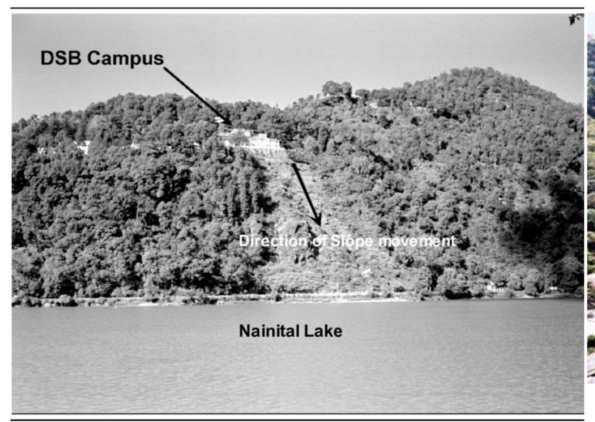
Applying Mathematical Modeling to Geophysical & Meteorological Data for Landslide Analytics & Forecasting

Scienteer Project #171603

Introduction and Background

Landslides are unpredictable natural disasters caused by changes in slope stability. The main causes of landslides are rainfall, human activity, and erosion.

- **18,000 deaths** and **4.8 million people** affected in past 2 decades
- **$3.5 billion** in annual damages in the U.S.



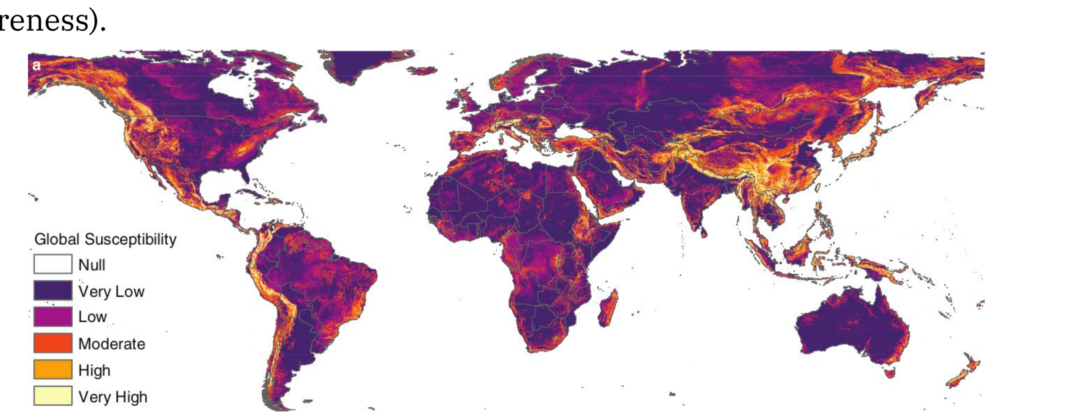
A photograph of a landslide near the DSB college campus in Nainital, India

A photograph of a landslide destroying homes and other infrastructure in its path.

Related Work

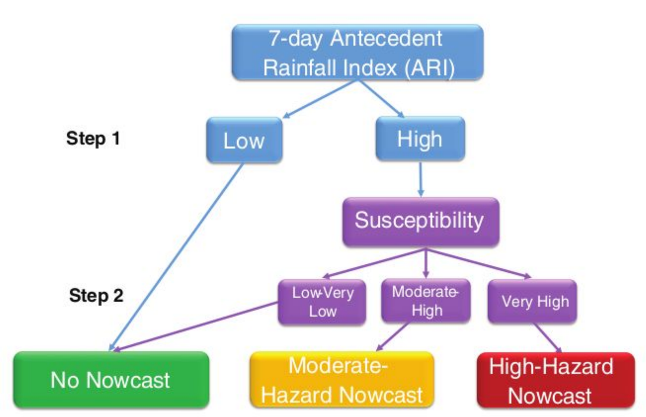
Landslide forecasting is a surprisingly difficult task due to lack of an obvious build-up and limited landslide-related data.

The only global-scale system for landslide forecasting that uses easily acquirable data is NASA's **LHASA** (Landslide Hazard Assessment for Situational Awareness).



A global landslide susceptibility map produced by NASA's LHASA system [Courtesy of LHASA]

LHASA uses a decision tree for generating landslide hazard "nowcasts". The ARI is calculated using IMERG data and a global susceptibility map is generated. These are used together to issue "nowcasts" of varying importance.



LHASA's decision tree framework [Courtesy of LHASA]

1-day	3-day	7-day	FPR (%)
27	39	47	1
24	35	40	1
10	14	18	0.2
8	14	16	0.2

LHASA's evaluation statistics (above) and a snippet from the LHASA paper (below) [Courtesy of LHASA]

However, LHASA suffers from problems.

- TPR < 50%
- Model latency
- Uses modeling methods from 1990s

> issue and generate their own action plans. This system is not intended for local planning or to inform detailed infrastructure projects due to its geographic scope and spatial resolution. LHASA is also *not* meant to be used as a warning or forecasting system. This is due to the model latency (4–5 h from

Engineering Goal

Engineering Goal:

This research's goal is to create **GLAS: A Global Landslide Analytics System** for:

- Landslide forecasting (what's the risk of a landslide in the next 5 days?)
- Landslide severity analysis
- Terrain susceptibility analysis

Contributions:

- First ever publicly available dataset of landslide incidents + relevant features (location/time, weather, and terrain data)
- Empirically proven algorithm for finding rainfall-induced landslides
- More accurate, lower-latency landslide forecasting system

All graphics not cited on the board were produced by the student researchers

Methodology

Feature Selection → Data Compilation → Data Processing → Modelling → Analysis & Evaluation

Phase 1: Feature Selection

Phase 2: Dataset Compilation

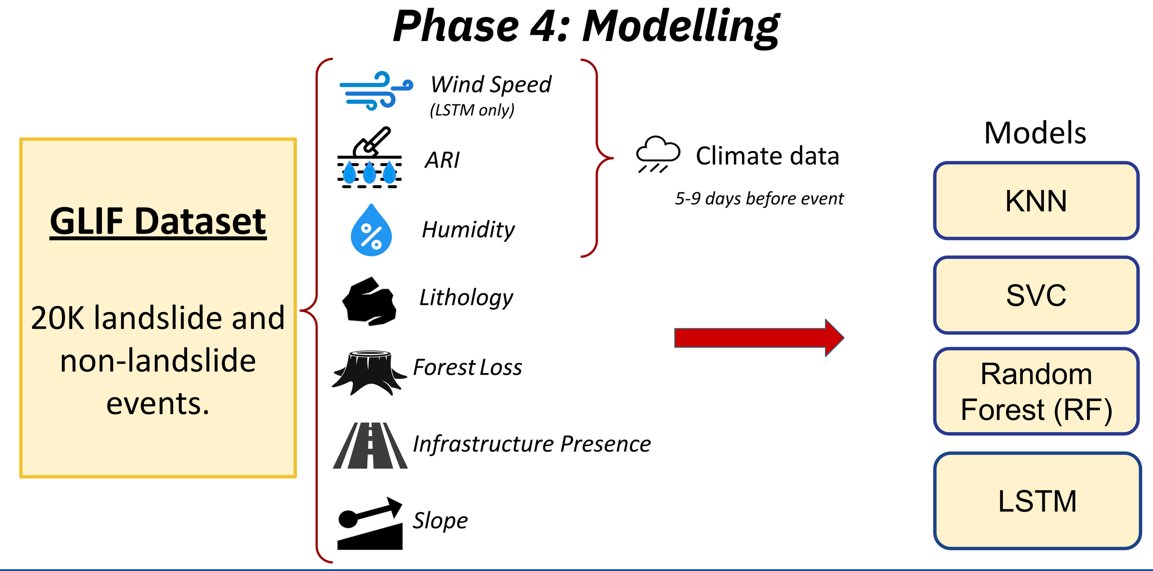
- NASA Global Landslide Catalog: catalog of ~11,000 landslides.
- Data was compiled for each day **15 days** before to the day of the landslide using:
  - worldweatheronline API: Climate data (precipitation, humidity, wind speed)
  - NASA's SRTM Dataset: Elevation data (Shuttle Radar Topography Mission)
  - Global Forest Change: Forest loss data (year of forest loss).
  - OpenStreetMap: Data on presence of roads and human infrastructure.
  - GLiM Dataset: Data about the types of rocks present around the world

Phase 3: Data Processing

- Antecedent **R**ainfall **I**ndex (**ARI**) was calculated using precipitation over 7 days.
- Slope calculated from elevation.
- Extensive map reprojection + smoothing required
- 90th percentile for slope values was taken

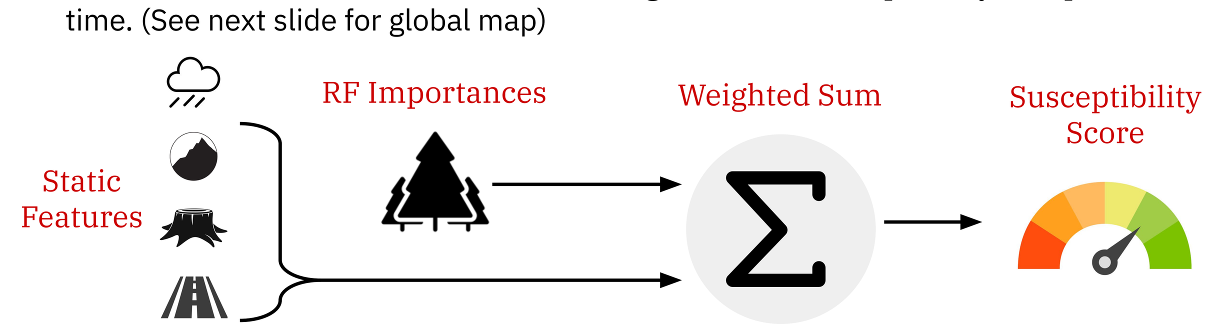
$$ARI = \frac{\sum_{t=0}^{6} p_t w_t}{\sum_{t=0}^{6} w_t}$$

Phase 4: Modelling

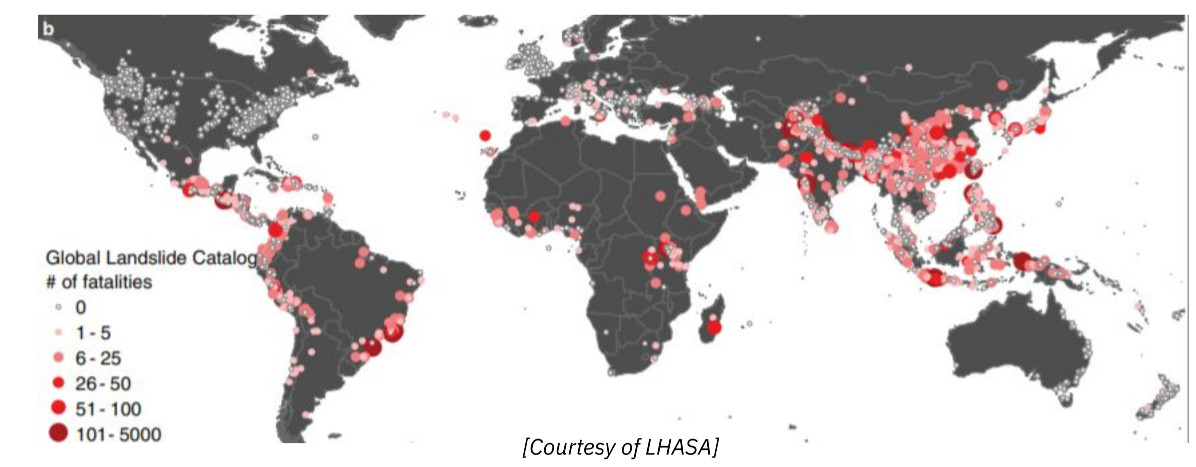


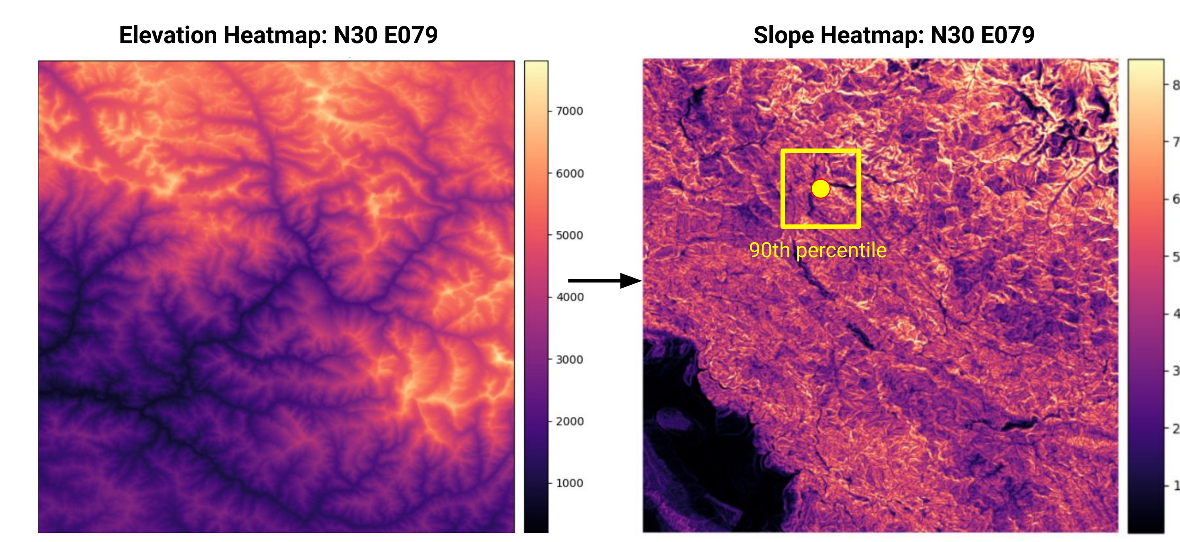
Phase 5: Analysis and Evaluation

- Recall, precision, accuracy + ablation study
- Susceptibility Score was constructed based on feature importances from the Random Forest. This is a metric indicating landslide susceptibility independent of time. (See next slide for global map)



Landslide Map



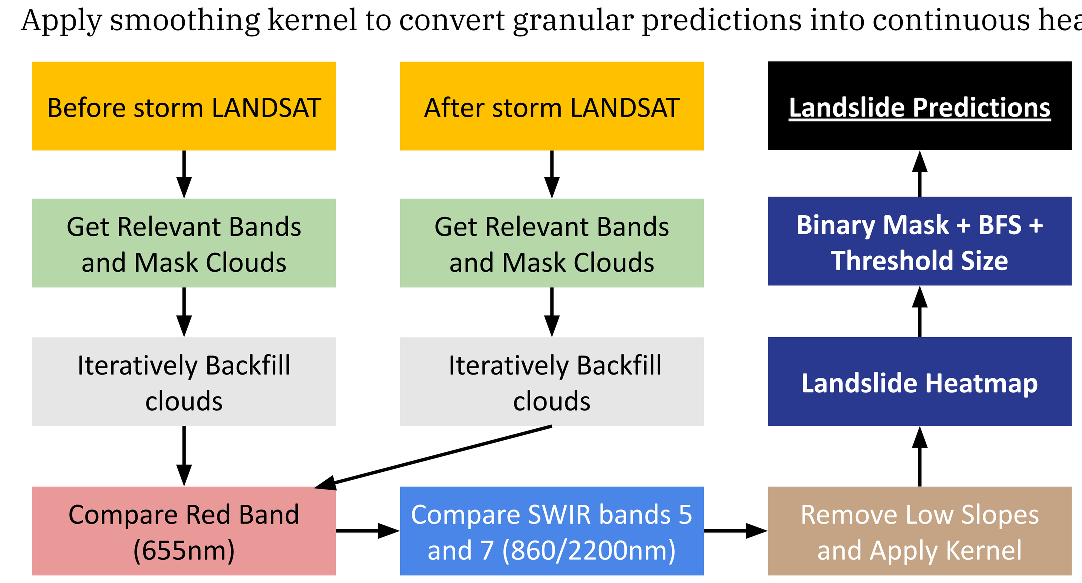


Landslide distribution visualization of the NASA Global Landslide Catalog (above), and an elevation map and slope map for a mountainous area in northern India generated from SRTM data

Detecting Unreported Rainfall-Induced Landslides using Historical Satellite Soil Moisture and Slope data

Key Principles:

- Reflectance in Red Band (655nm) indicates bare earth exposure
- Changes in the SWIR bands 5 and 7 (860 and 2200 nm), indicate soil moisture change
- Small slope values are not likely to be landslides
- Apply smoothing kernel to convert granular predictions into continuous heatmap



Data Evaluation and Result Analysis

- **Random Forest** model was the best model for both binary and severity classification with accuracies of **86.3% and 70.7%**, respectively. There was minim
- **Random Forest** model **forecasted 86.9% of landslides 5 days in advance**.
- **Landslide Detection Algorithm** achieved **93.5% TPR** & **12.1% FPR** on the GLC and **discovered 1345 uncatalogued landslides** in Nepal after a severe rainstorm.

Forecasting Landslide Day

A first generation of models were created for predicting in ***how many days*** a landslide is likely to occur (in a range from 6 to 9 days in the future or not at all). This is something that LHASA does not provide. Our initial Random Forest model obtained a raw accuracy of 37.4% and an early forecasting score (EFS) of 64.4%.

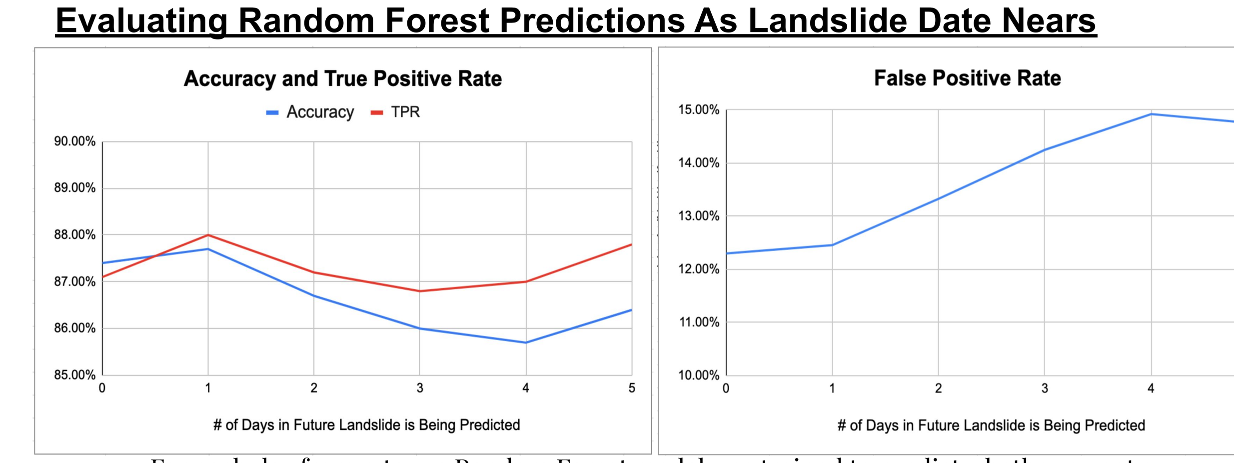
RF Landslide Date Prediction

True Date \ Predicted Date	no	6	7	8	9
no	0.67	0.066	0.075	0.078	0.13
6	0.086	0.39	0.2	0.14	0.17
7	0.1	0.24	0.27	0.18	0.2
8	0.12	0.26	0.21	0.19	0.24
9	0.15	0.2	0.17	0.18	0.29

Early Forecasting Score: Proportion of classifications that occurred on *or* before the actual day of the landslide.

Key:
- **No**: No landslide
- **6-9:** # of days until landslide
- **Blue outline** = correct prediction (contributes to EFS)

Evaluating Random Forest Predictions As Landslide Date Nears



For each day from 0 to 5, a Random Forest model was trained to predict whether or not a landslide would occur that many days into the future. When the number of days until the landslide decreased, the forecasting model's False Positive Rate also decreased. From days 1 to 4, accuracy and TPR decreased as the models had to predict the landslide further in advance.

Data and Results

- **Binary Classification**: Will it occur?
- **Severity Classification**: 0 - 3
  - 0 = no landslide. 3 = large landslide

	Binary	Severity
KNN	71.9%	62.3%
SVC	72.7%	62.6%
RF	86.3%	72.6%
LSTM	64.2%	55.8%

RF Binary Confusion Matrix

True \ Predicted	No	Yes
No	0.86	0.14
Yes	0.13	0.87

Recall: 86.9%
Precision: 85.9%

Subset A: 5k landslide events & 5k non-landslide events from random locations and times

Subset B: 5k landslide & 5k non-landslide events from known landslide locations but at different times

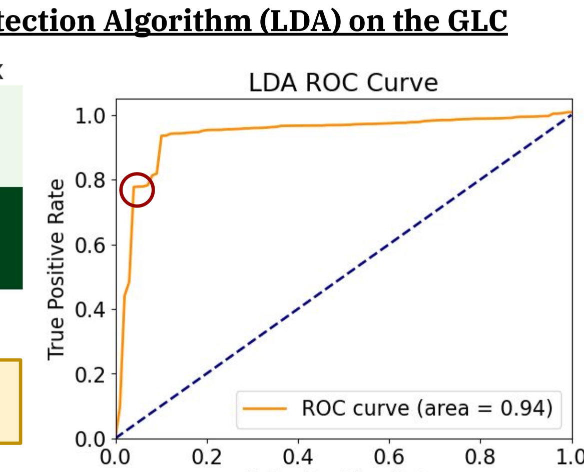
	Binary	Severity	EFS
Entire GLIF (RF)	86.3%	72.6%	64.4%
Subset A (RF)	90.4%	71.8%	69.0%
Subset B (RF)	87.5%	67.5%	58.3%

Validating Landslide Detection Algorithm (LDA) on the GLC

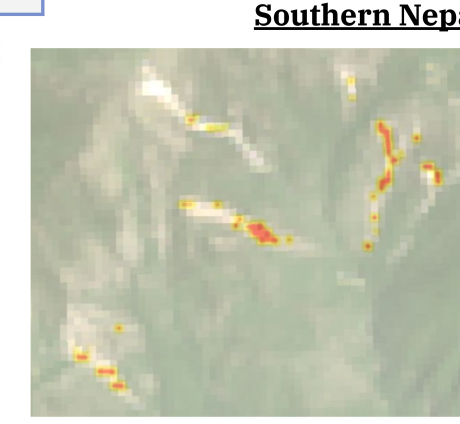
LDA Confusion Matrix

True \ Predicted	No	Yes
No	7603	1043
Yes	523	8123

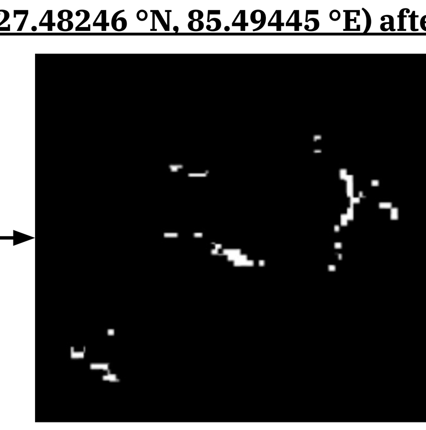
TPR: 93.9%
FPR: 12.1%



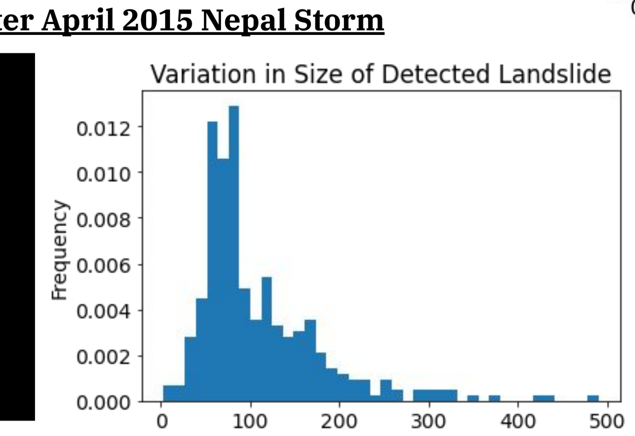
Southern Nepal (27.48246 °N, 85.49445 °E) after April 2015 Nepal Storm



LDA Landslide confidence heatmap



Binary mask of heatmap



- The LDA **detected 1436 rainfall-induced landslides** in Nepal due to the April 2015 storm
- **1345 landslides were not cataloged** in any existing landslide database

Conclusion

In this project, the researchers created:

- The **first comprehensive, open-sourced landslide dataset** with incidents and relevant features
- **Higher performance models** for landslide date and risk forecasting
- A **Landslide Detection Algorithm** using soil moisture and slope data

Our System	LHASA
True Positive Rates of **86.9%**.	True Positive Rates of **60%**.
False Positive Rates of **14.3%**	False Positive Rates of **3%**.
Uses satellite data to **find unreported landslides** with **94% TPR; 12% FPR**	No such capability
Forecasts landslides **5 days into the future with 86.3% accuracy**	Provides *"nowcasts"* with 4-5 hour latency.

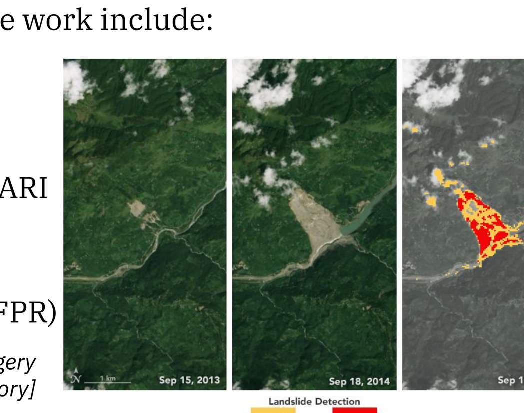
Our system outperforms the existing landslide hazard assessment models in **accuracy** and **latency**.

Landslides are a deadly disaster that affect the lives of millions of people and cause billions in damage annually. An early forecasting system that is more accurate and clairvoyant than existing systems can provide days in advance to prepare and evacuate, **saving lives and livelihoods**.

Future Work

In the future we'd like to incorporate more data sources and combine them in insightful ways. Some of the topics for future work include:

- Append data points from LDA to GLIF
- Additional Features
  - Fault lines
  - Combined feature from Lithology + ARI
- Deploy landslide web dashboard
- Computer Vision + Satellite Images
- Ensemble modeling techniques (lower FPR)



Landslide Detection using computer vision from satellite imagery [Courtesy of NASA Earth Observatory]

Random Forest Accuracies on Subsets

In addition to the whole GLIF dataset, two subsets of GLIF were created for the Random Forest model to train on to determine the model's performance with when non-landslide incidents were from random locations it hadn't seen before vs non-landslide incidents from former/future landslide locations.

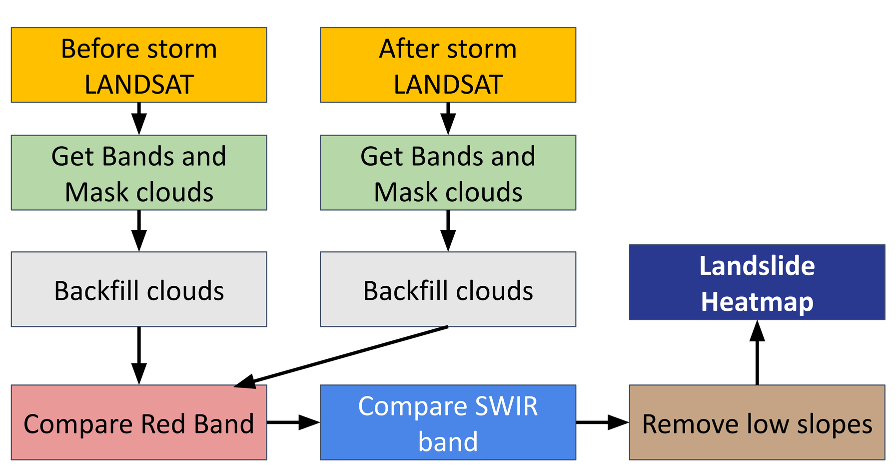
- **Subset A:** 5,000 landslide events and 5,000 non-landslide events from random locations and times.
- **Subset B:** 5,000 landslide and 5,000 non-landslide events from the same location, but at different times

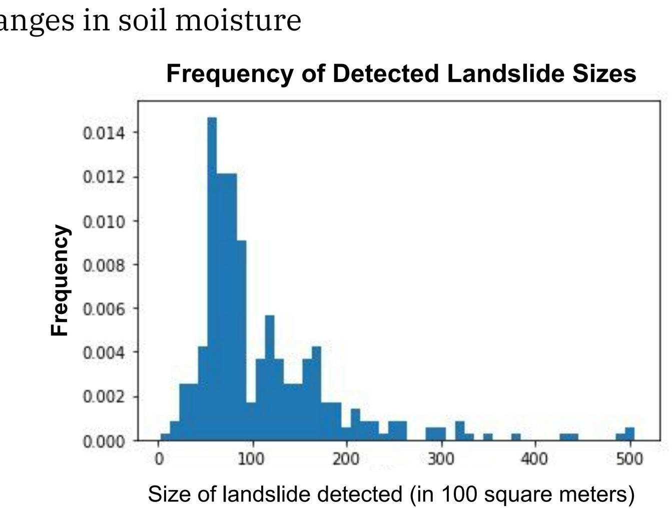
	Binary Classification Accuracy	Severity Classification Accuracy	Date Prediction Early Forecasting Score
Entire GLIF Dataset	86.3%	72.6%	64.4%
Subset A	90.4%	71.8%	69.0%
Subset B	87.5%	67.5%	58.3%

Algorithm for Mathematical Model Detecting Unreported Landslides using Historial Satellite Soil Moisture and Slope data

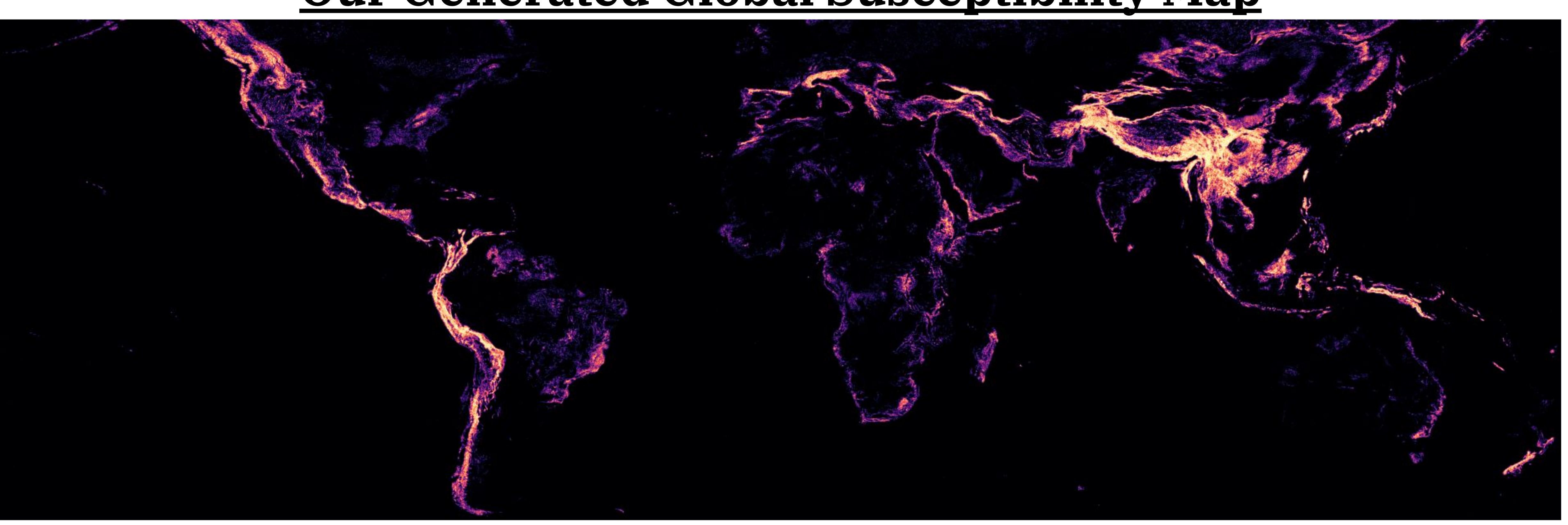
Algorithm Key Principles

- Reflectance in Red Band (655nm) indicates bare earth exposure
- Changes in the SWIR bands 5 and 7 (860 and 2200 nm), indicate changes in soil moisture
- small slope values are not likely to be landslides

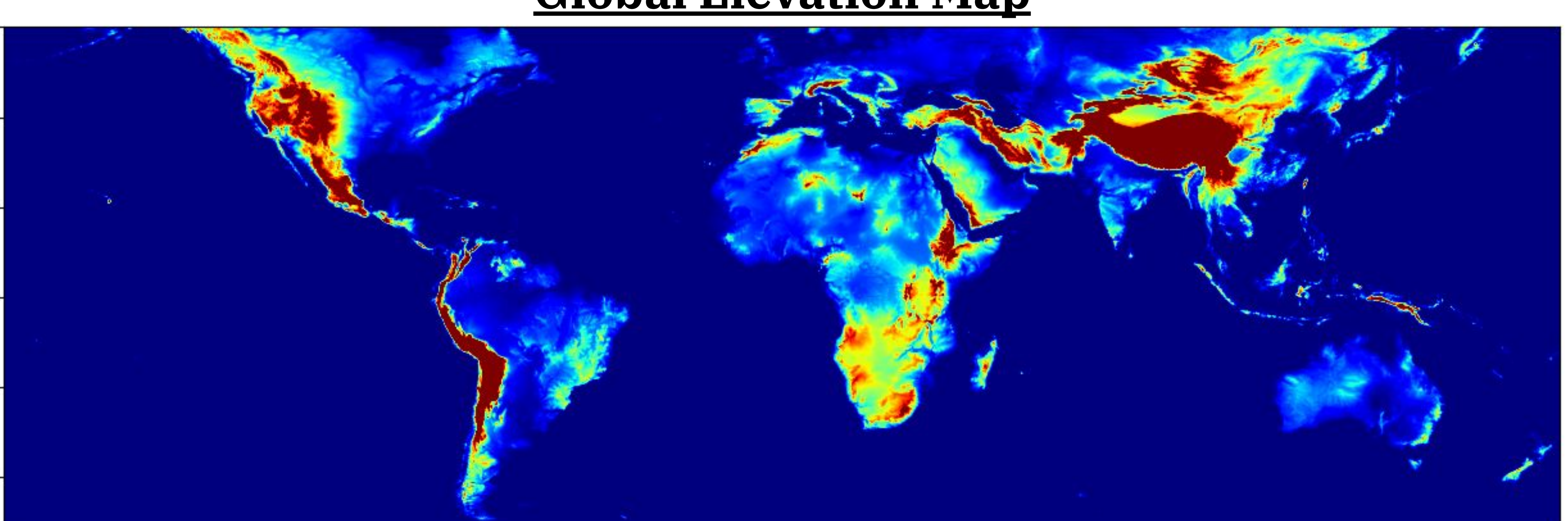




Our Generated Global Susceptibility Map



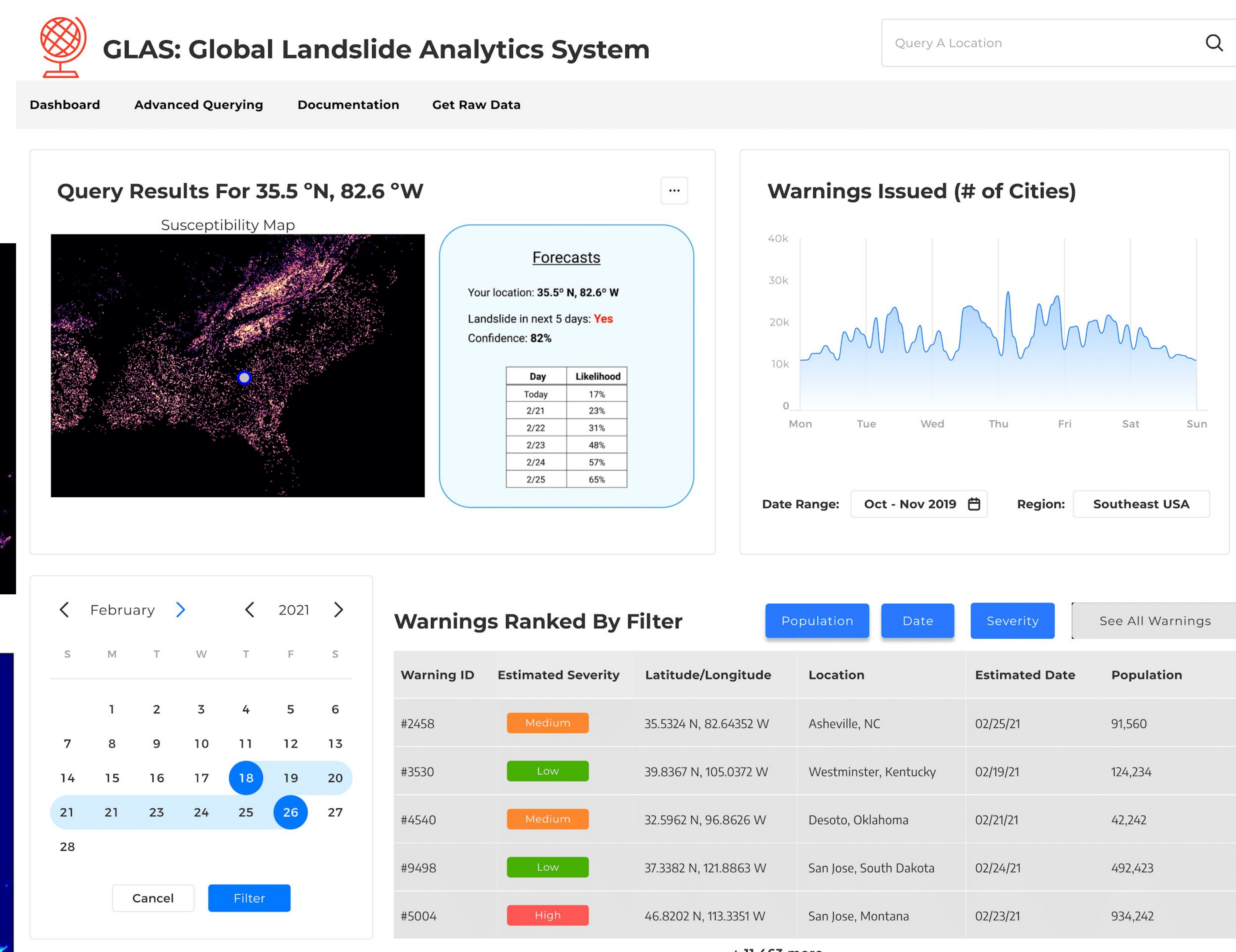
Global Elevation Map



Part of Dataset

id	date	lat	lon	country	fatalities	injuries	precip0	temp0	air0	humidity0	wind0	landslide	forest	ARI1	ARI0	slope	osm
1069	9/4/16	46.726906	13.787332	Austria	0	0	3.8	67	1019	96	6	1	1	0.24081206	2.62486829	22.012	433
1855	3/23/17	49.726406	-116.91183	Canada	0	0	0.1	33	1020	100	8	1	1	0.54643713	0.25723617	34.101	445
797	10/20/09	18.5347	-72.4097	Haiti	4	0	3.4	87	1013	96	13	1	1	4.71700196	3.62581061	25.38	672
12967	12/31/09	4.42142905	-75.220709	Colombia	0	0	0	84	1017	96	11	1	1	0.31050131	0.17407389	29.537	681
13089	5/1/14	39.2902	-76.6651	United_State	0	0	0.4	77	1010	99	18	1	1	9.22206381	2.6665098	0	3909
9036	7/12/13	36.2629	-115.6158	United_State	0	0	0.2	84	1013	52	12	1	1	0.52317938	0.28614712	4.063	3930
7990	5/10/09	4.44059512	-75.243905	Colombia	0	0	0.8	82	1015	82	11	1	1	0.72870658	0.7786029	30.007	4597
3954	10/25/10	15.5227	-85.265	Honduras	0	0	4.7	86	1012	100	8	1	1	3.67732159	4.09146546	12.026	4597
2288	3/29/14	42.3946	-122.2137	United_State	0	0	2.1	47	1015	99	19	1	1	2.28104177	2.00814981	13.668	4611
13419	12/2/12	36.9933	-122.0206	United_State	0	0	9.5	60	1019	97	41	1	1	4.19111223	8.07512121	7.014	5236
5865	2/6/17	32.9134621	-107.77359	United_State	0	0	0	55	1016	52	32	1	1	0	0	0.853	5261
10892	9/15/16	20.8939	-156.6464	United_State	0	0	0.6	88	1016	77	25	1	1	1.69351435	1.01735069	0	5265
11940	3/21/12	43.6609	-123.3327	United_State	0	0	7.7	37	1015	100	11	1	1	5.8608551	6.69390133	22.175	5290
8746	8/1/13	44.4305	-118.1417	United_State	0	0	2.2	75	1014	74	16	1	1	0.33526948	1.53925477	11.203	5298
12993	4/17/14	43.4756	-110.7826	United_State	0	0	0	37	1022	98	10	1	1	0.437135	0.18141525	9.15	5423
8110	3/30/11	40.1625	-123.7874	United_State	0	0	0	65	1028	97	24	1	1	1.59220117	0.84505983	1.595	5453
7964	7/13/15	38.499	-80.7205	United_State	0	0	3.6	84	1013	99	12	1	1	2.72345949	3.56428992	12.562	5463
12732	3/4/15	37.336	-83.1295	United_State	0	0	3.4	52	1019	99	13	1	1	0.49019915	2.48276968	21.632	5472
3365	3/22/17	48.736721	-122.35906	United_State	0	0	0.7	51	1016	92	17	1	1	0.17214387	0.51866912	23.914	5495
5092	12/9/15	45.88	-118.9662	United_State	0	0	2.6	57	1012	79	43	1	1	2.18294704	2.50415515	22.002	5499
4140	2/28/14	45.6721	-121.877	United_State	0	0	0	46	1006	97	19	1	1	1.04431399	0.32729628	4.727	5588
6425	2/9/16	54.8366	-2.7812	United_Kingc	0	0	0.7	41	984	92	39	1	1	0.84299744	0.81607201	0	5603
10633	8/5/13	-41.1212	146.1066	Australia	0	0	5.5	55	1007	93	42	1	1	2.32868108	4.67034097	0	5610
4716	9/7/08	30.0075	31.2774	Egypt	31	46	0	103	1009	90	12	1	1	0	0	12.823	5865
6041	8/29/15	45.2252	-122.3394	United_State	0	0	4.2	69	1011	97	25	1	1	0.19843933	2.82776051	18.61	5874
9835	12/8/15	47.488	-122.3633	United_State	0	0	7.9	56	1011	99	27	1	1	1.48247044	6.02616722	25.606	6197
7193	1/18/12	44.8746	-123.9314	United_State	0	0	12.5	50	1012	99	45	1	1	1.26321335	8.60157342	24.796	6208
12958	1/27/12	4.4405951	-75.249405	Colombia	0	0	0.3	82	1015	84	4	1	1	0.90172491	0.5571938	29.888	6477
9009	11/20/12	42.9171	-124.1018	United_State	0	0	14.2	53	1008	99	24	1	1	14.6064212	13.3673329	21.375	6480
3533	5/11/16	6.8586	37.7542	Ethiopia	42	0	0.1	84	1020	92	4	1	1	0.21856713	0.19425006	12.46	6610
12356	12/26/13	46.0539	9.4192	Italy	0	0	6.4	42	1004	99	19	1	1	11.4344382	7.28637848	29.621	6886
1056	2/21/17	35.4625036	-120.67602	United_State	0	0	0	59	1025	97	33	1	1	2.9149452	0.98760317	0.763	7111
11715	5/9/18	50.075462	-119.41801	Canada	0	0	7.7	57	1015	94	5	1	1	0.16812221	5.1748201	15.177	7151
1290	4/6/10	-22.9153	-43.0715	Brazil	0	0	2.7	77	1013	98	13	1	1	2.36511038	2.47487635	15.648	7486

Our Landslide Analytics Dashboard



OFFICIAL ABSTRACT and CERTIFICATION

Applying Mathematical Modeling to Geophysical & Meteorological Data for Landslide Analytics & Forecasting

Ishaan Javali and Shrey Joshi

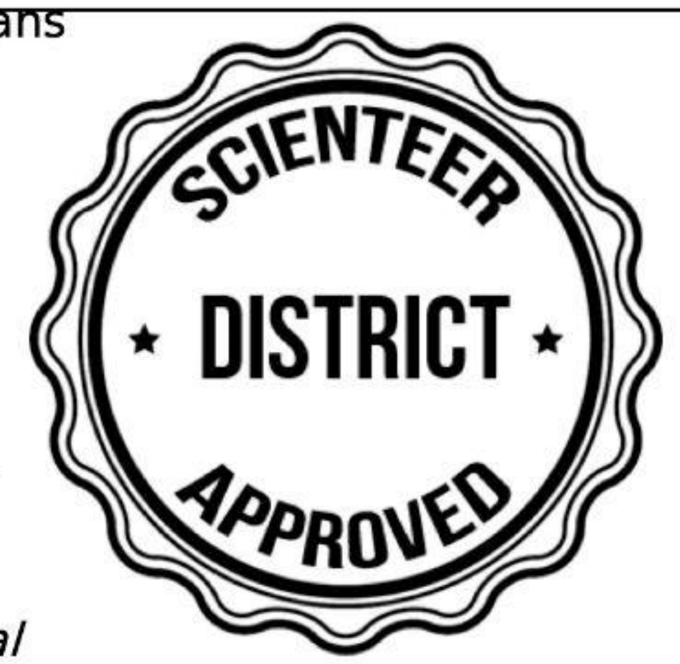
PLANO EAST SENIOR HIGH SCHOOL, Plano, Texas, US

Category

Mathematics

Landslides result in billions of dollars of damage annually in the U.S. and affected 4.8 million people over 2 decades. Landslide detection is a critical task for the protection of human life and livelihood in mountainous areas. Existing landslide warning systems are inefficient in predicting landslide occurrences, in terms of accuracy, latency, or applicability. NASA's LHASA system has a 4-5 hour latency for landslide prediction with an 8-60% Probability of Detection (POD) and uses a 5-feature, static susceptibility map along with precipitation data. Other research papers collected specific data for limited regions. We propose GLAS, a Global Landslide Analytics System featuring higher-performance, scalable landslide forecasting models along with the first publicly available dataset of landslide events and features. Landslide incidence reports for GLAS were collected from the Global Landslide Catalog. For each landslide, indicative features of landslides were collected over a 15-day period: elevation, climate, forest loss, and street presence data. These features were compiled into a first-of-its-kind, public global dataset for landslide and non-landslide events. KNN, SVC, and Random Forest algorithms and an LSTM neural network were trained on the dataset to forecast whether there would be a landslide 5 days in advance, yielding an accuracy of 86.3% and a detection rate of 86.9% on the test set. These results exceeded LHASA's POD of 60%, thus providing people days to prepare and evacuate. We are currently refining and testing a mathematical model for finding unreported landslides by parsing historical satellite imagery, providing additional training data for our forecasting models.

1. As a part of this research project, the student directly handled, manipulated, or interacted with (check ALL that apply):
   ☐ human Participants ☐ potentially hazardous biological agents
   ☐ vertebrate animals ☐ microorganisms ☐ rDNA ☐ tissue
2. I/we worked or used equipment in a regulated research institution or industrial setting:
   ☐ Yes ☑ No
3. This project is a continuation of previous research:
   ☑ Yes ☐ No
4. My display board includes non-published photographs/visual depictions of humans (other than myself):
   ☐ Yes ☑ No
5. This abstract describes only procedures performed by me/us, reflects my/our own independent research, and represents one year's work only:
   ☑ Yes ☐ No
6. I/we herby certify that the abstract and responses to the above statements are correct and properly reflect my/our own work:



This stamp or embossed seal attests that his project is in compliance with all federal and state laws and regulations and that all appropriate reviews and approvals have been obtained including the final clearance by the Scientific Review Committee.

Continuation/Research Progression Projects Form (7)

Required for projects that are a continuation/progression in the same field of study as a previous project.
This form must be accompanied by the previous year's abstract and Research Plan/Project Summary.

Student's Name(s): Ishaan Javali and Shrey Joshi

To be completed by Student Researcher

List all components of the current project that make it new and different from previous research. The information must be on the form; use an additional form for previous year and earlier projects.

Components	Current Research Project	Previous Research Project Year: 2020-2021
1. Title	Applying Mathematical Modeling to Geophysical & Metereological Data for Landslide Analytics & Forecasting	GLAS: A Global Landslide Analytics System
2. Change in goal/purpose/ objective	Goal: • Explore more landslide indicators, additional data sources, and perform landslide date forecasting • Introduce a susceptibility mapping approach with static terrain features • Parse through historical satellite data to find unreported landslides to use as additional training data with less self-report bias	Goal: • Create an initial dataset of global landslide indicators • Train an initial iteration of forecasting models for binary and severity classification models
3. Changes in methodology	Model robustness was demonstrated by analyzing the dataset in two subsets: one containing random non-landslide instances, the other with non-landslide instances from landslide areas (different times). Hyperparameter optimization was improved. A landslide analytics dashboard was created	Data on landslide indicators was collected with Python scripts for landslide and non-landslide events. Landslide forecasting was done by training SVC, KNN, Random Forest, and LSTM models.
4. Variables studied	A Lithology feature was added to the dataset. The focus is on precipitation and slope data and improving the performance of the forecasting models through additional training data found by parsing through historical satellite data to find past unreported landslides.	Elevation, infrastructure presence, forest loss, precipitation, humidity, wind speed, air pressure, temperature data was collected and compiled into the dataset.
5. Additional changes	The dataset consists of more accurate climate data. The model for forecasting the landslide date was improved. A mathematical algorithm was developed based on historical soil moisture and slope data to find unreported landslides to append to our dataset.	NA

Attached are:

☑ Abstract and Research Plan/Project Summary, Year 2020-2021

I hereby certify that the above information is correct and that the current year Abstract & Certification and project display board properly reflect work done only in the current year.

Ishaan Javali and Shrey Joshi	Ishaan Javali	11/01/2021
Student's Printed Name(s)	Signature	Date of Signature

Abstract and Research Plan/Project Summary, Year 2020-2021