Solar Energy Vol. 41, No. 3, pp. 215–226, 1988
Printed in the U.S.A.

0038-092X/88 $3.00 + .00

TIME SERIES ANALYSIS OF DAILY HORIZONTAL SOLAR RADIATION

J. M. GORDON* and T. A. REDDY†
Applied Solar Calculations Unit, Blaustein Institute for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus 84993, Israel

Abstract—An analysis of the stationary and sequential properties of daily global horizontal solar radiation, on a discrete monthly basis, is presented for a number of locations of widely varying climatic conditions. Such information is essential as input to analytic models for the long-term performance of solar energy systems and for the generation of synthetic daily radiation sequences that can serve as input to numerical simulations that model solar systems. The new aspects of our study include (1) analysis of a solar radiation database that is much larger than those considered heretofore and includes tropical low-latitude, as well as temperate middle-latitude, climates, (2) documentation of the magnitudes and correlations of generalized stationary and sequential radiation statistics for a wide range of climatic stations, (3) proposal of a simple functional form for the probability density function for daily radiation and comparison with actual data, (4) explicit consideration of confidence limits in predicting stationary radiation statistics from a limited number of years of data, (5) evidence that, contrary to the claims of most related studies, there do not seem to be universal values or universal correlations for either the persistence strengths or the persistence times of daily radiation, and, (6) discussion of the *practical* value of statistical studies of this nature for the design of solar energy systems.

1. INTRODUCTION

It has recently been recognized that the type of statistical information on solar radiation that has generally been published in the professional literature is insufficient for sizing and optimizing solar energy systems of very high solar fractions[1–5]. For example, the kinds of empirical and even analytic design methods that are appropriate for solar energy systems of intermediate solar fractions do *not* offer adequate accuracy in the optimal sizing of stand-alone photovoltaic systems.

Accordingly, solar radiation data should be "time-series" analyzed. Time-series analysis refers to a well-developed mathematical discipline that has been applied extensively in areas such as econometrics, hydrology and meteorology[6–8] and essentially involves analyzing data from the past in such a fashion that statistically meaningful projections of these data into the future can be made. Only recently has time-series analysis been applied to terrestrial solar radiation, since until recently extensive records of solar radiation have been scarce.

Previous studies involving time-series analysis of solar radiation data have typically considered either one location only or, at best, a few locations of similar climatic conditions[9–23]. The present study considers a large number of significantly different locations and climatic conditions, analyzes their solar radiation statistics, and checks for universal characteristics.

Our study of daily global horizontal solar radiation involves the following:

1. Explanation of our database and of the choice of the appropriate statistical variable.
2. Calculation of the monthly mean stationary statistical parameters (e.g., mean, variance, and skewness and consideration of correlations among them.
3. Proposal of a simple functional form for the stationary probability density for daily solar radiation, which requires a knowledge of the mean and variance only, and comparison with actual data from widely differing climatic stations.
4. Examination of confidence limits, namely, how one can ascertain the uncertainty range of the mean and the variance, given a limited number of yearly samples for a given month at a particular location.
5. Analysis of sequential behavior, which involves computation of stochastic components and determination of persistence times and persistence strengths.
6. Testing several types of auto-regressive stochastic models toward the generation of synthetic sequences which capture all the essential stationary and sequential statistical information for a broad range of climates.

Some of our results turn out to be "negative" in the sense that universal correlations are *not* found to exist. This in itself is an important finding, particularly in view of claims to the contrary, for example[17,19–21,24–25]. In addition, we believe that the present analyses are valuable in offering a procedure for characterizing each specific climatic station in a manner that subsequently can save considerable time and expense in the optimal sizing of solar energy systems.

*And Department of Mechanical Engineering, Ben-Gurion University of the Negev, Beersheva, Israel.
†Present address: Center for Energy & Environmental Studies, Princeton University, Princeton, NJ 08544, U.S.A.

The issue of time *scale* should be addressed at the outset. Although previous time-series studies for solar radiation have been performed on a yearly[17,18,20] or seasonal[9,11,12,14] basis, we choose to segment the year into *monthly* periods for practical reasons[15,16,19,21]. First, most solar system design calculations are performed on a monthly basis. Second, we prefer a time scale that is sufficiently short that, independent of location, annual trends in solar radiation can be approximated as constant (say, at their mean monthly values). In this way, a further filtering out of deterministic trends is not necessary. Such a procedure has also been used for the analysis of other meteorological variables such as cloud cover[26].

For a given location, however, a monthly time scale can result in an unnecessarily small statistical sample. For example, it may turn out for a particular site that deterministic trends exist for periods considerably longer than one month. In such a case, the period of analysis *could* contain a larger data set, with the associated improved statistical estimates.

2. DATABASE

Our database, which includes tropical low-latitude and temperate midlatitude locations, is summarized in Table 1. From the raw data, we eliminated: (1) all daily values with basic inconsistencies, as detailed in[27], (2) all months with more than two days of missing or spurious data, and (3) the entire month if less than three years of data were available after screening. For the Australian[39], Thai[40], and Israeli[41] locations, all 12 months of the year "survived" the filtering process. For the Indian locations, however, a total of 41 months only remained after screening (i.e., on average 6 months only, per site).

Although all the locations listed in Table 1 were included in our computations of *stationary* statistical properties, Bangkok, Chiang Mai, and Bet Dagan were *not* included in our analysis of *sequential* characteristics because the probability distributions *only* were available rather than actual raw data sequences.

Table 1. Locations considered in this study

Location	Country	Latitude	Period	Type of data
Darwin	Australia	12.4°S	1968–78	Rehabilitated
Hobart	Australia	42.9°S	1968–78	Rehabilitated
Melbourne	Australia	37.8°S	1968–78	Rehabilitated
Bhavnagar	India	21.7°N	1971–78	Raw
Bombay	India	18.9°N	1972–78	Raw
Jodhpur	India	26.3°N	1971–78	Raw
Madras	India	13.1°N	1971–78	Raw
Nagpur	India	21.1°N	1971–78	Raw
Poona	India	18.5°N	1971–78	Raw
Shillong	India	25.6°N	1971–78	Raw
Bangkok	Thailand	13.7°N	1968–72	p.d. only
Chiang Mai	Thailand	18.8°N	1968–72	p.d. only
Bet Dagan	Israel	32.0°N	1962–77	p.d. only

The probability distributions (p.d.) only were available for the last three locations listed.

3. SELECTION OF APPROPRIATE STATISTICAL VARIABLE

Because our analysis is performed on a *monthly* basis, we need not be concerned with annual or seasonal trends. We do, however, need to "extract" from the daily radiation data both the monthly *average* value and the effect of latitude. In prior studies[11,18,19,21], it has been suggested that extracting the influence of latitude be achieved, at least in part, by division of solar radiation values by the corresponding extraterrestrial values, to yield a daily "clearness index," K.

However, analysis of K itself is insufficient for our analyses because even monthly average clearness index, $\bar{K}$, can vary markedly from year to year. Hence, we wish to consider a variable via which we can retain year-to-year fluctuations in $\bar{K}$.

As an analogous problem, consider the tossing of a coin by 10 different people, each of whom tosses the coin 30 times. Assume that each person executes identical flips, but that there are differences in flipping technique among the 10 people. We are interested in the general statistics of "heads" and "tails" occurring. One analysis would be to simply consider the outcomes of all 300 events, without regard to who does the flipping. That would be analogous to considering K values for all years for which we have monthly data, without distinguishing among different years (e.g., the analysis of [22]).

A second analysis would be to first calculate the heads–tails statistics for *each* of the 10 people, and then average over those statistics. This would be analogous to first analyzing the statistics of monthly K values for *each* year and then averaging those statistics over all years, this procedure having been adopted in this study.

Consequently, we select as our variable, $X = K/\bar{K}$, where $\bar{K}$ is the monthly average K for a given month in a specific year. The stationary statistics of X are first calculated for each individual year, and *then* averaged over all years. There is no "best" variable for this analysis, to the best of our knowledge, and even selection of the variable X introduces small correlation errors. Specifically, because $\bar{K}$ depends on all K values for a given month, dividing K by $\bar{K}$ introduces an error of roughly $1/J$, where J is the number of days in a month.

4. GENERALIZED STATIONARY STATISTICS

We now define the stationary statistical variables that we have computed on a monthly basis for each of the locations noted in Section 2.

1. *Monthly mean clearness index*

For a specific year: $\bar{K}(r,t) = (1/J)\ \Sigma_{j=1}^{J}\ K(j,r,t)$

Grand average: $\langle \bar{K}(r) \rangle = (1/T)\ \Sigma_{t=1}^{T}\ \bar{K}(r,t)$

where (j,r,t) denotes (day, month, year), respectively, with day $j = 1, 2, \ldots J$; month $r = 1, 2, \ldots 12$; and year $t = 1, 2, \ldots T$.

2. *Monthly variance of the variable X*
where

$$X(j,r,t) = K(j,r,t)/\bar{K}(r,t)$$

For a specific year:

$$\sigma^2(X(r,t)) = (1/(J-1))\,\Sigma_{j=1}^{J}\,(X(j,r,t)-1)^2$$

3. Grand average:

$$\langle\sigma^2(X(r))\rangle = (1/T)\,\Sigma_{t=1}^{T}\,\sigma^2(X(r,t))$$

Monthly skewness of X
For a specific year:

$$S(X(r,t)) = (1/J)\,\Sigma_{j=1}^{J}\,((X(j,r,t)-1)/(X(r,t)))^3$$

Grand average: $\langle S(X(r))\rangle = (1/T)\,\Sigma_{t=1}^{T}\,S(X(r,t))$

A fundamental question is whether there are inherent and possibly universal correlations among the key statistical variables. For example, is it possible with a knowledge of $\langle\bar{K}\rangle$ to predict accurately corresponding values for the variance and skewness? Figure 1 presents a plot of the grand average of the variance, $\langle\sigma^2(X)\rangle$, vs. the grand average of the monthly average clearness index, $\langle\bar{K}\rangle$. A statistical test confirms the hypothesis that, at a significance level of .05, the spread in the data points is *not* due to chance variations.

There appears to be a correlation between these two variables, albeit with a rather large spread, which may be due primarily to the differences between temperate and tropical climates. Namely, at a given value of $\langle\bar{K}\rangle$, tropical climates seem to consistently exhibit a *lower* $\langle\sigma^2(X)\rangle$ than temperate climates.

This difference is probably a result of the higher atmospheric water vapor content in tropical climates. Whereas temperate climates can achieve an intermediate $\langle\bar{K}\rangle$ value via a combination of days with widely varying K values, tropical climates can achieve the same $\langle\bar{K}\rangle$ value with more uniform clear days whose intermediate K values are due to a consistently high level of precipitable water. A related observation was made by [28] in noting that the average

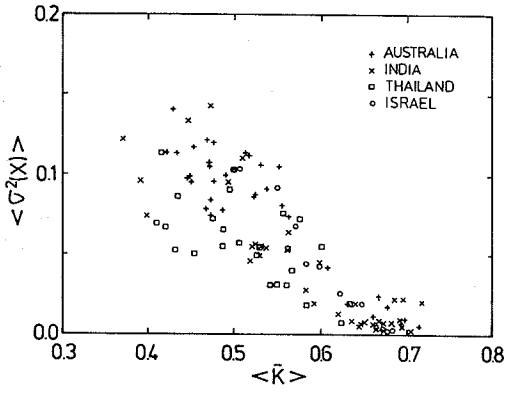

Fig. 1. Grand average of the variance, $\langle\sigma^2(X)\rangle$, vs. grand average of the monthly average clearness index, $\langle\bar{K}\rangle$.

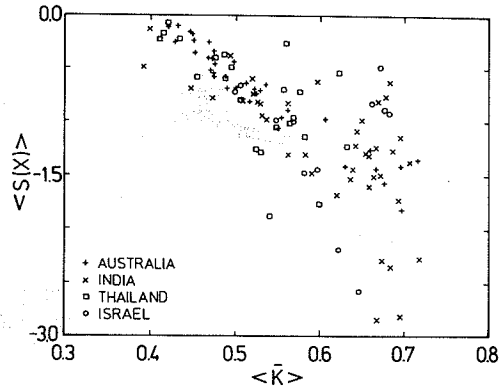

Fig. 2. Grand average of skewness, $\langle S(X)\rangle$, vs. $\langle\bar{K}\rangle$.

maximum values of K are lower for tropical, than for temperate, climates.

There is a limited predictive value to Fig. 1 in that if only $\langle\bar{K}\rangle$ is known for a given location, then the variance can be estimated from Fig. 1 if one can additionally classify the atmospheric conditions for that particular month. That is, the seemingly large spread in values of the variance becomes smaller (but far from negligible) provided one is able to categorize the climate of the location as temperate or tropical.

Figure 2 is a plot of skewness, $\langle S(X)\rangle$, vs. $\langle\bar{K}\rangle$. Although a rough correlation appears to exist here, there is not the clear distinction between temperate and tropical climates as in the case of variance. The relatively large scatter in Fig. 2 is due, in part, to the fact that statistical uncertainty increases with the order of the moment of X.

5. STATIONARY PROBABILITY DISTRIBUTION

The Probability Density Function (PDF), $P(X)$, for $X = K/\bar{K}$, summarizes the *stationary* statistics of importance. Several studies have considered explicitly prediction of the PDF of daily global horizontal terrestrial solar radiation[12,13,24,25,28–32]. However, to the best of our knowledge, none of these studies has considered a wide variety of climates and locations.

For purposes of analytic modeling, it is desirable to have a closed-form expression for $P(X)$ that has been compared favorably with actual data. This expression should preferably have no adjustable or arbitrary parameters. In addition, this expression should conform to observed universal characteristics of $P(X)$, for example: (1) $P(X)$ vanishes at sufficiently low X and sufficiently high X and (2) the variance of $P(X)$ decreases with increasing $\bar{K}$ (at least for the range $0.3 \leq \bar{K} \leq 0.75$).

Several closed-form functional forms have been proposed for $P(X)$[24,25,28,32]. Their major shortcoming is that they either assume the monthly maximum K value, K_{max}, to be a universal constant, or assume a knowledge of K_{max} as an input parameter. However, K_{max} is not a universal constant[28]. Furthermore, $P(X)$ is sensitive to the somewhat arbitrary

choice of K_{max}. For example, is K_{max} the highest value that occurs 1% of the time, or 0.1% of the time? A second minor point is that use of these proposed expressions requires solution of at least one transcendental equation.

We propose the following empirical functional form, which is plotted in Fig. 3:

$$P(X) = AX^n[1 - (X/X_{max})] \quad (1)$$

where the three parameters, A, n, and X_{max}, are determined from (1) the normalization of $P(X)$; (2) knowledge of $\bar{K}$ (i.e., $\bar{X} = 1$); and (3) knowledge of the variance of X, $\sigma^2(X)$:

$$n = -2.5 + 0.5\,[9 + (8/\sigma^2(X))]^{1/2} \quad (2)$$

$$X_{max} = (n + 3)/(n + 1) \quad (3)$$

$$A = (n + 1)(n + 2)/X_{max}^{n+1}. \quad (4)$$

More detailed discussion of the functional form of eqn (1) is presented in Appendix A. This includes exclusion of the skewness, $S(X)$, as an input parameter.

Equations (1) to (4) represent a functional form that can be evaluated without resorting to numerical solution of transcendental equations. In addition, eqn (1) can be integrated in closed form for system design problems where integrals of $P(X)$, such as the utilizability function, are required[34].

In Fig. 4 we present sample comparisons of the $P(X)$ of eqn (1) with actual data. The only two input parameters are $\langle\bar{K}\rangle$ and $\langle\sigma^2(X)\rangle$, which are nonadjustable. To characterize the "goodness of fit" quantitatively, we have performed the χ^2-test both for the $P(X)$ of eqn (1) and for the corresponding cumulative frequency distribution, $F(X)$. The latter is important both in the generation of synthetic sequences (see Section 8 below) and in solar system design methods. In our calculations of data-based $P(X)$, we have used 20 equistatistical bins (i.e., bins whose intervals are determined by an equal number of occurrences) for data-based X (with X varying in most cases between zero and two). In our χ^2-test calculations, the degrees of freedom are equal to the number of bins minus three, because for $P(X)$ we assume normalization and a knowledge of $\bar{X}$ and $\sigma^2(X)$.

The accuracy of eqn (1) is also illustrated by the fraction of the *total* data set (see Section 2) that corresponds to various significance levels, $\alpha(.05)$ and $\alpha(.01)$:

	$\alpha(.05)$	$\alpha(.01)$
$P(X)$	50%	60%
$F(X)$	80%	83%

In principle, the theoretical $P(X)$ is intended to pertain to daily K values for a given month in a *specific* year. However, we believe it unreasonable to assume that the average designer will have access to $\bar{K}$ and $\sigma^2(X)$ for every year of monthly data, but rather will at best know the grand averages $\langle\bar{K}\rangle$ and $\langle\sigma^2(X)\rangle$. Consequently, our "tests" of theory vs. data assume a knowledge of grand averages only as input to eqns (1) to (4). Generally, even $\langle\sigma^2(X)\rangle$ is not included in data available to designers. It is suggested that $\sigma^2(X)$ values (in addition to the presently recommended practice of specifying $\bar{K}$ only) be retained in the processing of solar radiation data for future application of the techniques discussed herein and for the generation of design year data sets.

Finally, we can now clarify what appears to be a misleading conclusion of several studies of daily solar radiation statistics: namely, that climates with the same $\langle\bar{K}\rangle$ exhibit essentially the same PDF[15,19,20,24,31,35]. As inspection of our data-based statistics above reveals, this conclusion is *not* universally valid. Because $\langle\bar{K}\rangle$ and $\langle\sigma^2(X)\rangle$ are strongly correlated in one way for temperate climates, and in a different way for tropical climates, an analysis of a small set of temperate climates only could lead to the conclusions of refs. [15,19,20,24,31,35]. One advantage of our approach is the ability to predict accurately $P(X)$ for *both* temperate and tropical climates *without* the need to distinguish between them.

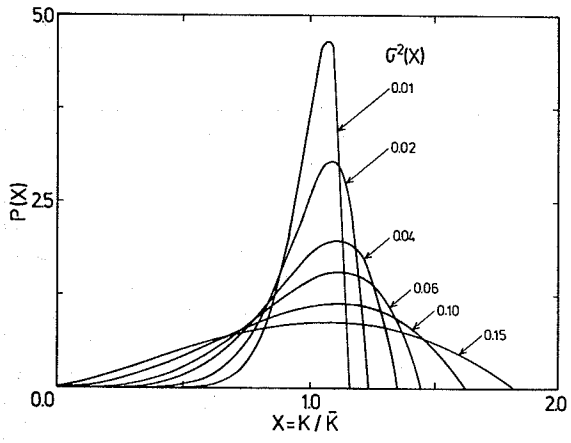

Fig. 3. Plots of the proposed $P(X)$ of eqn (1).

6. CONFIDENCE LIMITS

In statistics, the probability of occurrence of the stationary properties of specific samples drawn from a population set is usually expressed in terms of confidence limits. Here, the suitable question is: given daily solar radiation data for a specified month over a certain number of years, how can one determine upper and lower bounds for the mean of $\bar{K}$ and for the mean of $\sigma^2(X)$ at a specified confidence level? The answer to this question is useful in estimating the degree of uncertainty in the stationary PDF, since both $\bar{K}$ and $\sigma^2(X)$ are required as input parameters for use of eqn (1) for $P(X)$.

First, we test if the sampling distributions of the appropriate standardized random variables based on $\bar{K}$ and $\sigma^2(X)$ are normally distributed. This is nec-

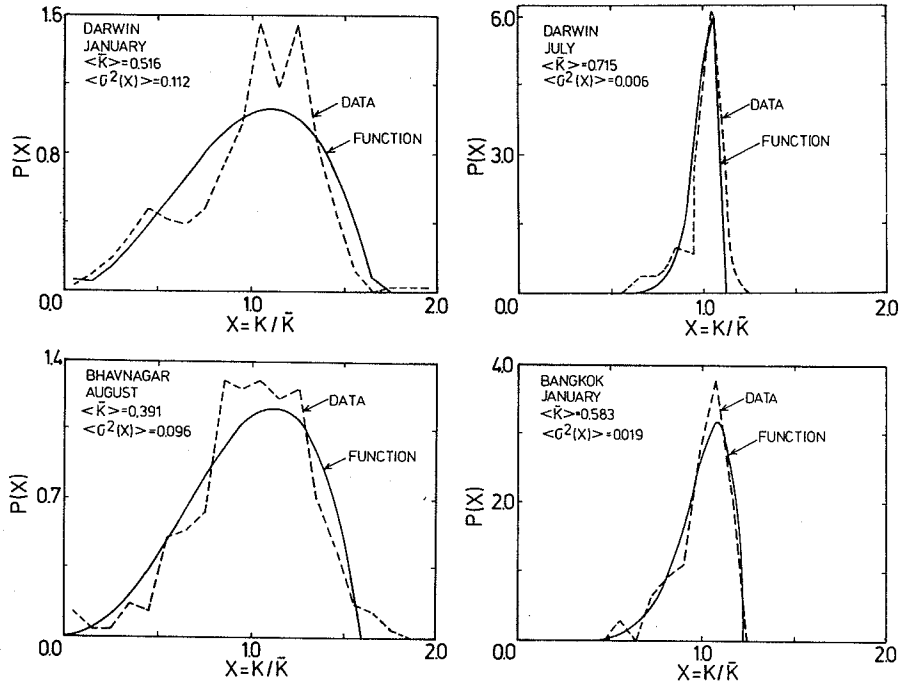

Fig. 4. Comparison of theoretical (eqn (1)) and data-based $P(X)$.

essary, in part, due to the limited record length, namely, 3 to 10 years only. Consider T years of daily solar radiation data for a given month r, and let j denote the day of the month. We define the standardized random variable Z:

$$Z(r,t) \equiv (Y(r,t) - \langle Y(r)\rangle)/\sigma(Y(r,t)) \qquad (5)$$

where Y represents either $\bar{K}$ or $\sigma^2(X)$.

Values of $\bar{K}$ and $\sigma^2(X)$ (i.e., Y of eqn (5)) were computed for the entire database delineated in Section 2. The probability distribution of the random variable Z based on these values, for both $\bar{K}$ and $\sigma^2(X)$, is presented in Fig. 5, together with the standard normal distribution. Both $\bar{K}$ and $\sigma^2(X)$ exhibit approximately normal distributions. This being the case, we

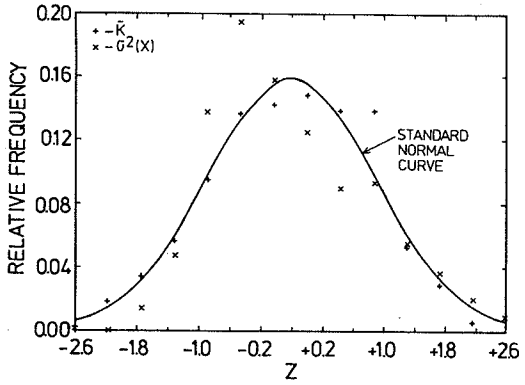

Fig. 5. Data-based probability distribution of the random variable Z (defined by eqn (5)) as a function of $Y = \bar{K}$ or $\sigma^2(X)$. For comparison, the standard normal distribution is plotted as a solid curve.

can now employ the probability tables of the Student t-distribution to compute upper and lower bounds for Z for specified confidence limits. The following example illustrates the method.

Example. For January in Darwin, Australia, the record length is $T = 10$ years. From the 10 values of $\bar{K}$ we calculate: $\langle\bar{K}\rangle = 0.516$ and $\sigma(\bar{K}) = 0.084$. For 9 degrees of freedom (i.e., $T - 1 = 9$) and a confidence limit of 90%, the Student t-distribution tables yield a critical value C of 1.83. Hence,

$$-C \le (\langle\bar{K}\rangle - \mu(\bar{K})\sqrt{T}/\sigma(\bar{K}) \le C$$

where $\mu(\bar{K})$ is the "expected mean" (i.e., the long-term mean in the absence of any long-term climatic changes). Thus, since $C\sigma(\bar{K})/\sqrt{T} = 0.049$, the expected mean of $\bar{K}$ is 0.516 ± 0.049, at the specified confidence limit of 90%.

We now present the magnitudes of the standard deviations of $\bar{K}$ and $\sigma^2(X)$ and examine any possible correlations between them and $\langle\bar{K}\rangle$ or between themselves. Figure 6 shows a plot of the sample standard deviation of the year-to-year variation in $\bar{K}$ vs. $\langle\bar{K}\rangle$. Figure 7 is a plot of the standard deviation of $\sigma^2(X)$ vs. $\langle\sigma^2(X)\rangle$. Figures 6 and 7 illustrate the point noted in Section 4 that statistical uncertainty increases with the order of the moment of X. Figures 6 and 7 may be useful, for example, in determining the uncertainty in system performance as estimated by analytic design methods.

7. SEQUENTIAL CHARACTERISTICS

Sequential or persistence effects on a day-to-day basis for daily global horizontal solar radiation se-

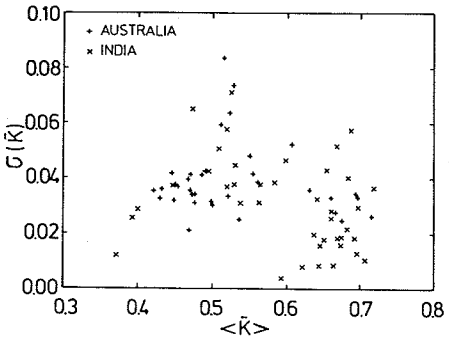

Fig. 6. Sample standard deviation of the year-to-year variation in $\bar{K}$ vs. $\langle \bar{K} \rangle$.

quences has been studied by [9,11,13–21] and involves computation of autocorrelation coefficients, $\rho(d)$, defined as[6,7]:

$$\rho(d) = \text{Cov}\,(Z(j), Z(j+d))/\{\sigma^2(Z(j)) \cdot \sigma^2[Z(j+d)]\}^{1/2} \quad (6)$$

where $\rho(d)$ is the dth day-lag autocorrelation coefficient, Cov denotes covariance, and Z is the random variable under consideration. Equation (6) can be reexpressed as

$$\rho(d) = C(d)/C(0) \quad (7)$$

where

$$C(d) = (1/(J-d)) \sum_{j=1}^{J-d} (Z(j) - \bar{Z}) (Z(j+d) - \bar{Z}) \quad (8)$$

for $d = 0,1,2, \ldots$.

The problem with analyzing daily data on a monthly basis is that a sequence of 30 values is not long enough for proper convergence of the summations in eqns (6) to (8). Hence, the random variable, Z, cannot simply be $X = K/\bar{K}$. We deal with this problem as follows. Since our objective is the calculation of autocorrelation coefficients that are representative of the mean for a particular month, over several years, and since the sequence of X values for a particular month has a different variance for each year, we define a standardized variable $Z(j,r,t)$ with zero mean and unit standard deviation:

$$Z(j,r,t) = (X(j,r,t) - 1)/\sigma(X(r,t)) \quad (9)$$

We then construct the following sequence—which is known as a "weak stationary stochastic sequence"[7]—for each month, r, for a specific location:

$$Z(1,r,1), Z(2,r,1), \ldots Z(J,r,1), Z(1,r,2), Z(2,r,2) \ldots, Z(J,r,2), \ldots, Z(1,r,T), Z(2,r,T), \ldots Z(J,r,T) \quad (10)$$

The autocorrelation coefficients $\rho(1)$ and $\rho(2)$ were then computed, as were higher-order coefficients. We find, however, that the first two autocorrelation coefficients are adequate to distinguish between actual persistence effects and random noise, and consequently have limited our attention to $\rho(1)$ and $\rho(2)$ only.

The two partial autocorrelation coefficients of interest, $\phi(n,n)$, are given by[6,7]:

$$\phi(1,1) = \rho(1) \quad (11)$$

$$\phi(2,2) = (\rho(2) - \rho^2(1))/(1 - \rho^2(1)). \quad (12)$$

Whereas autocorrelation coefficients, $\rho(d)$, for $d > 1$ retain the effects of persistence from lower values of d, the partial autocorrelation coefficient $\phi(d,d)$ "isolates" the correlation over d days independent of the implicit influence of correlations from lower values of d. Figures 8 and 9 present $\phi(1,1)$ and $\phi(2,2)$, respectively, vs. $\langle \bar{K} \rangle$.

Two points are worth noting. First, from Fig. 8, there *is* wide variation of $\phi(1,1)$ as a function of $\langle \bar{K} \rangle$. This stands in marked contrast to the conclusions of [17,19–21] that were restricted to a limited range of Canadian, Italian, and French locations. Second, we have performed a statistical test, at a significance level

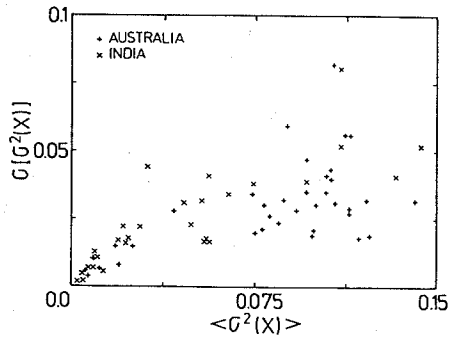

Fig. 7. Sample standard deviation of $\sigma^2(X)$ vs. $\langle \sigma^2(X) \rangle$.

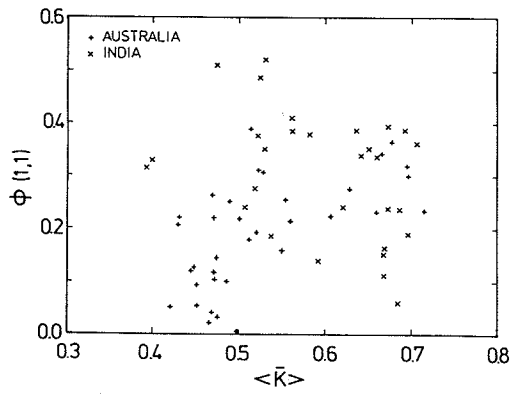

Fig. 8. Partial autocorrelation coefficient for one-day lag, $\phi(1,1)$, vs. $\langle \bar{K} \rangle$.

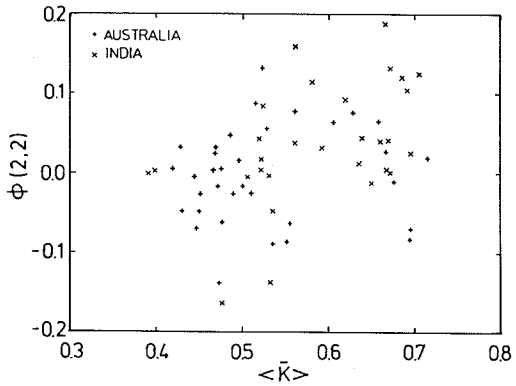

Fig. 9. Partial autocorrelation coefficient for two-day lag, $\phi(2,2)$, vs. $\langle \bar{K} \rangle$.

of .05, of the hypothesis that $\phi(1,1)$ and $\phi(2,2)$ are equal to zero (see Appendix B). We wish to ascertain whether any variation from zero is within the uncertainty levels. We find that in the vast majority of cases $\phi(1,1)$ cannot be taken as zero, whereas $\phi(2,2)$ can be taken as zero. However, for a nonnegligible number of months, $\phi(2,2)$ cannot be taken as zero, which is contrary to the conclusions of past studies[9,11,12,14,17,19–21,31].

8. GENERATION OF SYNTHETIC SEQUENCES

By "synthetic" sequences of daily solar radiation, we mean sequences that are simulated numerically from certain basic synoptic input parameters that characterize the statistical behavior of the actual data-based sequences. One motivation for this type of exercise is that one year's worth of such synthetic data, for example, could be used for simulating a solar energy system. Such a synthetic year would be superior to the so-called "Typical Meteorological Year"[36], which does *not* capture the correct statistics as accurately as the synthetic sequence generation method.

A second motivation is that the input synoptic parameters for the generation of synthetic sequences are easily "transportable" to other users, since a mere handful of synoptic parameters is required. Specifically, in the present study, for each month, only six synoptic parameters are required: $\langle \bar{K} \rangle$, $\langle \sigma^2(X) \rangle$, $\sigma(\bar{K})$, $\sigma(\sigma^2(X))$, $\phi(1,1)$ and $\phi(2,2)$. This stands in contrast to typically several orders of magnitude more data when years of daily solar radiation are involved.

The specific type of stochastic model we select for our analysis is influenced by the empirical fact that persistence times for daily solar radiation are relatively short (typically, 1–2 days). The general type of stochastic model we have selected is referred to as "autoregressive" (AR)[6,7], and is reviewed briefly in Appendix B.

For the specific objective of generating synthetic sequences, AR models are more flexible than the alternative of "Markov transition probability matrix" methods (e.g.) [11,12,14,33]) since the AR models can conveniently account for persistence effects that are longer than one day. Moreover, because of the relatively short persistence times for daily solar radiation, the complexity of "autoregressive moving average" models is not warranted[7].

The *order* of the model (i.e., the number of regressive autocorrelation coefficients to be retained in order to yield accurate results while remaining parsimonious) is still to be determined from our calculations. Related studies have found, however, that the AR models can be first[9,17,19–21,31], second[11], or third order[16,18].

An important problem with the AR model is that it inherently assumes that the residual error distribution is normal. However, this is *not* the case for daily solar radiation sequences. Therefore, straightforward use of the AR model would yield a synthetic sequence with the proper mean, variance, and autocorrelation coefficients, but with an *incorrect* stationary PDF[11,13,17,19–21]. We illustrate this inadequacy of AR models in our computations shortly. On a practical note, since long-term solar energy system performance is far more sensitive to an accurate stationary PDF than to accurate information on sequences[19], naïve use of AR models should be unsatisfactory for solar system design studies.

This shortcoming can be corrected via a Gaussian mapping technique[11,17,19–21]. To enable us to derive all expressions in closed form, we have used the empirical (but highly accurate) closed-form equation of [37] for the cumulative distribution of the error function $F(Z)$:

$$F(Z) = (1/2)\{1 \pm \sqrt{[1 - \exp(-2Z^2/\pi)]}\}$$

where the + sign applies to $Z > 0$ and the − sign to $Z < 0$.

We therefore consider three sets of results, for the broad range of locations and climates listed in Section 2, as follows:

1. AR1 = first-order autoregressive model *without* Gaussian mapping (to illustrate the inadequacy of this model in predicting the stationary PDF).
2. AR1MAP = first-order autoregressive model *with* Gaussian mapping.
3. AR2MAP = second-order autoregressive model *with* Gaussian mapping.

The AR1MAP and AR2MAP models require an inverse mapping procedure that involves determining synthetic radiation sequences, $X(j)$, by iteration from the cumulative frequency distribution implied by our closed-form expression for $P(X)$, eqn (1):

$$X(j)^{n+1} = (1/A)F(Z(j)) \Big/ \left[\frac{1}{n+1} - \frac{X(j)}{(n+2)\cdot X_{\max}}\right]$$

where the parameters A,n and $X_{\max}$ are given by eqns (2) to (4). An error tolerance of .001 for convergence was used in our iterative calculations. Also, the values needed to start the synthetic sequence generation were taken as the mean values of the series, although

Table 2. Comparison between actual synoptic parameters and the corresponding values as computed from synthetic data sequences of 600 values generated from three different auto-regressive (AR) models. Note that first-order models (AR1 and AR1MAP), by their nature, cannot generate a sequence with a correct partial autocorrelation coefficient for a lag of 2 days.

	Actual synoptic parameters	AR1	AR1MAP	AR2MAP
	January—Darwin, Australia—$\langle\bar{K}\rangle$ = .516			
$\bar{X}$	1.000	.991	.996	.995
$\sigma^2(X)$	.112	.107	.102	.102
$\phi(1,1)$	.388	.373	.359	.355
$\phi(2,2)$	.088	—	—	.056
	July—Darwin, Australia—$\langle\bar{K}\rangle$ = .715			
$\bar{X}$	1.000	.998	.998	.998
$\sigma^2(X)$	.006	.006	.007	.007
$\phi(1,1)$	.235	.229	.236	.236
$\phi(2,2)$	.019	—	—	−.044
	August—Bhavnagar, India—$\langle\bar{K}\rangle$ = .391			
$\bar{X}$	1.000	.992	.997	.997
$\sigma^2(X)$	.091	.088	.084	.084
$\phi(1,1)$	.315	.304	.292	.293
$\phi(2,2)$	.000	—	—	.017
	March—Bhavnagar, India—$\langle\bar{K}\rangle$ = .706			
$\bar{X}$	1.000	.998	.998	.998
$\sigma^2(X)$	.003	.003	.004	.004
$\phi(1,1)$	.364	.350	.341	.335
$\phi(2,2)$	.126	—	—	.036

this choice is not critical, especially for long series.

Our results are presented in two categories: (1) ability of a model to generate synthetic sequences with accurate mean, variance, and correlation coefficients and (2) ability of a model to predict the stationary PDF. Table 2 illustrates the former, while Table 3 and Fig. 10 illustrate the latter.

Table 3 and Fig. 10 exemplify the inadequacy of the AR1 model *without* Gaussian mapping in predicting the stationary PDF. Note that the differences in $P(X)$ between actual data and synthetically generated sequences stem from two sources. One is the fact that our eqn (1) for $P(X)$ is not perfect. The other is that the synthetic sequence is finite, and hence introduces uncertainties akin to selecting a finite sample from any population.

Table 3 illustrates the separate contributions of these two sources in terms of the χ^2 statistic. These separate contributions are presented in Table 3 in order to highlight the case where the user does not have actual data but rather must resort to using eqn (1) as the basis for selection of the appropriate sample (detailed below).

How sensitive are the results to sample size? Table 2 and Fig. 10 indicate that for the AR models *with* Gaussian mapping, even for a series of 600 values, the generated mean, variance, and stationary PDF *are* satisfactory. However, this is less clear for the generated autocorrelation coefficients, where 600 values would appear to be too small a sample size. These observations are similar to those from hydrological sequence studies[7].

If one wishes to generate only one year of synthetic data, how accurately will that one year represent the actual statistics of multiyear data? To illustrate this problem, we generated synthetic sequences for one month's worth of days for March in Bhavnagar, India, 10 separate times, each time with a different "seed" in the random number generator used. These results are presented in Table 4 for the AR2MAP model.

Ten separate generation runs of 30 values each,

Table 3. Accuracy of AR models in predicting stationary PDF, expressed as χ^2 values for 20 equistatistical bins (17 degrees of freedom), for synthetically generated daily solar radiation sequences 600 days long, vs. actual data. Numbers in parentheses refer to a comparison with the PDF of eqn (1).

Location	Month	$\langle\bar{K}\rangle$	$\langle\sigma^2(X)\rangle$	AR1	AR1MAP	AR2MAP
Darwin	January	.516	.112	103.0	66.9	70.3
				(76.7)	(27.6)	(27.5)
Darwin	July	.715	.006	94.6	112.3	82.2
				(103.9)	(33.9)	(31.6)
Melbourne	January	.555	.081	102.5	46.0	74.5
				(157.8)	(26.5)	(34.1)
Melbourne	June	.421	.114	40.5	27.4	35.9
				(79.2)	(23.8)	(23.5)
Bhavnagar	March	.706	.003	43.9	34.6	38.0
				(97.6)	(35.0)	(45.8)
Bhavnagar	August	.391	.091	33.8	26.2	26.3
				(118.1)	(22.7)	(21.2)
Nagpur	March	.636	.009	97.2	102.4	103.9
				(95.8)	(32.2)	(27.4)

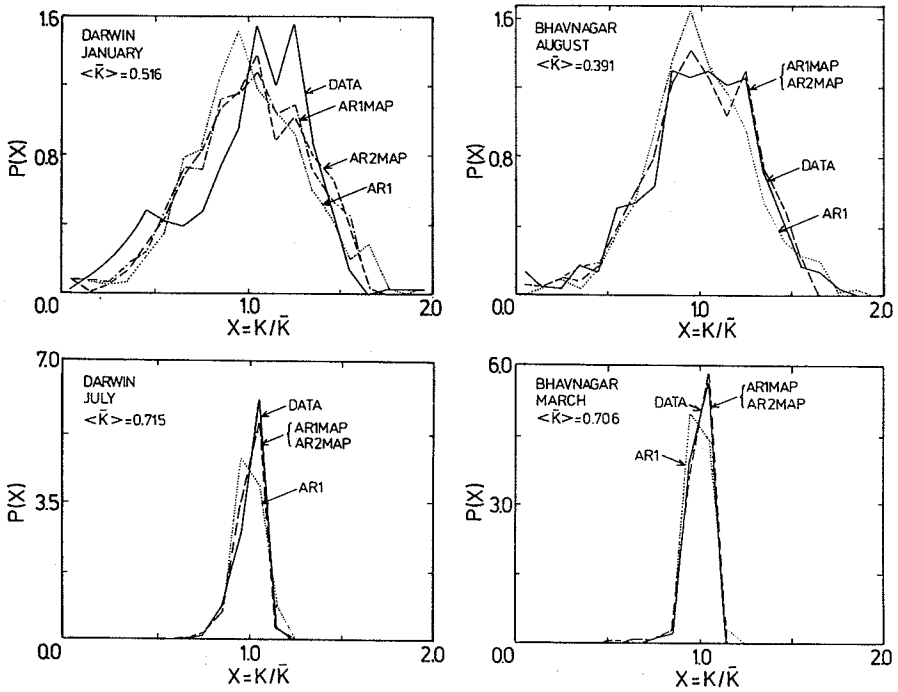

Fig. 10. Ability of AR models to generate accurate stationary PDF's, $P(X)$ (see Section 8).

and 5 generation runs of 600 values, were performed, the latter of which (understandably) exhibit far less variation from run to run than the former. The key observation is that, if one wants to generate one year only of synthetic data, then for each month one should perform several generation runs of synthetic sequences before selecting one run that most satisfactorily reproduces the actual synoptic input parameters.

Finally, the variable required for solar energy system design is the radiation, H, rather than X. For a given month, for a synthetic sequence of T years, the following procedure is recommended for converting X values into H values. Knowing $\langle \bar{K} \rangle$ and $\sigma(\bar{K})$, T

Table 4. Adequacy of "short-term" synthetic data. Comparison between actual synoptic parameters (Bhavnagar, March, $\langle \bar{K} \rangle$ = .706) and the corresponding values as computed from synthetic data sequences 30 values long from 10 separate generation runs, each run with a different "seed" for the random number generator. Results for five runs 600 values long are presented for comparison. Goodness of fit in predicting the stationary PDF is measured by the χ^2 statistic between $P(X)$ given by eqn (1) and synthetically generated daily solar radiation sequences (20 equistatistical bins and 17 degrees of freedom for the series of 600 values, and 5 equistatistical bins and 2 degrees of freedom for the series of 30 values).

	$\bar{X}$	$\sigma^2(X)$	$\phi(1,1)$	$\phi(2,2)$	χ^2
Actual synoptic data	1.000	0.003	0.364	0.126	—
Series of 30 values					
Run Number					
1	1.000	.005	.302	.061	2.67
2	1.005	.003	.440	.086	3.42
3	.988	.003	.402	−.126	4.76
4	.965	.003	.022	−.104	.52
5	.968	.005	.367	−.018	5.39
6	.967	.001	.279	.276	5.07
7	1.010	.002	.102	.525	1.78
8	1.004	.002	.176	.186	6.87
9	1.017	.001	.107	.050	14.31
10	1.004	.001	.284	.118	5.41
Series of 600 values					
Run Number					
1	.997	.003	.451	.139	26.78
2	.998	.003	.404	.148	18.62
3	.998	.003	.338	.080	18.48
4	1.004	.003	.353	.108	23.00
5	.998	.003	.396	.129	23.74

values of $\bar{K}$ are generated, which then multiply the respective X values for each of the T years to obtain the corresponding H values.

9. SUMMARY

The trends for optimal sizing procedures for solar energy systems in general, and stand-alone systems in particular, are (1) toward more accurate methods that take explicit account of the stochastic nature of solar radiation and its sequential characteristics on one hand, yet (2) toward relative fast, easy-to-use calculational tools on the other. We believe that the types of analysis and the results presented here are in the spirit of providing the kind of solar radiation input data necessary for analytic and/or time-saving calculational procedures.

One reason the results of time-series analysis are important in practical solar energy system design is that parameters such as the mean, the variance, and, to a lesser extent, the skewness of the stationary probability distribution of daily radiation, and the persistence (day-to-day correlation) strength, are essential input data for stochastic models (e.g., for the sizing and performance of stand-alone solar energy systems[1,2,4,38]).

A second reason is that even if such analytic design tools are not available, enough synoptic information can be distilled from daily radiation data for a given site that synthetic daily radiation sequences can be generated. These synthetic sequences would be superior to a so-called "Typical Meterological Year"[36]. Furthermore, since the generation of synthetic sequences is based on a handful of parameters, the synthetic sequence approach is easily "transportable" to other users.

Our analyses expand the range of climates and locations considered in solar radiation statistical studies, particularly for low-latitude tropical climates. We feel, however, that the value of our presentation goes beyond simply reporting on a broader database than past studies. Specifically, for daily horizontal global solar radiation, we have:

1. Segmented the year into monthly time scales and identified an appropriate variable for statistical analysis on the monthly basis
2. Calculated the mean monthly statistical parameters (e.g., mean, variance, and skewness), considered correlations among these parameters and the physical basis for such correlations
3. Noted key distinctions between temperate and tropical climates in terms of the stationary properties of daily solar radiation and presented evidence that runs contrary to the conclusions of previous studies regarding the universal nature of these statistical parameters
4. Proposed a simple functional form for the stationary probability density for daily solar radiation, which requires a knowledge of the mean and variance only and compared it against actual data from widely differing climatic stations
5. Suggested a procedure based on confidence limits that enables one to ascertain the uncertainty range of the mean and the variance, given a limited number of yearly samples for a given month at a particular location
6. Studied the *sequential* (day-to-day) behavior of daily solar radiation, which involves analysis of the stochastic components of daily solar radiation toward determining persistence times and persistence strengths and found a *lack* of general correlations between them and $\langle \bar{K} \rangle$
7. Examined several types of autoregressive stochastic models toward the generation of synthetic daily solar radiation sequences that capture all the essential statistical stationary and sequential information, for a broad range of climates, and noted how an appropriate synthetic sequence can be selected and used for solar system simulation.

Acknowledgments—We are grateful to A. Mani for kindly supplying us with radiation data of Indian locations. We are very appreciative to A. Zemel for a careful and critical reading of the manuscript, constructive comments, and useful discussions. We also acknowledge useful discussions with Y. Tsur and Y. Zarmi.

NOMENCLATURE

A	normalization constant (eqn (1))
d	index for persistence over consecutive days
$F(Y)$	cumulative frequency distribution of the variable Y
j	index for day of the month
J	number of days in the month
K	daily clearness index
n	exponent in eqn (1)
$P(X)$	Probability Density Function (PDF)
r	index for month of the year
s	index denoting the order of the autoregressive (AR) model
S	skewness
t	index for year
T	number of years of available data for a particular location
X	$K/\bar{K}$
X_{max}	maximum value of X
$\bar{Y}$	monthly mean value of the variable Y
$\langle \bar{Y} \rangle$	grand average of the variable Y (i.e., average of $\bar{Y}$)
Z	normalized random variable with zero mean and standard deviation of unity
α	significance level of hypothesis test
σ	standard deviation
σ^2	variance
$\mu(Y)$	population mean of the variable Y
ϵ	error or random noise term
$\rho(d)$	autocorrelation coefficient at lag of d days
$\phi(d)$	coefficient of autoregressive (AR) model at lag of d days
$\phi(d,d)$	partial autocorrelation coefficient at lag of d days
χ^2	chi-square statistic

REFERENCES

1. L. L. Bucciarelli, *Solar Energy* **32,** 205 (1984).
2. L. L. Bucciarelli, *Solar Energy* **36,** 11 (1986).

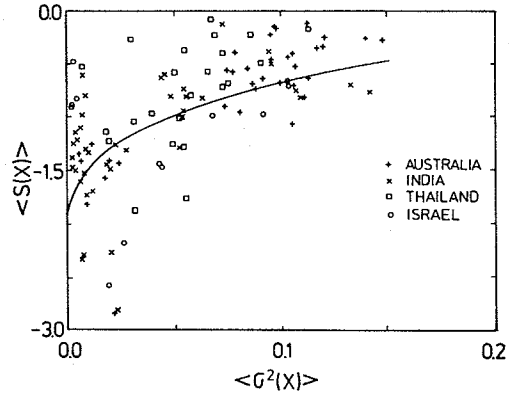

Fig. 11. Data-based plot of grand average skewness $\langle S(X)\rangle$ vs. grand average variance $\langle\sigma^2(X)\rangle$. All points below the solid curve do not satisfy eqn (A-2) (Appendix A).

APPENDIX B
AUTOREGRESSIVE MODELS

Autoregressive (AR) models[6,7] are stochastic models in which the current value of the process is expressed as a finite, linear sum of previous values plus random noise ϵ. In our analysis, the variable of interest is daily horizontal solar radiation. For the stochastic variable $Z(j) = (X(j) - 1)/\sigma(X)$, where j denotes the day, an AR process of order s is expressed as

$$Z(j) = \phi(1)\,Z(j-1) + \phi(2)\,Z(j-2) + \cdots + \phi(s)\,Z(j-s) + \epsilon(j).$$

The regression parameters, $\phi(1), \phi(2), \ldots \phi(s)$, are given by:

(1) for a first-order AR process.

$$\phi(1) = \rho(1)$$

(2) for a second-order AR process,

$$\phi(1) = \rho(1)(1 - \rho(2))/(1 - \rho^2(1))$$

$$\phi(2) = (\rho(2) - \rho^2(1))/(1 - \rho^2(1)),$$

where $\rho(1)$ and $\rho(2)$ are the autocorrelation coefficients for one- and two-day lags, respectively (eqns (6)–(8). Note that the autoregressive parameter over a two-day period, $\phi(2)$ is equal to the partial autocorrelation coefficient at a lag of two days, $\phi(2,2)$ (see eqn (12)).

The reliability of the parameters estimated from actual data is expressed in terms of the confidence interval for the parameter $\phi(j)$:

$$[\phi(j) - U(1-\alpha/2)\,\sigma(\phi(j)),\ \phi(j) + U(1-\alpha/2)\,\sigma(\phi(j))]$$

where $U(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of the standardized normal distribution. $\sigma(\phi(j))$ is the standard deviation of $\phi(j)$, which is estimated as follows:

First-order AR process:

$$\sigma^2(\phi(1)) = (1 - \phi^2(1))/TJ - 1)$$

Second-order AR process:

$$\sigma^2(\phi(2)) = \sigma^2(\phi(1)) = (1 - \phi^2(2))/(TJ - 2)$$

The random noise, ϵ, should have zero mean, be normally distributed, and have a variance as follows:

(a) for a first-order AR model,

$$\sigma^2(\epsilon(j)) = 1 - \rho^2(1)$$

(b) for a second-order AR model,

$$\sigma^2(\epsilon(j)) = 1 - \phi^2(1) - \phi^2(2) - 2\phi^2(1)\phi(2)/(1 - \phi(2))$$

3. J. M. Gordon and P. Zoglin, *Solar Cells* **17,** 285 (1986).
4. J. M. Gordon, *Solar Cells* **20,** 295 (1987).
5. T. A. Reddy, J. M. Gordon, and I. P. D. de Silva, *Solar Energy* **39,** 123 (1987).
6. G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco (1970).
7. J. D. Salas, J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrological Time Series*, Water Resources Publications, Colorado State University, Fort Collins, CO (1980).
8. C. E. P. Brooks and N. Carruthers, *Handbook of Statistical Methods in Meteorology*. Her Majesty's Stationery Office, London (1953).
9. B. J. Brinkworth, *Solar Energy* **19,** 343 (1977).
10. T. N. Goh and K. J. Tan, *Solar Energy* **19,** 755 (1977).
11. C. Mustacchi, V. Cena, and M. Rocchi, *Solar Energy* **23,** 47 (1979).
12. R. H. B. Exell, *Solar Energy* **26,** 161 (1981).
13. J. D. Engels, S. M. Pollock, and J. A. Clark, *Solar Energy* **26,** 91 (1981).
14. J. I. Jimenez, *Solar Energy* **26,** 497 (1981).
15. A. J. Biga and R. Rosa, *Solar Energy* **27,** 149 (1981).
16. R. H. Skaggs, D. G. Baker, and J. E. Ljungkull, *Solar Energy* **28,** 281 (1982).
17. E. Boileau, *Solar Energy* **30,** 333 (1983).
18. L. Vergara-Dominguez, R. Garcia-Gomez, A. R.. Figueiras-Vidal, J. R. Casar-Carredera, and F. J. Casajus-Quiros, *Solar Energy* **35,** 483 (1985).
19. U. Amato, A. Andretta, B. Bartoli, B. Coluzzi, V. Cuomo, and C. Serio, *Il Nuovo Cimento* **8,** 248 (1985).
20. U. Amato, A. Andretta, B. Bartoli, B. Coluzzi, V. Cuomo, F. Fontana, and C. Serio, *Solar Energy* **37,** 179 (1986).
21. V. A. Graham, K. G. T. Hollands, and T. E. Unny, Stochastic modeling of the daily solar atmospheric transmittance Kt. Intersol 85, 9th Biennial ISES Congress, p. 2473, ed. E. Bilgen and K. G. T. Hollands (1986).
22. A. Mani and S. Rangarajan, *Solar Radiation Over India*. Allied Publishers, New Delhi (1982).
23. C. W. Richardson, *Water Resources Research* **17,** 182 (1981).
24. P. Bendt, M. Collares-Pereira, and A. Rabl, *Solar Energy* **27,** 1 (1981).
25. K. G. T. Hollands and R. J. Huget, *Solar Energy* **30,** 195 (1983).
26. H. Madsen, H. Spliid, and P. Thyregod, *J. Climate and Appl. Meteorol.* **24,** 629 (1985).
27. J. M. Gordon and M. Hochman, *Solar Energy* **32,** 329 (1984).
28. G. Y. Saunier, T. A. Reddy, and S. Kumar, *Solar Energy* **38,** 169 (1987).
29. I. Bennett, *Solar Energy* **11,** 41 (1967).
30. J. D. Engels, J. A. Clark, and S. M. Pollock, *Solar Energy* **26,** 471 (1981).
31. A. A. Sfeir, *Solar Energy* **26,** 497 (1981).
32. J. A. Olseth and A. Skartveit, *Solar Energy* **33,** 533 (1984).
33. R. J. Aguiar, M. Collares-Pereira, and J. P. Conde, *Solar Energy* **40,** 269 (1988).
34. T. A. Reddy, *The Design and Sizing of Active Solar Thermal Systems*. Oxford University Press, Oxford (1987).
35. B. Y. H. Liu and R. C. Jordan, *Solar Energy* **7,** 53 (1963).
36. I. Hall, R. Prairie, H. Anderson, and E. Boes, Generation of typical meteorological years for 26 SOLMET stations. Sandia Laboratories Report SAND78-1601, Albuquerque, NM, August (1978).
37. A. A. Sfeir, *Solar Energy* **25,** 149 (1980).
38. M. Bida and J. F. Kreider, The suitability of the two and three-state models for representing actual distributions in the prediction of stand-alone photovoltaic systems performance. *Proc. 1987 A.S.E.S. Annual Meeting*, pp. 157–161, Portland, OR, 11–16 July (1987).
39. P. J. Walsh, M. C. Munro, and J. W. Spencer, An Australian climatic data bank for use in the estimation of building energy use. Division of Building Research, C.S.I.R.O., Highett, Australia (1983).
40. Meteorological Department of Thailand, Hourly and daily global insolation data for Bangkok and Chiang Mai for the period 1968–1972. Bangkok Thailand.
41. Probability distributions are based on analyses of raw data from A. Manes, A. Teitelman, and I. Freuhling. Solar Radiation and Radiation Balance at Bet Dagan, Central Meteorological Institute, Series A & B, 1962–1977, Israel Meteorological Service, Bet Dagan, Israel (1979), as reported in [24].

APPENDIX A
FUNCTIONAL FORMS FOR THE PROBABILITY DENSITY FUNCTION (PDF)

In arriving at the proposed empirical functional form for the stationary PDF given by eqn (1), we considered functional forms of a more general class, specifically:

$$P(X) = A(X/X_{max})^{m-1}\,(1 - (X/X_{max}))^{n-1} \qquad \text{(A-1)}$$

where $X = K/\bar{K}$. The functional form of eqn (A-1) is convenient because the normalization integral of $P(X)$ is the beta function, and further integrals can be expressed as gamma functions. In eqn (A-1), the nonadjustable parameters A, m, n and X_{max} are determined from normalization of the PDF plus a knowledge of the mean, variance, and skewness.

The problem arises, however, that when, in addition to the mean and the variance, the skewness is treated as a known input parameter, the value on n in certain instances turns out to be less than unity, which yields a $P(X)$ that diverges, rather than vanishes, at large X. In fact, in order to ensure than $n > 1$, the skewness $S(X)$ must satisfy the following inequality:

$$S(X) > (\sigma^2(X) - 1 - 4g\,\sigma(X))/(\sigma(X) - g + g\sigma^2(X)) \qquad \text{(A-2)}$$

where

$$g = (\sigma(X)/2)[1 - (1/\sigma^2(X)) + 1)^{1/2}] \qquad \text{(A-3)}$$

To illustrate how many months of daily data did *not* satisfy eqn (A-2), we present in Fig. 11 a plot of data-based $S(X)$ vs. $\sigma^2(X)$. Since eqn (A-2) could not handle all climates considered, we discarded eqn (A-1) as a potentially universal functional form in the present study. However, since the noise associated with the estimate of skewness from a limited series of random numbers is large, a possible approach would be to set all those data-based estimates of skewness values which do not satisfy eqns (A-2) to (A-3) equal to the right-hand side of eqn (A-2) and still resort to eqn (A-1).

We then considered the more limiting functional form:

$$P(X) = AX^n \cdot [1 - (X/X_{max})^N] \qquad \text{(A-4)}$$

where N is constrained to be a positive integer. It turns out that $N = 1$ yields the best fits when compared with our data-based $P(X)$ curves. Accordingly, eqn (1) was selected for our analyses.