Appendix III
Nonparametric Statistics for TMY Method

The basis of the nonparametric statistic used in the TMY method is described and procedure for calculating the statistical values is outlined.

Nonparametric inference

The Finkelstein-Schafer (FS) statistic used in the TMY method is a kind of nonparametric methods for testing statistical hypothesis. Established based on 'ranks', a nonparametric test use less information and has less restrictions than the normal theory statistical methods, such as a 'Student-t test' (Milton and Arnold, 1990, Chp. 8). However, they are generally less powerful statistically than their 'parametric' counterparts when the assumptions underlying the normal theory test are met. There are a wide range of nonparametric tests. Some are appropriate for comparing distributions which are different mainly in location (such as shifting of means); some are for differences in spread; some are designed for a quick check for differences in a more general nature. Although there is some overlap in sensitivity between the tests, by and large, none of them work well except in the specific situations for which they were designed (Neave and Worthington, 1988, Chp. 4).

The FS statistic is a 'catch-all' test that offer a general-purpose tool for quickly assessing the 'goodness-of-fit' between two distributions. The comparison of cumulative distribution function (CDF) in the TMY method is in fact testing the statistical hypothesis that "the yearly CDF is identical to the long-term CDF". This is the null hypothesis of the test. If the null hypothesis is rejected, an 'alternative hypothesis' will be the conclusion reached. For a two-

sided statistical test [1], the null hypothesis and alternative hypothesis can be written as:

$$H_0:\ F_Y(x) = F_X(x) \qquad \text{for all } x \tag{A3.1}$$

$$H_1:\ F_Y(x) \neq F_X(x) \qquad \text{for some } x \tag{A3.2}$$

where $F_Y(x)$ and $F_X(x)$ denote the two distributions being compared.

The normal procedure in statistical testing is to decide whether the hypothesis can be accepted under a certain significant level — a measure of confidence or uncertainty for the test statistic. However, when the nonparametric test is applied in the TMY method (NCC, 1981), no information has been provided on the 'significance' of the test.

Formulation of test statistic

The FS statistic was originally used for testing 'goodness-of-fit' for a small discrete sample *vis-á-vis* a continuous distribution function (Finkelstein and Schafer, 1971). But the values of the test statistic are used in the TMY method for comparing two 'discrete' distributions (the yearly and long-term CDF's). The formula to calculate the FS statistic (see Equation (5.2)) becomes ambiguous since the two distributions under comparison have different numbers of data (the long-term CDF has much more data than the yearly CDF). Proper formulation for the CDF comparison should take the number of elements of the long-term CDF, not the number of daily readings for that month.

Another unclear point in the TMY method is the treatment of 'ties' (observations with equal value) in the data and when selecting years. Ties are common in meteorological data because data are usually given in fixed decimal

[1] A 'two-sided' test looks for evidence contrary to the null hypothesis by comparing the realisation of the test statistic with two values, one of which would be considered to be unusually small and the other unusually large, if the null hypothesis were true. These values, called 'critical values', are chosen so that the test has a specified significance level that the null hypothesis will be rejected when it is in fact true. A 'one-side' test works in exactly the same way, except that the realisation of the test statistic is compared with only one critical value (Milton and Arnold, 1990).

places, such as temperature in 0.1 °C. It is proposed that when a tie occurs in the weather data, the difference between the two CDF's is to be calculated at the end of the tie. When a tie happens during selection of years, all the years with equal values will be taken. The procedure for comparing the CDF's and calculating the nonparametric statistics (including FS and KS statistics) is outlined in the following paragraphs.

Rank order values and CDF

Two sample populations of sizes *m* and *n* respectively are the input data. The sample space for each year (with *n* number of elements) is $X=\{x_1, x_2, ..., x_n\}$ and the sample space for the long term (with *m* number of elements) is $Y=\{y_1, y_2, ..., y_m\}$. First, the values are arranged in ascending order with the number of occurrence of repeated values counted and taken out. The rank order statistics of *X* and *Y* can be expressed as:

$$X_{(\,)}=\{x_{(1)} < x_{(2)} < ... < x_{(nx)}\} \text{ and } Y_{(\,)}=\{y_{(1)} < y_{(2)} < ... < y_{(my)}\} \qquad \text{(A3.3)}$$

where *nx* = total number of different values of *x* after ranking

n - *nx* = total nos. of repeated values of *x*

my = total number of different values of *y* after ranking

m - *my* = total nos. of repeated values of *y*

Subscript numbers enclosed in parenthesis is the convention used to represent ranked order variable. The empirical CDF for *X* is defined as:

$$S_n(x) = \begin{cases} 0 & \text{for } x < x_{(1)} \\ k \,/\, n & \text{for } x_{(k)} \le x \le x_{(k+1)} \\ 1 & \text{for } x \ge x_{(n)} \end{cases} \qquad \text{(A3.4)}$$

where *S n(x)* = value of the cumulative distribution function at *x*

n = total number of elements

k = rank order number (where *k* = 1,..., *n* - 1)

The original TMY method defined in the *TMY User's Manual* (NCC, 1981) uses *(k - 0.5)/n* for the empirical CDF. But it is believed that *k / n* is the more usual expression found in mathematics and statistics. Therefore, it is used here for the CDF formulation. The empirical CDF for *Y* can be defined by replacing *x* by *y* and *n* by *m* in Equation (A3.4). Their values at points $x_{(1)}$ to $x_{(nx)}$ and $y_{(1)}$ to $y_{(my)}$ can be calculated respectively using the following iterative schemes:

$$\text{For } X: \begin{cases} S_n(x_{(1)}) = NoR_{x_1} / n \\ S_n(x_{(i)}) = S_n(x_{(i-1)}) + NoR_{x_i} / n \quad \text{where } i = 2, 3, \ldots, nx\text{-}1 \\ S_n(x_{(nx)}) = 1 \end{cases} \tag{A3.5}$$

$$\text{For } Y: \begin{cases} S_m(y_{(1)}) = NoR_{y_1} / m \\ S_m(y_{(j)}) = S_m(y_{(j-1)}) + NoR_{y_i} / m \quad \text{where } j = 2, 3, \ldots, my\text{-}1 \\ S_m(y_{(my)}) = 1 \end{cases} \tag{A3.6}$$

where $NoRx_i$ = no. of occurrences counted at $x_{(i)}$ for $i = 1, 2, \ldots, nx$

$NoRy_j$ = no. of occurrences counted at $y_{(j)}$ for $j = 1, 2, \ldots, my$

Comparing CDF's

In general, the two distributions should be compared at every values within their ranges. The number of comparisons to be made depends on the size of the union product of the two distributions. Let $Z_{(\,)} \equiv X_{(\,)} \cup Y_{(\,)}$ be the union product of the two ordered variables to be compared. The total number of elements in $Z_{(\,)}$ is then determined by simple sets theory as follows:

- If $X_{(\,)}$and $Y_{(\,)}$ are completely distinct, total number of $Z_{(\,)}$ is $nx + my$.
- If $Y_{(\,)}$ contains all $X_{(\,)}$ (i.e. $X_{(\,)} \subseteq Y_{(\,)}$), then total number of $Z_{(\,)}$ is my , and vice versa.
- If $Y_{(\,)}$ only contains some elements of $X_{(\,)}$, then the total number of $Z_{(\,)}$ is determined from the sum ($nx + my$) minus the number of elements of the intersection product $X_{(\,)} \cap Y_{(\,)}$.

Let $Z_{(\,)} = \left\{ z_{(1)} < z_{(2)} < .... < z_{(p)} \right\}$ be the required union product where p is the total number of elements determined. The absolute difference between the two CDF's is computed as follows:

$$\delta_i = \left| \; S_n(z_{(i)}) - S_m(z_{(i)}) \; \right| \quad \text{for } i = 1, 2, \ldots, p \tag{A3.7}$$

The FS statistic is then calculated based on the p number of the difference (not the number of daily readings for the month, as mentioned in NCC (1981)):

$$FS = \frac{1}{p} \sum_{i=1}^{p} \delta_i \tag{A3.8}$$

Kolmogorov-Smirnov (KS) statistic

The Kolmogorov-Smirnov (KS) two-sample statistic is a well-known nonparametric test for general difference. The basic formulation of the KS statistic is defined as:

$$KS = \max \left| \, \delta_i \, \right| \qquad \text{for } i = 1, 2, \ldots, p \tag{A3.9}$$

The KS statistic is more sensitive to localised discrepancies or sharp differences while the FS statistic is sensitive to the average effect along the whole interval of values. Since the difference between two CDF's at a point is an accumulating difference for the probability functions, the maximum of these differences indicated by the KS statistic seems to be more logical and reasonable than the FS statistic.

Weighted-sum average

The weighted-sum average of the nonparametric statistic (FS or KS statistic) is determined from the weighted average of the values of the statistic calculated for the different parameters of interest. In general, it can be expressed as:

$$WS = \sum_{i=1}^{N} WF_i \cdot \beta_i \tag{A3.10}$$

where WS = weighted-sum average of the nonparametric test

WF_i = weighting factor for the i^{th} parameter

δ_i = value of nonparametric statistic calculated for the i^{th} parameter

N = total number of parameters considered

In this study, the same set of weighting factors are being used for the FS and KS statistics (see Table 5.4).