The Application of Bayesian Network Model

Quantifying Riverine Flood Hazard in the Java Island

Stephen Sanjaya

Additional Graduation Work - CE05050-09

TU Delft

The Application of Bayesian Network Model

Quantifying Riverine Flood Hazard in the Java Island

by

Stephen Sanjaya

as the completion of Additional Graduation Work
at the Delft University of Technology,

Student number:	4533313	
Project duration:	September 5, 2017 – January, 2018	
Assessment committee:	dr. ir. Oswaldo Moráles-Napoles,	TU Delft, supervisor
	dr. Markus Hracowitz,	TU Delft

TU Delft

Contents

1 Introduction

This chapter describes the main ideas of the study of riverine flood in Indonesia, precisely in Java Island.

1.1. Background

Flooding is a catastrophic natural disaster that leads to different type of damages. Such damages threaten many aspects of human lives, for instance environmental, societal and/or economic. Globally, absolute damage may increase by up to a factor of 20 by the end of century without any action taken [15]. Particularly, countries in Southeast Asia face a severe increase in flood risk [15]. This rises the need to understand the risk of flooding in order to alleviate the consequences of flooding, by quantifying the hazard.
By definition, flood risk is the combination of the probability of a flood event and of the potential adverse consequences to human health, the environment and economic activity associated with a flood event [12]. Flood risk management is divided into three big parts, namely risk analysis, risk assessment and risk reduction [14]. In the early stage, risk analysis involves the hazard assessment, which provides the information on the previous, current and future flood risks [14]. In a later stage, a better understanding toward flood hazard enables to an early prevention of flood prone areas, also to increase the preparedness in combating the impact of flooding. Accordingly, the role of flood hazard investigation is prominent to prevent more and more damages to occur.
As one of five big islands in Indonesia, Java is the busiest island in terms of economic activity, as well as

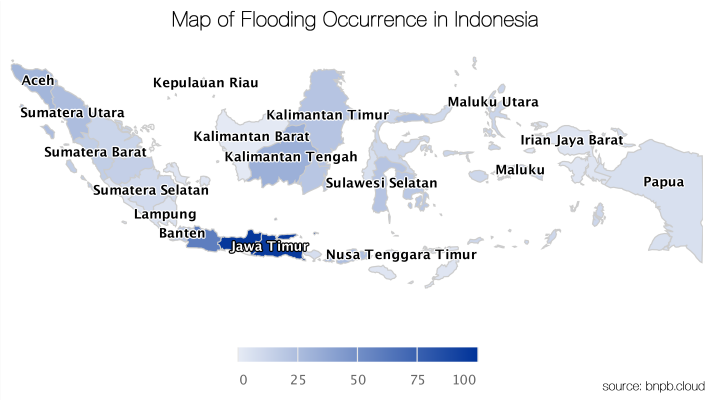


Figure 1.1: Map of flooding occurrence in Indonesia. It shows that Java Island experienced more flooding in comparison to any other islands. The scale shows number of occurence of flooding. (source:bnpb.cloud)

the most populous island in Indonesia. In contrast to that, flooding mainly happened in the Island of Java, as depicted in figure 1.1 compared to the other parts of the country. The scale is expressed in number of occurences, and East Java or in the graph *Jawa Timur* has more than 75 flooding cases. The expansion of population, inadequate spatial planning and land management, and urbanization are mentioned as the aggravated sources of flooding [1]. These are severely affected by the fact that climate change enhances the

uncertainty of the event. Currently, the flooding problem mainly was controlled by multipurpose reservoirs. There are sixteen reservoirs in the entire island of Java that have also a flood control function [2], apart from its main function as irrigation. Nevertheless, the application of flood hazard map in Indonesia, in general is not well explored and widely utilized. Certainly, a proper study needs to be carried out, to analyse the flood hazard potential in Java Island, especially riverine flood, as an early stage to have a better understanding regarding the flood risk.

1.2. Scope of the Research and Objectives

A stochastic model, like Bayesian Network, permits the flexibility to evaluate different datasets in order to estimate annual maximum daily discharge. This study is carried out to evaluate the performance of the Bayesian Network Model developed by Paprotny et al. (2017) in Europe, by implementing the BN configuration in the case of Java Island, Indonesia. By doing so, it is possible to take a measure of the adaptability of the model within a different environment as an input. Throughout this research, it is expected that it could be beneficial to give a preliminary idea regarding flood hazard potential in Java Island, particularly for ungauged rivers. The output of this research is a map of riverine flood hazard for the entire Island, which indicate extreme riverine discharge within the catchment.

1.3. Methodology

In general, the provision of spatial datasets are mainly compiled from open sources.The spatial datasets will be extracted using GIS as input parameters to the model, as depicted in figure 1.2. The detail of input parameters will be explained later in the section of spatial datasets.
The input parameters of will be bounded spatially to Java Island and temporally based on the availability of

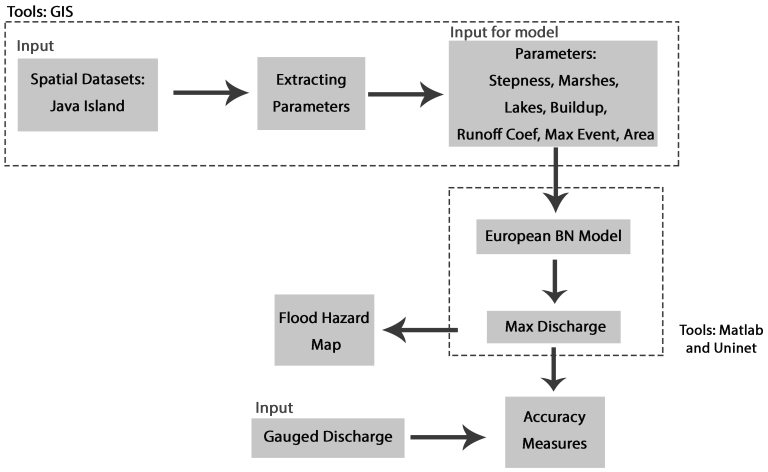


Figure 1.2: Flowchart diagram for the research methodology

the datasets. Thereafter, the parameters will be applied in the Bayesian Network Model to generate the maximum discharge of the gauged river in Java Island. Thus, there will be two discharge values, one estimated from the model and another one obtained from gauged from the site. In that sense, the validation through difference in output could be done using different measures such as the coefficient of determination, R^2 and the Nash-Sutcliffe efficiency coefficient, (NSE). Such quantification allows evaluating the performance of the BN in the Java Island. Finally, the whole process will be carried out again for the ungauged river, to generate the riverine flood hazard map for the entire Java Island.

1.4. Research Outline

In general, this report is divided into three parts. Chapter 2 presents the overview of the literature related to the case study and theoretical background that has been used in the research. Chapter 3 describes the application of the BN model in the particular case study and result and discussion of the model. Lastly, Chapter 4 presents the conclusion and recommendation for future research.

2

Overview of the Literature

2.1. System Description

Java island is one out of five major islands in Indonesia. It is situated on the equator precisely between E 105° and E 114°, as depicted in figure 2.1. The size of the island is approximately 132,107 km^2 [2]. Java Island is home to almost 140 million inhabitants, and considered as the most densely populated island in Indonesia [5]. It means that 50% of the population of the country is mainly concentrated only in Java Island.

In general, Indonesia experiences two distinctive seasons, wet season from December to March and a dry

Figure 2.1: Maps of Indonesia (source: www.freeworldmaps.net)

season from June to September, with temperature ranges between 21° C to 33°C and humidity between 75% and 100% [2]. These characteristics relatively reflect the general weather in Java Island. In addition to that, based on Köppen-Geiger Climate Classification, it is also categorised as AF which defines as equatorial rainforest and fully humid climate [8]. These bases generate a reasonably large precipitation flux all over the year. The average annual rainfall in Java Island is circa 2,680 millimetres/year [2]. The topography of the island in itself is very diverse, laying from east to west within the island's spine and is flanked by limestone ridges and lowlands [5]. Besides that, it is also surrounded by volcanoes, and some of them are still active, which result the soils are very fertile by the volcanic ash [5]. Due to abundant of rainfall events, most of the land use are occupied as a crop land. This in return, led to many construction of hydraulic structures, especially reservoirs. There are 87 reservoirs with irrigation as the main the function; moreover, sixteen of it built for multipurpose, including flood control [2].

2.2. Bayesian Network

Bayesian Networks (BN), as a probabilistic graphical model, are directed acyclic graph (DAG) that consists of nodes and arcs. The nodes resemble random variables that can uptake on two or more possible values in

discrete or continuous, whilst the arcs represent the existence of direct casual influences between the linked variables (nodes). The significances of the influence are shown by the quantification of conditional probabilities. The strength of these direct influences are quantified by assigning to each variable X_i a link matrix $P(x_i|\prod_{X_i})$, which represents judgemental estimates of the conditional probabilities of the events $X_i = x_i$, given any value combination $\prod_{X_i}$ of the parent set $\prod_{X_i}$.

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|\prod_{X_i}) \tag{2.1}$$

As a brief explanation of DAG and parent cells, figure 2.2 demonstrates several simple BNs. In general, these graphs present the acyclic relation, which imply the joint probability function between variables. In the hierarchical perspective, the immediate predecessors are the parents of the immediate successor nodes, whereas the successor nodes are the children. Referring to the same figure, the second graph shows that node F as the child, has two parent variables, namely node D and E in order of importance. Therefore, it is able to be concluded that node D has a direct dependence to node F. Meanwhile, the correlation between node E and F will be known, given we know the correlation between node D and F. This condition is so-called the conditional correlation. Another interesting information that could be inherited is from the third graph in

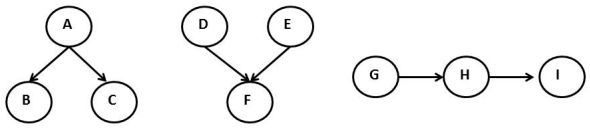


Figure 2.2: Example of directed acyclic graph of Bayesian Network. Courtesy of Couasnon [4]

figure 2.2. The information of Variable I will be obtained from G given one knows the H. It means that having more information about G will not give you more information about I without having H. In other word, G is conditionally independent of I. Later in the application, these type of graphs will be constructed in more robust networks. The advantages of network representation is that it allows people to express directly the fundamental qualitative of "direct dependency". The network then displays a consistent set of additional direct and indirect dependencies and preserves it as a stable part of the model, independent of the numerical estimates. Bayesian Networks have been applied in many projects in different fields [7], e.g. risk analysis, reliability analysis, prediction model, etc. BN has shown a compelling model in statistical analysis, in combination with the conceptual studies, to understand the risk.

2.2.1. Non-Parametric Bayesian Network

In Bayesian Network, Non-Parametric Bayesian Network (NPBN) is a part of continuous BN which was develop under the assumption without defining the marginal distribution [7], or in other word there is no joint distribution is assumed. Initially, the BN arcs are expressed using joint normal distribution for continuous variables. However, in many of the cases joint normal distribution fails to represent the arcs especially when normality does not hold. Therefore, Hanea et al. [7] proposed to use Copula to avoid this limitation.

COPULA– Hydrological modelling requires multivariate analysis which is really complex. In a smaller scale, it is common to use bivariate analysis to test the dependency between two random variables. Generally, this analysis is easier to perform if two variables follow the same distribution. However, this is not always the case. This problem could be solved using Copula, that avoid this restriction. The copula approach to dependence modelling is rooted in a representation theorem due to Sklar in 1959 [6]. Copula normalized the variables into a unit square distribution. Assuming under bivariate analysis of continuous random variables, cumulative distribution function (c.d.f) of this copula is formulated in the following equation

$$H(x, y) = C\{F(x), G(y)\}, \quad x, y \in \mathbb{R} \tag{2.2}$$

where $H(x, y)$ is the c.d.f.; $F(x)$ and $G(y)$ are the marginal distribution of two random variables; $C[0, 1]$ represents the copula.
Based on above definition, random variables are transformed into uniform square unit [0,1]. By doing so, the dependence of two variables in the model could be obtained using Pearson Correlation Coefficient, regardless its marginal distribution. Till date, the most renowned copula is Gaussian or Standard Normal Copula,

due to its rapid calculation in large multivariate model. The Bivariate Gaussian Copula is presented in the following equation:

$$C_\rho(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right) \tag{2.3}$$

where Φ is the standard normal distribution, Φ^{-1} is its inverse and Φ_ρ is the bivariate Gaussian cumulative distribution with (conditional) product moment correlation ρ between two marginal uniform variates u and v in the interval [0,1].

CORRELATION COEFFICIENT – As discussed regarding the relationship between nodes under conditional rank correlation, thus the concept of conditional rank correlation will be demonstrated in this section. The conditional rank correlation of x_i and x_j given random vector $Z = z$ is the rank correlation calculated in the conditional distribution $(x_i, x_j | Z = z)$. The arc J between variable x_i and its n parent $P_{a_1}(x_1), \ldots.., P_{a_n}(x_i), Pa_j(x_j) \rightarrow x_i$ is parametrized as follow:

$$\begin{cases} r(X_i, P_{a_j}(x_i)) & j = 1 \\ r(X_i, P_{a_j}(x_i) | P_{a_1}(x_j), \ldots, P_{a_{j-1}}(x_i)) & j = 2, \ldots, n \end{cases} \tag{2.4}$$

where j is the index in the non-unique sampling order. The empirical normal rank correlation $r(Z_i, Z_j)$ approaches the empirical rank correlation $r(x_i, x_j)$ if the normal copula assumption holds and given a large enough number of observations:

$$r\left(x_i, x_j\right) = \rho\left(F_{x_i}\left(x_i\right), F_{x_j}\left(x_j\right)\right) \tag{2.5}$$

The benefit of using joint normal distribution is that the conditional and partial correlation are equal [7]. Therefore the matrix correlation could be computed from the partial correlation. Partial correlation defines the correlation between two variables given certain circumstances to be hold, as this is relevant to the conditional rank correlation. Partial correlation is formulated in the following equation:

$$\rho_{12;3\ldots n} = \frac{\rho_{12;3\ldots n-1} - \rho_{1n;3\ldots n-1} \cdot \rho_{2n;3\ldots n-1}}{\sqrt{1-\rho^2_{1n;3\ldots n-1}} \cdot \sqrt{1-\rho^2_{2n;3\ldots n-1}}} \tag{2.6}$$

Under bivariate normal copula assumption, a simple characteristic relation between the Spearman's rank correlation r and the Pearson's product moment correlation coefficient ρ exist:

$$r = \frac{6}{\pi} \cdot \arcsin\left(\frac{\rho}{2}\right) \tag{2.7}$$

2.2.2. Inference

As defined previously in section 2.2.1 regarding the definition of non parametric model, thence, there is no analytical form of joint distributions involved in NPBN [7]. Performing inference in NPBNs could be handled by sampling it. In the case that BBN becomes more complicated with more DAGs involve, the numerical evaluation will be more time consuming. This condition may lead to a big drawback when it comes to a real-time decision making. However, this problem could be solved if one use the normal copula. Using normal copula allows the user to operate the joint normal distribution with the same rank correlation structure as the original one [7]. Therefore, the properties of the joint normal distribution could be used to transform the conditional rank correlation into the conditional product moment correlation. As mentioned earlier in section 2.2.1, the conditional and partial correlation are equal. Furthermore, one could sample from these normal joint distribution given the correlation matrix, then transform it back to the marginal distribution. By performing this method, a sample from the joint distribution of the original variables and the dependence structure realized by the normal copula become attainable. Another important remark is regarding the number of sample, in which related to the sensitivity of the result. The bigger the number of sample being generated, of course, will lead to higher confidence of mean of estimation . Nevertheless, it has to be compensated with the performance of the computer. For example, in the case of European BBN, 1000 samples were used each time of conditionalisation.

2.3. Method for Result Validation

As explained in previous section, the validation of the model will be quantified by estimating the true error value, as it is also aimed to set a threshold for indicating the realization of the model. There are many performance measures to validate the result of the model, however, in this research, two methods will be used,

namely coefficient of determination (R^2) and Nash-Sutcliffe Efficiency (NSE). The coefficient of determination represents the linear relation between the estimated and the observed data. It describes the proportion of the variance measured data explained by the model [9]. R^2 has 0 to 1 value range, with higher value indicates the less error variance. Nash-Sutcliffe Efficiency defines the relative values of the residual variance compared to the measured data variance [9]. NSE ranges between -∞ to 1.0, a value between 0 and 1.0 represents an acceptable performance, as 1.0 is the optimal value, meanwhile less than 0 signifies a poor performance [9]. The both error measurements are formulated in the following equations:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Q}_i - \bar{Q})^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2} \tag{2.8a}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2} \tag{2.8b}$$

where $\hat{Q}_i$ is the estimated value of the discharge from the model, Q_i is the measured value of the discharge, and $\bar{Q}$ is the mean value of the observed value.

2.4. Development of BN Model in Europe and USA

A Bayesian Network Model for quantifying riverine flood was first constructed based on the case of Europe; it has shown a satisfactory performance. This model later was also applied in the United States, where the result was not as well as the predecessor case. A brief description of both research area will be elaborated in the following section.

2.4.1. Extreme river discharge in Europe

BN model in Europe is used to statistically estimate the discharge, using all parameters, on a continental scale. The study of BN model in Europe would be the groundwork of the model for quantifying the riverine flood in this case. At the initial stage, many hydrological and climatic variables were involved in the extreme river discharge. However, using the sensitivity analysis, it has pinned down into seven nodes, such as precipitation and snowmelt flux, surface runoff, steepness, three different types of land use, and the size of the catchment; the model is shown in the figure 2.3. Those variables are the most influential variables in representing the estimation of the extreme discharge. It has also given an intuitive sequence and order of importance to the estimation of maximum discharge – as the benefit of using the Bayesian network. Due to its wellness most of the time, the Gaussian copula is also the ground basis of the conditionalisation of the river discharge. In addition to that, if the model performs accurately, it is possible to estimate maximum river discharge in the ungauged basin. The BN model that applied in Europe yielded an adequate result in estimating the annual

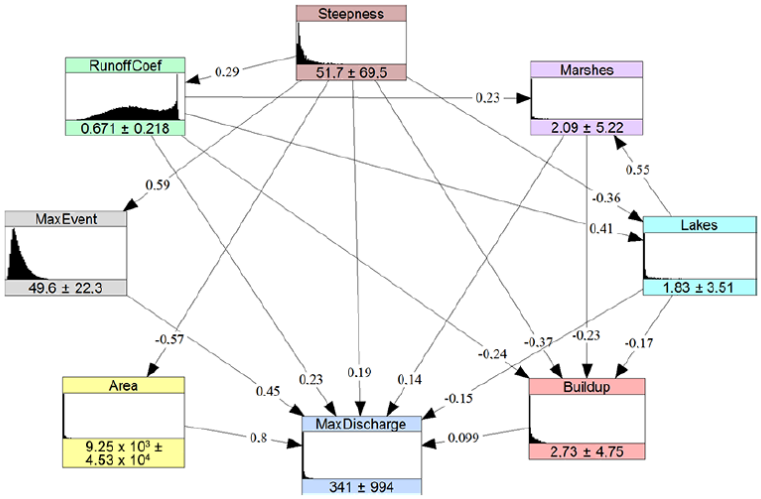


Figure 2.3: Bayesian Network Model for European Region (source: Paprotny and Morales-Nápoles [11])

maximum river discharge. In the table 2.1, the validation shows a very satisfactory measure, when compared

to the observation of the river discharge. Additionally, It is also emphasized that regions like Scandinavia, western Europe and the Danube basin, delivered a better result compared to southern Europe, especially in the Iberian Peninsula. However, since this model is designated for large catchment, it shows less precision

Table 2.1: Validation measure of Europe (source: Paprotny and Morales-Nápoles [11])

	Q_{MAXM}	Q per catchment area
R^2	0.92	0.52
NSE	0.92	0.41

of performance in the smaller catchment. In the following, table 2.2, the performance measures are getting lower proportionally to the size of the catchment. The study of BN in Europe has also proved reducing time

Table 2.2: Validation measure of Europe per catchment area (source: Paprotny and Morales-Nápoles [11])

Area	R^2	NSE
$< 100 km^2$	0.47	0.41
$100 - 1000 km^2$	0.55	0.38
$1000–10,000 km^2$	0.64	0.43
$> 10,000 km^2$	0.90	0.89

to perform a flood-hazard analysis, both continental-scale and local, as long as annual extremes are relevant for a particular study [11]. It is also possible to configure by adding another node that might be relevant to a particular case, in which has been done in the case of United States. This model shows the flexibility to integrate another aspect such as climate change or other land use variables.
Apart from its benefits, there are also several drawbacks of the model itself. Firstly, this model makes use a natural flow in the catchment, in which not always the case in the presence of the hydraulic structures. The BN model includes reservoirs indirectly as the contribution of the lakes and have a negative influence of the extreme discharge [11]. Nevertheless, the impact towards the result of the model is not demonstrated clearly in the study. Secondly, the temporal variability of the spatial datasets is in a daily time step. In consequences, this model is not applicable to such cases like flash flood or a flood of short duration. Thirdly, even though the availability of the data, in general, is sufficient, in contrary such data still contains slight inaccuracies or errors. This fact leads to an error generation of the final result, that might not be seen explicitly. Furthermore, in some developing countries, like Indonesia, the rare availability with poor quality data might have an unfortunate result.

2.4.2. Application of a Bayesian Network in the USA

After a satisfactory result of a Bayesian Network in Europe region in estimating river discharge, this model was also applied in the contiguous U.S. by Couasnon [4] in 2015. In principle, the structure of the model from Europe is the groundwork for the implementation in the U.S., which has a slightly decreased performance. However again, it is proved to be efficient regarding time process, and favourable method to obtain the annual maximum daily discharge. The result of the application of BN model in the U.S. is presented in the table 2.3 and table 2.4.

In the case of the U.S., several innovations, for instance, adding relevant nodes and saturating the struc-

Table 2.3: Validation measure of the United States (source: Couasnon [4])

	Q_{MAXM}	Q per catchment area
R^2	0.858	0.139
NSE	0.757	0.108

ture, were conducted to enhance the performance of the model. Nevertheless, the result did not showed a significant improvement. Later in the discussion of the research, several conclusion on the result were mentioned, as a result of unique characteristic of the U.S. topography. Concordant reasoning also took place on the available database for the U.S. for small drainage area. The delineation of a small catchment area is still inadequate, hence give an unfortunate result. As mentioned beforehand, the BN structure was built explicitly for riverine flood, which leads to the inability to predict other types of flooding. Precisely, more extensive

Table 2.4: Validation measure of the United States per catchment area (source: Couasnon [4])

Area	R^2	NSE
$< 100 km^2$	0.258	0.247
$100 - 1000 km^2$	0.319	0.262
1000–$10,000 km^2$	0.345	0.331
$> 10,000 km^2$	0.906	0.747

climatic and geographical diversity which favours other flooding mechanisms is not accounted for with the current structure [4]. Thus later, the model also considered the effect of seasonality in the run-off coefficient and did improve the model performance. The same conclusion also appears in the human influence of the model in getting the estimation, due to local control such as hydraulic structure. Furthermore, another human activity such as agricultural practices could also contribute to the error that occurred. Another remark is related to the probabilistic distribution for the frequency analysis. It is found that some stations demonstrated an unrealistic estimate of the high return period, due to its sensitivity towards the selection of the parametric value.

3

Quantifying Bayesian Network Model

This chapter describes spatial datasets that used in the model and the result and discussion of the model.

3.1. Spatial Data Sets

As mentioned in section 2.4.1, seven nodes were taken into account in quantifying the maximum discharge. In general, spatial datasets are extracted for two purposes, for validation measures and flood hazard map. The main difference is that for the former, drainage basins are delineated based on the location of the observed discharge; the latter, drainage basins are taken from HydroSHEDS database. Nevertheless, all procedures of extraction are identical, using Matlab and QGIS. The summary of the spatial datasets is tabulated in the following table.

Table 3.1: Table of Data Source

Data Type	Units	Nodes	Data Source	Year	Resolution
Maximum Daily Precipitation Flux	$kg/m^2/s$	MaxEvent	CORDEX EAS-44	1951-2005	1584 arc-second
Surface Runoff	$kg/m^2/s$	Runoff Coefficient	CORDEX EAS-44	1951-2005	1584 arc-second
Maximum Daily Discharge	m^3/s	Maximum Discharge	GRDC	varies between 1970 and 2010	N/A
Elevation	m	Steepness	SRTM (void-fill DEM)	2000	3 arc-second
Buildup	km^2	Buildup	0.5 km MODIS- based Global Land Cover Climatology (USGS)	2001-2010	15 arc-second
			0.5 km MODIS- based Global Land Cover Climatology (USGS)	2001-2010	15 arc-second
Marshes and Lake	km^2	Marshes and Lake	Globcover2009	2009	10 arc-second
			Global Lakes and Wetland Database	2004	30 arc-second

3.1.1. Precipitation Flux and Surface Runoff

Precipitation flux and surface runoff are both taken from CORDEX under the period of 1951 until 2005. There are several options for the products of these variables. However, the selection pinned down to model r3i1p1 HIRHAM5 from DMI (Danish Meteorological Institute), that has a more logical order of magnitude of precipitation flux compared to the updated product. HIRHAM5 is a regional atmospheric climate model that developed by the combination of HIRLAM and ECHAM [3]. The driving of the Regional Climate Model is based on ICHEC-EC-EARTH. The raw data is presented in three-dimensional matrices, in which the first and second dimension resembles the location in coordinate, longitude and latitude; whereas the third dimension is the timescale. The timescale of the raw data is on a daily basis, hence need to be clipped into annual maximum daily data. Conversion of the raw data from kg/m^2 is also conducted to obtain the real value of maximum yearly daily data [mm/day] for both variables. Since the domain of Indonesia lies in the East Asia Region, therefore the translated data will also be bounded spatially to the drainage basin. Throughout this process, the final data obtained with the temporal distribution of 50 years on the Java Island. These procedures are schematized in figure 3.1.

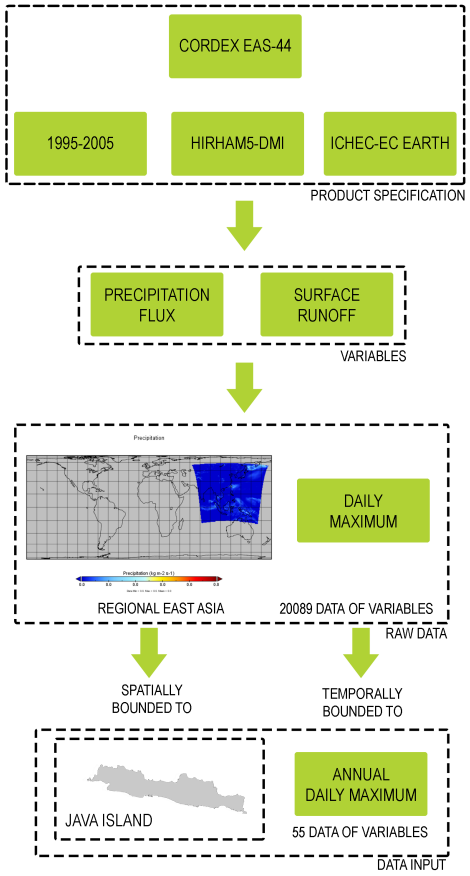


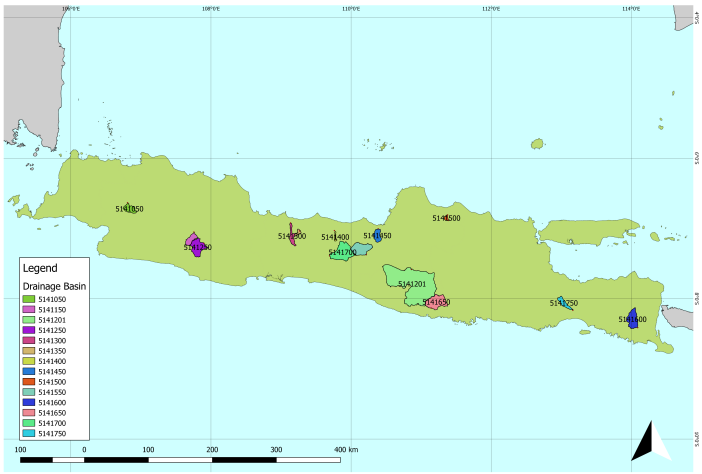
Figure 3.1: Flow diagram process of precipitation flux and surface runoff

3.1.2. Drainage Basin

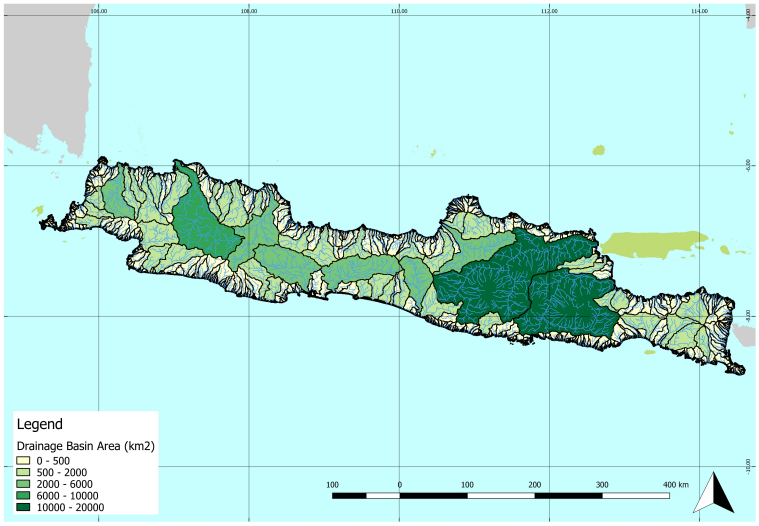
A drainage basin is the most crucial part in determining the boundary condition of the extraction of spatial datasets. Such data will give the information about the area and the elevation of the watersheds. On the part of validating the datasets in Java Island, the drainage basins are delineated based on the availability of the daily discharge data from GRDC. The delineation of these catchments is rendered using the flow accumulation raster file from HydroSHEDS with 3 arc-second resolution, with the measured location of the stream flow as the pour point of the watershed provided by GRDC. Moreover, in creating the flood hazard map, the drainage basins were taken from HydroSHEDS basin product with 15 arc-second resolution. These basins are also clipped with the existing river network, with resulting only 384 unit-catchments in estimating 100 years riverine flood discharge. However, the disadvantage of the product is that the basin is merely delimited in the primary river network, especially in the Java Island. It will affect a coarser estimation of the riverine flood discharge, and therefore the flood hazard map will be limited to at most the stream-flow per catchment area. In that sense, the result is not comparable to the previous study in Europe and the Contiguous of USA.

3.1.3. Land Use Data

The Europen BBN model makes use of three different types of land use; those are Build Up, Marshes and Lakes. The land use types are selected based on the sensitivity analysis that was conducted in the study of flood discharge in Europe by Paprotny and Morales-Nápoles [11] and have the higher influence to generate the maximum riverine flood discharge. Several land cover products will be utilised and combined to acquire for each variable, that is also validated visually with the real map data. Build up is an urbanised area and extracted from 0.5km MODIS-based Global Land Cover Climatology (USGS) with 15 arc-second resolution. This raster data is a combination of several land cover products in between the year 2001 and 2010. In contrast with precipitation flux and surface run-off, build up remains as a constant value in the BN model, also for other land use data. However, land cover does change in a certain period of time. Including the tempo-



(a) Delineation of unit catchment



(b) Drainage catchment of hydroSHEDS

Figure 3.2: Drainage basin on the top was delineated for validation purposes; On the bottom, the drainage basin was taken from HydroSHEDS

ral variation of the land cover might be a future reference for the development of the model. Based on the precursory study of Couasnon [4] for the case study of US, the lake and marshes land cover was solely taken from GLWD (Global Lake and Wetlands Database). On the other hand, marshes and lake data in Java Island are composed of the mixture of several products, due to the limitation of the outcome of GLWD. Hence the blend compilation between three different data sources is adapted to obtain a better land cover data. Despite the fact that 0.5 km MODIS-based GLCC does not show a clear distinction between marshes and lake in the layer of wetland permanent, nonetheless, it does cover most of the water bodies inside the island. Indeed, a validation to all compiled data is done to ensure the data by comparing the standard general maps from QGIS.

3.1.4. Discharge Data

Discharge data were taken from GRDC (Global River Discharge Center). Fourteen recording stations measured maximum daily discharge that ranges between 1970 and 2010 spread throughout the entire island, as depicted in figure 3.3. The raw data of discharge is a measure of maximum value on daily time scale per station. The annual maximum discharge could be obtained by taking the maximum amount of each year per

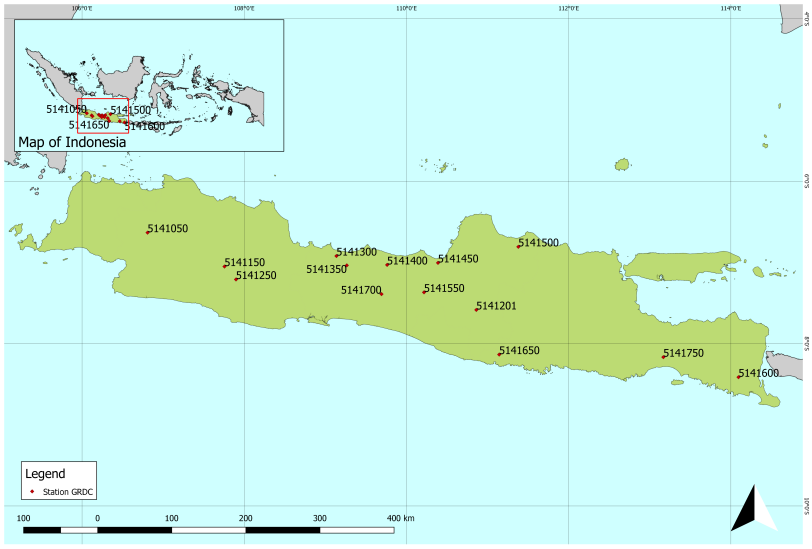


Figure 3.3: Station River Discharge Map of GRDC

station. It is critical to take into account that the missing record of the discharge is discarded during this process. The time series of data is available in the appendix A. This information used at a later stage to assess the performance of the BN. Consequently, the more time series available for the island, and widely spread out on the entire island, it should give a better representation for the validation of the model. In this case, there are 229 observations of discharge that are used to validate the performance of the model.

3.2. Result and Discussion

3.2.1. Inference Model

To start with the validation stage, one should ensure that the time scale of spatial datasets is as precise as the available annual maximum discharge data. In doing so, all spatial datasets should be imposed to the time of the annual maximum discharge data, especially the data that varies in time, such as rainfall flux, surface run-off. The remaining data will be merely altered based on spatial differences. In this case, it left the spatial datasets in a total of 229 data configurations to evaluate. Therefore, there will also be 229 estimated annual maximum river discharge to compare with the GRDC data. Once the time and spatial frame of each variable are identical, it is possible to proceed to the conditionalisation part. Table 3.2 shows the equation and unit of the variables, which hold in the assumption of the model. As stated in section 1.3 that the objective of the

Table 3.2: Equations table of spatial datasets

BN Node	Equation	Unit
Area	(-)	km^2
MaxEvent	max(Rainfall flux)	mm/day
RunoffCoef	max(Total Surface Runoff) / MaxEvent	(-)
Marsh	$\text{Area}_{\text{Marsh}}$/Area x 100	%
Lake	$\text{Area}_{\text{Lake}}$/Area x 100	%
Buildup	$\text{Area}_{\text{Buildup}}$/Area x 100	%
Max Discharge	(-)	m^3/s
Steepness	(Elev.Max - Elev.Min)/ sqrt(Area)	m/km

study is to evaluate the performance of the BN model that has been built previously, validation part would be a matter of conditionalisation, once the spatial datasets are ready. On the other hand, 100,000 samples were generated to ensure the consistency of the mean sampling value. In this way, the performance indicator will be more consistent every time the sampling is performed. Table 3.3 shows the summary of each node of the station based on the equation table before. It is noticeable that the size of the area is below 1,000 km^2.

The variation of the land use type is fairly diverse, with some of the stations were located in the urbanised area, viz. Station 5141300 and 5141450. Whereas another station, like Station 5141250 has thirteen percent of marshes within the catchment.

Table 3.3: Table of station

Station Code	Area $[m^2]$	Steepness $[m/km]$	Buildup [%]	Marshes [%]	Lakes [%]	# Data
5141050	183.92	148.95	0.81	9.22	0	25
5141150	183.56	138.02	0.76	2.59	0	19
5141201	3182.05	53.36	6.68	0.42	2.21	15
5141250	438.17	89.34	4.12	13.43	0	28
5141300	158.41	264.58	16.14	3.88	0	14
5141350	52.34	85.70	3.22	2.03	0	30
5141400	34.65	290.51	0	0	0	30
5141450	187.70	140.87	16.74	0.82	0.10	16
5141500	43.34	39.49	0	0	0	17
5141550	468.87	130.23	4.15	0.37	0	10
5141600	388.68	155.21	0.38	2.05	0	27
5141650	544.79	49.70	0	0.16	0	26
5141700	698.92	111.96	2.87	10.58	0.08	21
5141750	157.22	202.57	0.56	7.01	0	21

3.2.2. Result

Inference Result – Figure 3.4 shows the result of the annual maximum discharge of estimated plotted against the observed annual maximum discharge for all stations. From the normal plot, it could be concluded that the estimated value is overestimated. In general, the performance indicator shows an inferior result. It has 0.25 of the coefficient of determination, while -5.32 for NSE value. It results in a very poor Nash-Sutcliffe Value as it involves the mean value of the actual data. Based on the R^2 analysis, it has shown a low correlation due to the slope of its trendline. The slope resembles the overvaluation of the x-axis, which is the estimation of annual maximum discharge.

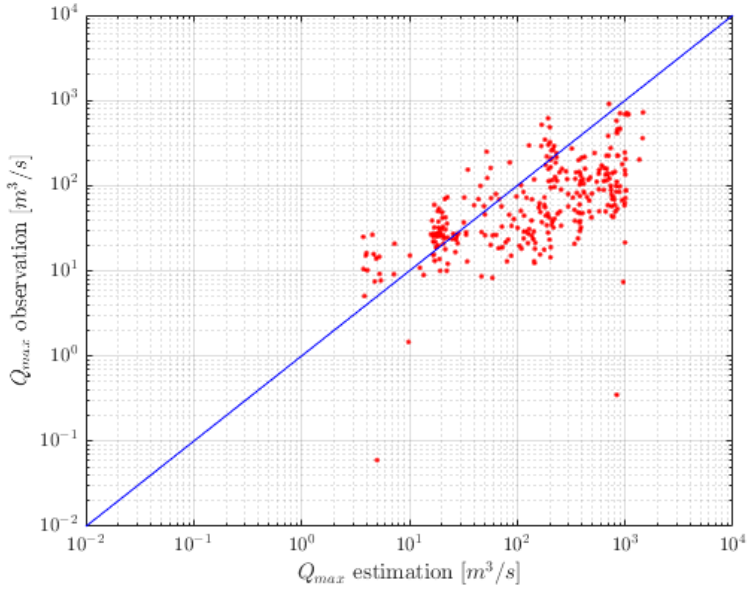
Comparison with the Previous Study – The study of BNs in Europe and USA has shown a satisfactory result. Hence it is essential to compare the result with the predecessor study. In table 3.4, it shows the performance measures for a different case study. The study in Java Island has shown the lowest measured compared to Europe and USA. All validation is done using the same European BBN model, which was built from the Europe datasets. However, that number of data that were used to validate the result is way less compared to Europe and the US.

Table 3.4: Compared validation measures

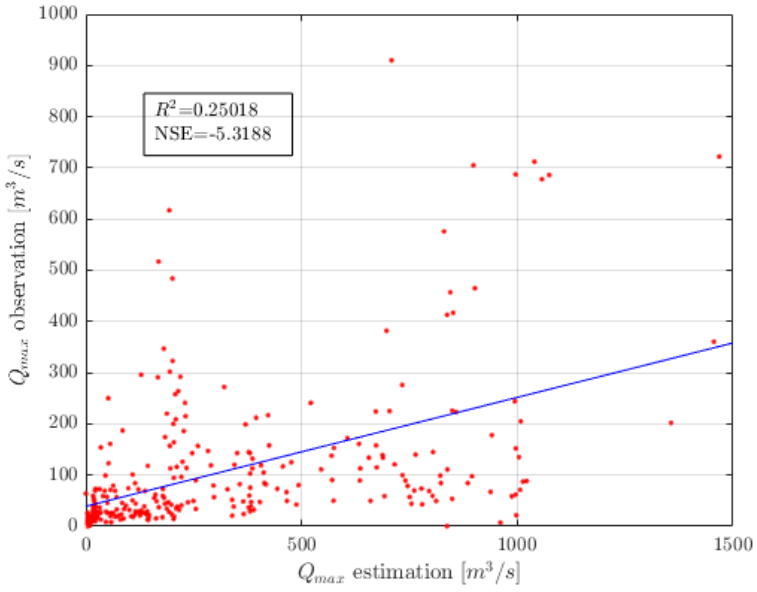
Measures	Europe	U.S.A	Java Island
R^2	0.92	0.858	0.25
NSE	0.92	0.757	-5.32

This comparison does not seem relevant due to the different size of the catchment. Therefore we need to indicate the measures of Europe and US based on the size of Java Island. The appropriate size to compare with the case of Java Island is the class area between 100 km^2 and 1,000 km^2, which is presented in the table 3.5. As the literature study conducted in the previous chapter, it is agreed that the smaller the catchment is, the less accuracy the model results will be. This hypothesis is reflected in the result in the following table. This follows with the fact that the model was initiated under the large size of catchment assumption.

Return Period of Each Station – This section will discuss the adequacy of the model result for each station by comparing the return period, as this is commonly used in measuring the result of the estimation. The



(a) Logarithm scale



(b) Normal scale

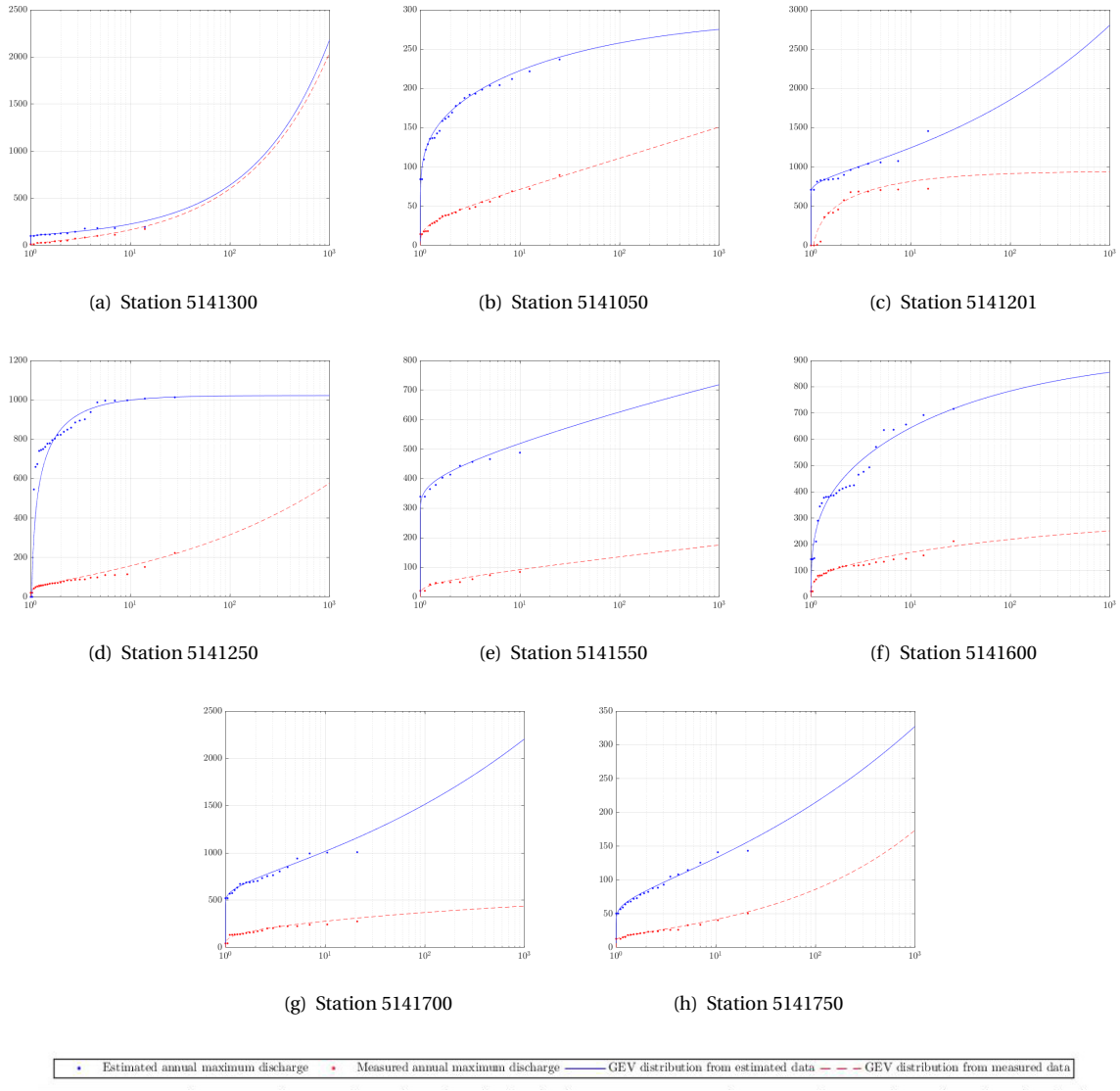
Figure 3.4: Plot of estimated and observed annual maximum discharge with x-axis represents the estimated Q_{max} and y-axis represent the observation Q_{max}

Table 3.5: Compared validation measures for related catchment size area

Measures	Europe $100-1000km^2$	U.S.A $100-1000km^2$	Java Island
R^2	0.55	0.319	0.25
NSE	0.38	0.262	-5.32

ground basis of the frequency analysis is using the generalised extreme value (GEV) distribution. GEV is the condensed parametric form for three limiting distribution for maxima, such as Gumbel, Weibull and Ferchet

[10]. In general, GEV distribution is preferable due to the importance of uncertainty, if the case is a shortfall of data, compared to Gumbel distribution [10]. Based on table 3.3, the number of data for each station is below 30. Figure 3.5 and 3.6 show the plot of maximum discharge on different return periods between the estimated and measured data for fourteen different stations.



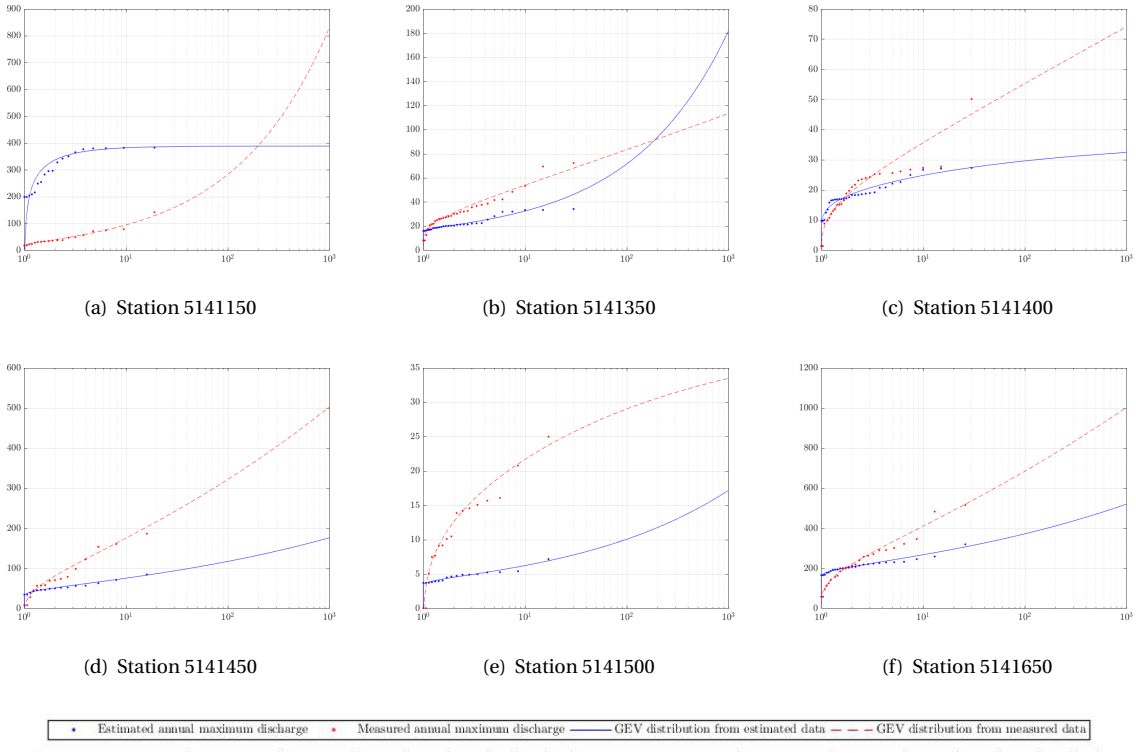
(a) Station 5141300 (b) Station 5141050 (c) Station 5141201

(d) Station 5141250 (e) Station 5141550 (f) Station 5141600

(g) Station 5141700 (h) Station 5141750

Figure 3.5: Plot of Discharge $[m^3/s]$ to Return Period in years for all stations with a better result

Figure 3.5 shows the station with a better result, despite the fact that it still depicts the overestimation of the annual maximum discharge, except for Station 5141300. Station 5141300 presents the best agreement of all, the result of the model has almost the same distribution compared to the observation value. Thus, a further optimisation towards the model is worth to be analysed to improve the result. On the other hand, figure 3.6 shows the station with low performance of the model. It could be concluded that almost all of the fit-distributions do not behave consistently between the model and the actual data. In contrast with the previous figure, figure 3.6 demonstrates the underestimation of the model; the red line is depicted above the estimated value.



(a) Station 5141150 (b) Station 5141350 (c) Station 5141400

(d) Station 5141450 (e) Station 5141500 (f) Station 5141650

Figure 3.6: Plot of Discharge $[m^3/s]$ to Return Period in years for all stations with a poor result

3.2.3. Discussion

This section present the discussion of the result form the model using the help from literature study and also to find an alternative improvement to gain a better result.

Application to the Ungauged Basin – Understanding the characteristic of the catchment through model would give a good representation of hydrological processes in the basin. Once the model is accurate enough to represent the entire area, it is possible to apply it to the ungauged basin, in which related to the purpose of modelling. It is also applicable in this case, after validating the result in the gauged basin, we could produce the flood hazard map in the entire area. In figure 3.7, it shows the annual discharge map for 100 year return period per catchment area in the Java Island. It exhibits that most of the basins in Java Island have 0.10 to 0.30 m^3/s per m^2, in which classified as fairly critical area. Due to the limitation of the delineation of the catchment area, as has been stated in the section of the drainage basin, this is the minimum result that could be generated from the data. One possible solution to improve the result is the delineation of secondary river network, to obtain higher resolution of the result.

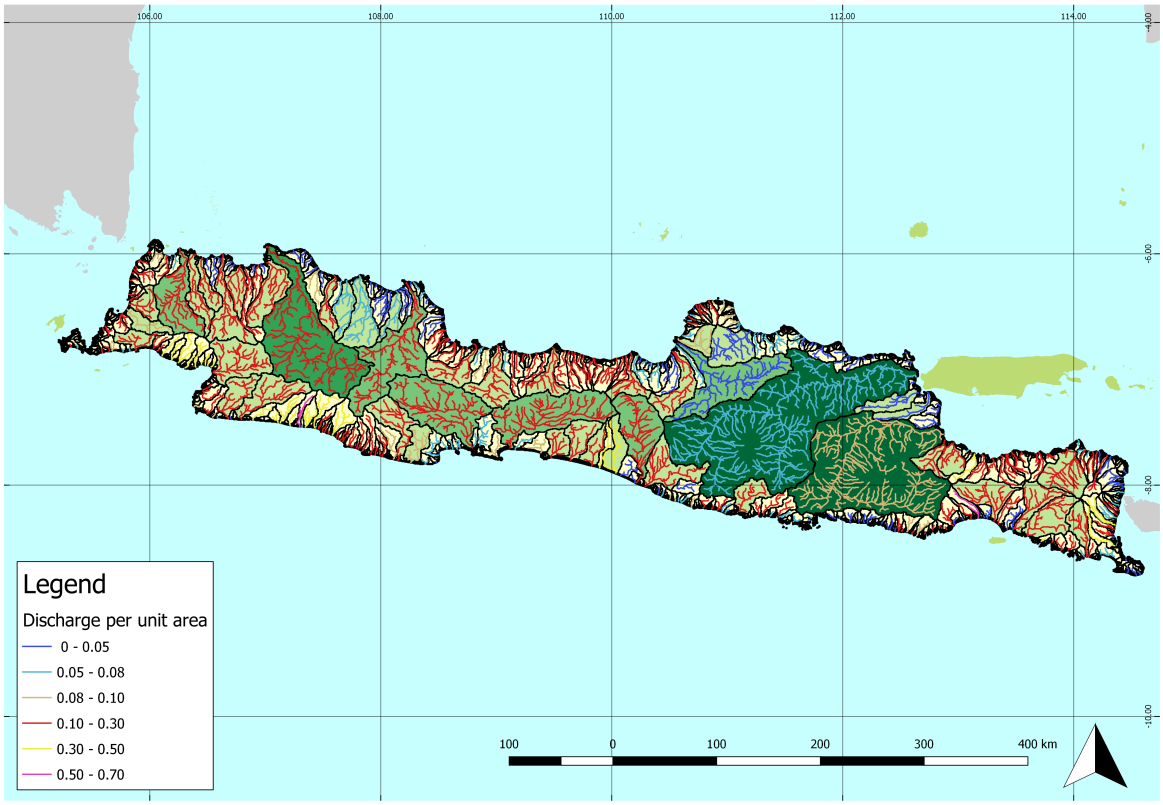
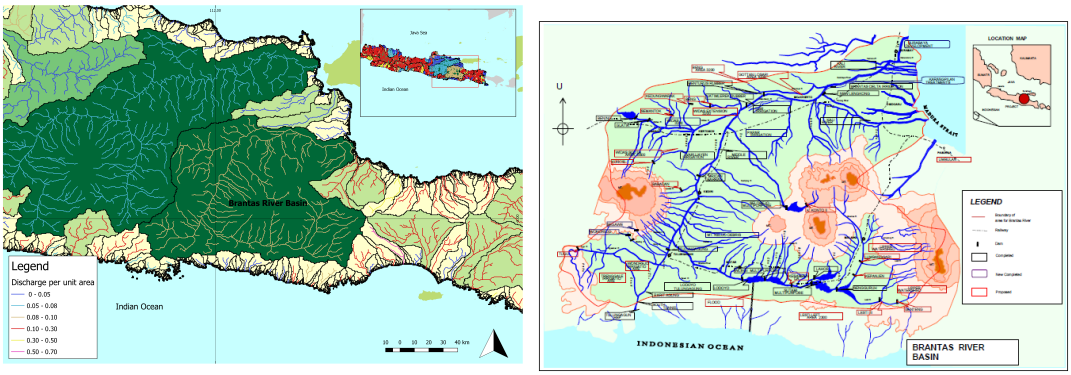


Figure 3.7: Discharge per catchment area of Java Island with 100 year return periods

One of a good example to discuss is the Brantas river basin, which is one of the most significant catchment in the Java island. Brantas river basin is a crucial basin in the East Java, as it has been invtervened by human for irrigation purposes and water supply, including the flood control. Within Brantas river basin, eight reservoirs and several hydraulic structures -such as weir and sluice- are still operating to control the river streamflow [13]. As depicted in figure 3.8, Brantas River Basin has about 0.08 to 0.10 m^3/s per m^2 riverine discharge for return period of 100 years, in which shows a moderate potential hazard compared to the other catchments. This condition becomes reasonable supported by the fact of many hydraulic structures were built in the catchment. In addition to that, an alternative to merely evaluate Brantas River Basin would be a valuable future research. Throughout this illustration, it could be concluded that this map is reasonably useful to present a preliminary flood risk assessment.

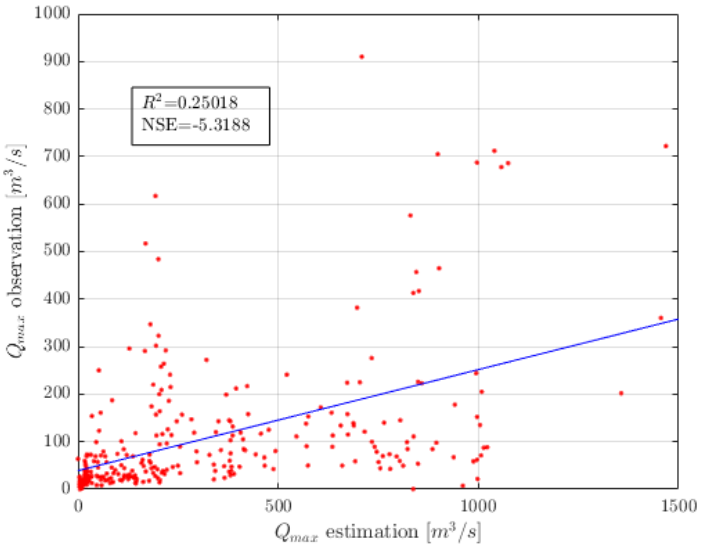


(a) Discharge map of Brantas River Basin (b) Network Brantas River Basin (Courtesy of Ramu [13])

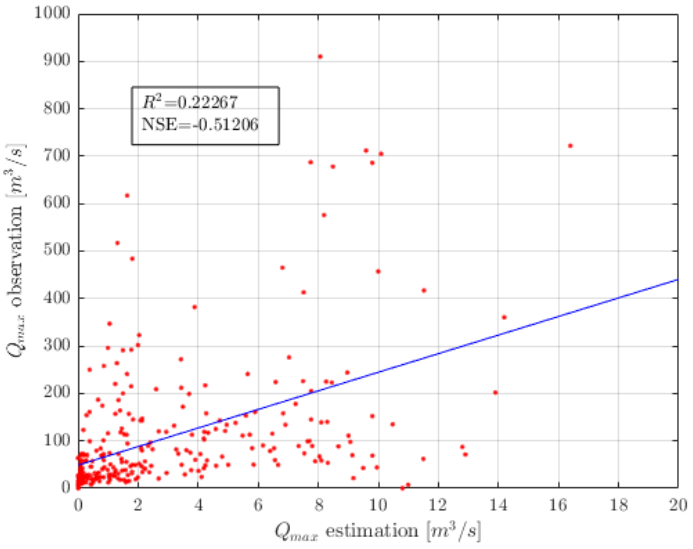
Figure 3.8: Brantas River Basin, where is located in the East Java, has about 0.08 to 0.10 m^3/s per m^2 riverine discharge for return period of 100 years, in which shows a moderate hazard compared to the other catchment. It could be evaluated by the presence of many hydraulic structures within the basin

MEAN AND MODE SAMPLING RESULT – In order to solve the analytical copula, one way is to do a sampling of the distribution and averaging the sampling result to get the final result. However, inferencing the model using the mean data of the estimation might not be the best representation of the result. Instead of the using mean sampling, the mode or most frequent value in the sampling would be used to represent the result of the model. Figure 3.9 shows the generation of mean and mode sample as the result of the model.

The provision of Annual of Maximum Discharge with the most common data compared to the observation shows a very similar coefficient of determination compared to the average value. However, the obtained order of magnitude for the estimated discharge is underestimated in the mode system. On the other hand in the mean system, the estimated discharge is overestimated. The NSE shows a significant change in the mode system, with ten times smaller result. In conclusion, using mode as the estimator does not give any satisfactory result compared to the mean with the same number of simulation generated in the script.



(a) Mean sampling value



(b) Mode sampling value

Figure 3.9: Plot discharge with different sampling estimator method with x-axis represents the estimated Q_{max} and y-axis represent the observation Q_{max}

CORRELATION BETWEEN AREA AND DISCHARGE – An alternative way to validate is to check the data distribution to ensure the quality of measured data, by plotting the area with respect to the discharge. Since it is assumed to use natural flow, hence the larger catchment area results in higher discharge. From the BN model, the highest hierarchy of another node for Maximum Discharge is the area of the catchment. It means that area gives the highest influence in estimating maximum discharge. Hence it is worth to analyse the dependency between two variables using semi-correlations to give an idea of the relationship between these two variables. This is done by transforming the set of data into standard normal and dividing them into four different parts. Then, computing the correlation of each part of the multivariate plot. Data is taken from the yielded datasets that have been mentioned earlier and plotted in figure 3.10.

Based on figure 3.10, in general, the partial correlation behaves in favour with the hypothesis, and shows a positive overall correlation of 0.78. Accordingly, it agrees that the larger the size of the catchment, it yields

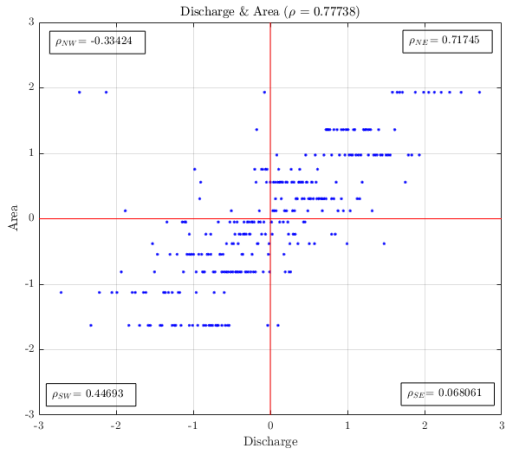


Figure 3.10: Correlation factor between area and maximum discharge

bigger river stream flow. On the other hand, a peculiar plotting data depicts on the North-West part, that has a negative correlation of -0.33. It is also noticeable that there are two outliers, that could be reconsidered to be included in the model. Apart from these variables, another nodes could be worth to be analysed in the future research, to have an insight toward the behaviour of the spatial datasets.

4

Conclusion and Recommendation

4.1. Conclusion

BN as a robust statistic method has a powerful methodology to relate many aspects, in this case, related to the estimation of maximum discharge. However, BN showed inferior performance on the application in Java Island, especially compared to its predecessor cases like Europe and USA. The estimation of annual maximum discharge is overly valued, with R^2 of 0.25 and NSE of -5.32. This might arise due to several reasons. Bayesian Network model for the riverine flood was initiated based on larger catchment scale-based. On the other hand, the size of Java Island is not as considerable as Europe and the US, in which against the assumption of the model. Therefore, the model does not execute at its maximum performance. Furthermore, merely 229 discharge were used to evaluate the performance of the model. The number of data that been used to validate is not adequate enough to represent the model. Apart from the performance of the model, several points could be derived, such as:

- Six out of fourteen stations shows an underestimation results in the frequency analysis, which relatively poor. On the other hand, the remaining stations show an overestimation.
- The resolution of the spatial datasets of Java Island is below par. In that sense, the derivation of the flood hazard map which depends on this data, yield an inferior resolution. Nevertheless, the flood hazard map is able to illustrate a very coarse representation of flood risk quantification.
- By changing the system of the sampling using mode – the value that appears most often – to generate the result does not increase the performance of the model, compared to the mean sampling value.

4.2. Recommendation

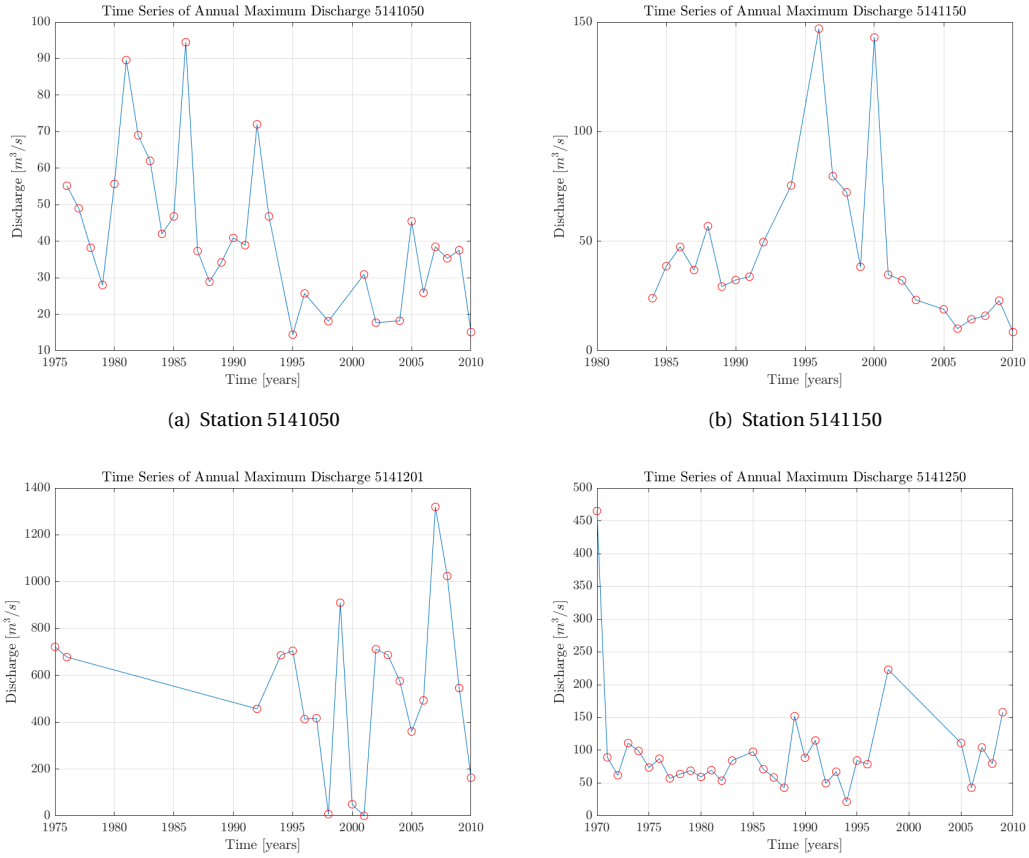
There are still rooms for improvement for the further research. Throughout the process, several recommendations could be mentioned out, as follow:

- BN model lies on the assumption of Standard Normal Copula or Gaussian due to its rapid calculations; however, it may affect the behaviour of the model. Evaluating another type of copula would be another solution to enhance the model.
- The exisiting land use nodes might not well contributed in Java Island. Adding more node that in favour with the mechanism of the flood in Java Island would be another alternative to improve the performance of the model. This approach was performed in the case of USA and did slightly improve the model performance.
- Another flooding mechanism that takes place in Java Island might be different compared to the built model. Most of the flooding problem that occurred was a flash flood that happens in a very short duration. This is against the assumption of the model that is not suitable for the short duration.
- The most common issue working with the spatial datasets is the resolution that leads to the accuracy of the model. It bestows more uncertainty to the model performance. Thus, quality assurance towards the spatial datasets is a rewarding solution to ensure better performance of the model.

A

Appendix - Time Series of River Discharge

In the chapter 3.1.4, it is mentioned that the discharge observation data were taken from GRDC (Global River Dicharge Centre). In Java Island, there are fourteen recording stations spread throughout the island. The initial time scale of the data is daily measurement, in contrary, we are intrested in annual maxima. Hence, an extraction of annual daily maximum was perfomed to all stations. Each station has different period of recording. In these following figure, the time series are plotted for each station with different time period. By doing so, it gives the user a broad idea towards the order of the magnitude of annual daily river discharge in the respective station.



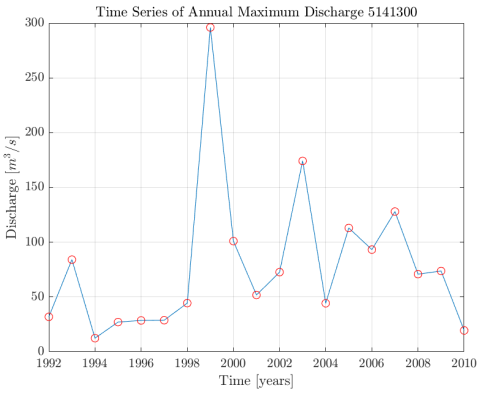
(a) Station 5141050

(b) Station 5141150

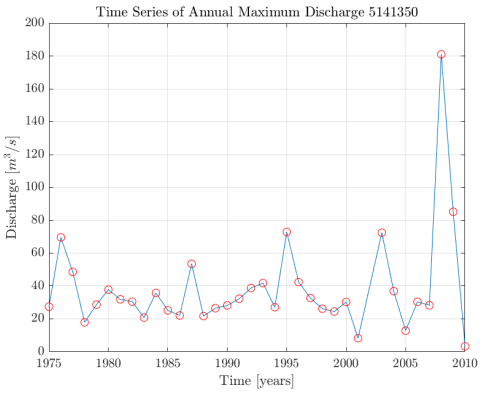
(c) Station 5141201

(d) Station 5141250

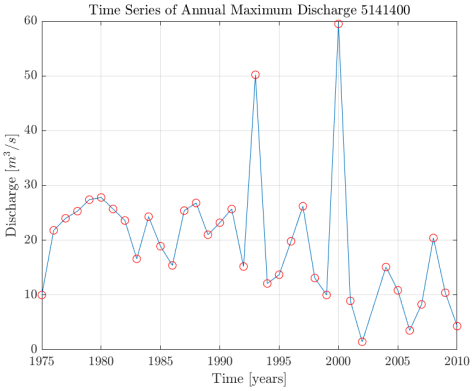
Figure A.1: Time series of annual maximum river discharge



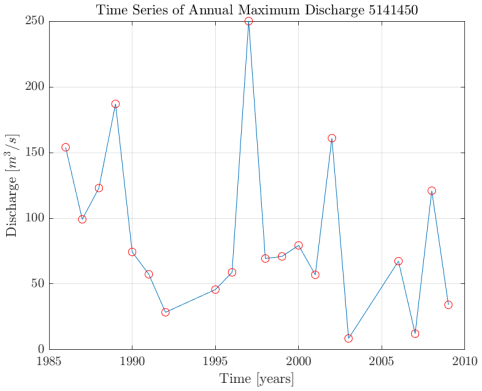
(a) Station 5141300



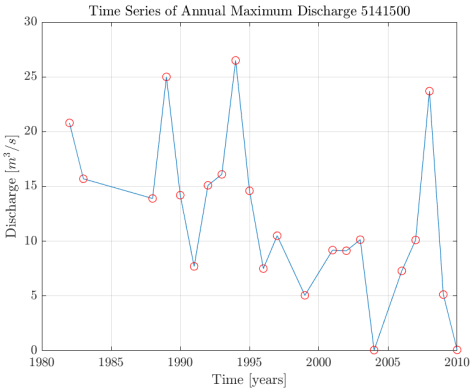
(b) Station 5141350



(c) Station 5141400

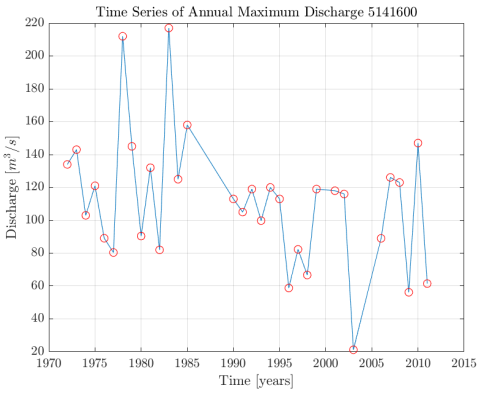


(d) Station 5141450

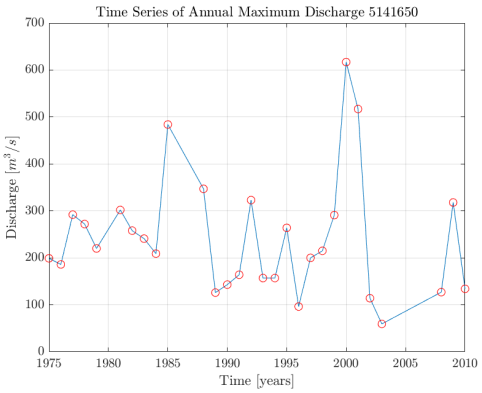


(e) Station 5141500

(f) Station 5141550

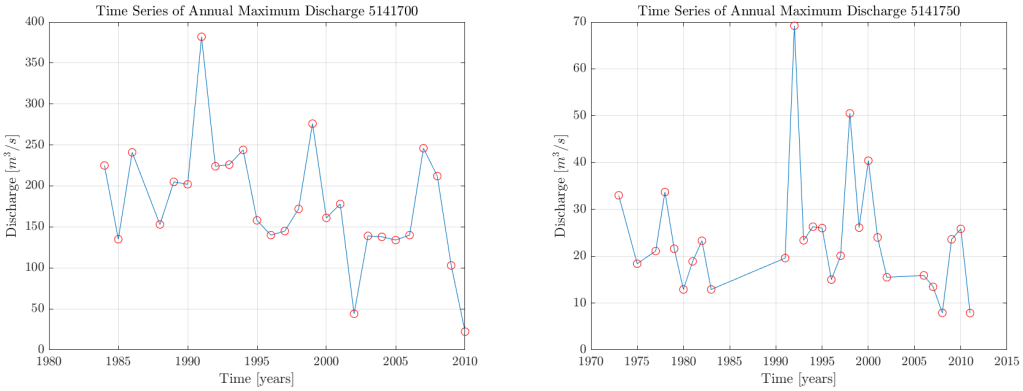


(g) Station 5141600



(h) Station 5141650

Figure A.2: Time series of annual maximum river discharge



(a) Station 5141700

(b) Station 5141750

Figure A.3: Time series of annual maximum river discharge

Bibliography

[1] Flood management in selected river basins sector project. Technical report, Asian Development Bank, 08 2016.

[2] A.D. Bank. *Indonesia: Country Water Assessment*. Country Sector and Thematic Assessments. Asian Development Bank, 2016. ISBN 9789292573614. URL https://books.google.nl/books?id=liTIDAAAQBAJ.

[3] Ole Bøssing Christensen, Martin Drews, Jens Hesselbjerg Christensen, Klaus Dethloff, Klaus Ketelsen, Ines Hebestadt, and Anette Rinke. Technical report 06-17 the hirham regional climate model version 5 (β). Technical report, Danish Meteorological Institute (DMI), 2007.

[4] A. A. O. Couasnon. Characterizing flood hazard a two spatial scales with the use of stochastic models(An application to the contiguous United States of America and the Houston Ship Channel). Master's thesis, Delft University of Technology, the Netherlands, 2017.

[5] New World Encyclopedia. Java — new world encyclopedia,, 2017. URL http://www.newworldencyclopedia.org/p/index.php?title=Java&oldid=1008106. [Online; accessed 14-December-2017].

[6] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

[7] Anca Hanea, Oswaldo Morales Napoles, and Dan Ababei. Non-parametric bayesian networks: Improving theory and reviewing applications. *Reliability Engineering & System Safety*, 144:265–284, 2015.

[8] Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. World map of the köppen-geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3):259–263, 2006.

[9] Daniel N Moriasi, Jeffrey G Arnold, Michael W Van Liew, Ronald L Bingner, R Daren Harmel, and Tamie L Veith. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900, 2007.

[10] Mauro Naghettini. *Fundamentals of statistical hydrology*. Springer, 2016.

[11] Dominik Paprotny and Oswaldo Morales-Nápoles. Estimating extreme river discharges in europe through a bayesian network. *Hydrology and Earth System Sciences*, 21(6):2615, 2017.

[12] P Prinos. Review of flood hazard mapping. *T03-07-01*, 2009.

[13] Kikkeri V. Ramu. Brantas river basin case study indonesia. Technical report, World Bank, 12 2004.

[14] Jochen Schanze. Flood risk management–a basic framework. *Flood risk management: hazards, vulnerability and mitigation measures*, pages 1–20, 2006.

[15] Hessel C Winsemius, JCJH Aerts, Ludovicus PH van Beek, Marc FP Bierkens, Arno Bouwman, Brenden Jongman, Jaap CJ Kwadijk, Willem Ligtvoet, Paul L Lucas, Detlef P van Vuuren, et al. Global drivers of future river flood risk. *Nature Climate Change doi*, 10, 2015.