Embedding Spatial Variability in Rainfall Field Reconstruction

Luis A. Duffaut Espinosa[a], Francisco Rosales[b], and Adolfo Posadas[c]

[a]Department of Electrical and Biomedical Engineering, University of Vermont, Burlington, Vermont 05041, USA.; [b]Academic Department of Finance, Universidad del Pacífico, Av. Salaverry 2020, Lima 11, Perú; [c]Agrosight, Allia Future Business Centre, Kings Hedges Road, Cambridge CB4 9HY, UK.

ARTICLE HISTORY
Compiled December 20, 2017

ABSTRACT

This manuscript provides a methodology for the reconstruction of a rainfall field when there are scarce rain-gauge stations available. This situation typically arises when measurements are taken from meteorological stations across time, and the information for the complete field is required as an input for larger scale models. The proposed method is based on a wavelet reconstruction technique that requires no distributional assumptions, but relies on the relation between rainfall and NDVI (Normalized Difference Vegetation Index) to account for the unobserved spatial variability of the field. The methodology is applied over a region of the southern Peruvian Andes where data gathered from meteorological stations provide enough statistical significance. A comparison with respect to an alternative source of spatial variability and common practices is provided.

KEYWORDS
Rainfall reconstruction, NDVI, Wavelets.

1. Introduction

Rainfall data collected from sparsely distributed meteorological stations is a common scenario faced by climate scientists. Whether it is to be used as an input in climate/weather models or to characterize its behavior in its original scale, a statistical reconstruction method of the signal is of great importance. Hydrological and general circulation models (GCM) are typical examples of this situation, where daily rainfall data is required as input to run simulations under specific scenarios. The results are then used to predict drought periods or to understand changes in climate systems (Lloyd-Hugues & Saunders 2002). In fact, there is a pressing need amongst policy makers (e.g. UN, USAID and FAO) for the reconstruction and correction of atmospheric datasets. Currently, these datasets are produced via global and regional models for rainfall outputs as well as ground/space-based rainfall observations such as those provided by NASA's earth-observing satellite missions. However, even though these measurements produce information having operational resolutions of 1km and finer, their accuracy is at best 10km due to the physical limitations such as cloud microphysics and terrain heterogeneity. These problems motivate the usage of correction

CONTACT Luis A. Duffaut Espinosa. Email: lduffaut@uvm.edu.

methods (Hwang & Graham 2013), and alternative reconstruction techniques (Duffaut Espinosa *et al.* 2017; Hartkamp *et al.* 1999; Heidinger *et al.* 2012; Hijmans *et al.* 2005; Hutchinson 1995; Lovejoy & Schertzer 2013; Price *et al.* 2000; Posadas *et al.* 2015; Quiroz *et al.* 2011).

The rainfall reconstruction technique proposed in this manuscript can be seen as an statistical method that exploits the well-known linear correlation between rainfall and the so-called Normalized Difference Vegetation Index (NDVI) in regions where the annual rainfall goes from 200 to 1200 mm Nicholson *et al.* (1996); Quiroz *et al.* (2011). The linearity limit corresponds to regions with low annual precipitations (Martiny *et al.* 2006). In (Heidinger *et al.* 2012; Quiroz *et al.* 2011), a point-wise version of the technique was used on low resolution data obtained from the Tropical Rainfall Measuring Mission (TRMM) (Simpson *et al.* 1996). However, it was found that when the reconstruction was applied on a point (spatial) sufficiently far from a meteorological station, the reconstruction showed certain bias towards the station and failed to embed the spatial heterogeneity of the surrounding area. The main contribution of the current method is that it incorporates the spatial heterogeneity of NDVI via: 1) a temporal wavelet reconstruction (Mallat 1998), and 2) a spatial prediction (Cressie N. 1991).

The method is applied to rainfall datasets for the Southern Peruvian Andes with data gathered from meteorological stations sparsely distributed in the region shown in Figure 1, and it is tested with relevant competing methods such as ANUSPLIN interpolation and wavelet-based closest station correction (Quiroz *et al.* 2011). The results show that the proposed method provides an improvement over common practices in that it gives better statistical metrics and goodness of fit of exceedance curves. It also remove fictitious boundaries introduced by Thiessen polygons when the rainfall influence at a spatial point is assumed to correspond to the closest meteorological station.

The structure of the manuscript is as follows. Section 2 describes the region of study and the datasets. Section 3 presents some mathematical tools, and the reconstruction method is discussed in terms of two algorithms. In Section 4 the method is applied to rainfall data collected from the southern Peruvian Andes, its performance is compared to other methods available in the literature, and the results are discussed. Section 5 provides conclusions and points of interest for future research.

2. Region of Study and Data

2.1. *Study area*

The studied region is defined by a grid with 225×225 cells of approximately 1km^2 each from the Southern Peruvian Andes (Figure 1). The majority of these cells are placed in the Peruvian Andean High-plateau or Altiplano, and a few cells, over the East side of the Andes. Geographical coordinates of the study area are between latitudes 14° to 16° South and longitudes 69° to 71° West constituting an area of approximately $225 \times 225\text{km}^2$. The altitudes range between 800 to 6500 m.a.s.l, approximately. The annual rainfall varies, on average, from $\sim$ 250mm in the arid southwest to $\sim$ 5000mm in the Amazon basin at the northeast corner of the study site (Garreaud *et al.* 2003). In this area of study, there are 19 meteorological stations from which observed rainfall measurements are obtained during a period of 8 years (from 1999 to 2006). The meteorological station data was obtained from the official Peruvian meteorological agency (SENAMHI 2017), which reports daily aggregations from tipping bucket rain gauges

at 0.1mm resolution. The measurements provided by SENAMHI amount to time series having 1 day temporal resolution.

2.2. *Data*

The main source of information for the methodology developed in section 3 is NDVI data. The NDVI dataset consists of 288 (dekad) composite images (225×225 pixels) with an approximate resolution of 1km corresponding to the area shown in Figure 1. NDVI data here correspond to the product VGT-S10 NDVI, which is derived from the vegetation instruments SPOT-4 and SPOT-5 over the time period starting in January 1999 and ending in December 2006. The period from January 2007 to December 2007 is also considered in this work for correction purposes (Quiroz *et al.* 2011). The spectral and spatial resolution of the vegetation instruments is the same. The spectral band 0.61-0.68 mm corresponding to red ($\mathrm{R_{Red}}$), and the band 0.78-0.89 mm corresponding to near-infrared ($\mathrm{R_{Infra}}$) were used to compute the NDVI index by employing the standard formula

$$\mathrm{NDVI} = \frac{\mathrm{R_{Infra}} - \mathrm{R_{Red}}}{\mathrm{R_{Infra}} + \mathrm{R_{Red}}}. \tag{1}$$

The final product has a spatial resolution of $\sim$ 1 km. The above formula for the NDVI index restricts the values to be in the interval $[-1, 1]$. In addition, the NDVI index is geometrically and radiometrically corrected producing the $S10$ NDVI product (Immerzeel *et al.* 2005). The dates assigned to the 288 dekadal samples were defined according to the civil calendar. In particular, every month was divided into 3 pieces: from the 1st day to the 10th; from the 11th to the 20th; and from the 21st to the end of each month. Each month therefore produces 3 NDVI data points per month for which the maximum value of NDVI in each piece of the month is assigned to the 10th, 20th and last day of the month. Since the NDVI product used in this paper is not evenly sampled, NDVI data has been re-sampled to an uniformly sampled 8 day resolution prior to its lag correction as described in (Duffaut Espinosa *et al.* 2017) given the inherent smoothness of the data. The 8 day resolution was chosen since it is convenient for the wavelet technique used in Section 3. In addition, NDVI contains a temporal lag with respect to the precipitation at a particular instant of time. The lag in the NDVI signal is the latency time that takes between rainfall reaching the ground and the time changes in the biomass index are registered in the red and infrared frequencies (Tucker 1979). The developments in (Immerzeel *et al.* 2005; Yarlequé *et al.* 2016) and (Duffaut Espinosa *et al.* 2017, Section 4) were followed in order to correct for this NDVI lag response, and therefore the NDVI dataset used in this paper accounts already for lag correction and re-sampling to a uniform 8 day resolution. To synchronize the time resolutions of the NDVI data and the ground measurements, the daily ground measurements were added every 8 days and the accumulated value is assigned at the end of each 8 day period.

A second source of information is obtained from the interpolation of the 19 meteorological stations data using the thin-plate smoothing spline algorithm implemented in the ANUSPLIN 4.36 package (Hutchinson 2006; The ANUSPLIN package 2007). This interpolation considers the latitude, longitude, and elevation of the area in addition to measured rainfall at each station (Hutchinson 1995). The method was chosen due to its higher accuracy compared to other methods in areas similar to the Andes high plateau, see Hijmans *et al.* (2005); Hartkamp *et al.* (1999); Jarvis & Stuart

(2001); Price *et al.* (2000). We also justify this selection since several climate products such as WorldClim ((Hijmans *et al.* 2005), `http://www.worldclim.org`) and IWMI Climate Atlas/CRU gridded data ((New *et al.* 2002), `http://www.iwmi.org`, `http://www.cru.uea.ac.uk`) have successfully applied the ANUSPLIN methodology.

3. Methods

Let $Y(s,t)$ denote a rainfall field, where $s \in \mathrm{S}$ with S denoting the set of all spatial coordinates of Y, $t \in \mathrm{T}$ with T denoting the set of all times related to Y at which the measurements are taken. Consider that Y is only known for $s \in \{s_1, s_2, \dots\} \subset \mathrm{S}$ and $t \in \{t_1, t_2, \dots\} \subset \mathrm{T}$, and that one is interested in estimating Y for all other $s \in \mathrm{S}$ and $t \in \mathrm{T}$. This situation typically arises in climate science when data is measured only at certain locations where there are meteorological stations and during certain period of time. However, climate modelers require information for the complete sets S and T, for instance, as input for models such as general circulation models (GCM) or hydrology transport models that require either better spatial and/or time resolutions. This is achieved by introducing auxiliary information in order to complete the unknown values of Y.

In what follows the field $Y(s,t)$ is handled fixing one of its two components. When this is the case, a bar will be placed above the component that is meant to be fixed. For example, $\boldsymbol{Y}(\bar{s},t)$ denotes a time series vector, while $\mathbf{Y}(s,\bar{t})$ denotes a day sample matrix of the field at time $\bar{t}$ (spatial point process).

3.1. *Temporal wavelet reconstruction*

3.1.1. Wavelet transforms

The wavelet transform of a function $f(t)$ with finite energy is defined as the integral transform with a family of functions $\psi_{\lambda,t}(u) := \frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-t}{\lambda}\right)$, and it is given by

$$\begin{aligned} \langle f, \psi_{\lambda,t}\rangle &= \int_{-\infty}^{\infty} f(u)\psi_{\lambda,t}(u)du, \quad \lambda > 0 && (2)\\ &= \int_{-\infty}^{\infty} f(u)\frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-t}{\lambda}\right), && (3) \end{aligned}$$

where λ is a scale parameter, t a location parameter and the functions $\psi_{\lambda,\tau}$ are called wavelets. The inverse wavelet transform is given by

$$\begin{aligned} f(t) &= \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{0}^{\infty} \lambda^{-2} W f(\lambda,u)\psi_{\lambda,u}(t)d\lambda du, && (4)\\ C_\psi &= 2\pi\int_0^{\infty}\frac{|\hat{\psi}(\omega)|^2}{\omega}d\omega < \infty, && (5) \end{aligned}$$

where $Wf(\lambda,u) := \langle f, \psi_{\lambda,u}\rangle$, $\hat{\psi}$ is the Fourier transform of ψ and ω is the frequency variable in the Fourier domain.

When the parameters λ and t in the wavelet transform $\langle f, \psi_{\lambda,t}\rangle$ lie in a continuum, the transform it is called continuous wavelet transform. To obtain a discrete wavelet

transform one can choose choose $\lambda = \lambda_0^m$, where m is an integer and $\lambda_0 > 1$ is a fixed dilation step. In particular, the following discretization is considered: $t = nt_0\lambda_0^m$, where $t_0 > 0$, and n is an integer. We can then re-parametrize the wavelet in terms of m and n as

$$\psi_{m,n}(t) = \frac{1}{\sqrt{\lambda_0^m}}\psi\left(\frac{t - nt_0\lambda_0^m}{\lambda_0^m}\right) \quad (6)$$

$$= \lambda_0^{-m/2}\psi\left(\lambda_0^{-m}t - nt_0\right), \quad (7)$$

and the discrete wavelet transform reads:

$$\langle f, \psi_{m,n}\rangle = \lambda_0^{-m/2}\int f(t)\psi\left(\lambda_0^{-m}t - nt_0\right)dt. \quad (8)$$

3.1.2. Wavelet approximation

Using wavelets that form a complete orthonormal basis (for instance, Haar wavelets), any finite energy function can be approximated up to arbitrary precision by a linear combination of basis functions $\psi_{m,n}(t)$. That is,

$$f(t) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} D_{m,n}\psi_{m,n}(t), \quad (9)$$

where $D_{m,n} := \langle f, \psi_{m,n}\rangle = \int_{-\infty}^{\infty} f(t)\psi_{m,n}(t)dt$. Moreover, by using an intermediate scale m_0, the last equation can be broken up into two sums

$$f(t) = \sum_{m=-\infty}^{m_0}\sum_{n=-\infty}^{\infty}\langle f, \psi_{m,n}\rangle\psi_{m,n}(t) + \sum_{m=m_0+1}^{\infty}\sum_{n=-\infty}^{\infty}\langle f, \psi_{m,n}\rangle\psi_{m,n}(t). \quad (10)$$

If in addition one considers the scaling functions $\phi_{m,n}(t)$ satisfying $\phi_{m,n}(t) = 2^{-m/2}\phi(2^{-m}t - n)$ and following the arguments in (Foufoula-Georgiou & Kumar 1994), then it can be shown that

$$\sum_{m=m_0+1}^{\infty}\sum_{n=-\infty}^{\infty}\langle f, \psi_{m,n}\rangle\psi_{m,n}(t) = \sum_{m=-\infty}^{\infty}\langle f, \phi_{m_0,n}\rangle\phi_{m_0,n}(t), \quad (11)$$

which, in fact, portrays the low frequency information (LFI) carried by the signal, whereas the first summation in the right hand side of (10) contains the high frequency information (HFI) of the signal. To make this separation clear, hereafter the following notation is used for the mentioned decomposition:

$$f(t) = L_f(t) + H_f(t) \quad (12)$$

$$H_f(t) = \sum_{m=-\infty}^{m_0}\sum_{n=-\infty}^{\infty}\langle f, \psi_{m,n}\rangle\psi_{m,n}(t) \quad (13)$$

$$L_f(t) = \sum_{m=-\infty}^{\infty}\langle f, \phi_{m_0,n}\rangle\phi_{m_0,n}(t). \quad (14)$$

An iterative decomposition can now be performed. Let $f^0 := f(= L_f + H_f)$ and make $f^1 := L_f$. Perform decomposition (12) on f^1 and make $f^2 := L_{f^1}$. The procedure can be continued an arbitrary number of times as long as the scale parameter allows it. The reverse process is known as *reconstruction* of the signal. That is, at decomposition level ℓ, the signal $\hat{f}^{\ell-1} := L_{f^\ell} + H_{f^\ell}$. The procedure is continued at level $\ell - 2$ by making $\hat{f}^{\ell-2} := \hat{f}^{\ell-1} + H_{f^{\ell-1}}$, and then repeated until reaching level 0. Note that signals f and $\hat{f} := \hat{f}^0$ are not necessarily the same, which makes decomposition and reconstruction noncomutative operations.

Recall that for a rainfall field $X(s,t)$ the variable s denotes the spatial coordinates of the field and t denotes the temporal variable of the field. Suppose there are two fields X and Z related to some climate variable (e.g., rainfall). The field X provides reliable information of the LFI of such climate variable whereas the field Z carries only information of the HFI of the same climate variable (e.g., X and Z are datasets obtained from two different satellite sensors). Therefore, decomposing $\boldsymbol{X}(\bar{s},t) = \boldsymbol{L_X}(\bar{s},t) + \boldsymbol{H_X}(\bar{s},t)$ and $\boldsymbol{Z}(\bar{s},t) = \boldsymbol{L_Z}(\bar{s},t) + \boldsymbol{H_Z}(\bar{s},t)$ at a fixed $\bar{s} \in \mathrm{S}$ allows one to reconstruct a signal $\boldsymbol{Y}$ at the same location by only using the good information of both sources. That is,

$$\boldsymbol{Y}(\bar{s},t) = \boldsymbol{L_X}(\bar{s},t) + \boldsymbol{H_Z}(\bar{s},t). \tag{15}$$

The reconstruction procedure can be performed after several levels of decomposition as described in the previous paragraph, which provides a proper embedding of the scaling properties of signals $\boldsymbol{X}$ (LFI) and $\boldsymbol{Z}$ (HFI). The level of decomposition is given by a number $\underline{\ell} \in \mathrm{N}$ indicating the level at which the LFI of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are closest to each other with respect to some statistical metric. Algorithm 1 summarizes the reconstruction procedure using two sources of information, and it is referred here simply as *temporal reconstruction* since it only conveys the time variable of the signals.

In (Heidinger *et al.* 2012), $\boldsymbol{X}(\bar{s},t)$ $\bar{s} \in \mathrm{S}$ was taken as the time series obtained from satellite information of The Tropical Rainfall Measuring Mission (TRMM), and $\boldsymbol{Z}(\bar{s},t)$ was provided from the closest meteorological station rainfall time series to the location $\bar{s}$. However, the spatial resolution of TRMM data is approximately 28 km, which, as reported in (Heidinger *et al.* 2012) introduces high spatial biases and does not take into account the spatial heterogeneity of the area.

In this manuscript, a natural candidate for the field X that provides a 1 km resolution is NDVI; while Z is obtained from two sources. The first is based on spline approximations of a finite number of local meteorological stations constructed using the topography map, longitude and latitude of the area. The main issue of using splines in this context is that splines are inherently smooth whereas rainfall fields exhibit spatial roughness (Kedem & Long 1987; Lin 1978). In the next section, it is argued that a linear predictor of the rainfall HFI based on a weighted sum of the surrounding meteorological stations provides a practical improvement over using splines or the closest meteorological station as the Z rainfall field. This amounts to a deterministic reconstruction model that preserve the temporal (8 days) and spatial (1km) resolutions of the data.

3.2. *Spatial reconstruction*

Two methodologies introducing spatial variability to the methodology in (Heidinger *et al.* 2012; Quiroz *et al.* 2011) are proposed. The first one (SR1) involves applying

Algorithm 1: Temporal Reconstruction

input : $\bar{s} \in \mathrm{S}$, $\boldsymbol{X}(\bar{s},t)$ and $\boldsymbol{Z}(\bar{s},t)$
output: $\boldsymbol{Y}(\bar{s},t)$
define : $\ell \leftarrow$ current decomposition level, $\bar{\ell} \leftarrow$ maximum allowed ℓ,
$\boldsymbol{X}^{\ell} \leftarrow \boldsymbol{X}(\bar{s},t)$ signal at level ℓ, $\boldsymbol{Z}^{\ell} \leftarrow \boldsymbol{Z}(\bar{s},t)$ signal at level ℓ,
dist $\leftarrow$ statistical metric between signals.

$\underline{\ell} \leftarrow 0$
for $\ell \leftarrow 0$ **to** $\bar{\ell}$ **do**
  Decompose $\boldsymbol{X}^{\ell} \rightarrow \boldsymbol{L_X}^{\ell} + \boldsymbol{H_X}^{\ell}$.
  Decompose $\boldsymbol{Z}^{\ell} \rightarrow \boldsymbol{L_Z}^{\ell} + \boldsymbol{H_Z}^{\ell}$.
  Assign $\boldsymbol{X}^{\ell+1} \leftarrow \boldsymbol{L_X}^{\ell}$
  Assign $\boldsymbol{Z}^{\ell+1} \leftarrow \boldsymbol{L_Z}^{\ell}$.
  if dist$(\boldsymbol{X}^{\ell+1}, \boldsymbol{Z}^{\ell+1}) <$ dist$(\boldsymbol{X}^{\ell}, \boldsymbol{Z}^{\ell})$ **then**
    $\underline{\ell} = \ell + 1$
  else
    continue
  end if
end for
for $i \leftarrow 0$ **to** $\underline{\ell}$ **do**
  Reconstruct $\boldsymbol{Y}^{i} \leftarrow \boldsymbol{L_X}^{i} + \boldsymbol{H_Z}^{i}$.
end for
Return $\boldsymbol{Y} \leftarrow \boldsymbol{Y}^{0}$

algorithm 1 to $X =$ NDVI and Z the rainfall field of spatially interpolated meteorological stations via splines (The ANUSPLIN package 2007; Hutchinson 1995, 2006). Although the rainfall fields obtained from the spline procedure are inherently *smooth* (spatially), these can still be used to extract the HFI component of the time series at each point in the area of study while NDVI, which is inherently smooth (temporally), provides the LFI part. A comparison of the spatial distribution between NDVI and ANUSPLIN fields is shown in Figure 7 for 28 January 2000 at the Andean region. The second methodology (SR2) considers HFI in time at every spatial point in the area of study as a weighted linear combination of all point-wise meteorological stations and the LFI is provided by $X =$ NDVI information. Specifically, at a location $\bar{s} \in \mathrm{S}$ and given that there is only knowledge of information at a finite number of places surrounding $\bar{s}$ (corresponding to meteorological stations locations), an estimation of rainfall is given by the linear predictor

$$\hat{\boldsymbol{X}}(\bar{s},t) = \sum_{i=1}^{K} \lambda_i \boldsymbol{X}_i(t), \tag{16}$$

where λ_i is a weight corresponding to the influence of the ith station around $\bar{s}$ and $\boldsymbol{X}_i(t)$ represents rainfall measured in time at the ith meteorological station. Naturally, these weights must add to 1, and they depend on their relative *distance* (Euclidean) of $\bar{s}$ to the meteorological stations. Thus, the weights $\{\lambda_i\}_{i=1}^{n}$ are calculated using a procedure borrowed from the *Kriging* interpolation procedure (Cressie N. 1991; Goovaerts 1997; Matheron 1965). In this manner, one can calculate $\hat{\boldsymbol{X}}(\bar{s},t)$ for every $\bar{s} \in \mathrm{S}$, and the rainfall field containing the high frequency in algorithm 1 is $Z = \hat{X}$.

3.2.1. Weight calculation

Calculating the weights in (16) requires spatial information of the area under study. A good source of spatial information is $X = \text{NDVI}$ due to its 1km spatial resolution. The first step is then to compute the semivariogram of $\mathbf{X}(s, \bar{t})$ for a fixed time $\bar{t} \in \text{T}$. This provides a notion of a zone of influence for the meteorological stations at each location in the area. Omitting time dependence, the experimental semivariogram is computed as

$$\gamma(h) = \frac{1}{2K(h)} \left(\sum_{\|s_j - s_i\|_2 = h} (\mathbf{X}(s_j, \bar{t}) - \mathbf{X}(s_i, \bar{t}))^2 \right), \tag{17}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, h is the separation distance between locations s_i and s_j, $K(h)$ is the number of location pairs separated by h, and $\mathbf{X}(s_i, \bar{t})$ is the value of the information (NDVI) at location $s_i \in \text{S}$ (Cressie N. 1991). For simplicity, only isotropic semivariograms are considered, however the information in heterogeneous terrain is very likely to be anisotropic. For illustration purposes, Figure 2 shows the averaged (in time) experimental semivariogram, which gives an idea of the average shape of a semivariogram obtained from an arbitrary NDVI day sample. The experimental semivariogram is then fitted to the function

$$\gamma(h) = \begin{cases} r + (c - r) \left\{ \frac{3}{2} \left(\frac{h}{a} \right) - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right\}, & h \geq a \\ c, & \text{otherwise} \end{cases} \tag{18}$$

where c is the *sill*, a is the range, and r is the nugget effect parameter (Cressie N. 1991). The results in Section 4 utilize only the semivariogram corresponding to the same NDVI day sample.

The task now is to find the explicit relationship between an arbitrary location $\bar{s} \in \text{S}$ and K fixed locations $\{\bar{s}_1, \ldots, \bar{s}_K\} \in \text{S}$. Assuming that NDVI is constituted as a random field that is weakly stationary whose mean μ and covariance function $C(i, j)$ with respect to the K fixed locations can be estimated from the data, then one can find the best linear relationship between NDVI at $\bar{s}$ and NDVI at the K fixed locations. That is, the linear predictor

$$X(\bar{s}, \bar{t}) = \sum_{i=1}^{K} \lambda_i X(\bar{s}_i, \bar{t}), \tag{19}$$

where $X(\bar{s}_i, \bar{t})$ is the value of the field at location $\bar{s}_i$ and time $\bar{t}$. The weights λ_i are

thus computed so that they minimize the mean squared error

$$
\begin{aligned}
e &= \mathbb{E}\left[\left(X(\bar{s},\bar{t}) - \sum_{i=1}^{K}\lambda_i X(\bar{s}_i,\bar{t})\right)^2\right] && (20)\\
&= Var\left(X(\bar{s},\bar{t}) - \sum_{i=1}^{K}\lambda_i X(\bar{s}_i,\bar{t})\right) && (21)\\
&= C(0,0) - 2\sum_{i=1}^{K}\lambda_i C(0,j) + \sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i\lambda_j C(i,j). && (22)
\end{aligned}
$$

Note that $C(i,j)$ is just the (i,j) element of the covariance matrix, and that the value $C(0,i)$ depends on the location $\bar{s}$. It is now only a matter of taking the derivative of e with respect to each λ_k. That is,

$$
\frac{\mathrm{d}\,\mathrm{e}}{\mathrm{d}\lambda_\mathrm{k}} = -2C(0,k) + 2\lambda_k C(k,k) + 2\sum_{i\neq k}^{K}\lambda_i C(i,k) = 0, \quad k = 1,2,\ldots,K, \tag{23}
$$

and since $C(i,j) = C(j,i)$ it follows that

$$
\underbrace{\begin{pmatrix} C(0,1)\\ \vdots \\ C(0,K)\end{pmatrix}}_{b} = \underbrace{\begin{pmatrix} C(1,1) & \cdots & C(1,K)\\ \vdots & \ddots & \vdots \\ C(N,1) & \cdots & C(N,K)\end{pmatrix}}_{C}\underbrace{\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_K\end{pmatrix}}_{\lambda}. \tag{24}
$$

Hence, the solution reads

$$
\lambda = C^{-1}b, \tag{25}
$$

which only depend on the covariance matrix values. The values of b and C can be directly obtained from the semivariogram computed out of the time averaged NDVI information since $C(h) = C(0) - \gamma(h)$, where $C(h) := C(i,j)$ such that $||s_i - s_j|| = h$ (Cressie N. 1991). Furthermore, one also require that $\sum_{i=1}^{n}\lambda_i = 1$ so that the system becomes

$$
\begin{pmatrix} C(0,1)\\ \vdots \\ C(0,K)\\ 1\end{pmatrix} = \begin{pmatrix} C(1,1) & \cdots & C(1,K) & 1\\ \vdots & \ddots & \vdots & \vdots\\ C(N,1) & \cdots & C(N,K) & 1\\ 1 & \cdots & 1 & 0\end{pmatrix}\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_K\\ \eta\end{pmatrix}, \tag{26}
$$

where η is a *Lagrange Multiplier*. From the procedure described above, one can easily find that the optimal Kriging variance is then

$$
\sigma^2 = C(0,0) - \sum_{i=1}^{K}C(0,i). \tag{27}
$$

Algorithm 2: Spatial Reconstruction

input : X and Z
output: Y

1: **for** $\bar{s} \in \mathrm{S}$ **do**
2: compute $\{\lambda_1(\bar{s}), \ldots, \lambda_K(\bar{s})\}$ using (25) for all $t \in \mathrm{T}$,
3: compute $\hat{\boldsymbol{X}}(\bar{s}, t)$ using (16),
4: apply Algorithm 1 to find $\boldsymbol{Y}(\bar{s}, t)$ with $\boldsymbol{X}(\bar{s}, t) = \mathrm{NDVI}(\bar{s}, t)$ and
5: $\boldsymbol{Z}(\bar{s}, t) = \hat{\boldsymbol{X}}(\bar{s}, t)$
6: **end for**
7: Return $Y(s, t)$ for all $s \in \mathrm{S}$ and $t \in \mathrm{T}$

Since matrix C can be singular, the pseudo inverse of C is used instead of the inverse. This fact, and numerical errors amounts for the possibility of having values of lambda slightly over 1, which at the same time causes other lambda values (corresponding to the same location) to be negligibly negative. Clearly, this is a consequence of the condition that forces the sum of weights to be equal to 1.

3.2.2. Reconstruction algorithm

The computation of weights presented in the previous section have to be performed for all times $t \in \mathrm{T}$. This provides a complete set of day samples containing the region of influence of each of the K locations where there is full rainfall information taken at meteorological stations. It is thus that one can compute the field $\hat{X}$ as a function of time and space via (16). Finally, Algorithm 1 is applied $X = \mathrm{NDVI}$ and $Z = \hat{X}$. The overall process is described in Algorithm 2.

4. Results and Validation

Methodologies SR1 and SR2 based on algorithms 1 and 2 and described in section 3 are now applied to the region of the Andes high plateau shown in Figure 1.

4.1. *Results*

A key calculation for algorithms 1 and 2 is computing the decomposition level $\underline{\ell}$. As described in Algorithm 1, $\underline{\ell}$ is computed by comparing the low frequency components of NDVI and either spline interpolation using the ANUSPLIN software (SR1) or $\hat{X}$ obtained using (16) (SR2). Then $\underline{\ell}$ is the level that provides a better goodness of fit statistic. That is, the statistical metric dist is chosen to be either 1-Nash-Sutcliffe Efficiency or 1-correlation coefficient. For the SR1 methodology, Figure 3 shows 6 levels of decomposed NDVI and ANUSPLIN time series for location (50,160) in the grid of 225×225 cells covering Figure 1 with location (0,0) being the upper left corner of the figure. Figure 5 shows 6 levels of decomposition for the SR2 methodology where NDVI and $\hat{X}$ obtained using the weighted linear predictor are compared. Tables 1 and 3 show the goodness of fit statistics for the time series at this point used in selecting $\underline{\ell}$ for SR1 and SR2, respectively. At location (200,20) in the grid, the decomposition levels are shown in Figures 4 and 5 as well as the corresponding goodness of fit statistics in

Tables 2 and 4 for the SR1 and SR2 methodologies, respectively. From these tables, it is clear that $\underline{\ell}$ is either 3 or 4. In general, it was observe that the level of decomposition for all grid points falls between levels 3 and 4, corresponding to a time dilation of 64 and 128 days, respectively. These time resolutions are within the range of NDVI lags reported for the studied region in (Duffaut Espinosa *et al.* 2017; Yarlequé *et al.* 2016). Therefore, a reasonable choice for $\underline{\ell}$ for all locations in the grid is assumed to be 3.

The SR1 methodology employs the 19 meteorological stations in the area of study as reference points to compute a $225 \times 225 \times 288$ datasets of rainfall values using the ANUSPLIN software. Then algorithm 2 is applied directly.

For the SR2 methodology, the same 19 meteorological stations are used for the calculation of the field $\hat{X}$ with weights computed using the spatial information provided by the NDVI dataset. Figure 8 shows the weights of 4 particular stations. Note that the spatial distribution of the weights provides a notion of the range of influence of meteorological stations. The results of the reconstructed rainfall are showed in Figure 9 for 4 randomly chosen days during the 8 year period of available NDVI information. The left column gives the results of using the SR1. The right column shows the result of the SR2 methodology. For comparison purposes, the middle column shows the case in which $\hat{X}$ at a specific location is given by the single closest meteorological station to the location of interest. Note that the area of influence in this case becomes the well-known Thiessen polygons (Voronoi 1908). It is observed in this case that there is a clear bias with respect to the station at the centre of each polygon, which produces an obvious fictitious rainfall boundary with the same shape of the Thiessen polygons.

4.2. *Validation*

Given that the reconstruction algorithms employs MRA on time series, a temporal validation on two grid points in the region is performed. For this purpose, two meteorological stations in the region of study were kept out of the process for generating ANUSPLIN and $\hat{X}$ fields. From this perspective, the validation is unbiased towards the algorithms 1 and 2. These stations are Pucara and Santa Rosa. Pucara is geographically located at longitude 70.37° W, 15.03° S and Santa Rosa is at 70.79° W, 14.62° S both have an altitude above 3900 m.a.s.l. Pucara and Santa Rosa stations were picked due to the fact that they are located on heterogeneous terrain and therefore they are representative of the spatial variability of the area.

Figure 10 shows the comparison of the data measured and data reconstructed at Pucara and Santa Rosa stations, and the corresponding statistics are shown in Table 5, where one notices the good agreement of the Hurst index, mean, maximum value, quantiles and variance. The discrepancies can be attributed to the stochastic nature of the time series as well as the difference in spatial scales. For example, the Hurst index being close to 0.5 indicates that there is not much correlation between the current and future observations, which is also indicated by the reconstructed time series. The exceedance probability curves were also computed and shown in Figure 11. This is validated by the goodness of fit statistics provided in Table 6 only for the weighted reconstruction procedure. The table shows the following indicators of the goodness of fit: MAE (mean average error) , RMSE (root mean squared error), CORR (correlation coefficient), PBIAS (Percent Bias), NSE (Nash-Sutcliffe Efficiency) and RSR (ratio of RMSE to the standard deviation of the observations). As a rule of thumb, the fitting or reconstruction can be considered satisfactory if the indicator NSE is greater than 0.50 and the indicator RSR is around 0.80 or below. This is indeed the case for both

stations; see Table 5. In addition, one can clearly see that the exceedance curves for the reconstructed NDVI data overlap better with the observed exceedance curves in comparison with the reconstructed curves obtained using spline interpolations. Discrepancies and errors are expected in this application. Some reasons for that can be attributed to the nature of NDVI in that it is a ratio of spectral measurements that relate to vegetation intake of water and nutrients. The latter are affected, for example, by the region's topography, which is not the scope of this manuscript. Also, NDVI does not provide information of the rainfall range, therefore the methodology (SR1) gets the range from the ANUSPLINE data whereas SR2 obtains the rainfall range from the linear combination of the point of interest surrounding stations. Potentially, the latter could be a source of error if the point of interest is uncorrelated to the surrounding stations, which imply that there were not enough stations around that point in addition to the fact that the stations' spatial dependence is time varying. This could explain the under estimation of rainfall for Pucara station shown in the top part of Figure 10 for the periods around 2000 and 2003. The only statistic in Table 5 that appears to be off compared to the others is $Q50$ (the 50th percentile). An explanation of this is that when transforming NDVI into a rainfall measure through the methods in this manuscript then the rainfall time series can sometimes get shifted up a tiny amount, and thus small values of rainfall measurements are introduced. This also observed in Figure 11, where the exceedance plots of reconstructed rainfall show discrepancies for very small values, but matches nicely for larger values of rainfall. Finally, the region located in the upper right corner of Figure 1 can not be accurately reconstructed using the methodology presented in this manuscript. The reason is that such region is located in the forest, which has annual rainfall exceeding the range under which NDVI correlates approximately linearly with rainfall (200mm to 1200mm).

5. Conclusion

In this manuscript a method for spatial reconstruction for rainfall was presented using ANUSPLIN and NDVI data as supporting/auxiliary corresponding to SR1 and SR2 methodologies. The procedure was validated spatially by testing the time series at two locations (corresponding to Pucara and Santa Rosa stations) that were not used in the reconstruction procedure but where there is on-site precipitation data available. One of the main drawbacks of using NDVI for spatial precipitation reconstruction is that it can only be applied over regions that correlate linearly with precipitation, i.e., where annual rainfall is between 200mm and 1200mm. The upper-right corner of the study area is above this threshold, and therefore the reconstruction is not reliable in that region. This constitutes one of the main drawbacks of using NDVI for spatial precipitation reconstruction. Naturally, the information carried by the auxiliary data is key for the method to work. Therefore, adding more sources of information such as cloud distributions will potentially improve the result. In principle this additional data can be embedded in similar manner as how the *co-Kriging* method handle several sources of data. This is a direction in which this method can be improved. The application also improves over what is available in the literature with respect to spatial reconstruction using Wavelets. For instance, the method was able to remove the fictitious boundaries resulting from using Thiessen polygons as shown in Figure 9.

As in any reconstruction procedure, the more data is available the better the reconstruction. Adding more meteorological station will improve the results in this manuscript. However, introducing more stations is not an easy task, e.g. it is expensive

to maintain meteorological stations, and the situation becomes even more troublesome when social aspects are added to the process of acquiring data in far reachable regions such those in the Andes.

Lastly, it is convenient to note that even though in this manuscript a few stations and satellite imagery were enough for a reliable reconstruction, the question of what is the minimum number of stations needed remains open, and left for future investigation.

References

Aceituno, P., & Montecinos, A. 1993. Circulation anomalies associated with dry and wet periods in the South American Altiplano. *Am. Meteor. Soc.*, 330–331.

Cressie N. 1991. *Statistics for Spatial Data.* Wiley-Interscience.

Duffaut Espinosa, L. A., Posadas, A., Carbajal, M., & Quiroz, R. 2017. Multifractal downscaling of rainfall using normalized difference vegetation index (NDVI) in the Andes plateau. *PLOSONE*, **12**(1), 1–25.

Foufoula-Georgiou, E., & Kumar, P. 1994. *Wavelets in Geophysics.* Editorial Academic Press Inc.

Garreaud, R., Vuille, M., & Clement, A. 2003. The climate of the Altiplano: Observed current conditions and mechanisms of past changes. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, **194**, 5–22.

Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation.* Oxford University Press.

Hartkamp, A.D., De Beurs, K., Stein, A., & White, J.W. 1999. *Interpolation techniques for climate variables.*

Heidinger, H., Yarlequé, C., Posadas, A., & Quiroz, R. 2012. TRMM rainfall correction over the Andean Plateau using wavelet multi-resolution analysis. *International Journal of Remote Sensing*, **33**(14), 4583–4602.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**(15), 1965–1978.

Hutchinson, M. F. 1995. Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographic Information Systems*, **9**, 305–403.

Hutchinson, M. F. 2006. *Centre for Resource and Environmental Studies.* Australian National University.

Hwang, S., & Graham, W. D. 2013. Development and comparative evaluation of a stochastic analog method to downscale daily GCM precipitation. *Hydrology and Earth Systems Science*, **17**, 4481–4502.

Immerzeel, W. W., Quiroz, R. A., & De Jong, S. M. 2005. Understanding precipitation patterns and land use interaction in Tibet using harmonic analysis of SPOT VGT-S10 NDVI time series. *International Journal of Remote Sensing*, **26**(11), 2281–2296.

Isaaks, E. H., & Srivastava, R. M. 1989. *An Introduction to Applied Geostatistics.* Oxford University Press.

Jarvis, C. H., & Stuart, N. 2001. A Comparison among Strategies for Interpolating Maximum and Minimum Daily Air Temperatures. Part II: The Interaction between Number of Guiding Variables and the Type of Interpolation Method. *Journal of Applied Meteorology*, **40**(6), 1075–1084.

Kedem, B., & Long, C. 1987. On the lognormality of rain rate. *Proc. Natl. Acad. Sci. USA*, **84**, 901–905.

Lin, S. H. 1978. More on rain rate distributions and extreme value statistics. *The Bell Systems Technical Journal*, **57**(5), 1545–1568.

Lloyd-Hugues, B., & Saunders, M.A. 2002. A drought climatology for Europe. *Journal of International Climatology*, **22**(13).

Lovejoy, S., & Schertzer, D. 2013. *The Weather and Climate: Emergent Laws and Multifractal Cascades*. Cambridge University Press.

Mallat, S. 1989. A theory of multiresolution signal decomposition: the wavelet representation. *EEE Transactions on Pattern Analysis and Machine Intelligence*, **1**(11), 674–693.

Mallat, S. 1998. *A wavelet Tour of Signal Processing*. Academic Press.

Martiny, N., Camberlin, P., Richard, Y., & Philippon, N. 2006. Compared regimes of NDVI and rainfall in semi-arid regions of Africa. *International Journal of Remote Sensing*, **27**(23), 5201–5223.

Matheron, G. 1965. *Les Variables Régionalis ées et leur Estimation*. Masson et Cie.

New, M., Lister, D., Hulme, M., & Makin, I. 2002. A high-resolution data set of surface climate over global land areas. *Climate research*, **21**(1).

Nicholson, S. E., R., Lare A., A., Marengo J., & P., Santos. 1996. A revised version of Lettau's evapoclimatonomy model. *Journal of Applied Meteorology*, **35**(4), 549–561.

Perica, S., & Foufoula-Georgiou, E. 1996. Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions. *J. Geophys. Res.*, **101**(D21), 26347–26361.

Posadas, A., Duffaut Espinosa, L. A., Yarlequé, C., Heidinger, H., Carvalho, L., Jones, C., & Quiroz, R. 2015. Spatial Random Downscaling of Rainfall Signals in Andean Heterogeneous Terrain. *Nonlinear Processes in Geophysics*, **22**(4), 383–402.

Price, D.T., McKenney, D.W., Nalder, I. A., Hutchinson, M.F., & Kesteven, J.L. 2000. A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agricultural and Forest Meteorology*, **101**(2–3), 81–94.

Quiroz, R., Yarlequé, C., Posadas, A., Mares, V., & Immerzeel, W. W. 2011. Improving daily rainfall estimation from NDVI using wavelet transform. *Environmental Modeling & Software*, **26**(2), 201–209.

SENAMHI. 2017. *Servicio nacional de meteorología e hidrología del Perú*.

Simpson, J., Kummerow, C., Tao, W. K., & Adler, R. F. 1996. On the tropical rainfall measuring mission (TRMM). *Meteorology and Atmospheric Physics*, **60**, 19–36.

The ANUSPLIN package. 2007. *Fenner School of Environment and Society, Australian National University*.

Tucker, C. J. 1979. Red and photographic infrared linear combination for monitoring vegetation. *Remote Sensing of Environment*, **8**, 127–150.

Vera, C., Baez, J., Douglas, M., Emmanuel, B. C., Marengo, J., Meitin, J., Nicolini, M., Nogues-Paegle, J., Penalba, O., Salio, P., Saulo, C., Silvia Dias, P., & Zipser, E. 2006. The south American low-level jet experiment. *Bull. Amer. Met. Soc.*, **87**(1), 63–77.

Voronoi, G. 1908. Recherches sur les paralléloèdres Primitives. *J. reine angew. Math.*, **134**, 1908.

Yarlequé, C., Vuille, M., Hardy, D. R., Posadas, A., & Quiroz, R. 2016. Multiscale assessment of spatial precipitation variability over complex mountain terrain using a high-resolution spatiotemporal wavelet reconstruction method. *Journal of Geophysical Research: Atmospheres*, **121**(20), 12,198–12,216.

Tables

Table 1. ANUSPLINE goodness of fit statistics at location $(50, 160)$.

Statistics	Lev. 1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	Lev. 6
NSE	0.15	0.36	**0.63**	**0.74**	0.47	0.016
Correlation Coefficient	0.52	0.64	**0.79**	**0.86**	0.68	0.44

Table 2. ANUSPLINE goodness of fit statistics location $(200, 20)$.

Statistics	Lev. 1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	Lev. 6
NSE	0.63	0.76	**0.83**	**0.83**	0.60	0.50
Correlation Coefficient	0.80	0.87	**0.92**	**0.92**	0.81	0.72

Table 3. Kriging goodness of fit statistics location $(50, 160)$.

Statistics	Lev. 1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	Lev. 6
NSE	0.28	0.42	**0.68**	**0.73**	0.58	0.15
Correlation Coefficient	0.62	0.70	**0.84**	**0.87**	0.76	0.52

Table 4. Kriging goodness of fit statistics location $(200, 20)$.

Statistics	Lev. 1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	Lev. 6
NSE	0.60	0.73	**0.81**	**0.82**	0.56	0.55
Correlation Coefficient	0.78	0.86	**0.91**	**0.91**	0.77	0.76

Table 5. Time series statistics.

Station	H	Mean*	Max*	Q50*	Q75*	Var**
Pucara (Observed)	0.50	16.82	120.60	6.70	23.80	546.20
Pucara (NDVI)	0.47	23.12	133.97	13.24	38.90	731.82
Santa Rosa (Observed)	0.56	17.18	99.1	7.7	25.32	474.52
Santa Rosa (NDVI)	0.53	19.01	97.37	11.57	32.61	421.90

*__Unit is mm.__

**__Unit is mm^2.__

Table 6. ECDF goodness of fit statistics.

Statistics	Pucara	Santa Rosa
MAE	0.082	0.051
RMSE	0.09	0.061
CORR	0.98	0.99
PBIAS	23.19	15.09
NSE	0.81	0.90
RSR	0.42	0.30

Figures

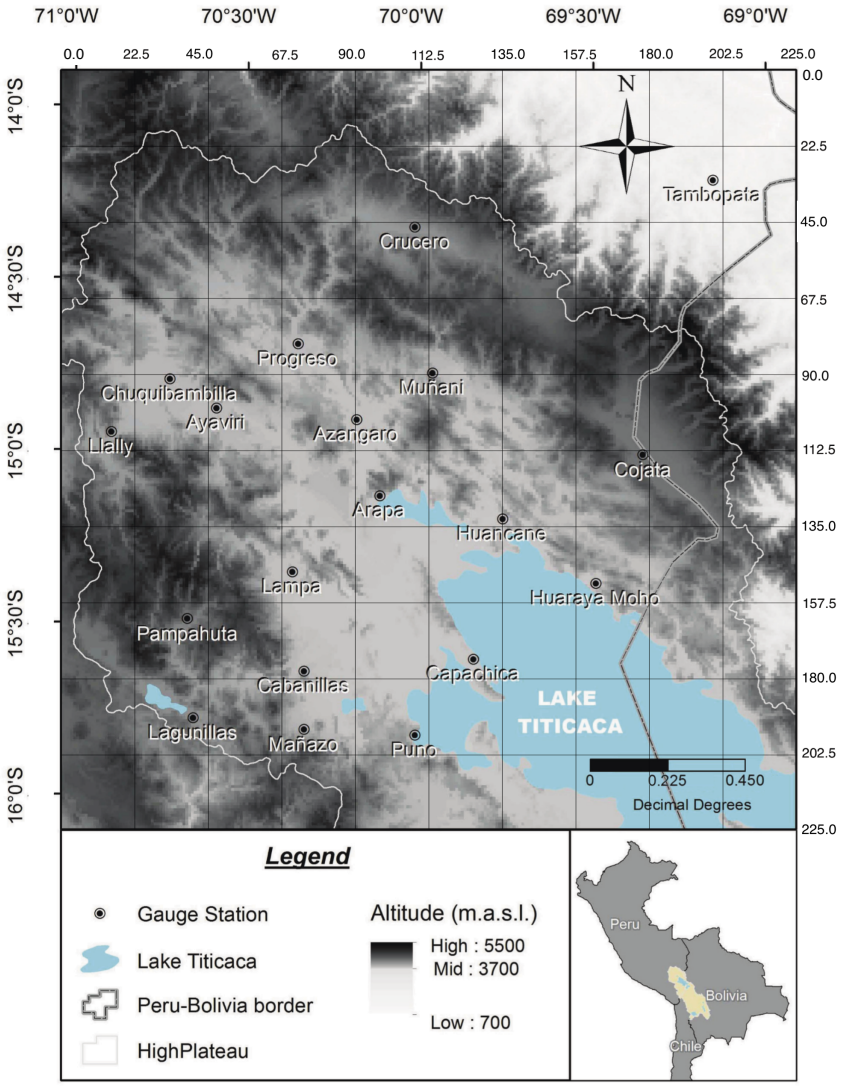

Figure 1.: Study region, where gray levels indicate altitude according to the legend, and the inner horizontal and vertical labels correspond to kilometer units encompassing a 225×225 km^2 area.

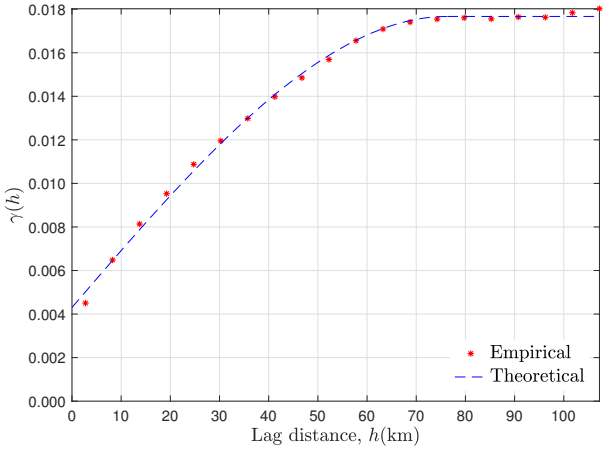

Figure 2.: Averaged in time NDVI semivariogam $\gamma(h)$.

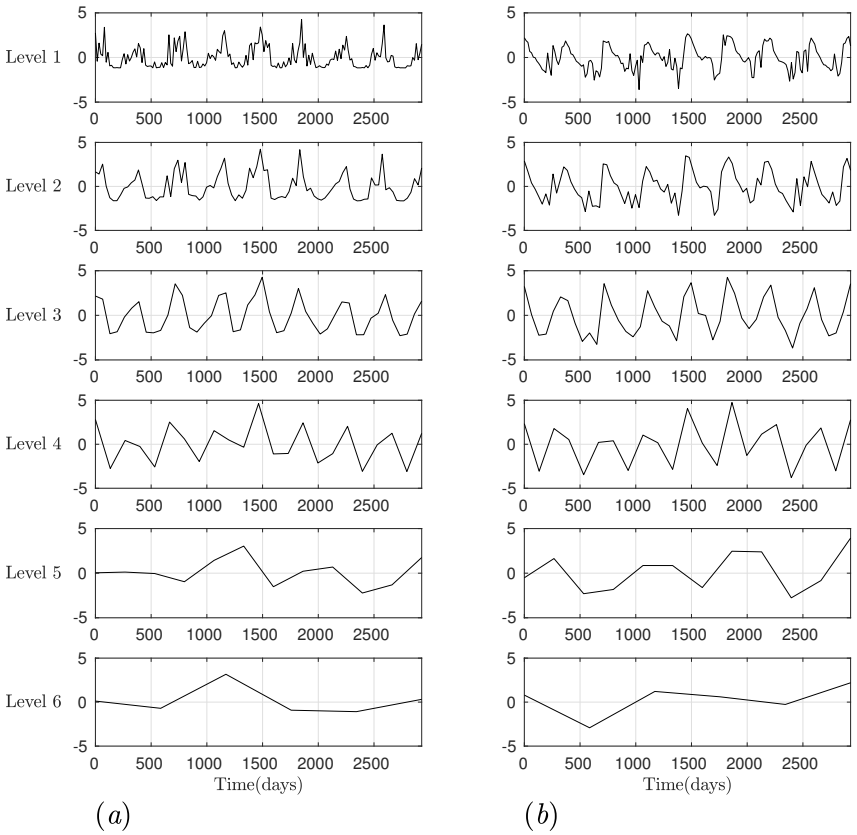

Figure 3.: SR1-Wavelet decomposition: grid location $(50, 160)$. Column (a) shows standardized ANUSPLINE and column (b) standardized NDVI.

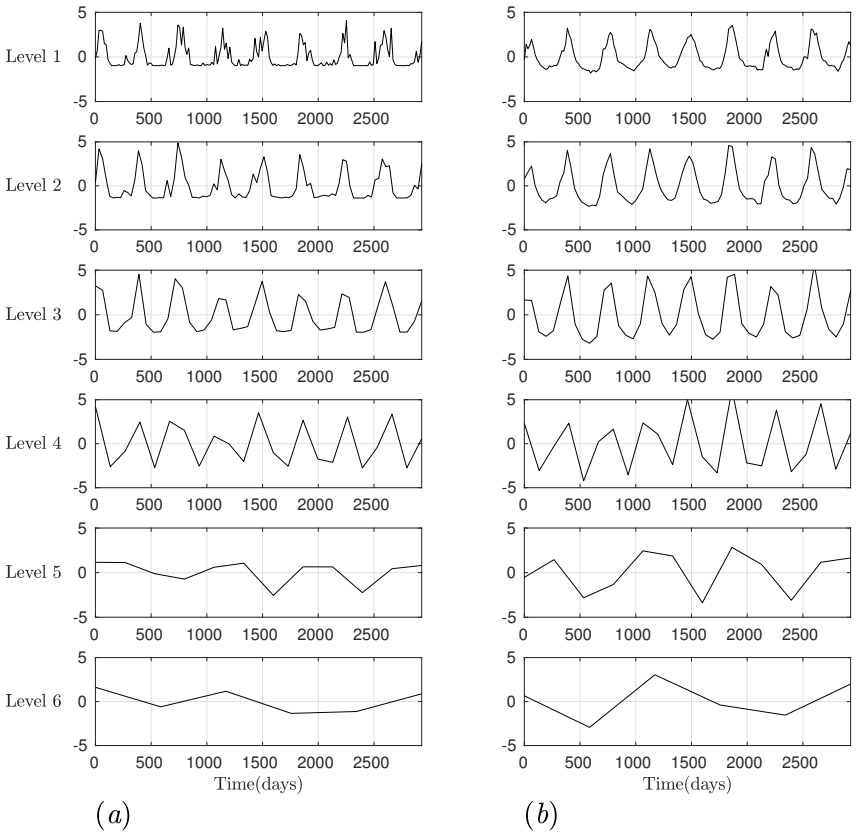

Figure 4.: SR1-Wavelet decomposition: grid location $(200, 20)$. Column (a) shows standardized ANUSPLINE and column (b) standardized NDVI.

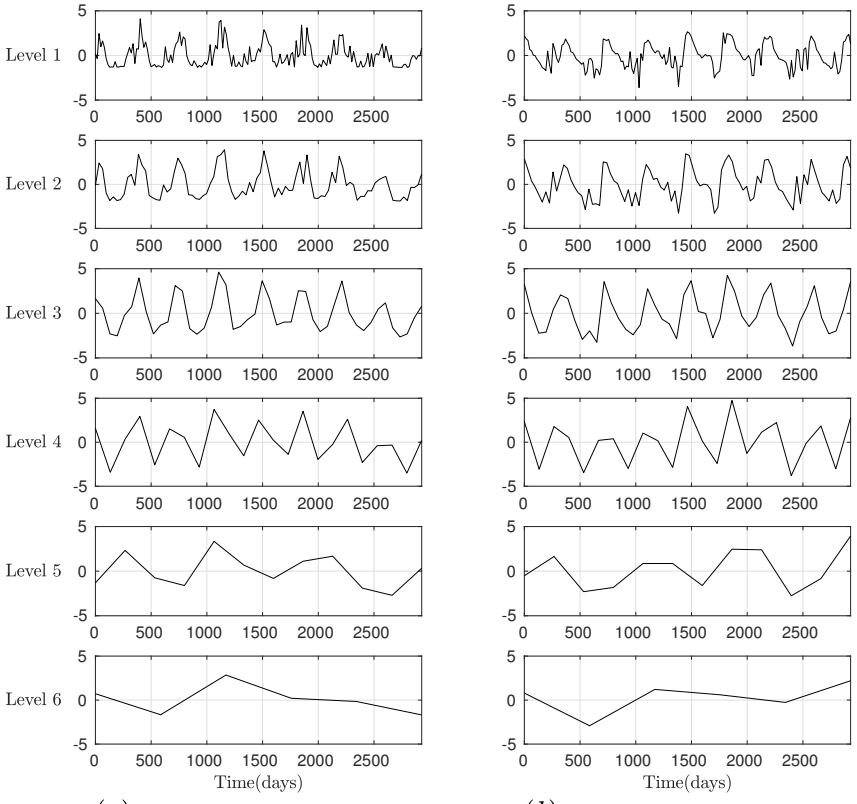

Figure 5.: SR2-Wavelet decomposition: grid location $(50, 160)$. Column (a) shows standardized $\hat{X}$ and column (b) standardized NDVI.

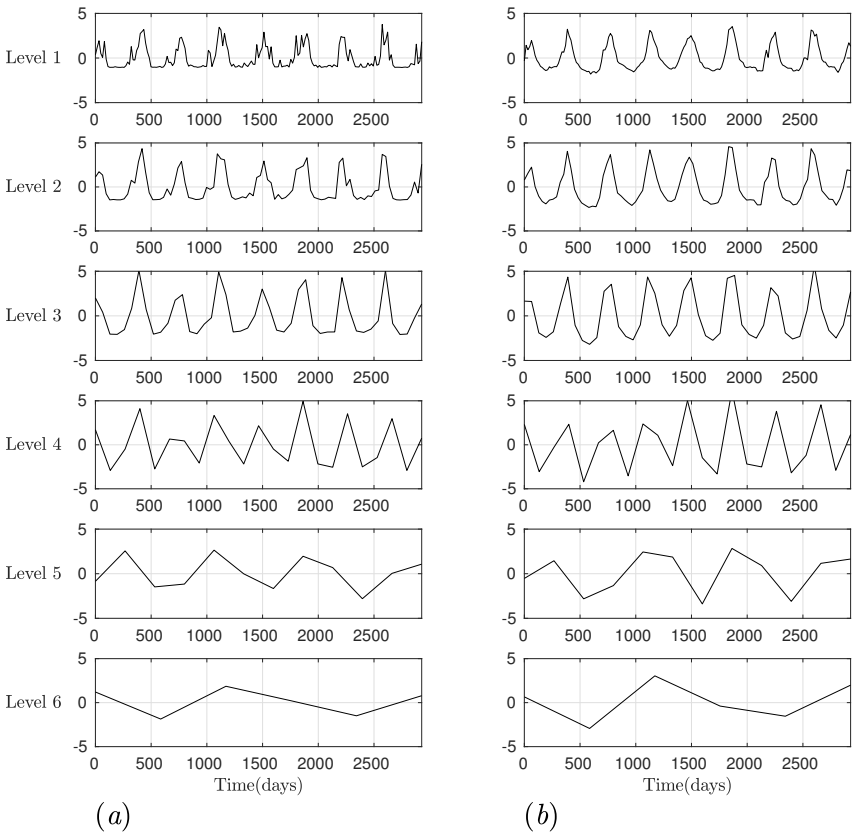

Figure 6.: SR2-Wavelet decomposition: grid location $(200, 20)$. Column (a) shows standardized $\hat{X}$ and column (b) standardized NDVI.

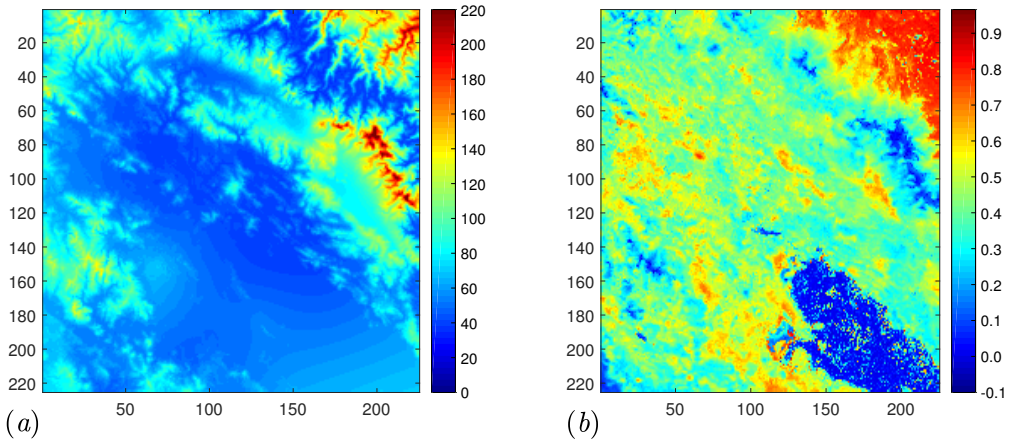

Figure 7.: (a) ANUSPLINE day sample with colorbar unit (mm). (b) NDVI day sample with adimensional colorbar units obtained from (1). Both figures correspond to 28 January 2000. The axes of (a) and (b) indicate adimensional grid locations.

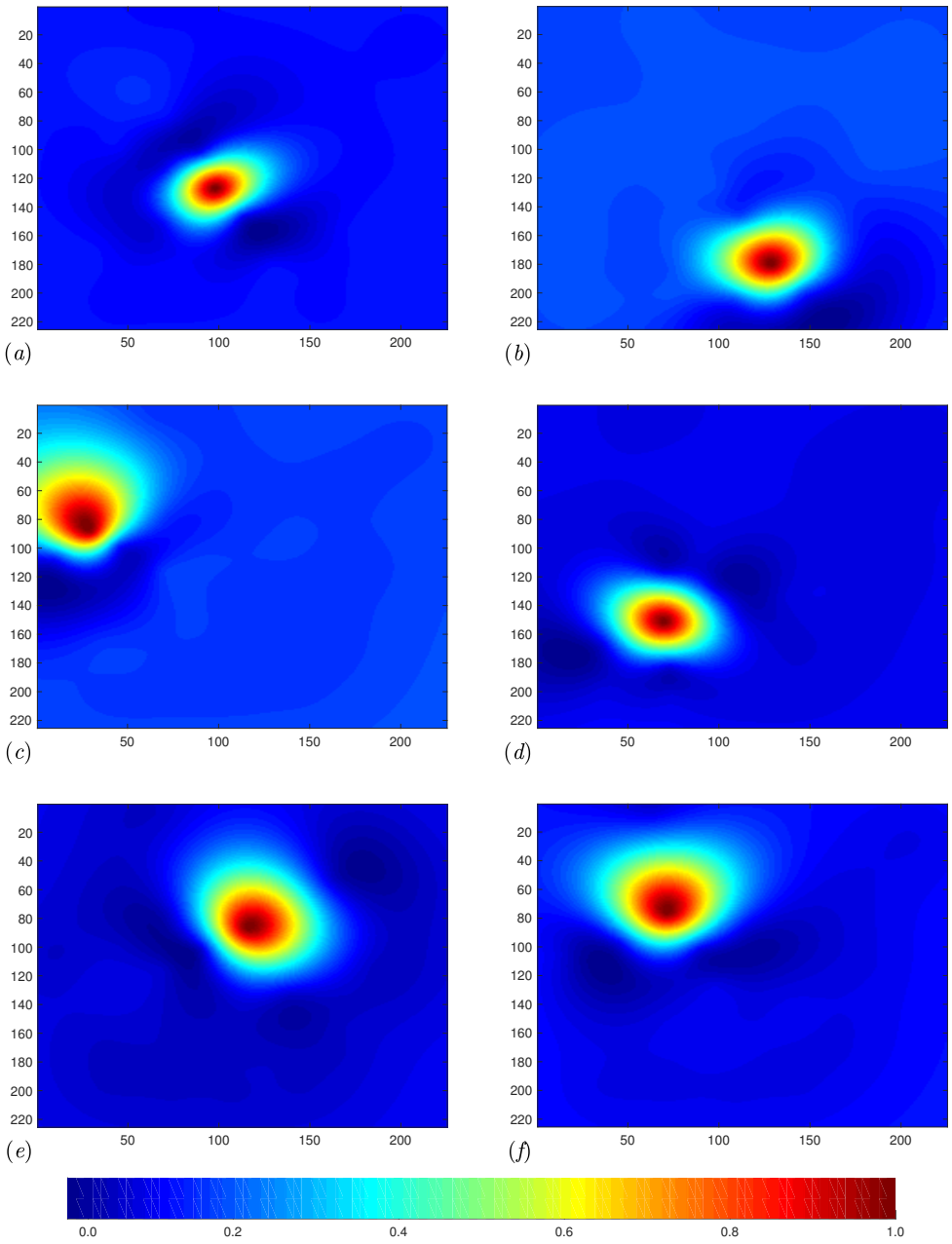

Figure 8.: Spatial lambda values. Each image plots the spatial distribution of λ_i corresponding to station i. First row shows influence regions for (a) Arapa and (b) Capachica, second row shows influence regions for (c) Chuquibambilla and (d) Lampa stations, and third row shows influence regions for (e) Munani and (f) Progreso stations. The axes of all figures indicate adimensional grid locations.

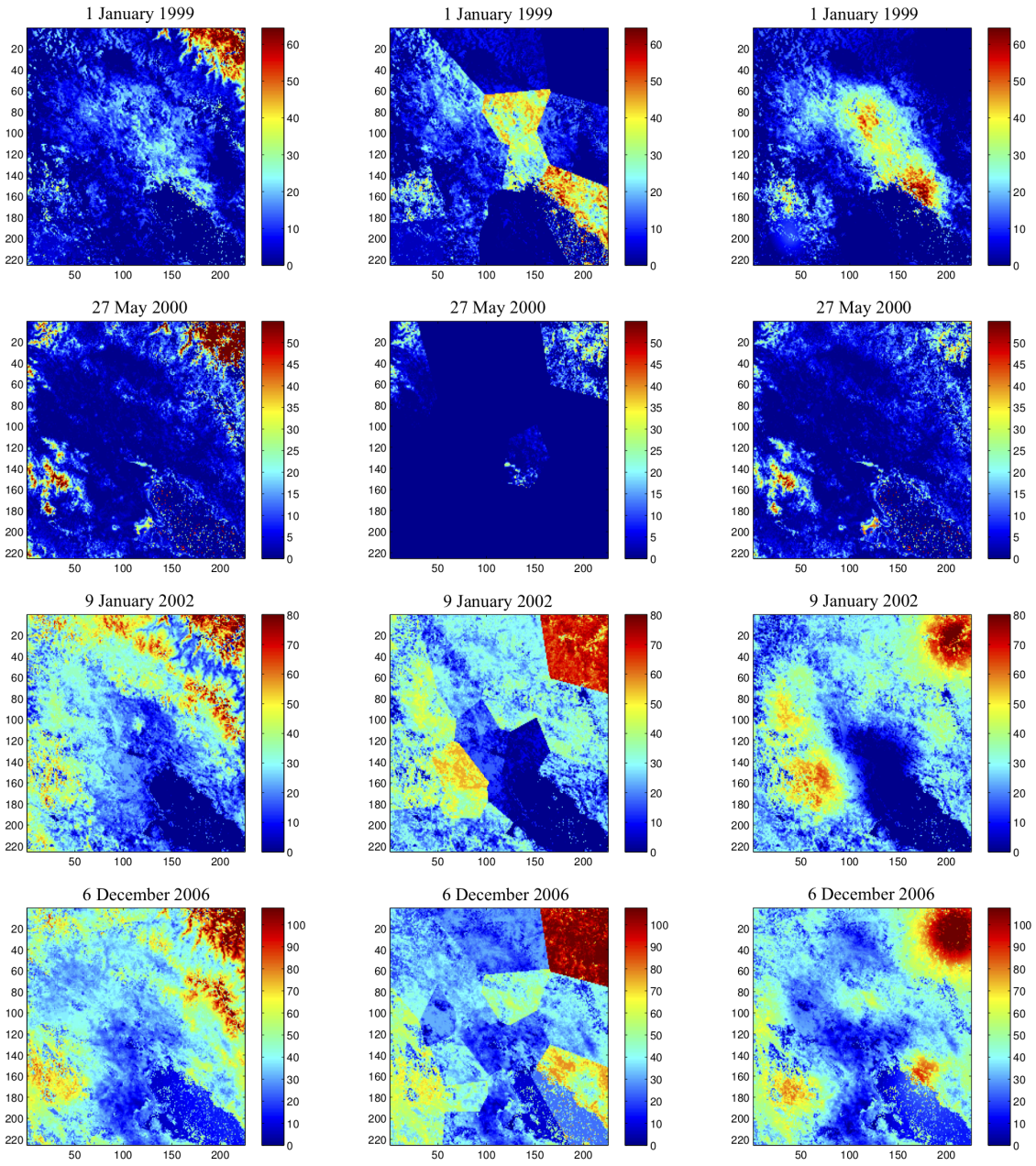

Figure 9.: Comparison of methodologies: Column (a) SR1, column (b) gives the result using Thiessen polygons spatial influence and column (c) SR2. The colorbars unit is mm and the axes of all figures indicate adimensional grid locations.

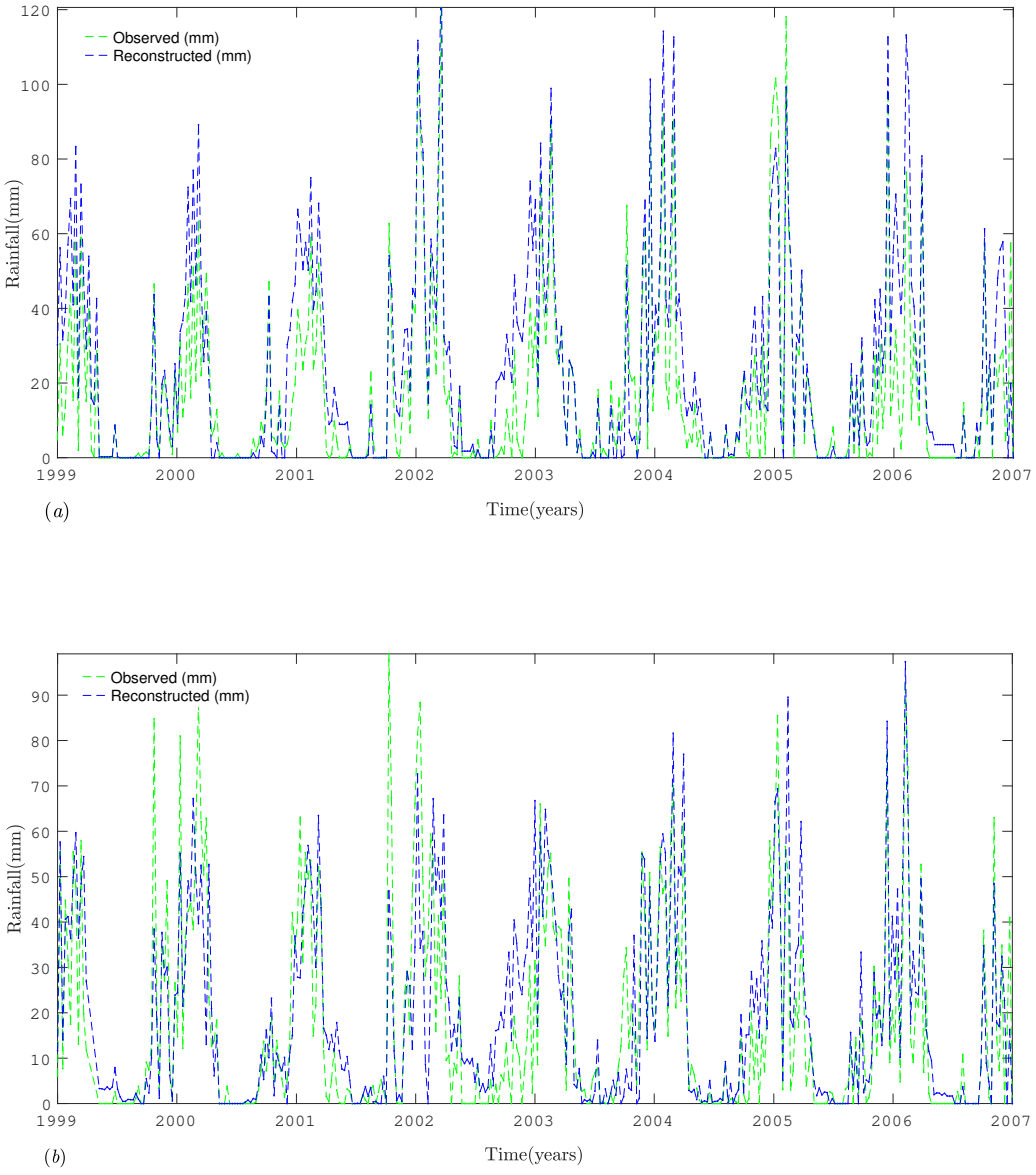

Figure 10.: Observed and reconstructed precipitation time series: Pucara station (a) and Santa Rosa station (b). Each rainfall data point is the result of an 8 day accumulation period as described in Section 2.2.

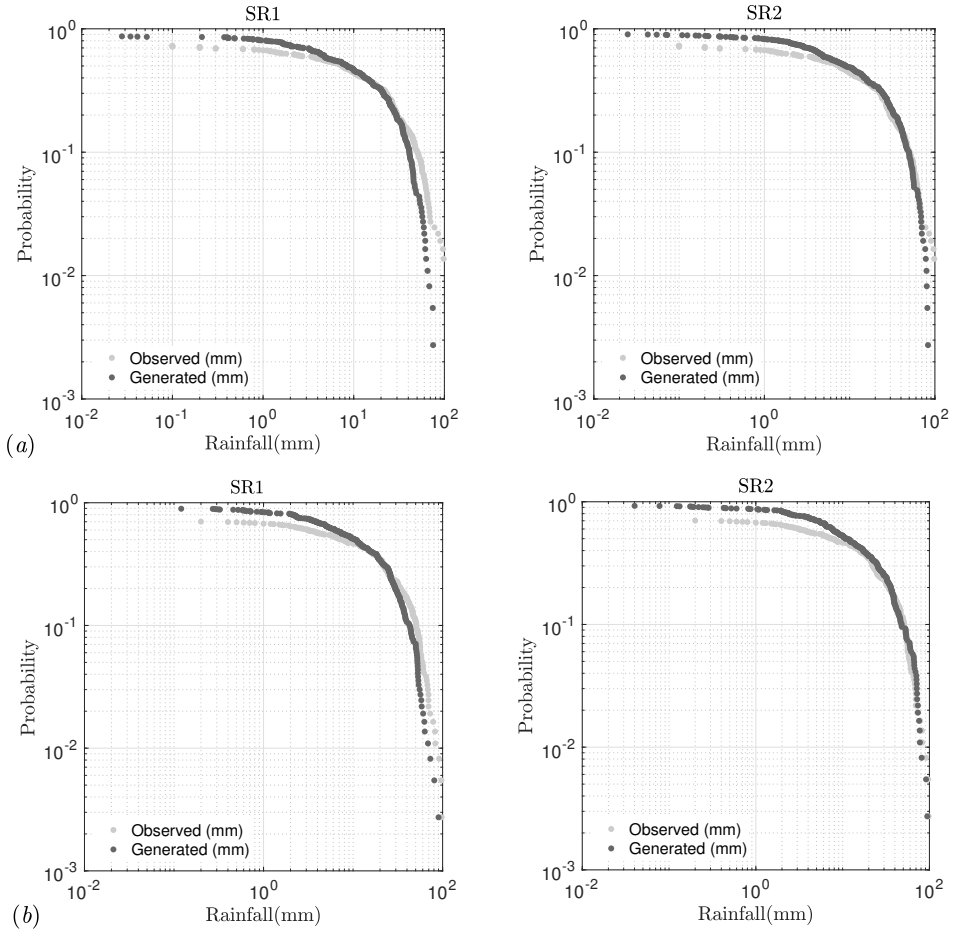

Figure 11.: Exceedance probability comparison (8 day resolution): (a) Pucara and (b) Santa Rosa.