

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Ecole Nationale Polytechnique



المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Hydraulic Department

Final year's project thesis

To obtain the State Engineer Diploma in Hydraulic

**Rainfall-Runoff modeling using Deep Learning
Application to Mediterranean climate**

Presented by: **Rania MOKHTARI & Maria AMEDDAH**

Under the supervision of **Mr. Abdelmalek BERMAD**, Professor at ENP

Co-directed by **Mr. Rafik OULEBSIR**, MCB at USTHB

Defended on the 07th of July 2022

Before the jury composed of:

President	Mr OULD HAMOU Malek,	Professor	ENP
Advisor	Mr BERMAD Abdelmalek,	Professor	ENP
Co-advisor	Mr OULEBSIR Rafik	MCB	USTHB
Examinator	Mr LEFKIR Abdelouahab,	Professor	ENSTP

ENP2022

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Ecole Nationale Polytechnique



المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Hydraulic Department

Final year's project thesis

To obtain the State Engineer Diploma in Hydraulic

**Rainfall-Runoff modeling using Deep Learning
Application to Mediterranean climate**

Presented by: **Rania MOKHTARI & Maria AMEDDAH**

Under the supervision of **Mr. Abdelmalek BERMAD**, Professor at ENP

Co-directed by **Mr. Rafik OULEBSIR**, MCB at USTHB

Defended on the 07th of July 2022

Before the jury composed of:

President	Mr OULD HAMOU Malek,	Professor	ENP
Advisor	Mr BERMAD Abdelmalek,	Professor	ENP
Co-Advisor	Mr OULEBSIR Rafik	MCB	USTHB
Examinator	Mr LEFKIR Abdelouahab,	Professor	ENSTP

ENP2022

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Ecole Nationale Polytechnique



المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Département Hydraulique

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'état en Hydraulique

**Modélisation Pluie-Débit en utilisant le Deep Learning
Application au climat méditerranéen**

Présenté par : **Rania MOKHTARI** et **Maria AMEDDAH**

Sous la direction de **Mr. Abdelmalek BERMAD**, Professeur A l'ENP

Co-dirigé par **Mr. Rafik OULEBSIR**, MCB A l'USTHB

Présenté et soutenu publiquement le (07/07/2022)

Composition du Jury

Président	Mr OULD HAMOU Malek	Professeur	ENP
Promoteur	Mr BERMAD Abdelmalek,	Professeur	ENP
Co-promoteur	Mr OULEBSIR Rafik	MCB	USTHB
Examineur	Mr LEFKIR Abdelouahab,	Professeur	ENSTP

ENP2022

ملخص:

تعد نمذجة جريان المطر أداة مهمة لإدارة الموارد المائية في مستجمعات المياه والتنبؤات بالمخاطر الهيدرولوجية مثل الفيضانات. تم إجراء العديد من الأبحاث من قبل علماء الهيدرولوجيا لإنتاج نماذج فعالة تولد استجابات مستجمعات المياه لهطول الأمطار. بشكل عام، تتضمن هذه النماذج معايير غالبًا ما تكون غير متوفرة، وحتى يصعب قياسها. لذلك، قد يكون من العملي التركيز على أساليب Deep Learning الجديدة، وهي أدوات قوية يمكنها فهم تعقيد العلاقة غير الخطية بين المدخلات والمخرجات دون الحاجة إلى اللجوء إلى عدة معايير.

في هذه الدراسة، استخدم المؤلفون نموذجين مختلفين RNN و LSTM على البيانات اليومية من 5 مستجمعات ذات طابع مناخي متوسطي اين اظهر نموذج LSTM نتائج أفضل لما تم تقييمه بـ NSE. أجريت تقييمات أخرى على نموذج LSTM بواسطة RSR و PBIAS أين اتضح أن كمية هطول الأمطار وتدفق السوابق هي المعطيات الأكثر تأثيرًا على النموذج.

الكلمات الدالة: النمذجة، هيدرولوجيا، Deep Learning, LSTM, RNN,

Résumé :

Les modèles pluie-débit sont des outils importants pour la gestion des ressources en eau dans les bassins versants et pour la prédiction des risques hydrologiques comme les crues. Plusieurs travaux de recherche ont été menés par des hydrologues pour produire des modèles performants générant le débit qui représente la réponse des bassins versants aux précipitations. Généralement, ces modèles impliquent des paramètres qui ne sont pas souvent disponible, parfois difficile à mesurer. Par conséquent, il peut être pratique de se focaliser sur les nouvelles méthodes de Deep Learning, qui sont performants, pour comprendre la complexité de relation entre les inputs et output sans avoir recours à plusieurs paramètres. Dans cette étude, les auteurs ont utilisé deux modèles différents (RNN et LSTM) sur des données journalières de 5 bassins versant avec un climat méditerranéen où le modèle LSTM a montré de meilleurs résultats pour ce qui a été évalué par le NSE. D'autres évaluations ont été faites sur le modèle LSTM par la RSR et la PBIAS où la précipitation et le débit précédent se sont avérés être les données les plus influentes sur le modèle.

Mots clés : modèle, hydrologique, Deep Learning, LSTM, RNN.

Abstract:

Rainfall-runoff modeling is an important tool for water resources management in watersheds and hydrological hazard predictions such as floods. Several research has been carried out by hydrologists to produce efficient models that generate the watersheds' responses to precipitation. Generally, these models involve parameters that are often unavailable, and even difficult to measure. Therefore, it may be practical to focus on new Deep Learning methods, which are powerful tools that can understand the complexity of the non-linearity relationship between inputs and outputs without having to resort to several parameters.

In this study, the authors used two different models RNN and LSTM on daily data from 5 catchments with a Mediterranean climate where the LSTM model showed better results for what was evaluated by the NSE. Other assessments were made on the LSTM model by RSR and PBIAS where the precipitation and antecedent flow being the parameters that most influenced the model.

Key words: modeling, hydrological, Deep Learning, LSTM, RNN.

Acknowledgment

We would like to thank Allah the Almighty for guiding us on this life path and for giving us health and the willingness to start and finish this thesis.

*We want to express our gratitude to our supervisor **Mr. Abdelmalek BERMAD**, for the guidance and orientation during the working process, and for his help, his trust, and advice.*

*We would like to acknowledge and give our warmest thanks to **Dr. Rafik OULEBSIR**, for his time, assistance, and supply of the essential methodological tools for the accomplishment of this work.*

We would like to thank the members of the jury for taking the time and patience to examine this work, as well as for all their comments and criticisms that will help improve our thesis.

Our deep thanks also go to all the members of our families who have been of great support in all the difficult moments. We also thank all our dear friends and classmates with whom we shared five (5) unforgettable years.

To all the people who have helped us reach the end of our journey, we offer our thanks, respect, and gratitude.

Dedication

This work is dedicated in the first place to my parents for standing by my side during this journey. To my dad an early polytechnic graduate himself, who cultivated in me a deep sense of curiosity developing my great interest in research, for his precious support, for being there for me despite all circumstances, for having a constant faith in me to achieve great things, a dad like no other. To my mom, who would get out of her way to help me. Them teaching me to never settle for less than what I deserve.

To my primary school teacher Hafida Abdessamed, may her soul rest in peace, the one who noticed the spark in me at a young age and whose memory keeps inspiring me until today.

To Mohamed, the kiddos of the family who outsmarts us all.

To Akram, Achraf my dear brothers living abroad, whom I'd loved to share my joy with today, for them to see what their little sister has become.

To my peer Rania, whose meeting has been such a blessing in my life, I couldn't have asked for a better partner.

To Aymen, for his constant support and presence in my life. Nesrine for being such an understanding roomie, Yousra, Wafa, and Halla for them campus times, we shared genuine laughter and sincere moments together.

To all the VIC family who marked my journey at school, Nabila, Iko, Akram the ones we look up to. Abdou, Abdallah, Asma, Joe the domino squad.

To all sincere people who accompanied me on this journey, I'm grateful.

Maria AMEDDAH

Dedication

Je dédie ce travail en premier lieu à mes chers parents qui m'ont soutenu tout au long de ma carrière, qui ont sacrifié pour faire de moi la personne que je suis aujourd'hui.

A ma chère maman, qui m'as encouragé dans chaque étape avec son amour, sa tendresse et surtout la grande confiance qu'elle a en moi. A mon cher père, mon idole, une grande école de vie qui m'inspire avec son expérience et ses connaissances ainsi que ses qualités et ses valeurs qui m'ont aidé à construire ma personnalité et à fixer mes objectifs.

A ma petite sœur SARAH, futur journaliste qui m'était toujours une grande source de motivation et d'énergie positive.

A mes deux petits frères YOUNES et ILYES.

A mon binôme et ma meilleure amie MARIA, que j'ai l'honneur d'avoir rencontré. Avec laquelle on est arrivé à réaliser cet honorable travail après avoir passé trois merveilleuses années ensemble.

A mes chers ami(e)s LARBI, ABDALLAH, YOUSRA et Wafa qui présentent une grande part de ma vie.

A mes ami(e)s polytechniciens que j'ai connu le long de mes années à l'école.

Rania MOKHTARI

Summary

List of tables

List of figures

List of abbreviations

General introduction	14
1. Chapter 1 Rainfall-Runoff modeling.....	17
1.1 <i>Why modeling?</i>	18
1.2 <i>The modeling processes</i>	21
1.2.1 The perceptual model.....	21
1.2.2 The conceptual model	21
1.2.3 The procedural model	21
1.2.4 Model calibration.....	21
1.2.5 Model validation	22
1.3 <i>Classification of models</i>	23
1.3.1 Model Structure.....	24
1.3.2 Spatial Processes.....	28
1.3.3 Mathematical models.....	30
1.4 <i>Rainfall-Runoff models</i>	31
1.4.1 Evaluation and Selection of Models	31
1.5 <i>Conclusion</i>	32
2. Chapter 2 Deep Learning	33
2.1 <i>General definitions</i>	34
2.1.1 Artificial intelligence	34
2.1.2 Machine learning	34
2.1.3 Deep learning	34
2.1.4 Artificial neural networks.....	35
2.2 <i>How deep learning works</i>	35
2.2.1 Input layer.....	35
2.2.2 Hidden Layers	35
2.2.3 Output layer.....	36
2.3 <i>Deep Learning applications</i>	36
2.4 <i>Recurrent Neural Network</i>	37
2.4.1 Definition	37
2.4.2 Commonly used activation functions in RNN	38
2.5 <i>Long short-term memory (LSTM)</i>	39
2.5.1 Why is LSTM an upgraded version of RNN	39

2.5.2	Core concept.....	40
2.6	<i>Deep learning and hydrology</i>	42
2.6.1	Brief history of ANN-Hydrology applications	43
2.6.2	LSTM and rainfall-runoff modeling research	43
2.6.3	Model Implementation in Rainfall-runoff prediction.....	44
2.7	<i>Conclusion</i>	44
3.	Chapter 3 The Study Zone - geography and data	46
3.1	<i>Datasets choice criteria</i>	47
3.1.1	Geographical area: Köppen–Geiger climate classification system.....	47
3.1.2	Data availability:	49
3.1.3	Hydrometric station, meteorology stations, watershed	50
3.2	<i>Study cases - geography</i>	50
3.2.1	Watersheds chosen for the study	50
3.3	<i>Study cases: data</i>	64
3.4	<i>Conclusion</i>	71
4.	Chapter 04 Tools and Methodology	72
4.1	<i>Tools</i>	73
4.1.1	Jupyter	73
4.1.2	Anaconda	73
4.1.3	Python	73
4.2	<i>Notions</i>	75
4.2.1	Nan	75
4.2.2	Overfitting.....	75
4.2.3	Outliers	76
4.3	<i>Processing approach</i>	76
4.3.1	Phase 01 - Collecting data.....	76
4.3.2	Phase 02 - Treating and data cleansing.....	77
4.3.3	Phase 03 - Determining appropriate inputs/output	78
4.3.4	Phase 04 - Data preprocessing for the model.....	80
4.3.5	Phase 05 - Model architecture.....	81
4.3.6	Validation and performance Monitoring (Numerical and graphical):.....	84
4.4	<i>Methodology summary</i>	87
4.5	<i>Conclusion</i>	88
5.	Chapter 05 Results & discussion	89
5.1	<i>Statistical parameters for numerical and graphical performance</i>	90
5.2	<i>The model Inputs</i>	90
5.3	<i>Results</i>	91
5.3.1	RNN Model	91

5.3.2	LSTM model	93
5.3.3	Comparison between RNN and LSTM.....	96
5.4	<i>Evaluation recap for the 3 studies (LSTM)</i>	99
5.5	<i>Discussion</i>	101
5.6	<i>Conclusion</i>	102
	General conclusion	104
	Bibliography	107
	APPENDIX	114

List of tables

Table 1.1 : Comparison of the spatial structures in rainfall-runoff models	29
Table 3.1 : Köppen climate classification scheme symbols description table	48
Table 3.2 : Geographical coordinates of Duero's hydrometric station	52
Table 3.3 : Morphometric characteristics of the Douro River sub-catchment.....	53
Table 3.4 : Geographical coordinates of Turia's hydrometric station	54
Table 3.5 : Morphometric characteristics of the TURIA River sub-catchment.....	55
Table 3.6 : Geographical coordinates of the San Joaquin hydrometric station	56
Table 3.7 : Morphometric characteristics of the San Joaquin River watershed.....	57
Table 3.8 : The San Joaquin River basin's major floods.....	58
Table 3.9 : Geographical coordinates of the hydrometric station of Bouchegouf.....	60
Table 3.10 : Morphometric characteristics of Bouchegouf sub-catchment.....	61
Table 3.11 : Geographical coordinates of the hydrometric station of Zardezas.....	62
Table 3.12 : Morphometric characteristics of Zardezas sub-catchment.....	63
Table 3.13 : List of hydrometric stations	64
Table 3.14 Statistical description of streamflow data	64
Table 3.15 : Rainfall data of Station (1)	66
Table 3.16 : Rainfall data of Station (2).....	67
Table 3.17 : Rainfall data of Station (3).....	67
Table 3.18 : Rainfall data of Station (4)	69
Table 3.19 : Rainfall data of Station (5).....	69
Table 4.1 model performance ratings.....	85
Table 5.1 : The model Inputs	91
Table 5.2: Statistical parameters of our results (RNN)	91
Table 5.3: Statistical parameters of our results (LSTM).....	94
Table 5.4 RMSE Values	98
Table 5.5 RSR Values.....	99
Table 5.6 Statistical parameters Study (01)	100
Table 5.7 Statistical parameters study (02)	100
Table 5.8 Statistical parameters study (03)	101

List of figures

Figure 1-1 Hydrological modeling	19
Figure 1-2 : Schema of the modeling process.....	23
Figure 1-3 : data use degree for each model type.....	24
Figure 1-4 : Artificial Neural Network.....	26
Figure 1-5 : Schematic diagram of Dawdy and O'Donnell's conceptual rainfall-runoff model	27
Figure 1-6 : From Lumped to distributed	28
Figure 2-1 : Artificial intelligence, Machine learning, Deep learning	35
Figure 2-2 : Deep Neural Network	36
Figure 2-3 RNN cell	37
Figure 2-4 : RNN vs ANN	38
Figure 2-5 : Comparison of the 3 main activation functions for RNN	39
Figure 2-6 : LSTM and GRU internal gates	40
Figure 2-7 : LSTM Cell	42
Figure 3-1 : Word map of Koppen-Geiger Climate Classification	47
Figure 3-2 : GRDC website	49
Figure 3-3 : NCEI Website.....	50
Figure 3-4 : Location map illustrating the Duero River watershed.....	51
Figure 3-5 : Geographical localization of the rainfall stations within the sub-catchment	52
Figure 3-6 : Location map illustrating the TURIA River sub-catchment (Haro et al. 2014)	54
Figure 3-7 : Geographical localization of the rainfall station within the sub-catchment	55
Figure 3-8 : Map of the San Joaquin River watershed	56
Figure 3-9 : Geographical localization of the rainfall stations within the watershed .	57
Figure 3-10 : Floodwaters surround a farm along the San Joaquin River in Sacramento County, January 21, 1969.	58
Figure 3-11 : Map of the hydrographic network of the Seybouse watershed.....	59
Figure 3-12 : Map of the hydrographic network of the Bouchegouf sub-catchment..	60
Figure 3-13 : Map of the hydrographic network of the Saf-Saf watershed.....	62
Figure 3-14 : Map of the hydrographic network of the Zardezas sub-catchment	63
Figure 3-15 : Streamflow distribution of the 5 stations	65
Figure 3-16 : Rainfall data distribution Station (S1).....	66
Figure 3-17 : Rainfall data distribution of S2.....	67
Figure 3-18 : Rainfall data distribution of S3.....	68
Figure 3-19 : Rainfall data distribution of S4.....	69
Figure 3-20 : Rainfall data distribution of S5	70
Figure 3-21 : Example of maximum temperature data of station (3).....	71

Figure 4-1 : Comparison between fitting types in different machine learning algorithms	76
Chapter 04 Tools and Methodology	
Figure 4-2 : Example of missing data treatment in streamflow	77
Figure 4-3 Example of treating outliers of maximum temperature	77
Figure 4-4 Rainfall-runoff plots of S1.....	78
Figure 4-5 Methodology summary	87
Figure 5-1 : Comparison of the predicted values and the observed values using ETP and precipitation inputs for the 5 regions during the test period of the RNN model. S1(1) , S2(1) , S3(1) , S4(1) , and S5(1) are the predicted and observed values for S1 to S5. S1(2) , S2(2) , S3(2) , S4(2) , and S5(2) are the scatter plots with the trendline of the predicted and observed values for S1 to S5.	93
Figure 5-2: Comparison of the predicted values and the observed values using ETP and precipitation inputs for the 5 regions during the test period of the LSTM model. S1(1) , S2(1) , S3(1) , S4(1) , and S5(1) are the predicted and observed values for S1 to S5. S1(2) , S2(2) , S3(2) , S4(2) , and S5(2) are the scatter plots with the trendline of the predicted and observed values for S1 to S5.	95
Figure 5-3 Comparison of RNN and LSTM (Study 01) - Figure 5-4 Comparison of RNN and LSTM (Study 02)	96
Figure 5-5 Comparison of RNN and LSTM (Study 03)	97

List of abbreviations:

LSTM: Long short-term memory

RNN: Recurrent neural network

NaN: Not a number

ANRH : Agence National des Ressources Hydriques

RMSE: Root Mean Squared Error

GRDC: Global Runoff Data Center

General introduction

Water is a vital resource for existing on earth. It is essential to human daily life, irrigation, and drinking water. The need to quantify it and monitor this resource became essential.

Rainfall-runoff models came in this context of quantifying. They are the type of equations that represent the water balance and the transfer equations of the water flows involved in the hydrological cycle at the catchment area scale. These models anticipate runoff in watercourses as well as other soil-atmosphere interactions like evapotranspiration, infiltration, and percolation. They are especially important in decision-making for integrated water resource management, particularly in the analysis of hydrological risk and flood forecasting. (Moulahoum 2019)

With rising demands on water resources throughout the world, it is becoming increasingly critical to plan, design, and manage water resources systems carefully and intelligently using efficient models. Especially that the mentioned rainfall-runoff modeling is complicated due to numerous complex interactions and feedback in the water cycle among precipitation and evapotranspiration processes, as also geophysical characteristics. Consequently, the lack of geophysical characteristics such as soil properties leads to difficulties in developing physical and analytical models when traditional statistical methods cannot simulate rainfall-runoff accurately. (Van et al., 2020)

This thesis falls in the concept of making improvements in the rainfall-runoff modeling by utilizing nowadays modern technology which essentially doesn't require involving complex and hard-to-get data parameters. It serves in understanding and reducing the complexity of various hydrological processes.

The main objective of this study is to develop and implement a methodological and practical approach that uses deep learning models (RNN & LSTM) to generate rainfall-runoff models of Mediterranean climate regions with the aim of managing the floods and the hydrological risks by predicting the runoff volume on a short term. This study relied on different meteorological parameters in the selected catchment areas located in the USA, Spain, and Algeria.

The present report is structured into five (5) chapters.

The first chapter presents generalities about rainfall-runoff modeling, including definitions, concepts, and the different steps to build the model.

The second chapter introduces deep learning notions with their history to use in hydrology.

The third chapter presents the study zone including the geography of the chosen catchment areas and the datasets used.

The fourth chapter explains the tools and methods used for our study.

The fifth chapter summarizes and discusses the obtained results.

Chapter 1

Rainfall-Runoff modeling

1.1 Why modeling?

Global advances in economies and standards of living have resulted in a growing dependency on water resources. Many societies have experienced water scarcity as a result of current patterns with societal advances; these are associated with factors such as (Daniel 2011)

- Population growth
- Increased urbanization and industrialization
- Increased energy use
- Increased irrigation associated with advances in agriculture productivity
- Desertification
- Global warming
- Poor water quality

Due to the limitations of hydrological measurement techniques, we are not able to measure everything we would like to know about hydrological systems. We have, in fact, only a limited range of measurement in space and time with limited techniques.

To predict the anticipated impact of future hydrological change, we require a method of extrapolating from known data in both space and time, particularly to ungauged catchments (where measurements are not accessible) and into the future (where measurements are not possible). Modeling the rainfall-runoff processes of hydrology in different types provides a means of quantitative extrapolation or prediction that will hopefully be helpful in decision making.

In its global sense, hydrological modeling concerns the simulation through a set of techniques and equations representing the water balance during the hydrological cycle at the scale of the catchment area, of all the hydrological processes, which define the behavior of a catchment area.

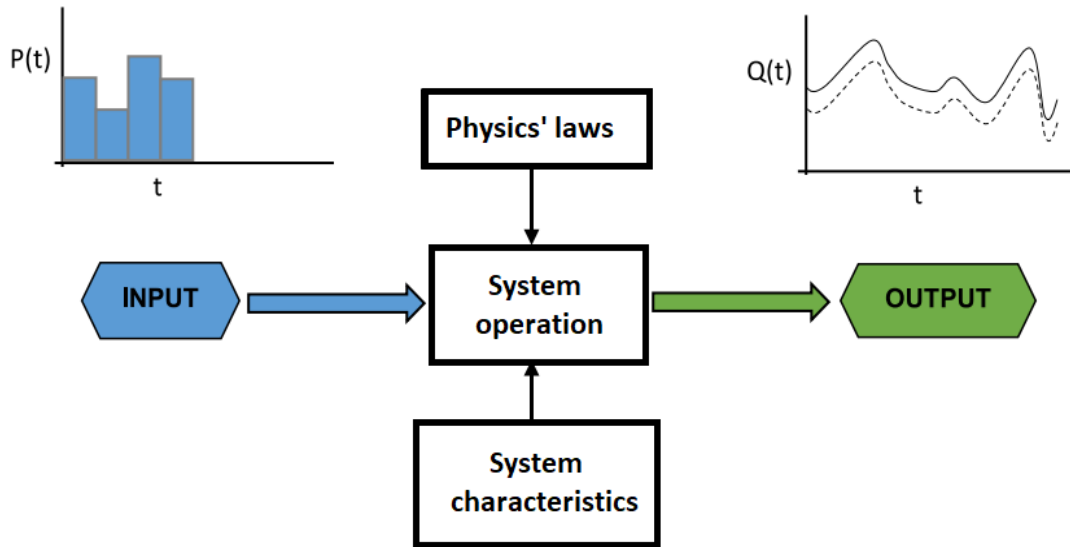


Figure 1-1 Hydrological modeling

Runoff modeling aids in the comprehension of hydrologic processes and how changes affect the hydrological cycle. Runoff models show what happens in water systems because of changes in previous surfaces, vegetation, and weather. As a function of numerous characteristics used to represent the watershed, a runoff model is defined as a collection of equations that aid in the estimation of the quantity of rainfall that converts into a runoff.

The application of rainfall-runoff models covers a variety of fields and applications in hydrologic research. These models are used for:

- Water resources planning.
- Flood forecasting.
- Investigation of the quality of natural waters.
- Extending time series of in-stream flow through space and time.
- Calculation of project floods.

Hydrological models can be grouped according to the modeling approach used. They can also be classified according to the nature of the algorithm used (empirical, conceptual, and physical). The method being deterministic or stochastic also differentiates the models according to the approach of the input or the specification of the parameters. The spatialization of the model between global, semi-distributed, and

distributed varies according to the size of the catchment area and the accuracy of the desired simulation.

In general, modeling is difficult due to a variety of uncertainties that are inadvertently conveyed into model output. These uncertainties are caused by a variety of factors, the most common of which are: (Wagener et al., 2004)

- Data uncertainty, which is caused by measurement errors or data pre-processing flaws.
- Model specification uncertainty, i.e., the inability to converge to a single ‘best’ parameter set (model) using the information provided by the available data. This is often referred to as the identifiability problem.
- Model structural uncertainty introduced through simplifications and/or inadequacies in the description of real-world processes.

Furthermore, even if those uncertainties could be eliminated, the natural processes themselves would still contain some (unmeasurable) randomness. Uncertainty is introduced by this randomness, which cannot be eliminated.

Moreover, modeling necessitates a good knowledge and understanding of the water cycle. Furthermore, it takes time and hard work to create these models. It also requires comprehensive soil profiles of study regions, which are impossible to obtain using existing survey and remote sensing approaches.

Data-driven methods, on the other hand, are frequently less expensive, more accurate, precise, and, most importantly, more flexible.

Because of its great ability to handle nonlinear and non-stationary problems, artificial neural networks (ANN) have become frequently used in water resource evaluations in recent years. Hydrological and hydraulic variables such as rainfall and runoff, as well as sediment loads, have been successfully simulated and predicted using ANN designs. ANN outperformed traditional statistical modeling techniques in many circumstances, and it has been used to forecast rainfall and runoff. As a result, in the next chapters, we will employ the ANN technique to develop our model.(Van et al. 2020)

1.2 The modeling processes

1.2.1 The perceptual model

The rainfall-runoff modeling process is composed of several essential steps that we will briefly explain.

In the first step, we establish our perceptual model, which represents the summary of the perceptions of how the catchment reacts to rainfall under various situations. A perceptual model differs from one person to another depending on the hydrologist's knowledge and analysis, his data sets, and experiences. All mathematical descriptions used to make predictions will be simplifications of the perceptual model. Therefore, it is very important to have a good understanding of this phase.

1.2.2 The conceptual model

At this point, the hypotheses and assumptions that have been made to simplify the description of the processes must be stated explicitly. Many models, for example, are based on Darcy's law, which states that flow is proportional to a gradient of hydraulic potential. Because measurements show that hydraulic potential gradients in structured soils can vary significantly over short distances, it is implicitly assumed that if Darcy's law is applied at the scale of a soil profile or larger, some average gradient can be used to characterize the flow and that the effects of preferential flow through macro pores in the soil can be ignored. It is therefore important to choose and establish wisely the assumptions and equations for the model calculation.

1.2.3 The procedural model

It represents the preparation of the code to be launched in the computing machine in order to obtain quantitative forecasts for a particular catchment. A stage of calibration of the model's parameters must be completed before proceeding to this level.

1.2.4 Model calibration

This step is essential to adjust the procedural model to the behavior of the catchment previously recorded and measured. Any hydrological model consists of

several state and input parameters. The idea is to find the best set of parameters by comparing the model output to real field measurements.

1.2.5 Model validation

After specifying the model parameter values, a simulation can be performed and quantitative predictions regarding the response obtained. The validation or evaluation of those predictions is the next step. This evaluation can also be done quantitatively, by generating one or more indices of the model's performance about the available (if any) runoff response observations.

This modeling technique, which appears to be simple, is filled with complications. As previously stated, the modeler's subjectivity in the perception process causes uncertainty in the user's choice of model.

As a result, a huge number of models that can be used have been observed. Another challenge comes from the availability of numerous combinations of model structures and parameter sets that provide an acceptable fit to measured flows in the scenario where the choice is important. As a result, distinguishing between the several alternative models and consequently validating a model in flow prediction research is difficult.

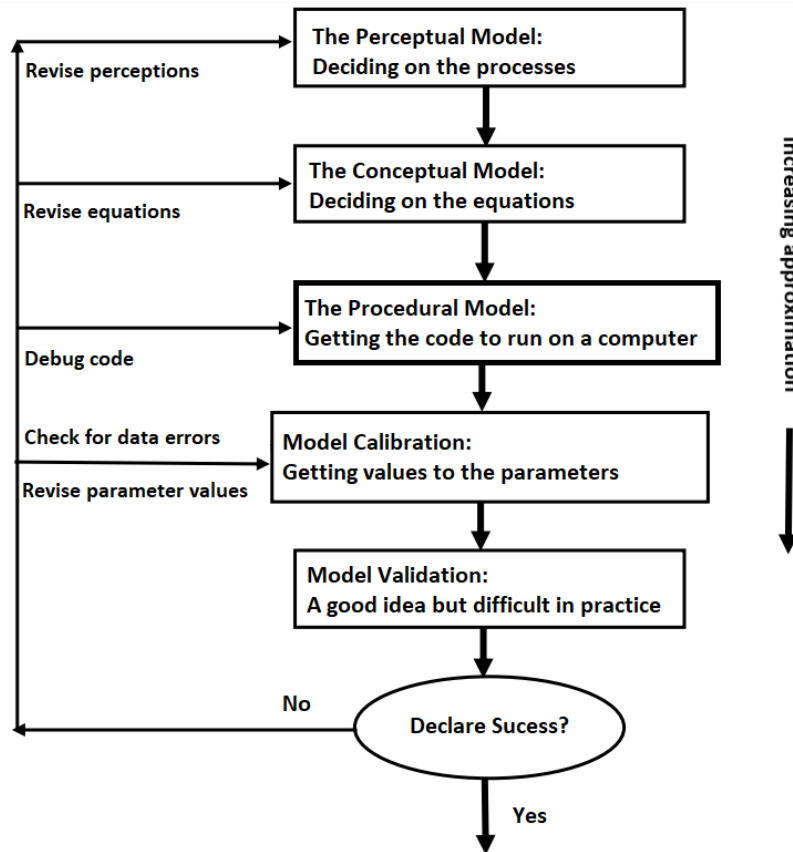


Figure 1-2 : Schema of the modeling process (Beven 2011)

1.3 Classification of models

The hydrological (rainfall-runoff) models were classified into several categories based on the modeling approach and the modeler's needs and goals, such as understanding and answering specific questions about the hydrological process, measuring the frequency of runoff occurrences, and predicting runoff yield for management purposes.

In this study, we divide models into three categories, each of which estimates runoff differently. The categories are empirical, conceptual, and physical, according to the model structure. Models are defined and divided in a variety of ways by researchers, including spatial resolution, input/output type, model simplicity, mathematical method, and so on. This section presents two additional classifications, the first of which is based on a geographical interpretation of the catchment region of the model. Models are classified as lumped, semi-distributed, or distributed. The second is based on a mathematical approach that distinguishes between deterministic and stochastic models.

1.3.1 Model Structure

The structure of a model determines how runoff is evaluated. Some can be used with a small number of variables, while others require a large number of interconnected variables. The governing equations determine the model structure, which ranges from simple to complicate. The following models are presented in order of increasing complexity, with empirical models being the simplest and physical mechanistic models being the most complex. Physical and conceptual models require a full understanding of the physics involved in the hydrological cycle's surface water movement.

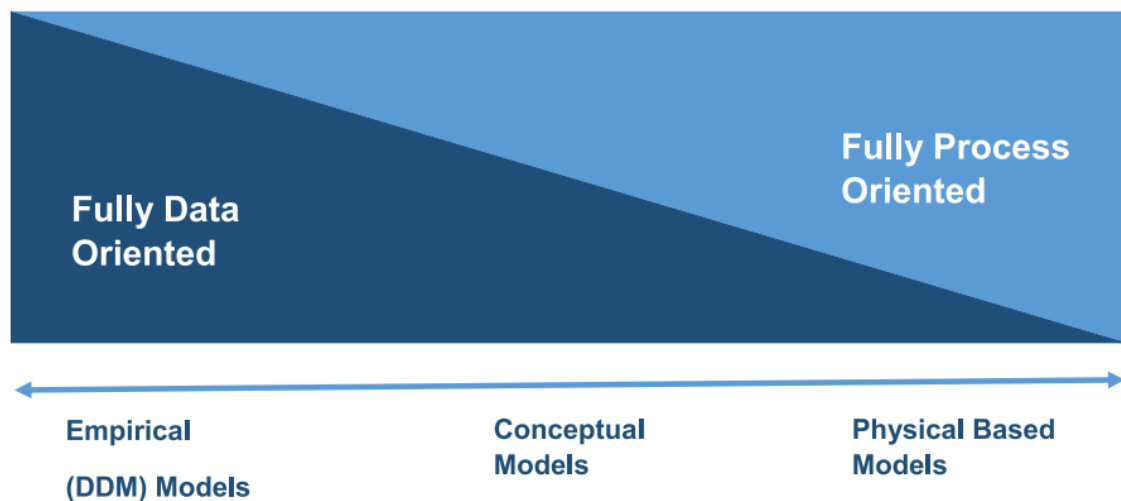


Figure 1-3 : data use degree for each model type (Moulaoum 2019)

1.3.1.1 Empirical model

These are observation-oriented models that merely employ data from existing sources without taking into account the hydrological system's characteristics and processes; hence, these models are also known as data-driven models. Rather than actual catchment processes, it uses mathematical equations derived from simultaneous input and output time series. These models are only applicable within the boundaries. The unit hydrograph exemplifies this method. In statistical approaches, regression and correlation models are used to determine the functional relationship between inputs and outputs. Simple rainfall-runoff regression models use rainfall and historical runoff as inputs, with runoff at a specific location as the output.

Some of the machine learning techniques utilized in hydro informatics methodologies include artificial neural networks and fuzzy regression.

Some of the machine learning techniques utilized in hydro informatics methodologies include artificial neural networks and fuzzy regression. The functions that convert rainfall to runoff are either an unknown mechanism (as in machine learning) or one that is not related to physical processes (as in the curve number method)

Empirical runoff models are best employed when no other outputs are required; for example, this model type cannot determine the distribution of runoff values between upstream and downstream areas. Due to a lack of particular knowledge about the watershed, ungauged watersheds are best represented using an empirical method.

Some cases of empirical models are the SCS-Curve Number used in SWAT2, regression equations, and machine learning used by Artificial and Deep Neural Networks. The machine learning techniques employ data-driven artificial neural networks that self-train to learn rainfall-runoff connection behaviors. Neural Networks employs machine learning to anticipate output based on data learned during the training period.

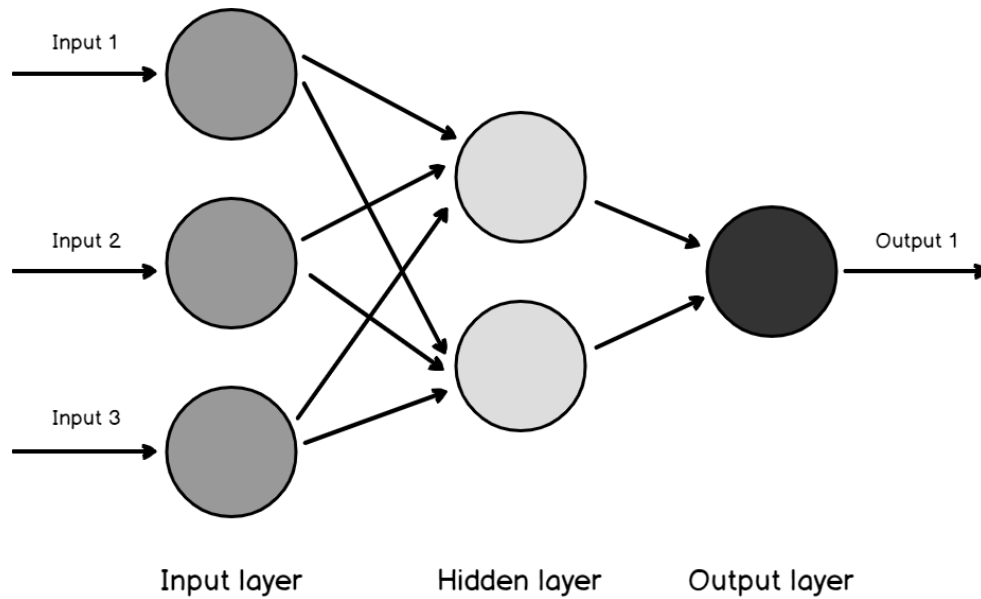


Figure 1-4 : Artificial Neural Network (Asanka 2020)

1.3.1.2 Conceptual models

Conceptual models are used to analyze runoff processes, which connect simplified components in the complete hydrological process. They are based on reservoir storage and simplified equations of physical hydrological processes that provide an overall picture of how a watershed operates.

The water balance equation is represented in conceptual models by the conversion of rainfall to runoff, evapotranspiration, and groundwater. The mathematical formulae that distribute the precipitation input data estimate each factor in the water balance equation (Sitterson et al. 2017). The general governing equations for conceptual models, which are versions of the water balance equation shown below, manage surface water and storage fluctuations.

$$\frac{dS}{dt} = P - ET - Q_s \pm GW \quad \text{Equation 1.1}$$

Where Ds/dt is the change in reservoir storage, P is precipitation, ET is evapotranspiration, Q_s is surface runoff, and GW is groundwater.

The complexity of conceptual models is affected by the intricacy of the balance equations used to describe hydrological components. Because of this volatility, these models necessitate a wide variety of parameters and meteorological input data. Another disadvantage is the absence of physical meaning in regulating equations and

parameters. Nonetheless, conceptual models are the most common since they are easy to use and calibrate.

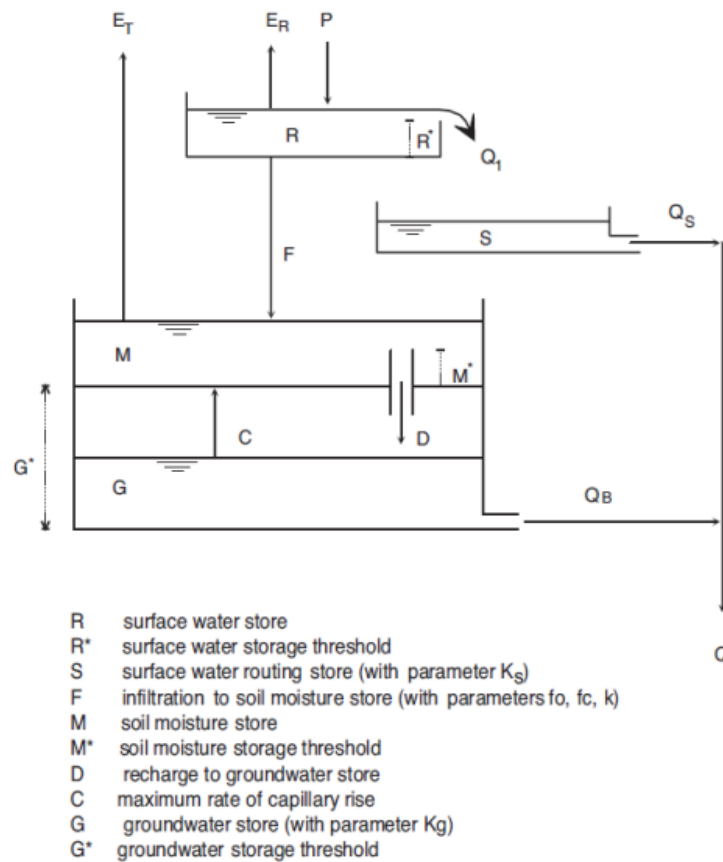


Figure 1-5 : Schematic diagram of Dawdy and O'Donnell's conceptual rainfall-runoff model (Beven, 2012)

1.3.1.3 Physical models

Physical models are based on physics information related to hydrological processes. The model is governed by physical equations that represent various features of the catchment's hydrologic reactions.

Water balance equations, conservation of mass and energy, momentum, and kinematics are among the general physics laws and concepts employed. Physical models use equations such as Saint Venant's, Boussinesq's, Darcy's, and Richard's. (2011) (Pechlivanidis et al.)

The physical model has a logical framework that is comparable to the real-world system, which is reinforced by the existing link between model parameters and physical catchment features. They are most effective when accurate data is available, the physical features of hydrological processes are well known, and they are employed

at tiny scales owing to computing time. A large number of physical and process factors are necessary to calibrate the model. Physical parameters are catchment qualities that can be measured; process parameters, on the other hand, are physical features that cannot be defined, such as average water storage capacity.

Because of the massive amounts of data necessary to run these models, over-parameterization occurs, resulting in vast computations that are incompatible with flood forecasts, even with current technical breakthroughs in calculating machinery.

As a result, rather than flood forecasting, physics-based models are used in solid transport and pollutant dissemination problems. These models may also be used to simulate groundwater circulation and interactions in the watershed with sediments, nutrients, and contaminants.

1.3.2 Spatial Processes

The geographical distribution of variables and factors involved in a catchment's behavior may be used to classify rainfall-runoff models (hydrological models). The lumped models are then distinguished from the semi-distributed or fully distributed models.

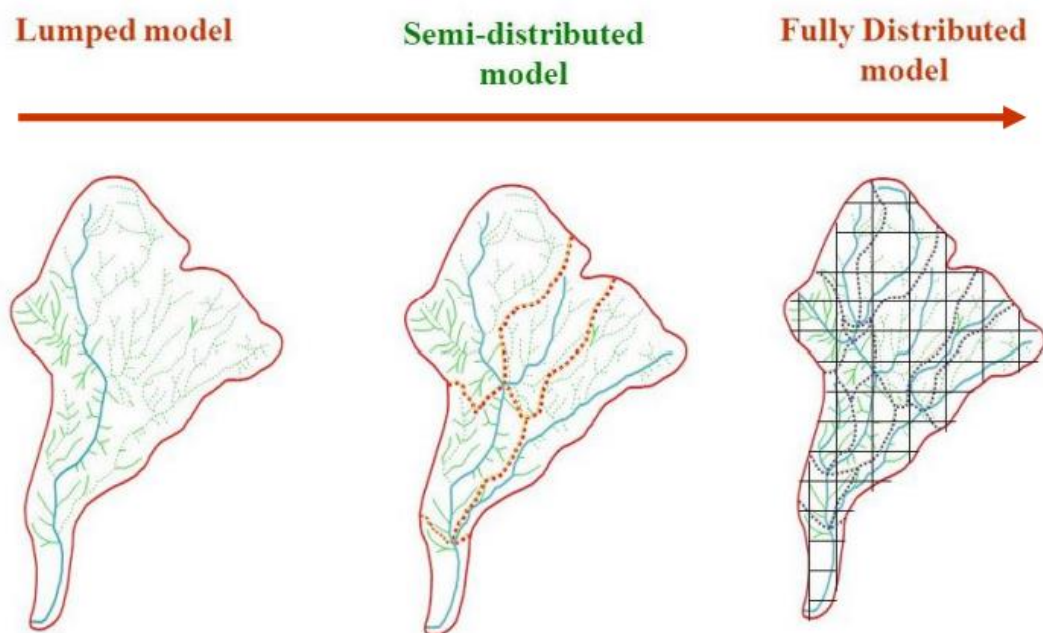


Figure 1-6 : From Lumped to distributed (Deltares USA)

Table 1.1 : Comparison of the spatial structures in rainfall-runoff models (Sitterson et al., 2017)

	Lumped	Semi-Distributed	Distributed
Method	Spatial variability is disregarded; entire catchment is modeled as one unit	Series of lumped and distributed parameters	Spatial variability is accounted for
Inputs	All averaged data by catchment	Both averaged and specific data by sub-catchment	All specific data by cell
Strengths	Fast computational time, good at simulating average conditions	Represents important features in catchment	Physically related to hydrological processes
Weaknesses	A lot of assumptions, loss of spatial resolution, not ideal for large areas	Averages data into sub-catchment areas, loss of spatial resolution	Data intense, long computational time

1.3.2.1 *Lumped models*

The catchment region is represented as a single homogenous unit in lumped models. In lumped models, the spatial heterogeneity of catchment parameters is ignored. Uniform precipitation quantities are utilized as averaged values over the watershed. The catchment characteristics are all defined to be the same, resulting in over- or under-parameterization. In these models, a single runoff output value is calculated at the catchment area's river outflow point. For regulatory reasons that include long-term circumstances, average and yearly computed data are employed. Lumped models do not correctly portray huge watersheds and catchments due to their various assumptions and averaged conditions. (Hamid et al., 2008)

1.3.2.2 *Semi-distributed models*

Semi-distributed models are variations of lumped models, with features of distributed models. They can consist of a series of lumped parameters applied in a quasi-spatially distributed manner. The model process divides the catchment into smaller areas, with different parameters for each. (Rinsema 2014)

Sub-areas can be divided in many ways; by slope, soil group, vegetation zones, or a combination called Hydraulic Response Units (HRUs) in which the region within the

HRU responds to rainfall the same way, based on overlaying maps of land cover, soil group, and elevation (Beven 2011; Devia et al., 2015). Semi-distributed models calculate runoff at the pour point for each sub-catchment. (Ocio et al., 2019)

1.3.2.3 *Fully distributed models*

Small components or grid cells divide the model process in fully distributed models. (Ocio et al., 2019)

They are also organized in the manner of a physically based model, which makes them more analogous to the actual hydrologic process. Spatially distributed models have transformed management practices by giving precise data for tiny pieces. Each tiny element (or cell) has a specific hydrological response and is computed individually, although interactions with neighboring cells are taken into account (Rinsema 2014). The model calculates runoff for each grid cell and provides comprehensive runoff data at various sites within the catchment area based on physical equations used to predict flow direction and natural time delays.

Although this technique has been criticized for being unduly deterministic, the uncertainty is considerably decreased when model parameters are obtained from physical attributes using many linear regressions. While this method has shown to be effective in a variety of situations, the industry's adoption of distributed models is limited due to their relative complexity.(Ocio et al., 2019)

1.3.3 **Mathematical models**

Rainfall-Runoff models can be categorized according to the techniques introduced in the modeling process. They can be deterministic or stochastic.

1.3.3.1 *Deterministic models*

Deterministic models are mathematical models that generate results based on predefined connections between states and occurrences. Their behavior is completely predictable, and they allow for a single simulation output to be generated using the same inputs and parameters.

1.3.3.2 *Stochastic models*

A model is stochastic if it has random_statistical distribution parameters as inputs, and consequently, its outputs are random. (Leonelli et al., 2017)

Despite the fact that most models are deterministic, stochastic models provide two key benefits. For starters, their theoretically basic form enables them to show variability when geographical or temporal data are few. Second, they enable decision-makers to evaluate the level of uncertainty in forecasts.

1.4 Rainfall-Runoff models

The USGS in its review of rainfall-runoff modeling for flood management has selected nine widely known models to illustrate the most commonly used models in practice and their classification of the model types detailed above. (Knapp et al., 1991)

However, there are indeed a large number of hydrological models. This list has been selected as a guideline.

- HEC -1 Flood Hydrograph Package (U.S. Army Corps of Engineers, Hydrologic Engineering Center, 1990): Conceptual, event model with fitted and/or empirical, lumped parameters
- TR-20 (Soil Conservation Service Technical Release 20, 1978): Conceptual, event model with empirical, lumped parameters
- HSPF -- Hydrologic Simulation Program - FORTRAN (Johanson et al., 1984): Conceptual, CS model with fitted and/or physical, HRU parameters
- SWMM -- Storm Water Management Model (Huber et al., 1981): Conceptual, event model with fitted, HRU parameters
- ANSWERS -- Areal Nonpoint Source Watershed Environment Response Simulation (Beasley and Huggins, 1982): Conceptual, event model with fixed, distributed parameters
- SHE -- Systeme Hydrologique European (Abbott et al., 1986b): Hydrodynamic, CS model with physical, distributed parameters
- PRMS -- Precipitation-Runoff Modeling System (Leavesley et al., 1983): Conceptual, CS model with physical and fitted, HRU parameters
- NWS RFSFS -- National Weather Service River Forecast System (NWS, 1983) Conceptual, CS model with fitted, lumped parameters
- GR -- Rural Engineering Models: Global continuous/events conceptual model with calibrated parameters.

1.4.1 Evaluation and Selection of Models

The goal of the model application should be clearly defined before selecting a model or modeling approach.

It is important to note that each type of model has its field of effectiveness in rainfall-runoff modeling, and its application is dependent on the study's goal and desired accuracy. Several criteria are usually recommended for use in the model selection: (Knapp et al., 1991)

- Ease of model use,
- Accuracy,
- Consistency of parameters,
- Sensitivity of output to changes in parameters,
- Theoretical limitations of the model,
- Data limitations.

It is recommended that the models be chosen based on their suitability for resolving specific problems. The goal of the simulation should be determined first, and the technique selection should follow rather than lead this decision. However, connecting modeling aims to an appropriate model is not always easy.

1.5 Conclusion

In this chapter, we have seen in general the basics of modeling the relation between rainfall-runoff by explaining the main steps to establish and how to use a model. Moreover, we discussed the different types of these models and their categories. Later, by giving examples of rainfall-runoff models, we can observe that there is a wide range of techniques and technologies used in it.

In the next chapter, we will elaborate the method we have chosen to establish our model.

Chapter 2 Deep Learning

Introduction

Neural technologies continue to make significant progress in their quest to be acknowledged as tools that provide efficient and effective solutions for modeling and analyzing the behavior of complex dynamical systems. Time series forecasting has received special attention, and improved models have been reported in a variety of disciplines, including rainfall-runoff modeling. (Abrahart et al., 2007)

This chapter introduces deep learning notions with their history of use in hydrology.

2.1 General definitions

2.1.1 Artificial intelligence

It is a branch of study that refers to any approach that enables computers to mimic human behavior and replicate or outperform human decision-making to complete complicated tasks autonomously or with minimum human intervention.

2.1.2 Machine learning

It is a collection of approaches that let machines automatically learn patterns from data. In contrast to programming, which consists of following established rules.

2.1.3 Deep learning

It is a subtype of machine learning that consists of neural network layers which tend to imitate the human brain's functionalities by enabling it to "learn" from large datasets. While a single-layer neural network may produce approximate predictions, more hidden layers can assist to improve and tune for accuracy. (IBM Cloud Education 2022)

Deep learning models outperform shallow machine learning models and traditional data analysis methodologies in many areas.(Janiesch, et al. 2021)

2.1.4 Artificial neural networks

Artificial neural networks (ANNs) are made up of node layers, each of which has an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, is linked to another and has its weight and threshold. (IBM Cloud Education 2021)

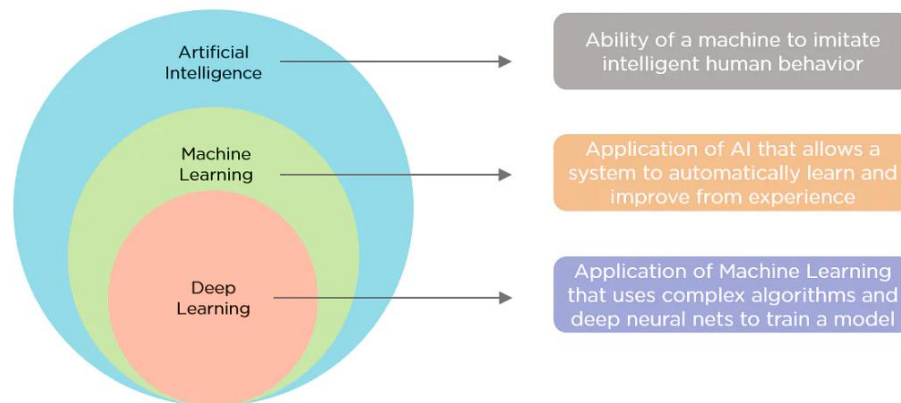


Figure 2-1 : Artificial intelligence, Machine learning, Deep learning (IBM Cloud Education 2022)

2.2 How deep learning works

Neural networks are layers of nodes, similar to how neurons make up the human brain. Individual layer nodes are linked to nodes in neighboring layers. The number of layers in the network indicates how deep it is. A single neuron gets hundreds of impulses from other neurons. Similarly, Signals pass between nodes in an artificial neural network and are assigned weights. (Reyes, 2022)

2.2.1 Input layer

It receives the observation's input data. This data is broken down into numbers and bits of binary data that a computer can interpret. To be within the comparable range, values must be standardized or normalized. (Data flair, 2020)

2.2.2 Hidden Layers

It computes mathematical functions on input data. Choosing the number of hidden layers and the number of neurons in each layer is difficult. It is responsible for non-linear processing units for feature extraction and transformation. Each

subsequent layer takes the output of the previous layer as input. It creates the learning hierarchy principles. Each level of the hierarchy grasps the ability to change the incoming data into a more abstract and composite representation. (Data flair, 2020)

2.2.3 Output layer

The layer is responsible for producing the final result.

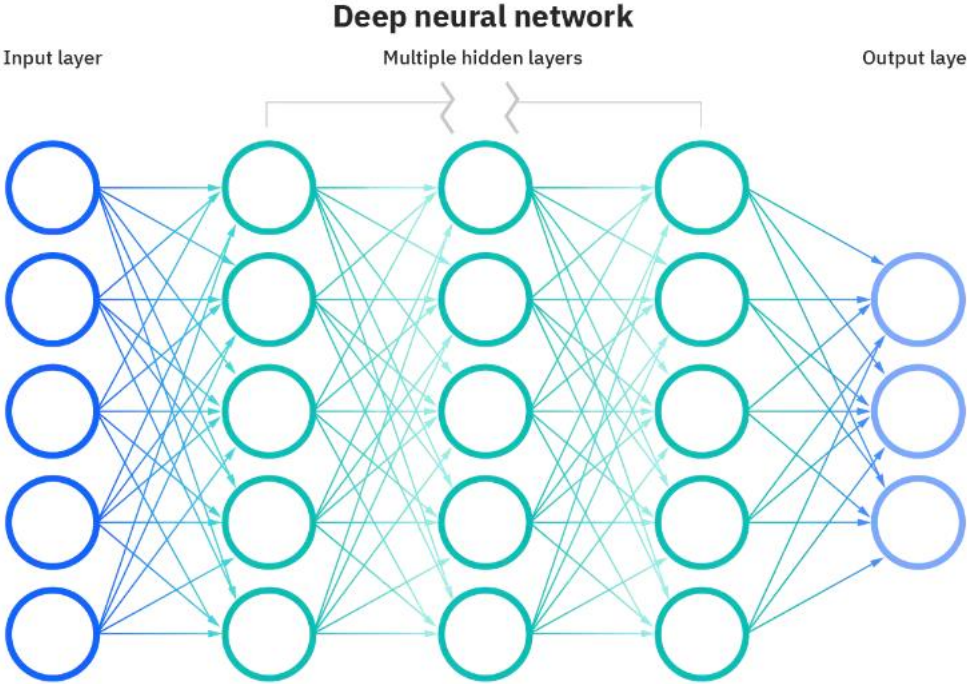


Figure 2-2 : Deep Neural Network (IBM, 2020)

2.3 Deep Learning applications

Deep learning has different architectures, this includes deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks... Each of these has its characteristics that can be used in different fields. We can state: computer vision, speech recognition, natural language processing, machine translation, and even more science-related fields such as bioinformatics, drug design, medical image analysis, climate research, and material inspection, generating results comparable to, and in some cases superior to, traditional approaches. (Deep learning 2022)

2.4 Recurrent Neural Network

2.4.1 Definition

Recurrent neural networks also abbreviated as RNN, are artificial neural networks that are extensively utilized in voice recognition and natural language processing and for our case of study: Time series data. RNNs identify the sequential properties of input and utilize patterns to forecast the next probable situation. (Pai 2020).

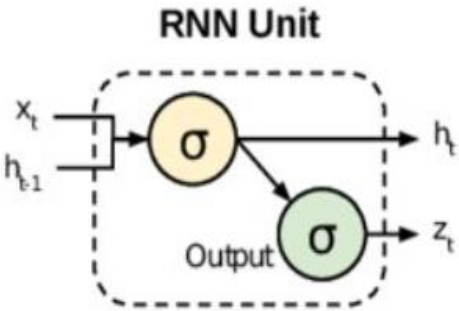


Figure 2-3 RNN cell (Donahue et al., 2014)

They feature an extra parameter matrix for connections between time steps in their structure, which improves training in the temporal domain and exploitation of the sequential character of the input. RNNs are trained to provide output in which predictions are made at each time step based on current input and knowledge from prior time steps. (Singh et al., 2022)

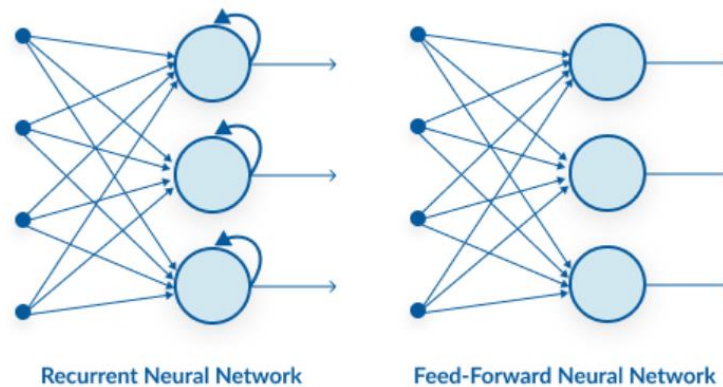


Figure 2-4 : RNN vs ANN (analyticsvidhya)

2.4.2 Commonly used activation functions in RNN

An activation function is used to produce or specify a specific output for a certain node based on the input. That is, the activation function will be applied to the summation results. There are several types of activation functions, including linear activation functions, heavy side activation functions, sigmoid functions, Tanh functions, and RELU activation functions.

Activation functions are a crucial element of any neural network in deep learning since they are capable of performing extremely complex and critical tasks such as object identification, picture categorization, language translation, and so on. We cannot envision performing these activities without the use of deep learning. (Goswami 2020)

2.4.2.1 Sigmoid function

The sigmoid function, also known as the logistic function, serves to standardize the outcome of any entry in the range of 0 to 1. The activation function's main aim is to keep the output or anticipated value within a specific range, which improves the model's efficiency and accuracy.

Range: 0 to 1

2.4.2.2 Tanh function

The primary distinction between the Tanh and Sigmoid activation functions lies in the range, the Tanh interval varies from -1 to 1. Rest functionality is identical to sigmoid functionality in that both may be utilized on a feed-forward network.

Negative values are also included here, whereas the sigmoid's minimum range is 0, and Tanh's minimum range is -1. As a result, the Tanh activation function is often referred to as the zero-centered activation function.

2.4.2.3 *Relu function*

ReLU is the best and most advanced activation function right now compared to the sigmoid and TanH because all the drawbacks like Vanishing Gradient Problem is completely removed in this activation function which makes this activation function more advanced compared to other activation functions.

Range: 0 to infinity

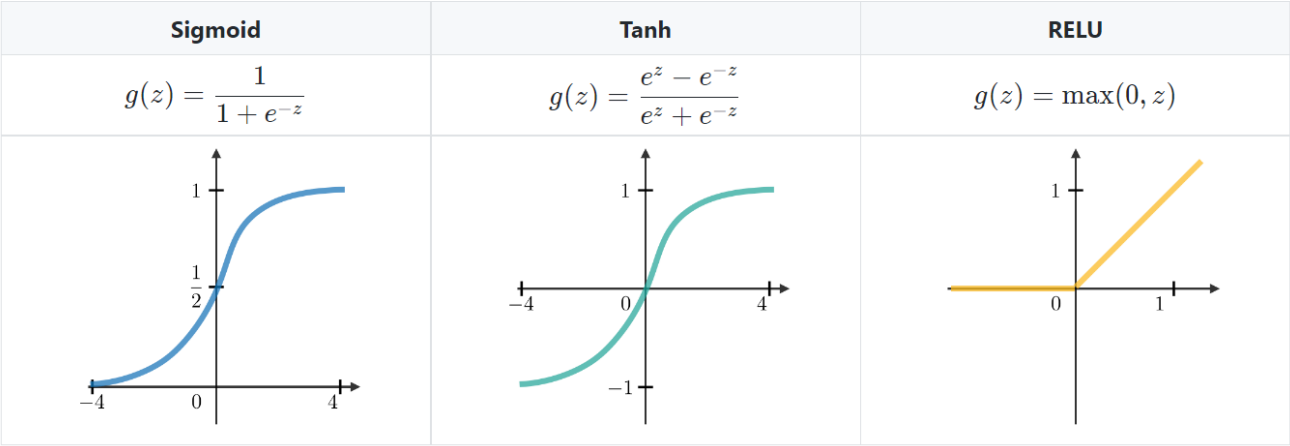


Figure 2-5 : Comparison of the 3 main activation functions for RNN (<https://stanford.edu>)

2.5 Long short-term memory (LSTM)

2.5.1 Why is LSTM an upgraded version of RNN

The vanishing gradient problem affects recurrent neural networks during backpropagation. Gradients are values that are used to update the weights of a neural network. The vanishing gradient problem occurs when the gradient declines as it propagates backward in time. When a gradient value becomes incredibly tiny, it does not contribute much to learning. (Phi 2020)

As a solution to short-term memory, LSTMs and GRUs were developed. They feature internal systems known as gates that allow them to control the flow of information.

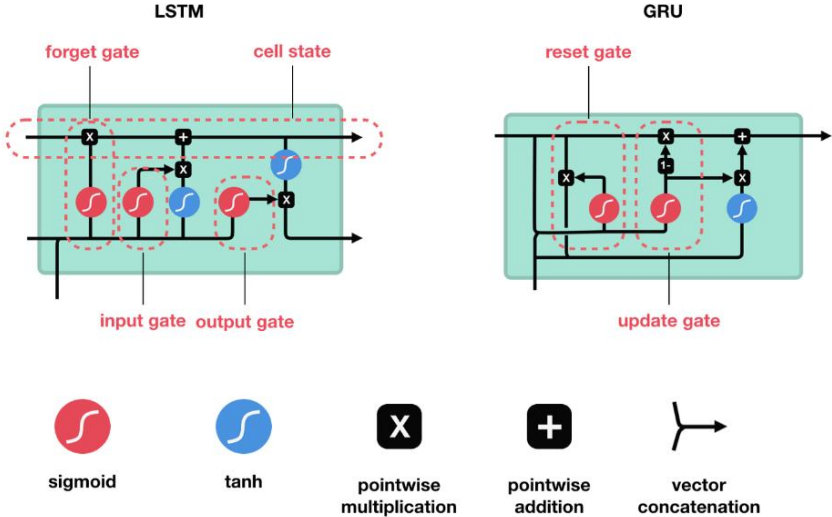


Figure 2-6 : LSTM and GRU internal gates (Towardsdatascience)

2.5.2 Core concept

The cell state and its multiple gates are central to LSTMs. The cell state serves as a transportation channel for relative information down the sequence chain. You might think of it as the network's "memory." In principle, the cell state can carry meaningful information throughout the sequence's processing. As a result, information from earlier time steps might travel to later time steps, diminishing the impact of short-term memory. As the cell state travels, information is added or deleted from the cell state via gates. The gates are several neural networks that determine whether information about the cell state is permitted. (Phi 2020)

To delve a bit more into what the various gates are doing. In an LSTM cell, we have three separate gates that control information flow. A forget gate, an input gate, and an output gate.

2.5.2.1 Forget gate

First, there's the forget gate. This gate determines whether information should be discarded or saved. The sigmoid function is used to process information from the current input $X(t)$ and the hidden state $h(t-1)$. The values range from 0 to 1. Closer to 0 implies that the value will be forgotten and closer to 1 means the value will be maintained.

2.5.2.2 *Input gate*

To update the cell state, the input gate conducts the following processes. First, the second sigmoid function receives the current state $X(t)$ and the previously concealed state $h(t-1)$. The values are changed between 0 (important) and 1 (unimportant). The exact information about the hidden and current states will then be supplied through the tanh function. To control the network, the tanh operator will generate a vector ($C(t)$) containing all possible values between -1 and 1. The activation function output values are set for point-by-point multiplication.

- Cell state:

The network has enough data from the forget gate and the input gate. The following step is to decide on and save the information from the new state in the cell state. $C(t-1)$ is multiplied by the forget vector $f(t)$ to recover the prior cell state. If the result is 0, the values in the cell state are removed. The network then takes the output value of the input vector $i(t)$ and conducts point-by-point addition, updating the cell state and providing the network with a new cell state $C(t)$.

2.5.2.3 *Output gate*

The value of the next hidden state is determined by the output gate. This state holds data from earlier inputs. First, the current and prior hidden state values are supplied into the third sigmoid function. The new cell state produced by the cell state is then transmitted to the tanh function. These two outputs are multiplied point by point. The network determines which information the hidden state should carry based on the final value. This hidden state is employed in prediction. Finally, the new cell state and hidden state are passed forward to the following time step. (Gaurav 2020)

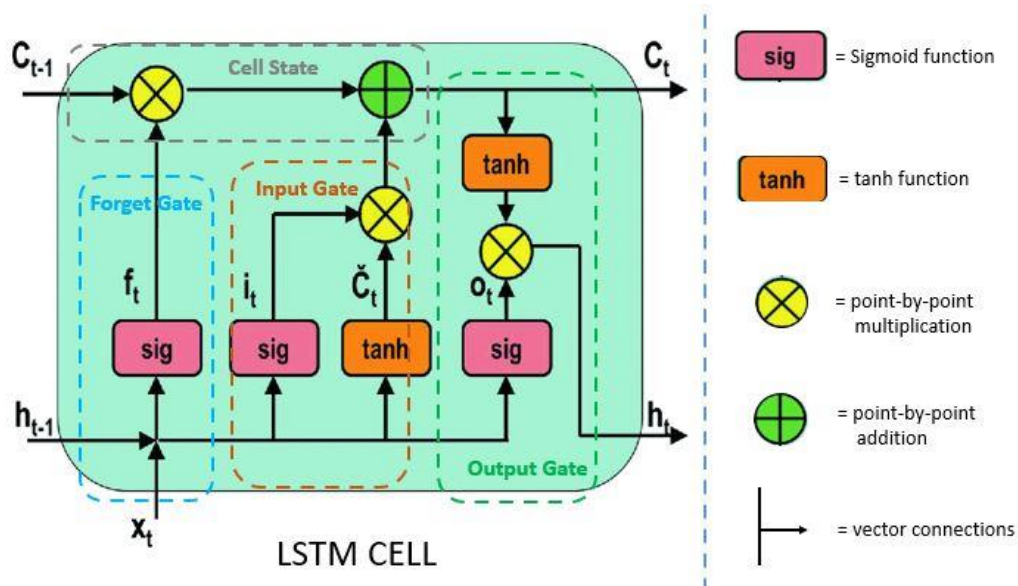


Figure 2-7 : LSTM Cell (Towards data science)

2.6 Deep learning and hydrology

Rainfall-runoff modeling is challenged by various complex interactions and feedback in the water cycle between precipitation and evapotranspiration processes, as well as geophysical properties. Runoff carries rainfall water to a catchment region that is geographically widespread, temporally changing, and non-linear (Tanty et al., 2015). As a result, the lack of these different parameters such as soil qualities makes it difficult to create physical and analytical models when typical statistical approaches cannot effectively predict rainfall-runoff.

The last decade has witnessed a virtual explosion of neural network (NN) modeling activities throughout the hydrological sciences. The use of the ANN algorithm in rainfall-runoff modeling is a recent advancement in the system-conceptual modeling approach. The benefit of the ANN approach is that it does not require a thorough knowledge of catchment features; it simply builds a link between input (rainfall, meteorological data) and output (runoff) based on learning during the neural network training process. It can capture the nonlinear relationship between prediction and predictors. Thus, while physical features are not evaluated individually, they are an essential element of the model. (Tanty et al., 2015)

2.6.1 Brief history of ANN-Hydrology applications

Various ANN designs have been used effectively to simulate and forecast hydrological and hydraulic variables such as rainfall, runoff, and sediment loads. In numerous instances, ANN outperformed traditional statistical modeling approaches. (Coulibaly *et al.*, 2000);(Dawson, Wilby 2001) ; (Sudheer *et al.*, 2002), This network has also been employed as an alternative for predicting rainfall–runoff. A feed-forward with three layers could primarily represent the rainfall-runoff process (Halff *et al.*, 1993) at first. Following the success of this model, several research has been conducted to apply other ANN architectures for rainfall-runoff prediction. (e.g., (MINNS, HALL 1996); (Shamseldin 1997) ; (de Vos, Rientjes 2005)). (Hsu *et al.*, 1995) suggested a linear least-squares simplex approach for training ANN models. The results demonstrated that the rainfall-runoff connections were better represented than in previous time series models. (Mason *et al.*, 1996) employed a radial basis function network for rainfall-runoff modeling, which allows quicker training than the traditional back-propagation approach. (Birikundavyi *et al.*, 2002), again, examined ANN models for daily streamflow forecasting and came to the conclusion that ANN outperforms other models such as deterministic models and traditional autoregressive models. (Toth, Brath 2007) and figured that ANN is an effective instrument for continuous period rainfall- rainfall, assuming that a large collection of hydro-meteorological data is available for calibration. (Bai *et al.*, 2016) using deep belief networks, he anticipated daily reservoir inflows. (DBNs).

Most of the experiments described above relied on a kind of ANN known as a multilayer feed-forward neural network (FNN), with only a few studies using recurrent neural networks (RNNs). Even though FNN offers various advantages in modeling statistical data, there are still significant challenges, such as the selection of optimum neural network parameters and the overfitting problem. As a result, the performance of ANN forecasts is also heavily influenced by the user's prior experience. (Dawson, Wilby 2001; de Vos, Rientjes 2005) (Van *et al.*, 2020)

2.6.2 LSTM and rainfall-runoff modeling research

The present LSTM design comprises multiple gates with various functions to govern neurons and maintain information. LSTM memory cells can store significant information for a longer period of time. (Gers, *et al.* 2000). This information-holding

capability enables LSTM to perform effectively while processing or forecasting a complicated dynamic sequence. (Hu et al. 2018) propose LSTM deep learning for rainfall-runoff modeling and rainfall-runoff ANN and LSTM are both appropriate for rainfall-runoff from conceptual and physical-based models. (Kratzert, et al. 2018) utilized LSTM to estimate rainfall-runoff, demonstrating its potential as a regional hydrological model wherein one model predicts flow for a range of catchments.

Several more research have demonstrated that LSTM outperforms the Hidden Markov Model and other RNNs in terms of capturing long-range relationships and nonlinear dynamics. (Baccouche, et al. 2011) (Graves, Jaitly 2014) (Van, et al. 2020)

2.6.3 Model Implementation in Rainfall-runoff prediction

When using neural networks for rainfall runoff models several decisions must be taken. First, a suitable neural network type must be chosen. Second, one must pick an adequate training method, proper training intervals, and an appropriate network layout. Finally, select how to pre- and post-process input-output While some of these procedures may be automated by making suitable changes to training algorithms, many judgments should still be made by trial and error. (Dawson et al., 2001)

The number of hidden layers and neurons in each hidden layer must be specified by the neurophysiologist. If the hidden layers have too few neurons, the network may be unable to explain the underlying function because it lacks enough parameters (or 'degrees of freedom') to map all locations in the training data. In contrast, if there are too many neurons, the network has too many free parameters and may overfit the data, resulting in a loss of generalizability. Furthermore, an 'excessive' amount of hidden neurons might stymie the training process to the point that a network takes an abnormally long period to learn. (Dawson, et al. 2001)

More specific details about data handling and our model implementation will be discussed in Chapters 04 and 05.

2.7 Conclusion

As shown in this chapter, hydrological modeling using deep learning (more specifically rainfall-runoff modeling) has shown great results over time, especially when involving recurrent neural networks with their improved version. For that, we

have chosen to work with this technique in our thesis to check its accuracy in different datasets across the Mediterranean climate regions.

Chapter 3

The Study Zone - geography and data

Introduction

Data is a key factor in Deep learning. For that, the models’ accuracy tends to increase with the increasing amount of training datasets. Hence, this chapter introduces the catchment areas chosen for the study, the dataset, its origins, and the way it has been processed to prepare it for modeling.

3.1 Datasets choice criteria

In choosing our datasets, we relied on the key factors listed below:

3.1.1 Geographical area: Köppen–Geiger climate classification system

Wanting to implement our model in different parts of Algeria and due to the lack of unreliable data in the region, we decided to choose regions with similar climate characteristics. For that, we relied on the Köppen–Geiger climate classification system.

The system categorizes the world into five climatic zones based on characteristics such as temperature, allowing for distinct vegetation growth. Köppen's map used various colors and shades to represent the world's climate zones. (*National Geographic Society*)

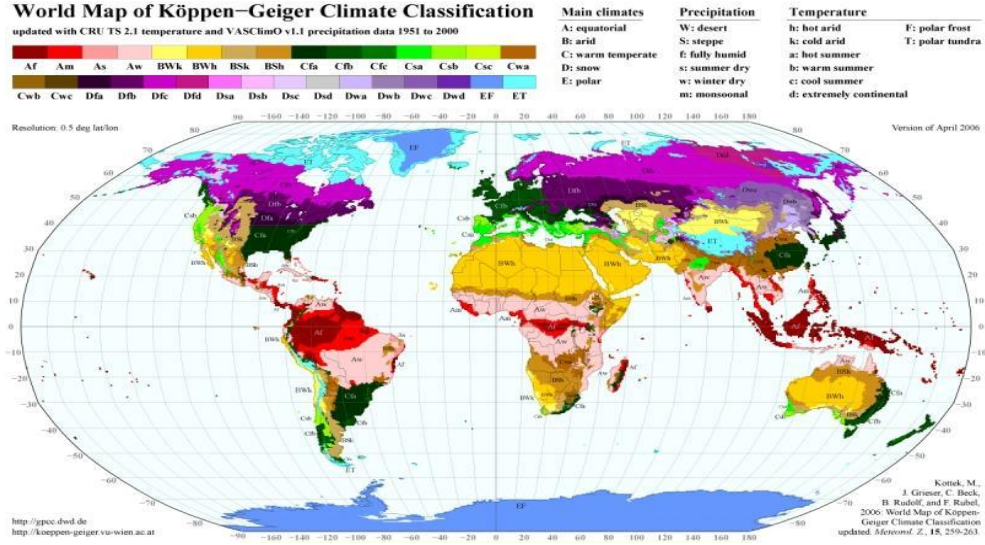


Figure 3-1 : World map of Köppen-Geiger Climate Classification (Kottek et al. 2006)

From Figure 3-1 we can notice that these regions exist in the northern part of Algeria, west side of USA, southern part of Europe, Australia and south America. Regions with a Mediterranean climate are referred to as: Csa, Csb, Csc accordingly, each one of these has its own climatic characteristics:

- **Csa = Hot-summer Mediterranean climate:** coldest month averaging above 0 °C (or -3 °C), at least one month's average temperature above 22 °C, and at least four months averaging above 10 °C
- **Csb = Warm-summer Mediterranean climate:** coldest month averaging above 0 °C (or -3 °C), all months with average temperatures below 22 °C, and at least four months averaging above 10 °C
- **Csc = Cold-summer Mediterranean climate:** coldest month averaging above 0 °C (or -3 °C) and 1–3 months averaging above 10 °C.

Table 3.1 : Köppen climate classification scheme symbols description table (Köppen climate classification - Wikipedia)

1st	2nd	3rd
A (Tropical)	f (Rainforest)	
	m (Monsoon)	
	w (Savanna, Dry winter)	
	s (Savanna, Dry summer)	
B (Arid)	W (Desert)	
	S (Steppe)	
		h (Hot)
		k (Cold)
C (Temperate)	w (Dry winter)	
	f (No dry season)	
	s (Dry summer)	
		a (Hot summer)
		b (Warm summer)
		c (Cold summer)
D (Continental)	w (Dry winter)	
	f (No dry season)	
	s (Dry summer)	
		a (Hot summer)
		b (Warm summer)
		c (Cold summer)
		d (Very cold winter)
E (Polar)	T (Tundra)	
	F (Eternal frost (ice cap))	

3.1.2 Data availability:

For high accuracy of the results, it is important to work with a dataset that is both large and reliable. Thus, we picked our stations carefully to satisfy both conditions, identifying the statistics of each dataset in every suitable region. We mainly relied on these two well-known websites to extract the information:

- **Global Runoff Data Centre (GRDC)**

The GRDC is an international archive of data up to 200 years old and fosters multinational and global long-term hydrological studies. Originally established three decades ago, the GRDC aims to help earth scientists analyze global climate trends and assess environmental impacts and risks. Operating under the auspices of WMO the database of quality controlled “historical” mean daily and monthly discharge data grows steadily and currently comprises river discharge data of well over 10,000 stations from 159 countries. (BfG - The GRDC)

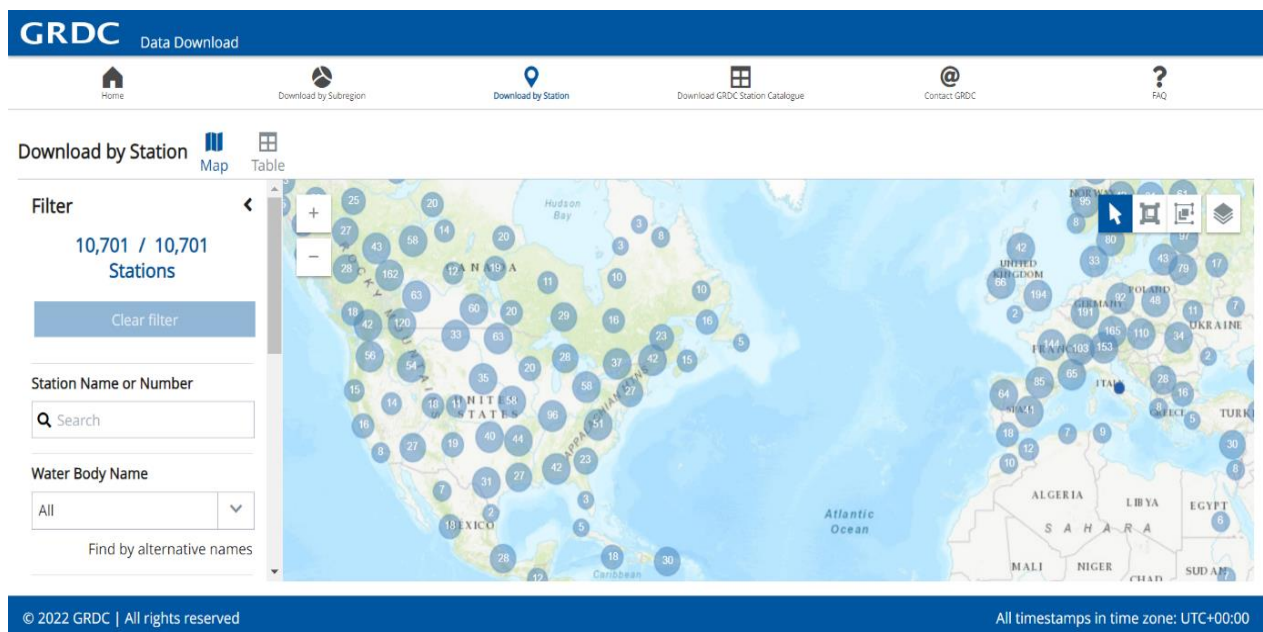


Figure 3-2 : GRDC website (The GRDC)

- **National centers for environmental information**

It is a US leading authority for environmental data and manages one of the largest archives of atmospheric, coastal, geophysical, and oceanic research in the world (National Centers for Environmental Information)

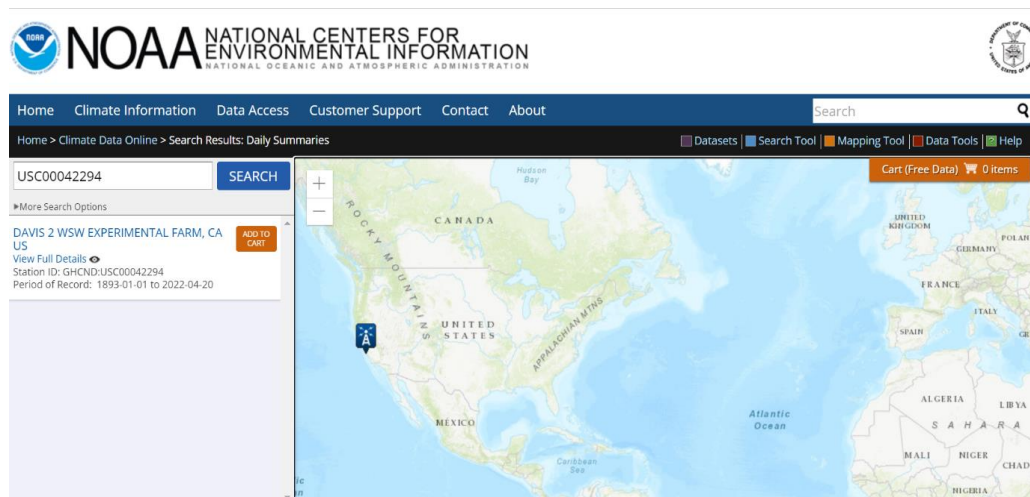


Figure 3-3 : NCEI Website

However, for the Algerian cases (Bouchegouf and Zardezas) we relied on the data provided by ANRH. The datasets we received were pretty limited compared to the ones we obtained from the above-mentioned websites.

3.1.3 Hydrometric station, meteorology stations, watershed

Hydrometric stations are placed on a river, lake, estuary, or reservoir where data on water quantity and quality are collected and recorded. Whilst rainfall and weather stations collect the amount of precipitation in a specific area by a rain gauge. This process is done on a daily basis. To establish a certain coherence between the 2 types of stations' datasets and after going through the 2 steps mentioned above, we took 1 to 4 rainfall measuring stations surrounding a main hydrometric station belonging to the same catchment area.

3.2 Study cases - geography

3.2.1 Watersheds chosen for the study

3.2.1.1 Duero - Spain

The sub-catchment area studied is part of the DUERO watershed situated within The Duero River Figure 3-4. Duero basin coincides almost exactly with the North Submeseta. The circle of mountains surrounding the basin is the area with the highest intensity of rainfall. The central area is much drier, but that is where the major aquifers, main cities, industry, and the most important agricultural production area are located.

Its predominant climate is the continental Mediterranean, with dry summers and cold winters. (Douro Hydrographic Confederation | Hispagua)

The headwaters of the watershed are found in the surrounding Urbion Mountains (Cordillera Ibérica), which reach a height of 1260 meters. The primary waterway has a length of 927 km and empties into the Atlantic Ocean in Porto, Portugal. (Cortes et al., 2019)

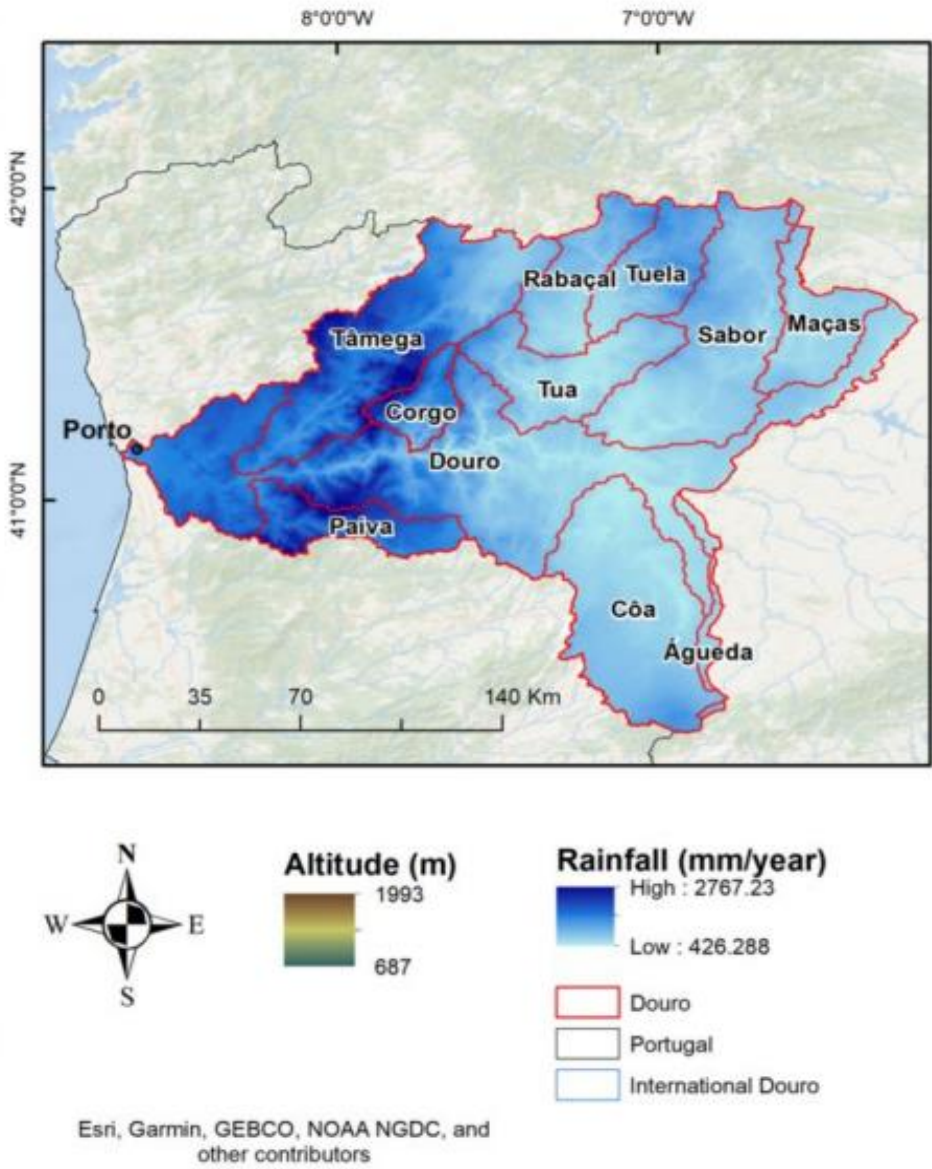


Figure 3-4 : Location map illustrating the Douro River watershed (Douro Hydrographic Confederation | Hispagua)

Chapter 03 The study zone – Geography and data

The hydrometric (runoff) station is situated in HERRERA DE DUERO near Valladolid inside the sub-catchment in the Duero watershed mentioned in Figure 3-4. Its geographical coordinates are as follows:

Table 3.2 : Geographical coordinates of Duero's hydrometric station

Latitude (DD)	Longitude (DD)	Elevation (m ASL)
41.56543	-4.66628	690.0

The three rainfall stations are all within a 50-kilometer radius of the hydrometric station to provide good precipitation data that match the streamflow values.

In Figure 3-5 we represented the stations using their coordinates on google maps, the blue one represents the hydrometric station while the yellow ones represent the meteorological stations. This color code is adapted for all other figures below.

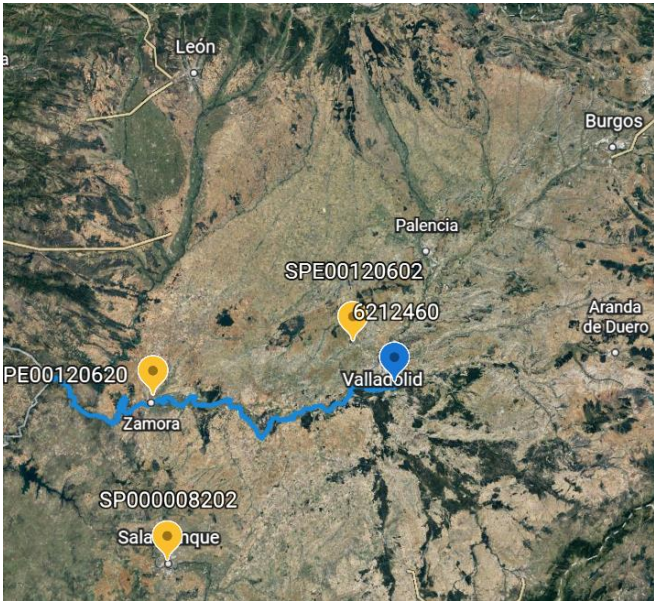


Figure 3-5 : Geographical localization of the rainfall stations within the sub-catchment

The sub-catchment delimited by HERRERA DE DUERO station N°6212460 contains the morphological features listed in the table below, which were obtained from the Spain - Centro de Estudios Hidrográficos (CEDEX) website.

Table 3.3 : Morphometric characteristics of the Douro River sub-catchment

Parameters	Douro River sub-catchment
Catchment area (km ²)	12740
Length (km)	927
Max. elevation (m)	2315
Min. elevation (m)	0
Average altitude (m)	700
Tributaries	<ul style="list-style-type: none"> • Left : Adaja, Águeda, Cega, Côa, Duratón, Huebra, Riaza, Tormes, Trabancos. • Right : Arandilla, Esla, Hornija, Pisuerga, Sabor, Tâmega.

3.2.1.2 Turia - Spain

The sub-catchment area studied is part of the Confederation JUCAR watershed situated within The TURIA (Guadalaviar) River. The climate described in the territory of the Confederation Júcar is a typical Mediterranean climate with warm summers and mild winters. It is situated within the thermo-Mediterranean bioclimatic and meso-Mediterranean dry (*Júcar Hydrographic Confederation | Hispagua*). The Montes Universales, located in the mountain rants in the far-western tip of the Sistema Ibérico, is the source of the TURIA River. It is known as the Guadalaviar River from its source to roughly the city of Teruel. The main water course is 280 km long and travels through Teruel, Cuenca, and Valencia provinces before discharging into the Mediterranean Sea not far from Valencia as shown in Figure 3-6.

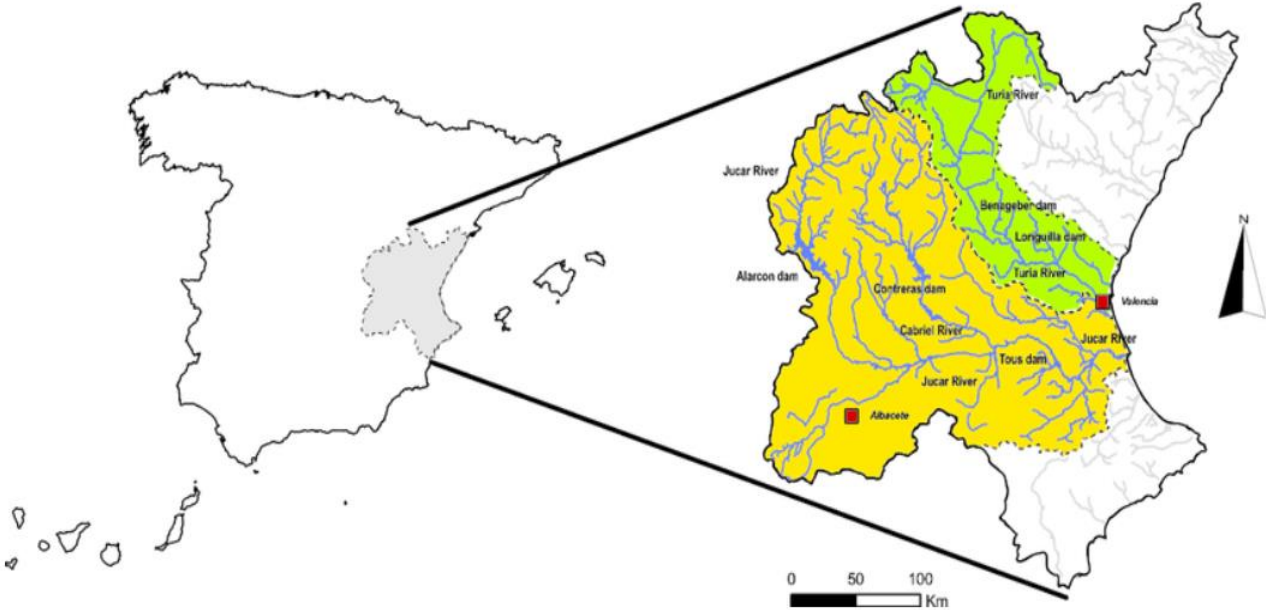


Figure 3-6 : Location map illustrating the TURIA River sub-catchment (Haro et al. 2014)

The hydrometric (runoff) station is situated in LA PRESA near Valencia at the outlet of the watershed.

Table 3.4 : Geographical coordinates of Turia's hydrometric station

Latitude (DD)	Longitude (DD)	Elevation (m ASL)
39.5177	-0.50406	50.0

The selected rainfall station is the only one that is located within a few kilometers of the hydrometric station (3 km), in order to provide good precipitation data that match the streamflow values. Figure 3-7

is the second largest river in California at 589 km. It starts in the high Sierra Nevada and flows through the rich agricultural region of the northern San Joaquin Valley before reaching Suisun Bay, San Francisco Bay, and the Pacific Ocean. An important source of irrigation water as well as a wildlife corridor, the San Joaquin is among the most heavily dammed and diverted of California's rivers. (*San Joaquin River* 2022).



Figure 3-8 : Map of the San Joaquin River watershed (SWAMP - San Joaquin River Basin)

The hydrometric (runoff) station is situated near VERNALIS next to the San Joaquin River mentioned in Figure 3-8. It measures the water of the Delta-Mendota canals, an aqueduct in central California that carries water diverted from the San-Joaquin River and passing by San Luis Reservoir.

Table 3.6 : Geographical coordinates of the San Joaquin hydrometric station

Latitude (DD)	Longitude (DD)	Elevation (m ASL)
37.676	-121.2663	7.62

The four stations are all within a 50-kilometer radius of the hydrometric station in order to provide good precipitation data that match the streamflow values. The difference between the distributions can be seen by comparing their graphs (rainfall VS runoff).

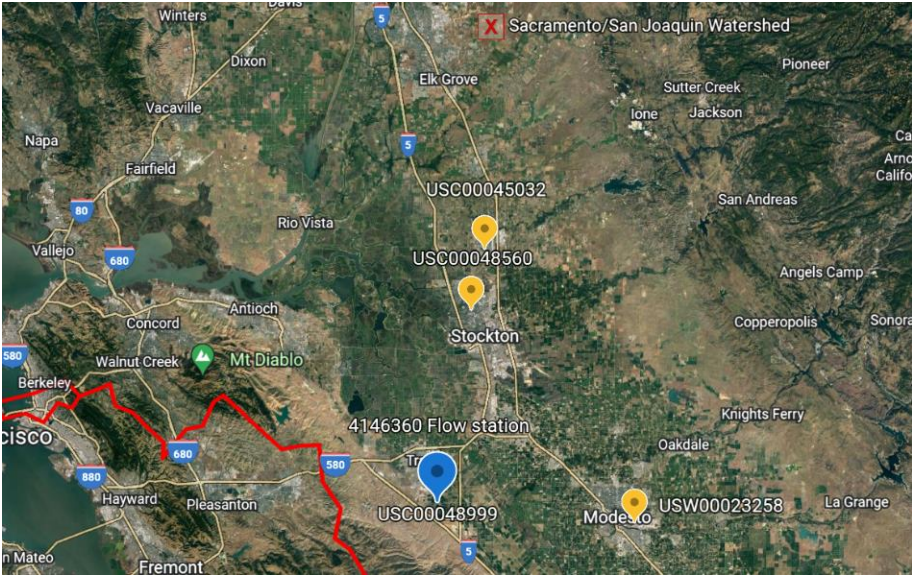


Figure 3-9 : Geographical localization of the rainfall stations within the watershed

The sub-catchment area delimited by VERNALIS station N° 4146360 contains the morphological features listed in the table below, which were obtained from the USGS website.

Table 3.7 : Morphometric characteristics of the San Joaquin River watershed

Parameters	San Joaquin River watershed
Catchment area (km ²)	35058.2
Length (km)	589
Max. elevation (m)	4410.1
Min. elevation (m)	0
Average altitude (m)	841.1
Average slope (degrees)	9.5
Tributaries	<ul style="list-style-type: none"> • Left: Fresno Slough • Right : Merced River , Tuolumne River, Stanislaus River, Mokelumne River

- Major events at the area:

Table 3.8 : The San Joaquin River basin's major floods

Date	Runoff (m ³ /s)	Event
26/12/1955	1442.216	Floods in the San Joaquin River Basin reflected those in the Sacramento River Basin. Flows on the San Joaquin River were completely controlled by Friant Dam. <i>(History of Flooding and Flood Protection 1983)</i>
30/04/1967	733.40	A vast amount of snowmelt from April to July compounded the flood damage experienced. The San Joaquin River Basin experienced a snowmelt volume of (9 621 144 m ³) to the valley floor. <i>(Chapter 2 - History of Flooding and Flood Protection 1983)</i>
03/01/1969	1347.88	Floodwaters produced by the January 1969 storms.
01/05/1997	1537.602	Floods in the Sacramento and San Joaquin Valleys resulted from local heavy runoff and flows from rivers that originate in the central Sierra Nevada. <i>(report.pdf 1999)</i>
23/02/2017	1138.335	Water from heavy storms in California throughout January finally started to reach the San Joaquin Valley over the past several weeks. The delay in the flood peak was due to the last 5 years of drought. <i>(The San Joaquin Demonstrates the Importance Of Floodplain Restoration 2017)</i>



Figure 3-10 : Floodwaters surround a farm along the San Joaquin River in Sacramento County, January 21, 1969. (NBC Los Angeles 2017)

Figure 3-10 presents a picture showing the damages caused at the area because of a major flood. This falls in correspondence with our data.

3.2.1.4 *Boucheougouf - Algeria*

The sub-catchment area studied is that of BOUCHEGOUF, it belongs to the basin of the SEYBOUSE (middle Seybouse) in the Northeast of Algeria in the territories of the wilaya of Guelma. It has a Mediterranean climate with a seasonal behavior, where we observe a succession of dry and rainy seasons very contrasted on an annual scale and variable flow regimes on an internal scale. The Boucheougouf watershed is located on the right side of the Seybouse basin, it is limited to the North by the Constantinois-East coastal basin, to the East by the Medjerdah basin, to the South by the basins of Oued Cherf and Oued Bouhamdane, to the west by the two basins of Guelma (average Seybouse) and Ain Berda (Oued Ressoul). (IZERROUKYENE 2017)

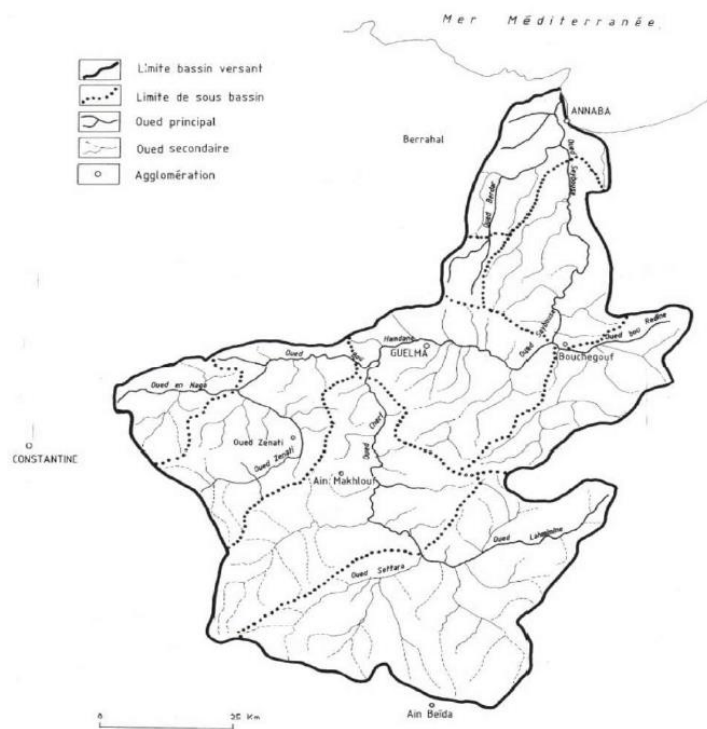


Figure 3-11 : Map of the hydrographic network of the Seybouse watershed (IZERROUKYENE 2017)

The rainfall and hydrometric data used were collected from the National Water Resources Agency (ANRH). However, the data of temperatures were collected from the "Global Weather Data for SWAT".

Table 3.10 : Morphometric characteristics of Bouchegouf sub-catchment

Parameters	Bouchegouf sub-catchment
Catchment area (km ²)	550
Length of main thalweg (km)	53
Max. elevation (m)	1317
Min. elevation (m)	95
Average altitude (m)	641

3.2.1.5 Zardezas - Algeria

Zardezas watershed coded (03-09-02) by ANRH, is located in the North-East of Algeria in the territories of the wilaya of Skikda.

The basin of Oued Saf-Saf, coded (0309) to which the Zardezas sub-basin belongs, results from the conjunction of two rivers: Oued Bouhadjeb and Oued Khemkhem, and part of the large Constantinois coastal watershed. It is limited by the Mediterranean Sea in the North, the basin of Rhumel Kebir in the East and Southeast, and the basin of Soummam in the West. A Mediterranean climate predominates the region, with cold and relatively humid winter and hot summer. (Bouhoun 2020)

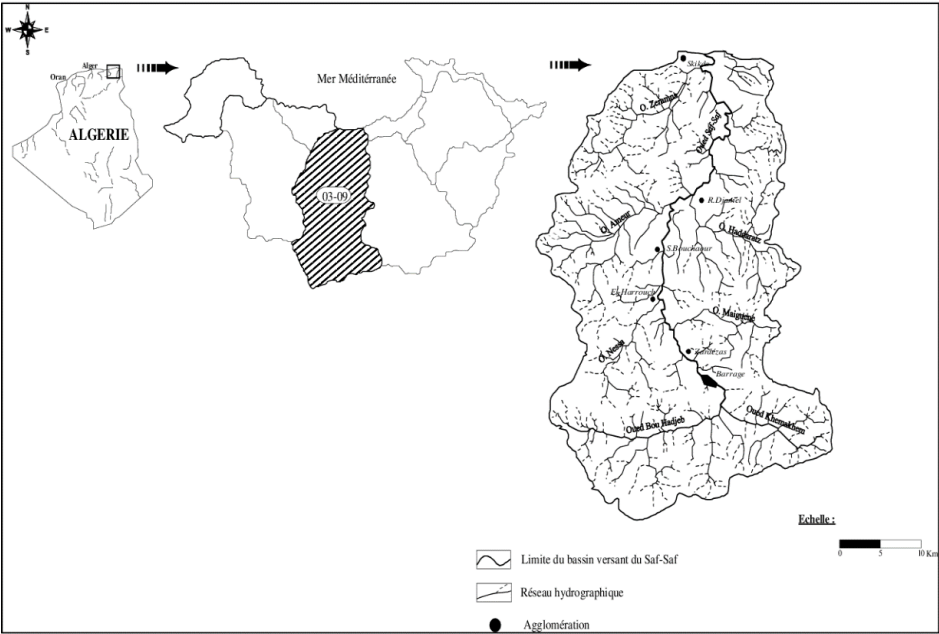


Figure 3-13 : Map of the hydrographic network of the Saf-Saf watershed (Khelfaoui, Zouini 2010)

The geographical coordinate of the hydrometric station is presented in the table below:

Table 3.11 : Geographical coordinates of the hydrometric station of Zardezas (ANRH)

X	Y
878.75	370.67

The rainfall station N°030903 (ANRH) is situated next to the Zardezas dam mentioned in Figure 3-14.

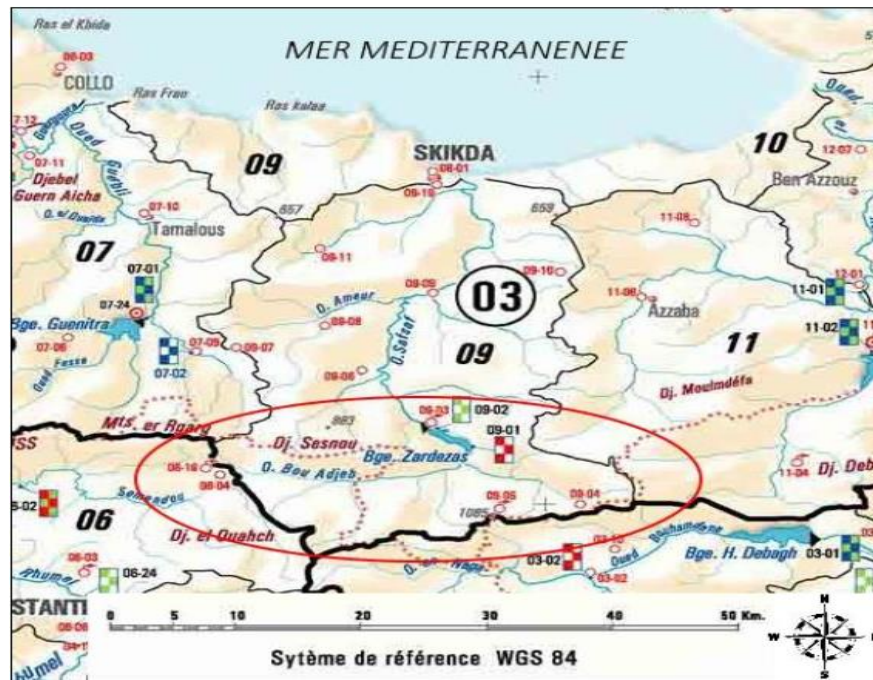


Figure 3-14 : Map of the hydrographic network of the Zardezas sub-catchment (ANRH)

The sub-catchment area delimited by Zardezas station N°030901 that is located in the outlet contains the morphological features listed in the table below:

Table 3.12 : Morphometric characteristics of Zardezas sub-catchment (ANRH)

Parameters	Zardezas sub-catchment
Catchment area (km ²)	345
Length of main thalweg (km)	24
Max. elevation (m)	1220
Min. elevation (m)	206
Average altitude (m)	641

3.3 Study cases: data

3.3.1.1 Streamflow data:

The streamflow data collected from the hydrometric stations of our chosen study zones are presented in the table below.

Table 3.13 : List of hydrometric stations

Hydrometric Station	Station's code	Country	Years
S1	6212460	Spain	[1912-2017]
S2	6227130	Spain	[1912-2017]
S3	4146360	USA	[1923-2021]
S4	140501	Algeria (Bouchegouf)	[1985-1995]
S5	030902	Algeria (Zardezas)	[1990-1996]

The range of our data covers several years of observations, we have up to 100 years of consecutive daily data.

Table 3.14 Statistical description of streamflow data

Hydrometric station	Number of observations	Missing values	Maximum Value (m ³ /s)	Minimum Value (m ³ /s)	Mean Value
S1	38625	4999	767.2	0	101.4
S2	38625	5342	310	0	128.6
S3	35611	1827	1982.17	0	66.37
S4	3652	0	124.4	0	1.91
S5	2557	0	124.3	0	1.58

The dataset differs in consistency and accuracy. Data downloaded from websites contain a higher number of values compared to the ANRH. In the other hand, ANRH data has no missing data compared to the other values.

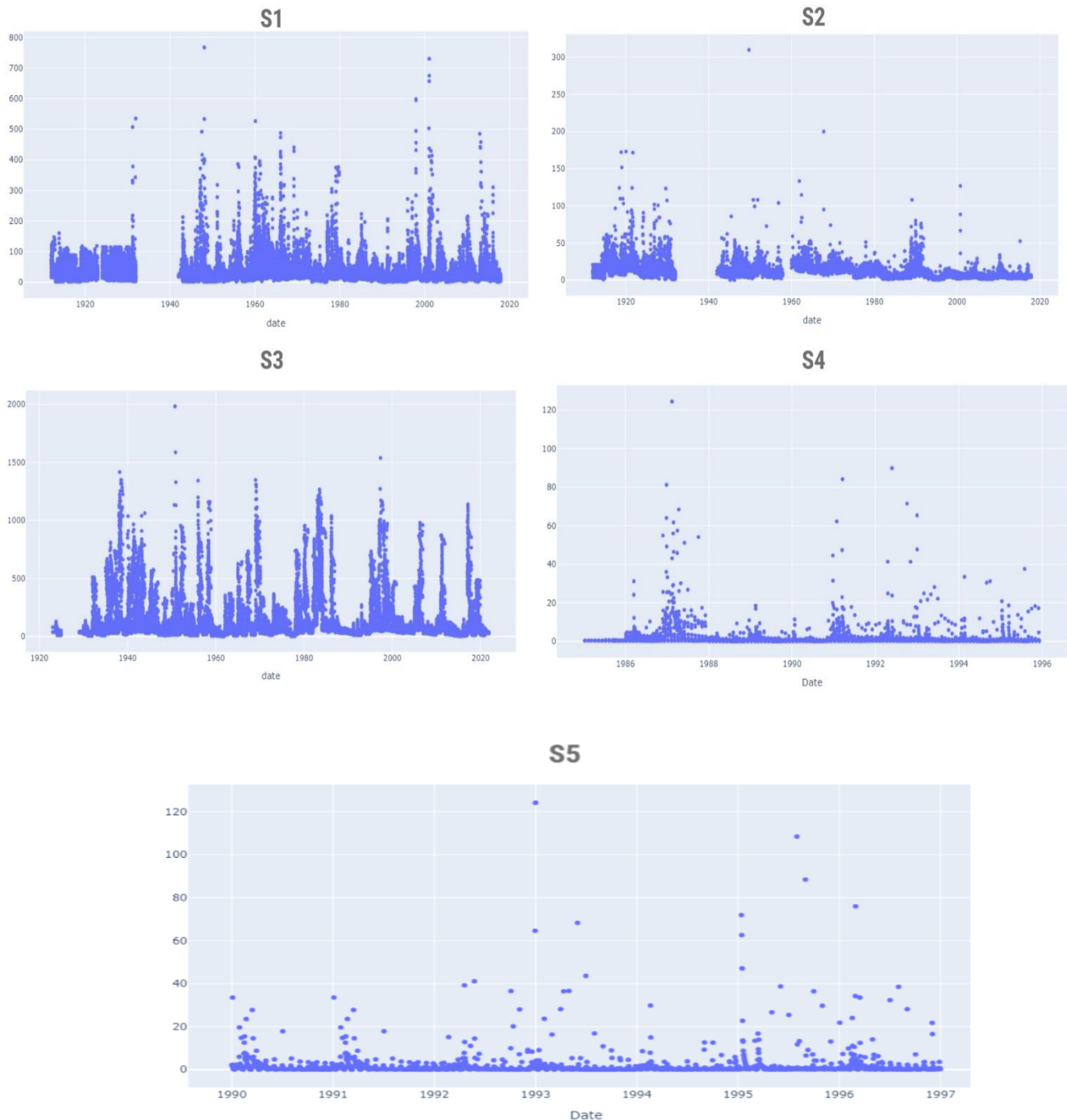


Figure 3-15 : Streamflow distribution of the 5 stations

The runoff observations fluctuate to the point where there are significant peaks in each period with different scales. Most of the cases correspond to major floods. Comparing the data variation intervals, it can be divided into high flow, medium flow and low flow data.

The gaps shown in S1, S2 and S3 represent missing values at those periods. There had been no recorded data.

3.3.1.2 *Precipitation data:*

In the tables presented blow, for each hydrometric station we picked 1 to 4 corresponding rainfall stations with daily recorded observations. This choice of course was based on the criteria mentioned earlier.

Table 3.15 : Rainfall data of Station (1)

Hydrometric station	Rainfall station	Number of rainfall observations	Missin g values	Maximu m Value (mm)	Minimu m Value (mm)	Mean Valu e
S1	SP000008202	28207	2	59	0	1.03
	SPE00120602	31067	63	90.8	0	1.24
	SPE00120620	36353	158	66.1	0	1.01

The precipitation data measured by the 3 stations are presented in the graphs below:



Figure 3-16 : Rainfall data distribution Station (S1)

Table 3.16 : Rainfall data of Station (2)

Hydrometric station	Rainfall station	Number of rainfall observations	Missing values	Maximum Value (mm)	Minimum Value (mm)	Mean Value
S2	SPM00008284	15066	2975	230.4	0	1.23

The precipitation data measured by the station is presented in the graph below:

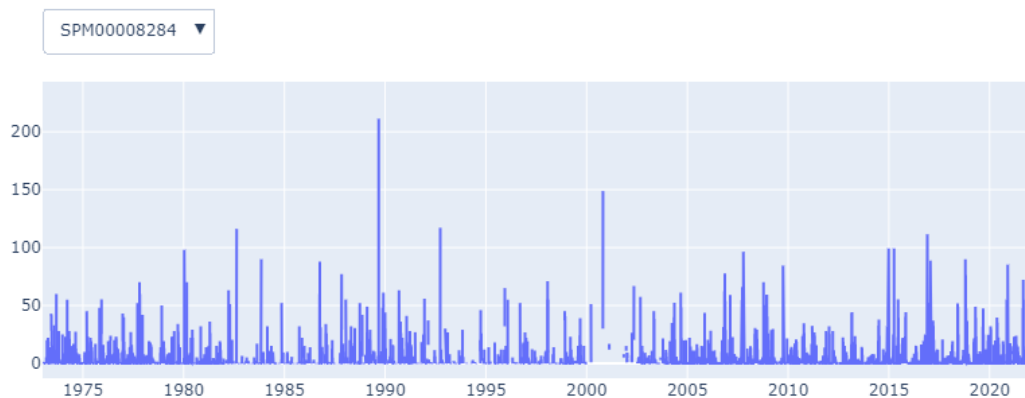


Figure 3-17 : Rainfall data distribution of S2

Figure 3-17 shows the precipitation distribution of the weather station near S2, it varies from 0 to 100 mm with few peaks showing up every 5 years. We can also notice a lack of data from 1993 to 2003 which will be particularly taken care of in the modeling process afterward.

Table 3.17 : Rainfall data of Station (3)

Hydrometric station	Rainfall station	Number of rainfall observations	Missing values	Maximum Value (mm)	Minimum Value (mm)	Mean Value
S3	USC00045032	37590	863	95.5	0	1.23
	USC00048560	31311	599	81.3	0	1.20
	USC00048999	28400	967	63.2	0	0.69
	USW00023258	37895	2454	69.1	0	0.85

The precipitation data measured by the 4 stations are presented in the graphs below:

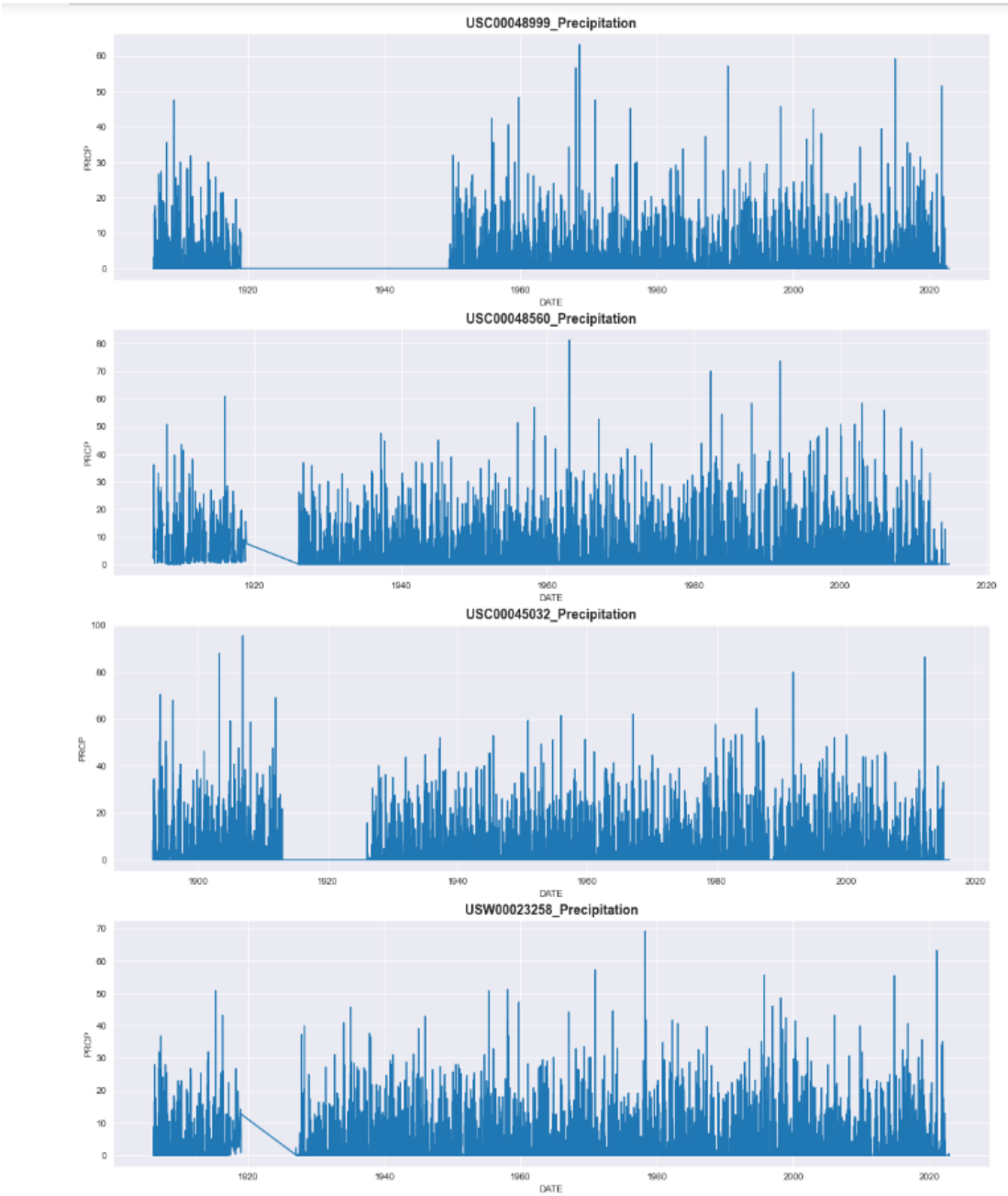


Figure 3-18 : Rainfall data distribution of S3

The rainfall at the San Joaquin River is measured by different stations, and each station shows significant changes in the daily concentration of precipitation from 1900 to 2021, the precipitation has extended wet periods with frequent and intense rains causing floods like the Storm of December 11, 1906. While, in other periods, the rainfall has decreased remarkably with quick fluctuations that can be defined as “jumps”. Missing values extend for almost 20 years.

Table 3.18 : Rainfall data of Station (4)

Hydrometric station	Rainfall station	Number of rainfall observations	Missing values	Maximum Value (mm)	Minimum Value (mm)	Mean Value
S4	140505	3651	1	74.1	0	1.34

The precipitation data measured by the station is presented in the graph below:

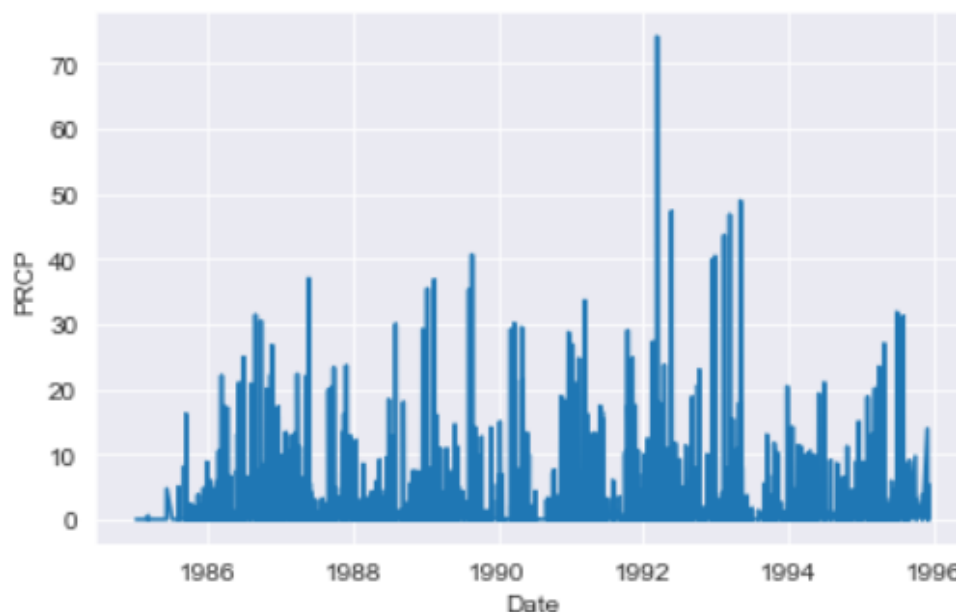


Figure 3-19 : Rainfall data distribution of S4

The rainfall at Bouchegouf station distribution shows significant changes in the daily concentration of precipitation during the period 1986- 1996, The maximum peak of 73.1 m³/s corresponds to the date of 03/11/1992.

Table 3.19 : Rainfall data of Station (5)

Hydrometric station	Rainfall station	Number of rainfall observations	Missing values	Maximum Value (mm)	Minimum Value (mm)	Mean Value
S5	030903	2557	0	76.7	0	1.65

The precipitation data measured by the station is presented in the graph below:

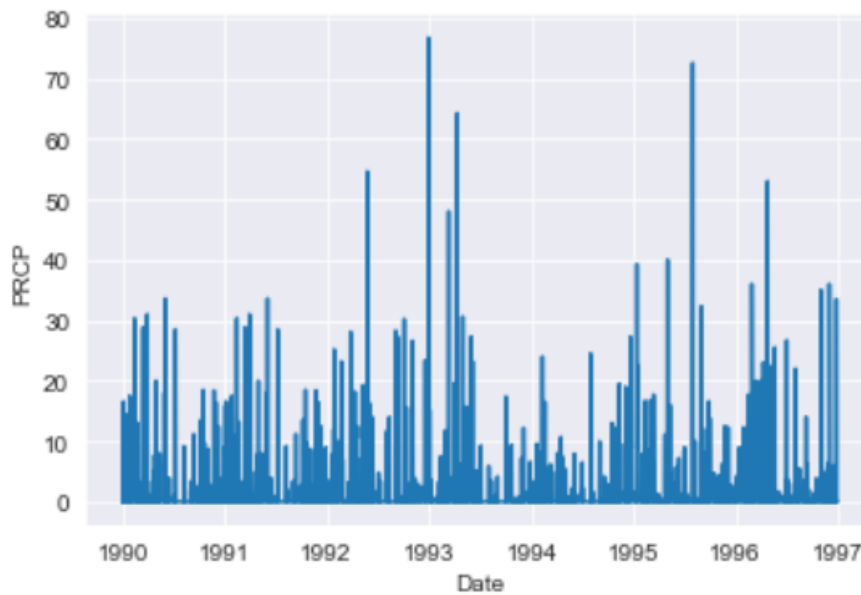


Figure 3-20 : Rainfall data distribution of S5

The rainfall at Zardezas station distribution shows significant changes in the daily concentration of precipitation during 7 years of observation.

3.3.1.3 *Temperature data:*

The temperature data were collected from the "Global Weather Data for SWAT and the "National Oceanic and Atmospheric Administration ". It is presented by three categories of data:

- Minimum temperature
- Average temperature
- Maximum temperature

The temperature varies in a cyclic way known as "cyclic thermal fluctuations" influenced by differences in topographical surface and altitude.

The maximum temperature measured by the 4 rainfall stations of S3 is presented in the graphs below:

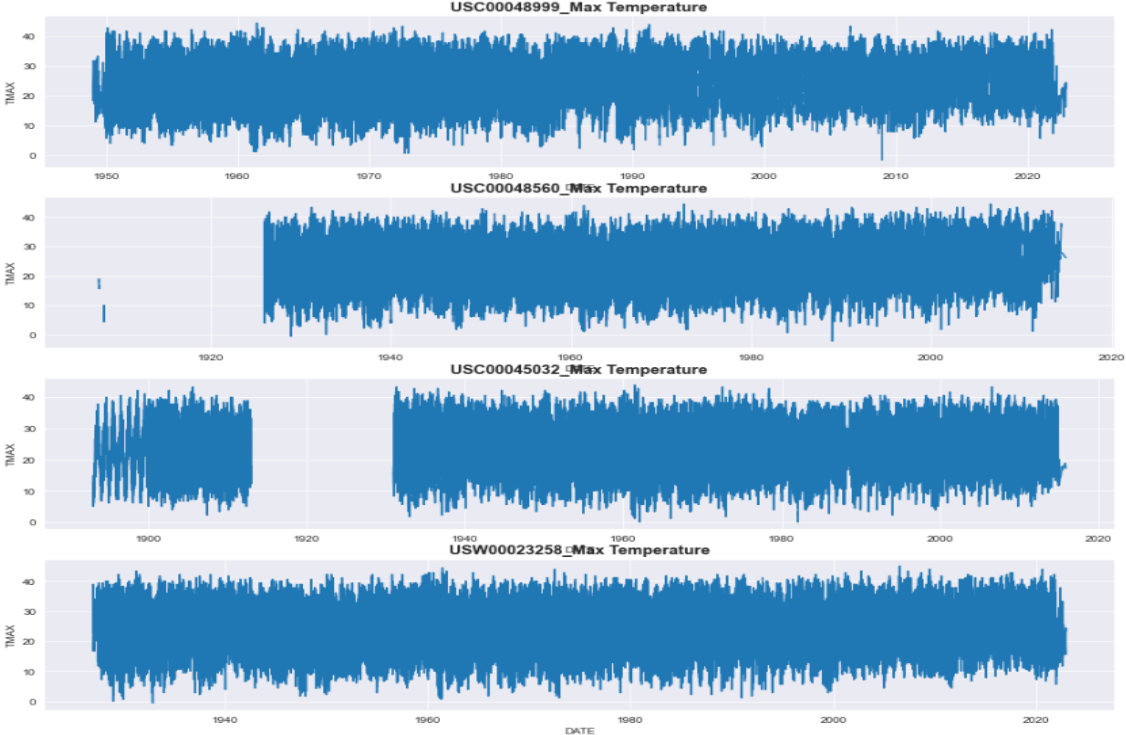


Figure 3-21 : Example of maximum temperature data of station (3)

3.4 Conclusion

In this chapter, a detailed description of the data and their studied watersheds, as well as their morphological characteristics, was stated. On the other hand, we also presented the hydro-meteorological data used by specifying their sources, their statistical characteristics as well as their graphical distributions.

Chapter 04 Tools and Methodology

Introduction

In recent decades, the advancement of computer science technology and research on complex natural systems has resulted in a multitude of mathematical models in hydrology such as Machine learning techniques with data-driven methods.

This chapter presents the used tools inspired by these technologies and the applied methodology to treat the data presented in the previous chapter and implement it in our model.

4.1 Tools

4.1.1 Jupyter

Jupyter is a web application that allows you to write in over 40 different programming languages, including Python, Julia, Ruby, R, and Scala². It is a collaborative effort whose purpose is to provide free software, open formats, and interactive computing services. It supports the development of notebooks, which are programs that contain both text in Markdown and code (*Jupyter 2022*). We used these notebooks feature to explore and analyze our data.

4.1.2 Anaconda

Anaconda is a free and open-source distribution of the Python and R programming languages for the creation of data science and machine learning applications, with the goal of simplifying package management and deployment. (*Anaconda 2022*)

4.1.3 Python

Python is a programming language that stands the test of time due to its flexibility, simplicity, and effective tools for creating modern applications. (*CFI Team [2021]*). It was conceived in the late 1980s by Guido van Rossum at Centrum Wiskunde & Informatica (CWI) in the Netherlands (*Python 2022*). It is an object-oriented programming language with simple syntax and modern scripting. It comes with a large

number of libraries, some of which are specially built for Machine learning and deep learning applications, which makes it our chosen programming language for the work.

4.1.3.1 *Pandas*

Pandas is a popular open-source Python library for data science/data analysis and machine learning activities. It is created based on Numpy, another library that supports multidimensional arrays. Pandas, being one of the most commonly used data wrangling tools, integrates well with many other Python data science modules. Among its various functionalities, we can mention:(*Pandas 2021*)

- The DataFrame object, which permits for easy and efficient data manipulation with indexes that can be strings of characters.
- Intelligent data alignment and missing data handling (NaN = not a number). Labels are used to align data (strings of characters). Sorting completely jumbled data using multiple criteria.
- Resizing and pivot tables (also known as pivot tables).
- Merging and joining massive amounts of data.
- Time series analysis.

4.1.3.2 *Tensorflow*:

TensorFlow is an open-source machine learning platform. It offers a rich and adaptable ecosystem of tools, libraries, and community resources that help experts to advance the field of machine learning and programmers quickly create and implement applications that use this technology. It facilitates the creation and training of machine learning models by utilizing intuitive high-level APIs such as Keras with eager execution. (*TensorFlow*).

4.1.3.3 *Keras*:

Keras is a high-level Python open-source library that is used to manage deep neural networks (DNN). It gives users a convenient interface for manipulating DNNs by including libraries like as Theano, CNTK, and TensorFlow. (*Keras*) Which allows us to manipulate multidimensional data tables. We combined Keras and TensorFlow for our application. We worked with Keras version 2.9 in our project.

4.1.3.4 *Sci-kit learn*

Scikit-learn (Sklearn) is Python's most usable and robust machine learning package. It offers a set of fast tools for machine learning and statistical modeling, such

as classification, regression, clustering, and dimensionality reduction, via a Python interface. This library, which is essentially written in Python, is built upon NumPy, SciPy and Matplotlib. (*Scikit-learn* 2021)

4.1.3.5 *Matplotlib*

Matplotlib is a Python package that allows you to create static, animated, and interactive visualizations. (*Matplotlib*) It helped us to:

- Produce plots suitable for publishing.
- Create interactive figures that can be zoomed in, panned, and updated.
- Change the visual design and layout.
- Export to a variety of file formats.
- Incorporate JupyterLab and Graphical User Interfaces.
- Use a diverse set of third-party programs based on Matplotlib.

4.1.3.6 *Seaborn*:

Seaborn is a matplotlib-based Python data visualization package. It offers a high-level interface for creating visually appealing and useful statistics visuals. (Waskom 2021)

4.2 Notions

4.2.1 Nan

NaN, which stands for Not a Number in computing, is a member of a numeric data type that can be regarded as an undefined or unrepresentable value. (*NaN* 2022) In our case, we symbolised missing values with NAN.

4.2.2 **Overfitting**

Overfitting is a statistical modeling mistake that arises when a function is too strongly fitted to a small collection of data points. As a result, the model is only helpful in reference to its initial data set and not in relevance to any additional data sets.

Overfitting the model often entails creating an overly complicated model to explain anomalies in the data under examination. In actuality, the data being investigated frequently contains some degree of mistake or random noise. Attempting to force the model to adapt to somewhat erroneous data might thereby infect the model with significant flaws and impair its prediction effectiveness. (Twin [2021])

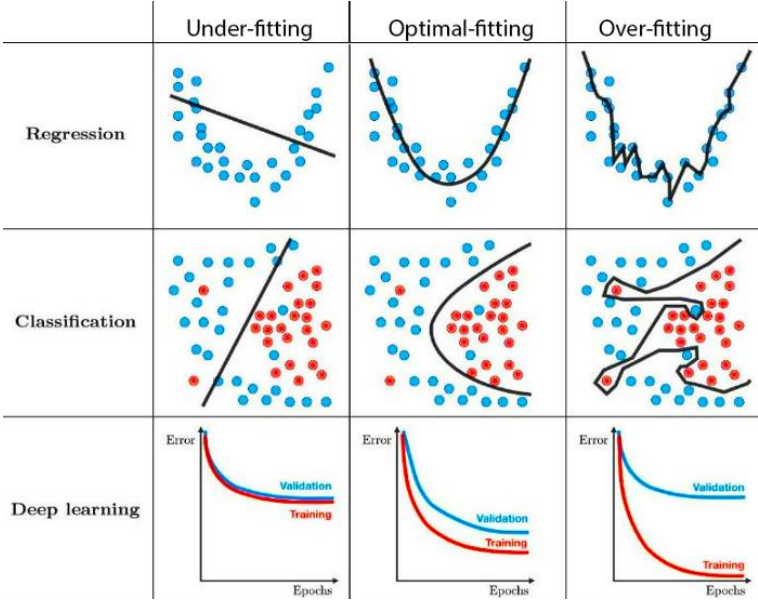


Figure 4-1 : Comparison between fitting types in different machine learning algorithms (Minhas 2021)

4.2.3 Outliers

An outlier in statistics is a data point that deviates considerably from other observations. It can be caused by measurement variability or by experimental mistake; the latter is sometimes eliminated from the data set. In statistical analysis, an outlier can generate major consequences.(Grubbs 1969)

4.3 Processing approach

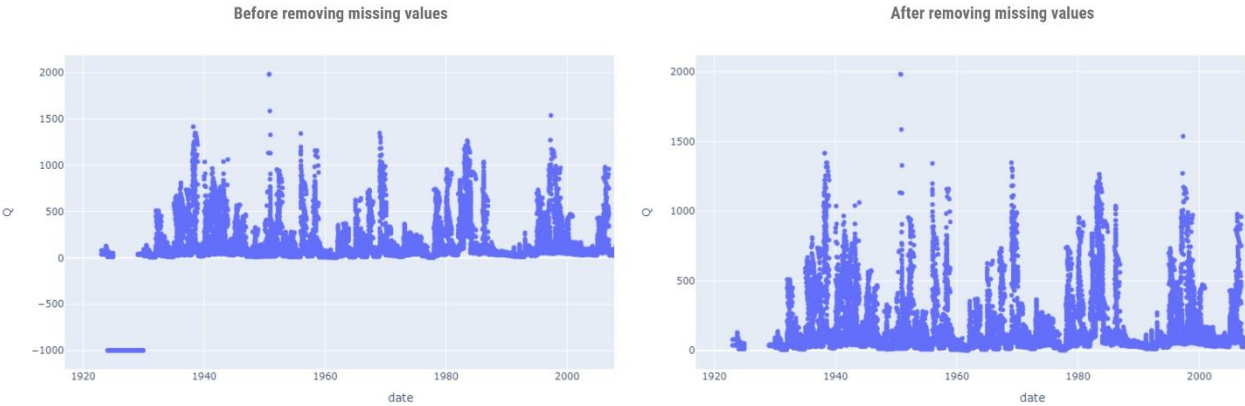
4.3.1 Phase 01 - Collecting data

The first step of modeling in deep learning models would be to gather data. We followed the criteria mentioned in the previous chapter to download the data from open-source websites for foreign countries while we got the ones of Algeria from ANRH. The data is daily measures from meteorology and hydrometric stations for varying periods of time. We got larger datasets from the online websites compared to the Algerian cases due to the lack of data availability in Algeria.

Data included daily: streamflow, precipitation, maximum temperature, minimum temperature, average temperature, spatial coordinates of the stations.

4.3.2 Phase 02 - Treating and data cleansing

While going through the datasets, we spot the existence of missing values and some outliers within. Thus, all data went through a cleansing phase where we replaced these values with NAN. This is considered a crucial part of the modeling process since any sort of erroneous values can affect our model’s performance and accuracy. Some examples of data treatment are shown below in figures:



Chapter 04 Tools and Methodology Figure 4-2 : Example of missing data treatment in streamflow

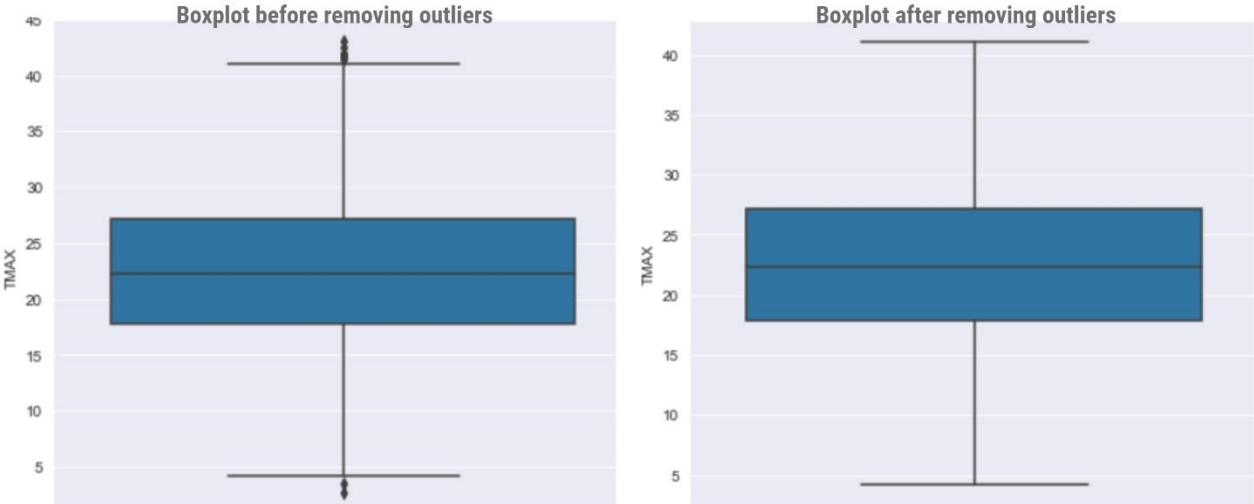


Figure 4-3 Example of treating outliers of maximum temperature

To ensure the rainfall-runoff data coherence we made the plot below representing the precipitation data of 3 meteorological stations surrounding the hydrometric station S1.

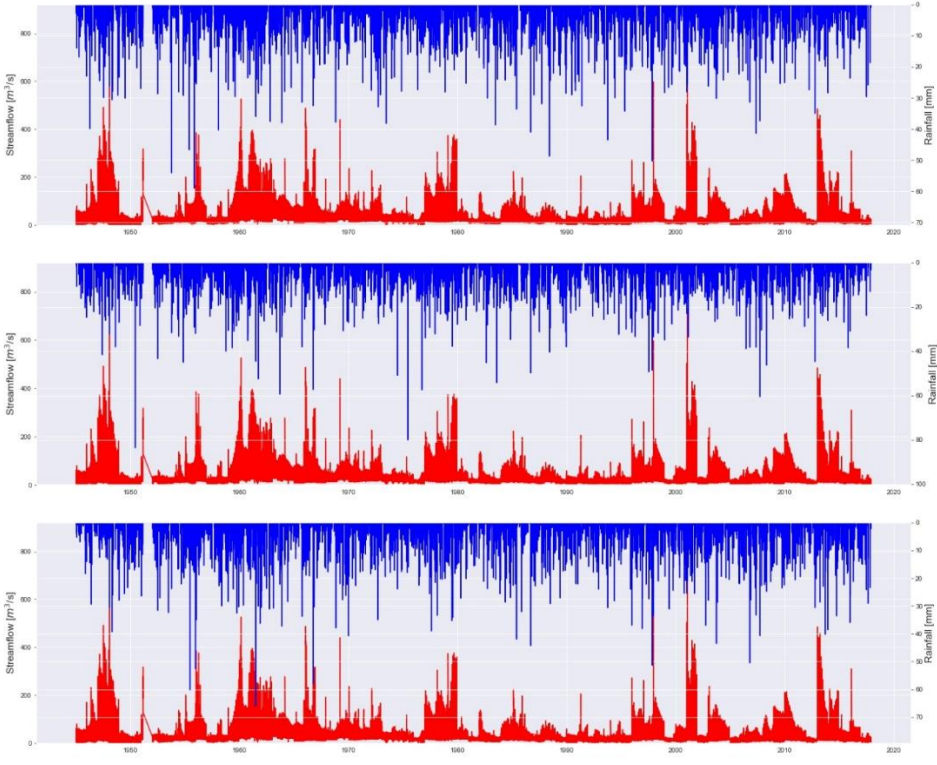


Figure 4-4 Rainfall-runoff plots of S1

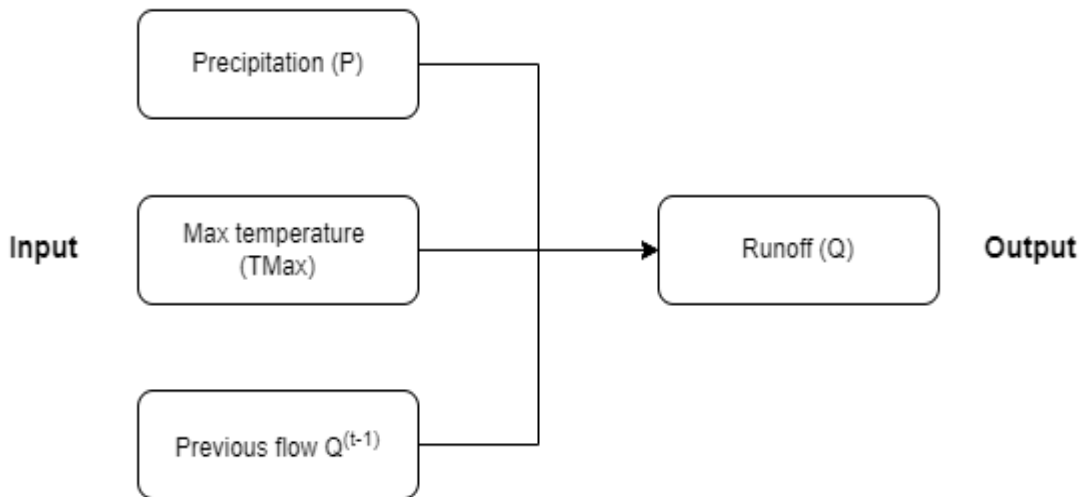
Figure 4-4 represents one example of the treated cases, the rest could be found in APPENDIX.

4.3.3 Phase 03 - Determining appropriate inputs/output

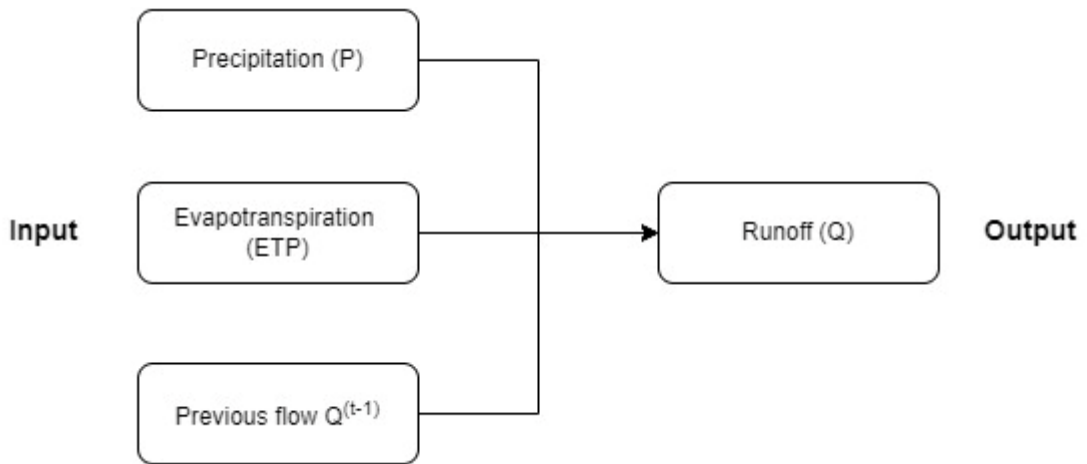
Before starting the modeling process, we must first identify the suitable inputs for our model, and what we are expecting as an output. A usual rainfall-runoff prediction will include precipitation as an input, and streamflow as output. In our study, for the sake of identifying the exact influencing factor of the flow output, we conducted 3 studies by varying the input variables shown below.

Since we are working on short term prediction, the expected output will be $Q^{(t)}$ where t stands for next day.

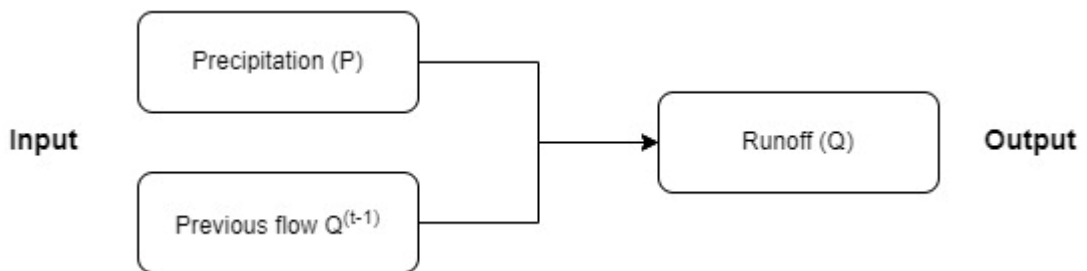
- 1st case:



- 2nd case:



- 3rd case:



- The evapotranspiration data are obtained from the temperature using the following Oudin equation: (Oudin 2006)

$$\left\{ \begin{array}{ll} ETP = \frac{R_e T_a + 5}{\lambda \rho \cdot 100} & \text{If } T_a + 5 > 0 \\ ETP = 0 & \text{Otherwise} \end{array} \right.$$

Where R_e is the extraterrestrial radiation ($\text{MJ m}^{-2} \text{d}^{-1}$) which depends only on the latitude and the Julian day (whose calculation is detailed by (Morton 1983)), λ is the latent heat flux (MJ kg^{-1}), ρ is the density of water (kg L^{-1}), to obtain ETP in mm d^{-1}) and T_a is the air temperature ($^{\circ}\text{C}$),

4.3.4 Phase 04 - Data preprocessing for the model

Working with RNN-LSTM models require a certain data preprocessing. Thus, our datasets went through a second stage of preprocessing of: Scaling, Reshaping, Cleansing.

4.3.4.1 *Standardization*

It is the process of transforming a numerical feature's real range of values into a standard range of values (Burkov 2019), for our case we arranged the scaling in the interval $[0, 1]$.

4.3.4.2 *Selecting inputs*

This is the step where we select our features' according to phase 03. The use of RNN-LSTM models gives us the chance to include streamflow from previous consecutive days as an input, which is what we referred to when mentioning $Q^{(t-1)}$. We chose to work with a lag time equals to 5 days $Q_{\text{prev}} = (Q(t-1) Q(t-2) Q(t-3) Q(t-4) Q(t-5))$ as an approximate estimation of the time the rainfall would take to reach the outlet.

4.3.4.3 *Data cleansing*

Data goes through another round of cleansing where all NAN values are removed, the dataset numbers decrease significantly in this phase.

4.3.4.4 *Data split*

The train-test split is a strategy for assessing a machine learning algorithm's performance on new data that was not used to train the model.

This is how we anticipate using the model in practice. To put it another way, we want to fit it to existing data with known inputs and outputs and then make predictions on new cases in the future where we don't have the expected output or goal values. It is also used to avoid overfitting.

We proceed by dividing the original data into three groups:

- **Training set:** the collection of data used to train the model and teach it to discover the hidden features/patterns in the data. The same training data is provided to the neural network design repeatedly in each epoch, and the model continues to learn the data's attributes. The training set includes a diverse and a random collection of inputs so that the model may be trained in all settings and forecast any previously unknown data sample that may arise in the future. It represents 70% of all data.

This portion of the total data to be used for training should contain enough patterns for the network to appropriately simulate the underlying connection between input and output variables. Initially, the weights and threshold values are allocated in all random numbers. These are changed during training according to the error, or the gap between ANN output and target answers. This change can be repeated recursively until a weight space with the minimum overall prediction error is obtained.

- **Validation set:** it is a different collection of data from the training set that is used to assess the performance of our model during training. This validation method provides information that enables us to fine-tune the model's hyper-parameters and settings. It's like a critic informing us whether our training is progressing in the proper path. The model is trained on the training set while being evaluated on the validation set at the end of each epoch. It represents 30% of the training set picked earlier.
- **Test set:** it is a distinct collection of data that is used to test the model after it has been trained. The testing set serves as an assessment of the final model and algorithm. It gives unbiased final model performance indicating the accuracy.

4.3.5 Phase 05 - Model architecture

Our RNN and LSTM models consist of these types of layers:

- Input layer (RNN or LSTM depending on the model) where our 5 inputs are introduced. RNN and LSTM identify long-term dependencies among time steps in time series and sequence data.
- Flatten layer: Data is converted into a 1-dimensional array for inputting it to the next layer.
- Dense layer: A dense layer is one that is deeply linked to the layer before it, which indicates that the neurons in the layer are connected to every neuron in the layer before it.
- Dropout layer: Dropout refers to disregarding units (i.e. neurons) during the training phase of a random group of neurons. Meaning that these units are not considered during a certain forward or backward pass. This layer contributes to avoid overfitting. Because a fully linked layer fills the majority of the parameters, neurons acquire co-dependency amongst themselves during training, which limits the individual power of each neuron and leads to over-fitting of training data.

Overall, each model consists of 9 layers, with 100 neurons in RNN and 200 neurons in LSTM as shown in figures below:

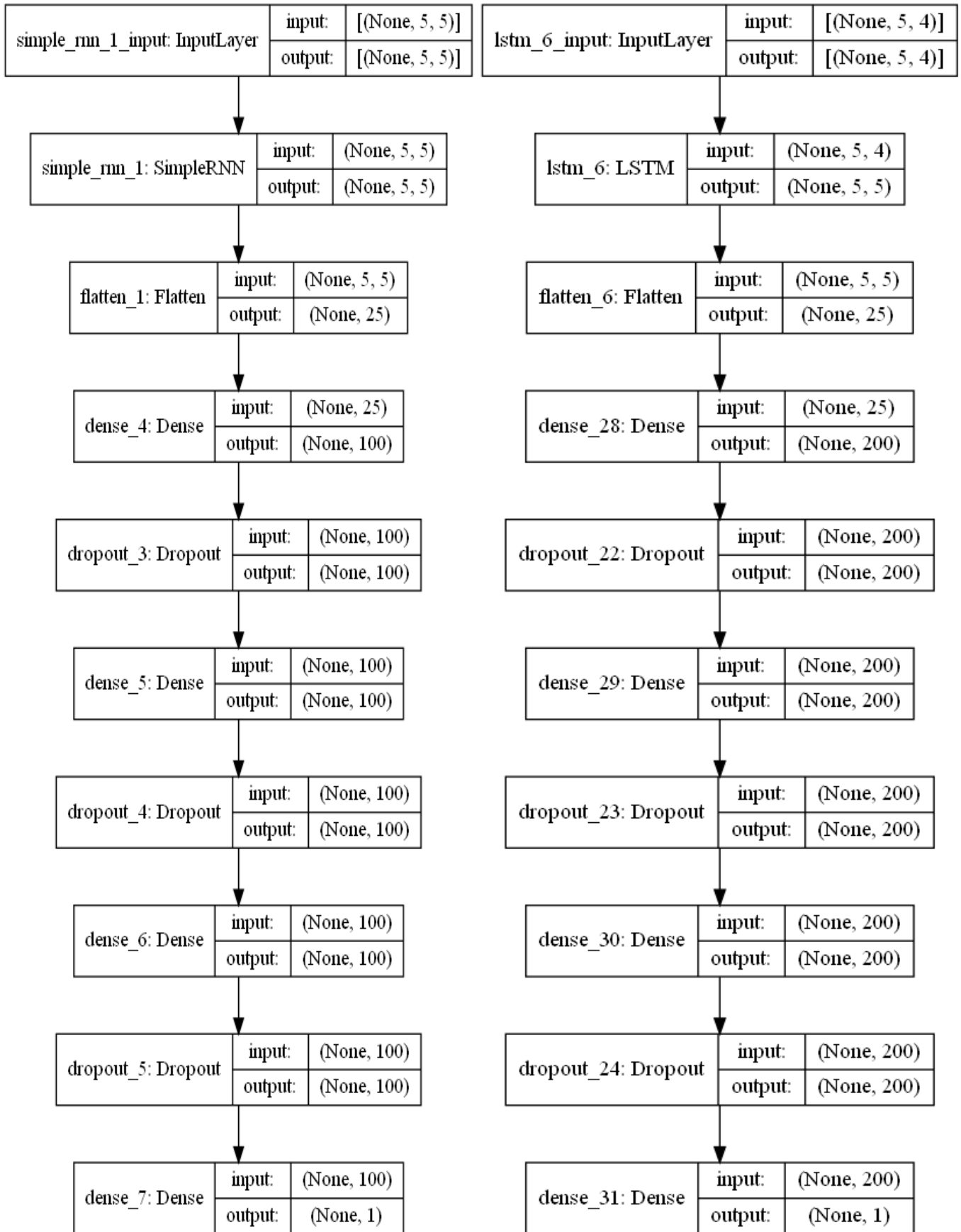


Figure 4-6 RNN & LSTM architectures

4.3.6 Validation and performance Monitoring (Numerical and graphical):

- Numerical performance:

The evaluation of the model's accuracy is an important phase in any machine learning model. Nevertheless, determining the validity of a model and its parameters based on indicators that have the same units as the variables remain challenging, because the magnitude of each indicator will always rely on the data utilized and the specific situation. The issue of identifying a maximum or reference point appears to be significant. Thus, standardized indicators create a reference performance value in each indicator to standardize model evaluation. The main virtue of normalized criteria is that they are dimensionless, allowing for model comparison.

Consequently, we used these research-proven standardized indicators to provide further information about our model's relevance: Percent Bias (PBIAS), Nash-Sutcliffe efficiency (NSE) also known as R-squared and RSR during training and testing.

4.3.6.1 *Percent bias (PBIAS)*

It measures the average trend of the simulated values (bigger or lower) in comparison to the observed values (Gupta, et al. 1999) As a result, it estimates the simulation's under or overestimation. Its optimal value is "0." Positive values of this criteria suggest an underestimating of the bias, whereas negative values indicate an overestimation of the bias.

$$PBIAS = 100 \cdot \frac{\sum_{i=1}^N (Q_o - Q_p)}{\sum_{i=1}^N Q_o} \quad \text{Equation 4.1}$$

This criterion is recommended for its ability to indicate model performance.

4.3.6.2 *RMSE-observations standard deviation ratio (RSR)*

RMSE is the square root of the Mean Squared error. It measures the standard deviation of residuals (Chugh 2022) :

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Q_o - Q_p)^2} \quad \text{Equation 4.2}$$

It handles the penalization of large errors done by square rooting it.

Although it is widely assumed that the lower the RMSE, the better the model performance, only (Singh et al., 2004) have provided a guideline to define what is deemed a low RMSE based on the standard deviation of the data. Based on (Singh et al., 2004)'s approach, a model evaluation statistic known as the RMSE-observations standard deviation ratio (RSR) was created. RSR standardizes RMSE by utilizing the standard deviation of the data:

$$RSR = \frac{RMSE}{\text{standard deviation}_o} = \frac{\sqrt{\sum_{i=1}^N (Q_o - Q_p)^2}}{\sqrt{\sum_{i=1}^N (Q_o - \bar{Q}_o)^2}} \quad \text{Equation 4.3}$$

4.3.6.3 Nash–Sutcliffe model efficiency coefficient (R-squared)





It represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e., irrespective of the values being small or large, the value of R-squared (NSE) will be less than one (Chugh 2022):

$$NSE = 1 - \frac{\sum_{i=1}^N (Q_o - Q_p)^2}{\sum_{i=1}^N (Q_o - \bar{Q}_o)^2} \quad \text{Equation 4.4}$$

Where : \bar{Q} – mean value of Q_o

NSE ranges between $-\infty$ and 1.0 (1 inclusive), with NSE =1 being the optimal value. (Moriasi et al. 2007)

Table 4.1 model performance ratings(Moriasi et al. 2007)

Parameter	Expression	Level of performance				
		 Very good	 Good	 Satisfactory	 Unsatisfactory	
NSE	$1 - \frac{\sum_{i=1}^N (Q_o - Q_p)^2}{\sum_{i=1}^N (Q_o - \bar{Q}_o)^2}$	$0,75 < NSE \leq 1$	$0,65 < NSE \leq 0,75$	$0,5 < NSE \leq 0,65$	$NSE \leq 0,5$	
PBIAS	$100 \cdot \frac{\sum_{i=1}^N (Q_o - Q_p)}{\sum_{i=1}^N Q_o}$	$PBIAS < \pm 10$	$\pm 10 \leq PBIAS < \pm 15$	$\pm 15 \leq PBIAS < \pm 25$	$PBIAS \geq \pm 25$	
RSR	$\frac{\sqrt{\sum_{i=1}^N (Q_o - Q_p)^2}}{\sqrt{\sum_{i=1}^N (Q_o - \bar{Q}_o)^2}}$	$0 \leq RSR \leq 0,5$	$0,5 < RSR \leq 0,6$	$0,6 < RSR \leq 0,7$	$RSR > 0,7$	

- Graphical performance:

The evaluation of the model's accuracy can be done graphically using **the Q-Q plot** graphs presenting the observed and the predicted values compared to the line equation $y=x$ to identify the correlation between the two variables by its closeness degree to the line.

The hydrographs containing both values at the same time (observed and predicted) are also used to observe the over or underestimation of the model.

4.4 Methodology summary

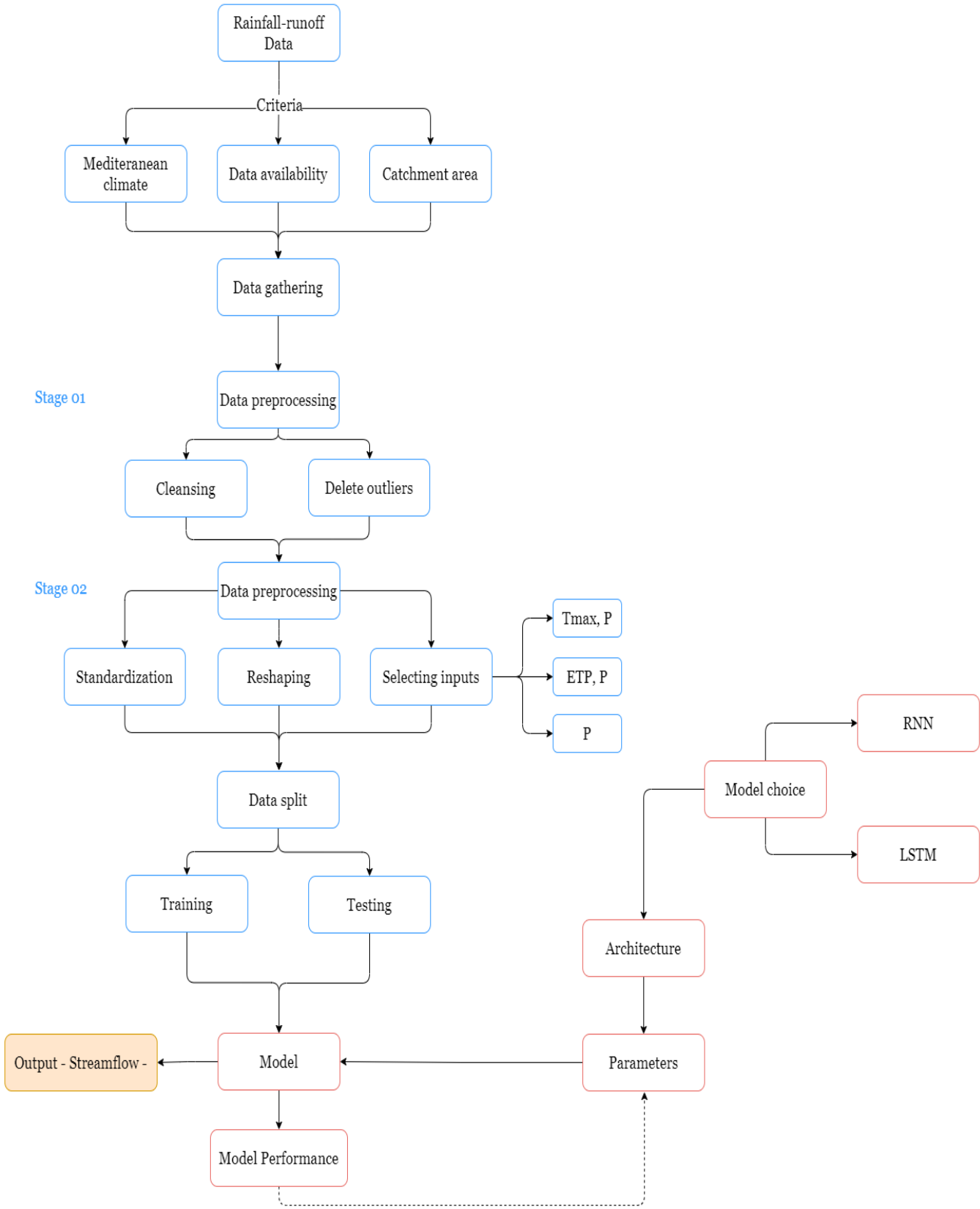


Figure 4-5 Methodology summary

4.5 Conclusion

This chapter represents the core work of the thesis. It summarizes the applied tools and the followed methodology for building our model. Overall, we used Jupyter-notebook to run our python-code with its various libraries for the purpose of data treatment and the development of deep-learning models that makes short-term rainfall-runoff predictions.

Chapter 05 Results & discussion

Introduction

In this chapter, we will apply the methodology presented in the previous chapter, namely the RNN and LSTM model using different input parameters, to observe their influence on the model. The analytical results are accompanied by graphical representations allowing a direct comparison of the predicted and observed hydrographs. The purpose of the performance comparison between the two model types is to know which one is the best.

This application of deep learning to model the Rainfall-Runoff is a first in Algeria.

5.1 Statistical parameters for numerical and graphical performance

To evaluate the level of performance of the developed models, the following statistical parameters (defined in 4.3.6) were used to compare the results with the given data:

- NSE: The Nash–Sutcliffe model efficiency
- RSR: standard deviation ratio RMSE
- PBIAS: Percent bias

The graphical performance is evaluated by the equation-line $y=x$ on the Q-Q plot graphs and the hydrographs representing the observed and predicted values.

5.2 The model Inputs

The model receives each time two parameters: the rainfall data and another one. The second input was essentially the evapotranspiration which is the most representative parameter related to the physical phenomenon. Then, since the access to this parameter is quite difficult and hard to find in each region, we replace it with temperature data and test the two cases. Not to forget, that the RNN and LSTM models automatically use the previous streamflow data as an input too.

To know the precipitation influence degree on our model's performance, we decided to keep only the precipitation datasets as input and compare it to the other studies.

Table 5.1 : The model Inputs

$F(T_{max}, P, Q_{prev}) = Q$	$F(ETP, P, Q_{prev}) = Q$	$F(P, Q_{prev}) = Q$
Maximum temperature and precipitation	Evapotranspiration and precipitation	Precipitation

5.3 Results

5.3.1 RNN Model

The results of the Rainfall-Runoff modeling using the RNN model with the 3 different inputs within the studied stations are presented through the statistical parameter: NSE in the table below Table 5.2:

Table 5.2: Statistical parameters of our results (RNN)

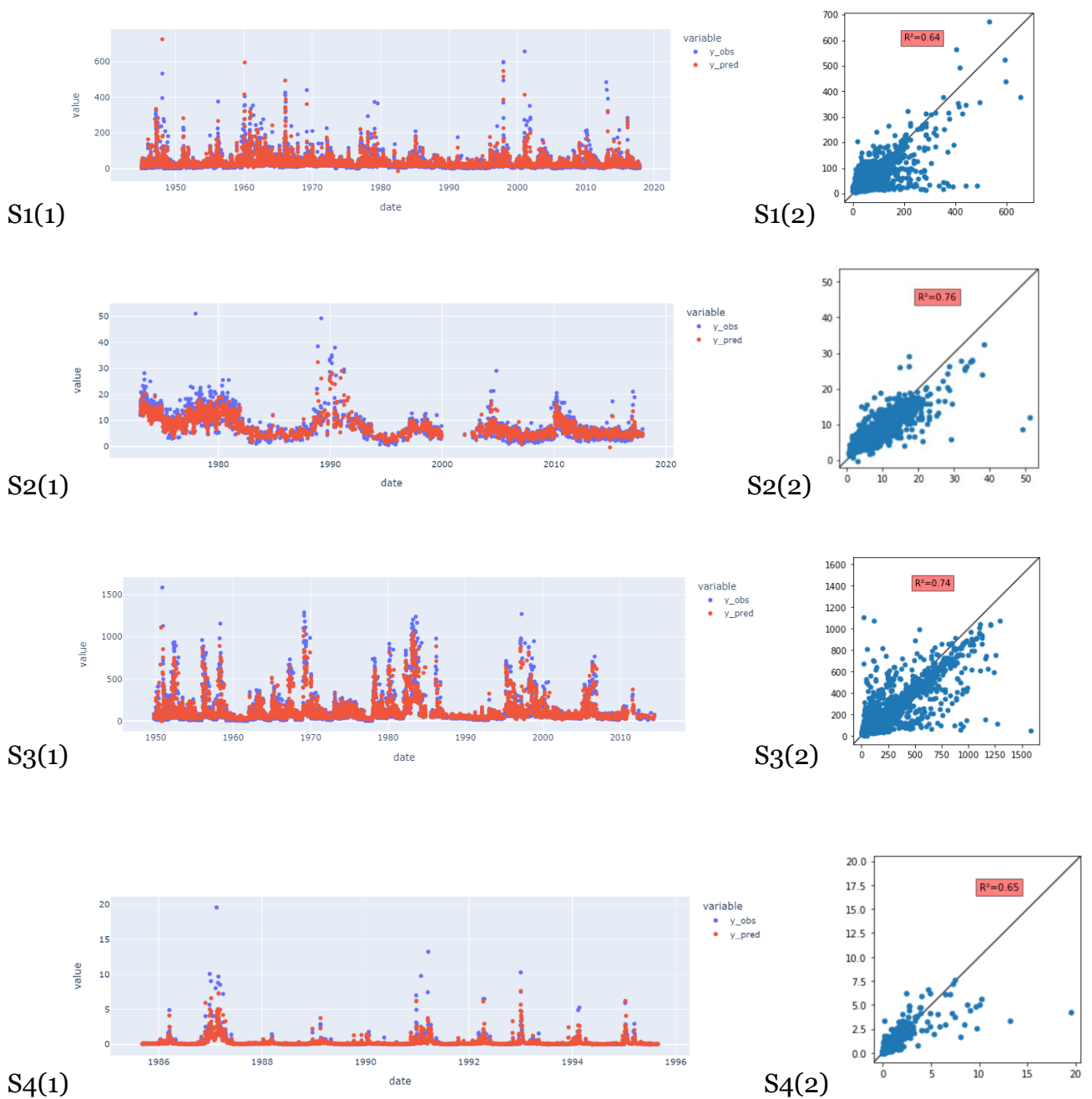
		Study 01		Study 02		Study 03	
Inputs	Country	Tmax ,P, Q _{prev}		ETP,P, Q _{prev}		P, Q _{prev}	
NSE(%)		Train	Test	Train	Test	Train	Test
S1	Spain	60%	64%	64%	64%	66%	65%
S2	Spain	72.5%	73.3%	76%	76.2%	77.6%	77.6%
S3	USA	73.4%	73.5%	73.5%	73.6%	74.5%	75.1%
S4	Boucheougouf	54.4%	57.2%	63.3%	64.7%	56%	56%
S5	Zardezas	73.4%	73%	78.75%	74%	75%	76.6%

Table 5.2 represents NSE results for our 5 stations in 3 studies showing statistics for training and testing periods using RNN model. As shown in red, S4 performed poorly compared to other stations with an NSE<60. While S1 had a medium performance of NSE not exceeding 66% as a max value. The best results were recorded for S2, S3, and S5 with S2 outperforming the last two with a slight difference of approximately 2-3% for both training and testing. Surprisingly, S5 referring to Zardezas gave good results although the overfitting problems faced during the coding part contrary to the other stations. We managed eventually to find suitable hyper-parameters for the model to avoid these problems.

Comparing the 3 studies, we noticed that relying on precipitation only as an input gave better results than the other 2 studies. Not to forget that the physics-based approach (2nd study) gave similar results as the 1st study based on temperature. However, the difference between the 3 studies stays minor, ranging from 1 to 3%. Bouchegouf remains an exception in this, ETP results were the outstanding ones.

5.3.1.1 Graphical performance

The output hydrographs comparing the observed and predicted values using the RNN model in the 2nd study (ETP, P, Q_{prev}) are presented below:



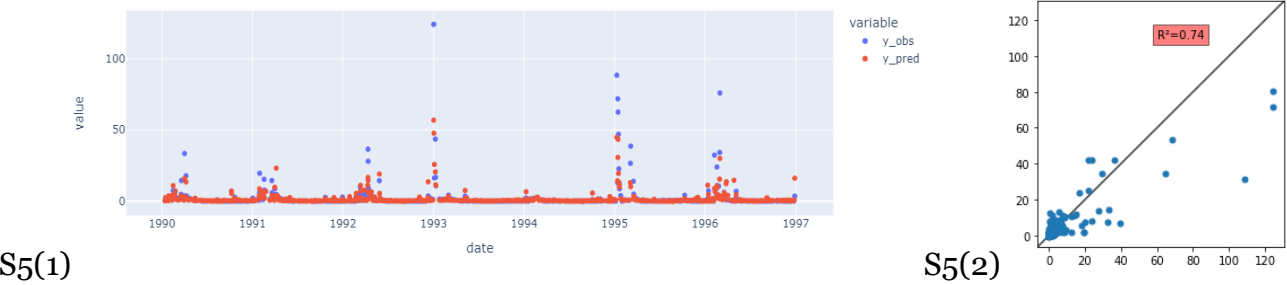


Figure 5-1 : Comparison of the predicted values and the observed values using ETP and precipitation inputs for the 5 regions during the test period of the RNN model. **S1(1), S2(1), S3(1), S4(1), and S5(1)** are the predicted and observed values for S1 to S5. **S1(2), S2(2), S3(2), S4(2), and S5(2)** are the scatter plots with the trendline of the predicted and observed values for S1 to S5.

As illustrated in figure, the 5 stations can be classified into 3 categories depending on the streamflow values: S4 low flow, S2, and S5 medium flow, S1 and S3 high flow. The RNN model underestimates the flow peaks in most cases, as some of these peaks are outliers that haven't been detected before like shown in (S1(1), S2(1), S3(1), S4(1), S5(1)). This is further explained in the scatter plots (S1(2), S2(2), S3(2), S4(2), and S5(2)) where we demonstrate the correlation between the observed and predicted data with the trendline. The unpredicted peaks have a direct influence on our model's performance.

The results in S2, S3, and S5 present a better correlation than the rest of the stations which explains the high value of NSE (NSE > 70%) and indicates the good performance of the fit regarding their fluctuations.

5.3.2 LSTM model

The results of the Rainfall-Runoff modeling using the LSTM model with the 3 different inputs within the studied stations are presented through the statistical parameter NSE, results are given in the table below:

- **NSE**

Table 5.3: Statistical parameters of our results (LSTM)

Inputs	Country	Study 01		Study 02		Study 03	
		Tmax ,P, Q _{prev}		ETP,P, Q _{prev}		P, Q _{prev}	
NSE(%)		Train	Test	Train	Test	Train	Test
S1	Spain	64%	66%	64%	64%	65.1%	64%
S2	Spain	79.7%	79.2%	75.3%	78%	76.2%	79.2%
S3	USA	80.2%	80.9%	76.9%	76.9%	77.2%	77.6%
S4	Boucheougouf	67.7%	68.9%	64.3%	64.7%	65%	65%
S5	Zardezias	79.2%	79.5%	78.75%	80%	79%	79.4%

Table 5.3 represents NSE results for our 5 stations in 3 studies showing statistics for training and testing periods using the LSTM model. As shown, S1 performed poorly compared to other stations with an NSE<66%. While S4 had an average performance of NSE not exceeding 69% as a max value. The best results were recorded for S2, S3, and S5 with S3 outperforming the last two with a slight difference of approximately 2-3% in the 1st study using Temperature and precipitation inputs with NSE of 80%. S5 referring to Zardezias gave good results although the overfitting problems faced during the coding part contrary to the other stations.

Comparing the 3 studies, S1 and S2 remained mainly the same for the testing period, but we noticed that relying on precipitation and the temperature as inputs for the case of S5 gave better results than the other 2 studies ranging from 1 to 6%. However, the difference between the 3 studies stays minor.

5.3.2.1 Graphical performance

The output hydrographs comparing the observed and predicted values using the LSTM model in the 2nd study (ETP, P, Q_{prev}) are presented below:

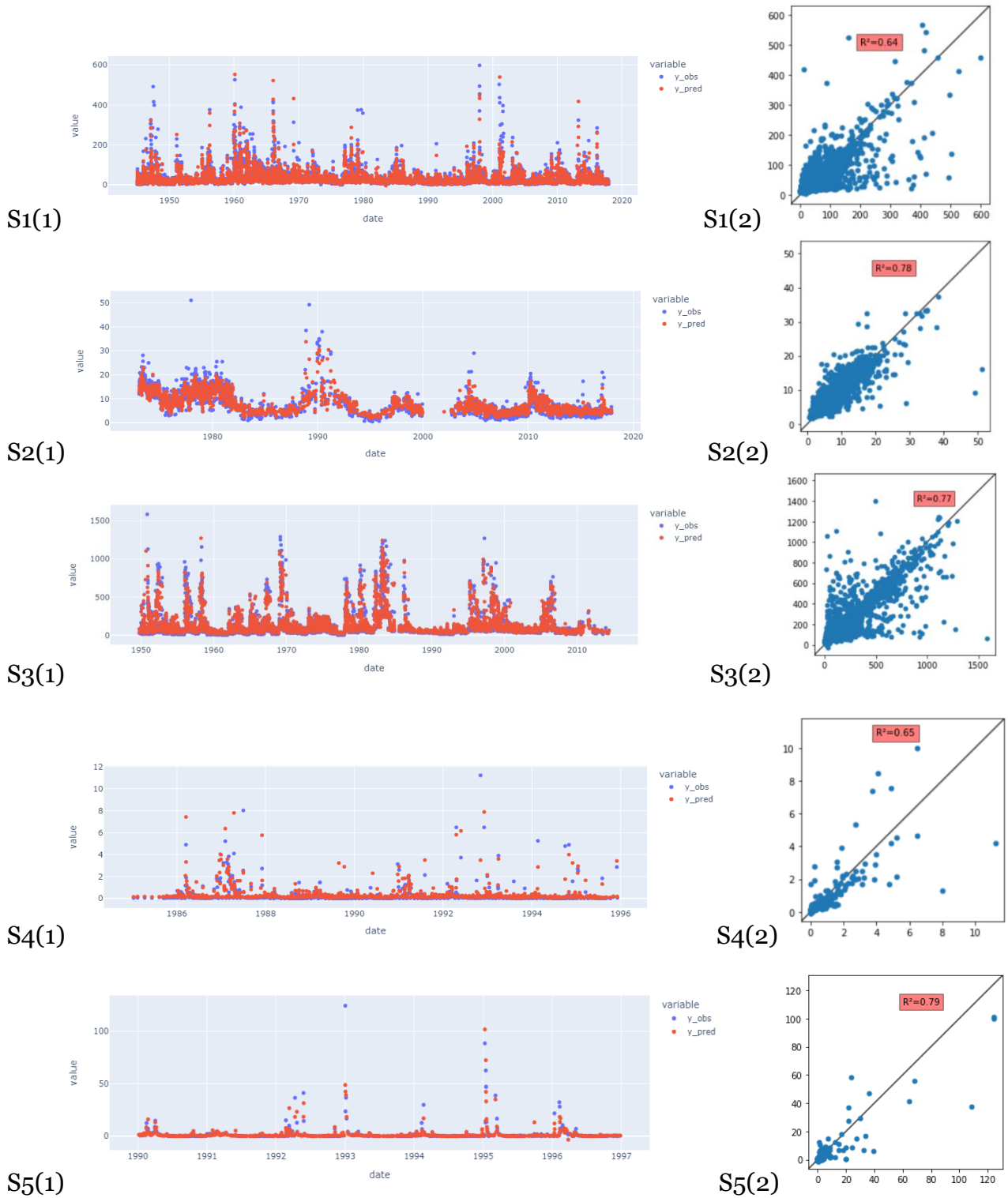


Figure 5-2: Comparison of the predicted values and the observed values using ETP and precipitation inputs for the 5 regions during the test period of the LSTM model. **S1(1)**, **S2(1)**, **S3(1)**, **S4(1)**, and **S5(1)** are the predicted and observed values for S1 to S5. **S1(2)**, **S2(2)**, **S3(2)**, **S4(2)**, and **S5(2)** are the scatter plots with the trendline of the predicted and observed values for S1 to S5.

5.3.3 Comparison between RNN and LSTM

As shown in Table 5.2 and Table 5.3, the values of each LSTM metric improved as compared to the RNN model in each region for both training and testing data. The NSE values of LSTM for training ranged between 0.6 and 0.8, and for testing, they ranged between 0.5 and 0.8. Both models had relatively good NSE values, with LSTM outperforming RNN for training and testing samples. It means that the LSTM predicted streamflow volumes were close to the observed values.

The graphical results show that the LSTM model's streamflow predictions were close to the observations.

The comparison between RNN and LSTM model in the 3 studies is illustrated in the next bar plots:

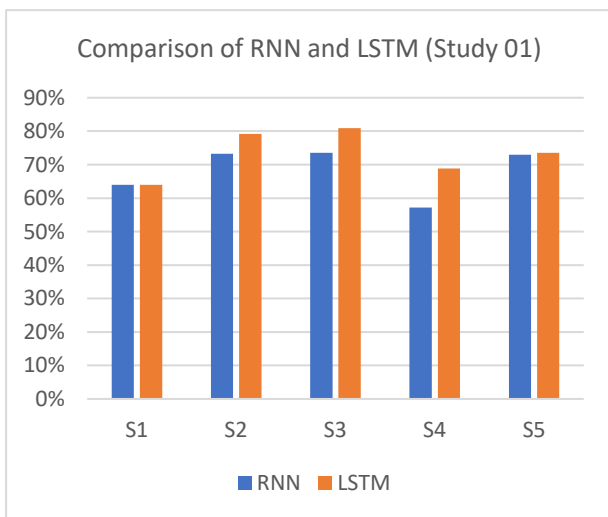


Figure 5-3 Comparison of RNN and LSTM (Study 01)

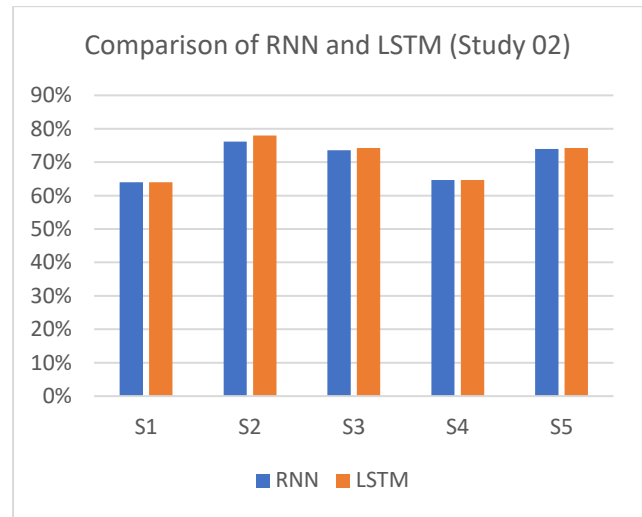


Figure 5-4 Comparison of RNN and LSTM (Study 02)

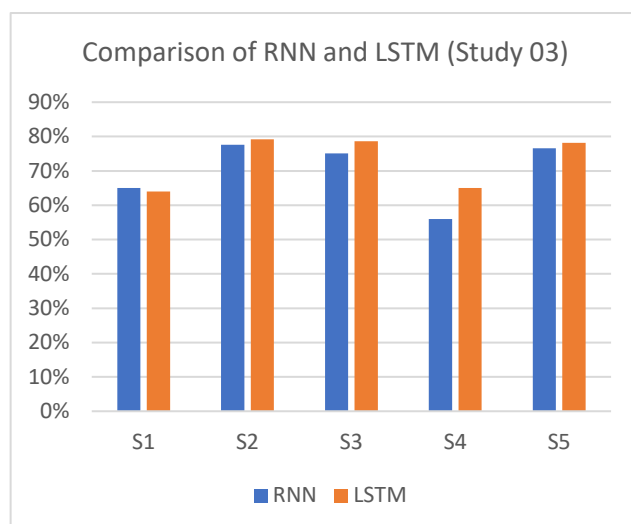


Figure 5-5 Comparison of RNN and LSTM (Study 03)

Since the NSE of the LSTM model has great values, indicating a solid linear relationship between the observed and predicted values, we present the remaining parameters below, as further studies on the model, which quantify the deviation in the units of our data to see the influence of the extreme values our model’s stability and whether the model underestimates or overestimates the peaks.

The results of the Rainfall-Runoff modeling using the LSTM model with the 3 different inputs within the studied stations are presented through the statistical parameter: PBIAS, RMSE and RSR as given in the tables below:

- PBIAS:

Table 5.5 PBIAS Values

		Study 01		Study 02		Study 03	
Inputs	Country	Tmax ,P, Q _{prev}		ETP,P, Q _{prev}		P, Q _{prev}	
		Train	Test	Train	Test	Train	Test
S1	Spain	-10.67	-10.40	5.13	4.12	7.19	6.14
S2	Spain	0.35	0.54	-0.16	-0.37	-4.63	-4.63
S3	USA	4,19	2,89	-10,3	-9,56	0,84	1,75
S4	Boucheougouf	13.45	12.07	17.19	11.76	14.32	13.72
S5	Zardezas	-11.02	-2.53	-10.64	-1.24	-19.05	-0.98

The overall PBIAS values has improved in the testing compared to the training.

Most importantly, the testing values fit in the very good estimation category, close to 0, where positive results indicate under-estimation and negative results indicate over-estimation. This notice doesn't apply to S4 Bouchegouf whose values are underestimated with a larger difference.

- RMSE:

Table 5.4 RMSE Values

Inputs	Country	Study 01		Study 02		Study 03	
		Tmax ,P, Q _{prev}		ETP,P,Q _{prev}		P, Q _{prev}	
RMSE		Train	Test	Train	Test	Train	Test
S1	Spain	25.09	22.63	24.04	24.87	24.28	24.99
S2	Spain	2.31	2.39	2.55	2.35	2.57	2.34
S3	USA	82,04	76,87	83,81	85,56	81,08	87,48
S4	Bouchegouf	3.47	3.2	4.14	2.9	3.5	2.74
S5	Zardezas	2.76	3.95	2.38	4.01	2.29	4,19

RMSE is reported in the same units as the model output (m3/s). It presents some high values since it penalizes large errors. However, for a better understanding and better interpretation of these values, we resorted to calculating RSR as a standardization of RMSE values:

- RSR:

Table 5.5 RSR Values

Inputs	Country	Study 01		Study 02		Study 03	
		Tmax ,P, Q _{prev}		ETP,P, Q _{prev}		P, Q _{prev}	
RSR		Train	Test	Train	Test	Train	Test
S1	Spain	0.62	0.56	0.58	0.60	0.59	0.60
S2	Spain	0.43	0.45	0.50	0.46	0.50	0.46
S3	USA	0.48	0.44	0.47	0.48	0.45	0.49
S4	Boucheougouf	0.70	0.65	0.84	0.59	0.71	0.56
S5	Zardezaz	0.3	0.45	0.26	0.44	0.25	0.46

We notice that RSR results are in correlation with NSE results presented in the table above in terms of station performance. RSR adds to these results by indicating the peaks performance.

S2 and S3 and S5 continue to give a great performance with an RSR equals to or inferior to 0,5, which falls into the very good category. This indicates a very good simulation performance overall, with a good peak prediction. The model was able to predict the peaks properly.

S1 stands in the good category, while S4 Boucheougouf has demonstrated mediocre results in the training period, that have improved a bit in testing to stand in satisfactory for study case 1 and good for case 2 and 3.

5.4 Evaluation recap for the 3 studies (LSTM)

To observe clearly the difference between the results of the LSTM model in each study case, we made this tables to compare:

Study (01) : Tmax,P, Q_{prev}


Table 5.6 Statistical parameters Study (01)

Station	Statistical parameters		
	NSE	PBIAS	RSR
S1	0.66	-10.4	0.56
S2	0.79	0.54	0.45
S3	0.81	2.89	0.44
S4	0.68	12.07	0.65
S5	0.73	-2.53	0.45

 Very good

 Good

 Satisfactory

 Unsatisfactory

Study(2): ETP,P, Q_{prev}

Table 5.7 Statistical parameters study (02)

Station	Statistical parameters		
	NSE	PBIAS	RSR
S1	0.64	4.12	0.60
S2	0.78	-0.37	0.46
S3	0.77	-9.56	0.48
S4	0.65	11.76	0.59
S5	0.8	-1.24	0.44

Study(3): P, Q_{prev}

Table 5.8 Statistical parameters study (03)

Station	Statistical parameters		
	NSE	PBIAS	RSR
S1	0.64	6.14	0.60
S2	0.79	-4.63	0.46
S3	0.77	1.75	0.49
S4	0.65	13.72	0.56
S5	0.79	-0.98	0.46

We can say regarding Table 5.6, Table 5.7 and Table 5.8 that overall, all stations kept the same level of performance in each study with small differences between the values.

S2, S3, S5 represent the stations in which our model performed the greatest, NSE, PBIAS, RSR indicated positive results in all of them. The model was able to fit the predicted values to the observed ones with very close over- or under-estimation values. It could also fit the peaks well.

S1 and S4 had lower fitting results seeing their NSE values, they couldn't get the hydrographs shapes properly, while it could approximate some peaks.

5.5 Discussion

We compared the performance of two deep learning models (RNN and LSTM) applied to 5 different hydrometric stations through daily observations studying the runoff response of the watersheds. The study cases we worked on varies in term of flow rate, from low, and medium to high streamflow. The differentiation was important to check the model's performance on different river types.

The data gathered from the five cases studied range in observation period from 7 to 100 years. This allowed us to test various types of cases and examine how the time scale affected the model's performance. Despite the short amount of data, we had good

results in S5, but not so much for S2, which has a big number of observations. Consequently, while it is advantageous to have large datasets to ensure the model learning rate, it can be applied to small datasets (few years) and still expect good outcomes.

Moreover, while treating the data used, we faced several problems, such as the missing values and the outliers. In most cases, we have had many missing values which led us to lose inputs' information that could have helped the model's performance. Furthermore, the peaks of hydrological events are difficult to identify and manage (true values or outliers) because removing these extreme values can probably underestimate true values that had occurred. So, we tried to analyze and detect as many values as possible and search them up historically. Then, to evaluate the model performance regarding these peaks, we relied on PBIAS to give us an approximation error of their values to identify the occurred under-, over-estimation, where our model has proven its efficiency in capturing these peaks.

In the matter of checking the first established hypothesis in our studies: whether the model needs to be physics-based. We compared the results of (Tmax, P, Q_{prev}) and (ETP, P, Q_{prev}) inputs and figured that there's no significant difference between the two models. Instead, the input (P, Q_{prev}) is what makes the difference in the results. The predicted flow strongly depends on (P, Q_{prev}) while Tmax and ETP can add a slight contribution to the performance.

5.6 Conclusion

The study has allowed us to see the efficiency of deep learning models on hydrological phenomena that are generally quite complex and difficult to predict using traditional linear methods. We created a model for Mediterranean climate data with daily steps to analyze the hydrological risks such as floods. Then, we extended our experiments by employing multiple input parameters to examine how the model responded to both Algerian and foreign datasets.

The LSTM model, a developed version of RNN, provided better results.

To get coherent data, we had to ensure that the measured rainfall contributes directly to the runoff, a task that had been difficult in our study. We initially worked on 12 main hydrometric stations where 7 have been eliminated in the process because of

this matter. The model performed poorly in these cases since it couldn't elaborate on the relation between input and output. Consequently, it is important to check the stations' locations before proceeding to the modeling part. Another important note is that even the station's surroundings have to be verified. As an example, the USA has a considerable large amount of data (more than 100 years) which would be perfect for modeling, but the existing dams in the hydrometric station area made it difficult to estimate the original streamflow volume without including dam release.

General conclusion

The rainfall-runoff modeling is a difficult and important nonlinear time-series problem in hydrology. It represents an essential means for flood prediction and water resources management. Algeria, being one of the many countries facing flood events, in many regions, makes it necessary to make effective streamflow predictions.

The present study aimed to develop and test deep learning models (RNN and LSTM) in several foreign and local regions with a Mediterranean climate to predict the runoff in a daily manner. To evaluate the impact of the inputs on the model's performance, daily datasets of different inputs such as temperature, evapotranspiration, and rainfall data were implemented. The obtained results were rated with numerical and graphical statistical parameters.

This study has demonstrated the ability and utility of deep learning models in daily-term prediction to find nonlinear relations between the input and the output. It proved that the model does not need a thorough grasp of a catchment's physical properties, nor does it necessitate substantial data preprocessing. The prediction has been made without recurring to geophysical characteristics which can be usually hard to provide.

We came up with the conclusion that the real influencing factors on predicted flow are precipitation and precedent streamflow. Inserting temperature or evapotranspiration as additional inputs would improve the model's results slightly (~3%).

The quality of the data inserted into deep learning models has a significant impact on the results. Erroneous values can cause a noticeable decrease in the model's performance not permitting the model to determine the appropriate patterns between input and output.

The results of our study allowed us to compare the performance of the RNN and LSTM models in the 5 chosen regions for the 3 study cases. After reviewing the NSE values of the 2 models where LSTM outperformed RNN which is coherent with previous research theories. Thus, we recommend using LSTM to predict rainfall-runoff for better performance and accurate results.

The work done on 12 stations where 7 were eliminated showcases the fact that checking the stations' locations and surroundings has a crucial impact on the outcome

of the model. While in rare cases, the model was able to find the connection between the input/output although the unsatisfaction of the above conditions.

For the perspectives, we suggest:

- Deployment of the model on a real Algerian case at the time to predict floods.
- Improving the LSTM model's performance by using these coding techniques and algorithms:
 - Customize an LSTM model for each study case.
 - Using DTW (Dynamic Time Warping Network) to determine the rainfall station that has the greater influence on the runoff.
 - Using Cross validation to avoid overfitting
 - Finding the optimum value of the lag time in the model by making several experiences using ACF and ACPF (autocorrelation analysis)

Bibliography

About | National Centers for Environmental Information (NCEI),
<https://www.ncei.noaa.gov/about>

ABRAHART, R. J. et SEE, L. M., 2007. Neural network modelling of non-linear hydrological relationships. *Hydrology and Earth System Sciences*. Vol. 11, n° 5, pp. 1563-1579.
DOI 10.5194/hess-11-1563-2007.

AMERICAN RIVERS, 2017. The San Joaquin Demonstrates The Importance Of Floodplain Restoration. *American Rivers* <https://www.americanrivers.org/2017/02/flooding-san-joaquin-floodplain-restoration/>

Anaconda (distribution Python), 2022.
[https://fr.wikipedia.org/w/index.php?title=Anaconda_\(distribution_Python\)&oldid=192837269](https://fr.wikipedia.org/w/index.php?title=Anaconda_(distribution_Python)&oldid=192837269)

BACCOUCHE, Moez, MAMALET, Franck, WOLF, Christian, GARCIA, Christophe et BASKURT, Atilla, 2011. Sequential Deep Learning for Human Action Recognition. In : SALAH, Albert Ali et LEPRI, Bruno (éd.), *Human Behavior Understanding*. Berlin, Heidelberg : Springer. 2011. pp. 29-39. Lecture Notes in Computer Science. ISBN 978-3-642-25446-8.

BAI, Yun, CHEN, Zhiqiang, XIE, Jingjing et LI, Chuan, 2016. Daily reservoir inflow forecasting using multiscale deep feature learning with hybrid models. *Journal of Hydrology*. Vol. 532, pp. 193-206. DOI 10.1016/j.jhydrol.2015.11.011.

BAJAJ, Aayush, 2021. Performance Metrics in Machine Learning [Complete Guide]. *neptune.ai*. <https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide>

BEVEN, Keith J., 2011. *Rainfall-Runoff Modelling: The Primer*. John Wiley & Sons. ISBN 978-1-119-95101-8.

BfG - The GRDC, https://www.bafg.de/GRDC/EN/Home/homepage_node.html

BIRIKUNDAVYI, S., LABIB, R., TRUNG, H. T. et ROUSSELLE, J., 2002. Performance of Neural Networks in Daily Streamflow Forecasting. *Journal of Hydrologic Engineering*. Vol. 7, n° 5, pp. 392-398. DOI 10.1061/(ASCE)1084-0699(2002)7:5(392).

BOUHOUN, Idris, 2020. Modélisation de la relation Pluie-Débit par les modèles: GR4J, Tank et Multi-Tank couplé au filtre de Kalman (Application à deux bassins versants de climats différents). *MCB U*. 2020. pp. 94.

BURKOV, Andriy, 2019. The hundred-page machine learning book. <https://book.africa/book/3710356/c888od>

CFI TEAM, 2021. Python (in Machine Learning). *Corporate Finance Institute*. <https://corporatefinanceinstitute.com/resources/knowledge/other/python-in-machine-learning/>

Chapter 2 - History of Flooding and Flood Protection, 1983. . pp. 25.

CHUGH, Akshita, 2022. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Analytics Vidhya*. <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>

CORTES, Rui, TERÊNCIO, Daniela, MOURA, João, JESUS, Joaquim, MAGALHÃES, Marco, FERREIRA, Pedro et PACHECO, Fernando, 2019. Undamming the Douro River Catchment: A Stepwise Approach for Prioritizing Dam Removal. *Water*. Vol. 11, n° 4, pp. 693. DOI 10.3390/w11040693.

COULIBALY, P., ANCTIL, F. et BOBÉE, B., 2000. Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology*. Vol. 230, n° 3, pp. 244-257. DOI 10.1016/S0022-1694(00)00214-6.

DANIEL, Edsel B., 2011. Watershed Modeling and its Applications: A State-of-the-Art Review. *The Open Hydrology Journal*. Vol. 5, n° 1, pp. 26-50. DOI 10.2174/1874378101105010026.

DATA FLAIR, 2020. How Deep Learning Works with Different Neuron Layers. *DataFlair* . <https://data-flair.training/blogs/how-deep-learning-works/>

DAWSON, C. W. et WILBY, R. L., 2001. Hydrological modelling using artificial neural networks. *Progress in Physical Geography: Earth and Environment*. mars 2001. Vol. 25, n° 1, pp. 80-108. DOI 10.1177/030913330102500104.

Deep learning, 2022. *Wikipedia* . https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=1088786364

DEVIA, Gayathri, BIGGANAHALLI PUTTASWAMIGOWDA, Ganasri et DWARAKISH, G.S., 2015. A Review on Hydrological Models. *Aquatic Procedia*. 31 décembre 2015. Vol. 4, pp. 1001-1007. DOI 10.1016/j.aqpro.2015.02.126.

DE VOS, N. J. et RIENTJES, T. H. M., 2005. Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth System Sciences*. 5 juillet 2005. Vol. 9, n° 1/2, pp. 111-126. DOI 10.5194/hess-9-111-2005.

DONAHUE, Jeff, HENDRICKS, Lisa, GUADARRAMA, Sergio, ROHRBACH, Marcus, VENUGOPALAN, Subhashini, SAENKO, Kate et DARRELL, Trevor, 2014. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *Arxiv*. 1 novembre 2014. Vol. PP. DOI 10.1109/TPAMI.2016.2599174.

Duero Hydrographic Confederation | Hispagua, <https://hispagua.cedex.es/en/instituciones/confederaciones/duero>

GAURAV, Singhal, 2020. Introduction to LSTM Units in RNN. <https://www.pluralsight.com/utilities/promo-only?noLaunch=true>

GERS, Felix A., SCHMIDHUBER, Jürgen et CUMMINS, Fred, 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*. Vol. 12, n° 10, pp. 2451-2471. DOI 10.1162/089976600300015015.

GOSWAMI, Divyang, 2020. Comparison of Sigmoid, Tanh and ReLU Activation Functions. *AITUDE* . <https://www.aitude.com/comparison-of-sigmoid-tanh-and-relu-activation-functions/>

GRAVES, Alex et JAITLEY, Navdeep, 2014. Towards End-To-End Speech Recognition with Recurrent Neural Networks. In : *Proceedings of the 31st International Conference on Machine Learning* . PMLR. <https://proceedings.mlr.press/v32/graves14.html>

GRUBBS, Frank E., 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics*. Vol. 11, n° 1, pp. 1-21. DOI 10.1080/00401706.1969.10490657.

GUPTA, Hoshin Vijai, SOROOSHIAN, Soroosh et YAPO, Patrice Ogou, 1999. Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration. *Journal of Hydrologic Engineering*. Vol. 4, n° 2, pp. 135-143. DOI 10.1061/(ASCE)1084-0699(1999)4:2(135).

HALFF, Albert H., HALFF, Henry M. et AZMOODEH, Masoud, 1993. Predicting Runoff from Rainfall Using Neural Networks. In : *Engineering Hydrology*. <https://cedb.asce.org/CEDBsearch/record.jsp?dockkey=0083448>

HAMID, Moradkhani et SOROOSHIAN, Soroosh, 2008. General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis, in hydrological modeling and water cycle, coupling of the atmospheric and hydrological models. *Water Sci. Technol. Libr.* 1 janvier 2008. Vol. 63, pp. 1-23.

HARO, David, SOLERA, Abel, PEDRO-MONZONÍS, María et ANDREU, Joaquín, 2014. Optimal Management of the Júcar River and Turia River Basins under Uncertain Drought Conditions. *Procedia Engineering*. Vol. 89. DOI 10.1016/j.proeng.2014.11.432.

HSU, Kuo-lin, GUPTA, Hoshin Vijai et SOROOSHIAN, Soroosh, 1995. Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*. Vol. 31, n° 10, pp. 2517-2530. DOI 10.1029/95WR01955.

HU, Caihong, WU, Qiang, LI, Hui, JIAN, Shengqi, LI, Nan et LOU, Zhengzheng, 2018. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water*. Vol. 10, n° 11, pp. 1543. DOI 10.3390/w10111543.

IBM CLOUD EDUCATION, 2021. What are Neural Networks? <https://www.ibm.com/cloud/learn/neural-networks>

IBM CLOUD EDUCATION, 2022. What is Deep Learning? <https://www.ibm.com/cloud/learn/deep-learning>

IZERROUKYENE.Abelkader, 2017.

JANIESCH, Christian, ZSCHECH, Patrick et HEINRICH, Kai, 2021. Machine learning and deep learning. *Electronic Markets*. Vol. 31, n° 3, pp. 685-695. DOI 10.1007/s12525-021-00475-2.

Júcar Hydrographic Confederation | Hispagua. <https://hispagua.cedex.es/en/instituciones/confederaciones/jucar#uno>

Jupyter, 2022. *Wikipédia*. <https://fr.wikipedia.org/w/index.php?title=Jupyter&oldid=194206414>

KHELFAOUI, Fayçal et ZOUINI, Derradji, 2010. Gestion intégrée et qualité des eaux dans le bassin versant du Saf-Saf (wilaya de Skikda, nord-est algérien). . 2010. pp. 7.

KNAPP, H. Vernon, DURGUNOĞLU, Ali et ORTEL, Terry W., 1991. A review of rainfall-runoff modeling for stormwater management. *ISWS Contract Report CR 516*. 1991.

Köppen climate classification - Wikipedia. https://en.wikipedia.org/wiki/K%C3%B6ppen_climate_classification

Köppen Climate Classification System | National Geographic Society.<https://education.nationalgeographic.org/resource/koppen-climate-classification-system>

KOTTEK, Markus, GRIESER, Jürgen, BECK, Christoph, RUDOLF, Bruno et RUBEL, Franz, 2006. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*. Vol. 15, n° 3, pp. 259-263. DOI 10.1127/0941-2948/2006/0130.

KRATZERT, Frederik, KLOTZ, Daniel, BRENNER, Claire, SCHULZ, Karsten et HERRNEGGER, Mathew, 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*. Vol. 22, n° 11, pp. 6005-6022. DOI 10.5194/hess-22-6005-2018.

LEONELLI, Manuele et GÖRGEN, Christiane, 2017. Sensitivity analysis in multilinear probabilistic models. *Information Sciences*. Vol. 411, pp. 84-97. DOI 10.1016/j.ins.2017.05.010.

Linear Regression Example. *scikit-learn*. https://scikit-learn/stable/auto_examples/linear_model/plot_ols.html

MASON, J C, PRICE, R.K. et TEM'ME, A., 1996. A neural network model of rainfall-runoff using radial basis functions. *Journal of Hydraulic Research*. Vol. 34, n° 4, pp. 537-548. DOI 10.1080/00221689609498476.

Matplotlib – Visualization with Python. <https://matplotlib.org/>

MINNS, A. W. et HALL, M. J., 1996. Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal*. Vol. 41, n° 3, pp. 399-417. DOI 10.1080/02626669609491511.

MORIASI, Daniel, ARNOLD, Jeff, VAN LIEW, Michael, BINGNER, Ron, HARMEL, R.D. et VEITH, Tamie, 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*. Vol. 50. DOI 10.13031/2013.23153.

MORTON, F.I., 1983. Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology. *Journal of Hydrology*. Vol. 66, n° 1, pp. 1-76. DOI 10.1016/0022-1694(83)90177-4.

MOULAHOU, 2019 .

NaN, 2022. *Wikipedia*. <https://en.wikipedia.org/w/index.php?title=NaN&oldid=1092139284>

OCIO, D., BESKEEN, T. et SMART, K., 2019. Fully distributed hydrological modelling for catchment-wide hydrological data verification. *Hydrology Research*. Vol. 50, n° 6, pp. 1520-1534. DOI 10.2166/nh.2019.006.

LOUDIN, Ludovic, 2006. Une formule simple d'évapotranspiration potentielle pour la modélisation pluie-débit à l'échelle du bassin versant. *La Houille Blanche*. Vol. 92, n° 6, pp. 113-120. DOI 10.1051/lhb:2006109.

PAI, Aravindpai, 2020. ANN vs CNN vs RNN | Types of Neural Networks. *Analytics Vidhya* <https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/>

- Pandas, 2021. *Wikipédia* .
<https://fr.wikipedia.org/w/index.php?title=Pandas&oldid=188466798>
- PHI, Michael, 2020. Illustrated Guide to LSTM's and GRU's: A step by step explanation. *Medium*.<https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>
- Python (programming language), 2022. *Wikipedia*.
[https://en.wikipedia.org/w/index.php?title=Python_\(programming_language\)&oldid=1093611065](https://en.wikipedia.org/w/index.php?title=Python_(programming_language)&oldid=1093611065)
- RAKESH TANTY, TANWEER S. DESMUKH, et MANIT BHOPAL, 2015. Application of Artificial Neural Network in Hydrology- A Review. *International Journal of Engineering Research and*. Vol. V4, n° 06, pp. IJERTV4ISO60247. DOI 10.17577/IJERTV4ISO60247.
report.pdf. <https://pubs.usgs.gov/fs/1999/0073/report.pdf>
- REYES, Kate, 2022. What is Deep Learning and How Does It Works [Updated]. *Simplilearn.com* .<https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning>
- RINSEMA, Jan Gert, 2014. *Comparison of rainfall runoff models for the Florentine Catchment*. University of Twente.
- San Joaquin River, 2022. *Wikipedia*.
https://en.wikipedia.org/w/index.php?title=San_Joaquin_River&oldid=1085343777
- Scikit-learn, 2021. *Wikipédia* . <https://fr.wikipedia.org/w/index.php?title=Scikit-learn&oldid=181501393>
- SHAMSELDIN, Assad Y., 1997. Application of a neural network technique to rainfall-runoff modelling. *Journal of Hydrology*. Vol. 199, n° 3, pp. 272-294. DOI 10.1016/S0022-1694(96)03330-6.
- SINGH, Eshan, KUZHAGALIYEVA, Nursulu et SARATHY, S. Mani, 2022. Chapter 9 - Using deep learning to diagnose preignition in turbocharged spark-ignited engines. In : BADRA, Jihad, PAL, Pinaki, PEI, Yuanjiang et SOM, Sibendu (éd.), *Artificial Intelligence and Data Driven Optimization of Internal Combustion Engines* . Elsevier. pp. 213-237. ISBN 978-0-323-88457-0. <https://www.sciencedirect.com/science/article/pii/B9780323884570000059>
- SINGH, Jaswinder, KNAPP, H Vernon et DEMISSIE, Misganaw, 2004. Hydrologic Modeling of the Iroquois River Watershed Using HSPF and SWAT. 2004. pp. 24.
- SITTERSON, Jan, KNIGHTES, Chris, PARMAR, Rajbir, WOLFE, Kurt, MUCHE, Muluken et AVANT, Brian. An Overview of Rainfall-Runoff Model Types. pp. 30.
- SUDHEER, K. P., GOSAIN, A. K. et RAMASASTRI, K. S., 2002. A data-driven algorithm for constructing artificial neural network rainfall-runoff models. *Hydrological Processes*. 2002. Vol. 16, n° 6, pp. 1325-1330. DOI 10.1002/hyp.554.
- TensorFlow, <https://www.tensorflow.org/?hl=fr>
- TOTH, Elena et BRATH, Armando, 2007. Multistep ahead streamflow forecasting: Role of calibration data in conceptual and neural network modeling. *Water Resources Research* [en ligne]. 2007. Vol. 43, n° 11. DOI 10.1029/2006WR005383.
<https://onlinelibrary.wiley.com/doi/abs/10.1029/2006WR005383>

TWIN, Alexandra. How Overfitting Works. *Investopedia*.
<https://www.investopedia.com/terms/o/overfitting.asp>

U.S. DEPARTMENT OF THE INTERIOR BUREAU OF RECLAMATION, 2006. *Chapter 8: Sacramento and San Joaquin River Basins*. 2006.
<https://www.usbr.gov/climate/secure/docs/2016secure/2016SECUREREport-chapter8.pdf>

VAN, Song Pham, LE, Hoang Minh, THANH, Dat Vi, DANG, Thanh Duc, LOC, Ho Huu et ANH, Duong Tran, 2020. Deep learning convolutional neural network in rainfall–runoff modelling. *Journal of Hydroinformatics*. Vol. 22, n° 3, pp. 541-561.
DOI 10.2166/hydro.2020.095.

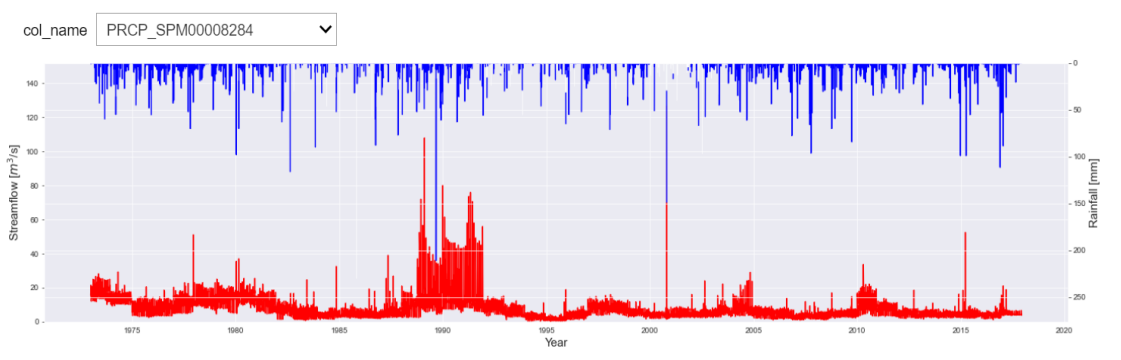
VAN, Song Pham, MINH LE, Hoang et THANH, Dat Vi, 2020. *Deep learning convolutional neural network in rainfallrunoff modelling*. 2020.

WAGENER, Thorsten, WHEATER, Howard et GUPTA, Hoshin Vijai, 2004. *Rainfall-runoff Modelling in Gauged and Ungauged Catchments*. World Scientific. ISBN 978-1-86094-466-6.

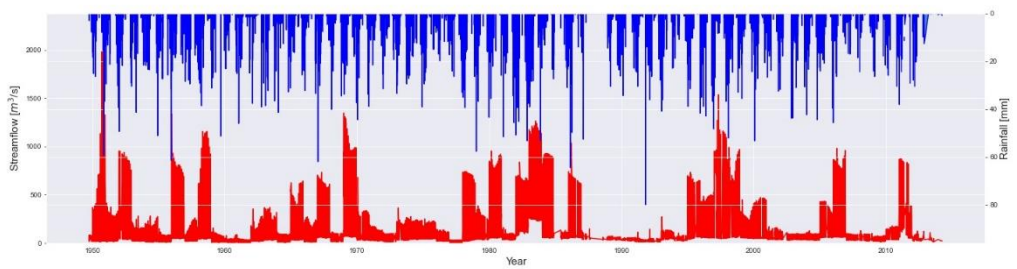
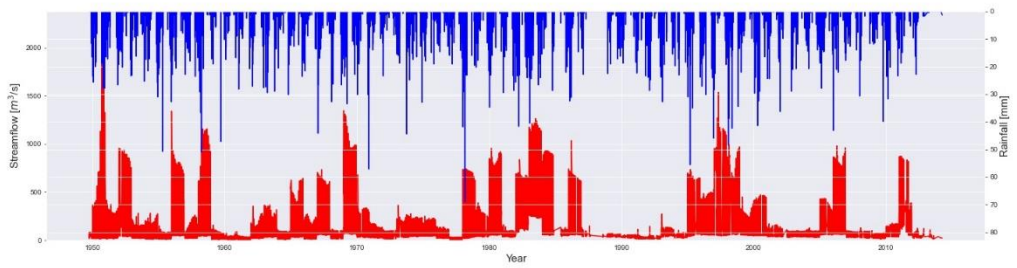
WASKOM, Michael, 2021. seaborn: statistical data visualization. *Journal of Open Source Software*. Vol. 6, n° 60, pp. 3021. DOI 10.21105/joss.03021.

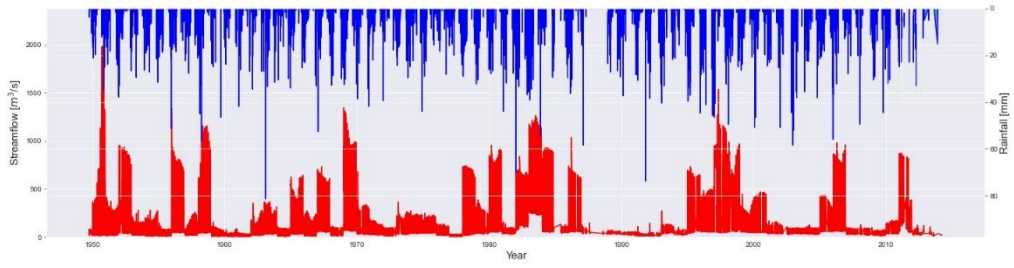
APPENDIX

Rainfall VS Runoff graphs

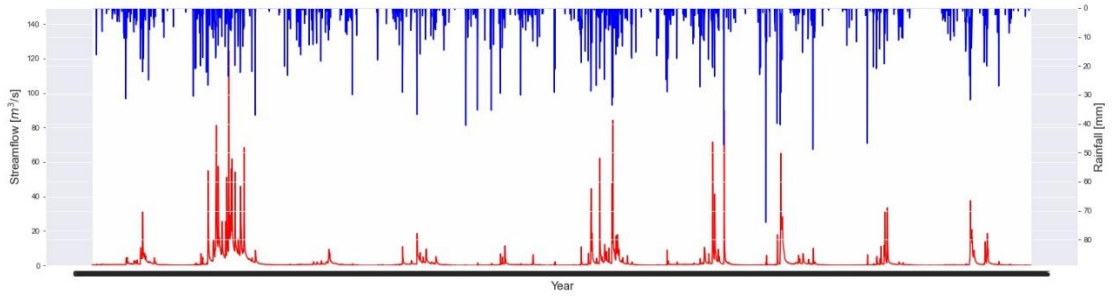
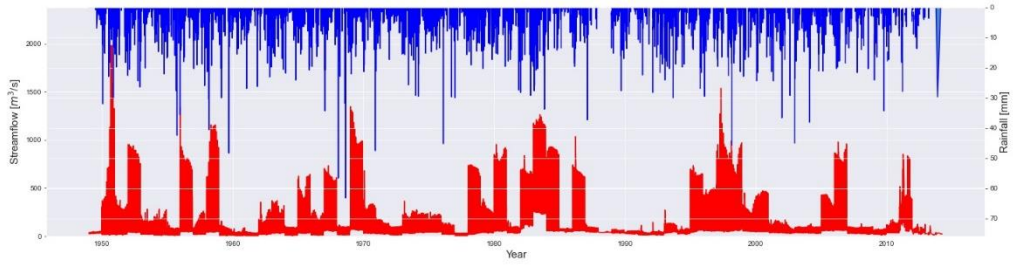


S (2)

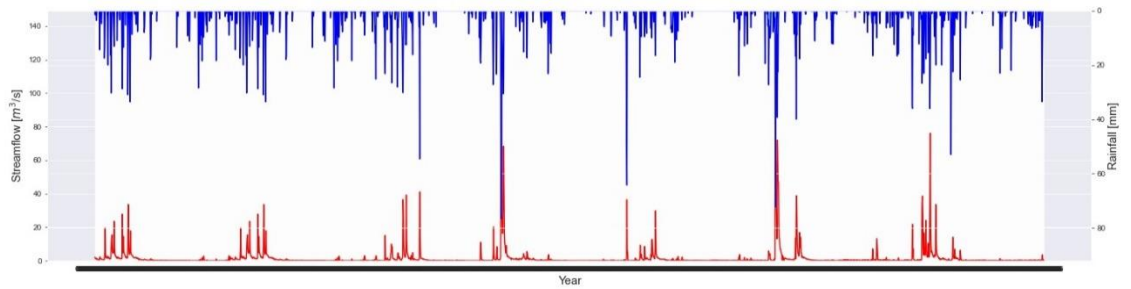




S(3)



S(4)



S(5)