Environmental Modelling & Software 26 (2011) 201–209

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Improving daily rainfall estimation from NDVI using a wavelet transform

Roberto Quiroz [a,*], Christian Yarlequé [a], Adolfo Posadas [a], Víctor Mares [a], Walter W. Immerzeel [b]

[a] International Potato Centre, Lima, Peru
[b] FutureWater, Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 18 May 2010
Received in revised form
7 July 2010
Accepted 24 July 2010
Available online 25 August 2010

Keywords:
Rainfall
NDVI
Transforms
Wavelets
Fourier
Reconstruction

ABSTRACT

Quantifying rainfall at spatial and temporal scales in regions where meteorological stations are scarce is important for agriculture, natural resource management and land-atmosphere interactions science. We describe a new approach to reconstruct daily rainfall from rain gauge data and the normalized difference vegetation index (NDVI) based on the fact that both signals are periodic and proportional. The procedure combines the Fourier Transform (FT) and the Wavelet Transform (WT). FT was used to estimate the lag time between rainfall and the vegetation response. Subsequently, third level decompositions of both signals with WT were used for the reconstruction process, determined by the entropy difference between levels and R^2. The low-frequency NDVI data signal, to which the high frequency signal (noise) extracted from the rainfall data was added, was the base for the reconstruction. The reconstructed and the measured rainfall showed similar entropy levels and better determination coefficients (>0.81) than the estimates with conventional statistical relations reported in the literature where this level of precision is only found for comparisons at the seasonal levels. Cross-validation resulted in ≤10% entropy differences, compared to more than 45% obtained for the standard method when the NDVI was used to estimate the rainfall in the same pixel where the weather station was located. This methodology based on high resolution NDVI fields and data from a limited number of meteorological stations improves spatial reconstruction of rainfall.

1. Introduction

Rainfall is a major driving force in the water cycle and the most important factor in promoting vegetation growth in rain-fed agriculture and natural grasslands and forests of the world. Accurate rainfall data with sufficient spatial resolution are of key importance in assessing basin scale hydrology but in many developing countries, adequate gauged data is seldom available.

Remote sensing can provide spatial precipitation patterns. Ground radar systems can also provide spatial precipitation information but validation of its data products is a major challenge for general hydrologic applications (Krajewski and Smith, 2002). Ground radar systems also have a limited range and are generally aimed at monitoring extreme events over limited time spans, making their use less suitable for long term assessments. Satellite remote sensing is a better source of spatial precipitation data, which are generally readily available over longer periods and large areas. Many different algorithms and types of sensors aboard a variety of satellites exist. Adler et al. (2001) provide an extensive overview and inter-compare 25 satellite based products to four model based, and to two climatological products. As many of these products have either a poor spatial resolution (~100 km) or a poor temporal resolution (~1 month), there is a need to develop a robust downscaling methodology for precipitation.

Many studies have used the intuitive correlation between rainfall and plant biomass, particularly in arid and semi-arid environments, to fill in this rainfall data gap (see Richard and Poccard, 1998; Kawabata et al., 2001; Lotsch et al., 2003; Nicholson and Farrar, 1994; Farrar et al., 1994; Nicholson et al., 1990; Eklundh, 1998; Martiny et al., 2006; Chamaille-Jammes et al., 2006; Dinku et al., 2008). However, the vegetation response to precipitation is highly variable in space, mainly due to soil and other influencing factors. There is also a delayed response in time, termed lag time (Farrar et al., 1994), which is defined as the time required for a volume of water equal to the annual mean of exchangeable soil moisture to be depleted by the combined processes of runoff and evapotranspiration. This lag time varies for different agro-ecologies; in semi-arid regions it is usually of the order of 2–3 months (Nicholson and Lare, 1990). This process is also described by Entekhabi et al. (1996) and Brunsell and Young (2008),

* Corresponding author. Tel.: +511 3175312; fax: +511 3175329.
E-mail addresses: r.quiroz@cgiar.org (R. Quiroz), c.yarleque@cgiar.org (C. Yarlequé), a.posadas@cgiar.org (A. Posadas), v.mares@cgiar.org (V. Mares), w.immerzeel@futurewater.nl (W.W. Immerzeel).

1364-8152/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.07.006

who stated that the surface affects the role of soil moisture by acting as a low-pass filter to the high frequency rainfall signal into a near immediate moisture effect and then slowly into a vegetation effect.

Most studies have sought statistical relationships between rainfall and NDVI but they seldom go beyond simple correlation analysis (Brunsell and Young, 2008). A linear relationship between rainfall and NDVI has been reported for areas with precipitation ranging from 200 to 1200 mm per year (Nicholson et al., 1996). In other areas the upper limit is attained at lower annual precipitations (Martiny et al., 2006). Above the upper threshold, the index "saturates" and NDVI increases only very slowly with increasing rainfall or becomes constant. Statistical relationships to estimate rainfall seem to be meaningless when applied to dekadal or daily data, especially when autocorrelations are removed (Eklundh, 1998). In spite of the lack of accuracy of the statistical relationship between these two signals, this is the standard method of using satellite information to estimate rainfall in arid and semi-arid regions. Thus, current procedures for estimating rainfall from NDVI are of limited use in applications in modeling agricultural production and land-atmosphere interactions studies, where dekadal or daily rainfall is required.

The present study aimed at 1) developing a procedure for generating daily precipitation data, using reconstructions that combine wavelet-filtered signals containing the low-frequency lag-corrected vegetation greening signal, extracted from NDVI time series, and the high frequency portion of the daily rainfall data; and, 2) test whether daily rainfall events can be approximated for neighboring areas where only NDVI data is available.

2. Materials and methods

2.1. Study area

The Altiplano is a high Andean plateau centered around Lake Titicaca in the Peruvian-Bolivian border. The plateau rises from the lake level at 3800 meters (m) to over 4500 m altitude. The rainfall varies from less than 400–600 mm yr^{-1}; average minimum temperature drops to −10 °C, droughts can last up to 150 d yr^{-1}, while frost-free days are around 150. The dominant vegetation is natural grasslands with cultivated areas mainly near the lake (Quiroz et al., 2003).

Convective activity and precipitation in the Altiplano occur almost exclusively during the austral summer and are associated with the seasonal expansion of the upper-air easterlies and related near-surface moisture influx from continental lowlands to the east. Precipitation from the west is rare because Pacific moisture is trapped vertically by large-scale subsidence and a stable low-level inversion at 900 hPa, and laterally by the coastal escarpment (Vuille and Keimig, 2004).

2.2. Climate data

Rain-gauge daily data from 10 weather stations (Fig. 1) were obtained from the Peruvian national meteorology and hydrology service (SENAMHI). Data for the period from January 1st 1999 through December 31st 2002 was included in the analysis. The raw data (Fig. 2) was checked for inconsistency and outliers. The analysis was conducted for the ten sites where the weather stations were located.

2.3. NDVI data

A dataset containing 180 10-day (dekad) composite NDVI images derived from the SPOT-4 and SPOT-5 VEGETATION instruments was used, spanning the period January 1999–December 2003. The VGT1 sensor aboard the SPOT-4 satellite provided the data for the January 1999–January 2003 period whereas the remaining period was covered with data from the VGT2 sensor aboard the SPOT-5 satellite. Both sensors have the same spectral and spatial resolution. The red spectral band (0.61–0.68 mm) and the near-infrared (NIR) spectral band (0.78–0.89 mm) were used to calculate the NDVI (NIR − RED/NIR + RED) and the imagery had a spatial resolution of 1 km. The synthesized preprocessed S10 NDVI product, which is a geometrically and radiometrically corrected 10-day composite image (Immerzeel et al., 2005), was used. The periods were defined according to the civil calendar: from the 1st day to the 10th; from the 11th to the 20th; and from the 21st to the end of each month.

The GPS coordinates of the weather stations were co-registered with the NDVI imagery data corresponding to each site. Therefore, for each location a vector containing 180 NDVI values, one for each civil calendar dekad, was extracted and used in the analysis. The dekadal NDVI value was repeated for each day within the respective dekad to match the daily observations in the rainfall data, generating 1826 NDVI daily values for the entire period. Given the difference in magnitude of the two signals, and for visual comparison purposes, the NDVI values were multiplied by the ratio of the mean value of both signals to generate magnitudes comparable to those registered for rainfall (Yarlequé, 2009).

2.4. Data pre-processing

2.4.1. Fourier analysis

Fourier or harmonic analysis is a mathematical technique used to decompose a complex static signal into a series of individual cosine waves, each characterized by a specific amplitude and phase angle. Several authors have successfully applied Fourier analysis in analyzing time series of NDVI imagery (e.g. Azzali and Menenti, 2000; Roerink and Menenti, 2000; Jakubauskas et al., 2001, 2002; Moody and Johnson, 2001; Immerzeel et al., 2005).

A stationary process can be represented by a series of harmonic functions, whose frequencies are multiples of a base frequency. This series of harmonic functions is called a Fourier series. Assuming that the process can be described by a function S, the usual form of the Fourier series is (Pipes and Harvill, 1971):

$$S(t) = \frac{A_0}{2} + \sum_{n=1}^{n=\infty} A_n \cos(n\omega t) + \sum_{n=1}^{n=\infty} B_n \sin(n\omega t) \quad (1)$$

Where t = time and A and B are the Fourier coefficients.

The constant term in Eq. (1) is always equal to the mean value of the S(t), (the mean NDVI value in a series of satellite imagery) and $\omega = 2\pi f_0$, where f_0 is the base frequency. Eq. (1) can be written in different forms, following basic mathematical laws (Pipes and Harvill, 1971). In this research it was decided to transfer Eq. (1) to a form that only contains cosine terms, which facilitates interpretation. Eq. (1) can also be written as

$$S(t) = \frac{A_0}{2} + \sum_{n=1}^{n=\infty} C_n \cos(n\omega t - \theta_n) \quad (2)$$

Eq. (2) now has a convenient form with only cosine terms. The signal is decomposed in a series of cosine terms, each with its own amplitude (C_n) and phase angle (θ_n), and a constant term ($A_0/2$). When a signal is described using Fourier analysis the values for the coefficients C_n need to be found. An algorithm to recover those coefficients from a discrete signal is the Fast Fourier Transform (FFT). In this case we analyzed a signal comprised of 1826 discrete NDVI daily values to estimate the Fourier coefficients C_n. The result of the FFT is a complex vector, with a real part containing the A coefficients and an imaginary part containing the B coefficients of Eq. (1). The coefficients C of Eq. (2) can be derived from A and B by calculating the length of the vector. There are a few limitations to the FFT related to the underlying mathematics. Firstly, to correctly recover a signal from the Fourier transform of its samples, the signal must be sampled with a frequency of at least twice its bandwidth (Nyquist frequency). Secondly, the signal needs to be sufficiently static to permit the analysis under the assumption that the wave is static (intrinsic assumption of the FFT) which means that both amplitude and phase of the individual terms should not vary significantly over time.

2.4.2. Determination of the lag time

The lag time between the onset of the rainy season and the greening of the vegetation was assessed with the Fourier analysis. Both rainfall (S_{Rain}) and NDVI (S_{NDVI}) signals were reconstructed with the six first harmonic components ($n = 1$ to 6 in Eq. (2)) of the Fourier series, with sizes N and M, respectively. By including six harmonics in the simulation of rainfall and NDVI signals, most of the variance in the original data is explained (Immerzeel et al., 2005). These smoothed Fourier transform (SFT) signals were used to estimate the lag time between the two primary signals. A new independent variable was generated through the simulation of the SFT for different periods T (where $T \varepsilon Z^+$). Out of all possible periods, T = 15, 30, 91, 121, 182, and 365 d were used for the analysis. Partitions $P_T = \{0, T, 2T, \ldots, kT\}$, $k \varepsilon Z^+$, with respect to T and $kT < N,M$, were defined. Each partition divided both signals (S_{Rain} and S_{NDVI}) into several sub-intervals. These intervals were used to search for the lags. Fig. 3 shows this concept with a sample of two sub-intervals; i and $i + 1$. For each sub-interval the time difference between the peaks of the signals S_{Rain} and S_{NDVI} were registered as the lag time for that sub-interval (lag_i and lag_{i+1}). Then the average lag time over the k-sub-intervals was obtained as:

$$Lag(T) = \langle lag_k \rangle = \langle \Delta t_k \rangle \quad (3)$$

where the <> symbolizes average over k. Thus, we are estimating the lag time as a new function Lag(T) (Eq. (3)), of the period T. The best coefficient of determination was used for estimating the lag time for each meteorological station. Once the lag time was considered, only four complete raining seasons spanning five years were suitable for the analysis, comprising a data set of 1421 daily data pairs (NDVI, rainfall)