Imperial College London

Department of Earth Science and Engineering

MSc in Environmental Data Science and Machine Learning

Independent Research Project
Final Report

Flood Probability Prediction Based on Machine Learning

by

Chaofan Wu

Email: chaofan.wu22@imperial.ac.uk
GitHub username: edsml-cw522
Repository: https://github.com/ese-msc-2022/irp-cw522

Supervisors:

Sesinam Dagadu(MEng)

Yves Plancherel(PhD)

Company'name: SnooCODE Limited
Company's address: 18 Royal Palm Avenue, McCarthy Hill, Accra-Ghana

August 2023

Contents

Abstract

In this research, a machine learning-based methodology is used for predicting the probability of flood occurrences in the Ghanaian region. Intended to offer valuable insights for urban planners, emergency response teams, and residents, this approach seeks to alleviate and manage the recurrent severe flooding challenges in Ghana. By amalgamating rainfall data, geospatial information (such as DEM, soil type, land cover etc.), machine learning models (ML), and Geographic Information Systems (GIS), the methodology produces flood probability maps tailored for specific target areas within Ghana. Comprehensive evaluations and analyses of the model were conducted, with experimental outcomes suggesting that the final model (Random Forest [1, 2]) provides prescriptive recommendations, instrumental in curtailing the potential impact areas of floods. Additionally, throughout the modelling process, data from various platforms were collated, culminating in a flood dataset apt for machine learning modelling. This dataset offers convenience for future flood studies in Ghana. In essence, this foundational research in flood prediction furnishes the field with fresh avenues for further exploration.

1 Introduction

During the rainy season spanning May to July, Ghana witnesses almost annual seasonal floods, which pose severe threats to human lives and property [3, 4, 5]. The nation has been subjected to some of the gravest flood events in its historical records, specifically in the years 1955, 1960, 1963, 1973, 1986, 1991, 1995, and 2001. These calamities have inflicted considerable damage to its infrastructure, industrial zones, and residential regions [6, 7]. Consequently, undertaking a comprehensive flood risk assessment for Ghana becomes an imperative and significant endeavour. Studies highlight that flooding in Ghana can be primarily attributed to three factors: 1. climatic determinants, which encompass large-scale rainfall, particularly during the wet season from May to July; 2. geographical elements, inclusive of the country's low-lying terrains and the clayey composition of its soil; 3. anthropogenic influences, such as inadequate drainage systems, the disposal of waste into drainage conduits and water bodies, imprudent development and utilisation of environmentally sensitive areas [3, 4, 5]. If a methodology were devised to visualise and evaluate regions within Ghana that are susceptible to flooding based on these three factors, it would significantly aid urban planners and administrators in devising strategies to counteract this recurrent flood menace [3, 4, 5]. Furthermore, this would also empower emergency response teams to pinpoint areas that necessitate focused interventions during flood incidents, thereby diminishing future damages and risks for the populace. Such measures would play a pivotal role in mitigating the enduring flood challenges faced by Ghana [3, 4, 5].

Currently, one-dimensional (1D) shallow water equation models (1D-SWEs), two-dimensional shallow water equation models (2D-SWEs), and machine learning-based models (ML) are the mainstream methods for modelling and evaluating flood risk [8]. They all have their own advantages and disadvantages. For example, the 1D-SWEs was initially used to simulate floods. It is relatively intuitive and easy to calculate, but its performance is poor for complex urban flood environments (such as multiple river intersections or obstacles such as bridges and dams) [8]. This is because the model assumes that the velocity of water flow is uniform across the entire cross section and simplifies the terrain into a one-dimensional line [8]. The 2D-SWEs was optimised based on the 1D-SWEs, improving the limitations of the 1D-SWEs and better reflecting the complex situations in urban flood environments, such as nonlinear water flow paths [8]. In addition, the 2D-SWEs can now use high-precision digital terrain models (DTMs) obtained from Light Detection and Ranging (LIDAR), which can clearly represent small-scale structural elements and small terrain changes [8]. Furthermore, the use of high-performance computing

(HPC) technology can improve the computational efficiency of the 2D-SWEs [8]. Although the computational burden of 2D-SWEs can be reduced by using coarser grid and sub-grid models, these models are still relatively heavy to compute [8]. ML-based models are very effective when quick results are needed or when the model needs to be run with input parameters that change multiple times [8]. The input parameters can also come from different dimensions, such as climate, geography, and human factors, and can be simultaneously input into the model for training. When ML models are properly calibrated and tuned, ML method can produce prediction results comparable to the physical models [8, 9].

ML models have advantages in prediction accuracy and computational efficiency [8]. However, this method requires a large amount of data for training. Therefore, choosing to use machine learning as a modelling method for this project will mean that the collection of data in the early stages will need special attention. Methods for modelling and evaluating flood risk based on ML have become a mainstream direction in recent years, and these methods are often using Geographic Information Systems (GIS) to extract geographic information (such as slope, distance from rivers and altitude, etc.) from geospatial Data (such as satellite maps) [10, 11]. For those who wish to predict the impact of different spatial geographical locations on flood risk, ML based approaches have been proven to be effective and accurate [5, 10, 11].

This study aims to address the challenge of leveraging various geospatial data sources, flood maps and topographic information, in conjunction with daily precipitation data, to construct a ML model capable of assessing and predicting the likelihood of flooding occurrences in regions of the African country of Ghana. This model seeks to provide a comprehensive representation of the spatial distribution of flood probabilities close the Northern and eastern area of Ghana. Under appropriate circumstances, the predictive outcomes of this model might offer auxiliary guidance for flood mitigation strategies, such as the implementation of green infrastructure and the construction of flood barriers, thereby assisting stakeholders in alleviating flood-related risks. The four most critical objectives of this project are listed below:

- Objective 1 (Data Collection): This encompasses, but is not limited to, the gathering of historical flood data, precipitation data, and geospatial information in Ghana.
- Objective 2 (Data Processing): The focus will be on exploratory data analysis (EDA), the design and computation of specific data features such as maximum continuous rainfall, maximum number of continuous rainfall days during the floods period, and data visualisation such as historical flood interactive maps with targeted boundaries and sampling coordinates.
- Objective 3 (Model Construction): Experiments with various machine learning models for training and prediction with a consistent training strategy across models. Subsequent evaluations and performance comparisons of the trained models will be conducted.
- Objective 4 (Model Application): Utilising the saved best-performing model to execute flood predictions and evaluate the predictive performance. The probability distribution maps of flooding needs to be generated and consulted for the target regions.

For objectives 1 2 and 3, they are all elaborate on in Section 2 ”Methods”, and for objective 4, it is elaborate on in Section 3 ”Results”.

2 Proposed Methodology

In accordance with the objectives outlined in Section 1 ”Introduction”, a comprehensive workflow of the entire project can be delineated by the flow diagram depicted in Figure 1. At the project’s inception (corresponds to the Figure 1, Objective 1, ”start”), the identification of rele-

vant data sources is paramount, as all subsequent steps rely heavily on this foundation. The primary datasets used for this study include global historical flood data [12, 13, 14], daily precipitation records [15, 16, 17], and various geospatial datasets, such as the digital elevation model (DEM) [18], soil types [19], and land cover [20].

Upon securing these essential data sources (corresponds to the Figure 1, Objective 1, "Confirm Data Source"), the initial task involves filtering out flood events specific to Ghana from the global flood data, given that Ghana is the focal region of this project. Following this, the creation of an interactive flood map (corresponds to the Figure 1, Objective 2, "Build Interactive Flood Maps to Visualise"), tailored to meet visualisation demands, is indispensable. This map helps in pinpointing the precise geographic regions that will be utilised for model training and deployment, granting a more intuitive understanding of the flood data's visual representation.

Once these regions are demarcated (corresponds to the Figure 1, Objective 2, "Determine Regions used for modelling and model application"), it necessitates a re-indexing of the historical flood data to extract flood events exclusive to our defined target areas. Subsequently, the datasets for model training and application can be curated. For the training dataset, stratified sampling of geographic coordinates (longitude and latitude) is employed (corresponds to the Figure 1, Objective 2, "Stratified Random Sampling"), while for the application dataset, a uniform sampling strategy is adopted (corresponds to the Figure 1, Objective 2, "Uniform Sampling"). With these coordinates in hand, integration with geospatial data in a Geographic Information System (GIS) environment allows for the extraction of additional geographical features, such as elevation, slope, and soil type (corresponds to the Figure 1, Objective 2, "Geographic Information System Data Processing and Acquisition for each Coordinate Point"). Concurrently, these coordinates are assigned new 'Flood ID' identifiers, and features like 'Duration', 'Max Duration', 'Event Start Date', and 'Event End Date' are computed for each point during each historical flood event, paving the way for the acquisition of daily rainfall data (corresponds to the Figure 1, Objective 2, "For each Coordinate Point"). Upon completion of these calculations, daily rainfall data during the flood events can be procured. This data, in turn, facilitates the computation of various statistical features, including mean, median, the maximum number of continuous rainy days, and maximum continuous rainfall, among others. By this juncture, the data preparation phase, aligning with objectives 1 and 2, is essentially complete. The subsequent step involves merging the geospatial data with the rainfall dataset, resulting in distinct datasets for model training and application.

Figure 1, beginning of objective 3: Prior to the construction of machine learning models, the data typically undergoes several preprocessing tasks. These tasks often encompass the removal of irrelevant columns, handling missing values, encoding essential categorical variables (e.g., one-hot encoding [21]), and slicing the data into training, validation, and test subsets, in addition to setting a consistent model training strategy. The next phase involves selecting and training models. For classification tasks, potential candidates include Logistic Regression [22], Random Forest Classifier [1, 2], Support Vector Machines (SVM) [23], XGBoost [24, 2], Gradient Boosting Classifier [25, 2], and Neural Networks [26] etc. Each of these models is trained, evaluated, and compared. The optimal model, based on performance metrics, is then employed to make predictions on the application dataset. The outcomes are then visualised via a flood probability map, and the model's predictive performance is thoroughly assessed.

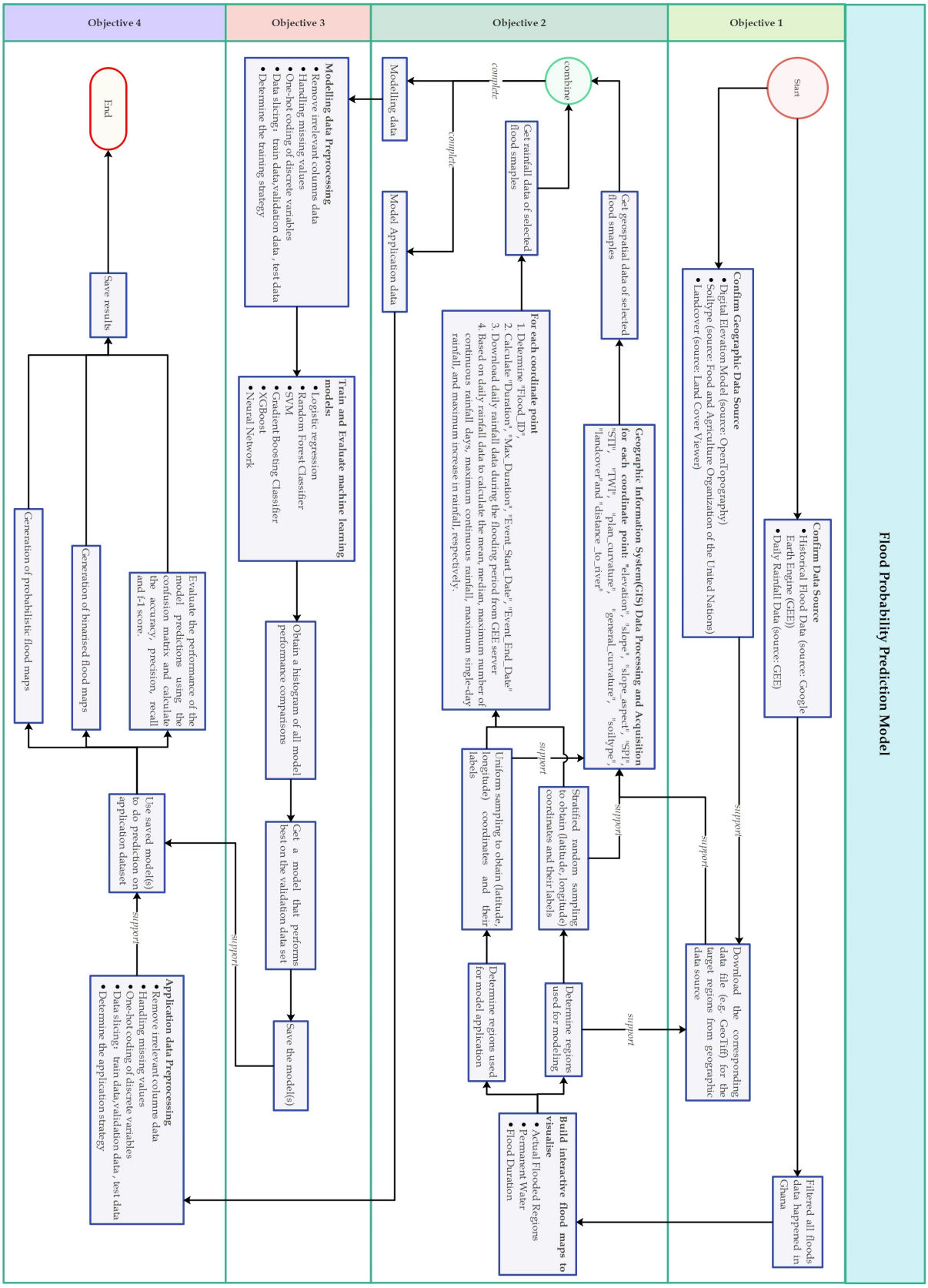

Figure 1: Flow Diagram of the Proposed Methodology.

2.1 Data Sources and Interpretation

The purpose of this section is to elucidate and visually present the data sources implicated in the project, alongside an exploration of their utilisation and processing methodologies. And for ease of exposition, the data sources utilised:

- Data Source 1: Global Flood Database [12, 13, 14].
- Data Source 2: Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) [15, 16, 17].
- Data Source 3: ALOS Global Digital Surface Model (AW3D30) [18].
- Data Source 4: Digital Soil Map of the World [19].
- Data Source 5: Land Cover Viewer [20].

Primarily, the Global Flood Database is a collaborative venture between Cloud to Street and the Dartmouth Flood Observatory (DFO) [12, 27]. This Global Flood Database product has been financially backed by the Google Earth Outreach Program and has been made open-source within the Google Earth Engine framework [12, 13, 14]. The database encapsulates 913 significant flood incidents from around the globe, spanning the period of 2000-2018 [12, 13, 14]. Each flood incident is represented as a distinct flood map. Within each map, multiple event features are embedded, such as “id”, “countries”, “system:index”, among others [12, 13]. The “countries” feature can act as a filter to pinpoint specific flood events happened on that given country or countries, while the “system:index” feature column encompasses the distinct commencement dates of each flood incident, serving as an essential precursor for deriving daily precipitation data [12, 13].

These flood maps are generated with a resolution of 250 meters, leveraging satellites from NASA’s Terra and Aqua MODIS fleet [12, 13, 14]. For each flood event, the database employs disparate bands to delineate various flood-related information [13]. The foremost is the “flooded” band, primarily illustrating the maximum inundation extent during the event [13]. Under this specific band, each pixel within a map undergoes a binary classification, denoted as either “water” or “non-water” [13]. Complementary bands include ”jrc permanent water” and “duration”. The former re-samples the 30-meter JRC global surface water dataset [28] to a 250-meter resolution, serving as a binary descriptor for whether a pixel within the flood event’s extent represents a permanent water body [13]. The latter provides insights into the duration, in days, for which the pixels identified as “water” under the “flooded” band persisted [13].

Moreover, additional bands indicative of data quality have been incorporated, elucidating cloud conditions during the event, exemplified by “clear views” and “clear percent” [13]. To enhance the presentation and elucidation of the data, an interactive flood map has been adeptly developed for this project. In Figure 2, a flood incident occurring between August 10, 2003, and October 19, 2003, (id=2320) [12, 13] serves as a visualisation exemplar, displayed using the interactive flood map.

The map in Figure 2 consists of three sub-maps, aligned from left to right, which represent the actual inundated area during the flood event, ”jrc permanent water” [12, 13], and ”duration” [12, 13] respectively. In Figure 2, the grey regions of all three sub-maps denote the maximum affected extent of the flood event. Notably, the actual inundated area of the flood is defined as those pixels which are marked as True in the ”flooded” band and simultaneously as False in the ”jrc permanent water” band. This definition is primarily adopted to ensure that, during the subsequent acquisition of flood coordinates, one can directly obtain the coordinates of areas inundated by floods whilst simultaneously ascertaining their exclusion from permanent water bodies. With the actual inundated flood area clearly defined, it then becomes feasible to perform

stratified random sampling within a predetermined geographical area, specifically sampling within the actual inundated area when it equals True and False. For instance, in the Left sub-map of Figure 2, random sampling should be conducted on both the red (where the actual inundated are equals True) and grey(where the actual inundated area equals False) pixel points. This methodology effectively confirms the coordinate points of the data used for modelling. Within this project, the regions for modelling and model application are visually represented in Figure 3, wherein the two larger regions are designated for modelling (modelling ROI1 and ROI2), while the smaller regions within them are reserved for model application (application sub-ROI1 and sub-ROI2). The model application involves the prediction and generation of a flood map. Here, the acronym "ROI" stands for "region of interest", indicating the areas of focus.

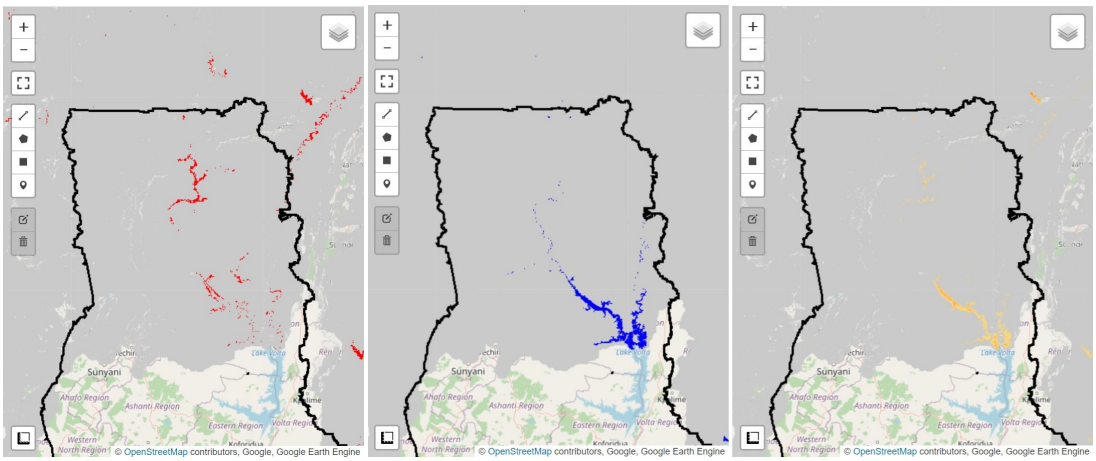

Figure 2: Visualisation of Flood Event (id=2320). (Left) Actual Inundation Area (Red Pixels); (Middle) Permanent Water Sources (Blue Pixels); (Right) Flood Duration in Days (Yellow Pixels, Darker indicates longer days).

During the project's progression, to ascertain the accuracy of point selection, it becomes indispensable to visually verify both flood and non-flood points within the targeted ROI region. This is foundational, as ensuring the model is trained on accurate samples is the initial step to its success. Consequently, in Figure 4, the flood event that transpired between August 10 and October 19, 2003 (id=2320) [12, 13] is once again utilised as a visual example to illustrate the specific sampling scenario within the Modelling ROI1 region. In Figure 4, the blue coordinate points represent flood points randomly selected from the actual inundated area, while the yellow points on the map denote the non-flood point selections. Evidently, as anticipated, the flood points are randomly sampled from the actual flood areas, while the non-flood points are sampled from the non-inundated areas.

Subsequently, the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) [15, 16] warrants mention as the source of daily precipitation data within this project. CHIRPS is a quasi-global precipitation dataset spanning 42 years, from 1981 to 2023 [15, 16, 17]. It encompasses data ranging from 50°S to 50°N latitude across all longitudes [15, 16, 17]. By amalgamating 0.05° resolution satellite imagery (precisely 5566 meters) with local station data, it crafts a gridded rainfall time series [15, 16, 17]. The CHIRPS dataset has also been made open-source within the Google Earth Engine framework [16]. Within this dataset, there exists a singular band denoted as "precipitation", and acquiring this band directly furnishes the requisite rainfall data [16]. Compared to pure satellite imagery precipitation datasets such as "Global Precipitation Measurement (GPM) v6" [29, 30], "WAPOR Actual Evapotranspiration and Interception 2.0" [31, 32, 33] and "WorldClim BIO Variables V1" [34, 35] these solely satellite-derived datasets are often vulnerable to uncontrollable factors, such as atmospheric conditions, cloud

cover, and sensor malfunctions, leading to potential omissions of rainfall data for specific periods. However, CHIRPS mitigates this concern by integrating with local meteorological stations, maximally preventing data omissions due to these uncontrollable factors.

Moreover, a note on the geospatial data sources utilised in this project is warranted. The ALOS Global Digital Surface Model (AW3D30) [18] was employed as the primary source for obtaining DEM data. This global dataset, generated from imagery collected between 2006 and 2011 by the Advanced Land Observing Satellite (ALOS) and its Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM), is maintained and managed by the Japan Aerospace Exploration Agency (JAXA) [18]. This dataset stands out as the most precise elevation data on a global scale, with its elevation accuracy for the 30-metre grid version being world-leading [18].

Additionally, the Digital Soil Map of the World, developed by the Food and Agriculture Organisation of the United Nations (FAO), was used as the primary source for procuring land type data for the target region [19]. The Digital Soil Map of the World is computed and stored using a geographical projection method (latitude-longitude) intersected with a template containing water-related features such as coastlines, lakes, glaciers, and double-line rivers, all at a scale of 1:5,000,000 [19]. The requisite data files can be readily accessed and downloaded from the FAO's official data repository [19].

Lastly, the project also engaged the Land Cover Viewer as the platform for obtaining Land Cover data. This platform is a fully open-source data acquisition platform, offering a clear and straightforward user interface [20]. Users, through the Land Cover Viewer platform, can swiftly acquire a Land Cover Map with a resolution of 100 metres for their target country or region [20]. The primary input for these maps is the PROBA-V satellite observation data, which is then organised into millions of Sentinel-2 equivalent tiles of 110x110km [20]. These data are subsequently processed in both a tiling grid and Universal Transverse Mercator (UTM) projection to produce the final 100-metre resolution Land Cover Map [20]. The advantage of this processing methodology lies in ensuring high quality of the map and facilitating continuity with Sentinel-2 observations [20].

The meticulous choice of geospatial data sources, combined with rigorous processing and computation using advanced GIS tools, ensures the extraction of pertinent data features such as “slope”, “slope aspect”, ”plan curvature”, ”soil type” and ”land cover”, etc. This approach, grounded in precision and accuracy, lays a solid foundation for subsequent analyses and investigations in the project [5, 8, 10, 11].

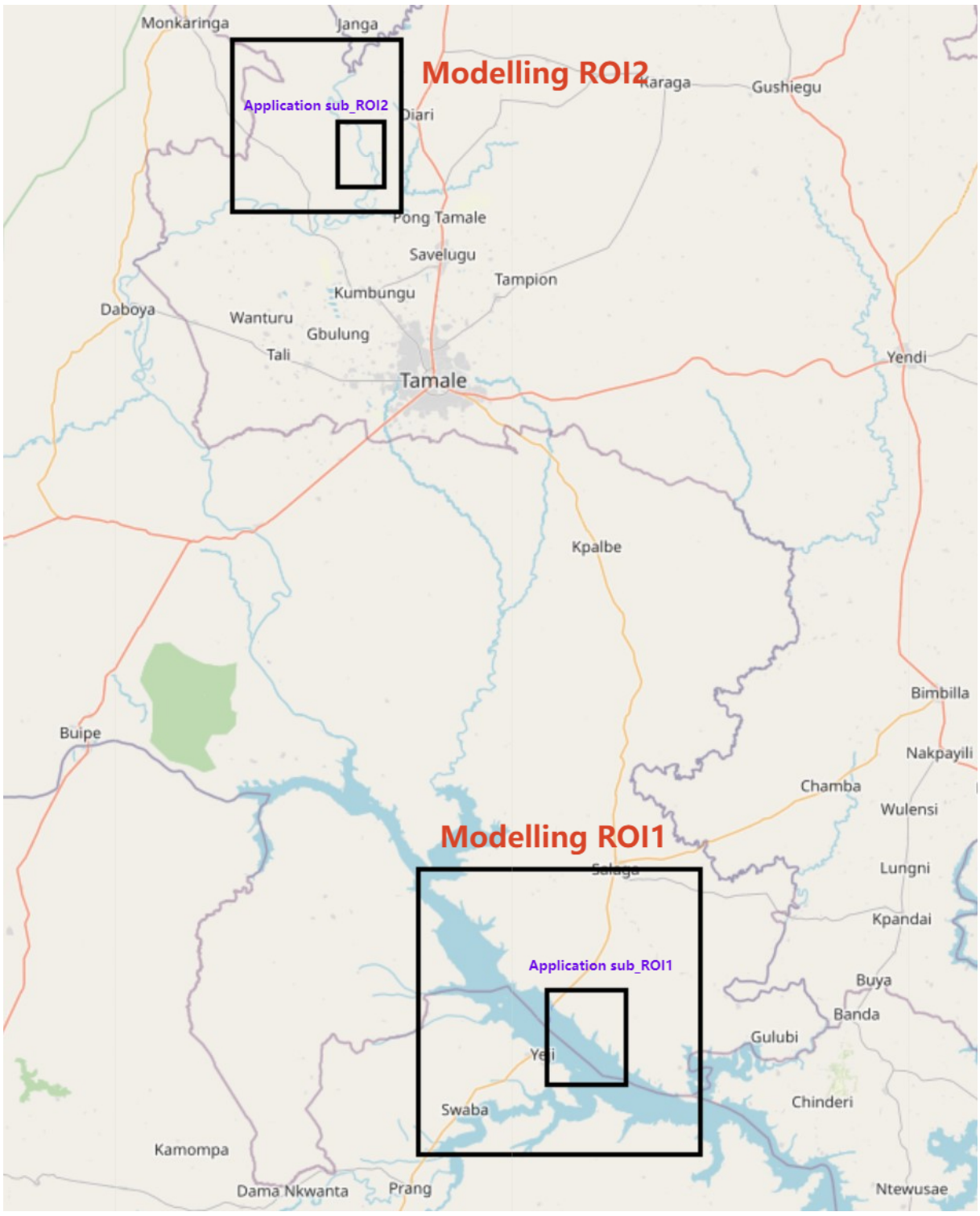

Figure 3: Geographical Representation of Modelling and Application ROI (region of interest) Areas. Legend: Large Rectangles = Modelling Areas; Small Inner Rectangles = Application Areas.

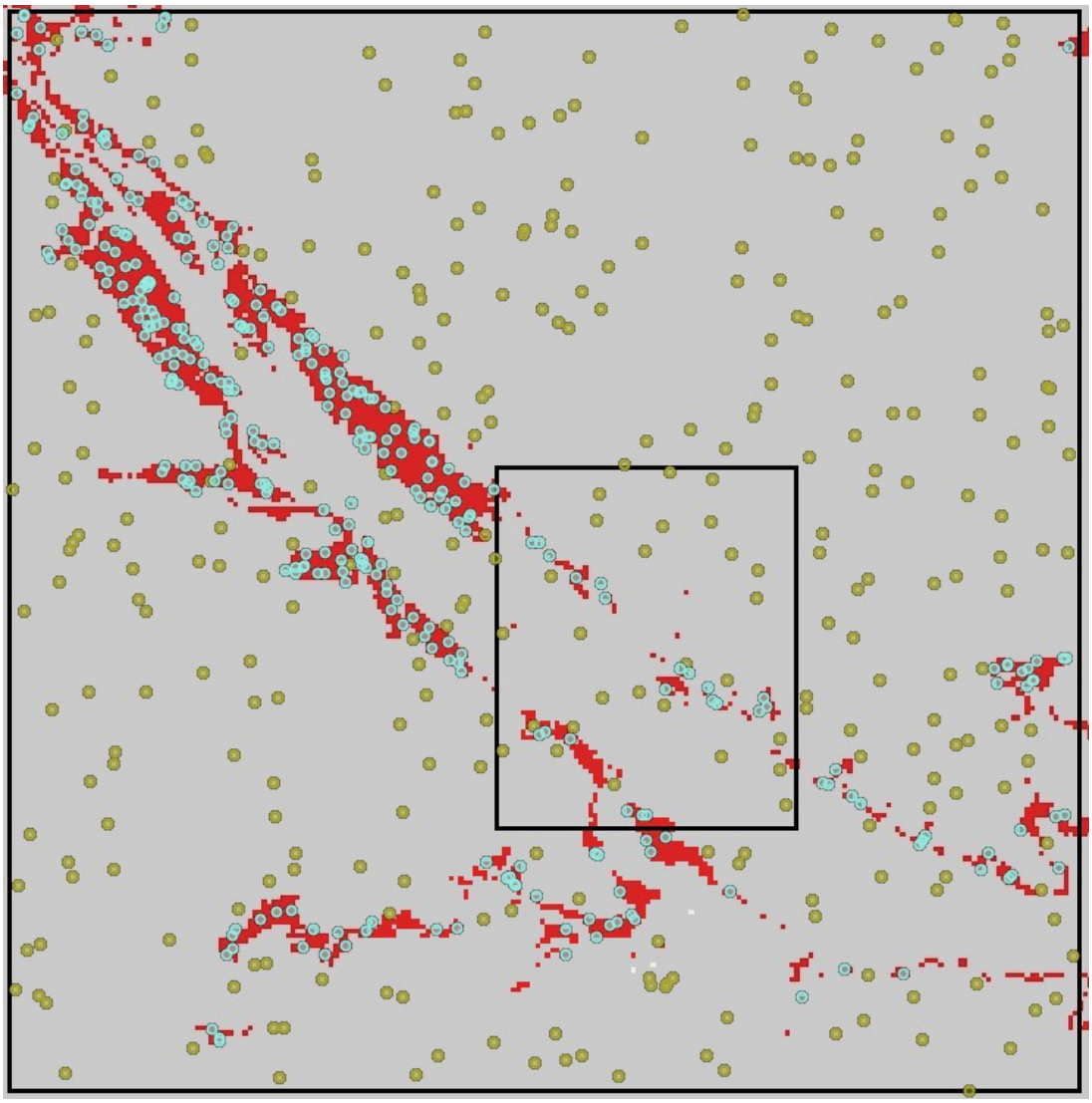

Figure 4: Stratified Random Sampling in Modelling ROI1 for Flood Event (id=2320). Legend: Blue Points = Flood Samples; Yellow Points = Non-flood Samples; Red Pixels = Actual Inundation.

2.2 Modelling

As delineated in Figure 1 under Objective 3, the process can essentially be restructured into three primary sub-phases: the data preprocessing phase (corresponding to ”modelling data preprocessing” in Figure 1), the model training and evaluation phase (corresponds to ”train and evaluate machine learning models” and its subsequent ”obtain histogram of model performances” in Figure 1), and the saving final model phase.

During the data preprocessing phase, besides executing an array of preprocessing actions (such as addressing missing values, discarding irrelevant data columns, and implementing one-hot encoding on specific features, as mentioned in Figure 1), establishing an appropriate training strategy emerges as a quintessential step. This is because varying training strategies often dictate the quantity of models to be preserved subsequently and their respective evaluation methodologies. This project has adopted the ”Leave-One-Out Cross Validation” [36] strategy as training strategies. The details of this approach within the context of this project are expounded upon in Figure 5. Prior to comprehending this strategy, it’s imperative to grasp

the origin of the ”Flood ID” identifier in the dataset. According to the regions of interest delineated in Figure 3 (ROI1 and ROI2), a total of six flood events, which had transpired in these two regions, can be extracted from the global flood database [12, 13, 14]. Their respective IDs in the database are: 2320, 3166, 3534, 3663, 3747, 4683. Of these, flood events 2320, 3166, and 4683 manifested in both regions (ROI1 and ROI2), implying that their impact was extensive, covering both regions. However, due to the inherent independence between regions, this project treated the six flood events as nine distinct events. This elucidates the rationale behind the assignment of new ”Flood ID” identifiers post data coordinate sampling in Objective 2 of Figure 1. This distinction is pivotal to differentiate instances where a single flood event had repercussions in multiple regions within the global flood database [12, 13, 14]. In Figure 5, the nine flood datasets merged to form a comprehensive modelling dataset. The main idea of the Leave-One-Out Cross Validation strategy is the iterative assignment of one of the nine flood datasets as the test set, whilst the remaining eight are shuffled and partitioned into training and validation subsets (e.g., ”80%” to ”90%” training data and ”20%” to ”10%” validation data. This kind of ratio splitting is more common in the field of machine learning data preprocessing [37]). Contrasted against conventional machine learning training approaches, this strategy proffers two potential advantages: 1. It aligns more congruently with real-world application scenarios. 2. It ensures a more holistic model evaluation. For the first advantage, the practical utility of a flood model often hinges on its prowess to accurately forecast future, yet-to-occur floods. Given that data on future floods remains perpetually elusive, this strategy attempts to circumvent this limitation by simulating past flood events as potential future occurrences. While this simulation may not epitomise rigorousness, it guarantees that the model is evaluated using a complete set of flood data, which it has not encountered during training. This bolsters the rigour and real-world relevance of the evaluation. Concerning the second advantage, if a model is solely trained on a single random partition of the nine flood datasets, its performance would be contingent on this particular partition, rendering conventional training strategies somewhat restrictive. However, the iterative partitioning inherent to Leave-One-Out Cross Validation mitigates this limitation, as the collective performance metrics over nine iterations better represent the model's competence.

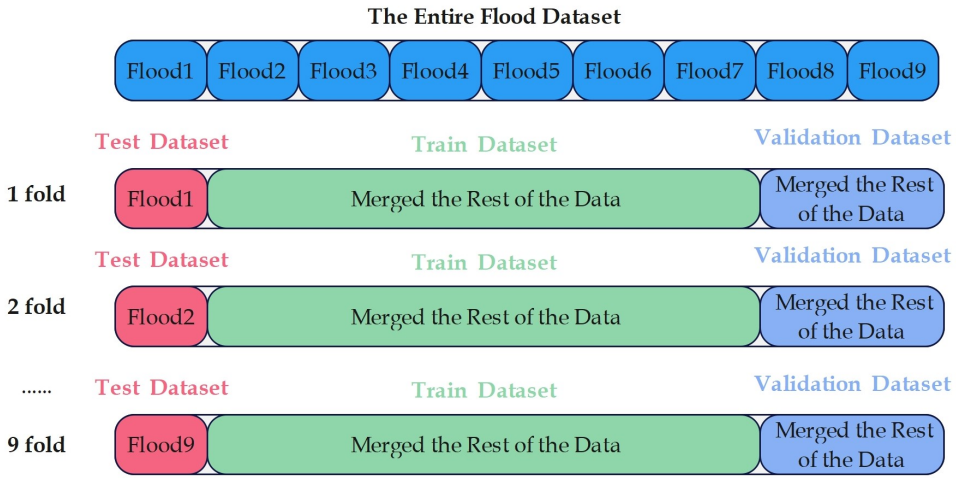

Figure 5: Illustration of Leave-One-Out Cross Validation Strategy. Legend: Test Dataset = Final Model Testing Dataset; Train/Validation Dataset = Model Training and Validation Dataset.

In the model training and evaluation phase, as alluded to in Figure 1, this project experimented with six distinct machine learning models. They were all subjected to identical data for training, evaluation, and testing, ensuring parity in performance assessments. The project utilised metrics such as accuracy, precision, recall, and F-1 score to furnish a comprehensive evaluation of

model performance [38]. These metrics can be derived from a binary classification confusion matrix. Typically, in binary classification tasks, post-classification, each data instance can be categorised into one of four scenarios: True Positives (TP, The number of cases where the true value is POSITIVE and the model considers it to be POSITIVE), False Negatives (FN, The number of cases where the true value is POSITIVE and the model considers it to be NEGATIVE), True Negatives (TN, The number of cases where the true value is NEGATIVE and the model considers it to be NEGATIVE), and False Positives (FP, The number of cases where the true value is NEGATIVE and the model considers it to be POSITIVE) [38]. The binary confusion matrix (CM) [38] is thus formulated as: $CM = \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$ Subsequently, the metrics can be defined as:

- **Accuracy:**

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:**

$$\frac{TP}{TP + FP}$$

- **Recall:**

$$\frac{TP}{TP + FN}$$

- **F1-score:**

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Drawing insights from the model performance exhibited in Figure 6, it's evident that XGBoost, Random Forest, Neural Net, and Gradient Boosting Classifier manifest commendable efficacy among the six models. In pursuit of the optimal model, the Random Forest model's performance on the validation set emerges as superior, with the only exceptions being a slightly subpar precision compared to the XGBoost model and a lower recall than the SVM. However, the SVM model, despite its stellar performance in recall on both validation and test sets, falters in other metrics, revealing its inherent limitations for this project's objectives. Furthermore, the Random Forest model's performance on the test set remains commendable, ranking second among all six models. However, considering the adoption of the Leave-One-Out Cross Validation strategy in this project, which treats test data as a simulation of a future flood event, the primary consideration when selecting the optimal model should be its performance on the validation set rather than the test set [36]. Thus, considering the aforementioned factors, electing the Random Forest as the final model for application emerges as a judicious choice.

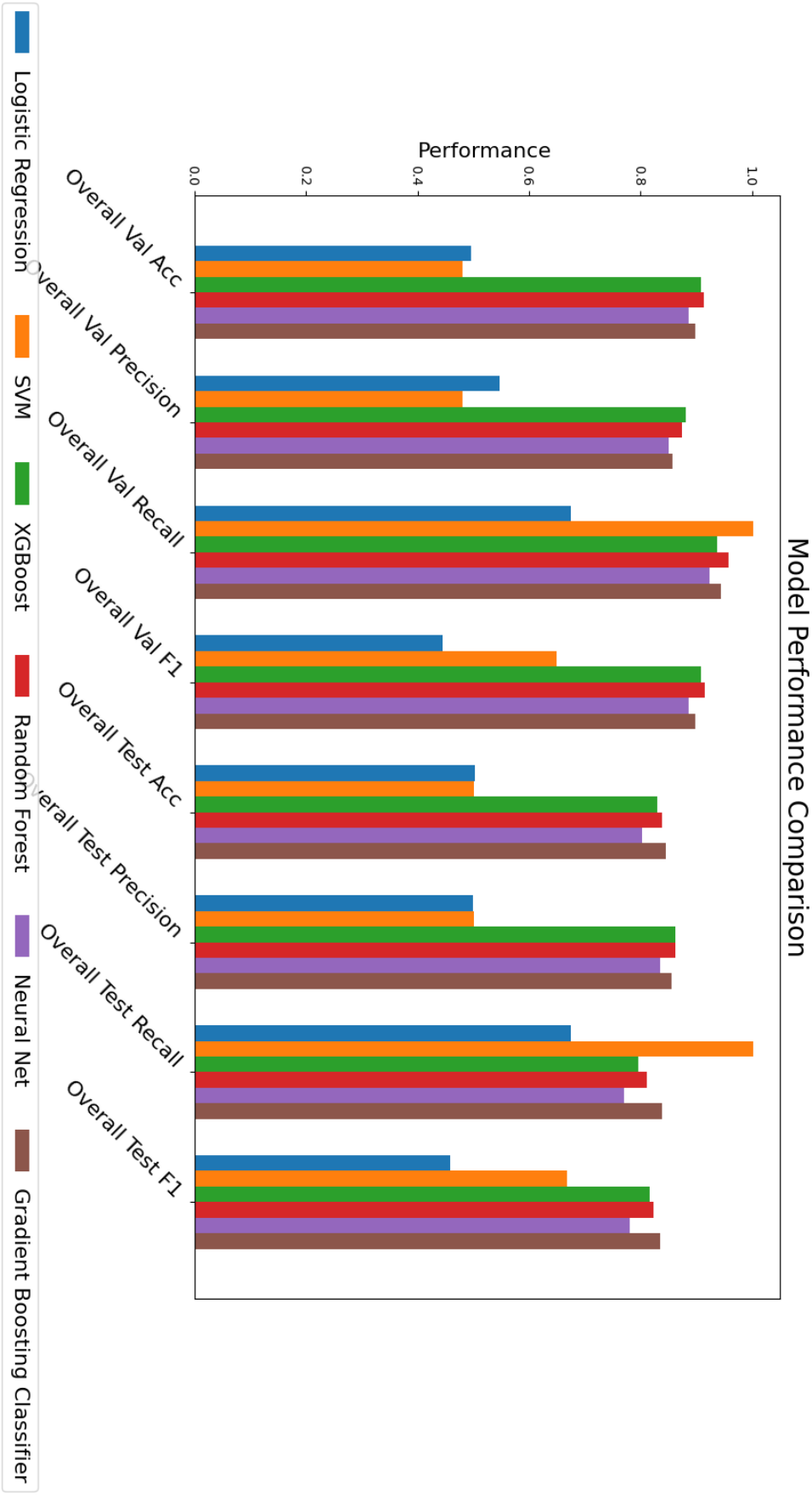

Figure 6: Comparative Analysis of Model Prediction Performances

3 Results

The objective of this section is to articulate the outcomes of the applied model on real-world regions (sub-ROI1 and sub-ROI2, as shown in Figure 3). This aligns with Objective 4 in Figure 1. A suite of nine distinct models have employed, all of which utilise the Random Forest machine learning algorithm. These models were constructed using a unique training strategy (see Figure 5). In the following subsections, the visualised outcomes are presented, followed by qualitative and quantitative assessments of the model's performance.

3.1 Visualisation of Flood Maps

Figures 7 and 8 present the actual and model-predicted binary flood maps, respectively, for nine historical flood events. These visualisations draw upon a global flood database with a pixel resolution of 250 meters. Consequently, each pixel corresponds to a geographical area of 250m×250m. Flooded areas are distinctly marked with red pixels. An initial qualitative analysis of Figures 7 and 8 provides a preliminary insight into the model's capability to accurately predict flood zones.

3.2 Quantitative Assessment

Although visual evaluations offer some level of insight, they lack the rigor of statistical analysis. Therefore, Figure 9 showcases a comprehensive statistical evaluation of the models' predictive performance, incorporating confusion matrices and pertinent binary classification metrics. A detailed analysis reveals that most models exhibit robust "recall" scores, albeit accompanied by less impressive "precision" metrics. As delineated in Section 2.2 "Modelling" this suggests that the model's predictions are generally inclusive of the actual flood-affected coordinates, albeit at the cost of a higher rate of false positives.

3.2.1 Regional Variation in Model Precision

A comparative analysis between the initial six and the final three confusion matrices (Figure 9) indicates a discernible improvement in precision for sub-ROI2 over sub-ROI1. This observation is further corroborated by a visual assessment of Figures 7 and 8.

3.3 Probabilistic Flood Risk Assessment

To better align with the intricacies of real-world flood management, it is imperative to understand the probabilistic nature of flood events. Figures 10 illustrate the model's flood probability predictions for the regions under study. These maps reveal a higher susceptibility to flooding in areas proximate to river boundaries, corroborating conventional understanding of flood dynamics [4, 5].

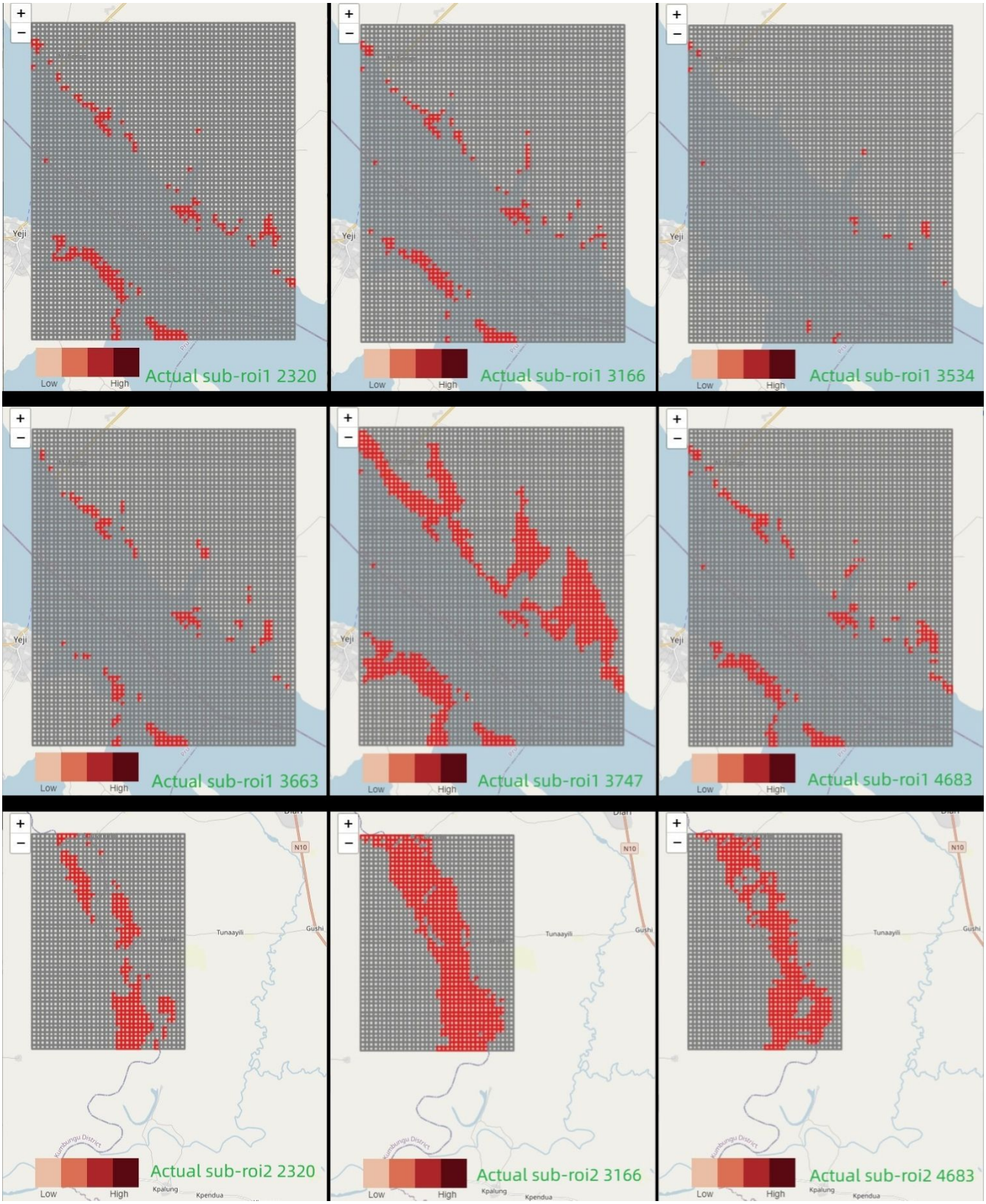

Figure 7: Binary Flood Maps from Global Flood Database [13], Identified by Unique Flood IDs (Geographic Area + Event ID). Legend: Geographic Area Correspond to Application Areas in Figure 3; Red Pixels = Actual Inundation. Pixel Resolution: 250 meters.

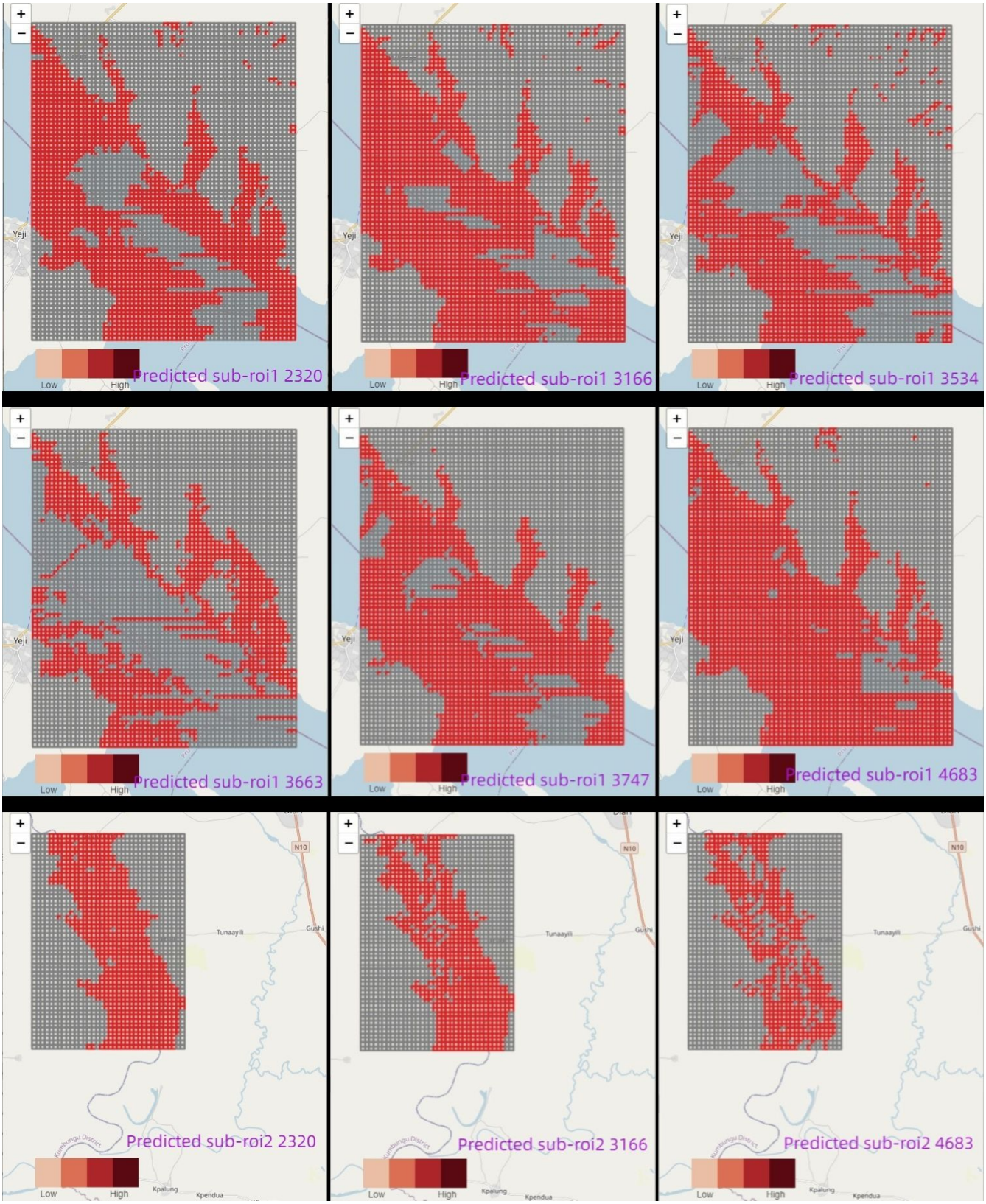

Figure 8: Predictive Binary Flood Maps, Identified by Unique Flood IDs (Geographic Area + Event ID from Database [13]). Legend: Geographic Area Correspond to Application Areas in Figure 3; Red Pixels = Predicted Inundation. Pixel Resolution: 250 meters.

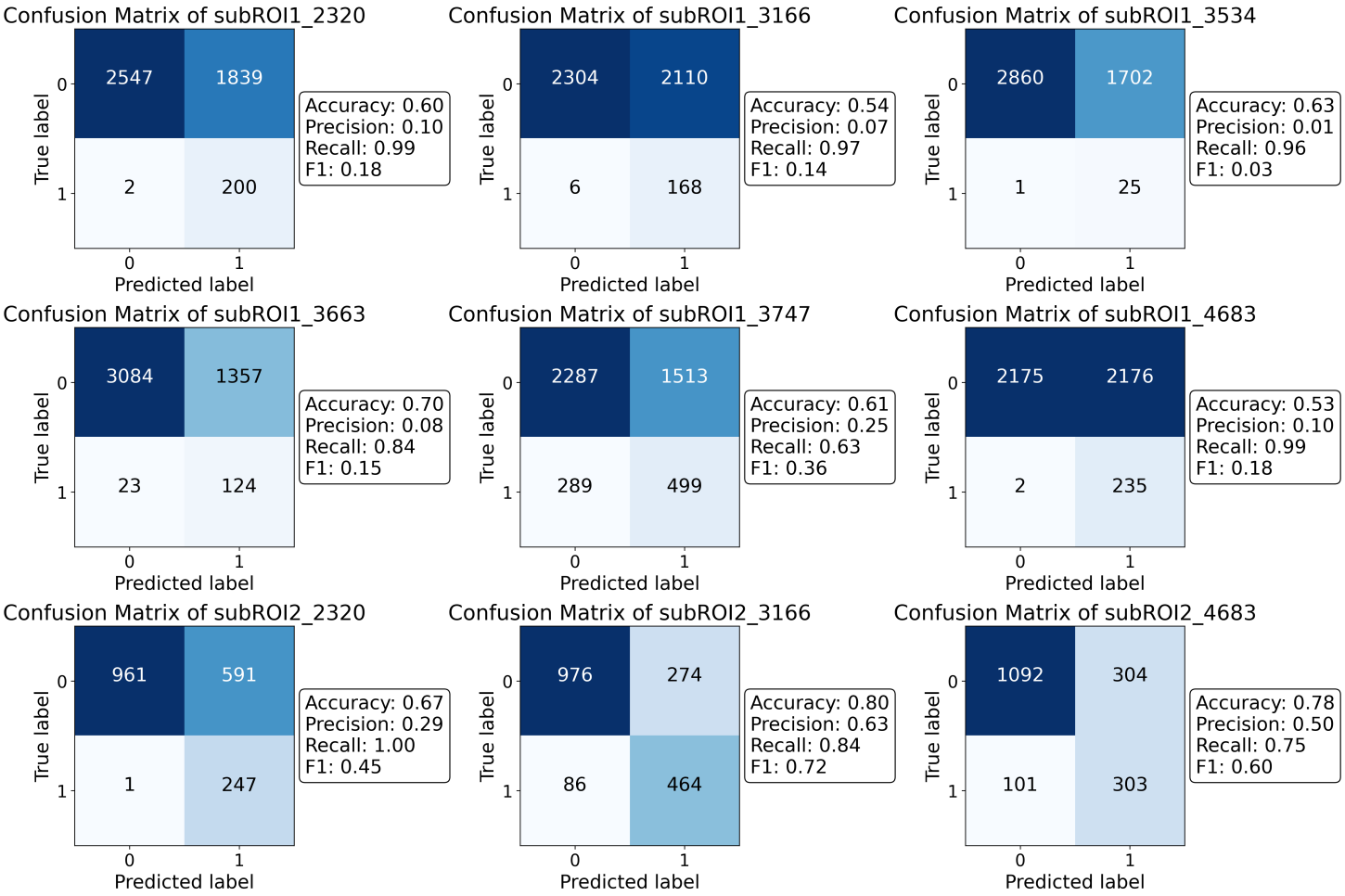

Figure 9: Confusion Matrices for Binary Flood Prediction, Identified by Unique Flood IDs (Geographic Area + Event ID from Database [13]). Legend: Geographic Area Correspond to Application Areas in Figure 3.

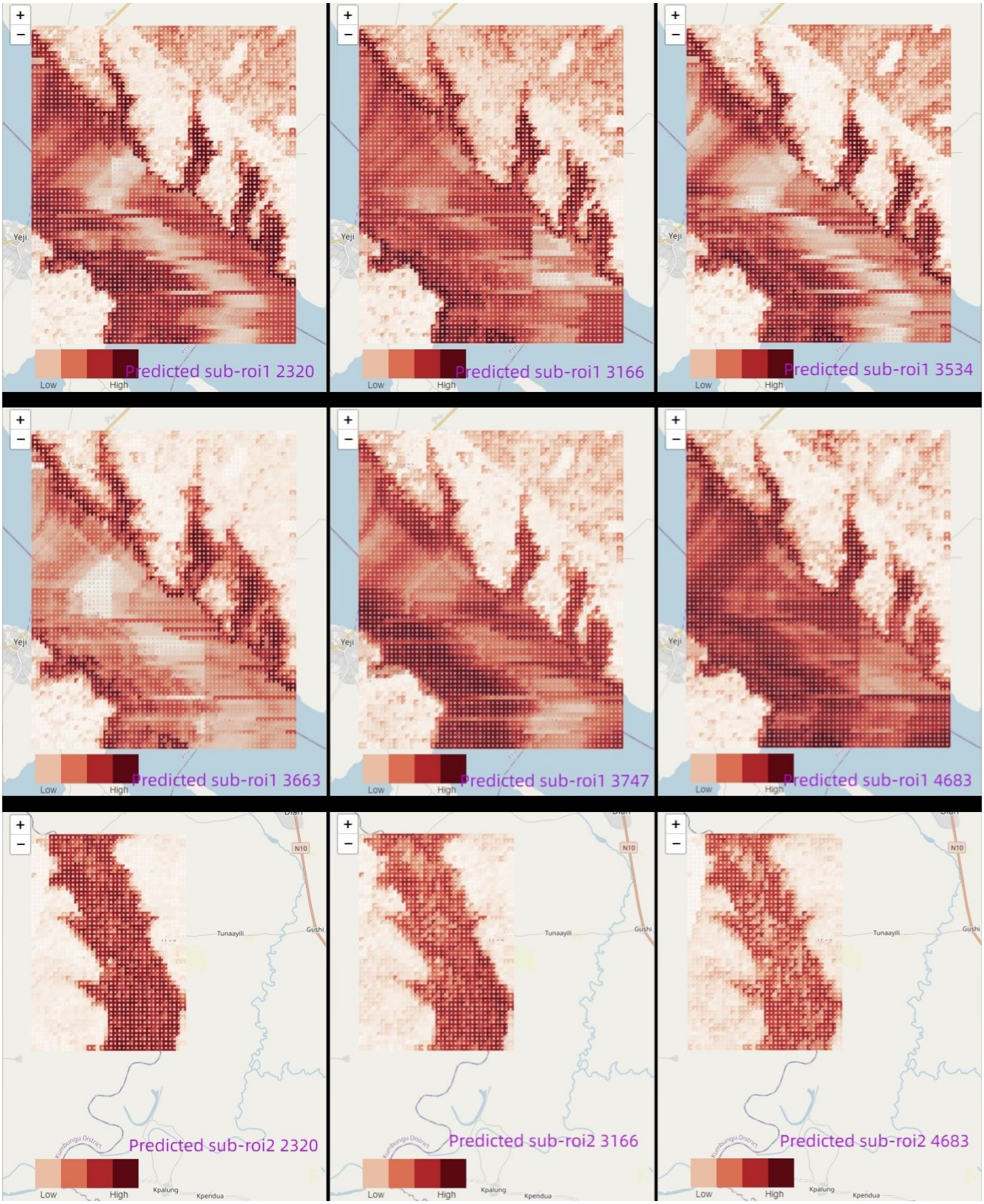

Figure 10: Probabilistic Flood Maps from Model Prediction, Identified by Unique Flood IDs (Geographic Area + Event ID from Database [13]). Legend: Regions Correspond to Application Areas in Figure 3; Red Pixels = Predicted Inundation (Darker = Higher Probability). Pixel Resolution: 250 meters.

4 Discussion

This section aims to elucidate the implications and limitations of the model's predictive performance as highlighted in Section 3, "Results". It also seeks to identify potential areas for methodological improvements. The discussion is predicated on the foundational assumption that the actual flood data in Figure 7 is devoid of errors and that similar flood events are likely to recur.

4.1 Model Strengths and Comparison with Previous Work

A noteworthy advantage of our model is its high 'recall' metric, which indicates its proficiency in identifying high-risk flood areas. This strength could facilitate significant advancements in flood risk mitigation if preemptive measures are applied based on the model's forecasts. Compared to other machine learning-based flood prediction projects, this study uniquely combines geospatial and rainfall data for model training. In contrast, most prior research tends to utilise either of the two [8, 9, 10, 11]. While focusing on a single data type may alleviate the burden during the data preparation phase and improve model efficiency, it may also limit the model's sensitivity to various data types, rendering the predictions more generalised and less accurate in real-world applications.

4.2 Limitations in Precision and Resource Allocation

Despite its strengths, the model's subpar 'precision' performance represents a significant limitation. The elevated rate of false positives could lead to resource misallocation and economic inefficiencies if preparations are executed based on these imprecise forecasts. Therefore, future research must aim to improve the model's precision without compromising its recall performance.

4.2.1 Spatial Variability in Predictive Accuracy

The model's inconsistent performance across different sub-regions (sub-ROI1 and sub-ROI2, in Figure 3) necessitates further investigation. Preliminary analyses suggest that this discrepancy may be attributed to the granularity introduced by resampling the permanent water source from a 30m to a 250m resolution. Further empirical studies are essential to validate this hypothesis.

4.3 Operational and Scalability Challenges

The current data acquisition system, which employs batch-processing for collecting rainfall data, poses logistical challenges. Network latency and server congestion are among the variables that introduce inefficiencies, thereby hindering the model's broader applicability. Future iterations of this model should focus on streamlining the data acquisition process to overcome these bottlenecks.

4.4 Constraints in Real-time Prediction and Potential Solutions

The model's reliance on historical flood data for training poses a significant limitation for real-time prediction. The historical database provides temporal markers like 'start time' and 'duration' of past floods, facilitating efficient retrieval of corresponding rainfall data. While this approach is advantageous for training and validating the model using past events, it introduces significant constraints when attempting to generalise the model for real-time or future flood prediction scenarios.

In practice, precise timing for imminent floods is may rarely available, limiting the model's utility in dynamic, real-world contexts. Even if some temporal data exists, such as short-term weather forecasts, its granularity and accuracy possibly fall short for precise prediction. Therefore, relying solely on historical flood data and associated temporal markers limits the model's ability to adapt to new, unobserved conditions and decreases its efficacy as a real-time flood predictive tool.

Incorporating a predictive rainfall model probably offers a solution by enabling real-time or future rainfall pattern prediction, eliminating the need for historical temporal markers. This enhances the model's real-world robustness and facilitates timely, targeted preemptive measures. However, it should be noted that integrating a predictive rainfall model would also likely introduce a new set of complexities, including the need for real-time data collection and processing, as well as potential challenges in model calibration and validation. These additional considerations warrant in-depth future research to optimise the model for real-time applicability.

4.5 Challenges in Probabilistic Assessment

Validating the probabilistic flood maps generated by the model, as illustrated in Figure 10, is a non-trivial task due to the absence of ground-truth probability maps. Future research should prioritise the development and validation of these ground-truth maps to refine the model's predictive capabilities.

5 Conclusion

Through meticulous data retrieval, computation, analysis, processing, and amalgamation from various public data sources, this project has prepared a flood dataset suitable for machine learning modelling for select regions in Ghana (Figure 3). After contrasting the performances of multiple machine learning models on this dataset, the Random Forest model was designated as the final applicative model. Based on its applicative performance, the following deduction was made: Despite the model's evident lack of precision during predictions (corresponding to a low binary classification precision metric), it exhibits a propensity to be sensitive in most scenarios when predicting potential flood points (corresponding to a high binary classification recall metric). Thus, this provides guiding recommendations and assistance in realistically narrowing down potential flood-affected areas.

References

[1] M. Pal. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.*, 26:217–222, 2005.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2021.

[3] S. K. M. Agblorti D. Babanawo, P. A. D. Mattah, E. K. Brempong, M. M. Mattah, and D. W. Aheto. Local indicator-based flood vulnerability indices and predictors of relocation in the ketu south municipal area of ghana. *Sustainability*, 2022.

[4] J. Abazaami C. Anab and S. A. Achanso. Moving beyond the ad hoc responses in flood management to a localization approach in ghana. *Ghana Journal of Development Studies*, 2022.

[5] Y. A. Twumasi and R. Asomani-Boateng. Mapping seasonal hazards for flood management in accra, ghana using gis. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 5, pages 2874–2876, 2002.

[6] S. A. Adinku. Disaster preparedness: A sociological study of the flood problem in the odaw catchment in accra. *Google Statistics*, 1994.

[7] Y. A. Twumasi and R. Asomani-Boateng. Planning and managing urban organic solid waste in an african city: linking organic solid waste composting to urban cultivation in accra, ghana. *Google Statistics*, 1999.

[8] L. Cea and P. Costabile. Flood risk in urban areas: Modelling, management and adaptation to climate change. a review. *Hydrology*, 9, 2022.

[9] B. Jamali, E. Haghighat, A. Ignjatovic, J. P. Leitão, and A. Deletic. Machine learning for accelerating 2d flood models: Potential and challenges. *Hydrological Processes*, 35, 2021.

[10] Z. Wang, X. Chen, Z. Qi, and et al. Flood sensitivity assessment of super cities. *Scientific Reports*, 13:5582, 2023.

[11] W. Fenglin, I. Ahmad, M. Zelenakova, and et al. Exploratory regression modeling for flood susceptibility mapping in the gis environment. *Scientific Reports*, 13:247, 2023.

[12] Global flood database, 2023.

[13] Global flood database v1 (2000-2018), 2023.

[14] B. Tellman, J. A. Sullivan, C. Kuhn, A. J. Kettner, C. S. Doyle, G. R. Brakenridge, T. Erickson, and D. A. Slayback. Satellites observe increasing proportion of population exposed to floods. *Nature*, 2021.

[15] Chirps: Rainfall estimates from rain gauge and satellite observations, 2023.

[16] Chirps daily: Climate hazards group infrared precipitation with station data (version 2.0 final), 2023.

[17] C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen. The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. *Sci. Data*, 2:150066, 2015.

[18] Alos world 3d-30m, 2023.

[19] Digital soil map of the world, 2021.

[20] Land cover viewer - ghana 2019, 2019.

[21] S. C. Bagui S. Bagui, D. Nandi and R. J. White. Machine learning and deep learning for phishing email classification using one-hot encoding. *Journal of Computer Science*, 2021.

[22] D. W. Hosmer and S. Lemeshow. Introduction to the logistic regression model. In *Applied Logistic Regression, Second Edition*, 2000.

[23] W. S. Noble. What is a support vector machine. *Nature Biotechnology*, 24:1565–1567, 2006.

[24] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[25] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[26] S. S. Haykin. Introduction. In *Neural Networks: A Comprehensive Foundation*, 1998.

[27] Dartmouth Flood Observatory. Space-based measurement, mapping, and modeling of surface water: For research, humanitarian, and water resources applications, 2023.

[28] Global surface water explorer, 2021.

[29] G. J. Huffman, E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan. Gpm imerg final precipitation l3 half hourly 0.1 degree x 0.1 degree v06, 2019.

[30] Gpm: Global precipitation measurement (gpm) v6, 2023.

[31] FAO. Wapor database methodology: Level 1. Technical report, Remote Sens. Water Prod. Tech. Rep., Rome, 2018.

[32] FAO. Wapor v2 database methodology. Technical report, Remote Sens. Water Prod. Tech. Rep., Rome, 2020.

[33] Gpm: Global precipitation measurement (gpm) v6, 2023.

[34] J. Parra R. Hijmans, S. Cameron, P. Jones, and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965–1978, 2005.

[35] Worldclim bio variables v1, 2023.

[36] T. T. Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48:2839–2846, 2015.

[37] M. G. Pecht and M. Kang. Machine learning: Data pre-processing. *Prognostics and Health Management of Electronics*, 2018.

[38] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6):1–13, 2020.