Available online at www.sciencedirect.com
ScienceDirect
Advances in Water Resources xxx (2007) xxx–xxx
Advances in Water Resources
www.elsevier.com/locate/advwatres

Accuracy evaluation of weather data generation and disaggregation methods at finer timescales

Bekele Debele [a,*], R. Srinivasan [b], J. Yves Parlange [a]

[a] Cornell University, Biological and Environmental Engineering Department, Ithaca, NY 14853, USA
[b] Texas A&M University, Spatial Sciences Laboratory, College Station, TX 77845, USA

Received 19 May 2006; received in revised form 30 August 2006; accepted 19 November 2006

Abstract

Availability of weather data at finer timescales such as hourly is vital in the application of dynamic physical and biological models. In this study, we have examined the suitability of various approaches (deterministic periodic versus stochastic) of disaggregating daily weather data into hourly data in the Cedar Creek watershed, TX, USA. We found the cosine function suitable to disaggregate daily maximum and minimum temperatures and wind speed data into respective hourly data. We also used a common logarithmic equation to compute vapor pressures from temperature data, and hence relative humidity (the ratio between actual and saturated vapor pressures multiplied by 100). Disaggregation following uniform distribution of daily rainfall over 24 h did not reproduce most statistical parameters computed from observed hourly rainfall data onsite. Conversely, both stochastic models formulated based on univariate (Hyetos) and multivariate (MuDRain) processes mimicked the measured hourly rainfall distributions very well. Overall, we found the MuDRain model superior, compared to other models to disaggregate daily rainfall data into hourly data.

Keywords: Disaggregation; Hyetos; MUDRain; Multivariate; Rainfall; Univariate

1. Introduction

Applications of dynamic models of physical and biological systems have been on the rise [1–3]. Most of these models require different weather data as state variables. For instance, simulations of plant physiological processes under field conditions and stream water quality processes (e.g., chlorophyll a, dissolved oxygen) should involve time varying temperature data as an input [1]. Similarly, more realistic and process-based hydrological and water quality models require their input variables on detailed time intervals than are currently available at most places [3]. Often, the only input data available are daily maxima and minima for temperature, daily totals for rainfall, and daily averages for solar radiation, wind speed and relative humidity. The effect of temporal and spatial variations in temperature, solar radiation, wind speed, and relative humidity on evapotranspiration calculation (a large component in the hydrological balance); and on algae growth, dissolved oxygen concentration, and the rate of chemical reaction that subsequently affects water quality simulations cannot be undermined [3,4]. Similarly, there is a significant effect of temporal and spatial variability in rainfall data on runoff and water quality simulations [5,6]. Thus, we initiated this study with the following objectives: (1) to assess and evaluate available daily weather data disaggregation methods, and (2) to develop appropriate methodologies to disaggregate daily weather data into hourly data, and examine their applicability in our study area and implications on subsequent hydrological and water quality models.

Efforts have been put into disaggregating weather data temporally (e.g., monthly to daily, daily to sub-daily, etc.)

* Corresponding author. Present address: 8750 Georgia AVe #802B, Silver Spring, MD 20910, USA. Tel.: +1 607 342 4870; fax: +1 301 589 3719.
E-mail addresses: bd58@cornell.edu (B. Debele), r-srinivasan@tamu.edu (R. Srinivasan), jp58@cornell.edu (J. Yves Parlange).

0309-1708/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2006.11.009

and spatially (e.g., from point measurements to aerial distributions), adopting different approaches [3,7–9]. Most often, practitioners tend to distribute daily available weather data to sub-daily data assuming uniform distributions [6]. Yet, numerous works in related areas have shown that such assumptions of uniform weather data distributions could be far from accurate representation of the reality [2,6]. Some have developed functional relationships between time of a day and some weather data (e.g., temperature and solar radiation). For example, Baker et al. [1] have compared the performances of three widely used methods of estimating the time dependency of air temperature (the sine/exponential model of Parton and Logan [10], the sinusoidal model of de Wit [11], and a linear model similar to that of Sanders [12]) using daily maxima and minima. Their study indicates that the results from all the models were, on average, comparable. However, in the application of specific process-oriented models, such as photosynthesis and respiration which are also the causes and effects of dissolved oxygen production and depletion in waterbodies, respectively, they recommend the sinusoidal model of de Wit [11] because it better reproduces the important criterion (the midday clear summer days' temperatures).

Similarly, Green and Kozek [2] have developed approximate regression equations using M-functional as quantiles of probability distributions to describe the hetroscedastic nature of weather data. They have used polynomial functions to describe radiation data; cosine functions to explain wet and dry bulb temperatures, soil temperature, soil heat flux and relative humidity; and maximum of constant and cosine functions to depict the wind speed data. The polynomial equations were assumed when variation was mostly confined to the daylight hours, as is the case with solar radiation data (Eq. (1)). Cosine functions, on the other hand, were used where variations occurred over a 24 h period and in a roughly periodic pattern, (e.g., in the case of temperature and wind speed data – Eqs. (2) and (3)). The following are sample equations (after Baker et al. [1] and Green and Kozek [2]):

1. polynomial function:

$$Y = \begin{cases} a((t - t_{SR})(t_{SS} - t))^b + \varepsilon & \text{for daylight hours, and} \\ \varepsilon & \text{otherwise} \end{cases} \quad (1)$$

2. cosine function:

$$Y = a * \cos\left(\frac{\pi(t+b)}{12}\right) + c + \varepsilon \quad (2)$$

3. maximum of constant and cosine function:

$$Y = \max\{c * \cos(at + b) + \varepsilon, 0.2\} \quad (3)$$

where Y is the recorded weather data at time t; t_{SR} and t_{SS} are the average monthly times of sunrise and sunset, respectively; a, b, and c are constants to be determined; and ε is the residual error due to randomness.

Precipitation data should also be well represented both temporally and spatially to calibrate diurnally varying hydrological model parameters [3,13,14]. Yet, the availability of detailed rainfall data, such as hourly data, is very limited in most areas [15]. To overcome such limitations, some of the plausible solutions are: (1) uniformly distribute daily rainfall data into hourly data, (2) stochastically generate hourly rainfall data, or disaggregate daily rainfall data into hourly data, (3) transfer detailed rainfall data from a nearby weather station to the area of interest, and (4) employ a multivariate disaggregation scheme – a combination of options 2 and 3 above.

Uniform distribution of daily rainfall data into hourly data is the simplest option available [6]. Unfortunately, hourly rainfall distribution is more non-uniform than it is uniform in most practical situations [16]. Rainfall distribution has been characterized by extreme variability in space and time [15,17]. Transferring rainfall data from a station near an area of interest is also a common practice among hydrologists who deal with diurnally varying hydrological models. This too has been critiqued for many argue that rainfall distribution is spatially highly variable [18,19]. In their study, Habib et al. [18] and Bradely et al. [19] have reported that spatial raingage distributions as dense as one raingage every 100 m to few 100 m away did not produce uniform readings, implying that denser raingage distributions should be used to accurately represent the reality on the ground. They have reported that the spatial correlation between hourly rainfall and separation distances between stations rapidly decreased with increases in the separation distances between stations.

On the other hand, many have studied the temporally stochastic nature of rainfall distributions (e.g., seasonality, inter and intra-storm variability, etc.) and developed various equations to help generate and/or disaggregate rainfall data from longer time-steps to shorter time-steps [16,20–22]. Some others [23,25] argue that in most practical situations rainfall distributions across smaller to medium sized watersheds can be assumed to have come from a population of similar distributions. Koutsoyiannis and Onof [23] have stressed the practical significance of using a multivariate approach (spatial and temporal) to disaggregate rainfall as opposed to a univariate approach (temporal alone) given that hourly rainfall data exist in the neighboring stations, and that there is significant cross-correlation between rainfall distributions at the gage stations – a case frequently met in practice for small and medium sized watersheds [23,25].

The reason is that an appropriate univariate disaggregation model would generate a synthetic hourly series, fully consistent with the known daily series and, simultaneously, statistically consistent with the actual hourly rainfall series. However, a synthetic series obtained by such a disaggregation scheme would not coincide with the actual one but would be a likely realization [26]. Nonetheless, using the multivariate approach, one could utilize the available hourly rainfall information at the neighboring station to

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

generate spatially and temporally consistent hourly rainfall series at the rain gauge of interest. That is, the spatial correlation becomes an advantage since (in combination with the available single-site hourly rainfall information) it enables a more realistic generation of the synthesized hyetographs. Thus, the location of a rainfall event within a day and the maximum intensity would not be arbitrary, as in the case of univariate disaggregation, but would resemble their actual values.

A wealth of literature is available on disaggregation techniques of hydrological data both spatially and temporally, though some of the methods are limited in their applications to specific rainfall conditions [5,23–30]. Koutsoyiannis and Manetas [7] have assumed a seasonal autoregressive model of lag one, AR(1), representing the historical monthly rainfall data and disaggregated annual to monthly rainfall data. Glasby et al. [30], Socolofsky et al. [5], and many more have disaggregated daily rainfall totals to hourly data. Grygier and Stedinger [13], Santos and Salas [21,27,29] and others have adopted a step-wise disaggregation approach where disaggregation is done at different levels: first by disaggregating the whole observed value into two, one for the first sub-period and another for the rest of the periods, and so on until all sub-periods are covered. Koutsoyiannis and Manetas [7] and Koutsoyiannis [26] have developed a simplified multivariate rainfall model by assuming that hourly rainfall distributions follow an AR(1) process. Koutsoyiannis [25] also reported the possibility of using higher order processes to represent hourly rainfall distributions. An AR(1) model is given by:

$$X_t = aX_{t-1} + bV_t \tag{4}$$

where $X_t := \left[X_t^1, X_t^2, \ldots, X_t^n\right]^T$ represents the hourly rainfall at time (t) and n locations, a and b are $(n \times n)$ matrices of parameters, and $V_t(t = \ldots,0,1,2,\ldots)$ is an independent identically distributed (IID) sequence of size n vectors of innovation random variables (so that the innovations are both spatially and temporally independent). Alternatively, the model can be expressed in terms of some nonlinear transformations (X_t^*) of the hourly depths X_t, in which case (4) is replaced by:

$$X_t^* = aX_{t-1}^* + bV_t \tag{5}$$

The most widely used rainfall disaggregation procedures can be summarized, after Koutsoyiannis and Manetas [7], as:

(1) Fit a suitable time series function with either observed (linear) or transformed (non-linear) variables;
(2) Use the fitted functions to generate sub-period values without reference to the given higher-level variables of that period, and;
(3) Apply an adjusting procedure to correct for the chosen sub-period values such that their total over the whole period of disaggregation be equal to observed values at the higher-level, without affecting the first and second order properties of the process.

The Levy-Stable [15] and the Bartlett–Lewis rectangular pulses (BLRP) [31–33] models are widely used viable options employed to reproduce individual rainfall event distributions. In this work, we adopted the modified Bartlett–Lewis rectangular pulse (BLRP) model to reproduce event rainfall distributions because of its wide application under different climates. The BLRP model, depicted in Fig. 1, is formulated based on the following assumptions:

(1) Storm origins (t_i) occur following a Poisson process with rate λ;
(2) Cell origins (t_{ij}) arrive following a Poisson process with rate β;
(3) Cell arrivals terminate after a time (v_i) exponentially distributed with parameter γ;
(4) Each cell has a duration (w_{ij})exponentially distributed with parameter η; and
(5) Each cell has a uniform intensity (X_{ij}) with a two-parameter (μ_x and σ_x) gamma distribution. In the original version of the model, all model parameters are assumed constant. In the modified version, the parameter η is randomly varied from storm to storm with a gamma distribution with shape parameter α and scale parameter v. Subsequently, parameters β and γ also vary in a manner that the ratios $\kappa = \beta/\eta$ and $\phi = \gamma/\eta$ be constant.

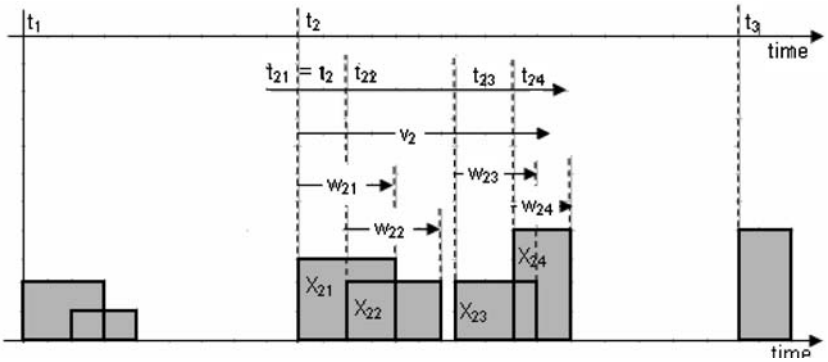

Fig. 1. The general assumptions of the Bartlett–Lewis rectangular pulse model (after Rodriguez-Iturbe et al., [31,32]; Onof and Wheater, [33]).

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

2. Study area and input data

The Cedar Creek watershed is located in the northeastern section of Texas (Fig. 2), which is part of the Trinity basin. Two weather stations: Terrell and Tyler (Fig. 2 and Table 1) were used as sources of detailed hourly weather data against which model disaggregations were compared. Detailed weather data (hourly rainfall, temperature, relative humidity, and wind speed) were taken from the National Climate Data Center (NCDC) website (http://lwf.ncdc.noaa.gov/oa/ncdc.html) for those two stations for the years from January 1998 to May 2004 (total of 77 months). Moreover, daily weather data for six other daily stations located in and around the watershed (Fig. 2 and Table 1) were obtained from the NCDC website. The primary weather data were checked for errors and missing values, and corrected whenever encountered based on the long-term trend of respective weather data distributions in the area. Weather data from Canton 5W station were excluded from the analyses for the lack of adequate data.

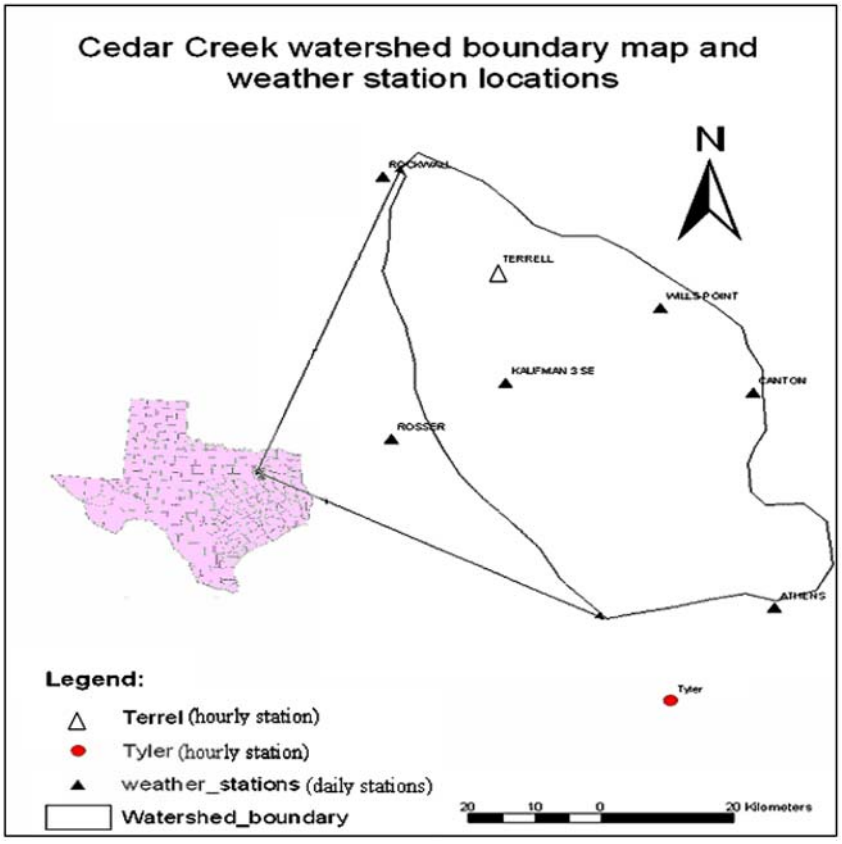

Fig. 2. The study area and locations of weather stations.

Table 1
Geographical location, elevation, distance from Terrell gauge station, and average annual rainfall data (1998–2003) of the gauge stations in the Cedar Creek watershed, Texas

Station name	Latitude	Longitude	Elevation (m)	Rainfall (mm)	Distance (km)
Athens	32.17	−95.83	136.6	1133.3	78.95
Canton 5 W	32.57	−95.95	147.8	1092.0	38.32
Kaufman 3SE	32.57	−96.27	128.0	975.5	22.29
Rockwall	32.93	−96.47	165.5	935.4	25.23
Rosser	32.47	−96.45	103.6	997.2	36.83
Terrell	32.77	−96.28	157.0	1114.4	0.00
Wills point	32.70	−96.02	159.1	1170.0	26.02
Tyler	32.35	−95.40	165.8	936.3	91.15
Statistics	Mean		145.4	1044.3	
	Standard deviation		21.6	93.6	

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

3. Approaches

3.1. Distribution and disaggregation models

Hourly weather data at the Terrell and Tyler stations were used to examine the characteristics of weather distributions in the watershed. We aggregated hourly weather data at these two hourly stations to create daily weather datasets – sum of 24 h for rainfall, and average for temperature, wind speed and relative humidity. Furthermore, daily maximum and minimum air temperatures were also determined from hourly dry bulb temperatures assuming that temperature distributions were uniform over each hour. These daily datasets were subjected to various disaggregation models to generate hourly weather data and the results were compared against measured hourly weather data. In addition, weather data from five daily stations were disaggregated into hourly data based on the study using detailed hourly data at the Terrell and Tyler stations.

3.1.1. Temperature, relative humidity and wind speed

Relative humidity, defined as the ratio of vapor pressures (actual to saturated vapor pressure), is a measure of the degree of saturation of the air at a given temperature [4]. The actual vapor pressure might be relatively constant throughout the day, but relative humidity fluctuates between a maximum near sunrise and a minimum around early afternoon (Fig. 3). On the other hand, the process of vapor removal depends to a large extent on wind and air turbulence, which transfers large quantities of air over the evaporating surface. Vapor removal by wind speed highly depends on a combined effect of climatic factors [4]. For example, wind speed has a far lesser effect on vapor removal in humid and warm climates, compared to dry and hot climates.

We distributed temperature data (maximum and minimum) using the methods recommended by Baker et al. [1] while that of wind speed was disaggregated using methods similar to that of Green and Kozek [2] (using a cosine function) and Neitsch et al. [34] (using a random function). In the Soil and Water Assessment Tool (SWAT) model, Neitsch et al. [34] used the following function to disaggregate monthly average wind speed into daily data:

$$w = w_{\text{mon}} * [-\ln(\text{rnd}[0,1))]^{0.3} \quad (6)$$

where w is the wind speed for the day (m s^{-1}), w_{mon} is the average wind speed for the month (m s^{-1}), and rnd[0, 1) is a random number between 0.0 and 1.0. On the other hand, hourly relative humidity was determined from hourly temperature distributions following the method described in Allen et al. [4].

We first fitted measured hourly temperature, relative humidity and wind speed data at the Terrell weather station, Texas (data from 1998 to 2003) to those selected models that represent each weather data and calibrated for corresponding models' parameters. We used the values of those models' parameters to subsequently disaggregate/distribute daily temperature, relative humidity, and wind speed data into respective hourly distributions at all seven stations (two hourly stations and five daily stations). We validated the models using hourly weather data at the Tyler station.

3.1.2. Rainfall disaggregation methods

While the uniform method of daily rainfall distribution involves no stochastic methodology except dividing the daily rainfall amount into 24 equal hourly values, the univariate (Hyetos) and Multivariate Disaggregation of Rainfall (MuDRain) models involve procedures that are more complicated. The Hyetos model is developed based on the modified BLRP model to disaggregate rainfall data at a single site. Where as, the MuDRain model is developed to disaggregate rainfall at single site or multiple sites based on temporal and spatial relationships between two or more sites' rainfall data. The essential statistics that should be preserved in the generated hourly series are: (1) the means,

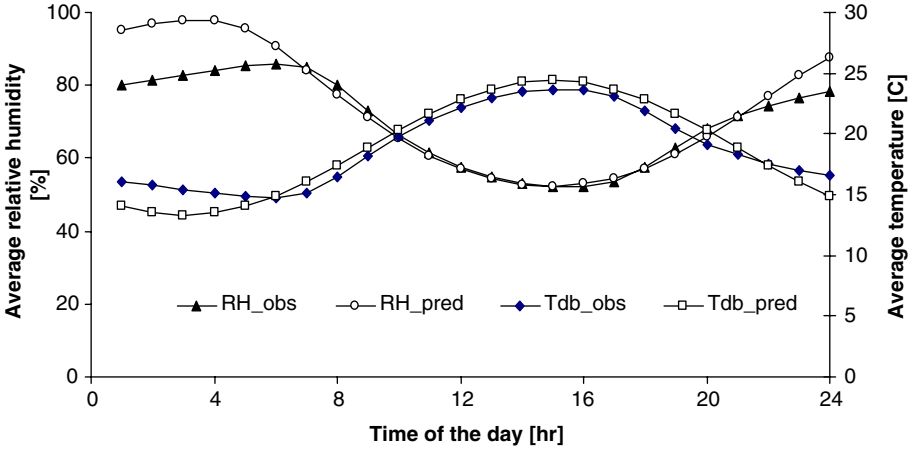

Fig. 3. Average hourly air temperature and RH distributions at the Terrell weather station, Texas (data from 1998 to 2003). Tdb_obs and Tdb_pred are the measured and predicted average hourly dry bulb temperatures, respectively; RH_obs and RH_pred are average hourly observed and predicted relative humidity data (%), respectively.

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

variances, and coefficients of skewness; (2) the temporal correlation structure (autocorrelations); (3) the spatial correlation structure (lag zero cross-correlations); and (4) the proportions of dry intervals. For more information about Hyetos and MuDRain models, readers may refer to the works by Koutsoyiannis et al. [24] and Koutsoyiannis [25], and on-line at: http://www.itia.ntua.gr.

3.1.2.1. The hyetos model. The Hyetos model disaggregates daily rainfall data at a single site into hourly data based solely on a temporal stochastic disaggregation scheme. It uses the Bartlett–Lewis rainfall model as a background stochastic model for rainfall generation. Then it uses a repetition scheme to derive a synthetic rainfall series, which resembles the given series at the daily timescale, and, subsequently, the proportional adjusting procedure, to make the generated hourly series fully consistent with given daily series [7,24]. The Hyetos model currently does not support model parameters' estimations. We developed a separate model to estimate model parameters based on the method of moments. To maintain seasonality in rainfall distribution, we computed model parameters on a monthly basis and the simulations were also done as such (Table 2). The Hyetos model operates under different modes where users can provide information at various details. We ran the model under two modes: full test mode and later under operational mode. The full test mode was run to test the suitability of the BLRP model and its parameters by comparing theoretically expected and predicted statistics (Table 3). The operational mode, which requires daily rainfall data and model parameters as input, is the most commonly used mode of operation for this specific application of daily rainfall disaggregation into hourly data. We calibrated the Hyetos model by generating various statistics and model parameters using hourly rainfall data from the Terrell station, and validated the model using hourly data at Tyler station.

3.1.2.2. The MuDRain model. The MuDRain disaggregation approach consists of two models that provide the required hourly rainfall series. The first model is a simplified multivariate rainfall model (in the form of either Eq. (4) or (5)) of hourly rainfall that can preserve the statistics of the multivariate rainfall process and, simultaneously, incorporate the available hourly information at hourly site(s), without any reference to the known daily totals at the other sites. The second model is a transformation model (commonly known as adjusting procedures) that modifies the series generated by the first model, so that the daily totals are equal to the given ones [25,26]. This uses a multivariate transformation, which does not affect the stochastic properties of the series.

To run the MuDRain model, we need two rainfall datasets: (1) daily rainfall data from the stations to be disaggregated into hourly data, and (2) hourly rainfall data from a reference weather station. In addition, cross-correlations between hourly rainfall data are also required to run the model. Computation of cross-correlations, and thus disaggregation runs, are recommended on monthly bases to maintain seasonality in rainfall distributions [26]. Hourly rainfall data at the Terrell and Tyler weather stations were arranged on monthly bases and cross-correlations between hourly rainfall data were computed. Cross-correlations were again computed based on daily rainfall data for the two stations. An empirical relationship similar to the one recommended by Koutsoyiannis [25] was developed to compute hourly cross-correlations for the rest of the six daily stations.

$$r_{ij,\mathrm{h}} = (r_{ij,\mathrm{d}})^m \tag{7}$$

Where $r_{ij,\mathrm{h}}$ and $r_{ij,\mathrm{d}}$ are the hourly and daily cross-correlation coefficients between stations i and j, respectively, and m is a constant determined based on the data at the Terrell and Tyler stations. In the cases where no hourly data is available, the value of m can be assumed approximately in the range of 2–3 [35] in [25]. The following procedures were adopted to prepare the data for MuDRain model run:

(1) We determined cross-correlations between hourly rainfall distributions at Terrell and Tyler stations – this was done month by month to maintain seasonality.

Table 2
Summary of the modified BLRP model parameters' estimates by month

Month	Λ (day^{-1})	κ (−)	ϕ (−)	α (−)	ν (day)	μ_x (mm/day)	σ_x (mm/day)
Jan	0.0101	1.5090	4.8723	15.4031	4.7532	1.3210	0.0541
Feb	0.0241	1.0845	5.6384	4.3364	4.9268	0.0959	0.0508
Mar	0.0078	1.7255	16.0978	17.4646	6.0242	0.2010	0.0068
Apr	0.0517	0.9713	7.8727	8.1949	2.4853	0.1559	0.0242
May	0.0073	0.9609	15.0618	9.0065	3.1151	0.3815	0.0063
Jun	0.0205	4.0829	85.4320	11.6315	2.5272	0.5651	0.0957
Jul	0.1488	0.6741	39.6472	5.9034	0.8581	0.1448	0.0947
Aug	0.0155	0.7881	9.6860	8.4716	5.0572	0.1990	0.0138
Sep	0.0436	2.1533	52.8035	13.7712	1.4325	0.9316	0.0541
Oct	0.0125	6.7191	118.7818	14.5528	2.5669	0.7492	0.0660
Nov	0.1159	1.1611	17.6791	14.7997	4.0112	0.4461	0.0762
Dec	0.0004	1.3311	15.0719	4.9929	3.3343	0.8485	0.0022
Ave.	*0.0382*	*1.9301*	*32.3870*	*10.7107*	*3.4243*	*0.5033*	*0.0454*
STD	*0.0470*	*1.7644*	*36.2038*	*4.4723*	*1.5630*	*0.3876*	*0.0342*

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

Table 3
Goodness-of-fit statistics produced by the Hyetos model using model parameters estimated based on hourly rainfall data at the Terrell weather station, Texas (data from 1998 to 2003)

Goodness-of-fit statistics	Weather stations						
	Terrell	Athens	Kaufman 3 SE	Rockwall	Rosser	Wills point	Tyler
Proportion of dryness	0.864	0.864	0.864	0.864	0.864	0.864	0.864
	0.878	0.845	0.921	0.889	0.834	0.810	0.887
Standard deviation	0.967	0.967	0.967	0.967	0.967	0.967	0.967
	1.107	1.111	0.930	1.320	0.998	0.941	1.238
Skew	19.308	19.308	19.308	19.308	19.308	19.308	19.308
	18.756	14.357	15.843	22.012	15.896	10.685	14.568
Lag -1 Auto-correlation	0.770	0.770	0.770	0.770	0.770	0.770	0.770
	0.756	0.773	0.738	0.764	0.779	0.737	0.735
Cross-correlation	0.742	0.781	0.807	0.740	0.860	1.000[a]	0.360
	0.653	0.682	0.674	0.605	0.846	1.000[a]	0.329

The first rows in each goodness-of-fit statistics category represent the theoretical expectations while the second rows denote that of generated data using the Hyetos model.
[a] Cross-correlations were computed against data at the Wills point gauge station.

(2) We aggregated hourly rainfall distributions at Terrell and Tyler stations into daily datasets, and determined cross-correlations between daily rainfall distributions, again on monthly basis separately.
(3) We fitted Eq. (7) between hourly and daily cross-correlations for each month based on data at Terrell and Tyler stations, and the values of m were determined (Table 4).
(4) We determined cross-correlations between daily rainfall data at Terrell and other five daily stations plus one hourly station (Tyler).
(5) Based on the values of m under step 3 and cross-correlations determined under step 4, hourly cross-correlations were back calculated for the five daily stations to be used in the MuDRain model (Table 5).

The MuDRain model was run with hourly rainfall data at the Terrell station as input and daily rainfall data for the rest of the six stations (five daily stations and Tyler station). The hourly rainfall datasets generated by MuDRain and observed onsite at the Terrell and Tyler stations were used for model calibration and validation, respectively. We also evaluated the applicability of the MuDRain model by comparing model generated theoretically expected and simulated statistics (Table 6).

3.2. Measure of prediction accuracy

The accuracy of the models' predictions was assessed using goodness-of-fit statistics. Both systematic and non-systematic error variations were calculated. Bias is a mea-

Table 4
Seasonal distribution of percent wet hour (%W) and hourly cross-correlation determination coefficient, m (Eq. (6)) at the Terrell weather station, Texas

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Ave
%W	1	5.3	6.9	6.4	3.4	3.3	3.9	1.5	1.9	5.2	5.0	6.1	7.1	4.7
	2	23.5	29.8	35.5	20.9	25.3	25.6	13.4	13.4	24.4	29.6	32.8	28.5	25.2
	3	7.2	9.3	7.0	3.3	4.8	5.5	2.3	2.6	5.7	5.4	5.8	9.0	5.7
	4	6.8	8.6	6.3	4.5	3.7	4.3	2.2	1.6	5.4	5.3	6.6	7.4	5.3
M		2.73	2.51	3.22	2.67	4.57	5.67	3.49	2.83	1.43	3.33	1.51	1.77	2.98

1, 2, 3 and 4 represent observed hourly rainfall data, uniformly distributed, generated by Hyetos and MuDRain models, respectively.

Table 5
Cross-correlations determined between hourly rainfall data for the month of December

DEC	Terrell	Athens	Kaufman 3 SE	Rockwall	Rosser	Wills point	Tyler
Terrell	1.000	0.509	0.530	0.624	0.627	0.600	*0.363*
Athens	0.509	1.000	0.539	0.551	0.676	0.651	*0.471*
Kaufman 3 SE	0.530	0.539	1.000	0.503	0.658	0.663	*0.403*
Rockwall	0.624	0.551	0.503	1.000	0.689	0.669	*0.436*
Rosser	0.627	0.676	0.658	0.689	1.000	0.790	*0.407*
Wills point	0.600	0.651	0.663	0.669	0.790	1.000	*0.360*
Tyler	*0.363*	*0.471*	*0.403*	*0.436*	*0.407*	*0.360*	1.000

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

Table 6
Goodness-of-fit statistics produced by the MuDRain model using data at seven weather stations, Texas (data from 1998 to 2003)

Goodness-of-fit statistics	Weather stations						
	Terrell	Athens	Kaufman 3 SE	Rockwall	Rosser	Wills point	Tyler
Proportion of dryness (−)	0.864	0.864	0.864	0.864	0.864	0.864	0.864
	0.864	0.831	0.890	0.869	0.857	0.822	0.842
Standard deviation (mm)	0.967	0.967	0.967	0.967	0.967	0.967	0.967
	0.967	1.031	0.910	1.22	1.048	0.911	0.918
Skewness (mm)	19.308	19.308	19.308	19.308	19.308	19.308	19.308
	19.308	13.157	13.878	21.311	16.236	10.334	13.345
Lag -1 auto-correlation (−)	0.770	0.770	0.770	0.770	0.770	0.770	0.770
	0.770	0.673	0.688	0.726	0.776	0.733	0.698
Cross-correlation (−)	0.742	0.781	0.807	0.740	0.860	1.000[a]	0.360
	0.673	0.672	0.670	0.618	0.813	1.000[a]	0.344

The first rows in each goodness-of-fit statistics category represent the theoretical expectations while the second rows denote that of generated hourly rainfall data using the MuDRain model.
[a] Cross-correlations were computed against data at the Wills point gauge station.

sure of systematic error that reflects consistent under- or over-prediction. In disaggregation aspects where hourly values are consistent with daily averages, bias is not theoretically anticipated. Non-systematic error variations, on the other hand, measure the expected magnitude of errors about the true values. The goodness-of-fit statistics [36–38] used to assess accuracy include the bias $\overline{e}$, Eq. (8); the standard error of the difference S_e, Eq. (9); the modified standard error of the difference S_{em}, Eq. (10); the relative bias R_b, Eq. (11); the relative standard error R_s, Eq. (12), the relative difference between observed and predicted hourly standard deviations (ΔS), Eq. (13), and significance of difference test Eq. (14). Statistical values were computed using measured and predicted hourly weather data with the GLM and ARIMA procedures in SAS [39].

Bias $\overline{e}$, standard error of the estimate S_e, and modified standard error of the estimate S_{em} are given by:

$$\overline{e} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{Y}_i - Y_i) \tag{8}$$

$$S_e = \sqrt{\frac{1}{v}\sum_{i=1}^{N}(\widehat{Y}_i - Y_i)^2} \tag{9}$$

$$S_{em} = \sqrt{\frac{1}{v}\sum_{i=1}^{N}(\widehat{Y}_i - \overline{e} - Y_i)^2} \tag{10}$$

where N is the total number of hours during which weather data were observed and generated, v is the degree of freedom ($v = N - 1$), Y and $\widehat{Y}$ are observed and predicted hourly weather data, respectively, and i is the counter for hourly weather dataset. The goodness-of-fit statistics can be standardized to yield dimensionless indices such as the relative bias R_b, relative standard error R_s, ΔS, and Test, which are given by:

$$R_b = \frac{\overline{e}}{\overline{Y}} \tag{11}$$

$$R_s = \frac{S_e}{S_y} \tag{12}$$

$$\Delta S = \frac{\widehat{S}_y - S_y}{S_y} \tag{13}$$

$$\text{Test} = \frac{\text{Abs}(\overline{Y} - \overline{\widehat{Y}})}{2S_e} \tag{14}$$

where $\overline{Y}$ and S_y are the mean and standard deviation (computed using Eq. (9) substituting $\overline{Y}$ for $\widehat{Y}$) of hourly observed weather data, respectively; $\widehat{S}_y$ is the standard deviation of the hourly predicted weather data; Test and $\overline{\widehat{Y}}$ are the test of significance of difference between observed and model predicted datasets and mean of predicted values, respectively; Abs(−) is the absolute value of the expression in the bracket. The closer the values of R_b, R_s, and ΔS to zero, the better the models are. Values of R_b, R_s, and ΔS greater than unity indicate reasonable differences between measured and model predicted values. The statistical significance of the differences was evaluated employing a criterion of $2 \times S_e$, which roughly corresponds to a two-sided test at a 5% significance level [40]. That is, if the value of Test (Eq. (14)) is larger than 1.0, then the difference is statistical significant. The probability of dryness/wetness was also computed and compared against those observed onsite.

4. Results and discussions

4.1. Temperature (T)

The cosine function of the following form has been fitted to estimate hourly temperature distributions from daily maximum and minimum temperature datasets (after de Wit [11]):

$$T_t = \frac{T_{max} - T_{min}}{2} * \cos\left(\frac{\pi(t-15)}{12}\right) + T_{av} \tag{15}$$

where T_t is temperature (°C) at time t (h) starting from midnight (in the range 1–24); T_{max}, T_{min}, and T_{av} are the maximum, minimum and average daily temperatures, respectively, (°C). Most weather stations record only daily

T_{max} and T_{min}. We fitted another equation to relate these daily measurements with the daily average (T_{av}) temperature. Finally, the following equation was best-fitted to measured temperature datasets ($N = 2311$):

$$T_{av} = 0.525T_{max} + 0.464T_{min} - 0.229 \text{ with } r^2 = 0.997 \text{ and RMSE} = 0.513. \quad (16)$$

Combining equations 15 and 16 together, temperature at time t (T_t) is given by:

$$T_t = \left(\frac{T_{max} - T_{min}}{2}\right) * \cos\left(\frac{\pi(t-15)}{12}\right) + 0.525T_{max} + 0.464T_{min} - 0.229 \quad (17)$$

with $r^2 = 0.94$. Table 7 shows the goodness-of-fit statistics computed for each weather variable from data at the Terrell weather station. Figs. 3 and 4, and Table 7 show that the fitted model has reproduced observed hourly temperature distributions in the area very well. Fig. 3 depicts the average hourly air temperature (5 year data – 1998–2003) over 24 h period. On the other hand, Fig. 4 depicts detailed hourly air temperature for a specific duration (over two days). Results from both graphs (Figs. 3 and 4) indicate that the model reproduced observed air temperature data very well. The high correlation coefficients between model estimates and observed data ($r = 0.97$), low values for $\Delta S = 0.009$ and $R_s = 0.25 \ll 1.0$, and statistically non-significant differences (Test $< 0.001 \ll 1.0$) also confirm that the cosine function could be used in the study area with great confidence. Table 7 also shows that the equation fitted to generate hourly temperature data from daily maximum and minimum temperatures has a very low bias (−0.008%), and thus is a dependable model.

4.2. Relative humidity (RH)

We fitted the following equation to compute hourly RH values based on actual (e_a) and saturated (e°) vapor pressures (after Allen et al. [4]), which themselves were derived from temperature distributions:

$$\text{RH} = \frac{e_a}{e^\circ} \quad (18)$$

The actual vapor pressure is usually assumed equal to the saturated vapor pressure at dew point temperature (T_{dp}) [4]. Thus, RH at time t (RH_t) can finally be rewritten as:

$$RH_t = \frac{e^\circ(T_{dp})_t}{e^\circ(T_{db})_t}, \text{ where } e^\circ(T) = 0.6108\exp\left(\frac{17.27T}{T+237.3}\right) \quad (19)$$

where e° is the saturated vapor pressure (mbar) at temperature T (°C), and T_{db} is the dry bulb/actual air temperature (°C). In the computation of relative humidity, and thus evapotranspiration, one also needs to know the dew point temperature (T_{dp}) corresponding with the actual air temperature (T_{db}). Since dew point temperature measurement is rare at many weather stations, we also decided to fit a functional relationship between commonly available weather data and dew point temperature. Most commonly, dew point temperatures are related to daily minimum air temperatures [4,41–43]. In addition, we also made two assumptions to determine hourly dew point temperatures: (1) that dew point temperature varies linearly between consecutive days, and (2) that mean daily dew point air temperature occurs at around sunrise, i.e., 6:00AM on average, at which time minimum air temperature is highly

Table 7
Summary of the goodness-of-fit statistics of various weather disaggregation schemes at the Terrell weather station, Texas (data from 1998 to 2003)

Goodness-of-fit statistics		T_{db}	T_{dp}	RH	Wind		Precipitation		
					SWAT	Cosine	Unif	MuVar	UnVar
Correlation		0.969	0.918	0.506	0.483	0.770	0.379	0.549	0.352
RMSE		5.47	11.842	365.73	14.01	5.91	1.332	0.948	1.140
Bias		−8.7E−5	−2.7E−3	−4.802	−0.765	0.265	9.37E−5	8.28E−5	−9.04E−6
SEE (S_e)		2.370	3.441	19.124	3.743	2.45	0.875	0.582	0.614
S_{em}		2.370	3.441	19.118	3.665	2.44	0.875	0.582	0.614
R_b		−4.7E−6	−2.2E−4	−6.9E−3	−0.102	0.035	7.77E−4	7.6E−4	−8.48E−5
STD	Obs	9.494	8.701	19.82	3.150	3.150	0.995	0.995	0.995
	Pred	9.585	8.044	18.59	3.138	3.153	0.359	0.597	0.619
	ΔS	0.009	−0.076	−0.062	−0.004	0.0009	− 0.639	−0.400	−0.378
R_s		0.250	0.396	0.965	0.979	0.778	2.435	0.975	0.992
Test		<0.001	<0.001	0.013	0.102	0.022	*1.218*	0.029	0.333

RMSE, STD and ΔS are the residual mean square error, standard deviation and change in standard deviation, respectively; Unif, MuVar and UnVar are the hourly rainfall distributions disaggregated using uniform, multivariate and univariate techniques, respectively; R_s, R_b, S_e and S_{em} are the relative standard error, relative bias, standard error of estimate (SEE) and modified SEE, respectively; SWAT and Cosine represent hourly wind speed data disaggregated following the SWAT model approach and Cosine functions, respectively; and Test is the test of significance of differences (a value greater than one shows a statistically significant difference between observed and model predicted datasets). RMSE, Bias, S_e, S_{em}, and STD all have the same units of measurements as their respective weather data category in each column (e.g., STD will have °C for T_{db} and T_{dp}, % for RH, m/s for wind speed, and mm for precipitation).

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009

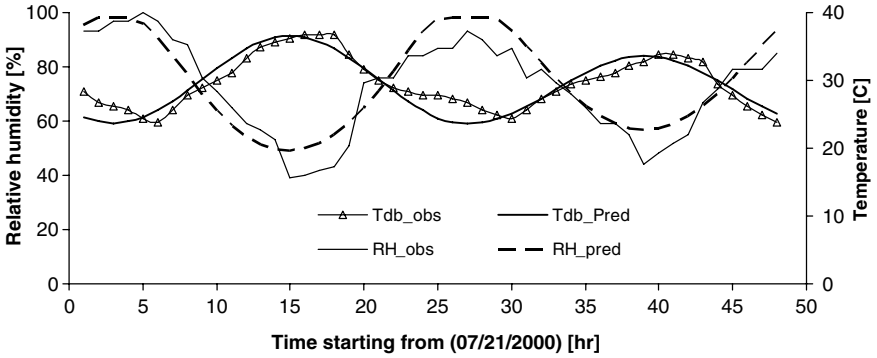

Fig. 4. Hourly air temperature and RH distributions at the Terrell weather station, Texas (starting from 07/21/2000). Tdb_obs and Tdb_pred are the measured and predicted hourly dry bulb temperatures, respectively; RH_obs and RH_pred are hourly observed and predicted relative humidity data (%), respectively.

correlated with dew point temperature [41–43]. Then, we fitted the following equations (after Meteotest [43]):

$$T_{\mathrm{dp\_hr}} = (T_{\mathrm{dp\_day}})_{\mathrm{d}} + \frac{hr}{24}[(T_{\mathrm{dp\_day}})_{\mathrm{d}} - (T_{\mathrm{dp\_day}})_{\mathrm{d+1}}] + T_{\mathrm{dp\_\Delta}} \tag{20}$$

where $T_{\mathrm{dp\_hr}}$ is the hourly dew point air temperature at time (h) (°C); $(T_{\mathrm{dp\_day}})_{\mathrm{d}}$, and $(T_{\mathrm{dp\_day}})_{\mathrm{d+1}}$ are the mean daily dew point air temperature on current day (d) at 6:00AM and next day (d+1) at 6:00AM, respectively (°C), and $T_{\mathrm{dp\_\Delta}}$ is the hourly fluctuations in dew point air temperature within a day (°C), which is given by:

$$T_{\mathrm{d\_\Delta}} = 0.5 * \sin\left[(hr+1) * \frac{\pi}{k_{\mathrm{r}}} - \frac{3*\pi}{4}\right] \tag{21}$$

where k_{r} is a constant that varies based on the average amount of monthly radiation. For average monthly radiation higher than 100 W m^{-2} (equal to an average of 8.64 MJ m^{-2} d^{-1}), $k_{\mathrm{r}} = 6$; else, $k_{\mathrm{r}} = 12$. We used $k_{\mathrm{r}} = 6$ for our analyses since average monthly radiation for all 12 months were higher than the 100 W m^{-2} threshold in our study area. The mean daily dew point air temperature (assumed to be equal to the dew point air temperature at sunrise – 6:00AM on average) is given by:

$$T_{\mathrm{dp\_day}} = 0.9153 * T_{\mathrm{min}} + 0.2021 \text{ with } R^2 = 0.903 \text{ and RMSE} = 6.96 \tag{22}$$

Table 7, and Figs. 3 and 4 depict the goodness-of-fit statistics and scatter plots between hourly observed and simulated RH datasets, respectively. The relative bias ($R_{\mathrm{b}} = -4.8\%$), relative standard error ($R_{\mathrm{s}} = 0.97 < 1.0$), $\Delta S = -0.06$, and Test $= 0.013 \ll 1.0$ (statistically non-significant difference between observed and model predicted datasets) show that the model's estimates were reasonably close to measured values, which confirms that the model can be used to simulate hourly RH in the area. Figs. 3 (average measured and predicted hourly RH distribution – data from 1998 to 2003) and four (hourly measured and simulated RH distributions on specific days) also show that our RH model performed well. Table 7 also shows the goodness-of-fit statistics between observed and simulated dew point air temperature datasets. The results indicate that our dew point temperature model predicted observed hourly dew point temperature data very well ($r = 0.918$ and Test $< 0.001 \ll 1.0$). However, compared to the performances of other models (e.g., dry bulb and dew point air temperature models), the RH model, despite its foundation on the relationship between dry bulb ($r = 0.969$) and dew point ($r = 0.918$) temperatures, did not perform so well ($r = 0.506$). The reason may be because of error propagation from the assumptions and errors made in temperature models and logarithmic relationships between RH on one hand and dry bulb and dew point air temperatures on the other hand (Eq. (19)).

4.3. Wind speed (w)

We best-fitted the following cosine function to the data at Terrell weather station (after Green and Kozek [2]):

$$W_{\mathrm{t}} = aW_{\mathrm{day}} * \cos\left(\frac{\pi(t-13)}{12}\right) + bW_{\mathrm{day}} \tag{23}$$

where W_{t} is the wind speed (m/s) at time t (h), W_{day} is average daily wind speed (m/s), and a and b are constants. The optimum value of b was computed to be 0.5 while that of a was determined to be within the following range:

$$a = \begin{cases} 0.9\left(1 - \frac{\mathrm{Abs}(13-t)}{14}\right) & \text{for } 6 < t < 20;\ t \text{ is time(h), and} \\ & \mathrm{Abs}\,(f) \text{ is the absolute value of } f. \\ 0.3 & \text{otherwise} \end{cases} \tag{24}$$

The values of the constants a and b were optimized for the best fit of Eq. (23) by constraining that the daily average of the Wt's from this equation should be equal to the observed daily averages. Statistical analyses between observed and simulated wind speed data are depicted in Table 7. A relative bias ($R_{\mathrm{b}} = 0.035\%$), relative standard error ($R_{\mathrm{s}} = 0.778$), $\Delta S < 0.001$, and Test $= 0.022 \ll 1.0$ (non-significant difference) indicate that statistically, wind speed data generated employing the cosine function were comparable with those measured onsite.

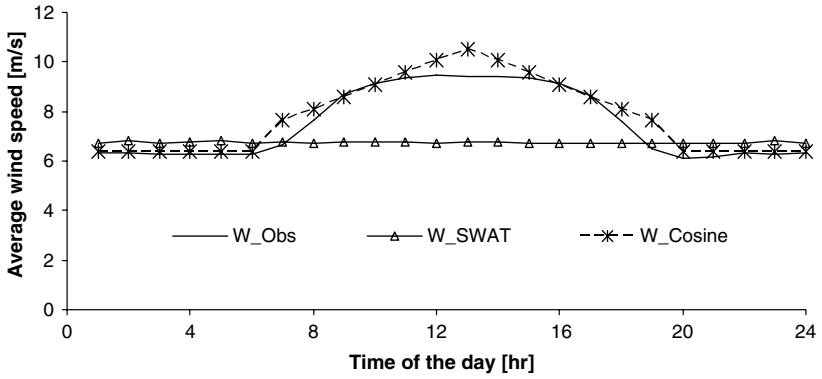

Fig. 5. Average hourly wind speed distribution at the Terrell weather station in Texas, (data from 1998 to 2003). W_Obs, W_Cosine, and W_SWAT are the measured, and simulated average hourly wind speed data following a cosine function and random distribution (SWAT approach), respectively.

Moreover, based on the assumption that wind speed can be modeled following random distributions, a modified exponential equation (Eq. (25)) was fitted by generating random numbers, $u\varepsilon$ [0,1) from uniform distribution. A similar approach has been used in the SWAT model to generate daily wind speed data from monthly averages [34]. Given daily average wind speed data, hourly wind speed can be computed using:

$$W_t = W_{day} * [-\ln(\mathrm{rnd}[0,1))]^{0.3} \quad (25)$$

where W_t is the hourly wind speed at time t (m s^{-1}), W_{day} is the average wind speed for the day (m s^{-1}), and rnd[0,1) is a random number between 0.0 and 1.0. The goodness-of-fit statistics (Table 7) and scatter plots (Figs. 5 and 6) depict that both approaches (cosine and random (SWAT) functions) produced reasonable results. However, comparing the two approaches, Table 7 and Fig. 5 establish that the cosine approach reproduced measured wind speed data better ($r = 0.77$ compared to $r = 0.48$, RMSE = 5.9 compared to RMSE = 14.0, and $S_e = 2.45$ compared to $S_e = 3.74$ for cosine and SWAT approaches, respectively). In addition, the scatter plots of measured versus simulated hourly wind speeds on specific days (Fig. 6) depict that both methods (cosine and SWAT) did not reproduce measured hour-to-hour wind speed data very well. However, both methods picked up the general trends of hourly wind speed in a day.

4.4. Rainfall (RF)

Average contribution of rainfall by each hour over the years at the Terrell weather station is shown in Fig. 7. A similar trend was observed at the Tyler weather station (not shown here). Fig. 7 depicts that over the years in the majority of the months, the percent contribution of rainfall by each hour of the day fluctuates very closely along the uniform line where each hour, on average, contributed equally. This shows that rainfall distribution in the area is not concentrated in certain period of the day over the years but is instead distributed over 24 h nearly equally, which is a prerequisite for the assumption of stationarity of rainfall distributions in each hour of the day. However, the months of July and May follow somewhat a different scenario. In the case of July, in all the six years studied, there was no rainfall observed in the hours between one and five. For the month of May there was a higher upward deviation from the uniform line during early hours of the day – around 70% of the rain fell in the hours ranging from midnight to midday. It is important to note, however, that this uniformity of rainfall distribution is not the same as the uniform distribution of rainfall over 24 h. The former is only about the contribution of rainfall by each hour of the day over the years while the latter illustrates whether or not the total daily rainfall is assumed to be uniform over the 24-h period.

4.4.1. Disaggregation using uniform distribution

Number of storms[1] and corresponding durations (data from 1998 – 2003) were determined for the Terrell weather station (Fig. 8). Fig. 8 depicts that more than 50% of the storms had durations less than or equal to 1 h, and more than 92% of the storms had durations barely equal to or less than 6 h. Only 33 storms out of 992 storms over six years (i.e., 3.33%) had durations stretching beyond 10 h. This result clearly contradicts any approach that assumes uniform distribution of daily rainfall over 24 h. Statistical values computed from historical versus disaggregated (using uniform distribution of daily rainfall over 24 h) hourly rainfall distributions are given in Tables 4 and 7. Table 4 depicts the percent wetness (%W) of hourly rainfall distribution for each month at the Terrell weather station. As expected, the uniform distribution method over-predicted the number of non-zero rainfall hours as volumes of short duration storms spread over 24 h. Table 7 also shows various statistical analyses made between historical hourly rainfall data and simulated data using the uniform distribution method (given under column Unif). The statistics confirm that hourly rainfall distribution generated by the uniform disaggregation method is statistically significantly different

[1] A storm is defined as rainfall duration with total rainfall > 0.25 mm and separated from another storm by = 6 h.

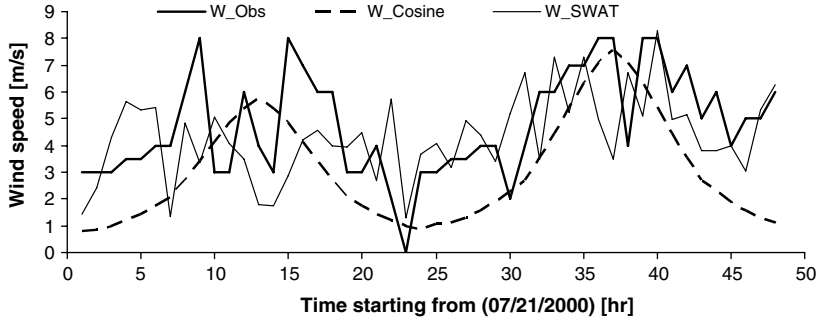

Fig. 6. Hourly wind speed distribution at the Terrell weather station in Texas, (starting from 07/21/2000). W_Obs, W_Cosine, and W_SWAT are the measured, and simulated hourly wind speed data following a cosine function and random distribution (SWAT approach), respectively.

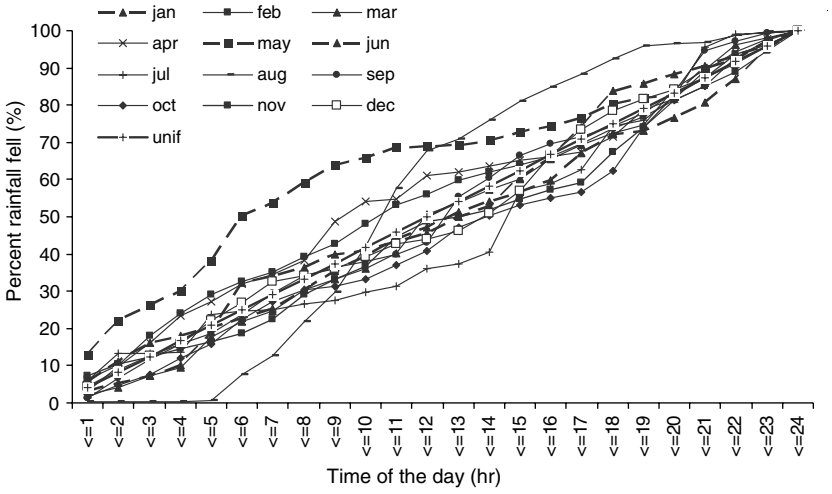

Fig. 7. Average diurnal distributions of total daily rainfall (%) by month at the Terrell weather station, Texas (data from 1998 to 2003).

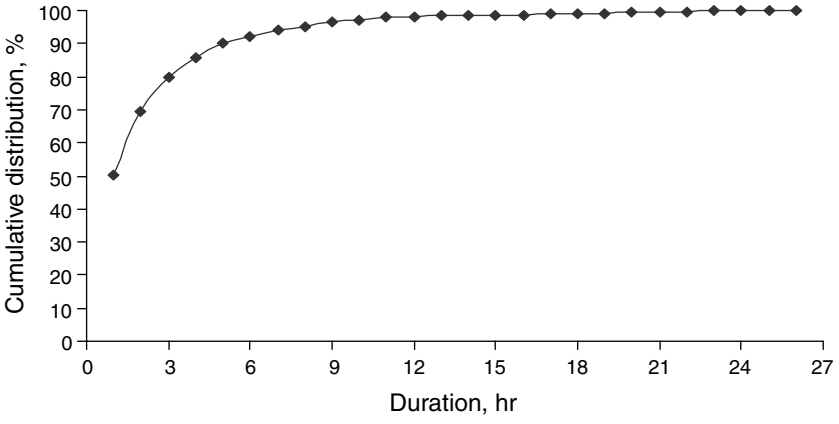

Fig. 8. Cumulative distribution of number of storms (%) and corresponding durations at the Terrell weather station, Texas (data from 1998 to 2003).

(Test = 1.218 > 1.0) from measured rainfall data. Thus, disaggregation based on uniform distribution cannot be an effective choice for use in the Cedar Creek watershed to simulate dynamic hydrological and water quality models.

4.4.2. Disaggregation using univariate stochastic methods (Hyetos model)

The summary of the modified Bartlett–Lewis model parameters estimates is present in Table 2. Parameters' estimates vary considerably over months indicating that separate monthly simulations are required to accurately reproduce actual rainfall distributions in the area. A review of the statistics produced by running the Hyetos model under the 'full test mode' employing the parameters estimated in Table 2 is also given in Table 3. Table 3 depicts the goodness-of-fit statistics, such as proportion of dryness, standard deviation, skewness coefficient, lag-1 autocorrelation and cross-correlation coefficients between theoretical

and disaggregated hourly rainfall data at the Terrell and six other stations for the month of December. Similar results were achieved for other months as well (not included). The expected statistics were closely reproduced by the model, compared with the statistics determined using the historical data, which is a clear indication of the appropriateness of the model in the study area.

Moreover, the statistical values computed from historical versus disaggregated hourly rainfall distributions employing the Hyetos model are shown in Table 7 (under column UnVar). Percent wetness (%W) as computed from observed and simulated (using Hyetos model) rainfall datasets is also shown in Table 4. The %W values were, on average, comparable. Values of the statistical analyses in Tables 3 and 7 confirm that the Hyetos model produced acceptable results, compared to measured values. The values of R_s and $\Delta S < 1$, and Test $= 0.333 < 1$ (statistically non-significance of the differences between measured and simulated hourly rainfall distributions) in Table 7 confirm the usefulness of this model to disaggregate daily rainfall into hourly data at the study area. Despite its effectiveness in reproducing average hourly rainfall characteristics, the univariate approach (Hyetos model) did not generate high correlations with observed hourly rainfall data, when compared with the uniform distribution. This can be explained by the assumption that in the Hyetos model, the beginning and durations of rainstorms follow random distributions, which may not coincide with the observed ones although the statistics, on average, are consistent with the observed rainfall data (Fig. 9).

4.4.3. Disaggregation using multivariate schemes (MuDRain model)

The values of m determined in our study range between 2 and 6 (Table 4) unlike what was reported in the study by Fytilas [35] in Koutsoyiannis [25]. They reported the value of m to approximately range between 2 and 3. The synopsis of cross-correlation coefficients computed between hourly rainfall distributions is given in Table 5. The result for only one month (December) is reported here for brevity. The cross-correlation coefficient between the Terrell and Tyler stations was very low, compared to other stations. This could be attributed to the long distance between the stations (91.15 km; Table 1) [18,19].

A summary of statistical analyses between theoretical and simulated (using MuDRain model) hourly rainfall distributions is given in Table 6. The table illustrates the comparison of the proportion of dryness, standard deviation, coefficient of skewness, lag-one autocorrelation, and cross-correlations between the theoretical and simulated datasets. The MuDRain model reproduced most of the statistics very well, and thus the model can be used with confidence [25] to disaggregate daily rainfall data into hourly data. A closer look at the comparison of the goodness-of-fit statistics in Tables 3 and 6 show that the MuDRain model reproduced expected statistics better, compared with the Hyetos model. The summary of statistics determined from the comparison of observed and model simulated (using MuDRain) hourly datasets is also shown in Table 7 under column MuVar. The overall statistical analyses indicate that the MuDRain model outperformed other rainfall disaggregation schemes. In addition, the limitations experienced with the Hyetos model (consistence in occurrence of storm events) was improved by employing a multivariate disaggregation scheme (MuDRain model) (Table 7 and Fig. 9). The higher correlation coefficient, non-significant differences (Test $= 0.029 \ll 1.0$) and smaller RMSE values for the linear equation between observed and simulated data obtained by the use of the MuDRain model, compared to other alternative disaggregation schemes, confirm its superiority in the study area (Table 7). Our result is consistent with those by Koutsoyiannis et al. [24] and Koutsoyiannis [25]. In addition, percent wetness (%W) computed by the MuDRain model, as shown in Table 4, also shows that the MuDRain model reproduced measured proportion of dry intervals very well.

Fig. 9 depicts the diurnal distribution of rainfall on a selected day (December 25–26, 2000). The MuDRain model reproduced both the peak and temporal distribution of rainfall (dry intervals) very well. However, the Hyetos

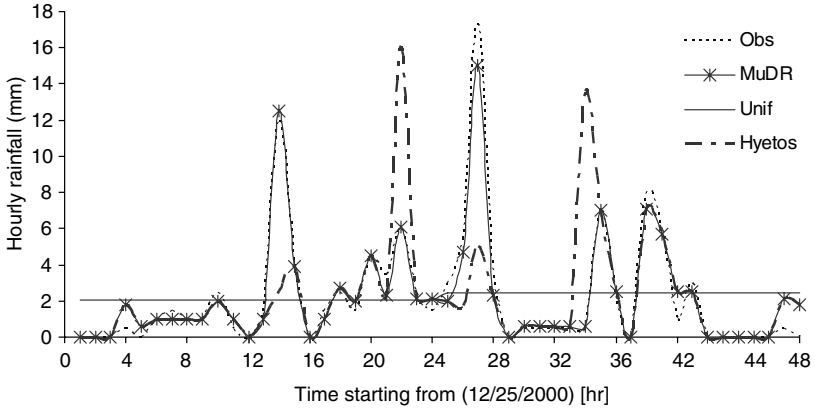

Fig. 9. Hourly measured and predicted rainfalls in Tyler, Texas. Obs, MuDR, Unif and Hyetos are the measured, and predicted hourly rainfall data using the multivariate, uniform and univariate disaggregation methods, respectively.

model did not reproduce time of occurrence of events although it had mimicked average rainfall characteristics very well. The effects of such mismatch in the temporal distributions of rainfall may be profound on small watersheds whose time of concentration is short (<24 h). The temporal consistency between observed and simulated hourly rainfall distributions can have a significant effect on further hydrological and water quality processes in the watershed [5]. Socolofsky et al. [5] reported that accurate placement of events during the day is necessary to accurately reproduce measured time series of water quality data because the first flush of contaminants during a storm can give a concentration spike of only a few events. Such effects govern the peak runoff rate and contaminant load that eventually also dictate watershed management decisions.

5. Conclusions

Weather data at finer timescales are vital inputs to simulate detailed physical and biological models. Yet, the dearth of detailed temporal and spatial weather data in most areas exacerbates the problem of running such models. In this study, we have examined the applicability of various approaches (deterministic periodic versus stochastic) of disaggregating daily weather data into hourly data. We found the cosine function appropriate to distribute daily maximum and minimum temperatures, and daily wind speed data into respective hourly datasets. In addition, we also fitted a functional relationship (linear equation) between daily minimum air temperature and mean daily dew point temperature. Furthermore, we used a sine function to represent dew point temperature fluctuations over a day whereby the combination of mean daily dew point temperature and hourly fluctuations produced hourly distributions of dew point air temperature. We also fitted a common logarithmic equation between temperature data (predicted hourly dry bulb and dew point air temperatures) and vapor pressures to compute hourly distributions of relative humidity (the ratio between actual and saturated vapor pressures multiplied by 100) data.

In addition, we also examined the use of different approaches of disaggregating daily rainfall data into hourly data. Disaggregation following uniform distribution of daily rainfall over 24 h did not reproduce most statistical parameters computed from observed hourly rainfall data onsite. Conversely, both stochastic models formulated based on univariate (Hyetos) and multivariate (MuDRain) processes mimicked measured hourly rainfall distributions, on average, very well. The MuDRain model reproduced better statistics in most cases including the time of coincidence between observed and simulated rainfall events, compared with other methods. The better temporal coincidence of events produced by the MuDRain model can be even more advantageous in further studies of the consequences of rainfall distributions on hydrological and water quality processes in the watershed. Whenever there is a reference hourly station in the neighboring watershed, the multivariate rainfall disaggregation scheme (MuDRain model) would be the appropriate choice. However, when reference hourly station is not in the vicinity, the univariate rainfall disaggregation method (Hyetos model) is also equally useful.

Acknowledgements

We would like to thank Ms. Marie-Laure (Imperial College, London) for providing some of the FORTRAN codes used for parameters estimation in the Hyetos model. We would also like to sincerely thank Dr. Koutsoyiannis for his detailed review and inputs towards the improvement of this manuscript. Great professional review by another anonymous reviewer is also acknowledged.

References

[1] Baker JM, Reicosky DC, Baker DG. Estimating the time dependence of air temperature using daily maxima and minima: A comparison of three methods. J Atmos Ocean Technol 1988;5:736–42.
[2] Green HM, Kozek AS. Modeling weather data by approximate regression quantiles. J ANZIAM 2003;44(E):C229–48.
[3] Connolly RD, Schirmer J, Dunn PK. A daily rainfall disaggregation model. Agric Forest meteorol 1998;92:105–17.
[4] Allen RG, Periera LS, Raes D, Smith M. Crop evapotranspiration; guidelines for computing crop water requirements. FAO irrigation and drainage paper no. 56, Rome, Italy, 1998. p. 300.
[5] Socolofsky S, Adams EE, Entekhabi D. Disaggregation of daily rainfall for continuous watershed modeling. J Hydrol Eng, ASCE 2001;6(4):300–9.
[6] Gutierrez-Magness AL, McCuen RH. Accuracy evaluation of rainfall disaggregation methods. J Hydrol Eng, ASCE 2004;9(2):71–8.
[7] Koutsoyiannis D, Manetas A. Simple disaggregation by accurate adjusting procedures. Water Resour Res 1996;32(7):2105–17.
[8] Cutrim EMC, Martin DW, Butzow DG, Silva IM, Yulaeva E. Pilot analysis of hourly rainfall in Central and Eastern Amazonia. J Climate, AMS 2000;13:1326–34.
[9] Santanello Jr JA, Friedl MA. Diurnal covariation in soil heat flux and net radiation. J Appl Meteorol, AMS 2003;42:851–62.
[10] Parton wj, Logan JA. A model for diurnal variation in soil and air temperature. Agric Meteorol 1981;23:205–16.
[11] de Wit CT. Simulation of assimilation, respiration and transpiration of crops. wageningen: Pudoc; 1978. p. 148.
[12] Sanders CG. untitled note. Hortscience 1975;10:560–1.
[13] Grygier JC, Stedinger JR. Condensed disaggregation procedures and conseravation corrections for stochastic hydrology. Water Resour Res 1988;24(10):1574–84.
[14] Salas JD. Analysis and modeling of hydrologic time series. In: Maidment D, editor. Hand book of hydrology. New York: McGraw-Hill; 1993 [chapter 19].
[15] Menabde M, Sivapalan M. Modeling of rainfall time series and extremes using bounded random cascades and Levy-stable distributions. Water Resour Res 2000;36(11):3293–300.
[16] Lall U, Sharma A. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resour Res 1996;32(3):679–93.
[17] Waichler SR, Wigmosta MS. Development of hourly meteorological values from daily data and significance to hydrological modeling at H.J. Andrews Experimental forest. J Hydrometeorol 2003;4:251–63.
[18] Habib E, Krajewski WF, Ciach GJ. Estimation of rainfall interstation correlation. J Hydrometeorol Am Meteorol Soc 2001;2:621–9.
[19] Bradley A, Peters-Lidard AC, Nelson BR, Smith JA, Young CB. Raingage network design using NEXRAD precipitation estimates. J Am Water Resour Assoc 2002;38(5):1393–408.

[20] Stedinger GR. Estimating correlations in multivariate stream-flow models. Water Reosur Res 1981;17:200–8.
[21] Santos EG, Salas JD. Stepwise disaggregation scheme for synthetic hydrology. J hydraul Eng 1992;118(5):765–84.
[22] Tarboton DG, Sharama A, Lall U. Disaggregation procedures for stochastic hydrology based nonparametric density estimation. Water Resour Res 1998;34(1):107–19.
[23] Koutsoyiannis D, Onof C. Rainfall disaggregation using adjusting procedures on a Poisson cluster model. J Hydrol 2001;246:109–22.
[24] Koutsoyiannis D, Onof C, Wheater HS. Multivariate rainfall disaggregation at a fine time scale. Water Resour Res 2003;39(7):1–18.
[25] Koutsoyiannis D. Multivariate disaggregation of rainfall – MuDRain, 2003. <http://www.itia.ntua.gr/>.
[26] Koutsoyiannis D. Coupling stochastic models of different time scales. Water Resour Res 2001;37(2):379–92.
[27] Koutsoyiannis D. A nonlinear disaggregation method with a reduced parameter set for simulation of hydrologic series. Water Resour Res 1992;28(12):3175–91.
[28] Koutsoyiannis D, Foufoula-Georgiou E. A scaling model of storm hyetograph. Water Resour Res 1993;29(7):2345–61.
[29] Koutsoyiannis D, Xanthopoulos T. A dynamic model for short-scale rainfall disaggregation. Hydrol Sci J 1990;35(3):303–22.
[30] Glasbey CA, Cooper G, McGechan MB. Disaggregation of daily rainfall by conditional simulation from a point-process model. J Hydrol 1995;165:1–9.
[31] Rodriguez-Iturbe, Cox DR, Isham V. Some models for rainfall based on stochastic point processes. Proc Royal Soc London A 1987;410:269–98.
[32] Rodriguez-Iturbe, Cox DR, Isham V. A point process model for rainfall: Further developments. Proc Royal Soc London A 1988;417:283–98.
[33] Onof C, Wheater HS. Modeling of British rainfall using a Random parameter Bartlett–Lewis rectangular pulse model. J Hydrol 1993;149:67–95.
[34] Neitsch SL, Arnold JG, Kiniry JR, Williams JR. Soil and water assessment tool theoretical documentation. Blackland Research Center, Texas Agricultural Experiment Station, Temple, Texas; 2001.
[35] Fytilas P. Multivariate rainfall disaggregation at a fine time scale, Diploma thesis submitted at the University of Rome, La Sapienza, 2002.
[36] McCuen RH. Microcomputer application in statistical hydrology. Upper Saddle River, NJ: Prentice-Hall; 1993.
[37] Ayyub BM, McCuen RH. Probability, statistics and reliability for engineers and scientists. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC press; 1997.
[38] Wojcik R, Buishand TA. Simulation of 6-hourly rainfall and temperature by two resampling schemes. J Hydrol 2003;273:69–80.
[39] SAS. Statistical Analyses Software – version 9.0. Cary, NC, USA: SAS Institute Inc.; 2002.
[40] Buishand TA, Brandsma T. Multi-site simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. Water Resour Res 2001;37:2761–76.
[41] Kenneth GH, Rezaul M, Carlson C. Estimating daily dew point temperature for the Northern Great Plains using maximum and minimum temperature. J Agronomy 2003;95:323–8.
[42] Sparks S, Changnon D, Starke J. Changes in the frequency of extreme warm-season surface dew points in Northeastern Illinois: implications for cooling-system design and operation. J Appl Meteorol 2002;41:890–8.
[43] Meteotest, Meteonorm version 5.0. The global meteorological database for engineers, planners and education. Software and data on CD-ROM. James and James, London, United Kingdom; 2003.

Please cite this article in press as: Debele B et al., Accuracy evaluation of weather data generation and ..., Adv Water Resour (2007), doi:10.1016/j.advwatres.2006.11.009