Analysis of spatiotemporal rainfall objects in hydrological ensemble forecast predictions: Application in the Dapoling-Wangjiaba catchment in the Huai River Basin

Andrés Julián Ruiz Gomez

MSc Thesis Identifier WSE-HI.22-01
April 2022

Analysis of spatiotemporal rainfall objects in hydrological ensemble forecast predictions: Application in the Dapoling-Wangjiaba catchment in the Huai River Basin

Master of Science Thesis
by

Andrés Julián Ruiz Gomez

Supervisors

Dimitri Solomatine (UNESCO-IHE)

Mentors

Gerald Augusto Corzo Pérez (UNESCO-IHE)
Schalk Jan Van Andel (UNESCO-IHE)
Germán Ricardo Santos Granados (Escuela Colombiana de Ingeniería Julio Garavito)

Examination committee

Dimitri Solomatine (UNESCO-IHE)
Gerald Augusto Corzo Pérez (UNESCO-IHE)
Schalk Jan Van Andel (UNESCO-IHE)
Germán Ricardo Santos Granados (Escuela Colombiana de Ingeniería Julio Garavito)
Albrecht Weerts (Deltares)

This research is done for the partial fulfilment of requirements for the Master of Science degree at the
IHE Delft Institute for Water Education, Delft, the Netherlands and Escuela Colombiana de Ingeniería Julio Garavito, Bogotá, Colombia

Delft
19/04/2022

Abstract

Extreme rainfall events have been a constant throughout human history. Throughout history humanity has developed tools that have allowed us to be better prepared for these phenomena, aiming at longer lead times that can diminish their consequences. Forecasting has attempted to provide decision-makers with enough information, that leads to actions that can minimize economic costs and human losses during extreme events. However, these developed tools are not perfect, and gaps can be found in previous works, such as the lack of object-based approaches to identify upcoming critical events.

This study focuses its interest on the Dapoling-Wangjiaba sub-basin located in the upstream zone of the Huaihe River catchment, which is considered one major region in the country due to its economic importance and its high population density. The region has experienced some major flood events in the last years, and hence, the Huaihe river commission has raised interest in finding improvements for the already existing flood forecasting tools, to increase the reliability of models and to offer better preparedness for these events.

This research made use of three components: Ensemble Prediction System, Object-based methodology ST-CORA and HBV hydrological model. Ensemble rainfall forecasting data was acquired from ECMWF and then transformed into rainfall objects through ST-CORA, and then this information was routed through an HBV lumped conceptual model to establish a relationship between objects' characteristics and their hydrological response.

Through the proposed methodology it was possible to analyse extreme rainfall events, based on rainfall object characteristics, such as volume, intensity, duration and area, as well as its location's centroid. Clear spatial patterns over the catchment were observed that change with seasonality. It was found that major critical objects tend to be concentrated in the south zone of the catchment, specifically at latitude 32°N and between longitudes 114.5°W-115°, where 46% of all critical hydrological responses took place. 67.9% Of major critical events take place during the rainy season (June-July-August), while more intense hydrological responses tend to occur in the first 10 days of September, which can be attributed to the state of the catchment after a rainy season, meaning that extreme events are not fully dominated by high intensity rainfall. Finally, it was found that critical events are not characterised for its length in time or volume, such as in the last months of the year, but by its maximum voxel intensity, such as during rainy season, meaning that for the events studied in this research, voxel intensity is the decisive characteristic for high hydrological responses.

Finally, and aiming to provide ideas for future improvements in flood forecasting, a classification algorithm was trained, with the task of predicting the occurrence or not occurrence of high discharge based on object's characteristics. The results were acceptable (73% average precision) when only considering rainfall objects' properties such as: start date, centroid location, maximum voxel intensity, duration and volume, and they were almost perfect when adding observed discharge from three days before the start date as attribute to represent the state of the river. Object-based methodologies are much faster in terms of computing time than other approaches such as hydrological modelling, when used for predicting flood events. Based on the results of this research, further work on the precision of arrival peak time with the help of a distributed conceptual model is recommended, as well as exploring a weighting scheme to evaluate EPS perturbed members based on the object's properties contained in them and its expected hydrological response.

Acknowledgements

After almost two years, this amazing experience comes to an end. Many people were involved in this process. First, I would like to thank Mr Santos for allowing me to join this shared programme between La Escuela Colombiana de Ingeniería and IHE. Being here representing my alma mater is a true honour.

Secondly, I want to thank my supervisory team. To my supervisor Dimitri, I want to thank him, for his willingness to share his experience and vast knowledge with me, always with his particular sense of humour. Even before having him as a supervisor, I had the fantastic opportunity to have him as a lecturer, where I started to learn and hear about concepts that before were not even imaginable.

I had the opportunity to meet constantly with both of my mentors: Mr. Jan van Andel and Mr. Corzo. Even with their tight schedules they always found time to meet with me almost every week, giving me their thoughts, opinions, and feedback, which help me to understand a topic that six months ago was completely new to me. Our discussions, even if sometimes stressful and confusing, will become memories that are worth keeping. Besides that, I want to thank them for all their support during challenging times. I always felt supported by them. They were always there to listen, even with non-academic related matters. To both, I am truly thankful for all what you did for me.

To my dearest friend, Mafe, I do not have enough words to thank you for your support. You are the proof that neither distance nor time will be an obstacle for being there when the people you love need you.

I want to thank my parents, for all their arduous work that led me here. Thanks for all your sacrifices. This is for them.

I made many friends during these two years. But three people will be always in my heart: Alejandra, Betina and Danna. You made a new and strange place feel like home. Our memories will always be remembered with a smile.

Table of Contents

List of Figures

List of Tables

Abbreviations

DDM – Data Driven Model

ANN – Artificial Neuronal Network

ST-CORA – Spatial Temporal Contiguous Object-based Rainfall Analysis

ECMWF – European Centre for Medium-Range Weather Forecasts

EPS – Ensemble Prediction System

GloFAS – Global ensemble streamflow forecasting and flood early warning system

EFAS – European Flood Awareness System

HAPI – Hydrological library for Python

MASL – Meters Above Sea Level

THORPEX – The Observing System Research and Predictability Experiment

TIGGE – THORPEX Interactive Grand Global Ensemble/ The International Grand Global Ensemble

HBV – Hydrologiska Byrans Vattenbalansavdelning

CHIRPS - Climate Hazards Group Infrared Precipitation Satellite

EPSG – European Petroleum Survey Group

Chapter 1 Introduction

1.1 Background

Rainfall events are something with which humankind has dealt since the beginning of times. Even with all the advances and technologies developed in the last century, extreme rainfall events still represent a source of hazard as it could be seen during the floods in Europe during the summer of 2021 (Ibebuchi, 2022).Unfortunately, these flood events tend to cause bigger impacts in developing nations due to the lack of adequate infrastructure and resources (World Bank, 2020). According to the World Bank (2020), even if all countries face at some point exposure to flood risk, 89% of the affected population are in low- and middle-income countries, representing 600 million people.

Considering the economical and human lives lost, that these events represent, it can be easily seen the importance of being prepared for them. This preparation is not only represented in structural measures such as dikes or retention basins, but also in non-structural measures such as an accurate and trust-worthy forecast system. This is because the earlier the planning starts, preparation and coordination would have more time to be professionally designed, and therefore, it can limit the consequences of extreme events (Alfieri, et al., 2013). This preparation can be represented in lead time, and it depends on the system with which one is working. Actions in an irrigation channel might not require the same lead-time as a catchment that has a fast response to extreme events (Andel, 2009).

Flood forecasting has a long history in hydrology and its beginnings can be traced back up to 160 years ago (Beven, 2012 as cited in (Nguyet, 2021). Before the era of computers, classical conceptual models were used such as the unit hydrograph or gage to gage correlation (Solomatine and Xue, 2004). However, with the power gained by personal computers, Data Driven Models started to be an alternative to conceptual models.

Solomatine and Xue (2004) and Solomatine and Dulal (2003) showed some examples on how M5 Model Trees and Artificial Neuronal Networks could be an option to generate discharge forecasts. In recent years, works on deep learning techniques have led to the use of Long-Short Term Memory Networks in forecasting application. Hu, et al. (2018) and Le, et al. (2019) used in-situ measurements as input to predict streamflow for the next 6 hours and 3 days, respectively. These four examples have in common that they used only the temporal data and did not consider other spatial variables (1D approach).

Akhtar, et al. (2009) addressed the issue of non-spatiotemporal variability in rainfall using flow length and travel information as pre-processed input information for an ANN to produce river flow forecasting. Their results showed that in the short range, future discharges were highly correlated to previous discharge time steps and that in time horizon from 7 up to 10 days rainfall information could improve the performance of the ANN. However, it did not show acceptable results for a period time longer than three days.

Efforts have been made to generate streamflow forecasting. Some previous works have focused on one-dimensional forecasting. However, due to increase in magnitude and frequency of convective rainfall objects, due to climate change, and human intervention, *"simple extrapolation of forecasts and warning from gauged sites may no longer be sufficient"* (Moore et al.,2006, as cited in (Nguyet, 2021). To address this issue, an integrated approach is needed, where not only spatiotemporal rainfall variability is considered, but also the uncertainty that is present in forecasting, aiming at being a valid option for any region, gaged or ungagged and that can be a reliable tool for decision-making.

A different approach on how precipitation is spatially understood and analysed has been developed in recent years with promising results. This new methodology is object-based solutions, that according to Laverde-Barajas, et al. (2020b, p.2) *"Aim to represent the cells and integrate spatiotemporal information from an event as an interconnected mass figure which can be used to characterize the storm events"*. This methodology focusses on capturing the intrinsic characteristics of precipitation such as intensity, location, and duration to better understand the event and its response in the region of interest.

Laverde-Barajas, et al. (2020b) considering the potential of the object-based methodology proposed a new framework denominated ST-CORA, that allows the identification and analysis of rainfall objects. The methodology consists in two steps: the identification of spatiotemporal convective objects and the storm segmentation and its classification.

Rainfall forecasting depends on the conditions of a chaotic, complex, and non-linear system such as the atmospheric one (Cloke and Pappenberger, 2009), which had led to difficulties when predicting its actual state. To tackle the uncertainty present in the numerical representation of physical processes, as well as the selected parameters, Ensemble Prediction System (EPS) started to be developed around 1994 (Cloke and Pappenberger, 2009), aiming to extend lead-time and quantify predictability of possible extreme events.

EPS had its origin in meteorology, and then it started to be used in the hydrological field since it was an alternative to face the uncertainty and chaos produced in water systems and that deterministic approaches were not able to handle (Demeritt, et al., 2007). EPS has the advantage that produces "perturbed" members, which can reproduce changes on the initial conditions of the system. One example of these systems is the generated by the ECMWF, which currently produces 50 members daily at midnight and noon. This forecasting is generated for a time-lapse of 6h with a lead-time of 15 days (Andel, 2009).

EPS is currently used in many forecasting systems and scenarios. Andel (2009) used it to produce ensemble water level forecasts for Anticipatory Water Management in Rijnland, Netherlands. Additionally, there is the world-recognized warning system known as GloFAS, which can generate daily flood forecasts since 2011 and monthly seasonal streamflow outlooks since 2017 (Alfieri, et al., 2013, Global Floods, 2021) based on EPS. Additionally, there is the EFAS, which can emit early warnings for flood events based on critical flood thresholds (Alfieri, et al., 2013).

Nevertheless, and up to the knowledge of the author, object-based methodologies have not yet been used with ensemble forecast members information for streamflow forecasting and it is there where the author sees a terrific opportunity to contribute to the field. This is because the generation of rainfall objects might help to capture better the how, when, and where of precipitation events and could contribute to provide an alternative on how flood warnings are understood and issued through the fact that not only the intensity threshold is considered, but also the spatiotemporal variability of precipitation events, which can be a base for future works on flood warning systems and flood forecasting.

The purpose of this study will be to analyse and evaluate the response to different forecasted rainfall events with the help of the ST-CORA object characterisation methodology and a conceptual HBV model, seeking to contribute to the development of future works on a more accurate forecasting system. Additionally, and once, the conceptual model is correctly calibrated it is expected that regions that can be a source of floods can be recognized and analysed.

1.2 Objectives

1.2.1 General objective

To use, test and analyse spatiotemporal methodologies in order to acquire a better understanding on the importance of including spatiotemporal information of rainfall events and its effect on the catchment response.

1.2.2 Specific objectives

- To identify, analyse and evaluate the relationship between the characteristics of rainfall objects generated by the methodology ST-CORA from ensemble forecast members, and their hydrological response.
- To develop and calibrate a hydrological model that can reproduce river discharge peaks of the studied area, and can provide information on the possibility of a flood event based on the hydrological response of the catchment.
- To characterize the case study area, in order to see which regions tend to respond more severely to rainfall events and be a source of possible flood events.

1.3 Research Questions

- What is the relationship between rainfall objects' properties and their hydrological response?
- Can areas inside the case study catchment be identified as origin of extreme events based on rainfall objects' properties?
- Could spatiotemporal rainfall objects generated from ensemble forecast members using methodologies such as ST-CORA, help to contribute to flood forecasting and early warning?

1.4 Innovation

The use of rainfall Ensemble Prediction information is widely used in streamflow forecasting, as it can be seen in the GloFAS system. However, and up to the knowledge of the author, the use of object-based methodologies in combination with EPS in order to represent rainfall and its routing have not been yet explored, which is the main objective and contribution of this research.

Existing systems of streamflow forecasting do not explicitly consider the possibilities of spatiotemporal representation of rainfall. This research focuses on understanding and capturing the intrinsic characteristics of precipitation events through the generation of spatiotemporal objects that help to understand better the how, when, and where of rainfall and its hydrological effect on a catchment through the discharge at the outlet.

Additionally, this research aims to generate new ideas in order to develop alternative on how flood warnings are currently issue and understood, by exploring the idea of not only considering they intensity rainfall threshold as main parameter, but also the spatiotemporal variability of precipitation events; which can be a basis for future works on flood warning systems and flood forecasting.

1.5 Practical Value

This research proposes an innovative approach for the use of ensemble forecast information in streamflow routing using object-based methodologies such as ST-CORA. This is a contribution on how precipitation and its hydrological response is understood and analysed. Which is why it is expected that results can be the basis for future works on early warning systems and data-driven models (DDM), with the objective of producing better, more accurate and trustable forecast, that allows a better preparation when facing high runoff quantities. As well as be a motivation for public entities such as weather national agencies to use object-based methodologies for their forecasts and flood systems.

Developing nations do not have enough resources for covering their territories with in-situ measurement stations and, hence, they tend to deal with data scarcity. Since this research will be working with products with quasi-global coverage, this could represent a wonderful opportunity for these countries to update and be better prepared for flood events, seeking to improve their forecasting systems, and hence, have a positive impact on their population.

1.6 Thesis outline

The following document is organized as it follows:

- **Chapter 1 – Introduction:** A brief description of the problem background and what it has been done to address the issue, as well as the motivation, innovation, practical value, objectives and research questions of this thesis.
- **Chapter 2 – Literature Review:** A brief conceptual review of the main theorical cores of this research: Flood forecasting, Hydrological Modelling, EPS, Object-based methodologies.

- **Chapter 3 – Case study:** It describes the area of interest and the available data for it.
- **Chapter 4 – Research methodology:** It explains the framework that it was followed in order to accomplish this research.
- **Chapter 5 – Results and discussion:** It analyses the obtained results and explore the relationship between rainfall objects and their hydrological response.
- **Chapter 6 – Conclusions and recommendations:** it summarizes the results of this research, as well as presents some recommendations and comments for future works on the field.

Chapter 2 Literature Review

This chapter presents the main concepts used in this research. They can be grouped in the following topics: Flood forecasting, hydrological modelling, and object-based methodologies.

2.1 Flood Forecasting

Flood forecasting have been implemented either at a continental scale, through systems such as GloFAS (Alfieri, et al., 2013), which is based on distributed hydrological simulation of EWP or at basin scale such as the one developed by Hopson and Webster (2010), that can generate a 1-10 days ensemble river discharge forecast in the Brahmaputra and Ganges Rivers in Bangladesh. Simple hydrological models have been used for the task. For example, HEC-HMS was used by Oleyiblo and Li (2010) to assess its suitability in forecasting for different catchments in China, providing acceptable results. Forecasting deterministic modelling was limited to one-dimensional, but in recent years, 2D approaches started to gained popularity and 3D approaches are used more and more for specific scenarios (Jain, et al., 2018).

In contrast to deterministic models, there is also stochastic models that make use tools such as Auto-regressive Moving Average (ARMA) and Auto-Regressive Integrated Moving Average (ARIMA). Usually these models are used in order to correct and improve forecasting accuracy, as it was used in the Sieve River basin, located in Italy (Toth, et al., 1999).

In more recent years, and with the development of new computational tools, Data-Driven Models started to become start-of-the-art in flood forecasting, where some authors such as Nguyet (2021), make use of Long Short-Term Memory, Long Short-Term Memory Sequence to Sequence, Bidirectional Long Short-Term Memory and Gated Recurrent Units in order to generate future discharge based in previous discharge and rainfall.

2.2 Hydrological Model

A model can be understood as a simplification of a physical and natural process. Models in the hydrological domain help to understand and predict the systems that they represent. These representations of the reality can be grouped into two main categories: Deterministic and stochastic, and depending on its physical insight and their data requirement, they can be further grouped In Empirical, Conceptual and Physically based (Lescouflair, 2021).

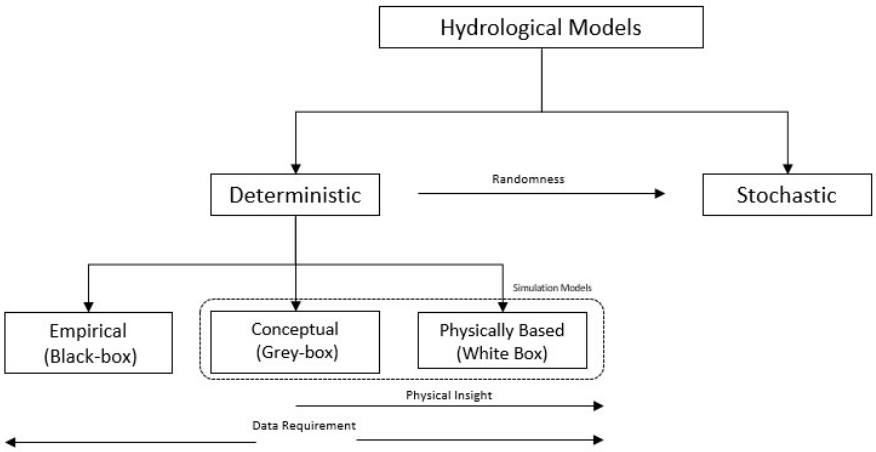


Figure 1 Classification of Hydrological Models
Source : Lescouflair (2021)

- **Empirical (Black box)**

These models are data-driven, based on parameters that do not reflect the physical processes in the field. These models attempt to catch a casual relationship between its inputs and outputs (Corzo Perez, 2009, Farrag, 2018). These types of models can be used in the field of problem prediction and to represent non-linear functions (Lescouflair, 2021).

- **Conceptual (Grey-box)**

These models are composed by storages that represent the physical processes and the relationship between the components of the hydrological cycle (Rainfall, percolation, evapotranspiration, etc). These relationships can be grouped for the whole area of study (lumped model) or be separated in space (distributed model). These storages represent up to some extent physical processes. However, its parameters are more related to the filling and depleting of the thanks than to the physical properties of the catchment.

For this research, a conceptual HBV model (Bergström, 1992) modified through the HAPI library (Farrag, 2018) would be calibrated and used for the analysis.

- **Physically based model**

These models are based on the physical understanding of the relationship and fluxes between the components of the hydrological cycle and uses a point representation of the flow, based on the conservation of mass, momentum, and energy in order to represent the system (Corzo Perez, 2009). These models tend to demand substantial amounts of data, and its calibration might be challenging due to the number of parameters to be considered (Lescouflair, 2021).

2.2.1 HBV Model

This model has its origins in 1973 and it was developed by the Swedish Meteorological and Hydrological Institute. It can be considered as a semi-distributed model, since the area of interest, can be sub-grouped into smaller regions, each with their own conceptual model

parameters. These sub-regions can be grouped based on elevation and vegetation (Corzo Perez, 2009).

One of the advantages of conceptual models such as HBV is that they represent the major processes inside the hydrological model, which can represent the reality, but without compiling with data requirements from a physical model. A schematic representation of the model can be seen in Figure 2.

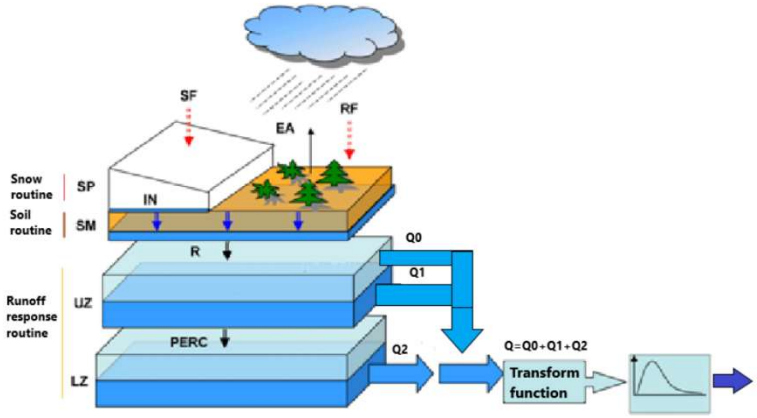


Figure 2 HBV Schematic Representation
Source: (Lescouflair, 2021)

The main inputs for the HBV are precipitation, evapotranspiration, temperature, and discharge, where the first two are used in any type of scenario. The third one is used for an area where the snow routine is considered and the last one, is used during the calibration and validation stage. These variables are governed by the water balance equation as it follows (Corzo Perez, 2009):

$$P - E_a - \frac{\Delta s}{\Delta t} = Q$$

Where:

- P: Precipitation (mm/d)
- E_a: Actual Evapotranspiration (mm/d)
- Δs: Change in storage
- Δt: timestep (d)
- Q: Runoff (mm/d)

When a distributed model or semi-distributed is proposed, it means that each cell (or region) of the catchment would perform as an individual HBV conceptual model. Each of these components, would have three tanks that represent: soil moisture, upper zone, and lower tank, and 5 routines: precipitation, soil moisture, response, routing, and snow.

- **Precipitation:** It can be in two directions, either from rainfall to snow, or snow to rainfall based on temperature thresholds (LT and UT). If the temperature is above the upper limit, then precipitation will be considered as pure snow. On the other hand, if it is below LT, it will be considered as rainfall. If temperature is between the given range, then it will a linear combination between the two states.

$$Rainfall = \begin{cases} P, & if\ T > LT \\ \left(\frac{T - LT}{UT - LT}\right) * P * Rfsc, & if\ UF < T < LT \\ 0, & if\ T < UT \end{cases}$$

$$Snowfall = \begin{cases} 0, & if\ T > LT \\ \left(1 - \frac{T - LT}{UT - LT}\right) * P * Sfsc, & if\ UF < T < LT \\ P, & if\ T < UT \end{cases}$$

Where T, is temperature at certain timestep, P is the precipitation, Rfsc is a correction factor for rainfall, Sfsc is a correction factor for snow. UT and LT correspond to the upper and lower temperature threshold, respectively.

- **Snow**: This routine is based on two states: Snowpack and water content. The snowpack accounts for water in solid state, while the water content would account for the liquid part. If temperature is above a threshold (melting point (MT)), the snowpack will be transformed into liquid. On the other hand, temperatures below will produce an accumulation of the snowpack.

$$Snow\ pack = Cfmax * (T - MT)$$

$$Water\ content = Cfr * Cfmax * (T - MT)$$

Where, Cfmax is day degree factor, T is the temperature, MT is the melting threshold temperature and Cfr is the refreezing factor.

- **Soil moisture:** Runoff in the surface (Rd) occurs when the soil moisture (Sm) exceeds the field capacity (Fc). If direct Runoff does not occur, this water accumulated in the soil can be either stored in the soil layer and then transported to the upper zone tank (Rind) or directly converted into evapotranspiration.

  Indirect runoff (Rind) is calculated as follows:

$$Rind = In * \left(\frac{SM}{FC}\right)^{\beta}$$

  Where Inf represent infiltrated water (Figure 2), and β is a shape coefficient that accounts for the non-linearity of indirect runoff from the soil.

  In case of evapotranspiration losses (El), these would depend on the potential evapotranspiration (Ep). Actual Evapotranspiration (ET). The relationship is represented by the next equation:

$$E_l = \begin{cases} Ep, if\ Sm \geq (Lp * Fc) \\ Ep * \frac{SM}{LP * FC}, if\ Sm < (Lp * Fc) \end{cases}$$

  Where Lp is the limit wilting point and represents the maximum potential evapotranspiration and FC represents the field capacity (Figure 3).

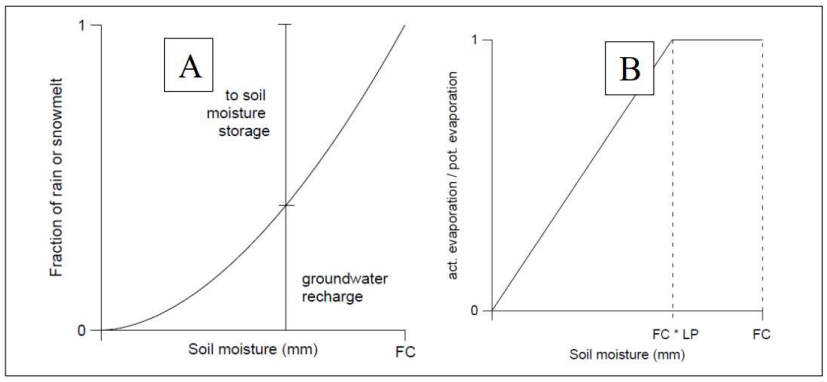


Figure 3 (A) Soil Moisture, (B) Evapotranspiration
Source:(Farrag, 2018)

Ep is first corrected using the long-term mean temperature (Seibert, 2005 as cited in (Lescouflair, 2021)) by the following equation:

$$Ep = 1 + \left((T - T_m) * e_{cor}\right) * Ep$$

Where:

- T_m: is the long-term average temperature.
- e_{cor}: Correction factor.

To represent percolation and capillarity in the soil, additional tanks must be created, one for the soil moisture and two others for the capillarity in the soil. Percolation transport (Perc) is a constant during the whole simulation time. Capillary (Cp) occurs only when there is water availability in the upper zone tank, hence it depends on soil moisture as it can be seen in the following equation:

$$C_f = Cflux * (\frac{FC - SM}{FC})$$

Where Cflux represents the capillary flux in the root zone.

- **Response:** This process is accountable for converting the water excess in the soil into runoff. It consists in two tanks: the upper one represents fast subsurface response and lower one slow groundwater response. For the quick flow, it depends on the following relation:

$$Q_0 = \begin{cases} K1 * UZ, & if\ UZ \leq UZL \\ K0 * (UZ - UZL) + K1 * UZ, & if\ UZ > UZL \end{cases}$$

Where:

- Qo: Outflow from the upper tank.
- UZ: Storage in the upper zone.
- UZL: Threshold parameter from the upper zone.
- K0 and K1: Recession coefficient from the upper tank.

For the lower storage:

$$Q_1 = K1 * LZ$$

Where:

- K1: recession coefficient from the lower tank.
- LZ: Lower reservoir storage.

- **Routing:** The routing can be done either through a triangular weighting function (MAXBAS) or through Muskingum-Cunge method (Cunge, 1969). For the latter, it is carried out through the flow direction and flow accumulation, and it is governed by the following equation:

$$Q^{j+1} = C_1 I^{j+1} + C_2 I^j + C_3 Q^j$$

Where:

$$C_1 = \frac{\Delta t - 2kx}{2k(1-x) + \Delta t}, C_2 = \frac{\Delta t + 2kx}{2k(1-x) + \Delta t}, C_3 = \frac{2\mathrm{K}(1-\mathrm{x}) - \Delta t}{2k(1-x) + \Delta t} \; and \; C_1 + C_2 + C_3 = 1$$

- Δt: Incremental time
- k: Corresponds to the lag time between two cells in the grid.
- x: It is a weighing factor that can varies from 0 to 0.5.

2.2.2 Model Calibration

By calibration, one can understand the process of adjusting certain (model) parameters, so that the model output matches some criteria, i.e., observed discharge value at the outlet catchment.

To do this, the error, between simulated and observed discharge, must be minimized, which would transform the calibration process into an optimization one, where:

- The objective function would be the error to be minimized.
- The decision variables are the parameters of the model.
- The constraints would be the boundaries or limits within the parameters have an actual meaning.

To address this issue, many algorithms can be used, among others genetic algorithms. These algorithms were inspired by the Darwinian theory of evolutionary (Mirjalili, 2019) and consist in a random search of variables of possible candidates for the solution, then these populations evolved to a next generation, in which only the best candidates are considered to create a new population equal to the original size. Solutions with deficient performance are eliminated. The process stops once a condition is met, such as the error converges, and it does not decrease much more.

For this, the algorithm uses three operators: selection, crossover, and mutation. Selection is the process where the best candidates are chosen, crossover is the partial exchange between two solutions previously chosen and mutation is slight changes in the parameters of the candidates from generation to generation.

2.2.3 Model Perform Criteria

To assess the model performance and its selection of parameters, many criteria can be chosen. Since this research will be making use of a open-source Python library developed by Farrag (2018). The metrics provided inside the same will be used. These are:

Root-mean squared error (RMSE)

$$RMSE: \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{obs} - Q_{sim})^2}$$

Where Q_{obs} is the observed discharge and Q_{sim} represents the model output. The ideal value is 0.

Weighted Root Mean Square Error

$$RMSE_{High\ flow} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(Q_{obs} - Q_{sim})^2\right) * w_{hf,i}}$$

$$RMSE_{Low\ flow} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(Q_{obs} - Q_{sim})^2\right) * w_{lf,i}}$$

Where w_{hf} places stronger weights on high flows and weaker ones in flows, while w_{lf} does it the other way around.

Nash and Sutcliffe (NSE)

Where $\overline{Q_{obs}}$, represents the average of the observed discharge. A value of 1 would represent a perfect match between the simulated and observed data. Aye, et al. (2017) mentioned a criterion to assess the performance of the model based on NSE:

Table 1 NSE Criteria
Source : (Aye, et al., 2017)

Performance	Efficiency criteria
Very Good	NSE > 0.8
Good	0.7 < NSE ≤ 0.8
Satisfactory	0.5 < NSE ≤ 0.7
Unsatisfactory	NSE ≤ 0.5

Logarithmic Nash and Sutcliffe (LNSE)

The classic NSE tends to overestimate high values (Farrag, 2018), hence applying the logarithmic function will flat high values and would give a major weight in the error calculation to the low flows. Its ideal value is also 1.

$$LNSE = 1 - \frac{\sum_{i=1}^{n}(\ln(Q_{obs}) - \ln(Q_{sim}))^2}{\sum_{i=1}^{n}(\ln(Q_{obs}) - \ln(\overline{Q_{obs}}))^2}$$

Kling-Gupta Efficiency (KGE)

Gupta, et al. (2009) proposed a new performance parameter to assess the results of a model calibration process based on the decomposition of RMSE into three components: correlation, bias, and measure of relative variability in the observed and simulated values. The ideal value is 1.

$$KGE = 1 - \sqrt{(c-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

Where:

- C: Correlation between Q_{obs} and Q_{sim}.
- α:Variability in the data ($\sigma Q_{sim}/\sigma Q_{obs}$)
- $\beta = (\frac{\overline{Q_{sim}}}{\overline{Q_{obs}}})$

2.2.4 Remote Sensing

As defined by USGS (2022) remote sensing aims to detect and monitor physical variables of a region by measuring the reflected and emitted radiation from a satellite or airplanes. This tool became an alternative for distant or dangerous areas, where in-situ measurements are not available. Remote sensing helps to overcome the data availability problem, while providing a wide range of meteorological parameters for different periods of time, as well for different spatial resolutions (Cheng, 2019).

The methodologies to obtain meteorological parameters can be numerous. For example, Follansbee proposed a technique that could estimate the reflectance in clouds, based on visible and thermal infrared channels (Follansbee, 1973 as cited in Cheng (2019)). In the case of Evapotranspiration, it cannot be estimated directly, but it can be calculated with some other parameters and then validate it through the relationship between vapour pressure and temperature (Granger, 2010 as cited in Cheng (2019)).

2.3 Ensemble Forecasting

As defined by ECMWF (2017): *"An ensemble weather forecast is a set of forecasts that present the range of future weather possibilities"*. In these systems, multiple runs are performed for the same model, each with different initial condition and slight variation in the original model. The aim is to represent how chaotic can be the systems that it mirrors and to represent, up to some extent, the uncertainty present in the system and the models themselves, as well as the lack of knowledge by us, humans, to interpret atmospheric processes, define proper models, and initial conditions.

EPS has the major advantage when compared to "classic" determinist forecast that represent the different scenarios that can occur and that the latter fails to represent (Demeritt, et al., 2007). It also contributes to handle uncertainty, since if all members produced a short range of outcomes, then the forecast can be considered more accurate. On the other hand, if the outcomes have a wide range, the forecast can be considered highly uncertain (Andel, 2009).

The history of EPS in the ECMWF can be traced back to 1979, where the first operational forecasts with an uncertainty estimate were produced, and in 1992 the first EPS was developed (ECMWF, 2017).

The EPS is produced first, with a control forecast that is calculated with the best available data (the one used for a deterministic forecast) and unperturbed models. Then, the models and the initial conditions are slightly perturbed, in order to produce different outcomes, these perturbed models and their outcomes are known as perturbed members. These perturbations are done in a way that each perturbed member has the same likelihood to occur as the others. A scheme of how the process looks like can be seen in Figure 4.

The process to produce the perturbation varies depending on the institution. They are based, mainly on dynamic estimates of analysis error and error growth (Andel, 2009). At ECMWF, it starts with a sample of initial state errors and use them as an input for a simplified and lower resolution model from the atmosphere for a time short forecast (36 hours), and through iterations it finds for which perturbations the maximum change development of the atmospheric system is reached. This method to find the rapid changing in the atmosphere's state is called Singular Vectors, hence the approach used by the ECMWF for its EPS is called Single Vector Approach (Mureau et al., 1993 and Molteni and Buizza, 1996 as cited in Andel (2009)). Each institution has its own method of preference. For example: the NCEP in the United Stated uses the Breeding method and the CMA in China uses Bred Vector. The objective of looking for the biggest change in the outcome is to cover the full probability distribution of the upcoming state of the atmosphere (Andel, 2009).

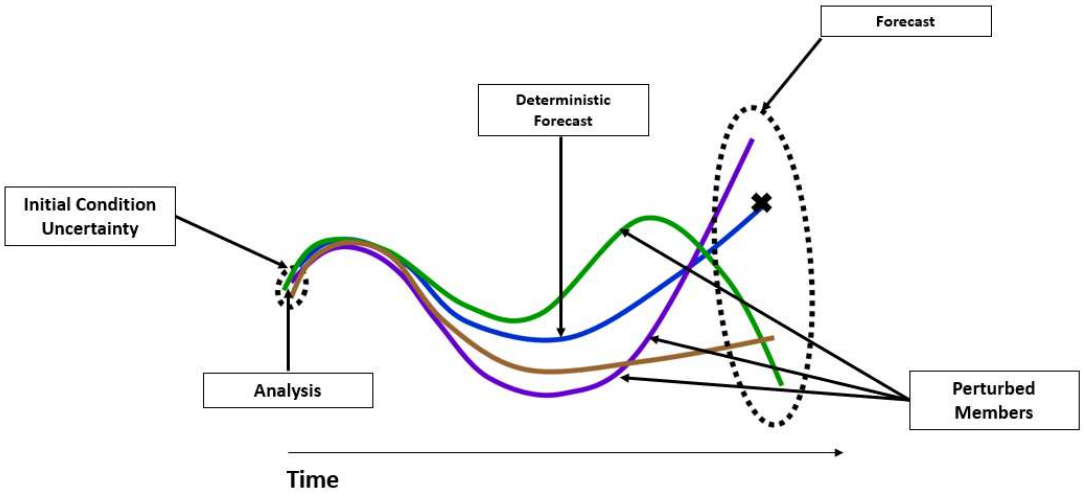


Figure 4 Schematic Illustration of an ensemble forecast
Source: Adapted from (UKMO, 2022)

The number of perturbed members produced varies from organization to organization. For example, the United Kingdom Meteorological Offices produces 17 perturbed members plus one control forecast. Deutsche Wetterdienst, produces 40 perturbed members plus one control forecast. The one of interest for this research, it is the one produced by the ECMWF, which consists of 50 perturbed members and 1 control (or deterministic) forecast.

The final result is a number (equal to the number of perturbed members) of time-series space by a time-step for each of the cells of the grid-extension and for all the atmospheric variables

modelled. For the ECMWF and the NCEP this time-series have an extension from 0 to 15 days ahead.

The fact that these information is available and produced in an operational way has allowed to contribute to the decision making, because when considering the uncertainty in weather forecasting have been proofed to be more cost-effective than when just considering determinist forecast (Rounlin, 2007 as cited by Andel (2009)).

2.4 Object-based Methodologies

An x and y plane can represent the space dimension. In this space, single cells can have values. These values positioned in space can be, and for this research, rainfall values. This pixel can be extended in a third dimension, which could be the time. Once the pixel acquires volume, through its expansion in the time dimension, it becomes what it would be define as a voxel.

Voxels can be grouped in space and in time, while groups of voxels can also be connected in the same dimensions. Once, a link is made between these isolated voxels groups, then it would become an object. The entire process can be seen in Figure 5.

For this research, a rainfall object would be defined as a group of voxels connected in space and time by a common attribute that all together form a 4D object with spatial reference (longitude and latitude), spatial extension and intensity.

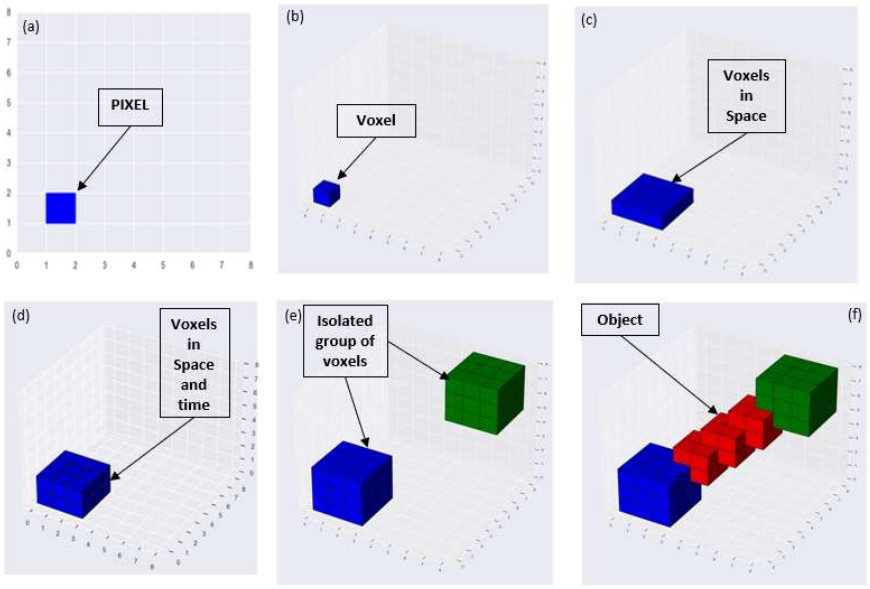


Figure 5 Object Creation (a) Pixel, (b) The pixel Acquired Volume, (c) Voxels grouped in space, (d) Voxels grouped in space and time, (e) Isolated group of voxels and (f) Interconnected voxels in space and time

In literature, many uses for object-based methodologies can be found. Davis, et al. (2009) used it to evaluate and compare climatic models. Skok, et al. (2009) used them to characterize and analyse the temporal and spatiotemporal behaviour of precipitation systems over the equatorial Pacific Ocean. Li, et al. (2015) used object-based methodologies to validate rainfall information from satellite-products against ground radar observations. More recently, Laverde-Barajas

(2022) developed an operational rainstorm analyser, based on a novel object-oriented methodology known as ST-CORA to monitor and alert rainstorm events in Thailand.

2.4.1 Literature Review Summary

Along the last chapter, basic concepts and examples of the use of object-based methodologies, EPS, hydrological modelling and flood forecasting were shown. However, there seems to be room for the use of rainfall objects and their characterisation in order to attempt to predict high hydrological responses, which is the starting point and motivation from this research. Following recent developed methodologies such as ST-CORA, EPS rainfall information will be used to create rainfall objects, that can be routed through a hydrological model, expecting to stablish a relationship between the former with the latter.

Chapter 3 Case Study

This chapter gives and overall picture of the Dapoling-Wangjiaba Catchment and the in-situ available data.

3.1 Description

The Huai River has a length of 1110 km, a drainage area of 174 000 km2 and a basing area of 270 000 km2 that is distributed across five provinces: Shandong, Henan, Hubei, Anhui and Jiangusu. It is located between the two biggest basins in China, the Yellow river and the Yangtze River basin and it begins in the Tongbai Mountains in the east of China.

It used to discharge in the yellow sea, but back in 1194 (until 1897) the Yellow river breached its own dikes, changed its course, and merged with the Huai River. As a result, salting processes occurred and the river failed to keep following to the lower part of the basin and finished by forming two lakes: Hongze and Gaoyou.

From 246 B.C. to 2010, 1008 flood disasters and 938 droughts were recorded in the Huai River Basin zone and the frequency of disaster produced by floods in the region is one time in ten years (Kai, et al., 2016). Additionally, the zone has special interest because:

- 2/3 of the middle and downstream of the major rivers of the catchments are prone to floods.
- 13% of China population lives in this zone (around 185 000 000 people).
- 12% of cultivated land area of China.
- 1/6 of food product in China.

However, during 2020 concerns regarding flood events raised again due to torrential precipitation that forced the government to raise its alert system to III, the second highest (Reuters, 2020) and as stated by Kai, et al. (2016, p. 707) from the Huaihe River Commission: “in recent years, the extreme weather events have increased significantly from global

perspective. The sudden strong rainfall is unforeseeable, and it becomes a new issue for flood prevention work. In the future, we should continue to strengthen early warning system, to optimize flood forecasting model and to improve forecast accuracy and quality".

In Figure 6 it can be seen some of the consequences of recent flood events in the region. Giving the importance of the region and the consequences of previously natural disasters, the Chinese government has implemented some structural measures in the region, among others: 38 large reservoirs (5700 reservoirs in total), 21 Storage areas, 1716 km embankments and some diversion rivers such as Huaihongxinhe and Ruhaishuidao.

Year	Disaster Area (10^4 hm^2)	Affected people (10^4)	Destroyed houses (10^4)	Immigrant (10^4)	direct economic losses(10^4 RMB)
1991	401.6	5423	196	226.1	339.6
2003	259.1	3730	77	207	286
2007	158.7	2472	11.53	80.9	155.2

Figure 6 Statistics of flood disasters in 1991, 2003 and 2007 in HRB (Kai, et al., 2016)

For this research, the Dapoling-Wangjiaba was selected as area on interest. This sub-basin is located at the upstream part of the Huai River (see Figure 7) and it is located between the coordinates 31.4°N - 33.3°N and 113.25°E – 115.8°E. This particular catchment accounts for 16% of the total area of the Huaihe River Basin and within it two major cities are located: Zhu Madian and Xin Yang.

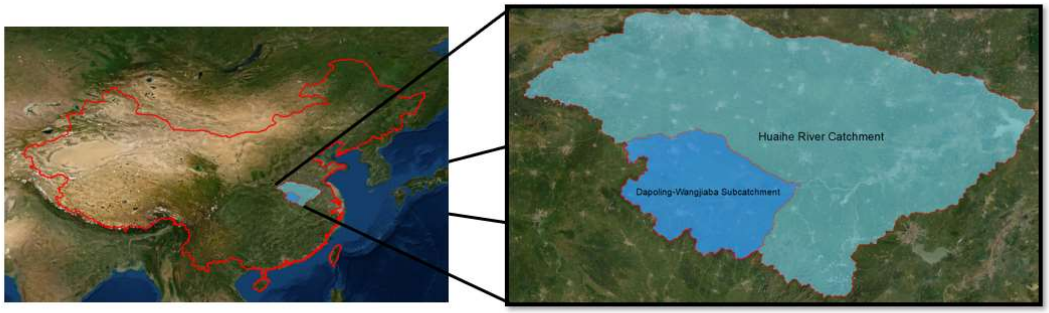


Figure 7 Location of the Dapoling-Wangjiaba Catchment

3.2 In-situ Measurements

A general picture of the available in-situ data can be se seen in Figure 8.

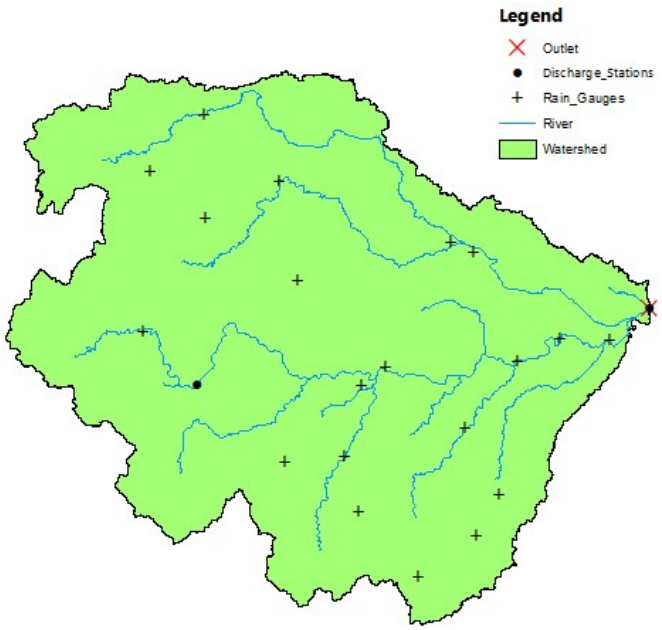


Figure 8 In-Situ Data for the Dapoling-Wangjiaba Catchment

3.2.1 Rainfall Stations

Daily precipitation data is available for 20 rain gauges for the period from 01/01/2001 to 28/10/2011. However, the data from the first four months of 2006 are missing. This data can be completed using the average values of the other years (Cheng, 2019). As it can be seen in Figure 9, the months with most average rainfall are June, July, and August, which correspond to rainy season.

Table 2 General Overview In-Situ Rain Gauge Stations

Name	Longitude (°)	Latitude (°)	Elevation (m.a.s.l.)	Max daily Rainfall (mm)	Average Daily Rainfall (mm)	Average Anual Rainfall (mm)	Std. Of Daily records
SuiPing	114.01	33.14	61	134	2.35	863.45	8.64
BanQiao	113.8	32.94	134	210.3	2.57	940.83	9.87
XiaTun	114.31	32.92	57	169.5	2.13	781.24	8.58
QueShan	114.03	32.8	88	113	2.24	822.53	8.49
ZhengYang	114.39	32.6	76	166.8	2.33	851.43	9.41
DaPoLing	113.8	32.42	102	254.1	2.65	962.4	10.45

Name	Longitude (°)	Latitude (°)	Elevation (m.a.s.l.)	Max daily Rainfall (mm)	Average Daily Rainfall (mm)	Average Anual Rainfall (mm)	Std. Of Daily records
BanTai	115.07	32.71	35	144.2	2.64	979.32	9.26
XinCai	114.98	32.74	36	220.7	2.69	1001.15	10.52
XiXian	114.74	32.33	44	181	2.82	1032.77	9.90
ZhuGanPu	114.65	32.26	46	120.8	2.71	1013.99	9.09
ZheZiJi	115.24	32.35	25	148.8	2.36	871.65	9.24
HuaiBin	115.41	32.43	34	131.8	2.70	1008.32	9.17
WangJiaBa	115.6	32.42	28	164.5	2.32	855.89	8.81
HuangChuan	115.05	32.13	34	169.1	2.61	979.32	9.27
NanLiDian	114.59	32.03	55	164.8	2.38	871.94	9.00
ShiShanKou	114.36	32.01	72	143.6	2.94	1094.4	9.87
ShuangLiuShu	115.18	31.91	66	188	2.68	983.86	10.30
WuYue	114.65	31.85	87	196.5	2.80	1030.94	10.05
BaiQueYuan	115.1	31.78	76	203.8	3.20	1188.69	10.56
XinXian	114.88	31.64	94	127	3.31	1233.02	10.49

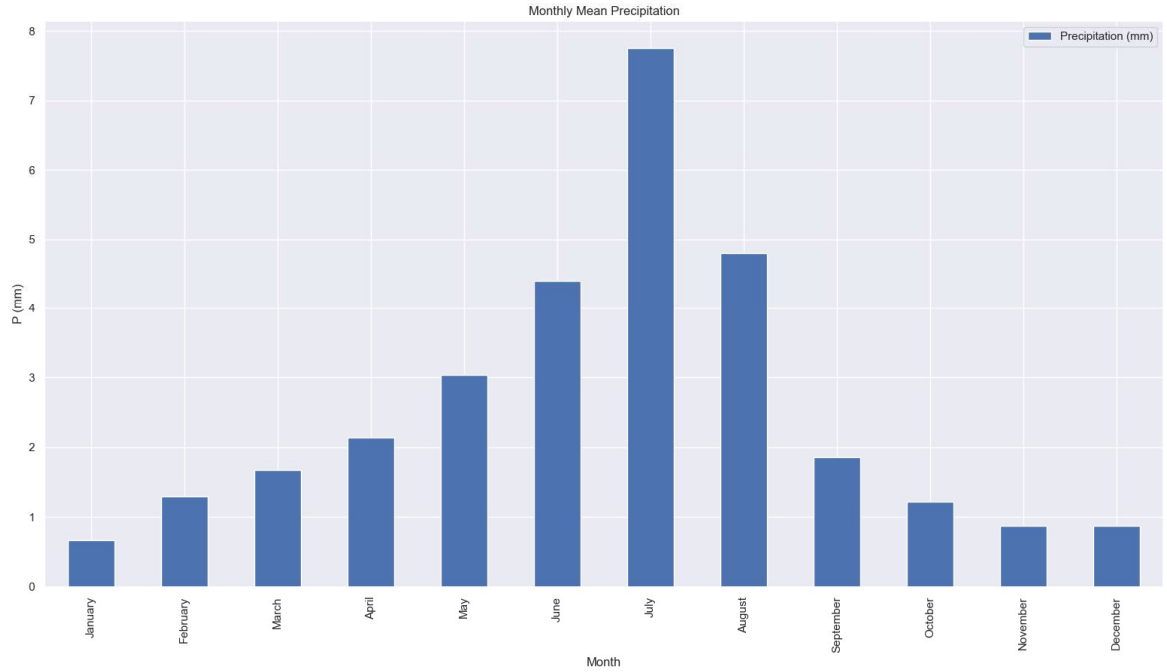


Figure 9 Monthly Mean Precipitation – In situ Data

3.2.2 Discharge Stations

Daily discharge data corresponds to the station Wangjiaba located at the outlet (Figure 8) Of the Sub-catchment. Data is available from 01/01/2001 to 31/12/2010. Its maximum value is 7700 m^3/s, and its maximum monthly average values are in the period from June-August (Figure 11).

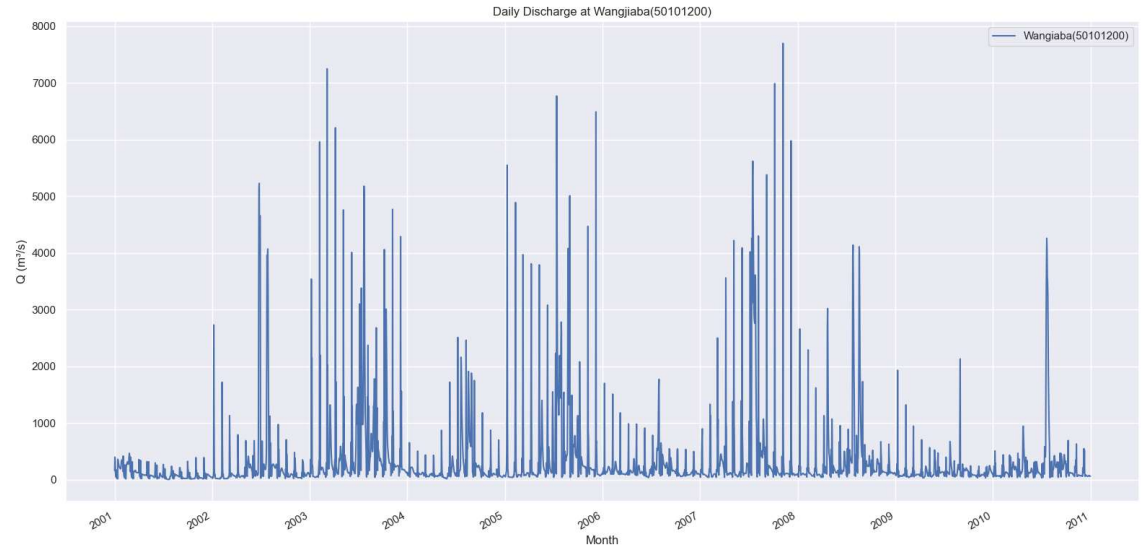


Figure 10 Observed Daily Discharge at Dapoling-Wangjiaba catchment outlet

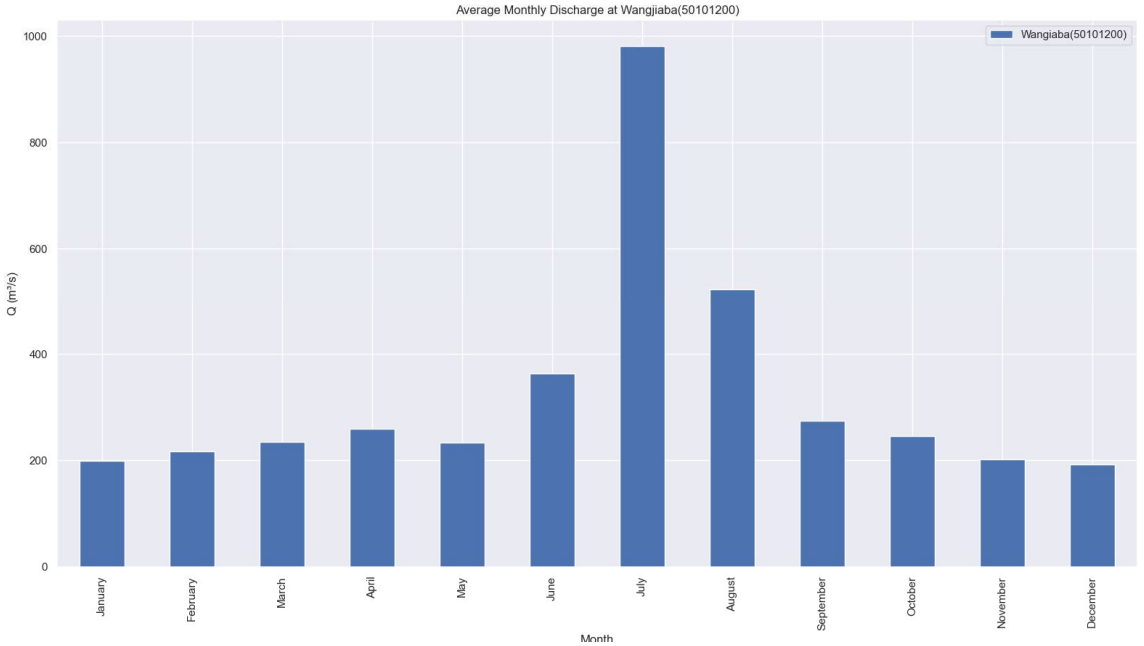


Figure 11 Average Monthly Discharge at Dapoling-Wangjiaba Catchment Outlet

Chapter 4 Research Methodology

This chapter introduces the methodology followed for this research. A general overview is first presented, then the data acquisition and its pre-processing are explained. After this, the rainfall objects' creation is explained, which is followed by the hydrological model calibration and the routing of the created objects.

4.1 General Overview

In Figure 17 an overall view from the methodology followed in this research is presented. This contains four stages: Data acquisition and its pre-processing, object generation, hydrological model calibration and the runoff generation and the analysis of results. In the following sections each of these stages are introduced in detail.

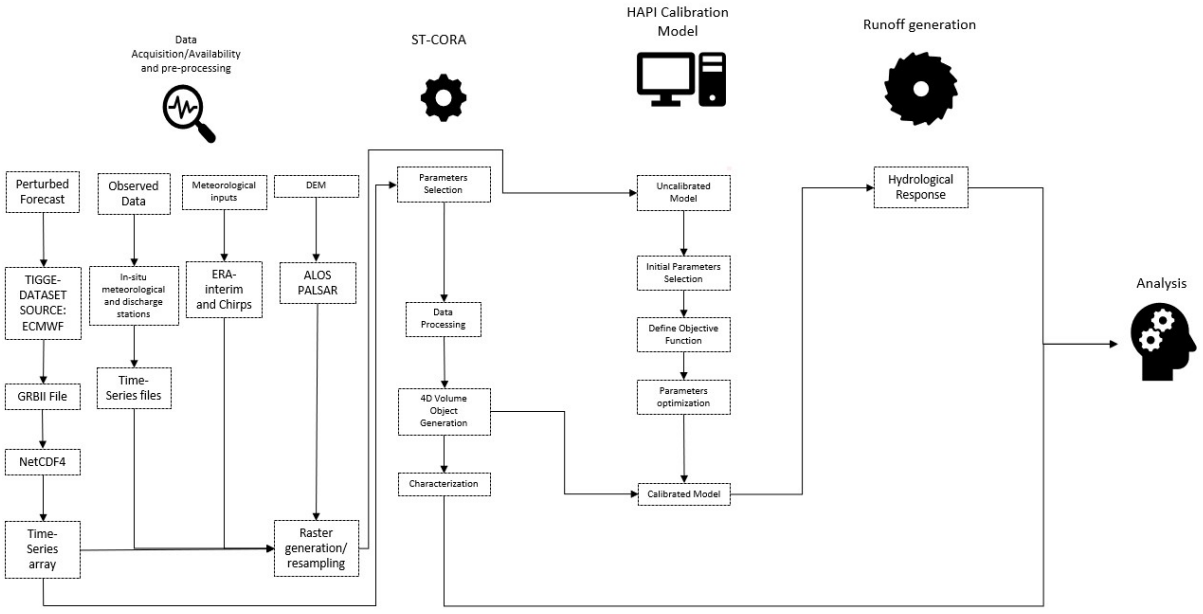


Figure 12 Research Methodology

4.2 Data acquisition

4.2.1 Topographic data

Due to lack of knowledge of the zone, free accessible information can help to grasp a better understanding of the region. For this, a DEM is used. Many satellite products including this information are available. For this research, ALOS-PALSAR is used. This information is freely available (https://search.asf.alaska.edu). The product resolution is 30x30 m. As it can be seen in Figure 13, elevation in the zone varies from 3 to 1131 m.

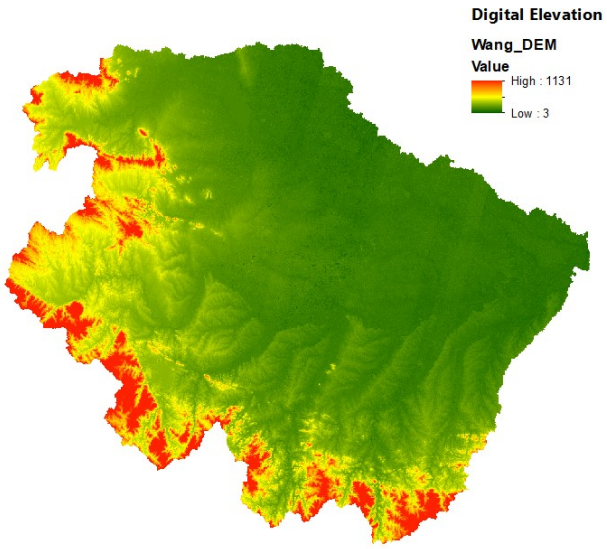


Figure 13 Digital Elevation Model for the Dapoling-Wangjiaba Sub-Catchment

4.2.1.1 Precipitation - CHIRPS

Precipitation can be also obtained, compared, and analysed through CHIRPS and the HAPI library. This product merges three types of information: global climatology, satellite estimates and in-situ observations at a great resolution (0.05°) and covering the period from 1981 to present. Additionally, and according to Bai, et al. (2018) CHIRPS can be considered a suitable rainfall information source for the zone of interest.

In Figure 14 can be seen the mean monthly average precipitation for the period from 2001 to 2011. Rainy season can be detected from June to august.

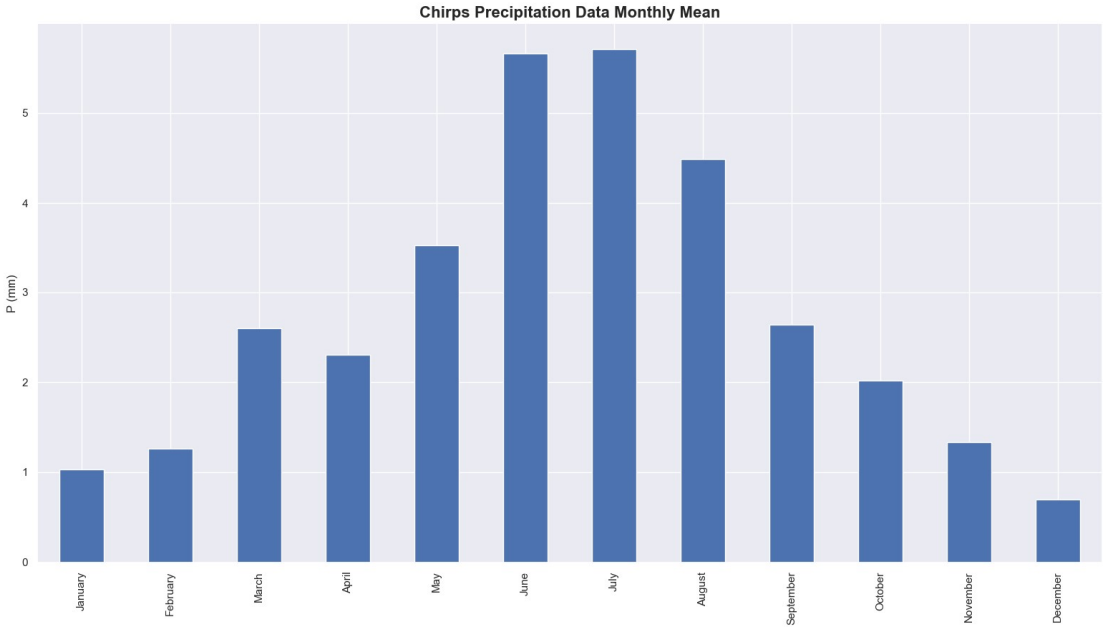


Figure 14 Monthly Mean Precipitation – CHIRPS

4.2.1.2 Temperature data

Temperature data have been acquired through the HAPI library, which uses as data source the ERA-Interim dataset. This reanalysed dataset has global coverage and contain atmospheric information since 1 January 1989 (Dee, et al., 2011). The resolution of the information is 12.5 m, and it covers the period, for which in-situ rainfall information is available. As it can be seen Figure 15, the region has a minimum daily temperature of -6.11 °C, a maximum of 34.62 °C and an average for the period of interest of 16.65 °C. These low temperatures led to believe that the snow subroutine from the hydrological model should be used.

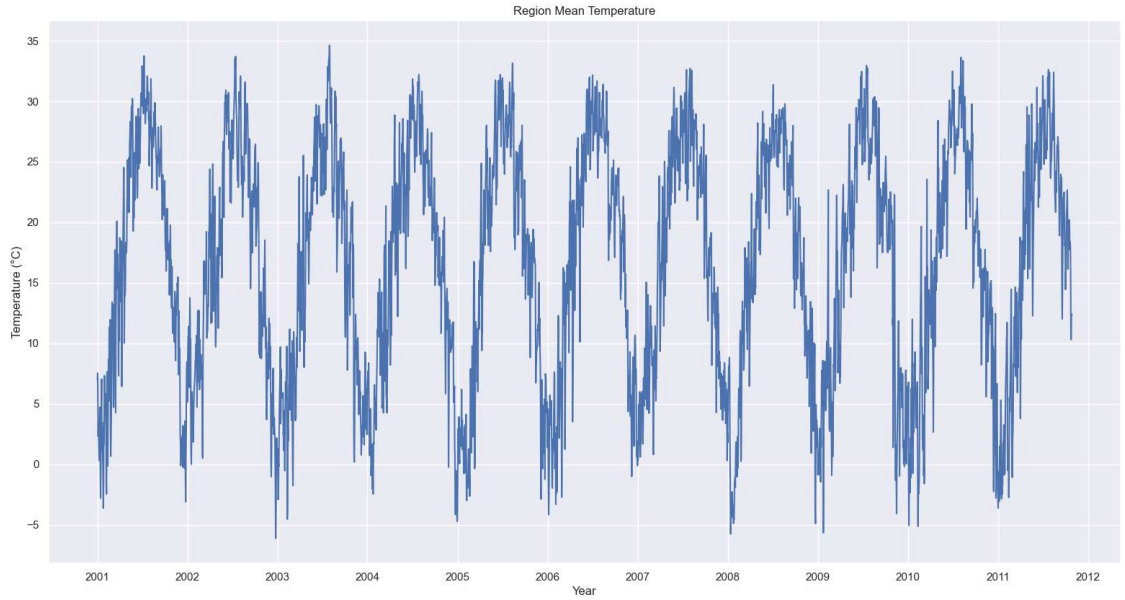


Figure 15 Dapoling-Wangjiaba mean daily temperature

4.2.1.3 Evapotranspiration data

This data is also acquired through the HAPI library. Since the data is provided by ECMWF, it has a particular connotation regarding the parameter of interest: Negative values in the information would represent evaporation and positive values will represent condensation. Hence, an additional step must be done, and it is to obtain the absolute value of all the original data. As it can be seen in Figure 16, the area presents a minimum daily evapotranspiration of 0.10 mm, a maximum of 3.59 mm and an average of 1.22 mm/day for the period of interest.

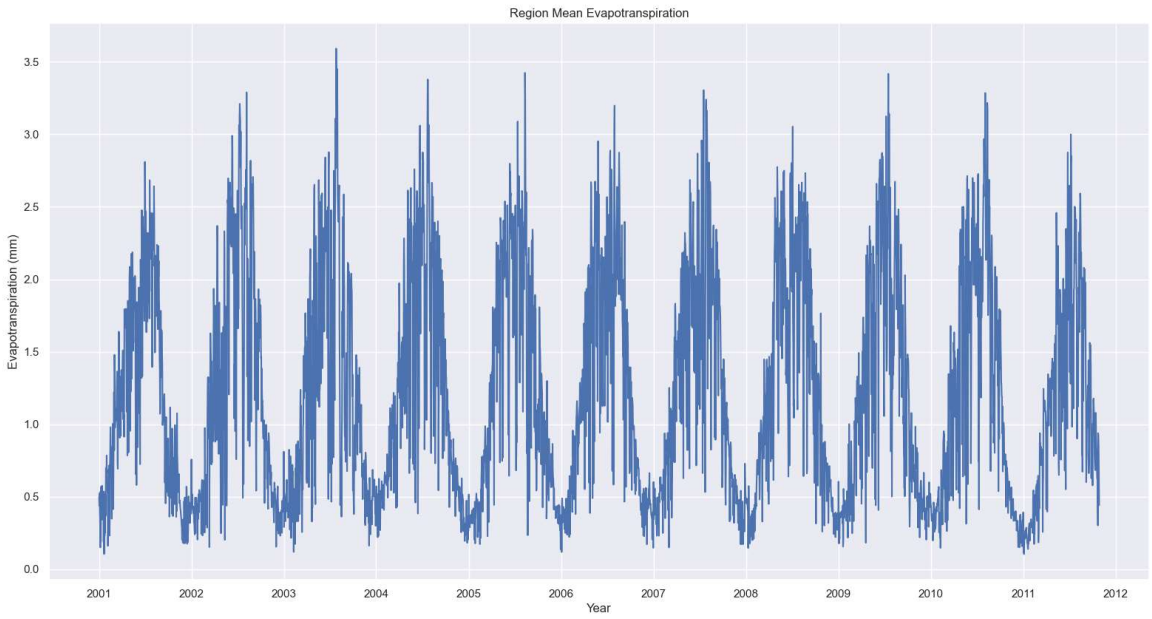


Figure 16 Dapoling-Wangjiaba mean daily evapotranspiration

4.2.1.4 Ensemble Prediction System

The dataset used in this research is part of the THORPEX project and its origins can be traced back to 2004, with the main objective to accelerate and improve the accuracy and reliability of medium-range forecast (up to two weeks) (Swinbank, et al., 2016). The TIGGE dataset aims to provide global operation ensemble forecast and it gathers data from 13 global numerical weather prediction centres (Table 3). The data can be easily accessible from the ECMWF Web-API and is stored in gridded binary (GRIB2) format. This format was standardized and stablished by the WHO, in order to manage perturbed members information.

The data is available from October 2006, which means that only from this month until October 2011, there will be data that matches the in-situ provided data. For this research, the ECMWF data will be selected for the EPS, this is because some previous studies have shown satisfactory results in the region when using this forecasted information(Liu, et al., 2021, Ran, et al., 2018)

Table 3 TIGGE Project Partners

Center	Country	Acronym
Bureau of Meteorology	Australia	BoM
China Meteorological Centre	China	CMA
Canadian Meteorological Centre	Canada	CMC
Centro de Previsão de Tempo e Estudos Climáticos	Brazil	CTPEC
European Centre for Medium-Range Weather Forecasts	Europe	ECMWF
Japan Meteorological Agency	Japan	JMA
Korea Meteorological Administration	Korea	KMA
Météo-France	France	MF
Met Office	United Kingdom	UKMO

Center	Country	Acronym
National Center for Atmospheric Research	United States	NCAR
National Centers for Environmental Prediction	United States	NCEP
National Climatic Data Center	United States	NCDC
Deutscher Wetterdienst	Germany	DWD

Source: Swinbank, et al. (2016)

4.3 Pre-processing

4.3.1 Perturbed member arrays

In Figure 17 can be seen how relatively easy the data from perturbed members can be managed and store. The process includes the transformation of GRIB to NetCDF format, since the author has some experience with the latter type of file, and from there the perturbed members can be easily identified and store individually as NumPy arrays.

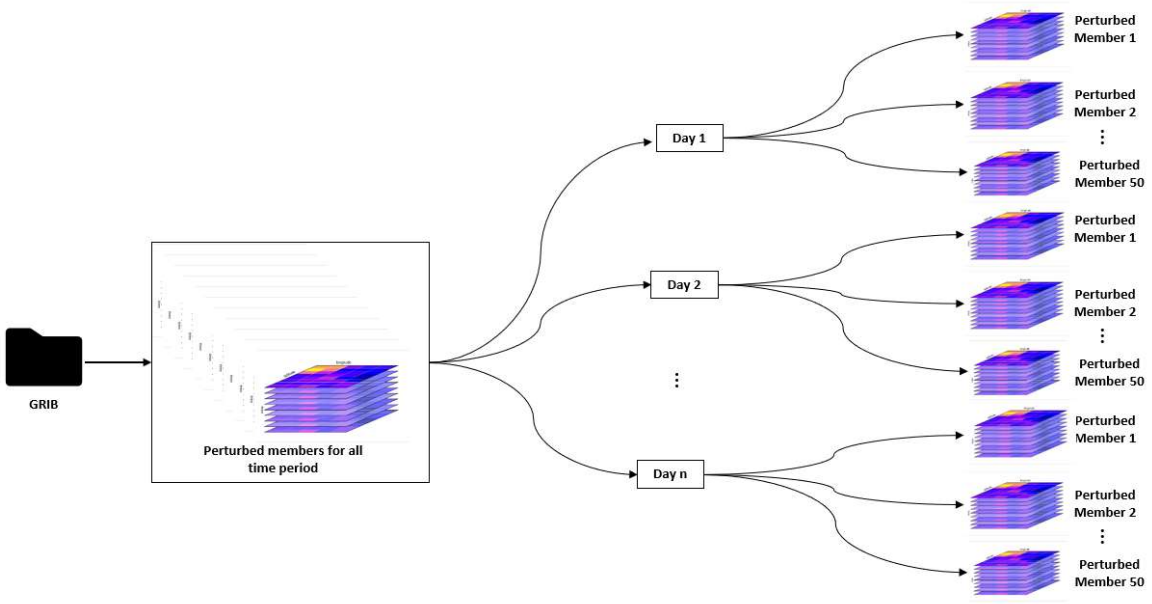


Figure 17 Forecast Ensemble Members Composition

4.3.2 Spatial gauge precipitation

To determine, whether CHIRPS information can be used as input for the hydrological model, a simple comparation with in-situ information can be carried out. For this, both datasets must have the same data format and spatial distribution. Meaning that from point-information such as the one given by the gauges, they must be transformed and distributed spatially. This can be done through different interpolation methodologies, such as the Inverse Distance Weighting (IDW). IDW is obtained through the following relation:

$$u(x) = \begin{cases} \frac{\sum_{i=1}^{n} w_i(x) * u(x_1)}{\sum_{i=1}^{n} w_i(x)}, if\ d(x, x_1) \neq 0 \\ u_i, \quad if\ d(x, x_1) = 0 \end{cases}$$

$$w_i(x): \frac{1}{d(x, x_1)^p}$$

where:

- $u(x_1)$ represents the rainfall gauge at x_1.
- $w_i(x)$ represents the weighting factor calculated based on the distance between the unknown rainfall information at point x and x_1.
- n is the total number of points.
- p is the power parameter.

4.3.3 Raster calculation and generation

All the input information used in this research is spatially distributed, hence some pre-processing must be done before its use. The first step is to re-project all spatial information to a single reference system. For this research, the EPSG 32650 (WGS 84/UTM zone 50N) seemed more suitable.

After this, units must be arranged, as it follows:

- Precipitation (mm)
- Temperature (°C)
- Evapotranspiration (mm)

It is important to remember that based on the ECMWF conventions, Evapotranspiration is considered negative, due to convention for fluxes, where all meteorological parameters that have an upwards direction would be considered negative, and those who have a downwards direction (e.g., rainfall) will be considered positive.

Once, this is done. All rasters must have the same spatial resolution. In Table 4 can be seen the original spatial resolution of the data used. Since all rasters were re-projected, they can be re-sample. For this, an according to the ArcGIS references, a bilinear resampling would be more suitable since the data is continuous.

Due to the fact, that precipitation is the most important input in the hydrological cycle. It would not be advisable to resample this data to a greater resolution. Hence, it is decided that all data shall be re-calculated into 5x5 km grid.

Table 4 Satellite-Products Used

Parameter	Product	spatial Resolution	Temporal Resolution
Precipitation	CHIRPS v2.0	0.05°x0.05°	Daily
Temperature	ERA5	0.125°x0.125°	Daily
Evapotranspiration	ERA5	0.125°x0.125°	Daily

Finally, the rasters must be clipped to fit the catchment shape. Once, this is done, the information is ready for its use in the hydrological model.

4.4 ST-CORA

As described by Laverde-Barajas, et al. (2019), object-based methodologies aim to analyse data from physical processes, its attributes, and the relationship between them. The same author aiming at introducing the spatial variability present in rainfall events, proposed a new spatiotemporal object-based methodology, which denominated ST-CORA.

Rainfall events have intrinsic characteristics, such as intensity, duration, location, and spatial extension. These characteristics can determine how severe or how much damage can produce an extreme event (Corzo Perez and Varouchakis, 2019). Object-based methodologies such as ST-CORA is an alternative approach, to the common 1D approach, to analyse precipitation events.

4.4.1 Intensity Threshold definition

First, an intensity threshold (IT) must be defined. This threshold would allow the algorithm to identify cells with precipitation values high enough to be considered as rainfall, instead of noise produced either by the NWP (Numerical Weather Prediction) forecast or by the used of satellite-product. The algorithm would identify each cell with true or false values as it follows:

$$S_{x,y,t} \coloneqq \begin{cases} 1, if\ R_{x,y,t} \geq IT \\ 0, \quad otherwise \end{cases}$$

Were $R_{x,y,t}$ represents rainfall voxel value and 1 and 0, represent true or false, respectively.

4.4.2 Connected-component labelling algorithm

Once all $S_{x,y,t}$ have been properly labelled, a connected-component labelling algorithm starts identifying preliminary connected objects as follows:

1. It scans all voxels from top to bottom and left to right assigning preliminary labels.

$$C(S_{x,y,t}) = \{N_{[x,y,t]} \in \partial s: S_s = S_N\}$$

2. If a neighbour has more than two C, it is identified as the lower label recording the label equivalences in a union-find table.
3. Resolve the table of equivalences classes using the union-find algorithm (Sedgewick, 1998, as cited in (Laverde-Barajas, et al., 2020b)).
4. Proceed to a second iteration relabelling on the resolved equivalences classes.

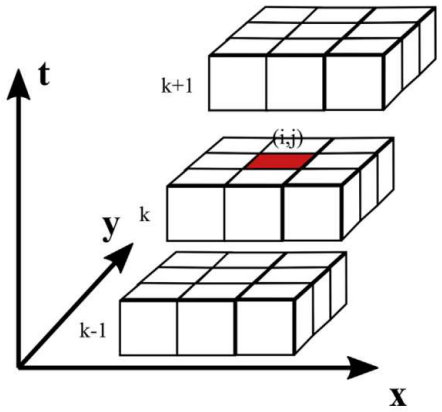


Figure 18 Neighbour System in Space and Time (26 Voxel Neighbours)
Source: (Corzo Perez and Varouchakis, 2019)

4.4.3 Noise Removal and Morphological Closing

After the connected-component labelling has been applied, the next step is to remove isolated events and do a morphological closing in order to delineate the created objects. Based on selected-used parameters, a minimum object size "T" parameter is defined. All created objects that are below this parameter, would be consider as noise and hence, removed.

After the noise removal, the algorithm resolves false merging. Meaning that the objects would be divided or merged depending on their level of connectivity (Figure 19). This fall-merging process follows a dilation process similar to the one employed to separate zones with a weak connection in weather radar (e.g., Han et al., 2009, as cited in(Laverde-Barajas, et al., 2020b)). This algorithm first performs a morphological dilatation, where the boundaries of the objects are expanded, and then it proceeds to a morphological erosion, in order to delineate the objects.

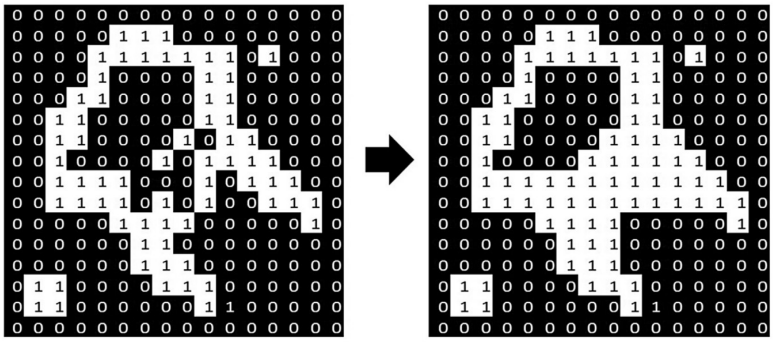


Figure 19 Morphological Closing
Source: (Corzo Perez and Varouchakis, 2019)

However, the previously mentioned process is based on binary objects and do not consider the voxel's precipitation values, hence Laverde-Barajas, et al. (2020a) introduced a multivariate Kernel Density Estimation (KDE) approach to consider the fours dimensions of the object,

when segmenting and merging voxels' groups. KDE has found many uses in the domain of image detection and object tracking (Laverde-Barajas, et al., 2020a).

KDE is estimated at point x from a random sample $X_1, X_2, \dots X_n$ from a density function, f,

$$\widehat{f_k}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x - x_i)$$

Where K corresponds to the kernel function and h is the bandwidth matrix. Since, the bandwidth value can range between a vast number of values, ST-CORA calculates it assuming that f, follows a Gaussian distribution (Henderson et al., 2012 as cited in (Laverde-Barajas, et al., 2020a)).

The process to detect the edges using KDE is based on the method developed by Pereira, et al. (2014). For this, the objects are segmented by a user-selected parameter density threshold *u.* This parameter identifies the edges that are lower than a probability percentage.

Once, this process is done, the cells return to precipitation values and the 4D-Object is created:

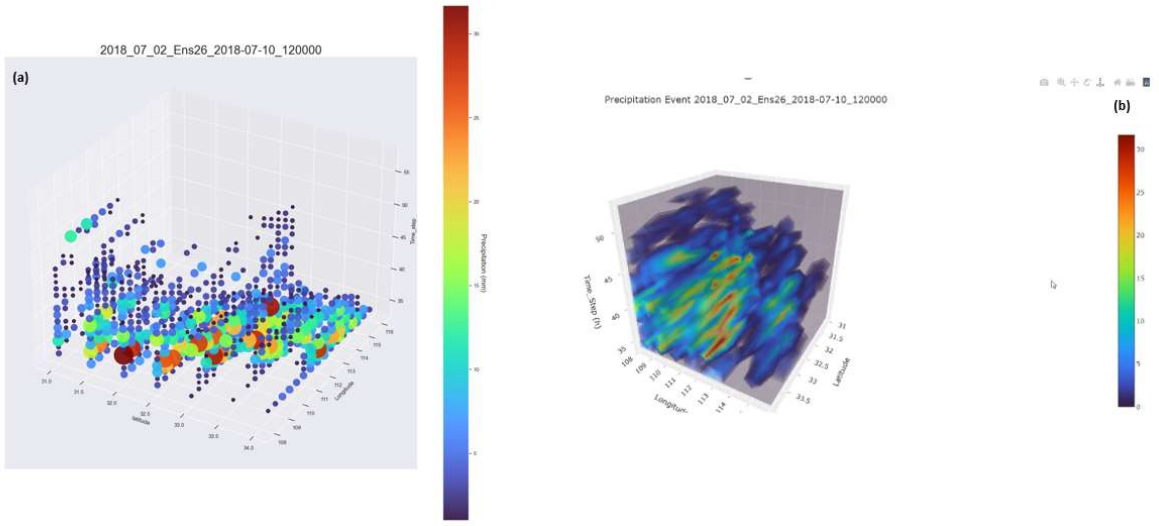


Figure 20 (a) Object as Spherical Voxels Group and (b) 4D Interconnected in Space and Time Object

Once the object is created, it acquired the following characteristics:

- Centroid
- Weighted centroid (based on the value of rainfall voxels)
- Start time
- Duration
- End time
- Maximum precipitation
- Spatial coverage
- Volume

4.4.4 ST-CORA parameterisation

Most of the parameters used by ST-CORA, are parameters that are defined by the used data source or have a relation with the study area, such as: location of the basin, spatial resolution, time resolution, among others. However, there are two user-selected parameters that define what is considered an object, and what not by the algorithm: Intensity threshold and minimum object size.

- **Intensity Threshold**

The EPS data has a coarse resolution of 0.5°x0.5° meaning approximately 50 km in the study zone. First, and to get and overview of the data, different quantiles are calculated for all the period, meaning from October 2006 to December 2010:

Table 5 Precipitation Quantiles for Raw Data

Quantile	Precipitation (mm/6h)
2	-2.47E-04
6	5.94E-07
8	3.22E-06
10	9.40E-06
12	2.21E-05
14	4.55E-05
16	8.52E-05
17	1.13E-04
18	1.48E-04
19	1.93E-04
20	2.47E-04
21	3.15E-04
22	3.98E-04
23	4.97E-04

Since we are dealing with information at a coarse resolution, it was expected that some of the data would show some non-physical values, such as negative precipitation. This can be already considered as noise or error in the NWP equations used in the model.

EPS has a resolution of 0.5°x0.5° which in the zone of interest it would represent approximately 50 km and an area of 2500 $km^2$for each cell. The time step for this EPS dataset is 6 hours. Ideally, there should be enough noise removed to focus the analysis on relatively important events. Percentile 2 contains negative values, which do not make much sense, and percentile 6 would represent less than a cubic meter per hour in each pixel, which is why the quantile 8 seems suitable for this parameter. 3.22E-06 mm/6h would account for a volume of 1.34 m^3/h in each cell.

- **Minimum Object Size**

This value was found through a trial-and-error approach, considering the intensity threshold previously defined. The aim is to generate enough objects, that can afterwards produce a

considerable runoff. For this, it was defined that 10 voxels would be considered as an object, and below that, it would become noise and hence, will be removed, due to low rainfall volume and low connection between its elements.

4.5 Hydrological Model Calibration

This stage of the research is focused on mainly the following steps:

- Process the data acquired and split it into calibration and validation set.
- Define the following and upper boundaries for the conceptual model parameters.
- Define a suitable objective function that allows the model to replicate as best as possible the hydrological behaviour in the catchment.
- Asses the model performance.

4.5.1 Calibration and Validation data set

Since the data was, as explained previously, already pre-processed. The calibration process can be started. For this, the data must be split it into calibration and validation data. From a visual inspection (see Figure 21), can be seen the two major peaks are located between 2001 and 2007, which can be used for the calibration process, while the peak located in 2010 can be used for the validation process.

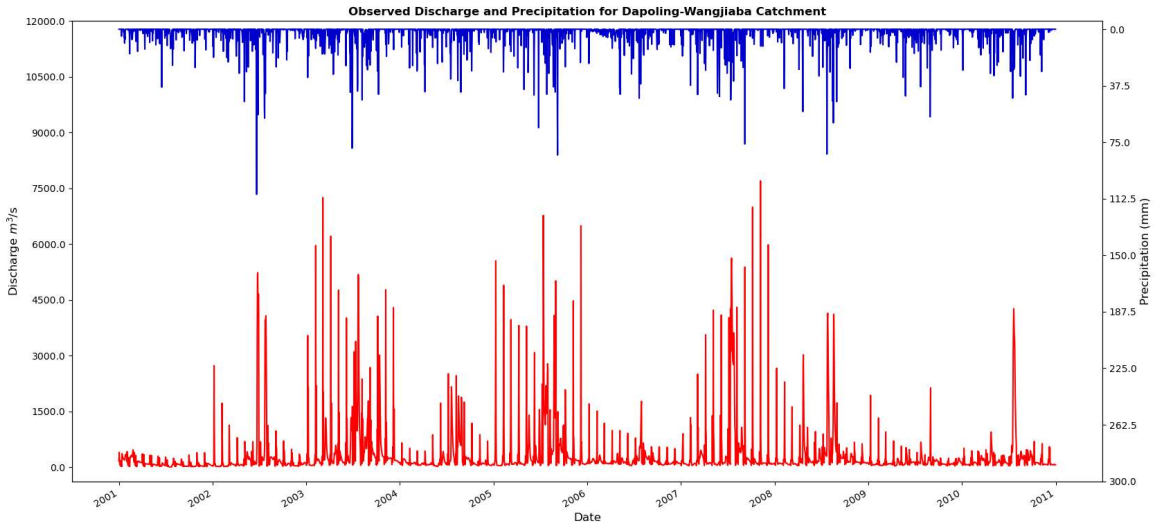


Figure 21 Recorded Rainfall and Runoff in the Dapoling-Wangjiaba Sub-Catchment

A basic description of the split datasets can be seen in the following table:

Statistical Properties	Complete Dataset	Calibration Dataset	Verification Dataset
Date	01/01/2001 - 31/12/2010	01/01/2001 - 31/12/2007	01/01/2008 - 31/12/2010
Considered days	3652.0	2556.0	1096.0
Mean (m^3/s)	328.7	359.4	257.2

Statistical Properties	Complete Dataset	Calibration Dataset	Verification Dataset
Standard Deviation (m^3/s)	705.6	775.9	498.1
Maximum (m^3/s)	7700.0	7700.0	4260.0
Minimum (m^3/s)	3.3	3.3	31.5

Figure 22 Statistical Properties from the Calibration and Validation Dataset

4.5.2 Lumped Model

To calibrate the model, lower and upper boundaries from the parameters must be provided. This information can be obtained from experience of from literature. Since the zone have been widely studied and models have been calibrated under different conditions, for a first attempt, these references (Cheng, 2019, Shao, 2018, Wang, 2017, Xu, 2014) can be used. Once, the first calibration is done, parameters can be refined.

Parameter	Name	Range
ttm	Threshold temperature	0-0.01
rcfc	Rainfall correction factor	0.8-1
sfcf	Snowfall correction factor	0-0.0006
Cfmax	Day Degree Factor	0-1
Cwh	Capacity for Water Holding in the Snowpack	0.1-0.8
Cfr	Refreezing Factor	0.2-0.9
fc	Field Capacity	250-600
beta	Shape Coefficient for Effective Precipitation Separation	2-4.2
e_corr	Evapotranspiration Correction Factor	0.009-0.5
lp	Limit for Potential Evapotranspiration	0.3-2
k0	Recession coefficient upper zone	0.1-0.6
k1	Recession coefficient upper zone	0.01-0.3
k2	Recession coefficient lower zone	0-0.1
uzl	Threshold parameter outflow from upper zone	50-90
perc	Percolation from upper to lower response box	0-1
maxbas	Transfer Function - MaxBas	3-7
k	Storage Constant - Muskingum	15-30
x	Attenuation Factor - Muskingum	0-0.5

Figure 23 Range of Parameters for HBV Model

The snow routine is present due to temperatures below zero Celsius degrees. Additionally, two routing routines will be used: Muskingum and Maxbas.

4.5.3 Model Calibration

The chosen objective function for the calibration is the Kling-Gupta Efficiency (KGE), since it combines the correlation, bias, and ratio of variances from NSE and RMSE and it has per objective the improvement of hydrological modelling (Gupta, et al., 2009).

4.5.4 Variables Initial State

Once the model is calibrated, it will be used to route all rainfall objects. Since the catchment is not isolated, it has some previous conditions that would contribute to its hydrological response, such as groundwater levels or moisture in the soil. For this, it is required to have a warm-up period that leads to replicate as best as possible the state of the catchment once the rainfall object occurs. For this, the following is proposed:

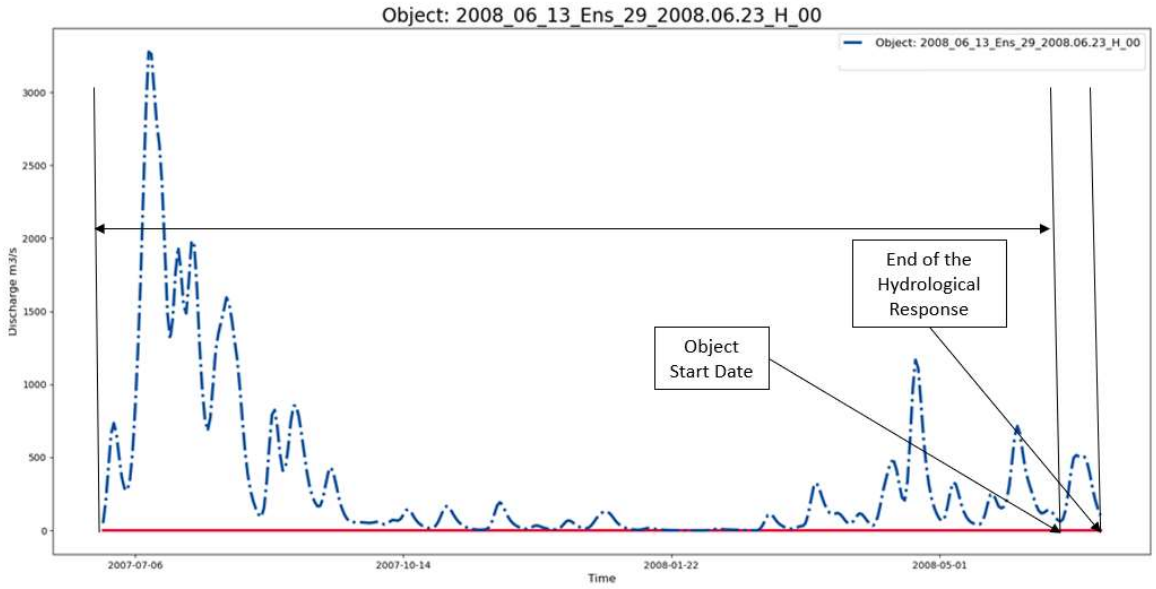


Figure 24 Variables Initial State

- Warm-up period: For this period, one year before the start of the event is considered, the data used is:
    - Rainfall: Observed.
    - Evapotranspiration and Temperature: Observed.

- Start date: Once the event occurs, rainfall data would change to the one provided by the ECMWF Forecasting.
- End date: Events tend to be short in time, but the catchment takes some time in order to respond to precipitation events, due to travelling distances. For this, the Maxbas parameter can be used. Meaning, that once it stops raining, a few days to the simulation will be added. This amount corresponds to the mentioned parameter and represents the time that it would take the last drop of rainfall to reach the outlet.

4.5.5 Critical Discharge Definition

To focus the analysis on those objects that can be considered as critical for the catchment, due to its high hydrological response, a discharge threshold must be defined. Severe events occur

in less magnitude than moderate events, which is why the attention should be drawn to those hydrological events that can represent a possible flood. One way to establish a relationship between the magnitude of extreme events and their frequency is through probability distributions. For this research two criteria would be use in order to define the critical discharge threshold:

- Several probabilistic distributions such as: Gumbel, normal distribution, logarithmic normal distribution, Pearson III, and logarithmic Pearson III. The analysis would be carried out for the observed discharge located at the outlet of the catchment and for different return periods.
- A visual inspection of the simulated discharge in the HBV Model with observed rainfall as input, in order to determine whether the defined threshold, is above low flows.

4.5.6 Classification Algorithm

Since this research aims to provide new approaches to improve forecasting, the first step would be to predict the occurrence of critical events based on rainfall object's properties, where the possible output would be either 1 for flood event or 0 for no flood event, using the critical discharge definition as threshold. Many classification algorithms are available, such as K-nearest neighbour, Logistic Regression or Support Vector Machine. However, and based on previous positive experience from the author with decision trees, this algorithm will be used.

All this research was executed using open-source resources, hence for this step of the methodology, the library Scikit will be used (Pedregosa, et al., 2011).

The science behind this algorithm, is that it will build regression models in the form of a tree with decision points along the branches. Each node represents a decision for one of the attributed introduced.

The main objective is to reduce the entropy to 0, so one class is fully dominated.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Where S represent the members, c the classes and p represent a fraction of examples given in a class.

In order to do this, the algorithm will attempt to maximize the gained information at each split:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} ID_{left} - \frac{N_{right}}{N} ID_{right}$$

Where:

- f: feature
- D_p: Original dataset
- D_{left}, D_{right}: spllited dataset at left and right node respectively
- I: Gini function (function that assess the likelihood of an incorrect classification)
- N: Original number of samples before the splitting
- N_{left}, N_{right}: number of samples at each node after the splitting

Chapter 5 Results and Discussion

This chapter has per objective presented the results found, after following the previous mentioned methodology. First, the satellite product is compared with the ground information, to assess its suitability. Then rainfall objects are described and divided into time periods or seasons, after which the relationship between objects' properties and hydrological response is established and analysed. Finally, the rainfall objects' properties are used as input to a classification algorithm for predicting high discharge events.

5.1 Satellite Product and In-Situ Data

For this research, there were two rainfall information sources. First there is point-gauge data interpolated through kriging methodology in order to distributed spatially through the catchment and data obtained directly from CHIRPS. In Figure 25, can be seen a comparation between their monthly mean value for these two data sources.

As it can be seen CHIRPS tend to show slightly higher precipitation levels except for the month of July. When calculating the MAE, 0.65 mm/month is obtained, which could be seen as an acceptable error.

Since CHIRPS data has already been tested in the region (Bai, et al., 2018) and through the simple analysis shown here is corroborated its suitability, it is decided to use CHIRPS rainfall information for the calibration of the model and the further analysis.

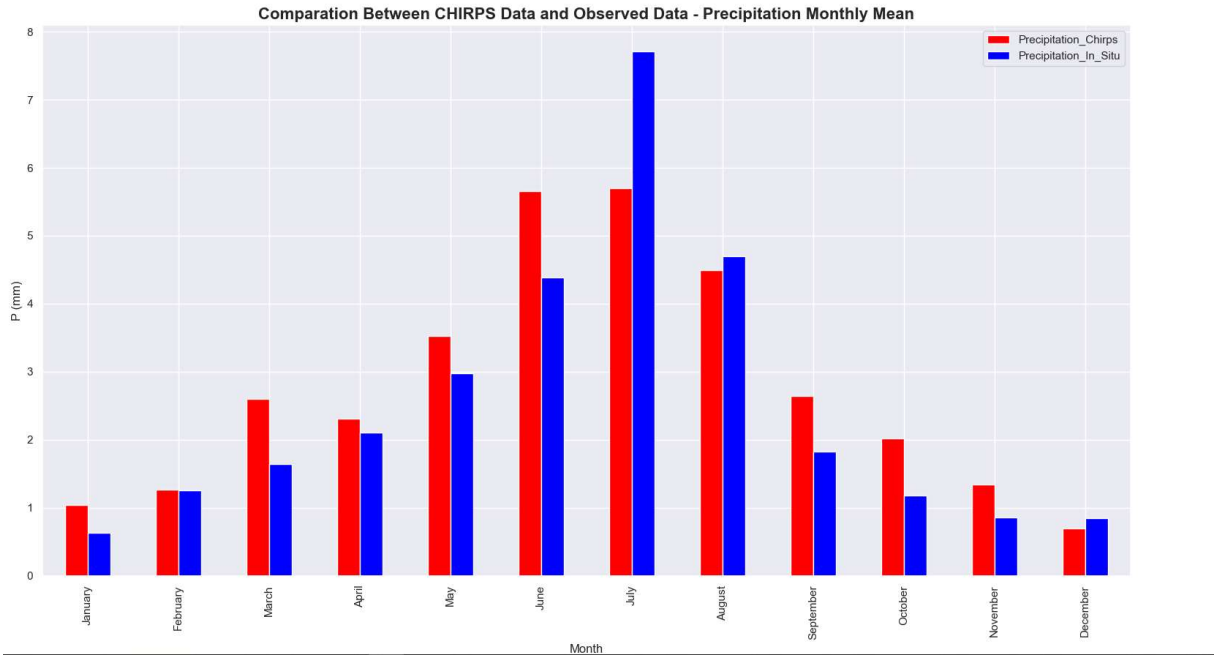


Figure 25 Precipitation Monthly Mean Comparation between Satellite Product and Observed Data

This analysis cannot be carried out for temperature and evapotranspiration since there is not enough in-situ information for these variables.

5.2 Model Calibration

For this research two conceptual models were calibrated, one considering a Max-Bas function and the other one Muskingum for the routing. The parameter to be optimized is KGE. Below there is a comparation between visual results, performance criteria and finally, the optimized parameters.

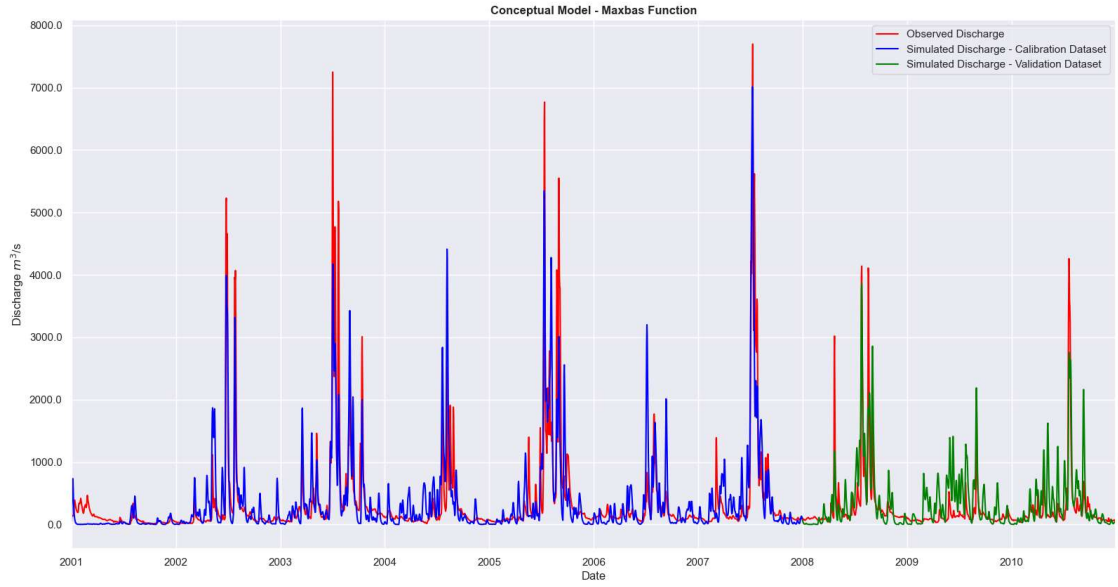


Figure 26 Observed and Simulated discharge Maxbas function - Calibration and Validation Dataset

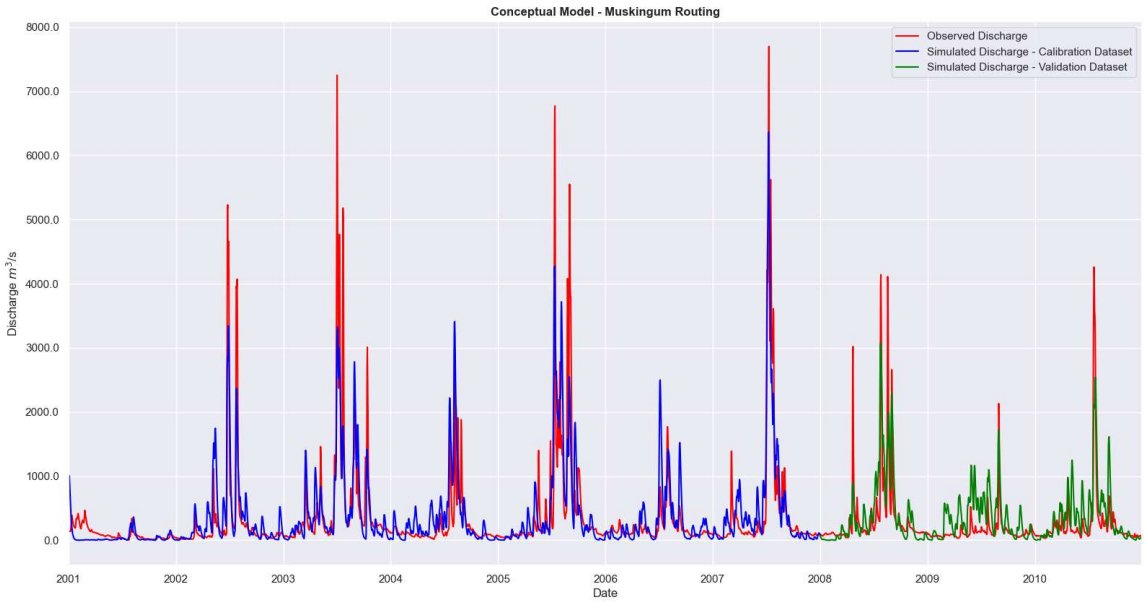


Figure 27 Observed and Simulated discharge Muskingum Routing - Calibration and Validation Dataset

As it can be seen in the previous figures, both conceptual models can reproduce reasonably good the observed runoff. However, the performance seems to be better for high flows than for low flows for both models. At the beginning of both simulations, there is a peak, and that the simulation takes almost 6 months, to start showing the same behaviour as the observed data, which is why this initial period will not be considered for the performance metrics, as it would be considered as a warm-up period.

Regarding the lows flows, there seems to be "water realising" pattern. As it can be seen in Figure 21, which is probably a result of the Nanwan reservoir, next to Xinjiang Shishankou reservoir at the south of the catchment, or the Suyahu reservoir next to the city of Zhumaian. Such human intervention will prevent the model to capture the relationship between rainfall and discharge at the outlet.

During the calibration period, both models show overall the same behaviour, and they manage to capture the hydrology in the catchment. However, during the validation dataset the Maxbas model shows a bigger overestimation for the low flows than the Muskingum one. This might be due to first the percolation coefficient and second the recession coefficient from the upper tank.

Percolation has a relative low value (see Table 7), which would mean that rainfall would be converted faster into direct runoff at the outlet than it should be. In order to allow a lower response during dry periods, this water should go to the groundwater tanks and should be stored there.

On the other hand, parameters K0 and K1 might be too high, and hence, this tank produces faster contribution to the outcoming discharge, instead of storing this water and releasing it, during rainy periods, when the thanks are more saturated. In Table 6, the performance metrics for the validation period can be seen.

Table 6 Performance Metrics for HBV Model

	Model	
Parameter	**MaxBas**	**Muskingum**
RMSE (m^3/s)	410.32	407.2
NSE	0.66	0.67
KGE	0.77	0.73
WB	85.19	84.87
R2	0.66	0.67
MAE (m^3/s)	219.11	214.26

Overall, both models show almost the same performance. However, as it was explained in section 4.5.4, once the rainfall objects reach their end, the optimized Maxbas parameter is added to the simulation to allow the catchment to respond to the precipitation event, and since the KGE (optimized parameter during calibration) has a slightly better performance for the Maxbas function, it is decided to use this model as the new representation of the reality for the object routing.

Table 7 Optimized Values for HBV Model

Parameter	Value
ttm	0.006
rcfc	0.810
sfcf	0.00030
Cfmax	0.036
Cwh	0.608
Cfr	0.884
fc	250.511
beta	2.860
e_corr	0.013
lp	0.302
k0	0.374
k1	0.297
k2	0.090
uzl	71.511
perc	0.100
maxbas	6.369

In Figure 28, there is a comparation between the simulated discharge and the observed discharge. Overall, the correlation coefficient between observed and simulated discharge is 82%.

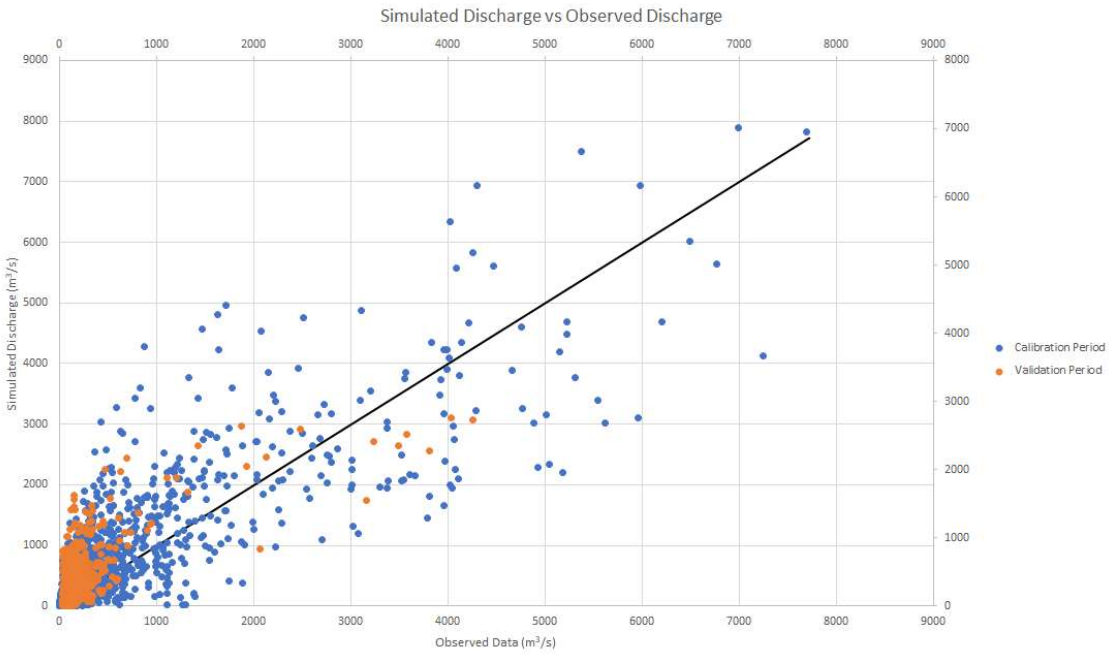


Figure 28 Simulated Discharge vs Observed Discharge - Maxbas Function

5.3 Critical Discharge

As explained in section 4.4.1, a frequency analysis is carried out in order to determine the critical discharge threshold. A summary table can be seen below:

TR	Excedencia	F(X)	Corrected Gumbel			Normal		Log Normal		Log Pearson III		Mean
	P ≥		Y	λ	Q (m^3/s)	Z	Q (m^3/s)	Z	Q (m^3/s)	U-0.81	Q (m^3/s)	
		1 - P ≥			μ+λ* σ		μ+Z* σ		μ+Z* σ		μ+U*σ	Q (m3/s)
2	0.5	0.5	0.37	-0.2	212.8	0.00	328.7	0.0	148.22	-0.1	128.1	204.48
5	0.2	0.8	1.50	0.7	836.4	0.84	922.6	0.8	370.56	0.8	346.0	618.87
10	0.1	0.9	2.25	1.3	1249.2	1.28	1233.0	1.3	598.22	1.3	635.3	928.92
20	0.1	1.0	2.97	1.9	1645.2	1.64	1489.3	1.6	888.46	1.8	1100.6	1280.89
50	0.0	1.0	3.90	2.6	2157.8	2.05	1777.9	2.1	1386.67	2.5	2154.2	1869.11
100	0.0	1.0	4.60	3.1	2541.9	2.33	1970.2	2.3	1865.79	2.9	3479.9	2464.45

Figure 29 Discharge Values for Different Return Periods

Now a visual inspection for the simulated discharge with CHIRPS precipitation information as input is conducted, giving the following results:

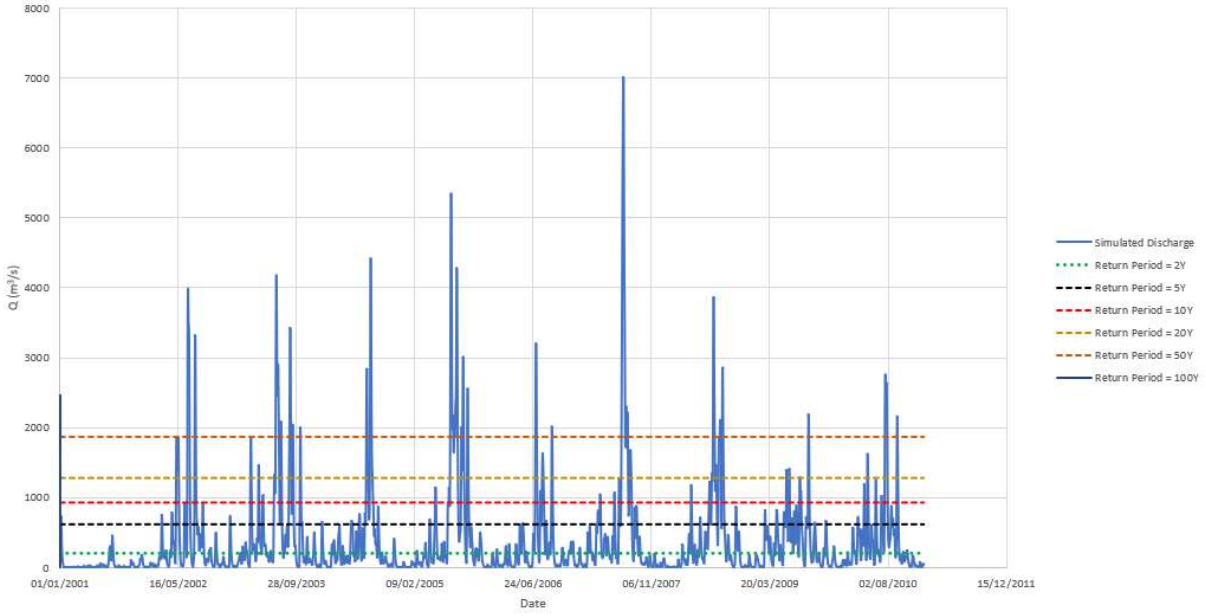


Figure 30 Simulated Discharge and Critical Discharge Possibilities

As it can be seen in Figure 30, return periods of 2 and 5 years, still cover some of the low flows. On the other hand, a return period of 10 years will be above these flows, and will hence, allow to focus the analysis on what it will be defined as "critical events".

As an additional part of this research, rainfall objects were created with CHIRPS data from 1990 to 2020. In order to do a similar process, as the done with the critical discharge definition, the characteristic of rainfall volume was used in order to carry out a frequency analysis for a 10-year return period. With this value, a matching object with a volume matching this return period was routed through the HBV model. The hydrological response can be seen in Figure 31.

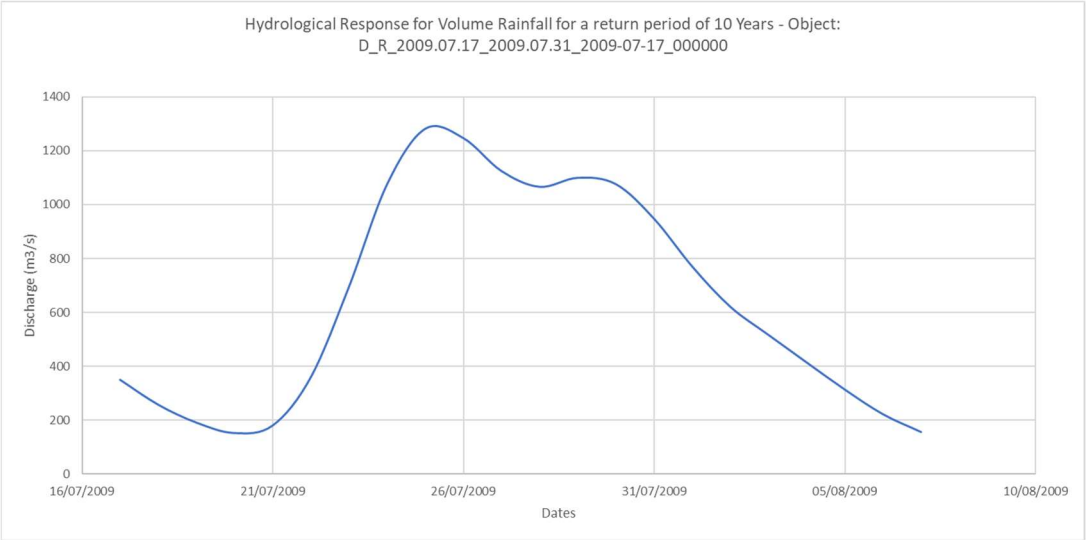


Figure 31 Hydrological Response for Volume Rainfall for a return period of 10 Years

This hydrograph has a peak on the 25/07/2009 with a value of 1271 m^3/s, which would be almost the same value as the critical discharge for the return period of 10 years for Gumbel distribution (1249 m^3/s) and the normal distribution (1233 m^3/s). This result is quite interesting, from the point of view that rainfall objects are considering connected precipitation in space and in time, which is how this phenomenon actually occurs in nature, instead of just a single instant value as can be the observed discharge in one station.

5.4 Rainfall Objects Characterisation

Before routing the objects, these must be created following the algorithm described in section 4.4. First, and as it can be seen in Figure 9, rainfall in the region tend to be group in periods. Hence, objects will be grouped in the following way:

- S1: (Driest season): December, January, and February
- S2: March, April, and May
- S3: (rainy season): June, July, and August
- S4: September, October, and November

In total, 25725 rainfall objects were identified in the period from October 2006 to December 2010. These rainfall objects are distributed in the following way:

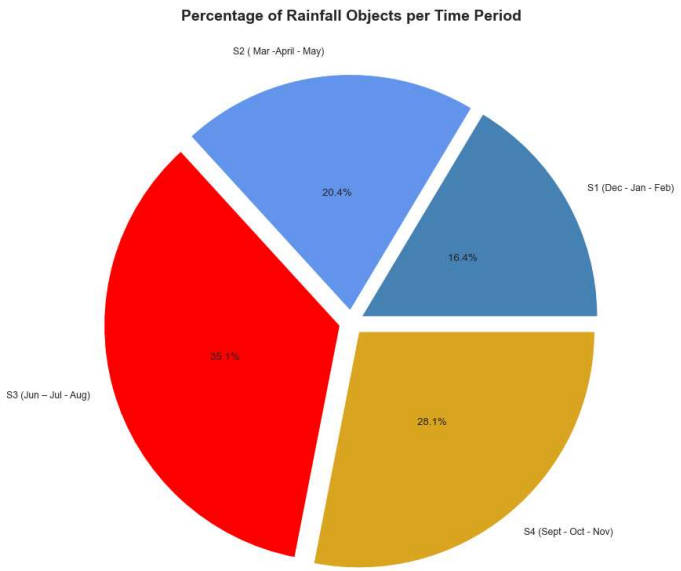


Figure 32 Percentage of Rainfall Objects per Time Period

As it can be seen in Figure 32, most rainfall objects occur during rainy season, which it was an expected result. However, there are more objects in S4 than in S2, even if during S4, there is less monthly precipitation than in S2 (see Figure 9), which might lead to think that during the last season there is more rainfall frequency but with less intensity. In order to assess this, a simple histogram comparing these two seasons can be done:

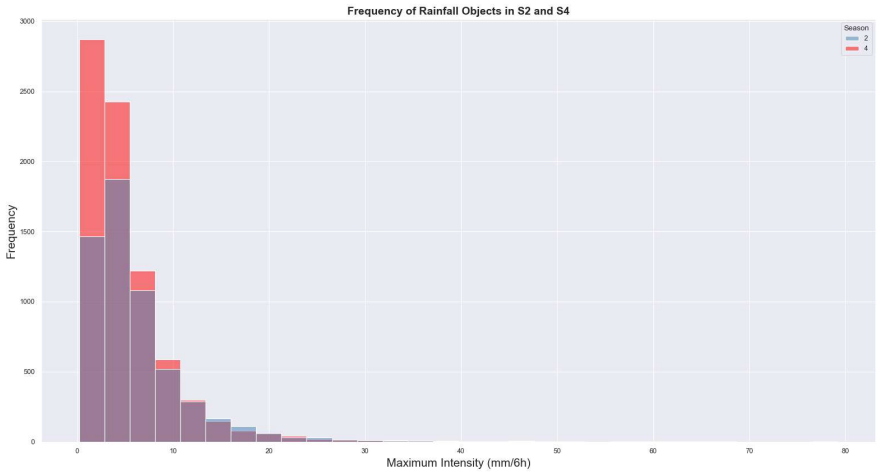


Figure 33 Frequency of rainfall objects in S2 and S4

As expected, during S2 there are less events but more intense, on the other hand during S4 there will be more events but less intense.

Spatial Localization

Regarding the spatial distribution of the rainfall objects:

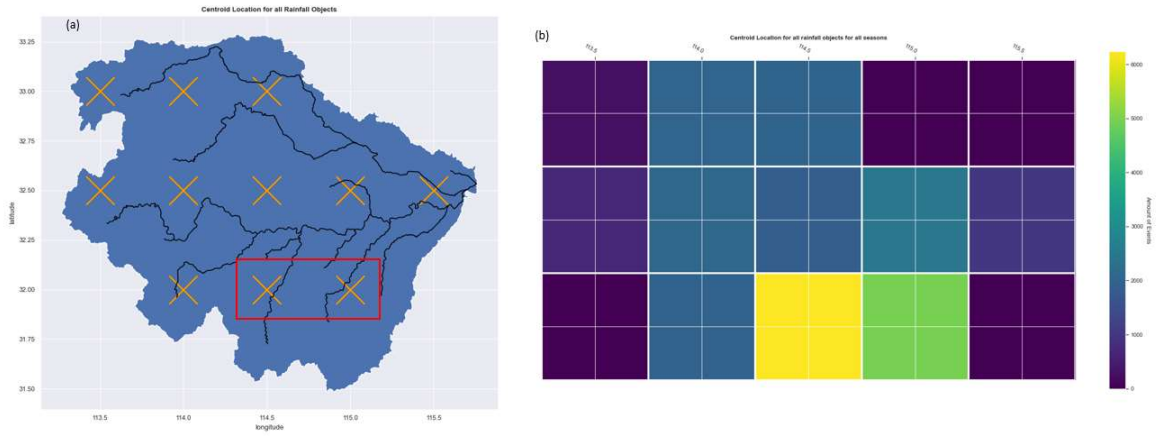


Figure 34 (a) Centroid Location Rainfall Objects – (b) Heatmap Amount of Rainfall Objects for all Time Periods

When looking at Figure 34, it can be seen that rainfall events tend to be located at 32°N and between 114.5°W – 115°W. For the study period 11203 rainfall events had place between these coordinates, accounting for 44% of the total. Repeating this analysis, but now for each season, the following results are found:

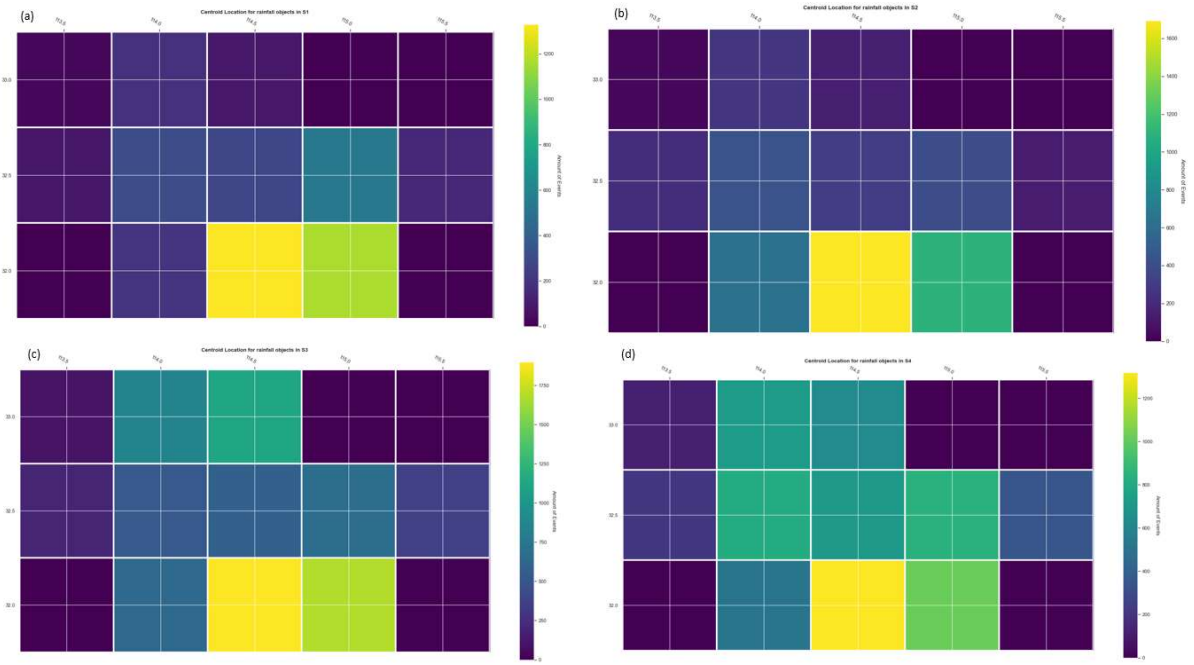


Figure 35 (a) Heatmap for Amount of Rainfall Objects for S1, (b) Heatmap for Amount of Rainfall Objects for Time Period 2, (c) Heatmap for Amount of Rainfall Objects for Time Period 3, (d) Heatmap for Amount of Rainfall Objects for Time Period 4

Overall, all events tend to concentrate in the same location. However, it is worth noticing that events during S1 and S2 tend to take place at the south of the catchment. During rainy season (S3), almost the same amount locates at the south (1800 at 32°N-114.5W), but also events appear at the north of the catchment (33°N), and during S4, and to difference to all other seasons, major events take place in the centre of the catchment at latitude 32.5°.

Properties Relationship

In order to see whether there is some relationship between the properties of the objects, a simple correlation matrix can be generated:

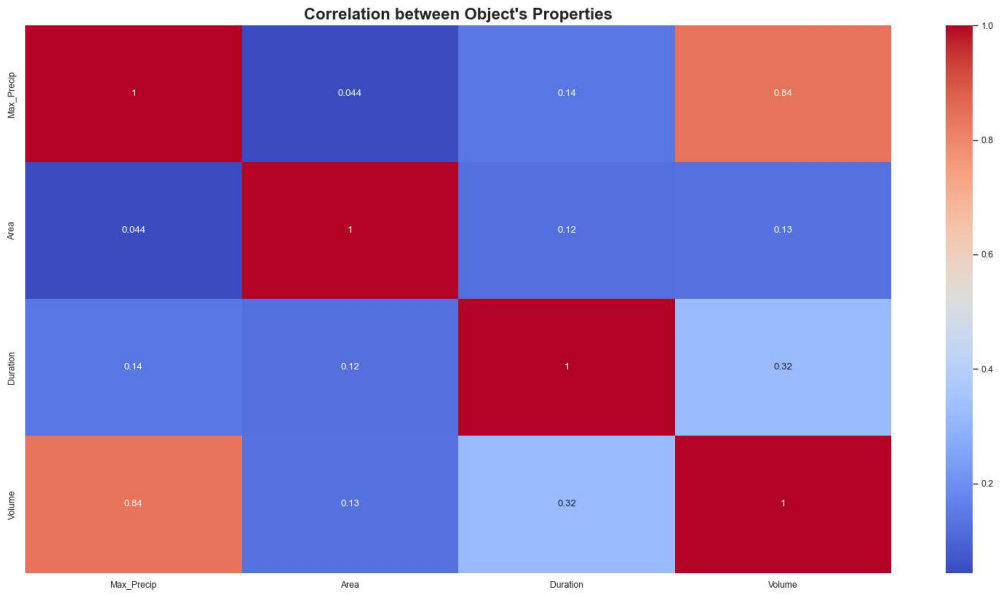


Figure 36 Correlation Matrix for Object's Properties

As it can be seen in Figure 36, the best correlation between the variables is the one between, the maximum precipitation intensity and the volume of rainfall, which has a Pearson coefficient value of 0.84. The duration of the rainfall objects shows some weak correlation with the object volume and, finally the covered area seems to show some random results, meaning that the spatial extent of the object does not imply stronger or weaker events in terms of volume, duration, and intensity.

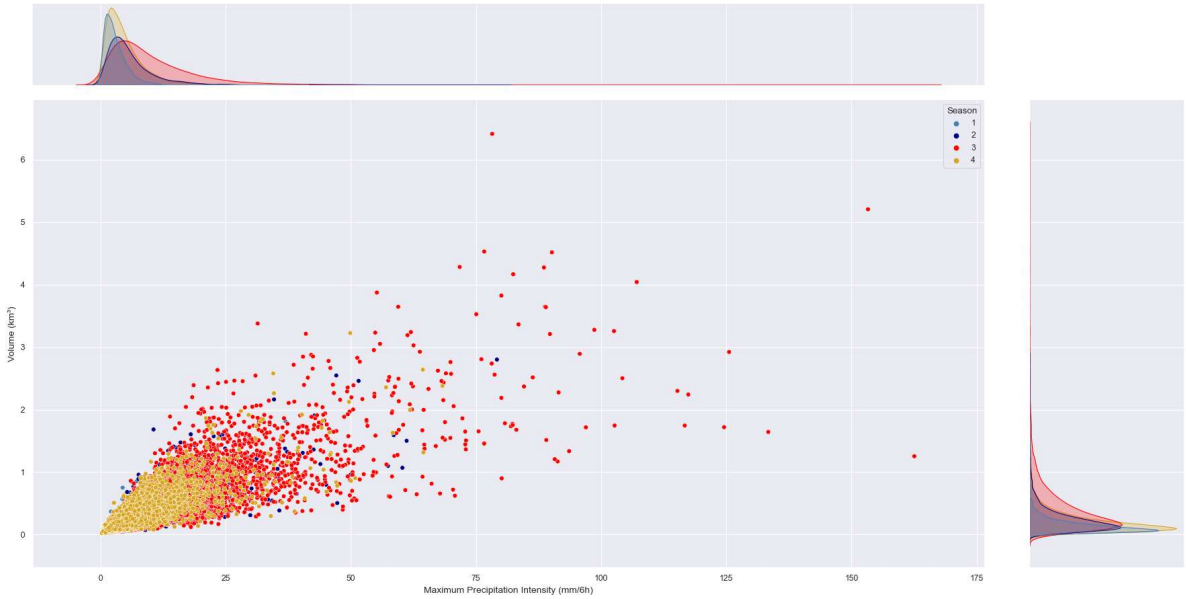


Figure 37 Object Volume vs Maximum Precipitation Intensity

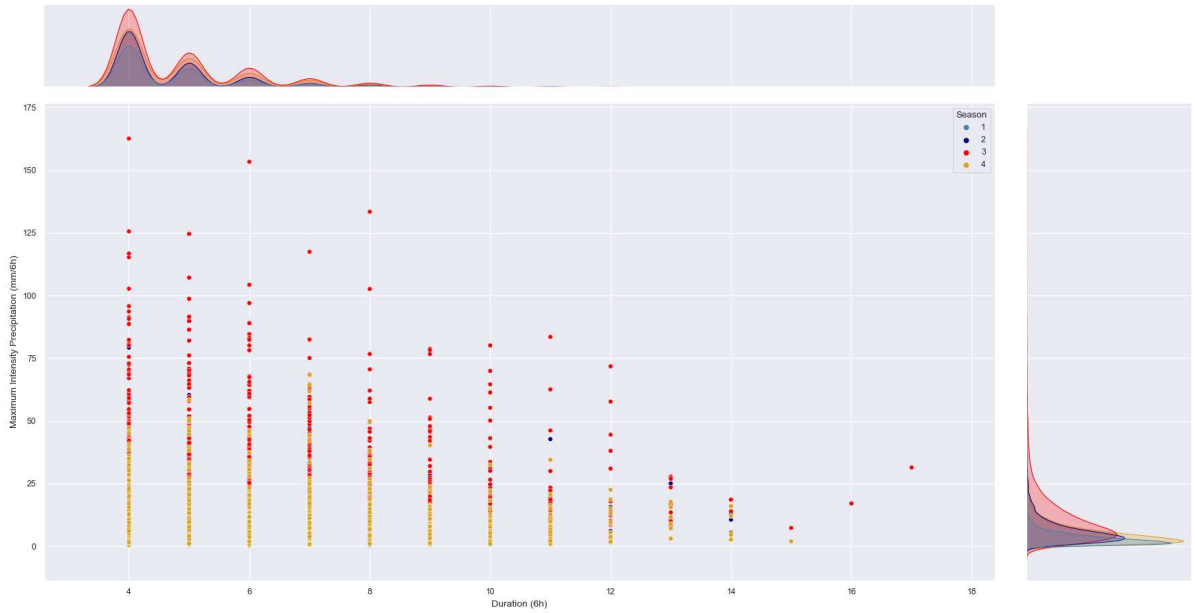


Figure 38 Maximum Precipitation Intensity vs Volume Object

In Figure 37 can be seen the correlation between volume and precipitation intensity. From the graph, it can be noticed that objects during S1 tend to have a high frequency for low values, and this increases with the season until reaching S3, where higher volumes and intensities appear, these results are expected, since it corresponds to rainy season. It is worth noticing that for S4, objects tend to be in the left part of the graph, but having much more frequency, which would confirm what it was noticed in Figure 33.

In Figure 38, during S3, shorter events occur but with higher intensity (and hence higher volume), and the frequency is much bigger than for all other seasons. Once the duration starts to increase (around 5 timesteps or 30 hours), the gap between the frequency S4 and S3 starts to decrease, and the objects tend to occur more during S4 (longer events but less intense).

5.5 Hydrological Response Characterisation

Once the rainfall objects are created, these can be routed through the calibrated model, obtaining the following properties:

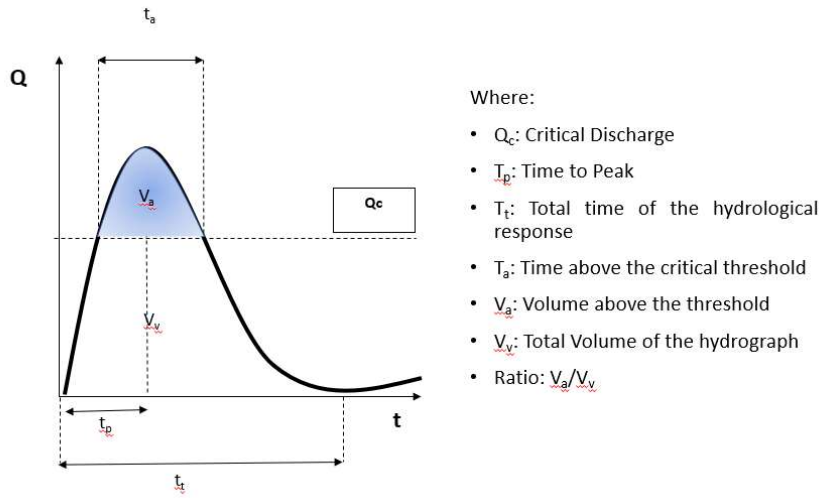


Figure 39 Hydrological Response Properties

Since the objective is to focus on objects, that can exceed the critical threshold (objects of interest). A first filter can be done to see how events are distributed per time period:

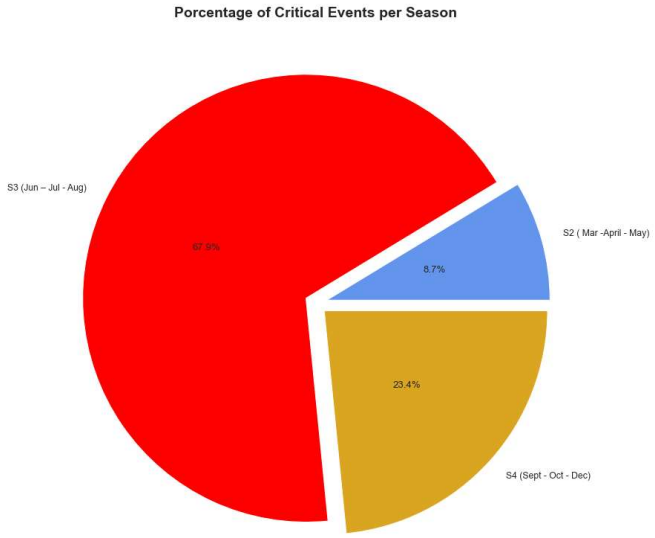


Figure 40 Distribution of Critical Events Frequency per Season

From the 27873 initial objects, only 2540 exceeded the critical threshold (9.11% of the total). When comparing Figure 40 with Figure 32, it can be noticed that critical hydrological responses do not occur during S1. The reason for this is due to their low intensity and rainfall volume. The same reason applies to S2, where even if events during this period account for 20.4% of all the total events, it only accounts for 8.7% of all critical hydrological responses associated to those events.

When looking at the correlation between all objects' properties and all hydrological responses the following is found:

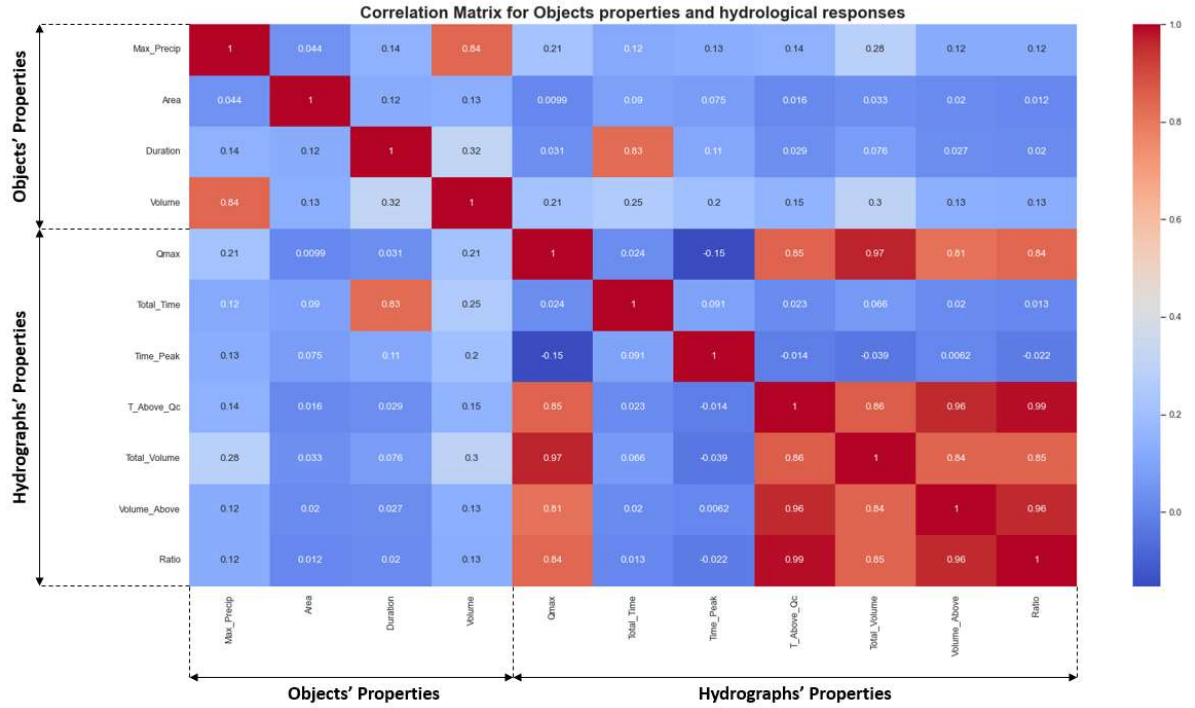


Figure 41 Correlation Matrix for Objects Properties and Hydrological Responses

The correlation matrix shown in Figure 41, confirms some expected results such as:

- The longer the duration of the rainfall object, the longer it will take the hydrological response.
- Maximum precipitation intensity and total volume of rainfall have a direct correlation with the volume of the hydrograph, and hence with its peak (they have a correlation value of 0.97). What it would mean that these two properties tend to assess how severe will be the hydrological response and, therefore, can be good indicators to predict a critical response at the catchment outlet.

Additionally, and as previously mentioned, covered area seems to lack correlation with other variables, and does not contribute too much to define whether a critical event would occur or not.

Finally, and regarding the time to maximum discharge it was found, that 92% of the total amount of events will have a fast response, 72 hours or less since the start date of the rainfall object (see Figure 42). 44% of rainfall events will be translated into peak discharge in 24h, while in 72h, will account for another 29% of the analysed events.

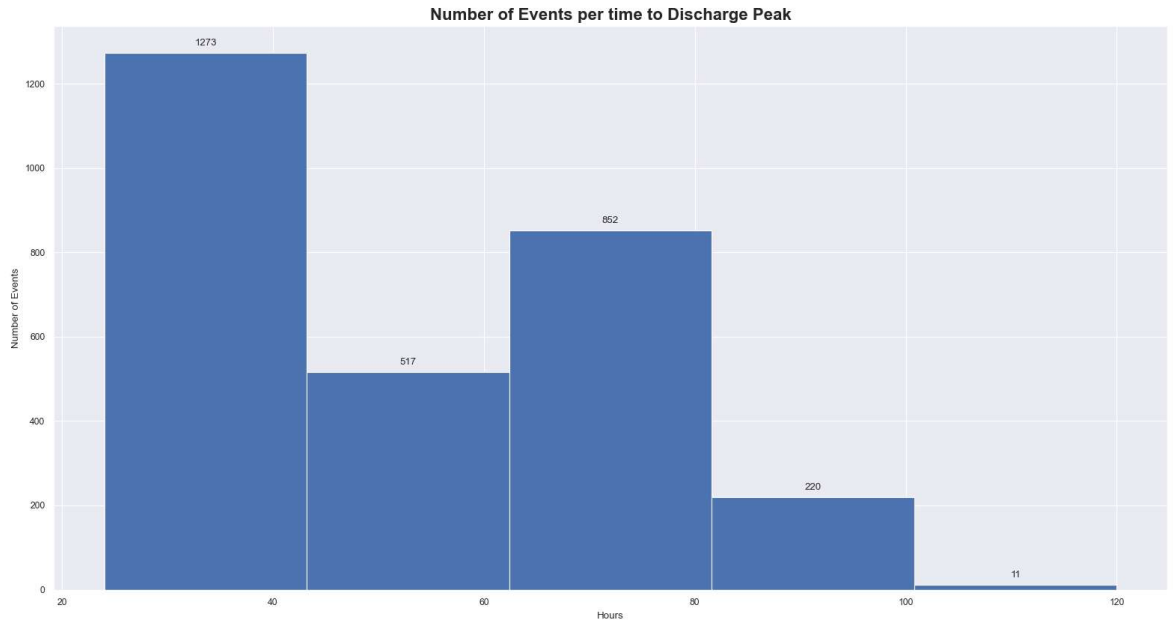


Figure 42 Number of Events per time to Maximum Discharge

If this characteristic is then distributed by time periods, it can be seen that during S4, short response times will be translated into high responses, but during S3 longer time catchment response will produce higher peaks during the summer.

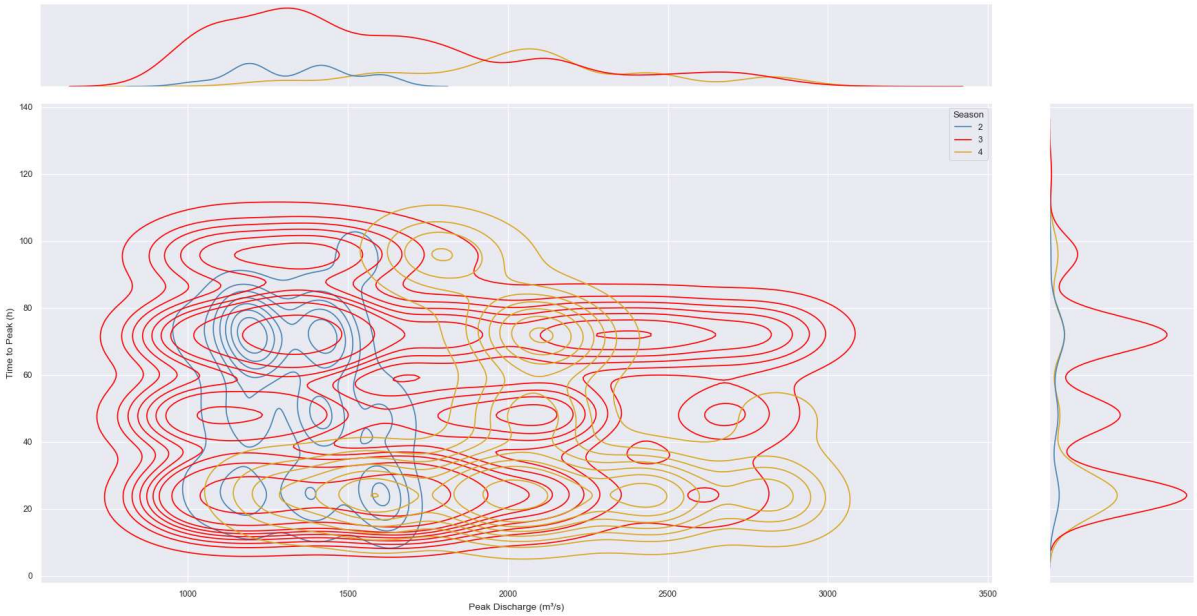
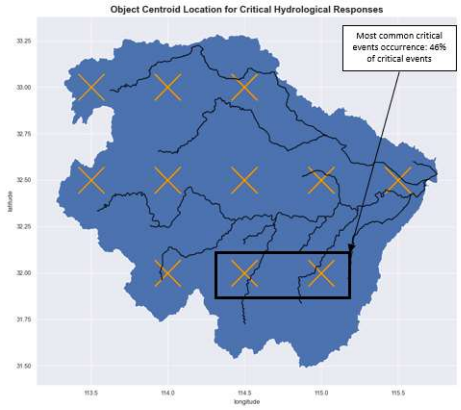


Figure 43 Comparation between Time to Peak and Peak Discharge per Season

Spatial Distribution

Regarding the spatial location of the objects' centroid that generate events above the critical threshold, the following is found:



Centroid Longitude (°)	Centroid Latitude (°)	Frequency	Percentage
114.5	32.0	632	25%
115	32.0	529	21%
115	32.5	258	10%
114.5	33.0	209	8%
114	33.0	197	8%
114	32.0	182	7%
114.5	32.5	161	6%
114	32.5	157	6%
115.5	32.5	140	6%
113.5	32.5	48	2%
113.5	33.0	27	1%
	Σ	2540	100%

Figure 44 Objects' Centroid Location for Critical Hydrological Responses

When comparing Figure 34 and Figure 44, it can be seen that the pattern remains, meaning that most objects of interest (those who produce responses above the threshold) are located at latitude 32°. Doing the distribution per seasons, the following is obtained:

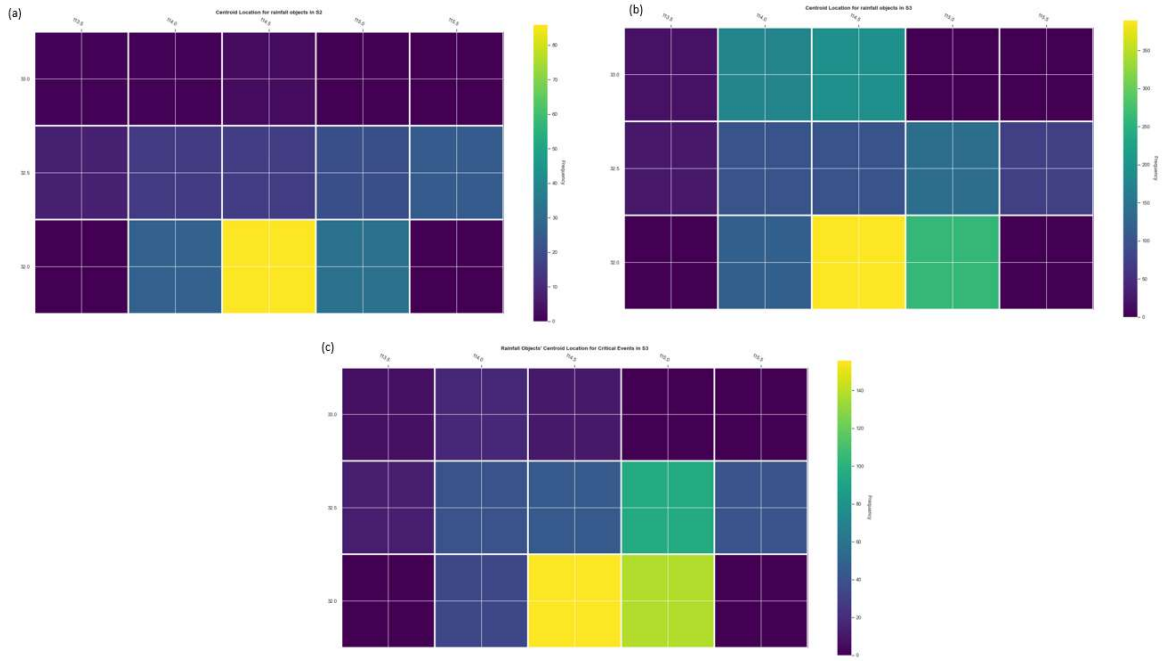


Figure 45 (a) Heatmap for Amount of Rainfall Objects Location Centroid that produce critical events for S2, (b) Heatmap for Amount of Rainfall Objects Location Centroid that produce critical events for S3, (c) Heatmap for Amount of Rainfall Objects Location Centroid that produce critical events for S4

When comparing Figure 45 with Figure 35, it is noticed that overall, critical events will be produced by rainfall events located at the south of catchment. For S3, the distribution is more along the catchment while, for S2 and S4 it is concentrated. In Figure 35, it was shown that rainfall objects' centroids locations tend to be distributed along the study area for S4, but in Figure 45 can be seen that the most intense ones will take place at latitude 32° between longitudes 114.5° and 115°, and in latitude 32.5° and latitude 115°.

If a new threshold of 2800 m^3/s were defined, the distribution of the objects would change as it follows:

Extreme hydrological responses for the new threshold only take place in S4 and S3, and even if in S4 object rainfall objects tend to be less in intensity and volume that those in S3, for this new threshold, more critical events occur in S4 (59) than in S3 (24).

In Figure 41, it was observed, how the peak discharge of the hydrological response is a good indicator on the severity of itself, hence this parameter will be the reference to see how the events change along the time periods.

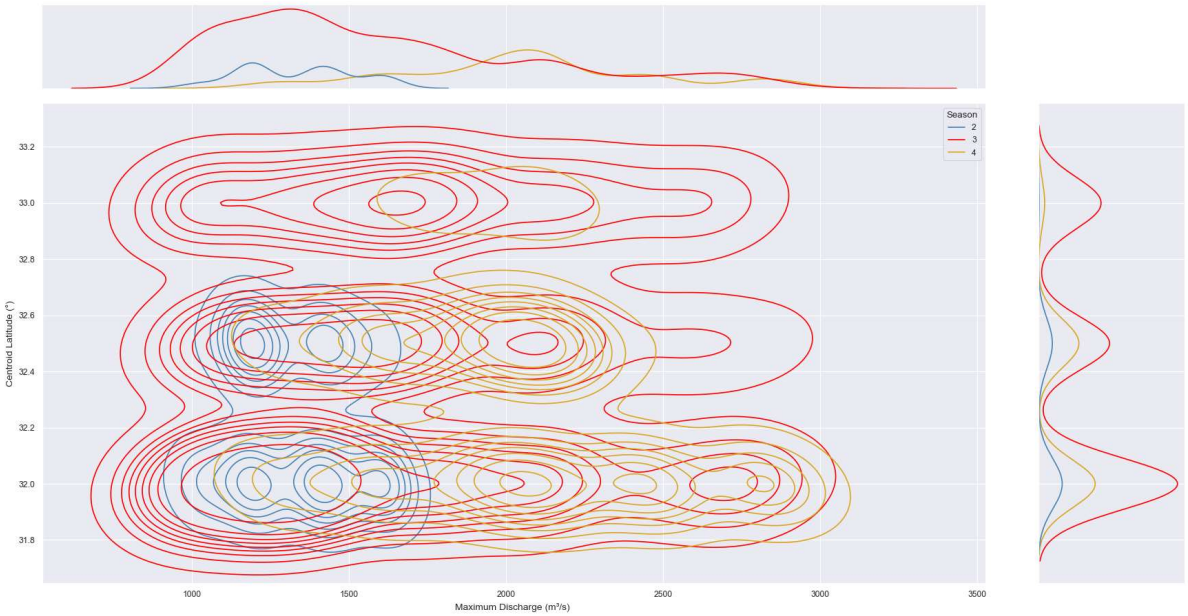


Figure 46 Centroid Latitude vs Maximum Discharge for Critical Events

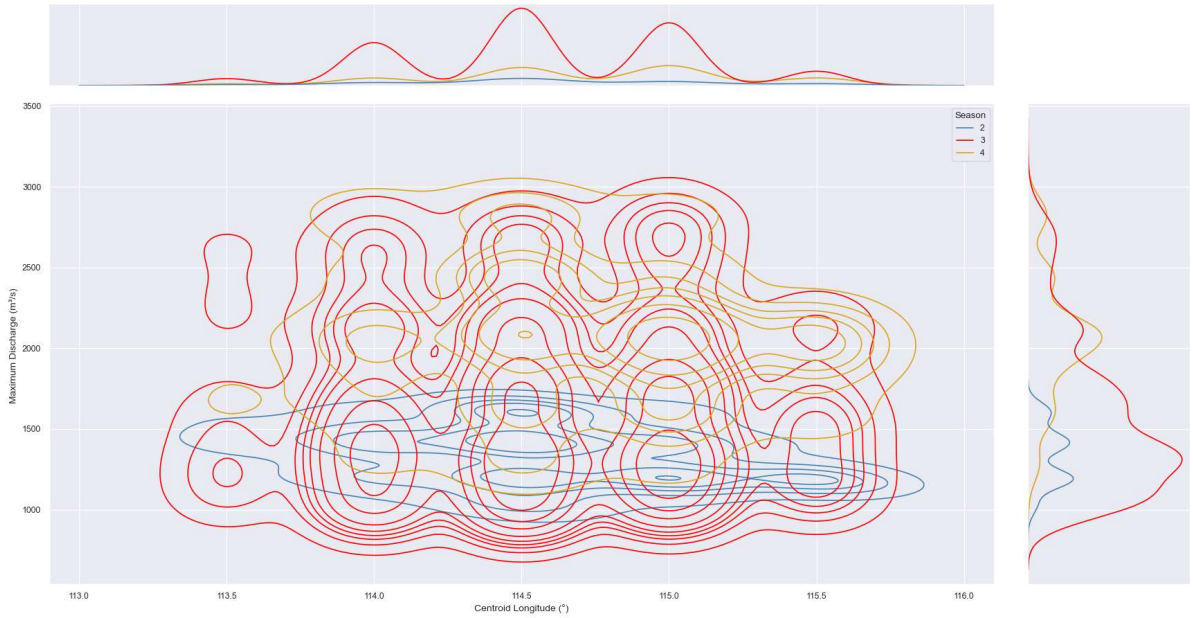


Figure 47 Centroid Longitude vs Maximum Discharge for Critical Events

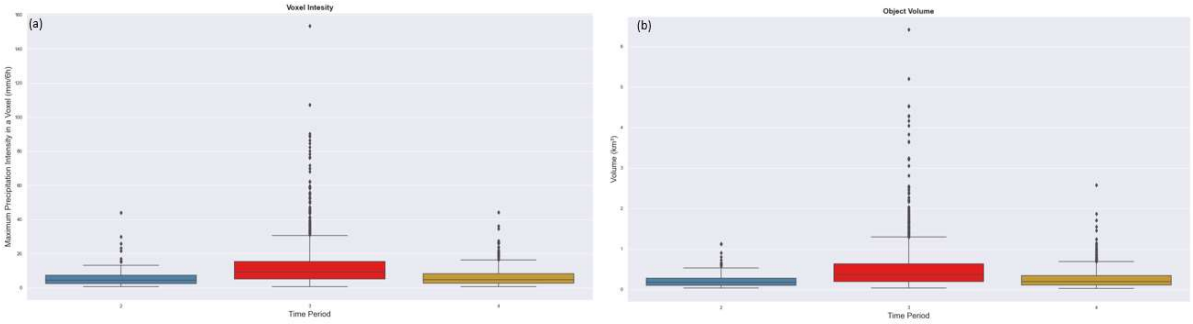
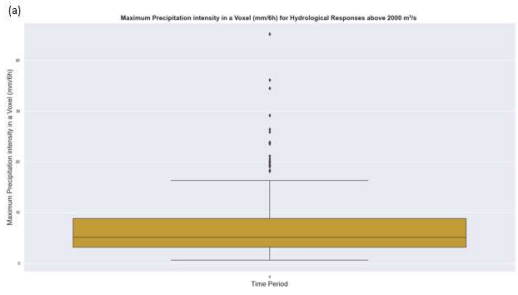


Figure 48 Main Object Properties Values Distribution per Time Period

From Figure 46 and Figure 47, the following can be observed:

- S3 and S4 show the most severe events, while S2 does not reach discharges above 2000 m^3/s.
- Peak discharge tends to increase for S2, S3 and S4, while the objects centroid moves to the south of the catchment.
- Along the longitude axis, no tendency is recognized. However, there is a concentration of objects along the coordinate 114.5°, which would indicate that objects' centroid located at 32°N – 114.5° tend to produce the biggest frequency of critical events.
- Figure 48 shows that for intensity and volume, S3 displays bigger values. However, and as it can be noticed in Figure 46 and Figure 47, S4 shows bigger peak discharges, which can be contradictory. In Figure 49(a), it is shown the intensity for objects that are above the critical threshold during S4. As it can be seen these values are quite low and should not be converted into high discharges. However, in Figure 49(b), the days on which these events occurred show a pattern and it is that they all occur in the first 14 days of September, which means that it is not the rainfall what triggers the high response, but the fact that the soil is saturated from the rainy season and less intense events can produce high flows in the river.



(b)

Day	Frequency
01/Sep	14
02/Sep	53
03/Sep	93
06/Sep	1
07/Sep	4
08/Sep	96
09/Sep	52
10/Sep	75
14/Sep	1

Figure 49 (a) Maximum Precipitation intensity in a voxel (mm/6h) for Hydrological Responses Above 2000 m^3/s, (b) Start date of hydrological response with peak discharge above 2000 m^3/s for S4.

A general overview from the properties' objects that generate high flows can be seen in Table 8.

Table 8 General Overview Objects' Properties that generate High Flows

	Time Period								
	S2			S3			S4		
	Maximum Intensity (mm/6h)	Duration (h)	Volume (km^3)	Maximum Intensity (mm/6h)	Duration (h)	Volume (km^3)	Maximum Intensity (mm/6h)	Duration (h)	Volume (km^3)
Mean	5.8	27.7	0.218	12.0	30.0	0.500	6.2	30.4	0.271
Min	0.7	24.0	0.033	0.5	24.0	0.030	0.6	24.0	0.021
25%	2.5	24.0	0.097	5.1	24.0	0.188	2.6	24.0	0.107
50%	4.5	24.0	0.178	9.2	24.0	0.360	4.6	30.0	0.191
75%	7.3	30.0	0.280	15.4	36.0	0.631	8.2	36.0	0.341
Max	43.9	54.0	1.130	153.3	78.0	6.415	44.2	84.0	2.581

5.6 Classification Algorithm

For this step, a trial-and-error procedure was followed, first all data from all seasons was used as input and a data splitting 70/30 was carried out. The used attributed were the followings:

- Centroid coordinates (Longitude and Latitude)
- Maximum Voxel Intensity
- Duration
- Volume
- Season

The results were extremely poor, obtaining an overall precision (for predicting and not predicting critical events) of 17%. However, and it was presented in section 5.4 and 5.5, rainfall objects tend to show similar behaviours seasonally, which is why they were separated into 3 decision trees, each for each season. For S1, since there are no critical events occurring during this period, it did not make sense to train a model. For the other threes, metrics started to improve as it can be seen in the following classification report:

Table 9 Classification Report Decision Tree – Second Attempt

	S2				S3				S4			
	Precision	recall	f1-score	support	Precision	recall	f1-score	support	Precision	recall	f1-score	support
0	0.97	0.93	0.95	1559	0.8	0.78	0.79	2083	0.92	0.92	0.92	1995
1	0.07	0.14	0.09	58	0.22	0.24	0.23	524	0.17	0.17	0.17	185
accuracy			0.9	1617			0.67	2067			0.86	2180
macro avg	0.52	0.53	0.52	1617	0.51	0.51	0.51	2067	0.55	0.55	0.55	2180
weighted avg	0.93	0.9	0.92	1617	0.69	0.67	0.68	2067	0.86	0.86	0.86	2180

Where 1 means critical event and 0 means no critical event. An improvement from the first attempt was noticed. However, the algorithm still fails to predict a critical event, and it performs poor when predicting high catchment responses. In Figure 49 was shown, how major events to occur during the first days of September, which is mainly due to the state of the catchment, which due to seasonality tend to be the same year after year. Keeping this in mind a new feature

was added, and it was the start date of the rainfall object. When adding this new attribute, the metrics improved in the following way:

Table 10 Classification Report - Decision Tree - Third Attempt

	S2				S3				S4			
	Precision	recall	f1-score	support	Precision	recall	f1-score	support	Precision	recall	f1-score	support
0	0.98	0.98	0.98	1553	0.9	0.9	0.9	2087	0.96	0.95	0.96	1998
1	0.48	0.47	0.48	64	0.6	0.6	0.6	520	0.53	0.58	0.55	182
accuracy			0.96	1617			0.84	2607			0.92	2180
macro avg	0.73	0.72	0.73	1617	0.75	0.75	0.75	2607	0.74	0.77	0.76	2180
weighted avg	0.96	0.96	0.96	1617	0.84	0.84	0.84	2607	0.93	0.92	0.92	2180

As it can be seen in Table 10, the metrics improved much more compared to the previous attempt, generating an average precision of 73%. Confirming, what it was expected, regarding the state of the catchment along the year.

Finally, and making an analogy where the river is a bucket. A bucket that is almost full will need a small amount of water to spill. On the other hand, an empty bucket can take a bigger amount of water before overtopping. Keeping this in mind another variable will be considered, and it is the observed discharge three days before the start date of the event. Since observed data is considered, three days seem a suitable value for data collected from the station to be ready for its use. The results when considering this new attribute are the following:

Table 11 Classification Report - Decision Tree - Fourth Attempt

	S2				S3				S4			
	Precision	recall	f1-score	support	Precision	recall	f1-score	support	Precision	recall	f1-score	support
0	1	1	1	1540	0.99	0.99	0.99	2093	1	1	1	2008
1	0.97	0.94	0.95	77	0.96	0.94	0.95	514	0.99	0.99	0.99	172
accuracy			1	1617			0.98	2607			1	2180
macro avg	0.98	0.97	0.98	1617	0.97	0.97	0.97	2607	0.99	0.99	0.99	2180
weighted avg	1	1	1	1617	0.98	0.98	0.98	2607	1	1	1	2180

As it can be seen in Table 11, performance is almost perfect. Meaning that rainfall objects' characteristics can predict the occurrence of critical events based on their intrinsic characteristics plus the observed discharge in the station three days before the rainfall event object start date.

Chapter 6 Conclusions and Recommendations

6.1 Conclusions

This research had as main objective to explore the relationship between rainfall objects and their hydrological response, with input data from EPS and the help of a conceptual lumped hydrological model, in order to explore its potential use in predicting critical events.

First, preparatory analyses were of precipitation data, preparing a hydrological model, and defining critical discharge levels were done (Sections 5.1, 5.2, and 5.3). It was found that data from CHIRPS matches relatively well the observed data in the study area, producing a MAE of less than 1 mm/d for the period time of interest, which would be in line with what other authors have found regarding this satellite-product. Unfortunately, for the other parameters (Evapotranspiration and Temperature) there was not enough in-situ data in order to generated grid data and make do the analysis. This part is presented in Section 5.1.

In order to represent the hydrological processes inside the catchment, an HBV lumped conceptual model was calibrated through the python open-source HAPI library. Both models performed quite well and were able to react at the same time as the observed data, meaning that they capture well enough the hydrological behaviour. However, there were some issues with overestimation, especially during low flows, which could be caused by high recession coefficients that produce discharge during dry seasons, instead of storing the water to release it during winter periods, and hence, improve the peak values during the former. Additionally, it was noticed that some discharge patterns do not correspond to rainfall responses, which might be due to human intervention in the zone and the considerable number of reservoirs in the zone, which might difficult the calibration process, since this conceptual model does not consider reservoir releases. This part is shown in section 5.2.

Regarding the critical discharge definition, some evidence was found that volume from generated objects can be used as parameter for frequency analysis as shown in section 5.3. This could lead to further investigations on how frequency analysis is carried out, since objects consider the spatiotemporal connectivity in rain, instead of being a single isolated value such as precipitation intensity or observed discharge, which are common approaches nowadays when using frequency analysis.

The generated rainfall objects were distributed along 4 time periods, seasons, using as criterion monthly mean precipitation values. Objects showed more frequency during S3 and S4, which made up for more than a half of all rainfall objects. These objects' centroids tend to locate in the south of the catchment and between longitudes 114.5° and 115°, where almost half of all generated objects took place. Regarding, the correlation between the objects' properties it was found that only three of them: Intensity, volume and duration show correlation between them, while the covered area for the objects shows poor relationship with the first three. This can be found in section 5.4.

From the relationship between rainfall objects and its hydrological response, it was found that the leading parameter is maximum voxel intensity, which has a strong correlation with the

object volume, leading to higher hydrological responses **(Research question 1)**. It was found that high responses tend to be produced by objects concentrated at the south of the catchment **(Research question 2)**. Additionally, It was also noticed that the most critical season for major catchment's response will be S3 and the first 14 days of S4, this not due to the intensity of the objects, but because of the state of the catchment. Finally, it was found that from 92% of all generated objects will reach peak discharge in a time period from 24h-72h since the start date of the object. This was presented in section 5.5.

Finally, and as it was presented in section 5.6, it was possible to predict the occurrence of critical flood events based on a classification algorithm such as decision trees and the object's properties **(Research question 3)**. When using only intrinsic properties of the rainfall objects, such as: Centroid location, maximum voxel intensity, duration, volume and start day it was possible to predict critical events with more than 70% of average precision for all seasons, and when adding observed discharge from three days before the object's start date as variable that represents the state of the river, the performance was almost perfect. This is interesting, because: first, it is already known that the catchment will respond most likely (92% of studied cases) in a time period from 1-3 days, second EPS has a time window of 15 days from the day it was generated. Third, no parameter from the hydrological model was used for the prediction and finally object-based methodologies such as ST-CORA are much faster in terms of computing time than the set-up and running of a calibrated hydrological model. All these reasons, allow the author to believe that object-based methodologies have a true potential for improving and expanding leading and preparation time when facing forecasted major rainfall events.

6.2 Outlook and recommendations

In this study, a lumped conceptual model was used. However, and ideally a distributed model that considers land use should be used. In combination, with a distributed model, a bigger study case should be considered. EPS has a coarse resolution of 0.5°x0.5°, meaning that for the study area only 12 cells could be used, which shorts a bit the potential of the research. A bigger region would allow a better spatial analysis. In case, this is carried out, a bigger region, would imply a more complex distributed model, which should be considered in time planning. Additionally, the hydrological model was calibrated for timesteps of a day. Acknowledging the potential of the topic developed in this research, it would be ideal to use a finer time resolution, to understand better the response of the catchment.

The idea behind volume's objects as a variable for frequency analysis, should be further explored.

Due to time limitations, this research reached a stage where major flood events could be predicted, knowing that they will occur, most likely, in a time window from 1-3 days from the start date of the object that generates them, but a detailed hindcast analysis with specific lead times and comparison to, for example, standard rainfall-runoff model-based ensemble flood forecasting, could not be implemented. This is recommended for future research.

Other recommendations for further research include:

- The classification algorithm used in this research was a first step. However, it is still important to see how the results are governed by the 3-days-before observed discharge, since this is where the metrics became almost perfect. For this future work, should attempt to predict the exact day on which the threshold is exceeded and how it is related to the rainfall object's properties.
- Use a distributed model with a finer timestep resolution that can provide more information for the time where the peak is reached (this research was carried out with daily information, hence, more precision in time to peak is not possible).
- Make use of this finer model in order to attempt to find a relationship between the time centroid of the object and the time to peak of the hydrological response.
- The state-of-the-art in the forecasting field is EPS, meaning that objects could be used in order to weight the perturbed members to which they belong, based on their expected hydrological response (as it was done in this research), in order to "reduce interest" on those perturbed members that do not contain objects of interest (that will not produce high hydrological response), and hence reducing the uncertainty when predicting a major event when using EPS, since not all members might contain objects of interest.

Chapter 7 References

Akhtar M, Corzo G, Van Andel S, Jonoski A (2009) River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin. Hydrology and Earth System Sciences 13: 1607-1618

Alfieri L, Burek P, Dutra E, Krzeminski B, Muraro D, Thielen J, Pappenberger F (2013) GloFAS – global ensemble streamflow forecasting and flood early warning. Hydrology and Earth System Sciences 17: 1161-1175 DOI 10.5194/hess-17-1161-2013

Andel S-Jv (2009) Anticipatory water management : using ensemble weather forecasts for critical events. CRC ;

[Taylor & Francis [distributor] [in Englisch]

Aye PP, Koontanakulvong S, Long TT (2017) Estimation of Groundwater Flow Budget in the Upper Central Plain, Thailand from Regional Groundwater Model. Ssms Jp 11: 90-100

Bai L, Shi C, Li L, Yang Y, Wu J (2018) Accuracy of CHIRPS satellite-rainfall products over mainland China. Remote Sensing 10: 362

Bergström S (1992) The HBV model–its structure and applications SMHI

Cheng C (2019) Can machine learning reduce uncertainties of distributed hydrological models? Case study of Dapoling- Wangjiaba catchment Delft : IHE Delft Institute for Water Education;.

Cloke H, Pappenberger F (2009) Ensemble flood forecasting: A review. Journal of hydrology 375: 613-626

Corzo Perez GA (2009) Hybrid models for hydrological forecasting: Integration of data-driven and conceptual modelling techniques

Corzo Perez GA, Varouchakis EA (2019) Spatiotemporal analysis of extreme hydrological events Elsevier, Amsterdam, Netherlands

Cunge J (1969) On the subject of a flood propagation computation method (Musklngum method). Journal of Hydraulic Research 7: 205-230

Davis CA, Brown BG, Bullock R, Halley-Gotway J (2009) The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. Weather and Forecasting 24: 1252-1267

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda M, Balsamo G, Bauer dP (2011) The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society 137: 553-597

Demeritt D, Cloke H, Pappenberger F, Thielen J, Bartholmes J, Ramos M (2007) Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. Environmental Hazards 7: 115-127 DOI 10.1016/j.envhaz.2007.05.001

Fact sheet: Ensemble weather forecasting (2017). Cited 25/02 2022

Farrag M (2018) Spatio-temporal simulation of catchment response based on dynamic weighting of hydrological models : case study, Jiboa catchment in El Salvador. UNESCO-IHE [in Onbepaald]

Global Flood Awareness System (GloFAS) (2021) https://www.globalfloods.eu/general-information/about-glofas/. Cited October 2021

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377: 80-91

Hopson TM, Webster PJ (2010) A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. Journal of Hydrometeorology 11: 618-641

Hu C, Wu Q, Li H, Jian S, Li N, Lou Z (2018) Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. Water 10: 1543

Ibebuchi CC (2022) Patterns of atmospheric circulation in Western Europe linked to heavy rainfall in Germany: preliminary analysis into the 2021 heavy rainfall episode. Theoretical and Applied Climatology: 1-15

Jain SK, Mani P, Jain SK, Prakash P, Singh VP, Tullos D, Kumar S, Agarwal S, Dimri A (2018) A Brief review of flood forecasting techniques and their applications. International Journal of River Basin Management 16: 329-344

Kai W, Deyi C, Zhaohui Y (2016) Flood Control and Management for the Transitional Huaihe River in China. Procedia Engineering 154: 703-709 DOI 10.1016/j.proeng.2016.07.572

Laverde-Barajas M (2022) Improving Satellite-Based Precipitation Estimates A Spatiotemporal Object-Oriented Approach to Error Analysis and Correction. Doctor Dissertation, TU Delft

Laverde-Barajas M, Corzo G, Bhattacharya B, Uijlenhoet R, Solomatine DP (2019) Spatiotemporal Analysis of Extreme Rainfall Events Using an Object-Based ApproachSpatiotemporal Analysis of Extreme Hydrological Events:95-112.

Laverde-Barajas M, Corzo GA, Poortinga A, Chishtie F, Meechaiya C, Jayasinghe S, Towashiraporn P, Markert A, Saah D, Son LH, Khem S, Boonya-Aroonnet S, Chaowiwat W, Uijlenhoet R, Solomatine DP (2020a) ST-CORAbico: A Spatiotemporal Object-Based Bias Correction Method for Storm Prediction Detected by Satellite. Remote Sensing 12 DOI 10.3390/rs12213538

Laverde-Barajas M, Corzo Perez GA, Chishtie F, Poortinga A, Uijlenhoet R, Solomatine DP (2020b) Decomposing satellite-based rainfall errors in flood estimation: Hydrological responses using a spatiotemporal object-based verification method. Journal of Hydrology 591 DOI 10.1016/j.jhydrol.2020.125554

Le X-H, Ho HV, Lee G, Jung S (2019) Application of long short-term memory (LSTM) neural network for flood forecasting. Water 11: 1387

Lescouflair A (2021) Model-based analysis of Nature-based Solutions impacts on spatio-temporal hydrological extremes Delft: IHE Institute for Water Educatioon; .

Li J, Hsu K, AghaKouchak A, Sorooshian S (2015) An object-based approach for verification of precipitation estimation. International Journal of Remote Sensing 36: 513-529

Liu C, Sun J, Yang X, Jin S, Fu S (2021) Evaluation of ECMWF Precipitation Predictions in China during 2015–18. Weather and Forecasting 36: 1043-1060

Mirjalili S (2019) Genetic algorithmEvolutionary algorithms and neural networks:43-55.

Nguyet LM (2021) Comparative analysis of Gated Recurrent Units, Long Shortterm Memory and Convolutional network architectures for flow forecasting in Vietnam. UNESCO-IHE

Oleyiblo JO, Li Z-j (2010) Application of HEC-HMS for flood forecasting in Misai and Wan'an catchments in China. Water Science and Engineering 3: 14-22

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12: 2825-2830

Pereira O, Torre E, Garcés Y, Rodríguez R (2014) Edge detection based on kernel density estimation. arXiv preprint arXiv:14111297

Ran Q, Fu W, Liu Y, Li T, Shi K, Sivakumar B (2018) Evaluation of quantitative precipitation predictions by ECMWF, CMA, and UKMO for flood forecasting: Application to two basins in China. Natural Hazards Review 19: 05018003

Shao Y (2018) Model-based analysis of the Dapoling-Wangjiaba catchment (Huai river basin) : scenarios of flood hazard change. UNESCO-IHE [in Onbepaald]

Skok G, Tribbia J, Rakovec J, Brown B (2009) Object-based analysis of satellite-derived precipitation systems over the low-and midlatitude Pacific Ocean. Monthly weather review 137: 3196-3218

Solomatine DP, Dulal KN (2003) Model trees as an alternative to neural networks in rainfall—runoff modelling. Hydrological Sciences Journal 48: 399-411

Solomatine DP, Xue Y (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. Journal of Hydrologic Engineering 9: 491-501

Swinbank R, Kyouda M, Buchanan P, Froude L, Hamill TM, Hewson TD, Keller JH, Matsueda M, Methven J, Pappenberger F (2016) The TIGGE project and its achievements. Bulletin of the American Meteorological Society 97: 49-67

Toth E, Montanari A, Brath A (1999) Real-time flood forecasting via combined use of conceptual and stochastic models. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 24: 793-798

What is an ensemble forecast? (2022) https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/mogreps. Cited 25/02 2022

What is remote sensing and what is it used for? (2022). Cited 28/02 2022

Wang X (2017) Exploring DYNIA and VARS methods for identifiability and sensitivity analysis of a hydrological model : case study of Huai River catchment. UNESCO-IHE [in Onbepaald]

1.47 billion people face flood risk worldwide: for over a third, it could be devastating (2020) https://blogs.worldbank.org/climatechange/147-billion-people-face-flood-risk-worldwide-over-third-it-could-be-devastating. Cited October 2021 2021

Xu Y (2014) Error and uncertainty prediction in hydrological modelling : case of Dapoling-Wangjiba catchment in Huai River basin Delft : UNESCO-IHE Institute for Water Education;.

Appendices

Appendix A. Research Ethics

United Nations Educational, Scientific and Cultural Organization

IHE DELFT Institute for Water Education under the auspices of UNESCO

Research Ethics Committee
IHE Delft Institute for Water Education
E ResearchEthicsCommitee@un-ihe.org

Date:	2022-03-02
To:	Andrés Julián Ruiz Gómez
MSc Programme:	HI
Approval Number:	IHE-RECO 2021-229

Subject: Research Ethics approval

Dear Andrés Julián Ruiz Gómez

Based on your application for Ethical Approval, the Research Ethics Committee (RECO), IHE Delft RECO gives ethical clearance for your research topic Analysis of spatiotemporal rainfall objects in hydrological ensemble forecast predictions

This approval valid until April 19, 2022. Please notify the RECO if your research protocol is modified in any way. If you do not complete your research by the specified date, contact RECO to request an extension for the ethical clearance.

Please keep this letter for your records and include a copy in the final version of your MSc. thesis, together with your personal ethics reflection.

On behalf of the Research Ethics Committee, I wish you success in the completion of your research.

Yours sincerely,

Dr. Angeles Mendoza Sammet
Coordinator, Research Ethics Committee IHE Delft

Copy to: Archive.

Appendix B. Personal Ethics Declaration

I, **Andrés Julián Ruiz Gomez**, the author behind the master thesis: "Analysis of spatiotemporal rainfall objects in hydrological ensemble forecast predictions: Application in the Dapoling-Wangjiaba catchment in the Huai River Basin", hereby declare that I have read and understood the principles and standards described in chapter two and three in the Netherlands Code of Conduct for Research Integrity. I declare then the following:

- I did not collect data from any individuals through questionnaires, interviews or anything similar.
- I was not required to get any ethics clearance from any kind of institution.
- I did not face any kind of conflict of interest.
- All the information used in this research, was authorized for its use or it was obtained from public databases.
- I did not receive any type of funding for the development of the research.

The principles applied in this research were:

1. Honesty
2. Transparency
3. Scrupulousness
4. Independence
5. Responsibility

These principles were applied in each stage of this work in the following way:

- **Design of research:** All ideas and previous work on which this research was based, were properly cited and mentioned. The data was acquired through public databases, or authorized by the owner of it. The methodology was clearly presented, in such a way, that anyone can follow it, and if desired, can replicate it.

  This research was developed always independently, with constant feedback from my supervisors and aiming, to provide new ideas for forecasting and objects-based methodologies. My sponsor did not have any influence in the chosen topic or in the way on how to better approach it. Data was acquired either through satellite products available to all public or by data used in previous MSc thesis, which the consent of the data owner.

  This research was drawn by the motivation of allowing, eventually, low-income countries to be better prepared for floods, based on the fact that it uses information publicly available and with almost global coverage.

- **Conduct research:** This work always followed the principle of honesty, meaning that when simplifications or assumptions needed to be made, they were informed and justified. Additionally, all steps of the research were conducted with the best of my capabilities. Based on my knowledge, experience and also the ones from my mentors, I chose the best methodologies and concepts to acquire the results. All the scripts, that allow me to get my results, will be stored and can be used by any third-party who wants to use them and validate them.

- **Report Results:** Even if the results were not as good as expected, they were informed and when it was possible, they were explained. When reporting the results, it was clearly done, following the described methodology, so these could be, if needed, replicated, judged, validated and that other people can make their own conclusions based on them. All authors read and approved this document, as well as agreed up to what extent they are valid. My supervisory team was always aware of my assumptions and simplifications. Additionally, every result presented was genuinely generated following the methodology, meaning that no results was elaborated, in such a way that could help me or benefit me.

- **Assessment and peer review:** The results obtained, followed a strict methodology supported by journals' information or other reliable sources of information. Additionally, all my expertise in the field was used for this research, as well as the one from my mentors: Dr. Gerald Corzo and Dr. Schalk Jan van Andel. All parties acted neutral and with honesty when giving me their comments and opinions. This document represents the work from constant meetings, presentations, emails, drafts' submission and independent work. I was not involved into "citation pushing", in order to benefit an external party. The ideas here presented and cited, were, up to my knowledge and expertise, the best possible.

- **Communication of research:** The author was completely honest, regarding the limitations of this document and hence, were explained and showed in the recommendations chapter. This research did not look to please any particular organism or person, but it was led by the willing to contribute to the flood forecasting field.

  In **Appendix A,** my Ethics' approval letter from Research Ethics Committee from IHE was attached as it was required on the same.