Available online at www.sciencedirect.com
SCIENCE DIRECT®

Ecological Modelling 192 (2006) 473–498

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Predicting the temporal and spatial probability of orographic cloud cover in the Luquillo Experimental Forest in Puerto Rico using generalized linear (mixed) models

Wei Wu[a,*], Charles Hall[b], Lianjun Zhang[c]

[a] *Graduate Program in Environmental Science, State University of New York, College of Environmental Science and Forestry, 301 Illick Hall, 1 Forestry Drive, Syracuse, NY 13210, USA*
[b] *Department of Environmental and Forest Biology, State University of New York, College of Environmental Science and Forestry, 354 Illick Hall, 1 Forestry Drive, Syracuse, NY 13210, USA*
[c] *Department of Forest and Natural Resource Management, State University of New York, College of Environmental Science and Forestry, 323 Bray Hall, 1 Forestry Drive, Syracuse, NY 13210, USA*

Received 8 December 2004; received in revised form 24 June 2005; accepted 18 July 2005
Available online 30 September 2005

Abstract

We predicted the spatial pattern of hourly probability of cloud cover in the Luquillo Experimental Forest (LEF) in North-Eastern Puerto Rico using four different models. The probability of cloud cover (defined as "the percentage of the area covered by clouds in each pixel on the map" in this paper) at any hour and any place is a function of three topographic variables: aspect, slope and the difference between elevation and lifting condensation level. We chose the best models based on multiple statistics including the Akaike Information Criterion (AIC), scaled deviance and extra-dispersion scale. As a result, the generalized linear model (GLM) and one generalized linear mixed model (GLMM) with exponential spatial structure were the best candidate models. The probabilities of cloud cover in both our simulations and the observations increased with elevation, and were higher at night. They decreased in the morning after the sun rose until early afternoon, and then increased again for the rest of the day until night, apparently in response to the movement of the lifting condensation level. Two types of satellite images were available to calibrate our models: the higher spatial resolution, but expensive and infrequent Landsat-7 Enhanced Thermal Mapper plus (ETM+) images and the frequent, free, but low spatial resolution Moderate Resolution Imaging Spectroradiometer (MODIS) images. The derived probabilities of cloud cover when calibrated to the two types of remote sensing images were very similar, which justifies our using the free MODIS images instead of the Landsat images to calibrate the models. We applied the model to all months and the results indicated in agreement with the data that the probability of cloud cover is less during the dry season,

* Corresponding author. Present address: 151 Link Hall, Department of Civil and Environmental Engineering, Syracuse University, Syracuse, NY 13244, USA.
E-mail address: wwu01@mailbox.syr.edu (W. Wu).

doi:10.1016/j.ecolmodel.2005.07.031

higher in wet season and moderate for the rest of the year. We found our models could usually predict the probability of cloud cover for each 100 m in elevation level at a certain time with an index of agreement (IoA) of 0.560–0.919 and at a certain location over a day with an IoA of 0.940–0.994, indicating a medium to good model simulation at that particular time or location.

Keywords: Generalized linear model (GLM); Generalized linear mixed model (GLMM); Spatial autocorrelation; probability of cloud cover; MODIS; Landsat-7 ETM+

1. Introduction

Tropical forests account for 32–36% of global terrestrial net primary production (Melillo et al., 1993; Field et al., 1998; Graham et al., 2003) and a large part of the world's actual evapotranspiration, which, with related processes, can explain 70% of the global energy transport through latent heat (Mauser and Schädlich, 1998). Clouds are extremely important to the carbon, water and energy budget of tropical forests since heavy cloud cover can reduce solar radiation on the earth and limit photosynthesis, sap flow and transpiration of tropical trees (Graham et al., 2003). Recent eddy covariance studies indicate that day-to-day variation in cloud cover and irradiance also affects the net carbon uptake by forests (Graham et al., 2003). Ongoing land use and global climate changes will induce changes in cloud formation. Lawton et al. (2001) and Nair et al. (2003) found that clouds became less abundant over deforested areas in the Atlantic lowlands of Costa Rica by analyzing satellite imagery. They also suggested that the lifting condensation level increased after deforestation due to an assumed reduction of the evapotranspiration rate and an enhanced sensible heat flux. However, these changes in the surface energy balance were not validated (Van der Molen, 2002). A relation between land cover and cloud formation is also suggested by the observations at the LEF after massive defoliation occurred during the passage of hurricane Hugo in 1989 (Van der Molen, 2002). The cloud base was lifted temporarily to a level above the highest peaks, which are normally surrounded by clouds. A permanent change in the lifting condensation levels would be likely to put already endangered ecosystems of tropical montane cloud forest into danger of extinction (Pounds et al., 1999; Bruijnzeel and Hamilton, 2000). This is because the trees in the tropical montane cloud forest are surrounded by clouds most of time and hence receive a considerable amount of their water supply and nutrients through the process of cloud stripping (Bruijnzeel and Procter, 1995; Bruijnzeel, 2000). They are very sensitive to the lifting of the cloud base (Still et al., 1999). Other studies indicate a significant decrease in insolation from the 1960s to 1990 and subsequently an even larger increase (Pinker et al., 2005; Wild et al., 2005). This dimming and subsequent brightening could have resulted from changes in cloud coverage, the abundance of atmospheric aerosols, or atmospheric transparency after explosive volcanic eruptions (Wild et al., 2005). Therefore it is vital to model the probability of cloud cover in tropical forests, and we are unaware of any studies that have done this both spatially and temporally.

Clouds develop in any air mass that becomes saturated with water (i.e. where relative humidity becomes 100%) in general. Saturation can occur by way of any one of a number of atmospheric mechanisms that cause the temperature of an air mass to be cooled to its dew point or frost point. The following mechanisms or processes can cause clouds to develop (derived from http://www.physicalgeography.net/fundamentals/8e.html):

(1) *Orographic uplift* occurs when air is forced to rise because of the physical presence of elevated land. As the air parcel rises it cools as a result of adiabatic expansion at a rate of approximately 6 °C/1000 m until saturation.
(2) *Convectional lifting* is associated with surface heating of the air at the ground surface. If enough heating occurs, the mass of air becomes warmer and lighter than the air in the surrounding environment, and it begins to rise, expand, and cool until saturation.
(3) *Convergence or frontal lifting* takes place when two masses of air come together. The leading edge of the cold and dry air mass acts as an inclined wall or front causing the moist warm air to be lifted and become cool until saturation.

(4) *Radiative cooling* occurs when the earth loses energy in the form of longwave radiation, which causes the ground and air above it to cool down until saturation. This process usually happens at night.

In mountainous areas such as the Luquillo Experimental Forest in North-Eastern Puerto Rico, orographic uplift is the main driver to form clouds for most days. However, clouds can also be caused by convergence or tropical waves. In this paper, we focused on predicting the probability of orographic cloud cover, thus the cloud coverage is related closely to topographic variables such as elevation and aspect.

The simplest way to model cloud cover would be to use a random number generator calibrated to weather station information. However, based on the mechanisms of cloud formation given above, the existence of clouds is not random. On the other hand, it is very difficult to model the probability of cloud cover using deterministic models due to the random behavior implicit in clouds. Statistical models – one type of stochastic models – are a better candidate than deterministic models because statistical models contain random factors in addition to deterministic mechanisms. Hence, regression techniques can be applied to remote sensing images in statistical models to quantify the relation between topographic variables on the earth and the associated remotely-sensed cloud cover (Foody, 2003). Multivariate logistic regression models (also known as logit models), a type of generalized linear models (GLMs), provide one approach to analyzing binary response variables (e.g. clouds' absence or existence) to estimate the probability of the question of interest. GLMs are an extension of the classical linear statistical models to response distribution that belong to the exponential family. The exponential family contains the Bernoulli, Binomial, Poisson, Negative binomial, Gamma, Gaussian, Beta, Weibull, and other distributions. GLMs combine elements from linear and nonlinear models, and they contain three components: the link function, the linear predictor, and the random component (Schabenberger and Pierce, 2002, P301). The basic idea of GLMs is to estimate the parameters of a linear model using maximum likelihood based on the distribution of the data (Littell et al., 1996, P423). However, if the observations are not independent from each other, one of the assumptions of regression – that the data are statistically independent – is violated. Cloud cover belongs to this type of observations since a cloud is more likely to exist if neighboring areas are covered by clouds. The effects of spatial dependency on conventional statistical methods include biased estimation of error variance and t-test significance levels, and an overestimation of R^2 (Anselin and Griffith, 1988). Thus spatial dependency could be seen as a methodological disadvantage. But if we quantify the spatial correlation in the right model, we can get more accurate estimates as well as standard errors.

Remote sensing techniques are very useful tools to model phenomena at a large scale, such as cloud cover, since the imagery covers a large area and can provide estimates at high spatial and temporal resolution (Kite and Droogers, 2000). The mean global cloud cover per day may exceed 60% (Bussieres and Goita, 1997), so that satellite images in many geographic locations are often contaminated with clouds. Although these cloudy scenes limit the usefulness of satellite measurements for other parameters, they can be very useful to help interpret temporal changes and spatial patterns of cloud coverage. Utilization of satellite imagery to distinguish cloudy areas from clear areas is known as cloud masking (Logar et al., 1998). Typical algorithms for cloud masking assume that clouds can be detected using thresholds derived from both the visible and the infrared channel (Welch et al., 1999). Other approaches rely upon "bispectral thresholding" and a variety of statistical methods (Welch et al., 1999). Concepts from neural network classification provide an improvement in both classification speed and accuracy over the traditional threshold and statistical techniques, but the procedures are more complicated and difficult to apply (Welch et al., 1999).

In this paper, we estimated the average hourly probability of cloud cover for each month with and without consideration of spatial autocorrelation in the data of cloud cover. In order to do that, we developed one generalized linear model (GLM) which did not consider spatial autocorrelation in the cloud data and three generalized linear mixed models (GLMM) that incorporated three different spatial structures: exponential, spherical and Gaussian to account for spatial autocorrelation in the cloud data. We calibrated the models by using two sets of remote sensing images, Landsat-7 Enhanced Thermal Mapper plus (ETM+) images and Moderate Resolution Imaging Spectroradiometer (MODIS) images on the Terra satellite. Landsat-7 ETM+ images have higher spatial resolution, 30 m in visible bands and infrared bands compared with MODIS level 1B images with 250 m resolution in both bands, but the MODIS image has higher temporal resolution, 2 days compared with 16 days for Landsat-7 ETM + images. The ETM+ image also has higher spectral resolution with eight bands including visible bands, infrared bands and a thermal band, while MODIS level 1B images with 250 m resolution have only two bands, visible and infrared. In addition, Landsat-7 ETM+ images have three visible bands—blue, green and red, and the MODIS images have only one integrated visible band. Finally, MODIS images are free while one scene of ETM+ imagery costs about $600. So we compared the derived probability of clouds calibrated by both images to each other to determine if the free, more frequent MODIS images can be used to calibrate the GL(M)Ms and reach a similar degree of precision as the Landsat ETM+ images.

2. Data

We were able to find the Landsat-7 ETM+ images provided by Global Land Cover Facility free of charge and the corresponding MODIS images on the same days for 27 March 2000, and 20 July 2001, which were relatively clear days, and on 9 January 2001, and 4 March 2003 which were relatively cloudy days, so that we have eight scenes altogether (Table 1). The scenes of the MODIS images are geo-referenced. The scene of Landsat-7 ETM+ images captured on 27 March 2000 is geo-rectified, which indicates a high degree of absolute geometric accuracy for analytical applications. The scene was radiometrically calibrated, corrected for sensor, platform induced, geometric and topographic distortions and mapped to Universal Transverse Mercator (UTM) cartographic projection. The Landsat-7 ETM+ scenes on 9 January 2001 and 20 July 2001 are at the level of “terrain corrected” which is radiometrically and geometrically corrected using the satellite model and platform-ephemeris information. It is rotated and aligned to UTM map projection using ground control points and a digital terrain model to improve the satellite model and remove geodetic inaccuracy caused by the parallax error that occurs because of local terrain elevation. The Landsat-7 ETM+ scene on 4 March 2003 is at level 1G which is radiometrically and systematically corrected (geometric correction) with UTM projection.

The other data we used include air temperature and dew temperature data from the continuous records at the El Verde station (18°19′22″N, 65°49′13″W) at 350 m in elevation in the LEF in order to calculate lifting condensation level. A digital elevation model (DEM) of the LEF at state plane reference system was also used for deriving topographic variables such as elevation, slope and aspect.

3. Study area

The Luquillo Experimental Forest (LEF), located in the Luquillo Mountains in the north eastern part of Puerto Rico, between 18°14′45.78″ and 18°20′58.23″N latitude and between 65°42′46.56″ and 65°53′53.33″W longitude (Wang, 2001), has elevations ranging from about 100 to 1075 m above mean sea level (Weaver and Murphy, 1990). The mean annual rainfall increases with elevation from approximately 2450 mm/year at lower elevations to over 4000 mm/year at higher elevations (Wang, 2001). Rainfall is distributed fairly evenly throughout the year in the LEF, with May and September–December being relative wet and February to April being relatively dry (Scatena, available at http://luq.lternet.edu/research/projects/climate_hydrology_description.html; Schellekens et al., 2000). The mean annual temperature declines from 23° to 19° from 200 to 1050 m in

Table 1
Time when the eight remote sensing images were taken

Time	27 March 2000	9 January 2001	20 July 2001	4 March 2003
Landsat-7 ETM+	10:36 local time (geo-rectified)	10:30 local time (terrain corrected)	10:30 local time (terrain corrected)	10:32 local time (L1G)
MODIS level 1B	11:15 local time	11:10 local time	11:05 local time	11:00 local time

elevation (Brown et al., 1983; Weaver and Murphy, 1990; Scatena and Lugo, 1995; Silver et al., 1999). As a result of changes in rainfall, temperature and cloud with elevations, the forests become shorter, denser, less-species-rich and less productive (Waide et al., 1998). Four life zones occur in the LEF: subtropical wet forest, subtropical rain forest, lower montane wet forest, and lower montane rain forest (Ewel and Whitmore, 1973), and four major vegetation types occupy these life zones, which is stratified roughly by the altitude, soil moisture and clouds. Below 600 m the dominant tree is the tabonuco (*Dacryodes excelsa*), which is best developed on protected, well-drained ridges and occupies nearly 70% of the LEF. Above the average cloud condensation level (600 m), palo colorado (*Cyrilla racemiflora*) is the dominant tree, which covers about 17% of the LEF, except in areas of steep slope and poorly drained and saturated soils, where the sierra palm (*Prestoea montana*) occurs in nearly pure stands (11% of the LEF). The elfin forest (also called cloud or dwarf forest, 2% of the LEF) occupies ridge lines above 750 m in elevation and is composed of dense stands of short, small diameter trees and shrubs that are almost continually exposed to winds and clouds. Both the palm and dwarf forests are dominated by only a few plant species (Brown et al., 1983). The elfin forest, Colorado forest and upper elevation palm forest are considered as endangered “tropical montane cloud forests”. The trees in those forests are surrounded by clouds most of time, receive a considerable amount of their water supply through the process of cloud stripping (Bruijnzeel and Procter, 1995; Bruijnzeel, 2000), and are very sensitive to the lifting of the cloud base (Still et al., 1999). A change in the cloud base may affect the biotope of the already endangered tropical montane cloud forest (Pounds et al., 1999; Bruijnzeel and Hamilton, 2000).

4. Methods

We predicted the hourly probability of cloud cover of each 240 m × 240 m grid cell for the entire LEF by using one GLM and three GLMMs that incorporated three different spatial structure models. The three spatial structure models are spherical, exponential and Gaussian models, and they have been found to be the most useful for geo-referenced data analysis to account for the spatial autocorrelation (Griffith and Layne, 1999, pp. 134–135). They can be applied in the

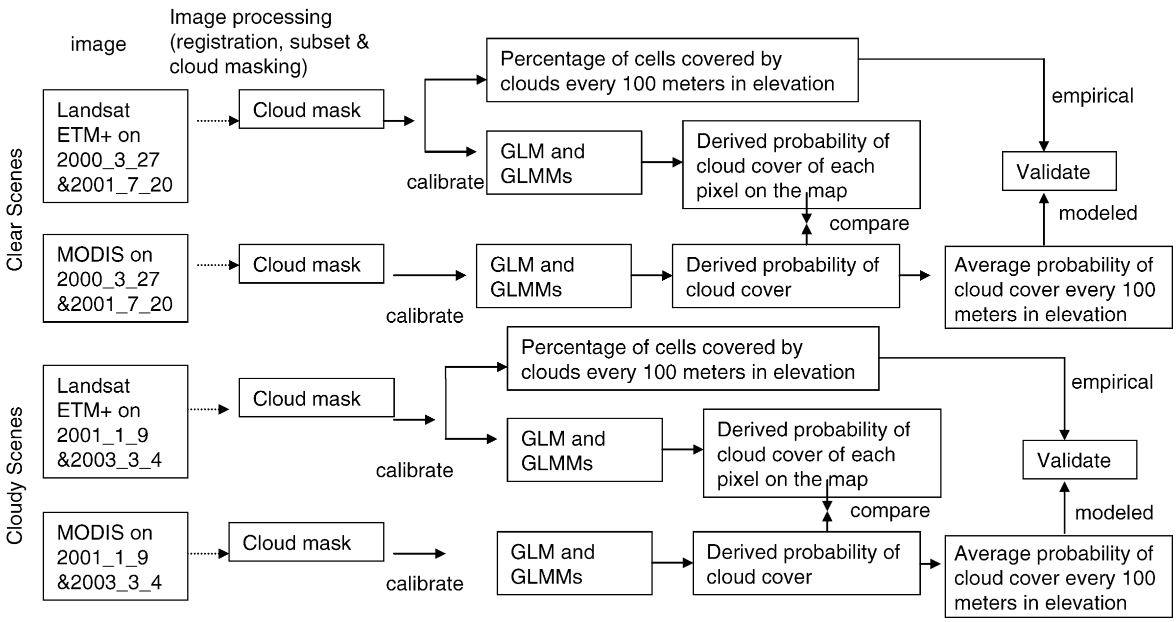

Fig. 1. Flow chart for deriving the probability of cloud coverage from the GLM and GLMMs calibrated using two sources of satellite data.

"mixed" procedure that accounts for spatial autocorrelation in the Statistical Analysis Software (SAS) we used. The hourly probability of cloud cover is a function of three topographic variables: aspect, slope and the difference between elevation and lifting condensation level. In order to complete the predictions, we followed the steps shown in the flow chart (Fig. 1). First, all the eight remote sensing images were resampled to a consistent state plane 27 reference system, and then registered spatially to the Landsat-7 ETM+ 2000-3-27 scene, which, as explained in the "data" section, has a high degree of absolute geometric accuracy. Then the registered images were subset to cover only the area of the LEF. Next we derived the cloud masks from the eight registered and subset images and then used them to calibrate the models. The coefficients of the independent variables in the GLM and GLMMs were estimated through pseudo-likelihood procedures using a SAS macro called "GLIMMIX".

4.1. Registration

Due to the presence of a number of distortions in remote sensing images that occur as a result of variations in platform positions, rotation of the earth, relief displacements, etc. (Chen et al., 2003), the images must be registered spatially before the location of the clouds can be represented accurately relative to the earth. Registration is a very time- and labor-consuming process, so automatic registration becomes an important topic to study. We tried to use the REGEEMY automatic registration system (Fedorov et al., 2002) but by inspection it did not give as good results as manual registration, probably because the MODIS images have only two bands and the cloud cover of some of the scenes was too high. So we decide to use the manual registration procedure from the "ERDAS Imagine" software. The reference image we used is the landsat ETM+ 2000-3-27 scene which was already geo- and ortho-rectified when we got it. GeoCover-Ortho scenes are the most accurate, commercially available base maps of the world created from Landsat imagery and have a better positional accuracy (50 m RMS) than the vast majority of the world's 1:200,000 maps (http://www.geocover.com/gc_ortho). When we registered the MODIS images (250 m resolution) to the ETM+ reference image (30 m resolution), we resampled both images to 240 m for each pixel. The relative mean square errors (RMSEs) of registration are all within one pixel.

4.2. Subset

The entire scenes were cut to a rectangular area that includes the LEF using the "subset" module in "ERDAS Imagine".

4.3. Cloud identification

We used the "unsupervised classification" module in Erdas Imagine to classify the registered and subset images to 5–10 undefined classes, and then we compared the classified images to the unclassified images by inspection to decide which classes should be combined to form a "cloud" class, which was then labeled as 1, with the other areas which were all labeled as 0 to represent "non-cloud" areas.

We also tried the threshold method to identify cloudy areas (Wen et al., 2001) for Landsat images obtained on 9 January 2001 (a relatively cloudy scene) and on 20 July 2001 (a relatively clear scene). Then we compared the cloud masks derived from "unsupervised classification" and the threshold method.

4.4. Model description

We calculated join-count statistics using "ROOKCASE" software (Sawada, 1999) to determine the spatial autocorrelation among pixels in the cloud masks. The commonly used measures of spatial autocorrelation include Moran's I statistics, Geary's C, and join-count statistics. Moran's I and Geary's C are for interval and ratio data, while join-count statistics is for testing binary and nominal data. In join-count statistics, spatially structured data sets are treated as mosaics of areas with different colors. For binary data, the two values of the variable in binary data are referred to as "black" (B) and "white" (W). Elements with a common boundary are said to be linked by a join. There are three definitions for a common boundary: Rook's Case, Queen's Case and Bishop's Case (Fig. 2). We used Rook's case to determine adjacency in this paper. The possible types of joins are black–black (BB), black–white (BW), and white–white (WW). Join counts are counts of the numbers of BB, BW, and WW joins in the study area, and these numbers are compared to the expected numbers

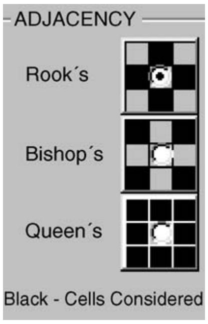

Fig. 2. Three types of contiguity.

of BB, BW and WW joins under the null hypothesis of no spatial autocorrelation to determine if spatial autocorrelation exists.

We then implemented one GLM that ignores the influence of spatial autocorrelation, referred to as model 0, and three GLMMs that incorporate three different spatial covariance structure models to derive the probability of cloud cover. The three spatial structure models which account for spatial autocorrelation are the spherical model (referred to as Model I, Eq. (1)), the exponential model (referred to as Model II, Eq. (2)) and the Gaussian model (referred to as Model III, Eq. (3)), which have been found to be the most useful ones for georeferenced data analysis purposes (Griffith and Layne, 1999, pp. 134–135). The independent variables we used are all related to topography—aspect, slope and the difference between elevation above sea level and lifting condensation level (lcl). For each remote sensing image, we derived a cloud mask using the procedures described in the "cloud identification" section (cloudy areas are represented by the number '1', non-cloudy areas are assigned a value of '0'), and then we used the cloud mask to calibrate the four models to derive the probability of cloud cover. We then compared Akaike Information Criterion (AIC), scaled deviance and extra-dispersion statistics of these four models to decide the best fit model. The range and sill estimates in the spatial structure models were derived in the GIS software "IDRISI" from the semivariograms based on the residuals of the standard logistic regression model and were put in the statistical models so the models started in the vicinity of plausible values of sill and range. Otherwise, the model may converge to local maxima and provide grossly unreasonable estimates (Littell et al., 1996).

$$\gamma(h) = \begin{cases} 0 & \text{if } |h| = 0 \\ C_0 + C_1[1.5(h/r) - 0.5(h/r)^3] & \text{if } 0 < |h| < r \\ C_0 + C_1 & \text{if } |h| \geq r \end{cases} \quad (1)$$

$$\gamma(h) = \begin{cases} 0 & \text{if } |h| = 0 \\ C_0 + C_1[1 - \exp(-h/r)] & \text{if } |h| > 0 \end{cases} \quad (2)$$

$$\gamma(h) = \begin{cases} 0 & \text{if } |h| = 0 \\ C_0 + C_1[1 - \exp(-h^2/r^2)] & \text{if } |h| > 0 \end{cases} \quad (3)$$

where γ is the semivariance, h the distance class or lag, r the range parameter, C_0 the nugget effect and C_1 is the sill.

4.4.1. Deriving the lifting condensation level (lcl)

We calculated the temperature at which all air parcels will condense according to the AWIPS method (http://meted.ucar.edu/awips/validate/li_as.htm). The required air temperature and dew temperature data are from the continuous records at the El Verde station at 350 m in elevation in the Luquillo Experimental Forest. Then we derived the lifting condensation level (LCL) by calculating how high the elevation would be necessary to reach that dew point temperature (saturation temperature) using an adiabatic lapse rate of 0.54°/100 m for January and March and 0.64°/100 m for July. We got the three values from a linear relation between daily mean air temperature and elevation using temperature data at 10 locations along a windward elevation gradient in the LEF in January, March and July. (http://luq.lternet.edu/data/temp/bistempdata/Bis-temp.htm).

4.4.2. Calculating the probability of cloud coverage

The coefficients of the independent variables in the GLM(M)s were estimated through pseudo-likelihood procedures described in Wolfinger and O'Connell (1993) using the SAS macro "GLIMMIX". The macro uses the procedure "PROC MIXED" in SAS and the output delivery system, requiring SAS/STAT and SAS/IML release 6.08 or later.

Since LCL changed with each hour, we got hourly probabilities of cloud cover (P) for each cell over a day. The relation between the probability of the cloud cover and the difference between the elevation and the lcl, the aspect and the slope in the GLM(M)s which are linked to a Binomial function is formed as Eq. (4) (Agresti, 1990)

$$\ln \frac{P}{1-P} = \beta_0 + \beta_1(\text{Elev} - \text{LCL}) + \beta_2(\text{Asp}) + \beta_3(\text{Slp}) \quad (4)$$

Thus the probability of the cloud cover can be calculated by Eq. (5)

$$P = \frac{e^{\beta_0+\beta_1(\text{Elev}-\text{LCL})+\beta_2(\text{Asp})+\beta_3(\text{Slp})}}{1 + e^{\beta_0+\beta_1(\text{Elev}-\text{LCL})+\beta_2(\text{Asp})+\beta_3(\text{Slp})}} \quad (5)$$

(P is the probability of cloud cover as a function of the difference between elevation and lifting condensation level, aspect and slope; β_0, β_1, β_2 and β_3 are the regression coefficients in the GL(M)Ms.)

4.4.3. Validation

We are more interested in the feasibility of calibrating the models using MODIS images to derive the probability of cloud cover than using Landsat-7 ETM+ images, due to the high temporal resolution and free distribution of the MODIS images. Thus we focused on validating the probability of cloud cover calibrated by MODIS images rather than the Landsat-7 ETM+ images.

We validated the model in two ways: first we calculated the proportion of areas covered by clouds at each 100 m elevation interval in the LEF from Landsat-7 ETM+ images. We interpreted this as the probability of cloud cover at each 100 m elevation at about 10:30 am when the Landsat images were obtained. Then we compared these observed results to the probability of cloud cover at 10:30 am derived from the best fit GL(M)M calibrated by the MODIS images at the same elevation intervals. Second, we calculated the hourly solar radiations available for photosynthesis and evapotranspiration during the day using a solar radiation model (Wu et al., 2005) that incorporated the derived probability of cloud cover calibrated by the MODIS images and compared them to the measured data at the El Verde meteorological station in the LEF.

We used "Index of Agreement (IoA)" (Eq. (6)) to evaluate the model performance (Janssen and Heuberger, 1995; Willmott, 1981)

$$\text{IoA} = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P'_i| + |O'_i|)^2} \quad (6)$$

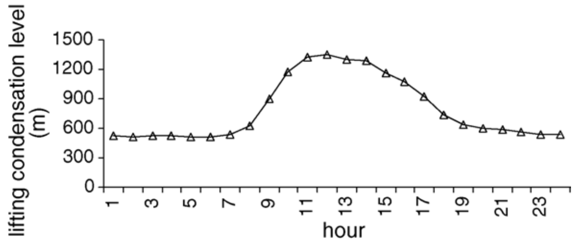

Fig. 3. Modeled lifting condensation level over a day in the Luquillo Mountains in January.

where P_i is the predicted values, O_i the observed value, $\overline{P}$ the mean of the predicted values, $\overline{O}$ the mean of the observed values, $P'_i = P_i - \overline{P}$, $O'_i = O_i - \overline{O}$.

IoA measures the agreement between predictions and observations on an individual level. This dimensionless index has limits of 0 (indicating no agreement) and 1 (indicating perfect agreement).

5. Results

The best fit model is the GLMM with the exponential spatial structure and the GLM according to the statistics of AIC, extra-dispersion scale and scaled deviance. The probabilities of cloud cover derived from the models was high at night, decreased in the morning after the sun rose until early afternoon and then increased in the afternoon and through night, in response to the movement of the lifting condensation level (Fig. 3). Since the variable "the difference between the elevation and the lifting condensation level" is always the most significant factor and the coefficient associated with it is positive, the probability of cloud cover increased with elevation, in agreement with the empirical data on the ground and the independent cloud masks derived from the Landsat-7 ETM+ images.

For spatial validation, we found our model could predict the probability of cloud cover every 100 m in elevation at a certain time with an index of agreement (IoA) of 0.560–0.919 (Table 6), which indicates a medium to high agreement between model simulation and observations at that particular time. For temporal validation, the IoA is from 0.940 to 0.994 (Table 7), which indicates a good model simulation at that particular location.

Table 2
Summary of join-count statistics on the cloud masks derived from the MODIS images. (B represents 0, i.e. non-cloud; W represents 1, i.e. cloud)

	BW joins/expected number/z-statistics	BB joins/expected number/z-statistics	WW joins/expected number/z-statistics
27 March 2000	615/2687.07/−67.81	5794/4772.24/63.81	1428/377.69/68.94
9 January 2001	961/3852.20/−66.46	2892/1478.98/64.19	3984/2505.82/66.59
20 July 2001	929/2647.03/−57.06	5680/4827.66/53.96	1228/362.31/57.69
4 March 2003	952/3124.73/−61.34	5189/4120.64/58.15	1696/591.31/62.24

5.1. Spatial autocorrelation in the cloud mask

Significant positive spatial autocorrelations existed in the cloud masks derived from both the MODIS images and the Landsat-7 ETM+ images on the 4 days sampled. We can say this since the cloud–cloud neighborhood pairs are beyond the 95% confidence intervals of the expected numbers of cloud–cloud neighbor-pairs under the assumption of spatial randomness according to the join-count statistics (Table 2). Significant spatial autocorrelations are also found in the significant z-statistics (Table 2).

5.2. Comparison of cloud masks from "unsupervised classification" and the threshold method

The cloud masks derived from the two methods are at moderate to high similarity according to the kappa statistics of 0.75 and 0.70 on 9 January 2001 and 20 July 2001, respectively (derived from Tables 3 and 4). Since the threshold method gives similar cloud masks, we were able to use the cloud masks derived from unsupervised classification in Erdas to calibrate the models. However we cannot eliminate the uncertainty introduced in the model based on different methods of cloud masking.

5.3. Assessment of the different forms of the statistical models

We needed to select the best fit model from among the four models we tried for predicting the probability of cloud cover. However, there are no formal criteria we can use to choose the best model in the case of using the GLIMMIX macro to estimate the parameters (Phil Gibbs, consultant of SAS Institute, personal contact). Usually the less the scaled deviance is, the better the model is, but there is no guarantee the model we choose based on scaled deviance is the best model (Phil Gibbs, personal contact). AIC statistics are generally used to compare the fitness of the models: the smaller the AIC, the better fit the model. We also needed to consider extra-dispersion statistics from the output of GLIMMIX, if it is larger than 1, there exists overdispersion, the condition by which the variability of the data exceeds the variability expected under a particular probability distribution, which is a problem because it affects the estimated precision of the parameter estimates.

Among the four models, Model I (spherical) has the lowest scaled deviance, indicating we should choose

Table 3
Correspondence table between the two cloud masks derived from the two methods on 9 January 2001

Threshold	Unsupervised		
	Cloud absence	Cloud presence	Total
Cloud absence	101886	5498	107384
Cloud presence	26986	123294	150280
Total	128872	128792	257664

The row represents the cloud mask derived from unsupervised classification in Erdas, and the column repsrents the cloud mask derived from the threshold method. The number represents cell numbers.

Table 4
Correspondence table between two methods on the cell numbers of cloud presence and absence on 20 July 2001

Threshold	Unsupervised		
	Cloud absence	Cloud presence	Total
Cloud absence	163145	7967	171112
Cloud presence	24479	62073	86552
Total	187624	70040	257664

Table 5
Coefficient estimates for Model 0 and Model II

Image	Model	Intercept (β_0)	Slp (β_3)	Asp (β_2)	Elev-LCL (β_1)	AIC	Extra-dispersion	Scaled deviance
27 March 2000 Landsat ETM+ obtained at 10:36 am	Non-spatial	0.1746 (0.5597)	−0.04029 (<0.0001)	0.002182 (0.0030)	0.001732 (<0.0001)	10403.5	0.9950	2153.02
	Spatial exp (442.85, 0.9438)	−0.6661 (0.2172)	−0.00611 (0.5429)	0.000032 (0.9656)	0.000829 (0.1199)	8740.0	0.9438	2302.66
27 March 2000 MODIS on Terra obtained at 11:15 am	Non-spatial	1.3318 (<0.0001)	−0.08732 (<0.0001)	0.004970 (<0.0001)	0.002440 (<0.0001)	10317.6	0.9838	2270.67
	Spatial exp (155.60,0.9934)	−.0.7403 (0.2054)	−0.00023 (0.9711)	0.000046 (0.9195)	0.000376 (0.3390)	6625.4	0.9934	2432.78
9 January 2001 Landsat ETM+ obtained at 10:30 am	Non-spatial	0.7300 (0.0019)	−0.02826 (<0.0001)	0.004915 (<0.0001)	0.000716 (0.0009)	9623.7	1.0050	2752.86
	Spatial exp (428.48, 0.9433)	1.0760 (0.0119)	0.003007 (0.7257)	0.000258 (0.6823)	0.001045 (0.0219)	8023.5	0.9433	3011.54
9 January 2001 MODIS on Terra obtained at 11:10 am	Non-spatial	2.7111 (<0.0001)	0.01526 (0.0427)	−0.00236 (0.0003)	0.002552 (<0.0001)	9865.5	0.9867	2583.85
	Spatial exp (700.79, 0.9980)	1.2139 (0.0192)	0.003156 (0.6773)	0.000176 (0.7478)	0.000823 (0.0668)	7379.6	0.9980	2637.04
20 July 2001 Landsat ETM+ obtained at 10:30 am	Non-spatial	−0.5023 (0.0718)	0.01729 (0.0163)	−0.00190 (0.0034)	0.000470 (0.0330)	9752.3	1.000	2634.29
	Spatial exp (405.37, 1.0008)	0.009335 (0.9858)	−0.00892 (0.3465)	−0.00017 (0.8034)	0.000935 (0.0552)	8499.9	2652.83	1.0008
20 July 2001 MODIS on Terra obtained at 11:05 am	Non-spatial	−0.2562 (0.4021)	0.01839 (0.0172)	−0.00260 (0.0002)	0.001031 (<0.0001)	10112.4	2353.28	0.9930
	Spatial exp (516.70, 1.0194)	−0.5534 (0.3469)	−0.01786 (0.0638)	0.000475 (0.4901)	0.000488 (0.3441)	8407.2	2341.85	1.0194
4 March 2003 Landsat ETM+ obtained at 10:32 am	Non-spatial	3.1246 (<0.0001)	−0.02545 (0.0004)	−0.00070 (0.2680)	0.003608 (<0.0001)	9647.1	1.0042	2733.52
	Spatial exp (430.13, 0.9059)	0.7599 (0.0895)	−0.00813 (0.3151)	0.001001 (0.0940)	0.001242 (0.0041)	7780.8	0.9059	3166.21
4 March 2003 MODIS on Terra obtained at 11:00 am	Non-spatial	1.7182 (<0.0001)	0.01002 (0.1771)	0.001724 (0.0088)	0.003394 (<0.0001)	9821.7	1.0010	2621.92
	Spatial exp (483.00, 0.9631)	0.3467 (0.6905)	−0.00186 (0.8246)	0.000070 (0.9098)	0.001269 (0.0066)	7950.9	0.9631	2894.60

Spatial exp(#1, #2): exponential spatial structure with #1 as range and #2 as sill., The numbers in the parentheses under coefficient estimates are the *p*-values of the estimates.

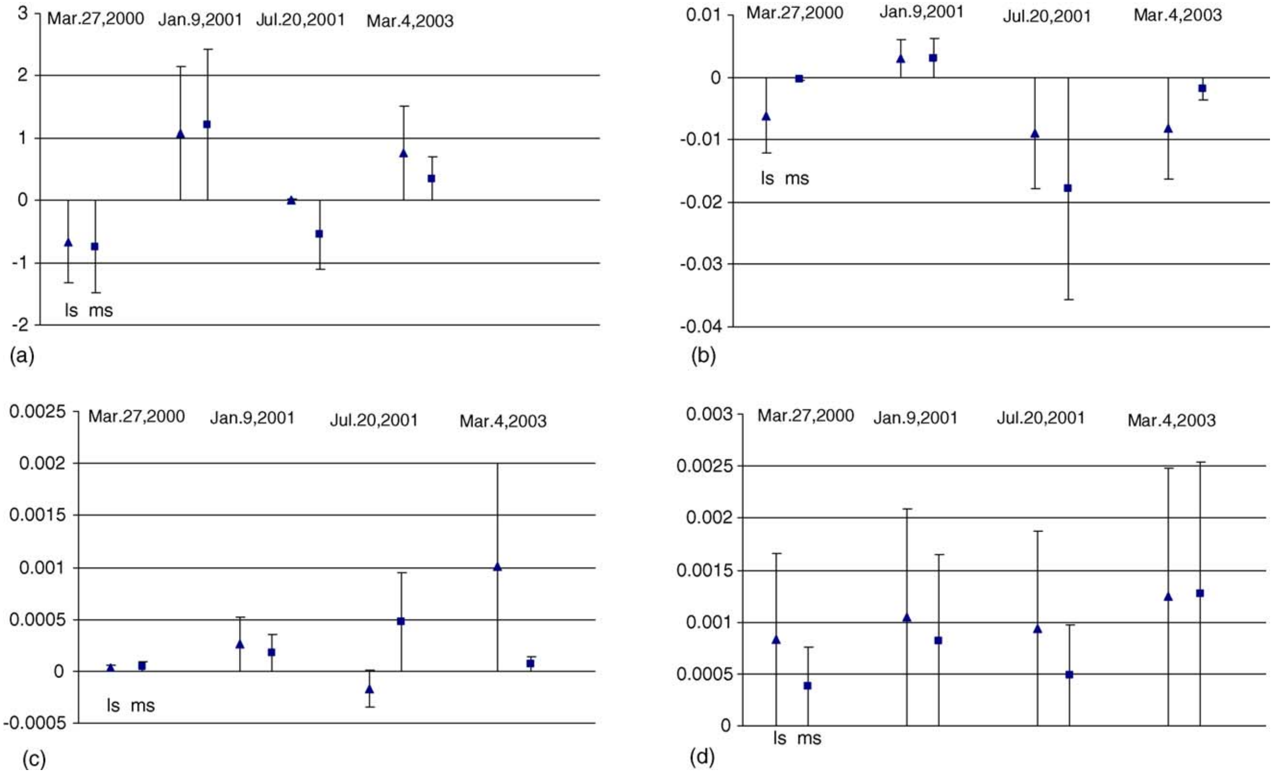

Fig. 4. (a) Comparison of the estimates of the intercepts in the model with associated standard errors of Model II calibrated by Landsat images and MODIS images on four different days. (Triangle: Landsat images, i.e. ls; square: MODIS images, i.e. ms). (b) Comparison of the estimates of the "slope" coefficients in the model with associated standard errors of Model II calibrated by Landsat images and MODIS images on four different days. (Triangle: Landsat images, i.e. ls; square: MODIS images, i.e. ms). (c) Comparison of the estimates of the "aspect" coefficients in the model with associated standard errors of Model II calibrated by Landsat images and MODIS images on four different days. (Triangle: Landsat images, i.e. ls; square: MODIS images, i.e. ms). (d) Comparison of the estimates of the coefficients of "difference between elevation and lcl" in the model with associated standard errors of Model II calibrated by Landsat images and MODIS images on four different days. (Triangle: Landsat images, i.e. ls; square: MODIS images, i.e. ms).

it, however, the extra-dispersion statistics is larger than 1, showing the existence of overdispersion. In addition, the AIC statistics from Model I are the highest among the three GLMMs. Model 0 can be a candidate for best fit model because it has next lowest scaled deviance among the four GL(M)Ms, and the extra-dispersion statistics are very close to 1, showing no problem of overdispersion, although the AIC statistics are the highest among the four models. Among the three GLMMs, Model II (exponential) has the lowest AIC statistics, shows no problem of overdispersion and has the next lowest scaled deviance, so it too can be a candidate for best fit model. Thus we chose Model 0 and Model II as the best models, but we cannot tell which is better. From the coefficient estimates of the two best models, the most significant parameter is consistently the difference between elevation and lifting condensation. In addition, the higher the elevation above the lifting condensation level, the higher the probability of cloud cover according to the positive sign of the derived coefficients (Table 5). The absolute values of the estimates of the coefficients in Model II are always smaller than, and not as significant as (*p*-values are larger), the estimates in Model 0 without spatial structure, because part of the variation is already explained by the spatial autocorrelation (Overmars et al., 2003). The exceptions are (1) the estimate of the coefficient of the aspect variable calibrated by the Landsat-7 ETM+ image on March 4 of 2003, and 2) the estimates of the coefficients of "Elev-LCL"s (the difference between elevation and LCL) calibrated by the Landsat-7 ETM+ image on 9 January 2001 and on 20 July 2001. The derived cloud probabilities in the following sections of this paper are all derived from these two "best" models except when we specify otherwise.

Comparisons of the coefficient estimates and the associated standard errors shows that the confidence

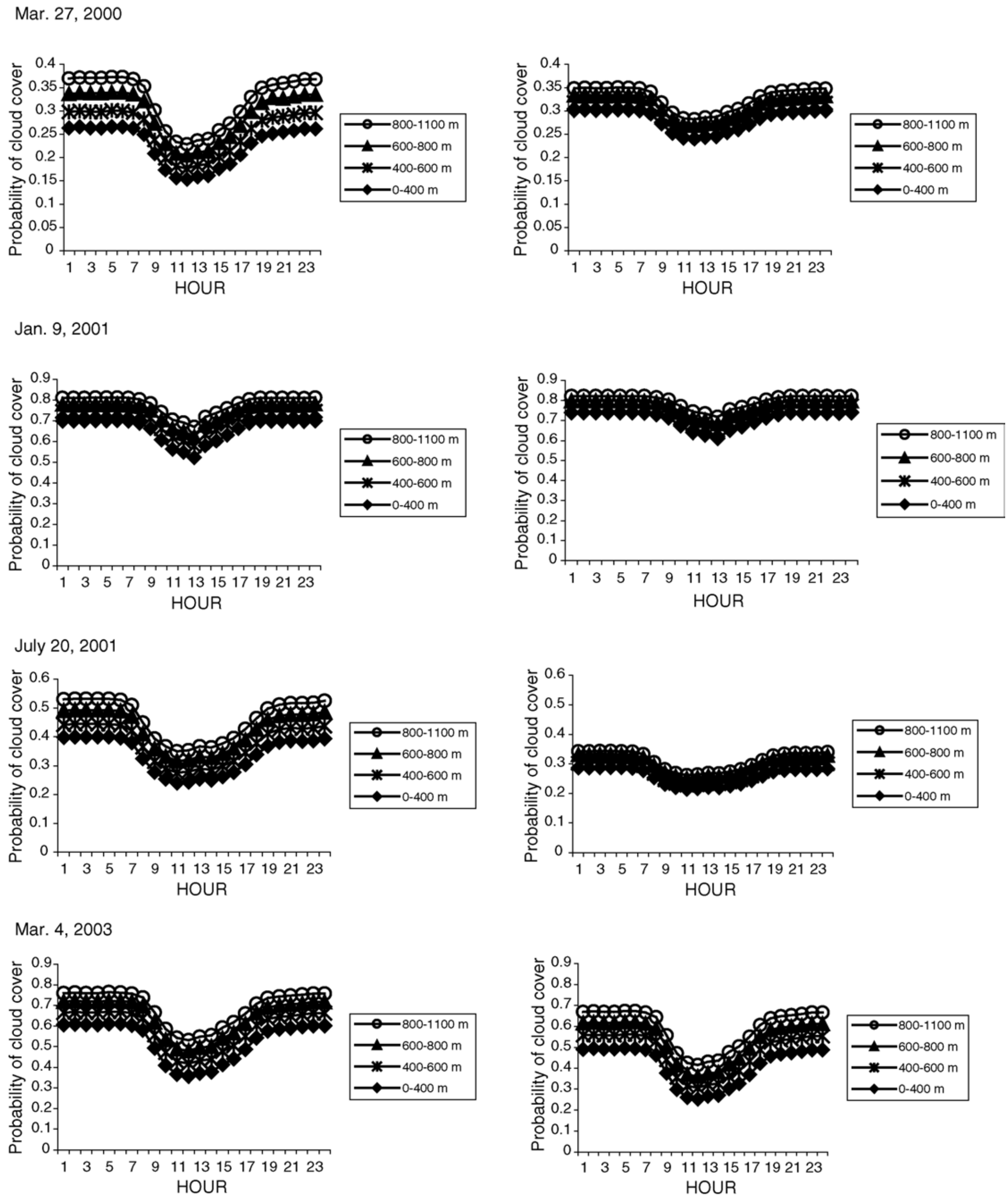

Fig. 5. Hourly probability of cloud cover over a day derived from the GLMM with the exponential spatial structure model calibrated by the Landsat image (left) and MODIS image (right) in the LEF.

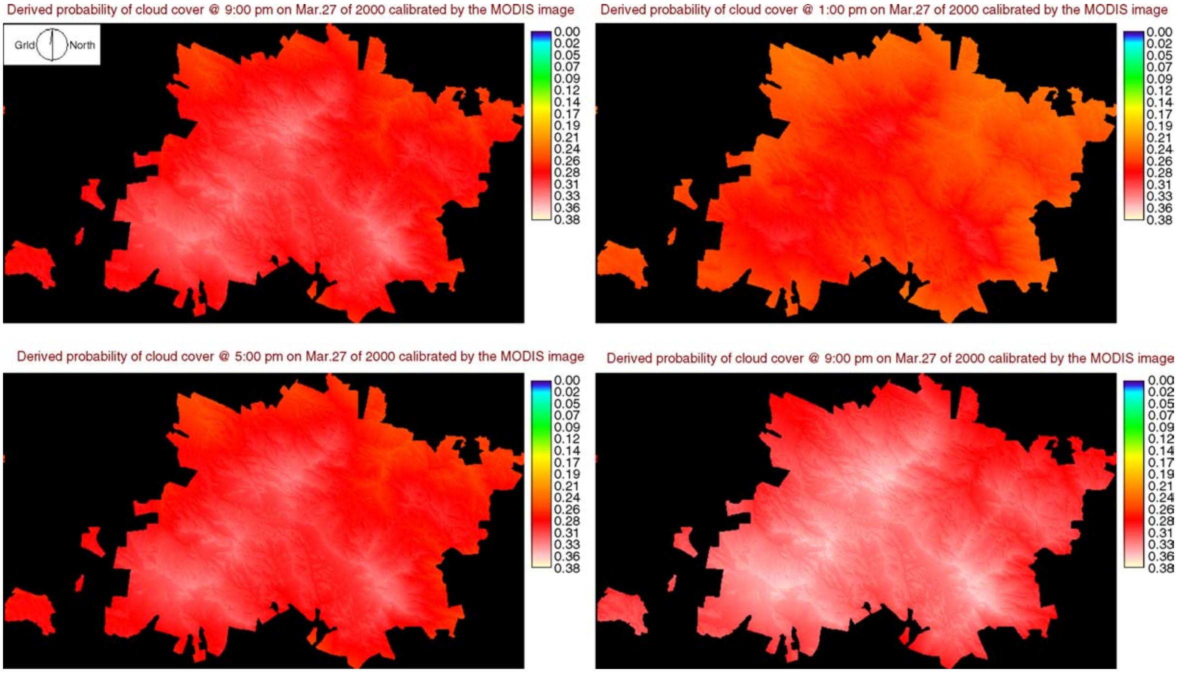

Fig. 6. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from the GLMM with exponential structure on 27 March 2000 calibrated by the MODIS image.

intervals of most of the coefficient estimates calibrated from MODIS images included the values of the estimates of the same coefficients calibrated from ETM+ images except the coefficients of "slope" from 27 March 2000 (p-value = 0.97) and from 4 March 2003 (p-value = 0.82), and the coefficient of "aspect" on 20 July 2001 (p-value = 0.49) and 4 March 2003 (p-value = 0.91). The exceptions are to the non-significant coefficients. This indicates no statistically significant difference overall in the estimated coefficients between the two calibrations (Fig. 4a and d), which justified our using more frequent and free MODIS images instead of Landsat images to calibrate the model. But we need to keep in mind this result is only for LEF and certain days, and it is not certain if it can apply to other places of the world and other days.

5.4. Results of the hourly probability of cloud cover

The derived lifting condensation level (LCL) moved up the mountain until early afternoon and then went down the mountain in the afternoon (Fig. 3) and at night. This is consistent with Brisco's derivation that was based on the independent meteorological data at Cape San Juan, which is the most northeastern portion of Puerto Rico, 12 miles northeast of El Yunque peak in the LEF and adjacent to the light station on a low hill (Briscoe, 1966, pp. 9a, Odum et al., 1970, pp. B349-B350). The derived probabilities of cloud cover decreased in the morning until early afternoon and increased in the afternoon and through night, apparently in response to the movement of the lifting condensation level. This pattern was also shown in the figures of hourly probability of cloud cover over a day derived from Model II calibrated by the Landsat-7 ETM+ image and the MODIS image on the 4 days temporarily (Fig. 5) and spatially (Figs. 6–9). This is also true for the hourly probability of cloud cover derived from Model 0.

6. Validation

We used two methods of validation since the derived probability of cloud cover is distributed both spatially

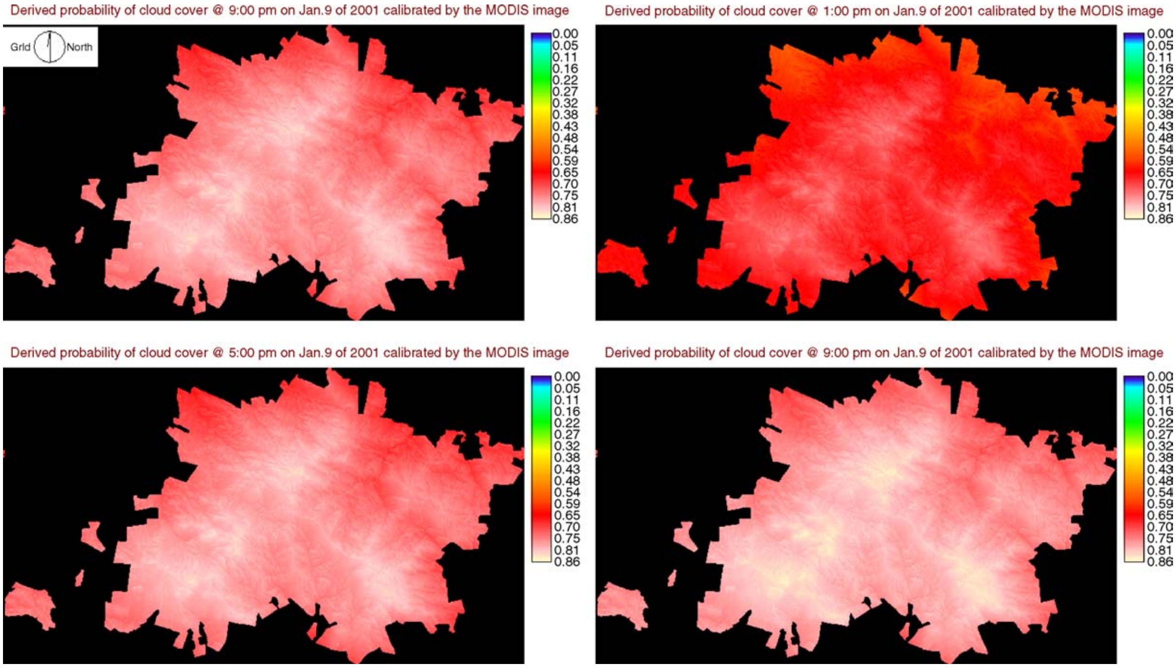

Fig. 7. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from the GLMM with exponential structure on 9 January 2001 calibrated by the MODIS image.

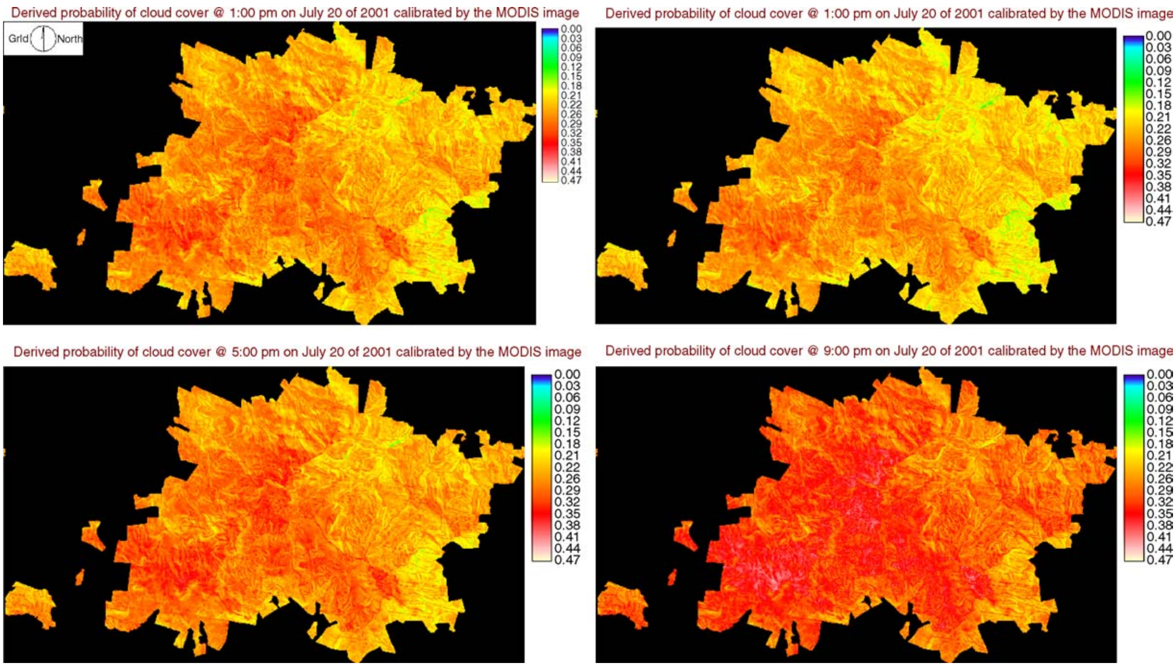

Fig. 8. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from the GLMM with exponential structure on 20 July 2001 calibrated by the MODIS image.

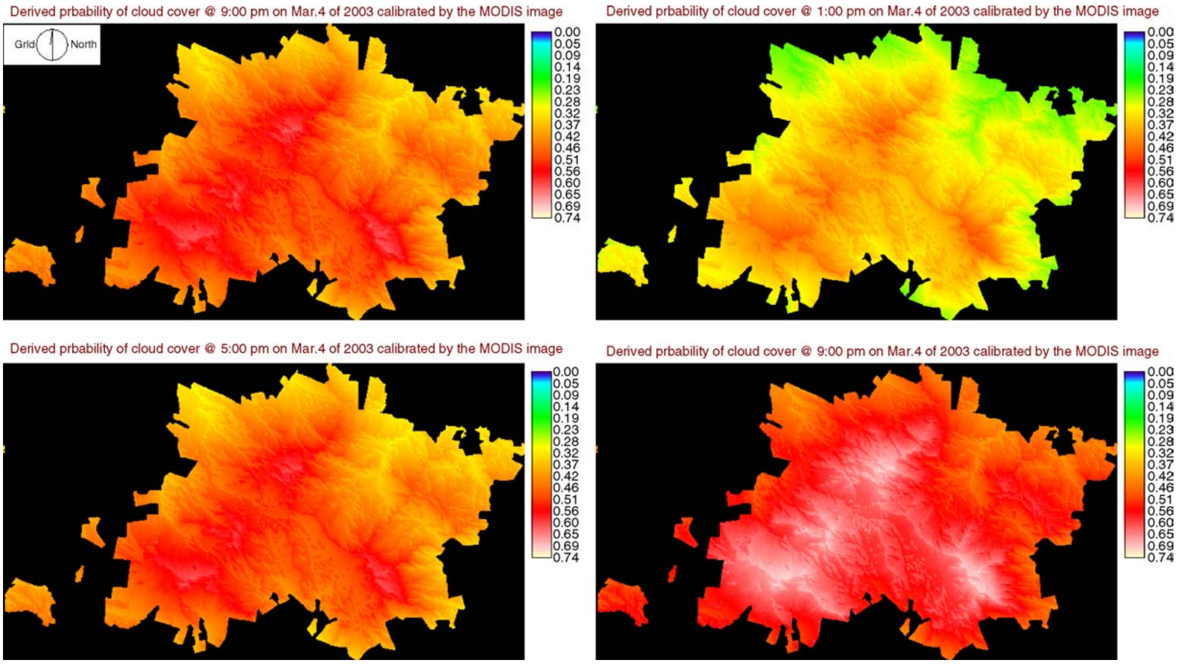

Fig. 9. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from the GLMM with exponential structure on 4 March 2003 calibrated by the MODIS image.

and temporally. One method measures spatial accuracy but only at a certain time, the other measures temporal accuracy but only at a certain location in the LEF. It was not possible to do validation both spatially and temporally simultaneously because of the limited independent data available for validation.

For spatial validation, we found our model could predict the probability of cloud cover every 100 m in elevation at a certain time with an index of agreement of 0.560–0.919 (Table 6), which indicates a medium to high agreement between model simulation and observations at that particular time. We calculated the proportions of the areas that are covered by clouds for every 100 m interval in elevation based on the cloud masks derived from Landsat-7 ETM+ images on each of the 4 days, and then we compared the results observed to the derived probabilities of cloud cover at the same elevation intervals calibrated by MODIS images (Fig. 10a and d) at the same time. Unfortunately the comparison does not look great since our modeled results are not about the presence or absence of clouds, but the probabilities of cloud cover. For example, the mountain tops have "80% probability of cloud cover" our results show a consistent 80% cover, not 80% clouds and 20% clear, as is the case from the Landsat images. However, we could get the impression that the higher probability of cloud cover was where clouds existed. On 27 March 2000 (Fig. 11a), the probabilities of cloud cover derived from Model 0 were closer to the

Table 6
The IoA of Model II and Model 0 in spatial validation using Landsat images

		Model II	Model 0
27 March 2000	All elevations	0.725	0.722
	Below 600 m	0.0319	0.928
	Above 600 m	0.940	0.658
9 January 2001	All elevations	0.767	0.632
	Below 600 m	0.970	0.938
	Above 600 m	0.599	0.493
20 July 2001	All elevations	0.560	0.585
	Below 600 m	0.695	0.698
	Above 600 m	0.0223	0.0175
4 March 2003	All elevations	0.679	0.919
	Below 600 m	0.972	0.941
	Above 600 m	0	0.875

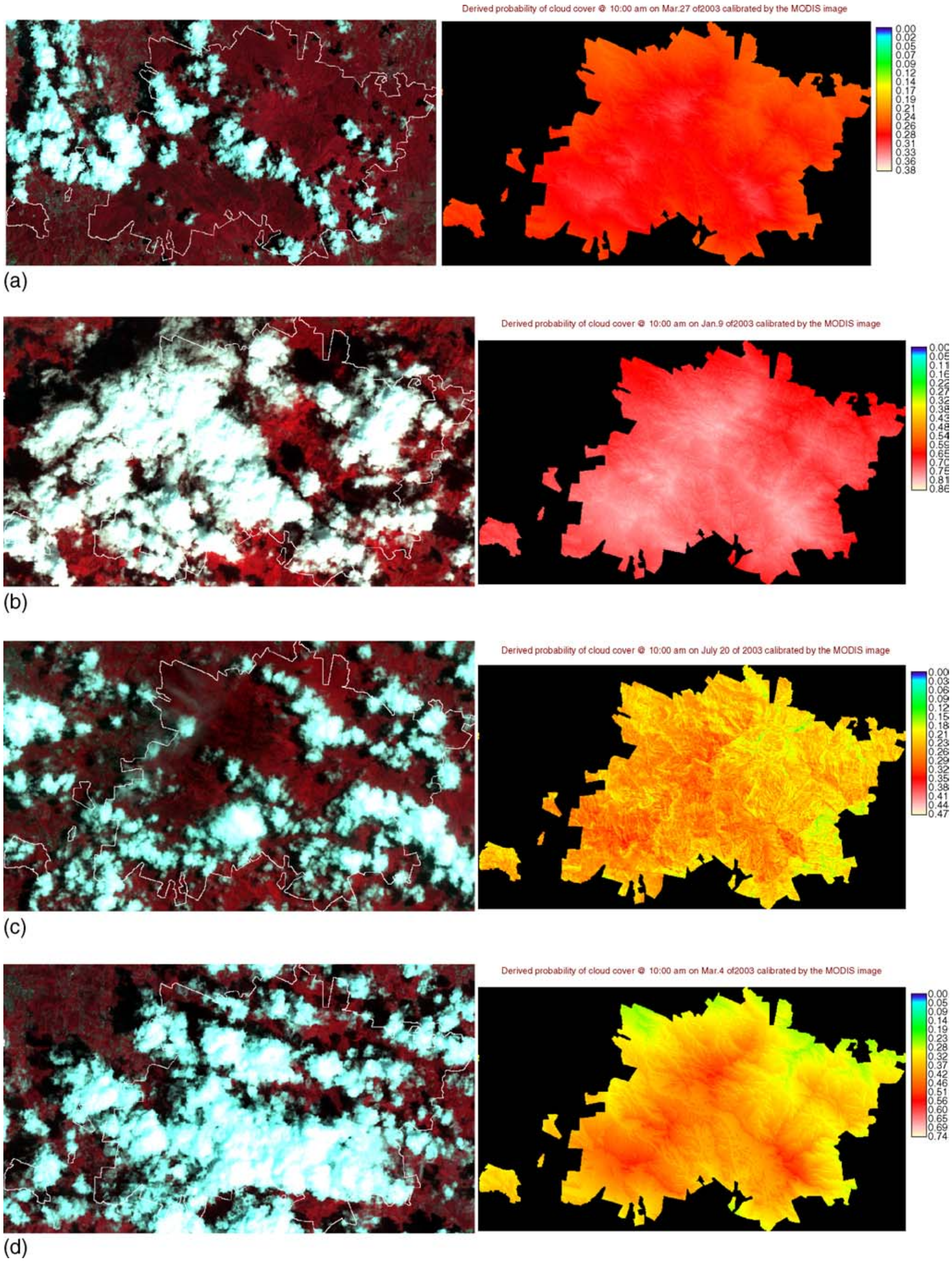

Fig. 10. Comparison between the original Landsat image (combinations of band 4, 3 and 2 displayed as blue, green and red) and the simulated probability of cloud cover calibrated by the MODIS image at the time when the Landsat image was obtained (cloudy areas had the higher probability of cloud cover; white lines on the Landsat images are the boundary of the LEF). (a) 20 March 2000, (b) 9 January 2001, (c) 20 July 2001, (d) 4 March 2003.

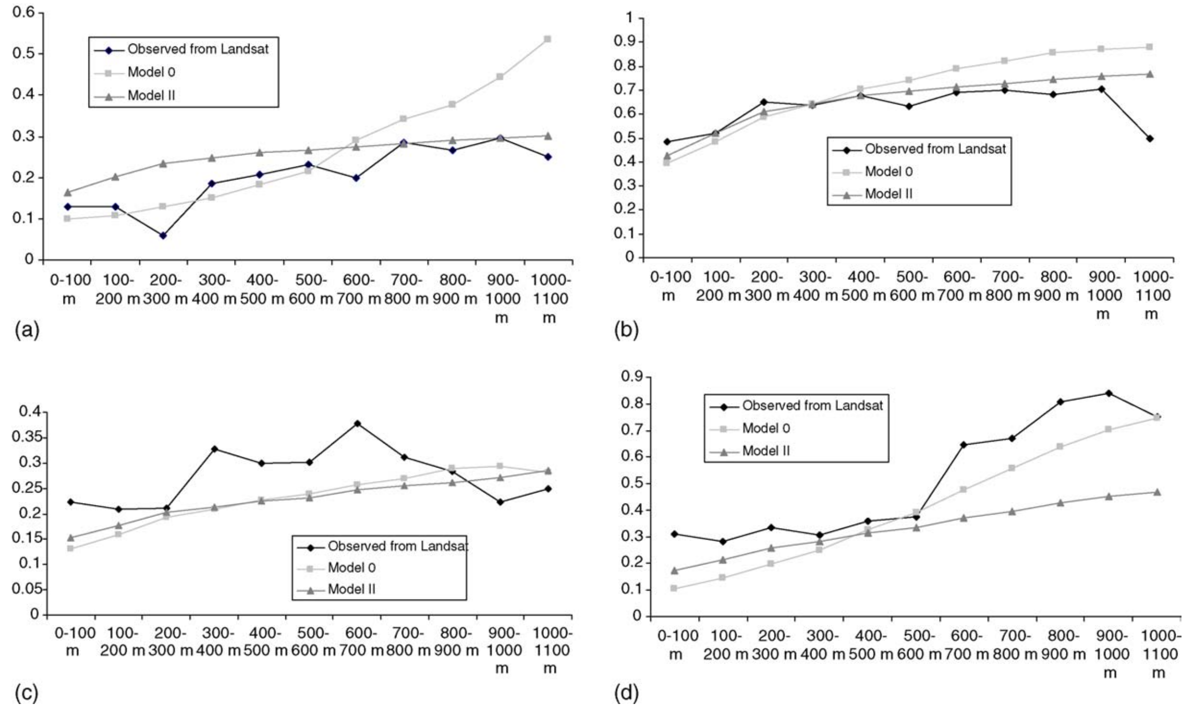

Fig. 11. Comparison of the probability of cloud cover as a function of elevation calculated directly from the cloud mask derived from the Landsat-7 ETM+ images and the ones predicted from Model II and Model 0 calibrated by the MODIS image. (a) 27 March 2000, (b) 9 January 2001, (c) 20 July 2001, (d) 4 March 2003.

actual proportions of cloud cover under 600 m in elevation, but the results from Model II gave a better estimate above 600 m elevation. Overall, the derived probabilities of cloud cover from Model II followed hourly empirical cloud cover proportions more closely than Model 0. On 9 January 2001 (Fig. 11b), the probabilities of cloud cover derived from Model II and calibrated with the MODIS images were closer to the observed percentages of cloud cover over the entire elevation ranges than was Model 0. On 20 July 2001 (Fig. 11c), the probabilities of cloud cover derived from Model 0 and II calibrated with the MODIS image were very similar (Model II a little bit better) and they both underestimated the proportion of cloud cover in the low to middle elevations (0–800 m) and overestimated a little in the high elevations. On 4 March 2003 (Fig. 11d), the probabilities of cloud cover derived from Model 0 and Model II both underestimated the empirical cloud cover proportion, but the underestimates were at a smaller scale below 600 m elevation and a much larger scale above 600 m elevation for the results from Model II compared with Model 0, so that Model 0 captured the observed pattern above 600 m better. Since the validation is only at one time, we could not draw a general conclusion about which model is better between Model 0 and Model II. The index of agreement for Model 0 and Model II are from 0.585 to 0.919 and from 0.560 to 0.767, respectively (Table 6), indicating a fair to good model simulation at that particular time. Neither Model 0 nor Model II consistently gave a better estimate than the other for both below and above 600 m in elevation. For example, on 27 March 2000, the IoA below 600 m is 0.928 for Model 0, much larger than the IoA of 0.0319 for Model II, but the IoA above 600 m is 0.658, smaller than the IoA of 0.940 for Model II, indicating Model II behaves better than Model 0 above 600 m, but not below 600 m. But the IoA too should be interpreted with caution. The IoA will invariably yield 0 if P'_i and O'_i (Eq. (6)) have opposite signs, irrespective of the sizes of deviations (Janssen and Heuberger, 1995). The IoA then falsely indicates that there is absolute no agreement between model predictions and observations, as in the case of above 600 m for Model II on 4 March 2003.

For the temporal validation, we calculated monthly average net solar radiation occurring on the surface using a solar radiation model (Wu et al., 2005) that included the monthly average cloud probability and compared it to the measured net solar radiation available for photosynthesis and evapotranspiration on the

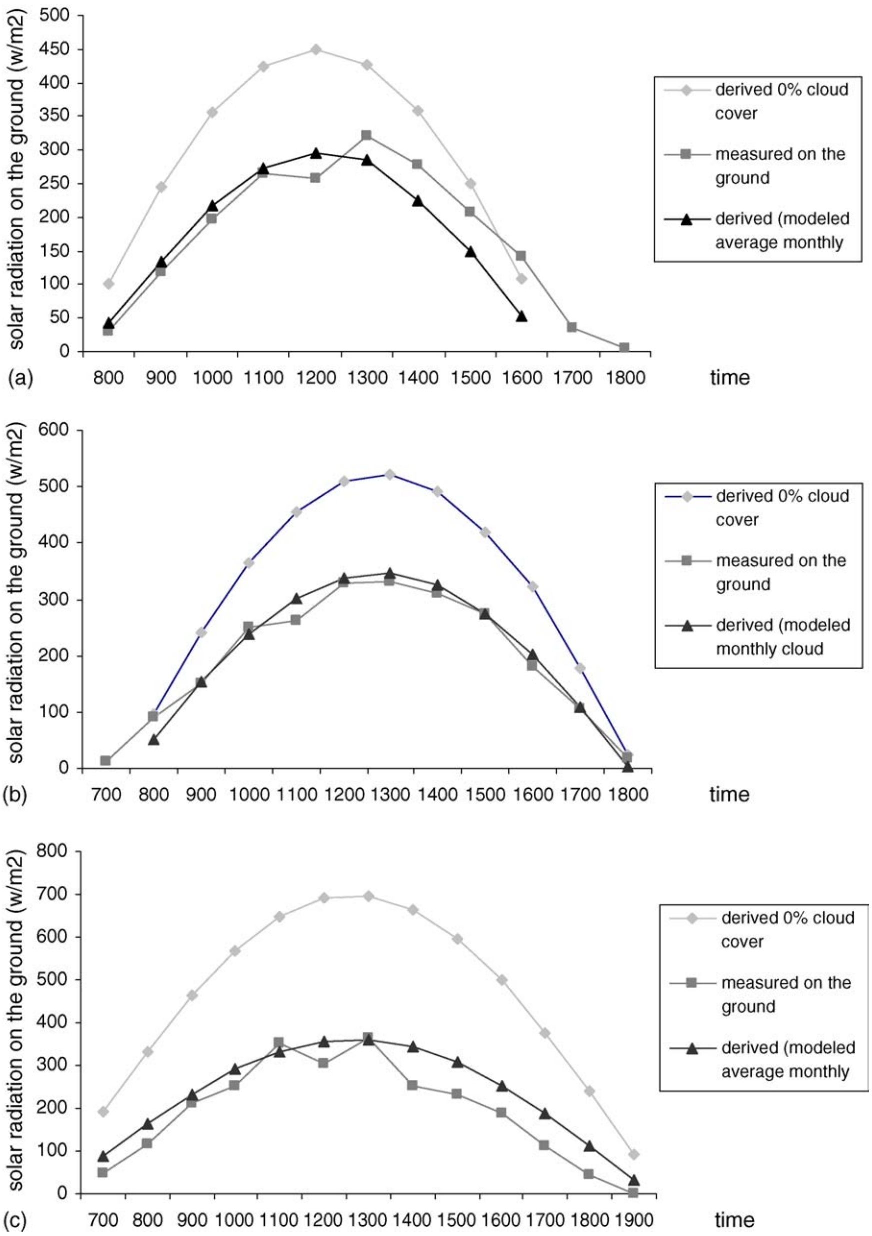

Fig. 12. (a) Validation of the modeled probability of cloud cover using the measured solar radiation on the ground in January (dry season). (b) Validation of the modeled probability of cloud cover using the measured solar radiation on the ground in March (dry season). (c) Validation of the modeled probability of cloud cover using the measured solar radiation on the ground in July (wet season).

ground. We could do this validation over time but only at one location: El Verde station (18°19′22″N, 65°49′13″ W) in the LEF.

Although Model II did not necessarily give better estimates than Model 0 for the probability of cloud cover at the time when the Landsat images were obtained, most of the time Model II worked better because the data on cloud cover had a significant positive spatial autocorrelation. Thus we used Model II when we calculated the average monthly probability of cloud cover in order to derive net solar radiation. We averaged the derived probability of cloud cover from Model II on 27 March 2000 and 4 March 2003 to get the average monthly probability of cloud cover for March. In order to calculate the monthly average probability of cloud cover in January, and July, we needed more MODIS images, so we added the scenes obtained on 8 January 2003, 4 January 2004, 18 January 2004, and 1 July 2000. The solar radiation that we derived by incorporating modeled monthly average probability of cloud cover is very close to the measured solar radiation on the ground (Fig. 12a and c). However our cloud model did not capture the frequent high cloud cover around noon, which can be seen in the measured data (i.e. the dip around noon). The brief higher cloud cover at noon was because of convectional lifting associated with the surface heating of the air at the ground surface which our cloud model did not incorporate. The solar radiation model did not predict the cloud-free solar radiation around the time of sunrise and sunset well. This caused the simulated cloud-free sunlight to be less than the measured sunlight. The IoA is from 0.940 to 0.994 (Table 7), which indicates a good model simulation at that particular location.

Table 7
The IoA of Model II in temporal validation using ground measured data at the El Verde station in the LEF

January	0.963
March	0.994
July	0.940

We also applied Model II to the other months to estimate the average probability of cloud cover of each month of a year. We found the derivation is intuitively justified since January to March, the relatively dry season, has less probability of cloud cover than in May, which is relatively rainy season (Figs. 13–19).

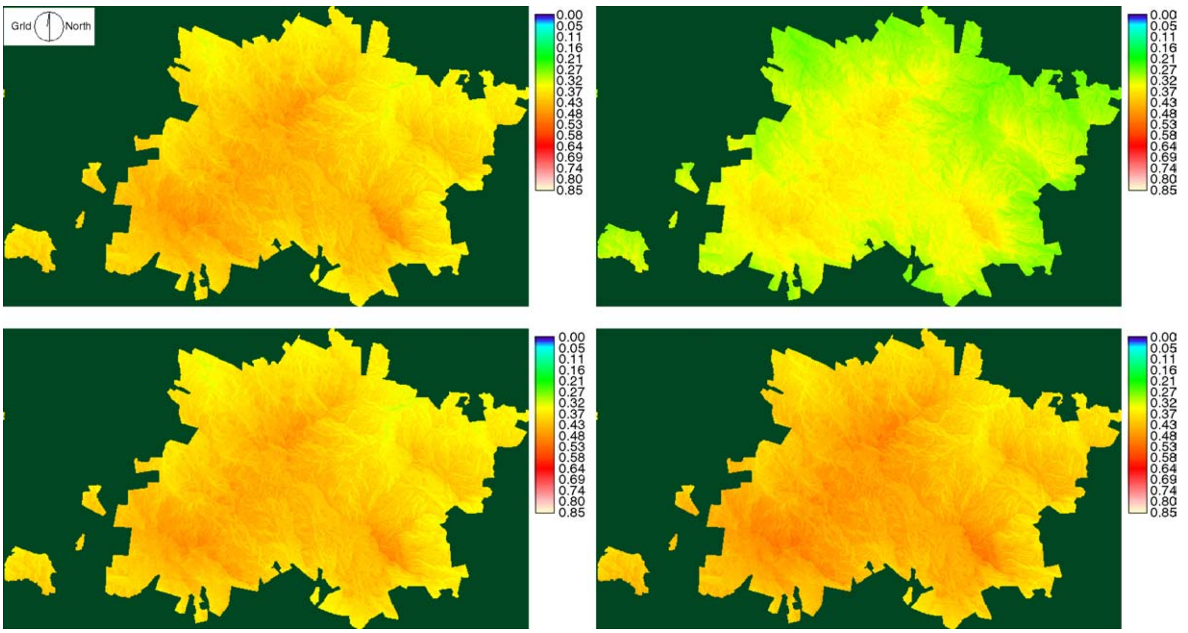

Fig. 13. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in January calibrated by the MODIS image.

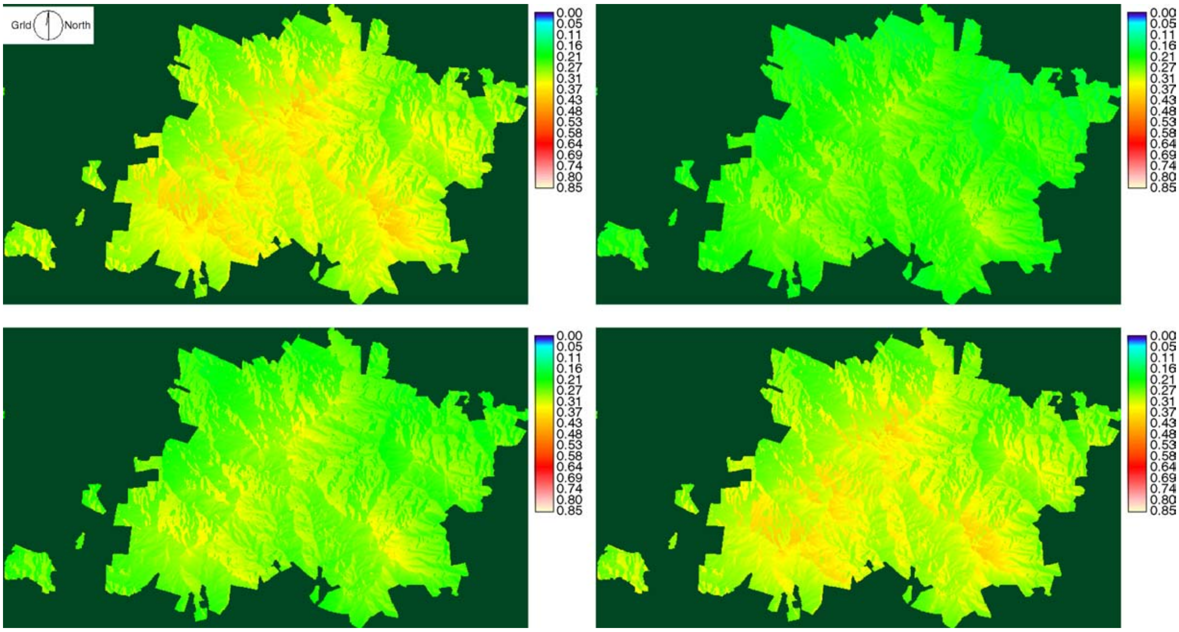

Fig. 14. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in February calibrated by the MODIS image.

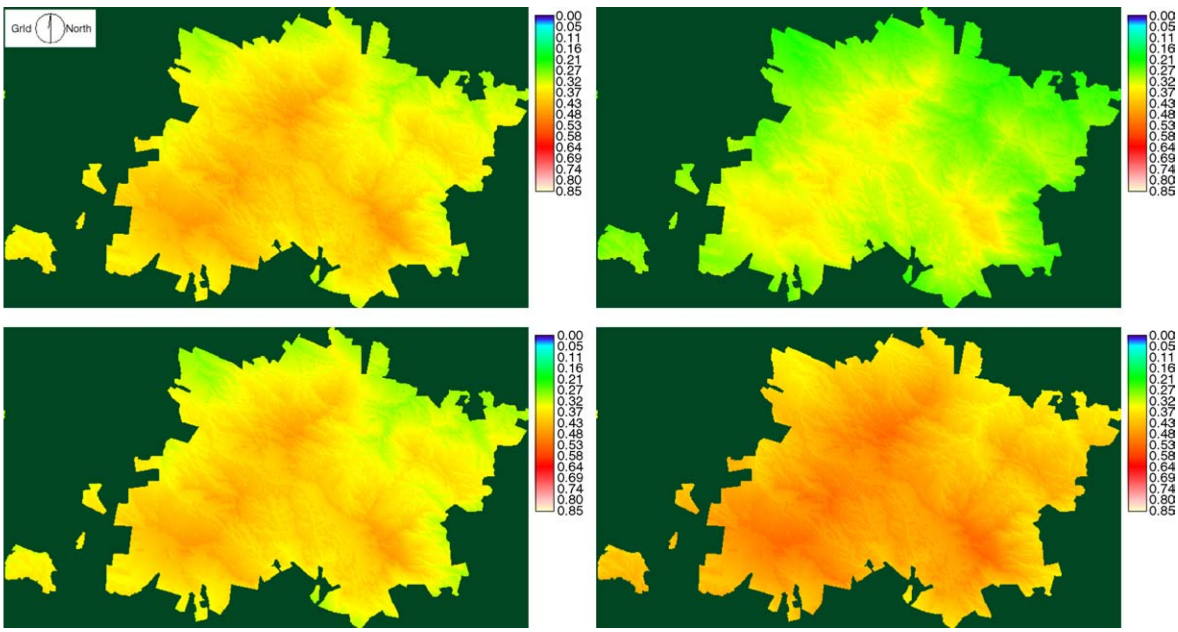

Fig. 15. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in March calibrated by the MODIS image.

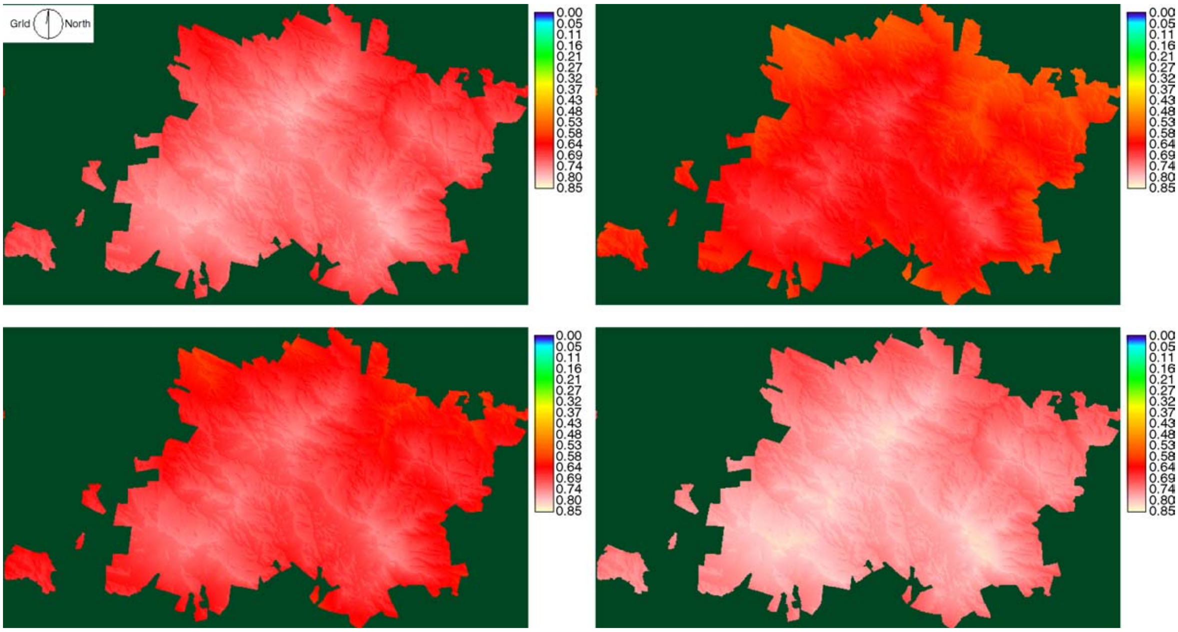

Fig. 16. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in May calibrated by the MODIS image.

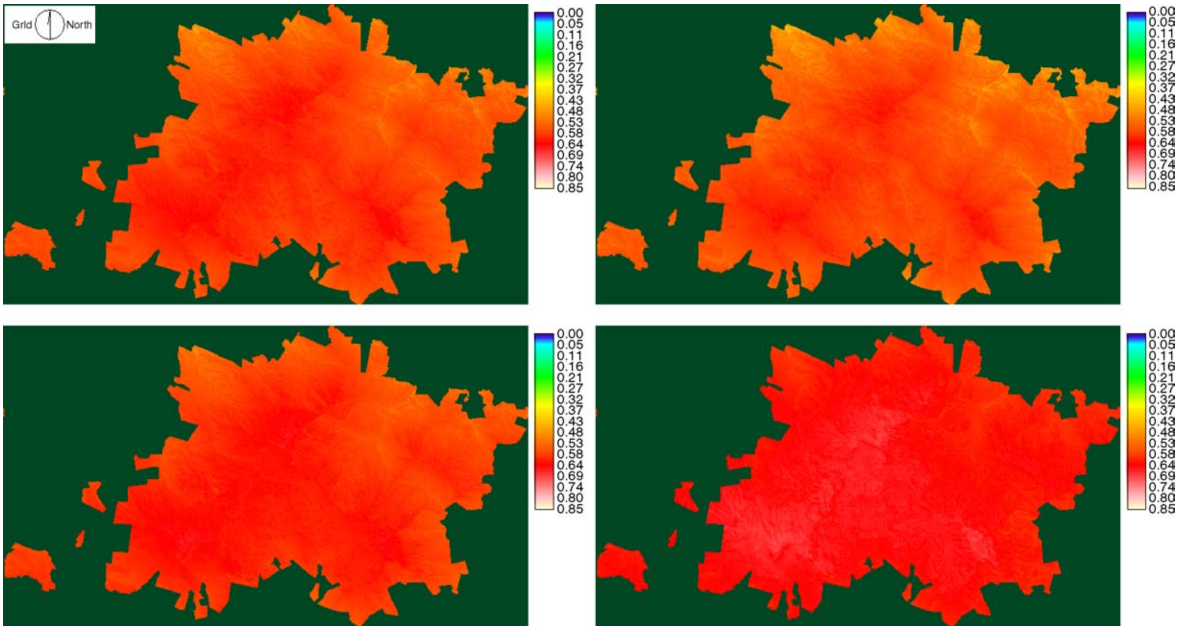

Fig. 17. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in July calibrated by the MODIS image.

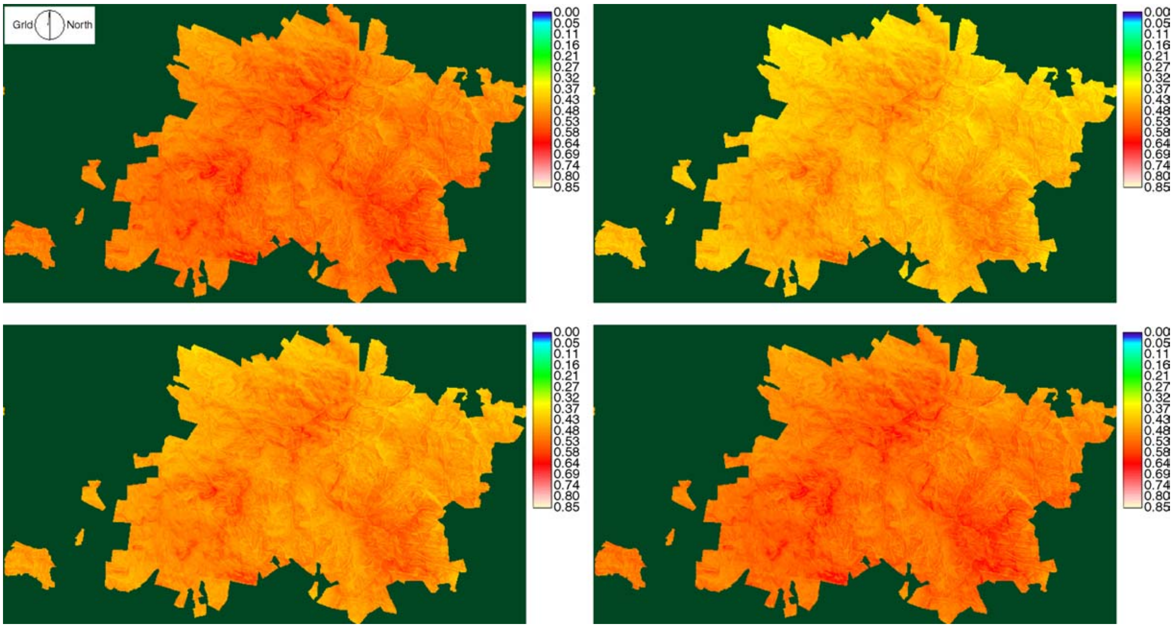

Fig. 18. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in September calibrated by the MODIS image.

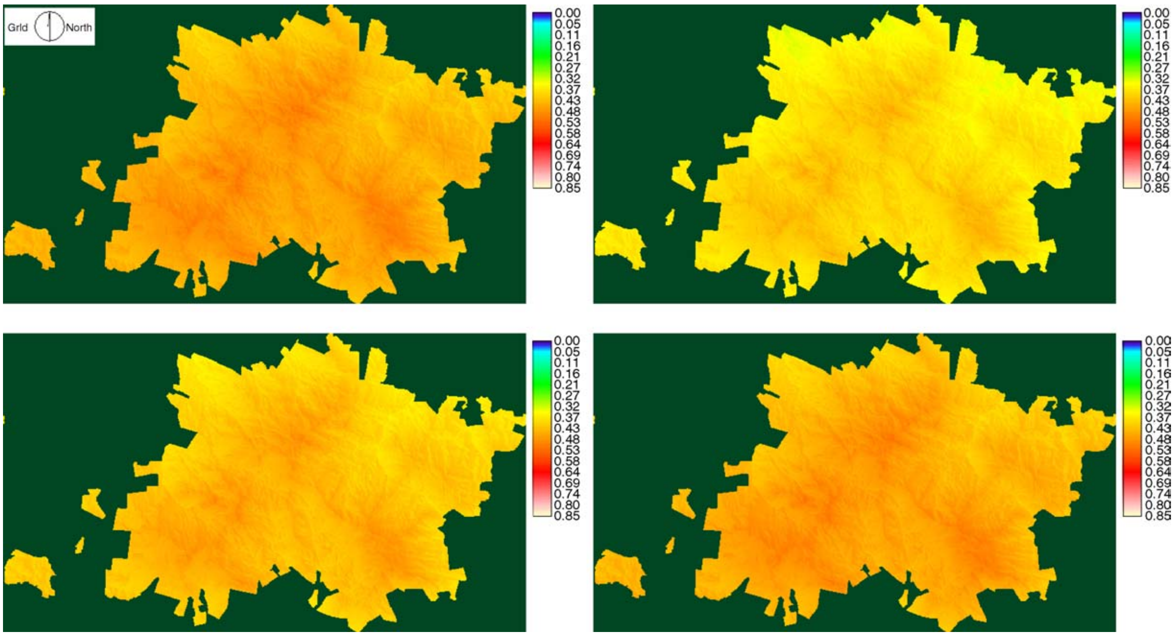

Fig. 19. Derived probability of cloud cover at 9:00 am, 1:00 pm, 5:00 pm, 9:00 pm from Model II in November calibrated by the MODIS image.

7. Discussion

The major reason that most of the coefficient estimates in Model II and Model 0 were not significantly different is based on the fact that the visible and infrared bands, which exist in both MODIS L1B images and Landsat-7 ETM+ images, are the two most important bands to differentiate clouds from vegetation. Clouds have a high reflectance in both visible and infrared bands while vegetation has a high reflectance in the near-infrared band but not in the visible band.

There are two main causes of spatial structure. First, spatial structure can be caused by a dependency of a dependent variable upon one or several independent variables, which are structured spatially, so that the pattern observed is in response to another variable. Second, spatial structure can result from the process that has produced the values of a dependent variable, and the variable is spatial in itself, so it reflects interactions among sites. Therefore it is useful to distinguish whether the dominant effects are caused by a reaction to external forces or by the interaction among neighboring individuals. When the major cause of spatial autocorrelation is a reaction to external forcing, a traditional regression is appropriate, whereas if the spatial autocorrelation is due to interactive effects there is a need for a model with a spatially-dependent covariance structure (Overmars et al., 2003). The probabilities of cloud cover on 4 March 2003 derived from Model 0 (which was not corrected for spatial autocorrelation) were better than from Model II especially at 600 m elevation and above (Fig. 11d), and the probabilities of cloud cover on 20 July 2001 derived from Model 0 and Model II were very similar even though there were significant spatial autocorrelations in the data of cloud cover on both days (Fig. 11c). But on the other 2 days, Model II gave better estimates of the probabilities of cloud cover (Fig. 11a and b). There are three reasons that may explain this. First, 600 m is a division elevation" in the LEF since it is the average lifting condensation level. Above that elevation, clouds exist nearly constantly, and below, the sky is relatively clear, so the requirement for homogeneity in the data set may be violated. In this situation, it is better to use an anisotropic (different spatial variation in different directions) model with two orthogonal direction components. But we used only the omni spatial model to quantify the spatial autocorrelation implicit in the cloud distribution, thus perhaps making the results from Model II projections not best. That also would explain why Model II gave good estimates most of time only in the elevations under 600 m or above 600 m but not both. Secondly, the proportion of the cloud cover on 4 March 2003 is much higher above 600 m elevation than below 600 m elevation, so it seems that elevation (which also has significant spatial structure) is such an important factor contributing to cloud formation that the spatial correlation in the cloud data is more a reaction to the independent variables such as elevation than the interaction among individual sites. Therefore the non-mixing model, Model 0, which does not consider spatial autocorrelation but is very sensitive to elevation is more appropriate. Thirdly, the smaller z values from join-count statistics on 20 July 2001 and 4 March 2003 show that spatial autocorrelations in the cloud data were not as strong as on 27 March 2000 and 9 January 2001 (Table 2). It is possible that the higher spatial autocorrelations of the cloud data on the latter 2 days were not only a reaction to elevation, but also due to the interaction among sites, so Model II that incorporated spatial autocorrelation was more appropriate to model the probabilities of cloud cover on those 2 days. On the other hand, join-count statistics assumes first-order homogeneity, that is, the probability of different categories is assumed to be uniform across the map. However, in our data of cloud cover, the assumption may be violated and so join-count statistics can produce misleading conclusion. Kabos and Csillag (2002) developed a new method for handling first-order heterogeneity on a regular lattice, but this type of analysis requires the information on the probability distribution of the category of interest at each location, which we do not have.

The spherical model is one of the most frequently used models in geostatistics. It represents transition features that have a common extent and that appear as patches. The exponential model represents a transition process in which the structure has random extents, i.e. it can quantify randomness in space. Cloud cover falls into this category, which explains why the exponential model behaves best among the three mixed models we have assessed for predicting the probability of cloud cover. The Gaussian model is rarely found to be useful to model environmental phenomena, so statistical models that use the Gaussian model as spatial covariance structure could not give a good fit.

Even though we have Landsat images and MODIS images with different spatial resolution, we cannot assess the effect of spatial resolution on the derivation since we have to integrate the pixels in the Landsat images to the spatial resolution of 240 m to reduce the total number of pixels in the image, or we could not run the models in SAS due to the limited memory of our PC computers.

Although the strategy of sampling MODIS images is random, we had to discard the totally cloudy images since there is no way to do spatial registering in totally cloudy scenes, which may cause the underestimate of the derived probability of cloud cover.

From Fig. 10, we can see that cloud shadow was not vertically under cloudy areas if the sun was not overhead. However, we could not correct these offsets due to solar angle since we do not know the height of cloud. This would cause overestimate or underestimate of solar radiation at a particular location. It needs further research to determine the exact shadow locations.

8. Conclusion

Because join-count statistics show that there exists significant positive spatial autocorrelation in the empirical data of cloud cover, we used one generalized linear model (GLM) and three generalized linear mixed models (GLMMs) that incorporated three different spatial structure models—spherical, exponential and Gaussian models to account for the spatial autocorrelation implicit in cloud distribution. It turned out that the GLMM with the exponential spatial structure model and the GLM were the best fit models based on multiple statistics including AIC, scaled deviance and extra-dispersion scale. The most significant variable in the models was always the difference between the elevation and the lifting condensation level based on the statistical outputs from the models. The coefficient estimates from the best models were similar when calibrated by two different types of remote sensing images, which justifies our using the MODIS images instead of the Landsat images to calibrate the GL(M)Ms in order to predict the hourly probabilities of cloud cover later when we calculated monthly average cloud cover. Our cloud models have the potential to lead to the more precise simulation of the daily variation in cloud cover, net solar radiation on the earth's surface and the ecological models driven by solar radiation anywhere anytime possible due to the high temporal resolution, moderate spatial resolution and the free distribution of the MODIS images. However, we could access only the agreement between the modeled results and the observations at a certain time or at a certain location due to the limitation of cloud information available. We found our models could usually predict the probability of cloud cover every 100 m in elevation at a certain time with an IoA of 0.560–0.919 and at a certain location over a day with an IoA of 0.940–0.994, indicating a medium to good model simulation at that particular time or location. Most of time in this study, Model II, with an exponential spatial structure, produced better estimates of cloud cover than Model 0 without spatial structures. However, when elevation is an extremely important factor contributing to cloud formation, the spatial autocorrelation in the cloud data results more from response to the independent variable "elevation" rather than to the interaction among sites. So Model 0 – the traditional regression method – is more appropriate to predict the cloud cover at the high elevations. It is important to consider spatial autocorrelation implicit in the dataset when we use regression techniques with remotely sensed images, but we need to keep in mind that there are two main causes of spatial structure, one is reaction to the independent variables with implicit spatial structure, the other is interaction within the dependent variable. When reaction is the dominant reason, the traditional regression models are appropriate. Otherwise, we need to apply mixed model that account for spatial autocorrelation. Thus the generalized linear mixed model does not necessarily give a better estimate compared to the traditional generalized linear model that does not incorporate spatial structure into considerations.

We believe that we have created a valuable tool that could have extensive use in the tropics. In areas that do not have large elevation ranges and orographic lifting is not the main driver for cloud formation, elevation may not be a suitable independent variable used in the models. In that case we would need to find new independent variables according to the specific environmental conditions at certain locations. We are also considering the addition of frontal system and other secondary drivers for cloud formation in the model. In subsequent papers, we plan to apply it to modeling the water budget of the

LEF and extend it to the entire San Juan (capital of Puerto Rico, north-western of the LEF) region.

Acknowledgements

This work was made possible thanks to the research grants BSR-8811902, DEB 9411973, DEB 0080538, and DEB 0218039 from NSF to the Institute for Tropical Ecosystem Studies, University of Puerto Rico, and to the International Institute of Tropical Forestry USDA Forest Service, as part of the Long-Term Ecological Research Program in the Luquillo Experimental Forest. The International Institute of Tropical Forestry USDA Forest Service, the University of Puerto Rico, and National Aeronautics and Space Administration (NASA) through Dr. Ramakrishna Nemani gave additional support.

The authors want to thank Dr. Ariel Lugo and Dr. Fred Scatena for providing the meteorological data and critical comments, Dr. Stephen Stehman for giving critical comments, Eda Melendez-Colom for the meteorological data, Dr. Eileen Helmer and University of Maryland's Global Land Cover Facility for providing the Landsat-7 ETM+ images used in this paper, Phil Gibbs for helping with model evaluation, and Pauline Stephen for the help of MODIS image processing.

References

Anselin, L., Griffith, D.A., 1988. Do spatial effects really matter in regression analysis? Papers Regional Sci. Assoc. 65, 11–34.

Agresti, A., 1990. Categorical Data Analysis. John Wiley and Sons.

Briscoe, C.B., 1966. Weather in the Luquillo Mountains of Puerto Rico. USDA Forest Service Research Paper ITF-3. Institute of Tropical Forestry. Rio Piedras Puerto Rico.

Brown, S., Lugo, A.E., Silander, S., Liegel, L., 1983. Research history and opportunities in the Luquillo Experimental Forest. US Forest Service. General Technical Report SO-44. New Orleans, Louisiana.

Bruijnzeel, L.A., 2000. Hydrology of tropical montane cloud forests: a reassessment. In: Gladwell, J.S. (Ed.), Proceedings of the Second International Colloquium on Hydrology in the Humid Tropics, CATHALAC. Panama City, Panama.

Bruijnzeel, L.A., Hamilton, L.S., 2000. Decision Time for Cloud Forest, vol. 13 of IHP Humid Tropics Programme Series. UNESCO-IHP, Paris, France.

Bruijnzeel, L.A., Procter, J., 1995. Hydrology and biochemistry of tropical montane cloud forests: what do we know? In: Hamilton, L.S., Juvik, J.O., Scatena, F.N. (Eds.), Tropical Montane Cloud Forests, No. 110 in Ecological Studies. Springer-Verlag, New York, pp. 38–78 (Chapter 3).

Bussieres, N., Goita, K., 1997. Evaluation of strategies to deal with cloudy situation in satellite evapotranspiration algorithm. In: Proceedings of the Third International Workshop NHRI Symposium, vol. 17, 16–18 October 1996, NASA, Goodard Space Flight Center, Greenbelt, Maryland, USA, pp. 33–43.

Chen, H., Varshney, P.K., Arora, M.K., 2003. Performance of mutual information similarity measure for registration of multitemporal remote sensing images. IEEE Trans. Geosci. Remote Sens. 41 (11), 2445–2454.

Ewel, J.J., Whitmore, J.L. 1973. The ecological life zones of Puerto Rico and the U.S. Virgin Islands. Research Paper ITF-18. USDA Forest Service, International Institute of Tropical Forestry, Rio Piedras, Puerto Rico.

Fedorov, D., Fonseca, L.M.G., Kenney, C., Manjunath, B.S., 2002. System for automatic registration of remote sensing images. In: IEEE International Geoscience and Remote Sensing Symposium-IGARSS02, 24–28 June 2002, Toronto, Canada.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281, 237–240.

Foody, G.M., 2003. Geographical weighting as a further refinement to regression modeling: an example focused on the NDVI–rainfall relationship. Remote Sens. Environ. 88, 283–293.

Graham, E.A., Mulkey, S.S., Kitajima, K., Phillips, N.G., Wright, S.S., 2003. Cloud cover limits net CO_2 uptake and growth of a rainforest tree during tropical rainy seasons. PNAS 100 (2), 572–576.

Griffith, D.A., Layne, L.J., 1999. A Casebook for Spatial Statistical Data Analysis—A Compilation of Analysis Different Thematic Datasets. Oxford University Press, New York, p. 506.

Janssen, P.H.M., Heuberger, P.S.C., 1995. Calibration of process-oriented models. Ecol. Model. 83, 55–66.

Kabos, S., Csillag, F., 2002. The analysis of spatial association on a regular lattice by join-count statistics without the assumption of first-order homogeneity. Comput. Geosci. 28, 901–910.

Kite, G.W., Droogers, P., 2000. Comparing evapotranspiration estimates from satellites, hydrological models and field data. J. Hydrol. 229, 3–18.

Lawton, R.O., Nair, U.S., Pielke, R.A., Welch, R.M., 2001. Climatic impact of tropical deforestation on nearby montane cloud forests. Science 294 (5542), 584–587.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., 1996. SAS Systems for Mixed Models. SAS Institute Inc, p. 309, 426.

Logar, A.M., Lloyd, D.E., Corwin, E.M., Penaloza, M.L., Feind, R.E., Berendes, T.A., Kuo, K., Welch, R.M., 1998. The ASTER polar cloud mask. IEEE Trans. Geosci. Remote Sens. 36 (4), 1302–1312.

Mauser, M., Schädlich, S., 1998. Modelling the spatial distribution of evapotranspiration on different scales using remote sensing data. J. Hydrol. 212–213, 250–267.

Melillo, J.M., McGuire, A.D., Kicklighter, D.W., Moore, B., Vörösmarly, C.J., Schless, A.J., 1993. Global climate change and terrestrial net primary production. Nature 363, 234–240.

Nair, U.S., Lawton, R.O., Welch, R.M., Pielke, R.A., 2003. Impact of land use on Costa Rican tropical montane cloud forests: sensitivity of cumulus cloud field characteristics to lowland deforestation. J. Geophys. Res. D: Atmos. 108 (7), ACL 4-1–ACL 4-13.

Odum, H.T., Drewry, G., Kline, K.R., 1970. Climate at El Verde 1963-1966. In: Odum, H.T., Pidgeon, R.F. (Eds.), A Tropical Rain Forest. U.S. Atomic Energy Communications, Washington, D.C, pp. B-347–B-418.

Overmars, K.P., De Koning, G.H.J., Veldkamp, A., 2003. Spatial autocorrelation in multi-scale land use models. Ecol. Model. 164 (2-3), 257–270.

Pinker, R.T., Zhang, B., Dutton, E.G., 2005. Do satellites detect trends in surface solar radiation? Science 308 (5723), 850–853.

Pounds, J.A., Fogden, M.P.A., Campbell, J.H., 1999. Biological response to climate change on a tropical mountain. Nature 398, 611–615.

Sawada, M., 1999. ROOKCASE: An Excel 97/2000 Visual Basic (VB) Add-in for exploring global and local spatial autocorrelation. Bull. Ecol. Soc. Am. 80 (4), 231–234.

Scatena, F.N.A.E., Lugo, 1995. Geomorphology, disturbance, and the soil and vegetation of two subtropical wet steepland watersheds of Puerto Rico. Geomorphology 13, 199–213.

Schabenberger, O., Pierce, F.J., 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press LLC, Boca Raton, FL.

Schellekens, J., Bruijnzeel, L.A., Scatena, F.N., Bink, N.J., Holwerda, F., 2000. Evaporation from a tropical rain forest, Luquillo Experimental Forest, Eastern Puerto Rico. Water Resour. Res. 36 (8), 2183–2196.

Silver, W.L., Lugo, A.E., Keller, M., 1999. Soil oxygen availability and biogeochemistry along rainfall and topographic gradients in upland wet tropical forest soils. Biochemistry 44, 301–328.

Still, C.J., Foster, P.N., Schneider, S.H., 1999. Simulating the effect of climate change on tropical cloud forests. Nature 398, 608–610.

Van der Molen, M.K., 2002. Meteorological impacts of land use change in the maritime tropics. Ph. D. dissertation. Vrije University, Amsterdam, Netherlands.

Waide, R.B., Zimmerman, J.K., Scatena, F.N., 1998. Controls of primary productivity: lessons from the Luquillo Mountains in Puerto Rico. Ecology 79, 31–37.

Wang, H., 2001. Dynamic modeling of the spatial and temporal variations of forest carbon and nitrogen inventories, including their responses to hurricane disturbances, in the Luquillo Mountains, Puerto Rico. Ph.D. thesis, State University of New York College of Environmental Science and Forestry, Syracuse, New York.

Weaver, P.L.P., Murphy, 1990. Forest structure and productivity in Puerto Rico's Luquillo Mountains. Biotropica 22 (1), 69–82.

Welch, R.M., Berendes, D., Berendes, T.A., Kuo, K., Logar, A.M., 1999. The ASTER Polar Cloud Mask Algorithm Theoretical Basis Document. http://eospso.gsfc.nasa.gov/eos_homepage/for_scientists/atbd/docs/ASTER/atbd-ast-13.pdf.

Wen, G., Cahalan, R.F., Tsay, S., Oreopoulos, L., 2001. Impact of cumulus cloud spacing on Landsat atmospheric correction and aerosol retrieval. J. Geophys. Res. 106 (D11), 12, 129–12, 138.

Wild, M., Gilgen, H., Roesch, A., Ohmura, A., Long, C.N., Dutton, E.G., Forgan, B., Kallis, A., Russak, V., Tsvetkov, A., 2005. From dimming to brightening: decadal changes in solar radiation at earth' surface. Science 308 (5723), 847–850.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. Period. 2, 184–194.

Wolfinger, R.D., O'Connell, M., 1993. Generalized linear mixed models: a pseudo-likelihood approach. J. Statistical Comput. Simul. 48, 233–243.

Wu, W., Hall, C.A.S., Scatena, F., Quackenbush, L., 2005. Spatial modelling of evapotranspiration in the Luquillo Experimental Forest of Puerto Rico using remotely-sensed data, J. Hydrol., submitted for publication.