Developing a geospatial model to identify private well contamination risk in Alabama

by

Sk Nafiz Rahaman

A thesis submitted to the Graduate Faculty of
Auburn University in
partial fulfillment of the
requirements for the Degree of
Master of Science in Geography

Auburn, Alabama
August 3, 2024

Keywords:
Private well, contamination risk, risk-scape, geospatial method, remote sensing, flood monitoring

Approved by

Dr. Jake R. Nelson, Chair, Assistant Professor, Department of Geosciences
Dr. Ann S. Ojeda, Assistant Professor, Department of Geosciences
Dr. Stephanie R. Rogers, Assistant Professor, Department of Geosciences

Abstract

In the United States, many people rely on private wells, and a constellation of risk factors affect the nature and severity of well water pollution, including the capacity of the well user to effectively manage their water supply. Identifying well user communities at risk of contaminant exposure remains a complex and underexplored area of research. The primary objective of this research is to better understand the spatial distribution and correlation of risk factors associated with the potential contamination of private well water. We developed a framework to evaluate this "risk-scape" using an unsupervised multivariate clustering approach and spatial autoregressive models to evaluate three key risk factors - socio-economic vulnerability, flood risk, and anthropogenic activity – with well water dependence. Our findings show that approximately 15% of Alabama's communities with high well dependence also have a higher flood risk and a large minority group with population below poverty line while 29% of high well use communities are composed of a high percentage of agricultural land with a large child population. This framework highlights where policy intervention or targeted resource allocation should be focused to mitigate well contamination in these communities. The framework's flexibility allows for application to any geographical area, offering a pathway for broad adoption.

Acknowledgements

I would like to express my heartfelt gratitude to Dr. Jake Nelson for the incredible opportunity to conduct this study and work as a research assistant under his guidance. Dr. Nelson has been an exceptional mentor and my first guardian in the United States. His unwavering belief in my abilities has been a source of inspiration, and I have learned immensely from his expertise in both research and teaching.

I extend my sincere thanks to my committee members, Dr. Ann Ojeda and Dr. Stephanie Rogers. Their invaluable support and significant contributions to the manuscript's editing have greatly enhanced the quality of this work. Their insightful comments and thoughtful feedback were instrumental in shaping the study.

I am deeply grateful to my friends at Auburn University who stood by me and supported me throughout my journey. Special thanks to Artat, Ansleigh, and Dikshya, who became my first friends in the United States and provided unwavering support.

I also want to express my profound appreciation to my family. Despite the physical distance, their love and encouragement were always with me. My girlfriend, Nishat, has been a pillar of support and motivation during my application process to Auburn University, and I am deeply thankful for her presence. My parents have always made the right decisions for me, guiding me to this pivotal moment in my life.

Lastly, I thank God for providing me with the right opportunities at the right time, allowing me to reach this significant milestone.

Table of Contents

List of Tables

List of Figures

Abbreviations

ADECA - Alabama Department of Economic and Community Affairs

ALOS - Advanced Land Observing Satellite

ANOVA - Analysis of Variance

BAI - Built-up Area Index

CAFOs - Confined Animal Feeding Operations

CBGs - Census Block Groups

CRA - Contamination Risk Assessment

DEM - Digital Elevation Models

DRASTIC - Depth to groundwater, Recharge rate, Aquifer, Soil, Topography, Vadose zone's impact, Aquifer's hydraulic conductivity

EPA - Environmental Protection Agency

EVI - Enhanced Vegetation Index

GCI - Green Chlorophyll Index

GEE - Google Earth Engine

GRD - Ground Range Detected

HAND - Height Above Nearest Drainage

HHRA - Human Health Risk Assessment

HSD - Honestly Significant Difference

LM - Lagrange Multiplier

MANOVA - Multivariate Analysis of Variance

MCC - Matthews Correlation Coefficient

ML – Machine Learning

MNDWI - Modified Normalized Difference Water Index

MSE - Mean Squared Error

NASS - National Agricultural Statistics Service

NDFI - Normalized Difference Flood Index

NDTI - Normalized Difference Turbidity Index

NDVI - Normalized Difference Vegetation Index

NDWI - Normalized Difference Water Index

NHGIS - National Historical Geographic Information System

NHU - Net Housing Unit

NIR - Near-infrared

NLCD - National Land Cover Database

OOB - Out-of-Bag

PFAS - Poly- and Perfluorinated Alkyl Substances

PRISM - Parameter-elevation Regressions on Independent Slopes Model

PRISM - Parameter-elevation Regressions on Independent Slopes Model

RF - Random Forest

RRM - Relative Risk Model

SARM - Spatial Auto Regressive Model

SAR - Synthetic Aperture Radar

SAVI - Soil Adjusted Vegetation Index

SEM - Spatial Error Model

SLM - Spatial Lag Model

SMI - Soil Moisture Index

SRTM - Shuttle Radar Topography Mission

SSE - Sum of Square Error

SST - Total Sum of Square

SVM - Support Vector Machine

SWIR - Short-wave infrared

TOA – Top of Atmosphere

TRI - Toxic Release Inventory

UAVs - Unmanned Aerial Vehicles

US – United States

USGS - United States Geological Survey

VH - Vertical-Horizontal

VIF - Variance Inflation Factor

VV - Vertical-Vertical

WRI - Water Ratio Index

Chapter 1

Perspective on private well water contamination risk

1.1 Introduction

Ensuring access to a safe and reliable water supply is recognized globally as a human right (United Nations, 2016). For many, the water used for drinking, bathing, and cooking is regulated by a government agency such as Environmental Protection Agency (EPA) that provides oversight and regulations to ensure that public water systems are readily available and free of harmful impurities (US EPA, 2019). For others living in areas that are not serviced by municipal supplies, a primary water source can be groundwater extracted by a private well. According to the United States Geological Survey (USGS), over 43 million people in the US rely on private wells as their primary source of drinking water (DeSimone et al., 2009). Importantly, private wells in the US are not subject to the same regulations as municipal water supplies, and it is the responsibility of the well users to ensure that their well is functioning properly and that water is in sufficient supply and of sufficient quality to serve the household needs (EPA, 2015).

Overall, private wells are more vulnerable than public supplies to a range of contaminants, including nitrate (Wheeler et al., 2015), pathogenic bacteria (Mapili et al., 2022), viruses (Borchardt et al., 2003), and parasites (Borchardt et al., 2021). The general hypothesis is that contaminant sources in the environment can be mobilized and transported to aquifers in which private well users draw from. These contaminant sources can be both geogenic or anthropogenic in nature. For many anthropogenic sources, fate and transport pathways are dominated by precipitation that drives groundwater infiltration, moving the pollutants from the surface into the aquifer. The groundwater is then extracted and consumed by well users. Environmental factors like geologic recharge potential (the rate that water moves down through the sub-surface) and climate events, such as floods, significantly impact the risk of groundwater contamination by facilitating the transport of pollutants from surface to groundwater sources (Geological Survey of Alabama, 2007). Human activities, especially in agriculture and industry, are major contributors to groundwater pollution, with proximity to these activities increasing the risk of contamination by substances like *E. coli*, poly- and perfluorinated alkyl substances (PFAS), and nitrate (Nolan & Hitt, 2006; Resek, 1996).

Floodwater can carry a variety of contaminants, including bacteria, viruses, chemicals, and heavy metals, into private wells. This contamination can lead to serious health risks for residents relying on these wells for drinking water (Ramesh et al., 2023). Areas with high rates of groundwater infiltration are particularly vulnerable to contamination during flooding. The soil and geological conditions in these regions allow floodwaters to seep quickly into the ground, carrying contaminants directly into the aquifers that feed private wells (Ledien et al., 2017). Monitoring flood impacts and well contamination over large and often remote areas is challenging. Remote sensing technologies and machine learning models can provide data and improve the ability of policymakers to identify and respond to high-risk zones effectively. These tools can be crucial in prioritizing areas for intervention and ensuring the safety of drinking water supplies.

Previous research has linked water contaminants to a variety of health problems such as gastrointestinal illnesses, skin infections, and conjunctivitis (Craun et al., 2010). Poor health outcomes among private well users may be exacerbated by financial constraints, limited knowledge of maintenance strategies, and skewed risk perceptions around contamination (Fizer et al., 2018; Imgrund et al., 2011). Additionally, these challenges disproportionately impact low-income and marginalized communities due to socioeconomic factors such as income, education, and access to educational resources that limit a households ability to mitigate contaminant exposure from well water (Martinez-Morata et al., 2022). Studies show that demographic factors, including age, can influence well management practices, with older adults being more vigilant about water quality (Flanagan et al., 2016). Furthermore, Rowles III et al. (2020) recently identified counties with high rates of arsenic (environmental risk) and flooding (infiltration risk), and compared that to areas with high concentrations of mobile homes (socio-economimc vulnerability indicator) to uncover the areas most susceptible to experiencing (negative) shifts in groundwater quality.

1.2 Research questions and hypothesis

Generally speaking, research regarding exposure risks for private well users focus on one of three domains: the socioeconomic characteristics of the well user (Malecki et al., 2017), the anthropogenic contaminant sources (Roostaei et al., 2021), or the natural/environmental contamination risks (Eccles et al., 2017). On their own, each has been shown to contribute to an increased risk of contaminant exposure, however, it is at the confluence of these factors that more

attention is needed. The first objective for this research is to evaluate the spatial distribution of known risk factors across the landscape and determine the areas where these risks may be most likely to influence private well water resources. The objective answers the following research question:

Research Question 1: How do the spatial distributions of socioeconomic vulnerabilities, anthropogenic contamination sources, and natural/environmental contamination risks converge to risk the private well water in Alabama?

Hypothesis 1: Areas in Alabama where high socioeconomic vulnerabilities, dense anthropogenic contamination sources, and significant natural/environmental contamination risks intersect will exhibit poorer private well water quality compared to areas where these risk factors are not as prevalent.

In order to address this objective, a spatially explicit analysis framework has been applied to develop what has been coined as a "risk-scape" of well water quality. Risk-scape, in the context of this research, refers to the spatial landscape where various risk factors such as anthropogenic contamination sources, socioeconomic vulnerabilities, and natural/environmental contamination risks converge to influence well water quality.

The second objective of the research shifts towards utilizing remote sensing methods rather than spatial analysis, adopting machine learning models to detect flooding in Mobile and Baldwin counties in Alabama. Since flooding is a major environmental risk factor for private well contamination, this research extensively explores flood monitoring in smaller scale areas and identifies zones where high flooding coincides with high private well use. The objective answers the following two research question:

Research Question 2: How multiple open-source remote sensing datasets can support each other and adopt machine learning models to improve inland flooding?

Hypothesis 2: Machine learning models will enhance the ability to monitor floods by effectively analyzing large volumes of open-source remote sensing data that will lead to more precise identification and prediction of flooded areas in the study region.

Research Question 3: To what extent do high flood zones overlap with high private well use zones?

Hypothesis 3: There will be a significant overlap between high flood zones and areas of high private well use in Mobile and Baldwin counties, indicating that these regions are at an increased risk of private well water contamination due to flooding.

Chapter 2

Applying a geospatial modeling framework to evaluate private well water contamination risk

2.1 Background

2.1.1 Contamination Risk Factors

2.1.1.1 Socioeconomic characteristics

While the risks associated with contaminated well water are important considerations for all well water users, they disproportionately affect historically underrepresented communities (Martinez-Morata et al., 2022). Socioeconomic factors, including income levels, race, and access to information, have been identified as significant determinants of the likelihood and severity of well contamination incidents. Specifically, the ability to monitor private well water quality and take steps to mitigate any contamination depends, at least in part, on a household's demographics and socioeconomic status (Flanagan et al., 2016). Recent studies reveal that although people may recognize the factors contaminating well water quality, they are not confident enough to manage their wells effectively (Osidach, 2021). For example, they may lack a complete understanding of the extent to which contaminants threaten their water supply (Flanagan et al., 2015, 2016; Osidach, 2021). Mooney et al. (2022) also revealed that demographic factors such as the age of household members also influence well management. Specifically, older adults tend to be more aware of their household's water quality, which acts as a mitigating factor against well contamination. Conversely, young people are often more apathetic, leading to a possible risk of improper well stewardship.

The study by Flanagan et al. (2016) indicates that while socially vulnerable groups may not inherently be at a higher risk of water contamination, their increased vulnerability is largely due to socio-economic factors. These groups, which often include low-income families and marginalized communities, may lack sufficient education or resources to fully understand the importance of proper well maintenance. Additionally, financial constraints can hinder their ability to regularly test water quality and maintain well systems. This combination of limited knowledge and financial barriers means that these groups are more likely to consume contaminated water as they might not have the means to ensure their drinking water is safe. Considering recent research

that finds minority populations (e.g., Hispanic/Latino and American Indian/Alaskan Native residents) tend to reside closer to areas with high concentrations of well-known groundwater contaminants such as arsenic and uranium (Martinez-Morata et al., 2022; Rowles et al., 2020), it is increasingly important to identify socially vulnerable well owner communities in relation to known well water contamination risks.

2.1.1.2 Environmental considerations

Contaminants on the ground surface can infiltrate into the subsurface, which is largely driven by gravitational forces, permeability of the soil, and existing moisture content (Geological Survey of Alabama, 2007). As the water moves through the soil profile, contaminants can be attenuated by processes such as adsorption, biodegradation, and chemical reactions, reducing their concentration. However, some contaminants, especially those that are highly soluble in water or non-reactive, can accumulate in groundwater over time. For instance, in agricultural areas, excessive use of fertilizers can lead to high concentrations of nitrate in the groundwater, posing a risk to drinking water sources (Nolan & Hitt, 2006). Similarly, improper disposal of industrial waste can introduce harmful chemicals into the subsurface, which can eventually migrate to groundwater resources (Resek, 1996).

Environmental factors like geography and climate significantly influence the risk of well water contamination. For example, in a study examining a 2013 flood event in Alberta, Canada, Eccles et al., 2017 reported a significant increase in the presence of *Escherichia coli* (E. coli) in private drinking wells post-flood, demonstrating the connection between flooding and the transport of contaminants to well water. Similar work by Rowles III et al. (2020) characterized the same effect, finding that seasonal contamination of wells in flood prone areas was signifncant and, importantly, was dependent on the amount of precipitation. A U.S.-based study focused on the impact of four major natural disasters— the 2016 Louisiana Floods, Hurricane Harvey in 2017, Hurricane Irma in 2017, and Hurricane Florence in 2018 found elevated levels of bacteria—such as Legionella, Mycobacterium, L. pneumophila, and M. avium—that are responsible for waterborne diseases in private well samples taken after these events (Mapili et al., 2022; Pieper et al., 2021). Together, this research suggests that areas prone to frequent flooding are at a higher risk for contaminant exposure compared to regions with lower flood potential.

2.1.1.3 Anthropogenic contaminant sources

Given the previously discussed mechanisms of contaminant transport to private wells, human activities, particularly those involving toxic release facilities and agriculture, have a long history of increasing risks to groundwater contamination. For example, a 1997 study conducted in Ontario, Canada, found an inverse relationship between the presence of E. coli in household wells and the distance of these wells from adjacent farmland (Goss et al., 1998). The transformation of grasslands into cultivated fields has also been shown to increase the occurrence of nitrate in private well water which was attributed to the application of nitrate-rich agricultural chemicals on the fields (Keeler & Polasky, 2014). More recent studies note similar spatial relationship related to PFAS emitters. Private wells closer in proximity to PFAS facilities were more likely to test positive for PFAS (Hu et al., 2021; Roostaei et al., 2021) and maintain higher concentrations of PFAS compared to background wells. Roostaei et al. (2021) indicates that PFAS contamination risk was driven by air deposition rate, wind direction, and by the distance the well was to the emission source. They also noted that PFAS occurrence in the wells was driven less by groundwater recharge potential or high-water events, highlighting the importance of considering multiple transport mechanisms in the evaluation of contamination risk.

It is important to acknowledge that pollutants related to anthropogenic processes have been known to interact synergistically with natural processes that exacerbate overall contamination risk (X. Li et al., 2020). For instance, industrial activities can introduce contaminants like sulfonamide antimicrobials into groundwater, as evidenced by their detection in private wells in areas near confined animal feeding operations (CAFOs) (Batt et al., 2006). Agrichemicals, such as herbicides and nitrate, are significant sources of diffuse pollution in groundwater, particularly in agricultural regions. This type of pollution is caused by runoff from agricultural fields and can lead to elevated levels of nitrates and other harmful chemicals in groundwater sources (Burkart et al., 1999). In addition, industrial activities, particularly those related to petrochemical enterprises, can discharge polluted wastewater into the environment. This wastewater often infiltrates the soil, contaminating the groundwater with various pollutants like total petroleum hydrocarbons, total dissolved solids, chlorides, and sodium (Radelyuk et al., 2021).

2.1.2 Well contamination risk assessment methods

In evaluating the risk of well contamination, researchers have adopted a multi-disciplinary approach, integrating insights from both social and environmental sciences. The primary objective of these studies is to predict and map the locations of contamination risks, providing critical support for understanding the broader patterns of risk. However, despite the progress, there is a notable gap in fully integrating these diverse approaches into a unified model.

Regression analysis, notably logistic regression, stands as a cornerstone in this domain. For instance, Eccles et al. (2017) employed regression models to uncover a significant link between flood conditions and *E. coli* presence in groundwater wells, revealing that well maintenance, often influenced by the socioeconomic status of users, plays a more crucial role in contamination risk than geographic location. Similarly, Ayotte et al. (2017) achieved over 80% accuracy in predicting geogenic contamination risks in private wells using logistic regression, emphasizing the method's effectiveness in identifying risk areas.

Emerging machine learning (ML) techniques, such as boosted regression trees and random forest classification, have shown even greater promise. Lombard et al. (2021) reported accuracy rates exceeding 90% in predicting contaminant exposure, while sensitivity analyses have enhanced the reliability of these assessments (Spaur et al., 2021). These advanced methods underscore the potential of ML in refining risk assessment, although they often lack integration with socio-economic data. The U.S. EPA's Priority Setting Approach represents a more holistic framework, considering various factors such as contaminant release, transport, and toxicity. However, this approach falls short in integrating hydrogeological variables like hydraulic conductivity, that others have identified as a significant contamination factor elsewhere in the U.S (Harman et al., 2001).

The integration of the Groundwater Relative Risk Model (RRM), Groundwater Contamination Risk Assessment (CRA), and Human Health Risk Assessment (HHRA) in studies by Teng et al. (2019) and Sresto et al. (2021) demonstrates the potential of cumulative approaches. These models utilize risk values and the Groundwater Vulnerability Index based on the Depth to groundwater, Recharge rate, Aquifer, Soil, Topography, Vadose zone's impact, Aquifer's hydraulic conductivity (DRASTIC) model to yield regional assessments (Aller et al., 1987). Additionally, statistical cluster analysis and Binary Hierarchical Logistic Regression, as explored in studies by

Hynds et al. (2014), Krolik et al. (2013), and Mooney et al. (2021), pinpoint ground-level risk areas, providing insights into the distribution of individual pathogenic contamination risk factors such as *E. coli*.

Despite these advancements, existing studies have not fully explored how social, natural, and anthropogenic risk sources combine to form comprehensive patterns of risk across landscapes. The deficiency in integrating all these elements into a single, comprehensive model hinders the complete understanding of contamination risk sources and the communities vulnerable to exposure. This highlights the need for a more inclusive approach that combines data describing social condition, environmental and anthropogenic contamination sources to form a holistic understanding of well water contamination, thereby aiding in the effective prediction and management of the risks. Such integration is essential for addressing the current gap in knowledge and for developing robust strategies to mitigate well contamination risks.

2.2 Methods

2.2.1 Study area

This study focuses on the state of Alabama in the US. Alabama has a total population of around 5.04 million people, with approximately 16% of the population using private well water

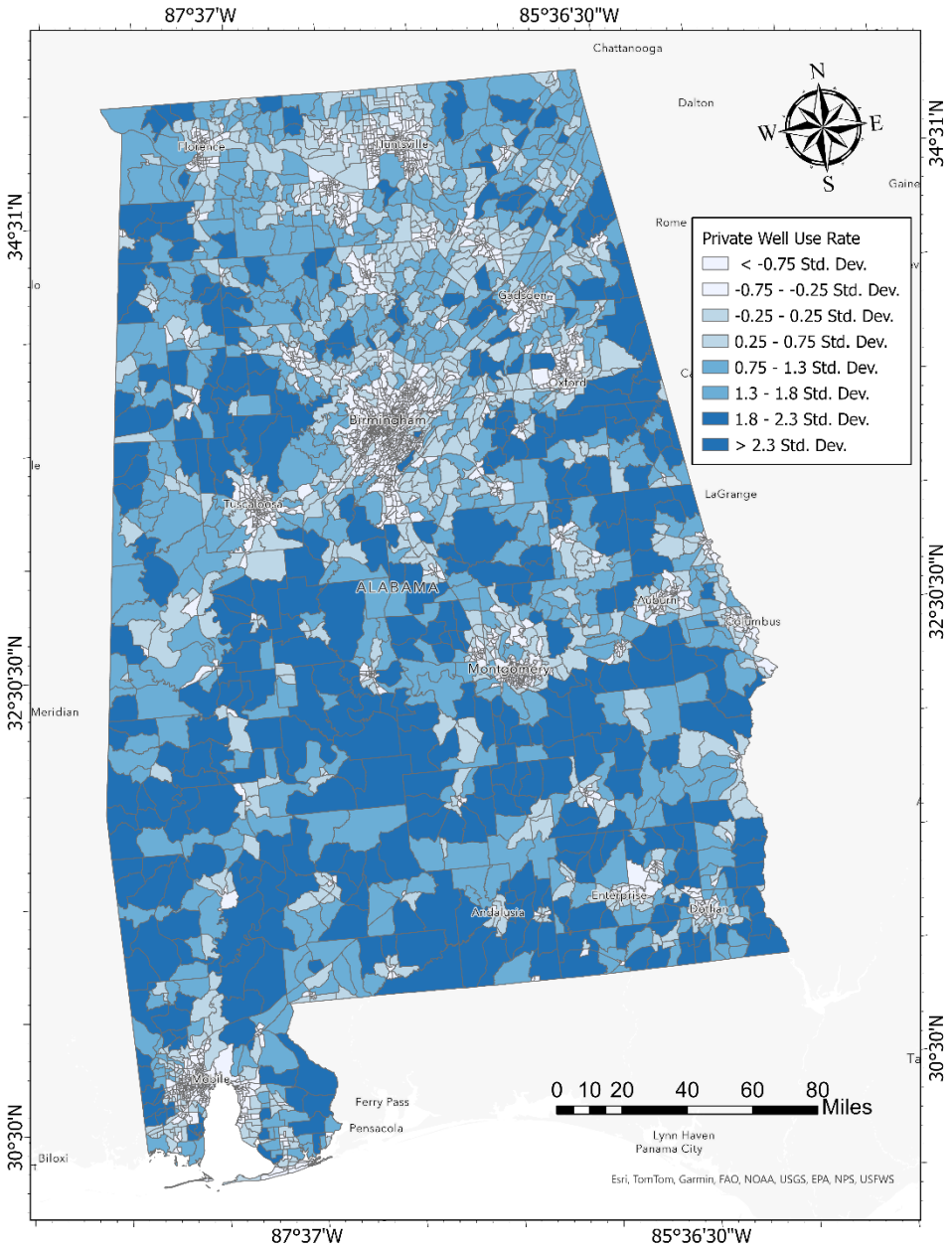


Figure 1 Block groups of Alabama and estimated rate (categorized using standard deviation) of housing units using private water supply sources. (A. H. Murray & Kremer., 2023)

for domestic use (ADPH, 2021). Figure 1 shows the study extent with the block group wise estimated distribution of private well use rate. The use rate is the estimated number of households reliant on a private well divided by the total number of households in the unit of analysis. Approximately 14% of people in Alabama have a high school education or less, 32% of the population is of minority groups, and 16% of people in Alabama are living below the poverty line (US Census, 2021). According to the Alabama Department of Economic and Community Affairs (ADECA), the southwest coastal regions of Alabama has the highest water surface elevation areas, which increases the risk of floods (ADECA, 2020).

2.2.2 Data background and preprocessing

We leveraged a recently published dataset of estimated counts of housing units using private well water sources from Murray & Kremer (2023) for the analysis. The original approach for creating this data set estimated well use rate for all US Census block groups (CBGs) in the US based on information from the 2010 US census. The approach utilized two methods. First, a reported wells method combined housing unit change with private well drilling logs for a sample of 20 states and considered variable well use rates from 1990 (the last time the US census asked about household water supply) to 2010. Second, the net housing unit (NHU) method was developed that assumed a constant well use rate based on household density. Their analysis showed a reliable relationship between well users and house hold density (R^2 of .78) which was used to extrapolate well estimates to all CBGs across the US (A. Murray et al., 2021). The 2010 model employed a simple linear regression that worked to an extent but struggled in areas with high housing unit growth. To address these complexities the authors have more recently adopted a machine learning approach that uses decision trees to estimate where census blocks have access to public water systems while validating it against known public water system boundaries (Murray & Kremer, 2023). The new model is more robust and has been recently applied to the 2020 census data. We use this new data set to estimate the percentage of housing units reliant on private wells across our study extent (A. H. Murray & Kremer., 2023).

Generally speaking, few CBGs contained the majority of the well users in a region, resulting in a highly right-skewed distribution. The vast majority of CBGs will not have any households using well water. To prepare for spatial and statistical analysis that require near-normal distributions, we applied a Box-Cox transformation to stabilize the variance (Box & Cox, 1964).

This transformation not only made the dataset more Gaussian but also paved the way for a more robust statistical interpretation, in line with the principles outlined by Johnson (2000).

Socioeconomic data for each CBG were sourced from the National Historical Geographic Information System (NHGIS), a project funded by the National Science Foundation for a five-year term (Manson et al., 2023). NHGIS aggregates U.S. census data across various geographic aggregations offering both geographic and attribute data that come equipped with the same unique IDs provided in the well use dataset (McMaster & Noble, 2005; Schroeder & McMaster, 2007). A variety of socioeconomic variables were considered, each of which related in some way to the social vulnerability of the CBG. We detail the complete list of socioeconomic variables in Table 1.

To investigate flood risk factors, we considered a variety of conditioning variables such as Digital Elevation Models (DEM), streamline density, and net recharge rates, as described by Tehrany et al., (2019) and presented in Table 1. In hydrology, a streamline is a path that water follows in a stream or river, typically influenced by the topography and shape of the riverbed. The connection between streamlines and floods is significant, as changes in streamlines, such as blockages or alterations in riverbed geometry, can influence the flow of water and potentially lead to flooding, especially during heavy rainfalls or sudden water input events. For instance, the disruption of streamlines can increase water contributions during flooding and affect the transport times of water flows, magnifying the risks and impacts of flood events (Hasenmueller & Robinson, 2016).

We also incorporated 30-year average net recharge data from the Geological Survey of Alabama and combined it with 30-year average precipitation data from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) (Geological Survey of Alabama, 2007; PRISM Climate Group, 2022). The PRISM raster, initially at an 800-meter spatial resolution (PRISM Climate Group, 2022), was resampled with the Nearest Neighbor method to 30 meters (Ver Hoef & Temesgen, 2013) because there are block groups with sizes of less than 800 square meters. To extract data accurately for a feature, it is important to have the raster grid cell smaller than the feature (Lechner et al., 2009). We additively integrated block group wise average of all these factors to calculate flood risk.

We utilize two primary indicators of anthropogenic risk in our analysis which include areas of agricultural activity and the location of facilities that emit toxic chemicals into the environment.

Agricultural activities utilize fertilizer that contains known groundwater contaminants such as nitrate. To account for these areas we collected 2020 agriculturally active areas for Alabama from the National Agricultural Statistics Service (NASS) (USDA, 2020) which is provided at 30 m resolution. We recognize that groundwater contamination from agricultural land use practices is a time-delayed effect that is difficult to predict at large (and heterogeneous) spatial scales. Still, we follow the precedent set by Kolpin (1997) using land use classifications to define agricultural areas as proxy for contaminants associated with land use practices.

Next, we leveraged the Toxic Release Inventory (TRI) dataset made available through the EPA to identify facilities and industrial sites known for releasing toxic chemicals (US EPA, 2023). The TRI program is a key resource for tracking and managing environmental releases of toxic chemicals in the US. Established under the Emergency Planning and Community Right-to-Know Act of 1986, the TRI program mandates that facilities in certain industries report annually on the quantity of toxic chemicals they release into the air, water, and land. This data includes information on waste management activities and pollution prevention efforts. The TRI dataset provides a comprehensive overview of environmental releases, but more importantly, offers valuable insights into the types, quantities, and locations of chemicals released by these facilities. Following previous work on the strong distance decay effect related to toxic emitters and groundwater contaminant concentration (Roostaei et al., 2021), we assume that distance and prevalence of these facilities is a strong indicator of potential well contamination. We calculated the kernel density of the facilities across Alabama as a proxy for contamination risk.

All contaminant risk factors were aggregated to Alabama CBGs for further analysis. Before proceeding with the analysis, it is essential to standardize the values of each factor from 0 to 1. This standardization process ensures uniformity in the scale of measurement for each factor, facilitating a more accurate and comparable assessment across different variables. It also minimizes potential biases arising from different units of measurement and varying ranges of data. Table 1 details all the variables that have been used in this study along with the source of the data.

Table 1 Variables for the model, description, and source of the data

Data	Description (Per block groups)	Source
Domestic Well Data		
Well Data	Estimated rate of housing units using private water supply sources within the census block in 2020.	(A. H. Murray & Kremer., 2023)
Socio-economic and Demographic Data		
Minority Percentage	Estimated percentage of Black or African American, American Indigenous, Asian, Native Hawaiian, Hispanic, and Other	National Historical Geographic Information System (NHGIS) database from Integrated Public Use Microdata Series (IPUMS). American Community Survey (ACS) 2018-2022 (U.S. Census Bureau, 2022).
Education less than high school	Estimated percentage of people who have attained not more than high school	
Income less than poverty line	Estimated percentage of people who have less than 0.99 ratio of income to poverty level in the past 12 months.	
Child population	Estimated total percentage of children (Age<18)	
Flood Risk Data		
Precipitation	30-year average (800 meter)	PRISM Climate Group
DEM	Average DEM (30 meter) (Reversed)	NASA Shuttle Radar Topography Mission (SRTM)
Streamline	Average streamline density	(U.S. Census Bureau, 2020)
Net recharge	30-year average (30 meter) (Reversed)	(Geological Survey of Alabama, 2018)
Anthropogenic Risk Data		
TRI	TRI Density	(US EPA, 2023)
Agriculturally active areas	Percentage of agriculturally active areas per block groups (Herbaceous,	(USDA, 2020)

	Pasture/Hay, and Cultivated Crops)	

2.2.3 Spatial Analysis

Different configurations of contaminant risk variables can come together to influence contaminant exposure differently. For example, some CBGs may have high well use, high environmental risk, and high anthropogenic risk but occur in CBGs that are well equipped from a socio-economic standpoint and therefore more likely (or able) to mitigate the risks. On the other hand, you could have the same high values of well use, environmental risk, and anthropogenic risk but in an area that is less likely to be able to mitigate the risks due to socio-economic constraints. There are many configurations of variables that contribute to differences in contaminant exposure and in order to capture these dynamics we leverage an unsupervised cluster segmentation process to group CBGs of "like" values to create a typology for evaluating how communities vary in their level of contaminant exposure.

2.2.4 Multivariate clustering

We have used Multivariate clustering analysis (Caliñski & Harabasz, 1974; Jain, 2010), a form of unsupervised classification, to aggregate CBGs into distinct groups that share similar attribute patterns that contribute to varying levels of private well contamination risk. Specifically, we were interested in identifying CBGs associated with relatively high well use rate and their relationship to each of the factors identified as contributing to well contamination risk. The goal of multivariate clustering is to identify groupings where the features within each cluster exhibit maximum similarity, while the clusters themselves remain as distinct as possible. Of particular importance is the identification of well use communities associated with low socioeconomic status, high flood potential, and a comparatively high density of TRI facilities or surrounding agricultural land. We used the K-Medoids algorithm for the clustering process and implement it in ArcGIS Pro (ESRI, 2024). This form of multivariate clustering uses Calinski-Harabasz pseudo-F-statistic (Caliñski & Harabasz, 1974) to determine the optimal number of clusters in the dataset. It is a ratio reflecting within group similarity and between group differences:

$$Ratio = \frac{\frac{R^2}{n_c - 1}}{\frac{1 - R^2}{n - n_c}} \quad (1)$$

Where,

$$R^2 = \frac{SST - SSE}{SST} \quad (2)$$

SST is a reflection between cluster differences and *SSE* reflects within-cluster similarity.

$$SST = \sum_{i=1}^{n_i} \sum_{j=1}^{n_c} \sum_{k=1}^{n_v} \left(V_{ij}^k - \overline{V^k}\right)^2 \quad (3)$$

$$SSE = \sum_{i=1}^{n_i} \sum_{j=1}^{n_c} \sum_{k=1}^{n_v} \left(V_{ij}^k - \overline{V_t^k}\right)^2 \quad (4)$$

n = the number of features

n_i = the number of features in cluster i

n_c = the number of classes (clusters)

n_v = the number of variables used to cluster features.

V_{ij}^k = the value of the k^{th} variable of the j^{th} feature in the i^{th} cluster

$\overline{V^k}$ = the mean of the k^{th} variable

$\overline{V_t^k}$ = the mean of the value of the k^{th} variable in cluster i

The clusters are generated by iteratively grouping each CBG with the cluster centroid in which it is closest. When subsequent iterations no longer improve the Total Sum of Square (SST) or the Sum of Square Error (SSE) – that is differences between and similarities within are maximized, the iterations stop and you are left with k-groupings of CBGs based on their attribute values. Following the pseudo-F-statistics our analysis resulted in two clusters. However, given our desire to distinguish CBGs based on our list of contaminant risk factors (Table 1) and in an effort to provide more nuance into where and why CBGs were at risk of contaminant exposure,

we opted to increase the final k-cluster number to 5. Given this choice we provide a deeper interrogation of the clusters to determine whether they are significantly different from one another using a series of statistical tests and exploratory spatial regression.

2.2.5 Spatial Autoregressive Model

The interpretation of clusters has an inherent limitation due to the aggregation. While we are able to assess the characteristics of the cluster members on average, it is difficult to tease apart the specific intra-cluster relationships across their characteristics. To that end, after identifying the clusters we employ several regressions to investigate the relationship between social and environmental factors and private well water use within the different clusters. The autocorrelative nature of private well use necessitates the use of spatial auto regressive models (SARM). We model the spatial relationship of the cluster members using a Queen Contiguity weighted matrix and combine that with Lagrange Multiplier (LM) tests to determine the specific model form spatial lag model (SLM) or a spatial error model (SEM) that is appropriate for the data. The SLM is expressed as:

$$Y = \rho WY + X\beta + \epsilon \quad (5)$$

where Y is the dependent variable vector representing well water use, ρ is the spatial lag coefficient, W is the spatial weights matrix, X is the matrix of independent variables, β is the vector of coefficients, and ϵ is the error term. The spatial weights matrix W is a key component, defining the spatial relationship between observations based on proximity, with closer observations assigned higher weights. The Spatial Error Model (SEM) is formulated as follows:

$$Y = X\beta + \mu \quad (6)$$

$$\mu = \lambda W\mu + \epsilon \quad (7)$$

Where μ is the vector of spatially autocorrelated error terms. λ is the spatial autoregressive coefficient for the error term, and β is a matrix of explanatory variables of interest.

2.3 Results

2.3.1 Multivariate clustering map

Figure 2 displays the multivariate clustering map illustrating well use in conjunction with various risk variables outlined in Table 1 and Table 2. The characteristics of each cluster were summarized through the box plot presented in Figure 3. Each box represents the overall distribution of the variable, the points represent the cluster medoid for that attribute (median), and each line connecting the cluster centers helps to illustrate high/low changes across clusters centers. In addition, we performed additional statistical evaluations (ANOVA, MANOVA, and Tukey HSD) to determine the uniqueness of variables making up the clusters included, which are detailed in Appendix I.

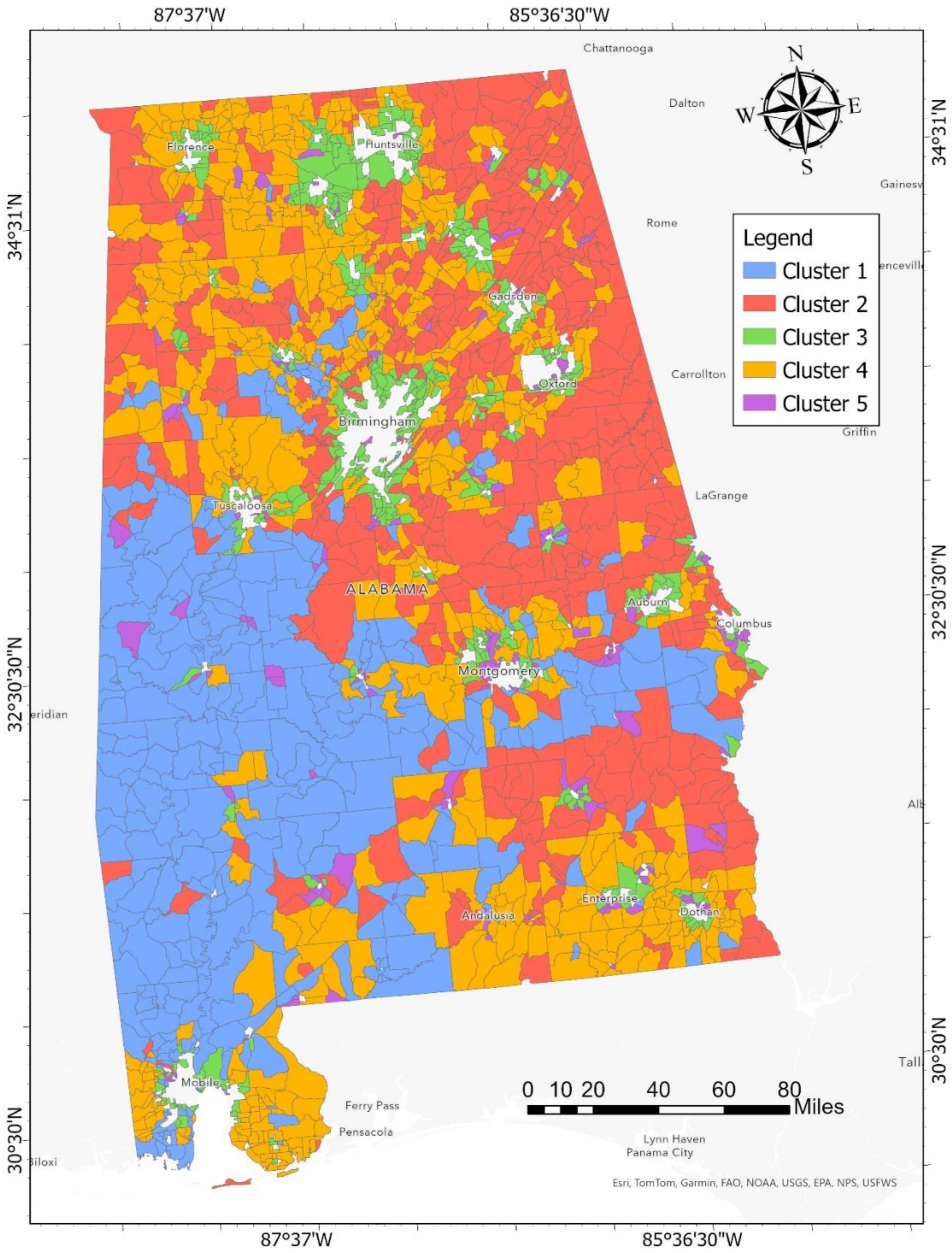


Figure 2 Multivariate clustering map of well use and all the risk factors.

Clusters 1, 2, and 4 are associated with the highest median well use rate denoted with the overlapping points on well use in Figure 2. The distribution of well use rate in each cluster is illustrated in Appendix I. We can therefore assume that the CBGs making up these clusters have some of the highest numbers of well users in Alabama on average. Hence, they warrant further attention.

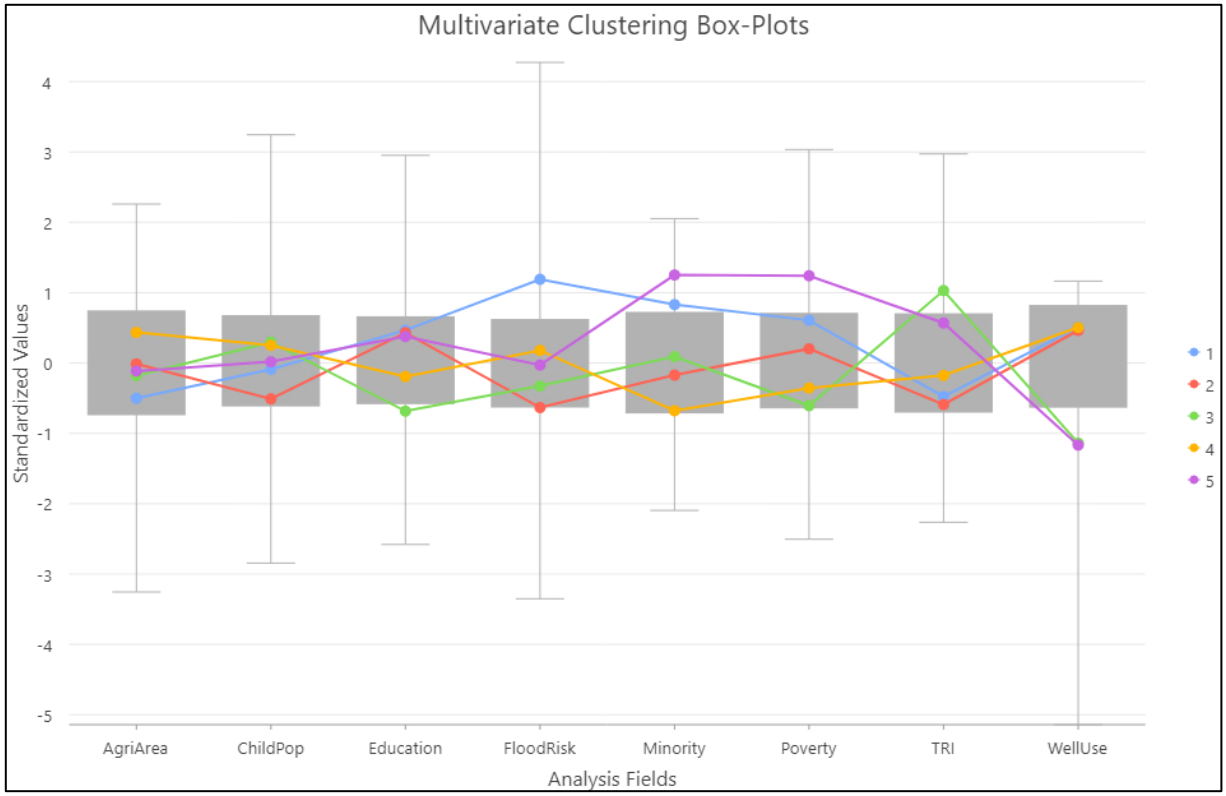


Figure 3 Multivariate clustering boxplot. Each box shows the distribution of each variable, and the five lines represent each cluster in the map.

CBGs within cluster 1 are associated with relatively high minority populations, high poverty, less education and a large percentage of these CBGs are at heightened flood risk. The CBGs in cluster 1 contain a low TRI density on average and high well use. There are comparatively fewer children making up the population, and low agricultural land use. From these groupings we'd expect households in these CBGs to have a high cumulative risk among the high well use clusters.

CBGs making up cluster 2 are less at risk of contaminant exposure than CBGs in cluster 1. This is predominately due to these CBGs having the lowest flood risk combined with moderate

agricultural land, minority and poverty. However, households in cluster 2 CBGs have the second highest percentage of people with less education.

CBGs in cluster 4 make up the highest well-use cluster. Compared to cluster 1, we view this cluster as having a low risk of contaminant exposure. Flood risk is not as prevalent among cluster 1 CBGs, however, the child population is relatively high compared to cluster 1 and 2. The percentage of CBG area devoted to agricultural land use practices is the highest. TRI density is moderate among these cluster along with the number of people with less education. CBGs in cluster 4 also have low minority community and poverty as well.

The R-squared values depicted in Table 2 offer insightful information regarding each variable's contribution to separating and assigning CBGs to each cluster. These values range between 0 and 1, where a higher value denotes that a greater proportion of variance in the model is accounted for by the variable. In other words, high values indicate importance for distinguishing the clusters from one another. The variable for well use rate demonstrates the highest R-squared value indicating that approximately 56.3% of the variance in the clustering is accounted for by the well use rate variable. Following well use, the TRI density accounts for roughly 39.3% of the variance within the model followed by minority population that explains approximately 37.2% of the variance in the model. Other variables such as flood risk, poverty, level of education, and child population have R-squared values of 35.1%, 30.4%, 20.6%, and 10.6%, respectively. Agricultural area has the lowest discriminatory power, only explaining approximately 10.2% of the variance.

Table 2 R-squared value of multivariate clustering

Variable	**R-squared value**
Well use	0.563
TRI density	0.393
Minority population	0.372
Flood risk	0.351
Household below poverty level	0.304
Education	0.206
Child population	0.106
Agricultural area	0.102

2.3.2 Intra-cluster well use rate and risk factors.

The use of clusters and their centroids provide a broad, aggregate understanding of the risks associated with each cluster. However, the aggregation can mask some of the specific relationships between the risk variables of interest and private well dependence within each cluster. Moreover, any inferences made at the cluster level risks violating the ecological fallacy. To help address this deficiency, we performed several regressions using those CBGs that make up each cluster. Prior to running the analysis, we perform several spatial diagnostic tests which indicated the presence of spatial autocorrelation and therefore required the use of a SARM which can account for this spatial dependence. We performed LM tests to identify the specific SARM form most suitable for addressing the autocorrelation in the data set. The results of the tests are indicated in Table 3. The LM test that was most significant was chosen for the analysis. The Variance Inflation Factor (VIF) for each independent variable indicates no significant multicollinearity among the models (Appendix I).

Table 3 Spatial dependence diagnostic test results. LM and Robust LM.

Cluster	**Moran's I (error)**	**LM (lag)**	**Robust LM (lag)**	**LM (error)**	**Robust LM (error)**
1	0.242***	14.192**	7.893**	31.545***	25.246***
2	0.275***	7.098**	1.419	58.087***	52.407***
3	0.113**	0.010	0.435-	7.024**	7.449**
4	0.236***	0.570	4.645*	47.336***	51.411***
5	-0.083	0.418	0.172	1.279	1.032
Significance levels are indicated as: *<.05, **<.01, ***<.001.					

Table 4 displays the results from spatial regression models that provide additional nuance to the characteristics of the population that fall within each of the clusters. As previously mentioned, clusters 1, 2, and 4 contain the highest well users while clusters 3 and 5 have relatively modest well user populations. Because the clustering algorithm produces unequal numbers of cluster members, we report the total number of CBGs used for each regression along with R^2, Rho or Lambda. Coefficients for poverty are negative for cluster 1 and 3, however, positive for cluster 2, 4, and 5. Coefficients for education is significant 3 ($p < 0.05$) and 4 ($p < 0.01$) with a positive

values, however, not significantly positive for cluster 2 and 5. The number of minority households relying on private wells does not significantly increase on average across cluster 2 and 5 at the level of the CBG, while the portion of the population composed of children varies across clusters with no significant positive or negative relation with well use rate.

The TRI variable, which measures the density of known toxic chemical emitters, consistently shows a significant negative relationship with well use rates in Clusters 1 ($p < 0.001$), 2 ($p < 0.001$), and 4 ($p < 0.001$) and cluster 5 ($p < 0.05$). However, in Cluster 3, the effect is non-significant. The variable representing agricultural area within each CBG is significant and positive in Cluster 3 ($p < 0.001$) and Cluster 5 ($p < 0.001$), indicating that within these regions private well locations are commonly found among agriculture areas. In other clusters, the effects are non-significant, with coefficients ranging from -0.010 to 0.014. Flood risk does not significantly affect well use rates in any of the clusters. The coefficients range from -0.047 to 0.032, all of which are non-significant, suggesting no statistical evidence for an impact of flood risk on well use. That said, the relationship between flood risk and private well use, although insignificant, is positive in clusters 2, 4, and 5.

Table 4 Spatial regression model results

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
CBGs Count	321	531	469	621	162
Model Type	SEM	SEM	SEM	SEM	OLS
Constant	1.069***	0.928***	0.729***	0.945***	0.574***
Poverty	-0.043	0.015	-0.045	0.009	0.049
Education	-0.007	0.033	0.060*	0.049**	0.038
Minority	-0.046*	0.007	-0.028	-0.025*	0.042
Child Pop	-0.030	-0.029	0.008	-0.016	-0.020
TRI	-0.209***	-0.098***	-0.024	-0.073***	-0.092*
Agri Area	0.014	-0.007	0.119***	-0.010	0.243***
Flood Risk	-0.026	0.032	-0.047	0.025	0.028
Lambda/Rho	0.402***	0.374***	0.148**	0.307***	-
R-squared	0.358	0.189	0.111	0.158	0.251

Significance levels are indicated as: *<.05, **<.01, ***<.001.

2.4 Discussion

This study takes a geospatial approach to generate a risk landscape, or risk-scape, associated with private well use across Alabama. We evaluated three different risk categories with this approach: socio-economic, flood potential and anthropogenic. The cluster-based approach allows one to understand the risk-scape across the study area by showing where, for example, socio-economically vulnerable well users are co-located with areas of high flood risk and toxic release facilities. One of the benefits of this approach is that it does not make any assumptions regarding the magnitude of risk posed by any of the indicators. Rather, it provides a mechanism for clustering well communities by similar risk categories while disentangling the most likely source of contamination that well users within that community may face. We therefore see this approach as a prospective decision support tool that enables one to communicate and deploy groundwater contamination mitigation strategies that meet the needs of the specific well user community more effectively. There are several important facets of this work worth further discussion.

First, one key contribution of this study is the development of a mechanism for evaluating the interconnections of several risk factors and associated demographics simultaneously. In effect, it allows one to determine the number and type of risk factors a community may face. For instance, socio-economically disadvantaged communities are often more vulnerable to environmental risks and less equipped to respond to contamination incidents (Flanagan et al., 2016). The approach used here helps clarify where and to what extent this might be taking place. It is also worth mentioning that in segmenting the population by risk variables, this approach does not make any assumptions about which risk factors contribute more to well water contamination than others. It merely provides an illustrative risk-scape that decision makers may use to identify and deploy resources that reflect the likely risks and needs of the underlying populations.

Second, by incorporating both a cluster analysis and regression, this framework is at once exploratory and explanatory. As noted previously, the location of private wells is uncertain. The best we can currently do is rely on estimates generated from statistical relationships (A. Murray et al., 2021). However, what we know from those relationships is that well users tend to cluster in

space, so it is therefore important to incorporate well use estimates directly into a broader risk-based approach to develop a baseline understanding of well dependence. In addition, while the clustering of well use rate and socio-economic characteristics do not actually *explain* well use prevalence, it nevertheless provides an excellent starting point for conducting explanatory analyses regarding the factors that may pose contamination risks. In this sense, our combined approach of generating clusters and subsequently evaluating them through regressions minimizes the potential for violating the ecological fallacy while confirming (or refuting) the larger trends at the highest level of CBG aggregation (clusters). The value of this multi-method approach was demonstrated by our results that revealed some mismatch between the aggregated risk factors at the cluster level, and the coefficient estimates generated from the intra-cluster regression analysis. However, it also provided additional evidence that well users in clusters 2, 4, and 5 are especially at risk from flood related well water issues. This relationship was partially revealed in the cluster analysis but was made clear when performing the regressions.

Finally, and perhaps the most important point, is that in the US the location of private well users is difficult to know for certain. Many estimates rely on well water location data spread across multiple data sets of varying quality, age, and information. This, in turn, makes the use of a cluster-based risk approach even more important. As detailed previously, we determined that patterns of well users cluster in space. This suggests that even if we do not know exactly where well users are located, it is likely that they are within the same general area as other well users. Moreover, it means that they face the same risks (if any) as the well users within each cluster. Developing a baseline for knowing where private well use is most prevalent and the risks that those well users are likely to face can help inform preventative mitigation strategies for well water contamination. Although it is difficult to speculate the degree that well use rate may be under- or overreported, there is enough evidence from this analysis to help private well water programs and other stakeholder groups to create and deploy outreach activities in the cluster segments with the highest well use rate and multiple contaminant threats.

2.5 Conclusion

The study, while comprehensive in its approach to understanding private well water contamination risk, presents certain challenges that open avenues for future research and improvement. One of the primary concerns is the reliance on existing datasets, which may not fully

capture the current state of well water usage and contamination risks. For example, the use of historical data might not accurately reflect recent demographic and land use changes which can change the resulting patterns of where contaminant exposure risk is highest. In addition, we chose to demonstrate this approach using communities across Alabama and as a result the specific findings may not be directly transferable to other regions with different socio-economic and environmental contexts. That said, the methodology and approach is adaptable. Researchers and policymakers in other regions can gain similarly nuanced insights into the spatial dynamics of environmental risks and their intersection with socio-economic vulnerabilities across well user communities by employing similar methods.

Another limitation for this study is the inherent complexity and variability of natural processes and human activities that influence groundwater quality. While the study attempts to account for various factors, there are numerous other confounding variables that were not included. For example, emerging contaminants such as pharmaceuticals and personal care products, which are increasingly detected in groundwater, were not considered (Khan et al., 2022; Stuart et al., 2012; Stuart & Lapworth, 2013). Similarly, the study does not account for geogenic risks like arsenic or uranium contamination or the impact of climate change, which could significantly alter precipitation patterns, flooding risks, and consequently, groundwater contamination dynamics (Amini et al., 2008; Coyte et al., 2018; Lemonte et al., 2017). Some studies suggest that characteristics like well depth, which relate to well contamination, have not been considered (Wheeler et al., 2015). Furthermore, resampling the raster data to align with other datasets may be suitable for small-sized CBGs (Lechner et al., 2009). However, the unavailability of all the data at the same spatial resolution is another limitation of the study.

Regarding future opportunities, there is a clear need for more dynamic and real-time data collection methods to accurately monitor well water use and contamination levels. Advances in remote sensing technologies could provide more precise and up-to-date information on flood risk. Additionally, expanding the scope of the study to include a broader range of contaminants and risk factors, especially in the context of climate change, would enhance the understanding of well water contamination risks. Finally, there is an opportunity for more participatory research approaches that involve local communities in monitoring and managing their well water resources. Engaging well users in data collection, risk assessment, and decision-making processes can lead to more

sustainable and community-centric solutions that help ensure a sustainable water future for users of private wells.

Chapter 3

A machine learning approach to predict flooding in Mobile and Baldwin Counties, Alabama

3.1. Introduction

Flooding remains one of the most pervasive natural disasters across the globe, affecting millions of lives and causing extensive damage to property, infrastructure, and ecosystems (Aldardasawi & Eren, 2021; W. Du et al., 2010; Tingsanchali, 2012; Watson et al., 2016). Flooding can significantly impact private and public well water quality primarily through contamination by pathogens and pollutants when the well head becomes overtopped by flood water or through natural infiltration. Prior research has shown that, in the aftermath of Hurricane Harvey, a quantitative microbial risk assessment for private wells in flood-impacted areas of Texas revealed increased contamination from fecal indicator bacteria. This finding underscores the need for improved testing and management practices for well water, as well as greater public awareness about the impact of floodwater on wells. (Gitter et al., 2023). A cross-sectional study in the Republic of Ireland revealed that private well users are largely unprepared to cope with flood-triggered contamination risks (Musacchio et al., 2021). It is therefore important to accurately identify flood-prone areas for educating and informing vulnerable populations about the potential for flood-induced well water contamination. This proactive approach helps communities take necessary precautions to ensure safe drinking water. Moreover, the frequency and severity of flooding events is expected to increase as a result of climate change which will increase the vulnerability of groundwater sources to contamination (Andrade et al., 2018; Musacchio et al., 2021).

There are several established datasets that represent flood boundaries, flooding extent, and inundation. The FEMA 100-year flood hazard boundary represents areas with a 1% annual chance of flooding, focusing on regulatory standards and floodplain management (Drewry et al., 2024). In contrast, Height Above Nearest Drainage (HAND) models use elevation data to predict inundation extents based on terrain and proximity to drainage channels, which provide a more dynamic flood prediction tool tailored to the local topography (Jafarzadegan et al., 2018). The Dartmouth Flood Observatory utilizes satellite-derived extents to monitor real-time flood events globally, offering crucial data for regions lacking ground-based hydrological data (Awadallah &

Tabet, 2015). These datasets are designed to predict and map flood extents based on river stream behaviors, making them well-suited for understanding and managing flood risks in areas close to rivers (Jafarzadegan et al., 2018). However, they may not fully capture the extent of inland flooding (van Leeuwen et al., 2017), particularly in areas where flooding is caused by intense localized rainfall that does not immediately flow into river systems.

The dynamic and often rapid onset of inland flood events, compounded by climate change underscores the urgent need for effective monitoring of flooded areas and management strategies for flood-induced challenges. In this context, remote sensing technologies have emerged as invaluable tools. Moderate spatial resolution (10-30 m) earth observation satellites like Landsat and Sentinel provide critical regional data for flood monitoring, management, and damage assessment, enabling timely and informed decision-making. Studies highlight the advancement in remote sensing methods, including the integration of multispectral, radar, and optical data for enhanced flood mapping and monitoring, underscoring their importance in disaster management and mitigation efforts (Domeneghetti et al., 2019; Lo et al., 2015; Schumann, 2015).

The integration of remote sensing techniques for flood monitoring has improved with the advent of threshold-based models and the comparison of pre- and post-flood imagery, specifically from satellite imagery gathered by the Sentinel mission (Liang & Liu, 2020). Sentinel-1 stands out from other freely available satellite data sources, such as Landsat and Sentinel-2, due to its unique capabilities. As a Synthetic Aperture Radar (SAR) system, Sentinel-1 can acquire data regardless of weather conditions and daylight, making it particularly invaluable for emergency response and disaster management. This all-weather, day-and-night imaging capability ensures continuous monitoring, which is critical during flood events when optical sensors may be obstructed by clouds or darkness (DeVries et al., 2020a). In comparison, Landsat and Sentinel-2 rely on optical sensors, which can be limited by cloud cover and the need for daylight. This limitation can hinder their effectiveness in capturing real-time data during floods, especially under adverse weather conditions (Solovey, 2020). Furthermore, Sentinel-2, while offering high spatial resolution (10-20 meters) and frequent revisits (every 5 days), still faces challenges in continuous monitoring due to cloud cover (Bontemps et al., 2015). The revisit frequency of Sentinel-1 is significantly enhanced due to the constellation of two satellites, Sentinel-1A and Sentinel-1B, which together offer a revisit interval of 6 days globally. This high revisit frequency ensures more consistent and timely

data acquisition for flood monitoring (J. Li & Roy, 2017). In contrast, Landsat-8, despite its long historical record and valuable data continuity, has a revisit interval of 16 days when considered alone, though it can be combined with Landsat-7 to achieve an 8-day interval (Chastain et al., 2019).

Unmanned Aerial Vehicles (UAVs) offer valuable capabilities for flood monitoring, including high-resolution data acquisition and rapid deployment. However, they come with notable limitations. The cost of operating UAVs can be significant, making widespread deployment expensive (Song et al., 2022). Furthermore, flying UAVs over disaster-impacted areas poses challenges, including navigation in adverse weather conditions and potential damage to the drones themselves (Guo et al., 2021). UAVs also struggle to cover larger flood-affected regions effectively due to their limited battery life and range (Casaseca-de-la-Higuera et al., 2018). In contrast, the use of Sentinel-SAR offers distinct advantages over UAVs, particularly in its ability to provide extensive coverage and operate under all weather conditions, including night-time and cloudy scenarios, which are often encountered during floods (Goudarzi et al., 2021). Therefore, while UAVs are beneficial for localized and detailed monitoring, Sentinel-SAR is more effective for comprehensive flood assessments.

Threshold-based models, especially those relying on local thresholding approaches using Sentinel-1 SAR imagery, have proven effective in delineating water extents by considering the complexity and variability of different land surface types within an image, thus offering a more accurate and detailed analysis of flood events (Liang & Liu, 2020). A threshold-based model in remote sensing refers to a technique used to separate different types of land cover, such as water and land, based on specific pixel intensity or backscatter value thresholds in satellite imagery (Liang & Liu, 2020). Furthermore, methodologies incorporating multi-temporal SAR statistics with historical surface water class probabilities have been utilized to distinguish unexpected floods from permanent or seasonally occurring surface water, offering a new method for near-real-time flood monitoring (Sharifi, 2020).

Despite these advantages, the reliance on threshold-based models for flood detection using Sentinel-1 data presents limitations. For example, threshold-based methods require the selection of specific backscatter values (i.e. pixel values) to distinguish between water and non-water surfaces, which can be challenging due to the variability of land surface conditions and the

presence of vegetation or urban structures that also influence backscatter. Such models often struggle with the heterogeneous nature of the terrain, where the same threshold may not apply uniformly across different regions or under different ambient conditions. This can lead to over- or under-detection of flooded areas, especially in complex environments that features of the built environment and dense vegetation. For these locations the backscatter response might be similar for water and non-water surfaces due to interference from the surrounding environment (Liang & Liu, 2020). The development of more sophisticated, automated approaches that can adapt to these challenges is essential for improving the accuracy and reliability of flood maps derived from Sentinel-1 and other SAR data.

Beside the limitations of threshold-based model, Sentinel-1 SAR lacks the ability to generate a wide range of exploratory variables. While Sentinel-1 is excellent for capturing flood extents due to its all-weather capabilities, it does not provide the same level of detail and variety of data as Sentinel-2, making it less versatile for comprehensive flood mapping and analysis (Solovey, 2020). Therefore, the integration of Sentinel-2's multispectral imagery with Sentinel-1's SAR data is often necessary to achieve the best results in flood monitoring and mapping (Tuo et al., 2022). Sentinel-2 has the ability to generate various exploratory indices to predict flooding, such as NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), NDWI (Normalized Difference Water Index), GCI (Green Chlorophyll Index) and so on. These indices are valuable for machine learning models as they provide comprehensive data on vegetation health, water bodies, and chlorophyll concentration, which are essential for accurate flood detection and monitoring (Y. Du et al., 2016).

The use of ML models, such as Random Forest (RF) regression, for spatial prediction of flooding represents a significant advancement over traditional threshold-based models. For example, a study by Mosavi et al. (2018) emphasized the significant contributions of ML methods, including Random Forest, in enhancing the prediction systems for flood risks. The study showcased the effectiveness and efficiency of ML models in capturing the complex dynamics of flood processes, thus providing a more accurate and cost-effective solution compared to traditional methods like using thresholds (Mosavi et al., 2018). Similarly, Tayfur et al. (2018) discussed the application of various ML methods, including RF, for predicting flood hydrographs. The study highlighted the power of ML models in offering high accuracy in flood prediction using less and

easily measurable data, overcoming significant parameter estimation problems associated with conventional models (Tayfur et al., 2018). Additionally, Sampurno et al. (2022) used an integrated hydrodynamic and ML approach to predict compound flooding in an estuarine delta in Indonesia, identifying RF as the most accurate algorithm for flood hazard prediction (Sampurno et al., 2022). Moreover, Rajab et al. (2023) leveraged historic climatic records to apply ML models, including RF, for flood forecasting in Bangladesh, highlighting its superior performance in predicting rainfall and flood risks compared to other models (Rajab et al., 2023). These studies collectively underscore the advanced capabilities of ML models, particularly RF, in providing accurate and efficient flood prediction, thus offering substantial improvements over traditional threshold-based methods.

This study aims to expand on these advances to develop a ML-based model to predict flooding in Mobile and Baldwin Counties in Alabama; two counties where rate of households using private well is moderately high on average (A. H. Murray & Kremer., 2023). The primary objective of this research is to enhance the accuracy of flood detection in suburban and rural areas by integrating the strengths of threshold-based models with ML-based models. This approach aims to leverage the robustness of threshold-based techniques in initial water delineation while employing ML models to refine and improve these predictions. By combining the simplicity and efficiency of threshold methods with the predictive power and adaptability of ML algorithms, this integrated model seeks to provide more accurate and detailed mapping of flooded area estimates, particularly in environments with complex land surface characteristics.

3.2. Methods

3.2.1 Conceptual diagram

Figure 4 illustrates the concept diagram of datasets used, analysis, and respective outcomes. Details of each diagram component are discussed from section 3.2.2.

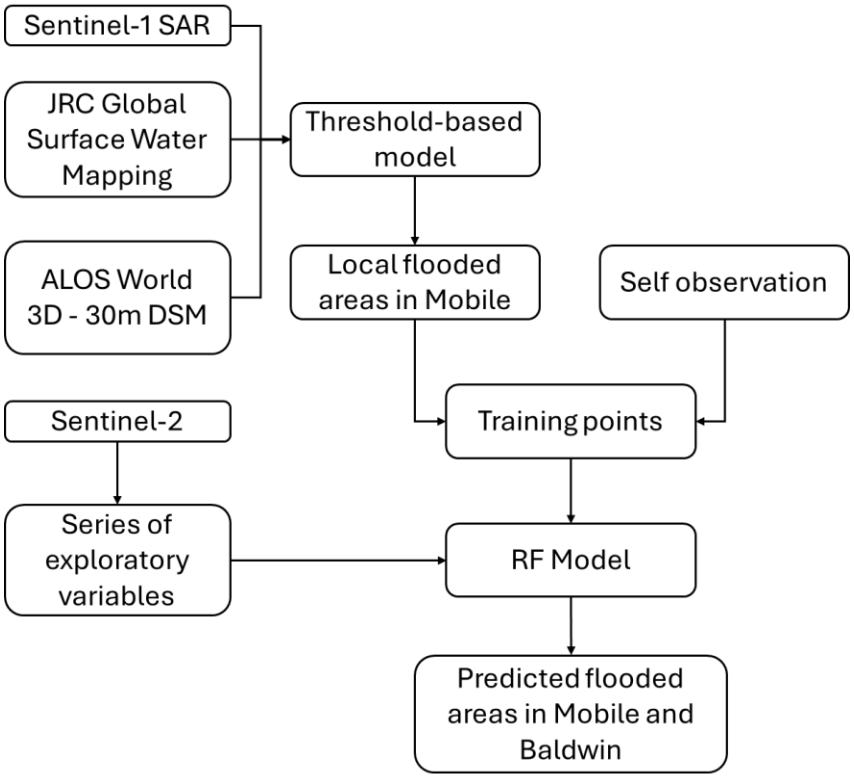


Figure 4 Concept diagram.

3.2.2 Data

This study uses satellite imagery datasets collected through Google Earth Engine (GEE), specifically the Sentinel-1 SAR imagery, which were used to develop the primary threshold-based flood model (Google Developers, 2022). The Sentinel-1 mission provides data from a C-band SAR operating at 5.405GHz and capable of dual-polarization. These data are encapsulated in the Sentinel-1 Ground Range Detected (GRD) scenes collection on GEE, which undergoes daily updates. A "scene" in the context of Sentinel-1 SAR imagery refers to a single, discrete image captured by the Sentinel-1 satellite's SAR over a specific geographic area at a particular moment in time. The Sentinel-1 image collection is filtered to include only ascending pass (covering Alabama) images with VH polarization and a resolution of 10 meters. Shamshiri et al. (2018) analyzed the use of VH channel in Sentinel-1 data, highlighting its role in improving spatial density (Shamshiri et al., 2018). The images are then clipped to the study area and median composites are created for the before and after periods. These composites are processed through the Refined Lee filter, and the resulting images are converted back to dB units. Refined Lee filter is known for its effectiveness in reducing noise while preserving image detail. Sun and Li (2020) improved

denoising methods for Sentinel-1 data, addressing both additive and multiplicative noise, emphasizing the need for advanced filtering techniques like the Refined Lee filter for high-quality SAR image analysis (Sun & Li, 2021). This filter uses a combination of 3x3 and 7x7 kernels to compute local statistics and gradients, ultimately refining the image quality.

The high-resolution multispectral imagery provided by the Sentinel-2 mission was used to support the analysis of terrestrial phenomena such as vegetation dynamics, soil and water cover, and the assessment of inland and coastal waters. This dataset encompasses 13 spectral bands in UINT16 format, capturing top-of-atmosphere (TOA) reflectance values. These TOA values are scaled by a factor of 10,000 following procedure outlined in the Sentinel-2 User Handbook (Google Developers, 2022). Furthermore, the Sentinel-2 dataset includes three quality assurance (QA) bands that are crucial for ensuring the accuracy and usability of the imagery. Among these, the QA60 band is particularly significant as it functions as a bitmask for cloud detection. This capability is essential for delineating pixels covered by clouds from those that are not, ensuring that analyses such as vegetation health assessments and soil moisture calculations are based on clear and accurate observations (ESA, 2022).

Finally, the JRC Global Surface Water Mapping Layers available in GEE were used for identifying permanent water bodies. This dataset includes maps of surface water locations and their temporal distribution from 1984 to 2021, along with statistical analyses of water surface extents and their changes over time. Researchers derived this compilation from 4,716,475 scenes from Landsat 5, 7, and 8, captured between March 16, 1984, and December 31, 2021. An expert system classified each pixel in these scenes as water or non-water, leading to a comprehensive monthly historical record (Pekel et al., 2016). This record enables change detection analysis for two periods: 1984-1999 and 2000-2021 (Pekel et al., 2016). The Advanced Land Observing Satellite (ALOS) World 3D - 30m (AW3D30) provided elevation data for the original data set derivation. AW3D30dataset offers a horizontal resolution of approximately 30 meters (1 arcsec mesh), based on the 5-meter mesh version of the World 3D Topographic Data (Tadono et al., 2014; Takaku et al., 2014). Table 5 shows the details of the four data sources.

Table 5 Details of data sources, sensor, spatial, spectral and temporal resolution

Data Source	**Sensor**	**Spatial Resolution**	**Spectral Resolution**	**Temporal Resolution**
Sentinel-1 SAR	C-band SAR	10, 25, 40 meters	Dual-polarization (VV, HH, VV+VH, HH+HV)	6 days
Sentinel-2	Multispectral Imager	10, 20, 60 meters	13 spectral bands	5 days at equator, and 2-3 days at mid-latitudes
JRC Global Surface Water Mapping	Landsat 5, 7, 8	30 meters	N/A (monthly surface water extent from 0-12 where 0 means there is no water over the years and 12 means water stays there for 12 months in a year.)	Monthly historical record from 1984 to 2021
ALOS World 3D - 30m (AW3D30)	Digital Surface Model (DSM)	Approx. 30 meters	N/A (elevation data)	N/A (based on acquisition period)

3.2.3 Study area

Mobile and Baldwin Counties in Alabama have been selected as the study area due to their highest average precipitation over the past 30 years, along with a noticeable rate of private well use. According to the Parameter-elevation Regressions on Independent Slopes Model (PRISM) Climate Group data, Baldwin County has the highest average precipitation in Alabama from 1991 to 2020, followed by Mobile County with the second highest (Figure 5a). Recently published block group level private well data from the EPA reveals that around 70% of households on average use private wells in Baldwin County, along with 57% of households in Mobile County. Figure 5b shows the geographic map of Mobile and Baldwin Counties along with the private well use rate.

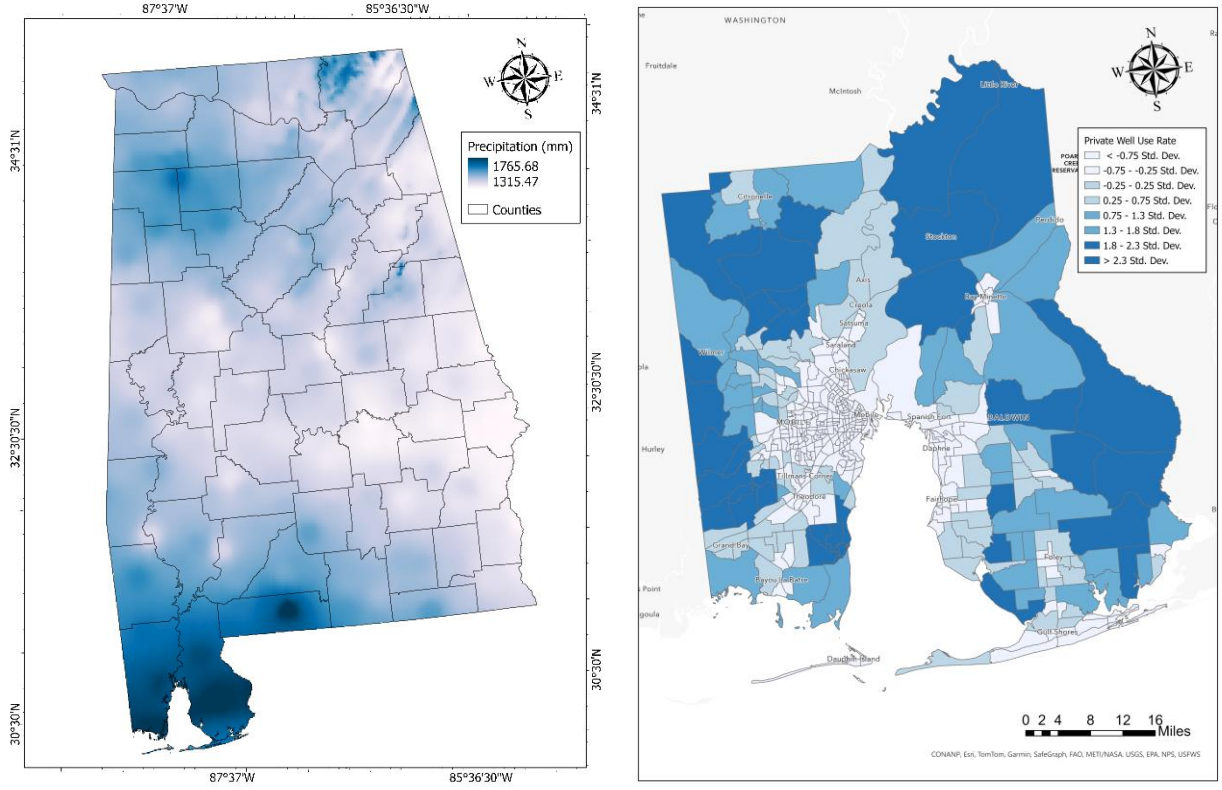


Figure 5 Study area map. (a) Average 30 years precipitation of Alabama where Mobile and Baldwin has the highest precipitation; (b) Block group level private well use rate in households of Mobile and Baldwin categorize with standard deviation.

The development of the flood model requires satellite imagery from a flooded time period and a non-flooded time period to compare data from wet vs. dry conditions. Although Sentinel-1 is capable of cloud penetration, Sentinel-2 operates as a conventional multispectral band satellite. To secure Sentinel-2 imagery with minimal cloud cover, the study was conducted during a period

of low flooding in the region of interest. The selected study period encompasses November and December 2019, coinciding with an unexpected flash flood event that stretched from Alabama to Tennessee, as reported by FloodList (2019). Record-breaking rainfall hit parts of Tennessee and Alabama starting on December 29, leading to severe flash floods that resulted in the deaths of two individuals (FloodList, 2019). In Tennessee, floodwaters prompted the closure of multiple roads within Williamson County, where the local Emergency Management Agency conducted two water rescues. Similar conditions in Alabama claimed a life when a vehicle was overtaken by the floodwaters in the city of Anderson, situated in Lauderdale County, AL, on December 29.

3.2.4 Threshold-based flood mapping

Pre-flooding and post-flooding imagery along with basic flood identification factors like permanent water bodies and a DEM has been used in the threshold-based model to delineate areas inundated by flood waters. The pre-flooding imagery was obtained from the median imagery of the whole month of November 2019. This imagery is considered as normal "blue sky" conditions for the area. The post-flooding imagery was sourced from median imagery of the whole month of December 2019. The application of averaging median filters in remote sensing has been shown to preserve fine details while attenuating impulse noise, indicating the utility of median-based approaches in maintaining the integrity of spatial information (Vassiliou et al., 1988). At its core, the threshold approach assesses the differences in SAR backscatter values between the pre- and post-flood imagery. SAR that interacts with water scatters the wavelengths differently than SAR interacting with non-water surfaces. Therefore, by comparing these two values one can identify locations where the SAR backscatter values have changed. The threshold value is used to select when that change is large enough to suggest that a previously dry location is now wet, indicating a flood. To assess that change a threshold value that indicates the difference between pre-flooding and post-flooding imagery (eq .8) has been specified while masking out high elevation areas and permanent water bodies (Table 5). Equation 8 shows the flood extraction calculation with a specific threshold (θ).

$$\text{Flooded Area} = \left(\frac{\text{After Image}}{\text{Before Image}} > \theta\right) \cap (\text{PermanentWater} < 5) \cap (\text{Slope} < \sigma) \tag{8}$$

In the equation for extracting flooded areas, a comparison of satellite images before and after a flooding event is used to identify where the changes in SAR backscatter occurred between the two time periods. For this analysis, a backscatter ratio value (after image / before image) greater than 1.25 (θ) would suggest the presence of water in the after-event image at those pixel locations, thus signaling potential flooding. This threshold value is chosen because water reflects radar signals more than dry land, leading to a significant difference in the backscatter values. It is important to mention that the threshold value can vary over space and land cover. Liang and Liu (2020) proposed a local thresholding approach to delineate water extent using Sentinel-1 SAR imagery which highlights the importance of adjusting thresholds based on local conditions (Liang & Liu, 2020). Manjusree et al. (2012) undertook a study to optimize threshold ranges for classifying flood water in SAR images across different polarizations. The study provides detailed insights into how different threshold values, including 1.5, 1.6, and 2.5 ratio, can be used for effective flood detection (Manjusree et al., 2012). To determine the most effective threshold value for identifying changes in overland water presence, soil moisture images were compared before and after the flood event while changing the threshold value. The goal was to balance regional generalizability with sensitivity to changes in water levels. A threshold value of 1.25 was selected as the optimal point. This high threshold ensures that the areas identified reflect the highest increase in soil moisture, effectively pinpointing the regions most likely to have experienced flooding. This approach, while primarily used for training purposes, allows for a focused identification of areas with significant changes in soil moisture, suggesting a strong potential for actual flooding.

To ensure that permanent water bodies are not mistakenly classified as flooding the equation also excludes areas or permanent or semi-permanent water indicated by a seasonality score of 5 or more from the Global Surface Water dataset (Table 5). It also incorporates slope measurement to exclude areas with steep slopes. Specifically, pixels with a slope greater than 5 percent are excluded. In the regions of Mobile and Baldwin, the average elevation is approximately 39 meters. 5 percent of this elevation amounts to around 2 meters, equivalent to more than 6 feet. Steeper slopes are less likely to retain floodwaters, so only regions with a slope less than 5 percent (σ) are considered. This method isolates flood-affected regions by leveraging the unique properties of radar satellite imagery to detect surface water changes, even under challenging conditions like cloud cover. Figure 6 depicts the results of threshold-based flood mapping in Mobile and Baldwin

County in December 2019 along with soil moisture scenario from Sentinel-1. The soil moisture scenario is a false color composite where the VV polarization which reflects wetlands has been run though the red channel while VH polarization has been run through blue and green channel. Typically, the VV polarization is sensitive to surface roughness and moisture, while VH polarization, due to its sensitivity to the structure and orientation of objects on the ground (Balenzano et al., 2011; Joseph et al., 2008). As a result, soil with higher moisture content will appear redder in the image.

Although the threshold-based model can isolate flooded pixels by comparing pre-flood and post-flood images, its accuracy faces certain limitations. As previously mentioned, selecting a specific threshold value—for SAR images, permanent water, and slope—requires careful consideration of the geography of the study area. A threshold that works in one region may not be suitable for another. For example, figure 6a demonstrates the flooded areas extracted using the threshold value of 1.25. While the model performed reasonably well for southern Mobile, it only identified a few flooded areas in Baldwin, which does not reflect the actual situation, as the pre-flood and post-flood imagery show significant changes in the soil moisture content after the flooding event. A comparison of pre-flood and post-flood images in southern Baldwin (Figures 4e and 4f, respectively) highlights flooded pixels in red that the threshold-based model failed to extract.

Moreover, the simplicity of the threshold-based model, which relies on relatively simple calculations between pixels from two images, often fails to capture the full extent of flooded areas due to variations in pixel reflectance. This issue is evident in Figure 6d, where the model identified some flooded areas, but the reddish pixels, indicating high soil moisture, cover a much larger area in the post-flood image (Figure 6f) than the model extracted. While it cannot be said for certain whether these are flooded areas, it is confirmed that the soil moisture content is especially high compared to the surrounding area which may indicate a higher affinity for flooding that was not captured by using the threshold approach for this period. This limitation suggests the need for adopting a prediction-based model capable of accurately determining flooded areas.

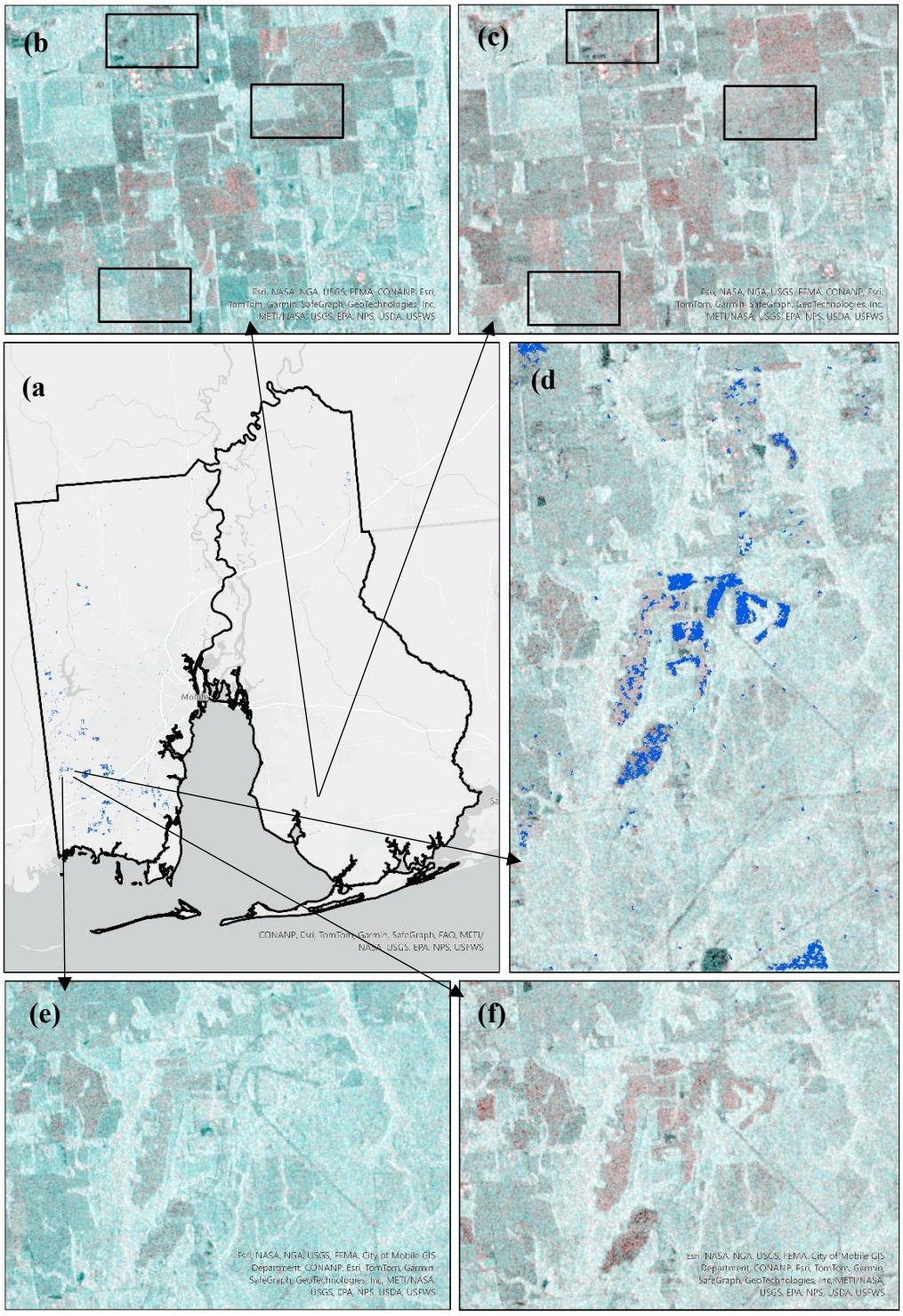


Figure 6 Threshold-based modeling result of Mobile and Baldwin to extract flood pixels; (a) Extracted flooded area over the whole study region; (b) pre-flood soil moisture scenario of a specific area in southern Baldwin where red indicates high soil moisture; (c) post-flood soil moisture scenario of the same area in Baldwin where flood pixels can be seen in red; (d) extracted flooded area of a specific region in Mobile; (e) pre-flood soil moisture scenario of Figure 6(d); (f) post-flood soil moisture scenario of Figure 6(d). Black boxes in (b) and (c) is to help identify flooding in micro scale.

3.2.5 Random Forest classification and regression

To predict flooded areas, a random forest ML procedure was used. The RF classification and regression uses data of known values to generate a training dataset. Relationships between explanatory variables of interest and a target value are generated and used to predict target values at unknown locations. The specific version of the tool implemented here creates models and generates predictions using an adaptation of the random forest algorithm, which is a supervised ML method developed by Breiman et al. (2001b; 2017). During the modeling process, the tool creates many decision trees, called an ensemble or a forest. Each tree generates its own prediction and is used as part of a voting scheme to make final predictions. The final predictions are not based on any single tree but rather on the entire forest. The use of the entire forest rather than an individual tree helps avoid overfitting the model to the training dataset, as does the use of both a random subset of the training data and a random subset of explanatory variables in each tree that constitutes the forest. In the context of flooding, RF model has been used to predict the areas of flooding that the threshold-based model fails to identify in both Mobile and Baldwin County. The model is trained using points from flooded areas (i.e., those pixels where the threshold model did a good job of delineating flooded pixels) collected manually. A suite of derived remote sensing indices associated with flood potential are used as the explanatory variables.

3.2.6 Training and testing the model

RF model associated with the Forest-based Classification and Regression tool from ArcGIS Pro has been utilized for this study. The outputs from the threshold-based model (detailed previously) along with pre- and post-flooding soil moisture scenario are used to identify known locations of flooding (target variable) in Mobile. It has been assumed that flooded pixels classified as flooded from the threshold model and areas of high soil moisture were most probably flooded during the study period. In the post-flood images, areas that have flooded show higher moisture reflectance than in pre-flood images, as observed in Figures 4b, 4c, 4e, and 4f. The RF model also requires data on the location of non-flooded regions which necessarily include permanent water bodies, high elevated lands, and wetland areas. 1329 training points have been collected, of which 465 points were from areas assumed to have flooded, and 864 points indicate non-flooded regions.

Predicting regional-level flooding from local training points is one of the objectives of this study. In this case, a local threshold-based model has been developed in the southwestern part of Mobile, assuming that this area is the closest to the coast with low elevation. However, the local model can be based on any other location within the predicted extent. The flooded training points are not only based on the threshold-based model but also on self-observation. The threshold-based

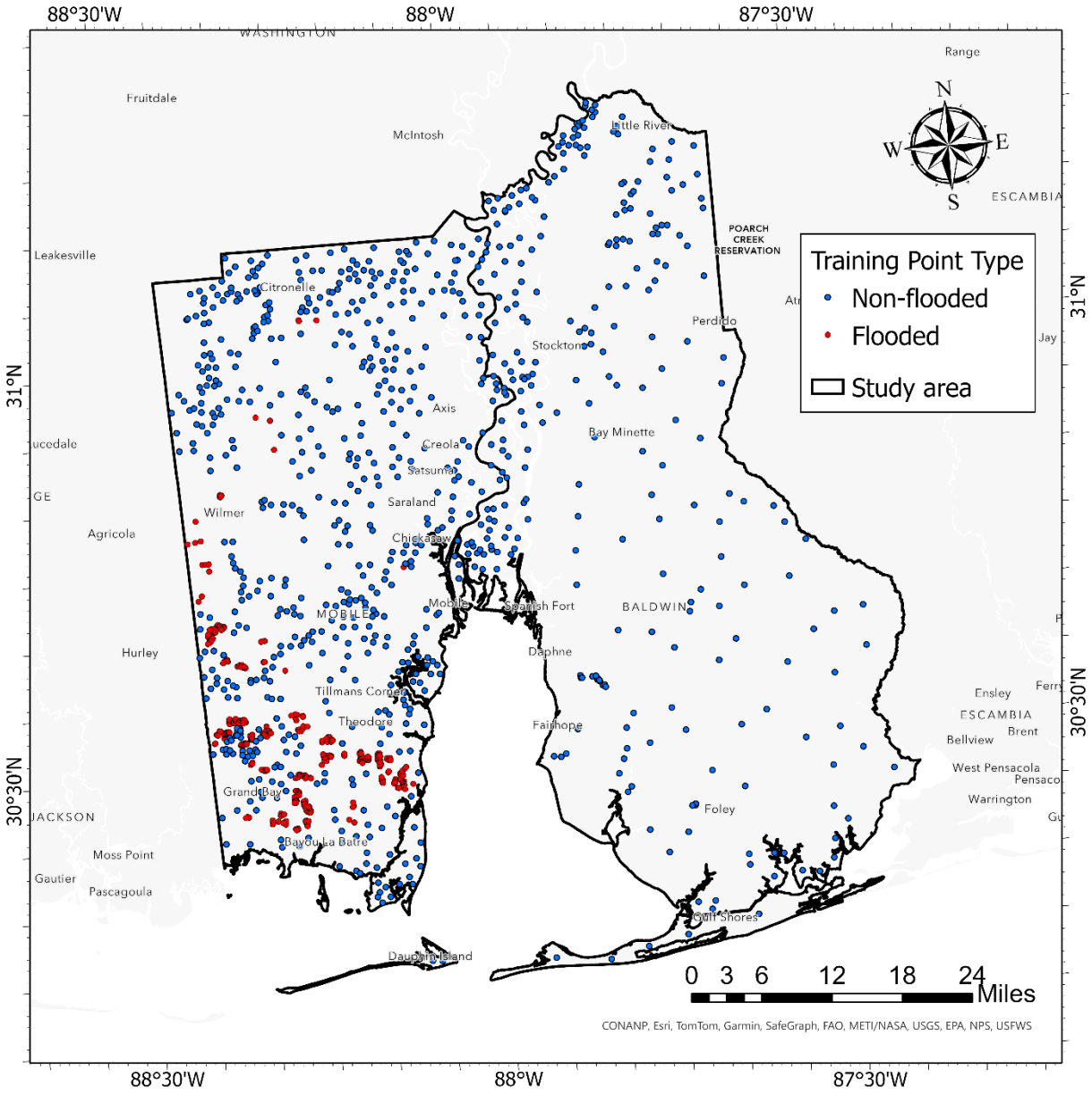


Figure 7 Training points over Mobile and Baldwin Counties, AL. Red points are the flooded training points and blue points are the non-flooded training points.

model only supports the process of identifying potential areas where floods have occurred and is merely an initial indicator of flood prone regions.

30% of the training data has been reserved, which is approximately 399 points from the total of 1329 training points, for validating the Random Forest model. It is a form of Out-of-Bag error (OOB) validation (Bylander, 2002; Cho et al., 2019). OOB errors serve as indicators of the model's accuracy. They are derived from the model's capability to predict the target variable accurately using the observations in the training dataset. The OOB error, which includes metrics such as the mean squared error (MSE) and the percentage of variation explained, is determined by the predictive performance of the model for the portion of the training data not used by certain trees within the forest. This method provides an internal validation mechanism, allowing for an assessment of the model's prediction accuracy using the training data itself, without the need for a separate validation set.

Figure 7displays the training points, with blue point representing the non-flooded training points and red point representing the flooded training points. Due to the strong performance of the threshold-based model in Mobile County (i.e., more flooded areas were easier to identify following the flood event), there were more training points collected from Mobile than Baldwin. The rationale behind this selection approach is twofold: First, it allows for the application of learned patterns from Mobile County to identify similar flood characteristics in Baldwin County, where direct observations or historical flood data is sparse or non-existent. Second, it aims to create a robust flood prediction tool that can be applied across different regions, enhancing our capacity to manage and respond to flood events more effectively.

3.2.7 Explanatory variables

Choosing the right explanatory variables is critical for the success of ML models in predicting floods. The variables directly influence the model's ability to learn from the data, recognize patterns, and make accurate predictions. Selecting relevant variables ensures that the model considers all significant factors that contribute to flooding, such as topography, land cover, soil moisture, and water bodies' presence (Mosavi et al., 2018). In addition, multiple derived variables were also included as explanatory variables.

The Normalized Difference Flood Index (NDFI), derived from Sentinel-1 SAR data, uses the difference and sum of vertical-vertical (VV) and vertical-horizontal (VH) polarization to identify water-covered surfaces, particularly useful in flood detection and monitoring. This index effectively distinguishes floodwaters from other land covers by exploiting the differential scattering properties of water surfaces under various polarizations, making it a vital variable for accurate flood mapping (Xue et al., 2022).

Similarly, the Normalized Difference Water Index (NDWI) from Sentinel-2, which calculates the contrast between the green and near-infrared (NIR) bands, and the Modified Normalized Difference Water Index (MNDWI), which substitutes the green for a short-wave infrared (SWIR) band, are instrumental in identifying and delineating water bodies. These indices exploit the spectral reflectance properties of water versus vegetation and built-up areas, enhancing the capability to monitor surface water extent and changes over time. Again, this index is crucial for flood risk assessment and management (Kashyap et al., 2022; Solovey, 2020). The Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) provide insights into vegetation health and biomass, which can be affected by flooding. Specifically, following a flood both the NDVI and EVI may decrease and serve as an additional proxy for where flooding has occurred. These indices are calculated from the red and NIR bands, reflecting the photosynthetic capacity and biomass of vegetation. Healthy vegetation reflects more NIR and less visible light, whereas flooded or stressed vegetation shows the opposite pattern. Understanding vegetation health and changes can help in assessing flood impacts on ecosystems and agricultural lands (Goffi et al., 2020).

The Normalized Difference Turbidity Index (NDTI) uses the SWIR bands to measure water turbidity, a parameter that often increases in water bodies during flood events due to the suspension of sediments. Monitoring turbidity is essential for assessing the quality of water during and after floods, impacting water treatment and public health (Solovey, 2020). The Soil Adjusted Vegetation Index (SAVI) modifies NDVI to minimize soil brightness effects, providing a more accurate representation of vegetation in areas with mixed land cover. This index is particularly relevant in flood-prone regions where vegetation cover is sparse or irregular, offering insights into how land cover affects flood dynamics (Goffi et al., 2020). Further, the Water Ratio Index (WRI), the Green Chlorophyll Index (GCI), and the Built-up Area Index (BAI) utilize various band combinations to

differentiate between water, vegetation, and urban areas. These indices contribute to a comprehensive understanding of land cover and land use, key factors in flood vulnerability and risk mapping (Andreo et al., 2019; Goffi et al., 2020).

Finally, the Soil Moisture Index (SMI), leveraging the contrast between NIR and SWIR bands, is indicative of soil moisture levels, a critical parameter in predicting flood potential and understanding the hydrological conditions leading up to flood events (Wanders et al., 2014). Land cover classification data, such as that from the National Land Cover Database (NLCD), and DEM, like NASA's Shuttle Radar Topography Mission (SRTM), provide essential contextual information on the natural and built environment. These data sources offer critical insights into terrain, elevation, slope, and land use patterns, all of which influence flood behavior, susceptibility, and impact (Sankaranarayanan et al., 2020).

The RF model incorporates the above 13 explanatory variables (Table 6). The majority of these variables are derived from different band combinations associated with Sentinel-1 and Sentinel-2 imagery. All the indices derived from Sentinel are generated at a 10-meter resolution, while land use and DEM data are provided at a 30-meter resolution. Table 5 presents a list of all explanatory variables, along with their equations, sources, and references.

Table 6 Explanatory variables, source, equations, and references.

Variable	Source	Equation	Reference
Normalized Difference Flood Index	Sentinel-1 SAR	$\frac{VV - VH}{VV + VH}$	(Xue et al., 2022)
Normalized Difference Water Index (NDWI)	Sentinel-2	$\frac{B3 - B8}{B3 + B8}$	(Kashyap et al., 2022)
Modified Normalized Difference Water Index (MNDWI)		$\frac{B3 - B11}{B3 + B11}$	(Solovey, 2020)
Normalized Difference Vegetation Index (NDVI)		$\frac{B8 - B4}{B8 + B4}$	(Goffi et al., 2020)
Normalized Difference Turbidity Index (NDTI)		$\frac{B11 - B12}{B11 + B12}$	(Solovey, 2020)
Soil Adjusted Vegetation Index (SAVI)		$\left(\frac{B8 - B4}{B8 + B4 + 0.5}\right) \times 1.5$	(Goffi et al., 2020)
Enhanced Vegetation Index (EVI)		$2.5 \times \frac{B8 - B4}{B8 + 6 \times B4 - 7.5 \times B2 + 1}$	(Goffi et al., 2020)
Water Ratio Index (WRI)		$\frac{B3 + B4}{B8 + B11}$	(Goffi et al., 2020)
Green Chlorophyll Index (GCI)		$\frac{B8}{B3} - 1$	(Andreo et al., 2019)

Built-up Area Index (BAI)		$\frac{1}{(0.1 - B4)^2 + (0.06 - B8)^2}$	(Goffi et al., 2020)
Soil Moisture Index (SMI)		$\frac{B11 - B8}{B11 + B8}$	(Wanders et al., 2014)
Land cover classification	NLCD 2019		(Sankaranarayanan et al., 2020)
DEM	NASA SRTM		(Sankaranarayanan et al., 2020)
$B2$, $B3$, $B4$, $B8$, $B11$, and $B12$ are Sentinel-2 bands corresponding to Blue, Green, Red, Near Infrared (NIR), and two Short Wave Infrared (SWIR) wavelengths, respectively. *VV* and *VH* are Sentinel-1 bands representing vertical-vertical and vertical-horizontal polarization, respectively.			

3.2.8 Correlation and Local Moran's I

To understand the extent of the predicted flooded area affecting private well users, Pearson correlation and Local Moran's I analysis have been conducted. Local Moran's I is a statistical measure used to identify local clusters or spatial autocorrelation within a given dataset. It helps in detecting local patterns by breaking down the global Moran's I into individual contributions from each location, thereby highlighting areas of significant spatial clustering or dispersion. Block group-level well use rates and total flooded areas per block group have been analyzed to identify local clusters of these two variables separately, which include five clusters: high-high (HH), high-low (HL), low-high (LH), low-low (LL), and not significant. Finally, a confusion matrix shows how many HH, LL, and not significant block groups are common for both variables in the study area.

3.3. Results

3.3.1 Predicted flooded areas

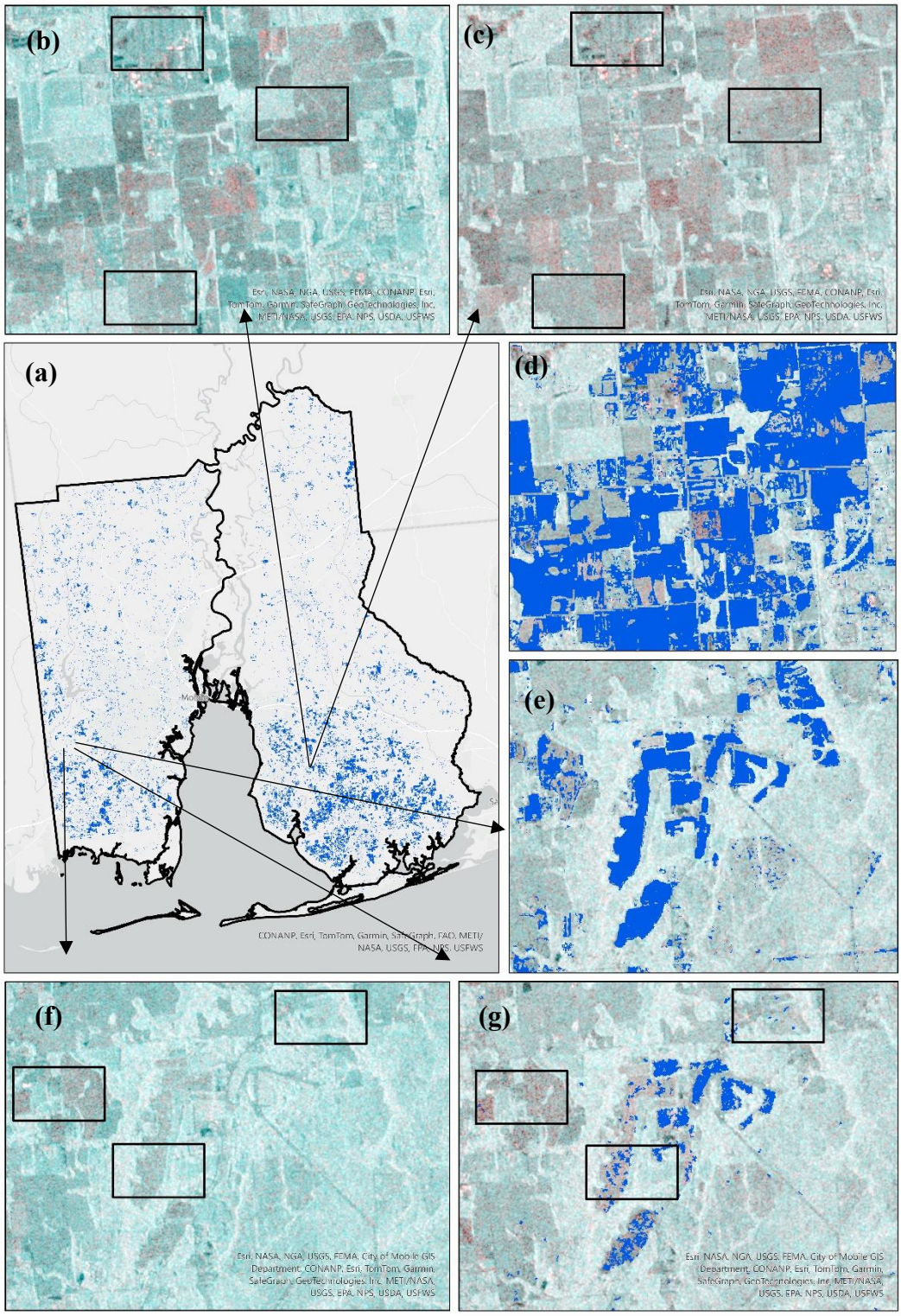


Figure 8 Predicted flooding from RF model; (a) predicted flooded area over the whole study area; (b) pre-flood soil moisture scenario in a specific region of Baldwin; (c) post-flood soil moisture along with threshold-based model flooded area in the same region in Baldwin; (d) RF model predicted flooded area in the same region in Baldwin; (e) RF model predicted flooded area in the same region in Mobile; (f) pre-flood soil moisture scenario in a specific region of Mobile; (g) post-flood soil moisture scenario along with threshold-based model flooded area in the same region in Mobile;.

Figure 8 displays the outcomes of the RF model, highlighting the predicted flooded areas. A comparison between Figure 8a and Figure 6a demonstrates that the RF model has indeed predicted that flooding is likely to occur in Baldwin County, a result that was clearly not captured by the threshold-based model. A closer examination of the southern region in Mobile, as shown in Figure 8e, reveals that the RF model identified additional flooded regions not detected by the threshold-based model. A detailed observation of the smaller boxes provides a clearer view of the changes in soil moisture before and after the flood. Despite observing an increase in soil moisture in the pre-flood scenario (Figure 8b), the threshold-model failed to identify any flooded areas in Baldwin County (Figure 8c). In contrast, the RF model predicted many flooded areas in this region, as evident in Figure 8d.

Tables 7, 8, and 9 assess the performance of the RF model, offering evidence of the importance of explanatory variables. Utilizing 100 trees and employing 30% of the training points for validation, the model's OOB errors indicate a relatively low MSE, which shows a decrease compared to the model with 50 trees. This trend is consistent for both non-flooded and flooded regions, with the latter showing a particularly low MSE. The decrease in MSE when the number of trees in the RF model is increased from 50 to 100 is generally a positive indicator of model performance (Breiman, 2001a; Probst & Boulesteix, 2018). Lower MSE values suggest that the model's predictions are closer to the actual values, indicating improved accuracy. This indicates that the model becomes more adept at identifying flooded areas correctly as the complexity of the forest increases, which is crucial for applications in flood prediction and assessment.

Table 7 Model out of bag errors

Number of Trees	MSE	Non-flooded	Flooded
50	0.663	0.935	0.157
100	0.52	0.757	0.08

Table 8 ranks each explanatory variable's importance for predicting flood vs non-flood areas. The values in the Importance column represent the cumulative Gini coefficients for each variable, aggregated from all the trees in the model. The percentages indicate the proportion of the overall Gini coefficient sum contributed by each variable. The Gini coefficient measures the importance of each variable in making accurate predictions. It is derived from decision trees within the random forest algorithm. For each variable, the Gini coefficient represents the sum of its contributions to splitting nodes across all trees in the model. A higher Gini coefficient indicates that the variable is more influential in improving the model's predictive accuracy. The percentage values reflect the proportion of the total Gini coefficients accounted for by each variable, highlighting their relative importance. For this case, GCI, NLCD, and SMI, each with a 9% contribution, are among the most significant predictors in the model.

Table 8 Variable importance (Gini coefficients) and percentages related to the proportion of the Gini Coefficient sum accounted for by each explanatory variable. Variable acronyms are defined in Table 6.

Variable	Importance	%
GCI	0.3	9
NLCD	0.3	9
SMI	0.3	9
SAVI	0.3	9
NDVI	0.29	9
WRI	0.29	8
NDTI	0.28	8
Elevation	0.28	8
NDWI	0.27	8
BAI	0.26	8

NDFI	0.21	6
EVI	0.17	5
MNDWI	0.14	4

Table 9 presents classification diagnostics for a flood prediction model, evaluated separately on training and validation datasets. The metrics reported are F1-Score, Matthews Correlation Coefficient (MCC), Sensitivity, and Accuracy, for both the non-flooded and flooded categories. The F1-Score is the harmonic mean of precision and recall. It is particularly useful when dealing with imbalanced datasets, where one class is significantly underrepresented. The F1-Score reaches its best value at 1 (perfect precision and recall) and worst at 0 (Takahashi et al., 2022). MCC is a correlation coefficient between the observed and predicted binary classifications. It returns a value between -1 and +1 where +1 indicates a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation. MCC is considered a balanced measure which can be used even if the classes are of very different sizes (Chicco, Warrens, et al., 2021). Sensitivity, also known as the true positive rate or recall, measures the proportion of actual positives that are correctly identified as such (Chicco, Starovoitov, et al., 2021). Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined (Zhu, 2020).

Table 9 Model diagnostics

Category	F1-Score	MCC	Sensitivity	Accuracy
Training data				
Non-flooded	1	0.99	0.99	0.99
Flooded	0.99	0.99	1	0.99
Validation data				
Non-flooded	0.97	0.92	0.97	0.96
Flooded	0.95	0.92	0.96	0.96

For the training data, the model demonstrates near-perfect performance across all metrics for both categories. Specifically, the non-flooded category achieves a perfect F1-Score of 1 and very high scores in MCC (0.99), Sensitivity (0.99), and Accuracy (0.99). The flooded category

also shows high performance with an F1-Score of 0.99, indicating a balanced precision and recall, and identical scores for MCC, Sensitivity, and Accuracy as the non-flooded category, highlighting the model's strong ability to correctly identify both categories during training. Even though the scores seem unusually high, this is expected since this validation is based on training points that are already known by the forest (ESRI, 2022).

On the validation data, the model exhibits slightly lower, yet strong, performance metrics, which suggests good generalization to unseen data. The non-flooded category scores are slightly reduced to 0.97 for both F1-Score and Sensitivity, and 0.92 for MCC, with accuracy slightly dropping to 0.96. The flooded category shows a similar trend with a reduced F1-Score of 0.95 and identical MCC and Sensitivity scores as the non-flooded category, with the same Accuracy. These results indicate the model's robustness and its effectiveness in predicting flood events, with a high level of reliability and consistency between training and validation, ensuring its utility in practical flood forecasting applications.

3.3.2 Flooded area and private well use rate

The correlation between block group-level total flooded area and well use rate was found to be approximately 0.587, indicating a moderate positive relationship. Figure 9 shows the confusion matrix of local Moran's I cluster type between total flooded area and household-level well use rate. The confusion matrix was applied to assess the change in geographic patterns between clusters of high and low well use and flooding risk identified by the model. Entries on only the diagonal in a confusion matrix indicate that cluster membership HH or LL resulting from well use and flood risk align in geographic space. Off-diagonal entries indicate the degree the well use and flood risk result in a different geographic configuration of clusters across the study area. Numerical values within the confusion matrix correspond to the total number of block groups that correspond to the pairwise cluster group. Thirty-two block groups out of 450 have HH clusters for both variables, while 236 block groups have LL clusters for both. Other cluster includes 8 HL cluster for flooded area and 30 HL cluster for well use. Figure 10 illustrates the map of the cluster results, revealing that the southeastern rural areas have the highest well use rates along with the largest flooded areas, whereas the city areas of Mobile have the lowest private well use rates along with the smallest flooded areas.

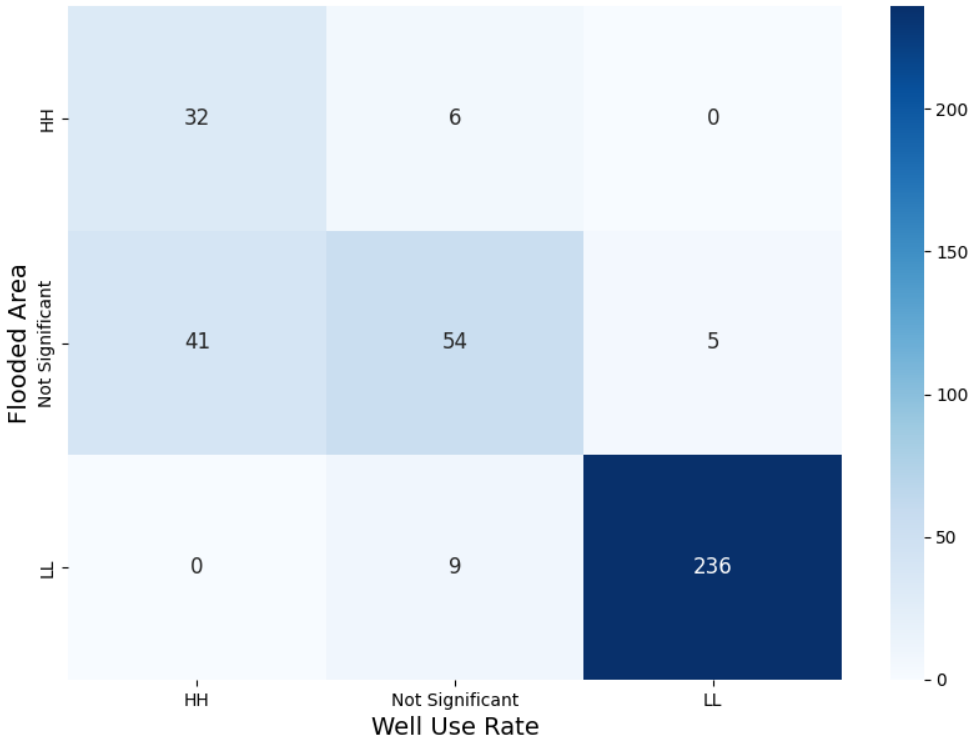


Figure 9 Confusion matrix between private well use cluster and flooded area cluster.

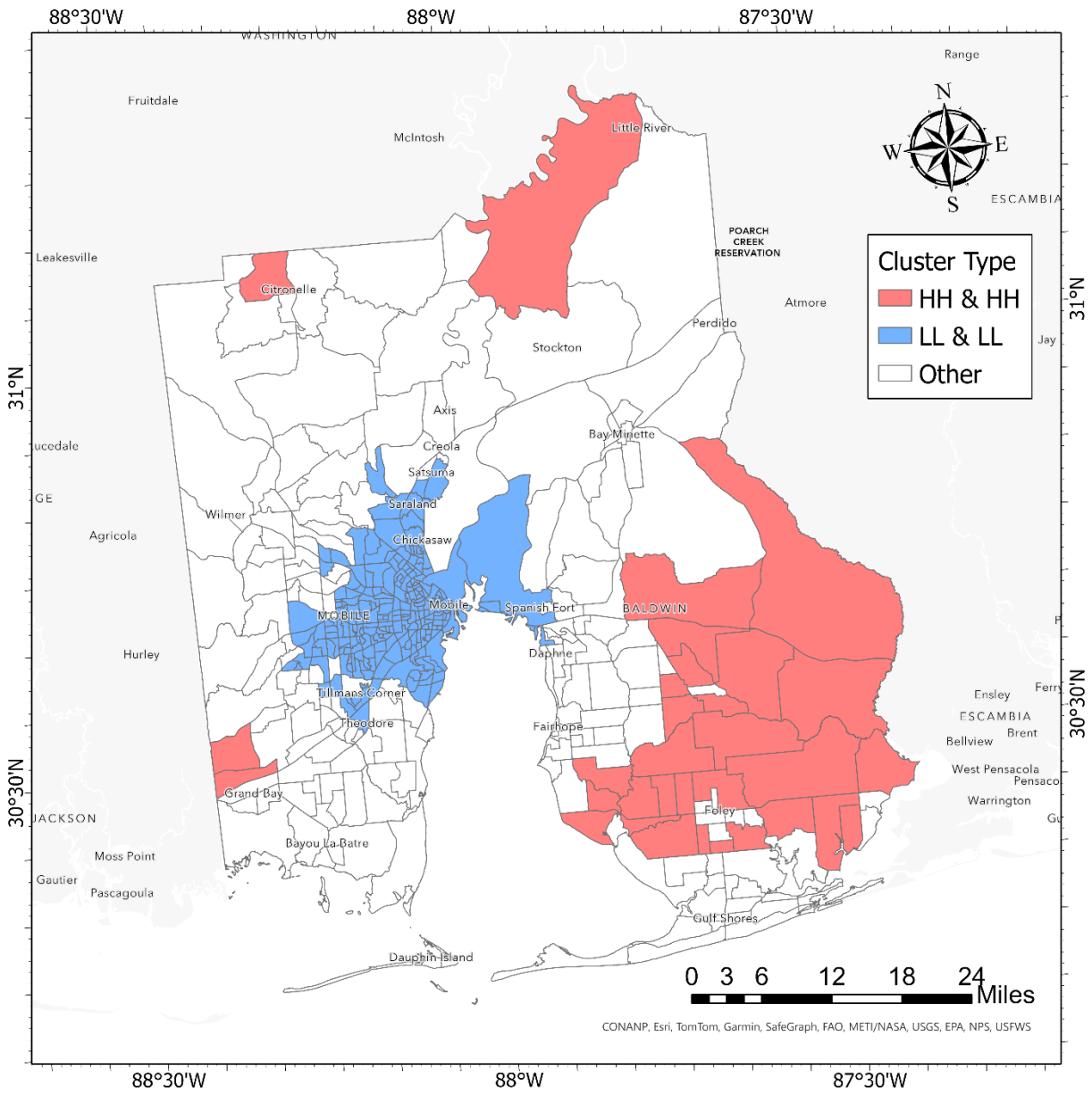


Figure 10 Cluster map. HH & HH is the block groups that have both flooded area and well use rate HH. LL & LL is the block groups that have both variables LL.

3.4 Discussion

Flood monitoring has attracted considerable interest, especially with advances in remote sensing and ML approaches. The availability of high-resolution data has significantly enhanced flood monitoring and accuracy assessment (Jenifer & Natarajan, 2022). Despite the clear methodological advancements in flood monitoring, major concerns arise regarding the collection

of in-situ data and the funding required for high-resolution imagery. Although conventional open-source satellite imagery cannot compare with drone and high-resolution satellite imagery, Sentinel and Landsat remain the only viable options for monitoring floods on a large scale with minimal funding, which is crucial for developing countries. However, both have their limitations. For example, one of the primary challenges with Sentinel-1 is its reliance on SAR data, which, while advantageous for cloud-penetrating capabilities, results in a lower spatial resolution compared to high-resolution optical sensors. This characteristic can limit the detail and accuracy of flood extent mapping, especially in complex urban or vegetated terrains where fine-scale distinctions are crucial. Furthermore, the processing of SAR data requires specialized algorithms to accurately interpret water extents, which can be a barrier for rapid and automated flood monitoring applications (DeVries et al., 2020b). Landsat imagery, while offering a longer historical record useful for change detection and trend analysis, suffers from a relatively lower temporal resolution and is susceptible to cloud cover, limiting its utility for timely flood event monitoring (Twele et al., 2016). The optical nature of Landsat data means that flood events under cloudy conditions or during nighttime cannot be effectively captured, posing significant challenges for real-time or near-real-time flood monitoring efforts (Ogilvie et al., 2018). This study adopts a method mindful of these concerns and shows potential for future improvements.

This study contributes to various aspects of flood monitoring using open-source satellite imagery. For instance, it goes beyond relying solely on Sentinel-1 SAR imagery, the most common open-source dataset for flood monitoring, by incorporating SAR imagery to identify training points through a threshold-based model and employing ML models to extract flooding information. Moreover, the predictive capability of this study also identifies important explanatory variables, offering a foundation for future improvements in model accuracy. Interestingly, all variables chosen for this study play a closely significant role with the importance percentage ranging from 4% to 9%. GCI, NLCD, SMI, and SAVI are considered the most important followed by NDVI, WRI, NDTI, Elevation, NDWI, BAI, and NDFI. EVI and MNDWI are the least important. These findings align with other research on the application of ML in flood prediction (Farhadi & Najafzadeh, 2021; Khosravi et al., 2018; McGrath & Gohl, 2022). The emphasis on variables such as GCI, NLCD, and SMI among the top-ranked variables reflects their direct or indirect influence on flood dynamics. For example, McGrath and Gohl (2022) highlighted the importance of integrating meteorological datasets with hydro-geomorphological variables to improve flood

prediction accuracy using ML models. Their results suggested that variables reflecting land cover, soil moisture, and vegetation health are important predictors of floods which confirms some of the results of RF model (McGrath & Gohl, 2022). The study by Farhadi and Najafzadeh (2021) utilized indices like Elevation, NDVI, and NDWI among others to map flood risk in the Galikesh River basin, demonstrating the indices' substantial contribution to understanding flood dynamics (Farhadi & Najafzadeh, 2021). Similarly, Singh et al. (2015) evaluated NDWI and MNDWI for assessing waterlogging, a key aspect of flood events, underscoring the effectiveness of these indices in delineating water features mixed with vegetation (Singh et al., 2015).

The important findings of this study highlight specific block groups with both high well use rates and extensive flooding. Although only 32 block groups have high well use rates along with extensive flooding and there are significantly more block groups (236) with low well use rates and low flooding, however, it is important to note that these 32 block groups cover a larger total area than the 236 block groups. The high well use and high flood block groups are primarily located in rural settings, accounting for 25% of the total study area, whereas the low flood and low well use block groups cover only 8% of the total study area and are mainly situated in urban settings. Although flood water may dissipate relatively quickly following a heavy precipitation event, it is important to know where the flood water accumulated. In the case of private well water, users may become victims of unexpected drinking water contamination following a flooding event. This can be caused by a number of reasons. For example, a study assessing the risk of drinking well contamination following the 2013 Calgary flood found that environmental factors, rather than the degree of submergence, played a crucial role in well water contamination, underscoring the complexity of predicting and managing flood-induced water quality issues (Eccles et al., 2017). Pre-flood vs post flood comparisons revealed that this contamination occurred after the flood which highlights the importance of knowing where and to what extent areas become inundated. The study by Masciopinto et al. (2019) found that severe flooding can lead to significant microbiological contamination of drinking water sources from wells. The research utilized a mathematical model to predict the fate and transport of viruses in groundwater, demonstrating that floodwaters can carry viruses several kilometers from their point of origin. The study underscores the importance of implementing additional water disinfection measures and regular monitoring of enteric viruses to ensure the safety of well water in flood-prone areas (Masciopinto et al., 2019). These studies collectively illustrate the critical need for targeted education and infrastructure

improvements to enhance resilience against flood-induced well water contamination which is possibly a major implementation of the predictive model proposed in this study.

3.5 Conclusion

Despite several advantages of the techniques used in this study, the model faces limitations due to its methodology and data quality. Primarily, it suits only minimal flooding scenarios, as it still relies on conventional multispectral imagery from Sentinel-2 for explanatory variables. Obtaining a post-flood image with acceptable cloud coverage, often unavailable during large-scale flooding, remains a challenge. Additionally, the training data for this model comes from a threshold-based model and compares pre-flood and post-flood imagery at 10-meter resolution, which identifies the most probable rather than certain flooded areas. Increased soil moisture does not always indicate flooding, but given the timing of the post-flood imagery, it most likely results from flooding.

One of the major challenges of this study is considering the protected wetland areas of the Mobile delta. Even though the model includes exploratory variables like GCI that can consider wetlands when predicting flood (Andreo et al., 2019), the temporal and spatial variability of these landscapes makes reliable data collection and model training difficult. The inherent variability in wetland hydrology, influenced by factors such as precipitation, evaporation, soil saturation, and plant uptake, complicates the development of consistent and reliable prediction models which is evident by several other similar studies (Herath et al., 2023; Jayathilake et al., 2023).

Future opportunities exist to address some limitations. For example, this study employs only one ML method, which could expand by incorporating other methods like Gradient Boosting, Naïve Bayes, and Support Vector Machine (SVM). Employing multiple methods could provide insights through a comparative analysis of each model to determine the most effective combination with the threshold-based training model.

This model excels by relying solely on open-source data and eliminating the need for additional field visits. It can serve as an initial step to identify potential flooding zones and act as a trail for investigating other important flooding-related factors, such as groundwater contamination. Future research could correlate flooding with contamination by collecting in-situ groundwater data before and after flood events. Additionally, applying this model to assess land-

use-specific flood risks and predict flooded areas from one region to another offers promising directions. For instance, this study predicted flooding in Baldwin County based on data from Mobile County flooding.

Chapter 4

4.1 Key findings

The first objective of this study has successfully utilized a geospatial approach to generate a comprehensive risk landscape, or "risk-scape," associated with private well use across Alabama, focusing on socio-economic vulnerability, flood potential, and anthropogenic risks. The innovative cluster-based methodology enables a nuanced understanding of the interplay between various risk factors and demographics, identifying areas where socio-economically disadvantaged well users are co-located with high flood risk and proximity to toxic release facilities. By not presuming the magnitude of risk posed by any single indicator, this approach offers a holistic mechanism to categorize well user communities based on similar risk profiles, thereby highlighting the most probable sources of contamination. The dual nature of the framework, incorporating both exploratory cluster analysis and explanatory regression, ensures a robust assessment of risk, demonstrating that private well users in specific clusters are particularly susceptible to flood-related contamination. This methodology serves as a potent decision support tool, aiding in the strategic deployment of groundwater contamination mitigation resources tailored to the needs of different well user communities.

The second objective of this study demonstrates significant advancements in flood monitoring by integrating remote sensing and ML techniques. By incorporating high-resolution data and drone imagery, the study enhances the accuracy of flood detection and assessment, despite the traditional reliance on open-source datasets like Sentinel-1 and Landsat. The research reveals that variables such as the GCI, NLCD, SMI, and SAVI are crucial for accurate flood prediction. The methodology employed not only utilizes Sentinel-1 SAR imagery for threshold-based model training but also integrates ML models to extract detailed flooding information, demonstrating a robust approach to identifying flood-prone areas. The results highlight specific block groups with high well use rates and significant flooding, predominantly located in rural areas. This geographic clustering underscores the potential vulnerability of these communities to flood-induced water contamination, necessitating targeted mitigation strategies.

4.2 Limitations

For the first objective, despite the nature of this study, several limitations warrant consideration. A significant challenge lies in the reliance on existing datasets, which may not accurately capture the current state of well water usage and contamination risks. Historical data might not reflect recent demographic shifts or land use changes, potentially skewing the risk assessments. The study's geographical focus on Alabama means that specific findings may not be directly applicable to other regions with different socio-economic and environmental contexts, though the methodology itself is adaptable. Additionally, the study does not encompass all possible contamination factors, such as emerging contaminants like pharmaceuticals and personal care products, or geogenic risks like arsenic and uranium. The impact of climate change, which could alter precipitation patterns and exacerbate flooding risks, is another critical factor not fully addressed. Furthermore, there are technical constraints related to aligning datasets of varying spatial resolutions, which could affect the precision of the risk assessments, particularly in smaller CBGs.

The second objective faces several limitations primarily related to data quality and methodological constraints. The reliance on conventional multispectral imagery from Sentinel-2 poses challenges, particularly in obtaining cloud-free post-flood images during large-scale flooding events. The model's training data, derived from a threshold-based approach, identifies likely rather than definite flooded areas, which may lead to inaccuracies in flood extent mapping. Additionally, increased soil moisture does not always equate to flooding, complicating the interpretation of post-flood imagery. The study's applicability is somewhat constrained to minimal flooding scenarios, and the use of only one ML method limits the exploration of potentially more effective techniques. The findings, while significant, may not fully capture the complexity of flood dynamics and their impacts on groundwater contamination, highlighting the need for more comprehensive approaches.

4.3 Future opportunities

Looking ahead, there are several promising avenues for advancing this research. Enhancing data collection methods to provide more dynamic and real-time monitoring of well water use and contamination levels is crucial. Advances in remote sensing technologies could offer more precise and current information on flood risks, thereby improving risk assessments, which has been further discussed in the second objective. Broadening the scope of the study to include a wider range of contaminants, particularly in the context of climate change, would yield a more comprehensive understanding of well water contamination risks. Incorporating participatory research approaches that engage local communities in the monitoring and management of their well water resources could lead to more sustainable and effective solutions. By involving well users in data collection, risk assessment, and decision-making processes, the study could foster community-centric strategies that ensure the long-term sustainability and safety of private well water supplies. Such approaches not only enhance the accuracy of risk assessments but also empower communities to take proactive steps in safeguarding their water resources.

For the second objective, expanding the range of ML methods, such as incorporating Gradient Boosting, Naïve Bayes, and SVM, could enhance model performance through comparative analysis. This multi-method approach would provide deeper insights into the most effective combinations for flood prediction. The study's reliance on open-source data and elimination of field visits make it a valuable initial tool for identifying potential flooding zones. Future research could build on this by correlating flooding events with groundwater contamination through the collection of in-situ data before and after floods. Additionally, applying this model to various land-use scenarios and different geographic regions could refine its predictive capabilities. For example, assessing flood risks in Baldwin County using data from Mobile County demonstrates the model's potential for broader application, paving the way for more comprehensive flood risk assessments and mitigation strategies.

References

ADECA. (2020). Floodplain management. *Renewable Resources Journal*, *34*(1). https://doi.org/10.1201/noe0849396274.ch93

ADPH. (2021). *Well Water*.

Aldardasawi, A. F. M., & Eren, B. (2021). Floods and Their Impact on the Environment. *Academic Perspective Procedia*, *4*(2), 42–49. https://doi.org/10.33793/acperpro.04.02.24

Aller, L., Bennett, T., Lehr, J. H., Petty, R. J., & Hackett, G. (1987). DRASTIC : A Standardized Method for Evaluating Ground Water Pollution Potential Using Hydrogeologic Settings. *NWWA/Epa-600/2-87-035*, 455.

Amini, M., Abbaspour, K. C., Berg, M., Winkel, L., Hug, S. J., Hoehn, E., Yang, H., & Johnson, C. A. (2008). Statistical modeling of global geogenic arsenic contamination in groundwater. *Environmental Science and Technology*, *42*(10), 3669–3675. https://doi.org/10.1021/es702859e

Andrade, L., O'Dwyer, J., O'Neill, E., & Hynds, P. (2018). Surface water flooding, groundwater contamination, and enteric disease in developed countries: A scoping review of connections and consequences. In *Environmental Pollution* (Vol. 236, pp. 540–549). https://doi.org/10.1016/j.envpol.2018.01.104

Andreo, V., Belgiu, M., Hoyos, D. B., Osei, F., Provensal, C., & Stein, A. (2019). Rodents and satellites: Predicting mice abundance and distribution with Sentinel-2 data. *Ecological Informatics*, *51*, 157–167. https://doi.org/10.1016/j.ecoinf.2019.03.001

Awadallah, A. G., & Tabet, D. (2015). Estimating flooding extent at high return period for ungauged braided systems using remote sensing: a case study of Cuvelai Basin, Angola. *Natural Hazards*, *77*(1), 255–272. https://doi.org/10.1007/s11069-015-1600-6

Ayotte, J. D., Medalie, L., Qi, S. L., Backer, L. C., & Nolan, B. T. (2017). Estimating the High-Arsenic Domestic-Well Population in the Conterminous United States. *Environmental Science and Technology*, *51*(21), 12443–12454. https://doi.org/10.1021/acs.est.7b02881

Balenzano, A., Mattia, F., Satalino, G., & Davidson, M. W. J. (2011). Dense Temporal Series of C- and L-band SAR Data for Soil Moisture Retrieval Over Agricultural Crops. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *4*(2), 439–450. https://doi.org/10.1109/JSTARS.2010.2052916

Batt, A. L., Snow, D. D., & Aga, D. S. (2006). Occurrence of sulfonamide antimicrobials in private water wells in Washington County, Idaho, USA. *Chemosphere*, *64*(11), 1963–1971. https://doi.org/10.1016/j.chemosphere.2006.01.029

Bontemps, S., Arias, M., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Matton, N., Morin, D., Popescu, R., Rabaute, T., Savinaud, M., Sepulcre, G., Valero, S., Ahmad, I., Bégué, A., Wu, B., de Abelleyra, D., Diarra, A., … Defourny, P. (2015). Building a data set over 12 globally distributed sites to support the development of agriculture monitoring applications with Sentinel-2. *Remote Sensing*, *7*(12), 16062–16090. https://doi.org/10.3390/rs71215815

Borchardt, M. A., Bertz, P. D., Spencer, S. K., & Battigelli, D. A. (2003). Incidence of enteric viruses in groundwater from household wells in Wisconsin. *Applied and Environmental Microbiology*, *69*(2), 1172–1180. https://doi.org/10.1128/AEM.69.2.1172-1180.2003

Borchardt, M. A., Stokdyk, J. P., Kieke, B. A., Muldoon, M. A., Spencer, S. K., Firnstahl, A. D., Bonness, D. E., Hunt, R. J., & Burch, T. R. (2021). Sources and risk factors for nitrate and microbial contamination of private household wells in the fractured dolomite aquifer of northeastern wisconsin. *Environmental Health Perspectives*, *129*(6). https://doi.org/10.1289/EHP7813

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Breiman, L. (2001a). Random forests. Random Forests, 1–122. *Machine Learning*, *45*(45), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. CRC Press. https://doi.org/10.1201/9781315139470

Burkart, M. R., Kolpin, D. W., & James, D. E. (1999). Assessing groundwater vulnerability to agrichemical contamination in the Midwest US. *Water Science and Technology*, *39*(3), 103–112. https://doi.org/10.1016/S0273-1223(99)00042-6

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, *48*(1–3), 287–297. https://doi.org/10.1023/A:1013964023376

Caliñski, T., & Harabasz, J. (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Casaseca-de-la-Higuera, P., Tristán Vega, A., Merino-Caviedes, S., Wang, Q., Luo, C., Wang, X., Wang, Z., & Hoyos-Barceló, C. (2018). Compressed UAV sensing for flood monitoring by solving the continuous travelling salesman problem over hyperspectral maps. *Spiedigitallibrary.OrgP Casaseca-De-La-Higuera, A Tristán-Vega, C Hoyos-Barceló, S Merino-Caviedes, Q WangRemote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large …, 2018•spiedigitallibrary.Org*, *10784*(10), 12. https://doi.org/10.1117/12.2325645

Chastain, R., Housman, I., Goldstein, J., & Finco, M. (2019). Empirical cross sensor comparison of Sentinel-2A and 2B MSI, Landsat-8 OLI, and Landsat-7 ETM+ top of atmosphere spectral characteristics over the conterminous United States. *Remote Sensing of Environment*, *221*, 274–285. https://doi.org/10.1016/j.rse.2018.11.012

Chicco, D., Starovoitov, V., & Jurman, G. (2021). The Benefits of the Matthews Correlation Coefficient (MCC) over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. *IEEE Access*, *9*, 47112–47124. https://doi.org/10.1109/ACCESS.2021.3068614

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification

Assessment. *IEEE Access*, *9*, 78368–78381. https://doi.org/10.1109/ACCESS.2021.3084050

Cho, G., Jung, K., & Hwang, H. (2019). Out-of-bag Prediction Error: A Cross Validation Index for Generalized Structured Component Analysis. *Multivariate Behavioral Research*, *54*(4), 505–513. https://doi.org/10.1080/00273171.2018.1540340

Coyte, R. M., Jain, R. C., Srivastava, S. K., Sharma, K. C., Khalil, A., Ma, L., & Vengosh, A. (2018). Large-Scale Uranium Contamination of Groundwater Resources in India. *Environmental Science and Technology Letters*, *5*(6), 341–347. https://doi.org/10.1021/acs.estlett.8b00215

Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., Calderon, R. L., Roberts, J. M., Beach, M. J., & Roy, S. L. (2010). Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. In *Clinical Microbiology Reviews* (Vol. 23, Issue 3, pp. 507–528). American Society for Microbiology. https://doi.org/10.1128/CMR.00077-09

DeSimone, L. A., Hamilton, P. A., & Gilliom, R. J. (2009). Quality of water from domestic wells in principal aquifers of the United States, 1991–2004—Overview of major findings. *Water*, *Circular 1*, 1991–2004.

DeVries, B., Huang, C., Armston, J., Huang, W., Jones, J. W., & Lang, M. W. (2020a). Rapid and robust monitoring of flood events using Sentinel-1 and Landsat data on the Google Earth Engine. *Remote Sensing of Environment*, *240*. https://doi.org/10.1016/j.rse.2020.111664

DeVries, B., Huang, C., Armston, J., Huang, W., Jones, J. W., & Lang, M. W. (2020b). Rapid and robust monitoring of flood events using Sentinel-1 and Landsat data on the Google Earth Engine. *Remote Sensing of Environment*, *240*(October 2018), 111664. https://doi.org/10.1016/j.rse.2020.111664

Domeneghetti, A., Schumann, G. J. P., & Tarpanelli, A. (2019). Preface: Remote sensing for flood mapping and monitoring of flood dynamics. In *Remote Sensing* (Vol. 11, Issue 8). https://doi.org/10.3390/rs11080940

Drewry, K. R., Jones, C. N., Hayes, W., Beighley, R. E., Wang, Q., Hochard, J., Mize, W., Fowlkes, J., Goforth, C., & Pieper, K. J. (2024). Using Inundation Extents to Predict Microbial Contamination in Private Wells after Flooding Events. *Environmental Science and Technology*, *58*(12), 5220–5228. https://doi.org/10.1021/acs.est.3c09375

Du, W., Fitzgerald, G. J., Clark, M., & Hou, X. Y. (2010). Health impacts of floods. In *Prehospital and Disaster Medicine* (Vol. 25, Issue 3, pp. 265–272). Cambridge University Press. https://doi.org/10.1017/S1049023X00008141

Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., & Li, X. (2016). Water bodies' mapping from Sentinel-2 imagery with Modified Normalized Difference Water Index at 10-m spatial resolution produced by sharpening the swir band. *Remote Sensing*, *8*(4). https://doi.org/10.3390/rs8040354

Eccles, K. M., Checkley, S., Sjogren, D., Barkema, H. W., & Bertazzon, S. (2017). Lessons learned from the 2013 Calgary flood: Assessing risk of drinking water well contamination. *Applied Geography*, *80*, 78–85. https://doi.org/10.1016/j.apgeog.2017.02.005

EPA. (2015). *Safe Drinking Water Act (SDWA) Safe Drinking Water Act US EPA*. https://www.epa.gov/sdwa

ESA. (2022). Sentinel-2 - Missions - Sentinel Online - Sentinel Online. In *European Space Agency*. https://sentinel.esa.int/web/sentinel/missions/sentinel-2

ESRI. (2022). *How Forest-based Classification and Regression works—ArcGIS Pro | Documentation*. https://pro.arcgis.com/en/pro-app/3.1/tool-reference/spatial-statistics/how-forest-works.htm

ESRI. (2024). *Multivariate Clustering (Spatial Statistics)—ArcGIS Pro | Documentation*. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/multivariate-clustering.htm

Farhadi, H., & Najafzadeh, M. (2021). Flood risk mapping by remote sensing data and random forest technique. *Water (Switzerland)*, *13*(21), 3115. https://doi.org/10.3390/w13213115

Fizer, C., de Bruin, W. B., Stillo, F., & Gibson, J. M. (2018). Barriers to managing private wells and septic systems in underserved communities: Mental models of homeowner decision making. *Journal of Environmental Health*, *81*(5), 8–15.

Flanagan, S. V., Marvinney, R. G., & Zheng, Y. (2015). Influences on domestic well water testing behavior in a Central Maine area with frequent groundwater arsenic occurrence. *Science of the Total Environment*, *505*, 1274–1281. https://doi.org/10.1016/j.scitotenv.2014.05.017

Flanagan, S. V., Spayd, S. E., Procopio, N. A., Marvinney, R. G., Smith, A. E., Chillrud, S. N., Braman, S., & Zheng, Y. (2016). Arsenic in private well water part 3 of 3: Socioeconomic vulnerability to exposure in Maine and New Jersey. *Science of the Total Environment*, *562*, 1019–1030. https://doi.org/10.1016/j.scitotenv.2016.03.217

FloodList. (2019). *USA – Deadly Flash Floods in Tennessee and Alabama – FloodList*. https://floodlist.com/america/usa/floods-tennessee-alabama-december-2019

Geological Survey of Alabama. (2007). *Groundwater Assessment Program*. https://www.gsa.state.al.us/gsa/groundwater/currentprojects

Geological Survey of Alabama. (2018). *2010-16 Statewide Groundwater Assessment*. https://www.gsa.state.al.us/gsa/groundwater/assessment

Gitter, A., Boellstorff, D. E., Mena, K. D., Gholson, D. M., Pieper, K. J., Chavarria, C. A., & Gentry, T. J. (2023). Quantitative Microbial Risk Assessment for Private Wells in Flood-Impacted Areas. *Water (Switzerland)*, *15*(3). https://doi.org/10.3390/w15030469

Goffi, A., Stroppiana, D., Brivio, P. A., Bordogna, G., & Boschetti, M. (2020). Towards an automated approach to map flooded areas from Sentinel-2 MSI data and soft integration of water spectral features. *International Journal of Applied Earth Observation and Geoinformation*, *84*. https://doi.org/10.1016/j.jag.2019.101951

Google Developers. (2022). *Sentinel-2: Cloud Probability | Earth Engine Data Catalog*. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR

Goss, M. J., Barry, D. A. J., & Rudolph, D. L. (1998). Contamination in Ontario farmstead

domestic wells and its association with agriculture: 1. Results from drinking water wells. *Journal of Contaminant Hydrology*, *32*(3–4), 267–293. https://doi.org/10.1016/S0169-7722(98)00054-0

Goudarzi, S., Soleymani, S. A., Anisi, M. H., Ciuonzo, D., Kama, N., Abdullah, S., Azgomi, M. A., Chaczko, Z., & Azmi, A. (2021). Real-time and intelligent flood forecasting using UAV-assisted wireless sensor network. *Computers, Materials and Continua*, *70*(1), 715–738. https://doi.org/10.32604/cmc.2022.019550

Guo, D., Bai, Y., Svinin, M., & Magid, E. (2021). Robust Adaptive Multi-Agent Coverage Control for Flood Monitoring. *SIBCON 2021 - International Siberian Conference on Control and Communications*. https://doi.org/10.1109/SIBCON50419.2021.9438872

Harman, W. A., Allan, C. J., & Forsythe, R. D. (2001). Assessment of potential groundwater contamination sources in a wellhead protection area. *Journal of Environmental Management*, *62*(3), 271–282. https://doi.org/10.1006/jema.2001.0436

Hasenmueller, E. A., & Robinson, H. K. (2016). Hyporheic zone flow disruption from channel linings: Implications for the hydrology and geochemistry of an urban stream, St. Louis, Missouri, USA. *Journal of Earth Science*, *27*(1), 98–109. https://doi.org/10.1007/s12583-016-0632-5

Herath, M., Jayathilaka, T., Hoshino, Y., & Rathnayake, U. (2023). Deep Machine Learning-Based Water Level Prediction Model for Colombo Flood Detention Area. *Applied Sciences (Switzerland)*, *13*(4). https://doi.org/10.3390/app13042194

Hu, X. C., Ge, B., Ruyle, B. J., Sun, J., & Sunderland, E. M. (2021). A Statistical Approach for Identifying Private Wells Susceptible to Perfluoroalkyl Substances (PFAS) Contamination. *Environmental Science and Technology Letters*, *8*(7), 596–602. https://doi.org/10.1021/acs.estlett.1c00264

Hynds, P., Misstear, B. D., Gill, L. W., & Murphy, H. M. (2014). Groundwater source contamination mechanisms: Physicochemical profile clustering, risk factor analysis and multivariate modelling. *Journal of Contaminant Hydrology*, *159*, 47–56. https://doi.org/10.1016/j.jconhyd.2014.02.001

Imgrund, K., Kreutzwiser, R., & de Loë, R. (2011). Influences on the water testing behaviors of private well owners. *Journal of Water and Health*, *9*(2), 241–252. https://doi.org/10.2166/wh.2011.139

Jafarzadegan, K., Merwade, V., & Saksena, S. (2018). A geomorphic approach to 100-year floodplain mapping for the Conterminous United States. *Journal of Hydrology*, *561*, 43–58. https://doi.org/10.1016/j.jhydrol.2018.03.061

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jayathilake, T., Sarukkalige, R., Hoshino, Y., & Rathnayake, U. (2023). Wetland Water Level Prediction Using Artificial Neural Networks—A Case Study in the Colombo Flood Detention Area, Sri Lanka. *Climate*, *11*(1). https://doi.org/10.3390/cli11010001

Jenifer, A. E., & Natarajan, S. (2022). DeepFlood: A deep learning based flood detection

framework using feature-level fusion of multi-sensor remote sensing images. *Forum for Nordic Dermato-Venerology*, *28*(3), 329–343. https://doi.org/10.3897/jucs.80734

Johnson, R. A. (2000). An Analysis of Transformations on JSTOR. *Biometrika*, *6*, 954–959. https://www.jstor.org/stable/2984418

Joseph, A. T., Van Der Velde, R., O'Neill, P. E., Lang, R. H., & Gish, T. (2008). Soil moisture retrieval during a corn growth cycle using L-band (1.6 GHz) radar observations. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(8), 2365–2374. https://doi.org/10.1109/TGRS.2008.917214

Kashyap, M., Bhatt, C. M., & Rawat, J. S. (2022). Application of Sentinel-2 Data for Extraction of Flood Inundation along Ganga River, Bihar. *International Journal for Research in Applied Science and Engineering Technology*, *10*(3), 1983–1991. https://doi.org/10.22214/ijraset.2022.41015

Keeler, B. L., & Polasky, S. (2014). Land-use change and costs to rural households: A case study in groundwater nitrate contamination. *Environmental Research Letters*, *9*(7), 074002. https://doi.org/10.1088/1748-9326/9/7/074002

Khan, H. K., Rehman, M. Y. A., Junaid, M., Lv, M., Yue, L., Haq, I. ul, Xu, N., & Malik, R. N. (2022). Occurrence, source apportionment and potential risks of selected PPCPs in groundwater used as a source of drinking water from key urban-rural settings of Pakistan. *Science of the Total Environment*, *807*. https://doi.org/10.1016/j.scitotenv.2021.151010

Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., & Tien Bui, D. (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. *Science of the Total Environment*, *627*, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266

Kolpin, D. W. (1997). Agricultural Chemicals in Groundwater of the Midwestern United States: Relations to Land Use. *Journal of Environmental Quality*, *26*(4), 1025–1037. https://doi.org/10.2134/jeq1997.00472425002600040014x

Krolik, J., Maier, A., Evans, G., Belanger, P., Hall, G., Joyce, A., & Majury, A. (2013). A spatial analysis of private well water Escherichia coli contamination in Southern Ontario. *Geospatial Health*, *8*(1), 65–75. https://doi.org/10.4081/gh.2013.55

Lechner, A. M., Stein, A., Jones, S. D., & Ferwerda, J. G. (2009). Remote sensing of small and linear features: Quantifying the effects of patch size and length, grid position and detectability on land cover mapping. *Remote Sensing of Environment*, *113*(10), 2194–2204. https://doi.org/10.1016/j.rse.2009.06.002

Ledien, J., Sorn, S., Hem, S., Huy, R., Buchy, P., Tarantola, A., & Cappelle, J. (2017). Assessing the performance of remotelysensed flooding indicators and their potential contribution to early warning for leptospirosis in Cambodia. *PLoS ONE*, *12*(7). https://doi.org/10.1371/journal.pone.0181044

Lemonte, J. J., Stuckey, J. W., Sanchez, J. Z., Tappero, R., Rinklebe, J., & Sparks, D. L. (2017). Sea Level Rise Induced Arsenic Release from Historically Contaminated Coastal Soils. *Environmental Science and Technology*, *51*(11), 5913–5922.

https://doi.org/10.1021/acs.est.6b06152

Li, J., & Roy, D. P. (2017). A global analysis of Sentinel-2a, Sentinel-2b and Landsat-8 data revisit intervals and implications for terrestrial monitoring. *Remote Sensing*, *9*(9). https://doi.org/10.3390/rs9090902

Li, X., Wu, H., & Qian, H. (2020). Groundwater contamination risk assessment using intrinsic vulnerability, pollution loading and groundwater value: a case study in Yinchuan plain, China. *Environmental Science and Pollution Research*, *27*(36), 45591–45604. https://doi.org/10.1007/s11356-020-10221-4

Liang, J., & Liu, D. (2020). A local thresholding approach to flood water delineation using Sentinel-1 SAR imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *159*, 53–62. https://doi.org/10.1016/j.isprsjprs.2019.10.017

Lo, S. W., Wu, J. H., Lin, F. P., & Hsu, C. H. (2015). Visual sensing for urban flood monitoring. *Sensors (Switzerland)*, *15*(8), 20006–20029. https://doi.org/10.3390/s150820006

Lombard, M. A., Bryan, M. S., Jones, D. K., Bulka, C., Bradley, P. M., Backer, L. C., Focazio, M. J., Silverman, D. T., Toccalino, P., Argos, M., Gribble, M. O., & Ayotte, J. D. (2021). Machine Learning Models of Arsenic in Private Wells throughout the Conterminous United States As a Tool for Exposure Assessment in Human Health Studies. *Environmental Science and Technology*, *55*(8), 5012–5023. https://doi.org/10.1021/acs.est.0c05239

Malecki, K. M. C., Schultz, A. A., Severtson, D. J., Anderson, H. A., & VanDerslice, J. A. (2017). Private-well stewardship among a general population based sample of private well-owners. *Science of The Total Environment*, *601–602*, 1533–1543. https://doi.org/https://doi.org/10.1016/j.scitotenv.2017.05.284

Manjusree, P., Prasanna Kumar, L., Bhatt, C. M., Rao, G. S., & Bhanumurthy, V. (2012). Optimization of threshold ranges for rapid flood inundation mapping by evaluating backscatter profiles of high incidence angle SAR images. *International Journal of Disaster Risk Science*, *3*(2), 113–122. https://doi.org/10.1007/s13753-012-0011-5

Manson, S., Schroeder, J., Van Riper, D., Knowles, K., Kugler, T., Roberts, F., & Ruggles, S. (2023). IPUMS National Historical Geographic Information System: Version 18.0 [dataset]. *IPUMS*. https://experts.umn.edu/en/publications/ipums-national-historical-geographic-information-system-version-1

Mapili, K., Rhoads, W. J., Coughter, M., Pieper, K. J., Edwards, M. A., & Pruden, A. (2022). Occurrence of opportunistic pathogens in private wells after major flooding events: A four state molecular survey. *Science of the Total Environment*, *826*, 153901. https://doi.org/10.1016/j.scitotenv.2022.153901

Martinez-Morata, I., Bostick, B. C., Conroy-Ben, O., Duncan, D. T., Jones, M. R., Spaur, M., Patterson, K. P., Prins, S. J., Navas-Acien, A., & Nigra, A. E. (2022). Nationwide geospatial analysis of county racial and ethnic composition and public drinking water arsenic and uranium. *Nature Communications*, *13*(1), 1–12. https://doi.org/10.1038/s41467-022-35185-6

Masciopinto, C., De Giglio, O., Scrascia, M., Fortunato, F., La Rosa, G., Suffredini, E., Pazzani, C., Prato, R., & Montagna, M. T. (2019). Human health risk assessment for the occurrence

of enteric viruses in drinking water from wells: Role of flood runoff injections. *Science of the Total Environment*, *666*, 559–571. https://doi.org/10.1016/j.scitotenv.2019.02.107

McGrath, H., & Gohl, P. N. (2022). Accessing the Impact of Meteorological Variables on Machine Learning Flood Susceptibility Mapping. *Remote Sensing*, *14*(7). https://doi.org/10.3390/rs14071656

McMaster, R. B., & Noble, P. (2005). The U.S. national historical geographic information system. In *Historical Geography* (Vol. 33, pp. 134–136). https://dev.icaci.org/files/documents/ICC_proceedings/ICC2003/Papers/099.pdf

Mooney, S., Boudou, M., O'Dwyer, J., & Hynds, P. D. (2022). Behavioral pathways to private well risk mitigation: A structural equation modeling approach. *Risk Analysis*, 1–28. https://doi.org/10.1111/risa.14021

Mooney, S., O'Dwyer, J., & Hynds, P. (2021). Groundwater Contamination and Extreme Weather Events: Perception-Based Clusters of Irish Well Users. In *Advances in Science, Technology and Innovation* (pp. 331–334). Springer Nature. https://doi.org/10.1007/978-3-030-59320-9_68

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. In *Water (Switzerland)* (Vol. 10, Issue 11). https://doi.org/10.3390/w10111536

Murray, A. H., & Kremer., F. (2023). *U.S. Private Domestic Wells 2020 (Well Density)*.

Murray, A., Hall, A., Weaver, J., & Kremer, F. (2021). Methods for Estimating Locations of Housing Units Served by Private Domestic Wells in the United States Applied to 2010. *Journal of the American Water Resources Association*, *57*(5), 828–843. https://doi.org/10.1111/1752-1688.12937

Musacchio, A., Andrade, L., O'Neill, E., Re, V., O'Dwyer, J., & Hynds, P. D. (2021). Planning for the health impacts of climate change: Flooding, private groundwater contamination and waterborne infection – A cross-sectional study of risk perception, experience and behaviours in the Republic of Ireland. *Environmental Research*, *194*. https://doi.org/10.1016/j.envres.2021.110707

Nolan, B. T., & Hitt, K. J. (2006). Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. *Environmental Science and Technology*, *40*(24), 7834–7840. https://doi.org/10.1021/es060911u

Ogilvie, A., Belaud, G., Massuel, S., Mulligan, M., Le Goulven, P., & Calvez, R. (2018). Surface water monitoring in small water bodies: Potential and limits of multi-sensor Landsat time series. *Hydrology and Earth System Sciences*, *22*(8), 4349–4380. https://doi.org/10.5194/hess-22-4349-2018

Osidach, V. Z. (2021). *Water Contaminant Risk Awareness Among Northeastern Ohio Well Water Users: A Qualitative Study*. Northcentral University.

Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

Pieper, K. J., Rhoads, W. J., Jones, C. N., Rome, M., Gholson, D. M., Katner, A., Boellstor, D. E., & Beighley, R. E. (2021). Microbial contamination of drinking water supplied by private wells after hurricane harvey. *Environmental Science and Technology*, *55*(12), 8382–8392. https://doi.org/10.1021/acs.est.0c07869

PRISM Climate Group. (2022). *PRISM Climate Group, Oregon State University. http://prism.oregonstate.edu.* https://prism.oregonstate.edu/

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*, 1–8. https://www.jmlr.org/papers/v18/17-269.html

Radelyuk, I., Tussupova, K., Persson, M., Zhapargazinova, K., & Yelubay, M. (2021). Assessment of groundwater safety surrounding contaminated water storage sites using multivariate statistical analysis and Heckman selection model: a case study of Kazakhstan. *Environmental Geochemistry and Health*, *43*(2), 1029–1050. https://doi.org/10.1007/s10653-020-00685-1

Rajab, A., Farman, H., Islam, N., Syed, D., Elmagzoub, M. A., Shaikh, A., Akram, M., & Alrizq, M. (2023). Flood Forecasting by Using Machine Learning: A Study Leveraging Historic Climatic Records of Bangladesh. *Water (Switzerland)*, *15*(22). https://doi.org/10.3390/w15223970

Ramesh, B., Callender, R., Zaitchik, B. F., Jagger, M., Swarup, S., & Gohlke, J. M. (2023). Adverse Health Outcomes Following Hurricane Harvey: A Comparison of Remotely-Sensed and Self-Reported Flood Exposure Estimates. *GeoHealth*, *7*(4). https://doi.org/10.1029/2022GH000710

Resek, J. E. B. & E. A. (1996). Pesticides in Ground Water. Distribution trends and governing factors. *INFO-NYT*, *14*, pp 588.

Roostaei, J., Colley, S., Mulhern, R., May, A. A., & Gibson, J. M. D. (2021). Predicting the risk of GenX contamination in private well water using a machine-learned Bayesian network model. *Journal of Hazardous Materials*, *411*, 125075. https://doi.org/10.1016/j.jhazmat.2021.125075

Rowles, L. S., Hossain, A. I., Ramirez, I., Durst, N. J., Ward, P. M., Kirisits, M. J., Araiza, I., Lawler, D. F., & Saleh, N. B. (2020). Seasonal contamination of well-water in flood-prone colonias and other unincorporated U.S. communities. *Science of the Total Environment*, *740*, 140111. https://doi.org/10.1016/j.scitotenv.2020.140111

Sampurno, J., Vallaeys, V., Ardianto, R., & Hanert, E. (2022). Integrated hydrodynamic and machine learning models for compound flooding prediction in a data-scarce estuarine delta. *Nonlinear Processes in Geophysics*, *29*(3), 301–315. https://doi.org/10.5194/npg-29-301-2022

Sankaranarayanan, S., Prabhakar, M., Satish, S., Jain, P., Ramprasad, A., & Krishnan, A. (2020). Flood prediction based on weather parameters using deep learning. *Journal of Water and Climate Change*, *11*(4), 1766–1783. https://doi.org/10.2166/wcc.2019.321

Schroeder, J. P., & McMaster, R. B. (2007). The creation of a multiscale national historical geographic information system for the United States Census. *Proceedings of the 23rd*

International Cartographic Conference. Moscow, Russia, August 4-10, 2007. https://icaci.org/files/documents/ICC_proceedings/ICC2007/documents/doc/THEME 10/Oral 3/The Creation of a Multiscale National Historical Geographic .doc

Schumann, G. J. P. (2015). Preface: Remote sensing in flood monitoring and management. In *Remote Sensing* (Vol. 7, Issue 12, pp. 17013–17015). https://doi.org/10.3390/rs71215871

Shamshiri, R., Nahavandchi, H., & Motagh, M. (2018). Persistent Scatterer Analysis Using Dual-Polarization Sentinel-1 Data: Contribution from VH Channel. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(9), 3105–3112. https://doi.org/10.1109/JSTARS.2018.2848111

Sharifi, A. (2020). Flood Mapping Using Relevance Vector Machine and SAR Data: A Case Study from Aqqala, Iran. *Journal of the Indian Society of Remote Sensing*, *48*(9), 1289–1296. https://doi.org/10.1007/s12524-020-01155-y

Singh, K. V., Setia, R., Sahoo, S., Prasad, A., & Pateriya, B. (2015). Evaluation of NDWI and MNDWI for assessment of waterlogging by integrating digital elevation model and groundwater level. *Geocarto International*, *30*(6), 650–661. https://doi.org/10.1080/10106049.2014.965757

Solovey, T. (2020). Flooded wetlands mapping from sentinel-2 imagery with spectral water index: A case study of kampinos national park in central Poland. *Geological Quarterly*, *64*(2), 492–505. https://doi.org/10.7306/gq.1509

Song, Y., Lee, H., Kang, D., Kim, B., & Park, M. (2022). A Study on the Determination Methods of Monitoring Point for Inundation Damage in Urban Area Using UAV and Hydrological Modeling. *Water (Switzerland)*, *14*(7). https://doi.org/10.3390/w14071117

Spaur, M., Lombard, M. A., Ayotte, J. D., Harvey, D. E., Bostick, B. C., Chillrud, S. N., Navas-Acien, A., & Nigra, A. E. (2021). Associations between private well water and community water supply arsenic concentrations in the conterminous United States. *Science of the Total Environment*, *787*, 147555. https://doi.org/10.1016/j.scitotenv.2021.147555

Sresto, M. A., Siddika, S., Haque, M. N., & Saroar, M. (2021). Groundwater vulnerability assessment in Khulna district of Bangladesh by integrating fuzzy algorithm and DRASTIC (DRASTIC-L) model. *Modeling Earth Systems and Environment*, *8*(3), 3143–3157. https://doi.org/10.1007/s40808-021-01270-w

Stuart, M., & Lapworth, D. (2013). Emerging organic contaminants in groundwater. In *Smart Sensors, Measurement and Instrumentation* (Vol. 4, pp. 259–284). Springer International Publishing. https://doi.org/10.1007/978-3-642-37006-9_12

Stuart, M., Lapworth, D., Crane, E., & Hart, A. (2012). Review of risk from potential emerging contaminants in UK groundwater. In *Science of the Total Environment* (Vol. 416, pp. 1–21). https://doi.org/10.1016/j.scitotenv.2011.11.072

Sun, Y., & Li, X. M. (2021). Denoising Sentinel-1 Extra-Wide Mode Cross-Polarization Images over Sea Ice. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(3), 2116–2131. https://doi.org/10.1109/TGRS.2020.3005831

Tadono, T., Ishida, H., Oda, F., Naito, S., Minakawa, K., & Iwamoto, H. (2014). Precise Global

DEM Generation by ALOS PRISM. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II–4*, 71–76. https://doi.org/10.5194/isprsannals-ii-4-71-2014

Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022). Confidence interval for micro-averaged F 1 and macro-averaged F 1 scores. *Applied Intelligence*, *52*(5), 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Takaku, J., Tadono, T., & Tsutsui, K. (2014). Generation of high resolution global DSM from ALOS PRISM. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *40*(4), 243–248. https://doi.org/10.5194/isprsarchives-XL-4-243-2014

Tayfur, G., Singh, V. P., Moramarco, T., & Barbetta, S. (2018). Flood hydrograph prediction using machine learning methods. *Water (Switzerland)*, *10*(8). https://doi.org/10.3390/w10080968

Tehrany, M. S., Jones, S., & Shabani, F. (2019). Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques. *Catena*, *175*(December 2018), 174–192. https://doi.org/10.1016/j.catena.2018.12.011

Teng, Y. G., Zuo, R., Xiong, Y., Wu, J., Zhai, Y. Z., & Su, J. (2019). Risk assessment framework for nitrate contamination in groundwater for regional management. *Science of the Total Environment*, *697*, 134102. https://doi.org/10.1016/j.scitotenv.2019.134102

Tingsanchali, T. (2012). Urban flood disaster management. *Procedia Engineering*, *32*, 25–37. https://doi.org/10.1016/j.proeng.2012.01.1233

Tuo, T., Fang, Z., & Yue, P. (2022). Flood Mapping from Sentinel-1 Imagery Using Index Composition and HSI Transformation. *2022 10th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2022*. https://doi.org/10.1109/Agro-Geoinformatics55649.2022.9858964

Twele, A., Cao, W., Plank, S., & Martinis, S. (2016). Sentinel-1-based flood mapping: a fully automated processing chain. *International Journal of Remote Sensing*, *37*(13), 2990–3004. https://doi.org/10.1080/01431161.2016.1192304

U.S. Census Bureau. (2020). *TIGER/Line Shapefiles*. U.S. Census Bureau. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

U.S. Census Bureau. (2022). *Selected housing characteristics, 2018-2022 American Community Survey 5-year estimates*.

United Nations. (2016). Human Rights to Water and Sanitation | UN-Water. *UN Water*. https://www.unwater.org/water-facts/human-rights-water-and-sanitation

US Census. (2021). *U.S. Census Bureau QuickFacts: Alabama*. US Census. https://www.census.gov/quickfacts/fact/table/AL/BZA210220

US EPA. (2019). Drinking Water Regulations | US EPA. In *EPA "United States Environmental protection Agency'protection Agency."* https://www.epa.gov/dwreginfo/drinking-water-regulations

US EPA. (2023). *Acetonitrile Petition | Toxics Release Inventory (TRI) Program | US EPA*. https://www.epa.gov/toxics-release-inventory-tri-program

USDA. (2020). *Cropland Data Layer*. USDA NASS. https://nassgeodata.gmu.edu/CropScape/

van Leeuwen, B., Tobak, Z., Kovács, F., & Sipos, G. (2017). Towards a continuous inland excess water flood monitoring system based on remote sensing data. *Journal of Environmental Geography*, *10*(3–4), 9–15. https://doi.org/10.1515/jengeo-2017-0008

Vassiliou, A. A., Boulianne, M., & Blais, J. A. R. (1988). On the Application of Averaging Median Filters in Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *26*(6), 832–838. https://doi.org/10.1109/36.7714

Ver Hoef, J. M., & Temesgen, H. (2013). A Comparison of the Spatial Linear Model to Nearest Neighbor (k-NN) Methods for Forestry Applications. In *PLoS ONE* (Vol. 8, Issue 3, p. e59129). Public Library of Science. https://doi.org/10.1371/journal.pone.0059129

Wanders, N., Karssenberg, D., De Roo, A., De Jong, S. M., & Bierkens, M. F. P. (2014). The suitability of remotely sensed soil moisture for improving operational flood forecasting. *Hydrology and Earth System Sciences*, *18*(6), 2343–2357. https://doi.org/10.5194/hess-18-2343-2014

Watson, K. B., Ricketts, T., Galford, G., Polasky, S., & O'Niel-Dunne, J. (2016). Quantifying flood mitigation services: The economic value of Otter Creek wetlands and floodplains to Middlebury, VT. *Ecological Economics*, *130*, 16–24. https://doi.org/10.1016/j.ecolecon.2016.05.015

Wheeler, D. C., Nolan, B. T., Flory, A. R., DellaValle, C. T., & Ward, M. H. (2015). Modeling groundwater nitrate concentrations in private wells in Iowa. *Science of the Total Environment*, *536*, 481–488. https://doi.org/10.1016/j.scitotenv.2015.07.080

Xue, F., Gao, W., Yin, C., Chen, X., Xia, Z., Lv, Y., Zhou, Y., & Wang, M. (2022). Flood Monitoring by Integrating Normalized Difference Flood Index and Probability Distribution of Water Bodies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 4170–4179. https://doi.org/10.1109/JSTARS.2022.3176388

Zhu, Q. (2020). On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters*, *136*, 71–80. https://doi.org/10.1016/j.patrec.2020.03.030

Appendix I

ANOVA and MANOVA result

Analysis of variance (ANOVA) was conducted to evaluate the differences among clusters for each of the studied variables (Table 5). The ANOVA results revealed significant differences among the clusters for all variables (all p-values < 0.05), indicating that the cluster means varied more than could be expected by chance. Multivariate analysis of variance (MANOVA) was also employed to assess the multidimensional means of the clusters. The combined variables exhibited a significant Wilks' Lambda indicating that the clusters have distinct multivariate means.

ANOVA and MANOVA result

Test	Variable	F-Statistic	P-Value
ANOVA	WellUse	676.98	<0.01
	Poverty	229.52	<0.01
	Education	136.48	<0.01
	Minority	310.18	<0.01
	ChildPop	61.99	<0.01
	TRI	340.08	<0.01
	AgriArea	59.80	<0.01
	FloodRisk	283.38	<0.01
MANOVA	Combined All	82.1898	<0.01

Tukey’s HSD Result

The multivariate analysis through Tukey's HSD test (Table 6) shows the pairwise patterns of difference among the clusters with respect to several key variables. For TRI, several comparisons show significant positive or negative differences. For instance, the comparison between clusters 3 and 1 shows a significant positive difference of 0.2639, while clusters 4 and 1 also show a significant positive difference of 0.0525. These differences suggest that the clusters differ considerably in terms of toxic release levels, with some clusters having higher or lower levels of toxicity than others. Regarding AgriArea, most comparisons show significant differences, both positive and negative. For example, the comparison between clusters 2 and 1 shows a positive

difference of 0.0885, while the difference between clusters 3 and 1 is 0.0580. These differences indicate variations in agricultural area coverage among the clusters, highlighting that some areas have more agricultural land than others.

The FloodRisk variable also shows significant differences across clusters. The comparison between clusters 2 and 1 shows a significant negative difference of -0.2391, indicating that cluster 2 has a significantly lower flood risk compared to cluster 1. Similarly, the comparison between clusters 3 and 1 shows a negative difference of -0.1991, pointing to variations in flood risk levels across different clusters. For Minority, all comparisons exhibit significant differences, suggesting that the proportion of minority populations varies considerably among the clusters. For example, the difference between clusters 2 and 1 is -0.2418, while the difference between clusters 4 and 1 is -0.3640. These results indicate that some clusters have higher concentrations of minority populations compared to others.

Education levels also show significant differences, with several comparisons indicating variations in educational attainment across clusters. The comparison between clusters 3 and 1 shows a significant negative difference of -0.1840, while clusters 4 and 1 show a difference of -0.1053. These differences suggest disparities in education levels among the clusters, with some clusters having higher or lower educational attainment. ChildPop, or the population of children, exhibits significant differences in several cluster comparisons. For instance, the comparison between clusters 2 and 1 shows a significant negative difference of -0.0630, indicating that cluster 2 has a lower population of children compared to cluster 1. Other comparisons, such as clusters 3 and 1, show a significant positive difference, highlighting the variations in child populations across clusters.

Poverty levels also vary significantly across clusters. The comparison between clusters 2 and 1 shows a significant negative difference of -0.0682, while the difference between clusters 3 and 1 is -0.2043. These differences indicate that poverty levels differ among the clusters, with some clusters experiencing higher or lower poverty rates. Finally, WellUse, representing well water use, shows significant differences in several comparisons. The comparison between clusters 3 and 1 shows a significant negative difference of -0.1746, indicating lower well use in cluster 3 compared to cluster 1. Other comparisons, such as clusters 5 and 1, show similar trends, suggesting variations in well water use across different clusters.

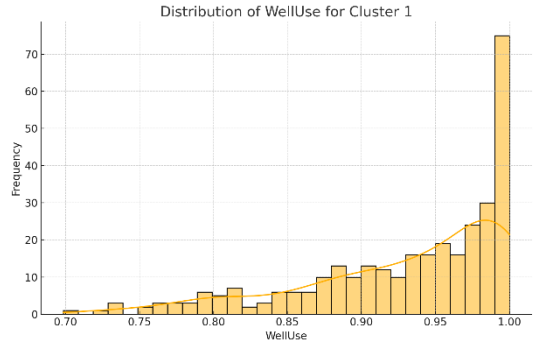
Variable wise Tukey's HSD test result for each cluster pair

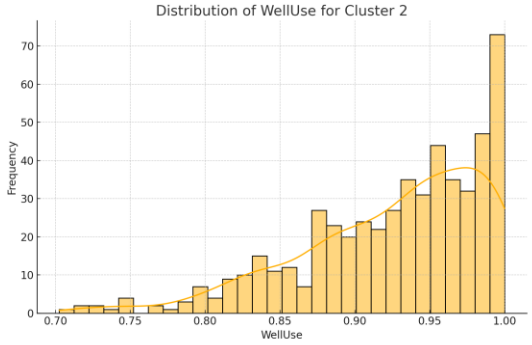
Cluster Pair	TRI	AgriArea	FloodRisk	Minority	Education	Child Pop	Poverty	WellUse
2-1	-0.0199-	0.0885***	-0.2391***	-0.2418***	-0.0057-	-0.0630***	-0.0682***	-0.0045-
3-1	0.2639***	0.0580***	-0.1991***	-0.1786***	-0.1840***	0.0577***	-0.2043***	-0.1746***
4-1	0.0525***	0.1696***	-0.1325***	-0.3640***	-0.1053***	0.0510***	-0.1621***	-0.0003-
5-1	0.1832***	0.0700***	-0.1600***	0.1014***	-0.0144-	0.0159-	0.1068***	-0.1773***
3-2	0.2837***	-0.0306*	0.0400***	0.0632***	-0.1783***	0.1206***	-0.1361***	-0.1701***
4-2	0.0723***	0.0810***	0.1066***	-0.1222***	-0.0995***	0.1140***	-0.0939***	0.0042-
5-2	0.2031***	-0.0186-	0.0791***	0.3432***	-0.0086-	0.0789***	0.1750***	-0.1728***
4-3	-0.2114***	0.1116***	0.0666***	-0.1854***	0.0788***	-0.0067-	0.0422***	0.1743***

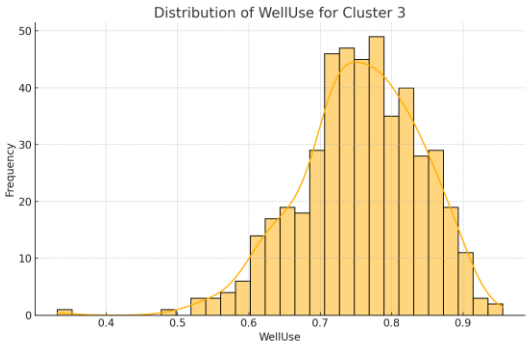
5-3	-0.0806***	0.0120-	0.0391***	0.2800***	0.1697***	-0.0417*	0.3111***	-0.0027-
5-4	0.1308***	-0.0996***	-0.0275*	0.4653***	0.0909***	-0.0351*	0.2689***	-0.1770***
Significance levels are indicated as: *<.05, **<.01, ***<.001. "-" indicates a non-significant result.								

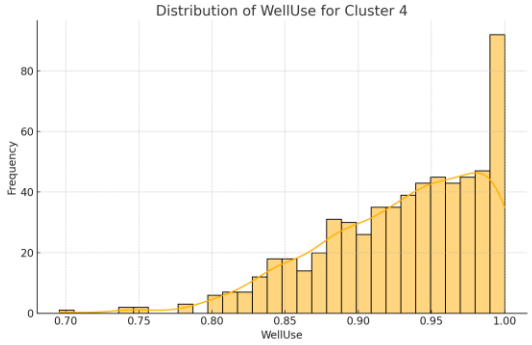
Distribution of well use rate in the cluster

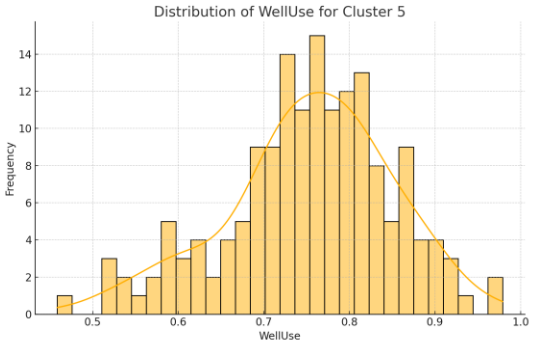
The cluster-wise distribution of well use rate suggests that Clusters 1, 2, and 4 have captured the highest count of block groups with a high well use rate. These clusters have more than 70 block groups with a well use rate close to 1. Clusters 3 and 5 have a high frequency of block groups with a well use rate ranging from 0.7 to 0.8; however, the count is less than 50 block groups for Cluster 3 and around 14 block groups for Cluster 5.


Distribution of WellUse for Cluster 1
Frequency
WellUse


Distribution of WellUse for Cluster 2
Frequency
WellUse


Distribution of WellUse for Cluster 3
Frequency
WellUse


Distribution of WellUse for Cluster 4
Frequency
WellUse


Distribution of WellUse for Cluster 5
Frequency
WellUse

Multicollinearity test result

Table 7 presents the Variance Inflation Factor (VIF) for each independent variable. A VIF value greater than 10 indicates significant multicollinearity. In this case, all variables have VIF values well below 10, suggesting no significant multicollinearity issues among the variables.

Multicollinearity test result using VIF

Variable	**VIF**
Minority	1.198648
Education	1.232746
ChildPop	1.068141
Poverty	1.25715
TRI	1.0876
AgriArea	1.02944
FloodRisk	1.087167