ELSEVIER
CLOUDS · PRECIPITATION · AEROSOLS · RADIATION · WEATHER MODIFICATION
ATMOSPHERIC
RESEARCH

Infilling methods for monthly precipitation records with poor station network density in Subtropical Argentina

Santiago I. Hurtado, Pablo G. Zaninelli, Eduardo A. Agosta, Lorenzo Ricetti

PII: S0169-8095(21)00034-X

DOI: https://doi.org/10.1016/j.atmosres.2021.105482

Reference: ATMOS 105482

To appear in: *Atmospheric Research*

Received date: 25 September 2020

Revised date: 14 January 2021

Accepted date: 19 January 2021

Please cite this article as: S.I. Hurtado, P.G. Zaninelli, E.A. Agosta, et al., Infilling methods for monthly precipitation records with poor station network density in Subtropical Argentina, *Atmospheric Research* (2021), https://doi.org/10.1016/j.atmosres.2021.105482

Infilling methods for monthly precipitation records with poor station network density in Subtropical Argentina.

Santiago I. Hurtado[a,b,*], Pablo G. Zaninelli[c,a,d], Eduardo A. Agosta[a,b], Lorenzo Ricetti[a]

[a]*Facultad de Ciencias Astronómicas y Geofísicas, Universidad de La Plata, La Plata, Buenos Aires, Argentina*

[b]*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*

[c]*Centro de Investigaciones del Mar y la Atmósfera (CIMA) CONICET-UBA, Universidad de Buenos Aires, Buenos Aires, Argentina*

[d]*Instituto Franco-Argentino sobre Estudios de Clima y sus Impactos (UMI3351-IFAECI/CNRS-CONICET-UBA), Buenos Aires, Argentina*

[*]*Corresponding author.*

Abstract

Precipitation plays a crucial role from a social and economic perspective in Subtropical Argentina (STAr). Therefore, it renders the need for continuous and reliable precipitation records to develop serious climatological researches. However, precipitation records in this region are frequently inhomogeneous and scarce, which makes it necessary to deal with data filling methods. Choosing the best method to complete precipitation data series relies on rain gauge network density and on the complexity of orography, among other factors. Most comparative-method studies in the literature are focused on dense station networks while, contrastingly, the STAr's station network density is remarkably poor (between 10 and 1000 times lower). The research aims at assessing the performance of several interpolation methods in STAr. In this sense, the performance of a large number of interpolation methods was evaluated for dry and wet seasons, interpolating raw monthly data and their anomalies applied to different time-series subsets. In general, most methods performances improve when applied to anomalies in the seasonal time-series subset. Multiple Linear Regression (MLR) stands out as the method with the best performance for infilling

[*] Corresponding Author

Email addresses: santiagoh719@gmail.com (Santiago I. Hurtado), pzaninelli@cima.fcen.uba.ar (Pablo G. Zaninelli), eduardo.agosta@gmail.com (Eduardo A. Agosta), lorenzoricetti@gmail.com (Lorenzo Ricetti)

precipitation records for most of the regions regardless of orography or season. Despite the bibliography invokes that kriging interpolation methods are the best ones, in this work the performance of kriging methods was similar to the one of the Inverse Distance Weighted method (IDW) and the Angular Distance Weighted method (ADW, the method used to generate CRU precipitation dataset).

Keywords: Interpolation methods; missing data; monthly precipitation; time series; scarce data

1.1 1. Introduction

Subtropical Argentina (STAr) extends from central to north Argentina, east of the Andes, roughly to the north of 34°S. In STAr the economic production represents more than 80% of the national gross domestic product. Its climate encompasses six different climate regimes, ranging from a monsoon-influenced humid subtropical climate in the east to a desertic climate in the west (Beck et al., 2018). STAr is located inside the southern "La Plata" basin where most of the national hydroelectric energy production takes place and it is also one of the major food production regions all over the world (Magrin et al., 2005, Penalba and Vargas, 2008, Cuya et al., 2013). In the last years, it has been observed that the hydropower and electricity produced in LPB, which represent about 73% of the demand, have been strongly modified not only by the population growth but also by climate change (Popescu et al., 2014). In addition, a significant reduction of about 30% of the hydropower production is expected for Argentina according to projections for 2100 and for the worst scenario of climate change, which could entail a cost in investments of the order of 30 billion dollars (Turner et al., 2017). Moreover, crop production has been altered by climate change (Magrin et al., 2005) while large uncertainty exists in regard to its projected changes for both the near and far future (Rolla et al., 2018). This makes the development of adaptation strategies remain a challenge under study, in particular for large periods of drought (Wehbe et al., 2018). Thus, continuous and homogeneous precipitation data records are necessary to correctly characterize the changes that precipitation has suffered due to climate change and so planning policies that allow the efficient use of hydrological resources (Kalteh and Berndtsson, 2007, Sattari et al., 2017). However, precipitation data records in STAr show several missing values, and even gaps, which is a drawback that extends to the whole of South America's network of stations (Skansi et al., 2013).

Therefore, in order to achieve a complete dataset, missing values should be infilled, for example, via interpolation methods.

There exist several interpolation methods that we classified into three big groups: spatial methods (based on spatial variations), temporal methods (which use the co-variability between time series) and spatio-temporal methods (a combination of the above, see Kyriakidis and Journel, 1999, for a review on space-time methods). The suitability of any of the well-known and commonly used interpolation methods is not guaranteed for every region of the globe (de Amorim Borges et al., 2016), and the choice of the technique is key since poorly performing interpolation methods could introduce significant errors in hydrological model outputs (such as water balances, streamflow and runoff; Vieux, 2001, Bárdossy and Pegram, 2014), in environmental model simulations (such as crop yield estimations and drought severity: Kajornrit et al., 2012) and may lead to an inadequate climatological analysis (Price et al., 2000)

1.1.1*1.1.* *Motivation*

The literature is plenty of interpolation method studies and a few of them assess intercomparison of methods for monthly precipitation data. In our exploration, some peer-reviewed research works were selected for relevance in order to assess such a goal and are displayed in Table 1. For the sake of summarizing their main outcomes, location of study regions, density of stations per study region as well as best selected methods found by each study are shown in Figure 1. Methods and their acronyms are displayed in Table 2. It is apparent that Kriging's family of methods is mostly chosen among the best (i.e., OK, KED, CoK, RK, SKlm and KEM), followed by the methods: MLR, IDW and NR. It is also evident that the good performance of a method is dependent on the study region and network density. Though there is no consensus about the best performance interpolation method for a region, it has been already accepted that regardless of the method type, the performance of a method depends on the sample density, sample design, climatological characteristics and topography (Collins, 1995, De Silva et al., 2007, Li and Heap, 2008, Di Piazza et al., 2011, Burrough et al., 2015). Therefore, specific local studies are necessary to determine the most indicated interpolation method since generalization is not plausible (Aguilera et al., 2020).

Papers listed in Table 1 show a wide range of network density with differences of up to four orders of magnitude (see Fig. S1). Most of the studies were carried out on regions with more than 1 station per $1000km^2$. The study of Mair and Fares (2010) depicts the densest network for the Makhaka Valley (ca. 76 stations per $1000km^2$). In the opposite extreme, the study of Sattari et al. (2017) depicts the least dense network for southern Iran (ca. 0.08 stations per $1000km^2$). The network density of STAr exhibits a mean network density of about 0.03 stations per $1000km^2$ which is less than half of the network density used in Sattari et al. (2017). This represents a real challenge since the interpolation of precipitation in scarce measurement areas is not only more important but also more difficult (Wagner et al., 2012).

To our knowledge, a few studies assessing the performance of interpolation methods for precipitation belongs to the Southern Hemisphere, and less in particular, focused on precipitation networks in South America. One example is the study of Barrios et al. (2018) who examined a high network density in the region of central Chile (which is two orders of magnitude higher than the one in STAr). Another example is the study of de Amorim Borges et al. (2016) who assessed a quite dense network within the Distrito Federal of Brazil (which is one order of magnitude higher than the one in STAr). Note that despite some other researchers have used interpolation methods either to infill station data or to generate gridded precipitation datasets in South America (see, for example, Liebmann and Allured, 2005, Zotelo et al., 2008, González et al., 2012, Jones et al., 2013), none of them have conducted a comparative analysis of at least two different methods' performance.

Table 1: List of research works that study the performance of different interpolation methods. Corresponding study region and reference number (#) from Figure 1. Acronyms of the best methods found by every work (see the acronyms in Table 2). The current research is also included in the list.

#	Research work	Study region	Best methods
1	Bárdossy and Pegram (2014)	Southern Cape (South Africa)	CP and MLR
2	Barrios et al. (2018)	Biobio basin (Chile)	ANN, MLR and IDW_h
3	de Amorim Borges et al. (2016)	Distrito Federal (Brazil)	OK, IDW, RK and

			RIDW
4	Delbari et al. (2013)	northeast of Iran	KED, OK and CoK
5	De Silva et al. (2007)	Sri Lanka	IDW, NR and AA
6	Di Piazza et al. (2011)	Sicily (Italy)	RK
7	Hwang et al. (2012)	Animas and Alapaha basin (USA)	MLR
8	Kurtzman et al. (2009)	Yarkon-Taninim Basin (Israel)	IDW
9	Mair and Fares (2010)	Mākaha Valley (USA)	NR
10	Morales et al. (2019)	Tabasco state (Mexico)	GCIDW
11	Pellicone et al. (2018)	Calabria region (Italy)	KED
12	Presti et al. (2010)	Candelaro River Basin (Italy)	TSLR and SBE
13	Sattari et al. (2017)	Southern Iran	AA, MLR and NIPALS
14	Tang et al. (1996)	Klang River basin (Malaysia)	NR, MNR and IDW
15	Teegavarapu and Chandramouli (2005)	Kentucky (USA)	CWM, ANN and KEM
16	Teegavarapu et al. (2009)	Kentucky (USA)	FFSGAM
17	Terzi (2012)	Turkey	MLR
18	Wagner et al. (2012)	Mula and the Mutha Rivers (India)	RIDW
19	Westerberg et al. (2010)	Choluteca River basin (Honduras)	UK and CWM
20	Xia et al. (1999)	Bavaria (Germany)	UK and MLR
21	Xu et al. (2015)	Sichuan Province (China)	OK and CoK
22	Yavuz and Erdoğan (2012)	Turkey	OK
23	Young (1992)	Arizona and New Mexico (United States)	MDA
24	Yozgatligil et al. (2013)	Turkey	NR, MP and MCMC
25	Zhang and Srinivasan (2009)	Luohe Rive (China)	KED and Sklm
26	Current research work	Subtropical Argentina	

Figure 1: Global map with the 25 peer-reviewed research works assessed in the current research, as displayed in Table 1. Dot-marker denotes the study region corresponding to every research work. The marker size is proportional to the natural logarithm of the amount of stations per $1000km^2$ for

each region. Numbers around markers refer to numbered works displayed in Table 1. Gray inset roughly corresponds to the STAr region.

In addition, the requirements of continuity and homogeneity for precipitation time series are not achieved in most of the precipitation rain gauge records of STAr whose rain gauge stations are insufficient, sparse (see Fig. 1, Kidd et al., 2017) and, in general, inhomogeneous (Hurtado et al., 2020b). Regarding the latter, Hurtado et al. (2020b) found that from all the observed inhomogeneities in the precipitation time series of STAr, only two can be certainly identified as climatic jumps (corresponding to the 1950s and 1970s) while the remaining breakpoints could be erratic in essence. Even more, the authors discourage the use of one station located in the north of STAr to carry out climatological studies because they suspect that its records are not trustful in almost the whole period, but primarily in the early period.

In summary, to the best knowledge of the authors, so far there has been no research work assessing the performance of interpolation methods neither in STAr nor in any region that comprises it. Moreover, in the revised literature no research work presents a region with such poor network density as the one of STAr, which could be an important factor in the interpolation method performance (Wagner et al., 2012). Therefore, the present work aims at complementing Hurtado et al. (2020b)' results from a data-quality-control perspective through the data filling analysis of precipitation time series in STAr. This is performed by assessing the best method for interpolating missing values of monthly precipitation records in STAr. With this aim, the performances of 19 different methods and 32 different sub-methods are compared. The evaluated interpolation methods are a selection of the most commonly used for precipitation, plus a selection of methods that, to the authors' knowledge, has not been assessed yet for interpolating missing precipitation data, such as GAM or RMLR. Secondly, an open-source R package was designed with all the used methodologies in this work to make them available to anybody for future applications to any dataset. In Section 2, the used data are presented and the different methods and sub-methods are described; the errors of each method are evaluated in Section 3.1; the variability is analyzed in Section 3.2; the spatial distribution of errors for MLR is presented in Section 3.3; the method's errors in the extremes precipitation values are analyzed in Section 3.4; a discussion of the results is offered in Section 4; and, finally, a summary and conclusions are presented in Section 5.

Table 2: Interpolation method, its acronym, brief description and corresponding reference. Methods used in the current work are highlighted in bold.

Method	Acronym	Description	References
Nearest Neighbor	**NN**	The nearest neighbor value.	Sattari et al. (2017)
Single Best Estimator	**SBE**	The value of the neighbor station with the highest linear correlation.	Teegavarapu and Chandramouli (2005)
Arithmetic Average	**AA**	The spatial average of nearby stations.	Sattari et al. (2017)
Climatological Mean	**Clim_AA**	Monthly mean of all records for the same month.	
Arithmetic Median	**AM**	The spatial median of nearby stations.	
United Kingdom traditional method	**UK**	As SBE but multiplied by the ratio of the mean precipitation at the target station and the reference station.	Sattari et al. (2017)
Normal Ratio (original)	**NR_1952**	The mean of the spatial precipitation values weighted by the ratio of the mean precipitation at the target station and each reference station.	Paulhus and Kohler (1952)
Normal Ratio (modified)	**NR**	The mean of the spatial precipitation values weighted by the student t statistic of the correlation between the target station and each reference station.	Young (1992)
Correlation Weighting Method	**CWM**	The mean of the spatial precipitation values weighted by the correlation between the target station and each reference station.	Teegavarapu and Chandramouli (2005)
Inverse Distance Weighting Method	**IDW**	The mean of the spatial precipitation values weighted by the inverse of the distance between the target station and each reference station elevated to K.	Di Piazza et al. (2011)

Angular Distance Weighted	**ADW**	Similar to IDW but add a weight corresponding to isolation of station in a direction	New et al. (2000), Harris et al. (2020)
Revised Nearest Neighbor Weighting Method	**RNNWM**	As IDW but with a redefined measure of distance.	Teegavarapu and Chandramouli (2005)
Ordinary Kriging	**OK**	Weighted mean method based on the spatial dependence structure of the data. The weights are given by a theoretical variogram model fitted to the semivariogram of the data.	Erxleben et al. (2002), Vicente-Serrano et al. (2003)
Kriging with External Drift	**KED**	The KED assumes that the interest variable mean depends on auxiliary variables. It is useful when the variable itself is related to other spatially known variables, such as elevation.	Snepvangers et al. (2003), Vicente-Serrano et al. (2003)
Theil Sehn Linear Regression	**TSLR**	Linear regression that estimates the slope by the median of the slopes of all lines through all pairs of points.	Sen (1968), Theil (1992)
Multi-Linear Regression	**MLR**	Multi-linear regression based on ordinary least squares.	Young (1992), Simolo et al. (2010)
Robust Multi-Linear Regression	**RMLR**	Multi-linear regression based on the MM-estimator.	Hampel et al. (2011), Venables and Ripley (2013)
Generalized Additive Model	**GAM**	Statistical model that combines the essence of General Linear Models and Additive models.	Hastie and Tibshirani (1987)
Optimal Interpolation	**OI**	Statistical model based on the estimation of a first guest value to each station and then the computation of the weighted mean of the first guest error to correct the estimated value.	Eischeid et al. (2000)

Artificial Neural Network	ANN	Methods based on neural networks	Teegavarapu and Chandramouli (2005)
Co-Kriging	CoK	The CoK assumes that the interest variable mean depends on auxiliary variables, such as elevation.	Delbari et al. (2013)
Copula Based	CP	Method based on multivariate cumulative distribution functions.	Bárdossy and Pegram (2014)
Fixed function set genetic algorithm method	FFSGAM	Method based on predefined functional forms whose coefficients are then estimated by optimization procedures.	Teegavarapu et al. (2009)
Generalization of the modified correlation coefficient with the inverse distance weighting method	GCIDW	Generalization of IDW modified to add height	Morales et al. (2019)
IDW modified to add height	IDW_h	IDW modified to add height	Barrios et al. (2018)
Locally Weighted Polynomial regression	LWP	Spatial regression based on nearest neighbors	Hwang et al. (2012)
Monte Carlo Markov Chain	MCMC	Method based on multiple imputations making a Markov Chain.	Yozgatligil et al. (2013)
Multiple Discriminant Analysis	MDA	Method based on empirical orthogonal functions	Young (1992)
Modified Normal Ratio	MNR	NR modified with a weighted parameter	Singh (1988)
Multilayer Perceptron	MP	Specific Neural Network	Yozgatligil et al. (2013)
Nonlinear estimation by Iterative Partial Least Square	NIPALS	Iterative method based on Principal Component Analysis	Sattari et al. (2017)

Residual IDW	RIDW	IDW applied over the residuals of another method.	Wagner et al. (2012)
Residual Kriging	RK	OK applied over the residuals of another method.	Di Piazza et al. (2011)
Simple Kriging with Local Means	SKlm	Simple Kriging applied over the residuals of another method.	Zhang and Srinivasan (2009)
Kriging Estimation Method	KEM	Normally KEM is referred to OK or Simple Kriging.	

1.2 2. Data and Methods

1.2.1*2.1.* *Data and study region*

Monthly precipitation records from a total of sixty-two weather stations provided by the Argentine National Meteorological Service (SMN, after its Spanish abbreviation, https://www.smn.gob.ar/) were used (see Table 2). These weather stations follow the World Meteorological Organization (WMO) standards and they are part of the WMO Global Telecommunication System (GTS). As it is shown in Figure 2, the stations are distributed all over STAr. The records started between the end of the 19^{th} century and the end of the 20^{th} century. The study region is characterized by a low weather station density (62 weather stations in an area of about $1,541,898km^2$), which represents a density of one weather station per $24,869km^2$. Two stations (87127 and 87360, see Table 3) were not used to analyze its interpolated time series since they have less than 100 records.

Most of the stations present a marked wet season from October to April and a dry one from May to September (see Fig. S2, and Hurtado et al., 2020a,b). Besides, a zonal gradient of annual accumulated precipitation is observed, presenting drier conditions to the west and wetter conditions to the east (Barros and Silvestri, 2002). According to the global climate classification from Beck et al. (2018), the region encompasses monsoon-influenced humid subtropical climate (CWa), humid subtropical climate (CFa), cold semi-arid climate (BSk), hot semi-arid climate (BSh), cold desert climate (BWk), and hot desert climate (BWh).

Table 3: Subtropical Argentina Rain Gauge Stations

Stations	OMM number	Lat	Lon	Z (m)
La Quiaca Obs.	87007	-22°6’	-65°36’	3459
Orán Aero	87016	-23°9’	-64°19’	357
Tartagal Aero	87022	-22°39’	-63°49’	450
Jujuy UN	87043	-24°10’	-65°11’	1302
Jujuy Aero	87046	-24°23’	-65°5’	905
Salta Aero	87047	-24°51’	-65°29’	1221
Metán	87050	-25°29’	-64°48’	855
Rivadavia	87065	-24°10’	-62°54’	205
Las Lomitas	87078	-24°42’	-60°35’	130
Iguazú Aero	87097	-25°44’	-54°28’	270
Tucumán Aero	87121	-26°51’	-65°6’	450
Termas de Río Hondo	87127	-27°29’	-64°56’	280
Santiago del Estero Aero	87129	-27°46’	-64°18’	199
Presidencia Roque Sáenz Peña	87148	-26°45’	-60°24’	93
Resistencia Aero	87155	-27°27’	-59°3’	52
Formosa Aero	87162	-26°12’	-58°14’	60
Bernardo de Irigoyen	87163	-26°15’	-53°39’	815
Corrientes Aero	87166	-27°27’	-58°46’	62
Ituzaingó	87173	-27°35’	-56°40’	72
Posadas Aero	87178	-27°22’	-55°58’	125
Oberá Aero	87187	-27°29’	-55°8’	303
Tinogasta	87211	-28°4’	-67°34’	1201
Chilecito Aero	87213	-29°14’	-67°26’	947
La Rioja Aero	87217	-29°23’	-66°49’	429
Catamarca Aero	87222	-28°36’	-65°46’	454
Villa María del Río Seco	87244	-29°54’	-63°41’	341
Ceres Aero	87257	-29°53’	-61°57’	88
Reconquista Aero	87270	-29°11’	-59°42’	53
Mercedes	87281	-29°13’	-58°6’	107
Paso de los Libres Aero	87289	-29°41’	-57°9’	70

Jáchal	87305	-30°14’	-68°45’	1175
San Juan Aero	87311	-31°34’	-68°25’	598
Chamical Aero	87320	-30°22’	-66°17’	461
Chepes	87322	-31°20’	-66°36’	658
Villa Dolores Aero	87328	-31°57’	-65°8’	566
Córdoba Aero	87344	-31°18’	-64°12’	495
Córdoba Observatorio	87345	-31°24’	-64°11’	425
Escuela Aviación Militar	87347	-31°27’	-64°16’	502
Pilar Observatorio	87349	-31°40’	-63°53’	338
Sunchales	87356	-30°58’	-61°20’	92
Rafaela	87360	-31°16’	-61°30’	99
Sauce Viejo / Santa Fe Aero	87371	-31°42’	-60°49’	18
Paraná Aero	87374	-31°47’	-60°29’	78
Monte Caseros Aero	87393	-30°16’	-57°39’	54
Concordia Aero	87395	-31°18’	-58°1’	38
Uspallata	87405	-32°36’	-69°20’	1891
San Carlos	87412	-33°46’	-69°2’	940
San Martín	87416	-33°5’	-68°25’	653
Mendoza Aero	87418	-32°50’	-68°47’	704
Mendoza Observatorio	87420	-32°53’	-68°51’	827
San Luis Aero	87436	-33°16’	-66°21’	713
Santa Rosa de Conlara Aero	87444	-32°23’	-65°11’	620
Villa Reynolds Aero	87448	-33°44’	-65°23’	486
Río Cuarto Aero	87453	-33°7’	-64°14’	421
Marcos Juárez Aero	87467	-32°42’	-62°9’	114
Venado Tuerto	87468	-33°40’	-61°58’	112
El Trébol	87470	-32°12’	-61°40’	96
Rosario Aero	87480	-32°55’	-60°47’	25
Gualeguaychú Aero	87497	-33°0’	-58°37’	23
Malargüe Aero	87506	-35°30’	-69°35’	1425
San Rafael Aero	87509	-34°35’	-68°24’	748

Laboulaye Aero	87534	-34°8’	-63°22’	137

Figure 2: Location of the 62 weather stations (open dots) in Subtropical Argentina and its elevation (Z). Red squares mark three regions that are evaluated in Section 4.

1.2.2*2.2.* *Preprocessing data treatment*

Before analyzing the interpolation method performances, a thorough quality control analysis was performed. First, values greater than the median plus three times the interquartile range were regarded as atypically extreme. These extremes, as well as zero values, were verified by contrasting records with their neighbors’ stations. For data after 1979, this verification included also the inspection of monthly outgoing longwave radiation anomalies to assess its reliability (not shown). From this analysis, few records were considered as errors and so were set as missing values. It is also relevant to make a breakpoint analysis to guarantee the homogeneity of the precipitation time-series, which is an essential requirement for the interpolation procedures (Štěpánek et al., 2009). Otherwise, breakpoints in time-series will alter the co-variability of the data and, consequently, the performance of the interpolation methods would be impaired. The breakpoint analysis for the current precipitation dataset in STAr was performed by Hurtado et al. (2020b). As was mentioned in the introduction, the authors have shown two main climate jumps (natural breakpoints) in 1956 and other in 1976. Given these climatic jumps and the fact that the data records are scarce in the earlier periods of the records in STAr, the current research was carried out by using the data only from 1979 to 2017.

1.2.3*2.3.* *Interpolation Methods*

Table 2 shows the interpolation methods considered in this work with a brief explanation of each one, corresponding acronyms used hereinafter and the reference of some works that applied them. The Kriging methods (OK and KED) need to fit a theoretical variogram to the empirical variogram in order to perform the interpolation. So, four theoretical variogram models were used: Gaussian (Gau), Exponential (Exp), Spherical (Sph) and Matern (Mat). Both OK and KED methods were used in five different forms, selecting the best fit to the empirical variogram using maximum likelihood (referred as OK and KED) and using one of the fixed models mentioned above (referred as OK_*model* and KED_*model*, in combination with Exp, Gau, Mat, and Sph). For KED, the secondary variable used was elevation since Ly et al. (2013) found that KED applied with elevation outperforms the other methods. Kriging methods were not used to interpolate precipitation values in "La Quiaca Obs." station (see Table 3) since there are few stations in the surroundings and the computation systematically failed.

Considering that the OI method requires a first guest estimator, two different estimators NN (OI_NN) and SBE (OI_SBE) were used. The regression methods (MLR, TSLR, and RMLR) were computed through the raw data as well as with the ranked-data, which produces a ranked regression (Presti et al., 2010). These variations in the regression methods are denoted as MLR_rk, TSLR_rk, and RMLR_rk.

1.2.4*2.4.* *Statistical Analysis*

Statistical and mathematical analysis and the different graphics presented in this work were all made in R. Among the R's packages used we can mention *ggplot2* (Wickham, 2016), *ggsn* and *knitr* ((Xie, 2014, 2015) for visualization, and *gam* (Hastie, 2011), *MASS* (Venables and Ripley, 2013), *sp* (Pebesma and Bivand, 2005, Bivand et al., 2013), *RobustLinearReg* (Hurtado, 2020) and *gstat* (Pebesma, 2004, Pebesma and Heuvelink, 2016) for calculus. All the computed methods were documented in an open-source R package made for this research work, available in the following link: https://github.com/santiagoh719/MissingData.

Every observed value of each station was interpolated with every method in order to calculate the error of each one, implementing a Leave-One-Out Cross-Validation (LOOCV) method (Sammut and Webb, 2010). Since precipitation can not be negative, all the negative interpolated values were set to zero. To assess the performance of each method the Standardized Root Mean Square Error (*SRMSE*) and the Standardized Mean Error (*SME*, Haberlandt, 2007) were used as measure of error:

$$SRMSE_{i,j} = \frac{\sqrt{\dfrac{\sum_{t=1}^{n_t}[Int_{i,j}(t) - Obs_i(t)]^2}{n_t}}}{\dfrac{\sum_{t=1}^{n_t} Obs_i(t)}{n_t}} \qquad (1)$$

$$SME_{i,j} = \frac{\frac{\sum_{t=1}^{n_t} Int_{i,j}(t) - Obs_i(t)}{n_t}}{\frac{\sum_{t=1}^{n_t} Obs_i(t)}{n_t}} \qquad (2)$$

Being *Obs* the observed value and *Int* the estimated value by the interpolation method with i corresponding to the *i-th* station, j the *j-th* interpolation method and n_t is the length (time steps) of *Obs* and *Int*.

The election of the SRMSE lies in the fact that not onlyis it a typical measure of error, but also it provides an estimation of the average error. The SME was selected in order to assess the *bias* of the methodologies, since it gives a notion of the systematic over- (under-) estimation error.

In order to apply an interpolation method, a subset of data (predictors) must be selected. The subset is usually taken from stations near the targeted station. To objectively select the best subset of predictors for every targeted station and method, 5 different subsets of predictor stations were used, consisting of all the stations at a distance lower than 100km, 200km, 300km, 400km and 500km, respectively. Then, for every method and targeted station, the subset with the lowest SRMSE was selected. In addition, IDW, ADW and RNNWM methods depend on a free parameter *k*. To select the best k for each station and method, they were computed with k varying from 0.1 to 10 with a step of 0.1, and then the subset and k value with the lowest SRMSE was selected for that method and station. Moreover, ADW was also applied with the parameters used in the globally monthly precipitation dataset CRU (Harris et al., 2020), which is $k = 4$, and a search radius of $450km$, this is noted in the manuscript as ADW_CRU.

For clarity, the interpolation methods were separated and ordered into groups: regression methods (MLR, RMLR, TSLR, MLR_rk, RMLR_rk and TSLR_rk), GAM, single estimation (NN, SBE and UK), average estimation (AA, AM, Clim_AA), inverse distance (ADW, ADW_CRU and IDW), weighted mean (RNNWM, CWM, NR and NR_1952), Optimal Interpolation (OI_NN and OI_SBE), kriging with external drift (KED, KED_Gau, KED_Exp, KED_Sph and KED_Mat) and ordinary kriging (OK, OK_Gau, OK_Exp, OK_Sph and OK_Mat).

Furthermore, the interpolation methods were applied to four different time series for every season (dry and wet): the full-year observed data (absolute value) time series (denoted in the text as "Full-yr. Series"), the seasonal subset of the observed data (absolute value) time series (denoted in the text as "Season Subset") and their corresponding time series of anomalies (absolute values minus monthly means of the 30 yr. period 1980-2010). Thus, on the one hand, we obtain two sets of interpolation coefficients: one for the observed full-year time series and another for their anomalies, which are the same sets used to evaluate in both wet and dry seasons. On the other hand, we obtain two sets of interpolation coefficients: one for the wet subset time series and another for the dry subset times series; and two more sets for their corresponding time series of anomalies.

To further explore the interpolation methods performances, Pearson's first-moment correlation coefficient between the interpolated data and the observations was calculated. Also, the Kolmogorov-Smirnov test (Conover and Conover, 1980) was used to assess the difference between the empirical probability distribution of the observed and the interpolated data. Its null hypothesis states that both the theoretical (interpolated) and the real (observed) data follow the same probability distribution.

Finally, to investigate the regional performance of Kriging methods, the Moran's I (Moran, 1950), a measure of the spatial autocorrelation, was computed to test if the data are either spatially autocorrelated or randomly distributed.

1.3 3. Results

1.3.13.1. Methods error

In Figure 3 it is shown the method's SRMSE boxplot for every applied time-series and season. It can be noticed that in general, the SRMSE is greater for the dry season, where the error can be 10 times greater than the mean value. The methods MLR, MLR_rk, RMLR and RMLR_rk (NN, SBE, OI_NN, OI_SBE, RNNWM and Clim_AA) present the best (worst) performance in all cases, being the SRMSE values in general lower (greater) than 0.5 for the wet season and lower (greater) than 0.7 (1) in median for the dry season. In addition, GAM presents a notable improvement when the anomalies' series are used while methods like NR_1952 and UK worsen. In general, the use of anomalies for interpolation gets smaller SRMSE high values, especially in spatial methods such as IDW, ADW and Kriging methods, but does not present much difference in regression methods (see for example MLR or RMLR). Further, regardless of the time series used as input, there are no many differences between AA, AM, IDW, ADW, ADW_CRU and Kriging methods. The only relevant difference is that Kriging methods, in comparison, present greater (lower) high values in the wet (dry) season. Note that, for both wet and dry season MLR followed by GAM present the lowest SRMSE values, when applied with the season subset over anomalies, presenting SRMSE values mostly between $0.3-0.5$ ($0.3-0.8$) for the wet (dry) season.

Figure 3: Boxplot of the spatial variation of the Standardized Root Mean Square Error (SRMSE), discriminating each method described in Table 2, wet and dry seasons, the used time-series (the full-yr. time-series or the season subset time series) and if computation was made using observed data (raw) or their anomaly. The scale is logarithmic for better visualization. The filled colors of boxes represent each method group. Vertical black dashed lines separate the method groups.

Figure 4, alike Figure 3 but with SME instead of SRMSE. In this case, for SME, the usage of anomalies represents an outstanding improvement, for almost all methods, reducing the error variability around zero with respect to the observed series. The methods that stand out for their good performance, in most cases, are GAM, MLR, NR_1952, UK and Clim_AA. The almost zero value of SME for Clim_AA is expected, given the methods' characteristics. However, all methods present difficulties in reproducing monthly precipitation for the dry season, in concordance with the results for SRMSE (see Fig. 3). In this sense, almost all methods tend to overestimate the precipitation values showing bias up to 0.1 SME in median, with the exception of the robust regression ones (MLR_rk, RMLR, RMLR_rk, TSLR, TSLR_rk) and AM that tend to underestimate the real values, reaching up values lower than 0.5 SME of median.

Figure 4: As Figure 3, but showing the Standardized Mean Error (SME) and using an arcosinushiperbolic scale. The scale is for better visualization. It is used in place of the logarithmic scale since the latter is not defined for negative values and it shows problems with values close to zero.

In summary, from Figures 3 and 4, it can be seen that in general the methods perform better when they are applied with series of anomalies to the corresponding season time series. The regression-based methods (MLR, MLR_rk, RMLR, RMLR_rk, TSLR and TSLR_rk) show little differences between the usage of the anomaly or the observed data. MLR outperforms the other methods for the wet season, except for GAM that shows a better performance in terms of SME. For the dry season, MLR and, in general, GAM present the best performance when the seasonal time series are considered as input. In addition, in almost all the stations, MLR presents the lowest SRMSE being the method that is selected most times as the best while GAM is the second-best for the anomalies’ case (see Fig. S3). Since the usage of the full-yr. time-series for the dry season presents greater errors regardless of the method used (see Fig. S4), henceforth the analysis of the methods’ performance using the full-yr. time-series will not be considered for the dry season.

1.3.2*3.2.* *Variability*

For climatology studies, it is imperative that the infilled values reproduce the time series’ variability. To assess this, the correlation between the interpolated time series and the real-time series was computed (see Fig. 5). In general, the most of the methods present correlations greater than 0.6 in median. MLR, and its ranked version, show a good performance reproducing the precipitation variability independently of the considered time series (anomalies or observed) and on the season. Moreover, MLR presents correlation values over 0.8 for more than 50% of the stations for all cases, which represent more than 64% of explained variability. Besides, GAM presents very high correlation values (around 0.8) for the anomaly time series. On the other hand, lower correlations are observed especially for Clim_AA, that present correlations mostly under 0.4, and also for the methods NR_1952 and UK when applied to the anomalies time series, showing in general correlations under 0.4. RNNWM presents a poor performance in all cases, reaching negative correlations for several studied stations. Finally, little or no difference is shown among the performance of the other methods. In general, the median of the correlation is around 0.7 for the remaining methods, which implies around 50% of explained variance.

Figure 5: Boxplot of the correlation between the interpolated time series and the observed ones for each method, each box contains one value per weather station showing spatial variations. Results for observed input are shown in the left column and for anomalies in the right column. Rows in the upper (lower) panels correspond to wet (dry) season using the season subset time-series and the full-yr. time-series subset as input; for dry season only for the season subset was used as input. The filled colors of the boxes represent each method group. Vertical black dashed lines separate the method groups.

To complete the previous analysis, Figure 6 shows the percentage of stations in which the Kolmogorov-Smirnov's test null hypothesis was rejected for each method at different significance levels. The better (worse) the method, the less (more) times the null hypothesis is rejected. As expected, the higher the level of significance, the lower the percentage of stations where the null hypothesis is rejected, being more evident for the anomaly precipitation series. It can be seen that most of the methods have difficulties reproducing the distribution of the precipitation time series for the dry season, particularly for the observed one. Contrastingly, NR_1952 and UK show a higher level of rejection for the anomalies' time series. In general, MLR, GAM and RMLR present the best performances.

Figure 6: Heatmap of the percentage of stations in which the null hypothesis of the Kolmogorov-Smirnov test is rejected at a significance level of 0.01% (first row), 0.05% (second row) and 0.1% (third row), type of serie and season. The values over 40% are filled with black. Vertical white dashed lines separate the method groups.

1.3.3*3.3. Spatial distribution of MLR's SRMSE*

In this section, the spatial distribution of MLR's SRMSE is analyzed since it is the method that presents the best performance. Figure 7 (left panels) shows the MLR's SRMSE presenting a marked difference between the west and east region. Lower errors are found in the east region (in general bellow 0.4) and higher toward the west near the Andes (reaching values greater than 0.8). Larger errors are observed for the dry season rather than for the wet season. SRMSE presents high values (that can reach values up to ca. 3) in the north-western region for the dry season, while for the wet season the greatest errors are located in the central-western and south-western of STAr. The lowest errors (lower than 0.3) are around $32^o S$ and between $65^o - 60^o W$. When selecting the best implementation option ("Anomaly" and "Season Subset"), in wet season, the SRMSE is always lower than 1.1, and in dry season lower than 1.5 with the exception of "La Quiaca Obs." station. Moreover, for the wet (dry) season the SRMSE is greater than 0.8 (1.1) only in four stations and it is mostly under 0.4 (0.6) being the median 0.37 (0.48).

"La Quiaca Obs" presents large SRMSE values in the dry season for all methods, being the lowest of 2.93 for MLR, applied with "Season Subset" over anomalies. SRMSE values in "La Quiaca Obs" in dry season are greater than 10 for half of the methods and greater than 5 for three-quarters of them. These large SRMSE values are due to the actual lack of rainfall in the dry season for this particular station ($1.15mm$ of the monthly mean for dry season), making errors of $3mm$ be really huge relative errors.

Moreover, in the right panels, it is shown the average value of SRMSE for each station regarding all methods divided by the MLR's SRMSE, in order to detect the regions where MLR represents an improvement in comparison with the other methods. It can be observed that MLR represents a considerable improvement in the whole region, especially in central-western STAr from the Andes to the Córdoba mountain range, outstanding in the dry season, where the average SRMSE of the other methods can exceed for more than 2 times the MLR's error. It is worth saying that the spatial distribution of GAM's SRMSE is as the one of MLR but with greater values, especially in central and south-central STAr (see Fig. S5).

Figure 7: Map of the MLR's SRMSE (left panels) and the ratio of the average SRMSE of all methods for each station and the MLR SRMSE (right panels) for every form of application and season. Discrete color scales displayed the ratio (upper scale) and the MLR's SRMSE (lower scale). 'La Quiaca Obs.' station is denoted by an asterisk marker. Stations with values of the ratio lower than 1.25 are not shown.

1.3.4*3.4.* *Errors in observed extremes*

Every method's errors were analyzed for high and low extremes since all interpolation methods tend to underestimate (overestimate) high (low) extremes (Teegavarapu, 2014). In order to summarize, each method was evaluated using the best configuration (observed or anomalous time series, full yr. or season subset time series) according to the results found in previous sections. Figure 8 shows the global (considering all interpolated values) root mean squared error (RMSE; panel a) and also the global mean error (ME; panel b) for different subsets of data according to the percentile of observed values. It is notable that for both wet and dry seasons, MLR (GAM) has the lowest (second lowest) global RMSE considering all the data and also considering only observed values over the percentile 75^{th} and 90^{th}. For low precipitation values (observed precipitation lower than the percentile 25^{th} and 10^{th}), MLR_rk stands out with the lowest RMSE values followed by RMLR_rk for both dry and wet seasons. In general, MLR and GAM present the best results (in terms of RMSE), with values among the five lowest in all cases, followed by MLR_rk, RMLR and RMLR_rk. The ME for all methods are positive for low observed precipitation values (overestimate low extremes) and is negative for high observed values (underestimate high extremes), in concordance with Teegavarapu (2014). For the low precipitation values, the ME values are in general similar and with little differences among methods, but for both dry and wet season UK, Clim_AA, RNNWM and NR_1952 stand out with large ME values. Also, for the low precipitation values the ME values MLR, RMLR, MLR_rk and RMLR_rk show the lowest ME values for both seasons. For high precipitation values, MLR, OI_NN, OI_SBE, NN and SBE present the best ME values followed by GAM (which is the 6^{th} best value, not marked) for both seasons. Despite the good performance in terms of ME of these methods, MLR and GAM are the only ones that present a good performance in terms of global RMSE (see Fig. 8 panel a) and also in terms of SRMSE (see Fig. 3).

Figure 8: Global root mean squared error (RMSE, panel **a**) and global mean error (ME, panel **b**) of each method for dry (left panels) and wet (right panels) season. The RMSE and ME were calculated for the best configuration (observed or anomalous time series, full yr. or season subset time series) for each method. The markers denote the five best values (those that are closest to cero). Each color line refers to a different subset of data used to compute the RMSE, being *<P10* (*<P25*) the subset of observed values lower than the 10^{th} (25^{th}) percentile, *>P75* (*>P90*) the subset of observed values greater than the 75^{th} (90^{th}) percentile, and **All** no subset.

Figure 9 shows the smooth curves of absolute error corresponding to the observed values. The smoothing method used was general additive models, implemented in the package *mgcv* (Wood, 2011). The GAM, MLR, MLR_rk, RMLR and RMLR_rk methods are in color to highlight that they have presented the best global RMSE values (see Fig. 8). In general, MLR tends to have the lowest absolute errors, followed by GAM and RMLR. In fact, for the wet (dry) season, it can be observed that MLR's absolute errors are the lowest for observed values greater than 155 (30) mm. As expected, MLR_rk and RMLR_rk present the lowest errors for the low observed values (see the zoom panels). Furthermore, the inter-method spread is greater (lower) for the wet season than for the dry season for low (high) observed values. This implies that there are less differences between the methods' error for low (high) precipitation values in the dry (wet) season.

Figure 9: Smooth curves for each method of absolute errors regarding the true observed value. The smooths were calculated for the best configuration (observed or anomalous time series, full yr. or season subset time series) for each method. In colors, the GAM, MLR, MLR_rk, RMLR and RMLR_rk methods are highlighted. Lower panels show a zoom of the lowest precipitation values.

1.4 4. Discussion

Previous studies found that Kriging methods, in particular methods with elevation as a drift variable such as KED or CoK, are among the best interpolation methods (Teegavarapu and Chandramouli, 2005, Li and Heap, 2011, Di Piazza et al., 2011, Ly et al., 2013, ; see Table 1). Other works emphasize the inclusion of elevation as a secondary variable in low network density areas (Di Piazza et al., 2011, de Amorim Borges et al., 2016). However, according to our results for STAr, Kriging methods do not show an outstanding performance in comparison with the other spatial methods and have similar performance to IDW, ADW and even AA. All these spatial methods present poor performances with mean SRMSE values around 0.45 (0.8) for the wet (dry) season.

Furthermore, there was no much difference between OK and KED (with height as a secondary variable). And even, OK outperformed KED in regions of complex orography (see Fig. 10). These results are supported by the findings of Kajornrit et al. (2011), who showed that OK performs, in general, better than CoK with elevation as a secondary variable. In addition, the different statistics analyzed in the present work show that the best-fit variogram model used for Kriging methods does not guarantee the best performance, being the exponential variogram model the best one (see Fig. 10 and Fig. 3, 4, and 5).

Figure 10: Map of the SRMSE difference between KED_Exp and OK_Exp (left panels), KED and OK (mitle panels), and OK and OK_Exp (right panels) for each season applied over anomalies because it shows lower errors than applied over the observed data (see Fig. 3, 4).

A factor that could threaten the good performance of Kriging methods in STAr (and also of spatial methods) could be the low spatial autocorrelation of precipitation in the region. In Figure 11 it is shown the Moran's I (spatial autocorrelation) calculated for every month and for three different regions within STAr: region 1 (R1) at the south, region 2 (R2) at the northeast and region 3 (R3) at the northwest of STAr (all these regions are marked in Fig. 2). These domains represent different regions in terms of orography; R1 in the Córdoba mountain range, R2 in the subtropical plains, and R3 in the Andes. According to Figure 11, precipitation time series are not spatially autocorrelated for any month at confidence levels of 99%, 95% or 90% for regions R2 and R3, while for R1 only a few months present significant spatial autocorrelations but most do not. Thus, such poor spatial autocorrelations of observed precipitation could be due to the poor station network density, and perhaps to the climatological characteristics typical of the STAr's precipitation. In this sense, better performances of the Kriging methods, as good as those shown by previous works, could be expected if the station network were denser than it currently is.

Figure 11: Boxplot of the p-value of the Moran's I spatial autocorrelation of monthly precipitation for all months in the period 1979-2017, in the regions R1, R2 and R3 shown in Figure 2. The three dashed lines are the boundaries for confidence levels of 90%, 95% and 99%.

1.5 5. Summary and Conclusions

This work completed the data-quality-control analysis of precipitation time series in STAr started by Hurtado et al. (2020b). An exhaustive comparative analysis has been done in STAr, a region with few stations and scarce data. The performances of 19 different methods and 32 different sub-methods to fill missing values of monthly precipitation records have been assessed.

Our results show that the evaluated interpolation methods present better performance for wet season and in wet regions than for dry season and in dry regions (see Fig. 3, Fig. 7, Fig. S4, Fig. S5). In general, most methods perform better when applied to precipitation anomalies using the corresponding season's time series subset, in concordance with the findings of New et al. (2000).

Overall MLR presents the best performance in terms of SRMSE, SME, correlation, and KS test. In addition, GAM shows a good performance mostly when precipitation anomalies are considered and has proven to be good at reproducing the monthly precipitation distribution (Fig. 6). The good performance of GAM is remarkable since, to the best of our knowledge, until the date there has not been any research work that has analyzed GAM's capacity at interpolating missing precipitation records. Analysis of GAM performance in other regions is encouraged since not only presents an outstanding performance but also its implementation is effortless. MLR_rk, RMLR and RMLR_rk also present good performances but their implementation requires a greater computational effort. Even though MLR is the method with the lowest errors, it presents SRMSE values around 0.37 (0.48) that represent the 37% (48%) of the observed monthly mean precipitation for the wet (dry) season. Considering these large errors (especially for the dry season), in future works the usage of satellite and reanalysis data for interpolating missing values in rain gauges will be examined.

The methods that were shown to be the most efficient in reproducing monthly precipitation, i. e. MLR and GAM, can only be used to interpolate station data but they are unable to generate a gridded dataset. From the analyzed methods that allow the generation of regional gridded datasets, there is no clear overall preference among IDW, ADW and Kriging methods. Thus, choosing among these methods will depend on the study season, since IDW and ADW have a lower maximum SRMSE value than Kriging method for wet season, and the opposite occurs for dry season (see Fig. 3). On the other hand, according to Ruelland et al. (2008) it may be preferable to use IDW rather than OK for hydrological modeling since they found that these methods are the most efficient for imputing monthly precipitation records but that IDW yields the most realistic model results. Nevertheless, it is notable that the SRMSE values for these methods are, in general, rather large being the median around 0.45 (0.8) and the 75^{th} percentile around 0.6 (1.1) for the wet (dry) season.

On balance, the current study has done a thorough analysis of interpolation methods for infilling monthly precipitation records using a weather station network located in STAr, which shows a low and heterogeneous spatial distribution. The aforementioned methods performances could differ if the study is replicated using different region data and/or through the use of a denser station network. Hurtado et al. (2020b) and the present work were carried out to generate continuous, homogeneous and reliable precipitation data records in the subtropics of Argentina.

Future works will examine the spatial and temporal variability of precipitation patterns, their potential climate forcing mechanisms, which would complement previous works such as Hurtado and Agosta (2020).

1.6 Acknowledgments

Our tribute to Professor Luis Eduardo Castro, Ph.D., who passed away during the time of writing the manuscript. Our gratitude for the fruitful discussion we shared. The current research was funded by the PPID-G006 from the National University of La Plata title: “Variaciones multiescala de la precipitación y sus extremos sobre Argentina y su relación con el almacenamiento de agua continental”. We wish to express gratitude to the Carmelite NGO for financial support.

References

Aguilera, H., Guardiola-Albert, C., Serrano-Hidalgo, C., 2020. Estimating extremely large amounts of missing precipitation data. Journal of Hydroinformatics 22, 578–592. URL: https://iwaponline.com/jh/article/22/3/578/72493/Estimating-extremely-large-amounts-of-missing , doi:10.2166/hydro.2020.127.

de Amorim Borges, P., Franke, J., da Anunciação, Y.M.T., Weiss, H., Bernhofer, C., 2016. Comparison of spatial interpolation methods for the estimation of precipitation distribution in distrito federal, brazil. Theoretical and applied climatology 123, 335–348.

Bárdossy, A., Pegram, G., 2014. Infilling missing precipitation records–a comparison of a new copula-based method with other techniques. Journal of hydrology 519, 1162–1170.

Barrios, A., Trincado, G., Garreaud, R., 2018. Alternative approaches for estimating missing climate data: application to monthly precipitation records in south-central chile. Forest Ecosystems 5, 28.

Barros, V.R., Silvestri, G.E., 2002. The relation between sea surface temperature at the subtropical south-central pacific and precipitation in southeastern south america. Journal of climate 15, 251–267.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future köppen-geiger climate classification maps at 1-km resolution. Scientific data 5, 180214.

Bivand, R.S., Pebesma, E., Gómez-Rubio, V., 2013. Hello world: introducing spatial data, in: Applied spatial data analysis with R. Springer, pp. 1–16.

Burrough, P.A., McDonnell, R., McDonnell, R.A., Lloyd, C.D., 2015. Principles of geographical information systems. Oxford university press.

Collins, F.C., 1995. A comparison of spatial interpolation techniques in temperature estimation. Ph.D. thesis. Virginia Tech.

Conover, W.J., Conover, W.J., 1980. Practical nonparametric statistics .

Cuya, D.G.P., Brandimarte, L., Popescu, I., Alterach, J., Peviani, M., 2013. A gis-based assessment of maximum potential hydropower production in la plata basin under global changes. Renewable energy 50, 103–114.

De Silva, R., Dayawansa, N., Ratnasiri, M., 2007. A comparison of methods used in estimating missing rainfall data. Journal of agricultural sciences 3.

Delbari, M., Afrasiab, P., Jahani, S., 2013. Spatial interpolation of monthly and annual rainfall in northeast of iran. Meteorology and Atmospheric Physics 122, 103–113.

Di Piazza, A., Conti, F.L., Noto, L.V., Viola, F., La Loggia, G., 2011. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for sicily, italy. International Journal of Applied Earth Observation and Geoinformation 13, 396–408.

Eischeid, J.K., Pasteris, P.A., Diaz, H.F., Plantico, M.S., Lott, N.J., 2000. Creating a serially complete, national daily time series of temperature and precipitation for the western united states. Journal of Applied Meteorology 39, 1580–1591.

Erxleben, J., Elder, K., Davis, R., 2002. Comparison of spatial interpolation methods for estimating snow distribution in the colorado rocky mountains. Hydrological Processes 16, 3627–3649.

González, M.H., Cariaga, M.L., Skansi, M.d.l.M., 2012. Some factors that influence seasonal precipitation in argentinean chaco. Advances in Meteorology 2012.

Haberlandt, U., 2007. Geostatistical interpolation of hourly precipitation from rain gauges and radar for a large-scale extreme rainfall event. Journal of Hydrology 332, 144–157.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 2011. Robust statistics: the approach based on influence functions. volume 196. John Wiley & Sons.

Harris, I., Osborn, T.J., Jones, P., Lister, D., 2020. Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset. Scientific data 7, 1–18.

Hastie, T., 2011. gam: Generalized additive models. r package version 1.14.

Hastie, T., Tibshirani, R., 1987. Generalized additive models: some applications. Journal of the American Statistical Association 82, 371–386.

Hurtado, S.I., 2020. Package RobustLinearReg: Robust Linear Regressions. URL: https://CRAN.R-project.org/package=RobustLinearReg . r package version 1.2.0.

Hurtado, S.I., Agosta, E.A., 2020. El niño southern oscillation-related precipitation anomaly variability over eastern subtropical south america: Atypical precipitation seasons. International Journal of Climatology .

Hurtado, S.I., Agosta, E.A., Godoy, A., 2020a. Estudio exploratorio de forzantes de la variabilidad en baja frecuencia de la precipitacion en el chaco, argentina. Meteorologica 45, 71–92.

Hurtado, S.I., Zaninelli, P.G., Agosta, E.A., 2020b. A multi-breakpoint methodology to detect changes in climatic time series. an application to wet season precipitation in subtropical argentina. Atmospheric Research , 104955.

Hwang, Y., Clark, M., Rajagopalan, B., Leavesley, G., 2012. Spatial interpolation schemes of daily precipitation for hydrologic modeling. Stochastic environmental research and risk assessment 26, 295–320.

Jones, P., Lister, D., Harpham, C., Rusticucci, M., Penalba, O., 2013. Construction of a daily precipitation grid for southeastern south america for the period 1961–2000. International journal of climatology 33, 2508–2519.

Kajornrit, J., Wong, K.W., Fung, C.C., 2011. Estimation of missing rainfall data in northeast region of thailand using spatial interpolation methods. Australian Journal of Intelligent Information Processing Systems 13.

Kajornrit, J., Wong, K.W., Fung, C.C., 2012. A comparative analysis of soft computing techniques used to estimate missing precipitation records .

Kalteh, A.M., Berndtsson, R., 2007. Interpolating monthly precipitation by self-organizing map (som) and multilayer perceptron (mlp). Hydrological sciences journal 52, 305–317.

Kidd, C., Becker, A., Huffman, G.J., Muller, C.L., Joe, P., Skofronick-Jackson, G., Kirschbaum, D.B., 2017. So, how much of the earth’s surface is covered by rain gauges? Bulletin of the American Meteorological Society 98, 69–78. doi:10.1175/BAMS-D-14-00283.1.

Kurtzman, D., Navon, S., Morin, E., 2009. Improving interpolation of daily precipitation for hydrologic modelling: spatial patterns of preferred interpolators. Hydrological Processes: An International Journal 23, 3281–3291.

Kyriakidis, P.C., Journel, A.G., 1999. Geostatistical space–time models: a review. Mathematical geology 31, 651–684.

Li, J., Heap, A.D., 2008. A review of spatial interpolation methods for environmental scientists .

Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Ecological Informatics 6, 228–241.

Liebmann, B., Allured, D., 2005. Daily precipitation grids for south america. Bulletin of the American Meteorological Society 86, 1567–1570.

Ly, S., Charles, C., Degré, A., 2013. Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review. Biotechnologie, Agronomie, Société et Environnement 17, 392–406.

Magrin, G.O., Travasso, M.I., Rodríguez, G.R., 2005. Changes in climate and crop production during the 20th century in argentina. Climatic change 72, 229–249.

Mair, A., Fares, A., 2010. Assessing rainfall data homogeneity and estimating missing records in mākaha valley, o ‘ahu, hawai‘i. Journal of Hydrologic Engineering 15, 61–66.

Morales, J.L., Horta-Rangel, F.A., Segovia-Domínguez, I., Morua, A.R., Hernández, J.H., 2019. Analysis of a new spatial interpolation weighting method to estimate missing data applied to rainfall records. Atmósfera 32, 237–259.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

New, M., Hulme, M., Jones, P., 2000. Representing twentieth-century space–time climate variability. part ii: Development of 1901–96 monthly grids of terrestrial surface climate. Journal of climate 13, 2217–2238.

Paulhus, J.L., Kohler, M.A., 1952. Interpolation of missing precipitation records. Monthly Weather Review 80, 129–133.

Pebesma, E., Bivand, R.S., 2005. S classes and methods for spatial data: the sp package. R news 5, 9–13.

Pebesma, E., Heuvelink, G., 2016. Spatio-temporal interpolation using gstat. RFID Journal 8, 204–218.

Pebesma, E.J., 2004. Multivariable geostatistics in s: the gstat package. Computers & geosciences 30, 683–691.

Pellicone, G., Caloiero, T., Modica, G., Guagliardi, I., 2018. Application of several spatial interpolation techniques to monthly rainfall data in the calabria region (southern italy). International Journal of Climatology 38, 3651–3666.

Penalba, O.C., Vargas, W.M., 2008. Variability of low monthly rainfall in la plata basin. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling 15, 313–323.

Popescu, I., Brandimarte, L., Peviani, M., 2014. Effects of climate change over energy production in la plata basin. International journal of river basin management 12, 319–327.

Presti, R.L., Barca, E., Passarella, G., 2010. A methodology for treating missing data applied to daily rainfall data in the candelaro river basin (italy). Environmental monitoring and assessment 160, 1.

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, 81–94.

Rolla, A.L., Nuñez, M.N., Guevara, E.R., Meira, S.G., Rodriguez, G.R., de Zárate, M.I.O., 2018. Climate impacts on crop yields in central argentina. adaptation strategies. Agricultural Systems 160, 44–59.

Ruelland, D., Ardoin-Bardin, S., Billen, G., Servat, E., 2008. Sensitivity of a lumped and semi-distributed hydrological model to several methods of rainfall interpolation on a large basin in west africa. Journal of Hydrology 361, 96–117.

Sammut, C., Webb, G.I. (Eds.), 2010. Leave-One-Out Cross-Validation. Springer US, Boston, MA. pp. 600–601. URL: https://doi.org/10.1007/978-0-387-30164-8_469, doi:10.1007/978-0-387-30164-8_469.

Sattari, M.T., Rezazadeh-Joudi, A., Kusiak, A., 2017. Assessment of different methods for estimation of missing data in precipitation studies. Hydrology Research 48, 1032–1044.

Sen, P.K., 1968. Estimates of the regression coefficient based on kendall's tau. Journal of the American statistical association 63, 1379–1389.

Simolo, C., Brunetti, M., Maugeri, M., Nanni, T., 2010. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. International Journal of Climatology 30, 1564–1576.

Singh, V.P., 1988. Hydrologic systems: watershed modeling. volume 2. Prentice Hall.

Skansi, M.d.l.M., Brunet, M., Sigró, J., Aguilar, E., Arevalo Groening, J.A., Bentancur, O.J., Castellón Geier, Y.R., Correa Amaya, R.L., Jácome, H., Malheiros Ramos, A., Oria Rojas, C., Pasten, A.M., Sallons Mitro, S., Villaroel Jiménez, C., Martínez, R., Alexander, L.V., Jones, P., 2013. Warming and wetting signals emerging from analysis of changes in climate extreme indices over South America. Global and Planetary Change 100, 295–307. URL: https://www.sciencedirect.com/science/article/pii/S0921818112002172 https://linkinghub.elsevier.com/retrieve/pii/S0921818112002172 , doi:10.1016/j.gloplacha.2012.11.004.

Snepvangers, J., Heuvelink, G., Huisman, J., 2003. Soil water content interpolation using spatio-temporal kriging with external drift. Geoderma 112, 253–271.

Štěpánek, P., Zahradníček, P., Skalák, P., 2009. Data quality control and homogenization of air temperature and precipitation series in the area of the czech republic in the period 1961–2007. Advances in Science and Research 3, 23–26.

Tang, W., Kassim, A., Abubakar, S., 1996. Comparative studies of various missing data treatment methods-malaysian experience. Atmospheric Research 42, 247–262.

Teegavarapu, R.S., 2014. Statistical corrections of spatially interpolated missing precipitation data estimates. Hydrological Processes 28, 3789–3808.

Teegavarapu, R.S., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of hydrology 312, 191–206.

Teegavarapu, R.S., Tufail, M., Ormsbee, L., 2009. Optimal functional forms for estimation of missing precipitation data. Journal of hydrology 374, 106–115.

Terzi, Ö., 2012. Monthly rainfall estimation using data-mining process. Applied computational intelligence and soft computing 2012.

Theil, H., 1992. A rank-invariant method of linear and polynomial regression analysis, in: Henri Theil's contributions to economics and econometrics. Springer, pp. 345–381.

Turner, S.W., Hejazi, M., Kim, S.H., Clarke, L., Edmonds, J., 2017. Climate impacts on hydropower and consequences for global electricity supply investment needs. Energy 141, 2081–2090.

Venables, W.N., Ripley, B.D., 2013. Modern applied statistics with S-PLUS. Springer Science & Business Media.

Vicente-Serrano, S.M., Saz-Sánchez, M.A., Cuadrat, J.M., 2003. Comparative analysis of interpolation methods in the middle ebro valley (spain): application to annual precipitation and temperature. Climate research 24, 161–180.

Vieux, B.E., 2001. Distributed hydrologic modeling using gis, in: Distributed Hydrologic Modeling Using GIS. Springer, pp. 1–17.

Wagner, P.D., Fiener, P., Wilken, F., Kumar, S., Schneider, K., 2012. Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. Journal of Hydrology 464, 388–400.

Wehbe, M.B., Seiler, R.A., Vinocur, M.G., Tarasconi, I.E., 2018. Is it possible to completely adapt agriculture production to the effects of climate variability and change in central argentina? new approaches in face of new challenges, in: Theory and Practice of Climate Adaptation. Springer, pp. 443–463.

Westerberg, I., Walther, A., Guerrero, J.L., Coello, Z., Halldin, S., Xu, C.Y., Chen, D., Lundin, L.C., 2010. Precipitation data in a mountainous catchment in honduras: quality assessment and spatiotemporal characteristics. Theoretical and applied climatology 101, 381–396.

Wickham, H., 2016. ggplot2: elegant graphics for data analysis. springer.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 3–36.

Xia, Y., Fabian, P., Stohl, A., Winterhalter, M., 1999. Forest climatology: estimation of missing values for bavaria, germany. Agricultural and Forest Meteorology 96, 131–144.

Xie, Y., 2014. knitr: a comprehensive tool for reproducible research in r. Implement Reprod Res 1, 20.

Xie, Y., 2015. Dynamic Documents with R and knitr. volume 29. CRC Press.

Xu, W., Zou, Y., Zhang, G., Linderman, M., 2015. A comparison among spatial interpolation techniques for daily rainfall data in sichuan province, china. International Journal of Climatology 35, 2898–2907.

Yavuz, H., Erdoğan, S., 2012. Spatial analysis of monthly and annual precipitation trends in turkey. Water resources management 26, 609–621.

Young, K.C., 1992. A three-way model for interpolating for monthly precipitation values. Monthly Weather Review 120, 2561–2569.

Yozgatligil, C., Aslan, S., Iyigun, C., Batmaz, I., 2013. Comparison of missing value imputation methods in time series: the case of turkish meteorological data. Theoretical and applied climatology 112, 143–167.

Zhang, X., Srinivasan, R., 2009. Gis-based spatial precipitation estimation: A comparison of geostatistical approaches 1. JAWRA Journal of the American Water Resources Association 45, 894–906.

Zotelo, C.H., Martín, S.L., Camilloni, I.A., 2008. Estimación del tiempo de retardo de la onda de crecida en la cuenca superior del río uruguay. Meteorologica 33, 19–30.

1.7 Supplementary Materials

Figure S1: Boxplot of density of stations (amount of stations per $1000km^2$) in the peer-reviewed works from Table 1 (marked as “Other Studies”), and the density of stations in subtropical Argentina calculated for each station with a radius of 500km (marked as "500km Ratio"). The graph is in logarithmic scale for better visualization.

Figure S2: Monthly mean precipitation (annual cycle) for every station used in this study from Subtropical Argentina. Solid curve in color represents the annual cycle for each station. Each boxplot for the spatial distribution of monthly mean precipitation among stations is shown.

Figure S3: Percentage of the total number of times, regarding all stations, in which each method has been selected as the best (in red color), the second-best (in yellow color), the third-best (in green color), the third-worst (in cyan color), the second-worst (in pink color) and the worst (in blue

color) in every station according to the SRMSE. Vertical black dashed lines separate the method groups.

Figure S4: Boxplot of the minimum SRMSE regarding all methods for each station, for each season and form of application. The graph is in logarithmic scale for better visualization.

Figure S5: Map of SRMSE of GAM for every form of application and season. The GAM SRMSE is represented by the discrete color-dots scale.