الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Réf :……

Centre Universitaire
Abdelhafid Boussouf Mila

Institut des sciences et de la technologie ***Département de Génie civil et d'hydraulique***

Mémoire préparé En vue de l'obtention du diplôme de Master

Spécialité: Hydraulique urbaine

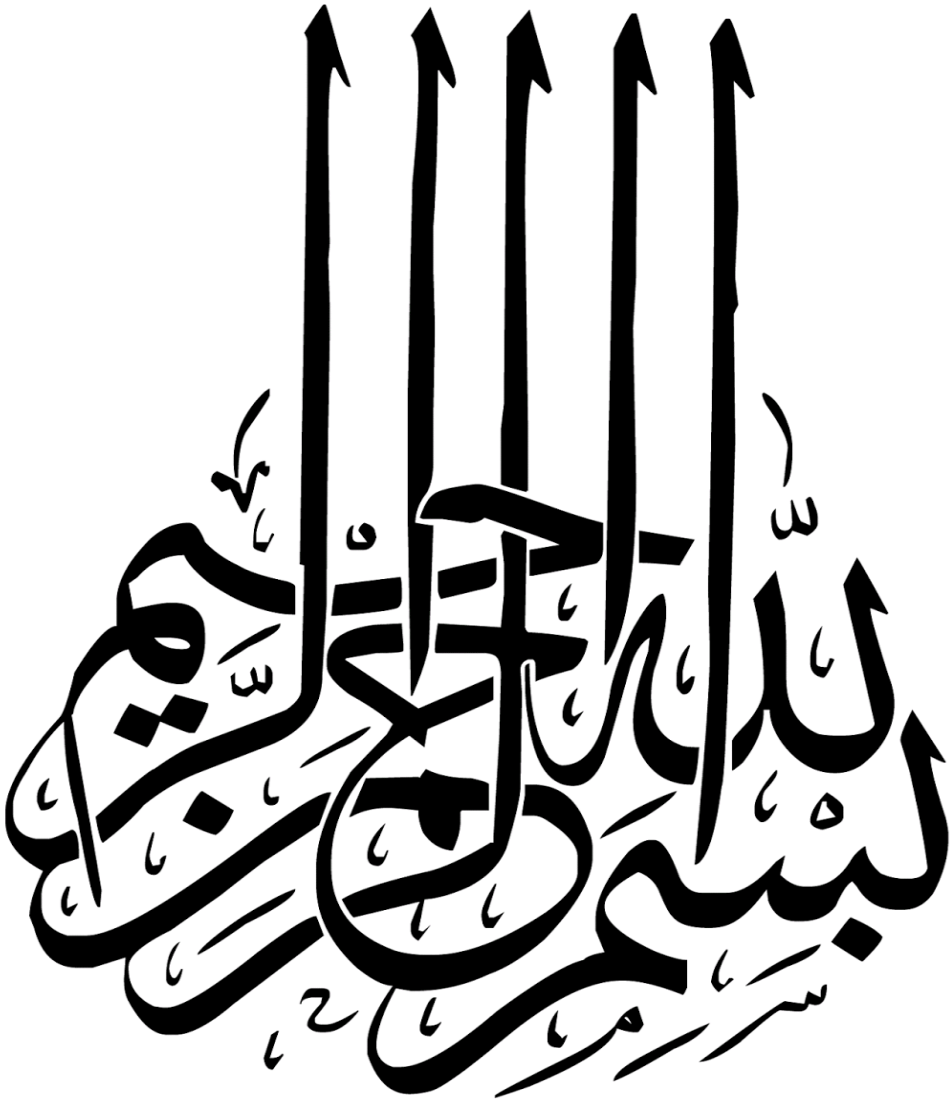
The estimation of climatic missing data using machine learning, case of Algerian stations

Préparé par :
- **Ahmed Islam Boucenna**
- **Abd El Madjid Boudouda**
- **Hadil Azzam**

Soutenue devant le jury

Dr.	**KOUSSA .M**	**Président**
Dr.	**BERHAIL.S**	**Examinateur**
Dr.	**KEBLOUTI.M**	**Encadrant**

Année universitaire : 2022/2023


بسم الله الرحمن الرحيم

DEDICATION

In the name of God, the most gracious, the most merciful

My Lord, who does not make the night sweet except with Your thanksgiving, and the moments are not sweet except with Your remembrance, and the Hereafter does not become sweet except with Your pardon, and Paradise does not become sweet without seeing You.

To my second mother, my dear grandmother, may God have mercy on her, my homeland, a candle that lights up my night and a light that illuminates my life, to my dear aunts, Djamila , Sakina and Samia, who has the kindness and fragrance of the world you.

To the two blessings that God has bestowed on me To the most important treasure in this world and the hereafter To the one when I bow down to kiss her hands and pour the tears of my weakness on her chest To my soul and heart, my beloved mother, and to my beloved father who is the light that illuminates my life, you are the father is referred to and proud of among the people.

To those with whom I knew the meaning of life, my brothers, to the rib of the soul and half of the mother's heart, my sister Maryem Rayan, and to my quarrelsome brother Mohammed Nidal, and to my little brother, my soul, whom I love more than anything, Rida Abd al-Rahman.

To my soul mate Marwa Sarah, thank you very much for your support, encouragement, sincerity and continuous sacrifice throughout my academic career. I did not imagine that I would be able to get through this difficult year without your presence, and I will not be able to return this debt to you as long as I live, with all the love in the world.

BOUCENNA AHMED ISLAM

DEDICATION

In the name of God, the most gracious, the most merciful

My Lord, who does not make the night sweet except with Your thanksgiving, and the moments are not sweet except with Your remembrance, and the Hereafter does not become sweet except with Your pardon, and Paradise does not become sweet without seeing You.

To my second mother, my dear grandmother, my homeland, a candle that lights up my night and a light that illuminates my life. All thanks and credit to you are the fragrance of the world.

To the two blessings that God has bestowed on me To the most important treasure in this world and the hereafter To the one when I bend down to kiss her hands and pour my tears of weakness on her chest To my soul and heart my love. My dear mother, and to my dear father, who is my role model.

To those with whom I learned the meaning of life and the family gathering and its joys, my brothers, to my sisters Rania Qatar Al-Nada and Yasmine, and to my spoiled brother Zakaria Amir

To asma my beautiful, my companion, and my soul mate, who supported me throughout my academic career, my homeland, and all my life, my dear love

BOUDOUDA ABD EL MADJID

DEDICATION

Praise is to God, enough, and prayers and peace be upon the beloved Chosen One, his family, and those who fulfilled. As for what follows: Praise be to God, who has enabled us to value this step in our academic journey with this note of ours, the fruit of effort and success, thanks to Him Almighty, dedicated to the most honorable of creation and to those who gave me life, hope, and upbringing On the passion of learning and knowledge, and those who taught me to ascend the ladder of life with wisdom and patience, righteousness, kindness and loyalty to them: My dear father and my beloved mother, Azzam Lakhdar, and Belhi Maryam, may God Almighty protect them and keep them as a light for my path.

To those whom God gave me the blessing of having in my life, to the strong knot that helped me in my research journey, which is still going on, brothers and sisters, especially my little brother Saif El-Din, my support in life.

To those who supported me as we pave the way together towards success in our scientific career, to my friend and my fiancé and future husband: Bouchaour Fares.

Finally, to everyone who helped me and had a role from near or far in completing this study, asking God Almighty to reward everyone with the best reward in this world and the hereafter. Then to every seeker of knowledge who strives with his knowledge to benefit Islam and Muslims with all that God has given him of knowledge and knowledge.

Azzam Hadil

Thanks

We thank the Almighty God for giving us the health and the will to start and finish this thesis. If it takes a lot of motivation, rigor and enthusiasm to carry out this thesis, then m,this research work needed the contribution of several people, whom I would like to thank! First of all, this work would not be as rich and would not have been possible without the help and supervision of Mr keblouti mehdi, we thank him for the quality of his exceptional supervision, for his patience, his rigor and availability during our preparation of this dissertation. Our thanks also go to all our teachers for their generosity and the great patience they have shown despite their academic and professional workload.

Abbreviation

Radar: Radio detection and ranging.

ML: Machine learning.

AI: Artificial intelligence.

$R^{2:}$ R-squared.

RMSE: Root Mean Squared Error.

MAE: Mean Absolute Error.

MRE: Mean relative error.

ANRH: Agence nationale des ressources hydrique.

N : number of months.

$\bar{x}$: Moyne.

S: standard deviation.

LR: Linear regression.

GP: Gaussian processes.

AR: additive Regression

MP: multilayer preceptor.

MS: Multi Shame.

RF: Random Forest.

IMC: Input Mapped Classifier.

WIHW: Weighted Instances Handler Wrapper.

RSS: Random Sub Space.

RBD: Regression by Discretization.

DT: Decision Table.

DS: Decision Stump.

Index

Chapter II Methods for estimating missing data

Chapter III Collected and critique of data

Chapter IV: Results and discussions

List of Table

List of Figures

Abstract

The rainfall recording process is facing a major problem in our country, Algeria, which is the loss of data, and this is due to various reasons. The process of predicting missing data is difficult and complex, because it requires time, especially in the case of a large number of missing data, and the temporal and spatial variation of the precipitation phenomenon and its complexity from the physical point of view intervene . in this work, we rely on a new and periodic method in this field, that of artificial intelligence and machine learning, using LR, DS, AR, GP, RF and other models to estimate the missing data of the rain gauge located in the northeast of Algeria in three different watersheds (Seybousse, Oued Ruhmel, central coastal basin of Constantine); Each basin has its own rain gauge. We studied monthly data over the period from 1970 to 2008 for the watersheds of Seybousse and the central coastal basin of Constantine and from 1986 to 2012 for the watersheds of Oued Ruhmel. We first confirmed the efficiency of the models by training them, and after they gave us good results and a small margin of error, we used these models to predict the missing rain data at our three stations. Finally, we concluded that the majority of these models were good and gave us excellent and accurate results. It allowed us to predict the missing data at the three stations.

Keywords: rainfall; missing data; artificial intelligence; machine learning; the Northeast of Algeria

Résumé

Le processus d'enregistrement des précipitations se heurte à un problème majeur dans notre pays, l'Algérie, qui est la perte de données, et cela est dû à diverses raisons. Le processus de prédiction des données manquantes est difficile et complexe, car il nécessite du temps, surtout dans le cas d'un grand nombre de données manquantes, et la variation temporelle et spatiale du phénomène des précipitations et sa complexité du point de vue physique interviennent. dans ce travail, nous sommes appuyés sur une méthode nouvelle et périodique dans ce domaine, celle de l'intelligence artificielle et l'apprentissage automatique, en utilisant les modèles LR, DS, AR, GP, RF et autres pour estimer les données manquantes des stations pluviométrique qui se trouve dans le nord-est de l'Algérie dans trois bassins versants différents (Seybousse, Oued Ruhmel, bassin Côtier Constentinois Centre) ; Chaque bassin a ses propres stations. Nous avons étudié des données mensuelles sur la période de 1970 à 2008 pour les bassins versants de Seybousse et du bassin Côtier Constentinois Centre et de 1986 à 2012 pour les bassins versants d'Oued Ruhmel. Nous avons d'abord confirmé l'efficacité des modèles en les entraînant, et après qu'ils nous aient donné de bons résultats et une petite marge d'erreur, nous avons utilisé ces modèles pour prédire les données de pluie manquantes à nos trois stations. Enfin, nous avons conclu que la majorité de ces modèles étaient bons et nous ont donné des résultats excellents et précis. Il nous a permis de prédire les données manquantes aux trois stations.

Mots-clés : précipitations ; données manquantes; intelligence artificielle ; apprentissage automatique ; le nord-est de l'Algérie

ملخص

تواجه عملية تسجيل هطول الأمطار مشكلة كبيرة في بلادنا الجزائر وهي ضياع البيانات وهذا يعود لأسباب مختلفة. عملية التبؤ بالبيانات المفقودة عملية صعبة ومعقدة ، لأنها تحتاج إلى وقت ، خاصة في حالة الكثير من البيانات المفقودة ، ويقف أمامها الاختلاف الزماني والمكاني لظاهرة التساقط وتعقيدها من الناحية الفيزيائية. لذلك اعتمدنا في هذا العمل على طريقة جديدة ودورية في هذا المجال وهي الذكاء الاصطناعي و التعلم العميق حيث اعتمدنا على نماذج LR و DS و AR و GP و RF وغيرها من النماذج لمحاكاة البيانات المفقودة، أجريت دراستنا في شمال شرق الجزائر في ثلاثة مستجمعات مائية مختلفة ، (سيبوس وادي الرمال والحوض القسنطيني الساحلي) ؛ كل حوض له محطاته الخاصة . لقد أجرينا محاكاة على البيانات الشهرية في الفترة الزمنية من 1970 إلى 2008 لكل من مستجمعات المياه سيبوس والحوض القسنطيني الساحلي ومن 1986 إلى 2012 لمستجمعات مياه وادي الرمال. حيث تأكدنا أولاً من فعالية النماذج من خلال تدريبها، وبعد أن أعطتنا نتائج جيدة وهامش خطأ صغير، استخدمنا هذه النماذج لتنبؤ ببيانات المطر المفقودة في محطاتنا الثلاث. وأخيراً توصلنا إلى أن غالبية هذه النماذج كانت جيدة وأعطتنا نتائج ممتازة ودقيقة. مكنتنا من التنبؤ بالبيانات المفقودة في المحطات الثلاثة.

الكلمات المفتاحية: هطول الأمطار؛ بيانات مفقودة؛ الذكاء الاصطناعي ؛ تعلم الآلة ؛ شمال شرق الجزائر

General Introduction

General Introduction

The availability of rainfall data offers considerable advantages in terms of weather forecasting, water resource management, natural risk prevention, scientific research and informed decision-making. These data are essential for understanding and adapting to climate variations, ensuring the safety of populations and promoting sustainable development.

Rainfall data can be lacking for various reasons. Among the main causes of incompleteness are technical problems such as equipment failures or measurement errors, as well as failures of meteorological stations. Additionally, the limited number of meteorological stations, especially in rural areas, may result in insufficient spatial coverage. Data loss can also occur due to problems with data storage, transmission or processing, while access restrictions and budgetary constraints can limit the availability of rainfall data. Finally, extreme weather events and harsh climatic conditions can damage data collection equipment and lead to interruption of collection or loss of data.

To estimate the missing rainfall data, several classical methods are used. One such method is spatial interpolation, which involves estimating missing values based on data from nearby measurement stations. Different interpolation techniques, such as distance-weighted inverse interpolation, Kriging interpolation or spline interpolation, can be used. Another common approach relies on the use of statistical analyses, where statistical models are used to characterize relationships between meteorological variables and estimate missing data based on predictors.

The use of artificial intelligence (AI) techniques has become increasingly popular for estimating missing rain data. AI algorithms can analyze existing rainfall data patterns and relationships with other meteorological variables to make predictions and fill in the gaps.

It's worth noting that the accuracy of AI-based rainfall estimation depends on the quality and quantity of available data, the chosen algorithm, and the specific characteristics of the region being studied. Additionally, incorporating domain expertise and validating the AI models against ground truth observations is crucial to ensure reliable results.

This study aims to complete the missing data on the maximum monthly precipitation in the northeast of Algeria using several large families of LR, RP, AR and REPT algorithms our work is composed of four chapters namely:

Chapter I Bibliography:

Chapter II: Methods for estimating missing data;

Chapter III: Collected and critique of data;

Chapter IV: Results and discussions;

Chapter I
Bibliography

Introduction

The study of the climatic variations of the terrestrial globe is one of the main objectives of climatology, because the impact of these changes on the biophysical and societal environment does not stop growing. Climatology therefore studies the families of meteorological conditions likely to affect different regions over long periods of time. This science calls on the various disciplines of nature likes geography, geology, physics and chemistry. Under the name of climate, we distinguish two different notions. The most classic that learned in geography lessons results from a spatial approach: the Earth is divided into climatic zones according to the weather conditions prevailing there in the different seasons. The other meaning of the word climate corresponds to a global temporal approach: we are interested in the modifications of meteorological conditions integrated over the whole of the globe and over the long term (30 years**).**

I. General information on climate

1. Definition

The climate is defined as the set of series of states of the atmosphere above a place in their usual succession. The study of short-term weather in specific areas is the domain of meteorology. The climate is determined using averages established from annual and monthly statistical measurements of local atmospheric data: temperature, precipitation, sunshine, and humidity, wind speed, etc. **[1].**

2. The elements of the climate

2.1 The atmosphere

The atmosphere is the gaseous layer that envelops the terrestrial globe **(Fig.I.1).** The air in which Man usually lives **[2]**, the Earth's atmosphere is essentially composed of nitrogen (78%) and oxygen (21%), as well as many inert gases: methane, hydrogen and ozone. In addition to these gases, the atmosphere contains varying proportions of water vapor and airborne aerosols (tiny particles solid or liquid of variable origin) **[3].**The atmosphere is the seat of a large number of phenomena such as radiation, electricity, vertical and horizontal movements of the wind **[2].**

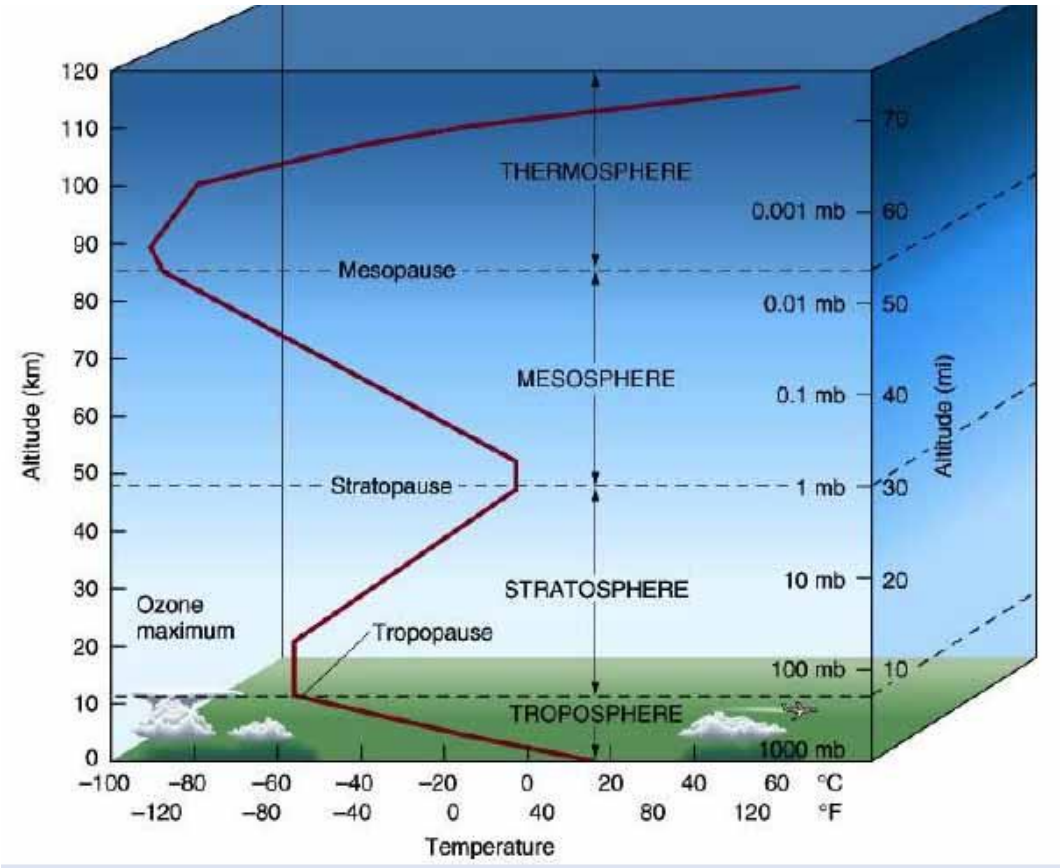


Figure I.1: Stratification of the different layers of the atmosphere **[4]**

2.2 Solar-radiation

It is our planet's only external energy source. The sun, due to its temperature of approximately 6000°K, radiates mainly in the visible and near infrared (between 300 nm and 1200 nm) with a maximum around 500 nm. It is the solar flux reaching the limit of the atmosphere, depending on the effects of transmission and diffusion of the atmosphere **[5].**

The average distance from the earth to the sun is close to 150 million km. solar radiation takes about 8 minutes to reach us **[6].** This flow of external origin is filtered by the atmosphere which absorbs, reflects and diffuses part of it **[7].** The earth receives different quantities of energy at the equator and at the poles.

Overall, half of the solar energy reaching the earth is absorbed by the continents and oceans which it heats. Part of this energy is returned, essentially in the form of infrared radiation. Certain gases present in small quantities in the atmosphere (water vapor, CO2, methane) absorb infrared radiation: only 10% of the radiation emitted by the surface escapes directly into space **[4].**

2.3 The Cloud

The visual appearance of clouds varies greatly depending on the altitude at which they form, the season, the location and general atmospheric conditions. The colors they display are only due to the lighting conditions (white when their surface is subjected to direct solar radiation, grey/black when their thickness is such that it attenuates the intensity of solar radiation. **[8]**

Clouds are not only made up of vapor but also of tiny water droplets and ice crystals (0.02 mm in diameter) which make the humidity in the air visible **[9].** The height development of a cloud corresponds to the condensation of the excess vapor released by the pseudo-adiabatic cooling of saturated air **[10].**

Fog is formed by condensation of water vapor in the air. The most common, radiation fog, is due to night cooling of the ground **[9].** Fog forms when the temperature remains positive. On the other hand, if the temperature becomes negative, condensation occurs in the form of hoar frost or frost **[8].**

2.4 Air humidity

The degree of humidity in the air depends on the amount of water in it. The oceans are the main source of water vapor in the air, covering three quarters of the planet's surface. Other sources are rivers, lakes and streams, soil and vegetation **[11]**.There is two types of humidity:

- Specific humidity: measures the exact weight of water vapor contained in an air mass;
- Relative humidity: expresses the ratio between the quantity of water vapor contained in an air mass and that which is necessary to saturate it **[9].**

3. Climatic factors

These are the meteorological elements that make up the climate and make it possible to characterize it, these are:

- **Temperature** is considered as a physical quantity linked to the immediate notion of hot and cold. Temperature is the manifestation, on a macroscopic scale, of the movement of atoms and molecules.
- **Precipitation** it is the set of different forms and states under which atmospheric water:

  Solid (snow), liquid (rain) and gaseous (fog, thrashing), moves on the surface of the globe. The rains originate from the vaporization of terrestrial waters.

- **The Wind** is the physical parameter representative of air movements. It arises mainly

  From the difference in atmospheric pressure between latitudes and the rotation of the earth.

- **Humidity** is the amount of water vapor in the air, not including liquid water and ice. We must distinguish between relative humidity and absolute humidity: Relative humidity is not really a measure of the amount of water vapor in the air but rather a ratio between the amount of water vapor in the area and its capacity.

II. The precipitation

It is the most important element in our study and research, precipitation is one of the most variable hydrological processes of the climate **[12].** It is characterized by three main parameters: its volume, its intensity and its frequency which vary according to the places, the days, the months and also the years **[13].** To identify and classify the various rainfall regions of the globe, we usually use the average monthly or annual precipitation (evaluated over a long period) and its variations **[12].** Air can only hold a limited amount of water vapor, which decreases with temperature. **[14].** Precipitation occurs when water vapor present in the atmosphere condenses into clouds and falls to earth. They are the only "entrance" to the main continental hydrological systems, which are the watersheds.

They are a physical phenomenon that describes the transfer of water in its liquid (rain) or solid (snow, hail) phase between the atmosphere and the ground. They are caused by a change in temperature or pressure **[15].**

1. Origin of precipitation

The water cycle, also known as the hydrologic cycle or the hydrological cycle, is a biogeochemical cycle that describes the continuous movement of water on, above and below the surface of the Earth. The mass of water on Earth remains fairly constant over time but the partitioning of the water into the major reservoirs of ice, fresh water, saline water (salt water) and atmospheric water is variable depending on a wide range of climatic variables. The water moves from one reservoir to another, such as from river to ocean, or from the ocean to the atmosphere, by the physical processes.

of evaporation, transpiration, condensation, precipitation, infiltration, surface runoff, and subsurface flow. In doing so, the water goes through different forms: liquid, solid (ice) and vapor. The ocean plays a key role in the water cycle as it is the source of 86% of global evaporation**. [16].**

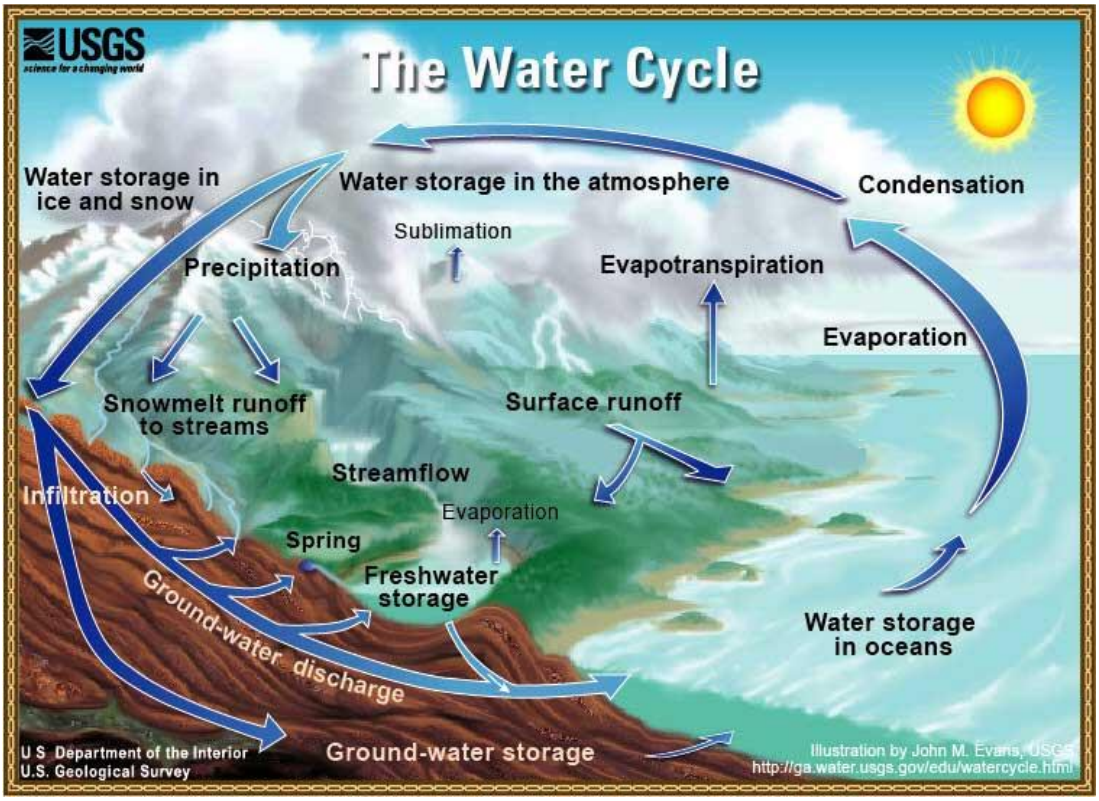


Figure I.2: (Water Cycle | Science Mission Directorate, s. d.) **[16].**

2. Different form of Precipitation

2.1 Rain

Rain is any liquid that drops from the clouds in the sky. Rain is described as water droplets of 0.5 mm or larger. Droplets less than half a millimeter are defined as a drizzle. Raindrops frequently fall when small cloud particles strike and bind together, creating bigger drops. As this process continues, the drops get bigger and bigger to an extent where they become too heavy to suspend on the air. As a result, the gravity pulls then down to the earth.

When high in the air, the raindrops start falling as ice crystals or snow but melt when as they proceed down the earth through the warmer air. Rainfall rates vary from time to time, for example, light rain ranges from rates of 0.01 to 0.1 inches per hour, moderate rain from 0.1 to .3 inches per hour, and heavy rain above 0.3 inches per hour. Rain is the most common component of the water cycle and replenishes most of the freshwater on the earth**. [17].**

2.2 Snow

Snow occurs almost every time there is rain. However, snow often melts before it reaches the earth's surface. It is precipitation in the form of flakes of ice water falling from the clouds. Snow is normally seen together with high, thin, and weak cirrus clouds. Snow can at times fall when the atmospheric temperatures are above freezing, but it mostly occur in sub-freezing air. When the temperatures are above freezing, the snowflakes can partially melt but because of relatively

Warm temperatures, the evaporation of the particles occurs almost immediately. **[17].**

2.3 Sleet (Ice Pellets)

Sleet takes place in freezing atmospheric conditions. Sleet, also known as ice pellets, form when snow falls into a warm layer then melts into the rain and then the rain droplets fall into a freezing layer of air that is cold enough to refreeze the raindrops into ice pellets. Hence, sleet is defined as a form of precipitation composed of small and semi-transparent balls of ice. They should not be confused with hailstones as they are smaller in size. **[17].**

2.4 Hail

Hailstones are big balls and irregular lumps of ice that fall from large thunderstorms. Hail is purely solid precipitation. As opposed to sleets that can form in any weather when there are thunderstorms, hailstones are predominately experienced in the winter or cold weather. Hailstones are mostly made up of water ice and measure between 0.2 inches (5 millimeters) and 6 inches (15 centimeters) in diameter. This ranges in size of a pea's diameter to that larger than a grapefruit. **[17].**

2.5 Graupel

The word graupel is Germanic in origin; it is the diminutive of Graupe, meaning "pearl barley." According to etymologists, there does seem to be a grain of truth in the assumption that the word grew from the Slavic word krupa, which has the same meaning. Graupel was first seen in an 1889 weather report and has been whirling around in the meteorology field ever since to describe "pellets of snow" or "soft hail" (the latter phrase is an actual synonym of graupel),Are soft, small pellets formed when super cooled water droplets (at a temperature below 32°F) freeze on to a snow crystal, a process called riming , If the riming is particularly intense, the rimed snow crystal can grow to an appreciable size, but remain less than 0.2 inches. is also called snow pellets or soft hail, as the particles are particularly fragile and generally disintegrate when handled. **[18]**

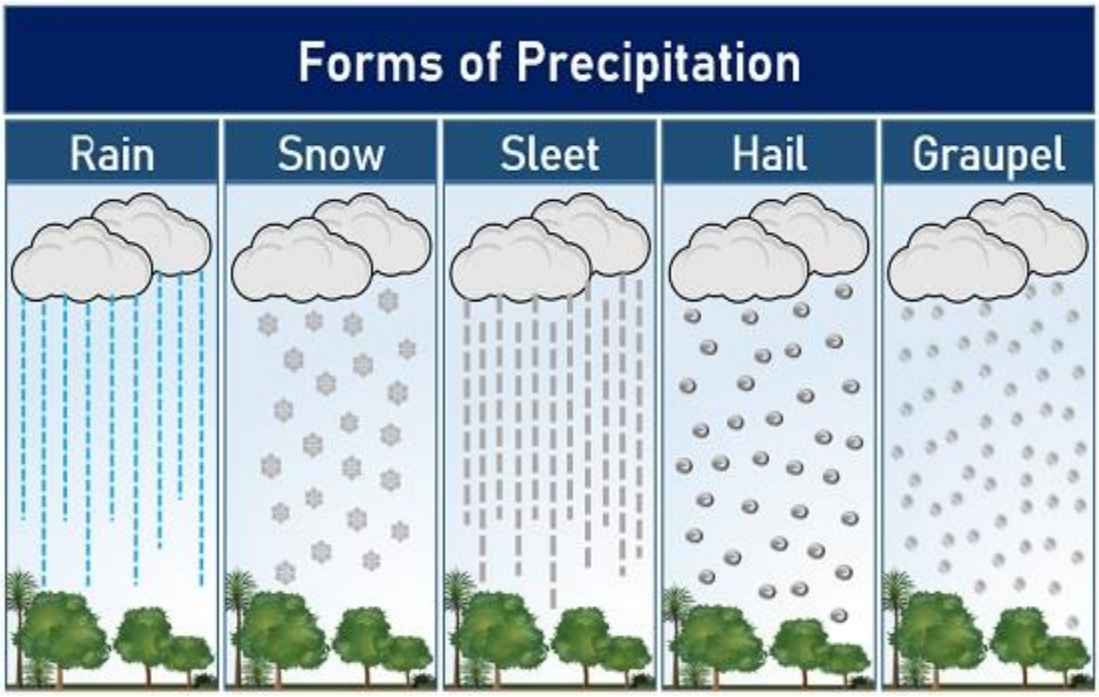


Figure I.3: form of precipitation **[19].**

3. Types of precipitation

The study of precipitation shows that the simplistic diagram according to which the clouds would be born above the oceans, and then pushed by the winds would fall as rain on the continents is false.

Indeed, we can assume that a cloud originating above the Atlantic pours about 100 mm over Europe and Western Russia before dissipating beyond the Urals. However, a cloud only contains a maximum of about 20 mm of water. It is therefore only at most 20 mm that come from the oceans and 80 mm that come from the atmosphere above the continents.(…..)

3.1 Convective precipitation

The convective system appears when two masses of air of different temperatures come into contact and more particularly a mass of cold air covers a warmer ground or when the lower layers are heated by solar radiation, the air of these lower layers then expands, becomes lighter and rises under the effect of Archimedes' thrust, it rises while cooling down to the level of condensation, the altitude at which the base of the cloud is formed. The air continues to rise, thus condensing, to the level of thermal equilibrium, the altitude of the top of the cloud.

This can reach altitudes of the order of 12 km for the most convective situations. This system is described by 03 phases: development, maturity and dissipation as it is represented in figure 04 **[20]** and **[21].**

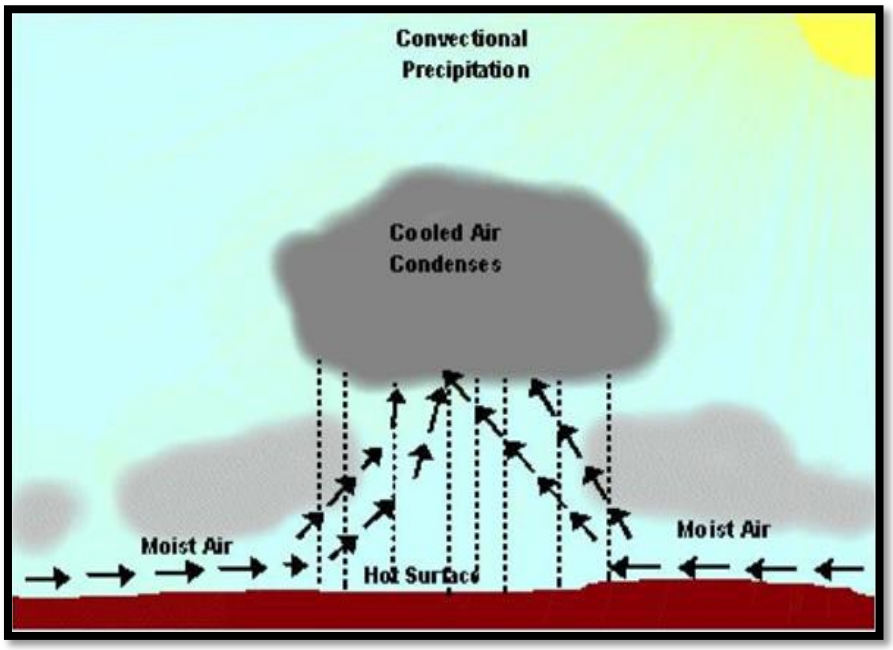


Figure I.4: Convective precipitation **[22]**

3.2 Orographic precipitation

The presence of a relief (a mountain range for example) on the trajectory of an air mass causes it to rise. The induced cooling can cause the formation of cloud cover and triggering precipitation. This type of system is related to cyclonic disturbances **[21],** Precipitation, of varying intensity and extent, mainly affects the slope facing the wind. The leeward side is on the contrary drier, because the relative humidity of the air mass and therefore the rain is less or even zero when it descends on the leeward side **[23].**

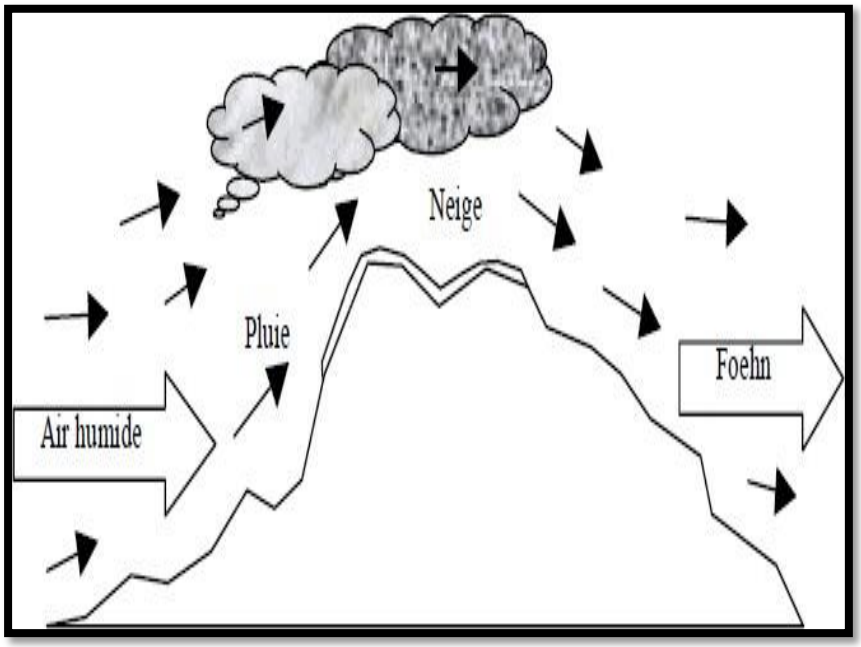


Figure I.5: Orographic precipitation **[24]**

3.3 Cyclonic precipitation

These precipitations are generated in the vicinity of the contact surfaces between two masses of air of different temperature and humidity, which is called a front. The hot air mass is always lifted aloft by the air mass. Cold Depending on whether the warm air mass follows or precedes

The cold air mass, there is a warm or cold front. In the case of a cold front, the clouds have a significant vertical development and the precipitation is intense. In the warm front case, the clouds have a significant horizontal extension and the precipitation is lower than for the cold front **[23].**

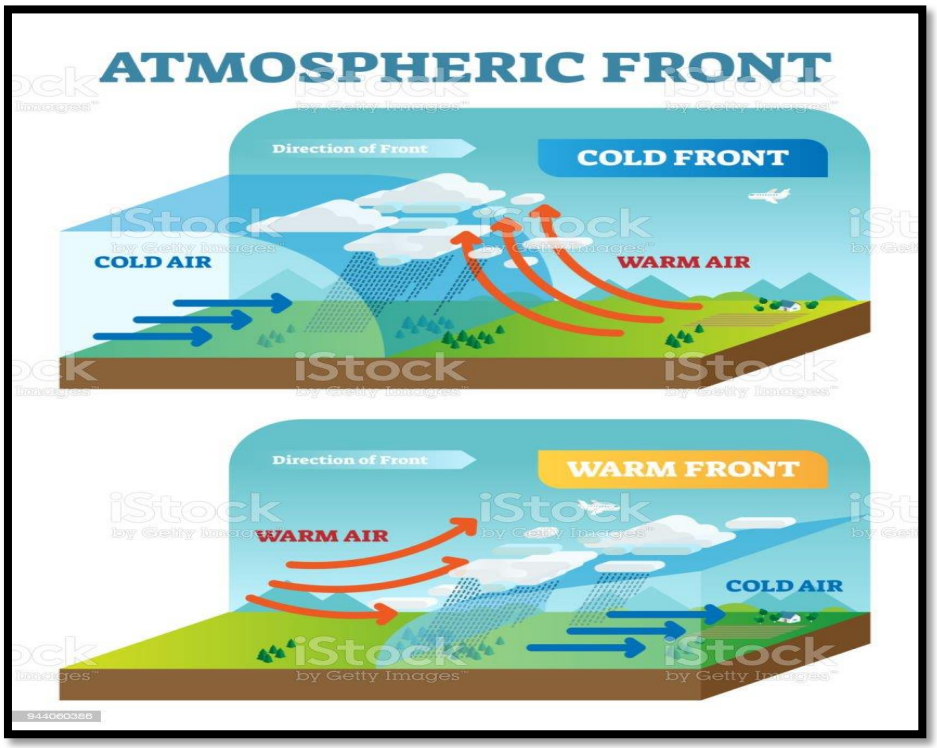


Figure I.6: Warm and cold front [24].

III. Measurement methods and principles of operation

There are two main categories of precipitation measuring devices: rain gauges and rain gauges. The rain gauges indicate the height of total precipitated water, in mm (1mm=1L/m2=10m3/ha), over periods generally equal to 24 hours. Rain gauges make it possible to determine the cumulative height of precipitation over time, and therefore to determine the intensity of the rain, in mm/h, over short time steps of the order of 1 to 6 minutes in general, the time steps being fixed or variable **[25].**

1 Indirect measurement method

1.1 The rain gauge

The rain gauges, whose models are very diverse, indicate the height of water precipitated during a given time interval, generally 24 hours (**Fig.I.7).**The water collected by the receiving surface is kept in a container graduated directly in millimeters of water. The reading of the volume makes it possible to know the height of precipitate drain. After reading, the container is emptied and replaced **[25].**

Figure I.7: rain gauge **[25]**

Rain gauges are intended to record the cumulative rainfall as a function of time, unlike rain gauges which record the Cumulative rainfall over a given period. There are three types of gauges: water level, tipping bucket and weighing **[25].**

1.2 Bucket rain gauge

The principle of this device is very simple (**Fig.I.8**), Rain water is collected in a receiving cone called an impluvium and flows through a calibrated nozzle to a bucket. When this bucket is filled, it tilts under the effect of the displacement of its Centre of gravity: the water flows outside the device and the opposite bucket fills in turn until the next tilting. On each fail over **[25].**

Figure I.8: tipping bucket rain gauge **[25].**

1.3 The rain gauge measuring the water level

This type uses either a float whose level varies with the height of water (**Fig.I.9**), or an acoustic direction of the distance from the level, or the conductivity of the water with several detectors positioned one above the other. This type of device includes a very precise and fragile emptying mechanism that requires careful maintenance. This is why these siphon gauges are only very rarely installed **[26].**

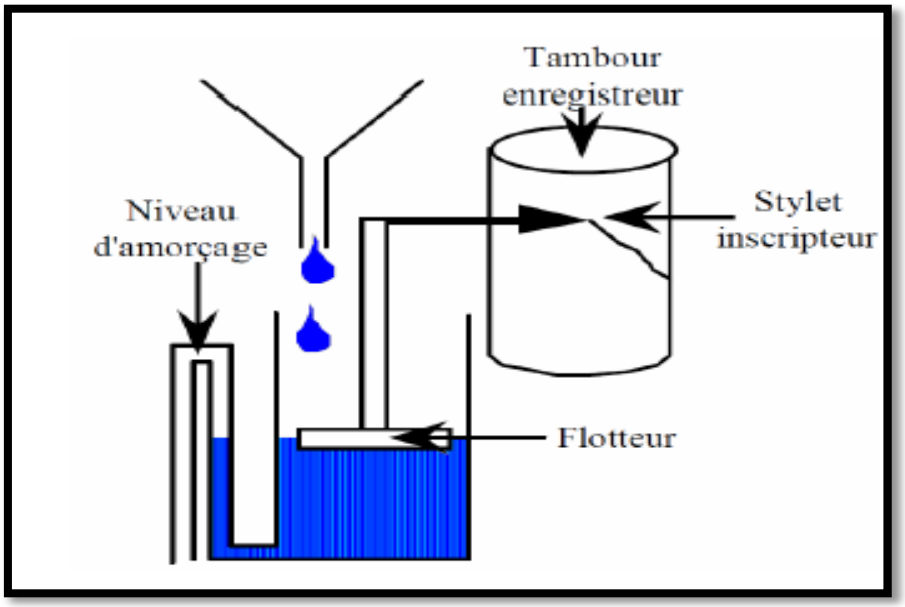


Figure I.9: Diagram of water level rain gauge with siphon **[26]**

1.4 Weighing rain gauge

In a conventional weighing rain gauge **(Fig.I.10),** the water collected in the receiving cone flows towards a single bucket which gradually fills and which empties by shifting its Centre of gravity as soon as a fixed mass of water is reached (150 to 200 g for current models).The water is evacuated outside the Pluviograph and the trough returns to its initial position. Between two tilts, the mass of the bucket and of the water it contains is measured continuously **[25].**

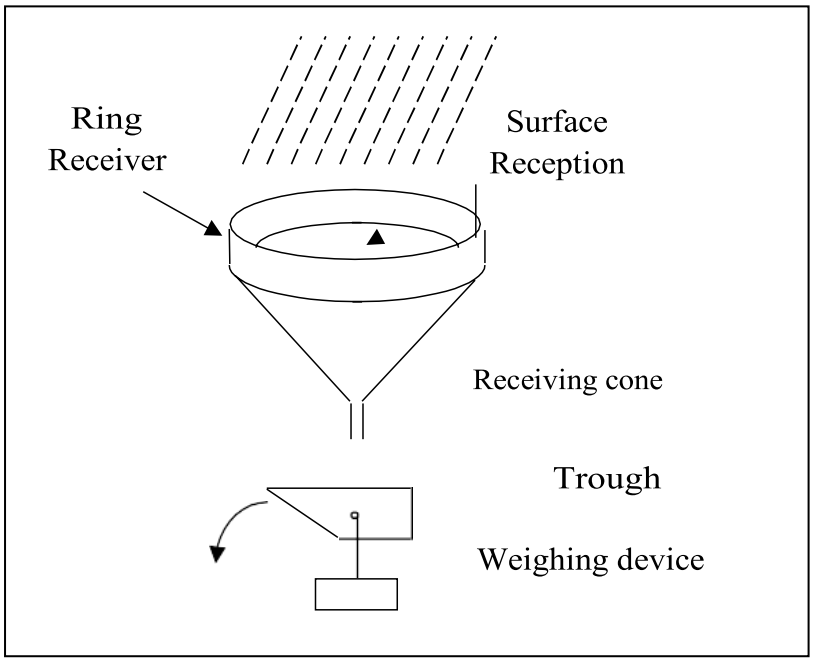


Figure I.10: Operating principle of a weighing rain gauge **[25].**

2. Indirect measuring instruments

2.1 Definition of radar

The term “radar” is the acronym of the English expression “Radio detection and ranging” (detection by radio waves and distance measurement). The radar is an indirect measurement instrument; it emits electromagnetic waves which propagate in the atmosphere at the speed of light. When these encounter an obstacle, they are partly reflected and the radar antenna picks up an echo in return **[27].**

2.2 Principle of measurement

Radar is an active sensor that emits electromagnetic pulses into the environment. Backscattered energy that is reflected from objects in its path is received by the radar. The typical radar system consists of at least four of the following components: a transmitter which sends out high frequency signals, an antenna which sends the output signal and receives the echoes, a receiver which processes the return signals so that they are ready to use, and a data display system **[28].**The radar part which is visible to the general public from the antenna covered with a protective dome, a Radom, and installed above an observation tower **(Fig.I.11)**

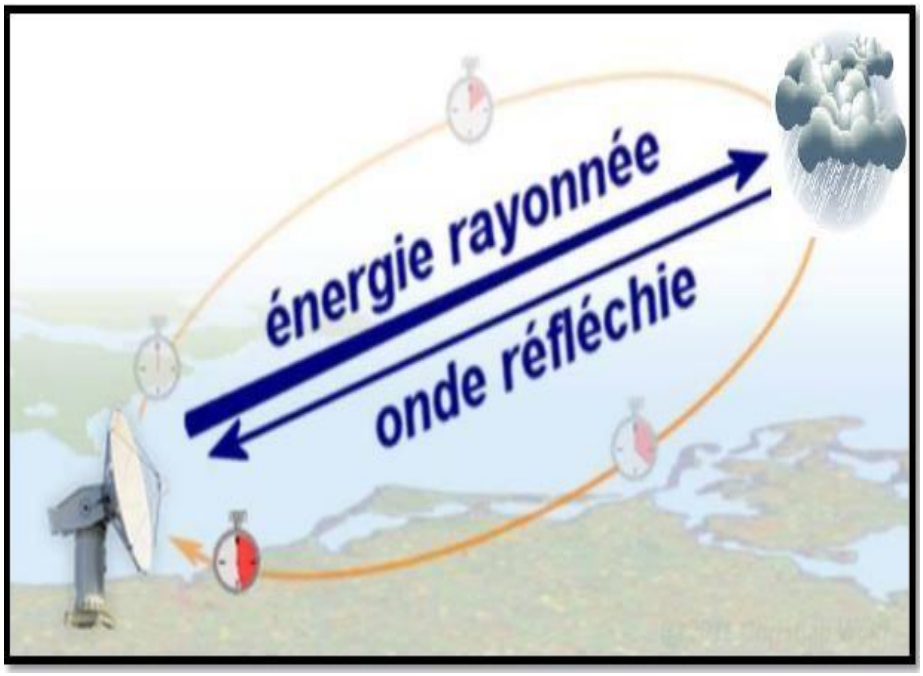


Figure I.11: Principle of radar emission **[25]**

3. Possible errors in precipitation measurements

Errors that can occur in precipitation measurements are:

3.1 Observation errors

Errors due to evaporation;

- ❖ Overflow of the rain gauge when the rains are very intense;
- ❖ Pierced rain gauge;
- ❖ Water losses during the transfer of the test tube into the pail;
- ❖ Rain gauge under a tree, etc. **[29].**

3.2 Systematic errors

Systematic errors include:

- The graduation of the test tube does not correspond to the opening of the rain gauge;
- A change in the operation of the rain gauge due to;
- A displacement of the rain gauge;
- Modification of the rain gauge environment;
- A change of observer;
- A broken test piece replaced by an unsuitable one **[29].**

3.3 Transcription and calculation errors

If it is a copy, you may encounter additional errors:

Figures that are not very legible may have been misinterpreted by the copier, the comma may have been omitted, the chronological order of the monthly sheets may have been incorrectly reproduced, etc. Also, errors may be encountered in the summation of readings **[29].**

4. Detection of errors and correction of anomalies

The visual control of rainfall data always proves to be effective and makes it possible to detect at first glance the gross heterogeneities that may exist and to correct them. Other less obvious heterogeneities may exist and do not appear during this check. For these, it is mandatory to use certain statistical methods to detect them **[30].**

IV. The use of the artificial intelligence models to estimate missing data:

Missing rain data can have implications for various applications, including hydrological modeling, climate research, and water resource management. To address this issue, efforts are made to minimize data gaps through regular maintenance and calibration of measurement equipment, training and supervision of data collectors, and implementing quality control procedures to identify and rectify missing or erroneous data.

In recent years, there has been an increased interest in using artificial intelligence (AI) models to estimate missing rain data. AI models, such as machine learning algorithms, can analyze the available data and learn patterns and relationships to make predictions and fill in the gaps

in the rainfall dataset. The advantage of using AI models for estimating missing rain data is their ability to capture complex patterns and nonlinear relationships in the data. They can adapt and improve their predictions as more data becomes available and can handle large datasets with numerous variables.

1. Artificial Intelligence

1.1 Definition of Artificial Intelligence

As a scientific endeavour, machine learning grew out of the quest for artificial intelligence (AI). In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptions and other models that were later found to be reinventions of the generalized linear models of statistics. **[31]** Probabilistic reasoning was also employed, especially in automated medical diagnosis. **[32].**

The prospect of creating intelligent computers has fascinated many people for as long as computers have been around and, as we shall see in the historic overview, the first hints in the direction of Artificial Intelligence date even before that. But what do we mean by Artificial Intelligence, if even the term intelligence itself is difficult to define? The precise definition and meaning of the word intelligence, and even more so of Artificial Intelligence, is the subject of much discussion and has caused a lot of confusion. One dictionary alone, for example gives four definitions of Artificial Intelligence:

- An area of study in the field of computer science. Artificial intelligence is concerned with the development of computers able to engage in human-like thought processes such as learning, reasoning, and self-correction
- The concept that machines can be improved to assume some capabilities normally thought to be like human intelligence such as learning, adapting, self-correction, etc.
- The extension of human intelligence through the use of computers, as in times past physical power was extended through the use of mechanical tools.
- In a restricted sense, the study of techniques to use computers more effectively by improved programming techniques. **[33].**

1.2 Applications of AI

A list of branches of AI is given below. However some branches are surely missing, because no one has identified them yet. Some of these may be regarded as concepts or topics rather than full branch **[33].**

-Logical

-Search

-Pattern Recognition

-Inference

-Common sense knowledge and Reasoning

AI has applications in all fields of human study, such as finance and economics, environmental engineering, chemistry, computer science, and so on. Some of the applications of AI are listed below: **[33].**

- Perception:
  - ❖ Machine vision
  - ❖ Speech understanding
  - ❖ Touch(tactile or haptic)sensation
- Robotics
- Natural Language Processing:
  - ❖ Natural Language Understanding
  - ❖ Speech Understanding
  - ❖ Language Generation
  - ❖ Machine Translation
- Planning
- Expert Systems
- Machine Learning
- Theorem Proving
- Symbolic Mathematics
- Game Playing

1.3 Problems of AI

Intelligence does not imply perfect understanding; every intelligent being has limited perception, memory and computation. Many points on the spectrum of intelligence versus cost are viable, from insects to humans. AI seeks to understand the computations required

from intelligent behavior and to produce computer systems that exhibit intelligence. Aspects of intelligence studied by AI include perception, communicational using human languages, reasoning, planning, learning and memory.

1.4 AI Technique

Artificial Intelligence research during the last three decades has concluded that Intelligence requires knowledge. To compensate overwhelming quality, knowledge possesses less desirable properties.

A. It is huge.
B. It is difficult to characterize correctly.
C. It is constantly varying.
D. It differs from data by being organized in a way that corresponds to its application.
E. It is complicated.

An AI technique is a method that exploits knowledge that is represented so that:

- The knowledge captures generalizations that share properties, are grouped to get her, rather than being allowed separate representation.
- It can be understood by people who must provide it—even though for many programs bulk of the data comes automatic ally from readings.
- In many AI domains, how the people understand the same people must supply the knowledge to a program.
- It can be easily modified to correct errors and reflect changes in real conditions.
- It can be widely used even if it's incomplete or inaccurate.
- It can be used to help overcome its own sheer bulk by helping to narrow the range of possibilities that must be usually considered **[33].**

2. Machine learning (ML)

2.1 Definition

Machine learning is an application that uses AI and allows system automation **[34] [35].**Machine learning focuses on computer data programs to access and analyze the data. Machine learning will enable computers to learn various things automatically without human assistance **[36].** Machine learning uses computer algorithms, which enables the computer to learn from examples and experiences. Machine learning algorithms occur in two categories,

namely supervised and unsupervised algorithms **[35].**Supervised machine learning algorithms apply to basics learned previously to new data by the use of examples that allow the machine to predict the future.

The machine undergoes the training of datasets, and the algorithm produces an inferred function to predict output values **[35],** after training, the system provides new input. The learning output can perform a comparison between the correct and the intended output to modify the model in case there are errors.

2.2 Approaches

Machine learning approaches are traditionally divided into three broad categories, which correspond to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system:

- **Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- **Unsupervised learning**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
- **Reinforcement learning:** A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize **[37].**

2.3 Models

Performing machine learning involves creating a model, which is trained on some training data and then can process additional data to make predictions. Various types of models have been used and researched for machine learning systems we mention some of them:

- Artificial neural networks
- Decision trees
- Support-vector machines
- Regression analysis
- Bayesian networks
- Gaussian processes

V Estimation of missing data and correction of precipitation

Rainfall is the main downward forcing of the hydrologic cycle and has a key role in many applications like the modeling and prediction of basin stream flow, reservoir inflow, simulation of climate change, and particularly, dealing with flash floods. Hence, the availability and quality of daily rainfall, particularly in regions dominated by monsoons (MS), rainfall records often suffer from missing data values because of technical defects in measuring equipment. The estimation of missing rainfall values is a difficult subject in hydrology partly because of the insufficient high-quality daily data in some stations of study sites, the presence of spatiotemporal variability's, and irregularities (nonlinearity) in rainfall characteristics, anomalous weather conditions, and climate change impact **[38].**

-In the following, we mention some of the studies conducted in this field**:**

1. Prediction of Rainfall in Australia Using Machine Learning

In this study, the locations that were studied were the regions of Victoria and Sydney, and the models used were, generally, Neural Networks and Random Forest. In line with these previous studies, in this work, a set of meteorological data from 49 different cities in Australia was taken. In the set, there is a variable that indicates whether or not it has rained on the day of taking the sample, and there are also other variables that show meteorological properties on the day of taking the sample, such as cloudiness, wind, sunlight, humidity, pressure, or temperature. In this case, it was observed that the efficiency of the algorithms was higher. Finally, the possible improvement of the results by modifying the data used to carry out the training was studied but without improvement compared with previous analyses. Likewise, it can be seen that algorithms based on Neural Networks work quite well to model nonlinear natural phenomena. Finally, the locality of the phenomenon can be observed, since, by considering the data independently by city, the algorithms work and are more efficient. Finally, another very interesting future study related to the one described would be to study the problem of predicting, several days in advance, which models are the most interesting or how many days in advance are optimal for making a prediction **[39].**

2. Multilayer Perceptron-based Predictive Model for the Reconstruction of Missing Rainfall-Data

In this study, used a new MLP structure was proposed to reconstruct missing rainfall data using the daily rainfall data of six rainfall stations located in Seoul by reconstructing missing

rainfall data in different seasons, I. In this study, station number two was selected as the station with missing rainfall data which were reconstructed with the proposed model using the dataset of the other five stations. By reconstructing missing daily rainfall values accurately, we can more accurately predict the occurrence of flash floods when compared to the use of monthly or annual data. This model can be extended to gap fill other meteorological data.

Selecting the appropriate long-term input data is essential to train the model to predict the missing rainfall data over different seasons. In addition, we plan to test other types of DL algorithms such as LSTM and convolution neural networks because further research should be undertaken to determine if similar results can be obtained from data-driven models, and with data from other catchment areas subjected to different climatic regimes than that of Seoul of Korea.**[40].**

3. Filling missing meteorological data with Computational Intelligence methods

In this research they have presented herein used the Multi-Layer Perceptron and Radial Basis Function networks and Support Vector Regression to predict the daily values of average, maximum and minimum air temperature as well as relative air humidity and saturation deficit in a 3.5year period of measurement. One of the aims of the research work was to evaluate the possibility to apply the a fore mentioned Computational Intelligence methods to supplement incomplete time series basing on data obtained from different stations, with use of various measurement equipment types. All the created models were characterized by good adjustment of the predicted values to observed ones, $r>0.9$. The highest correlation coefficients, and thus the smallest modeling errors were obtained for mean and maximum air temperature, while weaker correlations were noted for minimum temperature and relative air humidity. The differences in quality of models obtained for specific elements result mainly from two reasons. First of all, the station that provided input data, is located in the city Centre, within the Urban Heat Island, while output data were obtained from a station located outside it For the purposes of the present paper, the authors intentionally selected two stations located in completely different areas in order to highlight the differences between them and to verify the hypothesis that computational intelligence methods are applicable in such cases as well **[41].**

Conclusion

Climate is the state of the weather that extends over a long period of time that can reach months or an entire season, and the climatic characteristics of a place are taken by observing the area, and through that everything that happens there is recorded moment by moment, in this chapter we have tried to present the types of precipitation, their direct and indirect measuring instrument and the measurement errors, the latter we have tried to present some solutions by proposing the methods of artificial intelligence, and summaries of the few works that have been done to estimate missing values.

Chapter II
Methods for estimating missing data

Introduction

Precipitation estimation is a method of approximating the amount of precipitation that has fallen at a location or region. Maps of the estimated amount of precipitation that has fallen over a certain area and over a certain time period are compiled using several different data sources, including manual and automatic field observations and radar and satellite data. A number of Different algorithms can be used to estimate precipitation amounts from data collected by rain gauges, weather radar, and satellites or other remote sensing platforms.

I. Classic methods for estimating missing data

1. Simple methods

If complete data from neighboring stations are available, then the following methods can be used **[42].**

- Replace the missing value with that of the nearest station. This method is generally used to supplement the annual rainfall;
- Replace the missing value with a simple arithmetic mean of neighboring stations. This method is used when the average annual precipitation of station X (whose information we want to complete) is equal to the annual averages of the neighboring stations within 10%;
- If the difference between the annual average precipitation of station X and that of neighboring stations is greater than 10%, then the missing precipitation of X can be estimated by the averages weighted by the annual trends of neighboring stations, given by the following formula :

$$\mathbf{PX = 1n \sum(\overline{P}\, X \overline{P}\ in\ i = 1\ Pi)} \qquad \mathbf{(II.1)}$$

Where: PX: Missing value;

n : Number of reference stations;

P i: Precipitation at station i, corresponding to PX;

$\overline{P}$ i : Mean annual precipitation at station i;

$\overline{P}X$: Mean annual precipitation at station X

2. Method based on correlation

Regression and correlation consist of the study of the links between two or more variables. In hydrology, they are the oldest and most widely used mathematical tool, with multiple purposes:

- Extension in time of the series of hydrological observations of short duration;
- Forecast of hydrological quantities (flow based on observed hydro meteorological conditions: rainfall, temperatures, etc.);
- Geographical extension to unobserved basins of hydrological characteristics determined on various basins of analogous regime;
- Study of the dependence between the successive values of a series of hydrological data (time series).

Certain hydrological quantities may not be independent and yet not be linked by a functional relationship: it is said that there is a stochastic dependence between them (i.e. processes subject to chance) and make the object of a statistical study.

A strictly functional dependence corresponds to a theoretical conception is never verified in hydrology.

We say that there is a correlation between two observed variables, when the variations of the two variables occur in the same direction (positive correlation), or when the variations are in the opposite direction (negative correlation) **[29] [30].**

2.1 Definitions

- **Regression**

  Linear regression is a modeling method that makes it possible to establish a linear relationship between a continuous variable called "explained variable" or dependent and a set of other continuous variables called "explanatory variables" or independent. More specifically, it proposes an explanatory model which makes it possible to predict the dependent variable according to the independent variables **[43].**

- **Correlation**: Correlation: it is a method of research of the connection which exists between two random variables.

  The correlation between any random variables can be calculated. Very high but meaningless correlations are very common, so one only undertakes a correlation when the dependence between the variables can be explained **[29] [30].**

2.2 Choice of regression model

When the scatter diagram is linear or approximately linear, one can try to find the equation of the line that fits it best. This regression line from Y to X is generally determined by the method of least squares.

In practice, one always tries to find a linear regression even if it is necessary to make a transformation in the functional relation.

The different existing models are **[30]**:

The linear model represented by the equation of the line:

$$Y=A+BX \quad \textbf{(II.2)}$$

Curvilinear models, namely:

The power model: $Y=AXB$ **(II.3)**

The exponential model: $Y=AE\ BX$ **(II.4)**

The parabolic model: $Y=A+BX+CX2$ **(II.5)**

2.3 Conducting calculations for the extension of the series of annual rainfall totals

Let x and y be two variables, x observed n times and y observed k times with n > k. Let k be the number of pairs (x, y). We propose, from these k pairs to establish the line of regress y into x then, from the values of x, reconstruct the (NK) unobserved values of y.

Let: $\bar{X}K$; $\bar{Y}K$; ΣKX; ΣKY

the means and the standard deviations determined from the k pairs as well as the corresponding correlation coefficient rk **[29].**

The regression of y in x is written:

$$\hat{y}_j = rk\frac{k\sigma y}{k\sigma x}(xj - \bar{x}k) + \bar{y} \quad \textbf{(II.6)}$$

$k < j \leq n$

Thus will be reconstituted the (NK) values of y which are missing.

The estimate of the mean y of the extended sample $\hat{y}\bar{n}$ can be obtained directly from $\bar{x}n$ as follows:

$$\hat{y}_j = rk\frac{\sigma\frac{k}{y}}{\sigma\frac{k}{x}}(\bar{x}n - \bar{y}k) + \bar{y}k \quad \textbf{(II.7)}$$

2.4 Means of assessing the gain obtained by the extension

The benefit of extending the Y-series using the X-series is greater the higher the correlation coefficient. This benefit was translated by R, into relative efficiency E, which is expressed according to the following equation:

$$E = 1 + (1 - \frac{k}{n})(\frac{1-(k-2)r2}{k-3}) \quad \textbf{(II.8)}$$

Or:

R: This is the correlation coefficient calculated over k years;

E: Relative efficiency of which varies from k/n ton.

This benefit is expressed, using E in the form of a real gain of information that is expressed using the number of "effective" or "fictitious" years ń, to which corresponds the sample extended there.

Ń varies from k (no gain, because zero correlation between y and x with r = 0) to n (maximum gain, functional link between x and y and r = 1).

$$ń = \frac{k}{E} \qquad \textbf{(II.9)}$$

It is assumed that the extended y-series corresponds in weight of information to what a y-series actually observed over ń years would give. **[29].**

3. Principal component analysis method

Principal Component Analysis (PCA) is one of the most widely used multivariate data analysis methods. It consists of transforming interrelated variables (known as “correlated” in statistics) into new variables uncorrelated from each other. These new variables are called "principal components", or principal axes.

It allows the practitioner to reduce the number of variables and to make the information less redundant. It also makes it possible to explore multidimensional datasets made up of quantitative variables. It is widely used in marketing biostatistics, social sciences and many other fields **[44].**

Principal Component Analysis can be considered as a projection method that allows observations to be projected from the p-dimensional space of the p variables to a k-dimensional space (k < p) such that a maximum of information is conserved (the information is measured here through the total variance of the scatter plot) on the first dimensions.

If the information associated with the first 2 or 3 axes represents a sufficient percentage of the total variability of the point cloud, the observations can be represented on a 2 or 3-dimensional graph, thus greatly facilitating interpretation **[44].**

This method is applied to annual precipitation data, including us, by forming an initial rectangular matrix with the annual precipitation values of 5 stations in rows and the 43 years of observation in columns. The steps of principal component analysis are:

- ❖ Creation of the raw data matrix;
- ❖ Calculation of statistical parameters;

- ❖ Transformation of raw data into reduced centered data;
- ❖ Determination of the correlation matrix of the reduced centered data;
- ❖ Determination of the Eigen values from the correlation matrix of the reduced centered data;
- ❖ Determination of principal components (PC);
- ❖ Determination of reduced principal components (CP);
- ❖ Determination of the regression coefficient

4. Spatial interpolation methods for estimating missing data:

Spatial interpolation methods are commonly used to estimate missing rain data by inferring values at unobserved locations based on the available data from neighboring or nearby locations. These methods leverage the spatial correlation or relationship between rainfall measurements to interpolate and estimate the missing values. There are several methods but in the rest of the work we will cite only the most used methods in the hydrological field**:**

4.1 Spline

Spline (completely regularized) interpolation consists of the approximation of a function by means of series of polynomials over adjacent intervals with continuous derivatives at the end-point of the intervals. Smoothing spline interpolation enables to control the variance of the residuals over the data set. The solution is estimated by an iterative process. It is also referred to as the basic minimum curvature technique or thin plate interpolation as it possesses two main features: (a) the surface must pass exactly through the data points, and (b) the surface must have minimum curvature **[45].**

4.2 Interpolation methods by space partitioning

The methods of interpolation by partitioning of space are based on the use of a partitioning of the territory of study they determine the weight of the observations but especially their neighborhood. The type of terrain partitioning most used in climatology and hydrology is the division into distinct regions by polygons.

It was in 1911 that Thiessen proposed this method of spatial interpolation which now bears his name (Thiessen, 1911). This technique is based on the law of the nearest neighbor. Each observation from the rain gauges is assigned a polygon of influence constructed in such a way that each point in the polygon is closer to its observation site than to any other site.

The polygons are obtained by drawing the perpendicular bisectors of the segments connecting the observation sites. The study area is then partitioned into convex polygons, also referred to by the terms "Dirichlet cells" or "Voronoi diagram". (**Fig. II.1**) **[46].**

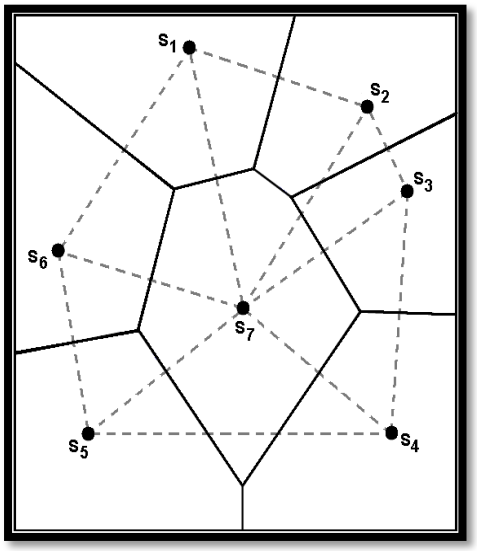


Figure II.1: Thiessen polygons (solid lines) accompanied by the associated Delaunay triangulation (dotted lines). **[46].**

4.3 Krigeage

Kriging is a geostatistical interpolation method that considers both the distance and the degree of variation between known data points when estimating values in unknown areas. geostatistic is a branch of statistics that explores natural phenomena that are both random and structural, kriging recognizes that the simple smooth mathematical function cannot be used in modeling the spatial variation of any continuous attribute value. The variation can be better described by a stochastic surface with an attribute known as a regionalized variable. The regionalized variable theory assumes that the value of a random variable Z at a location x. **[47].**

4.4. IDW method (Inverse distance weighting)

This method is often referred to in the literature by the acronym IDW (Inverse Distance Weighted). It is based on the principle that, for the same variable, the relative influence of an observation point decreases with the distance that separates it from the point in space whose value we want to estimate **.**We calculate the average of the measurements of the surrounding observation points (climate stations, in our case), with a greater weight given to the nearest points. The predicted value for a point in spaces:

$$z = [\sum_{i=1}^{N} \frac{Zi}{d_i^k}] \; / \; [\sum_{i=1}^{N} \frac{1}{d_i^k}] \qquad \textbf{(II.10)}$$

With:

Z = the estimated variable;

Z_i = the known value at measurement point i;

d = the distance between the point of unknown value and the measurement point i;

N= is the number of sites used for the interpolation;

k = the power to which the distance is raised.

In most cases k =2. However, it may be relevant to use other values of k, depending on the time step studied. It seems that for rain, in particular, the best value of k for annual accumulations is less than 1.5 **[48].**

II Machine learning methods for estimating missing data

1. Learning algorithms

On the Classify panel, when you select a learning algorithm using the Choose button the command-line version of the classifier appears in the line beside the button, including the parameters specified with minus signs. To change them, click that line to get an appropriate object editor

1.1Trees

Binary decision trees for datasets with a categorical or numeric class, dealing with missing values by treating them as a separate value and extending a third branch from the stump. Trees built by Random Tree consider a given number of random features at each node, performing no pruning. Random Forest constructs random forests by bagging ensembles of random trees, REP Tree builds a decision or regression tree using information gain/variance reduction and prunes it using reduced-error pruning. Optimized for speed, it only sorts values for numeric attributes once. It deals with missing values by splitting instances into pieces, as C4.5 does. You can set the minimum number of instances per leaf, maximum tree depth (useful when boosting trees), minimum proportion of training set variance for a split (numeric classes only), and number of folds for pruning.M5P is the model tree learner M50.

LMT builds logistic model trees. It can deal with binary and multiclass target variables, numeric and nominal attributes, and missing values. When fitting the logistic regression functions at a node using the Log it Boost algorithm, it uses cross-validation to determine how many iterations to run just once and employs the same number throughout the tree instead of

cross validating at every node. Machine learning is an application that uses AI and allows system automation **.**Machine learning focuses on computer data programs to access and analyze the data. **[49].**

1.2 Rules

Decision Table builds a decision table classifier. It evaluates feature subsets using best-first search and can use cross-validation for evaluation ,An option uses the nearest-neighbor method to determine the class for each instance that is not covered by a decision table entry, instead of the table's global majority, based on the same set of features.

One R is the 1R classifier with one parameter: the minimum bucket size for discretization. The Classifier model part shows that wage-increase-first-year has been identified as the basis of the rule produced; with a split at the value 2.9 dividing bad outcomes from good ones (the class is also good if the value of that attribute is missing). Beneath the rules the fraction of training instances correctly classified by the rules is given in parentheses **[49].**

1.3 Functions

Algorithms that fall into the functions category include an assorted group of classifiers that can be written down as mathematical equations in a reasonably natural way. Other methods, such as decision trees and rules, cannot (there are exceptions: Naive Bayes has a simple mathematical formulation). Four of them implement linear regression.

Simple Linear Regression learns a linear regression model based on a single attribute it chooses the one that yields the smallest squared error. Missing values and nonnumeric attributes are not allowed. The attribute that has the smallest squared error in this case is MMAX.

SMO implements the sequential minimal optimization algorithm for training a support vector classifier using kernel functions such as polynomial or Gaussian kernels. Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default—note that the coefficients in the output are based on the normalized data. Normalization can be turned off, or the input standardized to zero mean and unit variance. Pair wise classification is used for multiclass problems models can be fitted to the support vector machine output to obtain probability estimates. In the multiclass case the predicted probabilities will be combined using pair wise coupling **.**When working with sparse instances, turn normalization off for faster operation **[49].**

1.4 Lazy classifiers

Lazy learners store the training instances and do no real work until classification time. The simplest lazy learner is the k-nearest-neighbor classifier, which is implemented by IBK A variety of different search algorithms can be used to speed up the task of finding the nearest neighbors. A linear search is the default, but other options include KD-trees, ball trees and so-called "cover trees", the distance function used is a parameter of the search method. The default is the same as for IB1, that is, the Euclidean distance; other options include Chebyshev, Manhattan and Minkowski distances. The number of nearest neighbors (default k= 1) can be specified explicitly in the object editor or determined automatically using leave-one out cross-validation, subject to an upper limit given by the specified value. **[49].**

1.5 Miscellaneous classifiers

The "Misc." category includes just two classifiers. Serialized Classifier loads a model that has been serialized to a file and uses it for prediction. Providing a new training dataset has no effect, because it encapsulates a static model. Similarly, performing cross-validation using Serialized Classifier makes little sense. Input Mapped Classifier wraps a base classifier (or model that has been serialized to a file) constructs a mapping between the attributes present in the incoming test data and those that were seen when the model was trained.

Values for attributes present in the test data but not in the training data are simply ignored. Values for attributes present at training time but not present in the test data receive missing values. Similarly, missing values are used for incoming nominal values not seen during training **[49].**

1.6 Meta learning algorithms

Meta learning algorithms take classifiers and turn them into more powerful learners. One parameter specifies the base classifier(s); others specify the number of iterations for iterative schemes such as bagging and boosting and an initial seed for the random number generator. We already met Filtered Classifier it runs a classifier on data that has been passed through a filter, which is a parameter. The filter's own parameters are based exclusively on the training data, which is the appropriate way to apply a supervised filter to test data **[49].**

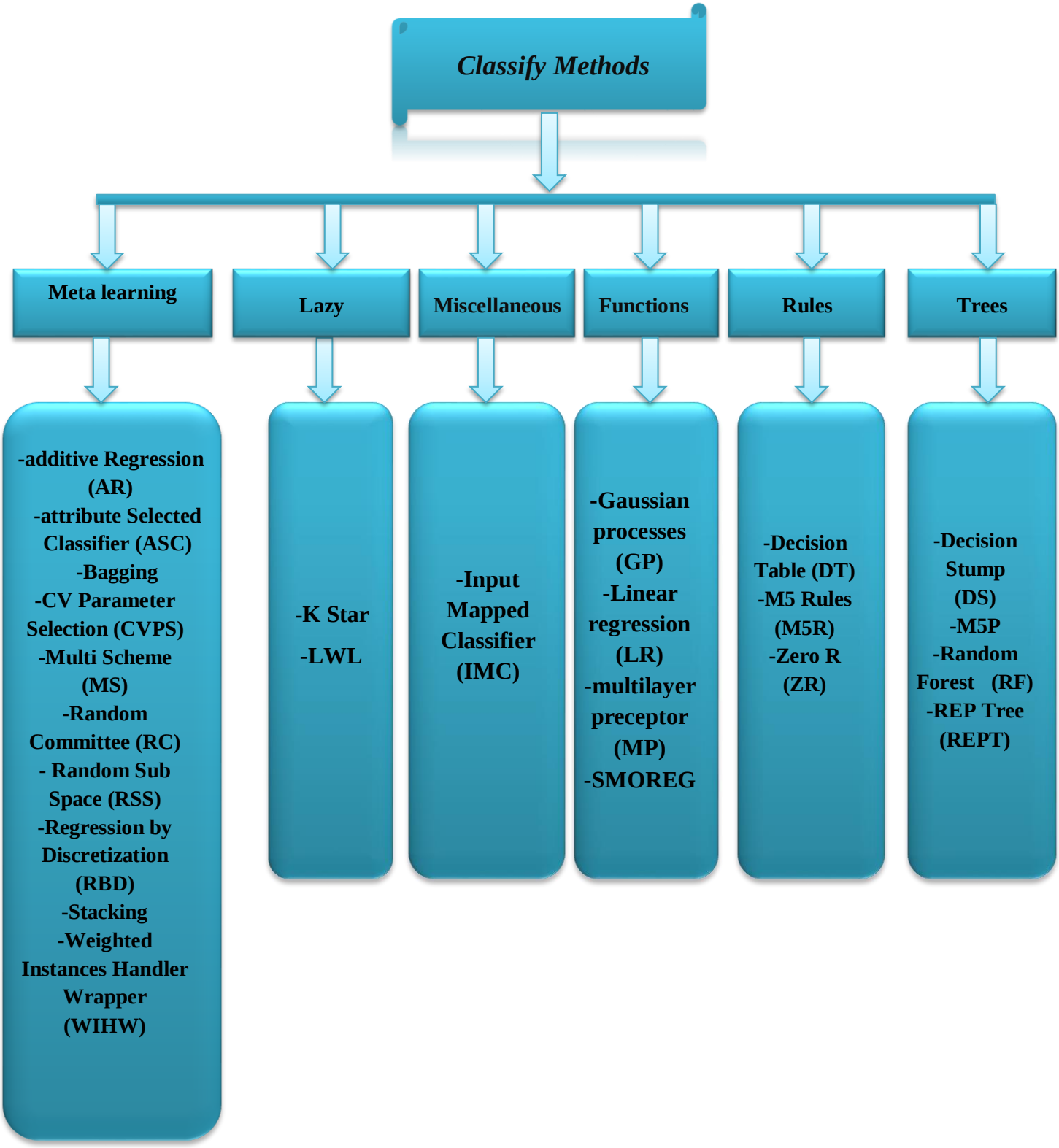


Figure II 02: Family of Classify Methods chart and its branches that we relied on in our work

III Ways to Evaluation of the effectiveness of estimation methods

We mention the methods that we will adopt in our research (the fourth chapter):

1. R-squared (R2)

Is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model? Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. **[50]**

2. Root Mean Squared Error (RMSE)

In statistical modeling and particularly regression analyses, a common way of measuring the quality of the fit of the model is the RMSE (also called Root Mean Square Deviation**) [51] .**

3. Mean Absolute Error (MAE)

In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of *Y* versus *X* include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement. versus an alternative technique of measurement. It has the same unit as the original data, and it can only be compared between models whose errors are measured in the same units. It is usually similar in magnitude to RMSE**. [50]**

4. Mean Relative Error (MRE)

The relative error is defined as the ratio of the absolute error of the measurement to the actual measurement. Using this method we can determine the magnitude of the absolute error in terms of the actual size of the measurement. If the true measurement of the object is not known, then the relative error can be found using the measured value. The relative error gives an indication of how good measurement is relative to the size of the object being measured. [**51]**

Table II.01: Evaluation Methods Equations

Numerical criteria	The equations	The meaning of the symbols
R-squared (R2) :	$R^2 = \frac{[\sum_{i=1}^{n}((\hat{P}_i-\bar{\hat{p}})(p_i-\bar{p}))]^2}{\sum_{i=1}^{n}(\hat{P}_i-\bar{\hat{p}})^2 \sum_{i=1}^{n}(p_i-\bar{p})^2}$	$\hat{P}i$: Estimated value; Pi : Observed value; n : Total number of observations
Root Mean Squared Error (RMSE):	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{P}_i - P_i)^2}$	
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\left\|(\hat{P}_i - P_i)\right\|$	
Mean Relative Error (MRE)	$MRE = \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{(\hat{P}_i - P_i)}{P_i}\right\|$	

Conclusion:

In this chapter, we presented the different methods used to estimate missing rainfall data, and these methods ranged from classical and spatial interpolation methods to machine learning-based methods, where artificial intelligence led to massive transformation of man in terms of technology. The models that work in artificial intelligence include machine learning are well presented finally, we have provided evaluation methods that allow us to see the effectiveness of these methods.

Chapter III
Collected and critique of data

Introduction

Precipitation is of interest to a number of different scientific communities such as the atmospheric and environmental communities and its monitoring and proper measurements has great economic and scientific value. For climate studies, accuracy of measurements and the homogeneity of the data records are crucial for properly and accurately assessing the change in environmental factors brought by climate change **[52].** However, accurately measuring precipitation is extremely difficult because of its highly variable properties and so consequently is one of the most poorly monitored environmental parameter specially on a global scale **[52].**The total amount of precipitation is the sum of the collected liquid **[53].** The concept of the gauge is simple, but its importance and practical use has led to a large number of different types of rain gauges **[52].** There are currently more than fifty different types of rain gauges in use around the world with different designs and each with its own associated mechanical errors **[53].** As a result, this has led to much variation among precipitation records over time across the globe.

I. Data criticism

Any climate study is based on the collection of rainfall data from sources different. for our study The rain data were collected from the National Resources Agency Hydraulics (ANRH, Algiers), in the form of tables of monthly and annual accumulations, they were very heterogeneous in some stations from the point of view of the reliability of measurements and the observation period, due to the absence of measurements over several years, The disparity of these resources often poses a problem of data quality.

The types of problems most often encountered when processing data in hydrology are**[54]:**

In general, the climatic elements in time do not occur in the same way and the corresponding series is not purely stationary. The causes of the homogeneity of the observations are:

- The anomaly or the defective state of the measuring devices.
- The permutation of the observer.
- Changing and or moving the station.
- The installation conditions (height above the ground).

The series of data from the Algerian stations are most often heterogeneous and has several gaps in the data series. This obliged us to make a study of the homogeneity of these last, and also the detection of the maximum and minimum singular values on the data of the monthly rains of the rain gauge stations of the three Algerian watersheds, such as seybouse, Oued Ruhmel and coastal constantinois center.

1. Seybouse watershed

The Seybouse watershed is located in the North-East of Algeria. It covers an area of 6452Km2. The basin is limited to the north by the coastal constantinois center and East watershed, to the south by the basin of the Constantinois highlands, to the west by the Kebir-Rhumel basin, and to the east by the Medjerda basin. The Seybouse basin covers seven wilayas "Oum El Bouaghi - Skikda-Annaba -El Tarf-Constantine-Guelma-Souk Ahras".

The Seybouse basin brings together six sub-watersheds with a very branched and dense hydrographic network; it includes three large wadis "CherfBouhamdane -Seybouse". It is a northern basin, with a relatively low flow **(fig.III.1)**

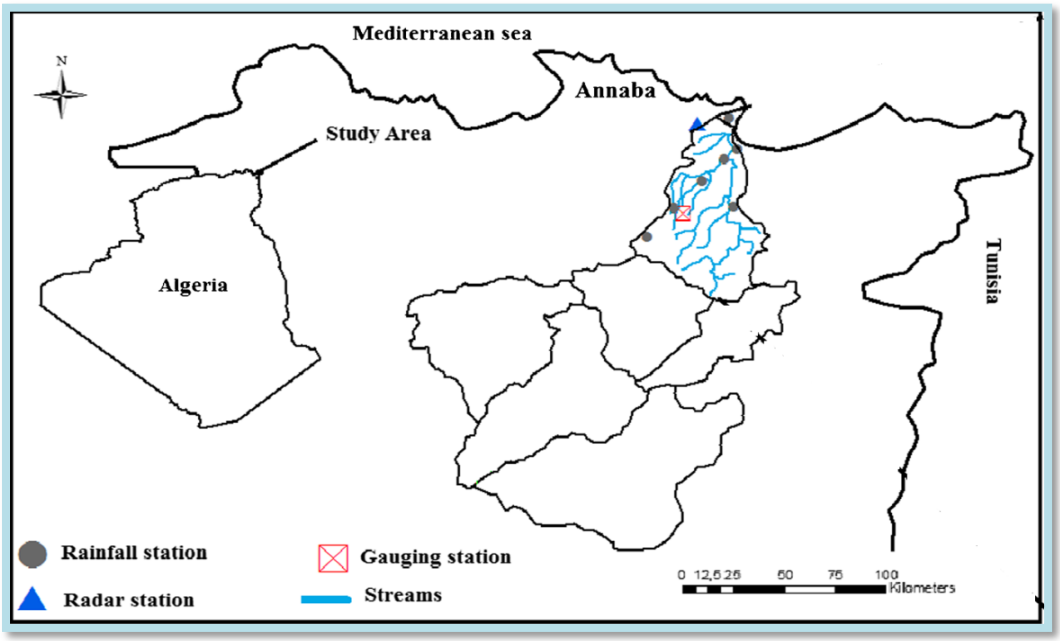


Figure III.1: Location of the study area and positions of the rain gauge stations of Seybouse watershed

1.2 Presentation of data

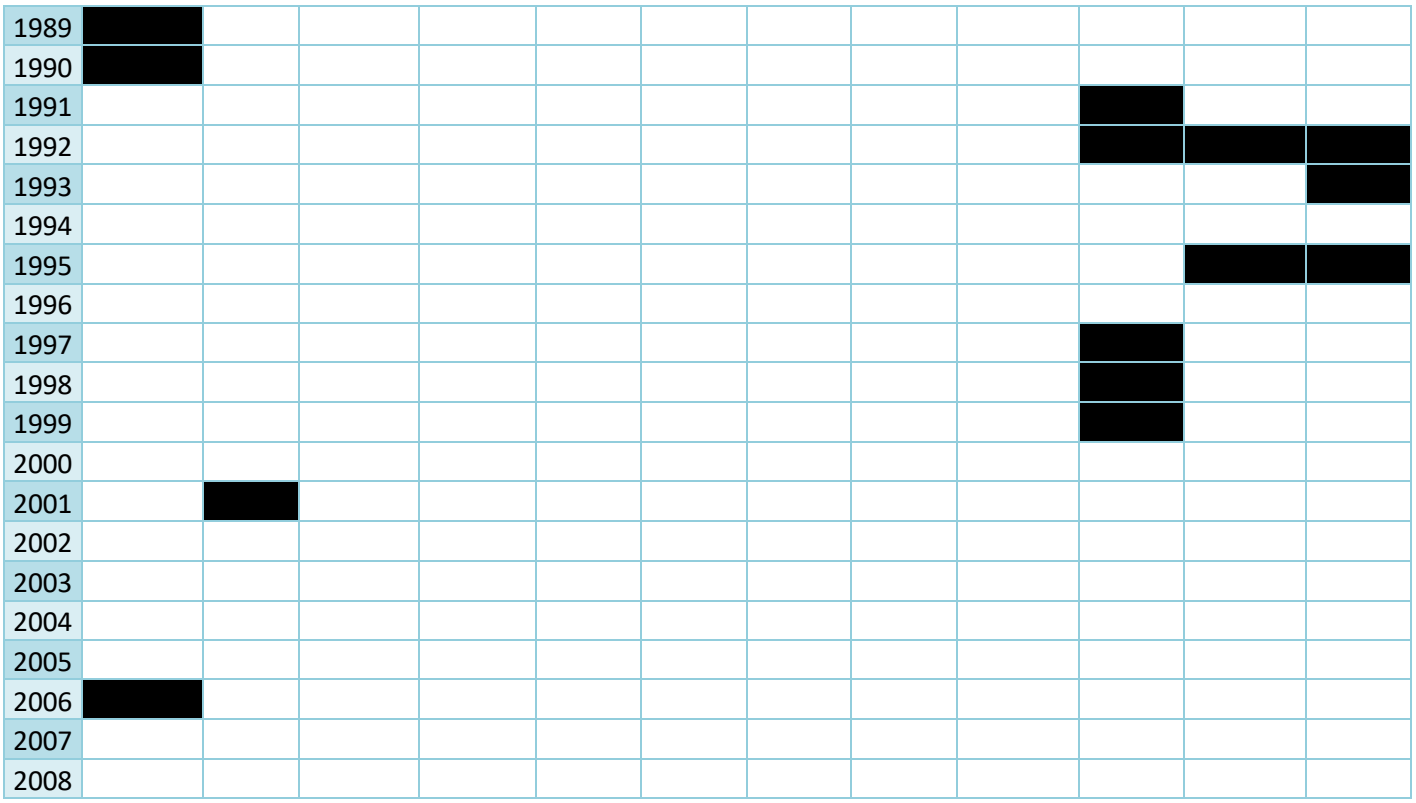
The Seybouse watershed has twenty rainfall stations, among them out of service stations and stations with gaps **(fig.III.1)**, for our case, we used only three stations (**table III.1, table III2**)

Table III.1: Rainfall station of Seybouse watershed

Station name	Coded	Coordinates			Observation Period
		X(KM)	Y (KM)	Z(M)	
Ain Barda	140666	937,288	387.737	55	1970-2008
Kef Mourad	140611	953.288	389.544	-	1970-2008
Pont Bouchet	140631	349,976	402,826	47	1970-2008

Table III.2: El Karma station observation time with gaps

	September	October	November	December	January	February	march	April	may	June	July	august
1970	■	■	■	■	■	■	■	■	■		■	
1971											■	
1972												
1973												
1974							■	■	■	■		■
1975	■	■	■	■	■	■	■	■	■	■	■	
1976	■											
1977												
1978												
1979												■
1980	■	■	■		■							
1981										■		
1982												■
1983									■	■		■
1984	■											
1985												■
1986												■
1987												
1988												



Existing data

Missing data

2. Oued Ruhmel watershed

The oued Rhumel is located in Eastern Algeria, It extends between altitudes 36°15' – 36° 35' and longitudes 6° 10' – 6° 20' it is bordered to the south by the Tellian highlands, to the West by the mountain of Petite Kabylie, to the North by the border of division of the waters formed by the Atlas Tellien, and to the east by the basin of the Seybousse. The Kébir Rhumel basin has a total area of 8735 km2. It splits into two large distinct parts. The western part, consisting of the basin of the Enndjawadi of an area of 2,169 km2, is characterized by relatively high rainfall (700mm/year on average) and mountainous topography. In this basin, the ratings reach 1,400, or any further. The Kebir-Rhumel basin covers six wilayas "Oum El Bouaghi - Skikda- Constantine- Mila-Sétif- Jijel". **(fig.III.2)**

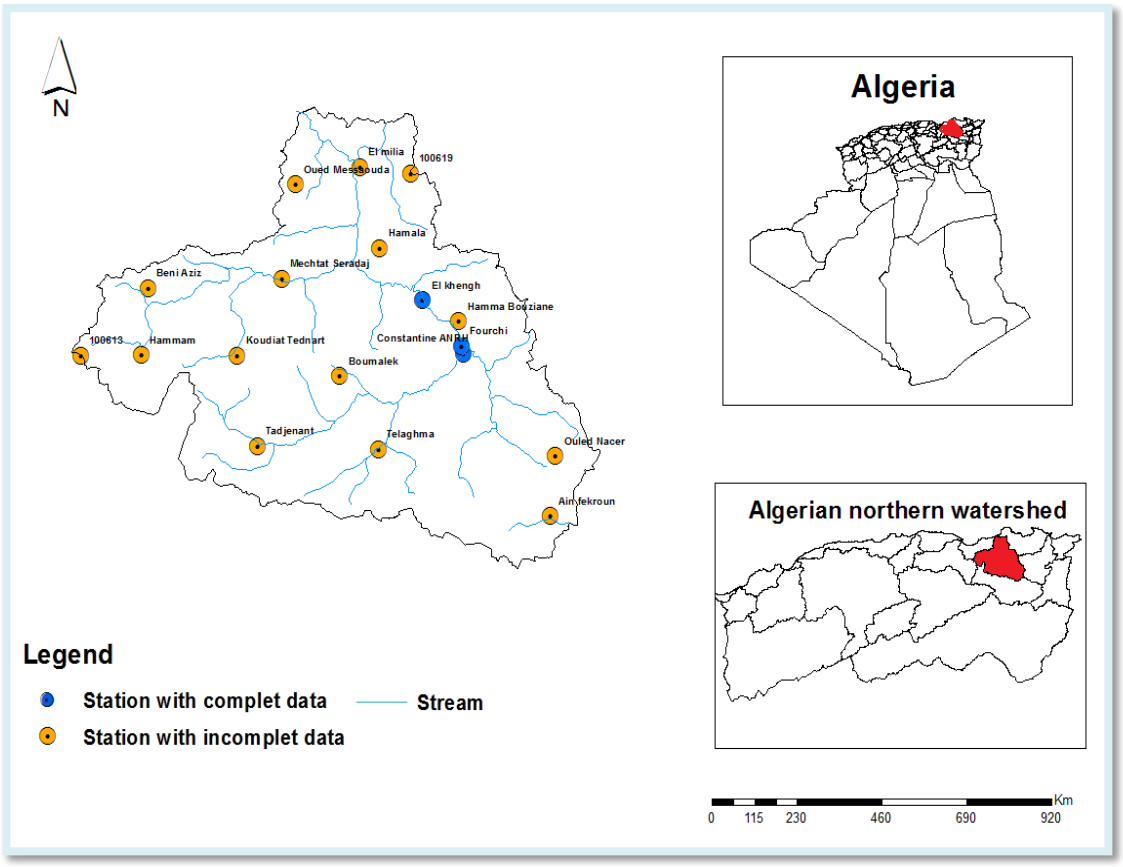


Figure III.2: Location of the study area and positions of the rain gauge stations of Oued Ruhmel watershed

2.1 Presentation of data

The Oued Rhumel watershed has eighteen rainfall stations, among them out of service stations and stations with gaps **(figIII.2),** for our case, we used only two stations **(table.III.3**,**table III.4).**

TABLE III.3: Rainfall station of Oued Ruhmel watershed

Station Name	Code	Coordinates			Observation Period
		X (KM)	Y (KM)	Z(M)	
Constantine ANRH	100410	850,350	344,750	595	1984-2012
El Kheneg	100620	849,850	357,450	300	1984-2012

Table III.4: Hamma Bouziane station observation time with gaps

	September	October	November	December	January	February	march	April	may	June	July	august
1984												
1985												
1986												
1987												
1988												
1989												
1990												
1991												
1992												
1993												
1994												
1995												
1996												
1997												
1998												
1999												
2000												
2001												
2002												
2003												
2004												
2005												
2006												
2007												
2008												
2009												
2010												
2011												
2012												

3. Coastal Constantinois Centre watershed

The Coastal Constantinois Centre watershed is located in the northeast of Algeria. H decomposes of three large basins; West, Center and East coastal Constantinois. This poolcovers an area of 11119 Km2, and limited to the North by the Mediterranean Sea, to the South by the Kebir Rhumel basin (10), and that of Seybouse (14) and the Medjerda basin (12), to the west by the Soummam basins (15), and to the east by the Tunisian border.The Constantine coastal basin covers ten wilayas "Bejaia -Jijel - Skikda-Annaba -El Tarf-Setif-Mila-Constantine-Guelma-Souk Ahras".**(fig.III.3).**

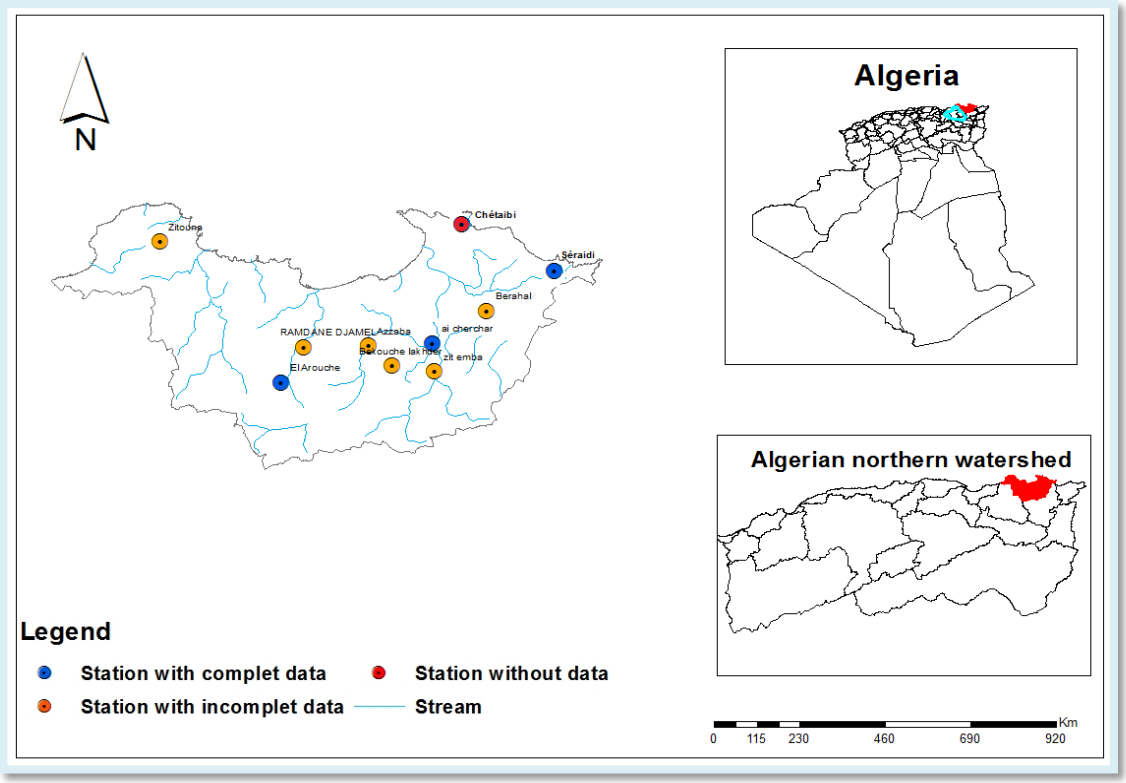


Figure III.3: Location of the study area and positions of the rain gauge stations of coastal Constantinois watershed

3.1 Presentation of data

The Coastal Constantinois Centre watershed has ten rainfall stations, among them out of service stations and stations with gaps **(figIII.3),** for our case, we used only two stations **(table III.5, tableIII.6)**

TABLE III.5: Rainfall station of coastal Constantinois watershed

Station Name	Code	Coordinates			Observation Period
		X(KM)	Y (KM)	Z(M)	
Azzaba	031106	892.3	391.35	91	1970-2008
Ain Cherchar	031201	909.5	393.2	34	1970-2008

Table III.6 Bekouche lakhdar station observation time with gaps

	September	October	November	December	January	February	march	April	may	June	July	august
1970												
1971												
1972												
1973												
1974												
1975												
1976												■
1977												
1978												■
1979												
1980								■				
1981												
1982	■											
1983	■											
1984												
1985												
1986										■	■	■
1987	■											
1988												
1989							■	■	■	■	■	■
1990	■	■									■	
1991										■	■	■
1992									■	■	■	■
1993	■							■		■	■	■
1994			■									
1995										■	■	
1996										■		
1997												
1998								■	■	■	■	■
1999										■	■	■
2000												
2001												
2002												
2003												
2004												
2005												
2006					■							
2007										■		
2008									■			

4. Data standardization

Several approaches are used for the homogenization of climate data. Some authors use direct methods based on metadata, side-by-side instrument comparisons or statistical studies of instrument changes **[55].**These approaches do not consist in detecting a change with a statistical method; they rather allow to directly correcting the data series. They are advantageous and should be used when the change information is accurate. However, they are inapplicable when the information is vague, incomplete or non-existent. Since this is often the case, so-called 'indirect' homogenization methods, which use neighboring series to detect discontinuities, have been developed. Literature reviews of the different approaches are presented in **[56]**and **[57].**

4.1Metadata

Reliable and complete metadata are important to ensure that the datasets have been collected under consistent conditions and therefore that the conclusions resulting from their analysis will be valid **[56].** In the case where the observation conditions have changed over time, the metadata make it possible to retrace the history of the measuring station and thus to correct any breaks artificially induced by changes in observation conditions or even to qualify the results of analyzes carried out on these data. Metadata are made up of recordings or photographs of a station, its weather directories, inspection reports or interviews with the people responsible for them **[55].**

4.2 Neighboring reference and comparison series

The amplitude of the in homogeneities present in the climatic series can be of the same order of magnitude as that of the true fluctuations of the climate **[56].**As a result, the series to be homogenized (base series) is often compared with neighboring series (which belong to the same climatic region).

Neighboring series must be homogeneous and exposed to the same climate fluctuations as the base series. The use of neighboring series is not effective when all stations in the same region have undergone the same change **[58].** The use of several neighboring series makes it possible to minimize the risk that they are all affected by the same inhomogeneity. Some authors rather use a reference series (function of one or more neighboring series) to compare with the base series, to prevent inhomogeneity from affecting the correlation coefficient. To create the reference series, weights are assigned to neighboring series using the correlation coefficients obtained with the differentiated series.

To prevent in homogeneity from affecting the correlation coefficient. To create the reference series, weights are assigned to neighboring series using the correlation coefficients obtained with the differentiated series.

4.3 Methods used to homogenize climate data

Various homogenization techniques have been developed to accommodate different types of factors such as the variable to be homogenized, the spatial and temporal variability of the data depending on where the stations are located, the length of the series and the number of missing data, the available metadata and the density of the observation network **[56].** Homogenization techniques also vary according to the objective for which they are applied and the philosophy of each work team, despite the great diversity of homogenization methods, they can nevertheless be classified into two main categories that join them all: subjective or objective methods. When the location of a discontinuity is detected by the naked eye on a graph, the method belongs to the subjective class even if statistical tests are applied afterwards. On the other hand, objective methods do not depend on the judgment of the user to locate in homogeneities, among some of the well-known methods we mention:

- standard normal homogeneity test [Alexandersson, 1986; Khaliq and Ouarda,2007].
- multiple regression [Vincent, 1998].
- two-phase regression [Easterling and Peterson, 1995; Lund and Reeves, 2002].
- bivariate test [Maronna and Yohai, 1978; Potter, 1981].
- sequential Wilcoxon test [Karl and Williams, 1987; Lanzante, 1996; Ducré-Robitaille eal, 2003].
- sequential t-test [Gullett et al., 1990].
- Jaruskova's method [Jaruskova, 1996].
- Bayesian approach [Rasmussen, 2001].

4.4 Main causes of in homogeneities

The main causes of in homogeneities in the climatologically series depend on the parameter that is measured; changes in the hours of observations or calculation methods can cause inhomogeneity in the series of average temperature and humidity, but do not affect precipitation or pressure. Breaks in precipitation series can be due to changes in instrumentation, instrument height, immediate station surroundings, station exposure, and relocations. Displacements and changes in exposure are likely to introduce very significant breaks in the precipitation series **[59].**

Generally, the main cause of break in most climatological series is probably station displacement. Moreover, the displacement of a station is often accompanied by a change of instrumentation, observer and environment **[60].**

5. Verification of homogeneity

5.1 The Wilcoxon test

The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples **[61].** The test was popularized by Sidney Siegel (1956) in his influential textbook on non-parametric statistics **[62].** Siegel used the symbol *T* for the test statistic, and consequently, the test is sometimes referred to as the Wilcoxon *T*-test, The Wilcoxon test can be used to assess whether there is a significant difference between the mean of a series before the jump and the mean of the series after the jump. Obviously, it is necessary to know a priori the moment of the jump to be able to carry out this test **[63].**

The Wilcoxon test is the most effective in detecting and correcting anomalies or erroneous values **[29];** it is a nonparametric test which uses the series of the ranks of the observations, instead of the series of their values. If the sample (of rain for example) X comes from the same population Y, the sample X UY (union of X and Y) is also derived from it. We proceed as follows: Given an observation series of length N from which we draw two samples X and Y: N1 and N2 are respectively the sizes of these samples, with N = N1 + N2 and N1 ≤ N2.Then ranks the values of our series in ascending order.

Subsequently, we will only be interested in the rank of each of the elements of the two samples in this series. If a value is repeated several times, we associate the corresponding average rank with it. We then calculate the sum W x of the ranks of the elements of the first sample in the common series:

$$W\ x = \Sigma\ \text{Rank}\ x.$$

Wilcoxon a constitute a homogeneous series, the quantity W x is between two bounds W max and W min given by the following formulas:

❖ W min= $\frac{(N1+N2+1)N1-1}{2} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{N1N2(N1+N2+1}{12}}$ **(III.1)**

And :

❖ W max =$N1 + N2 + 1)N1 - Wmin$ **(III.2)**

❖ $Z_{1-\frac{\alpha}{2}}$ Represents the value of the reduced centered variable of the normal law corresponding to:

- $1-\frac{\alpha}{2}$[At the 95% confidence level we have].
- $Z_{1-\frac{\alpha}{2}}$=1.96

5.2 Wilcoxon test results

The Wilcoxon test results on the stations of the three watersheds are shown in the following tables:

TABLE III.7: Wilcoxon Test result on watershed station Seybouse

Station Name	Wilcoxon Test			Inequality	Observation
	W min	W max	Σ Rank (X)		
Ain Berda	40755.79	45563.21	41314	40755.79<41314<45563.21	Homogeneous
Kef Mourad	40452.98	45243.02	44124	40452.98<44124 <45243.02	Homogeneous
Pont Bouchet	40981.17	45753.83	41470	40981.17<41470 <45753.83	Homogeneous

TABLE III.8: Wilcoxon Test result on watershed station Oued Ruhmel

Station Name	Wilcoxon Test			Inequality	Observation
	W min	W max	Σ Rank (X)		
Constantine ANRH	28096,29	31759,71	22568	28096,29>22568<31759,71	inhomogeneous
El Kheneg	27114,54	30683,46	21749	27114,54>21749<30683,46	inhomogeneous

TABLE III.9: Wilcoxon Test results on watershed stations Coastal Constantinois Centre

Station Name	Wilcoxon Test			Inequality	Observation
	W min	W max	Σ Rank (X)		
Azzaba	51750.45	57992.55	58815	51750.45<58815>57992.55	inhomogeneous
Ain Cherchar	55295.27	61830.73	57599	55295.27<57599 <61830.73	Homogeneous

6 Grubbs and Beck horsains test

it is necessary to identify the outliers in the data. The presence of the outliers can affect the effective forecasting and estimation. Therefore, several tests have been applied for detecting and removing the outliers from the data. Among these tests, Grubbs's test which is introduced by **[64].** and recommended by ISO and has been widely applied for the detecting of outliers in the data. Grubbs's test is easy to apply and operated using the mean and standard deviation of the data. This test is used to test the null hypothesis that a suspected value is an outlier versus

the alternative hypothesis that the suspected value is not an outlier. The Grubbs's test value is compared with the tabulated value at a fixed value of the level of significance. The null hypothesis that the suspected value is an outlier is accepted if the Grubbs's test value is smaller than the tabulated value **[65].**applied Grubbs's test for detecting outlier in flood series ,Grubbs' tests (1950, 1969, 1972) were developed to help determine whether the largest value, the smallest value, the largest or smallest value, or in the case of the test of Grubbs doubles, if the two larger values, or if the two smaller ones can be considered extremes (or outliers). This test assumes that the data correspond to a sample from a population that follows a normal distribution.

The singular values called horsains, are deduced by calculating the following statistics:

$$X\ max=(\bar{x}+s)*kn \quad \textbf{(III.3).}$$

$$X\ min=(\bar{x}-s)*kn \quad \textbf{(III.4).}$$

$\bar{x}$ And S: Are, respectively, the mean and the standard deviation of the natural logarithms of the elements constituting the sample.

Kn: Statistical value of the Grubbs and Beck test, tabulated for different sample sizes and levels of significance.

At the 5% risk level, the following polynomial approximation was proposed by **[42].**

6.1 Grubbs and Beck test results

The Grubbs and Beck test results on the stations of the three watersheds are shown in the following tables:

TABLE III.10: Test of representativeness of Grubbs and Beck case of Seybouse watershed

	Ain Berda	**Kef Mourad**	**Pont Bouche**
N	416	415	416
X max	118,35	115,67	117,34
X min	11,26	8,82	11,62
Kn, 0.05	3.21	3.21	3.21

TABLE III.11: Test of representativeness of Grubbs and Beck case of Oued Ruhmel watershed

	Constantine ANRH	**El Kheneg**
N	347	341
X MAX	98,44	98,81
X MIN	7,93	6,64
KN, 0.05	3,21	3,21

TABLE III.12: Test of representativeness of Grubbs and Beck case of coastal constantinois Centre watershed

	Azzaba	**Ain Cherchar**
N	470	485
X MAX	94,95	116,40
X MIN	2,97	5,94
KN, 0.05	3.21	3.21

Conclusion

The collection and review of rainfall data play a crucial role in numerous fields, including meteorology, hydrology, agriculture, and water resource management. Rainfall data provide essential information about the temporal and spatial patterns of precipitation, which is vital for understanding weather patterns, climate trends, and water availability. Accurate and comprehensive rainfall data serve as the foundation for climate studies, forecasting models, and water management strategies. By collecting data from rain gauges, weather radars, satellites, and other observation systems, meteorologists and hydrologists can analyze rainfall characteristics, such as intensity, duration, and frequency, to assess climate variability and identify potential climate change impacts.

Chapter IV
Results and discussions

Introduction

In this chapter, we present an analysis for the estimation of monthly precipitation data of a Mediterranean climate in eastern Algeria through a case study of a North-East measurement network that includes the data of three watersheds, for the estimation of these rains; we used and compared several approaches LR, RP, AR and RPET. The effectiveness of these approaches was evaluated and calculated in order to find the best model for estimating rainfall data for all the stations in question.

I. Verify the effectiveness of the models

1. Explanation of the process

Before starting to estimate the missing in the stations used, we will start by ensuring the effectiveness of the artificial intelligence models we have, the principle is based on the elimination of 20% of rainfall data from each station that suffers from loss of data, after the models estimate their values finally the Correlation coefficient the errors of all the models are calculated using the evaluation measures commonly used in the field of regression and predictive modeling such as R^2, RMSE, MAE,MRE the results were as shown in the following tables:

Table IV.1: Results of the performance criteria of the established models in the testing phase, case of el karma station

Model	R2	RMSE	MAE	MRE (%)
LR	0.80	5.17	0.31	0.25
GP	0.79	0.28	0.31	6.22
MP	0.80	1.53	0.14	6.89
SMORG	0.80	0.25	0.27	5.66
KSTAR	0.80	0.27	0.33	5.61
LWL	0.79	7.17	0.2	0.29
AR	0.78	0.72	0.31	12.33
Bagging	0.80	5.68	0.19	0.28
CVPS	0.76	0.55	0.36	14.51
MS	0.76	0.55	0.36	14.51
RSS	0.80	0.29	0.23	6.43
RBD	0.79	0.32	0.27	9.14
Stacking	0.76	0.55	0.36	14.51
WIHW	0.76	0.55	0.36	14.51
IMC	0.76	0.55	0.36	14.51
DT	0.80	0.29	0.47	6.44
M5R	0.80	0.26	0.28	5.58
ZR	0.76	0.55	0.36	14.51
DS	0.79	0.31	0.20	8.89
M5P	0.80	5.58	0.28	0.26
RF	0.80	0.29	0.20	6.06
REPT	0.79	0.37	0.14	8

Table IV.2: Results of the performance criteria of the established models in the testing phase, case of el Hamma Bouziane station

Model	R2	RMSE	MAE	MRE (%)
LR	0.81	4,99	0,03	0,73
GP	0.81	5,60	0,10	0,32
MP	0.81	4,99	0,03	0,73
SMORG	0.81	4,47	0,02	0,39
KSTAR	0.81	4,76	0,13	0,28
LWL	0.81	7,49	0,25	0,44
AR	0.81	10.28	0.01	0.88
Bagging	0.81	4,06	0,10	0,26
CVPS	0.78	12,33	0,02	0,52
MS	0.78	12,33	0,03	0,53
RSS	0.80	14,60	0,57	1,70
RBD	0.81	4,74	0,25	0,29
Stacking	0.78	12.32	0.03	0.53
WIHW	0.78	13,06	0,01	0,54
IMC	0.78	12,33	0,03	0,53
DT	0.81	6,74	0,07	0,40
M5R	0.81	4,99	0,08	0,29
ZR	0.78	13,06	0,01	0,54
DS	0.80	8,41	0,42	0,48
M5P	0.81	3,99	0,03	0,30
RF	0.75	15,94	0,38	15,95
REPT	0.81	3,98	0,17	0,28

Table IV.3: Results of the performance criteria of the established models in the testing phase, case of Bekouche Lakhder station

Model	R2	RMSE	MAE	MRE (%)
LR	0.79	14,29	0,49	1,72
GP	0.79	13,97	0,44	1,90
MP	0.79	15,16	0,15	10,78
SMORG	0.79	14,16	0,39	36,51
KSTAR	0.79	14,05	0,46	1,87
LWL	0.77	15,31	0,80	0,69
AR	0.78	15,42	0,58	1,06
Bagging	0.76	15,23	0,45	2,21
CVPS	0.74	13,06	0,01	0,54
MS	0.74	13,06	0,01	0,54
RSS	0.78	14,60	0,57	1,70
RBD	0.78	15.43	0.5	0.75
Stacking	0.74	13,06	0,01	0,54
WIHW	0.74	13,06	0,01	0,54
IMC	0.74	13,06	0,01	0,54
DT	0.79	15,07	0,39	1,10
M5R	0.79	14,53	0,49	1,88
ZR	0.74	13,06	0,01	0,54
DS	0.77	15,55	0,56	0,69
M5P	0.79	14,10	0,44	1,88
RF	0.78	15,94	0,38	15,95
REPT	0.78	14,35	0,08	1,41

2. Observe and analyze tables

After testing the models on data from the three watersheds, we noticed that:

- Concerning the R2, the models gave good results in all the stations with values of 0.74-0.81 which means the good choice of the reference stations.
- Concerning the RMSE, the results showed that the following models GP, SMORG, KSTAR, AR, CVPS, MS, RSS, RBD, STACKING, WIHW, IMC, DT, M5R, ZR, DS, RF, REPT, have a good performance on the data from the el kerma station with values lower than 1 mm. for the other stations, the RMSE values are a little high for some models with error values ranging between 10.28 and 15.43 mm.
- Concerning MAE, the models gave excellent results in the three stations where the results were lower than 1.
- For MRE the values of 27.20 % of the models used are less than 1, for the rest of the models the values range between 1 and 36.51 mm.

Finally, the model evaluation results showed good results, so they are ready to be used to estimate the missing values for the stations of the three watersheds.

III Estimate the missing data

1. Case of SEYBOUSE watershed

1.1 Data Analysis

Table IV.4: Monthly Data characteristic of Seybouse watershed stations

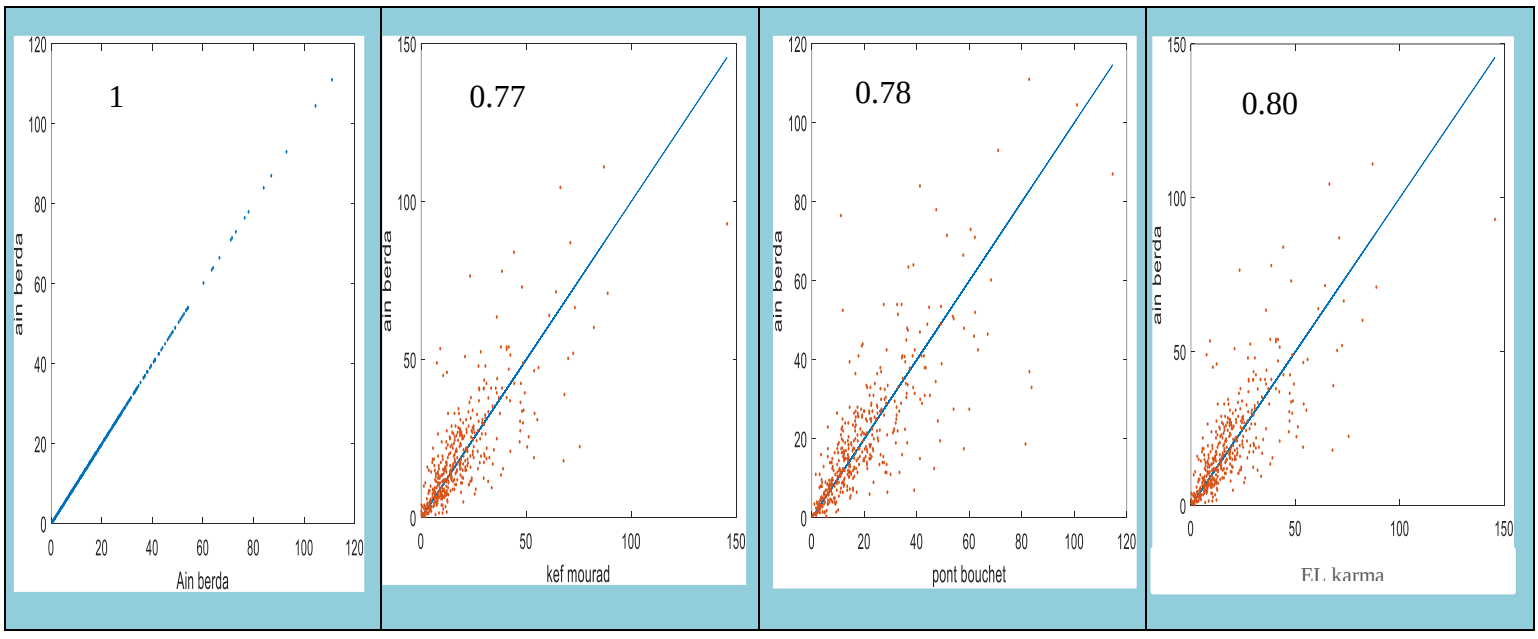
station	Minimum (mm)	Maximum (mm)	Mean (mm)	missing data	period observation
Ain Barda	0.4	111	20.118	0	1970-2008
Kef Mourad	0.1	145.6	19.392	0	1970-2008
Pont Bouchet	0.2	114.6	20.469	0	1970-2008
El Karma	0.9	106.3	20.443	55	1970-2008

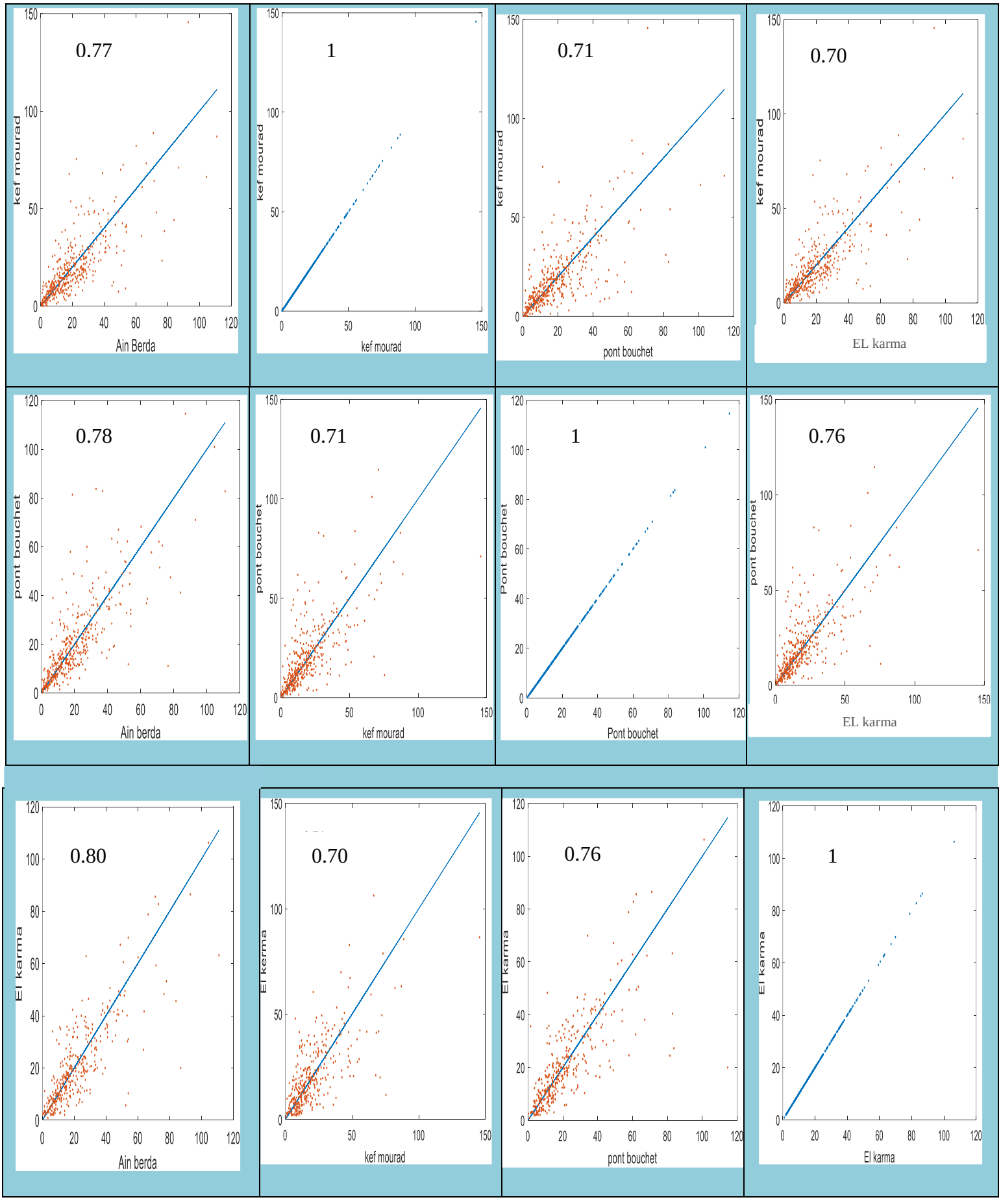
1.2 Correlation matrix

Table IV.5: Correlation matrix for stations of Seybouse watershed

Station	Ain Berda	Kef Mourad	Pont Boucher	El Karma
Ain Berda	**1**	0.77	0.78	0.80
Kef Mourad	0.77	**1**	0.71	0.70
Pont Boucher	0.78	0.71	**1**	0.76
El Karma	0.80	0.70	0.76	**1**

Figure IV.1: Curves of correlation matrix for stations of Seybouse watershed




0.77
kef mourad
Ain Berda
1
kef mourad
kef mourad
0.71
kef mourad
pont bouchet
0.70
kef mourad
EL karma
0.78
pont bouchet
Ain berda
0.71
pont bouchet
kef mourad
1
Pont bouchet
Pont bouchet
0.76
pont bouchet
EL karma
0.80
El karma
Ain berda
0.70
El karma
kef mourad
0.76
El karma
pont bouchet
1
El karma
El karma

1.3 Perform evaluation of estimation models

Table IV.6: Results of the performance criteria of the established models (El kerma station)

Model	R2	RMSE	MAE	MRE (%)
LR	0.85	8.32	5.57	47.87
GP	0.85	8.43	5.9	50.7
MP	0.89	8.82	7.2	61.83
SMORG	0.84	8.51	5.42	46.52
KSTAR	0.99	1.50	0.83	7.11
LWL	0.78	9.85	7.31	62.79
AR	0.87	7.75	5.47	46.95
Bagging	0.89	6.99	4.65	39.98
CVPS	0	15.65	11.64	100
MS	0	15.65	11.64	100
RSS	0.89	7.61	5.12	43.99
RBD	0.89	7.20	4.92	42.24
Stacking	0	15.65	11.64	100
WIHW	0	15.65	11.64	100
IMC	0	15.65	11.64	100
DT	0.81	9.14	6.55	56.24
M5R	0.85	8.32	5.57	47.87
ZR	0	15.65	11.64	100
DS	0.66	11.7	8.61	73.9
M5P	0.85	8.32	5.57	47.87
RF	0.97	3.61	2.38	20.46
REPT	0.82	8.99	6.20	53.28

2. Case of a Oued Rhumel watershed

2.1 Data Analysis

Table IV.7: Monthly Data characteristic of Oued Rumhel watershed stations

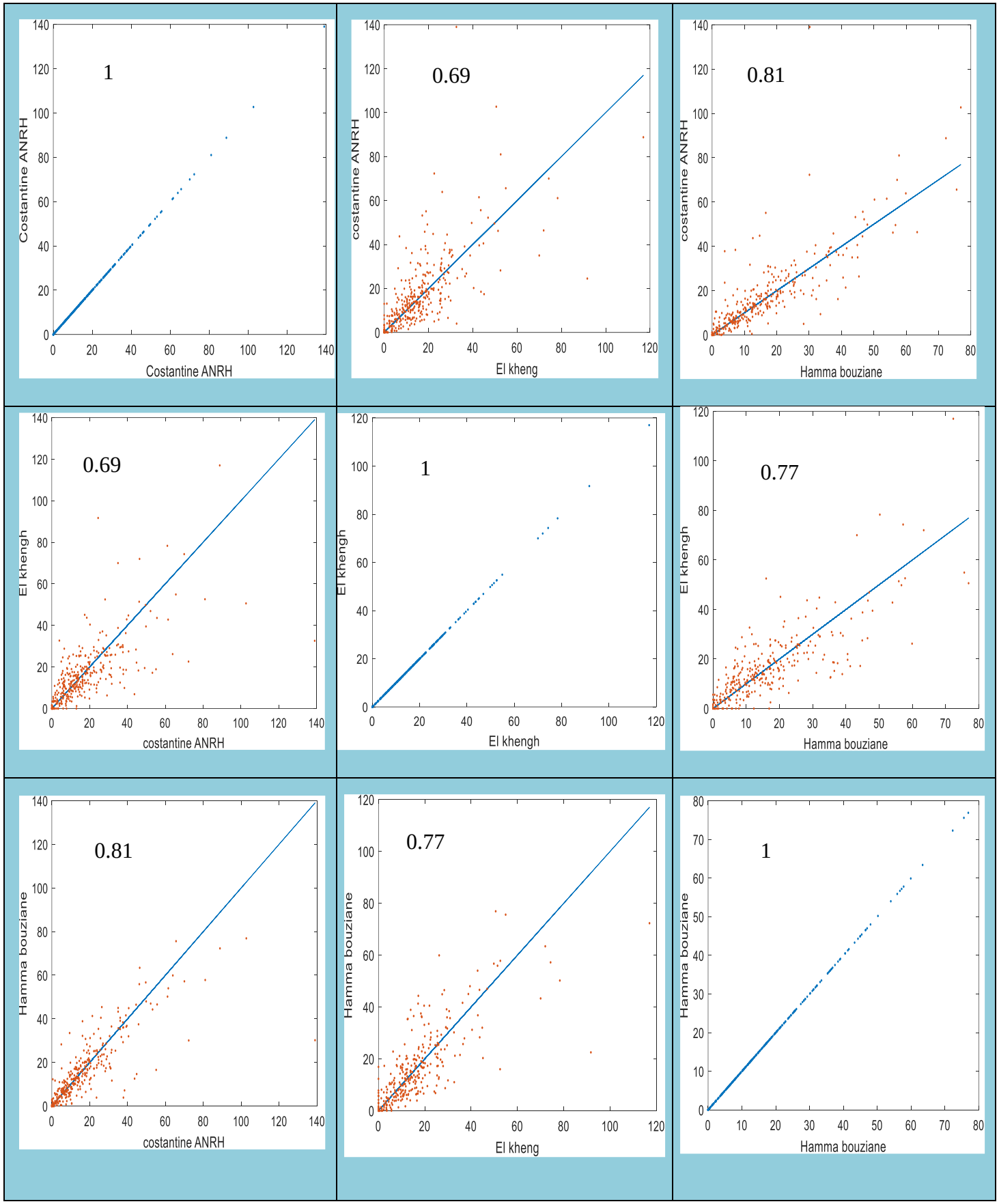
station	Minimum (mm)	Maximum (mm)	Mean (mm)	Missing data	Period observation
Hamma Bouziane	0	76.9	16.441	35	1970-2008
Constantine ANRH	0	139	17.384	0	1970-2008
El Khengh	0	117	16.426	7	1970-2008

2.2 Correlation matrix

Table IV. 8: Correlation matrix for stations of Oued Ruhmel watershed

Station	**Constantine ANRH**	**El Khengh**	**Hamma Bouziane**
Constantine ANRH	**1**	0.69	0.81
El Khengh	0.69	**1**	0.77
Hamma Bouziane	0.81	0.77	**1**

Figure IV.2: Curves of correlation matrix for stations of Oued Ruhmel watershed



2.3 Perform evaluation of estimation models

Table IV.9: Results of the performance criteria of the established models (Hamma Bouziane station)

Model	R^2	RMSE	MAE	MRE (%)
LR	0.86	7.42	4.93	47.13
GP	0.86	7.33	5.41	51.68
MP	0.93	6.11	4.53	43.34
SMORG	0.86	7.54	4.53	45.30
KSTAR	0.98	2.71	1.87	17.84
LWL	0.82	8.12	6.2	59.25
AR	0.88	6.73	4.89	46.75
Bagging	0.93	5.25	3.64	34.74
CVPS	0	14.10	10.46	100
MS	0	14.10	10.46	100
RSS	0.92	7.63	3.9	37.36
RBD	0.92	5.48	3.69	35.22
Stacking	0	14.10	10.46	100
WIHW	0	14.10	10.46	100
IMC	0	14.10	10.46	100
DT	0.83	7.79	5.77	55.18
M5R	0.92	5.55	3.88	37.12
ZR	0	14.10	10.46	100
DS	0.69	10.26	7.74	73.94
M5P	0.91	5.94	4.1	39.18
RF	0.98	2.63	1.82	17.42
REPT	0.89	6.43	4.73	45.19

3. Case of COASTAL Constantinois Center watershed

3.1 Data Analysis

Table IV.10: Monthly Data characteristic of Coastal Constantinois Center watershed Stations

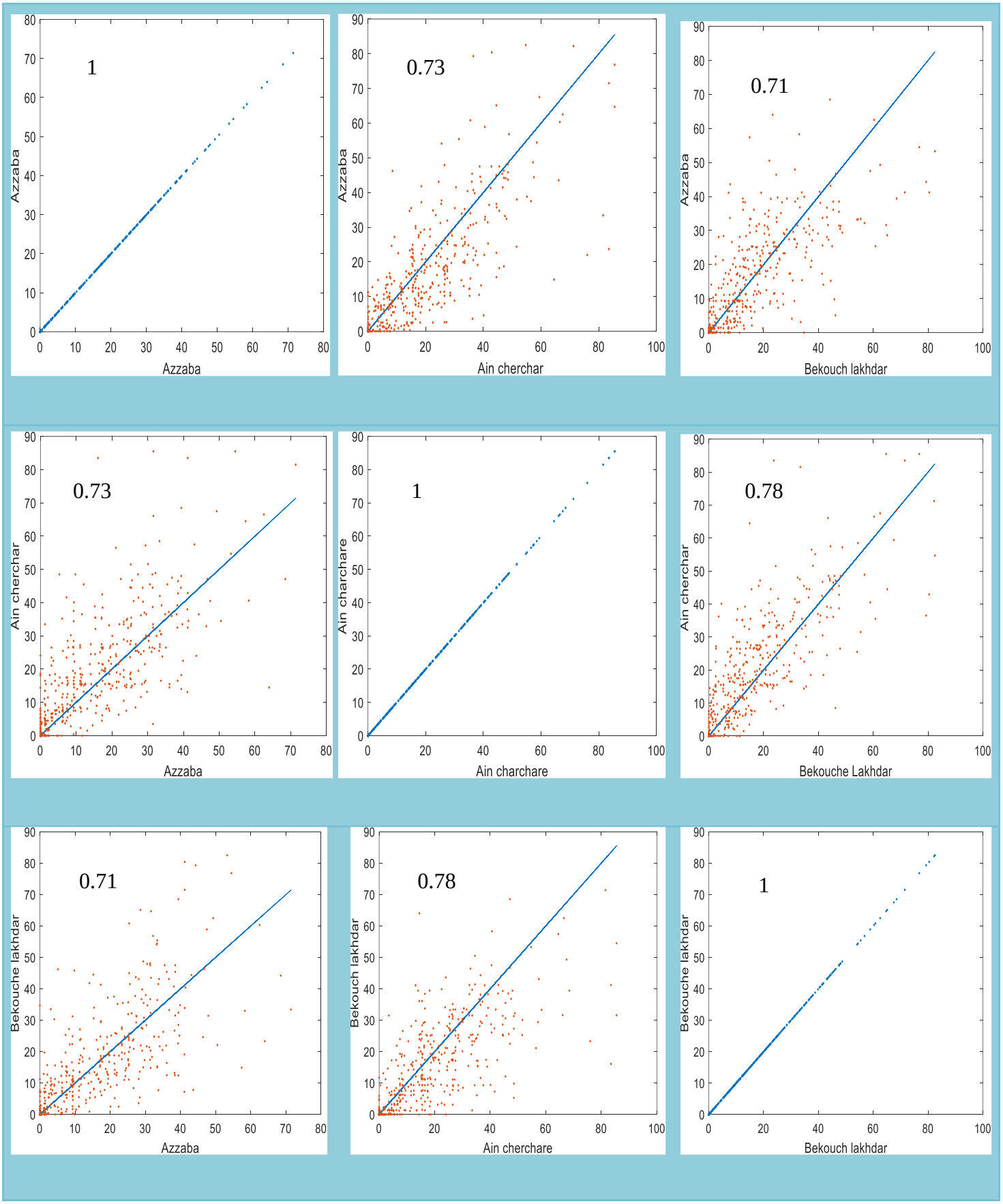
station	Minimum (mm)	Maximum (mm)	Mean (mm)	Missing data	Period Observation
Azzaba	0	71.4	15.252	16	1970-2008
Bekouche Lakhdar	0	82.5	17.136	45	1970-2008
Ain Cherchar	0	85.5	19.056	9	1970-2008

3.2 Correlation matrix

Table IV.11: Correlation matrix for stations of Coastal Constantinois Center Watershed

Station	**Azzaba**	**Ain Cherchar**	**Bekouche Lakhdar**
Azzaba	**1**	0.73	0.71
Ain Cherchar	0.73	**1**	0.78
Bekouche Lakhdar	0.71	0.78	**1**

Figure IV.3: Curves of correlation matrix for stations of Coastal Constantinois Center watershed



3.3 Perform evaluation of estimation models

Table IV.12: Results of the performance criteria of the established models (Bekouche Lakhder)

Model	R2	RMSE	MAE	MRE (%)
LR	0.81	9.78	6.75	53.12
GP	0.81	9.82	6.93	54.57
MP	0.86	12.34	9.59	75.5
SMORG	080	9.91	6.55	51.56
KSTAR	0.95	5.58	3.71	29.18
LWL	0.78	10.42	7.74	60.91
AR	0.81	9.7	6.94	54.68
Bagging	0.88	7.83	5.35	42.1
CVPS	0	16.54	12.7	100
MS	0	16.54	12.7	100
RSS	0.84	9.34	6.47	50.97
RBD	0.87	8.01	5.47	43.04
Stacking	0	16.54	12.7	100
WIHW	0	16.54	12.7	100
IMC	0	16.54	12.7	100
DT	0.77	10.58	7.32	60.80
M5R	0.81	9.72	6.6	51.99
ZR	0	16.54	12.7	100
DS	0.66	12.43	9.45	74.32
M5P	0.81	9.76	6.62	52.1
RF	0.97	4.15	2.85	22.46
REPT	0.83	9.27	6.54	51.51

4. Classification of results

After we tested several models in the stations that suffer from a lack of data, we categorized the results in the following table:

Table IV. 13: Classification of results for all stations

Excellent results	Good results	Average results	Bad results
KSTAR	Linear regression (LR)	Decision Stump (DS)	CV Parameter Selection (CVPS)
Random Forest (RF)	Gaussian processes (GP)		Multi Scheme (MS)
	multilayer preceptor (MP)		Stacking
	SMOREG		Weighted Instances Handler Wrapper (WIHW)
	LWL		Input Mapped Classifier (IMC)
	additive Regression (AR)		Zero R (ZR)
	Bagging		
	Random Sub Space (RSS)		
	Regression By Discretization (RBD)		
	Decision Table (DT)		
	M5 Rules (M5R)		
	M5P		
	REP Tree (REPT)		

5. Prediction of missing data

To choose the models to estimate the lost rainfall values in all stations, we used the K-STAR and RF models because they showed their efficiency compared to other models. The estimation results are shown in the following tables:

5.1 KSTAR model for El Karma station

Table IV.14: Missing data simulation results in El Karma station

N°	Data Prediction (mm)	N°	Data Prediction (mm)	N°	Data Prediction (mm)
01	4.13	21	19.97	41	4.71
02	18.55	22	16.11	42	6.05
03	5.87	23	29.56	43	7.58
04	18.71	24	13.35	44	7.02
05	20.6	25	23.32	45	4.1
06	26.98	26	6.14	46	3.27
07	23.69	27	9.78	47	3.18
08	28.69	28	32.04	48	4.50
09	23.86	29	6.77	49	4.34
10	4.86	30	27.82	50	4.21
11	4.31	31	15.02	51	4.27
12	20.36	32	29.15	52	4.04
13	4.01	33	12.15	53	4.56
14	31.36	34	3.84	54	3.64
15	3.77	35	3.78	55	6.30
16	5.92	36	5.92		
17	4.37	37	4.6		
18	4.12	38	3.27		
19	13.48	39	11.93		
20	13.75	40	3.38		

3.2 RF model for Hamma Bouziane station

Table IV. 15: Missing data simulation results in Hamma Bouziane station

N°	Data Prediction (mm)	N°	Data Prediction (mm)	N°	Data Prediction (mm)
01	8.06	**13**	6.39	**25**	1.57
02	31.66	**14**	15.34	**26**	2.12
03	16.78	**15**	13.76	**27**	15.45
04	9.04	**16**	34.74	**28**	7.36
05	12.76	**17**	23.08	**29**	11.59
06	25.49	**18**	18.29	**30**	19.52
07	14.41	**19**	30.47	**31**	20.82
08	20.99	**20**	24.07	**32**	14.77
09	13.14	**21**	10.55	**33**	16.19
10	11.3	**22**	20.83	**34**	12.33
11	10.2	**23**	12.99	**35**	11.5
12	0.74	**24**	7.19		

5.3 RF model for Bekouche Lakhdar

Table IV. 16: Missing data simulation results in Bekouche Lakhdar station

N°	Data Prediction (mm)	N°	Data Prediction (mm)	N°	Data Prediction (mm)
01	3.23	**21**	1.72	**41**	12.47
02	0.37	**22**	19.07	**42**	1.17
03	6.97	**23**	0.58	**43**	0.4
04	30.11	**24**	8.27	**44**	3.29
05	10.63	**25**	1.06	**45**	11.20
06	1.11	**26**	4.4	**46**	10.59
07	1.15	**27**	20.48	**47**	2.34
08	0.53	**28**	1.1	**48**	0.4
09	27.25	**29**	0.4	**49**	3.18
10	16.18	**30**	2.01	**50**	1.27
11	16.74	**31**	13.73	**51**	0.4

12	7.59	**32**	14.62	**52**	3.7
13	3.85	**33**	0.29	**53**	15.95
14	0.4	**34**	0.4	**54**	1.51
15	11.61	**35**	7.04	**55**	22.39

Conclusion

The availability of complete rainfall data is crucial for understanding the climate, predicting extreme weather events, managing water resources and scientific research. This data is essential for making informed decisions and implementing adaptation measures in the face of climate change and water-related challenges. This study aims to complete the missing data on the maximum monthly precipitation in the northeast of Algeria using several large families of LR, RP, AR and REPT algorithms, the test of these models has shown that the latter have shown a large performances concerning the estimation of the missing values, after we used the K-STAR model for the estimation of the values of the missing rains of the stations in question.

General Conclusion

General conclusion

Hydrometeorological data plays a critical role in numerous fields and applications. Its diverse uses span across various sectors and contribute to informed decision-making and effective management. In weather forecasting, hydrometeorological data, encompassing variables like rainfall, temperature, humidity, wind speed, and atmospheric pressure, are essential inputs for models that predict short-term and long-term weather conditions. This information enables meteorologists to issue forecasts and warnings, aiding sectors such as transportation, agriculture, and emergency preparedness.

The uncertainty of rainfall data can have a significant influence on the hydrological domain. Accurate rainfall data are essential for estimating water inflows to watersheds, groundwater recharge, stream flow and water resource availability. When there are gaps or errors in rainfall data, this can lead to inaccurate estimates and limited understanding of hydrological processes.

in this study we tried to complete the missing data of the maximum monthly precipitation in the north-east of Algeria using several large families of algorithms LR, RP, AR and REPT, the test of these models showed that the latter gave great performance in estimating missing values in northern Algeria.

In the perspective, it is essential to apply these methods for other areas with different types of climate and also to use combinations between the models and test their performance.

Bibliographic References

Bibliographic references

[1] HUNT JAMES Y JOHN SCHERMERHORN. (2004), Behavior Organizational México D.F, Limusa Wiley, p.278.

[2] DEVUYST P., 1972. La météorologie : comprendre, interpréter, appliquer la météorologie. Édition, Eyrolles. Paris, France, 164p.

[3] DECONINCK J.F., 2014. Paléoclimats : l'enregistrement des variations climatiques .2ème édition, Vuibert. Paris, France, 240p.

[4] CEA, 2013. Le climat : Observé le passé préservé l'avenir. Maya presse. France, 40p

[5] MERDAOUI Z., 2007. Caractérisation radiométrique des sites de Bouzaréah et de Ghardaïa. Mémoire de magister, Université de Blida, 127p.

[6] NIA M., 2010. Etudes comparatives des méthodes d'estimation du rayonnement solaire. Mémoire de magister, Université de Sétif, 107p

[7] KOLI BI ZUELI B., et PAULINE DIBI K., 2001. Initiation a la climatologie : climatologie générale. Cour pédagogique, Université d'Abidjan, 47p.

[8] DELMAS R., CHAUZY, S., et VERSTRAETE, J. M., 2007. Atmosphère, océan et climat. Belin pour la science, Paris, France, 287p.

[9] QA INTERNATIONAL COLLECTIF, 2007. La météo : Comprendre le climat et l'environnement. Québec Amérique. Canada, 128p

[10] VIGNEAU J.P., 2005. Climatologie. 2ème édition, ARMAND COLIN. Paris, France, 200p.

[11] DOUCET R., 2006. La science agricole : le climat et les sols agricoles. Bereger A.C. Québec, 444p.

[12] AKSOUH A., 2017. Caractérisation des intensités de pluie dans la région centre-est de l'Algérie en termes d'évolution temporelle et spatiale. Mémoire de master, Ecole nationale supérieure d'hydraulique, Blida, 58p.

[13] GUYOT G., 1999. Climatologie de l'environnement. 2ème édition, Dunod, Paris, France, 544p.

[14] FOUCAULT A., 2009. Climatologie et paléoclimatologie. 2ème édition, Dunod, Paris, France, 320p.

[15] BRAHMI D., 2014. Analyse spatio-temporelle des pluies en Algérie. Mémoire de

master, Université de Tlemcen, 63p.

[20] Sanchez-Diesma R, I., Zawadski and Semper-torres., (1970).Identification of the bright band through the analysis of volumetric radar data, journal of the atmospheric sciences 27, pp 299-307.

[21] ROCHE PA, (1963).Surface hydrology, OVERSEAS research and technical office (Paris), Gauthier-Villars Editor.

[22] ABABSA. B, (2018), Etude de l'évolution des précipitations dans la région de Guelma ; University Mohamed Kheider-Biskra.

[23] MORELL M., (1999).Acquisition and compilation of basic hydrological information. HGA Bucharest edition. p203.

[24] Laborde. J.P, (2000), Eléments d'Hydrologie de Surface ; Université de Nice - Sophia Antipolis, Centre National de la Recherche Scientifique.

[25] KEBLOUTI M, (2018).Les instruments des mesures en hydrologie, centre university Mila

[26] J.P. LABORDE, 2009.cours éléments d'hydrologie de surface université de Nice – Sophia Antipolis, 202 pages.

[27] DELOBBE, L., 2006. Estimation des précipitations à l'aide d'un radar météorologique, Institut Royal Météorologique de Belgique.

[28] RINEHART, R.E., 1997. Radar for meteorologists: Third edition. Rinehart Publishing.

[29] ABD ELWAHAB, SARI AHMED, 2002. initiation à l'hydrologie de surface, office des publications universitaires.

[30] TOUAIBIA IMAN.(2011) Département hydraulique faculté de technologie .université de Telmcen BP.230-1300 Algérie. ESTIMATION BIAIS DU MODELE REGRISSIF PUISSANCE CAS DU BASSIN VERSANT DU K'SOB

[31] SARLE, WARREN S. (1994). "Neural Networks and statistical models" . SUGI 19: proceedings of the Nineteenth Annual SAS Users Group International Conference. SAS Institute. pp. 1538–50

[32] RUSSELL, STUART; Norvig, Peter (2003) et (1995). Artificial Intelligence: A,Modern Approach (2nd ed.). Prentice Hall

[34] ALPAYDIN, E. (2020). Introduction to machine learning. MIT press.

[35] MARSLAND, S. (2015). Machine learning: an algorithmic perspective. CRC press.

[36] ACEMOGLU, D , RESTREPO, P. (2018). Artificial Intelligence, automation, and work (No. w24196).National Bureau of Economic Research.

[37] BISHOP, C. M. (2006), Pattern Recognition and Machine Learning, Springer

[39] SARASA-CABEZUELO, A. (2022). " Prediction of Rainfall in Australia Using Machine Learning." Information 13(4): 163.

[40] Narimani, R., C. Jun, et al. (2022).

[41] ZIELINSKI, W., J. KAJEWSKA-SZKUDLAREK, et al. (2018). "Filling missing meteorological data with Computational Intelligence methods." ITM Web of Conferences 23: 00015.

[42] LLAMAS J ,1992. Hydrologie générale. Principes et applications, 2ème édition, gaëtanmorin, québec .

[43] PAF-2010 Analyze quantitative de problème de gestion Louis Houde Department de Mathématiques ET d'informatique University du Québec à Trois-Rivières

[44] Regression multiple: principles ET examples d'application Dominique Laffly

UMR 5 603 CNRS University de Pau ET des Pays de l'Adour October 2006

[45] Naoum.S, Tsanis.I.K.Ranking Spatial Interpolation Techniques Using a Gis-Based Dss. Global Nest 2004; 06:1-20.

[48] DIRKS, K. N., HAY, J. E., Stow, C.D. Et Harris, D.High-resolution studies of rainfall on Norfolk Island. Part II: interpolation of rainfall data. Journal of Hydrologie. 1998; 208(3-4): 187-193.

[49] EIBE FRANK, Mark A. Hall, and Ian H. Witten 2016 Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann

[52] KIDD, C., A. BECKER, G. J. HUFFMAN, C. L. MULLER, P. JOE,G. SKOFRONICK-JACKSON, D. B. KIRSCHBAUM, 2017:So, how much of the Earth's surface is covered byrain gauges? Bull. Amer. Meteor. Soc., 98, 69–78

[53] TAPIADOR, F. J., and COAUTHORS, 2017: Global precipitation measurements for validating climate models. Atmos. Res., 197, 1–20.

[54] AIDA TADILA. , 2008 .typologie de rapports entre la ville de constantine et son rhumel-boumerzoug. Mémoire pour l'obtention du diplôme de magistère

[55] PETERSON, T.C., EASTERLING, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S.,Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T.,Salinger, J., Forland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., Parker, D. (1998) Homogeneity adjustments of in situ atmospheric climate data: areview. International Journal of Climatology 18(13), 1493-1517.

[56] AGUILAR, E., AUER, I., BRUNET, M, PETERSON, T.C., WIERINGA, J. (2003) Guidelines on climate meta data and homogenization. WMO- TD No. 1186, World Meteorological Organization, Geneva, Switzer land.

[57] BEAULIEU, C., OUARDA, T.B.M.J., SEIDOU, O. (2007) Synthèse des techniques d'homogénéisation desséries climatiques et analyse d'applicabilité aux séries de précipitations. Hydrological Sciences-Journal-des Sciences Hydrologiques 52 (1), 18-37.

[58] STAUDT, M., ESTEBAN PARRA, M.J., CASTRO DIEZ, Y. (2007) Homogenization of long-term monthly. Spanish temperature data. International Journal of Climatology 27 (13), 1809-1823.

[59] DALY, C., GIBSON, W.P., TAYLOR, G.H., DOGGETT, M.K., SMITH, J.I. (2007) Observer bias in daily precipitation measurements at United States cooperative network stations. Bulletin of the American Meteorological Society 88 (6), 899-912.

[60] MESTRE, O. (2000) Méthodes statistiques pour l'homogénéisation des données climatiques. Thèse, Université Paul Sabatier, Toulouse, France.

[61] WILCOXON, FRANK (Dec 1945). "Individual comparisons by ranking methods" (PDF). Biometrics Bulletin. 1 (6): 80–83.

[62] SIEGEL, SIDNEY (1956). Non-parametric statistics for the behavioral sciences. New York: McGraw-Hill. pp. 75–83.

[63] DANY FAUCHER., Juin 1997. Revue bibliographique des tests des stationnarités, rapport de recherche n R-499, INRS 2-89146-393-5.

[64] GRUBBS, F.E., 1950. Sample criteria for testing outlying observations. Ann. Math. Stat.21 (1), 27–58.

[65] COHN, T, ENGLAND, J., BERENBROCK, C., MASON, R., STEDINGER, J.,

LAMONTAGNE, J., 2013. Ageneralized GrubbsBeck test statistic for detecting multiple potentially influential low outliers in flood series. Water Resour. Res. 49 (8), 5047–5058.

Internet website

[16] https://science.nasa.gov/earth-science/oceanography/ocean-earth-system/ocean-water-cycle

[17] https://www.eartheclipse.com

[18] https://www.nssl.noaa.gov/education/svrwx101/hail/types/

[19] https://biologyreader.com/types-of-precipitation-in-hydrology.html

[33] http://www.eolss.net/Eolss-sampleAllChapter.aspx

[38] https://www.researchgate.net/publication/368307495

[46] https://www.xmswiki.com/wiki/GMS:Natural_Neighbor

[47] https://iri.columbia.edu/~rijaf/CDTUserGuide/html/interpolation_methods.html

[50] https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70

[51] https://c3.ai/glossary/data-science/root-mean-square-error-rmse