JOURNAL OF ENVIRONMENTAL HYDROLOGY

Open Access Online Journal of the International Association for Environmental Hydrology

VOLUME 29 2021

STATISTICAL MODELS FOR CLIMATE CHANGE IN THE BODIES OF WATER OF THE MEDITERRANEAN

Xavier Jiménez-Albán[1]
Antonio Monleón-Getino[2]

[1]Polytechnic University of Catalonia, School of Mathematics and Statistics, Barcelona, Spain
[2]Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain

The Paris Agreement brought with it an important global commitment: Prevent the global temperature from exceeding 1.5°C. Barcelona has signed multiple internal commitments such as reducing greenhouse gas emissions and becoming a carbon-neutral city. However, the water supply in the city will have negative consequences. This research presents novel statistical models based in machine learning that allow predicting different variables and their interactions in a system that is interrelated: the temperature in the city, the rainfall or the flow level of the Llobregat river that is the source of drinking water for the most of Barcelona.

L'Accord de Paris a entraîné un important engagement mondial: empêcher que la température mondiale ne dépasse 1,5 ° C. Barcelone a signé de multiples engagements internes tels que la réduction des émissions de gaz à effet de serre et le développement d'une ville neutre en carbone. Cependant, l'approvisionnement en eau de la ville aura des conséquences négatives. Cette recherche présente de nouveaux modèles statistiques, basés sur l'apprentissage automatique, qui permettent de prédire différentes variables et leurs interactions dans un système interdépendant: la température dans la ville, les précipitations ou le niveau de débit de la rivière Llobregat qui est la source d'eau potable l'eau pour la plupart de Barcelone.

INTRODUCTION

This study will focus on some climatic variables and analyze their behavior and interaction. We will focus on variables such as temperature and precipitation of the city of Barcelona and the stream flow variables of the Llobregat river that partly supplies the city of Barcelona. It is of great importance to foresee how climate change will affect the Mediterranean region and in particular the flow of the Llobregat River and potential impacts on Barcelona (Barcelona, Ajuntament de, 2014)

An analysis and a comparison is made with the so-called Representative Concentration Pathway (RCP) scenarios. Four pathways have been considered for climate modeling and research. These describe four possible future climate scenarios, all of which are considered possible taking into account the number of greenhouse gases that are emitted in the coming years. The four RCPs: RCP2.6, RCP4.5, RCP6, and RCP8.5, are labelled after a possible range of radiative forcing values in the year 2100.

METHODOLOGY

Information and data from different sources have been collected and efforts have been made to ensure that all variables correspond in time and place to the one of interest. Different techniques and functions were used with R and Python. These data contain daily information since 1996 from 2018 about streamflow (m^3/s) of the stream gauges of Balsareny, Castellar de n'Hug, Castelbell and el Vilar and Guardiola de Berguedà and reservoir capacity of the La Baells Reservoir (hm^3). belonging to the Llobregat river basin (Rodríguez et al, 2014). All this information has been filtered to obtain a mean of all stream gauges and in this way we have obtained the streamflow variable (caudal).

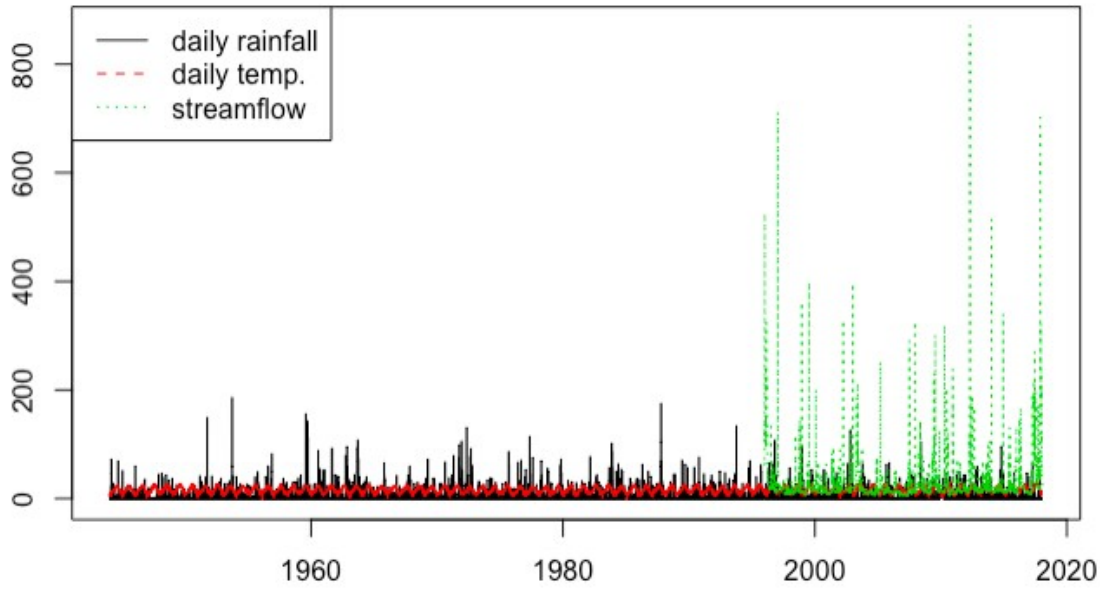

Figure 1. Data collected from AEMET, 1944-2018 and ACA 1996-2018

The variables and their interactions have been analyzed in a univariate and multivariate manner. For the univariate part, we study the data series with the Box–Jenkins method, analyzing different autoregressive moving average (ARMA) or autoregressive integrated moving average (ARIMA) models to find the best fit of a time-series model to past values of a time series. Having done this, the results have been compared with the analysis of the Representative Concentration Pathway (RCP) scenarios. The dependent variable is the streamflow, measured in (m^3/s). These data correspond to measurements taken in the stream gauges of Balsareny, Castellar de n'Hug, Castelbell i el Vilar and

Guardiola de Berguedà. Independent variables such as the level of monthly rainfall measured in *mm*, temperature measured in ^{o}C and the entry of water to the La Baells reservoir measured in m^3/s give the ability to observe interactions and correlations in the study.

For the multivariate part, an attempt was made to carry out an analysis with the series of precipitation, Llobregat river flow and temperature.

For the machine learning part, we used the so called Ensemble methods that are meta algorithms. We have explicitly used Gradient Boosting, a generalization of Boosting that trains many models in a sequential manner (Schapire, 2013). For this end we analize the dataset with **R**, specifically with the package `BDSbiost3` and the `Anuket()` function with the purpose of performing a prediction (Regression type). Obtaining prediction models with so much data is very computationally demanding and must be done on suitable computers. For the processing of this data, we have used the BIOST3 server which is a high-performance Linux (Ubuntu 18) (HPC: 40 cores Xeon SP 4114 2.2 GHz)

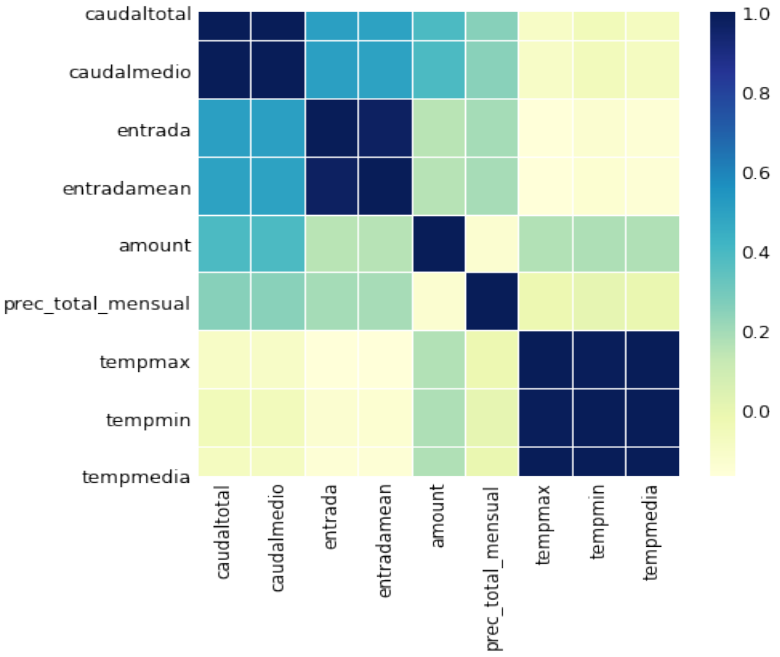

Figure 2. Correlation matrix of the variables under study

The correlation between variables of different nature is not strong, so we face a problem to perform a multivariate time series analysis. This seems to make sense since the streamflow of the Llobregat river not only depends on precipitation, but on factors such as the snow, ice, and glaciers. The rivers of snow melt are characterized because the precipitations are in the form of snow, and therefore the waters are retained during the winter. The rivers fed by snow melt are characteristic of the high mountains. On the other hand, the variables entrada and streamflow are moderately correlated because entrada corresponds to the entry in m^3/s to reserves in the La Baells reservoir, one of the three reservoirs that is part of the Ter-Llobregat system. The reservoir's function is to control the waters of the Llobregat river. When there is rain, the floodgates open and let the excess water pass, consequently the flow increases.

Stationary time series are those that oscillate around a constant level. If a time series has an overlapping behavior, that is to say, repeated over time, we will say it is seasonal. Hence, time series with trends, or with seasonality, are non-stationary. The next plot shows the monthly rainfall in Barcelona from 1996 to 2018. It is observed that the values of the series oscillate around a constant value, but some months have systematically more rainfall than others.

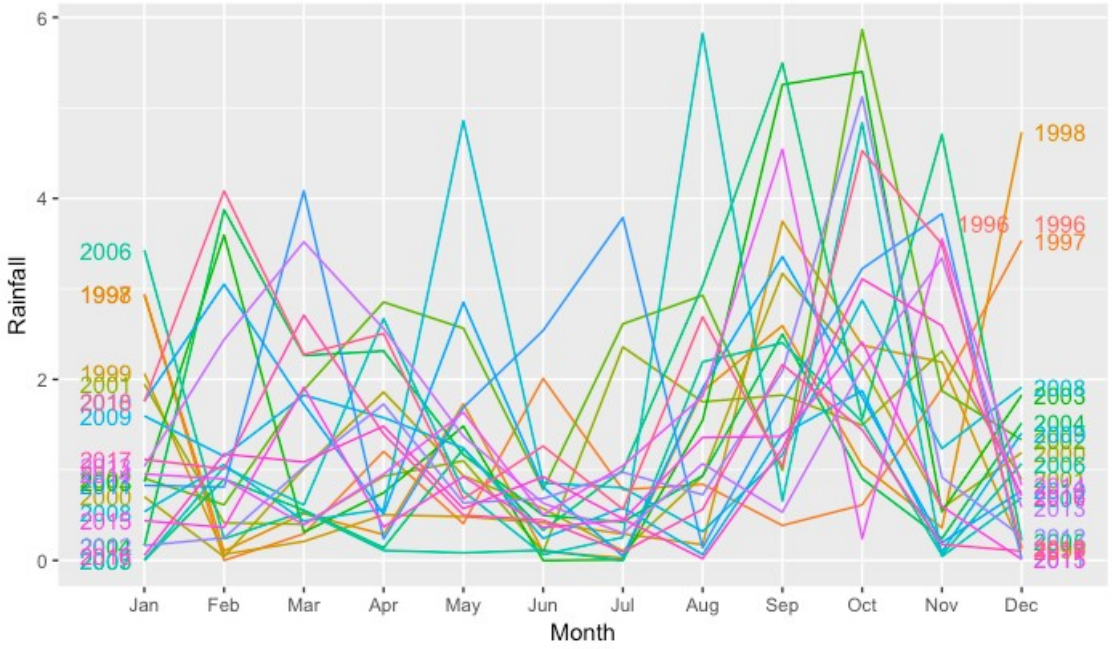

Figure 3. Monthly series of the amount of rainfall in Barcelona in the years 1996 to 2018

We say that z_t follows a first order autoregressive process or an AR(1) if has been generated by:

$$z_t = c + \phi z_{t-1} + \epsilon_t \quad (1)$$

where c and $-1 < \phi < 1$ are constants to be determined, ϵ_t is a white noise process with σ^2 variance. ϵ_t is known as the innovation. For example, consider z_t is the amount of water at the end of the month in the La Baells reservoir. During the month an amount of $c + \epsilon_t$ arrives at the reservoir, where c is the mean of this amount and ϵ_t is the innovation, a variable with zero mean and constant variance that makes the entrance varies from one period to another. if a proportion $(1-\phi) z_{t-1}$ is spent every month and the proportion ϕz_{t-1} remains. The amount of water at the end of the month follows the process $z_t = c + \phi z_{t-1} + \epsilon_t$

There is an ARIMA stationary model (also called Box-Jenkins model) when differencing with autoregression and a moving average model are combined. The full model can be written as:

$$z't = c + \phi_1 z'_{t-1} + \ldots + \phi_p z'_{t-p} + \theta 1 \epsilon_{t-1} + \ldots + \theta q \epsilon_{t-q} + \epsilon_t \quad (2)$$

where z' is the differenced series. In the right hand the predictors included lagged values of z_t and lagged errors. An ARIMA(p,d,q) model consists of p, order of autorregresive part; q, order of move average part, and d, the degree of first differencing. For example an ARIMA$(0,1,0)$ with a constant, is given by $z'_t = c + z'_{t-1}$

A seasonal ARIMA model is formed by the original ARIMA model plus an additional seasonal term. It is written as ARIMA$(p,d,q)(P,D,Q)_m$

We have used the Akaike information criterion to estimate the best model for the streamflow. Given a set of candidate models for the data, the preferred model is the one with the minimum value in the AIC. The analysis of this model along with some more and the value of its AIC can be summarized in the following table:

ARIMA order	AIC
ARIMA(1,0,0)(2,1,1)[12]	343.71
ARIMA(2,0,0)(2,1,1)[12]	344.01
ARIMA(1,0,1)(0,1,1)[12]	341.96
ARIMA(2,0,1)(1,0,0)[12]	344.85
ARIMA(1,0,4)(2,1,0)[12]	367.32

Taking into account this criterion, the preferred model would be the ARIMA(1,0,1)(0,1,1)[12], however, there are significant spikes in the residuals ACF and the model fails the Ljung-Box test.

We verify the residuals with two more models with 36 lags and both fail the test. None of the models considered here pass all residual tests. However, this happens regularly and it is common to use the model that best fits, even if neither of them passes the tests.

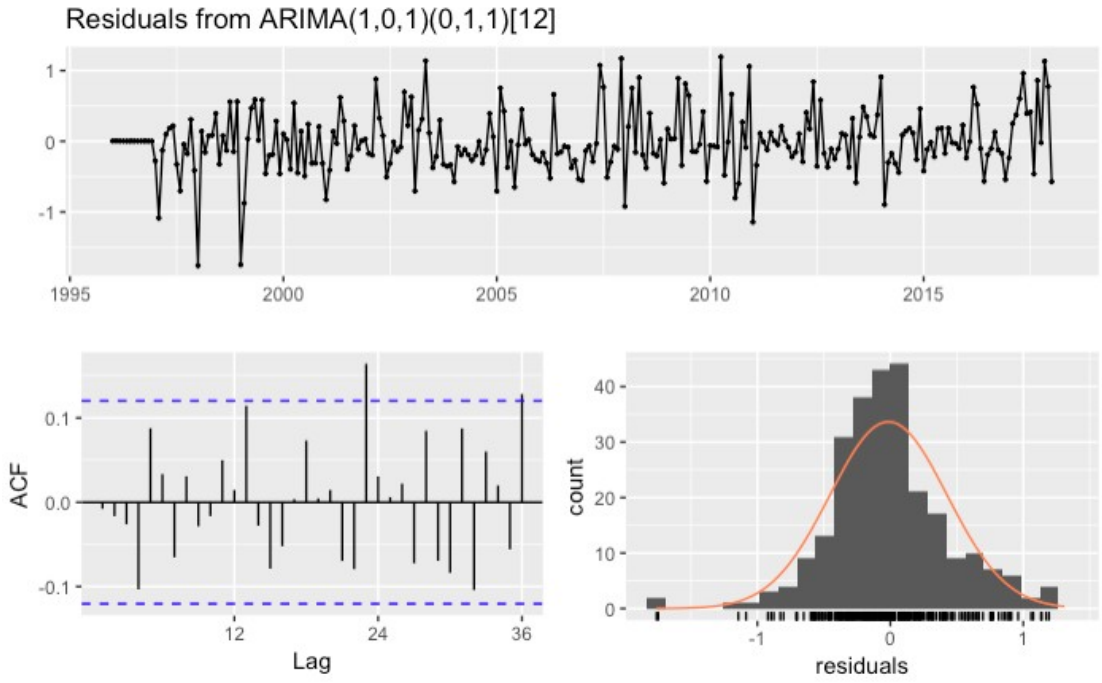

Figure 4. Residuals from ARIMA(1,0,1)(0,1,1)[12]

To help us understand the accuracy of the models, we compare predicted streamflow to real streamflow of the time series, and we set forecasts to start at 2015–01–01 to the end of the data.

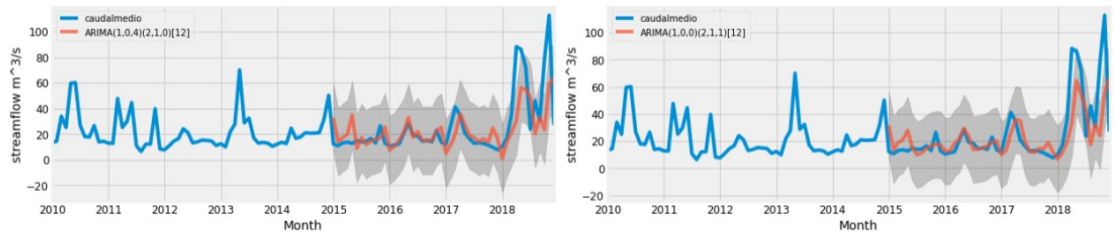

Figure 5. Forecasts: ARIMA(1,0,4)(2,1,0)[12] and ARIMA(1,0,0)(2,1,1)[12], lag = 36

We have the two models here. The line plot is showing the observed values compared to the rolling forecast predictions. To a large extent, both models fit the true values very well

Two models are presented here that seem to fit well with the reality of the data, Forecasts of the ARIMA(1,0,4)(2,1,0)[12] and ARIMA(1,0,0)(2,1,1)[12] are shown in the Figure:

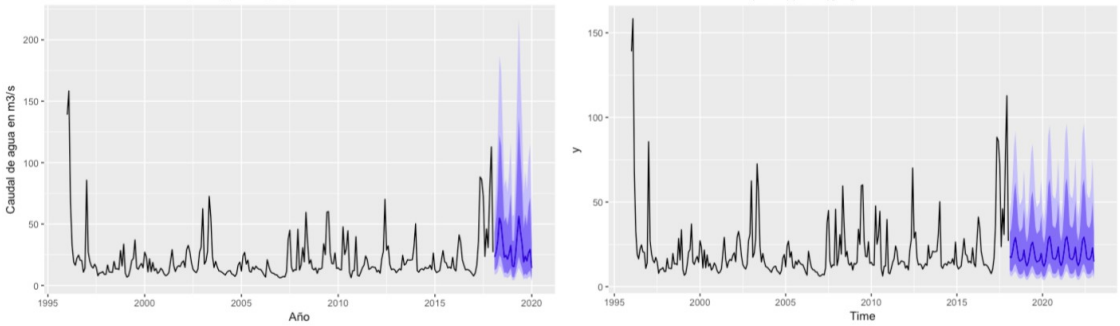

Figure 6. Forecasts: ARIMA(1,0,4)(2,1,0)[12] and ARIMA(1,0,0)(2,1,1)[12], lag = 36

Multivariate time series analysis considers simultaneously multiple time series. Each variable depends not only on its past values but also has some dependency on other variables. For this analysis let $z_t=(z_{1t},\ldots,z_{kt})$ be a k-dimensional time series observed at equally spaced time points. For example, let z_{1t} be the variable (Entry in m^3/s to the La Baells reservoir) the and z_{2t} the variable . By studying z_{1t} and z_{2t} jointly we can observe the dependence between these two variables. In this particular case k = 2 and it seems that variables and are positively correlated. The sample correlation is 0.51.

In the case of precipitation and streamflow variables the correlation is 0.25 and in the first instance this makes us think that the streamflow could not be predicted based on precipitation because as we said in the introduction, the flow of a river depends on many other factors besides the precipitation.

For a Vector Auto Regression analysis, we can suppose that each time series in the system influences each other, it means that the series can be predicted with past values of itself along with other series in the system. To test, if this relationship is possible, we can use the Granger's Causality Test, before even building the model. Granger's causality tests the null hypothesis that the coefficients of past values in the regression equation is zero.

The four pathways selected for climate modeling and research, describe different climate future, all of which are considered possible depending on how much greenhouse gases are emitted in the years to come. Four pathways have been selected for climate modeling and research, which describe different climate futures, all of which are considered possible depending on how much greenhouse gases are emitted in the years to come. The four RCPs, namely RCP2.6, RCP4.5, RCP6, and RCP8.5 (Van Vuuren et al. 2011)

RCP2.6: assumes that global annual GHG emissions (measured in CO_2 -equivalents) peak between 2010–2020, with emissions declining substantially thereafter

RCP4.5: emissions peak around 2040, then decline

RCP6: emissions peak around 2080

RCP8.5: emissions continue to rise throughout the 21st century

For this research we got data from the RCP4.5 and RCP8.5 scenarios, which correspond to intermediate and high emissions respectively. RCP8.5 was developed using the IIASA Integrated Assessment Modeling Framework that encompasses detailed representations of the principal GHG-emitting sectors—energy, industry, agriculture, and forestry (Clarke et al. 2007).The framework combines a careful blend of rich disciplinary models that operate at different spatial resolutions that are interlinked and integrated into an overall assessment framework. Integration is achieved through a

series of hard and soft linkages between the individual components, to ensure internal scenario consistency and plausibility. In the *RCP4.5* scenario CO_2 concentrations are slightly above than those of RCP6.0 until after mid-century; however emissions peak earlier (around 2040). Thus the CO_2 concentration reaches 540 ppm by 2100.

To obtain a prediction model of the future streamflow and to perform the cross-validation we have used the BIOST3 server. It is a Linux high-performance server (HPC: 40 cores Xeon SP 4114 2,2 GHz) (Monleon-Getino 2018). The model that has been implemented as follows:

```
streamflow = XGBoost(precipitation, temperature, Temp_Max*precipitation,
Temp_Min*precipitation, month)
```

We have performed two iterations with the function and and we have concluded that the best model is *Xgboost*

ANN-reg prediction & 0.02570466 & 0.07475244 \

Model	Iteration 1	Iteration 2
SVM	0.2637772	0.6714696
LM	0.01917422	0.05439776
XGBoost	0.800298	0.7961679
kernel-reg prediction	0.1957225	0.411959
ANN-reg prediction	0.02570466	0.07475244

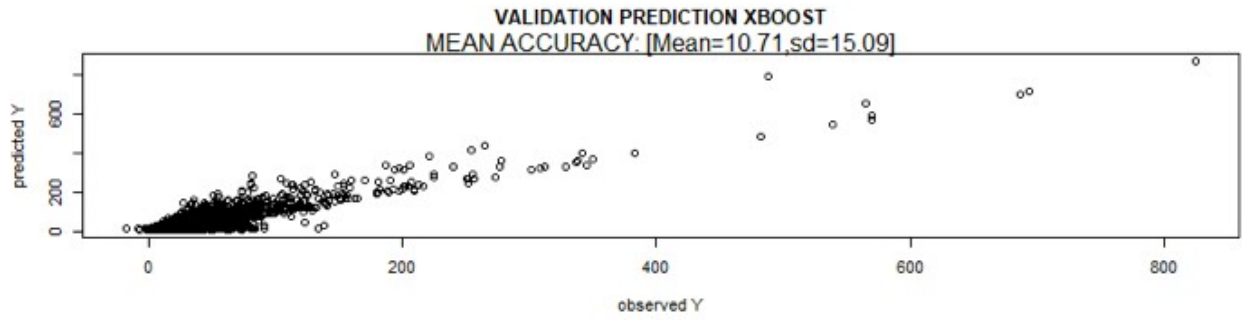

Figure 7. Mean absolute percentage error using Xboost

We can see that the XGBoost model predicts the values in a better way. The accuracy in this model is about 80%, and although the other models improved a bit, the XgBoost is still the best one. We measure the mean square error (RSME) and the absolute prediction error and it is observed that the best models are SVM, Lm, and XGBoost. Based on the coefficient of determination that provides a value of how well the observed results are replicated by the model, XGBoost has been selected as a forecasting method ($R_1^2 = 0.800298$, $R_2^2 = 0.7961679$) and a prediction model has been constructed using this method that has allowed to calculate the stream-flow from 2020 to 2100 from temperatures and precipitation.

RESULTS

For all the models studied we have obtained an increase in the average temperature in the city of Barcelona. Taking into account historical data, we have been able to make a monthly forecast with an

ARIMA(1,0,1)(1,0,1)[12]. With this model, an increase of 0.6% is expected for the year 2030 and a 6% increase in the annual average temperature for the year 2100 compared to the year 2018. That is, the temperature would increase half a degree by the year 2030 and a degree by the year 2100. These results would be more in line with a *RCP4.5*, in which the emission reduction targets of the 2015 Paris agreement are reached, that is, it would be assumed that carbon dioxide emissions would remain constant or even decrease with over the years.

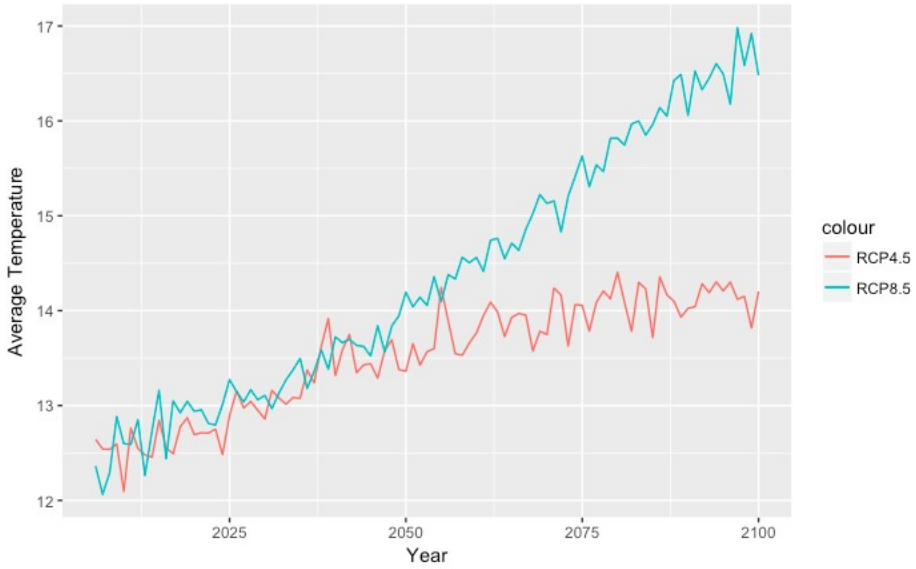

Figure 8. Forecasted average temperatures, scenarios RCP4.5 and RCP8.5

The average annual temperature in Barcelona was 17.7 °C in 2018. Taking the *RCP4.5* scenario into consideration, the temperature for the year 2050 will be 18.32 ° C, that is, it will increase slightly more than half a degree with respect to 2018. By the year 2100, it will increase to 19.5 ° C, that is to say, 8.5 % with respect to the year 2018. If we take the *RCP8.5* scenario into account, the temperature will be much higher, namely, it will rise 7.36% for the year 2050 and 20% for the year 2100, if we consider the year 2018 as a reference in which the temperature was 17.73 °C

Forecasted streamflow through univariate models: A negative trend is not so obvious for this univariate analysis, it seems that the values will remain constant over time. This can be verified by the decomposition of the additive time series shown below.

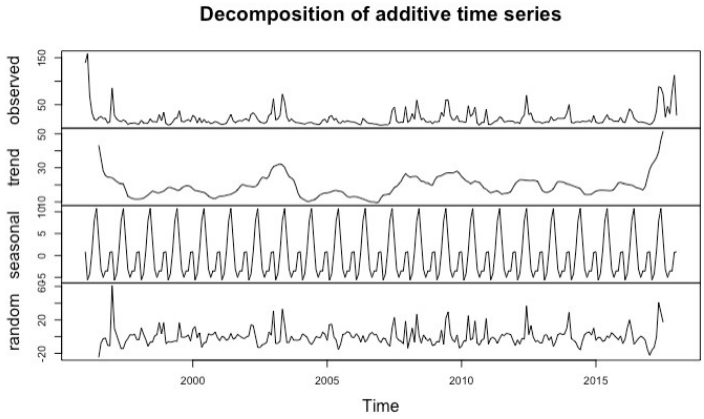

Figure 9. Decomposition of additive time series, Llobregat streamflow

Forecasted streamflow through univariate models: It must be taken into account that the correlation between precipitation and flow is 25% thus, the results could be biased. A constant tendency in the flow can be observed if the precipitation remains constant, but we cannot affirm by this model if the flow will decrease or increase if precipitation decreases or increases

Forecasted streamflow through the library `BDSbiost3`*:* Obtaining prediction models with so much data is very computationally demanding and must be done on suitable computers. For the processing of this data, we have used the BIOST3 server which is a high-performance Linux (Ubuntu 18) (HPC: 40 cores Xeon SP 4114 2.2 GHz) The XGBoosting algorithm has been used to forecast the streamflow and we can conclude that the streamflow could maintain a constant level over time. It can be said that by 2050 if we take into account the variables analyzed, Llobregat will have a flow rate of 32.5 m^3/s and by 2100, this value would have a slight downward trend since we would be talking about a flow of 32.1 m^3/s:

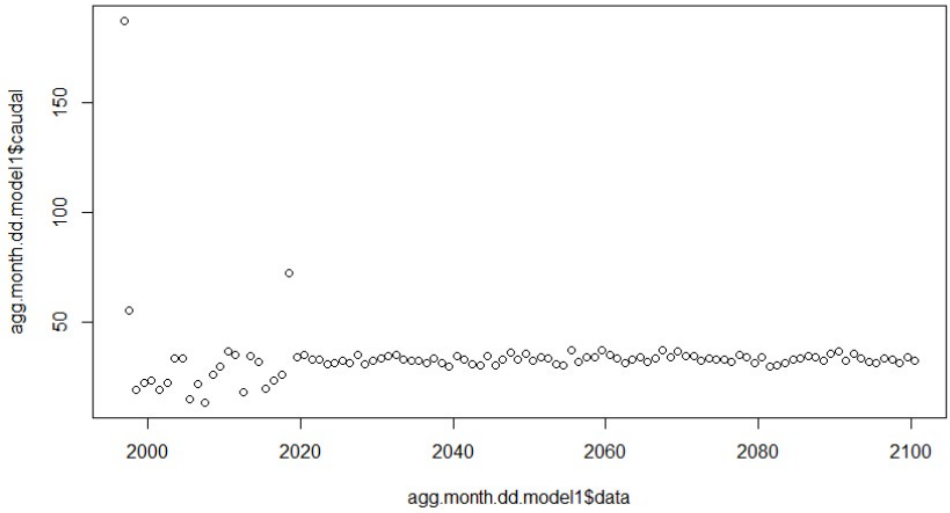

Figure 10. Forecasted streamflow through BDSbiost3 and `anuket()

Apparently, precipitation will decrease regardless of the scenario. For the RCP4.5 scenario, rainfall will decrease by 2.46% in the 1950s and 7.14% in 2100. For the RCP8.5 scenario, rainfall will decrease by 7.61% in the 1950s and by 18% in 2100

CONCLUSIONS

According to the temperature analysis, it can be corroborated that the temperature forecast in Barcelona corresponds to the RCP4.5 scenario. That is, whatever the scenario, the temperature will increase around 1.5ºC until 2050. If this year the emission reduction targets of the Paris agreement are reached, the temperature would not continue to rise, otherwise, it would rise to 3°C in 2100.

Apparently, the river flow rate will not decrease with the decrease in rainfall as it was appreciated in in the both univariate and the analysis using the High-Performance Computing BIOST3 server. This is due to the fact that the flow of the Llobregat depends on multiple factors that are outside the analysis of this study.

Although the flow of the river would apparently not decrease, this does not mean that the water resources that supply the city will not decrease. The overpopulation facing the city should be taken into account, and there are other sources of supply, whether surface and underground, which have not been part of this analysis.

As for precipitation, the results will be the other way around if we take into account the scenarios. While it is true that precipitation will decrease, a drastic change will not be observed until after the middle of the century in which precipitation will decrease by 2.5% and 7.6% for scenarios RCP4.5 and RCP8.5 respectively.

REFERENCES

Barcelona, Ajuntament de. n.d. "De Dónde Viene El Agua de La Ciudad, 2014" https://ajuntament.barcelona.cat.

Brosa, Pere, Monleon-Getino, Antonio, Mendez, Javier, and Gutiérrez, Francisco. 2019. "Turbidity Forecasting in the Delaware River" 5 (August): 01–09.

Clarke, Leon, James Edmonds, Henry Jacoby, Hugh Pitcher, John Reilly, and Richard Richels. 2007. "Scenarios of Greenhouse Gas Emissions and Atmospheric Concentrations."

Herschy, Reginald W. 2008. *Streamflow Measurement*. CRC Press.

Mauricio, José Alberto. 2007. "Análisis de Series Temporales." *Universidad Complutence de Madrid*.

Monleon-Getino, Antonio. 2018. "Machine Learning Algorithms Applied to Biosciences Ii: Examples of Classification (Discriminant) and Regresion Using the Library(BDSbiost3)."

Milly, Paul CD, Kathryn A Dunne, and Aldo V Vecchia. 2005. "Global Pattern of Trends in Streamflow and Water Availability in a Changing Climate." *Nature* 438 (7066). Nature Publishing Group: 347.

Nash, J Eamonn, and Jonh V Sutcliffe. 1970. "River Flow Forecasting Through Conceptual Models Part I—A Discussion of Principles." *Journal of Hydrology* 10 (3). Elsevier: 282–90.

Olcina Cantos, Jorge, and others. 2009. "Cambio Climático Y Riesgos Climáticos En España." Universidad de Alicante. Instituto Universitario de Geografía.

Peña, Daniel. 2005. *Análisis de Series Temporales*. Alianza.

Rodríguez, Raul, Xavier Navarro, M Carmen Casas, Jaime Ribalaygua, Beniamino Russo, Laurent Pouget, and Angel Redaño. 2014. "Influence of Climate Change on Idf

Sabater, Sergi, Isabel Muñoz, Emili García-Berthou, and Damià Barceló i Cullerés. 2014. "Multiple Stressors in Mediterranean Freshwater Ecosystems: The Llobregat River as a Paradigm." *Contributions to Science*, 161–69.

Van Vuuren, Detlef P, Jae Edmonds, Mikiko Kainuma, Keywan Riahi, Allison Thomson, Kathy Hibbard, George C Hurtt, et al. 2011. "The Representative Concentration Pathways: An Overview." *Climatic Change* 109 (1-2). Springer: 5.

Velasco Droguet, Marc, Pierre Antonie Versini, Ángels Cabello Gómez, and Antoni Barrera-Escoda. 2013. "Assessment of Flash Floods Taking into Account Climate Change Scenarios in the Llobregat River Basin." *Natural Hazards and Earth System Sciences* 13 (12). Copernicus Publications: 3145–56.

ADDRESS FOR CORRESPONDENCE
Xavier Jiménez-Albán
Universidad Central del Ecuador
Av. Universitaria
170129 Quito, Ecuador
Email: xmjimenez@uce.edu.ec