Climate Changes and River Level Monitoring: Variations in Rainfall and Temperatures on Major Italian Rivers Discharges

Francesco Biancucci, Mario Sessa
Alma Mater Studiorum, Master in Computer Science, Bologna

Project of Human Data Science A.A. 2021/22
francesco.biancucci2@studio.unibo.it, mario.sessa@studio.unibo.it

Abstract: This study proposes an application to visually analyze river discharge changes over the last ten years in the North of Italy. Rivers like Po and Adige are crucial to sustaining the Po valley's agriculture and food factories. Recent events argue about the rivers drying in some specific areas affected by high temperatures and low precipitations due to climate changes. The proposed analysis aims to study variations of temperatures, rainfall, and river discharges in some monitored locations on the entire courses of northern Italian rivers and try to focus on some causal dependencies or correlations to archive a prediction model implementation used in a final analysis application.

1. Introduction

Last month, residents of Boretto in northern Italy discovered that the vast stretch of the Po River, running just to the north of their small town, had transformed into a beach. The pale golden sand extended for around 10 meters towards the center of the river, and inhabitants took advantage of the newly formed terrain to take a stroll and walk their dogs. In other areas, the water level dropped so low that the remains of a tank from WWII were revealed, and the ruined walls of a medieval town emerged. Italy's Po River flows some 650km from the snowy Alps in the northwest to the wild Po mouth in the east before rushing out into the Adriatic Sea. During its course, the great waterway nourishes the expansive fertile plains of northern Italy, where farmers have thrived for generations. Dubbed Italy's breadbasket, these flatlands covered with crops are responsible for some 40% of Italy's GDP. At the moment, however, the ordinarily life-giving waters of the Po River have suddenly become an unexpected threat [1]. The dramatically low water levels of the river have been causing seawater lowering to back upstream. Po is not an isolated case, rivers in the north of Italy continue to drop down the current river level, and the total mean discharge can affect water usage in the entire Italian agriculture sector. Our study aims to retrieve some statistical patterns on the discharge behaviors and define if the current reduction in the river level and the minimum values of its discharge is an isolated case or a statistical trend. Our experiments will extend this analysis procedure by considering every river flows in the north and the center of Italy and try to find some common correlations on the behavior with the temperatures and precipitations provided by Copernicus Climate Data Store on the past ten years (2021-2011).

2. Study Area

This sections provides a brief introduction to the initial data set and

2.1. Geographical Domain

The study area is between (45.3345° N, 11.3725° E), and (45.1321° N, 11.4020° E) from the bottom of the Lazio Regions to the Alps in the Trentino Alto Adige. EFAS Data Station retrieves data according to their coverage around the Italian waterfalls. The European Flood Awareness System (EFAS), jointly developed by the European Commission and the European Centre for Medium-Range Weather Forecasts (ECMWF), is a hydrological forecast and monitoring system independent of administrative and political boundaries in the European domain. EFAS aims to support preparatory measures before significant flood events. EFAS is the first operational European system monitoring and forecasting floods across Europe. It is a component of the Copernicus Emergency Management Service used to retrieve geographical measurements of the discharge, precipitations, temperatures, and other features related to the river waterfall and its possible variation over time due to the climate change effects. Due to the research interests and high variability, we will concentrate only on the ten major italian rivers geographically located in the north and center of Italy.

2.2. *Temperatures and Precipitations*

These data provide additional information to determine the discharge behavior of the climate variations according to the seasonal temperatures and rainfall accomplished in a specific river area. The initial data set provides a gridded system in NetCTF-4 format with squares of $2.5 \times 2.5 km^2$ associated with temperatures (°C) and precipitations (mm) values over the last ten years. Data were retrieved from Italian ESGF[1] nodes. NetCTF-4 is an embedded format for raster and gridded data in a multi-dimensional data source. The ambient air temperature is a daily mean close to the ground at 2 meters from the local surface level, and precipitations determine the mean quantity of rainfall in a $2.5 \times 2.5 km^2$ area. Figure 1 shows some results from the initial data visualization on the temperatures and precipitations in the section with the current area of interest using Panoply, an open-source software provided by NASA to visualize gridded systems.

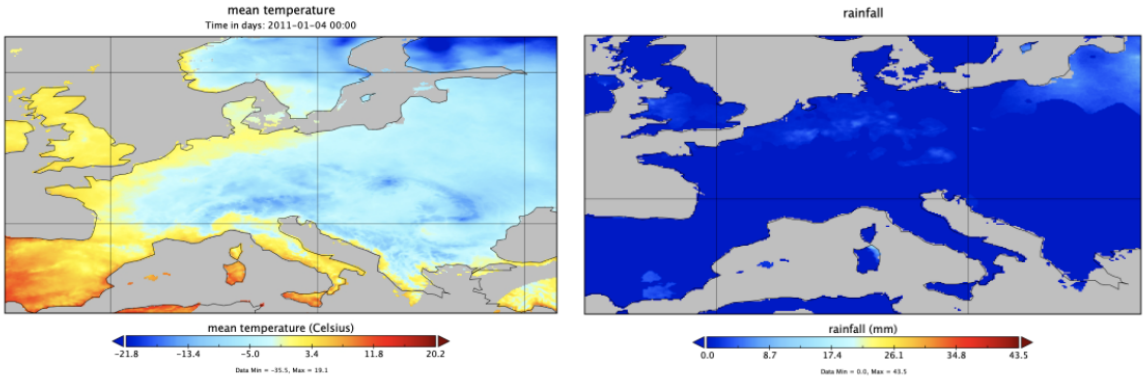

Fig. 1. Panoply visualization of the daily mean temperature and rainfall on the gridded system with an area of approximation equals to $2.5 \times 2.5 km^2$

.

2.3. *Discharges*

The initial data set provides a gridded daily hydrological time series forced with satellites meteorological observations. The data set consistently represents the essential hydrological features across the European Flood Awareness System (EFAS) domain on the major rivers in the European surface. Data format is NetCTF-4, provided by a multi-dimensional views capable to analyze data in the temporal and spatial dimensions. Due to the final goal, we filtered the entire data set to retrieve only the daily discharge values ($m^3 s$) on the north and center of Italy. This data set was produced by forcing the LISFLOOD hydrological model[2] with gridded observational data of precipitation and temperature matching data sets on a mean resolution of $2.5 \times 2.5 km^2$. So, we are able to match directly every area segmentation between our hydrological and meteorological data. Due to the final goal, we need to filter the final aggregated data only on matched rivers locations with coordinates of the ten major italian rivers, more information will be define in next sections.

3. Exploratory Spatial Data Analysis

We will focus on analyzing the data format during the initial exploratory of spatial data. This notebook is essential for data management, and it is an introduction to the data aggregation phase due to metadata representations. Furthermore, we can specify the internal structure of a NETCDF-4 file representing a well-known n-dimensional collection. Information mined from the data structures will be vital for critical decisions on the next steps. In the previous section, we analyzed the initial data composition given by the Copernicus Climate Data Store; these data sets are initially split between two logical data sources, River Discharge History (RDH) and Temperatures and Precipitations Indicators (TPI). RDH is the initial data set corresponding to the discharge values on the European rivers; meanwhile, TPI defines the precipitations and temperature measurements over the European geographical area. This section will introduce a brief introduction to data management according to the NetCTF-4 format. So, we test some data structures and functions to retrieve and better comprehend the data set composition.

[1]Earth System Grid Federation (ESGF) is a P2P enterprise system developed by NASA, NOAA, ESIP, and the European IS-INES that provides observation models on Earth intending to facilitate advancements in Earth System Science [2].

[2]LISFLOOD model is a grid-based hydrological rainfall-runoff-routing model capable of simulating the hydrological processes that occur in a catchment. It is used in large and transnational river basins on national, continental, or global scale for a variety of applications, including flood forecasting, water resources assessments. [3]

3.1. *Exploratory on RDH Dataset*

RDH is the first of the two datasets we will use during the project. The original name is "River discharge and related historical data from the European Flood Awareness System." It represents a geographical distribution of river discharges on the European surface in m^3s. The amount of data on this dataset overextends the accurate and not nullable information. The original dataset is unique; we have been split it due to the huge dimensions of files and the data limits for downloading on the Climate Data Store (CDS). During the exploratory, we will discuss only a subset of information for a time interval of 2 years inside the 2021-2022 events due to the expensive time spent on big data management and because data don't change metadata and layout over time; it is a good practice specified that the amount of data i two years available for the bound 2021 and 2022 isn't complete because the analysis is on the present and past data. We will match geographical locations with precipitations and temperatures after possible data shifting to match the measurements during the data aggregation phase. The following table displays the results of the initial analysis in a single year; we specify that, during the analysis, experimentation is done over a couple of years according to the original data set splitting.

Fields	**Descriptions**
Data Type	Gridded with geographical spatial references given by Lambert Azimuthal Equal-Area Projection on the ETRS-LAEA bound.
File Format	NetCDF-4.
Dimensions Description	Spatial References given by x and y pairs to display the Azimuth Equal-Area Projection and temporal dimension given by $time$ values.
Dimensions Shape	$(x, y, time) = (950, 1000, 365)$.
Horizontal Resolution	$5 \times 5km^2$ per Projection.
Spatial Variables	$latitude$ and $longitude$ on geographical references.
Discharge	Volume rate of water flow, including sediments, chemical and biological material, in the river channel averaged over a time step through a cross-section. The value is an average over each 24-hour time step measured in m^3s.

Table 1. Results on the Initial Exploratory on the River Discharge History (RDH) Dataset.

We obtained some information about each of the three dimensions distributions and what they are formalized. The $time$ is related to the initial moment of the forecast of the discharges per single year. The dataset also has 2022 as a year; due to the no-prediction measurements, the years 2021-2022 have 458 days instead of 730; during the data aggregation phase, we will exclude 2022 from the merging mechanisms to maintain consistency in data over time dimension. Meanwhile, x and $y$are the coordinates in meters of the geographical projection, and the dimensions are in spatial perspective (2D). It is essential to specify that (x, y) is a projection of the point in a geographical environment, and they do not represent any coordinates on the final data. Projection points are the total size of points in a matrix (N, M) bound given by the ETRS-LAEA bound. Meanwhile, $latitude$ and $longitude$ are related to the exact geographical coordinates on the globe inside the ETRS-LAEA section.

3.2. *Exploratory on TPI Dataset*

TPI is the second of the two used datasets. The original name is "Temperature and precipitation climate impact indicators from 1970 to 2100 derived from European climate projections". Due to the vast dimensions of files and the data limits for downloading on the Climate Data Store (CDS), original file analysis is split into different years. We analyzed only one of them to consider some structural characteristics and complete the exploratory data structure; we will merge them during the aggregation phase. The original data set has two types of data: monitored data related to the Global Climate Observing System[3] (GCOS) measurements, and predicted data from

[3]GCOS expert panels maintain definitions of Essential Climate Variables (ECVs) which are required to observe Earth's changing climate systematically. The observations supported by GCOS contribute to solving challenges in climate research and underpin climate services and adaptation measures.

eight model simulations included in the Coordinated Regional Climate Downscaling Experiment[4] (CORDEX). In our analysis, we will consider only real measurements referred to the last ten years as time-bound that we will form the TPI data set. During the initial exploratory, we need to retrieve additional information about the geometric distribution of the measurement projections and find a way to match TPI and RDH. Exploratory pattern focuses on the exact mechanisms of the TPI. The following table will show some results of the exploratory data phase.

Fields	**Descriptions**
Data Type	Gridded with geographical spatial references given by Lambert Azimuthal Equal-Area Projection given by Global Climate Observing System (GCOS).
File Format	NetCDF-4.
Dimensions Description	Spatial References given by x and y pairs to display the Azimuth Equal-Area Projection and temporal dimension given by $time$ values.
Dimensions Shape	$(time, x, y) = (365, 950, 1000)$.
Horizontal Resolution	$5 \times 5km^2$ per Projection.
Spatial Variables	$latitude$ and $longitude$ on geographical references.
Precipitations	Time mean flux of rain, snow and hail measured as the height of the equivalent liquid water in a square meter per time interval measured in $mm\ day^{-}1$
. Temperatures	Temperature of air at a height of 2 metres above the Earth's surface in Celsius.

Table 2. Results on the Initial Exploratory on the Temperature and Precipitation Indicators (TPI) Dataset.

We can see that the dataset has the exact dimensions of the RDH. It is good information because we can focus directly on our statistical hypothesis after the data aggregation if data references are the same for each point. This process should be much easier than procedures with clustering and geographical approximations like k-means techniques. Unfortunately, geographical and projections have different associations between RDH and TPI. So, we will describe our solution for geographical matching between RDH and TPI coordinates.

3.2.1. Comparison between RDH and TPI on coordinates

During the exploratory on RDH and TPI and from the dataset description on the CDS, we saw how data are distributed and which dimensions they are. Furthermore, we saw that the size for each dimension is the same over time, but projections coordinates are not.So, we had. to find a function $f : f(x,y) = (x',y') : x,y \in RDH \wedge x',y' \in TPI$ or vice-versa. We will test only on some samples, and if we have a positive result, we will extend verification to the entire domain. Our hypothesis is related to possible rotations on coordinates on the x and y dimensions because we had positive feedback on the spatial dimensions size comparisons. After some experiments, we found the rotation function to match coordinates over RDH and TPI projections:

$$f : C_{rd} \rightarrow C_{rd} : f(x,y) = (-x,y), \forall (x,y) \in C_{rd}$$

C_{rd} is the RDH data set's coordinates. Finally, according to the official datasets documentation, each measurement on TPI and RDH are values related to the same dimension time (one time per day at 12:00 a.m. each day in a year). In conclusion, we can analyze the main features for each measurement in the same way and compares possible results as additional information which should be helpful for critical decision during data management.

[4]Experiments to improves the Global Climate Model (GCM) resolution over projections in regional area implementing a Regional Climate Model (RCM) and Empirical Statistical Downscaling (ESD), applied over a limited area and, driven by GCMs, it can provide information on much smaller scales supporting more detailed impact and adaptation assessment and planning, which is vital in many vulnerable regions of the world.

4. Data Aggregation

We found the geographical correlation function that can match the TPI coordinates with the RDH dataset, which we passed to carry out the data aggregation phase. Data has three dimensions, two spatial (given by the projection on the Lambert Azimuthal Equal-Area Projection) and one temporal (time interval in the last ten years. This phase will aggregate data over time, so we transform the original dataset into a set of time series mathematically defined by the following set:

$$T = \{x | x = \{x_1, x_2, ..., x_n\} \wedge x_1, x_2 ... x_n \in T_{(x,y)} \wedge (x,y) \in C\}$$

Where T is the set of time series, x is a singular time series composed by $x_1, x_2, ..., x_n$ values correlated to different timestamp for a time series $T_{(x,y)}$ with (x,y) a coordinate of the set of locations C. Due to the approximation on $2.5km^2$ on the initial dataset, we don't need to aggregate on spatial dimensions. So, we aggregate data from the temperatures and precipitations dataset according to discharge locations in a geographical pattern matching each location in the study area described in section 2.1. We can filter each point over time by its coordinates to define a specific time series related to a river monitoring location. In addition, we can retrieve data like precipitation, temperature, and river discharge directly during the same filtering process. Figure 2 shows distributions related to

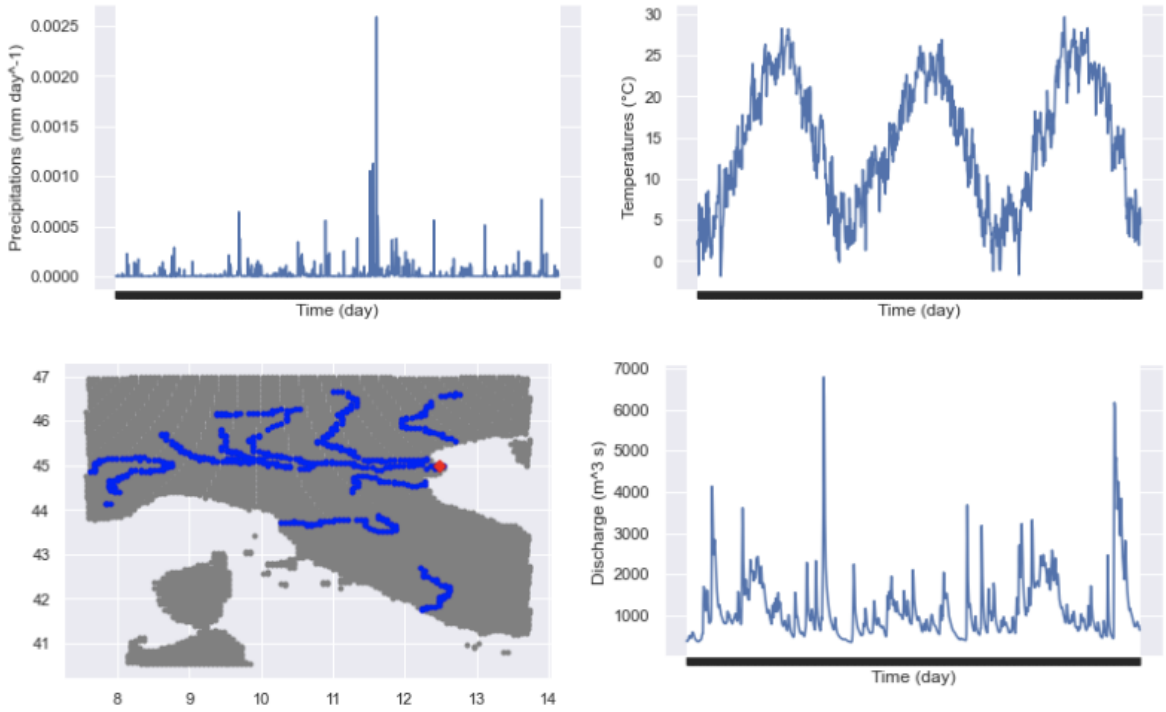

Fig. 2. Precipitations, Temperatures and Discharge on the Po's mouth in coordinates showed as red marker on the third plot.

the main monitored attributes for a specific location. During the data analysis, we will use the sampling approach, a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the analysis performed. Our experiments used a sampling technique based on the high semantic variance on river monitoring locations strictly related to a single river. In other words, we randomly chose a location at the source and at the mouth of the river to determine a possible geographical behavior extendable to each river's monitored station.

5. Data Analysis

During this phase, we will try to find some correlation and auto-correlation on the final dataframe as the result of data aggregation. According to the time series techniques, we will find a sampling of data related to coordinates, and then we will use the result dataset as the input for this phase. In the data analysis procedures, we will analyze some features of the dataset variables, their auto-correlations, correlations, and aspects related to the distribution of their values over time.

5.1. Pre-Processing

5.1.1. Coordinates Filtering

The Lambert Azimuthal Equal Area Projection considers every coordinates inside the $[x_{min}, y_{min}]$ to $[x_{max}, y_{max}]$ aggregated projections. This side-effect is due to the not null element of temperatures and precipitations in the aggregation phase that considers every geographical location in the projection bound. So, we need to filter only locations that match with a no-zero value for the discharge value. Filtering process is possible due to the continues discharge variations with minimum values different from zero.

5.1.2. Major Rivers Filtering

The output of the previous process is formed by a set of locations corresponding to every river or water flow in the study area. Due to dimensions, we apply a filter to our dataframe to determinate values that is coordinates of one of the major italian rivers: Tevere, Ticino, Oglio, Tanaro, Adige, Adda, Piave, Po, Arno, Reno. Due to the lack of river name association with a specific coordinate on the dataset, we implemented a supported script that generate a Json document starting from external ArcGIS Spatial Dataset formed by rivers in geometric lines. Firstly, we extrapolate points approximation on these lines maintaining a minimum distance to obtain a systematic sampling; than, we match coordinate of the river's reference dataset with our data to obtain a logic association between geographical locations and rivers references. Finally, we filter locations only if the name of the row matches with one of the ten major river's name.

5.2. Statistical Analysis

In this section, we will shows distributions over dataset variables on previous filtered locations. We will discover the possible seasonality, auto-correlations, correlations and hidden patterns inside the data variations over time for each of ten selected major rivers on their sources and mouths. Sources and mouths contains the time series related to the source and mouths points chosen by each north Italian rivers. Given an statistical hypothesis, this collection is used to verify some properties for different Italian rivers. So, we will analyze in the sections below variations on discharges, temperatures and precipitations and, finally, we concludes with some observations and related tests on the correlations and auto-correlations between their values. Due to avoiding verbose description, we will show only analysis on the Po's mouths but, during the project, we extended it to each river.

5.3. Seasonal Decomposition

We performed a seasonal decomposition using additive approach to find some analysis on the trends, seasonality and residual values over years and try to find some years or monthly variation on the discharge value. These information can obtains some observations:

1. The trend component is supposed to capture the slowly-moving overall level of the series.
2. The seasonal component captures patterns that repeat in a specific period.
3. The residual is what is left. It may or may not be auto-correlated.

In next subsections, we will show some related results on discharge, temperatures and river variations using additive seasonal decomposition.

5.3.1. Discharge Variations

Considering the ten major Italian rivers, discharge variation is between $1.7401733e-14 m^3s$ on the Arno source (43.781845° N, 11.010019° E) to $9156.39 m^3s$ on the Po's mouth (45.0442° N, 11.032715° E) and a general mean discharge of $250.55 m^3s$ in the period 2011-2021. In the Figure 3, we can see an example of how the trends is positive in the last months of most considered years (around November and December) on the Po's mouth and can influence mean trends. Initially, we suppose that this behavior was due to the decrement of the maximum temperatures and correlated water evaporation reduction.

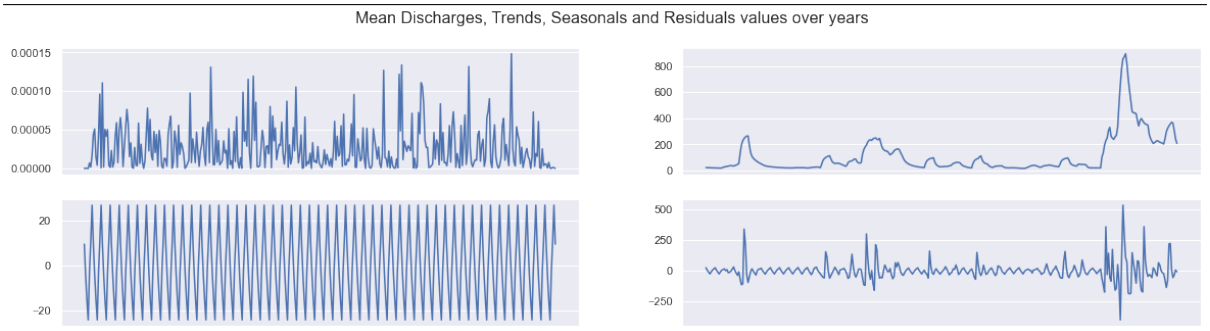

Fig. 3. Mean of seasonal decomposition discharge values on the period 2011-2021 for the Po's mouth. In order: Time series (up-left), trend (up-right), seasonality (down-left) and residual distribution (down-right). The x axes is equal to the 365 days with no consideration for additive leap days.

5.3.2. Temperature Variations

Temperatures rivers variation is between $-20.38°$ C on the Piave source in the Alps (46.63714° N, 12.694899° E) and 33.55° C on Reno mouth near the Comacchio Valley in Emilia-Romagna and a mean temperature of 13.46° C in the period 2011-2021. The seasonal decomposition is done on the same location of the discharge described in the previous section. Figure 4 shows how the mean annual temperature changes over a year considering measurements on the past 10 years. We know that temperatures focus the seasonality over years with local maxima in the summer

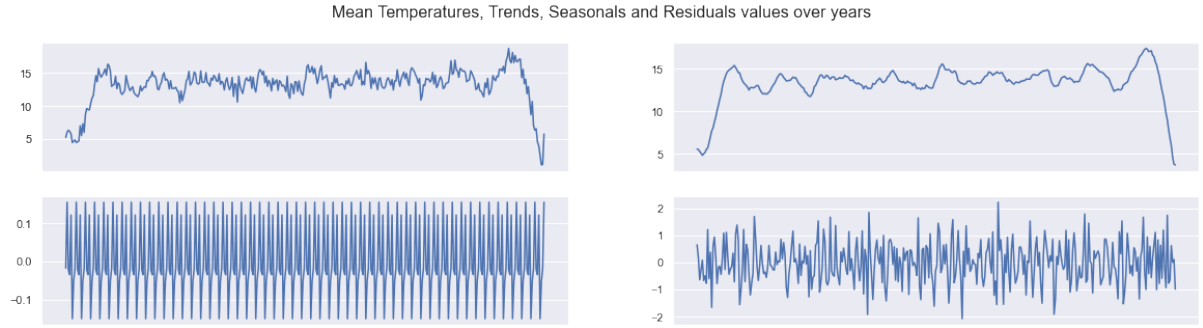

Fig. 4. Mean of seasonal decomposition temperature values on the period 2011-2021 for the Po's mouth. In order: Time series (up-left), trend (up-right), seasonality (down-left) and residual distribution (down-right). The x axes is equal to the 365 days with no consideration for additive leap days.

and local minima in the winter on the North of Italy. We can see how the mean trend is lower on the winter months but is similar for other seasons. The location analyzed before is related to the Po mouths, where the sea level and constant Adriatic winds can mitigate the temperature variation during spring and autumnal period.

5.3.3. Precipitations Variations

The precipitations bound is between a minimum of and a maximum of with a mean of . Precipitations have irregular distribution given by the seasons, area of interest, and temperature influence but also unavailable data like the air's humidity. In any case, we can consider it a variable that could cause a variation in the discharge of a given river. This test will be considered in the following sections. Now, we consider the distribution in different years and if it has a regular trend that may have interesting information with exciting data. Generally, precipitations have high variability influenced by temperature variations that can increment the probability of precipitations; an example of the rainy seasons is, in the northern hemisphere, between May and September, with some variations strictly related to the geographical locations weather compliance. As average temperatures at the Earth's surface rise, more evaporation occurs, increasing overall precipitation with the maximum pitch in months with high temperatures. [4] Figure 5 shows the mean annual precipitation distribution on one of the selected, monitored locations. The high variance affects the trend plot and makes it irregular according to a visual comparison with the uniform distributions of discharge and temperatures.

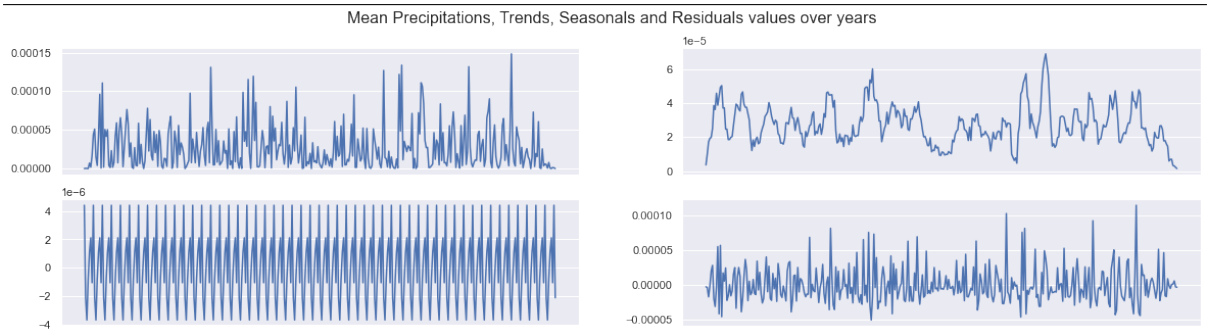

Fig. 5. Mean of seasonal decomposition precipitations values on the period 2011-2021 for the Po's mouth. In order: Time series (up-left), trend (up-right), seasonality (down-left) and residual distribution (down-right). The x axes is equal to the 365 days with no consideration for additive leap days.

5.4. Data Correlations

Data Correlation Analysis will focus on representative samples of mouths and sources for each of the ten major rivers. According to the previous data filtering, the total number of locations in our dataset is 462 different time series.

5.4.1. Lag Plots

A lag plot checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot. The Non-random structure in the lag plot indicates that the underlying data are not random. We use them for each source and mouth in our sample set. Due to the oversized dimensions, we cannot show the complete results in the documentation. So, we will invite you to check it on our Google Colab notebook or directly on the GitHub Repository.

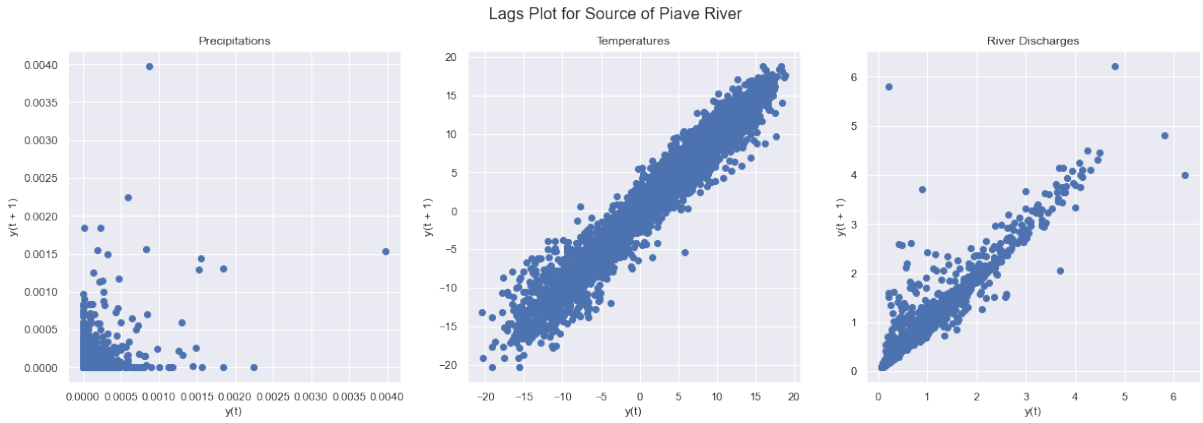

Fig. 6. It represents a sample of rivers lag plot, the image above defines the Piave sources values lag distributions on temperatures, discharges and precipitations. Generally, this kind of plot is similar to every italian rivers.

5.4.2. Augmented Dickey-Fuller (ADF) Tests

Suppose a seasonal decomposition is an analysis tool that provides us a framework for how to think about the different components of our time series. Stationary Tests observe only the trend behavior and specify if there is a unit root that causes a stochastic trend and make the time series non-stationary. Testing data for stationary is essential in research where the underlying variables are based on time. The Augmented Dickey-Fuller (ADF) test, which we address as a test for stationary, is a test looking for unit roots by attempting to fit an auto-regressive model to the data. In other words, ADF expands the Dickey-Fuller Test, including high order regressive process in the model.

$$y_t = c + \beta t + \alpha y_{t-1} + \Theta_1 \Lambda Y_{t-1} + \Theta_2 \Lambda Y_{t-2} \ldots + \Theta_p \Lambda Y_{t-p} + e_t \; ; \; p \in [1..t]$$

The Dickey-Fuller test is a unit root test that has a null hypothesis of the existence of the unit root that can make the time series not stationary, which is equal to that $\alpha = 1$ in the previous equation. In our cases, every river rejects the null hypothesis, so we can affirm that discharges, temperatures, and precipitations are stationary. According to the official definition of the Dickey-Fuller Test, the augmented version is used for complex and high variable time-dependent series. However, it has a relatively high false-positive error [5]. It is a good practice to combine it with the KPSS to secure no false positives. The table below shows some results from the Dickey-Fuller tests and statistics output on the selected samples for the significant Italian rivers. The time series follows the daily distribution of the discharge values.

5.4.3. Augmented Dickey-Fuller on Discharge Variations

River's Source	ADF Statistics	H0	p-value	Critic 1%	Critic 5%	Critic 10%
Adda	-8.815548	rejected	1.938805e-14	-3.432356	-2.862426	-2.567242
Adige	-6.621481	rejected	6.021646e-09	-3.432357	-2.862427	-2.567242
Arno	-5.304214	rejected	0.000005	-3.432356	-2.862426	-2.567242
Oglio	-5.493860	rejected	2.145171e-06	-3.432358	-2.862427	-2.567242
Piave	-7.385005	rejected	8.281093e-11	-3.432344	-2.862421	-2.567239
Po	-8.302083	rejected	3.993242e-13	-3.432347	-2.862422	-2.567240
Reno	-6.412963	rejected	1.871070e-08	-3.432358	-2.862427	-2.567242
Tanaro	-5.962068	rejected	2.027976e-07	-3.432358	-2.862425	-2.567242
Tevere	-7.051691	rejected	5.505515e-10	-3.432351	-2.862424	-2.567241
Ticino	-7.377117	rejected	8.663842e-11	-3.432346	-2.862422	-2.567239

Table 3. Trend Statistics, critical and p values for **Discharge** time series on the major rivers in the **source** locations considering daily distribution.

River's Mouth	ADF Statistics	H0	p-value	Critic 1%	Critic 5%	Critic 10%
Adda	-7.982572	rejected	2.599649e-12	-3.432351	-2.862424	-2.567241
Adige	-6.621481	rejected	6.021646e-09	-3.432357	-2.862427	-2.567242
Arno	-5.330118	rejected	4.741506e-06	-3.432356	-2.862426	-2.567242
Oglio	-5.493860	rejected	2.145171e-06	-3.432358	-2.862427	-2.567242
Piave	-9.624170	rejected	1.677881e-16	-3.432344	-2.862421	-2.567239
Po	-6.606066	rejected	6.552152e-09	-3.432354	-2.862425	-2.567241
Reno	-11.726975	rejected	1.370195e-21	-3.432345	-2.862421	-2.567239
Tanaro	-8.134079	rejected	1.070844e-12	-3.432352	-2.862425	-2.567241
Tevere	-7.725319	rejected	1.163210e-11	-3.432346	-2.862422	-2.567240
Ticino	-6.466863	rejected	1.398260e-08	-3.432351	-2.862424	-2.567241

Table 4. Trend Statistics, critical and p values for **Discharge** time series on the major rivers in the **mouth** locations considering daily distribution.

5.4.4. Augmented Dickey-Fuller on Temperatures Variations

River's Source	ADF Statistics	H0	p-value	Critic 1%	Critic 5%	Critic 10%
Adda	-3.559840	rejected	6.576883e-03	-3.432351	-2.862424	-2.567241
Adige	-3.641482	rejected	5.015748e-03	-3.432351	-2.862424	-2.567241
Arno	-3.297295	rejected	1.499981e-02	-3.432351	-2.862424	-2.567241
Oglio	-3.739062	rejected	3.596078e-03	-3.432350	-2.862424	-2.567240
Piave	-3.345214	rejected	1.297447e-02	-3.432351	-2.862424	-2.567241
Po	-3.464987	rejected	8.933177e-03	-3.432349	-2.862423	-2.567240
Reno	-3.379716	rejected	1.167011e-02	-3.432347	-2.862422	-2.567240
Tanaro	-3.210524	rejected	1.938354e-02	-3.432353	-2.862425	-2.567241
Tevere	-3.658147	rejected	4.741891e-03	-3.432348	-2.862423	-2.567240
Ticino	-3.279573	rejected	1.581663e-02	-3.432349	-2.862423	-2.567240

Table 5. Trend Statistics, critical and p values for **Temperatures** time series on the major rivers in the **source** locations considering daily distribution.

River's Mouth	ADF Statistics	H0	p-value	Critic 1%	Critic 5%	Critic 10%
Adda	-3.033705	3.186447e-02	rejected	-3.432349	-2.862423	-2.567240
Adige	-3.227309	1.845733e-02	rejected	-3.432348	-2.862423	-2.567240
Arno	-3.297295	1.499981e-02	rejected	-3.432351	-2.862424	-2.567241
Oglio	-3.103997	2.625880e-02	rejected	-3.432349	-2.862423	-2.567240
Piave	-3.171265	2.170939e-02	rejected	-3.432349	-2.862423	-2.567240
Po	-3.147182	2.325291e-02	rejected	-3.432348	-2.862423	-2.567240
Reno	-3.250171	1.725816e-02	rejected	-3.432347	-2.862422	-2.567240
Tanaro	-3.210524	1.938354e-02	rejected	-3.432353	-2.862425	-2.567241
Tevere	-3.507856	7.787621e-03	rejected	-3.432349	-2.862423	-2.567240
Ticino	-3.279573	1.581663e-02	rejected	-3.432349	-2.862423	-2.567240

Table 6. Trend Statistics, critical and p values for **Temperatures** time series on the major rivers in the **mouth** locations considering daily distribution.

5.4.5. Augmented Dickey-Fuller on Precipitations Variations

River's Source	ADF Statistics	H0	p-value	Critic 1%	Critic 5%	Critic 10%
Adda	-9.426132	rejected	5.340659e-16	-3.432354	-2.862426	-2.567241
Adige	-9.779641	rejected	6.787353e-17	-3.432354	-2.862426	-2.567241
Arno	-18.889900	rejected	0.000000	-3.432345	-2.862421	-2.567239
Oglio	-18.457039	rejected	2.149530e-30	-3.432345	-2.862421	-2.567239
Piave	-30.587298	rejected	0.000000e+00	-3.432342	-2.862420	-2.567239
Po	-42.748592	rejected	0.000000e+00	-3.432342	-2.862420	-2.567238
Reno	-20.725085	rejected	0.000000e+00	-3.432345	-2.862421	-2.567239
Tanaro	-17.111024	rejected	7.365509e-30	-3.432346	-2.862422	-2.567240
Tevere	-13.477790	rejected	3.286121e-25	-3.432348	-2.862423	-2.567240
Ticino	-18.957490	rejected	0.000000e+00	-3.432345	-2.862422	-2.567239

Table 7. Trend Statistics, critical and p values for **Precipitations** time series on the major rivers in the **source** locations considering daily distribution.

River's Mouth	ADF Statistics	H0	p-value	Critic 1%	Critic 5%	Critic 10%
Adda	-19.930490	rejected	0.000000e+00	-3.432344	-2.862421	-2.567239
Adige	-26.700276	rejected	0.000000e+00	-3.432343	-2.862421	-2.567239
Arno	-9.367602	rejected	7.526570e-16	-3.432354	-2.862426	-2.567241
Oglio	-18.878146	rejected	0.000000e+00	-3.432346	-2.862422	-2.567239
Piave	-27.622870	rejected	0.000000e+00	-3.432343	-2.862421	-2.567239
Po	-12.254129	rejected	9.394215e-23	-3.432351	-2.862424	-2.567241
Reno	-12.616197	rejected	1.614895e-23	-3.432350	-2.862424	-2.567240
Tanaro	-16.690653	rejected	1.489486e-29	-3.432346	-2.862422	-2.567240
Tevere	-19.626983	rejected	0.000000e+00	-3.432345	-2.862421	-2.567239
Ticino	-18.957490	rejected	0.000000e+00	-3.432345	-2.862422	-2.567239

Table 8. Trend Statistics, critical and p values for **Precipitations** time series on the major rivers in the **mouth** locations considering daily distribution.

5.5. Kwiatkowski–Phillips–Schmidt–Shin (KPSS) Tests

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) is a stationary test that determines if a time series is stationary around a mean or linear trend or is non-stationary due to a unit root. We will use it as a confirmation test due to the possible false-positive given by the Augmented Dickey-Fuller test. Contrary to the unit root tests, the KPSS has a null hypothesis that confirms the stationary of a time series, and it has as an alternative the presence of a unit root. [6] The KPSS test is based on linear regression. It breaks up a series into three parts: a deterministic trend (β_t), a random walk (r_t), and a stationary error (ε_t), with the regression equation:

$$x_t = r_t + \beta_t + \varepsilon_1$$

If the data is stationary, it will have a fixed element for an intercept or the series will be stationary around a fixed level. The test uses OLS[5] find the equation, which differs slightly depending on whether you want to test for level stationarity or trend stationarity.

5.5.1. KPSS on Discharge Variations

River's Source	KPSS Statistics	H0	p-value	# Lag	Critic 10%	Critic 5%	Critic 2.5%	Critic 1%
Adda	0.495335	**rejected**	0.042	33	0.347	0.463	0.574	0.739
Adige	0.100317	accepted	0.1	35	0.347	0.463	0.574	0.739
Arno	0.218688	accepted	0.1	32	0.347	0.463	0.574	0.739
Oglio	0.815674	**rejected**	0.01	36	0.347	0.463	0.574	0.739
Piave	0.518490	**rejected**	0.037	36	0.347	0.463	0.574	0.739
Po	0.624961	**rejected**	0.020	34	0.347	0.463	0.574	0.739
Reno	0.631529	**rejected**	0.019	30	0.347	0.463	0.574	0.739
Tanaro	0.370061	accepted	0.09	34	0.347	0.463	0.574	0.739
Tevere	0.122100	accepted	0.1	34	0.347	0.463	0.574	0.739
Ticino	0.183403	accepted	0.1	36	0.347	0.463	0.574	0.739

Table 9. Trend Statistics, lag value, critical and p values for **Discharge** time series on the major rivers in the **source** locations considering daily distribution.

[5] Ordinary Least Squares (OLS) Regression is a way to find the line of best fit for a set of data. It creates a model that minimizes the sum of the squared vertical distances (residuals). The Ordinary Least Squared Estimator's formula is $\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$ with $\mathbf{X}$ the regressor matrix, $\mathbf{X}^\top$ its transpose and y the vector with values of the response variable. [7]

River's Source	KPSS Statistics	H0	p-value	# Lag	Critic 10%	Critic 5%	Critic 2.5%	Critic 1%
Adda	0.229110	accepted	0.1	35	0.347	0.463	0.574	0.739
Adige	0.254982	accepted	0.1	35	0.347	0.463	0.574	0.739
Arno	0.157816	accepted	0.1	34	0.347	0.463	0.574	0.739
Oglio	0.899200	**rejected**	0.01	35	0.347	0.463	0.574	0.739
Piave	0.458329	accepted	0.052	35	0.347	0.463	0.574	0.739
Po	0.302339	accepted	0.1	34	0.347	0.463	0.574	0.739
Reno	0.169223	accepted	0.1	33	0.347	0.463	0.574	0.739
Tanaro	0.176998	accepted	0.1	32	0.347	0.463	0.574	0.739
Tevere	0.191386	accepted	0.1	35	0.347	0.463	0.574	0.739
Ticino	0.215012	accepted	0.1	36	0.347	0.463	0.574	0.739

Table 10. Trend Statistics, lag value, critical and p values for **Discharge** time series on the major rivers in the **mounth** locations considering daily distribution.

5.5.2. KPSS on Temperatures Variations

River's Source	KPSS Statistics	H0	p-value	# Lag	Critic 10%	Critic 5%	Critic 2.5%	Critic 1%
Adda	0.059025	accepted	0.1	36	0.347	0.463	0.574	0.739
Adige	0.057343	accepted	0.1	36	0.347	0.463	0.574	0.739
Arno	0.076314	accepted	0.1	36	0.347	0.463	0.574	0.739
Oglio	0.073241	accepted	0.1	36	0.347	0.463	0.574	0.739
Piave	0.052111	accepted	0.1	36	0.347	0.463	0.574	0.739
Po	0.076274	accepted	0.1	36	0.347	0.463	0.574	0.739
Reno	0.072082	accepted	0.1	36	0.347	0.463	0.574	0.739
Tanaro	0.098632	accepted	0.1	36	0.347	0.463	0.574	0.739
Tevere	0.086362	accepted	0.1	36	0.347	0.463	0.574	0.739
Ticino	0.076150	accepted	0.1	36	0.347	0.463	0.574	0.739

Table 11. Trend Statistics, lag value, critical and p values for **Temperature** time series on the major rivers in the **source** locations considering daily distribution.

River's Source	KPSS Statistics	H0	p-value	# Lag	Critic 10%	Critic 5%	Critic 2.5%	Critic 1%
Adda	0.073973	accepted	0.1	36	0.347	0.463	0.574	0.739
Adige	0.067674	accepted	0.1	36	0.347	0.463	0.574	0.739
Arno	0.157816	accepted	0.1	34	0.347	0.463	0.574	0.739
Oglio	0.072908	accepted	0.1	36	0.347	0.463	0.574	0.739
Piave	0.060004	accepted	0.1	36	0.347	0.463	0.574	0.739
Po	0.071562	accepted	0.1	36	0.347	0.463	0.574	0.739
Reno	0.072082	accepted	0.1	36	0.347	0.463	0.574	0.739
Tanaro	0.076827	accepted	0.1	36	0.347	0.463	0.574	0.739
Tevere	0.073275	accepted	0.1	36	0.347	0.463	0.574	0.739
Ticino	0.067206	accepted	0.1	36	0.347	0.463	0.574	0.739

Table 12. Trend Statistics, lag value, critical and p values for **Temperature** time series on the major rivers in the **mouth** locations considering daily distribution.

River's Source	KPSS Statistics	H0	p-value	# Lag	Critic 10%	Critic 5%	Critic 2.5%	Critic 1%
Adda	0.184276	accepted	0.1	24	0.347	0.463	0.574	0.739
Adige	0.199667	accepted	0.1	23	0.347	0.463	0.574	0.739
Arno	0.145617	accepted	0.1	23	0.347	0.463	0.574	0.739
Oglio	0.129654	accepted	0.10	24	0.347	0.463	0.574	0.739
Piave	0.225768	accepted	0.1	19	0.347	0.463	0.574	0.739
Po	0.085556	accepted	0.1	14	0.347	0.463	0.574	0.739
Reno	0.058867	accepted	0.1	17	0.347	0.463	0.574	0.739
Tanaro	0.370061	accepted	0.09	34	0.347	0.463	0.574	0.739
Tevere	0.122724	accepted	0.1	25	0.347	0.463	0.574	0.739
Ticino	0.087664	accepted	0.1	23	0.347	0.463	0.574	0.739

Table 13. Trend Statistics, lag value, critical and p values for **Precipitations** time series on the major rivers in the **source** locations considering daily distribution.

River's Source	KPSS Statistics	H0	p-value	# Lag	Critic 10%	Critic 5%	Critic 2.5%	Critic 1%
Adda	0.191511	accepted	0.1	23	0.347	0.463	0.574	0.739
Adige	0.094076	accepted	0.1	19	0.347	0.463	0.574	0.739
Arno	0.120443	accepted	0.1	23	0.347	0.463	0.574	0.739
Oglio	0.116544	accepted	0.1	19	0.347	0.463	0.574	0.739
Piave	0.089738	accepted	0.1	17	0.347	0.463	0.574	0.739
Po	0.094542	accepted	0.1	19	0.347	0.463	0.574	0.739
Reno	0.075756	accepted	0.1	22	0.347	0.463	0.574	0.739
Tanaro	0.111841	accepted	0.1	18	0.347	0.463	0.574	0.739
Tevere	0.097576	accepted	0.1	21	0.347	0.463	0.574	0.739
Ticino	0.215738	accepted	0.1	20	0.347	0.463	0.574	0.739

Table 14. Trend Statistics, lag value, critical and p values for **Precipitations** time series on the major rivers in the **mouth** locations considering daily distribution.

5.5.3. KPSS on Precipitations Variations

5.6. Auto-correlations and Partial Auto-correlations Functions

Auto-correlation and partial auto-correlation are measures of association between current and past series values and indicate which past series values are most useful in predicting future values. With this knowledge, we can determine the order of processes in a possible usage of ARIMA or SARIMA model during the modelling phase. More specifically:

1. Autocorrelation function (ACF). At lag k, this is the correlation between series values that are k intervals apart.

2. Partial autocorrelation function (PACF). At lag k, this is the correlation between series values that are k intervals apart, accounting for the values of the intervals between.

So, ACF and PACF determinate also the lag order for the Auto-regressive (AR) and Moving Average (MA) used in ARIMA[6] in order to find the optimal values in relation to the best correlations between two lagged value in the time series corresponding to the maximum values on the no-zero series from the starting point ($lag = 0$). ACF and PACF work only on stationary time series. So, due to the non-stationary behavior of some mouths or sources, we carried out ACF and PACF on the difference with previous values that smoothes the curve and make the time series stationary. The number of times we carried differential is related to the value of parameter d given in input to the ARIMA training. According to the KPSS results, we perform different functions on the non-stationary time series. They are strictly referred to on the discharge values distribution over time for rivers like Adda, Piave, Po,

[6] ARIMA is an acronym that stands for Auto-regressive Integrated Moving Average. It is a generalization of the simpler Auto-regressive Moving Average (ARMA) and adds the notion of integration corresponding to the number of difference carried out to make the time series stationary.

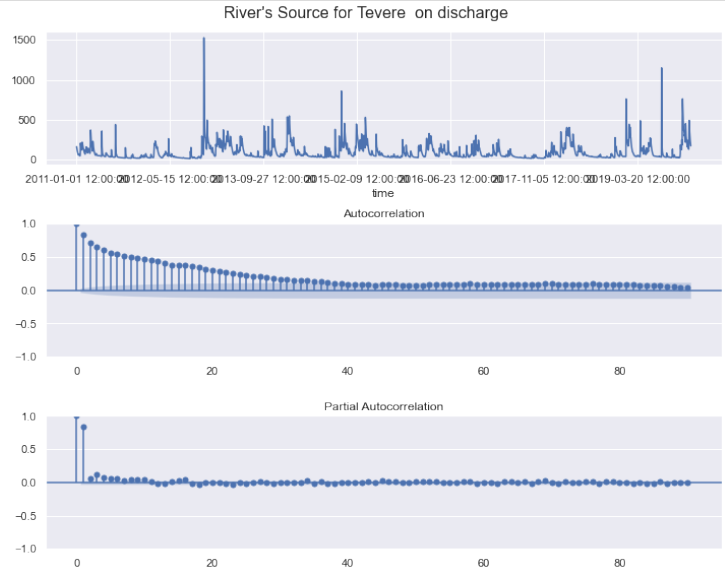

Fig. 7. Time-series, ACF and PACF plots of Tevere river's source for discharge distribution.

and Reno sources and Oglio source and mouth. Finally, due to the large plots, we avoided showing every plot related to discharges, precipitations, and temperatures time series on mouths and sources of the ten major Italian rivers. In the following figure, we can show one of the plots we displayed to calculate the p, q, and d values for a possible ARIMA model defining the auto-correlation and partial-auto-correlation of our time series.

5.7. Granger Causality

The Granger causality test is a statistical hypothesis test for determining whether one time series is a factor, offers helpful information in forecasting another time series, and determines how a time series with, a specific lag value, can be causal-dependant to another. In other words, according to Granger causality definition, if a signal X "Granger-causes" (or "G-causes") a signal Y, then past values of X should contain information that helps predict Y above and beyond the information contained in past values of Y alone. The null hypothesis for the test is that lagged X does not explain the variation in Y.

5.7.1. Lag Order

Granger causality always has a p past values of each of the two variables in the bivariate test. This value is the lag order which is equal to the number terms back down the auto-regressive process we want to test for serial correlation applied to the prediction given by $y \in Y$:

$$y_{i,t} = \alpha + \sum_{l=1}^{p} \beta_l y_{i,t-l} + \gamma_l x_i t - l + \varepsilon_{i,t}$$

A conventional way to choose the optimal p value corresponding to the best lag is given by some experiments to an auto-regression like vector autoregressive (VAR) [8] with various p lag and keep track of the AIC[7] for each chosen lag.

Generally, the lag order is more than 30 values, so we find the best fit considering the time series's length of 30 days, which has one value per day for ten consecutive years. We can also change the level of detail that can be transformed from daily order to a monthly approximation simplifying time series and its distribution and observing if there are some changing.

5.7.2. Performing Test on Daily Distribution

Due to internal Granger functionality, some results are equal to 0.000; this is interpreted as a high significance between X on Y. During our observations, we compute the Granger matrix using F-Test and compare how any monitored attributed can "G-cause" another one. The previous table gives the lag order used during tests.

[7]The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. The number of independent variables calculates it and the maximum likelihood estimate on the model training [9]

Rivers	Sources Lag Orders	Sources AIC	Mouths Lag Orders	Mouths AIC
Adda	32	-146.484	33	-127.986
Adige	31	-141.397	33	-125.643
Arno	33	-157.461	32	-130.121
Oglio	33	-158.697	33	-132.128
Piave	33	-159.434	32	-130.440
Po	31	-137.747	33	-114.45
Reno	31	-136.953	33	-160.606
Tanaro	32	-152.134	31	-124.248
Tevere	33	-132.257	32	-125.109
Ticino	33	-125.035	32	-126.586

Table 15. Lag Orders and AIC values for sources and mouths on the ten major italian rivers.

Discharge Causality on Sources

Rivers Sources	Temperatures	Precipitations
Adda	0.0053	0.0359
Adige	0.0	0.1389
Arno	0.0	0.5376
Oglio	0.0	0.4689
Piave	0.0041	0.0
Po	0.2808	0.0168
Reno	0.0000	0.0648
Tanaro	0.0176	0.0000
Tevere	0.0000	0.5516
Ticino	0.0006	0.0064

Discharge Causality on Mouths

Rivers Mouths	Temperatures	Precipitations
Adda	0.6319	0.0002
Adige	0.0000	0.0025
Arno	0.0000	0.1732
Oglio	0.0572	0.1873
Piave	0.7236	0.0436
Po	0.3418	0.0177
Reno	0.0000	0.0000
Tanaro	0.0032	0.0
Tevere	0.0000	0.7546
Ticino	0.0155	0.0483

In previous tables, we can check the results of the Granger Causality Test using the F-Test approach on the causal dependency between discharges and temperatures or precipitations. Columns represent the X values of the test. Meanwhile, the row is the Y value. In every cell, we put the p-value less than 0.05 in case of rejection. Rejecting the null hypothesis means the X value "G-causes" the Y values. We can observe that sources and mouths causal dependencies change. Generally, the discharge has a causal correlation with temperatures also given to the melting glaciers on the Alps or Apennines and the increment of the year temperature mean that can decrease the level of the river flow for summer period every year. Precipitations can also influence the river discharge, especially if the mean discharge value is small. However, we can observe that precipitation has different causal relations with the discharge; maybe it is due to the nature of the river, or it is a topology in terms of water usage, presence of lakes, or aggregation with tributaries. Due to the large lag, we can affirm that the effects and correlations between data are to the monthly approximation level. So, we will perform a new data aggregation using the monthly mean of the current daily distributions and use the result dataset in a new statistical analysis focusing on the same procedure.

5.8. Monthly Approximation

The Monthly approximation is a simple procedure of aggregation. We perform a simple algorithm to calculate each sample location's arithmetic mean of the month's days. We will test some analyses on the distribution at a different approximation level. Monthly behavior may have some hidden pattern that can decrease the variations of values over time. In other words, we can extrapolate some data from a different point of temporal view and check if some interesting hidden patterns exist. In this section, we will filter information related to the re-execution of the analysis procedure only on the exciting part. For more details, we suggest checking results in the notebook on the GitHub repository.

5.9. ADF and KPSS Approximation

This section will get some effects on the ADF and KPSS on the monthly distribution to check if there is a monthly trend given by a specific unit root or not. During the daily tests, we found stationary behavior on daily variation. However, changing the level of approximation can affect results.

5.9.1. Augmented Dickey Fuller Test on Monthly Dataset

ADF on Rivers Sources Discharges

Rivers Sources	ADF Statistics	P-Value
Adda	-6.567967	8.069253e-09
Adige	-6.121443	8.835477e-08
Arno	-2.839437	5.285754e-02
Oglio	-6.127887	8.541239e-08
Piave	-2.208336	**0.203234**
Po	-7.201822	2.354657e-10
Reno	-1.540703	**5.132888e-01**
Tanaro	-7.043791	5.756039e-10
Tevere	-5.011453	2.106357e-05
Ticino	-7.024598	6.412734e-10

ADF on River Mouths Discharges

Rivers Mouths	ADF Statistics	P-Value
Adda	-7.721226	1.191165e-11
Adige	-5.771534	5.378169e-07
Arno	-4.224085	0.000600
Oglio	-4.866510	0.000041
Piave	-2.997592	3.511992e-02
Po	-2.469483	**1.230770e-01**
Reno	-6.066460	1.178585e-07
Tanaro	-6.561707	8.349688e-09
Tevere	-4.744245	6.944057e-05
Ticino	-7.052417	5.483035e-10

ADF on Rivers Sources Temperatures

Rivers Sources	ADF Statistics	P-Value
Adda	-2.193495	**2.086148e-01**
Adige	-2.077884	**2.535001e-01**
Arno	-2.633550	**8.625943e-02**
Oglio	-9.027230	5.567719e-15
Piave	-2.271250	**0.181408**
Po	-2.633381	**8.629238e-02**
Reno	-2.005648	**2.841020e-01**
Tanaro	-2.182530	**2.126471e-01**
Tevere	-2.607534	**9.145019e-02**
Ticino	-2.395004	**1.431516e-01**

ADF on River Mouths Temperatures

Rivers Mouths	ADF Statistics	P-Value
Adda	-2.175270	**2.153430e-01**
Adige	-2.589741	**9.513610e-02**
Arno	-2.216527	**0.200302**
Oglio	-2.445217	**0.129379**
Piave	-2.810984	**5.672183e-02**
Po	-2.378454	**1.479102e-01**
Reno	-1.913159	**3.258964e-01**
Tanaro	-2.330310	**1.623787e-01**
Tevere	-10.636880	5.045647e-19
Ticino	-2.209746	**2.027274e-01**

ADF on Rivers Sources Precipitations

River Sources	ADF Statistics	P-Value
Adda	-6.455610	1.486084e-08
Adige	-6.804305	2.196596e-09
Arno	-7.872429	4.944170e-12
Oglio	-6.119698	8.916831e-08
Piave	-8.170576	8.644934e-13
Po	-3.932638	1.807356e-03
Reno	-6.708752	3.726896e-09
Tanaro	-9.479044	3.917762e-16
Tevere	-6.598285	6.837191e-09
Ticino	-5.794538	4.785909e-07

ADF on River Mouths Precipitations

River Sources	ADF Statistics	P-Value
Adda	-5.313436	5.135450e-06
Adige	-3.932894	1.805669e-03
Arno	-6.760010	2.807806e-09
Oglio	-5.639245	1.045707e-06
Piave	-6.703442	3.837617e-09
Po	-7.077587	4.757544e-10
Reno	-6.712822	3.644166e-09
Tanaro	-8.688959	4.089926e-14
Tevere	-7.077587	4.757544e-10
Ticino	-4.622027	1.176179e-04

Precipitations are similar to the daily distribution; we do not have trends. The majority of the climate models agree in expecting, in the Mediterranean area, an increase in the frequency of extreme precipitation events while intensity will be almost unchanged [1], so the precipitation means does not vary on the monthly approximation. However, temperatures have a unit root on monthly approximation. So, we need to difference it and update the d value on possible forecasting problems like ARIMA during the modeling phase. The lack of data gives another problem. During the monthly approximation, we need to focus if the ADF results were obtained without false positives due to the dataset reductions. So, we need also to perform the KPSS and determine if the KPSS cannot reject the null hypothesis caused by the lack of observations on the monthly approximation.

5.9.2. Kwiatkowski–Phillips–Schmidt–Shin Test on Monthly Dataset

KPSS on Rivers Sources Discharges

Rivers Sources	KPSS Statistics	P-Value	# Lags
Adda	0.408544	0.073472	2
Adige	0.060179	0.1	4
Arno	0.138444	0.1	3
Oglio	0.413383	0.071386	4
Piave	0.353740	0.097095	3
Po	0.433793	0.062589	2
Reno	0.393776	0.079838	2
Tanaro	0.239712	0.1	3
Tevere	0.069325	0.1	4
Ticino	0.126551	0.1	3

KPSS on River Mouths Discharges

KPSS Statistics	P-Value	# Lags
0.171122	0.1	2
0.147833	0.1	4
0.084010	0.1	4
0.363838	0.092742	5
0.254067	0.1	4
0.204909	0.1	3
0.080345	0.1	4
0.125055	0.1	2
0.101328	0.1	4
0.146455	0.1	3

KPSS on Rivers Sources Temperature

Rivers Sources	KPSS Statistics	P-Value	# Lags
Adda	0.030071	0.1	5
Adige	0.029390	0.1	5
Arno	0.038300	0.1	5
Oglio	0.037008	0.1	5
Piave	0.026814	0.1	5
Po	0.039340	0.1	5
Reno	0.036568	0.1	5
Tanaro	0.049974	0.1	5
Tevere	0.043536	0.1	5
Ticino	0.038711	0.1	5

KPSS on River Mouths Temperature

KPSS Statistics	P-Value	# Lags
0.037882	0.1	5
0.034600	0.1	5
0.042094	0.1	5
0.037135	0.1	5
0.030636	0.1	5
0.036511	0.1	5
0.038438	0.1	5
0.039018	0.1	5
0.037527	0.1	5
0.034299	0.1	5

ADF on Rivers Sources Temperature

Rivers Sources	ADF Statistics	P-Value	# Lags
Adda	0.121494	0.1	2
Adige	0.121501	0.1	3
Arno	0.098835	0.1	3
Oglio	0.085763	0.1	2
Piave	0.176357	0.1	3
Po	0.090423	0.1	1
Reno	0.051041	0.1	1
Tanaro	0.179420	0.1	1
Tevere	0.086374	0.1	1
Ticino	0.086385	0.1	0

ADF on River Mouths Temperature

ADF Statistics	P-Value	# Lags
0.231055	0.1	3
0.068880	0.1	3
0.076742	0.1	2
0.094330	0.1	1
0.071886	0.1	2
0.069340	0.1	3
0.064576	0.1	3
0.125354	0.1	2
0.059663	0.1	2
0.238620	0.1	0

If the KPSS cannot reject the null hypothesis and ADF, the total number of observations is insufficient to determine if the time series is stationary. In other words, we cannot use stochastic models like auto-regressive or ARIMA to generate some forecasting prediction because these methods has a pre-condition that the time series is stationary.

5.10. Granger Causality Test Adaptation

After updating the best lag for each time series, we can perform the Granger Causality and determine if the test obtains different results than the daily distributions.

Discharge Causality on Sources

Rivers Sources	Temperatures	Precipitations
Adda	0.0001	0.0752
Adige	0.0000	0.0859
Arno	0.0	0.5376
Oglio	0.0000	0.2479
Piave	0.0002	0.0051
Po	0.0874	0.0158
Reno	0.0201	0.5582
Tanaro	0.1198	0.0301
Tevere	0.0000	0.2934
Ticino	0.0000	0.0020

Discharge Causality on Mouths

Rivers Mouths	Temperatures	Precipitations
Adda	0.1568	0.0369
Adige	0.0001	0.0097
Arno	0.0028	0.4386
Oglio	0.2674	0.3709
Piave	0.1654	0.1208
Po	0.2862	0.0538
Reno	0.0005	0.0231
Tanaro	0.0505	0.1022
Tevere	0.0000	0.1019
Ticino	0.0005	0.0055

Due to the mean relations between daily and monthly values, the Granger Causality Test made similar results. Lag is set to 3 months to see a different lag order and maintains similar correlations between means of months and days in the short term. Furthermore, except for Po and Tanaro, every source has the temperatures that "G-causes" the discharge value. However, Tanaro and Po have their discharge with some causal dependencies on the precipitations. So, every source has at least one causal dependency between precipitations or temperatures. On the mouths side, the Apennines rivers have causal dependencies with temperatures; most Po and its tributaries do not. Ticino and Adige have both causal dependencies. Piave lost its high causal correlations with temperatures and precipitations, maybe due to the lack of atmospheric side effects like the Alps melting glaciers.

5.11. Statistical Conclusions

Our initial hypothesis was related to finding the correlation between temperatures and precipitations on the discharges. Using causality tests like Granger Causality Test, we determined temperatures and precipitations "G-cause" the discharge over its variations. Results show that the causal-dependency of temperatures and precipitations is relative to the specific time series, according to the distributions of the daily values. However, monthly approximations show that mean temperature and precipitations "G-cause" the mean discharge over time for some locations. In other words, discharges have causality dependencies with precipitations or temperatures in the short-term period depending on the rivers or their geographical locations. Furthermore, the proximity to the Alps or Apennines may be a spatial correlation with the discharge behavior because the temperatures can directly affect melting glaciers or snow on the mountains near the sources of the rivers.

6. Data Modelling

According to the Data Analysis assumption, we will define some prediction models related to the correlation with temperatures and precipitations where these variables are causal dependant on the river location of our interest. In other words, we can predict the future discharge values of a given point using different approaches based on the impact of temperatures and precipitation variations on the river flow. In this notebook, we will analyze different algorithms and choose the best one for both approaches specified at the end of the data analysis notebook. In detail, we can test the prediction efficiency using RMSE and other evaluations methods on the testing set (composed of samples on the 2021-2022 periods). During the application usage, models will be trained on the complete dataset to improve the future prediction of the period above 2022 and will be validate on only the 10% of the total data related to the near present period. Furthermore, due to the lack of pre-condition verification, ADF and KPSS make monthly data inapplicable for the traditional statistic forecasting method like ARIMA. So, we will use the daily distribution on ARIMA techniques and let the monthly version on the LSTM.

6.1. Data Splitting

For each time series, dataset coverage is a period equals to 10 years and 5 months (2011-2022). The Training and Validation set is composed by the 90% (70% and 20% respectively) of the entire time series starting for the first measurements in the 01-01-2011 until 27-03-2021. Motivations behind this split ratio is correlated to the short and long term predictions support that we want to obtain using the same model for a specific location.

6.2. *ARIMA Modelling*

ARIMA is a statistical technique to forecast values on a time series proposed by Box and Jenkin in 1976 [11]. It is a model derived from the combinations of Auto-regressive (AR) and Moving Average (MA) models. It also uses an integrated (I) approach where the use of differencing of raw observations [11] [12] (e.g., subtracting an observation from observation at the previous time step) to make the time series stationary. So, ARIMA has the prerequisite to have in input a training time series that is stationary. It has three different parameters used in the following equation:

$$\Phi_p(B)W_t = \Theta_q(B)e_t$$

$\Phi_p(B)$ is the auto-regressive operator of order p which is the lag order that defines the lag observations included in the model. $W_t = \Lambda dX_t$ that is the result equations proportionally to the d length which is the number of times that observations are differenced to obtain a stationary time series. Finally, $\Theta_q(B)$ is the moving average operator of order q that is equal to the size of the moving windows. The d values are given by the non-stationary test retrieved by KPSS and ADF results; meanwhile, p and q are defined with ACF and PACF applications. During the Data Modelling, we will discuss only a sample time series to evaluate the proposed model's effectiveness and efficiency. Then, if the results have good performances, we will extend these models to every time series of our dataset with opportune parameters. Due to the autocorrelation approach, ARIMA has in input only the time series of the discharges.

6.2.1. Training Phase

Due to the adaptability of the model in a runtime environment suitable in the final application, we will pre-configure p, q and d using the *auto_arima* function that retrieves automatically the best fit configuration defined during the ACF, PACF and stationary or unit root tests to support and propagate a model definition for every location on our dataset. In other words, the advantage of using Auto ARIMA over the ARIMA model is that after the data preprocessing step, we can skip the next steps and directly fit our model. It uses the AIC (Akaike Information Criterion) and BIC(Bayesian Information Criterion) values generated by trying different combinations of p, q, and d values to fit the model. Furthermore, the result model can plot some diagnostics for standardized residuals of one endogenous variable. The standardized residual measures the strength of the difference between

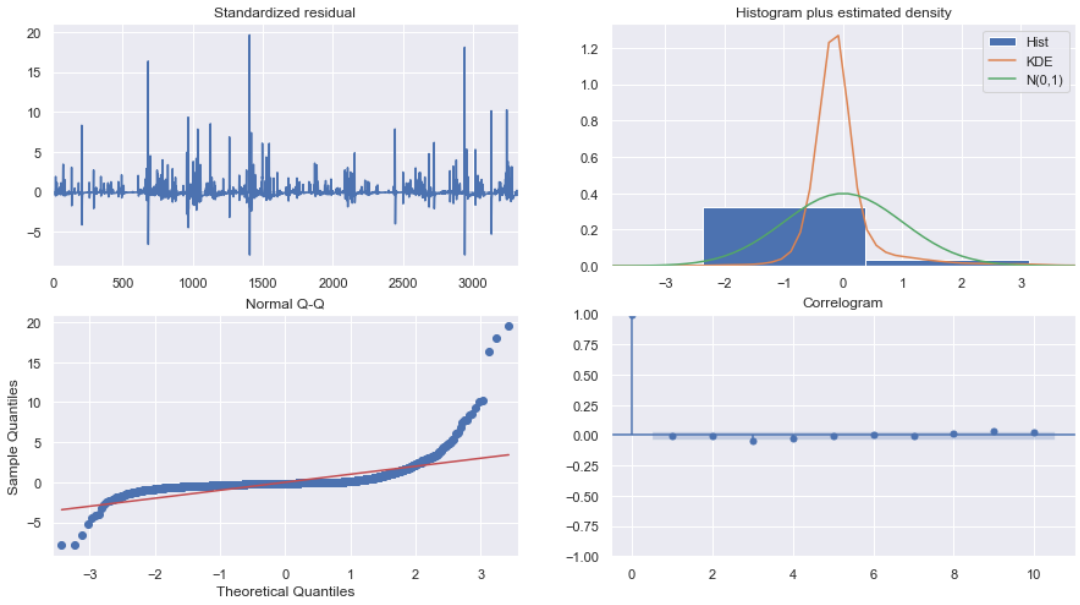

Fig. 8. Matrix of 2×2 plots. In order: standardized residuals over time, histogram plus estimated density of standardized residuals, along with a Normal(0,1) density plotted for reference, normal Q-Q plot, with Normal reference line, correlogram

.

observed and expected values for normalized or no-normalized distributions for outliers detection. The density histogram of the residuals shows the KDE of the standardized residuals and the normal distribution with 0 mean and one standard deviation. A normal Q-Q plot is a graphical tool to help us assess if a set of data plausibly came from a normal distribution. A correlogram is commonly used to detect randomness in the time series [10]. Figure

8. shows the presence of outliers and how the time series is considered random from the ARIMA results given by the zero value of the correlogram.

6.2.2. Testing Phase

Initially, the testing phase measures only on the selected sample tying to propose a forecasting series inside the testing interval without considering true values. In other words, we performs a one-step forecasting starting from the training set tail until the end of the testing set.

6.3. LSTM

LSTM stands for Long Short-Term Memory and is an artificial neural network that uses feedback connections to create a kind of short-term memory that makes it able to not only process single data points but series of points [16]. The LSTM is composed of a cell and three gates, the cell is the component assigned to remembering the data over a time interval while the three gates decide the input, the output and what to forget to avoid overfitting the model. The ability of the LSTM to remember data over an arbitrary interval of time makes it suitable for predicting or classifying time series data with an unknown span of time between events of interest in a time series. The strength of the LSTM with respect to the ARIMA model is that it is not a statistical model, so it can consider patterns that stay hidden in the ARIMA model.

6.3.1. Data Adaptation

Two models are created to use the LSTM for predicting the discharge information. The first one is a model that given a series o temperatures predict the next value for the temperature. The second one is the model that given a series of temperature and discharge values predicts the next value for the discharge.
To make the LSTM work the data is grouped by averaging the values for each measure over the time span of a month and then the data itself is enriched with the temporal information of the period to which the data refers to. So the values for the months are modelled over the sin and cos functions to transport the temporal information inside the model. After a number of tests we found out that the best configuration we could get was with a LSTM that uses 32 neurons and two dense layers; the first layer has 12 features and the second layer has 1. This choice is made because the model takes into consideration the previous 12 months to predict the next one, so it tries to consider the 12 months and condense it into the final feature that is then used to predict the future value. The precipitation values are not used for two reasons. The first one is that trying to predict the discharge using the precipitation values with the others showed a lower precision; the second is a more practical reason for which is really difficult to predict the future values for the precipitation since we are basically trying to do a more accurate weather forecasting.

6.3.2. Training Phase

The model uses the 70% of the dataset for the training phase, and it goes on until the loss function doesn't change significantly. In the training phase the 20% of the dataset is used for the validation. The two models are trained on data scaled in a different way; the temperature model is trained using the original data for the temperatures, because it shows that it can predict the values for the temperature in a more precise way. In the case of the discharge model the data needs to be normalized because the values for the discharge have a high variance so without the normalization the model will have trouble in learning to predict the future values.

6.3.3. Testing Phase

The testing phase is done using the 10% of the dataset and while the temperature model shows a good precision the model for the discharge is highly influenced by the characteristics of the river such as if the river has a lot of tributaries, the position of the considered point of the river and other factors that add to these.

6.4. Models Evaluation

Models evaluations provide information about the general performances of proposed forecasting models described in the section above. During this phase, we perform RMSE and MAE from the difference between the predictions given by the forecasting procedure and the real values from the testing set. Due to the different forecasting approaches, we normalize metrics using the interquartile range method:

$$NRMSE = \frac{RMSE}{||Q_1 - Q_3||} \text{ with } RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n}(d_i - f_i)^2}{n}\right)}$$

$$NMAE = \frac{MAE}{||Q_1 - Q_3||} \text{ with } MAE = \frac{\sum_{i=1}^{n}||d_i - f_i||}{n}$$

d_i is the prediction value at the time i, f_i is the true value simultaneously, and Q_1 and Q_3 are the 25th and 75th percentile of the true values. Finally, we can show in the table below metrics results for ARIMA and LSTM model on the sources and mouths samples and, finding the best fit for each model on the samples set; we can choose the final modeling approaches for the final application.

ARIMA Models Results on Sources

Rivers Sources Input Series	NRMSE	NMAE
Adda	2.159	2.941
Adige	1.266	0.797
Arno	12.030	4.836
Oglio	1.434	0.763
Piave	0.883	0.642
Po	2.311	2.059
Reno	2.862	1.104
Tanaro	1.603	0.89
Tevere	1.406	0.800
Ticino	0.867	0.616

ARIMA Models Results on Mouths

Rivers Mouths Input Series	NRMSE	NMAE
Adda	1.417	0.839
Adige	0.872	0.600
Arno	2.016	1.052
Oglio	0.961	0.804
Piave	1.00	0.656
Po	0.990	0.781
Reno	3.971	2.600
Tanaro	2.827	2.077
Tevere	1.254	0.744
Ticino	0.810	0.590

LSTM Models Results on Sources

Rivers Sources Input Series	NRMSE	NMAE
Adda	1.031	0.838
Adige	0.488	0.352
Arno	41.85	40.49
Oglio	0.727	0.474
Piave	1.123	0.940
Po	3.286	2.030
Reno	0.939	0.790
Tanaro	0.602	0.534
Tevere	1.178	0.708
Ticino	0.488	0.427

LSTM Models Results on Mouths

Rivers Mouths Input Series	NRMSE	NMAE
Adda	1.102	0.880
Adige	0.576	0.528
Arno	0.918	0.774
Oglio	1.222	0.986
Piave	0.905	0.765
Po	0.444	0.370
Reno	4.424	4.348
Tanaro	1.732	1.484
Tevere	1.017	0.816
Ticino	0.720	0.673

From the previous table understanding that the LSTM performs better than the ARIMA model is not immediate. The reason why the LSTM is better is that it shows lower errors where the model works fine, while it has a higher error where the model finds more difficulties. The LSTM also shows a better ability to follow the tendency of the value of the discharge. So it at least helps in understanding if the trend is increasing or decreasing.

7. Final Application

The final Application focuses on the Client-Server design pattern. A client has an interface to select a specific location related to one of the ten major Italian rivers. Then, it visualizes discharges, temperatures, and precipitations with the possibility of filtering data in a specific interval from 2011 to 2022 and hiding or showing graphs interactively. A map disposes rivers locations on a map built using Open Layer framework; forecasting procedures define a request to the server-side with two different modes: Online and Offline. A selector allows the user to select the mode. If the flag is set to online, the client requests the server to load the dataset corresponding to the history of the selected location, train the model and then predict future values using one-step forecasting; the maximum number of predictions is set to 12 months. However, due to the lengthy procedures to set up the forecasting model, we also built an offline forecasting mode. If the mode flag is set to offline, the client performs a data request to the server, retrieving predictions directly from the server file system and putting them in the response. The entire application is deployed also in Heroku at the following link. Finally, the forecasting is then showed in a line plot like all the other values, of course since the forecasting is done by aggregating the data in a monthly fashion the plot showed is on a monthly scale. Due to the Heroku deployment environment constraints, we can't modify the timeout of the request in an online environment, so we can't perform the online approach on the deployed version. A future solution should be to transfer in a new environment different from Heroku which has configurable metadata.

8. Conclusions

So many works are done to retrieve information about the discharge on the major Italian rivers, especially for essential rivers like Po [13] [14] that are vital for Italian agriculture and the water availability in the North area of Italy. Varying temperatures and precipitations are only two causal dependant attributes on the river discharges; tributary flows usually have a considerable impact, especially in rivers like Po [15]. Another factor we did not consider is the presence of lakes in the river path that can change discharges due to the atmospheric features on a larger area. So, temperatures and precipitations are correlated to external factors that can transform these features into causality elements on the discharge over time. For instance, temperatures have a major impact in locations related to the sources of the river than mouths; we suppose that is correlated to an additional external factor: the ice melting. Meanwhile, precipitations causality is independent of the river's nature and G-causes only in particular locations, but, generally, there is not a strong causality relation with the discharge. In conclusion, we can say that the increment of the mean temperatures over time can negatively affect the ice melting in the source locations with a temporal increment of the discharge values. However, the increment of temperature can even reduce the discharge values, and when it is low, like Oglio, Arno, Ticino, or Tevere, it can have a huge negative impact on total water availability.

References

1. *R. Vezzoli, P. Mercogliano, S. Pecora, A.L. Zollo, C. Cacciamani, Hydrological simulation of Po River (North Italy) discharge under climate change scenarios using the RCM COSMO-CLM, Science of The Total Environment, fVolumes 521–522, 2015, doi:10.1016/j.scitotenv.2015.03.096*
2. *L. Cinquini et al., "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data," 2012 IEEE 8th International Conference on E-Science, 2012, pp. 1-10, doi:10.1109/eScience.2012.6404471.*
3. *J. M. Van Der Knijff, J. Younis & A. P. J. De Roo (2010) LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24:2, 189-212, doi:10.1080/13658810802549154*
4. *Madden, Roland A. and Jill Williams. "The Correlation between Temperature and Precipitation in the United States and Europe." Monthly Weather Review 106 (1978): 142-147.doi:10.1175/1520-0493(1978)106¡0142:TCBTAP¿2.0.CO;2*
5. *Mushtaq, Rizwan, Augmented Dickey Fuller Test (August 17, 2011) doi10.2139/ssrn.1911068*
6. *Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, Yongcheol Shin, Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?, 1992, doi:10.1016/0304-4076(92)90104-Y.*
7. *Kuchibhotla, Arun K. and Brown, Lawrence D. and Buja, Andreas, Modelfree Study of Ordinary Least Squares Linear Regression, 2018, doi:10.48550/arxiv.1809.10538*
8. *Qin, Duo, Rise of VAR Modelling Approach. Journal of Economic Surveys, Vol. 25, Issue 1, pp. 156-174, 2011, doi:10.1111/j.1467-6419.2010.00637.x*
9. *Aho, Ken & Derryberry, DeWayne & Peterson, Teri. (2014). Model selection for ecologists: The worldviews of AIC and BIC. Ecology. 95. 631-6. doi:10.1890/13-1452.1*
10. *Michael Friendly (2002) Corrgrams, The American Statistician, 56:4, 316-324, doi:10.1198/000313002533*
11. *Khashei, M.; Bijari, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl. Soft Comput. 2011, 11, 2664–2675 doi:10.1016/j.asoc.2010.10.015*
12. *Xu, L.; Gao, P.; Cui, S.; Liu, C. A hybrid procedure for MSW generation forecasting at multiple time scales in Xiamen City, China. Waste Manag. 2013, 33, 1324–1331. doi:10.1016/j.wasman.2013.02.012*
13. *Hydrology of the Po River: looking for changing patterns in river discharge, A. Montanari, DICAM, University of Bologna, 2012, doi:10.5194/hessd-9-6689-2012*
14. *Zanchettin, D., Traverso, P. & Tomasino, M. Po River discharges: a preliminary analysis of a 200-year time series. Climatic Change 89, 411–433 (2008). doi:10.1007/s10584-008-9395-z*
15. *Turco, Marco and Vezzoli, Renata and Da Ronco, Pierfrancesco and Mercogliano, Paola, Variation in Discharge, Precipitation and Temperature in Po River and Tributaries Basins (December 2013). CMCC Research Paper No. 185, doi:10.2139/ssrn.2490966*
16. *S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735*