Gap Filling of Monthly Temperature Data and Its Effect on Climatic Variability and Trends

SANTIAGO BEGUERÍA, MIQUEL TOMAS-BURGUERA, AND ROBERTO SERRANO-NOTIVOLI

Estación Experimental de Aula Dei–Consejo Superior de Investigaciones Científicas, Zaragoza, Spain

DHAIS PEÑA-ANGULO AND SERGIO M. VICENTE-SERRANO

Instituto Pirenaico de Ecología–Consejo Superior de Investigaciones Científicas, Zaragoza, Spain

JOSÉ-CARLOS GONZÁLEZ-HIDALGO

Departamento de Geografía, Universidad de Zaragoza, Zaragoza, Spain

(Manuscript received 3 April 2019, in final form 9 August 2019)

ABSTRACT

Observational datasets of climatic variables are frequently composed of fragmentary time series covering different time spans and plagued with data gaps. Most statistical methods and environmental models, however, require serially complete data, so gap filling is a routine procedure. However, very often this preliminary stage is undertaken with no consideration of the potentially adverse effects that it can have on further analyses. In addition to numerical effects and trade-offs that are inherent to any imputation method, observational climatic datasets often exhibit temporal changes in the number of available records, which result in further spurious effects if the gap-filling process is sensitive to it. We examined the effect of data reconstruction in a large dataset of monthly temperature records spanning over several decades, during which substantial changes occurred in terms of data availability. We made a thorough analysis in terms of goodness of fit (mean error) and bias in the first two moments (mean and variance), in the extreme quantiles, and in long-term trend magnitude and significance. We show that gap filling may result in biases in the mean and the variance of the reconstructed series, and also in the magnitude and significance of temporal trends. Introduction of a two-step bias correction in the gap-filling process solved some of these problems, although it did not allow us to produce completely unbiased trend estimates. Using only one (the best) neighbor and performing a one-step bias correction, being a simpler approach, closely rivaled this method, although it had similar problems with trend estimates. A trade-off must be assumed between goodness of fit (error minimization) and variance bias.

1. Introduction

Observational datasets covering long time periods are one of the most important data sources for climate research. Typically, these datasets are composed of a (frequently large) number of climatic records from

a network of meteorological stations, over time periods that usually span over several decades. Ideally, these datasets should be stable in time and free of missing values, but this is usually not the case. On the contrary, the number and geographical location of the stations where the meteorological variables are recorded usually change largely over time as new stations are created while others are abandoned, so the resulting time series differ in their temporal coverage. These changes in the dataset structure are usually not random, since systematic trends are frequent such as a progressive increase in

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-19-0244.s1>.

Corresponding author: Santiago Beguería, santiago.begueria@csic.es

This article is licensed under a [Creative Commons Attribution 4.0 license](http://creativecommons.org/licenses/by/4.0/) (<http://creativecommons.org/licenses/by/4.0/>).

DOI: 10.1175/JCLI-D-19-0244.1

the number of stations in the network with time. Recently, the substitution of traditional (manual) stations by automatic ones is causing large modifications in the observational networks. Moreover, very few climate records are serially complete, since missing values (data gaps) are frequent. Data gaps occur for a variety of reasons, ranging from not reported observations and transcription mistakes (in manual stations) to instrumental failure (more frequent in automatic stations), data transmission, or archiving.

These effects add together and result in temporally sparse datasets that are not well suited for many statistical analyses that require serially complete data, or to provide climatic input to environmental models such as ecohydrological simulation codes. As a direct effect, the existence of data gaps in the dataset increases the uncertainty of the mean and variance climatologies calculated upon those data, as shown by [Stooksbury et al. \(1999\)](#). For that reason, most studies include a data reconstruction stage prior to further analyses, as it is frequently the case in the contexts of hydrological modeling (e.g., [Kim and Pachepsky 2010](#); [Lee and Kang 2015](#)), or temperature rise assessments (e.g., [Coulibaly and Evora 2007](#); [Eischeid et al. 2000](#); [Kashani and Dinpashoh 2012](#)).

One of the most important steps in data reconstruction pipelines is the imputation of missing values, or gap filling. A wide variety of methods for handling data gaps in climatic datasets have been proposed. Among the simplest and most widely used methods are several variants of the K -nearest neighbors (KNN) algorithm, which make a local imputation of data gaps based on the data observed at K neighboring stations. Other popular approaches are regression methods (including KNN regression), geographically weighted regression, kriging, or thin-plate splines. Recently, more complex machine learning methods such as regression trees and random forests, genetic algorithms, or artificial neural networks have also been used.

These methods were developed with the aim of maximizing the goodness of fit (i.e., providing the highest R^2) or minimizing the error (i.e., the lowest sum of absolute or squared errors) while producing unbiased results (i.e., preserving the mean). While large efforts have been made with regards to minimizing the error of gap-filling methods, little or no attention has been given to other relevant aspects such as preserving the variance of the data. As a consequence, the reconstructed series are usually deflated, that is, they have a reduced variance with respect to the real climate, which may cause problems if this effect is not considered in further analyses. For instance, variance deflation is an important problem for any study looking at climate variability or extreme quantiles, or for those studies based on anomalies, that is, deviations around a mean value ([Beguéría](#)

et al. 2016). Another aspect that might be affected by gap-filling procedures is the magnitude and significance of temporal trends, an issue of the greatest relevance in many climatic studies.

The goal of the present study is to examine the effect of data reconstruction (gap filling) in terms of goodness of fit (error minimization), bias in first- and second-order moments (mean and variance), in the extreme quantiles, and in long-term trend estimation. Temporal effects of systematic changes in the dataset structure such as the number of records available are investigated, and spurious effects on long-term trend magnitude and significance estimation are assessed.

Unlike most previous studies that used relatively small datasets, this study uses a very large dataset of monthly maximum and minimum temperature records over mainland Spain. We use a version of the KNN method, and different reconstruction options are described that aim at minimizing the variance bias and preserving the extremes on the reconstructed data. In particular, two weighting approaches (based on geographical and correlation distances) and two variance bias correction methods (pre- and post-bias correction) are compared, using a leave-one-out approach that allows comparison of reconstructed against observed values. The effect of the number of neighbors used in the reconstruction is also assessed.

The main hypothesis is that a trade-off exists between goodness of fit and variance preservation, and that no single solution is able to provide optimum results regarding both aspects, although a good compromise can be found. Specific hypotheses regarding the different reconstruction options are formulated in [section 3](#).

2. Reconstruction approaches

KNN weighted averaging methods are among the most popular gap-filling approaches in climate studies. They are, for instance, used in the elaboration of several widely used gridded datasets, such as the University of East Anglia Climatic Research Unit land surface air temperature dataset (CRUTEM; [Osborn and Jones 2014](#)). They are also the preferred approach for the development of reference series for the detection of inhomogeneities.

The basis of this family of methods consists of creating, for each location of interest, a reference series formed as a weighted average of the data observed at some neighboring stations. In a gap-filling context, the values of this reference series are then used for imputing each missing data point existing in the series of interest (usually termed the “candidate” series). Different variants of this basic method exist such as the single best estimator, arithmetic averaging, inverse distance and

FIG. 1. Different weighting schemes. The station for which new data are being calculated is represented as a red dot, while neighboring stations are represented as black dots (stations with data) or gray dots (stations that do not have data for that time and therefore do not participate in the calculation). The width of the segments represents the weights assigned to each station, which may be dependent on (a) the geographical distance (closest stations get a higher weight), or other criteria, such as (b) correlation distance (highly correlated stations get a higher weight). A restriction can be set to reduce the number of stations participating in the calculation, such as (c) a maximum distance

other weighted averages, or the ratio and difference methods. We shall cover all these variants and a few new ones in the following formal explanation of the approaches used on this study.

We will assume that we have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with n observations of the variable of interest on p stations, a matrix $\mathbf{G} \in \mathbb{R}^{n \times p}$ identifying the observations and the missing data, and a matrix of weights $\mathbf{W} \in \mathbb{R}^{p \times p}$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} g_{11} & \cdots & g_{1p} \\ g_{21} & \cdots & g_{2p} \\ \vdots & \ddots & \vdots \\ g_{n1} & \cdots & g_{np} \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pp} \end{bmatrix}.$$

Using $i = (1, 2, \dots, n)$ and $j = (1, 2, \dots, p)$ as row and column indices, respectively, we can represent the time series of data at a given station as the column vector $\mathbf{x}_{*j} = (x_{1j}, x_{2j}, \dots, x_{nj})$. The values on the matrix $\mathbf{G}$ are defined as

$$g_{ik} = \begin{cases} 1 & \text{if } \exists x_{ik}, \text{ and} \\ 0 & \text{otherwise;} \end{cases}$$

and

$$\text{diag}(\mathbf{W}) = (w_{11}, w_{22}, \dots, w_{pp}) = 0.$$

Estimates $\hat{x}_{ij}$ can thus be computed for any x_{ij} as a weighted average of the observations at the remaining stations:

$$\hat{x}_{ij} = \sum_{k=1}^p \left(\frac{w_{kj}}{\sum_{k=1}^p g_{ik} w_{kj}} \right) = \frac{\sum_{k=1}^p x_{ik} w_{kj}}{\sum_{k=1}^p g_{ik} w_{kj}}, \quad (1)$$

for $\sum_{k=1}^p g_{ik} > 0$. Estimates $\hat{x}_{ij}$ constitute the elements of matrix $\hat{\mathbf{X}}$, which can then be used to fill in the gaps (missing data) in $\mathbf{X}$. Moreover, the estimates $\hat{x}_{ij}$ that coincide with observed values x_{ij} can be used to compute various evaluation statistics.

The time series obtained as column vectors of $\hat{\mathbf{X}}$, $\hat{\mathbf{x}}_{*j}$, are occasionally referred to as the “reference series,” while the observed time series $\mathbf{x}_{*j}$ that are being reconstructed are referred to as the “candidate series.” We would use this terminology occasionally on the remaining of this article.

It is noteworthy to mention that the data used for reconstructing the missing values in a given column of $\mathbf{X}$ may change in time depending on the data availability at the remaining stations, as represented by the column vector $\mathbf{g}_{*j}$.

a. Computation of the weights

Determination of the weights w_{kj} is central to the method, and several criteria may be applied (Fig. 1). A frequent approach is to assume that stations that are

geographically closer should have stronger weights than those located farther apart, leading to

$$w_{kj} = \frac{1}{D(j, k)^2}, \quad (2)$$

where $D(j, k)$ is the geographical distance between the stations, what makes $\mathbf{W}$ an inverse distance matrix. Another common approach is to use the correlation distance:

$$w_{kj} = \rho(j, k)^4, \quad (3)$$

where $\rho(j, k)$ is the Pearson's correlation coefficient between data at stations j and k based on the common data between them.

As mentioned earlier, the diagonal elements of $\mathbf{W}$ are zero in order to prevent the observed data at the candidate station (if there is an observation) from being used in the computation of $\hat{x}_{ij}$. In this way, estimates can also be computed for the observed data, what may be interesting for a number of things including (but not restricted to) goodness-of-fit estimation, if it can be assumed that the statistics computed for the periods with data are valid also for the period without data.

To reduce memory allocation and computation time, zero weights w_{ij} are typically assigned to a large part of the station pairs, such as

- pairs of stations located at distances higher than a specified threshold;
- pairs of stations with correlation lower than a specified threshold;
- pairs of stations with no overlapping period ($\sum_i g_{ij}g_{ik} = 0$), or with overlapping period below a certain threshold ($\sum_i g_{ij}g_{ik} < \tau$).

This can dramatically reduce the computation time, while having a minimum impact on the reconstruction since stations that are very far apart in terms of geographic or correlation distance will have almost negligible weights. In spatially sparse datasets, on the other hand, it may make sense to reduce the computation to those cases where observational data exist within a reasonable distance in order to prevent using data from stations located very far apart.

It is also possible to limit the number of neighbor series used to the K closest (or best correlated) ones. This may help to prevent some effects that are linked to the number of neighbors used, as we shall see later.

Given the strong seasonal variability present in most climatic variables, Eq. (1) and therefore the determination of weights needs to be applied after segmenting the dataset $\mathbf{X}$ by months, that is, for each monthly

dataset separately. However, we shall not make a distinction (i.e., use subindices or other symbology to represent the different months) here or in the remaining discussion for sake of simplicity.

1) RESCALING OF THE NEIGHBOR SERIES (PRE-BIAS CORRECTION)

Equation (1) can yield unbiased estimates $\hat{x}_{ij}$ only if the data on the columns of $\mathbf{X}$ were identically distributed. However, this is often not the case with climate data, since the statistical characteristics (e.g., mean and variance) usually change across stations. If the data to compute the reference series $\hat{\mathbf{x}}_{*j}$ were used without any correction, the result will not be homogeneous with respect to the observed data series $\mathbf{x}_{*j}$, that is, their statistical properties will differ. Moreover, since the stations used in the process may change in time, temporal inhomogeneities could be introduced in $\hat{\mathbf{x}}_{*j}$ as data with different statistical properties were used in the reconstruction.

A bias correction procedure is therefore required before applying Eq. (1). One common approach is to rescale the data of the neighbor stations by taking differences or ratios with respect to their mean, that is, converting the original data to anomalies, and then summing or multiplying by the mean of the candidate series:

$$x_{ik}^j = (x_{ik} - \mu_{\mathbf{x}_{*k}}^j) + \mu_{\mathbf{x}_{*j}}^k, \quad \text{and} \quad (4)$$

$$x_{ik}^j = \left(\frac{x_{ik}}{\mu_{\mathbf{x}_{*k}}^j} \right) \mu_{\mathbf{x}_{*j}}^k, \quad (5)$$

where x_{ik}^j stands for the data at station k rescaled to station j , and $\mu_{\mathbf{x}_{*k}}^j$ ($\mu_{\mathbf{x}_{*j}}^k$) is the arithmetic mean of the data at station k (j), considering only the overlapping data between stations k (j) and j (k):

$$\mu_{\mathbf{x}_{*k}}^j = \frac{\sum_{i=1}^n g_{ij}g_{ik}x_{ik}}{\sum_{i=1}^n (g_{ij}g_{ik})}, \quad \text{and} \quad (6)$$

$$\mu_{\mathbf{x}_{*j}}^k = \frac{\sum_{i=1}^n g_{ik}g_{ij}x_{ij}}{\sum_{i=1}^n (g_{ij}g_{ik})}. \quad (7)$$

The differences method [Eq. (4)] is suited to unbounded variables such as temperature, while the ratio method [Eq. (5)] is best suited to bounded variables such as precipitation, which can only take zero or positive values. Both methods have been widely applied.

However, they can only correct for differences in the mean between $\mathbf{x}_{*k}$ and $\mathbf{x}_{*j}$. Equality of variances between stations must therefore be assumed, which is often a not valid assumption. Consider, for instance, the temperature variability of two nearby stations located at the coast and inland: it is to be expected that the station near the coast will have much lower temperature variability than the inland one, in addition to a different mean. In fact, differences in the variance across stations are probably the norm and not an exception, so a true normalization of the series (i.e., converting them to standardized anomalies before proceeding with the gap-filling process) is probably a better option than the methods of differences or ratios. Thus,

$$x_{ik}^j = \left(\frac{x_{ik} - \mu_{\mathbf{x}_{*k}}^j}{\sigma_{\mathbf{x}_{*k}}^j} \right) \sigma_{\mathbf{x}_{*j}}^k + \mu_{\mathbf{x}_{*j}}^k, \quad (8)$$

where $x_{ik}^j = [(x_{ik} - \mu_{\mathbf{x}_{*k}}^j)/\sigma_{\mathbf{x}_{*k}}^j] \sigma_{\mathbf{x}_{*j}}^k + \mu_{\mathbf{x}_{*j}}^k$ and $\sigma_{\mathbf{x}_{*j}}^k$ are the sampling standard deviations of $\mathbf{x}_{*k}$ and $\mathbf{x}_{*j}$, respectively, computed for the overlapping data between both stations:

$$\sigma_{\mathbf{x}_{*k}}^j = \sqrt{\frac{\sum_{i=1}^n (g_{ij} g_{ik} x_{ik} - \mu_{\mathbf{x}_{*k}}^j)^2}{\sum_{i=1}^n (g_{ij} g_{ik} - 1)}}, \quad \text{and} \quad (9)$$

$$\sigma_{\mathbf{x}_{*j}}^k = \sqrt{\frac{\sum_{i=1}^n (g_{ik} g_{ij} x_{ij} - \mu_{\mathbf{x}_{*j}}^k)^2}{\sum_{i=1}^n (g_{ij} g_{ik} - 1)}}. \quad (10)$$

Equation (8) assumes that the data are normally distributed, but the method can be generalized to other distributions. For instance, the Gamma or the Pearson III distributions are suitable options for precipitation data (McKee et al. 1993; Guttman 1999).

Thus, after bias correction Eq. (1) can be rewritten as

$$\hat{x}_{ij} = \frac{\sum_{k=1}^p x_{ik}^j w_{kj}}{\sum_{k=1}^p g_{ik} w_{kj}}. \quad (11)$$

2) RESCALING OF THE REFERENCE SERIES (POST-BIAS CORRECTION)

In most occasions, the weighted estimates resulting from Eq. (11) are directly used to fill in the missing data.

FIG. 2. Method overview: red, candidate series to reconstruct (observed data); blue, neighbor series (observed data); gray, candidate and neighbor series (missing data); dotted, bias-corrected series; purple, final reference series.

Pre-bias correction ensures that the weighted averages $\hat{x}_{ij}$ are unbiased estimators of the missing observations, that is, that their average will not systematically under or overestimate the average of the observations in the data series being reconstructed. However, since they are averages over a number of stations, they will have by definition a reduced variance. Spatial and temporal fluctuations in the variance of the reconstructed series are not desirable for many climatological applications, as they may affect the analysis of extreme values or climate variability will be affected. It is also not desirable for any statistical technique based on the computation of standardized anomalies, such as computation of climatic indices, or empirical orthogonal functions (EOFs).

For this reason, a second bias correction can be applied in order to correct for bias in the variance of the reconstructed data, for what only full standardization can guarantee a proper correction. We call this process post-bias correction, consisting of applying Eq. (8) to the reconstructed (reference) series resulting from the application of Eq. (11).

A step-by-step overview of the method is provided in Fig. 2.

b. Error estimation

Obtaining a reliable estimation of the uncertainty together with the estimates $\hat{x}_{ij}$ is of paramount importance. However, and unlike the simple arithmetic mean, there is no universally accepted expression for the standard error of a weighted mean. Avoiding computation-intensive bootstrap procedures that may be prohibitively expensive in computation time, a good approximation to

the standard error of a weighted mean is the one proposed by Cochran (1977):

$$\text{se}_{\hat{\mathbf{x}}_{*j}} = \frac{n}{(n-1) \left(\sum_{i=1}^n w_{ij} \right)^2} \left[\sum_{i=1}^n (w_{ij} x_{ij} - \mu_{w_{*j}} \hat{\mathbf{x}}_{*j})^2 - 2 \hat{\mathbf{x}}_{*j} \sum_{i=1}^n (w_{ij} - \mu_{w_{*j}}) (w_{ij} x_{ij} - \mu_{w_{*j}} \hat{\mathbf{x}}_{*j}) + \hat{\mathbf{x}}_{*j}^2 \sum_{i=1}^n (w_{ij} - \mu_{w_{*j}})^2 \right] \quad (12)$$

where $\mu_{w_{*j}}$ is the weighted mean of the vector of weights:

$$\mu_{w_{*j}} = \frac{\sum_{i=1}^n g_{ij} g_{ik} w_{ik}}{\sum_{i=1}^n g_{ij} g_{ik}}. \quad (13)$$

The method of Cochran (1977) assumes that the data are independent, but in the case of missing data imputation a high correlation usually exists between the observations at neighboring stations. Therefore, it is necessary to account for the reduction in the effective sample size as a consequence of correlation. Otherwise we would underestimate the standard error, by a factor that increases with the degree of correlation between the samples. Hence, a modified version of the original equation by Cochran was used, by considering the effective sample size n' instead of the actual number of observations involved in computing the weighted mean n :

$$n' = \frac{n}{[1 + (n-1)\mu_{\rho_j}]}, \quad (14)$$

where μ_{ρ_j} is the weighted mean of the vector of correlations with station j :

$$\mu_{\rho_j} = \frac{\sum_{i=1}^n g_{ij} g_{ik} \rho_{ik}}{\sum_{i=1}^n g_{ij} g_{ik}}. \quad (15)$$

The values $\text{se}_{\hat{\mathbf{x}}_{*j}}$ quantify the uncertainties in determining the (weighted) mean $\hat{\mathbf{x}}_{*j}$. A prediction interval for $\hat{\mathbf{x}}_{*j}$ may easily be calculated from

$$[\hat{\mathbf{x}}_{*j} - z \text{se}_{\hat{\mathbf{x}}_{*j}}, \hat{\mathbf{x}}_{*j} + z \text{se}_{\hat{\mathbf{x}}_{*j}}], \quad (16)$$

where z is a standard normal score. For instance, a value of $z = 1.96$ can be used for calculating a 95% prediction interval around the estimated values $\hat{\mathbf{x}}_{*j}$.

FIG. 3. Spatial structure of the dataset: geographic distribution of the maximum temperature observatories, with indication of eleven serially complete time series. UTM coordinates (m); reference system: WGS84/UTM zone 30N (EPSG:32630).

3. Case study and research hypotheses

A dataset of monthly minimum and maximum temperature ($^{\circ}\text{C}$) was compiled from daily records at 3030 stations of the Spanish meteorological service [La Agencia Estatal de Meteorología (AEMET)]. The spatial distribution of the stations is provided in Fig. 3 (and in Fig. S1 in the online supplemental material). Quality control of the daily records and monthly maximum and minimum temperature averages were computed by AEMET according to general WMO guidelines. Due to undocumented changes in location, instruments, local environment, and observing practices, the original dataset was subject to a semiautomated homogenization process, as described in Gonzalez-Hidalgo et al. (2015).

The period of study spanned between January 1951 and December 2014. The number of stations with data varied largely within this period, peaking at nearly 1700 during the 1990s starting from around 400 stations in 1951 (Figs. 4 and S2). After the peak, a decreasing number of stations is observed in the last two decades. Overall, the dataset contained 39.7% observations and 60.3% data gaps, with a higher concentration of data gaps in the first two decades of the period.

For each variable (maximum and minimum temperature) the reconstruction method was applied to fill the missing data, and serially complete datasets were obtained. Reconstructed time series $\hat{\mathbf{x}}_{*j}$ were computed for each station, including the missing data but also the periods with observed data. This allowed comparing the reconstructed data with the observations, and

FIG. 4. Temporal structure of the dataset: time series of the number of maximum temperature series available at each month. Gray horizontal lines in the background indicate the data availability at each station.

calculating a number of evaluation statistics for each station in the dataset. These included the coefficient of determination (R^2) and the mean absolute error (MAE) as measures of goodness of fit; the mean error (ME) and the percent bias (PBIAS) as measures of bias; the ratio of means (rM) and the ratio of standard deviations (rSD) as measures of bias in the first two moments of the distribution; and the ratio of the 5th (rQ05) and 95th quantiles (rQ95), as measures of bias in the estimation of extremes; and the ratio of the trend estimation (rTrend), using linear regression against time. The statistics were calculated for each month separately, and then averaged for each station. The median values over the set of stations were then calculated, and their kernel distributions were plotted.

In addition, a validation dataset was created consisting on 11 data series that were complete (i.e., contained no missing data) during the whole study period. The location and identifying codes of these stations are shown in Fig. 3. This serially complete dataset was used

for validating the results obtained using the whole dataset, by calculating the same set of evaluation statistics. Additionally, and since they covered the whole period of study, the magnitude and significance of linear temporal trends over the period 1951–2014 was computed and compared between the observed and the reconstructed series. The Mann–Kendall trend test (Kendall 1975) was used to determine the significance of the trends, while the Theil–Sen estimator (Theil 1950; Sen 1968) was used for determining their magnitude. All the computations, including the reconstruction methods and the statistical tests, were programmed in R language, version 3.4.0 (R Core Team 2017).

Alternative reconstruction options were used and their effect on the resulting reconstructed datasets was assessed. The different options have been explained in the previous section, and the acronyms used thereafter are given in Table 1.

Two weighting methods were compared, based on the inverse Euclidean distance (labeled Dis) and on the

TABLE 1. Summary of reconstruction options and acronyms used in this article.

Acronym	Weighting	Rescaling	Bias correction	Neighbors
DstInf	Euclidean distance	No rescaling	No correction	All
CorInf	Pearson's correlation	No rescaling	No correction	All
DstDifPreInf	Euclidean distance	Differences	Pre-bias correction	All
CorDifPreInf	Pearson's correlation	Differences	Pre-bias correction	All
DstStdPreInf	Euclidean distance	Standardization	Pre-bias correction	All
CorStdPreInf	Pearson's correlation	Standardization	Pre-bias correction	All
DstDifPostInf	Euclidean distance	Differences	Pre- and post-bias correction	All
CorDifPostInf	Pearson's correlation	Differences	Pre- and post-bias correction	All
DstStdPostInf	Euclidean distance	Standardization	Pre- and post-bias correction	All
CorStdPostInf	Pearson's correlation	Standardization	Pre- and post-bias correction	All
DstOne	Euclidean distance	No rescaling	No correction	One
CorOne	Pearson's correlation	No rescaling	No correction	One
DstDifPreOne	Euclidean distance	Differences	Pre-bias correction	One
CorDifPreOne	Pearson's correlation	Differences	Pre-bias correction	One
DstStdPreOne	Euclidean distance	Standardization	Pre-bias correction	One
CorStdPreOne	Pearson's correlation	Standardization	Pre-bias correction	One

Pearson's correlation (labeled Cor) between station pairs. The expectation was that the Pearson's correlation method would produce better results, at least in terms of goodness of fit (hypothesis H1).

To assess the effect of different rescaling methods for the neighbor series, we alternatively used the traditional differences (Dif) method versus full standardization (Std). It was expected that the traditional differences method would only correct the mean bias, while generating problems in the estimation of the variance (hypothesis H2).

Bias correction of the neighboring series prior to computing the weighted average (Pre bias correction) was expected to enhance the goodness of fit of the reconstructed series, as compared to no bias correction (hypothesis H3).

The effect of the number of neighbors K was assessed by performing different reconstructions with values between 1 and 10, and also by using all the available neighbors that met the conditions (coded as $K = \text{Inf}$). In all cases, only the neighbor series meeting the following criteria were considered:

- being at a distance no greater than 100 km;
- having a correlation $\rho \geq 0.6$ with the series being reconstructed;
- having an overlapping period with the series being reconstructed of at least seven observations.

Increasing the number of neighbors was expected to improve the goodness of fit between observed and reconstructed data (or reducing the error), but also to deflate the variance of the reconstructed data with respect to the observed set (hypothesis H4).

In addition to pre-bias correction, a second bias correction after computing the weighted average (post-bias correction) was expected to help correcting the variance bias when a large number of neighbors was used for the reconstruction, hence combining an improved goodness of fit with a lower variance bias (hypothesis H5).

To check for temporal changes, the same set of evaluation statistics and kernel density plots were computed for each decade during the study period. Since the number of stations with available data largely changed during the study period, it was expected that temporal changes would be found in the validation statistics (hypothesis H6).

Finally, the effects of the reconstruction in trend estimation were assessed by computing the ratio of the linear trend magnitudes computed with reconstructed and observed data. Since any reconstruction introduces noise in the data, it was expected that it would also deteriorate the estimation of the trend magnitude, with a tendency toward underestimating the absolute magnitudes of any existing trends (hypothesis H7).

Overall, reconstruction based on the Pearson's correlation weighting, with a large number of neighbors and pre- and post-bias correction based on full standardization was expected to yield the best results considering all these aspects, although it was not clear if these results would be consistent in time due to the strong changes in the dataset structure.

Another open question was how would this scheme compare to using just one (the best correlated) neighbor.

4. Results

A main summary of the results is provided in Table 2 for maximum temperature, for all the reconstruction options defined in the previous section. The table shows median values of evaluation statistics over the whole set of stations. Additionally, kernel density curves are provided in Fig. 5 for a subset of evaluation statistics and combinations of methods, for maximum temperature only. The kernel density curves complement the information provided by the median values, as they show the range of variability and the skewness of each statistic. The corresponding results for minimum temperature are provided in the supplemental material (Table S1 and Fig. S5).

All methods achieved satisfactory values of goodness of fit, with mean values of R^2 over the set of stations exceeding 0.75 for maximum temperature and above 0.70 for minimum temperature when all the available stations were used, and above 0.60 when only one station was used. The MAE was lower than 0.7°C for both variables and all reconstruction methods, with the exception of the reconstructions with no bias correction that yielded errors in excess of 1°C. Between the two variables analyzed, a better fit and lower errors were found for maximum than for minimum temperatures, probably due to the more local character of the latter.

As expected, correlation weighting achieved consistently better fit than distance weighting (hypothesis H1), while full standardization outperformed the differences method (hypothesis H2). Similarly, reconstructions with no bias correction resulted in much worse fit and higher errors (hypothesis H3). These results were consistent across the different method combinations, irrespective of the number of neighbors used. The number of neighbors, as expected, resulted in slightly worse results in terms of goodness of fit and MAE when only one neighbor was used, although the difference with respect to using all the available neighbors was very low when correlation weighting and full standardization was also used (hypothesis H4). In the case of the minimum temperature, using only one neighbor achieved

TABLE 2. Median values of several evaluation statistics comparing reference and observed maximum temperature series: coefficient of determination (R^2), mean absolute error (MAE), mean error (ME), percent bias (PBIAS), ratio of means (rM), ratio of standard deviations (rSD), ratio of the 5th quantiles (rQ05), ratio of the 95th quantiles (rQ95) and ratio of trends (rTrend). The best results for each evaluation statistic and block of methods are in bold.

Method	R^2 (—)	MAE (0.1°C)	ME (0.1°C)	PBIAS (%)	rM (—)	rSD (—)	rQ05 (—)	rQ95 (—)	rTrend (—)
DstInf	0.763	10.70	−0.364	−0.146	0.999	0.930	1.010	0.990	0.719
CorInf	0.795	12.40	−1.240	−0.629	0.994	0.892	1.010	0.981	0.711
DstDifPreInf	0.803	5.83	0.000	0.000	1.000	0.919	1.020	0.989	0.819
CorDifPreInf	0.821	5.64	−0.007	0.000	1.000	0.893	1.020	0.984	0.822
DstStdPreInf	0.811	5.49	−0.001	0.000	1.000	0.914	1.010	0.990	0.817
CorStdPreInf	0.826	5.26	−0.004	0.000	1.000	0.896	1.020	0.986	0.824
DstDifPostInf	0.803	5.83	0.000	0.000	1.000	0.919	1.020	0.989	0.819
CorDifPostInf	0.821	5.65	0.000	0.000	1.000	0.893	1.020	0.984	0.822
DstStdPostInf	0.811	5.61	0.000	0.000	1.000	1.000	1.000	1.000	0.888
CorStdPostInf	0.826	5.39	0.000	0.000	1.000	1.000	0.999	0.999	0.917
DstOne	0.612	13.60	0.562	0.321	1.000	1.040	0.999	1.010	0.713
CorOne	0.678	13.20	0.520	0.417	1.000	1.020	1.000	1.010	0.796
DstDifPreOne	0.696	7.71	0.000	0.000	1.000	0.998	1.000	0.998	0.809
CorDifPreOne	0.810	5.93	0.000	0.000	1.000	0.956	1.010	0.994	0.904
DstStdPreOne	0.704	7.34	0.000	0.000	1.000	0.999	1.000	1.000	0.824
CorStdPreOne	0.818	5.60	0.000	0.000	1.000	0.995	1.000	0.999	0.927

the highest R^2 , albeit the minimum error was attained by postcorrection using the highest number of neighbors. The density curves of R^2 were positively skewed, with values ranging between 0.4 and 1. The density curves of the rM statistic were almost equal, symmetrical, and centered around 1, with a range of variation between 0.99 and 1.01, showing that the mean bias was correctly addressed by all methods compared in the figure.

Including bias correction was crucial to achieve unbiased errors, as shown by the ME and PBIAS that were always very close to zero except when no bias correction was used. In that case, differences were found depending on the number of neighbors used: while using all available neighbors yielded a negative bias, using only one neighbor had the opposite effect. This could be a random effect, since in the case of the minimum temperature the bias was also positive when using all available neighbors. The magnitude of the mean bias was smaller for maximum temperature (between 0.3° and 1.2°C, or 0.15% and 0.63%) than for minimum temperature (reaching even −5.7% in the worst case). The rM, which also informs about systematic bias in the reconstruction, was equal or very close to one in all cases, except again for no bias correction.

The median rSD varied between less than 0.9 (implying underestimation of the standard deviation of around 10%) and 1, and once again yielded slightly worse results (i.e., higher underestimation) for minimum temperature.

Strong differences were apparent depending on the bias correction method used and the number of neighbors.

Only the reconstructions based on all neighbors available with post-bias correction (StdPostInf) and one neighbor (StdPreOne), both with full standardization, achieved values of the rSD equal (or very close) to 1, in accordance with hypotheses H4 and H5, respectively. The kernel density curves of rSD showed, for the differences method especially, a large dispersion of the rSD statistic and negatively skewed distributions with most of the density concentrated in values lower than 1, indicating that variance under estimation was prevailing and that it could be very severe in some cases. Post-bias correction, by definition, yielded values of exactly 1 in all cases, while using only one neighbor with full standardization yielded a distribution of this statistic very tightly concentrated around the ideal value of 1.

The rQ05 was higher than one for several methods, indicating a positive bias (i.e., tendency toward overestimating) in the estimation of the lower values of the distribution of both maximum and minimum temperatures. Again, only the reconstructions including post-bias correction or based in only one neighbor yielded values very close to 1, indicating no bias. Similarly, the rQ95 was lower than one, indicating a tendency toward underestimation of the higher values of monthly temperature, except for the same reconstruction options referred above.

The median rTrend was also lower than one in all cases, indicating that the reconstructed series tended to have reduced temporal trends than the observations. Unlike with the previous statistics, in this case the results were slightly better for minimum than for maximum temperature.

FIG. 5. Kernel density curves and median values (vertical lines) of evaluation statistics between observed and reference maximum temperature series over the whole dataset, computed with different correction methods: only one (best) neighbor and pre-bias correction, $K = 1$; and all available neighbors with pre- and post-bias corrections, $K = Inf$. Coefficient of determination (R^2), ratio of means (rM), ratio of standard deviations (rSD), and ratio of trends ($rTrend$).

The options that obtained the best results so far (i.e., CorStdPreOne and CorStdPostInf) also obtained median values of $rTrend$ closest to one among all the methods compared, although they did not solve completely the problem. This is best seen in the kernel distribution of the statistic, where a negative bias is evident for all methods. The distribution curves also show that the range of variation of this statistic was very large (between 0 and 2), indicating that the temporal trends can be very much affected by the reconstruction process. However, due to the fact that the statistic is a ratio and the trend magnitudes can achieve very low values in some cases, which would lead to large values of the ratio for relatively small differences, this last result has to be taken with care.

Significant differences were thus found between different reconstruction options. Relevant differences were found as a consequence of the type of bias correction applied, but also of the number of neighbors used in the reconstruction. Two methods, CorStdPreOne and CorStdPostInf, stood out from the rest regarding the different evaluation statistics, with a marginal advantage toward the second one. In-depth comparison between methods is undertaken in the following sections, as well as the analysis of temporal changes in the validation statistics as related to changes in the structure of the dataset (hypothesis H6) and on trend estimation (hypothesis H7).

a. Effect of the weighting method: Distance versus correlation

Substantial differences existed among the two weighting schemes compared: inverse distance Dst and correlation Cor. In general, either methods produced unbiased results, with PBIAS very close to zero and rM close to one, and nonskewed kernel densities of rM. There were very few to no differences between the two schemes in terms of mean bias (ME, PBIAS, and rM).

However, the correlation scheme yielded consistently higher R^2 and lower MAE than inverse distance reconstructions, that is, it had better goodness of fit. However, the opposite was found with relation to variance bias, since distance weighting yielded values of rSD that were always closer to one than correlation weighting, except when post-bias correction was also used. This is not surprising, due to the mentioned trade-off between goodness of fit and variance bias. In agreement with this trade-off, the method providing the best fit also resulted in a higher variance bias.

b. Effect of bias correction

Bias correction had a strong influence on the results of the reconstruction. The reconstructions with no bias

FIG. 6. Maps of the ratio between the standard deviations (rSD) of reconstructed and observed maximum temperature, for different reconstructions with pre-bias correction, based on all the neighbors available. Colors represent the mean value of all the stations contained in each hexagonal bin.

correction were poor in terms of goodness of fit, mean bias, and variance bias, and also in terms of reproducing the observed temporal trends. Both pre- and post-bias correction, on the contrary, resulted in much better statistics in comparison.

Two different bias correction strategies, pre- and postcorrection, were tested. While in the first case the neighboring series were matched to the candidate one before computing the weighted average, in the second case an additional correction of the resulting series was performed. In addition to this, two alternative bias correction methods were tested, named the differences method and full standardization. The results were different for different combinations of pre- and post-bias correction and the two bias correction methods.

When using pre-bias correction only, reconstructions based on standardization (StdPre) performed only marginally better those based on the differences method (DifPre) in terms of goodness of fit and error (higher R^2 and lower MAE), and there were no apparent differences in terms of rM, rSD, or $rTrend$. Interestingly, and unlike other data reconstruction options, however, the effect of the bias correction method in the variance bias was not spatially random, as shown in Fig. 6. The differences method, independently of the weighting scheme used, that is, distance or correlation, resulted in inflated variance (rSD higher than one) in areas with reduced temperature variability compared to their surroundings, such as

FIG. 7. Kernel density curves of maximum temperature (data in 0.1°C) at four stations, comparing the original data and reconstructed data with pre-bias correction and post-bias correction.

it is the case of coastal stations with respect to inland ones. Full standardization, on the other hand, showed a good ability at matching the variance of the observations, even in those difficult cases. This result suggests that full standardization is a better method even when using pre-bias correction only.

When post-bias correction was also performed the advantages of full standardization became more obvious, since it allowed correction of the variance bias, as shown by the values of the rSD statistic equal to one. This can be better seen in the kernel densities of Fig. 5, where pre- and post-bias correction methods are compared to each other. Thus, pre-bias correction only (dashed lines in the figure) did not yield unbiased results for rSD, independently of the weighting scheme used (differences or standardization). In all cases, a negatively skewed distribution centered in values lower than one were obtained, indicating a tendency toward underestimating the standard deviation of the data.

The effect of bias correction is illustrated by the examples in Fig. 7, depicting kernel densities of observed and reconstructed maximum temperature data at four randomly selected stations. A similar figure for minimum temperature is provided in the supplemental material (Fig. S4). The reduced variance of the data reconstructed with pre-bias correction contrasts with a better coincidence between the observed data and post-bias reconstruction.

Since the effect of post-bias correction is expanding the variance of the reconstructed data symmetrically around the mean, the R^2 is not affected by this transformation and the same values are obtained with respect to prior bias correction. MAE, on the other hand, is slightly increased in post-bias correction with respect to the pre-bias correction, due precisely to this variance expansion. Thus, obtaining unbiased variance data reconstructions is achieved in post-bias correction at the expense of an increase in the mean error.

c. Effect of the number of neighbors

The number of neighbors used also had a noticeable effect on the performance of the reconstruction method, as seen in the comparison between two extreme situations: using only one neighbor [the closest or the most correlated one (One)] or using all the neighbors available (Inf). Additional analyses with varying number of neighbors K illustrate the effects of modifying this parameter in greater detail (Figs. 8 and S5). The figure shows results using correlation weighting and full standardization (CorStdPre and CorStdPost), but very similar results were obtained for other options.

The goodness of fit (R^2) increased as did the number of neighbors used in the reconstruction, while the MAE (not shown) decreased. This effect was more pronounced if the distance method was used for computing the weights, but it was also present with correlation weighting (results only shown for correlation weighting). Based on this result alone, one could conclude that using the highest possible number of neighbors available, or a high number of them, would be the best option in any occasion. But there is a downside to this, because an opposite effect was found with respect to the variance bias: thus, the rSD decreased noticeably as the number of neighbors used in the reconstruction increased. When only one (the best) neighbor was used, in fact, the median rSD was very close to one, even if only pre-bias correction was performed, implying almost no variance bias. As a downside, however, the goodness of fit (R^2) also was lowest when only one neighbor was used.

This illustrates, again, the trade-off between goodness of fit and variance bias. While using $K = 1$ yields suboptimal results in terms of goodness of fit (or error magnitude), on the other hand it is beneficial for preserving the variance of the reconstructed time series.

As we shall see in the next section, and also shown in Fig. 8, post-bias correction also had a strong effect as it eliminated completely the variance bias, while preserving the R^2 and only marginally increasing the MAE. Hence, both post-bias correction and reducing the number of neighbors used in the reconstruction to only one (the best correlated) could be used for reducing the

FIG. 8. Kernel density curves of evaluation statistics comparing observed and reference maximum temperature series as a function of the maximum number of neighbors, computed with pre- and post-bias corrections, full standardization, and correlation weighting: coefficient of determination (R^2), ratio of means (rM), ratio of standard deviations (rSD), and ratio of trend magnitudes (rTrend).

variance bias in the reconstructed series. In both cases, preserving the variance is obtained at the expense of assuming a reduction in the goodness of fit, which interestingly had a similar magnitude in both cases.

There were no bias effects in the estimation of temporal trends with respect to the number neighbors used. The only effect of increasing the number of neighbors was slightly reducing the variance of the distribution of the rTrend statistic.

d. Temporal effects of changes in the dataset

As has been stressed several times, reconstruction based on correlation weighting and full standardization with pre- and post-bias correction and using the maximum number of neighbors available at each time (StdCorPostInf) was the best option overall, once a trade-off was accepted between unbiasedness at the cost of (slightly) increased error. Only the option of reconstructing based on correlation weighting with pre-bias correction and full standardization based on one (the best correlated) neighbor (StdCorPreOne) rivaled this method in terms of goodness of fit and variance bias, with the computational advantage of not requiring a postcorrection step. The only price of using only one neighbor with respect to using all and performing a post-variance correction was a slightly reduced

goodness of fit (median R^2 of 0.818 vs 0.826 and 0.758 vs 0.777 for maximum and minimum temperature, respectively) and increasing error (MAE of 5.60 vs 5.39 and 5.47 vs 5.18). The difference between the two methods, thus, was marginal.

However, there are other effects that could affect the performance of reconstruction methods and that have not been considered yet. In particular, variations in the dataset structure may have an impact in the performance of the reconstruction methods, as they could introduce temporal changes in the error or goodness of fit, or in the variance bias. This, in turn, could influence further climatological analyses.

As described earlier, there are substantial temporal changes in the number of stations available in the dataset (Figs. 4 and S2), with significantly fewer data available during the first decades of the study period. As a consequence, the number and correlations of the neighbors used in the reconstruction will also change with time. As it has been shown, the number of neighbors used has a strong influence in the results of the reconstruction process, and the correlation (or distance) with these neighbors would also have an effect on the reconstruction. In particular, if the number of stations is substantially reduced during a certain temporal period (as it is usually the case during the first decades of most datasets), an increased uncertainty (higher

FIG. 9. Decadal variation of evaluation statistics (kernel densities) between observed and reference maximum temperature series for pre- and post-bias correction: coefficient of determination (R^2), ratio of means (rM), and ratio of standard deviations (rSD). The vertical lines indicate the mean values of each density curve.

errors) are to be expected, but also an increased variability of the reconstructed data, as the number of neighboring stations available for reconstruction becomes smaller.

When using only one neighboring station (for instance in StdCorPreOne), decreasing the number of stations would imply that the mean correlation will also be reduced, with an immediate effect on the goodness of fit and the mean absolute error. This is also true for reconstructions based on many neighbors (such as StdCorPostInf), although the effect might be alleviated since there are more data used for reconstruction.

The results of temporally segmented analysis are shown in Fig. 9 (and Fig. S6). Kernel density curves are shown of the three main evaluation statistics for the best performing reconstruction methods, for each decade during the study period (except the last period, which consisted of only five years of data). Interpretation of these results needs to be done in relation with the

number of observations in the dataset on each decade, as shown in Fig. 4.

As expected (hypothesis H6), some changes in the goodness of fit and in the variance bias were apparent between decades, and some unexpected results also appeared. In general terms, the main differences between decades affected the range of values of the statistics (i.e., the width of the kernel density curves), while no significant changes in the mean biases were found. An evolution of the R^2 statistic as the study period advances is apparent, with lower median values and higher spread in early decades, coinciding with lower data availability. It is interesting to notice that, despite the fact that the maximum number of stations available was achieved during the 1990–99 decade and decreased afterward, the best fit was achieved later, in the 2000–09 decade. While a slight reduction in the goodness of fit could be expected in this last period, it is possible that

TABLE 3. Validation statistics: median values of several evaluation statistics comparing reference and observed maximum temperature series in the validation dataset and magnitude (Theil–Sen estimator) of the temporal trend. Values are provided for reconstructed series computed with no correction, pre–bias and post–bias correction, and an infinite maximum number of neighbors, for the whole dataset (100%) and for two reduced datasets (50% and 10%). The best results for each dataset block are in bold.

	R^2 (—)	MAE (1°C)	ME (0.1°C)	PBIAS (%)	rM (—)	rSD (—)	rQ05 (—)	rQ95 (—)	rTrend (—)
100% dataset									
CorNoInf	0.863	10.60	−1.360	−1.2800	0.987	0.969	0.959	0.983	0.514
StdCorPreInf	0.930	3.60	0.168	0.0833	1.000	0.900	1.020	0.988	0.893
StdCorPostInf	0.930	3.34	0.000	0.0000	1.000	1.000	1.000	1.000	0.987
StdCorPreOne	0.922	3.48	0.0734	0.0333	1.000	0.999	1.000	0.998	0.984
50% dataset									
CorNoInf	0.808	10.30	−1.450	−1.6900	0.983	0.995	0.963	0.996	0.324
StdCorPreInf	0.915	3.69	0.218	0.1170	1.000	0.905	1.020	0.988	0.870
StdCorPostInf	0.915	3.49	0.000	0.0000	1.000	1.000	1.000	0.999	0.949
StdCorPreOne	0.873	4.43	0.1250	0.0750	1.000	0.997	1.000	1.000	0.964
10% dataset									
CorNoInf	0.750	10.60	−3.580	−2.6000	0.974	1.120	0.951	0.973	0.203
StdCorPreInf	0.825	5.00	0.240	0.2080	1.000	1.000	0.996	1.000	1.070
StdCorPostInf	0.825	4.52	0.000	0.0000	1.000	1.000	1.000	0.999	1.184
StdCorPreOne	0.825	5.00	0.2400	0.2080	1.000	1.000	0.996	1.000	1.070

the spatial distribution of the set of stations was improved resulting in a better fit.

The rM did not show changes in their median values among decades, although the dispersion of the distribution showed a progressive reduction as the number of stations in the dataset increased (later decades). A similar effect was found with respect to the rSD: no changes were found in the median values of this statistic, although the dispersion of the values decreased with increasing number of observations in the dataset with time.

It is interesting to notice that post–bias correction, which yielded values of rSD equal to one when the whole period was analyzed (Fig. 5 and Table 2), did not perform in the same way when analyzed per decades. On the contrary, post–bias correction did not yield perfectly unbiased rSD statistics anymore, and a slight tendency toward overestimating the variance of the reconstructed data was found. In particular, post–bias correction did not outperform the reconstruction based on only one neighbor anymore, as it did when the whole time period was analyzed. This can be attributed to the stationarity assumption the bias correction methods relies upon, which implies assuming that the only source of variance in the data is random interannual variability. In the case of long-term temporal trends (or cycles) in the data, this assumption does not hold true since there is more than one source of variation besides random variability. As an effect, and since the bias correction methods assume stationarity, the resulting data have an increased random variance. Interestingly, the median value of the rSD was slightly higher than 1 for StdCorPostInf than

that for StdCorPreOne, revealing a tendency toward overestimating the variability of the data.

We have come to a point where we need to pay attention to possible temporal trends present in the data. It is not possible (or recommended) to perform trend analysis on fragmentary data series such as those in the complete observational dataset. Further analysis, therefore, needs to be performed using the validation dataset, as it is done in the following section.

e. Validation dataset: Main statistics

As described in the methods section, we created a validation dataset consisting on 11 serially complete records for further validation. The primary goal of this analysis was to validate the main results shown in the previous sections, on a reduced and more controlled dataset. The availability of serially complete records covering the whole period allowed also conducting a trend analysis, and further assess the effect of reconstruction on long-term trend estimation. Finally, in addition to a reconstruction based on all the data available two other reconstructions were done after removing 50% and 90% of the data series, in order to reproduce situations of data scarcity.

Evaluation statistics calculated on the validation dataset are presented in Table 3 for maximum temperature, and similar results for minimum temperature can be found in the supplemental material (Table S2). Results are only shown for correlation distance weighting and full standardization (StdCor), since this combination yielded the best results in previous

FIG. 10. Observed and reconstructed time series of the regional mean annual maximum temperature and residuals of the reconstructed series.

analyses, but the three bias correction options (no correction, pre-bias correction only, and pre- and post-bias correction) are retained, as well as using only one neighbor (StdCorPreOne).

Goodness-of-fit statistics were very good, even better than those obtained with the whole dataset, with R^2 values above 0.9 and mean absolute errors around 0.35°C in the 100% dataset, except when no bias correction was used (CorNoInf). Reconstruction with pre-bias correction only (StdCorPreInf) was capable of producing unbiased results, with values of the rM equal to one, but only the reconstruction with pre- and post-bias correction (StdCorPostInf) was able to also produce results with the same variance as the observations (rSD equal to one), resulting also in unbiased low and high quantiles (rQ05 and rQ95). Very close results were obtained with only one neighbor (StdCorPreOne). All the reconstructions tended to underestimate the temporal trends on average, although post-bias correction yielded rTrend values closer to one. These results were consistent with those obtained using the complete dataset.

Degradation of the dataset by removing 50% of the series did not degrade the validation statistics in the same proportion. The best R^2 , for instance, varied from 0.930 to 0.915, while the mean absolute error increased only slightly, from 0.334° to 0.349°C . This suggests that the dataset used in this study is very dense and contains a significant amount of redundant information.

Interestingly, the R^2 was more affected in the case of one neighbor methods (StdCorPreOne), which went from 0.922 to 0.873.

When the dataset was strongly degraded by removing 90% of the series the evaluation statistics were affected in a more significant way, with values of R^2 dropping to 0.825 (both for StdCorPostInf and StdCorPreOne) and larger errors. Interestingly, however, the ratio of trends improved slightly on this reduced dataset.

Reconstruction with bias correction seemed to do a very good job with respect to goodness of fit and reproducing the observed mean and standard deviation of the data, but some problems still existed related to long-term trend estimation, as shown by the rTrend values. Given the importance of estimation of trends in climate records, this merits further attention.

f. Validation dataset: Trend analysis

Since the records in the validation set are serially complete it was possible to compare the trend estimations over the whole period based on the observed and the reconstructed data. Additionally, it was possible to create a regional series by computing the mean of all the stations, and compare the trend estimations.

The regional time series is shown in the upper part of Fig. 10 for maximum temperature, while a similar plot for minimum temperature can be found in the supplemental material (Fig. S7). Inspection of these plots

TABLE 4. Long-term linear trend magnitude according to the Theil–Sen estimator of the regional and individual station series of the annual mean maximum temperature using the validation dataset ($0.1^{\circ}\text{C decade}^{-1}$). An asterisk (*) indicates that the trend is significant at the $\alpha = 0.05$ confidence level (Man–Kendall tau test), and the closest significant approximations to the observed trend are in bold.

	Regional	1024E	1387	2867	3195	7228	8175	8416	9170	9246	9390	9981A
Observed dataset												
	2.73*	1.41	2.71*	3.06*	2.74*	1.42*	3.10*	2.90*	2.56*	1.24	4.30*	3.74*
Reconstructed dataset (100%)												
CorInf	1.54*	0.90	3.40*	1.70*	1.30*	0.73	1.10	3.28*	0.96	0.83	1.22	0.81
StdCorPreInf	2.28*	1.52*	2.10*	2.30*	2.73*	1.30*	2.15*	2.88*	1.94*	1.67*	3.84*	3.26*
StdCorPostInf	2.53*	1.64*	2.30*	2.40*	2.98*	1.50*	2.51*	3.21*	2.19*	1.89*	4.07*	3.69*
StdCorPreOne	2.19*	2.04*	1.95*	3.11*	2.64*	1.36*	2.31*	2.78*	2.66*	1.94*	4.32*	3.68*
Reconstructed dataset (50%)												
CorInf	1.44	−0.39	3.46*	1.28*	1.94*	0.46	3.10*	4.17*	0.06	−0.57	0.48	1.11
StdCorPreInf	2.22*	1.55*	2.02*	2.08*	2.69*	1.44*	2.42*	2.62*	1.95*	1.65*	3.74*	3.14*
StdCorPostInf	2.49*	1.59*	2.19*	2.31*	2.91*	1.74*	2.80*	2.93*	2.09*	1.77*	4.05*	3.55*
StdCorPreOne	2.06*	1.90*	2.12*	2.23*	2.64*	1.12*	2.79*	3.05*	2.84*	1.79*	3.66*	3.81*
Reconstructed dataset (10%)												
CorInf	2.17*	−0.19	—	—	—	0.53	—	0.59	0.77	0.14	—	—
StdCorPreInf	2.62*	1.76*	—	—	—	1.52*	—	2.60*	2.71*	1.93*	—	—
StdCorPostInf	2.68*	1.67*	—	—	—	1.94*	—	3.04*	2.07*	1.77*	—	—
StdCorPreOne	2.21*	1.76*	—	—	—	1.52*	—	2.60*	2.71*	1.93*	—	—

reveals that the reconstruction algorithm did a good job in approximating the observed data, irrespective of the method used (StdCorPreOne, StdCorPreInf, and StdCorPostInf are shown). The observed and reconstructed series exhibit an apparent upward trend, in particular since the early 1970s. A closer inspection, however, reveals subtle differences between the observed and the reconstructed data, and it seems that the

reconstructed series had a systematic deviation with respect to the observed data. In particular, it seems that the reconstructed regional series of maximum temperature tended to overestimate the observed values at the beginning of the period and to underestimate them at the end, while the contrary is observed for minimum temperatures. To further assess this, a plot of the residuals between the observed and the reconstructed

FIG. 11. Time series of the mean absolute error (MAE) and of the standard error (SEM) of the maximum temperature on the validation set, for a reconstruction based on all the neighbors available (StdCorPostInf).

TABLE 5. Main characteristics of gap-filling comparison studies, sorted by publication year. Variables: precipitation (PCP), mean temperature (TMP), maximum temperature (TMX), minimum temperature (TMN). Methods: nearest neighbor (NN), single best estimator (SBE), arithmetic averaging (AA), inverse distance weighting (IDW), other weighted averages (WA), normal ratio (NR) and difference (ND), simple and multiple K -nearest-neighbor regression (OLS), multiple K -nearest-neighbor generalized linear regression (GLM), truncated total least squares regression (TTLS), quantile mapping (QM), geographically weighted regression (GWR), kriging (K), optimal interpolation (OI), Bayesian model averaging (BMA), thin-plate/cubic splines (SP), principal component analysis reconstruction (PCA), empirical orthogonal function reconstruction (EOF), artificial neural network (ANN), genetic algorithm (GA), adaptive neuro-fuzzy inference system (ANFIS), Monte Carlo Markov chain simulation (MCMC), regression tree (RT), random forest (RF). Goodness-of-fit (GOF) and error statistics: Pearson's correlation coefficient (r), coefficient of determination (R^2), coefficient of efficiency (CE), Nash-Sutcliffe efficiency (NSE), reduction of error (RE), linear error in probability space (LEPS), mean absolute error (MAE), median absolute error (mAE), mean absolute percent error (MAPE), mean square error (MSE), root-mean-square error (RMSE), absolute error (AE), mean relative error (MRE), skill scores (SS), correlation dimension (CD), Akaike Information Criterion (AIC). Bias statistics: comparison of means (RM), mean error (ME). Variance statistics: comparison of standard deviations (SD), comparison of interquartile ranges (IQR), comparison of deciles (DEC), frequencies of extremes (EX). Probability distribution function (PDF) comparison: t test, Kolmogorov-Smirnov (KS), Shapiro-Wilks (SW), Kendall's tau (KT), Friedman (F), Wilcoxon (W), Ansari-Bradley (AB), comparison of skewness (SK), quantile-quantile ($Q-Q$) plots. Trend analysis: Mann-Kendall (MK), Sen's slope estimator (SS).

Reference	Variable	Method	GOF/error	Mean bias	Variance bias	PDF	Trend
Richardson (1981)	Daily PCP, TMX, TMN	MCMC, OLS	RM	SD	—	—	—
Young (1992)	Monthly PCP	NR, OLS	r , MAE	—	—	—	—
Xia et al. (1999)	Daily PCP, SBE, OLS	AA, IDW, NR, MAE	MAE	—	—	—	—
Kuligowski and Barros (1998)	Daily PCP	ANN	RMSE, r	RM	—	—	—
Eischeid et al. (2000)	Daily PCP, TMX, TMN	SBE, NR, IDW, OI, OLW	r , RMSE	—	—	t test	—
Jeffrey et al. (2001)	Daily PCP, TMX, TMN	SP	R^2 , MAE, RMSE	ME	—	—	—
Teegavarapu and Chandramouli (2005)	Daily PCP	IDW, ANN, K	RMSE, MRE, MAE, R^2 , r	—	—	—	—
Coulibaly and Evora (2007)	Daily PCP, TMX, TMN	ANN	MAE, r	RM	SD	—	—
Di Luzio et al. (2008)	Daily PCP, TMX, TMN	IDW, OLS	MAE, r	RM	—	—	—
Ramos-Calzado et al. (2008)	Monthly PCP	OLS	—	ME	—	—	—
You et al. (2008)	Daily TMX, TMN	IDW, OLS	RMSE, R^2 , CE	—	—	—	—
Aly et al. (2009)	Monthly PCP	OLS	SS, r , MAE, RMSE	ME	—	—	—
Teegavarapu et al. (2009)	Daily PCP	IDW, GA	RMSE, MAE, AE, r	—	—	—	—
Kim and Pachepsky (2010)	Daily PCP	OLS, ANN	RMSE, r	—	—	KS	—
Simolo et al. (2010)	Daily PCP	OLS	MAE, RMSE	ME, RM	SD	—	—
Vicente-Serrano et al. (2010)	Daily PCP	NN, IDW, OLS	RMSE	—	EX	—	—
Di Piazza et al. (2011)	Monthly PCP	IDW, OLS, GWR, ANN, K	MSE, MAE, r	ME	—	—	—
Eccel et al. (2012)	Daily PCP, TMP	IDW	RMSE, MAE	ME	—	—	—
Kashani and Dinpashoh (2012)	Monthly PCP, TMX, TMN	AA, IDW, NR, SBE, OLS, K , SP, ANN, ANFIS, GA	MAE, CE, SS	—	—	—	—
Tardivo and Berti (2012)	Daily TMP	OLS	—	ME	—	—	—
Yozgatligil et al. (2013)	Monthly PCP, TMP	AA, NR, WA, ANN, MCMC	R^2 , RMSE, CD	—	—	—	—
Feng et al. (2013)	Seasonal PCP	TTLS	r , R^2 , RE, CE	—	—	—	—
Henn et al. (2013)	Daily TMP	OLS, K , EOF	RMSE, MAE	—	—	SW, F , W	—
Teegavarapu (2014)	Daily PCP	OI	SS, MAE, NS, RMSE, LEPS	—	—	KS, AB	—
Lee and Kang (2015)	Daily PCP	WA, OLS	—	—	—	SW, F , W	—
Londhe et al. (2015)	Daily PCP	ANN	RMSE, r , MAE, CE	—	—	—	—
Fang and Li (2016)	Annual TMP	BMA	r , MAE, RMSE	—	—	—	—
Zhang et al. (2016)	Annual TMP	BMA	r , CE, RE	—	—	—	—

TABLE 5. (Continued)

Reference	Variable	Method	GOF/error	Mean bias	Variance bias	PDF	Trend
Serrano-Notivol et al. (2017a), Serrano-Notivol et al. (2017b)	Daily PCP	GLM	SS, r , MAE	ME, RM	SD, DEC	—	—
Teegavarapu and Nayak (2017)	Daily PCP	OLS	—	—	—	KS	MK, SS
Woldesenbet et al. (2017)	Daily PCP, TMX, TMN	IDW, WA	RMSE, MAPE	—	—	KT	—
Zhou et al. (2018)	Daily PCP, TMP	PCA	RMSE	ME	—	—	—
Li et al. (2018)	Daily TMX, TMN	GWR	RMSE, R^2 , MAE	—	—	—	—
Diez-Sierra and del Jesus (2019)	Sub-daily PCP	RF	R^2	—	SD	SK	—
Singh and Xiaosheng (2019)	Daily PCP	IDW, WA, OLS, QM	R^2 , RMSE, AIC	PBIAS	—	Q-Q	—

series has been included at the bottom of Fig. 10 (and Fig. S8). A long-term pattern seems to exist in the residuals, with positive values predominating in the first half of the period and negative ones in the second half for maximum temperature, and opposite pattern in the case of minimum temperature. This pattern is especially noticeable if the period between 1970 and 2014 is considered. It is difficult to extract conclusions about the comparison between the three reconstruction methods based on the residuals plot, although it seems that both StdCorPostInf and StdCorPreOne did a better job than StdCorPreInf. Since this has a direct consequence on the estimation of the magnitude and significance of temporal trends based on reconstructed data, a closer look was taken.

Table 4 (and Table S3) provides the estimated trend magnitudes according to the Theil–Sen estimator and the significance of these trends according to the Mann–Kendall test, for the series included in the validation set and for a regional series constructed as the mean of those series, for the three reconstruction methods plus no bias correction (CorNoInf). Some of the series could not be reconstructed entirely in the most restrictive data scenario (10%), so the trends were not computed. We will not discuss here the legitimacy of the observed trends or the representativity of this reduced set of stations to represent the climate of the whole territory, since that is not the goal of this work. But it is interesting to compare the results of trend analysis based on different reconstruction methods.

The observed regional trend, considering the whole study period, was $+0.273^{\circ}\text{C decade}^{-1}$ for maximum temperature and $+0.230^{\circ}\text{C decade}^{-1}$ for minimum temperature, both significant at the $\alpha = 0.05$ confidence level. At the individual station level the trends varied largely, both for maximum and minimum temperature, but positive and significant values were found for both variables, with the exception of two stations that had not significant trends. The magnitude of the trends was strongly underestimated in reconstructed datasets when no bias correction was used (CorNoInf), and occasionally resulted in nonsignificant or even reversed sign trends. For instance, the regional trend was estimated at $0.154^{\circ}\text{C decade}^{-1}$ for maximum temperature, and $0.151^{\circ}\text{C decade}^{-1}$ for minimum temperature. When bias correction was applied, the estimated trend magnitudes were closer to the observed ones but still not very accurate. Pre-bias correction (StdCorPreInf) estimated the regional trend of maximum temperature at $0.228^{\circ}\text{C decade}^{-1}$, while post-bias correction (StdCorPostInf) estimated it at 0.253° , and 0.209° and $0.236^{\circ}\text{C decade}^{-1}$ for minimum temperature. At the individual station level, in most cases the magnitude of

the trend was underestimated, although in a few cases it was overestimated. Although post-bias correction provided the best estimation of the temporal trend at the regional level both for maximum and minimum temperature, for individual stations there was no clear winner, and all the methods provided the best estimation for some stations with the exception of no bias correction, which was clearly the worst option.

g. Validation dataset: Error estimation

The validation dataset also allowed us to check the accuracy of the error estimation, which provides a measure of the uncertainty associated to the reconstructed data. Of greatest practical interest is the evaluation of temporal changes in the uncertainty of the reconstruction, as the density of the dataset changes. A time series of the estimated standard error of the reconstructed data (SEM) for the validation dataset and the MAE is presented in Fig. 11 (and Fig. S8) over the validation dataset, for maximum and minimum temperatures.

The temporal evolution of the MAE shows a downward trend that is in agreement with the time evolution in the number of data series available in the dataset, as we saw earlier in the results segmented by decades. The MAE varied between 0.5°C at the beginning of the study period and around 0.25°C at the end. The time evolution of the SEM shows a similar downward pattern, ranging between 0.5° and 0.2°C . When comparing both time series, it is clear that the time evolution of the MAE becomes stationary earlier (around the 1970s) than the SEM (1990s), and the error values during the latter period are slightly higher than those estimated by the SEM. This may reveal that the SEM tends to underestimate the true error (uncertainty) of the reconstructed data, which recommends that further effort must be devoted to improving the error estimation in this kind of approach.

5. Discussion

The presence of data gaps and incomplete series is ubiquitous in climatological studies, and a data reconstruction step is mandatory in most cases. The gap-filling process, however, may alter the data and have consequences on the analyses undertaken on them. Despite its importance, there are still many studies that provide no details on the treatment of missing data, as reviewed by Woldesenbet et al. (2017).

On the other hand, there are a growing number of publications that compare different reconstruction methods. An analysis of the quality assessment of the reconstructed data on these studies (Table 5) reveals a strong focus on goodness-of-fit statistics such as the

Pearson's correlation coefficient or the coefficient of determination, or on error statistics such as the mean absolute error or the root-mean-square error, which in many occasions are the only statistics reported. Other effects that can be highly relevant to climatic analyses are thus often ignored. Only a few studies checked the mean bias (systematic deviations) of the reconstructed data, for instance by reporting the mean error or the percent bias before and after data reconstruction, or by comparing the means of both datasets (Coulibaly and Evora 2007; Di Piazza et al. 2011; Ramos-Calzado et al. 2008; Simolo et al. 2010). When these statistics were reported, they were used mainly as goodness-of-fit indicators, and little was said about systematic biases that are apparent in some studies. Similarly, only a few studies compared the probability distributions before and after reconstruction using tests such as the t test, Shapiro-Wilks, Friedman, or Wilcoxon, or by graphical methods such as Q-Q plots. Again, when these were reported they were mostly used as a general measure of goodness of fit (Eischeid et al. 2000; Lee and Kang 2015), although (Singh and Xiaosheng 2019) specifically reported on the quality of the reconstructions in the extreme quantiles of the variable.

Very few studies, too, looked at variance bias (systematic deviations in the variance) of the reconstructed data, that is, by comparing the standard deviation or the interquartile range before and after data reconstruction, or at changes in the extreme values, that is, by comparing the 10th and 90th percentiles. However, it is a well-known fact that data reconstruction tends to lead to systematic underestimation of the variance that constrains the extremes of the variable and affects the variability of the time series. Coulibaly and Evora (2007) analyzed not only changes in the mean but also in the standard deviation of reconstructed precipitation and temperature, and stressed that a trade-off existed between goodness of fit (mean bias) and variance preservation (variance bias) that prevents optimizing both at the same time. In a rather thorough review, Simolo et al. (2010) remarked on the underestimation of precipitation extremes (variance bias) and presented results grouped by deciles. Serrano-Notivol et al. (2017a,b) also reported the ratio of standard deviations and presented results by deciles, and commented on the compression of the variance that occurred on the reconstructed data. Teegavarapu and Nayak (2017) concluded that data reconstruction introduced bias in estimations of extreme precipitation, resulting in underestimation of heavy events.

The temporal effects of data reconstruction have also received little attention. However, this is a topic of

paramount importance, most notably if temporal trends are present in the variables analyzed, or when there are structural changes in the dataset such as a systematic change (e.g., increase) in the number of stations with time. Very few studies reported on temporal changes on the validation statistics. [Simolo et al. \(2010\)](#) showed the temporal evolution of goodness of fit (MAE, RMSE) and mean bias (ME), but they only commented on the magnitudes of these statistics and did not address the likely existence of a temporal trend, which seems a plausible hypothesis on some of their plots. The consequences of trends in fit and bias statistics on further analyses performed on the reconstructed dataset, which can be especially important on trend analysis, were also not addressed. [Feng et al. \(2013\)](#) also reported time series of goodness-of-fit statistics (r and RE) and noticed long-term patterns that they could relate to changes in the number of observations available. The article by [Teegavarapu and Nayak \(2017\)](#) is the only one of the reviewed studies that focused on the effect of reconstruction of rainfall data on long-term trend estimates of extreme precipitation. However, they did not compare between trends computed with filled and observed data (using, for instance, a leave-one-out approach).

Our results show that some frequently used gap-filling methods had undesired effects on some of the data properties, especially with regards to variance bias and low and high quantiles, and on trend magnitude and significance. Performing a bias correction that accounts for both mean and variance bias (i.e., by using full standardization) of both the neighbor series and the reconstructed series (pre- and post-bias correction) was the only method that allowed using a high number of neighbors in the reconstruction while preserving the variance of the reconstructed series. Similar results have been recently shown by [Singh and Xiaosheng \(2019\)](#), who performed a post-bias correction based on quantile mapping, that is, correcting the whole distribution of the reconstructed variable and not just the mean.

Using only one (the best correlated) neighbor, with pre-bias correction based on full standardization, also achieved very good results in that respect. Other bias correction methods such the differences method, or not performing a post-bias correction if more than one neighbor was used, resulted in a reduced variance and affected the extremes of the reconstructed variable. The worst option was to not perform a bias correction of any sort, which lead to not only higher absolute error but also to bias in the mean and a strong variance compression. On many occasions, however, only a simple differences pre-bias correction is used or no bias correction is performed at all.

The number of neighbors used also had a large influence on the reconstruction, especially if a post-bias correction is not performed. Several authors addressed the issue of the selection of neighbors, leading to a variety of recommendations. Most gap-filling and data reconstruction studies applied correlation and distance thresholds to select the neighboring series to be used, with values that differ among them. [Woldesenbet et al. \(2017\)](#) made a review of correlation and distance threshold values used by different authors.

The most relevant variable, however, was the number of neighbors used. This has also been subject to scientific debate. For instance, [Easterling and Peterson \(1995\)](#) suggested to use around five neighbors and never less than two, while more recently [Tardivo and Berti \(2014\)](#), using a multiple regression approach, found that the optimum number of neighbors tended to oscillate between one and four. Other authors used different numbers, and there are even several studies where the exact number is not provided. As shown by [Teegavarapu and Chandramouli \(2005\)](#) and by [Henn et al. \(2013\)](#), incrementing the number of neighbors has the effect of reducing the mean error of the reconstructed series. This was expected and confirmed by our results, which also show that this improvement is achieved at the expense of increasing the compression of the variance. The main advantage of using only one neighbor is that it reduces the variance compression to a minimum. This is an aspect that has been very seldom studied, but it has relevant consequences on the climatological studies performed upon the reconstructed data, most especially those related with climatic variability (e.g., thermal range), or with extremes (e.g., heat waves, frost).

Our results show that using only one neighbor (the best correlated) and performing a bias correction with full standardization leads to results that are only slightly worse than using a high number of neighbors in terms of goodness of fit, but on the contrary it allows for eliminating the variance compression and therefore reducing the bias of high and low quantiles of the variable. This was already noticed by [Vicente-Serrano et al. \(2010\)](#), who used a direct attribution of daily precipitation data from the closest neighbor in order to not reduce the frequency of high and extreme precipitation events, and it has also been shown recently by [Teegavarapu and Nayak \(2017\)](#).

All the methods tested influenced the estimation of long-term trends. Therefore, this effect seems to be unavoidable, since any imputation process introduces a certain amount of noise in the data and therefore increases the difficulty of detecting a trend. In any case, this is an aspect of the greatest importance and needs to be accounted for when long-term trends are estimated based on reconstructed data. In particular, there is a

relevant bias toward underestimating the magnitude and significance of the trends, as was already warned by [Teegavarapu and Nayak \(2017\)](#). Given the relevance and potential consequences of this effect on climate change studies, this is an issue that merits further study and probably new techniques should be developed to eliminate this effect. In particular, bias-correction techniques that preserve observed trends in the mean of the data have been developed for correcting GCM outputs ([Cannon et al. 2015](#)), but it remains unclear how these techniques could be adapted to the reconstruction of incomplete instrumental records.

Our results belong to a potentially favorable study case in which a high density of stations and data are available. If this is not the case, the results of the reconstruction might be worse. The validation results with only 50% and 10% of the data are probably more representative of many studies.

Similarly, a word of caution needs to be made with respect to the experiment done here, since our validation results are based on a leave-one-out procedure that covers the whole observed data series. That means that reconstructed series equal in length to the whole observed series were computed and evaluated. In a real case, however, only the gaps in the series would be subject to reconstruction, so any effects of the reconstruction would be restricted to those gaps.

The time structure of data gaps within the dataset, therefore, becomes a relevant issue. In an ideal case where there are no significant changes in the number of stations during the period of interest and where data gaps occur randomly in time, these effects are also random and may have little effect on posterior analysis besides increasing the uncertainty of the results, as shown by [Stooksbury et al. \(1999\)](#). In most real cases, however, the situation is far from this scenario and the occurrence of data gaps is not random. For instance, as we move back in the past the number of stations recording data usually becomes lower, and data gaps in the series also tend to be more frequent, so a higher fraction of the data would need to be imputed. Also, temporal biases are related to changes in the structure of the dataset such as the number of neighboring stations available, the correlation or the distance between stations ([Tardivo 2015](#)). As a result of both effects, the consequences of data reconstruction are not temporally random and may introduce bias in the analyses, which may include the generation of spurious trends.

To finish, it must be mentioned that although we have focused our study on gap-filling methods there are other issues that may affect the mean, variance, and trends of climatic records. In particular, inhomogeneities in the time series due to changes in

station location, instrumentation or measuring protocols may have an impact on further climatological studies ([Peterson et al. 1998](#); [Reeves et al. 2007](#); [Venema et al. 2013](#)). Thus, it is usual that the data undergo a homogenization process before (or sometimes after) gap filling, as for instance in the case of the dataset used in our case study. Available techniques, however, are not perfect and suffer from a number of issues. Although we did not deal with homogenization issues here and it is a topic that merits a dedicated study, it is worth mentioning that the same of very similar approaches as the ones discussed on this article are also used when reference series are developed for inhomogeneity detection and correction. Thus, the main results of this study could also apply in this context.

Similarly, in the creation of gridded datasets from station observations the KNN method is very often applied, and the results of this study could also be of relevance. However, in this case the goal is not to fill missing data or large gaps in an incomplete observational series but to estimate an entire data series for a location or surface for which no direct observations are available, which makes several of the methods tested here not applicable. In particular, the correlation weighting method cannot be used, nor the bias correction methods developed here. This leaves us in the worst scenario, as shown by our results. Some gridded datasets such as CRUTEM4 ([Osborn and Jones 2014](#)) include a “variance corrected” version in which an adjustment has been made to account for temporal changes in the number of stations available for a given grid cell. The correction procedure, which is based in the number of stations and the correlation between them, is described in [Osborn et al. \(1997\)](#) and [Jones et al. \(2001\)](#). The success of this approach in preserving the variance and the long-term trend of the data would merit further analysis.

6. Conclusions

Between the different weighting methods tested, the following general conclusions can be highlighted:

- 1) Correlation weighting clearly outperformed inverse geographical distance. Inverse distance also had issues in coastal regions, which could also appear in other areas such as at political borders or climatic barriers, where the geographical proximity does not necessarily imply a climatic similarity.
- 2) Full standardization of the neighboring series before reconstruction should also be recommended over the more simple (and frequently used) differences method, as it allowed for correcting not only biases in the mean but also in the variance of the reconstructed series.

- 3) Bias correction was also very relevant, and only a two-step procedure with pre- (neighboring data series) and post- (reconstructed series) bias corrections was able to control the variance bias, albeit at the expense of slightly increasing the mean absolute error over applying only pre-bias correction. Using no bias correction resulted in clearly worse fit and bias statistics, and should not be recommended in any case.
- 4) When the analysis was segmented by decades, however, it was found that post-bias correction while using all the stations available tended to overestimate the variance of the data due to the presence of a temporal trend in the temperature data that was not accounted for in the post-bias correction.
- 5) Using only one (the best) neighbor series with a pre-bias correction yielded results that were almost as good as using all the neighbor series available and performing pre- and post-bias corrections, at least when the whole time period was considered. While using only one neighbor decreased the goodness of fit (increased the mean absolute error), on the other hand, it was good at preserving the variance of the resulting time series. That is, the trade-off between goodness of fit and variance bias is a compromise that needs to be made no matter the reconstruction method used, and it was made apparent again when checking the effect of the number of neighbors used in the reconstruction.
- 6) Not one of the methods evaluated was able to produce totally unbiased estimates of the long-term trend. In a context where most observed temperature series exhibited an increasing trend over the period of study, a tendency toward underestimating the magnitude of the trends was found for all reconstruction methods. Particularly, it was found that performing no bias correction at all underestimated the regional trend of the monthly maximum temperatures as much as $0.19^{\circ}\text{C decade}^{-1}$ (-43% over the observed trend of $0.273^{\circ}\text{C decade}^{-1}$), while the underestimation was of $0.045^{\circ}\text{C decade}^{-1}$ (-16.5%) for pre-bias correction, $0.020^{\circ}\text{C decade}^{-1}$ (-7.3%) for post-bias correction, and $0.054^{\circ}\text{C decade}^{-1}$ (-19.7%) in the case of using only one neighbor. For minimum temperature the underestimation was of $0.79^{\circ}\text{C decade}^{-1}$ (-34% over the observed trend of $0.230^{\circ}\text{C decade}^{-1}$) for no bias, $0.21^{\circ}\text{C decade}^{-1}$ (-9.1%) for pre-bias correction, $0.06^{\circ}\text{C decade}^{-1}$ ($+2.6\%$) for post-bias correction, and $0.18^{\circ}\text{C decade}^{-1}$ (-7.8%) for pre-bias correction and one neighbor. These differences may seem small, but they might be very relevant when long-time climatic changes are evaluated and may have important consequences in a context of climate change evaluation.

Acknowledgments. This work has been supported by the research projects CGL2014-52135-C3-1-R, CGL2014-52135-C3-3-R, CGL2017-83866-C3-1-R, CGL2017-83866-C3-3-R, and PCIN-2017-020, financed by the Spanish Commission of Science and Technology and EU ERDF, and INNOMED financed by the ERA-NET WaterWorks 2015 cofunded call of the European Commission, which is an integral part of the 2016 Joint Activities developed by the Water Challenges for a Changing World Joint Programme Initiative (Water JPI) as a result of a joint collaborative effort with the Joint Programming Initiative on Agriculture, Food Security and Climate Change (FACCE JPI), and funded by Agencia Estatal de Investigación of Spain, Research Promotion Foundation of Cyprus, Agence Nationale de la Recherche and Office national de l'eau et des milieux aquatiques of France, Ministry for Education, University and Research of Italy, Center of International Projects of Moldova, and Foundation for Science and Technology of Portugal with cofunding by the European Union. R.S.N. is funded by postdoctoral Grant FJCI-2017-31595 of the Juan de la Cierva Programme, funded by the Spanish Ministry of Science, Innovation and Universities, and EU ERDF. The authors declare that they do not have competing interests.

REFERENCES

- Aly, A., C. Pathak, R. S. V. Teegavarapu, J. Ahlquist, and H. Fuelberg, 2009: Evaluation of improvised spatial interpolation methods for infilling missing precipitation records. *World Environmental and Water Resources Congress*, ASCE, 1–10, [https://doi.org/10.1061/41036\(342\)598](https://doi.org/10.1061/41036(342)598).
- Beguiría, S., S. M. Vicente-Serrano, M. Tomás-Burguera, and M. Maneta, 2016: Bias in the variance of gridded data sets leads to misleading conclusions about changes in climate variability. *Int. J. Climatol.*, **36**, 3413–3422, <https://doi.org/10.1002/joc.4561>.
- Cannon, A. J., S. R. Sobie, and T. Q. Murdock, 2015: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *J. Climate*, **28**, 6938–6959, <https://doi.org/10.1175/JCLI-D-14-00754.1>.
- Cochran, W. G., 1977: *Sampling Techniques*. 3rd ed. John Wiley & Sons, 413 pp.
- Coulbaly, P., and N. Evora, 2007: Comparison of neural network methods for infilling missing daily weather records. *J. Hydrol.*, **341**, 27–41, <https://doi.org/10.1016/j.jhydrol.2007.04.020>.
- Diez-Sierra, J., and M. del Jesus, 2019: Subdaily rainfall estimation through daily rainfall downscaling using random forests in Spain. *Water*, **11**, 125, <https://doi.org/10.3390/w11010125>.
- Di Luzio, M., G. L. Johnson, C. Daly, J. K. Eischeid, and J. G. Arnold, 2008: Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States. *J. Appl. Meteor. Climatol.*, **47**, 475–497, <https://doi.org/10.1175/2007JAMC1356.1>.
- Di Piazza, A., F. L. Conti, L. V. Noto, F. Viola, and G. La Loggia, 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. *Int. J. Appl. Earth Obs. Geoinf.*, **13**, 396–408, <https://doi.org/10.1016/j.jag.2011.01.005>.

- Easterling, D. R., and T. C. Peterson, 1995: A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol.*, **15**, 369–377, <https://doi.org/10.1002/joc.3370150403>.
- Eccel, E., P. Cau, and R. Ranzi, 2012: Data reconstruction and homogenization for reducing uncertainties in high-resolution climate analysis in alpine regions. *Theor. Appl. Climatol.*, **110**, 345–358, <https://doi.org/10.1007/s00704-012-0624-z>.
- Eischeid, J. K., P. A. Pasteris, H. F. Diaz, M. S. P. Lantico, and N. J. Lott, 2000: Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteor.*, **39**, 1580–1591, [https://doi.org/10.1175/1520-0450\(2000\)039<1580:CASCND>2.0.CO;2](https://doi.org/10.1175/1520-0450(2000)039<1580:CASCND>2.0.CO;2).
- Fang, M., and X. Li, 2016: Application of Bayesian model averaging in the reconstruction of past climate change using PMIP3/CMIP5 multimodel ensemble simulations. *J. Climate*, **29**, 175–189, <https://doi.org/10.1175/JCLI-D-14-00752.1>.
- Feng, S., Q. Hu, Q. Wu, and M. E. Mann, 2013: A gridded reconstruction of warm season precipitation for Asia spanning the past half millennium. *J. Climate*, **26**, 2192–2204, <https://doi.org/10.1175/JCLI-D-12-00099.1>.
- Gonzalez-Hidalgo, J. C., D. Peña-Angulo, M. Brunetti, and N. Cortesi, 2015: MOTEDAS: A new monthly temperature database for mainland Spain and the trend in temperature (1951–2010). *Int. J. Climatol.*, **35**, 4444–4463, <https://doi.org/10.1002/joc.4298>.
- Guttman, N. B., 1999: Accepting the standardized precipitation index: A calculation algorithm 1. *J. Amer. Water Resour. Assoc.*, **35**, 311–322, <https://doi.org/10.1111/j.1752-1688.1999.tb03592.x>.
- Henn, B., M. S. Raleigh, A. Fisher, and J. D. Lundquist, 2013: A comparison of methods for filling gaps in hourly near-surface air temperature data. *J. Hydrometeorol.*, **14**, 929–945, <https://doi.org/10.1175/JHM-D-12-027.1>.
- Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick, 2001: Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environ. Modell. Software*, **16**, 309–330, [https://doi.org/10.1016/S1364-8152\(01\)00008-1](https://doi.org/10.1016/S1364-8152(01)00008-1).
- Jones, P., T. Osborn, K. Briffa, C. Folland, E. Horton, L. Alexander, D. Parker, and N. Rayner, 2001: Adjusting for sampling density in grid box land and ocean. *J. Geophys. Res.*, **106**, 3371–3380, <https://doi.org/10.1029/2000JD900564>.
- Kashani, M. H., and Y. Dinpashoh, 2012: Evaluation of efficiency of different estimation methods for missing climatological data. *Stochastic Environ. Res. Risk Assess.*, **26**, 59–71, <https://doi.org/10.1007/s00477-011-0536-y>.
- Kendall, M., Ed., 1975: *Rank Correlation Methods*. 4th ed. Charles Griffin, 199 pp.
- Kim, J.-W., and Y. A. Pachepsky, 2010: Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *J. Hydrol.*, **394**, 305–314, <https://doi.org/10.1016/j.jhydrol.2010.09.005>.
- Kuligowski, R. J., and A. P. Barros, 1998: Using artificial neural networks to estimate missing rainfall data 1. *J. Amer. Water Resour. Assoc.*, **34**, 1437–1447, <https://doi.org/10.1111/j.1752-1688.1998.tb05443.x>.
- Lee, H., and K. Kang, 2015: Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Adv. Meteor.*, **2015**, 935868, <https://doi.org/10.1155/2015/935868>.
- Li, X., Y. Zhou, G. R. Asrar, and Z. Zhu, 2018: Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. *Remote Sens. Environ.*, **215**, 74–84, <https://doi.org/10.1016/j.rse.2018.05.034>.
- Londhe, S., P. Dixit, S. Shah, and S. Narkhede, 2015: Infilling of missing daily rainfall records using artificial neural network. *ISH J. Hydraul. Eng.*, **21**, 255–264, <https://doi.org/10.1080/09715010.2015.1016126>.
- McKee, T. B., N. J. Doesken, and J. Kleist, 1993: The relationship of drought frequency and duration to time scales. *Proc. Eighth Conf. on Applied Climatology*, Anaheim, CA, Amer. Meteor. Soc., 179–183.
- Osborn, T., and P. Jones, 2014: The CRUTEM4 land-surface air temperature data set: Construction, previous versions and dissemination via Google Earth. *Earth Syst. Sci. Data*, **6**, 61–68, <https://doi.org/10.5194/essd-6-61-2014>.
- , K. Briffa, and P. Jones, 1997: Adjusting variance for sample size in tree-ring chronologies and other regional mean timeseries. *Dendrochronologia*, **15**, 89–99.
- Peterson, T. C., and Coauthors, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol.*, **18**, 1493–1517, [https://doi.org/10.1002/\(SICI\)1097-0088\(19981115\)18:13<1493::AID-JOC329>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13<1493::AID-JOC329>3.0.CO;2-T).
- Ramos-Calzado, P., J. Gómez-Camacho, F. Pérez-Bernal, and M. Pita-López, 2008: A novel approach to precipitation series completion in climatological datasets: Application to Andalusia. *Int. J. Climatol.*, **28**, 1525–1534, <https://doi.org/10.1002/joc.1657>.
- R Core Team, 2017: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>.
- Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu, 2007: A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteor. Climatol.*, **46**, 900–915, <https://doi.org/10.1175/JAM2493.1>.
- Richardson, C. W., 1981: Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour. Res.*, **17**, 182–190, <https://doi.org/10.1029/WR017i001p00182>.
- Sen, P. K., 1968: Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.*, **63**, 1379–1389, <https://doi.org/10.1080/01621459.1968.10480934>.
- Serrano-Notivoli, R., S. Beguería, M. Á. Saz, L. A. Longares, and M. de Luis, 2017a: SPREAD: A high-resolution daily gridded precipitation dataset for Spain—An extreme events frequency and intensity overview. *Earth Syst. Sci. Data*, **9**, 721–738, <https://doi.org/10.5194/essd-9-721-2017>.
- , M. de Luis, M. Saz, and S. Beguería, 2017b: Spatially based reconstruction of daily precipitation instrumental data series. *Climate Res.*, **73**, 167–186, <https://doi.org/10.3354/cr01476>.
- Simolo, C., M. Brunetti, M. Maugeri, and T. Nanni, 2010: Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *Int. J. Climatol.*, **30**, 1564–1576, <https://doi.org/10.1002/JOC.1992>.
- Singh, V., and Q. Xiaosheng, 2019: Data assimilation for constructing long-term gridded daily rainfall time series over Southeast Asia. *Climate Dyn.*, **53**, 3289–3313, <https://doi.org/10.1007/S00382-019-04703-6>.
- Stooksbury, D. E., C. D. Idso, and K. G. Hubbard, 1999: The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States: A spatial and temporal study. *J. Climate*, **12**, 1524–1533, [https://doi.org/10.1175/1520-0442\(1999\)012<1524:TEODGO>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<1524:TEODGO>2.0.CO;2).
- Tardivo, G., 2015: Spatial and time correlation of thermometers and pluviometers in a weather network database. *Theor. Appl. Climatol.*, **120**, 19–28, <https://doi.org/10.1007/s00704-014-1148-5>.
- , and A. Berti, 2012: A dynamic method for gap filling in daily temperature datasets. *J. Appl. Meteor. Climatol.*, **51**, 1079–1086, <https://doi.org/10.1175/JAMC-D-11-0117.1>.

- , and —, 2014: The selection of predictors in a regression-based method for gap filling in daily temperature datasets. *Int. J. Climatol.*, **34**, 1311–1317, <https://doi.org/10.1002/joc.3766>.
- Teegavarapu, R. S., 2014: Statistical corrections of spatially interpolated missing precipitation data estimates. *Hydrol. Processes*, **28**, 3789–3808, <https://doi.org/10.1002/hyp.9906>.
- , and V. Chandramouli, 2005: Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.*, **312**, 191–206, <https://doi.org/10.1016/j.jhydrol.2005.02.015>.
- , and A. Nayak, 2017: Evaluation of long-term trends in extreme precipitation: Implications of in-filled historical data use for analysis. *J. Hydrol.*, **550**, 616–634, <https://doi.org/10.1016/j.jhydrol.2017.05.030>.
- , M. Tufail, and L. Ormsbee, 2009: Optimal functional forms for estimation of missing precipitation data. *J. Hydrol.*, **374**, 106–115, <https://doi.org/10.1016/j.jhydrol.2009.06.014>.
- Theil, H., 1950: A rank-invariant method of linear and polynomial regression analysis, 3; confidence regions for the parameters of polynomial regression equations. *Indag. Math.*, **1**, 467–482, https://doi.org/10.1007/978-94-011-2546-8_20.
- Venema, V. K., and Coauthors, 2013: Benchmarking homogenization algorithms for monthly data. *AIP Conference Proceedings*, Vol. 1552, AIP, 1060–1065.
- Vicente-Serrano, S. M., S. Beguería, J. I. López-Moreno, M. A. García-Vera, and P. Stepanek, 2010: A complete daily precipitation database for northeast Spain: Reconstruction, quality control, and homogeneity. *Int. J. Climatol.*, **30**, 1146–1163, <https://doi.org/10.1002/joc.1850>.
- Woldesenbet, T. A., N. A. Elagib, L. Ribbe, and J. Heinrich, 2017: Gap filling and homogenization of climatological datasets in the headwater region of the upper Blue Nile basin, Ethiopia. *Int. J. Climatol.*, **37**, 2122–2140, <https://doi.org/10.1002/joc.4839>.
- Xia, Y., P. Fabian, A. Stohl, and M. Winterhalter, 1999: Forest climatology: Estimation of missing values for Bavaria, Germany. *Agric. For. Meteorol.*, **96**, 131–144, [https://doi.org/10.1016/S0168-1923\(99\)00056-8](https://doi.org/10.1016/S0168-1923(99)00056-8).
- You, J., K. G. Hubbard, and S. Goddard, 2008: Comparison of methods for spatially estimating station temperatures in a quality control system. *Int. J. Climatol.*, **28**, 777–787, <https://doi.org/10.1002/joc.1571>.
- Young, K. C., 1992: A three-way model for interpolating for monthly precipitation values. *Mon. Wea. Rev.*, **120**, 2561–2569, [https://doi.org/10.1175/1520-0493\(1992\)120<2561:ATWMFI>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<2561:ATWMFI>2.0.CO;2).
- Yozgatligil, C., S. Aslan, C. Iyigun, and I. Batmaz, 2013: Comparison of missing value imputation methods in time series: The case of Turkish meteorological data. *Theor. Appl. Climatol.*, **112**, 143–167, <https://doi.org/10.1007/S00704-012-0723-X>.
- Zhang, X., X. Yan, and Z. Chen, 2016: Reconstructed regional mean climate with Bayesian model averaging: A case study for temperature reconstruction in the Yunnan–Guizhou Plateau, China. *J. Climate*, **29**, 5355–5361, <https://doi.org/10.1175/JCLI-D-15-0603.1>.
- Zhou, H., E. Aizen, and V. Aizen, 2018: Constructing a long-term monthly climate data set in central Asia. *Int. J. Climatol.*, **38**, 1463–1475, <https://doi.org/10.1002/JOC.5259>.