Estimation of WGEN Weather Generation Parameters in Arid Climates

M. E. Taulis[a *], M. W. Milke[b]

[a] Department of Civil Engineering, University of Canterbury, Private Bag 4800, Christchurch, NEW ZEALAND. [*] Corresponding author. Tel.: +64 3 3642987 ext 7332; fax: +64 3 3642774; email: mauricio@taulis.com. Contact mauricio@taulis.com for details on how to get the modified weather generation parameters software.

[b] Department of Civil Engineering, University of Canterbury, Private Bag 4800, Christchurch, NEW ZEALAND

Abstract

This research discusses some of the issues encountered while developing a set of WGEN parameters for Chile and advice for others interested in developing WGEN parameters for arid climates. The WGEN program is a commonly used and a valuable research tool; however, it has specific limitations in arid climates that need careful consideration. These limitations are analysed in the context of generating a set of WGEN parameters for Chile. Fourteen to twenty-six years of precipitation data are used to calculate precipitation parameters for eighteen locations in Chile, and three to eight years of temperature and solar radiation data are analysed to generate parameters for seven of these locations. Results indicate that weather generation parameters in arid regions are sensitive to erroneous or missing precipitation data. Research shows that the WGEN-estimated gamma-distribution shape parameter (α) for daily precipitation in arid zones will tend to cluster around discrete values of 0 or 1, masking the high sensitivity of these parameters to additional data. Rather than focus on the length in years when assessing the adequacy of a data record for estimation of precipitation parameters, researchers should focus on the number of wet days in dry months in a dataset. Analysis of the WGEN routines for the estimation of temperature and solar radiation parameters indicates that errors can occur when individual 'months' have fewer than two wet days in the dataset. Recommendations are provided to improve methods for estimation of WGEN parameters in arid climates.

Keywords

WGEN, synthetic weather generation, arid and semi-arid climates, Chile.

Manuscript

1. Introduction

Weather data are needed to assess policy and design decisions, and to take into account future events that can impact on these decisions. For example, weather data are used when carrying out designs for dams, landfills, and stormwater projects, to mention a few. The longer the weather record, the better the risks associated with these projects or policies can be assessed. However, many projects and policies will have effects over a much longer period of time than the weather record itself. Also, these records might be incomplete and contain erroneous data from transcription and computational mistakes. The use of solely historic data provides an analysis that is based on just one of many weather possibilities. To address these problems, stochastic models are often used to create different sets of synthetic data having the same statistical characteristics as historic data. These data span longer periods of time than the original weather record, they are more reliable than historic data because they are free from mistakes, and they provide a wider range of weather possibilities within the historic bounds.

There are several stochastic models for generating synthetic climatic data (Bond 1979, Bruhn et al. 1980, Nicks and Harp 1980, Richardson 1981). All of them generate one or more climatic variables from an existing set of data. Wilks and Wilby (1999) provide a valuable review of the topic of weather generation models. The WGEN weather generator model (Richardson and Wright 1984) is a stochastic model used to generate daily weather variables. WGEN uses a first-order Markov chain to decide whether a day is wet or dry. This is done by analysing historic precipitation records for each location and by using simple probability to define the wet or dry status of a given day. A wet day is defined as a day in which the occurrence of rain has been recorded. Thus, the following probabilities are defined:

$P_i\,(W/W)$: probability of a wet day on day i given a wet day on day i-1. (1)
$P_i\,(W/D)$: probability of a wet day on day i given a dry day on day i-1. (2)

The WGEN model uses a two-parameter gamma distribution to synthesize the distribution of rainfall amounts. Richardson (1982a) has shown that this density function generally fits the distribution of daily precipitation amounts. The two-parameter gamma distribution density function is given by:

$$f(p) = \frac{p^{\alpha-1} \cdot e^{-p/\beta}}{\beta^{\alpha} \cdot \Gamma(\alpha)}, \qquad p, \alpha, \beta > 0 \tag{3}$$

where: p is a random variable for daily precipitation; $f(p)$ is the density function for daily precipitation; α is the shape parameter; β is the scale parameter; e is the exponential function; $\Gamma(\alpha)$ is the gamma function of α. The typical values of α and β for weather generation are $0 < \alpha < 1$ and $\beta < 25.4$ *mm (1 inch).*

The α and β parameters are determined by analysing historic records for each location, and are defined for each month taking into account the whole data record and using the parameter generation procedure described by Haan (1977). This procedure calculates maximum likelihood estimators for the parameters of the gamma distribution (α and β) by taking into account the precipitation amount for wet days on a given month.

In WGEN, the temperature and solar radiation procedure is based on a weakly stationary process described by Matalas (1967). This process is used to develop a multivariate model considering the wet or dry status of a given day. The daily temperature and solar radiation values are generated from historic data using the following equations:

$$t_i(k) = m_i(k) \cdot \left[\chi_i(k) \cdot c_i(k) + 1\right] \tag{4}$$

where: $t_i(k)$ = the daily value (for day i) of maximum temperature (k = 1), minimum temperature (k = 2), or solar radiation (k = 3); $m_i(k)$ = the mean of maximum temperature, minimum temperature or solar radiation for day i; $\chi_i(k)$ = the residual element for day i; $c_i(k)$ = the coefficient of variation of maximum temperature (k = 1), minimum temperature (k = 2), or solar radiation (k = 3).

The means and the coefficients of variation are expected to experience seasonal changes. In WGEN the year is divided into thirteen 28-day 'seasons'. The seasonal changes are accounted for using the following harmonic equation:

$$u_i = \bar{u} + C \cdot \cos\left[\frac{2\pi}{365} \cdot (i - T)\right] \qquad i = 1...365 \tag{5}$$

where: $u_i = m_i(k)$ or $u_i = c_i(k)$; $\bar{u}$ = the mean value of u_i; C = the amplitude of the harmonic; T = the position of the harmonic in days. In this case, $\bar{u}$, C and T must be determined for wet and dry days from actual weather data. The residual elements are obtained from the following equation:

$$\chi_i(j) = A \cdot \chi_{i-1}(j) + B \cdot \varepsilon_i(j) \qquad (6)$$

where: $\chi_i(k)$ = 3x1 matrix of residual elements for day *i* (with k = 1 for maximum temperature, k = 2 for minimum temperature, and k = 3 for solar radiation); $\chi_{i-1}(k)$ =3x1 matrix of residual elements for day *i-1*; *A* and *B* are 3x3 matrices calculated from the lag 0 cross correlation coefficients, lag 1 cross correlation coefficients, and lag 1 serial correlation coefficients between maximum temperature, minimum temperatures, and solar radiation. The *A* and *B* matrices are calculated using equations (7) and (8):

$$A = M_1 M_0^{-1} \qquad (7)$$

$$B = M_0 - M_1 M_0^{-1} M_1^T \qquad (8)$$

Where matrices M_1 and M_0 are defined as follows:

$$M_0 = \begin{bmatrix} 1 & \rho_0(1,2) & \rho_0(1,3) \\ \rho_0(1,2) & 1 & \rho_0(2,3) \\ \rho_0(1,3) & \rho_0(2,3) & 1 \end{bmatrix} \qquad (9)$$

$$M_1 = \begin{bmatrix} \rho_1(1) & \rho_1(1,2) & \rho_1(1,3) \\ \rho_1(2,1) & \rho_1(2) & \rho_1(2,3) \\ \rho_1(3,1) & \rho_1(3,2) & \rho_1(3) \end{bmatrix} \qquad (10)$$

In these matrices, $\rho_0(i, j)$ is the lag 0 cross correlation coefficient between variables *i* and *j* (where *i* and *j* can have values of 1 for maximum temperature, 2 for minimum temperature, or 3 for solar radiation). In this context, the term "lag" refers to the number of time steps between the observations being considered. Thus, lag 0 refers to the correlation between two variables taken at the same time, and lag 1 refers to the correlation between two variables taken apart with a time step of 1. Therefore, at each location, the residuals of maximum temperature, minimum temperature, and solar radiation are serially correlated and cross-correlated through matrices *A* and *B*.

Richardson and Wright (1984) calculated correlation coefficients for 31 locations in the United States. They divided the year into thirteen, 28-day seasons and calculated the coefficients for each season and for each location. Richardson and Wright found that the seasonal and spatial variations of these coefficients were very small, so they averaged them in order to come up with generic values for matrices M_0, M_1, *A*, and *B*. This simplification can be inaccurate when considering areas presenting extreme latitudes because most of these coefficients are significantly correlated to latitude (Richardson 1982b).

Historic data are necessary to calculate the parameters for each location. These are daily precipitation, maximum temperature, minimum temperature, and solar radiation data collected over a period of time. Richardson (2000) has suggested that at least 20 years of daily precipitation data and 10 years of daily temperature (maximum and minimum) and solar radiation are needed to carry out the parameter generation calculations. In the case of arid regions, more years of daily data (30 years for example) would be needed (Richardson and Wright 1984). Unfortunately, the large number of years required, along with a limited research budget, can make the data collection process a difficult one.

The quality of data is equally as important as the quantity (the number of years used in the calibration procedure). Unfortunately, it is impossible to obtain a set of data files without missing days or suspicious measurements. Therefore it is desired to filter the data for any outliers and errors, and to replace these values with appropriate ones.

Care is needed when examining the assumptions of synthetic weather generation programs. For example, February 29 is deleted from leap years when generating the parameters for WGEN. This is necessary because, for practical purposes, the parameter generation algorithm considers a fixed number of days for each year.

The WGEN model has been thoroughly tested and documented for locations in the United States (Richardson and Wright 1984); it has also been studied for locations in Alaska (Skiles and Richardson 1998), Europe, and Asia (Semenov *et al.* 1998). Consequently, it is of interest to test the model in regions presenting extreme weather types, particularly the arid weather type. Chile, a country located on the west coast of South America, covers a wide variety of climates ranging from arid desertic to cold wet temperate climate, over a wide range in latitude. This research describes issues arising from the development of WGEN parameters for Chile with special emphasis in the arid North of Chile, and with broader value to those examining weather generation in other arid regions. Therefore this work does not focus on the development of a new stochastic model for synthetic weather generation; instead it improves the existing WGEN model by expanding its applicability to Chile while giving special emphasis to its usage in arid climates.

2. Background

Chile is located between the parallels 18° and 56° S and longitudes 68° and 76° W, having over 4200 km of length with just 240 km of average breadth. In addition, its territory includes several islands (including Easter Island) and the Antarctic territory enclosed between meridians 53° and 90°W. The country is divided into thirteen administrative regions including a metropolitan district (*Región Metropolitana*) (Figure 1). The crest-line of the Andes mountains forms the eastern border of the country. Latitude is the principal variable affecting climate with an arid desertic climate in the north and a cold wet climate in the south. Climate also changes according to elevation and distance from the coast. In coastal cities, winters are less cold and summers are less warm than in cities located in inland regions. There are several transversal valleys that run from the Andes Mountains to the sea across the country. Temperatures are more extreme in the transversal valleys than in coastal areas, and precipitation tends to be higher due to the proximity to the Andes Mountains (Sánchez Muñoz and Morales 1993). Table 1 gives annual average values for precipitation, temperature and solar radiation for the locations given in Fig. 1. These values show considerable variability between locations, particularly in the precipitation values.

Chilean climate can be divided into four classifications for the purposes of this paper (Instituto Geográfico Militar (Chile) 1998):

- Desertic without cold season
  This type of climate is common in the lowlands of northern Chile, from the border with Peru to Coquimbo (30°S). Near the coast, the weather in this region, in spite of being virtually rainless, can be cloudy and relatively cool for latitudes (eg, Arica, Iquique, Antofagasta, and La Serena) (Pearce and Smith 1993). While in the inland desert, the climate is arid and lacking in cloudiness or humidity (eg, Copiapó and Vallenar), although some locations (eg, Copiapó) may experience precipitation because of local conditions such as the proximity of high mountains.

- Mediterranean
  This type of climate characterises the territory of central Chile between latitudes 32° and 38°S (Pearce and Smith 1993), and consists of a long dry season followed by a mild and moderately wet winter. Temperatures differ significantly between seasons. At the coast, daily hours of sunshine average 2 to 3 hours in the winter and 8 to 9 hours in the summer. Inland, where there are fewer clouds, daily hours of sunshine increase by at least 1 hour during each season. Santiago and Curicó present this type of climate.

- Cool temperate climate with precipitation throughout the year
  This type of climate is present in Chile south of 38°S. Precipitation is frequent throughout the year but the amount of rain can vary depending on local factors such as elevation or proximity to the Andes. Winters are not very cold on the coast, but the summers are cool

and cloudy (Pearce and Smith 1993). Balmaceda, Concepcion, Temuco, Valdivia, Puerto Montt, and Puerto Aisén present this type of climate. Punta Arenas has a low annual rainfall because it is sheltered from the wet westerly winds by the southern Andes (Pearce and Smith 1993).

- Temperate islands
  Easter Island and the Juan Fernández archipelago have a warm-temperate climate with precipitation throughout the year.

The data for the precipitation component part of the model were obtained from the US National Oceanic and Atmospheric Administration (NOAA) and the Chilean Water Management Council (*Dirección General de Aguas).* Table 2 provides a summary of the sources for precipitation data and specific details.

To calculate temperature and solar radiation parameters, Richardson (1981) used five years of daily temperature and solar radiation data. Elsewhere, Richardson and Wright (1984) recommend at least ten years of temperature and solar radiation data. However, solar radiation data are unavailable for most of the locations in Chile, therefore the calculations had to be carried out with fewer years than recommended. Temperature data used were from the GDS CD-ROM and TD 9956 database, and solar radiation data were obtained from the World Radiation Data Centre (WRDC). Table 3 summarises the years and source for the temperature and solar radiation data.

3. Precipitation Data Quality Concerns

A mistaken large precipitation value in a database can have significant consequences in accurate weather generation for arid regions. For instance, the NOAA data files (GDS CD ROM) include an unusually large precipitation value of 701 mm for the 5th of April, 1987 at Vallenar. Not counting this value, the maximum daily precipitation value for April from 1978 to 1991 was 14 mm. Also, according to the Chilean *Instituto Nacional de Estadísticas,* the monthly precipitation value for April in 1987 was 0.1 mm (INE 1999). Therefore, the 701 mm precipitation value in the GDS CD ROM must constitute a mistake in the data file. This value was replaced with 0 mm of precipitation because this is the expected precipitation in an arid location like Vallenar. In this way, the calculated scale parameter value (β) for this month is 7mm, however if the error had not been spotted, the scale parameter (β) would have been 362 mm. Consequently, it is necessary to exhaustively screen data files for the possibility of errors when dealing with often short records of daily precipitation data available for arid regions around the world.

In addition to the problem of erroneous data, missing data can also cause problems in estimating weather-generation parameters. One approach to solving the problem of missing data is to generate weather using random and stochastic algorithms based on available data. First, the wet or dry status of the missing day can be decided using a random based algorithm that takes into account the type of weather at the desired location. After the wet or dry status of a given day has been resolved it is necessary to create a new value in the case that the missing day was a wet one. This can be done using a creation algorithm that uses past historic data (daily average precipitation per day in a given month for example) transformed by a random component. The final created values can then be checked against historic data to verify that they are within range.

Because precipitation is so rare in arid locations, any extra data can have significant consequences when generating weather parameters. Fig. 2 presents two sets of weather generation parameters for Antofagasta calculated with daily precipitation values from 1978 to 2000, which had 86 missing daily precipitation values. In this figure, set A represents parameters calculated using a data set in which the missing values have been replaced by synthetic wet or dry days. Set B represents the calculated parameters in which missing values have been replaced by zero rainfall. Set A has an excess of 31 wet days over set B (in 23 years of daily data). When comparing the results obtained from using sets A and B, the P(W/W) changes depending on the number of consecutive wet days in each data record, and the P(W/D) tends to increase when more wet days are introduced. Because the number of wet days is so small, the P(W/W) and the P(W/D) can change substantially when just a few wet days are added. Set A overestimates the amount of rainfall because the missing values which are replaced by wet days can generate unusually high results which are not representative of arid regions. For example, the extra 7 generated wet days for January in set A increases the average precipitation for this month (7 mm in set A against 1 mm in set B) which in turn generates a large value of β (15 mm in set A in contrast to 8 mm using set B) for this month. This value of β is not within the limits of what could be expected for this arid location during a particularly dry month. Therefore, in arid regions the parameter calculation procedure is sensitive both to the variations in the quality of the daily precipitation data, and to the procedures used in generating missing data.

4. WGEN Precipitation Parameter Estimates in Arid Locations

In WGEN, it is not possible to calculate weather generation parameters with fewer than two wet days in a month over the length of the record because at least two data points are required to estimate the two parameters of the Gamma distribution. Therefore, whenever there are fewer than two wet days in a month, the P (W/W) and P (W/D) are assigned a value of zero within WGEN. In this situation, the values of the scale and shape parameters (β and α) are assigned a nominal value of zero to state that they are undefined and WGEN will never generate precipitation during this month when calculating synthetic weather.

For a record of 10 years, with 10% of the days being wet days, around 30 wet days would be expected, far above the 2 needed to generate WGEN parameters. However, in arid locations, a 10 year record with one wet day in a given month only once every 10 years, only one wet day would be expected. One data point would result in WGEN parameters of zero, meaning that even a 100 year simulation would predict no precipitation during that month for all 100 years. This inadequacy of the WGEN system becomes more severe as the number of years of data available for parameter estimation decreases or as the frequency of precipitation decreases.

Even when there are two or more wet days in a given month over the length of the data record, there can be difficulties in estimating suitable WGEN parameters. The shape parameter (α_j) for a given month j is calculated in the WGEN system using the maximum likelihood estimator ($\hat{\alpha}_j$) described by Greenwood and Durand (Haan 1977):

$$\hat{\alpha}_j = \frac{8.898919 + 9.05995 \cdot y + 0.9775373 \cdot y^2}{y \cdot (17.79728 + 11.968477 \cdot y + y^2)}, \qquad 0.5772 < y \le 17.0 \tag{11a}$$

$$\hat{\alpha}_j = \frac{0.5000876 + 0.1648852 \cdot y - 0.0544274 \cdot y^2}{y}, \qquad 0 \le y \le 0.5772 \tag{11b}$$

$$if\ \hat{\alpha}_j \ge 1 \Rightarrow \hat{\alpha}_j = 0.998 \tag{11c}$$

Where y is a function of the precipitation amount in a given month as stated in equations 12, 13, and 14. In this set of equations, r_{ji} is the precipitation amount for wet day i on month j, and n is the number of rainy days in month j.

$$\bar{r}_j = \sum_{i=1}^{n} \frac{r_{ji}}{n} \tag{12}$$

$$\overline{\ln r_j} = \frac{\sum_{i=1}^{n} \ln r_{ji}}{n} \tag{13}$$

$$y = \ln \bar{r}_j - \overline{\ln r_j} \tag{14}$$

The moment estimator for the scale parameter, $\hat{\beta}$, is calculated in the WGEN system using equation 15.

$$\hat{\beta}_j = \frac{\bar{r}_j}{\hat{\alpha}_j} \tag{15}$$

Once these values have been calculated, it is possible to estimate the scale and shape parameters for each month by making $\alpha_j \approx \hat{\alpha}_j$ and $\beta_j \approx \hat{\beta}_j$.

In arid months y tends to be zero and the resulting shape parameter estimator, $\hat{\alpha}_j$, can sometimes be a very large number (the denominator in equations 11a and 11b tends to zero). The distribution function resulting from values of $\hat{\alpha}$ >>1 is skewed so that it can generate many low rainfall events. A distribution function with $\hat{\alpha}$ =1 would give the same result and with less computational effort than when using $\hat{\alpha}$ >>1. Therefore, for values of $\hat{\alpha}$ >1, $\hat{\alpha}$ = 1 is used instead (in the actual algorithm 0.998 is used rather than using 1 for computational purposes).

In a given location, if a given precipitation record has fewer than 2 wet days in a month then the corresponding shape parameter, α, is undefined for that month. However, as more years of data are added to the weather record, the previously rainless month may now have more than 2 wet days due to the new data contribution. As a result, the estimated shape parameter for arid zone climates tends to cluster around either 0 or 1. Similarly, scale parameters tend to be estimated as zero when there are fewer than two rainy days. Those who generate WGEN parameters in arid climates need to appreciate that these trends in estimated parameters result from the parameter generation process rather than the climate.

5. Length of Precipitation Records

The adequacy of the length of a precipitation record will be difficult to judge without a long record of data for analysis. One way to help judge this effect is, for a given location, to calculate synthetic precipitation parameters using a historic data record, and then use these parameters to generate a synthetic data record. One year of the random process for the weather will be either very similar or very different from long-term averages by chance, while an infinite number of simulated years would have an average very similar to the historic averages. In this paper, 30 simulated years are used to generate estimated precipitation averages to help see the potential for problems with inadequate precipitation records at different locations. In arid locations, the differences between historic and synthetically generated precipitation data can be significant in comparison to wet locations (Fig. 3). Differences for Antofagasta, for example, amount to about 42% because the historic mean annual precipitation value is 5.12 mm, whereas the calculated synthetic mean annual precipitation value for this location is 2.95 mm. Therefore, differences for Antofagasta and Arica (Fig. 3) suggest that it is more difficult to calculate adequate synthetic precipitation parameters for arid regions than for wet locations when using the WGEN model. A similar problem arises when generating dry-season weather in semi-arid locations.

Another way of considering the adequacy of precipitation data record lengths for arid locations is through bootstrapping. This technique has successfully been used in determining the accuracy of various prediction models (Rao 2000). Bootstrapping in this situation means subdividing the data set and evaluating the weather parameters as if the subset were the only available data. In this case, after screening for outliers and substituting missing data with generated values, the precipitation record was subdivided into 4 groups of 5 years, 3 groups of 7 years, and 2 groups of 10 years. These individual groups (including the whole record of 21 years) were then used to calculate the synthetic precipitation parameters. Fig. 4 shows the calculated parameters for a wet month (June) in Santiago. With five or seven years of data there is a significant variation in the precipitation parameters calculations. However, this variation decreases between the 10 year pairs. Although the estimated parameters are not independent between series lengths, the decrease in variability between groups with an increase in data length gives some confidence that the length of the data set is adequate for estimating WGEN parameters.

For arid months, however, the situation is different. Fig. 5 shows the precipitation parameters calculated for the same period lengths as before, but for a dry month in Santiago (January). In this case, variations in the precipitation parameters calculated using different record lengths are significant. As the length of the precipitation record increases, these variations decrease, but some of the values tend to persist with no or little variation. For the shape parameter, α, this variation appears to be either zero (undefined) or 1 from the start, and with a final value of 1 after 21 years. This situation takes place because some short records of precipitation data have fewer than 2 wet days in a month, and thus the corresponding shape parameter is undefined for that month. When these records are merged together, the previously dry months may now have more than 2 wet days. However, these records still have a small number of wet days which produce a shape parameter of 1. In a particular month, when α and ß are undefined, the P(W/D) = P(W/W) = 0 and the synthetically generated precipitation data for that month will in turn be zero.

Making data fit a gamma distribution can be complicated in arid regions and in the dry season at semiarid locations. Fig. 6 shows how precipitation data (1978 - 1998) are fitted to different gamma distributions throughout the year for Santiago. Gamma distributions are calculated using their corresponding scale and shape parameters. Each gamma distribution is then plotted for each month along with the precipitation frequency (calculated from historic data) for each 5 mm interval. For wet months, daily precipitation can have a wide range of values having high mean values (mean over 10mm) and maximum values of over 80mm. Dry months have a low mean precipitation value (≤ 7 mm) and their maximum precipitation per day is also low (≤ 30 mm). Most importantly, these months have a low frequency of precipitation which is characterised by only a few points in the gamma distribution graphs. The gamma distribution for dry months tends to zero quite rapidly (generally in the 25 mm interval) and has a skewed appearance. This is consistent with how the gamma distribution should look for arid months so that it is possible to generate low amounts of rainfall. However, small amounts of precipitation and a very low frequency can make it difficult to generate appropriate gamma distributions (eg,, February in Santiago). The fit parameters in arid situations are highly sensitive to the available rainfall data; therefore, if more data were available, the gamma distribution parameters could change considerably. On the other hand, for wet months (May thru August) there are more data points available to generate the gamma distribution. A large number of points makes it possible to describe the gamma distribution appropriately providing a better fit.

6. Temperature and solar radiation

Temperature and solar radiation take place continuously throughout time, whereas precipitation is an intermittent event that happens on certain days. This makes it easier to deal with any problems regarding the quality of the original temperature and solar radiation data sets. However, as with precipitation data, temperature and solar radiation data need to be checked for mistakes, and missing values need to be corrected before using these data. For this study, missing values were estimated by simple interpolation considering the average of the previous five days and the average of the five days following the missing day. This procedure was carried out for the maximum temperature, minimum temperature, and solar radiation data. In addition, suspicious values were detected using a visual inspection method. With regards to the quantity of data, in some cases it was impossible to collect more than 3 years of data on which to carry out the calculations.

In WGEN, when there are fewer than 2 wet days in a given month over the length of the data record , the parameter calculation procedure devised by Richardson and Wright (1984) fails to generate the means of maximum temperature (TXMW) and solar radiation (RMW) for wet days. When the number of wet days in a season is less than or equal to 2, the mean, standard deviation, and coefficient of variation for this season are automatically assigned a value of 0°F. The final values of TXMW and RMW are calculated by averaging the maximum temperature and solar radiation means for the 13 seasons throughout the year. Therefore, if one of these seasons has arbitrarily been assigned a value of 0°F, a bias is introduced and the final values of TXMW and RMW are erroneous. In actual fact, when there are 1 or 2 wet days in a given data record there certainly is enough information to determine the corresponding mean and standard deviations (i.e. if there is only 1 wet day, its mean would be this same amount, and with 2 days it is possible to calculate the mean and standard deviation). The only instance in which there is not enough information to estimate the means (maximum temperature or solar radiation) is when there are no wet days in a season.

To solve this problem the original algorithm can be corrected so that it assigns a value of zero to the means only when there are no wet days in a season. Also, this algorithm should assign a value of zero to the standard deviation only when there are 1 or 0 wet days. Table 4 shows that if this correction is not taken into consideration, the resulting TXMW and RMW values can have errors that amount to up to 71%.

After generating the annual temperature and solar radiation parameters taking into account the seasonal change in the means and coefficients of variation, it is necessary to calculate the residual elements. To do this, matrices *A* and *B* should be calculated for each station rather than by using the matrices embedded in the original model, which uses US values. A calculation of *A* and *B* matrices for all locations becomes more important for shorter data records where imprecision is already high. Matrices *A* and *B* for seven chilean locations presenting climates ranging from arid to wet are presented in Tables 5 and 6.

Richardson and Wright (1984) showed that the correlation coefficients for daily temperature and solar radiation were not season-dependent so they averaged them for each station. Following Richardson's procedure, the different seasonal correlation coefficients for Chile were then averaged for each location. The lag 0 cross correlation coefficients (Table 7) vary considerably from station to station (average standard deviation of 23%). In addition, the average lag 0 cross correlation coefficient between maximum and minimum temperatures (ρ_0 (1,2)) is 0.225, which is significantly different to the one used by the U.S. WGEN model (0.633).

Table 8 shows the calculated values for the lag 1 serial correlation and lag 1 cross correlation coefficients. The serial correlation coefficients have significant variations throughout the different stations; the highest variations are in the serial correlation coefficients for maximum temperature (8% std. deviation) and solar radiation (13.7% std. deviation). The average values for the serial correlation coefficients for maximum and minimum temperature (0.417 and 0.377) were significantly different to the ones used in the U.S. WGEN model (0.621 and 0.674). The $\rho_1(1,3)$ coefficients experienced differences throughout the different locations (std. deviation 9.2%), but more significant were the results for $\rho_1(2,1)$ and the results for $\rho_1(3,2)$ (std. deviations 14.3% and 17.8% respectively). The averages calculated for the lag 1 cross correlation coefficients for $\rho_1(1,2)$, $\rho_1(2,3)$ and $\rho_1(2,1)$ are considerably different to the ones used in the U.S. WGEN model (Table 8).

7. Conclusions

This paper presents calibration results that expand the applicability of the existing WGEN model to Chilean climate. This involved dealing with location-dependant data sets and modifying the model to customise it to specific locations in Chile. Although this work does not result in a new model, it adds value to the original WGEN model by expanding its applicability and accuracy. In addition, this paper has examined the dangers involved in estimating weather generation parameters for Chile. Many of these dangers would apply in other locations, especially in arid regions. The estimation of precipitation parameters is sensitive to erroneous and missing data. Researchers need to consider the use of methods for filtering data for errors, and consider methods for generating representative data to replace missing data that can allow for the generation of unbiased estimation of weather parameters.

The WGEN method for the generation of parameters for the WGEN model has deficiencies when applied to arid climates. If there is only one wet day in a given month over the length of the data record, the model will assume no rainfall will ever occur in that month. This deficiency becomes more severe when the data record is shorter and the frequency of rain is lower. For monthly data records with few rainfall events, the gamma distribution is difficult to fit to the data and the WGEN estimation method will limit the shape parameter for the gamma distribution to a maximum of 1. In some cases, this could lead to a poor fit between generated and actual precipitation depths. In addition, the generation of constant shape parameters for multiple months might lead the researcher to incorrectly assume that the climate is constant over these months, when in fact, the constant shape parameters result from a simplification in the parameter generation procedure.

This paper is unable to add to the advice of Richardson (2000) that at least 20 years of daily precipitation data are needed for accurate WGEN parameter estimates. The bootstrapping method in this paper relies on only 20 years of data and shows there is little confidence in the accuracy of parameters based on this number of years of data, but the results cannot help answer how many years of data would be needed. It seems evident that longer data records will be needed for drier climates; however, it might be more helpful for researchers to focus on the total number of wet days per month rather than the total length of the data record when assessing the adequacy of the dataset in estimating parameters. Chilean climate has numerous locations where there are very dry months, even though the average annual rainfall might indicate a non-arid climate. In these cases, twenty years of daily data would be adequate for wet months, but not adequate for dry months.

The WGEN technique for generating temperature and solar radiation can also give improper results for arid climates. This algorithm will arbitrarily assign a default value of zero to seasonal values when there are fewer than 2 wet days in a season. These default values in turn are used to estimate annual average seasonal values and a bias is introduced.

Although the WGEN program is commonly used and is a valuable research tool, there is a need to recognise its limitations. The specific limitations that can arise in arid climates have been highlighted by research into appropriate WGEN parameters for Chilean locations. It is hoped that this will assist future researchers who look to generate weather in Chile or other arid locations.

Acknowledgments

A number of different organisations have supplied the necessary weather data to carry out the calculations at special prices and sometimes free of charge. These organisations deserve particular acknowledgments, especially NOAA (US), WRDC (WMO, Russia), FAO, Dirección General de Aguas (MOP, Chile), and the Instituto Nacional de Estadísticas (Chile). The authors also wish to thank Dr. Stuart Larsen (Lincoln University) for his invaluable help in retrieving the weather data and his advice on how to deal with missing values. Thanks to Dr. Crile Doscher for helping to decode solar radiation data from FAO.

References

Bond, D. C. 1979. Generating daily weather values by computer simulation techniques, for crop yield forecasting models. U.S. Department of Agriculture, Economic Statistics and Cooperative Service, Statistical Research Division:38.

Bruhn, J. A., Fry W.E., and Fick. G.W. 1980. Simulations of daily weather data using theoretical probability distributions. Journal of Applied Meteorology 19:1029-1036.

FAO. 1985. Agroclimatological Data for Latin America and the Caribbean. Food and Agriculture Organization of the United Nations, Rome.

Haan, C. T. 1977. Statistical methods in hydrology, 1st edition. Iowa State University Press, Ames.

INE. 1999. Estadísticas del Medio Ambiente. Instituto Nacional de Estadísticas, Santiago.

Instituto Geográfico Militar (Chile). 1998. Atlas geográfico de Chile para la educación. *in*. Instituto Geográfico Militar, Santiago, Chile.

Matalas, N. C. 1967. Mathematical Assessment of Synthetic Hydrology. Water Resources Research 3:937-945.

Nicks, A. D., and J. F. Harp. 1980. Stochastic generation of temperature and solar radiation data. Journal of Hydrology 48:1-17.

Pearce, E. A., and C. G. Smith. 1993. The world weather guide, 3rd . edition. Helicon, Oxford.

Rao, J. S. 2000. Bootstrapping to assess and improve atmospheric prediction models. Data Mining and Knowledge Discovery 4:29-41.

Richardson, C. W. 1981. Stochastic Simulation of Daily Precipitation, Temperature, and Solar Radiation. Water Resources Research 17:182-190.

Richardson, C. W. 1982a. A comparison of three distributions for the generation of daily rainfall amounts. In Singh, V.P. (ed.), Statistical Analysis of Rainfall and Runoff, pp. 67-78, Proc. Int. Symp. On Rainfall-Runoff Modeling, Water Resources Pubilications:700.

Richardson, C. W. 1982b. Dependence structure of daily temperature and solar radiation. Transactions of the ASAE 25:735-739.

Richardson, C. W. 2000. Data requirements for estimation of weather generation parameters. Transactions of the ASAE 43:877-882.

Richardson, C. W., and D. A. Wright. 1984. WGEN: A model for generating daily weather variables. U.S. department of agriculture, Agricultural Research Service ARS-8:83.

Sánchez Muñoz, A., and R. Morales. 1993. Las regiones de Chile : espacio físico y humano-económico, 1. edition. Editorial Universitaria, Santiago de Chile.

Semenov, M. A., R. J. Brooks, E. M. Barrow, and C. W. Richardson. 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10:95-107.

Skiles, J. W., and C. W. Richardson. 1998. A stochastic weather generation model for Alaska. Ecological Modelling 110:211-232.

Wilks, D. S., and R. L. Wilby. 1999. The weather generation game: a review of stochastic weather models. Progress in Physical Geography 23:329-357.

Appendix –Tables and Figures

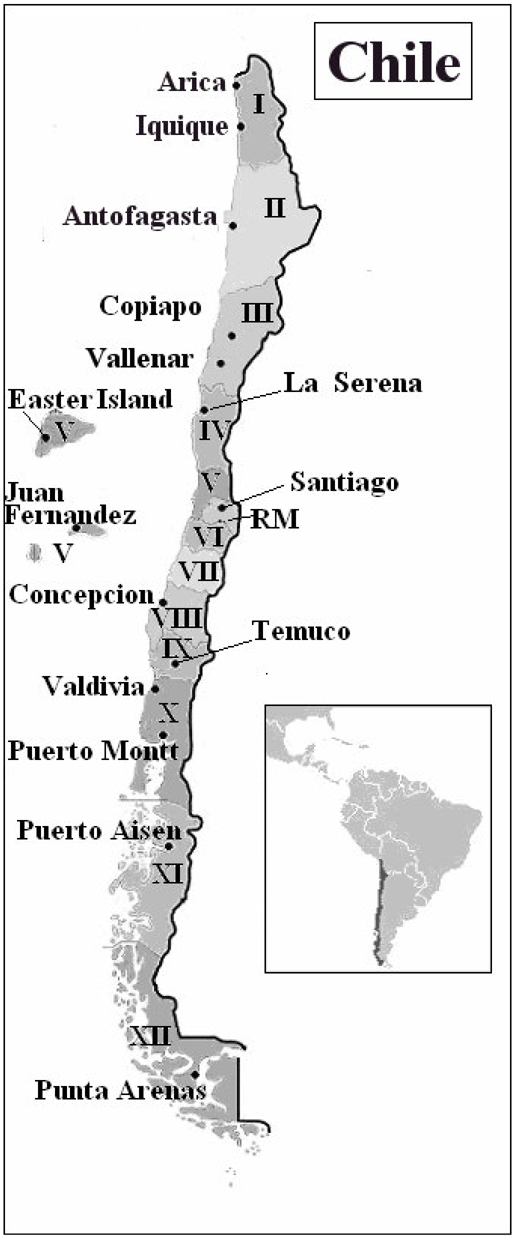

Fig. 1 Locations in Chile for which FAO climatic data is presented. Chile's administrative regions are indicated in the map (I-XII & RM).

Table 1
Annual climatic data for locations in Chile

Location (region)	Av. Prec. [mm]	Av. Temp. °C	Solar Radiation cal/cm^2
Arica (I)	0	18.8	372
Iquique (I)	2	17.9	424
Antofagasta (II)	7	16.5	423
La Serena (IV)	78	14.9	340
Easter Isl. (V)	1091	20.4	389
Santiago (RM)	335	14.2	310
Juan F. (V)	998	15.3	297
Concepción (VIII)	1294	13.1	327
Temuco (IX)	1191	12.0	290
Valdivia (X)	2540	11.9	323
P. Montt (X)	1982	11.0	265
P. Aisén (XI)	2941	9.0	186
P. Arenas (XII)	400	6.7	195

Source: (FAO 1985)

Table 2.
Source of data for weather stations

LOCATION	GDS			TD9956			DGA			Tot. Years
	Start	Finish	Yrs	Start	Finish	Yrs	Start	Finish	Yrs	Yrs
Santiago / Pudahuel	1978	1991	14	1992	1998	7				21
La Serena / La Florida	1978	1991	14	1994-	1998 & 2000	6	1992	1993	2	22
Arica / Chacalluta	1978	1991	14	1992	2000	9	1975	1977	3	26
Antofagasta	1978	1991	14	1992	2000	9				23
Vallenar	1978	1991	14							14
Copiapó	1978	1991	14				92 & 94-	1998	6	20
Temuco	1978	1991	14	1994-	1998 & 2000	6	1992		1	21
Puerto Montt / El Tepual	1978	1991	14	1992	1998	7				21
Punta Arenas / Chabunco	1978	1991	14	1994	1998	5	1992	1993	2	21
Easter Island	1978	1991	14	1994-	1998 & 2000	6				20
Curicó	1978	1991	14	1995	1998	4	1992	1993	2	20
Concepción	1978	1991	14	1992	1998	7				21
Balmaceda	1978	1991	14	1994-	1998 & 2000	6				20
Valdivia	1978	1991	14	1994	1998	5	1992	1993	2	21
Juan Fernández	1978	1991	14	1994-	1998 & 2000	6				20
Puerto Aisén	1978	1991	14							14
Chañaral	1978	1991	14	1994	1999	6				20
Iquique	1985	1991	7	1994	2000	7	1992	1993	2	16

GDS = Global Daily Summary CD ROM purchased from NOAA
D9956 data base = data available fom NOAA
DGA = *Dirección General de Aguas* (Chilean Water Management Council).

Table 3.
Number of years and sources of data for Temperature and Solar Radiation

Locations	Maximum and minimum temperature							Solar radiation		
	GDS CD ROM			TD 9956				WRDC		
	Start	Finish	Yrs	Start	Finish	Yrs	Total yrs	Start	Finish	Yrs
SANTIAGO/PUDAHUEL	1988	1991	4	1992	1993	2	6	1998	1993	6
LA SERENA/LA FLORIDA	1978	1980	3				3	1978	1980	3
ANTOFAGASTA	1978	1980	3	1992	1993	2	8	1978	1980	8
	1988	1990	3					1988	1990	
								1992	1993	
VALLENAR	1988	1990	3				3	1988	1990	3
CONCEPCION	1978	1980	3	1992	1993	2	5	1978	1980	5
TEMUCO	1978	1980	3				3	1978	1980	3
PUNTA ARENAS/CHABUNCO	1988	1990	3				3	1988	1990	3
EASTER ISLAND	1988	1990	3				3	1988	1990	3

GDS = Global Daily Summary CD ROM purchased from NOAA
TD9956 data base = data available fom NOAA
WRDC = World Radiation Data Center

Table 4.
Temperature and solar radiation parameter correction

Location	TXMW (°F)	RMW (Ly)
Antofagasta		
Original WGEN value	20.8	110.5
Corrected value	67.8	385.5
Error	69%	71%
Vallenar		
Original WGEN value	39.9	211.5
Corrected value	64.1	315.9
Error	38%	33%
La Serena		
OriginalWGEN value	24	90.9
Corrected value	63.2	265.5
Error	62%	66%
Santiago		
Original WGEN value	41.4	177.8
Corrected value	58.9	221
Error	30%	20%

Note: Data are presented in °F and Ly because the original algorithm works with these units.

Table 5.
Matrix *A* for different locations in Chile

Location	a_{11}	a_{12}	a_{13}	a_{21}	a_{22}	a_{23}	a_{31}	a_{32}	a_{33}
Antofagasta	0.439	-0.114	-0.039	0.137	0.329	0.035	-0.014	0.064	0.534
Concepción	0.395	-0.054	-0.104	-0.036	0.552	0.229	-0.032	-0.260	0.215
Easter I.	0.606	-0.173	-0.124	0.039	0.343	0.044	-0.044	0.067	0.264
P. Arenas	0.501	-0.188	0.008	0.191	0.202	0.036	0.068	-0.157	0.085
Santiago	0.570	0.068	-0.114	-0.213	0.518	0.286	-0.311	-0.040	0.542
Temuco	0.464	-0.118	-0.116	-0.117	0.604	0.268	0.093	-0.326	0.128
Vallenar	0.354	-0.131	-0.023	0.132	0.245	-0.080	-0.160	0.097	0.245

Table 6.
Matrix *B* for different locations in Chile

Location	b_{11}	b_{12}	b_{13}	b_{21}	b_{22}	b_{23}	b_{31}	b_{32}	b_{33}
Antofagasta	0.908	0.000	0.000	0.272	0.879	0.000	0.012	-0.125	0.837
Concepción	0.928	0.000	0.000	0.054	0.870	0.000	0.329	-0.385	0.764
Easter I.	0.828	0.000	0.000	0.344	0.869	0.000	0.227	-0.262	0.902
P. Arenas	0.894	0.000	0.000	0.408	0.849	0.000	0.049	-0.161	0.970
Santiago	0.862	0.000	0.000	-0.150	0.852	0.000	0.750	-0.298	0.385
Temuco	0.895	0.000	0.000	0.246	0.829	0.000	0.062	-0.461	0.786

Vallenar	0.951	0.000	0.000	0.415	0.855	0.000	0.585	-0.182	0.763

Table 7.
Lag 0 cross correlation coefficients

Location	ρ_0 (1,2)	ρ_0 (1,3)	ρ_0 (2,3)
Antofagasta	0.309	-0.012	-0.083
Vallenar	0.424	0.552	0.089
Easter I.	0.318	0.167	-0.131
Santiago	-0.166	0.606	-0.362
Concepción	0.053	0.299	-0.441
Temuco	0.182	0.085	-0.517
P. Arenas	0.455	0.072	-0.014
Average	0.225	0.253	-0.208
sd deviation	0.221	0.243	0.231
WGEN (US values)	0.633	0.186	-0.193

Notes:
ρ_0 = lag 0 cross correlation coefficient.
1 = maximum temperature
2 = minimum temperature
3 = solar radiation

Table 8. Lag 1 serial correlation and cross correlation coefficients

Location	ρ_1 (1)	ρ_1 (2)	ρ_1 (3)	ρ_1 (1,2)	ρ_1 (1,3)	ρ_1 (2,3)	ρ_1 (2,1)	ρ_1 (3,1)	ρ_1 (3,2)
Antofagasta	0.404	0.368	0.529	0.025	-0.035	0.006	0.238	-0.001	0.015
Vallenar	0.286	0.294	0.165	0.017	0.161	0.015	0.192	0.016	0.051
Easter I.	0.530	0.350	0.248	0.036	0.000	0.006	0.156	0.022	0.019
Santiago	0.490	0.450	0.368	0.015	0.207	-0.030	-0.125	0.024	-0.185
Concepción	0.361	0.449	0.320	0.013	0.038	-0.025	0.062	0.019	-0.356
Temuco	0.433	0.444	0.305	0.027	-0.016	-0.054	0.016	0.044	-0.376
P. Arenas	0.416	0.284	0.112	0.039	0.070	0.022	0.286	0.003	-0.138
Average	0.417	0.377	0.292	0.025	0.061	-0.009	0.118	0.018	-0.139
st deviation	0.080	0.072	0.137	0.010	0.092	0.028	0.143	0.015	0.178
WGEN (US val)	0.621	0.674	0.251	0.445	0.087	-0.100	0.563	0.015	-0.091

Notes:
ρ_1 = lag 1 serial or cross correlation coefficient.
1 = maximum temperature
2 = minimum temperature
3 = solar radiation

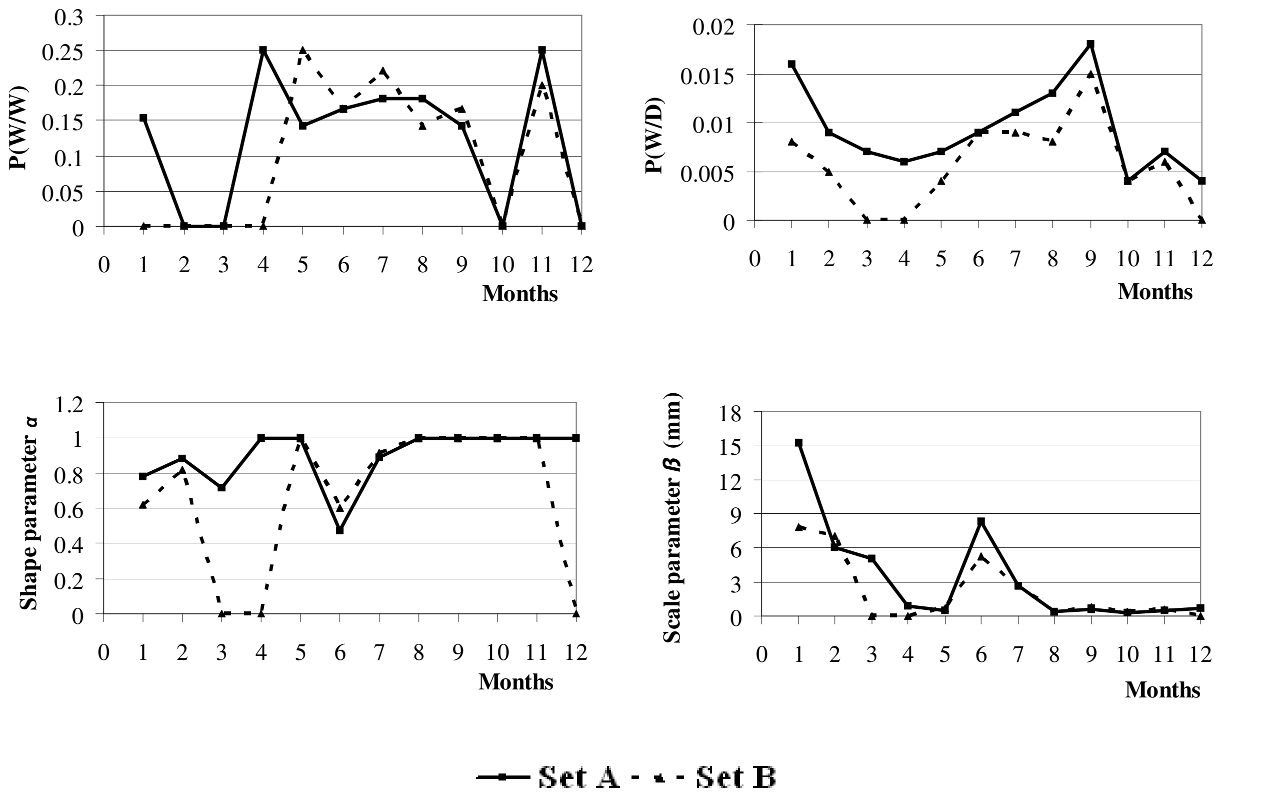

Fig. 2 Precipitation parameters for Antofagasta calculated using two methods to treat missing precipitation data (Set A = use of created precipitation; Set B = use of zero rainfall for all missing days).

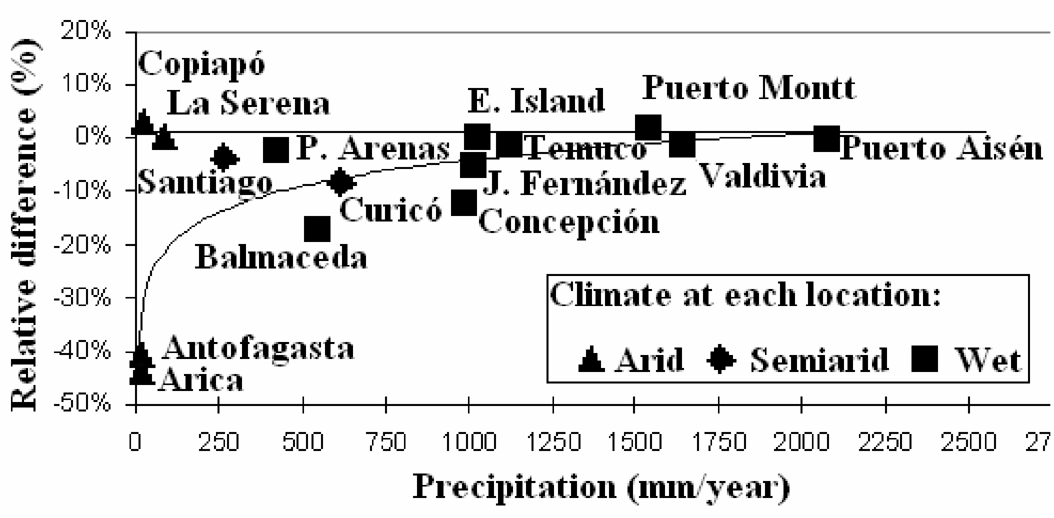

Fig. 3. Differences between historic and synthetically generated average annual precipitation rates for Chilean locations

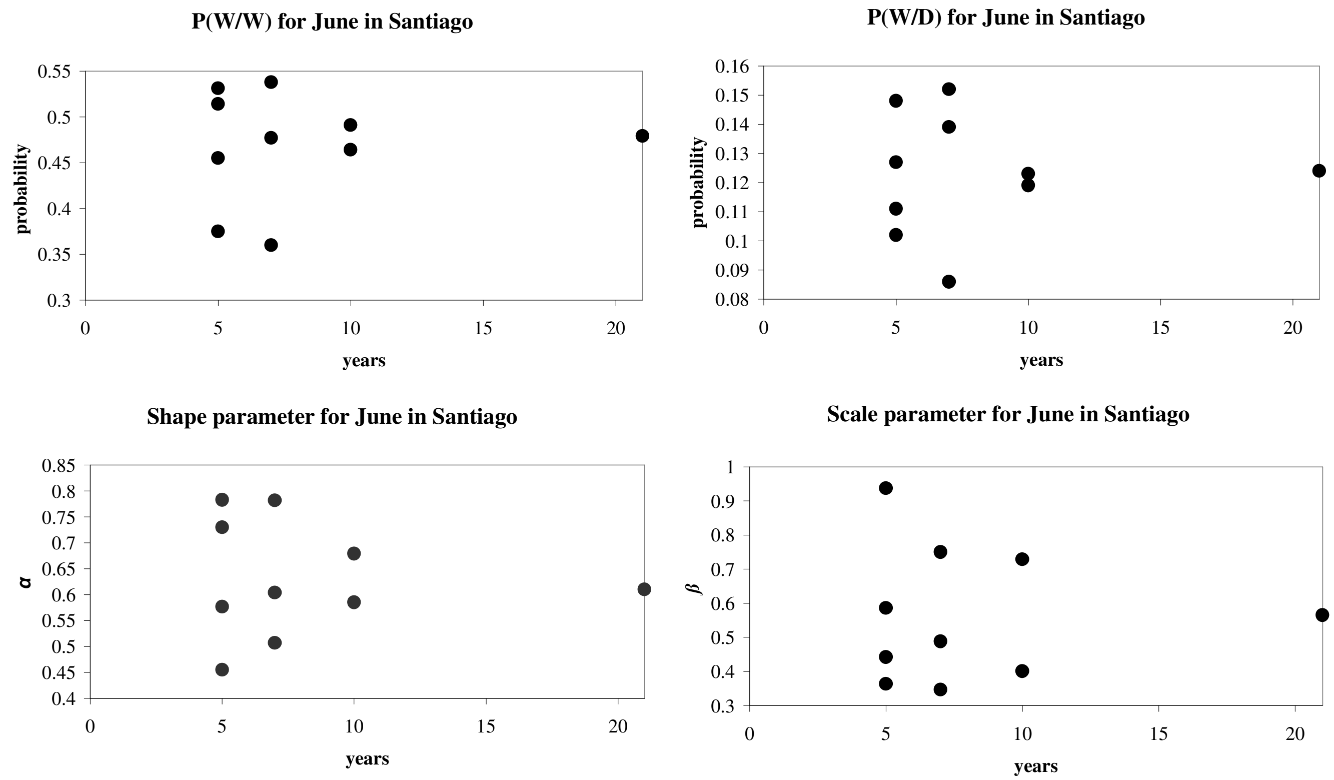

Figure 4. Estimates of synthetic precipitation parameters with period lengths of 5 to 21 years for a wet month in Santiago.

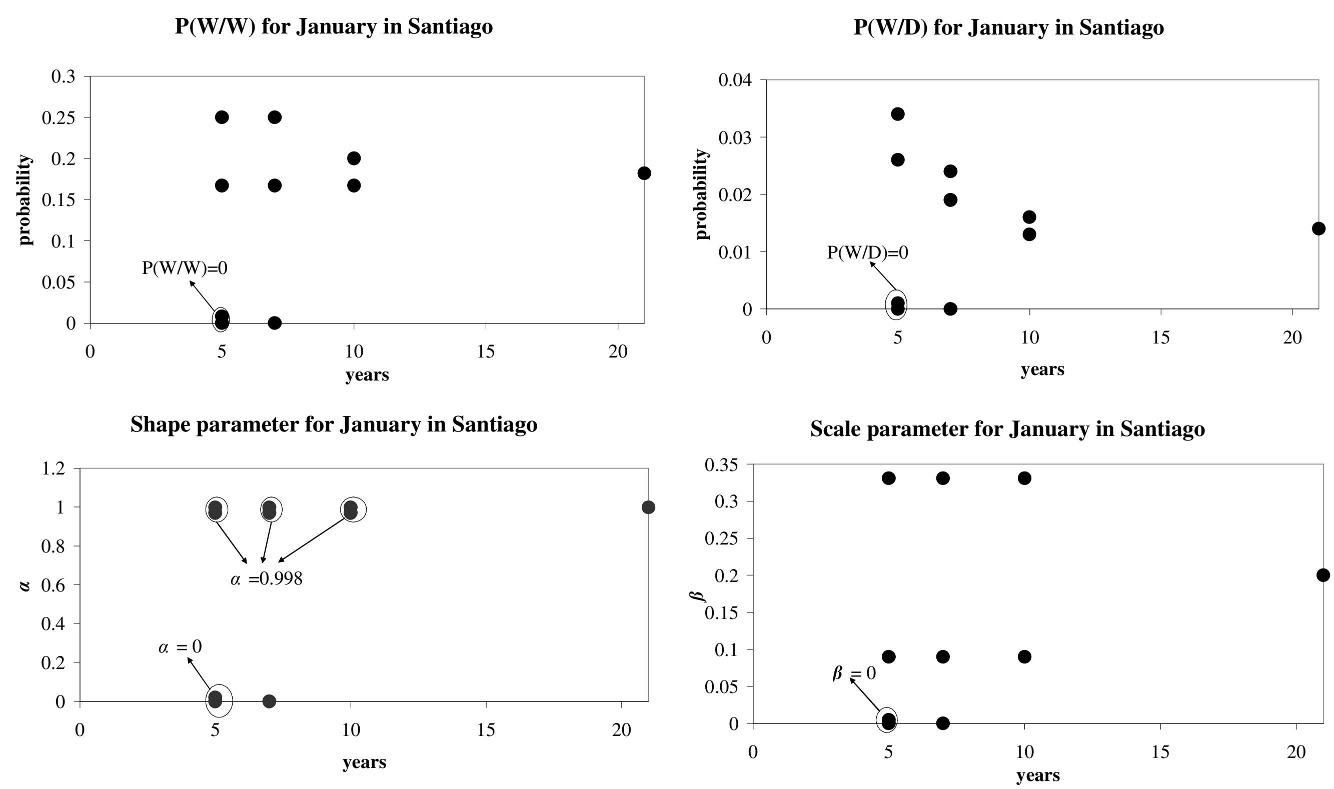

Figure 5. Estimates of synthetic precipitation parameters with period lengths of 5 to 21 years for a dry month (January) in Santiago.

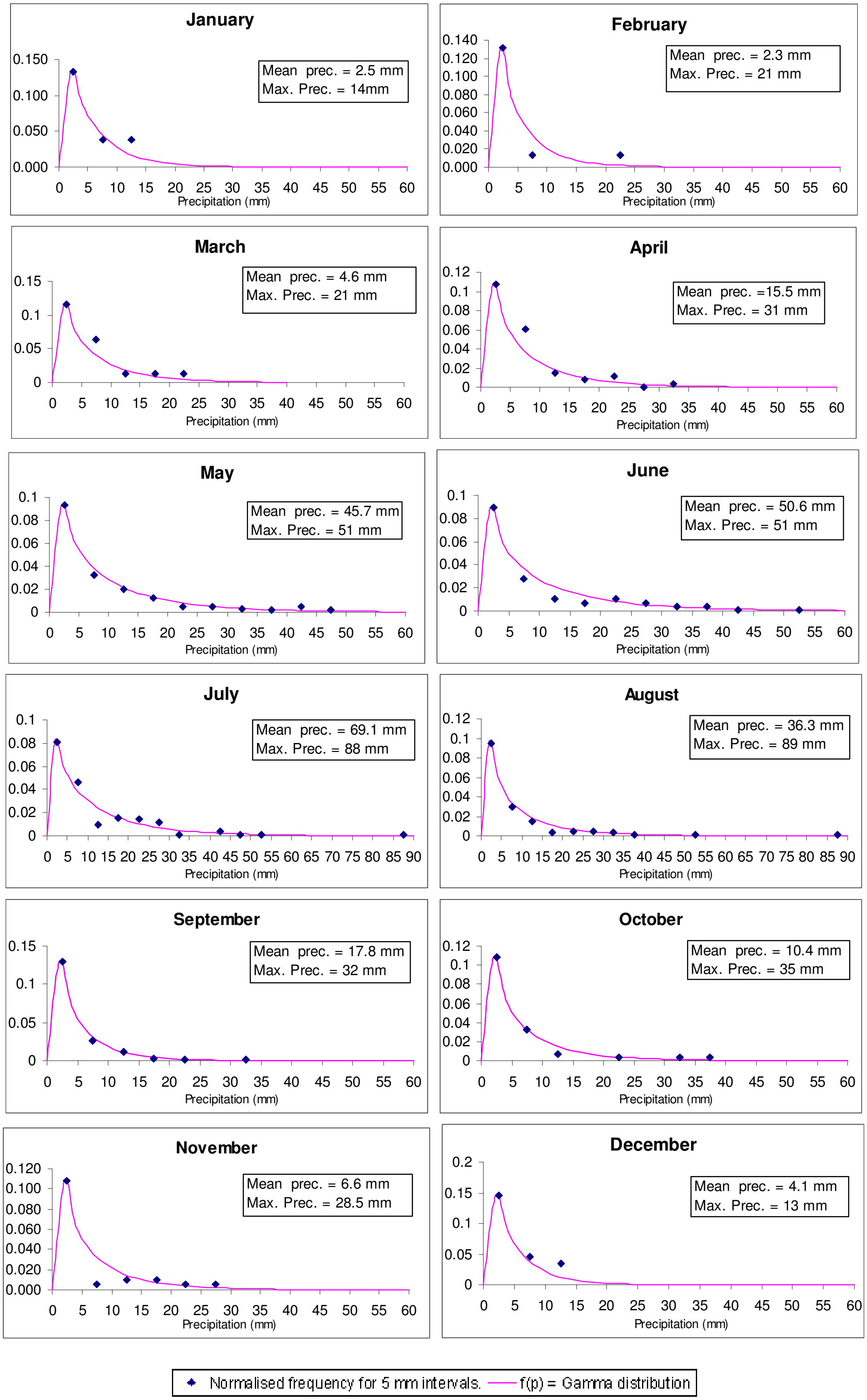

Figure 6. Gamma distribution fit for precipitation data in Santiago (the precipitation frequency is normalised against the maximum value of the gamma distribution to aid in visual comparison plot them in the same graph).