Final Insights Report

Intelligent Data Ecosystem for Assam - Flood Response and Management (IDEA-FRM)

Under **Open Contracting India Initiative**

By **CivicDataLab** in Collaboration with **Open-Contracting Partnership**

With the support of **Patrick J. McGovern Foundation**

open contracting india

OPEN CONTRACTING PARTNERSHIP

civic data lab

Patrick J McGovern FOUNDATION

Authors and Project Team:

Bianca Shah, Gaurav Godhwani, Jeeno George, Kabeer Arora , Liz Mariya Jacob, Mohak Sharda, Phanidatta Surampudi, Sai Krishna Dammalapati, Shreya Agarwal, Bernadine Fernz, Camila Salazar

About Us:

Open Contracting Partnership: A silo-busting global organisation working across governments, businesses, civil society and technologists to open up and transform public procurement worldwide. Bringing open data and open government together, we make sure public money is spent openly, fairly and effectively on public contracts, the single biggest item of spending by most governments. Using meaningful participatory approaches, we help our partners think through how to make the whole contracting process simple, accessible and inclusive. Over 50 countries, states and cities are already working with us to implement open contracting best practices to deliver better outcomes for people, planet and prosperity.

CivicDataLab: Founded in 2018, **CivicDataLab (CDL)** is a private research lab working at the intersection of data, tech, design and social science to strengthen access to public information and improve citizen participation in governance in India. We work closely with governments, non-profits, think-tanks, media houses, universities, and other actors to grow their data and tech capacity to enable data-driven decision-making at scale. We have harnessed the potential of open knowledge movements to co-create Open Budgets India, Justice Hub, Open Contracting India, Open City and more with 20,000+ public interest datasets and an active user base of more than half a million citizens.

About Open Contracting India initiative

Open Contracting India is an initiative by CivicDataLab and the Open Contracting Partnership aimed at making public procurement processes in India more efficient, effective and inclusive.

We do this by onboarding governments to international best practice for curating and releasing data including open data standards, developing tools and platforms to analyse data to inform decision-making, creating relatable case studies and resource material to promote improved performance and innovative reforms.

Contents

Executive summary

Extreme weather events are arguably the most visible manifestation of the climate change crisis. According to the World Meteorological Organisation (WMO) report 2021, the world has experienced a weather or climate-related disaster almost every day over the past 50 years, leading to daily losses amounting to over USD 200 million.

The Sendai Framework for Disaster Risk Reduction 2015-2030 and the New Urban Agenda (for local governments) places special emphasis on countries and cities to commit to mainstreaming holistic and data-informed disaster risk reduction strategies and policies to reduce vulnerabilities and improve climate resilience.

However, in India, the necessary good quality, machine-readable and interoperable data to inform this strategy and to correctly plan for flood response and relief is currently unavailable. Data that could enable more effective disaster-risk response and management is scattered or siloed across different agencies, at different scales and in different formats, making it difficult for the decision-makers and relevant stakeholders to make data-informed decisions. This often results in inefficient processes and policies or ad-hoc responses that fail to adequately cater to urgent, often life-saving needs in times of emergency.

Assam, a state of north-east India, is one of the most flood prone regions in the world where annual floods acutely affect 40% of the state. The 2022 floods caused over 190 deaths and affected over 8 million people (as per data compiled from ASDMA's daily FRIMS reports). Despite multiple initiatives by state bodies, the comprehensive, usable information needed is amiss. Through our project, '**Intelligent Data Ecosystem in Assam - Flood Response and Management (IDEA-FRM)**', we aimed to:

1. identify, collate and process all flood related information in AI ready format onto a single point; and
2. develop advanced data models to identify most vulnerable regions and their preparedness levels.

Our overarching objective was to make a functional data model that can ultimately be adopted by the Assam State Disaster Management Authority to better respond to and prepare for floods.

Following a comprehensive literature review, field study and stakeholder interviews we were able to identify relevant flood related data in five broad categories, namely:

1. **Satellite and weather data**: This data helps us understand floods as a function of various natural factors like rainfall trends, distance to rivers, elevation, slope, drainage density, vegetation density, built density, soil and lithology;
2. **Demographic Data:** This data helps us understand how floods interact with settlements and determine the impact on human lives and livelihoods looking at various social and economic factors;

3. **Access to infrastructure**: This helps us understand the vulnerability of regions as a function of infrastructure access to cope with floods;
4. **Damages**: This data helps us understand the trends of flood impacts in the regions historically; and
5. **Government Response**: And finally we need this dataset to understand how the government has responded to floods and where the gaps might be. We use finance/ spending data as seen through public procurement.

Figure 1 & 2: The office of District Disaster Management Agency

Fig 3: Meeting with 30 village heads in Darrang District

This work was not easy. Apart from identifying the different data sources and standardising the data, a major challenge in preparing the dataset was geocoding the procurement data (which does not have a field indicating the location of work). We had to apply open-source text mining algorithms across multiple fields to identify the work location and then validate them. Additionally, the damages data were available as daily reports in pdf format and needed to be converted to digital, tabular format for analysis. Another challenge was in working across changing numbers of districts. When the census was done In 2011, Assam had 27 districts which increased to 33 by 2018 and 35 in 2022. There was an enormous amount of work done on the dataset to seamlessly reflect this change across the years of study. In fact, all the raw data from the five identified sources were cleaned, formatted and transformed for ingestion into the analytics platform. This process includes consolidating or in certain cases, separating fields and columns, changing formats, assigning unique identifiers, deleting unnecessary data and making corrections to data. This data is now available to the public on Github - IDEA-FRM Repository.

Then, we experimented with two approaches:

The first approach demonstrates flood risk assessment that integrates a machine learning model to predict probability of flood occurrences. Whereas in ***the second approach***, we employ a statistical multivariate model to assess the preparedness to floods by combining all the datasets to identify the places which need to be better prepared to face floods.

Conventionally, the output from the flood prediction model is the input for calculating the flood risk to populations, enabling decision makers to be better prepared to manage and respond to floods. Even though the first approach gave us a relatively accurate prediction model, it also relies on validation of the predictions by third parties or by using the dataset ‘Damages’ and also does not use the dataset 'Government Response’. It gives a prospective vision rather than synthesising retrospective learnings from the past.

Additionally, although the first approach does predict probability and then risk, it does not establish the relationship between parameters which contribute to how severity of risk is determined. This relationship between the variables is established by the second method, which uses a multivariate statistical model. Furthermore, the projected results correlate with the actual data. We were able to identify revenue circles that have high flood impact but low response from the government.

The results from the first approach highlight that revenue circles with high flood impact are often those that have better access to infrastructure, leading to increased infrastructure damage from floods which is often cited as a cause for detrimental impact on recovery. Damages to infrastructure widens the impact of flooding beyond the immediate area to the surrounding regions. This analysis helps inform decision-makers on where and how to channel infrastructure preparedness and response activities. The model can be further fine-tuned to improve understanding on what infrastructure to build and where, channelling the right kind of investments in most needed places.

The second approach successfully identifies the relationship between variables that determine the preparedness of the revenue circles[1] to floods. We identify the top 5 out of 35 districts in Assam where preparation with respect to vulnerability was low.

The model has wide applicability and potential for scaling up, not only in other states of India but also to other countries. It can also be zoomed into the panchayat, village or ward levels.[2] We hope to expand this research and close the flood response and management data gap wherever the opportunity presents. Furthemore, the model is robust enough to incorporate other subjective and qualitative parameters for a comprehensive score of preparedness.

[1] As of 2022, Assam is administratively divided into 35 districts with 184 revenue circles. Bajali and Tamalpur are the newly added districts. Each district is subdivided into revenue circles for collection of taxes. NB:. We are yet to receive the updated official revenue circle map and hence the map we use for the projects shows only 180 revenue circles.

[2] The districts are divided into subdivisions for administrative purposes and revenue circles for tax collections. The subdivisions are further subdivided into development blocks for development purposes and Tehsils for revenue purposes. The development blocks are composed of gram panchayats. The gram panchayats are made of a large village or a cluster of smaller villages. The village is the lowest level of subdivisions in India.

Introduction

In India, floods account for more than 50% of all climate-related disasters and cause INR4.69 trillion/US$64 billion in economic losses. Assam, a state in north east India, has historically been prone to annual flooding. These yearly floods acutely affect the state of Assam, impacting over 100,000 people and killing hundreds of people, livestock and animals in three to four waves yearly. However, in the last few years, we have seen the increasing impact and unpredictability in its patterns, owing to changes in human settlement and natural factors like climate change. In 2022 alone, Assam recorded deaths of 197 people from floods that affected 5 million people.

The Sendai Framework for Disaster Risk Reduction is a framework for combating disaster risk, recognising the potential of data and technology in combating disasters. However, correct and relevant data is often inaccessible promptly.

Through the support of The Patrick J. McGovern Foundation's (PJMF) 2022 Accelerator Grant Program, the Open Contracting Partnership (OCP) and CivicDataLab (CDL) are collaborating to develop an intelligent data model combining fiscal, geospatial and demographic data to improve the equitable distribution of flood preparedness and relief infrastructure, goods and services in Assam, India.

Through this project, **'Intelligent Data Ecosystem for Assam- Flood Response and Management (IDEA-FRM)',** we focus on identifying and collating the relevant data related to Assam floods, making them interoperable and publishing them in open access for better collective analyses and innovative solutions. We have published these AI-ready datasets on GitHub with the hope that other stakeholders can collaborate and co-create data-driven solutions to flood crises in Assam. Additionally, we use this data to develop a decision-making framework for authorities to ensure resource allocation to the most vulnerable areas to increase the preparedness levels of districts in Assam. We continue to collect real-time feedback from the decision-makers and all concerned stakeholders to make it more exhaustive and more accurate over the coming months.

Hypothesis

Good data is fundamental for creating effective tools to manage disasters and reduce risks, a responsibility distributed across multiple actors and government agencies in India. Data on how much is spent and in which communities, is critical to understanding whether the intended investments reach the areas most affected by floods. However, such data does not yet exist in a form that can be accessed,

analysed, and used to inform decision-making by responsible government agencies. It is siloed and scattered across different online and offline platforms.

For a complete understanding of flood issues, it is necessary to examine multiple datasets concurrently. Our core hypothesis is that one can combine these datasets to provide insights into whether the government's response to floods is being focussed in places with the greatest need in a timely way and, in turn, guide future actions. The proposed method ensures the interoperability of freely available datasets for evidence-based decisions.

Approach

In this project, we aim to plug these gaps by linking different datasets-geospatial and satellite data, with socio-economic indicators and fiscal data for insights into the effectiveness of past actions and future needs for flood response and mitigation.
We make use of a framework (Figure 1) that unifies the variables in the datasets for:

Approach 1 : Flood Risk Assessment - This combines machine learning for flood prediction with datasets that indicate socio-economic characteristics of the population to assess the flood risk.

Approach 2 : Flood Preparedness Model - This approach applies structural equation modelling (SEM) to model the relationship between the various variables of flood conditioning factors, demographic vulnerability, access to infrastructure and government response to estimate the flood preparedness index.

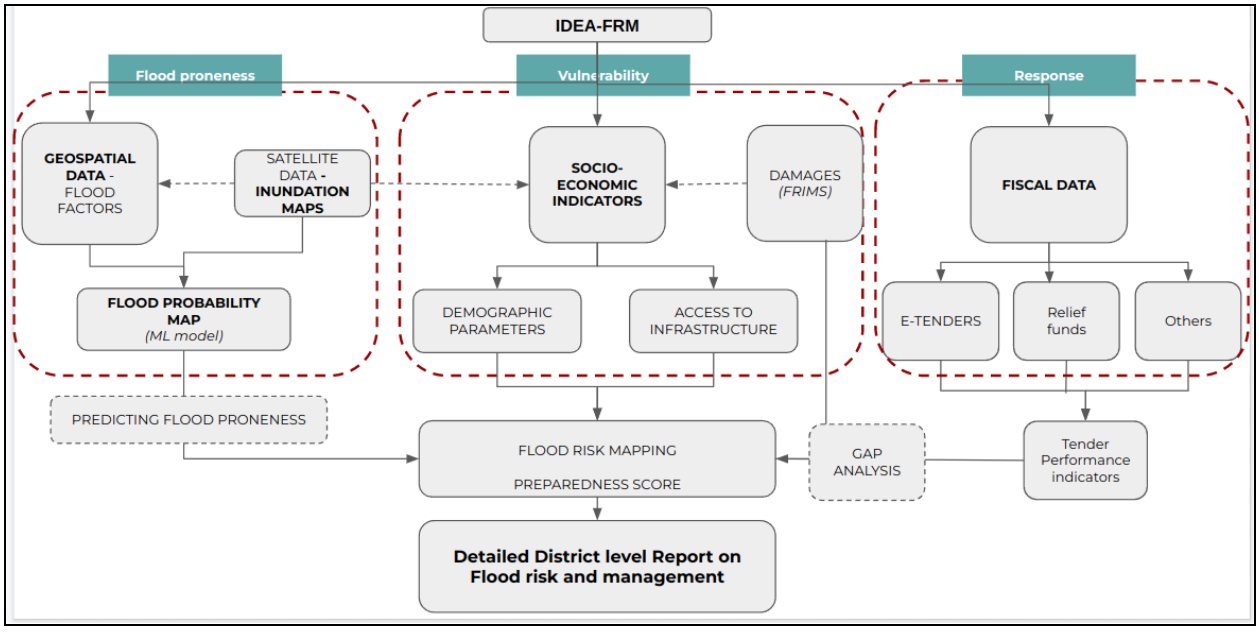


Figure 1: Framework of IDEA-FRM

The steps involved in the approach are as follows:

1. **Project and Data Scoping** - Our starting point was to understand the context of flood management in Assam and the data generated/used. We did this through expert interviews and panel discussions. This helped us to enrich our understanding of the context, testing of our hypothesis and established the need for a data model. It also helped us to select variables for the model so that it reflects the on-ground realities and needs. All of this laid the foundations for our project. We undertook a detailed **data scoping** exercise firstly based on desk study and then refined it on the basis of our understanding from the field research and interviews with state and non-state stakeholders.

2. **Data Preparation** - While our tech team mined data available online, our researchers simultaneously carried out field visits in Assam, meeting with government offices to onboard them onto the project and get important understanding of the way data is collected. In parallel to collecting data, we also started processing data to get it ready for ingesting into the data model. This entailed collating, cleaning and geocoding the socio-economic and fiscal datasets.

3. **Model Building Phase 1 -** As we started curating the data, we experimented with different methodologies for the first approach, that was a flood prediction probability model. The predicted flood probability map was then combined with socio-economic indicators to prepare flood risk maps. After that, we validated the effectiveness of flood-related procurements in addressing flood damages, ranking the vulnerability of administrative units against support/resources received to lessen flood impacts. Then we moved to the second approach where we used structural equation modelling to evaluate the relationship between the different independent and dependent variables that affect the preparedness of any area to floods. The smallest scale of work for the second approach is the revenue circle (the aerial extent ranges from 5 sq km to 3000 sq km), while a grid of 30mx30m is used for the first approach.

4. **Model Optimization** - Once the methodology was finalised, the results were validated with actual data (where present). We also undertook detailed discussions with stakeholders and/or decision-makers to validate the findings.

5. **Model Building Phase 2** - With the feedback on the initial outcomes, we revised the models to address the issues and then scaled them.

6. **Model Deployment** - Finally, we demonstrated the approaches to decision-makers and who are very keen to use them for actual flood management purposes in Assam. We will be moving the model to their servers by March 2023.

Datasets

The dataset collated falls under five broad categories and has a total of 72 variables:

1. **Satellite and weather data on flood proneness**: Flood is a function of various conditioning and triggering factors such as past rainfall trends, distance to rivers, elevation, slope, drainage density, vegetation density, built density, soil and lithology. Additionally, inundation maps are collected to analyse the actual flooded areas and validate results. Lastly, we need weather forecasts for the flood prediction models. These data are found through SRTM, Bhuvan , meteorological data, and weather forecasts. A detailed metadata is available here: Satellite and weather data
2. **Demographic Data:** How floods interact with settlements determines the impact on human lives and livelihoods. To understand this we need to see the demographic vulnerability which can be assessed through studying the population, sex ratio, child and elderly population, deprived population, household access to drinking water and sanitation, etc. available from SECC, NFHS and State Statistical Handbook. A detailed metadata is available here: Demographic Data.
3. **Access to infrastructure**: Access to infrastructure like hospitals, road networks, embankments etc largely influence response capacity of a region to floods and hence this becomes critical in understanding preparedness levels. By facilitating the movement of people and goods, transportation networks play a crucial role in daily economic operations. During floods, transport infrastructure is crucial in rescue and evacuation operations. Absence of transportation infrastructure increases the vulnerability of the population exposed to the extreme weather event. Infrastructure Access is measured by calculating proximity of a revenue circle to roads, rails, embankments and hospitals. This data is sourced from public resources like GeoSadak of Pradhan Mantri Gram Sadak Yojana. A detailed metadata is available here: Infrastructure Access
4. **Past Damage**: If we understand what kind of damage and in which areas they have occurred during previous flood cycles, it can help us understand where and what kinds of interventions are needed to mitigate the worst impacts of future floods. This is reported by ASDMA on a daily basis during floods (May 15th to October 15th) using Flood Reporting and Information Management System (FRIMS). Data available on the FRIMS portal is ingested and geocoded at the revenue circle level to identify a number of people affected, crops affected, roads damaged etc., and is specific to each revenue circle. The datasets are made available here: past damages data
5. **Government Response**: How the government has responded to combat floods, before, during and after is among the most important things to understand before proposing any solution. We looked at it through an expenditure lens including government procurement data

(assamtenders.gov.in), budget data and relief distributed (as reported in FRIMS). Government tenders related to floods were isolated and geocoded at revenue circle level. We have identified 4,400 e-tenders related to floods that have been published by 30 different departments from April 2016 to June 2022. These projects have been funded by different sources, with the State Disaster Relief Funds (SDRF) taking the largest share of tenders by count, followed by the State Owned Priority Development (SOPD) scheme, then the Rural Infrastructure Development Fund (RIDF). In terms of total tender value, SOPD is highest followed by SDRF. The datasets are made available here: Government Response

The sources for the identified indicators are listed here. Using this dataset various statistical, ML and AI models are possible to derive insights and suggest innovative solutions to the issue of floods.

Approach 1 : Flood Risk Assessment

The first approach aims to use the dataset to feed into a machine learning (ML) model to predict the probability of flooding across the state of Assam and subsequently to assess the flood risk. Flood risk is the likelihood of a flood event occurring and its associated negative consequences, which may include various types of impacts (Schanze, 2006). Conventionally risk is expressed as the product of hazard and vulnerability (Matsimbe, 2003).

$$Risk = Hazard\ x\ Vulnerability$$

Flood risk is the perceived danger due to floods on population or infrastructure or a combination of both. It depends on the type of risk one is planning to access. Here the datasets on demography and access to infrastructure are weighted using the probability of being in a flood prone area to quantify their vulnerability to floods and overlaid with flood hazard to assess the overall flood risk.

For the ML model, the first step is to collate data regarding the inundation along with the conditioning and triggering factors of floods. For data on inundation, we scraped inundation maps from 2018 to 2022 from Bhuvan. The inundation map, which takes up 411.6 GB, is stored in the S3 bucket of the AWS server. (see Figure 2)

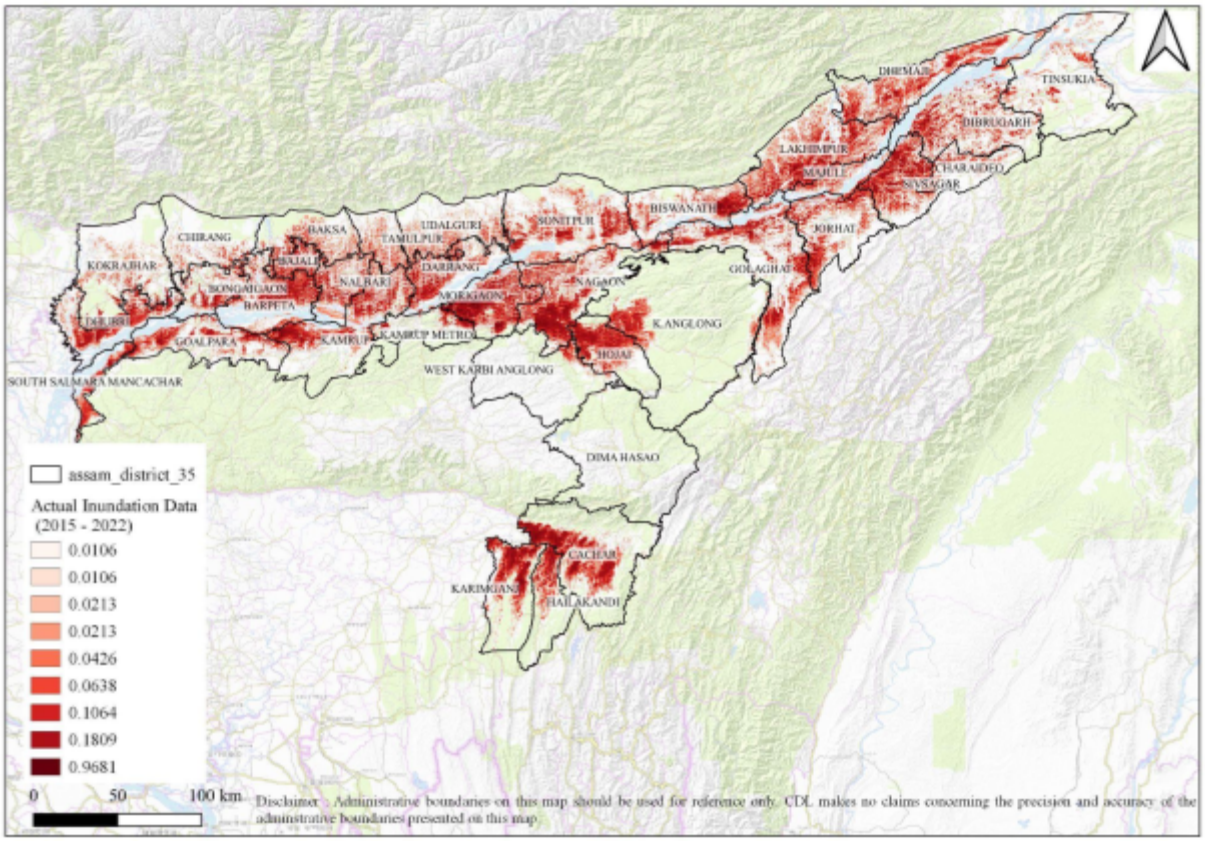


Figure 2 : Assam Inundation Map (2015-2022)

The conditioning factors selected are the distance from the river, drainage density, slope, and elevation, lithology, soil texture, and vegetation density. The triggering factor is the rainfall. The collected data was rasterized, wherever necessary. The conditioning and triggering factors were combined together using a method called multi-criteria decision analytics to get a flood-proneness map.

The conditioning factors are weighted using the Frequency Ratio (FR) model. The FR model is a bivariate statistical method and a simple geospatial assessment tool for understanding the probabilistic relationship between dependent and independent variables. The FR model is a useful method to assign weights to different classes of a map based on the interaction with flooded cells.

Methodology

The variables of the training data for the ML model comprises inundation maps along with the conditioning and triggering factors in the form of maps. The inundation maps are the response variables and the rest are the predictor variables. The predictor variables are :

1. Slope

2. Elevation
3. Distance from river
4. Drainage density
5. Surface Runoff (GCN)
6. Soil
7. Lithology
8. Land use
9. Built density
10. Vegetation density

Frequency ratio method is applied on all the predictor variables so as to assign weights to different classes of each variable based on the probability of being in a flooded area. This step also converts the categorical variables, such as soil, lithology and land use, to continuous variables.

These data are collected for five years from 2018 to 2022. A set of 10,000 random points are generated for each year across each district and applying the criteria that the minimum distance between points should be 100m. Next, data variables are extracted from the raster maps for all the points of the years to train the model.

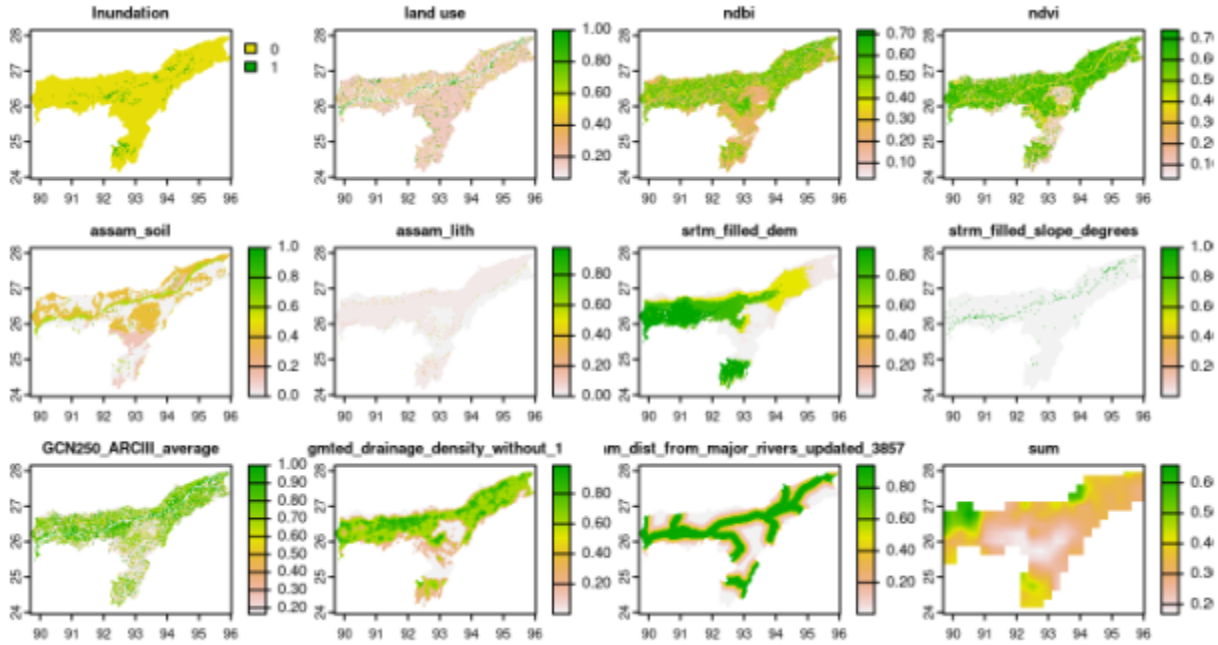


Figure 3: Predictor and Response variables for the ML model

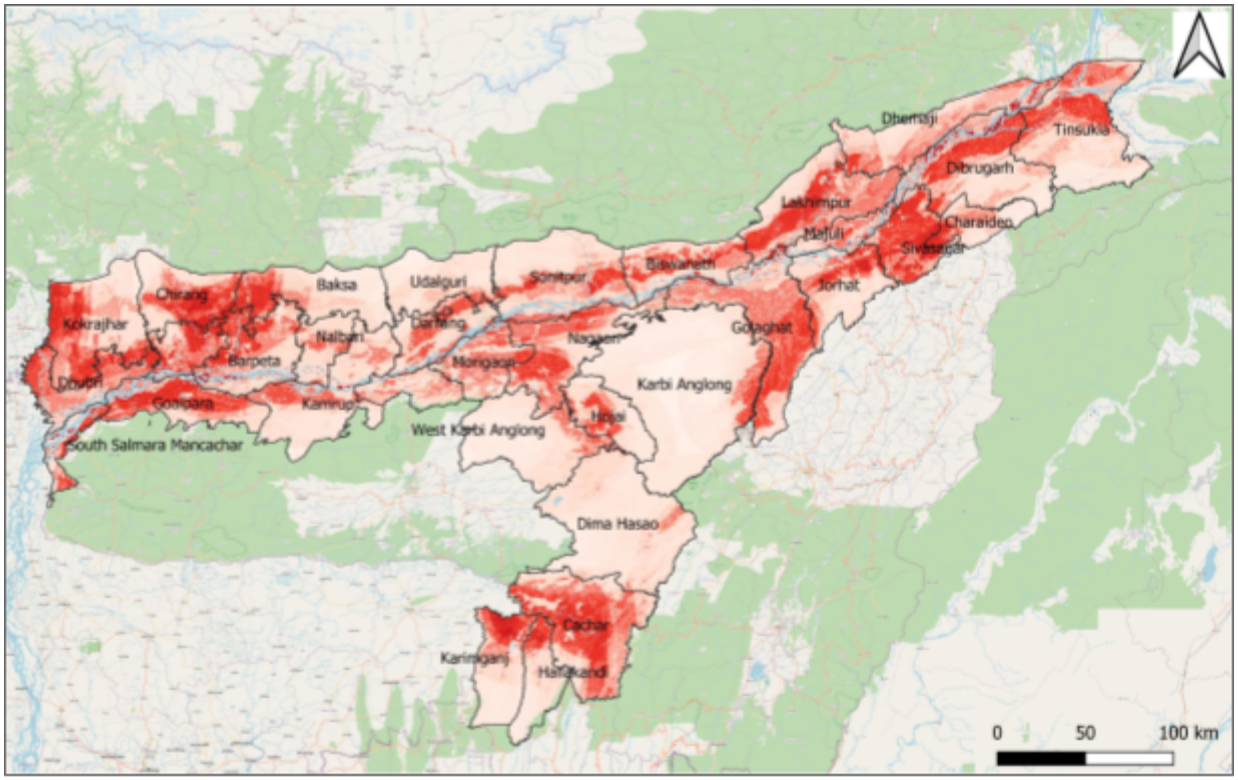
The NA values of the training and testing dataset are filled using random forest imputation and to eliminate the imbalance between non-inundated cells which are represented by '0' and inundated cells which are '1', we randomly select the non-inundated cells so that the number of inundated and non-inundated cells are similar.

Once the data preparation stages are over, we split the data into training and testing sets and tried out multiple ML methods. After comparing the model results, we decided to go forward with the Support Vector Machine (SVM) Machine Learning model. The validated trained model can be used to predict the flood probability for weather forecasts.

Upon calculating the probability of floods in different regions of the state, the next step is to identify the vulnerability of population and access to infrastructure. This is crucial to highlight the pockets of high risk. The vulnerability is estimated based on socio-economic factors such as demographic characteristics and access to infrastructure. Here, we again applied the FR model to weigh the classes of each factor based on the probability of being in a flood-prone area.

Outputs and Outcomes

We started by preparing a flood-proneness map using Multi-criteria Decision analytics and validated the results with the actual Flood Hazard Map - prepared using the collected inundation map from 2015 to 2022 (Figure 7). The method gives an accuracy of 86%, which further endorses the selected conditioning and triggering factors for its use in the machine learning model to predict the probability of flooding.



Disclaimer : Administrative boundaries on this map should be used for reference purposes only. CDL makes no claims concerning the precision and accuracy of the administrative boundaries presented on this map

Figure 4: Flood Proneness Map prepared using MCDA

Further, the SVM ML model is used to predict the flood proneness probability for each district. For Morigaon, the model shows a precision and accuracy of 0.42 and 0.80 respectively, while Kamrup has precision of 0.38 and accuracy of 0.80 (see Figure 5 and 6).

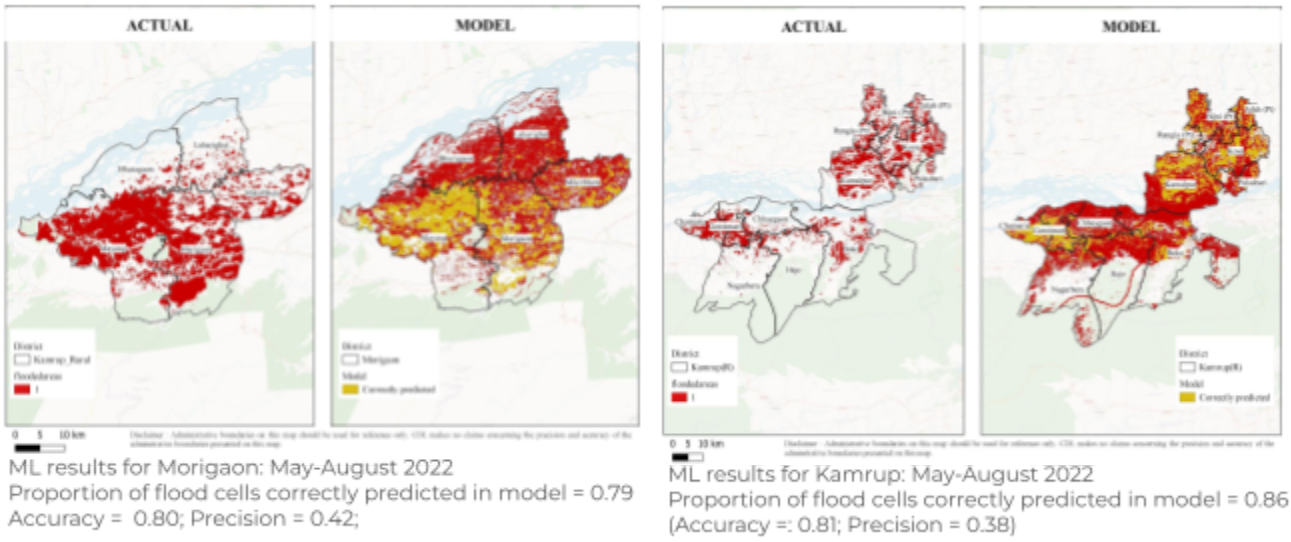
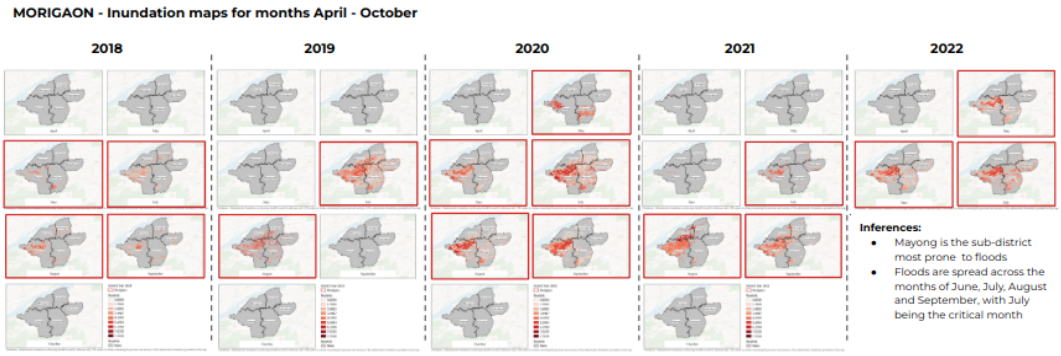


Figure 5 & 6: ML results for Morigaon & Kamrup

Additionally, the flood inundation data helped to identify the months that are more flood-prone and require additional attention. For example, in Morigaoan district, Mayong is the most flood prone area and July is the critical month.



Year	Revenue Circle	Overall Flood Vulnerability
2022	Mayong	16.64
2021	Bhuragaon	6.42
2020	Mayong	14.55
2019	Mayong	11.01
2018	Mayong	4.13

	2018			2019	2020	2021		2022
Month	June	July	August	July	July	July	September	June
Bhuragaon								
Laharighat								
Marigaon								
Mayong								
Mikirbheta								

Figure 7: Evaluating critical months and revenue circles of districts

The weighted demographic data and access to infrastructure are aggregated to get the vulnerability map (see Figure 8).

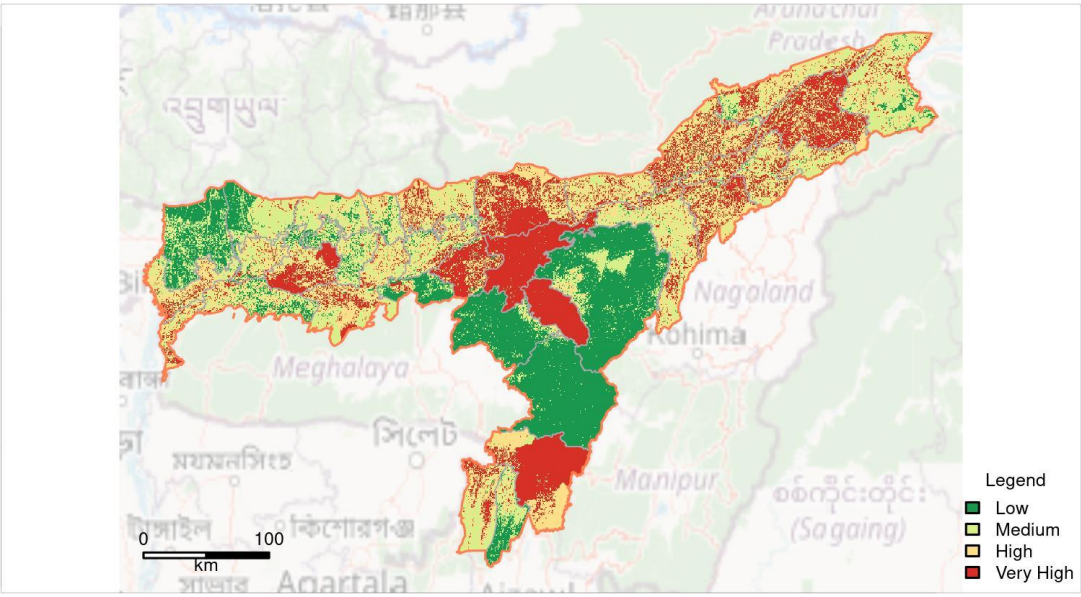
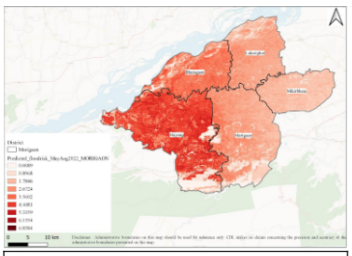


Figure 8: Vulnerability Map (2022) prepared using datasets - demographic and access to infrastructure

The product of the vulnerability map cropped to each district and the flood proneness map (actual or predicted) gives us the flood risk map. In Morigoan, we can see that the flood risk predicted for revenue circles in May-August 2022 corresponds to the actual flood impact in the revenue circles for the same period.



Actuals

Year	Revenue Circle	Risk
2022	Mayong	16.64
2021	Bhuragaon	6.42
2020	Mayong	14.55
2019	Mayong	11.013
2018	Mayong	4.13

Predicted risk map of May-Aug 2022

Revenue Circle	Risk	Year
Mayong	3.65	2022
Bhuragaon	2.97	2022
Laharighat	2.11	2022
Marigaon	2.04	2022
Mikirbheta	1.63	2022

Figure 9: Validating the flood risk of Morigaon estimated by aggregating vulnerability and ML generated flood predictions with actual flood damages for May-August 2022

Limitations

Conventionally, the actual or predicted flood inundation maps are combined with individual variables or combined variables of dataset 2 and 3 to determine the respective flood risk. Flood risk is the product of hazard and vulnerability. The vulnerability has to be defined in each case. For example, it can be the vulnerability of the population, vulnerability of infrastructure, vulnerability of populations due to lack of access to infrastructure, vulnerability of crops, overall vulnerability and so on. Flood risk is a popular measure to indicate the areas or attributes of any area that are at high risk from damage due to floods. Flood risk is a popular measure also because reliable and consistent data on flood damages are not available in most cases.

In our study we noticed that the flood damaged places are often not in the areas that are assessed as high risk. A possible explanation is that damage is a complex phenomenon resulting from the interactions of multitude of factors and hence can not be defined simply as a product of flood hazard and vulnerability. Another explanation stems from a limitation in the datasets. Firstly, flood hazard is measured as the number of times a particular area is flooded while the intensity or the flood water depth is not considered as the information is not recorded anywhere. Secondly, the demographic dataset is estimated from the last census in 2011 and may significantly vary from current on-ground realities. Moreover, the smallest scale at which damages are recorded is the revenue circle. Lastly, the infrastructure data is limited to roads, railway lines, hospitals and embankments while several other critical flood infrastructure such as schools that can turn into relief and distribution centres are excluded due to limitations in data availability.

Additionally, the flood risk assessment does not allow the further integration with other subjective and qualitative parameters that measures the coping capacity of the administrative circles, which may be villages, panchayat, blocks, revenue circle or districts (whichever scale the study is done), to improve flood resilience. Therefore, we experimented with a multivariate statistical model which can address these limitations to explain the relationship between the variables and can predict the latent variables.

Approach 2: Preparedness Score using Structural Equation Modelling

Structural Equation Modelling (SEM) is a multivariate, hypothesis-driven technique used to assess structural relationships. Structural relationships are the relations we hypothesise between the observable variables and unobservable (latent) variables. For instance, the “preparedness” of a revenue circle is unobservable – we cannot

directly measure preparedness. Hence, we use observable data of tenders, flood damages etc., to estimate this unobservable.

The dataset described in section 'Dataset' is used to measure 5 latent variables – Flood Proneness, Demographic Vulnerability, Infrastructure Access, Flood Impact and Preparedness. Data is captured from the years 2018 to 2022 for the modelling. Altogether, 72 variables are used to create the model.

Methodology

The following structural relationship is hypothesised for the model. We hypothesise that **"Flood Impact"** is understood directly by the reported flood damages on the FRIMS portal and indirectly by **flood proneness, demographic vulnerability** and **infrastructure access** of the region, which in turn are measured by the variables described in the above section. '**Preparedness'** is measured, under a measurement or outer model, using observed variables pertaining to money spent by the government through public procurement. Preparedness is regressed on four latent variables -

1. flood impact
2. flood proneness
3. demographic vulnerability
4. infrastructure access

under structural or inner models. We hypothesise the latter could directly or indirectly impact Preparedness.

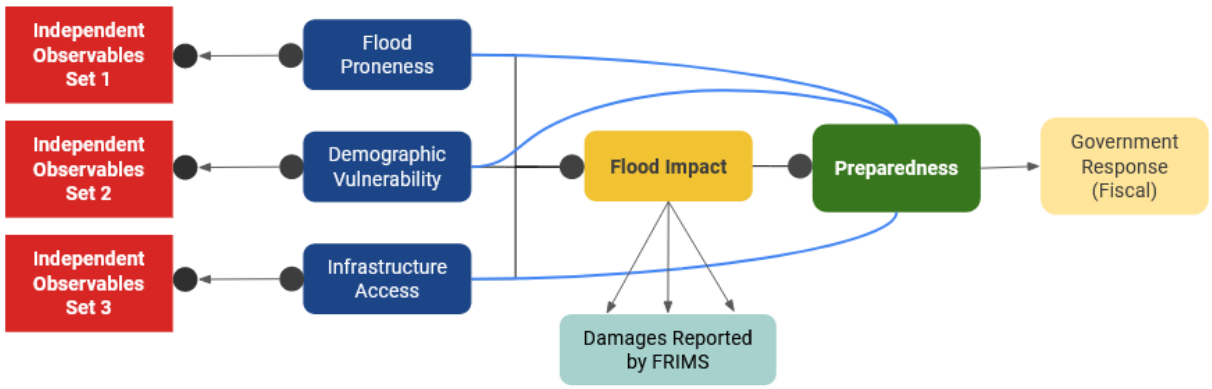


Figure 10: Schematic diagram for Structural Equation Model methodology

These hypotheses are tested during the SEM. Multivariate regression is done to fit the input data with the structural model we hypothesised.

The benefit of using an SEM over other multivariate techniques is that SEM gives us the opportunity to construct unobservable latent variables and predict their values. After the model is fit, we can predict values of each of the latent variables from the input data. While this can be done without SEM by manually giving weightages to each of the variables, SEM generates these weightages from the data and relationships between variables itselves. At the same time, SEM allows for more explainability between the variables than a prediction based Machine Learning model.

We validate the results of SEM comparing actual damages against the total tender value in the revenue circles. The detailed process involves normalising for all the years, the total damages and the total tender value across the revenue circles. The difference between the total tender value and the total damages is again normalised and compared against the damages. Thus, the greater the damage and the lower the efficiency score, the higher the susceptibility, while the greater the damage and higher the efficiency score, then the lower the susceptibility. These values are mapped and compared with the predictions from the SEM.

Outputs and Outcomes

The results from the SEM are:

Estimates: The SEM presents the regression estimates between each variable pair and also between the latent variables and the measured variables used to measure them. For instance, the latent variable **Preparedness is positively** associated with the Sum of Total Tender value in a revenue circle, followed by the count of tenders related to SOPD, SDRF and RIDF schemes. Which means the higher the amount of procurements made for a region, the better prepared it is. Among latent variables, Preparedness interacted with **flood_impact** and **infrastructure_access**, but not that strongly, indicating that while procurement happened in areas with more damages and infrastructure, other areas also would have seen significant procurement.

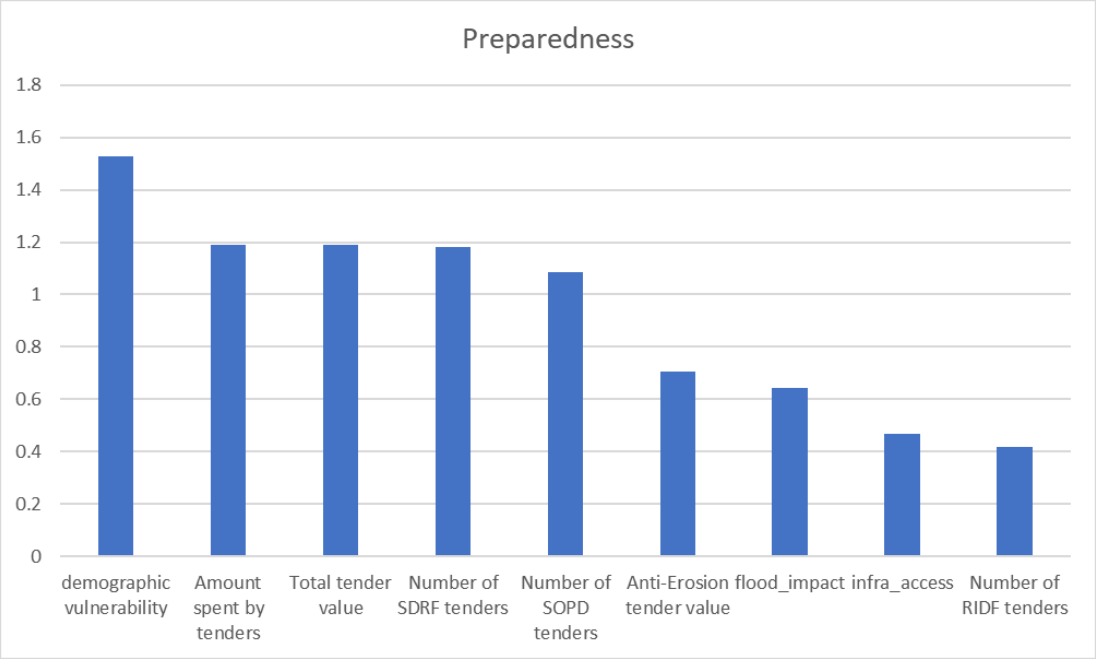


Table 1: Coefficients when Preparedness is regressed with other measured and latent variables.

Note: Reported only those variables that are statistically significant.

Factor Scores: With the model fitted above, we proceed to estimate the values of latent variables in each revenue circle. The values are then normalised and presented below.

Top ten revenue circles most impacted by floods:

district	revenue_circle	flood_impact	demographic_ vulnerability	flood_proneness	infra_access	preparedness
Barpeta	Barpeta	1.00	0.56	0.62	0.88	1.00
Barpeta	Sarthebari	0.89	0.36	0.55	0.91	0.22
Barpeta	Kalgachia	0.87	0.34	0.70	0.84	0.21
Bajali	Sarupeta	0.86	0.30	0.55	0.86	0.20
Barpeta	Chenga	0.79	0.30	0.84	0.74	0.21
Barpeta	Baghbor	0.72	0.53	0.73	0.73	0.34
Bajali	Bajali	0.70	0.15	0.49	0.92	0.24
Barpeta	Barnagar	0.60	0.27	0.58	0.86	0.21
Bajali	Jalah	0.54	0.02	0.50	0.94	0.18
Cachar	Silchar	0.38	0.71	0.39	0.79	0.45

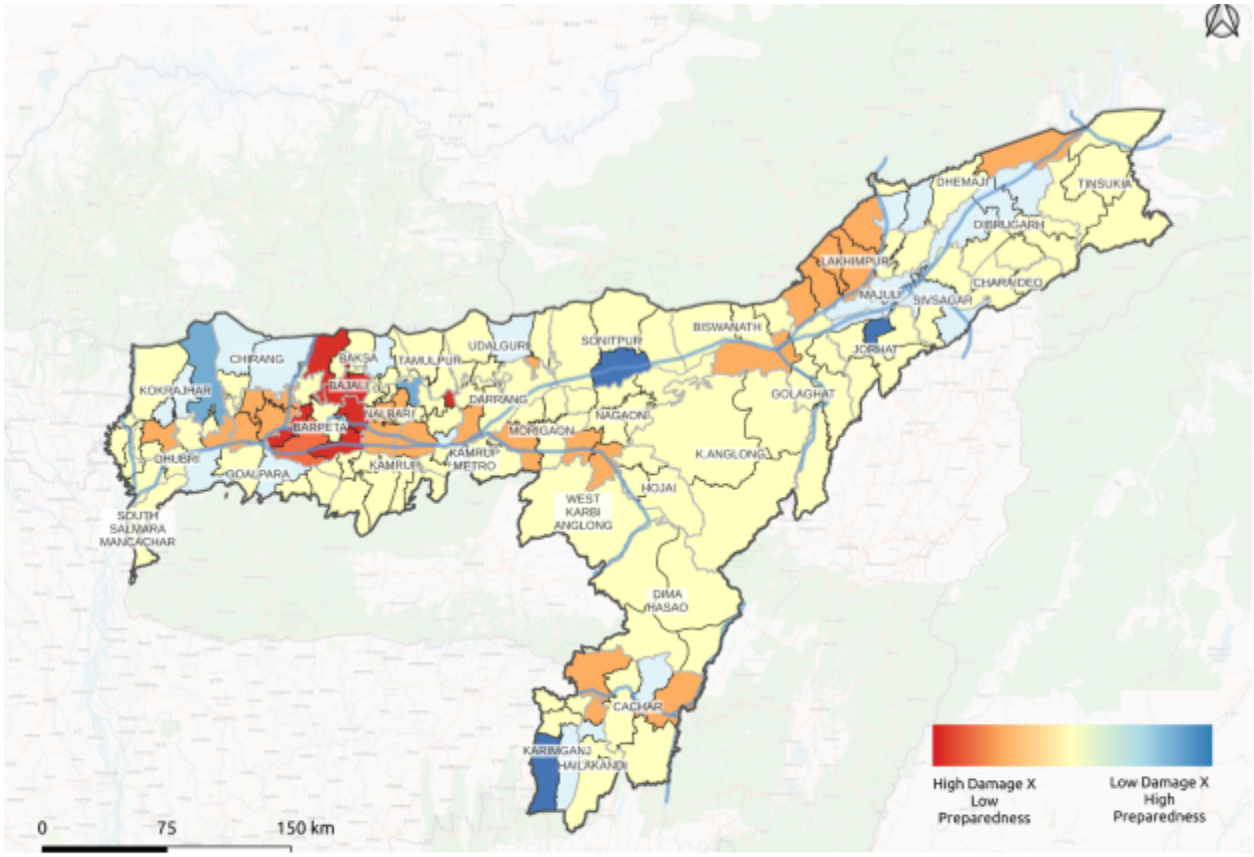
Table 2: SEM predicted scores for latent variables of the top 10 most flood impacted revenue circles

The scores for all revenue circles are available here: Link

a. It can be observed that, between 2018-2022, revenue circles in Barpeta and Bajali districts were most impacted by floods followed by Silchar. Of which, Barpeta, Baghbor and Silchar are also vulnerable demographically.
b. We can also observe that **infrastructure access** is high in these most impacted revenue circles. As we considered infrastructure damages (roads, embankments, bridges) from FRIMS portal in the **flood_impact** assessment, we can hypothesise that significant flood impact is actually the impact on infrastructure.
c. The troubling observation is that most of these most impacted revenue circles are relatively poor in **preparedness.**

Preparedness x Impact: To identify where government intervention is needed, we try to identify the revenue circles that are most impacted, but where relatively less preparatory expenditure took place over the years.

a. In the following map, we can observe the interaction between preparedness and flood impact. All the three revenue circles of Bajali are severely under-prepared given the flood impact it has seen. Even Barpeta district has 4 of its 6 revenue circles in this category. Lakhimpur is another district in which most revenue circles are under prepared.
b. Similarly, a few revenue circles are relatively better prepared given the flood impact they had seen. These are Patherkandi in Karimganj, Tezpur in Sonitpur, Kaliabor in Nagaon, Ghograpar in Nalbari, East Jorhat in Jorhat. It is interesting to note that while High-Impact, low prepared revenue circles are clustered together, low-impact, high prepared revenue circles are scattered across the state.



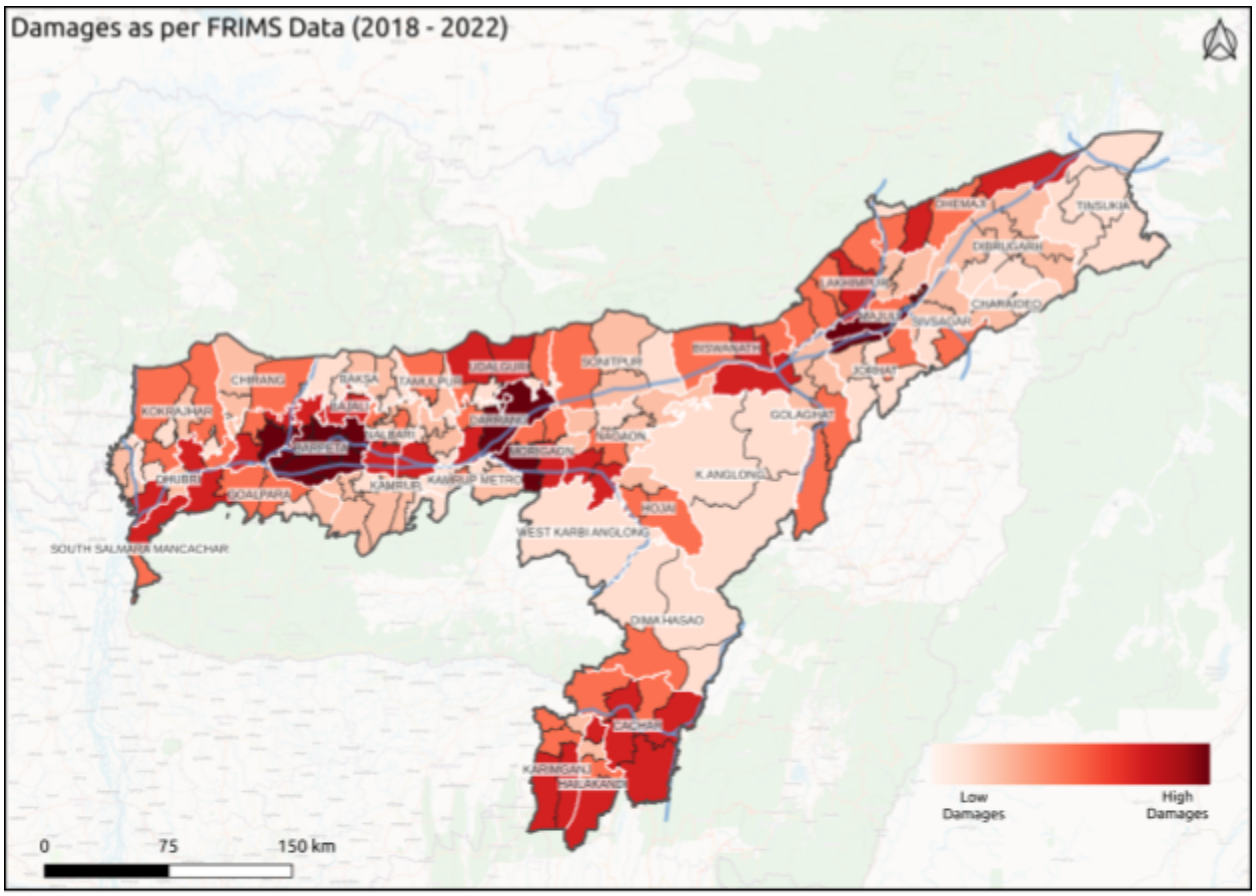
Disclaimer : Administrative boundaries on this map should be used for reference purposes only. CDL makes no claims concerning the precision and accuracy of the administrative boundaries presented on this map

Figure 11 : Map showing preparedness of revenue circle against flood impact score.

Model also estimates other factors like access to infrastructure, flood proneness and demographic vulnerability for each revenue circle.

Validation of SEM predictions

We compare the preparedness against flood impact scores from the SEM with the actual government response against reported flood damages in the last five years.

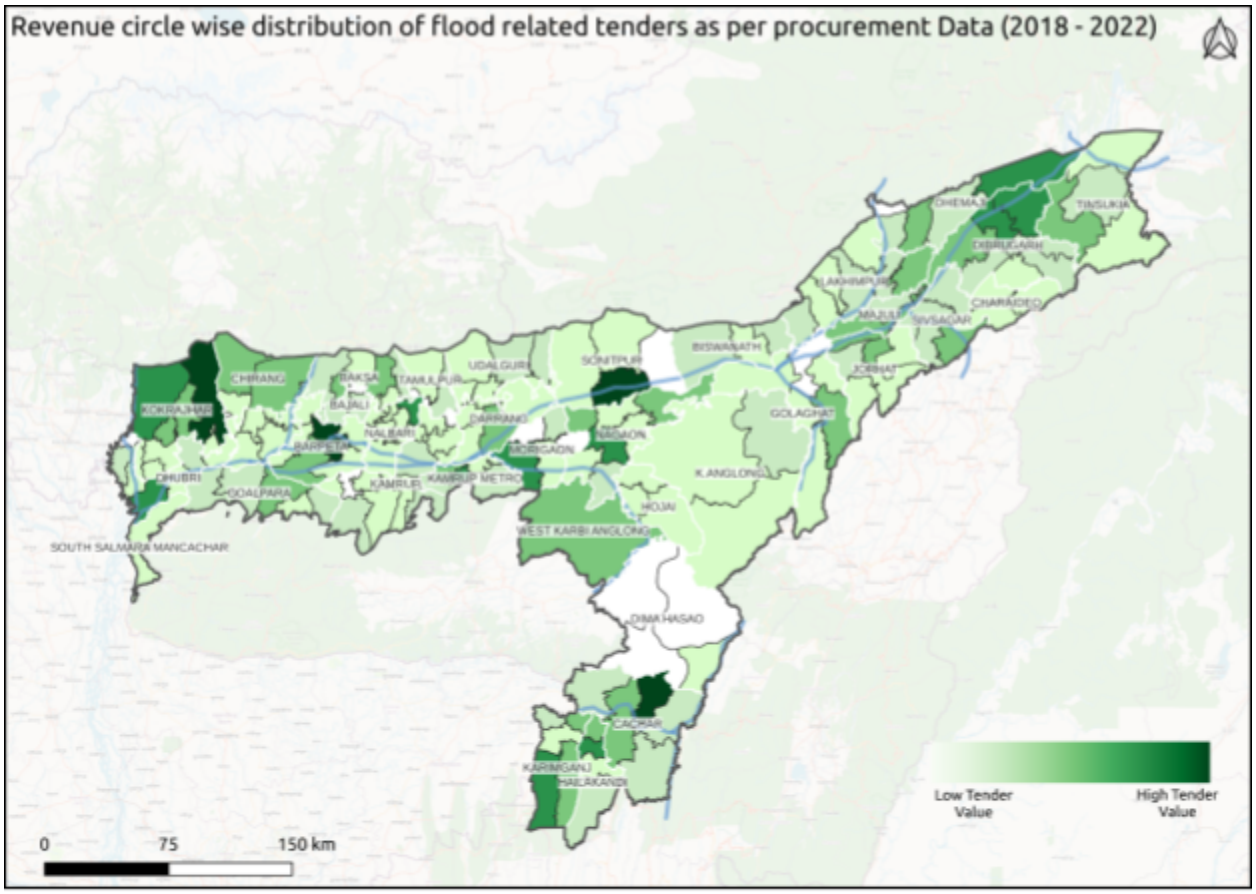


Disclaimer : Administrative boundaries on this map should be used for reference purposes only. CDL makes no claims concerning the precision and accuracy of the administrative boundaries presented on this map

Figure 12 : Five years of damages as per FRIMS mapped to revenue circles

The above map shows the **cumulative damages** recorded at revenue circle level **from 2018 to 2022 date as per ASDMA's records.** We have taken all the parameters mapped in the FRIMS for preparing the map. Each damage is normalised and added up to get the final aggregated damage. If a damage is available only at district level, the data is disaggregated to the revenue circle using the relative area of the revenue circle in the district as the weight to distribute the damage. It is crucial to disaggregate the data at revenue circle level as information on houses damaged by floods and animals affected or washed away are recorded at district level.

Having looked at the damages, we now look at flood related projects tendered in the last five years, to understand how the efforts in response to floods are distributed.

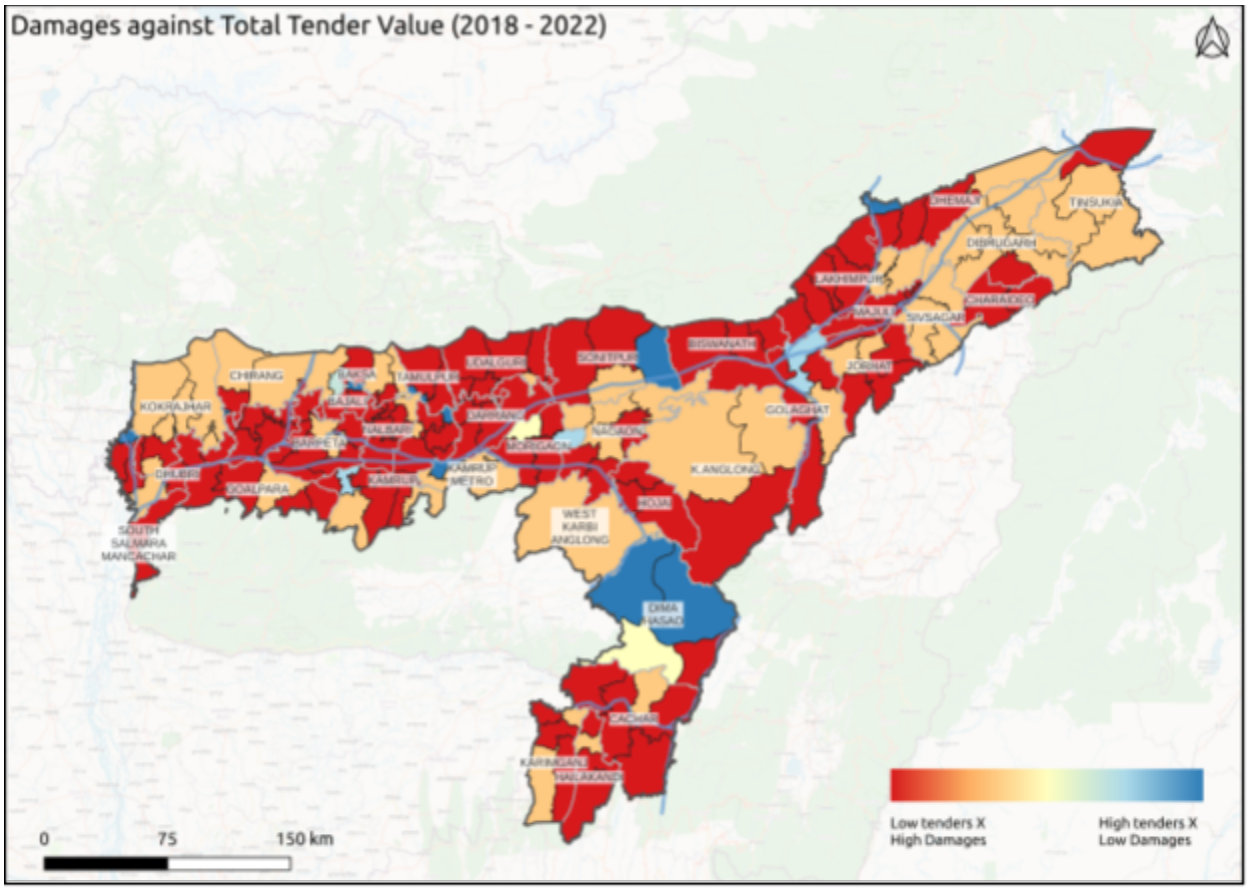


Disclaimer : Administrative boundaries on this map should be used for reference purposes only. CDL makes no claims concerning the precision and accuracy of the administrative boundaries presented on this map

Figure 13: Flood related tenders from last five years mapped to revenue circles

We can see that districts Dibrugarh has received the highest number of flood-related tenders followed by Kokrajhar, Cachar, Nagaon and Karimganj from 2018 to 2022.

Next, we layer the above two maps to compare with the map showing preparedness of revenue circle against flood impact score (Figure 14).



Disclaimer : Administrative boundaries on this map should be used for reference purposes only. CDL makes no claims concerning the precision and accuracy of the administrative boundaries presented on this map

Figure 14: Flood tenders mapped against damages for 2018-2022

The above map provides a consolidated picture of the distribution of total tenders against the aggregated flood damages. The map can be interpreted as follows:

(1) Dark red revenue circles are those which require more attention as the relative damages are more than the relative funds allocated through tenders.

(2) Medium yellow revenue circles are those that are next in line in terms of requiring more attention to resolve flood damages.

(3) Light yellow to blue revenue circles are those that are fairing relatively well compared to the others in terms of damages against fund allocations.

RELATIVE LOW TENDER VALUE X HIGH DAMAGES (2018-2022)		RELATIVE HIGH TOTAL TENDER VALUE x LOW FLOOD DAMAGES (2018-2022)	
Revenue Circle	**District**	**Revenue Circle**	**District**
Kalgachia	BARPETA	North Guwahati (Pt)	KAMRUP METRO
Pachim Nalbari	NALBARI	Sarupeta (Pt)	BAKSA
Chenga	BARPETA	Silonijan	K.ANGLONG
Morangi	GOLAGHAT	Tezpur	SONITPUR
Dhakuakhana (Pt-II)	DHEMAJI	Dibrugarh East	DIBRUGARH
Sarthebari	BARPETA	Ghograpar (Pt)	NALBARI
Barnagar (Pt)	BARPETA	Chabua	DIBRUGARH
Barkhetri	NALBARI	Kokrajhar (Pt)	KOKRAJHAR
Boitamari	BONGAIGAON	Jorhat East	JORHAT
Chapar (Pt)	DHUBRI	Udarbond	CACHAR

Table 3: Critical revenue circles based on mapping of tender value against damages

From the map above, it can be interpreted that the relative fund allocation through tenders to Kamrup Metro, Baksa, Dibrugarh, Kokrajhar and Karbi Anglong is much more than the relative flood damages in the other districts of the state. Conversely, the districts Barpeta, Nalbari, Golaghat, and Dhemajii are where the relative flood damages are much higher than the relative fund allocation of tenders in comparison to other districts of the state. These top revenue circles from Barpeta, Cachar, Bajali, and Lakhimpur are captured by the SEM model providing a validation for the model.

Limitations

1. Administrative boundaries are dynamic and information recorded is in a latest version which is not public yet, causing challenges in mapping damages accurately. We are however working closely with the Assam State Disaster Management Authority now to rectify this.
2. The entries in Flood Reporting and Information Management System (FRIMS marked) as ‘U/A’ are taken as a null value.

3. Some damages recorded do not give details of severity and type, for example, we can see the record of road damages but cannot see the road type which is damaged, limiting our inferences.
4. From the year 2021 onwards, we used the granular damages data available at the revenue circle level. But for years 2018-20, district level damages data was used.
5. The FRIMS data are recorded in two different spatial scales - district and revenue circles, which becomes challenging to consolidate data. The problem is felt the most for infrastructure damages which could be overcome by adding geographical coordinates.
6. As procurements related to floods are not tagged as such, we have identified them using keywords and filters and then validated them. Similarly, locations for tenders are not mentioned accurately for the place the proposed project is for. As such, around 818 out of 4,400 tenders remain without geocoding at revenue circles level instead have been mapped to district level only.
7. Currently, we are working with 72 variables but aim to include more in consultation with the ASDMA team and other experts.

Conclusion Summary:

The major contributions of the project are:

1. Single point of access for all updated (near-real time) flood related information across different departments and categories, namely-
    a. Geo-spatial and weather data
    b. Access to infrastructure data
    c. Socio - economic data
    d. Flood damages data
    e. Finance data
2. Machine learning model to predict the probability of flooding across the state of Assam for weather forecasts and risk assessment at revenue circle level.
3. The results from the models are useful to produce detailed district-level flood reports for decision makers to help prioritisation. These districts are where there is a gap in preparedness level with respect to impact observed.:

1. Bajali

Flood Proneness Score: 0.48

Flood Impact Score: 0.70

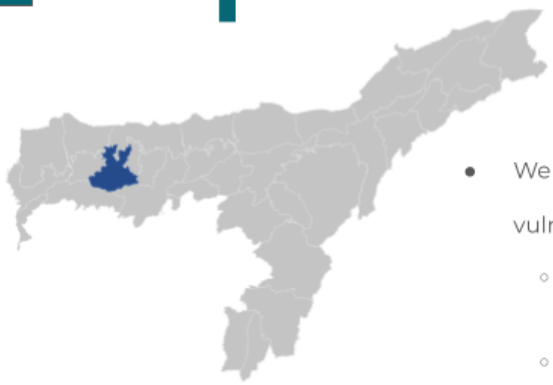
Preparedness Score: 0.21

Note : Bajali was part of Barpeta till 2021.

District Summary

- Bajali scores higher on the flood impact score and low on preparedness.
  - A total of **10,12,601 people were affected by floods** between 2018 to 2022
  - Between 2018-2022, **3341 road damages, 402 bridge damages and 323 embankment damages occured**
- In terms of government response
  - Despite of access to infrastructure being decent, we see high number of damages each year calling for repeated measures to repair infrastructure.
  - This can be seen from the procurement data as well – 50% of the tenders (61% by value) in Bajali in this period were related to repair and restoration
  - **Over ₹80 crores were spent on Repair, Restoration and Improvement tenders** .
  - **9/10 tenders in the district came from Sarupeta** revenue circle

2. Barpeta



Flood Proneness Score: 0.33

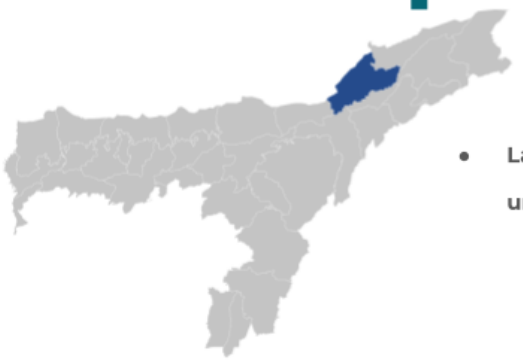
Flood Impact Score: 0.81

Preparedness Score: 0.36

District Summary

- We see high flood impact scores as well as high demographic vulnerability in the district of Barpeta.
  - **Barpeta also ranks number one in most damages reported as per ASDMA** records in the past 5 years.
  - **32,96,224 people were affected by floods in Barpeta** between 2018-2022.
- In terms of government response
  - There have been **108 flood-related tenders in Barpeta valuing to about ₹355 crores**
  - Of these tenders, over **60% of the tenders were identified as repair, restoration and improvement tenders.**
  - This indicates the need for more long term measures to reduce the impact of floods.
  - Most **tenders were concentrated in the Barpeta revenue circle** (HQ) of the district.

3. Lakhimpur



Flood Proneness Score: 0.42

Flood Impact Score: 0.21

Preparedness Score: 0.25

District Summary

- **Lakhimpur is another district where 6 of its 7 revenue circles are underprepared**.
  - **The flood impact in Lakhimpur is due to the high loss of human lives.**
  - 280 people (0.05% of the population affected by floods) lost their lives due to floods in Lakhimpur, as compared to 0.001% of the population affected in Bajali and Barpeta (13 and 52 human lives lost).
  - Thus **more measures that help in reducing vulnerability are needed** in Lakhimpur.
- In terms of the government response
  - **88 flood related tenders were issued in Lakhimpur between 2018-2022 valuing to over ₹300 crores**
  - Most tenders (59 out of the 88 tenders) seem to be Immediate measure / SDRF tenders indicating the high flood impact and lack of preventive measures in the district

4. Bongaigaon

Flood Proneness Score: 0.41

Flood Impact Score: 0.20

Preparedness Score: 0.21

District Summary

- **4 out of Bongaigaon's 5 revenue circles are underprepared**.
  - Flood impact is high due to **population, livestock affected and crop area damaged**
  - Despite being a small district in terms of area, **Bongaingaon saw 1137 road damage occurrences due to floods between 2018 and 2022**.
  - Of these, Srijangram scores high on demographic vulnerability.
- Government Response
  - The district performs highly on infrastructure access but **75% tenders (23 out of the 30 tenders)** valuing about ₹24 crores in the district have been **repair and restoration tenders.**

5. Nagaon

Flood Proneness Score: 0.46

Flood Impact Score: 0.17

Preparedness Score: 0.29

District Summary

- **In Nagaon,** Raha and Kampur revenue circles are relatively under-prepared compared to other revenue circles when seen along with the flood impact they face.
    - **9,70,956 people got affected in Nagaon with 1559 road damages occurring due to floods during 2018-22.**
- Government Response
    - The preparedness score of Nagaon is relatively higher due to the **high number of tenders since 2018**.
    - **There have been 113 tenders in the district valuing over ₹490 crores.**
    - Most tenders in the districts are concentrated in Nagaon revenue circle

Other Districts in need of Intervention

Cachar, Nalbari, Darrang, Dhubri and Dhemaji ranked closely.

	District	Flood Proneness Score	Flood Impact Score	Preparedness Score
6.	**Cachar**	0.73	0.18	0.40
7.	**Nalbari**	0.43	0.15	0.25
8.	**Darrang**	0.39	0.15	0.26
9.	**Dhubri**	0.33	0.15	0.29
10.	**Dhemaji**	0.43	0.12	0.26

Early Impacts and the Way Forward

The project ideates a data ecosystem based on accessible, interoperable datasets suitable for further analytics or the creation of tools to improve flood management and response in Assam. The data-driven approach aims for improved efficiency and can contribute to reducing the loss of lives and livelihoods caused by floods in Assam. It also enables the generation of targeted reports for administrative units at different scales.

This **model is already being institutionalised within the Assam State Disaster Management Authority (ASDMA).** In our most recent field visit this November, the CEO GD Tripathi committed to **using our data model to make decisions** on how funds from the State Disaster Response Fund (SDRF) will be spent at the next State Executive Committee Meeting. **By the 30th December 2022**, we will be able to measure how many of these decisions match our recommendations, including any

- increase in funds allocated to districts with high vulnerability
- increase in projects approved in districts with high vulnerability
- number of beneficiaries benefitting from these allocations
- decrease in funds allocated to districts with low vulnerability
- decrease in projects allocated to districts with low vulnerability

We are also expecting significant efficiency gains within ASDMA, primarily from the reduced administrative burden (time and cost saved) in assessing district preparedness. At present the Minimum Preparedness Scorecard is a manual process requiring over 10,000 government staff (district information officers) and volunteers to collect and report information. This process takes at least 4 months to complete and is run twice a year. As the CEO has committed to using our data model to create the 2023 Preparedness Index, **by January 2023**, we expect our data model to deliver

- time savings, reduced from 4 months to 2-3 days (inc. testing)
- human capital savings, reduced from 10,000 people to 5 people
- cost savings from reduced expenditure for the 10,000 district officers and volunteers to collect/report information (wages, travel, etc)

Furthermore, our model will improve accuracy and granularity and can be deployed more frequently (as opposed to only twice a year).

Moving forward, we would like to ultimately increase the granularity of the analysis by taking it to the village level, subject to the availability of data. Currently, this model uses the cumulative data from 2018-2022 at the revenue circle level for the analysis.

The variables used to measure flood proneness, demographic vulnerability and infrastructure access are mostly derived from satellite imagery, and they can be estimated at the village/block level. If the tender's data and damages data in FRIMS can be geo-tagged, this hyper-local analysis would be possible.

We used only the procurement data as measured variables for the **preparedness** of a revenue circle. We would like to add other variables like SDRF allocation and survey responses from field officers about preparedness measures taken.

Qualitative data like community perceptions, survey data etc., on each of the latent variables can be ingested in the data in further iterations.

We would ultimately need to develop a front-end, interactive tool to make it easier for decision makers to use this model to ensure better preparedness against floods and in turn more equitable outcomes for the most vulnerable communities. We will continue our work to strengthen their capacities so that they can go even further to protect lives and livelihoods.