Detection of satellite data-based flood-prone areas using logistic regression in the central part of Java Island

G Pratidina¹, Suroso²,⁴, P B Santoso³
¹,²,³Department of Civil Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Purwokerto

E-mail: ⁴surosotsipil@unsoed.ac.id

Abstract. The history of natural disasters recorded in BNPB (2019) explains that the total number of natural disaster events in the central part of Java (Central Java Province and Special Region of Yogyakarta Province) ranks highest in terms of the number of frequency of occurrences nationally. Of the total natural disasters that have occurred in Central Java, the number of floods is ranked third after the landslide and tornado disaster, which is around 1500 disasters. Various factors that can cause flooding cannot be eliminated. However, what is more necessary is how to control the impacts caused by floods so that they can be managed and monitored appropriately. One effort to overcome the problem of the threat of flooding is to develop a detection model for flood-prone areas. In this study, the detection of flood-prone areas was carried out by using a logistic regression method that takes into account the variables that cause flooding such as elevation, land slope, river distance, flow accumulation, rainfall, and runoff coefficients. The results of the modelling, obtained coefficients of the variables/parameters mentioned earlier, namely intercept (5.05766 – 16.13210), rainfall (-0.01547 – 0.04075), elevation (-0.02173 – -0.00592), slope (-0.28108 – -0.01940), runoff coefficient (-9.10476 – -7.15039), river distance (0.00038 – 0.00783), and flow accumulation (-9.26342E-06 – 0.00309). The level of success in this modelling testing was 93.47826% - 98.26087% of 329 flood event data points and not floods.

1 Introduction
The history of natural disasters recorded in the National Disaster Management Agency (BNPB) explains that the total number of natural disasters in the central part of Java island (Central Java Province and Special Region of Yogyakarta Province, DIY) rank highest in terms of the number of occurrences of frequency nationally. More than 7000 recorded various natural disasters ranging from floods, landslides, tornadoes, earthquakes, tsunamis, forest fires, droughts, and other natural disasters, which have occurred in this area in the period up to 2019. This indicates that Central Java has a very high level of vulnerability to the threat of natural disasters.

According to the BNPB, natural disasters that occurred in the central part of Java Island were dominated by floods, landslides, and tornadoes. From the total natural disasters that have occurred in those regions, the number of floods occurred in third place after landslides and tornadoes, which were around 1500 disaster events as can be seen in table 1. Furthermore, there are more than 25 districts/cities in the provinces of Central Java and DIY that are vulnerable to flood hazards due to occurring every year during the rainy season [1].
Table 1. Frequency of natural disasters in the central part of Java Island (1815-2019).

| No | Type of Disaster          | Central Java | Special Region of Yogyakarta |
|----|---------------------------|--------------|------------------------------|
| 1  | Floods                    | 1429         | 68                           |
| 2  | Landslides                | 2098         | 124                          |
| 3  | Floods and Landslides     | 1            | 0                            |
| 4  | Tidal / Abrasion Wave     | 23           | 9                            |
| 5  | Tornado                   | 1961         | 177                          |
| 6  | Drought                   | 433          | 48                           |
| 7  | Forest and Land Fires     | 88           | 1                            |
| 8  | Earthquake                | 35           | 14                           |
| 9  | Tsunami                   | 7            | 0                            |
| 10 | Volcanic Eruption         | 28           | 15                           |

A flood is an event or situation, in which an (urban) area is submerged since the water level in the rivers increase and overflow to plains around the river due to heavy rainfall intensity [2]. Recently, there is an increase in extreme rainfall resulting from global warming phenomena, namely increased in the surface temperature of the earth leading to climate change. The United Nations Office for the Coordination of Humanitarian Affairs report states that Indonesia is one of the countries most vulnerable to extreme climate-related disasters [3].

In addition to extreme rainfall factors, the cause of the increase in floods is human activities because of increasing population growth and rapid infrastructure developments [4]. It is a logical consequence that the many development activities will have an impact on changes in land use from infiltration areas to build areas reducing the flood retention capacity.

Various factors that can cause flooding cannot be completely eliminated. However, what is more, necessary is how to control the impacts of floods so that they can be managed and monitored appropriately. One effort to overcome the problem of the threat of flooding is to develop a model that can detect flood-prone areas. Accurate detection of flood-prone areas can strengthen the flood disaster early warning system so that it can minimize both material and non-material losses caused by floods.

This study aims to develop a model for flood prediction using logistic regression.

![Figure 1. Map of Central Java and Special Region of Yogyakarta Provinces.](image-url)
2 Data and study location

2.1 Study location

The location of the study was conducted in the central part of Java, as shown in figure 1. Geographically, the Province of Central Java is located at 5° 40’ and 3° 30’ South Latitude and between 108° 30’ and 111° 30’ East Longitude (including Karimunjawa Island). The farthest distance from West to East is 263 km and from North to South 226 km.

According to the Regional Disaster Management Agency (BPBD) in Central Java Province, a number of regions were listed as prone to flash flood. The area is on the banks of a large river. Areas prone to flooding include Banyumas, Purwokerto, Pati, Demak, Kudus, Brebes, Cilacap. Areas are prone to large rivers such as Karanganyar, Solo and Sukoharjo.

2.2 Data collection

The data needed in this study can be explained as follows:

2.2.1 Topographic data obtained from digital elevation model (DEM). This topographic data uses the most recent Shuttle Radar Topography Mission (SRTM) Data with the spatial resolution of 30m x 30m [5]. This DEM data will be processed to find out the slope and configure pixels where flow. The following is a map of Central Java DEM which can be seen in Figure 2. slope map in Figure 3., and flow accumulation map in Figure 4.

Figure 2. DEM map of Central Java and Special Region of Yogyakarta Provinces.

Figure 3. Slope map of Central Java and Special Region of Yogyakarta Provinces.
Figure 4. Flow accumulation map of Central Java and Special Region of Yogyakarta Provinces.

2.2.2 Precipitation data. Daily rainfall data used in this study is the TRMM (Tropical Rainfall Measuring Mission) satellite from the range of 1998 to 2018 [6], [7], [8]. The satellite data is processed using ArcGis, Microsoft Excel, and RStudio software. The following is an example of the TRMM rainfall map of Central Java Province on May 16, 2016, as can be seen in Figure 5.

Figure 5. TRMM rainfall map of Central Java and Special Region of Yogyakarta Provinces.

2.2.3 Spatial data on flood events (historical flood)

Figure 6. Historical flood map of Central Java and Special Region of Yogyakarta Provinces.
The historical flood data used is flood data that has occurred in the Central Java and DIY provinces. The data is then combined with daily rainfall data, so that rainfall is known in the flooded area. Map of flood events in Central Java and DIY are shown in Figure 6.

2.2.4 Land use land cover data (LULC). LULC data was obtained from TERRA MODIS satellites and data were used from 2001 to 2017 [9]. The data will show changes in land use caused by development in the Central Java Province. The following is an example of the 2017 Central Java LULC map which can be seen in Figure 7.

![Figure 7. LULC map of Central Java and Special Region of Yogyakarta Provinces.](image)

2.2.5 River network data. River network data is obtained from http://tanahair.indonesia.go.id per district of Central Java. This data will configure the river network in Central Java and display the distance of flood-prone areas with the river network. Map of the river network in Central Java Province can be seen in Figure 8.

![Figure 8. Stream map of Central Java and Special Region of Yogyakarta Provinces.](image)

The data that has been mentioned, will later be used for data processing as a logistic regression modeling factor. The following is a summary of the data types that will be used in Table 2.
### Tabel 2. The data layer of study area.

| Parameter                      | Sub-Classification         | Data Type | Scale               |
|-------------------------------|----------------------------|-----------|---------------------|
| Historical Flooded Area       | Flood Extent               | Point     | -                   |
|                               | Topographic Map in meter   | Grid      | 30 m x 30 m         |
| DEM                           | Slope in degree            | Grid      | 30 m x 30 m         |
|                               | Flow Accumulation          | Grid      | 30 m x 30 m         |
| TMPA                          | Precipitation in mm        | Grid      | 27.5 km x 27.5 km   |
| MODIS                         | Land Cover                 | Grid      | 500 m x 500 m       |
| Stream Map                    | Distance from the Nearest Stream | Grid | 15 x 15 m |

### 3 Method

As mentioned above, the purpose of this study is to predict or detect flood-prone areas in Central Java and DIY using data from certain agencies and satellites and logistic regression models. The stages of data analysis and processing to be carried out are as follows:

#### 3.1 Data processing

##### 3.1.1 Flood record point data analysis (historical flood).

Data on flood events are the main data in this study. The data was obtained from the BNPB (National Disaster Management Agency) from 2011 to 2018. For data on flood recording points that occurred in Central Java and DIY provinces, there were 215 data on flood events with 9 flood events in 2011 floods, in 2012 there were 16 floods, in 2013 there were 49 floods, in 2014 there were 49 floods, in 2015 there were 23 floods and 2018 as many as 69 floods.

##### 3.1.2 Analysis of TRMM daily rainfall data.

TRMM rainfall data was obtained after knowing the location of the flood event through historical flood data. Before data processing is carried out, the thing that needs to be done is the time adjustment of the historical flood data. Because TRMM data uses time with the format GMT + 0, then the time of the historical flood data needs to be changed first by reducing 6 hours from the original time.

##### 3.1.3 Elevation data analysis.

In this study, elevation data were obtained from the DEM (Digital Elevation Model) data obtained from the Indonesian Geospatial Information Agency (http://tanahair.indonesia.go.id). DEM data is in the form of pixels and can be processed with ArcGIS software. The elevation obtained refers to data on flood recording points. After processing it in ArcGIS software, the output that is produced is the elevation data in meters at each coordinate data point of the flood record.

##### 3.1.4 Slope data analysis.

Like elevation data, slope or slope data is also obtained from DEM data. DEM data is processed into land slope data using ArcGIS software. The output produced is the magnitude of the slope of the land in Central Java Province and DIY in the form of a percentage (%).

##### 3.1.5 Analysis of LULC data.

LULC data obtained from TERRA MODIS satellites were processed using ArcGIS software to obtain the type of land use at flood recording points in Central Java and DIY Provinces. Land use classification uses land use classification data from the Earth Observation and Modelling Facility (EOMF) which can be accessed at http://www.eomf.ou.edu/. This land use classification is obtained from the International Geosphere-Biosphere Program (IGBP) which is a research program that studies the phenomenon of global change that took place from 1987 to
2015. After obtaining LULC data, then the classification is changed according to the run-off coefficient of each flood point.

3.1.6 River network data analysis. River network data is obtained from the Geospatial Information Agency (BIG) and is used to identify river networks in Central Java and DIY Provinces. This data will result in the distance of the nearest river network to flood recording points in Central Java and DIY Provinces. Following is a summary of some data samples or parameters that will be used to do logistic regression modelling can be seen in Table 3.

Table 3. Sample of logistic regression parameter data.

| No | Disaster | Long  | Lat  | Precipitation (mm) | Elevation (m) | Slope (%) | LULC | C   |
|----|----------|-------|------|--------------------|---------------|-----------|------|-----|
| 55 | Flood    | 110.201 | -7.25961 | 0.000000            | 650           | 3.409910  | 9    | 0.45|
| 121| Flood    | 110.201 | -7.25961 | 0.000000            | 650           | 3.409910  | 9    | 0.45|
| 226| Flood    | 111.073 | -6.88254 | 9.389348            | 49            | 3.607020  | 14   | 0.175|
| 239| Flood    | 111.097 | -6.72001 | 3.904148            | 45            | 3.985060  | 14   | 0.175|
| 243| Flood    | 111.058 | -6.89586 | 5.830828            | 650           | 3.409910  | 9    | 0.45|
| 286| Flood    | 110.201 | -7.25961 | 11.828920           | 650           | 3.409910  | 9    | 0.45|
| 293| Flood    | 110.986 | -6.89914 | 50.549920           | 20            | 3.011980  | 14   | 0.175|
| 305| Flood    | 111.038 | -6.74867 | 47.854560           | 14            | 0.462433  | 13   | 0.5 |
| 306| Flood    | 111.038 | -6.74867 | 47.854560           | 14            | 0.462433  | 13   | 0.5 |
| 307| Flood    | 110.924 | -6.89947 | 52.878580           | 5             | 1.178840  | 12   | 0.175|
| 309| Flood    | 111.046 | -6.74496 | 2.456600            | 13            | 1.178840  | 13   | 0.175|
| 498| Flood    | 109.692 | -7.69878 | 0.000000            | 19            | 5.551110  | 12   | 0.175|
| 500| Flood    | 110.959 | -6.45553 | 80.877680           | 52            | 8.068660  | 14   | 0.5 |
| 503| Flood    | 110.834 | -6.81191 | 52.647260           | 19            | 3.807780  | 13   | 0 |
| 504| Flood    | 110.510 | -6.92071 | 18.382400           | 0             | 0.000000  | 17   | 0.5 |
| 505| Flood    | 109.200 | -6.87534 | 60.252830           | 3             | 1.906000  | 13   | 0.175|
| 509| Flood    | 109.634 | -6.91320 | 24.360450           | 10            | 2.287760  | 12   | 0.45|
| 511| Flood    | 109.462 | -7.64387 | 7.204940            | 122           | 17.299000 | 9    | 0.175|
| 536| Flood    | 108.677 | -7.30763 | 18.757860           | 140           | 8.854100  | 14   | 0.175|
| 537| Flood    | 108.807 | -7.48678 | 43.601220           | 6             | 1.906000  | 12   | 0.175|
| 542| Flood    | 109.250 | -7.60460 | 31.781600           | 7             | 1.178840  | 12   | 0.175|

3.2 Logistic regression modelling

Logistic regression analysis is a regression method used to find the relationship between categorical response variables with the nominal, ordinal scale with one or more continuous and categorical explanatory variables [10]. The response variable \( y \) from binary logistic regression consists of two categories namely "success" and "failure", where the notation for the "success" category and for the "fail" category. So that the response variable follows the Bernoulli distribution for every single observation. The probability functions for each observation are as follows:

\[
f(y) = \pi^y (1-\pi)^{1-y}; y = 0,1
\]  

(1)

where if \( y = 0 \) then \( f(y) = 1-\pi \) and \( y = 1 \) then \( f(y) = \pi \), so that the logistic regression function is obtained as follows:
\[ f(z) = \frac{e^z}{1+e^z}; \quad z = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \]  

(2)

where p is many predictor variables. The value \( f(z) \) is between 0 and 1 for each \( z \) value given because the value of \( z \) itself is between antara and \( \infty \). The logistic regression model actually describes the probability of an object. The logistic regression model is as follows:

\[ f(z) = \frac{e^z}{1+e^z}; \quad z = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \]  

(3)

The function \( \pi(x) \) is a nonlinear function so it needs to be transformed by using logic transformations to obtain linear functions in order to see the relationship between the response variable \( y \) and the predictor variable [11].

\[
  g(x) = \logit \pi(x) \\
  = \log \frac{\pi(x)}{1 - \pi(x)} \\
  = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \\
  = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \\
  = X\beta
\]

4 Results and discussion

4.1 Model fitting

Development of a logistic regression model was carried out with Rstudio software using a file that has a .csv format that contains Flood data (rainfall, elevation, slope, run-off coefficient, distance to the river). Flood event data was used as non-independent variables and the Flood event data used for this modeling was 70% of the randomized Flood occurrence data.

The data used in logistic regression modeling are 215 Flood record coordinate data and regional coordinate points which have a low probability of 114 data floods. Then the amount of data used in this modeling is 230 data taken randomly. After randomly collected data, the coefficients of logistic regression were obtained. In this study, the modeling that has been done is 500 times modeling to find out the range of the coefficients of logistic regression. The following coefficients from the logistic regression generated from the modeling are shown in Table 4.

| Parameter          | Coefficient of Logistic Regression |
|--------------------|-----------------------------------|
|                    | 5%      | 50%     | 95%     | Average |
| (Interception)     | 5.05766 | 7.5110  | 16.13210| 9.56692 |
| Precipitation      | -0.01547| 0.01210 | 0.04075 | 0.01246 |
| Elevation          | -0.02173| -0.00831| -0.00592| -0.01199|
| Slope              | -0.28108| -0.10251| -0.01940| -0.13433|
| Coefficient Runoff | -9.10476| -0.38548| 7.15039 | -0.77995|
| DNS                | 0.00038 | 0.00423 | 0.00783 | 0.00415 |
| Flow Accumulation  | -9.26342| 0.00137 | 0.00791 | 0.00309 |
4.2 Model validation

In logistic regression modeling validation, the data used is all data minus data that has been used for logistic regression modeling or 30% of all data. The data used in the validation is 99 data. Logistic regression modeling validation aims to determine the reliability of the model that has been done. So that it will be known how much success is obtained and the value of reliability from the modeling itself. Of the 500 experiments that have been conducted, data that deviate approximately 4 data from 99 validation data are generated. And it can be concluded that the accuracy for the validation of logistic regression modeling is 93.47826% - 98.26087%.

4.3 Prediction of flood prone areas

After finding out the coefficients of each flood parameter, the next step is to map flood-prone areas in Central Java and DIY Provinces. The mapping of flood-prone areas was made based on the results obtained from logistic regression modeling by mapping the results of binary numbers, namely 1 means flooding and 0 means no flooding in Central Java and DIY. By changing the map of Central Java and DIY into a grid size of 1000 m x 1000 m, the results obtained can be seen in table 5, below.

Table 5. Area of flood-prone areas in Central Java and DIY Provinces

| Districts   | Capital City | Province       | Area (km²) | Area of Flood (km²) | Area of Not Flood (km²) |
|-------------|--------------|----------------|------------|---------------------|-------------------------|
| Banjarnegara| Banjarnegara | Central Java   | 1160       | 904                 | 256                     |
| Bantul      | Bantul       | Special Region of Yogyakarta | 490 | 24 | 466 |
| Banyumas    | Banyumas     | Central Java   | 1413       | 1133                | 280                     |
| Batang      | Batang       | Central Java   | 742        | 265                 | 477                     |
| Blora       | Blora        | Central Java   | 1975       | 1402                | 573                     |
| Brebes      | Brebes       | Central Java   | 1726       | 1160                | 566                     |
| Cilacap     | Cilacap      | Central Java   | 2314       | 1623                | 691                     |
| Demak       | Demak        | Central Java   | 955        | 653                 | 302                     |
| Grobogan    | Purwodadi    | Central Java   | 2024       | 1742                | 282                     |
| Gunung Kidul| Wonosari     | Special Region of Yogyakarta | 1460 | 177 | 1283 |
| Jepara      | Jepara       | Central Java   | 1048       | 411                 | 637                     |
| Boyolali    | Boyolali     | Central Java   | 1060       | 917                 | 143                     |
| Karanganyar | Karanganyar  | Central Java   | 833        | 569                 | 264                     |
| Kebumen     | Kebumen      | Central Java   | 1305       | 1094                | 211                     |
| Kendal      | Kendal       | Central Java   | 1012       | 366                 | 646                     |
| Klaten      | Klaten       | Central Java   | 660        | 351                 | 309                     |
| Magelang City| Magelang   | Central Java   | 13         | 4                   | 9                       |
| Pekalongan City | Pekalongan | Central Java   | 77         | 63                  | 14                      |
| Salatiga City| Salatiga    | Central Java   | 17         | 13                  | 4                       |
| Semarang City| Semarang    | Central Java   | 380        | 187                 | 193                     |
| Surakarta City| Surakarta  | Central Java   | 56         | 24                  | 32                      |
| Tegal City  | Tegal        | Central Java   | 42         | 34                  | 8                       |
| Yogyakarta City| Yogyakarta| Special Region of | 32         | 1                   | 31                      |
From the results obtained, the area is the area of the approach in grid units. For example, the area of Banjarnegaras is 1160 grids, meaning that Banjarnegaras has an area of 1160 grids with each grid having an area of 1,000,000 m² and flood-prone areas which are predicted to have an area of 904 grids and non-flooded areas predicted to have 256 grids. For more details, the following map produced for mapping flood-prone areas can be seen in Figure 9.

![Map of flood-prone areas in Central Java and DIY Provinces](image)

**Figure 9.** Mapping detection of flood-prone areas in Central Java and DIY Provinces.
5 Conclusions
Based on the results of the research that has been done, it can be concluded that the results obtained are as follows:

a. Parameter coefficients obtained from logistic regression modelling, each parameter has its own coefficients, namely intercept (5.05766 – 16.13210), rainfall (-0.01547 – 0.04075), elevation (-0.28108 – 0.01940), runoff coefficient (-9.10476 – 7.15039), river distance (0.00038 – 0.00783), and flow accumulation (-9.26342E-06 – 0.00309).

b. From the parameter coefficients obtained, it can be concluded that the greater the coefficient obtained the greater the chance for a Flood disaster.

c. Of the 500 experiments that have been conducted, data that deviate approximately 4 data from 99 validation data are generated. In addition, it can be concluded that the accuracy for the validation of logistic regression modelling is 93.47826% - 98.26087%.

d. There are 25 districts that have the potential to be prone to flooding. This is concluded from the broader area 1 (flood prediction) than 0 (prediction of not flooding) from each district.

6 Acknowledgments
Financial support for completing this research is graciously provided by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia with contact number: 176/SP2H/LT/DPRM/2019.

7 References
[1] Marfai M A King L Singh L P Mardiatno D Sartohadi J Hadmoko D S and Dewi A 2008. Natural hazards in Central Java Province Indonesia: an overview Environmental Geology 56(2) 335-351
[2] Lim J and Lee K S 2018 Flood mapping using multi-source remotely sensed data and logistic regression in the heterogeneous mountainous regions in north korea Remote Sensing 10(7) 1036
[3] BAPPENAS 2011. Indonesia adaptation strategy-improving capacity to adapt climate change The Ministry of National Development Planning/National Development Planning Agency.
[4] Haghizadeh A Siahkamari S Haghiabi A H and Rahmati O 2017 Forecasting flood-prone areas using Shannon’s entropy model. Journal of Earth System Science 126(3) 39
[5] Mukul M Srivastava V and Mukul M 2015 Analysis of the accuracy of shuttle radar topography mission (SRTM) height models using international global navigation satellite system (IGS) network. Journal of Earth System Science 124(6) 1343-1357
[6] Huffman G J Adler R F Bolvin D T and Nelkin E J Nelkin. 2010 The TRMM multi-satellite precipitation analysis (TMPA) Chapter 1 in Satellite Rainfall Applications for Surface Hydrology
[7] Huffman G J 1997 Estimates of root-mean-square random error for finite samples of estimated precipitation J Appl. Meteor
[8] Ouma Y O Owiti T Kipkorir E Kibiay J and Tateishi R. 2012 Multitemporal comparative analysis of TRMM-3B42 satellite-estimated rainfall with surface gauge data at basin scale: daily, decadal and monthly evaluations. International Journal of Remote Sensing 33(24), 7662-7684.
[9] Friedl M and Sulla-Menashe D 2015 MCD12Q1 MODIS/terra+aqua land cover type yearly L3 global 500m SIN grid V006 NASA EOSDIS Land Processes DAAC
[10] Pradhan B 2010 Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing. Journal of Spatial Hydrology 9(2)
[11] Hosmer D W dan S Lemeshow (2000) : *Applied Logistic Regression* Second Edition John Willey & Sons New York.

[12] Tehrany M S, Shabani F, Jebur M N 2017 *GIS-based spatial prediction of flood areas using standalone frequency ration, logistic regression, weight of evidence and their ensemble techniques.* Geomatics, Natural Hazards and Risk. 8:2. 1538-1561. DOI : 10.1080/19475705.2017.1362038.

[13] Trigila A, Iadanza C, Esposito C and Scarascia-Mugnozza G. 2015. *Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in giampilieri(NE Sicily, Italy).* Geomorphology, 249, 119-136.