Accurate classification of bioclimatic data: temporal analysis

Giuseppe Passarella, Emanuele Barca, Delia Bruno,
Sabino Maggi and Rita Masciale
CNR-IRSA Water Research Institute,
Via F. De Blasio 5, 70132 Bari, Italy
Email: sabino.maggi@cnr.it

Aimé Lay-Ekuakille
Department of Innovation Engineering,
Università del Salento,
Via Monteroni, 73100 Lecce, Italy

Abstract— This paper presents a general methodology for processing bioclimatic data in the temporal domain. Two different methods are used to assess the presence of temporal trends in the time-series of bioclimatic indices at each measurement station. A preliminary stage checks for the statistical homogeneity in the data set and for the presence of serial autocorrelation in the data, applying the proper methods to remove these effects. The methodology has been applied to a case study in Apulia, Italy, using the popular De Martonne index as a bioclimatic indicator.

Index Terms—Bioclimatic indices, time-series, trend test, temporal analysis.

I. INTRODUCTION

Bioclimatic indices are well-established tools to characterize a region from the climatic point of view, in particular for its degree of moisture or dryness. These indices provide useful information about the irrigation requirements of the region or the dynamics of recharge of groundwater, an important parameter in coastal areas to prevent or control salinization of aquifers.

The accurate determination of the spatial distribution of the bioclimatic indices is of great importance for water delivery and management authorities and also for land planners. Likewise, the knowledge of the temporal evolution of these indices is of paramount importance to find trends that might lead to severe environmental risks.

In a companion paper, we have presented a general methodology to process bioclimatic indices in the spatial domain, where weather-related measurements taken at local meteorological stations randomly spread across the area are transformed into regular maps calculated over a fine grid, superimposed over the region under consideration [1]. The methodology has been applied to a case study in Apulia, southern Italy, where the widely used De Martonne index has been used as the bioclimatic indicator [2].

In this paper, we show how the same measurements can be analyzed in the temporal domain, showing how the De Martonne index changes with time over the whole area of the region. Taken together, the two papers can give accurate indications about the dryness level of a region, helping water delivery and management authorities and land planners to best manage the irrigation requirements of a given area. A broader goal of this work is to contribute to the study of the effects of global warming in a typical Mediterranean environment.

II. TEMPORAL ANALYSIS

In this paper, we have used two different methods for assessing the presence of temporal trends in the time-series of bioclimatic indices at each measurement station.

The first method is recommended by the World Meteorological Organization (WMO) as the standard method for the evaluation of temporal trends on meteoclimatic time-series series [3]. The second method, based on the F-test, is often applied in different research areas to study of the significance of the parameters of regressive models. Even if the two methods differ in their basic assumptions and complexity of calculations, in most cases they produce very similar results, as will be shown in the case study. Nevertheless, the use of two separate methods is still useful in the particular cases where one of the two methods might fail or give meaningless results.

Trend analysis is usually a very sensitive statistical task that is performed according to a methodology involving the application of a clever succession of statistical tests (Figure 1).

The first step verifies that the time-series data are statistically homogeneous. It the test fails, it applies one of the available statistical methods to homogenize the data set. The second test checks for the presence of serial autocorrelation in the data and, if needed, uses pre-whitening methods to remove it. Lastly, the presence of a trend is verified, calculating the slope of the trend when the result is significant. The trend test can be either parametric or non-parametric (i.e., distribution free), depending on whether it is necessary or not to take into account the probability distribution associated with the values of the variable of interest.

A data set is homogeneous when it can be considered to be taken from the same statistical population. The presence of heterogeneity in the data set can lead to the assessment of an wrong trend. The heterogeneity of a set of measurements can be related to factors such as the physical displacement of the measuring station (change of coordinates) or to the replacement of the measuring instrument with another one
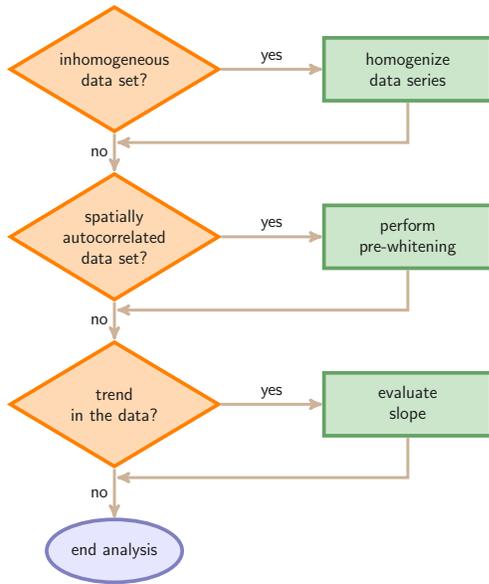
Fig. 1: Basic methodology for trend analysis.

having a different sensitivity or accuracy (e.g., the replacement of a manual pluviometer with an automatic instrument).

Several statistical tests are available to test the departure of homogeneity of the data set, among them the Buishand range test [4] used in this paper, the standard normal homogeneity test (SNHT) [5] and the von Neumann ratio test [6]. The first two tests are location-specific tests, as they can locate the year where a break in the data set is likely to occur, while the latter von Neumann test does not give this information. If heterogeneity is established, the data set should be subjected to a homogenization procedure, usually based on the SNHT test [5].

Another preliminary check is related to serial autocorrelation in the time-series. To avoid getting a false positive in the trend test, if some form of correlation in the data is found it is thus necessary to perform a pre-whitening procedure that removes the serial correlation in the time-series through a modification of the data values [7], [8].

A. Trend test and F-test

To check the presence of a trend in a time-series, several effective tests are available in the scientific literature. Among them, the two most popular tests are the Mann-Kendall test and the Spearman rho test [3], [9]–[11]. Both tests are rank-based and therefore independent of the distribution of the data (distribution free) and are already implemented in the main statistical software packages.

Unfortunately, statistical significance does not always coincide with the practical significance of the observed phenomenon. In other words, statistics tools are often able to discover effects of such a small size that do no make any sense in practical terms. At the end of the analysis, it is then important to quantify the slope of the trend, verifying its meaningfulness in practical terms. For such a calculation, the

Mann-Kendall test usually uses the Theil-Sen non-parametric estimator, that is in principle insensitive to the distribution of the data to be analyzed [12].

From what has been noted above, it is clear that the assessment of a trend in a time-series follows a quite complex procedure and requires a considerable knowledge of statistical methodologies and software. A faster and simpler alternative for checking the trend in a time-series is represented by the so-called zero slope F-test, a parametric test based on the assumption of gaussianity and homoscedasticity of the data, that can be easily implemented even within a simple spreadsheet application.

The F-test is robust for moderate deviations from the above assumptions. Its main limit is that it can be used only to verify the significance of linear relationships, not for other types of monotonic functions. When using the F-test, the calculation of the slope is usually performed using the ordinary least-squares (OLS) method.

III. CASE STUDY

In this paper we have used the same time-series containing precipitation and temperature data used for the spatial analysis [1]. The selected data sets contain measurements from 1931 to 2010, taken at the 82 weather monitoring stations of the Civil Protection of Apulia, Italy, where such extensive time-series data are available [13]–[15]. Missing data have been reconstructed using the effective MICE method, implemented in the R `mice` package [16], [17]. The De Martonne index has been chosen as the proper bioclimatic indicator for this case study and the annual De Martonne index has been calculated for each reconstructed time-series. A more detailed description of the De Martonne index and the MICE method can be found in the accompanying paper.

The Buishand range test does not detect any significant heterogeneity in the available data sets. Similarly, the serial autocorrelation test shows no evidence of autocorrelation.

The significance of the temporal trends at each weather monitoring station has been verified at the $p = 0.05$ level, applying the rigorous Mann-Kendall test method and the faster and simpler F-test method. Among the 82 weather monitoring stations, only 11 have shown a significant temporal trend. For all the other stations the significance hypothesis for the temporal trend has been rejected.

Table I list the 11 monitoring stations where a significant temporal trend has been assessed. From the Table it is apparent that the two tests yield very similar results, except in two stations, Fasano and Nusco, where the results of the two tests are opposite.

For these 11 stations, we have estimated the associated slope, where a positive slope corresponds to an increasing trend while a negative slope is related to a decreasing trend. As already noted, when using the Mann-Kendall test, the slope is calculated with the Theil-Sen estimator [12], while the F-test is intrinsically associated to the classical OLS method for the evaluation of the slope. The two methods give very similar slopes, with a ratio always very close or equal to 1.00, with

TABLE I: Significant temporal trends.

Monitoring station	Mann-Kendall test				F-test				Slope ratio
	p-value	Significance	Theil-Sen slope	Trend	p-value	Significance	OLS slope	Trend	
Ascoli Satriano	0.024	yes	-0.05	P	0.012	yes	-0.07	P	0.71
Bosco Umbra	0.015	yes	-0.15	P	0.009	yes	-0.16	P	0.94
Fasano	0.054	no	-0.06	—	0.017	yes	-0.07	P	—
Forenza	0.001	yes	-0.11	P	0.000	yes	-0.12	P	0.92
Gioia del Colle	0.015	yes	0.07	M	0.010	yes	0.07	M	1.00
Lagopesole	0.013	yes	0.10	M	0.005	yes	0.10	M	1.00
Maglie	0.040	yes	-0.08	P	0.037	yes	-0.08	P	1.00
Monte Sant'Angelo	0.014	yes	-0.12	P	0.007	yes	-0.12	P	1.00
Monticchio	0.000	yes	0.19	M	0.000	yes	0.19	M	1.00
Nusco	0.018	yes	-0.13	P	0.064	no	-0.10	—	—
Ruvo di Puglia	0.030	yes	-0.05	P	0.021	yes	-0.06	P	0.83

the only exception of the Ascoli Satriano monitoring station, where the ratio between the two calculated slopes significantly differs from 1.00 (Table I).

The monitoring stations of Gioia del Colle, Lagopesole and Monticchio, the latter two located outside the regional territory but managed by the Civil Protection of Apulia and for this reason included in the case study, are characterized by a moderately increasing trend, that is they are heading towards a more humid bioclimatic profile. All other stations appear to be moving very slowly towards a drier bioclimatic situation.

It is worth noting that the calculation of the slope can always be carried out, regardless of the outcome of the trend significance test for the trend. Unfortunately, this may often lead to an improper use of the results, since the presence of a positive or negative slope in a time-series does not automatically assure the presence of a real increasing (or
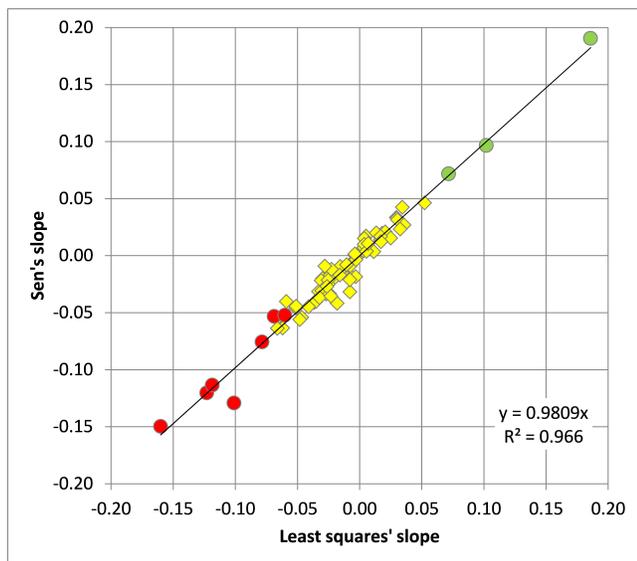

Fig. 2: Relation between the slopes calculated with the Theil-Sen and OLS methods, for the time-series measured at the 82 weather monitoring stations: (squares) non-significant trend, (red circles) significant negative slope, (green circles) significant positive slope.

decreasing) trend in the data set.

To better clarify this point, we show the relation between the slopes calculated with the Theil-Sen and OLS methods, for all the 82 available time-series (Figure 2). The pairs of slopes characterized by a significant positive or negative value are shown as green or red circles, respectively, while the non-significant trends are represented by yellow squares. All significant trends are characterized by positive or negative slopes located at the ends of the regression line, while all slopes located around the origin are not significant. The regression line, $y = 0.9809x$, and the coefficient of determination, $R^2 = 0.966$, confirm the excellent agreement between the two methods for calculating the slope.

The results of Figure 2 can be easily transposed on the spatial map of municipalities in Apulia, as shown in Figure 3, where the left map represents the results obtained with the Theil-Sen estimator, and the right map those with the OLS method. As in Figure 2, the municipalities characterized by a significant positive or negative slope of the bioclimatic index are shown in green and red, respectively, while the municipalities marked in yellow do not show any significant trend in the bioclimatic indicator.

Even if mathematically equivalent, positive or negative slopes have a profoundly different meaning from a climatic and environmental point of view. The simple assessment of a temporal trend cannot ignore the average bioclimatic conditions of the region. In Europe, an area that becomes more arid is in general more worrisome than an humid area that tend to a more humid climate, but the same could not always be true elsewhere. As an example, Figure 4 shows the time-series of two weather-stations characterized by a decreasing (i.e., more arid) or increasing (more humid) trend of the bioclimatic index.

Only the proper intersection of the information derived from the bioclimatic maps, built on the basis of the time-series of bioclimatic indexes, with the information about the trends and the transition from one bioclimatic class to another, allows to build a consistent and reliable knowledge base on which to base an effective and efficient territory management strategy.

An alternative and effective view of the temporal behavior of the bioclimatic index is obtained by representing for each year the most frequent bioclimatic class (modal class) occurring
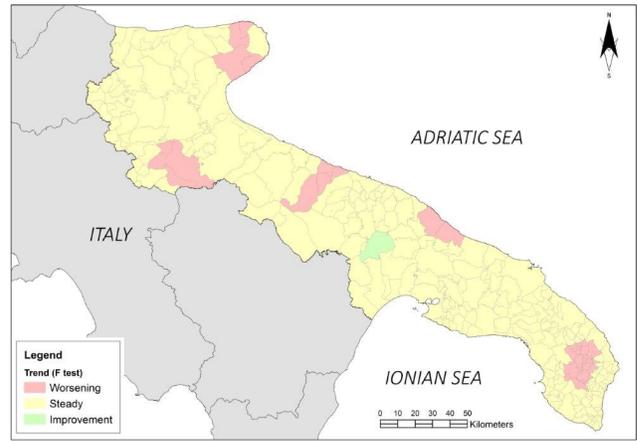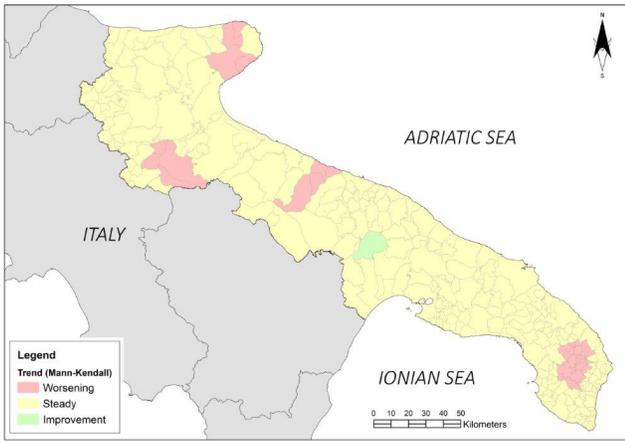
Fig. 3: Spatial map of municipalities in Apulia, showing the significativity of the trends calculated with the (left) Theil-Sen estimator and (right) ordinary least-squares.
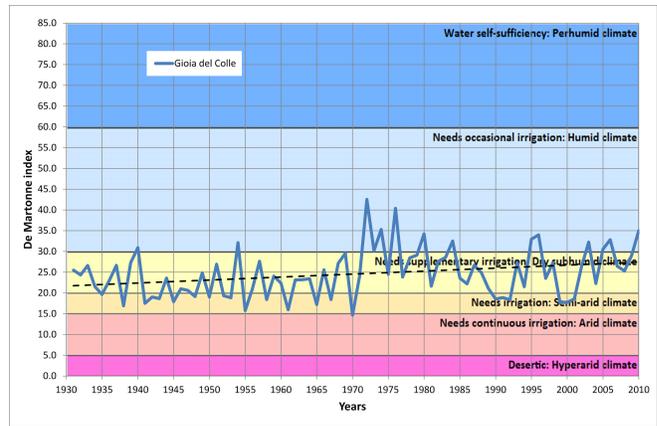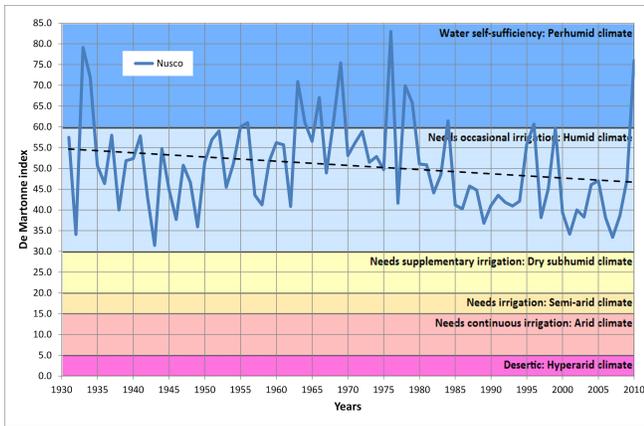

Fig. 4: Examples of time-series characterized by a significant (left) negative or (right) positive slope.

in Apulia between 1931 and 2010 (Figure 5). From the Figure, it is apparent that, at least in the relatively long time frame considered in this work, the Apulia region was characterized by a predominant ad persistent dry subhumid climate, interrupted by four occasional years – 1977, 1989, 1992 and 2000 – of arid climate, without any evidence of a transition to a more arid climate, as could in principle be expected in a southern mediterranean region.

A more detailed representation of the temporal behavior of the bioclimatic classes in Apulia is shown in Figure 6, which depicts the percentage of regional area belonging to each bioclimatic zone, from 1931 to 2010. In the time frame considered here, the areas characterized by a perhumid climate represented less than 5 % of the territory of the region, while areas with a humid and sub-humid climate predominated, covering roughly $50-60$ % of the area of the region. Desertic climate is totally absent in Apulia, at least in the period considered in this work, although large portions of regional territory have been characterized by a semi-arid and arid climate, in particular in the 1940s, in the 1960s, and in the last 15 years of the 20th century. Arid zones never exceed 30 %

of the region, except than in the four single years mentioned above (1977, 1989, 1992 and 2000), when they covered up to almost 50 % of the territory.

IV. CONCLUSION

In this and in the accompanying paper we have presented a methodology for characterizing a regional territory from the bioclimatic point of view that, starting from historical time-series of temperature and precipitation, allows to produce detailed maps of the persistent climatic zones, and to infer indications about their evolution over time.

The proposed methodology has been applied to a case study in Apulia, Italy, using thermo-pluviometric data measured at 82 stations of the monitoring network managed by the Regional Civil Protection from 1931 to 2010.

The De Martonne index has been chosen for the bioclimatic indicator for its intrinsic simplicity and ease of use. Nevertheless, it allows a reliable classification of the region from a bioclimatic point of view (aridity and irrigation).

The proposed methodology analyzes the characteristics and significance of the temporal trend of the bioclimatic index with
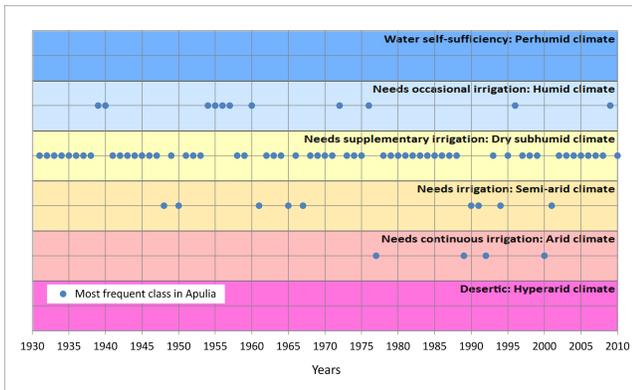
Fig. 5: Most frequent bioclimatic class of Apulia, from 1931 to 2010.

two separate methods, a first more rigorous method based on the Mann-Kendall test and on the Theil-Sen estimator, and a second more immediate method based on the F-test and on OLS. For the case study described here, the results of the two methods are equivalent.

The outcome of the work is represented by detailed maps of Apulia and bioclimatic plots, that show the evolution of the bioclimatic index with time.

The proposed methodology can be useful for the integrated management of the territory and in particular of its water resources. It allows to gather qualitative and quantitative information on the state of aridity, its possible evolution with time and the irrigation requirements at the regional and sub-regional level. Crossing this information with other related regional characteristics, such land use or water availability, it is thus possible to find the best compromise between water requirements and availability, avoiding potential imbalances and optimizing distribution costs and helping managing drought risks.

This spatial and temporal analysis of the climatic zones of Apulia could also be an useful tool to study the effects of global warming in a Mediterranean region.
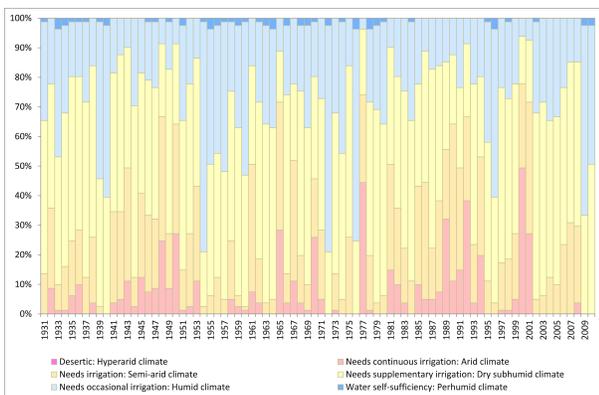

Fig. 6: Percentage of the area of the region belonging to each bioclimatic class, from 1931 to 2010.

REFERENCES

[1] G. Passarella, E. Barca, D. E. Bruno, S. Maggi, and A. Lay-Ekuakille, "Accurate classification of bioclimatic data: spatial analysis," *these Proceedings*, 2017.

[2] E. de Martonne, *Trait de Gographie Physique*. Librarie Arman Colin, Paris, 1925-1927.

[3] H. B. Mann, "Nonparametric tests against trend," *Econometrica*, vol. 13, no. 3, pp. 245–259, 1945. [Online]. Available: http://www.jstor.org/stable/1907187

[4] T. A. Buishand, "Some methods for testing the homogeneity of rainfall records," *Journal of Hydrology*, vol. 58, no. 1, pp. 11–27, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/002216948290066X

[5] H. Alexandersson, "A homogeneity test applied to precipitation data," *Journal of Climatology*, vol. 6, no. 6, pp. 661–675, 1986. [Online]. Available: http://dx.doi.org/10.1002/joc.3370060607

[6] J. V. Neumann, "Distribution of the ratio of the mean square successive difference to the variance," *The Annals of Mathematical Statistics*, vol. 12, no. 4, pp. 367–395, 1941. [Online]. Available: http://projecteuclid.org/euclid.aoms/1177731677

[7] K. H. Hamed and A. R. Rao, "A modified mann-kendall trend test for autocorrelated data," *Journal of Hydrology*, vol. 204, no. 1, pp. 182–196, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S002216949700125X

[8] M. Bayazit and B. Önöz, "To prewhiten or not to prewhiten in trend analysis?" *Hydrological Sciences Journal*, vol. 52, no. 4, pp. 611–624, 2007. [Online]. Available: http://dx.doi.org/10.1623/hysj.52.4.611

[9] S. Yue, P. Pilon, and G. Cavadias, "Power of the mannkendall and spearman's rho tests for detecting monotonic trends in hydrological series," *Journal of Hydrology*, vol. 259, no. 1, pp. 254–271, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022169401005947

[10] ——, "Corrigendum to power of the mann-kendall and spearman's rho tests for detecting monotonic trends in hydrological series [j. hydrol. 259 (2002) 254271]," *Journal of Hydrology*, vol. 264, no. 1, pp. 262–263, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022169402000781

[11] I. Ahmad, D. Tang, T. Wang, M. Wang, and B. Wagan, "Precipitation trends over time using mann-kendall and spearmans rho tests in swat river basin, pakistan," *Advances in Meteorology*, vol. 2015, pp. 431 860/1–15, 2015. [Online]. Available: https://www.hindawi.com/journals/amete/2015/431860/

[12] P. K. Sen, "Estimates of the regression coefficient based on kendall's tau," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389, 1968. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1968.10480934

[13] R. L. Presti, E. Barca, and G. Passarella, "A methodology for treating missing data applied to daily rainfall data in the candelaro river basin (italy)," *Environmental Monitoring and Assessment,*, vol. 160, no. 1, pp. 1–22, 2010. [Online]. Available: https://link.springer.com/article/10.1007%2Fs10661-008-0653-3

[14] E. Barca, L. Berardi, D. B. Laucelli, G. Passarella, and O. Giustolisi, "Evolutionary polynomial regression application for missing data handling in meteo-climatic gauging stations," in *Proceedings of GRASPA 2015*, 2015.

[15] E. Barca, D. E. Bruno, and G. Passarella, "Similarity indices of meteo-climatic gauging stations: definition and comparison," *Environmental Monitoring and Assessment*, vol. 188, no. 7, pp. 403/1–12, 2016. [Online]. Available: https://link.springer.com/article/10.1007/s10661-016-5407-z

[16] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. [Online]. Available: https://www.jstatsoft.org/v045/i03

[17] N. F. A. Radi, R. Zakaria, and M. A. Azman, "Estimation of missing rainfall data using spatial interpolation and imputation methods," *AIP Conference Proceedings*, vol. 1643, no. 1, pp. 42–48, 2015. [Online]. Available: http://aip.scitation.org/doi/abs/10.1063/1.4907423