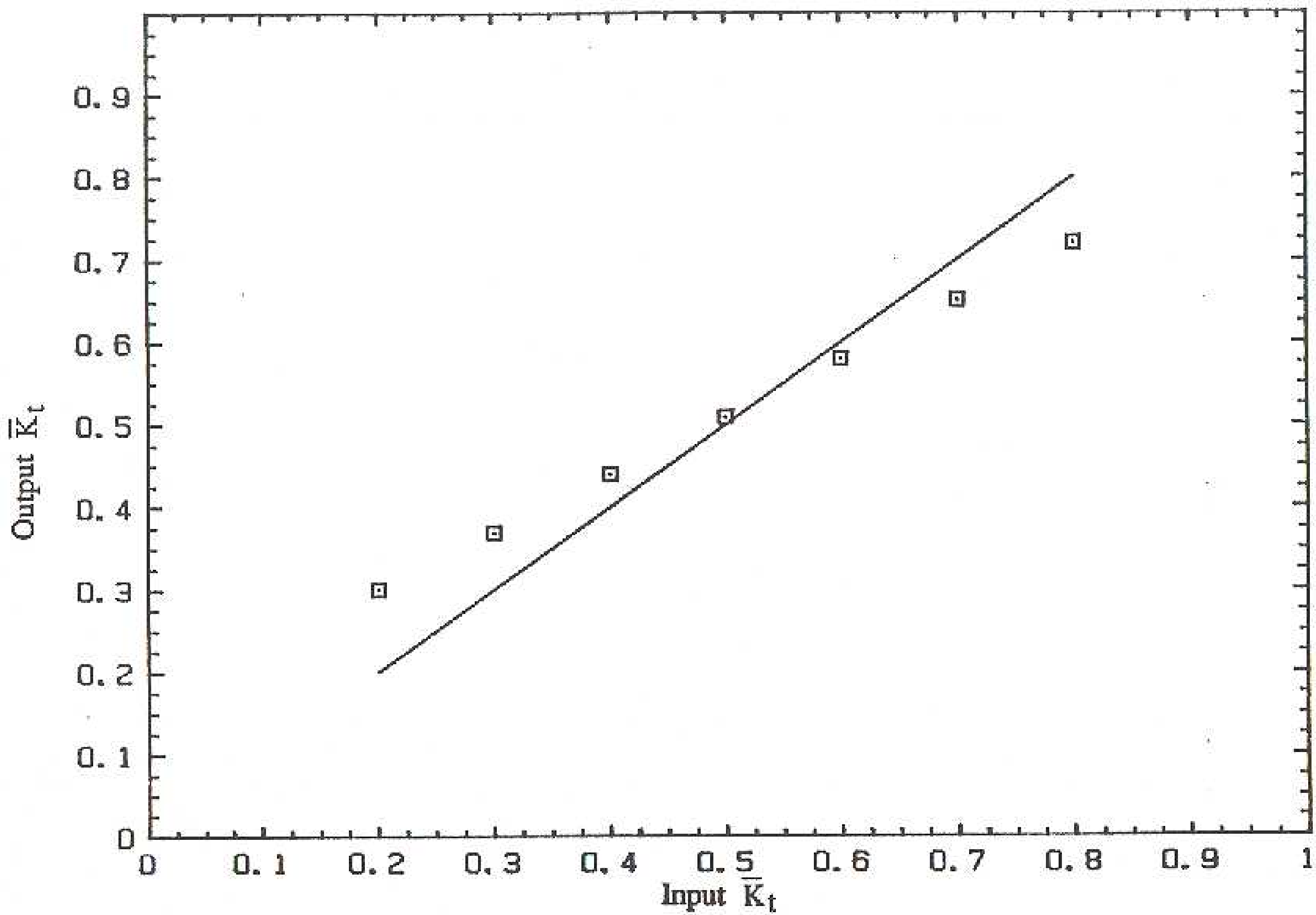

Figure 3.3 $\overline{K}_t$ Values As Obtained from Degelman's Expression for the K_t Distribution

Table 3.1
Variation of lag one daily autocorrelation when K_t values are ordered by a single sequence

K_t	Lag One Autocorrelation
0.2	0.28
0.3	0.27
0.4	0.31
0.5	0.39
0.6	0.44
0.7	0.44
0.8	0.44

been used to obtain three sequences to be used for three ranges of $\overline{K}_T$, namely 0.0 to 0.45, 0.45 to 0.55, and 0.55 to 1.0. These three sequences eliminate, to a large extent, the variation of the autocorrelation with $\overline{K}_T$.

Another minor change made to Degelman's original daily K_t generation process is the use of a 28 or 30 day distribution for those months of length other than 31 days, rather than the 31-day distribution. The original algorithm has a minor problem in that a day number in the sequence could exceed the number of days in the month, for example, the number 31 might be assigned to the sixth day of February, and there would not exist a day 31 for February. If this case arises, the number in the sequence is simply skipped. While this does have a slight affect upon the daily autocorrelation, it is small and considered less important than maintaining the long-term distribution.

3.1.3 Hourly Radiation Generation

Once the daily values of K_t have been generated for each day of the month, the hourly radiation is literally generated one day at a time. Take a day, for example; the daily K_t is known, and now it is necessary to generate hourly values which combine to give that value of K_t.

The general premise of Degelman's hourly radiation generation scheme is to first generate the direct normal radiation for the hour, then the diffuse radiation, and then combine them to obtain the total solar radiation on a horizontal surface for the hour.

The hourly direct normal radiation is calculated from the following commonly used expression:

$$I_{dn} = I_{sc} e^{(-\alpha / \cos\theta_z)} \tag{3.1.1}$$

I_{sc} is the apparent solar constant and θ_z is the zenith angle of the sun; both are easily calculated from knowledge of the latitude, date and time combined with known earth-sun geometry [Duffie and Beckman, 1980].

α in Equation (3.1.1) is the atmospheric extinction coefficient for the hour; it is an indication of the sky condition. Degelman examines the value of K_t for the day, and uses that along with several correlations to calculate two atmospheric extinction coefficients for the day, one for clear sky conditions and one to represent cloudy sky conditions. The cloud cover fraction (CCF) for the day is estimated from:

$$\begin{aligned} CCF &= 1 && K_t < 0.333 \\ CCF &= 2 - 3\,Kt && 0.333 \le K_t \le 0.667 \\ CCF &= 0 && K_t > 0.667 \end{aligned} \qquad (3.1.2)$$

To determine whether the clear or cloudy sky extinction coefficient is to be used for an hour, each hour, a random number between 0 and 1 is selected; if it is greater than the CCF for the day, the clear sky coefficient is used, else the cloudy sky coefficient is chosen for the hour. For very high (K_t>0.667) and very low (K_t<0.333) values of K_t, the two extinction coefficients are assumed identical, resulting in a constant sky condition throughout the day. For a more detailed explanation of the process by which the extinction coefficients are determined, see [Degelman,1970].

The relations between the *daily* diffuse and *daily* direct normal transmittance coefficients for cloudy days presented by Liu and Jordan [1960] are applied by Degelman on an *hourly* basis to obtain the diffuse radiation each hour. The diffuse transmittance coefficient (K_d) is defined as the ratio of the diffuse radiation on a horizontal surface to the extraterrestrial radiation on a horizontal surface; the direct transmittance coefficient (K_D) is defined as either the ratio of the direct (or beam) radiation on a horizontal surface to the extraterrestrial radiation on a horizontal surface or as the ratio of the direct normal radiation to the extraterrestrial normal radiation. Liu and Jordan [1960] presented a plot of K_d versus K_D for cloudy skies, from which Degelman developed the following relations:

$$K_d = 1.15\,(0.315 - 0.29K_D) \qquad K_D \ge 0.4$$

$$K_d = 1.15\,(13.3798K_D^3 - 10.679K_D^2 + 2.415K_D + 0.094) \qquad K_D < 0.4 \qquad (3.1.3)$$

Equation (3.1.3) is the equation used by Degelman to estimate the hourly diffuse transmittance coefficient from the direct transmittance coefficient; the values obtained are significantly different from the results presented by Liu and Jordan (Figure 3.4).

The diffuse radiation for the hour is obtained simply by multiplying the diffuse transmittance coefficient by the extraterrestrial radiation on a horizontal surface (I_o); likewise the direct normal radiation for the hour is obtained by multiplying the direct transmittance coefficient by the extraterrestrial normal radiation for the hour ($I_{o,n}$). The direct normal radiation (I_{dn}) is converted to the direct (or beam) radiation on a horizontal surface and summed with the diffuse radiation (I_d) to yield the total radiation for the hour.

$$I = I_d + I_{dn}\cos\theta_z \qquad (3.1.4)$$

It is important to note that once the hourly radiation values have been generated for a day and are consequently summed to yield the day's total radiation, their sum may not be equal to the daily value previously generated in the daily generation portion of the program.

3.1.4 Critique of the Degelman Hourly Radiation Generation

The single most important statistic to reproduce is undoubtedly the monthly-average mean. To check the accuracy of the radiation model, January radiation data were generated for input $\overline{K}_T$ values ranging from 0.2 to 0.8 at latitudes of 10, 20,

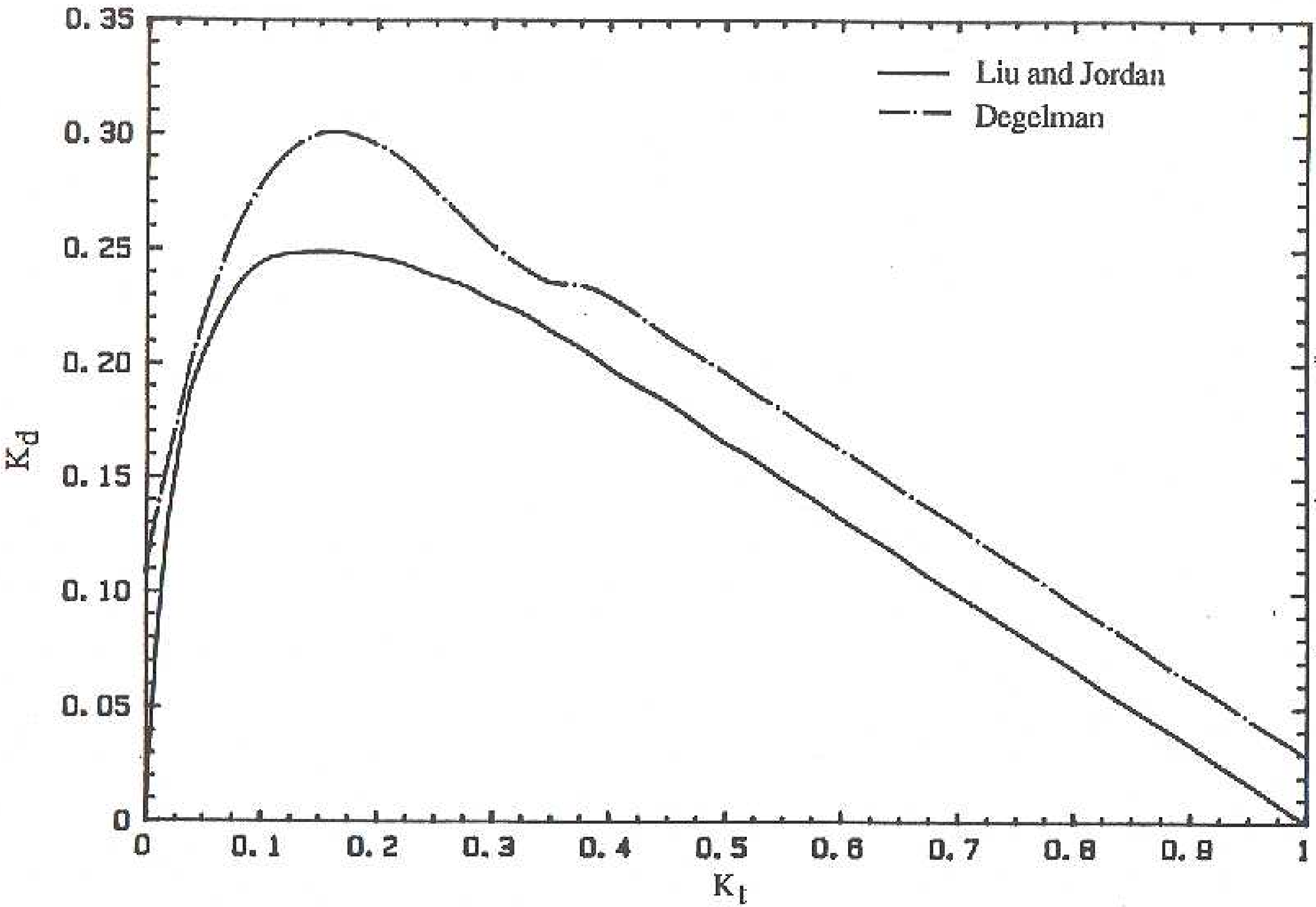

Figure 3.4 Comparison of Degelman's Diffuse Transmittance Correlation to the Liu and Jordan [1960] Curve

30, 40, 50, and 60 degrees. The hourly radiation values were summed to compute the generated $\overline{K}_T$ values. A plot of the generated versus input $\overline{K}_T$ (Figure 3.5) revealed not only errors in the means, but additionally a dependence of the error on latitude. As was previously demonstrated in Figures 3.2 and 3.3, the form of the equation which Degelman was using to represent the Liu and Jordan K_t distribution curves used in generating the daily K_t values was inaccurate, and this error is reflected to a great extent in Figure 3.5. Data were obtained in a similar manner, however this time substituting the expression of Bendt et al [1981] for Degelman's equation for the K_t distribution; it was expected that this would correct a large portion of the error. As can be seen in Figure 3.6, this is indeed the case. There are, however, still significant differences between the input and generated values of $\overline{K}_T$ at a latitude of 60.

The average diurnal variation of the total radiation on a horizontal surface (divided by the daily total radiation on a horizontal surface) of the Degelman generated data for Madison, WI can be seen in Figure 3.7. Each point represents the generated monthly-average value for that hour; the solid lines are the correlation for the long-term average [Collares-Pereira and Rabl, 1979]. While at a first glance the data appear to agree well with the correlation, further study shows a definite bias; the values at the hours around noon are too high, while the values at the early and late hours are low. This indicates that the Degelman model is proportioning the radiation incorrectly.

The average diurnal variation of the diffuse radiation (divided by the daily diffuse radiation) shows much better agreement with the long-term and is quite

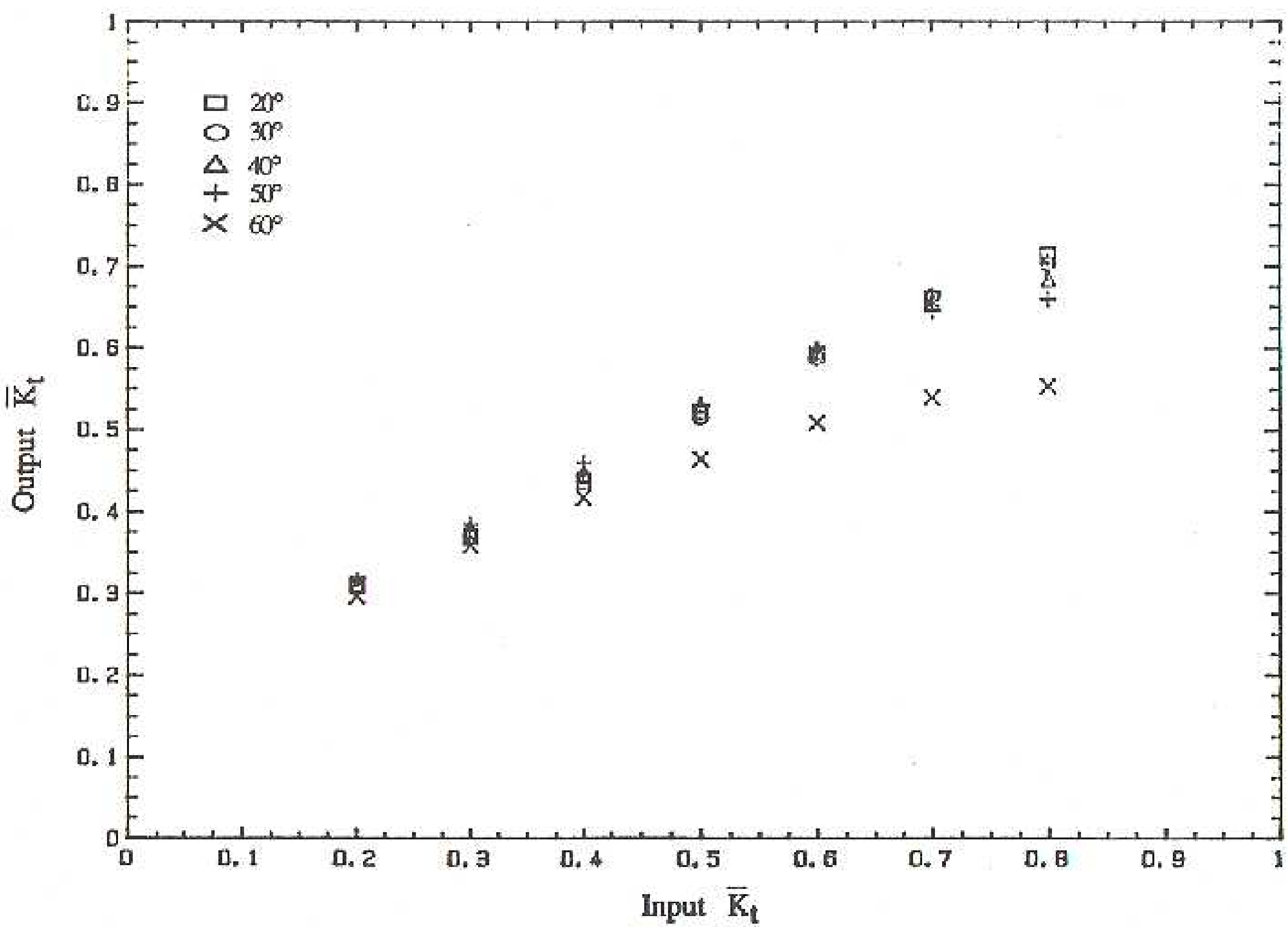

Figure 3.5 Output Versus Input $\overline{K}_T$ from Degelman's Radiation Model

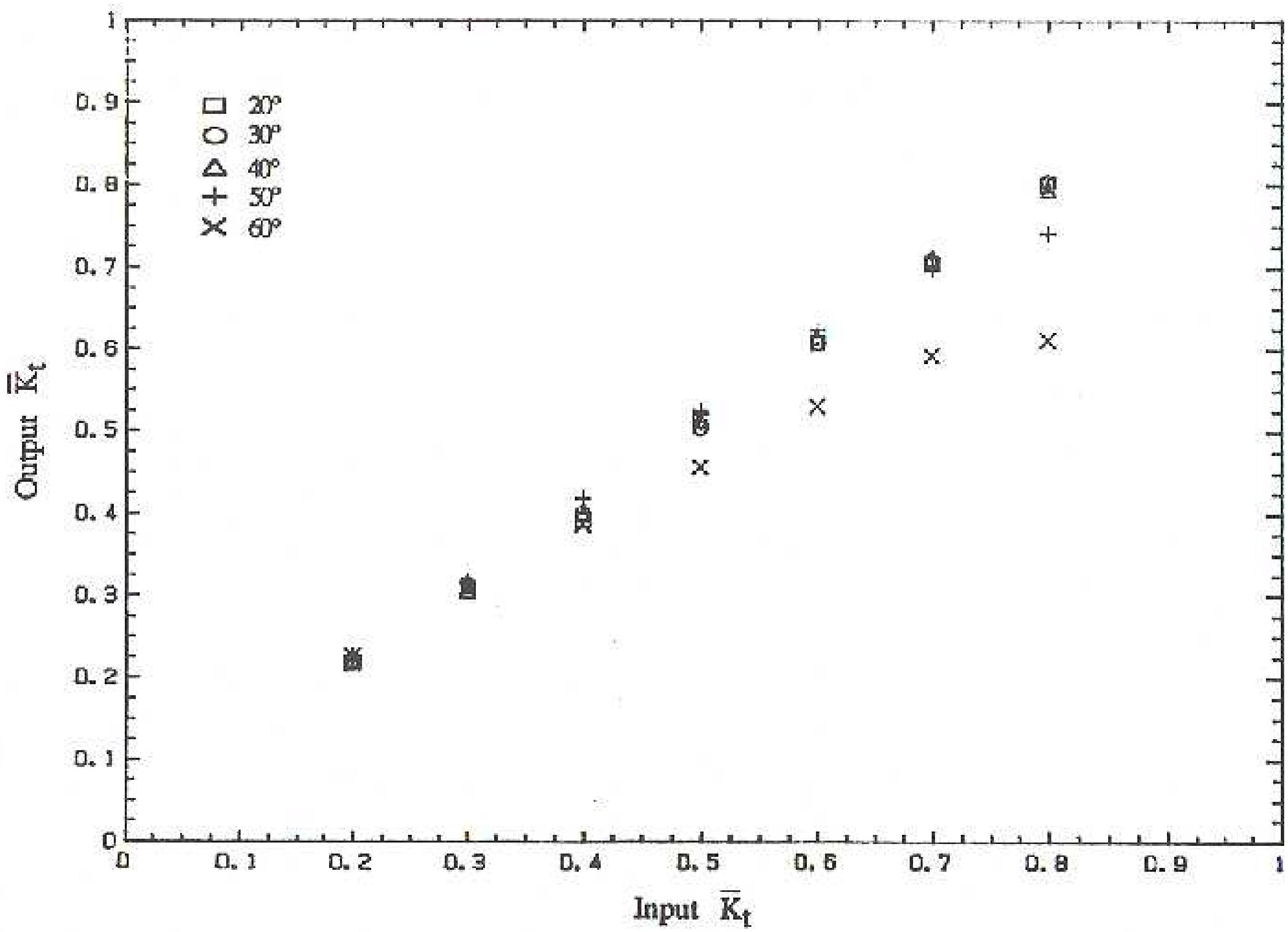

Figure 3.6 Output Versus Input $\overline{K}_T$ from Degelman's Radiation Model with Bendt's Correlation Substituted for the Daily K_t Distribution

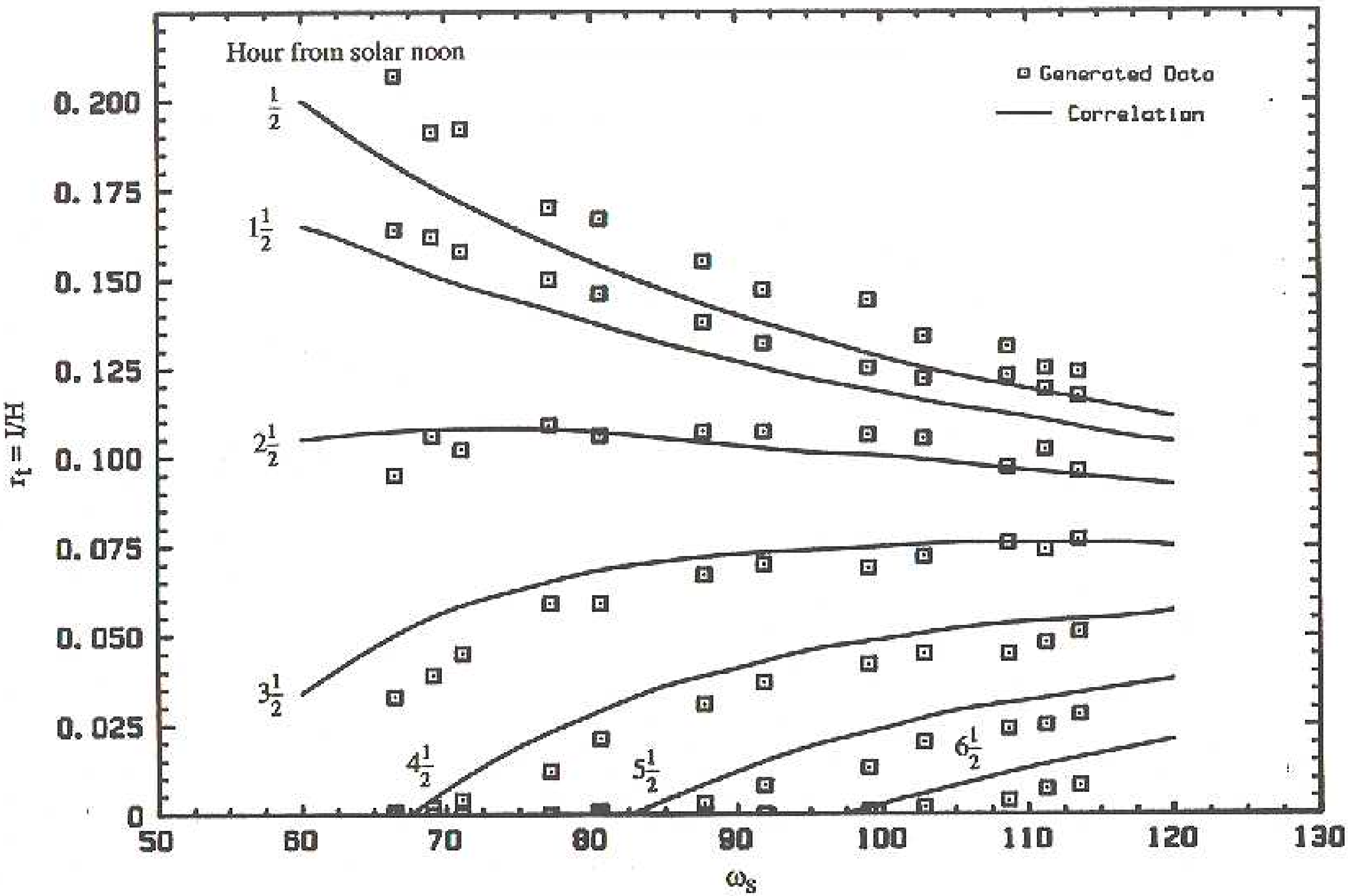

Figure 3.7 Average Diurnal Variation of I/H Obtained from Degelman's Weather Generation Program (Madison, WI)

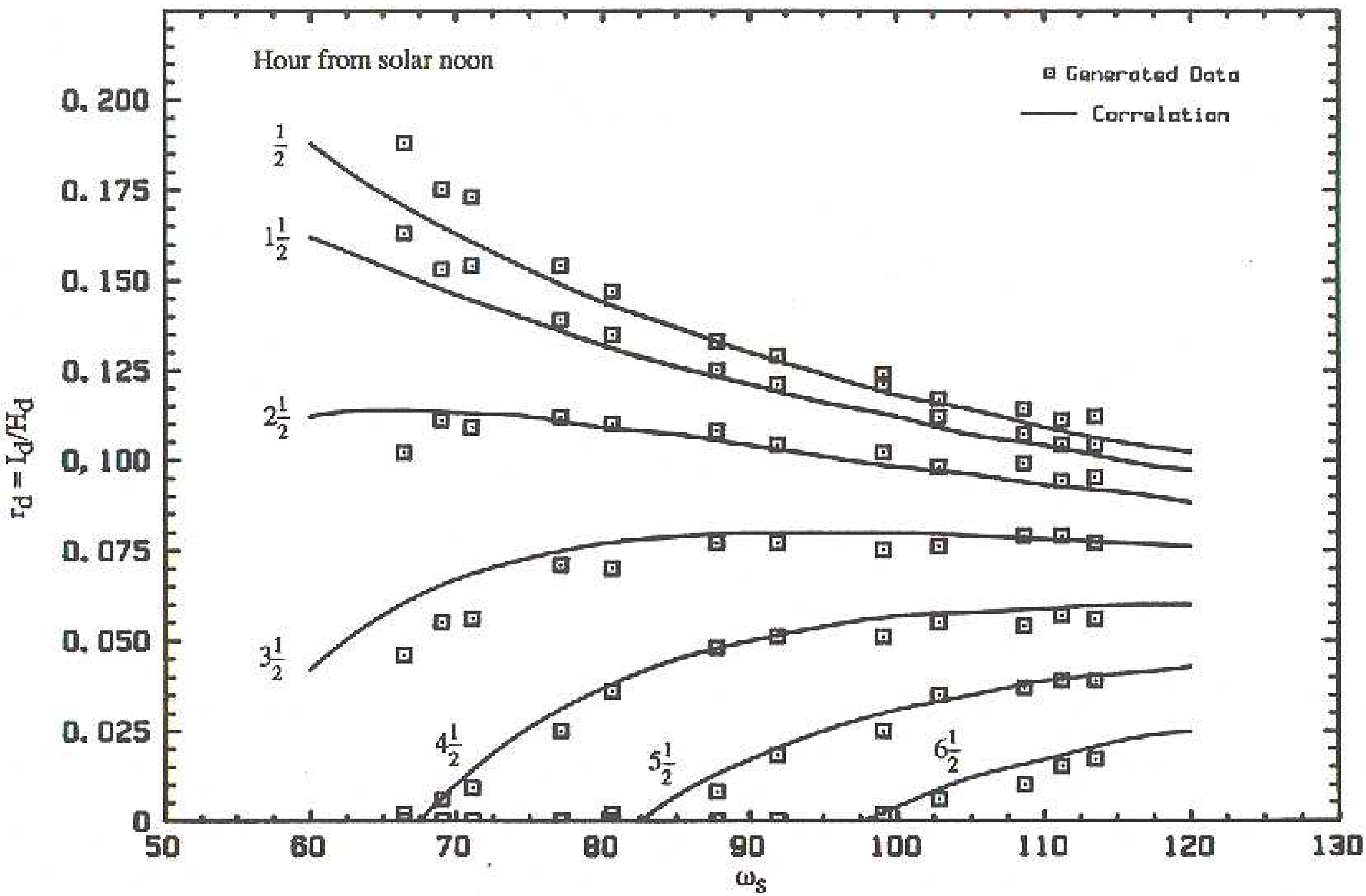

Figure 3.8 Average Diurnal Variation of I_d/H_d Obtained from Degelman's Weather Generation Program (Madison, WI)

satisfactory (Figure 3.8). This indicates a difficult problem within the model. Since the total radiation is obtained by combining the diffuse and direct normal radiation, it is the direct normal radiation which is in error. However, the diffuse radiation values are obtained from the direct normal radiation values, so that an attempt to correct the direct normal radiation would undoubtedly adversely affect the diffuse radiation values. The problem therefore lies in both the direct normal model *and* the diffuse generating component.

While the diurnal variation of the diffuse radiation indicated the correct spread of the diffuse throughout the day, it in no way yielded any information about the total amount of diffuse radiation being generated. Since diffuse radiation is not one of the inputs to the generation model, there is no way to compare it to the input data as was done with the total radiation. An alternative is to compare it with the correlations developed by Erbs[1982] representing the long-term average diffuse fraction (diffuse radiation divided by total radiation) as a function of clearness index, both on an hourly and daily scale.

Figures 3.9 and 3.10 are the hourly and daily diffuse fraction plots, respectively. Both clearly show that the fractional amount of diffuse radiation being generated relative to the clearness index is incorrect, even though the manner in which it is distributed throughout the day (the diurnal variation) is correct. The wrong percentage of an hour's radiation is diffuse, and likewise, the wrong percentage of a day's generated radiation is diffuse.

In addition, the random selection of the atmospheric extinction coefficient each

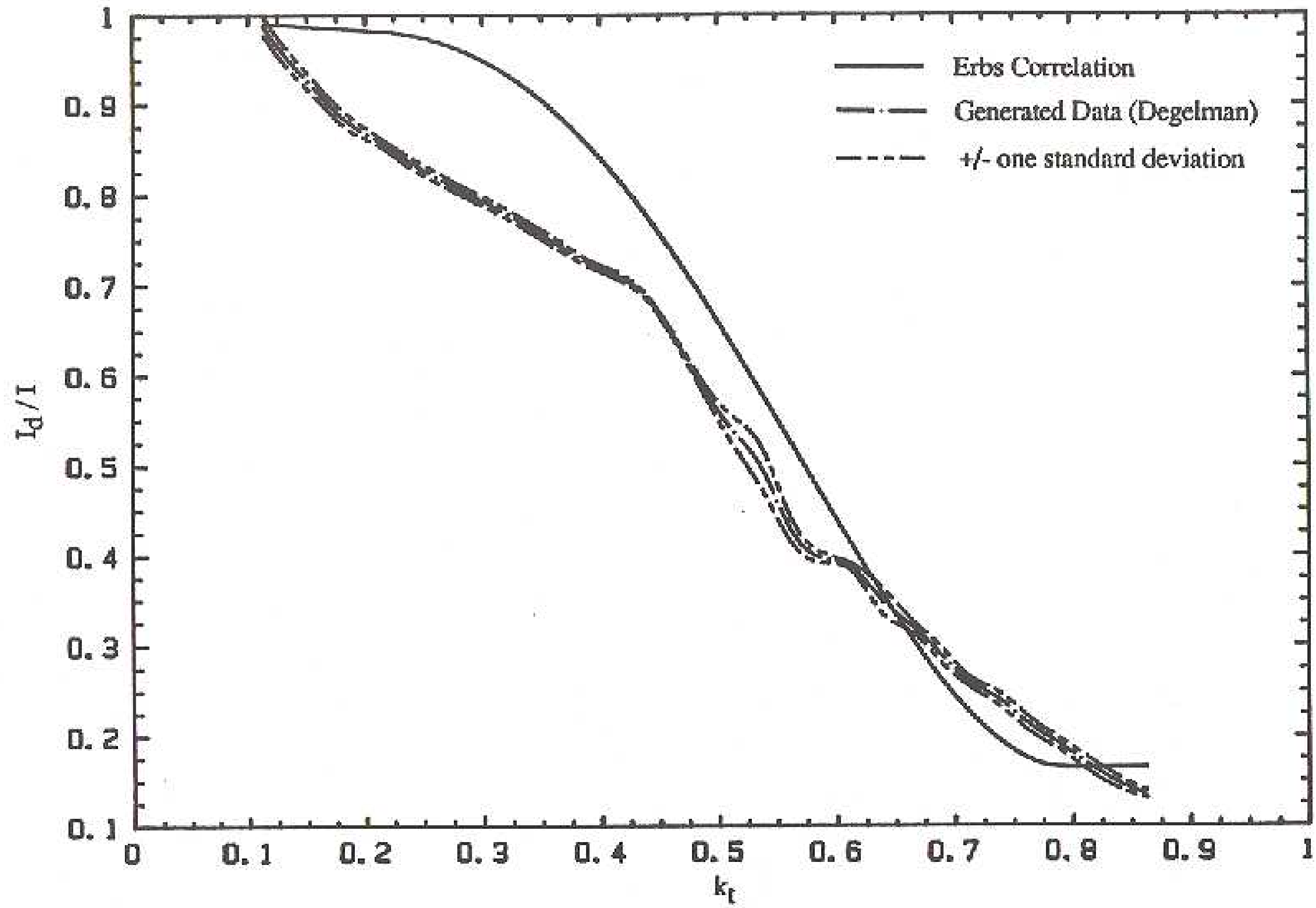

Figure 3.9 Hourly Diffuse Fraction Generated by Degelman's Weather Generation Program (Madison, WI)

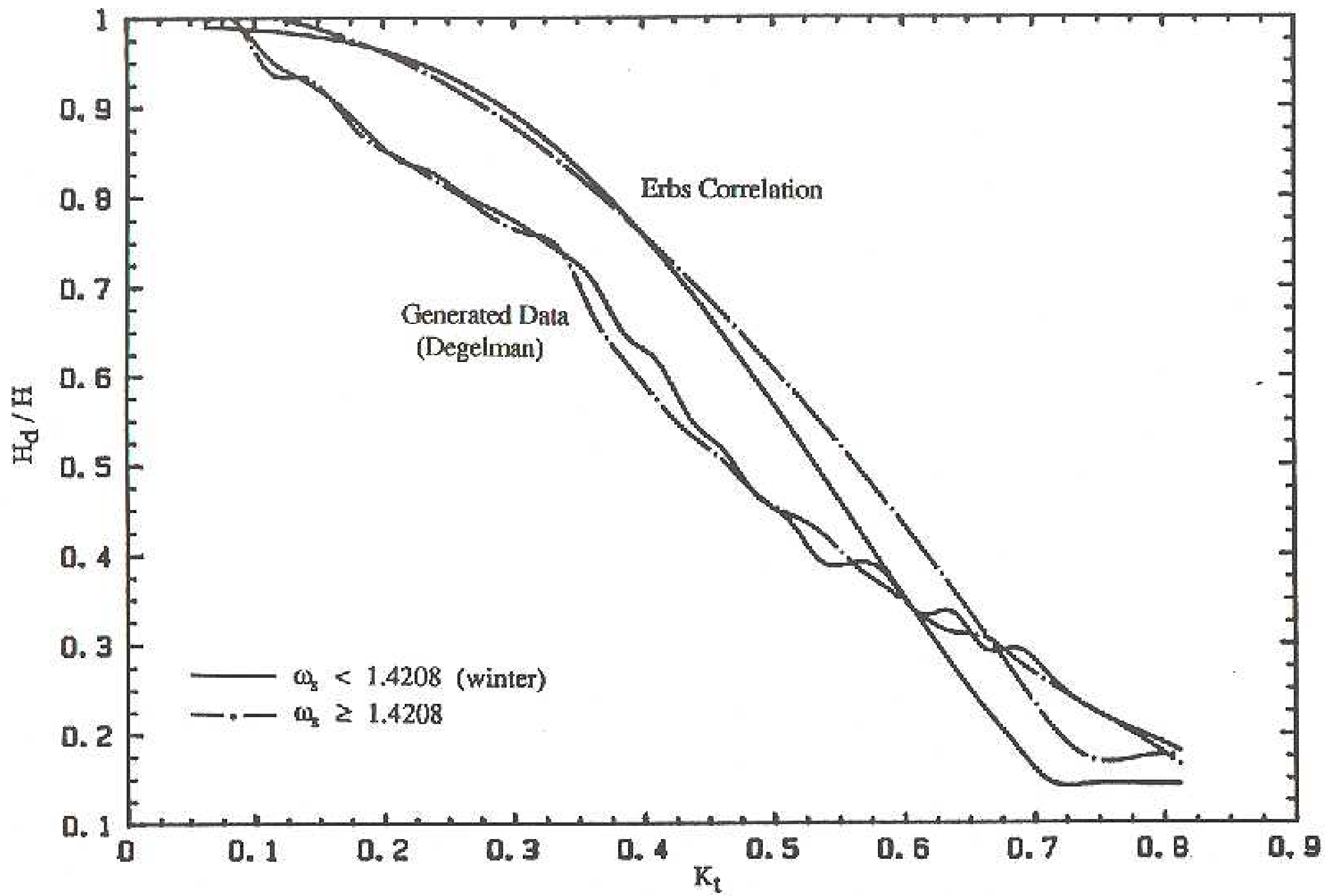

Figure 3.10 Daily Diffuse Fraction Generated by Degelman's Weather Generation Program (Madison, WI)

hour makes no attempt to reproduce any autocorrelation existing between hourly k_t values.

Due to the inconsistencies revealed in the analysis of the generated radiation data, a replacement of the existing hourly radiation model was deemed necessary.

3.2 Modification of the Radiation Model

3.2.1 Model Selection

In selecting an appropriate hourly radiation model to replace Degelman's model, several characteristics were considered, namely:

1. Location independence. The model should be location independent to the extent that all necessary parameters and correlations can be determined from the inputs to the generator program, specifically latitude, monthly-average daily radiation, ambient temperature, and humidity ratio.

2. Simplicity. The model should require as little computation as possible; one of the aims of the weather data generator is to decrease simulation time. However, the amount of required computation would be weighed against any increase in accuracy.

3. Reproducibility of hourly autocorrelation. The use of the long-term average hourly radiation values, such as determined by Equation (2.2.2), can lead to significant performance prediction errors. It is therefore important to maintain some degree of randomness within the hourly radiation structure. In particular, a realistic

representation of the hourly autocorrelation would appear important for solar systems with a critical amount of storage and systems requiring a minimum level of radiation for operation.

The performance of a system which has no storage would have a strong dependence on the distribution of the hourly radiation values, but would be independent of the hourly autocorrelation. Since there is no storage, the previous hour's value has no effect on the present hour's value. If a system had a large amount of storage, any effects of hourly autocorrelation would be dampened out. In between no storage and a large amount of storage, there is some "critical" amount of storage for which performance would be affected by the hourly autocorrelation. While the effect of hourly autocorrelation is probably much smaller than that of daily autocorrelation, it should be realistically represented if possible.

4. Ability to generate hourly values from daily values. While not essential, this allows the retention of the daily portion of Degelman's radiation generation model, which is a convenient way to ensure that each generated month will have the long-term daily radiation distribution and the appropriate daily autocorrelation.

The first objective of obtaining a location independent model suggests finding a model which replicates the hourly k_t sequence rather than the hourly radiation sequence. The amount of incident radiation at a site is location dependent, and would require considerable complexity to model directly. However, the state of the atmosphere, such as indicated by the transmittance (the clearness index), would be expected to possess a more universal character, such that the behavior of the k_t sequence could be modeled independent of location.

Stochastic models, such as the AR and MA models described in Chapter 2, are convenient for reproducing sequences of autocorrelated values. A logical choice therefore is to employ a stochastic model to generate the hourly k_t sequences.

Graham[1985] completed a comprehensive analysis of the K_t sequence, developed a stochastic model for replicating the sequence of daily clearness indices and extended the results to include the sequence of hourly k_t values. Since the K_t variable is not normally distributed [Liu and Jordan, 1960], direct modeling by an AR model results in non-normally distributed residuals. To eliminate this problem, Graham transforms the K_t values to a normally-distributed variable, χ, with a mean of 0 and a variance equal to 1, by equating the cumulative distribution functions and solving for the corresponding value of χ. An AR1 model was found to be an accurate means for reproducing the autocorrelation inherent in the sequence of χ values, in which the autoregressive parameter was either set to a constant value of 0.29 or estimated monthly from Equation (2.2.5). While these parameter values are the lag one autocorrelations of the K_t values rather than the χ values, Graham observed that the lag one autocorrelations of the two series were approximately equivalent, and therefore used the lag one autocorrelation of K_t as an estimate of ϕ_1 rather than the lag one autocorrelation of χ. To generate a sequence of K_t values, as each value of χ is obtained, it is transformed back to the non-normal K_t variable, again by equating the F values of the cumulative distribution functions.

The majority of Graham's work dealt with the development of the model for the daily K_t, however one chapter was devoted to applying an analogous procedure to replicate the hourly k_t sequence; this hourly model is the one used in the modified generator.

The daily model developed by Graham is an example of a Type I or "realistic" generation model, not suitable for the generator being developed in this study. Likewise, this type of procedure will produce Type I hourly values; that is the hourly k_t values will follow a typical, realistic pattern, *not* necessarily a long-term average pattern. Days in which the hourly values of k_t are equal to their long-term average values are smooth and symmetric; such behavior is not typical and generated data with this behavior can lead to significant errors in performance estimates. The goal is to produce an average month of weather data, but the days within this month should be realistic in the sense that the hours should display the same variability observed in actual days.

3.2.2 The k_t Sequence

Graham suggested that the sequence of k_t values can be considered to consist of a deterministic component and a random component. The deterministic trend is due to the geometric relation of the earth and the sun; the random component is brought about by various uncertainties such as cloud movement and pollution.

3.2.2.1 The Deterministic Component

The deterministic portion of the hourly k_t sequence is equivalent to the long-term average estimate of k_t for an hour, and can be computed by combining the daily K_t value with Equations (2.2.1) and (2.2.2):

$$k_{tm} = K_t \frac{r_t}{r_d} = \frac{H}{H_o}\frac{I}{H}\frac{H_o}{I_o} = \frac{I}{I_o} \tag{3.2.1}$$

where k_{tm} represents the long-term mean value of k_t for an hour.

Graham apparently did not recognize that Equation (3.2.1) can be used to estimate the mean k_t value, and instead developed a regression model from 185 days of data (approximately 2000 hours) from three Canadian cities: Toronto, Swift Current and Vancouver. The model presented by Graham expressed the variation in k_{tm} as a function of K_t and zenith angle.

$$k_{tm\ Graham} = K_t - 1.167\,K_t^3\,(1\text{-}K_t) + 0.979\,(1\text{-}K_t)\exp\left[\frac{-1.141\ (1\text{-}K_t)}{K_t\cos\theta_z}\right] \tag{3.2.2}$$

Equation (3.2.2) generates different values of k_{tm} than those obtained from Equation (3.2.1). Comparison of the two models on June 15 at latitudes of 30, 40 and 50 degrees for several K_t values is shown in Figures 3.11 a-e. While the two models appear similar for most values of K_t, Graham's model suggests the diurnal variation of k_{tm} for a very high value of K_t (e.g., 0.9) to be almost nonexistent, whereas the values of k_{tm} obtained from Equation (3.2.1) show a very marked variation on such a day. At very low values of K_t (e.g., 0.1) the same is true; Graham's model predicts an almost constant value of k_{tm}, while Equation (3.2.1) shows a slight variation. More disconcerting, however, is the comparison at K_t=0.5 (Figure 3.11 c). While the shapes are similar, the values as predicted by Graham are higher for all hours of the day. Therefore, for one of the two models, the value of K_t as calculated by summing the hourly values obtained from the k_{tm} model must not

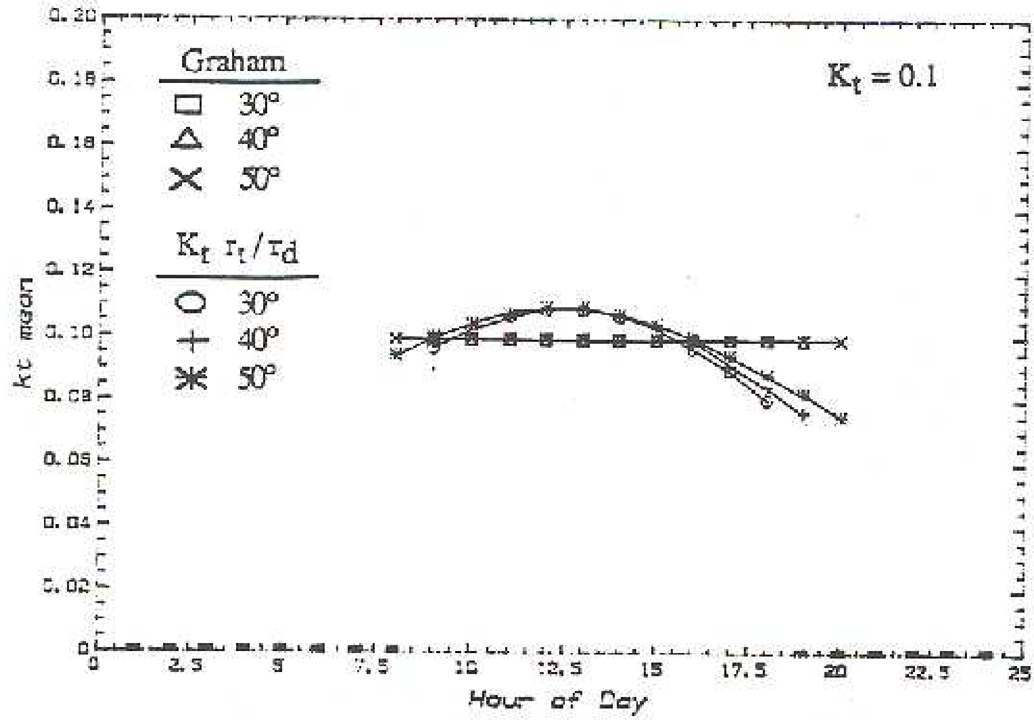

Figure 3.11A Comparison of Graham's Model for Estimating k_{tm} and the Model $k_{tm} = K_t\, r_t/r_d$ on June 15 for $K_t = 0.1$

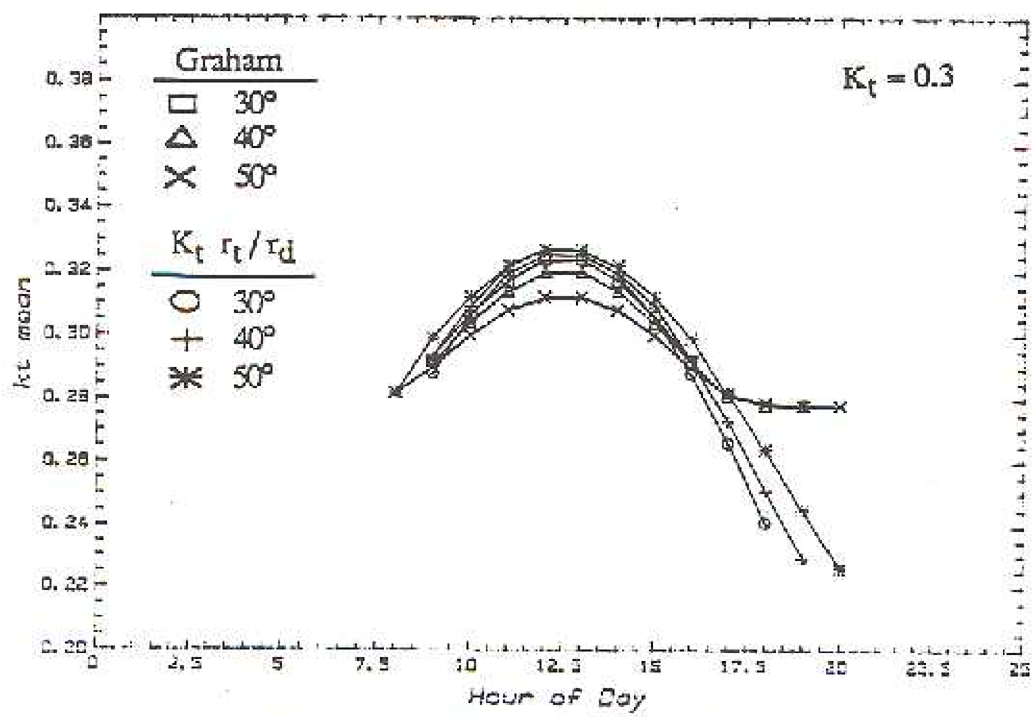

Figure 3.11B Comparison of Graham's Model for Estimating k_{tm} and the Model $k_{tm} = K_t\, r_t/r_d$ on June 15 for $K_t = 0.3$

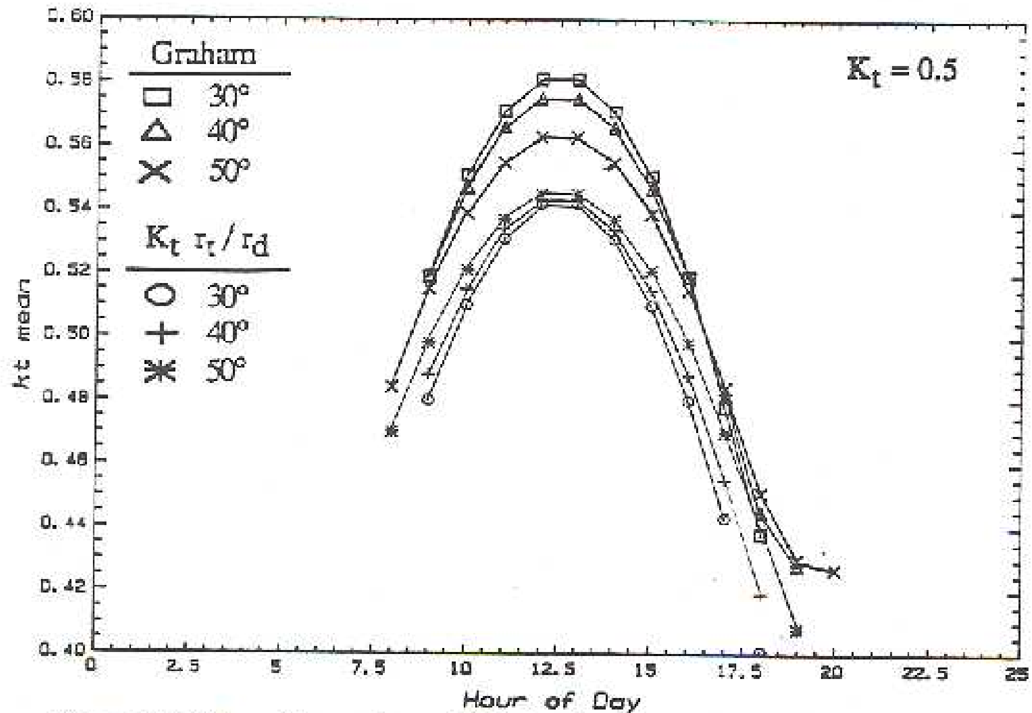

Figure 3.11C Comparison of Graham's Model for Estimating k_{tm} and the Model $k_{tm} = K_t\, r_t/r_d$ on June 15 for $K_t = 0.5$

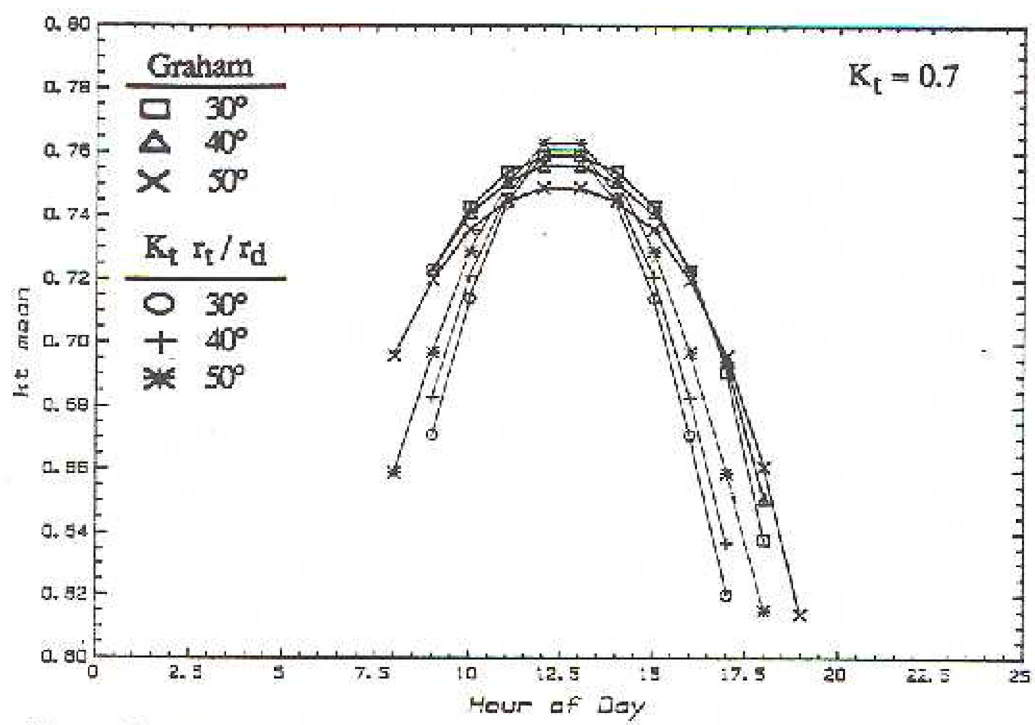

Figure 3.11D Comparison of Graham's Model for Estimating k_{tm} and the Model $k_{tm} = K_t\, r_t/r_d$ on June 15 for $K_t = 0.7$

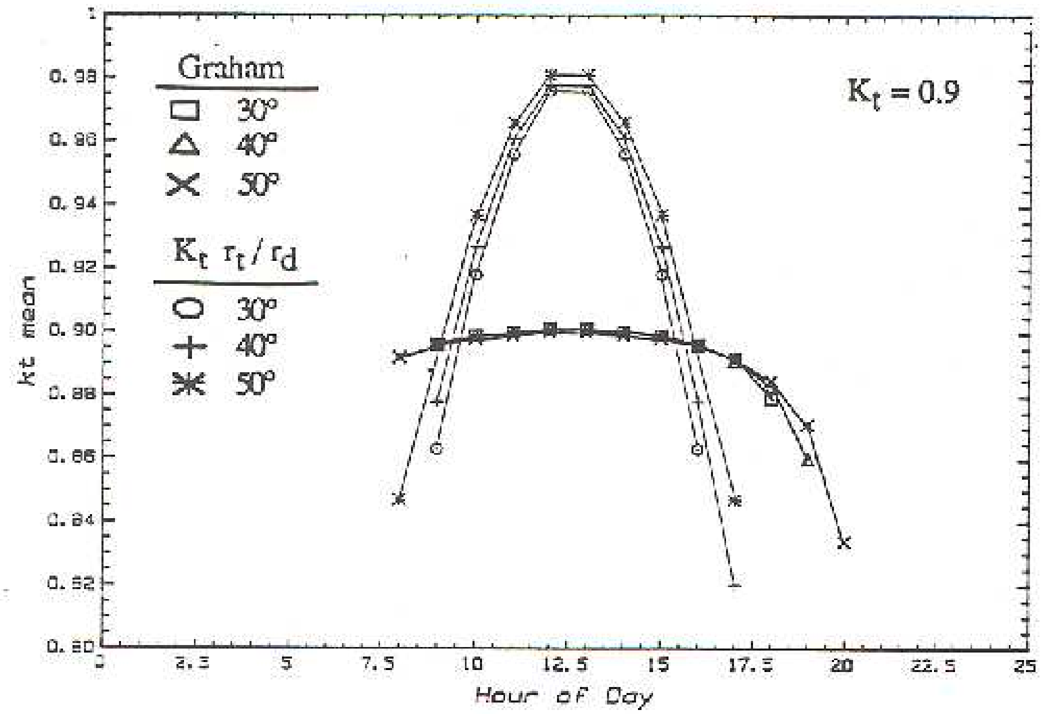

Figure 3.11E Comparison of Graham's Model for Estimating k_{tm} and the Model $k_{tm} = K_t\, r_t/r_d$ on June 15 for $K_t = 0.9$

Table 3.2

Comparison of the K_t values obtained from a day in which $k_t = k_{tm}$

K_t	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Jan 15									
$K_t\, r_t/r_d$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.89
Graham	0.10	0.19	0.28	0.36	0.46	0.55	0.66	0.77	0.88
Mar 15									
$K_t\, r_t/r_d$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.79	0.89
Graham	0.10	0.19	0.29	0.39	0.50	0.60	0.70	0.80	0.89
Jun 15									
$K_t\, r_t/r_d$	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Graham	0.10	0.20	0.30	0.42	0.53	0.63	0.72	0.81	0.90

be equal to the value of K_t from which the k_{tm} values were obtained. To check which model was in error, K_t values were generated from both models for input K_t values of 0.1, 0.2, ..., 0.9 at 40° latitude on the 15th day of each month. The abbreviated results presented in Table 3.2 clearly indicate that Graham's model inaccurately recreates K_t. Therefore, further reference to the variable k_{tm} will be to the value of k_{tm} as defined in Equation (3.2.1), unless explicitly stated otherwise.

3.2.2.2 The Random Component

The random component of the k_t sequence, a_{kt}, is simply the difference between the actual value of k_t and the long-term mean value:

$$a_{kt} = k_t - k_{tm} \tag{3.2.3}$$

As an indicator of the size and spread of the deviations, Graham computed the mean and standard deviation of the a_{kt} values, σ_a, as a function of K_t and developed a model for estimating σ_a:

$$\sigma_a = 0.1557 \sin\left[\frac{\pi K_t}{0.9330}\right] \tag{3.2.4}$$

Similarly, the mean and standard deviation of the a_{kt}'s as a function of K_t was calculated from the long-term hourly records for Albuquerque, Madison, and New York. The results are listed in Table 3.3.

Table 3.3
Mean and standard deviation of a_{kt}

	Mean			Standard Deviation		
K_t	Alb	Mad	NYC	Alb	Mad	NYC
0.1	0.0097	0.0059	0.0087	0.058	0.050	0.065
0.2	0.0051	0.0082	0.0068	0.099	0.099	0.107
0.3	0.0135	0.0104	0.0047	0.151	0.138	0.141
0.4	0.0097	0.0055	0.0006	0.172	0.160	0.154
0.5	0.0068	-0.0016	-0.0037	0.203	0.158	0.131
0.6	-0.0002	-0.0030	-0.0030	0.185	0.124	0.091
0.7	-0.0009	0.0022	-0.0016	0.139	0.068	0.066
0.8	0.0088	-0.0012	0.0053	0.057	0.094	0.064

Figure 3.12 shows the mean a_{kt} values as a function of K_t; a definite dependence on K_t is apparent. It is also evident that the k_{tm} value as predicted from Equation (3.2.1) is somewhat more accurate than the k_{tm} value predicted from Equation (3.2.2). However, the size of the deviations is very small; for the model of Equation (3.2.1), the maximum deviation is less than 0.015. If the k_{tm} model is correct, the mean a_{kt} value should be zero; it is very close.

A comparison of the values of σ_a as calculated in this study to the values as predicted by Equation (3.2.4) is shown in Figure 3.13. Agreement is quite good for the smaller values of K_t, however, particularly for values of K_t greater than 0.4, there is a considerable amount of location dependent spread within the data. Albuquerque consistently has a larger standard deviation, while New York has a consistently smaller standard deviation. Yet if a single expression is necessary to

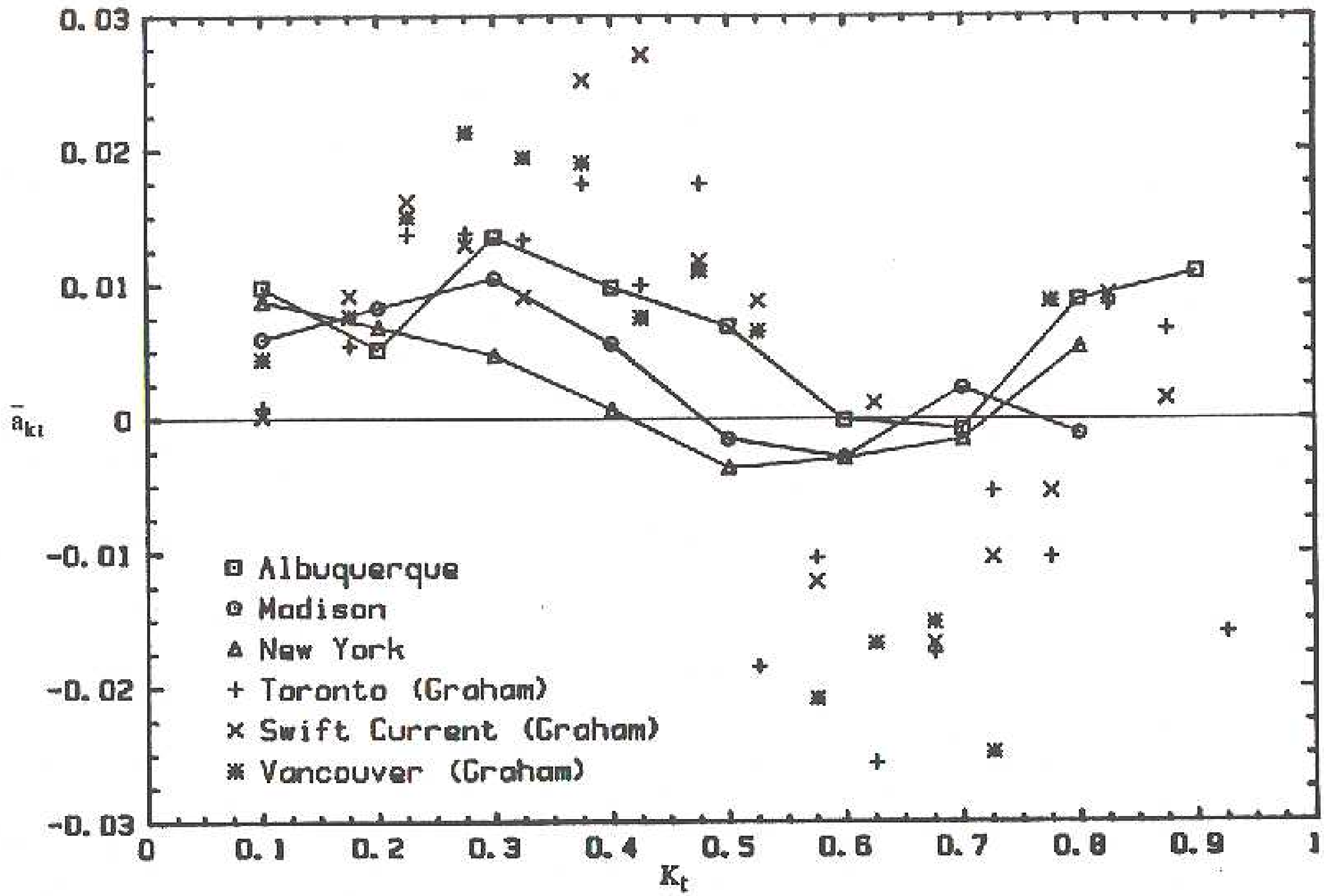

Figure 3.12 Mean Value of $a_{kt} = k_t - k_{tm}$ as a Function of K_t

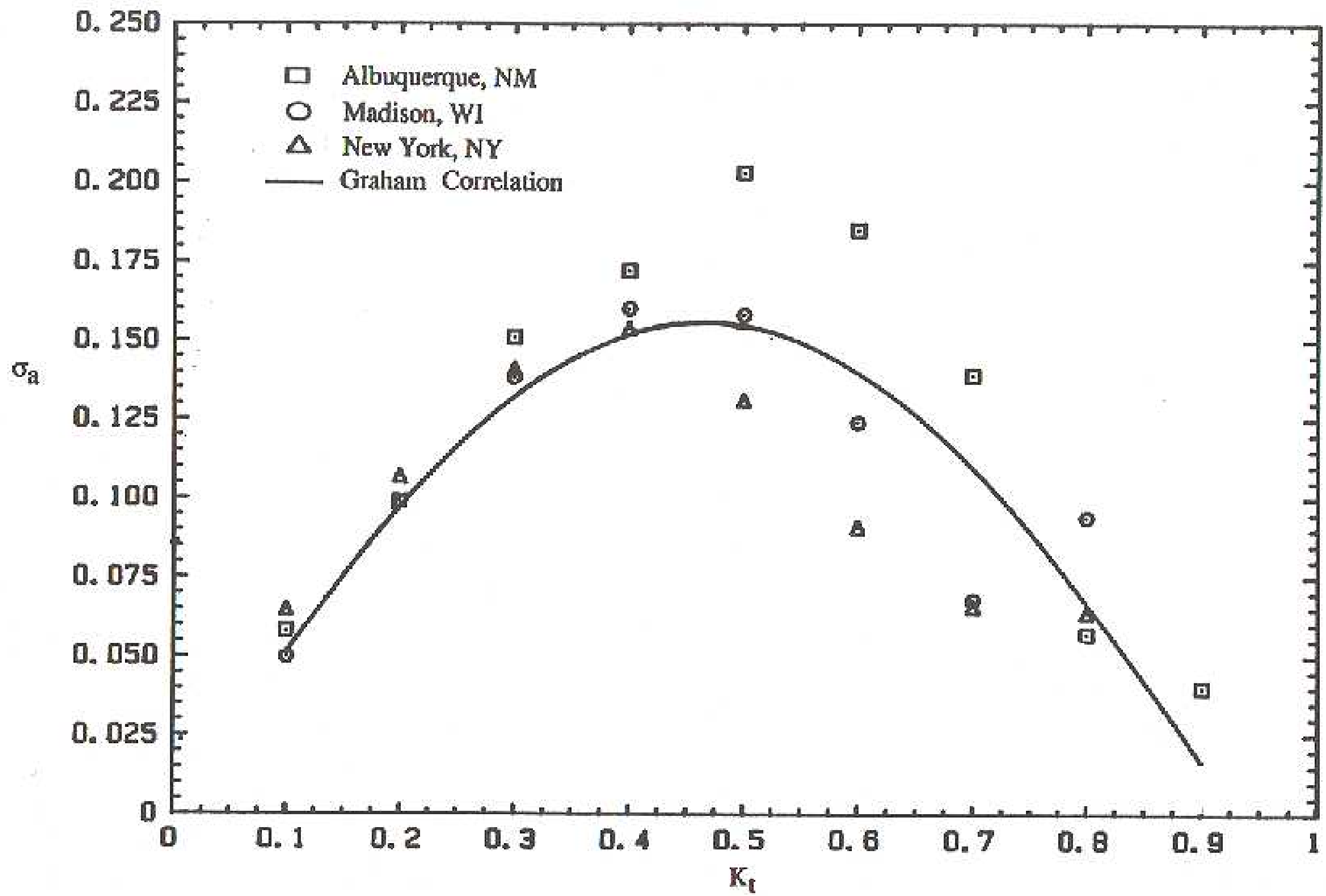

Figure 3.13 Standard Deviation of the a_{kt} Values as a Function of K_t

represent σ_a for all locations, Equation (3.2.4) appears adequate.

A common technique in stochastic modeling involves first estimating any deterministic trends and then reproducing the residuals (i.e., the random component) with an AR or MA model. It is therefore the autocorrelation structure of the deviations, a_{kt}, which is of interest in applying a time series model.

The sequence of a_{kt} values, however, unlike that of hourly temperature or daily K_t values, is discontinuous. No values are recorded during nighttime periods; an attempt to make a continuous k_t series by catenating the daily sequences of hourly k_ts would introduce erroneous terms into the estimation of the autocorrelation. While undoubtedly there is some degree of correlation between the last k_t value of one day and the first k_t value of the following day, it is different than that between consecutive hourly k_t values within a day.

3.2.2.3 Autocorrelation of the Hourly Disturbances, $a_{kt}=k_t-k_{tm}$

To obtain a reliable estimate of the autocorrelation coefficient, approximately 50 sequential values of a variable are necessary [Box and Jenkins, 1976]. Obviously when a single day is considered, there are only somewhere in the neighborhood of 10 a_{kt} values, definitely too few to obtain a reliable estimate of the autocorrelation. However, within just one year of data there are 365 days, creating many short sets of consecutive hourly a_{kt} values, but as previously noted they are not continuous. There should be a way to utilize the many sets of sequences to obtain a useful estimate of the autocorrelation.

The standard expression for estimating the lag one autocorrelation coefficient consists of the sum of the lag one product terms in the numerator and the sum of the lag zero product terms in the denominator.

$$\rho_1 = \frac{\sum_{t=1}^{n-1} (y_t - \bar{y})(y_{t+1} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2} \tag{3.2.5}$$

A "pooled estimate" of the lag one autocorrelation, applicable for the existing situation of many sets of a short sequence, can be obtained as follows:

$$\rho_1 = \frac{\sum_{T=1}^{n_d} \sum_{t=1}^{n_h-1} (y_t - \bar{y})(y_{t+1} - \bar{y})}{\sum_{T=1}^{n_d} \sum_{t=1}^{n_h-1} (y_t - \bar{y})^2} \tag{3.2.6}$$

where n_d is the number of days and n_h is the number of hours each day (for which there are k_t values).

The pooled estimate, Equation (3.2.6), can be thought of as the autocorrelation of the catenated k_t series, but *omitting* the product terms in the numerator which cross over the day boundaries. In other words, the numerator for each day is computed separately, and all the numerators are then combined to yield the "total" numerator for the autocorrelation coefficient. Additionally, since there is one less term in the numerator *each day*, one term must also be omitted in the denominator each day; the particular formulation of Equation (3.2.6) omits the last term each day. Normally, when using Equation (3.2.5), there is only one additional term in the denominator and due to the large number of terms, the effect of its inclusion is small.

Graham obtained estimates of the lag one autocorrelation using a similar expression (instead of discarding the last value of each day in the denominator, the first value of the day was discarded). However, the values of a_{kt} were computed using k_{tm} as calculated in Equation (3.2.2). No dependence on location was evident, although there appeared to be some dependence on K_t. Graham fit a regression model to represent the variation in ρ_1 as a function of K_t, although its ability to predict ρ_1 was not statistically different from the mean value of ρ_1=0.54.

$$\rho_{1a_{kt}} = 0.3455 + 1.0745\, K_t - 1.1327\, K_t^2 \tag{3.2.7}$$

The hourly a_{kt} lag one autocorrelations were computed from the 22 years of Albuquerque, Madison and New York data using Equation (3.2.6). Days with K_t values of 0.09 to 0.11, 0.19 to 0.21, etc. were grouped together. The mean, a_{kt}, was assumed to be zero, a reasonable approximation based on the data of Table 3.3. The first and last hour for which radiation was recorded each day were discarded, as the radiation amounts are small and often inaccurate.

The ρ_1 values computed in this study are displayed in Figure 3.14, along with the values as computed by Equation (3.2.7). The agreement between the data from this study and Graham's data is good, particularly when in consideration of the

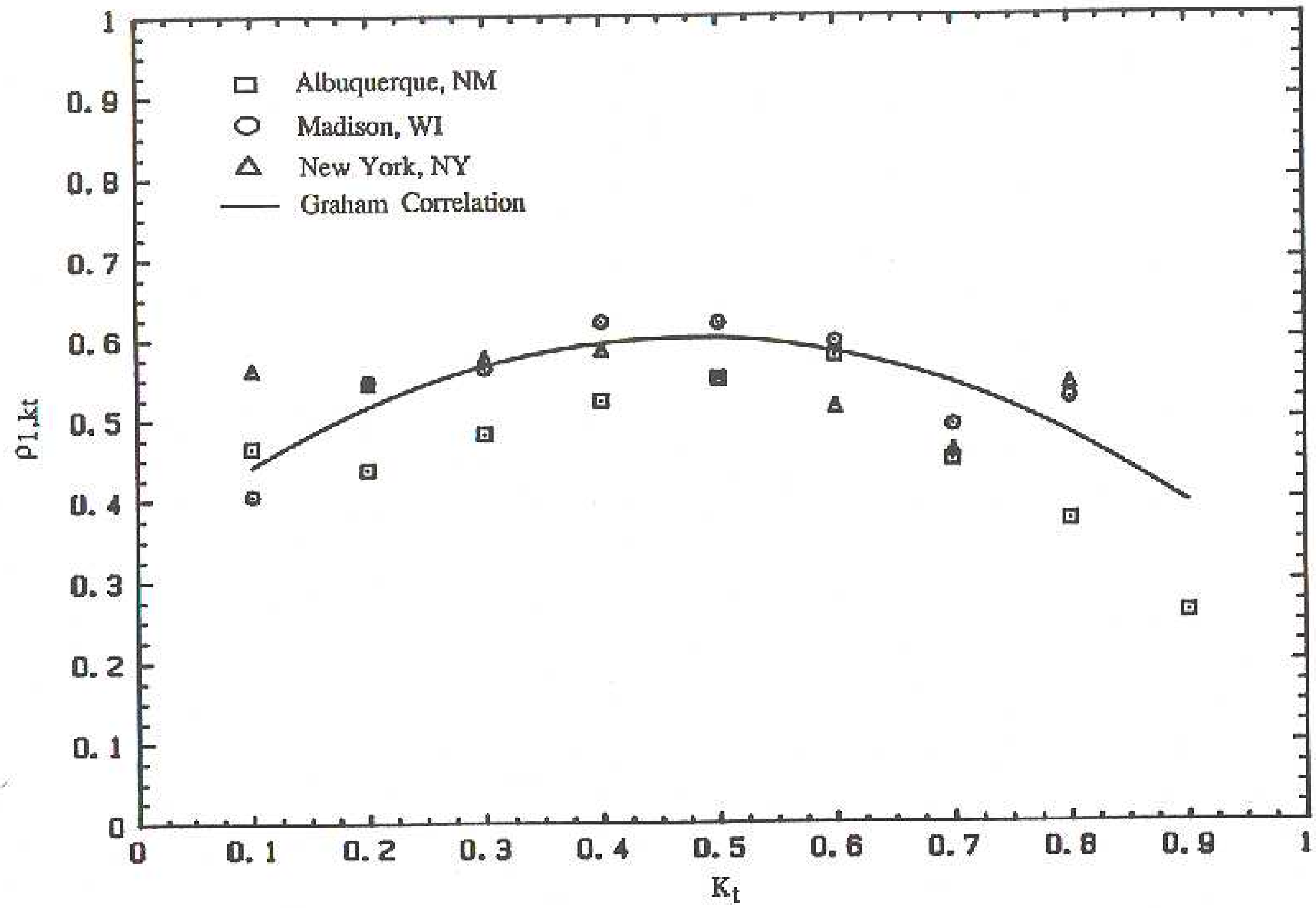

Figure 3.14 Lag One Autocorrelation of the a_{kt} Values as a Function of K_t

different k_{tm} models used. Unfortunately, some location dependence is apparent; the values calculated for Albuquerque are consistently lower, however, only by 0.1 or less.

3.2.3 Distribution of k_t about k_{tm}

One possible technique for modeling the k_t series would be to calculate k_{tm} for an hour, and obtain the deviation, a_{kt}, from the a_{kt} distribution (using a stochastic model to ensure the correct autocorrelation). The a_{kt} value would then be added to the k_{tm} value to produce the value of k_t for an hour. However, Graham concluded that the distribution of the k_t's about k_{tm} is similar to the distribution of the a_{kt} values about zero and more convenient to work with since it is bounded by 0 and 1. Therefore, Graham recommends instead modeling the k_t series directly, that is, to calculate k_{tm} for an hour, and obtain the k_t value from the distribution of k_t about k_{tm}.

The distribution of k_t about the long-term average value for an hour was first considered by Graham, and is distinctly different from the hourly k_t distributions described by Hollands and Huget[1983] and Theilacker[1980]. The shape of this particular k_t distribution is dependent on both the hour of the day and the value of K_t (see for example, Figures 3.16, 3.17 , and 3.18).

For a particular value of K_t, the variance associated with a k_t value at an hour far from noon will be larger then the variance associated with a k_t value at an hour near noon. Assume that the total amount of radiation has been calculated for a day, but the exact hourly values are unknown. There are virtually an infinite number of possible combinations of hourly radiation values which will sum to yield the specified amount of radiation for the day. One possible combination is that in which all hourly values are assumed equal to their long-term mean values. In selecting another possible set, if there is a deviation from the mean radiation value at some hour in the day, some other hour or combination of hours must also deviate from their mean by an equivalent amount so that the total radiation for the day is still the same. However, constraints are placed on the size of the deviation for an hour, as the maximum radiation for an hour cannot exceed the extraterrestrial radiation nor can it be less than zero.

Thinking in terms of the variable k_t instead of the radiation, a large deviation of k_t from k_{tm} for an early morning hour represents only a small deviation in radiation. If this radiation deviation was to be offset at the noon hour, it would translate into only a small deviation in k_t. Conversely, a large deviation of k_t from k_{tm} at the noon hour represents a large radiation deviation; the number of possible combinations which could offset this deviation is much smaller, and hence the probability of a large k_t deviation at the noon hour is much smaller than the probability of a large k_t deviation at an early morning hour. This is essentially the same as saying the variance of the k_t values at noon is smaller than the variance at an hour far from noon.

The value of K_t for the day also has an effect on the shape of the k_t distribution. Because the distribution is bounded by k_t=0 and k_t=1, for days with a high value of K_t, the distribution should be skewed to the right; similarly, for days with a low value of K_t, the distribution should be skewed to the left. For the mid-range K_t values, a more symmetrical distribution would be expected. It would also be expected that the variance of the k_t distribution would be smaller for extreme values of K_t and larger for the mid-range values. For example, on days with a high K_t value, every hour must be clear (a small variance), while for days with a mid-range value of K_t, every hour could be overcast or half of the hours could be clear and the other half cloudy (a larger variance).

Graham did not complete a detailed study of the distribution, but based upon reasoning similar to that above and histogram plots of the data for Toronto, Swift Current, and Vancouver, approximated the distribution by a beta distribution. Instead of analyzing the k_t values directly, the normalized variable $u=(k_t-x_l)/(x_u-x_l)$ was used, where x_l and x_u are the lower and upper bounds, respectively, and were estimated from the histogram plots.

$$\begin{aligned} x_l &= 0.0 && Kt < 0.5 \\ x_l &= 2.68\,K_t - 1.29 && 0.5 \geq K_t \leq 0.705 \\ x_l &= 0.6 && Kt > 0.705 \end{aligned} \tag{3.2.8}$$

$$x_u = 0.964\,K_t^{0.2} \tag{3.2.9}$$

Graham did not believe that the variation in variance with hour of the day would have a significant enough effect upon the generated data to warrant its inclusion in the distribution model.

The form of the pdf used by Graham is:

$$p(k_t) = c\,(x_u - x_l)^{\alpha+\beta-2} u^{\alpha-1} (1-u)^{\beta-1} \tag{3.2.10}$$

$$c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \frac{1}{(x_u - x_l)^{\alpha+\beta-2}}$$

Expressions for the mean and variance of Equation (3.2.10) were equated to the mean, k_{tm}, and the variance, σ^2, of the data, and then solved to yield equations for the parameters:

$$\alpha = \frac{k_{tm}^{\,2}(1-k_{tm})}{\sigma_a^2} - k_{tm}$$

$$\beta = \alpha\,\frac{1-k_{tm}}{k_{tm}} \tag{3.2.11}$$

k_{tm} was estimated by Graham from Equation (3.2.2) , and σ^2 was calculated by squaring σ as obtained from Equation (3.2.4). While σ^2 computed in this manner is the variance of the deviations, a_{kt}, rather than that of the k_t values about k_{tm}, Graham stated that the two distributions are similar in shape, and the variances should also be similar. This representation of the distribution includes the dependence on K_t through the value of $\sigma_a{}^2$, while neglecting any variation in shape due to hour.

Unfortunately, the beta distribution has no analytic representation for the cumulative distribution function, which is used in the transformation of the normally distributed variables to the k_t variables. Therefore, to calculate each hourly k_t value, interpolation in a table is required.

In an attempt to develop an explicit expression for the k_t distribution, preliminary plots of the cumulative distribution function were collected from the Albuquerque and Madison TMY data. Because of the small sample size, and so that all of the available data could be utilized, the size of the K_t bins was large (0.1). Rather than examining each hour separately, the data was sorted by hour pairs (e.g., 10 am & 3 pm), as no difference was expected for hours symmetric about solar noon.

Because of the different locations and groups of days making up a curve for a specified K_t value, the mean value of all of the k_t values for each curve, representing k_{tm}, was used to describe the curve rather than the value of K_t. In other words, although the curves were binned by K_t, they are used as distribution curves for a specified value of k_{tm} rather than for a specified value of K_t.

Examination of the cdf curves for the different hour pairs and k_{tm} values followed those expectations described previously. However, the curves for Madison and Albuquerque for equivalent hours and k_{tm} values were not as similar as had been hoped, indicating a dependence on some other variables besides k_{tm} and hour. The Albuquerque curves appeared to have a larger variance than those for Madison. The k_t cumulative distribution functions of the Seattle, WA TMY Data were also plotted; they agreed well with the Madison curves.

Several transformations of the k_t variable were tried in an attempt to find a more location-independent form. The distribution of the variable, u, as studied by

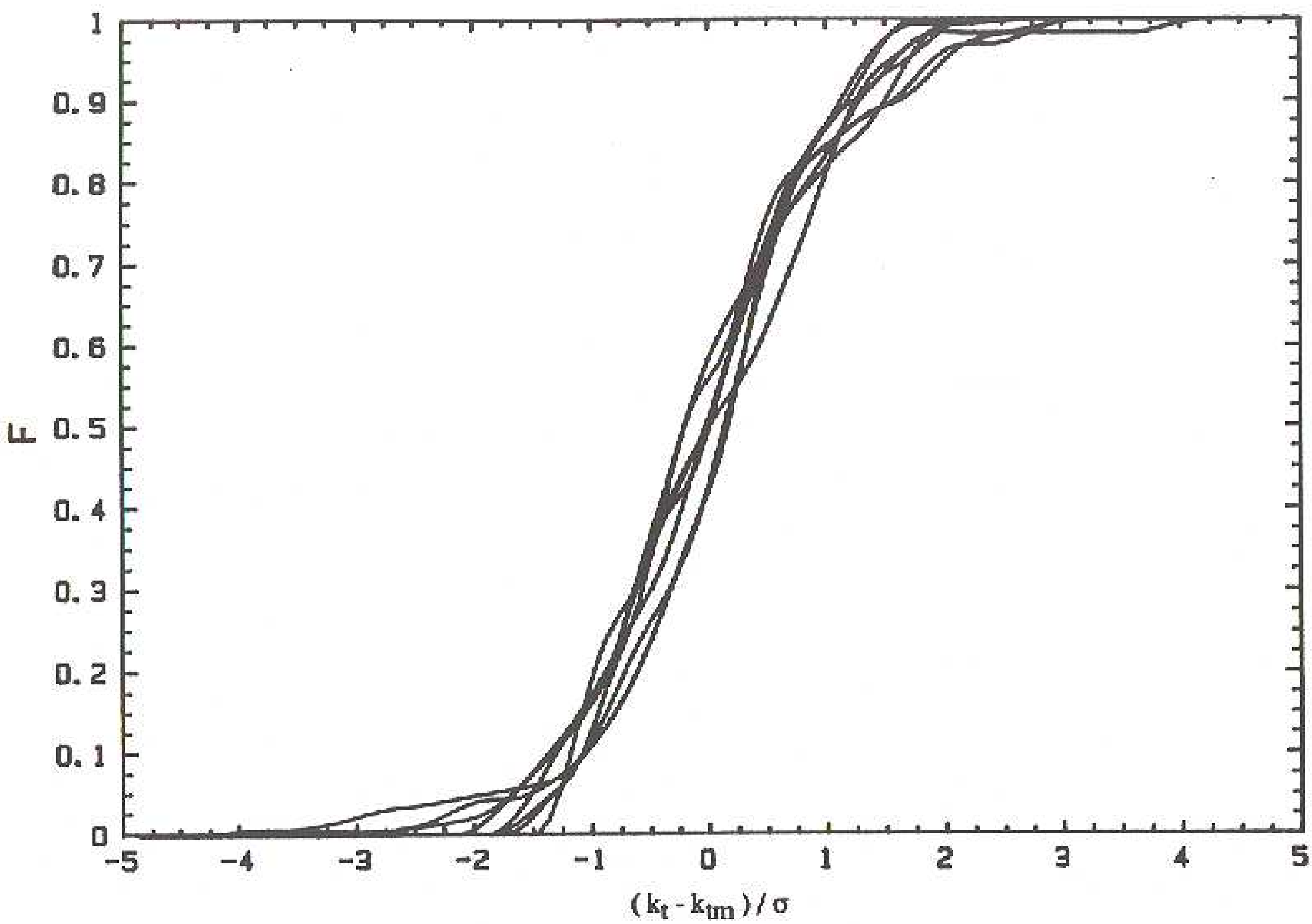

Figure 3.15 The Distribution of $(k_t-k_{tm})/\sigma$ for Madison, WI at the Hour Pair 11am, 2pm

Graham, was plotted for Albuquerque and Madison; the problem of different variances was still apparent. Plotting the distribution of the variable $h_{kt} = (k_t - k_{tm})/\sigma$, where k_{tm} and σ are the mean and standard deviation calculated for each curve, collapsed all of the curves (all hours, all k_{tm}s, and all locations) such that they could be represented by a single expression. An example of the h_{kt} distribution is shown in Figure 3.15. While to an extent, the plotting of many curves so close to each other hides any variation in skewness, it is also evident that the curves do not differ considerably.

Similarity to the normalized ambient temperature distribution [Erbs, 1984] and the hyperbolic tangent function prompted a linear least squares regression to estimate the parameter α in the following expression.

$$F_{kt} = \frac{1 + \tanh(\alpha h_{kt})}{2} \tag{3.2.12}$$

$$h = (k_t - k_{tm}) / \sigma$$

The Albuquerque and Madison data was regressed in three ways: separately for each hour and each location, for each location at all hours, and for all hours and all locations combined. Tables 3.4 and 3.5 summarize the results, listing the estimated value of the coefficient and the residual sum of squares for each case. Since the form of Equation (3.2.12) is not linear, the coefficient α was estimated by regressing $\tanh^{-1}(2F-1)$ on h_{kt}; therefore the units associated with the residual sum of squares is not the same as the F-values, but rather $\tanh^{-1}(2F-1)$. The increase in residual sum of squares is very small, while in the worst case, R^2 only dropped from 96% to 93%, indicating that a single parameter value is reasonable to use for all hours and locations.

Table 3.4
ALBUQUERQUE, NM (TMY Data)
Residual sum of squares for model fits to k_t distribution

K_t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	All K_t
Hour = 12, 1										
L,H	0.8	2.0	1.2	8.5	4.5	8.6	39.2	10.7	1.8	77.3
L	0.8	2.0	1.1	8.7	4.4	8.1	40.1	10.5	1.8	77.5
All	0.8	2.1	1.0	9.0	4.1	7.3	41.8	10.3	1.9	78.3
Hour = 11, 2										
L,H	0.8	1.2	10.7	4.8	3.2	9.6	19.1	21.9	2.9	74.2
L	0.8	1.2	11.2	4.6	2.5	9.0	19.5	23.8	3.0	75.6
All	0.8	1.3	11.6	4.6	2.2	8.7	19.9	25.3	3.1	77.5
Hour = 10, 3										
L,H	0.8	0.7	2.0	0.9	5.1	8.9	32.2	16.3	1.6	68.5
L	0.8	0.7	2.2	0.9	5.6	10.2	30.7	16.7	1.5	69.3
All	0.8	0.7	2.1	0.9	5.2	9.3	31.7	16.4	1.6	68.7
Hour = 9, 4										
L,H	0.8	4.3	3.8	3.0	0.7	7.3	5.8	53.1	0.3	79.1
L	0.8	4.2	4.0	3.1	0.7	7.8	5.9	52.5	0.3	79.3
All	0.8	4.2	3.8	3.0	0.7	7.4	5.9	53.0	0.3	79.1

L,H indicates coefficients specific to the location and hour pair
L indicates coefficients specific to the location
All indicates coefficients for all locations and all hour pairs

Table 3.5
MADISON, WI (TMY Data)
Residual sum of squares for model fits to k_t distribution

K_t	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	All
Hour = 12, 1										
L,H	9.2	5.1	5.7	4.5	3.8	12.7	7.7	2.5	1.3	52.5
L	8.3	5.0	5.8	5.2	3.9	13.3	7.8	2.7	1.2	53.2
All	7.9	5.0	5.9	5.8	4.1	13.9	8.0	2.8	1.1	54.5
Hour = 11, 2										
L,H	15.6	7.2	4.1	2.5	5.4	15.7	17.5	1.1	1.0	70.1
L	14.3	7.5	4.2	3.0	6.1	14.1	19.4	1.3	0.9	70.8
All	13.6	7.8	4.3	3.4	6.6	13.2	20.8	1.5	0.9	72.1
Hour = 10, 3										
L,H	3.0	7.4	2.9	8.6	8.8	11.9	36.7	1.5	1.1	81.9
L	3.0	7.5	3.1	8.9	9.1	12.2	35.3	1.5	1.1	81.7
All	3.2	7.8	3.4	9.5	9.8	12.9	33.2	1.7	1.1	82.6
Hour = 9, 4										
L,H	1.2	11.6	2.7	4.7	12.2	10.0	27.0	5.2	1.3	75.9
L	1.0	14.3	1.8	3.8	10.1	8.6	33.4	5.5	1.4	79.9
All	1.0	13.3	2.0	4.0	10.7	8.9	31.2	5.4	1.3	77.8

L,H indicates coefficients specific to the location and hour pair
L indicates coefficients specific to the location
All indicates coefficients for all locations and all hour pairs

Therefore, the form of the k_t distribution used to represent the hourly k_t data is the single parameter model:

$$F_{kt} = \frac{1 + \tanh(0.7925\, h_{kt})}{2} \tag{3.2.13}$$

An equivalent representation which is computationally more convenient is

$$F_{kt} = \frac{1}{1 + \exp(-1.585\, h_{kt})} \tag{3.2.14}$$

To use this distribution model for generating data, k_{tm} is estimated from Equation (3.2.1) , and the standard deviation can be estimated from the expression for σ_a developed by Graham, Equation (3.2.4).

While Figure 3.13 indicates discrepancies between the calculated σ_a and that predicted by Equation (3.2.4), it is not believed that effect of these differences will be significant. A new expression could possibly be developed, including some other parameters, however, using Equation (3.2.4) has the advantage of not being developed from the data used in the comparison. This distribution expression, like the beta distribution used by Graham, neglects any hour dependence in the shape of the distribution.

k_t distributions were also developed from the long-term data for Albuquerque, Madison, and New York. Since a large amount of data were available, the size of the K_t bins used in selecting the data for each curve was 0.02 (e.g., 0.09 to 0.11, 0.19 to 0.21, etc.). 40 bins of width 0.025 were used to group the k_t values. Distributions were compiled for four hour pairs (9am&4pm, 10am&3pm, 11am&2pm, 12pm&1pm). The total number of k_t values collected for each individual distribution curve ranged from 2 (Albuquerque, K_t=0.1) to 2248 (Albuquerque, K_t=0.8), with most of the curves consisting of between 100 and 700 values.

The long-term k_t distributions for the three locations are plotted in Figures 3.16, 3.17, and 3.18, along with the cumulative distribution functions as obtained from Equation (3.3.7) with k_{tm} and σ estimated from Equations (3.2.1) and (3.2.4). Agreement tends to be better for the curves corresponding to the lower K_t values; this is in part due to the greater inaccuracy of the σ_a model at the higher K_t values. While the fit is not exact, it does provide an approximation to the distribution curves suitable for this generation model; the exact shape of the distribution is not expected to have a significant effect.

In the analysis leading to Equation (3.3.7) and the work completed by Graham, no formal check of the distributions to determine whether k_{tm}, K_t and hour were sufficient to describe the variations was made. That is, can the standard deviation, skewness and kurtosis be sufficiently represented as functions of some combination of K_t, k_{tm}, and hour pair?

The modeling of the distribution of k_t about k_{tm} causes some confusion, in that while the individual curves were created by taking days with only a specified value of K_t, the mean value of k_t as calculated from the curves is not the same as K_t. In the generation process, the value of k_{tm} is known for an hour, and the actual value

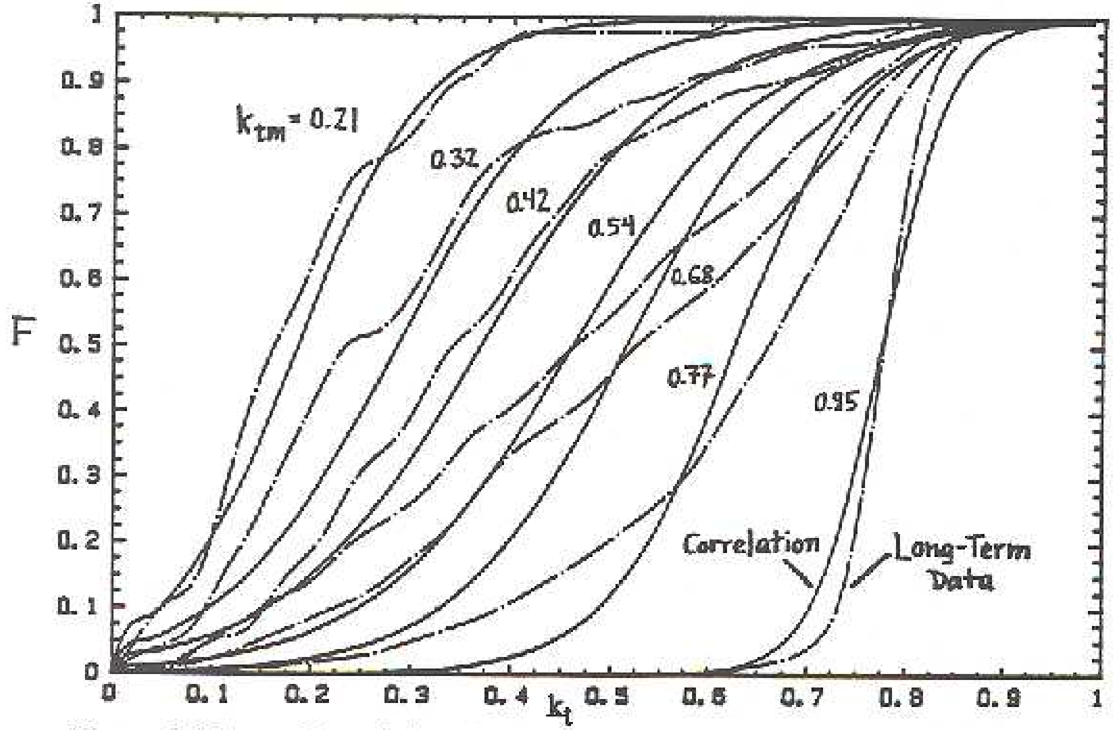

Figure 3.16A Cumulative Distribution of k_t about k_{tm} for Albuquerque, NM at the hour pair 12noon, 1pm

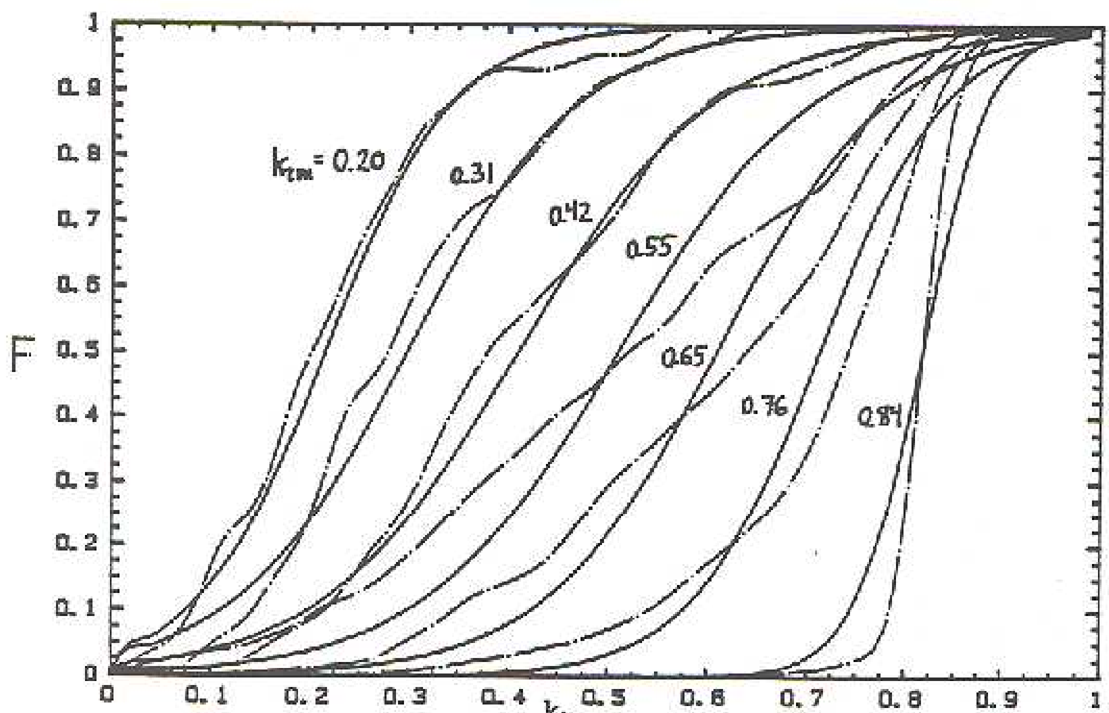

Figure 3.16B Cumulative Distribution of k_t about k_{tm} for Albuquerque, NM at the hour pair 11am, 2pm

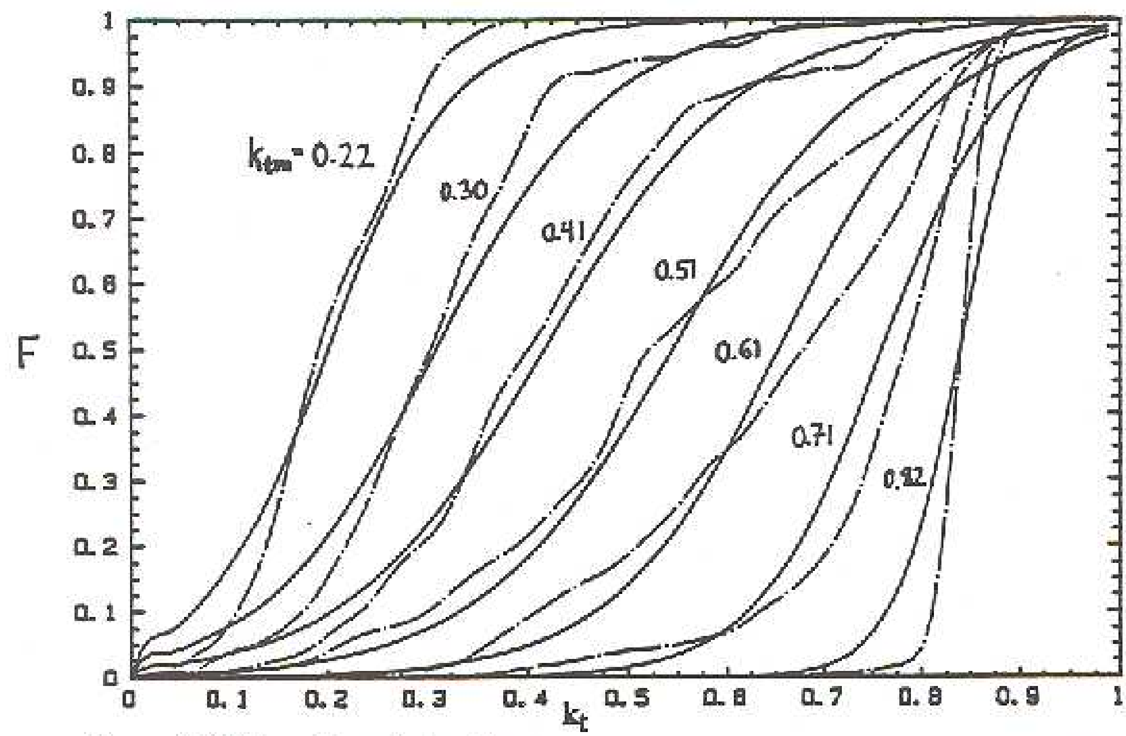

Figure 3.16C Cumulative Distribution of k_t about k_{tm} for Albuquerque, NM at the hour pair 10am, 3pm

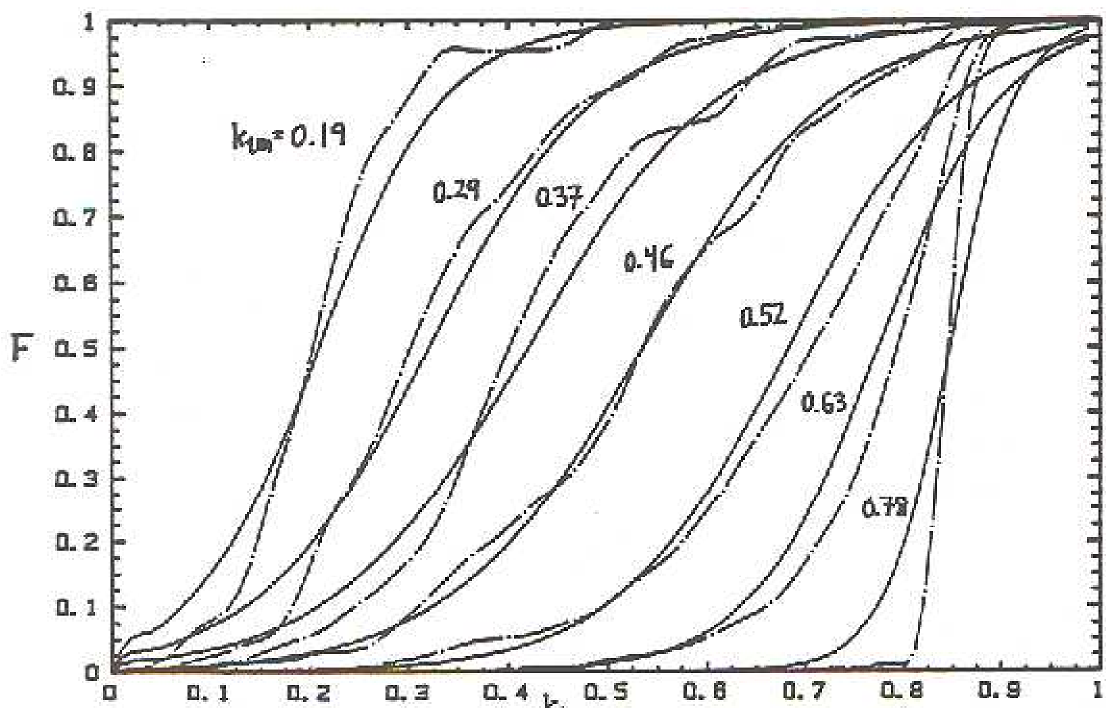

Figure 3.16D Cumulative Distribution of k_t about k_{tm} for Albuquerque, NM at the hour pair 9am, 4pm

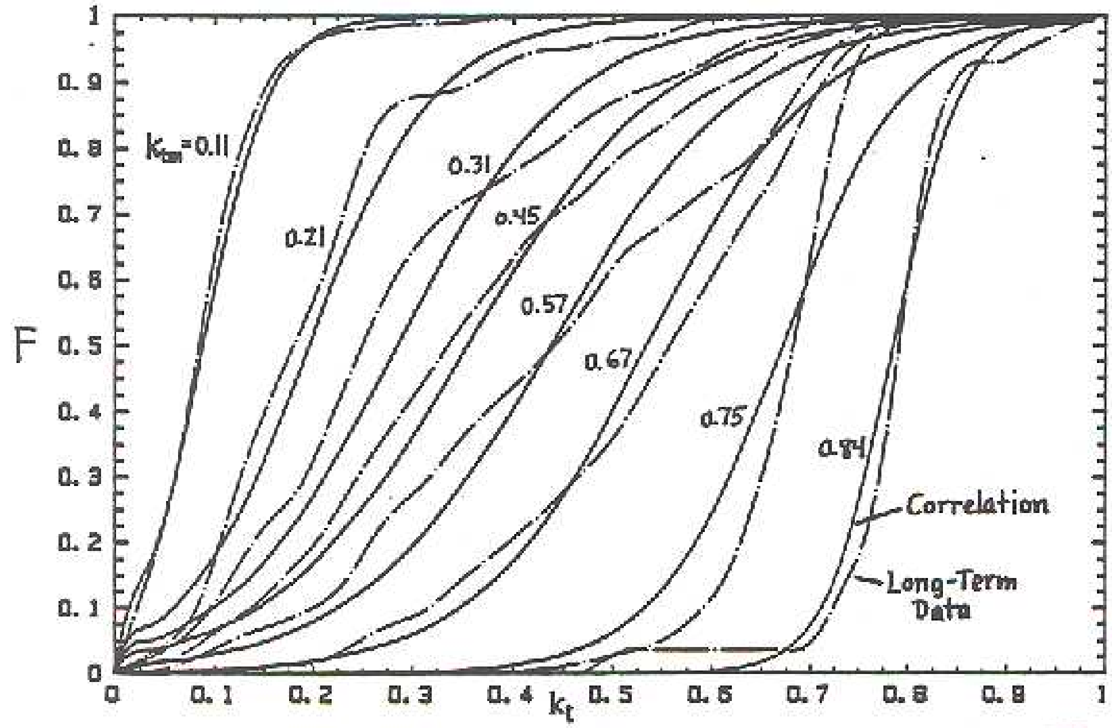

Figure 3.17A Cumulative Distribution of k_t about k_{tm} for Madison, WI at the hour pair 12noon, 1pm

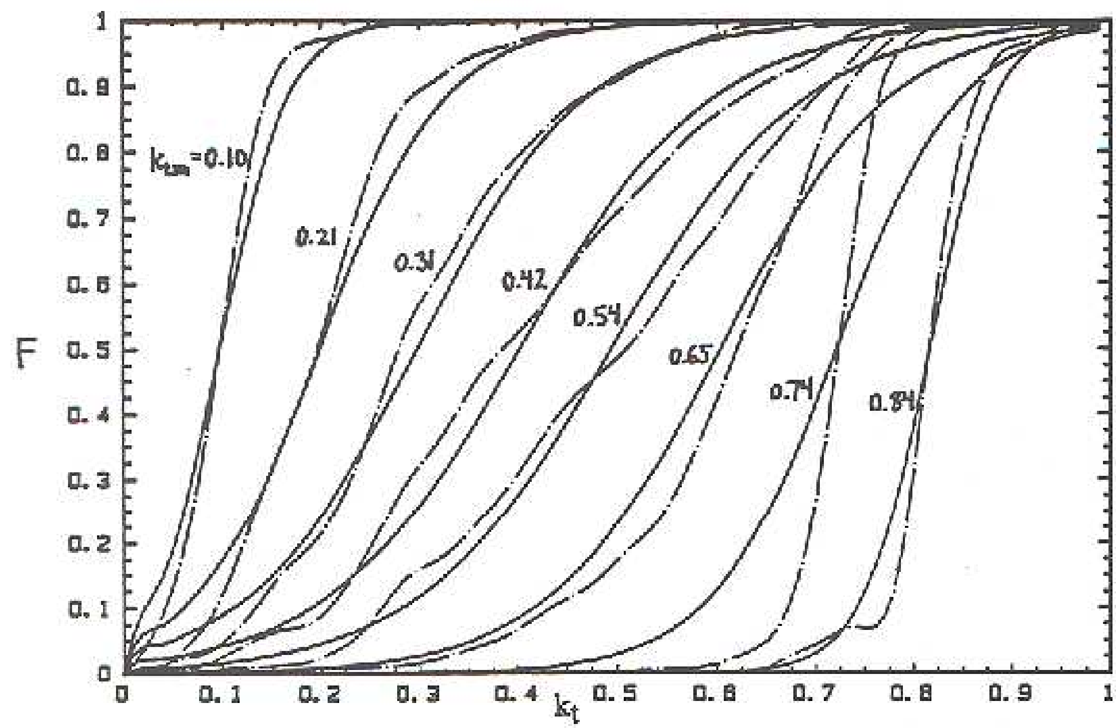

Figure 3.17B Cumulative Distribution of k_t about k_{tm} for Madison, WI at the hour pair 11am, 1pm

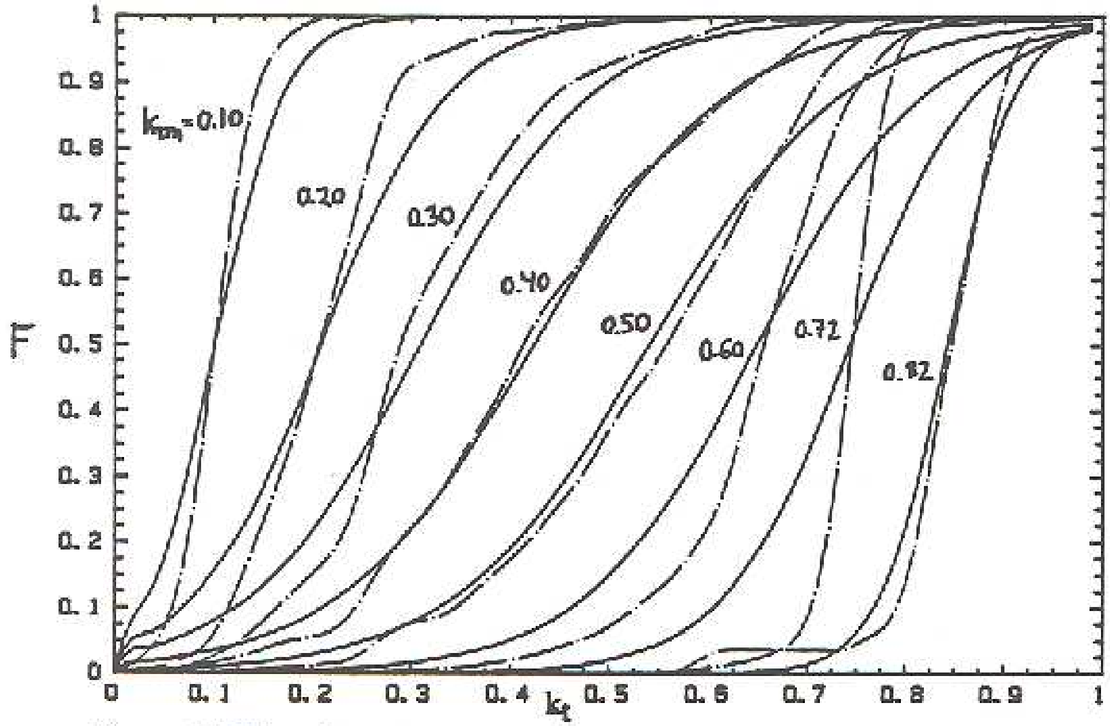

Figure 3.17C Cumulative Distribution of k_t about k_{tm} for Madison, WI at the hour pair 10am, 2pm

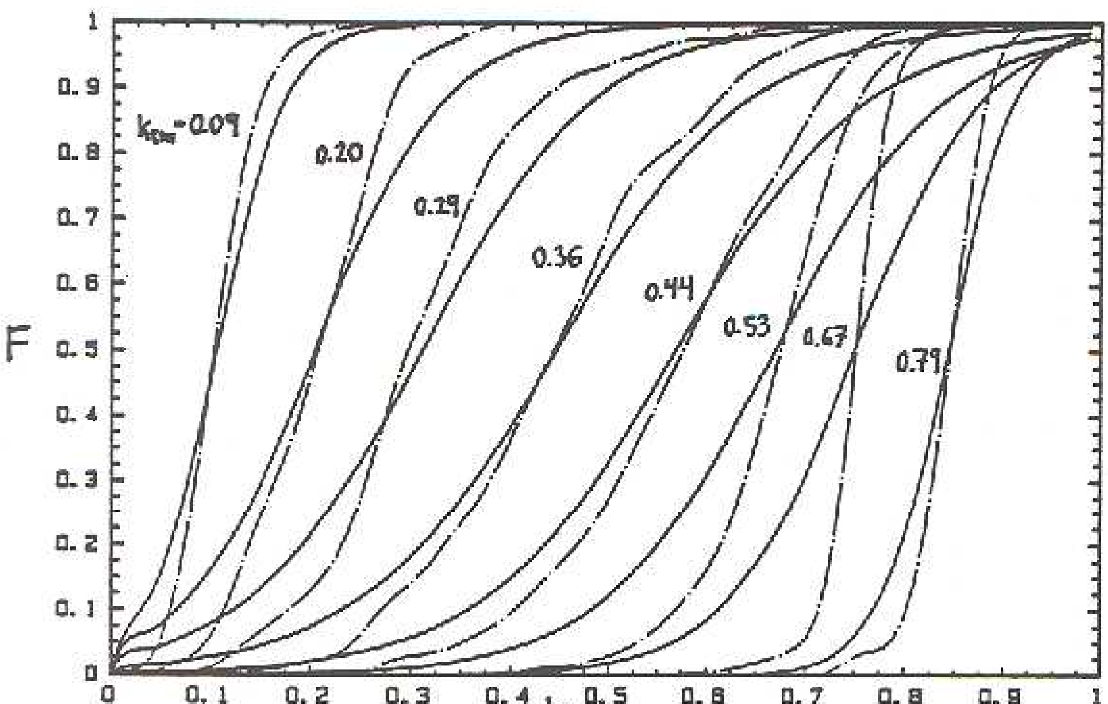

Figure 3.17D Cumulative Distribution of k_t about k_{tm} for Madison, WI at the hour pair 9am, 4pm

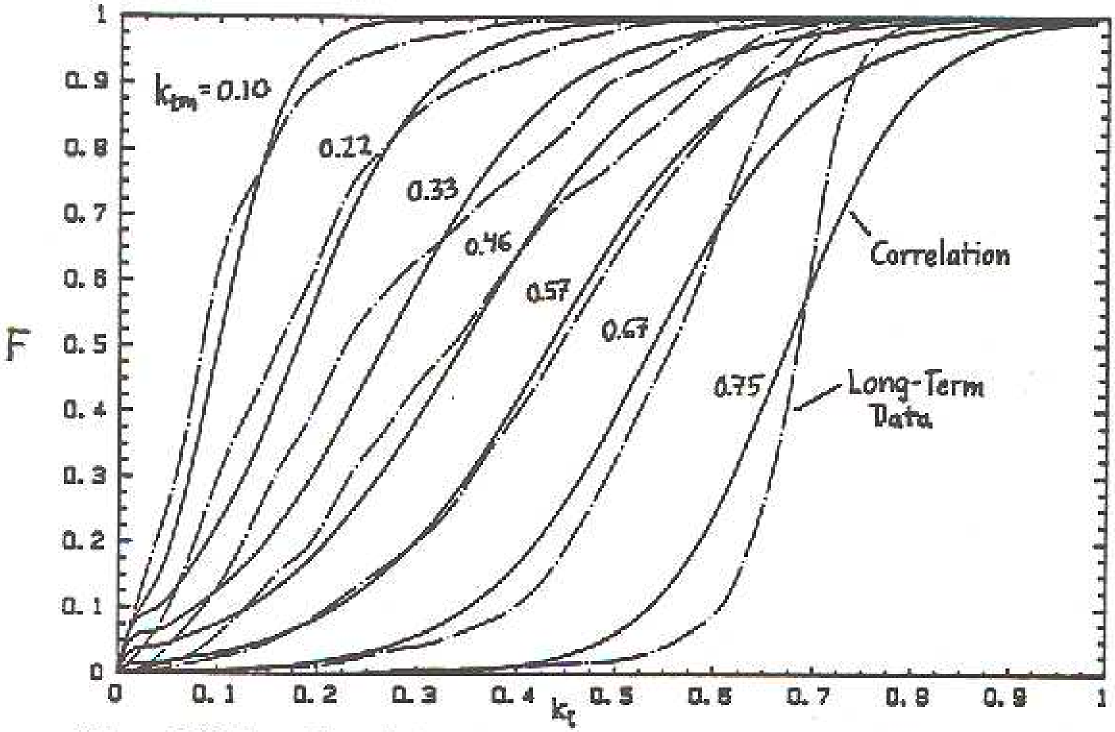

Figure 3.18A Cumulative Distribution of k_t about k_{tm} for New York, NY at the hour pair 12noon, 1pm

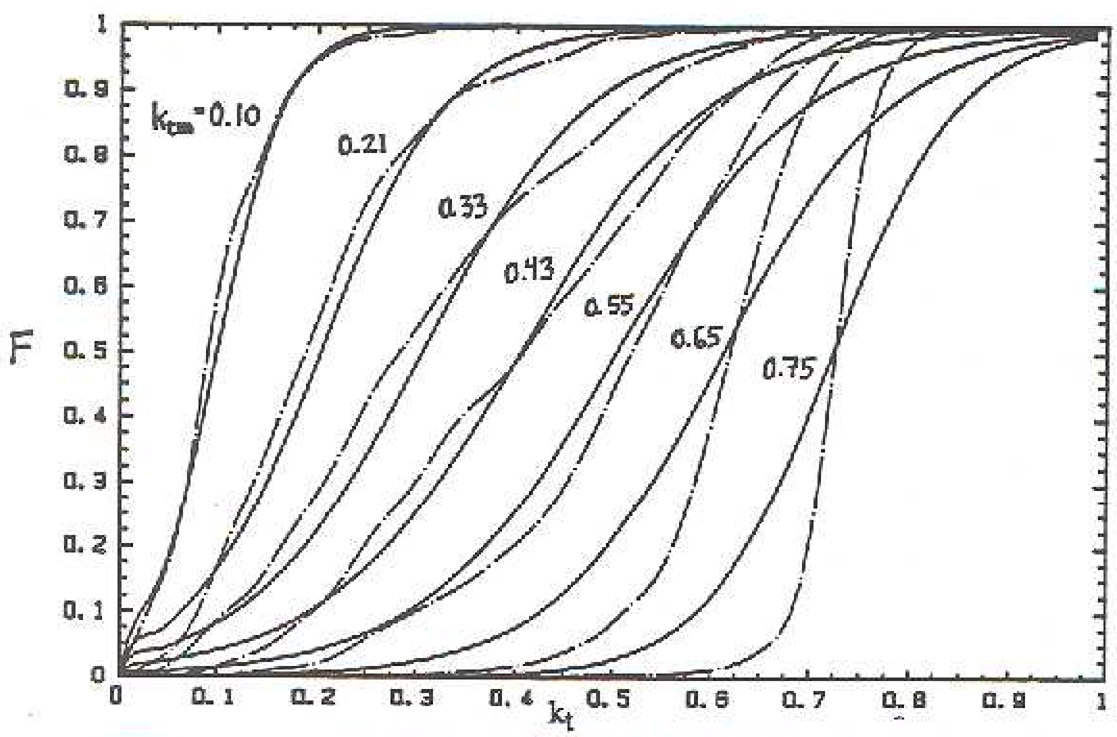

Figure 3.18B Cumulative Distribution of k_t about k_{tm} for New York, NY at the hour pair 11am, 2pm

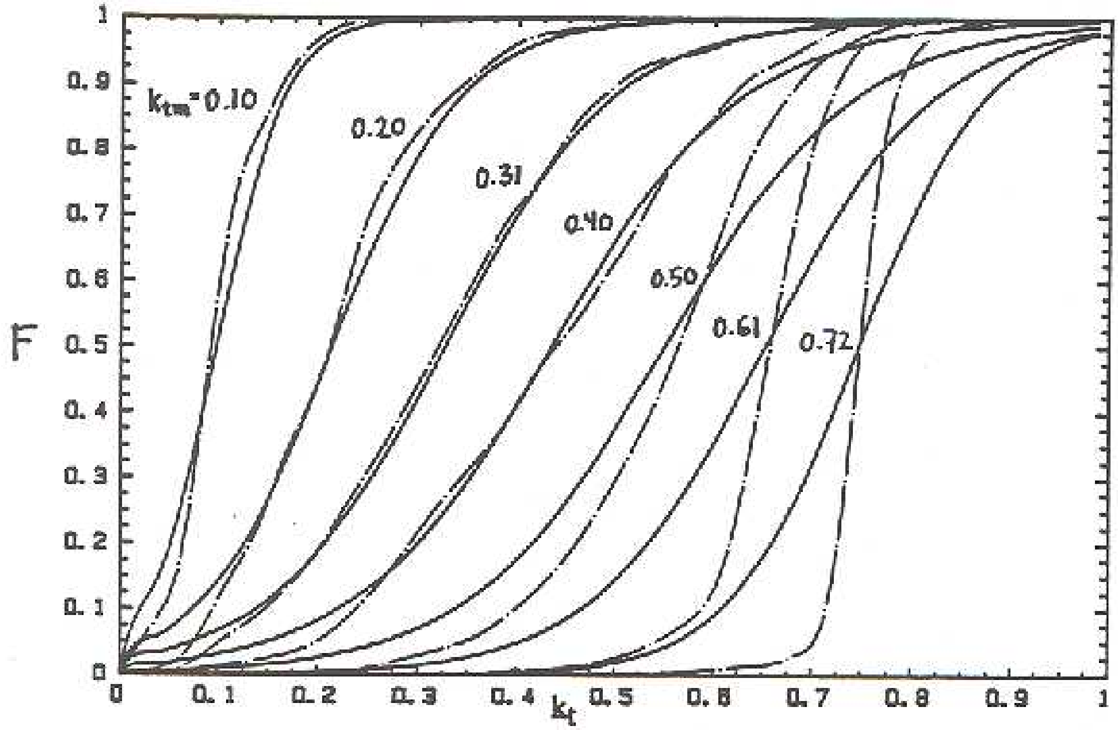

Figure 3.18C Cumulative Distribution of k_t about k_{tm} for New York, NY at the hour pair 10am, 3pm

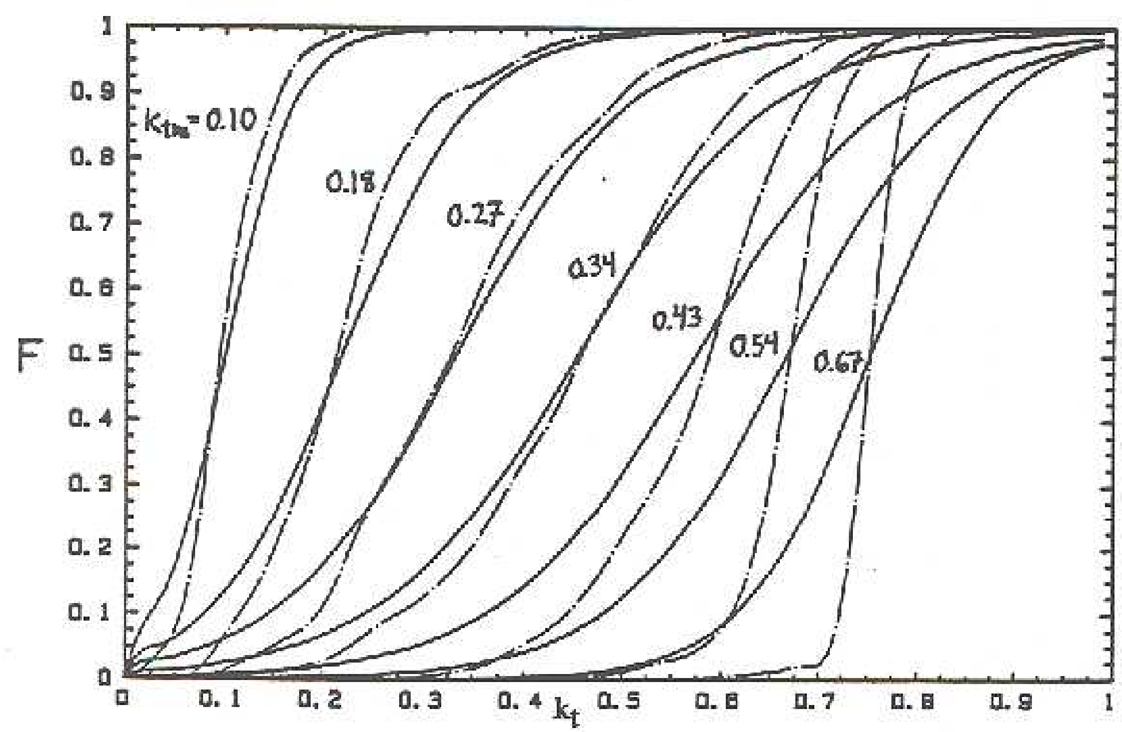

Figure 3.18D Cumulative Distribution of k_t about k_{tm} for New York, NY at the hour pair 9am, 4pm

of k_t must be found from these distribution curves. Therefore, the curves must be used as k_{tm} curves rather than K_t curves, that is, each curve must represent the distribution of k_t for that particular value of k_{tm}, *not* for that particular value of K_t, even though the curves were created for a specific K_t value. If the curves were used such that the particular curve was specified by K_t instead of k_{tm}, an inaccurate reproduction of the diurnal variation would occur (and if the hour differences were ignored as in Equation (3.2.10) or (3.2.13), no distinct diurnal variation would be generated at all). Modelling of the k_t distribution also clouds the issue as to whether K_t or k_{tm} is the appropriate variable of which the standard deviation, skewness, and kurtosis should be dependent upon. It was realized upon completion of this study that if the distribution of the deviations, a_{kt}, had been examined instead of the distribution of the k_t values about k_{tm}, this discrepancy would have been removed. In the generation process, the mean value of a_{kt} would always be zero, and the appropriate curve could be selected as a function of K_t. The k_t distribution was selected initially instead of the a_{kt} distribution as this was the approach taken by Graham.

The standard deviation of the k_t distributions, as shown in Figure 3.19 a-c, indicate a definite dependence upon hour and K_t, as expected. There are, however, significant differences between the different locations. While all appear to have the same general shape, somewhat similar to a sine wave, the size of the amplitude and the location of the peak differ between the three locations examined. This indicates that K_t and the hour are not sufficient to indicate the standard deviation of k_t about k_{tm}. This should have been expected from both visual inspection of the distribution curves and the location dependence apparent in the σ_a values (Figure 3.13).

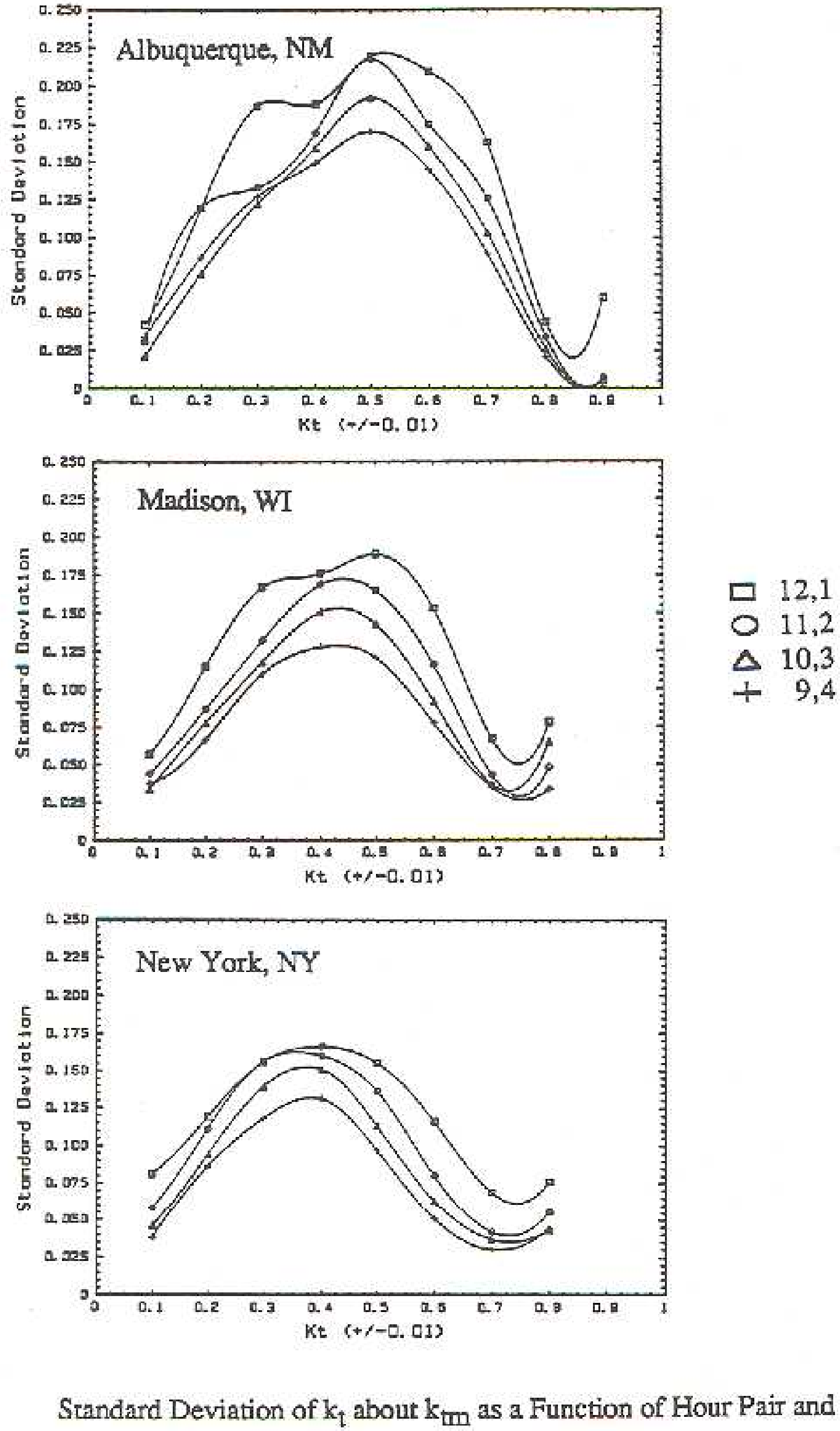

Figure 3.19 Standard Deviation of k_t about k_{tm} as a Function of Hour Pair and K_t

Possibly the inclusion of $\overline{\sigma}_{kt,yr}$, the standard deviation of the 12 $\overline{K}_t$ values about their average value or $\overline{K}_{t\ max}$ could explain these differences.

The skewness and kurtosis of the k_t distribution do not show the strong hour dependence evident in the standard deviation (Figures 3.20 and 3.21). They would seem to be reasonably represented solely by the value of K_t, especially in consideration of the large uncertainty in these estimates (they are extremely sensitive to rounding errors and outliers). Both the skewness and kurtosis do show some dependence on K_t, which the simplified models of Equation (3.2.10) and (3.2.13) neglect.

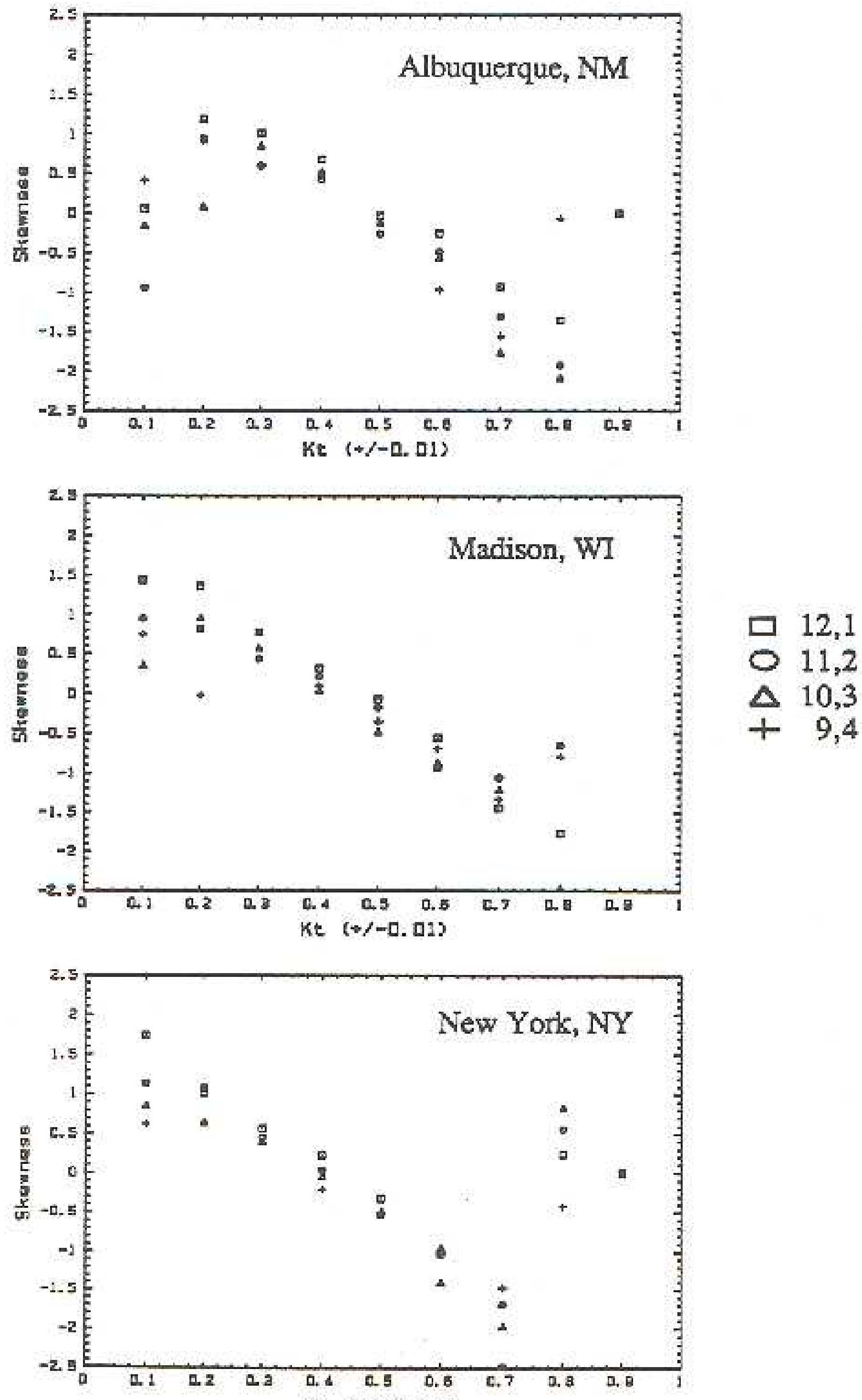

Figure 3.20 Skewness of k_t about k_{tm} as a Function of Hour Pair and K_t

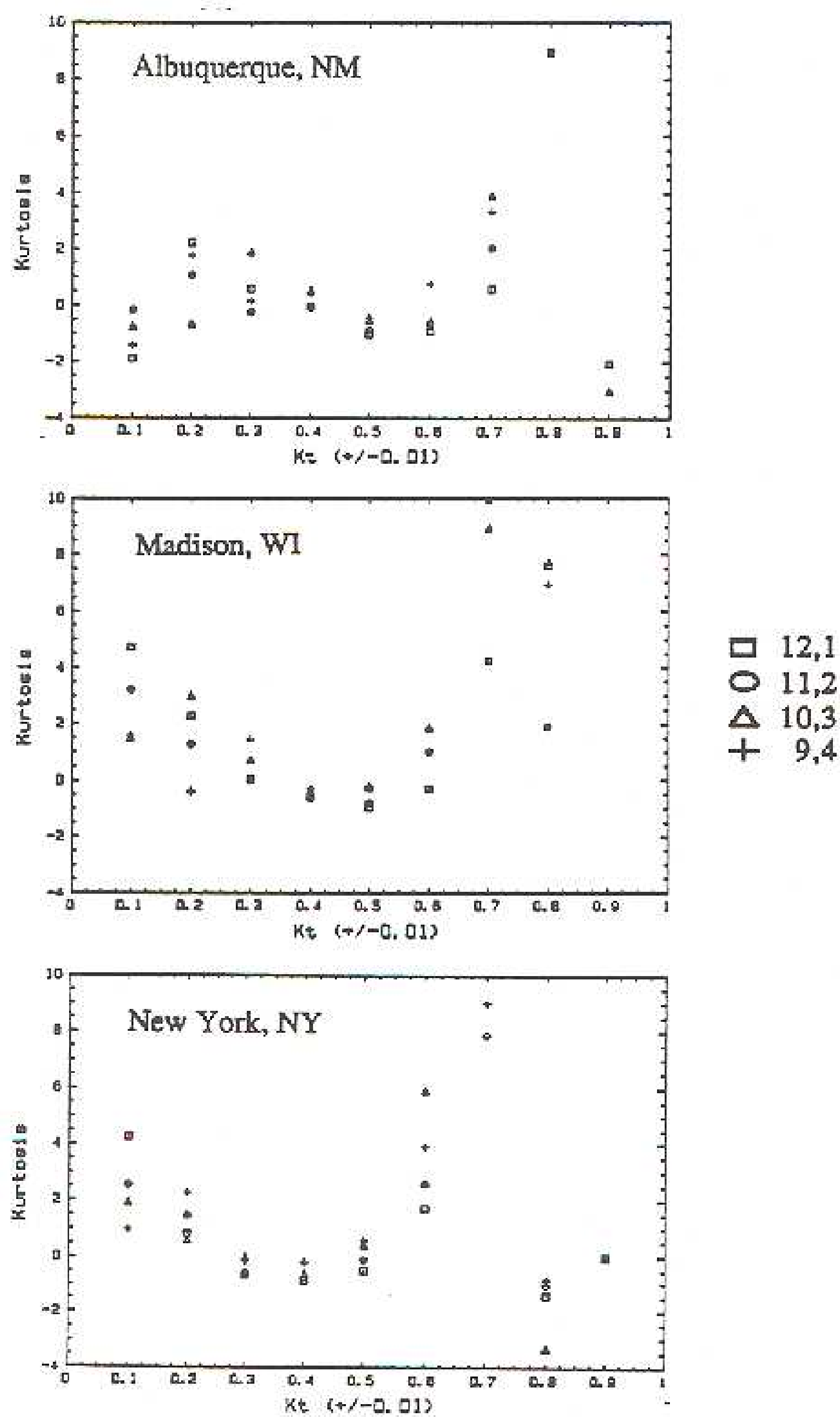

Figure 3.21 Kurtosis of k_t about k_{tm} as a Function of Hour Pair and K_t

3.2.4 The Graham k_t Model

The stochastic k_t model as developed by Graham[1985], rather than modeling the series of k_t values directly, transforms the k_t values to a sequence of normally distributed values which can be represented by an AR1 model.

As was demonstrated in Section 3.2.3, the distribution of k_t about k_{tm} is dependent on both the hour of the day and the daily value of K_t. Modeling the k_t's therefore requires modeling a variable whose probability structure changes each hour and each day. The k_t sequence is not stationary, and due to its non-Gaussian distribution would surely have non-Gaussian residuals when modeled by an AR process. To eliminate this problem, Graham transforms the k_t values through their cumulative distribution function to a normally distributed variable, χ, with mean 0 and variance 1. This transformed variable can then be represented by an AR1 model:

$$x_t = \phi x_{t-1} + \varepsilon_t \tag{3.2.15}$$

In Graham's modeling of the daily K_t sequence in a similar manner, the AR parameter, ϕ, was found to be not significantly different from the lag one autocorrelation of the K_t sequence. Assuming that again the effect of the transformation on the lag one autocorrelation value will be small, Graham recommends the use of ρ_1 as calculated from Equation (3.2.7) to represent the parameter ϕ.

To generate the k_t values, each hour a χ value is obtained by randomly selecting a value for ε_t from a Gaussian distribution and applying Equation (3.2.15), where the mean of ε_t is zero and the variance is equal to $1-\phi^2$. χ is transformed to the non-Gaussian k_t by equating the cumulative distribution functions.

The expression for the normal (or Gaussian) cumulative distribution function with a mean of 0 and variance of 1 is

$$F_{normal} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\chi} e^{-t^2/2}\, dt = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\chi}{\sqrt{2}}\right)\right] \qquad (3.2.16)$$

where

$$\mathrm{erf}\,(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2}\, dt \qquad (3.2.17)$$

The cumulative distribution function of the k_t distribution is represented by

$$F_{kt} = \frac{1}{1 + \exp\,(-1.585\, h_{kt})} \qquad (3.2.18)$$

where

$$h_{kt} = (k_t - k_{tm}) / \sigma$$

which is not the distribution equation used by Graham, but the expression developed in Section 3.2.3. k_{tm} is estimated from Equation (3.2.1) and σ from Equation (3.2.4).

Equating the cumulative distribution functions and solving for k_t yields:

$$k_t = k_{tm} - \frac{\sigma_a}{1.585} \ln\left[\frac{1}{0.5\left[1 + \operatorname{erf}\left(\frac{\chi}{\sqrt{2}}\right)\right]} - 1\right] \tag{3.2.19}$$

Since the sequence of k_t's is not continuous, a new sequence of χ's must be generated each day (the last χ of a day should not be used as the χ_{t-1} for the first hour of the next day).

3.2.5 Adaption to the Degelman generator

The hourly k_t model of Section 3.2.4 requires as inputs only the daily value of K_t and various quantities which are calculated from latitude and time. The Graham stochastic k_t model is easily substituted in place of Degelman's hourly radiation model, as the daily value of K_t is conveniently generated in the daily portion of the Degelman generation program.

3.2.6 Diffuse Radiation

The diffuse radiation each hour is computed deterministically using the hourly diffuse fraction correlation developed by Erbs[1982]. While realistically there is considerable variation of the diffuse fraction about the long-term average values , a study by Hollands and Crha[1987] indicates that the error due to neglecting this variation is on the order of 5% or less.

3.2.7 Correction of hourly k_t values

The stochastic approach used to generate the hourly k_t values does not ensure that a day's total radiation as obtained by summing the hourly values is equivalent to the daily radiation value as set forth in the daily portion of the generation process. To correct this, the entire day's k_t values are generated on the first hour of each day, converted to radiation values and summed to yield the total daily radiation. A correction factor equal to the ratio of the "target" daily radiation value (the value calculated in the daily portion) to the actual generated daily radiation value is computed. Each hourly k_t value is then multiplied by the correction factor, such that the generated daily radiation and the "target" daily radiation value are the same. Checks are included to make sure that no hour's k_t value exceeds a maximum value of 0.964 (the maximum k_t value observed by Graham; see Equation (3.2.9)). This correction has an insignificant effect on the diurnal variation and the hourly autocorrelation of the radiation values, while ensuring that the long-term daily statistics are maintained.

3.3 Comparison

Much of the statistical data collected from the long-term weather records for Albuquerque NM, Madison WI, and New York, NY is presented along with the statistics of the generated and TMY data in the Appendices. Appendices A through D contain the numerical quantities used to describe the variables, that is, the means, standard deviations, skewness, kurtosis, lag one and two daily autocorrelations, lag zero daily cross-correlations, and lag one, two, and 24 hourly autocorrelations for clearness index, ambient temperature, relative humidity, and windspeed. Appendices E and F contain plots of the daily distributions and monthly-average diurnal variations for the four variables at the three locations. The mean has not been subtracted from the distribution curves, so that a difference in the mean will shift a curve and may give a curve the appearance of having the wrong shape. The diurnal variation curves, on the other hand, are presented with the mean subtracted.

The monthly-average K_t values, $\overline{K}_T$, of the generated data are listed in Appendix A along with the means as computed from the TMY data and the long-term data. For the long-term data, however, two values are presented: the long-term average and the average of the monthly-average values, called the average long-term value. For the means, these values will be the same, however, for a statistic such as the maximums, the long-term maximum is different from the average of the monthly maximums. Also listed is the standard deviation of the monthly values about their long-term average values.

Since the hourly radiation values were generated at the beginning of the day and corrected so as to match the daily value determined from the K_t distribution, the monthly-average K_t values are virtually identical for the generated and long-term data. Slight differences are apparent due to round-off errors and the limits imposed on the corrected hourly values ($0 \leq k_t \leq 1$); the correction process does not always yield the exact total. The differences, however, are negligible. The TMY values differ only slightly; most are within 0.02 of the long-term value.

The standard deviations of the K_t values obtained from the different data sets are all approximately equivalent. The skewness of the New York data is better replicated by the TMY data, as it reflects some of the location peculiarities unable to be replicated by the generator; for Albuquerque and Madison significant differences are apparent only for several months. The minimum and maximum K_t values for each generated month should be equal to the average yearly maximum value rather than the long-term value. The object of this weather data generator is to synthesize "average" weather conditions, not to test the extremes. The minimum values for

Albuquerque were close to the long-term yearly average values. For Madison, the generated minimums were consistently low (particularly for the summer months); the TMY data provided a better approximation. For the New York data, the generated minimums were high in the winter and low in the summer, although the differences are only about 0.02. The generated maximums were high for both Madison and New York, while for Albuquerque they were approximately equivalent.

The lag one autocorrelation values are better replicated by the generated data at all three locations. For example, the lag one autocorrelation value for September in Albuquerque is -0.20 while the average long-term value is +0.22. The year-to-year variation in the K_t autocorrelation coefficient is high, which accounts for the wide variety of autocorrelation values computed from the TMY data. The generated autocorrelations tend to be high, particularly for New York; the long-term values calculated for these locations tended to be lower than the "location average" value of approximately 0.29.

The plots of the average diurnal variation of the generated, TMY, and long-term data can be seen in Appendix E. For Albuquerque and Madison, the TMY data shows the most scatter about the correlation lines, however, for New York, the generated data exhibited considerable scatter, particularly at the noon hour. The reason for this has not been determined, however it is worth noting that it is still comparable in quality to the TMY data for New York.

Examination of the RMS error associated with the K_t distributions revealed that the generated distributions generally provided a slightly better approximation to the long-term curve than did the TMY data in Madison; in New York, the opposite was true, while for Albuquerque the TMY and generated data were equally able to reproduce the long-term distribution. The reason that the TMY data is at times better at reproducing the long-term distribution is due to the limitations of the Bendt correlation in describing the long-term distribution. An example is the August distribution in New York (Figure 3.22); the shape of the long-term distribution is noticeably different from that of the correlation (and therefore the generated data). The TMY data in this instance, while differing from the long-term, tends to better follow the shape of the long-term curve. February in Madison, however, is a case in which the generated distribution is a better replication than the TMY (Figure 3.23). Plots of the other months are included in Appendix F.

The lag one autocorrelations of the generated hourly a_{kt} values are slightly low in comparison to the long-term values. This result was surprising; since the values were generated from a stochastic model, the correct autocorrelation would be expected. Several possible explanations were considered. First, it is possible that the estimate of the autocorrelation as obtained from Equation (3.2.6) might not be an accurate estimate. A simple test was performed in which 5 years of hourly values were calculated from an AR1 model with ϕ_1=0.54. Fourteen values were removed every 10 values, creating a discontinuous sequence similar to that of the hourly k_t values. The autocorrelation of the full series was computed and compared to the autocorrelation of the fragmented series as computed by Equation (3.2.6). Both estimates were equal to 0.54, indicating that Equation (3.2.6) is a valid estimator. Another possible explanation, and the most probable, is that Graham's assumption that the transformation to the Gaussian domain has no effect on the AR1 parameter is

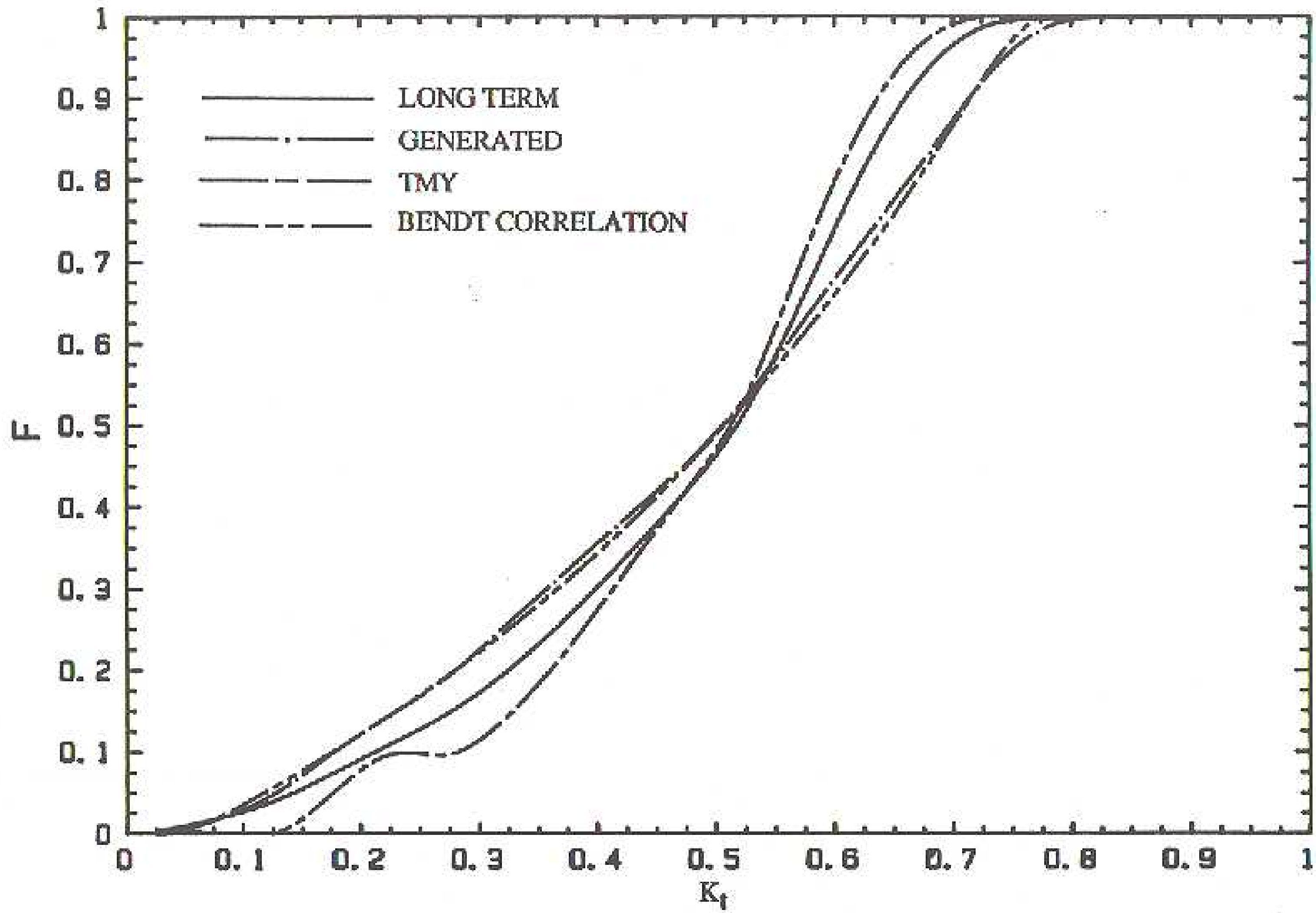

Figure 3.22 Comparison of Long-Term, Generated, and TMY Daily Clearness Index Distributions for August in New York, NY

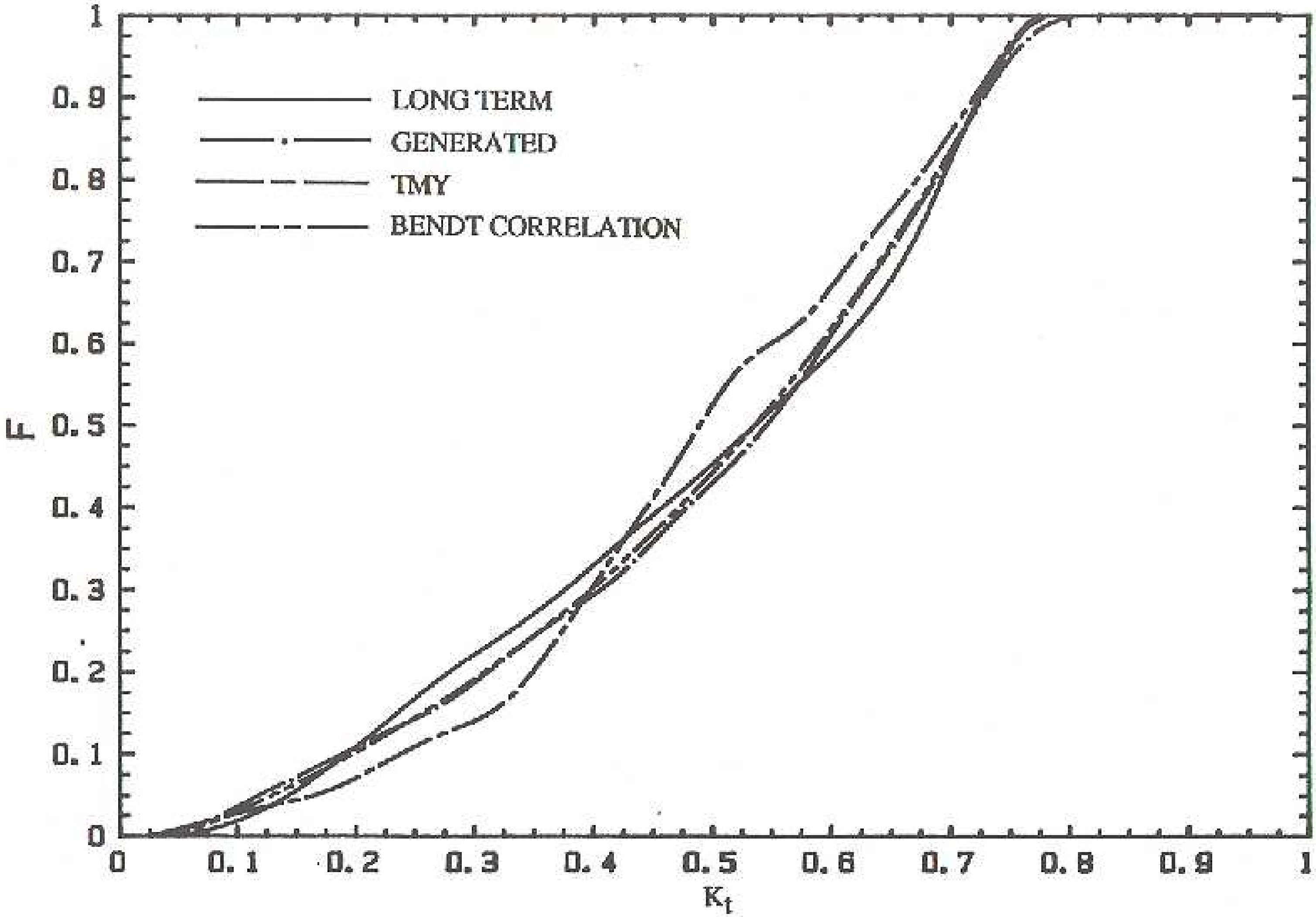

Figure 3.23 Comparison of Long-Term, Generated, and TMY Daily Clearness Index Distributions for February in Madison, WI

incorrect, and that the transformation causes a lower value of the autocorrelation to be computed upon generation.

The cross-correlation between K_t and $\bar{T}$ for the three locations examined was positive during the summer and negative during the winter, however the magnitude of the cross-correlations for a specific month differed between the locations. The TMY data tended to follow this pattern slightly better than did the generated data, however the magnitudes were often greater than those calculated from the long-term values. The standard deviation of the long-term estimates is high, so again a wide variation in the TMY values would be expected. This also indicates that the degree of uncertainty associated with the long-term value is greater.

Cross-correlations between daily-average windspeed and daily clearness index are small (0.25 or less) and negative for Albuquerque and Madison, but small and mostly positive for New York. The standard deviations of the monthly cross-correlations about their average value is also on the order of 0.25; there is considerable year-to-year variation. This is reflected in the TMY data, where values range from -0.52 to +0.57. The generated data, while not matching the long-term exactly, tends to be less than 0.30 in magnitude and negative, thus generally replicating the long-term values better than the TMY

The cross-correlations between daily-average relative humidity and daily clearness index calculated from the long-term data are strong and negative for all three locations. For Madison and New York, the values were approximately constant, ranging from -0.62 to -0.77, while the values for Albuquerque were similarly constant, however, at a value of -0.52. The standard deviations are relatively small (about 0.15), indicating that the values are significant. Intuitively this observation makes sense; the more water in the atmosphere, the smaller the transmittance. The generated values are small, indicating no cross-correlation within the generation model. The TMY values are similar to the long-term, but like the cross-correlations between other variables, the variation is wider.

Overall, after comparing the generated statistics with the long-term statistics, the generated radiation data is quite good. The only major weaknesses are in the hourly autocorrelation which is slightly low and in the cross-correlation between daily-average relative humidity and daily clearness index; the effect of both of these is unknown. Other differences between the generated data and the long-term data are due to the inability of the location-independent correlations to fully describe the weather pattern at a particular site; only by further improving on these correlations can improvements be made.

CHAPTER 4

Ambient Temperature Generation

This chapter describes the procedure by which the hourly ambient temperature values are generated in Degelman's model, including a critique of the resulting data, and presents an alternative stochastic model, with the results compared to long-term data, TMY data, and correlations.

4.1 Degelman Temperature Model

Similar to the generation of radiation values, daily-average ambient temperatures are generated first, and the hourly ambient temperatures are then obtained from a decomposition of the daily values.

4.1.1 Input Reducing Modifications

In the version of the weather data generator developed by Degelman, there were several additional required inputs, namely the standard deviation of the daily-average ambient temperatures about their monthly-average values, the monthly-average maximum daily temperature, and the standard deviation of the maximum daily temperatures about their monthly-average values. The particular version which was used in this study had been modified by Erbs to eliminate these inputs.

The standard deviation of the daily-average temperatures about their monthly-average values is different from the standard deviation estimated in Equation (2.2.7), but it is related. Erbs[unpublished] developed a similar expression relating the daily standard deviation, σ_d, to the monthly-average ambient temperature and the standard deviation of the monthly-average temperatures about the yearly-average value, σ_{yr}:

$$\sigma_d = 9.273 - 0.07952\,\overline{T}_m + 0.0097111\,\sigma_{yr} \tag{4.1.1}$$

The monthly-average maximum daily temperature can be estimated from the monthly-average ambient temperature and the monthly-average clearness index. The peak-to-peak amplitude of the monthly-average diurnal variation of ambient temperature is approximated from Equation (2.2.9); adding half of this amplitude to the monthly-average ambient temperature yields an estimation of the monthly-average maximum daily temperature.

$$\overline{T}_{max} = \overline{T}_m + \frac{1}{2}A \tag{4.1.2}$$

The standard deviation of the daily maximum temperatures about the monthly-average daily maximum temperature is estimated from the standard deviation of the daily-average ambient temperatures about the monthly-average value.

$$\sigma_{max}(i) = \frac{\sigma_d(i) - C_1(i)}{C_2(i)} \tag{4.1.3}$$

where i indicates the month and C_1 and C_2 are constants determined from Table 4.1.

Table 4.1
Coefficients used to estimate the standard deviation of the maximum daily ambient temperatures from the standard deviation of the average daily ambient temperature

Month	C_1	C_2
1	0.8033	0.7900
2	1.8151	0.6899
3	2.9018	0.5823
4	2.397	0.6053
5	0.6159	0.7343
6	0.3275	0.7920
7	1.2187	0.9717
8	1.4222	1.0269
9	0.2053	0.8326
10	3.0621	0.5167
11	3.0242	0.5811
12	5.0849	0.3777

Equations (4.1.1), (4.1.2), and (4.1.3) were included in the generator to replace the requirement of inputting the σ_d, $\overline{T}_{max}$, and σ_{max} values, respectively.

4.1.2 Daily Temperature Generation

The same procedure used to determine the daily K_t values is applied by Degelman to generate the daily-average ambient temperatures. Specifically, daily values are obtained from a distribution and ordered by a fixed sequence to replicate the daily autocorrelation.

The shape of the daily distribution is assumed to be normal; while this is not the case, and was recognized by Degelman, no formulation for the temperature distribution existed at the time (Degelman's model was completed prior to the work of Erbs[1984]). Due to the similarity in the two distributions, the assumption of normality is reasonable. Knowledge of both the mean and standard deviation is required to uniquely specify the normal distribution. The mean is simply the monthly-average ambient temperature, which is input, and the standard deviation is calculated internally from Equation (4.1.1).

Conceptually, generation of the daily-average ambient temperatures for a month consists of two steps, 1) obtaining N values from the distribution in a manner such that the N values will recreate the distribution, where N is the number of days in the month, and 2) ordering these N values such that they produce the correct daily autocorrelation. As was done with the K_t values, the order is determined from a fixed sequence, however the sequence used to order the ambient temperature values is a distinctly different sequence from the one used to order the K_t sequence.

Like the ordering of the K_t values, the starting position within the sequence is determined at the beginning of the generation process. However, for all of the sequences used in the generator, the starting position is the same. Therefore, by careful selection of the sequences, daily cross-correlations may be introduced.

In practice, rather than first obtaining all N temperature values and then ordering them, the temperature values are selected from the distribution in the order as specified by the sequence, allowing the daily-average value for each day to be determined on that day.

In addition to determining the daily-average ambient temperatures for each day of the month, the daily maximum temperatures are also generated in a similar manner. The distribution of the maximum temperatures is assumed to be normal, with the mean and standard deviation calculated from Equations (4.1.2) and (4.1.3), respectively.

A different sequence is used to order the daily maximum temperature values, however, the sequence was chosen such that significant correlation exists between the daily maximum and average values.

The minimum daily ambient temperatures are calculated by assuming the daily mean is equivalent to the median, meaning that the daily-average temperature is equal to the average of the daily maximum and minimum.

4.1.3 Critique of the Daily Generation Model

The normal distribution used by Degelman to generate the daily-average ambient temperatures is compared with the average long-term distribution in Figure 4.1; the difference is significant. Both distributions are plotted with a mean of 0 and standard deviation equal to 1, and both are symmetric about 0. The difference between the distribution curves is an example of a difference in kurtosis.

The daily lag one autocorrelation values as calculated from the generated data are in the range of 0.60 to 0.70, which are equivalent to values calculated from long-term data.

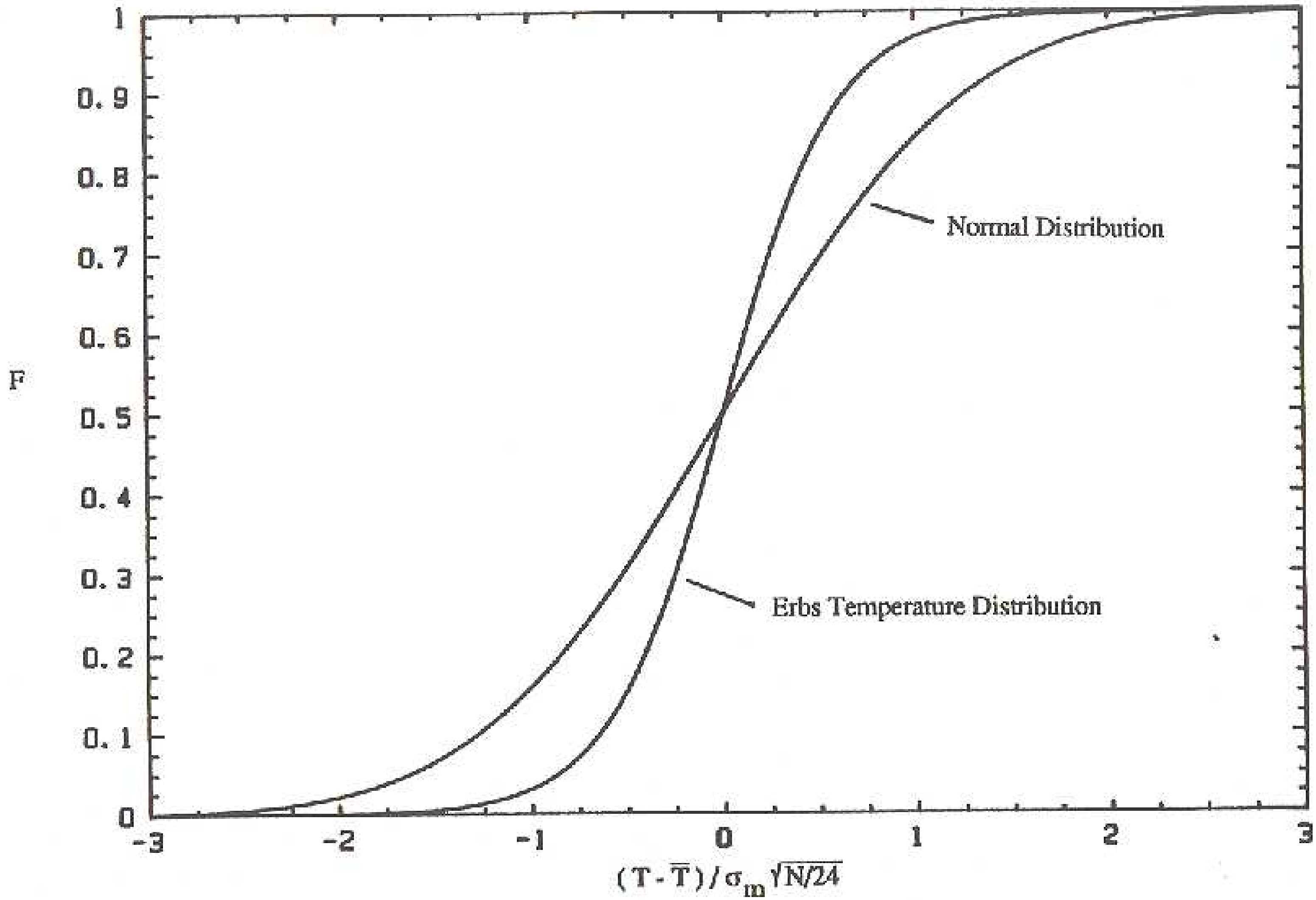

Figure 4.1 Comparison of Erbs Ambient Temperature Distribution to the Normal Distribution

4.1.4 Hourly Ambient Temperature Generation

To generate the hourly ambient temperature values, Degelman uses a cosine interpolation between the maximum and minimum daily temperatures. This produces a continuous series of hourly temperatures, in the sense that there are smooth transitions between adjacent days.

The minimum hourly ambient temperature is assumed to be the value for the hour in which the sun rises; the maximum hourly ambient temperature is assumed to be the value for 3 pm. All of the daily values are generated one day in advance, so that the following day's minimum temperature is available for the interpolation between the maximum of one day and the minimum of the next day.

The equation used to calculate the temperature each hour is:

$$T = T_{ave} + A \cos\left(\frac{\pi \Delta}{R}\right) \tag{4.1.4}$$

T_{ave} is the median ambient temperature for the particular portion of the day. For the hours between sunrise (the minimum temperature) and 3 pm (the maximum temperature), T_{ave} is equivalent to the daily-average ambient temperature. For the hours between 3 pm and midnight, T_{ave} is *not* the daily-average ambient temperature, but the average of the maximum temperature and the next day's minimum temperature. Likewise, for the hours between midnight and sunrise, T_{ave} is the average of the minimum temperature and the previous day's maximum temperature. In other words, as far as the hourly temperatures are concerned, the month consists not of 31 daily-average ambient temperatures, but of 31 sunrise-to-3pm average temperatures and 31 3pm-to-sunrise average temperatures.

R is the number of hours in the appropriate time period; Δ is equal to the number of hours into the time period.

A is the amplitude, likewise defined as the amplitude for the particular portion of the day. This is not the peak-to-peak amplitude, but one half of the difference between the appropriate maximum and minimum. To coincide with the definitions of R and Δ, the amplitude is negative during the sunrise-to-3pm period.

4.1.5 Critique of Generated Hourly Ambient Temperature Data

The cosine interpolation method produces a diurnal variation, but not the long-term monthly-average diurnal variation documented by Erbs[1984]. An example of the generated diurnal variation and the Erbs correlation is shown in Figure 4.2.

Additionally, the hourly ambient temperatures as generated by this method are purely deterministic. In the case of hourly radiation, simulation prediction errors can result from neglecting the random deviations from the long-term mean. It would be reasonable, then, to suspect that reproducing the randomness within the hourly ambient temperature structure might also be important.

The daily-average temperature as obtained by summing the hourly values is not identical to the "target" daily-average temperature value as set forth in the daily portion of the temperature generation. However, the differences are small, such that

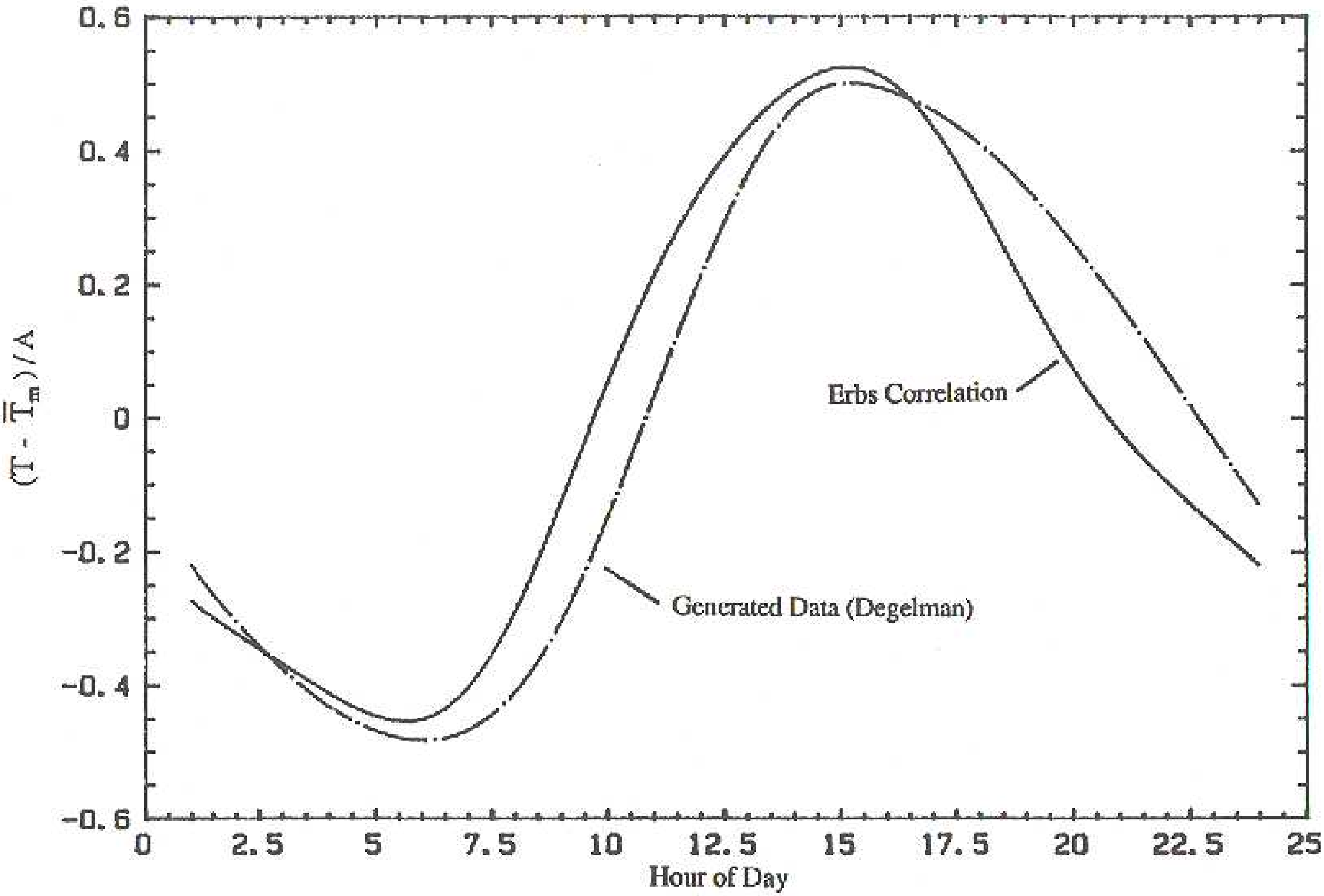

Figure 4.2 Comparison of Erbs Ambient Temperature Diurnal Variation to the Diurnal Variation Obtained from Degelman's Weather Generation Program

neither the shape of the distribution or the daily autocorrelation are significantly affected.

4.2 Modification of the Ambient Temperature Model

The deterministic approach employed by Degelman does not attempt to preserve the autocorrelation structure of the hourly ambient temperatures or the long-term distribution and diurnal variation documented by Erbs[1984]. While it is unknown exactly what the effect of neglecting the randomness within the hourly ambient temperatures will be, a truly representative weather model should include this aspect of the temperature variable.

Since one of the desired features of the temperature model includes reproducing the autocorrelated hourly deviations from the average diurnal variation, some type of stochastic model is indicated.

4.2.1 The Hourly Ambient Temperature Sequence

The hourly ambient temperature values from the Albuquerque NM, Madison WI, and Miami FL TMY data were analyzed to further understand the behavior of the hourly ambient temperature sequence and to attempt to improve upon the deterministic model.

The first step in the analysis of any data set is to plot the data. Rather then plotting one entire year of hourly temperature data, the data was plotted month by month and examined. Immediately apparent is the strong diurnal variation and the difference in amplitude of the diurnal variation throughout the year. As observed by Erbs[1984], the amplitude appears to vary as a function of $\overline{K}_T$. As expected, an annual variation is also apparent, meaning that the mean is dependent on the time of year.

Several days of hourly data were plotted at a time to allow a more detailed examination. While the diurnal variation is strong, there are noticeable deviations, particularly during the months in which the amplitude is small. There are many days in which the pattern of the ambient temperature values throughout the day does not resemble the monthly-average diurnal variation. However, since the degree to which the diurnal variation is degraded appears to be somewhat dependent on the size of the amplitude, this suggests that the size of the deviations from the monthly-average diurnal variation are roughly constant throughout the year. In other words, this suggests that the variance of the deviations (a_t) of the hourly ambient temperatures from their monthly-average hourly values may be approximately constant.

4.2.2 Autocorrelation of the Hourly Sequence

Due to the presence of the diurnal variation, calculating the autocorrelation of the raw hourly temperature data does not yield much information about the deviations from the monthly-average diurnal variation. More information can be gained by first removing the diurnal trend and then computing the autocorrelation.

The question then arises as to how to remove the diurnal variation, or more directly, how to compute the "average" value for an hour. Considering that the eventual goal is to be able to generate the hourly temperature values, somehow subtracting the average diurnal variation seems appropriate, since for generation it can be determined from Equation (2.2.8).

One possible technique involves computing an average diurnal variation for each day, using $\overline{T}$ (the daily-average ambient temperature) for $\overline{T}_m$ in Equation (2.2.6). The amplitude could be computed in one of two ways; by either using the monthly-average amplitude as predicted from Equation (2.2.9) or by using the actual amplitude as determined from the temperature data for the day. Thus, the "mean" value for each hour of a day could be determined from Equation (2.2.8). The reasoning here is that the temperature series could be reproduced simply by computing the daily-average and maximum temperatures from Degelman's fixed sequence method, with the amplitudes computed from the maximums and averages. The deviations could be computed from a time series model and added to the "mean" hourly values to produce simulated temperatures.

The problem with this method is that in calculating the average diurnal variation for each day from Equation (2.2.8), given the daily-average and the amplitude, a discontinuity results between the last temperature of one day and the first temperature of the next day. Such step changes have an insignificant effect when they occur every few days [Arens et al, 1980], however, when these discontinuities occur every day, they would undoubtedly have a more significant effect.

An alternative is to instead subtract the monthly-average hourly value from each hourly temperature. This technique largely removes the diurnal trend, however the autocorrelation of the hourly deviations as calculated in this manner is somewhat confounded with the autocorrelation of the daily-average temperatures. For example, if the average temperature for a day is higher then the monthly-average value, then it is likely all of the hours will also be higher than the monthly-average hourly values.

Subtracting the daily means would eliminate this problem, but as discussed previously, would introduce discontinuities. A stochastic model should be able to replicate this series (temperature minus hourly monthly-average values), and this was the technique used to remove the diurnal trend. The lag one hourly autocorrelation values were observed to vary between 0.95 and 0.98 independent of location or month.

Also to be considered is the annual variation. It was originally unclear as to whether or not neglecting this variation would leave a trend remaining within each month. As a check, Fourier coefficients were determined from the Madison and Albuquerque TMY hourly ambient temperature data. Yearly average values for each hour were then computed and subtracted from the data. Hourly monthly-average values were then calculated and subtracted. Lag one to twenty-four autocorrelations of this twice de-trended series were compared to those of a series de-trended only by subtracting the hourly monthly-average temperatures. No significant differences in the autocorrelation values were observed, therefore it was assumed that the annual variation could be neglected. That is, the long-term daily temperature can be assumed to have a constant value for all days within a month.

4.2.3 Model Development

A logical extension of the method for generating the hourly k_t values would be to generate the hourly ambient temperatures from the daily-average values in a similar manner. The temperature series, however, is different from the k_t series in that it is continuous. Once the daily-average is fixed, it is difficult to generate hourly values stochastically without introducing discontinuities between the last hour of one day and the first hour of the next day.

An alternative to predetermining the daily values and ordering then with a 31-day sequence is to predetermine the 24 hourly monthly-average temperatures from the diurnal temperature variation relation proposed by Erbs[1984]. The hourly temperatures can then be generated by a stochastic model which replicates the autocorrelated hourly deviations from the monthly-average diurnal variation. The autocorrelation function to be analyzed for such a model is the autocorrelation as calculated in Section 4.2.2, that is, the autocorrelation of the sequence of hourly temperatures minus their hourly monthly-average values.

In stochastic modeling, the series to be modeled must be stationary and the residuals must be Gaussian and uncorrelated. Stationarity means that the mean and variance of the series must be independent of time. Eliminating the diurnal variation by subtracting the hourly monthly-average values removes dependence of the mean on time, while the variance can be assumed constant throughout a month. To ensure Gaussian residuals, this implies that the series itself should be Gaussian; the distribution of the hourly ambient temperatures is not Gaussian, however it is similar.

TMY data from Albuquerque NM, Madison WI, and Miami FL were used in the identification and fitting of the model rather than the long-term data, so that the long-term data could be used for an evaluation of the model performance.

For a moving average (MA) model, the autocorrelation function (acf) can be used to tentatively identify the appropriate model, as the acf will be approximately zero after the n^{th} lag. Similarly , an AR model can be identified by examination of the partial autocorrelation function [Box and Jenkins, 1976]. An ARn model exhibits a partial autocorrelation function which "cuts off" after lag n. The values of the pacf can also be used as initial estimates of the AR parameters.

Monthly plots of the acf of the hourly time series (temperature - hourly monthly-average) show a slowly decaying function. This indicates that the series is close to unstationary; a better technique of detrending would help make the series more stationary. Examination of the acf does not indicate a MA model; neither is an AR1 model indicated, as the acf of an AR1 model should be an exponential decay. This indicates either a higher order AR or a mixed ARMA model. The pacf of the time series indicates that possibly an AR2 or AR3 model would suffice. Sample plots of the autocorrelation function, partial autocorrelation function, and the residual autocorrelation function are shown in Figures 4.3 A, B, and C. The set of plots representing August in Madison are an example of a month to which the model fits well. The pacf is essentially zero after lag 2, indicating an AR2 model, while the residuals are uncorrelated. January in Miami shows similar results, however the autocorrelation function tends to suggest examining an AR3 model. October in Albuquerque is an example of a month for which the fit was not particularly good; there are several significant autocorrelations apparent at lags past lag two, and the residuals show some autocorrelation. Overall, October in Albuquerque represents the worst months while January in Madison is indicative of the best months.

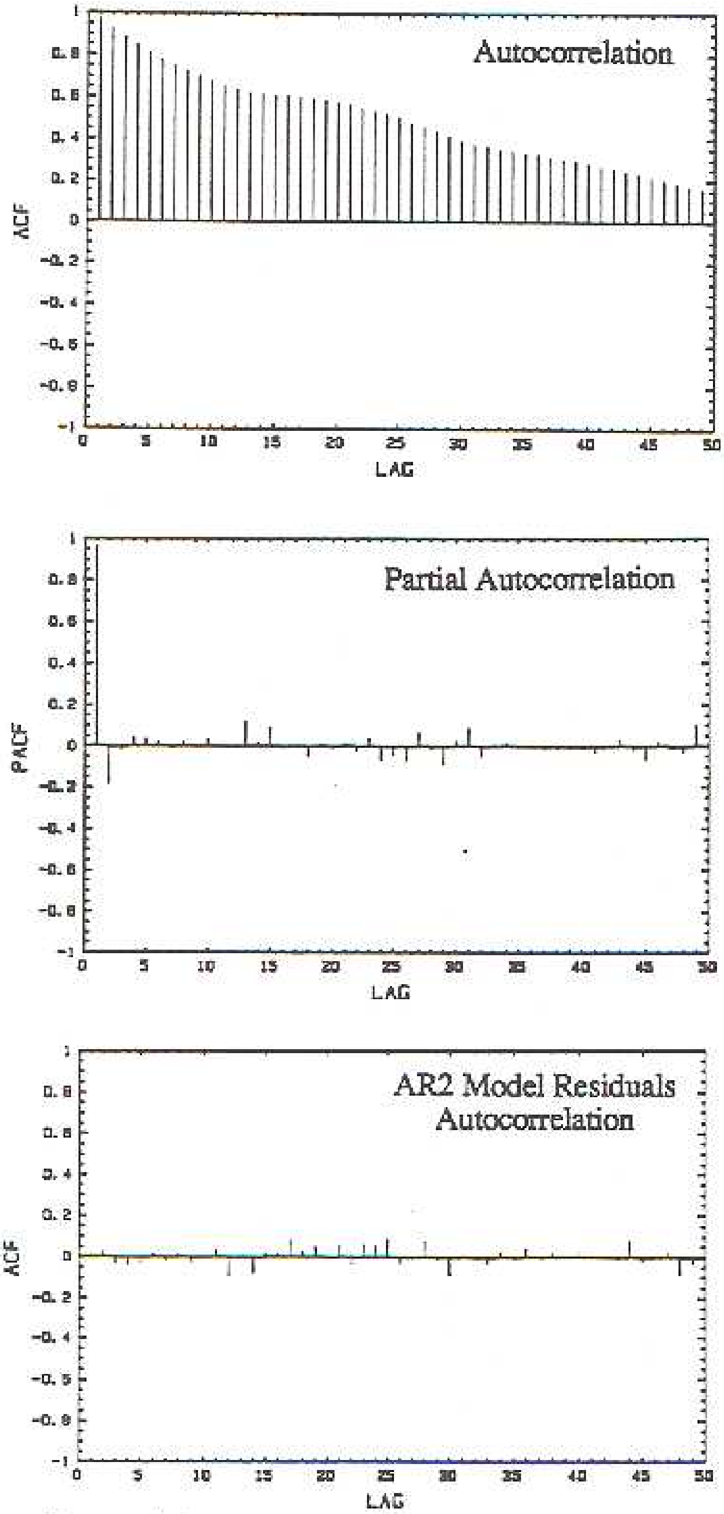

Figure 4.3A De-trended Hourly Ambient Temperature Autocorrelation Function, Partial Autocorrelation Function and Autocorrelation Function of the AR2 Model Residuals for August in Madison, WI (TMY Data)

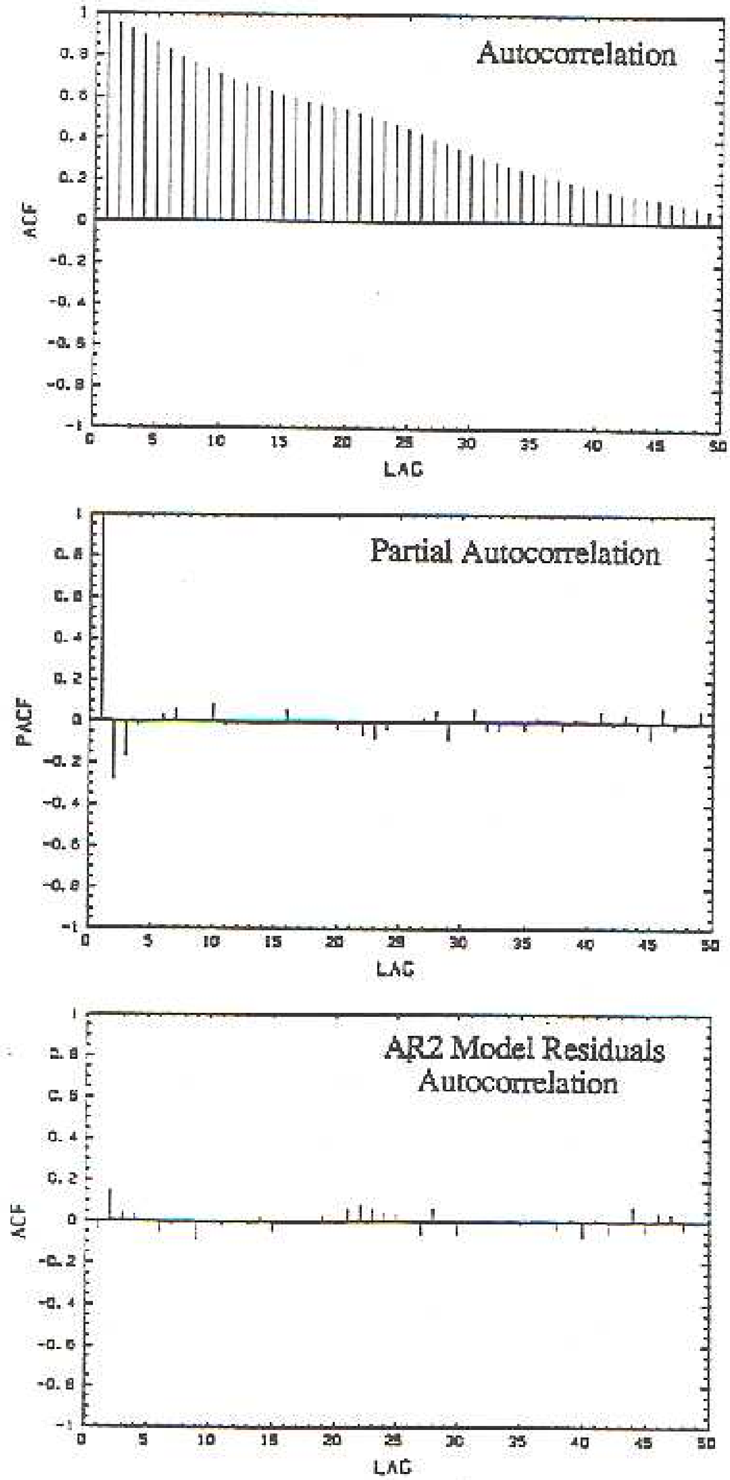

Figure 4.3B De-trended Hourly Ambient Temperature Autocorrelation Function, Partial Autocorrelation Function and Autocorrelation Function of the AR2 Model Residuals for January in Miami, FL (TMY Data)

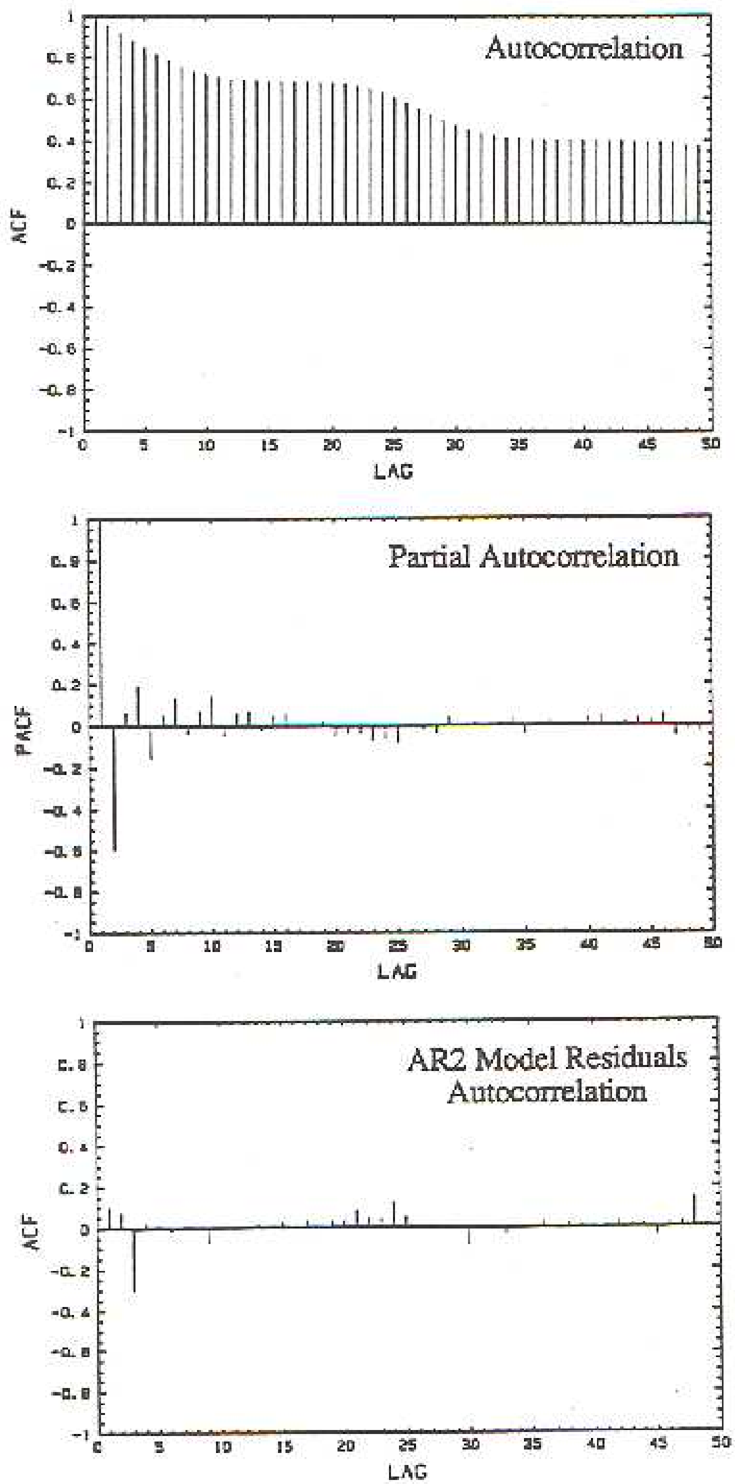

Figure 4.3C De-trended Hourly Ambient Temperature Autocorrelation Function, Partial Autocorrelation Function and Autocorrelation Function of the AR2 Model Residuals for October in Albuquerque, NM (TMY Data)

A previous study of the hourly ambient temperature sequence was conducted by Hittle and Pedersen[1981]. Hourly weather data from four U.S. locations, Charleston SC, Fort Worth TX, Santa Maria CA, and Madison WI were analyzed. The approach employed by Hittle and Pedersen consisted of three stages, 1) deterministic modeling through the use of the discrete Fourier transform, 2) stochastic modeling by an ARMA(2n,2n-1) model, and 3) modeling with a combined deterministic-stochastic model. Analysis of the ambient temperature sequence and the adequacy of the various fitted models indicated the third technique to be the most appropriate. The deterministic portion represented the diurnal variation, and was modeled by a Fourier series; the stochastic portion modeled the residuals obtained after removing the diurnal trend. An AR2 model was found by Hittle and Pedersen to be appropriate in describing the residual behavior at all of the locations. The autoregressive parameters, however, were different for the four locations and hence this model could not be directly inserted into the weather generation model.

The analysis of the Madison, Albuquerque, and Miami TMY data was similar to that of Hittle and Pedersen, except that rather than using a Fourier series to remove the diurnal variation, the hourly monthly-average values were simply subtracted. As previously noted, either an AR2 or AR3 model seemed to be indicated for modeling the residuals. Since Hittle and Pedersen also found the AR2 model to be appropriate, an AR2 model was developed to model the hourly ambient temperature deviations.

The AR parameters were determined for each month by a linear least squares regression. Similar to the results obtained by Hittle and Pedersen, the parameters were different for the various months and locations. Inspection of the parameters, however, indicated some degree of compensation between them, such that it might be possible to use a single set of parameters for all locations and all months without significantly affecting the fit of the model. The monthly parameters for the three locations are shown in Table 4.2.

Table 4.2
AR2 coefficients for each month at the three locations

	Madison WI		Albuquerque NM		Miami, FL	
	ϕ_1	ϕ_2	ϕ_1	ϕ_2	ϕ_1	ϕ_2
Jan	1.69	-0.70	0.96	0.00018	1.26	-0.28
Feb	1.25	-0.28	1.00	-0.018	1.71	-0.72
Mar	1.73	-0.74	1.67	-0.68	1.65	-0.70
Apr	1.34	-0.36	1.58	-0.61	1.03	-0.077
May	1.23	-0.25	0.98	-0.017	0.91	0.00052
Jun	1.19	-0.22	1.56	-0.59	0.82	-0.037
Jul	1.64	-0.67	0.92	0.0022	0.80	-0.00058
Aug	1.14	-0.18	0.86	0.038	0.89	-0.093
Sep	1.12	-0.15	1.46	-0.50	0.81	0.031
Oct	1.63	-0.65	1.65	-0.67	1.66	-0.69
Nov	1.73	-0.74	0.92	0.056	1.00	-0.047
Dec	1.18	-0.21	1.02	-0.035	1.73	-0.74

For some of the months, particularly summer months in Albuquerque and Miami, the second AR parameter is approximately equal to zero, indicating that an AR1 model might be adequate for those months. However, for many of the other months, the second parameter is significant, indicating the need for the AR2 model. It is interesting to note that there does not appear to be any pattern to the variation in the parameters. For example, in Madison, the January parameters are the same as the July parameters; no dependence on time of year is indicated. Also, the variation between months at one location is just as great as the variation between different locations, supporting the notion of location-independent parameters.

Parameters were then fit to the entire year's data at one time for each location. The results are listed in Table 4.3. Also included are parameter estimates from two years of Madison data, 1953 and 1970, and the values estimated by Hittle and Pedersen. While Madison, Albuquerque, and Miami are vastly different climates, the parameters are similar. Also, the parameters estimated from the different years of Madison data indicate that the variability from year to year in the estimates is considerable, more so than the variation between locations. The parameters estimated by Hittle and Pedersen are also very similar to those estimated in this study.

To test if any set of parameters could be used without significantly affecting the fit of the model, the different parameters listed in Table 4.3 were fit to the Albuquerque, Madison, and Miami TMY data. The ratio of the increase in sum of squares is listed in Table 4.3, along with the standard deviation of the residuals from each fit. While theoretically any ratio value over 1.00 is significant for such a large set of data, the increase in sum of squares is small and for engineering purposes can be ignored. Looking at the standard deviations shows the effect; at the worst, the standard deviation increases from roughly 0.7 degrees C to 1 degree C; this is very small, particularly in consideration of the accuracy of the measured temperature data.

A single set of parameters was then estimated for the TMY data at all three locations combined. These values are also listed in Table 4.3, along with the resulting sum of squares ratio and residual standard deviations. The increases are small, indicating that a single set of parameters can be used in the AR2 model to adequately

Table 4.3

AR2 parameter fits to the TMY data

Coefficients	ϕ_1	ϕ_2	Madison σ/SS ratio		Albuquerque σ/SS ratio		Miami σ/SS ratio	
Madison TMY	1.350	-0.369	0.79	1.00	0.90	1.07	0.74	1.08
Albuquerque TMY	1.089	-0.117	0.82	1.08	0..87	1.00	0.71	1.00
Miami TMY	1.064	-0.103	0.83	1.11	0.87	1.00	0.71	1.00
Madison 1953	1.172	-0.193	0.80	1.04	0.87	1.01	0.72	1.02
Madison 1970	1.695	-0.708	0.84	1.14	1.02	1.37	0.84	1.40
Fort Worth (Hittle)	1.363	-0.377	0.79	1.00	0.90	1.08	0.74	1.09
Madison (Hittle)	1.18	-0.206	0.80	1.04	0.87	1.01	0.72	1.01
Santa Maria (Hittle)	1.075	-0.283	1.27	2.59	1.11	1.64	0.85	1.42
All TMY	1.178	-0.202	0.80	1.04	0.87	1.01	0.72	1.02

describe the behavior of the temperature deviations at all locations. Tables 4.4 A, B, and C show the sum of squares ratio and the residual standard deviation for each month of the TMY data as the various coefficients are used in the AR2 model. Again, while theoretically the ratio increase is significant, it is minimal and the increase in residual standard deviation is small in comparison with the accuracy of the temperature measurements. The final parameters used in the AR2 model were those obtained from the fit to all of the TMY data, ϕ_1=1.178 and ϕ_2=-0.202.

Table 4.4A

Ratio of sum of squares and residual standard deviation for increasingly general parameter estimates MADISON,WI (TMY Data)

	Monthly TMY		Yearly TMY		Overall TMY	
January	1.000	0.531	1.249	0.593	1.550	0.662
February	1.000	0.820	1.011	0.825	1.009	0.823
March	1.000	0.412	1.365	0.482	1.762	0.548
April	1.000	0.958	1.000	0.959	1.029	0.972
May	1.000	0.992	1.015	0.999	1.004	0.993
June	1.000	0.906	1.027	0.918	1.000	0.906
July	1.000	0.461	1.158	0.497	1.386	0.543
August	1.000	0.892	1.044	0.912	1.003	0.894
September	1.000	0.945	1.052	0.969	1.002	0.946
October	1.000	0.642	1.134	0.684	1.351	0.745
November	1.000	0.449	1.317	0.516	1.669	0.582
December	1.000	0.827	1.034	0.841	1.002	0.828

Table 4.4B

Ratio of sum of squares and residual standard deviation for increasingly general parameter estimates ALBUQUERQUE, NM (TMY Data)

	Monthly TMY		Yearly TMY		Overall TMY	
January	1.000	0.884	1.017	0.892	1.048	0.905
February	1.000	0.973	1.011	0.978	1.035	0.989
March	1.000	0.440	1.664	0.567	1.469	0.533
April	1.000	0.478	1.378	0.562	1.256	0.536
May	1.000	1.124	1.011	1.130	1.037	1.145
June	1.000	0.577	1.357	0.672	1.244	0.643
July	1.000	1.215	1.037	1.237	1.072	1.258
August	1.000	1.059	1.066	1.094	1.113	1.118
September	1.000	0.526	1.197	0.575	1.121	0.556
October	1.000	0.434	1.583	0.544	1.410	0.515
November	1.000	0.859	1.030	0.872	1.068	0.888
December	1.000	0.878	1.007	0.882	1.028	0.891

Table 4.4C
Ratio of sum of squares and residual standard deviation for increasingly general parameter estimates MIAMI, FL (TMY Data)

	Monthly TMY		Yearly TMY		Overall TMY	
January	1.000	0.726	1.054	0.745	1.008	0.729
February	1.000	0.325	1.971	0.455	1.629	0.412
March	1.000	0.320	1.697	0.417	1.487	0.390
April	1.000	0.680	1.003	0.680	1.024	0.688
May	1.000	0.667	1.033	0.678	1.085	0.694
June	1.000	0.977	1.099	1.024	1.162	1.053
July	1.000	0.943	1.105	0.991	1.175	1.021
August	1.000	0.808	1.085	0.841	1.133	0.859
September	1.000	0.752	1.081	0.783	1.150	0.807
October	1.000	0.292	1.683	0.377	1.460	0.352
November	1.000	0.688	1.003	0.690	1.035	0.701
December	1.000	0.344	2.125	0.502	1.705	0.451

The residuals from this model are uncorrelated, but are not completely Gaussian. This problem was also apparent in the modeling of the k_t sequence, and to eliminate it, the k_t variable was transformed to a normally distributed variable, χ, and χ was modeled by an AR process. Likewise, the ambient temperature sequence can be transformed to a normally-distributed variable and modeled in the Gaussian domain. The transformation involves equating the cumulative distribution functions; an expression for the long-term average distribution of hourly ambient temperatures about their monthly-average value was developed by Erbs[1984]. Graham[1985] found the transformation of the K_t variable to a normally distributed variable to have only a small effect on the autocorrelation, and used the parameter estimates obtained from the untransformed data in the AR model. Since the distribution of T is more similar to the normal distribution than was the K_t distribution, the parameters used in the modeling of the χ series will be assumed equal to those obtained from fitting the AR2 model directly to the de-trended temperature data.

4.2.4 Ambient Temperature Generation

To use the AR2 model in the generation process, first, 24 hourly monthly-average temperatures are estimated using Equation (2.2.8). $\overline{T}_m$, the monthly-average daily temperature is known and the amplitude, A, is estimated from Equation (2.2.9).

For each hour of the month, a χ value is generated according to the AR2 model:

$$\chi_t = \phi_1 \chi_{t-1} + \phi_2 \chi_{t-2} + \varepsilon_t \tag{4.2.1}$$

where χ is a normally distributed variable with mean 0 and variance 1, χ_{t-1} is χ from the previous hour , χ_{t-2} is χ from 2 hours ago, and ε_t is a Gaussian disturbance with mean 0 and variance $\sigma^2=1-\phi_1\rho_1-\phi_2\rho_2$ (ρ_1 and ρ_2 are the lag 1 and 2 autocorrelations of χ).

The generated χ value is then transformed to an hourly temperature by equating the cumulative distribution function of χ (Equation 3.2.16) and the cumulative distribution function of hourly temperature. The cdf of the hourly ambient temperature is given by Equation (2.2.5), where for this application, $\overline{T}_h$ is substituted for $\overline{T}_m$ in Equation (2.2.6). The shape of the distribution is the same throughout the month, but the mean is dependent on the hour of the day. Solving for the hourly temperature, T, gives:

$$T = \overline{T}_h - \frac{\sigma_m}{1.698} \ln \left[\frac{1}{0.5 \left[1 + \mathrm{erf}\left(\frac{\chi}{\sqrt{2}} \right) \right]} - 1 \right] \tag{4.2.2}$$

The stochastic model just described does not guarantee that the generated monthly-average daily temperature is as originally specified. This discrepancy can be eliminated in a similar manner as was done with the radiation by comparing the generated monthly-average to the long-term average and adding the difference to each hourly temperature.

The hourly temperatures generated in this manner are actually Type I temperatures, in that they represent realistic weather data. Since the daily-average values were not predetermined, the daily-average values are not fixed so as to replicate the long-term distribution and autocorrelation. Rather the daily-average values are realistic daily values, creating a Type I daily sequence. This means that the distribution will not always exactly match the long-term distribution as predicted by Erbs correlation, and the daily autocorrelation may vary from the long-term average value, as a real month's autocorrelation value is observed to do. Since this was not the goal of this weather generator, both Degelman's deterministic hourly temperature model and the stochastic hourly temperature model were included in the weather generator, allowing the user to choose whichever model is most appropriate for the particular application. If the autocorrelation of hourly ambient temperature is thought to have an important effect, the stochastic model would be more appropriate, whereas if maintaining the long-term daily autocorrelation was thought to be more important, the deterministic model should be used.

4.3 Comparison

The monthly-average ambient temperature values for the generated data and the long-term data are identical due to the adjustment of the generated values. The TMY monthly-average values are also very similar, differing at the most by approximately one degree C. If the monthly-average values for a particular calendar month were assumed to be normally distributed independent random values, then the standard deviation of the long-term mean would vary from approximately 0.22 to 0.54 for the different months, indicating that for a 95% confidence interval, the actual long-term mean value could be within +/- one degree.

The standard deviation of the daily-average ambient temperatures as estimated by the generated and TMY data is generally within 2 degrees C of the long-term values, however, the TMY is often closer to the long-term. For example, for August in Albuquerque, the long-term standard deviation is 2.1, the TMY value is 1.8, and the generated value is 3.9. This two degree discrepancy causes the differences apparent between the long-term/TMY curves and the generated/correlation curves in Figure 4.4 The skewness and kurtosis vary slightly from the long-term for both the generated and TMY data, with neither the generated or TMY being significantly better. The average long-term monthly minimums are more closely reproduced by the TMY data; the largest deviations of the generated minimums from the average long-term values are 5 degrees C. Similarly, the maximums are better reproduced by the TMY data, however not as significantly as with the minimums.

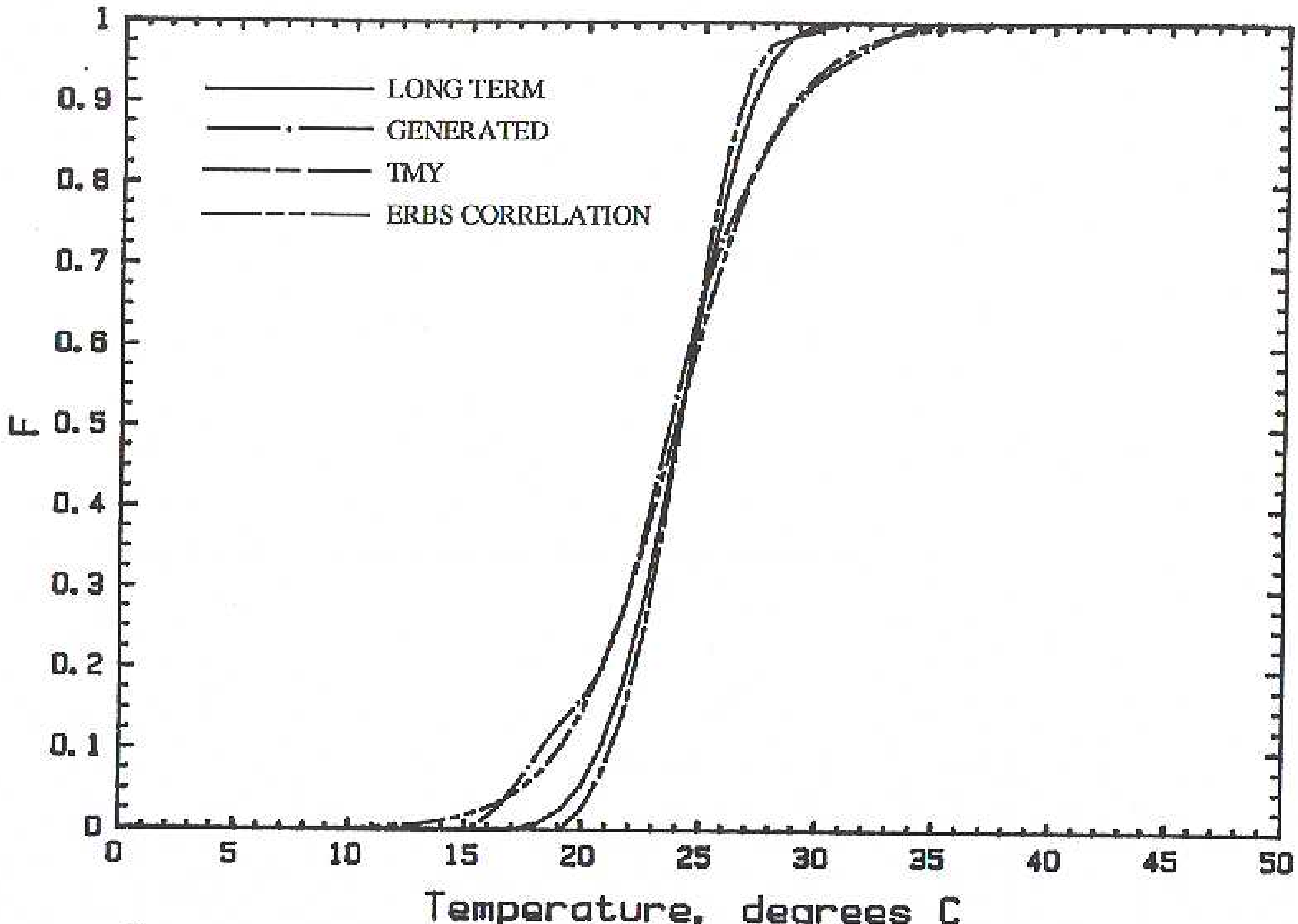

Figure 4.4 Comparison of Long-Term, Generated, and TMY Daily-Average Ambient Temperature Distributions for August in Albuquerque, NM

Examination of the RMS error associated with the monthly-average hourly diurnal variation of ambient temperature indicates that the TMY data often does a better job of simulating the long-term diurnal variation. The reason for this lies in the accuracy of the correlation. An example is the diurnal variation for November in Albuquerque (Figure 4.5). The long-term average curve is shifted in relation to the correlation; the generated curve lies on top of the correlation, indicating that this is the best that can be achieved without further research into developing a more detailed correlation. Some months, however, are quite accurately represented by the generated data, for example March in Madison (Figure 4.6). This month also indicates how the TMY is not always an accurate portrayal of the long-term.

The distributions of the daily-average ambient temperatures about their monthly-average value show one of the problems associated with having a "realistic" model. The May distribution in Madison (Figure 4.7) is an example of the generated data not replicating the correlation distribution. The correlation in this case approximated the long-term quite well, but due to the randomness in the stochastic temperature model, the generated data does not exactly recreate the correlation distribution. However, there are many months in which the distribution replication is very good, for example November in Madison (Figure 4.8).

The generated daily autocorrelations are slightly low; instead of yielding values of about 0.6 or 0.7, the generated values vary considerably and appear to have a mean of approximately 0.5. This is most likely due to an inadequacy in the AR2 model, such that the lag 24 hourly autocorrelations are not correctly reproduced. The lag one daily autocorrelation of the TMY data also exhibits a higher degree of variability than the

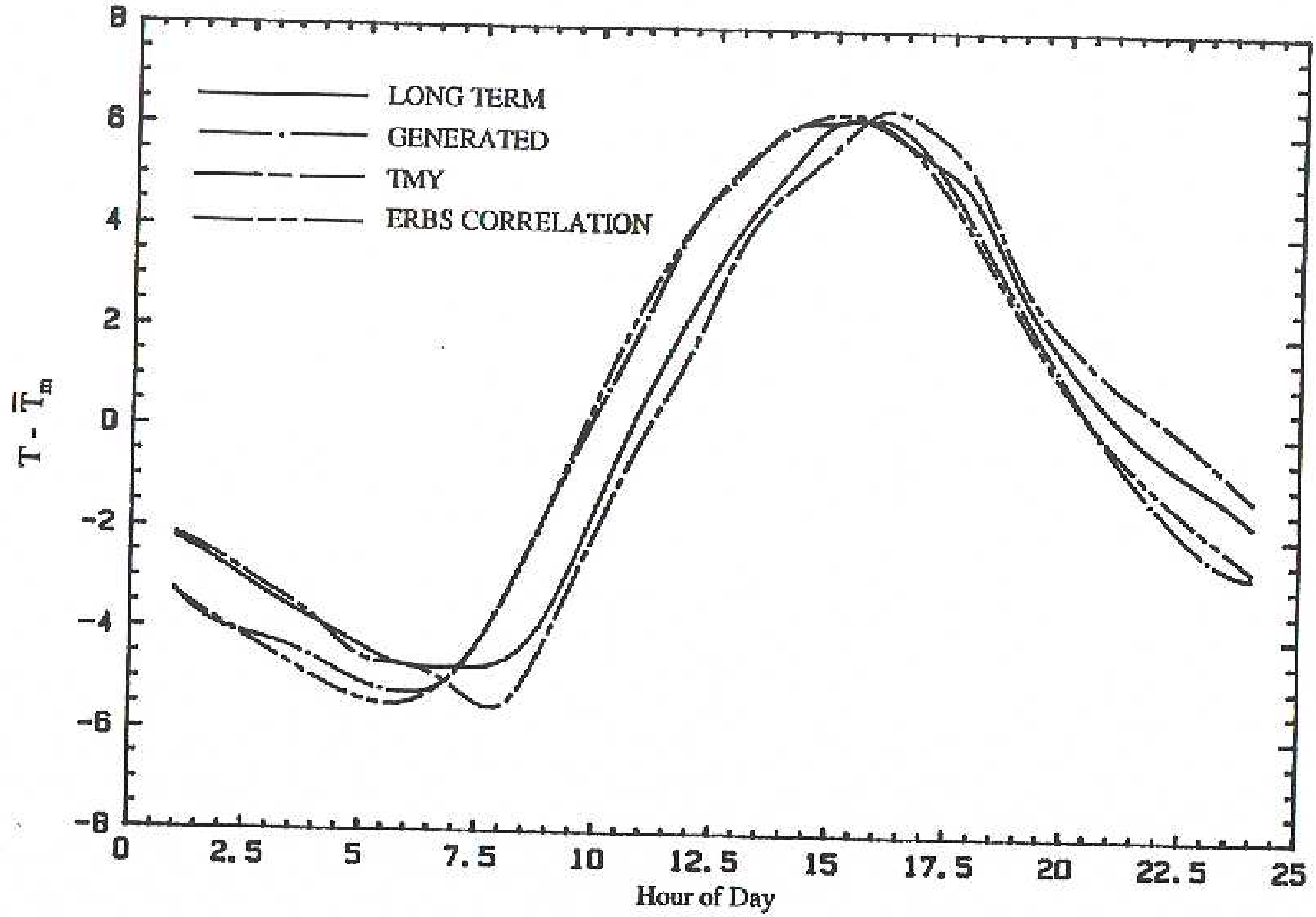

Figure 4.5 Comparison of Long-Term, Generated, and TMY Monthly-Average Ambient Temperature Diurnal Variations for November in Albuquerque, NM

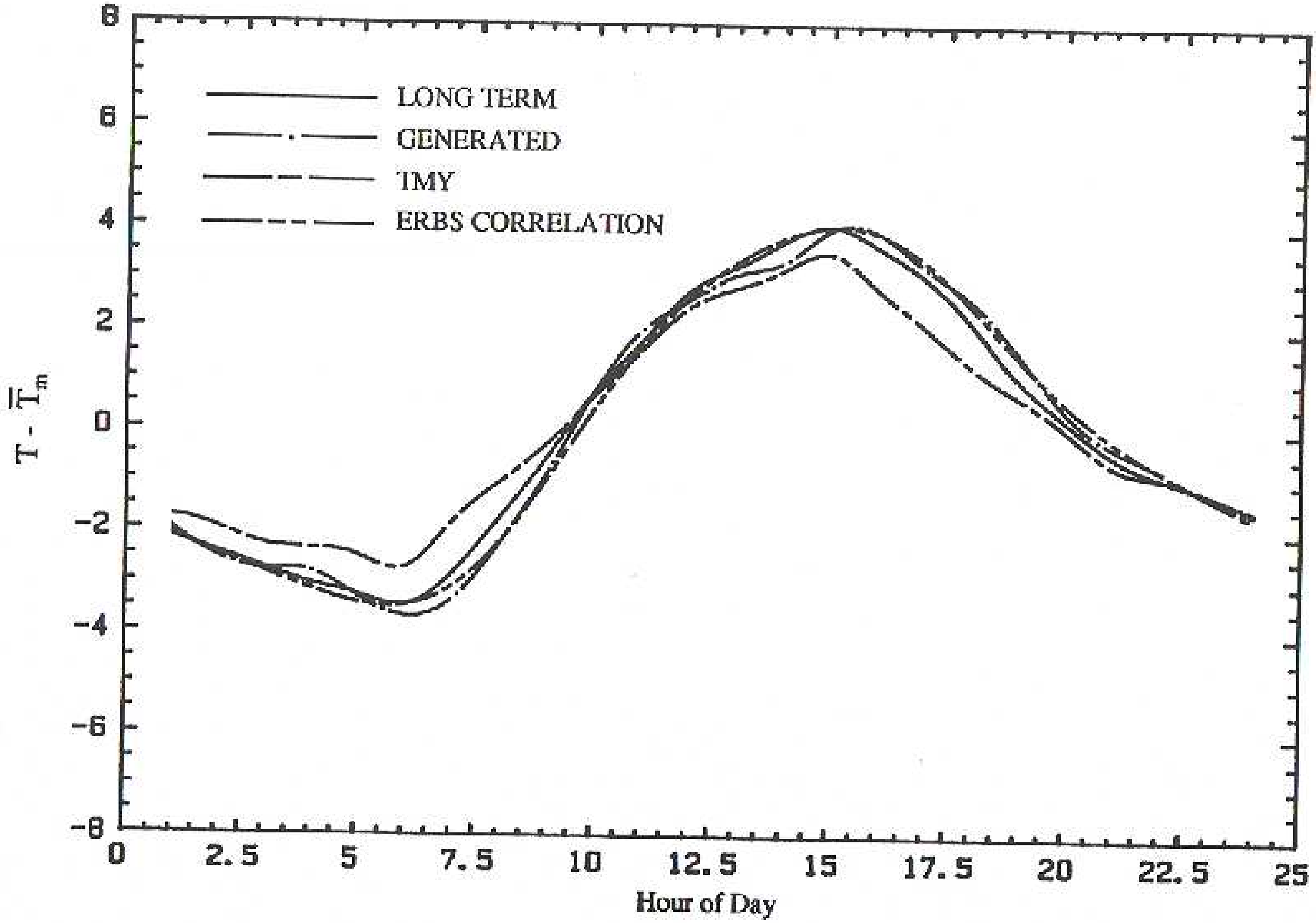

Figure 4.6 Comparison of Long-Term, Generated, and TMY Monthly-Average Ambient Temperature Diurnal Variations for March in Madison, WI

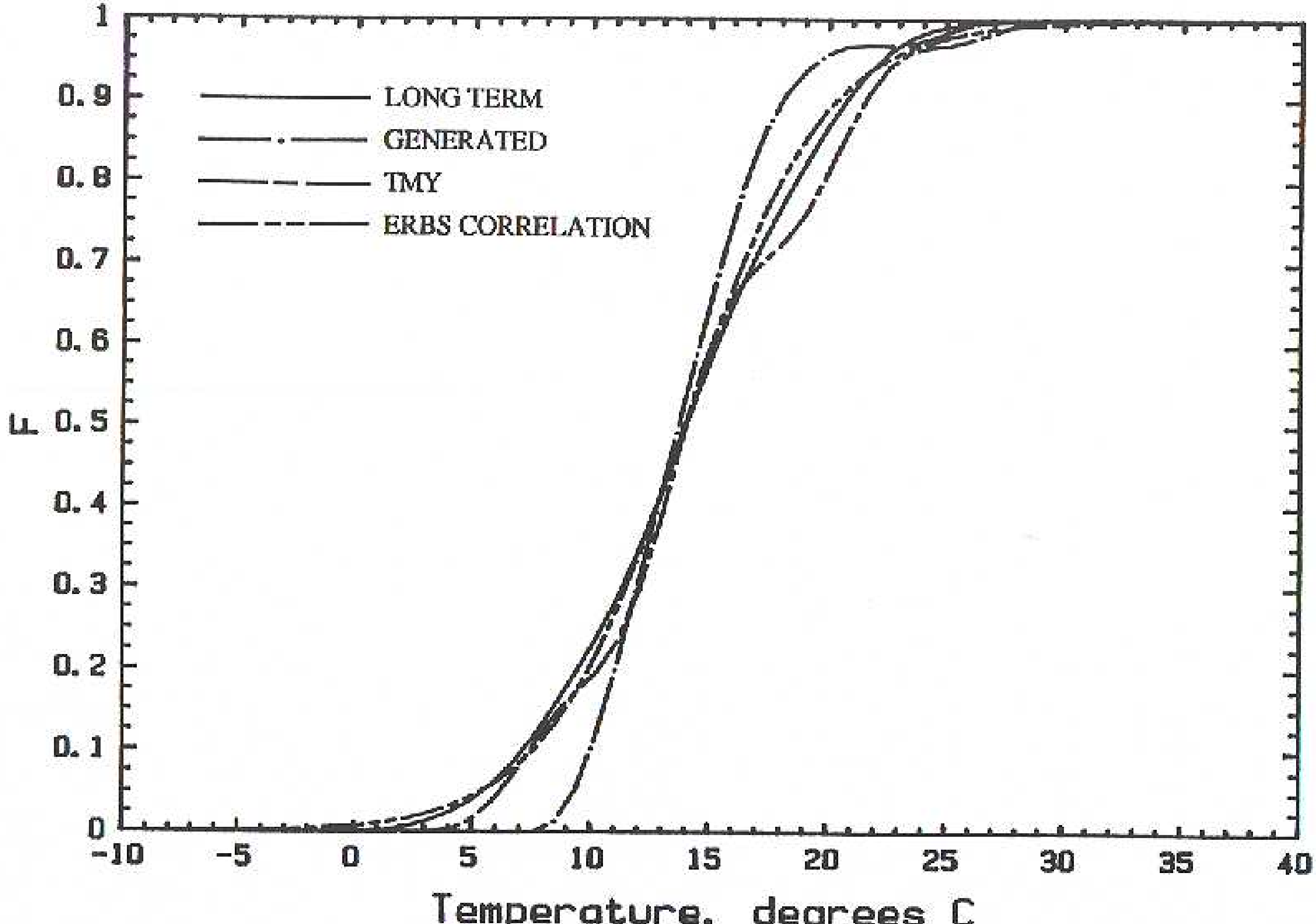

Figure 4.7 Comparison of Long-Term, Generated, and TMY Daily-Average Ambient Temperature Distributions for May in Madison, WI

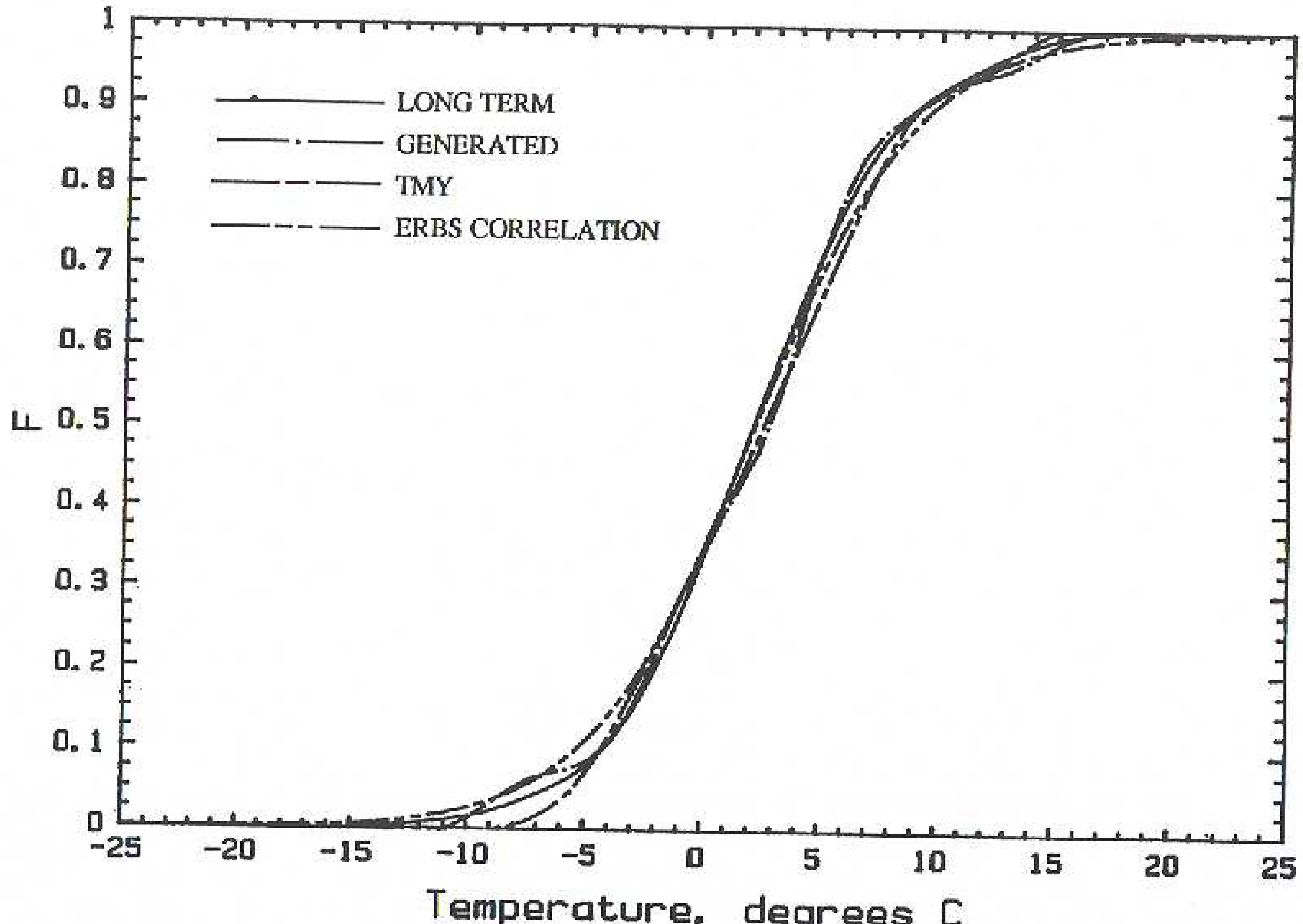

Figure 4.8 Comparison of Long-Term, Generated, and TMY Daily-Average Ambient Temperature Distributions for November in Madison, WI

long-term data.

The cross-correlation between daily-average ambient temperature and daily-average windspeed is negligible for Albuquerque and Madison; for the non-summer months in New York values ranged from -0.2 to -0.3, indicating a slight negative correlation. The standard deviation of these values was relatively large, on the order of 0.2 for all three locations. Both the TMY data and the generated data exhibited cross-correlations which varied considerably in sign and magnitude (-0.47 to +0.46), however, the magnitude of the correlations computed from the generated data tended to be smaller. This behavior would be expected because of the large standard deviations associated with the long-term values

The long-term average cross-correlation between daily-average ambient temperature and daily-average relative humidity varies considerably between the three locations. Madison and New York both exhibited strong positive correlations (0.6) for the winter months, changing to weaker negative correlations (-0.25) for the summer months. Albuquerque, on the other hand, shows strong negative values during the summer months, and weaker negative values during the winter months. The TMY data, while varying considerably, replicates the patterns evident at the three locations. The generated data produces strong negative cross-correlations for all months at all locations, mostly in the range -0.4 to -0.7.

Since the ambient temperature data generator is more of a Type I generator than a Type II generator, some of the statistical values computed are strongly a function of the random number seed. To determine the effect of the random number seed, ten years of weather data for Madison were generated using a different random number seed for each year. The first six months of the month-by-month results are listed in Table 4.5; the italicized values are the standard deviations. The monthly-average does not vary at all, as would be expected since the hourly values are scaled so as to match the monthly-average value. The average lag one autocorrelations for the different months are in the range of 0.48 to 0.60.

Table 4.5
Effect of the random number seed on the statistics associated with ambient temperature

	Mean	StDev	Skew	Kurt	Min	Max	ρ1	ρ2
Jan	-8.20	6.88	-0.23	0.40	-24.11	6.10	0.58	0.18
	0.00	*1.22*	*0.55*	*1.00*	*4.11*	*3.85*	*0.14*	*0.21*
Feb	-5.72	5.22	-0.06	-0.12	-16.96	5.17	0.51	0.17
	0.00	*0.83*	*0.36*	*0.66*	*2.65*	*2.95*	*0.09*	*0.13*
Mar	-0.24	5.7	-0.41	0.14	-13.43	10.09	0.54	0.11
	0.00	*0.56*	*0.59*	*1.47*	*3.30*	*2.12*	*0.07*	*0.17*
Apr	7.83	4.38	0.13	-0.09	-0.90	16.79	0.50	0.09
	0.00	*0.99*	*0.67*	*1.10*	*3.25*	*2.95*	*0.11*	*0.14*
May	13.96	4.26	0.21	0.34	5.14	24.12	0.52	0.17
	0.00	*0.61*	*0.41*	*0.97*	*1.89*	*2.18*	*0.12*	*0.27*
Jun	19.65	3.92	-0.05	-0.09	11.85	27.69	0.56	0.15
	0.00	*0.61*	*0.66*	*0.97*	*1.13*	*2.03*	*0.13*	*0.23*

CHAPTER 5

Relative Humidity and Windspeed Generation

This chapter covers the methods employed by Degelman to generate the hourly relative humidity values and hourly windspeed values. No modifications were attempted to either model, whether or not they were necessary, due to the time limitations of this project. A comparison of the generated statistics, distributions and diurnal variations with the long-term and TMY is included.

5.1 Relative Humidity Model Inputs

Rather than generating relative humidities directly, Degelman generates dewpoint temperatures which, along with knowledge of the dry bulb temperature and barometric pressure, are converted into relative humidities.

Several modifications to the required inputs dealing with humidity were made by Erbs. Originally, Degelman required inputs of monthly-average dewpoint temperature and the standard deviations of the daily-average dewpoint temperatures about their monthly-average values. Erbs changed and reduced the required inputs to include just the monthly-average humidity ratios. The monthly-average humidity ratios are then converted to monthly-average dewpoint temperatures using a relation developed by Erbs[1984].

The standard deviation of the daily-average dewpoint temperatures about their monthly-average value is assumed equal to the standard deviation of the daily maximum ambient temperatures about their monthly-average value.

The altitude of a location is an optional input, used to calculate the barometric pressure. If the altitude is not input, the altitude is assumed to be sea level.

5.2 Daily-Average Dewpoint Temperatures

Similar to the modeling of the daily clearness indices, daily-average ambient temperatures, and daily maximum ambient temperatures, the daily-average dewpoint temperatures are obtained from a distribution in a predetermined order. The maximum and minimum dewpoint temperature for each day is determined as a function of the average dewpoint and minimum and maximum dry bulb temperatures.

The distribution of the daily-average dewpoint temperatures about their monthly-average value is assumed to be normal. The mean is the monthly-average value and the standard deviation, as stated above, is set equal to the standard deviation of the daily maximum ambient temperatures about their monthly-average value.

5.3 Hourly Relative Humidities

Once the minimum, maximum, and average dewpoint and dry bulb temperature values are known for a day, the hourly relative humidity values are indirectly determined from hourly dewpoint temperatures. The dewpoint depressions (the dry bulb temperature minus the dewpoint temperature) at the maximum (3 pm) and minimum (sunrise) temperature hours are computed by assuming the maximum and minimum dry bulb and dewpoint temperatures coincide. The dewpoint depressions for each hour are found by linear interpolation between the depressions at the maximum and minimum hours. Dewpoint temperatures are simply computed by subtracting the dewpoint depression from the dry bulb temperature. A psychrometric routine written by Degelman is used to convert the hourly dry bulb and dewpoint temperatures (along with the barometric pressure) to hourly relative humidity values.

5.4 Comparison of the Relative Humidity Data

The data on the hourly weather tapes did not include relative humidity values, but instead contained hourly dewpoint temperatures. While it would have made some sense to output the dewpoint temperatures from the generator and compare them to the long-term dewpoint temperatures, all of the location-independent correlations developed by Erbs were for relative humidity. The TMY data consisted of hourly humidity ratios, which also had to be converted into relative humidities for comparison. The procedures used for both conversions are listed in Appendix H. Some of the recorded data, when converted into relative humidities, yielded values greater than 100%. In these cases, the relative humidity was assumed equal to 100%.

The generated monthly-average relative humidity values varied from the long-term values by as much as 16% relative humidity, however, most months were within 5%. The TMY data is generally more representative, with the exception of the summer months in Albuquerque for which the generated data provided a better approximation to the long-term value.

The TMY standard deviations are generally better than the standard deviations of the generated data at reproducing the long term standard deviations. This is expected, as the standard deviations obtained from the generated data are functions of the mean. Since the means are more accurately represented by the TMY data, it follows that the standard deviations would also be more accurate. Similarly, the skewness and kurtosis of the long-term data are generally better represented by the TMY data.

The average monthly maximum daily-average relative humidity values of the long-term data are slightly better emulated by the TMY data then the generated data for Madison and New York. In Albuquerque, however, the generated maximum values are more representative. Examination of the Albuquerque results indicates some discrepancies; the maximum TMY value is *greater* than the overall maximum long-term value for several months (October, November, and December). Since the TMY months are supposedly months taken from the long-term data, the TMY maximum should not exceed the long-term maximum.

The average monthly minimum daily-average relative humidity values of the long-term data are equally reproduced by the TMY and generated data in Albuquerque, within approximately 4 degrees C of the average long-term minimums. For Madison, the TMY values provide better approximations. More importantly, in June through December, the generated minimums are less than the overall long-term minimum values by a significant amount. In New York, the generated minimums, like the TMY minimums, are within roughly 5 degrees C of the average long-term values with the exception of November and December, in which the generated minimum values are again less than the overall long-term values.

The distributions of daily-average relative humidity for the long-term, generated and TMY data for Albuquerque, Madison, and New York are compared in Appendix E. They are sometimes difficult to compare due to significant differences in the mean values, which shift the distribution curves. Analysis of the bias error and RMS error indicates that Erbs correlation[1984] is better than both the TMY and the generated data at reproducing the long-term distribution. The main reason for this result, however, is probably due to the fact that the means of the correlation and the long-term data are virtually identical. For December in New York (Figure 5.1), the generated, long-term, and TMY means are approximately the same (the TMY is 2% low), so they can easily be compared. The generated distribution is quite different from the other three, which are all very similar; this is due to a larger standard deviation. The curve corresponding to the Erbs correlation, however, indicates the possible distribution which could be obtained from a better relative humidity generation model. October in Albuquerque (Figure 5.2) is another month in which all of the means are very similar, although the TMY mean is 4% higher. This is an example of a month in which the distribution is well reproduced by the generated data. Comparison of the Erbs correlation with the long-term distributions also indicates that the correlation provides a good approximation to the daily distribution in addition to the hourly distribution from which it was developed.

The long-term monthly-average diurnal variation of relative humidity is more accurately reproduced by the TMY data than the generated data. This is mainly due

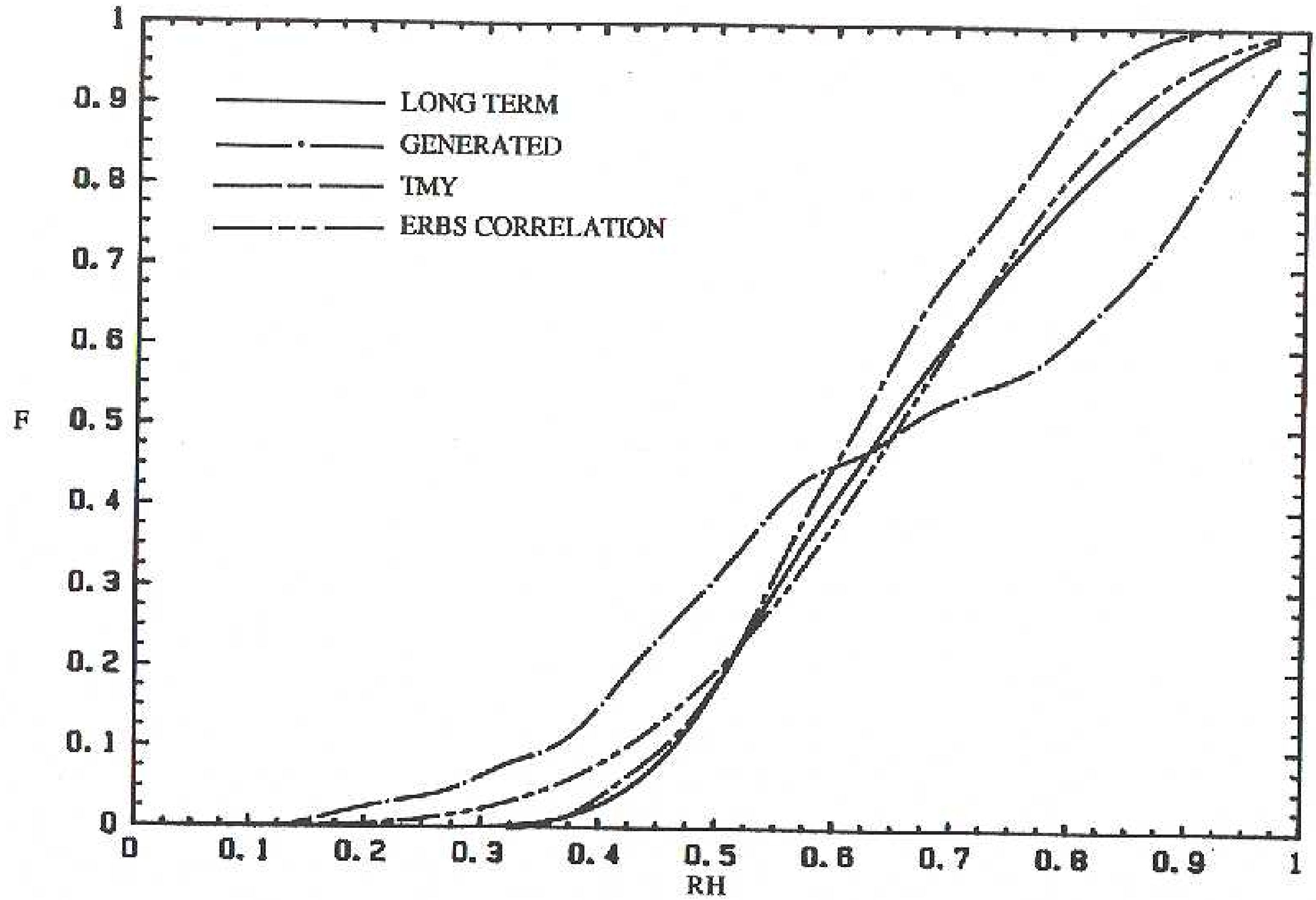

Figure 5.1 Comparison of Long-Term, Generated, and TMY Daily-Average Relative Humidity Distributions for December in New York, NY

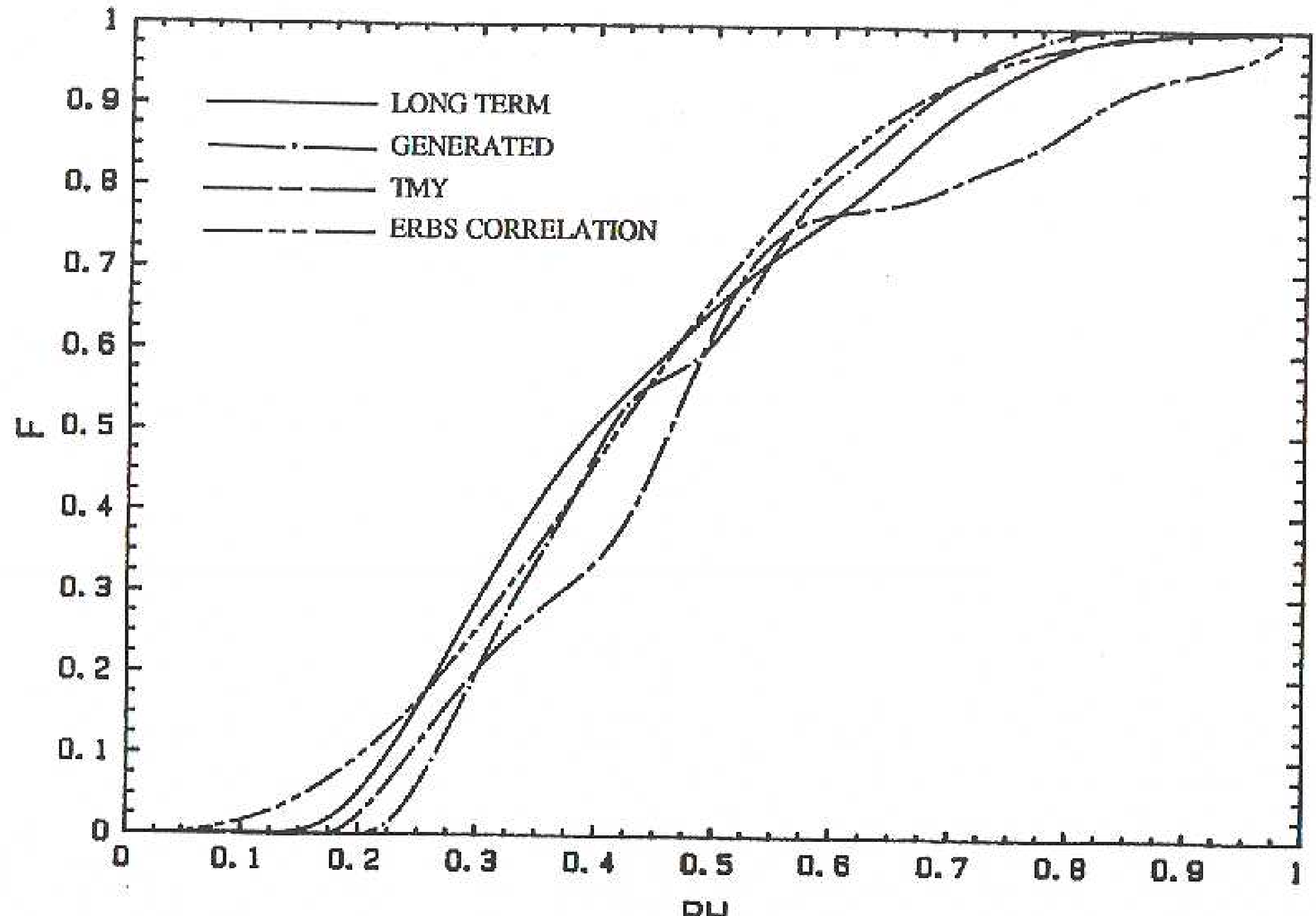

Figure 5.2 Comparison of Long-Term, Generated, and TMY Daily-Average Relative Humidity Distributions for October in Albuquerque, NM

to the straight line shape of the generated diurnal variations and differing amplitudes, for example, August in Madison (Figure 5.3). The TMY data is also generally better than the variation predicted by Erbs correlation. April in Albuquerque (Figure 5.4) is one of the months with the best generated diurnal variations; this happens to be a poor month for the TMY data, although it is interesting that Erbs correlation and the TMY data agree quite well.

The average long-term lag one autocorrelations of the daily-average relative humidity values show no systematic variation and are roughly constant (the range of values for each locations is about 0.17). The values for Madison and New York are approximately the same, with means of 0.45 and 0.40 respectively, however, the values for Albuquerque are significantly higher, with a mean of about 0.61. The mean of the generated autocorrelation values is close to that computed from the long-term values for Albuquerque, however the range is much greater, about 0.30. The TMY autocorrelation values are close to the long-term values, with the exception of several months for Madison and New York.

The average long-term cross-correlation between daily-average relative humidity and daily-average windspeed is essentially zero for Albuquerque and Madison, but in New York, for all of the months except those in late summer, cross-correlation values between -0.20 and -0.33 were computed. The standard deviation associated with each of the monthly values is relatively high, about 0.22, independent of location. The generated data produces no significant cross-correlations between daily-average relative humidity and daily-average windspeed; the cross-correlations computed from the TMY data vary widely, yet they are appropriately higher for the

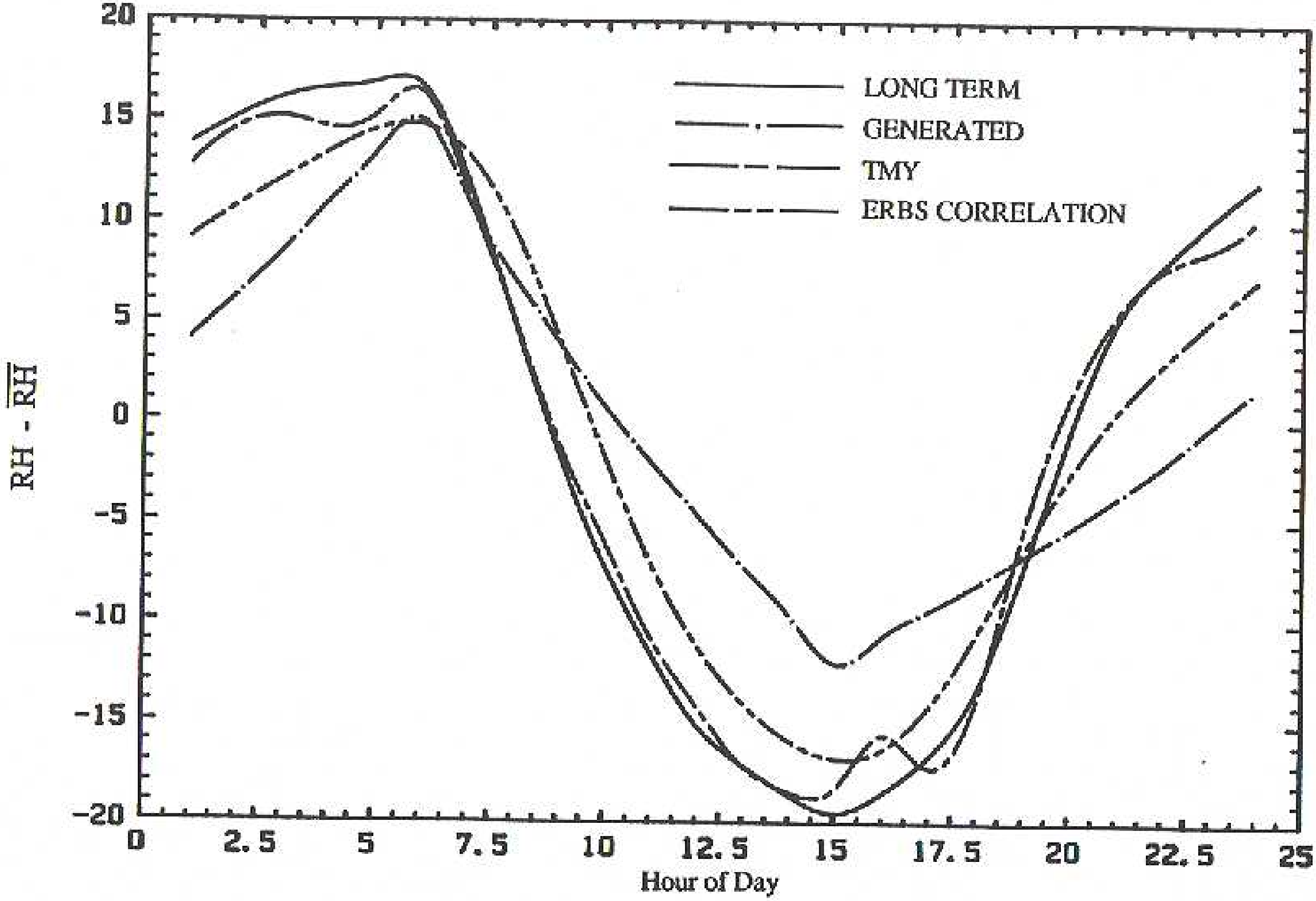

Figure 5.3 Comparison of Long-Term, Generated, and TMY Monthly-Average Relative Humidity Diurnal Variations for August in Madison, WI

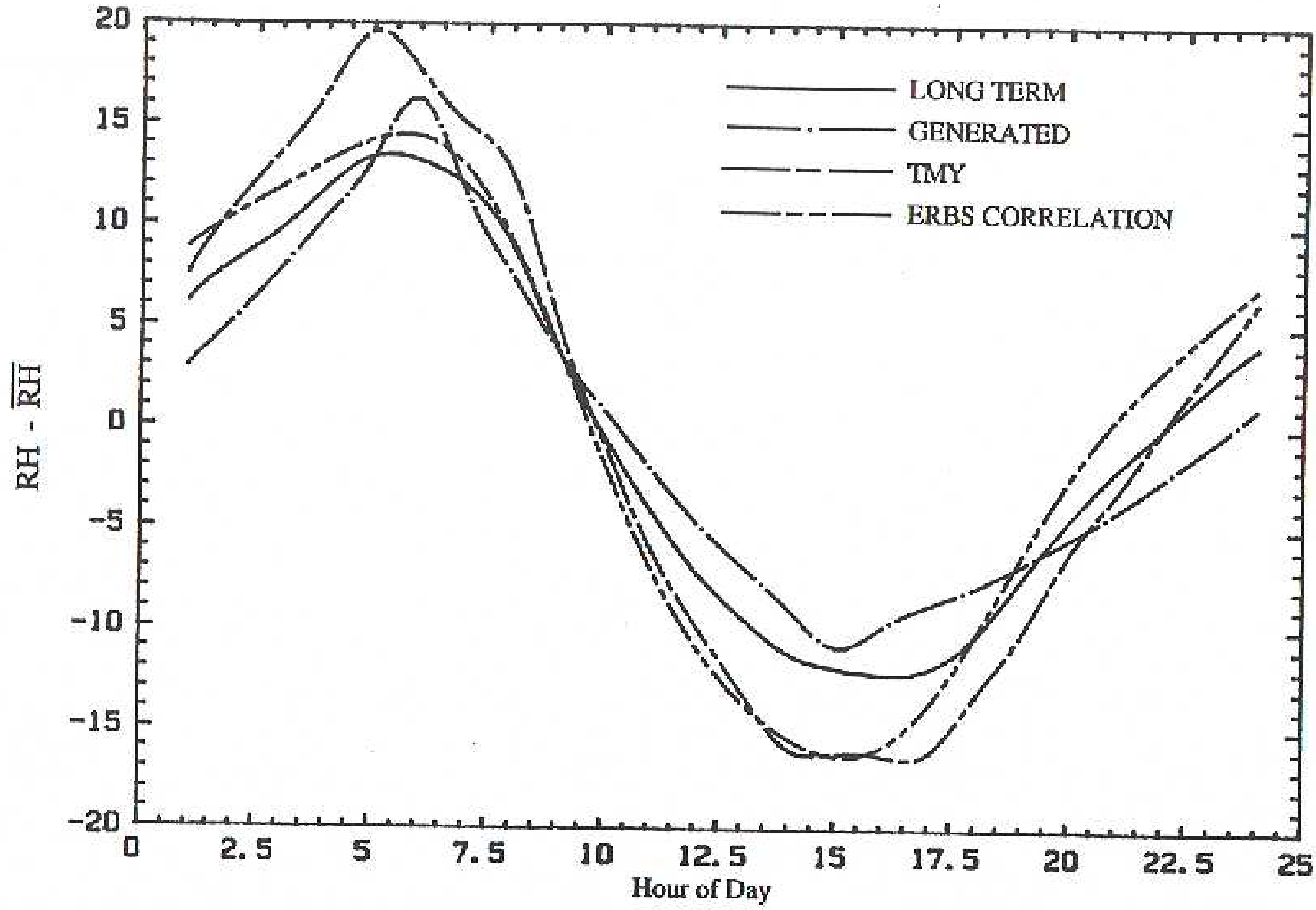

Figure 5.4 Comparison of Long-Term, Generated, and TMY Monthly-Average Relative Humidity Diurnal Variations for April in Albuquerque, NM

New York data.

5.5 Windspeed Simulation

Monthly-average windspeed is an optional input to the weather generator; if it is not available, an average value of 9 mph (4 m/s) is assumed.

Daily-average windspeed values are assumed to be normally distributed, with a standard deviation equal to 0.31 times the monthly-average value. A fixed 31-day sequence is used to order the values obtained from the distribution.

Hourly values are randomly selected from a normal distribution with a mean equal to the daily-average value and and a standard deviation equal to 0.35 times the daily-average value.

5.6 Windspeed Comparison

The distribution of hourly and daily-average windspeed has been observed to have either a Weibull or a χ^2 distribution [Corotis, 1977; Exell, 1985; Balouktsis, 1986] rather than a normal distribution as used by Degelman. There is a diurnal variation of hourly windspeeds and the hourly values are autocorrelated, both of which were neglected by Degelman.

The long-term windspeed data never exceeded 9 m/s; hourly values appeared to "cut off" at 9 m/s. Because of this, the long-term data characteristics presented are not necessarily accurate representations of the long-term.

No monthly-average windspeed values were input; the monthly-average value of 4 m/s was assumed. While this does not produce the most accurate generated data, this tests the generator with a minimum of inputs.

The monthly-average windspeed values, as expected, are all approximately 4 m/s. The yearly-average long-term values for Albuquerque, Madison, and New York are 3.6, 4.3, and 5.3 respectively. The range of the monthly-average values is approximately 1.5 for each location. The assumption of 4 m/s is reasonable, although it is better for Albuquerque and Madison than for New York. The TMY monthly-average values are more accurate than the generated averages, not surprising in consideration of the relatively small standard deviations (0.3 to 0.7) associated with the long-term means.

The average long-term lag one autocorrelation of daily-average windspeed is small and positive with no apparent month-to-month systematic variation. At each location, the monthly values are approximately constant at values of 0.17, 0.23, and 0.27 in Albuquerque, Madison, and New York, respectively. The standard deviation associated with the monthly values was approximately 0.17 at all three locations, which is very large in comparison with the autocorrelation values. The generated values are good replications of the long-term, however, they are most appropriate for Albuquerque, seeming to vary about a constant value of about 0.16. The TMY values vary considerably, ranging from -0.07 to +0.51.

The distributions of the daily-average windspeed values about their monthly-average values for the long-term, generated, and TMY data are compared in Appendix E. Similar to the distributions of clearness index, ambient temperature, and relative humidity, again the inclusion of the means complicates the comparison of the distribution shapes. An example of the best generated windspeed distributions is June in Albuquerque (Figure 5.5). An example of the difference between the long-term windspeed distribution and the normal distribution is December in Albuquerque (Figure 5.6).

The long-term monthly-average diurnal variations of windspeed generally show higher windspeeds during the day and lower speeds during the night. The amplitude of the diurnal variation is stronger during the summer months, becoming almost nonexistent during the winter. In Albuquerque, the diurnal variation peaks at about 5pm every month; the minimum occurs at approximately 7am (Figure 5.7). The diurnal variations computed from the New York data are similar, although the peaks tend to be less pronounced. In Madison, the diurnal variation has no well-defined peak; the windspeeds are roughly constant during the day, dropping quickly to a lower nighttime value (Figure 5.8). The generated data has no diurnal variation, but rather is a random sequence of windspeeds selected from a distribution. Therefore, the months in which the generated data most accurately reproduces the long-term diurnal variations are the winter months. The TMY data, while not producing very smooth curves, follows the pattern of the long-term data well.

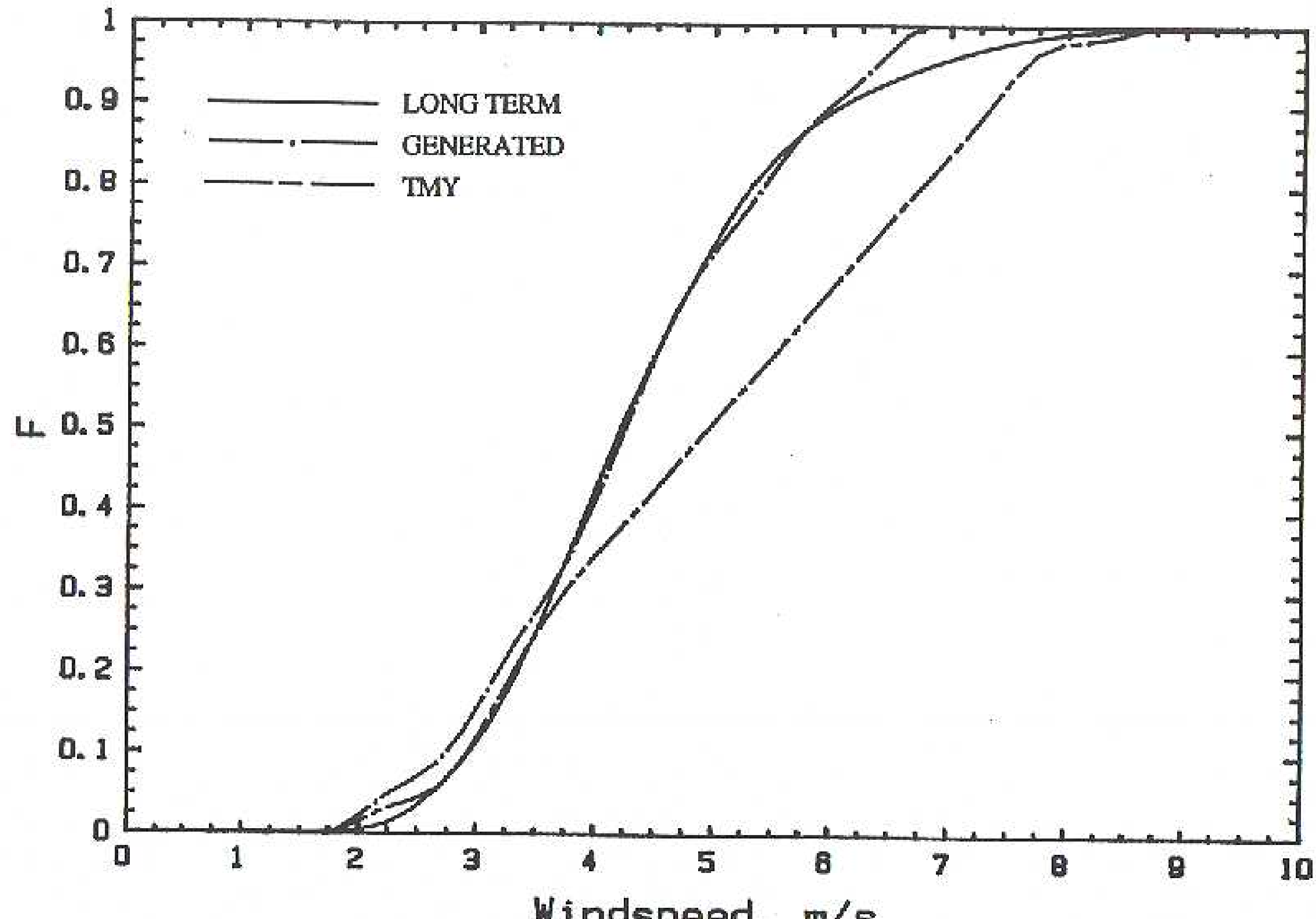

Figure 5.5 Comparison of Long-Term, Generated, and TMY Daily-Average Windspeed Distributions for June in Albuquerque, NM

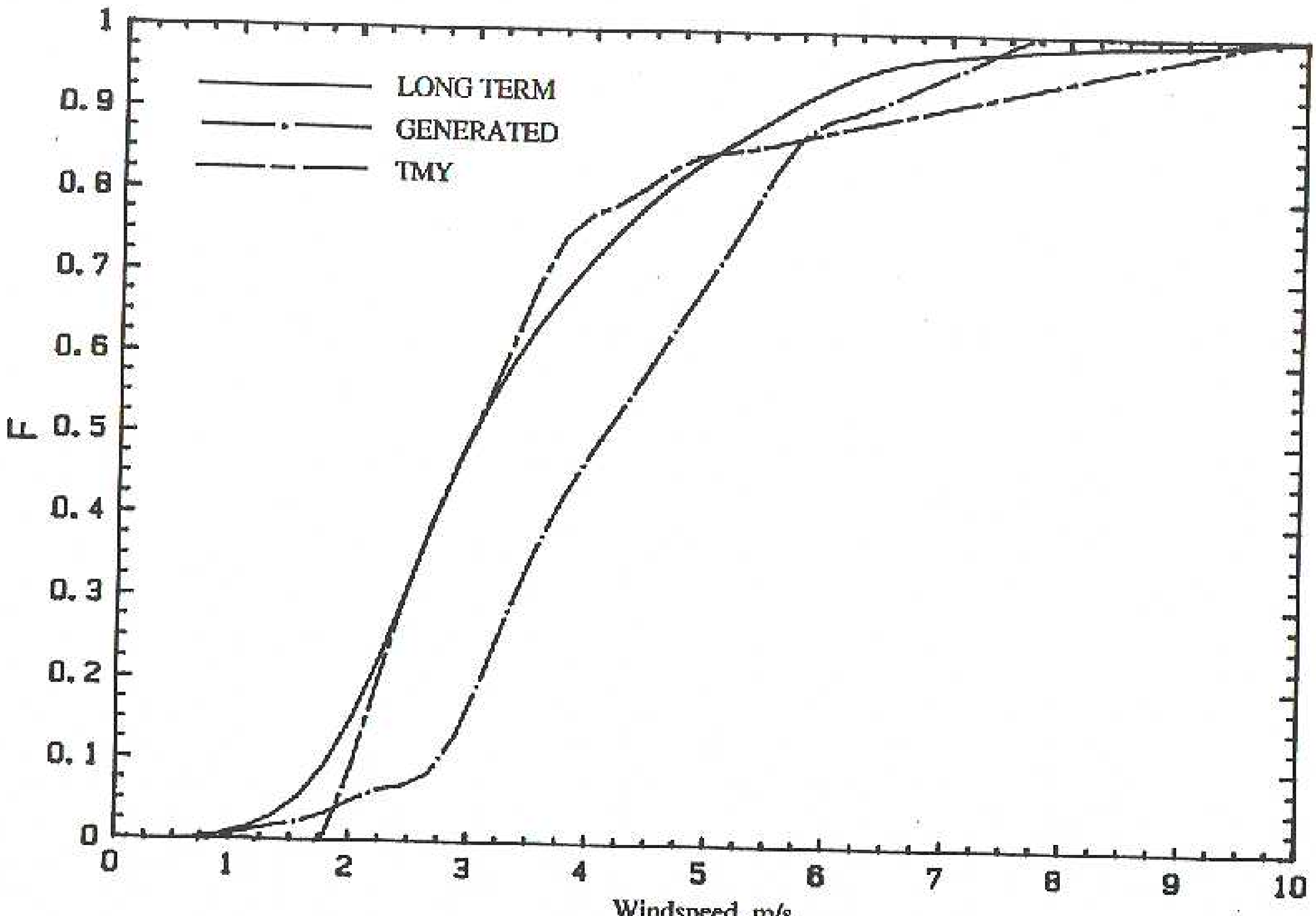

Figure 5.6 Comparison of Long-Term, Generated, and TMY Daily-Average Windspeed Distributions for December in Albuquerque, NM

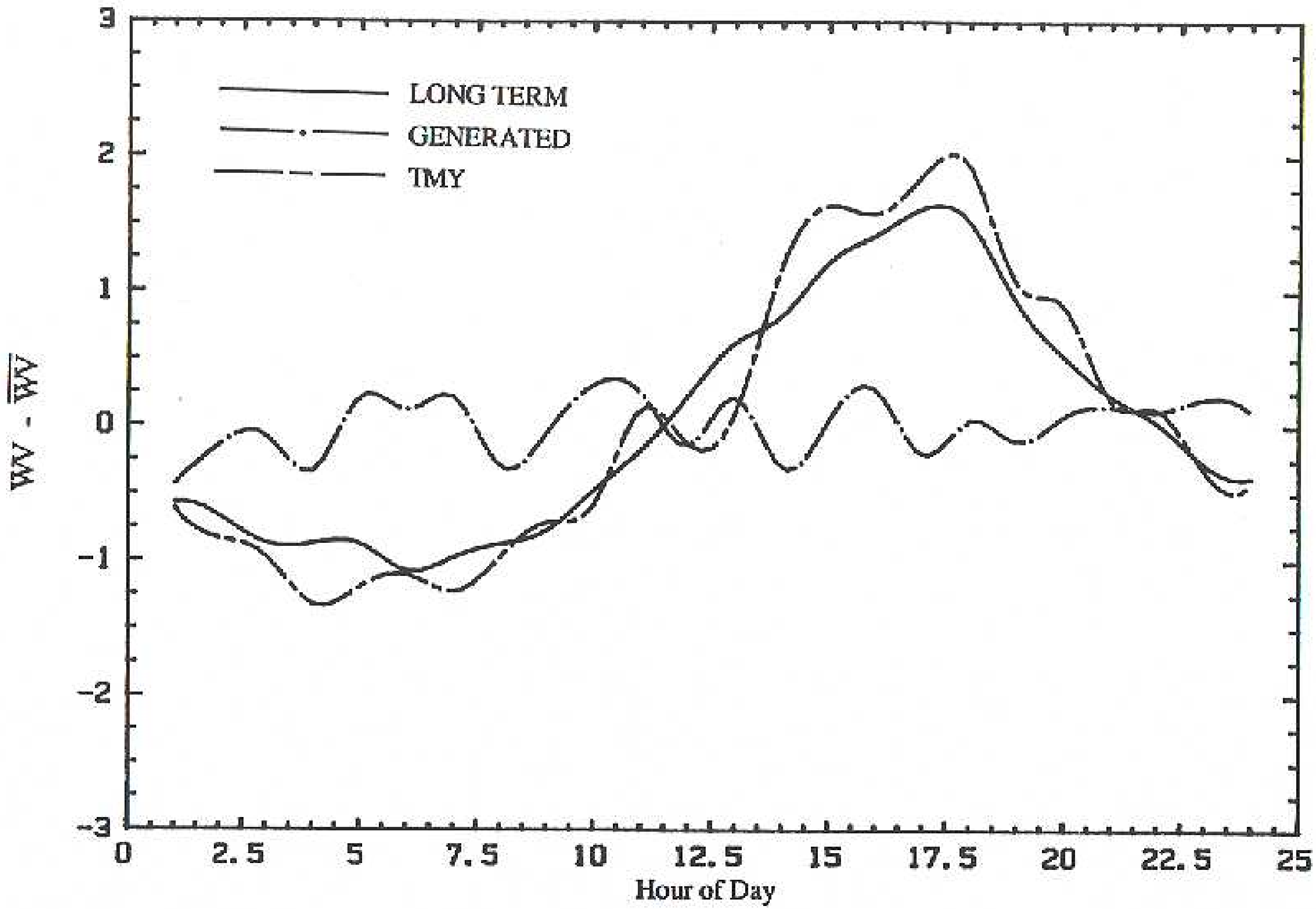

Figure 5.7 Comparison of Long-Term, Generated, and TMY Monthly-Average Windspeed Diurnal Variations for May in Albuquerque, NM

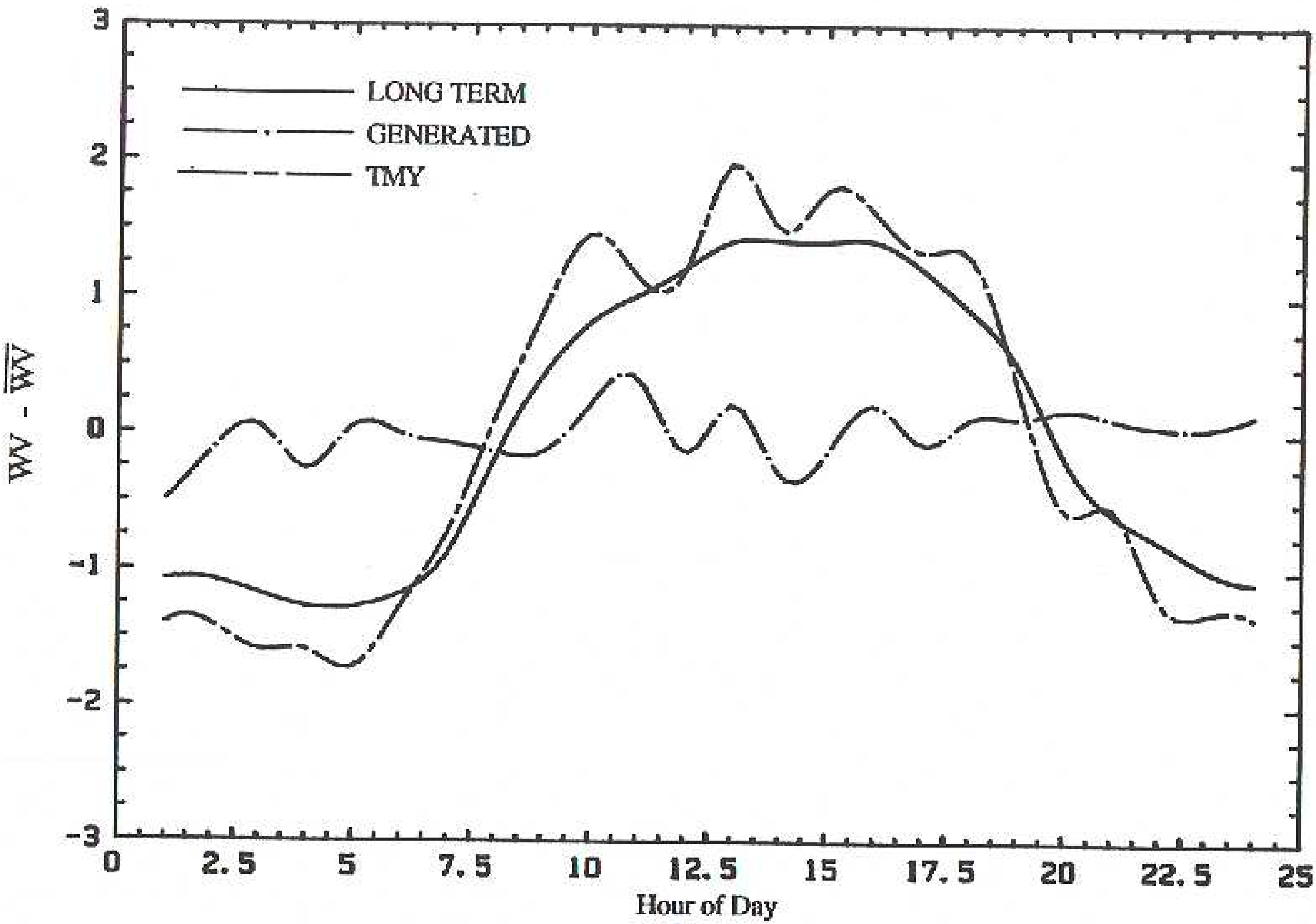

Figure 5.8 Comparison of Long-Term, Generated, and TMY Monthly-Average Windspeed Diurnal Variations for May in Madison, WI

CHAPTER 6

Conclusions and Recommendations

The goal of this project was to evaluate and modify an hourly weather generation program written by Degelman[1970] which generated not realistic weather data, but one "typical" year's data in which the statistics match those of the long-term data. The results presented in Chapters 3, 4, and 5 indicate that synthetically generating hourly weather data from limited monthly-average inputs is possible. The distributions and diurnal variations produced, while not always exact, are comparable to the TMY data in their ability to reproduce the long-term distributions and diurnal variations. Further refinement of some of the models is necessary, but long-term location-independent correlations already in existence illustrate the current potential of a weather generator. The development of correlations better able to describe the weather at a location will result in generated data more capable of reproducing the long-term statistics. The drawback of a weather generator is that it will only be able to reproduce those characteristics of a location's weather pattern which are able to be reproduced solely from location-independent correlations and the inputs. Cross-correlations and site specific tendencies, such as morning fog, are not able to be reproduced, as neither has yet to be classified in any location-independent manner.

The weather generation program written by Degelman showed a need for improvement. A stochastic model for replicating k_t values developed by

Graham[1985] was substituted into the hourly generation portion, yielding more realistic hourly radiation sequences. The hourly autocorrelation of the generated k_t sequence is slightly lower than the long-term, indicating a need to adjust the parameter of the AR1 model. The hourly ambient temperatures were generated deterministically in Degelman's original model; an AR2 model with location-independent parameters, similar to the model developed by Hittle and Pedersen[1981], was included as an option in the program. The hourly ambient temperature model was not completely replaced because the stochastic model generates a realistic sequence of hourly ambient temperatures, such that the long-term statistics are not necessarily well represented by one year's data.

This study has lead to many recommendations for further study. Most useful would be to study the effect of the different variables and variable interactions so as to determine what is important to reproduce accurately, and what can be roughly approximated. For example, do the differences in the monthly-average diurnal variation of ambient temperature shown in Figure 4.5 have an effect on simulation results? Is is necessary to reproduce the random variation in the hourly ambient temperatures or can a deterministic model be used?

Some manner of creating the sequences used to order the daily values obtained from the distributions should be found, such that the program can generate a sequence given an appropriate autocorrelation value. A more detailed investigation of the k_t distribution should be undertaken, possibly modeling instead the a_{kt} distribution. If possible, some attempt to model cross-correlations should be made.

The ambient temperature model should be investigated further and a better means of removing the diurnal trend found. This would yield a more stationary series, and most likely a better fitting model.. Improved models for relative humidity and windspeed should be developed; for relative humidity, the work of Erbs[1984] provides a basis. The k_t model should be investigated further to determine why the generated hourly autocorrelations are too low. A stochastic model of the diffuse radiation should be included, rather then the deterministic model used in the generator. A possible extension of the hourly generation process might include generating minute-by-minute data.