Theoretical and Applied Climatology
https://doi.org/10.1007/s00704-024-05079-2

RESEARCH

A methodological approach for filling the gap in extreme daily temperature data: an application in the Calabria region (Southern Italy)

Emanuele Barca[1] · Ilaria Guagliardi[2] · Tommaso Caloiero[3]

Received: 26 February 2024 / Accepted: 17 June 2024

Abstract

Regional studies are crucial for monitoring and managing the impacts of extreme climatic events. This phenomenon is particularly important in some areas, such as the Mediterranean region, which has been identified as one of the most responsive regions to climate change. In this regard, the analysis of large space-time sets of climatic data can provide potentially valuable information, although the datasets are commonly affected by the issue of missing data. This approach can significantly reduce the reliability of inferences derived from space-time data analysis. Consequently, the selection of an effective missing data recovery method is crucial since a poor dataset reconstruction could lead to misleading the decision makers' judgments. In the present paper, a methodology that can enhance the confidence of the statistical analysis performed on the reconstructed data is presented. The basic assumption of the proposed methodology is that missing data within certain percentages cannot significantly change the shape or parameters of the complete data distribution. Therefore, by applying several missing data recovery methods whose reconstructed dataset better overlaps the original dataset, larger confidence is needed. After the gap filling procedure, the temporal tendencies of the annual daily minimum temperature (T < 0 °C) were analysed in the Calabria region (southern Italy) by applying a test for trend detection to 8 temperature series over a 30-year period (1990–2019). The results showed that there was a constant reduction in the duration of frosty days, indicating the reliability of the effect of climate change.

1 Introduction

The Earth's climate has frequently changed over the past 4.5 billion years, with significant fluctuations in global average temperatures due to volcanic emissions, tectonic plate movements, changes in solar radiation, and several other factors (Caloiero et al. 2021; Lin and Qian 2022). However, since the last ice age, the Earth's climate has been relatively stable, with global temperatures varying by less than 1 °C over a century in the last 10,000 years. Over the last 100 years, human activities have started to play a major role in increasing temperature levels, with alarming acceleration over the last twenty years (Schmidt and Hertzberg 2011). The sixth report (AR6) by the Intergovernmental Panel on Climate Change (IPCC 2013) evidenced an increase in the Earth's surface temperature in the twentieth century that was around 1.1 °C above 1850–1900 in 2011–2020, with larger increases over land than over the ocean. An increase in temperature leads to greater evapotranspiration demand, accelerated hydrological cycle, an intensification, and an increase in extreme events, such as droughts, floods, cold spells and heat waves (Buttafuoco et al. 2016, 2018; Sirangelo et al. 2020). Several studies performed worldwide have shown an increase in the mean temperature. For example, in Eastern China, an increase of 1.52 °C over the past century has been detected (Zhao et al. 2014). Similarly, several increases in temperature have been detected in Switzerland (Ceppi et al. 2012), Pakistan (Nawaz et al. 2019), the United States, except for Pennsylvania and Maine (Martinez et al. 2012),

✉ Ilaria Guagliardi
ilaria.guagliardi@isafom.cnr.it

1 National Research Council of Italy, Water Research Institute (CNR-IRSA), Bari, Italy

2 Institute for Agriculture and Forest Systems in the Mediterranean (CNR-ISAFOM), National Research Council of Italy, Rende (CS), Italy

3 National Research Council of Italy, Research Institute for Geo-hydrological Protection (CNR-IRPI), Cosenza, Italy

Published online: 29 June 2024

and Senegal (Djaman et al. 2017). Several studies have also been carried out in the Mediterranean basin, a region highly susceptible to climate change since geographically, this region lies in a transition area between the hot, dry climate of Africa and the cold, humid air masses arriving from northern Europe (Goubanova and Li 2007; Vicente-Serrano et al. 2014). For instance, Giorgi (2002), at the annual scale, and Jacobeit (2000), in summer, demonstrated statistically significant warming trends over most parts of the Mediterranean area. Several authors have also detected a general increase in temperature in the Mediterranean basin over recent decades (Caloiero and Guagliardi 2020, 2021; Caloiero et al. 2016; Pellicone et al. 2019; Scorzini and Leopardi 2019; Gentilucci et al. 2020).

In these studies, climate change has been analysed considering variations in average temperature, but the frequency and severity of extreme events cause most of the social and economic costs linked with climate change (King et al. 2015). In fact, changes in climate extremes lead to more substantial impacts on human and natural systems than changes in the average climate (Infusino et al. 2022a; Ricca and Guagliardi 2015). For this reason, recently particular attention has been given worldwide to the analysis of extreme temperatures at different spatial scales, revealing differences in the characteristics of extreme temperatures, with marked variations in some areas and nonsignificant trends in others (Bonsal et al. 2001). Therefore, performing regional studies is crucial for evaluating the impacts of extreme events (Zarenistanak et al. 2014), even though this approach is not always possible due to the lack of complete and spatially coherent daily temperature series (Mishra and Singh 2011). Consequently, there is an absence of regional-scale studies on extreme changes in temperature (Murara et al. 2019).

Missing data are a common issue in practical data analysis, particularly for time series. In fact, it is well-known that for performing correct statistical analyses, working datasets must be complete. To address this issue effectively, ad hoc theoretical frameworks have been introduced to reconstruct incomplete datasets (Lo Presti et al. 2010). Many techniques for missing data reconstruction have been derived from these frameworks. With respect to temperature data, gap-filling methods can generally be separated into spatial and temporal groups. The methods of the first group are usually based on interpolation techniques such as inverse-distance weighting, kriging, multiple regressions, and thin-plate splines (Lompar et al. 2019). On the other hand, the methods used for the latter group depend on the autocorrelation of the meteorological time series (Lompar et al. 2019). For example, Claridge and Chen (2006) applied polynomial fitting and simple linear interpolation, while Liston and Elder (2006) reconstructed missing data using a linear combination of forecasts and hindcasts. Finally, there are several methods, such as empirical orthogonal functions that combine spatial and temporal interpolation (Von Storch and Zwiers 1999), even though the application of these methods is strictly related to the number of neighboring stations and the size of the data gap. Generally, older methods are single-valued, but more recently, multiple-valued methods have been introduced; consequently, multiple imputation methods have become available that are considered to be better than single-valued methods (Scheffer 2002). Another issue concerns how to perform an effective selection of the most appropriate method of missing data reconstruction. A relevant consequence of Wolpert's theorem is that no method can outperform all others across all possible datasets (Wolpert 1996).

Within this context, the aim of this study is to propose a multiple-valued methodology aimed at providing the best missing data reconstruction with respect to the case study and the selected methods. The gap-filling database was subsequently used to perform a reliable trend analysis of extreme temperature data in the Calabria region in southern Italy.

2 Study area and data

Due to its position in the center of the Mediterranean basin and its climatic characteristics, Calabria is considered a highly susceptible area to climate change (Guagliardi et al. 2021), where even a small temperature increase could lead to various environmental problems. The Calabria region is located between 37° 54' and 40° 09' N and between 15° 37' and 17° 13' E, with an area of 15,080 km² (Caroletti et al. 2019; Iovine et al. 2018). Although the region has a long north-south orientation and does not have many high peaks, it is one of the most mountainous Italian regions (Gaglioti et al. 2019; Infusino et al. 2022b), with mountains occupying 42% of the region and hills covering 49% of the territory (Fig. 1).

The Köppen-Geiger classification (Köppen 1936) identifies the climate of the region as a hot-summer Mediterranean climate, with relatively mild winters (with rain) and very hot summers (often very dry). With this climate, the coldest month generally has an average temperature above 0 °C, at least one month's average temperature reaches values higher than 22 °C, and at least four months present an average temperature above 10 °C. Moreover, due to its orography, the region exhibits sharp contrasts, with warm air currents coming from Africa affecting the Ionian side with short and heavy precipitation and western air currents affecting the Tyrrhenian side with high precipitation (Caroletti et al. 2019; Pellicone et al. 2018).

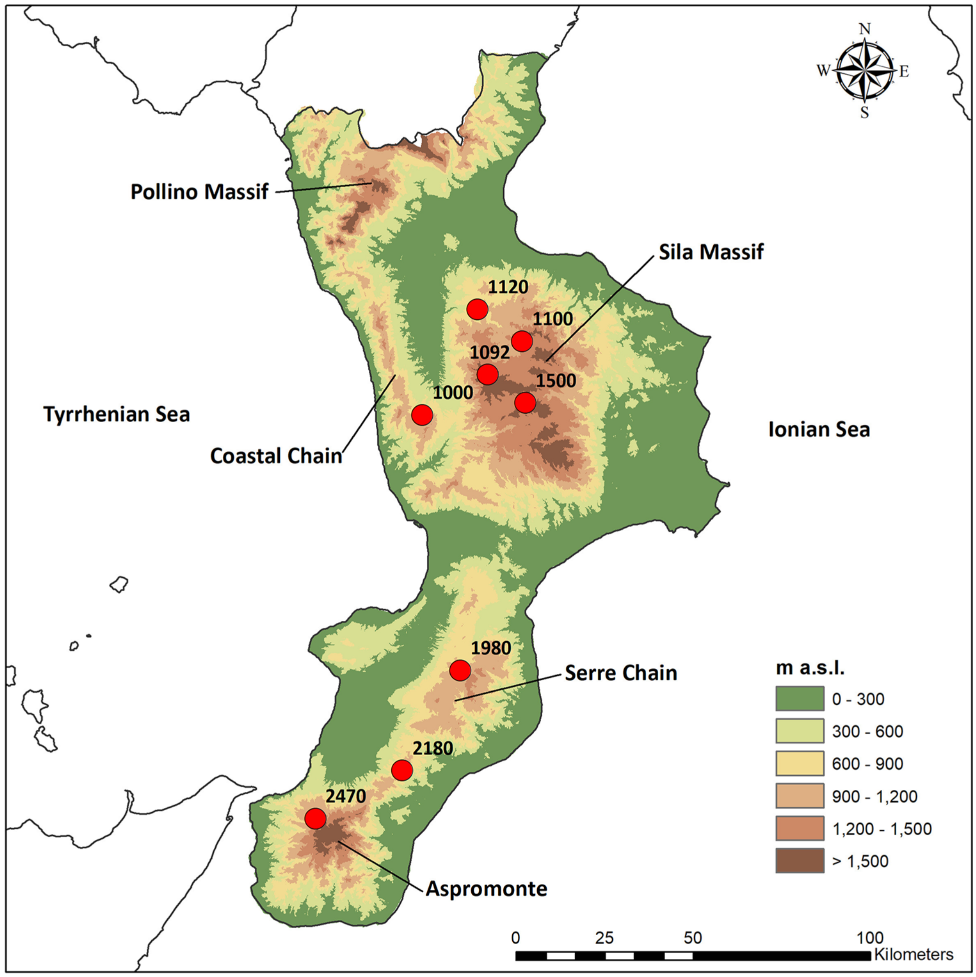

Fig. 1 Study area with selected temperature stations (the number close to the red dots corresponds to the code of the climatic series)

Given the findings of past studies showing that winter conditions are changing more rapidly than they are in any other season (Caloiero et al. 2017), in this study, particular attention was given to the temporal changes in frost days, i.e., the annual daily minimum temperature < 0 °C (Zhang et al. 2011). In fact, a decrease in frost days could have lasting impacts on ecosystems, especially in mountainous and forested areas such as the Calabria region. Indeed, frost days are linked to the forest health, by killing insect pests, and to the snowpack, which insulates soil providing subnivean refugia for some animals and preventing the freezing of roots and microbes (Contosta et al. 2019).

In Calabria, climatic data are managed by the Multi-Risk Functional Centre of the Regional Agency for Environmental Protection which publishes online daily temperature data.

The starting dataset was formed by daily minimum temperature related to 72 monitoring climatic gauges and covering an observation period ranging from 01/01/1990 to 31/12/2019, i.e. 30 years or 10957 days. The dataset was first analyzed to assess the missing data rate for each monitoring site, to check if there were time series with a missing data percentage too high to be reliably reconstructed. Percentage of missing data among sites ranged from 0.23 to 41.73% and, on average, the number of missing data rate was 8.4%. All the sites with a missing data rate larger than 10% have been dropped, consequently a reduced set of 28 sites have been retained for the analysis and, considering the elevation, 8 temperature series have been investigated with a remaining set of supporting sites of 20. These 8 series are well distributed in the region and fall within 4 of 5 of the main mountains of the region, the Sila Massif, the Coastal Chain, the Aspromonte Massif, the Serre Massif. In fact, in the Pollino Massif, no thermometric stations are available (Fig. 1; Table 1).

Following Caloiero et al. (2017), the 8 selected temperature series were checked for errors and inhomogeneity by means of a multiple application of the Craddock test (Craddock 1979). The homogenization procedure was applied to minimum series and, due to the low number of temperature stations available for this analysis, each series was tested against other ones, in subgroups of five series.

Table 1 Main characteristics of the target stations selected for this study

ID	Name	Elevation [m a.s.l.]	Location
1000	Domanico	736	Coastal Chain
1092	Camigliatello - Monte Curcio	1730	Sila Massif
1100	Cecita	1180	Sila Massif
1120	Acri	790	Sila Massif
1500	Nocelle - Arvo	1315	Sila Massif
1980	Serra San Bruno	790	Serre Chain
2180	Antonimina - Canolo Nuovo	880	Serre Chain
2470	Gambarie d'Aspromonte	1200	Aspromonte

3 Materials and methods

In this section, a new methodology aimed at providing the best missing data reconstruction with respect to the case study and the selected methods is presented (Fig. 2).

Following the linear regression approach, proposed by FAO (Fagandini et al. 2024), the series to be filled must have at least 70% of complete data; thus, the methodology assumes that the missing data cannot significantly change the shape or parameters of the complete data distribution if they are below a certain percentage (< 30%). Besides detecting the percentage of missing data, it is crucial to determine whether the missing data is isolated or chained together in a sequence of consecutive days without data. Therefore, the distribution shape and the characteristic parameters (mean and standard deviation) of the incomplete dataset can be considered within a certain approximation, the real distribution, and the real parameters. After applying several reconstruction methodologies, the distributions and parameters of the reconstructed datasets are compared with those of the original incomplete dataset. The reconstructed dataset that better overlapped the original dataset was considered the best. The core of the methodology is the application of several missing data recovery methods to an incomplete dataset, at least three and based on different assumptions, to account for all the specific features contained in the data. The need for using more than one method for data recovery derives from Wolpert's theorem (Wolpert 1996), which states the inexistence of a universally optimal machine learning method in absolute terms but only regarding to a

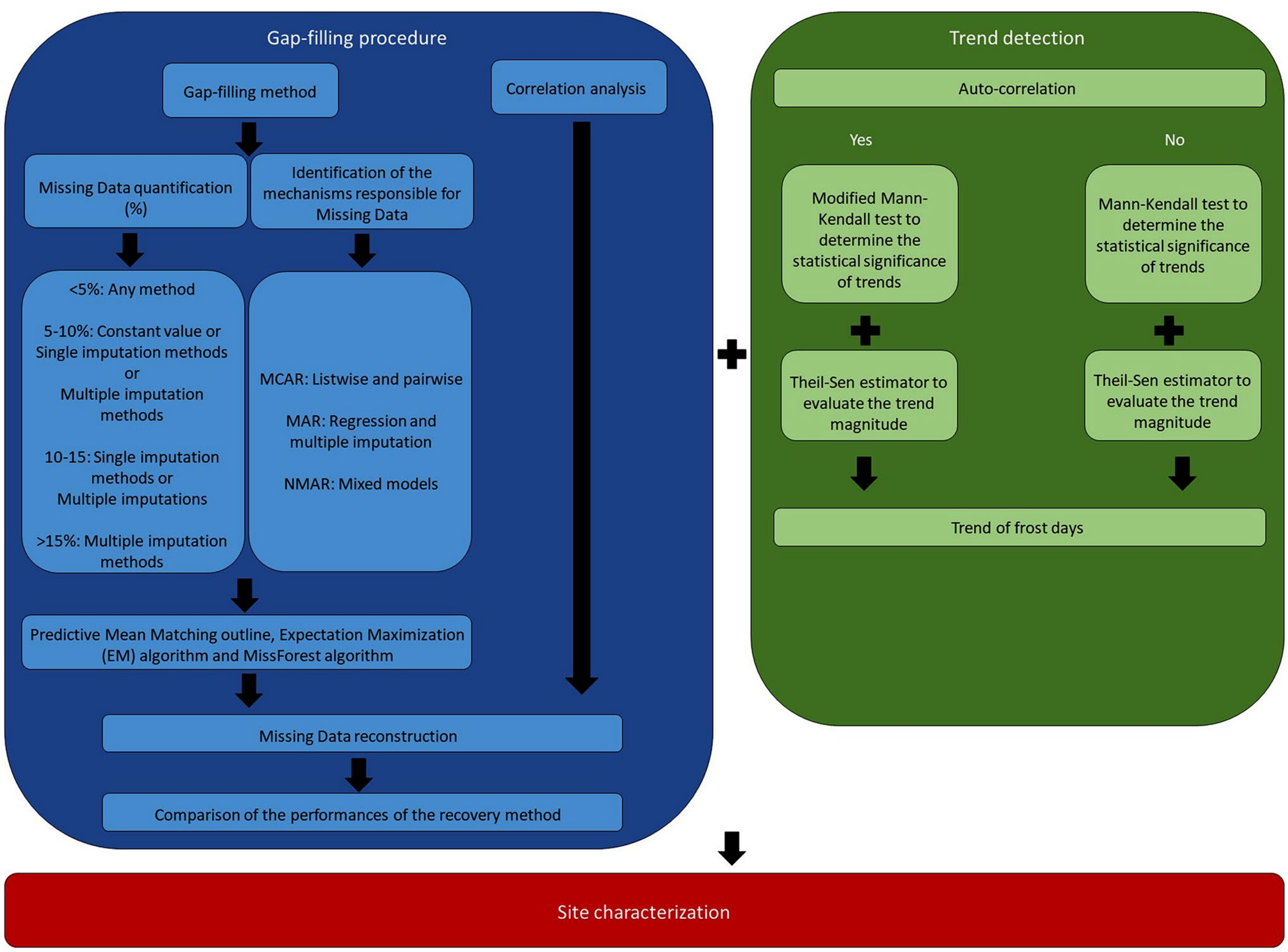

Fig. 2 Flowchart of the methodology

specific problem. After running the selected methods of data recovery, three (or more) different completed datasets are obtained. The analysis of the data loss mechanism is then carried out to check whether missing data can be classified as *ignorable*. In fact, the assumption of *ignorability* is essentially the belief that the available data are sufficient to estimate missing data; thus, this assumption is the basis of the missing data reconstruction concept.

After the missing data reconstruction, a trend analysis was performed to evaluate the possible trend in the gap-filled data.

3.1 Preliminary statistical analysis

First, descriptive statistics were computed to summarize the main features of the minimum temperature distribution. In addition, the hypothesis of normality was tested using the Anderson-Darling test (Barca et al. 2016, 2017). The descriptive statistics were generated by applying the R library *summarytools* and the related function *descr* (Comtois 2021).

3.2 Gap-filling procedure

The analysis and reconstruction of an incomplete dataset is a complex task, and to carry out this task correctly, a suitable methodology is needed. The methodology described hereafter is a multistep one.

The first step is preprocessing the incomplete dataset to obtain descriptive statistics useful for making correct decisions during the analysis. First, the quantification (in absolute and percentage terms) of the missing data (MD) for each incomplete station must be accomplished. Several numerical and graphical tools are currently available for performing such tasks. The aim of this analysis is manifold: (i) assess the percentages of MD to select the best reconstruction strategy (Table 2); (ii) assess wheter all the MDs are in principle recoverable; and (iii) recognize possible patterns among MDs by means of graphical tools, such as Margin Plots (Templ et al. 2012).

The second step is the correlation analysis, which helps in collecting all the time series (TS) related to sites that are better correlated with the target site. The TSs are the basic information of all the missing data recovery methods. Another important methodological step concerns the identification of the mechanisms responsible for data value losses. Once the support TSs for the considered target site are gathered and the loss mechanisms are identified, the next step is to select a group of suitable missing data recovery methods. As mentioned before, more than one methodology is recommended because a method that outperforms all the others for any class of data does not exist.

After applying the selected missing data recovery methods, the performances of each method must be compared to safely choose the best data reconstruction for the analyzing case. At this step, the probability density function (pdf) associated with the incomplete target site TS is considered and compared with the pdfs of the completed TSs provided by the applied recovery data methods. The comparison is carried out by superimposing the pdf associated with the incomplete TS with each of the densities of the completed datasets and computing the percentage of overlap. For all the completed datasets, the dataset that overcomes 90% of the overlap, and has the least deviation from the characteristic parameters extracted from the incomplete original data is selected as the optimal solution.

Table 2 Suggested methods considering the missing data rate (MD)

MD%	Suggested methods
< 5%	Any method
5–10%	Constant value, e.g. sample mean, is good only if the correlation among the variables is negligible Single imputation methods, e.g. regression Multiple imputation methods (best choice)
10–15%	Single imputation methods Multiple imputations strongly recommended
> 15%	Multiple imputation methods

3.2.1 Preprocessing and correlation analysis

In this study, space-time continuous data, i.e., time series data from different monitoring stations belonging to a permanent observational network, are analyzed. Generally, that kind of data is arranged as a bidimensional array endowed with the following structure:

- each row is uniquely associated with the date on which the observation was made according to the considered time aggregation level (daily, monthly, or yearly),
- each column is uniquely associated with a monitoring site.

Rows are often referred to as "items", and columns are often referred to as "sites".

As reported in Table 1, the missing data rate is the key factor for selecting the best method for rebuilding incomplete time series (Lo Presti et al. 2010). It is straightforward that not all incomplete datasets can be filled even if very powerful methods are currently available. In fact, some missing data percentage thresholds discourage any attempt to fill these gaps. Therefore, another important step is to drop all the columns of the matrix (i.e. the sites) affected by high missing data rates (> 30%) that could be severely distorted by reconstruction methods. Afterwards, the remaining columns are split into two groups: (i) the group containing the sites to be reconstructed (target sites) and (ii) the support site group, which provides the information needed to accomplish the reconstruction.

Once the target sites are recognized, a correlation analysis is performed to separate the columns (sites) from the support group that are not well correlated with the first group of sites.

3.2.2 Missing data method assessment

The subsequent methodological step concerns the identification of the mechanism responsible for data value losses. When the phenomenon of missing data is described in probabilistic terms, three different types of processes can be considered (Table 3): missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little and Rubin 1987; Lo Presti et al. 2010; Leurent et al. 2013). The substantial difference between these types of loss mechanisms relies on the degree of dependence between the probability of the data loss and the data value.

A comprehensive description of the theoretical framework can be found in Rubin (1976). If the loss mechanism is too difficult to assess due to the incompleteness of the supporting group sites, the usual assumption is MAR, and the consequent approach is multiple imputation. This approach is different from the others (e.g. uni-multivariate regression) in that it does not provide a single replacement for each missing entry in the dataset. In contrast, if a small distribution of realistic data is provided, afterwards, the user is burdened to select among the distributions the best suited replacement for the missing datum.

Once the features that the reconstruction approach should possess according to the percentage of missing data are identified, some algorithms for fixing the problem can be selected (see Table 1). The basic idea of the proposed methodology is to apply more than one algorithm, each endowed with different predictive characteristics (parametric, non-parametric, etc.).

3.2.3 Missing data reconstruction: predictive mean matching (PMM) outline

Predictive Mean Matching (PMM) is an algorithm used to perform multiple imputations for missing data; it is particularly useful for quantitative variables and has the advantage of being non-parametric (Rubin 1986; Little 1988). Such an approach to multiple imputations is known as multiple imputations by chained equations (MICE), sequential generalized regression, or the fully conditional specification (FCS) (Van Buuren 2007).

The algorithm follows this line: suppose there is a single variable x that has some cases with missing data and a set of variables z_i that are used to impute x. The steps of the PMM algorithm are as follows:

1. for cases with no missing data, estimating a linear regression of x on z_i, producing a set of coefficients b_i;
2. to perform a random extraction from the "posterior predictive distribution" of b, producing a new set of coefficients $b*$. This step is necessary to produce sufficient variability in the imputed values and is common to all methods for multiple imputations;
3. using $b*$, generate predicted values for x for both cases with missing data and those with no missing data;
4. for each case with missing x, identifying a set of cases with observed x whose true values are close to the predicted value for the case with missing data;
5. among those close cases, randomly choose one and assign its observed value to substitute for the missing value;
6. repeating steps 2 through 5 for each completed dataset.

The R library *mice* and function *mice* (Van Buuren 2007) was used for missing data estimation.

3.2.4 Missing data reconstruction: the expectation maximization (EM) algorithm

The package *norm* implements missing data procedures via the Multivariate Expectation-Maximization (EM) method.

Table 3 Formal definition of the data loss mechanisms (by Lo Presti et al. 2010)

Type	Definition	Suggested in-filling methods
MCAR	$p\,(M\vert\, Y,\varphi) = p(M\,\vert\varphi)\;forall Y,\varphi$	Listwise and pairwise
MAR	$p\,(M\vert\, Y,\varphi) = p(M\vert Y_{obs},\varphi) forall Y_{miss},\varphi$	Regression and multiple imputation
NMAR	$p\,(M\vert\, Y,\varphi) = p(M\vert Y_{miss},\varphi)$	Mixed models

p probability function, M missingness, φ probability function parameter set, matrix of all the potentially observed values $Y = Y_{obs} + Y_{miss}$; Y_{obs} *really observed data value matrix*, Y_{miss} missing data value matrix

The EM algorithm formalizes an elementary idea for dealing with missing data and consists of the following steps:

1. replacing missing values with estimated values (e.g. by averaging observations);
2. estimating the prediction model parameters;
3. re-estimating the missing data, assuming that the new estimated parameters are correct;
4. the parameters are re-estimated, and the procedure is repeated until convergence.

Therefore, the EM algorithm consists of two main stages, namely the Expectation and the Maximization, which are repeated until the missing data estimation changes significantly; otherwise, the process stops.

The EM algorithm is a parametric method and can be applied to any type of data, but it is particularly effective for distributions of the exponential family. The main advantage of using the EM algorithm is that it deterministically converges because it is simple to define pre-specified bounds that can be used to terminate the algorithm (King et al. 2001; Honaker and King 2010).

3.2.5 Missing data reconstruction: the missforest algorithm

MissForest is a random forest imputation algorithm for missing data, implemented in R in the *missForest* (Stekhoven and Buhlmann 2012) R package. Initially, all missing data were imputed using the mean/mode; subsequently, for each variable with missing values, MissForest was used to fit a random forest to the observed data and predict the missing data. This process of training and predicting repeats an iterative process until a stopping criterion is met, or a maximum number of user-specified iterations is reached. The reason for the multiple iterations is that, from iteration 2 onward, the random forests performing the imputation will be trained on better- and better-quality data that itself have been predictively imputed. The MissForest algorithm has an important drawback tied to the heavy computational cost that can be partially overcome by using the function in a parallel framework.

3.3 Time series serial correlation

To assess the significance of a time series trend, after data reconstruction, a test of serial correlation is needed. In fact, autocorrelation can adversely affect a trend test, lowering the significance threshold α and distorting the trend test outcomes. The serial correlation test is a graphical tool that is easily interpreted. The first quadrant of the Cartesian plane is shown, crossed by two dashed lines. From each lag, a vertical line is raised at lag 0, a line is generated that terminates at the 1.0 ordinate. This line is of little use because it indicates that the time series is correlated with itself. From a lag value of 1 on, if the vertical line overcomes the dashed line, the result can be considered significant. As an example, Fig. 3 shows that the line corresponding to lag 1 crosses the dashed line, which means that the time series is autoregressive of the first order (AR(1)) for the Nocelle-Arvo site. The functions *acf* and *pacf* of the R library stats were used to perform the time autocorrelation test.

3.4 Trend tests

The Mann-Kendall (MK) test is a nonparametric method for detecting trends in a time series. This test is widely used in environmental science because it is simple and robust. After the first versions were proposed by Mann (1945) and Kendall (1975), the test has been improved in many ways by introducing covariance matrix analysis (Dietz and Kileen 1981), filtering seasonality (Hirsch et al. 1982), multiple monitoring sites (Lettenmaier 1988) and covariates representing natural fluctuations (Libiseller and Grimvall 2002). If a serial correlation in the data is detected, a trend test robust against its adverse effects should be applied. The Modified Mann-Kendall (MMK) test is an implementation of the MK test that is free from serial correlation effects (Kundzewicz and Robson 2000). The R library used was *Kendall* and the corresponding function was *MannKendall* (Hipel and McLeod 1994).

In addition to the MK test, to evaluate the trend magnitude, the Theil-Sen estimator (TSE) was applied. In fact, this test is considered more powerful than linear regression methods in trend magnitude evaluation, because it is not subject to the influence of extreme values (Sen 1968).

4 Results

4.1 Basic statistics

Table 4 shows the basic statistics related to the eight target sites described above. The considered variable is the daily minimum temperature (°C).

Among the different statistics, the analysis of skewness, which yields values close to zero, is particularly important. This means that the empirical distributions of the target sites are all approximately Gaussian. This is a key feature because it guarantees the correct operation of parametric tests and applied methods (e.g. correlation). Unfortunately, this finding was not confirmed after the application of the Anderson-Darling Gaussianity test (Thode 2002; Barca et al. 2016), which rejects the Gaussianity hypothesis for all

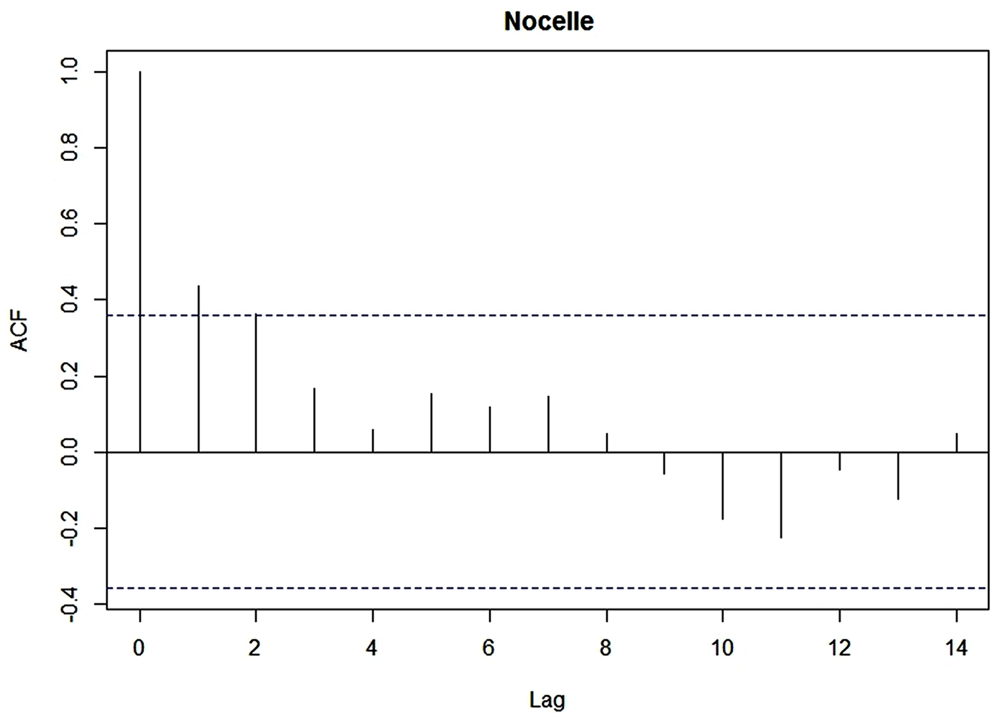

Fig. 3 Serial correlation test for the Nocelle-Arvo site (ID 1500). The time series appears to be AR(1)

Table 4 Basic statistics for the 8 target sites (the analysed period is 1990–2019 and the variable is daily minimum temperature)

Target Sites	*N* [count]	Mean [°C]	Sd [°C]	Absolute Min [°C]	Absolute Max [°C]	Skewness [dimensionless]	Kurtosis [dimensionless]
1000	9975	8.99	6.09	-9.1	28.1	0.03	-0.62
1092	10,110	3.71	6.54	-16.7	22.6	-0.05	-0.61
1100	10,154	3.85	5.93	-17	18.4	-0.17	-0.66
1120	10,553	8.99	6.51	-9	27.4	0.1	-0.78
1500	10,871	3.38	5.87	-24.6	17.9	-0.3	-0.45
1980	10,664	5.04	5.24	-13.2	21.1	-0.18	-0.64
2180	10,442	8.51	5.79	-8.6	26.4	0.04	-0.81
2470	10,659	7.12	5.91	-9	24.8	0.09	-0.83

Table 5 Values of quantile imbalance related to the target stations

	1000	1092	1100	1120	1500	1980	2180	2470
q_i [°C]	0.022	0.020	0.0078	0.030	0.0091	-0.027	0.046	0.019

the empirical distributions related to the different TSs. This is probably due to the large size of the TS associated with each target size. Table 5 reports the values of the *quantile imbalance* (q_i) that, if close to zero, assures that at least the central part of the distribution is approximately Gaussian (Brunsdon et al. 2002).

The obtained values confirm that the hypothesis can be tested safely, and thus, parametric tests can be safely carried out (e.g. Pearson correlation).

4.2 Missing data and data reconstruction: target and supporting sets

The eight target sites selected for the analysis and their percentages of missing data are reported in Fig. 4.

Figure 4 provides a specific visualization of the amount of missing data, with the location of missing values shown in black and providing information on the overall percentage of missing values. A careful analysis of the figure reveals that all the missing data are potentially recoverable. In fact, there is no black line crossing the entire gray area, and the missing data percentages are within the range (Table 1) to be reliably reconstructed by means of multiple imputation methods. It is evident that the missing data are evenly

Fig. 4 Missing data percentages

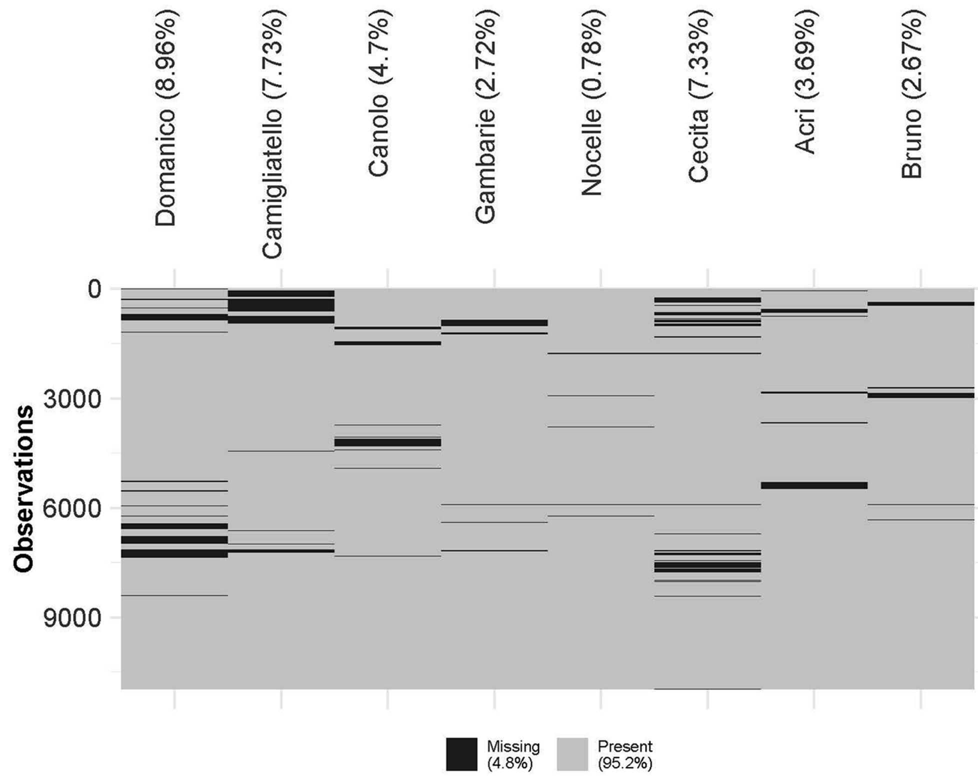

Table 6 Percentages of overlap between the original data density and the reconstructed data density according to the applied methods

NORM								
	1000	1092	2180	2470	1500	1100	1120	1980
Overlap	0.975	0.992	0.993	0.982	0.996	0.987	0.985	0.985
Deviation from the average [%]	1.73	0.45	0	1.28	0.46	0.57	0.59	1.66
Deviation from the std deviation [%]	0.57	0.45	0.12	0.3	0.19	0.52	0.09	0.37
MISS_FOREST								
	1000	**1092**	**2180**	**2470**	**1500**	**1100**	**1120**	**1980**
Overlap	0.983	0.988	0.989	0.986	0.984	0.984	0.982	0.985
Deviation from the average [%]	1.397	0.196	0.195	-1.322	0.252	-0.568	-0.607	1.585
Deviation from the std deviation [%]	-0.370	0.677	0.052	0.314	0.210	0.614	0.118	0.069
MICE								
	1000	**1092**	**2180**	**2470**	**1500**	**1100**	**1120**	**1980**
Overlap	0.984	0.990	0.990	0.986	0.984	0.983	0.982	0.985
Deviation from the average [%]	1.383	0.347	0.263	-1.400	0.276	-0.466	-0.595	1.436
Deviation from the std deviation [%]	-0.492	0.296	0.132	0.181	0.190	0.488	0.068	-0.084

distributed throughout the observation period, without any significant sequence of consecutive days without data. The plot was drawn by means of the R function *vis_miss* from the *naniar* library (Tierney et al. 2015).

Unfortunately, no time series among those analyzed appears to be complete, which raises difficulty in identifying the underlying mechanism. The MAR mechanism was then assumed to be main driver of the data loss. The consequence of such an assumption is that the rebuilt time series should have the same distribution as the original data with the same parameters (mean and standard deviation). The following reconstruction methods were applied: *mice*, *MissForest*and *imp.norm* from R packages *mice*, *missForest* and *norm*, respectively (Van Buuren 2007; Feng et al. 2020; Anderson et al. 1994). As a result, the rebuilt method reporting the best performance was *imp.norm* with an overlap rate greater than 97% (Table 6).

As mentioned above, the validation is structured in two stages. During the first stage, an overall comparison between densities is carried out; in the second stage, the main parameters, mean and standard deviation, before and after the data reconstruction are computed and compared. To perform the first stage, the two produced densities (the original and the rebuilt density) are overlapped, and the percentage of overlap is computed. A percentage larger than 90% indicates very good reconstruction performance, and a percentage larger than 95% can be considered excellent (De Girolamo et al. 2022). As shown in Table 7, the overlap values were always greater than 97%; therefore, the performed reconstruction was judged as excellent. For the second stage, the average and the standard deviation of the original and reconstructed data were compared to assess the deviation of these parameters from one dataset to the other. Table 7 shows that such deviations can be considered substantially negligible; therefore, this further validation stage can be considered successful.

4.3 Number of subzeros data points after reconstruction

Table 7 reports the number of subzeros days after dataset reconstruction for each year.

The thirty years of data were analysed to determine possible trends underlying the presence of climate change effects on the minimum temperature distribution. Figure 5 shows the relationship between elevation and the number of subzero days; as expected, the number of subzeros increases with increasing elevation.

Another interesting analysis carried out on the sub-zero values was the cluster analysis, which involved crossing the subzero values and the elevation by means of a *heatmap* plot. By analysing the heatmap in Fig. 6, it can be concluded that there are two homogeneous clusters of monitoring sites formed by the 4 sites. Nocelle – Arvo (ID 1500), Cecita (ID 1100), Camigliatello (ID 1092) and Serra S. Bruno (ID 1980) belong to the first group, while Acri (ID 1120), Domanico (ID 1000), Antonimina - Canolo Nuovo (ID 2180) and Gambarie d'Aspromonte (ID 2470) belong to the second group. These two groups are consistent with the elevation.

4.4 Trend assessment

Due to the nonsignificant degree of temporal autocorrelation, the Mann-Kendall trend without correction for

Table 7 Number of daily minimum temperature subzeros per year

Year	1000	1092	1100	1120	1500	1980	2180	2470
1990	15	107	139	37	144	46	8	28
1991	41	135	166	49	151	96	35	51
1992	31	121	119	42	137	88	18	45
1993	39	121	129	55	133	87	31	58
1994	17	109	123	21	118	72	12	21
1995	30	132	116	40	114	91	23	40
1996	37	129	110	44	108	80	20	36
1997	30	124	126	31	127	93	9	40
1998	38	131	121	42	121	92	29	49
1999	26	119	105	39	107	79	25	49
2000	30	103	107	41	106	69	24	50
2001	27	125	97	33	92	62	20	43
2002	18	104	83	25	91	67	14	28
2003	42	127	115	49	113	85	39	58
2004	28	113	91	31	98	61	18	40
2005	49	129	119	46	122	96	50	62
2006	32	115	108	30	107	80	28	49
2007	12	121	96	7	104	67	11	18
2008	16	123	95	12	110	68	12	23
2009	23	129	85	20	93	51	22	34
2010	29	112	79	24	97	46	26	34
2011	19	123	106	16	112	68	19	32
2012	30	113	109	29	117	70	39	51
2013	13	107	93	11	98	53	25	36
2014	5	99	89	5	102	56	9	16
2015	16	107	107	15	120	62	25	37
2016	15	104	83	14	99	54	17	23
2017	17	121	112	19	121	83	32	47
2018	9	107	80	9	83	39	18	31
2019	25	113	98	23	110	68	32	49
Average	**25**	**117**	**107**	**29**	**112**	**71**	**23**	**39**

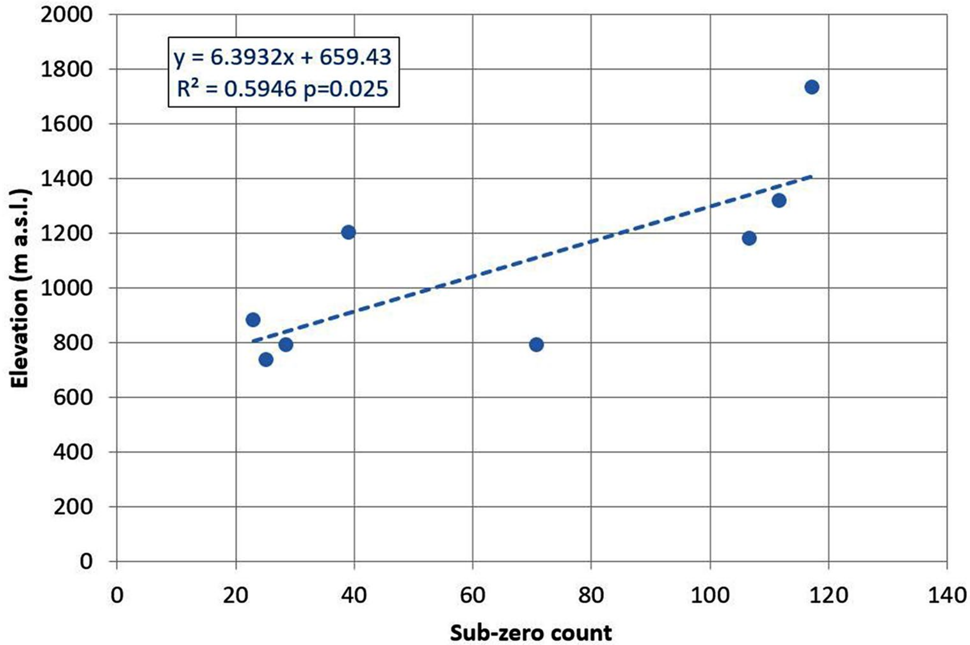

Fig. 5 Relationships between subzero count (x) and elevation (y). The relationship was significant ($p = 0.025$)

autocorrelation was applied. Table 8 shows the results of the trend tests.

The p-values in italics refer to the significant test outcomes. Only two out of eight series were nonsignificant: site Gambarie d'Aspromonte, (ID 2470) with a negative slope of -0.25, and site Antonimina - Canolo Nuovo (ID 2180), which had a slope of 0.14. This trend behavior can be influenced by the geographic position of the stations, that are the southernmost considered and are directly interested by African air currents, presenting high-temperature values with respect to the other stations. As regards the significant trends, the most marked negative values were detected in the Sila Massif and, specifically, in the Cecita, Acri and Nocelle - Arvo stations with slopes of -1.50, -1.21 and − 1.05, respectively. A relevant reduction of -1.14 has also been detected at the Serra San Bruno site, while less marked trends have been identified in Domanico (ID 1000, slope of -0.65) and Camigliatello (ID 1092, slope of -0.46).

The results of this study agree with the scientific findings related to projections on southern Europe's frost regime (Torma and Kis 2022) and with previous investigations on maximum and minimum temperatures at a planetary scale that evidenced an increase in the minimum values rather than the maximum values (Caloiero and Guagliardi 2020).

5 Discussion

The results of this study proved that the outcomes of a data reconstruction process can be confidently analysed via statistical analysis. It is sufficient that those outcomes be rebuilt following the precautions needed; thus, the results can be effectively applied for climate change analysis. Therefore, tools for data reconstruction are invaluable for carrying out complex data analysis and obtaining important results that are impossible to obtain otherwise. The results of the present study, undoubtedly prove that the Calabria region is interested in a warming process. This conclusion is reached by showing that the eight- time series of the count subzero values in six out of the eight sites have a decreasing trend.

It is necessary to develop adaptation measures and to stimulate their implementation by decision- makers in politics, administration, the economy, and society considering that climate change affects different sectors, such as agriculture, forestry, urban and regional planning, nature conservation, and water management, and that its impacts will inevitably be felt more in the future.

By analysing the scientific literature, it is possible to find papers about old and new missing data reconstruction methodologies that are claimed by authors as "optimal". Such papers are commonly structured as follows: an incomplete dataset, a method declared by the authors as the best, at least a couple of benchmark methods and a fine statistical analysis to confirm the claim. No one seems to realize about a paradoxical situation: often a benchmark in one paper, i.e., the weak method to be improved, becomes the best method in another paper. There is a straightforward explanation that solves the paradox: the optimality of a missing data reconstruction method must be considered not absolute but appropriately related only to the considered dataset. However, the crucial question to address is as follows: if the best absolute method does not exist, how can a suitable reconstruction method be selected at the start of the process of data reconstruction? The common approach is to

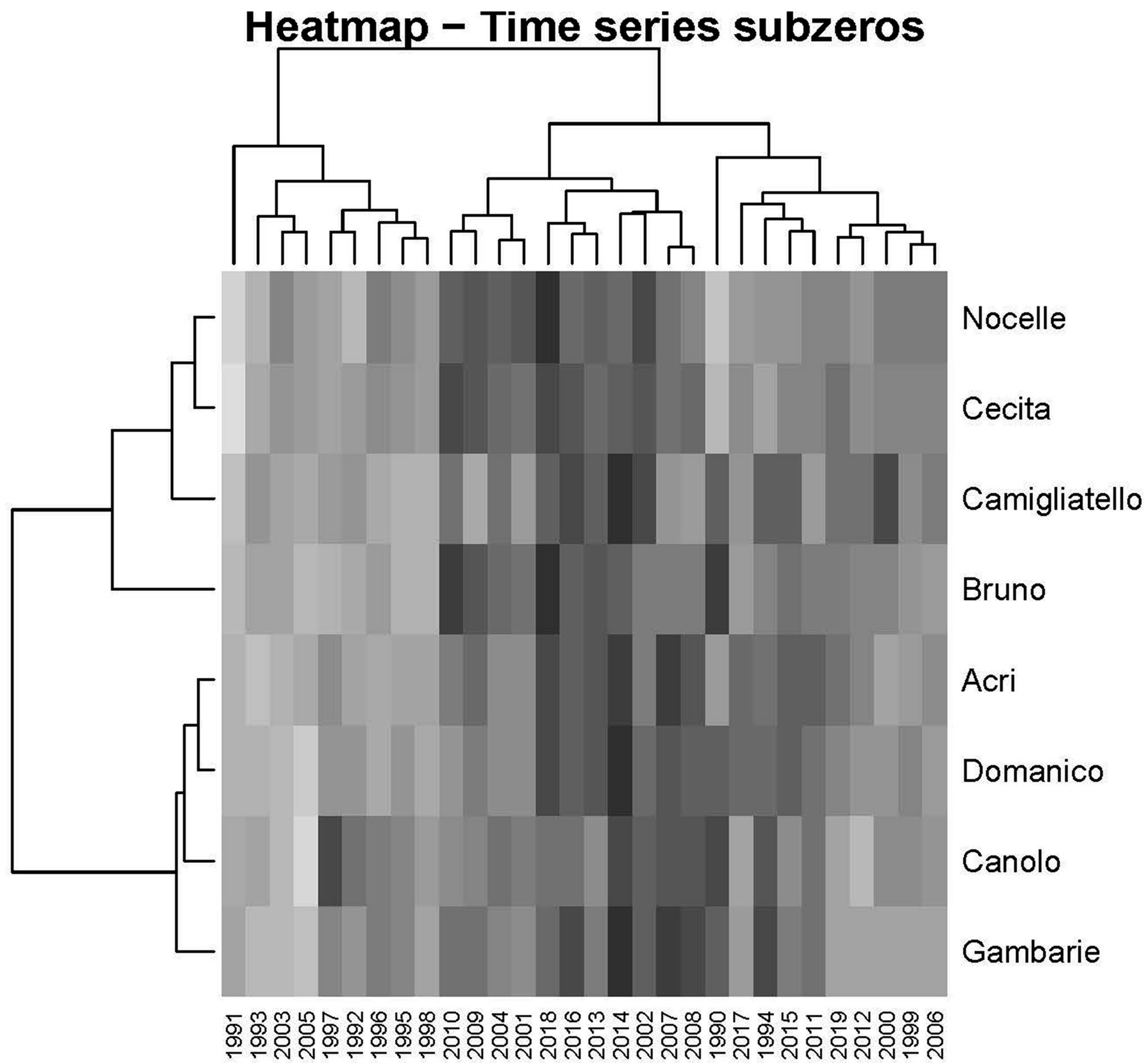

Fig. 6 Heatmap of the values reported in Table 8. The dendrogram on the left represents clusters between sites; the dendrogram on the top represents similarities between years

Table 8 Parameters of the Mann-Kendall trend test

Site	tau	Sen's slope	*p*-value
1000	-0.37	-0.65	0.0045
1092	-0.22	-0.46	0.029
1100	-0.52	-1.50	5.08e-05
1120	-0.54	-1.21	3.2e-05
1500	-0.36	-1.05	0.005
1980	-0.38	-1.14	0.003
2180	0.078	0.14	0.55
2470	-0.14	-0.25	0.29

Italics is for significant outcomes

select the most commonly used method, or one of the most recently introduced methods. However, the consequences of Wolpert theorem suggest a possible way to follow: by applying a suite of reconstruction methods and selecting the best method according to the rank of similarity between the incomplete dataset and the completed dataset. Obviously, the best-found method downstream of the proposed procedure cannot be considered the best absolute method but is the best with respect to the dataset used. In the present study, three methods were selected and compared, namely *mice*, *missForest* and *norm*; these methods have different technical characteristics, e.g. parametric vs. nonparametric. This approach was used to attempt to grasp, at most, the structure underlying the incomplete dataset. The goal of the proposed methodology is not to simply fill in the gaps in the incomplete dataset but accomplish this goal by maximizing confidence in the conclusive statistical analysis and in its results.

6 Conclusions

The present research attempts to show how assessments of high quality and consistency with temperature data are important for underpinning any analyses. This study aimed to explore the minimum temperature trend by means of a gap-filled database of 8 temperature series collected over a 30-year period (1990–2019) in the Calabria region in southern Italy and to provide gap filling of daily temperature data for a reliable trend analysis of frost days in the Calabria region (southern Italy).

The following conclusions can be drawn:

- the performances of the gap-filling technique in estimating daily observed temperature data is composed of strict criteria (specific tests), although it allows adjustments by expert users, according to their knowledge of the local climate. Specifically, the percentages of missing data were assessed. If all of the missing data were reconstructed, the possible patterns among missing data, the correlation analysis between sites and target sites, the mechanisms responsible for data value losses, and the selection of a group of suitable missing data recovery methods that must be compared to safely choose the best data reconstructionwere assessed to determine the best reconstruction strategy;
- after the missing data reconstruction, the Mann-Kendall nonparametric test showed a decreasing trend in the number of frost days for six out of the eight series, confirming an overall increase in the minimum temperature in the study area;
- that outcome can be interpreted as an actual measurement of the climate change effect over the region of interest;
- this approach includes novel and effective methodologies that, under suitable boundary conditions, are capable of reconstructing the missing data.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Emanuele Barca, Tommaso Caloiero and Ilaria Guagliardi. The first draft of the manuscript was written by Emanuele Barca, Tommaso Caloiero and Ilaria Guagliardi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Anderson E, Bai Z, Bischof C, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Ostrouchov S, Sorensen D et al (1994) LAPACK User's Guide, 2nd edition, SIAM, Philadelphia

Barca E, Bruno E, Bruno DE, Passarella G (2016) GTest: a software tool for graphical assessment of empirical distributions' gaussianity. Environ Monit Assess 188:138

Barca E, Porcu E, Bruno DE, Passarella G (2017) An automated decision support system for aided assessment of variogram models. Environ Model Softw 87:72–83

Bonsal BR, Zhang X, Vincent L, Hogg W (2001) Characteristics of daily and extreme temperatures over Canada. J Clim 14:1959–1976

Brunsdon C, Fotheringham AS, Charlton M (2002) Geographically weighted Summary statistics. Comput Environ Urban Syst 26:501–524

Buttafuoco G, Caloiero T, Guagliardi I, Ricca N (2016) Drought assessment using the reconnaissance drought index (RDI) in a southern Italy region. 6th IMEKO TC19 Symposium on Environmental Instrumentation and Measurements, 52–55

Buttafuoco G, Caloiero T, Ricca N, Guagliardi I (2018) Assessment of drought and its uncertainty in a southern Italy area (Calabria region). Meas: J Int Meas Confed 113:205–210

Caloiero T, Guagliardi I (2020) Temporal variability of temperature extremes in the Sardinia region (Italy). Hydrology 7:55

Caloiero T, Guagliardi I (2021) Climate change assessment: seasonal and annual temperature analysis trends in the Sardinia region (Italy). Arab J Geosci 14:2149

Caloiero T, Callegari G, Cantasano N, Coletta V, Pellicone G, Veltri A (2016) Bioclimatic analysis in a region of southern Italy (Calabria). Plant Biosyst 150:1282–1295

Caloiero T, Coscarelli R, Ferrari E, Sirangelo B (2017) Trend analysis of monthly mean values and extreme indices of daily temperature in a region of southern Italy. Int J Climatol 37:284–297

Caloiero T, Pellicone G, Modica G, Guagliardi I (2021) Comparative analysis of different spatial interpolation methods Applied to

Monthly Rainfall as support for Landscape Management. Appl Sci 11:9566
Caroletti GN, Coscarelli R, Caloiero T (2019) Validation of Satellite, Reanalysis and RCM Data of Monthly Rainfall in Calabria (Southern Italy). Remote Sens 11:1625
Ceppi P, Scherrer SC, Fischer A, Appenzeller C (2012) Revisiting Swiss temperature trends 1959–2008. Int J Climatol 32:203–213
Claridge DE, Chen H (2006) Missing data estimation for 1–6 h gaps in energy use and weather data using different statistical methods. Int J Energy Res 30:1075–1091
Comtois D (2021) summarytools: Tools to Quickly and Neatly Summarize Data. R package version 0.9.9. https://CRAN.R-project.org/package=summarytools
Contosta AR, Casson NJ, Garlick S, Nelson SJ, Ayres MP, Burakowski EA, Campbell J, Creed I, Eimers C, Evans C, Fernandez I, Fuss C, Huntington T, Patel K, Sanders-DeMott R, Son K, Templer P, Thornbrugh C (2019) Northern forest winters have lost cold, snowy conditions that are important for ecosystems and human communities. Ecol Appl 29:e01974
Craddock JM (1979) Methods of comparing annual rainfall records for climatic purposes. Weather 34:332–346
De Girolamo AM, Barca E, Leone M, Porto AL (2022) Impact of long-term climate change on flow regime in a Mediterranean basin. J Hydrology: Reg Stud 41:101061
Dietz EJ, Kileen A (1981) A nonparametric multivariate test for Monotone Trend with Pharmaceutical Applications. J Am Stat Assoc 76:169–174
Djaman K, Balde AB, Rudnick DR, Ndiaye O, Irmak S (2017) Long-term trend analysis in climate variables and agricultural adaptation strategies to climate change in the Senegal River Basin. Int J Climatol 37:2873–2888
Fagandini C, Todaro V, Tanda MG et al (2024) Missing Rainfall Daily Data: a comparison among gap-filling approaches. Math Geosci 56:191–217
Feng L, Moritz S, Nowak G, Welsh AH, O'Neill TJ (2020) imputeR: A General Multivariate Imputation Framework_. R package version 2.2, https://CRAN.R-project.org/package=imputeR.
Gaglioti S, Infusino E, Caloiero T, Callegari G, Guagliardi I (2019) Geochemical characterization of spring waters in the Crati River Basin, Calabria (Southern Italy). Geofluids 3850148
Gentilucci M, Materazzi M, Pambianchi G, Burt P, Guerriero G (2020) Temperature variations in Central Italy (Marche region) and effects on wine grape production. Theor Appl Climatol 140:303–312
Giorgi F (2002) Variability and trends of sub-continental scale surface climate in the twentieth century. Part I: Clim Dyn 18:675–691
Goubanova K, Li L (2007) Extremes in temperature and precipitation around the Mediterranean basin in an ensemble of future climate scenario simulations. Glob Planet Change 57:27–42
Guagliardi I, Caloiero T, Infusino E, Callegari G, Ricca N (2021) Environmental estimation of Radiation Equivalent Dose Rates in Soils and Waters of Northern Calabria (Italy). Geofluids 6617283. https://doi.org/10.1155/2021/6617283
Hipel KW, McLeod AI (1994) Time Series Modelling of Water resources and Environmental systems. Elsevier, Amsterdam, p 1013
Hirsch RM, Slack JR, Smith RA (1982) Techniques of Trend Analysis for Monthly Water Quality Data. Water Resour Res 18:107–121
Honaker J, King G (2010) What to do about missing values in time-series cross-section data. Am J Polit Sci 54:561–581
Infusino E, Guagliardi I, Gaglioti S, Caloiero T (2022b) Vulnerability to Nitrate occurrence in the Spring Waters of the Sila Massif (Calabria, Southern Italy). Toxics 10:137
Infusino E, Caloiero T, Fusto F, Calderaro G, Brutto A, Tagarelli G (2022a) Characterization of the 2017 summer heat waves and their effects on the Population of an area of Southern Italy. Int J Environ Res Public Health 18:970
Iovine G, Guagliardi I, Bruno C, Greco R, Tallarico A, Falcone G, Lucà F, Buttafuoco G (2018) Soil-gas radon anomalies in three study areas of Central-Northern Calabria (Southern Italy). Nat Haz 91:193–2191
IPCC (2013) Summary for policymakers. Fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
Jacobeit J (2000) Rezente Klimaentwicklung Im Mittelmeerraum. Petermanns Geogr Mittl 144:22–33
Kendall MG (1975) Rank correlation methods. Charles Griffin, London
King G, Honaker J, O'Connell AJ, Scheve K (2001) Analyzing Incomplete Political Science Data: an alternative algorithm for multiple imputation. Am Polit Sci Rev 95:4969
King AD, Donat MG, Fischer EM, Hawkins E, Alexander LV, Karoly DJ, Dittus AJ, Lewis S, Cand Perkins SE (2015) The timing of anthropogenic emergence in simulated climate extremes. Environ Res Lett 10:094015
Köppen W (1936) Das Geographische System der Klimate. Handbuch der Klimatologie; Köppen,W., Geiger, R., Eds.; Verlag von Gebrüder Borntraeger: Berlin, Germany, 1936; Vol. 1, pp. 1–44
Kundzewicz Z, Robson A (2000) Detecting trend and other changes in hydrological data. World Climate Programme Data and Monitoring. WMO/TD-No. 1013, Geneva
Lettenmaier DP (1988) Multivariate Nonparametric tests for Trend in Water Quality. Water Resour Bull 24:505–512
Leurent B, Crawford M, Gilbert H, Morris R, Sweeting M, Nazareth I (2013) Sensitivity analyses for trials with missing data, assuming missing not at random mechanisms. Trials 14:O97
Libiseller C, Grimvall A (2002) Performance of partial Mann-Kendall Test for Trend Detection in the Presence of Covariates. Environmetrics 13:71–84
Lin J, Qian T (2022) The atlantic multi-decadal oscillation. Atmos Ocean 60:307–337
Liston GE, Elder K (2006) A meteorological distribution system for high-resolution terrestrial modeling (MicroMet). J Hydrometeorol 7:217–234
Little RJA (1988) Missing-data adjustments in large surveys (with discussion). J Bus Econ Sta 6:287–301
Little RJA, Rubin DB (1987) Statistical analysis with Missing Data. Wiley, New York
Lo Presti RL, Barca E, Passarella G (2010) A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environ Monit Assess 160:1–22
Lompar M, Lalić B, Dekić L, Petrić M (2019) Filling gaps in Hourly Air Temperature Data using Debiased ERA5 Data. Atmosphere 10:13
Mann HB (1945) Nonparametric tests against Trend. Econometrica 13:245–259
Martinez CJ, Maleski JJ, Miller MF (2012) Trends in precipitation and temperature in Florida, USA. J Hydrol 452–453:259–281
Mishra AK, Singh VP (2011) Drought modeling - a review. J Hydrol 791:157–175
Murara P, Acquaotta F, Garzena D, Fratianni S (2019) Daily precipitation extremes and their variations in the Itajaí River Basin, Brazil. Meteorol Atmos Phys 131:1145–1156
Nawaz Z, Li X, Chen Y, Guo Y, Wang X, Nawaz N (2019) Temporal and spatial characteristics of precipitation and temperature in Punjab, Pakistan. Water 11:1916
Pellicone G, Caloiero T, Modica G, Guagliardi I (2018) Application of several spatial interpolation techniques to monthly rainfall data in the Calabria region (southern Italy). Int J Climatol 38:3651–3666
Pellicone G, Caloiero T, Guagliardi I (2019) The De Martonne aridity index in Calabria (Southern Italy). J Maps 15:788–7963

Ricca N, Guagliardi I (2015) Multi-temporal dynamics of land use patterns in a site of community importance in Southern Italy. Appl Ecol Environ Res 13:677–691
Rubin DB (1976) Inference and Missing Data. Biometrika 63:581–590
Rubin DB (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. J Bus Econ Stat 4:87–94
Scheffer J (2002) Dealing with missing data. Res Lett Inf Math Sci 3:153–160
Schmidt MW, Hertzberg JE (2011) Abrupt climate change during the last ice age. Nat Educ Knowl 3:11
Scorzini AR, Leopardi M (2019) Precipitation and temperature trends over central Italy (Abruzzo Region): 1951–2012. Theor Appl Climatol 135:959–977
Sen PK (1968) Estimates of the regression coefficient based on Kendall's tau. J Am Sta Assoc 63:1379–1389
Sirangelo B, Caloiero T, Coscarelli R, Ferrari E, Fusto F (2020) Combining stochastic models of air temperature and vapour pressure for the analysis of the bioclimatic comfort through the Humidex. Sci Rep 10:11395
Stekhoven DJ, Bühlmann P (2012) MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112–118
Templ M, Alfon A, Filzmoser P (2012) Exploring incomplete data using visualization tools. J Adv Data Anal Classif 6:29–47
Thode JHC (2002) Testing for normality. Marcel Dekker, New York
Tierney NJ, Harden FA, Harden MJ, Mengersen KA (2015) Using decision trees to understand structure in missing data. BMJ Open Jun 295:e007450. https://doi.org/10.1136/bmjopen-2014-007450
Torma CZ, Kis A (2022) Bias-Adjustment of High-Resolution Temperature CORDEX Data over the Carpathian Region: Expected Changes Including the Number of Summer and Frost Days. Int J Climatol 2022:1–16
Van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med Res 16:219–242
Vicente-Serrano SM, Lopez-Moreno JI, Beguería S, Lorenzo-Lacruz J, Sanchez-Lorenzo A, García-Ruiz JM et al (2014) Evidence of increasing drought severity caused by temperature rise in southern Europe. Environ Res Lett 9:044001
von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge, New York, p 484
Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural Comput 8:1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341
Zarenistanak M, Dhorde AG, Kripalani RH (2014) Trend analysis and change point detection of annual and seasonal precipitation and temperature series over southwest Iran. J Earth Syst Sci 123:281–295
Zhang X, Alexander L, Hegerl GC, Jones P, Tank AK, Peterson TC, Trewin B, Zwiers FW (2011) Indices for monitoring changes in extremes based on daily temperature and precipitation data. WIREs Clim Chang 2:51–870
Zhao P, Jones LJ, Cao Yan Z, Zha S, Zhu Y, Yu Y, Tang G (2014) Trend of surface air temperature in eastern China and associated large-scale climate variability over the last 100 years. J Clim 27:4693–4703