The 7th International Electronic Conference on Water Sciences

Daily Streamflow Modelling Using ML Based on Discharge and Rainfall Time Series in the Besós River Basin, Spain

Session E

Hydrological Modelling of Basins under Variable Conditions

By:

Mr. Mohamed Hamitouche, MSc | CIHEAM ZARAGOZA

Mr. Marc Ribalta, Sr. Data Scientist | Eurecat

Definitions

Model: “a simplified representation of a system at some particular point in time or space”

Modeling is the act of building a **model**.

Simulation: “Is fundamentally an imitation of a real process or system over time”

A **simulation** is the process of using a **model** to study the behavior and performance of an actual or theoretical system

(IGI Global dictionary)

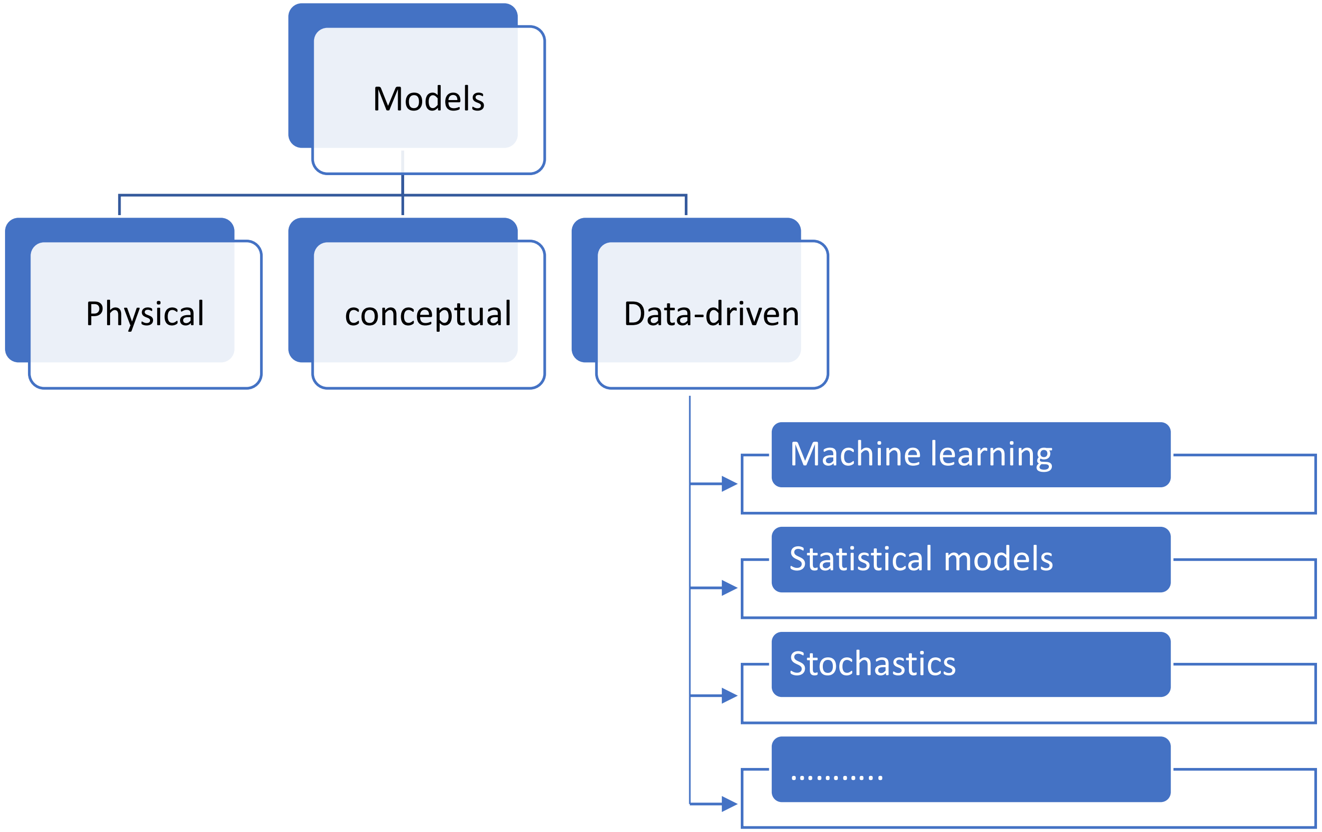
« **Prediction** is concerned with estimating the outcomes for unseen data. For this purpose, you fit a model to a training data set, which results in an estimator $\hat{f}(x)$ that can make predictions for new samples x.

Forecasting (prediction?) is a sub-discipline of prediction in which we are making predictions about the future, on the basis of time-series data. Thus, the only difference between prediction and forecasting is that we consider the temporal dimension. »

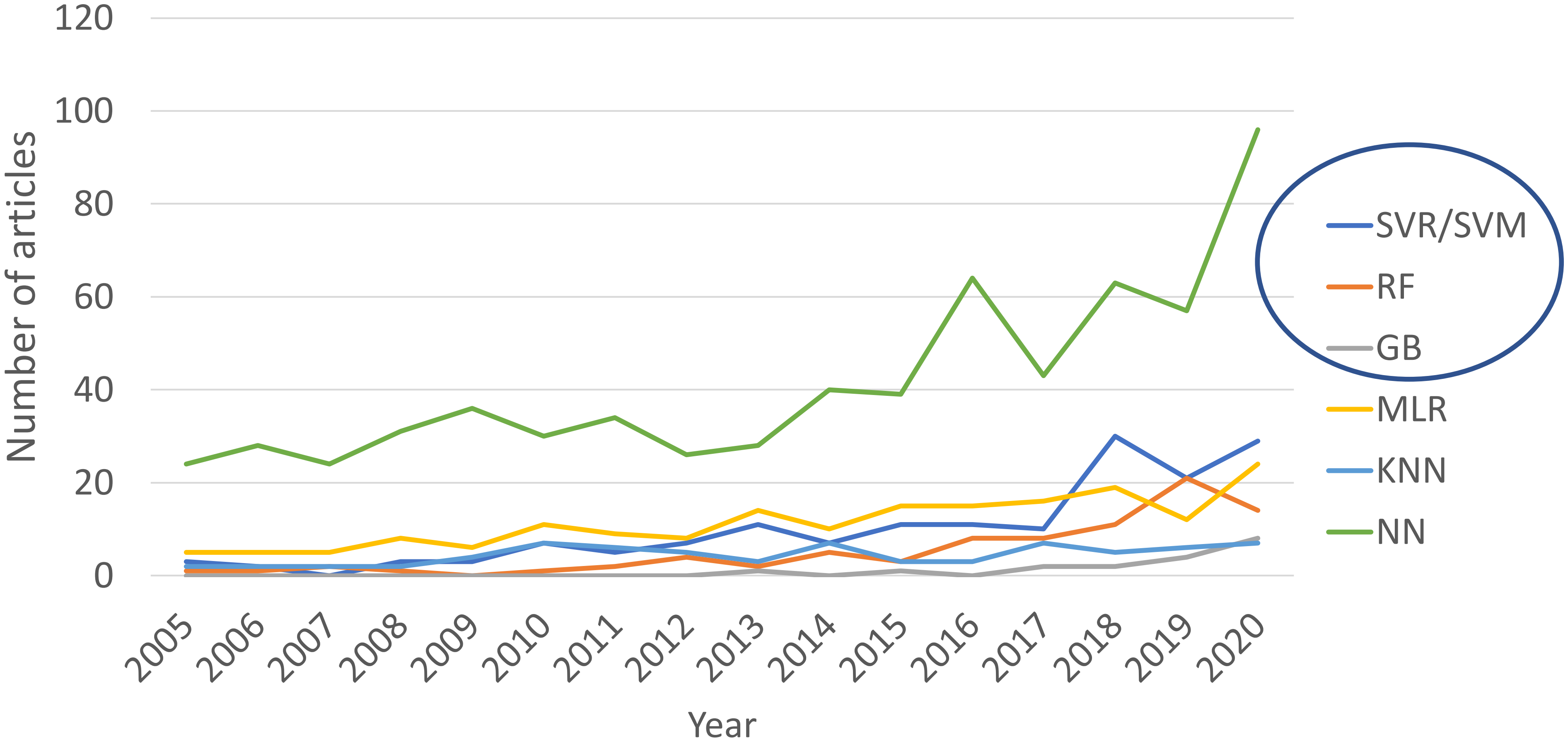
DATA SCIENCE BLOG
Understand. Implement. Succeed.

Introduction

- The development of flow forecasting or prediction models helps a lot in the management of water uses, the development of water resource distribution policies and risk management.



Literature



Machine learning models used in streamflow prediction/modelling/forecasting (Scopus)

Objective

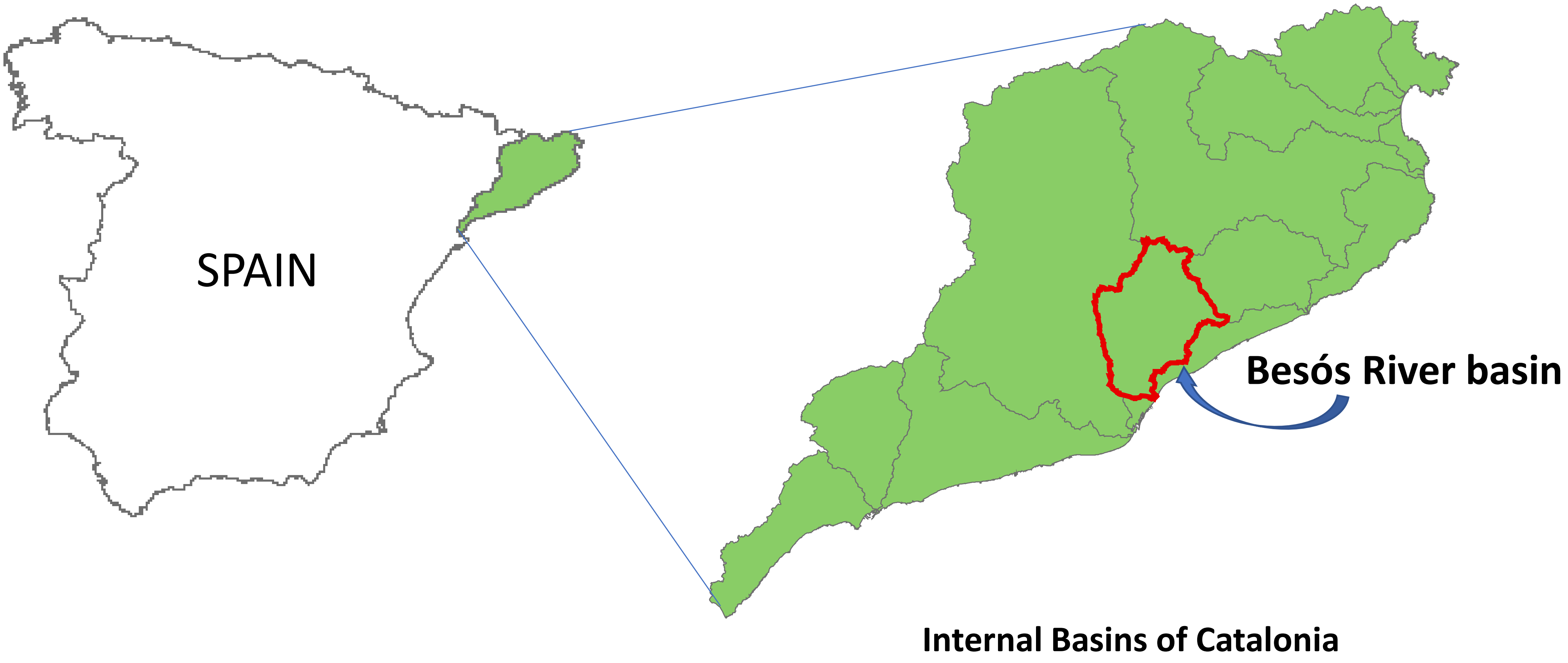
- Utilise / train the ML models to predict the daily flow at the gauging station (target) "Santa Coloma de Gramenet" in the Besós river based on rainfall and flow data

- Compare these models with MLR

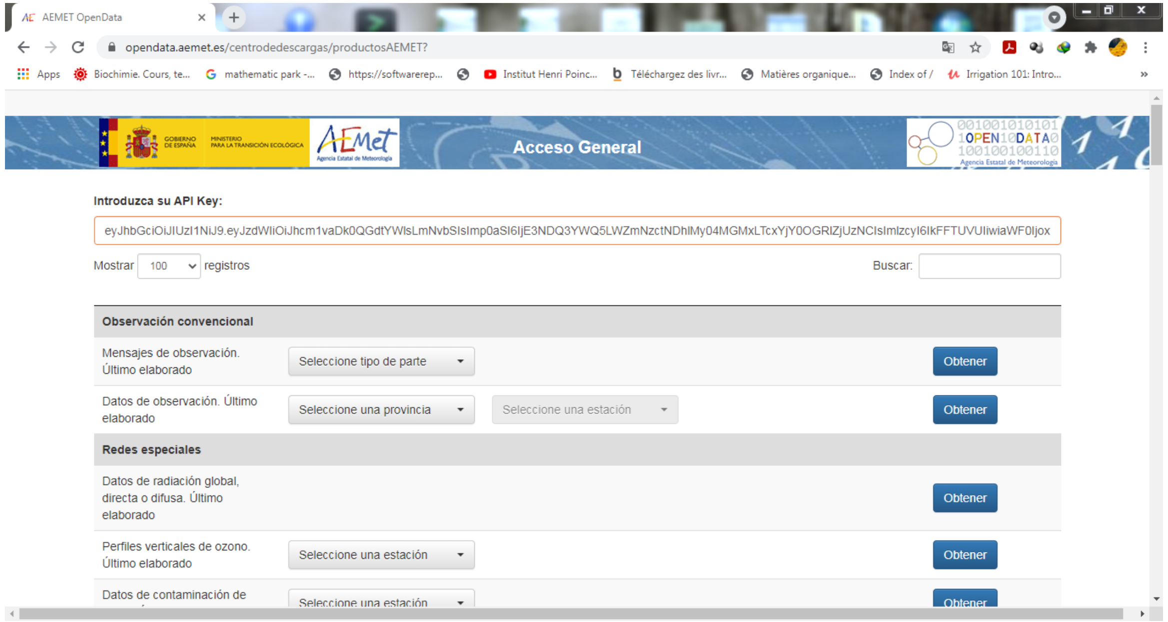
- Compare these models with each other

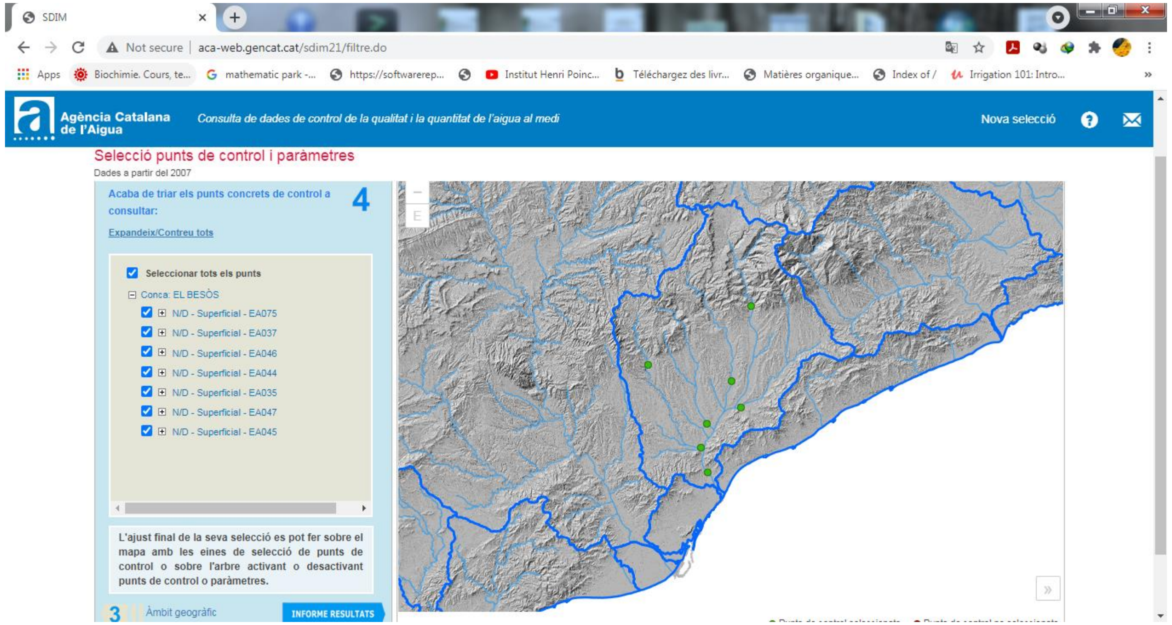
- Deduce which has been the best prediction model

Study zone



Data Collection





https://opendata.aemet.es/centrodedescargas/productosAEMET

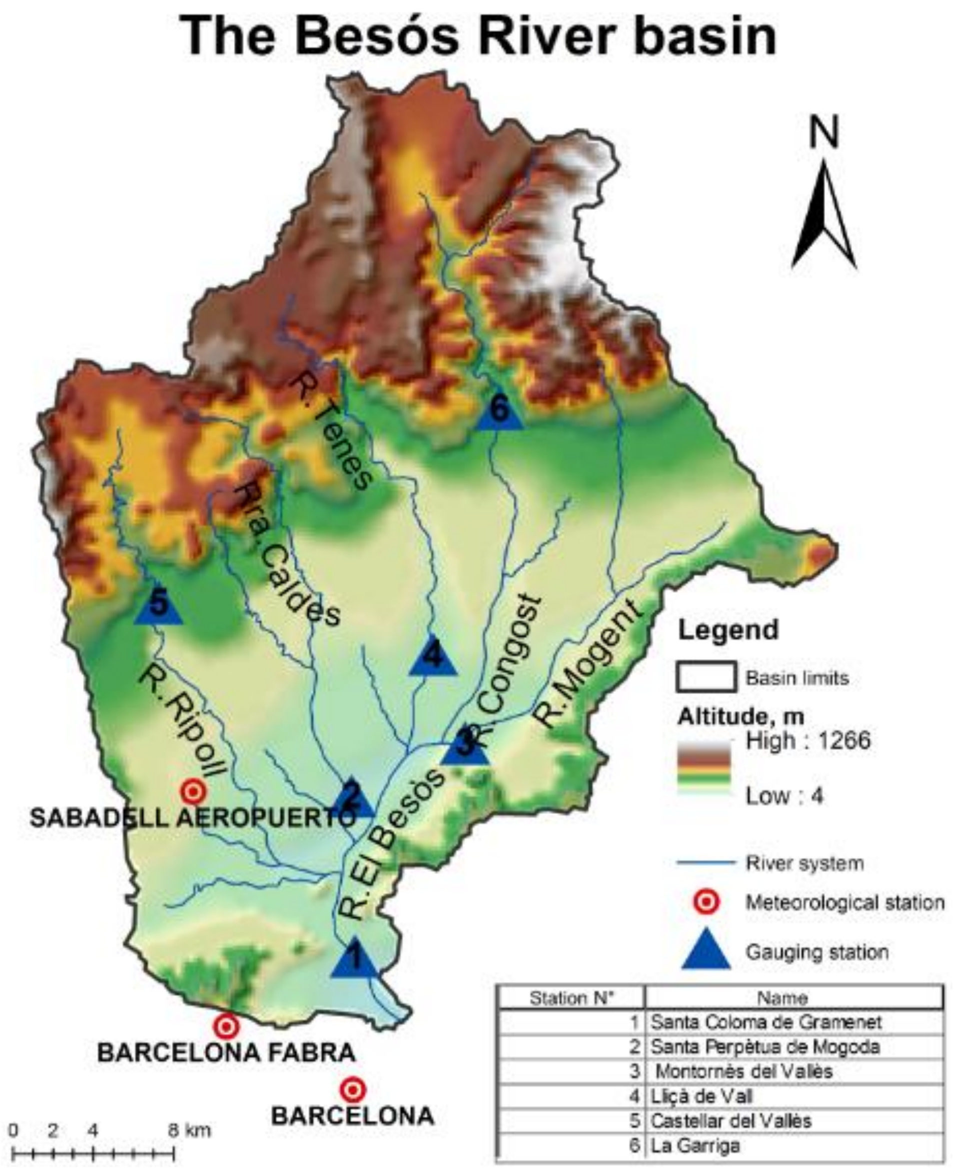
http://aca-web.gencat.cat/sdim21/

Spatial distribution

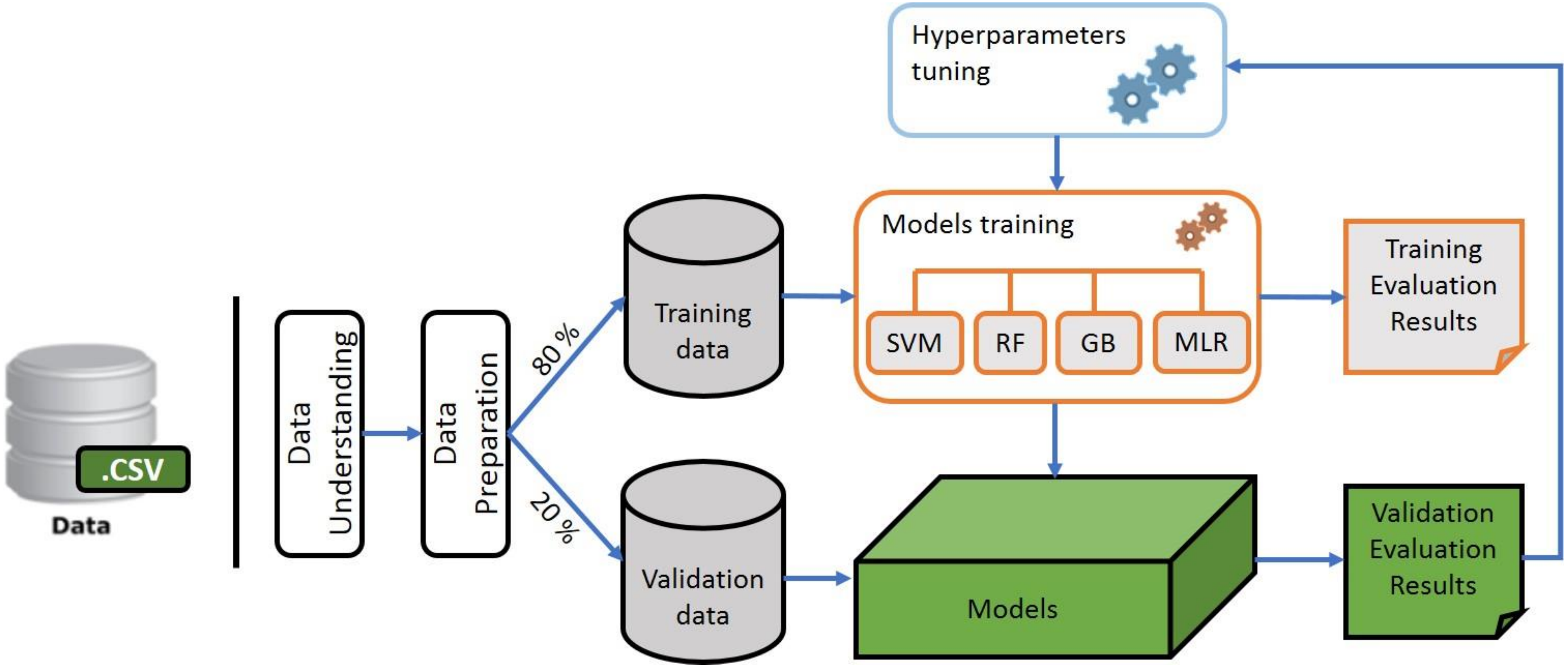
2 scenarios:

1. Without considering the antecedent flow in the target gauging station (outlet)

2. With considering the antecedent flows in the target gauging station (outlet)



Methodology



General Scheme

Prerequisite Libraries

sklearn is a free software machine learning library for Python

NumPy is a Python-based library that supports large, multi-dimensional arrays and matrices. Also, NumPy has a large collection of high-level mathematical functions that operate on these arrays.

Data Visualization

Pandas is a Python-based library written for data manipulation and analysis.

we've imported the required libraries into our Jupyter Notebook

jupyter Proyecto_indiv Dernière Sauvegarde : 08/06/2021 (modifié)

File Edit View Insert Cell Kernel Widgets Help

Exécuter Code

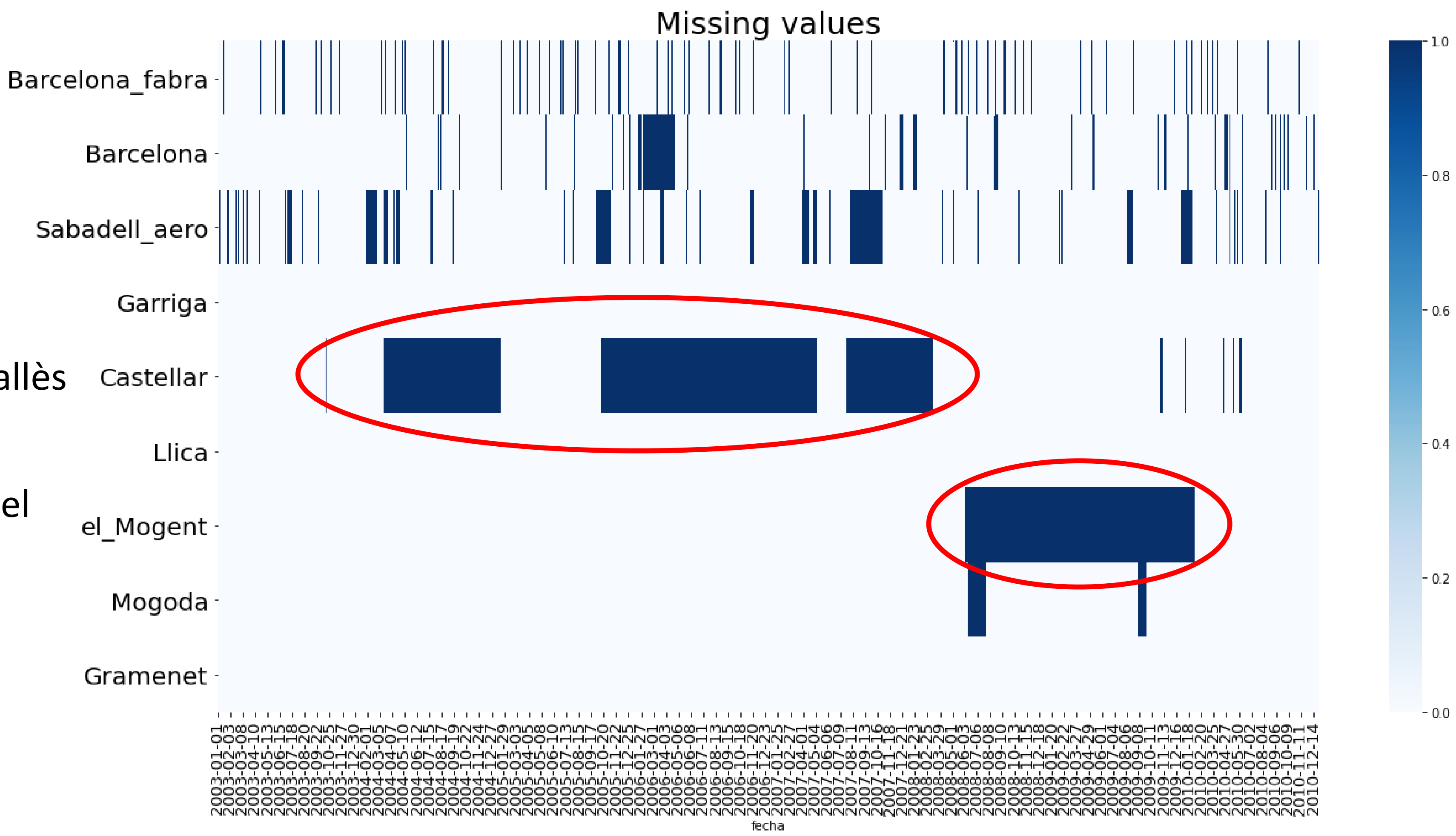
Entrée [2]:

```
import numpy as np
import pandas as pd
import os
import re
import seaborn as sns
import matplotlib.pyplot as plt
import json
%matplotlib inline
from statsmodels.graphics import tsaplots
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error, r2_score
import hydroeval as he
from sklearn.model_selection import RandomizedSearchCV, GridSearchCV
from scipy import stats
```

Data Exploration – Missing Values

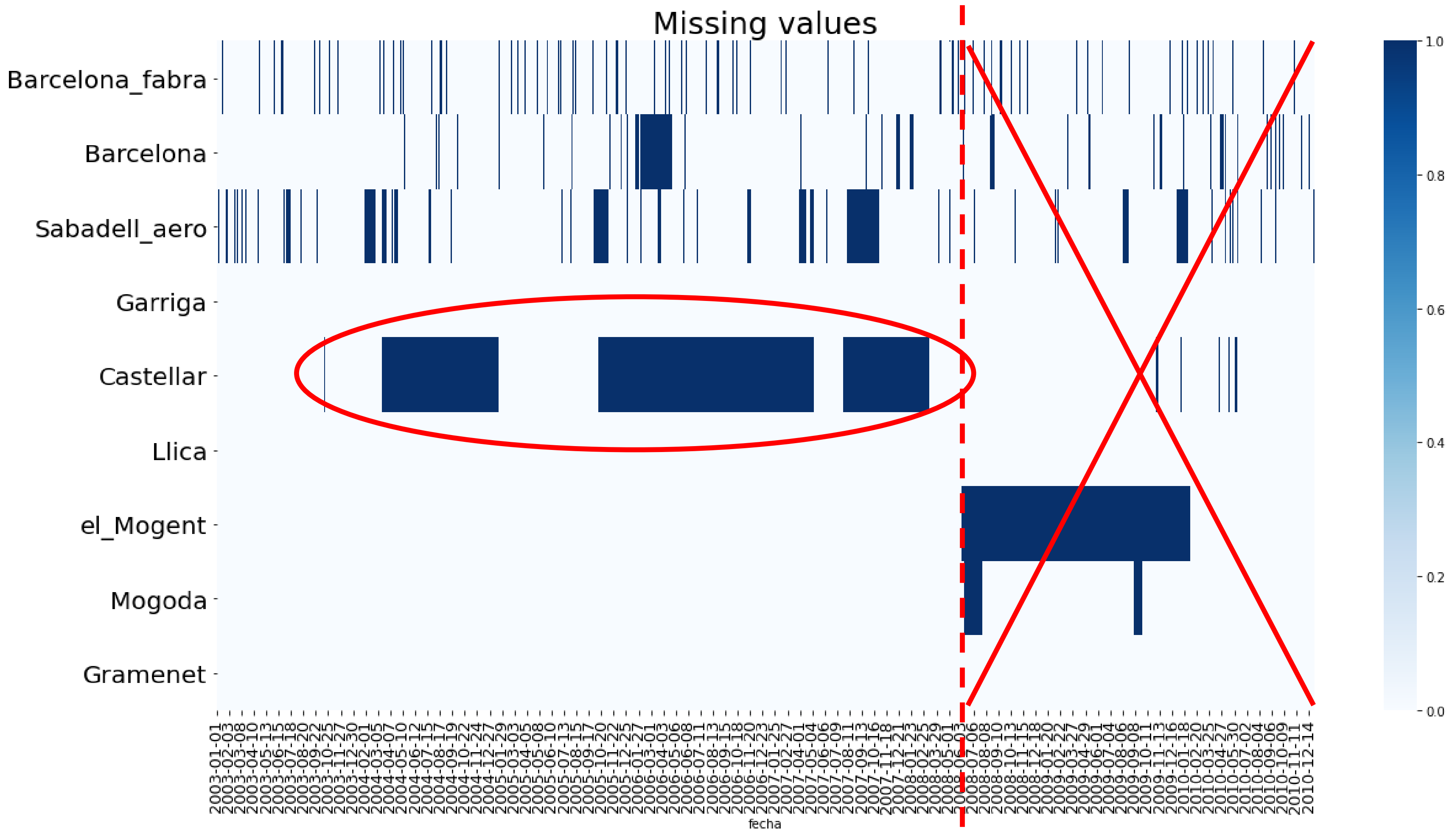
Castellar del Vallès

Montornès del Vallès



01/01/2003 ← → **31/12/2010**

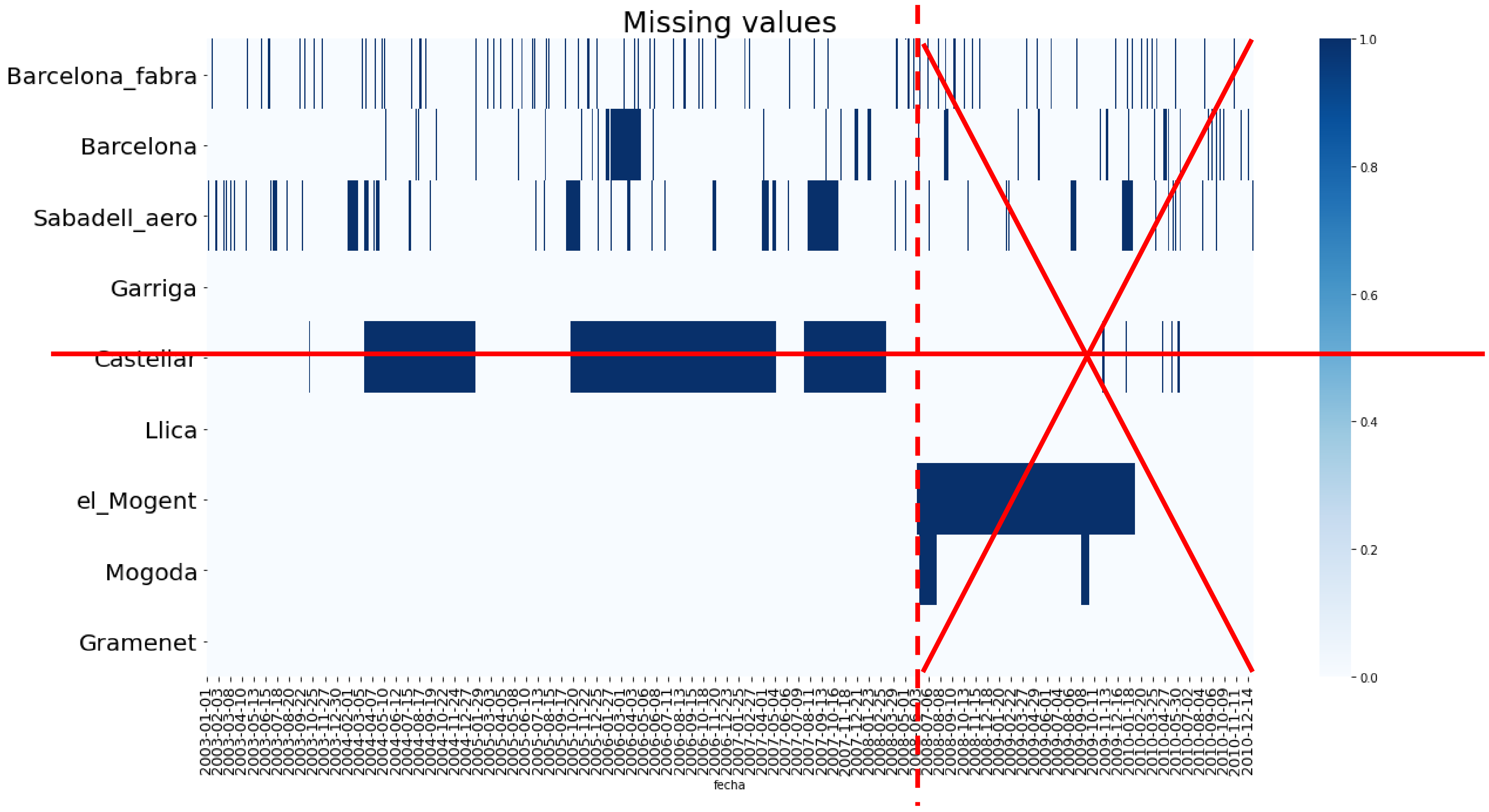
Data Exploration – Missing Values



05/06/2008

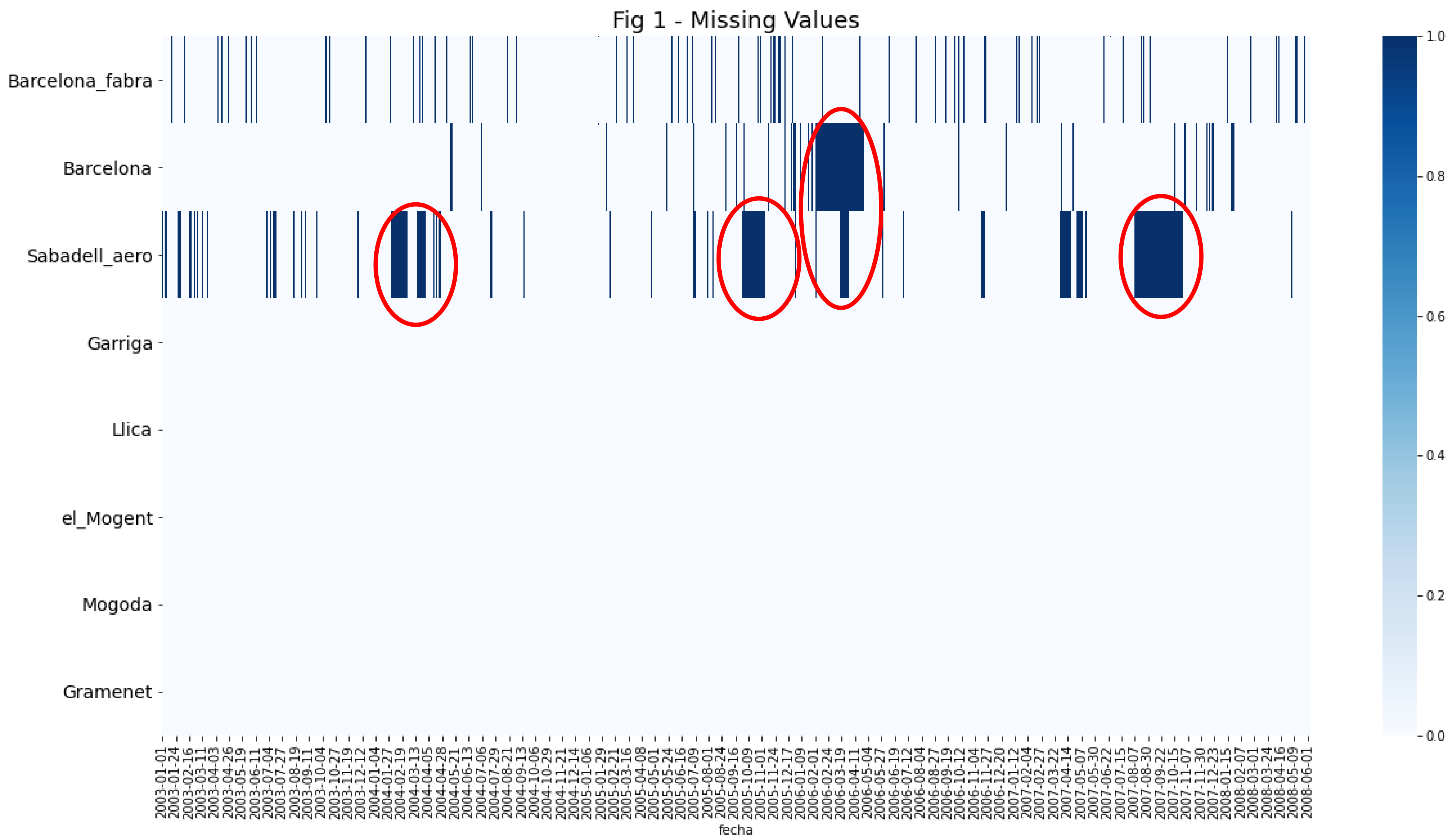
01/01/2003 **31/12/2010**

Data Exploration – Missing Values

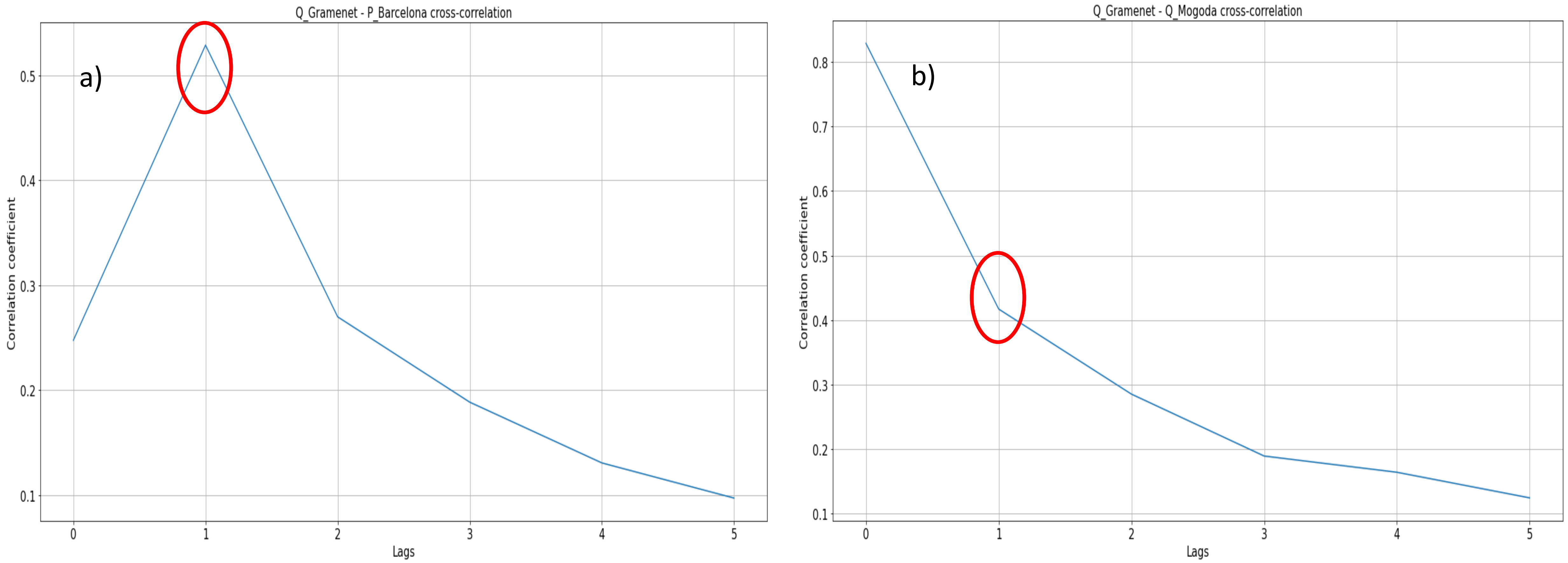


01/01/2003 ⟷ **31/12/2010**

Data Exploration – Missing Values

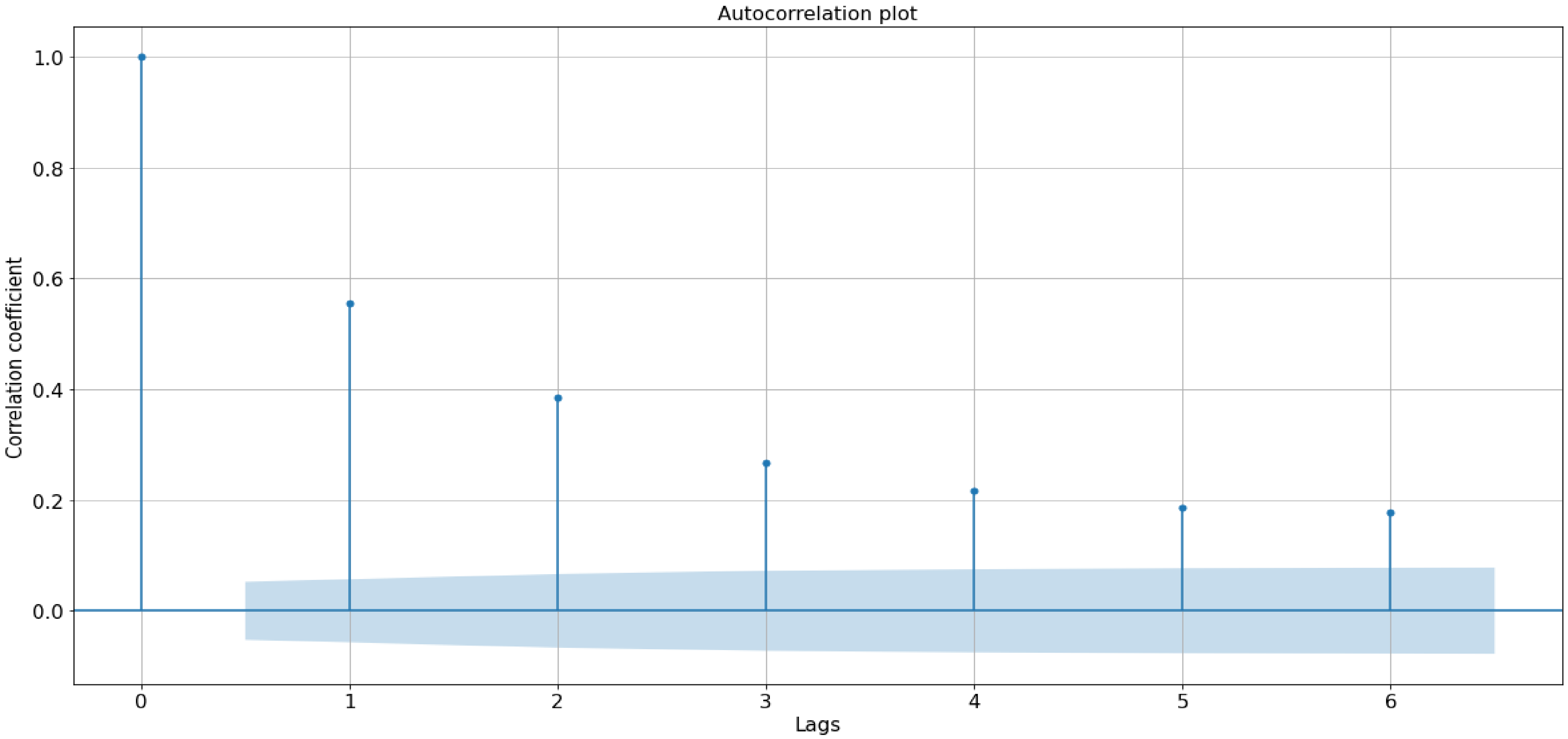


Relevant Lag times



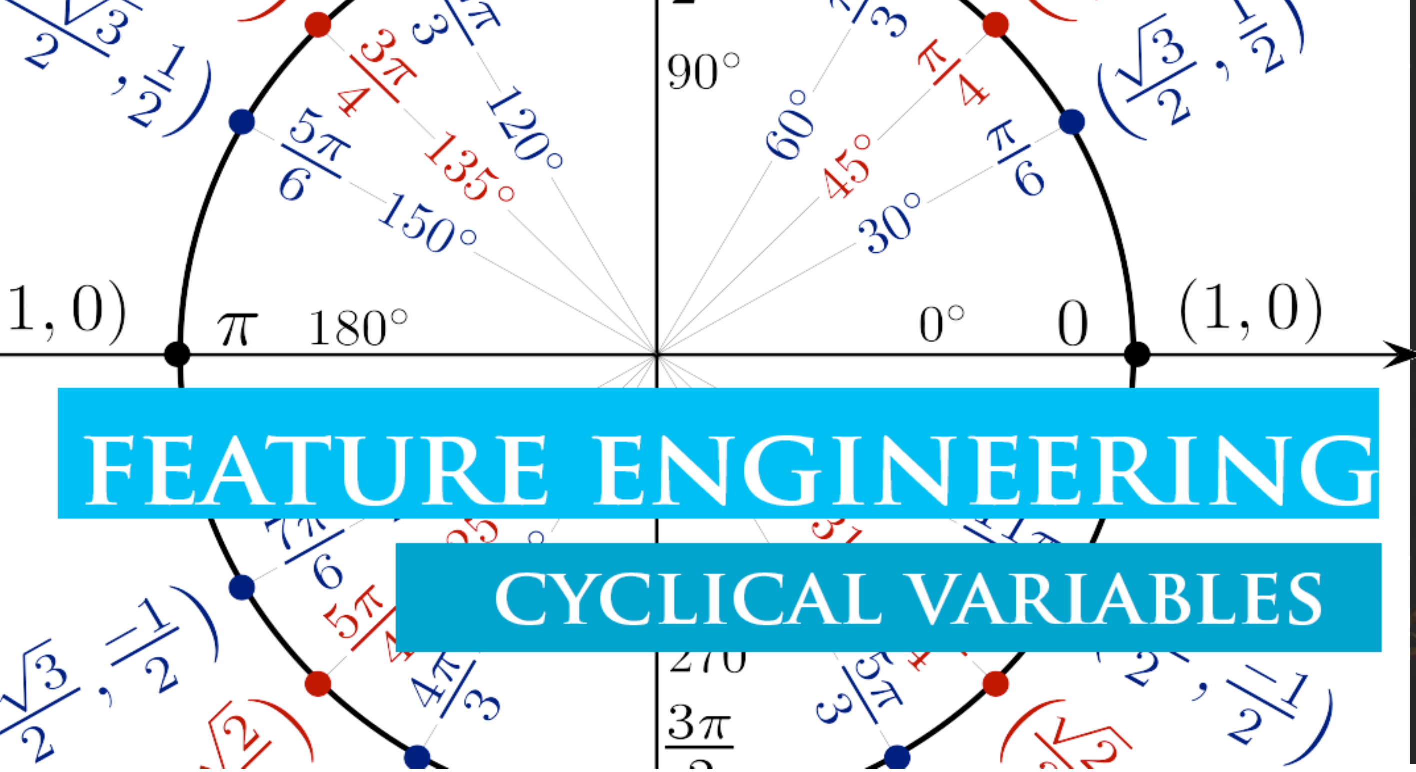
Cross-correlation between the flow at the outlet and the precipitation and flow at the input meteorological and gauging stations [a): Barcelona meteorological station; b): gauging station of Santa Perpètua d Mogoda]

Relevant Lag times



+
Trial and error method

Autocorrelogram



FEATURE ENGINEERING

CYCLICAL VARIABLES

The Cyclical Formula

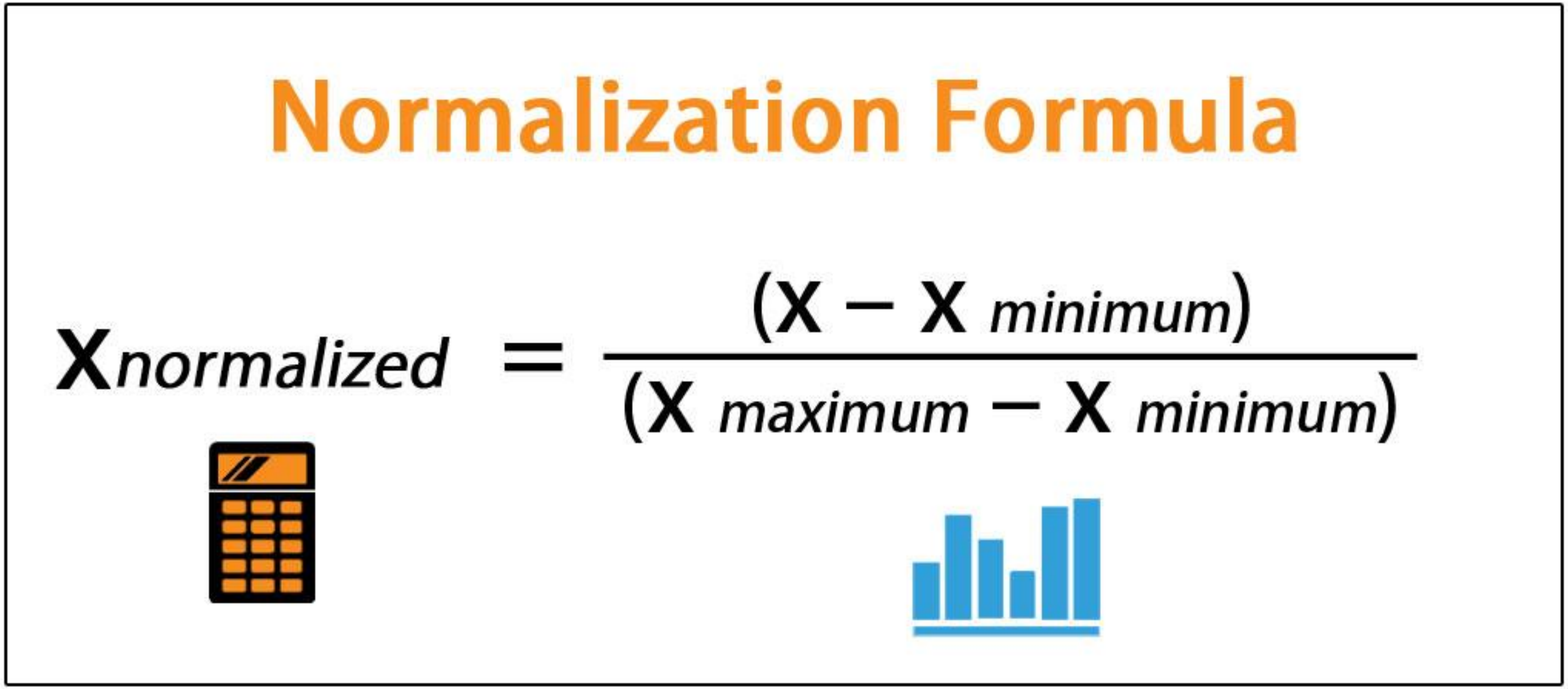
the general formula to convert a variable into a set of cyclical features:

$$x = \sin\left(\frac{a \times 2\pi}{\max(a)}\right)$$

$$y = \cos\left(\frac{a \times 2\pi}{\max(a)}\right)$$

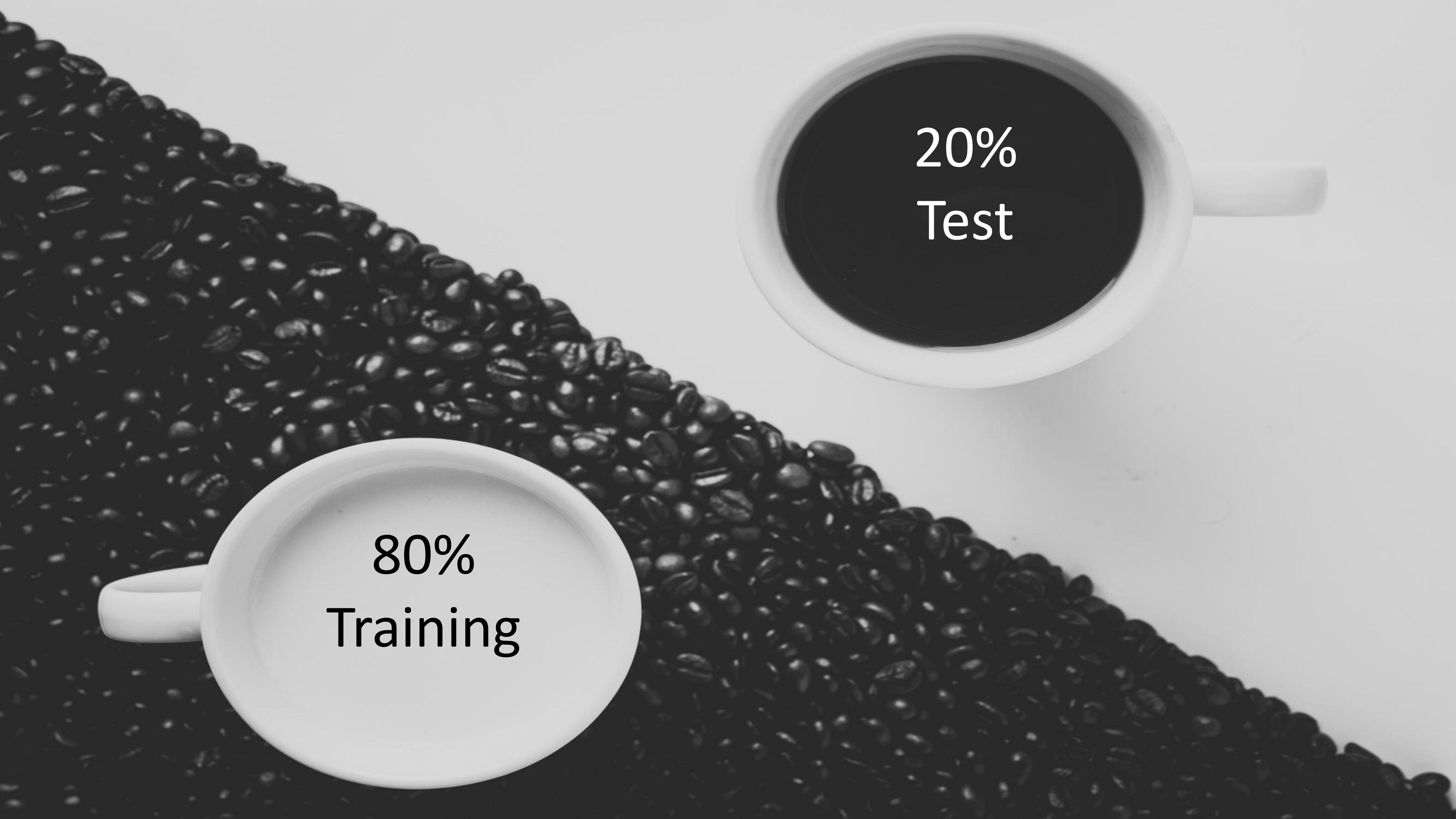
Wind direction, seasons, time, days (of a month, year, etc.) are all cyclical variables

Data Normalisation



https://www.wallstreetmojo.com/normalization-formula/

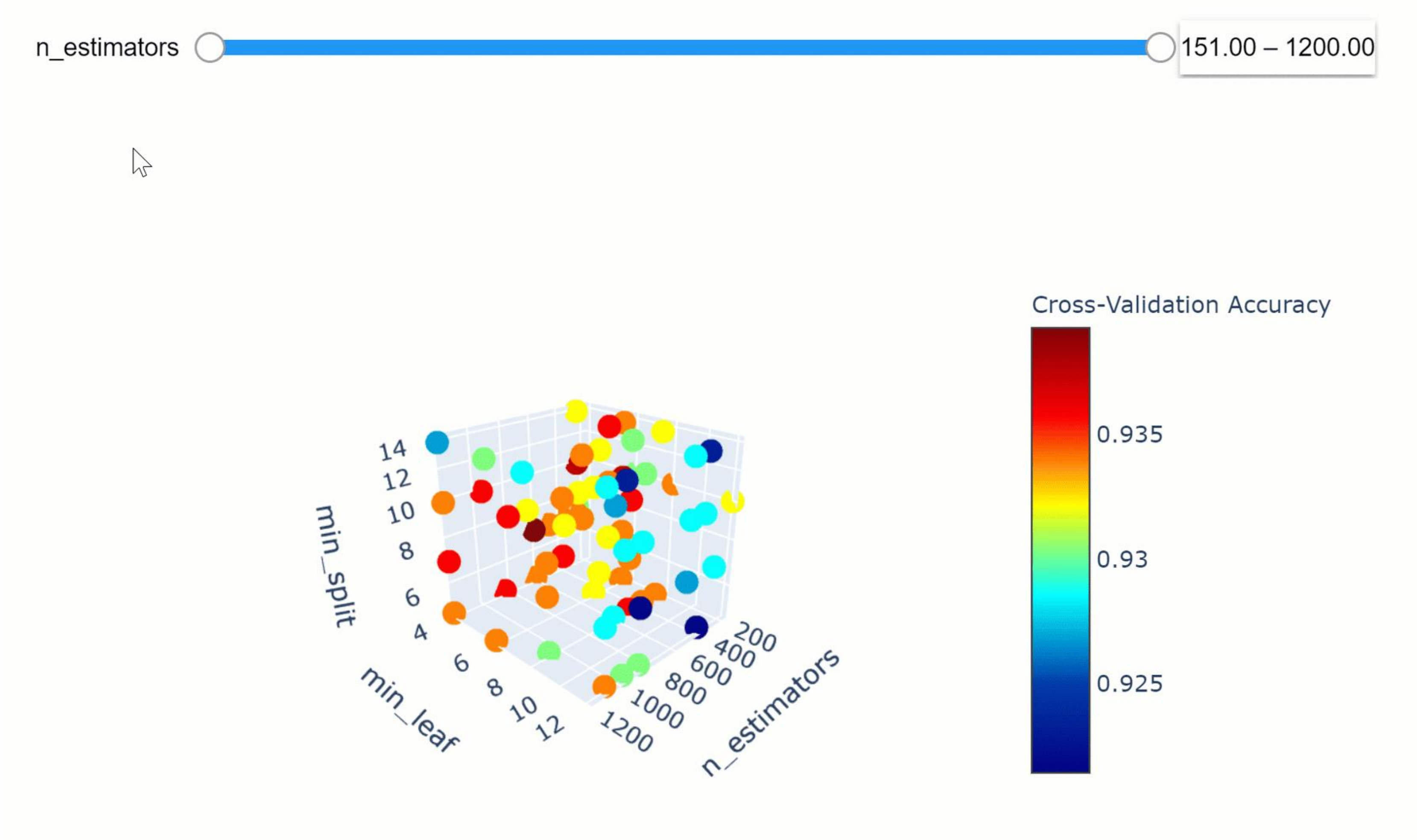
Min-Max Scaler


20%
Test
80%
Training

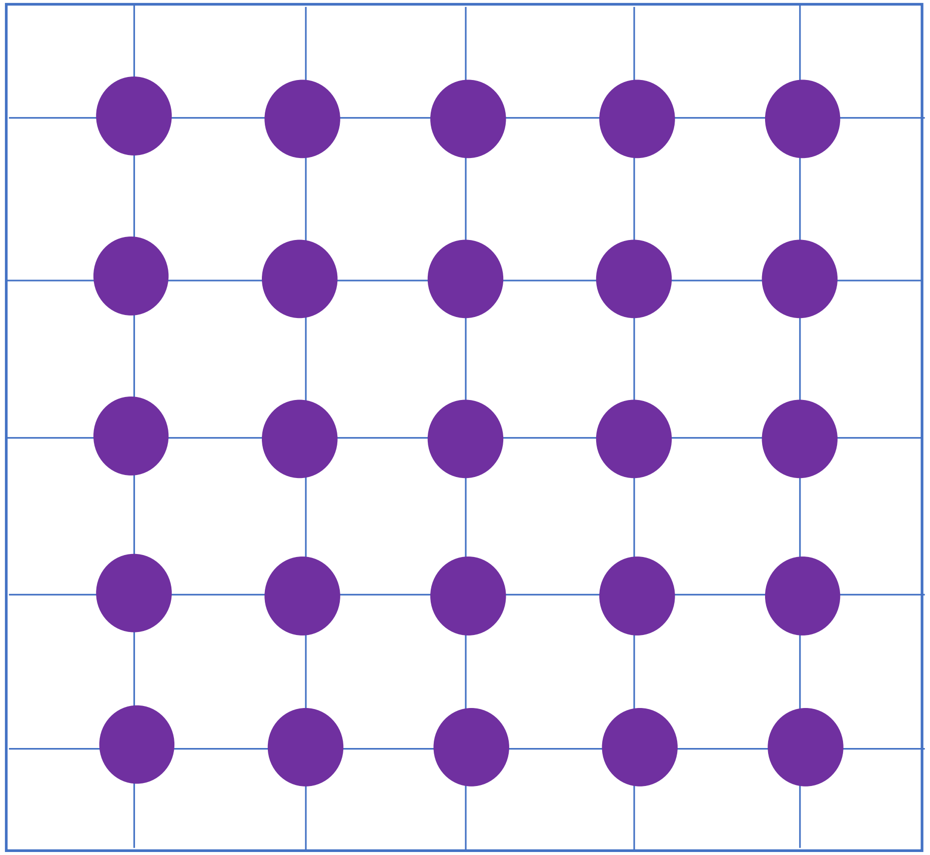
Model Hyperparameters

Model	hyperparameters
SVR	C
	ε
	γ
GBR	X_1 = Number of gradient boosted trees
	X_2 = The maximum depth of a tree
	X_3 = The number of variables to use in each node
	X_4 = learning rate parameter which controls the magnitude of each tree's contribution to the final result
	X_5 = the fraction of samples to select in each tree
RFR	n_{tree}
	m_{try}

Hyperparameter Optimisation



Grid Search

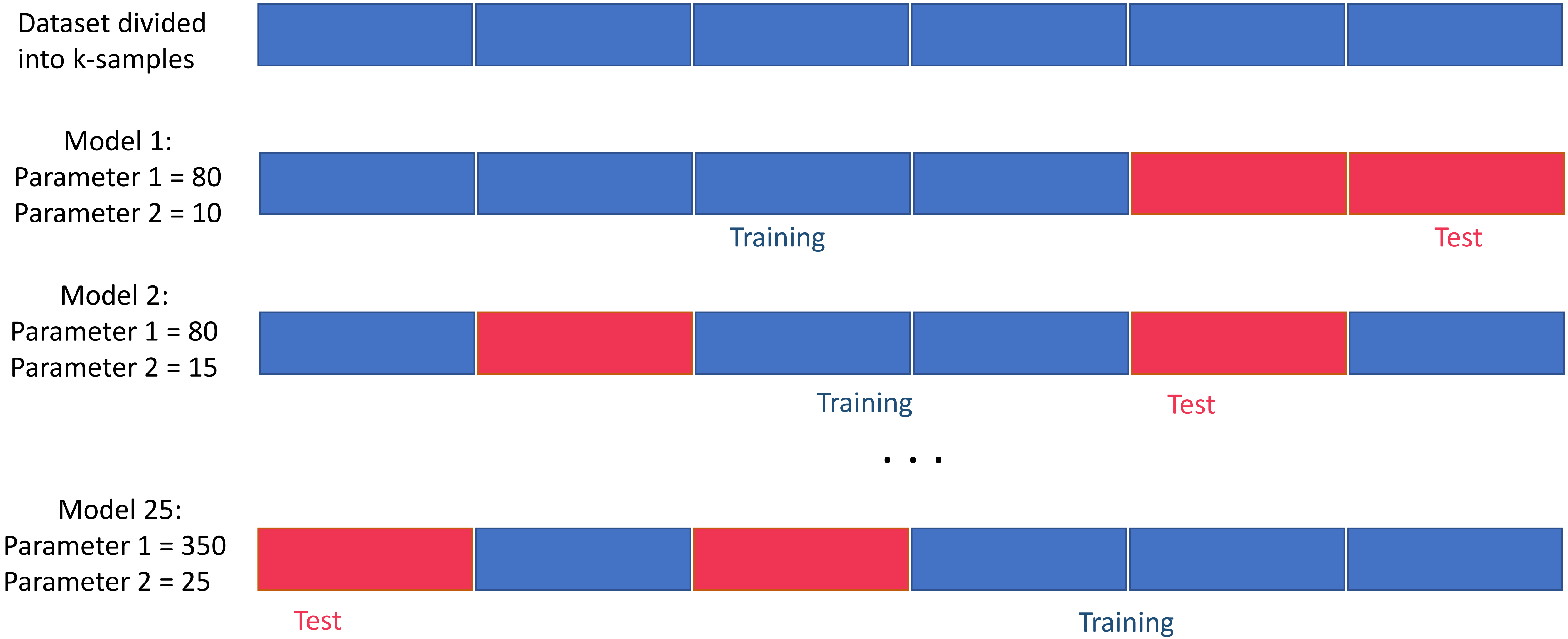


Second parameter

First parameter

- Set a range of values for each hyperparameter
- Try all possible combinations

Cross validation



Validation metrics

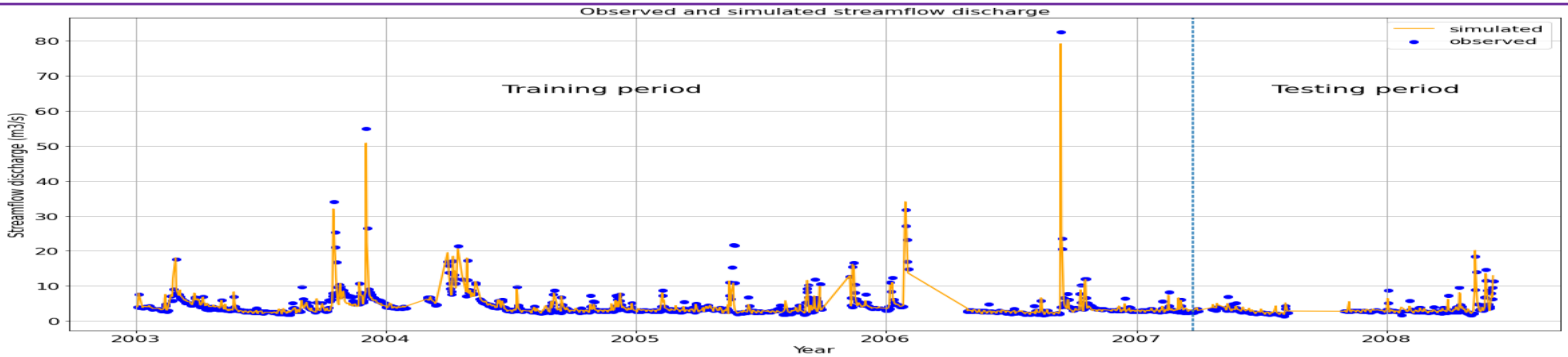
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Q_{t,obs} - Q_{t,pred}\right)^2} \qquad 0 < RMSE < \infty$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Q_{t,obs} - Q_{t,pred}\right| \times 100 \qquad 0 < MAE < \infty$$

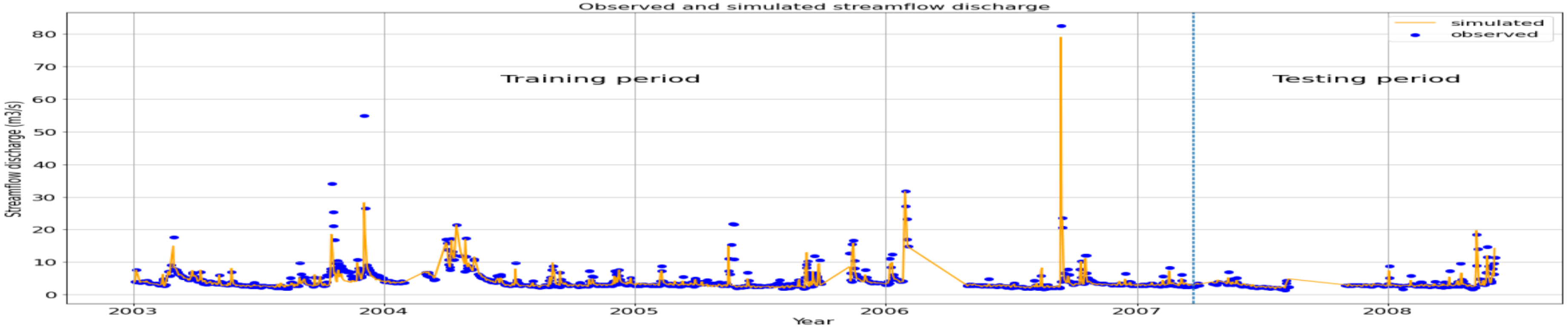
$$R^2 = \frac{N\sum_{i=1}^{N}(Q_{t,obs} - \bar{Q}_{t,obs})(Q_{t,pred} - \bar{Q}_{t,pred}) - \sum_{i=1}^{N}(Q_{t,obs})\sum_{i=1}^{N}(Q_{t,pred})}{\sqrt{\left[\left(N\sum_{i=1}^{N}({Q^2}_{t,obs}) - \left(\sum_{i=1}^{N}(Q_{t,obs})\right)^2\right)\left(N\sum_{i=1}^{N}\left({Q^2}_{t,pred}\right) - \left(\sum_{i=1}^{N}(Q_{t,pred})\right)^2\right)\right]}} \qquad 0 < R^2 < 1$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Q_{t,obs} - Q_{t,pred}\right)^2}{\sum_{i=1}^{N}\left(Q_{t,obs} - \bar{Q}_{t,obs}\right)^2}$$

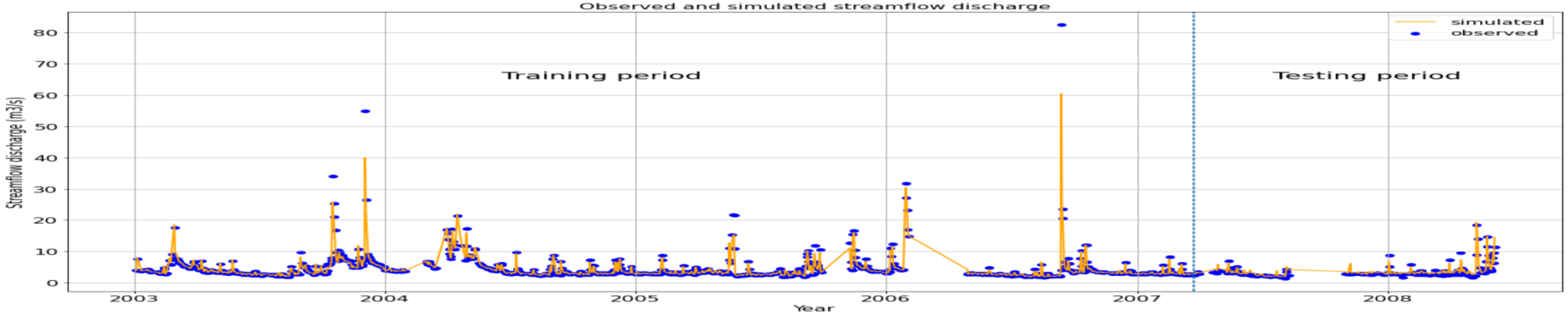
Results – 1st scenario



- GBR

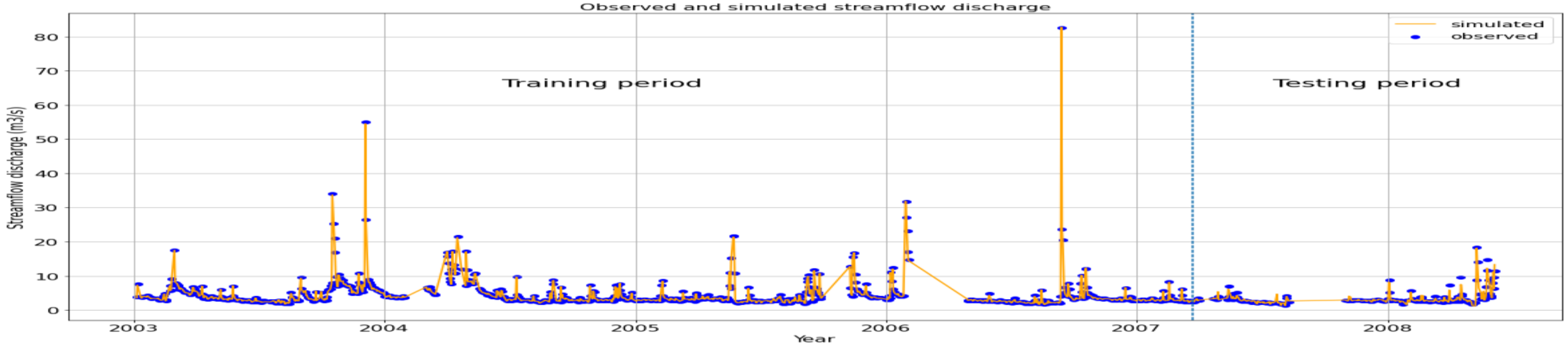


- SVR

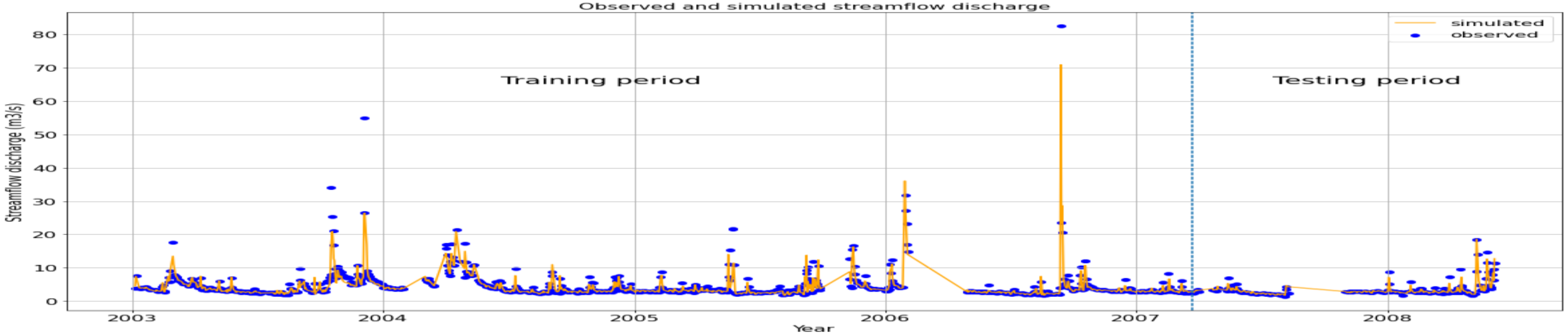


- RFR

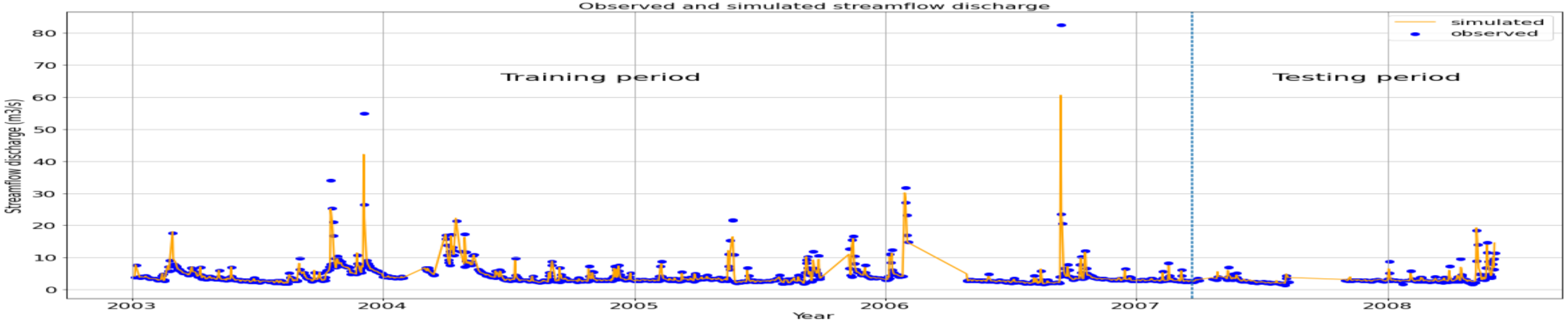
Results – 2nd scenario



- GBR



- SVR



- RFR

Validation metrics

Scenario	Model	Training					Test			
		RMSE	MAE	R^2	CE		RMSE	MAE	R^2	CE
1	MLR	1,783	0,780	0,819	0,779		0,806	0,502	0,819	0,808
	SVR	1,685	0,600	0,838	0,774		**0,604**	**0,369**	**0,898**	**0,877**
	GBR	1,062	0,558	0,936	**0,928**		0,720	0,480	0,856	0,844
	RFR	**0,983**	**0,259**	**0,945**	0,922		0,758	0,545	0,840	0,834
2	MLR	1,477	0,583	0,876	0,858		0,776	0,388	0,833	0,850
	SVR	1,563	0,492	0,861	0,805		**0,578**	**0,307**	**0,907**	**0,890**
	GBR	**0,171**	**0,131**	**0,998**	**0,998**		0,685	0,381	0,869	0,862
	RFR	0,921	0,207	0,952	0,933		0,624	0,368	0,892	0,890

Conclusions

- The SVR has been the best prediction model, whether it is for the first or second scenario
- In the training period, the RFR and GBR models have been with the best performance compared to other models.
- They have a better performance and so great than the test period
- The MLR, although it has the lowest (worst) performance, has given very acceptable results
- The use of previous flows at the target gauging station has improved the results for all models, both in the training period as well as in the test period.

Recommendations to improve the model performance

- Find other data sources, other types of data, spatial optimisation techniques
- Sensitivity analysis to input data
- Analysis of sensitivity to the length of the input data and the split ratio
- Using Evolutionary algorithms like GA, BO,… for a better optimisation of the hyperparameters
- Develop other types of ML models with better predictive power
- Xgboost $\longrightarrow$ GB

individual_project | Kaggle

kaggle.com/mohamedhamitouche/individual-project

Apps · Biochimie. Cours, te... · mathematic park -... · https://softwarerep... · Institut Henri Poinc... · Téléchargez des livr... · Matières organique... · Index of / · Irrigation 101: Intro...

kaggle

Search

Sign In · Register

- Home
- Competitions
- Datasets
- Code
- Discussions
- Courses
- More

View Active Events

individual_project

Python notebook using data from multiple data sources · 22 views · 1h ago · pandas, matplotlib, numpy, +3 more

0 · Copy and Edit · 0

Streamflow Modelling Using Machine Learning Based on Discharge and Precipitation Time series (Case Study: Santa Coloma de Gramenet Hydrometric Station)

In [6]:

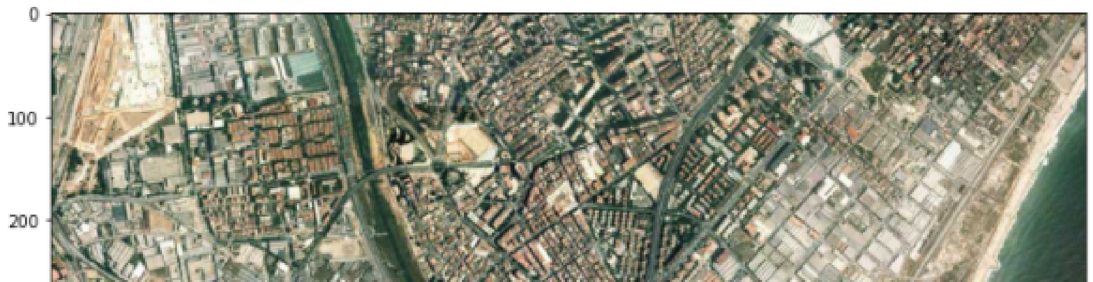
```
plt.rcParams['figure.figsize'] = (20, 10)
```

In [9]:

```
import matplotlib.pyplot as plt
image1 = plt.imread("../input/bess-river/figu002.jpg")
plt.imshow(image1)
```

Out[9]:

```
<matplotlib.image.AxesImage at 0x7fd0dbd90e10>
```



Version 3 of 3

Quick Version

Notebook

Streamflow Modelling Using Machine Learni...

Input (4)

Execution Info

Log

Comments (0)

Webgraphy

- https://scikit-learn.org/stable/index.html
- https://pypi.org/
- http://blog.davidkaleko.com/feature-engineering-cyclical-features.html
- https://towardsdatascience.com/
- https://github.com/
- https://stackoverflow.com/
- https://www.kaggle.com/
- https://medium.com/
- https://www.lovelyanalytics.com/
- https://www.analyticsvidhya.com/
- https://datascientest.com/

Thank you

For your attention!