Meteorological Time Series: An Exploratory Statistical and Critical Analysis

A. Manuela Gonçalves[1], F. Catarina Pereira[1], Marco Costa[2], and Celina P. Leão[3](✉)

[1] Department of Mathematics and Center of Mathematics, University of Minho, Braga, Portugal
mneves@math.uminho.pt, id9976@alunos.uminho.pt

[2] Águeda School of Technology and Management and Center for Research and Development in Mathematics and Applications, University of Aveiro, Aveiro, Portugal
marco@ua.pt

[3] ALGORITMI Centre, University of Minho, Guimarães, Portugal
cpl@dps.uminho.pt

Abstract. Increasingly, reduction of water availability has been a reality, and population growth, pollution, and climate change have contributed to exacerbating this problem. Dry periods, which occur when precipitation is lower than expected in a given territory, have become more frequent and prolonged, and therefore it is crucial to efficiently manage water use in response to environmental concerns. The main challenge in this work is to present the irrigation problem as an optimal control problem along with the presentation of preliminary results based on an exploratory statistical and critical analysis of daily meteorological variables. The variables considered are: maximum air temperature, minimum air temperature, and total precipitation recorded during the last ten years (2010–2019). The methodology followed, based on state-space models, shows flexibility to allow the integration of new data, updating in real time the model, and the incorporation of covariates that are important to explain the process in analysis.

Keywords: Exploratory data analysis · Meteorological variables · Imputation · Irrigation · Time series

1 Introduction

The multiple effects of climate change on the planet are a growing concern because dry periods have become more frequent, and the increasing threats associated with meteorological phenomena have deeply affected water availability and the water cycle, which has a fundamental role in regulating the ecosystem.

The rise in temperatures that have occurred in recent years and the drought episodes that have been faced constitute, in a global context, economic, social, and environmental unbalances.

J. Machado et al. (Eds.): icieng 2022, LNME, pp. 197–208, 2023.
https://doi.org/10.1007/978-3-031-09360-9_17

Agriculture is one of the sectors undergoing great impact from the disruption of weather systems because this economic sector is directly dependent on weather conditions: approximately 69% of the world's freshwater is used in agriculture [1]. There is much that can be done to save water, and this collective effort is of the utmost importance to our planet.

Some solutions have been proposed to increase water availability, such as desalination. However, this solution has a major disadvantage because, in most cases, this technique uses a lot of energy and thus contributes to a increasing greenhouse gas emissions if the energy source is not renewable [2,3].

The excess of water in the soil is responsible for significant water waste, and therefore we have to create solutions to mitigate the dwindling of water resources. Understanding the behavior of humidity in the soil allows, among other advantages, efficient planning of water use, particularly in irrigation systems.

This study is carried out in the context of the project "TO CHAIR - The Optimal Challenges in Irrigation" (`systec.fe.up.pt/projects/FCT-TOCHAIR/`), which aims to understand and analyze the behavior of humidity in the soil in order to find optimal solutions to improve the efficiency of daily water use in irrigation systems [4].

The main goal of this work is to study the behavior and perform an exploratory statistical and critical analysis of three daily weather variables, namely maximum air temperature, minimum air temperature, and total precipitation, to particularly determine evapotranspiration [5] (related to the irrigation planning problem). The ultimate goal is to predict these environmental variables at a given location (in this case, a farm in the Portuguese district of Bragança), within a time horizon of 7 days.

In environmental variables, it is common to encounter missing values [6,7]. In this work we will use two approaches to impute these missing values, namely the linear interpolation [8], and a linear regression analysis [9] where the independent variable has a strong association with the dependent variable.

2 Exploratory Data Analysis

Our data source are the records of the three weather variables (maximum air temperature, minimum air temperature, and total precipitation) observed at a farm located in the municipality of Bragança in northern Portugal, recorded at the farm by a portable station in the period from January 1st, 2010 to December 31st, 2019 on a daily basis.

In Fig. 1 are represented the time series of the meteorological variables under study. To begin with, the presence of a strong seasonal pattern in the series of maximum and minimum air temperatures is noticeable. In addition, the series of total precipitation shows great variability with several extreme values. Furthermore, all three meteorological variables present missing values that have to be addressed.

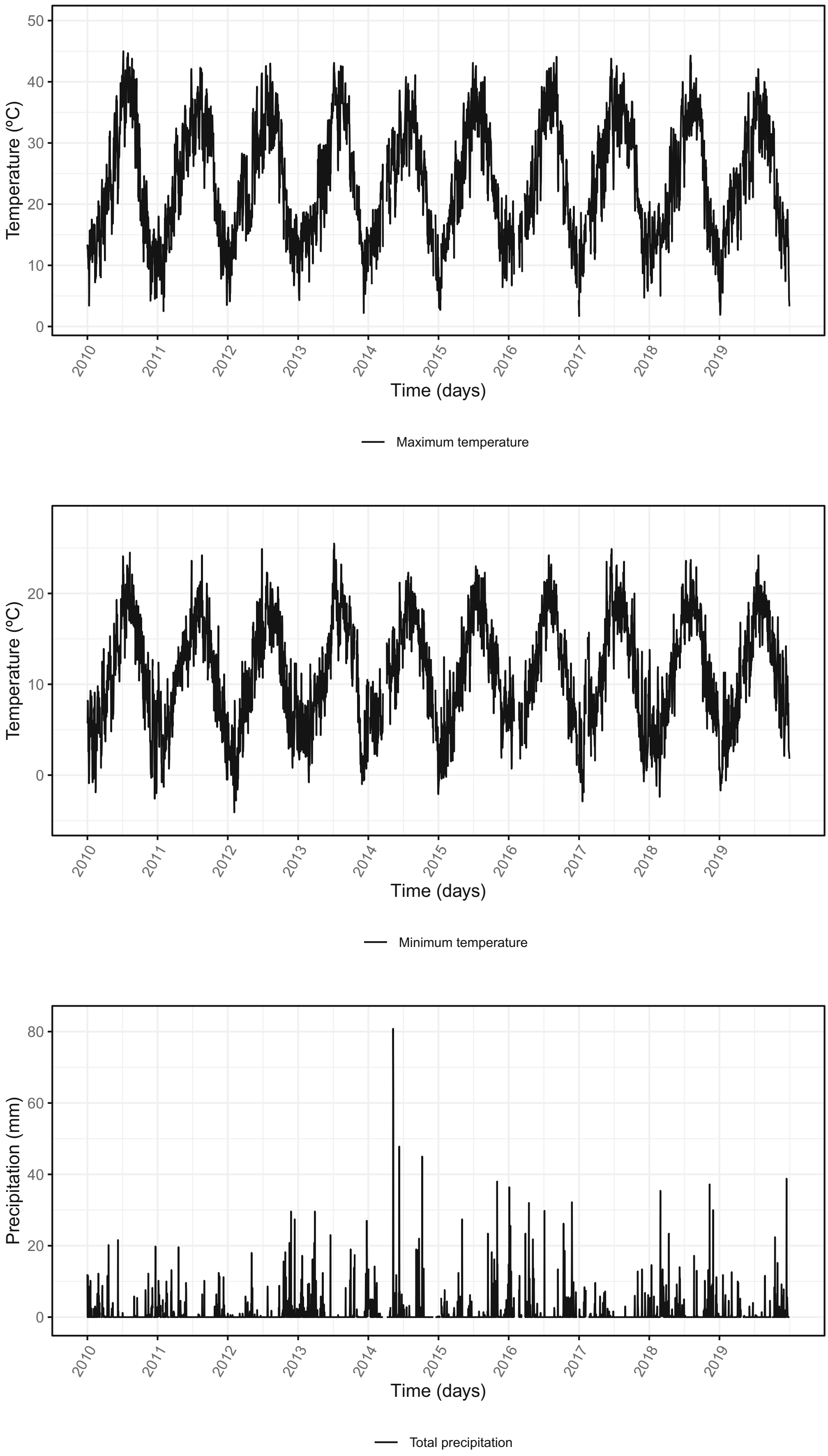

Fig. 1. Graphical representation of daily meteorological time series: maximum temperature, minimum temperature, and total precipitation.

The boxplots of each of the three variables under study are shown in Fig. 2, where both variables of maximum and minimum air temperature seem to have a symmetrical distribution, besides not presenting outliers, unlike the variable of total precipitation, which presents many outliers, highlighting the highest value on the order of 80 mm, and seems to have a positively skewed (right-skewed) distribution, which can be seen in the boxplot of total precipitation without outliers (Fig. 2, in the lower right corner). In the study, these outliers of total precipitation were not removed, they were just not displayed in Fig. 2 to be able to visualize and analyze the boxplot.

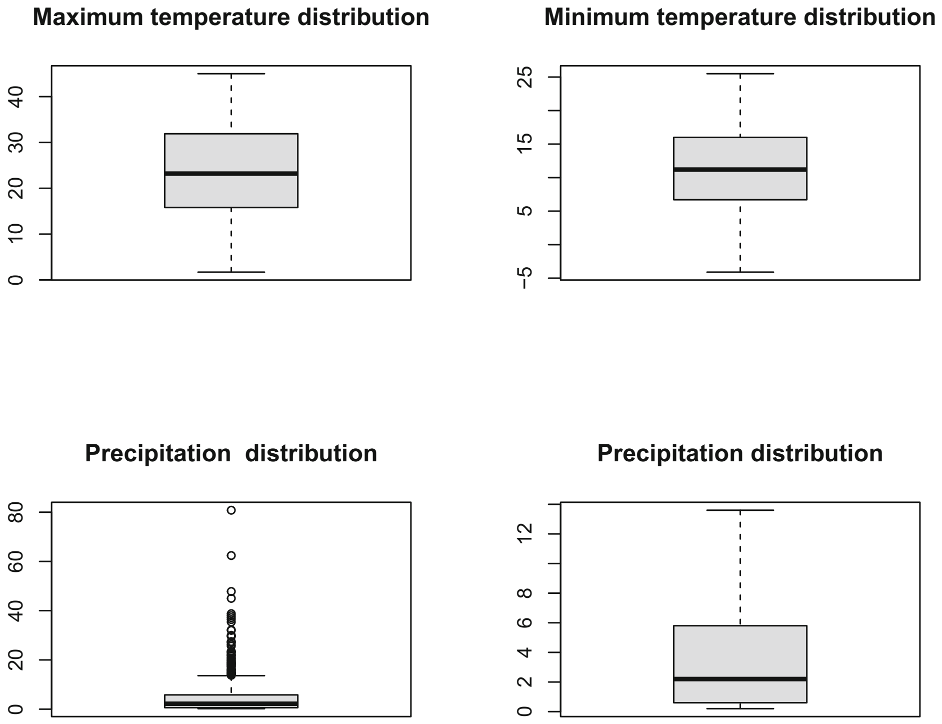

Fig. 2. Boxplot of meteorological time series: maximum temperature (°C), minimum temperature (°C), and total precipitation (mm) with and without outliers.

2.1 Exploratory Data Analysis by Year

Table 1 shows the descriptive statistics of maximum air temperature by year. In general, the maximum temperature averages between 2010 and 2019 do not present great differences, with 2017 being the hottest year in this period recording an average temperature of 25.34 °C. The highest temperature was recorded in 2010, a year which also presented a higher standard deviation (10.79 °C). 2018 was the only year where there are no gaps, unlike 2016, which presents 43 missing values (NA's). In total, this variable presents 122 NA's.

Table 1. Descriptive statistics of maximum air temperature by year.

Year	Min.	Max.	Range	1st Quart.	Median	3rd Quart.	Mean	Std deviation	NA's (122)
2010	3.40	45.00	41.60	14.15	20.65	33.48	23.32	10.79	5
2011	2.50	42.30	39.80	16.00	25.60	32.50	24.19	9.51	4
2012	4.10	43.00	38.90	16.30	22.80	31.20	23.54	9.12	1
2013	2.20	43.10	40.90	14.88	21.90	32.15	23.20	9.86	9
2014	5.90	41.10	35.20	16.75	25.30	30.35	24.09	8.33	38
2015	2.70	43.10	40.40	15.75	23.00	31.75	23.52	9.56	2
2016	3.20	44.10	41.40	16.65	23.10	32.75	24.52	9.70	43
2017	1.70	43.80	42.10	18.15	26.20	33.10	25.34	9.36	10
2018	5.00	44.30	39.30	15.20	21.10	32.30	23.37	9.33	0
2019	1.90	42.10	40.20	16.45	22.70	30.70	23.43	9.00	10

As regards the minimum temperature, it appears that 2012 presented the lowest minimum temperature recorded (−4.10 °C). Also in 2012, about 25% of the observations were below 5.70 °C. The pattern about missing values repeats itself as in the maximum temperature (see Table 2). The exploratory analysis of total precipitation (Table 3) only considered the days on which precipitation occurred. There was a total 882 days (25.75% of the data). From Table 3, the minimum value recorded was 0.20 mm for all observed years. 2014 stands out as presenting the maximum value of precipitation (80.80 mm), which is clearly visible in Fig. 1; the annual standard deviation indicates a higher variability in the year 2014 as well. In 2018, approximately 50% of the days had total precipitation greater than 2.90 mm.

Table 2. Descriptive statistics of minimum air temperature by year.

Year	Min.	Max.	Range	1st Quart.	Median	3rd Quart.	Mean	Std deviation	NA's (122)
2010	−2.60	24.50	27.10	6.68	10.65	16.10	10.99	6.08	5
2011	−1.30	24.20	25.50	6.90	11.40	15.00	11.10	5.36	4
2012	−4.10	24.90	29.00	5.70	10.50	15.70	10.30	6.18	1
2013	−1.00	25.50	26.50	5.80	10.10	15.80	10.56	6.28	9
2014	−2.10	22.30	24.40	7.95	12.10	16.10	12.01	5.06	38
2015	−1.50	23.00	24.50	7.15	11.20	15.90	11.29	5.84	2
2016	0.70	24.20	23.50	7.90	11.80	16.45	11.96	5.29	43
2017	−2.90	24.90	27.80	6.45	11.40	16.90	11.40	6.52	10
2018	−2.40	23.70	26.10	6.80	10.80	16.80	11.27	5.78	0
2019	−1.70	24.20	25.90	5.90	11.40	15.75	10.99	5.99	10

Table 3. Descriptive statistics of precipitation by year.

Year	Min.	Max.	Range	1st Quart.	Median	3rd Quart.	Mean	Std deviation	NA's (122)
2010	0.20	21.60	21.40	0.80	2.10	4.90	3.71	4.57	5
2011	0.20	19.60	19.40	0.40	1.60	4.70	3.18	3.79	4
2012	0.20	29.60	29.40	0.40	1.20	3.50	3.65	6.02	1
2013	0.20	29.60	29.40	0.60	2.60	7.60	5.10	6.36	9
2014	0.20	80.80	80.60	0.80	2.60	6.75	7.17	13.77	38
2015	0.20	38.00	37.80	0.40	1.90	4.50	4.05	6.43	2
2016	0.20	36.40	36.20	0.80	2.80	8.20	6.41	8.24	43
2017	0.20	13.40	13.20	0.80	2.20	5.15	3.39	3.28	10
2018	0.20	37.20	37.00	1.00	2.90	7.90	5.31	6.72	0
2019	0.20	38.80	38.60	0.60	2.00	4.70	3.94	2.50	10

2.2 Exploratory Data Analysis by Month

Another way to analyze time series data is to consider monthly subseries. Table 4 shows the descriptive statistics of maximum air temperature in degrees, by month. The average maximum temperature is higher in spring/summer. On the other hand, January presents the lowest average maximum temperature (12.07 °C). The standard deviation indicates higher variability in May, June, and October.

Table 4. Descriptive statistics of maximum air temperature by month.

Month	Min.	Max.	Range	1st Quart.	Median	3rd Quart.	Mean	Std deviation	NA's (122)
Jan.	1.70	20.50	18.80	9.80	12.70	14.70	12.07	3.80	15
Feb.	2.50	24.90	22.40	12.62	15.25	17.70	15.20	3.60	32
Mar.	7.90	30.30	22.40	15.70	18.40	22.22	18.85	4.30	30
Apr.	10.80	32.80	22.00	18.12	22.00	25.98	22.18	4.89	14
May	15.10	38.00	22.90	22.50	27.30	30.50	26.67	5.13	1
June	19.00	43.80	24.80	27.60	31.40	35.50	31.47	5.37	0
July	22.80	45.00	22.20	32.95	35.85	38.60	35.60	4.22	0
Aug.	24.60	44.30	19.70	32.82	36.15	38.70	35.78	3.94	0
Sept.	20.20	44.10	23.90	29.10	32.30	34.90	31.91	4.53	2
Oct.	10.60	36.80	26.20	20.60	24.30	28.00	24.52	5.29	10
Nov.	4.20	23.80	19.60	13.90	16.65	18.80	16.34	3.56	0
Dec.	2.20	21.40	19.20	9.70	12.85	14.93	12.33	3.70	18

Regarding amplitudes, October stands out with an amplitude of 26.20 °C. Only the months of June, July, August, and November have no missing values. Regarding the minimum air temperature (Table 5), January and February present the lowest average minimum temperature (4.41 and 4.38 °C, respectively). In July, about

25% of the days registered minimum temperatures above 20.30 °C. February presented the lowest minimum temperature recorded (−4.10°C). Again, the pattern of missing values repeats itself as in the maximum temperature.

Table 5. Descriptive statistics of minimum air temperature by month.

Month	Min	Max	Range	1st Quart.	Median	3rd Quart.	Mean	Std deviation	NA's (122)
Jan.	−2.90	13.80	16.70	1.70	4.40	6.90	4.41	3.47	15
Feb.	−4.10	15.70	19.80	2.10	4.35	6.57	4.38	3.46	32
Mar.	0.70	13.90	13.20	4.70	6.55	8.43	6.59	2.67	30
Apr.	3.00	16.80	13.80	7.90	9.50	11.38	9.53	2.36	14
May	4.80	23.50	18.70	10.50	12.40	14.40	12.34	2.93	1
June	8.70	24.90	16.20	13.38	15.45	17.80	15.67	3.08	0
July	10.70	25.50	14.80	16.92	18.60	20.30	18.60	2.65	0
Aug.	12.40	24.50	12.10	17.20	18.50	20.00	18.50	2.20	0
Sept.	10.20	22.90	12.70	14.43	16.30	17.88	16.00	2.39	2
Oct.	5.00	20.00	15.00	10.38	12.40	14.50	12.34	2.90	10
Nov.	−1.10	17.40	18.50	5.77	8.20	10.50	8.02	3.45	0
Dec.	−2.60	14.20	16.80	2.90	5.60	7.60	5.38	3.30	18

Table 6 shows the descriptive statistics of total precipitation by month. Similar to the analysis of this variable by year, only the days on which precipitation occurred were considered. Variability is higher in February, with a standard deviation of 13.46 mm. Regarding the amplitudes, May stands out with an amplitude of 80.60 mm. The highest average values of total precipitation were recorded in May and September (6.39 and 6.57 mm, respectively). The behavior of the missing values is repeated, with the months of February and March standing out as having the highest number of NA's.

Table 6. Descriptive statistics of total precipitation by month.

Month	Min.	Max.	Range	1st Quart.	Median	3rd Quart.	Mean	Std deviation	NA's (122)
Jan.	0.20	36.40	36.20	0.40	2.00	5.25	3.81	3.87	15
Feb.	0.20	35.40	35.20	0.80	2.20	6.30	3.97	13.46	32
Mar.	0.20	29.60	29.40	0.80	1.80	5.80	4.15	8.27	30
Apr.	0.20	32.00	31.80	0.75	2.00	4.60	4.01	6.11	14
May	0.20	80.80	80.60	0.80	2.40	6.80	6.39	4.44	1
June	0.20	47.80	47.60	0.60	1.60	6.20	5.04	6.44	0
July	0.20	29.80	29.60	0.60	1.40	3.55	3.90	6.78	0
Aug.	0.20	17.20	17.00	0.60	1.40	3.10	3.28	6.95	0
Sept.	0.20	23.40	23.20	1.05	4.90	9.85	6.57	7.51	2
Oct.	0.20	45.00	44.80	0.55	2.90	7.60	5.94	6.98	10
Nov.	0.20	38.00	37.80	0.70	2.20	5.80	5.20	6.28	0
Dec.	0.20	38.80	38.60	0.60	2.00	4.50	4.12	5.85	18

Figure 3 shows the boxplots of maximum air temperature, minimum air temperature, and total precipitation by month. It can be seen that precipitation shows several discrepant values, unlike the maximum/minimum temperature variables. The monthly subseries of maximum/minimum air temperatures show a clear seasonal pattern, reaching higher temperatures in July and August. The median of the maximum/minimum air temperatures gradually decreases from summer to winter. The median of total precipitation is higher in September.

The analysis revealed the existence of random missing values. This situation is quite common in real data and it may derive from multiple factors - e.g., sensors malfunction at the weather station, data loss, etc. Dealing with time series with missing data can be quite problematic because most modeling approaches are designed for complete time series. To solve this problem, a possible solution is to do imputation, which consists in filling these missing values based on some properties of the data.

2.3 Predictive Model

In this work, two types of imputation were chosen. One of them was conducted through a database with daily records of maximum air temperature and minimum air temperature between January 1 to December 31, 2019 in Vila Real, located about 50 km from the site under study (i.e., a farm in the Portuguese district of Bragança). The Vila Real database was selected because these are a good predictor of the location of the farm. The interpolation was done through a linear interpolation model fitted to a **linear regression model** given by $Y_t = \alpha + \beta X_t + e_t$, where Y_t and X_t are the maximum/minimum air temperature recorded at the farm and in Vila Real, respectively, and t is the time, in days. This method was not considered for imputation of missing data for total precipitation since there is no strong correlation for this variable between the two locations: the farm and Vila Real.

Therefore, the model was adjusted for the maximum and minimum air temperatures variables and the coefficients of determination obtained were 0.9340 and 0.9089, respectively. It should be noted that the imputation of data through this method was done only for the year 2019 given the availability of data. For all other missing values (between 2010 and 2018 for maximum/minimum air temperature, and for all total precipitation data), it was decided to impute through **linear interpolation** [8,10].

Figure 4 shows the series of maximum air temperature, minimum air temperature, and total precipitation with the imputed values (in red). Therefore, for the series of the maximum/minimum air temperatures from 2010 up to and including 2018, the imputation was done through linear interpolation, and in 2019 through the linear regression model between the farm data and the Vila Real data. For the total precipitation variable, imputation of all the missing data was done through linear interpolation.

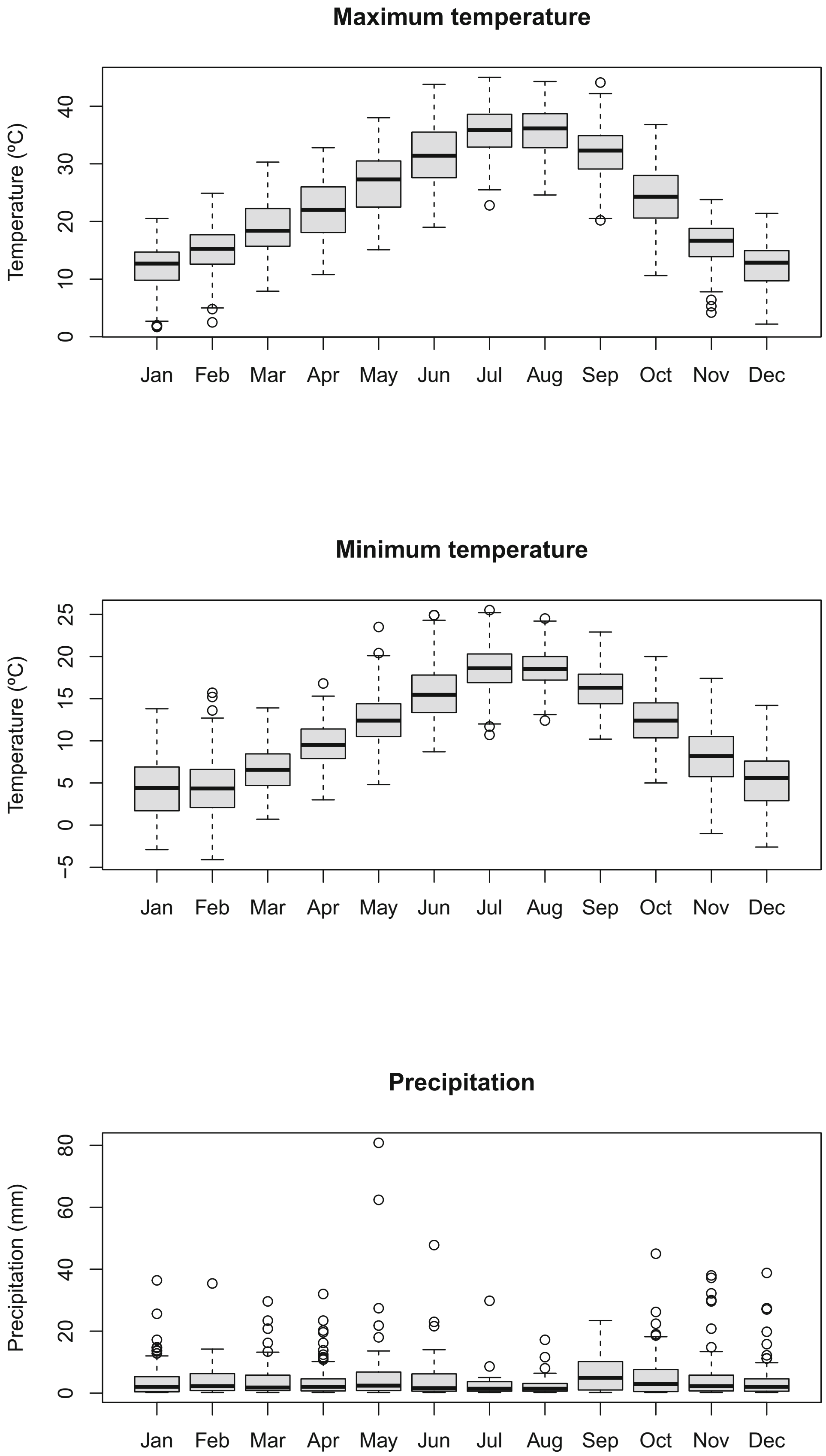

Fig. 3. Boxplots of maximum air temperature, minimum air temperature, and total precipitation by month.

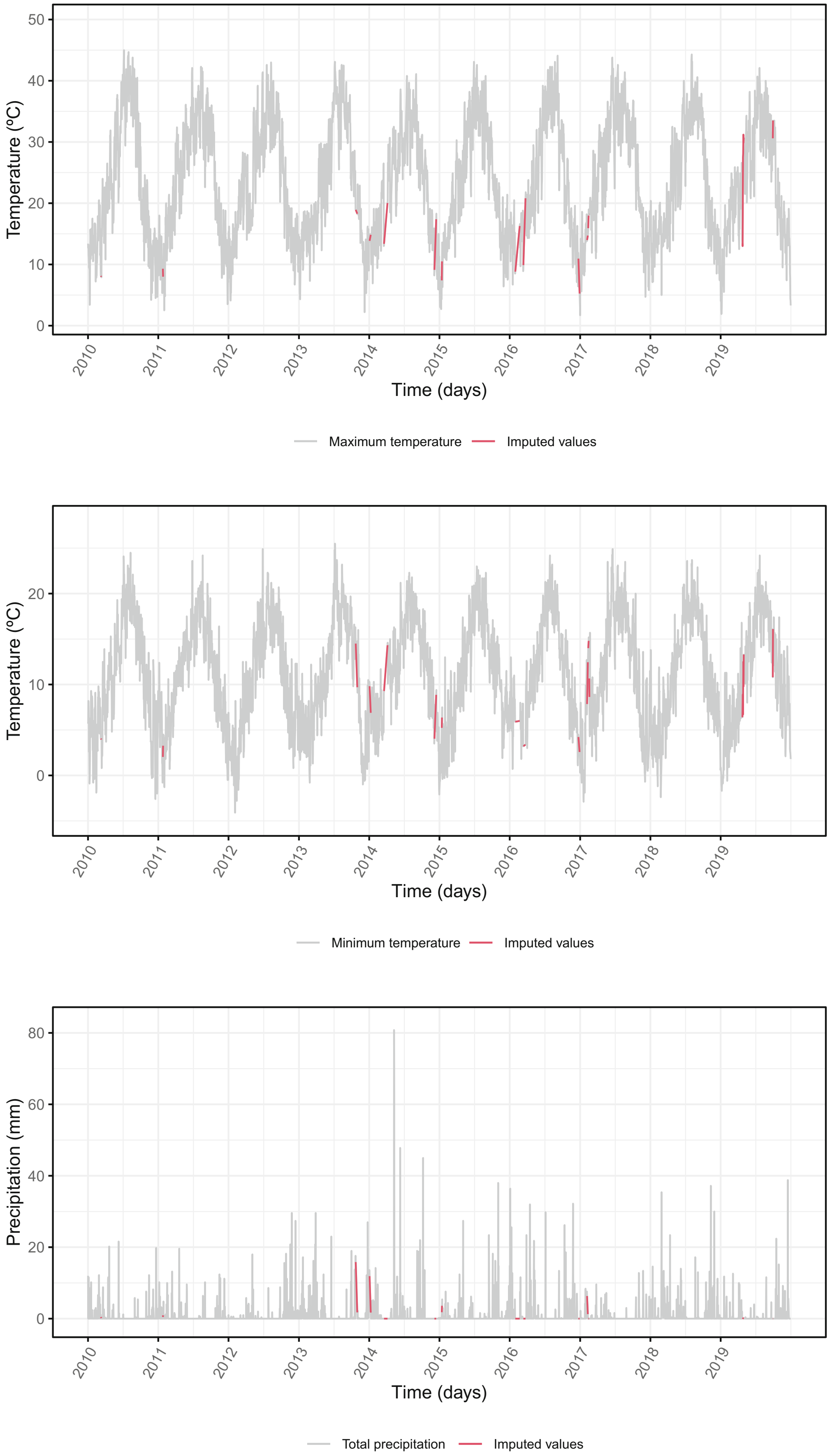

Fig. 4. Graphical representation of daily meteorological time series (in black), and imputed values (in red).

3 Discussion and Future Work

In this work, we address the daily irrigation problem of minimizing water consumption. For this, we selected three weather variables: maximum air temperature, minimum air temperature, and total precipitation. This type of preliminary data analysis performed in this work is fundamental in any time series analysis because it allows the study and identification of data behavior that will provide relevant information for the application of appropriate models, with the ultimate goal of forecasting, allowing for better modeling and forecasting.

After the exploratory analysis, the modeling and forecasting step follows. Given the behavior of meteorological time series, two types of models stand out: TBATS (Trigonometric Seasonal, Box-Cox Transformation, ARMA errors, Trend and Seasonal Components) and state-space models. These models have successfully been applied to modeling and foresight scenarios of environmental and weather variables.

TBATS models are suitable for modeling time series with strong trend and complex seasonal patterns (multiple seasonal patterns), which are predominant features in weather series. For example, in [11], they modeled the minimum air temperature series (recorded on the same farm) using TBATS models, which proved to be very appropriate for the daily forecasting approach.

State-space models are a very flexible class of models as they allow several features of the data to be integrated explicitly into the model structure itself, and they also allow predictions to be made and updated in real time whenever a new observation is integrated into the model [12–14]. In addition, they allow incorporating covariates that are important to explain the process. In [15], they proposed a model with a state-space representation that establishes a stochastic linear relationship between the maximum temperature observed at a farm (the same database) and the h-days ahead forecast ($h = 1, \ldots, 6$) produced from a weather website (as a covariate).

In order to have more precise results, for future work it is suggested to extend the study to other regions of Portugal.

Acknowledgements. This research was partially supported by The Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. This research was partially financed by Portuguese funds through Portuguese Foundation for Science and Technology (“Fundaçào para a Ciência e a Tecnologia” - FCT), within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM. This research was financed by national funds through FCT (Fundação para a Ciência e a Tecnologia) through the individual PhD research grant UI/BD/150967/2021 of CMAT-UM. This work was also supported by the Portuguese Foundation for Science and Technology (FCT) in the framework of the Strategic Funding UID/FIS/04650/2019, UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM. The authors are granted for the access to the meteorological data provided by the Direção Regional de Agricultura e Pescas do Norte. This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

1. United Nations: The United Nations World Water Development Report 2021: Valuing water. UNESCO, Paris (2021)
2. Ghalavand, Y., Hatamipour, M.S., Rahimi, A.: A review on energy consumption of desalination processes. Desalination Water Treat. **54**(6), 1526–1541 (2015). https://doi.org/10.1080/19443994.2014.892837
3. Zarzo, D., Prats, D.: Desalination and energy consumption. What can we expect in the near future? Desalination **427**, 1–9 (2018). https://doi.org/10.1016/j.desal.2017.10.046
4. Costa, C., Gonçalves, A.M., Costa, M., Lopes, S.: Forecasting temperature time series for irrigation planning problems. In: Proceedings of the 34th International Workshop on Statistical Modelling (IWSM 2019) (2019). http://hdl.handle.net/10773/26437
5. Lopes, S., Pereira, R., Pereira, P., Caldeira, A., Fonte, V.: Optimal control applied to an irrigation planning problem: a real case study in Portugal. Int. J. Hydrol. Sci. Technol. **9**(2), 173–188 (2019). https://doi.org/10.1504/IJHST.2019.098161
6. Gonçalves, A.M., Costa, M.: Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. Stoch. Environ. Res. Risk Assess. **27**, 1021–1038 (2012). https://doi.org/10.1007/s00477-012-0640-7
7. Costa, M., Monteiro, M.: Statistical modelling of water quality time series - the river vouga basin case study. In: Research and Practices in Water Quality. IntechOpen (2015). https://doi.org/10.5772/59062
8. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice, 2nd edn. OTexts, Melbourne (2018)
9. Di Franco, G.: An alternative procedure for imputing missing data based on principal components analysis. Qual. Quant. **48**(3), 1149–1163 (2013). https://doi.org/10.1007/s11135-013-9826-4
10. Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J.: Comparison of different Methods for Univariate Time Series Imputation in R (2015)
11. Gonçalves, A.M., Costa, C., Costa, M., Lopes, S.O., Pereira, R.: Temperature time series forecasting in the optimal challenges in irrigation (TO CHAIR). In: Gaspar-Cunha, A., Periaux, J., Giannakoglou, K.C., Gauger, N.R., Quagliarella, D., Greiner, D. (eds.) Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences. CMAS, vol. 55, pp. 423–435. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-57422-2_27
12. Harvey, A.C.: Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, New York (2009)
13. Costa, M., Alpuim, T.: Parameter estimation of state space models for univariate observations. J. Stat. Plan. Inference **140**(7), 1889–1902 (2010)
14. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications: With R Examples, 4th edn. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-52452-8
15. Costa, M., Pereira, F.C., Gonçalves, A.M.: Improving short-term forecasts of daily maximum temperature with the Kalman filter with GMM estimation. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12952, pp. 552–562. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86973-1_39