

Predicting Arsenic Contamination Hotspots inside Abandoned River Bends in Bangladesh: A Machine Learning Approach

J.P. Biesheuvel

Predicting Arsenic Contamination Hotspots inside Abandoned River Bends in Bangladesh: A Machine Learning Approach

by

Julian Peter Biesheuvel

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday, June 13, 2025 at 12:30 AM

Thesis committee:	Dr. R.C. (Roderik) Lindenbergh	TU Delft	Chair and daily supervisor
	Dr. Ir. J. (Jing) Sun	TU Delft	External committee member
Additional supervisor:	Dr. M.E. (Rick) Donselaar	TU Delft	

Student number:	4550188
Project duration:	November 1, 2024 – June 13, 2025

Faculty of Civil Engineering and Geosciences
Stevinweg 1, 2628 CN Delft, The Netherlands

Code and documentation related to this thesis are available at
`https://github.com/JulianPBiesheuvel/AsHotSpots`.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`

This thesis was partly supported by the FAST fund, TU Delft, which sponsored participation in the ESA Living Planet Symposium 2025 in Vienna.

Front cover image: Sentinel-2 mosaic (June 1 – December 1, 2024),
generated via Copernicus Dataspace Browser, centered near Bangladesh (lat: 23.1989, lon: 89.9950).

TU Delft FAST TU Delft University Fund

“Look again at that dot. That’s here. That’s home. That’s us.”

— Carl Sagan

Visual Summary

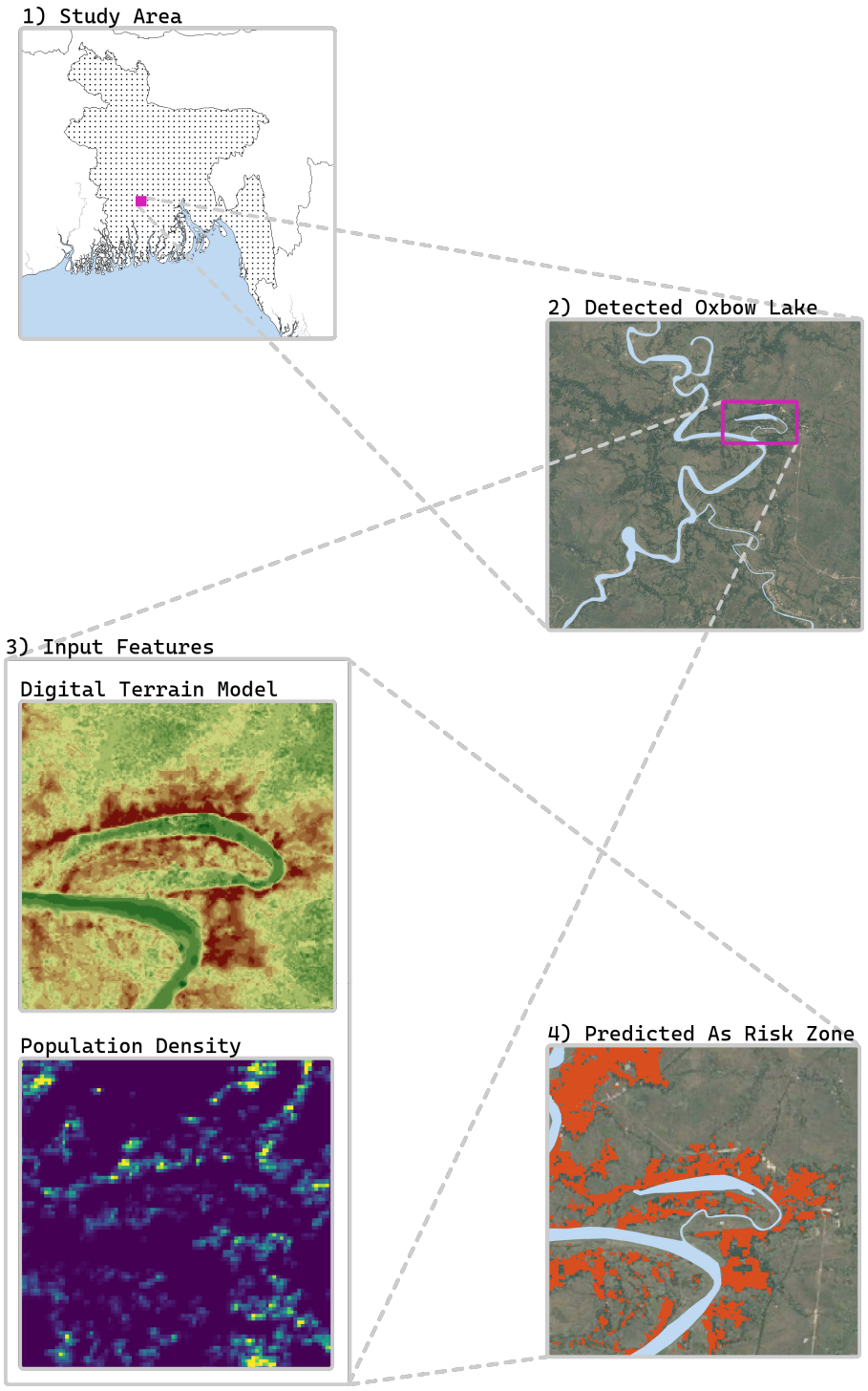


Figure 1: Visual summary of the methodology used for arsenic risk prediction in abandoned river bends in Bangladesh. **1)** The study area is located within the Ganges-Brahmaputra Basin. **2)** Oxbow lakes are detected from satellite imagery using the YOLO object detection algorithm. **3)** Input features, including population density, are extracted for the detected oxbow lake region. **4)** A Gaussian Mixture Model is applied to predict arsenic contamination risk zones, with results visualized as spatial risk categories. Red areas indicate high-risk zones, while potential and low-risk zones are omitted from this image for clarity. This figure presents a simplified overview of all major stages in the current proof-of-concept workflow.

Abstract

Arsenic contamination in groundwater is a major public health concern in the Ganges-Brahmaputra Basin, where millions rely on shallow aquifers for drinking water. Naturally occurring arsenic is mobilised under specific sedimentological and geochemical conditions, particularly in Holocene alluvial deposits. Although extensively studied, arsenic distribution remains highly variable and difficult to predict. This study investigates how geomorphological features, specifically oxbow lakes and point bars, can be used to improve arsenic risk prediction and mapping using machine learning. The approach offers a targeted and scalable method for identifying high-risk zones, particularly in data-scarce environments. The divergence between theoretical assumptions and dataset trends illustrates the challenges of generalising risk models without high-precision, ground-validated input data.

As a proof of concept, a two-stage workflow was implemented. In the first stage, a You Only Look Once object detection model was trained to locate oxbow lakes and point bars using satellite imagery. These landforms are key indicators of arsenic-prone zones due to their depositional history. The model performed well on well-isolated oxbow lakes and their associated point bars but struggled with hydrologically connected oxbow lakes and heavily vegetated areas, highlighting the need for more diverse training data and the potential value of false-colour imagery.

A case study was conducted using historical arsenic well measurements to evaluate model assumptions. A supervised classification with the eXtreme Gradient Boosting algorithm confirmed the predictive value of geomorphological variables, with sand content, elevation, and soil organic carbon emerging as dominant predictors. Vegetation and precipitation data were excluded due to low relevance and poor temporal alignment.

In the second stage, a Gaussian Mixture Model was applied to classify arsenic risk using the same geospatial variables. The model produced spatially coherent and interpretable risk zones, with high probability in most predictions. Areas of low probability were primarily located at transition zones between risk classes, indicating regions where higher-resolution or more precise input data may be necessary to reduce uncertainty and improve model reliability.

This study provides a practical and semi-automated framework for geospatial arsenic risk assessment. While the risk classification is relative, future work should incorporate population-weighted exposure metrics to better guide mitigation. The method developed here supports more efficient fieldwork planning and decision-making in complex fluvial environments.

Preface

Before you lies my master's thesis, which marks the culmination of my studies at Delft University of Technology. I began my academic journey in the bachelor's program in Computer Science and Engineering, followed by the master's in the same field. Out of curiosity, I took the courses Geology and Sensing Technologies, and those turned out to be a turning point. They sparked my curiosity about how technology can be used to better understand natural environments and be leveraged for positive impact. A fascination with natural processes and a love for the outdoors naturally drew me to this field. These interests ultimately led me to switch to the master's in Applied Earth Sciences. It was a decision I've never regretted and a journey I've thoroughly enjoyed. One of the highlights in the program was the unforgettable fieldwork in Iceland with Mats, Bart, and Jorrit.

Choosing this thesis topic felt like a natural step. I liked that it wasn't confined to just one discipline but instead brought different perspectives together to address a real-world problem. The aim was to develop something meaningful for people who, sadly, cannot take access to safe drinking water for granted. I want to thank Rick for proposing and posting this master thesis topic, and for the in-depth discussions about how to approach it. This master thesis helped me realise how much we take for granted and made the real-world relevance of this research very tangible to me.

Over the past six months, I've learned a lot about the interplay between natural processes that govern arsenic mobilisation and risk. At times, it felt like I was doing 'detective work', digging through data, testing hypotheses, and exploring models. I truly enjoyed this process, and I'm thankful for the weekly meetings with Roderik. Your support gave me the space to work independently while also offering valuable insights and fresh perspectives when I needed them. I want to thank Roderik in particular for suggesting the idea of presenting my work at the ESA Living Planet Symposium 2025, which has been a great opportunity to share my research with a broader audience. Thank you, Roderik, Jing, and Rick, for your feedback and for giving me the confidence to carry this research forward. And thanks to the Environment and Population Research Center (EPRC), Dhaka, Bangladesh for providing the extensive dataset on groundwater arsenic concentrations in the study area of Araihazar Upazila in Bangladesh, that was used in the present study.

Looking back on my 3 years at the faculty, the final six months were mostly spent in Room 3.34, which quickly became our office — shared between Yongxing, Isabel, and me. I'm grateful for the many hours spent there with them. Thank you both for your support, company, and all the good times, from endless coffee breaks and cake reviews to many laughs. Thanks also to Alkis for all those walks around the faculty or campus, especially when we were both stuck on our theses or just needed some fresh air.

Finally, I want to thank all my other friends who I haven't named here. Thank you for your continuous support, for always lending a listening ear when I needed one, and for providing the much-needed distraction from the thesis when it was most welcome. It truly meant more to me than I can say. And of course, I want to thank my parents and my brother, in particular my mom and dad, for their unconditional love and unwavering support.

- Julian Biesheuvel
Zuidland, June 2025

Contents

Visual Summary i

Abstract ii

Preface iii

List of Figures vii

List of Tables vii

Glossary viii

1. Introduction 1

1.1 Research Context . . . 1
1.2 Related Work and Problem Statement . . . 1
1.3 Research Significance and Methods . . . 1
1.4 Research Questions . . . 2
1.5 Thesis Outline . . . 3

2. Geospatial and Environmental Context 4

2.1 Arsenic Contamination in Bangladesh: Environmental and Geological Context . . . 4
2.1.1 Geological, Hydrological, and Environmental Setting of Bangladesh . . . 4
2.1.2 Geomorphological Traps: Oxbow Lakes, Clay Plugs, and Point Bars . . . 4
2.1.3 Human Interaction: Groundwater Usage and Exposure Risks . . . 5
2.2 Foundations for a Data-Driven Detection Framework . . . 6
2.3 Geomorphological Features Detection from Remote Sensing Data . . . 6
2.3.1 Object Detection Methods . . . 6
2.4 Risk-Based Arsenic Mapping Using Machine Learning . . . 7
2.4.1 Conceptual Framework for Arsenic Risk . . . 7
2.4.2 Predictive Risk Mapping Approach . . . 7
2.4.3 Risk Classification Method . . . 8

3. Methodology and Data 9

3.1 General Workflow Overview . . . 9
3.2 Evaluating (Auxiliary) Predictors for Arsenic Risk Mapping with a XGBoost Model . . . 9
3.2.1 Auxiliary Data Sources . . . 10
3.3 YOLO-OBB for Object Detection . . . 10
3.3.1 Data Sources and Retrieval . . . 11
3.3.2 Training and Evaluation Strategy . . . 12
3.3.3 Post-Processing for High-Resolution Inference: SAHI . . . 14
3.4 Gaussian Mixture Model for Risk Clustering . . . 14
3.4.1 Data Sources and Retrieval . . . 14
3.4.2 Cluster Assignment and Risk Level Mapping . . . 15

4. Case Study **17**
4.1 Historical Arsenic Data from Wells in Araihazar, Bangladesh . . . 17
4.2 (Auxiliary) Data Sources as Indicators of Arsenic Risk . . . 18
4.2.1 Auxiliary Data Sources . . . 19
4.3 Testing (Auxiliary) Data Sources for Arsenic Risk Prediction . . . 21
4.4 Model Interpretation and Feature Relevance . . . 24

5. Model Implementation and Results **25**
5.1 Historic Arsenic Dataset for Model Implementation Testing . . . 25
5.2 Testing Oxbow Lake Detection Using a Trained YOLOv11x-OBB . . . 25
5.3 Testing Arsenic Risk Classification using GMM . . . 29
5.4 Application of Oxbow Lake Detection in Risk Mapping . . . 32
5.5 Summary Results . . . 32

6. Discussion **33**
6.1 Interpretation of Key Results and Findings . . . 33
6.1.1 Arsenic-Risk Prediction by the Gaussian Mixture Model . . . 33
6.1.2 Data Quality Limitations . . . 33
6.1.3 Oxbow Lake Detection by the YOLO Model . . . 34
6.1.4 Other Approaches for Arsenic Risk Prediction . . . 34
6.2 Context within Literature . . . 35

7. Conclusion and Recommendations **36**
7.1 Conclusion . . . 36
7.1.1 Research Questions Revisited . . . 36
7.1.2 Summary . . . 38
7.2 Recommendations . . . 39
7.2.1 Geomorphological Feature Object Detection . . . 39
7.2.2 Arsenic-risk Prediction . . . 39
7.2.3 Incorporating Population Exposure into Risk Assessment . . . 40

A Supplementary Tables **41**
A.1 Environmental Variable Categories for Arsenic Risk Modeling . . . 41
A.2 Overview of Monitored Wells and Arsenic Levels . . . 41

B Supplementary Figures **42**
B.1 Relationship Between Arsenic Concentration and Well Depth . . . 42
B.2 Supporting Analysis of Elevation and Population Density Distributions . . . 43

List of Figures

1 Visual summary of the methodology used for arsenic risk prediction in abandoned river bends in Bangladesh. **1)** The study area is located within the Ganges-Brahmaputra Basin. **2)** Oxbow lakes are detected from satellite imagery using the YOLO object detection algorithm. **3)** Input features, including population density, are extracted for the detected oxbow lake region. **4)** A Gaussian Mixture Model is applied to predict arsenic contamination risk zones, with results visualized as spatial risk categories. Red areas indicate high-risk zones, while potential and low-risk zones are omitted from this image for clarity. This figure presents a simplified overview of all major stages in the current proof-of-concept workflow. i

2 Stages in oxbow lake formation: (a) early meander with outer bend erosion and inner bend deposition, (b) narrowing neck, (c) neck cut-off during high flow, and (d) complete isolation forming an oxbow lake. Red and yellow overlays indicate erosion and deposition, respectively. The fully developed oxbow lake has a maximum diameter of approximately 4 km. Madhumati River, Dhaka Division, Bangladesh. 23°14′26.97″N, 89°41′27.49″E. Map Data: Google, © 2025 Maxar Technologies. Image dates: see sub-figures. 5

3 Conceptual framework for arsenic risk mapping based on hazard and exposure. The diagram illustrates how geomorphological features such as point bars and sandy ridges (hazard layer) and population presence (exposure layer) combine to define arsenic risk. Risk levels are classified as High Risk for areas with both high elevation and high population density, Potential Risk for areas with high elevation and low population density, and Low Risk for areas with low elevation regardless of exposure. 7

4 Workflow for predictive arsenic-risk mapping. (1) A region of interest (ROI) containing oxbow lakes is manually selected in Bangladesh. (2) Oxbow lakes are automatically detected within the ROI using a YOLOv11x-OBB object detection model. (3) Population density and DTM, along with other auxiliary data sources, are retrieved for the ROI. (4) A GMM is used to perform unsupervised risk classification based on all data sources. 9

5 Example detection of oxbow lakes in a regional-scale true-colour mosaic acquired in May 2025 from Landsat/Copernicus sources, viewed at ∼150 km altitude. Oxbow lakes appear as small, curved water bodies within the broader river landscape. The Madhumati River is visible on the right. The prominent meander with a point bar, just before the river bifurcates, spans ∼10 km. Coordinates: 23°23′12.72″N, 89°0′26.19″E. Image exported via Google Earth Pro. Map data: Google, © 2025 Maxar Technologies. 10

6 Geographic distribution of oxbow lake locations used for model training and validation. Each point represents a site where satellite imagery of oxbow lakes was collected. 11

7 Both images were obtained from Sentinel-2 Level-2A true-color imagery through the Copernicus Dataspace Browser. North arrow and flow direction are omitted as these image are for illustrative purposes. 11

8 Comparison between ground-truth validation labels (a) and model predictions (b) for oxbow lake detection. 13

9 Spatial distribution of tube wells sampled for arsenic concentration in Araihazar Upazila, Bangladesh, overlaid on georeferenced Google Earth Pro imagery from April 2002. Each coloured marker represents a tube well [68]. Coordinates: 23°46′48.00″N, 90°38′26.00″E. Image exported via Google Earth Pro. Map data: Google, © 2025 Maxar Technologies. 17

10 Temporal variability of arsenic concentrations in 20 monitored wells in Araihazar Upazila, Bangladesh, from 2001 to 2003 [68]. 18

11 Aerial image of a meander belt in Bangladesh showing the location of the elevation transect (A–B) across a point bar, analysed in Figure 12. Image taken from Google Earth Pro, dated 1 September 2025. Coordinates: 23°08′27.10″N, 89°42′42.29″E. Map data: Google, © 2025 Maxar Technologies. . 20

12 Elevation profile across a point bar, highlighting variations in TPI and TWI. Labels A and B correspond to the endpoints of the transect shown in white in Figure 11 21

13 Training and validation log loss over 100 epochs for different soil depth feature sets. Each line represents a distinct soil depth range (Set 1 through Set 6), with solid lines indicating validation loss and dashed lines indicating training loss. Set 1 (0–5 cm) shows the lowest validation loss and smallest train-validation gap. Deeper soil sets exhibit progressively higher validation losses and larger gaps, suggesting increased overfitting. 22

14 Feature importance analysis for the best-performing XGBoost model using Set 1 variables (0–5 cm soil depth). **Left:** Model-derived feature importance. **Right:** SHAP-based feature importance. . . . 22

15 Model evaluation and feature correlation analysis for the best-performing XGBoost model using Set 1 variables (0–5 cm soil depth). **Left:** Confusion matrix. **Right:** Pearson correlation matrix of the entire dataset. 23

16 SHAP waterfall plot for the best-performing XGBoost model using Set 1 variables (0–5 cm soil depth). The plot shows how individual features contribute to a single high-risk prediction, with SOC_0_5, SAND_0_5, and elevation having the strongest positive influence. 24

17 Spatial distribution of historic arsenic sampling locations across the Upazilas of Kalia and Lohagara. The three panels represent a north-to-south segmentation of the study area. The satellite imagery was obtained from Google Earth Pro and georeferenced, captured in May, 2025. Map data: Google, © 2025 Maxar Technologies. 26

18 YOLO test cases (1–4). The satellite imagery was obtained from Google Earth Pro, captured in May 2025, and December 2026. Map data: Google, © 2025 Maxar Technologies. 27

19 YOLO test cases (5–8). The satellite imagery was obtained from Google Earth Pro, captured in May 2025, and December 2026. Map data: Google, © 2025 Maxar Technologies. 28

20 Geospatial layers supporting arsenic risk assessment in the area of interest, classified using Gaussian Mixture Model (initialized with `kmeans`). **A):** GMM-based arsenic risk classification based on **elevation** and **population density**. **B):** Population density map (for visualisation histogram-equalised), emphasizing spatial patterns of human settlement. **C):** Maximum prediction probability derived from the GMM model, representing the probability of the assigned risk class. **D):** DTM showing topographic variation across the study region. 30

21 Spatial overview of the arsenic risk classification alongside relevant geospatial predictor variables. These layers provide spatial context for interpreting the risk distribution and support the understanding of relationships between predictor variables and model outcomes. 31

22 Joint distribution of arsenic concentration (µg/L) and well depth (log-transformed, in meters) based on the historic dataset from Kalia and Lohagara Upazilas. 42

23 Distribution analysis of elevation and population density. The bottom row presents corresponding Q–Q plots to assess normality. While elevation approximates a normal distribution, population density remains highly skewed, indicating a non-Gaussian distribution structure. 43

24 Distribution analysis of elevation and population density after transformation to approximate Gaussian distributions. The bottom row presents corresponding Q–Q plots to assess normality. While the transformation effectively normalizes elevation, population density remains highly skewed, indicating residual non-Gaussian characteristics. 43

List of Tables

1 Overview of test images used for evaluating YOLO model performance across different observation altitudes and geomorphological settings. 25

2 Categorisation of environmental variables used for arsenic risk modelling. 41

3 Well information including coordinates, depth, installation year, sampling count, and mean arsenic concentration, from [68]. 41

Glossary

oxbow lake After the river abandons a major meander, it frequently remains behind as a crescent-shaped lake called an oxbow (lake) [1].

point bar A point bar develops along the inner banks of meanders, where reduced flow velocity promotes the gradual accumulation of clastic materials transported by the river [1].

1. Introduction

1.1 Research Context

Bangladesh is the most affected country worldwide by arsenic poisoning, with approximately 50 million people at risk due to chronic arsenic exposure through groundwater and food. Shallow aquifers, which serve as the primary source of drinking and cooking water for the majority of the population, are heavily contaminated [2], [3]. These individuals are chronically exposed to arsenic levels that exceed the World Health Organisation (WHO) recommended limit of 10 µg/L, raising serious public health concerns [2], [4]. In numerous cases, arsenic exposure exceeds 50 µg/L, which is the regulatory limit defined by the Government of Bangladesh [5]. Prolonged exposure to arsenic is associated with an increased incidence of cancers, cardiovascular diseases, and premature mortality in young adults, among others [6]. Alongside the physical impacts of arsenic poisoning, those with arsenicosis also endure social and economic hardships, including discrimination in both social and workplace environments [7]. Bangladesh's government started the Arsenic Risk Reduction Project in 2020 to actively tackle the widespread arsenic poisoning by raising awareness among the public and providing safe drinking water using technologies such as deep tube wells and rainwater harvesting [8]. Given the widespread threat of arsenicosis across Bangladesh, accurately identifying high-risk regions in urgent need of intervention remains a persistent challenge.

1.2 Related Work and Problem Statement

Conventional methods for measuring arsenic concentrations to identify areas of high risk, such as soil and groundwater sampling, often require costly and time-consuming fieldwork, which is further hindered by the Monsoon period. Consequently, no large, detailed, and up-to-date arsenic distribution information at a high sampling resolution is available for Bangladesh. For the available datasets, numerous studies have tried to predict arsenic hotspots using geospatial interpolation methods such as Kriging, Inverse-Weighting Distance, or using Random Forest (RF) estimators incorporating various features, including different soil types, to interpolate measured arsenic concentrations and risks [9]–[14]. A key limitation of Kriging and Inverse-Weighting Distance method is that it often produces an imprecise representation of high arsenic concentrations and risks, and generates artificial peaks, i.e. misleading 'bull's eye' patterns around data points. While RF models provide a more refined view of arsenic concentration and risk patterns, incorporating a large number of variables, such as different soil types, often relies on incomplete, coarse-resolution datasets or demands extensive fieldwork to obtain detailed data. A shared limitation of these approaches is their failure to account for the geomorphological features, such as point bars and oxbow lakes, that underpin the spatial distribution of arsenic. Integrating these geomorphological features provides a more robust and physically grounded framework for hotspot identification.

1.3 Research Significance and Methods

The key underlying factor among these studies and datasets is the geographical location of the groundwater and/or soil samples. Within the Ganges-Brahmaputra Basin, Bangladesh is characterised by sampling sites that frequently derive from Holocene alluvial basins. Recent research suggests that sand-rich point bars within abandoned meandering river bends, which are common features in Holocene alluvial basins, are potential hotspots for high arsenic concentrations in groundwater and soils [15], [16]. These abandoned meandering river bends, known as oxbow lakes, are characterised by their crescent shape. As suggested by Donselaar *et al.* [17], oxbow lakes can be distinguished by their shape, their size relative to other geomorphological features, and their contrast with the surrounding landscape, making them suitable candidates for automatic detection using machine- or deep-learning methods. Once the oxbow lakes (i.e. areas with high arsenic concentrations) and their associated point bars are identified, the urgency of community-level intervention remains uncertain. Each identified area is assigned a risk level based on a set of predictor variables, combined with a machine learning method and a study-specific definition of risk. In this way, a more precise predictive arsenic-risk map can be constructed, focusing on regions that require immediate action from local government authorities to mitigate arsenic poisoning.

1.4 Research Questions

The main objective of this research is to develop a machine-learning model that accurately identifies high-risk zones for arsenic contamination in Bangladesh, without the need for extensive fieldwork or direct in-situ arsenic measurements, by leveraging geomorphological features such as point bars and oxbow lakes, in line with the methodology proposed by Donselaar *et al.* [17]. The scope of this study is two-fold: first, to detect oxbow lakes and point bars across the landscape using satellite-based remote sensing, second, to conduct a targeted arsenic risk analysis focused on these geomorphological settings. Rather than predicting exact arsenic concentrations, this research aims to classify areas as either at risk or not at risk, enabling spatial prioritisation for intervention. Based on this objective and its components, the main research question of this project can be defined as:

How can geomorphological features, specifically oxbow lakes and point bars, be used to optimise machine-learning models for accurate arsenic risk prediction and mapping in the Ganges-Brahmaputra Basin?

The following sub-questions will support the main question:

1. *How are the environmental conditions and geological characteristics in the Ganges-Brahmaputra Basin related to groundwater arsenic contamination?*

   Because environmental and geological conditions substantially impact arsenic levels and related risks, it is necessary to understand their interactions and site-specific roles clearly. This sub-question will further deepen and investigate the complex interplay between the variables of interest in the Ganges-Brahmaputra Basin.

2. *Which method is effective for detecting geomorphological features such as oxbow lakes and point bars?*

   A machine-or deep-learning method can be used to pinpoint the point bars and oxbow lakes from satellite imagery. An essential condition the method should satisfy is that it must be able to generalise to other areas with similar geomorphological features. This sub-question aims to identify a method that is both easy to implement (i.e. with minimal effort, time, and resources) and efficient for detecting oxbow lakes and point bars.

3. *To what extent do auxiliary environmental and geomorphological variables explain the spatial and temporal distribution of arsenic contamination risk?*

   Donselaar *et al.* [17] identify geomorphological features, population density, and vegetation indices as key variables associated with high arsenic concentrations. For this sub-question, a dataset containing repeated measurements of arsenic concentrations from tube wells in Bangladesh, collected over a three-year period in a geomorphological setting characterised by oxbow lakes and point bars, will be used as a case study to examine the spatial and temporal distribution of arsenic contamination risk, with a focus on the contribution of key predictor variables.

4. *How should arsenic contamination risk be defined to align with both model outputs and the practical implications for public health and groundwater management?*

   Arsenic levels in groundwater that exceed the WHO's recommended limit of 10 µg/L are classified as high risk for human health. However, in order to implement effective mitigation strategies against arsenic poisoning on a community level, the population density also needs to be considered. A definition of arsenic contamination risk for this study will be established based on the combination of population density and other influential variables identified in Sub-question 3 as strongly correlated with elevated arsenic levels.

5. *What machine-learning method is suited for classifying arsenic risk?*

   The objective of the sub-questions is to identify what machine-learning method can classify different levels of risk, as defined in Sub-question 4, based on notable features such as a Digital Terrain Model (DTM), population density, and other influential variables found in Sub-question 3 that strongly correlate with high arsenic concentrations. Risk is categorised into clusters, with each classification associated with a probability. The method aims for pixel-level classification of arsenic risk, prioritising high-confidence probabilistic outputs.

1.5 Thesis Outline

This report is structured as follows. The **Geospatial and Environmental Context** chapter reviews arsenic contamination, relevant geomorphological processes, existing detection methods, risk theory, and object detection and clustering techniques. The **Methodology and Data** chapter outlines the proposed framework, describing the object detection and risk classification approaches and how available (auxiliary) data are processed and can be integrated. The **Case Study** chapter presents a specific case in which the auxiliary data sources are tested and the hypotheses outlined in Geospatial and Environmental Context are evaluated. The Model Implementation and Results chapter details the implementation and testing of the framework, and presents the outcomes of the experiments. Finally, the **Discussion and Conclusion** chapters interpret the results, answer the research questions, and offer recommendations for future work.

2. Geospatial and Environmental Context

This chapter outlines the environmental and geomorphological context of arsenic contamination, with a focus on Bangladesh. It examines natural processes, human activities, and climate impacts. The chapter also introduces a data-driven method for detecting oxbow lakes and point bars using remote sensing, and presents an unsupervised machine learning approach for classifying arsenic risk based on geomorphology and population exposure.

2.1 Arsenic Contamination in Bangladesh: Environmental and Geological Context

Essential to this research is the environmental and geological context in which arsenic contamination occurs in Bangladesh, where both natural processes and anthropogenic activities contribute to elevated arsenic levels in groundwater. The following sections explore arsenic's geogenic origin, the role of geomorphological features in its accumulation, and the interaction between human activity and arsenic-contaminated groundwater, including associated health risks.

2.1.1 Geological, Hydrological, and Environmental Setting of Bangladesh

Inorganic arsenic, a toxic metalloid, is a natural occurring component of the Earth's crust. In Bangladesh, arsenic primarily originates from geological sources, with the mineral arsenopyrite, an iron hydroxide bound to arsenic, identified as the main source [18], [19]. Since the last Ice Age, approximately 12,000 years ago, arsenic-bearing minerals eroded from the Himalaya-Arakan-Yoma mountain range have been transported downstream via the Ganges-Brahmaputra-Meghna river system [20]. As these minerals were deposited, the Holocene alluvial basin of the Bengal Delta gradually formed, with fluvial processes continuously reworking and sorting the sediments, leading to the accumulation of arsenic-rich layers with large lateral and vertical variability, driven by lithofacies heterogeneity and permeability contrasts that control groundwater flow within the aquifer [16], [20]–[23]. Biogeochemical processes then mobilise arsenic from sediments into groundwater through microbial redox transformations of iron phases, leading to arsenite [As(III)] presence in the groundwater [24]. Shallow aquifers, often with organic matter-rich peat layers at depths of less than 100 meters, are primarily affected by arsenic contamination [3], [16], [20], [22].

A recurring natural phenomenon in Bangladesh is the monsoon, which runs from early June to mid-October. The average rainfall varies from 1200 mm in the West to 3000 mm in the Northeast and Southeast [25]. In addition to recharging aquifers, monsoon-driven changes in hydraulic head also regulate the mobilisation of dissolved organic carbon (DOC) from surrounding clay–peat layers. The hydraulic head refers to the pressure and elevation that drive groundwater flow, influencing how water and dissolved substances move through the subsurface. During the dry season, declining groundwater levels promote the release of DOC into adjacent aquifers, while the onset of the monsoon and subsequent recharge suppress DOC release and induce aquifer flushing [26]. Climate change is projected to intensify these processes by altering monsoon patterns and increasing the frequency of extreme weather events. Such changes can exacerbate the mobilisation of arsenic into groundwater [27], [28].

2.1.2 Geomorphological Traps: Oxbow Lakes, Clay Plugs, and Point Bars

A review by Donselaar *et al.* [16], a preliminary study, identified that arsenic hotspots are often located within Holocene fluvial and deltaic flood basins and predominantly include geomorphological features such as oxbow lakes, clay plugs, and point bars. This study focuses exclusively on oxbow lakes. The formation process of oxbow lakes is illustrated in Figure 2. Sand-rich fluvial point bars form within shallow meandering river bends, which either remain water-filled as oxbow lakes or become infilled with fine-grained, low-permeable sediments to form clay plugs [16]. The fine-grained sediments and the stagnant bottom waters of oxbow lakes are both rich in organic carbon, creating oxygen-poor (anoxic) conditions that promote microbial activity. This microbial activity drives the reductive dissolution of iron oxyhydroxides, releasing arsenic into the groundwater [16]. Organic carbon, consumed by microbial communities, originates primarily from lake vegetation (macrophytes), such as *Eichhornia crassipes* sp. and *Hydrilla verticillata* sp., as well as suspended clay and silt [29]. Flooding during the monsoon season drowns the macrophytes, whose biomass deposits onto the lakebed, providing a carbon source that fuels microbial activity. Oxbow lakes are susceptible to extreme weather events; during droughts they may dry up, while monsoonal rains can cause them to overflow or reconnect temporarily with the main river.



(a) September 2001 (b) December 2006 (c) January 2015 (d) January 2025

Figure 2: Stages in oxbow lake formation: (a) early meander with outer bend erosion and inner bend deposition, (b) narrowing neck, (c) neck cut-off during high flow, and (d) complete isolation forming an oxbow lake. Red and yellow overlays indicate erosion and deposition, respectively. The fully developed oxbow lake has a maximum diameter of approximately 4 km. Madhumati River, Dhaka Division, Bangladesh. 23°14′26.97″N, 89°41′27.49″E. Map Data: Google, © 2025 Maxar Technologies. Image dates: see sub-figures.

In a later studies by Ghosh *et al.* [24], [29], the concept of geomorphological features such as oxbow lakes and point bars acting as arsenic hotspots was further developed. Differential compaction in sandy point bars leads to topographically elevated ridges, which are part of the lateral accretion patterns typical of point bar formation, where arsenic accumulates within the aquifer [16], [29]. Normally, aquifers are flushed, and their recharge efficiency is high due to groundwater flow following a gravity-driven gradient. In contrast, the stratigraphic confinement of point bars by surrounding oxbow lakes or clay plugs, both characterised by low permeability and low porosity, restricts horizontal groundwater flow and reduces recharge efficiency, resulting in the persistent accumulation of arsenic within the aquifer.

2.1.3 Human Interaction: Groundwater Usage and Exposure Risks

Point bars act as population nuclei for three main reasons: their elevated ridges protect against annual monsoon floods, the surrounding floodplain areas are highly fertile, supporting agriculture, and the sandy point bars serve as accessible aquifers, providing groundwater for drinking and irrigation. As a result, these areas attract dense settlements where groundwater is primarily accessed through shallow hand-pump tube wells, the most common form of water supply in rural Bangladesh [3]. These wells, typically 10 to 70 meters deep, are widespread due to their low installation costs and the absence of government regulation on well construction [20]. Once arsenic is mobilised within a nearby clay plug or oxbow lake, it migrates into the adjacent porous and permeable point bars [15]. Extensive groundwater extraction by local populations induces a pressure gradient, enhancing advective flow and drawing arsenic-contaminated water toward the tube wells, thereby exacerbating the health risk [15], [16], [24]. Ghosh and Donselaar [15] further suggest that this over-extraction in densely populated areas leads to more wells exceeding accepted arsenic safety limits.

Another important reason for the prompt identification of densely populated arsenic hotspots is the increasing impact of climate change on the release of arsenic into the environment. Rising sea levels, floods, and extreme weather events, driven by the warming climate, contribute to the release of increased amounts of arsenic from sediments into groundwater [30], [31]. The root causes of the elevated arsenic concentration levels are the salinisation of aquifers and increased reduction rates [30], [31]. Due to rising arsenic levels in the environment, arsenic concentrations in rice may also increase [32]. Bangladesh is the world's third-largest producer, with almost 7 % of the world's total production [33].

2.2 Foundations for a Data-Driven Detection Framework

Data-driven approaches are increasingly used in environmental risk modelling because these methods rely on patterns learned directly from the data, without predefined rules or assumptions. For arsenic risk mapping, this enables the identification of relationships between environmental variables and the likelihood of contamination. Machine learning models are well suited to this task due to the ability to handle complex, multi-dimensional input data and account for uncertainty. Donselaar *et al.* [17] show how remote sensing data combined with machine learning techniques can be applied to model arsenic contamination risk across large areas. The study presents a pipeline that incorporates geomorphological features alongside auxiliary datasets such as elevation and population density to identify potential hotspots. Feature selection in this context depends on factors like data availability, spatial resolution, and relevance to the target variable. Model selection often reflects a trade-off between predictive accuracy, interpretability, and computational efficiency. These methodological considerations form the basis for applying machine learning to geospatial arsenic risk analysis.

2.3 Geomorphological Features Detection from Remote Sensing Data

The Ganges-Brahmaputra Delta is the largest delta system in the world, with approximately two-thirds of it located in Bangladesh [34], [35]. Within this deltaic region, oxbow lakes are widespread, yet to the best of the author's knowledge, no existing literature systematically records or quantifies them in Bangladesh. Oxbow lakes are crescent-shaped water bodies formed when a river meander is cut off from the main channel. These lakes are typically hydrologically disconnected from the river but can become temporarily reconnected during seasonal flooding or high-flow events. During the wet season (June–October), monsoonal flooding brings nutrient-rich inflow that deposits organic matter into the lakes. However, high water levels and flushing often submerge or remove aquatic vegetation. In contrast, during the dry season (November–April), water levels recede, light penetration increases, and the previously deposited nutrients support the growth of aquatic vegetation, such as macrophytes. This seasonal cycle results in visible changes in oxbow lake characteristics, with vegetation cover peaking in the dry season and declining during the wet season. These dynamics influence the spectral and visual appearance of oxbow lakes in satellite imagery, affecting how distinctly they contrast with the surrounding landscape throughout the year.

2.3.1 Object Detection Methods

Empirical studies show that oxbow lakes can vary significantly in size, ranging from a few hundred meters to several kilometres. As a result, satellite imagery, which is widely available online, often open-access, and offering high spatial and temporal resolution, is the most suitable option for capturing both oxbow lakes and the point bars within them. Object detection in geoscience using satellite imagery is widely adopted [36]. Convolutional Neural Networks (CNNs) [37], particularly the Mask Region-based CNN (Mask R-CNN) [38], are among the most popular architectures due to their ability to detect objects of varying shapes and sizes, localise them within complex backgrounds, and delineate their exact boundaries. Mask R-CNN is a two-stage deep learning framework that first generates region proposals and then performs classification and pixel-level segmentation. One example that highlights the effectiveness of a Mask R-CNN deep learning model is a study in which a Mask R-CNN was trained for surface water mapping in the boreal forest-tundra [39]. Training Mask R-CNN models requires annotated images with one or multiple bounding boxes and segmented objects. Segmenting objects is the most time-consuming part of creating a training dataset, but this process can be accelerated using tools such as Segment Anything Model by Meta AI [40]. Another popular model for object detection is the You Only Look Once (YOLO) model, which predicts real-time bounding boxes and class probabilities directly from full images in one evaluation [41]. A feature that YOLO offers is oriented bounding boxes (OBB), which allow for a better fit for rotated or elongated objects and, therefore, achieve an improved detection performance. Furthermore, Ultralytics provides pre-trained YOLO models, with the largest model having 58.8 million parameters [42]. This enables training on a custom dataset by leveraging transfer learning and allows for fewer custom training samples compared to training from scratch. The Ultralytics package requires users to use only a few lines of code to train a YOLO model using a custom dataset, perform validation, and run predictions on images. Due to its single-stage design, the YOLO model generally has a lower localisation precision than Mask R-CNN. On the other hand, the YOLO model is more efficient than Mask R-CNN, which is computationally intensive and requires complex implementation [43]. With YOLO's high accuracy, swift implementation, and low computational resources required, it is the favoured model for this study.

2.4 Risk-Based Arsenic Mapping Using Machine Learning

2.4.1 Conceptual Framework for Arsenic Risk

In this project, risk is defined analogously to concepts used in fields such as flood risk analysis, where risk is expressed as the product of hazard, exposure, and vulnerability [44]. Here, hazard refers to the natural tendency for arsenic occurrence, primarily controlled by geomorphology and topography. Exposure is defined as the number of people present at a given location, typically quantified through population density. Vulnerability reflects the degree to which populations are affected upon exposure, representing the lack of resistance to harmful impacts.

It is assumed that high-elevation areas that are densely populated are likely arsenic hotspots. Vulnerability is considered relatively constant across the study area. This is a simplification, as variations exist: some households rely on deep tube wells, which are generally safer and less arsenic-contaminated than shallow tube wells, as deep aquifers are more frequently flushed [45]. However, since groundwater forms the primary drinking water source for nearly all households, it is assumed that, on average, the population has a similar vulnerability to arsenic exposure. Therefore, the risk model in this study primarily focuses on the interaction between hazard and exposure (see Figure 3).

The risk is then preliminarily defined based on the geomorphology and the human exposure. Risk is classified into three levels: High Risk (high elevation and high population density), Potential Risk (high elevation and low population density), and Low Risk (low elevation, independent of population density). This classification is used as an initial framework and may be adjusted based on the case study. Auxiliary data, such as vegetation indices, organic carbon contents, and land use and cover, will be tested to refine the model at a pixel level. Depending on the results of the case study and the performance of each auxiliary data source, the risk definition may be updated to represent local arsenic dynamics and exposure.

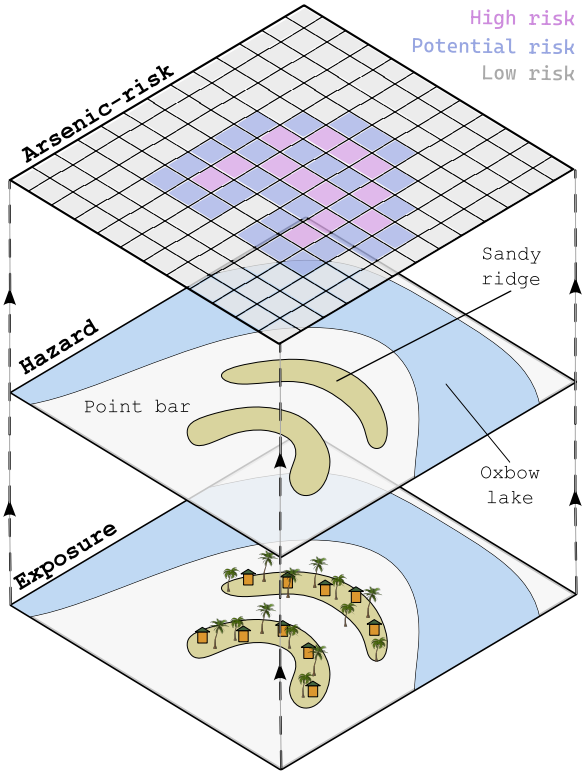


Figure 3: Conceptual framework for arsenic risk mapping based on hazard and exposure. The diagram illustrates how geomorphological features such as point bars and sandy ridges (hazard layer) and population presence (exposure layer) combine to define arsenic risk. Risk levels are classified as High Risk for areas with both high elevation and high population density, Potential Risk for areas with high elevation and low population density, and Low Risk for areas with low elevation regardless of exposure.

2.4.2 Predictive Risk Mapping Approach

After detecting the oxbow lakes, the next step is to conduct a risk analysis to develop a predictive arsenic-risk map, based on the definition and structure outlined in Section 2.4.1. As described by Donselaar *et al.* [17], arsenic risk must be predicted specifically for areas where people live on elevated point bars with access to groundwater. One option for a predictive arsenic-risk map is for each localised oxbow lake and associated point bar to assess its risk based on the available population per square kilometre and Digital Terrain Model (DTM). However, this method provides at best a global overview of risk per point bar but misses an accurate view of risk versus non-risk areas on the point bar itself.

Another option would be to perform pixel-based arsenic-risk prediction for each located point bar. This approach requires multiple data sources to preferably align in spatial and temporal resolution to achieve the best predictive efficiency. A clustering algorithm that can handle multivariate input, is unsupervised, provides probabilities, and can do pattern discovery, is preferred for the pixel-based approach. Unsupervised learning is the method of choice, as the goal is to avoid thresholding or reliance on predefined class boundaries. For example, elevation varies across a point bar, with ridges forming topographic high grounds, and also differs between point bars. Probabilities complement each prediction by providing a measure of certainty regarding the type of risk per pixel. Clustering algorithms are inherently designed for pattern discovery. Hence, in the case of predictive arsenic-risk mapping, the model's expected output is high risk prediction for highly populated point bar ridges and low risk for low topographic areas with barely any population.

2.4.3 Risk Classification Method

Based on all the requirements outlined at the beginning of this subsection, the Gaussian Mixture Model (GMM) [46]–[48] is the most suitable clustering algorithm as it satisfies all the conditions. An unsupervised clustering method such as GMM has been successfully applied in environmental and remote sensing studies [49], [50]. GMM is a probabilistic model that assumes data points are generated from a mixture of several Gaussian distributions, each representing a different cluster. For multivariate input data, each point is assigned to one or more clusters with a probability, with the assignments optimised to maximise the likelihood of the data under a mixture of Gaussian distributions. The number of clusters corresponds to the number of defined risk levels.

3. Methodology and Data

This chapter of the report outlines the methodology for the proof of concept and discusses each method in detail. For each method, the data sources are listed together with the (pre)-processing steps.

3.1 General Workflow Overview

This section outlines the methods introduced in the Chapter 2 and describes how they are integrated within the overall processing pipeline. Figure 4 presents a schematic overview of the three-stage workflow. The schematic will guide the implementation of the methods throughout the project, including the development of the initial proof of concept. The auxiliary data sources will be assessed in the Chapter 4. The following sections detail each method, including data sources and retrieval, how they handle data, their output, and evaluation strategies.

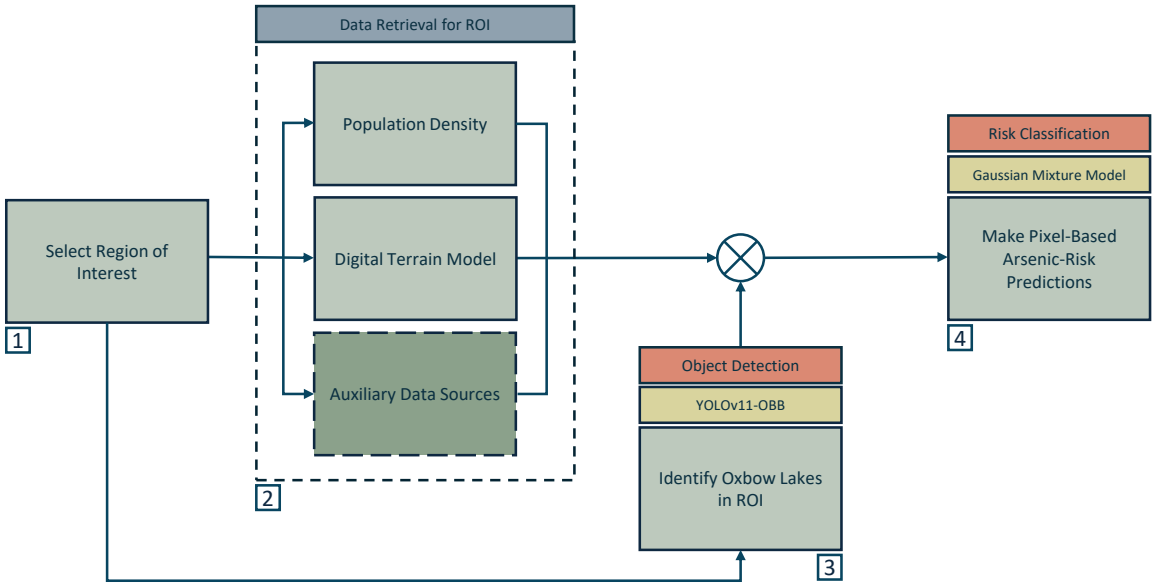


Figure 4: Workflow for predictive arsenic-risk mapping. (1) A region of interest (ROI) containing oxbow lakes is manually selected in Bangladesh. (2) Oxbow lakes are automatically detected within the ROI using a YOLOv11x-OBB object detection model. (3) Population density and DTM, along with other auxiliary data sources, are retrieved for the ROI. (4) A GMM is used to perform unsupervised risk classification based on all data sources.

3.2 Evaluating (Auxiliary) Predictors for Arsenic Risk Mapping with a XGBoost Model

In Chapter 4, a study will be performed to evaluate auxiliary data sources that can potentially contribute to predicting high-risk arsenic areas. The hypothesis, as outlined in Chapter 2, will be validated. Neither this section nor the methods discussed here are part of the general workflow overview and can be considered independent. For this purpose, a tree-based supervised machine learning model is used: eXtreme Gradient Boosting (XGBoost) [51]. XGBoost is an ensemble learning method well-suited for tabular data, constructing decision trees sequentially such that each new tree focuses on correcting the errors of the previous ones. The model can compute feature importance scores, providing insights and interpretability into which (auxiliary) data sources are influential model predictors, for example, identifying high-risk arsenic zones. Furthermore, the model can handle high-dimensional non-linear relations between features in the data without explicit transformations or prior domain assumptions, and includes regularisation to prevent overfitting. While it involves more hyperparameters than RF model, XGBoost offers greater flexibility for fine-tuning the model to better fit the data. This makes it a strong candidate for exploratory studies where the relationships between variables are complex and not well understood. Since this case study is not part of the core workflow, a full methodological breakdown is omitted.

3.2.1 Auxiliary Data Sources

Donselaar *et al.* [17] proposed in their paper the following data sources: geomorphic elements, oxbow lake vegetation intensity, regional climate maps, DEM point bar elevation, and population density maps. Geomorphic elements such as oxbow lakes and point bars, along with DEM (or DTM) and population density, have already been thoroughly discussed in Chapter 2, relevance and necessity for this research, and underpin both the detection and risk assessment components of the proposed methodology. In the Section 2.1.1, the impact of the local climate on arsenic contamination is highlighted. In Section 2.1.2, the interaction between macrophytes (oxbow lake vegetation) acting as an organic carbon source for microbial activity is underlined. Since arsenic concentrations cannot be directly measured using remote sensing and rely solely on in situ measurements, proxy variables serve as the next best alternative. The interplay between factors such as hydrology, geology, and biogeochemical processes makes it challenging to determine which proxy variables are effective predictors of elevated arsenic concentrations. Hence, various auxiliary data sources will be evaluated for their effectiveness in predicting elevated arsenic concentrations. Previous studies that aimed to predict arsenic concentrations and associated risk [10], [14], [30], [52] categorised their predictor variables into distinct groups, as detailed in Table 2 in AppendixA.1.

Soil-related variables are used only in Chapter 4 and are not included in the broader modelling pipeline, as their key characteristics are already inherently captured by the geomorphological features being analysed, particularly point bars. Distance to the river can be omitted, as the aim is to identify risk zones independently of in situ measurements. This study does not consider aridity due to time constraints and the limited availability of consistent data. All other listed variables will be considered as auxiliary data sources. Further details on each auxiliary data source are provided in the Chapter 4, where it will also be clarified which additional datasets are used in the model implementation for the final arsenic risk prediction.

3.3 YOLO-OBB for Object Detection

Once a region of interest is selected, the first stage of the workflow is to detect oxbow lakes from satellite imagery. For this study, we define two types of satellite imagery based on their spatial coverage and level of detail: zoomed-out (regional-scale) imagery refers to images captured at a high viewing altitude, covering large geographical areas where oxbow lakes appear as small features in the landscape; zoomed-in (local-scale) imagery refers to images where individual oxbow lakes occupy most of the image frame, allowing for more fine-grained detail. The objective is to use zoomed-out ([100, 150] km viewing altitude) optical satellite imagery as input to the YOLO model, allowing most oxbow lakes to remain visually identifiable while enabling the processing of large areas in a single pass; the model processes the imagery and outputs bounding boxes and associated detection probabilities for each detected oxbow lake (see Figure 5). However, zoomed-out imagery is not suitable for training the model, as oxbow lakes are typically much smaller than dominant landscape features and can be difficult to detect depending on the season, with additional background noise further complicating the identification process. Furthermore, one of YOLO's known limitations its difficulty in detecting small objects in scenes [41].

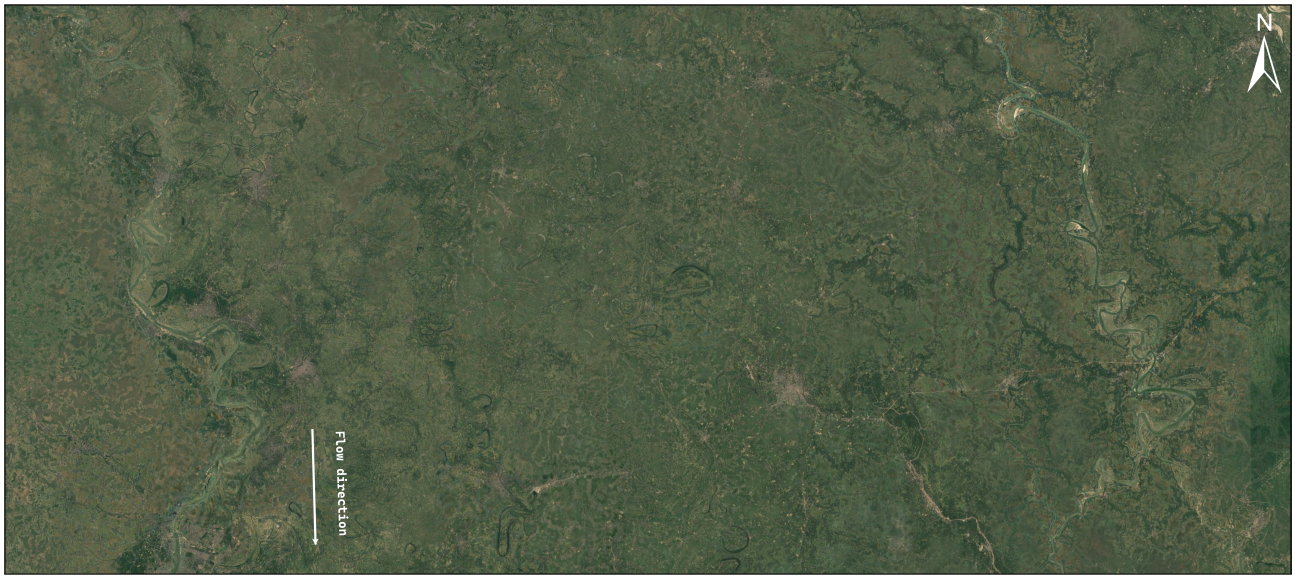


Figure 5: Example detection of oxbow lakes in a regional-scale true-colour mosaic acquired in May 2025 from Landsat/Copernicus sources, viewed at ~150 km altitude. Oxbow lakes appear as small, curved water bodies within the broader river landscape. The Madhumati River is visible on the right. The prominent meander with a point bar, just before the river bifurcates, spans ~10 km. Coordinates: 23°23′12.72″N, 89°0′26.19″E. Image exported via Google Earth Pro. Map data: Google, © 2025 Maxar Technologies.

YOLO is characterised by a single-shot architecture in which the input image is divided into a grid, and each grid cell predicts bounding boxes along with their associated class probabilities. Centre coordinates, width, height, and rotation parametrise the bounding boxes. Class probabilities are determined for each grid cell. For each predicted bounding box, a class-specific score is calculated by multiplying the confidence score of the box with the probability that it contains a specific object class. The next section provides a detailed explanation of how the YOLO model is trained in the context of this study.

3.3.1 Data Sources and Retrieval

The YOLO model was trained on zoomed-in (local-scale) optical satellite imagery to learn the characteristic features of oxbow lakes and their surrounding environments. The Copernicus Dataspace Browser[1] provided the optical satellite imagery used for both the training and test datasets. The images were primarily sourced from Sentinel-2 Level-2A products with a spatial resolution of 10 meters, covering acquisition dates between January 1st 2024 and December 31st 2024, to capture seasonal and inter-annual variability. Figure 6 shows all sampling locations from which training data were extracted, with each image captured under different seasonal and environmental conditions. These images contain one or multiple oxbow lakes, observed at zoomed-in scales (corresponding to viewing altitudes of approximately ~5 to 15 kilometres, depending on the scene and size of the oxbow lake(s)), and may also include other geomorphological features such as rivers or active river channels. Two example samples are shown in Figure 7. The majority of samples were taken outside Bangladesh (with a few within the country), in neighbouring regions that share a similar geomorphological and environmental setting.

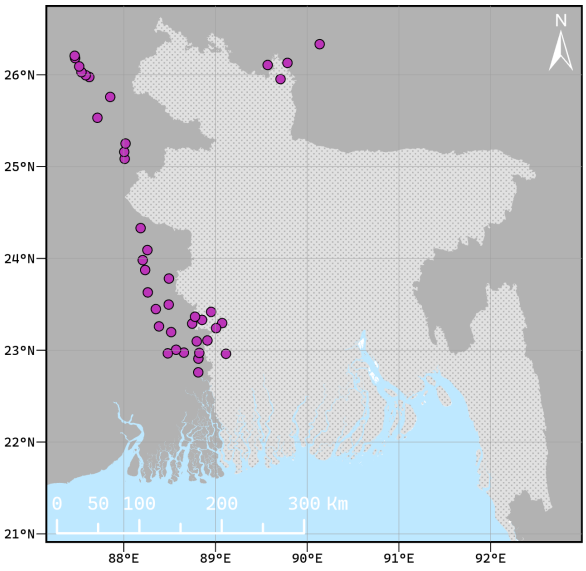


Figure 6: Geographic distribution of oxbow lake locations used for model training and validation. Each point represents a site where satellite imagery of oxbow lakes was collected.



(a) Example image used for ground-truth validation, captured on 9 November 2024.



(b) Example image used for model prediction, acquired on 3 May 2024

Figure 7: Both images were obtained from Sentinel-2 Level-2A true-color imagery through the Copernicus Dataspace Browser. North arrow and flow direction are omitted as these image are for illustrative purposes.

[1] https://browser.dataspace.copernicus.eu/

3.3.2 Training and Evaluation Strategy

In total, 104 samples were collected and subsequently augmented, resulting in a dataset of 208 images. Geometric transformations were applied to the images, including rotations of 90°and 180°, and colour augmentations such as RGB shifts, random gamma adjustments, and blurring. The images were then imported into Label Studio[2] for annotation. Only the class "oxbow" was labelled for this initial proof of concept, as each oxbow lake inherently contains an associated point bar within its structure. Additionally, only inactive oxbow lakes were labelled; oxbow lakes that were partially active or still connected to the main river on either side were excluded from annotation. Additionally, clay plugs were excluded from labelling in this manner. The YOLOv11x-OBB model pre-trained by Ultralytics was used as the base model and fine-tuned on the custom dataset [53], [54]. The input images do not have a standardised resolution, but are resized and normalised by the model to a fixed 640×640 pixel tensor with three colour channels. The dataset was split into 150 images for training, 15 for validation, and 15 for testing.

The YOLO model evaluates its performance using four metrics: precision, recall, IoU, mAP@0.5, and mAP@0.5:0.95. Below each metric, a brief explanation is provided to guide its interpretation.

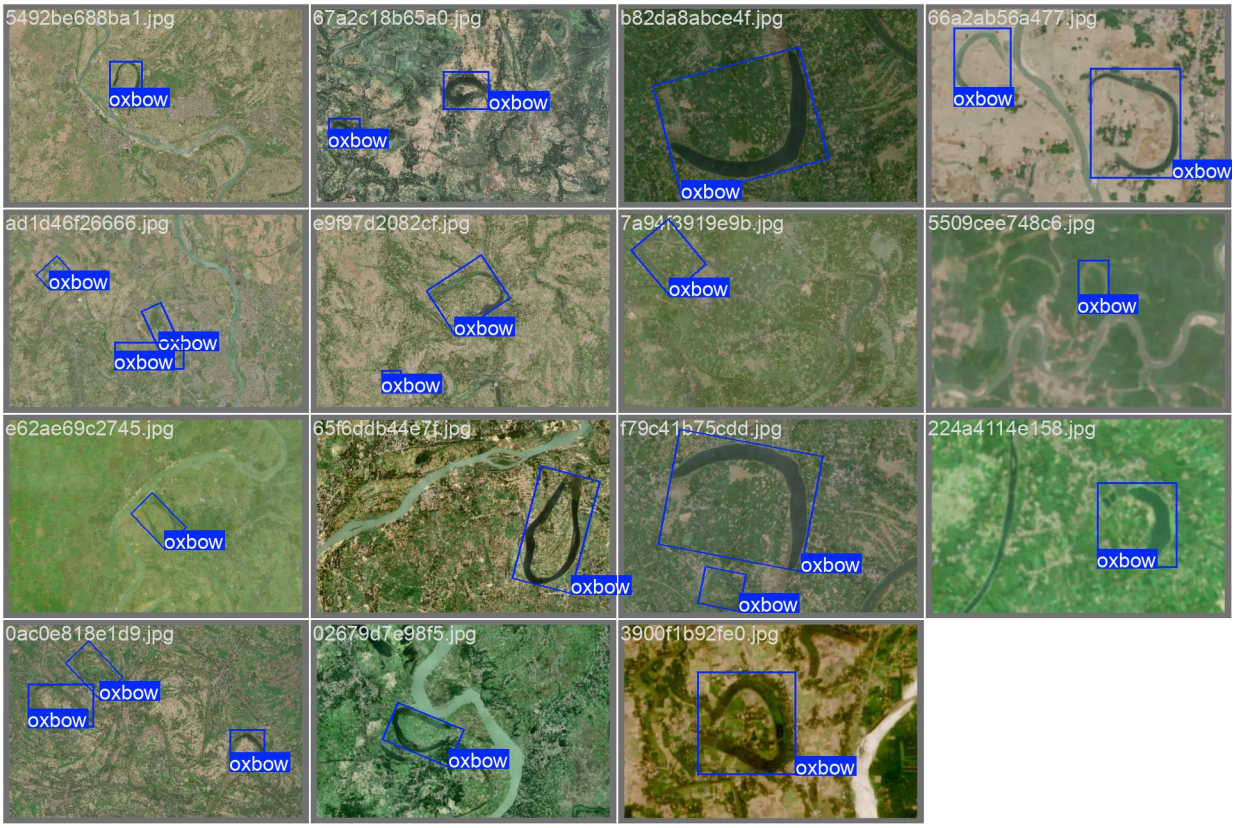
Metric	Formula	Description
Precision	$\frac{\text{TP}}{\text{TP} + \text{FP}}$	How many of the model's true positives were actually correct. Higher is better, meaning very few false positives.
Recall	$\frac{\text{TP}}{\text{TP} + \text{FN}}$	How many of the actual objects the model successfully detected. Higher is better; some objects may be missed, but most are detected correctly.
Intersection over Union (IoU)	$\frac{\text{Area of Overlap}}{\text{Area of Union}}$	The intersection over union between the predicted and ground-truth bounding boxes. Higher is better, meaning the predicted bounding boxes align well with the ground truth.
mAP@0.5	$AP_{\text{IoU=0.5}} = \int_0^1 \text{Precision}(r)\, dr$	
	$\text{mAP@0.5} = \frac{1}{N} \sum_{i=1}^{N} AP_i$	Average precision (area under the precision-recall curve), when IoU threshold = 0.5.
mAP@0.5:0.95	$\text{mAP@[0.5:0.95]} = \frac{1}{10} \sum_{t=0.5}^{0.95} AP_{\text{IoU}=t}$	Mean average precision across IoU thresholds from 0.5 to 0.95, in 0.05 increments. Again, higher is better.

Table: Evaluation metrics used to assess YOLO model performance, including formula and interpretation.

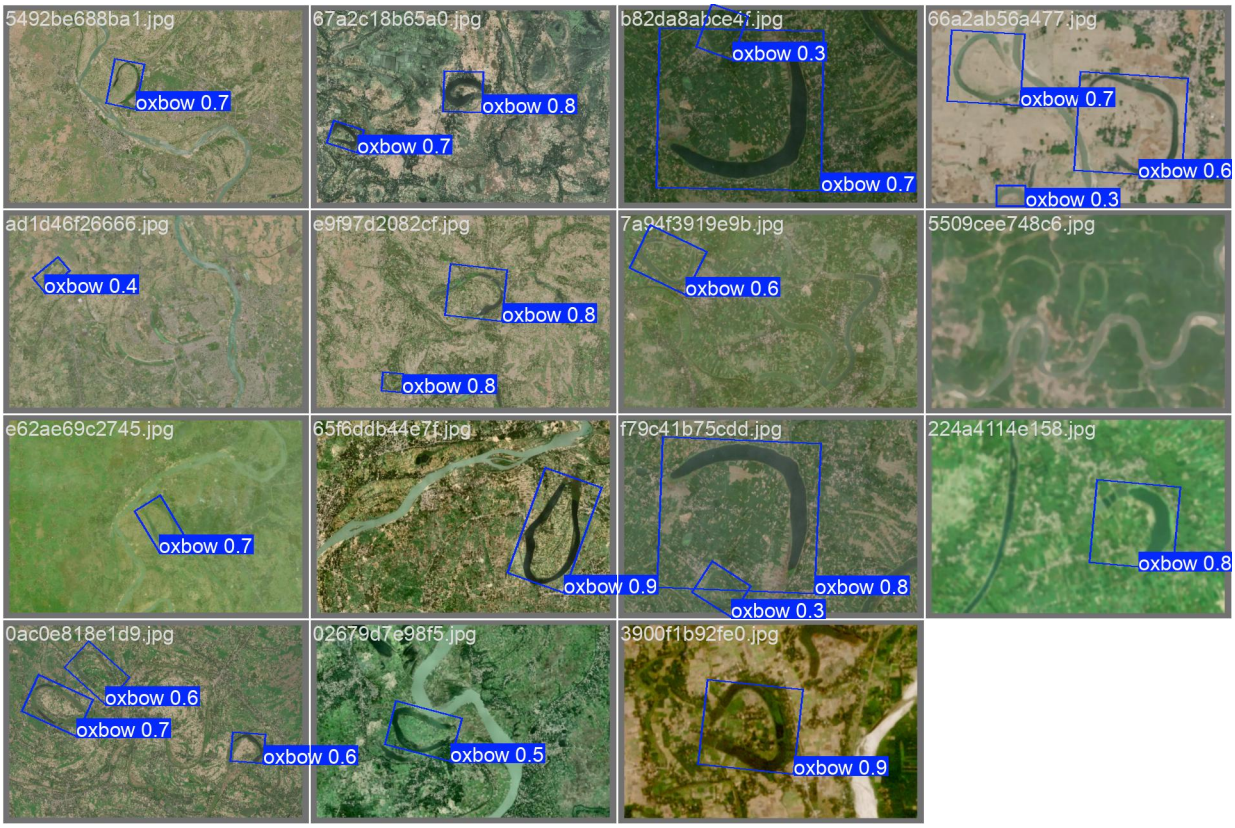
The pre-trained YOLOv11x-OBB model was fine-tuned using the Ultralytics package (version 8.3.130) in Python 3.11.12 with PyTorch 2.6+cu124. Training was conducted on an NVIDIA A100 GPU (12 compute units, 83.5 GB RAM) via Google Colab, for 200 epochs using the Adam optimiser, a batch size of 16, and a fixed random seed of 42. To enhance generalisation, various data augmentation techniques were applied, including rotation, translation, scaling, shearing, flipping, and adjustments in hue, saturation, and brightness, in order to mimic seasonal and landscape changes. These augmentations were applied in addition to the already augmented training dataset, as an extra measure to prevent overfitting given the limited number of available samples.

The fine-tuned model was evaluated on a validation set consisting of 15 images with 27 labelled oxbow lake instances. The model achieved a high detection performance with a precision of 0.953, recall of 0.815, mAP@0.4 of 0.905, and mAP@0.5:0.95 of 0.683. The model performs very well in precision and general detection accuracy; it can reliably identify oxbow lakes with few false positives. Recall is lower, meaning some oxbow lakes are missed in the detection. Both mAP scores are relatively high, showing the model can localise oxbow lakes accurately. Figure 8 illustrates this comparison between the ground-truth labels (top) and the model's predictions (bottom) on the validation set.

[2] https://labelstud.io/



(a) Ground-truth validation labels for oxbow lakes.



(b) Predicted bounding boxes for oxbow lakes by the trained YOLO model on the validation set.

Figure 8: Comparison between ground-truth validation labels (a) and model predictions (b) for oxbow lake detection.

3.3.3 Post-Processing for High-Resolution Inference: SAHI

After fine-tuning the model for detecting oxbow lakes, it will be tested on regional-scale satellite imagery viewed at an approximate altitude of 150 km to evaluate its performance in identifying previously unseen oxbow lakes across a broader contextual area. In such imagery, oxbow lakes appear as small, curved water bodies embedded within a complex riverine landscape, and are much smaller and surrounded by significantly more background noise compared to the close-range images used for fine-tuning. Therefore, testing the model directly on this large image without additional processing would likely result in poor performance, as oxbow lakes cannot be reliably identified in the full-resolution scene alone. Slicing Aided Hyper Inference (SAHI) improves object detection in large-scale and high-resolution imagery by dividing the image into smaller overlapping slices, performing detection on each, and reconstructing the results into the original image. This method enhances the model's ability to detect small features, improves scalability across varying resolutions and scene sizes, and preserves detection accuracy[55], [56]. The application and outcomes of the corresponding SAHI model are presented and discussed in detail in Chapter 5.

3.4 Gaussian Mixture Model for Risk Clustering

When oxbow lakes are identified, the next step in the pipeline is to predict the risk of arsenic in the associated point bars. Pixel-level risk classification is performed using a GMM based on the selected predictor variables, which are determined in the preliminary study conducted in Chapter 4, where auxiliary data sources are evaluated.

3.4.1 Data Sources and Retrieval

Elevation and population density are consistently used as core input variables in the GMM-based risk classification, regardless of the presence or absence of additional auxiliary data, as they form the fundamental basis for all model predictions. Instead of a DEM, a DTM is preferred as it excludes vegetation and built structures, representing of the bare-earth surface and its structures, such as ridges, for geomorphological analysis. Open-access, (semi-)high-resolution DTM data is limited or unavailable for many regions worldwide. One solution to overcome this is to use GEDI and ICESat-2 data [57], [58]. However, these datasets are not well suited for application in Bangladesh. GEDI is optimised for forested areas and often yields unreliable ground elevation estimates in floodplains and cropland-dominated regions, where the waveform is ground-dominated and canopy cover is sparse. ICESat-2 offers higher vertical accuracy, but its spatial coverage is sparse, and its track-based sampling limits utility for high-resolution terrain modelling in low-relief and hydrologically complex landscapes like the Ganges-Brahmaputra delta.

An independent researcher has developed and publicly released TessaDEM, a near-global DTM at 30 meter resolution with tree height bias correction [59]. This was achieved by combining multiple elevation data from different sources according to tree height, urbanisation, and water presence [60]–[64]. TessaDEM provides an Elevation API to retrieve the elevation from latitude and longitude coordinates. To obtain the DTM data over Bangladesh, the data was retrieved square by square within a bounding box encompassing the country's full spatial extent. The data was requested using a grid of rows and columns configured to achieve an effective spatial resolution of approximately 30 meters, based on the proportional subdivision of latitude and longitude ranges described in the TessaDEM documentation. The almost 23,000 tiles were then clipped to Bangladesh's country borders.

For population density, the most important requirement is that the data is up to date, as the goal is to produce an accurate arsenic risk prediction and to assist as many people as possible. There are several data sources available for estimating population distribution. One option considered was the WorldPop database, which provides gridded population estimates at a resolution of 100 meters [65], but this was not used in the final analysis. Instead, this study used the GHS-POP R2023A dataset from the Global Human Settlement Layer (GHSL), produced by the Human Settlement Emergency initiative under the European Union Copernicus program. This open-access spatial raster dataset represents residential population as the number of people per cell. It spans from 1975 to 2020 in five-year intervals, with projections for 2025 and 2030, and is provided at a ~90 meter resolution [66]. Population estimates are derived by disaggregating census and administrative data from GPWv4.11 [67], guided by the built-up area classifications from the GHSL-BUILT layer. For this study, the projected 2025 population density data was retrieved and clipped to the borders of Bangladesh. The original ~90 meter resolution cells were resampled to a 30 meter resolution by dividing each cell into a 3 by 3 arrangement using nearest neighbour resampling in ArcGIS Pro, ensuring alignment with the DTM resolution. For the population dataset used in Chapter 4, data from the year 2020 was selected and resampled in the same manner to closely match the temporal context of the arsenic dataset and to assess how population density correlates with other variables in the analysis.

3.4.2 Cluster Assignment and Risk Level Mapping

The GMM clustering algorithm is used to identify arsenic-risk patterns governed by topographic high grounds and densely populated areas on point bars. Since ground-truth arsenic concentrations are not always available or scarce across Bangladesh, an unsupervised learning approach is preferred to uncover inherent groupings in available proxy data that could reflect potential risk zones. Although Donselaar *et al.* [17] proposed using an RF model for predictive arsenic-risk mapping, this approach may not be suitable when applied to a DTM dataset. Despite the interpretability of RF models, particularly their ability to reveal decision thresholds, these thresholds become less meaningful across broad spatial scales, even when feature normalisation is applied. This limits the model's explanatory nature and the geographic generalisation of its outputs. While RF models offer limited and non-trivial means of estimating uncertainty, GMMs provide a probabilistic framework that enables soft classification and interpretable measures of probability based on the underlying Gaussian distributions. For this reason, GMMs were selected as the model of choice.

GMMs assume that the data is generated from a mixture of several Gaussian distributions and, as universal approximators, can effectively model complex data distributions given a sufficient number of Gaussian components (mixture). In a mixture model, each data point is assumed to be generated by one of several components, but the identity of the generating component is unknown and treated as a latent variable. The latent variable is the hidden link between the different data sources and helps model this hidden structure in the data. As a result, several simpler Gaussian models are fitted to different parts of the data. Each point receives a soft assignment after clustering, which is associated with a probability of belonging to each cluster rather than being assigned a single hard label. The model estimates the parameters of these distributions. The probability of an observation x is modelled as:

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x \mid \mu_k, \Sigma_k) \tag{1}$$

Where π_k are the mixing coefficients (weights), μ_k and Σ_k are the mean and covariance of the k-th Gaussian component, and K is the number of clusters. Bayesian Information Criterion (BIC) can be used to compare GMMs trained with different values of K. BIC evaluates the model fit while penalising model complexity, helping to avoid overfitting. Lower BIC values indicate better models. Further details on the final choice of K and the methodology for assigning clusters to risk classes are provided in Section 5.3. The Expectation-Maximisation (EM) algorithm treats latent variables as missing data. The EM algorithm optimises model parameters accordingly by approximating the latent variable by taking the posterior distribution's expected value (mean). During the E-step (Eq 2), the Gaussian components are frozen to update their parameters and the posteriors are updated (i.e. which Gaussian generated each data point?). A soft number of points is assigned to each of the components.

$$\gamma_{nk} = \mathbb{E}[z_{nk}] = p(z_k = 1 \mid \mathbf{x}_n) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x}_n \mid \mu_j, \Sigma_j)}, \quad N_k = \sum_{n=1}^{N} \gamma_{nk} \tag{2}$$

In the M-step (Eq. 3), the posteriors are frozen, and the model parameters (such as means, variances, and mixing coefficients) are re-estimated based on the expectations computed in the E-step (i.e. the newly assigned points or updated soft assignments), under the assumption that the data was indeed generated in this way and that the components were responsible for generating it. GMM clustering involves finding the maximum likelihood solution for assigning data points to components. Although the EM algorithm converges to an optimal clustering assignment once stabilised, it is sensitive to initialisation and may converge to a local optimum, potentially resulting in suboptimal or unstable clustering outcomes if poorly initialised.

For this study, the `covariance_type` parameter was set to full, which allows each component to have its own general covariance matrix, providing the greatest flexibility in capturing elliptical and rotated cluster shapes. The `init_params` setting was initialised with `kmeans`, which uses the standard K-Means initialization strategy to estimate initial responsibilities or cluster means. Other initialisation methods such as `random` and `random_from_data` were not considered promising in this context, as they tend to produce less stable or poorly fitted models.

$$\pi_k^{\text{new}} = \frac{N_k}{N}, \quad \mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n, \quad \Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^\top \tag{3}$$

The input features tend to correlate with each other, for example, high population densities on topographic high grounds. The GMM assigns higher responsibilities to one particular component in regions where the joint distribution of these input features has a distinctive peak. The GMM then tries to model the co-occurrence in the multidimensional feature space as an elliptical region. This is also one of GMM's limitations, as it struggles to model non-Gaussian, skewed, or irregularly shaped data.

4. Case Study

This case study evaluates the predictive value of various (auxiliary) data sources for arsenic risk by analysing a dataset with arsenic concentrations from 20 wells in Bangladesh, collected over a 3-year time window, from 2001 to 2003 [68]. Consecutive measurements for all wells were taken at intervals of approximately 20 to 30 days. The aim is to assess which predictors support the hypotheses proposed by Donselaar *et al.* [17], which additional variables emerge as potentially relevant despite not being previously considered, and which of the hypothesised predictors show limited or no predictive power in practice. The case study begins with a brief overview of the arsenic dataset, followed by a discussion of the (auxiliary) data sources used as predictors. Subsequently, the XGBoost method and the corresponding dataset are introduced, and the model results are interpreted. Based on the results, a final set of predictor variables is assembled, which will be used in Chapter 5, where these variables of interest are also examined in further detail.

4.1 Historical Arsenic Data from Wells in Araihazar, Bangladesh

From a 3-year survey of 6,500 households in the Araihazar Upazila, Cheng *et al.* [68] selected 20 wells for monitoring to demonstrate the integrity of the data (see Table 3 in Appendix A.2). The wells cover an area of 25 km². The wells are divided into two groups: shallow tube wells that draw water from late Holocene aquifers with grey sediments at depths of 8 to 20 meters, and deeper tube wells that tap into older, presumed Pleistocene aquifers at depths ranging from 30 to 142 meters [68]. Some of the wells are privately owned and used by a dozen or so people, while other wells serve a few hundred villagers. Furthermore, wells 4110, 4101, and 4071 were surrounded by a mix of high- and low-arsenic shallow tube wells in close proximity, while high-arsenic tube wells predominantly surround wells 84, 816, 825, and 1651. The exact locations of the wells are shown in Figure 9. Some of the wells are located on abandoned point bars, others at old levees, or on the bank of the river.

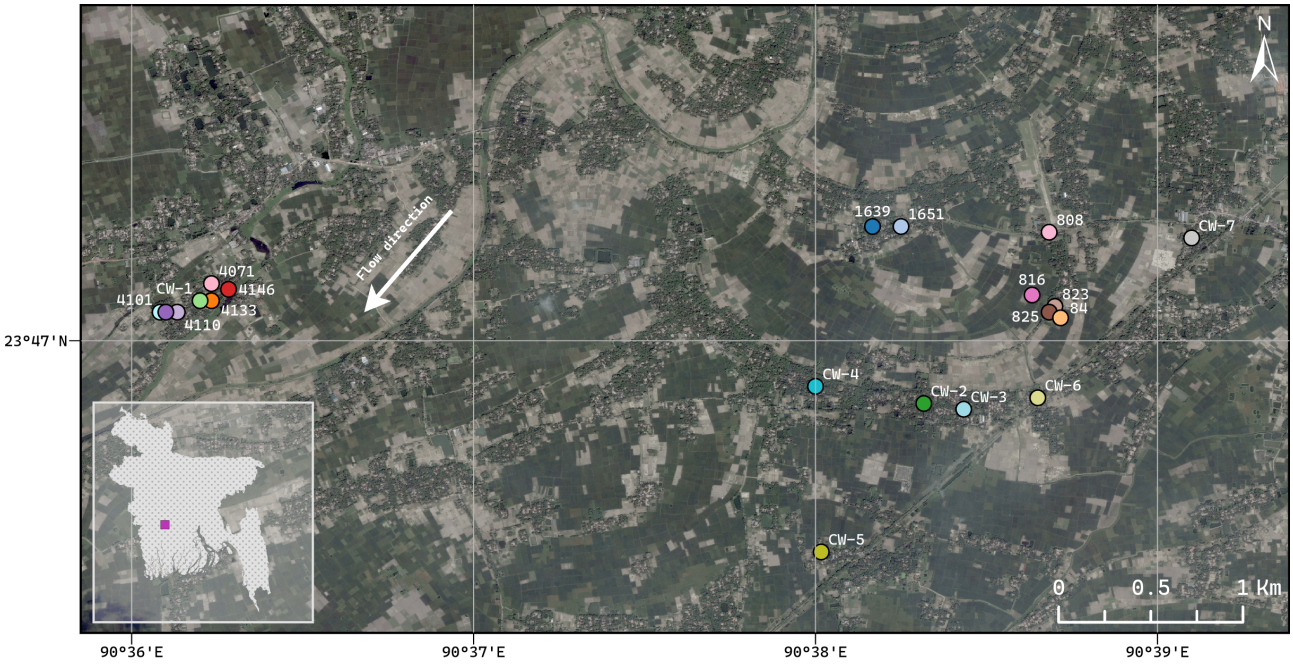


Figure 9: Spatial distribution of tube wells sampled for arsenic concentration in Araihazar Upazila, Bangladesh, overlaid on georeferenced Google Earth Pro imagery from April 2002. Each coloured marker represents a tube well [68]. Coordinates: 23°46′48.00″N, 90°38′26.00″E. Image exported via Google Earth Pro. Map data: Google, © 2025 Maxar Technologies.

As seen in Figure 10, the temporal variability of the arsenic concentrations in the tube wells remains stable over time with a few exceptions. This provides partly supportive evidence for the hypothesis by Donselaar *et al.* [17] that geomorphological features such as point bars act as persistent arsenic sinks and serve as arsenic hotspots where concentrations remain consistently high due to the local geological setting. Moreover, the available data suggest a tendency for higher arsenic concentrations in shallow aquifers compared to deeper ones. However, this observation should be interpreted with caution due to the limited sample size. In response to Cheng *et al.* [68] research, Ravenscroft *et al.* [69] rebut Cheng *et al.*'s assertion that arsenic concentrations in wells remain stable over time, arguing that the limited monitoring time is insufficient for drawing such definite conclusions, and that it may give well users a false sense of security ("safe" wells remain safe). Ravenscroft *et al.* [69] referenced studies where arsenic concentrations in tube wells increased over time. Detailed information about some of these sampling locations in these studies is unknown. Therefore, it is challenging to assess how Ravenscroft *et al.* [69] rebut is substantiated, given the lack of detailed contextual information about the sampling locations and conditions. However, Ravenscroft *et al.* [69] statement regarding increasing arsenic concentrations over time aligns with Donselaar *et al.* [17] findings. Since the rate of reductive dissolution and the subsequent release of arsenic into groundwater is controlled by organic carbon availability, McArthur and Cheng may be correct within their respective spatial contexts. Without detailed spatial and sedimentological information about the monitored wells, whether the observed stability or increase in arsenic concentrations reflects local biogeochemical conditions remains uncertain. This ambiguity highlights the importance of considering geomorphological features, such as point bars, which act as arsenic hotspots where reductive conditions may persist over time.

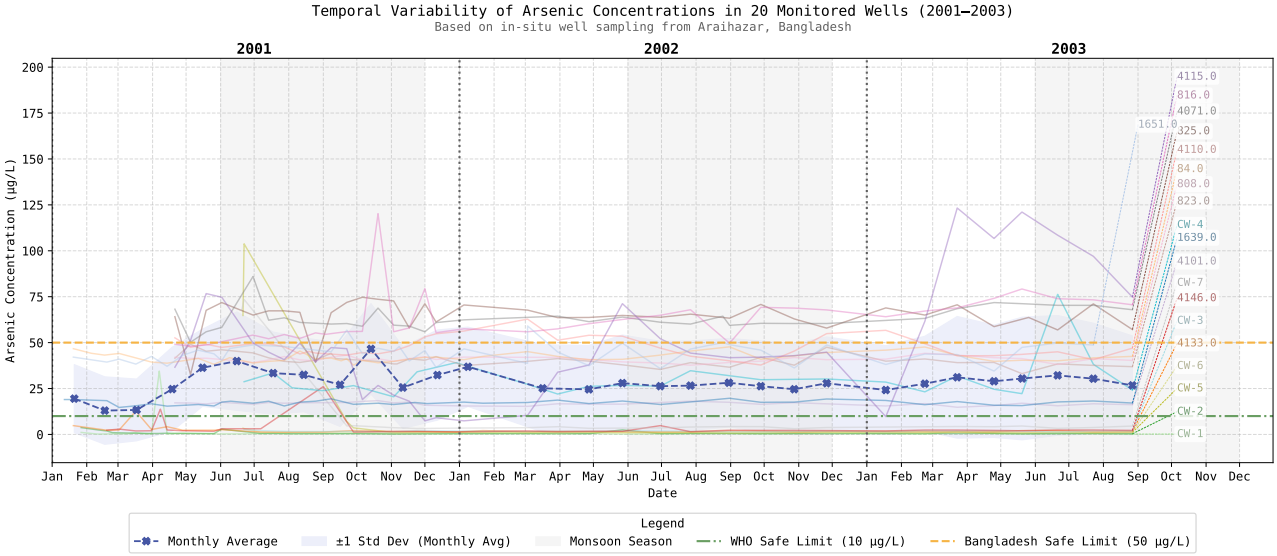


Figure 10: Temporal variability of arsenic concentrations in 20 monitored wells in Araihazar Upazila, Bangladesh, from 2001 to 2003 [68].

4.2 (Auxiliary) Data Sources as Indicators of Arsenic Risk

Donselaar *et al.* [17] outline in their study 4 predictor variables for arsenic risk prediction: local climate variables, more specifically precipitation, topographic elevation, population density, and oxbow lake vegetation indices. Topographic elevation and population density have already been addressed in Chapter 3. The other two variables, precipitation and oxbow lake vegetation indices, are less straightforward.

According to Donselaar *et al.* [17], precipitation drives lateral groundwater flow, flushing arsenic-rich water from non-confined sedimentary units with high permeability, and floods oxbow lake vegetation. Given that monsoonal precipitation is a recurring annual event and that high-temporal-resolution arsenic datasets are limited, the reliability of precipitation as a predictive variable is uncertain. While climate change may lead to more extreme rainfall events, the general consistency of monsoon patterns and the persistence of elevated arsenic concentrations in certain areas make it difficult to justify precipitation as a strong explanatory factor for temporal variability in arsenic levels.

As discussed in Chapter 2, there are cases where increased precipitation can enhance arsenic mobilisation, suggesting that precipitation is indeed an important contributing factor. Therefore, it will be included as a predictor in the XGBoost model; however, due to the complexity of the underlying processes and limited temporal data, its precise impact remains difficult to quantify. Total precipitation (tp) data can be obtained from the ERA5-Land Monthly Averaged Reanalysis dataset in the climate datastore, openly available online, by the European Union Copernicus program and ECMWF [70]. The data's spatial resolution is 9 km. Total precipitation accounts for the accumulated liquid and frozen water over a particular period.

Oxbow lake vegetation indices, proposed by Donselaar *et al.* [17] as proxies for organic carbon density, can be derived using various variables. The Normalised Difference Vegetation Index (NDVI) is widely used in remote sensing to monitor and assess vegetation presence, biomass [71], [72], and health from satellite imagery. NDVI is simple to compute and strongly correlates with photosynthetically active biomass. NDVI is calculated using the red and near-infrared (NIR) bands of the EM-spectrum, as such:

$$\text{NDVI} = \frac{(\text{NIR} - \text{Red})}{(\text{NIR} + \text{Red})} \tag{4}$$

NIR reflects strongly from healthy vegetation due to internal leaf structure, and red is absorbed by chlorophyll during photosynthesis. High NDVI values over oxbow lakes may suggest the presence of macrophytes. Accurate measurement of NDVI within oxbow lakes before and after the monsoon period requires precise delineation; however, in this study, oxbow lakes are identified but not explicitly delineated. Additionally, the XGBoost model requires tabular input, with each row representing a single arsenic measurement. Unlike variables such as DTM, assigning an appropriate NDVI value to each data point is not straightforward. NDVI's temporal variability and the lack of exact sampling dates for many arsenic measurements make it difficult to ensure meaningful correspondence. Yet, NDVI will be used as a predictor variable in the XGBoost model. Instead of measuring the NDVI for the oxbow lake, the NDVI is measured for the entire oxbow lake and the point bar within. NDVI can be computed from every optical satellite imagery with the red and NIR bands. Key considerations include cloud-free conditions, as clouds and shadows can distort imagery, temporal alignment with the period of interest, and the frequency of revisiting the satellite. The Landsat 7 satellite is the only satellite with the correct temporal alignment and has a revisit time of 16 days. Google Earth Engine provides the cloud-reduced Landsat Collection 2 Tier 1 Level 2 NDVI composites, which are created from all the scenes in each 32-day period beginning from the first day of the year and continuing to the 352nd day of the year[3]. However, these 32-day intervals do not align precisely with calendar months, which may limit temporal comparability with monthly datasets or field measurements. Additionally, Landsat 7 experienced a scan line corrector failure in late May 2003, resulting in gaps in the imagery. Despite these limitations, this dataset is used in the present study due to its ease of access, time efficiency, and suitability for a proof-of-concept analysis.

4.2.1 Auxiliary Data Sources

Chapter 3 identified four categories of variables commonly used in previous studies to predict arsenic concentrations and associated risk. A subset of these variables is selected for inclusion in this case study. Some variables, such as soil type, are inherently represented by geomorphological features like point bars. Nonetheless, this case study incorporates soil-related variables to examine their contribution to arsenic risk prediction. Using machine learning practices and national and local geographical databases, SoilGrids mapped the globe's soil properties at a 250 m spatial resolution [73]. The soil maps are not tied to a specific year but result from integrating soil profiles collected over multiple years; therefore, they provide a comprehensive, up-to-date representation of global soil properties rather than a snapshot from a particular year. The SoilGrids 250 m v2.0 dataset is available through the GEE Community Catalogue [74]. In this case study, the following variables will be used: sand, silt, organic carbon density (OCD), soil organic carbon content in the fine earth fraction (SOC), and organic carbon stocks (OCS). To align with the 30 meter resolution of the DTM, the original 250 meter cells were resampled by dividing each cell into smaller units using a nearest neighbour approach in ArcGIS Pro, resulting in a uniform spatial resolution of 30 meters across all datasets.

[3] `https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_COMPOSITES_C02_T1_L2_32DAY_NDVI`

The three recurring variables in prior research are Topographic Position Index (TPI), Topographic Wetness Index (TWI), and Land Use and Land Cover (LULC). Examples of how TPI and TWI relate to the elevation profile are visualised in Figure 12, with their spatial context shown in Figure 11. The TPI can conveniently be calculated from the DTM, as it is geomorphometric and quantifies a location's relative position on a landscape (e.g. whether a point is in a valley, on a slope, or on a ridge). TPI is calculated by comparing the elevation of a focal pixel to the mean elevation of surrounding pixels within a defined window, as such:

$$\mathrm{TPI}(x, y) = z_{x,y} - \hat{z}_{neighbourhood} \tag{5}$$

With $z_{x,y}$ the elevation at the focal point, and $z_{neighbourhood}$ the average elevation in a surrounding area. If TPI > 0, the location is higher than its surroundings and typically indicates a ridge or hilltop. If TPI < 0, the location is lower than its surroundings, such as a valley or channel. A TPI value close to 0 suggests a flat area or a uniform slope. TPI can help identify catchment areas and water routing and often strongly correlates with geomorphological features; for example, low TPI areas correspond to floodplains, riverbeds, and alluvial plains. From a sedimentology perspective, low TPI areas, such as point bars, are characterised by sediment deposition.

The TWI is a hydrological terrain index that quantifies how likely water will accumulate at a specific location. TWI uses the landscape's topographic features such as slope and upstream contributing area for modelling how water flows and accumulates over terrain, as such:

$$\mathrm{TWI} = \ln\left(\frac{\alpha}{\tan\beta}\right) \tag{6}$$

The formula reflects the balance between the flow accumulation (α) and the drainage capacity (steepness of slope, β). Steeper slopes have faster water run-off. The natural logarithm is used to normalise the index. High TWI values indicate wet areas prone to water accumulation, whereas low TWI values are usually dry and well-drained areas such as ridges and steep slopes. Primary use cases of TWI are mapping soil moisture and saturated zones, identifying wetlands and drainage networks, and predicting surface run-off and flood risk areas. High TWI values often correspond with low-energy zones where fine sediments settle, and the soil moisture is high.



Figure 11: Aerial image of a meander belt in Bangladesh showing the location of the elevation transect (A–B) across a point bar, analysed in Figure 12. Image taken from Google Earth Pro, dated 1 September 2025. Coordinates: 23°08′27.10″N, 89°42′42.29″E. Map data: Google, © 2025 Maxar Technologies.

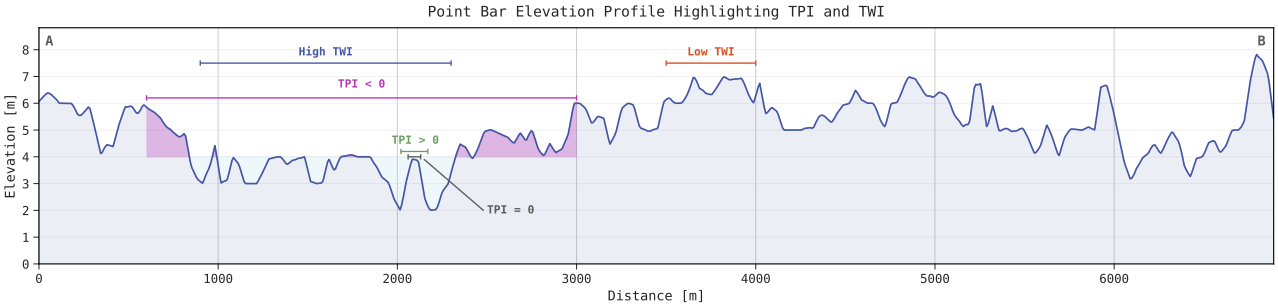


Figure 12: Elevation profile across a point bar, highlighting variations in TPI and TWI. Labels A and B correspond to the endpoints of the transect shown in white in Figure 11

Land use and land cover (LULC) is included as an explanatory variable due to its strong correlation with population density, as well as its influence on land management, water use, and environmental risk exposure. The Global Land Analysis & Discovery (GLAD) project provides a global LULC change dataset that quantifies transitions across various land cover classes, such as urban areas, agricultural land, forests, and wetlands, from 2000 to 2020, at a 30 meter spatial resolution [75]. For this case study, the LULC data from the year 2000 is used and retrieved from GEE. For each data point, the LULC value that is most represented within a radius of three cells is considered.

4.3 Testing (Auxiliary) Data Sources for Arsenic Risk Prediction

The subsection aims to test and verify the contributions of each of the (auxiliary) data sources, as described in previous sections, to arsenic risk prediction. First, a brief overview of the complete dataset, including variables, is provided. Second, the XGBoost model and the testing setup are explained. And lastly, results from the XGBoost model are interpreted and compared to the hypothesis posed by Donselaar *et al.* [17].

The dataset Cheng *et al.* [68] provided contains 681 entries from 20 wells sampled over a 3-year period. For each point in the dataset, the associated topographic elevation, TPI, TWI, OCD, OCS, SOC, sand, silt, LULC, tp, NDVI, and population density value were assigned. For population density, the sum within a 30 meter radius around each well was used. Different radii (30, 90, 150, and 210 meters) were tested, with the 30 meter radius yielding the best model performance. Since the objective is to predict arsenic risk rather than actual concentrations, each arsenic measurement is converted into a binary risk level. A value of 0 indicates no risk for concentrations below 10 µg/L, while 1 indicates risk for concentrations equal to or above 10 µg/L. Before training the XGBoost model, the dataset is preprocessed by splitting it into training and testing subsets based on Well ID, ensuring that all measurements from a given well are assigned exclusively to either the training or the testing set. To reduce the risk of spatial data leakage, a minimum distance of 300 meters between wells in different subsets is maintained. The data is randomly split into 70% training and 30% testing, with measurements from 14 wells assigned to the training set and those from 6 wells to the test set. This resulted in a training set of 479 samples and a test set of 202 samples. Of these, 288 training samples and 153 test samples were classified as high-risk, highlighting the overall class imbalance, with 441 out of 681 total samples labelled as high-risk. Stratified K-Fold splitting with five folds is applied to the training data to maintain a similar distribution of the binary target variable in each fold and reduce class imbalance during cross-validation. Additionally, the SAND, SILT, OCD, and SOC variables each consist of six bands representing depths of 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm. The OCS variable includes a single band covering the 0–30 cm depth range. All combinations of the soil variables, together with the other predictors, were considered and tested during the development of the XGBoost model (combinations labeled as sets 1 to 6, with set 1 depth 0-5 cm).

To optimise the performance of the XGBoost classifier, a randomised search was conducted over a predefined grid of hyperparameters. The search space included the following: `n_estimators` [195, 200], `learning_rate` [0.0475, 0.05], `max_depth` [3], `min_child_weight` [4], `subsample` [0.5], `colsample_bytree` [0.5], `reg_alpha` [0.1, 0.2, 0.4], and `reg_lambda` [0.1, 0.5, 1]. Cross-validation was carried out using the previously defined stratified training folds. The model was trained with a fixed `random_state` of 42 to ensure reproducibility, and all implementations were carried out using the Scikit-learn API with the XGBoost library in Python environment version 3.10.

The best-performing model was selected based on the highest balanced accuracy, which accounts for class imbalance by averaging the recall across all classes. The optimal hyperparameters found were: `subsample` = 0.5, `reg_lambda` = 0.1, `reg_alpha` = 0.2, `n_estimators` = 200, `min_child_weight` = 4, `max_depth` = 3, `learning_rate` = 0.05, and `colsample_bytree` = 0.5. This model, trained using soil depth features from 0–5 cm, achieved a training accuracy of 0.979 and a test accuracy of 0.812, resulting in an accuracy gap of 0.167. The precision was 1.000, recall was 0.752, and the F1 score was 0.858. As shown in Figure 13, the training and validation log loss curves are plotted for each feature set corresponding to different soil depths, using the hyperparameters specified above. The feature set based on the 0–5 cm soil depth consistently achieved the lowest validation loss, indicating better generalisation performance. In contrast, models using deeper soil features showed higher validation loss and signs of overfitting, as reflected by the increasing gap between training and validation losses.

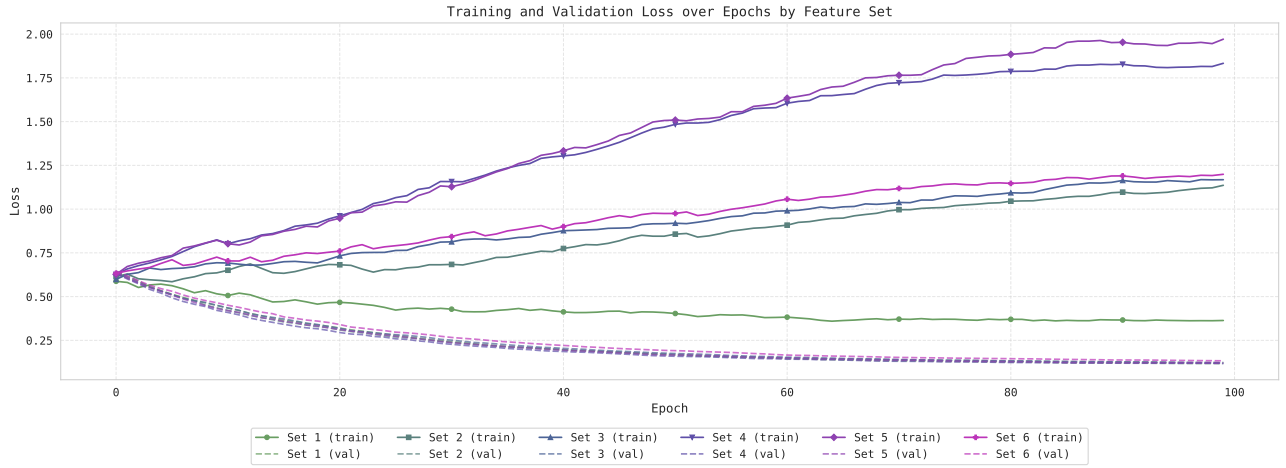


Figure 13: Training and validation log loss over 100 epochs for different soil depth feature sets. Each line represents a distinct soil depth range (Set 1 through Set 6), with solid lines indicating validation loss and dashed lines indicating training loss. Set 1 (0–5 cm) shows the lowest validation loss and smallest train-validation gap. Deeper soil sets exhibit progressively higher validation losses and larger gaps, suggesting increased overfitting.

These results indicate that the model generalised well to unseen data, as demonstrated by a test accuracy of 0.812 and a strong F1 score of 0.858. The perfect precision score (1.000) suggests that all locations predicted as high-risk were indeed high-risk in the test set. However, the recall score of 0.752 indicates that the model failed to identify a number of high-risk locations. This is reflected in the confusion matrix (Figure 15), where 38 out of 191 high-risk cases were misclassified as no-risk. No false positives are present, confirming that the model is highly reliable in predicting high-risk areas. This also suggests a conservative prediction tendency that prioritises precision over recall. While this reduces the likelihood of triggering unnecessary mitigation efforts in low-risk areas, it increases the risk of overlooking actual high-risk cases, which is not desirable in real-world scenarios due to potential public health consequences. The model correctly classified all low-risk cases. The accuracy gap of 0.167 between training and testing performance suggests some level of overfitting, but the model still performs robustly overall, given the imbalanced class distribution and the complexity of the input features.

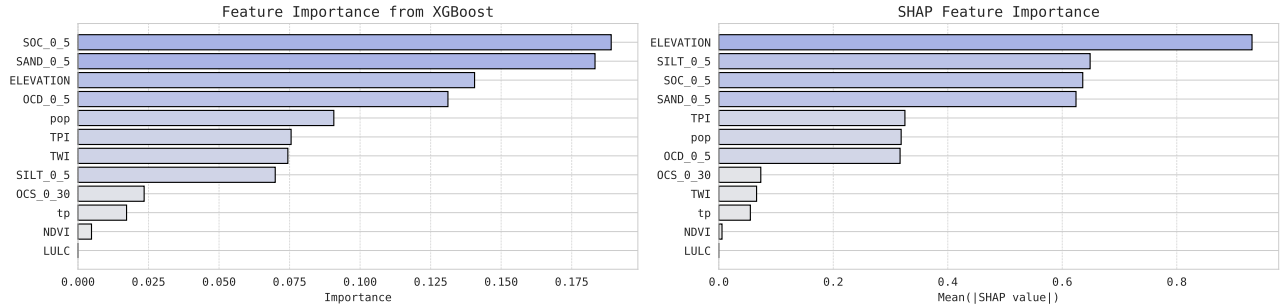


Figure 14: Feature importance analysis for the best-performing XGBoost model using Set 1 variables (0–5 cm soil depth). **Left:** Model-derived feature importance. **Right:** SHAP-based feature importance.

The feature importance plot on the left in Figure 14, highlights which variables contributed most to the model's decision-making process, offering insights into which predictors were most influential in identifying high-risk areas. SOC, sand, and elevation were the top three most effective variables for the model's performance in reducing classification error during training. NDVI, tp, and LULC contributed little to nothing in the model's decision-making. The correlation plot on the right in Figure 15 illustrates the linear relationship between the variables, calculated over the entire dataset used in the case study. The features tp and NDVI show little to no correlation with other variables, either because they are inherently uncorrelated or because their relationships with other features are non-linear and therefore not captured by the correlation matrix. LULC is omitted because it is a categorically encoded variable represented with numerical values, which makes correlation analysis inappropriate. Population density has virtually no correlation with silt content, which aligns with geomorphological expectations, as settlements are typically located on sandy ridges rather than in fine-grained floodplain environments. Additionally, organic carbon density (OCD_0_5) is strongly negatively correlated with sand content, further supporting the interpretation that sandy areas, where people tend to reside, have lower organic matter accumulation. This spatial pattern of habitation and sediment characteristics is consistent with the observations described by Donselaar *et al.* [17]. Sand content does not show any meaningful correlation with elevation and exhibits only a very weak negative correlation with population density. According to the conceptual model described by Donselaar *et al.* [17], stronger positive correlations between sand, elevation, and population density would be expected, assuming linear relationships. However, the dataset used in this study is compiled from various sources spanning a broad time range, from 1905 to 2016, which may not accurately reflect the present-day geomorphological and settlement patterns of the region. In particular, the data may not align well with the spatial distribution of sandy ridges that have formed and evolved over time. Additionally, no correlation is observed between elevation and population density, although this too would be anticipated based on the theoretical framework. One possible explanation for this discrepancy is the method used to assign population values: the sum within a fixed-radius buffer was used rather than the mean, and the chosen radius may have been either too small or too large to capture meaningful spatial associations. Arsenic (As) exhibits small negative correlations with both population density and elevation, which contradicts the conceptual model proposed by Donselaar *et al.* [17]. Their model suggests that arsenic hotspots are more likely to occur on elevated sandy ridges where people tend to settle. The observed negative correlation with population density may reflect limitations in how the variable was derived, as the use of summed population values within a fixed-radius buffer might not accurately capture the spatial distribution of settlements around the wells. Additionally, these results indicate that the relationship between arsenic concentration and these variables is likely non-linear or more complex than what can be captured by simple correlation measures.

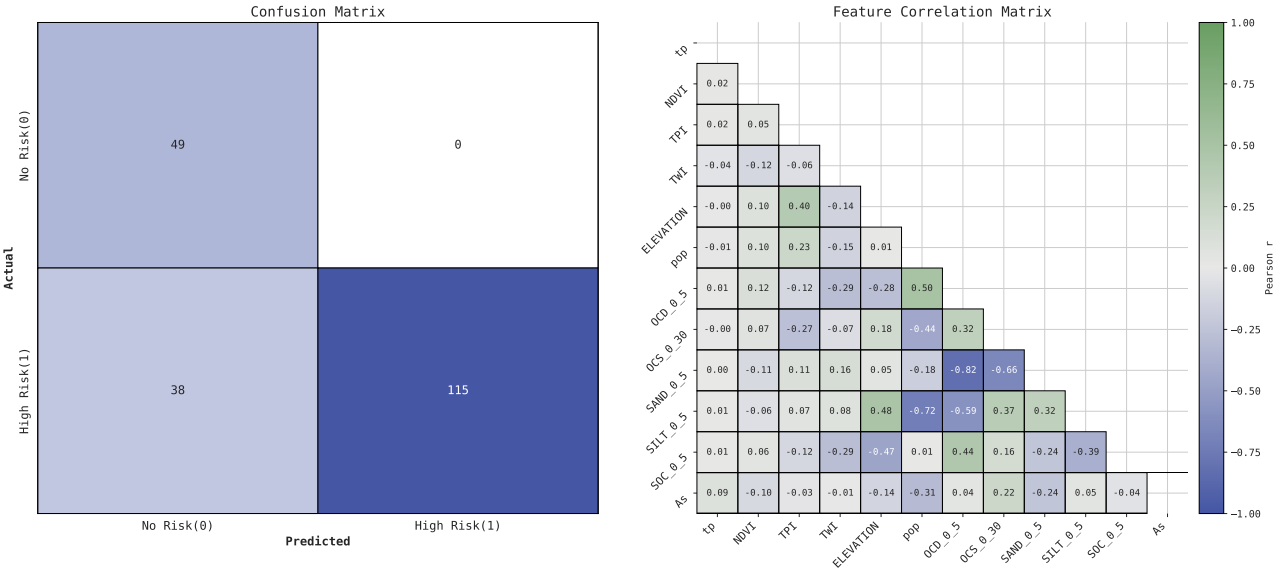


Figure 15: Model evaluation and feature correlation analysis for the best-performing XGBoost model using Set 1 variables (0–5 cm soil depth). **Left:** Confusion matrix. **Right:** Pearson correlation matrix of the entire dataset.

SHAP (SHapely Additive exPlanations) values are computed using the best-performing classifier to interpret the trained XGBoost model. SHAP quantifies the predictive impact of each variable within the trained model. SHAP values capture both linear and non-linear interactions between features. The SHAP feature importance plot (on the left of Figure 14) shows that elevation, SILT_0_5, SOC_0_5, and SAND_0_5 are the most influential predictors in the model. These findings are partially reflected in the correlation matrix (Figure 15), where some of these features exhibit moderate to strong correlations with each other. These global importance rankings are further illustrated in the SHAP waterfall plot (Figure 16), which shows how individual features contribute to specific model predictions. In this case, elevation, SOC_0_5, and population density contribute positively, pushing the prediction above the risk threshold.

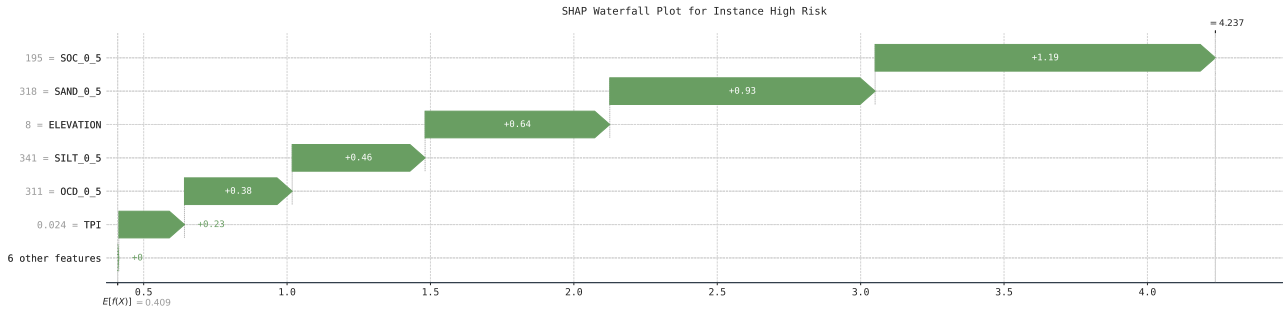


Figure 16: SHAP waterfall plot for the best-performing XGBoost model using Set 1 variables (0–5 cm soil depth). The plot shows how individual features contribute to a single high-risk prediction, with SOC_0_5, SAND_0_5, and elevation having the strongest positive influence.

4.4 Model Interpretation and Feature Relevance

The case study supports the hypothesis proposed by Donselaar *et al.* [17] that geomorphological features, particularly sandy and elevated point bars, serve as persistent arsenic hotspots. This interpretation is reinforced by the high importance of sand, elevation, and SOC in both the XGBoost feature importance and SHAP analyses. These findings align with the notion that topographically elevated, sandy ridges within point bars are preferred locations for human settlement and are strongly associated with an elevated risk of arsenic exposure. In contrast, variables such as NDVI and precipitation (tp) contributed minimally across all feature importance metrics. This may indicate limited predictive value in the context of this dataset or reflect limitations in the spatial or temporal resolution of these variables. As a result, NDVI and tp will not be included in the subsequent model implementation. While this case study is not without limitations, with known limitations in sampling, temporal coverage, and the construction of certain features, the overall results are consistent and sufficiently robust to inform further modelling. The agreement between model outputs, theoretical expectations, and the geomorphological context provides sufficient support to proceed with the selected variables. The next chapter will focus on elevation and population density, as these variables demonstrated theoretical relevance, consistent model contribution, and practical interpretability for arsenic risk prediction.

5. Model Implementation and Results

This chapter presents the practical implementation of the machine learning approaches, as discussed in Chapter 3, to predict arsenic-risk. Building on the conceptual frameworks introduced in earlier chapters, the implementation focuses on the two main components: 1) the detection of geomorphological features using the YOLO model applied to satellite imagery, and 2) the prediction and classification of arsenic-risk based on an GMM trained on geospatial and environmental features. For both parts, the results of the models are evaluated using domain knowledge (e.g. are the results geologically meaningful). Before testing the model implementations, a new historical arsenic dataset is introduced. This dataset will be used throughout the chapter.

5.1 Historic Arsenic Dataset for Model Implementation Testing

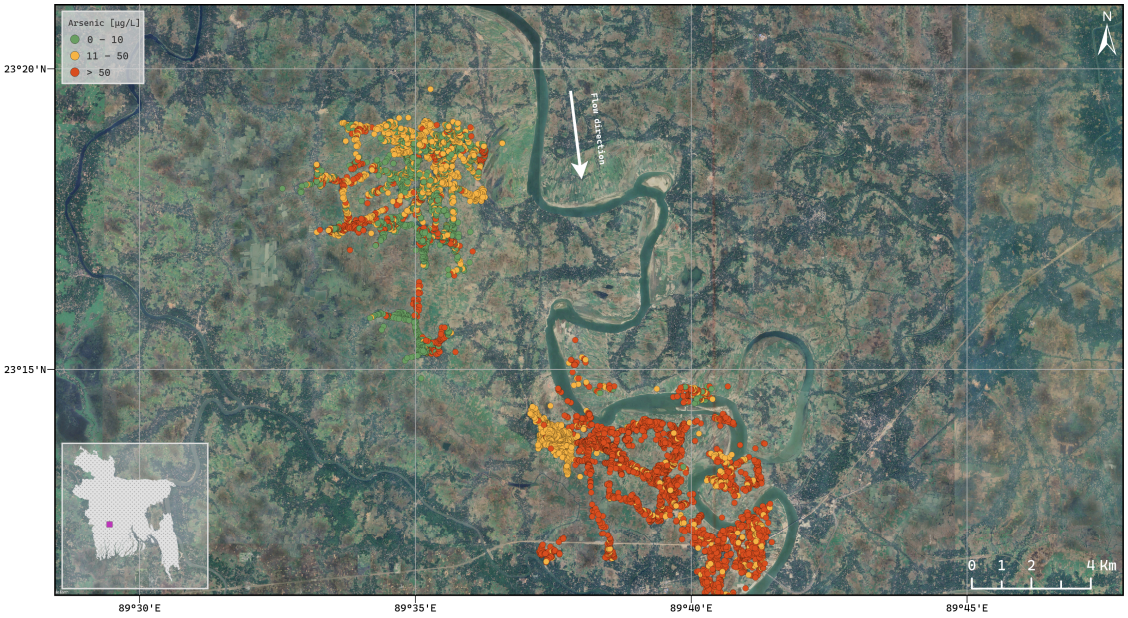
A new historic arsenic dataset, consisting of almost 20,000 unique records (sampling locations), provided by Donselaar [76], will be used to implement both parts of the machine learning pipeline. The historical arsenic dataset samples were taken in the Upazilas of Kalia and Lohagara at the end of 2013 and the start of 2014. Figures 17 illustrates the spatial distribution of the historic arsenic sampling locations across the Upazilas of Kalia and Lohagara. Compared to the dataset used in the Chapter 4, this dataset covers a broader geographic area within a similar geomorphological setting and includes an order of magnitude more data samples; however, it does not include repeated measurements from the same wells over time. Figure 22 shows the relationship between well depth and measured arsenic concentrations, and their distribution, with the mean and standard deviation for both variables indicated.

5.2 Testing Oxbow Lake Detection Using a Trained YOLOv11x-OBB

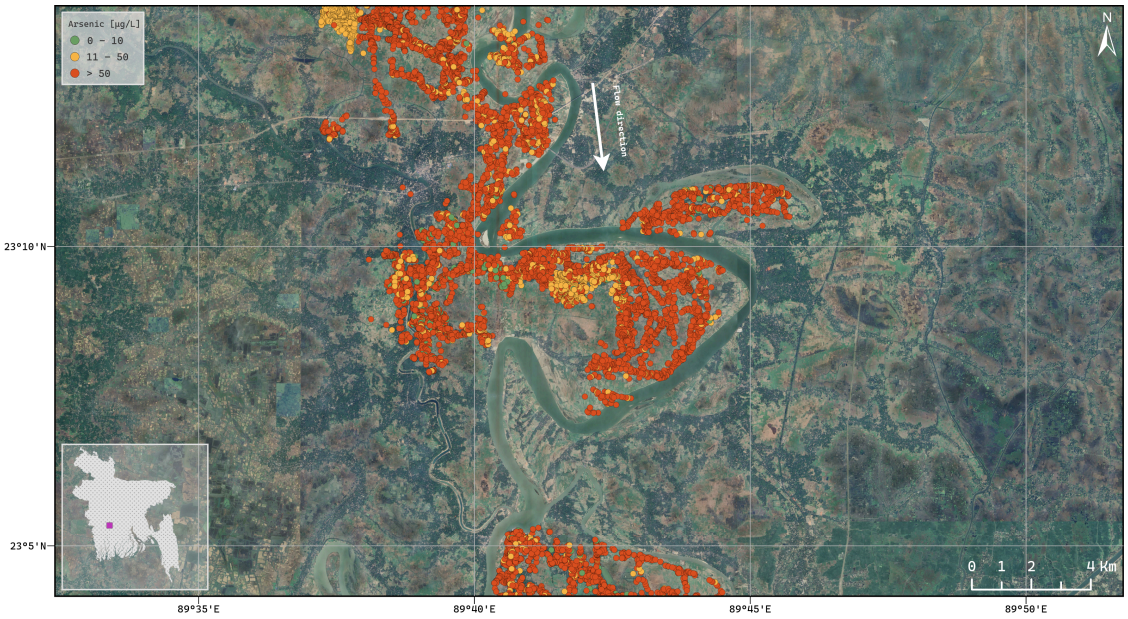
In this section, the trained YOLO model, from Section 3.3, will be used for inference on a regional-scale satellite image in the area defined by the new dataset, using YOLO SAHI. The objective of this section is to evaluate the performance of the trained model on large geospatial areas that include oxbow lakes. Specifically, the assessment focuses on the model's ability to accurately detect oxbow lakes within their broader landscape context. Section 3.3 introduced the use of regional-scale imagery and specified the image scales used for training. In this section, satellite images acquired under varying spatial scales, landscape settings, and seasonal conditions are evaluated to assess the generalisation performance of the trained YOLO model across different environmental settings. Images for the YOLO inference testing are taken from Google Earth Pro in HD resolution (1920x1080). The eight test cases used in this evaluation are summarised in Table 1, for different scales and geomorphological settings. The model inference results across all eight test cases are visualized in Figure 18 and 19. Since the model is tested on previously unseen scenes for which no ground-truth annotations are available, evaluation using standard quantitative metrics is not feasible. Therefore, visual inspection of inference results is provided, for example, by a domain expert evaluating detection accuracy under different environmental settings.

Date	Scale	Obs. Altitude (km)	Season / Setting	Test Case
May 2025	Regional	114.77	Wet season (isolated oxbow lakes, side-channel river)	1
		114.77	Wet season (dominant river system, few oxbow lakes)	2
		87.47	Wet season (isolated oxbow lakes, side-channel river)	3
		87.47	Wet season (dominant river system, few oxbow lakes)	4
Dec 2016	Regional	111.10	Dry season (isolated oxbow lakes, side-channel river)	5
		111.10	Dry season (dominant river system, few oxbow lakes)	6
		87.47	Dry season (isolated oxbow lakes, side-channel river)	7
		87.47	Dry season (dominant river system, few oxbow lakes)	8

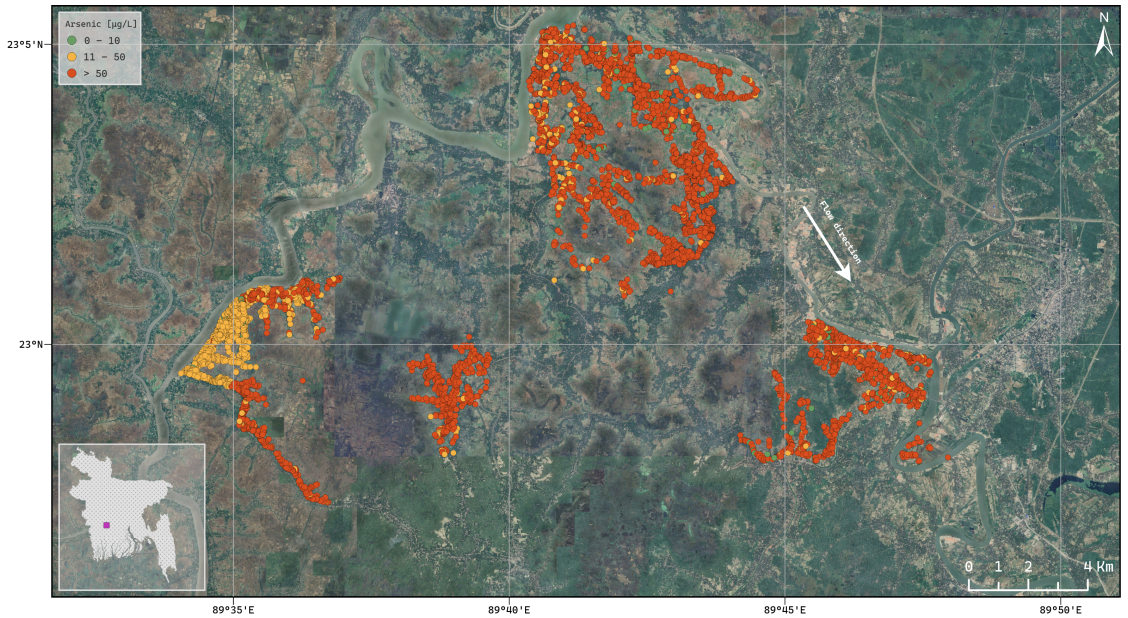
Table 1: Overview of test images used for evaluating YOLO model performance across different observation altitudes and geomorphological settings.



(a) Northern section

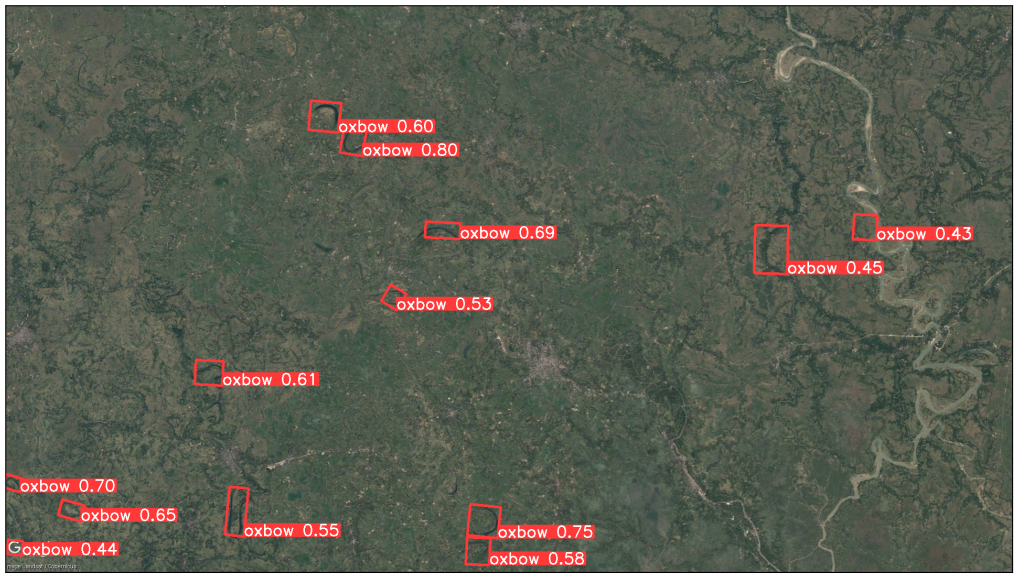


(b) Central section

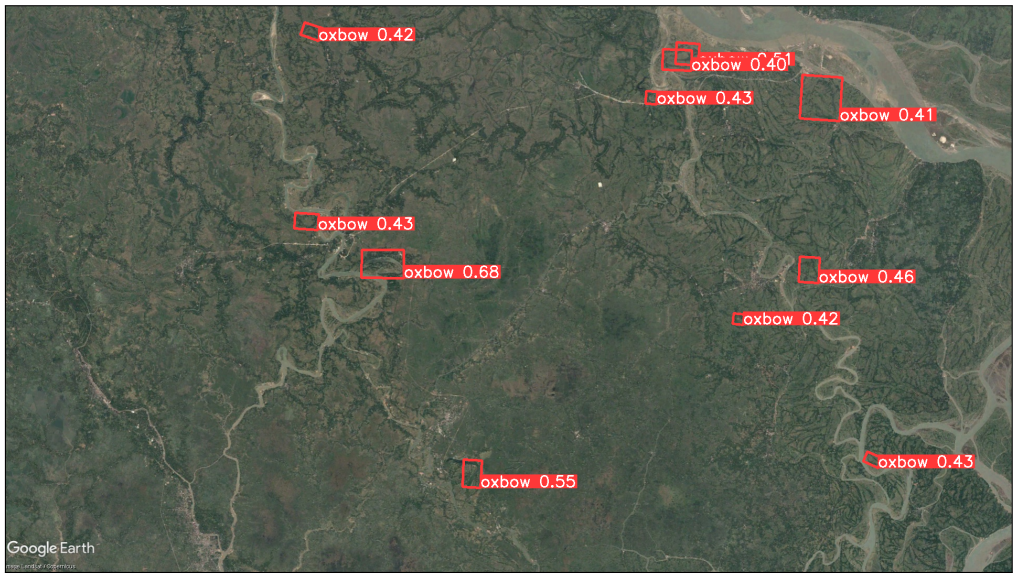


(c) Southern section

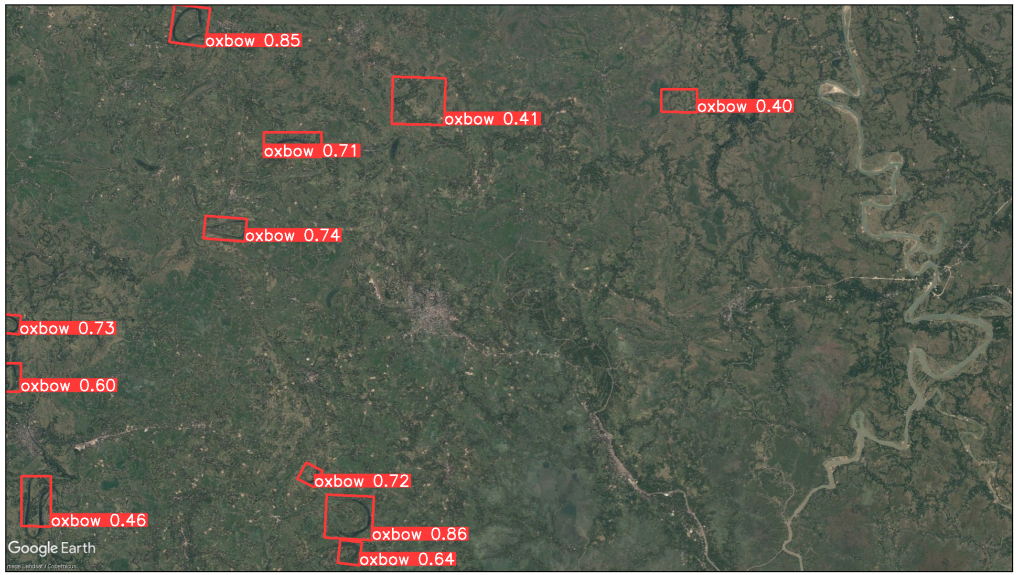
Figure 17: Spatial distribution of historic arsenic sampling locations across the Upazilas of Kalia and Lohagara. The three panels represent a north-to-south segmentation of the study area. The satellite imagery was obtained from Google Earth Pro and georeferenced, captured in May, 2025. Map data: Google, © 2025 Maxar Technologies.



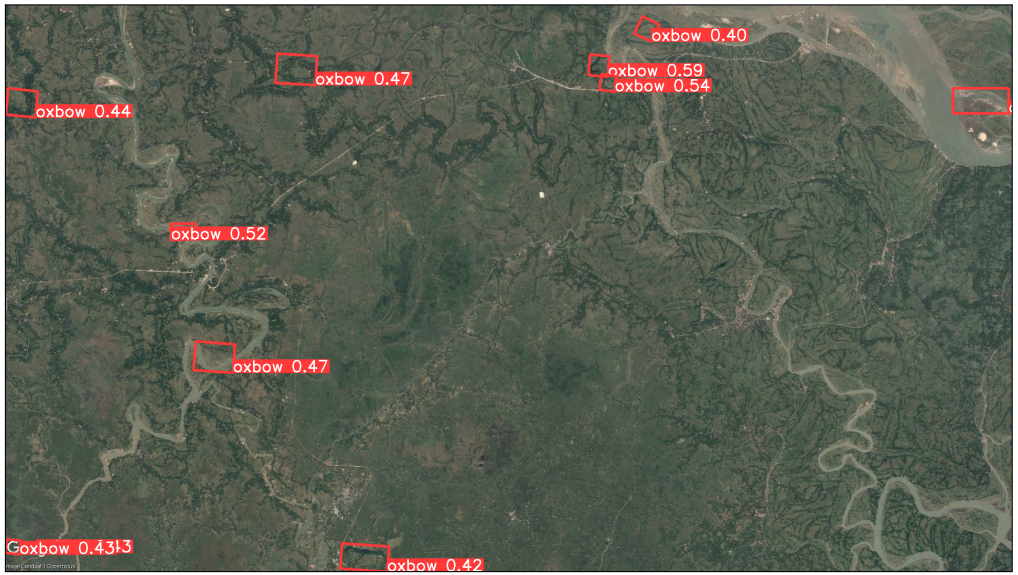
(a) Test case 1: Wet season, isolated oxbows



(b) Test case 2: Wet season, dominant river

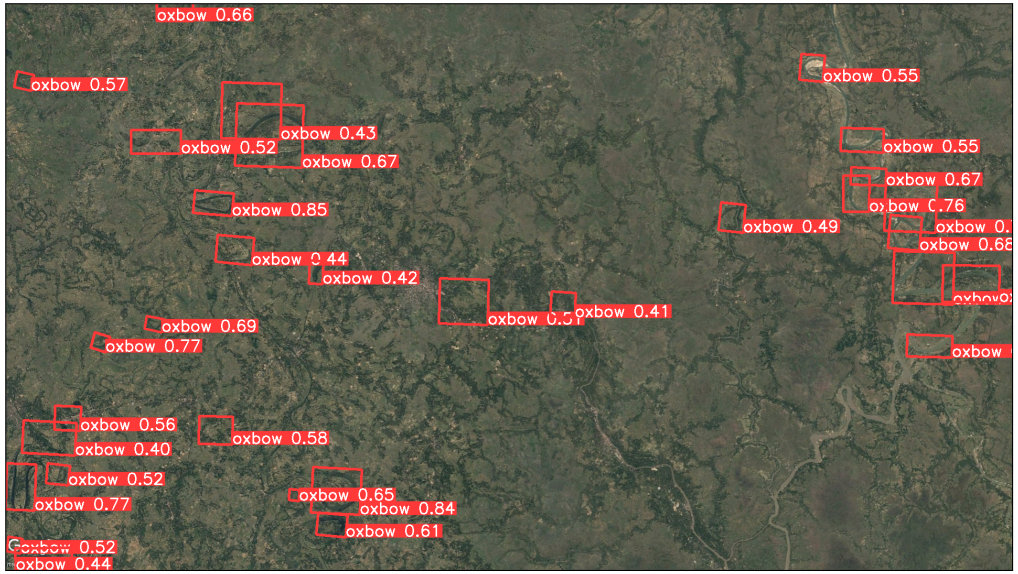


(c) Test case 3: Wet season, isolated oxbows

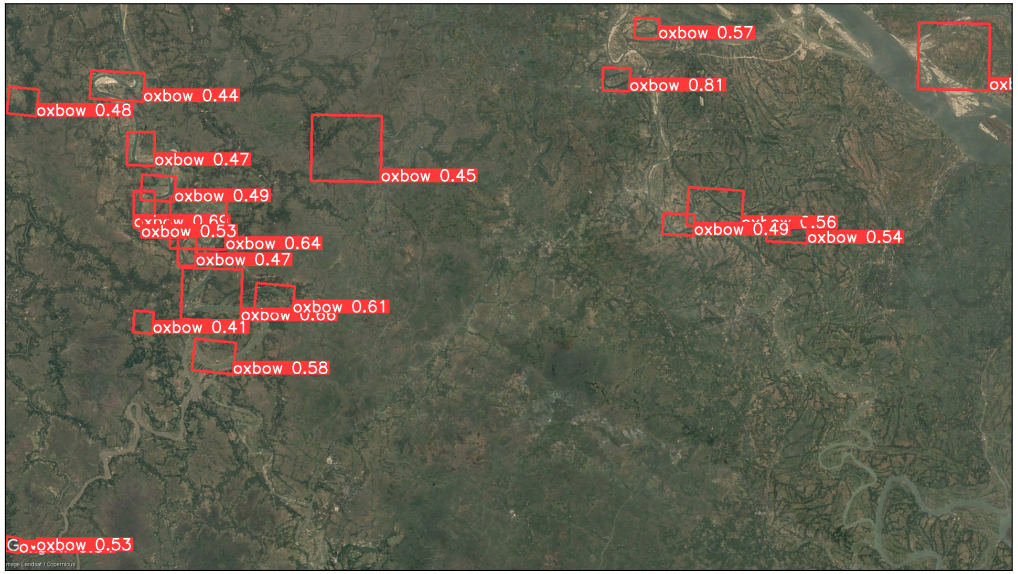


(d) Test case 4: Wet season, dominant river

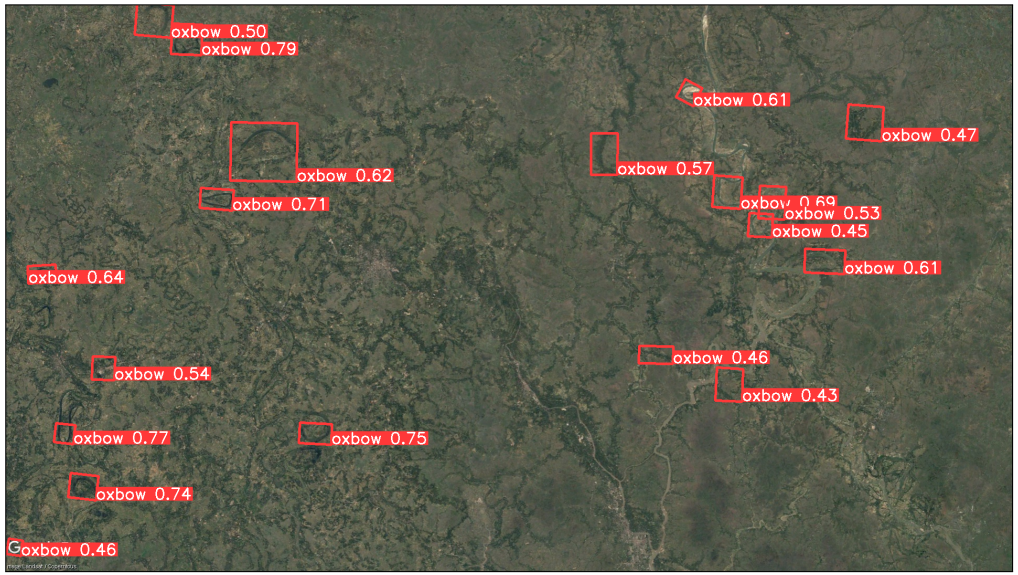
Figure 18: YOLO test cases (1–4). The satellite imagery was obtained from Google Earth Pro, captured in May 2025, and December 2026. Map data: Google, © 2025 Maxar Technologies.



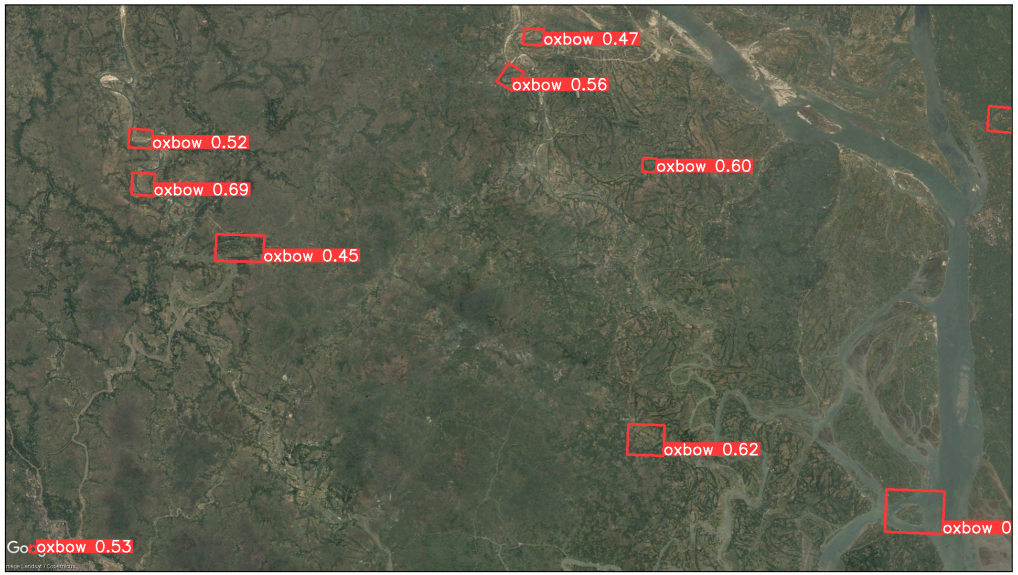
(a) Test case 5: Dry season, isolated oxbows



(b) Test case 6: Dry season, dominant river



(c) Test case 7: Dry season, isolated oxbows



(d) Test case 8: Dry season, dominant river

Figure 19: YOLO test cases (5–8). The satellite imagery was obtained from Google Earth Pro, captured in May 2025, and December 2026. Map data: Google, © 2025 Maxar Technologies.

Through empirical testing, the following parameters were found to perform optimally across the various test cases: a slice height and width of 200 pixels, and an overlap height and width ratio of 0.60 for the sliding window. The confidence threshold was set to 40%. There are no formal guidelines for selecting optimal slicing parameters, as these often depend on the specific characteristics of the dataset and the objects being detected. In this case, the trade-off involves balancing generalisation performance against spatial precision, particularly since oxbow lakes vary significantly in size and shape. The YOLO inference was conducted using the Ultralytics package (version 8.3.130) in Python 3.11.12 with PyTorch 2.6+cu124, running on an NVIDIA A100 GPU (12 compute units, 83.5 GB RAM) via Google Colab.

The most notable difference between the test cases is the seasonal variation. Test cases 5 and 6, which correspond to the dry season, show a higher number of (false positive) oxbow lake detections for similar scenes compared to test cases 1 and 2 in the wet season. This suggests that seasonal conditions may influence the model's ability to distinguish oxbow lakes, potentially due to differences in water visibility, vegetation cover, or lighting. In the dry season cases, many of the false positives appear as elongated shapes with dense vegetation, such as tree lines alongside canals or riverbanks, which contrast strongly with the surrounding landscape and visually resemble oxbow lakes. Additionally, a consistent pattern across the results is the reduced number of detections in scenes dominated by large river channels. This occurs even when the imagery contains prominent meander-like geomorphological features, indicating that the presence of a dominant river system may hinder the model's sensitivity to isolated oxbow lakes in such settings. However, the effect of these false positives is significantly reduced when moving from the regional scale in test case 5 to the more focused, local scale in test case 7, suggesting that scene context and spatial resolution play a role in moderating detection errors.

5.3 Testing Arsenic Risk Classification using GMM

The definition of risk used in this study was introduced in Chapter 2, while the GMM approach was described in Chapter 3. Chapter 4 validated and tested a set of predictor variables for use with the GMM. This section integrates these three components to generate an arsenic-risk prediction based on the variables identified in the case study. This section evaluates the performance of the GMM using topographic elevation and population density as predictor variables. At this stage, pixel-wise arsenic-risk predictions are computed across the entire area of interest, without distinguishing between the presence or absence of oxbow lakes and their associated point bars. Both the population density and elevation data are derived from the most recent available sources. Accurate knowledge of current population density on point bar ridges is important for precise risk prediction. The population density data for 2025 are based on projections from the GHSL database. Topographical elevation data are not associated with a specific year, given that elevation generally changes slowly over time. One of the elevation data sources used in this study is the ALOS World 3D 30m dataset, with the most recent version released in 2016. Figure 23 in Appendix B.2 illustrates the original distributions of elevation and population density. Elevation follows an approximately Gaussian distribution with sharp peaks, while population density is heavily skewed. This non-Gaussian nature complicates clustering with the GMM. To address this, a log-transformation is applied to the population data to reduce the long tail, and a quantile-based transformation is used for elevation to better approximate a normal distribution. The resulting distributions, shown in Figure 24, are better suited for multivariate clustering, although population density remains somewhat skewed and continues to pose challenges for the GMM.

First, the two data sources are spatially aligned and clipped to the defined area of interest. Subsequently, only the data points containing valid (non-NaN) values are retained for analysis. The topographical elevation and population density are not normalised. In GMMs, the assumption that all clusters have the same variance is relaxed, for example, compared to K-means clustering. The clustering by GMM is entirely data-driven and free of manually defined thresholding; post hoc interpretation of the resulting cluster centroids was performed. This allowed for a qualitative labelling of risk classes (e.g. 'High Risk', 'Potential Risk', 'Low Risk') based on a centroid-based approach using the relative position of each cluster centroid. The sum of the elevation and population values is used as a risk score for each centroid. These scores are then ranked, and the cluster labels are assigned based on their relative scores. The lowest score is assigned to the Low Risk category, and the highest score to the High Risk category. This method ensures stable, data-driven classification that adapts to varying distributions and avoids arbitrary decision boundaries. The GMM was trained for three components with a fixed random seed of 42, `covariance_type: full`, and initialisation method `kmeans`. To evaluate the optimal number of components for the GMM, both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) were computed across a range of component values. The analysis revealed that $k = 3$ components yielded the lowest BIC and AIC scores, indicating the best balance between model fit and complexity. Notably, both criteria began to converge after $k = 2$, and the log-likelihood also showed minimal improvement beyond this point, suggesting diminishing returns with additional components.

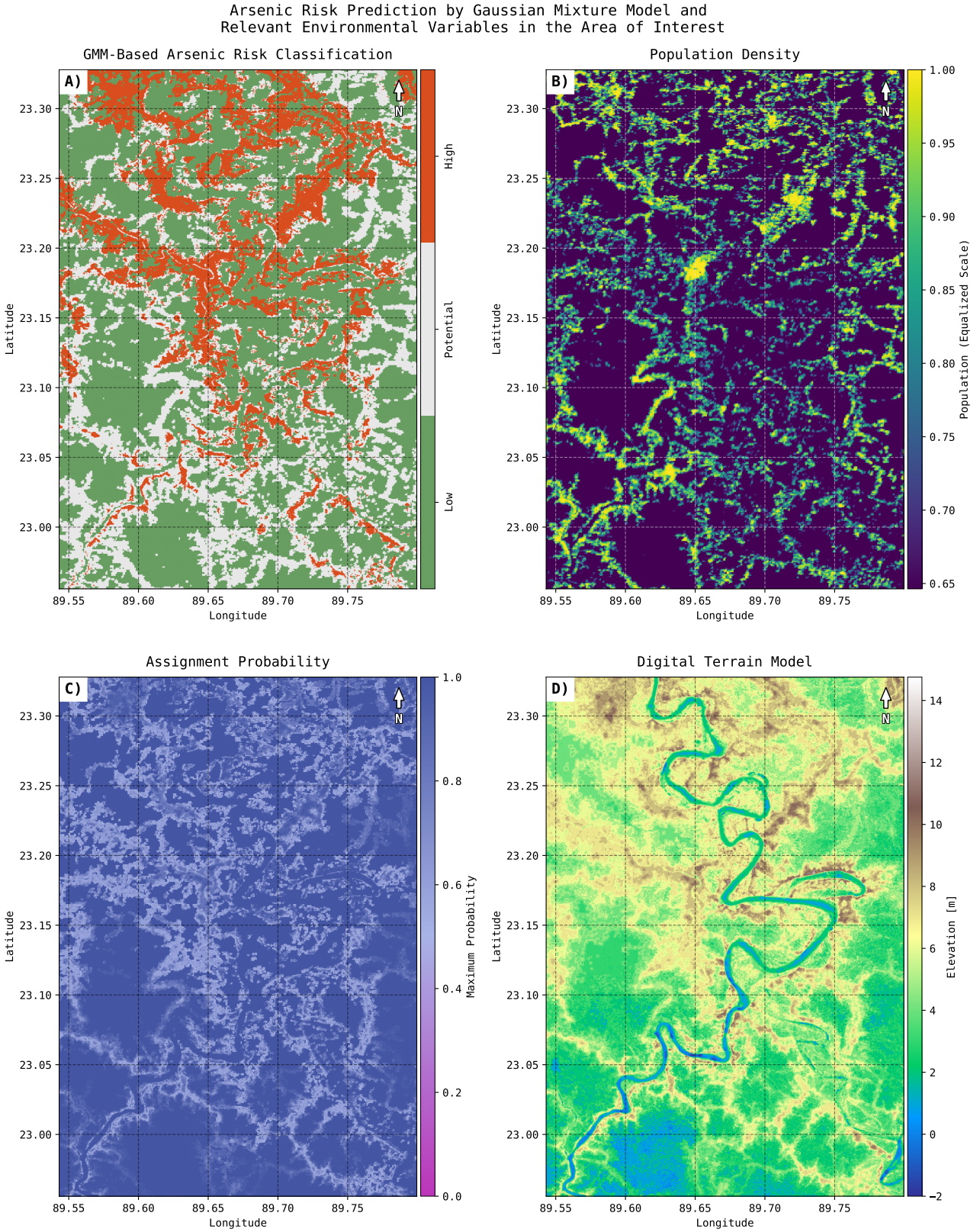
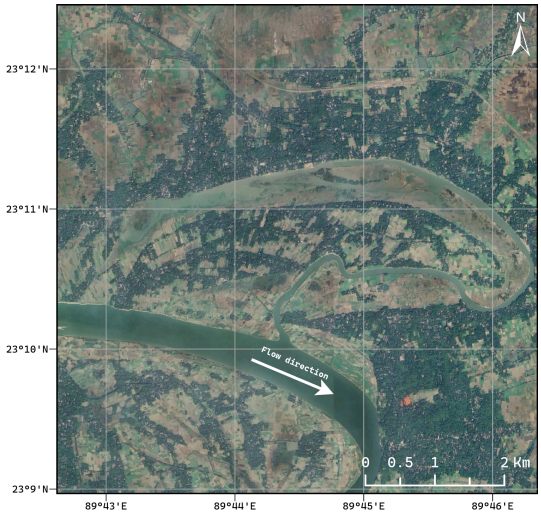
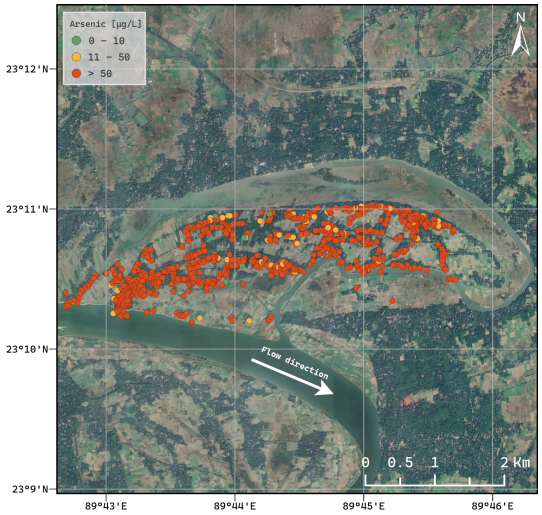


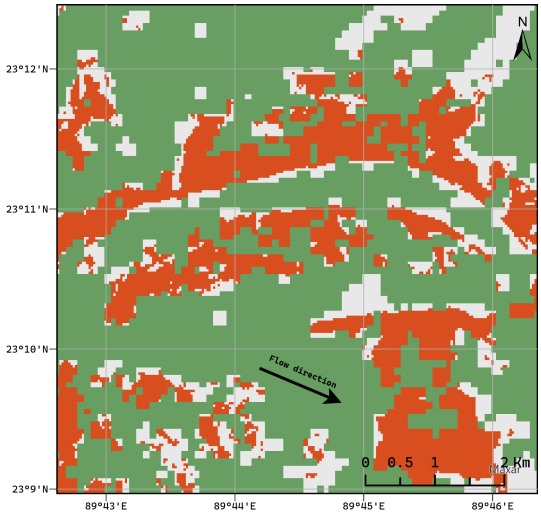
Figure 20: Geospatial layers supporting arsenic risk assessment in the area of interest, classified using Gaussian Mixture Model (initialized with `kmeans`). **A):** GMM-based arsenic risk classification based on **elevation** and **population density**. **B):** Population density map (for visualisation histogram-equalised), emphasizing spatial patterns of human settlement. **C):** Maximum prediction probability derived from the GMM model, representing the probability of the assigned risk class. **D):** DTM showing topographic variation across the study region.



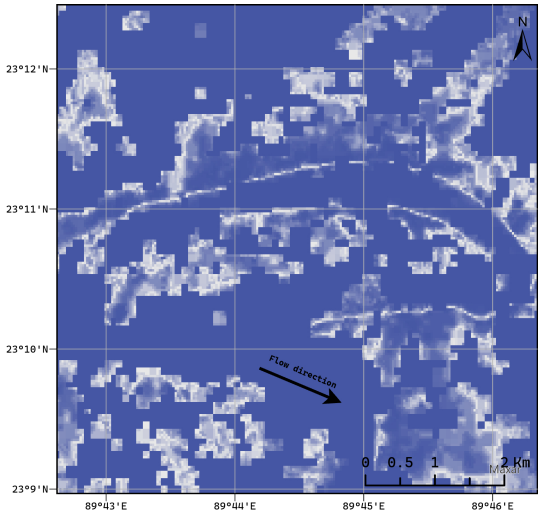
(a) Optical satellite image of an oxbow lake and the associated point bar in the area of interest.



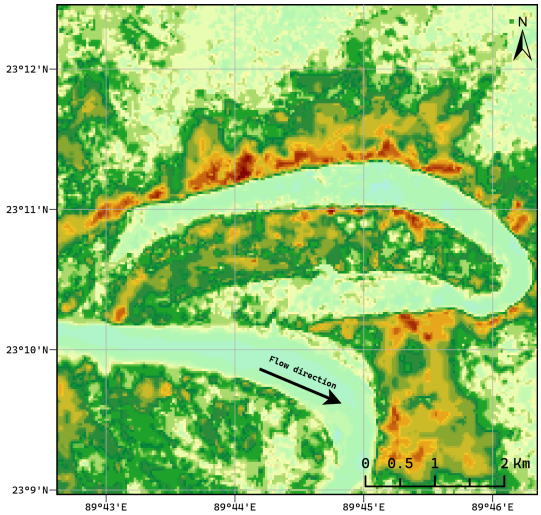
(b) Measured arsenic concentrations from groundwater wells within the study area, overlaid on satellite imagery.



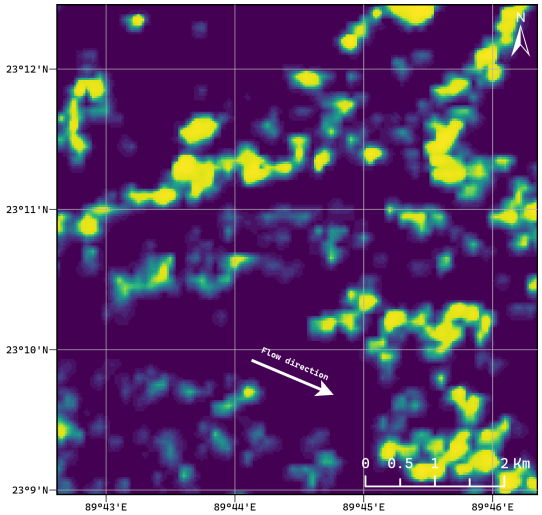
(c) GMM-based arsenic risk classification derived from topographic elevation and population density features



(d) Prediction certainty visualised as the maximum GMM class probability per pixel.



(e) DTM highlighting topographic variation across the oxbow lake and surrounding floodplain.



(f) Population density across the area of interest, displayed with histogram equalization applied solely for visualization purposes to enhance contrast and reveal settlement concentration patterns.

Figure 21: Spatial overview of the arsenic risk classification alongside relevant geospatial predictor variables. These layers provide spatial context for interpreting the risk distribution and support the understanding of relationships between predictor variables and model outcomes.

As shown in Figure 20, the top left plot displays the GMM's arsenic risk prediction results. When comparing the GMM-based arsenic risk predictions with the geospatial predictor variables (DTM, population density), all exhibit spatial patterns that almost completely correspond to those observed in the predicted risk, especially around point bars within oxbow lakes. Inspecting the model probability plot, the model is overall, for the area of interest, mostly certain in its clustering. This suggests the model is decisive in its classification. Lower probability zones surround the boundary between classes. Additionally, regions with lower model probability (approximately 0.5) are typically found where there is a mismatch between the predictor variables, such as areas with high population density and low elevation, or the opposite. One such example is located between longitudes 89.55 and 89.60 and latitudes 23.20 and 23.75. An additional noteworthy observation is found in the area between longitudes 89.75 and 89.80 and latitudes 23.20 and 23.25, where the model predicts a Potential Risk with a probability level between 0.6 and 0.8. In this region, the DTM indicates low-lying terrain within a populated area. A more detailed analysis of the arsenic risk prediction is presented in the following section, focusing on a single oxbow lake and its associated point bar in the area of interest.

5.4 Application of Oxbow Lake Detection in Risk Mapping

In this section, one of the oxbow lakes and its associated point bars, as detected by the YOLO model, is selected for closer examination to evaluate the spatial relationship between geomorphological features and arsenic risk predictions. Figure 21 shows a selected oxbow lake detected by the YOLO model. The top-left panel displays the satellite imagery of the area, while the top-right panel shows the measured arsenic concentrations from nearby wells. The middle-left panel presents the GMM-predicted arsenic risk classification, and the middle-right panel shows the corresponding prediction probability. From the satellite image (21a), the ridges on the point bar are visible, accentuated by the dense vegetation growing along them, supported by the DTM in Figure 21e. According to the hypothesis by Donselaar *et al.*, ridges on point bars are more likely to be populated. All available arsenic measurements from the well dataset are located directly on these ridges, as shown in Figure 21b. According to the safety standards of the WHO, all of these wells pose a high risk to the local population. To determine whether the prediction in Figure 21c is reasonable, the results should be compared with the topographic elevation and the area's population density, which defines risk in this study. The predictions in Figure 21c align with underlying geospatial patterns that appear to be governed by topography and population density variations, upon comparison with the DTM in Figure 21e and the population density in Figure 21f. When the probability is inspected in Figure 21d, the GMM model was the least certain in boundary cases, specifically in transitions from High Risk to Potential Risk, or areas where only one of the two prediction variables is high. Another notable pattern in the predictions by the GMM are the sharp boundaries and cut-offs between the High Risk class and the other classes.

5.5 Summary Results

A two-part machine learning pipeline for arsenic-risk prediction in Bangladesh was implemented, using: 1) YOLOv-11x-OBB for oxbow lake detection in satellite imagery, and 2) a GMM for arsenic-risk classification. Testing the YOLO model across eight scenarios with varying altitudes, seasons, and landscapes revealed better detection during wet seasons and in areas with isolated oxbow lakes (hydrologically disconnected to the main river) compared to those dominated by river systems. When limited to just elevation and population density, the GMM successfully classified areas into High, Potential, and Low Risk categories, with the highest probability away from class boundaries. The case study analysis of a selected oxbow lake confirmed a spatial relationship between geomorphological features and arsenic risk, supporting the hypothesis that populated ridges on point bars are associated with elevated arsenic risk. However, areas predicted to be at risk but lacking arsenic measurements from tube wells remain uncertain, as they cannot be validated without in situ data.

6. Discussion

6.1 Interpretation of Key Results and Findings

6.1.1 Arsenic-Risk Prediction by the Gaussian Mixture Model

The results of this study reveal distinct spatial patterns in arsenic risk, which appear to be influenced by underlying geospatial features and demographic factors. In particular, areas predicted to be High Risk zones by the GMM often coincide with zones of elevated topography and relatively high population density. This alignment supports the underlying hypothesis by Donselaar *et al.* [17] that certain physical features, such as ridges on point bars, are susceptible to a greater risk of arsenic exposure. The predicted arsenic risk maps revealed that High Risk areas were typically surrounded by zones of Potential Risk, reflecting a gradual transition in both elevation and population density. These transitional zones often occurred along the edges of geomorphological ridges, where elevation decreased and fewer people were present.

A valuable aspect of using the GMM is the availability of probabilistic outputs, which allow for a pixel-level interpretation of model certainty. Probability is defined here as the maximum posterior probability that a data point belongs to any of the GMM components. Areas with high probability (e.g. >0.9) indicate strong separation between classes and well-supported predictions. In the visualisation, these regions appear in deep blue and were frequently associated with High Risk zones, suggesting the model is highly decisive in identifying areas of greatest concern. In contrast, regions with probabilities values around 0.5 suggest that the model assigns nearly equal probabilities to two or more risk categories. These low-probability zones frequently occur along the boundaries between High and Potential Risk areas, or in regions where geospatial predictors, such as elevation and population density, provide conflicting signals. Visually, these areas often appear in grey and are commonly mapped as Potential Risk zones or transitional zones that bleed into High Risk areas.

6.1.2 Data Quality Limitations

A further limitation affecting the GMM's predictive capability was the inconsistent availability of high-resolution, online-accessible, and temporally appropriate input data. Population density data were resampled from their original 3 arcsecond (~90 meter) resolution to 30 meter resolution using nearest neighbour assignment, where each 90 by 90 meter cell was divided into three 30 by 30 meter sub-cells with the original value uniformly assigned. An alternative approach would have been linear interpolation. While nearest neighbour resampling preserves the original values and avoids introducing artificial gradients, it can lead to abrupt transitions and sharp class boundaries in the resulting maps, potentially distorting the spatial continuity of arsenic risk zones. In contrast, linear interpolation would have produced smoother transitions and reduced abrupt discontinuities in the data representation, but at the cost of possibly blurring sharp spatial contrasts or introducing artificial gradients, which may mask meaningful local variation. Furthermore, the DTM data had a spatial resolution of 30 meters, but its quality may be questionable due to its compilation from multiple sources. As no independent ground-truth elevation data were available, the accuracy of the DTM could not be directly validated. Inconsistent spatial granularity across features may have hindered the model's ability to detect fine-scale spatial patterns relevant to arsenic risk.

Additionally, both the historical arsenic dataset used in Chapter 4 and the data employed in Chapter 5 were collected exclusively from residential areas, which are typically situated on elevated ridges within point bars or other geomorphological features with higher topographic elevation. As a result, the absence of sampling in lower-lying clay-rich zones limits the ability to validate risk predictions across the full range of geomorphological contexts and may introduce a sampling bias toward known higher risk settings.

6.1.3 Oxbow Lake Detection by the YOLO Model

In order to perform arsenic risk classification at specific geomorphological features, it was first necessary to accurately identify the location of oxbow lakes and their associated point bars. To this end, a YOLO-based object detection model was employed to locate these features across the study area, as oxbow lakes and point bars are geomorphological structures considered critical to understanding the spatial distribution of arsenic contamination. The observed variation in detection performance across test cases is largely attributable to seasonal changes and landscape context. During the dry season, reduced water levels and clearer vegetation boundaries enhance the visibility of oxbow lakes, improving model accuracy. Conversely, high water coverage and reduced contrast between water and surrounding land in the wet season likely obscure key visual features, leading to detection shortfalls and misclassification. Moreover, the limited success in detecting oxbow lake-like features near active river channels suggests that the model was trained predominantly on isolated oxbow lakes and may not generalise well to similar geomorphological features, such as partly hydrologically connected oxbow lakes that are still connected to the river on one side. The lack of curated negative examples, such as features that are visually similar to oxbow lakes but are not, likely limited the model's ability to distinguish between true positives and false positives, thereby reducing its robustness. Additionally, macrophytes within oxbow lakes may alter the visual appearance of oxbow lakes, further affecting the model's ability to generalise across different environmental settings. Lastly, oxbow lakes come in many different shapes and sizes, which challenges the model's detection capabilities, as these features can appear in highly variable and often complex patterns within the landscape, making consistent identification more difficult. This underscores how landscape and seasonal variability critically influence the reliability of remote sensing-based geomorphological detection, while also highlighting that the limited availability of diverse training samples constrained the YOLO model's generalisation performance. In particular, the occurrence of false positives suggests that the model was likely overfitted to specific visual characteristics present in the limited training set, reducing its ability to generalise to unseen or more complex cases, with the effects of this overfitting becoming evident during validation.

6.1.4 Other Approaches for Arsenic Risk Prediction

In addition to the GMM presented in this thesis, several supervised machine learning methods were also explored as potential approaches for arsenic risk prediction. These included the Histogram Gradient Boosting Classifier and XGBoost, applied both with a minimal feature set (elevation and population density) and with an extended set that included terrain variables such as the TPI, TWI, and LULC. While some of these models appeared to capture spatial patterns reasonably well, they were ultimately excluded from the final report. This decision was made because the results did not align with the theoretical framework that guided this study.

Specifically, an inspection of the arsenic well dataset provided by Donselaar *et al.* [17] revealed that the average elevation and population density were lower for wells classified as Low Risk. This observation contradicts the hypothesis that higher elevation and higher population density increase the likelihood of arsenic contamination. However, this does not suggest that the original hypothesis is incorrect. Rather, it highlights the complex interplay between geospatial variables and measured arsenic concentrations. One possible explanation is that the digital elevation model and the population density data used in this study may lack the spatial precision needed to represent these relationships accurately. Another explanation lies in the role of well depth, which strongly influences arsenic exposure but cannot be transferred or easily modelled when making predictions in unseen areas. In fact, the dataset indicated that High Risk wells were often shallow, which supports the underlying hypothesis. It is also possible that low-lying points in the terrain function as drainage zones where arsenic-rich run-off accumulates, leading to high concentrations despite lower elevation. Meanwhile, elevated ridges on point bars, although theoretically at higher risk, may be sparsely inhabited and therefore under-represented in the training data. Additionally, more densely populated areas may have better access to alternative water sources, reducing their actual exposure.

It is important to note that the differences in mean elevation and population density between the risk categories were relatively small, with elevation differences often lower than a single meter. Therefore, the proposed supervised models should not be dismissed entirely, as they have demonstrated the ability to capture spatial patterns effectively. However, their interpretability and alignment with existing theory remain limited without further data analysis. These findings reinforce the need for more robust, high-resolution datasets and careful model validation when investigating environmental health risks. They also suggest that probabilistic frameworks with integrated uncertainty, such as the GMM used in this study, remain better suited for dealing with complex geospatial relationships in data-scarce settings, for this particular research.

6.2 Context within Literature

This study builds upon and extends existing research on arsenic contamination by introducing a spatially explicit, geomorphologically informed modelling approach that integrates machine and deep learning techniques. Direct comparison with prior studies is challenging, as this work does not rely on measured arsenic concentrations to predict precise hotspot locations. Instead, it focuses on identifying geomorphological features, such as point bars and oxbow lakes, to constrain areas of elevated risk spatially. Based on the findings of Donselaar *et al.* [17], these features are assumed to be associated with persistently high arsenic concentrations, making them reliable proxies for long-term contamination risk. By incorporating the YOLO model, geomorphological features such as oxbow lakes can be rapidly detected. These detections can then be refined using topographic elevation (DTM) and population density data to produce maps that precisely delineate potential arsenic hotspots. This methodology improves upon previous efforts that tended to generalise risk over large regions and were prone to producing unrealistic hotspot patterns due to spatial averaging.

7. Conclusion and Recommendations

7.1 Conclusion

7.1.1 Research Questions Revisited

This chapter synthesises the thesis's main findings, evaluates the effectiveness of the proposed methods, and revisits the research questions in light of the results. It also outlines the work's broader practical implications and proposes directions for future research. The main research question of this thesis was:

How can geomorphological features, specifically oxbow lakes and point bars, be used to optimise machine-learning models for accurate arsenic risk prediction and mapping in the Ganges-Brahmaputra Basin?

To address this question, a series of sub-questions was formulated. The answers to these sub-questions together provide a comprehensive evaluation of the proposed approach and demonstrate how the main research question has been answered. Each sub-question will be revisited, summarising the findings and outlining how the sub-question was answered, and how its findings contributed to resolving the main research question.

1. *How are the environmental conditions and geological characteristics in the Ganges-Brahmaputra Basin related to groundwater arsenic contamination?*

   In Chapter 2, an extensive overview is provided detailing the region's geomorphology, hydrology, and environmental settings. Particular emphasis is placed on geomorphological features such as point bars and oxbow lakes, and the ways in which human interaction with groundwater contributes to exposure risk. Much of the geoscientific basis for understanding arsenic mobilisation in this context has been established by [15]–[17], [24], [29], whose work forms the conceptual foundation for this thesis. Their research provided the guiding framework through which the literature was reviewed and the modelling strategy was developed. They also provided clear constraints and guidelines, such as which (auxiliary) proxy variables were likely to be informative.

2. *Which method is effective for detecting geomorphological features such as oxbow lakes and point bars?*

   Donselaar *et al.* [17] proposed using a Mask R-CNN to detect and delineate oxbow lakes and point bars. While a Mask R-CNN offers precise segmentation and is well-suited for delineating complex shapes, it requires extensive annotated datasets and is computationally intensive. In contrast, this study employed the YOLO model, which was ultimately chosen due to its efficient training process, suitability for smaller custom-labelled datasets through transfer learning, and ability to generate results rapidly. These characteristics make YOLO a practical and effective choice for a proof of concept, allowing for rapid development and initial validation of geomorphological detection without the overhead of exhaustive data annotation or model complexity.

   The trained YOLO model showed potential in detecting oxbow lakes and associated point bars across varied seasonal and landscape contexts. The model's performance indicates a promising capacity for generalisation to other regions with oxbow lakes. While the surrounding vegetation and landscape settings may differ from the study area, the geometric consistency of oxbow lakes enhances the model's transferability. However, several limitations were observed during inference. The model performed well for well-isolated oxbow lakes, shown for all test cases, but struggled with more ambiguous cases. Partly active oxbow lakes, which remain hydrologically connected to the main river on one side, were rarely detected, likely due to their reduced visual separation from the river channel. Similarly, point bars embedded within active river meanders were difficult to identify, as their boundaries were less distinct. These detection gaps show the need for further training on a more diverse set of samples, covering a wider range of oxbow lake morphologies, seasonal appearances, and hydrological states. The YOLO model used in this study proved to be an efficient and scalable detection tool and is a good fit for a proof-of-concept, but it must be expanded and refined to support broader and more robust applications.

3. *To what extent do auxiliary environmental and geomorphological variables explain the spatial and temporal distribution of arsenic contamination risk?*

   This sub-question was addressed in Chapter 4, where an XGBoost classifier was employed to evaluate the predictive power of a range of auxiliary environmental and geomorphological variables using arsenic concentration data from 20 wells collected between 2001 and 2003. The analysis demonstrated that sand content, elevation, and SOC were the most influential predictors of arsenic risk. These findings support the hypothesis by Donselaar *et al.* [17] that elevated, sandy point bars function as persistent arsenic hotspots. While population density contributed less than expected, its inclusion still improved model performance and aligned with broader geomorphological patterns.

   Feature importance and SHAP analysis confirmed the relevance of geomorphologically derived features, while variables such as NDVI and precipitation showed negligible contribution. Their limited impact is attributed to temporal misalignment, coarse spatial resolution, and conceptual mismatch, in this specific case and model, with the target variable. Consequently, these variables were excluded from the final model implementation. Although LULC exhibited no predictive power within the XGBoost framework, its spatial alignment with known arsenic-prone zones suggest that it remains a valuable proxy for human and environmental factors influencing arsenic mobilisation. In Chapter 4, the most frequently occurring LULC class within a three-cell radius was assigned to each data point, which may have oversimplified local land cover heterogeneity and, in turn, influenced the computed feature importance, potentially giving a skewed impression of its value.

4. *How should arsenic contamination risk be defined to align with both model outputs and the practical implications for public health and groundwater management?*

   This study defined arsenic contamination risk using a conceptual framework inspired by flood risk analysis. In practice, this framework was translated into a three-tier risk classification: High Risk for areas that are both elevated and densely populated, Potential Risk for elevated regions with fewer people, and Low Risk for low-lying areas, regardless of population. This categorisation closely matched the clustering results from the Gaussian Mixture Model and allowed for a straightforward interpretation that could be directly applied to spatial risk maps. This approach is relevant for public health and groundwater management, as it provides a practical means of identifying and prioritising high-risk zones for mitigation and monitoring. However, this framework does not currently incorporate a quantitative measure of population exposure per square kilometre, which limits its ability to assess the absolute human impact of each risk zone.

5. *What machine-learning method is suited for classifying arsenic risk?*

   The GMM was identified as a suitable method for arsenic risk classification due to its ability to uncover latent structure in multivariate data without requiring labelled samples. Unlike supervised approaches such as a RF model, which rely on ground-truth data and often produce geographically inconsistent thresholds when extrapolated over large areas, the GMM offers a flexible probabilistic framework for unsupervised clustering. In this study, the GMM was able to capture spatial arsenic-risk patterns that broadly aligned with known geomorphological risk zones such as point bar ridges. Its capacity to provide soft clustering and interpretable spatial risk maps supports its application in data-scarce, geospatially complex environments. However, the model showed limitations in handling highly skewed input data, particularly population density. It is expected that this affected the model's ability to form well-separated clusters and may have introduced uncertainty in class boundaries. Additionally, the GMM occasionally assigned a Potential Risk classification to low-elevation, populated areas, which does not align with the intended definition of Potential Risk as areas that exhibit intermediate arsenic exposure and are geomorphologically distinct from both low- and High Risk zones. While such misclassification is less critical than incorrectly assigning High Risk, since it avoids triggering unnecessary mitigation measures, it still introduces ambiguity in the interpretation of risk categories and highlights the need for careful validation of unsupervised classifications. Rather than disregarding these ambiguous or low-probability areas, their uncertainty should be acknowledged as an informative layer that can guide further investigation. In a decision-making context, low-probability zones could be prioritised for field sampling or targeted monitoring, as they represent regions where model uncertainty may mask potentially significant risk. These findings suggest that while GMM is mostly effective for certain aspects of arsenic risk mapping, attention must be paid to feature scaling, class semantics, and the implications of soft clustering in heterogeneous environments.

7.1.2 Summary

This thesis has demonstrated that geomorphological features, specifically oxbow lakes and point bars, can effectively guide and constrain machine-learning models for arsenic risk prediction in the Ganges-Brahmaputra Basin. By focusing on spatial zones most susceptible to arsenic accumulation, as supported by geological theory and prior research, the study narrowed the predictive modelling scope in a targeted way. The integration of geomorphological detection with probabilistic machine learning enabled a spatially explicit and data-efficient framework for risk classification, even in the absence of extensive in-situ arsenic measurements, and reduces the reliance on costly fieldwork while allowing the rapid identification of arsenic hotspots. This, in turn, can inform field sampling strategies and assist local authorities in prioritising high-risk areas for intervention. As a proof of concept, the study shows that incorporating geomorphological insight improves model interpretability, scalability, and practical relevance, (partly) successfully answering the main research question and laying the groundwork for future applications. These findings underscore the importance of developing harmonised, high-resolution datasets when applying to environmental health challenges. Future efforts should prioritise the development of integrated geospatial datasets that are temporally aligned, semantically consistent, and capable of capturing both natural and anthropogenic drivers of environmental risk.

7.2 Recommendations

While this research presents a promising framework for geomorphologically informed arsenic risk prediction, several paths remain for refining the methodology and expanding its application.

7.2.1 Geomorphological Feature Object Detection

Starting with the initial component of the pipeline, object detection of geomorphological features presents several opportunities for further development. The YOLO model remains a practical choice, having shown effectiveness even with a limited training set. Once oxbow lakes are identified, a next step could involve delineating their boundaries. This can be accomplished by using SAR imagery, which is capable of distinguishing water from land regardless of weather or lighting conditions. The resulting water mask can then be applied to optical imagery to isolate vegetation within the oxbow lakes before and after the monsoon, enabling a rough estimate of organic carbon content. Alternatively, preprocessed masks can be used directly to support this analysis.

Another option involves applying the Segment Anything Model (SAM) developed by Meta AI, which can automatically delineate the boundaries of detected oxbow lakes. While its effectiveness in this specific application has yet to be fully evaluated, it offers a promising direction for automating the delineation process. If greater computational resources are available, the Mask R-CNN model proposed by Donselaar *et al.* [17] could be considered. This model may be fine-tuned using existing datasets or trained from scratch specifically on oxbow lakes, with SAM potentially assisting in generating annotated training data more efficiently.

In future iterations, the detection scope could be broadened beyond well-isolated oxbow lakes and point bars to include partly active oxbow lakes, meandering river bends, levees, and clay plugs. These features share similar sedimentological and biogeochemical conditions that contribute to arsenic mobilisation and accumulation. Although their visual appearance may differ, particularly in active fluvial environments, the underlying processes are mostly comparable. Incorporating these additional landforms into the object detection framework would offer a more comprehensive basis for identifying geomorphological arsenic hotspots across a wider range of fluvial settings. raining data could also be expanded by including annotated imagery from other regions or countries with comparable oxbow lake systems, which would improve model robustness and generalisability to a broader range of fluvial morphologies. Moreover, using false-colour imagery for training YOLO or Mask R-CNN models could help reduce the impact of seasonal visual variability by enhancing spectral contrast between water, vegetation, and bare soil, thereby improving detection consistency across different environmental conditions.

7.2.2 Arsenic-risk Prediction

Building on the outputs of the geomorphological detection step, the next stage involves refining the arsenic-risk prediction component to improve spatial accuracy, model robustness, and interpretability. The GMM was selected for its probabilistic framework, which enables soft classification and interpretable uncertainty estimates. However, the model's performance was sensitive to input data quality, especially the inconsistencies in spatial resolution, interpolation methods, the derived nature of several variables, and skewed data. An alternative to GMM for future work could be using supervised or unsupervised models capable of handling non-Gaussian clusters, capturing complex geospatial patterns, and integrating diverse data types such as continuous population density and categorical land use and land cover data. Alternatively, population density may be approximated using proxy variables, such as building density. Another important direction is examining temporal dependencies between auxiliary variables such as precipitation, NDVI, and land cover transitions. Understanding which predictors remain temporally stable can help improve long-term risk assessments and inform the selection of more reliable features. This may lead to more robust and context-aware classifications that reflect both spatial and seasonal variability.

Although LULC did not demonstrate predictive performance in the case study results, it remains a conceptually relevant variable for arsenic risk prediction. For example, Donselaar *et al.* [17] hypothesised that tube wells positioned on ridges enhance both diffusive and advective groundwater flow toward local water sources, thereby increasing the likelihood of arsenic contamination. The LULC data reveal that vegetated areas frequently coincide with these elevated ridges on point bars. Vegetation, particularly dense or deep-rooted species, contributes to groundwater drawdown through transpiration, reinforcing subsurface flow dynamics. Similarly, built-up areas, which often correlate with human activity and increased water extraction, exert a comparable influence on groundwater movement. These patterns suggest that LULC encodes both ecological and anthropogenic signals that are directly relevant to arsenic mobilisation processes. Its native 30-meter spatial resolution offers a finer-grained, spatially explicit proxy for hydrogeologically meaningful land use patterns compared to interpolated datasets like population density. When properly preprocessed or used within models that can handle categorical data, LULC has clear potential to improve spatial predictions of arsenic risk.

7.2.3 Incorporating Population Exposure into Risk Assessment

While the current approach identifies areas of geogenic arsenic risk based on geomorphological and environmental indicators, it does not explicitly incorporate population density in terms of exposure per square kilometre. As a result, the High Risk classification remains a relative measure, based primarily on geospatial features rather than actual human vulnerability. To make mitigation strategies more effective and targeted, future work should assess the number of people affected per unit area for each oxbow lake or geomorphological unit. Integrating population exposure data into the risk framework would allow for prioritisation of interventions not only based on contamination potential but also on actual public health impact. This would move the model toward a more impact-aware arsenic risk assessment, better aligning risk classification with mitigation urgency. Implementing this addition would be straightforward and quick, as the necessary population data and spatial delineations are already in place within the current framework.

A Supplementary Tables

A.1 Environmental Variable Categories for Arsenic Risk Modeling

Category	Variables
Topographic & Geomorphological	Elevation, Topographic Slope, Topographic Wetness Index (TWI), Topographic Position Index (TPI), Distance to the river, Geotectonic
Soil & Land Use	Soil Drain, Soil Moisture Capacity, Soils, Organic Carbon, Land Use & Land Cover (LULC)
Climate & Meteorological	Temperature, Precipitation, Aridity, Potential Evapotranspiration, Evapotranspiration
Vegetation & Surface Indices	Normalised Difference Vegetation Index (NDVI)

Table 2: Categorisation of environmental variables used for arsenic risk modelling.

A.2 Overview of Monitored Wells and Arsenic Levels

Well ID	Longitude	Latitude	Depth (m)	Year Installed	No. of Samples	Mean As ($\mu g\,L^{-1}$)
816	N90°38.38′	E23°47.08′	8	1999	36	(63 ± 13)
4110	N90°36.08′	E23°47.06′	10	1999	36	(48 ± 5)
4115	N90°36.06′	E23°47.05′	10	1999	36	(50 ± 32)
808	N90°38.41′	E23°47.19′	8	1995	35	(41 ± 1)
823	N90°38.42′	E23°47.06′	10	1997	36	(40 ± 3)
4071	N90°36.14′	E23°47.06′	10	1999	34	(63 ± 6)
4101	N90°36.05′	E23°47.05′	12	1997	36	(17 ± 1)
825	N90°38.41′	E23°47.07′	12	1997	34	(16 ± 1)
84	N90°38.43′	E23°47.04′	20	1995	43	(42 ± 3)
1651	N90°38.15′	E23°47.20′	20	1994	43	(44 ± 5)
4133	N90°36.14′	E23°47.07′	30	2001	35	(2.2 ± 1.9)
4146	N90°36.11′	E23°47.06′	30	2001	33	(3.2 ± 4.6)
CW-1	N90°36.12′	E23°47.04′	40	2001	30	(0.4 ± 0.3)
CW-4	N90°38.39′	E23°47.00′	60	2001	30	(30 ± 11)
CW-6	N90°38.27′	E23°46.50′	60	2001	27	(1.4 ± 0.6)
CW-3	N90°38.26′	E23°46.48′	60	2001	25	(1.8 ± 0.3)
1639	N90°38.11′	E23°46.47′	70	2001	35	(1.3 ± 0.5)
CW-2	N90°36.19′	E23°46.49′	75	2001	35	(0.8 ± 0.8)
CW-7	N90°39.06′	E23°47.17′	123	2001	25	(3.9 ± 0.5)
CW-5	N90°38.01′	E23°46.23′	142	2001	25	(1.4 ± 1.4)

Table 3: Well information including coordinates, depth, installation year, sampling count, and mean arsenic concentration, from [68].

B Supplementary Figures

B.1 Relationship Between Arsenic Concentration and Well Depth

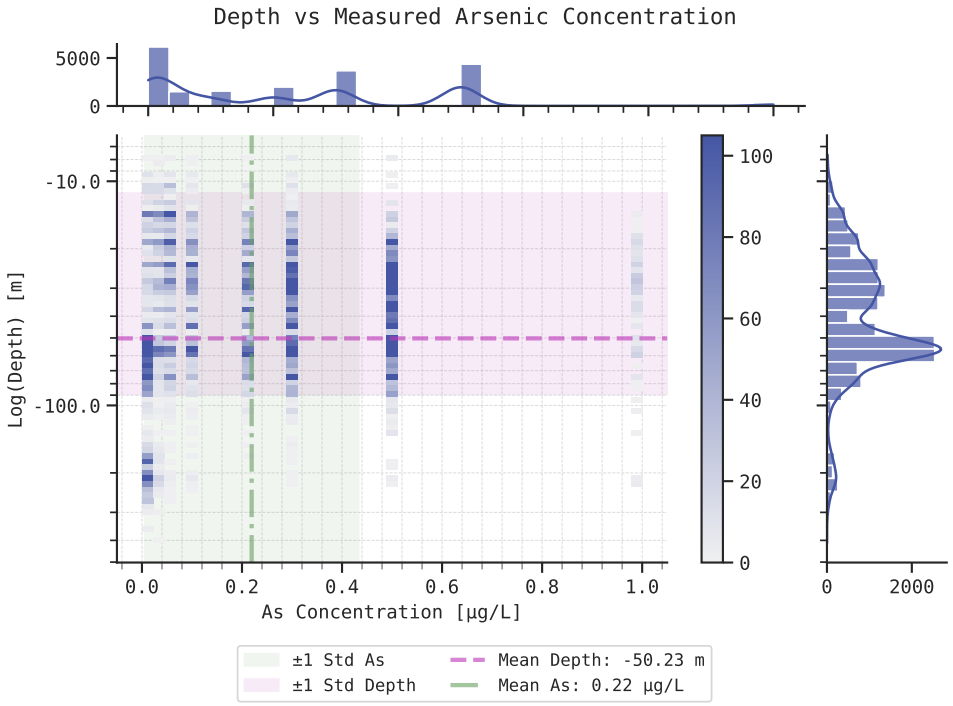


Figure 22: Joint distribution of arsenic concentration (µg/L) and well depth (log-transformed, in meters) based on the historic dataset from Kalia and Lohagara Upazilas.

B.2 Supporting Analysis of Elevation and Population Density Distributions

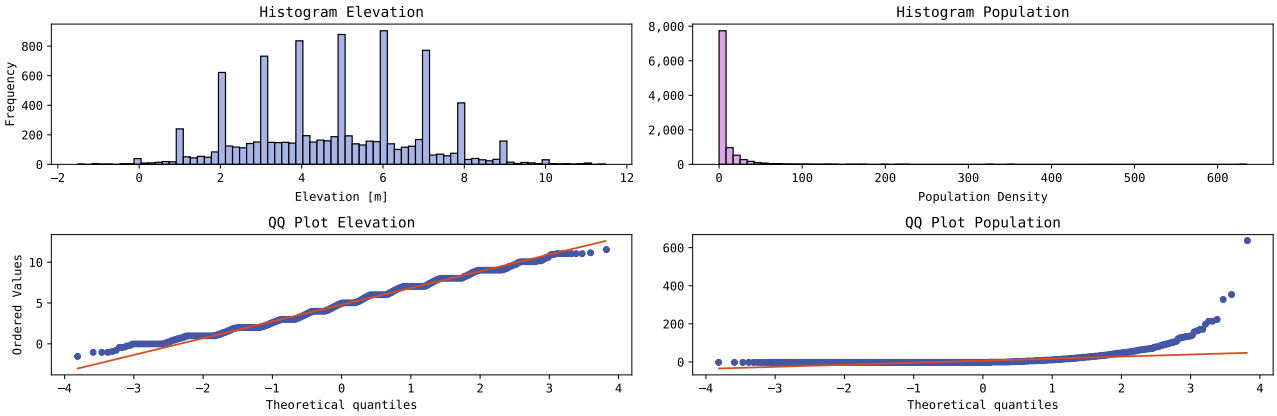


Figure 23: Distribution analysis of elevation and population density. The bottom row presents corresponding Q–Q plots to assess normality. While elevation approximates a normal distribution, population density remains highly skewed, indicating a non-Gaussian distribution structure.

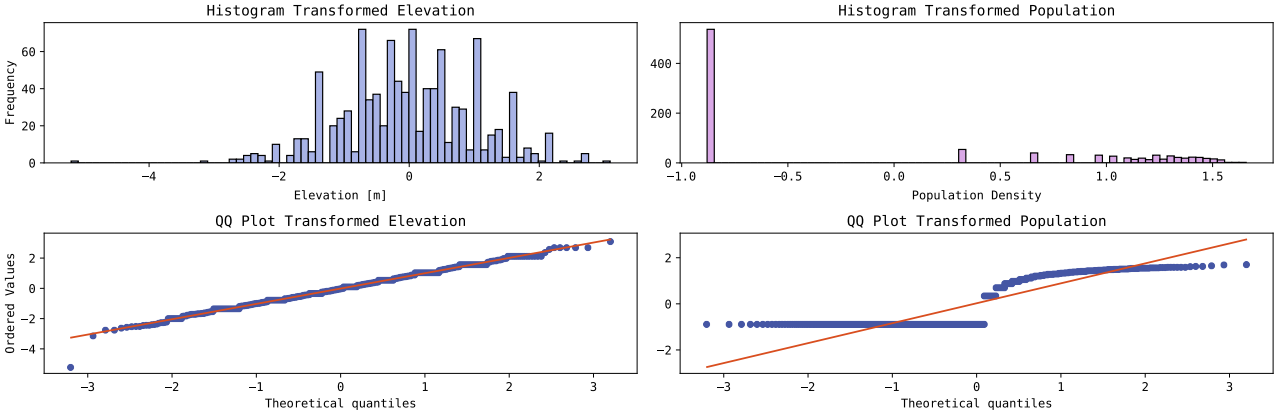


Figure 24: Distribution analysis of elevation and population density after transformation to approximate Gaussian distributions. The bottom row presents corresponding Q–Q plots to assess normality. While the transformation effectively normalizes elevation, population density remains highly skewed, indicating residual non-Gaussian characteristics.

Bibliography

[1] A. W. Grabau and G. Marshall, *A Textbook of Geology*. Boston, New York: D.C. Heath & Co, 1920, vol. 1, p. 417, Part I: General Geology. [Online]. Available at: https://openlibrary.org/books/OL7078265M/A_textbook_of_geology.

[2] S. A. Ahmad, M. H. Khan, and M. Haque, "Arsenic contamination in groundwater in bangladesh: Implications and challenges for healthcare policy," *Risk Management and Healthcare Policy*, vol. 11, pp. 251–261, 2018. DOI: 10.2147/rmhp.s153188. [Online]. Available at: https://doi.org/10.2147/rmhp.s153188.

[3] M. A. Ali, "Arsenic contamination of groundwater in bangladesh," *International Review for Environmental Strategies*, vol. 6, no. 2, pp. 329–360, 2006.

[4] Q. Y. Chen, S. Shen, H. Sun, *et al.*, "Pbmc gene expression profiles of female bangladeshi adults chronically exposed to arsenic-contaminated drinking water," *Environmental Pollution*, vol. 259, p. 113 672, 2019. DOI: 10.1016/j.envpol.2019.113672. [Online]. Available at: https://doi.org/10.1016/j.envpol.2019.113672.

[5] R. Uddin and N. H. Huda, "Arsenic poisoning in bangladesh," *Oman Medical Journal*, vol. 26, no. 3, p. 207, 2011. DOI: 10.5001/omj.2011.51. [Online]. Available at: https://doi.org/10.5001/omj.2011.51.

[6] World Health Organization. "Arsenic." Accessed on 2025-04-28. (2018), [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/arsenic Last accessed: 04/28/2025.

[7] M. S. Rahaman, N. Mise, and S. Ichihara, "Arsenic contamination in food chain in bangladesh: A review on health hazards, socioeconomic impacts and implications," *Hygiene and Environmental Health Advances*, vol. 2, p. 100 004, 2022. DOI: 10.1016/j.heha.2022.100004. [Online]. Available at: https://doi.org/10.1016/j.heha.2022.100004.

[8] Department of Public Health Engineering. "National strategy for water supply and sanitation 2014." Accessed on 2025-04-28. (2014), [Online]. Available at: https://dphe.portal.gov.bd/sites/default/files/files/dphe.portal.gov.bd/page/23471eb2_4fc7_4b55_92d1_0cb6e5d5e38d/2021-02-01-09-54-9fe8f0da5c4ad48f753bb8a509b992b1.pdf Last accessed: 04/28/2025.

[9] S. Pal, S. K. Singh, P. Singh, S. Pal, and S. R. Kashiwar, "Spatial pattern of groundwater arsenic contamination in patna, saran, and vaishali districts of gangetic plains of bihar, india," *Environmental Science and Pollution Research*, vol. 31, no. 41, pp. 54 163–54 177, 2023. DOI: 10.1007/s11356-022-25105-y. [Online]. Available at: https://doi.org/10.1007/s11356-022-25105-y.

[10] B. Nath, R. Chowdhury, W. Ni-Meister, and C. Mahanta, "Predicting the distribution of arsenic in groundwater by a geospatial machine learning technique in the two most affected districts of assam, india: The public health implications," *GeoHealth*, vol. 6, no. 3, 2022. DOI: 10.1029/2021gh000585. [Online]. Available at: https://doi.org/10.1029/2021gh000585.

[11] J. Podgorski, R. Wu, B. Chakravorty, and D. A. Polya, "Groundwater arsenic distribution in india by machine learning geospatial modeling," *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, p. 7119, 2020. DOI: 10.3390/ijerph17197119. [Online]. Available at: https://doi.org/10.3390/ijerph17197119.

[12] M. A. Rahman, M. A. B. Siddique, R. Khan, *et al.*, "Mechanism of arsenic enrichment and mobilization in groundwater from southeastern bangladesh: Water quality and preliminary health risks assessment," *Chemosphere*, vol. 294, p. 133 556, 2022. DOI: 10.1016/j.chemosphere.2022.133556. [Online]. Available at: https://doi.org/10.1016/j.chemosphere.2022.133556.

[13] M. Kabir, M. Salam, D. Paul, M. Hossain, N. Rahman, and M. Latif, "Spatial dependency of soil arsenic in bangladesh," *Journal of the National Science Foundation of Sri Lanka*, vol. 45, no. 2, pp. 179–191, 2017, DOI not available from document.

[14] A. Mukherjee, S. Sarkar, M. Chakraborty, *et al.*, "Occurrence, predictors and hazards of elevated groundwater arsenic across india through field observations and regional-scale ai-based modeling," *The Science of The Total Environment*, vol. 759, p. 143 511, 2020. DOI: 10.1016/j.scitotenv.2020.143511. [Online]. Available at: https://doi.org/10.1016/j.scitotenv.2020.143511.

[15] D. Ghosh and M. E. Donselaar, "Predictive geospatial model for arsenic accumulation in holocene aquifers based on interactions of oxbow-lake biogeochemistry and alluvial geomorphology," *The Science of The Total Environment*, vol. 856, p. 158 952, 2022. DOI: 10.1016/j.scitotenv.2022.158952. [Online]. Available at: https://doi.org/10.1016/j.scitotenv.2022.158952.

[16] M. E. Donselaar, A. G. Bhatt, and A. K. Ghosh, "On the relation between fluvio-deltaic flood basin geomorphology and the wide-spread occurrence of arsenic pollution in shallow aquifers," *The Science of The Total Environment*, vol. 574, pp. 901–913, 2016. DOI: 10.1016/j.scitotenv.2016.09.074. [Online]. Available at: https://doi.org/10.1016/j.scitotenv.2016.09.074.

[17] M. E. Donselaar, S. Khanam, A. K. Ghosh, C. Corroto, and D. Ghosh, "Machine-learning approach for identifying arsenic-contamination hot spots: The search for the needle in the haystack," *ACS ES&T Water*, vol. 4, no. 8, pp. 3110–3114, 2024. DOI: 10.1021/acsestwater.4c00422. [Online]. Available at: https://doi.org/10.1021/acsestwater.4c00422.

[18] M. A. Fazal, T. Kawachi, and E. Ichion, "Validity of the latest research findings on causes of groundwater arsenic contamination in bangladesh," *Water International*, vol. 26, no. 3, pp. 380–389, 2001. DOI: 10.1080/02508060108686930. [Online]. Available at: https://doi.org/10.1080/02508060108686930.

[19] R. M. Pradhan and A. K. Behera, "Geogenic arsenic in groundwater in india–a short review," *Journal of Geointerface*, vol. 2, no. 1, pp. 47–58, 2023.

[20] M. Hossain, "Arsenic contamination in bangladesh—an overview," *Agriculture Ecosystems & Environment*, vol. 113, no. 1-4, pp. 1–16, 2005. DOI: 10.1016/j.agee.2005.08.034. [Online]. Available at: https://doi.org/10.1016/j.agee.2005.08.034.

[21] Y. Zheng, M. Stute, A. Van Geen, *et al.*, "Redox control of arsenic mobilization in bangladesh groundwater," *Applied Geochemistry*, vol. 19, no. 2, pp. 201–214, 2003. DOI: 10.1016/j.apgeochem.2003.09.007. [Online]. Available at: https://doi.org/10.1016/j.apgeochem.2003.09.007.

[22] J. McArthur, D. Banerjee, K. Hudson-Edwards, *et al.*, "Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water: The example of west bengal and its worldwide implications," *Applied Geochemistry*, vol. 19, no. 8, pp. 1255–1293, 2004. DOI: 10.1016/j.apgeochem.2004.02.001. [Online]. Available at: https://doi.org/10.1016/j.apgeochem.2004.02.001.

[23] A. Van Geen, Y. Zheng, R. Versteeg, *et al.*, "Spatial variability of arsenic in 6000 tube wells in a 25 km2 area of bangladesh," *Water Resources Research*, vol. 39, no. 5, 2003. DOI: 10.1029/2002wr001617. [Online]. Available at: https://doi.org/10.1029/2002wr001617.

[24] D. Ghosh, S. Kumar, M. E. Donselaar, C. Corroto, and A. K. Ghosh, "Organic carbon transport model of abandoned river channels - a motif for floodplain geomorphology influencing biogeochemical swaying of arsenic," *The Science of The Total Environment*, vol. 762, p. 144 400, 2020. DOI: 10.1016/j.scitotenv.2020.144400. [Online]. Available at: https://doi.org/10.1016/j.scitotenv.2020.144400.

[25] R. Ahmed and I. Kim, "Patterns of daily rainfall in bangladesh during the summer monsoon season: Case studies at three stations," *Physical Geography*, vol. 24, no. 4, pp. 295–318, 2003. DOI: 10.2747/0272-3646.24.4.295. [Online]. Available at: https://doi.org/10.2747/0272-3646.24.4.295.

[26] B. Planer-Friedrich, C. Härtig, H. Lissner, *et al.*, "Organic carbon mobilization in a bangladesh aquifer explained by seasonal monsoon-driven storativity changes," *Applied Geochemistry*, vol. 27, no. 12, pp. 2324–2334, 2012. DOI: 10.1016/j.apgeochem.2012.08.005. [Online]. Available at: https://doi.org/10.1016/j.apgeochem.2012.08.005.

[27] M. A. T. Jihan, S. Popy, S. Kayes, G. Rasul, A. S. Maowa, and M. M. Rahman, "Climate change scenario in bangladesh: Historical data analysis and future projection based on cmip6 model," *Scientific Reports*, vol. 15, no. 1, 2025. DOI: 10.1038/s41598-024-81250-z. [Online]. Available at: https://www.nature.com/articles/s41598-024-81250-z.

[28] A. A. Fahad, M. Hasan, N. Sharmili, S. Islam, E. T. Swenson, and M. K. Roxy, "Climate change quadruples flood-causing extreme monsoon rainfall events in bangladesh and northeast india," *Quarterly Journal of the Royal Meteorological Society*, vol. 150, no. 760, pp. 1267–1287, 2023. DOI: 10.1002/qj.4645. [Online]. Available at: https://doi.org/10.1002/qj.4645.

[29] S. Kumar, D. Ghosh, M. Donselaar, F. Burgers, and A. Ghosh, "Clay-plug sediment as the locus of arsenic pollution in holocene alluvial-plain aquifers," *CATENA*, vol. 202, p. 105 255, 2021. DOI: 10.1016/j.catena.2021.105255. [Online]. Available at: https://doi.org/10.1016/j.catena.2021.105255.

[30] S. H. Frisbie, E. J. Mitchell, and A. R. Molla, "Sea level rise from climate change is expected to increase the release of arsenic into bangladesh's drinking well water by reduction and by the salt effect," *PLoS ONE*, vol. 19, no. 1, e0295172, 2024. DOI: 10.1371/journal.pone.0295172. [Online]. Available at: https://doi.org/10.1371/journal.pone.0295172.

[31] D. Gayle, "Climate crisis to increase cancer risk for tens of millions of people in bangladesh," 2024. [Online]. Available at: https://www.theguardian.com/world/2024/jan/17/climate-crisis-increase-cancer-risk-bangladesh-water.

[32] A. Ruggeri. "How climate change could affect arsenic in rice." (2025), [Online]. Available at: https://www.bbc.com/future/article/20250417-how-climate-change-could-affect-arsenic-in-rice.

[33] U.S. Department of Agriculture. "Production: Rice, milled." Accessed: 2025-04-29. (2024), [Online]. Available at: https://www.fas.usda.gov/data/production/commodity/0422110.

[34] Geological Society of London. "Ganges delta." Accessed: 2025-04-29. (2024), [Online]. Available at: https://www.geolsoc.org.uk/science-and-policy/plate-tectonic-stories/ganges-delta/.

[35] Delta Alliance. "Ganges-brahmaputra delta." Accessed: 2025-04-29. (2024), [Online]. Available at: http://www.delta-alliance.org/deltas/ganges-brahmaputra-delta.

[36] Satellite Image Deep Learning. "Techniques – satellite image deep learning." Accessed: 2025-04-29. (2024), [Online]. Available at: https://github.com/satellite-image-deep-learning/techniques.

[37] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv (Cornell University)*, 2015. DOI: 10.48550/arxiv.1511.08458. [Online]. Available at: https://arxiv.org/abs/1511.08458.

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *arXiv (Cornell University)*, 2017. DOI: 10.48550/arxiv.1703.06870. [Online]. Available at: https://arxiv.org/abs/1703.06870.

[39] P. Freitas, G. Vieira, J. Canário, W. F. Vincent, P. Pina, and C. Mora, "A trained mask r-cnn model over planetscope imagery for very-high resolution surface water mapping in boreal forest-tundra," *Remote Sensing of Environment*, vol. 304, p. 114 047, 2024. DOI: 10.1016/j.rse.2024.114047. [Online]. Available at: https://doi.org/10.1016/j.rse.2024.114047.

[40] N. Ravi, V. Gabeur, Y. Hu, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv (Cornell University)*, 2024. DOI: 10.48550/arxiv.2408.00714. [Online]. Available at: http://arxiv.org/abs/2408.00714.

[41] R. Joseph, D. Santosh, G. Ross, and F. Ali, "You only look once: Unified, real-time object detection," *arXiv (Cornell University)*, 2015. DOI: 10.48550/arxiv.1506.02640. [Online]. Available at: https://arxiv.org/abs/1506.02640.

[42] Ultralytics. "What are oriented bounding boxes (obb) and how do they differ from regular bounding boxes?" Accessed: 2025-04-29. (2024), [Online]. Available at: https://docs.ultralytics.com/tasks/obb/#what-are-oriented-bounding-boxes-obb-and-how-do-they-differ-from-regular-bounding-boxes.

[43] LabelVisor. "Yolov8 vs mask r-cnn: In-depth analysis and comparison." Accessed: 2025-04-29. (2023), [Online]. Available at: https://www.labelvisor.com/yolov8-vs-mask-r-cnn-in-depth-analysis-and-comparison/.

[44] W. Kron, "Flood zoning and loss accumulation analysis for germany," in *Flood Defence '2002, Wu et al. (eds)*, ISBN 1-880132-54-0, Science Press, New York Ltd., 2002, pp. 82–88.

[45] P. L. Smedley and D. G. Kinniburgh, "Sorption and transport," in *Arsenic Contamination of Groundwater in Bangladesh*, British Geological Survey, 2002, ch. 12, pp. 213–220.

[46] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894. DOI: 10.1098/rsta.1894.0003. [Online]. Available at: https://doi.org/10.1098/rsta.1894.0003.

[47] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available at: https://www.jstor.org/stable/2984875.

[48] E. Fetaya, J. Lucas, and E. Andrews, *Csc 411 lectures 15-16: Gaussian mixture model & em*, Lecture handout, 2023. [Online]. Available at: https://www.cs.toronto.edu/~jlucas/teaching/csc411/lectures/lec15_16_handout.pdf Last accessed: 04/29/2025.

[49] H. Guan, J. Huang, L. Li, *et al.*, "Improved gaussian mixture model to map the flooded crops of vv and vh polarization data," *Remote Sensing of Environment*, vol. 295, p. 113 714, 2023. DOI: 10.1016/j.rse.2023.113714. [Online]. Available at: https://doi.org/10.1016/j.rse.2023.113714.

[50] O. Okwuashi, E. Eyo, and A. Eyoh, "Supervised gaussian mixture model based remote sensing image classification," *Global Journal of Environmental Sciences*, vol. 10, no. 1&2, pp. 57–65, 2011, ISSN 1596-6194. [Online]. Available at: https://www.ajol.info/index.php/gjes/article/view/79235.

[51] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. [Online]. Available at: https://doi.org/10.1145/2939672.2939785.

[52] H. Cao, X. Xie, Y. Wang, and Y. Deng, "The interactive natural drivers of global geogenic arsenic contamination of groundwater," *Journal of Hydrology*, vol. 597, p. 126 214, 2021. DOI: 10.1016/j.jhydrol.2021.126214. [Online]. Available at: https://doi.org/10.1016/j.jhydrol.2021.126214.

[53] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv (Cornell University)*, 2024. DOI: 10.48550/arxiv.2410.17725. [Online]. Available at: http://arxiv.org/abs/2410.17725.

[54] G. Jocher and J. Qiu, *Ultralytics yolo11*, version 11.0.0, 2024. [Online]. Available at: https://github.com/ultralytics/ultralytics.

[55] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," *2022 IEEE International Conference on Image Processing (ICIP)*, 2022. DOI: 10.1109/icip46576.2022.9897990. [Online]. Available at: https://doi.org/10.1109/icip46576.2022.9897990.

[56] Ultralytics. "Slicing aided hyper inference (sahi) — tiled inference guide." Accessed: 2025-04-29. (2024), [Online]. Available at: https://docs.ultralytics.com/guides/sahi-tiled-inference/ Last accessed: 04/29/2025.

[57] NASA GEDI Mission Team. "Global ecosystem dynamics investigation (gedi)." Accessed: Apr. 30, 2025. (2025), [Online]. Available at: https://gedi.umd.edu/ Last accessed: 04/30/2025.

[58] NASA ICESat-2 Mission Team. "Ice, cloud, and land elevation satellite-2 (icesat-2)." Accessed: Apr. 30, 2025. (2025), [Online]. Available at: https://icesat-2.gsfc.nasa.gov/ Last accessed: 04/30/2025.

[59] "Tessadem: Near-global 30-meter digital elevation model." Accessed: Apr. 30, 2025. (2024), [Online]. Available at: https://tessadem.com/ Last accessed: 04/30/2025.

[60] Japan Aerospace Exploration Agency (JAXA). "ALOS World 3D (AW3D) Dataset." Accessed: Apr. 30, 2025. (2025), [Online]. Available at: https://www.eorc.jaxa.jp/ALOS/en/dataset/aw3d_e.htm Last accessed: 04/30/2025.

[61] D. Yamazaki, D. Ikeshima, R. Tawatari, *et al.*, "A high-accuracy map of global terrain elevations," *Geophysical Research Letters*, vol. 44, no. 11, pp. 5844–5853, 2017. DOI: 10.1002/2017GL072874.

[62] P. Potapov, M. C. Hansen, A. Pickens, *et al.*, "The global 2000–2020 land cover and land use change dataset derived from the landsat archive: First results," *Frontiers in Remote Sensing*, vol. 3, p. 856 903, 2022. DOI: 10.3389/frsen.2022.856903.

[63] European Space Agency (ESA). "Mapping our human footprint from space." Accessed: Apr. 30, 2025. (2025), [Online]. Available at: https://www.esa.int/Applications/Observing_the_Earth/Mapping_our_human_footprint_from_space Last accessed: 04/30/2025.

[64] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, pp. 418–422, 2016. DOI: 10.1038/nature20584.

[65] WorldPop, *Bangladesh 100m population, version 2*, Accessed: Apr. 30, 2025, University of Southampton, 2017. DOI: 10.5258/SOTON/WP00533.

[66] M. Schiavina, S. Freire, A. Carioli, and K. MacManus, *Ghs-pop r2023a - ghs population grid multitemporal (1975-2030)*, Accessed: Apr. 30, 2025, European Commission, Joint Research Centre (JRC), 2023. DOI: 10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE.

[67] C. for International Earth Science Information Network (CIESIN), *Gridded population of the world, version 4 (gpwv4): Population density, revision 11*, Accessed: Apr. 30, 2025, NASA Socioeconomic Data and Applications Center (SEDAC), 2020. DOI: 10.7927/H49C6VHW.

[68] Z. Cheng, A. Van Geen, A. A. Seddique, and K. M. Ahmed, "Limited temporal variability of arsenic concentrations in 20 wells monitored for 3 years in araihazar, bangladesh," *Environmental Science & Technology*, vol. 39, no. 13, pp. 4759–4766, 2005. DOI: 10.1021/es048065f. [Online]. Available at: https://doi.org/10.1021/es048065f.

[69] P. Ravenscroft, R. J. Howarth, and J. M. McArthur, "Comment on "limited temporal variability of arsenic concentrations in 20 wells monitored for 3 years in araihazar, bangladesh"," *Environmental Science & Technology*, vol. 40, no. 5, pp. 1716–1717, 2006. DOI: 10.1021/es058017a. [Online]. Available at: https://doi.org/10.1021/es058017a.

[70] J. Muñoz Sabater, *ERA5-Land monthly averaged data from 1950 to present*, https://doi.org/10.24381/cds.68d2bb30, Accessed: May. 07, 2025, 2019.

[71] Y. Zhang, L. Guo, Y. Chen, *et al.*, "Prediction of soil organic carbon based on landsat 8 monthly ndvi data for the jianghan plain in hubei province, china," *Remote Sensing*, vol. 11, no. 14, p. 1683, 2019. DOI: 10.3390/rs11141683. [Online]. Available at: https://doi.org/10.3390/rs11141683.

[72] C. Crapart, A. G. Finstad, D. O. Hessen, R. D. Vogt, and T. Andersen, "Spatial predictors and temporal forecast of total organic carbon levels in boreal lakes," *The Science of The Total Environment*, vol. 870, p. 161 676, 2023. DOI: 10.1016/j.scitotenv.2023.161676. [Online]. Available at: https://doi.org/10.1016/j.scitotenv.2023.161676.

[73] L. Poggio, L. M. De Sousa, N. H. Batjes, *et al.*, "Soilgrids 2.0: Producing soil information for the globe with quantified spatial uncertainty," *SOIL*, vol. 7, no. 1, pp. 217–240, 2021. DOI: 10.5194/soil-7-217-2021. [Online]. Available at: https://doi.org/10.5194/soil-7-217-2021.

[74] International Soil Reference and Information Centre (ISRIC), *Soilgrids data*, https://www.soilgrids.org, Data available from www.soilgrids.org. Publication date: 2020-05-04. Period: Mar 31, 1905 – Jul 04, 2016. License: CC BY 4.0, 2020. DOI: 10.17027/isric-soilgrids.713396fa-1687-11ea-a7c0-a0481ca9e724.

[75] P. Potapov, M. C. Hansen, A. Pickens, *et al.*, "The global 2000-2020 land cover and land use change dataset derived from the landsat archive: First results," *Frontiers in Remote Sensing*, vol. 3, 2022. DOI: 10.3389/frsen.2022.856903. [Online]. Available at: https://doi.org/10.3389/frsen.2022.856903.

[76] Environment and Population Research Center (EPRC), *Unpublished arsenic concentration dataset*, Courtesy of Environment and Population Research Center (EPRC), Dhaka, Bangladesh, 2024.