Theoretical and Applied Climatology (2019) 137:1089–1104
https://doi.org/10.1007/s00704-018-2653-8

ORIGINAL PAPER

Spatio-temporal estimation of climatic variables for gap filling and record extension using Reanalysis data

David Morales-Moraga[1,2] · Francisco J. Meza[1,2] · Marcelo Miranda[2] · Jorge Gironás[1,3,4,5]

Received: 18 April 2017 / Accepted: 1 October 2018 / Published online: 9 October 2018

Abstract

The availability of reliable meteorological records is crucial for the development of a number of environmental studies. Unfortunately, these records are not always complete, usually show errors and/or have an insufficient length. This paper presents a gap filling and data record extension methodology for minimum temperature, maximum temperature, and precipitation. It uses climatic information from the NCEP-NCAR Reanalysis project, identifying pixels (grid cells) within a Reanalysis domain that have the highest Pearson's correlation coefficient with the variable of interest. Nine stations in the Maipo River basin (Santiago, Chile) were selected for a reconstruction experiment (from 1950 to 1970) and a subsequent gap filling experiment (from 1970 to 2012). A generalized linear mixed model with a bidirectional stepwise fit procedure was used to model temperature, whereas precipitation occurrence was represented using a generalized linear mixed model with binomial distribution, and precipitation amount used an exponential generalized linear model. The performance of the algorithm was compared with inverse distance weighting and spline interpolation methods and further evaluated using the Standardized Precipitation Evapotranspiration Index, contrasting real versus modeled data. Values of the coefficient of determination averaged 0.76 (0.74–0.84) minimum temperature, 0.73 (0.73–0.81) for maximum temperature, and 0.68 (0.51–0.78) for precipitation. Root-mean-squared error was around 1.5 °C and 5 mm for temperature and precipitation, respectively. The model explains local variation of climatic variables and indicators and can be replicated anywhere, as the *Reanalysis* data are easily accessible and have a worldwide coverage.

Keywords Temperature · Precipitation · Record extension · Gap filling · Reanalysis data

1 Introduction

Access to reliable time series of climatic variables with sufficient length is critical for the study of several geophysical and environmental processes (Beniston et al. 2012). Changes in environmental conditions (Beniston et al. 2012; Kottek et al. 2006); vegetation patterns (Sandholt et al. 2002; Weng et al. 2004), as well as changes on growth, development; and the adaptation of plants and animals (Zavala 2004; Pörtner 2001) are usually explained by climatic variables.

Unfortunately, climate records with the desired length are not always available. Indeed, the densification of weather networks has only occurred recently with substantial investments in standard weather stations and the incorporation of automatic weather stations. According to the National Climatic Data Center, the number of weather stations in the USA varied from little more than 3000 in 1900 to almost 95,000 in 2000. In addition, available data sets usually show discontinuities due to sensor failure or human error.

Several methods have been developed to fill out discontinuities in temperature and precipitation time series. Some of them are deterministic (producing the same output for given initial conditions) and others are stochastic, which provide likely realizations of the variable of interest.

✉ Francisco J. Meza
fmeza@uc.cl

[1] Centro Interdisciplinario de Cambio Global, Pontificia Universidad Católica de Chile, Santiago, Chile

[2] Departamento de Ecosistemas y Medio Ambiente, Pontificia Universidad Católica de Chile, Santiago, Chile

[3] Departamento de Ingeniería Hidráulica y Ambiental, Pontificia Universidad Católica de Chile, Santiago, Chile

[4] CIGIDEN Centro Nacional de Investigación para la Gestión Integrada de Desastres Naturales, CONICYT/FONDAP/15110017, Santiago, Chile

[5] CEDEUS Centro de Desarrollo Urbano Sustentable, CONICYT/FONDAP/15110020, Santiago, Chile

Gap filling methods usually rely on time series for temporal estimation and spatial neighbors as regression variables to extrapolate values in space (Tardivo and Berti 2014). For the case of temperature, Daly (2006) compared six commonly used methods for estimating missing climatic data and showed that Daymet (local regression), PRISM (local regression), and a Regional regression perform better. Moreover, the author also noted a significant influence of the coast (included as distance from coastline) over minimum temperature, especially at scales ≤ 10 km. Tardivo and Berti (2012) used data from a set of stations within a regional domain to fill gaps in extreme temperature data, interpolating values using only meteorological variables and without the inclusion of geographic or spatial variables that could have improved estimations.

In the case of rainfall, literature shows several examples of gap filling and data generation methods according to climate projections from general circulation models (e.g., Diez et al. 2005; Maidment et al. 2012; Nagata 2011; Rojas et al. 2010; Schmidli et al. 2006; Fowler et al. 2007; Vrac and Naveau 2007). Another common approach is the use of stochastic downscaling techniques from global circulation models, but in this case, estimates of individual values are only likely realizations of climate conditions (Castro et al. 2013). In general, these methods usually capture trends but tend to be biased when are disaggregated in time (Casanueva et al. 2012). Other methods use algorithms based on existing data sets. For instance, Ramos-Calzado et al. (2008) presented a method based on error propagation theory taking into account the uncertainty of the precipitation measurements. Another example of this approach is found in the method named CUTOFF (Feng et al. 2014), where the estimate of a missing value is obtained using similar observed temporal information from the nearest spatial neighbors.

One common problem of simple gap filling methods is that they do not always consider local topography due to the relatively coarse grid size (Schmidli et al. 2006) having problems when representing rainfall variability observed as a consequence of specific local conditions (Colle 2004). Climatic processes respond to a number of topographic variables, such as a slope, aspect, latitude, and longitude, which are correlated with spatial gradients (e.g., the coastal influence on temperature) across the land (McCune 2007; Daly 2006). For this reason, Digital Elevation Models (DEM) are often the most useful variable for spatial estimation of precipitation (Hong et al. 2005; Lookingbill and Urban 2003; Ruiz-Arias et al. 2009), allowing us to capture the altitudinal gradient of these variables.

The use of information from remote sensing is another alternative to account for local scale effects (see examples in Bustos and Meza 2014, Sobrino et al. 2004; Suga et al. 2003; Wan and Li 1997), especially when satellite observations have a fairly good spatial resolution (i.e., between 30 m and 1 km), and is able to capture changes in temperature as a function of land physiography. Unfortunately, these satellite observations have moderate to low temporal resolution, and their use is limited by the presence of clouds. In addition, the moment in which the sensor captures land images may not be adequate to infer climatic conditions. For instance, if maximum temperatures are usually observed after midday, and the satellite passes before noon, the consequent estimate from the processed image can only be regarded as a proxy for maximum temperature (Sandholt et al. 2002).

Regarding record extension, only few methods deal with meteorological time series and their extension into the past (Begert et al. 2005; Perry and Hollis 2005; Sherwood et al. 2008), mainly due to the lack of covariates with the same spatial and temporal structure as the data series. In these cases, estimates not only tend to be biased, but their reliability is sometimes questionable (Chen and Hwang 2000). Some examples of record extension can be found in the case of frost days (Perry and Hollis 2005), the use of memory processes for the case of temperature (Blender et al. 2008), and an interesting approximation followed by Jung-Woo and Yakov (2010) to combine artificial neural networks and regression trees. The authors showed that this method improved accuracy and was more robust when compared to artificial neural networks and regression trees alone.

The NCEP/NCAR Reanalysis project corresponds to a systematic effort to produce data sets for climate monitoring and research that date back to 1948 (Kalnay et al. 1996; Kistler et al. 2001). This data set has been compared with meteorological observations that are not part of the assimilation network to develop forecast systems and/or describe local weather behavior (Bengtsson et al. 2004; Beniston et al. 2012; Bojanowski et al. 2014; Fuka et al. 2013; Harnik and Chang 2003; Kubik et al. 2012; Linares-Rodriguez et al. 2011; Maidment et al. 2012; Montecinos and Aceituno 2003; Wright et al. 2009). Bao and Zhang (2012) used four different Reanalysis products (i.e., NCEP/NCAR, NCEP–CFSR, ERA-Interim, and ERA-40) to evaluate relations between variables such as temperature, relative humidity, and wind speed on the Tibetan Plateau and the same variables measured in a network of 11 radiosondes stations (none of them assimilated into these four Reanalysis data sets). They found consistency between the values of temperature and wind speed measured in the field with those delivered by Reanalysis for an average of 3 months, but not for relative humidity, for which a significant bias was found. The Reanalysis NCEP/NCAR provided the most consistent results. Nevertheless, these relationships can be more consistent in certain places, given the amount of remotely sensed data that provide real information for assimilation interpolations (Ramella and Haimberger 2014).

The use of Reanalysis variables for climatic studies and applied meteorological models has been described before (Bao and Zhang 2012; Beniston et al. 2012; Betts et al. 2006; Fuka et al. 2013; Harnik and Chang 2003; Maidment et al. 2012; Saha et al. 2010; Wright et al. 2009), but few studies have tried to find temporal or spatial dependences of environment variables with predictors from the Reanalysis projects. Linares-Rodriguez et al. (2011) used four variables (total cloud cover, surface temperature, total water vapor column, and total ozone column) from the ERA-interim Reanalysis (Simmons et al. 2007) to generate global radiation data at a daily scale through artificial neural networks. Kubik et al. (2012) used information from MERRA Reanalysis (Rienecker et al. 2011) to obtain wind data through simulations.

This study proposes a methodology for the estimation of maximum and minimum temperatures and precipitation using variables from the NCEP/NCAR Reanalysis project at daily scale. Since the spatial correlation of climatic variables is somehow captured by large-scale features of the atmosphere detectable within the Reanalysis data, it should be possible to find statistical relationships between large-scale variables and local ground observations. In this context, a downscaling methodology for low spatial resolution Reanalysis data (2.5° Lat. × 2.5° Lon.) was generated to estimate daily maximum (Tx) and minimum (Tn) temperatures and Precipitation (Pp) in order to fill data gaps and extend time series back to 1950. The method was applied to nine locations in the Maipo River basin (Central Chile). The estimates are the result of the interaction of observed data at the station level and Reanalysis grid values with high statistical similarity within a region representing the local climate of the area. This joint approach has already been reported by some authors but only as a downscaling approach (Bastola and Misra 2014; Brands et al. 2012; Hwang et al. 2013; Misra et al. 2012; Yoshimura and Kanamitsu 2008); thus, its use in gap filling and record extension needs to be evaluated.

2 Methods

2.1 General characteristics of the study area

The region under study corresponds to the Maipo basin in Central Chile (33° S). Mean annual temperature values are around of 14 °C (20 °C in summer and 7.5 °C in winter). Rainfall is concentrated in the winter months (June to August) with 80% of the total ~350 mm falling during these months (Fig. 1). Snow accumulation occurs above 1500 m.a.s.l. during winter. The region is known to have a significant El Niño Southern Oscillation (ENSO) footprint. Montecinos and Aceituno (2003) reported a relationship between precipitation anomalies and ENSO, indicating a statistical dependence between the amount of rainfall in a year and the increase in sea surface temperature in the eastern equatorial Pacific.

2.2 Meteorological data

We selected ten meteorological stations with daily records of temperature and precipitation (Table 1), seven from the Dirección General de Aguas (National Water Directorate; DGA) and three from the Dirección Meteorológica de Chile (Meteorological Directorate of Chile; DMC). These stations are spread out and represent different locations within the Maipo basin (Fig. 2). Nine stations were used in the reconstruction experiment. The remaining one (Quinta Normal station) is the only one with a complete time series from January 1, 1914 to the present, so it was used as an independent data set in the algorithm. We evaluated the procedure with and without the inclusion of this reference station as a way to measure the effect of its presence and contrast results in the reconstructed time series against real data.

Prior to the development of the model, Tn and Tx data had been grouped by month and transformed into standardized anomalies (mean = 0 and standard deviation = 1) to avoid heteroscedasticity in residuals (Royer and Poirier 2010). In each month, we evaluated the presence of outliers, removing them following the methodology described by Laurikkala et al. (2000) for univariate data.

Since Pp is represented as an occurrence and amount process, we first transformed their values into a time series of binary variables representing occurrences (1) and non-occurrences (0) of the event (defined strictly as days with precipitation greater than zero or not, respectively). For the amount process, we selected only values greater than zero and expressed them as standardized anomalies.

We noted a non-homogeneous distribution of the frequency values registered by precipitation gauges over the ten analyzed stations (Fig. 3). Data shows extraordinarily large frequencies of values in 0.5 classes that suggest a bias in precipitation records, normally associated to measurement errors. Given the large number of observations that fall in each of the 0.5-mm classes, we grouped all values in 0.5-mm range intervals (i.e., those values between 0.1 and 0.4 were rounded to 0.5, those between 0.6 and 0.9 were rounded to 1.0, etc.). Alternatively, we could have introduced assumptions on the distribution of values to disaggregate the data with the risk of invalidating the model for gap filling. Note that for a general methodology, this step is not essential and is to be regarded as a convenient solution, given the nature of the observations available.

2.3 Reanalysis variables

We selected 14 variables from the Reanalysis NCEP/NCAR (Table 2). We used 12 of them for Tn and Tx estimation and the full set (including precipitable water surface and precipitation rate) in the model for the estimation of precipitation occurrence and amount at each station. We selected data from the 500-mbar pressure level and surface data. Some of the variables used are available at both levels, like air temperature and the east-west (u) and south-north (v) components of the Geostrophic Wind (Table 2). We selected the surface level,

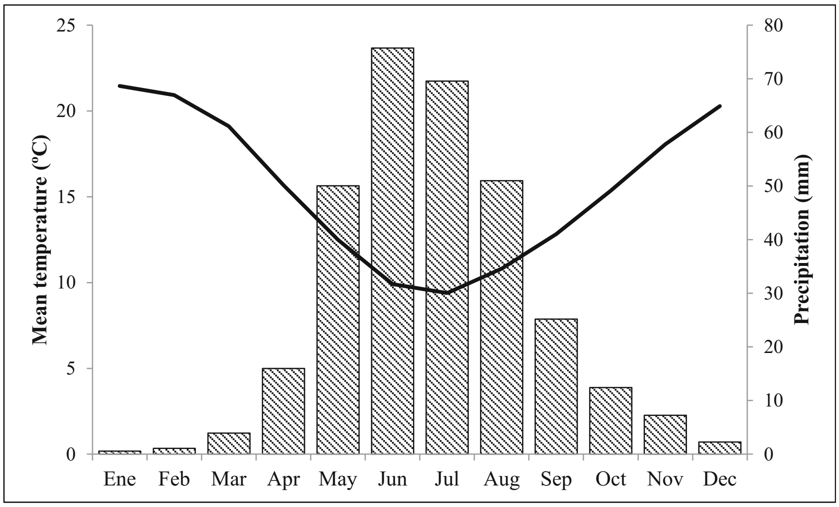


Fig. 1 Seasonal variation of mean temperature (black line) and precipitation (bars) at the Quinta Normal meteorological station

because it is the closest to the level at which ground observations are taken, whereas the 500-mbar level variables were chosen because some meteorological variables, like precipitation and frosts, are sensitive to large-scale phenomena such as the position and strength of the South American Anticyclone. Moreover, variables from the 500 mbar are normally included in weather forecasting models (Gershunov and Cayan 2003; Flannigan and Wotton 2001) as they usually show significant correlation with ground observations.

2.4 Proposed algorithm

We restricted the domain of Reanalysis data to the coordinates 17.5° to 50° S and 65° to 175° W (Fig. 4). The method is independent of the selected area as it is based on selecting the pixels with the highest level of correlation. However, the two major controllers of Central Chile's climate can be found in this region (the South American Anticyclone and the El Niño Southern Oscillation phenomenon). Although no mechanistic explanation is provided in this model, and any highly correlated pixel could be used, we believe that the use of this window reduces the risk of incorporating spurious relations. Nevertheless, the majority of the cells selected are found in the 5 × 5 square grid whose centroid corresponds to the cell where the basin is located. Therefore, the method could be applied using only the closest neighbors.

We extracted 546 time series (the number of pixels in the restricted Reanalysis domain) for the 14 NCEP/NCAR Reanalysis variables. Data of each Reanalysis variable were previously grouped by month and expressed as standardized anomalies over the period of interest (1950–2012). We correlated each variable of each pixel with the temperature and precipitation series of the nine locations for each month. Then, for each Reanalysis variable, we searched for the pixel (grid cell) having the best (i.e., highest in absolute value) Pearson's correlation coefficient and extracted its monthly data to create a vector of predictors with the highest linear individual correlation. In a few cases, a log or square root transformation was needed to obtain higher correlation values (Wang and Murphy 2004). The process is repeated independently for the three climatic variables (Fig. 5).

Table 1 General characteristics of selected stations in the Maipo basin

Station Name	Acronym	Lon.	Lat.	Altitude (m)	Starting date	Ending date
El Yeso Embalse	EYEm	70° 05′	33° 40′	2475	April 13, 1962	August 21, 2012
Pirque	Pir	70° 35′	33° 40′	659	October 10, 1967	August 31, 2012
Cerro Calán	Ccal	70° 32′	33° 23′	848	June 18, 1976	August 31, 2012
Rungue Embalse	Run	70° 54′	33° 01′	700	January 01, 1967	May 31, 2007
Melipilla	Mel	71° 12′	33° 41′	168	June 02, 1971	August 31, 2012
Los Panguiles	LP	71° 01′	33° 26′	195	November 03, 1980	August 31, 2012
Quinta Normal	QN	70° 40′	33° 26′	527	January 01, 1950	December 31, 2012
Tobalaba	Tob	70° 32′	33° 27′	650	January 01, 1969	December 31, 2012
Pudahuel	Pud	70° 47′	33° 32′	493	January 04, 1966	December 31, 2012
Los Cerrillos	Cerr	70° 42′	33° 28′	511	January 02, 1963	February 07, 2006

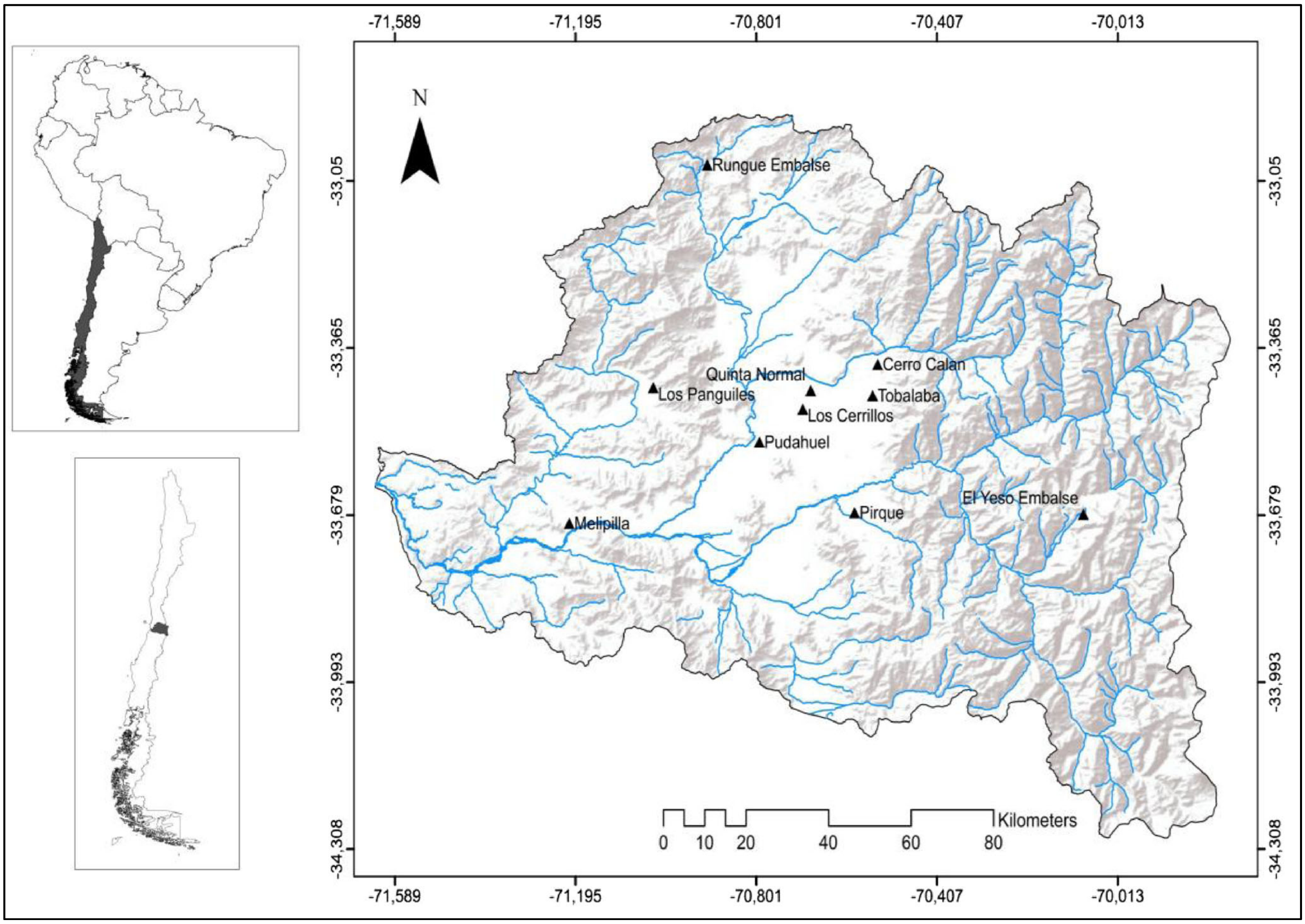


Fig. 2 Spatial distribution of ten selected stations over the Maipo basin

2.4.1 Temperature estimation

After selecting the Reanalysis standardized values of the corresponding months with maximum absolute correlation, a reconstructed time series of monthly dependent and independent variables was built. Orthogonal scores of a principal component analysis with Varimax rotation (Drosdowsky and Chambers 2001) were used to avoid multicollinearity and verify those Reanalysis variables that show the highest influence on the selected scores used in the regression model, according to the eigenvalues selection proposed by Peres-Neto et al. (2003). Finally, for Tn and Tx, a generalized linear mixed

Fig. 3 Histogram of precipitation in the Cerro Calán station for the 0–5-mm range

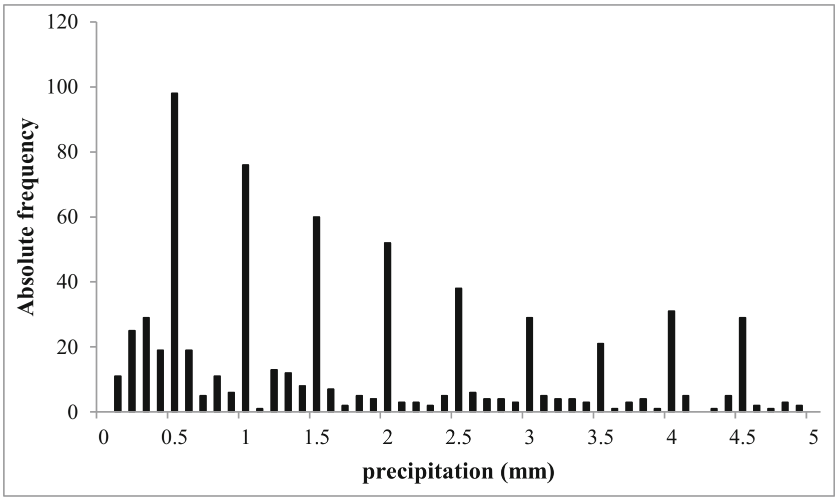


Table 2 Reanalysis variables used in the algorithm for gap filling and record extension

Variable	Unit	Short name
Air temperature at 2 m	K	AT2m
Air temperature at 500 mbar	K	AT500mbar
Geopotential height	m	HGT500mbar
Net longwave radiation	W m^{-2}	NLR2m
Precipitable water surface	Kg m^{-2}	PWS
Precipitation rate	Kg m^{-2} s^{-1}	PR
Relative humidity	%	RH500mbar
Sea level pressure	Kg m^{-1} s^{-2}	SLP
Sensible heat flux	W m^{-2}	SHF2m
Specific humidity	kg kg^{-1}	SH500mbar
U-wind at 2 m	m s^{-1}	UW2m
U-wind at 500 mbar	m s^{-1}	UW500mbar
V-wind at 2 m	m s^{-1}	VW2m
V-wind at 500 mbar	m s^{-1}	VW500mbar

model (identity link function) was fitted (Eq. 1) assigning month as an additional variable. The model is fitted following a bidirectional stepwise procedure using Akaike information criterion (Akaike 1974) to select the most parsimonious model.

$$\hat{\mu}_{ij} = \left(\beta_0 + \beta_j + \left(\sum_{h=1}^{12} \beta_h \times X_{ih}\right) + \theta \times \mu_{ij\_QN}\right) \qquad (1)$$

Here, $\hat{\mu}_{ij}$ corresponds to the estimate of the anomaly of Tn or Tx for day i in month j; β_0 is the model intercept; β_j corresponds to the coefficient associated to each month; β_h corresponds to fixed effect coefficients for each principal component, X_{ih}; θ is the coefficient associated to recorded value of the anomaly of temperature (either Tx of Tn) at Quinta Normal station (μ_{ij_QN}), which is the station with a complete record in this experiment; X_{ih} corresponds to the principal component scores of the standardized predictors from the Reanalysis data (i.e., the 12 variables from surface and 500-mbar levels).

The individual contribution of each Reanalysis variable is evaluated from the PCA loadings. We examined the frequency of absolute values of significant loadings ($p > 0.05$) over 0.3.

2.4.2 Precipitation estimation

We obtained the binary vector of occurrences (1) and non-occurrences (0) of precipitation for each station. Sometimes, a threshold different than zero is used to classify days at dry/wet days. This method is not sensitive to the selected threshold unless it is applied to a very dry region where the selection of a threshold modifies the relative frequency of days with precipitation. Since precipitation shows a strong seasonal behavior, and summer months usually have few rain events, we fitted a model by season, generating four subtables (in locations with a higher frequency of rainy days, it would be advisable to do this by month). Two more Reanalysis variables were included in this case: precipitable water surface (PWS) and precipitation rate (PR). Once again, we searched for the best linear relationship within station binary data and Reanalysis anomalies (Wang and Murphy 2004).

For the occurrence process, only standardized anomalies of predictors are needed. The algorithm applies a GLMM with a logit link function to estimate the probability of the occurrence of precipitation. In this case, the season enters as an additional variable in the model. The occurrence of a precipitation event on a day is determined to occur when the estimated probability exceeds the 0.5 threshold. Model coefficients and PCA best predictor vector scores were found using a bidirectional stepwise process.

We followed the same method used for temperature estimation (Eq. 1) to detect the presence of recurrent specific components in the model using loading analysis. In this case, we used the logit function of the predictand $\hat{\mu}_{ij}$ which is calculated as:

$$\text{logit}\left(\hat{\mu}_{ij}\right) = \log_e\left(\frac{\hat{\mu}_{ij}}{1-\hat{\mu}_{ij}}\right) \qquad (2)$$

Note that the variables involved in the estimation of precipitation occurrence would also be present in a model that predicts precipitation amounts. In fact, there is a strong (nonlinear) relationship between the estimates of the precipitation occurrence in the observed time series and the amount registered (Fig. 6). With this, we can estimate the daily precipitation amounts from the probabilities of occurrence obtained with the binomial GLMM. To achieve this, a generalized linear model with a gamma link function was fitted. Here, the predictor variables are the log values of the discretized rainfall amount for Quinta Normal (a way to assess the spatial local effect of a representative neighbor station) and the estimated probabilities of occurrence for each station obtained from the GLMM in Eq. 2.

2.5 Performance of the algorithm

To assess the performance of our method, we compared it against two common interpolation procedures, the inverse distance weighting and spline methods. We selected the values recorded in 1979 at the Pudahuel and Melipilla stations to evaluate the results of the applied methods, and we used Quinta Normal, Cerrillos, and Pirque station as neighbors for the IDW and spline methods. The year 1979 was selected, since temperature and precipitation records are complete for all five.

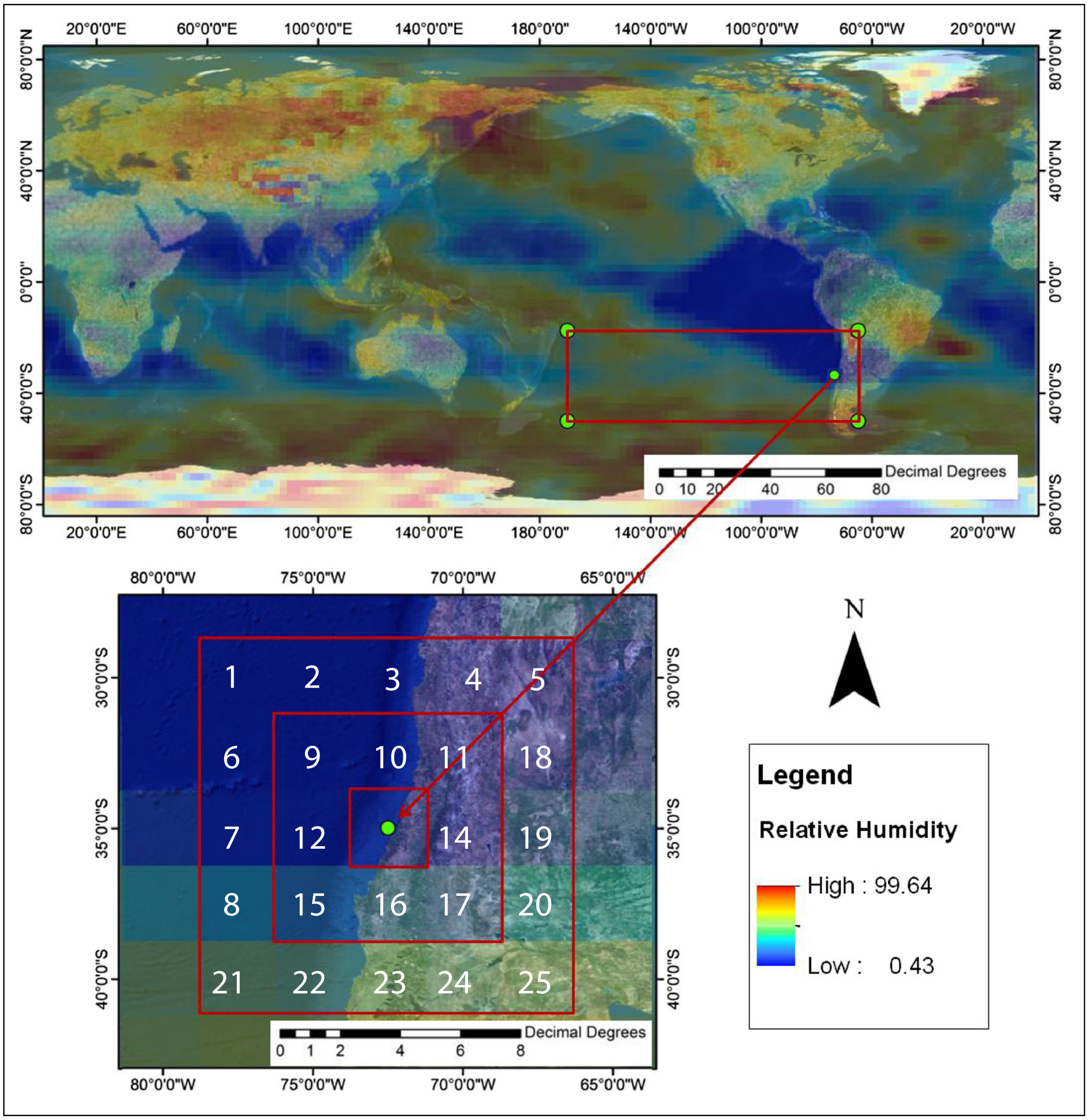
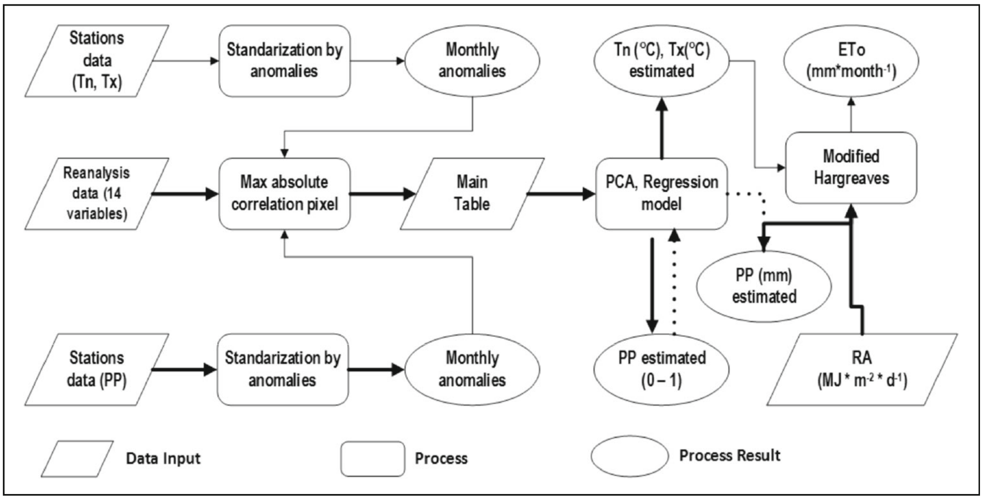


Fig. 4 Selected spatial window for searching correlations between station data and Reanalysis variables

Fig. 5 Diagram of the gap filling and record extension algorithm. Additional steps for SPEI calculations are shown



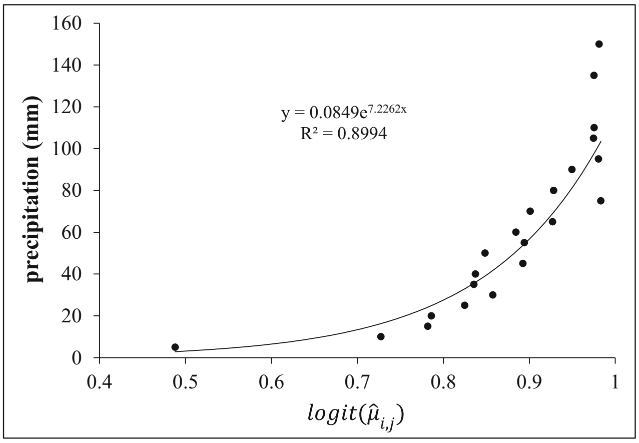


Fig. 6 Occurrence probabilities of a precipitation event calculated at 5-mm intervals from a generalized linear mixed model for the Cerro Calán station

In addition to the comparison described before, and as a way to evaluate how dependent the method is on pixel (grid cell) selection with the highest absolute correlation value, we applied the proposed method using not the best pixel in terms of absolute correlation coefficient, but the Reanalysis data for the pixel (grid) where the station was located. In this case, we compared the results obtained for both methods for the Cerro Calán station, using the model R^2 and RMSE values.

At each station, we performed a cross-validation experiment to test the robustness of the algorithm. We randomly selected a subsample containing 10% of the observations. The model was then fit using the remaining 90% of the observations and applied to the excluded data set to evaluate the performance of the method comparing observed vs estimated values and calculating goodness of fit statistics. The procedure was repeated 100 times to determine mean and variance of RMSE, MAE, Bias, etc. In addition, the Standardized Precipitation Evapotranspiration Index (SPEI) (Paulo et al. 2012; Vicente-Serrano et al. 2010) was used as a pseudo-independent validation method, as it depends on nonlinear combinations of the estimated variables. Potential evapotranspiration is needed for the SPEI calculation, which is estimated from the modified Hargreaves equation (Doggers and Allen 2002). We calculated extraterrestrial radiation (Ra) using the latitude of the stations. We computed a 3-month aggregated SPEI value for the time period 1980–2010 in all stations separately. R^2 and RMSE were obtained from the regression between two SPEI series and used to evaluate whether the algorithm was capable of capturing temporal trends and could be used for filling and extension of climatic variables at the basin level.

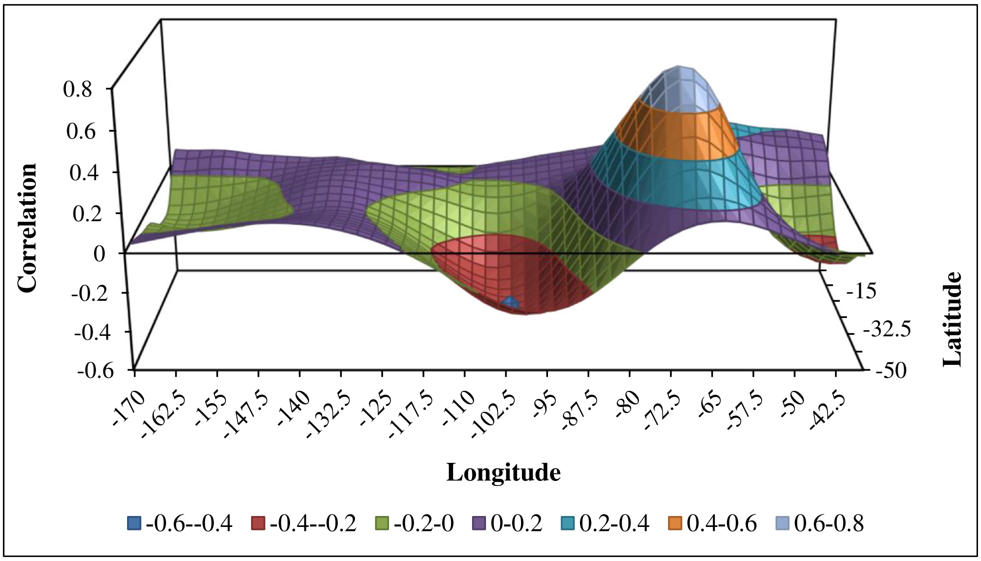
All the steps and procedures described were conducted in the statistical software R, using the NCEP/NCAR Reanalysis data. Required packages were “ncdf” (Pierce 2011), “RNCEP” (Kemp and Kemp 2012), “reshape” (Wickham 2007), “nlme” (Pinheiro et al. 2012), “lme4 (Bates et al. 2012), “lmerTest” (Kuznetsova et al. 2013), and SPEI (Beguería and Vicente-Serrano 2013).

3 Results

3.1 Minimum and maximum temperatures

The highest correlation values between the Reanalysis data and temperature variables were found in grid cells (pixels) that are located close to the stations. Figure 7 shows an example of Pearson’s correlation values for the Reanalysis variable air temperature at 500 mbar (ATM500) and Tn for the Cerro Calán station. Since Reanalysis values are based on several coarser observations, one would expect this behavior and eventually restrict the domain only to pixels that are in the surroundings of the location whose data record is to be

Fig. 7 Spatial distribution of correlation coefficients between the Cerro Calán station data (33.2° S, 70.5° W) and Reanalysis variable ATM500mbar for July



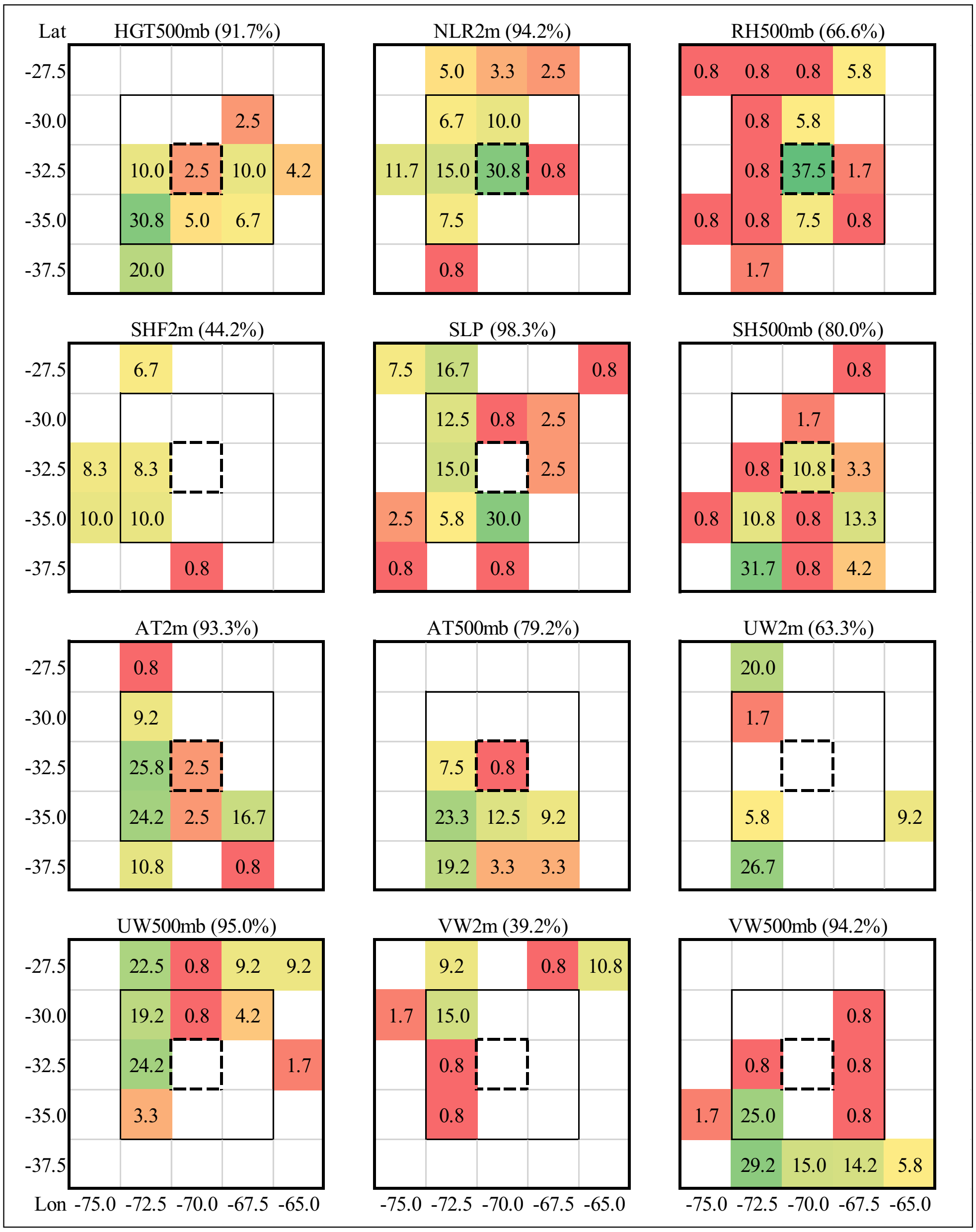


Fig. 8 Relative frequency of selection of pixels (grid cells) in the 5 × 5 grid whose centroid corresponds to the location of the basin (dashed grid) for the estimation of minimum temperature. See Table 2 for the description of the Reanalysis variables

expanded (or missing data is to be filled). In accordance with what is expected theoretically, other variables such as the 500-mbar geopotential height (HGT500mbar), relative and specific humidity at 500mbar (RH500mbar, SH500mbar), and air temperature at 2 m (AT2m), show positive associations. The opposite is verified for the remaining predictors.

As a way to generalize the previous results, we calculated correlations between a regional mean of minimum and maximum temperatures as well as mean precipitation and correlated them with the Reanalysis variables for a domain composed only of the closest 25 pixels. Figure 8 presents the frequency with which each pixel is selected as the best linear correlation and, thus, becomes a candidate for its inclusion in a GLMM for the case of minimum temperature. (The same conclusion was obtained when looking at the figures corresponding to maximum temperature and precipitation with a slight tendency of maximum temperature to select more pixels outside the closest domain; data not shown.) It is clear that pixels closer to the basin centroid (33.58° S, 70.59° W) are more likely to be selected based on highest absolute correlation.

3.1.1 Precipitation

In the case of precipitation, we found that the Reanalysis variables PWS, PR, RH500mbar, SH500mbar, AT2m, UW2m, UW500mbar, and VW500mbar have a consistent positive association with almost all station data. In the case of AT2m, the results were somewhat surprising, because in this type of

Table 3 Summary of model fit statistics for minimum (Tn) and maximum (Tx) temperature

		Without Quinta Normal			With Quinta Normal		
Variable	Station	R^2	AIC	RMSE	R^2	AIC	RMSE
Tn	Ccal	0.78	17,472.19	1.40	0.81	15,482.97	1.23
	Cerr	0.73	19,935.88	1.54	0.75	19,050.80	1.46
	EYEm	0.77	18,241.74	1.43	0.81	16,290.06	1.25
	LP	0.71	21,113.82	1.63	0.74	19,648.69	1.52
	Mel	0.70	21,226.81	1.63	0.74	20,145.84	1.52
	Pir	0.72	20,343.96	1.57	0.75	19,567.28	1.46
	Pud	0.70	21,280.35	1.63	0.74	19,344.51	1.49
	Run	0.71	26,112.92	1.62	0.76	19,211.03	1.46
	Tob	0.74	19,323.98	1.50	0.77	18,894.94	1.41
Tx	Ccal	0.71	21,122.31	1.63	0.72	20,705.48	1.59
	Cerr	0.72	20,436.50	1.58	0.73	20,049.29	1.55
	EYEm	0.75	18,796.50	1.48	0.76	19,435.48	1.46
	LP	0.73	20,231.07	1.57	0.74	19,689.49	1.54
	Mel	0.69	21,962.32	1.69	0.71	20,925.96	1.62
	Pir	0.70	21,331.65	1.64	0.72	20,796.26	1.61
	Pud	0.71	20,863.57	1.61	0.73	20,595.28	1.58
	Run	0.74	19,627.99	1.54	0.75	19,669.32	1.52
	Tob	0.72	20,569.52	1.59	0.73	20,008.75	1.57

regime, precipitation has a negative correlation with maximum temperature and a positive association with minimum temperature.

In this case, there is also a strong association between Reanalysis variables within the nearest pixels and the precipitation variable of the basin; however, only the variables NLR2m, PWS, PR, RH500mbar, and VW2m concentrated more than 60% of the maximum absolute correlations in the central pixel or its eight neighbors (a figure that increase up to 70% when we included 16 border pixels).

3.2 Estimations

3.2.1 Maximum and minimum temperatures

Since the algorithm for temperature estimation involves the development of one model per station, and that each variable can be estimated with up to 26 coefficients in Eq. 1, we only present the summary statistics of model performance rather than reporting the individual regression coefficients. The mean R^2 value for the estimation of both minimum and maximum temperatures was 0.72 for the case where the reference station of Quinta Normal (QN) was not included, and 0.75 when added (details by station are presented in Table 3). The Cerro Calán station showed the highest correlation ($R^2 = 0.78$) using only the Reanalysis variables; this value increases to 0.81 when QN station was included. The RMSEs for minimum and maximum temperatures were 1.40 and 1.22 °C, respectively. All coefficients are significant at 5% level.

We observe that, in general, the predictive power of the models increased by 2.4% when the reference station QN was included. For the case of minimum temperature, such increase is 3.3%, whereas the RMSE varied from 1.57 to 1.49 °C when QN was added. Furthermore, seasonality is very important. In fact, the use of a coefficient that accounts specific effects at monthly level increases the predictive power of the GLMM by 8.8% as compared to using non-grouped data.

Regarding the contribution of the NCEP/NCAR Reanalysis variables, in all fitted models, the first and second principal components were selected. For Tn, the variables HGT500mbar, SLP, AT500mbar, and VW5000mbar had significant loading values over 0.3 contrasted with the first component (PC1) and were part in five of the nine GLMMs fitted. Using the same indicator, variables NLR2m, RH500mbar, and VW2m are more relevant in PC2. For the case of Tx, the variables HGT500mbar, SHF2m, AT2m, AT500mbar, and UW500mbar are significant in PC1, participating in the estimation for all nine

Table 4 Summary of model fit statistics for the estimation of precipitation amount

	Without Quinta Normal		With Quinta Normal	
Station	R^2	Mean absolute error (MAE)	R^2	Mean absolute error (MAE)
Ccal	0.74	5.76	0.78	5.68
Cerr	0.78	4.22	0.82	3.98
EYEm	0.47	15.15	0.51	11.30
LP	0.65	6.52	0.72	6.49
Mel	0.54	11.32	0.55	8
Pir	0.68	6.24	0.74	5.91
Pud	0.78	3.43	0.83	2.98
Run	0.55	11.11	0.59	9.43
Tob	0.64	7.06	0.69	6.20

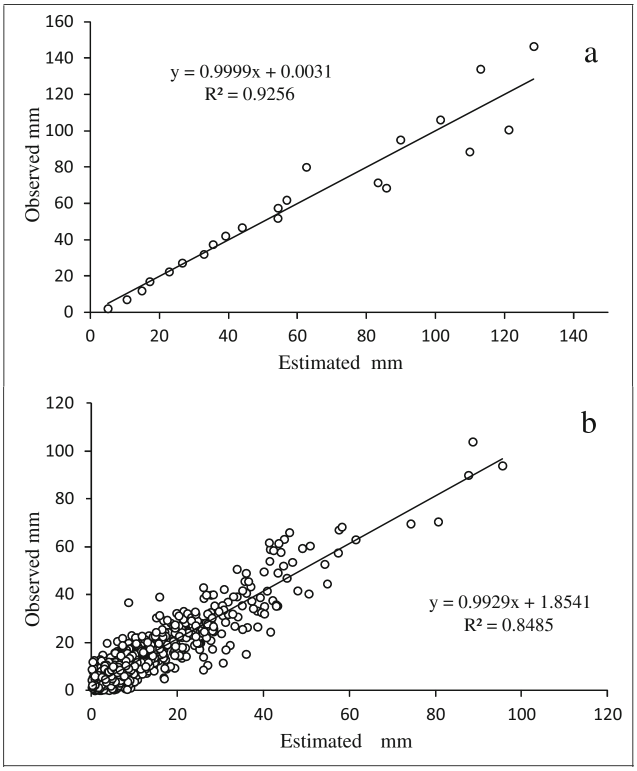


Fig. 9 **a** Estimated vs observed precipitation data for discretized values in 5-mm intervals. **b** Estimation of the amount of precipitation greater than 0.5 mm. Both graphics refer to the Cerro Calán station

stations. PC2 had NLR2m, RH500mbar, AT2m, and VW2m as the most significant variables.

HGT500mbar and AT500mbar had significant loading values with PC12. The scores of this principal component related with HGT500mbar participated in eight of the nine estimation for Tn, and in all stations for Tx. AR500mbar participated thought the PC12 in the 18 possible estimations. In 17 of the 18 estimations, at least nine of the 12 principal components participated. For the Embalse el Yeso station, only seven PCs were needed to obtain a model with the lowest AIC.

3.2.2 Precipitation

The principal components PC1 and PC4 were the only variables used to estimate the occurrence of precipitation events in the GLMM, explaining 14.38 and 8.64% of the variance. Individually, PC1 loadings were associated with HGT500mbar (above 0.3 in the all the 9 stations), UW500mbar (used in 8 stations), and AT500mbar (5 stations). Meanwhile, PC4 was related to RH500mbar (7 stations), SH500mbar (6 stations), SHF2m, and VW500 (5 stations each). Despite PC2 having significant loadings in eight of the nine stations, with the additional variable PWS, this principal component was selected in only three cases (Cerro Calán, Cerrillos, and Embalse el Yeso). The other added variable associated to precipitation (PR) had a presence in six out of nine of the loadings of PC6 and was used in the GLMM in six occasions.

Occurrence probabilities of precipitation were estimated for each station. The performance of the binomial GLMM increased by 25.5% on average (6.2–39.6%) when the reference station (QN) was added to the model. Mean hit rate values (Wilks 2006) were around 90% in all stations, with the lowest being Melipilla (87%) and the highest El Yeso Embalse (95%).

Unlike the estimation of extreme temperatures and the precipitation occurrence probabilities, the effect of the reference station in the estimation of precipitation amounts is very important (Table 4). The poorest result is observed in the station El Yeso Embalse probably due to the fact that the station has the highest elevation, and thus, some of the precipitation falls as snow in winter time. In addition, we noticed that in all precipitation estimations, we did not obtain an estimated value greater than the highest observed value during a single day. Figure 9 shows water precipitation amounts for discretized data and the final disaggregation to the complete time span for Cerro Calán station.

3.3 Comparison with other methods and validation

We compare the results with a simple and yet commonly used interpolation method. We used a regional regression model with several correlated stations and geographical variables derived from a digital elevation model as covariates (lat, long, distance from shoreline, and major water bodies, vegetation indices derived from remote sensing). The model was run for the period 2000–2012 at 1-km spatial resolution. We applied it to Pirque and Los Panguiles stations with the following results: min temperature (Tn), at los Panguiles had an RMSE value of 2.52 °C and R^2 of 0.827, while at Pirque, the values were 0.8 °C and 0.876, respectively. For the case of maximum temperature (Tx), values found for Los Panguiles were an RMSE value of 2.03 and R^2 of 0.907, and at Pirque, values were 1.67 °C and 0.905, respectively.

We then used IDW interpolation method to estimate Tx and IDW and spline interpolation for Pp. We estimated the values registered at Pudahuel and Melipilla stations during year 1979. These methods were compared with the results obtained using the method based on Reanalysis without the inclusion of the observations of Quinta Normal as covariate.

IDW outperforms the method based on reanalysis data. At Pudahuel station, RMSE values of 0.85 and 3.3 °C for IDW and the method based on Reanalysis were obtained, whereas RMSE values of 2.2 and 3.4 °C were obtained for these methods at the Melipilla station. IDW better reproduce heat waves (or extreme cold temperatures) as it uses observed values in the neighbor stations at the moment of interpolation. This advantage becomes more evident if the neighbor stations are close in space, as they tend to be highly correlated.

In the case of precipitation, the method based on the Reanalysis data outperforms IDW and spline methods for

Table 5 Summary of goodness of fit statistics of a cross validation experiment for each station. Values in parentheses correspond standard deviations of 100 simulations

Variable	Station	RMSE	MAE	Bias	R^2
Tn	Ccal	1.66 (0.05)	1.23 (0.03)	0 (0.05)	0.81 (0.01)
	Cer	1.91 (0.04)	1.45 (0.03)	− 0.01 (0.05)	0.75 (0.01)
	EYEm	1.7 (0.04)	1.25 (0.03)	0 (0.05)	0.81 (0.01)
	LP	1.94 (0.04)	1.49 (0.03)	0 (0.05)	0.75 (0.01)
	Mel	1.97 (0.05)	1.51 (0.03)	0 (0.06)	0.74 (0.01)
	Pir	1.92 (0.05)	1.46 (0.04)	0.02 (0.05)	0.75 (0.01)
	Pud	1.94 (0.05)	1.48 (0.04)	0 (0.06)	0.75 (0.01)
	Run	1.91 (0.05)	1.46 (0.03)	0 (0.05)	0.75 (0.01)
	Tob	1.86 (0.04)	1.41 (0.03)	0 (0.05)	0.77 (0.01)
Tx	Ccal	2.03 (0.04)	1.59 (0.03)	0.01 (0.07)	0.72 (0.01)
	Cer	1.99 (0.05)	1.55 (0.03)	0 (0.06)	0.73 (0.01)
	EYEm	1.88 (0.04)	1.46 (0.03)	0 (0.05)	0.76 (0.01)
	LP	1.97 (0.05)	1.55 (0.04)	0 (0.06)	0.74 (0.01)
	Mel	2.06 (0.04)	1.62 (0.03)	0 (0.07)	0.71 (0.01)
	Pir	2.05 (0.04)	1.61 (0.03)	0 (0.05)	0.72 (0.01)
	Pud	2.01 (0.04)	1.58 (0.03)	0.01 (0.06)	0.73 (0.01)
	Run	1.93 (0.04)	1.51 (0.03)	− 0.01 (0.06)	0.75 (0.01)
	Tob	1.99 (0.04)	1.56 (0.03)	0.01 (0.06)	0.73 (0.01)
Pp	Ccal	8.96 (0.36)	6.68 (0.41)	3.63 (0.99)	0.74 (0.02)
	Cer	7.38 (0.74)	4.42 (0.75)	2.29 (0.74)	0.78 (0.03)
	EYEm	25.48 (1.80)	18.25 (1.20)	9.57 (1.02)	0.40 (0.04)
	LP	10.71 (1.55)	7.76 (1.00)	3.30 (1.08)	0.64 (0.04)
	Mel	22.53 (3.41)	16.07 (2.43)	10.38 (2.54)	0.51 (0.06)
	Pir	11.16 (3.05)	7.74 (2.18)	3.09 (1.98)	0.65 (0.09)
	Pud	6.84 (0.61)	4.11 (0.59)	1.82 (0.35)	0.77 (0.03)
	Run	16.02 (0.74)	10.31 (0.52)	4.17 (0.82)	0.55 (0.03)
	Tob	10.82 (0.40)	7.00 (0.33)	2.67 (0.29)	0.61 (0.02)

the Pudahuel station (RMSE of 0.78, 1.24 and 3.44 mm, respectively) and gives better results than the spline method for the Melipilla station (RMSE of 3.94 for Reanalysis, 2.6 for IDW and 14.6 mm for spline). IDW highly depends on the occurrence of precipitation in one station the occurrence in the neighbor stations. Spline functions tend to smooth values, but the effect of distance is very relevant leading to inconsistencies when stations are not strongly correlated.

Results of the cross validation experiment are presented in Table 5. While maximum and minimum temperatures show little variation among stations in terms of RMSE and Bias and show the highest R^2 values, precipitation shows a higher variability. Results are more variable for the case of precipitation. In two stations (El Yeso Embalse and Melipilla), R^2 values are around or less than 0.5; nevertheless, the regression coefficients of the model are all significantly different than zero at 0.05 level contributing to partially explain the variance of the observations and, thus, significantly improve the estimations of missing data.

To test the performance of the method in generating secondary information, we calculated the Standardized Precipitation Evapotranspiration Index. SPEI is an indicator of the temporal changes between dry periods (SPEI < 0) and wet periods (SPEI > 0). We observed a good agreement when comparing SPEI values calculated using the gap filling/record extension method and when using real data (Fig. 10) in both series for the same time period. The average of R^2 was 0.87, varying between 0.78 for Melipilla station to 0.91 for Cerrillos station (Table 6). It is possible to detect climatic trends in all stations consistent with observed data.

4 Discussion

There is no limitation regarding of the plausible Reanalysis variables to be used in the presented methodology: however, the selection of those having the highest temporal correlation with climatic variables increase the predictive power of the

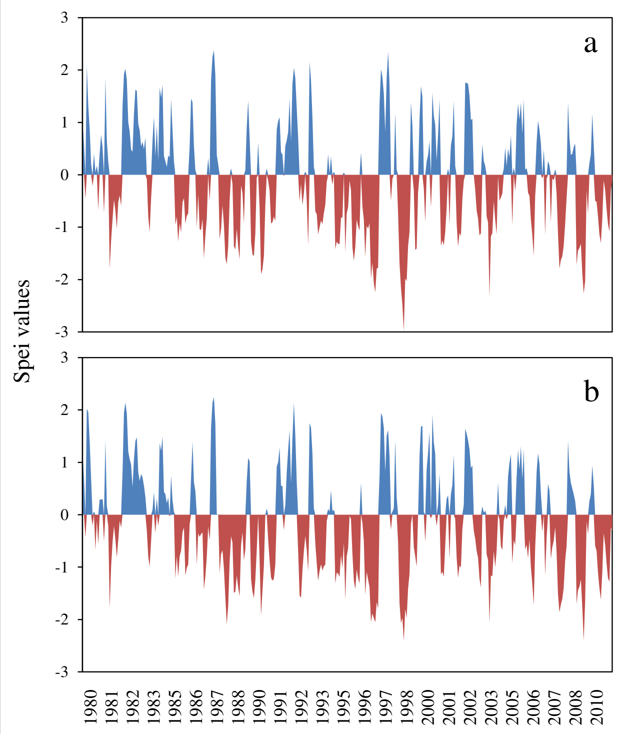


Fig. 10 **a** SPEI values calculated with real Tn, Tx, and Pp data. **b** SPEI values calculated with estimated data of Tn, Tx, and Pp. Both graphics refer to the Cerrillos station ($R^2 = 0.87$)

method, especially on a monthly basis. Therefore, any climatic variable could be potentially estimated for reconstruction of past time series. The main limitation is the initial extension of the candidate time series being reconstructed, considering that a large temporal database is crucial for detecting temporal trends and periodicity. Also, in this case, monthly decomposition needs a sufficient number of years to produce estimates with high goodness-of-fit. This work used data with at least 30 years of continuous data. The length of the time series is related to the consistence of the obtained products, independent of the estimation procedure applied (Bengtsson et al. 2004). Nevertheless, in most cases, this criterion is used for filling temporal data gaps with its own data (Tardivo and Berti 2012) than for data extension in any time direction.

Table 6 Coefficient of determination (R^2) between SPEI values calculated from real and estimated data

Station acronym	R^2
Ccal	0.78
Cerr	0.87
EYEm	0.70
LP	0.74
Mel	0.67
Pir	0.71
Pud	0.73
Run	0.81
Tob	0.76

At the meteorological station level, the values estimated for some particular variable to fill temporal gaps have been the focus in the use of neighbor stations (Sterl 2001; Tardivo and Berti 2014). In this case, the restriction lies in the time series length of the shortest neighbor station, which means that the final result is limited by data length within the neighborhood. With sufficiently large series for calibration, this method can be useful in expanding records and detecting temporal trends in variables like temperature, precipitation, and solar radiation (Betts et al. 2006; Bojanowski et al. 2014; Refslund et al. 2014).

Since 1979, the assimilation algorithm of the NCEP/NCAR Reanalysis began using remote sensing data (Kalnay et al. 1996), modifying the database slightly with respect to the 1948–1971 period. This could be a source of error that may introduce problems in trend detection and be misleading.

In precipitation, the error associated with the observer can be very important. The discretization approach followed here is a solution but might not be the best one, particularly if a particular study requires greater precision in the estimation of the amount of precipitation.

The selection of potential predictors is highly relevant in improving the results of the algorithm. The selection of the Reanalysis variables was conducted by choosing those with a seasonal and meteorological component; their combined effects increase model predictive power (Hwang et al. 2013; Yoshimura and Kanamitsu 2008). In addition, the use of a reference station has a positive effect in improving estimates, especially if a strong correlation with the candidate station is found (Daly 2006), because it provides non-explained local effects. This inclusion is more relevant in the estimation of precipitation than in the case of temperature.

The local effect in each station, mainly influenced by its topography and altitude, can be detected within the heterogeneous area of the Maipo basin. Three stations (Tobalaba, Quinta Normal, and Pudahuel) are located in the Santiago metropolitan area, with a marked heat island effect, which can affect estimates (Buyadi et al. 2013). However, monthly aggregated estimates in the case of extreme temperature and seasonal estimations of rainfall appear to mask this negative effect.

Not all complementary variables showed the same degree of contribution to the models in the algorithm. AT2m, AT500mbar, and RH500mbar were the most significant in models of extreme temperatures, while UW500mbar played a significant role in estimating occurrences of precipitation. The variables HGT500mbar and RH500mbar were found to have a high presence through their weight in the precipitation and both temperature models.

Concerning spatial correlation selection, over 70% of these were located in a 25-pixel neighborhood around the central pixel from the station location for almost all complementary variables. Some selected pixels had low geographical and physiographical association with the basin. However, we decided not to omit them, because their inclusion improved the final adjustment. Both positive and significantly negative correlations appeared along the defined spatial window, and although there were no spatial criteria, we selected only those pixels with a greater absolute correlation, no matter where they were.

One potential limitation of the method proposed here that arises when the length of the time series to be filled or expanded is significantly large is that the assumption of homogeneity of variance may no longer hold as well as other characteristics of the distribution. In our example, we tested the hypothesis that the reconstructed minimum temperature followed the normal distribution using the Kolmogorov-Smirnov test. The hypothesis is rejected only in 10% of the 108 time series (9 stations times 12 months). On the other hand, only half of the time series fulfill the assumption of variance homogeneity when looking at the Bartlett homogeneity test. Because of the statistical properties of a parametrical method, extreme temperature values are normally underestimated (we detected a deviation of 3.6 and 3.4 °C for maximum and minimum temperatures, respectively). If the number of data gaps or length of the time series to be reconstructed is large, there is a risk of obtaining a heteroscedastic model that fails in the estimation of extreme temperature values.

5 Conclusion

We have presented a methodology for filling gaps and expanding data records of minimum and maximum temperatures and precipitation within a GLMM framework, using climatic information from the Reanalysis project. Results show that the estimations of temperature and precipitation are good and correctly represent the spatio-temporal pattern observed in the basin under study. The use of a reference station improves the estimations of precipitation amounts, although the effect is not significant if the model is fitted considering only the probabilities of rainfall occurrences extracted from the binomial GLMM.

This methodology can be used to fill gaps and expand data records of information gathered since 1950. Since it is not exclusive, new weather variables may be included for estimation. It is important to emphasize that in the future, land use associated with the weather station may be considered in order to include the effect of urban heat islands, water bodies, and other land uses that significantly influence local climatic variables.

Acknowledgments Special thanks to Alvaro Paredes for his help on the algorithm generation and Shaw Lacy for edits. The authors would like to thank two anonymous reviewers for their comments and suggestions on how to improve the robustness of the method and for pointing out its limitations.

Funding information The authors would like to acknowledge the support from FONDECYT Project Nos. 1120713 and 1170429. This work was partly carried out with the aid of a grant from the Inter-American Institute for Global Change Research (IAI, CRN3056) which is supported by the US National Science Foundation (Grant GEO-1128040).

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Bao X, Zhang F (2012) Evaluation of NCEP–CFSR, NCEP–NCAR, ERA-Interim, and ERA-40 *Reanalysis* datasets against independent sounding observations over the Tibetan Plateau. J Clim 26:206–214

Bastola S, Misra V (2014) Evaluation of dynamically downscaled *Reanalysis* precipitation data for hydrological application. Hydrol Process 28:1989–2002

Bates D, Maechler M, Bolker B (2012) lme4: linear mixed-effects models using S4 classes. R package version 0.999999–0

Begert M, Schlegel T, Kirchhofer W (2005) Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. Int J Climatol 25:65–80

Beguería S, Vicente-Serrano SM (2013) SPEI: calculation of the Standardized Precipitation-Evapotranspiration Index. R package version 1.3

Bengtsson L, Hagemann S, Hodges KI (2004) Can climate trends be calculated from *Reanalysis* data? J Geophys Res Atmos 109: 1984–2012

Beniston M, Stoffel M, Harding R, Kernan M, Ludwig R, Moors E, Samuels P, Tockner K (2012) Obstacles to data access for research related to climate and water: implications for science and EU policy-making. Environ Sci Policy 17:41–48

Betts AK, Ball JH, Barr AG, Black TA, McCaughey JH, Viterbo P (2006) Assessing land-surface-atmosphere coupling in the ERA-40 *Reanalysis* with boreal forest data. Agric For Meteorol 140:365–382

Blender R, Fraedrich K, Sienz F (2008) Extreme event return times in long-term memory processes near 1/f. Nonlinear Process Geophys 15(4):557–565

Bojanowski JS, Vrieling A, Skidmore AK (2014) A comparison of data sources for creating a long-term time series of daily gridded solar radiation for Europe. Sol Energy 99:152–171

Brands S, Gutiérrez JM, Herrera S, Cofiño AS (2012) On the use of *Reanalysis* data for downscaling. J Clim 25:2517–2526

Bustos E, Meza FJ (2014) A method to estimate maximum and minimum air temperature using MODIS surface temperature and vegetation data: application to the Maipo basin, Chile. Theor Appl Climatol 120:211–226. https://doi.org/10.1007/s00704-014-1167-2

Buyadi SN, Mohd WM, Misni A (2013) Impact of land use changes on the surface temperature distribution of area surrounding the National Botanic Garden, Shah Alam. Procedia Soc Behav Sci 101:516–525

Casanueva A, Herrera S, Fernandez J, Frias MD, Gutierrez JM (2012) Comparison of statistical and dynamical downscaling methods in representing temperature extremes. 12th Annual Meeting of the European Meteorological Society (EMS) and the 9th European Conference on Applied Climatology (ECAC), Poland, 10-14 September 2012

Castro LM, Miranda M, Fernández B (2013) Evaluation of TRMM multi-satellite precipitation analysis (TMPA) in a mountainous region of the Central Andes range with a Mediterranean climate. Hydrol Res 46:89–105. https://doi.org/10.2166/nh.2013.096

Chen SM, Hwang JR (2000) Temperature prediction using fuzzy time series. IEEE Trans Syst Man Cybern B Cybern 30:263–275

Colle BA (2004) Sensitivity of orographic precipitation to changing ambient conditions and terrain geometries: an idealized modeling perspective. J Atmos Sci 61(5):588–606

Daly C (2006) Guidelines for assessing the suitability of spatial climate data sets. Int J Climatol 26:707–721

Diez E, Primo C, Garcia-Moya JA, Gutiérrez JM, Orfila B (2005) Statistical and dynamical downscaling of precipitation over Spain from DEMETER seasonal forecasts. Tellus A 57:409–423

Doggers P, Allen RG (2002) Estimating reference evapotranspiration under inaccurate data conditions. Irrig Drain Syst 16:33–45

Drosdowsky W, Chambers LE (2001) Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. J Clim 14(7):1677–1687

Feng L, Nowak G, O'Neill TJ, Welch AH (2014) CUTOFF: a spatio-temporal imputation method. J Hydrol 519:3591–3605

Flannigan MD, Wotton BM (2001) Climate, weather, and area burned. In: Johnson E, Miyanishi K (eds) Forest fires, behavior and ecological effects. EE.UU. Academic Press, New York, pp 351–373

Fowler HJ, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int J Climatol 27(12):1547–1578

Fuka DR, Walter MT, MacAlister C, DeGaetano AT, Steenhuis TS, Easton ZM (2013) Using the climate forecast system *Reanalysis* as weather input data for watershed models. Hydrol Process 28:5613–5623. https://doi.org/10.1002/hyp.10073

Gershunov A, Cayan DR (2003) Heavy daily precipitation frequency over the contiguous United States: sources of climatic variability and seasonal predictability. J Clim 16:2752–2765

Harnik N, Chang EK (2003) Storm track variations as seen in radiosonde observations and reanalysis data. J Clim 16(3):480–495

Hong Y, Nix H, Hutchinson M, Booth T (2005) Spatial interpolation of monthly mean climate data for China. Int J Climatol 25:1369–1379

Hwang S, Graham WD, Adams A, Geurink J (2013) Assessment of the utility of dynamically-downscaled regional *Reanalysis* data to predict streamflow in west central Florida using an integrated hydrologic model. Reg Environ Chang 13:69–80

Jung-Woo K, Yakov A (2010) Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J Hydrol 394:305–314

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year *Reanalysis* project. Bull Am Meteorol Soc 77:437–471

Kemp MU, Kemp MMU (2012) Package 'RNCEP'

Kistler R, Collins W, Saha S, White G, Woollen J, Kalnay E, Chelliah M, Ebisuzaki W, Kanamitsu M, Kousky V, van den Dool H, Jenne R, Fiorino M (2001) The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. Bull Am Meteorol Soc 82(2): 247–267

Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen-Geiger climate classification updated. Meteorol Z 15:259–263

Kubik M, Brayshaw D, Coker P (2012) *Reanalysis*: an improved data set for simulating wind generation?. In: WREF 2012. Denver, CO. http://tinyurl.com/c4ge72x

Kuznetsova A, Brockhoff PB, Bojesen RH (2013) lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 1.2–0

Laurikkala J, Juhola M, Kentala E, Lavrac N, Miksch S, Kavsek B (2000) Informal identification of outliers in medical data. In Fifth international workshop on intelligent data analysis in medicine and pharmacology (pp. 20–24)

Linares-Rodriguez A, Ruiz-Arias JA, Pozo-Vasquez D, Tovar-Pescador J (2011) Generation of synthetic daily global solar radiation data based on ERA-Interim *Reanalysis* and artificial neural networks. Energy 36:5356–5365

Lookingbill TR, Urban DL (2003) Spatial estimation of air temperature differences for landscape-scale studies in montane environments. Agric For Meteorol 114:141–151

Maidment RI, Grimes DI, Allan RP, Greatrex H, Rojas O, Leo O (2012) Evaluation of satellite-based and model re-analysis rainfall estimates for Uganda. Meteorol Appl 20:308–317

McCune B (2007) Improved estimates of incident radiation and heat load using non-parametric regression against topographic variables. J Veg Sci 18:751–754

Misra V, DiNapoli SM, Bastola S (2012) Dynamic downscaling of the twentieth-century *Reanalysis* over the southeastern United States. Reg Environ Chang 13:15–23

Montecinos A, Aceituno P (2003) Seasonality of the ENSO-related rainfall variability in Central Chile and associated circulation anomalies. J Clim 16:281–296

Nagata K (2011) Quantitative precipitation estimation and quantitative precipitation forecasting by the Japan Meteorological Agency. RSMC Tokyo –Typhoon Center Technical Review 13:37–50

Paulo AA, Rosa RD, Pereira LS (2012) Climate trends and behaviour of drought indices based on precipitation and evapotranspiration in Portugal. Nat Hazards Earth Syst Sci 12:1481–1491

Peres-Neto PR, Jackson DA, Somers KM (2003) Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. Ecology 84:2347–2363

Perry M, Hollis D (2005) The development of a new set of long-term climate averages for the UK. Int J Climatol 25:1023–1039

Pierce D (2011) ncdf: Interface to Unidata netCDF data files. R package version 16.6

Pinheiro J, Bates D, DebRoy S, Sarkar D, the R Development Core Team. (2012). Nlme: linear and nonlinear mixed effects models. R package version 3.1–104

Pörtner H (2001) Climate change and temperature-dependent biogeography: oxygen limitation of thermal tolerance in animals. Naturwissenschaften 88:137–146

Ramella L, Haimberger L (2014) A global radiosonde and tracked balloon archive on 16 pressure levels (GRASP) back to 1905–part 2: homogeneity adjustments for PILOT and radiosonde wind data. Earth Syst Sci Data Discuss 7:335–383

Ramos-Calzado P, Gomez-Camacho J, Perez-Bernal F, Pita-Lopez MF (2008) A novel approach to precipitation series completion in climatological datasets: application to Andalusia. Int J Climatol 28(11):1525–1534

Refslund J, Dellwik E, Hahmann AN, Barlage MJ, Boegh E (2014) Development of satellite green vegetation fraction time series for use in mesoscale modeling: application to the European heat wave 2006. Theor Appl Climatol 117:377–392

Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E, Bosilovich MG, Schubert SD, Takacs L, Kim GK, Bloom S, Chen J, Collins D, Conaty A, da Silva A, Gu W, Joiner J, Koster RD, Lucchesi R, Molod A, Owens T, Pawson S, Pegion P, Redder CR, Reichle R, Robertson FR, Ruddick AG, Sienkiewicz M, Woollen J (2011) MERRA: NASA's modern-era retrospective analysis for research and applications. J Clim 24:3624–3648

Rojas E, Arce B, Peña A, Boshell F, Ayarza M (2010) Quantization and interpolation of local trends in temperature and precipitation in the high Andean areas of Cundinamarca and Boyaca (Colombia). Corpoica 11:173–182

Royer A, Poirier S (2010) Surface temperature spatial and temporal variations in North America from homogenized satellite SMMR-SSM/I microwave measurements and *Reanalysis* for 1979–2008. J Geophys Res Atmos (1984–2012) 115:1–16

Ruiz-Arias JA, Tovar-Pescador J, Pozo-Vázquez D, Alsamamra H (2009) A comparative study of DEM-based models to estimate solar radiation on mountainous terrains. Int J Geogr Inf Sci 23(8):1049–1076

Saha S, Moorthi S, Pan HL, Wu X, Wang J, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D, Liu H, Stokes D, Grumbine R, Gayno G, Wang J, Hou YT, Chuang HY, Juang HMH, Sela J, Iredell M, Treadon R, Kleist D, van Delst P, Keyser D, Derber J, Ek M, Meng J, Wei H, Yang R, Lord S, van den Dool H, Kumar A, Wang W, Long C, Chelliah M, Xue Y, Huang B, Schemm JK, Ebisuzaki W, Lin R, Xie P, Chen M, Zhou S, Higgins W, Zou CZ, Liu Q, Chen Y, Han Y, Cucurull L, Reynolds RW, Rutledge G, Goldberg M (2010) The NCEP climate forecast system *Reanalysis*. Bull Am Meteorol Soc 91:1015–1057

Sandholt I, Rasmussen K, Andersen J (2002) A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. Remote Sens Environ 79:213–224

Schmidli J, Frei C, Vidale PL (2006) Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods. Int J Climatol 26:679–689

Sherwood SC, Titchner HA, Thorneb PW, McCarthyb MP (2008) Short communication how do we tell which estimates of past climate change are correct? Int J Climatol 29:1520–1523. https://doi.org/10.1002/joc.1825

Simmons A, Uppala S, Dee D, Kobayashi S (2007) ERA-Interim: new ECMWF *Reanalysis* products from 1989 onwards. ECMWF Newsl 110:25–35

Sobrino JA, Jiménez-Muñoz JC, Paolini L (2004) Land surface temperature retrieval from LANDSAT TM 5. Remote Sens Environ 90(4): 434–440

Sterl A (2001) On the impact of gap-filling algorithms on variability patterns of reconstructed oceanic surface fields. Geophys Res Lett 28:2473–2476

Suga Y, Ogawa H, Ohno K, Yamada K (2003) Detection of surface temperature from LANDSAT-7/ETM+. Adv Space Res 32:2235–2240

Tardivo G, Berti A (2012) A dynamic method for gap filling in daily temperature datasets. J Appl Meteorol Climatol 51:1079–1086

Tardivo G, Berti A (2014) The selection of predictors in a regression-based method for gap filling in daily temperature datasets. Int J Climatol 34:1311–1317

Vicente-Serrano SM, Beguería S, López-Moreno JI (2010) A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J Clim 23:1696–1718

Vrac M, Naveau P (2007) Stochastic downscaling of precipitation: from dry events to heavy rainfalls. Water Resour Res 43(7). https://doi.org/10.1029/2006WR005308

Wan Z, Li ZL (1997) A physics-based algorithm for retrieving land-surface emissivity and temperature from EOS/MODIS data. IEEE Trans Geosci Remote Sens 35:980–996

Wang D, Murphy M (2004) Estimating optimal transformations for multiple regression using the ACE algorithm. J Data Sci 2(4):329–346

Weng Q, Lu D, Schubring J (2004) Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies. Remote Sens Environ 89(4):467–483

Wickham H (2007) Reshaping data with the reshape package. J Stat Softw 21:1–20

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press/Elsevier, New York 627 pp

Wright CK, de Beurs KM, Akhmadieva ZK, Groisman PY, Henebry GM (2009) Reanalysis data underestimate significant changes in growing season weather in Kazakhstan [Internet]. Environ Res Lett 2009: 045020 Available from: http://iopscience.iop.org/1748-9326/4/4/045020

Yoshimura K, Kanamitsu M (2008) Dynamical global downscaling of global *Reanalysis*. Mon Weather Rev 136:2983–2998

Zavala MA (2004) Estructura, dinámica y modelos de ensamblaje del bosque mediterráneo: entre la necesidad y la contingencia. Ecología del bosque mediterráneo en un mundo cambiante. Organismo Autónomo de Parques Nacionales. Ministerio de Medio Ambiente, Madrid, pp 249–280